Essays on Using Shrinkage Estimators in Econometrics

Dissertation
zur Erlangung des Doktorgrades
Dr. rer. pol.
der Fakultät für Wirtschaftswissenschaften
der Universität Duisburg-Essen

vorgelegt
von
Stephan Johannes Hetzenecker

aus
Sulzbach-Rosenberg, Deutschland

Betreuer:
Prof. Dr. Christoph Hanck
Lehrstuhl für Ökonometrie

Essen, 31.03.2022

Gutachter:
Prof. Dr. Christoph Hanck
Prof. Dr. Florian Heiß
Prof. Dr. Heiko Jacobs

Tag der mündlichen Prüfung:
04.11.2022

DOI: 10.17185/duepublico/77356
URN: urn:nbn:de:hbz:465-20230413-062700-1

Contents

Acknowledgments

First and foremost, I would like to thank my supervisor, Christoph Hanck, for all his support during the completion of this thesis, for the invaluable freedom I was granted in my research, and for his fast feedback and many valuable comments on my projects, which substantially improved every part of this thesis.

I am very grateful to Prof. Dr. Florian Heiß for agreeing to be my second supervisor and for successful cooperation on the first project.

My thanks also go to my friend and co-author Maximilian Osterhaus. Working together with you was one of the most fun things about writing this thesis. I like your humble way of expressing smart thoughts. Moreover, I would like to thank Sascha Alexander Keweloh for the fruitful cooperation and the inspiring discussions about diverse research ideas and projects. Your creativity and passion is impressive.

In addition, I would also like to thank Prof. Toker Doganoglu, Ph.D., for introducing me into the discrete choice literature, for encouraging me to pursue a Ph.D., and for his advice on my first project.

I wrote this thesis while being a student at the Ruhr Graduate School of Economics. I gratefully acknowledge the financial support by the Ruhr Graduate School of Economics during my first three years as a Ph.D. student.

Furthermore, I would like to thank my colleagues at the Ruhr Graduate School of Economics and at the department of Economic at the University of Duisburg-Essen, who created a pleasant and supportive working environment.

Finally, I wish to express my deepest gratitude to my family for their encouragement and support during this time and in all the years before.

1 Introduction

This introduction serves the purpose to argue why all chapters of this thesis fit under the umbrella of shrinkage estimators. To this end, the first part of this introduction briefly motivates the usage of shrinkage estimators and illustrates the underlying key ideas of such estimators. The second part of the introduction summarizes the content and contribution of each chapter in this thesis and outlines the context in which shrinkage methods are used in each chapter.

1.1 Motivation for Using Shrinkage Estimators

The history of shrinkage estimators is closely connected to Charles Stein (cf. Gruber (2017) for a historical survey of shrinkage estimators). In 1956, Stein (1956) discovered a class of estimators whose estimators *can* achieve a lower mean squared error (MSE) than the maximum likelihood estimator, the least squares estimator, and even than the uniformly minimum variance unbiased estimator (Fourdrinier, Strawderman and Wells, 2018). To be precise, he considered the situation where there is only one observation from a multivariate normal distribution with unknown mean vector that is to be estimated and known variance-covariance matrix equal to the identity matrix (without loss of generality). In this case, it seems natural to use the value of the observation as an estimator of its mean. However, Stein (1956) proved that this estimator, which corresponds to the sample average, is not optimal in a mean squared error sense if the dimension of the multivariate normal distribution exceeds two. By combining the components of the multivariate observation in a specific way, we can improve the mean squared error (defined as the sum of the MSE of each component). This seems paradox since the components of the multivariate observation are independent in this example. It is astonishing and strikingly contrary to generally held belief that we can increase estimation accuracy by incorporating seemingly irrelevant information in each component of the multivariate observation (see, e.g, Fourdrinier et al. (2018), Efron and Morris (1977) and Efron (2017)). Stein's result (i) surprised many (Gruber, 2017), (ii) is one of the most influencial, certaintly most controversial results in statistics in the post-war period (Efron, 2017), and (iii), without exaggeration, one of the most discussed in statistiscal theory (Hoffmann, 2000).

As noted by Stein (1956), a shortcoming of his result is that it does not allow for immediate practical application. On the one hand, not all estimators in the discovered class improve the MSE compared to the minimum variance unbiased estimator and, on the other hand, there is no clear guidance on how to obtain those estimators, which

actually have lower MSE, in practice. However, in 1961, Charles Stein and his student Willard James (James and Stein, 1961) sharpened the result of Stein (1956) and derived a class of estimators, whose estimators *always* have smaller MSE than the minimum variance unbiased estimator, and, maybe even more importantly, the best estimator in this class (in terms of MSE). This best estimator is referred to as James-Stein estimator and shifts the estimate "closer" to the true mean vector, by shrinking each component of the observation towards zero (Hoffmann, 2000). In this case and in most applications of shrinkage estimators, zero is the shrinking target (Lehmann and Casella, 2006). If we have prior knowledge that the parameter vector is close to some value other than zero, we can use this value as shrinking target for the James-Stein estimator, without loosing any of the theoretical advantages (Hoffmann, 2000).[1]

By proposing the James-Stein estimator, James and Stein (1961) introduced the first classic shrinkage estimator (Hansen, 2016*a*). The James-Stein estimator and variations have proved useful in different settings such as predicting cure rates for operations at hospitals, outcomes of election nights, batting averages for baseball players, and fire alarm probabilites (see, e.g., Gruber (2017), Efron (2017), and Efron and Morris (1977)) and can reduce prediction errors by 50% and more in the aforementioned examples (Efron, 2017).

An important insight from the James-Stein estimator, which constitutes a biased estimator, is that shrinkage has two opposing effects. First, shrinkage typically leads to biased estimators. Second, it can lower the variance of the resulting estimator and, thereby, the MSE compared to unbiased estimators (Draper and Van Nostrand, 1979). By definition, the MSE of an estimator is equal to the the sum of the variance and the squared bias of the estimator. By tolerating a (small) bias, we can decrease the variance of an estimator to such an extent that the MSE decreases. That is, the James-Stein estimator illustrates that there is a bias-variance tradeoff and that we may prefer biased over unbiased estimates in some circumstances (Hoffmann, 2000).

To demonstrate that an unbiased estimator is not always desirable, Chernoff and Moses (1959) give the following simplified but yet insightful example, which we present in the slightly modified form of Jaynes (2003): A cable company wants to lay a telephone cable across Francisco Bay. Unfortunately, they do not know how much cable they will need and, therefore, they have to estimate the length of the cable. If they overestimate the length of the cable, the financial loss will be proportional to the excess amount of the cable length. However, if they underestimate the length of the cable, the cable end falls into the water and they have to deal with a financial disaster. In this example, we clearly want to overestimate the cable length and using an unbiased estimator could be described as foolhardy, as Jaynes (2003) put it.[2]

[1] In regression analysis, a shrinkage target of zero means that an independent variable has no influence on the dependent variable and, therefore, may be considered as the natural origin for a regression coefficient (Copas, 1983).

[2] Furthermore, Jaynes (2003) note that unbiased estimators can even violate elementary logic. In the following, we describe one of many possible examples where this undesirable feature of unbiased estimators appears. Given a sample of size one and that we observe two events, an unbiased estimator for the expected number of events λ of a Poisson distribution is two. Yet, an unbiased estimator for λ^3

A famous example for a shrinkage estimator, which illustrates the variance-bias trade-off, is the estimation of the variance of a normal distribution from a random sample of size n (Jaynes, 2003). Usually, the sample variance is calculated by one over $n-1$ times the sum of the squared deviations of each observation from the sample mean. This is an unbiased estimator of the population variance. However, using the scaling factor of one over $n+1$ instead of one over $n-1$ shrinks the estimate closer to zero and results in an estimator for the population variance, which is biased but has more than 50% lower MSE than the unbiased estimator. That is, it requires twice the amount of data for the unbiased estimator to obtain the same MSE, highlighting the relative cost which may be associated with an unbiased estimator.

The James-Stein estimator layes out the foundations for modern shrinkage estimation, a major theme in current research (Efron, 2017). In a linear regression context, two well-known shrinkage estimators are ridge regression (Hoerl and Kennard, 1970) and the **L**east **A**bsolute **S**hrinkage and **S**election **O**perator (Tibshirani, 1996). Both methods utilize the aforementioned variance-bias tradeoff and aim to reduce the MSE. Hoerl and Kennard developed ridge regression to address the problematic behavior of the ordinary least squares (OLS) estimator under multicollinearity, where OLS coefficients can have wrong signs, can be sensitive to minor changes in the underlying data, and exhibit high variance, leading to poor predictions (cf. Farrar and Glauber (1967), Vinod (1978), and Hastie, Tibshirani and Friedman (2009)). By adding a penalty on the sum of the squared coefficients to the OLS loss function, ridge regression shrinks the coefficients not only towards zero but also towards each other (Hastie et al., 2009), i.e., it assigns the same coefficients to highly (positively) correlated variables. Thereby, it stabilizes the regression coefficients and improves the precision of the estimates under multicollinearity. Similarly to ridge regression, LASSO adds a penalty to the OLS loss function. However, it penalizes the sum of the absolute value of coefficients. In contrast to ridge regression, which only shrinks coefficients towards zero, the LASSO can estimate certain coefficients to be zero and, therefore, can select variables. Hence, LASSO allows to determine the most important variables, which is helpful for the interpretation of a model.[3] However, the LASSO tends to randomly select only one out of a set of highly correlated variables (Zou and Hastie, 2005). To combine the advantages of ridge regression and LASSO, Zou and Hastie (2005) introduce the elastic net. On the one hand, elastic net can simultaneously estimate and select variables like LASSO. On the other hand, it can deal with highly correlated variables like ridge regression and can select groups of highly correlated variables. Ridge regression, LASSO, elastic net, and extensions are the main shrinkage estimators studied in this thesis.

---

is zero. That is, we observe two events and are still advised to estimate λ to be zero. This is absurd since if λ were zero, it would be impossible to ever observe two events. The (biased) maximum likelihood estimator is 2^3 and does not violate elementary logic.

[3]The James-Stein estimator can also be applied to linear regression models. However, it can neither solve the problems of multicollinearity nor select variables (Hoerl, 2020). Draper and Van Nostrand (1979) provide a review of the James-Stein estimator and ridge regression and Hansen (2016*b*) compares the MSE of the James-Stein estimator and the LASSO.

1.2 Overview of Thesis

This thesis consists of six chapters, including this introduction. The second and third chapter deal with the estimation of unobserved heterogeneity in discrete choice models while the fourth chapter studies the estimation of observed heterogeneity in discrete choice models. The fifth chapter proposes a method to identify and efficiently estimate structural vector autoregressions (SVARs). The sixth chapter concludes.

The second chapter, which is co-authored by Florian Heiss (Heinrich Heine University Düsseldorf) and Maximilian Osterhaus (Heinrich Heine University Düsseldorf), aims to model unobserved heterogeneity across agents in discrete choice models. In contrast to observed heterogeneity, unobserved heterogeneity cannot be linked to observed characteristics of the agents and arises due to different tastes and preferences of agents. To draw valid conclusions from a model, it is crucial to capture unobserved heterogeneity present in the population. A popular approach which takes this into account and allows the coefficients of the economic model to vary across agents is a random coefficients model. Our goal is to estimate the distribution of the random coefficients.

For that purpose, this chapter considers the simple and computationally fast nonparametric estimator of Fox, Kim, Ryan and Bajari (2011), hereafter FKRB. Their estimator approximates the distribution of random coefficients through a fixed number of grid points, representing prespecified types of heterogenous agents. The probability of each type occurring in the population is estimated by constrained least squares. To obtain a valid distribution function, the weights are constrained to be nonnegative and to sum up to one.

We show that these constraints induce unintended shrinkage of the estimated probability weights towards zero. That is, they transform the estimation problem into a special case of the nonnegative LASSO (Wu, Yang and Liu, 2014), explaining the estimator's sparse nature observed in many of its applications. In addition to its sparse nature, the connection to nonnegative LASSO reveals that the estimator may randomly select certain types of agents under strong correlation and, in consequence, that the interpretation of the estimates may be misleading. This property is especially relevant as Fox, Kim and Yang (2016) prove that their estimator identifies the true distribution if sufficiently many types of agents are included in the estimation, i.e., the grid of random coefficients becomes sufficiently dense. However, in practice, a sufficiently dense grid of random coefficients is typically accompanied by high correlation among grid points. Hence, the estimator's sparsity and "random" selection behavior can lead to inaccurate approximations of the true distribution function and have a drastic impact on the identification of the model.

To mitigate the random selection behavior and to allow for more accurate approximations of the true distribution function, we extend the estimator to a special case of the nonnegative elastic net (Wu and Yang, 2014). Since our estimator is a generalization

of the FKRB estimator, approximation accuracy of our estimator is guaranteed to be as least as high as the one of the FKRB estimator. Our theoretical results and Monte Carlo simulations show that our generalized estimator improves the recovery of the support of the random coefficients' distribution of the random coefficients and achieves a more accurate approximation. This chapter is published in the Journal of Econometrics.

Monte Carlo simulations conducted in Chapter 2 reveal that the generalized estimator of the second chapter still tends to shrink the weights too much and yields solution which are too sparse, especially if the grid of random coefficients becomes dense as required for identification. Recognizing this property, Chapter 3, which is single-authored, proposes to use a random elastic net estimator. The random elastic net estimator builds on the generalized estimator of Chapter 2 and is based on the random LASSO estimator developed by Wang, Nan, Rosset and Zhu (2011). It applies a bootstrap procedure which is similar to that of the random forest (Breiman, 2001). More concretely, the random elastic net estimator repeatedly estimates the nonnegative elatic net estimator of the Chapter 2 using only a randomly selected subset of prespecified grid points. The final estimates are obtained by averaging over the estimates across bootstrap repetitions. The key idea is to break a substantial part of the correlation among grid points before each estimation, by randomly selecting a subset of the grid points, and to average out the "random" selection behavior of the nonnegative elatic net estimator by repeating the estimation sufficiently often (with different subsets of the grid points).

Two Monte Carlo simulations presented in this chapter reveal that the random elastic net estimator substantially improves the recovery of the true support of the distribution function compared to FKRB estimator and, more importantly, also to the nonnegative elatic net estimator. The improved recovery of the true distribution's support also translates to more accurate approximations of the true distribution function. The application of the random elastic net estimator to the model developed by Blundell, Gowrisankaran and Langer (2020), who study the gains from dynamic enforcement of air pollution regulations, highlights that the elastic net and random elastic net estimator can recover complicated distribution functions whereas the estimated distribution function of the FKRB estimator does not seem very informative. It estimates only few positive weights and, consequently, approximates a potentially continuous distribution through a step function with only few steps.

The fourth chapter is joint work with Maximilian Osterhaus. In contrast to the preceding two chapters, the fourth chapter aims to model observed heterogeneity in discrete choice models. Observed heterogeneity refers to differences across agents which can be linked to differences in their observed characteristics such as their age, gender, and income. Typically, observed heterogeneity is modeled using parametric functional forms (such as linear interactions). However, relying on such parametric approaches seems restrictive in the era of ever growing data sets. Recently, Farrell, Liang and Misra (2021*a*)

proposed to combine the structure imposed by economic models with the flexibility of deep learning to estimate the expected value of heterogeneous quantities of interest. These quantities of interest are functions of observed characteristics of the agents in the population. The appealing feature of their approach is that, on the one hand, the results remain interpretable and, on the other hand, flexible functional forms of heterogeneity can be recovered. To conduct inference with the deep learning approach, Farrell et al. (2021*a*) build on the influence function approach of Chernozhukov, Chetverikov, Demirer, Duflo, Hansen, Newey and Robins (2018).

However, they do not analyze the impact of shrinkage, which is routinely applied when tuning deep learning methods, on their method. In this chapter, we intend to fill this gap and study the finite sample performance of the approach of Farrell et al. (2021*a*) in the context of discrete choice models. To this end, we conduct several Monte Carlo experiments. These experiments reveal that deep learning generally allows to precisely approximate the expected value of heterogenous parameters and that regular robust standard errors do not lead to valid inferential statements. When no shrinkage is employed, the influence function approach can suffer from substantial bias and large estimated standard errors driven by extreme outliers, resulting from overfitting. Shrinkage reduces overfitting and stabilizes the estimation approach, at the expense of inducing a regularization bias. The regularization bias invalidates the inferential statements obtained from the influence function procedure in our experiments. Unlike shrinkage, repeated sample splitting does not introduce additional bias while also stabilizing the estimation procedure. Consequently, our simulations indicate that repeated sample splitting allows the construction of valid inferential statements.

The fifth chapter, which is co-authored by Sascha Keweloh (Technical University of Dortmund), deals with estimating the simultaneous interaction of multiple time series variables. For that purpose, we consider SVARs which allow to assess how a shock to one variable affects all considered variables simultaneously. For identification of an SVAR, an a priori structure has to be imposed on the model. To this end, traditional approaches restrict the interaction of the variables in specific ways (see, e.g., Sims, 1980). However, oftentimes only some but not all of those restrictions are well-justified, e.g., by economic theory. In contrast, modern data-driven approaches do not restrict the interactions of the variables in the SVAR (see, e.g., Lanne, Meitz and Saikkonen, 2017 and Keweloh, 2021*b*). Rather, they identify the SVAR by solely relying on stochastic properties of the shocks, such as independence and non-Gaussianity. To be precise, the stochastic properties of the shocks imply higher-order moment conditions which allow to identify the SVAR. However, the imposed assumptions on the stochastic properties of the shocks – if correct – may still place challenges on finite sample estimation. For instance, higher-order moment conditions may be volatile in finite samples and lead to imprecise estimates.

We propose to combine block-recursive restrictions with higher-order moment conditions. Our approach exploits only the restrictions on the interactions that are well-justified

and relies only on those stochastic properties of the shocks that are additionally needed for identification. That said, our proposed block-recursive SVAR nests both the traditional approach based on restrictions for the interaction of the shocks and the data-driven approach based on non-Gaussian shocks as limiting cases. For a prespecified block-recursive structure, we derive identifying and overidentifying moment conditions. We prove that these overidentifying moment conditions can increase the asymptotic efficiency of the SVAR estimator. However, the number of overidentifying moment conditions increases quickly when the number of variables in the SVAR increases. As a result, many moment conditions can cause problems for finite sample estimation. A further problem is that, in practice, some moment conditions included in the estimation may be invalid, which would lead to inconsistent estimates.

To exploit potential efficiency gains of overidentifying moment conditions in finite samples and to safeguard against invalid moment conditions, we use a LASSO-type GMM estimator (Cheng and Liao, 2015) which does not shrink coefficients – as usually done by LASSO – but moment conditions to zero. Our LASSO-type SVAR GMM estimator consistently selects only relevant and valid overidentifying moment conditions and refrains from selecting invalid or redundant moment conditions. Furthermore, the LASSO-type SVAR GMM estimator is asymptotically normal. A Monte Carlo experiment and an application to oil market data illustrate the improved performance and the relevance of the proposed LASSO-type GMM estimator.

In summary, each of the four main chapters applies modern shrinkage methods and highlights that shrinkage estimators prove useful in various fields. In particular, while Chapters 2 – 4 apply shrinkage estimators in discrete choice models, Chapter 5 applies them in a time series context. In all chapters, we demonstrate the increased performance of our estimators using simulation methods. In Chapters 2 and 5, the simulations support the conclusions drawn from our theoretical analysis of the estimators. Readers will note that some aspects are mentioned more than once. These repetitions ensure that each chapter can be read independently.

2 Nonparametric Estimation of the Random Coefficients Model: An Elastic Net Approach

Co-authored by Florian Heiß and Maximilian Osterhaus

Abstract

This paper investigates and extends the computationally attractive nonparametric random coefficients estimator of Fox, Kim, Ryan and Bajari (2011). We show that their estimator is a special case of the nonnegative LASSO, explaining its sparse nature observed in many applications. Recognizing this link, we extend the estimator, transforming it into a special case of the nonnegative elastic net. The extension improves the estimator's recovery of the true support and allows for more accurate estimates of the random coefficients' distribution. Our estimator is a generalization of the original estimator and therefore, is guaranteed to have a model fit at least as good as the original one. A theoretical analysis of both estimators' properties shows that, under conditions, our generalized estimator approximates the true distribution more accurately. Two Monte Carlo experiments and an application to a travel mode data set illustrate the improved performance of the generalized estimator.

JEL codes:	*C14, C25, L.*
Keywords:	*Random Coefficients, Mixed Logit, Nonparametric Estimation, Elastic Net.*
Publication status:	*Forthcoming in the Journal of Econometrics. Available online: https://doi.org/10.1016/j.jeconom.2020.11.010.*

2.1 Introduction

Adequately modeling unobserved heterogeneity across agents is a common challenge in many empirical economic studies. A popular approach to address unobserved heterogeneity is the random coefficients model, which allows the coefficients of the economic model to vary across agents. The aim of the researcher is to estimate the distribution of the random coefficients.

Fox et al. (2011), hereafter FKRB, propose a simple and computationally fast estimator that can approximate distributions of any shape. The estimator uses a fixed grid where every grid point is a prespecified vector of random coefficients. The distribution function is obtained from the probability weights at the grid points, which are estimated with constrained least squares. In principle, the approach can approximate any distribution arbitrarily closely if the grid of random coefficients is sufficiently dense (McFadden and Train, 2000).

Applications of the estimator indicate, however, that it tends to estimate only few positive weights and, thus, sets the weights at many grid points to zero. As a consequence, the estimator lacks the ability to estimate smooth distribution functions but instead approximates potentially continuous distributions through step functions with only few steps. Our first contribution is to show that the estimator of FKRB is Nonnegative LASSO (Wu et al., 2014) (NNL) with a fixed tuning parameter to explain its sparse nature.

NNL, which was first mentioned in the seminal work of Efron, Hastie, Johnstone, Tibshirani and Others (2004) as positive LASSO, is a popular model selection method typically used in applications with supposedly sparse models. It is applied in various research fields, e.g., in vaccine design (Hu, Follmann and Miura, 2015), nuclear material detection (Kump, Bai, Chan, Eichinger and Li, 2012), document classification (El-Arini, Xu, Fox and Guestrin, 2013), and index tracking in stock markets (Wu et al., 2014). NNL shares the property of LASSO (Tibshirani, 1996) that it regularizes the coefficients of the model and shrinks some to zero. This property is observed for the FKRB estimator in different Monte Carlo studies (e.g., Fox et al., 2011 and Fox et al., 2016) and applications to real data (e.g., Nevo, Turner and Williams, 2016, Illanes and Padi, 2019, Blundell et al., 2020 and Houde and Myers, 2019). Nevo et al. (2016) study the demand for residential broadband and estimate that there are only 53 out of 8626 potentially heterogeneous consumer types. Illanes and Padi (2019) use the approach to estimate the demand for private pension plans in Chile and assign positive weights to only 194 of 83,251 grid points. Blundell et al. (2020) analyze firms' reaction to the regulation of air pollution and recover no more than 12 of the 10,001 potential points.

In addition to its sparse nature, the connection of the FKRB estimator to NNL reveals the estimator's potentially incorrect selection of grid points under strong correlation. The estimator "randomly" selects one out of a group of highly correlated points and sets the remaining weights to zero (see Zou and Hastie, 2005, and Hastie et al., 2009, for the

random behavior of LASSO).

The estimator's sparsity and "random" selection behavior can cause inaccurate approximations of the true distribution through non-smooth distributions with the estimated support possibly deviating from the true distribution's support. The latter can lead to misleading conclusions with respect to the heterogeneity of agents in the population. Fox et al. (2016) prove that the estimator identifies the true distribution if the grid of random coefficients becomes sufficiently dense. However, in practice, the correlation tends to increase with the density of the grid and can become so strong that the optimization problem to the FKRB estimator cannot be solved due to singularity (Nevo et al., 2016, Online Supplement). Therefore, the high correlation of a dense grid in combination with the incorrect grid point selection of the estimator under strong correlation can have a drastic impact on the identification of the model.

Our second contribution is to provide a generalization of the FKRB estimator that is able to accurately approximate continuous distributions even under strong correlation. Recognizing the link to NNL, we add a quadratic constraint on the probability weights. The constraint transforms the estimator to a special case of nonnegative elastic net (Wu and Yang, 2014). The extension mitigates the sparsity and improves the selection of the grid points. Due to the additional flexibility that is introduced with the extension, the estimator adjusts to the degree of correlation among grid points. Note that our generalization always includes the FKRB estimator as a special case such that the model fit cannot be worse for our estimator than the FKRB estimator.

We show theoretically, under conditions, that our estimator provides more accurate estimates of the true underlying distribution. For that purpose, we derive the selection consistency and an error bound on the estimated distributions. The analysis of the selection consistency examines the estimator's ability to estimate positive probability weights at grid points that lie inside the true distributions support, and zero weights at points outside the true support. The selection consistency is necessary to approximate the true distribution as accurately as possible. Since the estimated distribution recovers the existing heterogeneity in the population, i.e., agents' varying preferences, recovering the true support points is also important for the correct interpretation of the model.

The analysis reveals that our generalized estimator correctly selects the grid points under less restrictive conditions than the FKRB estimator. The error bounds on the estimated distribution functions illustrate the positive impact of our extension on the overall approximation accuracy. Two Monte Carlo experiments in which we estimate a random coefficients logit model confirm the superior properties of our generalized estimator.

Other nonparametric estimators for the random coefficients model include Train (2008), Train (2016), Burda, Harding and Hausman (2008) and Rossi, Allenby and McCulloch (2012). Train (2008) introduces three estimators that are, in principle, similar to the general approach of FKRB but employ a log-likelihood criterion instead of constrained least squares. Train (2016) suggests approximating the random coefficients'

distribution with polynomials, splines or step functions instead of with a fixed grid of preference vectors. The approach substantially reduces the number of required grid points if the researcher specifies overlapping splines and step functions. Due to the lower number of required grid points, the approach reduces the curse of dimensionality, which is a shortcoming of the fixed grid approach if the economic model includes a large number of random coefficients. However, Train (2008) estimates the respective model with the EM algorithm, which is sensitive to its starting values and is not guaranteed to converge to a global optimum, and Train (2016) uses simulated log-likelihood for the estimation. Burda et al. (2008) and Rossi et al. (2012) employ a Bayesian hierarchical model to approximate the random coefficients' distribution with a mixture of Normal distributions. Even though the estimator potentially has better finite sample properties, it uses a Markov Chain Monte Carlo technique with a multivariate Dirichlet Process prior on the coefficients, which is computationally more demanding.

The remainder of the paper is organized as follows. Section 2.2 describes the FKRB estimator and introduces our generalized version. Section 2.3 derives the condition on the estimators' sign consistency and an error bound on the estimated distribution functions. Section 2.4 presents two Monte Carlo experiments that investigate the performance of our generalized estimator in comparison to the FKRB estimator. Section 2.5 applies the estimators to the *Mode Canada* data set from the R package *mlogit* (Croissant, 2019). Section 2.6 concludes and provides an outlook.

2.2 Fixed Grid Estimators

To introduce our estimator, we consider the framework of a random coefficient discrete choice model. The approach, however, is not restricted to discrete choice models but can be applied to any model with unobserved heterogeneous parameters. Let there be an i.i.d. sample of N observations, each confronted with a set of J mutually exclusive potential outcomes. The researcher observes a K-dimensional real-valued vector of explanatory variables $x_{i,j}$ for every observation unit i and potential outcome j, and a binary vector y_i whose entry $y_{i,j}$ is equal to one whenever she observes outcome j for the ith observation, and zero otherwise. The goal is to estimate the unknown distribution of heterogeneous parameters $F_0(\beta)$ in the model

$$P_{i,j}(x) = \int g(x_{i,j}, \beta)\, dF_0(\beta) \tag{2.1}$$

where $g(x_{i,j}, \beta)$ denotes the probability of outcome j conditional on the random coefficients β and covariates $x_{i,j}$. The researcher specifies the functional form of $g(x_{i,j}, \beta)$. A prominent example of Equation (2.1) is the multinomial mixed logit model, the state-of-the-art model for demand estimation. For a detailed description of the multinomial mixed logit see Train (2009, pp. 134–150). In this model, consumer i realizes utility $u_{i,j} = x_{i,j}^T \beta_i + \omega_{i,j}$ from alternative j, given product characteristics $x_{i,j}$ and unobserved consumer-specific preferences β_i. $\omega_{i,j}$ denotes an additive, consumer- and choice-specific

error term. Consumer i chooses alternative j of J alternatives (and an outside good with utility $u_{i,0} = \omega_{i,0}$) if $u_{i,j} > u_{i,l}$ for all $l \neq j$. Under the assumption that $\omega_{i,j}$ follows a type I extreme value distribution, the unconditional choice probabilities, $P_{i,j}(x)$, are of the form

$$P_{i,j}(x) = \int \frac{\exp\left(x_{i,j}^T \beta\right)}{1 + \sum_{l=1}^{J} \exp\left(x_{i,l}^T \beta\right)} dF_0\left(\beta\right). \tag{2.2}$$

$F_0(\beta)$ represents the distribution of heterogeneous consumer preferences in the population and is to be estimated.

2.2.1 Fixed Grid Estimator by FKRB

In most applications, researchers place restrictive assumptions on the functional form of $F_0(\beta)$ in advance, and estimate its parameters from the data. FKRB propose a simple and fast mixture approach to estimate the underlying random coefficients' distribution without restrictive assumptions on its shape. The estimator is a special case of sieve estimators (Chen, 2007). It uses a finite and fixed grid of random coefficient vectors as mixture components to construct the distribution from the estimated probability weight of every component. The underlying idea of this fixed grid estimator is the transformation of the unconditional choice probabilities in Equation (2.1) into a probability model in which $F_0(\beta)$ enters linearly. FKRB derive the linear probability model in two steps: they transform Equation (2.1) into a regression model with the random coefficients' distribution as the only unknown term. Adding $y_{i,j}$ to both sides and moving $P_{i,j}$ to the right results in the probability model

$$y_{i,j} = \int g\left(x_{i,j}, \beta\right) dF_0\left(\beta\right) + \left(y_{i,j} - P_{i,j}\left(x\right)\right). \tag{2.3}$$

To exploit linearity in parameters, they use a sieve space approximation to the infinite-dimensional parameter $F_0(\beta)$. The sieve space approximation divides the support of the random coefficients β into R fixed vectors. Each vector has length K, the number of random coefficients included in the model. The location of these vectors is specified by the researcher. With the sieve space approximation, Equation (2.3) becomes a simple linear probability model with unknown parameters $\theta = (\theta_1, \ldots, \theta_R)^T$

$$y_{i,j} \approx \sum_{r=1}^{R} g\left(x_{i,j}, \beta_r\right) \theta_r + \left(y_{i,j} - P_{i,j}\left(x\right)\right) \tag{2.4}$$

where $g(x_{i,j}, \beta_r)$ denotes the conditional choice probability evaluated at grid point r. Given the fixed grid of random coefficients, $\mathcal{B}_R = (\beta_1, \ldots, \beta_R)$, the researcher estimates the probability weight θ_r at every point $r = 1, \ldots, R$. The linear relationship between the outcome variable and the unknown parameters θ allows to estimate the mixture weights with the least squares estimator. The linear regression, which regresses the binary dependent variable $y_{i,j}$ on the choice probabilities evaluated at $\mathcal{B}_R$, in total has NJ obser-

vations, J "regression observations" for every statistical observation unit $i = 1, \ldots, N$ and R covariates $z_{i,j} = (g(x_{i,j}, \beta_1), \ldots, g(x_{i,j}, \beta_R))$. By the definition of choice probabilities, the expected value of the composite error term $y_{i,j} - P_{i,j}(x_{i,j})$ conditional on $x_{i,j}$ is zero. Thus, the regression model satisfies the mean-independence assumption of the least squares approach (Fox et al., 2011).

The estimator of the random coefficients' joint distribution is constructed from the estimated weights

$$\hat{F}(\beta) = \sum_{r=1}^{R} \hat{\theta}_r \, 1[\beta_r \leq \beta],$$

where β is an evaluation point chosen by the researcher and the indicator function $1[\beta_r \leq \beta]$ is equal to one whenever $\beta_r \leq \beta$, and zero otherwise.

To ensure that $\hat{F}(\beta)$ is a valid distribution function, FKRB suggest estimating the weights with the least squares estimator subject to the constraints that the weights are nonnegative, and sum to one

$$\begin{aligned} \hat{\theta}^{FKRB} &= \arg\min_{\theta} \frac{1}{2NJ} \sum_{i=1}^{N} \sum_{j=1}^{J} \left(y_{i,j} - \sum_{r=1}^{R} \theta_r z_{i,j}^r \right)^2 \\ \text{s.t.} \quad & \theta_r \geq 0, \quad r = 1, \ldots, R, \quad \text{and} \quad \sum_{r=1}^{R} \theta_r = 1. \end{aligned} \tag{2.5}$$

Key to an accurate approximation of $F_0(\beta)$ is the precise estimation of the probability weights at every grid point. Basis to a precise estimation of the probability weights is the consistent selection of the relevant grid points. This requires the constrained least squares estimator to estimate positive weights at all grid points at which $F_0(\beta)$ has a positive probability mass, and zero weights otherwise. While zero weights at grid points inside $F_0(\beta)$'s support cause inaccurate approximations through step functions with only few steps, positive estimates at grid points outside $F_0(\beta)$'s support lead to unreliable estimates of the random coefficients' distribution.

2.2.2 Nonnegative LASSO vs. Nonnegative Elastic Net

To provide a more accurate non-parametric estimator with similar computational advantages, we suggest a simple generalization of the FKRB estimator. Our adjusted version includes the baseline estimator as a special case but allows for smoother estimates of $F_0(\beta)$ when necessary. To derive our estimator, we extend the optimization problem formulated in Equation (2.5) by a constraint on the sum of the squared probability weights. This additional constraint provides a straightforward way to mitigate the estimator's sparse nature. Our generalized estimator is still simple and computationally fast.

2.2.2.1 Connection to Nonnegative LASSO

We first illustrate the source of the FKRB estimator's sparsity, which helps to understand its behavior and the intuition behind our extension.

One explanation of the potential sparsity of the estimates is the effect of the nonnegativity constraint. Slawski and Hein (2013) show that nonnegative least squares estimators exhibit a self-regularizing property that yields sparse solutions. The FKRB estimator restricts the weights not only to be nonnegative but also to sum up to one. Taking both constraints into account, we recognize that the FKRB estimator is a special case of the nonnegative LASSO (NNL) (Wu et al., 2014).

To show the relation of the FKRB estimator to NNL, we transform the equality constrained problem formulated in Equation (2.5) into its inequality constrained form. The constraint that the probability weights sum to one allows us to reparametrize the optimization problem in terms of $R-1$ instead of R unknown parameters. Without loss of generality, one can rewrite the Rth weight as $\theta_R = 1 - \sum_{r=1}^{R-1} \theta_r$. Substituting θ_R in Equation (2.4) with $1 - \sum_{r=1}^{R-1} \theta_r$ gives the inequality constrained optimization problem

$$\begin{aligned} \hat{\theta}^{\text{FKRB}} &= \arg\min_{\theta} \frac{1}{2NJ} \sum_{i=1}^{N} \sum_{j=1}^{J} \left(\tilde{y}_{i,j} - \sum_{r=1}^{R-1} \theta_r \tilde{z}_{i,j}^r \right)^2 \\ &\text{s.t.} \quad \theta_r \geq 0, \quad r = 1, \ldots, R-1, \quad \text{and} \quad \sum_{r=1}^{R-1} \theta_r \leq 1 \end{aligned} \tag{2.6}$$

where $\tilde{y}_{i,j} = y_{i,j} - z_{i,j}^R$ and $\tilde{z}_{i,j}^r = z_{i,j}^r - z_{i,j}^R$ for every $r = 1, \ldots, R-1$. Because Equation (2.6) is an equivalent form of the optimization problem in Equation (2.5), the objective functions are minimized by the same vector of probability weights. The only difference in the inequality constrained problem is the estimation of the Rth weight, which is calculated after optimization as $\hat{\theta}_R = 1 - \sum_{r=1}^{R-1} \hat{\theta}_r$, and is not explicitly part of the optimization. By the constraints $\theta_r \geq 0, r = 1, \ldots, R-1$, and $\sum_{r=1}^{R-1} \theta_r \leq 1$, the Rth weight satisfies the property of a probability weight, $1 \geq \theta_R \geq 0$.

Comparing the FKRB estimator's transformed optimization problem with that of the NNL applied to the linear probability model formulated in Equation (2.4),

$$\begin{aligned} \hat{\theta}^{\text{NNL}} &= \arg\min_{\theta} \frac{1}{2NJ} \sum_{i=1}^{N} \sum_{j=1}^{J} \left(\tilde{y}_{i,j} - \sum_{r=1}^{R-1} \theta_r \tilde{z}_{i,j}^r \right)^2 \\ &\text{s.t.} \quad \theta_r \geq 0, \quad r = 1, \ldots, R-1, \quad \text{and} \quad \sum_{r=1}^{R-1} \theta_r \leq c, \end{aligned}$$

reveals that the baseline estimator is a special case of NNL with fixed tuning parameter $c = 1$. The constraint that the probability weights sum to one resembles an ℓ_1 penalty that regularizes the parameter estimates and shrinks some weights to zero if the sum of

unrestricted weights exceeds one.

The amount of regularization depends on the size of the unrestricted estimates. The more the sum of the $R-1$ unconstrained weights in Equation (2.6) exceeds one, the stronger the shrinkage imposed by the constraint, and the larger the number of potential zero weights (see, e.g., Hastie et al., 2009, p. 69, for the effect of the LASSO tuning parameter). According to Wu et al. (2014), NNL can result in very sparse models if the constraint is too restrictive. If the sum of the $R-1$ unconstrained weights is less than or equal to one, the constraint has no effect, and the estimated coefficients correspond to the nonnegative least squares solution.

In addition to its sparse nature, the relation to NNL reveals that the FKRB estimator exhibits a "random" selection behavior among grid points. Just like NNL, the estimator has no unique solution when the correlation among choice probabilities evaluated at $\mathcal{B}_R$ is strong. It tends to select one out of a group of highly correlated grid points at random and estimates the weights of the remaining grid points to zero (see Zou and Hastie, 2005, and Hastie et al., 2009, for the random behavior of LASSO).

The correlation is particularly strong in a dense grid among neighboring grid points which is why the random selection behavior becomes more severe if the number of grid points increases. The reason for the strong correlation in dense grids can be explained by the calculation of the regressor matrix $\tilde{Z} = (\tilde{z}^1, \ldots, \tilde{z}^{R-1})$: For every row in $\tilde{Z}$, the column entries are calculated with the same vector of characteristics $x_{i,j}$ and the only term that differs across columns is the vector of random coefficients β^r. If the grid becomes dense, the difference between the neighboring random coefficient vectors vanishes and the corresponding column entries for every row in $\tilde{Z}$ are evaluated at almost exactly the same point. As a consequence, $\tilde{Z}^T\tilde{Z}$ is at best near-singular if the number of grid points R approaches infinity. This contradicts the requirement of a dense grid for accurate approximations of $F_0(\beta)$ (Fox et al., 2016).

2.2.2.2 Elastic Net Estimator

Extending the FKRB estimator's optimization problem formulated in Equation (2.6) by a quadratic constraint on the probability weights alleviates the sparse nature and random selection behavior. The additional constraint is known from ridge regression (Hoerl and Kennard, 1970) and transforms the FKRB estimator into the nonnegative elastic net (Wu and Yang, 2014) with fixed constraint on the ℓ_1-penalty. Thus, our adjusted estimator minimizes

$$\begin{aligned} \hat{\theta}^{\text{ENET}} &= \arg\min_{\theta} \frac{1}{2NJ} \sum_{i=1}^{N}\sum_{j=1}^{J} \left(\tilde{y}_{i,j} - \sum_{r=1}^{R-1} \theta_r \tilde{z}_{i,j}^r\right)^2 \\ \text{s.t.} \quad \theta_r \geq 0, \quad r = 1, \ldots, R-1, \quad &\text{and} \quad \sum_{r=1}^{R-1} \theta_r \leq 1 \quad \text{and} \quad \sum_{r=1}^{R-1} \theta_r^2 \leq t \end{aligned} \tag{2.7}$$

where t is a nonnegative tuning parameter specified by the researcher. Having a linear and quadratic constraint on the probability weights ensures a more reliable selection of grid points: the quadratic constraint encourages a grouping effect, which allows us to recover highly correlated points inside the true support of $F(\beta)$ together and, hence, reduces the estimator's sparsity. The linear constraint, in turn, retains the LASSO property, which makes it possible to select weights inside the support of the true distribution and to estimate zero weights at points outside the true support (Zou and Hastie, 2005).

In addition to the improved selection consistency, our theoretical findings in Section 2.3 show that the quadratic constraint has the desirable property that it allows the specification of a substantially finer grid of random coefficients. While the FKRB estimator runs into almost perfect collinearity problems if the grid becomes finer (Fox et al., 2016), the quadratic constraint ensures that the optimization problem for our adjusted estimator always has a solution. The non-sparse solutions together with the possibility of a finer grid endow our estimator with the ability to provide more accurate and reliable estimated distribution functions.

When implementing the estimator in common statistical software (e.g., R, MATLAB), many quadratic optimization routines only allow for linear constraints. In order to incorporate the constraint on the sum of squared probability weights into these routines, consider the Lagrangian version of our generalized estimator in Equation (2.7)

$$\hat{\theta}^{\text{ENET}} = \arg\min_{\theta} \frac{1}{2NJ} \sum_{i=1}^{N} \sum_{j=1}^{J} \left(\tilde{y}_{i,j} - \sum_{r=1}^{R-1} \theta_r \tilde{z}_{i,j}^r \right)^2 + \frac{1}{2} \mu \sum_{r=1}^{R-1} \theta_r^2 + \lambda \left(\sum_{r=1}^{R-1} \theta_r - 1 \right) - \sum_{r=1}^{R-1} \nu_r \theta_r. \tag{2.8}$$

The first term in Equation (2.8) is the least squares objective function that minimizes the sum of squared residuals. The second term corresponds to the constraint on the sum of squared probability weights where $\mu \geq 0$ is the equivalent counterpart to t in Equation (2.7). The third and fourth terms with their nonnegative Lagrange multipliers λ and ν_r, $r = 1, \ldots, R-1$, enforce the constraints that the estimated weights sum to one and that they are nonnegative, respectively. λ and ν_r, $r = 1, \ldots, R-1$, are endogenously determined by the system through the formulation of the linear constraints. In particular, λ corresponds to an endogenous LASSO parameter. Adding the second term to the first term in Equation (2.8) transforms the loss function such that we can use quadratic optimization routines. The third and fourth terms can be supplied as linear constraints as stated in Equation (2.7) to these routines.

The tuning parameter μ is specified by the researcher before the optimization commences. It relates to t in opposite direction: large values of μ imply small values of t. The larger the value of the tuning parameter μ, the stronger is the penalty on the sum of squared probability weights, and, hence, the smaller is t. For every μ, there exists a t such that the estimated weights in Equation (2.8) and Equation (2.7) are the same (Hastie et al., 2009, p. 63).

The specification of the tuning parameter μ allows adjusting the estimator to the level of correlation among grid points. Larger (smaller) values of μ (t) give more weight to the quadratic constraint, which enables the joint recovery of grid points if the correlation is strong and, hence, reduces the sparsity of the estimator.

The specification of the tuning parameter μ allows adjusting the estimator to the level of correlation among grid points. Larger (smaller) values of μ (t) give more weight to the quadratic constraint, which enables the joint recovery of grid points if the correlation is strong and, hence, reduces the sparsity of the estimator. For increasing (decreasing) values of μ (t), the estimator shrinks the probability weights of highly correlated grid points toward each other and induces an averaging of the estimated weights. For $\mu = 0$ (any $t \geq 1$), the quadratic constraint does not bind, such that the adjusted estimator simplifies to the baseline estimator. Therefore, our estimator is a generalization of the FKRB estimator given in Equation (2.6), including it as a special case.

Based on our Monte Carlo experiments, we recommend choosing the tuning parameter μ with cross-validation and the one standard error rule based on the mean squared error (MSE) criterion. This approach ensures that our estimator achieves a model fit that is at least as high as the FKRB estimator's. If the model fit is highest for $\mu = 0$ ($t \geq 1$), the outcome of our generalized estimator is the same as that for the FKRB estimator, while it performs better if the model fit is highest for some $\mu > 0$ ($t < 1$). For decreasing values of t, the estimator shrinks the probability weights of highly correlated grid points toward each other and induces an averaging of the estimated weights.

The theoretical analysis in Section 2.3 and the Monte Carlo studies in Section 2.4 indicate that the improved selection property of our generalized estimator leads to more precise estimates of the probability weights. If the linear constraint on the sum of the probability weights is strictly binding, i.e., if the sum of unconstrained nonnegative weights is larger than one, the FKRB estimator leads to biased estimates of the probability weights. This follows from its equivalence to NNL (see, e.g., Hastie et al., 2009, p. 91). In comparison to the unconstrained solution, the estimator shrinks the weights at some grid points to zero despite the potential positive probability mass of $F_0(\beta)$ at these points. Due to the constraint that the estimated weights sum to one, the incorrect zero weights lead to downward biased estimates at points with positive weights. The FKRB estimator reallocates the probability mass from the points with incorrect zero weights to other points, which imposes an upward bias at these points.

The quadratic constraint potentially reduces the described distortions through its improved selection consistency. As a result of more correct positive probability weights, the quadratic constraint diminishes the reallocation of probability caused by the linear constraint and, therefore, reduces the bias both at points with incorrect zero weights and positive weights.

Remark 2.1. Our generalized estimator can be extended to a generalized least-squares and smooth basis densities version of our estimator analogous to Fox et al. (2011).[1] Furthermore, the proposed elastic net version is not the only possible way to address the sparse nature of the FKRB estimator. These extensions have to fit into the framework that the estimated probability weights are nonnegative and sum to one, which, e.g., excludes the adaptive LASSO (Zou, 2006) and post selection estimators. Among the suitable extensions, we considered the Factor-Adjusted Regularized Model Selection (FarmSelect) (Fan, Ke and Wang, 2020) and the nonnegative version of the S-LASSO (Hebiri and van de Geer, 2011).
FarmSelect is a LASSO extension that addresses highly correlated covariates. The underlying idea of the approach is the decorrelation of covariates via a factor model with few latent factors. In our context, Farm-Select requires the choice probabilities to follow an approximate factor model. S-LASSO is a different variant of the elastic net that uses a ℓ_2-fusion penalty, $\lambda \sum_{r=1}^{R-1} \theta_r + \mu \sum_{r=2}^{R-1} (\theta_r - \theta_{r-1})^2$, which penalizes the squared difference of neighboring probability weights. The penalty helps to smooth the solution which makes it particularly suitable for the estimation of continuous distributions.
Monte Carlo simulations suggest that S-LASSO is a promising alternative to the elastic net estimator.[2] Compared to the elastic net extension, the S-LASSO imposes additional restrictions on the shape of the distribution. We believe that the elastic net extension may be the most intuitive approach.

2.3 Theoretical Analysis of the Estimators' Properties

The requirement of a sufficiently fine grid, which potentially includes points outside the true support, transforms the fixed grid estimator into a high dimensional regression problem with potentially sparse solutions and highly correlated covariates. Recall that in such a context, an important element of an accurate estimation of $F_0(\beta)$ is the consistent selection of grid points. It guarantees the correct recovery of $F_0(\beta)$'s support, and therefore, is crucial to accurate estimation of the probability weights. In Subsection 2.3.1, we study both estimators' ability to select the correct weights. To evaluate the overall approximation accuracy of the estimators presented in Section 2.2, we derive an error bound for the estimated probability weights and the estimated distribution functions in Subsection 2.3.2.

We show that our generalized estimator is selection consistent under less restrictive conditions on the design matrix. While the estimator of FKRB is less likely to be selection consistent if the number of grid points becomes large (and hence, the correlation strong),

[1] The extensions adjust the calculation of the sum of squared residuals. For the generalized least-squares version, each observation is weighted to address the heteroscedasticity. The smooth basis densities estimator uses pre-specified parametric distributions instead of fixed random coefficient vectors to simulate the choice probabilities. The estimated probability weights denote the weight of every parametric distribution. For a more detailed description see Fox et al. (2011).

[2] The results are available from the authors on request.

the generalized estimator can satisfy the condition through an appropriate choice of the tuning parameter μ. Similarly, compared to the derived error bounds for the FKRB estimator, the error bounds for the generalized estimator can be decreased through the choice of the tuning parameter μ.

Due to the relation of the estimators to the NNL and nonnegative elastic net, respectively, we build on the literature on regularized regression. Our proof of the selection consistency mainly follows Jia and Yu (2010), who analyze selection consistency of the elastic net under i.i.d. Gaussian errors. Similarly to Jia and Yu (2010), Wu et al. (2014) and Wu and Yang (2014) derive selection consistency of the nonnegative LASSO and the nonnegative elastic net for i.i.d. Gaussian errors. We extend their proof to sub-Gaussian errors and allow for correlation among the J errors that belong to the same observation unit i. Thereby, we additionally contribute to the literature on the nonnegative elastic net. Neither Jia and Yu (2010) nor Wu and Yang (2014) calculate error bounds on the deviation between the estimated and the true coefficients. Our proof of the error bound on the estimated weights is drawn from Takada, Suzuki and Fujisawa (2017), who analyze a generalization of the elastic net. We modify their proof such that it is in line with the probability model in Section 2.2.

In line with Fox et al. (2016) and in addition to the tuning parameter μ, we also treat the specification of the grid points as tuning parameters specified by the researcher. In particular, we allow the number of grid points $R(N)$ to depend on the sample size N. That is, the larger N, the more grid points $R(N)$ can be included into the grid. To keep notation uncluttered, we drop the dependence on N and write R instead of $R(N)$ where not relevant in the subsequent analyses.

Suppose $\theta^* = (\theta_1^*, \dots, \theta_{R-1}^*)^T$ specifies the vector of probability weights that yields the most accurate discrete approximation, $F^*(\beta) = \sum_{r=1}^{R} \theta_r^* \mathbf{1}[\beta_r \leq \beta]$ with $\theta_R^* = 1 - \sum_{r=1}^{R-1} \theta_r^*$, of $F_0(\beta)$ which can be obtained with the estimators for a given grid $\mathcal{B}_R$.[3] In the following, the introduction of $F^*(\beta)$ allows us to study the selection consistency and the distance between $\hat{\theta}$ and θ^* for any number of grid points R. In addition, we use $F^*(\beta)$ as a benchmark to compare the estimated distribution function, $\hat{F}(\beta) = \sum_{r=1}^{R} \hat{\theta}_r \mathbf{1}[\beta_r \leq \beta]$ with $\hat{\theta}_R = 1 - \sum_{r=1}^{R-1} \hat{\theta}_r$, to the true underlying distribution $F_0(\beta)$. Fox et al. (2016) show that, under some regularity conditions, it holds that $|F_0(\beta) - F^*(\beta)| = O(R^{-\bar{s}/K})$ where $\bar{s} \geq 0$ measures the degree of smoothness of $F_0(\beta)$[4]

[3] For instance, the best discrete approximation θ^* can be chosen such that it minimizes the MSE of the true distribution and its best discrete approximation over all grid points. If the true distribution is continuous with density $f_0(\beta_r)$, θ_r^* can be calculated as the normalized weighted density at grid point β_r for $r = 1, \dots, R-1$, i.e., $\theta_r^* = w(\beta_r) f_0(\beta_r) / \left(\sum_{r=1}^{R-1} w(\beta_r) f_0(\beta_r) \right)$. E.g., the weights $w(\beta_r)$ can be obtained by quadrature methods (cf. Fox et al., 2016, Lemma 1). If the true distribution is discrete and the grid for the estimation includes the true mass points, θ^* corresponds to the probability mass of the true distribution at every point and the fixed grid estimator can, in principle, recover the true distribution without approximation error. Our subsequent results do not rely on the way the weights θ^* are calculated and hold for continuous and discrete true distributions.

[4] The density function of β is assumed to be $\bar{s}$-times continuously differentiable.

and K refers to the number of random coefficients. Thus, the difference of $F_0(\beta)$ and $F^*(\beta)$ becomes negligibly small for R going to infinity.

In order to analyze the selection consistency and to derive the error bounds on the estimated weights and distribution functions, we use the Lagrangian formulation of our generalized estimator stated in Equation (2.8). We exploit the structure of our data and make the following assumptions on the linear probability model corresponding to $F^*(\beta)$

$$y_{i,j} = \sum_{r=1}^{R} \theta_r^* z_{i,j}^r + \epsilon_{i,j}, \tag{2.9}$$

where $\epsilon_{i,j}$ is the linear probability error and $\theta_R^* = 1 - \sum_{r=1}^{R-1} \theta_r^*$, and on the data generating process.

Assumption 2.1.

(i) $\Big(\epsilon_i = (\epsilon_{i,1}, ..., \epsilon_{i,J})\Big)_{i=1}^{N}$ *are independent.*

(ii) $\epsilon_{i,j}$ *is sub-Gaussian:* $\mathbb{E}\left[\exp\left(t\epsilon_{i,j}\right)\right] \leq \exp\left(\frac{\sigma^2 t^2}{2}\right) \quad (\forall t \in \mathbb{R})$ *for* $\sigma > 0$.

(iii) $(\tilde{Z}_i)_{i=1}^{N}$ *are i.i.d. with a density bounded from above and each* $\tilde{z}_{i,j}^r \in [-1, 1]$.

(iv) $\mathbb{E}\left[\epsilon_i | \tilde{Z}_1, ..., \tilde{Z}_N\right] = 0$.

$\tilde{Z}$ refers to the regressor matrix of the transformed model in Equation (2.6) and $\tilde{Z}_i$ to the corresponding $J \times R - 1$ regressor matrix for observation unit i. Assumption 2.1(i) imposes independence across the vectors of errors for each observation unit. It does not assume independence of elements within each vector of errors. Assumption 2.1(ii) assumes that the errors are sub-Gaussian with variance proxy σ. The variance proxy σ serves as an upper bound of the variance of the errors and allows for (conditional) heteroscedasticity. Note that the error term in the linear probability model in Equation (2.9) is sub-Gaussian with variance proxy $\sigma \leq 1$. This follows from the fact that the error term in the linear probability model is bounded between -1 and 1 since $y_{i,j}$ is either 0 or 1, the weights θ_r are nonnegative and, by Assumption 2.1(iii), $\tilde{z}_{i,j}^r$ is also bounded between -1 and 1. $\tilde{z}_{i,j}^r \in [-1, 1]$ is satisfied by the logit kernel in Equation (2.2) and other examples such as the kernel of binary choice and of multinomial choice without logit errors (see, e.g., Fox et al., 2016). Assumption 2.1(iv) holds by the definition of linear probability models.

2.3.1 Selection Consistency

For our analysis of the selection consistency, we adapt the definition of Zhao and Yu (2006). An estimator is defined as equal in sign if $\hat{\theta}_r$ and θ_r^* have the same sign for every $r = 1, \ldots, R-1$. Due to the nonnegativity of the estimates, the definition implies that $\hat{\theta}$

must be positive at all points in $\mathcal{B}_R$ for which $\theta_r^* > 0$, and zero at those where $\theta_r^* = 0$. Therefore, the estimation of the correct signs is equivalent to the correct selection of grid points. If an estimate $\hat{\theta}$ of θ^* is equal in sign, we write $\hat{\theta} =_s \theta^*$.

Our definition only includes $R-1$ points of the transformed model in Equation (2.8). That is, we only identify whether the $R-1$ weights included in Equation (2.8) have the correct sign but not whether the last weight $\hat{\theta}_R = 1 - \sum_{r=1}^{R-1} \hat{\theta}_r$ has the correct sign.

Definition 2.1. *An estimate $\hat{\theta}$ is **sign consistent** if*

$$\lim_{N\to\infty} P\left(\hat{\theta} =_s \theta^*\right) = 1.$$

According to Definition 2.1, an estimator is sign consistent if it estimates a positive weight at every grid point at which $\theta_r^* > 0$, and zero weights otherwise with probability approaching one as N goes to infinity.

To derive the condition under which our generalized estimator is sign consistent, we assume that $\mathcal{B}_R$ includes both grid points inside the support of $F_0(\beta)$, i.e., points at which $\theta_r^* > 0$, and points outside the true support, i.e., at which $\theta_r^* = 0$. Let $S = \{r \in \{1,\dots,R-1\} | \theta_r^* > 0\}$ define the index set of grid points at which $\theta^* > 0$, and let $S^C = \{r \in \{1,\dots,R-1\} | \theta_r^* = 0\}$ denote its complement. The corresponding cardinalities are defined as $s :- |S|$ and $s^C :- |S^C|$. We refer to grid points in S as active grid points and to grid points in S^C as inactive grid points. $\tilde{Z}_S$ and $\tilde{Z}_{S^C}$ denote the sub-matrices of all columns of $\tilde{Z}$ that are in S and S^C, respectively.

Since we allow the number of grid points $R(N)$ to increase with the sample size N, we typically expect the number of active points $s(N)$ to increase with N as well if $F_0(\beta)$ is sufficiently smooth. We again drop the dependence on N for ease of notation and simply write s instead of $s(N)$.

Let λ denote the endogenous LASSO parameter given in Equation (2.8), that follows from the constraint $c = 1$ in Equation (2.7). μ is the exogenous tuning parameter that is specified by the researcher.

For the analysis in this subsection, we assume that $\lambda > 0$. This holds if the inequality constraint on the sum of probability weights is strongly active.[5] The assumption implies that (i) the left-out probability weight, θ_R, is equal to zero, which can be easily justified by the possibility to exclude a point that is located far outside the presumed true support, and that (ii) the remaining $R-1$ probability weights do not sum to exactly one when estimated without the linear constraint on the sum of probability weights.[6]

[5] A strongly active constraint requires strict complementary slackness of the KKT condition for the inequality constraint (cf. Nocedal and Wright, 2006, pp. 341–343).

[6] Note that for $\lambda = 0$, the generalized estimator simplifies to the nonnegative ridge estimator for $\mu > 0$ and to the nonnegative least squares estimator for $\mu = 0$. For the latter, we refer the interested reader to Slawski and Hein (2013) who study the selection consistency of the nonnegative least squares estimator.

Following Wu and Yang (2014), we then obtain the subsequent condition for the sign consistency of the generalized estimator:

Nonnegative Elastic Irrepresentable Condition (NEIC). *For $\lambda > 0$, there exists a positive constant $\eta > 0$ (independent of N) such that*

$$\max_{r \in S^C} \frac{1}{NJ} \tilde{Z}_{S^C}^T \tilde{Z}_S \left(\frac{1}{NJ} \tilde{Z}_S^T \tilde{Z}_S + \mu I_S \right)^{-1} \left(\iota_S + \frac{\mu}{\lambda} \theta_S^* \right) \leq 1 - \eta$$

where ι_S is a vector of s ones and I_S is the identity matrix.

The NEIC is a condition for the correct recovery of support points through our generalized estimator.

The term $\tilde{Z}_{S^C}^T \tilde{Z}_S$ restricts the linear dependency between active and inactive grid points. The term $\tilde{Z}_S^T \tilde{Z}_S$ measures the linear dependency among active grid points. The condition is less likely to be satisfied if the number of grid points R – and therefore, the correlation – increases. Besides the linear dependence of the regressor matrix, the condition takes into account the magnitude of the endogenously fixed LASSO parameter λ and the tuning parameter μ. For $\mu = 0$, the NEIC reverts to the Nonnegative Irrepresentable Condition (NIC), the corresponding condition for selection consistency of the FKRB estimator. In comparison to the NEIC, the NIC is more restrictive in two ways: First, it requires the inverse of $\tilde{Z}_S^T \tilde{Z}_S$ to exist, which is not necessary for the NEIC. Note that this restricts the number of points R the researcher can include into the grid for the FKRB estimator. Second, the researcher can ensure the NEIC to be met through an appropriate choice of the tuning parameter μ, which is not possible for the NIC.

In addition to the NEIC, we restrict the rate at which the number of active grid points $s(N)$ and total grid points $R(N)$ can increase with the sample size N. This accommodates the fact that the number of grid points specified by the researcher should diverge if $F_0(\beta)$ is continuous, which is necessary for the convergence of the estimated distribution $\hat{F}(\beta)$ to the true underlying distribution $F_0(\beta)$.

Rate Condition on Density of Grid (RCDG).

1. $\lim_{N \to \infty} 2\ s(N) J \exp\left(-\frac{N \xi_{\min}^S(\mu,N)^2 \rho(\mu,N)^2}{2\ s(N)} \right) = 0.$
2. $\lim_{N \to \infty} 2(R(N)-1) J \exp\left(-N\eta^2\lambda^2 \left(\frac{\xi_{\min}^S(\mu,N)}{s(N)\sqrt{s(N)} + \xi_{\min}^S(\mu,N)} \right)^2 /2 \right) = 0,$

where $\xi_{\min}^S(\mu,N)$ denotes the (unrestricted) minimal eigenvalue of $1/(NJ)\tilde{Z}_S^T\tilde{Z}_S + \mu I_S$ and $\rho(\mu,N) := \min_{i \in S} \left| \left(1/(NJ)\tilde{Z}_S^T\tilde{Z}_S + \mu I_S\right)^{-1} \left(1/(NJ)\tilde{Z}_S^T\tilde{Z}_S\theta_S^* - \lambda\iota_S\right) \right|$.

The RCDG can only be satisfied if $\xi_{\min}^S(\mu,N) > 0$.

This is only restrictive for the FKRB estimator and always holds for the generalized estimator as long as $\mu > 0$ since $1/(NJ)\tilde{Z}_S^T\tilde{Z}_S + \mu I_S$ is positive definite for $\mu > 0$ and only positive semidefinite for $\mu = 0$. The assumption $\xi^S_{\min}(\mu, N) > 0$ excludes the possibility of perfect collinearity to ensure that the solution to the FKRB estimator exists.

Theorem 2.1. *Suppose Assumption 2.1 holds. Suppose further that NEIC and RCDG hold. Then*

$$\lim_{N\to\infty} \mathbb{P}\left(\hat{\theta} =_s \theta^*\right) = 1.$$

Proof. See Appendix 2.C.2. □

Theorem 2.1 establishes the selection consistency of the generalized estimator, for which $\mu \geq 0$, and for the FKRB estimator, for which $\mu = 0$. The theorem relies on sufficient conditions for the estimators to select the true weights. These conditions are more restrictive for the FKRB estimator than for our generalization. That is, because the minimal eigenvalue $\xi^S_{\min}(\mu, N) = \xi^S_{\min}(0, N) + \mu$ is higher for the generalized than for the FKRB estimator and moreover, the NEIC holds whenever the NIC is satisfied.

This implies that our estimator consistently selects the true support whenever the FKRB estimator does. The converse is not true since the NEIC might hold even though the NIC does not. Thus, Theorem 2.1 reveals that our estimator can select the true weights in cases in which the FKRB estimator cannot.

Remark 2.2. Theorem 2.1 can also be applied to the smooth basis densities estimator proposed by Fox et al. (2011). The estimator is an extension of the fixed grid version for which the researcher specifies R parametric density functions $\phi(\beta|\Omega_r)$ with fixed distribution parameters instead of a fixed grid of random coefficients.[7] Regarding the analysis of the selection consistency, the only difference to the fixed grid approach lies in the calculation of the regressor matrix Z. For the smooth basis densities estimator, Fox et al. (2011) suggest to calculate the columns in Z with D i.i.d. simulation draws from the respective distribution function, i.e., $z^r_{i,j} = (1/D)\sum_{d=1}^{D} g(x_{i,j}, \beta_{r,d})$ where $\beta_{r,d}$ is drawn from a parametric distribution, e.g., with parameters $\Omega_r := (\mu_r, \Sigma_r)$, and $g(x_{i,j}, \beta_{r,d})$ denotes the logit kernel as in Equation (2.2). Since Since Assumptions 2.1(i)-(iv) also hold true for the smooth basis densities estimator, Theorem 2.1 also applies to the estimator whereby the selection consistency relates to the correct recovery of active and inactive basis densities.

[7] E.g., for fixed normal densities $\Omega_r = (\mu_r, \Sigma_r)$ where μ^r is $k \times 1$ mean vector and Σ_r a $k \times k$ variance–covariance matrix that are specified by the researcher before optimization. The probability weight for every basis density is estimated from the data using the estimator in Equation (2.5). The distribution function estimator for the smooth basis densities estimator is $\hat{F}(\beta) = \sum_{r=1}^{R} \hat{\theta}_r \Phi(\beta|\Omega_r)$ where $\Phi(\cdot)$ is the distribution function corresponding to $\phi(\cdot)$ (Fox et al., 2016).

2.3.2 Error Bounds

A key requirement for an accurate estimation of $F_0(\beta)$ – in addition to the correct support recovery discussed in Subsection 2.3.1 – is the precise estimation of the probability weights. In this section, we derive an error bound for the euclidean distance between the estimated probability weights and the weights that yield the best discrete approximation of $F_0(\beta)$.

Let $\mathcal{H}$ denote the set of vectors of length $R-1$ in $[-1,1]^{R-1}$ for which the ℓ_1-norm is no greater than 2

$$\mathcal{H} := \left\{ x \in [-1,1]^{R-1} \;\middle|\; \|x\|_1 \leq 2 \right\}.$$

The set $\mathcal{H}$ contains all possible values of $\Delta\hat{\theta} :- \hat{\theta} - \theta^*$ since $\hat{\theta}$ and θ^* are vectors of weights which sum up to at most 1. Therefore, it is sufficient to consider elements in $\mathcal{H}$ when analyzing the potential error $\Delta\hat{\theta}$.

Define the restricted minimum eigenvalue of the real symmetric $R-1 \times R-1$ matrix $1/(NJ)\tilde{Z}^T\tilde{Z} + \mu I_{R-1}$ over the set of vectors $\mathcal{H}$ as

$$\xi_{\min}(\mu) := \inf_{v \in \mathcal{H}} \frac{v^T[\frac{1}{NJ}\tilde{Z}^T\tilde{Z} + \mu I_{R-1}]v}{\|v\|_2^2}.$$

Because the restricted minimal eigenvalue is greater than or equal to the unrestricted minimal eigenvalue, we use the restricted eigenvalue to derive a tighter error bound. We still assume $\xi_{\min}(\mu) > 0$, which rules out perfect collinearity. By the same arguments as in Subsection 2.3.1, $\xi_{\min}(\mu) > 0$ is always satisfied for our generalized estimator with $\mu > 0$ and $\xi_{\min}(\mu) > 0$ is only restrictive for the FKRB estimator.

Following the proof in Takada et al. (2017), we obtain an error bound on the $R-1$ estimated probability weights.

Theorem 2.2. *Let $0 < \delta \leq 1$. Define $\gamma \equiv \gamma(N,\delta) := \sqrt{2\log\left(\frac{2(R-1)J}{\delta}\right)/N}$. Suppose Assumption 2.1 holds, and that $\xi_{\min}(\mu) > 0$ for $\mu \geq 0$. Then, it holds with probability $1-\delta$ that*

$$\|\hat{\theta} - \theta^*\|_2 \leq \frac{2\sqrt{R-1}\;\gamma + 2\mu\sqrt{s}\;\|\theta_S^*\|_\infty}{\xi_{\min}(\mu)}.$$

Proof. See Appendix 2.C.3. □

Theorem 2.2 holds with probability approaching one as $\delta \to 0$. The estimation error for the Rth weight, $\theta_R = 1 - \sum_{r=1}^{R-1}\theta_r$, which is not included in the bound, approaches zero whenever $\|\hat{\theta} - \theta^*\|_2$ is close to zero.

Because $\gamma(N,\delta)$ decreases in N, the error bound becomes tighter if the number of observation units increases. The number of grid points leads to a direct increase of the error bound, both through R and s, which is expected to increase with R, e.g., if

the true distribution is continuous. The number of grid points also has an indirect effect attributable to the stronger correlation typically associated with an increase in the number of grid points. This effect is captured through the restricted minimum eigenvalue $\xi_{\min}(\mu)$, which decreases if the correlation increases. Hence, an increase in the number of grid points R typically leads to a wider error bound on the estimated weights (for a given tuning parameter μ).

The researcher can affect the error bound on the estimated weights through the choice of the tuning parameter μ. For $\mu = 0$, the bound in Theorem 2.2 simplifies to the error bound for the FKRB estimator. A comparison of the bound for $\mu = 0$ and $\mu > 0$ reveals that the extension has two opposing effects on the estimator's precision. First, a direct increasing effect that is captured through the tuning parameter in the numerator of Theorem 2.2 and, second, an indirect decreasing effect via the restricted minimum eigenvalue since $\xi_{\min}(\mu) = \xi_{\min}(0) + \mu > \xi_{\min}(0)$ for $\mu > 0$.

While the direct effect becomes stronger with the number of true support points s, the indirect effect is especially relevant if the correlation among grid points is strong. In that case, the extension leads to an increase of $\xi_{\min}(\mu)$ and hence, to a tighter error bound. The indirect effect is most important if the design matrix $\tilde{Z}$ is almost singular, i.e., if the grid is sufficiently dense. In that case, the restricted minimum eigenvalue $\xi_{\min}(0)$ of the FKRB estimator is close to zero. The appropriate choice of μ offsets this effect and can lead to a tighter error bound.

Corollary 2.1 establishes the condition under which our extension provides a tighter error bound on the estimated weights than the FKRB estimator.

Corollary 2.1. *When* $\sqrt{s}\,\|\theta_S^*\|_\infty \xi_{\min}(0) < \sqrt{R-1}\,\gamma$*, then the error bound for* $\|\hat{\theta} - \theta^*\|_2$ *in Theorem 2.2 is tighter for the generalized estimator than for the FKRB estimator.*

Proof. See Appendix 2.C.3. □

Using the error bound on the estimated and true probability weights in Theorem 2.2, we derive a bound on the error of the estimated distribution function $\hat{F}(\beta)$ and the best discrete distribution $F^*(\beta)$.

Theorem 2.3. *Under the assumptions and conditions in Theorem 2.2, it holds at any point* $\beta \in \mathbb{R}^K$ *with probability* $1-\delta$ *that*

$$|\hat{F}(\beta) - F^*(\beta)| \leq \frac{4(R-1)\ \gamma + 4\mu\sqrt{(R-1)s}\ \left\|\theta_S^*\right\|_\infty}{\xi_{\min}(\mu)}.$$

Proof. See Appendix 2.C.3. □

The bound on the difference between the estimated distribution and the best discrete approximation of $F_0(\beta)$ increases in R and decreases in $\xi_{\min}(\mu)$. Similarly to Theorem 2.2, the difference in the distributions decreases in N since k may decrease when N increases.

Recall that the absolute difference $|F_0(\beta) - F^*(\beta)|$ becomes negligibly small as R increases (Fox et al., 2016). Therefore, the estimation error can be well captured by $|\hat{F}(\beta) - F^*(\beta)|$ which explains the relevance of Theorem 2.3.

Remark 2.3. Theorem 2.3 can be extended in a straightforward way to an error bound for the smooth basis densities estimator suggested by Fox et al. (2011) if the support of β is bounded and D i.i.d. simulation draws. Following the argumentation in Fox et al. (2016), the distribution function estimated with the smooth basis densities estimator, $\hat{F}_D(\beta) = \sum_{r=1}^R \hat{\theta}_r \Phi(\beta|\Omega_r)$, can be nested into the discrete approximation model by means of the simulation approximated distribution $\tilde{F}_D(\beta) = \sum_{r=1}^R \hat{\theta}_r (1/D) \sum_{d=1}^D 1[\beta_{r,d} \leq \beta]$ where $\hat{\theta}$ is estimated with the smooth basis densities estimator. Using the simulation approximated distribution, we obtain

$$\begin{aligned}\left|\hat{F}_D(\beta) - F^*(\beta)\right| &\leq \left|\tilde{F}_D(\beta) - F^*(\beta)\right| + \left|\hat{F}_D(\beta) - \tilde{F}_D(\beta)\right| \\ &\leq \left|\tilde{F}_D(\beta) - F^*(\beta)\right| + \sum_{r=1}^R \hat{\theta}_{r,D} \left|\frac{1}{D}\sum_{d=1}^D 1[\beta_{r,d} \leq \beta] - \Phi(\beta|\Omega_r)\right|\end{aligned}$$

For $D \to \infty$, $\tilde{F}_D(\beta)$ converges to $\hat{F}_D(\beta)$ such that the second expression goes to zero for any given r (by the Glivenko–Cantelli theorem) Fox et al. (2016). The first expression is the absolute difference between the fixed grid estimator and the best possible approximation that can be obtained with a mixture of smooth basis densities (Fox et al., 2016). The expression can be bounded by the error bound presented in Theorem 2.3. Consequently, the absolute difference between $\hat{F}(\beta)$ and $F^*(\beta)$ can also be bounded by Theorem 2.3 if $D \to \infty$.

2.4 Monte Carlo Simulation

We conduct two Monte Carlo experiments to examine the selection consistency and the approximation accuracy of our generalized estimator. The Monte Carlo simulation on the selection consistency uses a discrete distribution with a subset of grid points as support points. The second experiment generates the random coefficients from a mixture of two normal distributions. This allows us to study the estimators' ability to estimate smooth distributions. We use a random coefficients logit model as the true data generating process to generate individual-level discrete choice data. Each observational unit i chooses among $J = 4$ mutually exclusive alternatives and an outside option. For every alternative j and observation unit i, we draw the two-dimensional covariate vector $x_{i,j} = (x_{i,j,1}, x_{i,j,2})$ from $\mathcal{U}(0, 5)$ and $\mathcal{U}(-3, 1)$, respectively. To study the effect of the fixed grid and the number of observation units on the estimators' performance, we run every experiment for different sample sizes and numbers of grid points. We repeat the experiment for every combination of R and N 200 times to compare the performance of our estimator with the FKRB estimator in terms of selection consistency and accuracy for every setup. All calculations are conducted with the statistical software R (R Core Team, 2018).

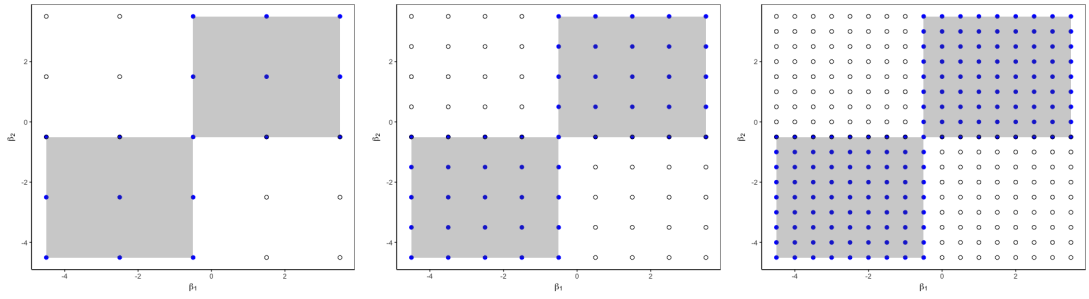
2.4.1 Discrete Distribution

To study the estimators' selection consistency, we generate the random coefficients β from a discrete probability mass function. The estimator successfully recovers the true support from the data if it estimates a positive weight at every support point of $F_0(\beta)$, and zero weights at all points outside its support. For the support points of $F_0(\beta)$, we select a subset of the grid points from the fixed grid we use for the estimation. The grid covers the range $[-4.5, 3.5] \times [-4.5, 3.5]$ with $R = \{25, 81, 289\}$ uniformly allocated grid points. We specify the support of our discrete data generating distribution on $[-4.5, -0.5] \times [-4.5, 0.5]$, and $[-0.5, 3.5] \times [-0.5, 3.5]$, whereby the number of support points varies due to the varying number of grid points. That is, we draw the random coefficients β from a discrete mass function with $S = \{17, 49, 161\}$ support points, each drawn with uniform probability weight $\theta_s = 1/S$.

In this setup, the data generating process exactly matches the underlying probability model of the fixed grid estimator. This way, we abstract from any approximation errors that can arise from the sieve space approximation of the true underlying distribution. Therefore, the experiment studies the estimators' selection consistency in the most simple framework possible. The two areas of the discrete distribution with positive probability mass simulate two heterogeneous groups of preferences in the population. We estimate every distribution for sample sizes $N = \{1000, 10,000\}$.

Figure 2.1 illustrates the setup of the Monte Carlo experiment for the three data generating distributions. The blue shaded area indicates the support of the discrete mass functions, and the filled blue points inside this area the active grid points. The hollow

Figure 2.1: Grid of Monte Carlo Study with Discrete Mass Points



(a) $R = 25$, $S = 17$ (b) $R = 81$, $S = 49$ (c) $R = 289$, $S = 161$

black points outside the blue shared areas are the inactive grid points that are not used for data generation.

We choose the optimal tuning parameter μ for the generalized estimator with 10-fold cross-validation from a sequence of 101 potential values. For 100 of these values, we use the sequence suggested by the R package *glmnet* for ridge regression with nonnegative coefficients. We also include $\mu = 0$ in the range of possible values to allow our estimator to simplify to the FKRB estimator if the model fit in the cross-validation is highest for $\mu = 0$. The selection of the optimal tuning parameter is based on the mean squared error (MSE) criterion. In addition to the tuning parameter with the lowest MSE, we report the tuning parameter that follows from the one-standard-error rule (OneSe).[8]

As robustness-checks, we consider the prediction accuracy of the predicted choice of every observation and the log-likelihood as a measure of fit in the cross-validation. We choose the μ based on the smallest average out-of-sample prediction error and based on the highest log-likelihood, respectively. The results of the Monte Carlo study for the log-likelihood and predicted choices as selection criteria can be found in Appendix 2.A. They indicate that the MSE and the one-standard-error rule give the best results.

To evaluate the estimators' selection consistency, we calculate the average share of sign consistent estimates. An estimate is sign consistent if it is positive at active grid points, and zero otherwise. A weight is defined as positive if it is greater than 10^{-3}. To illustrate the sparsity of the estimators' solutions, we report the average number of positive weights and the average share of true positive weights.

Beyond selection consistency, the discrete setup of the Monte Carlo experiment allows us to study the bias of the estimated probability weights. Denote the estimated

[8]We observe that the curve of the MSE in dependency of μ tends to be flat and that the μ chosen by OneSe often corresponds to the largest element of the sequence of tuning parameters suggested by the *glmnet* package. Therefore, a possible strategy is to choose the largest μ given by the *glmnet* package to obtain μ of OneSe if one wants to avoid cross-validation.

weight at grid point r in Monte Carlo run m by $\hat{\theta}_{r,m}$. We calculate the L_1 norm

$$L_1 = \frac{1}{M}\sum_{m=1}^{M}\frac{1}{R}\sum_{r=1}^{R}\left|\theta_r - \hat{\theta}_{r,m}\right|$$

to measure the average absolute bias of $\hat{\theta}$ in comparison to the true weights θ over all Monte Carlo runs M. In addition, we adopt the root mean integrated squared error (RMISE) from Fox et al. (2011) to provide a metric on the approximation accuracy of the estimated distribution. The RMISE averages the squared difference between the true and estimated distribution at a fixed set of grid points across all Monte Carlo runs

$$\text{RMISE} = \sqrt{\frac{1}{M}\sum_{m=1}^{M}\left[\frac{1}{E}\sum_{e=1}^{E}\left(\widehat{F}_m(\beta_e) - F_0(\beta_e)\right)^2\right]},$$

where $\widehat{F}_m(\beta_e)$ denotes the estimated distribution function in Monte Carlo run m evaluated at β_e. For the evaluation, we use $E = 10{,}000$ points uniformly distributed over the range $[-4.5, 3.5] \times [-4.5, 3.5]$.

Table 2.1 summarizes the results of the Monte Carlo experiment. The first three columns report the sample size N, the number of grid points R, and the number of true support points S. The upper part of the table presents the measures on the accuracy

Table 2.1: Summary Statistics of 200 Monte Carlo Runs with Discrete Distribution.

			RMISE			L_1			μ		ρ
N	R	S	FKRB	MSE	OneSe	FKRB	MSE	OneSe	MSE	OneSe	3rd Qu.
1,000	25	17	0.069	0.041	0.035	0.035	0.017	0.015	55.89	67.90	0.808
1,000	81	49	0.082	0.052	0.038	0.019	0.009	0.007	53.91	69.93	0.819
1,000	289	161	0.088	0.057	0.045	0.006	0.004	0.003	55.89	71.29	0.822
10,000	25	17	0.041	0.024	0.022	0.020	0.012	0.011	61.34	66.90	0.808
10,000	81	49	0.050	0.030	0.027	0.015	0.008	0.007	60.40	68.96	0.819
10,000	289	161	0.059	0.037	0.034	0.006	0.004	0.003	61.73	70.48	0.822
			Pos.			% True Pos.			% Sign		
N	R	S	FKRB	MSE	OneSe	FKRB	MSE	OneSe	FKRB	MSE	OneSe
1,000	25	17	13.10	20.77	22.25	67.32	95.23	99.79	71.18	78.44	78.70
1,000	81	49	15.29	46.56	54.44	26.88	77.39	89.81	53.15	75.65	80.95
1,000	289	161	16.00	103.37	123.63	8.38	54.58	65.56	48.10	69.34	74.56
10,000	25	17	17.38	19.46	19.77	91.56	98.71	99.88	87.02	88.38	88.74
10,000	81	49	23.32	45.22	48.02	42.07	82.00	87.14	61.62	82.89	85.65
10,000	289	161	24.34	96.81	105.33	13.24	54.79	59.62	50.62	71.83	74.27

Note: The table reports the average summary statistics over all Monte Carlo replicates for the FKRB estimator (FKRB), and for our generalized estimator with tuning parameter μ from a 10-fold cross-validation and the MSE criterion (MSE) and the one-standard-error rule (OneSe).

of the estimated weights, and the lower part the shares of positive, true positive, and sign consistent estimated weights. The final column in the upper part reports the third quantile of the absolute values of the correlation ρ among grid points.[9]

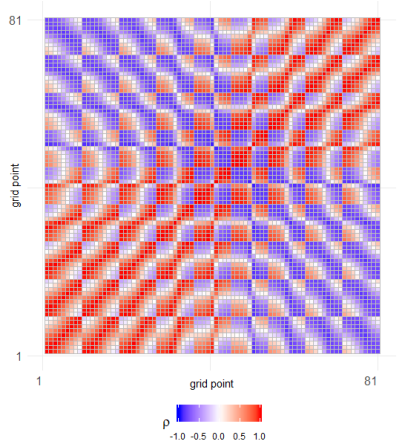
The results show that our generalized estimator outperforms the FKRB estimator for every combination of N and R, in particular when the tuning parameter μ is chosen based on the one-standard-error rule. With respect to the selection consistency, the generalized estimator recovers more true positive and sign consistent probability weights than the FKRB estimator. While the decrease in these shares is moderate for the generalized estimator when the discrete distribution becomes more complex, the correct recovery through the FKRB estimator significantly worsens.

This is best illustrated by the small number of positive weights, which changes only slightly alongside the increasing complexity. For $N = 1000$ ($N = 10,000$) and in the extreme case of $R = 289$, the FKRB estimator estimates positive weights at no more than 16 (24) of the grid points (in comparison to 124 (105) for the generalized estimator with OneSe).

In addition to its improved selection consistency, all measures on the estimated weights indicate that our generalized version provides substantially more accurate estimates of the probability weights than the FKRB estimator. The bias reduction persists for small and large sample sizes.

The plot of the correlation matrix in Figure 2.2 and the third quantile of the values of absolute correlation in Table 2.1 both illustrate that correlation among many grid points is strong.

Figure 2.2: Correlation Matrix for $N = 10,000$ and $R = 81$



[9] In addition, we also considered the mean and median to summarize the absolute correlation among grid points. We focus on the third quantile since it best illustrates the strong correlation in this setup.

2.4.2 Continuous Distribution

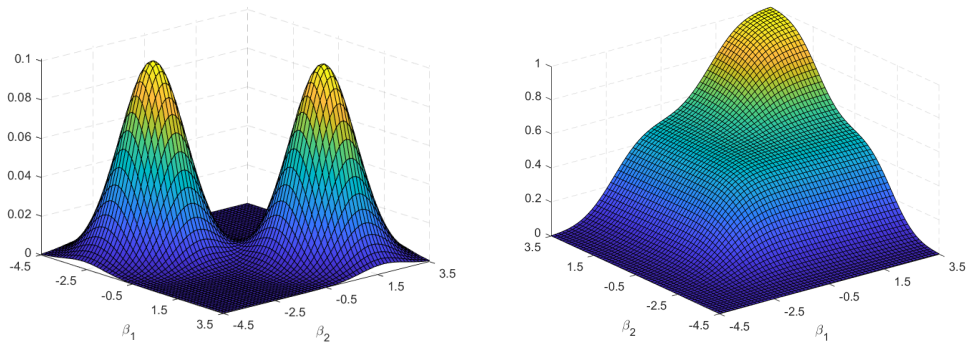
The second Monte Carlo experiment considers a mixture of two bivariate normal distributions for $F_0(\beta)$ to analyze how our generalized estimator accommodates more complex continuous distributions. This way, we can assess its ability to recover distributions that cannot be estimated with parametric techniques.

For the estimation, we use a fixed grid with points spread on $[-4.5, 3.5] \times [-4.5, 3.5]$. The fixed grid covers the support of the true distribution with coverage probability close to one (0.993). We keep the correlation among grid points as low as possible and generate the grid points with a Halton sequence. To study the convergence of the estimated distribution to $F_0(\beta)$ for an increasing number of grid points, we estimate the model with $R = \{25, 50, 100, 250\}$. The number of observation units N varies between 1000 and 10,000. The variance–covariance matrices of the two normals are $\Sigma_1 = \Sigma_2 = \left[\begin{smallmatrix} 0.8 & 0.15 \\ 0.15 & 0.8 \end{smallmatrix}\right]$. We generate the random coefficient vectors β from the following two-component bivariate mixture

$$0.5\ \mathcal{N}\Big([-2.2, -2.2], \Sigma_1\Big) + 0.5\ \mathcal{N}\Big([1.3, 1.3], \Sigma_2\Big).$$

The left panel in Figure 2.3 displays the bimodal joint density of the mixture of two normals, and the right panel the joint distribution function.

Figure 2.3: True Density and Distribution Function of Mixture of two Normals



(a) PDF (b) CDF

For the calculation of the RMISE, we use $E = 10{,}000$ evaluation points uniformly distributed over the range of the fixed grid. In addition, we report the average number of positive, true positive, and sign consistent estimated weights. For the number of true positive and sign consistent weights, we calculate the true density at every grid point and then normalize the density of each grid point by the sum of densities at all grid points. We define a true weight as positive if its normalized density is greater 10^{-3}.

Table 2.2 summarizes the average results over the $M = 200$ Monte Carlo replicates

Table 2.2: Summary Statistics of 200 Monte Carlo Runs with Mixture of Two Bivariate Normals.

			RMISE			Pos.			μ		ρ
N	R	S	FKRB	MSE	OneSe	FKRB	MSE	OneSe	MSE	OneSe	3rd Qu.
1,000	25	17	0.086	0.072	0.055	9.83	13.20	17.84	22.72	74.23	0.823
1,000	50	33	0.087	0.068	0.059	12.56	26.84	32.61	48.85	74.27	0.820
1,000	100	61	0.100	0.075	0.062	13.45	43.41	55.36	48.74	73.99	0.823
1,000	250	127	0.101	0.073	0.062	14.22	86.30	105.14	56.42	74.70	0.824
10,000	25	17	0.063	0.061	0.057	11.63	12.60	14.76	18.66	73.90	0.823
10,000	50	33	0.058	0.049	0.047	17.52	25.44	28.33	50.92	74.05	0.820
10,000	100	61	0.061	0.048	0.043	19.94	39.36	47.24	49.69	74.12	0.822
10,000	250	127	0.062	0.043	0.039	22.03	80.90	89.30	63.55	74.66	0.824

			% True Pos.			% Sign		
N	R	S	FKRB	MSE	OneSe	FKRB	MSE	OneSe
1,000	25	17	49.59	66.82	88.38	60.12	70.06	80.84
1,000	50	33	33.26	70.65	85.44	52.78	73.58	81.55
1,000	100	61	18.82	62.11	79.35	48.51	71.37	80.45
1,000	250	127	7.93	55.58	68.09	51.57	71.15	76.33
10,000	25	17	58.15	64.09	76.91	64.56	68.74	77.58
10,000	50	33	47.15	69.59	77.73	61.21	74.98	79.95
10,000	100	61	28.31	58.41	70.46	53.61	70.90	77.72
10,000	250	127	13.26	55.26	61.13	53.87	72.98	75.59

Note: The table reports the average summary statistics over all Monte Carlo replicates for the FKRB estimator (FKRB), and for our generalized estimator with tuning parameter μ from a 10-fold cross-validation and the MSE criterion (MSE) and the one-standard-error rule (OneSe).

for the FKRB estimator and our generalized estimator when μ is chosen with 10-fold cross-validation and the MSE and one-standard error rule, respectively. Results for the prediction accuracy of the predicted choices and the log-likelihood as criteria are reported in Appendix 2.A.

The RMISE shows that our generalized estimator provides more accurate estimates of the true underlying random coefficients' distribution than the FKRB estimator for every combination of N and R. For $N = 10{,}000$ the generalized version becomes more accurate with increasing number of grid points and approximates $F_0(\beta)$ quite well for $R = 250$. However, the FKRB estimator does not result in a lower RMISE for $N = 10{,}000$ when R increases.

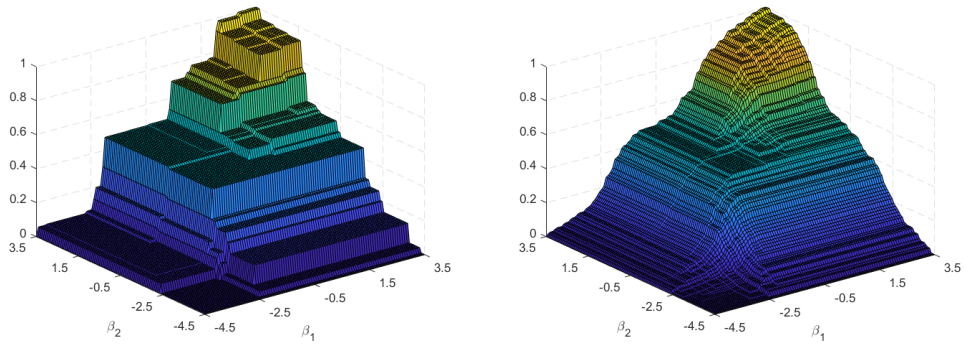
The improved performance of our estimator for every combination of N and R can be explained with the larger number of true positive and sign consistent estimated probability weights. Independently of the number of (relevant) grid points, the FKRB estimator estimates only a small number of positive weights and, hence, recovers only few relevant grid points. The share of true positive and sign consistent estimated weights

is substantially higher for our estimator.

Figure 2.4 plots an example of the joint distribution functions estimated with the FKRB estimator (Panel (a)) and our generalized estimator (Panel (b)). Figure 2.5 shows the corresponding estimated and true marginal distributions of β_1 and β_2. The distribution functions are estimated for $N = 10{,}000$ and $R = 250$.

The plots illustrate the impact of the FKRB estimator's sparse nature on the estimated marginal and joint distribution functions. Visual inspection shows that it approximates $F_0(\beta)$ through a step function with only few steps due to the small number of positive weights. In contrast, our generalized estimator provides a smooth estimate that is close to the true underlying distribution function.

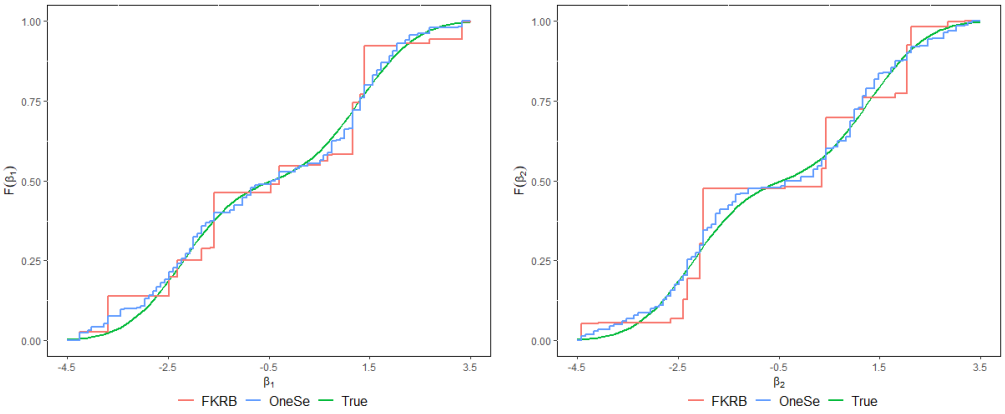
Figure 2.4: Estimated Joint Distribution Functions for $N = 10,000$ and $R = 250$



(a) FKRB

(b) Generalized with OneSe

Figure 2.5: True and Estimated Marginal Distribution Functions for $N = 10,000$ and $R = 250$



2.5 Empirical Application

To study the performance of our generalized estimator with real data, we apply it to the *ModeCanda* data set from the *R package mlogit*. Originally, the Canadian National Rail Carrier VIA Rail assembled the data in 1989 to analyze the demand for future intercity travel in the Toronto–Montréal corridor. The data contains information on travelers who can choose among the four intercity travel mode options car, bus, train, and air. Due to the small number of bus users (18), we follow Bhat (1997*a*) and drop bus as an alternative. Furthermore, we only consider travelers in our analysis that can choose among all three options. Thus, the analyzed data consists of 3593 business travelers who can choose among airplane, train, and car. In addition to the observed choices, the data includes information on traveler's income, the trip distance, the frequency of the service, total travel cost, an indicator that is one if either the city of arrival or departure is a big city and zero otherwise, and the in- and out-of-vehicle travel time. We construct the travel time variable by summing up in-vehicle travel time and out-of-vehicle time. This is done for two reasons: first, the data on out-of-vehicle time is always zero for car users and would therefore only capture the preferences of airplane and train users. Second, we think it is plausible that individuals care more about total travel time than the travel time inside and outside of a vehicle separately.

A detailed description of the data can be found in Marwick and Koppelman (1990). Among others, the data set has been studied by Bhat (1995,9,9,9), Koppelman and Wen (2000), Wen and Koppelman (2001). The only paper that analyzes the data with a random coefficients logit model is the study by Hess, Bierlaire and Polak (2005). However, they only use the explanatory variables as input for a Monte Carlo study and simulate travelers' mode choices.

We estimate a mixed logit model with a random coefficient on the travel time and fixed coefficient on all other variables to study the preferred travel mode of business travelers. We include all the above variables into the utility specification along with mode specific constants, where we specify car as the reference alternative. To apply the fixed grid approach to a model with fixed and random coefficients, we follow the recommendation of Fox et al. (2016) and Houde and Myers (2019) who suggest a two-step estimator to estimate the model with fixed and random coefficients.[10] In the first step, all coefficients are estimated using a semiparametric mixed logit. We assume that the random coefficient is normally distributed. In the second step, the fixed variables and their estimated coefficients from the first stage are treated as data and only the random coefficient of travel time is estimated with the FKRB and generalized estimator. Houde and Myers (2019) justify the procedure with the argument that a mixed logit can recover the means of a distribution fairly well despite the incorrect assumptions on the random coefficients' distribution. Thus, the fixed coefficients can be estimated consistently with

[10]We also provide an algorithm to update both the fixed and random coefficients in Appendix 2.B. The algorithm is a modification of the flexible grid estimator in Train (2008). Unfortunately, the algorithm seems to be very slow and we do not include its results in our comparison here.

the semiparametric approach. They illustrate this property in a Monte Carlo study.

We center the grid of the random coefficient around the mean estimate of the travel coefficient from the first step[11] and add three standard deviations to each side. We estimate the second step with different numbers of grid points. The preferred specification uses $R = 100$ uniformly spread points on the range $[-0.061, 0.027]$. We choose the tuning parameter with 10-fold Cross-Validation and the one standard error rule as criterion. Figure 2.6 summarizes the mass and the distribution functions estimated with the FKRB and the generalized estimator.

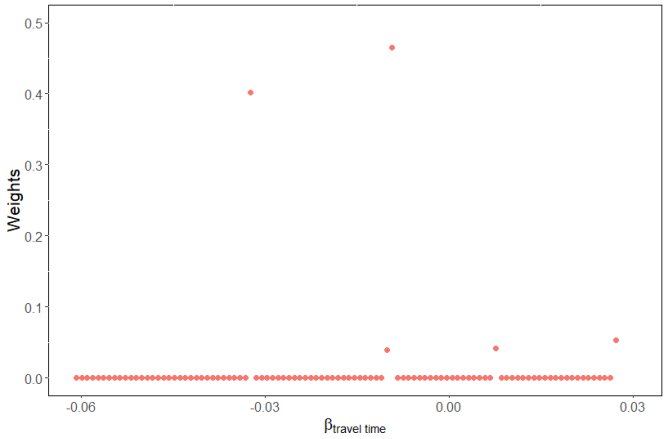
The generalized estimator estimates a smooth mass function whereas the FKRB exhibits LASSO-type behavior. The FKRB estimator only selects five out of 100 grid points whereas the generalized version selects 75 grid points.[12] Furthermore, it can easily be seen that the estimated mass function obtained by the generalized estimator does not seem to be normally distributed but rather looks like a mixture of two normal distributions. That is, specifying a normal or any other parametric distribution function does not seem appropriate in this example. A quite unexpected result is that there are positive weights at positive grid points implying that some people appreciate longer trips. Even though one might argue that this might be the case if such travelers accept additional travel time for, say, additional comfort when traveling, this might also be a sign of a misspecified model. For the FKRB estimator these weights sum up 9.5% and for the generalized estimator to 10.1%, which is lower than 12.6% for the mixed logit with a normal distribution. The weighted mean of the coefficient of travel time for the FKRB estimator is -0.01593 and -0.01631 for the generalized estimator. This is roughly the same as -0.01682, the mean coefficient obtained from the mixed logit model with normally distributed travel time coefficient which is in line with the justification of Houde and Myers (2019) for the two-step estimator. In addition to the estimated distributions, we report the mean (and median) over individuals' own- and cross-travel time elasticities for the FKRB estimator, the generalized estimator and the semiparametric mixed logit with a normal distribution in Appendix 2.A. We also calculate the ratio between elasticities estimated with the FKRB estimator and the semiparametric estimator in comparison to the elasticities estimated with the generalized estimator. The ratios show that most differences of the estimated own- and cross-travel time elasticities do not seem to be too large. Yet, few deviate from each other whereby the semiparametric estimator is up to 6.3 ($= 1/0.16$) times smaller and the FKRB estimator is up to 1.8 times larger than the generalized estimator. We also observe in the continuous Monte Carlo experiment that the estimated elasticities are rather similar for the FKRB estimator and the generalized estimator.[13] Therefore, it is not clear to what extent the generalized estimator outperforms the FKRB estimator in terms of the estimated elasticities, while it is very clear in terms of the estimated distribution.

[11] The estimated coefficients of the first stage are provided in Appendix 2.A.

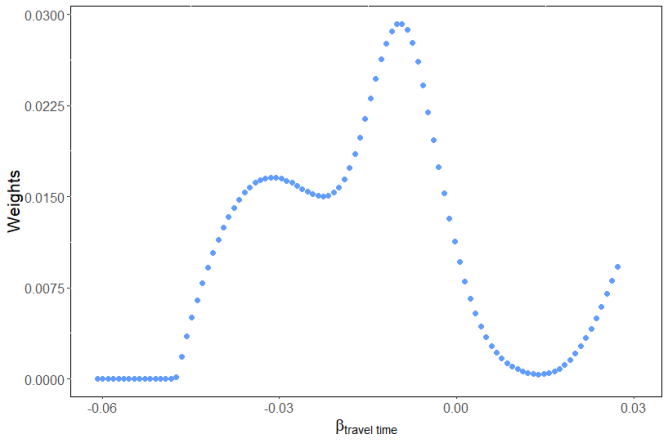
[12] We again define a weight as positive if it is greater than 10^{-3}.

[13] The results are available on request.

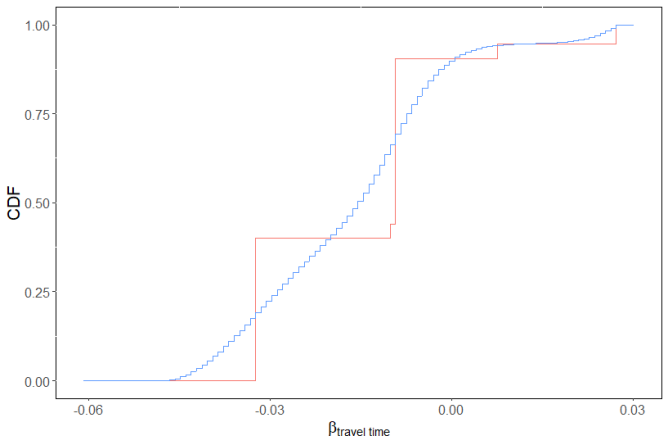
Figure 2.6: Estimated Distributions of Travel Time in Mode Canada Data with $R = 100$



(a) Mass Function for FKRB



(b) Mass Function for Generalized with OneSe



(c) CDFs for FKRB (red) and Generalized with OneSe (blue)

2.6 Conclusion

We extend the simple and computationally attractive nonparametric estimator of Fox et al. (2011). We illustrate that their estimator is a special case of NNL, explaining its sparse solutions. The connection to NNL reveals that the estimator tends to randomly select among highly correlated grid points. This behavior gives reason to doubt the precise estimation of the true distribution through the estimator.

To mitigate its undesirable sparsity and random selection behavior, we add a quadratic constraint on the probability weights to the optimization problem of the FKRB estimator. This simple and straightforward extension transforms the estimator to a special case of nonnegative elastic net. The combination of the linear and quadratic constraint on the probability weights enables a more reliable selection of the relevant grid points. As a consequence, our generalized estimator provides more accurate estimates of the true underlying random coefficients' distribution without substantially increasing computation time and complexity. We derive conditions for selection consistency and an error bound on the estimated distribution function to verify the improved properties of our estimator.

Two Monte Carlo studies illustrate the attractive theoretical properties of our estimator. They show that our generalized version estimates considerably more positive probability weights and recovers more grid points correctly. In addition to the improved selection consistency, the estimator provides more accurate estimates of the true underlying distributions.

Applying the FKRB and the generalized estimator to a data set of travel choices made in the Toronto–Montréal corridor confirms the sparsity of the FKRB estimator. In contrast, the generalized estimator selects substantially more grid points, resulting in a smooth distribution function. This illustrates the fact that our generalized estimator is able to approximate continuous distribution functions.

A challenging, but practically relevant topic is the development of an inference procedure. To this end, one has to take into account the relation of the FKRB and our generalized estimator to the nonnegative LASSO and nonnegative elastic net, respectively. Assuming random regression coefficients, Pötscher and Leeb (2009) prove that estimators of the distribution function of the LASSO, including resampling methods, cannot be uniformly consistent. Assuming fixed regression coefficients, Dezeure, Bühlmann and Zhang (2017) propose a de-biased LASSO estimator to conduct inference. However, it is not straightforward how to construct such a de-biased estimator in our setting.[14]

In addition, it might be a promising venue for future research to attempt to weaken

[14] Our experiments for inference regarding the estimated joint CDF and estimated elasticities suggest that the m-out-of-n-(block-)bootstrap might be a promising choice. Efron's (block-)bootstrap (Efron, 1979), in contrast, seems to have poor coverage. For the m-out-of-n-block-bootstrap, we base our simulation on block length J to take the correlation structure of our data into account. In these experiments, we followed the recommendation of Jentsch and Leucht (2016) for discrete data and chose $m = (NJ)^{2/3}$. The results are available on request.

some of our regularity conditions, such as the rate condition on the density of the grid. For a given number of observations, this would theoretically justify to increase the number of grid points used for the estimation. Moreover, our derived error bounds are non-asymptotic, so asymptotic results might provide further useful insights.

Appendix 2.A Supplementary Tables

Table 2.A.1: Detailed Summary Statistics of 200 Monte Carlo Runs with Discrete Distribution.

			RMISE					L_1					μ				ρ
N	R	S	FKRB	MSE	OneSe	LL	PredOut	FKRB	MSE	OneSe	LL	PredOut	MSE	OneSe	LL	PredOut	3rd Qu.
1,000	25	17	0.069	0.041	0.035	0.059	0.047	0.035	0.017	0.015	0.028	0.022	55.89	67.90	11.32	31.04	0.808
1,000	81	49	0.082	0.052	0.038	0.067	0.056	0.019	0.009	0.007	0.014	0.011	53.91	69.93	17.93	31.70	0.819
1,000	289	161	0.088	0.057	0.045	0.070	0.061	0.006	0.004	0.003	0.005	0.004	55.89	71.29	25.75	35.59	0.822
10,000	25	17	0.041	0.024	0.022	0.035	0.031	0.020	0.012	0.011	0.017	0.015	61.34	66.90	16.23	29.40	0.808
10,000	81	49	0.050	0.030	0.027	0.044	0.037	0.015	0.008	0.007	0.013	0.011	60.40	68.96	13.95	31.39	0.819
10,000	289	161	0.059	0.037	0.034	0.051	0.046	0.006	0.004	0.003	0.005	0.005	61.73	70.48	17.69	26.40	0.822

			Pos.					% True Pos.					% Sign				
N	R	S	FKRB	MSE	OneSe	LL	PredOut	FKRB	MSE	OneSe	LL	PredOut	FKRB	MSE	OneSe	LL	PredOut
1,000	25	17	13.10	20.77	22.25	15.93	18.84	67.32	95.23	99.79	79.62	90.35	71.18	78.44	78.70	76.56	79.52
1,000	81	49	15.29	46.56	54.44	29.14	38.95	26.88	77.39	89.81	50.46	66.22	53.15	75.65	80.95	64.58	71.55
1,000	289	161	16.00	103.37	123.63	62.08	83.70	8.38	54.58	65.56	33.01	44.46	48.10	69.34	74.56	59.59	64.87
10,000	25	17	17.38	19.46	19.77	18.11	18.73	91.56	98.71	99.88	94.62	96.88	87.02	88.38	88.74	88.24	88.86
10,000	81	49	23.32	45.22	48.02	29.57	37.70	42.07	82.00	87.14	53.94	68.73	61.62	82.89	85.65	68.26	76.12
10,000	289	161	24.34	96.81	105.33	50.80	63.70	13.24	54.79	59.62	28.49	35.99	50.62	71.83	74.27	58.46	62.35

Note: The table reports the average summary statistics over all Monte Carlo replicates for the FKRB estimator (FKRB), and for our generalized estimator with tuning parameter μ from a 10-fold cross-validation and the MSE criterion (MSE), the one-standard-error rule (OneSe), the log-likelihood criterion (LL) and the number of correctly predicted binary outcomes (PredOut). The predicted binary outcome is set to one for the alternative with the highest estimated choice probability.

Table 2.A.2: Detailed Summary Statistics of 200 Monte Carlo Runs with Mixture of Two Bivariate Normals.

			RMISE					Pos.					μ				ρ
N	R	S	FKRB	MSE	OneSe	LL	PredOut	FKRB	MSE	OneSe	LL	PredOut	MSE	OneSe	LL	PredOut	3rd Qu.
1,000	25	17	0.086	0.072	0.055	0.081	0.067	9.83	13.20	17.84	10.66	14.39	22.72	74.23	2.36	30.13	0.823
1,000	50	33	0.087	0.068	0.059	0.079	0.068	12.56	26.84	32.61	17.45	25.25	48.85	74.27	9.00	33.02	0.820
1,000	100	61	0.100	0.075	0.062	0.09	0.076	13.45	43.41	55.36	22.54	39.26	48.74	73.98	8.54	33.56	0.823
1,000	250	127	0.101	0.073	0.062	0.089	0.076	14.22	86.30	105.14	41.64	68.17	56.42	74.70	14.97	33.02	0.824
10,000	25	17	0.063	0.061	0.057	0.062	0.060	11.63	12.60	14.76	11.74	13.35	18.66	73.90	0.77	30.42	0.823
10,000	50	33	0.058	0.049	0.047	0.053	0.049	17.52	25.44	28.33	20.26	24.30	50.92	74.05	8.38	34.56	0.820
10,000	100	61	0.061	0.048	0.043	0.054	0.050	19.94	39.36	47.24	28.10	34.99	49.69	74.12	11.79	30.13	0.822
10,000	250	127	0.062	0.043	0.039	0.053	0.046	22.03	80.90	89.30	48.67	64.80	63.55	74.66	20.32	36.27	0.824

			% True Pos.					% Sign				
N	R	S	FKRB	MSE	OneSe	LL	PredOut	FKRB	MSE	OneSe	LL	PredOut
1,000	25	17	49.59	66.82	88.38	54.03	73.18	60.12	70.06	80.84	62.82	73.94
1,000	50	33	33.26	70.65	85.44	46.83	67.44	52.78	73.58	81.55	60.92	72.51
1,000	100	61	18.82	62.11	79.35	32.71	57.00	48.51	71.37	80.45	56.36	69.28
1,000	250	127	7.93	55.58	68.09	26.34	44.06	51.57	71.15	76.33	59.31	66.70
10,000	25	17	58.15	64.09	76.91	58.79	68.38	64.56	68.74	77.58	64.98	71.62
10,000	50	33	47.15	69.59	77.73	55.27	66.59	61.21	74.98	79.95	66.44	73.30
10,000	100	61	28.31	58.41	70.46	41.07	51.80	53.61	70.90	77.72	61.00	67.21
10,000	250	127	13.26	55.26	61.13	32.33	43.95	53.87	72.98	75.59	62.58	67.94

Note: The table reports the average summary statistics over all Monte Carlo replicates for the FKRB estimator (FKRB), and for our generalized estimator with tuning parameter μ from a 10-fold cross-validation and the MSE criterion (MSE), the one-standard-error rule (OneSe), the log-likelihood criterion (LL) and the number of correctly predicted binary outcomes (PredOut). The predicted binary outcome is set to one for the alternative with the highest estimated choice probability.

Table 2.A.3: First Stage Output of Mode Canada Data: Semiparametric Estimation with Normally Distributed Random Coefficient for the Total Travel Time.

	Dependent variable:
	Mode Choice
Intercept Train	-1.641^{***}
	(0.304)
Intercept Air	-7.153^{***}
	(0.913)
Frequency	0.077^{***}
	(0.008)
Cost	-0.009
	(0.009)
Income Train	-0.018^{***}
	(0.003)
Income Air	0.040^{***}
	(0.005)
Distance Train	0.002^{*}
	(0.001)
Distance Air	0.003^{***}
	(0.001)
Urban Train	1.722^{***}
	(0.163)
Urban Air	1.261^{***}
	(0.194)
Travel Time	-0.017^{***}
	(0.003)
sd.Travel Time	0.015^{***}
	(0.002)
Observations	3,593
Mc Fadden R^2	0.358
Log Likelihood	-2,340.700
LR Test	$2{,}615.034^{***}$ (df = 12) (p = 0.000)

Note: The table reports the mean estimates and standard errors (in brackets) obtained by the *mlogit package* for the semiparametric mixed logit model with normally distributed travel time.
*p<0.1; **p<0.05; ***p<0.01.

Table 2.A.4: Estimated Own- and Cross-Travel Time Elasticities in Mode Canada Data.

Elasticities estimated with FKRB:

	Car	Air	Train
Car	-0.8992 (-0.8444)	1.3982 (0.6692)	0.1164 (0.129)
Air	0.5895 (0.5943)	-1.2267 (-0.5079)	0.2049 (0.1589)
Train	-0.1622 (0.0346)	0.1840 (0.1352)	-0.6712 (-0.8861)

Elasticities estimated with ENet:

	Car	Air	Train
Car	-0.8382 (-0.7731)	1.4082 (0.682)	0.1473 (0.1009)
Air	0.5312 (0.5034)	-1.2581 (-0.5704)	0.1765 (0.1339)
Train	-0.0887 (0.036)	0.1900 (0.1118)	-0.6285 (-0.7691)

Elasticities estimated semiparametrically:

	Car	Air	Train
Car	-0.8567 (-0.7483)	1.4115 (0.7221)	0.2511 (0.1621)
Air	0.4938 (0.4481)	-1.3251 (-0.6791)	0.1595 (0.1051)
Train	0.0138 (0.0466)	0.2322 (0.1004)	-0.7057 (-0.8399)

Note: The table reports the mean and the median (in brackets) over individuals' own- and cross-travel time elasticities for the FKRB estimator, the elastic net estimator, and the semiparametric mixed logit with normal distribution. The reported numbers correspond to the percentage change of the choice probability of an alternative in a column after a one percent increase in the travel time of an alternative in a row.

Table 2.A.5: Ratio of Estimated Own- and Cross-Travel Time Elasticities in Mode Canada Data.

Estimated Elasticities of FKRB divided by those of ENet:			
	Car	Air	Train
Car	1.0728 (1.0922)	0.9929 (0.9813)	0.7908 (1.2783)
Air	1.1099 (1.1804)	0.9750 (0.8905)	1.1605 (1.1864)
Train	1.8291 (0.9611)	0.9685 (1.2098)	1.0680 (1.1521)
Semiparametrically estimated Elasticities divided by those of ENet:			
	Car	Air	Train
Car	1.0221 (0.9679)	1.0023 (1.0589)	1.7054 (1.6064)
Air	0.9296 (0.8901)	1.0533 (1.1906)	0.9032 (0.7846)
Train	-0.1559 (1.2961)	1.2221 (0.8984)	1.1230 (1.0920)

Note: The table reports the ratio of the mean and the median (in brackets) over individuals' own- and cross-travel time elasticities reported in Table 2.A.4 for (1) the FKRB estimator and elastic net estimator and (2) the semiparametric mixed logit with normal distribution and the elastic net estimator.

Appendix 2.B Algorithm to Update Fixed and Random Coefficients

The algorithm to update the fixed coefficients uses a modification of the flexible grid estimator in Train (2008). Let F denote the set of indices corresponding to the fixed coefficients and M to the set of indices corresponding to the random coefficients. The goal is to maximize with respect to the fixed coefficients β^F and the weights $\theta = (\theta_1, \ldots, \theta_R)$ corresponding to β^M. Therefore, define the vector which is to be maximized as $\pi = \{\beta_F, \theta\}$. Then, rewrite $z^r_{i,j}$ more explicitly:

$$z^r_{i,j} := z_{i,j}(\beta^F, \beta^M_r) = g(x_{i,j}, \beta^F, \beta^M_r) = \frac{\exp\left(x^F_{i,j}\beta^F + x^M_{i,j}\beta^M_r\right)}{1 + \sum_{l=1}^{J} \exp\left(x^F_{i,l}\beta^F + x^M_{i,l}\beta^M_r\right)}. \tag{2.B.1}$$

The likelihood criterion given in Train (2008) is

$$LL(\beta^F, \beta^M) = \frac{1}{N}\sum_{i=1}^{N} \log\left(\sum_{r=1}^{R} \theta_r z^r_{i,y_i}\right) = \frac{1}{N}\sum_{i=1}^{N} \log\left(\sum_{r=1}^{R} \theta_r z_{i,y_i}(\beta^F, \beta^M_r)\right).$$

The probability of agent i having coefficients π conditional on her observed choice y_i

and being type r is

$$h_{i,r}(\pi) = \frac{\theta_r z_{i,y_i}(\beta^F, \beta_r^M)}{\sum_{r=1}^{R} \theta_r z_{i,y_i}(\beta^F, \beta_r^M)}. \tag{2.B.2}$$

Based on Equation (2.B.2) one can derive the iterative EM update scheme which updates $\pi^{t+1} = \{\beta_F, \theta\}^{t+1} = \{\beta_F, (\theta_1, \ldots, \theta_R)\}^{t+1}$ by using a previous estimated trial π^t to maximize

$$\begin{aligned}\pi^{t+1} &= \arg\max_{\pi} Q\left(\pi|\pi^t\right) \\ &= \arg\max_{\pi} \sum_{i=1}^{N}\sum_{r=1}^{R} h_{i,r}\left(\pi^t\right) \log\left(\theta_r z_{i,y_i}(\beta^F, \beta_r^M)\right).\end{aligned} \tag{2.B.3}$$

Since $\log\left(\theta_r z_{i,j}(\beta^F, \beta_r^M)\right) = \log(\theta_r) + \log(z_{i,y_i}(\beta^F, \beta_r^M))$ one can maximize Equation (2.B.3) separately for β^F and θ. Since we use our generalized estimator given in Equation (2.7), we only maximize Equation (2.B.3) over β^F:

$$\{\beta^F\}^{t+1} = \arg\max_{\beta^F} \sum_{i=1}^{N}\sum_{r=1}^{R} h_{i,r}\left(\pi^t\right) \log\left(z_{i,y_i}(\beta^F, \beta_r^M)\right). \tag{2.B.4}$$

Plugging Equation (2.B.1) into Equation (2.B.4) gives

$$\{\beta^F\}^{t+1} = \arg\max_{\beta^F} \sum_{i=1}^{N}\sum_{r=1}^{R} h_{i,r}\left(\pi^t\right) \log\left(\frac{\exp\left(x_{i,y_i}^F \beta^F + x_{i,y_i}^M \beta_r^M\right)}{1 + \sum_{l=1}^{J} \exp\left(x_{i,l}^F \beta^F + x_{i,l}^M \beta_r^M\right)}\right) \tag{2.B.5}$$

or equivalently

$$\{\beta^F\}^{t+1} = \arg\max_{\beta^F} \sum_{i=1}^{N}\sum_{j=1}^{J}\sum_{r=1}^{R} y_{i,j} h_{i,r}\left(\pi^t\right) \log\left(\frac{\exp\left(x_{i,j}^F \beta^F + x_{i,j}^M \beta_r^M\right)}{1 + \sum_{l=1}^{J} \exp\left(x_{i,l}^F \beta^F + x_{i,l}^M \beta_r^M\right)}\right) \tag{2.B.6}$$

This is is the formula of a weighted (standard) logit model where only the coefficients β^F are to be maximized and the coefficients β^M are treated as constants. The weights $h_{i,r}(\pi^t)$, calculated as given in Equation (2.B.2), do not depend on the product j, but differ for different observations i and grid points r.

The whole update scheme is given by the following steps

Generalized Estimator of Equation (2.7) ***with fixed and random coefficients***

1. Estimate semi-parametric model with all regressors and store the coefficients of the fixed parameters β_0^F.
2. Choose the grid points β_r^M, $r = 1, ..., R$.
3. Calculate the logit kernel, $z_{i,j}(\beta_0^F, \beta_r^M)$, for each agent at each point.
4. Estimate θ_0 using the Generalized Estimator in Equation (2.7).
5. Calculate weights for each agent at each point with $\pi_0 = \{\beta_0^F, \theta_0\}$ as
$$h_{i,r}(\pi_0) = \frac{\theta_{r0} z_{i,y_i}(\beta_0^F, \beta_r^M)}{\sum_{r=1}^{R} \theta_{r0} z_{i,y_i}(\beta_0^F, \beta_r^M)}.$$
6. Update the fixed coefficients $\beta_0^F = \beta_1^F$ by estimating a weighted standard logit as specified in Equation (2.B.6) .
7. Repeat steps 3 and 6 until convergence, using the updated coefficients $\pi_0 = \pi_1$, where $\theta_0 = \theta_1$ is updated in step 4.
8. Use these estimated weights $\widehat{\theta}$ to calculate the estimated distribution
$$\hat{F}(\beta) = \sum_{r=1}^{R} \hat{\theta}_r \, 1[\beta_r \leq \beta].$$

Appendix 2.C Proofs of Results in Section 2.3

Below, we provide the proofs of the results presented in Section 2.3. For that purpose, we first introduce some additional notation. Let A be a $m \times n$ matrix and x be a $n \times 1$ vector. In the following, the $\|A\|_\infty$ norm refers to the matrix norm induced by the maximum norm of vectors. Then

$$\|A\|_\infty := \max_{||x||_\infty = 1} \|Ax\|_\infty = \max_{1 \le i \le m} \sum_{j=1}^{n} |a_{ij}|$$

denotes the maximum row sum of matrix A. $\|x\|_\infty$ refers to the largest absolute element of vector x. Similarly, $\|A\|_2$ is defined as the matrix norm induced by the euclidean vector norm. That is,

$$\|A\|_2 := \max_{||x||_2 = 1} \|Ax\|_2 ,$$

is called spectral norm. It can be shown that $\|A\|_2 = \max_{1 \le i \le n} \sqrt{\psi_i(A^T A)}$ where $\psi_i(A^T A)$ denotes the eigenvalues of $A^T A$.

2.C.1 Proof of Probability Bound

Lemma 2.C.1 uses Hoeffding's inequality to derive a probability bound for sub-Gaussian random variables. We use the lemma in the proofs of Theorems 2.1 - 2.3.

Lemma 2.C.1. *Suppose Assumption 2.1 holds. Then, for* $\gamma \ge 0$

$$\mathbb{P}\left(\left\| \frac{1}{NJ} \tilde{Z}^T \epsilon \right\|_\infty \ge \gamma \right) \le 2(R-1)J \exp\left(-\frac{N\gamma^2}{2} \right).$$

Proof. Notice that

$$\mathbb{P}\left(\left\| \frac{1}{NJ} \tilde{Z}^T \epsilon \right\|_\infty \ge \gamma \right) = \mathbb{P}\left(\max_{1 \le r \le R-1} \left| \frac{1}{NJ} \sum_{i=1}^{N} \tilde{Z}_i^{rT} \epsilon_i \right| \ge \gamma \right) \tag{2.C.1}$$

where $\epsilon_i = (\epsilon_{i,1}, \ldots, \epsilon_{i,J})$ denotes a random vector of J dependent variables such that Equation (2.C.1) can equivalently be written as

$$\begin{aligned} \mathbb{P}\left(\max_{1 \le r \le R-1} \left| \frac{1}{NJ} \sum_{i=1}^{N} \tilde{Z}_i^{rT} \epsilon_i \right| \ge \gamma \right) &= \mathbb{P}\left(\max_{1 \le r \le R-1} \left| \frac{1}{NJ} \sum_{i=1}^{N} \sum_{j=1}^{J} \tilde{z}_{i,j}^{r} \epsilon_{i,j} \right| \ge \gamma \right) \\ &= \mathbb{P}\left(\bigcup_{1 \le r \le R-1} \left\{ \left| \frac{1}{NJ} \sum_{i=1}^{N} \sum_{j=1}^{J} \tilde{z}_{i,j}^{r} \epsilon_{i,j} \right| \ge \gamma \right\} \right). \end{aligned}$$

From $\sum_{i=1}^{N}\sum_{j=1}^{J}\tilde{z}_{i,j}^{r}\epsilon_{i,j} \leq J \max_{1\leq j\leq J}\sum_{i=1}^{N}\tilde{z}_{i,j}^{r}\epsilon_{i,j}$, we obtain the upper bound

$$\begin{aligned}
\mathbb{P}\left(\bigcup_{1\leq r\leq R-1}\left\{\left|\frac{1}{NJ}\sum_{i=1}^{N}\sum_{j=1}^{J}\tilde{z}_{i,j}^{r}\epsilon_{i,j}\right|\geq\gamma\right\}\right) &\leq \mathbb{P}\left(\bigcup_{1\leq r\leq R-1}\left\{J\max_{1\leq j\leq J}\left|\frac{1}{NJ}\sum_{i=1}^{N}\tilde{z}_{i,j}^{r}\epsilon_{i,j}\right|\geq\gamma\right\}\right)\\
&\leq \sum_{r=1}^{R-1}\mathbb{P}\left(\max_{1\leq j\leq J}\left|\frac{1}{N}\sum_{i=1}^{N}\tilde{z}_{i,j}^{r}\epsilon_{i,j}\right|\geq\gamma\right)\\
&= \sum_{r=1}^{R-1}\mathbb{P}\left(\bigcup_{1\leq j\leq J}\left\{\left|\frac{1}{N}\sum_{i=1}^{N}\tilde{z}_{i,j}^{r}\epsilon_{i,j}\right|\geq\gamma\right\}\right)\\
&\leq \sum_{r=1}^{R-1}\sum_{j=1}^{J}\mathbb{P}\left(\left|\frac{1}{N}\sum_{i=1}^{N}\tilde{z}_{i,j}^{r}\epsilon_{i,j}\right|\geq\gamma\right)\\
&\leq (R-1)J\max_{\substack{1\leq r\leq R-1\\1\leq j\leq J}}\mathbb{P}\left(\left|\frac{1}{N}\sum_{i=1}^{N}\tilde{z}_{i,j}^{r}\epsilon_{i,j}\right|\geq\gamma\right).
\end{aligned}$$

Recall from Assumption 2.1(iii) and Equation (2.9) that $-1\leq\tilde{z}_{i,j}^{r}\leq 1$ and $-1\leq\epsilon_{i,j}\leq 1$. Therefore, $\xi := (\tilde{z}_{1,j}^{r}\epsilon_{1,j},\ldots,\tilde{z}_{N,j}^{r}\epsilon_{N,j})$ is a vector of independent uniformly bounded random variables since for every $i=1,\ldots,N$ it holds that $-1\leq\tilde{z}_{i,j}^{r}\epsilon_{i,j}\leq 1$. It follows from the assumption of conditional exogeneity (Assumption 2.1(iv)) that $\mathbb{E}[\xi]=0$. Due to the boundedness of ξ_i, $i=1,\ldots,N$, its moment generating function satisfies

$$\mathbb{E}\left[\exp(s\xi_i)\right]\leq\exp\left(\frac{\sigma^2 s^2}{2}\right).$$

For any $s\in\mathbb{R}$, ξ_i is said to be sub-Gaussian with variance proxy σ^2. Thus, using Hoeffding's inequality,

$$\max_{\substack{1\leq r\leq R-1\\1\leq j\leq J}}\mathbb{P}\left(\left|\frac{1}{N}\sum_{i=1}^{N}\tilde{z}_{i,j}^{r}\epsilon_{i,j}\right|\geq\gamma\right)\leq 2\exp\left(-\frac{N\gamma^2}{2\sigma^2}\right).$$

It follows from $\xi_i\in[-1,1]$ that $\sigma^2=1$. Therefore,

$$\begin{aligned}
\mathbb{P}\left(\left\|\frac{1}{NJ}\tilde{Z}^{T}\epsilon\right\|_{\infty}\geq\gamma\right) &\leq (R-1)J\max_{\substack{1\leq r\leq R-1\\1\leq j\leq J}}\mathbb{P}\left(\left|\frac{1}{N}\sum_{i=1}^{N}\tilde{z}_{i,j}^{r}\epsilon_{i,j}\right|\geq\gamma\right)\\
&\leq 2(R-1)J\exp\left(-\frac{N\gamma^2}{2}\right).
\end{aligned}$$

□

2.C.2 Proof of Selection Consistency

In the following, we provide the proof of Theorem 2.1. We first derive two sufficient conditions in Lemma 2.C.3 that ensure that the estimated weights are equal in sign, i.e. $\hat{\theta} =_s \theta^*$. Lemma 2.C.4 provides a bound on the probability of the first sufficient condition and Lemma 2.C.5 a bound on the probability of the second sufficient condition. Finally, we use Lemma 2.C.4 and Lemma 2.C.5 to prove Theorem 2.1. Both Lemma 2.C.4 and Lemma 2.C.5 employ Lemma 2.C.2. To keep notation uncluttered, we drop the dependence of $R(N)$, $s(N)$, $\xi^S_{\min}(\mu, N)$ and $\rho(\mu, N)$ on N and write R, s, $\xi^S_{\min}(\mu)$ and $\rho(\mu)$ in the subsequent proofs.

Lemma 2.C.2. *It holds that*

$$\left\| \left(\frac{1}{NJ} \tilde{Z}_S^T \tilde{Z}_S + \mu I_S \right)^{-1} \right\|_\infty \leq \sqrt{s} \frac{1}{\xi^S_{\min}(\mu)}.$$

Proof. Using Singular Value Decomposition (SVD), rewrite $\tilde{Z}_S$ as

$$\frac{1}{\sqrt{NJ}} \tilde{Z}_S = ADM^T \tag{2.C.2}$$

where A is a $NJ \times s$ matrix with orthogonal columns, i.e. $A^T A = I_S$.
M is a $s \times s$ orthogonal matrix satisfying $M^T M = MM^T = I_S$. D is a diagonal $s \times s$ matrix consisting of the singular values of $(1/\sqrt{NJ})\tilde{Z}_S$ on its diagonal. We apply the SVD in Equation (2.C.2) to rewrite

$$\left(\frac{1}{NJ} \tilde{Z}_S^T \tilde{Z}_S + \mu I_S \right)^{-1} = \left(MD^T A^T ADM^T + \mu I_S \right)^{-1} = \left(MD^2 M^T + \mu MM^T \right)^{-1}$$

$$= M \left(D^2 + \mu I_S \right)^{-1} M^T$$

Therefore,

$$\left\| \left(\frac{1}{NJ} \tilde{Z}_S^T \tilde{Z}_S + \mu I_S \right)^{-1} \right\|_\infty = \left\| M \left(D^2 + \mu I_S \right)^{-1} M^T \right\|_\infty \leq \sqrt{s} \left\| M \left(D^2 + \mu I_S \right)^{-1} M^T \right\|_2 \tag{2.C.3}$$

$$= \sqrt{s} \left\| \left(D^2 + \mu I_S \right)^{-1} \right\|_2 = \sqrt{s} \max_{i \in S} \sqrt{\psi_i}$$

$$= \sqrt{s} \max_{i \in S} \frac{1}{d_{ii}^2 + \mu} = \sqrt{s} \frac{1}{\min_{i \in S} d_{ii}^2 + \mu} = \sqrt{s} \frac{1}{\xi^S_{\min}(\mu)}$$

where ψ_i denotes the eigenvalues of $\left(\left(D^2+\mu I_S\right)^{-1}\right)^T \left(D^2+\mu I_S\right)^{-1} = \left(D^2+\mu I_S\right)^{-2}$. Thus, $\psi_i = \left(d_{ii}^2+\mu\right)^{-2}$, as the eigenvalues of a diagonal matrix are its diagonal entries. The (unrestricted) eigenvalues of $1/(NJ)\tilde{Z}_S^T\tilde{Z}_S + \mu I_S$ are defined as $\xi^S(\mu)$. $\xi^S_{\min}(\mu)$ corresponds to the minimal eigenvalue of the matrix. The first inequality in Equation (2.C.3) holds by the relation of the absolute row sum norm and the spectral norm. The transformation from the first to the second line follows from the invariance of the spectral norm to orthogonal transformations (Gentle, 2007, pp. 130-131). The equality in the second line follows from the spectral norm. The last equality in Equation (2.C.3) holds by the relation of singular values to eigenvalues. □

Lemma 2.C.3. *Sufficient conditions for* $\hat{\theta} =_s \theta^*$ *are*

$$\mathcal{M}(V) := \left\{ \max_{j \in S^C} V_j \leq \lambda \right\},$$

$$\mathcal{M}(U) := \left\{ \max_{i \in S} |U_i| < \rho(\mu) \right\}$$

where

$$V := \frac{1}{NJ}\tilde{Z}_{S^C}^T \left[\tilde{Z}_S \left(\frac{1}{NJ}\tilde{Z}_S^T\tilde{Z}_S + \mu I_S \right)^{-1} \left(\lambda \iota_S + \mu\theta_S^* - \frac{1}{NJ}\tilde{Z}_S^T\epsilon \right) + \epsilon \right],$$

$$U := \left(\frac{1}{NJ}\tilde{Z}_S^T\tilde{Z}_S + \mu I_S \right)^{-1} \frac{1}{NJ}\tilde{Z}_S^T\epsilon,$$

$$\rho(\mu) := \min_{i \in S} \left| \left(\frac{1}{NJ}\tilde{Z}_S^T\tilde{Z}_S + \mu I_S \right)^{-1} \left(\frac{1}{NJ}\tilde{Z}_S^T\tilde{Z}_S\theta_S^* - \lambda\iota_S \right) \right|.$$

Proof. The Lagrangian of our generalized estimator in Equation (2.8) formulated in matrix notation is given by

$$L(\theta) := \quad \frac{1}{2NJ}||\tilde{y} - \tilde{Z}\theta||_2^2 + \lambda\left(\iota^T\theta - 1\right) + \frac{1}{2}\mu\ \theta^T\theta - \nu^T\theta \tag{2.C.4}$$

which is minimized with respect to θ, i.e. $\theta = \arg\min_{\theta} L(\theta)$. λ and ν are Lagrangian multipliers that enforce that the estimated weights sum to one and that they are non-negative respectively. $\mu > 0$ is an additional tuning parameter. Note that for $\mu = 0$, Equation (2.C.4) corresponds to the objective function of the estimator by Fox et al. (2011).

To analyze the support recovery of our estimator, we follow the proof in Jia and Yu (2010). The estimator recovers the true support of the distribution if every estimated probability weight $\hat{\theta}$ has the same sign as the true weights θ^*, i.e. $\hat{\theta} =_s \theta^*$. This is the case if the Karush-Kuhn-Tucker (KKT) conditions to the optimization problem in

Equation (2.C.4) are satisfied. The KKT conditions are given by

$$-\frac{1}{NJ}\tilde{Z}^T\left(\tilde{y}-\tilde{Z}\hat{\theta}\right)+\lambda\iota+\mu\,\hat{\theta}-\nu=0, \tag{2.C.5}$$

$$\lambda\left(\iota^T\hat{\theta}-1\right)=0,$$

$$\nu_r\,\hat{\theta}_r=0,$$

$$\lambda\geq 0,\quad \nu_r\geq 0 \qquad\qquad \forall\quad r=1,\ldots,R-1.$$

Denote the set of grid points where the true distribution has positive probability mass by $S=\{r\in\{1,\ldots,R-1\}|\theta_r^*>0\}$ and let $S^C=\{r\in\{1,\ldots,R-1\}|\theta_r^*=0\}$ denote its complement set. The corresponding cardinalities are defined as $s:=|S|$ and $s^C:=|S^C|$. We refer to grid points in S as active grid points and to grid points in S^C as inactive grid points. Splitting $\hat{\theta}$, $\tilde{Z}$ and ν over S and S^C into two blocks gives

$$-\frac{1}{NJ}\left[\tilde{Z}_S\ \tilde{Z}_{S^C}\right]^T\left(\tilde{y}-\left[\tilde{Z}_S\ \tilde{Z}_{S^C}\right]\begin{pmatrix}\hat{\theta}_S\\ \hat{\theta}_{S^C}\end{pmatrix}\right)+\lambda\iota+\mu\begin{pmatrix}\hat{\theta}_S\\ \hat{\theta}_{S^C}\end{pmatrix}-\begin{pmatrix}\nu_S\\ \nu_{S^C}\end{pmatrix}=0.$$

Recall that $\theta_r^*=0$ for all grid points outside S, so that $\tilde{Z}\theta^*=\tilde{Z}_S\theta_S^*$. In order to recover the active grid points, it must hold that $\hat{\theta}=_s\theta^*$ which implies $\hat{\theta}_{S^C}=0$. The two conditions that follow from Equation (2.C.5) require

$$-\frac{1}{NJ}\tilde{Z}_S^T\left(\tilde{y}-\tilde{Z}_S\hat{\theta}_S\right)+\lambda\iota_S+\mu\hat{\theta}_S-\nu_S=0, \tag{2.C.6}$$

$$-\frac{1}{NJ}\tilde{Z}_{S^C}^T\left(\tilde{y}-\tilde{Z}_S\hat{\theta}_S\right)+\lambda\iota_{S^C}-\nu_{S^C}=0.$$

Note that $\hat{\theta}_S>0$ and $\hat{\theta}_{S^C}=0$ imply

$$\nu_r=0 \qquad\qquad \forall\quad r\in S, \tag{2.C.7}$$

$$\nu_r\geq 0 \qquad\qquad \forall\quad r\notin S. \tag{2.C.8}$$

It follows from Condition (2.C.7) that Condition (2.C.6) simplifies to

$$-\frac{1}{NJ}\tilde{Z}_S^T\left(\tilde{y}-\tilde{Z}_S\hat{\theta}_S\right)+\lambda\iota_S+\mu\hat{\theta}_S=0.$$

Substituting the true model $\tilde{y}=\tilde{Z}\theta^*+\epsilon$, we can re-express the required conditions as

$$-\frac{1}{NJ}\tilde{Z}_S^T\tilde{Z}_S\left(\theta_S^*-\hat{\theta}_S\right)-\frac{1}{NJ}\tilde{Z}_S^T\epsilon+\lambda\iota_S+\mu\hat{\theta}_S=0 \tag{2.C.9}$$

and

$$-\frac{1}{NJ}\tilde{Z}_{S^C}^T\tilde{Z}_S\left(\theta_S^*-\hat{\theta}_S\right)-\frac{1}{NJ}\tilde{Z}_{S^C}^T\epsilon+\lambda\iota_{S^C}-\nu_{S^C}=0. \tag{2.C.10}$$

Reformulating Condition (2.C.9) gives

$$\hat{\theta}_S = \underbrace{\left(\frac{1}{NJ}\tilde{Z}_S^T\tilde{Z}_S + \mu I_S\right)^{-1}\left(\frac{1}{NJ}\tilde{Z}_S^T\epsilon}_{=:U} + \frac{1}{NJ}\tilde{Z}_S^T\tilde{Z}_S\theta_S^* - \lambda\iota_S\right) > 0 \tag{2.C.11}$$

where the positivity constraint follows from the KKT conditions and the definition of $\hat{\theta}_S$.

Plugging Equation (2.C.11) into Equation (2.C.10) and using Condition (2.C.8) yields

$$\underbrace{\frac{1}{NJ}\tilde{Z}_{S^C}^T\left[\tilde{Z}_S\left(\frac{1}{NJ}\tilde{Z}_S^T\tilde{Z}_S + \mu I_S\right)^{-1}\left(\lambda\iota_S + \mu\theta_S^* - \frac{1}{NJ}\tilde{Z}_S^T\epsilon\right) + \epsilon\right]}_{=:V} \leq \lambda\iota_{S^C}. \tag{2.C.12}$$

U and V are defined in Equation (2.C.11) and Equation (2.C.12), respectively.

The vector U consists of s elements U_i, $i \in S$, and is constructed from the conditions on the positive weights, and vector V from the condition on the zero weights. Therefore, V has $R - s$ elements V_j, $j \in S^C$. Condition (2.C.12) is equivalent to the event

$$\mathcal{M}(V) := \left\{\max_{j \in S^C} V_j \leq \lambda\right\}.$$

The event $\mathcal{M}(U)$ defines a condition for the positive weights

$$\mathcal{M}(U) := \left\{\max_{i \in S} |U_i| < \rho(\mu)\right\}$$

where $\rho(\mu) := \min_{i \in S} |g_i|$ with $g_i := \left[\left(\frac{1}{NJ}\tilde{Z}_S^T\tilde{Z}_S + \mu I_S\right)^{-1}\left(\frac{1}{NJ}\tilde{Z}_S^T\tilde{Z}_S\theta_S^* - \lambda\iota_S\right)\right]_i$. Therefore, the event $\mathcal{M}(U)$ implies

$$0 < \rho(\mu) - \max_{i \in S}|U_i| < \rho(\mu) - |U_i| < |g_i| - |U_i| < |g_i + U_i| = |\hat{\theta}_{S_i}| = \hat{\theta}_{S_i}, \quad \forall i \in S$$

where g_i, U_i and $\hat{\theta}_{S_i}$ denote the ith element of the respective vectors g, U and $\hat{\theta}_S$. The second last equality holds by definition of g_i and U_i (see Equation (2.C.11)) and the last inequality by the reverse triangle inequality. Because the weights are constrained to be nonnegative by the KKT conditions, the absolute value $|\hat{\theta}_{S_i}|$ can be omitted. Consequently, $\mathcal{M}(U)$ is a sufficient condition for Equation (2.C.11) to hold and thus for $\hat{\theta}_S > 0$.

□

Lemma 2.C.4. *Suppose Assumption 2.1 holds. Suppose further that the NEIC holds. Let $\mathcal{M}^C(V)$ denote the complement of $\mathcal{M}(V)$. Then,*

$$\mathbb{P}\left(\mathcal{M}^C(V)\right) \leq 2(R-1)J \exp\left(-\frac{N\eta^2\lambda^2\left(\frac{\xi^S_{\min}(\mu)}{s\sqrt{s}+\xi^S_{\min}(\mu)}\right)^2}{2}\right).$$

Proof. V_j is sub-Gaussian with mean

$$\overline{V} := E(V) = \frac{1}{NJ}\tilde{Z}^T_{S^C}\tilde{Z}_S\left(\frac{1}{NJ}\tilde{Z}^T_S\tilde{Z}_S + \mu I_S\right)^{-1}(\lambda\iota_S + \mu\theta^*_S).$$

Recall the Nonnegative Elastic Net Irrepresentable Condition (NEIC) is

$$\max_{r\in S^C}\frac{1}{NJ}\tilde{Z}^T_{S^C}\tilde{Z}_S\left(\frac{1}{NJ}\tilde{Z}^T_S\tilde{Z}_S + \mu I_S\right)^{-1}\left(\iota_S + \frac{\mu}{\lambda}\theta^*_S\right) \leq 1-\eta.$$

Therefore, $\overline{V}_j \leq (1-\eta)\lambda$. Let $\tilde{V} := \frac{1}{NJ}\tilde{Z}^T_{S^C}\Big[-\tilde{Z}_S\left(\frac{1}{NJ}\tilde{Z}^T_S\tilde{Z}_S + \mu I_S\right)^{-1}\frac{1}{NJ}\tilde{Z}^T_S + I_{NJ}\Big]\epsilon$ such that $V = \overline{V} + \tilde{V}$. Consequently, it holds for the complement of $\mathcal{M}(V)$ that $\lambda < \max_{j\in S^C} V_j = \max_{j\in S^C}(\overline{V}_j + \tilde{V}_j) \leq \max_{j\in S^C}\overline{V}_j + \max_{j\in S^C}\tilde{V}_j \iff \max_{j\in S^C}\tilde{V}_j > \lambda - \max_{j\in S^C}\overline{V}_j \geq \lambda - (1-\eta)\lambda = \eta\lambda$.

We use the last inequality to derive an upper bound on $\mathcal{M}^C(V)$:

$$\begin{aligned}
\mathbb{P}\left(\mathcal{M}^C(V)\right) &= \mathbb{P}\left(\max_{j\in S^C} V_j > \lambda\right) \leq \mathbb{P}\left(\max_{j\in S^C}\tilde{V}_j > \eta\lambda\right) \leq \mathbb{P}\left(\max_{j\in S^C}|\tilde{V}_j| > \eta\lambda\right) \\
&= \mathbb{P}\left(\max_{j\in S^C}\left|\frac{1}{NJ}\tilde{Z}^T_{S^C}\Big[-\tilde{Z}_S\left(\frac{1}{NJ}\tilde{Z}^T_S\tilde{Z}_S + \mu I_S\right)^{-1}\frac{1}{NJ}\tilde{Z}^T_S + I\Big]\epsilon\right| > \eta\lambda\right) \\
&\leq \mathbb{P}\left(\max_{j\in S^C}\left|\frac{1}{NJ}\tilde{Z}^T_{S^C}\tilde{Z}_S\left(\frac{1}{NJ}\tilde{Z}^T_S\tilde{Z}_S + \mu I_S\right)^{-1}\frac{1}{NJ}\tilde{Z}^T_S\epsilon\right| + \max_{j\in S^C}\left|\frac{1}{NJ}\tilde{Z}^T_{S^C}\epsilon\right| > \eta\lambda\right) \\
&= \mathbb{P}\left(\left\|\frac{1}{NJ}\tilde{Z}^T_{S^C}\tilde{Z}_S\left(\frac{1}{NJ}\tilde{Z}^T_S\tilde{Z}_S + \mu I_S\right)^{-1}\frac{1}{NJ}\tilde{Z}^T_S\epsilon\right\|_\infty + \max_{j\in S^C}\left|\frac{1}{NJ}\tilde{Z}^T_{S^C}\epsilon\right| > \eta\lambda\right) \\
&\leq \mathbb{P}\left(\left\|\frac{1}{NJ}\tilde{Z}^T_{S^C}\tilde{Z}_S\right\|_\infty\left\|\left(\frac{1}{NJ}\tilde{Z}^T_S\tilde{Z}_S + \mu I_S\right)^{-1}\right\|_\infty\left\|\frac{1}{NJ}\tilde{Z}^T_S\epsilon\right\|_\infty + \max_{j\in S^C}\left|\frac{1}{NJ}\tilde{Z}^T_{S^C}\epsilon\right| > \eta\lambda\right).
\end{aligned}$$

The last inequality holds due the property of the absolute row sum norm that $\|ABx\|_\infty \leq \|A\|_\infty\|B\|_\infty\|x\|_\infty$ for arbitrary matrices A, B and a vector x.

By Lemma 2.C.2 and $\left\|\frac{1}{NJ}\tilde{Z}^T_{S^C}\tilde{Z}_S\right\|_\infty \leq s$ (since every entry in $\tilde{Z}$ is at most 1 in

absolute value, and thus the absolute row sum of $\frac{1}{NJ}\tilde{Z}_{S^C}^T\tilde{Z}_S$ at most $\frac{1}{NJ}sNJ = s$), we obtain

$$\begin{aligned}
\mathbb{P}\left(\mathcal{M}^C(V)\right) &\leq \mathbb{P}\left(s\sqrt{s}\frac{1}{\xi_{\min}^S(\mu)}\max_{j\in S}\left|\frac{1}{NJ}\tilde{Z}_{S^C}^T\epsilon\right| + \max_{j\in S^C}\left|\frac{1}{NJ}\tilde{Z}_{S^C}^T\epsilon\right| > \eta\lambda\right) \\
&\leq \mathbb{P}\left(s\sqrt{s}\frac{1}{\xi_{\min}^S(\mu)}\max_{j\in R}\left|\frac{1}{NJ}\tilde{Z}^T\epsilon\right| + \max_{j\in R}\left|\frac{1}{NJ}\tilde{Z}^T\epsilon\right| > \eta\lambda\right) \\
&= \mathbb{P}\left(\left(s\sqrt{s}\frac{1}{\xi_{\min}^S(\mu)}+1\right)\max_{j\in R}\left|\frac{1}{NJ}\tilde{Z}^T\epsilon\right| > \eta\lambda\right) \\
&\leq \mathbb{P}\left(\max_{j\in R}\left|\frac{1}{NJ}\tilde{Z}^T\epsilon\right| > \eta\lambda\frac{1}{s\sqrt{s}\frac{1}{\xi_{\min}^S(\mu)}+1}\right).
\end{aligned}$$

Applying Hoeffding's inequality with $\gamma = \eta\lambda\frac{1}{s\sqrt{s}\frac{1}{\xi_{\min}^S(\mu)}+1}$ as outlined in Lemma 2.C.1 gives

$$\begin{aligned}
\mathbb{P}\left(\mathcal{M}^C(V)\right) &\leq 2(R-1)J\exp\left(-\frac{N\left(\eta\lambda\frac{1}{s\sqrt{s}\frac{1}{\xi_{\min}^S(\mu)}+1}\right)^2}{2\sigma^2}\right) \\
&= 2(R-1)J\exp\left(-\frac{N\left(\eta\lambda\frac{\xi_{\min}^S(\mu)}{s\sqrt{s}+\xi_{\min}^S(\mu)}\right)^2}{2\sigma^2}\right) \\
&= 2(R-1)J\exp\left(-\frac{N\eta^2\lambda^2\left(\frac{\xi_{\min}^S(\mu)}{s\sqrt{s}+\xi_{\min}^S(\mu)}\right)^2}{2}\right).
\end{aligned}$$

□

Remark 2.C.1. The above calculations can be simplified to for the baseline estimator, i.e. if $\mu = 0$. Assume that the NIC condition for LASSO holds (NEIC with $\mu = 0$). Additionally, note that it holds for $\mu \geq 0$ that

$$\left(\frac{1}{NJ}\tilde{Z}_S^T\tilde{Z}_S + \mu I_S\right)^{-1}\tilde{Z}_S^T = \tilde{Z}_S^T\left(\frac{1}{NJ}\tilde{Z}_S\tilde{Z}_S^T + \mu I_N\right)^{-1}.$$

Using the above equality for $\mu = 0$, we obtain

$$\begin{aligned}
\mathbb{P}\left(\max_{j\in S^C} V_j > \lambda\right) &\leq \mathbb{P}\left(\max_{j\in S^C} \tilde{V}_j > \eta\lambda\right) \leq \mathbb{P}\left(\max_{j\in S^C} |\tilde{V}_j| > \eta\lambda\right) \\
&= \mathbb{P}\left(\max_{j\in S^C}\left|\frac{1}{NJ}\tilde{Z}_{S^C}^T\Big[-\tilde{Z}_S\left(\frac{1}{NJ}\tilde{Z}_S^T\tilde{Z}_S\right)^{-1}\frac{1}{NJ}\tilde{Z}_S^T + I_S\Big]\epsilon\right| > \eta\lambda\right) \\
&= \mathbb{P}\left(\max_{j\in S^C}\left|\frac{1}{NJ}\tilde{Z}_{S^C}^T\Big[-\frac{1}{NJ}\tilde{Z}_S\tilde{Z}_S^T\left(\frac{1}{NJ}\tilde{Z}_S\tilde{Z}_S^T\right)^{-1} + I_S\Big]\epsilon\right| > \eta\lambda\right) \\
&= \mathbb{P}\left(\max_{j\in S^C}\left|\frac{1}{NJ}\tilde{Z}_{S^C}^T\Big[-I_S + I_S\Big]\epsilon\right| > \eta\lambda\right) \\
&= \mathbb{P}\left(0 > \eta\lambda\right) = 0
\end{aligned}$$

since $\eta\lambda > 0$.

Lemma 2.C.5. *Suppose Assumption 2.1 holds. Let $\mathcal{M}^C(U)$ denote the complement of $\mathcal{M}(U)$. Then,*

$$\mathbb{P}\left(\mathcal{M}^C(U)\right) \leq 2sJ\exp\left(-\frac{N\xi_{\min}^S(\mu)^2\rho(\mu)^2}{2s}\right).$$

Proof. Because U is sub-Gaussian with mean 0, the probability of the complement of $\mathcal{M}(U)$ corresponds to

$$\begin{aligned}
\mathbb{P}\left(\mathcal{M}^C(U)\right) &= \mathbb{P}\left(\max_{i\in S}|U_i| \geq \rho(\mu)\right) \\
&= \mathbb{P}\left(\max_{i\in S}\left(\frac{1}{NJ}\tilde{Z}_S^T\tilde{Z}_S + \mu I_S\right)^{-1}\frac{1}{NJ}\tilde{Z}_S^T\epsilon \geq \rho(\mu)\right) \\
&\leq \mathbb{P}\left(\left\|\left(\frac{1}{NJ}\tilde{Z}_S^T\tilde{Z}_S + \mu I_S\right)^{-1}\right\|_\infty \left\|\frac{1}{NJ}\tilde{Z}_S^T\epsilon\right\|_\infty \geq \rho(\mu)\right).
\end{aligned}$$

In the next step Lemma 2.C.2 is applied again.

$$\begin{aligned}
\mathbb{P}\left(\mathcal{M}^C(U)\right) &\leq \mathbb{P}\left(\sqrt{s}\frac{1}{\xi^S_{\min}(\mu)}\left\|\frac{1}{NJ}\tilde{Z}_S^T\epsilon\right\|_\infty \geq \rho(\mu)\right) \\
&\leq \mathbb{P}\left(\left\|\frac{1}{NJ}\tilde{Z}_S^T\epsilon\right\|_\infty \geq \xi^S_{\min}(\mu)\frac{1}{\sqrt{s}}\rho(\mu)\right) \\
&\leq 2sJ\exp\left(-\frac{N\left(\xi^S_{\min}(\mu)\frac{1}{\sqrt{s}}\rho(\mu)\right)^2}{2\sigma^2}\right) = 2sJ\exp\left(-\frac{N\xi^S_{\min}(\mu)^2\rho(\mu)^2}{2s\sigma^2}\right) \\
&= 2sJ\exp\left(-\frac{N\xi^S_{\min}(\mu)^2\rho(\mu)^2}{2s}\right)
\end{aligned}$$

where the last inequality follows from Hoeffding's inequality in Lemma 2.C.1 with $\gamma = \xi^S_{\min}(\mu)\frac{1}{\sqrt{s}}\rho(\mu)$. □

We use the above lemmata to prove Theorem 2.1.

Proof of Theorem 2.1.

It holds that

$$\mathbb{P}\left(\hat{\theta} =_s \theta\right) \geq \mathbb{P}\left(\mathcal{M}(V) \cap \mathcal{M}(U)\right)$$

since $\mathcal{M}(U)$ is a sufficient condition for the selection of the true weights according to Lemma 2.C.3.

Under the condition that RCDG holds, applying Lemma 2.C.4 and Lemma 2.C.5 gives $\lim_{N\to\infty} \mathbb{P}\left(\mathcal{M}^C(V)\right) = 0$ and $\lim_{N\to\infty} \mathbb{P}\left(\mathcal{M}^C(U)\right) = 0$.

Thus,

$$\begin{aligned}
\lim_{N\to\infty} \mathbb{P}\left(\hat{\theta} =_s \theta\right) &\geq \lim_{N\to\infty} \mathbb{P}\left(\mathcal{M}(V) \cap \mathcal{M}(U)\right) \\
&\geq \lim_{N\to\infty} \left\{1 - \mathbb{P}\left(\mathcal{M}^C(V)\right) - \mathbb{P}\left(\mathcal{M}^C(U)\right)\right\} \\
&= 1.
\end{aligned}$$

□

2.C.3 Proof of Error Bounds

In the following, we first provide the proof of the error bound of the estimated weights presented in Theorem 2.2 and the proof of Corollary 2.1. We then use the derived bound to proof the error bound of the estimated random coefficients' distribution in Theorem 2.3. In the proofs of Theorem 2.2 and Theorem 2.3, we apply Lemma 2.C.1.

Proof of Theorem 2.2.

Note that if $\hat{\theta}$ is the solution to the Lagrangian in Equation (2.C.4), it must hold that it minimizes (2.C.4), i.e. $L(\hat{\theta}) \leq L(\theta)$ for any θ. Thus, it holds that $L(\hat{\theta}) \leq L(\theta^*)$ where θ^* are the true weights. Applying this to the objective function in (2.C.4), we obtain

$$\frac{1}{2NJ}\left\|\tilde{y}-\tilde{Z}\hat{\theta}\right\|_2^2+\lambda\left(\iota^T\hat{\theta}-1\right)+\frac{\mu}{2}\hat{\theta}^T\hat{\theta}\leq\frac{1}{2NJ}\left\|\tilde{y}-\tilde{Z}\theta^*\right\|_2^2+\lambda\left(\iota^T\theta^*-1\right)+\frac{\mu}{2}\theta^{*T}\theta^*.$$

Substituting the true model $\tilde{y}=\tilde{Z}\theta^*+\epsilon$ into the above condition and simplifying gives

$$\frac{1}{2NJ}\left\|\tilde{Z}\left(\theta^*-\hat{\theta}\right)+\epsilon\right\|_2^2+\lambda\left(\iota^T\hat{\theta}-1\right)+\frac{\mu}{2}\hat{\theta}^T\hat{\theta}\leq\frac{1}{2NJ}\left\|\epsilon\right\|_2^2+\lambda\left(\iota^T\theta^*-1\right)+\frac{\mu}{2}\theta^{*T}\theta^*.$$

Taking into account that

$$\left\|\tilde{Z}(\theta^*-\hat{\theta})+\epsilon\right\|_2^2=\left\|\tilde{Z}(\theta^*-\hat{\theta})\right\|_2^2+\left\|\epsilon\right\|_2^2+2\epsilon^T(\tilde{Z}(\theta^*-\hat{\theta}))$$

we obtain

$$\begin{aligned}&\frac{1}{2NJ}\left\|\tilde{Z}\left(\theta^*-\hat{\theta}\right)\right\|_2^2+\lambda\left(\iota^T\hat{\theta}-1\right)+\frac{\mu}{2}\hat{\theta}^T\hat{\theta}\leq\\&\frac{1}{NJ}\epsilon^T\tilde{Z}\left(\hat{\theta}-\theta^*\right)+\lambda\left(\iota^T\theta^*-1\right)+\frac{\mu}{2}\theta^{*T}\theta^*.\end{aligned}\tag{2.C.13}$$

Note that $\epsilon^T\tilde{Z}(\hat{\theta}-\theta^*)\leq\left\|\tilde{Z}^T\epsilon\right\|_\infty\left\|\hat{\theta}-\theta^*\right\|_1$.

Applying Lemma 2.C.1 with $\gamma\equiv\gamma(N,\delta):=\sqrt{2\log\left(\frac{2(R-1)J}{\delta}\right)\Big/N}$ we obtain

$$\mathbb{P}\left(\left\|\frac{1}{NJ}\tilde{Z}^T\epsilon\right\|_\infty \geq \gamma\right) \leq 2(R-1)J\exp\left(-N\left(\sqrt{\frac{2\log\left(\frac{2(R-1)J}{\delta}\right)}{N}}\right)^2 \Big/ 2\right)$$

$$= 2(R-1)J\exp\left(\log\left(\left(\frac{2(R-1)J}{\delta}\right)^{-1}\right)\right)$$

$$= \delta. \tag{2.C.14}$$

In the following, we assume that $\{(1/(NJ))||\tilde{Z}^T\epsilon||_\infty \leq \gamma\}$, which happens with probability at least $1-\delta$ according to Equation (2.C.14). Therefore, the rest of the proof holds with probability $1-\delta$. Using that the event $\{(1/(NJ))||\tilde{Z}^T\epsilon||_\infty \leq \gamma\}$ occurs, we can bound the the right hand side in Equation (2.C.13) from above by

$$\frac{1}{2NJ}\left\|\tilde{Z}\left(\theta^* - \hat{\theta}\right)\right\|^2 + \lambda\left(\iota^T\hat{\theta} - 1\right) + \frac{\mu}{2}\hat{\theta}^T\hat{\theta} \leq \gamma\left\|\hat{\theta} - \theta^*\right\|_1 + \lambda\left(\iota^T\theta^* - 1\right) + \frac{\mu}{2}\theta^{*T}\theta^*. \tag{2.C.15}$$

We split $\hat{\theta}$, $\tilde{Z}$ and ν over S and S^C into two blocks, whereby S again denotes the set of relevant grid points for which the true weights $\theta^* > 0$ and S^C the set of points for which $\theta^* = 0$. It follows that

$$\iota^T\theta = \iota_S^T\theta_S + \iota_{S^C}^T\theta_{S^C} = ||\theta_S||_1 + ||\theta_{S^C}||_1$$

and

$$\theta^T\theta = \theta_S^T\theta_S + \theta_{S^C}^T\theta_{S^C}.$$

Thus, we can reformulate Equation (2.C.15) as

$$\frac{1}{2NJ}\left\|\tilde{Z}\left(\theta^* - \hat{\theta}\right)\right\|_2^2 + \lambda\left(\left\|\hat{\theta}_S\right\|_1 + \left\|\hat{\theta}_{S^C}\right\|_1 - 1\right) + \frac{\mu}{2}\left(\hat{\theta}_S^T\hat{\theta}_S + \theta_{S^C}^{*T}\theta_{S^C}^*\right) \leq$$

$$\gamma\left\|\hat{\theta} - \theta^*\right\|_1 + \lambda\left(\left\|\theta_S^*\right\|_1 + \left\|\theta_{S^C}^*\right\|_1 - 1\right) + \frac{\mu}{2}\left(\theta_S^{*T}\theta^* + \theta_{S^C}^{*T}\theta_{S^C}^*\right).$$

It follows from $\theta_{S^C}^* = 0$ that $||\hat{\theta} - \theta^*||_1 = ||\hat{\theta}_S - \theta_S^*||_1 + ||\hat{\theta}_{S^C}||_1$ such that after some simple manipulations we obtain

$$\frac{1}{2NJ}\left\|\tilde{Z}\left(\theta^* - \hat{\theta}\right)\right\|_2^2 + \lambda\left(\left\|\hat{\theta}_S\right\|_1 + \left\|\hat{\theta}_{S^C}\right\|_1 - 1\right) + \frac{\mu}{2}\left(\hat{\theta}_S^T\hat{\theta}_S - \theta_S^{*T}\theta_S^* + \hat{\theta}_{S^C}^T\hat{\theta}_{S^C}\right) \leq$$

$$\gamma \left\| \hat{\theta} - \theta^* \right\|_1 + \lambda \left(\left\| \theta_S^* \right\|_1 - 1 \right). \tag{2.C.16}$$

Note that the terms in (2.C.16) that are multiplied by the Langrangian parameter λ drop out. Recall that by the definition of a linear probability model, $||\theta_S^*||_1 - 1 = 0$. With respect to the second term, $\lambda(||\hat{\theta}_S||_1 + ||\hat{\theta}_{S^C}||_1 - 1)$, there are two different cases to be considered due to the inequality constraint $\sum_{r=1}^{R-1} \theta_r \leq 1$: (1) the estimated probability weights sum to one (the constraint is binding), and (2) the sum of the estimated probability weights is less than one (the constraint is not binding). In the former case, $||\hat{\theta}_S||_1 + ||\hat{\theta}_{S^C}||_1 - 1 = 0$. In the latter case, the KKT conditions require $\lambda = 0$. Thus, Condition (2.C.16) simplifies to

$$\frac{1}{2NJ} \left\| \tilde{Z} \left(\theta^* - \hat{\theta} \right) \right\|_2^2 + \frac{\mu}{2} \left(\hat{\theta}_S^T \hat{\theta}_S - \theta_S^{*T} \theta_S^* + \hat{\theta}_{S^C}^T \hat{\theta}_{S^C} \right) \leq \gamma \left\| \hat{\theta} - \theta^* \right\|_1. \tag{2.C.17}$$

It follows from $||\hat{\theta}_S - \theta_S^*||_2^2 = \hat{\theta}_S^T \hat{\theta}_S - 2\theta_S^{*T} \hat{\theta}_S + \theta_S^{*T} \theta_S^*$ that

$$\hat{\theta}_S^T \hat{\theta}_S - \theta_S^{*T} \theta_S^* + \hat{\theta}_{S^C}^T \hat{\theta}_{S^C} = \left\| \hat{\theta}_S - \theta_S^* \right\|_2^2 + 2\theta_S^{*T} \hat{\theta}_S - 2\theta_S^{*T} \theta^* + \left\| \hat{\theta}_{S^C} \right\|_2^2$$

and from $\theta_{S^C}^* = 0$ that $||\hat{\theta}_{S^C}||_p = ||\hat{\theta}_{S^C} - \theta_{S^C}^*||_p$ for $p = 1, 2$.

Consequently, we can collect the terms over the index sets S and S^C to $||\hat{\theta}_S - \theta_S^*||_1 + ||\hat{\theta}_{S^C}||_1 = ||\hat{\theta} - \theta^*||_1$ and $||\hat{\theta}_S - \theta_S^*||_2^2 + ||\hat{\theta}_{S^C}||_2^2 = ||\hat{\theta} - \theta^*||_2^2$.

This yields

$$\hat{\theta}_S^T \hat{\theta}_S - \theta_S^{*T} \theta_S^* + \hat{\theta}_{S^C}^T \hat{\theta}_{S^C} = \left\| \hat{\theta} - \theta^* \right\|_2^2 + 2\theta_S^{*T} \hat{\theta}_S - 2\theta_S^{*T} \theta^*.$$

Therefore, Equation (2.C.17) can be equivalently expressed as

$$\frac{1}{2NJ} \left\| \tilde{Z} (\theta^* - \hat{\theta}) \right\|_2^2 + \frac{\mu}{2} \left\| \hat{\theta} - \theta^* \right\|_2^2 \leq$$

$$\gamma \left\| \hat{\theta} - \theta^* \right\|_1 + \frac{\mu}{2} \left(2\theta_S^{*T} \theta_S^* - 2\theta_S^{*T} \hat{\theta}_S \right). \tag{2.C.18}$$

Next, because $\theta_S^* > 0$ and $||\hat{\theta}_S - \theta_S^*||_1 \leq \sqrt{s} ||\hat{\theta}_S - \theta_S^*||_2$ it holds that

$$\theta_S^{*T} \left(\theta_S^* - \hat{\theta}_S \right) \leq \theta_S^{*T} \left| \hat{\theta}_S - \theta_S^* \right| \leq \left\| \theta_S^* \right\|_\infty \left\| \hat{\theta}_S - \theta_S^* \right\|_1 \leq \sqrt{s} \left\| \theta_S^* \right\|_\infty \left\| \hat{\theta}_S - \theta_S^* \right\|_2 \tag{2.C.19}$$

where $|\hat{\theta}_S - \theta_S^*|$ takes the absolute value of each element of the vector $\hat{\theta}_S - \theta_S^*$.

Substituting Condition (2.C.19) back into the error bound in Equation (2.C.18) and

using the fact that $||\hat{\theta} - \theta^*||_1 \leq \sqrt{(R-1)}\ ||\hat{\theta} - \theta^*||_2$, we can rewrite Equation (2.C.18) as

$$\frac{1}{2NJ}\left\|\tilde{Z}(\theta^* - \hat{\theta})\right\|_2^2 + \frac{\mu}{2}\left\|\hat{\theta} - \theta^*\right\|_2^2 \leq \gamma\sqrt{(R-1)}\left\|\hat{\theta} - \theta^*\right\|_2 + \mu\sqrt{s}\left\|\theta_S^*\right\|_\infty \left\|\hat{\theta}_S - \theta_S^*\right\|_2 \quad \text{(2.C.20)}$$

Recall that

$$\left\|\tilde{Z}(\hat{\theta} - \theta^*)\right\|_2^2 = (\hat{\theta} - \theta^*)^T \tilde{Z}^T \tilde{Z}(\hat{\theta} - \theta^*)$$

and that the left-hand-side in Condition (2.C.20) can be summarized as

$$\frac{1}{2}(\hat{\theta} - \theta^*)^T\left[\frac{1}{NJ}\tilde{Z}^T\tilde{Z} + \mu I\right](\hat{\theta} - \theta^*) \leq \left(\gamma\sqrt{(R-1)} + \mu\sqrt{s}\left\|\theta_S^*\right\|_\infty\right)\left\|\hat{\theta} - \theta^*\right\|_2 \quad \text{(2.C.21)}$$

Recall that $\xi_{\min}(\mu)$ defines the minimum eigenvalue of the real symmetric matrix $1/(NJ)\tilde{Z}^T\tilde{Z} + \mu I$ over the set of vectors $\mathcal{H}$ (see Subsection (2.3.2)).

It holds that $\xi_{\min}(\mu) > 0$ if $\mu > 0$ and that $\xi_{\min} \geq 0$ if $\mu = 0$. In the following, we assume $\xi_{\min}(\mu) > 0$.

Thus, multiplying the left-hand-side in Condition (2.C.21) by $||\hat{\theta} - \theta^*||_2^2/||\hat{\theta} - \theta^*||_2^2$ and using the restricted minimum eigenvalue definition gives the upper ℓ_2-error bound between the estimated and true probability weights:

$$\frac{\xi_{\min}(\mu)}{2}\left\|\hat{\theta} - \theta^*\right\|_2^2 \leq \left(\gamma\sqrt{(R-1)} + \mu\sqrt{s}\,\|\theta_S^*\|_\infty\right)\left\|\hat{\theta} - \theta^*\right\|_2$$

$$\Rightarrow \quad \left\|\hat{\theta} - \theta^*\right\|_2 \leq \frac{2\sqrt{(R-1)}\,\gamma + 2\mu\sqrt{s}\,\|\theta_S^*\|_\infty}{\xi_{\min}(\mu)}.$$

□

Proof of Corollary 2.1.

By assumption, it holds that

$$\left(\sqrt{(R-1)}\,\gamma + \mu\sqrt{s}\,\|\theta_S^*\|_\infty\right)\xi_{\min}(0) \leq \sqrt{(R-1)}\,\gamma\xi_{\min}(0) + \mu\sqrt{(R-1)}\,\gamma$$

$$= \sqrt{(R-1)}\,\gamma(\xi_{\min}(0) + \mu).$$

Using $\xi_{\min}(\mu) = \xi_{\min}(0) + \mu$ gives

$$\left(\sqrt{(R-1)}\,\gamma + \mu\sqrt{s}\,\|\theta_S^*\|_\infty\right)\xi_{\min}(0) \leq \sqrt{(R-1)}\,\gamma\xi_{\min}(\mu)$$

which is equivalent to

$$\frac{2\sqrt{(R-1)}\,\gamma + 2\mu\sqrt{s}\,\|\theta_S^*\|_\infty}{\xi_{\min}(\mu)} \leq \frac{2\sqrt{(R-1)}\,\gamma}{\xi_{\min}(0)}.$$

□

Proof of Theorem 2.3.

It holds that the difference of $\hat{F}(\beta)$ and $F^*(\beta)$ in any point $\beta \in \mathbb{R}^K$ can be bounded by

$$\left|\hat{F}(\beta) - F^*(\beta)\right| = \left|\sum_{r=1}^{R} \hat{\theta}_r\, 1[\beta_r \leq \beta] - \sum_{r=1}^{R} \theta_r^*\, 1[\beta_r \leq \beta]\right|$$

$$\leq \sup_\beta \left|\sum_{r=1}^{R} \left(\hat{\theta}_r - \theta_r^*\right)\, 1[\beta_r \leq \beta]\right|$$

$$\leq \sum_{r=1}^{R} \left|\hat{\theta}_r - \theta_r^*\right| = \sum_{r=1}^{R-1} \left|\hat{\theta}_r - \theta_r^*\right| + \left|\hat{\theta}_R - \theta_R^*\right|$$

where the last inequality holds by the triangle inequality.

Then,

$$\left|\hat{F}(\beta) - F^*(\beta)\right| \leq \sum_{r=1}^{R-1} \left|\hat{\theta}_r - \theta_r^*\right| + \left|1 - \sum_{r=1}^{R-1} \hat{\theta}_r - 1 + \sum_{r=1}^{R-1} \theta_r^*\right|$$

$$= \sum_{r=1}^{R-1} \left|\hat{\theta}_r - \theta_r^*\right| + \left|\sum_{r=1}^{R-1} \left(\theta_r^* - \hat{\theta}_r\right)\right| \leq 2\sum_{r=1}^{R-1} \left|\hat{\theta}_r - \theta_r^*\right|$$

$$= 2\left\|\hat{\theta} - \theta^*\right\|_1 \leq 2\sqrt{(R-1)}\left\|\hat{\theta} - \theta^*\right\|_2,$$

which, by Theorem 2.2, can be bounded by

$$|\hat{F}(\beta) - F^*(\beta)| \leq 2\sqrt{(R-1)}\,\frac{2\sqrt{(R-1)}\,\gamma + 2\mu\sqrt{s}\,\|\theta^*_S\|_\infty}{\xi_{\min}(\mu)}.$$

□

3 Nonparametric Estimation of the Random Coefficients Model: A Random Elastic Net Approach

Abstract

This paper extends the computationally attractive nonparametric random coefficients estimator of Heiss, Hetzenecker and Osterhaus (2021), which includes the nonnegative LASSO estimator of Fox, Kim, Ryan and Bajari (2011) as a special case, to a random elastic net estimator. The random elastic net estimator is a bootstrap method. It repeatedly estimates the estimator of Heiss et al. (2021), varying the potential support of the random coefficients' distribution across repetitions. Subsequently, it averages the results of the repetitions to obtain final estimates. The random elastic net estimator improves the estimator's recovery of the true support and allows for more accurate estimates of the random coefficients' distribution. Two Monte Carlo experiments and an application to the regulation of air pollution (Blundell et al., 2020) illustrate the improved performance of the random elastic net estimator.

JEL codes: *C14, C25, L.*

Keywords: *Random Coefficients, Mixed Logit, Nonparametric Estimation, Random Elastic Net.*

Publication status: *To be submitted.*

3.1 Introduction

A popular approach for modeling unobserved heterogeneity, which is a common challenge in many empirical studies, are random coefficient models. Random coefficient models allow the coefficients of the economic model to vary across agents according to an unknown distribution function, which the researcher aims to estimate.

For this purpose, Fox, Kim, Ryan and Bajari (2011), hereafter FKRB, introduce a simple and computationally fast nonparametric estimator for the random coefficients' distribution. The estimator approximates the distribution function using a fixed grid of random coefficients. Every grid point in the grid is a prespecified vector of random coefficients and represents a certain type of heterogenous agent. The distribution function is obtained from the estimated probability weights at the grid points, which are estimated with constrained least squares. In principle, the true distribution function can be approximated arbitrarily closely if the grid of random coefficients is sufficiently dense (McFadden and Train, 2000).

Monte Carlo studies (e.g., Fox et al., 2011 and Fox et al., 2016) and applications to real data (e.g., Nevo et al., 2016, Illanes and Padi, 2019, Blundell, Gowrisankaran and Langer, 2020 and Houde and Myers, 2019) indicate, however, that the estimator tends to estimate only few positive weights and that it sets the weights at many grid points to zero. For instance, Nevo et al. (2016) study the demand for residential broadband and estimate only 53 out of $8,626$ potentially heterogeneous consumer types. Illanes and Padi (2019) use the approach to estimate the demand for private pension plans in Chile and assign positive weights to only 194 of $83,251$ grid points. Blundell et al. (2020) develop a model to study firms' reactions to dynamic enforcement of air pollution regulations and recover no more than 12 out of $10,001$ grid points. These applications illustrate the sparse nature of the estimator of FKRB due to which it lacks the ability to estimate smooth distribution functions and instead approximates potentially continuous distributions through step functions with only few steps.

Heiss, Hetzenecker and Osterhaus (2021) explain the sparse nature by showing that the estimator of FKRB is a nonnegative LASSO (Wu et al., 2014) (NNL) estimator with a fixed tuning parameter. NNL was first mentioned in the seminal work of Efron et al. (2004) as positive LASSO. It shares the property of LASSO (Tibshirani, 1996), which is a popular model selection method typically used in applications with supposedly sparse models, that it regularizes the coefficients of the model and shrinks some to zero.

Besides its sparse nature, the connection of the FKRB estimator to NNL implies that the estimator potentially selects grid points incorrectly under strong correlation. This is due to the fact that the estimator "randomly" chooses one out of a group of highly correlated grid points and sets the remaining weights to zero (see Zou and Hastie, 2005, and Hastie et al., 2009, for the random behavior of LASSO).

A consequence of the estimator's sparsity and "random" selection behavior can be an inaccurate approximation of the true distribution through a non-smooth distribution. The

estimated support can deviate from the true distribution's support and, hence, conclusions with respect to the heterogeneity of agents in the population be misleading. Fox et al. (2016) show that identification of the true distribution requires a sufficiently dense grid of random coefficients. Yet, in practice, a more dense grid tends to be accompanied with higher correlation among grid points, leading to the incorrect grid point selection, and correlation can become so strong that the optimization problem to the FKRB estimator cannot be solved due to singularity (Nevo et al., 2016, Online Supplement). Thus, the high correlation associated with a dense grid in combination with the incorrect grid point selection of the estimator under strong correlation can cause problems for the identification of the model.

In order to address these shortcomings, Heiss et al. (2021) generalize the FKRB estimator to an elastic net estimator, which includes the FKRB estimator as a special case. They show that their proposed estimator is able to select the correct support of the distribution in cases where the FKRB estimator fails to do so. Furthermore, they prove that the elastic net estimator, under conditions, approximates the random coefficients' distribution more accurately than the FKRB estimator.

Even though the extension mitigates the sparsity and improves the selection of the grid points, the solutions still tend to be too sparse. Therefore, we propose an estimator based on the *random LASSO* estimator devoloped by Wang et al. (2011). The key idea of the random LASSO estimator is similar to that of the random forest (Breiman, 2001). By repeatedly sampling a random subset of regressors and estimating the model with these subsets, the random LASSO estimator aims to substantially decrease the correlation among the regressors within each repetition. The coefficients for the regressors not drawn in a repetition are set to zero. The final coefficients are the averages over the coefficients of all repetitions. Due to the random selection of the regressors in each repetition, the subset of regressors used for the estimation may only include some of the highly correlated regressors. This diminishes the random selection behavior of the LASSO and improves the chance that it selects all regressors correctly. Hence, the union of the selected regressors over all repetitions might provide the correct set of variables, including all of the highly correlated variables. To account for correlations within the subset of regressors in a repetition, we replace LASSO estimation with elastic net estimation and refer to the estimator as random elastic net estimator.

So far, the random LASSO estimator and random elastic estimator net prove most useful in computational biology and medical sciences where researchers often encounter high-dimensional problems with highly correlated variables – e.g., when aiming to identify cancer driver genes (see, e.g., Björn, Badam, Spalinskas, Brandén, Koyi, Lewensohn, De Petris, Lubovac-Pilav, Sahlén, Lundeberg et al., 2020, Kim, Hao, Gautam, Mersha and Kang, 2018, and Park, Imoto and Miyano, 2015 for applications of the random LASSO estimator and Park, Niida, Imoto and Miyano, 2017 and Yu, Cen, Chen, Markowitz, Shaw, Tsai, Conejo-Garcia and Wang, 2022 for applications of the random elastic net estimator).

More broadly, the random LASSO and random elastic net are related to model

averaging techniques (see Moral-Benito (2015) for an overview). Model averaging methods do not select a single model among a set of candidate models like model selection methods, such as the LASSO, but rather combine candidate models (Liu, Okui and Yoshimura, 2016). This strategy can be viewed as a type of insurance against selecting a single poor model (Leung and Barron, 2006). Furthermore, Hansen (2008) points out that model averaging methods balance specification error (bias) against overparameterization (variance) and are useful when we have a well defined goal such as minimizing the mean squared error of the estimation. A concrete example of a model averaging technique, which is similar to random LASSO, is random subset regression (Boot and Nibbering, 2019). Random subset regression randomly draws regressors, which is referred to as "feature bagging", to approximate the original model by combining many low-dimensional models. Boot and Nibbering (2019) apply random subset regression to lower the mean squared forecast error.

In high-dimensional settings with more variables than observations, the random LASSO estimator and random elastic net estimator solve a further limitation of the LASSO estimator in addition to the correlation problem. In these settings, the LASSO can select at most as many variables as there are observations. In contrast, the random LASSO estimator and random elastic net estimator can still select all variables in this case. In our discrete choice framework, this implies that the number of grid points is not limited by the sample size. Importantly, this allows us to specify a dense grid which is required for an accurate approximation of the underlying distribution – even if we increase the number of random coefficients included into the model. Hence, the curse of dimensionality, which is a shortcoming of the FKRB estimator if the model consists of a large number of random coefficients, is mitigated.

To study the finite sample performance of the random elastic net estimator, we conduct two Monte Carlo experiments in which we estimate a random coefficients logit model. The first Monte Carlo experiment uses a discrete distribution to study the estimators' ability to correctly recover the true distribution's support. The results demonstrate that our proposed random elastic net substantially improves the recovery of the true support of the distribution function compared to the FKRB estimator and the elastic net estimator – especially when the number of grid points is large relative to the sample size. The second Monte Carlo experiment considers a mixture of two bivariate normal distributions for the true distribution. The results highlight that the random elastic net estimator can recover smooth and possibly complex distribution functions. The estimator approximates these distribution functions more accurately than the FKRB estimator and the elastic net estimator. The application of the random elastic net estimator to the study of Blundell et al. (2020) confirms the property of the random elastic net estimator that it is able to estimate smooth and possibly complex distribution functions. More concretely, while Blundell et al. (2020) rely on the FKRB estimator and approximate a 5-dimensional distribution with only 12 out of $10,0001$ grid points, the random elastic net estimator recovers 154 grid points, which makes it easier to draw conclusions with respect to the shape of the underlying distribution.

The nonparametric estimators for the random coefficients model of Train (2008), Train (2016), and Bansal, Daziano and Achtnicht (2018) constitute alternatives to the FKRB estimator and the elastic net estimator of Heiss et al. (2021). Train (2008) proposes three estimators that use a log-likelihood criterion instead of constrained least squares but are, other than that, similar to the general approach of FKRB. To substantially reduce the number of required grid points, Train (2016) suggests approximating the random coefficients' distribution with (possibly overlapping) polynomials, splines or step functions instead of with a fixed grid of random coefficients' vectors. The approach lowers the number of required grid points and, thereby, constitutes another way to reduce the curse of dimensionality, arising for models with a large number of random coefficients. Bansal et al. (2018) extend the approach of Train (2016) to allow for fixed coefficients in addition to random coefficients. However, for the estimation of the respective model, Train (2008) relies on the EM algorithm, which is sensitive to its starting values and is not guaranteed to converge to a global optimum. While both Train (2016) and Bansal et al. (2018) uses simulated maximum likelihood for the estimation, the computation time for the latter approach increase up to 40 times compared to the former approach. Heiss et al. (2021) also note that their estimation procedure tends to be slow when including both random and fixed coefficients. To circumvent the computational burden, they follow the recommendation of Fox et al. (2016) and Houde and Myers (2019) who suggest a two-step estimator to estimate the model with fixed and random coefficients.

The remainder of the paper is organized as follows. Section 3.2 describes the FKRB estimator and the elastic net estimator and introduces the random elastic net estimator. In Section 3.3, we present the two Monte Carlo experiments that study the performance of the random elastic net estimator in comparison to the FKRB estimator and the elastic net estimator. Section 3.4 applies the proposed estimator to the study of Blundell et al. (2020). Section 3.5 concludes.

3.2 Fixed Grid Estimators

For the introduction of our estimator, we consider the framework of a random coefficient discrete choice model.[1] The approach, however, is not restricted to discrete choice models, but can be applied to any model with unobserved heterogeneous parameters. Let there be an i.i.d. sample of N observations, each confronted with a set of J mutually exclusive potential outcomes. The researcher observes a K-dimensional real-valued vector of explanatory variables $x_{i,j}$ for every observation unit i and potential outcome j, and a binary vector y_i whose entries are equal to one whenever she observes outcome j for the ith observation, and zero otherwise. The goal is to estimate the unknown distribution of heterogeneous parameters $F_0(\beta)$ in the model

[1]Note that the framework of this chapter is the same as in Chapter 1 and included to ensure that this Chapter can be read independently of Chapter 1. Readers, who want to avoid repetitions to Chapter 1, can skip these parts and continue with Subsection 3.2.2.

$$P\left(x_{i,j}\right) = \int g\left(x_{i,j}, \beta\right) dF_0\left(\beta\right) \tag{3.1}$$

where $g\left(x_{i,j}, \beta\right)$ denotes the probability of outcome j conditional on the random coefficients β and covariates $x_{i,j}$. The researcher specifies the functional form of $g\left(x_{i,j}, \beta\right)$. A prominent example of Equation (3.1) is the multinomial mixed logit model, the state-of-the-art model for demand estimation. For a detailed description of the multinomial mixed logit see Train (2009, pp. 134-150). In this model, consumer i realizes utility $u_{i,j} = x_{i,j}^T \beta_i + \omega_{i,j}$ from alternative j, given product characteristics $x_{i,j}$ and unobserved consumer-specific preferences β_i. $\omega_{i,j}$ denotes an additive, consumer- and choice-specific error term. Consumer i chooses alternative j of J alternatives (and an outside good with utility $u_{i,0} = \omega_{i,0}$) if $u_{i,j} > u_{i,l}$ for all $l \neq j$. Under the assumption that $\omega_{i,j}$ follows a type I extreme value distribution, the unconditional choice probabilities, $P_{i,j}(x)$, are of the form

$$P_{i,j}(x) = \int \frac{\exp\left(x_{i,j}^T \beta\right)}{1 + \sum_{l=1}^{J} \exp\left(x_{i,l}^T \beta\right)} dF_0\left(\beta\right).$$

$F_0(\beta)$ represents the distribution of heterogeneous consumer preferences in the population and is to be estimated. In most applications, researchers place restrictive assumptions on the functional form of $F_0(\beta)$ in advance, and estimate its parameters from the data.

3.2.1 Fixed Grid Elastic Net Estimator of Heiss et al. (2021)

FKRB propose a simple and fast mixture approach to estimate the underlying random coefficients' distribution without restrictive assumptions on its shape. The estimator is a special case of sieve estimators (Chen, 2007). It uses a finite and fixed grid of random coefficient vectors as mixture components to construct the distribution from the estimated probability weight of every component. The underlying idea of this fixed grid estimator is the transformation of the unconditional choice probabilities in Equation (3.1) into a probability model in which $F_0(\beta)$ enters linearly. FKRB derive the linear probability model in two steps: they transform Equation (3.1) into a regression model with the random coefficients' distribution as the only unknown term. Adding $y_{i,j}$ to both sides and moving $P_{i,j}$ to the right results in the probability model

$$y_{i,j} = \int g\left(x_{i,j}, \beta\right) dF_0\left(\beta\right) + \left(y_{i,j} - P_{i,j}\left(x\right)\right). \tag{3.2}$$

To exploit linearity in parameters, they use a sieve space approximation to the infinite-dimensional parameter $F_0(\beta)$. The sieve space approximation divides the support of the random coefficients β into R fixed vectors, resulting in a fixed grid of random coefficients $\mathcal{B}_R = (\beta_1, \ldots, \beta_R)$. Each vector β_r, $r = 1, \ldots, R$, has length K, the number of random coefficients included in the model. The location and values of these vectors are

specified by the researcher. With the sieve space approximation, Equation (3.2) becomes a simple linear probability model with unknown parameters $\theta = (\theta_1, \ldots, \theta_R)^T$

$$y_{i,j} \approx \sum_{r=1}^{R} g\left(x_{i,j}, \beta_r\right) \theta_r + \left(y_{i,j} - P_{i,j}\left(x\right)\right) \tag{3.3}$$

where $g(x_{i,j}, \beta_r)$ denotes the conditional choice probability evaluated at grid point r. Given the fixed grid of random coefficients $\mathcal{B}_R$, the researcher estimates the probability weight θ_r at every point $r = 1, \ldots, R$. The linear relationship between the outcome variable and the unknown parameters θ allows to estimate the mixture weights with the least squares estimator. The linear regression, which regresses the binary dependent variable $y_{i,j}$ on the choice probabilities evaluated at $\mathcal{B}_R$, in total has NJ observations, J "regression observations" for every statistical observation unit $i = 1, \ldots, N$ and R covariates $z_{i,j} = \left(z_{i,j}^1, \ldots, z_{i,j}^R\right)^T = (g(x_{i,j}, \beta_1), \ldots, g(x_{i,j}, \beta_R))^T$. By the definition of choice probabilities, the expected value of the composite error term $y_{i,j} - P_{i,j}(x)$ conditional on $x_{i,j}$ is zero. Thus, the regression model satisfies the mean-independence assumption of the least squares approach (Fox et al., 2011).

The estimator of the random coefficients' joint distribution is constructed from the estimated weights

$$\hat{F}\left(\beta\right) = \sum_{r=1}^{R} \hat{\theta}_r \; 1\left[\beta_r \leq \beta\right] \tag{3.4}$$

where β is an evaluation point chosen by the researcher and the indicator function $1[\beta_r \leq \beta]$ is equal to one whenever $\beta_r \leq \beta$, and zero otherwise.

To ensure that $\hat{F}(\beta)$ is a valid distribution function, FKRB suggest to estimate the weights with the least squares estimator subject to the constraints that the weights are greater than or equal to zero, and sum to one. Heiss et al. (2021) show that the corresponding estimator is a special case of nonnegative LASSO, explaining its sparse solutions. To mitigate the sparsity, Heiss et al. (2021) add a constraint on the sum of the squared weights. Their extension results in the following estimator

$$\begin{aligned} \hat{\theta}^{\text{ENet}} &= \arg\min_{\theta} \frac{1}{NJ} \sum_{i=1}^{N} \sum_{j=1}^{J} \left(y_{i,j} - \sum_{r=1}^{R} \theta_r z_{i,j}^r\right)^2 \\ \text{s.t.} \quad \theta_r \geq 0 \quad \forall r \quad &\text{and} \quad \sum_{r=1}^{R} \theta_r = 1 \quad \text{and} \quad \sum_{r=1}^{R} \theta_r^2 \leq t \end{aligned} \tag{3.5}$$

where $t \geq 1/R$ is a nonnegative tuning parameter specified by the researcher.[2]

[2]Note that, for technical reasons only, Heiss et al. (2021) penalize $R - 1$ weights in the quadratic constraint whereas here all R weights are penalized. In applications, the difference between both

Heiss et al. (2021) show that the estimator in Equation (3.5) is a special case of an elastic net estimator. To be precise, the tuning parameter t is the ridge penalty and determines the impact of the quadratic constraint. The LASSO tuning parameter is fixed since the weights are constrained to be nonnegative and to sum up to one.

For any $t \geq 1$, the quadratic constraint on the probability weights has no effect and, hence, the elastic net estimator reduces to the FKRB estimator. Thus, the elastic net estimator is a generalization of the FKRB estimator and includes the FKRB estimator as a special case.

The smallest possible value for t is $1/R$ since, for $t < 1/R$, the constraint that the weights sum up to one can not be fullfilled anymore. Given $1/R \leq t < 1$, the quadratic constraint shrinks the probability weights of highly correlated grid points towards each other in order to meet the quadratic constraint. That is, the quadratic constraint encourages a grouping effect which allows to recover highly correlated points inside the true support of $F(\beta)$ together. This reduces the estimator's sparsity. The linear constraint, in turn, retains the LASSO property, which makes it possible to select weights inside the support of the true distribution and to estimate zero weights for points outside the true support.

In particular, the specification of the tuning parameter t allows adjusting the estimator to the level of correlation among grid points and thus, controls the sparsity of the results. Smaller values of t put more weight on the the quadratic constraint and yield less sparse results. Heiss et al. (2021) recommend choosing the tuning parameter with cross-validation and the one standard error rule based on the mean squared error (MSE) criterion.

The quadratic constraint has the desirable property that it allows for the specification of a substantially finer grid of random coefficients which is necessary for accurate estimation of $F_0(\beta)$. While the FKRB estimator runs into almost perfect collinearity problems if the grid becomes finer (Fox et al., 2016), the quadratic constraint ensures that the optimization problem for the elastic net estimator always has a solution.

Key to an accurate approximation of $F_0(\beta)$ is the precise estimation of the probability weight at every grid point. Basis to a precise estimation of the probability weights is the consistent selection of the relevant grid points. This requires the elastic net estimator to estimate positive weights at all grid points at which $F_0(\beta)$ has a positive probability mass, and zero weights otherwise. While zero weights at grid points inside $F_0(\beta)$'s support cause inaccurate approximations through step functions with only few steps, positive estimates at grid points outside $F_0(\beta)$'s support lead to unreliable estimates of the random coefficients' distribution.

Heiss et al. (2021) illustrate in two Monte Carlo studies that the elastic net estimator substantially improves the approximation accuracy and the selection consistency compared to the FKRB estimator, i.e., the share of estimated weights with correct sign increases.

---

approaches turned out to have no substantial impact on the results of the estimator.

However, the selection consistency of the elastic net estimator seems to depend on the density of the grid used for estimation. If the grid becomes more dense, the results of Heiss et al. (2021) indicate that the share of correctly selected grid points decreases. In particular, the results of the elastic net estimator are still too sparse. This is due to the fact that the correlation among grid points becomes stronger when the grid becomes more dense and it becomes more challenging to correctly select the true weights.

3.2.2 Random Elastic Net Estimator

The sparsity of the elastic net estimator in a dense grid motivates us to propose an estimator based on the *random LASSO* procedure developed by Wang et al. (2011). The new estimator is called random elastic net estimator and incorporates the fixed grid random coefficient framework. The random elastic net estimator improves the selection consistency of the elastic net approach and is also effective for accurate estimation of the random coefficients' distribution if the number of grid points is large relative to the number of observations. In principle, the proposed method can even be used when the number of grid points is greater than the number of observations.

The key idea of the random LASSO estimator of Wang et al. (2011) is similar to that of the random forest (Breiman, 2001). The random forest draws B bootstrap samples of the data and fits a tree in every sample using only a randomly selected subset of all variables. The predictions of the random forest are obtained by averaging the predictions of all B trees. Using only a randomly selected subset of regressors instead of all regressors to fit each tree lowers the correlation between the trees. Therefore, the variance of the final prediction decreases (Hastie et al., 2009).

Random LASSO, like the random forest, draws B bootstrap samples from the data, consisting of the dependent variable y and the regressor matrix Z. Furthermore, each bootstrap sample includes only a randomly selected subset of regressors from Z. In each bootstrap sample, the drawn regressors are then used to fit a LASSO regression instead of a tree. The coefficients of the regressors which are not drawn in a bootstrap sample are set to zero. Finally, the estimated coefficients are obtained by averaging the coefficients over all bootstrap samples.

Intuitively, this strategy diminishes the random selection behavior of the LASSO when variables are highly correlated. That is, Wang et al. (2011) stress that "baseline" LASSO tends to randomly select only a few of a set of highly correlated variables. This implies that if one repeatedly draws data from the same distribution and estimates LASSO each time, one would expect LASSO to select different subsets of the highly correlated variables. Yet, the union of the selected regressors over all repetitions might provide the correct set of variables, including all of the highly correlated variables.

However, dealing with only one given data set, this strategy is not feasible and splitting the data set might not be desirable due to loss in efficiency, depending on the sample size.

Instead of using a subset of observations and repeatedly estimating the model, random LASSO instead aims to substantially decrease the correlations among the regressors before estimation of the model. This is achieved by randomly drawing a subset of regressors in each bootstrap sample. Consequently, each bootstrap sample may be expected to only include some of the highly correlated regressors and thus, the chance that those variables are selected by LASSO increases. Repeating this process sufficiently often, we expect that the union of the selected variables of all bootstrap samples includes all of the relevant highly correlated variables.

The random LASSO consists of two steps in total, which are outlined in the following. The previously described approach corresponds to the first step. The second step of random LASSO differs from the first step only in the probability with which the regressors are drawn in each bootstrap sample. In the first step, each regressor is drawn randomly, i.e., it is equally likely for each regressor to be included in the regression in a given bootstrap sample. In the second step, the probability of drawing a regressor is adjusted to be proportional to its coefficient obtained in the first step. Thereby, we focus on the important regressors in the second step and can estimate them more precisely.

For the random elastic net estimator, we replace the LASSO regression in each bootstrap sample by an elastic net regression. Hence, the random elastic net estimator includes the random LASSO estimator as a special case.

Algorithm 1 summarizes our random elastic net estimator. The estimator uses the same number of grid points R, the vector of observed choices y and the matrix of choice probabilities Z as the FKRB and the elastic net estimator. Further inputs to the algorithm are the number of variables q and h drawn in step 1 and step 2. We refer to step 1 as the bootstrap (BS) step and to step 2 as the bootstrap update (BSU) step.

In each bootstrap sample b in the BS step, we first set the weight at each grid point to zero and then randomly draw q out of R grid points. We denote the corresponding index set of grid points in bootstrap sample b by Q_b. We allow the ridge tuning parameter μ_b for the elastic net estimation to vary for each b. This allows us to adjust μ_b to the correlation among the columns of Z_{Q_b} where Z_{Q_b} includes only those columns which are in the index set Q_b. Choosing $\mu_b = 0$ for each $b = 1, \ldots, B$ yields the random LASSO estimator. However, we recommend to choose μ_b based on the sequence of tuning parameters suggested by, e.g., the *glmnet* package (Friedman, Hastie and Tibshirani, 2010) - for ridge regression with nonnegative coefficients given y and Z_{Q_b}.

Using y, Z_{Q_b} and μ_b, we estimate the weights $\hat{\theta}_{Q_b}^{(b)}$ with an elastic net regression whereby the $\hat{\theta}_{Q_b}^{(b)}$ correspond to the estimated weights of the selected grid points in bootstrap sample b. Finally, we obtain the estimated bootstrap weights $\hat{\theta}^{BS}$ by averaging the weights over all bootstrap samples.

In the subsequent BSU step, the number of variables drawn is h, which may differ from q in the BS step. In contrast to the BS step, the variables are chosen with probability corresponding to $\hat{\theta}^{BS}$ in this step. The remaining setup of the BSU step is the same as

Algorithm 1 Random Elastic Net

1: Choose the number of grid points R and compute y and Z.
2: Choose the number of grid points q and h.
3: Draw B bootstrap samples of size N without replacement from y and Z.

Step 1 – Bootstrap (BS)

4: **for each** bootstrap sample $b \in \{1, 2, \ldots, B\}$ **do**
5: Set $\hat{\theta}_r^{(b)} = 0, \quad r = 1, \ldots, R.$
6: Randomly select q grid points with probability proportional to $1/R$. Let Q_b denote the index set of the selected variables.
7: Choose μ_b for Elastic Net estimation.
8: By applying the Elastic Net estimator with y, μ_b and the selected grid points Z_{Q_b}, obtain

$$\hat{\theta}_{Q_b}^{(b)} = \arg\min_{\theta} \frac{1}{NJ} \sum_{i=1}^{N} \sum_{j=1}^{J} \left(y_{i,j} - \sum_{r \in Q_b} \theta_r z_{i,j}^r \right)^2 + \mu_b \sum_{r \in Q_b} \theta_r^2$$
$$\text{s.t.} \quad \theta_r \geq 0 \quad \forall r \in Q_b \quad \text{and} \quad \sum_{r \in Q_b} \theta_r = 1. \tag{3.6}$$

9: **end for**
10: Compute the BS estimator $\hat{\theta}_r^{BS}$, $r = 1, \ldots, R$, by averaging over all bootstrap runs

$$\hat{\theta}_r^{BS} = \frac{1}{B} \sum_{b=1}^{B} \hat{\theta}_r^{(b)}$$

Step 2 – Bootstrap Update (BSU)

11: **for each** bootstrap sample $u \in \{1, 2, \ldots, B\}$ **do**
12: Set $\hat{\theta}_r^{(u)} = 0, \quad r = 1, \ldots, R.$
13: Randomly select h grid points with selection probability of each grid point Z_j proportional to its weight $\hat{\theta}_j^{BS}, j = 1, \ldots, R$, obtained in Step 1. Let H_u denote the index set of the selected variables.
14: Choose μ_u for Elastic Net estimation.
15: By applying the Elastic Net estimator with y, μ_u and the selected grid points Z_{H_u}, obtain

$$\hat{\theta}_{H_u}^{(u)} = \arg\min_{\theta} \frac{1}{NJ} \sum_{i=1}^{N} \sum_{j=1}^{J} \left(y_{i,j} - \sum_{r \in H_u} \theta_r z_{i,j}^r \right)^2 + \mu_u \sum_{r \in H_u} \theta_r^2$$
$$\text{s.t.} \quad \theta_r \geq 0 \quad \forall r \in H_u \quad \text{and} \quad \sum_{r \in H_u} \theta_r = 1. \tag{3.7}$$

16: **end for**
17: Compute the BSU estimator $\hat{\theta}_r^{BSU}$, $r = 1, \ldots, R$, by averaging over all bootstrap runs

$$\hat{\theta}_r^{BSU} = \frac{1}{B} \sum_{u=1}^{B} \hat{\theta}_r^{(u)}.$$

Note: This algorithm describes the random elastic net estimator. Step 1 corresponds to the BS estimator and Step 2 to the BSU estimator. Inserting the weights $\hat{\theta}^{BS}$ and $\hat{\theta}^{BSU}$ into Equation (3.4), we get $\hat{F}^{BS}(\beta)$ and $\hat{F}^{BSU}(\beta)$, respectively.

in the BS step.

The probabilities used to draw each grid point in the BS step and the BSU step, respectively, can be viewed as a Bayesian prior on the distribution of the weights. In the BS step, one does not have any prior information on the importance of each grid point and therefore, assumes that each grid point is equally important. That is, each grid point is drawn with equal probability. The BSU step incorporates information about the weights at every grid point obtained in the BS step and draws each grid point with probability proportionally to its importance estimated in the BS step.

Note that grid points not drawn in a bootstrap sample are set to zero. Therefore, drawing grid points with high weights with higher probability implies that these grid points are estimated in more bootstrap samples. In turn, their final estimated weight in the BSU step, which is calculated as the average of all bootstrap samples, can increase. Furthermore, it seems reasonable to focus on the important grid points in the second step and to drop the unimportant grid points, which are estimated to be zero in the BS step, in the estimation of the BSU step. However, by including grid points with higher weights more often, it becomes more likely that the selected grid points are more correlated on average. Therefore, there is a trade-off between including relevant grid points more often and breaking the correlation among grid points in each bootstrap sample.

Remark 3.1. As an alternative specification of the random elastic net estimator, we could place less weight on the estimated BS weights, $\hat{\theta}^{BS}$, to draw each grid point in the BSU step. For instance, one could use a prior defined as the mixture of the probability $\hat{\theta}_j^{BS}$, $j = 1, \ldots, R$, and the uniform probability $1/R$ to draw each grid point in the second step, i.e., use $\hat{\theta}_j^* := m\,\hat{\theta}_j^{BS} + (1-m)\,1/R,\ j = 1, \ldots, R$, where $m \in (0, 1]$ controls the weight put on the uninformative uniform prior relative to the prior using the bootstrap weights of step 1.[3] However, we observe in our Monte Carlo studies that for $m = 0.5$ the results are worse compared to those where we use $\hat{\theta}^{BS}$, which corresponds to $m = 1$.

Remark 3.2. Another choice of prior in the BSU step is to select only among grid points which have a positive weight in the BS step and to draw each of those grid points proportional to $1/S$ where S denotes the number of positive weights in $\hat{\theta}^{BS}$. Applying this prior would mean loosing information on the importance of each weight in the BS step which does not seem to be useful according to our simulations, which we do not report for brevity.

Remark 3.3. A further refinement of the random elastic net estimator, which we consider as a robustness check in our subsequent Monte Carlo studies, is to relax the constraint in Equation (3.6) and in Equation (3.7) that the weights in each estimation of each bootstrap sample have to sum up to 1, i.e., $\sum_{r \in Q_b} \theta_r = 1$ and $\sum_{r \in H_u} \theta_r = 1$. That is, we require that the weights only have to be smaller or equal to one, i.e., $\sum_{r \in Q_b} \theta_r \leq 1$

[3] We exclude $m = 0$ since this would correspond to the BS step, i.e., we would get the same results in the BSU step as in the BS step.

and $\sum_{r \in H_u} \theta_r \leq 1$. The refinement allows to estimate zero weights at every grid point if only unimportant variables happen to be drawn in a given bootstrap sample. This is not possible in the random elastic net estimator presented in Algorithm 1. To ensure that the final weights $\hat{\theta}_j^{BS}, j = 1, \ldots, R$, and $\hat{\theta}_j^{BSU}, j = 1, \ldots, R$, sum up to one, we calculate them as a weighted average given by $\hat{\theta}_j^{BS} = \sum_{b=1}^{B} \hat{\theta}_j^{(b)} / \left(\sum_{j=1}^{R} \sum_{b=1}^{B} \hat{\theta}_j^{(b)} \right)$ and $\hat{\theta}_j^{BSU} = \sum_{u=1}^{B} \hat{\theta}_j^{(u)} / \left(\sum_{j=1}^{R} \sum_{u=1}^{B} \hat{\theta}_j^{(u)} \right)$. In our Monte Carlo studies, this modification and the baseline algorithm of the random elastic net estimator yield very similar results.

The choice of q and h plays an important role for the performance of the random elastic net estimator. Wang et al. (2011) recommend to consider a sequence of values for q and h and to find the optimal q and h with MSE-based cross-validation. Note that the sequences used for q and h do not have to be the same. Instead of performing the cross-validation across each combination of q and h, we first run a cross-validation to find the optimal q in the BS step and, subsequently, another cross-validation to find the optimal h in the BSU step. The second cross-validation uses the weights $\hat{\theta}^{BS}$ obtained by the optimal q found in the first cross-validation. Running the cross-validation for q and h in successive order substantially reduces the computational cost of the tuning process.

Wang et al. (2011) observe that the choice of the number of bootstrap samples B does not substantially influence the results of their algorithm when B is large, e.g., $B = 500$ or $B = 1,000$. Intuitively, choosing B sufficiently large ensures that each regressor is drawn in sufficiently many bootstrap samples and therefore, that their coefficients, which are an average of their estimated coefficients in all bootstrap samples b, are reliable.

A disadvantage of the random elastic net estimator is that its estimation can be time consuming - in particular, when the sample size is large - since one has to estimate B bootstrap models. However, if the number of grid points R is large, it might be faster to estimate several models which only include very few grid points repeatedly than estimating one big model.

3.3 Monte Carlo Simulation

We conduct two Monte Carlo experiments to examine the selection consistency and the approximation accuracy of the random elastic net estimator. The first Monte Carlo simulation uses a discrete distribution with a subset of grid points as support points. This allows us to study the estimators' ability to estimate discrete distributions and in particular, the selection consistency of the estimator.

The second experiment generates the random coefficients from a mixture of two normal distributions. Thereby, we aim to study the estimators' ability to estimate smooth distributions.

For both Monte Carlo simulations, we use a random coefficients logit model as the true data generating process to generate individual-level discrete choice data. Each

observational unit i chooses among $J = 4$ mutually exclusive alternatives and an outside option. For every alternative j and observation unit i, we draw the two-dimensional covariate vector $x_{i,j} = (x_{i,j,1}, x_{i,j,2})$ from $\mathcal{U}(0, 5)$ and $\mathcal{U}(-3, 1)$, respectively. To study the effect of the fixed grid and the number of observation units on the estimators' performance, we run every experiment for different sample sizes, and numbers of grid points. We repeat the experiment for every combination of R and N 100 times to compare the performance of our random elastic net estimator with the elastic net and the FKRB estimator in terms of selection consistency and accuracy for every setup. All calculations are conducted with the statistical software R (R Core Team, 2018).

3.3.1 Discrete Distribution

To study the estimators' selection consistency, we generate the random coefficients β from a discrete probability mass function. The estimator successfully recovers the true support from the data if it estimates a positive weight at every support point of $F_0(\beta)$, and zero weights at all points outside its support.

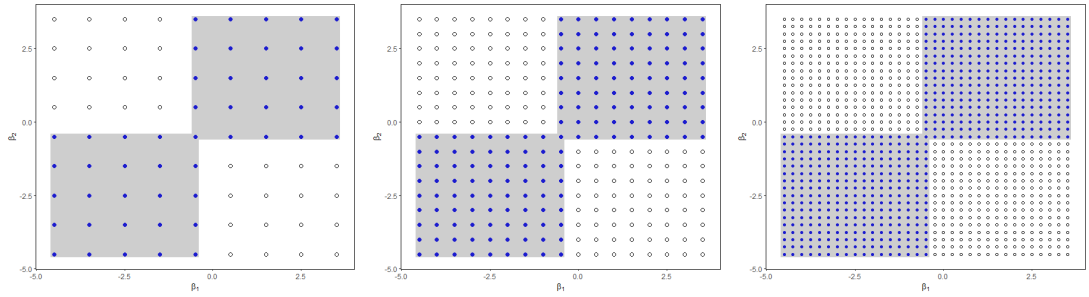
For the support points of $F_0(\beta)$, we select a subset of the grid points from the fixed grid we use for the estimation. The grid covers the range $[-4.5, 3.5] \times [-4.5, 3.5]$ with $R = \{81, 289, 1,089\}$ uniformly allocated grid points. We specify the support of our discrete data generating distribution on $[-4.5, -0.5] \times [-4.5, 0.5]$, and $[-0.5, 3.5] \times [-0.5, 3.5]$, whereby the number of support points varies due to the varying number of grid points. That is, we draw the random coefficients β from a discrete mass function with $S = \{49, 161, 577\}$ support points, each drawn with uniform probability weight $\theta_s = 1/S$. By considering these specifications, we can infer the effect of different number of grid and support points on the estimation of the distribution function. In particular, it allows to inspect the behavior of the estimators when the number of grid points increases. When the number of grid points is large, the grid becomes dense and the correlation among grid points becomes strong. Therefore, we analyze how this increase in correlation affects the considered estimators.

In this discrete setup, the data generating process exactly matches the underlying probability model of the fixed grid estimator since Equation (3.2) and Equation (3.3) coincide in this case. This way, we abstract from any approximation errors that can arise from the sieve space approximation of the true underlying distribution. Therefore, the experiment studies the estimators' selection consistency in the most simple framework possible.

The two areas of the discrete distribution with positive probability mass simulate two heterogeneous groups of preferences in the population. We estimate every distribution for sample sizes $N = \{1,000, 10,000\}$.

Figure 3.1 illustrates the setup of the Monte Carlo experiment for the three data generating distributions corresponding to the three values of R. The gray shaded area indicates the support of the discrete mass functions, and the filled blue points inside this

Figure 3.1: Grid of Monte Carlo Study with Discrete Mass Points



(a) $R = 81$, $S = 49$ (b) $R = 289$, $S = 161$ (c) $R = 1,089$, $S = 577$

area the active grid points. The hollow black points outside the gray shaded areas are the inactive grid points that are not used for data generation.

We choose the optimal tuning parameter μ for the elastic net estimator to be the maximum value of the sequence of 100 tuning parameteres suggested by the *glmnet* package (Friedman et al., 2010) for ridge regression with nonnegative coefficients. Heiss et al. (2021) note that this heuristic approach gives very similar results compared to the selection of μ_b by cross-validation and the one standard error rule based on the MSE as criterion. Thereby, we avoid performing a potentially time-consuming cross-validation when R is large.

The random elastic net estimator is estimated using $B = 1,000$ bootstrap samples. We consider the same sequence of variables $C := \{5+3k | \, k = 0, \ldots, 15\}$ for q and h in the cross-validation, i.e., $q \in C$ and $h \in C$. As described in Subsection 3.2.2, we sequentially apply 10-fold cross validation based on the MSE criterion to find the optimal number of grid points q and h drawn in the BS step and BSU step of the random elastic net estimator.

The sequential execution of the cross-validation reduces the evaluations from $15^2 = 225$ to $15 + 15 = 30$ combinations of tuning parameters q and h. Finally, both the BS estimator and the BSU estimator are calculated with the optimal q and h, respectively.

In the BS step, we choose the ridge tuning parameter μ_b in each bootstrap sample b such that it can adapt to the correlation among the chosen grid points Z_{Q_b} in each bootstrap sample b. Concretely, we specify μ_b as the maximum value of the sequence of tuning parameters suggested by the *glmnet* package (Friedman et al., 2010) for a given y and Z_{Q_b}. The ridge tuning parameter μ_u in the BSU step of the random elastic net estimator is chosen analogously.

To evaluate the estimators' selection consistency, we calculate the average share of sign consistent estimates. An estimate is sign consistent if it is positive at active grid points, and zero otherwise. A weight is defined as positive if it is greater than 10^{-3}.[4]

[4]Due to the optimization method which we use, we do not get estimates which are exactly zero like

To illustrate the sparsity of the estimators' solutions, we report the average number of positive weights and the average share of true positive weights.

Beyond selection consistency, the discrete setup of the Monte Carlo experiment allows us to study the bias of the estimated probability weights. Denote the estimated weight at grid point β_r in Monte Carlo run m by $\hat{\theta}_{r,m}$. We calculate the L_1 norm

$$L_1 = \frac{1}{M} \sum_{m=1}^{M} \frac{1}{R} \sum_{r=1}^{R} \left| \theta_r - \hat{\theta}_{r,m} \right|$$

to measure the average absolute bias of $\hat{\theta}$ in comparison to the true weights θ over all Monte Carlo runs M. In addition, we adopt the root mean integrated squared error (RMISE) from Fox et al. (2011) to provide a metric on the approximation accuracy of the estimated distribution. The RMISE averages the squared difference between the true and estimated distribution at a fixed set of grid points across all Monte Carlo runs

$$\text{RMISE} = \sqrt{\frac{1}{M} \sum_{m=1}^{M} \left[\frac{1}{E} \sum_{e=1}^{E} \left(\widehat{F}_m(\beta_e) - F_0(\beta_e) \right)^2 \right]},$$

where $\widehat{F}_m(\beta_e)$ denotes the estimated distribution function in Monte Carlo run m evaluated at grid point β_e. For the evaluation, we use $E = 10,000$ points uniformly distributed over the range $[-4.5, 3.5] \times [-4.5, 3.5]$.

Table 3.1 summarizes the results of the Monte Carlo experiment. The first three columns report the sample size N, the number of grid points R, and the number of true support points S. The upper part of the table presents the measures on the accuracy of the estimated weights along with the average number of positively estimated weights. The lower part of the table reports the average shares of true positive, and sign consistent estimated weights. The final three columns in the lower part report the average number of grid points q and h selected via the cross-validation in the BS step and BSU step of the random elastic net estimator and the third quantile of the absolute values of the correlation ρ among all grid points.[5]

The results show that the random elastic net estimator outperforms the FKRB and the elastic net estimator for almost every combination of N and R.[6] All measures indicate

with other LASSO-type optimizers. We use 10^{-3} as a threshold to define a weight to be positive as this is roughly $1/1,089$, i.e., one over the maximum number of grid points R included in the Monte Carlo simulation. Using 10^{-4} as a threshold, we obtain qualitatively similar results.

[5] In addition, we also considered the mean and median to summarize the absolute correlation among grid points. We focus on the third quantile since it best illustrates the strong correlation in this setup.

[6] Only for $R = 81$ and $N = 10,000$, the RMISE of the BSU estimator is 0.028 which is slightly larger than the RMISE of 0.027 of the elastic net estimator. However, even in this case the corresponding

Table 3.1: Summary Statistics of 100 Monte Carlo Runs with Discrete Distribution.

			RMISE				$100 \times L_1$				Pos.			
N	R	S	FKRB	ENet	BS	BSU	FKRB	ENet	BS	BSU	FKRB	ENet	BS	BSU
1,000	81	49	0.078	0.034	0.026	0.028	1.857	0.700	0.499	0.563	15.63	55.21	70.00	67.53
1,000	289	161	0.087	0.043	0.025	0.027	0.636	0.308	0.149	0.172	15.85	122.91	202.93	187.59
1,000	1,089	577	0.088	0.052	0.024	0.028	0.179	0.115	0.041	0.052	16.32	240.11	470.12	413.27
10,000	81	49	0.050	0.027	0.019	0.028	1.546	0.739	0.441	0.575	23.41	47.64	64.48	59.49
10,000	289	161	0.059	0.035	0.019	0.029	0.602	0.332	0.133	0.176	24.70	106.45	186.10	171.50
10,000	1,089	577	0.062	0.040	0.017	0.027	0.177	0.122	0.038	0.051	24.73	205.29	466.02	400.52

			% True Pos.				% Sign				q	h	ρ
N	R	S	FKRB	ENet	BS	BSU	FKRB	ENet	BS	BSU	BS	BSU	3rd Qu.
1,000	81	49	27.00	90.61	99.35	98.96	52.88	80.98	73.28	75.86	32.66	7.61	0.819
1,000	289	161	8.32	65.21	97.43	95.23	48.07	74.42	82.63	85.48	37.55	8.12	0.822
1,000	1,089	577	2.34	36.74	78.10	67.85	48.00	63.90	86.60	80.96	37.28	7.58	0.824
10,000	81	49	42.51	87.31	99.80	99.61	62.04	86.32	80.64	86.58	32.99	16.01	0.819
10,000	289	161	13.42	60.14	98.98	96.42	50.69	74.46	90.18	92.38	37.28	17.75	0.822
10,000	1,089	577	3.73	33.01	78.88	67.68	48.70	63.15	87.81	81.95	40.01	18.62	0.824

Note: The table reports the average summary statistics over all Monte Carlo replicates for the FKRB estimator (FKRB), for the elastic net estimator with tuning parameter μ from the maximum value of the sequence of tuning parameters suggested by *glmnet* (ENet) and for the random elastic net estimator with 10-fold cross-validation for the BS step (BS) and for the BSU step (BSU) based on the MSE as criterion.

that our random elastic net estimator provides substantially more accurate estimates of the probability weights than the FKRB estimator. The bias reduction persists for small and large sample sizes.

Additionally, the results for $N = 1,000$ and $R = 1,089$ illustrate that the random elastic net estimator yields good approximations of the true distribution function even if, first, the sample size is rather small and, second, the number of grid points R is large relative to the sample size N and thus, to the number of regression observations NJ. This property is particularly relevant for applications with many random coefficients which require a large number of grid points to accurately approximate the joint distribution of the random coefficients.

Surprisingly, the BSU estimator seems to perform worse than the BS estimator. This might be due to the fact that all active grid points have the same weight $1/S$ in this Monte Carlo setup and therefore, the uniform prior used in the BS step of the random elastic net estimator is already a decent choice, even though it also draws the inactive grid points with the same probability as the active grid points in each bootstrap sample.

estimated average absolute bias is still lower for the BSU than for the elastic net estimator.

Consequently, updating the prior in the BSU step of the random elastic net estimator might not be of great help here. The results of a discrete Monte Carlo study using different values for the positive weights, which we report in Table 3.A.1, support the argument that the BSU estimator seems to perform better than the BS estimator when the positive weights differ from each other.

With respect to selection consistency, the random elastic net estimator recovers more true positive and sign consistent probability weights from the data than the FKRB and the elastic net estimator. For instance, the BS and the BSU estimator select almost all positive grid points correctly for $R = 81$ and $R = 289$ and still select at least 67% for $R = 1,089$. The share of correctly selected grid points for the FKRB and for the elastic net estimator decreases substantially if R increases from $R = 81$ to $R = 1,089$. Especially in the extreme case of $R = 1,089$, the FKRB estimator selects no more than 4% of grid points with positive probability weights correctly, illustrating its LASSO-type behavior. This can also be seen in the small number of positive estimated weights, which does not increase for the FKRB estimator when R increases. In contrast, the number of positive estimated weights increases considerably for the elastic net estimator and even more substantially for the random elastic net estimator.

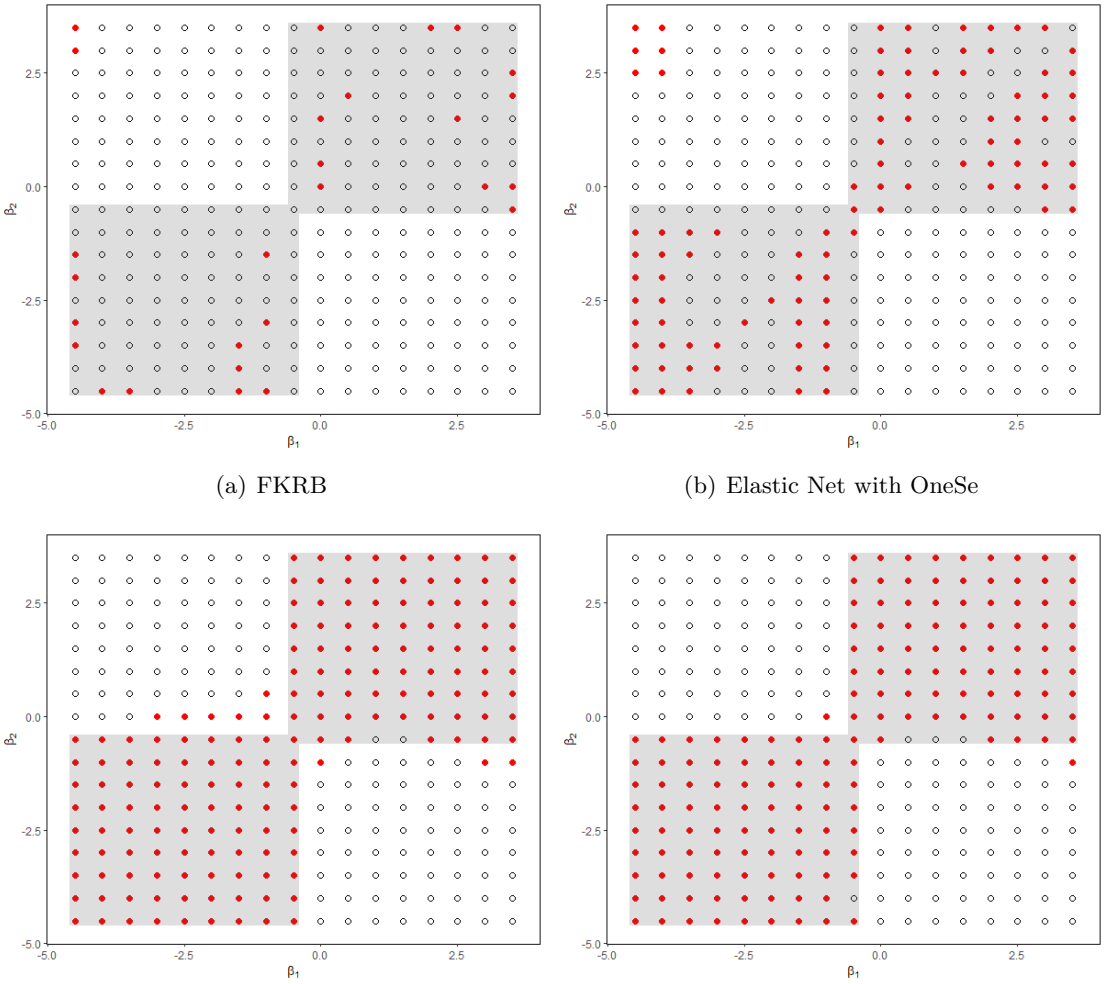
Figure 3.2 illustrates the improved selection consistency of the random elastic net estimator compared to the FKRB and elastic net estimator. It plots an example of the positive grid points estimated with the FKRB estimator (Panel (a)), the elastic net estimator (Panel (b)), the BS estimator of the random elastic net (Panel (c)) and the BSU estimator of the random elastic net (Panel (d)). The estimated positive grid points are displayed in red and should only lie inside the gray shaded area which highlights the support of the discrete mass function.[7]

The values for q and h reported in Table 1 indicate that the number of grid points sampled in the BS step of the random elastic net estimator is at least two times higher than the number of grid points sampled in the BSU step. While the BS estimator samples out of the full set of R grid points, for the BSU estimator this set reduces to those grid points with a positive estimated weight in the BS step. However, drawing less grid points might be sufficient to approximate the lower total number of grid points in the BSU step compared to the BS step. Furthermore, our results on the share of sign consistent (positive) weights indicate that the random elastic estimator tends to select the correct grid points in the BS step. Therefore, we expect the average correlation among grid points included in the BSU step to be higher than the average correlation among all R grid points since the positive grid points are neighbors to each other in our specification of the Monte Carlo study, as can be seen in Figure 3.1.[8] In turn, the optimal number of grid points h used in the BSU step has to decrease compared to q in order to be able to

[7]The two and six positive grid points in the upper left of the plotted grid for the FKRB and the elastic net estimator may indicate that they wrongly select grid points close to the border of the grid. The weights of these grid points sum up to 0.0178 and 0.0172 for the FKRB and the elastic net estimator, respectively.

[8]Note that if the true distribution function was continuous one would also expect the majority of neighboring grid points to share the same sign.

Figure 3.2: Estimated Positive Grid Points for $N = 10,000$ and $R = 289$



(a) FKRB

(b) Elastic Net with OneSe

(c) BS of Random Elastic Net

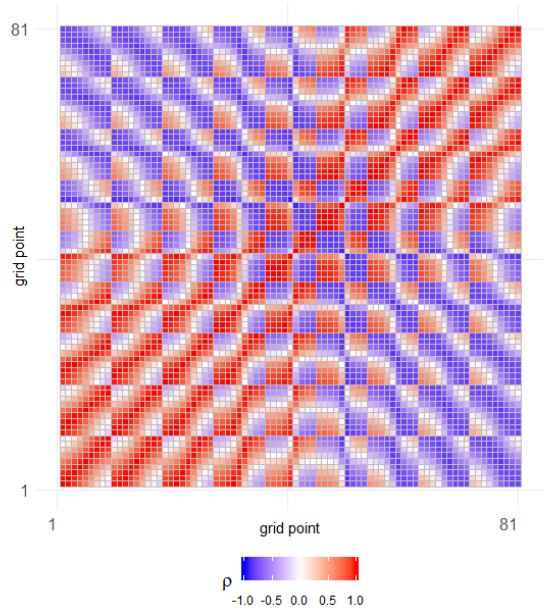
(d) BSU of Random Elastic Net

break down the increased average correlation among grid points.

Additionally, the plot of the correlation matrix in Figure 3.3 and the third quantile of the values of absolute correlation in Table 3.1 both illustrate that the correlation among many grid points is strong.

We report two robustness checks in Appendix 5.B. The first robustness check considers the modification of the random elastic net estimator described in Remark 3.3, which only requires the sum of the weights to be smaller or equal to one in each bootstrap sample. Subsequently, the final weights $\hat{\theta}^{BS}$ and $\hat{\theta}^{BSU}$ are normalized at the end of the BS step and the BSU step to ensure that they sum up to 1. The results of this modification of

Figure 3.3: Correlation Matrix for $N = 10,000$ and $R = 81$



the random elastic net are very similar to those of Table 3.1.

The second robustness check applies the FKRB estimator within each bootstrap sample instead of the elastic net estimator, i.e., $\mu_b = 0$, $b = 1, \ldots, B$ and $\mu_u = 0$, $u = 1, \ldots, B$. We refer to this estimator as random LASSO estimator. However, the summary statistics of the random LASSO estimator are worse than those of the random elastic net estimator. This indicates that taking the correlations among the grid points into account is important for the estimation within each bootstrap sample.

3.3.2 Continuous Distribution

The second Monte Carlo experiment considers a mixture of two bivariate normal distributions for $F_0(\beta)$ to analyze how our generalized estimator accommodates more complex continuous distributions. This way, we can assess its ability to recover distributions that cannot be estimated with parametric techniques (unless the mixture structure, e.g., the number of mixture components and the family of each mixing distribution, was known).

For the estimation, we use a fixed grid with points spread on $[-4.5, 3.5] \times [-4.5, 3.5]$. The fixed grid covers the support of the true distribution with probability close to one (0.993). We keep the correlation among grid points as low as possible and generate the grid points with a Halton sequence. To study the convergence of the estimated distribution to $F_0(\beta)$ for an increasing number of grid points, we estimate the model with

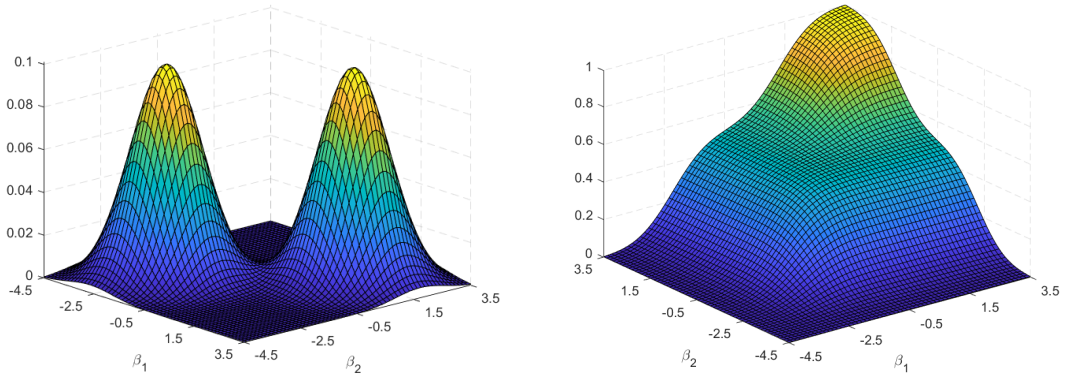
$R = \{100, 300, 500\}$. The number of observation units N varies from 1,000 to 10,000.

The variance-covariance matrices of the two normals are $\Sigma_1 = \Sigma_2 = \left[\begin{smallmatrix} 0.3 & 0.1 \\ 0.1 & 0.3 \end{smallmatrix}\right]$. We generate the random coefficient vectors β from the following two-component bivariate mixture

$$0.5\ \mathcal{N}\Big([-2.2, -2.2], \Sigma_1\Big) + 0.5\ \mathcal{N}\Big([1.3, 1.3], \Sigma_2\Big)$$

The left panel in Figure 3.4 displays the bimodal joint density of the mixture of the two normals, and the right panel the joint distribution function.

Figure 3.4: True Density and Distribution Function of Mixture of two Normals



(a) PDF

(b) CDF

For the calculation of the RMISE, we use $E = 10,000$ evaluation points uniformly distributed over the range of the fixed grid. Instead of the L_1 norm, we report the maximum difference between the true and estimated distribution at the evaluation points

$$\text{Max Dif} = \frac{1}{M} \sum_{m=1}^{M} \max_{e=1,\ldots,E} \left| \widehat{F}_m(\beta_e) - F_0(\beta_e) \right| \ ,$$

averaged over all Monte Carlo runs M.[9]

In addition, we track the average number of positive, true positive, and sign consistent estimated weights. For the number of true positive and sign consistent weights, we calculate the true density at every grid point and then normalize the density of each grid point by the sum of densities at all grid points. We define a true weight as positive if its

[9]We do not report the L_1 norm since the true weight θ_r at every fixed grid point β_r is not known for the continuous distribution and, hence, the value of the L_1 norm would depend on the way we approximate the true weights θ_r, $r = 1, \ldots, R$.

normalized density is greater 10^{-3}.[10]

Table 3.2 summarizes the average results over the $M = 100$ Monte Carlo replicates for the FKRB estimator, the elastic net estimator with μ equal to the maximum value of the sequence of tuning parameters suggested by the *glmnet* package (Friedman et al., 2010), and the random elastic net estimator with $B = 1,000$ bootstrap samples. The optimal number of draws q and h drawn in the BS step and BSU step of the random elastic net estimator is determined by 10-fold cross-validation and the MSE as criterion. We consider the same sequence of variables $C := \{5 + 3k | k = 0, ..., 15\}$ for q and h in the cross-validation, i.e., $q, h \in C$.

Table 3.2: Summary Statistics of 100 Monte Carlo Runs with Mixture of Two Bivariate Normals.

			RMISE				Max. Dif.				Pos.			
N	R	S	FKRB	ENet	BS	BSU	FKRB	ENet	BS	BSU	FKRB	ENet	BS	BSU
1,000	100	59	0.095	0.057	0.052	0.047	0.276	0.177	0.168	0.144	13.56	54.02	66.68	63.56
1,000	300	149	0.103	0.062	0.050	0.045	0.285	0.163	0.137	0.114	14.37	117.73	179.35	165.55
1,000	500	203	0.106	0.066	0.051	0.046	0.286	0.168	0.132	0.111	14.77	158.57	265.69	243.32
10,000	100	59	0.058	0.040	0.037	0.029	0.187	0.129	0.126	0.107	19.68	46.15	61.93	55.50
10,000	300	149	0.064	0.040	0.033	0.023	0.194	0.109	0.089	0.068	21.75	99.96	163.84	143.86
10,000	500	203	0.066	0.042	0.035	0.024	0.197	0.113	0.083	0.064	21.77	134.32	241.85	209.16

			% True Pos.				% Sign				q	h	ρ
N	R	S	FKRB	ENet	BS	BSU	FKRB	ENet	BS	BSU	BS	BSU	3rd Qu.
1,000	100	59	19.36	79.08	92.93	90.98	50.28	80.30	83.98	84.80	44.93	6.62	0.824
1,000	300	149	6.66	64.89	94.17	91.54	52.16	75.55	84.09	86.08	43.91	6.95	0.824
1,000	500	203	3.66	57.63	93.93	90.38	59.42	74.48	82.53	84.13	44.87	6.80	0.825
10,000	100	59	28.47	69.95	91.58	86.59	54.92	77.39	87.13	87.68	43.37	12.44	0.823
10,000	300	149	10.52	57.21	93.65	87.32	53.54	73.85	88.75	89.12	40.82	13.10	0.824
10,000	500	203	6.15	51.66	93.15	85.95	60.04	74.48	86.67	87.36	41.27	12.92	0.825

Note: The table reports the average summary statistics over all Monte Carlo replicates for the FKRB estimator (FKRB), for the elastic net estimator with tuning parameter μ from the maximum value of the sequence of tuning parameters suggested by *glmnet* (ENet) and for the random elastic net estimator with 10-fold cross-validation for the BS step (BS) and for the BSU step (BSU) based on the MSE as criterion.

[10] Using 10^{-4} as a threshold, we obtain qualitatively similar results.

The RMISE shows that the random elastic net estimator provides more accurate estimates of the true underlying random coefficients' distribution than the FKRB and the elastic net estimator for every combination of N and R. For every R, the RMISE of the random elastic net estimator for $N = 1,000$ is even lower than the RMISE of the FKRB estimator for $N = 10,000$. That is, the random elastic net estimator seems to perform well for small sample sizes. Nevertheless, as expected, the RMISE of all estimators decreases when the sample size increases from $N = 1,000$ to $N = 10,000$. Comparing the estimators, the maximum absolute difference between the true and estimated distribution shows a qualitatively similar pattern.

Furthermore, the BSU estimator has a lower RMISE than the BS estimator. This is in contrast to the results obtained in the discrete Monte Carlo study where the BS estimator performed better than the BSU estimator. A possible explanation of this observation might lie in the different distributions we use in the discrete and continuous Monte Carlo study. In the discrete Monte Carlo study, the probability weights of the discrete mass function are all $1/S$. In the continuous Monte Carlo study, there is substantial variation in the magnitude of the probability weights at every grid point (loosely speaking, some are more important than others). Due to this variation, updating the uniform prior of the BS step in the BSU step might have a bigger effect than in the discrete Monte Carlo experiment.[11]

In any case, in both Monte Carlo studies the BS as well as the BSU estimator perform better than the FKRB and elastic net estimator. Thus, both versions of the random elastic net estimator improve the considered benchmarks. It might be sufficient to calculate just the BS estimator in situations when it is computationally demanding to estimate the BSU step of the random elastic net estimator.

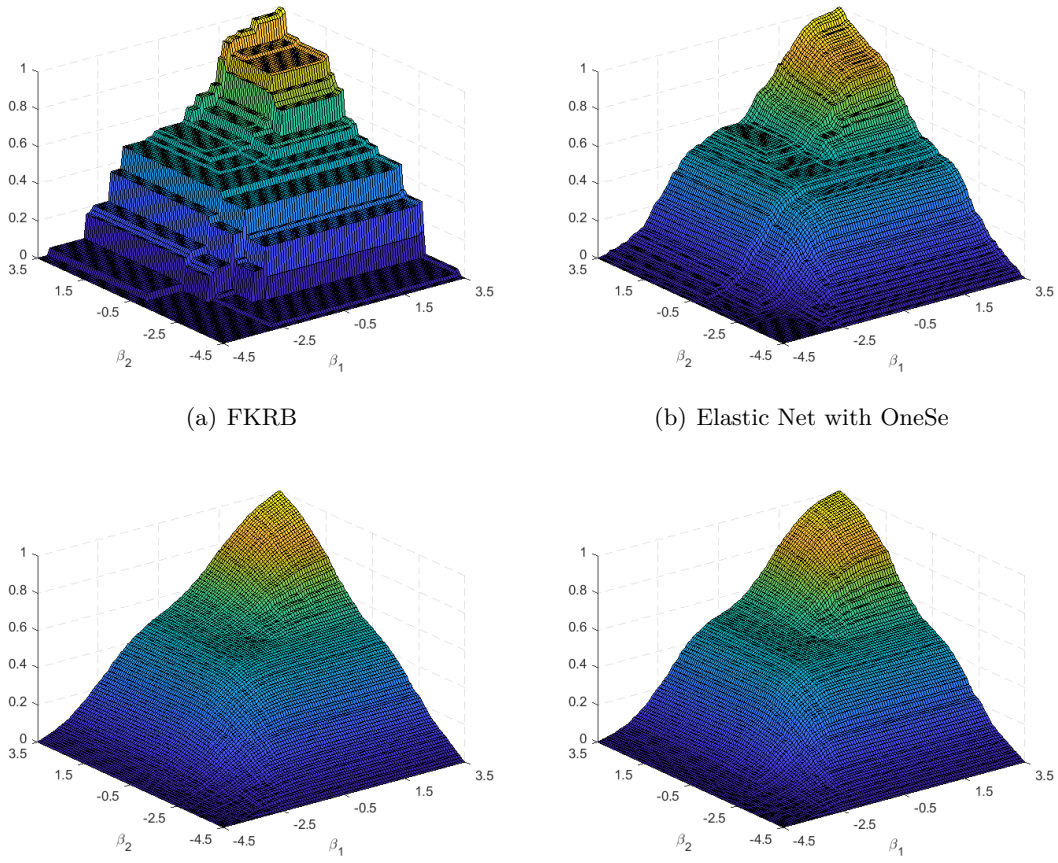
Even though no clear rule emerges when to use the BS or the BSU estimator, we tend to recommend to use the BSU estimator since it approximates the underlying distribution better than the BS if the distribution function is not uniform or close to it.

The improved performance of the random elastic net estimator for every combination of N and R can be explained with the larger number of true positive and sign consistent estimated probability weights. Independently of the number of (relevant) grid points, the FKRB estimator estimates only a small number of positive weights and, hence, recovers only few relevant grid points. While the share of true positive and sign consistent estimated weights is notably higher for the elastic net estimator, it is highest for the random elastic net estimator. For $R = 300$ and $N = 10,000$, the FKRB estimator estimates the sign of about 11% of the positive weights and 54% of all weights correctly. These numbers increase to 57% and 89% for the elastic net estimator and to at least 87% and 89% for the random elastic net estimator.

Figure 3.5 plots an example of the joint distribution functions estimated with the

[11] The results of a discrete Monte Carlo study using nonuniform positive weights (cf. Table 3.A.1) also suggest that the BSU estimator achieves a better fit than the BS estimator when there is variation in the magnitude of the positive weights.

Figure 3.5: Estimated Joint Distribution Functions for $N = 10,000$ and $R = 500$



(a) FKRB

(b) Elastic Net with OneSe

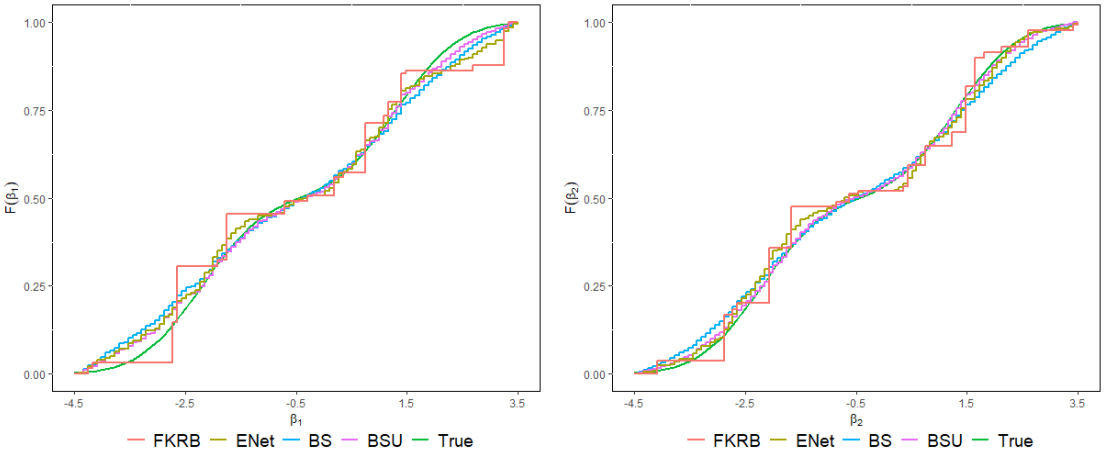
(c) BS of Random Elastic Net

(d) BSU of Random Elastic Net

FKRB estimator (Panel (a)), the elastic net estimator (Panel (b)), the BS estimator of the random elastic net (Panel (c)) and the BSU estimator of the random elastic net (Panel (d)). Figure 3.6 shows the corresponding estimated and true marginal distributions of β_1 and β_2. The distribution functions are estimated for $N = 10,000$ and $R = 500$.

The plots illustrate the impact of the FKRB estimator's sparse nature on the estimated marginal and joint distribution functions. Visual inspection shows that it approximates $F_0(\beta)$ through a step function with only few steps due to the small number of positive weights. In contrast, the elastic net and the random elastic net estimator provide smooth estimates that are close to the true underlying distribution function (cf. Figure 3.4 (b)). In particular, the BSU estimator of the random elastic net achieves the best fit. Figure 3.A.2, where we plot the error between estimated estimated joint distribution function and the true distribution function for each estimator, confirms this impression.

Figure 3.6: True and Estimated Marginal Distribution Functions for $N = 10,000$ and $R = 500$



For the continuous Monte Carlo study, we conducted the same two robustness checks as for the discrete experiment. The results are again summarized in Appendix 5.B. The first robustness check, which only requires the sum of the weights to be smaller or equal to one in each bootstrap sample and normalizes the final weights afterwards, yields very similar results to the random elastic net estimates reported in Table 3.2.

The second robustness check uses the random LASSO estimator. The results show that the random LASSO estimator performs better in the BS step and worse in the BSU step than the random elastic net estimator. However, we obtain the best overall result of all considered estimators with the BSU step of the random elastic net estimator. Thus, the BSU step of the random elastic net estimator in Table 3.2 gives the best approximation of the true distribution function.

3.4 Application

To study the performance of our random elastic net estimator with real data, we apply it to the model developed by Blundell et al. (2020).[12] We compare the results of the random elastic net estimator to the results of Blundell et al. (2020) which they obtain by applying the FKRB estimator.

[12]We would like to thank Blundell et al. (2020) for sharing their data and code online.

3.4.1 Empirical Framework

Blundell et al. (2020) study the gains from dynamic enforcement of air pollution regulations. The U.S. Environmental Protection Agency (EPA) uses dynamic enforcement of air pollution regulations to enforce the Clean Air Act Amendments (CAAA).

If a plant is not in compliance with the regulations, the EPA can classify the plant either as regular violator or as high priority violator (HPV). A plant is designated as HPV if a violation is particularly severe or, more importantly, if violations occur repeatedly.

While regular violators are more likely to be inspected and to violate regulations than plants that are in compliance, HPVs are even more likely to be inspected and to violate regulations than regular violators.

The EPA determines fines for violations based on the gravity of the violation and the economic benefit that the plant received from the violation. Additionally, fines also escalate with the the regulatory state of the plant. That is, fines for violations increase dramatically for plants in HPV status compared to those in regular violator status (cf. Blundell et al. (2020), Figure 1).

The dependence of fines on the violator status is a key element of dynamic enforcement. First, dynamic enforcement underpenalizes small violations and gives plants time to fix their violations. Second, it uses the threat of high fines for repeated violations, i.e., for plans in HPV status, as incentive for plants to invest in pollution abatement. Dynamic enforcement might not only be beneficial for plants but also for the regulator, e.g., if the imposition of fines is costly to the regulator or when the regulator cannot infer plants' compliance costs with its regulatory policies.

Besides issued fines, HPVs face costs from the increased level of regulatory oversight. According to Blundell et al. (2020), plants in HPV status receive concrete deadlines for EPA and plant actions to resolve outstanding violations. Also, more frequent inspections might be costly since plants might have to shut down production lines to allow for an inspection. HPVs might face additional costs from potential loss of reputation. Using a watchlist, HPVs that fail to resolve all of their violations on time are publicly disclosed by the EPA.[13]

To exit HPV status, plants have to resolve all outstanding violations and to invest in pollution abatement technologies. Without an investment, plants cannot leave HPV status. Even though some regular violators transition without investment to compliance, this is not the case for most of them. More precisely, most regular violators also need to make an investment to return to compliance. That being said, the probability of returning to compliance after an investment is lower than 50% for both regular violators and HPVs.

Blundell et al. (2020) build and estimate a dynamic model of plants and a regulator. In each period t, the actions of the regulator are assumed to depend only on the regulatory

[13]The watchlist was eliminated after the sample period analyzed by Blundell et al. (2020).

state Ω_t and a predictor of compliance issues e_t. The regulatory state Ω_t consists of the following six components: EPA regions, two-digit NAICS industrial sector[14], expected gravity of potential violations, which differs across counties and industries, depreciated accumulated violations, regular violator or high priority violator status and two quarterly lags of investment. The states of EPA region, industry, and gravity do not change over time. The states of compliance, lagged depreciated accumulated violation status, violation status, and lagged investment can change from period t to $t+1$.

Conditional on the regulatory state and an *i.i.d.* private information shock, the regulator first decides whether or not to inspect a plant. If a plant is inspected, the regulator finds out whether there is a violation or not. If violations occur, the regulator determines the fines for these violations. Furthermore, the compliance status of a plant can change due to an inspection, e.g., a plant may transition to regular violator or to HPV status or back to compliance.

Blundell et al. (2020) model the actions of the regulator using conditional choice probabilities, i.e., they estimate plants' expectations of regulatory actions conditional on the state Ω_t and the environmental compliance signal e_t. This avoids making an assumption on the utility function of the regulator.

Subsequent to the regulator actions, plants not in compliance can decide whether or not to invest in pollution abatement. Regular or high priority violators will only invest in pollution abatement if the expected regulatory costs from inspections, fines, violations, and designation as a high priority violator exceed the investment costs. Therefore, it is necessary to accurately recover the regulatory and investment costs to estimate the value of dynamic enforcement of the CAAA.

3.4.2 Estimation of the Model

Regulatory costs and investment costs may be heterogeneous. For instance, plants with low investment costs may have incentives to invest in pollution abatement when they are regular violators and fines are still low while plants with high investment costs will wait until they become HPV and fines are higher. Additionally to costs from inspections, violations, fines, and investments, plants are assumed to bear costs from being HPV but not from being regular violator.

Blundell et al. (2020) account for the potential heterogeneity of plants in regulatory and investment costs by modeling these costs as random coefficients. To be precise, they allow costs from inspection, β^I, violations, β^V, fines, β^F, investments, β^X, and HPV status, β^H to vary across plants, i.e., $\beta = (\beta^I, \beta^V, \beta^F, \beta^X, \beta^H)'$ is the vector of 5 random coefficients. For each plant, its own utility parameters, which are fully described by β, are assumed to be constant over time.

In order to estimate the joint distribution of β, Blundell et al. (2020) apply the

[14]The data of Blundell et al. (2020) includes the seven most polluting North American Industry Classification System(NAICS) industrial sectors.

FKRB estimator. More precisely, they generate a fixed grid for β consisting of $10,001$ 5-dimensional grid points.[15] The way Blundell et al. (2020) construct the dependent variable y and the regressors Z in the FKRB estimator in Equation (3.5)[16] differs substantially from the approach outlined in Section 3.2. The dependent variable y is not discrete anymore and Z is not calculated using a logit kernel. Instead, the dependent variable y represents a empirical moments observed in the data and the regressors z^r are structural moments implied by the dynamic model for given fixed grid points β_r, $r = 1, \ldots, 10,001$. In the FKRB and elastic net estimator in Equation (3.5), the number of products J becomes one and the number of observations N equals the number of moments K, i.e., $J = 1$ and $N = K$. This approach closely follows the GMM framework proposed by Nevo et al. (2016) which also adopts the FKRB estimator to match structural moments to observed moments, in order to estimate the distribution of random coefficients. We present the elastic net version of the estimator employed by Blundell et al. (2020):

$$\begin{aligned} \hat{\theta}^{\text{ENet-BGL}} = \arg\min_{\theta} & \left(m^d - \sum_{r=1}^{R} \theta_r m(\beta_r) \right)' W \left(m^d - \sum_{r=1}^{R} \theta_r m(\beta_r) \right) \\ \text{s.t.} \quad \theta_r \geq 0 \quad \forall r \quad \text{and} \quad & \sum_{r=1}^{R} \theta_r = 1 \quad \text{and} \quad \sum_{r=1}^{R} \theta_r^2 \leq t \end{aligned} \tag{3.8}$$

where m^d is a K-dimensional vector of empirical moments and $m(\beta_r)$ is a K-dimensional vector of structural moments of the model calculated at β_r. We use the same weighting matrix W as Blundell et al. (2020) in the first step.[17] It is updated in a second step to increase efficiency of the GMM estimation.[18] The tuning parameter $t \geq 0$ is again chosen by the researcher as in Equation (3.5), e.g., by using cross-validation. For $t = 1$, the estimator simplifies to the FKRB estimator used by Blundell et al. (2020).

For each β_r, $m(\beta_r)$ is recovered from the model using the Bellman equation, i.e., the dynamic model is solved by dynamic optimization for each grid point β_r, $r = 1, \ldots, R$, before the estimation of θ commences. Blundell et al. (2020) use $K = 14,374$ moments. The first $5,000$ moments represent the equilibrium share of plants in a particular time-varying state conditional on the non-time-varying states. The next $4,687$ moments, i.e., the second set of moments, multiply the first moments by the the conditional share of plants having an investment at that state.[19] The last $4,687$ moments are referred to as the third set of moments. They multiply the moments in the second set by the average

[15] They construct the first $10,000$ grid points using a Halton sequence. Additionally, they include the estimate of a quasi-likelihood model as a grid point.

[16] Recall that the elastic net estimator in Equation (3.5) reduces to the FKRB estimator for $t = 1$.

[17] In the first step, they calculate the weighting matrix as the inverse of the variance-covariance matrix of the moments $m(\beta_{QL})$ at the quasi-likelihood estimate β_{QL}.

[18] We follow Blundell et al. (2020) and calculate W in the same manner as they do in the second step. We refer the interested reader to Blundell et al. (2020) for the details.

[19] 313 of the 5000 moments in the first set are excluded in the second set of moments since they represent states in compliance and therefore, there is no investment in these states. For the same reason, those moments are also excluded in the third moments.

number of investments in the following six periods, corresponding to plants that invest at that state. The second and third set of moments should capture the important role of investment in order to return to compliance.

We estimate the elastic net estimator in Equation (3.8) by cross-validation where we use the sequence of tuning parameters generated by the *glmnet* package (Friedman et al., 2010).[20] For the random elastic net estimator, we use $B = 2,000$ bootstrap repetitions. In the cross-validation, we consider the sequence $q \in \{15k \,|\, k = 1, \ldots, 50\}$ for the BS step and $h \in \{5k \,|\, k = 1, \ldots, 50\}$ for the BSU step.[21] In contrast to the Monte Carlo simulations, we choose the minimum value of the sequence suggested by *glmnet* for μ_b. In this application, we observe that the maximum value of the sequence suggested by *glmnet* penalizes the weights too heavily such that they are all equal. However, using the optimal q from the cross-validation of the BS step, we repeat the BS step where we tune μ_b by cross-validation. In this procedure, also the smallest μ_b is selected most of the times. Furthermore, note that selecting the smallest instead of the largest element of the sequence for μ_b might give similar results as the random LASSO which also performs decently in the Monte Carlo studies presented in Section 3.3.

In line with Blundell et al. (2020), we define a weight to be positive if it is greater than 10^{-5}. The FKRB estimator only estimates 12 out of the $10,001$ weights to be positive, i.e., the result is sparse due to the LASSO property of the FKRB estimator. The elastic net estimator estimates 439 positive weights. For the random elastic net estimator, the BS step estimates 772 positive weights, which are subsequently used in the BSU step. The BSU step estimates 154 positive weights.[22] For the random elastic net estimator, we focus on the BSU step in the presentation of the subsequent results.

Figure 3.7 displays two-dimensional contour plots of the five dimensional random coefficients' distribution estimated with the FKRB, elastic net, and random elastic net estimator. Using heat maps, the figure highlights the correlation between the dimensions of the random coefficients' distribution. The plots for the elastic net and random elastic net estimator look similar and their estimated distribution functions suggest rather sophisticated correlation patterns between the random coefficients. Most plants seem to face costs from investment and from being in HPV status, as indicated by the negative range of values. However, a substantial share of plants finds inspections beneficial (36% for the FKRB estimator, 44% for the elastic net estimator, and 42% for the random elastic net estimator). For the FKRB estimator, the estimated contour plots do not seem very informative since there are only 12 positive weights. We provide the heat maps of the dimensions of the other random coefficients in Figure 3.A.7 and 3.A.8 in the Appendix.

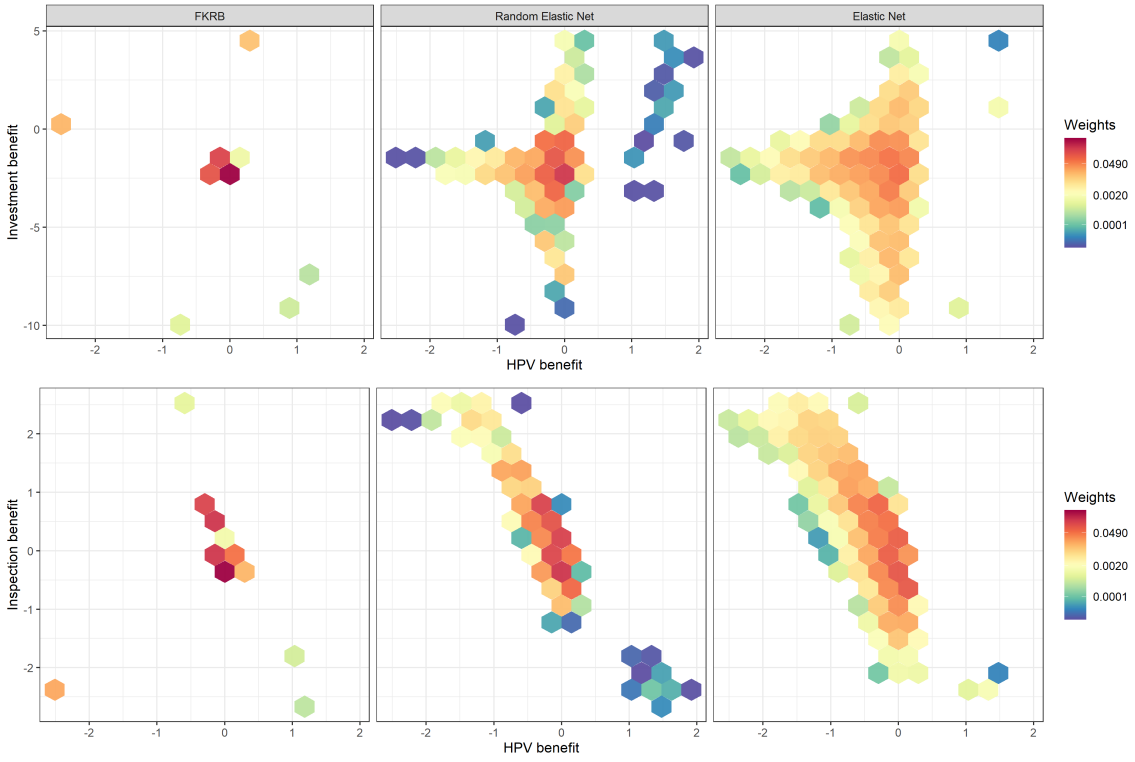
[20] To prevent data leakage, we sample those moments in the cross-validation together which are included in all of the three sets of moments.

[21] In the cross-validation, $q = 750$, which is the highest value included in the sequence, and $h = 130$ are selected. That said, the results of the BS step for, e.g., $q = 500$ and $q = 1,000$ seem similar.

[22] If we used a threshold of 10^{-3} to define a positive weight, the FKRB, elastic net and random elastic estimator would estimate 9, 275, and 75 positive weights, respectively. Using 10^{-5} as a threshold, we try to avoid omitting mass at any grid point in the subsequent counterfactuals.

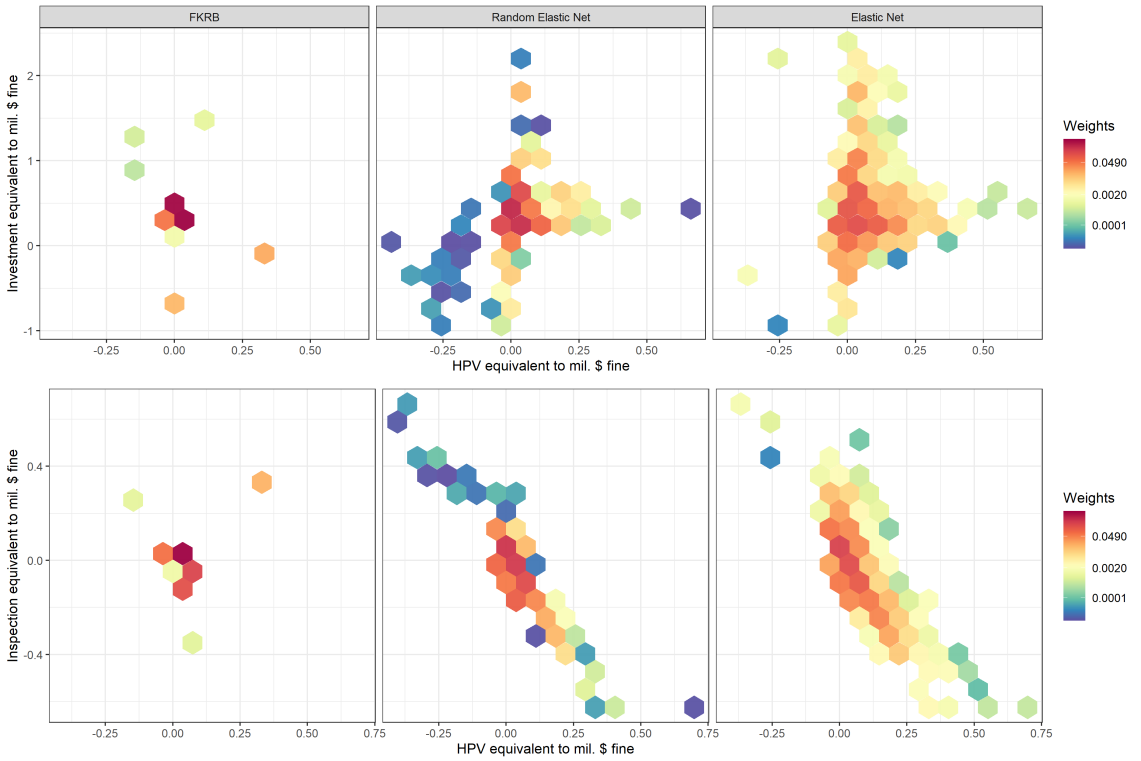
Additionally, we plot the estimated marginal CDFs, mass functions and corresponding histograms in Figure 3.A.3 – 3.A.5 in the Appendix. The estimated marginal CDFs of the random elastic net estimator lie between those of the FKRB and the elastic net estimator. Regarding the estimated histograms, the FKRB estimator does not allow to infer the underlying distribution function which is only possible for the histograms of the elastic net and of the random elastic net estimator (cf. Figure 3.A.5). Their histograms appear similar for most random coefficients. We interpret this as a sign of robustness of the results for these estimators. Furthermore, the weighted means of the random coefficients in Figure 3.A.4 do not deviate substantially for the FKRB, elastic net, and random elastic net estimator, suggesting that the means of the random coefficients can be recovered by each of these estimators.

Figure 3.7: Estimated Heat Maps for Blundell et al. (2020).



Given that the random coefficients represent utility parameters, it is difficult to interpret them. Therefore, Blundell et al. (2020) consider the ratios of the coefficients relative to fines. More precisely, β^X/β^F indicates how costly investments are for plants expressed in \$1 million fine per quarter. Similarly, β^V/β^F is the equivalent of an additional violation, β^I/β^F the equivalent of an inspection, and β^H/β^F the equivalent

Figure 3.8: Estimated Heat Maps for Equivalent in Fines for Blundell et al. (2020)



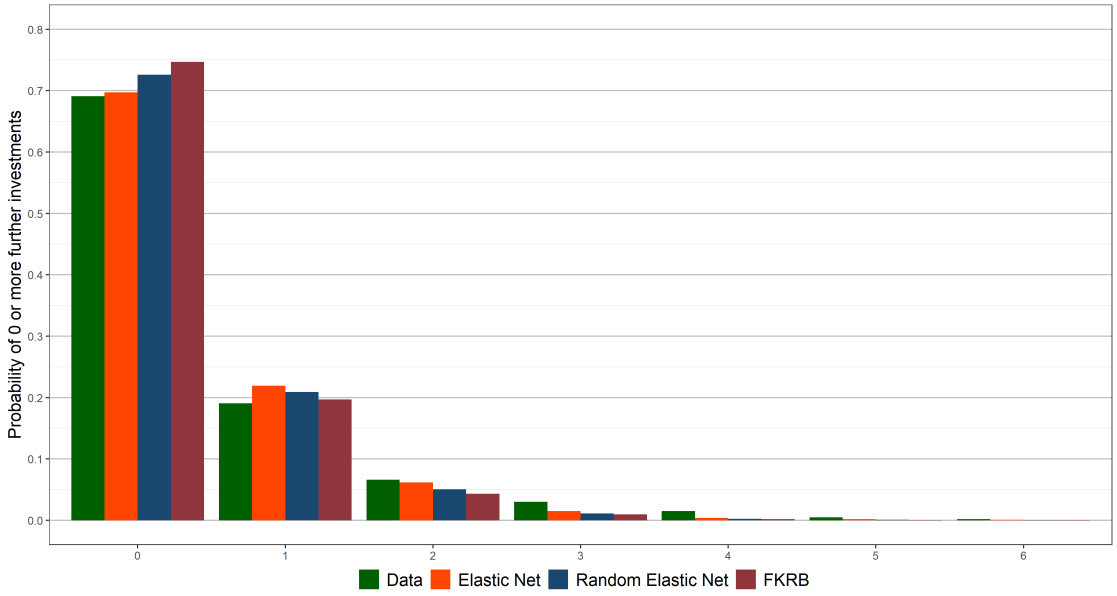
of HPV status to a \$1 million fine per quarter. Figure 3.8 shows the heat maps for those ratios, i.e., we plot β^X/β^F and β^I/β^F against β^H/β^F. Thereby, the figure visualizes the estimated distributions for three random coefficients in each heat map.

As expected, most plants find it costly to be in HPV status and to invest, i.e., they have a positive value for β^H/β^F and β^X/β^F.[23] However, the negative values of inspection, expressed in \$1 million fine per quarter, indicate that some plants find it beneficial to be inspected. An explanation of this surprising observation might be that those plants want to be inspected in order to leave regular or HPV status. Leaving violator status is only possible after an additional inspection, confirming all violations have been resolved. The weights of the random coefficients in \$1 million fine per quarter can also be inspected in the histograms in Figure 3.A.6 in the Appendix and in the heat maps for the remaining ratios in Figure 3.A.9 in the Appendix.

[23]The signs of the ratios β^X/β^F, β^I/β^F, and β^H/β^F are determined by the signs of β^X, β^I, and β^X, which can be positive or negative (cf. Figure 3.7), since β^F is negative for all grid points, i.e., fines are costly to plants. Hence, we expect a positive value for β^H/β^F and β^X/β^F since this implies negative values for β^H and β^X, i.e., plants dislike being in HPV status and face costs from investments and fines.

Figure 3.9 illustrates the model fit of the FKRB, elastic net, and random elastic net estimators in terms of further investments after an initial investment. Recall that investment is a key variable for plants since plants can only return to compliance if they invest. Yet, the success of investments is stochastic and therefore, repeated investments are necessary. All three estimators match the investment patterns in the data quite well. According to the MSE between the model fit and the data, the elastic net estimator is closest to the data and the random elastic net estimator is closer to the data than the FKRB estimator. Blundell et al. (2020) note that if the investments were *i.i.d.*, we would expect only 2.3% of additional investments compared to the 30% observed in the data. A random coefficients model can capture this stylized fact which might not be possible for models without random coefficients.

Figure 3.9: Further Investments in the Six Periods After Initial Investment for Blundell et al. (2020)



3.4.3 Counterfactuals

We estimate three counterfactuals presented in Blundell et al. (2020) using our estimates from the elastic net, and random elastic net estimator and compare them to the results of Blundell et al. (2020).The first counterfactual examines the value of dynamic enforcement. To this end, the cost of being in HPV status is removed and the regulator fines all plants in regular and HPV status identically for a given region, industry, and gravity state. However, the total assessed fines are kept the same as in the baseline model for each

region, industry, and gravity state. The second counterfactual is the same as the first one, except that pollution damages are kept the same as in the baseline model across region, industry, and gravity state instead of total fines.

The third counterfactual changes the escalation mechanism of fines. For plants in HPV status, fines are doubled compared to the baseline model.

Table 3.3 reports the results for these counterfactuals. The baseline column for each model shows long run mean values implied by the estimated structural parameters. The baseline estimates of each model are very similar and replicate the data quite well.

Table 3.3: Results of Counterfactuals

		Baseline			Same fines for all violators; fines constant		
	Data	FKRB	ENet	RENet	FKRB	ENet	RENet
Compliance (%)	95.62	95.11	95.17	95.17	66.72	57.14	65.16
Regular violator (%)	2.88	3.47	3.48	3.45	2.53	2.05	2.37
HPV (%)	1.50	1.42	1.35	1.37	30.75	40.82	32.47
Investment rate (%)	0.40	0.54	0.54	0.55	0.47	0.43	0.46
Inspection rate (%)	9.65	9.41	9.40	9.39	20.54	24.06	21.05
Fines (thousands $)	0.18	0.32	0.32	0.31	0.32	0.32	0.31
Violations (%)	0.55	0.54	0.54	0.53	5.00	6.05	4.85
Plant utility	—	0.006	0.020	0.022	0.077	0.171	0.117
Pollution damages (mil. $)	1.65	1.53	1.51	1.52	4.04	4.84	4.18
		Same fines for all violators; pollution damages constant			Fines for HPVs doubled relative to baseline		
	Data	FKRB	ENet	RENet	FKRB	ENet	RENet
Compliance (%)	95.62	94.49	94.73	94.93	95.52	95.57	95.57
Regular violator (%)	2.88	2.72	1.91	2.06	3.47	3.49	3.46
HPV (%)	1.50	2.79	3.36	3.01	1.01	0.95	0.97
Investment rate (%)	0.40	0.65	0.76	0.75	0.55	0.54	0.55
Inspection rate (%)	9.65	9.88	10.03	9.93	9.28	9.26	9.26
Fines (thousands $)	0.18	1.98	4.90	3.70	0.36	0.36	0.35
Violations (%)	0.55	0.74	0.80	0.76	0.49	0.48	0.48
Plant utility	—	0.001	0.006	0.010	0.005	0.019	0.021
Pollution damages (mil. $)	1.65	1.53	1.51	1.52	1.48	1.47	1.48

Note: The table reports summary statistics of the values in the data, the values predicted by the model estimates (baseline), and three counterfactuals for the FKRB, elastic net, and random elastic net estimator. Each summary statistic is per plant / quarter and calculated in the same way as in Blundell et al. (2020). The counterfactuals of FKRB correspond to those given in Blundell et al. (2020).

For the first counterfactual, the estimates of the FKRB and random elastic net estimator deviate only slightly from each other. In contrast, the elastic net estimator predicts a more drastic reaction if fines for HPVs are the same as for regular violators. For instance, the share of plants in HPV status rises from 1.35% to 41% for the elastic net estimator compared to a rise from 1.42% (1.37%) to 31% (33%) for the FKRB (random elastic net) estimator. For the elastic net estimator, pollution damages increase from \$1.51 to \$4.84 million per plant / quarter instead of \$1.53 to \$4.04 and 1.52 to \$4.18 million per plant / quarter for the FKRB and random elastic net estimator.

For the second counterfactual, the values of most estimated summary statistics of the three models are close to each other when pollution damages are held constant, but fines do not escalate with regulatory state. However, mean fines increase substantially with this regulatory policy. That is, the FKRB estimates imply that mean fines increase from \$320 to \$1,980 per plan / quarter and even to \$4,900 and \$3,700 per plan / quarter for the elastic net and random elastic net estimates. Therefore, mean fines are 2.5 and 1.9 times higher for the elastic net and random elastic net estimator compared to the FKRB estimator. Thus, the elastic net and random elastic net estimates imply that it may be more costly to keep the pollution damages constant if additional fines for HPVs are removed.

For the third counterfactual, the results of the elastic net and random elastic net estimator are in line with those of the FKRB estimator. All of the three models predict more or less the same shares for each variable when fines for HPVs are doubled relative to the baseline. In particular, mean fines increase slightly, pollution damages drop slightly and also the shares of plants in HPV status drops from 1.4% to 1% for the three estimators. For the FKRB estimator, Blundell et al. (2020) argue that the result of the third counterfactual shows that there is only limited benefit from increasing the escalation rate of fines. This conclusion is supported by the elastic net and random elastic net estimates.

Overall, the implications of the results of the estimators for the counterfactuals largely coincide. Therefore, it is not clear to what extent the more accurate estimation of the distribution functions translates to different conclusions for counterfactual policies.

3.5 Conclusion

We extend the simple and computationally attractive nonparametric elastic net estimator of Heiss et al. (2021), which includes the nonnegative LASSO estimator of Fox et al. (2011) as a special case. To this end, we propose a random elastic net estimator which mitigates the sparsity of the solutions and allows to estimate the random coefficients' distribution more accurately. The key idea is to repeatedly estimate the model by only using a subset of the regressors and to average these estimates in the end. Thereby, a substantial part of the correlation among the regressors is broken down before the estimation.

Two Monte Carlo studies illustrate the improved performance of the random elastic net estimator compared to the FKRB and the elastic net estimator. They show that the random elastic net estimator estimates considerably more positive probability weights and recovers more grid points correctly. In addition to the improved selection consistency, the estimator provides more accurate estimates of the true underlying distributions. Both steps of the random elastic net estimator, the BS step and the BSU step, perform better than the FKRB and elastic net estimator. While the BS estimator achieves the best fit in the discrete Monte Carlo study, the BSU yields the best approximation of the random coefficients' distribution in the continuous Monte Carlo study. We recommend to use the BSU estimator since the setup of the continuous Monte Carlo study might be more realistic, as it allows for a more diverse range of consumer types than the discrete Monte Carlo study.

We apply the random elastic net estimator to the study of Blundell et al. (2020) who analyze the gains from dynamic enforcement of air pollution regulations. The results highlight that the elastic net and random elastic net estimator can estimate rather complicated distribution functions. The estimated distribution functions of the elastic net and random elastic net estimator seem similar, suggesting that the results are robust. In contrast, the distribution function estimated with FKRB estimator is not very informative due to its sparse solution. Yet, the results of three conducted counterfactuals seem to qualitatively coincide for the FKRB, elastic net, and random elastic net estimator. Therefore, it is not clear how the more accurate estimation of the random coefficients' distribution function translates to different conclusions for counterfactual policies.

We do not discuss the development of an inference procedure for the random elastic net estimator, which is a practically relevant topic. Since we average all estimates of the bootstrap repetitions to calculate our final random elastic net estimate, we need to account for the uncertainty in each bootstrap sample to obtain a valid inference procedure. This is challenging since each bootstrap estimate is obtained using a nonnegative elastic net estimator. More precisely, for valid inference it is necessary to de-bias the nonnegative elastic net estimates of each bootstrap sample (cf. Dezeure et al. (2017) for a de-biasing procedure for LASSO). However, it is not straightforward how to construct such a de-biased estimator in our setting.

Appendix 3.A Supplementary Tables and Figures

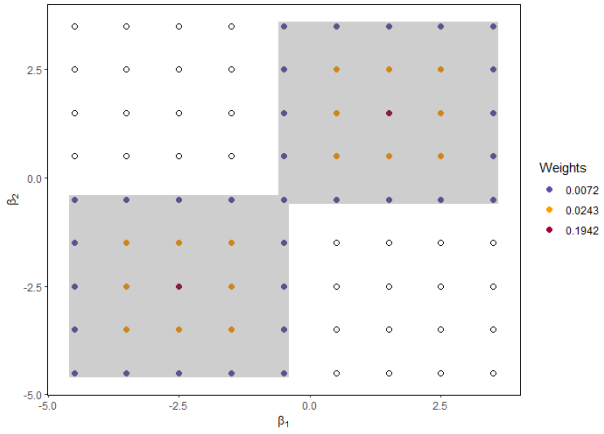
Table 3.A.1: Summary Statistics of 100 Monte Carlo Runs with Discrete Distribution as given in Figure 3.A.1.

			RMISE				$100 \times L_1$				Pos.			
N	R	S	FKRB	ENet	BS	BSU	FKRB	ENet	BS	BSU	FKRB	ENet	BS	BSU
1,000	81	49	0.106	0.075	0.071	0.067	1.778	1.072	0.978	0.983	14.10	47.54	64.32	61.54
1,000	289	161	0.107	0.063	0.050	0.046	0.630	0.321	0.229	0.226	14.81	120.94	223.26	209.33
1,000	1,089	577	0.111	0.069	0.044	0.041	0.179	0.107	0.058	0.061	15.15	279.38	712.94	646.43
10,000	81	49	0.069	0.062	0.059	0.048	1.359	1.058	0.901	0.889	19.96	40.89	65.09	60.45
10,000	289	161	0.067	0.042	0.035	0.028	0.591	0.313	0.188	0.188	22.32	100.22	212.85	191.70
10,000	1,089	577	0.072	0.047	0.031	0.026	0.176	0.107	0.048	0.050	22.70	231.23	675.74	596.40

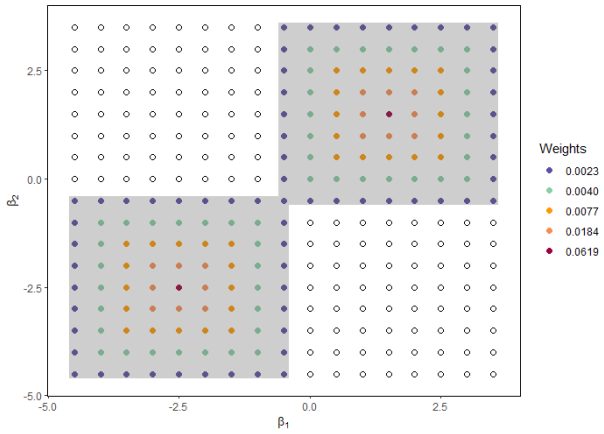
			% True Pos.				% Sign				q	h	ρ
N	R	S	FKRB	ENet	BS	BSU	FKRB	ENet	BS	BSU	BS	BSU	3rd Qu.
1,000	81	49	25.73	84.45	98.20	97.10	53.23	82.99	78.91	81.01	43.61	6.17	0.819
1,000	289	161	7.99	65.80	98.93	97.54	48.07	75.75	77.26	80.54	44.87	6.08	0.822
1,000	1,089	577	2.28	44.01	97.21	93.11	48.04	68.00	84.56	86.32	46.79	6.86	0.824
10,000	81	49	36.12	75.47	99.49	98.47	58.57	80.33	79.52	84.01	39.44	12.98	0.819
10,000	289	161	12.26	57.88	99.68	98.19	50.23	74.10	81.70	87.36	40.64	16.55	0.822
10,000	1,089	577	3.46	37.82	98.29	92.99	48.60	65.86	89.12	90.80	40.73	15.98	0.824

Note: The table reports the average summary statistics over all Monte Carlo replicates for the FKRB estimator (FKRB), for the elastic net estimator with tuning parameter μ from the maximum value of the sequence of tuning parameters suggested by *glmnet* (ENet) and for the random elastic net estimator with 10-fold cross-validation for the BS step (BS) and for the BSU step (BSU) based on the MSE as criterion. Weights outside the gray shaded area in Figure 3.A.1 are zero. In each shaded area, the weights are calculated by $\tilde{\theta}_i = 1/(1+d_i)^3$ where d_i is the Manhattan distance to the center for β_i. We normalize the weights such that they sum up to one, i.e., $\theta_i = \tilde{\theta}_i / \big(\sum_{r=1}^{R} \tilde{\theta}_r\big)$. Weights are defined to be positive if they are greater than 10^{-4}.

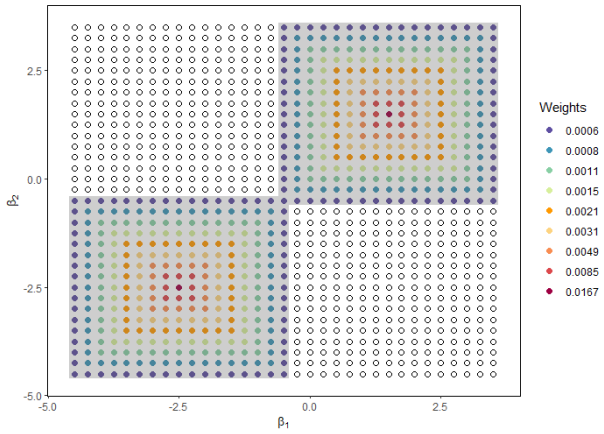
Figure 3.A.1: Grid of Monte Carlo Study with Discrete Mass Points But Different Weights



(a) $R = 81$, $S = 49$



(b) $R = 289$, $S = 161$



(c) $R = 1,089$, $S = 577$

Table 3.A.2: Summary Statistics of 100 Monte Carlo Runs with Discrete Distribution and Normalized Weights of Random Elastic Net.

			RMISE				$100 \times L_1$				Pos.			
N	R	S	FKRB	ENet	BS	BSU	FKRB	ENet	BS	BSU	FKRB	ENet	BS	BSU
1,000	81	49	0.078	0.034	0.026	0.028	1.857	0.700	0.473	0.552	15.63	55.21	70.43	64.47
1,000	289	161	0.087	0.043	0.024	0.027	0.636	0.308	0.143	0.171	15.85	122.91	198.65	176.48
1,000	1,089	577	0.088	0.052	0.023	0.025	0.179	0.115	0.041	0.053	16.32	240.11	470.89	404.04
10,000	81	49	0.050	0.027	0.019	0.028	1.546	0.739	0.427	0.577	23.41	47.64	65.14	58.55
10,000	289	161	0.059	0.035	0.019	0.028	0.602	0.332	0.131	0.178	24.70	106.45	185.69	168.57
10,000	1,089	577	0.062	0.040	0.017	0.025	0.177	0.122	0.038	0.052	24.73	205.29	466.13	397.61

			% True Pos.				% Sign				q	h	ρ
N	R	S	FKRB	ENet	BS	BSU	FKRB	ENet	BS	BSU	BS	BSU	3rd Qu.
1,000	81	49	27.00	90.61	99.43	99.08	52.88	80.98	72.85	79.79	27.56	5.27	0.819
1,000	289	161	8.32	65.21	97.86	95.11	48.07	74.42	84.59	89.20	32.75	5.48	0.822
1,000	1,089	577	2.34	36.74	78.36	66.99	48.00	63.90	86.82	80.90	32.12	5.48	0.824
10,000	81	49	42.51	87.31	99.90	99.51	62.04	86.32	79.95	87.62	29.90	19.16	0.819
10,000	289	161	13.42	60.14	99.02	95.79	50.69	74.46	90.37	92.69	34.61	22.04	0.822
10,000	1,089	577	3.73	33.01	79.01	67.32	48.70	63.15	87.93	81.84	37.70	23.60	0.824

Note: The table reports the average summary statistics over all Monte Carlo replicates for the FKRB estimator (FKRB), for the elastic net estimator with tuning parameter μ from the maximum value of the sequence of tuning parameters suggested by *glmnet* (ENet) and for the random elastic net estimator with 10-fold cross-validation for the BS step (BS) and for the BSU step (BSU) based on the MSE as criterion. The weights in each bootstrap sample have only to be smaller or equal to one, i.e., $\sum_{r \in Q_b} \theta_r \leq 1$ and $\sum_{r \in H_u} \theta_r \leq 1$. At the end of step 1 and step 2, the weights $\hat{\theta}_j^{BS}, j = 1, \ldots, R$, and $\hat{\theta}_j^{BSU}, j = 1, \ldots, R$, are normalized by $\hat{\theta}_j^{BS} = \sum_{b=1}^{B} \hat{\theta}_j^{(b)} / \left(\sum_{j=1}^{R} \sum_{b=1}^{B} \hat{\theta}_j^{(b)} \right)$ and $\hat{\theta}_j^{BSU} = \sum_{u=1}^{B} \hat{\theta}_j^{(u)} / \left(\sum_{j=1}^{R} \sum_{u=1}^{B} \hat{\theta}_j^{(u)} \right)$.

Table 3.A.3: Summary Statistics of 100 Monte Carlo Runs with Discrete Distribution using Random LASSO instead of Random Elastic Net.

			RMISE				$100 \times L_1$				Pos.			
N	R	S	FKRB	ENet	BS	BSU	FKRB	ENet	BS	BSU	FKRB	ENet	BS	BSU
1,000	81	49	0.078	0.034	0.036	0.050	1.857	0.700	0.702	0.868	15.63	55.21	69.83	63.75
1,000	289	161	0.087	0.043	0.035	0.050	0.636	0.308	0.213	0.260	15.85	122.91	188.05	165.16
1,000	1,089	577	0.088	0.052	0.035	0.052	0.179	0.115	0.059	0.076	16.32	240.11	377.64	315.91
10,000	81	49	0.050	0.027	0.025	0.039	1.546	0.739	0.544	0.732	23.41	47.64	65.11	59.01
10,000	289	161	0.059	0.035	0.024	0.037	0.602	0.332	0.165	0.221	24.70	106.45	183.25	163.56
10,000	1,089	577	0.062	0.040	0.022	0.035	0.177	0.122	0.048	0.064	24.73	205.29	415.44	342.32

			% True Pos.				% Sign				q	h	ρ
N	R	S	FKRB	ENet	BS	BSU	FKRB	ENet	BS	BSU	BS	BSU	3rd Qu.
1,000	81	49	27.00	90.61	98.27	97.31	52.88	80.98	72.19	78.53	20.27	6.98	0.819
1,000	289	161	8.32	65.21	93.20	87.32	48.07	74.42	83.07	84.43	23.81	6.38	0.822
1,000	1,089	577	2.34	36.74	62.00	51.50	48.00	63.90	78.04	72.58	23.60	6.44	0.824
10,000	81	49	42.51	87.31	99.76	99.04	62.04	86.32	79.81	86.48	28.31	14.69	0.819
10,000	289	161	13.42	60.14	98.18	92.96	50.69	74.46	90.27	91.27	32.81	15.53	0.822
10,000	1,089	577	3.73	33.01	70.33	57.88	48.70	63.15	83.39	76.92	34.70	16.88	0.824

Note: The table reports the average summary statistics over all Monte Carlo replicates for the FKRB estimator (FKRB), for the elastic net estimator with tuning parameter μ from the maximum value of the sequence of tuning parameters suggested by *glmnet* (ENet) and for the random elastic net estimator with $\mu_b = 0\,, b = 1, \ldots B$, for the BS step (BS) and $\mu_u = 0\,, u = 1, \ldots B$, for the BSU step (BSU).

Table 3.A.4: Summary Statistics of 100 Monte Carlo Runs with Mixture of Two Bivariate Normals and Normalized Weights of Random Elastic Net.

			RMISE				Max. Dif.				Pos.			
N	R	S	FKRB	ENet	BS	BSU	FKRB	ENet	BS	BSU	FKRB	ENet	BS	BSU
1,000	100	59	0.095	0.057	0.052	0.049	0.276	0.177	0.169	0.154	13.56	54.02	66.12	60.91
1,000	300	149	0.103	0.062	0.050	0.047	0.285	0.163	0.137	0.120	14.37	117.73	178.54	159.57
1,000	500	203	0.106	0.066	0.051	0.048	0.286	0.168	0.133	0.116	14.77	158.57	265.2	235.71
10,000	100	59	0.058	0.040	0.037	0.030	0.187	0.129	0.127	0.109	19.68	46.15	61.77	54.96
10,000	300	149	0.064	0.040	0.033	0.024	0.194	0.109	0.091	0.070	21.75	99.96	163.58	143.24
10,000	500	203	0.066	0.042	0.035	0.025	0.197	0.113	0.084	0.066	21.77	134.32	241.88	208.57

			% True Pos.				% Sign				q	h	ρ
N	R	S	FKRB	ENet	BS	BSU	FKRB	ENet	BS	BSU	BS	BSU	3rd Qu.
1,000	100	59	19.36	79.08	93.29	90.81	50.28	80.30	84.96	87.25	41.75	5.12	0.824
1,000	300	149	6.66	64.89	94.58	91.27	52.16	75.55	84.77	87.80	40.79	5.18	0.824
1,000	500	203	3.66	57.63	94.23	89.70	59.42	74.48	82.88	85.09	42.29	5.18	0.825
10,000	100	59	28.47	69.95	91.95	86.31	54.92	77.39	87.73	87.88	41.69	12.35	0.823
10,000	300	149	10.52	57.21	93.84	86.77	53.54	73.85	89.02	88.77	39.53	12.53	0.824
10,000	500	203	6.15	51.66	93.28	85.15	60.04	74.48	86.76	86.83	40.40	11.12	0.825

Note: The table reports the average summary statistics over all Monte Carlo replicates for the FKRB estimator (FKRB), for the elastic net estimator with tuning parameter μ from the maximum value of the sequence of tuning parameters suggested by *glmnet* (ENet) and for the random elastic net estimator with 10-fold cross-validation for the BS step (BS) and for the BSU step (BSU) based on the MSE as criterion. The weights in each bootstrap sample have only to be smaller or equal to one, i.e., $\sum_{r\in Q_b}\theta_r \leq 1$ and $\sum_{r\in H_u}\theta_r \leq 1$. At the end of step 1 and step 2, the weights $\hat{\theta}_j^{BS}, j = 1,\ldots,R$, and $\hat{\theta}_j^{BSU}, j = 1,\ldots,R$, are normalized by $\hat{\theta}_j^{BS} = \sum_{b=1}^{B}\hat{\theta}_j^{(b)} / \left(\sum_{j=1}^{R}\sum_{b=1}^{B}\hat{\theta}_j^{(b)}\right)$ and $\hat{\theta}_j^{BSU} = \sum_{u=1}^{B}\hat{\theta}_j^{(u)} / \left(\sum_{j=1}^{R}\sum_{u=1}^{B}\hat{\theta}_j^{(u)}\right)$.

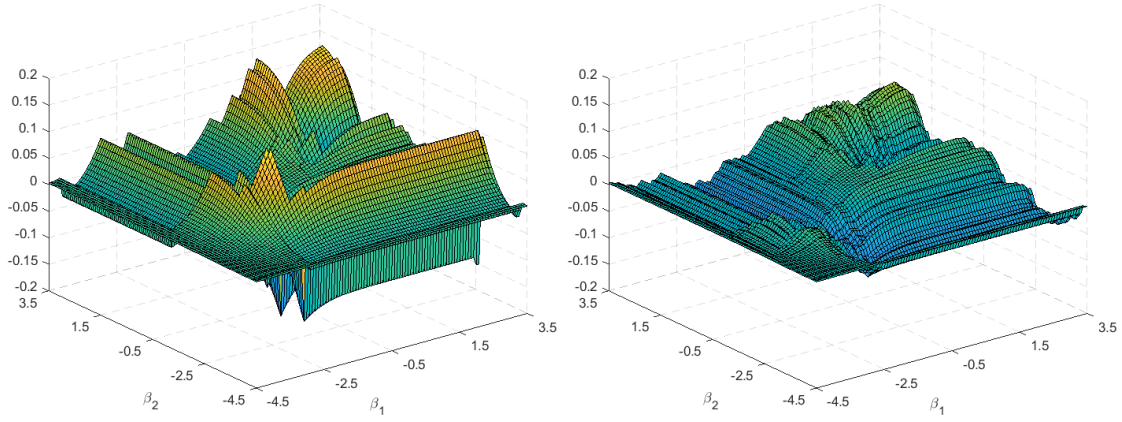
Table 3.A.5: Summary Statistics of 100 Monte Carlo Runs with Mixture of Two Bivariate Normals using Random LASSO instead of Random Elastic Net.

			RMISE				Max. Dif.				Pos.			
N	R	S	FKRB	ENet	BS	BSU	FKRB	ENet	BS	BSU	FKRB	ENet	BS	BSU
1,000	100	59	0.095	0.057	0.052	0.050	0.276	0.177	0.159	0.151	13.56	54.02	65.10	57.09
1,000	300	149	0.103	0.062	0.046	0.045	0.285	0.163	0.124	0.121	14.37	117.73	164.01	140.21
1,000	500	203	0.106	0.066	0.047	0.044	0.286	0.168	0.119	0.114	14.77	158.57	231.04	192.80
10,000	100	59	0.058	0.040	0.033	0.035	0.187	0.129	0.119	0.138	19.68	46.15	62.10	53.22
10,000	300	149	0.064	0.040	0.027	0.029	0.194	0.109	0.078	0.095	21.75	99.96	161.38	133.09
10,000	500	203	0.066	0.042	0.027	0.028	0.197	0.113	0.072	0.084	21.77	134.32	230.92	184.37

			% True Pos.				% Sign				q	h	ρ
N	R	S	FKRB	ENet	BS	BSU	FKRB	ENet	BS	BSU	BS	BSU	3rd Qu.
1,000	100	59	19.36	79.08	89.75	84.49	50.28	80.30	81.80	83.61	27.74	5.51	0.824
1,000	300	149	6.66	64.89	88.40	80.92	52.16	75.55	83.47	83.98	26.45	5.33	0.824
1,000	500	203	3.66	57.63	85.01	75.89	59.42	74.48	82.22	82.46	27.05	5.51	0.825
10,000	100	59	28.47	69.95	90.69	83.00	54.92	77.39	85.92	85.72	37.49	11.90	0.823
10,000	300	149	10.52	57.21	92.23	81.66	53.54	73.85	88.16	87.09	34.67	12.83	0.824
10,000	500	203	6.15	51.66	90.00	78.22	60.04	74.48	86.3	86.04	35.30	12.68	0.825

Note: The table reports the average summary statistics over all Monte Carlo replicates for the FKRB estimator (FKRB), for the elastic net estimator with tuning parameter μ from the maximum value of the sequence of tuning parameters suggested by *glmnet* (ENet) and for the random elastic net estimator with $\mu_b = 0\,, b = 1, \ldots B$, for the BS step (BS) and $\mu_u = 0\,, u = 1, \ldots B$, for the BSU step (BSU).

Figure 3.A.2: Error of True Joint Distribution Function Minus Estimated Joint Distribution Functions for $N = 10,000$ and $R = 500$



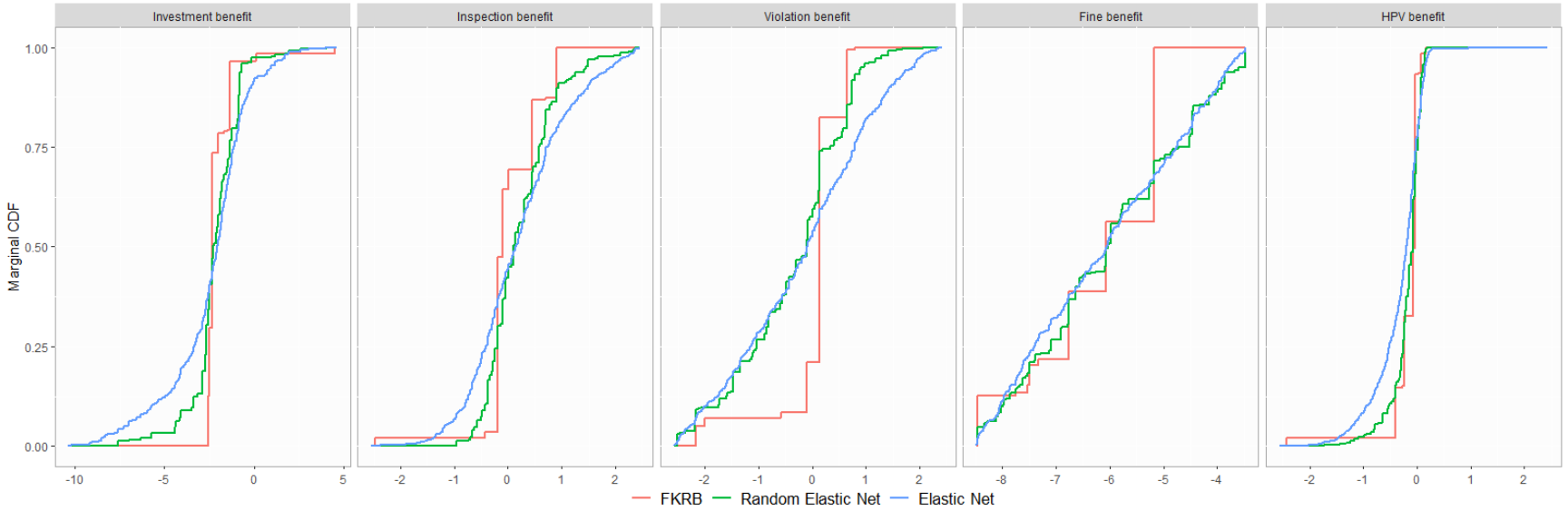
(a) FKRB

(b) Elastic Net with OneSe

(c) BS of Random Elastic Net

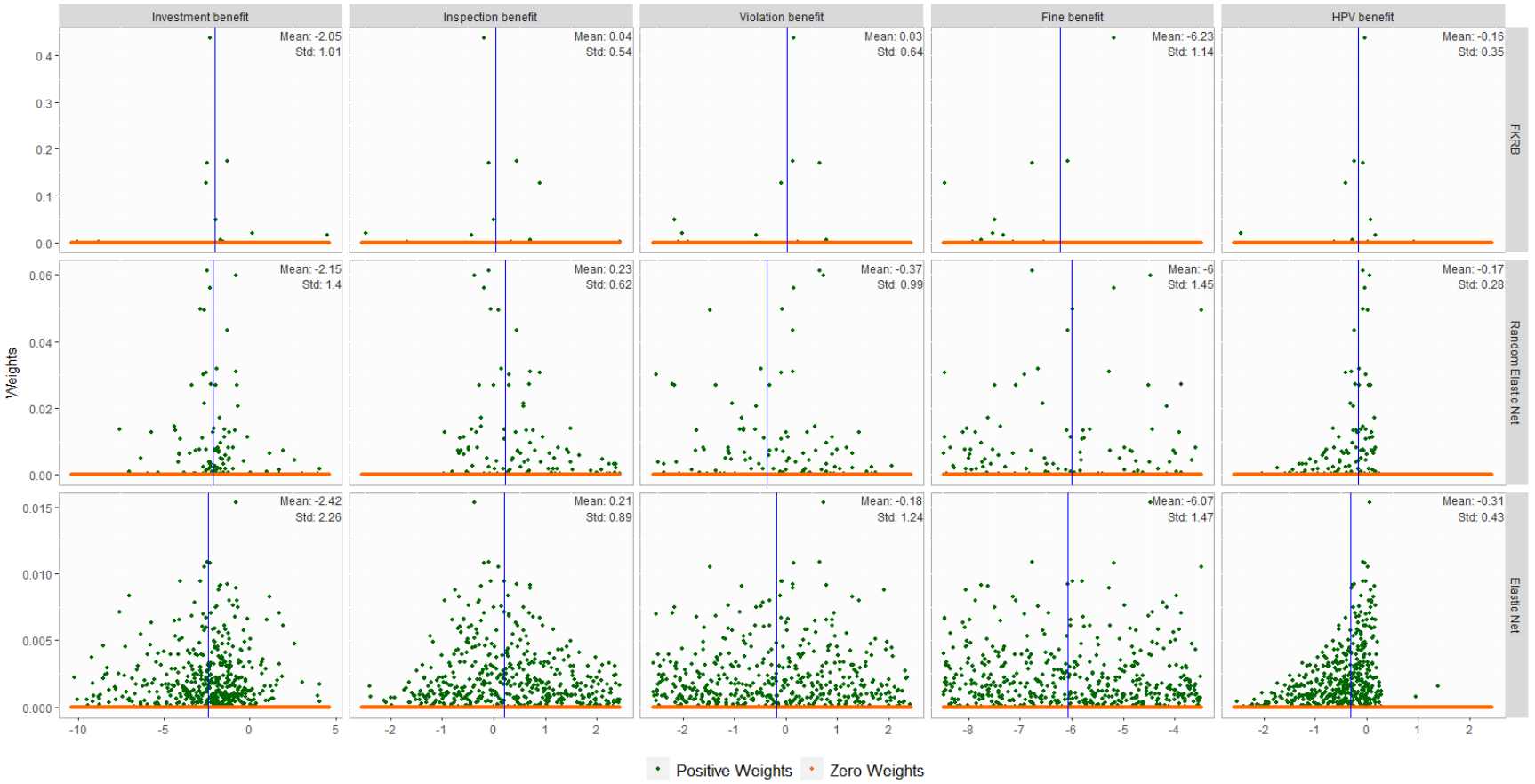
(d) BSU of Random Elastic Net

Figure 3.A.3: Estimated Marginal Distribution Functions for the FKRB, Elastic Net, and Random Elastic Net Estimator.



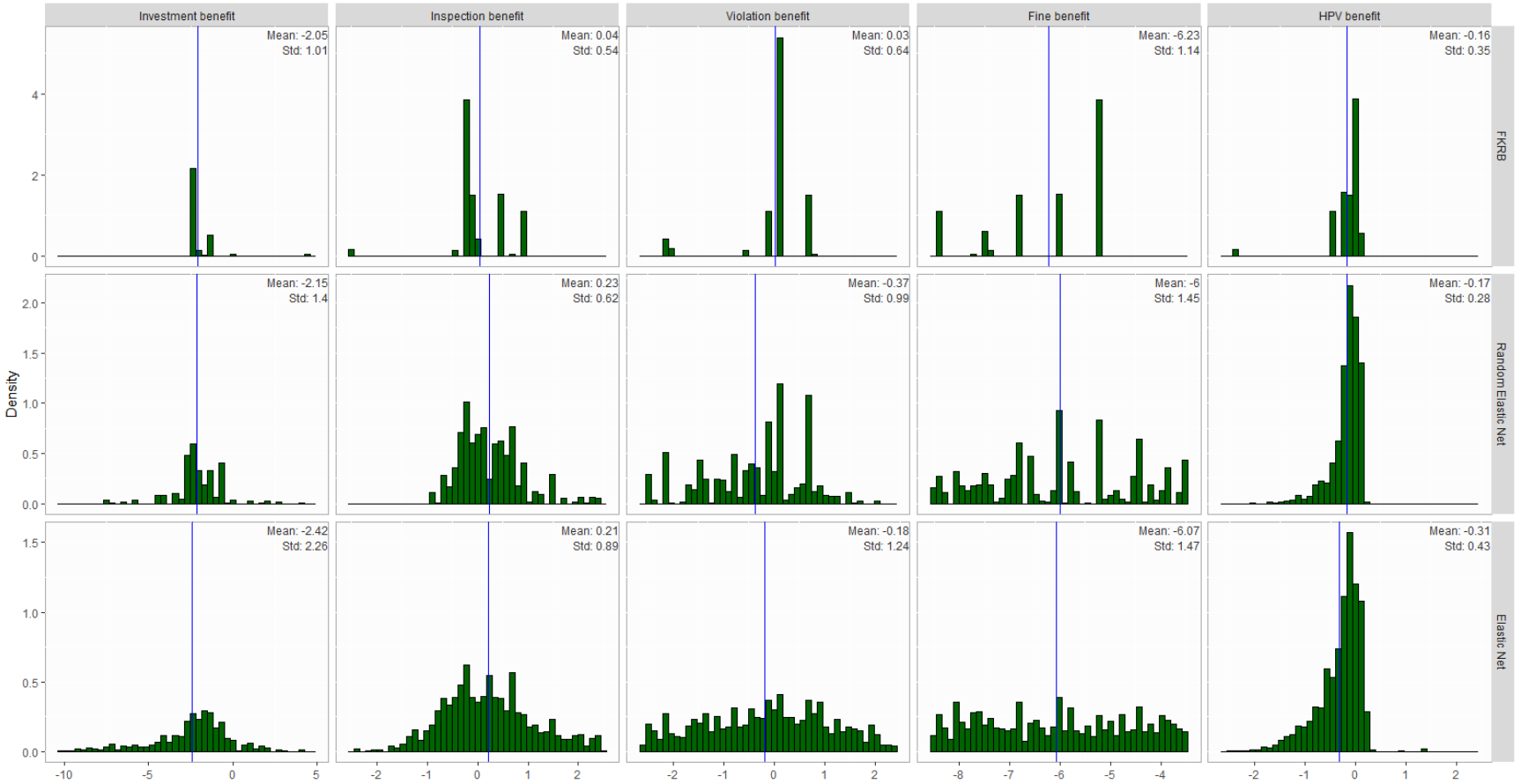
Note: The figure shows the marginal distribution functions estimated with the FKRB estimator, elastic net estimator, and random elastic net estimator. The random elastic net estimator is the BSU estimator. For each estimator, we use the same weighting matrix W as Blundell et al. (2020) in the first step. Subsequently, we update W for each estimator in the same manner as Blundell et al. (2020) do in the second step. The marginal distribution function of the FKRB estimator correspond to the results of Blundell et al. (2020).

Figure 3.A.4: Estimated Mass Points for the FKRB, Elastic Net, and Random Elastic Net Estimator.



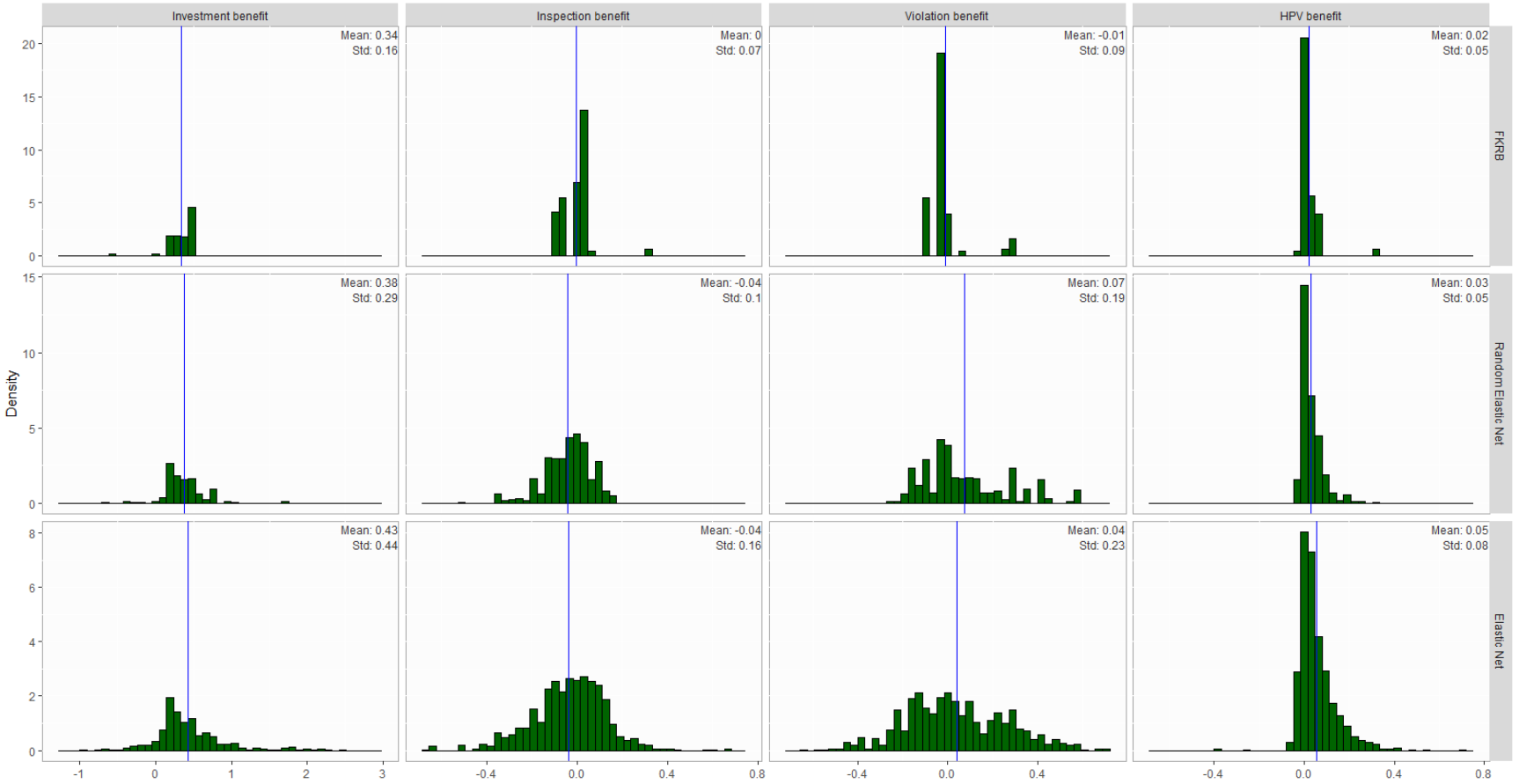
Note: The figure shows the weights at each random coefficient estimated with the FKRB estimator, elastic net estimator, and random elastic net estimator. The random elastic net estimator is the BSU estimator. The vertical blue line is drawn at the weighted mean of each random coefficient. Additionally, we report the weighted mean (Mean) and weighted standard deviation (Std) in the top right corner. For each estimator, we use the same weighting matrix W as Blundell et al. (2020) in the first step. Subsequently, we update W for each estimator in the same manner as Blundell et al. (2020) do in the second step. A weight is defined to be positive if it is greater than 10^{-5}. The weights of the FKRB estimator correspond to those given in Blundell et al. (2020).

Figure 3.A.5: Estimated Histograms for the FKRB, Elastic Net, and Random Elastic Net Estimator.



Note: The figure shows histograms (green bars) for the weights in Figure 3.A.4 at each random coefficient. The weights are estimated with the FKRB estimator, elastic net estimator, and random elastic net estimator. The random elastic net estimator is the BSU estimator. The vertical blue line is drawn at the weighted mean of each random coefficient. Additionally, we report the weighted mean (Mean) and weighted standard deviation (Std) in the top right corner. For each estimator, we use the same weighting matrix W as Blundell et al. (2020) in the first step. Subsequently, we update W for each estimator in the same manner as Blundell et al. (2020) do in the second step. The histograms of the FKRB estimator correspond to the resulst of Blundell et al. (2020).

Figure 3.A.6: Estimated Histograms with Random Coefficients in Fine Equivalent for the FKRB, Elastic Net, and Random Elastic Net Estimator.



Note: The figure shows histograms (green bars) for the weights in Figure 3.A.4 at the equivalent of each random coefficient to a $1 million fine per quarter. The weights are estimated with the FKRB estimator, elastic net estimator, and random elastic net estimator. The random elastic net estimator is the BSU estimator. To calculate the equivalent of each random coefficient to a $1 million fine per quarter, we divide each random coefficient by the the random coefficient of fine, i.e., β^i/β^F, where $i \in \{X, I, V, H\}$. The vertical blue line is drawn at the weighted mean of each random coefficient. Additionally, we report the weighted mean (Mean) and weighted standard deviation (Std) in the top right corner. For each estimator, we use the same weighting matrix W as Blundell et al. (2020) in the first step. Subsequently, we update W for each estimator in the same manner as Blundell et al. (2020) do in the second step. The histograms of the FKRB estimator correspond to the resulst of Blundell et al. (2020).

Table 3.A.6: Counterfactual Results Including the FKRB, Elastic Net, and the BS and BSU step of the Random Elastic Net estimator.

		Baseline				Same fines for all violators; fines constant			
	Data	FKRB	ENet	BS	BSU	FKRB	ENet	BS	BSU
Compliance (%)	95.62	95.11	95.17	95.23	95.17	66.72	57.14	59.65	65.16
Regular violator (%)	2.88	3.47	3.48	3.49	3.45	2.53	2.05	2.15	2.37
HPV (%)	1.50	1.42	1.35	1.28	1.37	30.75	40.82	38.20	32.47
Investment rate (%)	0.40	0.54	0.54	0.54	0.55	0.47	0.43	0.44	0.46
Inspection rate (%)	9.65	9.41	9.40	9.37	9.39	20.54	24.06	23.09	21.05
Fines (thousands $)	0.18	0.32	0.32	0.30	0.31	0.32	0.32	0.30	0.31
Violations (%)	0.55	0.54	0.54	0.53	0.53	5.00	6.05	5.68	4.85
Plant utility	—	0.006	0.020	0.016	0.022	0.077	0.171	0.147	0.117
Pollution damages (mil. $)	1.65	1.53	1.51	1.51	1.52	4.04	4.84	4.63	4.18

		Same fines for all violators; pollution damages constant				Fines for HPVs doubled relative to baseline			
	Data	FKRB	ENet	BS	BSU	FKRB	ENet	BS	BSU
Compliance (%)	95.62	94.49	94.73	94.85	94.93	95.52	95.57	95.59	95.57
Regular violator (%)	2.88	2.72	1.91	1.97	2.06	3.47	3.49	3.49	3.46
HPV (%)	1.50	2.79	3.36	3.18	3.01	1.01	0.95	0.92	0.97
Investment rate (%)	0.40	0.65	0.76	0.75	0.75	0.55	0.54	0.54	0.55
Inspection rate (%)	9.65	9.88	10.03	9.97	9.93	9.28	9.26	9.25	9.26
Fines (thousands $)	0.18	1.98	4.90	4.61	3.70	0.36	0.36	0.34	0.35
Violations (%)	0.55	0.74	0.80	0.78	0.76	0.49	0.48	0.48	0.48
Plant utility	—	0.001	0.006	0.000	0.010	0.005	0.019	0.015	0.021
Pollution damages (mil. $)	1.65	1.53	1.51	1.51	1.52	1.48	1.47	1.47	1.48

Note: The table reports summary statistics of the values in the data, the values predicted by the model estimates (baseline), and three counterfactuals for the FKRB, elastic net, the BS step and the BSU step of the random elastic net estimator. Each summary statistic is per plant / quarter and calculated in the same way as in Blundell et al. (2020). The counterfactuals of FKRB correspond to those given in Blundell et al. (2020).

Figure 3.A.7: Estimated Heat Maps for Blundell et al. (2020).

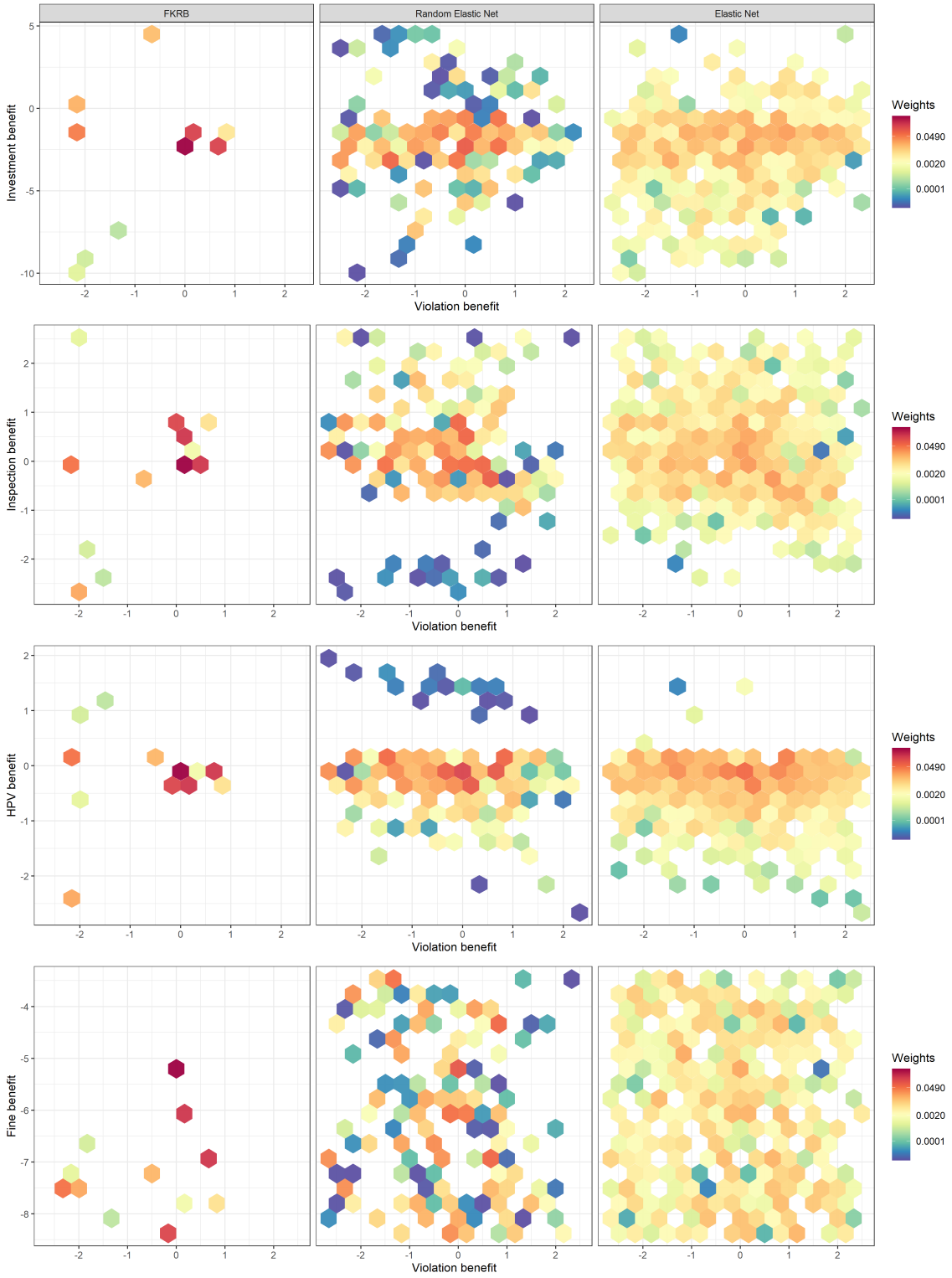


Figure 3.A.8: Estimated Heat Maps for Blundell et al. (2020).

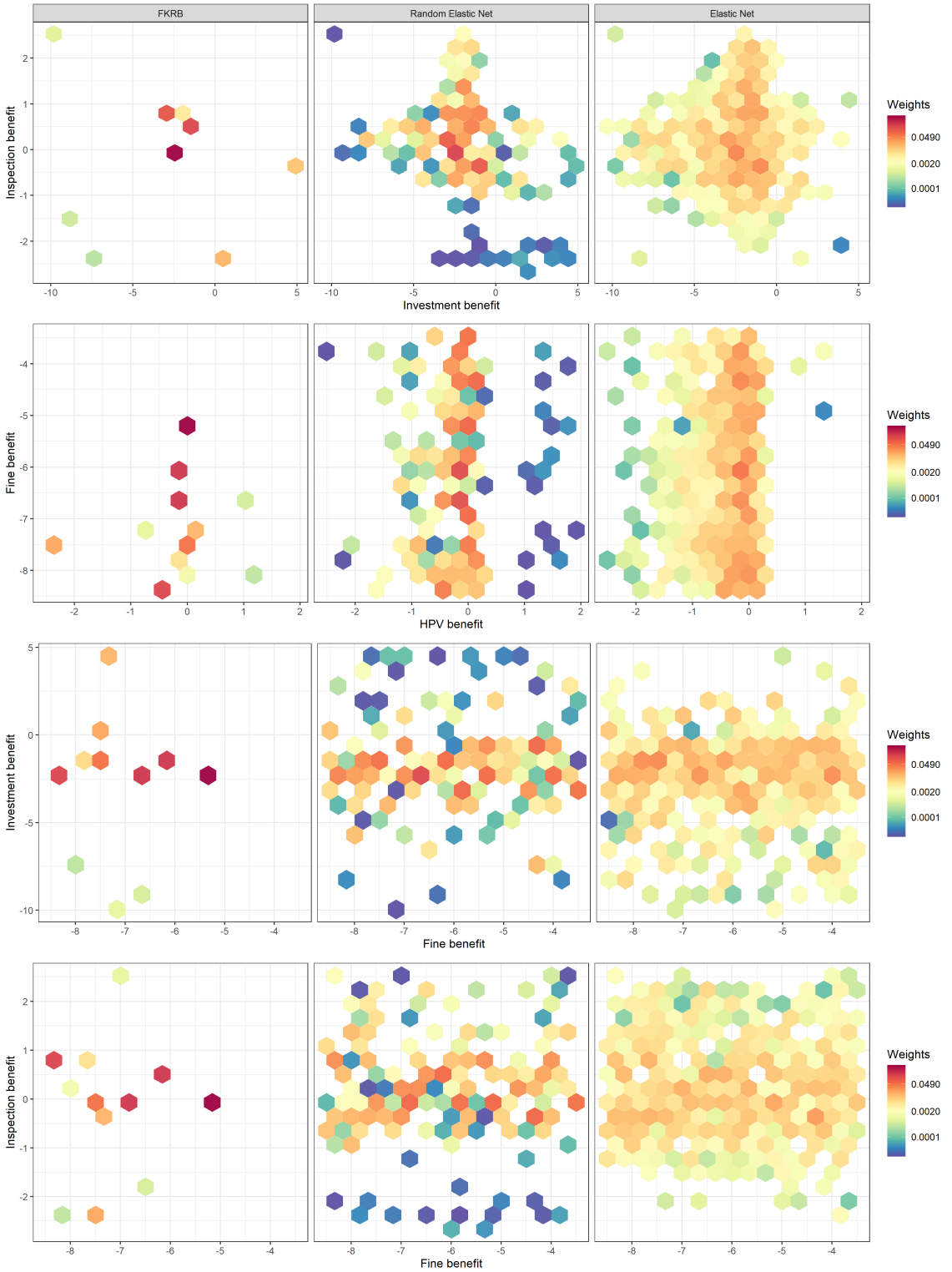
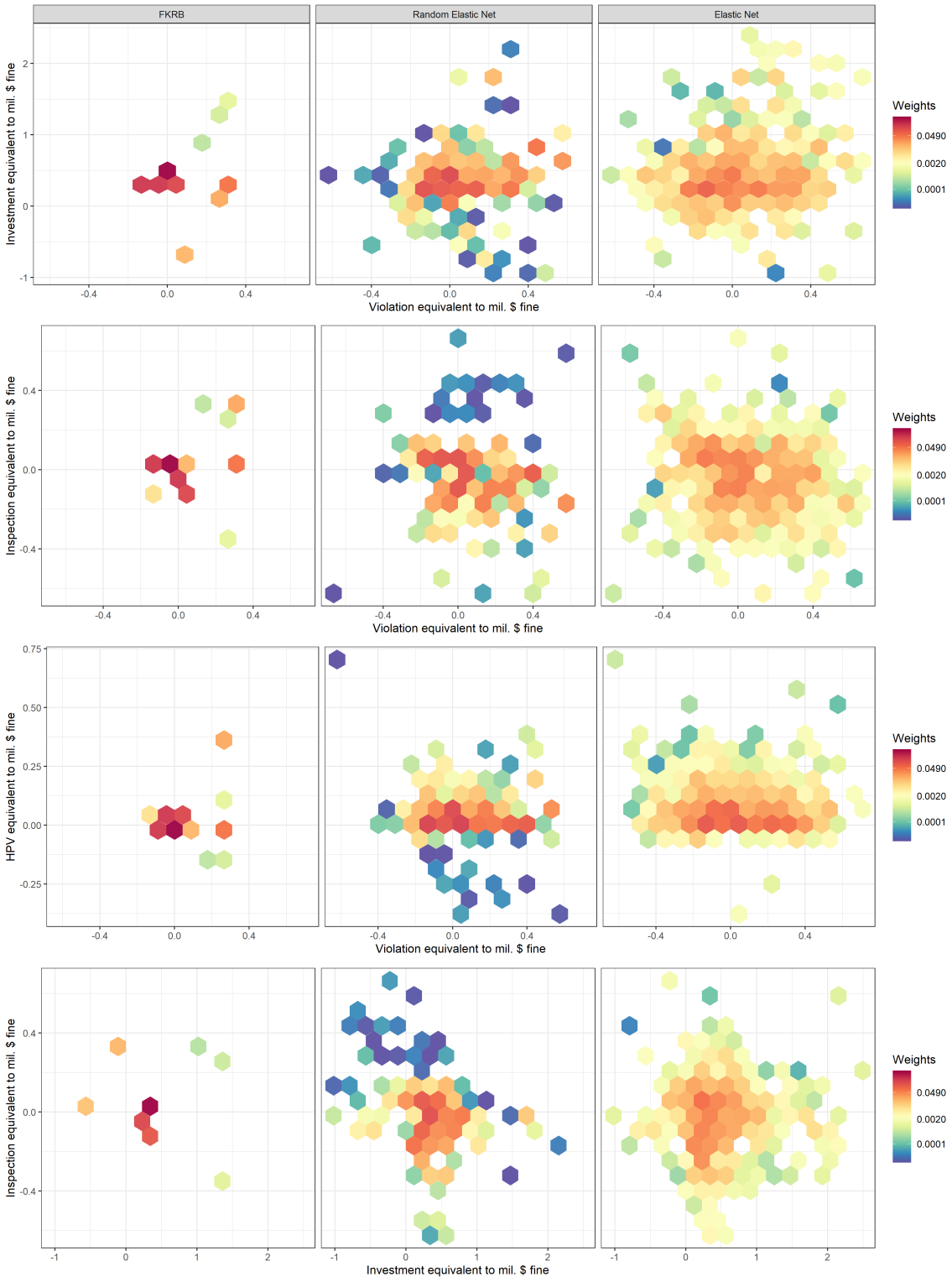


Figure 3.A.9: Estimated Heat Maps with Random Coefficients in Fine Equivalent for Blundell et al. (2020).



4 Deep Learning for the Estimation of Heterogeneous Parameters in Discrete Choice Models

Co-authored by Maximilian Osterhaus

Abstract

This paper studies the finite sample performance of the flexible estimation approach of Farrell et al. (2021*a*), who propose to use deep learning for the estimation of heterogeneous parameters in economic models, in the context of discrete choice models. The approach combines the structure imposed by economic models with the flexibility of deep learning, which assures the interpretebility of results on the one hand, and allows estimating flexible functional forms of observed heterogeneity on the other hand. For inference after the estimation with deep learning, Farrell et al. (2021*a*) derive an influence function that can be applied to many quantities of interest. Focusing on discrete choice models, we conduct a series of Monte Carlo experiments that investigate the impact of regularization on the proposed estimation and inference procedure. The results of these experiments show that deep learning for the estimation of heterogeneous parameters generally leads to precise estimates of the true average parameters and that regular robust standard errors lead to invalid inference results. Without regularization, the influence function approach can lead to substantial bias and large estimated standard errors caused by extreme outliers. Regularization reduces this property and stabilizes the estimation procedure, but at the expense of inducing an additional bias. The bias in combination with decreasing variance associated with increasing regularization leads to the construction of invalid inferential statements in our experiments. Repeated sample splitting, unlike regularization, stabilizes the estimation approach without introducing an additional bias, thereby allowing for the construction of valid inferential statements.

JEL codes:	*C14, C25, C45*
Keywords:	*Deep Learning, Conditional Logit Model, Observed Heterogeneity, Inference.*
Publication status:	*To be submitted.*

4.1 Introduction

Appropriately modeling heterogeneity across economic agents is a key challenge in many empirical economic studies. Often, the heterogeneity can be linked to observed characteristics of agents. This is typically achieved using parametric specifications in the form of linear interactions of only few observed characteristics with the variables of interest. Even restrictive functional forms like linear functions rapidly lead to a large number of parameters, especially if the heterogeneity is modeled as a function of multiple characteristics (Cranenburgh, Wang, Vij, Pereira and Walker, 2021). Furthermore, limiting the heterogeneity to linear functions of only few characteristics can lead to misspecification of the true shape and extent of heterogeneity, and to potentially incorrect results for quantities of interest, such as elasticities or willingness-to-pay measures.

The increasing availability of large data sets makes it possible to reduce the reliance on parametric methods and to apply more flexible approaches to study heterogeneity. A promising tool for this task is deep learning, which is known for its ability to flexibly model functional forms and to handle large amounts of data. While deep learning so far has been applied with great success for pure prediction tasks (LeCun, Bengio and Hinton, 2015), Farrell et al. (2021*a*) propose to employ deep learning for the estimation of heterogeneous parameters. They incorporate the heterogeneity across economic agents into the economic model specified by the researcher through coefficients that are functions of agents' observed characteristics. The approach combines parametric approaches – which impose structure on the model grounded in economic principles and reasoning – with deep learning – which lets the data speak for itself with its flexibility.

To derive theoretically valid inferential statements after estimating the coefficient functions with deep learning, Farrell et al. (2021*a*) extend the deep learning theory for generic regression approaches developed by Farrell, Liang and Misra (2021*b*) to M-estimators. Building on Chernozhukov, Chetverikov, Demirer, Duflo, Hansen, Newey and Robins (2018), they derive an influence function that makes inference feasible in a wide range of settings – the provided inferential statements cover any parameter of interest that is a function of the heterogeneous coefficient functions. Farrell et al. (2021*a*) show that the inference procedure allows to construct valid inferential statements under fairly weak conditions. However, they leave the role of regularization and its consequences for estimation and subsequent inference for future research.

Conducting a series of Monte Carlo experiments, we intend to fill this gap and study the finite sample properties of the proposed inference procedure in the context of discrete choice models. The results of these experiments show that deep learning generally is well suited for the estimation of heterogeneous parameters, especially if the sample size is sufficiently large, and that naive inference after estimating the parameters with deep learning leads to invalid inference. Further, the proposed estimation procedure is sensitive to overfitting when no regularization is used. We observe that estimation without regularization can results in substantial bias and large estimated standard errors. The sensitivity to overfitting is more pronounced in small samples but does not completely

disappear with increasing sample size. Regularization in form of l_2-penalties on the weights tuned in the network reduces the sensitivity to overfitting and rapidly decreases the average estimated standard errors. However, it also appears to introduce a new source of bias which in combination with the decreasing variance explains the poor coverage of the estimated confidence intervals observed in our experiments. Finally, the experiments show that substantially better results are obtained when repeated sample splitting is used. Unlike regularization, repeated sample splitting substantially reduces the bias arising from overfitting without inducing a new bias, this way leading to valid inferential results in out experiments.

Our paper contributes to a growing literature on the combination of deep learning and structural modeling in discrete choice models.[1] Among others, Sifringer, Lurkin and Alahi (2020) and Wong and Farooq (2021) apply deep learning to estimate demand for travel modes in a logit framework. To avoid model misspecification in discrete choice models, Sifringer et al. (2020) propose to decompose the utility into two parts: a knowledge-driven part which includes the variables of interest and is specified by the researcher, and a data-driven part, which is estimated with deep learning using the remaining explanatory variables that are not of primary interest. Separating those two parts of the utility assures that the parameters of interest can be interpreted. However, as the knowledge-driven part needs to be fully specified, its coefficients are constant across agents. Therefore, this approach seems more restrictive than the approach of Farrell et al. (2021*a*) which allows for heterogeneous coefficients. In contrast, Wong and Farooq (2021) allow for a knowledge-driven part of the utility and an additional random component of the utility which can depend on the characteristics of all alternatives. That is, their approach captures unobserved heterogeneity and cross-effects of non-linear utilities across all alternatives. Thus, their model relaxes the IIA property. Both have in common that they do not provide a theoretically valid inference procedure for parameters of interest but rely on approximations of the confidence intervals based on the Hessian of the estimated model, which are not guaranteed to have the correct size. Wang, Wang and Zhao (2020) focus on estimating economic quantities of interest, e.g., market shares, elasticities and changes in social welfare, with deep learning using a completely unstructured utility. Similarly to Sifringer et al. (2020) and Wong and Farooq (2021), they do not present a valid approach for inference on the quantities of interest.[2] They rely on the predicted choice probabilities and the gradient of the estimated model and do not take into account that the considered quantities are accompanied with additional uncertainty when no structure is imposed on the utility.

The remainder of this paper is organized as follows. Section 4.2 illustrates how deep learning can be employed to estimate heterogeneous parameters in economic models and outlines the inference and estimation procedure. Section 4.3 presents Monte Carlo

[1] For recent surveys of the application of machine learning and deep learning for the estimation of discrete choice models, see, e.g., Karlaftis and Vlahogianni (2011), Wang, Mo, Hess and Zhao (2021), and Cranenburgh et al. (2021).

[2] For example, they calculate the standard deviation of the average elasticity as the standard deviation of the elasticity of each individual.

experiments that study the inference procedure and Section 4.4 applies the influence function approach to real data. Section 4.5 concludes.

4.2 Deep Learning for Heterogeneity

This section introduces the methodical framework of Farrell et al. (2021*a*) who propose to estimate heterogeneous parameters in econometric models using deep learning in the form of multi-layer feed-forward neural networks. The flexibility of deep neural networks (DNNs) makes them ideally suited for the estimation of economic models with individual heterogeneity. Subsection 4.2.1 explains the design of the network which directly integrates the economic model specified by the researcher into the network architecture. Subsection 4.2.2 explains the inference approach which is based on the concept of influence functions, and Subsection 4.2.3 lays out the estimation procedure. While the estimation and inference procedure is applicable to a wide range of models, we focus on multinomial discrete choice models when introducing the estimation procedure.

4.2.1 Deep Learning

Starting point of the estimation approach is the economic model specified by the researcher. The model relates the outcome $\boldsymbol{Y}$ to the variables of interest $\boldsymbol{X}$, and to socio-demographic characteristics $\boldsymbol{W}$ that are included to capture the heterogeneity across individuals.[3] We are interested in analyzing consumers' preferences. For that purpose, we consider a conditional logit model to model individuals' choices over a set of J mutually exclusive alternatives. In this context, let $\boldsymbol{x}_{i,j}$ denote a K-dimensional real-valued vector of observed product characteristics for consumer $i = 1, \ldots, N$ and alternative $j = 1, \ldots, J$, $\boldsymbol{w}_i$ a D-dimensional vector of observed socio-demographics of consumer i, and $\boldsymbol{y}_i$ a J-dimensional vector with entry 1 if alternative j is chosen by consumer i and zero otherwise. Consumers choose the alternative that maximizes their utility. Given the unobserved individual parameters $\alpha_j(\boldsymbol{w}_i)$, $j = 1, \ldots, J$, and $\boldsymbol{\beta}(\boldsymbol{w}_i) = (\beta_1(\boldsymbol{w}_i), \ldots, \beta_K(\boldsymbol{w}_i))'$ consumer i realizes utility $u_{i,j} = \alpha_j(\boldsymbol{w}_i) + \boldsymbol{x}_{i,j}'\boldsymbol{\beta}(\boldsymbol{w}_i) + \omega_{i,j}$ from alternative j, where $\omega_{i,j}$ denotes an idiosyncratic, consumer- and choice-specific error term. Thus, consumer i chooses alternative j if $u_{i,j} > u_{i,l}$ for all $j \neq l$. Under the assumption that $\omega_{i,j}$ is independently and identically distributed type I extreme value, the probability that consumer i chooses alternative j conditional on the observed product characteristics and socio-demographics is

$$\mathbb{P}\left(y_{i,j} = 1 | \boldsymbol{x}_i, \boldsymbol{w}_i\right) = \frac{\exp\left(\alpha_j\left(\boldsymbol{w}_i\right) + \boldsymbol{x}_{i,j}'\boldsymbol{\beta}\left(\boldsymbol{w}_i\right)\right)}{\sum_{m=1}^{J} \exp\left(\alpha_m\left(\boldsymbol{w}_i\right) + \boldsymbol{x}_{i,m}'\boldsymbol{\beta}\left(\boldsymbol{w}_i\right)\right)}. \tag{4.1}$$

[3] *Notation:* The variables written in capital letters denote random variables and small letters observational units. All vectors and matrices are written in bold.

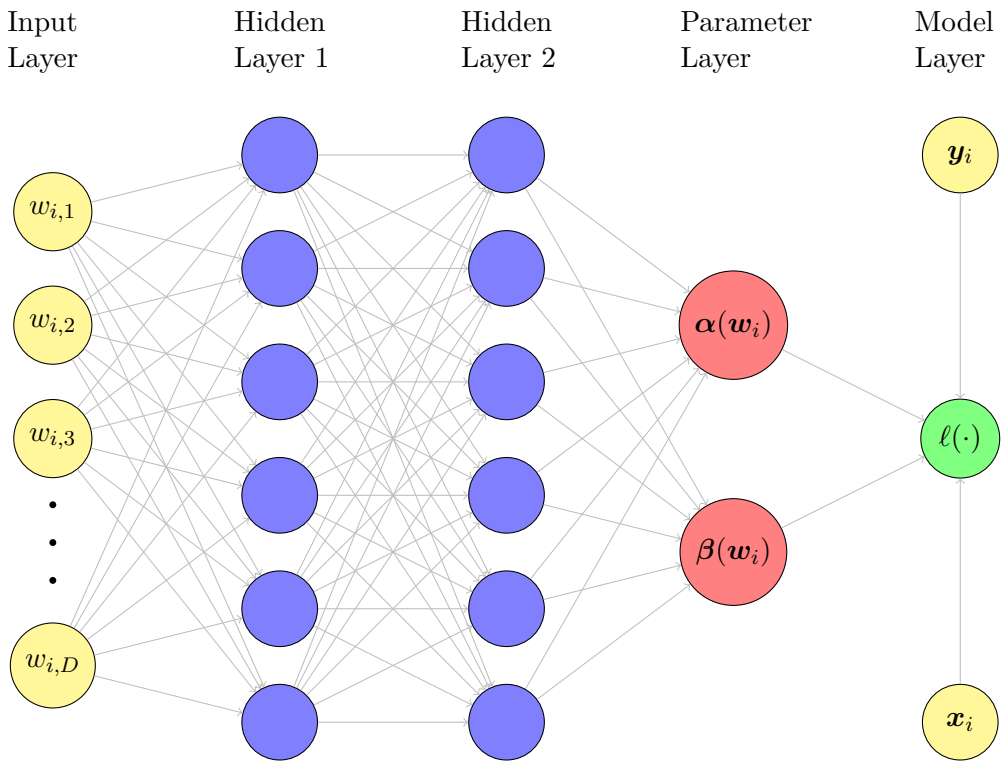
The goal of the researcher is to estimate the unknown heterogeneous coefficient functions $\boldsymbol{\alpha}(\boldsymbol{w}_i) = (\alpha_1(\boldsymbol{w}_i), \ldots, \alpha_J(\boldsymbol{w}_i))'$ and $\boldsymbol{\beta}(\boldsymbol{w}_i)$, which are functions of consumers' socio-demographic characteristics that capture the observed heterogeneity across consumers. Thus, the functions capture no unobserved heterogeneity, i.e., there are no random coefficients.[4]

For the estimation of $\boldsymbol{\alpha}(\cdot)$ and $\boldsymbol{\beta}(\cdot)$, Farrell et al. (2021*a*) advocate deep neural networks. The proposed network architecture allows to combine a standard fully-connected feedforward neural network – which is used to estimate the coefficient functions $\boldsymbol{\alpha}(\cdot)$ and $\boldsymbol{\beta}(\cdot)$ – with the economic structure imposed by the conditional logit model. The key idea of the network architecture is to be fully flexible in modeling the individual heterogeneity while retaining the structure which assures the interpretability of the results. Figure 4.1 illustrates such an architecture. Given consumers' observed socio-demographics, $\boldsymbol{w}_i$, $i = 1, \ldots, N$, in the input layer, the feedforward network learns the coefficient functions $\boldsymbol{\alpha}(\cdot)$ and $\boldsymbol{\beta}(\cdot)$ using two hidden layers, a parameter layer, and a model layer. The first part of the network, the input layer and the hidden layers, corresponds to the structure of a standard feedforward neural network. The number of hidden layers and the number of units per hidden layer determine the flexibility of the approach regarding the shape of the estimated coefficient functions. The coefficient functions $\boldsymbol{\alpha}(\cdot)$ and $\boldsymbol{\beta}(\cdot)$ returned in the parameter layer are then forwarded to the model layer, where they are combined with the variables of interest, $\boldsymbol{x}_i$, and the observed choices, $\boldsymbol{y}_i$, to minimize the individual loss function, $\ell\left(\boldsymbol{y}_i, \boldsymbol{x}_i, \boldsymbol{\alpha}\left(\boldsymbol{w}_i\right), \boldsymbol{\beta}\left(\boldsymbol{w}_i\right)\right)$. To be clear, the variables of interest, $\boldsymbol{x}_i$, are additional inputs provided only to the model layer but are not used as inputs to the coefficient functions $\boldsymbol{\alpha}(\cdot)$ and $\boldsymbol{\beta}(\cdot)$. The novelty of this network architecture is the model layer, which ensures that the coefficient functions $\boldsymbol{\alpha}(\cdot)$ and $\boldsymbol{\beta}(\cdot)$ are learned within the structure imposed by the specified model. This way, the estimated results have an economically meaningful interpretation, which typically is not the case for regular machine learning applications in economics (Farrell et al., 2021*a*).

The number of hidden layers (the depth of the network), and the number of units per layer (the width of each layer) are specified by the researcher. According to the universal approximation theorem (Hornik, Stinchcombe and White, 1989 and Cybenko, 1989), a feedforward network with only one hidden layer might be already sufficient to represent any function if the number of hidden units is sufficiently large. Networks with multiple hidden layers typically require less units per hidden layer – and hence total parameters – to represent the desired function, and in many circumstances generalize well in terms of out-of-sample performance. However, such networks tend to be harder to optimize (Goodfellow, Bengio and Courville, 2016). In Theorem 1, Farrell et al. (2021*a*) derive error bounds for the estimated coefficient functions $\hat{\boldsymbol{\alpha}}(\cdot)$ and $\hat{\boldsymbol{\beta}}(\cdot)$, where they allow the depth of the network to increase with the sample size, and the width of the network with the sample size and the number of continuous input variables, respectively. Beyond

[4]The parameters $\boldsymbol{\beta}(\boldsymbol{w}_i)$ and $\boldsymbol{\alpha}(\boldsymbol{w}_i)$ can be considered as the best approximations to some unobserved individual parameters $\boldsymbol{\alpha}_i$ and $\boldsymbol{\beta}_i$ that lie in an assumed function class.

Figure 4.1: Feedforward Neural Network for the Estimation of the Heterogeneous Parameters $\boldsymbol{\alpha}(\boldsymbol{w}_i)$ and $\boldsymbol{\beta}(\boldsymbol{w}_i)$



the number of hidden layers and units, the researcher needs to specify the activation function at every layer. The design of hidden layers is an active area of research which does not provide definite guidelines for the choice of activation functions yet. According to Goodfellow et al. (2016), rectified linear units are an excellent default choice, which are also recommended by Farrell et al. (2021*a*). Overall, specifying the network architecture is a trial-and-error process where the final architecture can be selected based on the best out-of-sample fit (Goodfellow et al., 2016).

When estimating the model, the coefficient functions $\boldsymbol{\alpha}(\boldsymbol{w}_i)$ and $\boldsymbol{\beta}(\boldsymbol{w}_i)$ are learned jointly. To simplify the notation, we write $\boldsymbol{\delta}(\boldsymbol{w}_i) := (\boldsymbol{\alpha}(\boldsymbol{w}_i)', \boldsymbol{\beta}(\boldsymbol{w}_i)')'$ and $L := J + K$ in the following. In our case, the individual loss function, $\ell\left(\boldsymbol{y}_i, \boldsymbol{x}_i, \boldsymbol{\delta}(\boldsymbol{w}_i)\right)$, following from the economic model of interest, is the empirical log-likelihood for individual i,

$$\ell\left(\boldsymbol{y}_i, \boldsymbol{x}_i, \boldsymbol{\delta}(\boldsymbol{w}_i)\right) = \sum_{j=1}^{J} y_{i,j} \log\left(\mathbb{P}\left(y_{i,j} = 1 | \boldsymbol{x}_i, \boldsymbol{w}_i\right)\right),$$

where $\mathbb{P}\left(y_{i,j} = 1 | \boldsymbol{x}_i, \boldsymbol{w}_i\right)$ is the conditional logit choice probability given in Equation (4.1).

Then, $\hat{\boldsymbol{\delta}}(\boldsymbol{w}_i) := (\hat{\boldsymbol{\alpha}}(\boldsymbol{w}_i)', \hat{\boldsymbol{\beta}}(\boldsymbol{w}_i)')'$ are determined such that they simultaneously maximize the log-likelihood

$$\hat{\boldsymbol{\delta}}(\boldsymbol{w}_i) = \arg\max_{\boldsymbol{\delta}} \sum_{i=1}^{N} \ell\left(\boldsymbol{y}_i, \boldsymbol{x}_i, \boldsymbol{\delta}(\boldsymbol{w}_i)\right),$$

where we optimize over the class of DNNs which use the type of architecture described in Figure 4.1. The log-likelihood loss function forces the DNN to learn the coefficient functions within the structure imposed by the conditional logit model. This has two advantages in comparison to naively applied prediction-focused machine learning methods, which predict the choice probabilities $\hat{\mathbb{P}}\left(y_{i,j} = 1 | \boldsymbol{x}_i, \boldsymbol{w}_i\right)$ using a completely unstructured nonparametric utility $\hat{u}(\boldsymbol{y}_i, \boldsymbol{w}_i, \boldsymbol{x}_i)$: First, it assures that the network provides economically meaningful results. For the unstructured approach, in contrast, it is not clear how estimates of $\boldsymbol{\alpha}(\boldsymbol{w}_i)$ and $\boldsymbol{\beta}(\boldsymbol{w}_i)$ can be separately recovered from $\hat{u}(\boldsymbol{y}_i, \boldsymbol{w}_i, \boldsymbol{x}_i)$, which, however, is often necessary for interpretation. And second, even if $\boldsymbol{\alpha}(\boldsymbol{w}_i)$ and $\boldsymbol{\beta}(\boldsymbol{w}_i)$ could be separately recovered in the unstructured approach, Farrell et al. (2021*a*) show that the additional structure of the model enables a faster rate of convergence for the estimated coefficient functions (given the model is correctly specified). For the structured approach, the rate of convergence only depends on the dimension of the socio-demographic characteristics, $\dim(\boldsymbol{w}_i)$, whereas for the naive prediction focused machine learning with unstructured $\hat{u}(\boldsymbol{y}_i, \boldsymbol{w}_i, \boldsymbol{x}_i)$, it depends on both the dimension of the socio-demographic characteristics and the dimension of the variables of interest, i.e., $\dim(\boldsymbol{w}_i) + \dim(\boldsymbol{x}_i)$. While the convergence rate in the structured network is fast enough for inference, the convergence rate of the unstructured model would often be too slow for inference (Farrell et al., 2021*a*).

4.2.2 Inference

Inference for machine learning methods for the estimation of economic models is challenging. For that reason, Farrell et al. (2021*a*) adopt the semiparametric inference procedure suggested by Chernozhukov et al. (2018) which allows to perform inference on expected values of heterogeneous quantities using an influence function approach. Due to the structure imposed by the economic model, the proposed procedure can be applied to any quantity of interest (e.g., expected value of coefficients, elasticticties, or measures for the willingness-to-pay) which are functions of the heterogeneous coefficient functions $\boldsymbol{\delta}(\cdot)$ (and a fixed vector $\boldsymbol{x}^*$ containing arbitrary values of the variables of interest).

Let the real-valued function $H(\cdot)$ specified by the researcher denote the function of interest. Then, the inference procedure described in the following allows to conduct inference on the expected value of $H(\cdot)$ given some $\boldsymbol{x}^*$,

$$\theta_0 = \mathbb{E}\left[H\left(\boldsymbol{W}, \boldsymbol{\delta}\left(\boldsymbol{W}\right); \boldsymbol{x}^*\right)\right].$$

Note that $H(\cdot)$ directly depends on the coefficient functions $\boldsymbol{\delta}(\cdot)$, making inference on θ_0 depend on how well $\hat{\boldsymbol{\delta}}(\cdot)$ approximates its true counterpart $\boldsymbol{\delta}(\cdot)$. Because the empirical

plug-in estimator of θ_0,

$$\hat{\theta}_{PI} = \frac{1}{N}\sum_{i=1}^{N} H\left(\boldsymbol{w}_i, \hat{\boldsymbol{\delta}}\left(\boldsymbol{w}_i\right); \boldsymbol{x}^*\right),$$

is only valid under strong conditions on $\hat{\boldsymbol{\delta}}(\cdot)$, which are unlikely to be satisfied if the functions are estimated with deep-neural networks, Farrell et al. (2021*a*) propose to use the concept of influence functions for inference. The approach builds on the seminal work of Newey (1994) and has the advantage that it provides results for valid inference under less restrictive conditions on the distributional approximations of $\boldsymbol{\delta}(\cdot)$. These assumptions are known to hold for many machine learning methods (Farrell et al., 2021*a*).

The influence function for θ_0 involves the gradient and Hessian corresponding to the loss function $\ell\left(\boldsymbol{y}_i, \boldsymbol{x}_i, \boldsymbol{\delta}(\boldsymbol{w}_i)\right)$ with respect to $\boldsymbol{\delta}(\boldsymbol{w}_i)$. Let $\boldsymbol{\ell}_{\boldsymbol{\delta}}\left(\boldsymbol{y}_i, \boldsymbol{x}_i, \boldsymbol{\delta}(\boldsymbol{w}_i)\right)$ denote the L-dimensional vector of first derivatives of $\boldsymbol{\ell}\left(\boldsymbol{y}_i, \boldsymbol{x}_i, \boldsymbol{\delta}(\boldsymbol{w}_i)\right)$ w.r.t. $\boldsymbol{\delta}(\boldsymbol{w}_i)$,

$$\boldsymbol{\ell}_{\boldsymbol{\delta}}\left(\boldsymbol{y}_i, \boldsymbol{x}_i, \boldsymbol{\delta}(\boldsymbol{w}_i)\right) = \left.\frac{\partial \ell\left(\boldsymbol{y}_i, \boldsymbol{x}_i, \boldsymbol{b}\right)}{\partial \boldsymbol{b}}\right|_{\boldsymbol{b}=\boldsymbol{\delta}(\boldsymbol{w}_i)},$$

and $\boldsymbol{\ell}_{\boldsymbol{\delta},\boldsymbol{\delta}}\left(\boldsymbol{y}_i, \boldsymbol{x}_i, \boldsymbol{\delta}(\boldsymbol{w}_i)\right)$ the $L \times L$-matrix of second order derivatives with entries $\{k_1, k_2\}$ defined as

$$\left[\boldsymbol{\ell}_{\boldsymbol{\delta},\boldsymbol{\delta}}\left(\boldsymbol{y}_i, \boldsymbol{x}_i, \boldsymbol{\delta}(\boldsymbol{w}_i)\right)\right]_{k_1,k_2} = \left.\frac{\partial^2 \ell\left(\boldsymbol{y}_i, \boldsymbol{x}_i, \boldsymbol{b}\right)}{\partial b_{k_1} \partial b_{k_2}}\right|_{\boldsymbol{b}=\boldsymbol{\delta}(\boldsymbol{w}_i)}.$$

Define $H_{\boldsymbol{\delta}}(\boldsymbol{w}_i, \boldsymbol{\delta}(\boldsymbol{w}_i); \boldsymbol{x}^*)$ as the L-dimensional vector of first derivatives of $H(\boldsymbol{w}_i, \boldsymbol{\delta}(\boldsymbol{w}_i); \boldsymbol{x}^*)$ w.r.t. $\boldsymbol{\delta}(\boldsymbol{w}_i)$. Further, define

$$\boldsymbol{\Lambda}(\boldsymbol{w}_i) := \mathbb{E}[\ell_{\boldsymbol{\delta},\boldsymbol{\delta}}(\boldsymbol{Y}, \boldsymbol{X}, \boldsymbol{\delta}(\boldsymbol{W}))|\boldsymbol{W} = \boldsymbol{w}_i],$$

corresponding to the expected individual Hessian for individual i conditional on her socio-demographic characteristics $\boldsymbol{w}_i$. Then, a valid and Neyman orthogonal score for the parameter of inferential interest, θ_0, is $\psi(\boldsymbol{w}_i, \boldsymbol{\delta}(\boldsymbol{w}_i), \boldsymbol{\Lambda}(\boldsymbol{w}_i)) - \theta_0$, where

$$\psi\left(\boldsymbol{w}_i, \boldsymbol{\delta}(\boldsymbol{w}_i), \boldsymbol{\Lambda}(\boldsymbol{w}_i)\right) = H\left(\boldsymbol{w}_i, \boldsymbol{\delta}\left(\boldsymbol{w}_i\right); \boldsymbol{x}^*\right) - H_{\boldsymbol{\delta}}\left(\boldsymbol{w}_i, \boldsymbol{\delta}\left(\boldsymbol{w}_i\right); \boldsymbol{x}^*\right)' \boldsymbol{\Lambda}\left(\boldsymbol{w}_i\right)^{-1} \boldsymbol{\ell}_{\boldsymbol{\delta}}\left(\boldsymbol{y}_i, \boldsymbol{x}_i, \boldsymbol{\delta}(\boldsymbol{w}_i)\right) \tag{4.2}$$

is the influence function when centered at θ_0. Hence, θ_0 can be identified from the condition $\mathbb{E}\left[\psi\left(\boldsymbol{W}, \boldsymbol{\delta}\left(\boldsymbol{W}\right), \boldsymbol{\Lambda}\left(\boldsymbol{W}\right)\right) - \theta_0\right] = 0$. In case of the conditional logit model stated in Equation (4.1), the gradient vector $\boldsymbol{\ell}_{\boldsymbol{\delta}}(\boldsymbol{y}_i, \boldsymbol{x}_i, \boldsymbol{\delta}(\boldsymbol{w}_i))$ for individual i is

$$\boldsymbol{\ell}_{\boldsymbol{\delta}}(\boldsymbol{y}_i, \boldsymbol{x}_i, \boldsymbol{\delta}(\boldsymbol{w}_i)) = (c_{i,1}, \ldots, c_{i,J}, \tilde{c}_{i,1}, \ldots, \tilde{c}_{i,K})'$$

with jth element $c_{i,j} = y_j - \mathbb{P}(y_{i,j} = 1|\boldsymbol{x}_i, \boldsymbol{w}_i)$ and $(J+k)$th element $\tilde{c}_{i,k} = \sum_{j=1}^{J}(y_{i,j} - \mathbb{P}(y_{i,j} = 1|\boldsymbol{x}_i, \boldsymbol{w}_i))x_{i,j,k}$. The matrix $\boldsymbol{\ell}_{\boldsymbol{\delta},\boldsymbol{\delta}}\left(\boldsymbol{y}_i, \boldsymbol{x}_i, \boldsymbol{\delta}(\boldsymbol{w}_i)\right)$ can be written as

$$\boldsymbol{\ell}_{\boldsymbol{\delta},\boldsymbol{\delta}}\left(\boldsymbol{y}_i, \boldsymbol{x}_i, \boldsymbol{\delta}(\boldsymbol{w}_i)\right) = \dot{\boldsymbol{G}}_i \tilde{\boldsymbol{x}}_i \tilde{\boldsymbol{x}}_i'$$

with $\dot{\boldsymbol{G}}_i$ being the derivative of the conditional logit choice probabilities with respect to the linear index $\tilde{\boldsymbol{x}}_i'\boldsymbol{\delta}(\boldsymbol{w}_i)$, and $\tilde{\boldsymbol{x}}_i = [\boldsymbol{e}_1, \ldots, \boldsymbol{e}_J, \boldsymbol{x}_i]$ where $\boldsymbol{e}_j$ is a unit vector with L elements where the jth element is equal to one and zero otherwise. Thus, the $L \times L$

matrix $\dot{\boldsymbol{G}}_i$ for individual i has entries $\dot{g}_{kk} = \mathbb{P}\left(y_{i,j} = 1|\boldsymbol{x}_i, \boldsymbol{w}_i\right)\left(1 - \mathbb{P}\left(y_{i,j} = 1|\boldsymbol{x}_i, \boldsymbol{w}_i\right)\right)$ on the main diagonal and $\dot{g}_{k,l} = -\mathbb{P}\left(y_{i,j} = 1|\boldsymbol{x}_i, \boldsymbol{w}_i\right)\mathbb{P}\left(y_{i,m} = 1|\boldsymbol{x}_i, \boldsymbol{w}_i\right)$ for all $k \neq l$ on the off-diagonal. A detailed derivation of the influence function for the conditional logit model presented in Equation (4.1) is given in Farrell et al. (2021*a*, v1 on arXiv.org).

The plug-in estimator $\hat{\theta}_{PI}$ takes only one source of uncertainty in $H(\boldsymbol{w}_i, \hat{\boldsymbol{\delta}}(\boldsymbol{w}_i); \boldsymbol{x}^*)$ into account: the direct effect of perturbations in the data on $H(\boldsymbol{w}_i, \hat{\boldsymbol{\delta}}(\boldsymbol{w}_i); \boldsymbol{x}^*)$ for a given $\hat{\boldsymbol{\delta}}(\boldsymbol{w}_i)$ estimated with the sample. In contrast, the influence function approach additionally accounts for the uncertainty in the estimated coefficient functions due to perturbations in the data when estimating θ_0 with machine learning. For illustrative purposes, assume there are estimates $\hat{\boldsymbol{\delta}}(\boldsymbol{w}_i)$ and $\hat{\boldsymbol{\Lambda}}(\boldsymbol{w}_i)$ for a given sample. Using $\hat{\boldsymbol{\delta}}(\boldsymbol{w}_i)$ and $\hat{\boldsymbol{\Lambda}}(\boldsymbol{w}_i)$ to calculate the influence function, $\psi(\boldsymbol{w}_i, \hat{\boldsymbol{\delta}}(\boldsymbol{w}_i), \hat{\boldsymbol{\Lambda}}(\boldsymbol{w}_i))$, presented in Equation (4.2), the sample analogue of $\mathbb{E}[\psi(\boldsymbol{W}, \hat{\boldsymbol{\delta}}\left(\boldsymbol{W}\right), \hat{\boldsymbol{\Lambda}}\left(\boldsymbol{W}\right))]$ is

$$\begin{aligned}
\hat{\theta}_{IF} =& \frac{1}{N}\sum_{i=1}^{N}\psi\left(\boldsymbol{w}_i, \hat{\boldsymbol{\delta}}(\boldsymbol{w}_i), \hat{\boldsymbol{\Lambda}}(\boldsymbol{w}_i)\right) \\
=& \frac{1}{N}\sum_{i=1}^{N} H\left(\boldsymbol{w}_i, \hat{\boldsymbol{\delta}}\left(\boldsymbol{w}_i\right); \boldsymbol{x}^*\right) && \text{(4.3a)} \\
&- \frac{1}{N}\sum_{i=1}^{N} H_{\hat{\boldsymbol{\delta}}}\left(\boldsymbol{w}_i, \hat{\boldsymbol{\delta}}\left(\boldsymbol{w}_i\right); \boldsymbol{x}^*\right)' \hat{\boldsymbol{\Lambda}}\left(\boldsymbol{w}_i\right)^{-1} \boldsymbol{\ell}_{\hat{\boldsymbol{\delta}}}\left(\boldsymbol{y}_i, \boldsymbol{x}_i, \hat{\boldsymbol{\delta}}(\boldsymbol{w}_i)\right). && \text{(4.3b)}
\end{aligned}$$

Similarly to $\hat{\theta}_{PI}$, the term in Equation (4.3a) captures the changes in the function $H(\boldsymbol{w}_i, \hat{\boldsymbol{\delta}}(\boldsymbol{w}_i); \boldsymbol{x}^*)$ in response to perturbations in the data, treating the coefficient functions $\hat{\boldsymbol{\delta}}(\boldsymbol{w}_i)$ as if they were known. This way, the term accounts for the uncertainty in the parameter of inferential interest due to changes in $H(\boldsymbol{w}_i, \hat{\boldsymbol{\delta}}(\boldsymbol{w}_i); \boldsymbol{x}^*)$. The term in Equation (4.3b) is an additional correction term that includes an estimate of the nuisance function $\boldsymbol{\Lambda}(\boldsymbol{w}_i)$ and, thereby, accounts for the uncertainty in the functional forms of the coefficient functions $\boldsymbol{\delta}(\boldsymbol{w}_i)$ arising from perturbations in the data. The correction term isolates the impact of the nonparametric estimation on the estimated parameters of inferential interest, which is enabled through the imposed structure of the economic model relating the outcome $\boldsymbol{Y}$ to the covariates $\boldsymbol{X}$ in a known way.

The correction terms $H_{\boldsymbol{\delta}}(\boldsymbol{w}_i)$, $\boldsymbol{\ell}_{\boldsymbol{\delta}}\left(\boldsymbol{y}_i, \boldsymbol{x}_i, \boldsymbol{\delta}(\boldsymbol{w}_i)\right)$ and $\boldsymbol{\ell}_{\boldsymbol{\delta},\boldsymbol{\delta}}\left(\boldsymbol{y}_i, \boldsymbol{x}_i, \boldsymbol{\delta}(\boldsymbol{w}_i)\right)$ can be calculated analytically and do not need to be estimated. In contrast, the matrix $\boldsymbol{\Lambda}(\boldsymbol{w}_i)$ consists of regression-type objects which must be estimated, i.e., the individual Hessian $\ell_{\boldsymbol{\delta},\boldsymbol{\delta}}(\boldsymbol{Y}, \boldsymbol{X}, \boldsymbol{\delta}(\boldsymbol{W}))$ is projected on $\boldsymbol{W}$. For this projection, DNNs can be used as well. Further, note that the product $\Lambda(\boldsymbol{w}_i)^{-1}\boldsymbol{\ell}_{\boldsymbol{\delta}}(\boldsymbol{w}_i, \boldsymbol{\delta}(\boldsymbol{w}_i))$ does not depend on the function $H(\cdot)$, which simplifies calculations if multiple parameters are of inferential interest.

An important assumption of the inference procedure is that the matrix $\boldsymbol{\Lambda}(\boldsymbol{w}_i)$ is invertible with bounded inverse. With respect to the conditional logit model in Equation (4.1), the assumption implies that the choice probabilities are bounded away from zero and one.[5]

4.2.3 Estimation

With the influence function in Equation (4.2), the estimator $\hat{\theta}$ of θ_0 and an corresponding estimator $\hat{\Psi}$ of its asymptotic variance can be formed using the semiparametric inference procedure of Chernozhukov et al. (2018). For the estimation, the influence function $\psi(\boldsymbol{w}_i, \boldsymbol{\delta}(\boldsymbol{w}_i), \boldsymbol{\Lambda}(\boldsymbol{w}_i))$ needs to be evaluated at every data point in the sample. In order to obtain a properly centered limiting distribution under weaker conditions on the first stage estimates $\hat{\boldsymbol{\delta}}(\boldsymbol{w}_i)$, the estimation procedure for θ_0 is based on sample splitting (Farrell et al., 2021*a*).

For the conditional expected individual Hessian of the conditional logit model, $\boldsymbol{\Lambda}(\boldsymbol{w}_i)$, the dependent variable $\boldsymbol{Z} := \dot{\boldsymbol{G}}\boldsymbol{X}\boldsymbol{X}'$ is regressed on the socio-demographic characteristics $\boldsymbol{W}$. Because $\dot{\boldsymbol{G}}$, and hence $\boldsymbol{Z}$, depend on the coefficient functions $\boldsymbol{\delta}(\boldsymbol{W})$, the estimation of the influence function requires three-way splitting of the sample. The first sub-sample is used to estimate the heterogeneous parameter functions $\hat{\boldsymbol{\delta}}(\boldsymbol{w}_i)$. These are subsequently treated as the inputs to calculate the "observed" matrix $\boldsymbol{z}_i$ of $\boldsymbol{Z}$, using $\boldsymbol{w}_i$ and $\boldsymbol{x}_i$ of the second sub-sample. Using $\boldsymbol{z}_i$ as the dependent variable and $\boldsymbol{w}_i$ as the independent variable, $\hat{\boldsymbol{\Lambda}}(\boldsymbol{w}_i)$ is estimated with the second sub-sample. The influence function is then calculated with the third sub-sample (Farrell et al., 2021*a*). The procedure thus consists of the following steps:

1. Split the observation units $\{1, \dots, n\}$ into S subsets, denoted by $\mathcal{S}_s \subset \{1, \dots, n\}$, $s = 1, \dots, S$.
2. For each $s = 1, \dots, S$, let $\mathcal{S}_s^c$ denote the complement of $\mathcal{S}_s$. For nonlinear models like the conditional logit model, the functions $\boldsymbol{\delta}_s(\boldsymbol{w}_i)$ and $\boldsymbol{\Lambda}_s(\boldsymbol{w}_i)$, corresponding to split s, cannot be estimated simultaneously. Instead, the complement $\mathcal{S}_s^c$ is split into two pieces to first estimate $\hat{\boldsymbol{\delta}}_s(\boldsymbol{w}_i)$ using the first piece, and then $\hat{\boldsymbol{\Lambda}}_s(\boldsymbol{w}_i)$ using the second piece together with the fixed functions $\hat{\boldsymbol{\delta}}_s(\boldsymbol{w}_i)$.
3. The final estimator of θ_0 is then

$$\hat{\theta} = \frac{1}{S}\sum \hat{\theta}_s, \qquad \hat{\theta}_s = \frac{1}{|\mathcal{S}_s|}\sum_{i \in \mathcal{S}_s} \psi\left(\boldsymbol{w}_i, \hat{\boldsymbol{\delta}}_s(\boldsymbol{w}_i), \hat{\boldsymbol{\Lambda}}_s(\boldsymbol{w}_i)\right), \tag{4.4}$$

   where $|\mathcal{S}_s|$ is the cardinality of $\mathcal{S}_s$ and is assumed to be proportional to the sample size.

   Furthermore, an estimator $\hat{\Psi}$ of the asymptotic variance of $\hat{\theta}$ is given by the

[5] In order to assure the numerical stability of the approach, Farrell et al. (2021*a*) propose trimming or regularization of $\Lambda(\boldsymbol{w}_i)$ by adding a positive constant to the main diagonal, e.g., $\Lambda(\boldsymbol{w}_i) + I$.

variance-analogue of Equation (4.4)

$$\hat{\Psi} = \frac{1}{S}\sum_{s=1}^{S}\hat{\Psi}_s, \qquad \hat{\Psi}_s = \frac{1}{|\mathcal{S}_s|}\sum_{i\in\mathcal{S}_s}\left(\psi\left(\boldsymbol{w}_i, \hat{\boldsymbol{\delta}}_s(\boldsymbol{w}_i), \hat{\boldsymbol{\Lambda}}_s(\boldsymbol{w}_i)\right) - \hat{\theta}\right)^2. \tag{4.5}$$

For $\hat{\theta}$ and $\hat{\Psi}$, Farrell et al. (2021*a*) provide inference results that establish asymptotic normality and validity of standard errors,

$$\sqrt{n}\hat{\Psi}^{-1/2}\left(\hat{\theta} - \theta\right) \to_d \mathcal{N}(0,1). \tag{4.6}$$

This allows a simple construction of confidence intervals for θ. Chernozhukov et al. (2018) prove that these are uniformly valid but not necessarily semi-parametrically efficient.[6]

A central input to the influence function, and hence to the estimated inference results, is the conditional expected individual Hessian $\boldsymbol{\Lambda}(\boldsymbol{w}_i)$ which is a nuisance function as it is required only for the calculation of the influence functions but not of interest per se. Estimating $\boldsymbol{\Lambda}(\boldsymbol{w}_i)$ is a prediction problem for which different machine learning methods can be used. In the Monte Carlo experiments and application presented below, we estimate $\boldsymbol{\Lambda}(\boldsymbol{w}_i)$ by another neural network using the mean squared error (MSE) as loss function. Because the matrix $\boldsymbol{\Lambda}(\boldsymbol{w}_i)$ is symmetric, we only need to estimate $L(L+1)/2$ entries. To keep the estimation procedure as simple as possible, we estimate the entries of $\boldsymbol{\Lambda}(\boldsymbol{w}_i)$ using a single network with $L(L+1)/2$ output units. Alternatively, one could estimate each entry with a separate network, which is more flexible but has the disadvantage that it is computationally more expensive.

The estimation procedure described above has some potential weaknesses that can lead to misleading results. The first one is potential overfitting when predicting the choice probability for each alternative, which can lead to estimated probabilities close to zero and one, respectively. As a consequence, the matrix $\hat{\boldsymbol{\Lambda}}(\boldsymbol{w}_i)$ might not be invertible (or close to not being invertible, leading to extremely large entries of the inverse) if the entries are estimated precisely. Related to the overfitting problem, a practical disadvantage of the sample splitting – beyond the computational cost – is that small sub-samples potentially provide imprecise estimates, which is particularly relevant for applications with small sample sizes (Farrell et al., 2021*a*).[7]

Remark 4.1. To increase finite sample precision, Chernozhukov et al. (2018) suggest to repeat the sample splitting procedure outlined above R times. To this end, let $\hat{\theta}_r$ and $\hat{\Psi}_r$ denote the estimators shown in Equation (4.4) and (4.5) for repetition $r = 1, \ldots, R$.

[6] However, the constructed standard confidence intervals for θ can be semi-parametrically efficient and Farrell et al. (2021*a*) also conjecture that they are semi-parametrically efficient but do not prove it.

[7] For the asymptotic results of the sample splitting procedure, Farrell et al. (2021*a*) treat S as fixed and therefore, the sample splitting is asymptotically negligible.

Then, the final estimator is the median over the repetitions,[8] i.e,

$$\hat{\theta}^{med} = median\left\{\hat{\theta}_r\right\}_{r=1}^{R}, \quad \text{and} \quad \hat{\Psi}^{med} = median\left\{\hat{\Psi}_r + \left(\hat{\theta}_r - \hat{\theta}^{med}\right)^2\right\}_{r=1}^{R}.$$

Chernozhukov et al. (2018) note that the choice of $R \geq 1$ does not affect the asymptotic distribution of $\hat{\theta}^{med}$. By Equation (4.6), each $\hat{\theta}_k$ is asymptotically normal and therefore, $\hat{\theta}^{med}$ is asymptotically normal, too. In our simulations, we set $R = 5$ and find that repeated sample splitting substantially improves the precision of the estimates.

4.3 Monte Carlo Experiments

This section presents different Monte Carlo experiments that study the performance of the deep learning estimation procedure and, in particular, the inference procedure presented in Section 4.2. To study the performance in a realistic setup, we use semi-synthentic data for the experiments. The data is taken from the Swissmetro dataset (Bierlaire, Axhausen and Abay, 2001), which is an openly available dataset collected in Switzerland during March 1998.[9] The data consists of survey data from 1,191 car and train travelers. It was collected to analyze the impact of a new innovative transportation mode, represented by the Swissmetro, against usual transportation modes, namely car and regular train connections.[10] For every respondent, nine stated choice situations were generated in which the respondents could choose between three travel mode alternatives: Swissmetro (abbreviated as sm), train, and car (only for car owners). In total, the data consists of 10,719 choice situations (Antonini, Gioia and Frejinger, 2007). When preparing the data, we follow the instructions of Sifringer et al. (2020) and remove all observations for which not all three alternatives – Swissmetro, train, car – are available. This reduces the number of travelers to 1,683 and thus, the final data set to 9,036 observations.[11]

For the data generation, we consider an individual-level discrete choice demand model of the form presented in Equation 4.1. The variables of interest in our Monte Carlo experiments are the travel cost (*cost*), the travel time (*time*), and the frequency (*freq*) of the train and Swissmetro connections (frequency is zero for car).[12] Each traveler chooses the travel mode among the three alternatives car, Swissmetro, and train that

[8] Chernozhukov et al. (2018) also consider taking the average across repetitions instead of the median. However, they recommend to use the median since it is less dependent on the outcome of a single repetition.

[9] We downloaded the test and training data from the github repository github.com/BSifringer/EnhancedDCM.

[10] The Swissmetro is a revolutionary mag-lev underground system operating at speeds up to 500 km/h in partial vacuum.

[11] For the estimation, we follow Sifringer et al. (2020) and ignore the panel structure of the data.

[12] The travel cost, travel time, and frequency variables are scaled downwards by factor one hundred (Sifringer et al., 2020). For those travelers that have an annual season pass, we set the travel cost of the train and Swissmetro to zero.

provides her with the highest utility,

$$u_{i,j} = \alpha_j(\boldsymbol{w}_i) + cost_{i,j}\beta^{\text{cost}}(\boldsymbol{w}_i) + time_{i,j}\beta^{\text{time}}(\boldsymbol{w}_i) + freq_{i,j}\beta^{\text{freq}}(\boldsymbol{w}_i) + \omega_{i,j},$$

for $j = \{\text{car}, \text{train}, \text{sm}\}$. We specify the true coefficients as functions of travelers' yearly income (*income*), age (*age*), gender (*male*), and a variable indicating who payed for the ticket (*who*). Income and age are categorical variables that assign travelers' income and age into four and six groups, respectively. The gender variable is equal to one if the traveler is male and zero otherwise. The variable *who* is a categorical variable that takes four values (0 if it is unknown who pays, 1 if the traveler payed herself, 2 if the employer pays, and 3 if the traveler and employer split half-half). In order to make the information represented by the categorical variable more easily accessible for the network, we transform *who* into three dummy variables denoted by who^1, who^2, and who^3, leaving out the category 0 as reference category.[13] We specify the observed consumer socio-demographics as $\boldsymbol{w}_i := (age_i, income_i, male_i, who_i^1, who_i^2, who_i^3)'$. The intercept functions for each alternative are

$$\alpha_{\text{train}}(\boldsymbol{w}_i) = -1 + 1 \cdot income_i,$$

$$\alpha_{\text{sm}}(\boldsymbol{w}_i) = -3 + 1 \cdot age_i,$$

and $\alpha_{\text{car}}(\boldsymbol{w}_i) = 0$, i.e., the alternative car serves as reference. The coefficient functions for the covariates of interest are specified as

$$\beta^{\text{cost}}(\boldsymbol{w}_i) = -6 + income_i - 0.8 \cdot who_i^1 - 1 \cdot who_i^2 - 1.2 \cdot who_i^3$$

$$\beta^{\text{freq}}(\boldsymbol{w}_i) = -5 + income_i + 0.9 \cdot male_i$$

$$\beta^{\text{time}}(\boldsymbol{w}_i) = -6 + 1 \cdot age_i.$$

To study the finite sample performance of the proposed inference procedure, we consider the expected value of the heterogeneous coefficients $\beta^{\text{cost}}(\boldsymbol{w}_i)$, $\beta^{\text{freq}}(\boldsymbol{w}_i)$, and $\beta^{\text{time}}(\boldsymbol{w}_i)$ as the parameters of inferential interest, i.e., $\theta_0^k = E[\beta^k(\boldsymbol{w}_i)]$, $k \in \{\text{cost}, \text{freq}, \text{time}\}$. Accordingly, the function $H(\cdot)$ corresponds to

$$H(\boldsymbol{w}_i, \boldsymbol{\delta}(\boldsymbol{w}_i); \boldsymbol{x}^*) = \beta^k(\boldsymbol{w}_i),$$

where $\boldsymbol{\delta}(\boldsymbol{w}_i) = \left(\alpha_{\text{train}}(\boldsymbol{w}_i), \alpha_{\text{sm}}(\boldsymbol{w}_i), \beta^{\text{cost}}(\boldsymbol{w}_i), \beta^{\text{freq}}(\boldsymbol{w}_i), \beta^{\text{time}}(\boldsymbol{w}_i)\right)'$. Thus, the gradient vector $H_{\boldsymbol{\delta}}(\boldsymbol{w}_i, \boldsymbol{\delta}(\boldsymbol{w}_i); \boldsymbol{x}^*)$ is equal to one for the element corresponding to the derivative with respect to β^k, and zero for all other entries.

[13] A detailed description of the data and summary statistics can be found here.

4.3.1 Small Data Set

We conduct 1000 Monte Carlo repetitions. In every repetition, we use the individual coefficients, the covariates, and an idiosyncratic error term $\omega_{i,j}$ to calculate the utility for each alternative and each individual. For that purpose, we draw $\omega_{i,j}$ from a Type I extreme value distribution for every traveler and alternative in every replicate and select the alternative that provides the largest utility.

To simulate deviations between the sample and the population values of the covariates, we split the data into two sets. We use all observations to calculate the true values, θ_0^k, $k \in \{\text{cost}, \text{freq}, \text{time}\}$, but use only three quarter of the data for the estimation. This way, we can test whether the proposed inference procedure adequately accounts for the uncertainty related to $H(\cdot)$, and for the uncertainty related to the functional form of the heterogeneous coefficient functions $\boldsymbol{\delta}(\boldsymbol{w}_i)$ which arises due to deviations between observations in the sample and the population.

We use the same network architecture to estimate the heterogeneous coefficient functions and to estimate the conditional expected individual Hessian $\boldsymbol{\Lambda}(\boldsymbol{w}_i)$ – except for the number of output units in both networks. More precisely, we choose one hidden layer with 100 units and rectified linear activation functions. For the units in the output layer, we use linear activation functions. The number of output units are five in the network for the heterogeneous coefficient functions, and 15 in the network for $\boldsymbol{\Lambda}(\boldsymbol{w}_i)$. Both networks use travelers' income, age, gender, and the dummy variables indicating who is paying for the ticket as inputs. When estimating the coefficient functions, we set the dropout rate to 0.2. For the network used to estimate $\boldsymbol{\Lambda}(\boldsymbol{w}_i)$, we test different regularizers to account for the difficulty of projecting $\boldsymbol{\Lambda}(\boldsymbol{w}_i)$. We consider the l_2-regularizers $\lambda = 0, 10^{-5}, 10^{-4}, 2{\cdot}10^{-3}$ which we use to avoid overfitting $\boldsymbol{z}_i$ and, thereby, to ensure that the predicted individual Hessian $\hat{\boldsymbol{\Lambda}}(\boldsymbol{w}_i)$ does not become collinear for any individual i. While using a l_2-regularizer $\lambda > 0$ ensures that we can invert $\hat{\boldsymbol{\Lambda}}(\boldsymbol{w}_i)$, we note that $\lambda > 0$ potentially introduces a bias in the estimation and is not covered by the inference results of Farrell et al. (2021*a*). When training the networks, we set the maximum number of epochs to $20,000$, and the batch size to 50. During the training, we track the in-sample log-likelihood and the in-sample mean squared error, respectively, and stop the training if the change in the loss function does not exceed 10^{-8} across epochs (with a patience of 100 epochs). We select the network with the best in-sample fits. For the estimation with the influence function approach, we split the training data into $S = 5$ folds. Furthermore, we split $\mathcal{S}_s^c$ into two equally sized pieces, using the first one to estimate $\boldsymbol{\delta}_s(\boldsymbol{w}_i)$, and the second one to estimate $\boldsymbol{\Lambda}_s(\boldsymbol{w}_i)$.

As a benchmark, we estimate the model with maximum likelihood using the true specification. We refer to this estimator as oracle logit estimator. In addition, we also estimate a conditional logit model where we do not account for any type of heterogeneity but instead include only two alternative-specific intercepts and the slope coefficients for *cost*, *freq*, and *time*. This allows us to study the potential consequences when one does not account for heterogeneity across travelers even though it is present in the data. Finally,

we also use a neural network to estimate the heterogeneous coefficient functions without the outlined inference procedure of Farrell et al. (2021*a*). Instead, we conduct naive inference using the average heterogeneous coefficient functions and the corresponding estimated Fisher information matrix to calculate robust standard errors. This allows us to assess the importance of an appropriate inference procedure after the estimation of the model parameters with machine learning.

Table 4.1 reports the coverage of the estimated 95% confidence intervals, the average estimated standard errors, and estimated bias across Monte Carlo replicates for all three covariates of interest. Furthermore, we present the share of Monte Carlo replicates in which the false null hypotheses that the coefficients are zero are correctly rejected at a significance level of 0.05. This is supposed to serve as an indicator for the power of the hypothesis tests when calculated with the different inference procedures. For the influence function approach, we additionally calculate the in-sample and out-of-sample MSE of the neural network for $\boldsymbol{\Lambda}_s(\boldsymbol{w}_i)$, and track the share of outliers across Monte Carlo replicates. We calculate the in-sample MSE with the part of $\mathcal{S}_s^c$ used for the estimation of $\boldsymbol{\Lambda}_s(\boldsymbol{w}_i)$, and the out-of-sample MSE with the left out fold. We treat a Monte Carlo replicate as outlier if the estimated standard error is larger than 5 for at least one of the three estimated parameters.

The reported average results for the oracle logit estimator across Monte Carlo replicates reveal that accounting for the correct (functional) form of heterogeneity provides precise estimates of the true average coefficients, and correct coverage of the true average coefficients through the estimated 95% confidence intervals. In addition, the hypotheses tests with the nulls that the average coefficients are zero have high power when calculated with the oracle logit estimator, as the null hypotheses are correctly rejected in every Monte Carlo replicate. In contrast, the basic logit estimator, which does not account for any heterogeneity across consumers, performs poorly both in terms of the estimated coefficients and in terms of the coverage of the confidence intervals. The estimated standard errors of the oracle logit and the basic logit seem similar but the confidence intervals do not cover the true values of interest in any of the Monte Carlo replicates when estimated with the basic logit. The poor coverage can be explained by the bias of the estimated coefficients, which implies confidence intervals centered around biased estimates.

The results for the influence function approach depend on the regularization parameter λ used for the estimation of $\boldsymbol{\Lambda}(\boldsymbol{w}_i)$. For $\lambda = 0$, the confidence intervals for all three parameters have a coverage of 93%, giving the impression that the influence function approach is a valid inference procedure when the heterogeneous coefficient functions are estimated with deep learning and without regularization in the network used to estimate $\boldsymbol{\Lambda}(\boldsymbol{w}_i)$. However, the estimated average coefficients deviate quite substantially from the true values – especially for the travel cost and travel time coefficients –, and the estimated standard errors are substantially larger than in the oracle logit estimator. The large estimated standard errors explain the correct coverage of the confidence intervals despite of the biased average coefficient estimates. Even though the confidence intervals are

Table 4.1: Average Summary Statistics of 1000 Monte Carlo Replicates for Small Data and without Repeated Sample Splitting

	Conditional Logit		Influence Function Approach with λ equal to				
	Oracle	Basic	0	10^{-5}	10^{-4}	$2 \cdot 10^{-3}$	NN
$\theta_{cost} \in \widehat{CI}_{cost}$	0.95	0.00	0.93	0.92	0.83	0.40	0.99
$\theta_{freq} \in \widehat{CI}_{freq}$	0.95	0.00	0.93	0.92	0.89	0.68	1.00
$\theta_{time} \in \widehat{CI}_{time}$	0.94	0.00	0.93	0.94	0.88	0.54	1.00
$\widehat{se}_{cost}$	0.07	0.05	6.85	3.67	1.70	0.75	0.61
$\widehat{se}_{freq}$	0.10	0.07	5.25	8.19	2.26	1.24	3.56
$\widehat{se}_{time}$	0.07	0.06	5.52	4.91	1.53	1.05	3.08
Bias_{cost}	-0.01	0.65	-4.95	0.07	-0.09	-0.51	-0.17
Bias_{freq}	-0.00	0.59	0.61	-4.45	-0.77	-0.61	-0.18
Bias_{time}	-0.01	0.80	-2.58	-1.49	-0.27	-0.57	-0.17
Rej. $\theta_{cost} = 0$	1.00	1.00	0.47	0.56	0.78	0.93	1.00
Rej. $\theta_{freq} = 0$	1.00	1.00	0.27	0.35	0.50	0.79	0.00
Rej. $\theta_{time} = 0$	1.00	1.00	0.60	0.69	0.84	0.93	0.03
$MSE(\Lambda)^{Train}$	.	.	5.04	5.18	5.41	5.99	.
$MSE(\Lambda)^{Test}$	.	.	5.30	5.38	5.51	6.04	.
Share Outlier	0.00	0.00	0.26	0.18	0.11	0.04	0.12

Note: The table reports the average summary statistics over all Monte Carlo replicates for the conditional logit using the true specification (Oracle), the conditional logit using the three variables of interest for the estimation (Basic), the influence function approach using five different values for λ for the estimation of $\Lambda_s(\boldsymbol{w}_i)$, and the neural network (NN), which uses robust standard erros and does not rely on the influence function approach.

centered around biased estimates, they are so large that they cover the true parameters in about 93% of the replicates for all three variables of interest. Moreover, the large estimated standard errors lead to low power of the hypotheses tests with the nulls that the true coefficients are zero as shown by the small share of rejections of the null hypotheses – at most in only about 60% of the Monte Carlo replicates.

Overall, choosing $\lambda > 0$ leads to more precise estimates of the true average coefficients (considering all three coefficients together, the estimates are most precise for $\lambda = 10^{-4}$), and to smaller estimated standard errors. However, the bias of the estimated average coefficients remains relatively large, so that the coverage of the confidence intervals gradually declines with increasing λ due to the smaller estimated standard errors with increasing λ. For instance for $\lambda = 2 \cdot 10^{-3}$, the confidence intervals have a coverage of only about 68% or less. The fact that the estimated coefficients tend to become more precise and the share of outliers decreases with increasing λ indicates that the large deviation of the estimated coefficients from the true values for $\lambda = 0$ are driven by

outliers. This is illustrated by the boxplot of $\hat{\theta}_{freq}$ in Panel (a) of Figure 4.2. The mean (red point) and median (horizontal line inside the colored boxes) values deviate quite substantially, which is due to the high minimum and maximum values of $\hat{\theta}_{freq}$ across Monte Carlo replicates. The median biases and estimated standard errors across Monte Carlo replicates reported in Table 4.A.1 in Appendix 4.A confirm this impression. The results show that the median of the estimated coefficients across Monte Carlo replicates are closer to the true values for $\lambda = 0$ and become less precise with increasing λ. More importantly, the median of the estimated standard errors are substantially smaller than the mean values across Monte Carlo replicates for each λ value. Overall, the median results for different values of λ are in line with the expected effect of regularization: The bias increases and the estimated standard errors decrease with increasing λ. The average of the MSEs of the neural network for $\boldsymbol{\Lambda}_s(\boldsymbol{w}_i)$ in the training and test sample are lowest for $\lambda = 0$ and therefore, the MSE may be used to choose an appropriate λ value.[14]

Estimating the average heterogeneous coefficients with a neural network without the influence function approach provides more accurate estimates than the influence function approach. However, the confidence intervals are too wide (the coverage is at least 99% for all three variables), implying that the naive inference procedure with the regular robust standard errors is not valid. This is also indicated by the poor power of the hypotheses tests with the nulls that the average travel time and frequency coefficients are zero, which are rejected in only 3% and 0% of the Monte Carlo replicates, respectively. The results on the share of outliers reveal that the issue is not unique to the influence function approach but also appears when the parameters are estimated with a neural network and without sample splitting. However, the share is substantially smaller in comparison to the influence function approach with $\lambda = 0$, indicating that the smaller samples used for the estimation of the networks due to sample splitting might be one of the reasons causing the issue. The Monte Carlo experiment in Subsection 4.3.2 studies the performance of the influence function approach for a larger sample size.

To resolve the sensitivity of the estimated results to potential outliers, we apply the repeated sample splitting procedure outlined in Remark 4.1. Table 4.2 reports the results for the sample splitting procedure with $R = 5$ repetitions.[15] The repeated sample splitting reduces the share of outliers substantially in comparison to the approach without repeated sample splitting. In fact, for $\lambda \geq 10^{-4}$, there are no outliers anymore. Comparing Panel (a) and (b) in Figure 4.2 illustrates that the estimates vary less across Monte Carlo replicates when estimated with repeated sample splitting. Furthermore, the less extreme minimum and maximum values indicate that the extreme outliers are removed. Accordingly, the mean and median values are closer to each other when the coefficient functions are estimated with repeated sample splitting. The reduced share of outliers leads to more precise estimates of the average coefficients and to smaller estimated

[14] Note that the MSE in the test sample is also available to the researcher since it is calculated with the left out fold.

[15] To reduce computation time, we only employ repeated sample splitting if we observe an outlier in the first repetition of each Monte Carlo run.

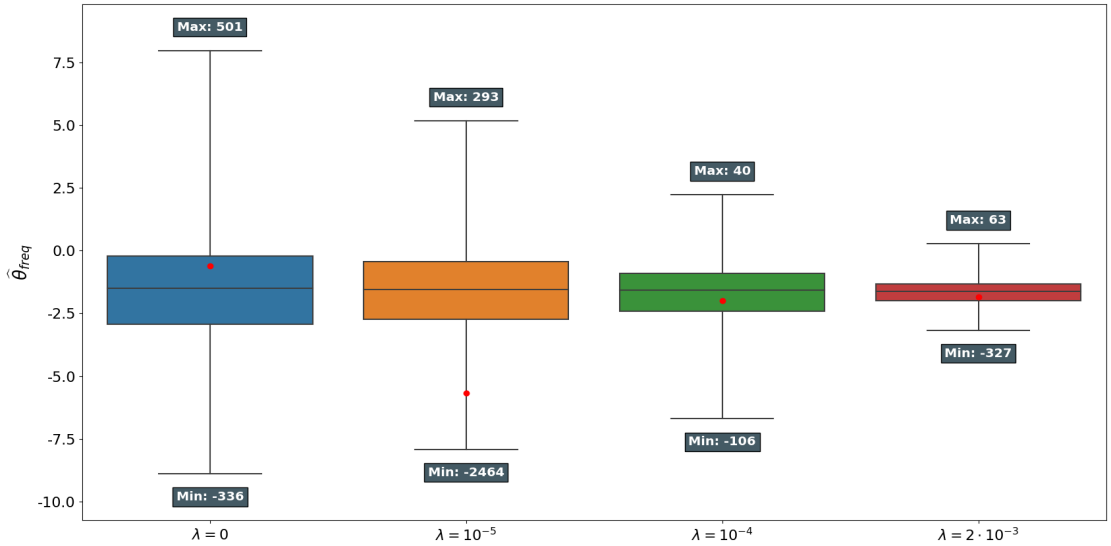
Table 4.2: Average Summary Statistics of 1000 Monte Carlo Replicates for Small Data and Repeated Sample Splitting with $R = 5$

	Conditional Logit		Influence Function Approach with λ equal to				
	Oracle	Basic	0	10^{-5}	10^{-4}	$2 \cdot 10^{-3}$	NN
$\theta_{cost} \in \widehat{CI}_{cost}$	0.94	0.00	0.94	0.93	0.82	0.42	0.99
$\theta_{freq} \in \widehat{CI}_{freq}$	0.96	0.00	0.94	0.92	0.90	0.66	1.00
$\theta_{time} \in \widehat{CI}_{time}$	0.96	0.00	0.95	0.95	0.87	0.59	1.00
$\widehat{se}_{cost}$	0.07	0.05	1.61	1.34	0.72	0.32	0.61
$\widehat{se}_{freq}$	0.10	0.07	1.91	1.64	1.10	0.58	3.65
$\widehat{se}_{time}$	0.07	0.06	1.59	1.19	0.68	0.43	3.13
Bias_{cost}	-0.01	0.65	-0.29	-0.30	-0.32	-0.41	-0.18
Bias_{freq}	-0.00	0.59	-0.26	-0.32	-0.28	-0.39	-0.19
Bias_{time}	-0.00	0.81	-0.02	-0.16	-0.23	-0.36	-0.17
Rej. $\theta_{cost} = 0$	1.00	1.00	0.48	0.61	0.84	0.96	0.99
Rej. $\theta_{freq} = 0$	1.00	1.00	0.25	0.32	0.54	0.81	0.00
Rej. $\theta_{time} = 0$	1.00	1.00	0.64	0.78	0.92	0.96	0.02
$MSE(\Lambda)^{Train}$	.	.	5.07	5.23	5.46	6.04	.
$MSE(\Lambda)^{Test}$	.	.	5.30	5.37	5.51	6.03	.
Share Outlier	0.00	0.00	0.05	0.02	0.00	0.00	0.12

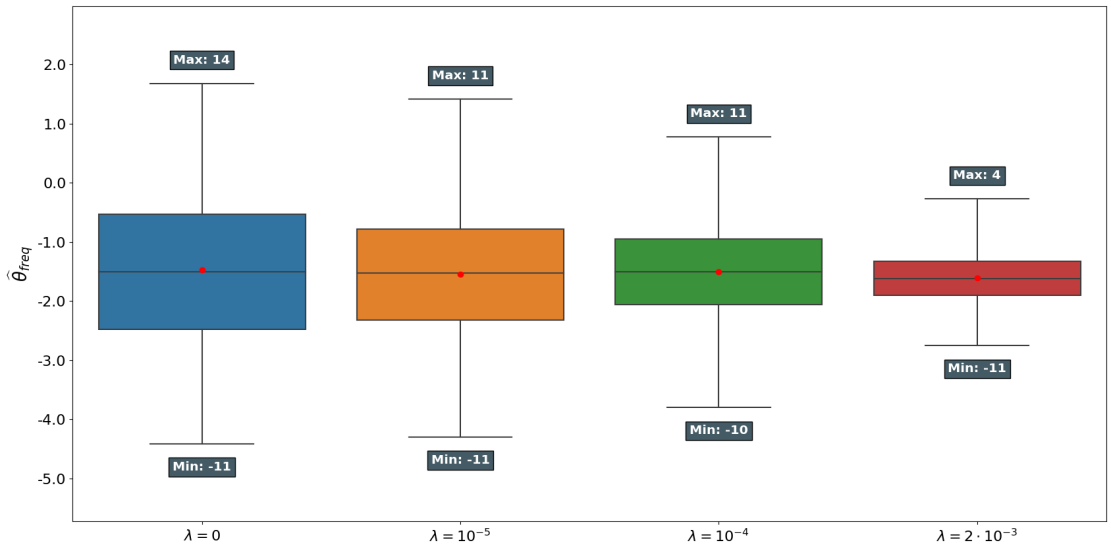
Note: The table reports the average summary statistics over all Monte Carlo replicates for the conditional logit using the true specification (Oracle), the conditional logit using the three variables of interest for the estimation (Basic), the influence function approach, using five different values of λ for the estimation of $\Lambda_s(\boldsymbol{w}_i)$, and the neural network (NN), which uses robust standard erros and does not rely on the influence function approach.

standard errors. In contrast to the influence function approach without repeated sample splitting, the overall average bias of the estimated average coefficients is smallest for $\lambda = 0$ and increases with increasing λ. With respect to the confidence intervals, the coverage for $\lambda = 0$ is 94% for the travel cost and frequency coefficients, and 95% for the travel time coefficient. The coverage of the confidence intervals gradually decreases with λ. While for $\lambda = 10^{-5}$ the coverage is below but still close to 95% (for the travel time it is exactly 95%), the coverage for $\lambda = 2 \cdot 10^{-3}$ is at most 66% (for the travel cost coefficient, the coverage of the confidence interval is just 42%). Thus, the influence function approach with repeated sample splitting and regularizer $\lambda = 0$ allows to precisely estimate average effects across travelers and provides a valid inference procedure. Using a regularizer $\lambda > 0$ increases the average bias and decreases the estimated variance of the coefficients. The combination of increasing bias and decreasing magnitude of the estimated standard errors with increasing λ leads to inappropriately small confidence intervals centered around biased estimates and, hence, to a poor coverage of the true values. Based on these results, we do not recommend using regularization in the form of

Figure 4.2: Boxplots of $\hat{\theta}_{freq}$ across Monte Carlo Replicates for Small Data and Different λ-values



(a) No repeated sample splitting ($R = 1$)



(b) Repeated sample splitting ($R = 5$)

Note: Panel (a) and (b) show boxplots for the influence function approach, using four different values of λ. The colored region within each boxplots highlights the interquartile range (IQR), the horizontal line within the IQR corresponds to the median, and the whiskers indicate the 0.05 and 0.95 quantile, respectively. The red dot is the mean across Monte Carlo replicates.

a l_2-penalty with $\lambda > 0$ in the network used to estimate $\boldsymbol{\Lambda}(\boldsymbol{w}_i)$ to stabilize the inference procedure but to rather rely on repeated sample splitting. However, even for the repeated

sample splitting, the estimated standard errors are substantially larger than those in the oracle logit model. This leads to a poor power as indicated by the rare rejection of the false null hypotheses that the true average coefficients are zero, which are rejected in only about 48%, 25%, and 64% of the Monte Carlo replicates for the travel cost parameter, the frequency parameter, and the travel time parameter, respectively, for $\lambda = 0$.

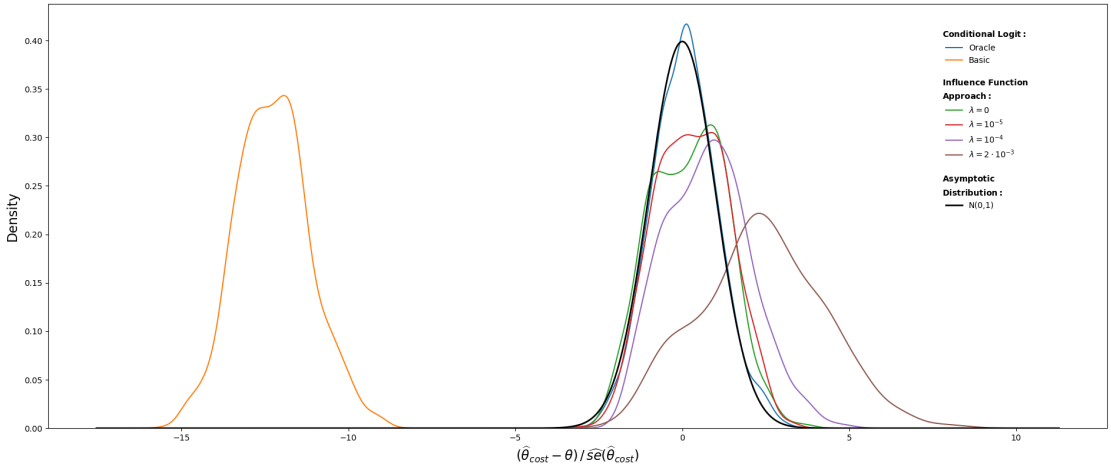
Figure 4.3 shows the estimated densities of $(\hat{\theta}_{cost} - \theta)/\hat{se}(\hat{\theta}_{cost})$ for the oracle logit estimator, the basic logit estimator, and the influence function approach for different values of λ. The limiting distribution of the influence function approach is the standard normal as stated in Equation (4.6). First, the figure illustrates the bias of the basic logit estimator and illustrates that the estimated t-statistics of the oracle logit estimator are well approximated by a standard normal distribution. Second, comparing Panel (a) and Panel (b) reveals that the estimates obtained with the influence function approach only seem to be close to the standard normal distribution when repeated sample splitting is used and $\lambda = 0$ or $\lambda = 10^{-5}$.

Remark 4.2. Beyond repeated sample splitting, we conduct several other adjustments of the estimation procedure that are intended to reduce outliers observed in some Monte Carlo replicates. We considered taking the median instead of the average in Equation (4.4) and (4.5), i.e., replacing $\hat{\theta} = \frac{1}{S}\sum \hat{\theta}_s$ by $\hat{\theta} = median\left\{\hat{\theta}_s\right\}_{s=1}^{S}$ and $\hat{\Psi} = \frac{1}{S}\sum_{s=1}^{S} \hat{\Psi}_s$ by $\hat{\Psi} = median\left\{\hat{\Psi}_s\right\}_{s=1}^{S}$. This leads to smaller estimated standard errors but also to a lower average coverage across Monte Carlo replicates (< 0.85), indicating that the bias remains large. Furthermore, we also apply the modification suggested by Farrell et al. (2021*a*) and add a constant c to the diagonal elements of $\hat{\mathbf{\Lambda}}_s(\boldsymbol{w}_i)$. For $c = 1$, the coverage is quite poor, and $c = 10^{-5}$ seems to have no impact on the results. That is, the choice of the constant c seems to require further tuning which we did not investigate further.

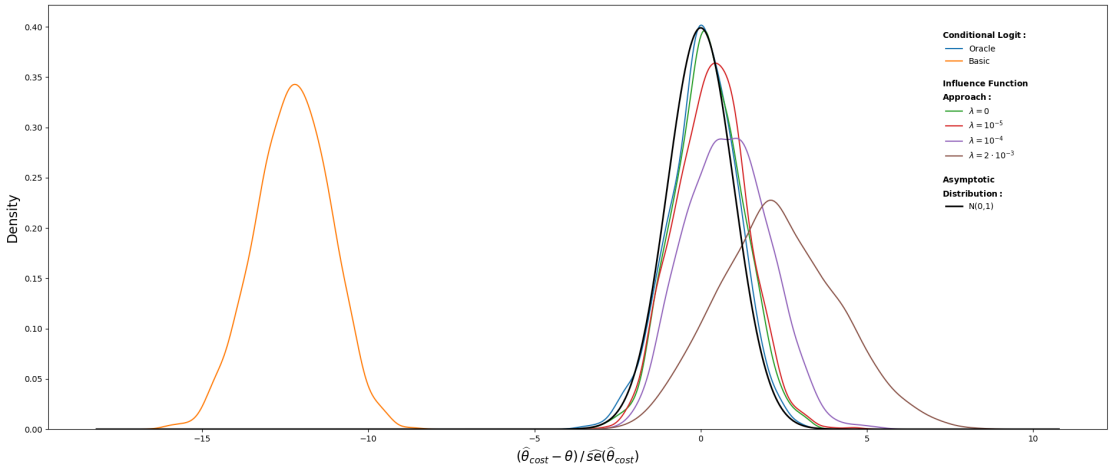
4.3.2 Large Data Set

The following Monte Carlo experiment aims to analyze whether the results of the previous experiment persist for larger sample sizes. For that purpose, we revisit the Swissmetro data set and use the same specification as before. However, we now sample the socio-demographic characteristics and the covariates of interest with replacement from the original data set such that we obtain $50,000$ travelers choosing among the three alternatives. With respect to the socio-demographic characteristics, we randomly generate new travelers by drawing from the values of *income*, *age*, *gender*, and *who*. Because we sample independently across characteristics, we create new types of travelers characterized through new combinations of socio-demographic variables.

Figure 4.3: Density of Estimated t-Statistic of $\hat{\theta}_{cost}$ for Different Estimators and Small Data



(a) No repeated sample splitting ($R = 1$)



(b) Repeated sample splitting ($R = 5$)

Note: The plot shows kernel density estimates of the estimated t-statistic for the conditional logit using the true specification (Oracle), the conditional logit using the three variables of interest for the estimation (Basic), the influence function approach, using four different values for λ for the estimation of $\Lambda_s(\boldsymbol{w}_i)$. Additionally, the standard normal distribution is included.

With respect to the covariates of interest, we make sure that we randomly draw the travel time, travel cost, and frequency for a specific alternative only from the the values for the specific alternative existing in the data (e.g., the cost variable for alternative car can only take values of existing values of the cost variable for cars). However, for a given alternative, we draw the covariates independently across variables from different choice situations. Otherwise, the Monte Carlo study is the same as the one presented above.

Table 4.3 reports the average Monte Carlo results for $N = 50,000$ and when the influence function approach is estimated with repeated sample splitting. The results for the oracle logit and the basic logit are similar to those obtained for the small sample size. For the oracle logit, the average estimated bias across Monte Carlo replicates is (almost) zero and the confidence intervals cover the true frequency and travel time coefficients in 95% of the Monte Carlo replicates, and the true travel cost coefficient in 94%. For the basic logit model, the standard errors of the estimated coefficients are similar to those of the oracle logit. Nevertheless, the confidence intervals have zero coverage due to the substantial bias of the estimated average coefficients.

For the influence function approach with repeated sample splitting, the estimated coefficients are almost as precise as those estimated with the oracle logit model, independent of λ (i.e., the average values vary only slightly across different values for λ), and the estimated standard errors are substantially smaller in comparison to the results

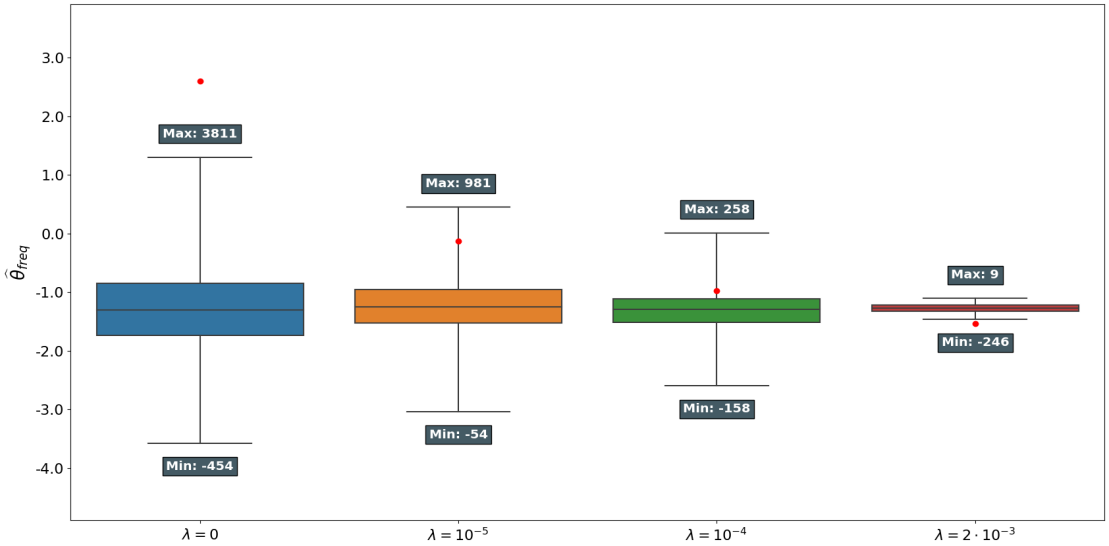
Table 4.3: Average Summary Statistics of 1000 Monte Carlo Replicates for Large Data and Repeated Sample Splitting with $R = 5$

	Conditional Logit		Influence Function Approach with λ equal to				
	Oracle	Basic	0	10^{-5}	10^{-4}	$2 \cdot 10^{-3}$	NN
$\theta_{cost} \in \widehat{CI}_{cost}$	0.95	0.00	0.94	0.92	0.91	0.76	1.00
$\theta_{freq} \in \widehat{CI}_{freq}$	0.95	0.00	0.95	0.93	0.90	0.83	1.00
$\theta_{time} \in \widehat{CI}_{time}$	0.94	0.00	0.95	0.94	0.92	0.86	1.00
$\widehat{se}_{cost}$	0.02	0.02	0.50	0.41	0.35	0.12	0.49
$\widehat{se}_{freq}$	0.04	0.04	0.80	0.59	0.44	0.14	1.72
$\widehat{se}_{time}$	0.03	0.02	0.45	0.36	0.29	0.14	1.27
Bias_{cost}	0.00	0.60	-0.05	-0.02	-0.05	-0.05	-0.03
Bias_{freq}	0.00	0.53	-0.09	-0.07	-0.06	-0.05	-0.03
Bias_{time}	0.00	0.78	0.01	0.02	-0.03	-0.04	-0.02
Rej. $\theta_{cost} = 0$	1.00	1.00	0.89	0.91	0.93	0.98	1.00
Rej. $\theta_{freq} = 0$	1.00	1.00	0.62	0.75	0.84	0.96	0.00
Rej. $\theta_{time} = 0$	1.00	1.00	0.95	0.96	0.97	0.99	0.94
$MSE(\Lambda)^{Train}$	.	.	8.80	8.84	8.90	9.23	.
$MSE(\Lambda)^{Test}$	.	.	8.88	8.87	8.91	9.23	.
Share Outlier	0.00	0.00	0.00	0.00	0.00	0.00	0.00

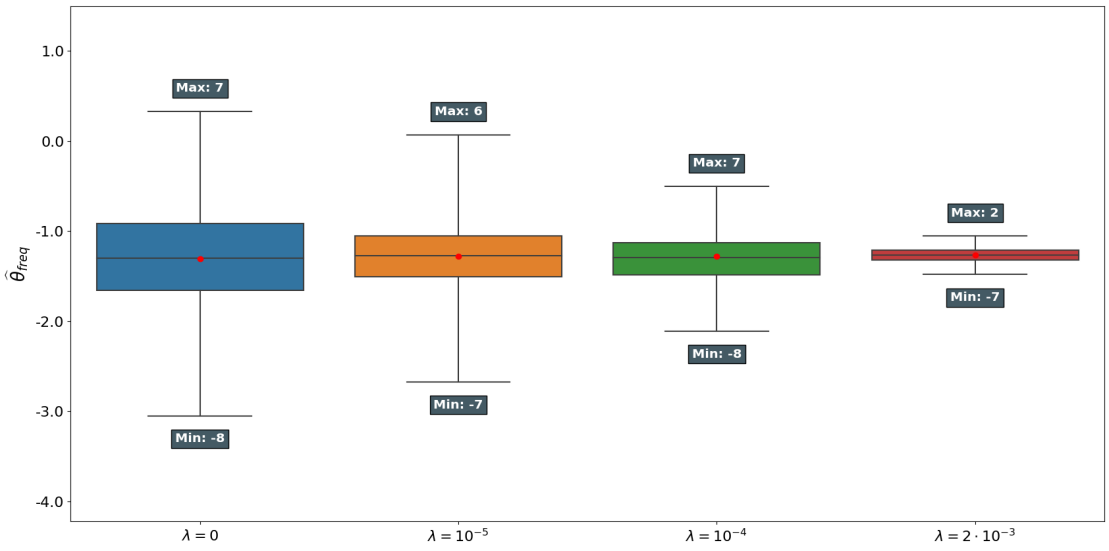
Note: The table reports the average summary statistics over all Monte Carlo replicates for the conditional logit using the true specification (Oracle), the conditional logit using the three variables of interest for the estimation (Basic), the influence function approach, using five different values for λ for the estimation of $\mathbf{\Lambda}_s(\boldsymbol{w}_i)$, and the neural network (NN), which uses robust standard erros and does not rely on the influence function approach.

for the small sample size. However, they are still larger than those estimated with the oracle logit estimator. For $\lambda = 0$, the confidence intervals have the correct coverage (they cover the true travel cost parameter in 94%, and the true frequency and travel time parameters in 95% of the Monte Carlo replicates). For $\lambda > 0$, the coverage of the confidence intervals decreases below 95%, which is the result of the declining estimated standard errors with increasing λ. However, the coverage declines not as rapidly with increasing λ as observed for the small sample size. With respect to the power of the hypotheses tests with the nulls that the coefficients are zero, the percentage of rejections of the incorrect null hypothesis are substantially larger for $\lambda = 0$ than for the small sample size – in 89% of the Monte Carlo replicates for the travel time coefficient, 62% for the frequency coefficient, and 95% for the travel time coefficient. Even though the share of outliers for the influence function approach decreases substantially compared to the Monte Carlo experiment with the small sample size, repeated sample splitting seems still necessary as the mean deviates substantially from the median when no repeated sample splitting is used (cf. Table 4.A.3 and Table 4.A.4 and Figure 4.4).

Figure 4.4: Boxplots of $\hat{\theta}_{freq}$ across Monte Carlo Replicates for Large Data and Different λ-values



(a) No repeated sample splitting ($R = 1$)



(b) Repeated sample splitting ($R = 5$)

Note: Panel (a) and (b) show boxplots for the influence function approach, using four different values of λ. The colored region within each boxplots highlights the interquartile range (IQR), the horizontal line within the IQR corresponds to the median, and the whiskers indicate the 0.05 and 0.95 quantile, respectively. The red dot is the mean across Monte Carlo replicates.

With respect to the estimation of the coefficient functions with a deep neural network and naive inference, we observe a similar improvement when increasing the sample size

as for the influence function approach. The estimated average coefficients become more precise – they are similarly precise as those obtained with the oracle logit – and the estimated standard errors become smaller. A potential explanation for the more precise coefficient estimates and the smaller standard errors might be the fact that the issue with the outlier disappears completely, both for the influence function approach with repeated sample splitting (even for $\lambda = 0$) and when only the coefficient functions are estimated with the neural network. However, the confidence intervals remain too wide, confirming the impression from the experiments with the small sample size that regular robust standard errors calculated with parameters estimated with deep learning are not a valid inference procedure.

4.4 Application

This sections applies the estimation procedure presented in Section 4.2 to the Swissmetro dataset. We consider the same utility specification as in the Monte Carlo experiments. That is, we include alternative-specific constants (car remains the reference category) along with the travel cost, frequency, and travel time, i.e,

$$\boldsymbol{\delta}(\boldsymbol{w}_i) = \left(\alpha_{\text{train}}\left(\boldsymbol{w}_i\right), \alpha_{\text{sm}}\left(\boldsymbol{w}_i\right), \beta^{\text{cost}}\left(\boldsymbol{w}_i\right), \beta^{\text{freq}}\left(\boldsymbol{w}_i\right), \beta^{\text{time}}\left(\boldsymbol{w}_i\right)\right)'.$$

We estimate the model with the influence function approach using

$$\boldsymbol{w}_i := (age_i,\, income_i,\, who_i^1,\, who_i^2,\, who_i^3,\, luggage_i)'$$

as the set of input variables to the network. The variable *luggage* is an ordinal variable with information on the pieces of luggage a traveler carries on her trip. It is zero if the traveler carries no luggage, 1 if she carries one piece, and 3 if she carries several pieces.

As a benchmark, we estimate a conditional logit model and a nested logit model. In comparison to the conditional logit model, the nested logit allows for more realistic substitution patterns across alternatives (it does not exhibit the IIA property with respect to alternatives across nests). For the nested logit model, we follow Bierlaire et al. (2001) and group the alternatives car and train in one nest (representing existing alternatives), and Swissmetro in another nest (representing the newly introduced alternative).[16] For both models, we use the same utility specification as for the influence function approach, except that we model the coefficients as linear functions of the the input variables $\boldsymbol{w}_i$. More precisely, in addition to alternative-specific constants and the variables travel cost, frequency, and travel time, we include interactions of the alternative-specific constants and the variables of interest with each of the variables in $\boldsymbol{w}_i$.[17] Similarly to the Monte Carlo experiments, we also include a neural network estimated with the full training sample as a benchmark. For the neural network, we conduct naive inference using robust standard errors for the estimated coefficient functions. For the influence function

[16] Since the nest including the alternative Swissmetro is a degenerate nest, we estimate an unscaled version of the nested logit in order to make the identification of the dissimilarity parameter feasible (see, e.g., Heiss, 2002).

[17] Interacting the alternative-specific constants with $\boldsymbol{w}_i$ yields multinomial coefficients for each variable in $\boldsymbol{w}_i$.

approach and for the neural network approach with naive inference, we use the same network architectures as in the Monte Carlo experiment. In line with the results from the Monte Carlo experiments, we use repeated sample splitting with $R = 5$ repetitions and set $\lambda = 0$ in the network for the estimation of $\boldsymbol{\Lambda}(\boldsymbol{w}_i)$ when estimating the model with the influence function approach, as $\lambda > 0$ provides incorrect coverage of the confidence intervals in the Monte Carlo experiments.

For the estimation, we follow Sifringer et al. (2020) and split the $9,036$ observations into a training and a test set which consist of three and one quarter of the total observations, respectively. We use the test set to compare the out-of-sample performance of the influence function approach to the benchmark models. Table 4.4 reports the average heterogeneous coefficient functions for the travel cost, frequency, and travel time and their corresponding estimated standard errors. Additionally, we calculate the

Table 4.4: Estimated Average Travel Cost, Frequency and TRavel Time Parameters and Corresponding Estimated Standard Errors

	CL	NL	IFA	NN
$\hat{\theta}_{cost}$	-1.144	-1.418	-1.849	-1.943
$\hat{\theta}_{freq}$	-0.891	-0.966	-1.040	-1.106
$\hat{\theta}_{time}$	-1.368	-1.728	-1.797	-2.172
$\hat{se}_{cost}$	0.061	0.078	0.954	1.343
$\hat{se}_{freq}$	0.129	0.154	2.440	2.476
$\hat{se}_{time}$	0.085	0.099	2.119	1.375
LL^{Train}	-0.763	-0.762	-0.655	-0.638
LL^{Test}	-0.777	-0.772	-0.753	-0.695

Note: The table reports the estimated average coefficients and the standard errors three variables of interest, and the in- and out-of-sample log-likelihood for the conditional logit (CL), the nested logit (NL), the influence function approach with $\lambda = 0$ and repeated sample splitting with $R = 5$ (IFA), and the neural network with naive inference using robust standard errors (NN).

in- and out-of-sample log-likelihood per observation. Both the in- and out-of-sample log-likelihood increases with increasing flexibility of the estimation approach. While there is only a slight improvement when going from the conditional logit to the nested logit model, the influence function approach has a substantially higher in-sample as well as out-of-sample log-likelihood. With respected to the estimated average coefficients, all four estimators estimate the same sign. Travelers find alternatives with higher travel cost, frequency, and travel time less attractive.[18] The estimated average coefficients are smallest in magnitude when the model is estimated with the conditional logit model and increase in magnitude with increasing out-of-sample log-likelihood, which is especially the

[18]Frequency is calculated as average minutes of waiting time for a given transportation mode, i.e., a higher frequency variable implies less frequent connections.

case for the travel cost coefficient. The results for the estimated standard errors are in line with the results from the Monte Carlo experiments, as the estimated standard errors of the influence function approach are substantially lager than those of the conditional and nested logit model. In fact, for the influence function approach, none of the estimated average coefficients is significantly different from zero, highlighting that larger samples might be needed for the influence function approach than for traditional logit models.

Figure 4.5 plots the histograms of the estimated coefficient functions predicted by the influence function approach (blue bars) and the nested logit model (green bars) using the test set. First, the plots reveal that there is substantial heterogeneity across travelers. Second, the heterogeneity in the intercept for Swissmetro and in the coefficient for travel cost across travelers appears to be similar when estimated with the influence function approach and the nested logit model, implying that the heterogeneity can be well captured by the linear approximation employed by the nested logit model. In contrast, the heterogeneity in the intercept for train and the coefficients for frequency and travel time predicted by the more flexible influence function approach deviates to a larger extent from the heterogeneity predicted by the nested logit model.

One advantage of the influence function approach is that it can be easily applied to any parameter of inferential interest that is a function of the heterogeneous coefficient functions. In addition to the estimated average coefficient for the travel time, travel cost, and frequency, we are interested in estimating mean elasticities. More precisely, we focus on the expected own- and cross-travel time elasticities with respect to changes in the travel time evaluated at the mean values of travel cost, frequency, and travel time of every alternative. Thus, the parameters of inferential interest calculated with the influence function approach are

$$\theta_0^{l,m} = E\left[H^{l,m}(\boldsymbol{w_i}, \boldsymbol{\delta}(\boldsymbol{w}_i); \boldsymbol{x}^*)\right]$$

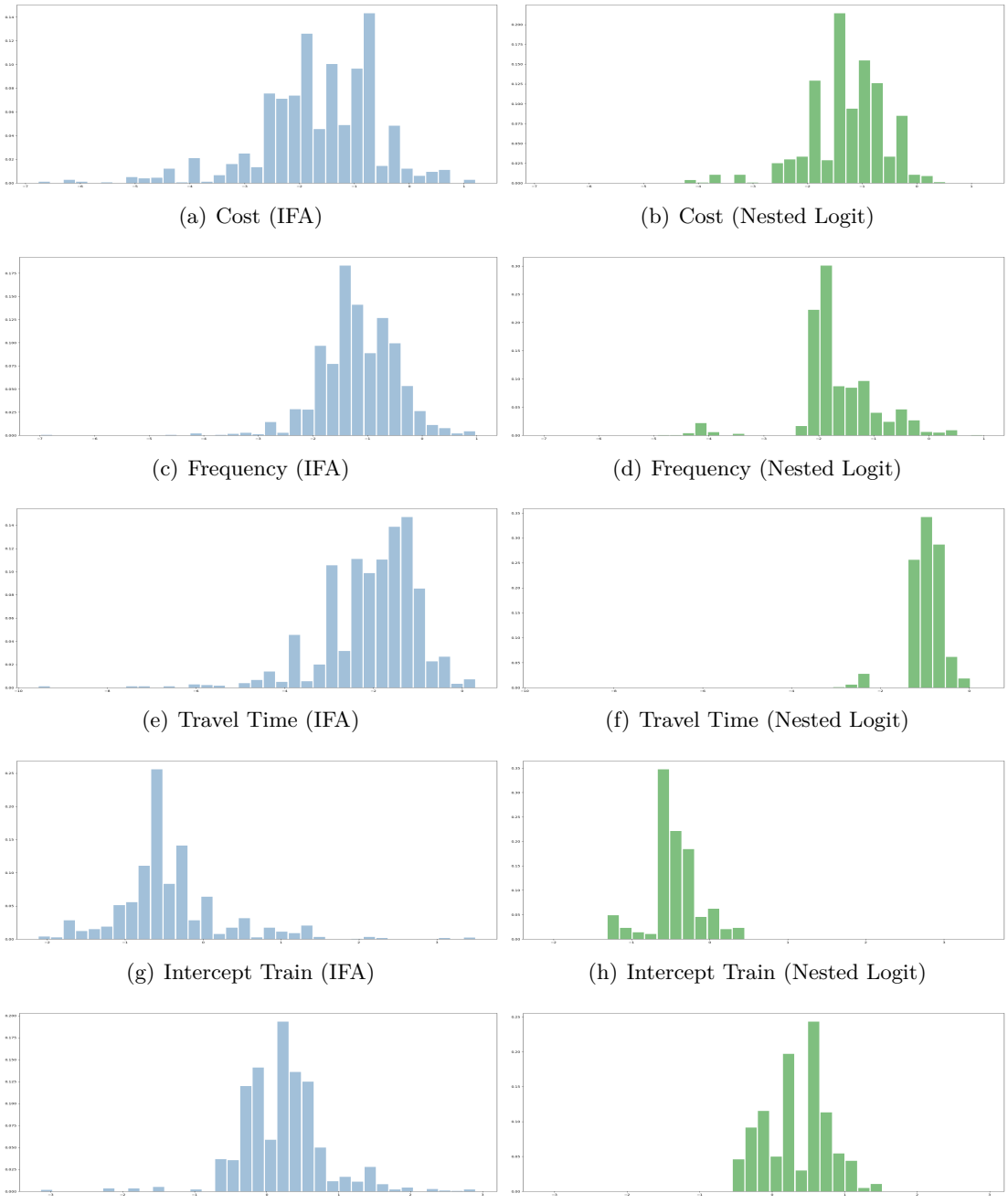
where $\boldsymbol{x}^*$ is a matrix with row entries $\bar{\boldsymbol{x}}_j'$ which contain the average travel time, travel cost and frequency for alternative $j \in \{\text{car}, \text{train}, \text{sm}\}$, and

$$H^{l,m}(\boldsymbol{w_i}, \boldsymbol{\delta}(\boldsymbol{w}_i); \boldsymbol{x}^*) = \beta^{\text{time}}(\boldsymbol{w}_i)\ \bar{x}_{m,\text{time}}\left(\mathbb{I}_{m,l} - \mathbb{P}\left(y_{i,m} = 1 | \boldsymbol{x}^*, \boldsymbol{w}_i\right)\right)$$

where $\mathbb{I}_{l,m}$ is an indicator that is equal to one when l is equal to m and zero otherwise for $l, m \in \{\text{car}, \text{train}, \text{sm}\}$.

Hence, $H^{l,m}(\boldsymbol{w_i}, \boldsymbol{\delta}(\boldsymbol{w}_i); \boldsymbol{x}^*)$ is the individual own- and cross-travel time elasticity calculated at the average travel cost, frequency, and travel time of every alternative, indicating the percentage change of choosing alternative l after a one percentage increase in the average travel time of alternative m. Consequently, $\theta_0^{l,m}$ corresponds to the

Figure 4.5: Histograms of Estimated Coefficient Functions for Influence Function Approach and Nested Logit Model



(a) Cost (IFA) (b) Cost (Nested Logit)

(c) Frequency (IFA) (d) Frequency (Nested Logit)

(e) Travel Time (IFA) (f) Travel Time (Nested Logit)

(g) Intercept Train (IFA) (h) Intercept Train (Nested Logit)

(i) Intercept SM (IFA) (j) Intercept SM (Nested Logit)

Note: The green bars represent the heterogeneous coefficients in the test set predicted with the nested logit model, and the blue bars the heterogeneous coefficients in the test set predicted with the influence function approach (IFA) with repeated sample splitting with $R = 5$.

expected own- and cross-travel time elasticity across individuals.

For the conditional logit, nested logit, and naive neural network approach, we use Efron's Bootstrap (Efron, 1979) with 1000 bootstraps iterations to calculate the estimated standard errors of the own- and cross-travel time elasticities evaluated at the means.[19]

Table 4.5: Estimated Own- & Cross-Travel Time Elasticities

	Influence Function:			**Neural Network:**		
	Car	SM	Train	Car	SM	Train
Car	-2.385 (0.274)	1.338 (0.167)	0.143 (0.105)	-2.313 (0.172)	1.315 (0.098)	0.237 (0.039)
SM	0.605 (0.274)	-0.463 (0.167)	0.143 (0.105)	0.876 (0.066)	-0.670 (0.054)	0.237 (0.039)
Train	0.605 (0.274)	1.338 (0.167)	-3.466 (0.107)	0.876 (0.066)	1.315 (0.098)	-3.526 (0.231)
	Conditional Logit:			**Nested Logit:**		
	Car	SM	Train	Car	SM	Train
Car	-1.71 (0.388)	0.791 (0.127)	0.211 (0.265)	-0.936 (0.051)	0.715 (0.061)	0.145 (0.02)
SM	0.559 (0.327)	-0.46 (0.126)	0.211 (0.265)	0.398 (0.023)	-0.45 (0.032)	0.075 (0.01)
Train	0.559 (0.327)	0.791 (0.127)	-1.83 (0.254)	1.097 (0.288)	0.715 (0.061)	-1.255 (0.068)

Note: The table reports estimated mean and the standard errors (in brackets) over individuals' own- and cross-travel time elasticities evaluated at the mean for the influence function approach, the neural network, the conditional logit, and the nested logit model. The reported numbers correspond to the percentage change of the choice probability of an alternative in a row after a one percent increase in the travel time of an alternative in a column.

Overall, the own- and cross-travel time elasticities estimated with the influence function approach and the neural network are quite similar. With respect to the own-travel time elasticities, both the influence function approach and the neural network predict that travelers respond more sensitively to an increase in the travel time than predicted by the conditional and nested logit model.

A disadvantage of the influence function approach, the neural network, and the conditional logit model is the restriction of the cross-elasticities through the IIA property imposed by the conditional logit model and the model specified in Equation (4.1), which restricts the cross-elasticities to be identical across alternatives. In contrast, the nested logit model, which allows for different cross-elasticities across alternatives in different nests, predicts that travelers are substantially more likely to substitute from car to train and vice versa in response to an increase in the travel time of either of the alternatives.

Moreover, the standard errors of the own- and cross-travel time elasticities estimated with the influence function approach remain larger than those of the nested logit model estimated with Efron's bootstrap – though the difference is not as large as for the

[19]For the nested logit model, we estimate the own- and cross-travel time elasticities at the mean using numerical derivatives of the choice probabilities with respect to the travel time.

estimated average coefficients – and are only slightly larger than in the conditional logit model and even smaller for some own- and cross-elasticities.

4.5 Conclusion

This paper investigates the finite sample performance of the estimation approach of Farrell et al. (2021*a*) in the context of discrete choice models, who propose deep learning for the estimation of heterogeneous parameters in econometric models. For the construction of valid second-stage inferential statements after the first-stage estimation of the heterogeneous parameters with deep learning, they provide an influence function approach that builds on Neyman orthogonal scores in combination with sample splitting.

To study the proposed estimation and inference procedure, we conduct several Monte Carlo experiments. First, the experiments reveal that deep learning generally allows to recover precise estimates of the true average heterogeneous parameters – especially if the number of observations is sufficiently large – and that naive inference with robust standard errors leads to incorrect inferential statements. Second, we observe that the influence function approach proposed for the construction of valid inferential statements is sensitive to overfitting when no l_2-regularization is employed. Overfitting results in substantial average estimated bias and extremely large average estimated standard errors across Monte Carlo replicates. The sensitivity to overfitting is more pronounced for small samples but does not disappear with increasing sample size in our experiments. Using l_2-regularization appears to stabilize the estimation as it reduces the number of Monte Carlo replicates with extreme outliers, but leads to poor coverage of the confidence intervals. This is a consequence of the decreasing magnitude of the estimated standard errors and the increasing bias induced with increasing regularization, which in combination lead to tighter confidence intervals that are centered around biased estimates. A tool that achieves substantially better results in our Monte Carlo experiments than regularization is repeated sample splitting. Unlike l_2-regularization, it substantially reduces the number of outliers across Monte Carlo replicates without inducing additional bias, enabling the construction of valid inferential statements. However, repeated sample splitting appears to have a less drastic effect on the estimated variance than l_2-regularization, which causes relatively large estimated standard errors.

Due to the complexity of neural networks, we restrict our Monte Carlo experiments to the impact of l_2-regularization on the inference procedure. An interesting avenue for future research is to consider different forms of regularization, such as dropout rates, and varying complexities of the network architecture used to estimate the influence function approach (e.g., to vary the number of neurons and hidden layers). In addition, both the influence function approach and the neural network combined with naive inference exploit that the variables of interest enter the utility linearly. An interesting comparison, however, is the estimation with a completely unstructured network (e.g., with Efron's bootstrap for inference) which could potentially further illustrate the advantage of the influence function approach.

Appendix 4.A Additional Tables

Table 4.A.1: Median Summary Statistics of 1000 Monte Carlo Replicates for Small Data and without Repeated Sample Splitting

	Conditional Logit		Influence Function Approach with λ equal to				
	Oracle	Basic	0	10^{-5}	10^{-4}	$2 \cdot 10^{-3}$	NN
$\theta_{cost} \in \widehat{CI}_{cost}$	0.95	0.00	0.93	0.92	0.83	0.40	0.99
$\theta_{freq} \in \widehat{CI}_{freq}$	0.95	0.00	0.93	0.92	0.89	0.68	1.00
$\theta_{time} \in \widehat{CI}_{time}$	0.94	0.00	0.93	0.94	0.88	0.54	1.00
$\widehat{se}_{cost}$	0.07	0.05	1.36	1.02	0.53	0.18	0.60
$\widehat{se}_{freq}$	0.10	0.07	1.62	1.34	0.84	0.34	3.29
$\widehat{se}_{time}$	0.07	0.06	1.28	0.96	0.46	0.24	2.98
Bias_{cost}	-0.01	0.65	-0.23	-0.21	-0.31	-0.44	-0.16
Bias_{freq}	-0.00	0.59	-0.28	-0.32	-0.35	-0.40	-0.19
Bias_{time}	-0.01	0.80	-0.09	-0.09	-0.23	-0.41	-0.17
Rej. $\theta_{cost} = 0$	1.00	1.00	0.47	0.56	0.78	0.93	1.00
Rej. $\theta_{freq} = 0$	1.00	1.00	0.27	0.35	0.50	0.79	0.00
Rej. $\theta_{time} = 0$	1.00	1.00	0.60	0.69	0.84	0.93	0.03
$MSE(\Lambda)^{Train}$	.	.	4.99	5.13	5.35	5.93	.
$MSE(\Lambda)^{Test}$	.	.	5.20	5.28	5.40	5.93	.
Share Outlier	0.00	0.00	0.26	0.18	0.11	0.04	0.12

Note: The table reports the median of the variables $\hat{se}_i$, $BIAS_i$, $MSE(\Lambda)^{Train}$, and $MSE(\Lambda)^{Test}$ and the average of the variables $\theta_i \in \hat{CI}_i$, Rej. $\theta_i = 0$, and Share Outlier, $i \in \{cost, freq, time\}$, over all Monte Carlo replicates for the conditional logit using the true specification (Oracle), the conditional logit using the three variables of interest for the estimation (Basic), the influence function approach, using five different values of λ for the estimation of $\Lambda_s(\boldsymbol{w})$, and the neural network (NN), which uses robust standard erros and does not rely on the influence function approach.

Table 4.A.2: Median Summary Statistics of 1000 Monte Carlo Replicates for Small Data and Repeated Sample Splitting with $R = 5$

	Conditional Logit		Influence Function Approach with λ equal to				
	Oracle	Basic	0	10^{-5}	10^{-4}	$2 \cdot 10^{-3}$	NN
$\theta_{cost} \in \widehat{CI}_{cost}$	0.94	0.00	0.94	0.93	0.82	0.42	0.99
$\theta_{freq} \in \widehat{CI}_{freq}$	0.96	0.00	0.94	0.92	0.90	0.66	1.00
$\theta_{time} \in \widehat{CI}_{time}$	0.96	0.00	0.95	0.95	0.87	0.59	1.00
$\widehat{se}_{cost}$	0.07	0.05	1.20	0.96	0.46	0.18	0.60
$\widehat{se}_{freq}$	0.10	0.07	1.48	1.16	0.73	0.33	3.45
$\widehat{se}_{time}$	0.07	0.06	1.21	0.85	0.42	0.24	3.04
Bias_{cost}	-0.01	0.65	-0.26	-0.29	-0.33	-0.41	-0.18
Bias_{freq}	-0.00	0.59	-0.28	-0.31	-0.28	-0.40	-0.18
Bias_{time}	-0.00	0.81	-0.03	-0.13	-0.24	-0.38	-0.17
Rej. $\theta_{cost} = 0$	1.00	1.00	0.48	0.61	0.84	0.96	0.99
Rej. $\theta_{freq} = 0$	1.00	1.00	0.25	0.32	0.54	0.81	0.00
Rej. $\theta_{time} = 0$	1.00	1.00	0.64	0.78	0.92	0.96	0.02
$MSE(\Lambda)^{Train}$	.	.	5.01	5.15	5.37	5.94	.
$MSE(\Lambda)^{Test}$	.	.	5.22	5.30	5.43	5.95	.
Share Outlier	0.00	0.00	0.05	0.02	0.00	0.00	0.12

Note: The table reports the median of the variables $\hat{se}_i$, $BIAS_i$, $MSE(\Lambda)^{Train}$, and $MSE(\Lambda)^{Test}$ and the average of the variables $\theta_i \in \hat{CI}_i$, Rej. $\theta_i = 0$, and Share Outlier, $i \in \{cost, freq, time\}$, over all Monte Carlo replicates for the conditional logit using the true specification (Oracle), the conditional logit using the three variables of interest for the estimation (Basic), the influence function approach, using five different values of λ for the estimation of $\Lambda_s(\boldsymbol{w})$, and the neural network (NN), which uses robust standard erros and does not rely on the influence function approach.

Table 4.A.3: Average Summary Statistics of 1000 Monte Carlo Replicates for Large Data and without Repeated Sample Splitting

	Conditional Logit		Influence Function Approach with λ equal to				
	Oracle	Basic	0	10^{-5}	10^{-4}	$2 \cdot 10^{-3}$	NN
$\theta_{cost} \in \widehat{CI}_{cost}$	0.95	0.00	0.94	0.93	0.90	0.75	1.00
$\theta_{freq} \in \widehat{CI}_{freq}$	0.95	0.00	0.92	0.94	0.91	0.83	1.00
$\theta_{time} \in \widehat{CI}_{time}$	0.95	0.00	0.94	0.94	0.92	0.85	1.00
$\widehat{se}_{cost}$	0.02	0.02	3.43	1.45	1.15	0.24	0.49
$\widehat{se}_{freq}$	0.04	0.04	8.24	2.03	1.60	0.30	1.73
$\widehat{se}_{time}$	0.03	0.02	3.02	3.24	0.98	0.26	1.28
Bias_{cost}	-0.00	0.60	1.38	0.87	0.05	-0.24	-0.03
Bias_{freq}	-0.00	0.53	3.82	1.08	0.24	-0.32	-0.03
Bias_{time}	-0.00	0.78	0.67	3.13	-0.01	-0.24	-0.02
Rej. $\theta_{cost} = 0$	1.00	1.00	0.83	0.88	0.89	0.98	1.00
Rej. $\theta_{freq} = 0$	1.00	1.00	0.58	0.69	0.81	0.97	0.00
Rej. $\theta_{time} = 0$	1.00	1.00	0.90	0.93	0.94	0.98	0.93
$MSE(\Lambda)^{Train}$	.	.	8.81	8.85	8.90	9.23	.
$MSE(\Lambda)^{Test}$	.	.	8.89	8.88	8.92	9.24	.
Share Outlier	0.00	0.00	0.07	0.05	0.04	0.01	0.00

Note: The table reports the average summary statistics over all Monte Carlo replicates for the conditional logit using the true specification (Oracle), the conditional logit using the three variables of interest for the estimation (Basic), the influence function approach, using five different values of λ for the estimation of $\Lambda_s(\boldsymbol{w})$, and the neural network (NN), which uses robust standard erros and does not rely on the influence function approach.

Table 4.A.4: Median Summary Statistics of 1000 Monte Carlo Replicates for Large Data and without Repeated Sample Splitting

	Conditional Logit		Influence Function Approach with λ equal to				
	Oracle	Basic	0	10^{-5}	10^{-4}	$2 \cdot 10^{-3}$	NN
$\theta_{cost} \in \widehat{CI}_{cost}$	0.95	0.00	0.94	0.93	0.90	0.75	1.00
$\theta_{freq} \in \widehat{CI}_{freq}$	0.95	0.00	0.92	0.94	0.91	0.83	1.00
$\theta_{time} \in \widehat{CI}_{time}$	0.95	0.00	0.94	0.94	0.92	0.85	1.00
$\widehat{se}_{cost}$	0.02	0.02	0.35	0.25	0.19	0.04	0.49
$\widehat{se}_{freq}$	0.04	0.04	0.58	0.37	0.23	0.06	1.72
$\widehat{se}_{time}$	0.03	0.02	0.27	0.19	0.16	0.05	1.27
Bias_{cost}	-0.00	0.60	-0.05	-0.01	-0.07	-0.05	-0.03
Bias_{freq}	-0.00	0.53	-0.09	-0.04	-0.08	-0.06	-0.03
Bias_{time}	-0.00	0.78	-0.00	0.00	-0.04	-0.03	-0.02
Rej. $\theta_{cost} = 0$	1.00	1.00	0.83	0.88	0.89	0.98	1.00
Rej. $\theta_{freq} = 0$	1.00	1.00	0.58	0.69	0.81	0.97	0.00
Rej. $\theta_{time} = 0$	1.00	1.00	0.90	0.93	0.94	0.98	0.93
$MSE(\Lambda)^{Train}$	.	.	8.81	8.85	8.90	9.24	.
$MSE(\Lambda)^{Test}$	.	.	8.89	8.88	8.91	9.24	.
Share Outlier	0.00	0.00	0.07	0.05	0.04	0.01	0.00

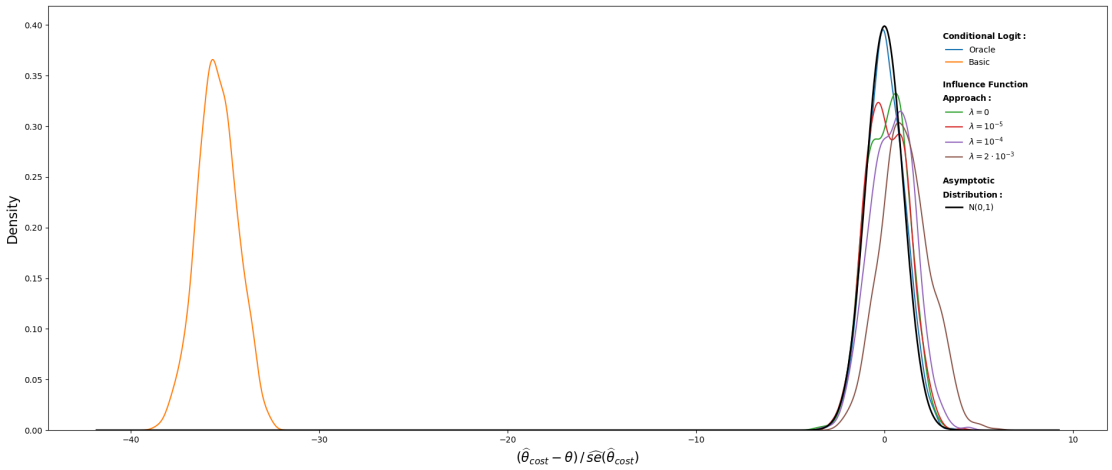
Note: The table reports the median of the variables $\hat{se}_i$, $BIAS_i$, $MSE(\Lambda)^{Train}$, and $MSE(\Lambda)^{Test}$ and the average of the variables $\theta_i \in \hat{CI}_i$, Rej. $\theta_i = 0$, and Share Outlier, $i \in \{cost, freq, time\}$, over all Monte Carlo replicates for the conditional logit using the true specification (Oracle), the conditional logit using the three variables of interest for the estimation (Basic), the influence function approach, using five different values of λ for the estimation of $\Lambda_s(\boldsymbol{w})$, and the neural network (NN), which uses robust standard erros and does not rely on the influence function approach.

Table 4.A.5: Median Summary Statistics of 1000 Monte Carlo Replicates for Large Data and Repeated Sample Splitting with $R = 5$

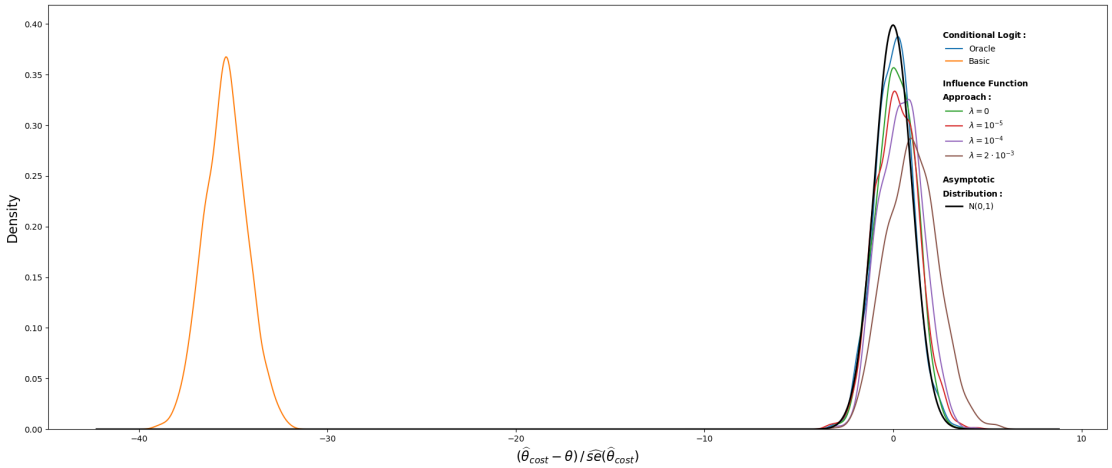
	Conditional Logit		Influence Function Approach with λ equal to				
	Oracle	Basic	0	10^{-5}	10^{-4}	$2 \cdot 10^{-3}$	NN
$\theta_{cost} \in \widehat{CI}_{cost}$	0.95	0.00	0.94	0.92	0.91	0.76	1.00
$\theta_{freq} \in \widehat{CI}_{freq}$	0.95	0.00	0.95	0.93	0.90	0.83	1.00
$\theta_{time} \in \widehat{CI}_{time}$	0.94	0.00	0.95	0.94	0.92	0.86	1.00
$\widehat{se}_{cost}$	0.02	0.02	0.30	0.23	0.17	0.04	0.49
$\widehat{se}_{freq}$	0.04	0.04	0.51	0.33	0.22	0.06	1.72
$\widehat{se}_{time}$	0.03	0.02	0.26	0.18	0.14	0.05	1.26
Bias_{cost}	-0.00	0.60	-0.04	-0.03	-0.06	-0.04	-0.03
Bias_{freq}	0.00	0.53	-0.08	-0.06	-0.08	-0.05	-0.02
Bias_{time}	-0.00	0.78	-0.01	-0.00	-0.03	-0.03	-0.02
Rej. $\theta_{cost} = 0$	1.00	1.00	0.89	0.91	0.93	0.98	1.00
Rej. $\theta_{freq} = 0$	1.00	1.00	0.62	0.75	0.84	0.96	0.00
Rej. $\theta_{time} = 0$	1.00	1.00	0.95	0.96	0.97	0.99	0.94
$MSE(\Lambda)^{Train}$	.	.	8.80	8.83	8.89	9.22	.
$MSE(\Lambda)^{Test}$	.	.	8.87	8.87	8.90	9.23	.
Share Outlier	0.00	0.00	0.00	0.00	0.00	0.00	0.00

Note: The table reports the median of the variables $\hat{se}_i$, $BIAS_i$, $MSE(\Lambda)^{Train}$, and $MSE(\Lambda)^{Test}$ and the average of the variables $\theta_i \in \hat{CI}_i$, Rej. $\theta_i = 0$, and Share Outlier, $i \in \{cost, freq, time\}$, over all Monte Carlo replicates for the conditional logit using the true specification (Oracle), the conditional logit using the three variables of interest for the estimation (Basic), the influence function approach, using five different values of λ for the estimation of $\Lambda_s(\boldsymbol{w})$, and the neural network (NN), which uses robust standard erros and does not rely on the influence function approach.

Figure 4.A.1: Density of Estimated t-Statistic of $\hat{\theta}_{cost}$ for Different Estimators and Large Data



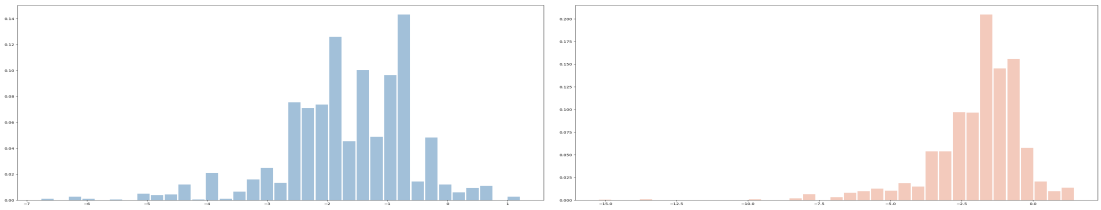
(a) No repeated sample splitting ($R = 1$)



(b) Repeated sample splitting ($R = 5$)

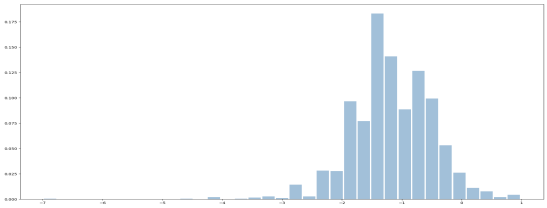
Note: The plot shows kernel density estimates of the estimated t-statistic for the conditional logit using the true specification (Oracle), the conditional logit using the three variables of interest for the estimation (Basic), the influence function approach, using four different values for λ for the estimation of $\Lambda_s(\boldsymbol{w})$. Additionally, the standard normal distribution is included.

Figure 4.A.2: Histograms of Estimated Coefficient Functions for Influence Function Approach and Neural Network

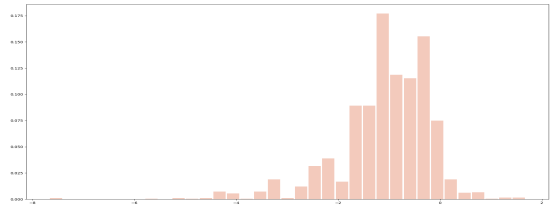


(a) Cost (IFA)

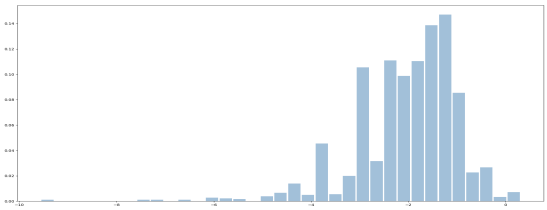
(b) Cost (NN)



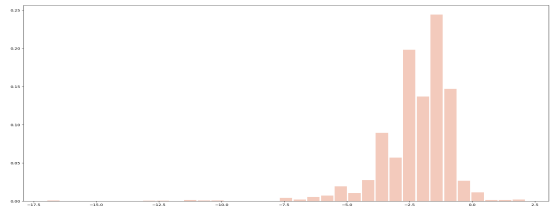
(c) Frequency (IFA)



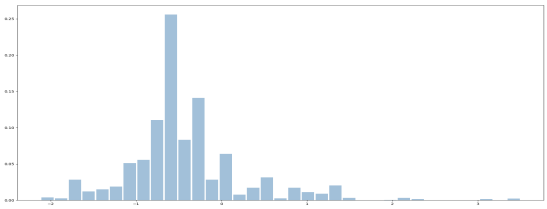
(d) Frequency (NN)



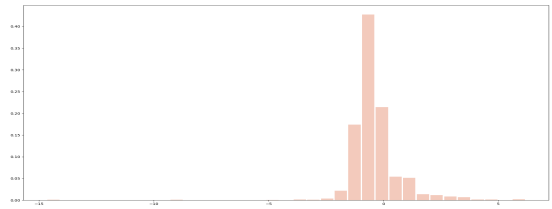
(e) Travel Time (IFA)



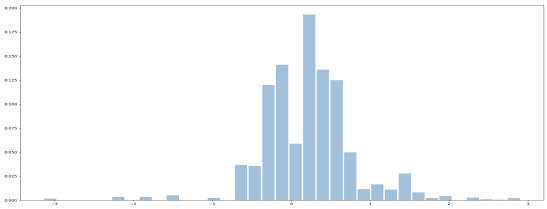
(f) Travel Time (NN)



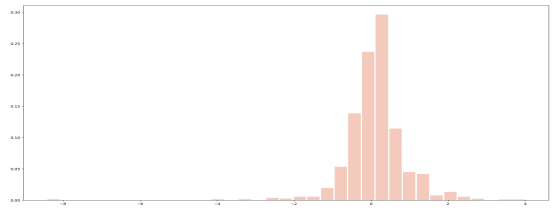
(g) Intercept Train (IFA)



(h) Intercept Train (NN)



(i) Intercept SM (IFA)



(j) Intercept SM (NN)

Note: The orange bars represent the heterogeneous coefficients in the test set predicted with the neural network (NN) model (without the influence function approach), and the blue bars the heterogeneous coefficients in the test set predicted with the influence function approach (IFA) with repeated sample splitting with $R = 5$.

Table 4.A.6

	Conditional Logit:		Nested Logit:	
	(1)	(2)	(1)	(2)
Const SM	1.248*** (0.183)	0.721* (0.424)	1.490*** (0.225)	1.109** (0.548)
Const Train	−1.110*** (0.417)	−0.191 (0.748)	−1.013** (0.411)	−0.007 (0.971)
Cost	−0.878*** (0.042)	−0.984** (0.429)	−0.976*** (0.046)	−1.258** (0.500)
Freq	−0.735*** (0.115)	−2.307* (1.190)	−0.778*** (0.122)	−2.603 (1.918)
Time	−1.216*** (0.051)	−2.710*** (0.586)	−1.449*** (0.048)	−3.138*** (0.657)
Age_{sm}	−0.234*** (0.030)	−0.198*** (0.045)	−0.262*** (0.036)	−0.262*** (0.059)
$\text{AGE}_{\text{train}}$	0.040 (0.047)	0.020 (0.087)	0.035 (0.046)	0.003 (0.088)
$\text{Income}_{\text{sm}}$	0.015 (0.030)	−0.009 (0.043)	0.036 (0.034)	0.006 (0.051)
$\text{Income}_{\text{train}}$	−0.279*** (0.041)	−0.150* (0.079)	−0.288*** (0.043)	−0.164* (0.087)
Who1_{sm}	−0.347** (0.161)	−0.012 (0.408)	−0.430** (0.197)	−0.198 (0.530)
$\text{Who1}_{\text{train}}$	1.305*** (0.390)	0.029 (0.714)	1.317*** (0.402)	−0.030 (0.957)
Who2_{sm}	0.047 (0.166)	0.497 (0.415)	0.024 (0.200)	0.448 (0.536)
$\text{Who2}_{\text{train}}$	1.160*** (0.398)	0.080 (0.730)	1.175*** (0.411)	0.062 (0.971)
Who3_{sm}	−0.072 (0.181)	0.904** (0.426)	−0.128 (0.214)	0.875 (0.547)
$\text{Who3}_{\text{train}}$	1.199*** (0.418)	−0.437 (0.762)	1.184*** (0.431)	−0.644 (0.998)
Male_{sm}	−0.322*** (0.077)	−0.302*** (0.111)	−0.327*** (0.084)	−0.354*** (0.137)
$\text{Male}_{\text{train}}$	−0.428*** (0.115)	−0.206 (0.213)	−0.423*** (0.114)	−0.133 (0.219)
$\text{Luggage}_{\text{sm}}$	0.132** (0.052)	0.211*** (0.076)	0.129** (0.058)	0.214** (0.102)
$\text{Luggage}_{\text{train}}$	0.541*** (0.079)	0.350** (0.144)	0.562*** (0.088)	0.346** (0.165)
Cost*Age		−0.429*** (0.050)		−0.531*** (0.047)
Freq*Age		0.088 (0.113)		0.089 (0.115)
Time*Age		−0.065 (0.055)		−0.127** (0.050)
Cost*Income		0.098** (0.042)		0.098* (0.052)
Freq*Income		−0.153 (0.098)		−0.134 (0.109)
Time*Income		−0.085 (0.054)		−0.116* (0.070)
Cost*Who1		1.018** (0.419)		1.362*** (0.494)
Freq*Who1		1.747 (1.154)		1.961 (1.905)
Time*Who1		1.739*** (0.568)		2.156*** (0.661)
Cost*Who2		1.028** (0.420)		1.327*** (0.495)
Freq*Who2		1.543 (1.171)		1.740 (1.916)
Time*Who2		1.768*** (0.574)		2.141*** (0.671)
Cost*Who3		1.234*** (0.433)		1.582*** (0.519)
Freq*Who3		1.779 (1.208)		2.060 (1.939)
Time*Who3		3.099*** (0.574)		3.837*** (0.673)
Cost*MALE		−0.536*** (0.097)		−0.644*** (0.119)
Freq*Male		−0.053 (0.277)		−0.124 (0.285)
Time*Male		−0.394*** (0.139)		−0.601*** (0.156)
Cost*Luggage		0.399*** (0.075)		0.525*** (0.096)
Freq*Luggage		0.081 (0.189)		0.095 (0.230)
Time*Luggage		0.459*** (0.093)		0.611*** (0.121)
iv:train			0.805*** (0.039)	0.738*** (0.048)
iv:car			0.872*** (0.039)	0.761*** (0.048)
Observations	7,234	7,234	7,234	7,234
R^2	0.123	0.148	0.124	0.150
Log Likelihood	−5,683.250	−5,520.814	−5,676.610	−5,512.196
LR Test	1,599.627*** (df = 19)	1,924.500*** (df = 40)	1,612.908*** (df = 21)	1,941.736*** (df = 42)

Note: *p<0.1; **p<0.05; ***p<0.01

5 Block-Recursive Non-Gaussian Structural Vector Autoregressions: Identification, Efficiency, and Moment Selection

Co-authored by Sascha Alexander Keweloh

Abstract

We combine block-recursive restrictions with higher-order moment conditions to identify and estimate non-Gaussian structural vector autoregressions. For a given block-recursive structure, we derive a set of identifying moment conditions based on the assumption of uncorrelated shocks across blocks and mean independent shocks within the blocks. We then obtain overidentifying moment conditions from the assumption of independent shocks and show that these conditions can decrease the asymptotic variance of the estimator. In particular, we derive conditions under which the frequently applied estimator based on the Cholesky decomposition is inefficient. We use a LASSO-type GMM estimator to select the relevant and valid overidentifying moment conditions in a data-driven way. A Monte Carlo experiment illustrates the improved performance of the proposed estimator. In the empirical illustration, we take advantage of the block-recursive framework to analyze the impact of speculative shocks in the oil market.

JEL codes:	*C32, Q43*
Keywords:	*SVAR, Identification, Non-Gaussianity, GMM, Block-Recursive, LASSO, Moment Selection.*
Publication status:	*Submitted. Earlier version available as SFB 823 Discussion Paper No. 26/2021: http://dx.doi.org/10.17877/DE290R-22447.*

5.1 Introduction

Identification of a structural vector autoregression (SVAR) requires to assume an a priori structure of the model. Traditionally, identification is based on imposing structure on the interaction of the variables, ideally derived from macroeconomic theory (e.g., short-run restrictions Sims (1980) or long-run restrictions Blanchard and Quah (1989)). However, uncontroversial theoretical restrictions are rare. More recently, data-driven approaches allow to identify the SVAR without imposing any restrictions on the interaction. Instead, identification is achieved by imposing structure on the stochastic properties of the shocks (e.g., time-varying volatility as discussed in Rigobon (2003), Lanne, Lütkepohl and Maciejowska (2010), Lütkepohl and Netšunajev (2017), and Lewis (2021) or non-Gaussian and independent shocks as discussed in Gouriéroux, Monfort and Renne (2017), Lanne et al. (2017), Lanne and Luoto (2021), Keweloh (2021*b*), and Guay (2021)).

Traditional identification approaches may appear unnecessarily restrictive compared to novel data-driven approaches. However, Olea, Luis, Plagborg-Møller and Qian (2022) stress that these data-driven approaches rely on information in higher moments, while traditional approaches only rely on second moments. The data-driven approaches are sensitive to the imposed statistical properties on the higher moments, while the traditional approaches are not and hence, are robust to these statistical properties. Additionally, they argue that using economic theory for identification is a feature and not a handicap and conclude that traditional identification approaches remain relevant.

We agree with their reasoning and recognize the advantages of identification approaches based on economic theory. However, in many applications we can derive some, but not sufficiently many convincing restrictions from economic theory to ensure identification. Therefore, with a traditional purely restriction based approach, even the most plausible restrictions are worthless if there are not sufficiently many. We propose a Generalized Method of Moments (GMM) estimator that combines the traditional identification approach based on restrictions with the more recent data-driven approach based on non-Gaussianity. Our approach allows to impose a block-recursive structure, meaning that shocks in a given block only influence variables in the same block or blocks ordered below. The block-recursive structure seems plausible in many macroeconomic applications. Examples include applications analyzing (i) the interaction of macroeconomic and financial variables, where the former respond sluggishly while the latter respond quickly, or (ii) the interaction of small and large open economies, where large economies may have an immediate impact on small economics but not vice versa. Additionally, the block-recursive structure nests two important special cases: a recursive and an unrestricted SVAR.

Identification based on higher moments and non-Gaussian shocks oftentimes relies on the assumption of independent shocks which is criticized as too restrictive (see, e.g., Kilian and Lütkepohl (2017, Chapter 14)). Importantly, our identification result does not rely on independent shocks but is robust in the sense that it allows for various kinds of dependencies of the shocks. In particular, for a given block-recursive structure

identification of the shocks within a given block is based on a small (subset) of cokurtosis conditions derived from mean independence of the shocks in the corresponding block.[1] Therefore, identification within a block follows from Lanne and Luoto (2021). Moreover, the impact of the shocks in one block on variables in another block is identified based only on covariance conditions and not on higher-order moment conditions and requires only uncorrelated shocks. Therefore, imposing a finer block-recursive structure reduces the dependency of identification on higher-order moment conditions.

However, if the shocks are independent, using only the set of identifying conditions, which are derived from mean independent shocks within blocks and uncorrelated shocks across blocks, can be inefficient. To demonstrate this, we prove that in a recursive SVAR with independent shocks the set of overidentifying higher-order moment conditions can contain additional information and allows to decrease the asymptotic variance of the GMM estimator.[2] Efficient estimation requires to detect and select the valid and relevant overidentifying conditions. To this end, Lanne and Luoto (2021) suggest to calculate the information and moment selection criteria proposed by Andrews (1999) and Hall, Inoue, Jana and Shin (2007) for all possible combinations of moment conditions. However, they note that this approach becomes infeasible in higher-dimensional SVARs.

In a general GMM setup, Cheng and Liao (2015) propose a LASSO-type GMM estimator, hereafter referred to as the penalized GMM estimator (pGMM), which consistently selects only relevant and valid overidentifying conditions in a data-driven way. We apply the pGMM estimator to the block-recursive SVAR to exploit potential efficiency gains from overidentifying moment conditions. Our block-recursive SVAR pGMM estimator is consistent, asymptotically normal and as efficient as the asymptotically efficient block-recursive SVAR GMM estimator, including all valid and relevant overidentifying moment conditions. Importantly, these properties also hold if there are invalid overidentifying moment condition which could arise due to dependent structural shocks. Additionally, the pGMM estimator refrains from selecting valid but redundant overidentifying conditions which would neither increase nor decrease the asymptotic variance of the estimator but lead to imprecise estimates in small samples due to a many moments problem.

Guay (2021) also proposes to combine restrictions with non-Gaussian identification. In particular, he tests which shocks of the SVAR are identified based on non-Gaussianity and subsequently, his approach only uses restrictions to identify the remaining part of the SVAR. In this approach, if all shocks are non-Gaussian, no restrictions have to be used and the SVAR can be estimated solely by higher-order moment conditions. Consequently, the identification approach relies as heavily on non-Gaussianity as possible and as little

[1] A common critique to the assumption of independent shocks is that it does not allow for multiple shocks to be driven by the same volatility process. Thereby, it rules out a case which may be encountered for some macroeconomic shocks. However, mean independent shocks and, in particular, the set of cokurtosis conditions used for identification allow for these kinds of dependencies.

[2] Note that this is not trivial. For example, in a linear regression model $y_t = \beta_1 x_t + \epsilon_t$ the GMM estimator with the moment condition $E[x_t \epsilon_t] = 0$ is identified and efficient under (conditional) homoscedastic errors. Therefore, including additional higher-order moment conditions like $E[x_t^2 \epsilon_t] = 0$ does not decrease the asymptotic variance of the GMM estimator even if the shocks or variables are non-Gaussian.

on restrictions as necessary. In contrast to that, our identification approach relies as much as possible on economically justified restrictions and on non-Gaussianity only when needed.To be precise, the more block-recursiveness restrictions the researcher imposes, the less identification depends on higher order-moment conditions.

We conduct two Monte Carlo experiments. In the first one, we demonstrate that the performance of a purely data-driven estimator based on non-Gaussianity deteriorates substantially with both a decreasing sample size and an increasing model size. However, exploiting the block-recursive order can mitigate this performance decline. In the second Monte Carlo experiment, we illustrate that the pGMM estimator successfully selects relevant moment conditions and increases the finite sample performance compared to other block-recursive SVAR estimators for a given block-recursive structure.

We use the block-recursive SVAR pGMM estimator to analyze the impact of oil supply and oil demand shocks, including speculative oil supply and demand shocks, on the oil price. In his seminal work, Kilian (2009) highlights that it is necessary to distinguish between oil supply and demand shocks rather than including solely an oil price shock in the SVAR for the oil market. However, oil prices are not only affected by supply and demand shocks, but also by speculative shocks causing shifts in the expectations of forward-looking traders (see, e.g., Baumeister and Kilian (2016)). In particular, new oil production technologies, anticipated wars, or news about oil discoveries or about the (future) state of the economy can shift expectations of future oil supply and future oil demand. The studies of Kilian and Murphy (2014), Juvenal and Petrella (2015), Byrne, Lorusso and Xu (2019), and Moussa and Thomas (2021) extend the original oil market SVAR from Kilian (2009) to include speculative shocks. We contribute to this literature by explicitly distinguishing between speculative supply and speculative demand shocks.

The remainder of the paper is organized as follows: Section 5.2 reviews the SVAR and different identification schemes. Section 5.3 introduces the block-recursive SVAR. Section 5.4 derives identifying and overidentifying moment conditions in a block-recursive SVAR, analyzes which of the overidentifying conditions are redundant or relevant in a recursive SVAR, and describes the block-recursive SVAR GMM estimator and the pGMM estimator. In Section 5.5, we present the Monte Carlo experiments. In Section 5.6, we use the proposed block-recursive estimator to analyze the impact of flow and speculative supply and demand shocks in the oil market. Section 5.7 concludes.

5.2 Overview SVAR

This section briefly recalls the identification problem and common identification approaches for SVAR models. A detailed overview can be found in Kilian and Lütkepohl (2017). Consider the SVAR

$$y_t = \nu + A_1 y_{t-1} + ... + A_p y_{t-p} + B_0 \varepsilon_t,$$

with parameter matrices $A_1, ..., A_p \in \mathbb{R}^{n \times n}$, an intercept term $\nu \in \mathbb{R}^n$, an invertible matrix $B_0 \in \mathbb{R}^{n \times n}$, an n-dimensional vector of time series $y_t = [y_{1,t}, ..., y_{n,t}]'$ and an n-dimensional vector of serially uncorrelated structural shocks $\varepsilon_t = [\varepsilon_{1,t}, ..., \varepsilon_{n,t}]'$ with mean zero and unit variance. The parameter matrices $A_1, ..., A_p$ need to satisfy $det(I - A_1 c - ... - A_p c^p) \neq 0$ for $c \leq 1$ to ensure a stable process.

W.l.o.g. we focus on the simultaneous interaction of the SVAR given by

$$u_t = B_0 \varepsilon_t,$$

with the reduced form shocks $u_t = y_t - A_1 y_{t-1} - ... - A_p y_{t-p}$, which can be estimated consistently by OLS. The reduced form shocks are an unknown mixture B_0 of the unknown structural shocks ε_t. So far, neither the mixing matrix B_0 nor the structural shocks ε_t are identified. To see this, define the unmixed innovations $e(B)$ as the innovations obtained by unmixing the reduced form shocks with some matrix B

$$e_t(B) := B^{-1} u_t.$$

Note that for $B = B_0$, the unmixed innovations are equal to the structural shocks, i.e., $e_t(B_0) = \varepsilon_t$. Additionally, given an estimate $\hat{B}$ of B_0 we refer to $e_t(\hat{B})$ as the estimated structural shocks. The true structural shocks ε_t and the true mixing matrix B_0 are unknown and without imposing further structure, we cannot verify whether the mixing matrix B and the unmixed innovations $e_t(B)$ are equal to the true mixing matrix B_0 and the true structural shocks ε_t.

To identify B_0 and the shocks ε_t, the researcher has to impose structure on the SVAR. The structure can be specified in two ways: We may

(i) impose more structure on the interaction of the shocks (see Sims (1980) for short-run restrictions, Blanchard (1989) for long-run restrictions, and Uhlig (2005) for sign restrictions),

(ii) impose more structure on the stochastic properties of the structural shock (see Lanne et al. (2010) for time-varying volatility or Gouriéroux et al. (2017), Lanne et al. (2017), Lanne and Luoto (2021) Keweloh (2021*b*), and Guay (2021) for non-Gaussian shocks).

Imposing structure on the stochastic properties of the shocks can be used to derive conditions for the unmixed innovations, while imposing structure on the interaction narrows the space of possible mixing matrices used to unmix the reduced form shocks.

In applied work, the probably most frequently imposed structure are uncorrelated structural shocks (meaning $\varepsilon_{i,t}$ is restricted to be uncorrelated with $\varepsilon_{j,t}$ for $i \neq j$) and a recursive interaction (meaning restricting B_0 such that $b_{ij} = 0$ for $i < j$ where b_{ij} denotes the element at row i and column j of B_0). Uncorrelated shocks with unit variance can be used to derive $(n+1)n/2$ (co-)variance conditions from $I = E\left[\varepsilon_t \varepsilon_t'\right] \overset{!}{=} E\left[e_t(B) e_t(B)'\right]$. A recursive interaction implies that $n(n-1)/2$ parameters of B_0 are known a priori, leaving only $(n+1)n/2$ unknown parameters in the mixing matrix B. It is then straightforward to show that, if the remaining $(n+1)n/2$ parameters of the restricted B matrix generate unmixed innovations $e_t(B)$ which satisfy the $(n+1)n/2$ (co-)variance conditions, the matrix B has to be equal to B_0 and, hence, the unmixed innovations are equal to the structural shocks, meaning the SVAR is identified.[3]

However, economic theory rarely allows to derive the required $n(n-1)/2$ parameter restrictions to ensure identification. More recently, identification methods based on non-Gaussian and independent shocks have been put forward in the literature (see Gouriéroux et al. (2017), Lanne et al. (2017), Lanne and Luoto (2021), Keweloh (2021*b*), or Guay (2021)). These identification schemes do not require to impose any restrictions on the impact of the shocks, in particular on the matrix B_0. Instead, the researcher has to impose structure on the stochastic properties of the shocks. If the structural shocks are not only mutually uncorrelated but mutually independent, we can derive additional moment conditions. For example, independent and mean zero shocks imply that all entries of coskewness matrices $E\left[\varepsilon_t \varepsilon_t' \varepsilon_{i,t}\right]$ for $i = 1, \ldots, n$ are zero except for the ith diagonal element, which contains the (unknown) skewness of the shock $\varepsilon_{i,t}$. Hence, we can exploit that the mixing matrix B has to generate unmixed innovations, which satisfy the coskewness moment conditions derived from $E\left[\varepsilon_t \varepsilon_t' \varepsilon_{i,t}\right] \overset{!}{=} E\left[e_t(B) e_t(B)' e_{i,t}(B)\right]$. Similarly, we can use that the mixing matrix B has to generate unmixed innovations which satisfy the cokurtosis moment conditions derived from $E\left[\varepsilon_t \varepsilon_t' \varepsilon_{i,t} \varepsilon_{j,t}\right] \overset{!}{=} E\left[e_t(B) e_t(B)' e_{i,t}(B) e_{j,t}(B)\right]$.

5.3 Imposing structure in a SVAR

This section introduces the framework of the block-recursive SVAR. First, we discuss various structures of the interaction of the shocks allowed in this framework and then, assumptions on the stochastic properties of the shocks.

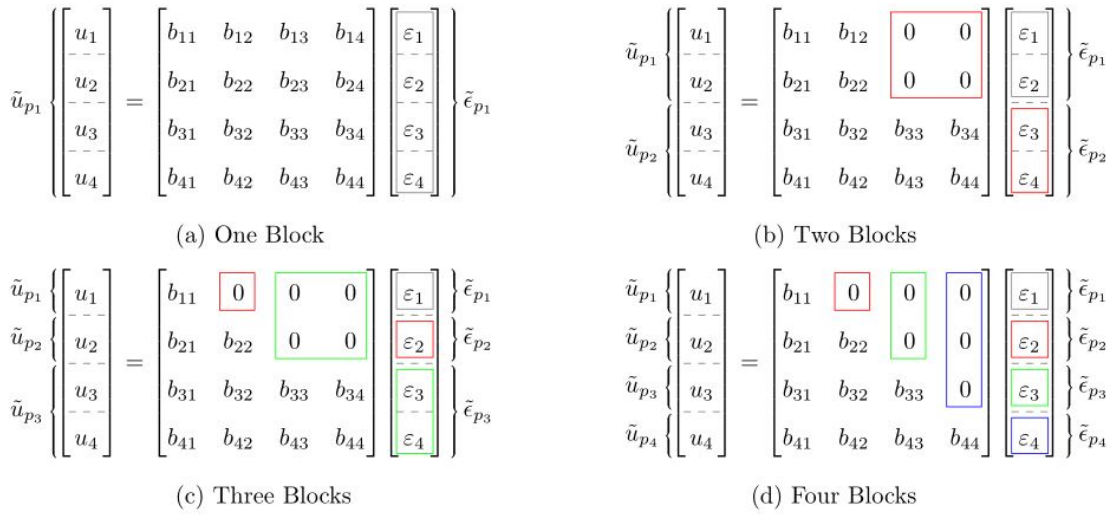
5.3.1 Imposing structure on the interaction of shocks

Traditionally, identification of a SVAR is based on the structure imposed on the interaction of the shocks (see Section 5.2). These restriction based approaches require restrictions on the interaction of the shocks to ensure identification, e.g., a recursive structure. The

[3] Note that this GMM approach is equivalent to the the frequently used estimator obtained by applying the Cholesky decomposition to the variance-covariance matrix of the reduced form shocks.

reasoning behind a recursive structure is oftentimes the prejudice that some variables, e.g., some macroeconomic variables like inflation, tend to move slowly, while other variables, e.g. financial variables like stock prices, react faster. However, in practice this intuitive reasoning oftentimes allows to order only some, but not all variables recursively. This motivates us to consider the block-recursive SVAR, meaning that the structural shocks are ordered in blocks of consecutive shocks and each structural shock can simultaneously affect all variables in the same block and in blocks ordered below but not variables in blocks ordered above.[4] Figure 5.1 shows different block-recursive structures in a SVAR with four variables. The examples show that a block-recursive structure generalizes the

Figure 5.1: Examples of Different Block-Recursive SVAR Models.



Note: The figure illustrate how the the block structure can be defined by the structural shocks and our definition of $\tilde{\varepsilon}_{p_i}$ and $\tilde{u}_{p_i}$, $i = 1, \ldots, m$.

unrestricted SVAR and the fully-recursive SVAR and includes both as extreme cases.

We now introduce the notation for the block-recursive SVAR. Suppose that the structural shocks can be ordered into $m \leq n$ blocks of consecutive shocks. Let the indices $p_1 = 1 < p_2 < \ldots < p_m \leq n$ denote the beginning of a new block and for a given block p_i let $\tilde{\varepsilon}_{p_i,t}$ and $\tilde{u}_{p_i,t}$ denote the vectors of all structural and reduced form shocks in the ith block, such that

$$\tilde{\varepsilon}_{p_i,t} := \left[\varepsilon_{p_i,t}, \varepsilon_{p_i+1,t}, \ldots, \varepsilon_{p_{i+1}-1,t}\right]' \quad \text{and} \quad \tilde{u}_{p_i,t} := \left[u_{p_i,t}, u_{p_i+1,t}, \ldots, u_{p_{i+1}-1,t}\right]',$$

where $p_{m+1} := n + 1$ for ease of notation. Moreover, let l_i denote the number of shocks in block i for $i = 1, ..., m$. The vector of all structural shocks ε_t can then be decomposed into the m blocks $\varepsilon_t = [\tilde{\varepsilon}'_{p_1,t}, \ldots, \tilde{\varepsilon}'_{p_m,t}]'$ and the reduced form shocks can be decomposed

[4] Zha (1999) derives identifying restrictions for the block-recursive SVAR. The author restricts not only the simultaneous interaction, but also the lagged interaction. Our proposed block-recursive structure affects only the simultaneous interaction, while the lagged interaction remains unrestricted.

analogously into $u_t = [\tilde{u}'_{p_1,t}, \ldots, \tilde{u}'_{p_m,t}]'$. The SVAR is block-recursive with $m \leq n$ blocks with $p_1 = 1 < p_2 < \ldots < p_m \leq n$, if shocks in the ith block have no simultaneous impact on reduced form shocks in blocks j with $j < i$ such that for $i = 1, \ldots, m$

$$b_{ql} = 0, \text{ for } l \geq p_i \text{ and } q < p_i.$$

Any block-recursive structure can be described by the following assumption.

Assumption 5.1. *(Block-recursive interaction.)*
For $m \leq n$ blocks with $p_1 = 1 < p_2 < \ldots < p_m \leq n$ and $q, l = 1, ..., n$ let
$B_0 \in \mathbb{B}_{brec} \equiv \mathbb{B}_{brec}(p_1, \ldots, p_m) := \{B \in \mathbb{R}^{n \times n} |\ det(B) \neq 0$ *and* $b_{ql} = 0$

if $\exists p_i \in \{p_1, ..., p_m\}$ *with* $l \geq p_i$ *and* $q < p_i\}$.

5.3.2 Imposing structure on the stochastic properties of shocks

Imposing structure according to Assumption 5.1 on the interaction is not sufficient to ensure identification and further assumptions on the dependence and potential non-Gaussianity of the shocks are required. In the following, we discuss different structures imposed on the mutual dependencies of the shocks.

Almost all identification approaches at least assume uncorrelated structural shocks such that $E[\varepsilon_{i,t}\varepsilon_{j,t}] = E[\varepsilon_{i,t}]\,E[\varepsilon_{j,t}]$ for $i \neq j$.[5] Uncorrelated shocks are justified by the idea that a given structural shock contains no information on other structural shocks, e.g., a structural monetary policy shock should not depend on other structural shocks. In general, imposing uncorrelated structural shocks does not rule out that the structural shocks are dependent. If they are dependent, the interpretation of the estimated SVAR via impulse response functions can be misleading. For example, consider the two random variables $\varepsilon_1 \sim \mathcal{N}(0, 1)$ and $\varepsilon_2 = \varepsilon_1^2 - 1$ such that both random variables are uncorrelated, but dependent. Policy analysis based on impulse response functions typically uses the ceteris paribus assumption that only a single shock varies, while the other shocks remain unchanged. In the example above, both shocks are uncorrelated, but nevertheless always move simultaneously. Therefore, uncorrelated structural shocks are not sufficient to guarantee that the ceteris paribus assumption holds.

A more rigorous implementation of the idea that a given shock contains no information on other shocks is to assume independent shocks such that $E[h(\varepsilon_{i,t})g(\varepsilon_{j,t})] = E[h(\varepsilon_{i,t})]\,E[g(\varepsilon_{j,t})]$ for $i \neq j$ and any bounded, measurable functions $g(\cdot)$ and $h(\cdot)$. If shocks are independent, a structural shock cannot contain any information on any other structural shock. Therefore, independent structural shocks justify the ceteris paribus interpretation used in policy analysis based on impulse response functions. However, several authors argue that the assumption of independent structural shocks is too strong

[5] Proxy-variable identification approaches are different and instead assume that structural shocks are uncorrelated with an external proxy variable (see, e.g., Stock and Watson (2012), or Mertens and Ravn (2013)).

(cf. Kilian and Lütkepohl (2017, Chapter 14), Lanne and Luoto (2021), Lanne, Liu and Luoto (2021), or Olea et al. (2022)). In particular, independence of the shocks implies that also the volatility processes of the shocks are independent, which may be too restrictive for some macroeconomic applications. For example, suppose that $\tilde{\varepsilon}_{1,t}$ and $\tilde{\varepsilon}_{2,t}$ are drawn independently of each other and represent unscaled structural shocks. Moreover, in each period an additional volatility shock v_t is drawn independently of the other shocks and the structural shocks are given by $\varepsilon_{1,t} = \tilde{\varepsilon}_{1,t}v_t$ and $\varepsilon_{2,t} = \tilde{\varepsilon}_{2,t}v_t$. These structural shocks are uncorrelated, but dependent since the variance of one shock contains information on the variance of the other shock.

A compromise between the two extreme cases of uncorrelated and independent shocks is the assumption of mean independent shocks, such that $E\left[\varepsilon_{i,t}g(\varepsilon_{j,t})\right] = E\left[\varepsilon_{i,t}\right]E\left[g(\varepsilon_{j,t})\right]$ for $i \neq j$ with a bounded, measurable function $g(\cdot)$. If shocks are mean independent, a structural shock cannot contain any information about the mean of other structural shocks. Mean independent shocks can justify the ceteris paribus assumption used in impulse response analysis and at the same time allow for dependent volatility processes. In particular, the two shocks $\varepsilon_{1,t} = \tilde{\varepsilon}_{1,t}v_t$ and $\varepsilon_{2,t} = \tilde{\varepsilon}_{2,t}v_t$ defined above are mean independent since a given shock contains no information on the mean of the other shock.

Imposing structure on the dependence of the structural shocks allows to derive moment conditions (see, e.g., Lanne and Luoto (2021), Keweloh (2021*b*), or Guay (2021)). For $i, j, k, l = 1, ..., n$ we define the following moment conditions:

$$\text{Variance:} \quad E[e(B)_{i,t}^2] - 1 = 0 \tag{5.1}$$

$$\text{Covariance:} \quad E[e(B)_{i,t}e(B)_{j,t}] = 0, \qquad \text{for } i < j \tag{5.2}$$

$$\text{Coskewness:} \quad E[e(B)_{i,t}^2 e(B)_{j,t}] = 0, \qquad \text{for } i \neq j \tag{5.3}$$

$$E[e(B)_{i,t}e(B)_{j,t}e(B)_{k,t}] = 0, \qquad \text{for } i < j < k \tag{5.4}$$

$$\text{Cokurtosis:} \quad E[e(B)_{i,t}^3 e(B)_{j,t}] = 0, \qquad \text{for } i \neq j \tag{5.5}$$

$$E[e(B)_{i,t}^2 e(B)_{j,t}e(B)_{k,t}] = 0, \qquad \text{for } i \neq j,\ i \neq k,\ j < k \tag{5.6}$$

$$E[e(B)_{i,t}e(B)_{j,t}e(B)_{k,t}e(B)_{l,t}] = 0, \qquad \text{for } i < j < k < l \tag{5.7}$$

$$E[e(B)_{i,t}^2 e(B)_{j,t}^2] - 1 = 0, \qquad \text{for } i < j \tag{5.8}$$

The variance conditions in Equation (5.1) follow from the unit variance normalization. The remaining conditions are derived from different assumptions on the dependence of the structural shocks. In particular, uncorrelated structural shocks only imply the covariance conditions in Equation (5.2). Mean independent shocks additionally imply the coskewness conditions in Equation (5.3) and (5.4) and the cokurtosis conditions in

Equation (5.5)-(5.7). In addition, the symmetric cokurtosis conditions in Equation (5.8) follow from independent shocks.

Moreover, note that if all structural shocks are Gaussian, the conditions in Equation (5.3)-(5.8) do not contain information beyond the information contained in the variance and covariance conditions.

5.4 Estimation of a block-recursive SVAR

In this section, we combine identification based on block-recursiveness restrictions and non-Gaussian shocks. First, for a given block-recursive structure we derive corresponding identifying asymmetric cokurtosis conditions based on mean independent shocks within the blocks. Importantly, identification is achieved without many higher-order moment conditions and holds under fairly general conditions on the dependencies of the shocks. Second, we show that additional overidentifying higher-order moment conditions – some of these conditions additionally require the assumption of independent shocks – can decrease the asymptotic variance of the estimator if the overidentifying conditions are valid. Third, we propose to use a LASSO-type GMM estimator to select the valid and relevant overidentifying higher-order moment conditions in a data-driven way. Consistency of the estimator only relies on the identifying moment conditions and, thus, is robust to various kinds of dependencies of the shocks. Furthermore, it can exploit efficiency gains from valid and relevant overidentifying conditions and ignore noise from valid but redundant overidentifying conditions.

5.4.1 Identification

In this section, we show that identification of a block-recursive SVAR can be achieved by the variance and covariance conditions in Equation (5.1) and (5.2) and the asymmetric cokurtosis conditions in Equation (5.5) corresponding to innovations in the same block. The identification result is robust in the sense that it allows for various sorts of dependencies of the shocks. To be clear, shocks in different blocks only need to be uncorrelated and shocks in the same block only need to fulfill the asymmetric cokurtosis conditions.

Let $E[f_{\mathbf{2}}(B, u_t)] = 0$ contain all variance and covariance conditions in Equation (5.1) and (5.2) and let $E[f_{\mathbf{4_{p_k}}}(B, u_t)] = 0$ contain all asymmetric cokurtosis conditions from Equation (5.5) corresponding to shocks in block k, e.g., $E[e(B)^3_{i,t} e(B)_{j,t}] = 0$ for $i, j = p_k, ..., p_{k+1} - 1$ and $i \neq j$. We define the identifying moment conditions as

$$E[f_{\mathbf{N}}(B, u_t)] := E \begin{bmatrix} f_{\mathbf{2}}(B, u_t) \\ f_{\mathbf{4_{p_1}}}(B, u_t) \\ \vdots \\ f_{\mathbf{4_{pm}}}(B, u_t) \end{bmatrix} = 0.$$

In the following, we simplify the notation for moment conditions, e.g., we write $E[f_{\mathbf{N}}(B, u_t)]$ instead of $E[f_{\mathbf{N}}(B, u_t)] = 0$. Note that the identifying moment conditions do not contain asymmetric cokurtosis conditions of shocks in different blocks, e.g., the moment conditions $E[e(B)_{i,t}^3 e(B)_{j,t}]$ for shocks $e(B)_{i,t}$ and $e(B)_{j,t}$ in different blocks are not contained in $E[f_{\mathbf{N}}(B, u_t)]$. The conditions $E[f_{\mathbf{N}}(B, u_t)]$ can be justified by the following assumption.

Assumption 5.2. *(Block-recursive mean independence.)*
For $m \leq n$ blocks with $p_1 = 1 < p_2 < \ldots < p_m \leq n$,

(i) all shocks are uncorrelated, i.e., $E[\varepsilon_{i,t}\varepsilon_{j,t}] = 0$ for $i \neq j$.

(ii) all shocks within the same block are mean independent, i.e., $E[\varepsilon_{i,t}|\varepsilon_{-i,t}] = 0$ for $i \in \{p_k, p_k + 1, ..., p_{k+1} - 1\}$ and $-i = \{p_k, p_k + 1, ..., p_{k+1} - 1\} \backslash i$ for $k = 1, \ldots, m$.

The identifying moment conditions contain n variance conditions, $n(n-1)/2$ covariance conditions and $\sum_{k=1}^{m} l_k(l_k - 1)/2$ asymmetric cokurtosis conditions, where $l_k := p_{k+1} - p_k$ denotes the number of shocks in block k. Therefore, each additional specified block refines the identifying moment conditions $E[f_{\mathbf{N}}(B, u_t)]$ such that they contain fewer higher-order moment conditions. In the extreme case when the SVAR is specified recursively, meaning each block contains only one variable, the identifying moment conditions contain no higher-order moment conditions. In the other extreme case of a single block containing all variables, the identifying moment conditions contain all $n(n-1)$ asymmetric cokurtosis conditions and are similar to the conditions proposed in Lanne and Luoto (2021).[6]

The following proposition shows that the identifying moment conditions are sufficient to locally identify the block-recursive SVAR.

Proposition 5.1. *(Identification in the block-recursive SVAR.)*
Let $u_t = B_0\varepsilon_t$ with $m \leq n$ blocks and $B_0 \in \mathbb{B}_{brec} \equiv \mathbb{B}_{brec}(p_1, \ldots, p_m)$ such that Assumption 5.1 holds. Moreover, suppose that Assumption 5.2 holds. If at most one structural shock in each block has zero excess kurtosis, the identifying moment conditions $E[f_{\mathbf{N}}(B, u_t)] = 0$ locally identify $B = B_0$ for $B \in \mathbb{B}_{brec}$.

Proof. The proof recursively applies the identification result from Lanne and Luoto (2021) and can be found in Appendix 5.A.3. □

[6] Lanne and Luoto (2021) propose to select $n(n-1)/2$ asymmetric cokurtosis conditions, which is sufficient for local identification if none of the asymmetric conditions does include the third power of a Gaussian shock. They advocate to rely on a moment selection criterion to avoid including redundant conditions or conditions of Gaussian shocks. Additionally, Lanne and Luoto (2021) note that including all $n(n-1)$ asymmetric cokurtosis conditions ensures local identification even if conditions related to Gaussian shocks are included. We argue that the degree of overidentification remains reasonably small even if we include all asymmetric cokurtosis conditions and therefore, including redundant conditions can be expected to be rather harmless. For example, in a SVAR with four variables and no restrictions the identifying moment conditions consists of 22 conditions to identify 16 parameters. Thus, we suggest to use all asymmetric cokurtosis conditions in order to avoid the cumbersome process of selecting a subset of the conditions.

In Proposition 5.1 the impact of shocks on variables in different blocks is identified based on covariance conditions. The interaction of shocks on variables within the same block is identified based on asymmetric cokurtosis conditions and the local identification result of Lanne and Luoto (2021). Local identification means that the moment conditions $E[f_{\mathbf{N}}(B, u_t)]$ identify B_0 in a small neighborhood of B_0 (see Hall (2005)). Importantly, the proposition also holds for different higher-order moment conditions ensuring identification within the blocks. For example, the identifying conditions $E[f_{\mathbf{N}}(B, u_t)]$ could contain all variance-covariance, coskewness and cokurtosis conditions implied by independent structural shocks for each block. In this case, global identification up to sign and permutation within each block follows from Keweloh (2021*b*).

Without further restrictions, data-driven approaches relying on non-Gaussian and independent shocks can only ensure identification up to sign and permutation. This means that the order and sign of the shocks in the impulse response functions is not identified. In practice, the researcher has to manually assign labels to the shocks. Restricting the solution to a given block-recursive structure simplifies the permutation or labeling problem. In particular, shocks can only be permuted inside blocks. For instance, in example (b) in Figure 5.1 shocks from the second block cannot be permuted into the first block since this violates the block-recursive structure. Therefore, specifying a finer block-recursive structure simplifies the labeling of the shocks.

Define the block-recursive SVAR GMM estimator which minimizes the variance, covariance and the asymmetric cokurtosis conditions over the set of block-recursive matrices as

$$\hat{B}_{\mathbf{N}} := \underset{B \in \mathbb{B}_{brec}}{\arg\min}\, g_{\mathbf{N}}(B)' W_{\mathbf{N}} g_{\mathbf{N}}(B), \tag{5.9}$$

with a suitable weighting matrix $W_{\mathbf{N}}$ and $g_{\mathbf{N}}(B) := 1/T \sum_{t=1}^{T} f_{\mathbf{N}}(B, u_t)$. Consistency and asymptotic normality follow from the identification result in Proposition 5.1 and standard assumptions including valid moment conditions implied by the dependence structure imposed in Assumption 5.2. That is,

$$\hat{B}_{\mathbf{N}} \xrightarrow{p} B_0$$

$$\sqrt{T}\left(\left(vec\left(\hat{B}_{\mathbf{N}}\right) - vec\left(B_0\right)\right) \xrightarrow{d} \mathcal{N}\left(0, V_{\mathbf{N}}\right),\right.$$

where the formula for the asymptotic variance, $V_{\mathbf{N}}$, is standard but lengthy and, therefore, deferred to Appendix 5.A.1. Moreover, under standard assumptions the weighting matrix $W_{\mathbf{N}}^{*} := S_{\mathbf{N}}^{-1}$ with $S_{\mathbf{N}} := \lim_{T\to\infty} E[g_{\mathbf{N}}(B) g_{\mathbf{N}}(B)']$ leads to the estimator $\hat{B}_{\mathbf{N}}^{*}$ with lowest possible asymptotic variance (see, e.g., Hall (2005)).

In many applications, the researcher is only interested in some structural shocks. For this case, we derive a partial identification result under weaker assumptions.

Proposition 5.2. *(Partial identification in the block-recursive SVAR.)*
Let $u_t = B_0\varepsilon_t$ with $m \leq n$ blocks and $B_0 \in \mathbb{B}_{brec} \equiv \mathbb{B}_{brec}(p_1, \ldots, p_m)$ such that Assumption 5.1 holds. Moreover, let $B_{i,0}$ denote the columns of B_0 representing impact of the structural shocks in the ith block. Let $\widetilde{\mathbb{B}}_{brec} := \mathbb{B}_{brec}(\tilde{p}_1, \ldots, \tilde{p}_{\tilde{m}})$ denote a potentially different block-recursive interaction. Assume that there exists a block $\tilde{p}_j$ of $\widetilde{\mathbb{B}}_{brec}$ which contains the shocks of block p_i, i.e., there exits a j, $1 \leq j \leq \tilde{m}$, such that $\tilde{p}_j = p_i$ and $\tilde{p}_{j+1} = p_{i+1}$.

The moment conditions $E\begin{bmatrix} f_{\mathbf{2}}(B, u_t) \\ f_{\mathbf{4_{\tilde{p}_j}}}(B, u_t) \end{bmatrix} = 0$ *locally identify $B_{i,0}$ for $B \in \widetilde{\mathbb{B}}_{brec}$ if the following conditions hold:*

1. *The shocks ε_t are uncorrelated.*
2. *The asymmetric cokurtosis conditions of block $\tilde{p}_j$ hold.*
3. *At most one shock in block $\tilde{p}_j$ has zero excess kurtosis.*

Proof. The proof can be found in Appendix 5.A.3. □

Proposition 5.2 reveals that we can identify a specific block of shocks by using only the second moments of all shocks and the asymmetric cokurtosis conditions of the shocks in the block of interest as long as the block of interest is specified correctly and contains at most one Gaussian shock. To see the advantages of the partial identification result, consider that we are only interested in the last two structural shocks in Figure 5.1 (b). In this example, Proposition 5.2 implies that the impact of the last two shocks is identified even if (i) the first and second shock are both Gaussian, (ii) the first and second shock do not satisfy the asymmetric cokurtosis conditions but are only uncorrelated, or (iii) the block-recursive structure is misspecified as the one displayed in Figure 5.1 (c). Additionally, Proposition 5.2 implies that the moment conditions used in Proposition 5.1 identify the shocks in a block of interest if the block of interest is specified correctly, contains at most one Gaussian shock, and there exists a B such that the moment conditions are fulfilled. However, the B matrix can differ from B_0, except for the columns corresponding to the block of interest.

5.4.2 Overidentification and efficiency gains

In the previous section, we proposed a block-recursive SVAR GMM estimator, which uses only a (small) subset of asymmetric cokurtosis conditions, and provide an identification result which does not require independent shocks. However, the excluded set of coskewness and cokurtosis conditions can decrease the asymptotic variance of the estimator and hence, increase the efficiency of the estimator. In this section, we define the overidentified block-recursive SVAR GMM estimator which contains all coskewness and cokurtosis conditions implied by independent shocks. Additionally, we derive conditions for the redundancy and relevance of the overidentifying coskewness and cokurtosis conditions in a recursive SVAR with independent structural shocks.

Assumption 5.3. *(Independent shocks.)*
All shocks are independent, i.e., $\varepsilon_{i,t}$ is independent of $\varepsilon_{j,t}$ for $i \neq j$.

For a given block-recursive SVAR, define the overidentifying moment conditions as

$$E[f_{\mathbf{D}}(B, u_t)] = E\begin{bmatrix} f_{\mathbf{3}\backslash\mathbf{N}}(B, u_t) \\ f_{\mathbf{4}\backslash\mathbf{N}}(B, u_t) \end{bmatrix},$$

where $E[f_{\mathbf{3}\backslash\mathbf{N}}(B, u_t)]$ contains all coskewness conditions from Equation (5.3)-(5.4), and $E[f_{\mathbf{4}\backslash\mathbf{N}}(B, u_t)]$ contains all cokurtosis conditions from Equation (5.5)-(5.8), implied by independent shocks and not included in the identifying moment conditions $E[f_{\mathbf{N}}(B, u_t)]$.

The overidentified block-recursive SVAR GMM estimator is defined as

$$\hat{B}_{\mathbf{N+D}} := \underset{B \in \mathbb{B}_{brec}}{\arg\min} \begin{bmatrix} g_{\mathbf{N}}(B) \\ g_{\mathbf{D}}(B) \end{bmatrix}' W_{\mathbf{N+D}} \begin{bmatrix} g_{\mathbf{N}}(B) \\ g_{\mathbf{D}}(B) \end{bmatrix}, \tag{5.10}$$

with a suitable weighting matrix $W_{\mathbf{N+D}}$ and $g_{\mathbf{D}}(B) := 1/T \sum_{t=1}^{T} f_{\mathbf{D}}(B, u_t)$. Note that the overidentified block-recursive SVAR GMM estimator uses all coskewness and cokurtosis conditions implied by independent shocks. That is, the moment conditions used for estimation are the same as in the SVAR GMM estimator proposed by Keweloh (2021*b*). However, the latter estimator neither uses restrictions nor distinguishes between identifying and overidentifying moment conditions. In contrast to that, we allow for block-recursive restrictions. These restrictions allow to transform identifying into overidentifying moment conditions.

Consistency and asymptotic normality of the overidentified block-recursive SVAR GMM estimator in Equation (5.10) require that not only the identifying but also the overidentifying moment conditions are valid, which holds if the shocks are independent as assumed in Assumption 5.3. That is,

$$\hat{B}_{\mathbf{N+D}} \xrightarrow{p} B_0$$

$$\sqrt{T}\left(\left(vec\left(\hat{B}_{\mathbf{N+D}}\right) - vec\left(B_0\right)\right) \xrightarrow{d} \mathcal{N}\left(0, V_{\mathbf{N+D}}\right),\right.$$

where the formula for the asymptotic variance, $V_{\mathbf{N+D}}$, is standard and can be found in Appendix 5.A.1. Again, under standard assumptions the weighting matrix $W^*_{\mathbf{N+D}} := S^{-1}_{\mathbf{N+D}}$ with $S_{\mathbf{N+D}} := \lim_{T\to\infty} E[g_{\mathbf{N+D}}(B_0) g_{\mathbf{N+D}}(B_0)']$, where $g_{\mathbf{N+D}}(B_0) := [g_{\mathbf{N}}(B_0)', g_{\mathbf{D}}(B_0)']'$, leads to the estimator $\hat{B}^*_{\mathbf{N+D}}$ with lowest possible asymptotic variance (see, e.g., Hall (2005)).

Adding additional valid moment conditions can never increase the asymptotic variance of the GMM estimator (see, e.g., Breusch, Qian, Schmidt and Wyhowski (1999)). Therefore, if the structural shocks are independent such that the overidentifying conditions hold, the asymptotic variance of $\hat{B}^*_{\mathbf{N+D}}$ is equal to or smaller than the asymptotic variance of $\hat{B}^*_{\mathbf{N}}$. If including an additional moment condition decreases the asymptotic variance of the estimator, the moment condition is called relevant, otherwise the moment condition

is called redundant. A moment condition is called partially relevant for a subset of parameters if it decreases the asymptotic variance of a subset of parameters. If this is not the case, the moment condition is called partially redundant.

In the following proposition, we show that overidentifying higher-order moment conditions in $E[f_{\mathbf{D}}(B, u_t)]$ can decrease the asymptotic variance of the estimator. To this end, we consider the special case of a recursive SVAR with independent shocks. In this case, the SVAR is identified solely by second-order moment conditions and all coskewness and cokurtosis moment conditions are overidentifying. The proposition highlights that some coskewness and cokurtosis conditions are always (partially) redundant, while other conditions are relevant if certain conditions for the skewness, excess kurtosis, and elements of the inverse of B_0 are fulfilled. The proposition also implies that if at least one shock has a non-zero skewness, at least one higher-order moment condition will be relevant and consequently, the recursive SVAR GMM estimator based solely on second-order moment conditions, which is equal to frequently used estimator obtained by applying the Cholesky decomposition, is inefficient.

Proposition 5.3. *(Redundant and relevant moment conditions in the recursive SVAR.) Let $A := B_0^{-1}$ and let a_{ql} denote the element at row q and column l of A. Additionally let $i, j, k, l \in \{1, ..., n\}$ and $i \neq j \neq k \neq l$. The impact of a shock $\epsilon_{q,t}$ is equal to the unrestricted elements in the q-th row of B_0. In a recursive SVAR with independent structural shocks the following redundancy statements hold w.r.t. the identifying second-order moment conditions $E[f_{\mathbf{2}}(B, u_t)]$.*

Coskewness condition:

1. $E[e(B_0)_i e(B_0)_j e(B_0)_k]$ *is redundant.*
2. $E[e(B_0)_i^2 e(B_0)_j]$ *is partially redundant for the impact of the shock $\epsilon_{q,t}$ with $q \neq j$.*
3. $E[e(B_0)_i^2 e(B_0)_j]$ *is partially redundant for the impact of the shock $\epsilon_{j,t}$ if and only if*

for $i < j$	*for $i > j$*
$\frac{2E[\epsilon_{j,t}^3]}{E[\epsilon_{j,t}^4]-1} a_{jj} = 0.$	$\begin{aligned} \frac{2E[\epsilon_{j,t}^3]}{E[\epsilon_{j,t}^4]-1} a_{jj} + E[\epsilon_{i,t}^3] a_{ij} &= 0, \\ E[\epsilon_{i,t}^3] a_{i,z} &= 0, \quad z = j+1, \ldots, i. \end{aligned}$

Cokurtosis condition:

1. $E[e(B_0)_i e(B_0)_j e(B_0)_k e(B_0)_l]$ *and* $E[e(B_0)_i^2 e(B_0)_j e(B_0)_k]$ *are redundant.*
2. $E[e(B_0)_i^3 e(B_0)_j]$ *is partially redundant for the impact of the shock $\epsilon_{q,t}$ with $q \neq j$.*
3. $E[e(B_0)_i^3 e(B_0)_j]$ *is partially redundant for the impact of the shock $\epsilon_{j,t}$ if and only if*

for $i < j$	*for $i > j$*
$\frac{2E[\epsilon_{j,t}^3]E[\epsilon_{i,t}^3]}{E[\epsilon_{j,t}^4]-1} a_{jj} = 0.$	$\begin{aligned} \frac{2E[\epsilon_{j,t}^3]E[\epsilon_{i,t}^3]}{E[\epsilon_{j,t}^4]-1} a_{jj} + (E[\epsilon_{i,t}^4] - 3) a_{ij} &= 0, \\ (E[\epsilon_{i,t}^4] - 3) a_{i,z} &= 0, \quad z = j+1, \ldots, i. \end{aligned}$

4. $E[e(B_0)_i^2 e(B_0)_j^2 - 1]$ *is partially redundant for the impact of the shock* $\epsilon_{q,t}$ *with* $q \neq i$ *and* $i < j$.
5. $E[e(B_0)_i^2 e(B_0)_j^2 - 1]$ *is partially redundant for the impact of the shock* $\epsilon_{i,t}$ *with* $i < j$ *if and only if*

$$E[\epsilon_{j,t}^3]E[\epsilon_{i,t}^3]a_{jz} = 0, \quad z = i, \ldots, j.$$

Proof. The proof can be found in Appendix 5.A.4. □

In practice, the conditions in Proposition 5.3 cannot be verified since the matrix B_0, the skewness, and the kurtosis of the structural shocks are unknown a priori. Furthermore, Proposition 5.3 only covers a recursive SVAR with independent shocks, i.e., if the shocks are only mean independent or the SVAR has a different block-recursive structure, we do not have a theoretical result on which moment conditions are relevant and which are not.

5.4.3 Data-driven moment selection

Section 5.4.1 provides an identification result for block-recursive SVARs only requiring a (small) subset of cokurtosis conditions which is robust in the sense that it allows for various kinds of dependencies of the shocks. Section 5.4.2 stresses that there is a trade-off between robustness and efficiency of the estimator. For robustness, we leave out overidentifying conditions, which has the downside that some of these conditions may be valid and relevant, i.e., decrease the asymptotic variance of the estimator. However, an advantage is that one does not include potentially invalid overidentifying conditions, which could lead to an inconsistent overidentified block-recursive SVAR GMM estimator in Equation (5.10). Additionally, valid but redundant overidentifying conditions can lead to a many moment problem and a poor finite sample performance of the overidentified block-recursive SVAR GMM estimator, compare Cheng and Liao (2015), Hall (2005), and Hall (2015). Therefore, we propose to use the pGMM estimator of Cheng and Liao (2015) to detect and include only the relevant and valid overidentifying moment conditions in a data-driven way. By including valid and relevant moment conditions in the estimation, we exploit the asymptotic efficiency gains of relevant moments. By leaving out invalid or redundant moment conditions, we can avoid inconsistent estimates and issues related to many moment conditions.

In general, the overidentifying higher-order moment conditions $E[f_{\mathbf{D}}(B, u_t)]$ can be separated into three sets: $E[f_{\mathbf{A}}(B, u_t)]$ contains valid and relevant moment conditions, $E[f_{\mathbf{R}}(B, u_t)]$ contains valid but redundant conditions, and $E[f_{\mathbf{I}}(B, u_t)]$ contains invalid moment conditions. The goal is to select the moments $E[f_{\mathbf{A}}(B, u_t)]$ and to leave out the moments $E[f_{\mathbf{R}}(B, u_t)]$ and $E[f_{\mathbf{I}}(B, u_t)]$. However, in practice the researcher does not know whether a given moment condition is invalid, redundant, or valid and relevant. Therefore, we propose to detect and select the relevant and valid overidentifying moment conditions in a data-driven way. Based on Cheng and Liao (2015), we define the block-

recursive SVAR pGMM estimator

$$\{\widehat{B}^{pGMM}_{\mathbf{N+D}}, \widehat{\beta}\} := \underset{\{B,\beta\}\in\Lambda}{\arg\min} \begin{bmatrix} g_{\mathbf{N}}(B) \\ g_{\mathbf{D}}(B) - \beta \end{bmatrix}' W_{\mathbf{N+D}} \begin{bmatrix} g_{\mathbf{N}}(B) \\ g_{\mathbf{D}}(B) - \beta \end{bmatrix} + \lambda \sum_{j\in\widetilde{D}} \omega_j \, |\beta_j|, \quad (5.11)$$

where $\lambda \geq 0$ is a tuning parameter specified by the researcher, $\beta \in \mathbb{R}^{k_{\mathbf{D}}}$ is the vector of slackness parameters, $\Lambda := \{\mathbb{B}_{brec}, \mathbb{R}^{1\times k_{\mathbf{D}}}\}$ is the parameter space of $\{B, \beta\}$, $\omega \in \mathbb{R}^{k_{\mathbf{D}}}$ is a vector of weights used in the penalty term, and $\widetilde{D} := \{1, \ldots, k_{\mathbf{D}}\}$ with $k_{\mathbf{D}}$ denoting the number of conditions in $E[f_{\mathbf{D}}(B, u_t)]$.

The vector of slackness parameters β allows the moment conditions $E[f_{\mathbf{D}}(B, u_t)]$ to deviate from zero without increasing the first part of the loss function and therefore, to decrease their impact on the estimation. However, each element of β gets penalized in the second part of the loss function and consequently, giving slack to overidentifying moments adds a cost, i.e., increases the loss function. The vector of weights ω and the tuning parameter λ govern the cost of giving slack to moment conditions. In particular, a smaller λ makes it cheaper to give slack to all overidentifying moments and a smaller ω_j makes it less costly to give slack to a specific overidentifying moment j.

The pGMM estimator in Equation (5.11) has two special cases. First, if $\lambda = 0$, adding slack to the overidentifying moments is not penalized. Therefore, the solution of the pGMM estimator is $\widehat{B}^{pGMM}_{\mathbf{N+D}} = \hat{B}_{\mathbf{N}}$ and $\widehat{\beta} = g_{\mathbf{D}}\left(\hat{B}_{\mathbf{N}}\right)$, where $\hat{B}_{\mathbf{N}}$ is the solution of the the block-recursive SVAR GMM estimator in Equation (5.9) using only the identifying moments $E[f_{\mathbf{N}}(B, u_t)]$ and the weighting matrix $W_{\mathbf{N}}$, equal to the block of the weighting matrix $W_{\mathbf{N+D}}$ corresponding to the identifying conditions $E[f_{\mathbf{N}}(B, u_t)]$. Second, if $\lambda = \infty$, deviations of $\widehat{\beta}$ from zero become infinitely costly for overidentifying moments with $\omega_j > 0$. Assuming $\omega > 0$, the pGMM estimator cannot give slack to any overidentifying moment condition. Thus, $\widehat{B}^{pGMM}_{\mathbf{N+D}} = \hat{B}_{\mathbf{N+D}}$ and $\widehat{\beta} = 0$ minimize the loss function of the pGMM estimator, where $\hat{B}_{\mathbf{N+D}}$ is the solution of the the overidentified block-recursive SVAR GMM estimator in Equation (5.10), using the weighting matrix $W_{\mathbf{N+D}}$. Choices of λ other than $\lambda = 0$ or $\lambda = \infty$ lead to solutions which lie between these extreme cases. In practice, we recommend using cross-validation to find the optimal value of λ.

The penalty term uses weights $\omega_j \geq 0, \ \forall j \in \tilde{D}$, to shrink the elements of β differently. Let $E[f_{\mathbf{D}_j}(B, u_t)]$ for $j \in \widetilde{D}$ correspond to one specific moment of $E[f_{\mathbf{D}}(B, u_t)]$. A higher ω_j leads to more shrinkage for β_j and consequently, makes it more likely that β_j becomes zero, meaning that the corresponding moment $E[f_{\mathbf{D}_j}(B, u_t)]$ gets selected. Furthermore, $\omega_j = 0$ implies that even if the tuning parameter λ is large, there is no cost for giving slack to the moment condition $E[f_{\mathbf{D}_j}(B, u_t)]$, implying that those moments do not influence the estimated $\widehat{B}^{pGMM}_{\mathbf{N+D}}$. Since we aim to select only the relevant and valid moment conditions $E[f_{\mathbf{A}}(B, u_t)$, and not the invalid $E[f_{\mathbf{I}}(B, u_t)]$ or redundant moment conditions $E[f_{\mathbf{R}}(B, u_t)]$, we would specify $\omega_j > 0$ for all valid and relevant conditions, and $\omega_j = 0$ for all invalid or redundant conditions. To achieve this without prior knowledge on $E[f_{\mathbf{A}}(B, u_t)]$, $E[f_{\mathbf{R}}(B, u_t)]$, and $E[f_{\mathbf{I}}(B, u_t)]$, Cheng and Liao (2015) construct ω_j

allowing information-based adaptive adjustment for each moment in $E[f_{\mathbf{D}}(B, u_t)]$. More precisely, they use

$$\omega_j = \frac{\mu_j^{r_1}}{|\beta_j^{*r_2}|},\ j \in \tilde{D}, \tag{5.12}$$

where μ_j is a measure for the empirical relevance of the moment condition $E[f_{\mathbf{D}_j}(B, u_t)]$, relative to the identifying moment conditions $E[f_{\mathbf{N}}(B, u_t)]$, and β_j^* is a preliminary consistent estimator of $E[f_{\mathbf{D}_j}(B_0, u_t)]$ and $r_1 \geq r_2 \geq 0$ are constants specified by the researcher. The use of $1/|\beta_j^{*r_2}|$ resembles an adaptive LASSO penalty (cf. Zou (2006)) and implies that moments with small β_j^* are subject to more shrinkage. Since β_j^* is a consistent estimator and the true value of β_j^* for a valid moment is zero, the adaptive penalty ensures that valid moments get selected. However, using only the adaptive penalty, we would unintendedly incentivize the estimator to select also redundant moments since, by definition, these are also valid. To avoid selecting redundant moments, Cheng and Liao (2015) suggest to multiply the adaptive penalty with

$$\mu_j = \rho_{\max}\left(\widehat{V}_{\mathbf{N}} - \widehat{V}_{\mathbf{N}+\mathbf{D}_j}\right),\ j \in \tilde{D}, \tag{5.13}$$

where $\rho_{\max}(A)$ is the maximum eigenvalue of a square matrix A and $\widehat{V}_{\mathbf{N}}$ and $\widehat{V}_{\mathbf{N}+\mathbf{D}_j}$ are consistent estimators of the efficient asymptotic variance-covariance matrices $V_{\mathbf{N}}^*$ and $V_{\mathbf{N}+\mathbf{D}_j}^*$, defined in Appendix 5.A.1. If the maximum eigenvalue of $V_{\mathbf{N}}^* - V_{\mathbf{N}+\mathbf{D}_j}^*$ is positive, then adding moment condition $E[f_{\mathbf{D}_j}(B, u_t)]$ to the conditions $E[f_{\mathbf{N}}(B, u_t)]$ decreases the asymptotic variance of the estimator and hence, moment condition $E[f_{\mathbf{D}_j}(B, u_t)]$ is relevant. Therefore, μ_j estimates the empirical relevance of the moment $E[f_{\mathbf{D}_j}(B, u_t)]$.[7]

Cheng and Liao (2015) show that, under conditions, the pGMM estimator consistently selects the valid and relevant moments, i.e., $\lim_{T\to\infty} P(\hat{\beta}_j = 0) = 1$ if the moment condition $E[f_{\mathbf{D}_j}(B, u_t)]$ is in $E[f_{\mathbf{A}}(B, u_t)]$, and does not select the invalid or redundant moments, i.e., $\lim_{T\to\infty} P(\hat{\beta}_j = 0) = 0$ if the moment condition $E[f_{\mathbf{D}_j}(B, u_t)]$ is in $E[f_{\mathbf{R}}(B, u_t)]$ or $E[f_{\mathbf{I}}(B, u_t)]$. They also derive that, under conditions, the pGMM estimator is a consistent estimator of B_0 and asymptotically normal with asymptotic variance $V_{\mathbf{N}+\mathbf{A}}$.[8] In our case, the conditions in particular require that Assumption 5.2 holds. However, consistency and asymptotic normality do not rely on independent shocks, i.e., Assumption 5.3. Even though the SVAR pGMM estimator uses the moment conditions $E[f_{\mathbf{N}}(B, u_t)]$ and $E[f_{\mathbf{D}}(B, u_t)]$ for estimation, its asymptotic variance only

[7] Cheng and Liao (2015) show that $V_{\mathbf{N}}^* - V_{\mathbf{N}+\mathbf{D}_j}^*$ is positive semidefinite for every $j \in \widetilde{D}$, implying that the maximum eigenvalue of $V_{\mathbf{N}}^* - V_{\mathbf{N}+\mathbf{D}_j}^*$ is nonnegative. Furthermore, note that both $\widehat{V}_{\mathbf{N}} \equiv \widehat{V}_{\mathbf{N}}\left(\hat{B}_{\mathbf{N}}\right)$ and $\widehat{V}_{\mathbf{N}+\mathbf{D}_j} \equiv \widehat{V}_{\mathbf{N}+\mathbf{D}_j}\left(\hat{B}_{\mathbf{N}}\right)$ are evaluated at $\hat{B}_{\mathbf{N}}$, which is obtained from Equation (5.9). Thereby, we do not rely on $\hat{B}_{\mathbf{N}+\mathbf{D}_j}$ to estimate $V_{\mathbf{N}+\mathbf{D}_j}^*$ since the moment associated with $\mathbf{D}_j$ may be invalid and hence, $\widehat{V}_{\mathbf{N}+\mathbf{D}_j}\left(\hat{B}_{\mathbf{N}+\mathbf{D}_j}\right)$ inconsistent for $V_{\mathbf{N}+\mathbf{D}_j}^*$.

[8] This result is not explicitly stated in Cheng and Liao (2015) but follows from their Remark 3.5 using the Cramér-Wold device, an arbitrary weighting matrix W and replacing the variance of the sample GMM estimator with the asymptotic variance. We prove the result in Appendix 5.A.5 under Assumption 5.1 and 5.2.

depends on the moments conditions $E[f_{\mathbf{D}}(B, u_t)]$ and $E[f_{\mathbf{A}}(B, u_t)]$. That is, the SVAR pGMM estimator successfully ignores the redundant and invalid moments and decreases the asymptotic variance by incorporating the information contained in the relevant and valid moments. The weighting matrix $W^*_{\mathbf{N}+\mathbf{D}} := S^{-1}_{\mathbf{N}+\mathbf{D}}$ leads to the estimator with the lowest possible asymptotic variance (Hall, 2005), corresponding to the asymptotic variance of the oracle estimator. The oracle estimator uses only moment conditions $E[f_{\mathbf{N}}(B, u_t)]$ and $E[f_{\mathbf{A}}(B, u_t)]$ and is infeasible in practice without prior knowledge of $E[f_{\mathbf{D}}(B, u_t)]$ and $E[f_{\mathbf{A}}(B, u_t)]$. However, the SVAR pGMM estimator is as efficient as the oracle estimator asymptotically.

5.5 Finite sample performance

In this section, we conduct two Monte Carlo studies. The first one illustrates that the performance of SVAR estimators can be improved substantially by exploiting the block-recursive structure. This is especially relevant for SVARs with a large number of variables. The second Monte Carlo study focuses on how to incorporate information in overidentifying higher-order moment conditions. More concretely, we demonstrate that the pGMM estimator selects relevant and does not select redundant moment conditions in a data-driven way and thereby, improves the finite sample performance.

For both Monte Carlo experiments, we consider three different sample sizes $T = \{100, 250, 1\,000\}$ to analyze the influence of the sample size on the performance of the estimators. We independently and identically draw each structural shock ϵ_{it}, $i = 1, \ldots, n$, $t = 1, \ldots, T$, from the two-component mixture

$$\epsilon_{it} \sim 0.79\,\mathcal{N}(-0.2, 0.7^2) + 0.21\,\mathcal{N}(0.75, 1.5^2),$$

where $\mathcal{N}(\mu, \sigma^2)$ indicates a normal distribution with mean μ and standard deviation σ. The shocks have skewness 0.9 and excess kurtosis 2.4.

We compare the finite sample performance of various SVAR estimators.[9] Based on the simulations presented in Keweloh (2021*a*), we use continuous updating estimators (CUEs) instead of GMM estimators and estimate the asymptotically efficient weighting matrix based on serially and mutually independent shocks.[10] Since CUE estimators are closely related to GMM estimators, we use both terms interchangeably. More specifically, we refer to the estimators as follows:

[9]The estimators are implemented in python and the pGMM estimator uses the solvers of Defferrard, Pena and Perraudin (2017).

[10]Keweloh (2021*a*) demonstrates that the inability to precisely estimate S, the long-run covariance matrix of the moment conditions, and as consequence the efficient weighting matrix leads to a poor small sample performance of two-step GMM and CUE estimators. Recognizing this downside, Keweloh (2021*a*) proposes a novel estimator for S exploiting serially and mutually independent shocks.Keweloh (2021*a*) illustrates that the estimator for S substantially increases the small sample performance of the two-step GMM and CUE estimator. Additionally, Keweloh (2021*a*) illustrates that CUE estimators are less biased than GMM estimator in small samples.

- GMM: Continuous updating estimator based on Equation (5.9) using only the identifying moment conditions $E[f_{\mathbf{N}}(B, u_t)]$.
- oGMM: Overidentified continuous updating estimator based on Equation (5.10) using the identifying moment conditions $E[f_{\mathbf{N}}(B, u_t)]$ and overidentifying moment conditions $E[f_{\mathbf{D}}(B, u_t)]$.
- GMM-Oracle: Overidentified continuous updating estimator based on Equation (5.10) using the identifying moment conditions $E[f_{\mathbf{N}}(B, u_t)]$ and the relevant overidentifying moment conditions $E[f_{\mathbf{A}}(B, u_t)]$.
- pGMM: Continuous updating LASSO estimator based on Equation (5.11).

We only indicate which block-recursive structure is imposed for estimation, when necessary (e.g., when comparing an GMM estimator without restrictions with a block-recursive GMM estimator).

5.5.1 Block-Recursive Structure

We simulate a SVAR with $n = 2$ and $n = 4$ variables. The mixing matrices B_0 are given by

$$B_0 = \begin{bmatrix} 10 & 5 \\ 5 & 10 \end{bmatrix} \quad \text{and} \quad B_0 = \begin{bmatrix} 10 & 5 & 0 & 0 \\ 5 & 10 & 0 & 0 \\ 5 & 5 & 10 & 5 \\ 5 & 5 & 5 & 10 \end{bmatrix}. \tag{5.14}$$

The Monte Carlo study analyzes the impact of imposing a block-recursive structure for GMM estimators. In the small SVAR with $n = 2$, we impose no restrictions. In the large SVAR with $n = 4$, we estimate the GMM estimator without restrictions and the block-recursive GMM estimator, using the block-recursive structure in Equation (5.14), i.e., we apply zero restrictions for all elements where B_0 is zero.[11]

Table 5.1 summarizes the results of $M = 3,500$ Monte Carlo simulations. The table shows the average of each estimated element $\bar{b}_{ij} = 1/M \sum_{m=1}^{M} \hat{b}_{ij}^m$ and the estimated mean squared error (MSE), $\hat{\sigma}_{i,j}^2 = 1/M \sum_{m=1}^{M} \left(\hat{b}_{ij}^m - b_{ij}\right)^2$, where b_{ij} denotes the element of B_0 in row i and column j and $\hat{b}_{ij}^m$ its estimated value in Monte Carlo run m. Moreover, we calculate the average over the empirical biases, $Bias := \sum_{i=1}^{n} \sum_{j=1}^{n} w_{i,j} \left(\bar{b}_{ij} - b_{ij}\right)$, and the average over the estimated MSEs, $Var := \sum_{i=1}^{n} \sum_{j=1}^{n} w_{i,j} \hat{\sigma}_{i,j}^2$, across estimated elements in $\hat{B}$, i.e., $w_{i,j}$ equals zero if $\hat{b}_{ij}^m$ is restricted to be zero and one over the number of estimated elements in $\hat{B}$ otherwise. Additionally, we report the number of moments used by each estimator.

[11] In this Monte Carlo study, we focus on GMM estimators. We include the oGMM, GMM-Oracle and pGMM estimator in the second Monte Carlo study.

Table 5.1: Finite sample performance of the GMM and block-recursive GMM estimator.

		n=2	n=4	
		GMM	GMM	block-recursive GMM
$T = 100$	$\hat{B}$	$\begin{bmatrix} 9.78 & 4.90 \\ (2.26) & (4.31) \\ 4.90 & 9.76 \\ (4.24) & (2.18) \end{bmatrix}$	$\begin{bmatrix} 9.28 & 4.63 & 0.04 & 0.07 \\ (3.24) & (4.82) & (5.31) & (5.27) \\ 4.70 & 9.23 & 0.08 & 0.05 \\ (4.87) & (3.20) & (5.32) & (5.14) \\ 4.68 & 4.62 & 9.27 & 4.74 \\ (6.54) & (6.74) & (5.01) & (6.54) \\ 4.67 & 4.65 & 4.66 & 9.33 \\ (6.67) & (6.53) & (6.48) & (4.93) \end{bmatrix}$	$\begin{bmatrix} 9.74 & 4.91 & . & . \\ (2.31) & (4.30) & & \\ 4.87 & 9.74 & . & . \\ (4.43) & (2.18) & & \\ 4.86 & 4.89 & 9.63 & 4.84 \\ (2.51) & (2.44) & (2.17) & (4.41) \\ 4.87 & 4.91 & 4.84 & 9.64 \\ (2.56) & (2.45) & (4.24) & (2.34) \end{bmatrix}$
	#Mo	5.00	22.00	14.00
	Bias	-0.1649	-0.3314	-0.1878
	MSE	3.25	5.41	3.03
		n=2	n=4	
		GMM	GMM	block-recursive GMM
$T = 250$	$\hat{B}$	$\begin{bmatrix} 9.88 & 4.90 \\ (1.10) & (2.30) \\ 4.98 & 9.85 \\ (2.22) & (1.13) \end{bmatrix}$	$\begin{bmatrix} 9.56 & 4.79 & 0.02 & 0.06 \\ (1.64) & (2.77) & (3.19) & (3.21) \\ 4.77 & 9.54 & -0.01 & 0.04 \\ (2.69) & (1.65) & (3.14) & (3.26) \\ 4.74 & 4.83 & 9.56 & 4.83 \\ (4.05) & (3.94) & (2.76) & (3.92) \\ 4.74 & 4.82 & 4.79 & 9.61 \\ (4.10) & (3.91) & (3.86) & (2.85) \end{bmatrix}$	$\begin{bmatrix} 9.87 & 4.91 & . & . \\ (1.07) & (2.41) & & \\ 4.94 & 9.83 & . & . \\ (2.33) & (1.15) & & \\ 4.93 & 4.91 & 9.81 & 4.92 \\ (1.16) & (1.20) & (1.13) & (2.30) \\ 4.94 & 4.92 & 4.91 & 9.84 \\ (1.14) & (1.21) & (2.32) & (1.09) \end{bmatrix}$
	#Mo	5.00	22.00	14.00
	Bias	-0.0982	-0.2065	-0.1069
	MSE	1.69	3.18	1.54
		n=2	n=4	
		GMM	GMM	block-recursive GMM
$T = 1000$	$\hat{B}$	$\begin{bmatrix} 9.96 & 5.00 \\ (0.24) & (0.46) \\ 4.97 & 9.97 \\ (0.48) & (0.22) \end{bmatrix}$	$\begin{bmatrix} 9.92 & 4.99 & 0.00 & 0.02 \\ (0.26) & (0.53) & (0.64) & (0.54) \\ 4.95 & 9.94 & 0.00 & 0.02 \\ (0.51) & (0.29) & (0.61) & (0.53) \\ 4.95 & 4.99 & 9.92 & 4.99 \\ (0.73) & (0.72) & (0.56) & (0.65) \\ 4.95 & 4.99 & 4.96 & 9.95 \\ (0.69) & (0.66) & (0.75) & (0.43) \end{bmatrix}$	$\begin{bmatrix} 9.97 & 5.02 & . & . \\ (0.22) & (0.48) & & \\ 4.97 & 9.99 & . & . \\ (0.46) & (0.24) & & \\ 4.98 & 5.01 & 9.96 & 4.99 \\ (0.25) & (0.28) & (0.21) & (0.40) \\ 4.98 & 5.01 & 4.98 & 9.97 \\ (0.25) & (0.27) & (0.41) & (0.20) \end{bmatrix}$
	#Mo	5.00	22.00	14.00
	Bias	-0.0262	-0.0295	-0.0124
	MSE	0.35	0.57	0.31

Note: The table reports the average $\bar{b}_{ij}$ and the corresponding estimated MSE (in parentheses) of each estimated element in $\hat{B}$ as well as the BIAS and MSE across estimated elements in $\hat{B}$ over $3,500$ Monte Carlo replicates. We estimate the GMM estimator without restrictions for $n = 2$ and $n = 4$, and the block-recursive GMM estimator for $n = 4$, which uses zero restrictions highlighted by the dots.

For each estimator, the average bias and MSE decreases with the sample size. Furthermore, the simulation highlights how the performance of the GMM estimators, which are based entirely on non-Gaussianity, decreases with an increasing model size (e.g., the average bias and MSE for each sample size is up to 2.1 and 1.9 times higher for the GMM estimator with $n = 4$ compared to the GMM estimator with $n = 2$). The Monte Carlo study illustrates how in a typical macroeconomic application, which rarely or if at all contains more than a few hundred observations, data-driven estimates based on non-Gaussianity become less reliable the more variables the SVAR contains. However, the simulation also stresses that exploiting the block-recursive structure annihilates the deterioration of the performance induced by a larger model. That is, the average bias and MSE for each sample size in Table 5.1 is at least 1.8 and 1.8 times higher for the GMM estimator with $n = 4$ compared to the block-recursive GMM estimator with $n = 4$. Using the block-recursive structure allows to identify the four elements on the lower left of B_0 (each with a value of 5) only by covariance moment conditions (which explains why the average MSE of the block-recursive GMM estimator with $n = 4$ even can be lower than or comparable to the GMM estimator with $n = 2$, which relies on higher-order moment conditions).

Our results suggest that if in a given application well-justified restrictions are available, these restrictions should be used as they substantially improve the performance of the estimator.

5.5.2 Recursive Structure

In this subsection, we simulate a recursive SVAR using $n = 4$ variables and

$$B_0 = \begin{pmatrix} 10 & 0 & 0 & 0 \\ 5 & 10 & 0 & 0 \\ 5 & 5 & 10 & 0 \\ 5 & 5 & 5 & 10 \end{pmatrix}.$$

For the estimation of B_0, we impose a recursive order for all considered estimators, i.e., we use zero restrictions for all elements where B_0 is zero. In this setup, the pGMM, GMM-Oracle, and the oGMM estimator are efficient estimators and have a smaller asymptotic variance than the GMM estimator, which is equivalent to the estimator obtained by applying a Cholesky decomposition. By using a recursive structure, we can apply Proposition 5.3 to calculate whether an overidentifying moment condition is relevant or redundant. Therefore, we can analyze whether the pGMM estimator selects relevant moment conditions and does not select redundant moment conditions. With the imposed recursive order, the identifying moment conditions $E[f_{\mathbf{N}}(B, u_t)]$ contain 10 and the overidentifying conditions $E[f_{\mathbf{D}}(B, u_t)]$ contain 47 conditions. All moment conditions in $E[f_{\mathbf{D}}(B, u_t)]$ are valid. More precisely, 17 of overidentifying conditions are redundant and 30 overidentifying conditions are relevant.

The construction of the weights for the pGMM estimator as in Equation (5.12)

requires an initial consistent estimate $\hat{B}$ to estimate β^* and the asymptotic variance in Equation (5.13). To this end, we apply the GMM estimator, which is the Cholesky estimator in this case. Moreover, we again use the assumption of independent shocks to estimate the asymptotic variance, as proposed by Keweloh (2021*a*). We use $r_1 = 2$ and $r_2 = 1$ in Equation (5.12) and additionally, we normalize the weights such that they sum to one, i.e., we use $\omega_j^* := \omega_j / \sum_{k \in \widetilde{D}} \omega_k$, allowing for straightforward comparison among the weights.

We choose the optimal λ for the pGMM estimator with 5-fold cross-validation from a sequence of 10 potential values. The maximum value of the sequence of λ's depends on the sample size, ensuring that it is large enough to select all moments j for which $\omega_j^* > 10^{-4}$.[12] We also include $\lambda = 0$ in the range of possible values to allow our estimator to simplify to the recursive SVAR. The selection of the optimal tuning parameter is based on the median of the GMM loss of each left-out fold.

Table 5.2 summarizes the results of $M = 3,500$ Monte Carlo simulations. We report the same summary statistics as in Table 5.1. In addition, we calculate the average number of moments selected by the pGMM estimator and the median of the chosen λ's for the pGMM estimator across Monte Carlo runs. In Appendix 5.B.1, we display results including the Post-pGMM estimator which uses the moments selected by pGMM in a second stage estimation.

The GMM estimator performs well in the smallest sample size in terms of bias and MSE. However, the GMM estimator is asymptotically inefficient and has the largest MSE among all considered estimators for $T = 250$ and $T = 1000$. Due to many moments, the oGMM estimator performs worst in terms of bias and MSE among the considered estimators for $T = 100$. Yet, its performance improves with sample size and it eventually outperforms the GMM estimator in terms of MSE. The bias is highest for the oGMM and GMM-Oracle estimator across sample sizes, which might be explained by the greater number of moments used by these estimators. Note that both estimators are asymptotically efficient. Nevertheless, many moment conditions can still lead to a finite sample bias. The MSE of the GMM-Oracle estimator is already comparable to the GMM estimator in small samples. Relative to the other estimators, its MSE further decreases with the sample size and it performs best in the largest sample size. In general, the GMM-Oracle estimator is infeasible since the redundant moments are unknown a priori.[13] In contrast to that, the pGMM estimator is feasible and uses a data-driven approach to select the relevant and valid moments. The pGMM estimator performs well across all sample sizes in terms of bias and MSE. For $T = 100$, its bias and MSE is notably smaller than the one of the oGMM and the GMM-Oracle estimator and surprisingly, also smaller than the one of the GMM estimator. In the largest sample, the

[12] We specify the maximum value of the sequence of λ's in a data-driven way using the subgradient of Equation (5.11) with respect to β. We give more details on how to construct the maximum value of the sequence of λ's in the cross-validation in Appendix 5.A.6.

[13] Even if we knew the non-Gaussianity of the shocks, we would not be able to derive the oracle estimator if the block-recursive structure was not just purely recursive. In this case, we still lack the information on which moments are redundant and which are relevant.

pGMM estimator performs similar to the oGMM and GMM-Oracle estimator in terms of MSE and best in terms of bias.[14] The simulation shows that the pGMM estimator can, without prior specification, distinguish informative from non-informative overidentifying moments, which solves the many moments problem of the oGMM estimator and allows

Table 5.2: Finite Sample Performance of the pGMM estimator.

		GMM	oGMM	GMM-Oracle	pGMM
$T = 100$	$\hat{B}$	$\begin{bmatrix} 9.93 & . & . & . \\ (1.09) & & & \\ 4.98 & 9.86 & . & . \\ (1.21) & (1.02) & & \\ 4.97 & 4.95 & 9.83 & . \\ (1.49) & (1.29) & (1.12) & \\ 4.96 & 4.93 & 4.91 & 9.78 \\ (1.71) & (1.46) & (1.27) & (1.08) \end{bmatrix}$	$\begin{bmatrix} 9.77 & . & . & . \\ (1.07) & & & \\ 4.90 & 9.71 & . & . \\ (1.31) & (1.01) & & \\ 4.89 & 4.88 & 9.70 & . \\ (1.69) & (1.43) & (1.10) & \\ 4.90 & 4.88 & 4.88 & 9.69 \\ (2.07) & (1.74) & (1.46) & (1.09) \end{bmatrix}$	$\begin{bmatrix} 9.76 & . & . & . \\ (1.07) & & & \\ 4.91 & 9.70 & . & . \\ (1.17) & (1.02) & & \\ 4.91 & 4.88 & 9.69 & . \\ (1.50) & (1.26) & (1.10) & \\ 4.92 & 4.88 & 4.88 & 9.67 \\ (1.81) & (1.51) & (1.25) & (1.10) \end{bmatrix}$	$\begin{bmatrix} 9.96 & . & . & . \\ (1.09) & & & \\ 5.00 & 9.88 & . & . \\ (1.15) & (1.01) & & \\ 4.98 & 4.96 & 9.85 & . \\ (1.46) & (1.22) & (1.11) & \\ 4.99 & 4.96 & 4.95 & 9.82 \\ (1.71) & (1.42) & (1.21) & (1.10) \end{bmatrix}$
	#Mo	10.00	57.00	40.00	24.22
	Bias	−0.0883	−0.1806	−0.1804	−0.0650
	MSE	1.27	1.40	1.28	1.25
	λ	.	.	.	71.08
		GMM	oGMM	GMM-Oracle	pGMM
$T = 250$	$\hat{B}$	$\begin{bmatrix} 9.97 & . & . & . \\ (0.43) & & & \\ 4.99 & 9.96 & . & . \\ (0.51) & (0.43) & & \\ 4.98 & 5.00 & 9.93 & . \\ (0.64) & (0.52) & (0.45) & \\ 4.98 & 4.99 & 4.98 & 9.91 \\ (0.72) & (0.61) & (0.51) & (0.45) \end{bmatrix}$	$\begin{bmatrix} 9.90 & . & . & . \\ (0.40) & & & \\ 4.96 & 9.90 & . & . \\ (0.49) & (0.40) & & \\ 4.96 & 4.97 & 9.87 & . \\ (0.65) & (0.51) & (0.42) & \\ 4.97 & 4.96 & 4.97 & 9.86 \\ (0.73) & (0.61) & (0.49) & (0.42) \end{bmatrix}$	$\begin{bmatrix} 9.90 & . & . & . \\ (0.40) & & & \\ 4.97 & 9.90 & . & . \\ (0.44) & (0.40) & & \\ 4.97 & 4.97 & 9.87 & . \\ (0.59) & (0.46) & (0.42) & \\ 4.98 & 4.97 & 4.96 & 9.85 \\ (0.65) & (0.54) & (0.44) & (0.42) \end{bmatrix}$	$\begin{bmatrix} 9.99 & . & . & . \\ (0.42) & & & \\ 5.01 & 9.97 & . & . \\ (0.45) & (0.41) & & \\ 5.01 & 5.02 & 9.94 & . \\ (0.59) & (0.46) & (0.42) & \\ 5.02 & 5.01 & 5.00 & 9.92 \\ (0.66) & (0.55) & (0.44) & (0.43) \end{bmatrix}$
	#Mo	10.00	57.00	40.00	27.20
	Bias	−0.0311	−0.0676	−0.0656	−0.0114
	MSE	0.53	0.51	0.48	0.48
	λ	.	.	.	118.92
		GMM	oGMM	GMM-Oracle	pGMM
$T = 1000$	$\hat{B}$	$\begin{bmatrix} 10.00 & . & . & . \\ (0.11) & & & \\ 5.00 & 9.99 & . & . \\ (0.13) & (0.11) & & \\ 4.99 & 4.99 & 9.99 & . \\ (0.15) & (0.13) & (0.11) & \\ 4.99 & 4.99 & 4.99 & 9.98 \\ (0.19) & (0.15) & (0.13) & (0.11) \end{bmatrix}$	$\begin{bmatrix} 9.98 & . & . & . \\ (0.10) & & & \\ 4.99 & 9.97 & . & . \\ (0.12) & (0.10) & & \\ 4.99 & 4.99 & 9.98 & . \\ (0.13) & (0.11) & (0.10) & \\ 4.99 & 4.99 & 4.99 & 9.97 \\ (0.16) & (0.14) & (0.11) & (0.10) \end{bmatrix}$	$\begin{bmatrix} 9.98 & . & . & . \\ (0.10) & & & \\ 4.99 & 9.97 & . & . \\ (0.11) & (0.10) & & \\ 4.99 & 4.99 & 9.98 & . \\ (0.13) & (0.10) & (0.10) & \\ 4.99 & 4.99 & 4.99 & 9.97 \\ (0.15) & (0.13) & (0.11) & (0.10) \end{bmatrix}$	$\begin{bmatrix} 10.00 & . & . & . \\ (0.11) & & & \\ 5.00 & 9.99 & . & . \\ (0.11) & (0.10) & & \\ 5.00 & 5.00 & 10.00 & . \\ (0.13) & (0.11) & (0.10) & \\ 5.00 & 5.00 & 5.00 & 9.98 \\ (0.16) & (0.13) & (0.11) & (0.10) \end{bmatrix}$
	#Mo	10.00	57.00	40.00	29.59
	Bias	−0.0076	−0.0158	−0.0158	−0.0021
	MSE	0.13	0.12	0.11	0.12
	λ	.	.	.	75.34

Note: The table reports the average $\bar{b}_{ij}$ and the corresponding estimated MSE (in parentheses) of each estimated element in $\hat{B}$ as well as the BIAS and MSE across estimated elements in $\hat{B}$ over $3,500$ Monte Carlo replicates for the GMM estimator, the oGMM estimator, the GMM-Oracle estimator, and the pGMM estimator. All estimator use zero restrictions which are highlighted by the dots.

[14]The Post-pGMM estimator reported in Appendix 5.B.1 performs similar to the pGMM estimator.

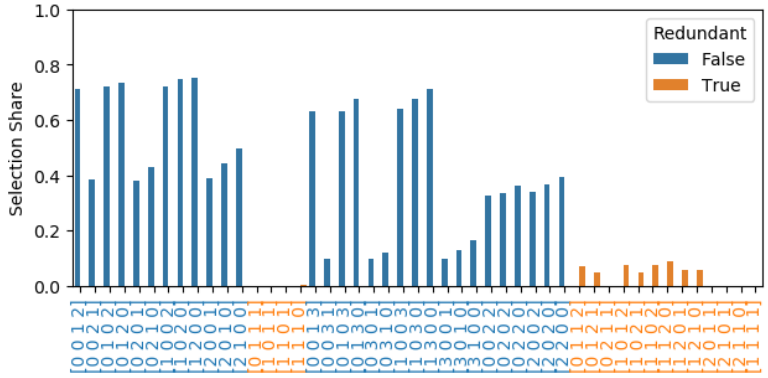
to exploit information in overidentifying higher-order moments already in small samples.

Table 5.2 indicates that the average number of selected moments increases only slightly as T increases. Even for $T = 1000$, the pGMM estimator only selects 20 out of 30 valid and relevant overidentifying moments in addition to the 10 identifying moments. That said, the remaining 10 moments would only decrease the MSE from 0.12 to 0.11, indicating that the moments not being selected would not lower the MSE much. Figure 5.2 illustrates that pGMM estimator only selects relevant moments and manages to leave out redundant moments, especially as T increases. Moreover, the share of selections of each moment across all Monte Carlo runs rises with the sample size for the majority of relevant moments. In Figure 5.B.2, we plot the average weight of each moment across Monte Carlo runs. By comparing Figure 5.2 and Figure 5.B.2, we argue that there is a clear correlation between the average weight and the number of selections of each moment. More precisely, all redundant moments have an average weight which is very close to zero and hence, they are not selected by the pGMM estimator.

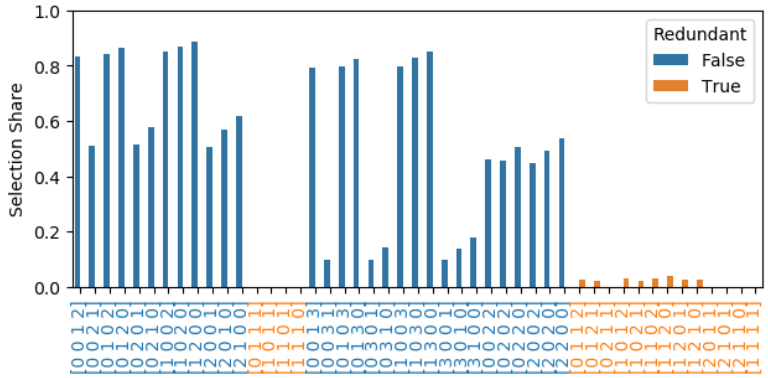
Figure 5.3 highlights the influence of λ on β and hence, on the number of selected moment conditions for one Monte Carlo run.[15] For instance, for $\log(\lambda) = -6$ no overidentifying moment conditions are selected and the solution of the pGMM estimator corresponds to the one of the GMM estimator. Further, the number of selected moments increases as λ increases, i.e., the penalty shrinks the elements of β to zero. Furthermore, the relevant moments get selected first when λ increases and we do not select any redundant moment until λ becomes very large.

[15] For the purpose of illustration, we use a wider range of of λ values for this plot.

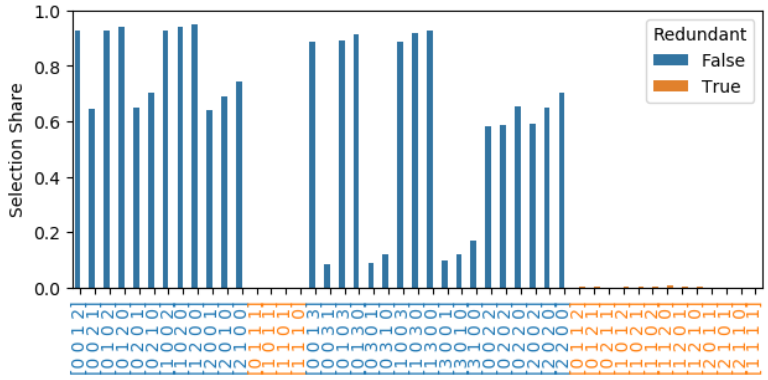
Figure 5.2: Share of Selections of Moments across Monte Carlo Runs



(a) $T = 100$



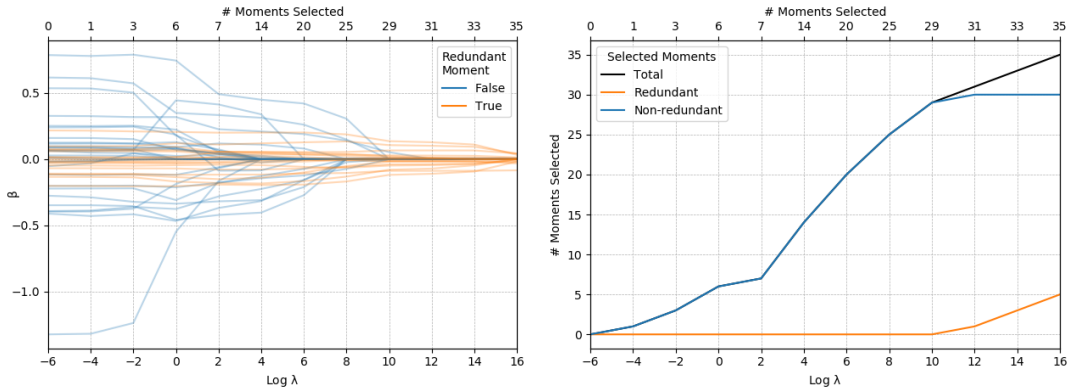
(b) $T = 250$



(c) $T = 1000$

Note: The figure shows how often each moment gets selected across $M = 3,500$ Monte Carlos simulations. Redundant moments (orange) and relevant moments (blue) are displayed on the x-axis. Each x-axis label abbreviates a moment condition, e.g., $[0, 1, 2, 1]$ corresponds to $E[e(B)^0_{1,t}\, e(B)^1_{2,t}\, e(B)^2_{3,t}\, e(B)^1_{4,t}]$.

Figure 5.3: Illustration of Influence of λ on β.



(a) Trace Plot

(b) Selected Moments

Note: Panel (a) of the figure shows the values of β in dependence on $log(\lambda)$ for one Monte Carlo run for $T = 100$ and the corresponding number of selected moments in $\widetilde{D}$. Panel (b) of the figure splits the number of selected moments into the number of selected redundant and the number of selected relevant moments for each $log(\lambda)$.

5.6 Application of the block-recursive SVAR: Disentangling speculative demand and supply shocks in the oil market

In this section, we propose a SVAR model for the oil market to analyze the impact of flow and speculative supply and of flow and speculative demand shocks on the real oil price. A flow supply shock for oil represents an exogenous deviation in the present amount of oil coming out of the ground and a flow demand shock for oil an exogenous deviation in the present amount of oil being consumed. A speculative oil supply shock represents a shift in the expected future oil supply and a speculative oil demand shock a shift in the expected future oil demand.

We consider a SVAR with monthly data from January 1974 to December 2019 of the form

$$\begin{bmatrix} O_t \\ Y_t \\ OP_t \\ SR_t \end{bmatrix} = \alpha + \sum_{i=1}^{12} A_i \begin{bmatrix} O_{t-i} \\ Y_{t-i} \\ OP_{t-i} \\ SR_{t-i} \end{bmatrix} + \begin{bmatrix} u_t^O \\ u_t^Y \\ u_t^{OP} \\ u_t^{SR} \end{bmatrix}. \tag{5.15}$$

The variable O_t is the log difference of global oil production, Y_t is the log difference of industrial production, measuring economic activity, OP_t is the growth rate of real oil price, and SR_t are real monthly stock returns.[16] We decompose the reduced form shocks

[16]Global oil production is given by the global crude oil including lease condensate production obtained from the U.S. EIA. We obtain industrial production by the monthly industrial production index in the OECD and six major other countries from Baumeister and Hamilton (2019). The real oil price is equal to

u_t into four structural shocks with

$$\begin{bmatrix} u_t^O \\ u_t^Y \\ u_t^{OP} \\ u_t^{SR} \end{bmatrix} = \begin{bmatrix} b_{11} & b_{12} & 0 & 0 \\ b_{21} & b_{22} & 0 & 0 \\ b_{31} & b_{32} & b_{33} & b_{34} \\ b_{41} & b_{42} & b_{43} & b_{41} \end{bmatrix} \begin{bmatrix} \varepsilon_t^s \\ \varepsilon_t^d \\ \varepsilon_t^{s-exp} \\ \varepsilon_t^{d-exp} \end{bmatrix}, \tag{5.16}$$

where ε_t^s is a flow supply shock for oil, ε_t^d is a flow demand shock for oil, ε_t^{s-exp} is a speculative oil supply shock, and ε_t^{d-exp} is a speculative oil demand shock. The block-recursive restrictions in Equation (5.16) imply that oil production and economic activity behave sluggishly and can contemporaneously only respond to flow supply and demand shocks, whereas oil prices and stock returns can immediately incorporate all available information and contemporaneously respond to flow and speculative supply and demand shocks.

The simultaneous relationship is estimated using the block-recursive SVAR pGMM estimator.[17] In line with the Monte Carlo simulations, we apply continuous updating for the weighting matrix and use the assumption of serially and mutually independent shocks to estimate the asymptotically efficient weighting matrix as proposed by Keweloh (2021*a*). With the imposed block-recursive structure, we can divide the moment conditions into 14 identifying conditions $E[f_{\mathbf{N}}(B, u_t)]$ and 43 overidentifying conditions $E[f_{\mathbf{D}}(B, u_t)]$. We use the same specifications to construct the weights as in the Monte Carlo simulation, i.e., we use $r_1 = 2$ and $r_2 = 1$ in Equation (5.12). For the cross-validation, we consider a range of 28 values for λ, including $\lambda = 0$. The maximum value of λ is chosen such that all conditions $E[f_{\mathbf{D}}(B, u_t)]$ for which $\omega_j / \sum_{k \in \widetilde{D}} \omega_k > 10^{-7}$ get selected. With the chosen $\lambda = 34679$, which is the 27th value of the considered sequence, 12 coskewness and 12 cokurtosis conditions are selected.[18]

For each estimated structural shock, Table 5.3 shows the estimated skewness, kurtosis and p-value of the Jarque-Bera test. To ensure identification, at most one structural shock in each block may be Gaussian. With our block-recursive structure, each block contains only two shocks and, therefore, it is sufficient for identification to show that at least one structural shock in each block is non-Gaussian. Furthermore, the block-recursive structure implies that each of the two unmixed innovations in the first block is equal to a linear combination of the two structural shocks in the first block, i.e., if both structural shocks are Gaussian, the two unmixed innovations have to be Gaussian as well. However,

the refiner's acquisition cost of imported crude oil from the U.S. EIA deflated by the U.S. CPI. Real stock prices correspond to the aggregate U.S. stock index constructed by the OECD deflated by the U.S. CPI.

[17] In Appendix 5.B.2, we conduct various robustness checks. In particular, we estimate the block-recursive SVAR using the GMM estimator from Equation (5.9) and the overidentified GMM estimator from Equation (5.10). Estimates using the white fast SVAR GMM estimator proposed by Keweloh (2021*b*) and the PML estimator proposed by Gouriéroux et al. (2017) are qualitatively similar and available on request. Additionally, we report results for different specifications of the variables in the block-recursive SVAR.

[18] Additionally, we compute the block-recursive SVAR pGMM estimator using the plugin rule $\lambda = k_{\mathbf{D}}^{r_2/4} T^{(-0.5-r_2/4)}$, where $k_{\mathbf{D}}$ denotes the number of overidentifying moment conditions, see Cheng and Liao (2015). The estimator selects 8 coskewness and 6 cokurtosis conditions.

the skewness, kurtosis, and the Jarque-Bera test suggest that the unmixed innovations in the first block are non-Gaussian and, hence, that at least one structural shock in the first block is non-Gaussian. Consequently, the first block is identified. Moreover, the unmixed innovations in the second block are equal to a linear combination of the structural shocks in the second block (the argument follows from Equation (5.A.3) in the proof of Proposition 5.2). Again, the skewness, kurtosis and the Jarque-Bera test suggest that the unmixed innovations in the second block are non-Gaussian, implying that at least one structural shock in the second block is non-Gaussian. Thus, the second block is also identified. Consequently, the block-recursive SVAR is identified.

Table 5.3: Non-Gaussianity of the estimated structural shocks

	ε_t^s	ε_t^d	ε_t^{s-exp}	ε_t^{d-ext}
Skewness	−0.97	−0.21	0.46	−0.82
Kurtosis	9.92	4.58	6.79	6.88
JB-Test	0.00	0.00	0.00	0.00

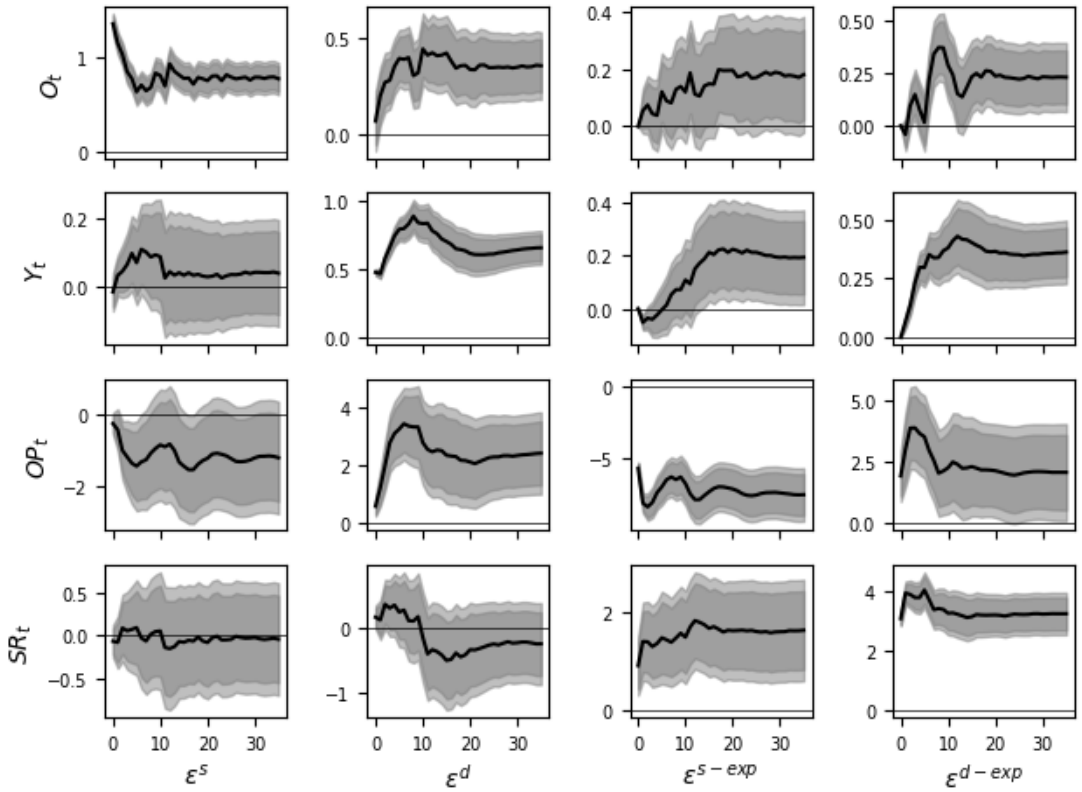
Note: Skewness, kurtosis and the p-value of the Jarque-Bera test.

In Figure 5.4, we show impulse response functions (IRFs). With the block-recursive structure, labeling of the shocks in the plot of the IRFs is straightforward. In the first block, there is only one shock which leads to a significant immediate increase of economic activity and, thus, an immediate increase in demand for oil. We label this shock as the flow demand shock and the remaining shock in the first block as the flow supply shock. In the second block, one shock leads to an immediate increase of the real oil price and to a long-run increase of economic activity. We label this shock as the speculative oil demand shock. The remaining shock in the second block leads to an immediate decrease of the oil price and to an increase of economic activity and oil production in the long-run, which corresponds to the speculative oil supply shock.

Our results show that flow supply shocks immediately increase oil production and decrease the real oil price and flow demand shocks increase economic activity and the real oil price. Moreover, oil production responds to the demand shock with a lagged increase. Interestingly, it seems that real stock returns do not respond significantly to flow demand and supply shocks. With respect to the speculative shocks, we find that a supply expectation shock leads to an increase of oil production and of economic activity after one year. Furthermore, it immediately and permanently decreases the real oil price and increases real stock returns. A speculative demand shock increases oil production and economic activity. Additionally, the speculative demand shocks leads to an immediate increase of the real oil price and of real stock returns.

Figure 5.5 shows the contribution of the estimated structural shocks to the evolution of the real oil price. Figure 5.B.3 in Appendix 5.B.2 shows the historical evolution of the real oil price. Figure 5.5 suggests that the increase of the real oil price from 1978 to 1981 is mainly driven by flow supply and speculative supply shocks. Moreover, we find that the decline of the real oil price from 1981 to 1985 is largely explained by speculative

Figure 5.4: Impulse Responses of the block-recursive SVAR pGMM estimator.

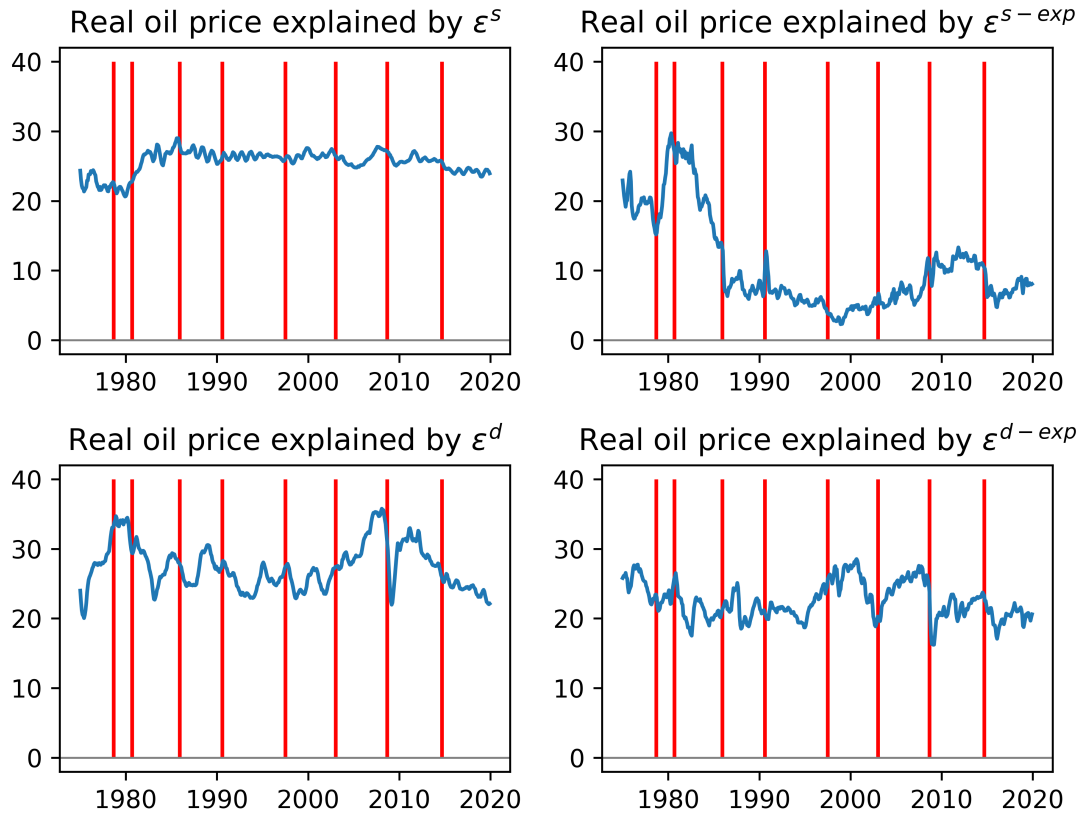


Note: Impulse responses to the estimated structural shocks for the block-recursive SVAR pGMM estimator. Confidence bands are symmetric 68% and 80% bands based on standard errors and 1000 replications. The rows show the cumulative responses. The x-axis displays monthly lags.

supply shocks. Additionally, the decrease in real oil prices after the collapse of OPEC in 1985 and the peak of real oil prices during the Persian Gulf War in 1990 can to a large extent be explained explained by speculative supply shocks. The run-up in the real oil prices from 2003 to 2008 is driven by flow demand, speculative demand, and speculative supply shocks. Flow demand and speculative demand shocks explain the plunge of the real oil price during the financial crisis in 2008. Additionally, most of the recovery of the real oil price after the financial crisis is explained by demand shocks. The collapse of the real oil price since mid 2014 is related to flow demand, speculative demand, and speculative supply shocks.

The IRFs in Figure 5.4 show no evidence against a recursive structure of the shocks in the first block. That said, our results clearly suggest that the second block does not have a recursive structure since the two structural shocks in the second block have an immediate impact on both reduced form shocks in the second block. As a robustness-check and to illustrate the impact of misspecification in the second block, we estimate a recursive specification as proposed in Kilian and Park (2009). That is, we restrict b_{12} and b_{34}

Figure 5.5: Real oil price evolution explained by the estimated structural shocks.



Note: In each of the panels, we simulate the real oil price (blue line) by setting all but one of the shocks to zero (and for ease of interpretation, we also set $\alpha = 0$ in Equation (5.15)). The red vertical bars indicate the following events: Iranian Revolution (1978 : 9), Iran Iraq War (1980 : 9), collapse of OPEC (1985 : 12), Persian Gulf War (1990 : 8), Asian Financial Crisis of (1997 : 7), Iraq War (2003 : 1), the collapse of Lehman Brothers (2008 : 9), and the oil price decline in mid 2014.

in Equation (5.16) to zero. In this case, the interpretation of the shocks changes and we refer to the third and fourth shock as speculative oil price shock and residual stock market shock, respectively.

Figure 5.B.5 in Appendix 5.B.2 displays the IRFs of the recursive SVAR. The response of the real oil price to flow supply and demand shocks in the recursive model is similar to the the one in the block-recursive model. The speculative oil price shock leads to an decrease of the real oil price. However, none of the remaining variables shows any significant response to the speculative oil price shock, except for economic activity which shows a small negative reaction in the first seven month. In the recursive SVAR for the oil market, we cannot distinguish between speculative supply and speculative demand shocks. Rather, the speculative oil price shock contains a mixture of the speculative supply and speculative demand shock. However, the impact of the speculative oil price

shocks on oil production and the economy should depend on the source of the speculative oil price shock and, thus, it is not surprising that we are unable to find a clear response of oil production, economic activity, and the stock market to the speculative oil price shock in the recursive specification.

As a further robustness-check, we estimate the SVAR without any restrictions on the interaction, i.e., we estimate the model without the zero restrictions given in Equation (5.16). In this case, the labeling of the shocks is the same as in Equation (5.16). However, the difference is that oil production and economic activity can now contemporaneously respond to speculative supply and demand shocks. Figure 5.B.6 and Figure 5.B.7 in Appendix 5.B.2 show the corresponding IRFs. Overall, the unrestricted responses in Figure 5.B.6 are comparable to the block-recursive responses in Figure 5.4. However, the confidence bands are broader and there is no significant response of the real oil price to flow supply and (almost) no significant response to flow demand shocks.

5.7 Conclusion

For a non-Gaussian block-recursive SVAR, we derive a small set of identifying moment conditions based on the assumption of mean independent shocks. Moreover, we derive overidentifying moment conditions, some of these require mean independent shocks and some of these additionally require independent shocks. We show that the overidentifying conditions can decrease the asymptotic variance of the block-recursive SVAR estimator. In particular, we prove that the frequently applied Cholesky estimator can be inefficient. Since some of the overidentifying moment conditions may be redundant, i.e., may not decrease the asymptotic variance, or be invalid, i.e., may lead to inconsistent estimates, we employ the block-recursive SVAR pGMM estimator to select only the relevant and valid overidentifying moment conditions.

We demonstrate in a Monte Carlo experiment that imposing a block-recursive structure substantially increases the finite sample performance compared to unrestricted estimators. Furthermore, a second Monte Carlo experiment highlights that, for a given block-recursive structure, the block-recursive SVAR pGMM estimator selects only relevant moment conditions and thereby, increases finite sample precision compared to the block-recursive SVAR GMM estimator and overidentified block-recursive SVAR GMM estimator.

Our application analyzes the impact of flow and speculative supply and flow and speculative demand shocks in the oil market. We argue that there are some but not enough well-justified restrictions available to identify the SVAR based on second moments. Traditional approaches would either rely on additional less credible restrictions or refrain from using any restrictions and solely rely on non-Gaussianity. The proposed block-recursive estimator allows to utilize only the well-justified restrictions and, therefore, offers a compromise between both approaches. The application illustrates that by combining data-driven identification with traditional zero restrictions we are able to gain deeper insights into the transmission of demand and supply shocks in the oil market.

Appendix 5.A Supplementary Notation and Proofs

We include the formulas in Appendix 5.A.1 and 5.A.2 for completeness, even though they are standard textbook results (cf. Hall (2005)).

5.A.1 Asymptotic variance of the block-recursive SVAR GMM estimator

The asymptotic variance of the block-recursive SVAR GMM estimator defined in Equation (5.9) is given by

$$V_{\mathbf{N}} := M_{\mathbf{N}} S_{\mathbf{N}} M_{\mathbf{N}}'$$

where

$$M_{\mathbf{N}} := \left(G_{\mathbf{N}}' W_{\mathbf{N}} G_{\mathbf{N}}\right)^{-1} G_{\mathbf{N}}' W_{\mathbf{N}}, \qquad S_{\mathbf{N}} := \lim_{T\to\infty} E\left[T g_{\mathbf{N}}(B_0) g_{\mathbf{N}}(B_0)\right],$$

$$G_{\mathbf{N}} := E\left[\frac{\partial f_{\mathbf{N}}(B_0, u_t)}{\partial vec(B)'}\right].$$

Consequently, using the weighting matrix $W_{\mathbf{N}}^* := S_{\mathbf{N}}^{-1}$ leads to the estimator $\hat{B}^*$ with the asymptotic variance

$$V_{\mathbf{N}}^* := (G_{\mathbf{N}}' S_{\mathbf{N}}^{-1} G_{\mathbf{N}})^{-1},$$

which is the lowest possible asymptotic variance (see Hall (2005)).

5.A.2 Asymptotic variance of the (overidentified) block-recursive SVAR GMM estimator

The asymptotic variance of the overidentified block-recursive SVAR GMM estimator defined in Equation (5.10) is given by

$$V_{\mathbf{N+D}} := M_{\mathbf{N+D}} S_{\mathbf{N+D}} M_{\mathbf{N+D}}', \tag{5.A.1}$$

where

$$M_{\mathbf{N+D}} := \left(G_{\mathbf{N+D}}' W_{\mathbf{N+D}} G\right)^{-1} G_{\mathbf{N+D}}' W_{\mathbf{N+D}}, \qquad S_{\mathbf{N+D}} := \lim_{T\to\infty} E\left[g_{\mathbf{N+D}}(B_0) g_{\mathbf{N+D}}(B_0)'\right],$$

$$G_{\mathbf{N+D}} := \begin{bmatrix} G_{\mathbf{N}} \\ G_{\mathbf{D}} \end{bmatrix}, \qquad g_{\mathbf{N+D}}(B_0) := \begin{bmatrix} g_{\mathbf{N}}(B_0) \\ g_{\mathbf{D}}(B_0) \end{bmatrix},$$

$$G_{\mathbf{D}} := E\left[\frac{\partial f_{\mathbf{D}}(B_0, u_t)}{\partial vec(B)'}\right].$$

Using the weighting matrix $W^*_{\mathbf{N+D}} := S^{-1}_{\mathbf{N+D}}$ leads to the estimator $\hat{B}^*_{\mathbf{N+D}}$ with the asymptotic variance

$$V^*_{\mathbf{N+D}} := (G'_{\mathbf{N+D}} S^{-1}_{\mathbf{N+D}} G_{\mathbf{N+D}})^{-1}, \tag{5.A.2}$$

which is the lowest possible asymptotic variance (see Hall (2005)). To construct $V_{\mathbf{N+D}_j}$ and $V^*_{\mathbf{N+D}_j}$, $j \in \widetilde{D}$, we replace the moment conditions $f_{\mathbf{D}_j}(B, u_t)$ by moment condition $f_{\mathbf{D}_j}(B, u_t)$, $j \in \widetilde{D}$, in Equation (5.A.1) and (5.A.2).

5.A.3 Identification in the block-recursive SVAR

Proof of Proposition 5.1.
For ease of notation, we omit the time index t and w.l.o.g., consider an example with two blocks[19]

$$\begin{bmatrix} u_{p_1} \\ u_{p_2} \end{bmatrix} = \begin{bmatrix} B_{11,0} & 0 \\ B_{21,0} & B_{22,0} \end{bmatrix} \begin{bmatrix} \varepsilon_{p_1} \\ \varepsilon_{p_2} \end{bmatrix} \quad \text{and} \quad B = \begin{bmatrix} B_{11} & 0 \\ B_{21} & B_{22} \end{bmatrix},$$

where u_{p_1} and u_{p_2} contain the reduced form shocks of the first and second block, ε_{p_1} and ε_{p_2} contain the structural shocks of the first and second block, and $B_{11,0}$, $B_{21,0}$, $B_{22,0}$, B_{11}, B_{21}, and B_{22} are the corresponding blocks of the matrices B_0 and B.

First, let $E[f_{\mathbf{2_{p_1}}}(B, u)] = 0$ contain all (co-)variance conditions of shocks in the first block. The block-recursive structure implies that $u_{p_1} = B_{11,0}\varepsilon_{p_1}$. If at most one structural shock in the first block has zero excess kurtosis, it follows from Lanne and Luoto (2021) that the conditions containing only shocks in the first block

$$E \begin{bmatrix} f_{\mathbf{2_{p_1}}}(B, u) \\ f_{\mathbf{4_{p_1}}}(B, u) \end{bmatrix} = 0$$

locally identify $B_{11} = B_{11,0}$, the impact of the shocks in the first block on the variables in the first block.

---

[19] If the SVAR contains more than two blocks, the procedure outlined in the proof can be repeated multiple times to identify arbitrary many blocks. For example, a SVAR with three blocks

$$\begin{bmatrix} u_{p_1} \\ u_{p_2} \\ u_{p_3} \end{bmatrix} = \begin{bmatrix} B_{11,0} & 0 & 0 \\ B_{21,0} & B_{22,0} & 0 \\ B_{32,0} & B_{32,0} & B_{33,0} \end{bmatrix} \begin{bmatrix} \epsilon_{p_1} \\ \epsilon_{p_2} \\ \epsilon_{p_3} \end{bmatrix} \text{ can be written as } \begin{bmatrix} u_{p_1} \\ \tilde{u}_{p_2} \end{bmatrix} = \begin{bmatrix} B_{11,0} & 0 \\ \tilde{B}_{21,0} & \tilde{B}_{22,0} \end{bmatrix} \begin{bmatrix} \epsilon_{p_1} \\ \tilde{\epsilon}_{p_2} \end{bmatrix},$$

with $\tilde{u}_{p_2} = [u'_{p_2}, u'_{p_3}]'$, $\tilde{B}_{22,0} = \begin{bmatrix} B_{22,0} & 0 \\ B_{32,0} & B_{33,0} \end{bmatrix}$, $\tilde{B}_{21,0} = \begin{bmatrix} B_{21,0} \\ B_{31,0} \end{bmatrix}$,and $\tilde{\epsilon}_{p_2} = [\epsilon'_{p_2}, \epsilon'_{p_3}]'$. Our proof then shows how to identify $B_{11,0}$, $\tilde{B}_{21,0} = \begin{bmatrix} B_{21,0} \\ B_{31,0} \end{bmatrix}$, and ϵ_{p_1}. Defining

$$\begin{bmatrix} z_{p_2} \\ z_{p_3} \end{bmatrix} := \begin{bmatrix} u_{p_2} \\ u_{p_3} \end{bmatrix} - \begin{bmatrix} B_{21,0} \\ B_{31,0} \end{bmatrix} \epsilon_{p_1} \text{ then yields } \begin{bmatrix} z_{p_2} \\ z_{p_3} \end{bmatrix} = \begin{bmatrix} B_{22,0} & 0 \\ B_{32,0} & B_{33,0} \end{bmatrix} \begin{bmatrix} \epsilon_{p_2} \\ \epsilon_{p_3} \end{bmatrix},$$

which is another block-recursive SVAR with two blocks.

Second, let $E\left[f_{\mathbf{2_{P1P2}}}(B,u)\right]=0$ contain all covariance conditions belonging to shocks in both blocks. At the local solution $B_{11}=B_{11,0}$, the covariance conditions containing shocks of both blocks only hold if $B_{21}=B_{21,0}$. To see this, rewrite the covariance conditions as $E\left[e_{p_2}(B)e_{p_1}(B)'\right]=0$. With the partitioned inverse of B and the block-recursive structure, it holds that $e_{p_2}(B) = -B_{22}^{-1}B_{21}B_{11}^{-1}B_{11,0}\varepsilon_{p_1} + B_{22}^{-1}\left(B_{21,0}\varepsilon_{p_1}+B_{22,0}\varepsilon_{p_2}\right)$. Therefore, with $B_{11}=B_{11,0}$ it holds that

$$E\left[e_{p_2}(B)e_{p_1}(B)'\right] = -B_{22}^{-1}B_{21}E\left[\varepsilon_{p_1}\varepsilon_{p_1}'\right] + B_{22}^{-1}B_{21,0}E\left[\varepsilon_{p_1}\varepsilon_{p_1}'\right] + B_{22,0}E\left[\varepsilon_{p_2}\varepsilon_{p_1}'\right].$$

With $E\left[\varepsilon_{p_1}\varepsilon_{p_1}'\right]=I$ and $E[\varepsilon_{p_2}\varepsilon_{p_1}']=0$, the condition $E\left[e_{p_2}(B)e_{p_1}(B)'\right]=0$ implies $0=-B_{22}^{-1}\left(B_{21}-B_{21,0}\right)$ at $B_{11}=B_{11,0}$. Therefore, at the local solution $B_{11}=B_{11,0}$ the covariance conditions $E\left[f_{\mathbf{2_{P1P2}}}(B,u)\right]$, globally identify $B_{21}=B_{21,0}$ the impact of shocks in the first block on variables in the second block.

Finally, let $E[f_{\mathbf{2_{P2}}}(B,u)]=0$ contain all (co-)variance conditions of shocks in the second block. At the solution $B_{11}=B_{11,0}$ and $B_{21}=B_{21,0}$ the unmixed innovations of the second block $e_{p_2}(B)$ are mixtures of the structural shocks in the second block and are not influenced by shocks from the first block. This follows from the partitioned inverse of B and the block-recursive structure such that $e_{p_2}(B)=B_{22}^{-1}B_{22,0}\varepsilon_{p_2}$. If at most one structural shock in the second block has zero excess kurtosis, it then again follows from Lanne and Luoto (2021) that at the solution $B_{11}=B_{11,0}$ and $B_{21}=B_{21,0}$ the remaining conditions containing only shocks in the second block

$$E\begin{bmatrix} f_{\mathbf{2_{P2}}}(B,u) \\ f_{\mathbf{4_{P2}}}(B,u) \end{bmatrix} = 0$$

locally identify $B_{22}=B_{22,0}$, meaning the impact of shocks in the second block on variables in the second block.

□

Proof of Proposition 5.2.
To simplify the notation let

$$\tilde{u}_1 := [u_1,...,u_{p_i-1}]', \qquad \tilde{e}_1(B) := [e_1(B),...,e_{p_i-1}(B)]', \qquad \tilde{\varepsilon}_1 := [\varepsilon_1,...,\varepsilon_{p_i-1}]',$$

$$\tilde{u}_2 := [u_{p_i},...,u_{p_{i+1}-1}]', \quad \tilde{e}_2(B) := [e_{p_i}(B),...,e_{p_{i+1}-1}(B)]', \quad \tilde{\varepsilon}_2 := [\varepsilon_{p_i},...,\varepsilon_{p_{i+1}-1}]',$$

$$\tilde{u}_3 := [u_{p_{i+1}},...,u_n]', \qquad \tilde{e}_3(B) := [e_{p_{i+1}}(B),...,e_n(B)]', \qquad \tilde{\varepsilon}_3 := [\varepsilon_{p_{i+1}},...,\varepsilon_n]',$$

such that $\tilde{u}_1$, $\tilde{e}_1(B)$, and $\tilde{\varepsilon}_1$ contain all reduce form shocks, unmixed innovations, and structural shocks in blocks preceding the ith block of $\mathbb{B}_{brec}$, $\tilde{u}_2$, $\tilde{e}_2(B)$, and $\tilde{\varepsilon}_2$ contain the innovations and shocks in the i-th block of $\mathbb{B}_{brec}$, and $\tilde{u}_3$, $\tilde{e}_3(B)$, and $\tilde{\varepsilon}_3$ contain the innovations and shocks following block i of $\mathbb{B}_{brec}$. Moreover, we denote parts of the B_0

matrix as follows

$$\begin{bmatrix} \tilde{u}_1 \\ \tilde{u}_2 \\ \tilde{u}_3 \end{bmatrix} = \begin{bmatrix} B_{11,0} & 0 & 0 \\ B_{21,0} & B_{22,0} & 0 \\ B_{31,0} & B_{32,0} & B_{33,0} \end{bmatrix} \begin{bmatrix} \tilde{\varepsilon}_1 \\ \tilde{\varepsilon}_2 \\ \tilde{\varepsilon}_3 \end{bmatrix},$$

and B_{11}, B_{21}, B_{31}, B_{22}, B_{32}, and B_{33} denote the respective parts of a given B matrix.

With the block-recursive structure and the partitioned inverse, it holds that

$$\tilde{e}_1(B) = B_{11}^{-1} B_{11,0} \tilde{\varepsilon}_1,$$

$$\tilde{e}_2(B) = -B_{22}^{-1} B_{21} B_{11}^{-1} B_{11,0} \tilde{\varepsilon}_1 + B_{22}^{-1} \left(B_{21,0} \tilde{\varepsilon}_1 + B_{22,0} \tilde{\varepsilon}_2 \right).$$

For any matrix B satisfying $E\left[f_{\mathbf{2}}(B, u_t)\right] = 0$ and, therefore, $0 = E\left[\tilde{e}_2(B)\tilde{e}_1(B)'\right]$ it holds that $0 = -B_{22}^{-1}\left(B_{21,0} - B_{21}B_{11}^{-1}B_{11,0}\right) B_{11,0}'(B_{11}^{-1})'$ and, thus, $B_{21} = B_{21,0}B_{11,0}^{-1}B_{11}$. Any B Matrix satisfying the condition $0 = E\left[\tilde{e}_2(B)\tilde{e}_1(B)'\right]$ thus yields innovations of the second block equal to

$$\tilde{e}_2(B) = B_{22}^{-1} B_{22,0} \tilde{\varepsilon}_2, \tag{5.A.3}$$

meaning the innovations of the second block are equal to a linear combination of the structural shocks in the second block. Applying the identification result from Lanne and Luoto (2021) yields that the conditions $E\left[f_{\mathbf{4_{\tilde{P}_i}}}(B, u_t)\right] = 0$ locally identify $B_{22,0}$.

Analogously, with the block-recursive structure and the partitioned inverse it holds that

$$\begin{aligned} \tilde{e}_3(B) = &- B_{33}^{-1} \begin{bmatrix} B_{31} & B_{32} \end{bmatrix} \begin{bmatrix} B_{11} & 0 \\ B_{21} & B_{22} \end{bmatrix}^{-1} \begin{bmatrix} B_{11,0} & 0 \\ B_{21,0} & B_{22,0} \end{bmatrix} \begin{bmatrix} \tilde{\varepsilon}_1 \\ \tilde{\varepsilon}_2 \end{bmatrix} \\ &+ B_{33}^{-1} \left(\begin{bmatrix} B_{31,0} & B_{32,0} \end{bmatrix} \begin{bmatrix} \tilde{\varepsilon}_1 \\ \tilde{\varepsilon}_2 \end{bmatrix} + B_{33,0} \tilde{\varepsilon}_3 \right). \end{aligned}$$

With $B_{21} = B_{21,0}B_{11,0}^{-1}B_{11}$ it follows that

$$\begin{aligned} \tilde{e}_3(B) = &- B_{33}^{-1} \begin{bmatrix} B_{31} & B_{32} \end{bmatrix} \begin{bmatrix} B_{11}^{-1} & 0 \\ -B_{22}^{-1} B_{21,0} B_{11,0}^{-1} B_{11} B_{11}^{-1} & B_{22}^{-1} \end{bmatrix} \begin{bmatrix} B_{11,0} & 0 \\ B_{21,0} & B_{22,0} \end{bmatrix} \begin{bmatrix} \tilde{\varepsilon}_1 \\ \tilde{\varepsilon}_2 \end{bmatrix} \\ &+ B_{33}^{-1} \left(\begin{bmatrix} B_{31,0} & B_{32,0} \end{bmatrix} \begin{bmatrix} \tilde{\varepsilon}_1 \\ \tilde{\varepsilon}_2 \end{bmatrix} + B_{33,0} \tilde{\varepsilon}_3 \right) \\ = &- B_{33}^{-1} \begin{bmatrix} B_{31} & B_{32} \end{bmatrix} \begin{bmatrix} B_{11}^{-1} & 0 \\ -B_{22}^{-1} B_{21,0} B_{11,0}^{-1} & B_{22}^{-1} \end{bmatrix} \begin{bmatrix} B_{11,0} & 0 \\ B_{21,0} & B_{22,0} \end{bmatrix} \begin{bmatrix} \tilde{\varepsilon}_1 \\ \tilde{\varepsilon}_2 \end{bmatrix} \\ &+ B_{33}^{-1} \left(B_{31,0} \tilde{\varepsilon}_1 + B_{32,0} \tilde{\varepsilon}_2 + B_{33,0} \tilde{\varepsilon}_3 \right) \end{aligned}$$

$$= - B_{33}^{-1} \left[B_{31} B_{11}^{-1} - B_{32} B_{22}^{-1} B_{21,0} B_{11,0}^{-1} \quad B_{32} B_{22}^{-1} \right] \begin{bmatrix} B_{11,0} & 0 \\ B_{21,0} & B_{22,0} \end{bmatrix} \begin{bmatrix} \tilde{\varepsilon}_1 \\ \tilde{\varepsilon}_2 \end{bmatrix}$$

$$+ B_{33}^{-1} \left(B_{31,0} \tilde{\varepsilon}_1 + B_{32,0} \tilde{\varepsilon}_2 + B_{33,0} \tilde{\varepsilon}_3 \right).$$

Hence, at $B_{22} = B_{22,0}$ the condition $E\left[f_{\mathbf{2}}(B, u_t) \right] = 0$ implies $0 = E[\tilde{e}_3(B) \tilde{e}_2(B)']$ and therefore,

$$0 = B_{33}^{-1} (-B_{32} B_{22}^{-1} B_{22,0} + B_{32,0})$$

which implies $B_{32} = B_{32,0}$. □

5.A.4 Redundant and relevant moment conditions in the recursive SVAR

The proof of Proposition 5.3 requires to verify the redundancy conditions from Breusch et al. (1999). However, verifying these conditions is a lengthy task. We derive analytical expressions for the conditions in Online Appendix C and summarize them in Lemma 5.C.14 in Online Appendix C. The following proof of Proposition 5.3 uses Lemma 5.C.8 and 5.C.14 in Online Appendix C.

Proof of Proposition 5.3.
In the recursive SVAR, the identifying moment conditions $E[f_{\mathbf{N}}(B, u_t)]$ only contain second-order moment conditions and therefore, are referred to as $E[f_{\mathbf{2}}(B, u_t)]$ in this proof.

Breusch et al. (1999) show that overidentifying moment conditions $E[f_{\mathbf{D}}(B, u_t)]$ are redundant w.r.t. the identifying moment conditions $E[f_{\mathbf{2}}(B, u_t)]$ if and only if

$$G_{\mathbf{D}} = S_{\mathbf{D2}} S_{\mathbf{2}}^{-1} G_{\mathbf{2}},$$

where

$$G_{\mathbf{D}} := E\left[\frac{\partial f_{\mathbf{D}}(B_0, u_t)}{\partial vec(B)'} \right], \qquad G_{\mathbf{2}} := E\left[\frac{\partial f_{\mathbf{2}}(B_0, u_t)}{\partial vec(B)'} \right],$$

$$S_{\mathbf{2}} := \lim_{T \to \infty} E\left[g_{\mathbf{2}}(B_0) g_{\mathbf{2}}(B_0)' \right], \qquad S_{\mathbf{D2}} := \lim_{T \to \infty} E\left[g_{\mathbf{D}}(B_0) g_{\mathbf{2}}(B_0)' \right].$$

Moreover, Breusch et al. (1999) show that overidentifying moment conditions $E[f_{\mathbf{D}}(B, u_t)]$ are partially redundant w.r.t. $E[f_{\mathbf{2}}(B, u_t)]$ for a subset of coefficients $b \subset vec(B)$ w.r.t. the moment conditions $E[f_{\mathbf{2}}(B, u_t)]$ if and only if

$$G_{\mathbf{D}}^{b} - S_{\mathbf{D2}} S_{\mathbf{2}}^{-1} G_{\mathbf{2}}^{b} = \left(G_{\mathbf{D}}^{\neg b} - S_{\mathbf{D2}} S_{\mathbf{2}}^{-1} G_{\mathbf{2}}^{\neg b} \right) \left(\left(G_{\mathbf{2}}^{\neg b} \right)' S_{\mathbf{2}}^{-1} G_{\mathbf{2}}^{\neg b} \right) \left(\left(G_{\mathbf{2}}^{\neg b} \right)' S_{\mathbf{2}}^{-1} G_{\mathbf{2}}^{b} \right) \quad (5.A.4)$$

where

$$G_{\mathbf{2}}^{b} := E\left[\frac{\partial f_{\mathbf{2}}(u_t, B_0)}{\partial b'} \right], \qquad G_{\mathbf{D}}^{b} := E\left[\frac{\partial f_{\mathbf{D}}(u_t, B_0)}{\partial b'} \right],$$

$$G_{\mathbf{2}}^{\neg b} := E\left[\frac{\partial f_{\mathbf{2}}(u_t, B_0)}{\partial(\neg b)'}\right], \qquad G_{\mathbf{D}}^{\neg b} := E\left[\frac{\partial f_{\mathbf{D}}(u_t, B_0)}{\partial(\neg b)'}\right],$$

and where $\neg b$ denotes all unrestricted elements of B not contained in b. With Lemma 5.C.8 it holds that $G_{\mathbf{2}}^{b_i\prime} S_{\mathbf{2}}^{-1} G_{\mathbf{2}}^{b_j} = 0$ for $i, j \in \{1, \dots, n\}$ with $i \neq j$. Therefore, for any vector $b_i = [b_{ii}, ..., b_{ni}]$ representing the impact of the ith structural shock $\epsilon_{i,t}$ it holds that $G_{\mathbf{2}}^{b_i\prime} S_{\mathbf{2}}^{-1} G_{\mathbf{2}}^{\neg b_i}$ is zero. Therefore, for any vector $b_i = [b_{ii}, ..., b_{ni}]$ the right hand side of Equation (5.A.4) is zero and hence the partial redundancy condition simplifies to

$$G_{\mathbf{D}}^{b_i} - S_{\mathbf{D2}} S_{\mathbf{2}}^{-1} G_{\mathbf{2}}^{b_i} = 0.$$

The statements then follow from Lemma 5.C.14. □

5.A.5 Asymptotic variance of the block-recursive SVAR pGMM estimator

We show how to derive the asymptotic variance of the pGMM estimator, $V_{\mathbf{N+A}}$, based on Remark 3.5 of Cheng and Liao (2015). We first show Lemma 5.A.1 and then apply the result in Remark 3.5 of Cheng and Liao (2015). Recall that $E[f_{\mathbf{I}}(B, u_t)]$ and $E[f_{\mathbf{R}}(B, u_t)]$ denote the sets of invalid and redundant moment conditions, respectively. Denote $E[f_{\mathbf{U}}(B, u_t)]$ as moment conditions either in $E[f_{\mathbf{I}}(B, u_t)]$ or $E[f_{\mathbf{R}}(B, u_t)]$ and the number of moment conditions $E[f_{\mathbf{U}}(B, u_t)]$ by $k_{\mathbf{U}}$. Similarly, we denote $k_{\mathbf{A}}$ as the number of moment conditions in $E[f_{\mathbf{A}}(B, u_t)]$. Further, define the number of unrestricted elements in $vec(B)$ as d_B. In the proof of Lemma 5.A.1, we use the indices $1 \equiv \mathbf{N+A}$, $2 \equiv (\mathbf{N+A}, \mathbf{U})$, $3 \equiv (\mathbf{U}, \mathbf{N+A})$, and $4 \equiv \mathbf{U}$ to keep notation uncluttered. Let $\iota^* = (\iota', \mathbf{0}'_{k_{\mathbf{U}}})'$ where $\iota = (1, \dots, 1)'$ is a $d_B \times 1$ vector, i.e., $\iota^{*\prime} A \iota^*$ gives the leading $d_B \times d_B$-upper west block of an arbitrary $(d_B + k_{\mathbf{U}}) \times (d_B + k_{\mathbf{U}})$ matrix A.

Lemma 5.A.1.

$$\iota^{*\prime} \left(\Gamma' W \Gamma\right)^{-1} \left(\Gamma' W S_{\mathbf{N+D}} W \Gamma\right) \left(\Gamma' W \Gamma\right)^{-1} \iota^* = V_{\mathbf{N+A}},$$

where

$$\Gamma := \begin{bmatrix} G_{\mathbf{N+A}} & \mathbf{0}_{(k_{\mathbf{N}}+k_{\mathbf{A}})\times k_{\mathbf{U}}} \\ G_{\mathbf{U}} & -I_{k_{\mathbf{U}}} \end{bmatrix}, \qquad V_{\mathbf{N+A}} := M_{\mathbf{N+A}} S_{\mathbf{N+A}} M'_{\mathbf{N+A}}$$

$$M_{\mathbf{N+A}} := \left(G'_{\mathbf{N+A}} W^{pi}_{\mathbf{N+A}} G_{\mathbf{N+A}}\right)^{-1} G'_{\mathbf{N+A}} W^{pi}_{\mathbf{N+A}}, \qquad S_{\mathbf{N+A}} := \lim_{T\to\infty} E\left[g_{\mathbf{N+A}}(B_0) g_{\mathbf{N+A}}(B_0)'\right],$$

$$G_{\mathbf{N+A}} := \begin{bmatrix} G_{\mathbf{N}} \\ G_{\mathbf{A}} \end{bmatrix}, \qquad W^{pi}_{\mathbf{N+A}} := \left(W_{\mathbf{N+A}} - W_{\mathbf{N+A},\mathbf{I\cup R}} W^{-1}_{\mathbf{I\cup R}} W_{\mathbf{I\cup R},\mathbf{N+A}}\right),$$

$$G_{\mathbf{A}} := E\left[\frac{\partial f_{\mathbf{A}}(B_0, u_t)}{\partial vec(B)'}\right], \qquad W_{\mathbf{N+D}} := \begin{bmatrix} W_{\mathbf{N+A}} & W_{\mathbf{N+A},\mathbf{I\cup R},} \\ W_{\mathbf{I\cup R},\mathbf{N+A}} & W_{\mathbf{I\cup R}} \end{bmatrix},$$

$$W_{\mathbf{N+A}} \in \mathbb{R}^{(k_{\mathbf{N}}+k_{\mathbf{A}})\times(k_{\mathbf{N}}+k_{\mathbf{A}})}, \qquad W_{\mathbf{N+A},\mathbf{I\cup R}} \in \mathbb{R}^{(k_{\mathbf{N}}+k_{\mathbf{A}})\times(k_{\mathbf{D}}-k_{\mathbf{A}})},$$

$$W_{\mathbf{I\cup R},\mathbf{N+A}} = W'_{\mathbf{N+A},\mathbf{I\cup R}}, \qquad W_{\mathbf{I\cup R}} \in \mathbb{R}^{(k_{\mathbf{D}}-k_{\mathbf{A}})\times(k_{\mathbf{D}}-k_{\mathbf{A}})}.$$

Proof. Recall that $G_{\mathbf{N}+\mathbf{A}}$ and $G_{\mathbf{U}}$ have dimension $(k_{\mathbf{N}}+k_{\mathbf{A}})\times d_B$ and $k_{\mathbf{U}}\times d_B$, respectively. We define

$$L := \begin{bmatrix} L_1 & L_2 \\ L_3 & L_4 \end{bmatrix} := \left(\Gamma' W \Gamma\right)^{-1}.$$

Additionally, let

$$N := \begin{bmatrix} N_1 & N_2 \\ N_3 & N_4 \end{bmatrix} := \left(\Gamma' W S_{\mathbf{N}+\mathbf{D}} W \Gamma\right),$$

and denote the inverse of W by

$$W^{ipi} := \begin{bmatrix} W_1^{ipi} & W_2^{ipi} \\ W_3^{ipi} & W_4^{ipi} \end{bmatrix} := W^{-1} = \begin{bmatrix} W_1 & W_2 \\ W_3 & W_4 \end{bmatrix}^{-1}.$$

Let $W_1^{pi} := \left(W_1 - W_2 W_4^{-1} W_3\right)$. Then, by the partitioned inverse, $W_1^{ipi} := \left(W_1^{pi}\right)^{-1}$. By similar arguments as leading to (2.18) in the Online Appendix of Cheng and Liao (2015), we get that

$$L_1 = \left(G_1' \left(W_1 - W_2 W_4^{-1} W_3\right) G_1\right)^{-1} = \left(G_1' W_1^{pi} G_1\right)^{-1}$$

and, by using the partitioned inverse formula again, and similar arguments as leading to (2.10), (2.11) and (2.18) in the Online Appendix of Cheng and Liao (2015), that

$$\begin{aligned} L_3 &= -W_4^{-1}\left(-G_1' W_2 - G_4' W_4\right)' \left(G_1' W_1^{pi} G_1\right)^{-1} \\ &= \left(W_4^{-1} W_3 G_1 + G_4\right) L_1 \\ &= X L_1, \end{aligned} \tag{5.A.5}$$

where we used that $W_4' = W_4$, $W_3 = W_2'$ and $X := \left(W_4^{-1} W_3 G_1 + G_4\right)$. Further, let

$$H := \begin{bmatrix} H_1 & H_2 \\ H_3 & H_4 \end{bmatrix} := W S_{\mathbf{N}+\mathbf{D}} W,$$

where

$$\begin{aligned} H_1 &:= W_1 S_1 W_1 + W_2 S_3 W_1 + W_1 S_2 W_3 + W_2 S_4 W_3 \\ H_2 &:= W_1 S_1 W_2 + W_2 S_3 W_2 + W_1 S_2 W_4 + W_2 S_4 W_4 \\ H_3 &:= W_3 S_1 W_1 + W_4 S_3 W_1 + W_3 S_2 W_3 + W_4 S_4 W_3 \\ H_4 &:= W_3 S_1 W_2 + W_4 S_3 W_2 + W_3 S_2 W_4 + W_4 S_4 W_4. \end{aligned}$$

Note that $H_3 = H_2'$ since $W_3 = W_2'$, $W_1 = W_1'$, $W_4 = W_4'$, $S_3 = S_2'$, $S_1 = S_1'$ and $S_4 = S_4'$. Hence, similar to (2.11) in the Online Appendix of Cheng and Liao (2015),

$$\begin{aligned}
N_1 &= G_1' H_1 G_1 + G_4' H_3 G_1 + G_1' H_2 G_4 + G_4' H_4 G_4 \\
&= G_1' H_1 G_1 + G_4' H_2' G_1 + G_1' H_2 G_4 + G_4' H_4 G_4 \\
N_2 &= -G_1' H_2 - G_4' H_4 \\
N_3 &= N_2' \\
N_4 &= H_4.
\end{aligned}$$

Then,

$$\begin{aligned}
\iota^{*\prime} \left(\Gamma' W \Gamma\right)^{-1} \left(\Gamma' W S_{\mathbf{N}+\mathbf{D}} W \Gamma\right) \left(\Gamma' W \Gamma\right)^{-1} \iota^* &= \iota^{*\prime} L N L \iota^* \\
&= L_1 N_1 L_1 + L_2 N_3 L_1 + L_1 N_2 L_3 + L_2 N_4 L_3 \\
&= L_1 N_1 L_1 + L_3' N_3 L_1 + L_1 N_2 L_3 + L_3' N_4 L_3 \\
&\overset{(5.A.5)}{=} L_1 N_1 L_1 + L_1' X' N_2' L_1 + L_1 N_2 X L_1 \\
&\quad + L_1' X' N_4 X L_1 \\
&= L_1 \left(N_1 + X' N_2' + N_2 X + X' N_4 X\right) L_1, \qquad (5.A.6)
\end{aligned}$$

where we used that $L_1' = L_1$, $L_3' = L_2$, and $N_3' = N_2$.

Next, define $Y := N_1 + X' N_2' + N_2 X + X' N_4 X$. Then, multiplying out gives

$$\begin{aligned}
Y =& G_1' H_1 G_1 + G_4' H_3 G_1 + G_1' H_2 G_4 + G_4' H_4 G_4 + \left(G_1' W_2 W_4^{-1} + G_4'\right) \left(-H_2' G_1 - H_4' G_4\right) \\
&+ \left(-G_1' H_2 - G_4' H_4\right) \left(W_4^{-1} W_2' G_1 + G_4\right) + \left(G_1' W_2 W_4^{-1} + G_4'\right) H_4 \left(W_4^{-1} W_2' G_1 + G_4\right) \\
=& G_1' W_2 W_4^{-1} H_4 W_4^{-1} W_2' G_1 + G_1' H_1 G_1 - G_1' W_2 W_4^{-1} H_2' G_1 - G_1' H_2 W_4^{-1} W_2' G_1 \\
=& G_1' \left(W_2 W_4^{-1} H_4 W_4^{-1} W_2' + H_1 - W_2 W_4^{-1} H_2' - H_2 W_4^{-1} W_2'\right) G_1 \\
=& G_1' \left(W_2 W_4^{-1} W_3 S_1 W_2 W_4^{-1} W_3 + W_1 S_1 W_1 - W_2 W_4^{-1} W_3 S_1 W_1 - W_1 S_1 W_2 W_4^{-1} W_3\right) G_1 \\
=& G_1' \left(W_1 - W_2 W_4^{-1} W_3\right) S_1 \left(W_1 - W_2 W_4^{-1} W_3\right) G_1
\end{aligned}$$

$$=G_1' W_1^{pi} S_1 W_1^{pi} G_1 \tag{5.A.7}$$

Plugging (5.A.7) into (5.A.6), we obtain
$\iota^{*\prime} \left(\Gamma' W \Gamma\right)^{-1} \left(\Gamma' W S_{\mathbf{N+D}} W \Gamma\right) \left(\Gamma' W \Gamma\right)^{-1} \iota^*$

$$= L_1 \left(G_1' W_1^{pi} S_1 W_1^{pi}\right) G_1 L_1$$

$$= \left(G_1' W_1^{pi} G_1\right)^{-1} \left(G_1' W_1^{pi} S_1 W_1^{pi} G_1\right) \left(G_1' W_1^{pi} G_1\right)^{-1}$$

$$= \left(G_{\mathbf{N+A}}' W_{\mathbf{N+A}}^{pi} G_{\mathbf{N+A}}\right)^{-1} \left(G_{\mathbf{N+A}}' W_{\mathbf{N+A}}^{pi} S_{\mathbf{N+A}} W_{\mathbf{N+A}}^{pi} G_{\mathbf{N+A}}\right) \left(G_{\mathbf{N+A}}' W_{\mathbf{N+A}}^{pi} G_{\mathbf{N+A}}\right)^{-1}$$

which was to show. □

Note that in the following proposition, we treat the number of valid and relevant moment conditions, $k_{\mathbf{A}}$, and the number of invalid moment conditions, $k_{\mathbf{I}}$, as fixed constants to keep our asymptotic results for the pGMM estimator in line with the asymptotic results for the block-recursive SVAR GMM estimator in Equation (5.10). Cheng and Liao (2015) allow both $k_{\mathbf{A}}$ and $k_{\mathbf{I}}$ to increase with the sample size. However, their results also hold when the number of moment conditions is fixed.

Proposition 5.A.1. *Assume that the Assumptions in Theorem 3.3 of Cheng and Liao (2015) hold. Further, assume that* $E\left[\frac{\partial f_{\mathbf{A}}(B_0,u_t)}{\partial vec(B)'}\right] = \frac{\partial E[f_{\mathbf{A}}(B_0,u_t)]}{\partial vec(B)'}$ *and Assumption 5.1 and 5.2 hold. Then,*

$$\sqrt{T}\left(vec(\hat{B}_{\mathbf{N+D}}) - vec(B_0)\right) \xrightarrow{d} \mathcal{N}(0, V_{\mathbf{N+A}})$$

Proof. Define $\Sigma_{CL} := \left(\Gamma' W \Gamma\right)^{-1} \left(\Gamma' W S_{\mathbf{N+D}} W \Gamma\right) \left(\Gamma' W \Gamma\right)^{-1}$ and $\gamma = \left(\nu', \mathbf{0}_{k_{\mathbf{U}}}'\right)'$ where $\nu \in \mathbb{R}^{d_B}$ is an arbitrary vector. Then, by Remark 3.5 of Cheng and Liao (2015),

$$\left|\left|\Sigma_{CL}^{1/2}\,\gamma\right|\right|^{-1} \sqrt{T}\nu'\left(vec(\hat{B}_{\mathbf{N+D}}) - vec(B_0)\right) \xrightarrow{d} \mathcal{N}(0,1),$$

where $||a|| := \sqrt{a'a}$ is the ℓ_2-norm of an arbitrary vector a.

Note that Lemma 5.A.1 immediately implies $\left|\left|\Sigma_{CL}^{1/2}\gamma\right|\right| = \sqrt{\gamma'\Sigma_{CL}\gamma} = \sqrt{\nu' V_{\mathbf{N+A}}(W)\nu}$. Hence,

$$\left|\left|V_{\mathbf{N+A}}(W)^{1/2}\nu\right|\right|^{-1} \sqrt{T}\nu'\left(vec(\hat{B}_{\mathbf{N+D}}) - vec(B_0)\right) \xrightarrow{d} \mathcal{N}(0,1),$$

where $V_{\mathbf{N+A}}(W)$ is the asymptotic variance of $vec(\hat{B}_{\mathbf{N+D}})$ since it holds that

$$\nu^{*\prime} V_{\mathbf{N+A}}(W)\nu^* = \left|\left|V_{\mathbf{N+A}}(W)^{1/2}\nu\right|\right|^{-2} \nu' V_{\mathbf{N+A}}(W)\nu = 1$$

where $\nu^* := \left|\left|V_{\mathbf{N+A}}(W)^{1/2}\nu\right|\right|^{-1} \nu$.

Consequently, using the Cramér-Wold device, we get

$$\sqrt{T}\left(vec(\hat{B}_{\mathbf{N+D}}) - vec(B_0)\right) \xrightarrow{d} \mathcal{N}(0, V_{\mathbf{N+A}}).$$

□

5.A.6 Choice of maximum λ in the cross-validation

In the following, we illustrate how to choose the maximum value of λ in the cross-validation. Define the loss function of the pGMM estimator as

$$L^*(B, \beta) := L(B, \beta) + \lambda \sum_{i \in \widetilde{D}} \omega_i \, |\beta_i|, \tag{5.A.8}$$

where $L(B, \beta) := \begin{bmatrix} g_{\mathbf{N}}(B) \\ g_{\mathbf{D}}(B, \beta) \end{bmatrix}' W \begin{bmatrix} g_{\mathbf{N}}(B) \\ g_{\mathbf{D}}(B, \beta) \end{bmatrix}$.

Further, let $z \in \partial||\beta||_1$, where $z \in \mathbb{R}^{k_{\mathbf{D}}}$, denote the subgradient for the ℓ_1-norm evaluated at β, i.e.,

$$\begin{aligned} z_i &= sign(\beta_i), \text{ if } \beta_i \neq 0, \\ z_i &\in [-1, 1], \quad \text{if } \beta_i = 0, \end{aligned} \tag{5.A.9}$$

for $i = 1, \ldots, k_{\mathbf{D}}$ (Wainwright, 2006). Then, the first order condition of the pGMM estimator with respect to β_i, $i = 1, \ldots, k_{\mathbf{D}}$, evaluated at β and B is

$$\frac{\partial L^*(B, \beta)}{\partial \beta_i} = \frac{\partial L(B, \beta)}{\partial \beta_i} + \lambda \omega_i \, z_i = 0 \tag{5.A.10}$$

Note that $\omega_i \geq 0$. However, if $\omega_i = 0$, β_i is not penalized and therefore, we only consider $i \in \widetilde{P} := \{j \in \tilde{D} | \ \omega_j > 0\}$ for which, by definition, $\omega_i > 0$ when choosing the maximum value of λ in the cross-validation. By (5.A.9) and (5.A.10), $\beta = \mathbf{0} = (0, \ldots, 0)'$ and $B = B_0$ minimize the loss function in (5.A.8) only if

$$\frac{1}{\omega_i} \frac{\partial L(B_0, \mathbf{0})}{\partial \beta_i} \in \lambda[-1, 1],$$

for $i \in \widetilde{P}$. Thus,

$$\max_{i \in \widetilde{P}} \left| \frac{1}{\omega_i} \frac{\partial L(B_0, \mathbf{0})}{\partial \beta_i} \right| \leq \lambda.$$

This motivates us to use

$$\lambda_{\max} = \max_{i \in \widetilde{P}} \left| \frac{1}{\omega_i} \frac{\partial L(B_0, \mathbf{0})}{\partial \beta_i} \right|.$$

as the largest value in the cross-validation. Note that any $\lambda > \lambda_{\max}$ would not have an effect on β as $\lambda_{\max}$ already shrinks all elements of β to zero. In practice, we replace B_0

and ω_i by consistent estimators to obtain $\lambda_{\max}$. Furthermore, we consider a weight ω_j to be positive and hence, $j \in \widetilde{P}$, if $\omega_j / \sum_{k \in \widetilde{D}} \omega_k > 10^{-4}$.

Appendix 5.B Supplementary Figures and Tables

5.B.1 Finite sample performance

Table 5.B.1: Finite sample performance including Post-LASSO.

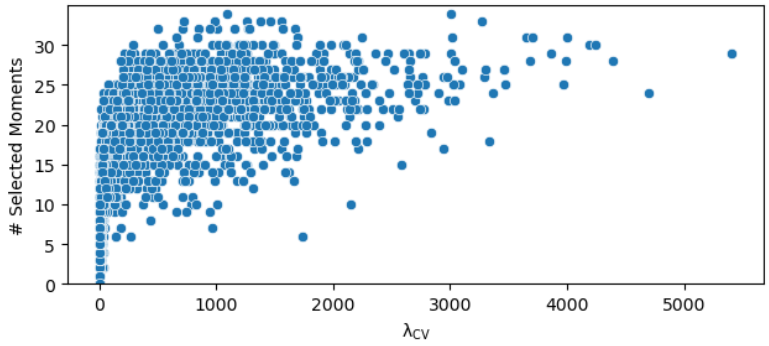
$T = 100$	GMM	oGMM	GMM-Oracle	pGMM	Post-pGMM
$\hat{B}$	$\begin{bmatrix} 9.93\ (1.09) & . & . & . \\ 4.98\ (1.21) & 9.86\ (1.02) & . & . \\ 4.97\ (1.49) & 4.95\ (1.29) & 9.83\ (1.12) & . \\ 4.96\ (1.71) & 4.93\ (1.46) & 4.91\ (1.27) & 9.78\ (1.08) \end{bmatrix}$	$\begin{bmatrix} 9.77\ (1.07) & . & . & . \\ 4.90\ (1.31) & 9.71\ (1.01) & . & . \\ 4.89\ (1.69) & 4.88\ (1.43) & 9.70\ (1.10) & . \\ 4.90\ (2.07) & 4.88\ (1.74) & 4.88\ (1.46) & 9.69\ (1.09) \end{bmatrix}$	$\begin{bmatrix} 9.76\ (1.07) & . & . & . \\ 4.91\ (1.17) & 9.70\ (1.02) & . & . \\ 4.91\ (1.50) & 4.88\ (1.26) & 9.69\ (1.10) & . \\ 4.92\ (1.81) & 4.88\ (1.51) & 4.88\ (1.25) & 9.67\ (1.10) \end{bmatrix}$	$\begin{bmatrix} 9.96\ (1.09) & . & . & . \\ 5.00\ (1.15) & 9.88\ (1.01) & . & . \\ 4.98\ (1.46) & 4.96\ (1.22) & 9.85\ (1.11) & . \\ 4.99\ (1.71) & 4.96\ (1.42) & 4.95\ (1.21) & 9.82\ (1.10) \end{bmatrix}$	$\begin{bmatrix} 9.84\ (1.06) & . & . & . \\ 4.96\ (1.09) & 9.79\ (1.01) & . & . \\ 4.94\ (1.39) & 4.93\ (1.19) & 9.77\ (1.11) & . \\ 4.94\ (1.61) & 4.93\ (1.38) & 4.91\ (1.18) & 9.73\ (1.11) \end{bmatrix}$
#Mo	10.00	57.00	40.00	24.22	24.22
Bias	−0.0883	−0.1806	−0.1804	−0.0650	−0.1256
MSE	1.27	1.40	1.28	1.25	1.21
λ	.	.	.	71.08	.

$T = 250$	GMM	oGMM	GMM-Oracle	pGMM	Post-pGMM
$\hat{B}$	$\begin{bmatrix} 9.97\ (0.43) & . & . & . \\ 4.99\ (0.51) & 9.96\ (0.43) & . & . \\ 4.98\ (0.64) & 5.00\ (0.52) & 9.93\ (0.45) & . \\ 4.98\ (0.72) & 4.99\ (0.61) & 4.98\ (0.51) & 9.91\ (0.45) \end{bmatrix}$	$\begin{bmatrix} 9.90\ (0.40) & . & . & . \\ 4.96\ (0.49) & 9.90\ (0.40) & . & . \\ 4.96\ (0.65) & 4.97\ (0.51) & 9.87\ (0.42) & . \\ 4.97\ (0.73) & 4.96\ (0.61) & 4.97\ (0.49) & 9.86\ (0.42) \end{bmatrix}$	$\begin{bmatrix} 9.90\ (0.40) & . & . & . \\ 4.97\ (0.44) & 9.90\ (0.40) & . & . \\ 4.97\ (0.59) & 4.97\ (0.46) & 9.87\ (0.42) & . \\ 4.98\ (0.65) & 4.97\ (0.54) & 4.96\ (0.44) & 9.85\ (0.42) \end{bmatrix}$	$\begin{bmatrix} 9.99\ (0.42) & . & . & . \\ 5.01\ (0.45) & 9.97\ (0.41) & . & . \\ 5.01\ (0.59) & 5.02\ (0.46) & 9.94\ (0.42) & . \\ 5.02\ (0.66) & 5.01\ (0.55) & 5.00\ (0.44) & 9.92\ (0.43) \end{bmatrix}$	$\begin{bmatrix} 9.93\ (0.41) & . & . & . \\ 4.98\ (0.44) & 9.92\ (0.41) & . & . \\ 4.98\ (0.57) & 4.99\ (0.45) & 9.89\ (0.43) & . \\ 4.99\ (0.64) & 4.98\ (0.54) & 4.97\ (0.44) & 9.87\ (0.44) \end{bmatrix}$
#Mo	10.00	57.00	40.00	27.20	27.20
Bias	−0.0311	−0.0676	−0.0656	−0.0114	−0.0480
MSE	0.53	0.51	0.48	0.48	0.48
λ	.	.	.	118.92	.

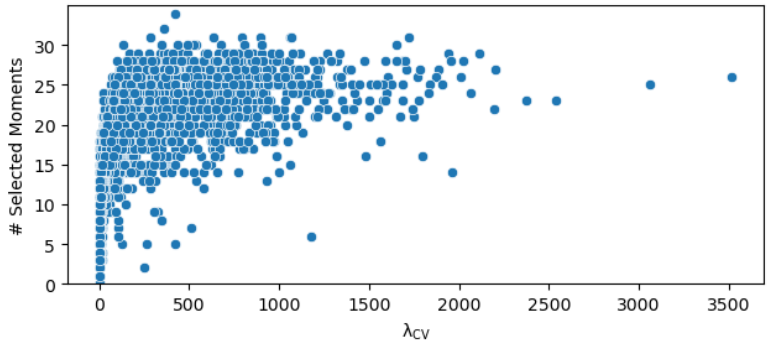
$T = 1000$	GMM	oGMM	GMM-Oracle	pGMM	Post-pGMM
$\hat{B}$	$\begin{bmatrix} 10.00\ (0.11) & . & . & . \\ 5.00\ (0.13) & 9.99\ (0.11) & . & . \\ 4.99\ (0.15) & 4.99\ (0.13) & 9.99\ (0.11) & . \\ 4.99\ (0.19) & 4.99\ (0.15) & 4.99\ (0.13) & 9.98\ (0.11) \end{bmatrix}$	$\begin{bmatrix} 9.98\ (0.10) & . & . & . \\ 4.99\ (0.12) & 9.97\ (0.10) & . & . \\ 4.99\ (0.13) & 4.99\ (0.11) & 9.98\ (0.10) & . \\ 4.99\ (0.16) & 4.99\ (0.14) & 4.99\ (0.11) & 9.97\ (0.10) \end{bmatrix}$	$\begin{bmatrix} 9.98\ (0.10) & . & . & . \\ 4.99\ (0.11) & 9.97\ (0.10) & . & . \\ 4.99\ (0.13) & 4.99\ (0.10) & 9.98\ (0.10) & . \\ 4.99\ (0.15) & 4.99\ (0.13) & 4.99\ (0.11) & 9.97\ (0.10) \end{bmatrix}$	$\begin{bmatrix} 10.00\ (0.11) & . & . & . \\ 5.00\ (0.11) & 9.99\ (0.10) & . & . \\ 5.00\ (0.13) & 5.00\ (0.11) & 10.00\ (0.10) & . \\ 5.00\ (0.16) & 5.00\ (0.13) & 5.00\ (0.11) & 9.98\ (0.10) \end{bmatrix}$	$\begin{bmatrix} 9.99\ (0.11) & . & . & . \\ 5.00\ (0.11) & 9.98\ (0.11) & . & . \\ 4.99\ (0.13) & 4.99\ (0.11) & 9.98\ (0.10) & . \\ 5.00\ (0.16) & 4.99\ (0.13) & 4.99\ (0.11) & 9.97\ (0.11) \end{bmatrix}$
#Mo	10.00	57.00	40.00	29.59	29.59
Bias	−0.0076	−0.0158	−0.0158	−0.0021	−0.0122
MSE	0.13	0.12	0.11	0.12	0.12
λ	.	.	.	75.34	.

Note: The table reports the average $\bar{b}_{ij}$ and the corresponding estimated MSE (in parentheses) of each estimated element in $\hat{B}$ as well as the BIAS and MSE across estimated elements in $\hat{B}$ over $3,500$ Monte Carlo replicates for the GMM estimator, the oGMM estimator, the GMM-Oracle estimator, the pGMM estimator, and the Post-pGMM estimator. The Post-pGMM estimator uses only the overidentifying moment conditions selected by the pGMM estimator for the estimation of the block-recursive SVAR. All estimator use zero restrictions which are highlighted by the dots.

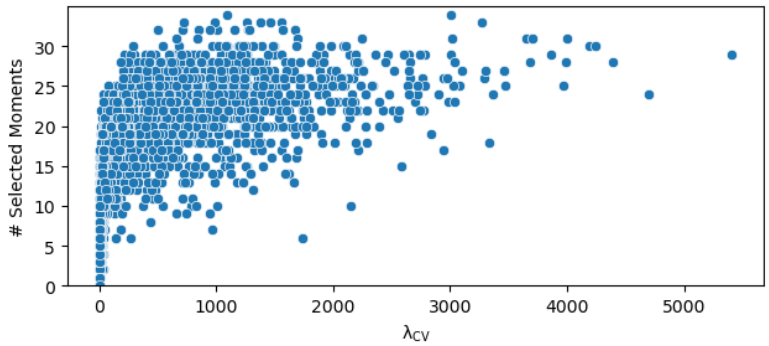
Figure 5.B.1: Relationship of chosen λ_{CV} and Number of Selected Moments across Monte Carlo runs.



(a) $T = 100$



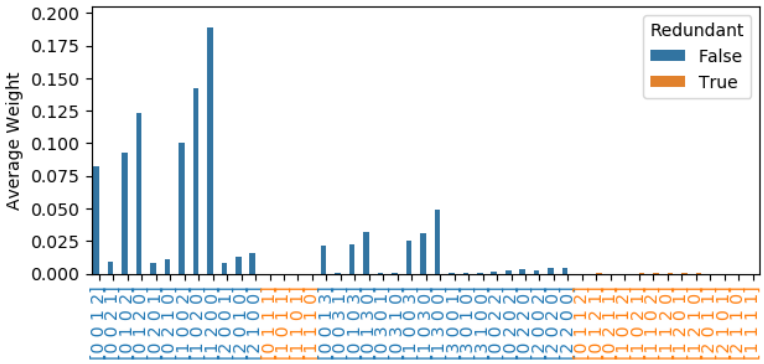
(b) $T = 250$



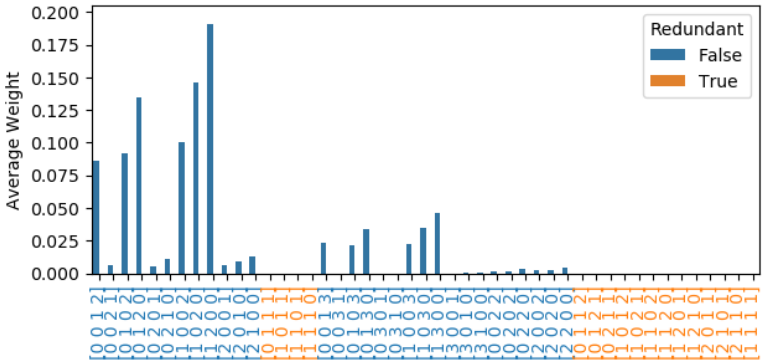
(c) $T = 1000$

Note: The figure shows the chosen λ_{CV} in the cross-validation and the corresponding number of selected moments for each of the $M = 3,500$ Monte Carlo simulations.

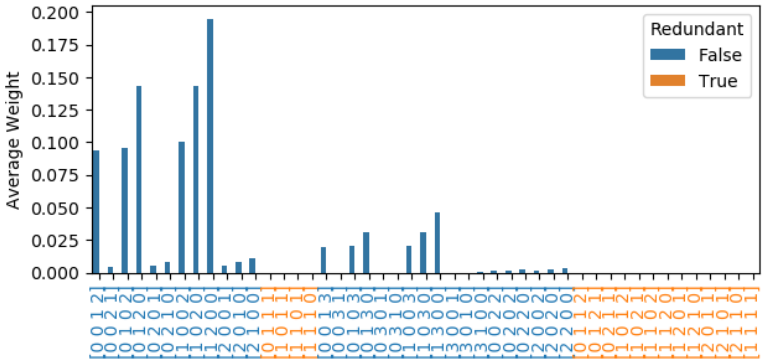
Figure 5.B.2: Average Weight of Moments across Monte Carlo runs.



(a) $T = 100$



(b) $T = 250$



(c) $T = 1000$

Note: The figure shows the average weight of each moment across $M = 3,500$ Monte Carlo simulations. Redundant moments (orange) and relevant moments (blue) are displayed on the x-axis. Each x-axis label abbreviates a moment condition, e.g., $[0, 1, 2, 1]$ corresponds to $E[e(B)^0_{1,t}\, e(B)^1_{2,t}\, e(B)^2_{3,t}\, e(B)^1_{4,t}]$.

5.B.2 Empirical illustration

This section contains supplementary material and robustness checks for the application presented in Section 5.6.

Table 5.B.2 shows descriptive statistics of the variables used in the SVAR. Table 5.B.3

Table 5.B.2: Descriptive statistics.

	Mean	Median	Std. deviation	Variance	Skewness	Kurtosis
O_t	0.078	0.19	1.5	2.26	-1.66	10.8
Y_t	0.20	0.29	0.60	0.37	-1.2	5.21
OP_t	0.32	0.03	7.31	53.4	0.06	4.46
SR_t	0.34	0.62	3.61	13.03	-0.82	3.67

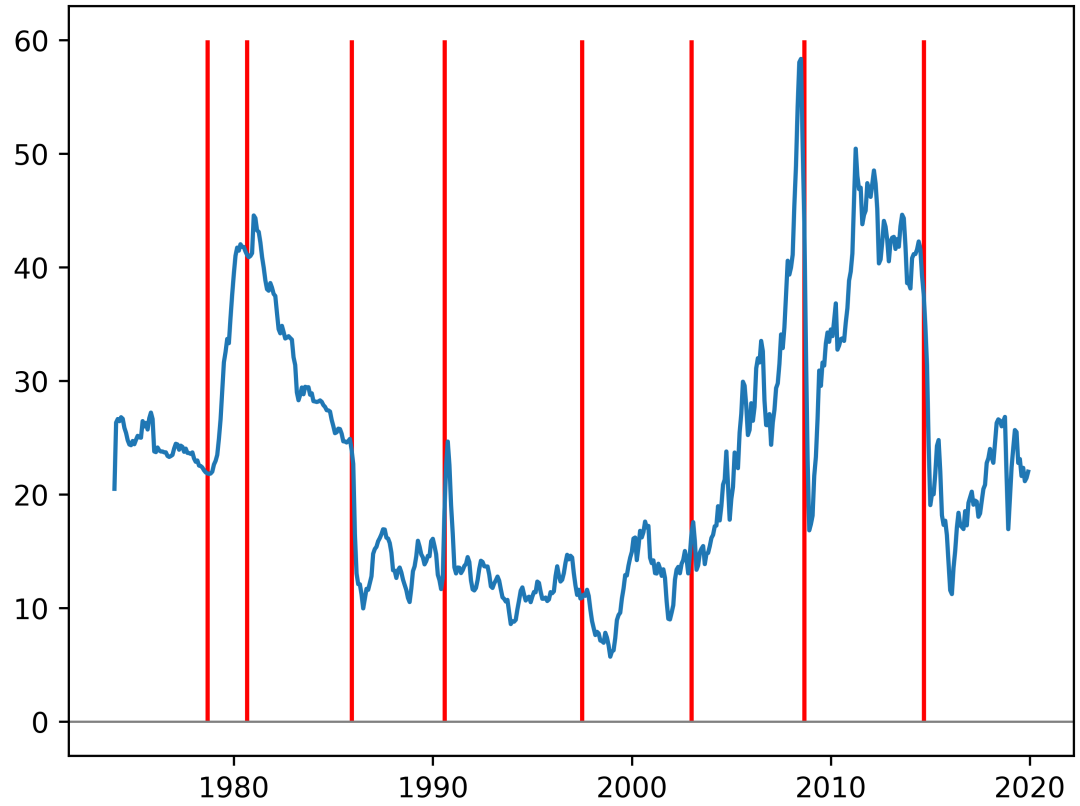
shows the correlation between the estimated structural shocks from the block-recursive SVAR pGMM estimator and the reduced form shocks. Figure 5.B.3 shows the historical

Table 5.B.3: Correlation of reduced form and estimated structural shocks.

	u^O	u^Y	u^{OP}	u^{SR}
ε^s	1	-0.03	-0.13	-0.05
ε^d	0.06	1	0.12	0.06
ε^{s-exp}	-0.08	0.02	0.94	-0.27
ε^{d-exp}	-0.05	0.02	0.33	0.96

evolution of the real oil price.

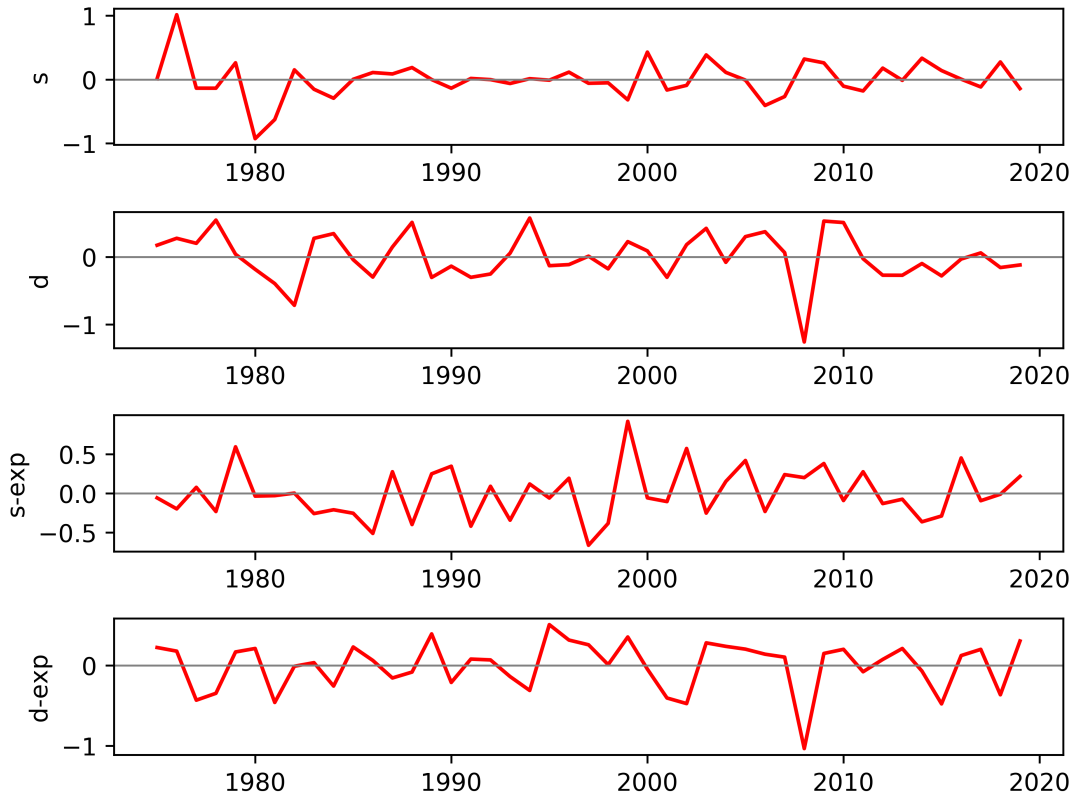
Figure 5.B.3: Real oil price.



Note: The vertical bars indicate the following events: Iranian Revolution 1978 : 9, Iran Iraq War 1980 : 9, collapse of OPEC 1985 : 12, Persian Gulf War 1990 : 8, Asian Financial Crisis of 1997 :7, Iraq War 2003 : 1, the collapse of Lehman Brothers (2008 : 9), and the oil price decline in mid 2014.

Figure 5.B.4 shows the estimated structural shocks across years.

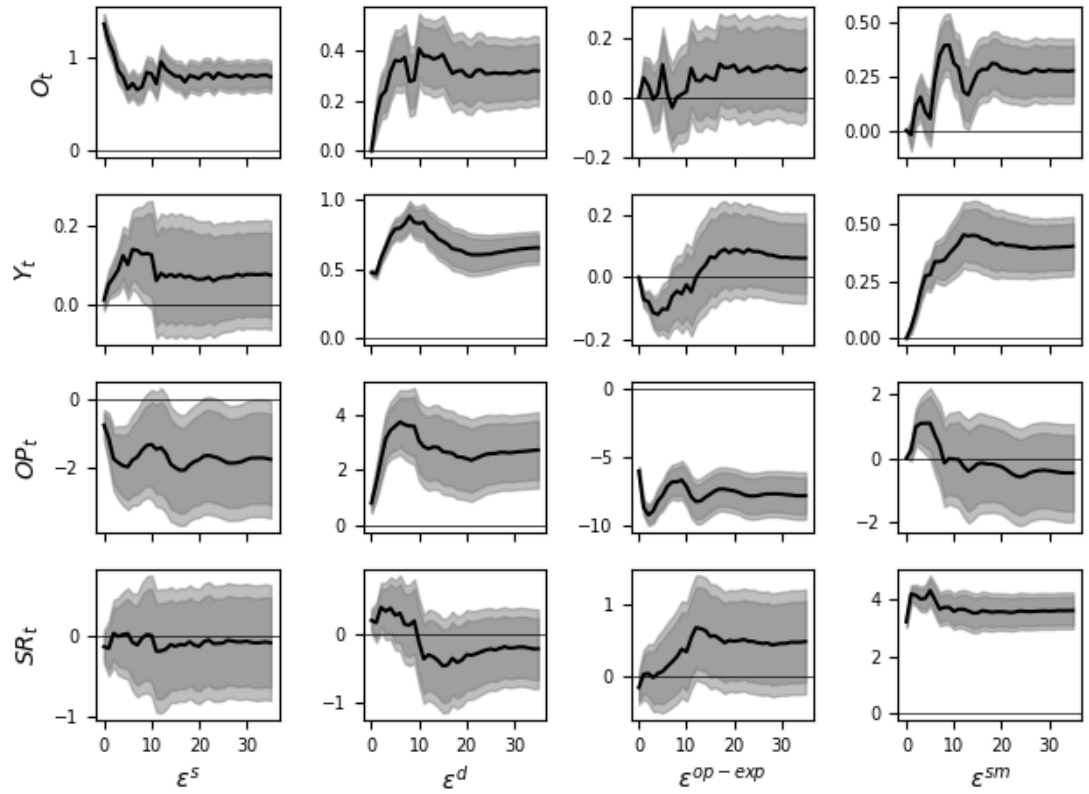
Figure 5.B.4: Estimated structural shocks, averaged to annual frequency.



Note: The figure shows the average across years for each estimated structural shocks of the block-recursive SVAR pGMM estimator.

Figure 5.B.5 shows the IRF for the recursive oil market SVAR from Section 5.6 estimated with the SVAR GMM estimator from Equation (5.9). In the recursive SVAR, the GMM estimator is just identified and equal to the estimator obtained by applying the Cholesky decomposition to the variance-covariance matrix of the reduced form shocks.

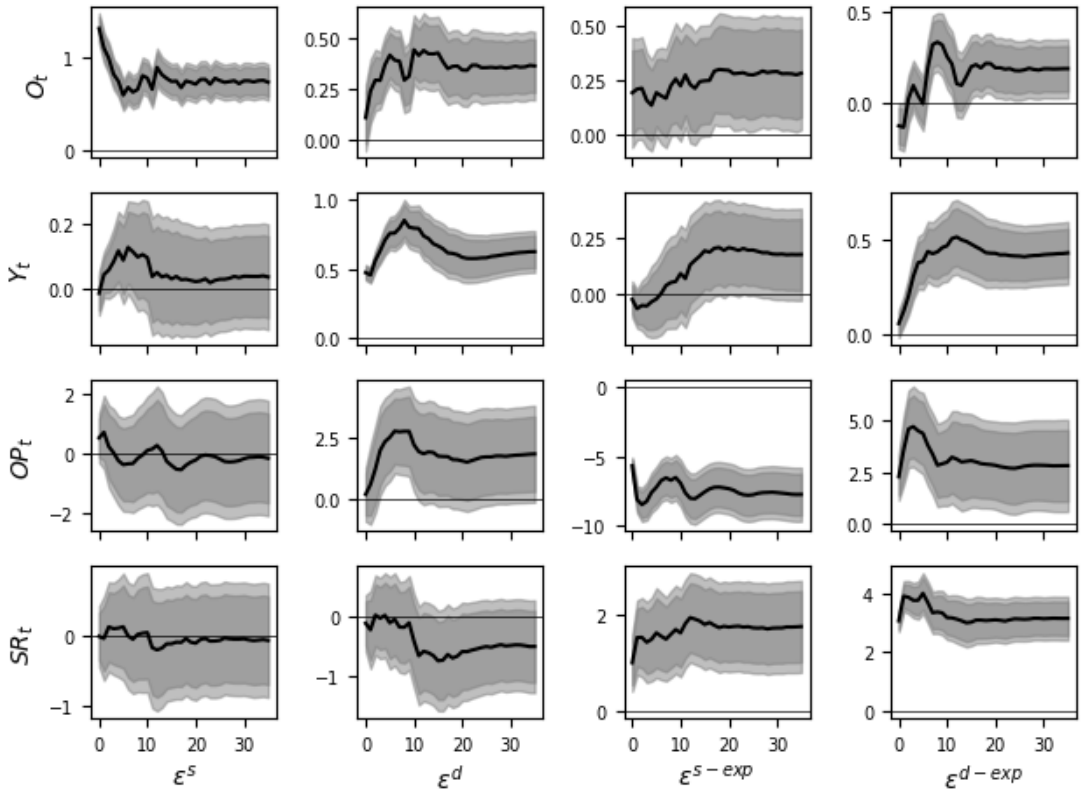
Figure 5.B.5: Impulse Responses of the recursive SVAR GMM estimator.



Note: Impulse responses to the recursive oil market SVAR from Section 5.6 estimated with the recursive SVAR GMM estimator from Equation (5.9), equal to the estimator obtained by applying the Cholesky decomposition to the variance-covariance matrix of the reduced form shocks. Confidence bands are symmetric 68% and 80% bands based on standard errors and 500 replications. The rows show the cumulative responses. The shock ε^{op-exp} denotes a speculative oil price shock and the shock ε^{sm} represents a residual stock market shock.

Figure 5.B.6 shows the IRF for the unrestricted oil market SVAR from Section 5.6 estimated with the unrestricted SVAR GMM estimator from Equation (5.9) where the weighting matrix is continuously updated and estimated based on the assumption of serially and mutually independent shocks.

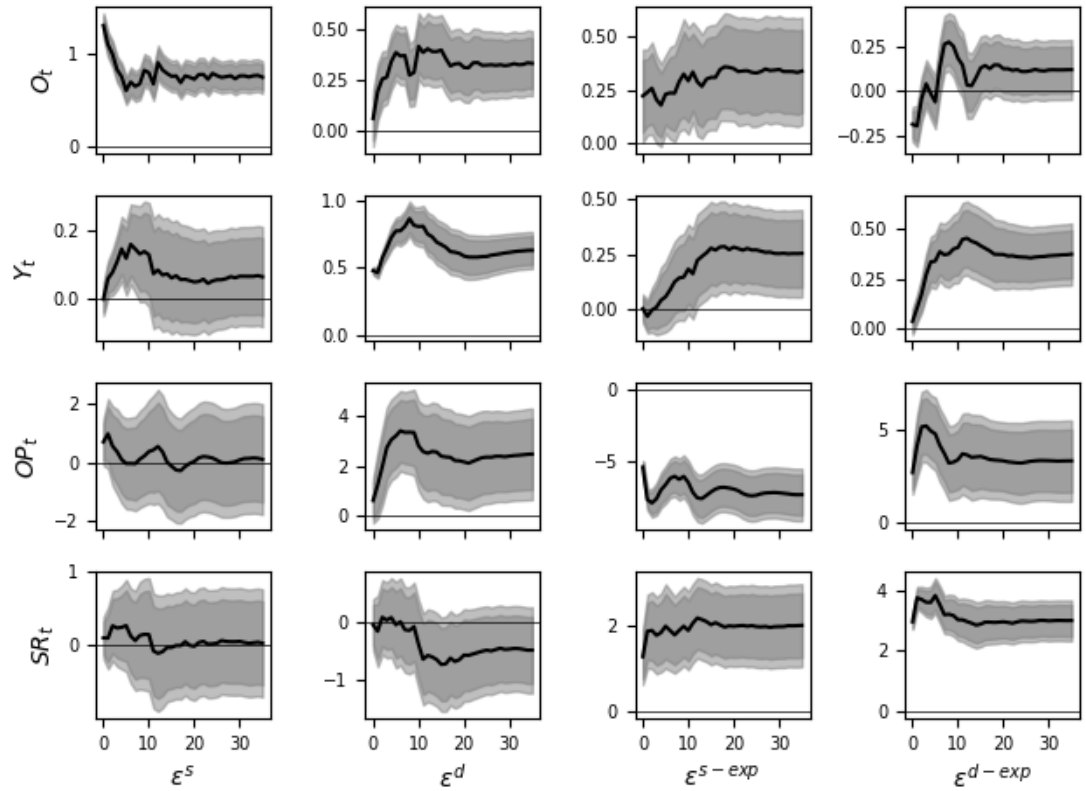
Figure 5.B.6: Impulse Responses of the unrestricted SVAR GMM estimator.



Note: Impulse responses to the estimated structural shocks for the unrestricted oil market SVAR from Section 5.6 estimated with the unrestricted SVAR GMM estimator from Equation (5.9) where the weighting matrix is continuously updated and estimated based on the assumption of serially and mutually independent shocks. Confidence bands are symmetric 68% and 80% bands based on standard errors and 500 replications. The rows show the cumulative responses.

Figure 5.B.7 shows the IRF for the unrestricted oil market SVAR from Section 5.6 estimated with the overidentified unrestricted SVAR GMM estimator from Equation (5.10) where the weighting matrix is continuously updated and estimated based on the assumption of serially and mutually independent shocks.

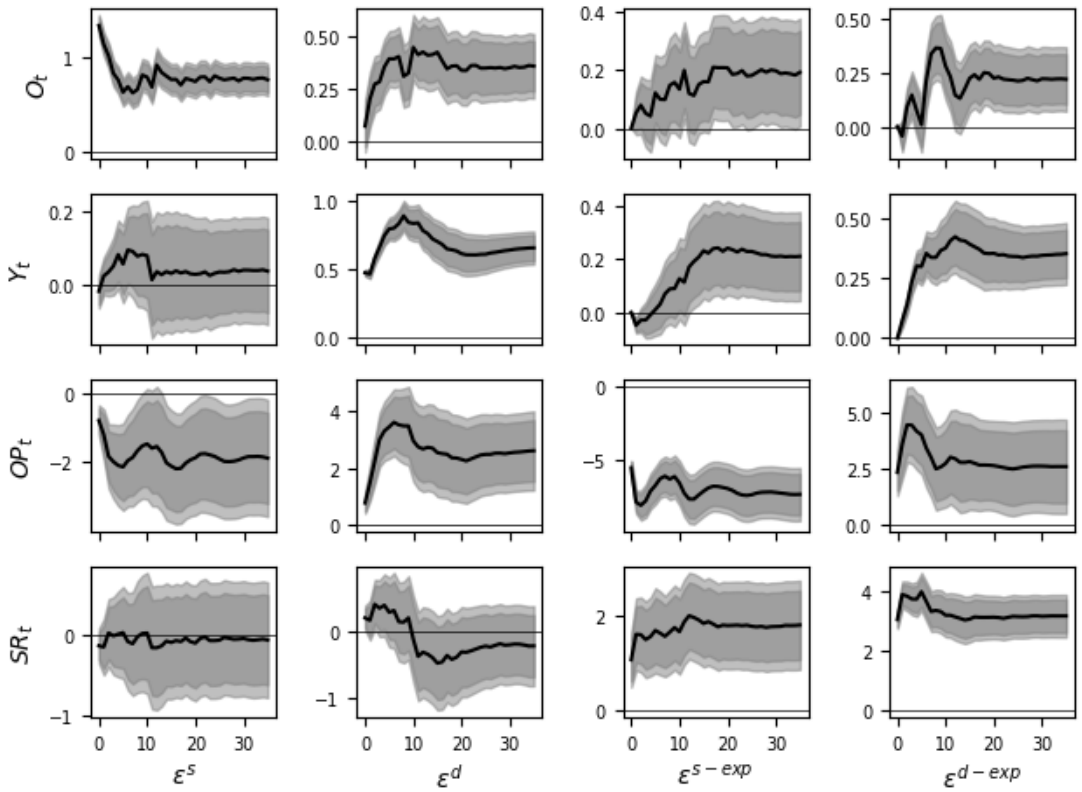
Figure 5.B.7: Impulse Responses of the unrestricted SVAR oGMM estimator.



Note: Impulse responses to the estimated structural shocks for the unrestricted oil market SVAR from Section 5.6 estimated with the overidentified unrestricted SVAR GMM estimator from Equation (5.10) where the weighting matrix is continuously updated and estimated based on the assumption of serially and mutually independent shocks. Confidence bands are symmetric 68% and 80% bands based on standard errors and 500 replications. The rows show the cumulative responses.

Figure 5.B.8 shows the IRF for the block-recursive oil market SVAR from Section 5.6 estimated with the block-recursive SVAR GMM estimator from Equation (5.9) where the weighting matrix is continuously updated and estimated based on the assumption of serially and mutually independent shocks.

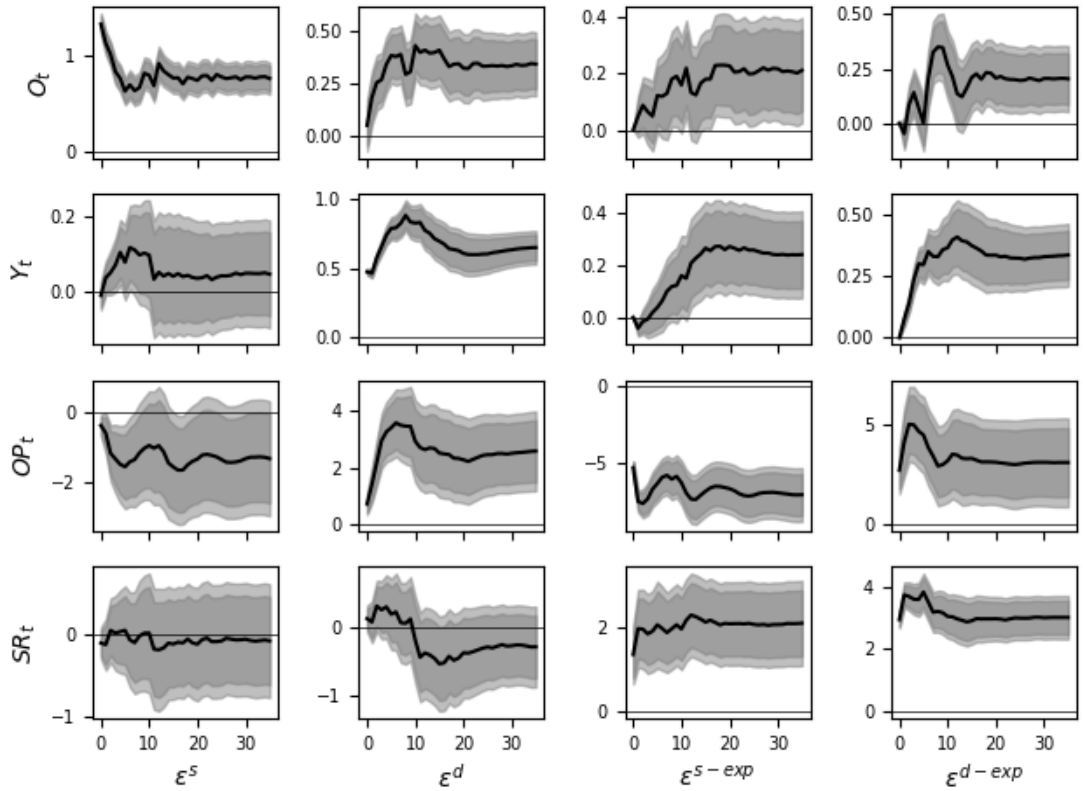
Figure 5.B.8: Impulse Responses of the block-recursive SVAR GMM estimator.



Note: Impulse responses to the estimated structural shocks for the block-recursive oil market SVAR from Section 5.6 estimated with the block-recursive SVAR GMM estimator from Equation (5.9) where the weighting matrix is continuously updated and estimated based on the assumption of serially and mutually independent shocks. Confidence bands are symmetric 68% and 80% bands based on standard errors and 500 replications. The rows show the cumulative responses.

Figure 5.B.9 shows the IRF for the block-recursive oil market SVAR from Section 5.6 estimated with the overidentified block-recursive SVAR GMM estimator from Equation (5.10) where the weighting matrix is continuously updated and estimated based on the assumption of serially and mutually independent shocks.

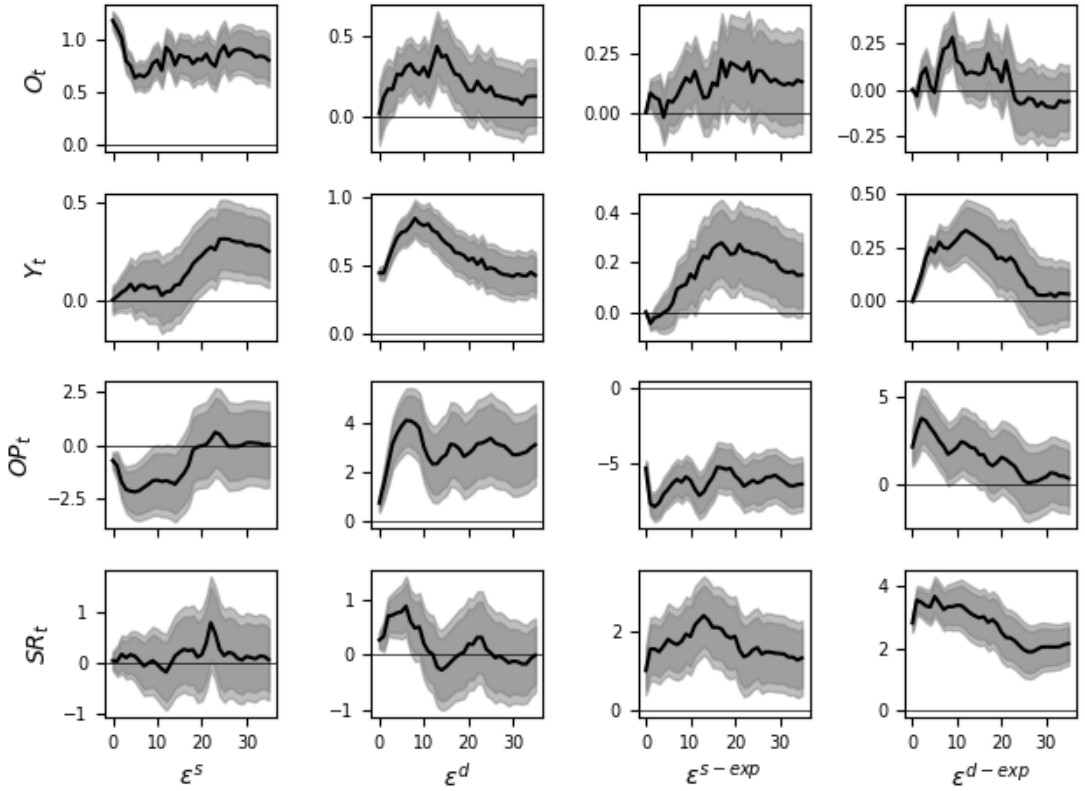
Figure 5.B.9: Impulse Responses of the block-recursive SVAR oGMM estimator.



Note: Impulse responses to the estimated structural shocks for the block-recursive oil market SVAR from Section 5.6 estimated with the overidentified block-recursive SVAR GMM estimator from Equation (5.10) where the weighting matrix is continuously updated and estimated based on the assumption of serially and mutually independent shocks. Confidence bands are symmetric 68% and 80% bands based on standard errors and 500 replications. The rows show the cumulative responses.

Figure 5.B.10 shows the IRF for the block-recursive oil market SVAR from Section 5.6 using 24 lags estimated with the block-recursive SVAR GMM estimator from Equation (5.9) where the weighting matrix is continuously updated and estimated based on the assumption of serially and mutually independent shocks.

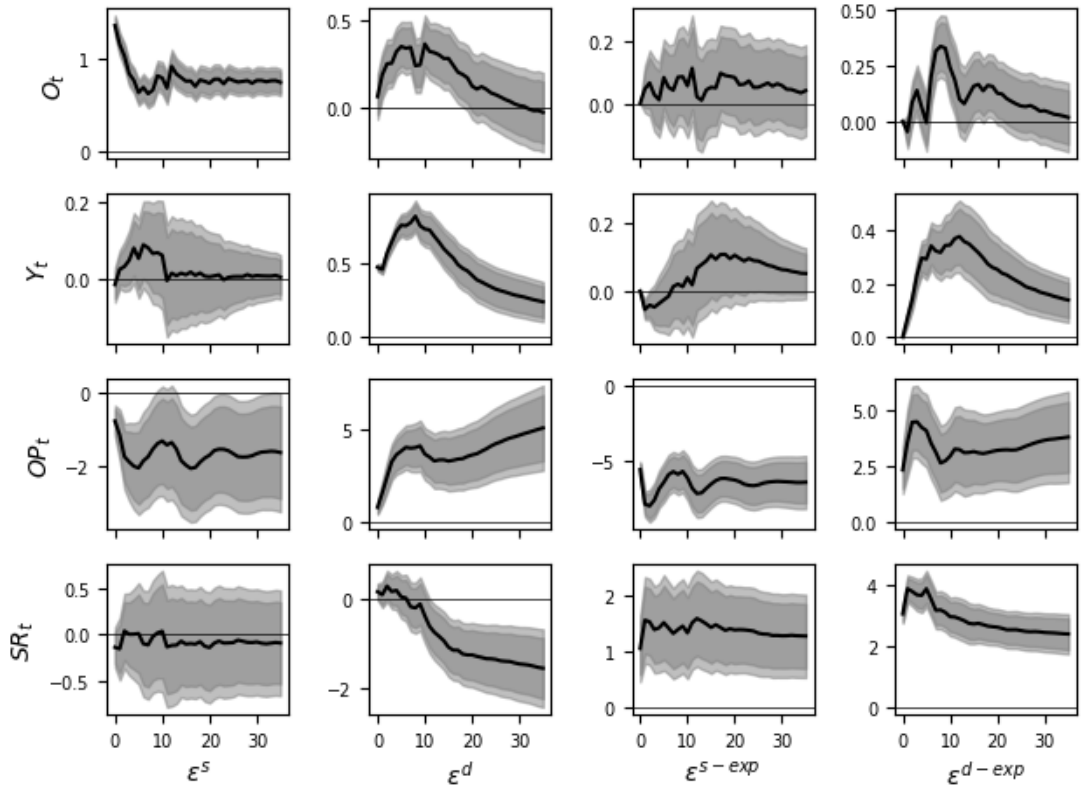
Figure 5.B.10: Impulse Responses of the block-recursive SVAR GMM estimator using 24 instead of 12 lags.



Note: Impulse responses to the estimated structural shocks for the block-recursive oil market SVAR from Section 5.6 using 24 lags estimated with the block-recursive SVAR GMM estimator from Equation (5.9) where the weighting matrix is continuously updated and estimated based on the assumption of serially and mutually independent shocks. Confidence bands are symmetric 68% and 80% bands based on standard errors and 500 replications. The rows show the cumulative responses.

Figure 5.B.11 shows the IRF for the block-recursive oil market SVAR from Section 5.6 using the percentage deviation of industrial production from a linear trend instead of the log difference of industrial production. The SVAR is estimated with the block-recursive SVAR GMM estimator from Equation (5.9) where the weighting matrix is continuously updated and estimated based on the assumption of serially and mutually independent shocks.

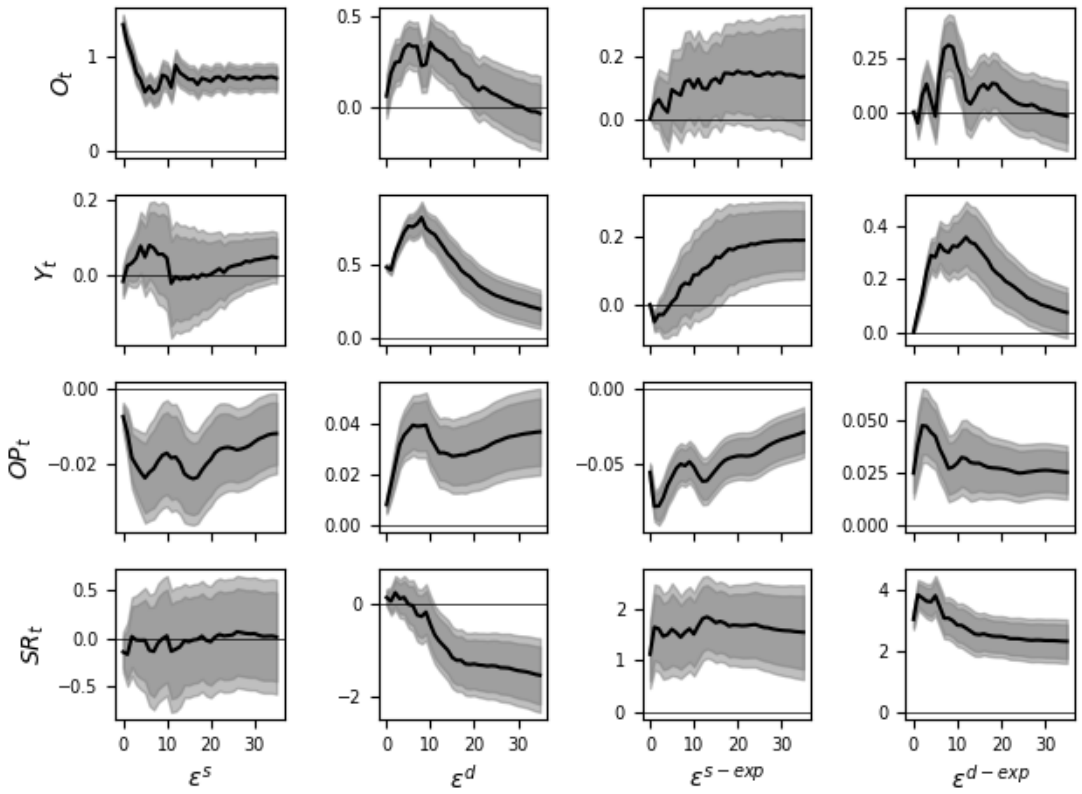
Figure 5.B.11: Impulse Responses of the block-recursive SVAR estimator using the percentage deviation of industrial production from a linear trend.



Note: Impulse responses to the estimated structural shocks for the block-recursive oil market SVAR from Section 5.6 using the percentage deviation of industrial production from a linear trend instead of the log difference of industrial production. The SVAR is estimated with the block-recursive SVAR GMM estimator from Equation (5.9) where the weighting matrix is continuously updated and estimated based on the assumption of serially and mutually independent shocks. Confidence bands are symmetric 68% and 80% bands based on standard errors and 500 replications. The rows O_t, OP_t, and SR_t show the cumulative responses.

Figure 5.B.12 shows the IRF for the block-recursive oil market SVAR from Section 5.6 using log of real oil price instead of real oil price growth and the percentage deviation of industrial production from a linear trend instead of the log difference of industrial production. The SVAR is estimated with the block-recursive SVAR GMM estimator from Equation (5.9) where the weighting matrix is continuously updated and estimated based on the assumption of serially and mutually independent shocks.

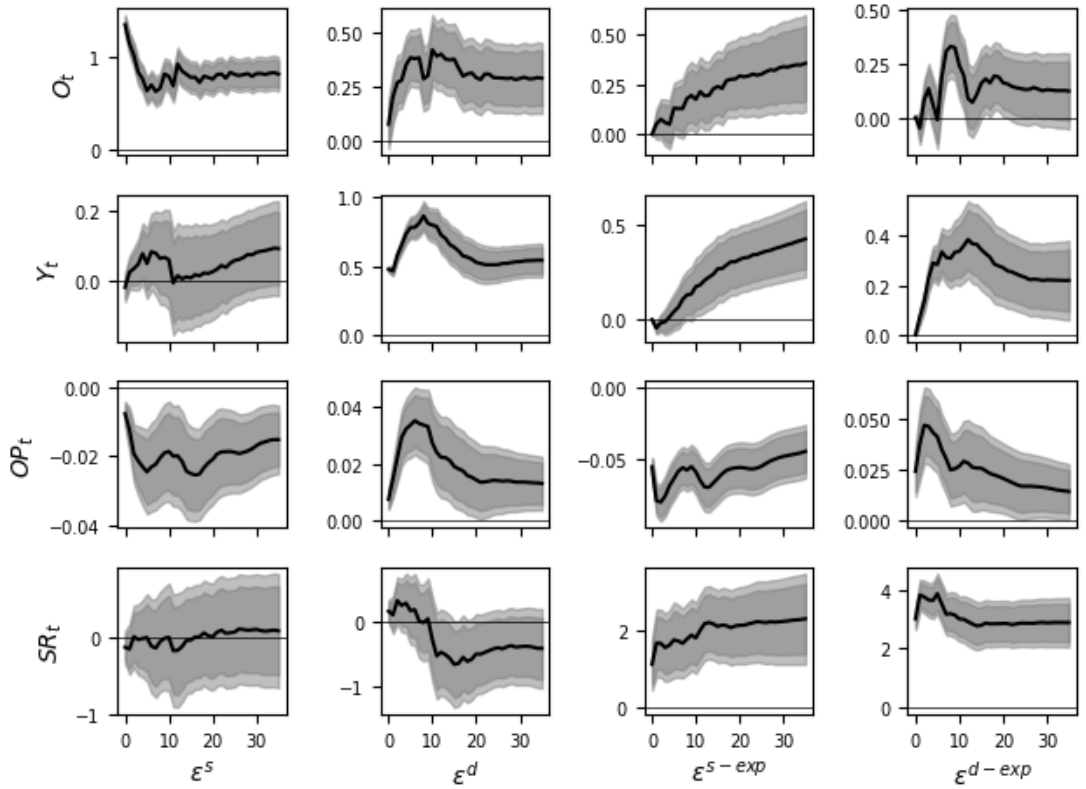
Figure 5.B.12: Impulse Responses of the block-recursive SVAR GMM estimator using the percentage deviation of industrial production from a linear trend and the log of the real oil price.



Note: Impulse responses to the estimated structural shocks for the block-recursive oil market SVAR from Section 5.6 using the log of the real oil price instead of real oil price growth and the percentage deviation of industrial production from a linear trend instead of the log difference of industrial production. The SVAR is estimated with the block-recursive SVAR GMM estimator from Equation (5.9) where the weighting matrix is continuously updated and estimated based on the assumption of serially and mutually independent shocks. Confidence bands are symmetric 68% and 80% bands based on standard errors and 500 replications. The rows O_t and SR_t show the cumulative responses.

Figure 5.B.13 shows the IRF for the block-recursive oil market SVAR from Section 5.6 using log of real oil price instead of real oil price growth. The SVAR is estimated with the block-recursive SVAR GMM estimator from Equation (5.9) where the weighting matrix is continuously updated and estimated based on the assumption of serially and mutually independent shocks.

Figure 5.B.13: Impulse Responses of the block-recursive SVAR GMM estimator using the log of the real oil price.



Note: Impulse responses to the estimated structural shocks for the block-recursive oil market SVAR from Section 5.6 using the log of the real oil price instead of real oil price growth. The SVAR is estimated with the block-recursive SVAR GMM estimator from Equation (5.9) where the weighting matrix is continuously updated and estimated based on the assumption of serially and mutually independent shocks. Confidence bands are symmetric 68% and 80% bands based on standard errors and 500 replications. The rows O_t, Y_t, and SR_t show the cumulative responses.

Appendix 5.C Derviation of the Lemmata used in the Proof of Proposition 5.3

5.C.1 Notation and preparations part 1

Consider a recursive SVAR $u = B_0\epsilon$ with independent structural shocks with mean zero and unit variance. Let $A := B_0^{-1}$ and a_{ql} $[b_{ql}]$ denote the element at row q and column l of A $[B_0]$. Moreover, let $\omega_{i^2} := \omega_{ii} := E\left[\epsilon_i^2\right]$, $\omega_{i^3} := \omega_{iii} := E\left[\epsilon_i^3\right]$, and $\omega_{i^4} := \omega_{iiii} := E\left[\epsilon_i^4\right]$ for $i = 1, \ldots, n$. Throughout this part of the appendix, the superscript $^{(*)}$ indicates that the equality follows from $e(B_0) = \epsilon$ with ϵ being mutually independent with mean zero

and unit variance. Additionally, the superscript $^{(**)}$ indicates that the equality follows from B_0 and hence A_0 being recursive.

We divide the variance-covariance conditions $E[f_{\mathbf{2}}(B, u_t)]$ into a set of variance conditions

$$E[f_{\mathbf{2_M}}(B, u_t)] := E \begin{bmatrix} e(B)_{1,t}^2 - 1 \\ \vdots \\ e(B)_{n,t}^2 - 1 \end{bmatrix},$$

and $n-1$ sets of covariance conditions $E[f_{\mathbf{2_{C_1}}}(B, u_t)], \ldots, E[f_{\mathbf{2_{C_{n-1}}}}(B, u_t)]$ where

$$E[f_{\mathbf{2_{C_i}}}(B, u_t)] := E \begin{bmatrix} e(B)_{i,t} e(B)_{i+1,t} \\ \vdots \\ e(B)_{i,t} e(B)_{n,t} \end{bmatrix}, \text{ for } i = 1, \ldots, n-1.$$

We divide the coskewness conditions $E[f_{\mathbf{3}}(B, u_t)]$ into n subsets

$$E[f_{\mathbf{3_{ii}}}(B, u_t)] := E \begin{bmatrix} e(B)_{1,t} e(B)_{i,t}^2 \\ \vdots \\ e(B)_{i-1,t} e(B)_{i,t}^2 \\ e(B)_{i,t}^2 e(B)_{i+1,t} \\ \vdots \\ e(B)_{i,t}^2 e(B)_{n,t} \end{bmatrix}, \text{ for } i = 1, \ldots, n,$$

and one additional subset $E[f_{\mathbf{3_{rest}}}(B, u_t)]$ containing all remaining coskewness conditions of $E[f_{\mathbf{3}}(B, u_t)]$ not contained in a subset $E[f_{\mathbf{3_{ii}}}(B, u_t)]$, which are all coskewness conditions of the type $E[e(B)_{i,t} e(B)_{j,t} e(B)_{k,t}]$ with $i \neq j \neq k$.

We divide the cokurtosis conditions $E[f_{\mathbf{4}}(B, u_t)]$ into n subsets

$$E[f_{\mathbf{4_{ii}}}(B, u_t)] = E \begin{bmatrix} e(B)_{1,t} e(B)_{i,t}^3 \\ \vdots \\ e(B)_{i-1,t} e(B)_{i,t}^3 \\ e(B)_{i,t}^3 e(B)_{i+1,t} \\ \vdots \\ e(B)_{i,t}^3 e(B)_{n,t} \end{bmatrix}, \text{ for } i = 1, \ldots, n,$$

$n-1$ subsets

$$E[f_{\mathbf{4_{iii}}}(B, u_t)] = E \begin{bmatrix} e(B)_{i,t}^2 e(B)_{i+1,t}^2 \\ \vdots \\ e(B)_{i,t}^2 e(B)_{n,t}^2 \end{bmatrix}, \text{ for } i = 1, \ldots, n-1,$$

and one additional subset $E[f_{\mathbf{4_{rest}}}(B, u_t)]$ containing all remaining cokurtosis conditions of $E[f_{\mathbf{4}}(B, u_t)]$ not contained in a subset $E[f_{\mathbf{4_{ii}}}(B, u_t)]$ or $E[f_{\mathbf{4_{iii}}}(B, u_t)]$, which are all cokur-

tosis conditions of the type $E[e(B_0)_i e(B_0)_j e(B_0)_k e(B_0)_l]$ and $E[e(B_0)_i^2 e(B_0)_j e(B_0)_k]$ with $i \neq j \neq k \neq l$.

Throughout this part of the appendix, we will use the following Lemmata.

Lemma 5.C.1. *The derivative of the i-th element of the unmixed innovations at B_0 with respect to an element b_{pq} is given by*

$$\frac{\partial e_{it}(B_0)}{\partial b_{pq}} = \begin{cases} -a_{ip}\epsilon_{qt}, & \text{if } i \geq p \\ 0, & \text{else} \end{cases}.$$

Proof.

$$\begin{aligned}
\frac{\partial e_t(B_0)}{\partial b_{pq}} &= \frac{\partial B_0^{-1}}{\partial b_{pq}} u_t \\
&= \left(-B_0^{-1} \frac{\partial B_0}{\partial b_{pq}} B_0^{-1} \right) u_t \\
&= -B_0^{-1} \frac{\partial B_0}{\partial b_{pq}} B_0^{-1} B_0 \epsilon_t \\
&= -B_0^{-1} \frac{\partial B_0}{\partial b_{pq}} \epsilon_t \\
&= -A_0 \frac{\partial B_0}{\partial b_{pq}} \begin{bmatrix} \epsilon_{1t} \\ \vdots \\ \epsilon_{nt} \end{bmatrix} \\
&= - \begin{bmatrix} a_{1p} \\ \vdots \\ a_{np} \end{bmatrix} \epsilon_{qt} \overset{\text{recursive SVAR}}{=} - \begin{bmatrix} 0 \\ \vdots \\ 0 \\ a_{pp} \\ \vdots \\ a_{np} \end{bmatrix} \epsilon_{qt}
\end{aligned}$$

□

Lemma 5.C.2. *For $i = 1, \ldots, n-1$ and $j = 1, \ldots, n$ let*

$$G_{\mathbf{2}}^{b_{j:,j}} := \left[G_{\mathbf{2_M}}^{b_{j:,j}}, G_{\mathbf{2_{C_1}}}^{b_{j:,j}}, \ldots, G_{\mathbf{2_{C_{n-1}}}}^{b_{j:,j}} \right]'$$

with

$$G_{\mathbf{2_M}}^{b_{j:,j}} := E\left[\frac{\partial f_{\mathbf{2_M}}(u_t, B_0)}{\partial (b_{jj}, \ldots, b_{nj})'}\right],$$

$$G_{\mathbf{2_{C_i}}}^{b_{j:,j}} := E\left[\frac{\partial f_{\mathbf{2_{C_i}}}(u_t, B_0)}{\partial (b_{jj}, \ldots, b_{nj})'}\right].$$

Then

$$G_{\mathbf{2_M}}^{b_{j:,j}} = -2\underbrace{\begin{bmatrix} & 0_{(j-1)\times(n-j+1)} & \\ a_{jj} & \cdots & a_{jn} \\ & 0_{(n-j)\times(n-j+1)} & \end{bmatrix}}_{n\times(n-j+1)} = -2\underbrace{\begin{bmatrix} & 0_{(j-1)\times(n-j+1)} & \\ a_{jj} & 0_{1\times(n-j)} & \\ & 0_{(n-j)\times(n-j+1)} & \end{bmatrix}}_{n\times(n-j+1)} \tag{5.C.1}$$

and

$$G_{\mathbf{2_{C_i}}}^{b_{j:,j}} = 0_{(n-i)\times(n-j+1)}, \qquad \text{for } i \neq j, \tag{5.C.2}$$

$$G_{\mathbf{2_{C_i}}}^{b_{i:,i}} = -\underbrace{\begin{bmatrix} a_{i+1,i} & a_{i+1,i+1} & & 0 \\ \vdots & & \ddots & \\ a_{ni} & a_{n,i+1} & \ldots & a_{nn} \end{bmatrix}}_{(n-i)\times(n-i+1)}. \tag{5.C.3}$$

Proof. Equation (5.C.1): The (q,r)-th entry of $G_{\mathbf{2_{C_i}}}^{b_{j:,j}}$ is equal to

$$G_{\mathbf{2_{C_i}}}^{b_{j:,j}}(q,r) = E\left[\frac{\partial(e(B_0)_q^2 - 1)}{\partial b_{j+r-1,j}}\right] \overset{(*)}{=} \begin{cases} -2a_{q,j+r-1}, & \text{if } q = j \\ 0, & \text{else} \end{cases}.$$

Equation (5.C.2) and (5.C.3): The (q,r)-th entry of $G_{\mathbf{2_{C_i}}}^{b_{j:,j}}$ is equal to

$$G_{\mathbf{2_{C_i}}}^{b_{j:,j}}(q,r) = E\left[\frac{\partial(e(B_0)_i e(B_0)_{i+q})}{\partial b_{j+r-1,j}}\right] \overset{(*)}{=} \begin{cases} -a_{i+q,j+r-1}, & \text{if } j = i \\ -a_{i,j+r-1} \overset{(**)}{=} 0, & \text{if } j = (i+q) \\ 0, & \text{else} \end{cases}.$$

□

Lemma 5.C.3. *For $i, j = 1, \ldots, n$ let*

$$G_{\mathbf{3_{ii}}}^{b_{j:,j}} := E\left[\frac{\partial f_{\mathbf{3_{ii}}}(u_t, B_0)}{\partial (b_{jj}, \ldots, b_{nj})'}\right],$$

$$G_{\mathbf{3_{rest}}}^{vec(B)} := E\left[\frac{\partial f_{\mathbf{3_{rest}}}(u_t, B_0)}{\partial vec(B)'}\right].$$

Then

$$G_{\mathbf{3_{ii}}}^{b_{j:,j}} = 0_{(n-1)\times(n-j+1)}, \qquad \text{for } i \neq j, \quad \text{(5.C.4)}$$

$$G_{\mathbf{3_{ii}}}^{b_{i:,i}} = -\omega_{iii} \underbrace{\begin{bmatrix} & 0_{(i-1)\times(n-i+1)} & & \\ a_{i+1,i} & a_{i+1,i+1} & & 0 \\ \vdots & & \ddots & \\ a_{ni} & a_{n,i+1} & \dots & a_{nn} \end{bmatrix}}_{(n-1)\times(n-i+1)},$$

and

$$G_{\mathbf{3_{rest}}}^{vec(B)} = 0. \qquad \text{(5.C.5)}$$

Proof. Equation (5.C.4): The (q,r)-th entry of $G_{\mathbf{3_{ii}}}^{b_{j:,j}}$ with $q < i$ is equal to

$$G_{\mathbf{3_{ii}}}^{b_{j:,j}}(q,r) = E\left[\frac{\partial(e(B_0)_i^2 e(B_0)_q)}{\partial b_{j+r-1,j}}\right] = \begin{cases} -a_{q,j+r-1}\omega_{iii}, & j = i, q \geq r+i-1 \\ -a_{q,j+r-1}\omega_{iii} \overset{(**)}{=} 0, & j = i, q < r+i-1 \\ 0, & \text{else} \end{cases}$$

and the (q,l)-th entry of $G_{\mathbf{3_{ii}}}^{b_{j:,j}}$ with $q \geq i$ is equal to

$$G_{\mathbf{3_{ii}}}^{b_{j:,j}}(q,r) = E\left[\frac{\partial(e(B_0)_i^2 e(B_0)_{q+1})}{\partial b_{j+r-1,j}}\right] = \begin{cases} -a_{q+1,j+r-1}\omega_{iii}, & j = i, q \geq r+i-1 \\ -a_{q+1,j+r-1}\omega_{iii} \overset{(**)}{=} 0, & j = i, q < r+i-1 \\ 0, & \text{else} \end{cases}.$$

Every element in $G_{\mathbf{3_{rest}}}^{vec(B)}$ can be written as $E[\frac{\partial(e(B)_{a,t}e(B)_{b,t}e(B)_{c,t})}{\partial b_{q,l}}]$ for some $a, b, c \in \{1, \dots, n\}$ with $a \neq b \neq c$. Equation (5.C.5) follows with Lemma 5.C.1, $e(B_0) = \epsilon$, and independence and mean zero of ϵ_t.

□

Lemma 5.C.4. *Let*

$$G_{\mathbf{4_{ii}}}^{b_{j:,j}} := E\left[\frac{\partial f_{\mathbf{4_{ii}}}(u_t, B_0)}{\partial (b_{jj}, \dots, b_{nj})'}\right], \qquad \text{for } i = 1, \dots, n-1,\ j = 1, \dots, n,$$

$$G_{\mathbf{4_{iii}}}^{b_{j:,j}} := E\left[\frac{\partial f_{\mathbf{4_{iii}}}(u_t, B_0)}{\partial (b_{jj}, \dots, b_{nj})'}\right], \qquad \text{for } i,\ j = 1, \dots, n,$$

$$G_{\mathbf{4_{rest}}}^{vec(B)} := E\left[\frac{\partial f_{\mathbf{4_{rest}}}(u_t, B_0)}{\partial vec(B)'}\right].$$

Then

$$G^{b_{i:,i}}_{\mathbf{4_{ii}}} = -2\underbrace{\begin{bmatrix} a_{ii} & a_{in} \\ & \\ a_{ii} & a_{in} \end{bmatrix}}_{(n-i)\times(n-i+1)} = -2\underbrace{\begin{bmatrix} a_{ii} & \\ & 0_{(n-i)\times(n-j)} \\ a_{ii} & \end{bmatrix}}_{(n-i)\times(n-i+1)}, \tag{5.C.6}$$

$$G^{b_{j:,j}}_{\mathbf{4_{ii}}} = 0_{(n-i)\times(n-j+1)}, \qquad \textit{for } i>j, \tag{5.C.7}$$

$$G^{b_{j:,j}}_{\mathbf{4_{ii}}} = -2\underbrace{\begin{bmatrix} & 0_{(j-i-1)\times(n-j+1)} & \\ a_{jj} & \cdots & a_{jn} \\ & 0_{(n-j)\times(n-j+1)} & \end{bmatrix}}_{(n-i)\times(n-j+1)} = -2\underbrace{\begin{bmatrix} & 0_{(j-i-1)\times(n-j+1)} & \\ a_{jj} & 0_{(1)\times(n-j)} & \\ & 0_{(n-j)\times(n-j+1)} & \end{bmatrix}}_{(n-i)\times(n-j+1)}, \quad \textit{for } i<j, \tag{5.C.8}$$

and

$$G^{b_{i:,i}}_{\mathbf{4_{iii}}} = -\omega_{iiii}\underbrace{\begin{bmatrix} a_{1i} & a_{1n} \\ & \\ a_{i-1,i} & a_{i-1,n} \\ a_{i+1,i} & a_{i+1,n} \\ & \\ a_{ni} & a_{nn} \end{bmatrix}}_{(n-1)\times(n-i+1)} = -\omega_{iiii}\underbrace{\begin{bmatrix} & 0_{(i-1)\times(n-i+1)} & \\ & & \\ a_{i+1,i} & a_{i+1,i+1} & 0 \\ \vdots & & \ddots \\ a_{ni} & \cdots & a_{nn} \end{bmatrix}}_{(n-1)\times(n-i+1)}, \tag{5.C.9}$$

$$G^{b_{j:,j}}_{\mathbf{4_{iii}}} = -3\underbrace{\begin{bmatrix} & 0_{(j-1)\times(n-j+1)} & \\ a_{ij} & \cdots & a_{in} \\ & 0_{(n-j-1)\times(n-j+1)} & \end{bmatrix}}_{(n-1)\times(n-j+1)} \tag{5.C.10}$$

$$= -3\underbrace{\begin{bmatrix} & 0_{(j-1)\times(n-j+1)} & & \\ a_{ij} & \cdots & a_{ii} & 0_{1\times(n-i)} \\ & 0_{(n-j-1)\times(n-j+1)} & & \end{bmatrix}}_{(n-1)\times(n-j+1)}, \qquad \textit{for } i>j,$$

$$G^{b_{j:,j}}_{\mathbf{4_{iii}}} = -3\underbrace{\begin{bmatrix} & 0_{(j-1-1)\times(n-j+1)} & \\ a_{ij} & \cdots & a_{in} \\ & 0_{(n-j)\times(n-j+1)} & \end{bmatrix}}_{(n-1)\times(n-j+1)} = 0_{(n-1)\times(n-j+1)}, \qquad \textit{for } i<j, \tag{5.C.11}$$

and

$$G^{vec(B)}_{\mathbf{4_{rest}}} = 0. \tag{5.C.12}$$

Proof. Equation (5.C.6), (5.C.7), and (5.C.8): The (q,r)-th entry of $G_{\mathbf{4_{ii}}}^{b_{j:,j}}$ is equal to

$$G_{\mathbf{4_{ii}}}^{b_{j:,j}}(q,r) = E\left[\frac{\partial(e(B_0)_i^2 e(B_0)_{i+q}^2 - 1)}{\partial b_{j+r-1,j}}\right] \overset{(*)}{=} \begin{cases} -2a_{i,j+r-1}, & \text{if } j = i, r = 1 \\ -2a_{i,j+r-1} \overset{(**)}{=} 0, & \text{if } j = i, r \neq 1 \\ -2a_{i+q,j+r-1}, & \text{if } j = i+q, r = 1 \\ -2a_{i+q,j+r-1} \overset{(**)}{=} 0, & \text{if } j = i+q, r \neq 1 \\ 0, & \text{else} \end{cases}.$$

Equation (5.C.9), (5.C.10), and (5.C.11): The (q,r)-th entry of $G_{\mathbf{4_{iii}}}^{b_{j:,j}}$ with $q < i$ is equal to

$$G_{\mathbf{4_{iii}}}^{b_{j:,j}}(q,r) = E\left[\frac{\partial(e(B_0)_i^3 e(B_0)_q)}{\partial b_{j+r-1,j}}\right] \overset{(*)}{=} \begin{cases} -a_{q,j+r-1}\omega_{iiii} \overset{(**)}{=} 0, & \text{if } j = i \\ -3a_{i,j+r-1}, & \text{if } j = q, r \leq i-j+1 \\ -3a_{i,j+r-1} \overset{(**)}{=} 0, & \text{if } j = q, r > i-j+1 \\ 0, & \text{else} \end{cases}.$$

The (q,r)-th entry of P_i^j with $q \geq i$ is equal to

$$G_{\mathbf{4_{iii}}}^{b_{j:,j}}(q,r) = E\left[\frac{\partial(e(B_0)_i^3 e(B_0)_{q+1})}{\partial b_{j+r-1,j}}\right] \overset{(*)}{=} \begin{cases} -a_{q+1,j+r-1}\omega_{iiii}, & \text{if } j = i, q-r \geq j-2 \\ -a_{q+1,j+r-1}\omega_{iiii} \overset{(**)}{=} 0, & \text{if } j = i, q-r < j-2 \\ -3a_{i,j+r-1}, & \text{if } j = q+1 \\ 0, & \text{else} \end{cases}.$$

Equation (5.C.12) follows analogously to Equation (5.C.5). □

Lemma 5.C.5. *The matrix $S_{\mathbf{2}} = E\left[f_{\mathbf{2}}(u_t, B_0) f_{\mathbf{2}}(u_t, B_0)'\right]$ can be written as*

$$S_{\mathbf{2}} = \begin{bmatrix} S_{\mathbf{2}_M} & S_{\mathbf{2}_M\mathbf{2}_C} \\ S_{\mathbf{2}_C\mathbf{2}_M} & S_{\mathbf{2}_C} \end{bmatrix}, \text{ with } \begin{aligned} S_{\mathbf{2}_M} &:= E\left[f_{\mathbf{2}_M}(u_t, B_0) f_{\mathbf{2}_M}(u_t, B_0)'\right], \\ S_{\mathbf{2}_M\mathbf{2}_C} &:= E\left[f_{\mathbf{2}_M}(u_t, B_0) f_{\mathbf{2}_C}(u_t, B_0)'\right], \\ S_{\mathbf{2}_C\mathbf{2}_M} &:= E\left[f_{\mathbf{2}_C}(u_t, B_0) f_{\mathbf{2}_M}(u_t, B_0)'\right], \\ S_{\mathbf{2}_C} &:= E\left[f_{\mathbf{2}_C}(u_t, B_0) f_{\mathbf{2}_C}(u_t, B_0)'\right], \end{aligned}$$

and

$$S_{\mathbf{2}_M} = \underbrace{\begin{bmatrix} \omega_{1111} - 1 & & 0 \\ & \ddots & \\ 0 & & \omega_{nnnn} - 1 \end{bmatrix}}_{n \times n}, \tag{5.C.13}$$

$$S_{\mathbf{2}_M\mathbf{2}_C} = 0_{n \times n(n-1)}, \tag{5.C.14}$$

$$S_{\mathbf{2}_C\mathbf{2}_M} = 0_{n(n-1)\times n},$$

$$S_{\mathbf{2}_C} = I_{n(n-1)\times n(n-1)}. \tag{5.C.15}$$

Therefore, $S_{\mathbf{2}}$ is equal to

$$S_{\mathbf{2}} = \begin{bmatrix} \omega_{1111} - 1 & & 0 & \\ & \ddots & & 0_{n\times n(n-1)} \\ 0 & & \omega_{nnnn} - 1 & \\ & 0_{n(n-1)\times n} & & I_{n(n-1)\times n(n-1)} \end{bmatrix}. \tag{5.C.16}$$

Proof. Equation (5.C.13), (5.C.14), (5.C.15), and (5.C.16): The (q,r)-th entry of $S_{\mathbf{2}_M}$ is equal to

$$S_{\mathbf{2}_M}(q,r) = E\left[(e(B_0)_q^2 - 1)(e(B_0)_r^2 - 1)\right] \overset{(*)}{=} \begin{cases} \omega_{qqqq} - 1, & \text{if } q = r \\ 0, & \text{else} \end{cases}.$$

Any entry of $S_{\mathbf{2}_M\mathbf{2}_C}$ can be written as

$$E\left[(e(B_0)_a^2 - 1)(e(B_0)_b e(B_0)_c)\right] \overset{(*)}{=} 0$$

for some $a, b, c \in \{1, \ldots, n\}$ with $b \neq c$. Any entry of $S_{\mathbf{2}_C\mathbf{2}_C}$ can be written as

$$E\left[(e(B_0)_a e(B_0)_b)(e(B_0)_c e(B_0)_d)\right] \overset{(*)}{=} \begin{cases} 1, & \text{if } a = c, b = d \\ 0, & \text{else} \end{cases},$$

for some $a, b, c, d \in \{1, \ldots, n\}$ with $a \neq b$ and $c \neq d$. The case $a = c$ and $b = d$ occurs at the diagonal of $S_{\mathbf{2}_C\mathbf{2}_C}$.

□

Lemma 5.C.6. *For $i = 1, \ldots, n$ and $j = 1, \ldots, n-1$ let*

$$S_{\mathbf{3_{ii}2}} = \begin{bmatrix} S_{\mathbf{3_{ii}2_M}} & S_{\mathbf{3_{ii}2_{C_1}}} & \cdots & S_{\mathbf{3_{ii}2_{C_{n-1}}}} \end{bmatrix},$$

$$S_{\mathbf{3_{rest}2}} = \begin{bmatrix} S_{\mathbf{3_{rest}2_M}} & S_{\mathbf{3_{rest}2_{C_1}}} & \cdots & S_{\mathbf{3_{rest}2_{C_{n-1}}}} \end{bmatrix}$$

with

$$S_{\mathbf{3_{ii}2_M}} := E\left[f_{\mathbf{3_{ii}}}(u_t, B_0) f_{\mathbf{2_M}}(u_t, B_0)'\right],$$

$$S_{\mathbf{3_{ii}2_{C_j}}} := E\left[f_{\mathbf{3_{ii}}}(u_t, B_0) f_{\mathbf{2_{C_j}}}(u_t, B_0)'\right],$$

$$S_{\mathbf{3_{rest}2_M}} := E\left[f_{\mathbf{3_{rest}}}(u_t, B_0) f_{\mathbf{2_M}}(u_t, B_0)'\right],$$

$$S_{\mathbf{3_{rest}2_{C_j}}} := E\left[f_{\mathbf{3_{rest}}}(u_t, B_0) f_{\mathbf{2_{C_j}}}(u_t, B_0)'\right],$$

and

$$S_{\mathbf{3_{ii}2_M}} = \underbrace{\begin{bmatrix} \omega_{111} & & 0 & 0 & & & \\ & \ddots & & 0 & & 0_{(n-i)\times(n-i)} & \\ 0 & & & \omega_{(i-1)^3} \; 0 & & & \\ & & & 0 \quad \omega_{(i+1)^3} & & & 0 \\ & 0_{(i-1)\times(i-1)} & & 0 & & \ddots & \\ & & & 0 \quad 0 & & & \omega_{nnn} \end{bmatrix}}_{(n-1)\times(n)} \tag{5.C.17}$$

and

$$S_{\mathbf{3_{ii}2_{C_i}}} = \omega_{iii} \underbrace{\begin{bmatrix} 0_{(i-1)\times(n-i)} \\ I_{(n-i)\times(n-i)} \end{bmatrix}}_{(n-1)\times(n-j)}, \qquad \text{for } i = 1, \ldots, n-1, \tag{5.C.18}$$

$$S_{\mathbf{3_{ii}2_{C_j}}} = 0_{(n-1)\times(n-j)}, \qquad \text{for } i < j, \tag{5.C.19}$$

$$S_{\mathbf{3_{ii}2_{C_j}}} = \omega_{iii} \underbrace{\begin{bmatrix} & 0_{(j-1)\times(n-j)} & \\ 0_{1\times(i-j-1)} & 1 & 0_{1\times(n-i)} \\ & 0_{(n-j-1)\times(n-j)} & \end{bmatrix}}_{(n-1)\times(n-j)}, \qquad \text{for } i > j, \tag{5.C.20}$$

and

$$S_{\mathbf{3_{rest}2}} = 0. \tag{5.C.21}$$

Proof. Equation (5.C.17): The (q,r)-th entry of $S_{\mathbf{3_{ii}2_M}}$ with $q < i$ is equal to

$$S_{\mathbf{3_{ii}2_M}}(q,r) = E\left[(e(B_0)_i^2 e(B_0)_q)(e(B_0)_r^2 - 1)\right] \overset{(*)}{=} \begin{cases} \omega_{qqq}, & \text{if } r = q \\ 0, & \text{else} \end{cases}$$

and the (q,r)-th entry of $S_{\mathbf{3_{ii}2_M}}$ with $q \geq i$ is equal to

$$S_{\mathbf{3_{ii}2_M}}(q,r) = E\left[(e(B_0)_i^2 e(B_0)_{q+1})(e(B_0)_r^2 - 1)\right] \overset{(*)}{=} \begin{cases} \omega_{(q+1)^3}, & \text{if } r = q+1 \\ 0, & \text{else} \end{cases}.$$

Equation (5.C.18), (5.C.19) and (5.C.20): The (q,r)-th entry of $S_{\mathbf{3_{ii}2_{C_j}}}$ with $q < i$ is equal to

$$S_{\mathbf{3_{ii}2_{C_j}}}(q,l) = E\left[(e(B_0)_i^2 e(B_0)_q)(e(B_0)_j e(B_0)_{j+r})\right] \overset{(*)}{=} \begin{cases} \omega_{iii}, & \text{if } i = j+r, j = q \\ 0, & \text{else} \end{cases}$$

and the (q,r)-th entry of $S_{\mathbf{3_{ii}2_{C_j}}}$ with $q \geq i$ is equal to

$$S_{\mathbf{3_{ii}2_{C_j}}}(q,l) = E\left[(e(B_0)_i^2 e(B_0)_{q+1})(e(B_0)_j e(B_0)_{j+r})\right] \overset{(*)}{=} \begin{cases} \omega_{iii}, & \text{if } i = j, q+1 = j+r \\ 0, & \text{else} \end{cases}.$$

Equation (5.C.21) holds, since every moment condition in $E[f_{\mathbf{3_{rest}}}(u_t, B_0) f_{\mathbf{2}_M}(u_t, B_0)']$ can be written as

$$E\left[(e(B_0)_a e(B_0)_b e(B_0)_c)(e(B_0)_d^2 - 1)\right]$$

for some $a, b, c, d = \{1, \ldots, n\}$ and $a \neq b \neq c$, which implies
$E\left[(e(B_0)_a e(B_0)_b e(B_0)_c)(e(B_0)_d^2 - 1)\right] = 0$ by independence and mean zero of ϵ.

□

Lemma 5.C.7. *Let*

$$S_{\mathbf{4_{ii}2}} = \begin{bmatrix} S_{\mathbf{4_{ii}2_M}} & S_{\mathbf{4_{ii}2_{C_1}}} & \cdots & S_{\mathbf{4_{ii}2_{C_{n-1}}}} \end{bmatrix}, \qquad \textit{for } i = 1, \ldots, n-1,$$

$$S_{\mathbf{4_{iii}2}} = \begin{bmatrix} S_{\mathbf{4_{iii}2_M}} & S_{\mathbf{4_{iii}2_{C_1}}} & \cdots & S_{\mathbf{4_{iii}2_{C_{n-1}}}} \end{bmatrix}, \qquad \textit{for } i = 1, \ldots, n,$$

$$S_{\mathbf{4_{rest}2}} = \begin{bmatrix} S_{\mathbf{4_{rest}2_M}} & S_{\mathbf{4_{rest}2_{C_1}}} & \cdots & S_{\mathbf{4_{rest}2_{C_{n-1}}}} \end{bmatrix}$$

with

$$S_{\mathbf{4_{ii}2_M}} := E\left[f_{\mathbf{4_{ii}}}(u_t, B_0) f_{\mathbf{2}_M}(u_t, B_0)'\right], \qquad \textit{for } i = 1, \ldots, n-1,$$

$$S_{\mathbf{4_{ii}2_{C_j}}} := E\left[f_{\mathbf{4_{ii}}}(u_t, B_0) f_{\mathbf{2}_{C_j}}(u_t, B_0)'\right], \qquad \textit{for } i, j = 1, \ldots, n-1,$$

$$S_{\mathbf{4_{iii}2_M}} := E\left[f_{\mathbf{4_{iii}}}(u_t, B_0) f_{\mathbf{2}_M}(u_t, B_0)'\right], \qquad \textit{for } i = 1, \ldots, n,$$

$$S_{\mathbf{4_{iii}2_{C_j}}} := E\left[f_{\mathbf{4_{iii}}}(u_t, B_0) f_{\mathbf{2}_{C_j}}(u_t, B_0)'\right], \qquad \textit{for } j = 1, \ldots, n-1,\ i = 1, \ldots, n,$$

$$S_{\mathbf{4_{rest}2_M}} := E\left[f_{\mathbf{4_{rest}}}(u_t, B_0) f_{\mathbf{2}_M}(u_t, B_0)'\right],$$

$$S_{\mathbf{4_{rest}2_{C_j}}} := E\left[f_{\mathbf{4_{rest}}}(u_t, B_0) f_{\mathbf{2}_{C_j}}(u_t, B_0)'\right], \qquad \textit{for } j = 1, \ldots, n-1,$$

and

$$S_{\mathbf{4_{ii}2_M}} = \underbrace{\begin{bmatrix} & \omega_{iiii} - 1 & \omega_{(i+1)^4} - 1 & & 0 \\ 0_{(n-i)\times(i-1)} & \vdots & & \ddots & \\ & \omega_{iiii} - 1 & 0 & & \omega_{nnnn} - 1 \end{bmatrix}}_{(n-i)\times(n)}, \tag{5.C.22}$$

$$S_{\mathbf{4_{iii}2_M}} = \omega_{iii} \underbrace{\begin{bmatrix} \omega_{111} & & & 0 & 0 & & & \\ & \ddots & & & 0 & & 0_{(n-i)\times(n-i)} & \\ 0 & & & \omega_{(i-1)^3} & 0 & & & \\ & & & & 0 & \omega_{(i+1)^3} & & 0 \\ & 0_{(i-1)\times(i-1)} & & & 0 & & \ddots & \\ & & & & 0 & 0 & & \omega_{nnn} \end{bmatrix}}_{(n-1)\times(n)} \tag{5.C.23}$$

and

$$S_{\mathbf{4_{ii}2_{C_i}}} = \omega_{iii} \underbrace{\begin{bmatrix} \omega_{(i+1)^3} & & 0 \\ & \ddots & \\ 0 & & \omega_{nnn} \end{bmatrix}}_{(n-i)\times(n-i)}, \tag{5.C.24}$$

$$S_{\mathbf{4_{ii}2_{C_j}}} = 0_{(n-i)\times(n-j)}, \qquad \textit{for } i \neq j, \tag{5.C.25}$$

$$S_{\mathbf{4_{iii}2_{C_i}}} = \omega_{iiii} \underbrace{\begin{bmatrix} 0_{(i-1)\times(n-j)} \\ I_{(n-i)\times(n-j)} \end{bmatrix}}_{(n-1)\times(n-i)}, \tag{5.C.26}$$

$$S_{\mathbf{4_{iii}2_{C_j}}} = 0_{(n-1)\times(n-j)}, \qquad \textit{for } i < j, \tag{5.C.27}$$

$$S_{\mathbf{4_{iii}2_{C_j}}} = \omega_{iiii} \underbrace{\begin{bmatrix} & 0_{(j-1)\times(n-j)} & \\ 0_{1\times(i-j-1)} & 1 & 0_{1\times(n-i)} \\ & 0_{(n-j-1)\times(n-j)} & \end{bmatrix}}_{(n-1)\times(n-j)}, \qquad \textit{for } i > j, \tag{5.C.28}$$

and

$$S_{\mathbf{4_{rest}2}} = 0. \tag{5.C.29}$$

Proof. Equation (5.C.22): The (q,r)-th entry of $S_{\mathbf{4_{ii}2_M}}$ is equal to

$$S_{\mathbf{4_{ii}2_M}}(q,r) = E\left[(e(B_0)_i^2 e(B_0)_{i+q}^2 - 1)(e(B_0)_r^2 - 1)\right]$$

$$\overset{(*)}{=} \begin{cases} \omega_{iiii} - 1, & \text{if } r = i \\ \omega_{(i+q)^4} - 1, & \text{if } r = i+q \\ 0, & \text{else} \end{cases}.$$

Equation (5.C.23): The (q,r)-th entry of $S_{\mathbf{4_{iii}2_M}}$ with $q < i$ is equal to

$$S_{\mathbf{4_{iii}2_M}}(q,r) = E\left[(e(B_0)_i^3 e(B_0)_q)(e(B_0)_r^2 - 1)\right] \overset{(*)}{=} \begin{cases} \omega_{iii}\omega_{qqq}, & \text{if } r = q \\ 0, & \text{else} \end{cases}$$

and the (q,r)-th entry of $SS_{\mathbf{4_{iii}2_M}}$ with $q \geq i$ is equal to

$$S_{\mathbf{4_{iii}2_M}}(q,r) = E\left[(e(B_0)_i^3 e(B_0)_{q+1})(e(B_0)_r^2 - 1)\right] \overset{(*)}{=} \begin{cases} \omega_{iii}\omega_{(q+1)^3}, & \text{if } r = q+r \\ 0, & \text{else} \end{cases}.$$

Equation (5.C.24) and (5.C.25): The (q,r)-th entry of $S_{\mathbf{4_{ii}2_{C_j}}}$ is equal to

$$S_{\mathbf{4_{ii}2_{C_j}}}(q,r) = E\left[(e(B_0)_i^2 e(B_0)_{i+q}^2 - 1)(e(B_0)_j e(B_0)_{j+r})\right] \overset{(*)}{=} \begin{cases} 0, & \text{if } i \neq j \\ \omega_{iii}\omega_{(i+q)^3}, & \text{if } i = j, q = r \\ 0, & \text{if } i = j, q \neq r \end{cases}.$$

Equation (5.C.26), (5.C.27), and (5.C.28): The (q,r)-th entry of $S_{\mathbf{4_{iii}2_{C_j}}}$ with $q < i$ is equal to

$$S_{\mathbf{4_{iii}2_{C_j}}}(q,r) = E\left[(e(B_0)_i^3 e(B_0)_q)(e(B_0)_j e(B_0)_{j+r})\right] \overset{(*)}{=} \begin{cases} 0, & \text{if } i < j \\ \omega_{iiii}, & \text{if } i = j, q = j+r \ {}^{(***)} \\ 0, & \text{if } i = j, q \neq j+r \\ \omega_{iiii}, & \text{if } i > j, q = j \\ 0, & \text{if } i > j, q \neq j \end{cases}$$

and note that for $q < i$ the case $^{(***)}$ never occurs since $i > q = j + r = i + r$ implies $r < 0$. Moreover, the (q,r)-th entry of $S_{\mathbf{4_{iii}2_{C_j}}}$ with $q \geq i$ is equal to

$$S_{\mathbf{4_{iii}2_{C_j}}}(q,r) = E\left[(e(B_0)_i^3 e(B_0)_{q+1})(e(B_0)_j e(B_0)_{j+r})\right] \overset{(*)}{=} \begin{cases} 0, & \text{if } i < j \\ \omega_{iiii}, & \text{if } i = j, q+1 = j+r \\ 0, & \text{if } i = j, q+1 \neq j+r \\ \omega_{iiii}, & \text{if } i > j, q+1 = j \ {}^{(***)} \\ 0, & \text{if } i > j, q+1 \neq j \end{cases}$$

and note that for $q \geq i$ the case $^{(***)}$ never occurs since $q + 1 = j < i \leq q$ implies $1 < 0$.

Equation (5.C.29)) holds, since every moment condition in $E[f_{\mathbf{4_{rest}}}(u_t, B) f_{\mathbf{2}_{C_j}}(u_t, B_0)']$ can be written as

$$E\left[(e(B_0)_a e(B_0)_b e(B_0)_c)(e(B_0)_d e(B_0)_f)\right]$$

$a, b, c, d, f = \{1, \ldots, n\}$ and $a \neq b \neq c$, which implies
$E\left[(e(B_0)_a e(B_0)_b e(B_0)_c)(e(B_0)_d e(B_0)_f)\right] = 0$ by independence and mean zero of ϵ.

□

5.C.2 Preparation part 2

Lemma 5.C.8. *For b_{pq} and $b_{\tilde{p}\tilde{q}}$ with $p, q, \tilde{p}, \tilde{q} \in \{1, \ldots, n\}$ and $p \geq q$ and $\tilde{p} \geq \tilde{q}$ and $q \neq \tilde{q}$ it holds that*

$$(G_{\mathbf{2}}^{b_{pq}})' S_{\mathbf{2}}^{-1} G_{\mathbf{2}}^{b_{\tilde{p}\tilde{q}}} = 0.$$

Proof. For b_{pq} and $b_{\tilde{p}\tilde{q}}$ with $p, q, \tilde{p}, \tilde{q} \in \{1, \ldots, n\}$ and $p \geq q$ and $\tilde{p} \geq \tilde{q}$ and $q \neq \tilde{q}$ it holds that

$$(G_{\mathbf{2}}^{b_{pq}})' S_{\mathbf{2}}^{-1} G_{\mathbf{2}}^{b_{\tilde{p}\tilde{q}}} = (G_{\mathbf{2_M}}^{b_{pq}})' S_{\mathbf{2}}^{-1} G_{\mathbf{2_M}}^{b_{\tilde{p}\tilde{q}}} + \sum_{i=1}^{n-1} (G_{\mathbf{2_{C_i}}}^{b_{pq}})' S_{\mathbf{2}}^{-1} G_{\mathbf{2_{C_i}}}^{b_{\tilde{p}\tilde{q}}}.$$

The statement then follows by plugging in Equation (5.C.1), (5.C.2), and (5.C.3). □

With Lemma 5.C.8 it holds that $G_{\mathbf{2}}^{b_i\prime} S_{\mathbf{2}}^{-1} G_{\mathbf{2}}^{b_j} = 0$ for $i, j \in \{1, \ldots, n\}$ with $i \neq j$. Therefore, for any vector $b_i = [b_{ii}, \ldots, b_{ni}]$ representing the impact of the i-th structural shock $\epsilon_{i,t}$ it holds that $G_{\mathbf{2}}^{b_i\prime} S_{\mathbf{2}}^{-1} G_{\mathbf{2}}^{\neg b_i}$ is zero. Therefore, for any vector $b_i = [b_{ii}, \ldots, b_{ni}]$ the right hand side of Equation (5.A.4) is zero and hence,

$$G_{\mathbf{D}}^{b_i} - S_{\mathbf{D2}} S_{\mathbf{2}}^{-1} G_{\mathbf{2}}^{b_i} = 0. \tag{5.C.30}$$

The following Lemmata yield analytic expressions for $S_{\mathbf{D2}} S_{\mathbf{2}}^{-1} G_{\mathbf{2}}^{b_i}$ in Equation (5.C.30).

Lemma 5.C.9. *For $i, j = 1, \ldots, n$ it holds that*

$$S_{\mathbf{3_{ii}2_M}} S_{\mathbf{2_M}}^{-1} G_{\mathbf{2_M}}^{b_{j:,j}} = -2 \frac{1}{\omega_{jjjj} - 1} \underbrace{\begin{bmatrix} \omega_{111} & & 0 & & 0 & & & & \\ & \ddots & & & 0 & & 0_{(n-i)\times(n-i)} & & \\ 0 & & & \omega_{(i-1)^3} & 0 & & & & \\ & & & & 0 & \omega_{(i+1)^3} & & & 0 \\ & 0_{(i-1)\times(i-1)} & & & 0 & & & \ddots & \\ & & & & 0 & 0 & & & \omega_{nnn} \end{bmatrix}}_{(n-1)\times(n)}$$

$$\times \underbrace{\begin{bmatrix} & 0_{(j-1)\times(n-j+1)} \\ a_{jj} & 0_{1\times(n-j)} \\ & 0_{(n-j)\times(n-j+1)} \end{bmatrix}}_{n\times(n-j+1)}.$$

For $i = j$

$$S_{\mathbf{3_{ii}2_M}} S^{-1}_{\mathbf{2_M}} G^{b_{j:,j}}_{\mathbf{2_M}} = O_{n-1\times n-j+1}.$$

For $i < j$

$$S_{\mathbf{3_{ii}2_M}} S^{-1}_{\mathbf{2_M}} G^{b_{j:,j}}_{\mathbf{2_M}} = -\frac{2\omega_{jjj}}{\omega_{jjjj}-1} \underbrace{\begin{bmatrix} & 0_{(j-2)\times(n-j+1)} & \\ a_{jj} & 0_{1\times(n-j)} & \\ & 0_{(n-j)\times(n-j+1)} & \end{bmatrix}}_{(n-1)\times(n-j+1)}.$$

For $i > j$

$$S_{\mathbf{3_{ii}2_M}} S^{-1}_{\mathbf{2_M}} G^{b_{j:,j}}_{\mathbf{2_M}} = -\frac{2\omega_{jjj}}{\omega_{jjjj}-1} \underbrace{\begin{bmatrix} & 0_{(j-1)\times(n-j+1)} & \\ a_{jj} & 0_{1\times(n-j)} & \\ & 0_{(n-j-1)\times(n-j+1)} & \end{bmatrix}}_{(n-1)\times(n-j+1)}.$$

Furthermore, for $i = j$

$$S_{\mathbf{3_{ii}2_{C_i}}} G^{b_{i:,i}}_{\mathbf{2_{C_i}}} = -\omega_{iii} \underbrace{\begin{bmatrix} 0_{(i-1)\times(n-i)} \\ I_{(n-i)\times(n-i)} \end{bmatrix}}_{(n-1)\times(n-i)} \underbrace{\begin{bmatrix} a_{i+1,i} & a_{i+1,i+1} & & 0 \\ \vdots & & \ddots & \\ a_{ni} & a_{n,i+1} & \dots & a_{nn} \end{bmatrix}}_{(n-i)\times(n-i+1)}$$

$$= -\omega_{iii} \underbrace{\begin{bmatrix} & 0_{(i-1)\times(n-i+1)} & & \\ a_{i+1,i} & a_{i+1,i+1} & & 0 \\ \vdots & & \ddots & \\ a_{ni} & a_{n,i+1} & \dots & a_{nn} \end{bmatrix}}_{(n-1)\times(n-i+1)}.$$

For $i < j$

$$S_{\mathbf{3_{ii}2_{C_j}}} G^{b_{j:,j}}_{\mathbf{2_{C_j}}} = 0_{(n-1)\times(n-j)} \underbrace{\begin{bmatrix} a_{j+1,j} & a_{j+1,j+1} & & 0 \\ \vdots & & \ddots & \\ a_{nj} & a_{n,j+1} & \dots & a_{nn} \end{bmatrix}}_{(n-j)\times(n-j+1)} = 0_{(n-1)\times(n-j+1)}.$$

For $i > j$

$$S_{\mathbf{3_{ii}2_{C_j}}} G^{b_{j:,j}}_{\mathbf{2_{C_j}}} = -\omega_{iii} \underbrace{\begin{bmatrix} & 0_{(j-1)\times(n-j)} & \\ 0_{1\times(i-j-1)} & 1 & 0_{1\times(n-i)} \\ & 0_{(n-j-1)\times(n-j)} & \end{bmatrix}}_{(n-1)\times(n-j)} \underbrace{\begin{bmatrix} a_{j+1,j} & a_{j+1,j+1} & & 0 \\ \vdots & & \ddots & \\ a_{nj} & a_{n,j+1} & \dots & a_{nn} \end{bmatrix}}_{(n-j)\times(n-j+1)}$$

$$= -\omega_{iii} \underbrace{\begin{bmatrix} & 0_{(j-1)\times(n-j+1)} & \\ a_{ij} & \dots & a_{ii} \quad 0_{1\times(n-i)} \\ & 0_{(n-j-1)\times(n-j+1)} & \end{bmatrix}}_{(n-1)\times(n-j+1)}.$$

Proof. Follows from Lemma 5.C.2, Lemma 5.C.5, and Lemma 5.C.6 and simple matrix algebra. □

Lemma 5.C.10. *For $i = 1, \dots, n$ and $j = 1, \dots, n-1$ it holds that*

$$S_{\mathbf{4_{iii}2_M}} S^{-1}_{\mathbf{2_M}} G^{b_{j:,j}}_{\mathbf{2_M}} = -2\frac{\omega_{iii}}{\omega_{jjjj}-1} \underbrace{\begin{bmatrix} \omega_{111} & & 0 & 0 & & & \\ & \ddots & & 0 & & 0_{(n-i)\times(n-i)} & \\ 0 & & & \omega_{(i-1)^3} & 0 & & \\ & & & 0 & \omega_{(i+1)^3} & & 0 \\ & 0_{(i-1)\times(i-1)} & & 0 & & \ddots & \\ & & & 0 & 0 & & \omega_{nnn} \end{bmatrix}}_{(n-1)\times(n)}$$

$$\times \underbrace{\begin{bmatrix} & 0_{(j-1)\times(n-j+1)} \\ a_{jj} & 0_{1\times(n-j)} \\ & 0_{(n-j)\times(n-j+1)} \end{bmatrix}}_{n\times(n-j+1)}.$$

For $i = j$

$$S_{\mathbf{4_{iii}2_M}} S^{-1}_{\mathbf{2_M}} G^{b_{j:,j}}_{\mathbf{2_M}} = O_{n-1\times n-j+1}.$$

For $i < j$

$$S_{\mathbf{4_{iii}2_M}} S^{-1}_{\mathbf{2_M}} G^{b_{j:,j}}_{\mathbf{2_M}} = -\frac{2\omega_{jjj}\omega_{iii}}{\omega_{jjjj}-1} \underbrace{\begin{bmatrix} & 0_{(j-2)\times(n-j+1)} \\ a_{jj} & 0_{1\times(n-j)} \\ & 0_{(n-j)\times(n-j+1)} \end{bmatrix}}_{(n-1)\times(n-j+1)}.$$

For $i > j$

$$S_{\mathbf{4_{iii}2_M}} S^{-1}_{\mathbf{2_M}} G^{b_{j:,j}}_{\mathbf{2_M}} = -\frac{2\omega_{jjj}\omega_{iii}}{\omega_{jjjj}-1} \underbrace{\begin{bmatrix} a_{jj} & 0_{(j-1)\times(n-j+1)} \\ & 0_{1\times(n-j)} \\ & 0_{(n-j-1)\times(n-j+1)} \end{bmatrix}}_{(n-1)\times(n-j+1)}.$$

Furthermore, for $i = j$

$$S_{\mathbf{4_{iii}2_{C_i}}} G^{b_{i:,i}}_{\mathbf{2_{C_i}}} = -\omega_{iiii} \underbrace{\begin{bmatrix} 0_{(i-1)\times(n-i)} \\ I_{(n-i)\times(n-i)} \end{bmatrix}}_{(n-1)\times(n-i)} \underbrace{\begin{bmatrix} a_{i+1,i} & a_{i+1,i+1} & & 0 \\ \vdots & & \ddots & \\ a_{ni} & a_{n,i+1} & \dots & a_{nn} \end{bmatrix}}_{(n-i)\times(n-i+1)}$$

$$= -\omega_{iiii} \underbrace{\begin{bmatrix} & 0_{(i-1)\times(n-i+1)} & & \\ a_{i+1,i} & a_{i+1,i+1} & & 0 \\ \vdots & & \ddots & \\ a_{ni} & a_{n,i+1} & \dots & a_{nn} \end{bmatrix}}_{(n-1)\times(n-i+1)}.$$

For $i < j$

$$S_{\mathbf{4_{iii}2_{C_j}}} G^{b_{j:,j}}_{\mathbf{2_{C_j}}} = 0_{(n-1)\times(n-j)} \underbrace{\begin{bmatrix} a_{j+1,j} & a_{j+1,j+1} & & 0 \\ \vdots & & \ddots & \\ a_{nj} & a_{n,j+1} & \dots & a_{nn} \end{bmatrix}}_{(n-j)\times(n-j+1)} = 0_{(n-1)\times(n-j+1)}.$$

For $i > j$

$$S_{\mathbf{4_{iii}2_{C_j}}} G^{b_{j:,j}}_{\mathbf{2_{C_j}}} = -\omega_{iiii} \underbrace{\begin{bmatrix} & 0_{(j-1)\times(n-j)} & \\ 0_{1\times(i-j-1)} & 1 & 0_{1\times(n-i)} \\ & 0_{(n-j-1)\times(n-j)} & \end{bmatrix}}_{(n-1)\times(n-j)} \underbrace{\begin{bmatrix} a_{j+1,j} & a_{j+1,j+1} & & 0 \\ \vdots & & \ddots & \\ a_{nj} & a_{n,j+1} & \dots & a_{nn} \end{bmatrix}}_{(n-j)\times(n-j+1)}$$

$$= -\omega_{iiii} \underbrace{\begin{bmatrix} & 0_{(j-1)\times(n-j+1)} & & \\ a_{ij} & \dots & a_{ii} & 0_{1\times(n-i)} \\ & 0_{(n-j-1)\times(n-j+1)} & & \end{bmatrix}}_{(n-1)\times(n-j+1)}.$$

For $i = 1, \ldots, n-1$ *and* $j = 1, \ldots, n-1$

$$S_{\mathbf{4_{ii}2_M}} S^{-1}_{\mathbf{2_M}} G^{b_{j:,j}}_{\mathbf{2_M}} = -2 \frac{1}{\omega_{jjjj} - 1} \underbrace{\begin{bmatrix} & \omega_{iiii} - 1 & \omega_{(i+1)^4} - 1 & & 0 \\ 0_{(n-i)\times(i-1)} & \vdots & & \ddots & \\ & \omega_{iiii} - 1 & 0 & & \omega_{nnnn} - 1 \end{bmatrix}}_{(n-i)\times(n)}$$

$$\times \underbrace{\begin{bmatrix} & 0_{(j-1)\times(n-j+1)} & \\ a_{jj} & & 0_{1\times(n-j)} \\ & 0_{(n-j)\times(n-j+1)} & \end{bmatrix}}_{n\times(n-j+1)}.$$

For $i = j$

$$S_{\mathbf{4_{ii}2_M}} S^{-1}_{\mathbf{2_M}} G^{b_{j:,j}}_{\mathbf{2_M}} = -2 \frac{\omega_{iiii} - 1}{\omega_{iiii} - 1} \underbrace{\begin{bmatrix} a_{jj} & \\ & 0_{(n-i)\times(n-j)} \\ a_{jj} & \end{bmatrix}}_{(n-i)\times(n-j+1)}$$

$$= -2 \underbrace{\begin{bmatrix} a_{jj} & \\ & 0_{(n-i)\times(n-j)} \\ a_{jj} & \end{bmatrix}}_{(n-i)\times(n-j+1)}.$$

For $i < j$

$$S_{\mathbf{4_{ii}2_M}} S^{-1}_{\mathbf{2_M}} G^{b_{j:,j}}_{\mathbf{2_M}} = -2 \frac{\omega_{jjjj} - 1}{\omega_{jjjj} - 1} \underbrace{\begin{bmatrix} & 0_{(j-i-1)\times(n-j+1)} & \\ a_{jj} & & 0_{1\times(n-j)} \\ & 0_{(n-j)\times(n-j+1)} & \end{bmatrix}}_{(n-i)\times(n-j+1)}$$

$$= -2 \underbrace{\begin{bmatrix} & 0_{(j-i-1)\times(n-j+1)} & \\ a_{jj} & & 0_{1\times(n-j)} \\ & 0_{(n-j)\times(n-j+1)} & \end{bmatrix}}_{(n-i)\times(n-j+1)}.$$

For $i > j$

$$S_{\mathbf{4_{ii}2_M}} S^{-1}_{\mathbf{2_M}} G^{b_{j:,j}}_{\mathbf{2_M}} = 0_{(n-i)\times(n-j+1)}.$$

Furthermore,

$$S_{\mathbf{4_{ii}2_{C_i}}} G^{b_{i:,i}}_{\mathbf{2_{C_i}}} = -\omega_{iii} \underbrace{\begin{bmatrix} \omega_{(i+1)^3} & & 0 \\ & \ddots & \\ 0 & & \omega_{nnn} \end{bmatrix}}_{(n-i)\times(n-i)} \underbrace{\begin{bmatrix} a_{i+1,i} & a_{i+1,i+1} & & 0 \\ \vdots & & \ddots & \\ a_{ni} & a_{n,i+1} & \dots & a_{nn} \end{bmatrix}}_{(n-i)\times(n-i+1)}$$

$$= -\omega_{iii} \underbrace{\begin{bmatrix} a_{i+1,i}\omega_{(i+1)^3} & a_{i+1,i+1}\omega_{(i+1)^3} & & 0 \\ \vdots & & \ddots & \\ a_{ni}\omega_{nnn} & a_{n,i+1}\omega_{nnn} & \dots & a_{nn}\omega_{nnn} \end{bmatrix}}_{(n-i)\times(n-i+1)}.$$

Proof. Follows from Lemma 5.C.2, Lemma 5.C.5, and Lemma 5.C.7 and simple matrix algebra. □

Lemma 5.C.11. *For $i, j = 1, \dots, n$ and $i = j$*

$$G^{b_{j:,j}}_{\mathbf{3_{ii}}} - S_{\mathbf{3_{ii}2}} S^{-1}_{\mathbf{2}} G^{b_{j:,j}}_{\mathbf{2}} = -\omega_{iii} \underbrace{\begin{bmatrix} & 0_{(i-1)\times(n-i+1)} & & \\ a_{i+1,i} & a_{i+1,i+1} & & 0 \\ \vdots & & \ddots & \\ a_{ni} & a_{n,i+1} & \dots & a_{nn} \end{bmatrix}}_{(n-1)\times(n-i+1)}$$

$$+\omega_{iii} \underbrace{\begin{bmatrix} & 0_{(i-1)\times(n-i+1)} & & \\ a_{i+1,i} & a_{i+1,i+1} & & 0 \\ \vdots & & \ddots & \\ a_{ni} & a_{n,i+1} & \dots & a_{nn} \end{bmatrix}}_{(n-1)\times(n-i+1)}$$

$$= \ 0_{(n-1)\times(n-i+1)}.$$

For $i, j = 1, \dots, n$ and $i < j$

$$G^{b_{j:,j}}_{\mathbf{3_{ii}}} - S_{\mathbf{3_{ii}2}} S^{-1}_{\mathbf{2}} G^{b_{j:,j}}_{\mathbf{2}} = \frac{2\omega_{jjj}}{\omega_{jjjj} - 1} \underbrace{\begin{bmatrix} & 0_{(j-2)\times(n-j+1)} & \\ a_{jj} & 0_{1\times(n-j)} & \\ & 0_{(n-j)\times(n-j+1)} & \end{bmatrix}}_{(n-1)\times(n-j+1)}.$$

For $i, j = 1, \ldots, n$ and $i > j$

$$G^{b_{j:,j}}_{\mathbf{3_{ii}}} - S_{\mathbf{3_{ii}2}} S^{-1}_{\mathbf{2}} G^{b_{j:,j}}_{\mathbf{2}} = \frac{2\omega_{jjj}}{\omega_{jjjj} - 1} \underbrace{\begin{bmatrix} & 0_{(j-1)\times(n-j+1)} & \\ a_{jj} & & 0_{1\times(n-j)} \\ & 0_{(n-j-1)\times(n-j+1)} & \end{bmatrix}}_{(n-1)\times(n-j+1)}$$

$$+ \omega_{iii} \underbrace{\begin{bmatrix} & 0_{(j-1)\times(n-j+1)} & & \\ a_{ij} & \cdots & a_{ii} & 0_{1\times(n-i)} \\ & 0_{(n-j-1)\times(n-j+1)} & & \end{bmatrix}}_{(n-1)\times(n-j+1)}.$$

Proof. For $i, j = 1, \ldots, n$ let

$$W^{b_{j:,j}}_{\mathbf{3_{ii}}} := S_{\mathbf{3_{ii}2}} S^{-1}_{\mathbf{2}} G^{b_{j:,j}}_{\mathbf{2}}.$$

Then, for $i = 1, \ldots, n-1$

$$W^{b_{i:,i}}_{\mathbf{3_{ii}}} = -\omega_{iii} \underbrace{\begin{bmatrix} & 0_{(i-1)\times(n-i+1)} & & \\ a_{i+1,i} & a_{i+1,i+1} & & 0 \\ \vdots & & \ddots & \\ a_{ni} & a_{n,i+1} & \cdots & a_{nn} \end{bmatrix}}_{(n-1)\times(n-i+1)},$$

for $i = n$

$$W^{b_{i:,i}}_{\mathbf{3_{ii}}} = 0_{(n-1)\times 1},$$

for $i < j$

$$W^{b_{j:,j}}_{\mathbf{3_{ii}}} = -\frac{2\omega_{jjj}}{\omega_{jjjj} - 1} \underbrace{\begin{bmatrix} & 0_{(j-2)\times(n-j+1)} & \\ a_{jj} & & 0_{1\times(n-j)} \\ & 0_{(n-j)\times(n-j+1)} & \end{bmatrix}}_{(n-1)\times(n-j+1)},$$

for $i > j$

$$W^{b_{j:,j}}_{\mathbf{3_{ii}}} = -\frac{2\omega_{jjj}}{\omega_{jjjj} - 1} \underbrace{\begin{bmatrix} & 0_{(j-1)\times(n-j+1)} & \\ a_{jj} & & 0_{1\times(n-j)} \\ & 0_{(n-j-1)\times(n-j+1)} & \end{bmatrix}}_{(n-1)\times(n-j+1)}$$

$$- \omega_{iii} \underbrace{\begin{bmatrix} & 0_{(j-1)\times(n-j+1)} & & \\ a_{ij} & \cdots & a_{ii} & 0_{1\times(n-i)} \\ & 0_{(n-j-1)\times(n-j+1)} & & \end{bmatrix}}_{(n-1)\times(n-j+1)}.$$

Moreover, it holds that

$$W_{\mathbf{3_{ii}}}^{b_{j:},j} = \begin{bmatrix} S_{\mathbf{3_{ii}2_M}} & S_{\mathbf{3_{ii}2_{C_1}}} & \cdots & S_{\mathbf{3_{ii}2_{C_{n-1}}}} \end{bmatrix} \begin{bmatrix} S_{\mathbf{2_M}} & 0 \\ 0 & I_{n(n-1)\times n(n-1)} \end{bmatrix}^{-1}$$

$$\times \left[G_{\mathbf{2_M}}^{b_{j:},j}, G_{\mathbf{2_{C_1}}}^{b_{j:},j}, \ldots, G_{\mathbf{2_{C_{n-1}}}}^{b_{j:},j} \right]'$$

$$= S_{\mathbf{3_{ii}2_M}} S_{\mathbf{2_M}}^{-1} G_{\mathbf{2_M}}^{b_{j:},j} + \sum_{q=1}^{n-1} S_{\mathbf{3_{ii}2_{C_q}}} G_{\mathbf{2_{C_q}}}^{b_{j:},j}.$$

From Lemma 5.C.2 it follows that $G_{\mathbf{2_{C_q}}}^{b_{j:},j} = 0$ for $i \neq j$ and, hence, for $i = 1, \ldots, n$, $j = 1, \ldots, n-1$

$$W_{\mathbf{3_{ii}}}^{b_{j:},j} = S_{\mathbf{3_{ii}2_M}} S_{\mathbf{2_M}}^{-1} G_{\mathbf{2_M}}^{b_{j:},j} + S_{\mathbf{3_{ii}2_{C_j}}} G_{\mathbf{2_{C_j}}}^{b_{j:},j}$$

and for $j = n$

$$W_{\mathbf{3_{ii}}}^{b_{j:},j} = S_{\mathbf{3_{ii}2_M}} S_{\mathbf{2_M}}^{-1} G_{\mathbf{2_M}}^{b_{j:},j}.$$

With Lemma 5.C.9 (implying $S_{\mathbf{3_{ii}2_M}} S_{\mathbf{2_M}}^{-1} G_{\mathbf{2_M}}^{b_{i:},i} = 0$ and $S_{\mathbf{3_{ii}2_{C_j}}} G_{\mathbf{2_{C_j}}}^{b_{j:},j} = 0$ for $i < j$) it follows that

$$W_{\mathbf{3_{ii}}}^{b_{i:},i} = S_{\mathbf{3_{ii}2_{C_i}}} G_{\mathbf{2_{C_i}}}^{b_{i:},i}, \qquad \text{for } i < n,$$

$$W_{\mathbf{3_{ii}}}^{b_{i:},i} = 0_{(n-1)\times 1}, \qquad \text{for } i = n,$$

$$W_{\mathbf{3_{ii}}}^{b_{j:},j} = S_{\mathbf{3_{ii}2_M}} S_{\mathbf{2_M}}^{-1} G_{\mathbf{2_M}}^{b_{j:},j}, \qquad \text{for } i < j,$$

$$W_{\mathbf{3_{ii}}}^{b_{j:},j} = S_{\mathbf{3_{ii}2_M}} S_{\mathbf{2_M}}^{-1} G_{\mathbf{2_M}}^{b_{j:},j} + S_{\mathbf{3_{ii}2_{C_j}}} G_{\mathbf{2_{C_j}}}^{b_{j:},j}, \qquad \text{for } i > j.$$

The statements then follow with Lemma 5.C.9 and Lemma 5.C.3.

□

Lemma 5.C.12. *For $i, j = 1, \ldots, n-1$ and $i = j$*

$$G_{\mathbf{4_{ii}}}^{b_{j:},j} - S_{\mathbf{4_{ii}2}} S_{\mathbf{2}}^{-1} G_{\mathbf{2}}^{b_{j:},j} = -2 \underbrace{\begin{bmatrix} a_{ii} & \\ \vdots & 0_{(n-i)\times(n-j)} \\ a_{ii} & \end{bmatrix}}_{(n-i)\times(n-i+1)} + 2 \underbrace{\begin{bmatrix} a_{ii} & \\ \vdots & 0_{(n-i)\times(n-i)} \\ a_{ii} & \end{bmatrix}}_{(n-i)\times(n-i+1)}$$

$$+\omega_{iii}\underbrace{\begin{bmatrix} a_{i+1,i}\omega_{(i+1)^3} & a_{i+1,i+1}\omega_{(i+1)^3} & & 0 \\ \vdots & & \ddots & \\ a_{ni}\omega_{nnn} & a_{n,i+1}\omega_{nnn} & \dots & a_{nn}\omega_{nnn} \end{bmatrix}}_{(n-i)\times(n-i+1)}$$

$$=\omega_{iii}\underbrace{\begin{bmatrix} a_{i+1,i}\omega_{(i+1)^3} & a_{i+1,i+1}\omega_{(i+1)^3} & & 0 \\ \vdots & & \ddots & \\ a_{ni}\omega_{nnn} & a_{n,i+1}\omega_{nnn} & \dots & a_{nn}\omega_{nnn} \end{bmatrix}}_{(n-i)\times(n-i+1)}.$$

For $i=1,\dots,n-1$, $j=1,\dots,n$, and $i>j$

$$G^{b_{j:,j}}_{\mathbf{4_{ii}}} - S_{\mathbf{4_{ii}2}}S^{-1}_{\mathbf{2}}G^{b_{j:,j}}_{\mathbf{2}} = 0_{(n-i)\times(n-j+1)} - 0_{(n-i)\times(n-j+1)}$$

$$= 0_{(n-i)\times(n-j+1)}.$$

For $i=1,\dots,n-1$, $j=1,\dots,n$, and $i<j$

$$G^{b_{j:,j}}_{\mathbf{4_{ii}}} - S_{\mathbf{4_{ii}2}}S^{-1}_{\mathbf{2}}G^{b_{j:,j}}_{\mathbf{2}} = -2\underbrace{\begin{bmatrix} & 0_{(j-i-1)\times(n-j+1)} & \\ a_{jj} & 0_{1\times(n-j)} & \\ & 0_{(n-j)\times(n-j+1)} & \end{bmatrix}}_{(n-i)\times(n-j+1)} + 2\underbrace{\begin{bmatrix} & 0_{(j-i-1)\times(n-j+1)} & \\ a_{jj} & 0_{1\times(n-j)} & \\ & 0_{(n-j)\times(n-j+1)} & \end{bmatrix}}_{(n-i)\times(n-j+1)}$$

$$= 0_{(n-i)\times(n-j+1)}.$$

Proof. For $i=1,\dots,n-1$, $j=1,\dots,n$ let

$$W^{b_{j:,j}}_{\mathbf{4_{ii}}} := S_{\mathbf{4_{ii}2}}S^{-1}_{\mathbf{2}}G^{b_{j:,j}}_{\mathbf{2}}.$$

Then, for $i=1,\dots,n-1$

$$W^{b_{i:,i}}_{\mathbf{4_{ii}}} = -2\underbrace{\begin{bmatrix} a_{ii} & \\ & 0_{(n-i)\times(n-i)} \\ a_{ii} & \end{bmatrix}}_{(n-i)\times(n-i+1)} - \omega_{iii}\underbrace{\begin{bmatrix} a_{i+1,i}\omega_{(i+1)^3} & a_{i+1,i+1}\omega_{(i+1)^3} & & 0 \\ \vdots & & \ddots & \\ a_{ni}\omega_{nnn} & a_{n,i+1}\omega_{nnn} & \dots & a_{nn}\omega_{nnn} \end{bmatrix}}_{(n-i)\times(n-i+1)},$$

for $i<j$

$$W^{b_{j:,j}}_{\mathbf{4_{ii}}} = -2\underbrace{\begin{bmatrix} & 0_{(j-i-1)\times(n-j+1)} & \\ a_{jj} & 0_{1\times(n-j)} & \\ & 0_{(n-j)\times(n-j+1)} & \end{bmatrix}}_{(n-i)\times(n-j+1)},$$

and for $i > j$

$$W_{\mathbf{4_{ii}}}^{b_{j:},j} = 0_{(n-i)\times(n-j+1)}.$$

Moreover, it holds that

$$\begin{aligned} W_{\mathbf{4_{ii}}}^{b_{j:},j} =& \begin{bmatrix} S_{\mathbf{4_{ii}2_M}} & S_{\mathbf{4_{ii}2_{C_1}}} & \cdots & S_{\mathbf{4_{ii}2_{C_{n-1}}}} \end{bmatrix} \begin{bmatrix} S_{\mathbf{2_M}} & 0 \\ 0 & I_{n(n-1)\times n(n-1)} \end{bmatrix}^{-1} \\ & \times \left[G_{\mathbf{2_M}}^{b_{j:},j}, G_{\mathbf{2_{C_1}}}^{b_{j:},j}, \ldots, G_{\mathbf{2_{C_{n-1}}}}^{b_{j:},j} \right]' \\ =& S_{\mathbf{4_{ii}2_M}} S_{\mathbf{2_M}}^{-1} G_{\mathbf{2_M}}^{b_{j:},j} + \sum_{q=1}^{n-1} S_{\mathbf{4_{ii}2_{C_q}}} G_{\mathbf{2_{C_q}}}^{b_{j:},j}. \end{aligned}$$

From Lemma 5.C.2 it follows that $G_{\mathbf{2_{C_q}}}^{b_{j:},j} = 0$ for $i \neq j$ and, hence, for $i = 1, \ldots, n$, $j = 1, \ldots, n-1$

$$W_{\mathbf{4_{ii}}}^{b_{j:},j} = S_{\mathbf{4_{ii}2_M}} S_{\mathbf{2_M}}^{-1} G_{\mathbf{2_M}}^{b_{j:},j} + S_{\mathbf{4_{ii}2_{C_j}}} G_{\mathbf{2_{C_j}}}^{b_{j:},j}$$

and for $j = n$

$$W_{\mathbf{4_{ii}}}^{b_{j:},j} = S_{\mathbf{4_{ii}2_M}} S_{\mathbf{2_M}}^{-1} G_{\mathbf{2_M}}^{b_{j:},j}.$$

From Lemma 5.C.7 it follows that $S_{\mathbf{4_{ii}2_{C_j}}} = 0$ for $i \neq j$ and, hence,

$$W_{\mathbf{4_{ii}}}^{b_{i:},i} = S_{\mathbf{4_{ii}2_M}} S_{\mathbf{2_M}}^{-1} G_{\mathbf{2_M}}^{b_{i:},i} + S_{\mathbf{4_{ii}2_{C_i}}} G_{\mathbf{2_{C_i}}}^{b_{i:},i}$$

$$W_{\mathbf{4_{ii}}}^{b_{j:},j} = S_{\mathbf{4_{ii}2_M}} S_{\mathbf{2_M}}^{-1} G_{\mathbf{2_M}}^{b_{j:},j}, \qquad \text{for } i \neq j.$$

The statements then follow with Lemma 5.C.10 and Lemma 5.C.4.

□

Lemma 5.C.13. *For $i, j = 1, \ldots, n$ and $i = j$*

$$G_{\mathbf{4_{iii}}}^{b_{j:},j} - S_{\mathbf{4_{iii}2}} S_{\mathbf{2}}^{-1} G_{\mathbf{2}}^{b_{j:},j} = -\omega_{iiii} \underbrace{\begin{bmatrix} & 0_{(i-1)\times(n-i+1)} & \\ a_{i+1,i} & a_{i+1,i+1} & 0 \\ \vdots & & \ddots \\ a_{ni} & \cdots & a_{nn} \end{bmatrix}}_{(n-1)\times(n-i+1)}$$

$$+\,\omega_{iiii}\underbrace{\begin{bmatrix} & 0_{(i-1)\times(n-i+1)} & & \\ a_{i+1,i} & a_{i+1,i+1} & & 0 \\ \vdots & & \ddots & \\ a_{ni} & a_{n,i+1} & \dots & a_{nn} \end{bmatrix}}_{(n-1)\times(n-i+1)}$$

$$= \; 0_{(n-1)\times(n-i+1)}.$$

For $i,j=1,\dots,n$ *and* $i>j$

$$G^{b_{j:,j}}_{\mathbf{4_{iii}}} - S_{\mathbf{4_{iii}2}}S^{-1}_{\mathbf{2}}G^{b_{j:,j}}_{\mathbf{2}} = -\,3\underbrace{\begin{bmatrix} & 0_{(j-1)\times(n-j+1)} & & \\ a_{ij} & \dots & a_{ii} & 0_{1\times(n-i)} \\ & 0_{(n-j-1)\times(n-j+1)} & & \end{bmatrix}}_{(n-1)\times(n-j+1)}$$

$$+\,\frac{2\omega_{jjj}\omega_{iii}}{\omega_{jjjj}-1}\underbrace{\begin{bmatrix} a_{jj} & 0_{(j-1)\times(n-j+1)} \\ & 0_{1\times(n-j)} \\ & 0_{(n-j-1)\times(n-j+1)} \end{bmatrix}}_{(n-1)\times(n-j+1)}$$

$$+\,\omega_{iiii}\underbrace{\begin{bmatrix} & 0_{(j-1)\times(n-j+1)} & & \\ a_{ij} & \dots & a_{ii} & 0_{1\times(n-i)} \\ & 0_{(n-j-1)\times(n-j+1)} & & \end{bmatrix}}_{(n-1)\times(n-j+1)}$$

$$= \; (\omega_{iiii}-3)\underbrace{\begin{bmatrix} & 0_{(j-1)\times(n-j+1)} & & \\ a_{ij} & \dots & a_{ii} & 0_{1\times(n-i)} \\ & 0_{(n-j-1)\times(n-j+1)} & & \end{bmatrix}}_{(n-1)\times(n-j+1)}$$

$$+\,\frac{2\omega_{jjj}\omega_{iii}}{\omega_{jjjj}-1}\underbrace{\begin{bmatrix} a_{jj} & 0_{(j-1)\times(n-j+1)} \\ & 0_{1\times(n-j)} \\ & 0_{(n-j-1)\times(n-j+1)} \end{bmatrix}}_{(n-1)\times(n-j+1)}.$$

For $i,j=1,\dots,n$ *and* $i<j$

$$G^{b_{j:,j}}_{\mathbf{4_{iii}}} - S_{\mathbf{4_{iii}2}}S^{-1}_{\mathbf{2}}G^{b_{j:,j}}_{\mathbf{2}} = 0_{(n-1)\times(n-j+1)} + \frac{2\omega_{jjj}\omega_{iii}}{\omega_{jjjj}-1}\underbrace{\begin{bmatrix} & 0_{(j-2)\times(n-j+1)} \\ a_{jj} & 0_{1\times(n-j)} \\ & 0_{(n-j)\times(n-j+1)} \end{bmatrix}}_{(n-1)\times(n-j+1)}$$

$$= \frac{2\omega_{jjj}\omega_{iii}}{\omega_{jjjj}-1} \underbrace{\begin{bmatrix} & 0_{(j-2)\times(n-j+1)} & \\ a_{jj} & 0_{1\times(n-j)} & \\ & 0_{(n-j)\times(n-j+1)} & \end{bmatrix}}_{(n-1)\times(n-j+1)}.$$

Proof. For $i, j = 1, \ldots, n$ let

$$W_{\mathbf{4_{iii}}}^{b_{j:,j}} := S_{\mathbf{4_{iii}2}} S_{\mathbf{2}}^{-1} G_{\mathbf{2}}^{b_{j:,j}}.$$

Then, for $i = 1, \ldots, n-1$

$$W_{\mathbf{4_{iii}}}^{b_{i:,i}} = -\omega_{iiii} \underbrace{\begin{bmatrix} & 0_{(i-1)\times(n-i+1)} & & \\ a_{i+1,i} & a_{i+1,i+1} & & 0 \\ \vdots & & \ddots & \\ a_{ni} & a_{n,i+1} & \cdots & a_{nn} \end{bmatrix}}_{(n-1)\times(n-i+1)}.$$

For $i = n$

$$W_{\mathbf{4_{iii}}}^{b_{j:,j}} = 0_{(n-1)\times 1}.$$

For $i < j$

$$W_{\mathbf{4_{iii}}}^{b_{j:,j}} = -\frac{2\omega_{jjj}\omega_{iii}}{\omega_{jjjj}-1} \underbrace{\begin{bmatrix} & 0_{(j-2)\times(n-j+1)} & \\ a_{jj} & 0_{1\times(n-j)} & \\ & 0_{(n-j)\times(n-j+1)} & \end{bmatrix}}_{(n-1)\times(n-j+1)}.$$

For $i > j$

$$W_{\mathbf{4_{iii}}}^{b_{j:,j}} = -\frac{2\omega_{jjj}\omega_{iii}}{\omega_{jjjj}-1} \underbrace{\begin{bmatrix} & 0_{(j-1)\times(n-j+1)} & \\ a_{jj} & 0_{1\times(n-j)} & \\ & 0_{(n-j-1)\times(n-j+1)} & \end{bmatrix}}_{(n-1)\times(n-j+1)}$$

$$- \omega_{iiii} \underbrace{\begin{bmatrix} & 0_{(j-1)\times(n-j+1)} & & \\ a_{ij} & \cdots & a_{ii} & 0_{1\times(n-i)} \\ & 0_{(n-j-1)\times(n-j+1)} & & \end{bmatrix}}_{(n-1)\times(n-j+1)}.$$

Moreover, it holds that

$$W_{\mathbf{4_{iii}}}^{b_{j:,j}} = \begin{bmatrix} S_{\mathbf{4_{iii}2_M}} & S_{\mathbf{4_{iii}2_{C_1}}} & \cdots & S_{\mathbf{4_{iii}2_{C_{n-1}}}} \end{bmatrix} \begin{bmatrix} S_{\mathbf{2_M}} & 0 \\ 0 & I_{n(n-1)\times n(n-1)} \end{bmatrix}^{-1}$$

$$\left[G^{b_{j:,j}}_{\mathbf{2_M}}, G^{b_{j:,j}}_{\mathbf{2_{C_1}}}, \ldots, G^{b_{j:,j}}_{\mathbf{2_{C_{n-1}}}}\right]'$$

$$= S_{\mathbf{4_{iii}2_M}} S^{-1}_{\mathbf{2_M}} G^{b_{j:,j}}_{\mathbf{2_M}} + \sum_{q=1}^{n-1} S_{\mathbf{4_{iii}2_{C_q}}} G^{b_{j:,j}}_{\mathbf{2_{C_q}}}.$$

From Lemma 5.C.2 it follows that $G^{b_{j:,j}}_{\mathbf{2_{C_q}}} = 0$ for $i \neq j$ and, hence, for $i = 1, \ldots, n$, $j = 1, \ldots, n-1$

$$W^{b_{j:,j}}_{\mathbf{4_{iii}}} = S_{\mathbf{4_{iii}2_M}} S^{-1}_{\mathbf{2_M}} G^{b_{j:,j}}_{\mathbf{2_M}} + S_{\mathbf{4_{iii}2_{C_j}}} G^{b_{j:,j}}_{\mathbf{2_{C_j}}}$$

and for $j = n$

$$W^{b_{j:,j}}_{\mathbf{4_{iii}}} = S_{\mathbf{4_{iii}2_M}} S^{-1}_{\mathbf{2_M}} G^{b_{j:,j}}_{\mathbf{2_M}}.$$

With Lemma 5.C.10 it follows that

$$W^{b_{i:,i}}_{\mathbf{4_{iii}}} = S_{\mathbf{4_{iii}2_{C_i}}} G^{b_{i:,i}}_{\mathbf{2_{C_i}}}, \qquad \text{for } i < n,$$

$$W^{b_{i:,i}}_{\mathbf{4_{iii}}} = 0_{(n-1)\times 1}, \qquad \text{for } i = n,$$

$$W^{b_{j:,j}}_{\mathbf{4_{iii}}} = S_{\mathbf{4_{iii}2_M}} S^{-1}_{\mathbf{2_M}} G^{b_{j:,j}}_{\mathbf{2_M}}, \qquad \text{for } i < j,$$

$$W^{b_{j:,j}}_{\mathbf{4_{iii}}} = S_{\mathbf{4_{iii}2_M}} S^{-1}_{\mathbf{2_M}} G^{b_{j:,j}}_{\mathbf{2_M}} + S_{\mathbf{4_{iii}2_{C_j}}} G^{b_{j:,j}}_{\mathbf{2_{C_j}}}, \qquad \text{for } i > j.$$

The statements then follow with Lemma 5.C.10 and Lemma 5.C.4.

□

5.C.3 Final Lemma

We now combine the conditions in Lemma 5.C.11 - 5.C.13 into conditions for specific moment conditions.

Lemma 5.C.14. *In a recursive SVAR with independent shocks, it holds that for $i, j, k, l \in \{1, \ldots, n\}$*

1. *coskewness moment condition $E[f_{\mathbf{D}}(B, u_t)] = E[e(B)_i e(B)_j e(B)_k]$ with $i \neq j \neq k$ satisfy*

$$G^b_{\mathbf{D}} - S_{\mathbf{D2}} S^{-1}_{\mathbf{2}} G^b_{\mathbf{2}} = 0$$

   for every unrestricted element b of B_0.

2. *coskewness moment condition $E[f_{\mathbf{D}}(B, u_t)] = E[e(B)_i^2 e(B)_j]$ with $i \neq j$ satisfy*

$$G^{b_{pq}}_{\mathbf{D}} - S_{\mathbf{D2}}S^{-1}_{\mathbf{2}}G^{b_{pq}}_{\mathbf{2}} = \begin{cases} \frac{2E[\epsilon^3_{j,t}]}{E[\epsilon^4_{j,t}]-1}a_{jp}, & \text{if } p=j,\ q=j,\ i<j, \\ \frac{2E[\epsilon^3_{j,t}]}{E[\epsilon^4_{j,t}]-1}a_{jp} + E[\epsilon^3_{i,t}]a_{ip}, & \text{if } q=j,\ p=j,\ i>j, \\ E[\epsilon^3_{i,t}]a_{ip}, & \text{if } q=j,\ p=j+1,\dots,i,\ i>j, \\ 0, & \text{else} \end{cases}.$$

3. *cokurtosis conditions* $E[f_{\mathbf{D}}(B,u_t)] = E[e(B_0)_i e(B_0)_j e(B_0)_k e(B_0)_l]$ *and* $E[f_{\mathbf{D}}(B,u_t)] = E[e(B_0)^2_i e(B_0)_j e(B_0)_k]$ *with* $i \neq j \neq k \neq l$ *satisfy*

$$G^{b}_{\mathbf{D}} - S_{\mathbf{D2}}S^{-1}_{\mathbf{2}}G^{b}_{\mathbf{2}} = 0$$

for every unrestricted element b *of* B_0.

4. *cokurtosis conditions* $E[f_{\mathbf{D}}(B,u_t)] = E[e(B_0)^2_i e(B_0)^2_j - 1]$ *with* $i \neq j$ *satisfy*

$$G^{b_{pq}}_{\mathbf{D}} - S_{\mathbf{D2}}S^{-1}_{\mathbf{2}}G^{b_{pq}}_{\mathbf{2}} = \begin{cases} E[\epsilon^3_{i,t}]E[\epsilon^3_{j,t}]a_{jp}, & \text{if } q=i,\ p=i,\dots,j \\ 0, & \text{else} \end{cases}.$$

5. *cokurtosis conditions* $E[f_{\mathbf{D}}(B,u_t)] = E[e(B_0)^3_i e(B_0)_j]$ *with* $i \neq j$ *satisfy*

$$G^{b_{pq}}_{\mathbf{D}} - S_{\mathbf{D2}}S^{-1}_{\mathbf{2}}G^{b_{pq}}_{\mathbf{2}} = \begin{cases} \frac{2E[\epsilon^3_{j,t}]E[\epsilon^3_{i,t}]}{E[\epsilon^4_{j,t}]-1}a_{jp}, & \text{if } p=j,\ q=j, \quad i<j, \\ \frac{2E[\epsilon^3_{j,t}]E[\epsilon^3_{i,t}]}{E[\epsilon^4_{j,t}]-1}a_{jp} + (E[\epsilon^4_{i,t}]-3)a_{ip}, & \text{if } q=j,\ p=j,\ i>j, \\ (E[\epsilon^4_{i,t}]-3)a_{ip}, & \text{if } q=j,\ p=j+1,\dots,i,\ i>j, \\ 0, & \text{else} \end{cases}.$$

Proof. The statements directly follow from Lemma 5.C.11 - 5.C.13. □

6 Conclusion

This thesis applies modern shrinkage methods in the context of discrete choice models and time series models, identifies shortcomings of existing estimators, and proposes methods for their improvement. Chapter 2 deals with a special case of the LASSO regression and elastic net regression in a discrete choice framework which models unobserved heterogeneity with a nonparametric approach. Building on the estimators of Chapter 2, Chapter 3 considers a random elastic net estimator. Chapter 4 estimates flexible forms of observed heterogeneity in discrete choice models using neural networks. To this end, an influence function approach is applied together with ℓ_2-regularization, which shrinks the weights of the neural network towards zero. Chapter 5 applies a LASSO-Type GMM estimator to select valid and relevant moment conditions in a SVAR model in a data-driven way.

This thesis also reveals that there are open questions with respect to the studied approaches to be solved in the future. For example, Chapter 2 relates the FKRB estimator to the LASSO estimator, implying that it might be challenging to construct a valid inference procedure for this estimator and its extensions. Neither Chapter 2 nor Chapter 3 develop an inference procedure for the proposed estimators, which, however, would be important for applications. Furthermore, the estimator presented in Chapter 2 does not allow for high-dimensional random coefficients. In principle, the estimator developed in Chapter 3 can deal with high-dimensional random coefficients. However, in this case it would be interesting to explore ways to reduce computation time when including a larger number of random coefficients. The estimation of the neural network in Chapter 4 involves many tuning parameters. Further simulations could help to guide the choice of those tuning parameters and shed light on the robustness of the applied influence function method. In theory, the LASSO-type GMM estimator in Chapter 5 is not distorted by invalid moment conditions. Further simulations could illustrate this property. Additionally, developing a shrinkage estimator which can identify potential zero restrictions on the interactions of the shocks of the SVAR from the data could complement and further justify the restrictions imposed by ecomonic theory.

References

Andrews, D. W. (1999), 'Consistent Moment Selection Procedures for Generalized Method of Moments Estimation', *Econometrica* **67**(3), 543–563.

Antonini, G., Gioia, C. and Frejinger, E. (2007), 'Swissmetro: description of the data'.

Bansal, P., Daziano, R. A. and Achtnicht, M. (2018), 'Extending the logit-mixed logit model for a combination of random and fixed parameters', *Journal of choice modelling* **27**, 88–96.

Baumeister, C. and Hamilton, J. D. (2019), 'Structural Interpretation of Vector Autoregressions with Incomplete Identification: Revisiting the Role of Oil Supply and Demand Shocks', *American Economic Review* **109**(5), 1873–1910.

Baumeister, C. and Kilian, L. (2016), 'Understanding the Decline in the Price of Oil since June 2014', *Journal of the Association of Environmental and resource economists* **3**(1), 131–158.

Bhat, C. R. (1995), 'A heteroscedastic extreme value model of intercity travel mode choice', *Transportation Research Part B: Methodological* **29**(6), 471–483.

Bhat, C. R. (1997*a*), 'An endogenous segmentation mode choice model with an application to intercity travel', *Transportation science* **31**(1), 34–48.

Bhat, C. R. (1997*b*), 'Covariance heterogeneity in nested logit models: econometric structure and application to intercity travel', *Transportation Research Part B: Methodological* **31**(1), 11–21.

Bhat, C. R. (1998), 'Accommodating variations in responsiveness to level-of-service measures in travel mode choice modeling', *Transportation Research Part A: Policy and Practice* **32**(7), 495–507.

Bierlaire, M., Axhausen, K. and Abay, G. (2001), The acceptance of modal innovation: The case of swissmetro.

Björn, N., Badam, T. V. S., Spalinskas, R., Brandén, E., Koyi, H., Lewensohn, R., De Petris, L., Lubovac-Pilav, Z., Sahlén, P., Lundeberg, J. et al. (2020), 'Whole-genome sequencing and gene network modules predict gemcitabine/carboplatin-induced myelosuppression in non-small cell lung cancer patients', *NPJ systems biology and applications* **6**(1), 1–15.

Blanchard, O. J. (1989), 'A Traditional Interpretation of Macroeconomic Fluctuations', *The American Economic Review* pp. 1146–1164.

Blanchard, O. J. and Quah, D. (1989), 'The Dynamic Effects of Aggregate Demand and Supply Disturbances', *The American Economic Review* **79**(4), 655–673.

Blundell, W., Gowrisankaran, G. and Langer, A. (2020), 'Escalation of Scrutiny: The Gains from Dynamic Enforcement of Environmental Regulations', *American Economic Review* **110**(8), 2558–2585.

Boot, T. and Nibbering, D. (2019), 'Forecasting using random subspace methods', *Journal of Econometrics* **209**(2), 391–406.

Breiman, L. (2001), 'Random forests', *Machine learning* **45**(1), 5–32.

Breusch, T., Qian, H., Schmidt, P. and Wyhowski, D. (1999), 'Redundancy of Moment Conditions', *Journal of Econometrics* **91**(1), 89–111.

Burda, M., Harding, M. and Hausman, J. (2008), 'A Bayesian mixed logit–probit model for multinomial choice', *Journal of Econometrics* **147**(2), 232–246.

Byrne, J. P., Lorusso, M. and Xu, B. (2019), 'Oil Prices, Fundamentals and Expectations', *Energy Economics* **79**, 59–75.

Chen, X. (2007), 'Large sample sieve estimation of semi-nonparametric models', *Handbook of econometrics* **6**, 5549–5632.

Cheng, X. and Liao, Z. (2015), 'Select the Valid and Relevant Moments: An Information-Based LASSO for GMM with Many Moments', *Journal of Econometrics* **186**(2), 443–464.

Chernoff, H. and Moses, L. E. (1959), 'Elementary decision theory'.

Chernozhukov, V., Chetverikov, D., Demirer, M., Duflo, E., Hansen, C., Newey, W. and Robins, J. (2018), 'Double/debiased machine learning for treatment and structural parameters', *The Econometrics Journal* **21**, C1–C68.

Copas, J. B. (1983), 'Regression, prediction and shrinkage', *Journal of the Royal Statistical Society: Series B (Methodological)* **45**(3), 311–335.

Cranenburgh, S. V., Wang, S., Vij, A., Pereira, F. and Walker, J. (2021), 'Choice modelling in the age of machine learning', *arXiv preprint arXiv:2101.11948* .

Croissant, Y. (2019), *mlogit: Multinomial Logit Models*.
URL: *https://cran.r-project.org/package=mlogit*

Cybenko, G. V. (1989), 'Approximation by superpositions of a sigmoidal function', *Mathematics of Control, Signals and Systems* **2**, 303–314.

Defferrard, M., Pena, R. and Perraudin, N. (2017), 'PyUNLocBoX: Optimization by Proximal Splitting'.

Dezeure, R., Bühlmann, P. and Zhang, C.-H. (2017), 'High-dimensional simultaneous inference with the bootstrap', *Test* **26**(4), 685–719.

Draper, N. R. and Van Nostrand, R. C. (1979), 'Ridge regression and james-stein estimation: review and comments', *Technometrics* **21**(4), 451–466.

Efron, B. (1979), 'Bootstrap methods: Another look at the jackknife', *The Annals of Statistics* **7**, 1 – 26.

Efron, B. (2017), 'Charles stein, 1920-2016', *Journal of the Royal Statistical Society. Series A (Statistics in Society)* pp. 923–925.

Efron, B., Hastie, T., Johnstone, I., Tibshirani, R. and Others (2004), 'Least angle regression', *The Annals of statistics* **32**(2), 407–499.

Efron, B. and Morris, C. (1977), 'Stein's paradox in statistics', *Scientific American* **236**(5), 119–127.

El-Arini, K., Xu, M., Fox, E. B. and Guestrin, C. (2013), Representing Documents Through Their Readers, *in* 'Proceedings of the 19th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining', KDD '13, ACM, New York,

NY, USA, pp. 14–22.
Fan, J., Ke, Y. and Wang, K. (2020), 'Factor-adjusted regularized model selection', *Journal of Econometrics* **216**(1), 71–85.
Farrar, D. E. and Glauber, R. R. (1967), 'Multicollinearity in regression analysis: the problem revisited', *The Review of Economic and Statistics* pp. 92–107.
Farrell, M. H., Liang, T. and Misra, S. (2021*a*), 'Deep learning for individual heterogeneity: An automatic inference framework', *arXiv preprint arXiv:2010.14694v2* .
Farrell, M. H., Liang, T. and Misra, S. (2021*b*), 'Deep neural networks for estimation and inference', *Econometrica* **89**, 181–213.
Fourdrinier, D., Strawderman, W. E. and Wells, M. T. (2018), *Shrinkage estimation*, Springer.
Fox, J. T., Kim, K., Ryan, S. and Bajari, P. (2011), 'A simple estimator for the distribution of random coefficients', *Quantitative Economics* **2**(3), 381–418.
Fox, J. T., Kim, K. and Yang, C. (2016), 'A simple nonparametric approach to estimating the distribution of random coefficients in structural models', *Journal of Econometrics* **195**(2), 236–254.
Friedman, J., Hastie, T. and Tibshirani, R. (2010), 'Regularization paths for generalized linear models via coordinate descent', *Journal of Statistical Software* **33**(1), 1–22.
Gentle, J. E. (2007), *Matrix Algebra: Theory, Computations, and Applications in Statistics*, 1st edn, Springer Publishing Company, Incorporated.
Goodfellow, I., Bengio, Y. and Courville, A. (2016), *Deep Learning*, MIT Press. http://www.deeplearningbook.org.
Gouriéroux, C., Monfort, A. and Renne, J.-P. (2017), 'Statistical Inference for Independent Component Analysis: Application to Structural VAR Models', *Journal of Econometrics* **196**(1), 111–126.
Gruber, M. H. (2017), *Improving efficiency by shrinkage: The James-Stein and ridge regression estimators*, Routledge.
Guay, A. (2021), 'Identification of Structural Vector Autoregressions Through Higher Unconditional Moments', *Journal of Econometrics* **225**(1), 27–46.
Hall, A. R. (2005), *Generalized Method of Moments*, Oxford University Press.
Hall, A. R. (2015), 'Econometricians Have Their Moments: GMM at 32', *Economic Record* **91**, 1–24.
Hall, A. R., Inoue, A., Jana, K. and Shin, C. (2007), 'Information in Generalized Method of Moments Estimation and Entropy-Based Moment Selection', *Journal of Econometrics* **138**(2), 488–512.
Hansen, B. E. (2008), 'Least-squares forecast averaging', *Journal of Econometrics* **146**(2), 342–350.
Hansen, B. E. (2016*a*), 'Efficient shrinkage in parametric models', *Journal of Econometrics* **190**(1), 115–132.
Hansen, B. E. (2016*b*), 'The risk of james–stein and lasso shrinkage', *Econometric Reviews* **35**(8-10), 1456–1470.
Hastie, T., Tibshirani, R. and Friedman, J. (2009), *The elements of statistical learning: data mining, inference and prediction*, 2 edn, Springer.

Hebiri, M. and van de Geer, S. (2011), 'The Smooth-Lasso and other $\ell_1 + \ell_2$-penalized methods', *Electronic Journal of Statistics* **5**, 1184 – 1226.

Heiss, F. (2002), 'Structural choice analysis with nested logit models', *The Stata Journal* **2**, 227–252.

Heiss, F., Hetzenecker, S. and Osterhaus, M. (2021), 'Nonparametric estimation of the random coefficients model: An elastic net approach', *Journal of Econometrics* (Forthcoming).

Hess, S., Bierlaire, M. and Polak, J. W. (2005), 'Estimation of value of travel-time savings using mixed logit models', *Transportation Research Part A: Policy and Practice* **39**(2-3), 221–236.

Hoerl, A. E. and Kennard, R. W. (1970), 'Ridge regression: Biased estimation for nonorthogonal problems', *Technometrics* **12**(1), 55–67.

Hoerl, R. W. (2020), 'Ridge regression: a historical context', *Technometrics* **62**(4), 420–425.

Hoffmann, K. (2000), 'Stein estimation—a review', *Statistical Papers* **41**(2), 127–158.

Hornik, K., Stinchcombe, M. and White, H. (1989), 'Multilayer feedforward networks are universal approximators', *Neural Networks* **2**, 359–366.

Houde, S. and Myers, E. (2019), Heterogeneous (Mis-) Perceptions of Energy Costs: Implications for Measurement and Policy Design, Working Paper 25722, National Bureau of Economic Research.

Hu, Z., Follmann, D. A. and Miura, K. (2015), 'Vaccine design via nonnegative lasso-based variable selection', *Statistics in medicine* **34**(10), 1791–1798.

Illanes, G. and Padi, M. (2019), Competition, Asymmetric Information, and the Annuity Puzzle: Evidence from a Government-Run Exchange in Chile, Technical report, Center for Retirement Research.

James, W. and Stein, C. (1961), Estimation with quadratic loss, *in* 'Proceedings of the Fourth Berkeley Symposium on Mathematical Statistics and Probability, Volume 1: Contributions to the Theory of Statistics', University of California Press, pp. 361–379.

Jaynes, E. T. (2003), *Probability theory: The logic of science*, Cambridge university press.

Jentsch, C. and Leucht, A. (2016), 'Bootstrapping sample quantiles of discrete data', *Annals of the Institute of Statistical Mathematics* **68**(3), 491–539.

Jia, J. and Yu, B. (2010), 'On model selection consistency of the elastic net when p $\gg$ n', *Statistica Sinica* **20**(2), 595–611.

Juvenal, L. and Petrella, I. (2015), 'Speculation in the Oil Market', *Journal of Applied Econometrics* **30**(4), 621–649.

Karlaftis, M. G. and Vlahogianni, E. I. (2011), 'Statistical methods versus neural networks in transportation research: Differences, similarities and some insights', *Transportation Research Part C: Emerging Technologies* **19**, 387–399.

Keweloh, S. A. (2021*a*), 'A Feasible Approach to Incorporate Information in Higher Moments in Structural Vector Autoregressions', *Discussion Papers SFB 823* .

Keweloh, S. A. (2021*b*), 'A Generalized Method of Moments Estimator for Structural Vector Autoregressions Based on Higher Moments', *Journal of Business & Economic*

Statistics **39**(3), 772–782.
Kilian, L. (2009), 'Not All Oil Price Shocks Are Alike: Disentangling Demand and Supply Shocks in the Crude Oil Market', *American Economic Review* **99**(3), 1053–69.
Kilian, L. and Lütkepohl, H. (2017), *Structural Vector Autoregressive Analysis*, Cambridge University Press.
Kilian, L. and Murphy, D. P. (2014), 'The Role of Inventories and Speculative Trading in the Global Market for Crude Oil', *Journal of Applied econometrics* **29**(3), 454–478.
Kilian, L. and Park, C. (2009), 'The Impact of Oil Price Shocks on the US Stock Market', *International Economic Review* **50**(4), 1267–1287.
Kim, Y., Hao, J., Gautam, Y., Mersha, T. B. and Kang, M. (2018), 'Diffgrn: differential gene regulatory network analysis', *International journal of data mining and bioinformatics* **20**(4), 362–379.
Koppelman, F. S. and Wen, C.-H. (2000), 'The paired combinatorial logit model: properties, estimation and application', *Transportation Research Part B: Methodological* **34**(2), 75–89.
Kump, P., Bai, E.-W., Chan, K.-S., Eichinger, B. and Li, K. (2012), 'Variable selection via RIVAL (removing irrelevant variables amidst Lasso iterations) and its application to nuclear material detection', *Automatica* **48**(9), 2107–2115.
Lanne, M., Liu, K. and Luoto, J. (2021), 'Identifying Structural Vector Autoregression via Large Economic Shocks', *Available at SSRN 3910532* .
Lanne, M. and Luoto, J. (2021), 'GMM Estimation of Non-Gaussian Structural Vector Autoregression', *Journal of Business & Economic Statistics* **39**(1), 69–81.
Lanne, M., Lütkepohl, H. and Maciejowska, K. (2010), 'Structural Vector Autoregressions with Markov Switching', *Journal of Economic Dynamics and Control* **34**(2), 121–131.
Lanne, M., Meitz, M. and Saikkonen, P. (2017), 'Identification and Estimation of Non-Gaussian Structural Vector Autoregressions', *Journal of Econometrics* **196**(2), 288–304.
LeCun, Y., Bengio, Y. and Hinton, G. (2015), 'Deep learning', *nature* **521**, 436–444.
Lehmann, E. L. and Casella, G. (2006), *Theory of point estimation*, Springer Science & Business Media.
Leung, G. and Barron, A. R. (2006), 'Information theory and mixing least-squares regressions', *IEEE Transactions on information theory* **52**(8), 3396–3410.
Lewis, D. J. (2021), 'Identifying Shocks via Time-Varying Volatility', *The Review of Economic Studies* **88**(6), 3086–3124.
Liu, Q., Okui, R. and Yoshimura, A. (2016), 'Generalized least squares model averaging', *Econometric Reviews* **35**(8-10), 1692–1752.
Lütkepohl, H. and Netšunajev, A. (2017), 'Structural Vector Autoregressions with Smooth Transition in Variances', *Journal of Economic Dynamics and Control* **84**, 43–57.
Marwick, K. P. and Koppelman, F. S. (1990), 'Proposals for analysis of the market demand for high speed rail in the Quebec/Ontario corridor', *Submitted to Ontario/Quebec Rapid Task Force* .
McFadden, D. and Train, K. (2000), 'Mixed MNL models for discrete response', *Journal*

of Applied Econometrics **15**(5), 447–470.

Mertens, K. and Ravn, M. O. (2013), 'The Dynamic Effects of Personal and Corporate Income Tax Changes in the United States', *American Economic Review* **103**(4), 1212–47.

Moral-Benito, E. (2015), 'Model averaging in economics: An overview', *Journal of Economic Surveys* **29**(1), 46–75.

Moussa, Z. and Thomas, A. (2021), 'Identifying Oil Supply News Shocks and their Effects on the Global Oil Market', *USAEE Working Paper No. 21-490* .

Nevo, A., Turner, J. L. and Williams, J. W. (2016), 'Usage-Based Pricing and Demand for Residential Broadband', *Econometrica* **84**(2), 411–443.

Newey, W. K. (1994), 'The asymptotic variance of semiparametric estimators', *Econometrica* **62**, 1349–1382.

Nocedal, J. and Wright, S. J. (2006), *Numerical optimization*, Springer series in operations research and financial engineering, 2. ed. edn, Springer, New York, NY.

Olea, M., Luis, J., Plagborg-Møller, M. and Qian, E. (2022), 'SVAR Identification From Higher Moments: Has the Simultaneous Causality Problem Been Solved?', *Working Paper* . Prepared for AEA Papers and Proceedings.

Park, H., Imoto, S. and Miyano, S. (2015), 'Recursive random lasso (rrlasso) for identifying anti-cancer drug targets', *PLoS One* **10**(11), e0141869.

Park, H., Niida, A., Imoto, S. and Miyano, S. (2017), 'Interaction-based feature selection for uncovering cancer driver genes through copy number-driven expression level', *Journal of Computational Biology* **24**(2), 138–152.

Pötscher, B. M. and Leeb, H. (2009), 'On the distribution of penalized maximum likelihood estimators: The LASSO, SCAD, and thresholding', *Journal of Multivariate Analysis* **100**(9), 2065–2082.

R Core Team (2018), *R: A Language and Environment for Statistical Computing*, R Foundation for Statistical Computing, Vienna, Austria.

Rigobon, R. (2003), 'Identification through Heteroskedasticity', *Review of Economics and Statistics* **85**(4), 777–792.

Rossi, P. E., Allenby, G. M. and McCulloch, R. (2012), *Bayesian statistics and marketing*, John Wiley & Sons.

Sifringer, B., Lurkin, V. and Alahi, A. (2020), 'Enhancing discrete choice models with representation learning', *Transportation Research Part B: Methodological* **140**, 236–261.

Sims, C. A. (1980), 'Macroeconomics and Reality', *Econometrica: Journal of the Econometric Society* pp. 1–48.

Slawski, M. and Hein, M. (2013), 'Non-negative least squares for high-dimensional linear models: Consistency and sparse recovery without regularization', *Electron. J. Statist.* **7**, 3004–3056.

Stein, C. (1956), Inadmissibility of the usual estimator for the mean of a multivariate normal distribution, *in* 'Proceedings of the Third Berkeley symposium on mathematical statistics and probability', Vol. 1, pp. 197–206.

Stock, J. H. and Watson, M. (2012), 'Disentangling the Channels of the 2007-09 Recession',

Brookings Papers on Economic Activity **43**(1), 81–156.

Takada, M., Suzuki, T. and Fujisawa, H. (2017), 'Independently Interpretable Lasso: A New Regularizer for Sparse Regression with Uncorrelated Variables', *arXiv preprint arXiv:1711.01796* .

Tibshirani, R. (1996), 'Regression shrinkage and selection via the lasso', *Journal of the Royal Statistical Society: Series B (Methodological)* **58**(1), 267–288.

Train, K. (2008), 'EM algorithms for nonparametric estimation of mixing distributions', *Journal of Choice Modelling* **1**(1), 40–69.

Train, K. (2009), *Discrete choice methods with simulation*, Cambridge university press.

Train, K. (2016), 'Mixed logit with a flexible mixing distribution', *Journal of Choice Modelling* **19**, 40–53.

Uhlig, H. (2005), 'What are the Effects of Monetary Policy on Output? Results from an Agnostic Identification Procedure', *Journal of Monetary Economics* **52**(2), 381–419.

Vinod, H. D. (1978), 'A survey of ridge regression and related techniques for improvements over ordinary least squares', *The Review of Economics and Statistics* pp. 121–131.

Wainwright, M. J. (2006), 'Sharp thresholds for sparsity recovery in the noisy and high-dimensional setting using ℓ_1-constrained quadratic programming', *44th Annual Allerton Conference on Communication, Control, and Computing 2006* **1**(5), 217–224.

Wang, S., Mo, B., Hess, S. and Zhao, J. (2021), 'Comparing hundreds of machine learning classifiers and discrete choice models in predicting travel behavior: an empirical benchmark', *arXiv preprint arXiv:2102.01130* .

Wang, S., Nan, B., Rosset, S. and Zhu, J. (2011), 'Random lasso', *The annals of applied statistics* **5**(1), 468.

Wang, S., Wang, Q. and Zhao, J. (2020), 'Deep neural networks for choice analysis: Extracting complete economic information for interpretation', *Transportation Research Part C: Emerging Technologies* **118**, 102701.

Wen, C.-H. and Koppelman, F. S. (2001), 'The generalized nested logit model', *Transportation Research Part B: Methodological* **35**(7), 627–641.

Wong, M. and Farooq, B. (2021), 'Reslogit: A residual neural network logit model for data-driven choice modelling', *Transportation Research Part C: Emerging Technologies* **126**, 103050.

Wu, L. and Yang, Y. (2014), 'Nonnegative Elastic Net and application in index tracking', *Applied Mathematics and Computation* **227**, 541–552.

Wu, L., Yang, Y. and Liu, H. (2014), 'Nonnegative-lasso and application in index tracking', *Computational Statistics and Data Analysis* **70**, 116–126.

Yu, X., Cen, L., Chen, Y., Markowitz, J., Shaw, T. I., Tsai, K. Y., Conejo-Garcia, J. R. and Wang, X. (2022), 'Tumor expression quantitative trait methylation screening reveals distinct cpg panels for deconvolving cancer immune signaturestumor expression quantitative trait methylation', *Cancer Research* .

Zha, T. (1999), 'Block Recursion and Structural Vector Autoregressions', *Journal of Econometrics* **90**(2), 291–316.

Zhao, P. and Yu, B. (2006), 'On model selection consistency of Lasso', *Journal of Machine*

learning research **7**, 2541–2563.

Zou, H. (2006), 'The adaptive lasso and its oracle properties', *Journal of the American statistical association* **101**(476), 1418–1429.

Zou, H. and Hastie, T. (2005), 'Regularization and variable selection via the elastic net', *Journal of the Royal Statistical Society. Series B (Statistical Methodology)* **67**(2), 301–320.

Eidesstattliche Erklärung

Ich gebe folgende eidesstattliche Erklärung ab:

Ich erkläre hiermit, dass ich die vorliegende Arbeit selbständig ohne unzulässige Hilfe Dritter verfasst, keine anderen als die angegebenen Quellen und Hilfsmittel benutzt und alle wörtlich oder inhaltlich übernommenen Stellen unter der Angabe der Quelle als solche gekennzeichnet habe.

Die Grundsätze für die Sicherung guter wissenschaftlicher Praxis an der Universität Duisburg-Essen sind beachtet worden.

Ich habe die Arbeit keiner anderen Stelle zu Prüfungszwecken vorgelegt.

Ort, Datum	Stephan Hetzenecker

Erklärung zur Koautorenschaft

Kapitel 2

Das zweite Kapitel "Nonparametric Estimation of the Random Coefficients Model: An Elastic Net Approach" ist in Zusammenarbeit mit Florian Heiß und Maximilian Osterhaus entstanden. Wir erklären hiermit, dass Stephan Hetzenecker an sämtlichen Teilen der Arbeit in etwa proportional beteiligt war.

Ort, Datum	Florian Heiß
Ort, Datum	Maximilian Osterhaus
Ort, Datum	Stephan Hetzenecker

Erklärung zur Koautorenschaft

Kapitel 4

Das vierte Kapitel "Deep Learning for the Estimation of Heterogeneous Parameters in Discrete Choice Models" ist in Zusammenarbeit mit Maximilian Osterhaus entstanden. Wir erklären hiermit, dass Stephan Hetzenecker an sämtlichen Teilen der Arbeit in etwa proportional beteiligt war.

Ort, Datum	Maximilian Osterhaus
Ort, Datum	Stephan Hetzenecker

Erklärung zur Koautorenschaft

Kapitel 5

Das fünfte Kapitel "Block-Recursive Non-Gaussian Structural Vector Autoregressions: Identification, Efficiency, and Moment Selection" ist in Zusammenarbeit mit Sascha Alexander Keweloh entstanden. Sascha Alexander Keweloh hatte die Idee zur Verwendung von blok-recursiven Restriktionen zur Identifikation von SVARs und hat die Beweise bezüglich der Identifikation maßgeblich ausgebarbeitet. Stephan Hetzenecker hatte die Idee zur Verwendung von LASSO zur Momentenauswahl und hat die Beweise zu LASSO maßgeblich ausgebarbeitet. Insbesondere haben beide Autoren an allen Beweisen mitgearbeitet. Wir erklären hiermit, dass Stephan Hetzenecker an sämtlichen Teilen der Arbeit in etwa proportional beteiligt war.

Ort, Datum	Sascha Alexander Keweloh
Ort, Datum	Stephan Hetzenecker