Three Essays on Estimating Heterogeneity in Discrete Choice Models

Dissertation
zur Erlangung des akademischen Grades
Doctor Rerum Politicarum
im Fach Volkswirtschaftslehre
durch die Wirtschaftswissenschaftliche Fakultät
der Heinrich-Heine-Universität Düsseldorf

von: Maximilian Osterhaus
geboren am 14.11.1989 in Osnabrück

Erstgutachter: Prof. Dr. Florian Heiß
Zweitgutachter: Prof. Dr. Joel Stiebale

Abgabedatum: 08.11.2021

Acknowledgment

Writing this thesis would not have been possible without the support of many people. I am very grateful for their support, feedback, guidance, and patience. I know that it is impossible to return all these favors, but I would still like to try as best I can.

First and foremost, I would like to thank my supervisors Prof. Dr. Florian Heiss and Prof. Dr. Joel Stiebale for their advice and support during the completion of this thesis. They have shown me confidence and granted me the invaluable freedom to pursue my own research interests.

Special thanks go to my friend and co-author Stephan Hetzenecker for the great collaboration and his patience. I learned so much from the numerous interesting and insightful discussions we had. I could not have asked for a better co-author.

In addition, I would also like to thank Prof. Toker Doganoglu, Ph.D.. He sparked my interest in empirical economics and gave me the optimal starting conditions for this thesis. He encouraged me to pursue a doctorate and supported and advised me far beyond my time in Würzburg.

I would also like to thank my colleagues who created a pleasant and stimulating working environment. I enjoyed working in this environment and benefited from the numerous discussions. In particular, I want to mention Hedieh Aghelmaleki, who was very supportive and made work a joyful experience.

To my family and close friends, I am grateful for the endless support I have received along the way. Without the invaluable support of my parents, this work would not have been possible. To my brother Christopher, I am grateful for his academic advice and the proofreading of countless pages. Finally, I would like to dedicate this work to my wife Carolin. She has always believed in me and always picked me up in the downs of doing research, giving me perspective, and putting a smile on my face. I am deeply grateful for her guidance, encouragement, and support, which I will never take for granted.

to Carolin

Contents

List of Tables

List of Figures

Introduction

Modeling heterogeneity across economic agents plays an important role in many empirical economic studies: Consumers have heterogeneous preferences for product characteristics, firms have heterogeneous production costs, or workers differ in their preferences concerning wages and commuting times. The consistent estimation of heterogeneous parameters in econometric models is not only important for the identification of the heterogeneity across economic agents but also for consistent counterfactuals and correct policy recommendations. For instance, firms must have an idea about consumers' preferences to assess how different consumers might respond to changes in prices and quality of existing alternatives, or regulatory agencies need to estimate firms' costs to evaluate the desirability of market outcomes.

The literature distinguishes between two sources of heterogeneity: observed and unobserved heterogeneity. While the former links the differences across agents to differences in their observed characteristics, the latter captures idiosyncratic variations across agents. Given constraints on the amount of available data, many empirical studies employ parametric estimators for heterogeneous parameters. These estimators, however, restrict the form of the observed and unobserved heterogeneity to the functional form specified by the researcher. Fortunately, the increasing availability of large datasets makes it possible to reduce the reliance on parametric methods and to study heterogeneity across economic agents at new levels of detail. The nuanced study of heterogeneity, however, requires sufficiently flexible estimation approaches that can handle large amounts of data while being computationally feasible. This thesis addresses this challenge in two lines of work.

Chapter 1 and Chapter 2 study the nonparametric estimation of random coefficient models, which are widely used to capture unobserved heterogeneity. In these models, the parameters vary across agents according to an unknown distribution that the researcher attempts to estimate from the data.[1] Parametric estimators for this model assume a family of distributions for the random coefficients prior to the estimation, thereby restricting the shape of the estimated distribution to the shape of the assumed family of distributions. Nonparametric estimators overcome this limitation as they do not require researchers to make such a priori assumptions but allow them to estimate distributions of any shape. However, this flexibility is usually accompanied by a high computational cost, emphasizing the need for computationally simple and fast nonparametric estimators.

Chapter 1, forthcoming in the Journal of Econometrics, proposes such a computationally simple and fast nonparametric estimator for random coefficient models. The estimator extends the fixed-grid estimator of Fox, Kim, Ryan, and Bajari (2011), who propose approximating the underlying

[1] The model can also be combined with observed heterogeneity by specifying the parameters of the random coefficients' distribution as a function of observed characteristics (cf. Greene, Hensher, and Rose, 2006).

random coefficients' distribution through a discrete distribution with fixed support points. We show that the estimator is a special case of nonnegative lasso (Wu, Yang, and Liu, 2014), which explains its sparsity, leading to inaccurate approximations of the true distribution through step functions with only a few steps. Recognizing this link, we extend the estimator by transforming it into a special case of the nonnegative elastic net (Wu and Yang, 2014). Our theoretical results as well as finite sample simulations demonstrate that the extension improves the selection of the "true" support points and provides more accurate approximations of the underlying distribution.

Many nonparametric estimators for random coefficient models, including the fixed grid estimator, face a severe limitation, which is the exponential increase of the number of parameters in the number of random coefficients included in the model. This property, known as the curse of dimensionality, limits the application of such estimators to moderately low-dimensional random coefficient models. Chapter 2 addresses this problem and presents a nonparametric sparse grid estimator for high-dimensional random coefficient models. The estimator uses a truncated tensor product of hierarchical basis functions for the approximation of the underlying distribution. Due to the truncation, the number of parameters increases substantially slower than exponentially, rendering the nonparametric estimation of high-dimensional random coefficient models feasible. Monte Carlo experiments show that the truncation deteriorates the approximation accuracy only slightly if the underlying distribution is sufficiently smooth. Moreover, the experiments show the good performance of the sparse grid estimator compared to existing nonparametric estimators – even when the distribution is not smooth. The superiority in performance of our estimator is particularly pronounced for models with moderately high-dimensional random coefficients, for which the sample size imposes restrictions on the number of parameters that can be estimated using existing estimators. For non-smooth distributions, we study a spatially adaptive refinement procedure. The spatially adaptive refinement gradually adds basis functions in those areas of the random coefficients' distribution where it exhibits a wiggly and steep curvature to further improve the approximation accuracy.

The second line of work, in form of Chapter 3, addresses the nonparametric estimation of models with observed heterogeneity. In this chapter, we study the finite sample performance of the flexible estimation approach of Farrell, Liang, and Misra (2021a), who propose to use deep learning for the estimation of heterogeneous parameters from observed characteristics of economic agents, in the context of discrete choice models. The approach combines the features of parametric approaches – which impose a structure on the model based on economic principles and reasoning – with deep learning – which allows estimating flexible functional forms of heterogeneity. For valid second-stage inference after first-stage estimation of econometric models with deep learning, Farrell et al. (2021a) adopt the influence function approach of Chernozhukov et al. (2018). We conduct a series of Monte Carlo experiments that investigate the impact of regularization, which is commonly employed when using deep learning – and machine learning in general – on the proposed inference procedure. The results of these experiments provide three main insights: First, deep learning for the estimation of heterogeneous parameters generally allows to recover precise estimates of the true average parameters but does not allow for valid inference statements when regular robust standard errors are used. Second, the inference procedure proposed by Farrell et al. (2021a) appears to

be sensitive to overfitting, expressing itself through substantial bias and large estimated standard errors. Regularization reduces the impact of overfitting on the estimation results but induces an additional bias. The bias in combination with decreasing variance associated with increasing regularization leads to the construction of invalid inference statements in our experiments. And third, the experiments show that much better results are obtained when repeated sample splitting is used. Unlike regularization, repeated sample splitting reduces the sensitivity to overfitting without introducing an additional bias, thereby allowing for the construction of valid inference statements.

Taken together, this thesis shows how heterogeneity, which is a fundamental concept in econometric models, can be recovered with less restrictive assumptions on its functional forms by nonparametric estimation approaches. While our simulations illustrate the ability of the estimators to recover complex forms of heterogeneity across economic agents, our work also documents some limitations and areas where the estimators result in poor estimation results. Uncovering these limitations is fundamental for the improvement of existing procedures. Therefore, this thesis paves new directions for the flexible estimation of heterogeneity in econometric models, which is substantial to explain the behavior of economic agents.

Chapter 1

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

Co-authored by Florian Heiss and Stephan Hetzenecker

1.1 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, Kim, and Yang, 2016) and applications to real data (e.g., Nevo, Turner, and Williams, 2016, Illanes and Padi, 2019, Blundell, Gowrisankaran, and Langer, 2020 and Houde and Myers, 2021). 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, Tibshirani, and Friedman, 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 1.2 describes the FKRB estimator and introduces our generalized version. Section 1.3 derives the condition on the estimators' sign consistency and an error bound on the estimated distribution functions. Section 1.4 presents two Monte Carlo experiments that investigate the performance of our generalized estimator in comparison to the FKRB estimator. Section 1.5 applies the estimators to the *Mode Canada* data set from the R package *mlogit* (Croissant, 2019). Section 1.6 concludes and provides an outlook.

1.2 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

$$P_{i,j}(x) = \int g(x_{i,j}, \beta)\, dF_0(\beta) \tag{1.1}$$

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 (1.1) 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

$$P_{i,j}(x) = \int \frac{\exp\left(x_{i,j}^T \beta\right)}{1 + \sum_{l=1}^{J} \exp\left(x_{i,l}^T \beta\right)} dF_0(\beta). \tag{1.2}$$

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

1.2.1 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 (1.1) into a probability model in which $F_0(\beta)$ enters linearly. FKRB derive the linear probability model in two steps: they transform Equation (1.1) 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

$$y_{i,j} = \int g\left(x_{i,j}, \beta\right) dF_0\left(\beta\right) + \left(y_{i,j} - P_{i,j}\left(x\right)\right). \tag{1.3}$$

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 (1.3) becomes a simple linear probability model with unknown parameters $\theta = (\theta_1, \ldots, \theta_R)^T$

$$y_{i,j} \approx \sum_{r=1}^{R} g\left(x_{i,j}, \beta_r\right) \theta_r + \left(y_{i,j} - P_{i,j}\left(x\right)\right) \tag{1.4}$$

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 observations, 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

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

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

$$\begin{aligned} \hat{\theta}^{FKRB} &= \arg\min_{\theta} \frac{1}{2NJ} \sum_{i=1}^{N} \sum_{j=1}^{J} \left(y_{i,j} - \sum_{r=1}^{R} \theta_r z_{i,j}^{r} \right)^2 \\ \text{s.t.} \quad & \theta_r \geq 0, \quad r = 1, \ldots, R, \quad \text{and} \quad \sum_{r=1}^{R} \theta_r = 1. \end{aligned} \tag{1.5}$$

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.

1.2.2 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 (1.5) 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.

1.2.2.1 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 (1.5) 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 (1.4) with $1 - \sum_{r=1}^{R-1} \theta_r$ gives the inequality constrained optimization problem

$$\begin{aligned} \hat{\theta}^{\text{FKRB}} = \arg\min_{\theta} \frac{1}{2NJ} \sum_{i=1}^{N} \sum_{j=1}^{J} \left(\tilde{y}_{i,j} - \sum_{r=1}^{R-1} \theta_r \tilde{z}_{i,j}^r \right)^2 \\ \text{s.t.} \quad \theta_r \geq 0, \quad r = 1, \ldots, R-1, \quad \text{and} \quad \sum_{r=1}^{R-1} \theta_r \leq 1 \end{aligned} \tag{1.6}$$

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 (1.6) is an equivalent form of the optimization problem in Equation (1.5), 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 optimiza-

tion 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 (1.4),

$$\begin{aligned} \hat{\theta}^{\text{NNL}} = \arg\min_{\theta} \frac{1}{2NJ} \sum_{i=1}^{N} \sum_{j=1}^{J} \left(\tilde{y}_{i,j} - \sum_{r=1}^{R-1} \theta_r \tilde{z}_{i,j}^r \right)^2 \\ \text{s.t.} \quad \theta_r \geq 0, \quad r = 1, \ldots, R-1, \quad \text{and} \quad \sum_{r=1}^{R-1} \theta_r \leq c, \end{aligned} \tag{1.7}$$

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 (1.6) 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).

1.2.2.2 Elastic Net Estimator

Extending the FKRB estimator's optimization problem formulated in Equation (1.6) 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

$$\begin{aligned} \hat{\theta}^{\text{ENET}} &= \arg\min_{\theta} \frac{1}{2NJ} \sum_{i=1}^{N} \sum_{j=1}^{J} \left(\tilde{y}_{i,j} - \sum_{r=1}^{R-1} \theta_r \tilde{z}_{i,j}^r \right)^2 \\ \text{s.t.} \quad \theta_r \geq 0, \quad r = 1, \ldots, R-1, \quad &\text{and} \quad \sum_{r=1}^{R-1} \theta_r \leq 1 \quad \text{and} \quad \sum_{r=1}^{R-1} \theta_r^2 \leq t \end{aligned} \tag{1.8}$$

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 1.3 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 (1.8)

$$\hat{\theta}^{\text{ENET}} = \arg\min_{\theta} \frac{1}{2NJ} \sum_{i=1}^{N} \sum_{j=1}^{J} \left(\tilde{y}_{i,j} - \sum_{r=1}^{R-1} \theta_r \tilde{z}_{i,j}^r \right)^2 + \frac{1}{2}\mu \sum_{r=1}^{R-1} \theta_r^2 + \lambda \left(\sum_{r=1}^{R-1} \theta_r - 1 \right) - \sum_{r=1}^{R-1} \nu_r \theta_r. \tag{1.9}$$

The first term in Equation (1.9) 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 (1.8). 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 (1.9) 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 (1.8) 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 (1.9) and Equation (1.8) 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 (1.6), 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 1.3 and the Monte Carlo studies in Section 1.4 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 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 non-negative 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.

1.3 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 1.3.1, we study both estimators' ability to select the correct weights. To evaluate the overall approximation accuracy of the estimators presented in Section 1.2, we derive an error bound for the estimated probability weights and the estimated distribution functions in Subsection 1.3.2.

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), 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.

[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.

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 1.2.

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 $\theta^* = (\theta_1^*, \ldots, \theta_{R-1}^*)^T$ 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] 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 (1.9). We exploit the structure of our data and make the following assumptions on the linear probability model corresponding to $F^*(\beta)$

$$y_{i,j} = \sum_{r=1}^{R} \theta_r^* z_{i,j}^r + \epsilon_{i,j}, \tag{1.10}$$

[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 $r = 1, \ldots, R-1$, 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.

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 1.

(i) $\left(\epsilon_i = (\epsilon_{i,1}, ..., \epsilon_{i,J})\right)_{i=1}^{N}$ *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) $\left(\tilde{Z}_i\right)_{i=1}^{N}$ *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 (1.6) and $\tilde{Z}_i$ to the corresponding $J \times R - 1$ regressor matrix for observation unit i. Assumption 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 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 (1.10) 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 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 (1.2) 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 1(iv) holds by the definition of linear probability models.

1.3.1 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 (1.9). That is, we only identify whether the $R - 1$ weights included in Equation (1.9) 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 1. *An estimate* $\hat{\theta}$ *is* ***sign consistent*** *if*

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

According to Definition 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 $S = \{r \in \{1, \ldots, R-1\} | \theta_r^* > 0\}$ define the index set of grid points at which $\theta^* > 0$, and let $S^C = \{r \in \{1, \ldots, R-1\} | \theta_r^* = 0\}$ 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 (1.9), that follows from the constraint $c = 1$ in Equation (1.8). 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]

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

[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.

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. $\lim_{N\to\infty} 2\ s(N)J \exp\left(-\frac{N\xi^S_{\min}(\mu,N)^2\rho(\mu,N)^2}{2\ s(N)}\right) = 0.$

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

where $\xi^S_{\min}(\mu,N)$ *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 $\xi^S_{\min}(\mu,N) > 0$.

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 1. *Suppose Assumption 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 1.6. □

Theorem 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 1 reveals that our estimator can select the true weights in cases in which the FKRB estimator cannot.

Remark 2. Theorem 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 (1.2). Since Since Assumptions 1(i)-(iv) also hold true for the smooth basis densities estimator, Theorem 1 also applies to the estimator whereby the selection consistency relates to the correct recovery of active and inactive basis densities.

1.3.2 Error Bounds

A key requirement for an accurate estimation of $F_0(\beta)$ – in addition to the correct support recovery discussed in Subsection 1.3.1 – 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

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

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 1.3.1, $\xi_{\min}(\mu) > 0$ 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.

[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 (1.5). 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).

Theorem 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 1 holds, and that* $\xi_{\min}(\mu) > 0$ *for* $\mu \geq 0$. *Then, it holds with probability* $1 - \delta$ *that*

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

Proof. See Appendix 1.6. □

Theorem 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 $\left\|\hat{\theta} - \theta^*\right\|_2$ 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 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 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 1 establishes the condition under which our extension provides a tighter error bound on the estimated weights than the FKRB estimator.

Corollary 1. *When* $\sqrt{s}\ \left\|\theta_S^*\right\|_\infty \xi_{\min}(0) < \sqrt{R-1}\ \gamma$, *then the error bound for* $\left\|\hat{\theta} - \theta^*\right\|_2$ *in Theorem 2 is tighter for the generalized estimator than for the FKRB estimator.*

Proof. See Appendix 1.6. □

Using the error bound on the estimated and true probability weights in Theorem 2, we derive a bound on the error of the estimated distribution function $\hat{F}(\beta)$ and the best discrete distribution $F^*(\beta)$.

Theorem 3. *Under the assumptions and conditions in Theorem 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 1.6. □

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, 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 3.

Remark 3. Theorem 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 3. Consequently, the absolute difference between $\hat{F}(\beta)$ and $F^*(\beta)$ can also be bounded by Theorem 3 if $D \to \infty$.

1.4 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).

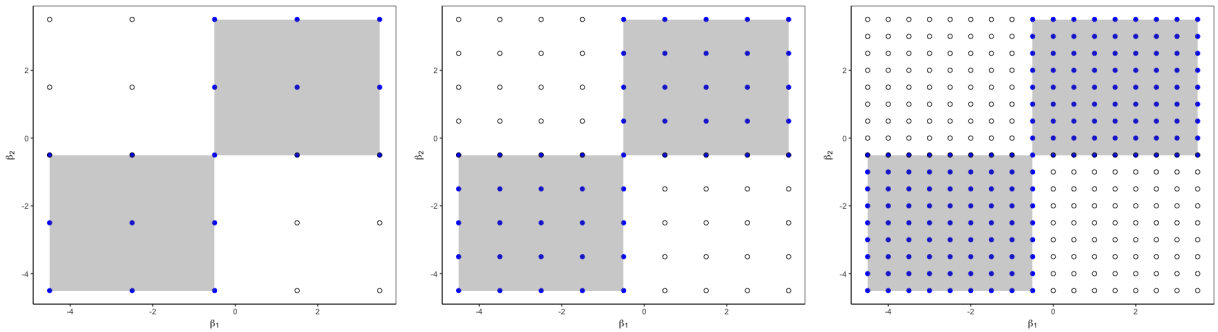
1.4.1 Discrete Distribution

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

Figure 1.1: Grid of Monte Carlo Study with Discrete Mass Points



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

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

$$L_1 = \frac{1}{M} \sum_{m=1}^{M} \frac{1}{R} \sum_{r=1}^{R} \left| \theta_r - \hat{\theta}_{r,m} \right| \tag{1.11}$$

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

$$\text{RMISE} = \sqrt{\frac{1}{M} \sum_{m=1}^{M} \left[\frac{1}{E} \sum_{e=1}^{E} \left(\widehat{F}_m(\beta_e) - F_0(\beta_e) \right)^2 \right]}, \tag{1.12}$$

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 1.1 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

[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.

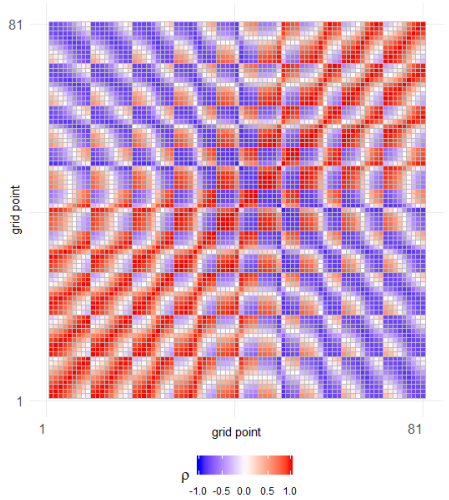
Table 1.1: Summary Statistics of 200 Monte Carlo Runs with Discrete Distribution.

			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).

upper part of the table presents the measures on the accuracy 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]

Figure 1.2: Correlation Matrix for $N = 10,000$ and $R = 81$



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-

[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.

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 1.2 and the third quantile of the values of absolute correlation in Table 1.1 both illustrate that correlation among many grid points is strong.

1.4.2 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

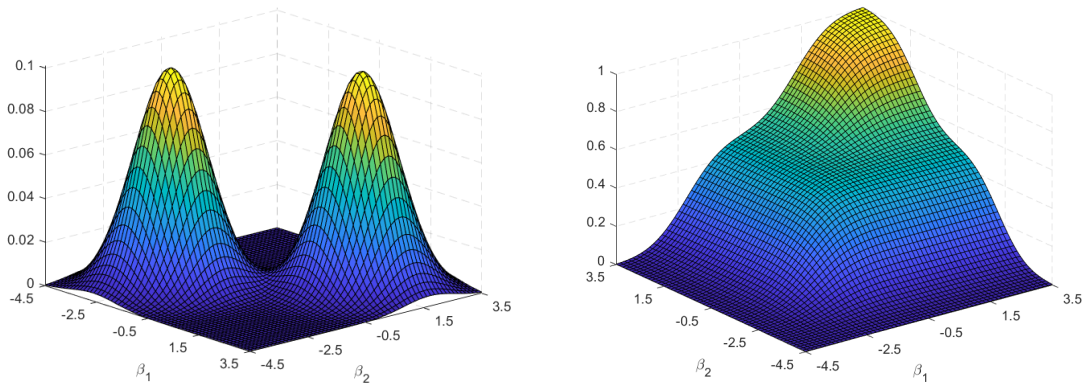
$$0.5\, \mathcal{N}\Big([-2.2, -2.2], \Sigma_1\Big) + 0.5\, \mathcal{N}\Big([1.3, 1.3], \Sigma_2\Big).$$

The left panel in Figure 1.3 displays the bimodal joint density of the mixture of two normals, and the right panel the joint distribution function.

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

Figure 1.3: True Density and Distribution Function of Mixture of two Normals



(a) PDF

(b) CDF

Table 1.2: 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

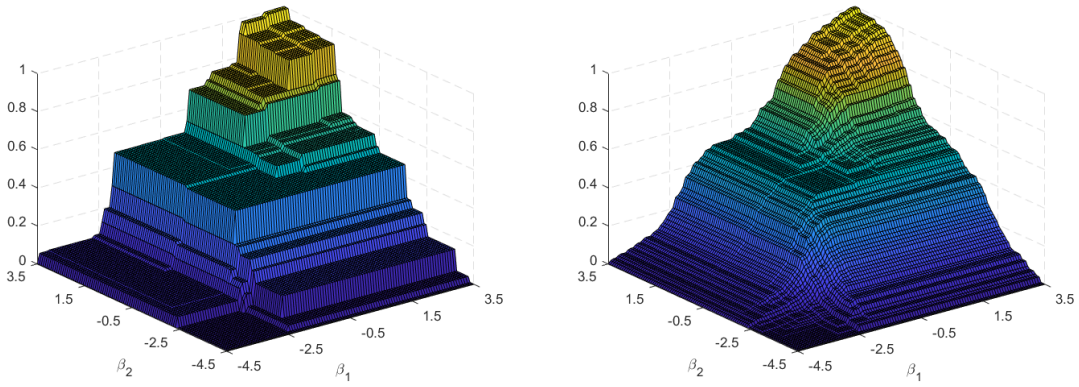
Note: The table reports the average summary statistics over all Monte Carlo replicates for the FKRB estimator (FKRB), and for our generalized estimator with tuning parameter μ from a 10-fold cross-validation and the MSE criterion (MSE) and the one-standard-error rule (OneSe).

predicted choices and the log-likelihood as criteria are reported in Appendix 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 1.4: Estimated Joint Distribution Functions for $N = 10,000$ and $R = 250$



(a) FKRB

(b) Generalized with OneSe

Figure 1.5: True and Estimated Marginal Distribution Functions for $N = 10,000$ and $R = 250$

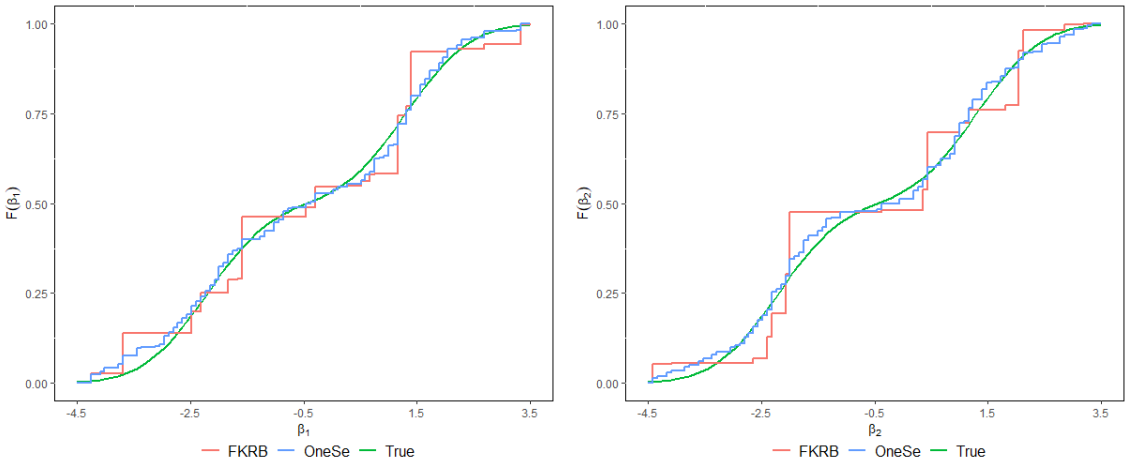


Figure 1.4 plots an example of the joint distribution functions estimated with the FKRB estimator (Panel (a)) and our generalized estimator (Panel (b)). Figure 1.5 shows the corresponding estimated

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.

1.5 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 (1997b) 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, 1997a, 1997b, 1998), 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 (2021) 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 (2021) justify

[10] We also provide an algorithm to update both the fixed and random coefficients in Appendix 1.6. 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.

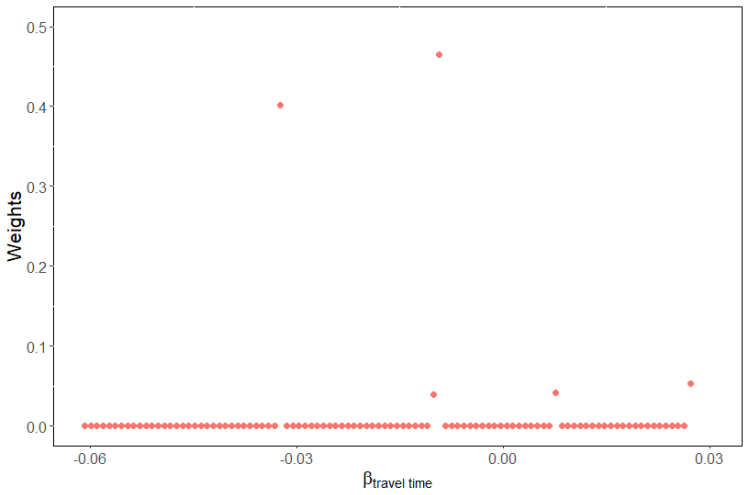
the procedure with the argument that a mixed logit can recover the means of a distribution fairly well despite the incorrect assumptions on the random coefficients' distribution. Thus, the fixed coefficients can be estimated consistently with the semiparametric approach. They illustrate this property 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 1.6 summarizes the mass and the distribution functions estimated with the FKRB and the generalized estimator.

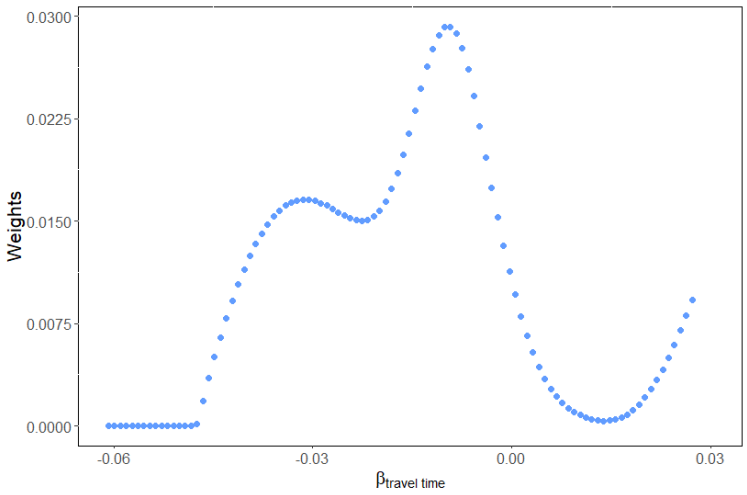
The generalized estimator estimates a smooth mass function whereas the FKRB exhibits LASSO-type behavior. The FKRB estimator only selects five out of 100 grid points whereas the generalized version selects 75 grid points.[12] Furthermore, it can easily be seen that the estimated mass function obtained 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 (2021) 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 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 A.
[12] We again define a weight as positive if it is greater than 10^{-3}.
[13] The results are available on request.

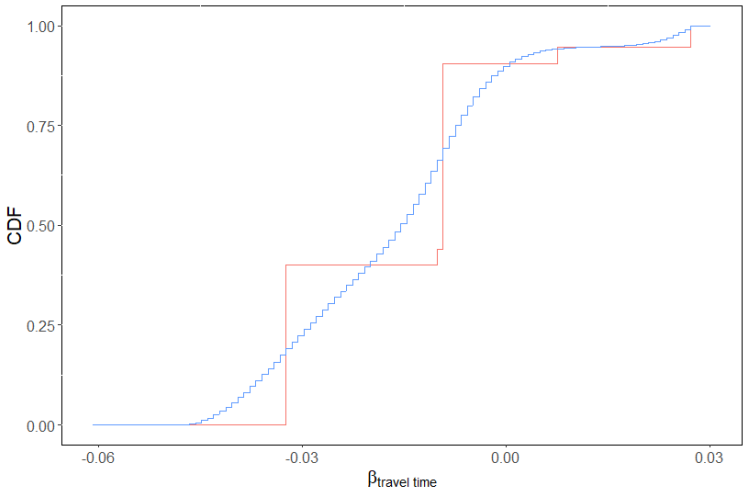
Figure 1.6: Estimated Distributions of Travel Time in Mode Canada Data with $R = 100$



(a) Mass Function for FKRB



(b) Mass Function for Generalized with OneSe



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

1.6 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 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.

[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.

Appendix A: Supplementary Tables

Table 1.3: 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 1.4: 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 1.5: 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 1.6: 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 1.7: 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 1.6 for (1) the FKRB estimator and elastic net estimator and (2) the semiparametric mixed logit with normal distribution and the elastic net estimator.

Appendix 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 $z_{i,j}^r$ more explicitly:

$$z_{i,j}^r := z_{i,j}(\beta^F, \beta_r^M) = g(x_{i,j}, \beta^F, \beta_r^M) = \frac{\exp\left(x_{i,j}^F \beta^F + x_{i,j}^M \beta_r^M\right)}{1 + \sum_{l=1}^{J} \exp\left(x_{i,l}^F \beta^F + x_{i,l}^M \beta_r^M\right)}. \tag{1.13}$$

The likelihood criterion given in Train (2008) is

$$LL(\beta^F, \beta^M) = \frac{1}{N} \sum_{i=1}^{N} \log\left(\sum_{r=1}^{R} \theta_r z_{i,y_i}^r\right) = \frac{1}{N} \sum_{i=1}^{N} \log\left(\sum_{r=1}^{R} \theta_r z_{i,y_i}(\beta^F, \beta_r^M)\right). \tag{1.14}$$

The probability of agent i having coefficients π conditional on her observed choice y_i and being type r is

$$h_{i,r}(\pi) = \frac{\theta_r z_{i,y_i}(\beta^F, \beta_r^M)}{\sum_{r=1}^{R} \theta_r z_{i,y_i}(\beta^F, \beta_r^M)}. \tag{1.15}$$

Based on Equation (1.15) 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

$$\begin{aligned} \pi^{t+1} &= \arg\max_{\pi} Q\left(\pi | \pi^t\right) \\ &= \arg\max_{\pi} \sum_{i=1}^{N} \sum_{r=1}^{R} h_{i,r}\left(\pi^t\right) \log\left(\theta_r z_{i,y_i}(\beta^F, \beta_r^M)\right). \end{aligned} \tag{1.16}$$

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 (1.16) separately for β^F and θ. Since we use our generalized estimator given in Equation (1.8), we only maximize Equation (1.16) over β^F:

$$\{\beta^F\}^{t+1} = \arg\max_{\beta^F} \sum_{i=1}^{N} \sum_{r=1}^{R} h_{i,r}\left(\pi^t\right) \log\left(z_{i,y_i}(\beta^F, \beta_r^M)\right). \tag{1.17}$$

Plugging Equation (1.13) into Equation (1.17) gives

$$\{\beta^F\}^{t+1} = \arg\max_{\beta^F} \sum_{i=1}^{N} \sum_{r=1}^{R} h_{i,r}\left(\pi^t\right) \log\left(\frac{\exp\left(x_{i,y_i}^F \beta^F + x_{i,y_i}^M \beta_r^M\right)}{1 + \sum_{l=1}^{J} \exp\left(x_{i,l}^F \beta^F + x_{i,l}^M \beta_r^M\right)}\right) \tag{1.18}$$

or equivalently

$$\{\beta^F\}^{t+1} = \arg\max_{\beta^F} \sum_{i=1}^{N}\sum_{j=1}^{J}\sum_{r=1}^{R} y_{i,j} h_{i,r}\left(\pi^t\right) \log\left(\frac{\exp\left(x_{i,j}^F\beta^F + x_{i,j}^M\beta_r^M\right)}{1+\sum_{l=1}^{J}\exp\left(x_{i,l}^F\beta^F + x_{i,l}^M\beta_r^M\right)}\right). \tag{1.19}$$

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 $h_{i,r}\left(\pi^t\right)$, calculated as given in Equation (1.15), 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 (1.8) 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 (1.8).
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 (1.19) .
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

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

Appendix C: Proofs of Results in Section 1.3

Below, we provide the proofs of the results presented in Section 1.3. 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 $A^T A$.

C.1: Proof of Probability Bound

Lemma 1 uses Hoeffding's inequality to derive a probability bound for sub-Gaussian random variables. We use the lemma in the proofs of Theorems 1 - 3.

Lemma 1. *Suppose Assumption 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

$$\mathbb{P}\left(\left\| \frac{1}{NJ} \tilde{Z}^T \epsilon \right\|_\infty \ge \gamma \right) = \mathbb{P}\left(\max_{1 \le r \le R-1} \left| \frac{1}{NJ} \sum_{i=1}^{N} \tilde{Z}_i^{rT} \epsilon_i \right| \ge \gamma \right) \tag{1.20}$$

where $\epsilon_i = (\epsilon_{i,1}, \ldots, \epsilon_{i,J})$ denotes a random vector of J dependent variables such that Equation (1.20) 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 1(iii) and Equation (1.10) 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 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,

$$\max_{\substack{1\leq r\leq R-1\\1\leq j\leq J}}\mathbb{P}\left(\left|\frac{1}{N}\sum_{i=1}^{N}\tilde{z}_{i,j}^{r}\epsilon_{i,j}\right|\geq\gamma\right) \leq 2\exp\left(-\frac{N\gamma^2}{2\sigma^2}\right). \tag{1.21}$$

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

$$\begin{aligned}
\mathbb{P}\left(\left\|\frac{1}{NJ}\tilde{Z}^{T}\epsilon\right\|_{\infty}\geq\gamma\right) &\leq (R-1)J\max_{\substack{1\leq r\leq R-1\\1\leq j\leq J}}\mathbb{P}\left(\left|\frac{1}{N}\sum_{i=1}^{N}\tilde{z}_{i,j}^{r}\epsilon_{i,j}\right|\geq\gamma\right)\\
&\leq 2(R-1)J\exp\left(-\frac{N\gamma^2}{2}\right).
\end{aligned} \tag{1.22}$$

□

C.2: Proof of Selection Consistency

In the following, we provide the proof of Theorem 1. We first derive two sufficient conditions in Lemma 3 that ensure that the estimated weights are equal in sign, i.e. $\hat{\theta} =_s \theta^*$. Lemma 4 provides a bound on the probability of the first sufficient condition and Lemma 5 a bound on the probability of the second sufficient condition. Finally, we use Lemma 4 and Lemma 5 to prove Theorem 1. Both Lemma 4 and Lemma 5 employ Lemma 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. *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

$$\frac{1}{\sqrt{NJ}} \tilde{Z}_S = ADM^T \tag{1.23}$$

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 (1.23) to rewrite

$$\begin{aligned} \left(\frac{1}{NJ} \tilde{Z}_S^T \tilde{Z}_S + \mu I_S \right)^{-1} = \left(MD^T A^T ADM^T + \mu I_S \right)^{-1} &= \left(MD^2 M^T + \mu MM^T \right)^{-1} \\ &= M \left(D^2 + \mu I_S \right)^{-1} M^T \end{aligned} \tag{1.24}$$

Therefore,

$$\begin{aligned} \left\| \left(\frac{1}{NJ} \tilde{Z}_S^T \tilde{Z}_S + \mu I_S \right)^{-1} \right\|_\infty &= \left\| M \left(D^2 + \mu I_S \right)^{-1} M^T \right\|_\infty \leq \sqrt{s} \left\| M \left(D^2 + \mu I_S \right)^{-1} M^T \right\|_2 \\ &= \sqrt{s} \left\| \left(D^2 + \mu I_S \right)^{-1} \right\|_2 = \sqrt{s} \max_{i \in S} \sqrt{\psi_i} \\ &= \sqrt{s} \max_{i \in S} \frac{1}{d_{ii}^2 + \mu} = \sqrt{s} \frac{1}{\min_{i \in S} d_{ii}^2 + \mu} = \sqrt{s} \frac{1}{\xi^S_{\min}(\mu)} \end{aligned} \tag{1.25}$$

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 (1.25) 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 (1.25) holds by the relation of singular values to eigenvalues. □

Lemma 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

$$\begin{aligned}
V &:= \frac{1}{NJ} \tilde{Z}_{S^C}^T \left[\tilde{Z}_S \left(\frac{1}{NJ} \tilde{Z}_S^T \tilde{Z}_S + \mu I_S \right)^{-1} \left(\lambda \iota_S + \mu \theta_S^* - \frac{1}{NJ} \tilde{Z}_S^T \epsilon \right) + \epsilon \right], \\
U &:= \left(\frac{1}{NJ} \tilde{Z}_S^T \tilde{Z}_S + \mu I_S \right)^{-1} \frac{1}{NJ} \tilde{Z}_S^T \epsilon, \\
\rho(\mu) &:= \min_{i \in S} \left| \left(\frac{1}{NJ} \tilde{Z}_S^T \tilde{Z}_S + \mu I_S \right)^{-1} \left(\frac{1}{NJ} \tilde{Z}_S^T \tilde{Z}_S \theta_S^* - \lambda \iota_S \right) \right|.
\end{aligned}$$

Proof. The Lagrangian of our generalized estimator in Equation (1.9) formulated in matrix notation is given by

$$L(\theta) := \quad \frac{1}{2NJ} ||\tilde{y} - \tilde{Z}\theta||_2^2 + \lambda \left(\iota^T \theta - 1 \right) + \frac{1}{2} \mu \ \theta^T \theta - \nu^T \theta \tag{1.26}$$

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 (1.26) 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 (1.26) are satisfied. The KKT conditions are given by

$$-\frac{1}{NJ} \tilde{Z}^T \left(\tilde{y} - \tilde{Z}\hat{\theta} \right) + \lambda \iota + \mu \ \hat{\theta} - \nu = 0, \tag{1.27}$$

$$\lambda \left(\iota^T \hat{\theta} - 1 \right) = 0, \tag{1.28}$$

$$\nu_r \ \hat{\theta}_r = 0, \tag{1.29}$$

$$\lambda \geq 0, \quad \nu_r \geq 0 \qquad\qquad \forall \quad r = 1, \ldots, R-1. \tag{1.30}$$

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 (1.27) require

$$-\frac{1}{NJ}\tilde{Z}_S^T\left(\tilde{y}-\tilde{Z}_S\hat{\theta}_S\right)+\lambda\iota_S+\mu\hat{\theta}_S-\nu_S=0, \tag{1.31}$$

$$-\frac{1}{NJ}\tilde{Z}_{S^C}^T\left(\tilde{y}-\tilde{Z}_S\hat{\theta}_S\right)+\lambda\iota_{S^C}-\nu_{S^C}=0. \tag{1.32}$$

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

$$\nu_r = 0 \qquad \forall \quad r \in S, \tag{1.33}$$

$$\nu_r \geq 0 \qquad \forall \quad r \notin S. \tag{1.34}$$

It follows from Condition (1.33) that Condition (1.31) simplifies to

$$-\frac{1}{NJ}\tilde{Z}_S^T\left(\tilde{y}-\tilde{Z}_S\hat{\theta}_S\right)+\lambda\iota_S+\mu\hat{\theta}_S=0. \tag{1.35}$$

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

$$-\frac{1}{NJ}\tilde{Z}_S^T\tilde{Z}_S\left(\theta_S^*-\hat{\theta}_S\right)-\frac{1}{NJ}\tilde{Z}_S^T\epsilon+\lambda\iota_S+\mu\hat{\theta}_S=0 \tag{1.36}$$

and

$$-\frac{1}{NJ}\tilde{Z}_{S^C}^T\tilde{Z}_S\left(\theta_S^*-\hat{\theta}_S\right)-\frac{1}{NJ}\tilde{Z}_{S^C}^T\epsilon+\lambda\iota_{S^C}-\nu_{S^C}=0. \tag{1.37}$$

Reformulating Condition (1.36) gives

$$\hat{\theta}_S=\underbrace{\left(\frac{1}{NJ}\tilde{Z}_S^T\tilde{Z}_S+\mu I_S\right)^{-1}\left(\frac{1}{NJ}\tilde{Z}_S^T\epsilon}_{=:U}+\frac{1}{NJ}\tilde{Z}_S^T\tilde{Z}_S\theta_S^*-\lambda\iota_S\right)>0 \tag{1.38}$$

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

Plugging Equation (1.38) into Equation (1.37) and using Condition (1.34) yields

$$\underbrace{\frac{1}{NJ}\tilde{Z}_{S^C}^T\left[\tilde{Z}_S\left(\frac{1}{NJ}\tilde{Z}_S^T\tilde{Z}_S+\mu I_S\right)^{-1}\left(\lambda\iota_S+\mu\theta_S^*-\frac{1}{NJ}\tilde{Z}_S^T\epsilon\right)+\epsilon\right]}_{=:V}\leq\lambda\iota_{S^C}. \tag{1.39}$$

U and V are defined in Equation (1.38) and Equation (1.39), 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 (1.39) 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

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

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 (1.38)) 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 (1.38) to hold and thus for $\hat{\theta}_S > 0$.

□

Lemma 4. *Suppose Assumption 1 holds. Suppose further that the NEIC holds. Let $\mathcal{M}^C(V)$ denote the complement of $\mathcal{M}(V)$. Then,*

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

Proof. V_j is sub-Gaussian with mean

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

Recall the Nonnegative Elastic Net Irrepresentable Condition (NEIC) is

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

Therefore, $\overline{V}_j \leq (1-\eta)\lambda$. Let $\tilde{V} := \frac{1}{NJ} \tilde{Z}_{S^C}^T \left[- \tilde{Z}_S \left(\frac{1}{NJ} \tilde{Z}_S^T \tilde{Z}_S + \mu I_S \right)^{-1} \frac{1}{NJ} \tilde{Z}_S^T + I_{NJ} \right] \epsilon$ 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 $\mathcal{M}^C(V)$:

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

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 and $\left\|\frac{1}{NJ}\tilde{Z}_{S^C}^T\tilde{Z}_S\right\|_\infty \leq s$ (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 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 4. 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

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

since $\eta\lambda > 0$.

Lemma 5. *Suppose Assumption 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 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 1 with $\gamma = \xi^S_{\min}(\mu)\frac{1}{\sqrt{s}}\rho(\mu)$. □

We use the above lemmata to prove Theorem 1.

Proof of Theorem 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 3.

Under the condition that RCDG holds, applying Lemma 4 and Lemma 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,

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

□

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 and the proof of Corollary 1. We then use the derived bound to proof the error bound of the estimated random coefficients' distribution in Theorem 3. In the proofs of Theorem 2 and Theorem 3, we apply Lemma 1.

Proof of Theorem 2.

Note that if $\hat{\theta}$ is the solution to the Lagrangian in Equation (1.26), it must hold that it minimizes (1.26), 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 (1.26), 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

$$\begin{aligned}&\frac{1}{2NJ}\left\|\tilde{Z}\left(\theta^*-\hat{\theta}\right)\right\|_2^2+\lambda\left(\iota^T\hat{\theta}-1\right)+\frac{\mu}{2}\hat{\theta}^T\hat{\theta}\leq\\&\frac{1}{NJ}\epsilon^T\tilde{Z}\left(\hat{\theta}-\theta^*\right)+\lambda\left(\iota^T\theta^*-1\right)+\frac{\mu}{2}\theta^{*T}\theta^*.\end{aligned}\tag{1.40}$$

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

$$\begin{aligned}\mathbb{P}\left(\left\|\frac{1}{NJ}\tilde{Z}^T\epsilon\right\|_\infty\geq\gamma\right)&\leq 2(R-1)J\exp\left(-N\left(\sqrt{\frac{2\log\left(\frac{2(R-1)J}{\delta}\right)}{N}}\right)^2\Big/2\right)\\&=2(R-1)J\exp\left(\log\left(\left(\frac{2(R-1)J}{\delta}\right)^{-1}\right)\right)\\&=\delta.\end{aligned}\tag{1.41}$$

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 (1.41). 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 (1.40) from above by

$$\frac{1}{2NJ}\left\|\tilde{Z}\left(\theta^*-\hat{\theta}\right)\right\|^2+\lambda\left(\iota^T\hat{\theta}-1\right)+\frac{\mu}{2}\hat{\theta}^T\hat{\theta}\leq\gamma\left\|\hat{\theta}-\theta^*\right\|_1+\lambda\left(\iota^T\theta^*-1\right)+\frac{\mu}{2}\theta^{*T}\theta^*.\tag{1.42}$$

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 (1.42) as

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

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

$$\begin{aligned}&\frac{1}{2NJ}\left\|\tilde{Z}\left(\theta^* - \hat{\theta}\right)\right\|_2^2 + \lambda\left(\left\|\hat{\theta}_S\right\|_1 + \left\|\hat{\theta}_{S^C}\right\|_1 - 1\right) + \frac{\mu}{2}\left(\hat{\theta}_S^T\hat{\theta}_S - \theta_S^{*T}\theta_S^* + \hat{\theta}_{S^C}^T\hat{\theta}_{S^C}\right) \leq \\ &\gamma\left\|\hat{\theta} - \theta^*\right\|_1 + \lambda\left(\left\|\theta_S^*\right\|_1 - 1\right).\end{aligned} \tag{1.43}$$

Note that the terms in (1.43) 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 (1.43) simplifies to

$$\frac{1}{2NJ}\left\|\tilde{Z}\left(\theta^* - \hat{\theta}\right)\right\|_2^2 + \frac{\mu}{2}\left(\hat{\theta}_S^T\hat{\theta}_S - \theta_S^{*T}\theta_S^* + \hat{\theta}_{S^C}^T\hat{\theta}_{S^C}\right) \leq \gamma\left\|\hat{\theta} - \theta^*\right\|_1. \tag{1.44}$$

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 (1.44) can be equivalently expressed as

$$\frac{1}{2NJ}\left\|\tilde{Z}\left(\theta^*-\hat{\theta}\right)\right\|_2^2+\frac{\mu}{2}\left\|\hat{\theta}-\theta^*\right\|_2^2\leq$$
$$\gamma\left\|\hat{\theta}-\theta^*\right\|_1+\frac{\mu}{2}\left(2\theta_S^{*T}\theta_S^*-2\theta_S^{*T}\hat{\theta}_S\right). \tag{1.45}$$

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

$$\theta_S^{*T}\left(\theta_S^*-\hat{\theta}_S\right)\leq\theta_S^{*T}\left|\hat{\theta}_S-\theta_S^*\right|\leq\left\|\theta_S^*\right\|_\infty\left\|\hat{\theta}_S-\theta_S^*\right\|_1\leq\sqrt{s}\left\|\theta_S^*\right\|_\infty\left\|\hat{\theta}_S-\theta_S^*\right\|_2 \tag{1.46}$$

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

Substituting Condition (1.46) back into the error bound in Equation (1.45) and using the fact that $||\hat{\theta}-\theta^*||_1\leq\sqrt{(R-1)}\;||\hat{\theta}-\theta^*||_2$, we can rewrite Equation (1.45) as

$$\frac{1}{2NJ}\left\|\tilde{Z}\left(\theta^*-\hat{\theta}\right)\right\|_2^2+\frac{\mu}{2}\left\|\hat{\theta}-\theta^*\right\|_2^2\leq\gamma\sqrt{(R-1)}\left\|\hat{\theta}-\theta^*\right\|_2+\mu\sqrt{s}\left\|\theta_S^*\right\|_\infty\left\|\hat{\theta}_S-\theta_S^*\right\|_2. \tag{1.47}$$

Recall that

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

and that the left-hand-side in Condition (1.47) can be summarized as

$$\frac{1}{2}\left(\hat{\theta}-\theta^*\right)^T\left[\frac{1}{NJ}\tilde{Z}^T\tilde{Z}+\mu I\right]\left(\hat{\theta}-\theta^*\right)\leq\left(\gamma\sqrt{(R-1)}+\mu\sqrt{s}\left\|\theta_S^*\right\|_\infty\right)\left\|\hat{\theta}-\theta^*\right\|_2. \tag{1.48}$$

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 (1.3.2)).

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 (1.48) 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:

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

□

Proof of Corollary 1.

By assumption, it holds that

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

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 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

$$\begin{aligned}\left|\hat{F}(\beta)-F^*(\beta)\right| &= \left|\sum_{r=1}^{R}\hat{\theta}_r\;1[\beta_r\leq\beta]-\sum_{r=1}^{R}\theta_r^*\;1[\beta_r\leq\beta]\right|\\ &\leq \sup_\beta\left|\sum_{r=1}^{R}\left(\hat{\theta}_r-\theta_r^*\right)\;1[\beta_r\leq\beta]\right|\\ &\leq \sum_{r=1}^{R}\left|\hat{\theta}_r-\theta_r^*\right| = \sum_{r=1}^{R-1}\left|\hat{\theta}_r-\theta_r^*\right|+\left|\hat{\theta}_R-\theta_R^*\right|\end{aligned}$$

where the last inequality holds by the triangle inequality.

Then,

$$\begin{aligned}\left|\hat{F}(\beta)-F^*(\beta)\right| &\leq \sum_{r=1}^{R-1}\left|\hat{\theta}_r-\theta_r^*\right|+\left|1-\sum_{r=1}^{R-1}\hat{\theta}_r-1+\sum_{r=1}^{R-1}\theta_r^*\right|\\ &= \sum_{r=1}^{R-1}\left|\hat{\theta}_r-\theta_r^*\right|+\left|\sum_{r=1}^{R-1}\left(\theta_r^*-\hat{\theta}_r\right)\right| \leq 2\sum_{r=1}^{R-1}\left|\hat{\theta}_r-\theta_r^*\right|\\ &= 2\left\|\hat{\theta}-\theta^*\right\|_1 \leq 2\sqrt{(R-1)}\left\|\hat{\theta}-\theta^*\right\|_2,\end{aligned}$$

which, by Theorem 2, can be bounded by

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

□

Chapter 2

A Sparse Grid Approach for the Nonparametric Estimation of High-Dimensional Random Coefficient Models

2.1 Introduction

Adequately modeling unobserved heterogeneous behavior of economic agents is a common challenge in many empirical economic studies. Random coefficient models are frequently applied to address this challenge. They allow the coefficients of the model to vary across agents according to an unknown distribution. Conventional parametric estimators typically assume that the random coefficients follow a certain family of distributions up to some unknown finite-dimensional parameters. However, such estimators lack flexibility as they are often limited to a few families of distributions, and are restrictive as they rely on the assumption that the assumed distribution is correct. Due to the increasing availability of large data sets, nonparametric estimators for random coefficient models become more and more attractive for applied research. These estimators allow to recover distributions from the data without such limiting prior assumptions on the shape of the distribution.

A popular nonparametric approach is the method of sieves (Chen, 2007). Sieve estimators approximate the underlying distribution using a finite number of basis functions that typically increase with the sample size. Unfortunately, sieve estimators can quickly become computationally unfeasible when the model includes multiple random coefficients. Because the standard way to extend one-dimensional basis functions to multi-dimensional functions is a tensor product construction, the number of parameters increases exponentially in the number of random coefficients. This property, known as the curse of dimensionality, limits the application of such estimators to models with only a few random coefficients – even if the number of basis functions in one dimension is moderately small (Chen, 2007).

This paper proposes and investigates a sparse grid approach for the nonparametric estimation of high-dimensional random coefficient models. The estimator approximates the underlying distribution using a linear combination of multi-dimensional hierarchical basis functions. The hierarchical structure of the basis functions has two major advantages: First, their local support makes it possible to accurately approximate the local peculiarities of the distribution without imposing certain functional forms in other regions. Second, they are particularly suited for the construction of sparse bases. For the construction of a sparse hierarchical basis, we adopt the sparse grid method suggested by Zenger (1991). The approach uses a truncated tensor product which reduces the number of basis functions substantially. Because smoother functions can typically be approximated by a smaller number of basis functions (Hansen, 2014), the truncated tensor product deteriorates the approximation accuracy only slightly if the underlying random coefficients distribution is sufficiently smooth (see, e.g. Bungartz and Griebel, 2004). In addition to the sparse tensor product construction, we study a spatially adaptive refinement procedure for estimating non-smooth distribution functions. Depending on the local shape of the underlying distribution, the spatially adaptive refinement incrementally adds basis functions in those areas of the distribution where the improvement in the overall approximation accuracy is highest. To provide a computationally simple and fast estimator, we exploit linearity using the linear probability model transformation suggested by Fox et al. (2011). This way, the parameters of the model can be estimated using constrained least squares.

We study the finite sample properties of our estimator in various Monte Carlo experiments. Using the nonparametric estimator of Fox et al. (2011) as a benchmark, our estimator provides

comparably accurate approximations of the true underlying distribution, even if the distribution has a steep and wiggly curvature. Moreover, the results confirm the theoretical properties of the sparse grid approach. The accuracy of the estimator slowly declines with an increasing number of random coefficients included into the model and with decreasing smoothness of the true distribution. For non-smooth distributions, the spatially adaptive refinement improves the approximation accuracy remarkably. Because the estimator becomes more accurate with increasing sample size if the number of basis functions is sufficiently large, the estimator can be viewed as a sieve estimator (Chen, 2007). An application to the model of dynamic regulation of air pollution in Blundell et al. (2020) emphasizes the advantage of our estimator. Blundell et al. (2020) estimate the five-dimensional distribution with the fixed grid estimator of Fox et al. (2011) using $10,001$ grid points. Even though our estimator requires substantially fewer parameters, the estimated results are similar to those of Blundell et al. (2020) – especially with respect to the estimated predictions of the conducted counterfactual experiments.

The underlying principle of sparse grids – a sparse tensor product decomposition – goes back to the seminal work of Smolyak (1963). Sparse grids for estimating nonlinear models in economics (including random coefficient models) have been studied by Heiss and Winschel (2008). In contrast to our estimator, the approach of Heiss and Winschel (2008) studies sparse grids in combination with quadrature rules for numerical integration, thereby restricting the approach to the parametric estimation of random coefficient models. Sparse grids in combination with hierarchical basis functions have been used in several research areas for function approximation and interpolation to overcome the curse of dimensionality. Among others, Ma and Zabaras (2009) employ the concept for the solution of stochastic differential equations (a frequent challenge in physics and engineering), Pflüger, Peherstorfer, and Bungartz (2010) for high dimensional classification problems (in data mining), and Peherstorfer, Pflüger, and Bungartz (2014) and Franzelin and Pflüger (2016) for nonparametric density estimation. The only application of sparse hierarchical bases in economics which we are aware of is by Brumm and Scheidegger (2017). They employ a sparse hierarchical basis for the interpolation of macroeconomic policy functions in dynamic optimization problems.

Our sparse grid estimator primarily relates to the nonparametric fixed grid estimator of Train (2008), Fox et al. (2011) and Heiss, Hetzenecker, and Osterhaus (2021), and to the nonparametric estimator of Train (2016). Both estimators use linear sieves to approximate the underlying random coefficients' distribution. The fixed grid approach uses a set of fixed support points and estimates the probability mass at every point from the data. The disadvantage of the approach is that the number of parameters equals the number of support points, leading to a large number of parameters if the estimated distribution is supposed to be smooth – especially if the model has multiple random coefficients. In fact, Fox et al. (2016) show that the fixed grid estimator suffers from the curse of dimensionality as the derived error bound of the estimated distribution function is less tight if the number of random coefficients increases. Train (2016) proposes to approximate the random coefficients' distribution using polynomials, splines or step functions as basis functions inside logit kernels to model the shape of the distribution. The logit kernel assures nonnegativity of the probability mass at each support point and summation to one. The parameters of the model are estimated with simulated maximum likelihood. In order to avoid an exponential increase in the number of basis functions, Train (2016) proposes to use mainly one-dimensional basis functions and

to include only few multi-dimensional basis functions to capture the correlation across dimensions. In contrast to the approach proposed in this paper, the approach proposed by Train (2016) lacks theoretical guidance on the choice of basis functions.

The remainder of the paper is organized as follows. Section 2.2 presents a nonparametric estimator for random coefficients distribution using a linear combination of basis functions. Section 2.3 explains the construction of sparse hierarchical bases, and Section 2.4 presents a spatially adaptive refinement procedure of the sparse hierarchical basis. Section 2.5 studies the performance of the estimator in several Monte Carlo experiments, and Section 2.6 presents an application to real data. Section 2.7 concludes.

2.2 Estimator

This section briefly lays out the random coefficients model and presents a computationally simple and fast nonparametric estimator that approximates the true distribution using a linear combination of basis functions. The estimator is general in the sense that it can be applied using any type of basis functions. The construction of sparse hierarchical bases is deferred until Section 2.2.

For the introduction of the estimator, consider the following random coefficient discrete choice model. Let there be an i.i.d. sample of N observations, each confronted with a set of J mutually exclusive potential outcomes, and an outside option. The researcher observes a D-dimensional real-valued vector of explanatory variables, $\boldsymbol{x}_{n,j} = (x_{n,j,1}, \ldots, x_{n,j,D})$, for every observation unit n and potential outcome j, and a unit vector $\boldsymbol{y}_i$ whose entries are equal to one when she observes outcome j for the nth observation unit, and zero otherwise.[1] Denote the probability of outcome j for a given covariate vector $\boldsymbol{x}_{n,j}$ and random coefficient vector $\boldsymbol{\beta}_n = (\beta_{n,1}, \ldots, \beta_{n,D}) \in \mathbb{R}^D$ by $g\left(\boldsymbol{x}_{n,j}, \boldsymbol{\beta}_n\right)$. The functional form of $g(\cdot)$ is specified by the researcher. Integrating the conditional outcome probability $g\left(\boldsymbol{x}_{n,j}, \boldsymbol{\beta}_n\right)$ over the distribution of random coefficients yields the unconditional probability that outcome j occurs for observation n given covariates $\boldsymbol{x}_{n,j}$,

$$P_{n,j}\left(\boldsymbol{x}\right) = \int_{\Omega_1} \cdots \int_{\Omega_D} g\left(\boldsymbol{x}_{n,j}, \boldsymbol{\beta}\right) f_0\left(\boldsymbol{\beta}\right) \mathrm{d}\beta_D \ldots \mathrm{d}\beta_1, \tag{2.1}$$

where $f_0(\boldsymbol{\beta}) : \Omega \to \mathbb{R}_+$ represents the joint probability density function of the unknown random coefficients' distribution with domain Ω.

The goal of the researcher is to estimate the unknown distribution from the data. A popular nonparametric approach for this task is the method of linear sieves. Linear sieves use a finite linear combination of prespecified basis functions (e.g., polynomials or splines) to approximate functions of unknown shapes. Define $\Phi_B := \{\phi_b\}_{b=1}^B$ as the finite set of such basis functions with corresponding approximation space V_B. The number of functions in Φ_B, B, and the shape are specified by the researcher.[2] Starting from the approximation of the true probability density function, $f_0(\boldsymbol{\beta})$,

[1] *Notation*: In the following, vectors and matrices will be written in bold.

[2] In order to select the domain of the basis, the researcher can use some preliminary estimates. For instance, estimating the distribution with a parametric approach first and then centering the grid at the mean estimates and taking multiple standard deviations to specify the domain.

through a linear combination of the basis functions in Φ_B,

$$f_0(\boldsymbol{\beta}) \approx \tilde{f}(\boldsymbol{\beta}) := \sum_{b=1}^{B} \alpha_b \phi_b(\boldsymbol{\beta}) \tag{2.2}$$

the approximated unconditional outcome probabilities are

$$P_{n,j}(\boldsymbol{x}) \approx \tilde{P}_{n,j}(\boldsymbol{x}) = \int_{\Omega_1} \cdots \int_{\Omega_D} g(\boldsymbol{x}_{n,j}, \boldsymbol{\beta}) \sum_{b=1}^{B} \alpha_b \phi_b(\boldsymbol{\beta}) \, \mathrm{d}\beta_D \ldots \mathrm{d}\beta_1, \tag{2.3}$$

where $\boldsymbol{\alpha} := (\alpha_1, \ldots, \alpha_B)'$ denotes the coefficient vector to be estimated. The $\approx$ arises from the approximation of the true joint probability density function $f_0(\boldsymbol{\beta})$ through $\tilde{f}(\boldsymbol{\beta})$. For the estimation of $\boldsymbol{\alpha}$, we adopt the approach of Fox et al. (2011) and transform Equation (2.3) into a linear probability model. Adding $y_{n,j}$ to both sides and moving $P_{n,j}$ to the right yields

$$y_{n,j} \approx \sum_{b=1}^{B} \alpha_b \int_{\Omega_1} \cdots \int_{\Omega_D} g(\boldsymbol{x}_{n,j}, \boldsymbol{\beta}) \phi_b(\boldsymbol{\beta}) \, \mathrm{d}\beta_D \ldots \mathrm{d}\beta_1 + (y_{n,j} - P_{n,j}(\boldsymbol{x})). \tag{2.4}$$

where we used the sum rule of integration, thereby restricting the summation terms in Equation (2.3) to be finite. Equation (2.3) reveals two computationally desirable properties: First, the coefficients α_b, $b = 1, \ldots, B$, are independent of the integral, implying that the integral needs to be simulated only once prior to the estimation. Second, the coefficients enter the unconditional choice probabilities linearly. For the estimation of $\boldsymbol{\alpha}$, we simulate the integral in Equation (2.4) using a finite set of nodes $\mathcal{B}_R := \{\boldsymbol{\beta}_r\}_{r=1}^{R}$ (e.g., using Halton or Sobol quasi-random sequences),

$$y_{n,j} \approx \sum_{b=1}^{B} \alpha_b \sum_{r=1}^{R} g(\boldsymbol{x}_{n,j}, \boldsymbol{\beta}_r) \phi_b(\boldsymbol{\beta}_r) + (y_{n,j} - P_{n,j}(\boldsymbol{x})). \tag{2.5}$$

The $\approx$ is now decomposed of the error from the approximation of $f_0(\boldsymbol{\beta})$ through $\tilde{f}(\boldsymbol{\beta})$, and the approximation error arising from the numerical simulation of the integral.[3] The property that $\boldsymbol{\alpha}$ enters Equation (2.5) linearly allows to estimate the coefficients with constrained least squares, which is easy to implement and computationally fast. The binary outcome vector $\boldsymbol{y} = (y_{1,1}, \ldots, y_{1,J}, \ldots, y_{N,J})$ denotes the dependent variable and $\sum_{r=1}^{R} g(\mathbf{x}_{n,j}, \boldsymbol{\beta}_r)\phi_b(\boldsymbol{\beta}_r)$ the bth regressor – the regression in total has NJ observations, J "regression observations" for every statistical observation unit $n = 1, \ldots, N$ and B regressors. In order to estimate a valid distribution function, we estimate the coefficient vector $\boldsymbol{\alpha}$ subject to the constrains that $\tilde{f}(\boldsymbol{\beta})$ is nonnegative and has unit integrand,

$$\hat{\boldsymbol{\alpha}} = \underset{\boldsymbol{\alpha} \in \Lambda}{\arg\min} \frac{1}{2NJ} \sum_{n=1}^{N} \sum_{j=1}^{J} \left(y_{n,j} - \sum_{b=1}^{B} \alpha_b \sum_{r=1}^{R} g(\boldsymbol{x}_{n,j}, \boldsymbol{\beta}_r) \phi_b(\boldsymbol{\beta}_r) \right)^2 \tag{2.6}$$

where $\Lambda := \{\boldsymbol{\alpha} \in \mathbb{R}^D : \sum_{b=1}^{B} \alpha_b \phi_b(\boldsymbol{\beta}_r) \geq 0 \; \forall \; \boldsymbol{\beta}_r \in \mathcal{B}_R, \sum_{b=1}^{B} \alpha_b \sum_{r=1}^{R} \phi_b(\boldsymbol{\beta}_r) = 1\}$.[4] By the

[3] By the strong law of large numbers, the approximated integral over of the basis function converges weakly to its analytic solution such that the latter approximation error approaches zero (Train, 2009) if R is sufficiently large.

[4] The estimator relates to the smooth basis densities estimator proposed in Fox et al. (2011). They propose to approximate the true distribution through a mixture of normal densities, and estimate the probability weight of each normal subject to the constraint that the weights are nonnegative and sum to one. The proposed estimator is a

definition of choice probabilities, the expected value of the composite error term $y_{n,j} - P_{n,j}(\boldsymbol{x}_{n,j})$ conditional on $\boldsymbol{x}_{n,j}$ is zero, such that the regression model satisfies the mean-independence assumption of the least squares approach (Fox et al., 2011). The optimization problem stated in Equation (2.6) is convex and has a single global optimum if the basis functions in Φ_B are linearly independent. It can be solved with common statistic software using specialized optimization routines (e.g., R's solve.QP function from the quadprog package or MATLAB's lsqlin function).

The estimated joint distribution function at point $\boldsymbol{\beta}$ is constructed from the weighted sum of the estimated coefficients and basis functions,

$$\hat{F}(\boldsymbol{\beta}) = \sum_{b=1}^{B} \hat{\alpha}_b \sum_{r=1}^{R} 1[\boldsymbol{\beta}_r \leq \boldsymbol{\beta}] \, \phi_b(\boldsymbol{\beta}_r), \tag{2.7}$$

where $1[\boldsymbol{\beta}_r \leq \boldsymbol{\beta}]$ is an indicator function that is equal to one whenever $\boldsymbol{\beta}_r \leq \boldsymbol{\beta}$, and zero otherwise. The term to the right of coefficient α_b corresponds to the simulated integral of the corresponding basis function $\phi_b(\cdot)$ with upper bound $\boldsymbol{\beta}$ using R simulation nodes. The estimated distribution approximates the true underlying distribution through a discrete distribution with R support points and probability weight $\hat{f}(\boldsymbol{\beta}_r) = \sum_{b=1}^{B} \hat{\alpha}_b \phi_b(\boldsymbol{\beta}_r)$ at every point $r = 1, \ldots, R$.

For the estimation of multi-dimensional random coefficients distributions, the multi-dimensional bases are typically constructed using a regular tensor product of one-dimensional basis functions. Starting from a one-dimensional basis with B basis functions, the D-dimensional regular tensor product basis includes B^D basis functions (Chen, 2007). Because the exponential dependency renders the approach computationally unfeasible for high-dimensional distributions, the above estimator with a regular tensor basis is limited to moderately low-dimensional random coefficient models.

Remark 1. The proposed estimator can be easily extended to a generalized least-squares version and a simulated maximum likelihood version. For the generalized least-squares version, each "regression observation" in Equation (2.6) is weighted by a weighting matrix to address the heteroscedasticity problem associated with linear probability models and the correlation across observations that belong to the same observation unit n. For a detailed description of the calculation of an efficient weighting matrix, see Fox et al. (2011).

As an alternative to constrained least squares, the coefficients $\boldsymbol{\alpha}$ can be estimated with simulated maximum likelihood using the approach of Train (2016), who proposes to model the probability weight at support point β_r using a linear combination of basis functions inside a logit kernel. The exponential function in the logit kernel assures that the estimated weights are positive. The denominator normalizes the probability weights such that they sum up to one.

Remark 2. When choosing the family of basis functions and the number of simulation draws R, it is important that the basis functions are linearly independent, and that the researcher chooses R to be sufficiently large such that the draws are sufficient to cover the domain densely. If the number of simulation draws is too small, there are only a few simulation draws inside the support

special case of nonnegative lasso (see Heiss et al. (2021) for more details), leading to sparse solutions. In contrast, our estimator does not relate to the lasso and, hence, does not suffer from sparsity.

of every basis functions with the consequence that most column entries are zero. This property can lead to an ill-conditioning of the least squares problem (Judd, Maliar, and Maliar, 2011).[5] One tool recommended in the literature to improve the numerical stability of least squares problems is Thikonov regularization (Hoerl and Kennard, 1970) (see, e.g., Judd et al., 2011, Cohen, Davenport, and Leviatan, 2013, or Pflüger et al., 2010), which is already successfully used by Heiss et al. (2021) to improve the performance of the nonparametric fixed grid estimator of Fox et al. (2011).[6]

2.3 Sparse Hierarchical Bases

This section explains the construction of sparse hierarchical bases. Because the sparse grid is based on a truncated tensor product of one-dimensional hierarchical basis functions, we start with the concept of hierarchical basis functions, and then explain how sparse grids can be constructed from multi-dimensional hierarchical bases. For a more comprehensive presentation of hierarchical bases and sparse grids, see, e.g., Bungartz and Griebel (2004) and Garcke (2013).

2.3.1 Hierarchical Multilevel Bases

Hierarchical bases are based on a decomposition of the approximation space into a finite number of hierarchically structured segments – intervals in the univariate case and hyper-rectangles in the multivariate case. These segments are constructed via a discretization of the domain Ω of the function under consideration using equidistant grids. In the following, we consider the D-dimensional unit cube, $\Omega = [0,1]^D$ for ease of notation. The construction of the hierarchical basis can be easily adapted to different domains via rescaling. Furthermore, we assume that f_0 is vanishing on the boundary of Ω ($f_0|_{\delta\Omega} = 0$).[7]

Let $l \in \mathbb{N}$ denote the discretization level specified by the researcher. In the one-dimensional case, the grid Ω_l with points $b_{l,i} := 2^{-l}\ i$ and mesh size $h_l := 2^{-l}$ splits the domain Ω into 2^l equally-sized intervals. The index $i \in \mathbb{N}$ indicates the location of a grid point. Every grid point is associated with a basis function $\phi : [0,1] \to \mathbb{R}$ that is centered at the corresponding grid point. For the construction of the sparse grid basis, we consider the piecewise-linear hat function

$$\phi(\beta) := \begin{cases} 1 - |\beta|\,, & \text{if } \beta \in [-1,1] \\ 0, & \text{otherwise.} \end{cases} \tag{2.8}$$

Using translation and scaling according to level l and index i, the basis function centered at grid

[5] In addition to the multicollinearity problem, choosing fewer simulation draws can also lead to poor scaling of the regressor matrix. If there are only a few simulation draws inside the support of every basis function, the columns of the regressor matrix have only a few very small entries, in which case they are treated as if they are columns of zeros (Judd et al., 2011).

[6] We noticed that the instability of the estimator for high levels when the distribution is estimated with ordinary least squares disappears when the coefficients are estimated with constrained least squares. To this end, the constraints seem to constitute a form of regularization that stabilizes the estimator and potentially makes additional regularization redundant.

[7] The restriction that f_0 is vanishing at the boundary of Ω can be overcome by adding basis functions that are nonzero at the boundaries (for more details, see, e.g., Pflüger (2010)). Train (2016) points out that it can be beneficial to restrict the function to be zero at the boundaries of the domain as this eliminates the long tails of some distributions, e.g., of the normal or lognormal distribution, which can be unrealistic in real-world applications.

point $b_{l,i}$ is

$$\phi_{l,i}(\beta) := \phi\left(\frac{\beta - b_{l,i}}{h_l}\right) \tag{2.9}$$

with $\phi_{l,i}(b_{l,i}) = 1$ and local support $[b_{l,i} - h_l, b_{l,i} + h_l]$.

To construct a basis with hierarchically arranged functions, the locations of the grid points – and the number of basis functions within a level – are determined by the index sets

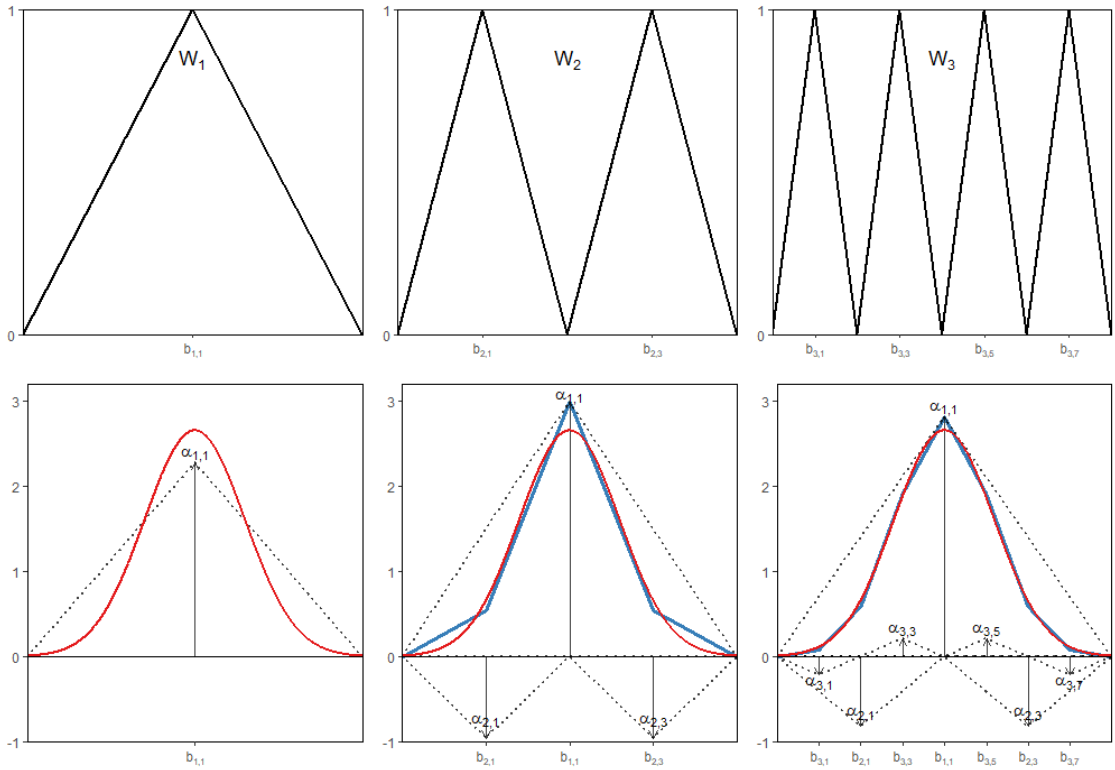
$$\mathcal{I}_l := \left\{ i \in \mathbb{N} : \ 1 \leq i \leq 2^l - 1, i \text{ odd} \right\}. \tag{2.10}$$

All basis functions with level l centered at the grid points corresponding to index set $\mathcal{I}_l$ span the hierarchical subspace W_l,

$$W_l := \text{span}\left\{\phi_{l,i} : i \in \mathcal{I}_l\right\}. \tag{2.11}$$

The upper panel in Figure 2.1 illustrates the one-dimensional piecewise-linear hierarchical subspaces W_l going from level $l = 1$ (left) to level $l = 3$ (right). All hierarchical basis functions with the same

Figure 2.1: One-Dimensional Piecewise-linear Hierarchical Basis Functions



Note: The top panel shows the one-dimensional hierarchical subspaces W_l for $l = 1$ (left), $l = 2$ (center), and $l = 3$ (right). The bottom panel illustrates the approximation of a univariate normal density (solid red line) with mean 0.5 and standard deviation 0.15, $f(\beta) = \phi(\beta|0.5, 0.15^2)$, through a one-dimensional piecewise-linear hierarchical basis Φ_l with levels $1, 2, 3$ (solid blue line). The contribution of every basis function to the approximation is indicated by the grey arrows.

level have the same size, shape and compact support. While the number of basis functions that span a subspace increase with the level l of the subspace, the support of each function decreases with l. The index sets $\mathcal{I}_l$ ensure that (i) different basis functions within the same level have mutually disjoint support, and (ii) that the support of a basis function with level l nests the support of two

basis functions of the next higher level, $l + 1$.

The hierarchical basis of level l is the set of all basis functions with level $1 \leq k \leq l$ and corresponding index $i \in \mathcal{I}_k$,

$$\Phi_l := \{\phi_{k,i} : \ i \in \mathcal{I}_k, 1 \leq k \leq l\}. \tag{2.12}$$

The bottom panel in Figure 2.1 illustrates the approximation of a univariate normal density using a one-dimensional piecewise-linear hierarchical basis. Due to the hierarchical structure, hierarchical bases of different levels are nested such that the basis of a level l refines a basis of the next lower level. The smaller support of basis functions with a higher level allows to approximate local peculiarities more accurately. The shape of the approximated function depends on the shape of the specified type of basis function. For instance, the one-dimensional piecewise-linear hierarchical basis approximates the true probability density function on every segment by a linear function.

Starting from the one-dimensional hierarchical basis, a D-dimensional hierarchical basis on $\Omega = [0, 1]^D$ is obtained via a tensor product construction. Let the multi-index $\boldsymbol{l} = (l_1, \ldots, l_D) \in \mathbb{N}^D$ denote the discretization level of the hierarchical basis in every dimension, and $\boldsymbol{i} \in \mathbb{N}^D$ indicate the spatial position of the D-dimensional grid points. In the following, all relational operations involving vectors are to be read component-wise. The D-dimensional grid $\Omega_{\boldsymbol{l}}$ with grid points $\boldsymbol{b_{l,i}} := (b_{l_1,i_1}, \ldots, b_{l_D,i_D})$ and mesh size $\boldsymbol{h_l} = (h_{l_1}, \ldots, h_{l_D})$ can be constructed from the cartesian product of one-dimensional grids in every dimension. Accordingly, the indices i_d and l_d can vary across d for a given grid point. The grid points are equidistant in each dimension but can differ across dimensions (e.g., the grid can be finer in more important dimensions).

As in the one-dimensional case, every grid point spans a basis function with support on the respective segment. The D-dimensional basis function centered at grid point $\boldsymbol{b_{l,i}}$ is defined as the product of one-dimensional basis functions,

$$\phi_{\boldsymbol{l},\boldsymbol{i}}(\boldsymbol{\beta}) := \prod_{d=1}^{D} \phi_{l_d,i_d}(\beta_d). \tag{2.13}$$

The left panel in Figure 2.2 illustrates the tensor product construction of a two-dimensional piecewise-bilinear hierarchical basis function with level $\boldsymbol{l} = (2, 1)$ and index $\boldsymbol{i} = (1, 1)$ from the one-dimensional piecewise-linear basis function $\phi_{2,1}$ in dimension $d = 1$ and $\phi_{1,1}$ in dimension $d = 2$ (dashed black lines).

The multivariate hierarchical subspaces are defined analogously to the univariate case,

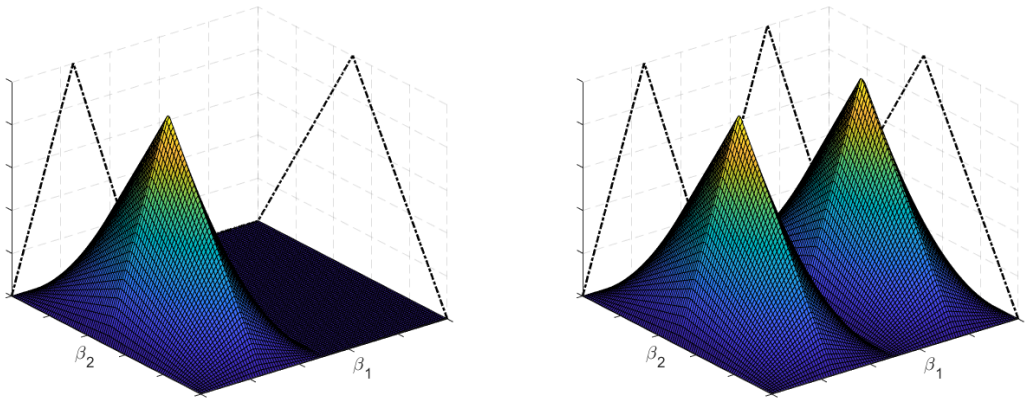
$$W_{\boldsymbol{l}} := \operatorname{span}\{\phi_{\boldsymbol{l},\boldsymbol{i}} : \ \boldsymbol{i} \in \mathcal{I}_{\boldsymbol{l}}\}, \quad \mathcal{I}_{\boldsymbol{l}} := \mathcal{I}_{l_1} \times \cdots \times \mathcal{I}_{l_D}, \tag{2.14}$$

where the D-dimensional index sets $\mathcal{I}_{\boldsymbol{l}}$ can be constructed as the cartesian product of the one-dimensional index sets. The right panel in Figure 2.2 shows the hierarchical subspace $W_{(2,1)}$ which is spanned by the basis functions $\phi_{(2,1),(1,1)}$ and $\phi_{(2,1),(3,1)}$.

The D-dimensional hierarchical basis is the set of all basis functions with level $\boldsymbol{1} \leq \boldsymbol{k} \leq \boldsymbol{l}$ and index $\boldsymbol{i} \in \mathcal{I}_{\boldsymbol{k}}$

$$\Phi_{\boldsymbol{l}} := \{\phi_{\boldsymbol{k},\boldsymbol{i}} : \boldsymbol{i} \in \mathcal{I}_{\boldsymbol{k}}, \boldsymbol{1} \leq \boldsymbol{k} \leq \boldsymbol{l}\}. \tag{2.15}$$

Figure 2.2: Two-Dimensional Piecewise-bilinear Hierarchical Basis Functions



Note: The left panel illustrates the tensor product construction of the two-dimensional piecewise-bilinear hierarchical basis function for level $\boldsymbol{l} = (2, 1)$ and index $\boldsymbol{i} = (1, 1)$ from the one-dimensional piecewise-linear basis function $\phi_{2,1}$ in dimension $d = 1$ and $\phi_{1,1}$ in dimension $d = 2$ (dashed black lines). The right panel shows the hierarchical subspace $W_{(2,1)} = \{\phi_{(2,1),(1,1)}, \phi_{(2,1),(3,1)}\}$.

The approximated function $\tilde{f}_{\boldsymbol{l}} \in V_{\boldsymbol{l}}$ is a linear combination of D-dimensional hierarchical basis functions with coefficients $\alpha_{\boldsymbol{k},\boldsymbol{i}} \in \mathbb{R}$,

$$f_0(\boldsymbol{\beta}) \approx \tilde{f}_{\boldsymbol{l}}(\boldsymbol{\beta}) := \sum_{\boldsymbol{k}=\boldsymbol{1}}^{\boldsymbol{l}} \sum_{\boldsymbol{i} \in I_{\boldsymbol{k}}} \alpha_{\boldsymbol{k},\boldsymbol{i}} \phi_{\boldsymbol{k},\boldsymbol{i}}(\boldsymbol{\beta}) \tag{2.16}$$

where $V_{\boldsymbol{l}}$ denotes the function space spanned by the hierarchical basis functions. Due to the linear independence across piecewise-linear hierarchical basis functions, the underlying function f_0 can be uniquely approximated through $\tilde{f}_{\boldsymbol{l}}$ (Valentin and Pflüger, 2016).

Bungartz and Griebel (2004) show that the full hierarchical basis includes $|V_{\boldsymbol{l}}| = (2^l - 1)^D = \mathcal{O}(2^{lD})$ basis functions. The full hierarchical basis has the same limitation as other existing nonparametric estimators for random coefficient models. Due to the regular tensor product construction of multi-dimensional basis functions from one-dimensional basis functions, the number of parameters increases exponentially in the number of random coefficients, prohibiting an accurate approximation of the underlying distribution if the model includes multiple random coefficients. For instance, for a 5-dimensional distribution and level $l = 3$ in each dimension, estimating the model using the full hierarchical basis involves the estimation of $(2^3 - 1)^5 = 16,807$ parameters.

2.3.2 Classical Sparse Grids

To alleviate the curse of dimensionality, sparse grids seek to construct an approximation space that is better than the full grid space $V_{\boldsymbol{l}}$ in the sense that the same number of basis functions leads to a higher approximation accuracy. The classical sparse grid approach (Zenger, 1991) takes advantage of the hierarchical nature of the basis functions. Starting from the definition of the approximation

space $V_{\boldsymbol{l}}$ as a direct sum of hierarchical subspaces,[8]

$$V_{\boldsymbol{l}} := \bigoplus_{\boldsymbol{k} \leq \boldsymbol{l}} W_{\boldsymbol{k}}, \tag{2.17}$$

the approach reduces the number of basis functions by selecting only those subspaces that contribute most to an accurate approximation. The selection of subspaces arises from a discrete optimization that weights the approximation benefit of a hierarchical subspace - measured in terms of its contribution to the overall approximation in the L_2 norm - against its cost - measured in terms of the number of parameters (Bungartz and Griebel, 2004).

The cost of subspace $W_{\boldsymbol{l}}$ can be immediately derived from its corresponding index set $\mathcal{I}_{\boldsymbol{l}}$, and is given by $|W_{\boldsymbol{l}}| = |\mathcal{I}_{\boldsymbol{l}}| = 2^{|\boldsymbol{l}-\mathbf{1}|_1}$. The contribution of a subspace to the approximation accuracy depends on the smoothness of the function under consideration, or more precisely, on its function class. The Classical sparse grid is derived for functions that are assumed to be sufficiently smooth, i.e., with bounded second-order mixed derivatives. This function class belongs to the mixed Sobolov space (of functions vanishing on the boundary)[9]

$$\mathcal{H}^2_{\text{mix}}(\Omega) := \left\{ f : \Omega \to \mathbb{R} : D^{\boldsymbol{r}} f \in L_2(\Omega), |\boldsymbol{r}|_\infty \leq c, f|_{\partial\Omega} = 0 \right\} \tag{2.18}$$

with $|\boldsymbol{r}|_\infty := \max_{1 \leq d \leq D} r_d$ and smoothness parameter $c = 2$, where $\mathcal{D}^{\boldsymbol{r}}$ denotes the differential operator defined by

$$\mathcal{D}^{\boldsymbol{r}} := \frac{\partial^{\boldsymbol{r}}}{\partial \beta_1^{r_1} \cdots \partial \beta_D^{r_D}} \tag{2.19}$$

given a D-tuple $\boldsymbol{r} = (r_1, \ldots, r_D)$ of nonnegative integers. Recall that in the representation of the probability density function as a weighted sum of hierarchical basis functions, the coefficient $\alpha_{\boldsymbol{k},\boldsymbol{i}}$ indicates the refinement of the local approximation constructed with those functions of the next lower level, $\boldsymbol{l} - \mathbf{1}$, through the function with level $\boldsymbol{l}$. Bungartz and Griebel (2004) show for functions $f \in \mathcal{H}^2_{\text{mix}}(\Omega)$ that the coefficients in the representation of the underlying function as a linear combination of piecewise-linear hierarchical basis functions decay rapidly as the level increases,

$$|\alpha_{\boldsymbol{l},\boldsymbol{i}}| = \mathcal{O}\left(2^{-2|\boldsymbol{l}|_1}\right), \tag{2.20}$$

where $|\boldsymbol{l}|_1 := \sum_{d=1}^{D} l_d$. Thus, the decreasing support of basis functions with increasing level together with the decay of the coefficients imply a decreasing contribution of subspaces with higher level if the underlying function is sufficiently smooth.

The classical sparse grid leaves out those subspaces within the full grid space $V_{\boldsymbol{l}}$ that contribute only little to the function approximation, i.e., for which the absolute values of the coefficients are small. This is done via an a priori optimization which minimizes the approximation error (measured

[8]The space $V_{\boldsymbol{l}}$ is the direct sum of subspaces $W_{\boldsymbol{k}}$, $\mathbf{1} \leq \boldsymbol{k} \leq \boldsymbol{l}$, if $V_{\boldsymbol{l}} = W_{\mathbf{1}} + \ldots + W_{\boldsymbol{k}}$ and the subspaces $\{W_{\mathbf{1}}, \ldots, W_{\boldsymbol{l}}\}$ are disjoint (Gentle, 2007, p. 48).

[9]Assuming a certain smoothness class of functions often required in the nonparametric estimation literature when studying the approximation accuracy of estimators (see, e.g., Chen (2007) for different smoothness classes). For example, when deriving the error bound for the nonparametric fixed grid estimator of Fox et al. (2011), Fox et al. (2016) assume that the density of the true underlying random coefficients distribution is a s-times continuously differentiable density function with all own and partial derivatives uniformly bounded (with respect to the L_2-norm) by a constant $\bar{C} < \infty$. Their derived error bound on the true and estimated distribution is tighter if the true probability density function is smoother.

by the L_2 norm) while keeping the number of grid points fixed. For functions in the mixed Sobolov space $\mathcal{H}_2^{\text{mix}}(\Omega)$, this yields the classical sparse grid space

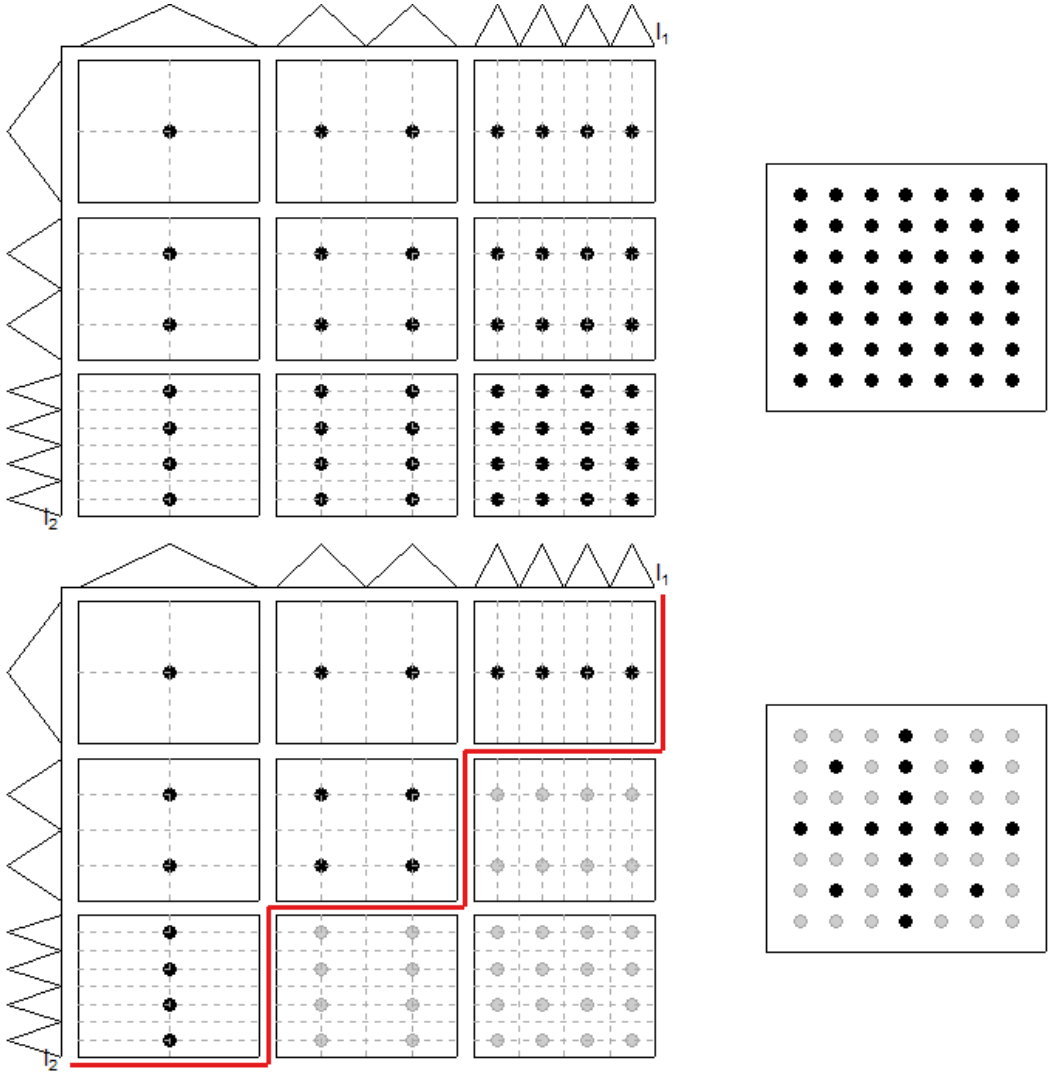
$$V_{l_S} := \bigoplus_{|\boldsymbol{k}|_1 \leq l_S + D - 1} W_{\boldsymbol{k}}. \tag{2.21}$$

of level $l_S \in \mathbb{N}$ specified by the researcher. Note that the set of basis functions that spans V_{l_S} is now decomposed of only those functions from subspaces for which $|\boldsymbol{k}|_1 \leq l_S + D - 1$ for every $\mathbf{1} \leq \boldsymbol{k} \leq \boldsymbol{l}$,

$$\Phi_{l_S} := \{\phi_{\boldsymbol{k},\boldsymbol{i}} : \boldsymbol{i} \in \mathcal{I}_{\boldsymbol{k}}, |\boldsymbol{k}|_1 \leq l_S + D - 1\}. \tag{2.22}$$

Thus, the sparse grid approximation space V_{l_S} is a subset of the full grid approximation space with discretization level l_S in every dimension, $V_{l_S} \subset V_{\boldsymbol{l}}$ with $\boldsymbol{l} = (l_S, \ldots, l_S)$. Figure 2.3 illustrates the construction of a two-dimensional classical sparse grid of level $l_S = 3$. The number of basis

Figure 2.3: Two-Dimensional Full and Sparse Grid for Level $l = 3$



Note: The upper panel illustrates the construction of a two-dimensional full Cartesian grid of level $\boldsymbol{l} = (3, 3)$, and the lower panel the construction of a classical sparse grid of level $l_S = 3$.

functions in the sparse hierarchical basis is

$$|V_{l_S}| = \sum_{i=0}^{l_S-1} 2^i \cdot \binom{D-1+i}{D-1} = 2^{l_S} \left(\frac{l_S^{D-1}}{(D-1)!} + \mathcal{O}\left(l_S^{D-2}\right) \right). \tag{2.23}$$

Expressed in terms of the number of basis functions in one dimension, $B = 2^l - 1$, Equation 2.23 implies that the function space spanned by the sparse hierarchical basis is of order $\mathcal{O}(B \log(B)^{D-1})$, compared to $\mathcal{O}(B^D)$ for regular tensor product bases (Bungartz and Griebel, 2004, Brumm and Scheidegger, 2017). Table 2.1 reports the number of basis functions in a sparse hierarchical basis for different dimensions and discretization levels and for a regular tensor product basis with $B = (3, 5, 7)$ basis functions in one dimension. Clearly, the estimation of the model with a tensor product basis rapidly becomes computationally unfeasible in five- and higher-dimensional problems for more than three basis functions in one dimension. The sparse grid approach renders the estimation of the corresponding random coefficients' distributions with such discretization levels computationally feasible.

Table 2.1: Number of Grid Points in Full Cartesian Grid vs. Sparse Grid

	Sparse Grid			Tensor Product Basis		
Dimension	$\|V_2\|$	$\|V_3\|$	$\|V_4\|$	$B = 3$	$B = 5$	$B = 7$
2	5	17	49	9	25	49
3	7	31	111	27	125	343
4	9	49	209	81	625	2401
5	11	71	351	243	3125	16807
6	13	97	545	729	15,625	117,649
8	17	161	1,121	6561	390,625	$5.76 \cdot 10^6$
10	21	241	2,001	59,049	$9,77 \cdot 10^6$	$2.82 \cdot 10^8$

Analogously to the full hierarchical basis, the approximated function $\tilde{f}_{l_S} \in V_{l_S}$ corresponds to the finite weighted sum of hierarchical basis functions centered at the sparse grid points,

$$f_0(\boldsymbol{\beta}) \approx \tilde{f}_{l_S}(\boldsymbol{\beta}) = \sum_{|\boldsymbol{k}|_1 \leq l_S + D - 1} \sum_{\boldsymbol{i} \in I_{\boldsymbol{k}}} \alpha_{\boldsymbol{i},\boldsymbol{k}} \phi_{\boldsymbol{i},\boldsymbol{k}}(\boldsymbol{\beta}). \tag{2.24}$$

Bungartz and Griebel (2004) show that the approximation accuracy of functions constructed with the sparse piecewise-linear hierarchical basis deteriorates only slightly from $\mathcal{O}(2^{-2\boldsymbol{l}})$ for the full hierarchical basis to $\mathcal{O}(2^{-2l_S} \cdot l_S^{D-1})$ if the function under consideration is sufficiently smooth.

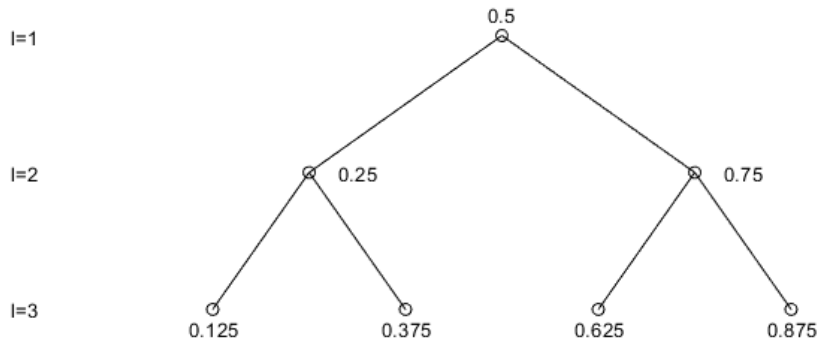
Remark 3. The construction of sparse grids is not restricted to piecewise-linear functions but can be constructed using several different types of basis functions. For instance, Valentin and Pflüger (2016) consider cardinal B-Splines. If the spline functions are of odd degree, the knots of the cardinal B-Spline basis coincide with the grid points of the hierarchical basis. In fact, the hat function corresponds to the cardinal B-Spline of degree one. For more information on alternative basis functions, see, e.g., Bungartz and Griebel (2004) and Pflüger (2010), respectively.

2.4 Spatially Adaptive Refinement

The classical sparse grid contains those basis functions that are optimal in the sense that they deteriorate the approximation accuracy only slightly if the function to be approximated has bounded second-order mixed derivatives. For functions outside this class, i.e., functions with a wigglier and steeper curvature, spatially adaptive refinement can be used to further increase the approximation accuracy. Starting from the sparse grid, the refinement procedure incrementally adds basis functions to subregions where the underlying distribution is characterized by a steep curvature (Pflüger, 2010).

Due to the nested support of hierarchical basis functions of different levels, hierarchical bases are particularly suited for spatially adaptive refinement procedures. Recall that the support of a basis function of level $l-1$ is subdivided among the basis functions of the next finer level l (for each function, multiple functions of the next finer level exist). Thus, by adding additional basis functions of the next finer level, one can refine the basis in some regions without affecting the approximation accuracy in others. Figure 2.4 illustrates this tree-like structure of a one-dimensional hierarchical basis of level $l = 3$. Each grid point in the tree always serves as origin for two new points. Suppose that the point 0.25, a point of level $l = 2$, is selected for refinement. The spatially adaptive approach adds the two neighboring grid points at location 0.125 and 0.375 of the next higher level, $l = 3$. These newly added points span basis functions with disjoint support that cover half of the support of the basis function spanned at 0.25 (cf. Section 2.3). For the adaptive

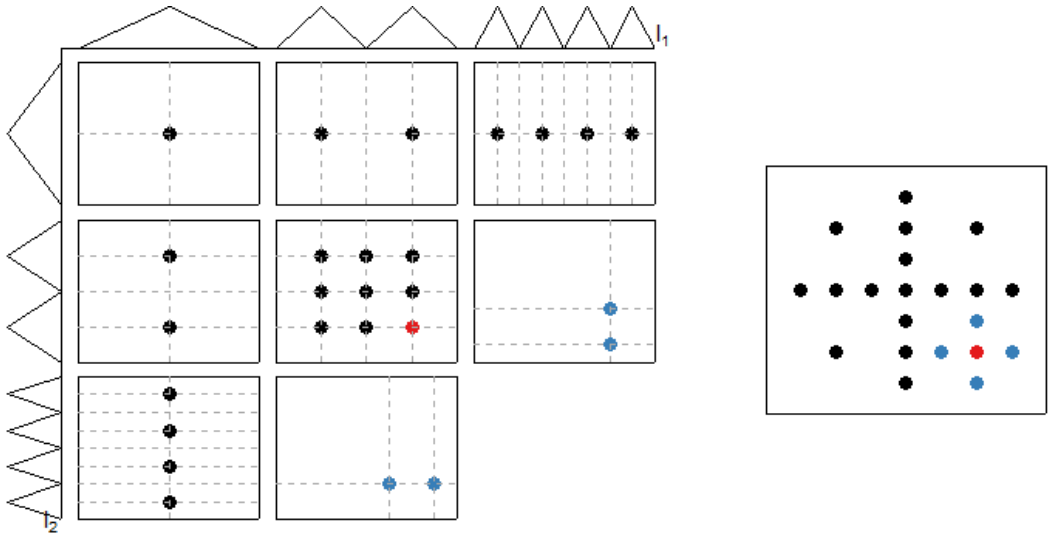
Figure 2.4: One-dimensional Tree-like Structure of Full Grid of Level $l_S = 3$



refinement of a D-dimensional hierarchical basis, in every dimension, all neighboring points of the next higher level that are not yet included into the grid are added, thus keeping the coordinates of the refined grid point in the remaining dimensions fixed. This way, at most $2D$ points are added to the current grid for every point refined. Figure 2.5 illustrates the spatially adaptive procedure for a two-dimensional sparse grid of level $l_S = 3$. The red point is the grid point selected for refinement, and the four blue points are those added to the current grid. The basis functions spanned at the newly added points extend the current sparse basis.

Refinable in the hierarchical structures are only grid points for which at least one neighboring point of the next higher level does not exist yet in any of the dimensions. To keep the hierarchical structure of the basis consistent, all originating points of the new grid that are not yet included have to be added if not already included. This can lead to a scenario where more than $2D$ points are added per refinement step (Pflüger, 2010).

Figure 2.5: Spatially Adaptive Refinement of Two-Dimensional Sparse Grid of Level $l_S = 3$



Note: The figure illustrates the spatially adaptive refinement of a sparse grid of level $l_S = 3$. The red point represents the grid point that is selected for refinement, and the blue points represent the grid points that are added to the initial sparse grid.

The challenge of the spatially adaptive refinement is to select those grid points for refinement that lead to the largest improvement in the approximation accuracy. A possible but computationally intensive and inefficient strategy is to consider every current grid point for refinement separately, and add only those candidates that contribute most to the problem's solution (according to a suitable error measure).[10] However, already in the one-dimensional case, there are as many grid points to be considered on the next higher level as potential new grid points as there are in the current grid. The model has to be estimated for each of these points, of which many are unlikely to be relevant (Pflüger, 2010).

A more efficient strategy is the identification of refinement candidates based on information available on the current grid. A refinement criterion commonly used in applications is the absolute value of the estimated coefficients $\alpha_{\boldsymbol{k},\boldsymbol{i}}$ (see, e.g., Pflüger et al., 2010 and Brumm and Scheidegger, 2017). Recall that in the reconstruction of the underlying function f_0 as weighted sum of hierarchical basis functions, the coefficient $\alpha_{\boldsymbol{k},\boldsymbol{i}}$ represents the local variation of f_0 at the area around the corresponding grid point $\boldsymbol{b}_{\boldsymbol{k},\boldsymbol{i}}$. Accordingly, refining grid points with the largest absolute value of the corresponding coefficient first promotes the refinement of those regions where the local variation of f_0 is strong (Pflüger, 2010).[11] Another criterion suggested by Pflüger (2010) specifically for regression tasks is the contribution of each basis function to the squared estimated local error. Let

$$\hat{\epsilon}^2_{n,j} := (y_{n,j} - \sum_{b=1}^{B} \hat{\alpha}_b \sum_{r=1}^{R} g(\boldsymbol{x}_{n,j}, \boldsymbol{\beta}_r)\phi_b(\boldsymbol{\beta}_r))^2$$

denote the squared estimated local error for observation unit n and alternative j following from

[10] The strategy relates to the concept of backward deletion in optimal knot search for spline functions (e.g., see Wand, 2000).

[11] Peherstorfer et al. (2014) suggest to weight the absolute value of the estimated hierarchical coefficients $\hat{\alpha}_{\boldsymbol{k},\boldsymbol{i}}$, $|\boldsymbol{k}|_1 \leq l_S + D - 1$ and $\boldsymbol{i} \in \mathcal{I}_{\boldsymbol{k}}$, by the function value of the basis function.

the regression in (2.6). The grid point that centers the basis function with the largest contribution to the squared local error,

$$c_{\boldsymbol{l},\boldsymbol{i}} := \sum_{n=1}^{N} \sum_{j=1}^{J} |\hat{\alpha}_{\boldsymbol{l},\boldsymbol{i}} \sum_{r=1}^{R} g(\boldsymbol{x}_{n,j}, \boldsymbol{\beta}_r) \phi_{\boldsymbol{l},\boldsymbol{i}}(\boldsymbol{\beta}_r) \hat{\epsilon}_{n,j}^2|,$$

is refined first, where R corresponds to the number of simulation draws,and $\hat{\boldsymbol{\alpha}} := (\hat{\alpha}_1, \ldots, \hat{\alpha}_B)\prime$ to the vector of estimated coefficients (the estimation of the coefficients is explained in Section 2.2). We employ the criterion in the Monte Carlo experiments presented in the subsequent section, where the spatially adaptive refinement substantially improves the approximation accuracy of the sparse hierarchical basis estimator.

In addition to the refinement criterion, the researcher has to specify the number of refinement steps alongside the number of grid points refined per step. The choice depends on the problem and the data at hand. On the one hand, refining more than one grid point at once leads to a broader refinement, which can help to circumvent the refinement from getting stuck in a single characteristic of the underlying function. On the other hand, refining too many points at once expedites the increase in the number of points (especially in cases of high dimensional problems) which can cause overfitting (especially for small data sets) (Pflüger, 2010).

Selecting the number of points and the refinement steps relates to a model selection task. For sieve series models, Hansen (2014) presents a variety of different model selection procedures, the most prominent being the Akaike information criterion (AIC) and cross-validation.[12] While cross-validation techniques have the advantage that they take the out-of-sample fit into account to avoid over-fitting, the AIC is computationally less expensive, which is an advantage in high-dimensional problems. In the Monte Carlo experiments presented in Section 2.5, we studied both k-fold cross-validation and the AIC for the selection of the number of refinement steps. The results indicate that both AIC and k-fold cross-validation appear to be suitable criteria leading to an improved approximation accuracy of the sparse hierarchical basis estimator when the local squared error is used for the selection of the grid points to be refined.

2.5 Monte Carlo Simulations

This section studies the finite sample properties of the sparse grid estimator in several Monte Carlo experiments using true random coefficient distributions of varying smoothness and dimensionality. The experiments apply the estimator to a random coefficients logit model with individual-level discrete choice data.[13] The model is widely used in applied econometrics to study discrete choices of economic agents among a finite number of alternatives. In this model, every observation unit

[12]Selecting the number of grid points and total refinement steps relate to the selection of the number of knots in spline regression. Another model selector than the AIC and cross-validation that is commonly used in this literature is generalized cross-validation (e.g., see Zhou and Shen, 2001, Ruppert, Wand, and Carroll, 2003). For the spatially adaptive refinement of sparse grids, Pflüger (2010) suggests using k-fold cross-validation and to refine the hierarchical basis as long as the out-of-sample fit decreases. This approach is also used in spline regression where it is known as the myopic algorithm (the out-of-sample fit is typically measured via generalized cross-validation). We employ a full-search algorithm (Ruppert et al., 2003, pp. 127-128) in our Monte Carlo experiments, which calculates the out-of-sample fit for every refined grid and then selects the model with the lowest out-of-sample fit.

[13]For a detailed description of the random coefficients multinomial logit, see Train, 2009, pp. 134-150.

n makes a single discrete choice among J mutually exclusive alternatives (and an outside option). Observation units pick the alternative that realizes the highest utility. Let $u_{n,j} = \boldsymbol{x}_{n,j}^T \boldsymbol{\beta}_n + \omega_{n,j}$ denote the utility from alternative j, given covariates $\boldsymbol{x}_{n,j}$ and unobserved individual-specific preferences $\boldsymbol{\beta}_n$. The random variable $\omega_{n,j}$ denotes an additive, consumer- and choice-specific error term. Observation unit n chooses alternative j if $u_{n,j} > u_{n,k}$ for all $k \neq j$ (and $u_{n,0} = \omega_{n,0}$). Under the assumption that $\omega_{n,j}$ is i.i.d. type I extreme value across alternatives and observation units, the unconditional choice probabilities, $P_{n,j}(\boldsymbol{x})$, are of the form

$$P_{n,j}(\boldsymbol{x}) = \int_{\Omega_1} \cdots \int_{\Omega_D} \frac{\exp\left(\boldsymbol{x}_{n,j}^T \boldsymbol{\beta}\right)}{1 + \sum_{j=1}^{J} \exp\left(\boldsymbol{x}_{n,j}^T \boldsymbol{\beta}\right)} f(\boldsymbol{\beta}) \, \mathrm{d}\beta_D \ldots \mathrm{d}\beta_1. \tag{2.25}$$

In our experiments, the observation units choose among $J = 5$ mutually exclusive alternatives and an outside option. We estimate the model for different sample sizes $N = 1000, 10000$, and number of random coefficients $D = 2, 4, 6$. We draw the entries of the D-dimensional vectors of alternative-specific characteristics, $\boldsymbol{x}_{n,j}$, independently from a $\mathcal{N}(0,1)$ for every observation unit n and alternative j. In order to study the performance of the sparse grid estimator for distributions of varying smoothness, we consider two alternative distributions for the true random coefficients distribution. The first experiment generates the random coefficients $\boldsymbol{\beta}$ from a mixture of two multivariate normals,

$$0.5 \cdot \mathcal{N}\left(\boldsymbol{\mu}^{(1)}, \boldsymbol{\Sigma}^{(1)}\right) + 0.5 \cdot \mathcal{N}\left(\boldsymbol{\mu}^{(2)}, \boldsymbol{\Sigma}^{(2)}\right),$$

where the entries of the D-dimensional mean vectors are $\mu_d^{(1)} = -1.5$ and $\mu_d^{(2)} = 1.5$ for $d = 1, \ldots, D$. The variance matrices $\boldsymbol{\Sigma}^{(1)} = \boldsymbol{\Sigma}^{(2)}$ have entries $\Sigma_{dd}^{(1)} = 0.4$ on the main diagonal and $\Sigma_{dk}^{(1)} = 0.1$ on the off-diagonal, i.e., for $d \neq k$. The second experiment considers a more sophisticated and less smooth distribution. It generates the random coefficients from a mixture of four multivariate normals,

$$0.25 \cdot \mathcal{N}\left(\boldsymbol{\mu}^{(1)}, \boldsymbol{\Sigma}^{(1)}\right) + \ldots + 0.25 \cdot \mathcal{N}\left(\boldsymbol{\mu}^{(4)}, \boldsymbol{\Sigma}^{(4)}\right),$$

with $\mu_d^{(1)} = -2.5$, $\mu_d^{(2)} = -0.8$, $\mu_d^{(3)} = 0.8$, and $\mu_d^{(4)} = 2.5$ for $d = 1, \ldots, D$. The variance matrices of the second design, $\boldsymbol{\Sigma}^{(m)}$, $m = 1, \ldots, 4$, are $1/4$ times the variance matrix of the first design, implying a steeper curvature. Figure 2.6 and Figure 2.7 display the bivariate joint distribution functions of the mixture of two normals, and the mixture of four normals, respectively. Due to the smaller variance and the higher number of mixture components, the mixture of four multivariate normals has a steeper and wigglier curvature, which, in theory, is more difficult to recover for the sparse grid estimator. Figure 2.10 in the appendix shows the true joint probability densities of true distributions.

For every distribution, we generate 200 data sets for every combination of N and D. For every data set, we estimate the random coefficients' distribution using the sparse grid and the spatially adaptive sparse grid estimator. The sparse grids are constructed on $\Omega = [-4, 4]^D$ for levels $l_S = (2, 3, 4)$. The support covers the true support with a coverage probability close to one (at least 0.998). The hierarchical bases are constructed using piecewise D-linear hat functions. We simulate the integral using quasi-random number sequences. To ensure proper coverage of the true distributions' support, we let the number of simulation draws R increase with the dimension of the

true distribution, i.e., we use $R = D \cdot 2000$ Halton draws. We also conducted the experiments using the mexican hat function (cf. Pflüger, 2010). Table 2.7 in the appendix presents the results which are similar to those obtained with the piecewise-linear hat function. For the spatially adaptive refinement, we conduct 10 refinement steps, whereby the maximum discretization level is 5. In every refinement step, we select the grid point (among those grid points that can be updated) with the largest contribution to the local squared error (c.f. Section 2.4). We select the number of refinement steps using 5-fold cross-validation, whereby the final spatially adaptive sparse grid estimator uses the refined grid that achieves the lowest out-of-sample MSE. In addition, we studied the performance of the spatially adaptive refinement when the number of refinement steps is selected based on the out-of-sample log-likelihood, and the AIC. The results are quite similar for all three criteria as indicated by the results in Table 2.5 in the appendix.

As a benchmark, we estimate the random coefficients distribution using the nonparametric estimator of Fox et al. (2011). The estimator uses a fixed grid of random coefficients instead of basis functions for approximating the underlying distribution. To assure a certain comparability across the estimators in terms of the number of parameters, we use the same number of grid points, $q = 3, 7, 15$, in every dimension as the full hierarchical basis has basis functions for $l = 2, 3, 4$. We construct the D-dimensional grid points from the cartesian product of the one-dimensional points. This is in line with Fox et al. (2011), who recommend increasing the number of grid points exponentially with D. For $D = 4$, we can only estimate the random coefficients distribution with $q = 3, 7$ points in every dimension, and for $D = 6$ with only $q = 3$ grid points. Using more grid points in these setups is not possible as the number of parameters exceeds the sample size.

We assess the estimators' approximation accuracy using the root mean integrated squared error (RMISE) from Fox et al. (2011). Denote the estimated distribution function in Monte Carlo run m evaluated at β_e by $\widehat{F}_m(\beta_e)$, and the true distribution by $F_0(\beta_e)$. The RMISE averages the squared difference between the true and estimated distribution at a fixed set of evaluation points across all Monte Carlo runs,

$$\text{RMISE} = \sqrt{\frac{1}{200}\sum_{m=1}^{200}\left[\frac{1}{E}\sum_{e=1}^{E}\left(\widehat{F}_m(\beta_e) - F_0(\beta_e)\right)^2\right]}.$$

We use a uniform grid with $E = 10^D$ points spread on $[-4, 4]^D$ for the evaluation. All calculations are conducted with the statistical software R (R Core Team, 2018).

The left part in Table 2.2 presents the average RMISE across the Monte Carlo replicates for the fixed grid estimator (FKRB), the sparse grid estimator (SG), and the spatially adaptive sparse grid estimator (ASG) for the mixture of two normals, while the right part the results for the mixture of four normals. The sparse grid and the spatially adaptive sparse grid estimator achieve more accurate approximations of the true random coefficients distributions than the FKRB estimator, independent of the dimension, sample size, and the refinement level/number of fixed grid points – even though the FKRB estimator uses a substantially greater number of parameters in higher dimensions. The difference in the approximation accuracy is particularly large for $D = 6$, where the FKRB estimator cannot use more than 3 grid points in every dimension.

Table 2.2: Average Number of Parameters and RMISE over 200 Monte Carlo Replicates for Mixture of 2 and Mixture of 4 Normals

		Mixture of 2 normals						Mixture of 4 normals					
		Parameters			RMISE			Parameters			RMISE		
N	q/l_S	FKRB	SG	ASG	FKRB	SG	ASG	FKRB	SG	ASG	FKRB	SG	ASG
Dimension $D = 2$													
1000	3/2	9	5	35.6	0.2067	0.0736	0.0514	9	5	39.1	0.1955	0.0881	0.0577
1000	7/3	49	17	41.6	0.0993	0.0479	0.0536	49	17	48.3	0.1022	0.0473	0.0589
1000	15/4	225	49	70.4	0.0912	0.0475	0.0561	225	49	72.9	0.0951	0.0549	0.0624
10,000	3/2	9	5	42.5	0.2039	0.0718	0.0280	9	5	50.1	0.1934	0.0863	0.0435
10,000	7/3	49	17	52.9	0.0843	0.0418	0.0290	49	17	63.6	0.0854	0.0434	0.0418
10,000	15/4	225	49	76.3	0.0581	0.0313	0.0303	225	49	84.5	0.0648	0.0519	0.0394
Dimension $D = 4$													
1000	3/2	81	9	149.7	0.2254	0.0983	0.0583	81	9	153.7	0.2328	0.1201	0.0634
1000	7/3	2401	49	213.8	0.1273	0.0620	0.0598	2401	49	226.9	0.1241	0.0850	0.0632
1000	15/4	.	209	341.7	.	0.0502	0.0569	.	209	341.2	.	0.0691	0.0577
10,000	3/2	81	9	144.1	0.2226	0.0979	0.0409	81	9	151.6	0.2316	0.1197	0.0536
10,000	7/3	2401	49	193.8	0.0787	0.0613	0.0386	2401	49	215.2	0.0915	0.0846	0.0511
10,000	15/4	.	209	341.5	.	0.0492	0.0361	.	209	379.6	.	0.0685	0.0480
Dimension $D = 6$													
1000	3/2	729	13	276.2	0.2156	0.0853	0.0569	729	13	266.4	0.2441	0.1099	0.0733
1000	7/3	.	97	386.9	.	0.0643	0.0547	.	97	416.8	.	0.0909	0.0626
1000	15/4	.	545	954.0	.	0.0610	0.0622	.	545	1024.9	.	0.0866	0.0643
10,000	3/2	729	13	246.9	0.2138	0.0849	0.0577	729	13	233.7	0.2441	0.1096	0.0774
10,000	7/3	.	97	207.1	.	0.0640	0.0625	.	97	208.1	.	0.0906	0.0888
10,000	15/4	.	545	1055.3	.	0.0605	0.0687	.	545	1099.1	.	0.0862	0.0568

Note: The table reports the total number of parameter and the RMISE for the FKRB estimator, the sparse grid estimator (SG), and the adaptive sparse grid estimator (ASG). The adaptive sparse grid estimator performs five refinement steps, whereby the final number of refinements is determined based on the lowest out-of-sample mean squared error calculated with five-fold cross-validation. The grid point to be refined in every refinement step is selected according to its contribution to the local squared error.

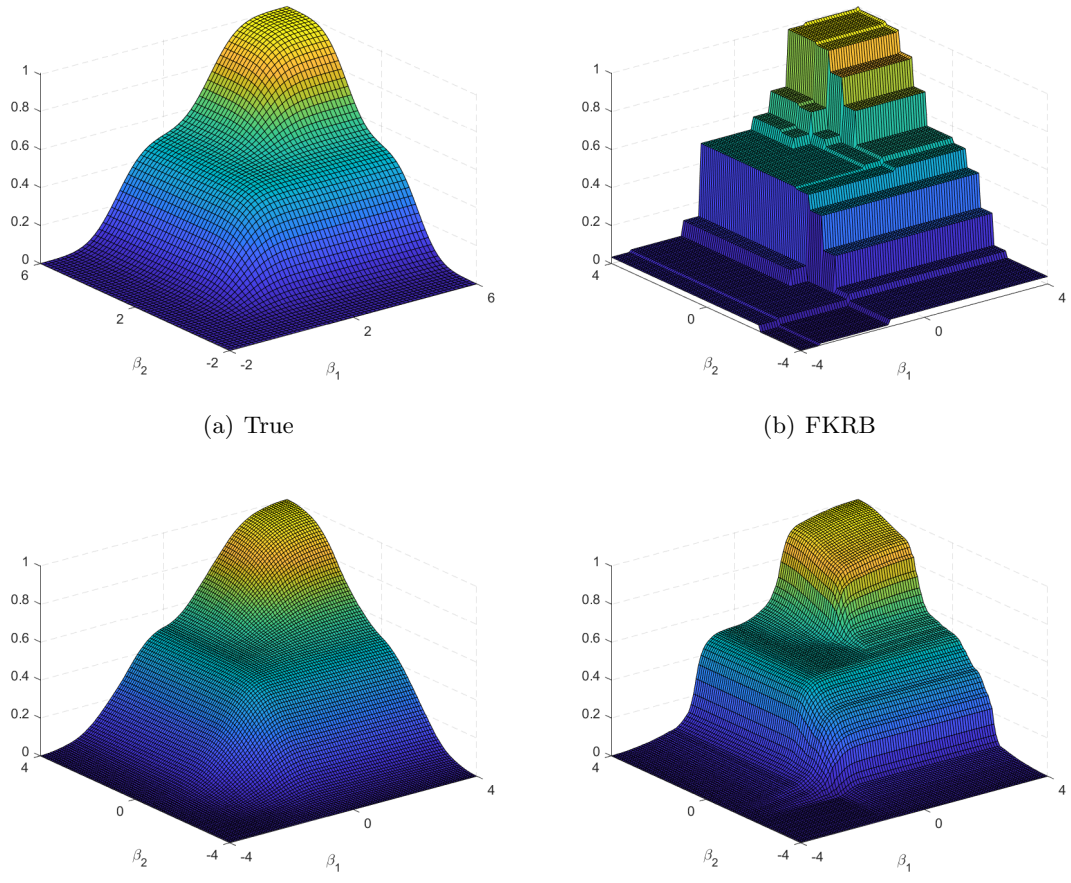
The discrepancy in the approximation accuracy can be explained by the FKRB estimator's relation to the lasso (cf. Heiss et al., 2021). Due to this relation, the estimator provides sparse solutions that lead to approximations through step functions with only a few steps. Figure 2.6 and Figure 2.7 illustrate this property for the bivariate joint mixture of two normals and bivariate joint mixture of four normals, respectively. In contrast to the FKRB estimator, the sparse grid and the spatially adaptive sparse grid estimator provide smooth approximations due to the substantially greater number of simulation draws compared to fixed grid points used by the FKRB estimator.

Overall, the results for the sparse grid estimator presented in Table 2.2 confirm the theoretical properties of the sparse hierarchical basis outlined by, e.g., Bungartz and Griebel (2004), as follows: (i) The estimator becomes more accurate with increasing levels – except for $D = 2$ and the mixture of four normals, where the RMISE is larger for $l_S = 4$ than for $l_S = 3$, which appears to be the consequence of over-fitting as indicated by the out-of-sample log-likelihood reported in Table 2.6.

(ii) The approximation accuracy declines with an increasing number of random coefficients (except for $l_S = 2$ and when going from $D = 4$ to $D = 6$). And (iii), the sparse grid estimator is less precise when approximating the mixture of four normals than the mixture of two normals due to the steeper and wigglier curvature of the former (except for $D = 2$ and $l_S = 3$) – which is also the case for the FKRB estimator.

The limited ability to accurately approximate non-smooth distributions is illustrated by the estimated bivariate joint CDFs in Figure 2.6 and Figure 2.7. The visual inspection of the estimated CDF of the mixture of two normals indicates that the sparse grid estimator is able to accurately approximate the smooth curvature of the true distribution. Inspecting the estimated mixture of four normals reveals that the estimator cannot recover the steep and wiggly shape of the true distribution. This can be explained by the limited number of basis functions with sufficiently small support in every dimension (i.e., basis functions with a high level in every dimension). The spatially adaptive sparse grid estimator, in contrast, which incrementally adds basis functions of higher levels, is able to approximate such a curvature accurately as illustrated by the estimated joint CDF.

Figure 2.6: True and Estimated Bivariate Joint CDF of Mixture of 2 Normals for $N = 10,000$



(a) True

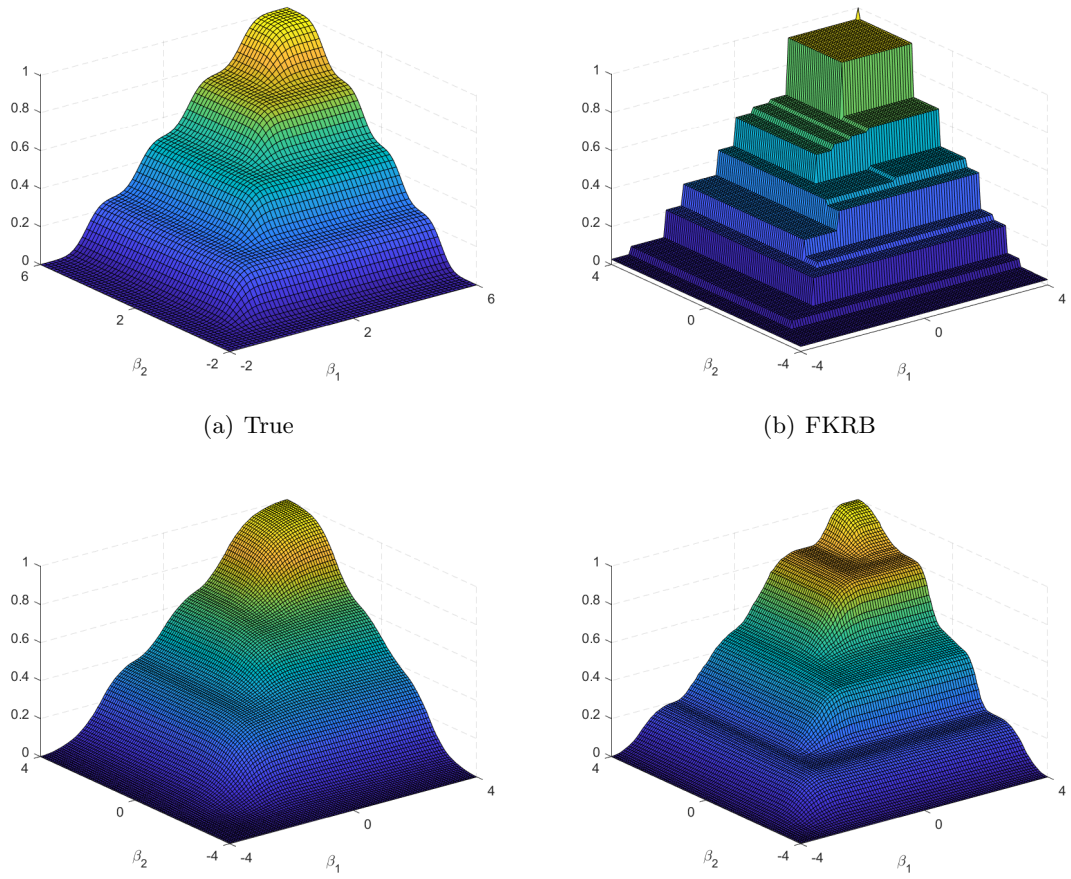
(b) FKRB

(c) SG

(d) ASG

Note: Estimated bivariate joint distribution function for the mixture of two normals estimated with the FKRB estimator using 225 grid points, with the sparse grid estimator of level $l_S = 4$ (SG), and the spatially adaptive sparse grid estimator (ASG). The number of refinement steps is selected using 5-fold cross-validation and based on the lowest out-of-sample mean squared error.

Figure 2.7: True and Estimated Bivariate Joint CDF of Mixture of 4 Normals for $N = 10,000$



(a) True

(b) FKRB

(c) SG

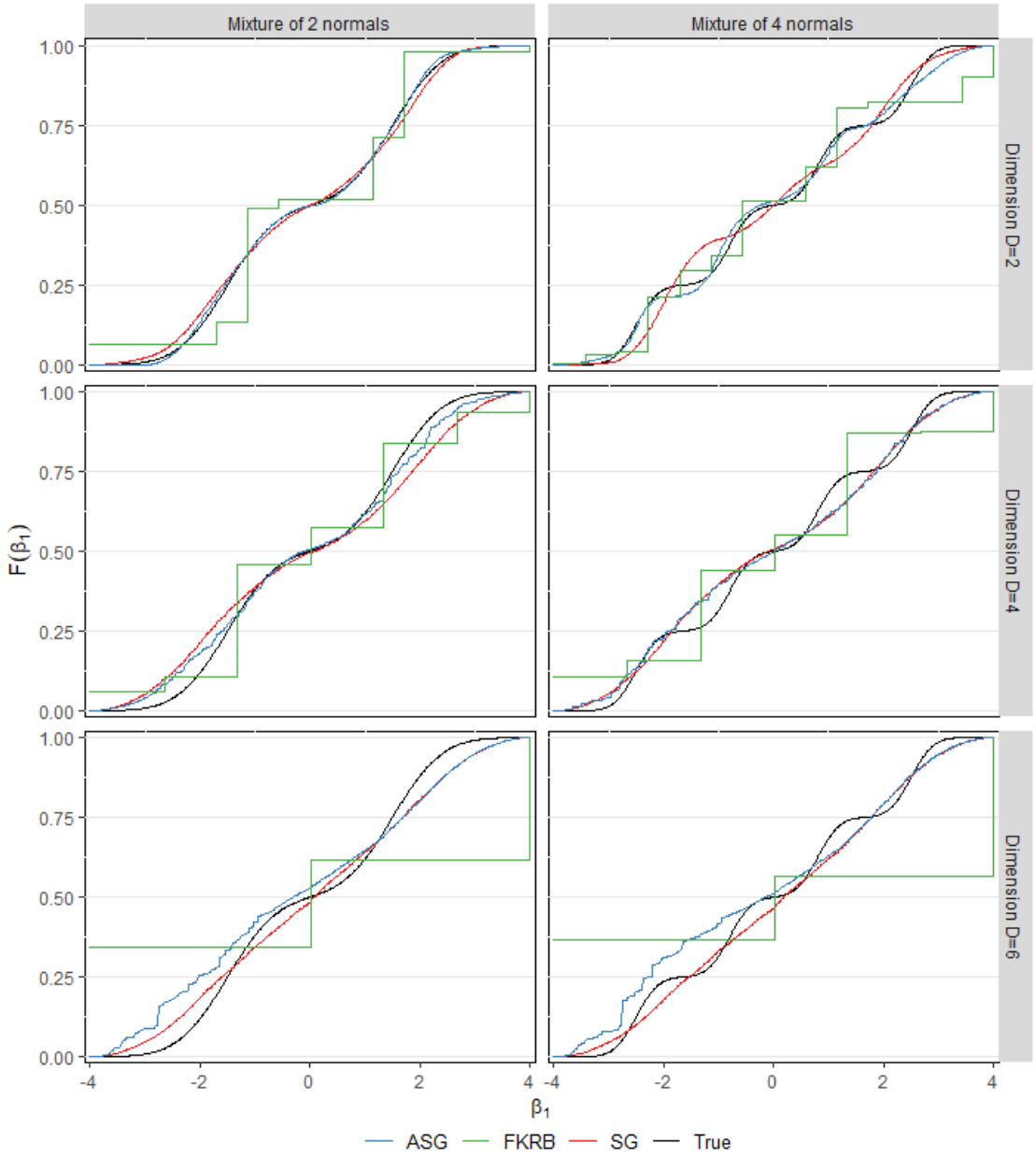
(d) ASG

Note: Estimated bivariate joint distribution function for the mixture of four normals estimated with the FKRB estimator using 225 grid points, with the sparse grid estimator of level $l_S = 4$ (SG), and the spatially adaptive sparse grid estimator (ASG). The number of refinement steps is selected using 5-fold cross-validation and based on the lowest out-of-sample mean squared error.

With respect to the finite sample properties, the sparse grid estimator improves only slightly with increasing sample size for small levels $l_S = 2$ and $l_S = 3$, independent of the true distribution. Due to the larger support of the hierarchical basis functions with smaller levels and the imposed shape restriction following from these basis functions, the sparse hierarchical bases of levels $l_S = 2$ and $l_S = 3$ restrict the flexibility of the estimator to rather smooth approximations, despite an increasing sample size. For sparse hierarchical bases of level $l_S = 4$, in contrast, the approximation accuracy of the sparse grid estimator improves stronger when increasing the sample size from $N = 1000$ to $N = 10,000$ due to the larger number of basis functions with small support. However, this effect declines as the number of random coefficients included in the model increases. In fact, the improvement in the RMISE is negligible for $D = 6$, independent of the level and true distribution.

The results indicate that the number of basis functions in the sparse hierarchical basis could potentially increase faster than in the classical sparse grid to obtain more accurate approximations, which would be possible with respect to the total number of parameters. This impression is confirmed by the estimated marginal CDFs of β_1 for the mixture of two normals and the mixture of four normals in Figure 2.8.

Figure 2.8: True and Estimated Marginal CDFs of β_1 for Mixture of 2 and Mixture of 4 Normals and $N = 10,000$



Note: The figure shows the true and estimated marginal CDFs of β_1 for the mixture of two and the mixture of four normals across models with different number of random coefficients for $N = 10,000$. The sparse grid estimator has level $l_S = 4$, and the spatially adaptive sparse grid estimator refines the sparse grid conducting 15 refinement steps whereby the final estimator is selected based on the lowest out-of-sample MSE. The FKRB estimator estimates the two-dimensional distributions with 15 grid points, the four-dimensional distribution with 7 grid points, and the six-dimensional distribution with 3 grid points in every dimension.

Considering the mixture of two normals, the sparse grid estimator succeeds to accurately approximate the true marginal CDF for $D = 2$. However, the approximation becomes less accurate as the number of random coefficients included in the model increases, indicating that there are too few hierarchical basis functions with sufficiently small support to recover the curvature of the true distribution. This effect is even stronger for the mixture of four normals, where the sparse grid estimator cannot recover the steep and wiggly curvature of the true distribution. While for $D = 2$, the sparse grid estimator can at least recover the curvature at the boundary of the domain,

for $D = 4$ and $D = 6$ it approximates the true marginal CDFs through a line – though correctly located.

The results for the spatially adaptive sparse grid estimator reported in Table 2.2 show that the performance of the refinement depends on the level of the sparse grid and the true distribution. First, the improvement is strongest for sparse hierarchical bases of level $l_S = 2$ and declines as the level increases, independent of the shape of the true distribution. In fact, for $D = 2$ and $N = 1000$, the refinement leads to an improvement only for $l_S = 2$, indicating that the refinement can rapidly lead to over-fitting if the dimension and sample size is small. This is also indicated by the out-of-sample MSE plotted in Figure 2.14, which remains more or less constant with increasing refinement steps. Second, the refinement is more effective for the mixture of four normals than for the mixture of two normals, as the approximation of the steep and wiggly curvature of the former requires more basis functions of smaller levels, i.e., with smaller support. For the mixture of two normals, the spatially adaptive refinement of the sparse grid of level $l_S = 4$ on average leads to less precise estimates than the sparse grid itself. A potential explanation is that this is the consequence of an over-fitting problem as indicated by the in-sample and out-of-sample MSE plotted in Figure 2.14 and Figure 2.15.

Figure 2.8 illustrates the improvement of the sparse grid estimator through the spatially adaptive refinement. Considering the mixture of four normals, the estimated marginal CDF almost perfectly approximates the shape of the true distribution for $D = 2$. However, the approximation accuracy declines with increasing dimensionality of the true distribution. Thus, increasing the number of refinement steps, which is close to the maximum number of 10 for $D = 6$ (see Table 2.6 in the appendix), might lead to more accurate approximations as more basis functions of smaller levels allow to recover the curvature of the mixture of four normals more precisely.

2.6 Empirical Application

In order to study the performance of the sparse hierarchical basis estimator on real data, we apply it to the setting of air pollution regulation from Blundell et al. (2020), hereafter referred to as BGL.[14] They study the gains from dynamic enforcement of air pollution regulations using a discrete-time dynamic model of regulator and plant interactions. In this model, the regulator makes decisions regarding inspections and fines, and plants decide whether and when to invest in pollution abatement technologies. The quantification of the gains from dynamic enforcement of the regulation crucially depends on the estimation of plants' costs arising from compliance with the regulation. BGL estimate a random coefficients model to accommodate the unobserved heterogeneity of costs across plants. They estimate the five-dimensional joint distribution using the nonparametric fixed grid estimator of Fox et al. (2011). We apply the sparse grid estimator and the spatially adaptive sparse grid estimator to this setting and compare the estimated distribution and the results of counterfactual experiments calculated with the estimated distributions to the results of BGL.

The Clean Air Act and its amendments (CAAA) restrict the pollution of criteria and hazardous

[14] We gratefully thank Blundell et al. (2020) for the provided data and code, and Stephan Hetzenecker, who provided a parallelized version of the code that substantially speeds up the calculations of the optimal weighting matrix and the counterfactuals.

air pollutants through plants' in the United States to be at or below thresholds that could be achieved with the best technologies and practices. The US Environmental Protection Agency (EPA) used a dynamic enforcement regime to ensure plants' compliance with the CAAA. The EPA's inspections aim to uncover possible violations. Plants detected as violators will be subject to further inspections. These inspections, in turn, might uncover additional violations, leading to potential fines. Among other factors, the magnitude of fines depends on the economic benefit of the violating plant, and on the gravity of the violation. The latter is calculated from the actual or potential harm and plants' history of noncompliance. Plants can only exit violator status if they resolve all outstanding violations. The total cost of noncompliance to a violator arise from the investment cost required to resolve outstanding violations, from an increased level of oversight through the EPA, and from fines.[15]

While plants with fewer and less severe violations are designated as "regular violators", those with particularly severe or repeated violations can be designated as "high priority violators" (HPVs). The idea of this regulatory regime is to make it more costly for plants to be in HPV status: HPV undergo a higher level of oversight – expressing itself through more frequent inspections – are exposed to higher fines, and have to fulfill explicit deadlines to resolve all outstanding violations. The higher cost for HPV in comparison to regular violator are intended to encourage plants that are out of compliance to return to compliance via investments in improved processes and technologies (Blundell et al., 2020).

BGL model the regulatory framework using a discrete-time dynamic model in which each plant plays a dynamic game with the regulator. In this game, the regulator decides whether or not to inspect a plant, and plants decide whether or not to invest in pollution abatement technologies. While the regulator wants plants to comply with the CAAA, which causes costs arising from inspections and issuing fines, plants seek to maximize their surplus. The actions of the regulator and the investment decision of a plant within period t are functions of the regulatory state Ω_t, which is known to the regulator and plant at the beginning of a period. The regulatory state lists (i) a plant's EPA region, (ii) two-digit NAICS industrial sector,[16] (iii) expected gravity of potential violations, as measured by county non-attainment status and potential environmental damages for plants based on the county and industry, (iv) depreciated accumulated violations with a 10 percent quarterly depreciation rate, (v) regular violator or high priority violator status, and (vi) two quarterly lags of investment. While the states addressing the EPA region, the industry and the gravity of fines do not change over time, the depreciated accumulated violations, the violator status, and the lagged investments can change from period t to period $t+1$.

To incorporate plants' history of violations into the regulator's inspection policy, BGL model the probability of a plant being inspected through the regulator as a function of the regulatory state, $\mathcal{I}(\Omega)$. The actual inspection decision Ins arises stochastically. In each period, the regulator first receives an i.i.d. private information shock to the value of an inspection and then decides whether or not to inspect the plant. The regulator and plant then receive a compliance signal $\boldsymbol{e}_t \equiv (e_t^1, \ldots, e_t^5)$

[15]Costs from an increased level of oversight, i.e., from more frequent inspections, are caused by the potential shut down of production lines.

[16]The data covers the seven most polluting industrial sectors in North America defined by the North American Industry Classification System (NAICS).

which provides information on the presence and severity of a violation. BGL assume that $\boldsymbol{e}_t$ is a function only of the regulatory state Ω_t, and the regulator's inspection policy and decisions (i.e., of the inspection probabilities $\mathcal{I}$, and the inspection decision Ins_t). Therefore, $\boldsymbol{e}_t$ is the predictor of compliance issues beyond the state, $Vio(\Omega, e^1)$. In addition, e^2 affects the fine chosen by the regulator through $Fine(\Omega, e^2)$, and e^3, e^4, and e^5 determines plants' transition to compliance, regular violator, and HPV status through $\tilde{\Omega} \equiv T(\Omega, e^3, e^4, e^5)$. Following the regulator's actions, Ins, Vio, $Fine$ and T, plants that are not in compliance under $\tilde{\Omega}$ make a binary decision $X \in \{0, 1\}$ of whether or not to invest in pollution abatement technologies.

In order to avoid assumptions on the regulator's objective function, BGL do not estimate the regulator's utility function. Instead, they estimate plants' expectations of regulator actions using conditional choice probabilities (CCPs), and then use these probabilities to estimate plants' utility functions. To condition on the state, they estimate the CCPs separately for plants in compliance, regular violators, and HPVs, and include indicators for two lags of investments; region; industry and gravity state dummies; and depreciated accumulated violations (for plants not in compliance).

The utility of a plant depends on the regulatory actions, the HPV status designation $HPV(\cdot)$, and the investment cost $\theta^X + \epsilon_{Xt}$,

$$\begin{aligned} U(\Omega, \boldsymbol{e}) = \theta^I Ins(\Omega) + \theta^V Vio\left(\Omega, e^1\right) + \theta^F Fine\left(\Omega, e^2\right) + \\ \theta^H HPV\left(T\left(\Omega, e^3, e^4, e^5\right)\right) + \theta^X + \epsilon_{Xt}, \end{aligned} \tag{2.26}$$

where ϵ_{Xt} is an idiosyncratic cost shock assumed to be known to the plant prior to its investment decision, and which is assumed to be i.i.d. type I extreme value. The plant chooses its investment decision in order to minimize its expected discounted sum of costs from inspections (I), violations (V), fines (F), designation as HPV (H), and investment (X).

To account for unobserved heterogeneity across plants, BGL specify a random coefficients model which allows the structural parameters of the model, $\boldsymbol{\theta} \equiv (\theta^I, \theta^V, \theta^F, \theta^H, \theta^X)$, to vary across plants (but not over time).[17] They estimate the distribution of the random coefficients using the non-parametric fixed grid estimator of Fox et al. (2011) with $10,0001$ five-dimensional grid points $\theta \equiv (\boldsymbol{\theta}_1, \ldots, \boldsymbol{\theta}_R)$ – $10,000$ points from a Halton sequence and one point which corresponds to the parameter estimates from a model with homogeneous utility parameters. They estimate the probability weights η_r, $r = 1, \ldots, R$ at the grid points from the data using a GMM estimator similar to the approach of Nevo et al. (2016). The estimator minimizes the squared distance between the value of some statistic in the data, m_k^{d}, and the weighted sum of the statistic estimated at every grid point, $m_k(\boldsymbol{\theta}_r)$, $k = 1, \ldots, K$, subject to the constraints that the weights are nonnegative and sum up to one. Let $G(\boldsymbol{\eta})$ denote the $K \times 1$ vector of moments with the kth entry being $G_k(\boldsymbol{\eta}) = m_k^{\mathrm{d}} - \sum_{r=}^{R} \eta_r m_k(\boldsymbol{\theta}_r)$. The GMM estimator solves the constrained optimization problem

$$\begin{aligned} \boldsymbol{\eta} &= \arg\min_{\boldsymbol{\eta}} G'(\boldsymbol{\eta})\ W\ G(\boldsymbol{\eta}) \\ \text{s.t.} \quad \eta_r \geq 0, \quad r = 1, \ldots, R, \quad &\text{and} \quad \sum_{r=1}^{R} \eta_r = 1 \end{aligned} \tag{2.27}$$

[17] With heterogeneous investment costs, the escalation mechanism may incentivize low-cost plants to invest in pollution abatement when they are regular violators and fines are low, while high-cost plants will wait until they become HPVs and fines are high.

where W is a $K \times K$ weighting matrix and G' the transpose of G. For the estimation of the probability weights, BGL calculate three sets of moments. The first set ($5,000$ moments) represents the equilibrium share of plants being in a particular time-varying state, conditional on fixed states of region, industry, and gravity states. The second set ($4,687$ moments) multiplies the first set by the share of plants investing in this state. And the third set ($4,687$ moments) multiplies the second set by the sum of investments in the following six periods.[18] The second and third set of moments are intended to capture the effect of plants' investments on compliance.

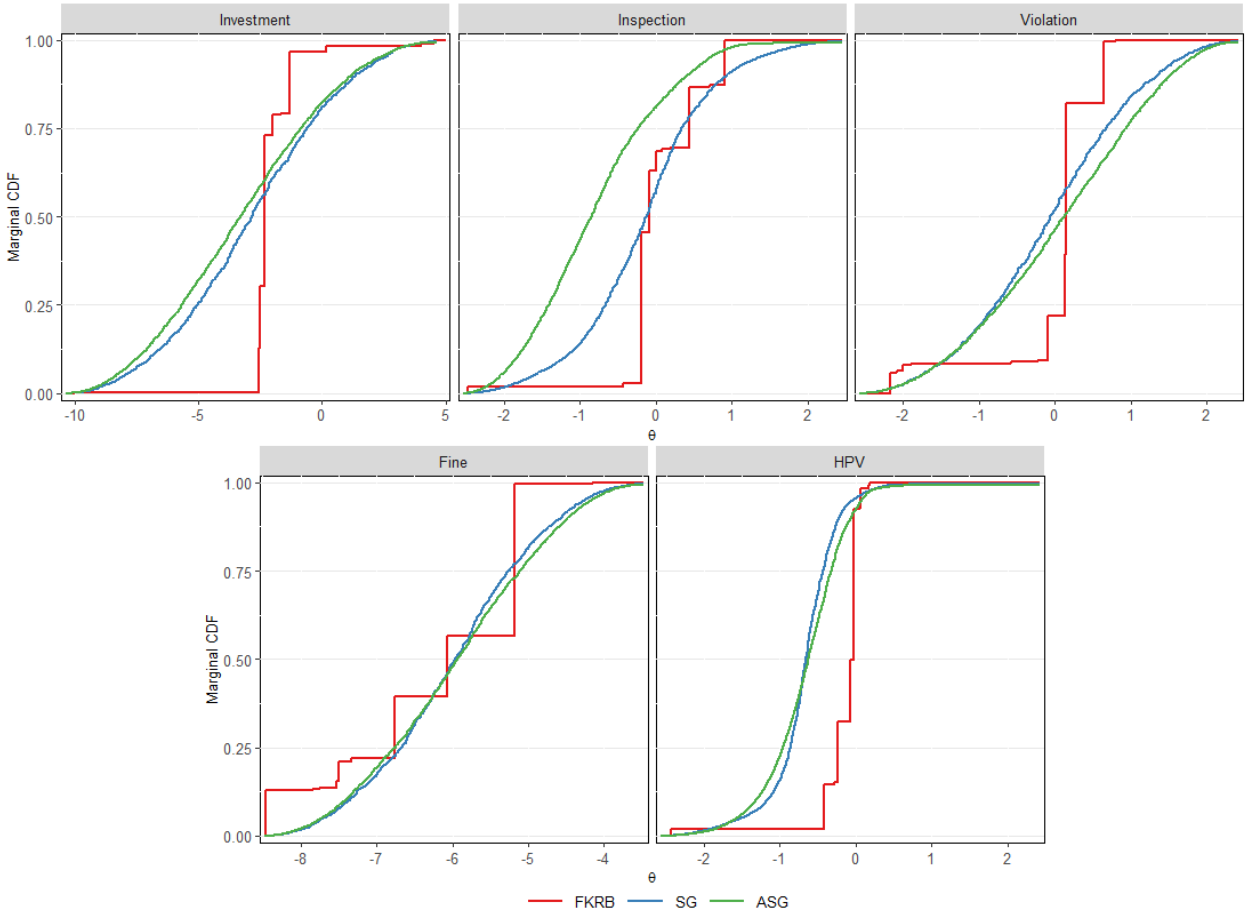
Because the fixed grid approach of Fox et al. (2011) treats the probability weight at every grid point as a parameter, the estimation of the random coefficients distribution involves the estimation of $10,001$ parameters. To reduce the computational burden, and in line with the results from the Monte Carlo experiments presented in the previous section, we estimate the distribution with the sparse grid estimator of level $l_S = 4$, and the spatially adaptive sparse grid estimator using the same GMM approach. The corresponding moments are of the form $G_k(\boldsymbol{\alpha}) = m_k^{\mathrm{d}} - \sum_{b=1}^{B} \alpha_b \sum_{r=1}^{R} \phi_b(\boldsymbol{\theta}_r) m_k(\boldsymbol{\theta}_r)$ for $k = 1, \ldots, K$. We minimize the weighted squared sum of moments subject to the constraints $\sum_{b=1}^{B} \alpha_b \phi_b(\boldsymbol{\theta}_r) \geq 0$, $r = 1, \ldots, R$, and $\sum_{b=1}^{B} \alpha_b \sum_{r=1}^{R} \phi_b(\boldsymbol{\theta}_r) = 1$. For the spatial refinement of the sparse grid, we make ten refinement steps. In every step, we select the grid point with the largest contribution to the squared estimated local error. The final number of refinement steps is determined using five-fold cross-validation.[19] The final adaptive estimator uses the grid for which the out-of-sample mean-squared-error is lowest, which is the case after eight refinements. Figure 2.17 in the appendix shows the change in the out-of-sample MSE with an increasing refinement of the sparse grid. The lowest out-of-sample MSE is achieved after 14 refinements. To increase the efficiency of the estimator, we adopt the two-step approach of BGL. In the first step, we calculate the weighting matrix using the homogeneous parameter estimates of $\boldsymbol{\theta}$ provided by BGL. In the second step, we update W using the estimated random coefficients distribution from the first step.

Figure 2.9 shows the estimated marginal CDFs for the FKRB estimator, the sparse grid estimator, and the spatially adaptive sparse grid estimator. The figure illustrates that the fixed grid estimator of Fox et al. (2011) approximates the random coefficients' distribution through a step function with only a few steps – which is a result of its sparse nature (it estimates only 14 positive weights). The estimated marginal CDFs of the fine, inspection and HPV utility parameters look relatively similar for all the three estimators, except that the distributions estimated with the sparse grid and spatially adaptive sparse grid estimator are much smoother. The estimated marginal CDFs of the HPV utility parameter illustrate that the sparse grid estimator can approximate a steep curvature – though not as steep as the curvature estimated by the FKRB estimator. For the remaining utility parameters, the marginal CDFs estimated with the FKRB estimator and with the sparse grid and spatially adaptive sparse grid estimator deviate to a larger extent. Except for the inspection utility parameter, where the estimated marginal CDFs seem to deviate to a larger extent, the sparse grid and spatially adaptive sparse grid estimator provide similar marginal distributions. The estimated histograms plotted in Figure 2.18 in the appendix

[18] For the calculation of the moments, BGL solve the relevant Bellman equation and calculate $m_k(\eta_r)$ for each of the R grid points. They provide a detailed description on the set of moments in their paper and more information on the calculation of the moments in the online appendix of the paper.

[19] To preserve all information contained in the data, we sample all moments together that use the same first and second moments, respectively, for the calculation of further moments.

Figure 2.9: Estimated Marginal CDFs of Five Utility Parameters



Note: The figure shows the marginal CDFs for the five utility parameters estimated with the FKRB estimator using 10,001 grid points, with the sparse grid estimator of level $l_S = 4$ (SG), and the spatially adaptive sparse grid estimator (ASG). For the spatially adaptive sparse grid estimator, the number of refinement steps is selected using 5-fold cross-validation and based on the lowest out-of-sample mean squared error.

confirm the impression.

The weighted means of the estimated random coefficients' distribution reported in Table 2.3 confirm the impression from the visual inspection of the estimated marginal CDFs. Maybe most noticeable, plants find investments, inspections, fines, being in HPV status, and violations costly on average when the random coefficients distribution is estimated with the sparse grid estimator. This result is in line with the quasi-likelihood estimates, i.e., if plants have homogeneous utility parameters. For the FKRB estimator, in contrast, violations and inspections increase plants' utility on average slightly. When estimated with the adaptive sparse grid estimator, plants on average receive a positive utility from violations. Given that the estimated mean parameters are utility parameters, we can express them as fine-equivalents to compare the magnitude of the estimated means in a meaningful way. When the model parameters are estimated with the FKRB estimator, plants' costs for being in HPV status are equivalent to an average fine of about $\$25,516$ (θ^H/θ^F multiplied by \$1 million), whereas the average costs for being designated as HPV are about $\$114,380$ in fine equivalents for the sparse grid estimator and about $\$112,839$ for the spatially adaptive sparse grid estimator. In line with that, the average costs from investments are equivalent to about $\$330,196$ in fines for the FKRB estimator, about $\$475,274$ in fines for the sparse grid estimator, and $\$545,950$

Table 2.3: Plants' Estimated Structural Mean Parameters

	Quasi-likelihood estimates	FKRB	SG	ASG
Negative of investment cost ($-\theta^X$)	-2.872	-2.051	-2.831	-3.232
Inspection utility (θ^I)	-0.049	0.047	-0.152	-0.797
Violation utility (θ^V)	-0.077	0.012	-0.062	0.080
Fine utility (mil. dollars, θ^F)	-5.980	-6.211	-5.956	-5.920
HPV status utility (θ^H)	-0.065	-0.158	-0.681	-0.668
Parameters	5	10001	351	462

Note: The table reports the homogeneous parameter estimates (QML), and the estimated weighted mean of each random coefficient together with the total number of parameter required for the estimation of the random coefficients distribution for the FKRB estimator, the sparse gird estimator (SG), and the spatially adaptive sparse grid estimator (ASG).

for the adaptive sparse grid estimator. Finally, plants' average costs from inspections and violations estimated with the FKRB estimator are equivalent to fines of \$-7,572, and \$-2,002, respectively, implying that inspections and violations do not decrease utility for some plants. In contrast, inspections are equivalent to about \$25,471 in fines, and violations are about \$10,405 on average for the sparse grid estimator, and \$134,628 and \$-13,514 for the adaptive sparse grid estimator.

To study how the difference in the estimated distributions translates to counterfactual statistics calculated with the respective estimated random coefficients' distributions, we replicate three counterfactual experiments conducted by BGL. The first experiment studies how regulatory states, pollution, and investments change when the regulator fines plants in regular violator and HPV status identically for a given region, industry, and gravity state, keeping the total assessed fines the same as the baseline model for each such group. Thus, the costs of HPV status are set to zero in this experiment to fully remove dynamic enforcement. The second experiment considers the same fine structure as the first experiment but keeps the total pollution damages the same as the baseline model within each region, industry, and gravity state group. The third experiment doubles the fines for firms in HPV status compared to the baseline model, which allows to study the effect of higher escalation rates of fines.[20]

Table 2.4 presents the results of the experiments in terms of the long-run mean value of regulatory states, regulatory actions, investment rates, plant utility, and pollution damages for the FKRB estimator, the sparse grid estimator, and the spatially adaptive sparse grid estimator. The first column reports the long-run mean values observed in the data. The baseline columns show the outcomes calculated at the structural parameters estimated with each estimator. The estimated mean values are similar for all three estimators, and replicate the data quite well. Overall, the predicted results of the counterfactual experiments are relatively similar – especially for the FKRB and sparse grid estimator. In the first counterfactual experiment, the share of plants in compliance predicted by the FKRB and sparse grid estimator decreases substantially from about 95% to 65% if

[20]The inspection policies in the experiments are the same as in the baseline model to assure the same state-contingent distribution of the compliance signal $\boldsymbol{e}$. Furthermore, the counterfactuals are based on surplus-optimizing plants given alternative regulatory policies and do not necessarily stem from the equilibrium of a dynamic game.

Table 2.4: Counterfactual Results for Different Fine Structures

		Baseline			Same fines for all violators; fines const.		
	Data	FKRB	SG	ASG	FKRB	SG	ASG
Compliance (percent)	95.62	95.11	95.39	95.35	66.76	65.47	80.60
Regular violator (percent)	2.88	3.47	3.55	3.61	2.52	2.34	3.00
HPV (percent)	1.50	1.42	1.06	1.04	30.72	32.19	16.40
Investment rate (percent)	0.40	0.54	0.53	0.52	0.47	0.47	0.48
Inspection rate (percent)	9.65	9.41	9.33	9.32	20.52	20.87	15.00
Fines (thousand dollars)	0.18	0.32	0.29	0.28	0.32	0.29	0.28
Violations (percent)	0.55	0.54	0.52	0.51	4.97	5.00	2.82
Pollution damages (mil. dollar)	1.65	1.53	1.48	1.47	4.03	4.12	2.77
		Same fines for all violators; pollution damages const.			Fines for HPVs doubled relative to baseline		
	Data	FKRB	SG	ASG	FKRB	SG	ASG
Compliance (percent)	95.62	94.49	95.09	95.39	95.52	95.73	95.65
Regular violator (percent)	2.88	2.72	2.27	2.70	3.47	3.56	3.62
HPV (percent)	1.50	2.78	2.64	1.91	1.01	0.72	0.73
Investment rate (percent)	0.40	0.65	0.70	0.64	0.55	0.53	0.52
Inspection rate (percent)	9.65	9.88	9.80	9.56	9.28	9.21	9.21
Fines (thousand dollars)	0.18	1.98	4.52	3.67	0.36	0.29	0.29
Violations (percent)	0.55	0.74	0.71	0.63	0.49	0.46	0.46
Pollution damages (mil. dollar)	1.65	1.53	1.48	1.47	1.48	1.44	1.44

Note: Each statistic is the long-run equilibrium mean per plant/quarter, weighted by the number of plants by region, industry, and gravity state. Baseline refers to the model predictions in the existing regulatory actions and outcomes. The other experiments vary the escalation of fines.

fines are identical for regular violators and HPVs and the total fines are kept constant compared to the baseline model. In contrast to the FKRB estimator (66.76% in compliance and 30.72% in HPV status) and the sparse grid estimator (65.47% in compliance and 32.19% in HPV status), the drop in the share of plants in compliance predicted with the spatially adaptive sparse grid estimator is less strong (80.60% plants are in compliance and 16.40% are designated as HPVs). In line with the higher share of plants in non-compliance, the predicted total pollution damages increase from \$1.53 mil. per plant/quarter to \$4.03 mil per plant quarter for the FKRB estimator and from \$1.48 mil. per plant/quarter to \$4.12 mil per plant/quarter for the sparse grid estimator. For the spatially adaptive sparse grid estimator, the total pollution damages are predicted to increase less strongly from \$1.47 mil. per plant/quarter to \$2.77 mil. per plant/quarter.

The results of the second counterfactual experiment deviate only slightly from each other when estimated with the three estimators, except for the total fines. If the fines are the same for regular violators and high priority violators and total pollution damages are kept constant compared to the baseline model, the total amount of fines increases from \$320 per plant/quarter to \$1,980 per plant/quarter for the FKRB estimator. For the sparse grid and spatially adaptive sparse grid, the predicted total fines increase stronger from \$290 (SG) and \$280 (ASG) per plant/quarter to \$4,520 and \$3,670 per plant/quarter in comparison to the baseline model, respectively.

For the third counterfactual experiment, the difference between mean values predicted by the

FKRB, the sparse grid, and the spatially adaptive sparse grid estimator are even smaller than in the previous experiments. When the fines are doubled for HPV in comparison to the baseline model, the FKRB estimator predicts a decrease in the share of plants in compliance from 95.11% to 95.52$, the sparse grid estimator a decrease from 95.39% to 95.73$, and the spatially adaptive sparse grid estimator from 95.35% to 95.65$. Thus, the substantial increase in the fines for HPVs only leads to a slight increase in the predicted share of plants in compliance. The strongest effect of the counterfactual regulatory policy is the change in the share of plants in HPV status. All three estimators predict a decrease to a similar extent (from 1.42$ to 1.01% for the FKRB estimator, from 1.42% to 0.53% for the sparse grid, and from 1.04% to 0.52% for the spatially adaptive sparse grid estimator). Most importantly, the increased escalation of fines does only lead to a slight decrease in the predicted total pollution damages. When predicted with the FKRB estimator, the total pollution damages decrease from $1.53 million to only $1.48 in response to the change in the fine scheme. The predicted change is similar for the sparse grid (from $1.48 million to $1.44 million) and the spatially adaptive sparse grid estimator (from $1.47 million to $1.44 million).

2.7 Conclusion

A common approach in the nonparametric literature is to approximate functions of unknown shapes using linear combinations of basis functions. For the approximation of multi-dimensional functions, the bases are typically constructed using regular tensor product constructions of one-dimensional basis functions. Such constructions lead to an exponential increase of the number of parameters in the number of dimensions, which restricts the approach to random coefficient models with only moderately few random coefficients. In order to circumvent this limitation, we propose to use sparse hierarchical bases for the nonparametric estimation of high-dimensional random coefficient models.

The proposed estimator approximates the true distribution using a linear combination of hierarchical basis functions, whereby the multi-dimensional basis functions are constructed from the one-dimensional functions using a truncated tensor product. The underlying idea goes back to Smolyak (1963) and has been frequently applied in mathematics and physics for the approximation of high-dimensional functions. The truncated tensor product reduces the number of basis functions substantially in comparison to a regular tensor product – thereby rendering the estimation of high-dimensional distributions feasible. The sparse hierarchical basis deteriorates the approximation accuracy only slightly if the underlying distribution is sufficiently smooth. For non-smooth distributions, we additionally propose a spatially adaptive refinement procedure, which incrementally adds basis functions in those areas of the true distributions' domain where it has a steeper and wigglier curvature.

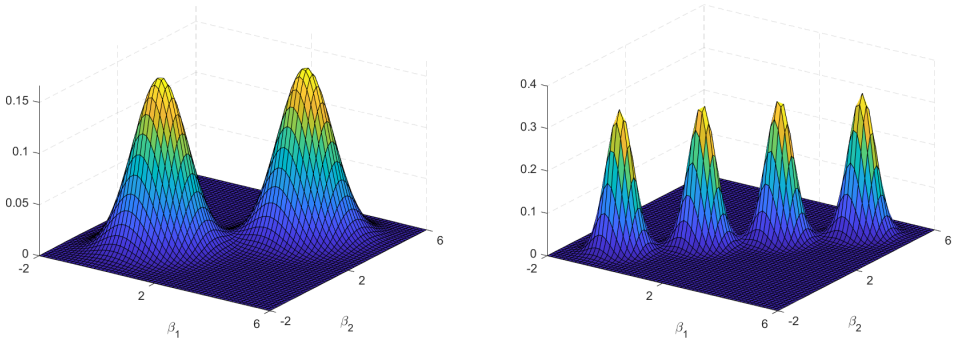
We study the properties of the sparse hierarchical basis estimator in various Monte Carlo experiments. Using the nonparametric fixed grid estimator of Fox et al. (2011) as a benchmark, the results show that our estimator provides more accurate approximations of the true distribution, even for models with only a few random coefficients, and especially for models with moderately many random coefficients. Moreover, the results confirm the theoretical properties of sparse hierarchical bases presented by Bungartz and Griebel (2004). The sparse grid estimator becomes less accurate if the true distribution has a steeper and wigglier curvature and if the number of

random coefficients included into the model increases. The spatially adaptive refinement of the sparse grid works particularly well for those distributions. Applying the estimator to a data set of plants' investments in pollution abatement technologies illustrates the advantage of the sparse hierarchical basis estimator. Even though the approach requires a substantially smaller number of parameters for the estimation of the five-dimensional random coefficients distribution, the counterfactuals predicted based on the estimated distribution deviate only slightly from those predicted by the estimator of Fox et al. (2011), which involves the estimation of $10,001$ parameters.

A practically relevant topic with respect to the application of the estimator in applied research is a valid inference procedure. Such a procedure has to take into account that the coefficients are estimated with constrained least squares, i.e., the coefficients on the boundary of the parameter space cannot have an asymptotic normal distribution. In addition, a promising avenue for future research is to consider different kinds of sparse grids. The Monte Carlo results show that for random coefficient models with four or six random coefficients, the number of basis functions could increase faster than the rate of the classical sparse grid. Studying different kinds pf sparse grid constructions and their theoretical properties when applied to the estimation of random coefficients' distributions would provide valuable insights.

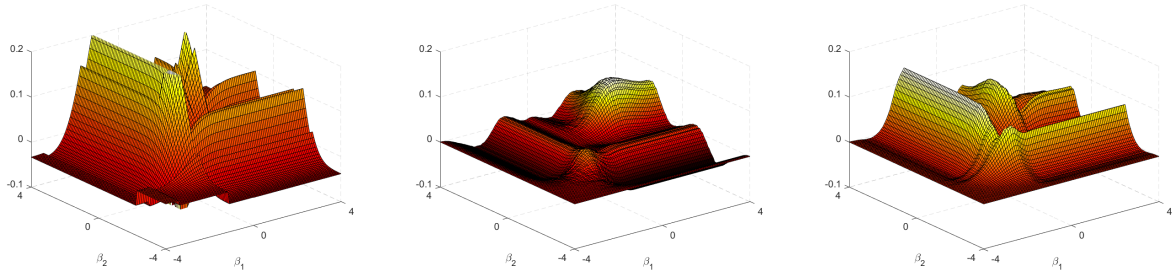
Appendix: Additional Tables and Figures

Figure 2.10: True Joint PDFs of Mixture of 2 and Mixture of 4 Normals



(a) Mixture of 2 (b) Mixture of 4

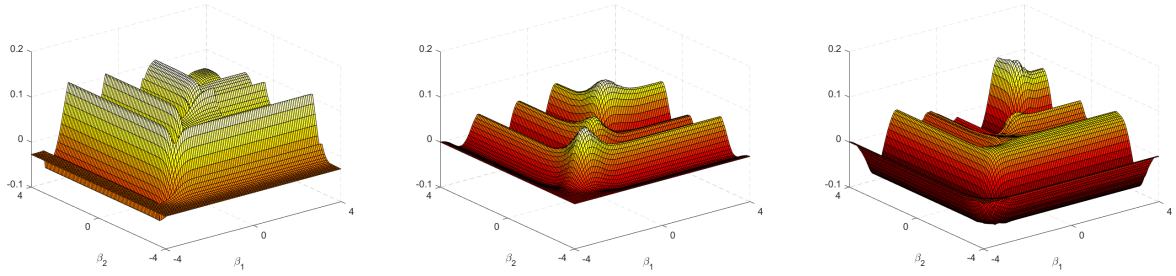
Figure 2.11: Approximation Error of Estimated Bivariate Joint CDF for Mixture of 2 Normals and $N = 10,000$



(a) FKRB (b) SG (c) ASG

Note: Approximation error of estimated bivariate distribution functions of mixture of two normals for $N = 10,000$, and estimated with the FKRB estimator with 225 grid points, with the sparse grid estimator with level $l_S = 4$, and the spatially adaptive sparse grid estimator. The number of refinement steps is selected using 5-fold cross-validation and based on the lowest out-of-sample mean squared error.

Figure 2.12: Approximation Error of Estimated Bivariate Joint CDF of Mixture of 4 Normals and $N = 10,000$



(a) FKRB (b) SG (c) ASG

Note: Approximation error of estimated bivariate distribution functions of mixture of four normals for $N = 10,000$, and estimated with the FKRB estimator with 225 grid points, with the sparse grid estimator with level $l_S = 4$, and the spatially adaptive sparse grid estimator. The number of refinement steps is selected using 5-fold cross-validation and based on the lowest out-of-sample mean squared error.

Table 2.5: Average Number of Parameters, Refinement Steps and RMISE across 200 Monte Carlo Replicates for Different Selection Criteria for Spatially Adaptive Refinement

		Mixture of 2 Normals									Mixture of 4 Normals								
		Refinements			Parameter			RMISE			Refinements			Parameter			RMISE		
N	l_S	MSE	LL	AIC	MSE	LL	AIC	MSE	LL	AIC	MSE	LL	AIC	MSE	LL	AIC	MSE	LL	AIC
Dimension $D = 2$																			
1000	2	6.8	5.9	3.9	35.6	30.3	19.1	0.0514	0.0500	0.0480	7.3	6.2	4.0	39.1	31.9	19.7	0.0577	0.0568	0.0531
1000	3	4.6	1.9	0.1	41.6	26.3	17.3	0.0536	0.0495	0.0484	5.6	2.3	0.1	48.3	28.7	17.4	0.0589	0.0547	0.0479
1000	4	4.7	2.7	0.0	70.4	61.0	49.0	0.0561	0.0527	0.0475	5.2	2.6	0.0	72.9	60.6	49.0	0.0624	0.0580	0.0549
10,000	2	8.0	4.4	4.2	42.5	21.0	20.1	0.0280	0.0342	0.0351	9.3	4.6	4.2	50.1	22.3	20.4	0.0435	0.0492	0.0485
10,000	3	6.8	0.3	0.3	52.9	18.3	18.2	0.0290	0.0399	0.0401	9.0	0.9	0.6	63.6	20.7	19.2	0.0418	0.0468	0.0457
10,000	4	6.3	0.5	0.0	76.3	50.9	49.1	0.0303	0.0294	0.0312	8.3	0.4	0.0	84.5	50.6	49.1	0.0394	0.0510	0.0518
Dimension $D = 4$																			
1000	2	9.1	9.1	6.9	149.7	149.4	99.8	0.0583	0.0586	0.0537	9.3	9.2	7.3	153.7	152.5	105.3	0.0634	0.0635	0.0624
1000	3	8.4	8.5	3.7	213.8	216.1	104.1	0.0598	0.0607	0.0530	8.5	8.7	3.8	226.9	231.3	105.1	0.0632	0.0635	0.0636
1000	4	6.5	6.5	1.3	341.7	339.3	222.3	0.0569	0.0582	0.0482	6.5	6.8	1.8	341.2	350.0	229.9	0.0577	0.0583	0.0545
10,000	2	9.3	8.9	8.6	144.1	133.6	126.2	0.0409	0.0423	0.0429	9.5	9.5	8.7	151.6	151.2	129.4	0.0536	0.0536	0.0529
10,000	3	8.3	7.3	6.6	193.8	167.5	151.3	0.0386	0.0409	0.0420	9.1	8.9	6.9	215.2	210.2	159.8	0.0511	0.0511	0.0501
10,000	4	6.8	5.2	3.6	341.5	303.3	265.0	0.0361	0.0381	0.0405	8.4	7.6	3.6	379.6	360.8	265.8	0.0480	0.0480	0.0472
Dimension $D = 6$																			
1000	2	9.7	9.7	7.0	276.2	280.0	154.9	0.0569	0.0573	0.0564	9.8	9.8	7.5	266.4	267.6	170.8	0.0733	0.0734	0.0795
1000	3	9.2	9.5	0.1	386.9	397.9	99.2	0.0547	0.0552	0.0642	9.4	9.6	1.0	416.8	421.6	127.1	0.0626	0.0627	0.0869
1000	4	7.9	8.2	2.2	954.0	974.6	613.3	0.0622	0.0642	0.0451	8.7	8.8	2.8	1024.9	1028.7	638.4	0.0643	0.0644	0.0609
10,000	2	10.0	10.0	9.5	246.9	247.6	225.2	0.0577	0.0577	0.0562	10.0	10.0	9.7	233.7	233.8	222.1	0.0774	0.0774	0.0774
10,000	3	9.7	9.8	0.8	207.1	208.3	108.7	0.0625	0.0625	0.0636	9.8	9.9	1.2	208.1	209.1	113.4	0.0888	0.0888	0.0899
10,000	4	9.5	9.4	3.1	1055.3	1054.3	657.6	0.0687	0.0688	0.0339	9.8	9.8	3.9	1099.1	1102.5	708.9	0.0568	0.0569	0.0429

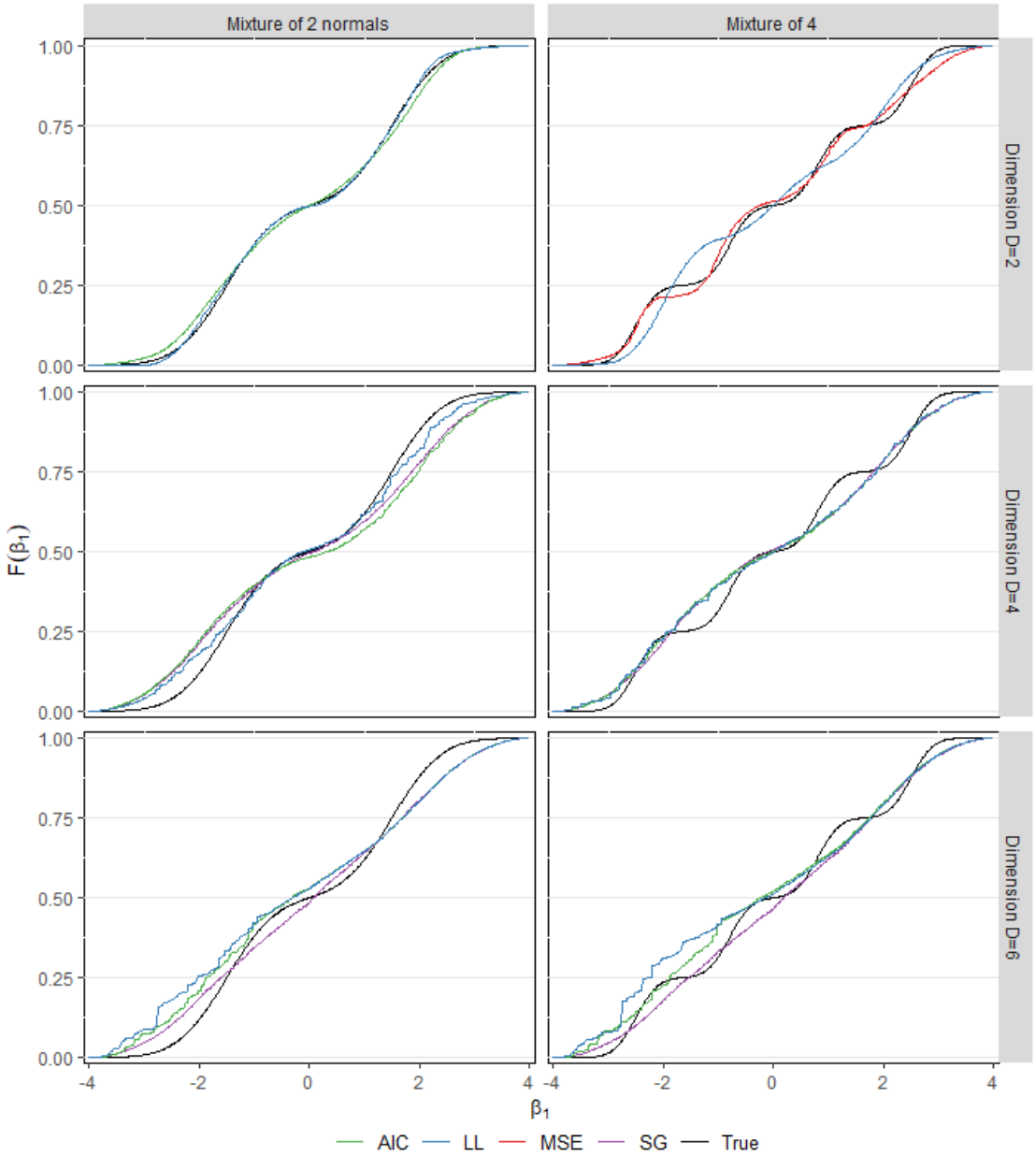
Note: The table reports the total number of parameter and the RMISE for the spatially adaptive sparse grid estimator using the out-of-sample mean squared error (MSE), the out-of-sample log-likelihood (LL), and the Akaike information criterion (AIC) for the selection of the final number of refinement steps. For every estimator five refinement steps are performed. The out-of-sample mean squared error and the out-of-sample log-likelihood are calculated with five-fold cross-validation. The grid point to be refined in every refinement step is selected according to its contribution to the local squared error.

Table 2.6: Average Out-of-sample MSE and Out-of-sample Log-likelihood across 200 Monte Carlo Replicates for Different Selection Criteria for Spatially Adaptive Refinement

		Mixture of 2 Normals								Mixture of 4 Normals							
		MSE				LL				MSE				LL			
N	l_S	SG	MSE	LL	AIC	SG	MSE	LL	AIC	SG	MSE	LL	AIC	SG	MSE	LL	AIC
Dimension $D = 2$																	
1000	2	1422.7	1367.0	1367.4	1369.1	-290.7	-276.2	-276.1	-276.4	1425.3	1358.9	1359.5	1361.5	-291.8	-274.1	-274.0	-274.4
1000	3	1368.7	1366.6	1367.3	1368.7	-276.2	-276.2	-275.9	-276.2	1361.0	1358.2	1359.2	1361.0	-274.2	-274.1	-273.9	-274.1
1000	4	1369.0	1366.8	1367.3	1369.0	-276.6	-276.4	-276.3	-276.6	1360.5	1358.1	1358.8	1360.5	-274.4	-274.3	-274.2	-274.4
10,000	2	1420.9	1366.5	1366.9	1367.0	-2902.1	-2766.4	-2763.2	-2763.3	1423.0	1357.3	1358.2	1358.4	-2911.0	-2741.5	-2738.2	-2738.3
10,000	3	1367.1	1366.4	1367.0	1367.1	-2762.6	-2767.0	-2762.5	-2762.5	1359.0	1357.0	1358.5	1358.7	-2738.2	-2743.4	-2737.9	-2738.0
10,000	4	1366.8	1366.4	1366.7	1366.8	-2765.9	-2767.8	-2765.8	-2765.9	1357.7	1356.8	1357.6	1357.7	-2740.9	-2744.9	-2740.8	-2740.9
Dimension $D = 4$																	
1000	2	1471.3	1351.1	1351.2	1356.7	-299.6	-266.2	-266.2	-267.9	1469.6	1333.4	1333.5	1338.4	-299.7	-262.2	-262.1	-263.6
1000	3	1390.3	1348.8	1348.9	1358.8	-278.0	-265.4	-265.4	-268.7	1384.4	1330.8	1331.0	1343.2	-276.7	-261.4	-261.3	-265.1
1000	4	1367.4	1345.6	1345.9	1350.9	-270.9	-264.6	-264.5	-266.1	1358.3	1326.2	1326.6	1332.9	-269.0	-260.0	-259.9	-261.9
10,000	2	1468.9	1344.7	1344.7	1344.8	-2986.8	-2645.8	-2645.6	-2645.7	1466.8	1326.6	1326.6	1326.9	-2991.1	-2601.8	-2601.7	-2602.4
10,000	3	1388.9	1344.2	1344.3	1344.5	-2776.7	-2645.7	-2645.3	-2645.7	1382.6	1324.9	1324.9	1325.4	-2762.6	-2597.7	-2597.6	-2598.9
10,000	4	1364.8	1343.5	1343.7	1344.0	-2703.8	-2645.4	-2644.8	-2645.6	1355.1	1323.2	1323.3	1324.0	-2682.5	-2595.2	-2594.9	-2596.5
Dimension $D = 6$																	
1000	2	1497.8	1384.8	1384.8	1404.6	-305.1	-274.5	-274.5	-280.0	1492.5	1369.0	1369.0	1385.0	-303.9	-270.6	-270.6	-275.0
1000	3	1419.0	1366.7	1366.8	1418.2	-284.1	-269.5	-269.5	-283.9	1409.9	1345.8	1345.9	1398.5	-281.8	-264.2	-264.2	-278.7
1000	4	1411.6	1356.6	1356.7	1375.3	-281.5	-266.7	-266.6	-272.0	1401.7	1335.2	1335.3	1349.0	-279.0	-261.3	-261.3	-265.3
10,000	2	1493.8	1390.6	1390.6	1390.9	-3043.6	-2760.5	-2760.5	-2761.7	1488.2	1379.1	1379.1	1379.4	-3033.8	-2733.3	-2733.3	-2734.1
10,000	3	1416.2	1413.1	1413.1	1415.1	-2834.2	-2823.9	-2823.8	-2830.7	1407.1	1403.5	1403.5	1405.5	-2811.5	-2799.9	-2799.9	-2806.8
10,000	4	1406.4	1356.7	1356.7	1359.7	-2802.3	-2667.6	-2667.5	-2677.7	1396.4	1334.9	1334.9	1340.1	-2777.8	-2612.6	-2612.6	-2628.2

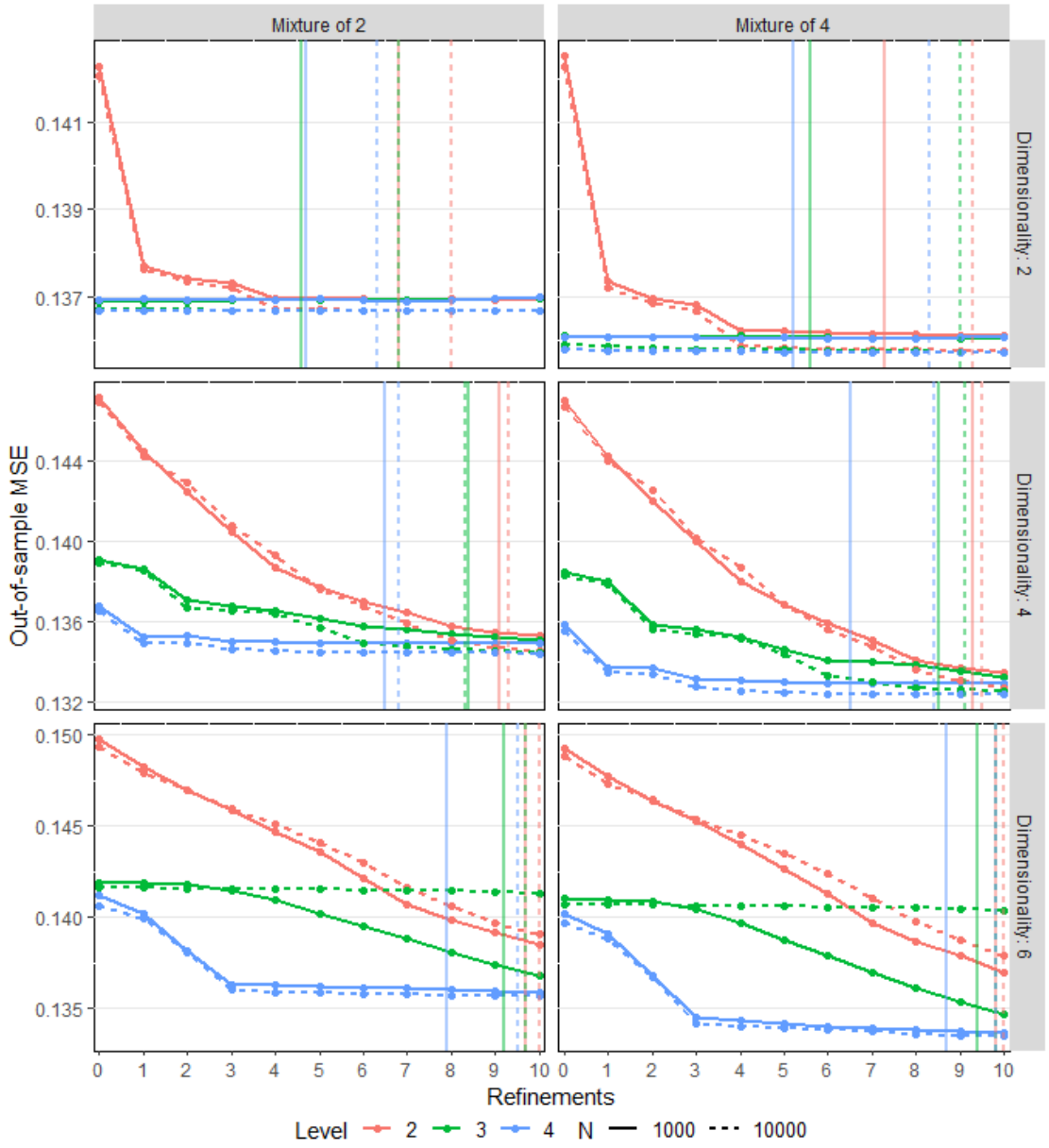
Note: The table reports the average out-of-sample mean squared error (MSE) and the average out-of-sample log-likelihood (LL) for the spatially adaptive sparse grid estimator using the out-of-sample mean squared error (MSE), the out-of-sample log-likelihood (LL), and the Akaike information criterion (AIC) for the selection of the final number of refinement steps. For every estimator five refinement steps are performed. The out-of-sample mean squared error and the out-of-sample log-likelihood are calculated with five-fold cross-validation. The grid point to be refined in every refinement step is selected according to its contribution to the local squared error.

Figure 2.13: True and Estimated Marginal CDFs of β_1 for Mixture of 2 and Mixture of 4 Normals for Spatially Adaptive Sparse Grid Estimator with Different Selection Criteria and $N = 10,000$



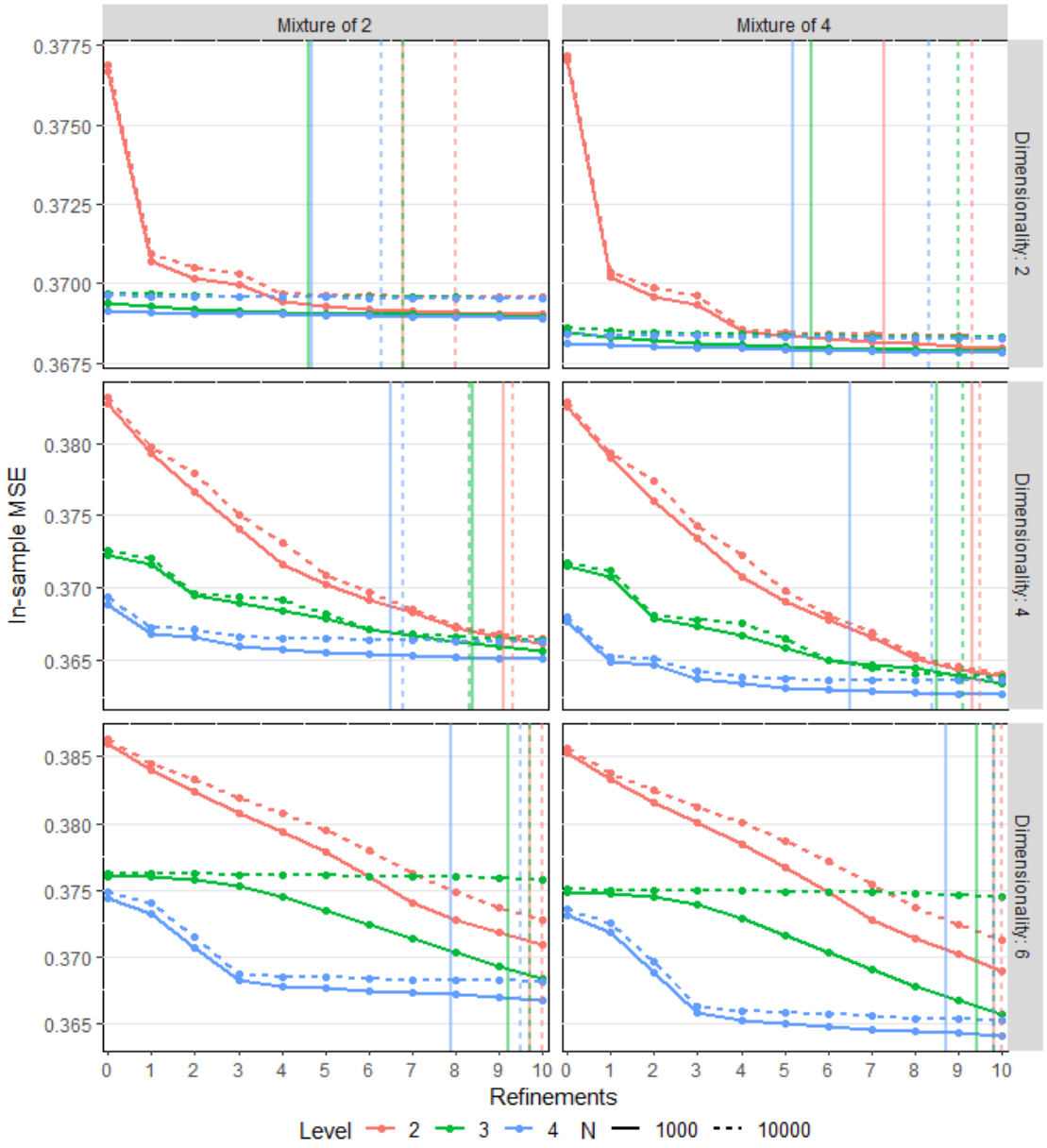
Note: The figure shows the true and estimated marginal CDFs of β_1 for the mixture of two and the mixture of four normals across models with different number of random coefficients for $N = 10,000$. The spatially adaptive sparse grid estimator refines a sparse grid of level $l_S = 4$ conducting 15 refinement steps and using the out-of-sample MSE, the out-of-sample log-likelihood (LL), and the Akaike information criterion (AIC) to select the number of refinement steps.

Figure 2.14: Average Out-of-sample MSE of Spatially Adaptive Refinement across 200 Monte Carlo Replicates



Note: The figure shows the average out-of-sample mean squared error (MSE) of the spatially adaptive refinement of sparse grids of different levels across refinement steps that is calculated via 5-fold cross-validation. The vertical lines report the average number of refinement steps that are selected based on the lowest out-of-sample MSE. The solid lines report the results for $N = 1000$, the dashed lines the results for $N = 10,000$.

Figure 2.15: Average In-sample Mean Squared Error of Spatially Adaptive Refinement across 200 Monte Carlo Replicates.



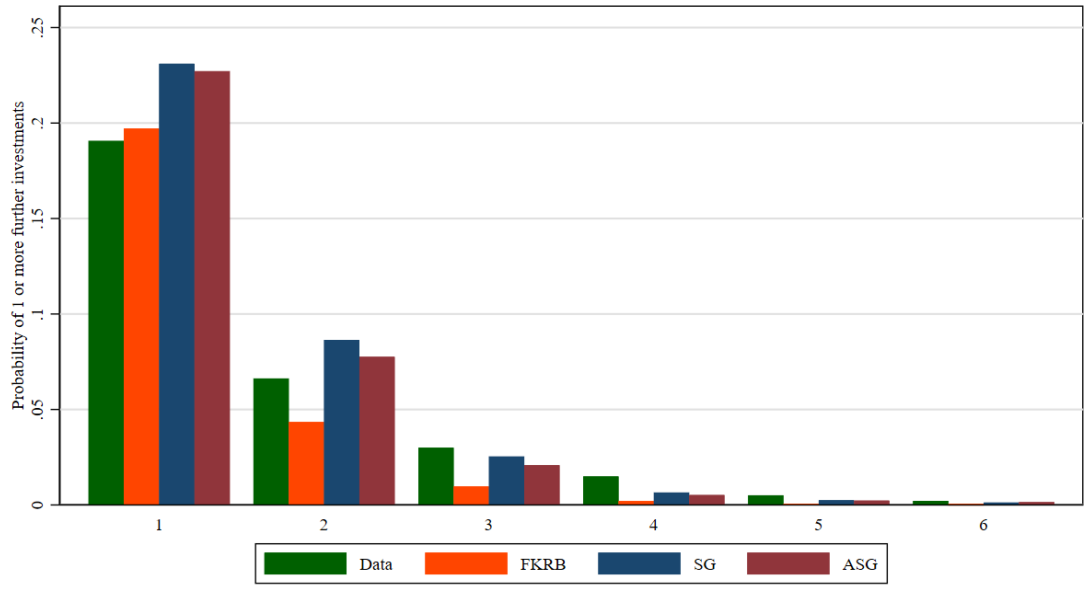
Note: The figure shows the average in-sample mean squared error (MSE) of the spatially adaptive refinement of sparse grids of different levels across refinement steps that is calculated via 5-fold cross-validation. The vertical lines report the average number of refinement steps that are selected based on the lowest out-of-sample MSE. The solid lines report the results for $N = 1000$, the dashed lines the results for $N = 10,000$.

Table 2.7: Average Number of Parameters and RMISE over 200 Monte Carlo Replicates for Sparse Grid with Mexican Hat Basis

		Mixture of 2 Normals						Mixture of 4 Normals					
		Parameters			RMISE			Parameters			RMISE		
N	q/l_S	FKRB	SG	ASG	FKRB	SG	ASG	FKRB	SG	ASG	FKRB	SG	ASG
Dimension $D = 2$													
1000	3/2	9	5	34.6	0.2068	0.0724	0.0516	9	5	39.6	0.1956	0.0882	0.0577
1000	7/3	49	17	41.2	0.0994	0.0484	0.0537	49	17	51.0	0.1021	0.0476	0.0595
1000	15/4	225	49	71.1	0.0912	0.0482	0.0571	225	49	75.1	0.0950	0.0530	0.0625
10000	3/2	9	5	43.1	0.2039	0.0706	0.0300	9	5	50.5	0.1934	0.0864	0.0435
10000	7/3	49	17	55.8	0.0843	0.0424	0.0311	49	17	65.4	0.0854	0.0439	0.0422
10000	15/4	225	49	77.4	0.0580	0.0326	0.0300	225	49	87.0	0.0646	0.0492	0.0397
Dimension $D = 4$													
1000	3/2	81	9	153.1	0.2254	0.0976	0.0578	81	9	153.5	0.2328	0.1199	0.0609
1000	7/3	2401	49	221.5	0.1273	0.0629	0.0589	2401	49	236.4	0.1241	0.0860	0.0613
1000	15/4	.	209	334.1	.	0.0495	0.0574	.	209	338.1	.	0.0684	0.0576
10000	3/2	81	9	147.8	0.2226	0.0972	0.0392	81	9	153.8	0.2316	0.1196	0.0488
10000	7/3	2401	49	211.2	0.0787	0.0623	0.0402	2401	49	228.2	0.0915	0.0857	0.0502
10000	15/4	.	209	339.9	.	0.0485	0.0364	.	209	380.5	.	0.0677	0.0472
Dimension $D = 6$													
1000	3/2	729	13	307.4	0.2156	0.0850	0.0554	729	13	303.2	0.2440	0.1098	0.0674
1000	7/3	.	97	435.6	.	0.0644	0.0540	.	97	451.9	.	0.0911	0.0605
1000	15/4	.	545	966.7	.	0.0609	0.0610	.	545	1029.4	.	0.0866	0.0643
10000	3/2	729	13	292.1	0.2138	0.0846	0.0544	729	13	276.7	0.2441	0.1095	0.0716
10000	7/3	.	97	215.0	.	0.0641	0.0606	.	97	221.0	.	0.0909	0.0859
10000	15/4	.	545	1052.2	.	0.0603	0.0680	.	545	1098.9	.	0.0861	0.0574

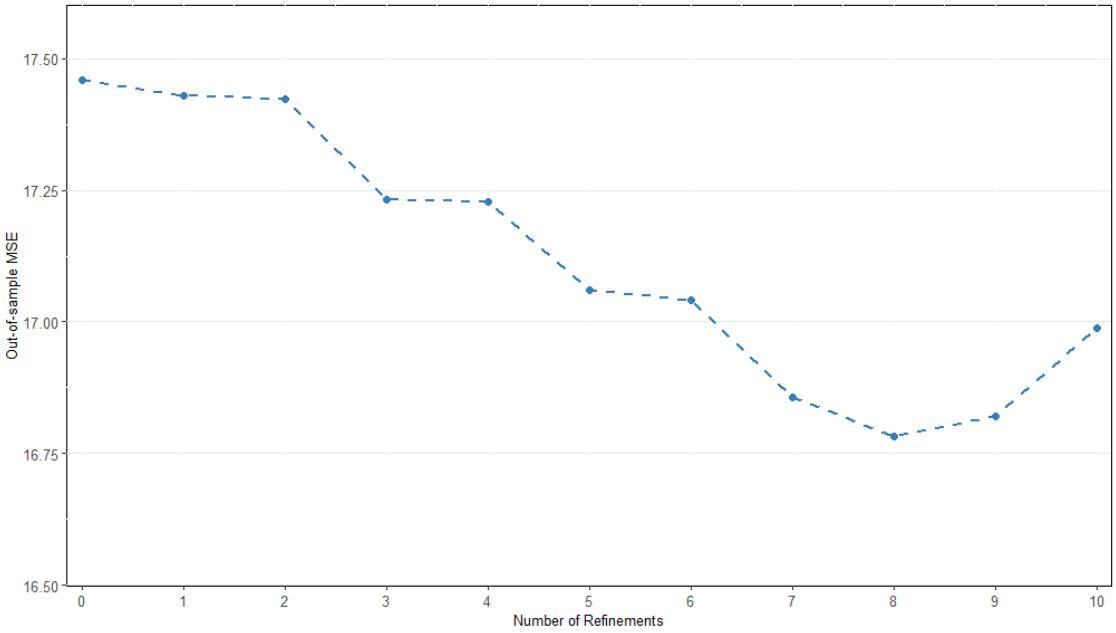
Note: The table reports the total number of parameter and the RMISE for the FKRB estimator, the sparse grid estimator (SG), and the adaptive sparse grid estimator (ASG). The adaptive sparse grid estimator performs five refinement steps, whereby the final number of refinements is determined based on the lowest out-of-sample mean squared error calculated with five-fold cross-validation. The grid point to be refined in every refinement step is selected according to its contribution to the local squared error.

Figure 2.16: Probabilities of Further Investments in the Next Six Periods



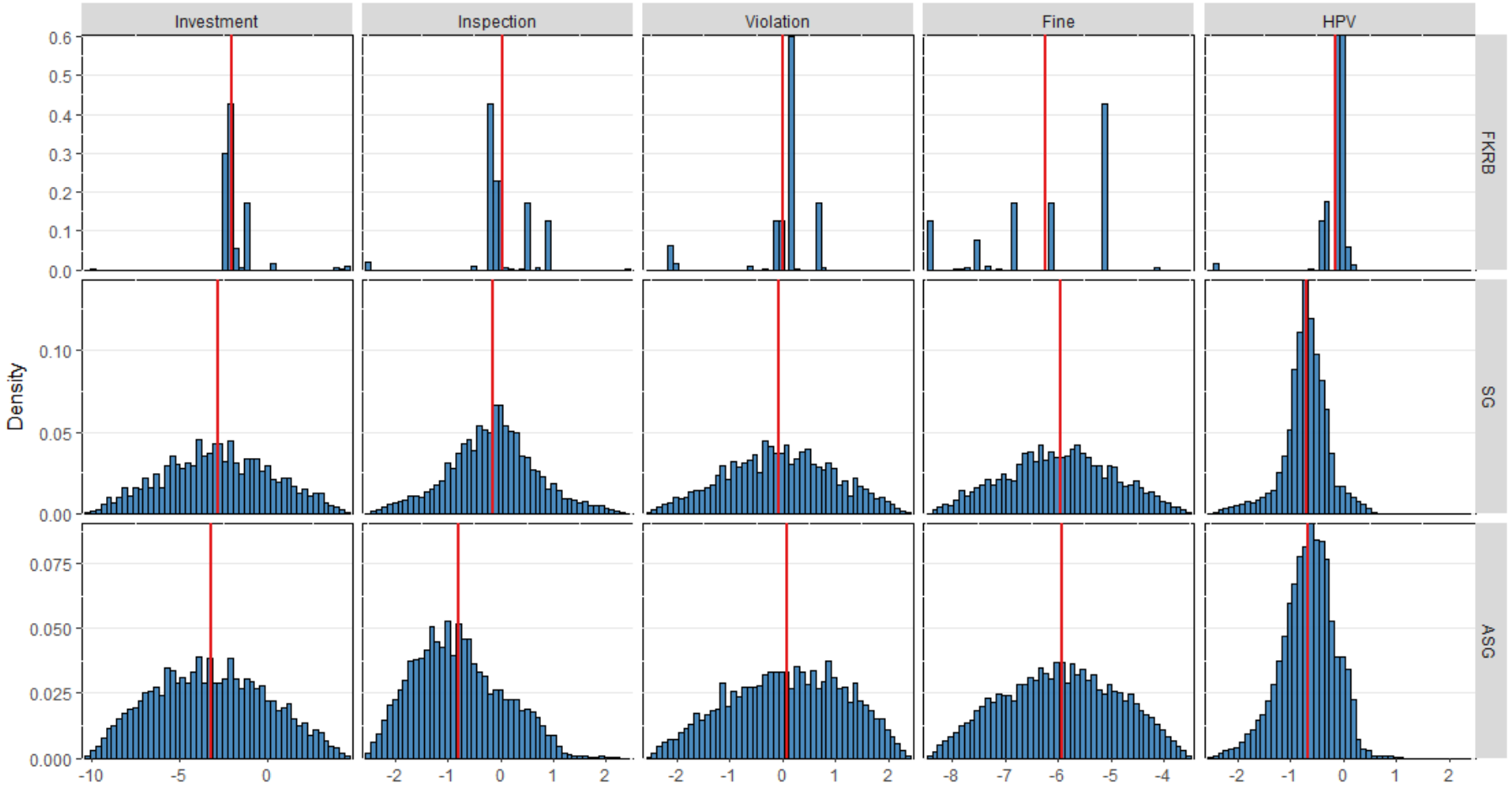
Note: The figure shows the probability of further investments of plants in the six periods after an initial investment observed in data, and predicted by the estimator of Fox et al. (2011), the sparse grid estimator of level $l = 4$, and the spatially adaptive sparse grid estimator. The spatially adaptive sparse grid estimator uses ten refinement steps whereby the finale number of refinements is determined using 5-fold cross-validation and is based on the lowest out-of-sample mean squared error.

Figure 2.17: Out-of-sample MSE of the Spatially Adaptive Refinement



Note: The figure shows the out-of-sample mean squared error (MSE) of the spatially adaptive refinement of the 5-dimensional sparse grid of level $l_S = 4$ calculated via 5-fold cross-validation. In every refinement step, the grid point with the largest contribution to the squared estimated local error is selected for refinement.

Figure 2.18: Estimated Histograms of the Five Utility Parameters



Note: The figure shows the histograms for the five utility parameter estimated with the estimator of Fox et al. (2011) (FKRB), wit the sparse grid estimator (SG), and the spatially adaptive sparse grid estimator (ASG). The red lines show the means of the estimated distribution in every of the five dimensions. The spatially adaptive sparse grid estimator uses ten refinement steps whereby the finale number of refinements is determined using 5-fold cross-validation and based on the lowest out-of-sample mean squared error.

Chapter 3

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

Co-authored by Stephan Hetzenecker

3.1 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 missspecify 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. (2021a) 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 inference statements after estimating the coefficient functions with deep learning, Farrell et al. (2021a) extend the deep learning theory for generic regression approaches developed by Farrell, Liang, and Misra (2021b) to M-estimators. Building on Chernozhukov et al. (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. (2021a) show that the inference procedure allows to construct valid inference 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 systematic part of individuals' utility into 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 as in a usual logit model. 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. (2021a) which allows for heterogeneous coefficients. In contrast, Wong and Farooq (2021) allow for a systematic 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 3.2 illustrates how deep learning can be employed to estimate heterogeneous parameters in economic models and outlines the inference and estimation procedure. Section 3.3 presents Monte Carlo experiments that study the inference procedure and Section 3.4 applies the influence function approach to real data. Section 3.5 concludes.

3.2 Deep Learning for Heterogeneity

This section introduces the methodical framework of Farrell et al. (2021a) 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 3.2.1 explains the design of the network which directly integrates the economic model specified by the researcher into

[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 elasticiy of each individual.

the network architecture. Subsection 3.2.2 explains the inference approach which is based on the concept of influence functions, and Subsection 3.2.3 lays out the estimation procedure. While the estimation and inference procedure is applicable on a wide range of models, we focus on multinomial discrete choice models when introducing the estimation procedure.

3.2.1 Deep Learning

The 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

$$\mathbb{P}\left(y_{i,j} = 1 | \boldsymbol{x}_i, \boldsymbol{w}_i\right) = \frac{\exp\left(\alpha_j\left(\boldsymbol{w}_i\right) + \boldsymbol{x}_{i,j}\boldsymbol{\beta}\left(\boldsymbol{w}_i\right)\right)}{\sum_{m=1}^{J} \exp\left(\alpha_m\left(\boldsymbol{w}_i\right) + \boldsymbol{x}_{i,m}\boldsymbol{\beta}\left(\boldsymbol{w}_i\right)\right)}. \tag{3.1}$$

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]

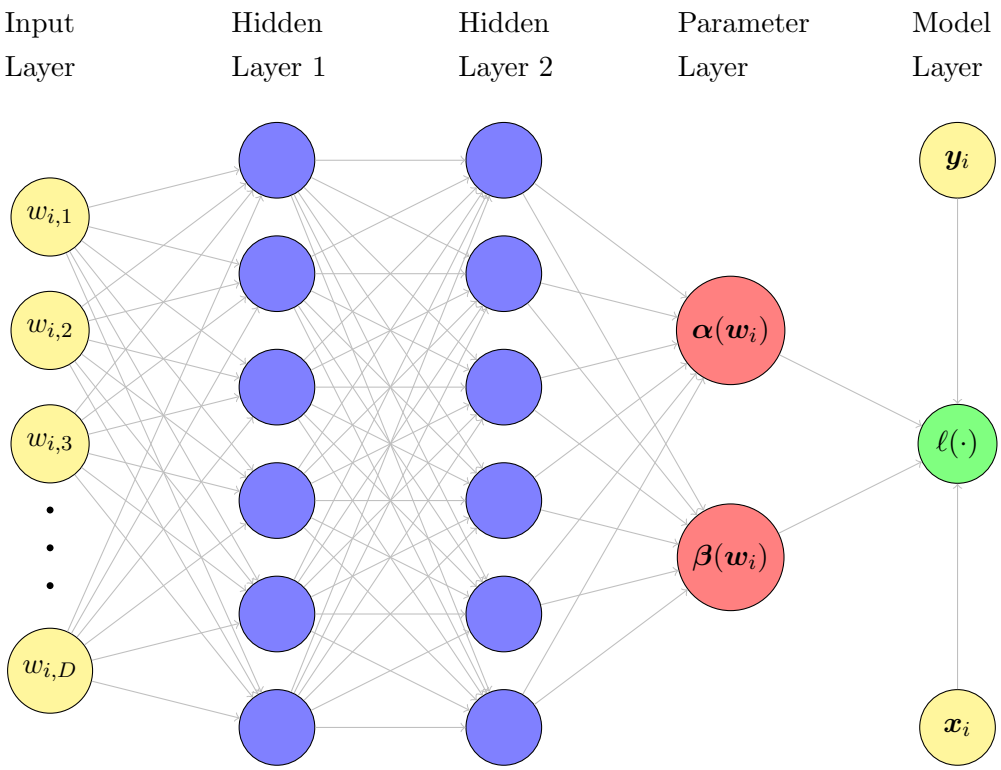
For the estimation of $\boldsymbol{\alpha}(\cdot)$ and $\boldsymbol{\beta}(\cdot)$, Farrell et al. (2021a) advocate deep neural networks. The proposed network architecture allows to combine a standard fully-connected feedforward neural network – which is used to estimate the coefficient functions $\boldsymbol{\alpha}(\cdot)$ 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 3.1 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

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

[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.

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., 2021a).

Figure 3.1: 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 (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, 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. (2021a) 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 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. (2021a). 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,

$$\ell\left(\boldsymbol{y}_i, \boldsymbol{x}_i, \boldsymbol{\delta}(\boldsymbol{w}_i)\right) = \sum_{j=1}^{J} y_{i,j} \log\left(\mathbb{P}\left(y_{i,j} = 1 | \boldsymbol{x}_i, \boldsymbol{w}_i\right)\right), \tag{3.2}$$

where $\mathbb{P}\left(y_{i,j} = 1 | \boldsymbol{x}_i, \boldsymbol{w}_i\right)$ is the conditional logit choice probability given in Equation (3.1). 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

$$\hat{\boldsymbol{\delta}}(\boldsymbol{w}_i) = \arg\max_{\boldsymbol{\delta}} \sum_{i=1}^{N} \ell\left(\boldsymbol{y}_i, \boldsymbol{x}_i, \boldsymbol{\delta}(\boldsymbol{w}_i)\right), \tag{3.3}$$

where we optimize over the class of DNNs which use the type of architecture described in Figure 3.1. 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. (2021a) 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., 2021a).

3.2.2 Inference

Inference for machine learning methods for the estimation of economic models is challenging. For inference on the coefficient functions estimated with deep learning, Farrell et al. (2021a) adopt the

semiparametric inference procedure suggested by Chernozhukov et al. (2018). The procedure 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 statistic 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 $\boldsymbol{x}^*$,

$$\theta_0 = \mathbb{E}\left[H\left(\boldsymbol{W}, \boldsymbol{\delta}\left(\boldsymbol{W}\right); \boldsymbol{x}^*\right)\right]. \tag{3.4}$$

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. (2021a) 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., 2021a).

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

$$\boldsymbol{\Lambda}(\boldsymbol{w}_i) := \mathbb{E}[\ell_{\boldsymbol{\delta},\boldsymbol{\delta}}(\boldsymbol{Y}, \boldsymbol{X}, \boldsymbol{\delta}(\boldsymbol{W}))|\boldsymbol{W} = \boldsymbol{w}_i], \tag{3.5}$$

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 $\psi(\boldsymbol{w}, \boldsymbol{\delta}, \boldsymbol{\Lambda}) - \theta_0$, where

$$\psi\left(\boldsymbol{w}_i, \boldsymbol{\delta}, \boldsymbol{\Lambda}\right) = H\left(\boldsymbol{w}_i, \boldsymbol{\delta}\left(\boldsymbol{w}_i\right); \boldsymbol{x}^*\right) - H_{\boldsymbol{\delta}}\left(\boldsymbol{w}_i, \boldsymbol{\delta}\left(\boldsymbol{w}_i\right); \boldsymbol{x}^*\right)' \boldsymbol{\Lambda}\left(\boldsymbol{w}_i\right)^{-1} \boldsymbol{\ell}_{\boldsymbol{\delta}}\left(\boldsymbol{y}_i, \boldsymbol{x}_i, \boldsymbol{\delta}(\boldsymbol{w}_i)\right) \tag{3.6}$$

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 (3.1), the gradient vector $\boldsymbol{\ell}_{\boldsymbol{\delta}}(\boldsymbol{y}_i, \boldsymbol{x}_i, \boldsymbol{\delta}(\boldsymbol{w}_i))$ for individual i is

$$\boldsymbol{\ell}_{\boldsymbol{\delta}}(\boldsymbol{y}_i, \boldsymbol{x}_i, \boldsymbol{\delta}(\boldsymbol{w}_i)) = (c_{i,1}, \ldots, c_{i,J}, \tilde{c}_{i,1}, \ldots, \tilde{c}_{i,K})' \tag{3.7}$$

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 (3.1) is given in Farrell et al. (2021a, 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 $H(\boldsymbol{w}_i, \hat{\boldsymbol{\delta}}(\boldsymbol{w}_i); \boldsymbol{x}^*)$, while treating $\hat{\boldsymbol{\delta}}(\boldsymbol{w})$ estimated with the sample as fixed. 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 (3.6), the sample analogue of $\mathbb{E}\big[\psi\big(\boldsymbol{W}, \hat{\boldsymbol{\delta}}\left(\boldsymbol{W}\right), \hat{\boldsymbol{\Lambda}}\left(\boldsymbol{W}\right)\big)\big]$ is

$$\begin{aligned}
\hat{\theta}_{IF} =& \frac{1}{N}\sum_{i=1}^{N}\psi\left(\boldsymbol{w}_i, \hat{\boldsymbol{\delta}}(\boldsymbol{w}_i), \hat{\boldsymbol{\Lambda}}(\boldsymbol{w}_i)\right) \\
=& \frac{1}{N}\sum_{i=1}^{N} H\left(\boldsymbol{w}_i, \hat{\boldsymbol{\delta}}\left(\boldsymbol{w}_i\right); \boldsymbol{x}^*\right) && \text{(3.8a)} \\
&- \frac{1}{N}\sum_{i=1}^{N} H_{\hat{\boldsymbol{\delta}}}\left(\boldsymbol{w}_i, \hat{\boldsymbol{\delta}}\left(\boldsymbol{w}_i\right); \boldsymbol{x}^*\right)' \hat{\boldsymbol{\Lambda}}\left(\boldsymbol{w}_i\right)^{-1} \boldsymbol{\ell}_{\boldsymbol{\delta}}\left(\boldsymbol{y}_i, \boldsymbol{x}_i, \hat{\boldsymbol{\delta}}(\boldsymbol{w}_i)\right). && \text{(3.8b)}
\end{aligned}$$

Similarly to $\hat{\theta}_{PI}$, the term in Equation (3.8a) 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 (3.8b) is an additional correction term that includes an estimate of the nuisance function $\boldsymbol{\Lambda}(\boldsymbol{w})$ 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 $\Lambda(\boldsymbol{w})^{-1}\boldsymbol{\ell}_{\boldsymbol{\delta}}(\boldsymbol{w}, \boldsymbol{\delta}(\boldsymbol{w}))$ 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 (3.1), the assumption implies that the choice probabilities are bounded away from zero and one.[5]

3.2.3 Estimation

With the influence function in Equation (3.6), 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 $\psi(\boldsymbol{w}_i, \boldsymbol{\delta}, \boldsymbol{\Lambda})$ 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., 2021a).

For the conditional expected individual Hessian matrix 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., 2021a). The procedure thus consists of the following steps:

1. Split the observation units $\{1, \ldots, n\}$ into S subsets, denoted by $\mathcal{S}_s \subset \{1, \ldots, n\}$, $s = 1, \ldots, S$.

2. For each $s = 1, \ldots, S$, 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

$$\hat{\theta} = \frac{1}{S}\sum \hat{\theta}_s, \qquad \hat{\theta}_s = \frac{1}{|\mathcal{S}_s|}\sum_{i \in \mathcal{S}_s} \psi\left(\boldsymbol{w}_i, \hat{\boldsymbol{\delta}}_s, \hat{\boldsymbol{\Lambda}}_s\right), \tag{3.9}$$

   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 variance-analogue

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

of Equation (3.9)

$$\hat{\Psi} = \frac{1}{S}\sum_{s=1}^{S} \hat{\Psi}_s, \qquad \hat{\Psi}_s = \frac{1}{|\mathcal{S}_s|}\sum_{i\in\mathcal{S}_s}\left(\psi\left(\boldsymbol{w}_i, \hat{\boldsymbol{\delta}}_s, \hat{\boldsymbol{\Lambda}}_s\right) - \hat{\theta}\right)^2. \tag{3.10}$$

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

$$\sqrt{n}\hat{\Psi}^{-1/2}\left(\hat{\theta} - \theta\right) \to_d \mathcal{N}(0,1). \tag{3.11}$$

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 $\hat{\boldsymbol{\Lambda}}(\boldsymbol{w}_i)$ is a prediction problem for which different machine learning methods can be used. In the Monte Carlo experiments and application presented below, we estimate $\hat{\boldsymbol{\Lambda}}(\boldsymbol{w}_i)$ 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., 2021a).[6]

Remark 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 (3.9) and (3.10) for repetition $r = 1, \ldots, R$. Then, the final estimator is the median over the repetitions,[7] 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 (3.11), 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.

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

[7] 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.

3.3 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 3.2. 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.[8] 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.[9] 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.[10]

For the data generation, we consider an individual-level discrete choice demand model of the form presented in Equation 3.1. 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).[11] Each traveler chooses the travel mode among the three alternatives car, Swissmetro, and train that provides her with the highest utility,

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

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.[12] 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

$$\begin{aligned}\alpha_{\text{train}}\left(\boldsymbol{w}_i\right) &= -1 + 1 \cdot income_i,\\ \alpha_{\text{sm}}\left(\boldsymbol{w}_i\right) &= -3 + 1 \cdot age_i,\end{aligned}$$

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

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

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

[11] The travel cost, travel time, and frequency variables are scaled downwards by factor 100 (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.

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

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

$$\begin{aligned}
\beta^{\text{cost}}(\boldsymbol{w}_i) &= -6 + income_i - 0.8 \cdot who_i^1 - 1 \cdot who_i^2 - 1.2 \cdot who_i^3 \\
\beta^{\text{freq}}(\boldsymbol{w}_i) &= -5 + income_i + 0.9 \cdot male_i \\
\beta^{\text{time}}(\boldsymbol{w}_i) &= -6 + 1 \cdot age_i.
\end{aligned} \tag{3.12}$$

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.

3.3.1 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. (2021a). 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 $\boldsymbol{\delta}_s(\boldsymbol{w})$, and the second one to estimate $\boldsymbol{\Lambda}_s(\boldsymbol{w})$.

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. (2021a). 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 3.1 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 at all, 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

Table 3.1: 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
$\theta_{cost} \in \hat{CI}_{cost}$	0.95	0.00	0.93	0.92	0.83	0.40	0.99
$\theta_{freq} \in \hat{CI}_{freq}$	0.95	0.00	0.93	0.92	0.89	0.68	1.00
$\theta_{time} \in \hat{CI}_{time}$	0.94	0.00	0.93	0.94	0.88	0.54	1.00
$\hat{se}_{cost}$	0.07	0.05	6.85	3.67	1.70	0.75	0.61
$\hat{se}_{freq}$	0.10	0.07	5.25	8.19	2.26	1.24	3.56
$\hat{se}_{time}$	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

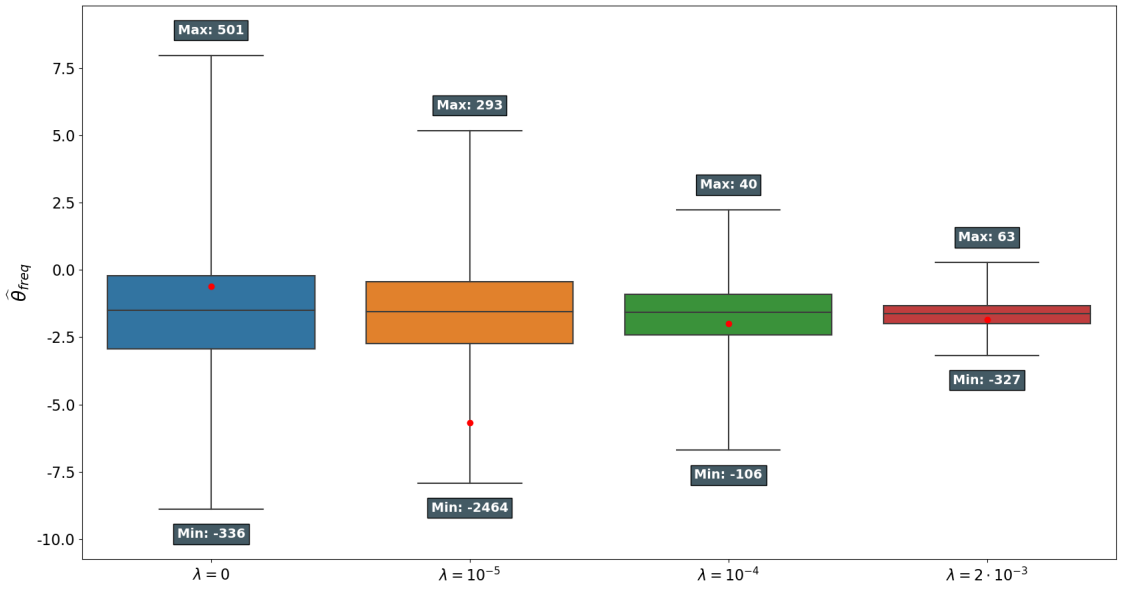
Note: The table reports the average summary statistics over all Monte Carlo replicates for the conditional logit using the true specification (Oracle), the conditional logit using the three variables of interest for the estimation (Basic), the influence function approach using five different values for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

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 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.

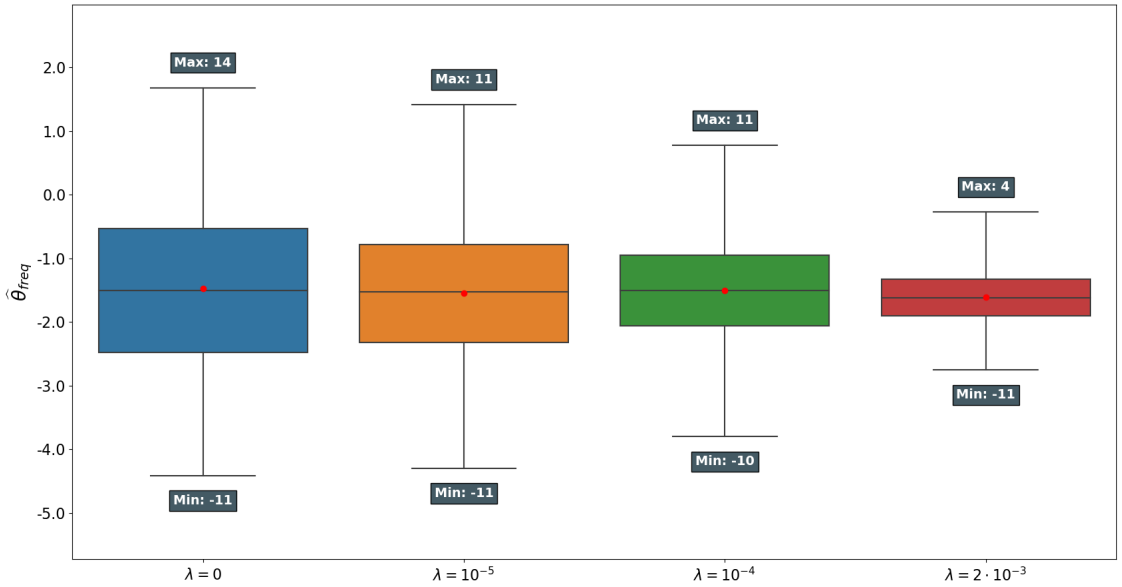
Overall, choosing $\lambda > 0$ leads to more precise estimates of the true average coefficients (considering all three coefficients together, the estimates are most precise for $\lambda = 10^{-4}$), and to smaller estimated standard errors. However, the bias of the estimated average coefficients remains 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 3.2. 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.

Figure 3.2: 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.

The median biases and estimated standard errors across Monte Carlo replicates reported in Table 3.7 in the appendix 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 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.[13]

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 3.3.2 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 1. Table 3.2 reports the results for the sample splitting procedure with $R = 5$ repetitions.[14] 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 3.2 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 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

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

[14] 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 3.2: Average Summary Statistics of 1000 Monte Carlo Replicates for Small Data and Repeated Sample Splitting with $R = 5$

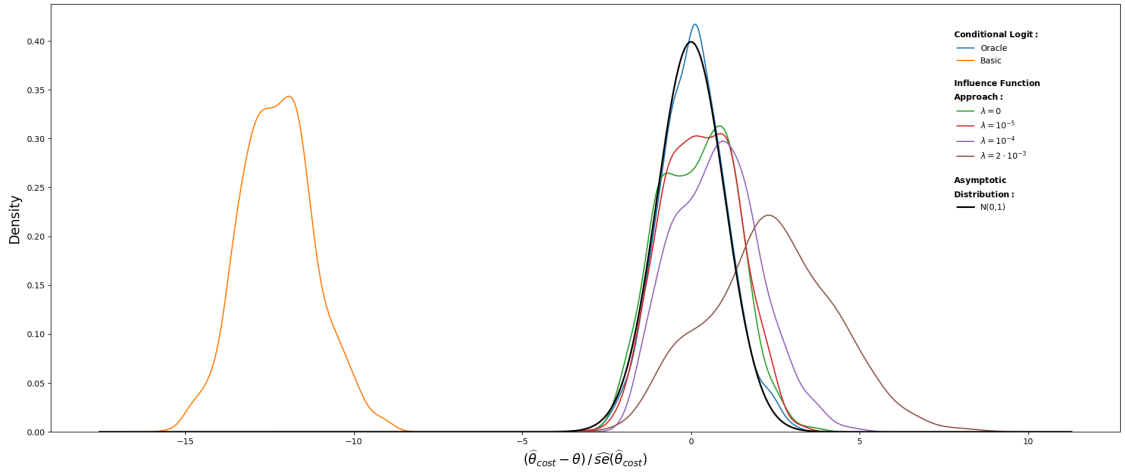
	Conditional Logit		Influence Function Approach with λ equal to				
	Oracle	Basic	0	10^{-5}	10^{-4}	$2 \cdot 10^{-3}$	NN
$\theta_{cost} \in \hat{CI}_{cost}$	0.94	0.00	0.94	0.93	0.82	0.42	0.99
$\theta_{freq} \in \hat{CI}_{freq}$	0.96	0.00	0.94	0.92	0.90	0.66	1.00
$\theta_{time} \in \hat{CI}_{time}$	0.96	0.00	0.95	0.95	0.87	0.59	1.00
$\hat{se}_{cost}$	0.07	0.05	1.61	1.34	0.72	0.32	0.61
$\hat{se}_{freq}$	0.10	0.07	1.91	1.64	1.10	0.58	3.65
$\hat{se}_{time}$	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 $\Lambda_s(\boldsymbol{w})$, and the neural network (NN), which uses robust standard erros and does not rely on the influence function approach.

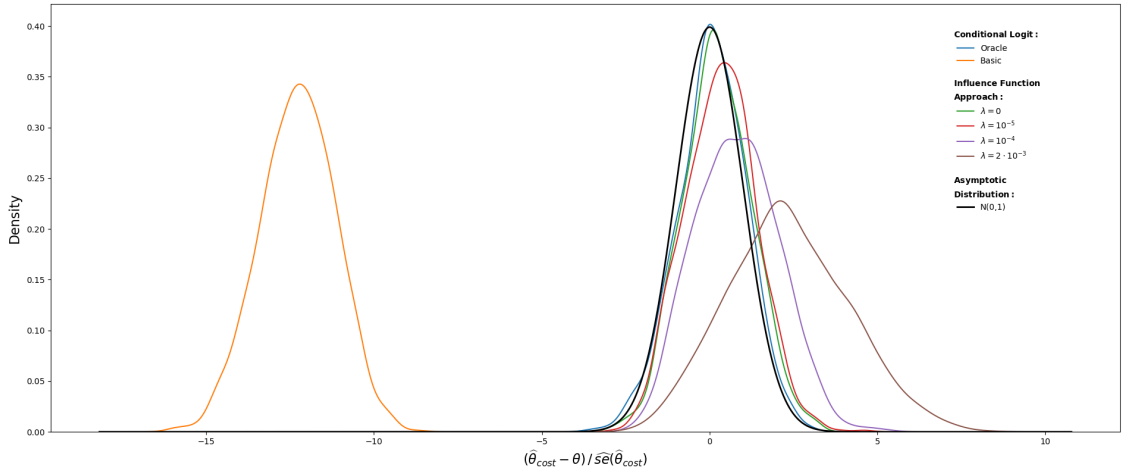
increasing λ leads to inappropriately small confidence intervals centered around biased estimates and, hence, to a poor coverage of the true values. Based on these results, we do not recommend using regularization in the form of a l_2-penalty with $\lambda > 0$ in the network used to estimate $\boldsymbol{\Lambda}(\boldsymbol{w}_i)$ 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 3.3 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 (3.11). 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 $\lambda = 10^{-5}$.

Figure 3.3: Density of Estimated t-Statistic of $\hat{\theta}_{cost}$ for Different Estimators



(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.

Remark 2. Beyond the repeated sample splitting, we conduct several other adjustments of the estimation procedure that are intended to reduce outliers in some Monte Carlo replicates. We considered taking the median instead of the average in Equation (3.9) and (3.10), 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 the the bias remains large. Furthermore, we also apply the modification suggested by Farrell et al. (2021a) and add a constant c to the diagonal elements of $\hat{\boldsymbol{\Lambda}}_s$. 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.

3.3.2 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 Swiss Metro 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.

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 3.3 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

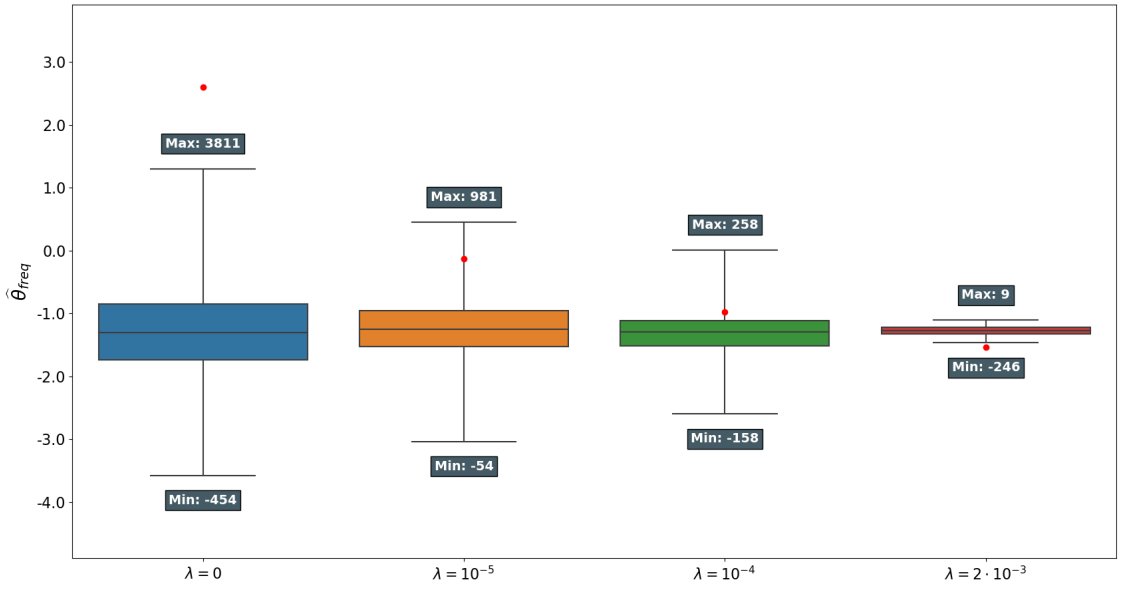
Table 3.3: Average Summary Statistics of 1000 Monte Carlo Replicates for Large Data and Repeated Sample Splitting with $R = 5$

	Conditional Logit		Influence Function Approach with λ equal to				
	Oracle	Basic	0	10^{-5}	10^{-4}	$2 \cdot 10^{-3}$	NN
$\theta_{cost} \in \hat{CI}_{cost}$	0.95	0.00	0.94	0.92	0.91	0.76	1.00
$\theta_{freq} \in \hat{CI}_{freq}$	0.95	0.00	0.95	0.93	0.90	0.83	1.00
$\theta_{time} \in \hat{CI}_{time}$	0.94	0.00	0.95	0.94	0.92	0.86	1.00
$\hat{se}_{cost}$	0.02	0.02	0.50	0.41	0.35	0.12	0.49
$\hat{se}_{freq}$	0.04	0.04	0.80	0.59	0.44	0.14	1.72
$\hat{se}_{time}$	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

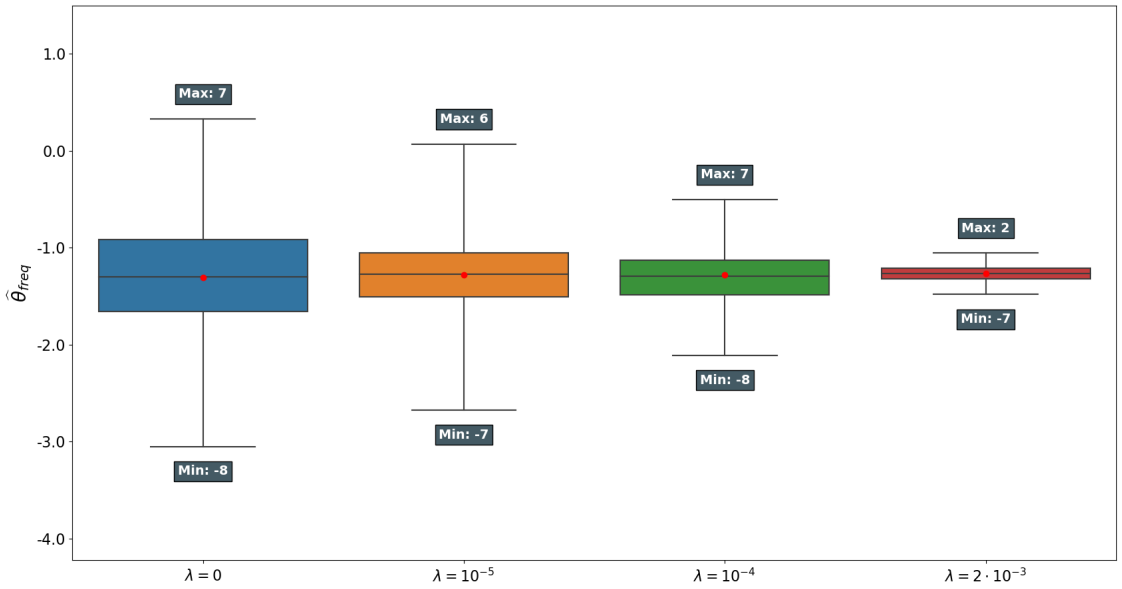
Note: The table reports the average summary statistics over all Monte Carlo replicates for the conditional logit using the true specification (Oracle), the conditional logit using the three variables of interest for the estimation (Basic), the influence function approach, using five different values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

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.

Figure 3.4: 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.

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 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 3.8 and Table 3.10 and Figure 3.4).

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.

3.4 Application

This sections applies the estimation procedure presented in Section 3.2 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,

$$\boldsymbol{\delta}(\boldsymbol{w}_i) = \left(\alpha_{\text{train}}(\boldsymbol{w}_i), \alpha_{\text{sm}}(\boldsymbol{w}_i), \beta^{\text{cost}}(\boldsymbol{w}_i), \beta^{\text{freq}}(\boldsymbol{w}_i), \beta^{\text{time}}(\boldsymbol{w}_i)\right)'.$$

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 other nest (representing the newly introduced alternative).[15] 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 $\boldsymbol{w}_i$.[16] 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 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 3.4 reports the average heterogeneous coefficient functions for the travel cost, frequency, and travel time and their corresponding estimated standard errors. Additionally, we calculate the 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 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.[17] 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 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

[15] 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).

[16] Interacting the alternative-specific constants with $\boldsymbol{w}_i$ yields multinomial coefficients for each variable in $\boldsymbol{w}_i$.

[17] Frequency is calculated as average minutes of waiting time for a given transportation mode, i.e., a higher frequency variable implies less frequent connections.

Table 3.4: 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	-1943
$\hat{\theta}_{freq}$	-0.891	-0.966	-1.040	-1106
$\hat{\theta}_{time}$	-1.368	-1.728	-1.797	-2172
$\hat{se}_{cost}$	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	-0638
LL^{Test}	-0.777	-0.772	-0.753	-0695

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).

might be needed for the influence function approach than for traditional logit models.

Figure 3.5 plots the histograms of the predicted coefficients using the test set for the influence function approach (blue bars) and the nested logit model (green bars). First, the plots reveal that there is substantial heterogeneity across travelers. Second, the heterogeneity in the intercept functions 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 coefficients for the travel cost, frequency, and travel time predicted by the more flexible influence approach deviates to a larger extent from the coefficients predicted by the nested logit model – especially for the travel time coefficient.

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

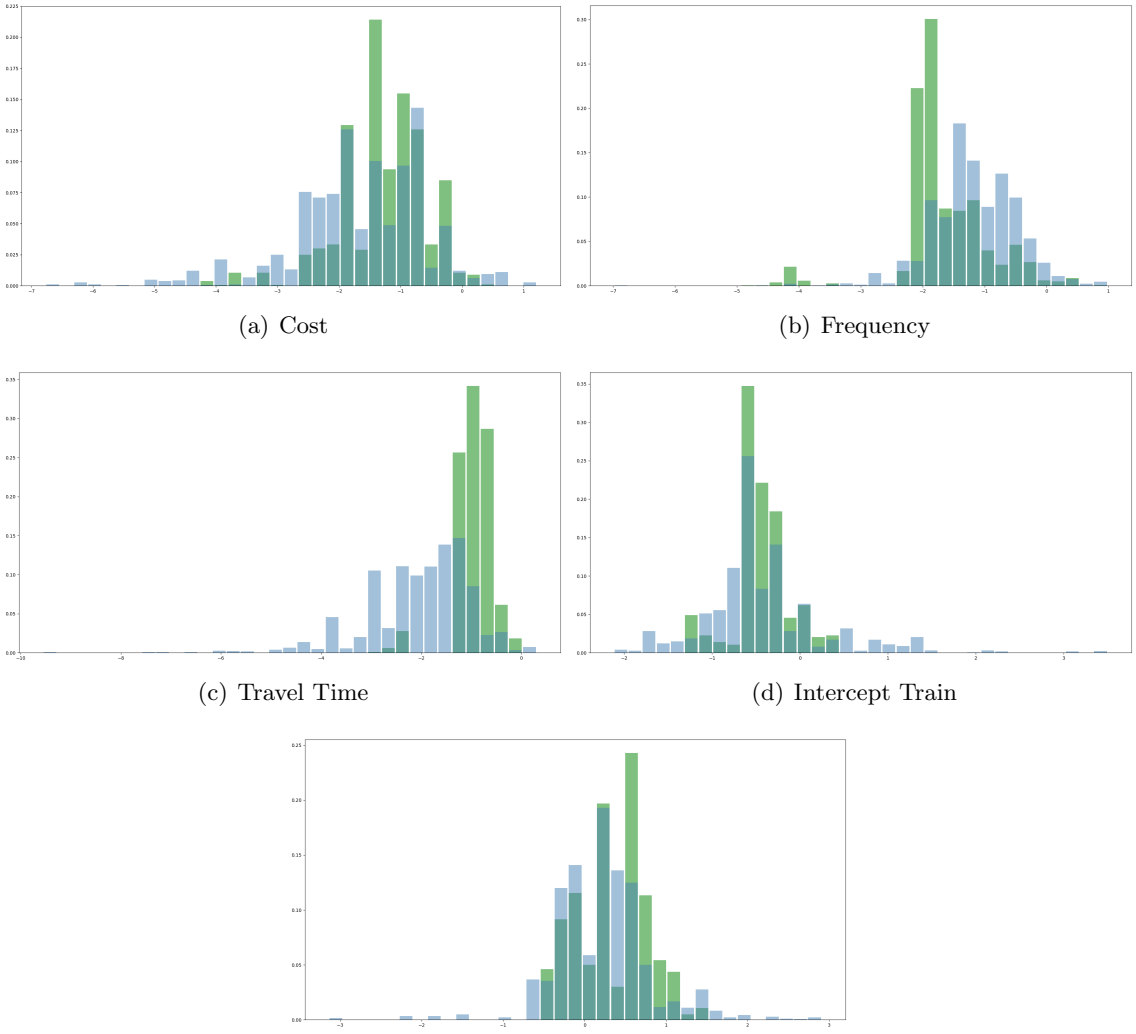
$$\theta_0^{l,m} = E\left[H^{l,m}(\boldsymbol{w_i}, \boldsymbol{\delta}(\boldsymbol{w}_i); \boldsymbol{x}^*)\right]$$

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 $l, m \in \{\text{car}, \text{train}, \text{sm}\}$.

Figure 3.5: Histograms of Estimated Coefficient Functions for Influence Function Approach and Nested Logit Model



(a) Cost (b) Frequency

(c) Travel Time (d) Intercept Train

(e) Intercept SM

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 with repeated sample splitting with $R = 5$.

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 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.[18]

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,

[18]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.

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.

Table 3.5: 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.

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 (3.1), 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 estimated average coefficients – and are only slightly larger than in the conditional logit model and even smaller for some own- and cross-elasticities.

3.5 Conclusion

This paper investigates the finite sample performance of the estimation approach of Farrell et al. (2021a) 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 inference 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 inference statements. Second, we observe that the influence function proposed for the construction of valid inference 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 inference 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).

Appendix: Additional Tables and Figures

Table 3.6: 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
$\theta_{cost} \in \hat{CI}_{cost}$	0.95	0.00	0.93	0.92	0.83	0.40	0.99
$\theta_{freq} \in \hat{CI}_{freq}$	0.95	0.00	0.93	0.92	0.89	0.68	1.00
$\theta_{time} \in \hat{CI}_{time}$	0.94	0.00	0.93	0.94	0.88	0.54	1.00
$\hat{se}_{cost}$	0.07	0.05	1.36	1.02	0.53	0.18	0.60
$\hat{se}_{freq}$	0.10	0.07	1.62	1.34	0.84	0.34	3.29
$\hat{se}_{time}$	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 3.7: 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
$\theta_{cost} \in \hat{CI}_{cost}$	0.94	0.00	0.94	0.93	0.82	0.42	0.99
$\theta_{freq} \in \hat{CI}_{freq}$	0.96	0.00	0.94	0.92	0.90	0.66	1.00
$\theta_{time} \in \hat{CI}_{time}$	0.96	0.00	0.95	0.95	0.87	0.59	1.00
$\hat{se}_{cost}$	0.07	0.05	1.20	0.96	0.46	0.18	0.60
$\hat{se}_{freq}$	0.10	0.07	1.48	1.16	0.73	0.33	3.45
$\hat{se}_{time}$	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 3.8: 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
$\theta_{cost} \in \hat{CI}_{cost}$	0.95	0.00	0.94	0.93	0.90	0.75	1.00
$\theta_{freq} \in \hat{CI}_{freq}$	0.95	0.00	0.92	0.94	0.91	0.83	1.00
$\theta_{time} \in \hat{CI}_{time}$	0.95	0.00	0.94	0.94	0.92	0.85	1.00
$\hat{se}_{cost}$	0.02	0.02	3.43	1.45	1.15	0.24	0.49
$\hat{se}_{freq}$	0.04	0.04	8.24	2.03	1.60	0.30	1.73
$\hat{se}_{time}$	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 3.9: 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
$\theta_{cost} \in \hat{CI}_{cost}$	0.95	0.00	0.94	0.93	0.90	0.75	1.00
$\theta_{freq} \in \hat{CI}_{freq}$	0.95	0.00	0.92	0.94	0.91	0.83	1.00
$\theta_{time} \in \hat{CI}_{time}$	0.95	0.00	0.94	0.94	0.92	0.85	1.00
$\hat{se}_{cost}$	0.02	0.02	0.35	0.25	0.19	0.04	0.49
$\hat{se}_{freq}$	0.04	0.04	0.58	0.37	0.23	0.06	1.72
$\hat{se}_{time}$	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

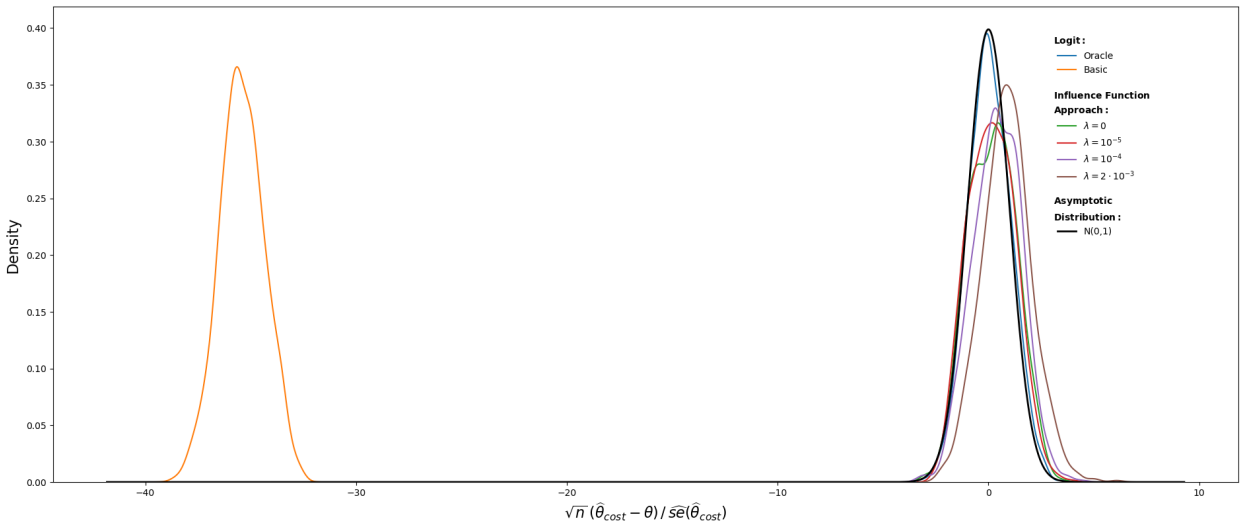
Note: The table reports the median of the variables $\hat{se}_i$, $BIAS_i$, $MSE(\Lambda)^{Train}$, and $MSE(\Lambda)^{Test}$ and the average of the variables $\theta_i \in \hat{CI}_i$, Rej. $\theta_i = 0$, and Share Outlier, $i \in \{cost, freq, time\}$, over all Monte Carlo replicates for the conditional 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 3.10: Median Summary Statistics of 1000 Monte Carlo Replicates for Large Data and Repeated Sample Splitting with $R = 5$

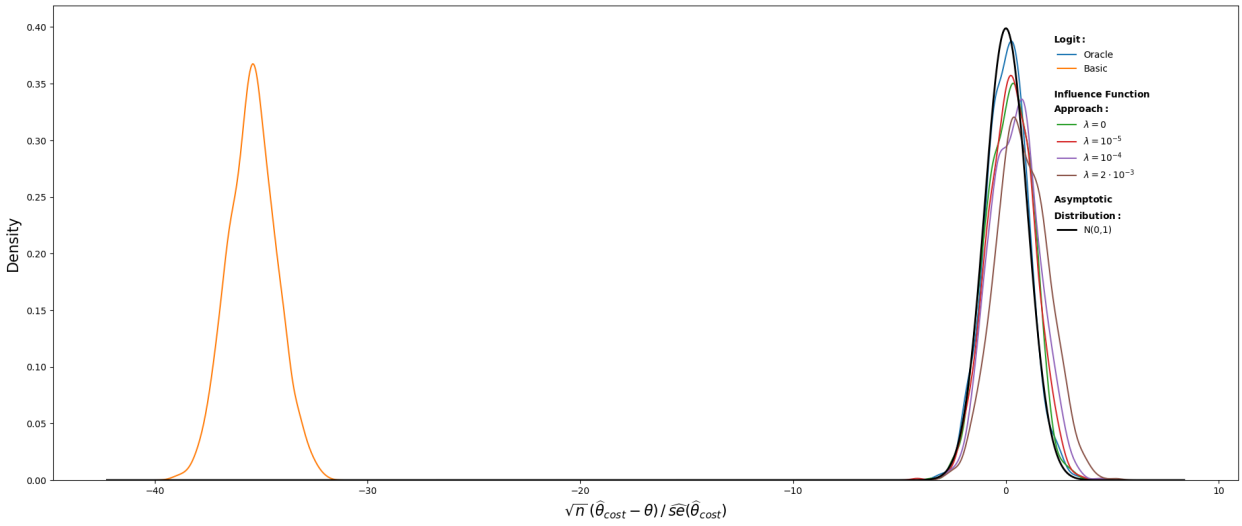
	Conditional Logit		Influence Function Approach with λ equal to				
	Oracle	Basic	0	10^{-5}	10^{-4}	$2 \cdot 10^{-3}$	NN
$\theta_{cost} \in \hat{CI}_{cost}$	0.95	0.00	0.94	0.92	0.91	0.76	1.00
$\theta_{freq} \in \hat{CI}_{freq}$	0.95	0.00	0.95	0.93	0.90	0.83	1.00
$\theta_{time} \in \hat{CI}_{time}$	0.94	0.00	0.95	0.94	0.92	0.86	1.00
$\hat{se}_{cost}$	0.02	0.02	0.30	0.23	0.17	0.04	0.49
$\hat{se}_{freq}$	0.04	0.04	0.51	0.33	0.22	0.06	1.72
$\hat{se}_{time}$	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 3.6: Density of Estimated t-Statistic of $\hat{\theta}_{cost}$ for Different Estimators and Large Data



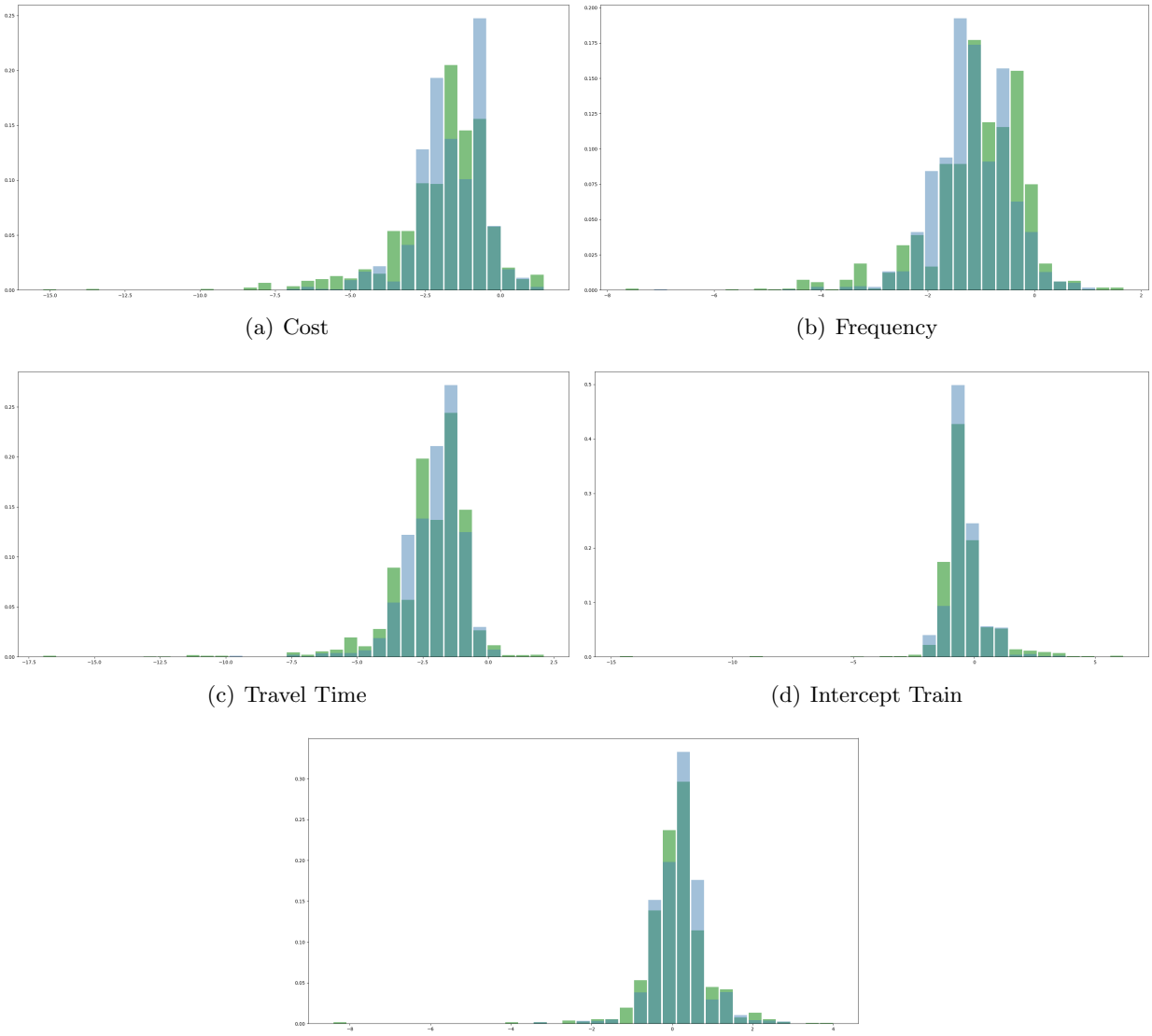
(a) No repeated sample splitting ($R = 1$)



(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 3.7: Histograms of Estimated Coefficient Functions for Influence Function Approach and Neural Network



(a) Cost

(b) Frequency

(c) Travel Time

(d) Intercept Train

(e) Intercept SM

Note: The green bars represent the heterogeneous coefficients in the test set predicted with the neural network model (without the influence function approach), and the blue bars the heterogeneous coefficients in the test set predicted with the influence function approach with repeated sample splitting with $R = 5$.

Table 3.11: Estimation Results Logit Models

	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

References

Antonini, G., Gioia, C., and Frejinger, E. (2007). Swissmetro: description of the data.

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. (1997a). 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. (1997b). An endogenous segmentation mode choice model with an application to intercity travel. *Transportation science*, *31*(1), 34–48.

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..

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.

Brumm, J., and Scheidegger, S. (2017). Using adaptive sparse grids to solve high-dimensional dynamic models. *Econometrica*, *85*(5), 1575–1612.

Bungartz, H.-J., and Griebel, M. (2004). Sparse grids. *Acta Numerica*, *13*, 147–269.

Burda, M., Harding, M., and Hausman, J. (2008). A Bayesian mixed logit–probit model for multinomial choice. *Journal of Econometrics*, *147*(2), 232–246.

Chen, X. (2007). Large sample sieve estimation of semi-nonparametric models. *Handbook of econometrics*, *6*, 5549–5632.

Chernozhukov, V., Chetverikov, D., Demirer, M., Duflo, E., Hansen, C., Newey, W., and Robins, J. (2018, 11). Double/debiased machine learning for treatment and structural parameters. *The Econometrics Journal*, *21*, C1-C68. doi: 10.1111/ectj.12097

Cohen, A., Davenport, M. A., and Leviatan, D. (2013). On the stability and accuracy of least squares approximations. *Foundations of computational mathematics*, *13*(5), 819–834.

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 [Computer software manual]. Retrieved from `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.

Dezeure, R., Bühlmann, P., and Zhang, C.-H. (2017). High-dimensional simultaneous inference with the bootstrap. *Test*, *26*(4), 685–719.

Efron, B. (1979). Bootstrap methods: Another look at the jackknife. *The Annals of Statistics*, 7, 1 – 26.

Efron, B., Hastie, T., Johnstone, I., Tibshirani, R., and Others. (2004). Least angle regression. *The Annals of statistics*, *32*(2), 407–499.

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* (pp. 14–22). New York, NY, USA: ACM.

Fan, J., Ke, Y., and Wang, K. (2020). Factor-adjusted regularized model selection. *Journal of Econometrics*, *216*(1), 71–85.

Farrell, M. H., Liang, T., and Misra, S. (2021a). Deep learning for individual heterogeneity: An automatic inference framework. *arXiv preprint arXiv:2010.14694v2*.

Farrell, M. H., Liang, T., and Misra, S. (2021b). Deep neural networks for estimation and inference. *Econometrica*, *89*, 181-213.

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.

Franzelin, F., and Pflüger, D. (2016). From data to uncertainty: An efficient integrated data-driven sparse grid approach to propagate uncertainty. In J. Garcke and D. Pflüger (Eds.), *Sparse grids and applications - stuttgart 2014* (pp. 29–49). Springer International Publishing.

Garcke, J. (2013). Sparse grids in a nutshell. In J. Garcke and M. Griebel (Eds.), *Sparse grids and applications* (pp. 57–80). Berlin, Heidelberg: Springer Berlin Heidelberg.

Gentle, J. E. (2007). *Matrix Algebra: Theory, Computations, and Applications in Statistics* (1st ed.). Springer Publishing Company, Incorporated.

Goodfellow, I., Bengio, Y., and Courville, A. (2016). *Deep learning*. MIT Press. (http://www.deeplearningbook.org)

Greene, W. H., Hensher, D. A., and Rose, J. (2006). Accounting for heterogeneity in the variance of unobserved effects in mixed logit models. *Transportation Research Part B: Methodological*, *40*(1), 75-92.

Hansen, B. E. (2014). Nonparametric sieve regression: Least squares, averaging least squares, and cross-validation. *Handbook of Applied Nonparametric and Semiparametric Econometrics and Statistics*.

Hastie, T., Tibshirani, R., and Friedman, J. (2009). *The elements of statistical learning: data mining, inference and prediction* (2nd ed.). 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*(none), 1184 – 1226.

Heiss, F., Hetzenecker, S., and Osterhaus, M. (2021). Nonparametric estimation of the random coefficients model: An elastic net approach. *Journal of Econometrics*.

Heiss, F., and Winschel, V. (2008). Likelihood approximation by numerical integration on sparse

grids. *journal of Econometrics*, *144*(1), 62–80.

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.

Hornik, K., Stinchcombe, M., and White, H. (1989). Multilayer feedforward networks are universal approximators. *Neural Networks*, *2*, 359-366.

Houde, S., and Myers, E. (2021). Are consumers attentive to local energy costs? evidence from the appliance market. *Journal of Public Economics*, *201*, 104480.

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* (Tech. Rep.). Center for Retirement Research.

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.

Judd, K. L., Maliar, L., and Maliar, S. (2011). Numerically stable and accurate stochastic simulation approaches for solving dynamic economic models. *Quantitative Economics*, *2*(2), 173-210.

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.

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.

LeCun, Y., Bengio, Y., and Hinton, G. (2015). Deep learning. *nature*, *521*, 436-444.

Ma, X., and Zabaras, N. (2009). An adaptive hierarchical sparse grid collocation algorithm for the solution of stochastic differential equations. *Journal of Computational Physics*, *228*(8), 3084 - 3113.

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.

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* (2. ed. ed.). New York, NY: Springer.

Peherstorfer, B., Pflüger, D., and Bungartz, H.-J. (2014). Density estimation with adaptive sparse grids for large data sets. In *Sdm* (p. 443-451).

Pflüger, D., Peherstorfer, B., and Bungartz, H.-J. (2010). Spatially adaptive sparse grids for high-dimensional data-driven problems. *Journal of Complexity*, *26*(5), 508-522.

Pflüger, D. (2010). *Spatially adaptive sparse grids for high-dimensional problems* (Dissertation). Technische Universität München, München.

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 [Computer software manual]. Vienna, Austria.

Rossi, P. E., Allenby, G. M., and McCulloch, R. (2012). *Bayesian statistics and marketing*. John Wiley & Sons.

Ruppert, D., Wand, M. P., and Carroll, R. J. (2003). *Semiparametric regression*. Cambridge University Press.

Sifringer, B., Lurkin, V., and Alahi, A. (2020). Enhancing discrete choice models with representation learning. *Transportation Research Part B: Methodological*, *140*, 236-261.

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.

Smolyak, S. A. (1963). Quadrature and interpolation formulas for tensor products of certain classes of functions. In *Doklady akademii nauk* (Vol. 148, pp. 1042–1045).

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.

Valentin, J., and Pflüger, D. (2016). Hierarchical gradient-based optimization with b-splines on sparse grids. In J. Garcke and D. Pflüger (Eds.), *Sparse grids and applications - stuttgart 2014* (pp. 315–336). Springer International Publishing.

Wand, M. (2000). A comparison of regression spline smoothing procedures. *Computational Statistics*, *15*, 443-462.

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., 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.

Zenger, C. (1991). Parallel algorithms for partial differential equations. In W. Hackbusch (Ed.), *Notes on numerical fluid mechanics* (Vol. 31, pp. 241–251). Vieweg.

Zhao, P., and Yu, B. (2006). On model selection consistency of Lasso. *Journal of Machine learning research*, *7*(Nov), 2541–2563.

Zhou, S., and Shen, X. (2001). Spatially adaptive regression splines and accurate knot selection schemes. *Journal of the American Statistical Association*, *96*(453), 247-259.

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 Versicherung

Ich, Maximilian Osterhaus, versichere an Eides statt, dass die vorliegende Dissertation von mir selbstständig und ohne unzulässige fremde Hilfe unter Beachtung der "Grundsätze zur Sicherung guter wissenschaftlicher Praxis an der Heinrich-Heine-Universität Düsseldorf" erstellt worden ist.

Düsseldorf, 07.11.2021 _____

Maximilian Osterhaus