RUHR
ECONOMIC PAPERS

Florian Heiss
Stephan Hetzenecker
Maximilian Osterhaus

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

#824

Imprint

Ruhr Economic Papers

Published by

RWI – Leibniz-Institut für Wirtschaftsforschung
Hohenzollernstr. 1-3, 45128 Essen, Germany

Ruhr-Universität Bochum (RUB), Department of Economics
Universitätsstr. 150, 44801 Bochum, Germany

Technische Universität Dortmund, Department of Economic and Social Sciences
Vogelpothsweg 87, 44227 Dortmund, Germany

Universität Duisburg-Essen, Department of Economics
Universitätsstr. 12, 45117 Essen, Germany

Editors

Prof. Dr. Thomas K. Bauer
RUB, Department of Economics, Empirical Economics
Phone: +49 (0) 234/3 22 83 41, e-mail: thomas.bauer@rub.de

Prof. Dr. Wolfgang Leininger
Technische Universität Dortmund, Department of Economic and Social Sciences
Economics – Microeconomics
Phone: +49 (0) 231/7 55-3297, e-mail: W.Leininger@tu-dortmund.de

Prof. Dr. Volker Clausen
University of Duisburg-Essen, Department of Economics
International Economics
Phone: +49 (0) 201/1 83-3655, e-mail: vclausen@vwl.uni-due.de

Prof. Dr. Roland Döhrn, Prof. Dr. Manuel Frondel, Prof. Dr. Jochen Kluve
RWI, Phone: +49 (0) 201/81 49-213, e-mail: presse@rwi-essen.de

Editorial Office

Sabine Weiler
RWI, Phone: +49 (0) 201/81 49-213, e-mail: sabine.weiler@rwi-essen.de

Ruhr Economic Papers #824

Responsible Editor: Manuel Frondel

ISSN 1864-4872 (online) – ISBN 978-3-86788-957-5

Ruhr Economic Papers #824

Florian Heiss, Stephan Hetzenecker, and Maximilian Osterhaus

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

Ruhr Graduate School in Economics

UNIVERSITÄT DUISBURG ESSEN

Bibliografische Informationen der Deutschen Nationalbibliothek

The Deutsche Nationalbibliothek lists this publication in the Deutsche Nationalbibliografie; detailed bibliographic data are available on the Internet at **http://dnb.dnb.de**

RWI is funded by the Federal Government and the federal state of North Rhine-Westphalia.

http://dx.doi.org/10.4419/86788957
ISSN 1864-4872 (online)
ISBN 978-3-86788-957-5

Florian Heiss, Stephan Hetzenecker, and Maximilian Osterhaus[1]

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

Abstract

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

JEL-Code: C14, C25, L

Keywords: Random coefficients; mixed logit; nonparametric estimation; elastic net

September 2019

1 Florian Heiss, Heinrich Heine University Düsseldorf; Stephan Hetzenecker, RGS Econ and UDE; Maximilian Osterhaus, Heinrich Heine University Düsseldorf. *– Financial support by the Deutsche Forschungsgemeinschaft (DFG) project 235577387/GRK1974 and the Ruhr Graduate School in Economics is gratefully acknowledged.* *- All correspondence to: Florian Heiss, Heinrich Heine University Düsseldorf, Universitätsstr. 1, 40225 Düsseldorf, Germany, e-mail: florian.heiss@hhu.de*

1 Introduction

Adequately modeling unobserved heterogeneity across agents is a common challenge in many empirical economic studies. A popular approach to address unobserved heterogeneity are random coefficient models, which allow 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 & Train, 2000).

Applications of the estimator indicate, however, that it tends to estimate only few positive weights and that it 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, Yang, & Liu, 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, et al. (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, & Miura, 2015), nuclear material detection (Kump, Bai, sik Chan, Eichinger, & Li, 2012), document classification (El-Arini, Xu, Fox, & 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, & Yang, 2016) and applications to real data (e.g., Nevo, Turner, & Williams, 2016, Illanes & Padi, 2019, Blundell, Gowrisankaran, & Langer, 2018 and Houde & Myers, 2019). Nevo et al. (2016) study the demand for residential broadband and estimate that there are only 53 out of 8,626 potentially heterogeneous consumer types. Illanes and Padi (2019) use the approach to estimate the demand for private pension plans in Chile and assign positive weights to only 194 of 83,251 grid points. Blundell et al. (2018) 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 & Hastie, 2005, and Hastie, Tibshirani, & 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 increases 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 & 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 analyze theoretically, under conditions, that our estimator provides more accurate estimates of the true underlying distribution. For that purpose, we show the selection consistency and derive 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 reveals 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 shows 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 different 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 fixed points if the researcher specifies overlapping splines and step functions. Due to the lower number of required fixed 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, both Train (2008) and Train (2016) estimate the respective model with the EM algorithm, which is sensitive to its starting values and is not guaranteed to converge to a global optimum. Burda et al. (2008) and Rossi et al. (2012) employ a Bayesian hierarchical model to approximate the random coefficients' distribution with a mixture of Normal distributions. Even though the estimator potentially has better finite sample properties, it uses a Markov Chain Monte Carlo technique with a multivariate Dirichlet Process prior on the coefficients, which is computationally more demanding.

The remainder of the paper is organized as follows. Section 2 describes the FKRB estimator and introduces our generalized version. Section 3 derives the condition on the estimators' sign consistency and an error bound on the estimated distribution functions. We present two Monte Carlo experiments in Section 4 that investigate the performance of our generalized estimator in comparison to the FKRB estimator. Section 5 applies the estimators to the *Mode Canada* data set from the R package *mlogit* (Croissant, 2019). 6 concludes.

2 Fixed Grid Estimators

For the introduction of 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 entries are equal to one whenever she observes outcome j for the ith observation, and zero otherwise. The goal is to estimate the unknown distribution of heterogeneous parameters $F_0(\beta)$ in the model

$$P_{i,j}(x) = \int g(x_{i,j}, \beta)\, dF_0(\beta) \tag{1}$$

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) is the multinomial mixed logit model, the state-of-the-art model for demand estimation. In this model, consumer i realizes utility $u_{i,j} = x_{i,j}^T \beta_i + \omega_{i,j}$ from alternative j, given product characteristics $x_{i,j}$ and unobserved consumer-specific preferences β_i. $\omega_{i,j}$ denotes an additive, consumer- and choice-specific error term. Consumer i chooses alternative j of J alternatives (and an outside good with utility

$u_{i,0} = \omega_{i,0}$) if $u_{i,j} > u_{i,l}$ for all $l \neq j$. Under the assumption that $\omega_{i,j}$ follows a type I extreme value distribution, the unconditional choice probabilities, $P_{i,j}(x)$, are of the form

$$P_{i,j}(x) = \int \frac{\exp\left(x_{i,j}^T \beta\right)}{1 + \sum_{l=1}^{J} \exp\left(x_{i,l}^T \beta\right)} dF_0\left(\beta\right). \tag{2}$$

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

2.1 Fixed Grid Estimator by FKRB

FKRB propose a simple and fast mixture approach to estimate the underlying random coefficients' distribution without restrictive assumptions on its shape. The estimator uses a finite grid of fixed 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) into a probability model in which $F_0(\beta)$ enters linearly. FKRB derive the linear probability model in two steps: they transform Equation (1) into a regression model with the random coefficients' distribution as the only unknown term. Adding $y_{i,j}$ to both sides and moving $P_{i,j}$ to the right results in the probability model

$$y_{i,j} = \int g\left(x_{i,j}, \beta\right) dF_0\left(\beta\right) + \left(y_{i,j} - P_{i,j}\left(x\right)\right). \tag{3}$$

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

$$y_{i,j} \approx \sum_{r=1}^{R} g\left(x_{i,j}, \beta_r\right) \theta_r + \left(y_{i,j} - P_{i,j}\left(x\right)\right) \tag{4}$$

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.

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

$$\hat{F}(\beta) = \sum_{r=1}^{R} \hat{\theta}_r \; 1[\beta_r \leq \beta] \tag{5}$$

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 greater than or equal to zero, and sum to one

$$\begin{aligned} \hat{\theta}^{FKRB} &= \arg\min_{\theta} \frac{1}{NJ} \sum_{i=1}^{N} \sum_{j=1}^{J} \left(y_{i,j} - \sum_{r=1}^{R} \theta_r z_{i,j}^r \right)^2 \\ \text{s.t.} \quad \theta_r &\geq 0 \quad \forall r \quad \text{and} \quad \sum_{r=1}^{R} \theta_r = 1. \end{aligned} \tag{6}$$

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

2.2 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 (6) by a constraint on the sum of the squared probability weights. This additional constraint provides a straightforward way to mitigate the estimator's sparse nature. Our generalized estimator is still simple and computationally fast.

2.2.1 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 (6) 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 (4) with $1 - \sum_{r=1}^{R-1} \theta_r$ gives the inequality constrained optimization problem

$$\begin{aligned} \hat{\theta}^{\text{FKRB}} = \arg\min_{\theta} \frac{1}{NJ} \sum_{i=1}^{N} \sum_{j=1}^{J} \left(\tilde{y}_{i,j} - \sum_{r=1}^{R-1} \theta_r \tilde{z}_{i,j}^r \right)^2 \\ \text{s.t.} \quad \theta_r \geq 0 \quad \forall r \quad \text{and} \quad \sum_{r=1}^{R-1} \theta_r \leq 1 \end{aligned} \tag{7}$$

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 (7) is an equivalent form of the optimization problem in Equation (6), the objective functions are minimized by the same vector of probability weights. The only difference in the inequality constrained problem is the estimation of the Rth weight, which is calculated after optimization as $\theta_R = 1 - \sum_{r=1}^{R-1} \theta_r$, and is not explicitly part of the optimization. By the constraints $\theta_r \geq 0 \quad \forall r$ 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 (4),

$$\begin{aligned} \hat{\theta}^{\text{NNL}} = \arg\min_{\theta} \frac{1}{NJ} \sum_{i=1}^{N} \sum_{j=1}^{J} \left(\tilde{y}_{i,j} - \sum_{r=1}^{R-1} \theta_r \tilde{z}_{i,j}^r \right)^2 \\ \text{s.t.} \quad \theta_r \geq 0 \quad \forall r \quad \text{and} \quad \sum_{r=1}^{R-1} \theta_r \leq s, \end{aligned} \tag{8}$$

reveals that the baseline estimator is a special case of NNL with fixed tuning parameter

$s = 1$. 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 (7) exceeds one, the stronger the shrinkage imposed by the constraint, and the larger the number of potential zero weights. 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 remaining to zero (see Zou & Hastie, 2005, and Hastie et al., 2009, for the random behavior of LASSO). Because the correlation is particularly strong in a dense grid among neighboring grid points, the random selection behavior is especially severe for dense random coefficient grids. This property conflicts with the requirement of a sufficiently fine grid for accurate approximations of $F_0(\beta)$.

2.2.2 Elastic Net Estimator

Extending the FKRB estimator's optimization problem formulated in Equation (7) 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 & Kennard, 1970) and transforms the FKRB estimator into the nonnegative elastic net (Wu & Yang, 2014) with fixed constraint on the ℓ_1-penalty. Thus, our adjusted estimator minimizes

$$\begin{aligned} \hat{\theta}^{\text{ENET}} &= \arg\min_{\theta} \frac{1}{NJ} \sum_{i=1}^{N} \sum_{j=1}^{J} \left(\tilde{y}_{i,j} - \sum_{r=1}^{R-1} \theta_r \tilde{z}_{i,j}^r \right)^2 \\ \text{s.t.} \quad \theta_r \geq 0 \quad \forall r \quad \text{and} \quad & \sum_{r=1}^{R-1} \theta_r \leq 1 \quad \text{and} \quad \sum_{r=1}^{R-1} \theta_r^2 \leq t \end{aligned} \tag{9}$$

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.

In addition to the improved selection consistency, 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.

The specification of the tuning parameter allows adjusting the estimator to the level of correlation among grid points. Smaller values of 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 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. For 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 (7), including it as a special case. We recommend choosing the tuning parameter 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. If the model fit is highest for $t \geq 1$, the outcome of our adjusted estimator is the same as that for the estimator by FKRB, while it performs better if the model fit is lowest for some $t < 1$.

Loosely speaking, the improved selection consistency of our generalized estimator leads to more precise estimates of the probability weights. We argue that the FKRB estimator can lead to potentially biased estimates if the linear constraint is binding. In that case, the estimator shrinks the weights at some grid points to zero despite the 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.

The results of the two Monte Carlo studies presented in Section 4 demonstrate that the quadratic constraint reduces both the sparsity and the bias. Moreover, we derive an error bound on the estimated probability weights in Section 3 which, under certain conditions, is tighter for our generalized estimator than for the FKRB estimator if the correlation between grid points is strong.

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 is fundamental to an undistorted estimation of the probability weights. Subsection 3.1 aims to analyze both estimators' ability to select the correct weights. To evaluate the overall approximation accuracy of the estimators presented in Section 2, we derive an error bound for the estimated probability weights in Subsection 3.2.

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 fixed grid $\mathcal{B}_R$. Furthermore, assume that $F^*(\beta)$ converges to $F_0(\beta)$ for R going to infinity. 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. The introduction of $F^*(\beta)$ allows us to study the selection consistency and the distance between $\hat{\theta}$ and θ^*.

The focus of our analysis is on the impact of the correlation among the grid points on the estimators. We show that our generalized estimator is selection consistent under less restrictive conditions on the design matrix.

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 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 draws from Takada, Suzuki, and Fujisawa (2017), who analyze a generalization of the elastic net. We adjust their proof such that it is in line with the probability model in Section 2.

For any $\mathcal{B}_R$, denote the linear probability model corresponding to $F^*(\beta)$ by

$$y_{i,j} = \sum_{r=1}^{R} \theta_r^* z_{i,j}^r + \epsilon_{i,j} \tag{10}$$

where $\epsilon_{i,j}$ is the linear probability error. For our analysis of the selection consistency and for the error bound on the estimated weights, we make the following assumptions on

the linear probability model in Equation (10), 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 (7) 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 (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 (2) and other examples such as the kernel of binary choice and of multinomial choice without logit errors (e.g., see Fox et al., 2016). Assumption 1 (iv) holds by the definition of linear probability models.

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 the true weights θ is equal in sign, we write $\hat{\theta} =_s \theta$.

Our definition only includes $R - 1$ points of the transformed model in Equation (9). That is, we only identify whether the $R - 1$ weights included in Equation (9) have the correct sign but not whether the last weight $\theta_R = 1 - \sum_{r=1}^{R-1} \theta_r$ 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 $\theta^* > 0$, and zero weights otherwise with probability approaching one as the number of observation units N goes to infinity.

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

Let λ denote the fixed LASSO parameter which corresponds to the Lagrange parameter for $s = 1$ in Equation (9) and μ the Lagrange version of the ridge tuning parameter t in Equation (9).

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

Nonnegative Elastic Irrepresentable Condition (NEIC). *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. In addition to the linear dependence of the regressor matrix, the magnitude of the fixed LASSO parameter and the tuning parameter μ is taken into account by the NEIC. For $\mu = 0$, the NEIC reverts to the Nonnegative Irrepresentable Condition (NIC), the corresponding condition for selection consistency through the estimator proposed by FKRB. In contrast to the NEIC, the NIC requires that the inverse of $\tilde{Z}_S^T \tilde{Z}_S$ exists, which is not a necessary condition for the NEIC to hold.

We exploit the special structure of our data by incorporating the fact that all θ are between zero and one, and all elements of $\tilde{Z}$ between minus one and one.

In line with Fox et al. (2016), we allow $R(N)$ to depend on the sample size N. That is, the larger N, the more grid points $R(N)$ can be included into the grid. If $R(N)$ increases, we typically expect the number of positive weights $s(N)$ to increase if the true distribution $F_0(\beta)$ is sufficiently smooth. The next condition restricts the rate at which $s(N)$ and $R(N)$ can increase with N. For convenience, we write s and R instead of $s(N)$ and $R(N)$ in the subsequent analyses.

Rate Condition on Density of Grid (RCDG).

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

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

RCDG requires that $\xi^S_{\min}(\mu) > 0$. Otherwise, the condition can never be satisfied. This is only restrictive for the FKRB estimator and always holds for its generalization as long as $\mu > 0$ since $1/(NJ)\tilde{Z}_S^T\tilde{Z}_S + \mu \mathrm{I}_S$ is positive definite for $\mu > 0$ and only positive semidefinite for $\mu = 0$. The assumption $\xi^S_{\min}(\mu) > 0$ 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 C.2. □

Theorem 1 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. Since $\xi^S_{\min}(\mu) = \xi^S_{\min}(0)+\mu$, the minimal eigenvalue $\xi^S_{\min}(\mu)$ is higher for the elastic net than for the LASSO estimator. Furthermore, 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 NIC does not. Thus, Theorem 1 reveals that our estimator can select the true weights in cases in which the FKRB estimator cannot.

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 3.1 - is the precise estimation of the probability weights. In this section, we derive the error bound for 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 in $[-1,1]^R$ for which the ℓ_1-norm is no greater than 2

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

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 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 $R \times R$ matrix $1/(NJ)\tilde{Z}^T\tilde{Z} + \mu \mathrm{I}_R$ over the set of vectors $\mathcal{H}$ as

$$\xi_{\min}(\mu) := \inf_{v \in \mathcal{B}} \frac{v^T \left[\frac{1}{NJ}\tilde{Z}^T\tilde{Z} + \mu \mathrm{I}_R\right] v}{\|v\|_2^2}.$$

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 3.1, $\xi_{\min}(\mu) > 0$ is always satisfied for our generalized estimator with $\mu > 0$ and $\xi_{\min}(\mu) > 0$ is only restrictive for the FKRB estimator.

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

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

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

Proof. See Appendix C.3. □

Theorem 2 holds with probability approaching one as $\delta \to 0$. Because $\gamma(N, \delta)$ decreases in N, the error bound becomes tighter if the number of observation units increases. This can be seen from the condition $\gamma(N, \delta) \leq k\lambda$ which requires a smaller constant k for a larger N (and fixed λ).

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 typically leads to a wider error bound on the estimated weights (for a fixed 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.

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 becomes particularly important if the design matrix tends to be almost singular, in which case the restricted minimum eigenvalue of the FKRB estimator approaches zero (and the error bound its maximum possible value 2). Also note that the estimation error for the weight θ_R, which is not included in the bound in Theorem (2) and calculated as $\theta_R = 1 - \sum_{r=1}^{R-1} \theta_r$, will approach zero whenever $\left\|\hat{\theta} - \theta^*\right\|_2$ is close to zero.

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* $\sqrt{s}\left\|\theta_S^*\right\|_\infty \xi_{\min}(0) < \sqrt{R-1}\ k\lambda$, *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 C.3. □

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*

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

Proof. See Appendix C.3. □

The bound on the difference between the estimated distribution and the best discrete approximation of $F_0(\beta)$ increases in R and decreases in $\xi_{\min}(\mu)$. Similarly to Theorem 2, the difference in the distributions decreases in N since k may decrease when N increases.

Additionally, 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 $F_0(\beta)$[1] and K refers to the number of random coefficients. This explains the relevance of Theorem 3 since the difference of $F_0(\beta)$ and $F^*(\beta)$ becomes negligibly small as R increases and the estimation error can then be well captured by $|\hat{F}(\beta) - F^*(\beta)|$.

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

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

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).

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 $[-0.5, 4.5] \times [-0.5, 3.5]$, 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 $\theta_s = 1/S$.

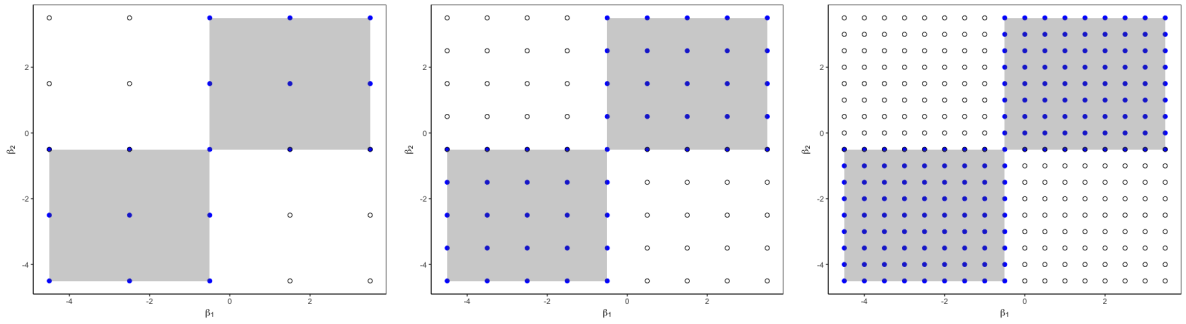
In this setup, the data generating process exactly matches the underlying probability model of the fixed grid estimator. This way, we abstract from any approximation errors that can arise from the sieve space approximation of the true underlying distribution. Therefore, the experiment studies the estimators' 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 $N = \{1000, 10000\}$.

Figure 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.

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

Figure 1: Grid of Monte Carlo Study with Discrete Mass Points



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

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).[2]

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

$$L_1 = \frac{1}{M} \sum_{m=1}^{M} \frac{1}{R} \sum_{r=1}^{R} \left| \theta_r - \hat{\theta}_{r,m} \right| \tag{11}$$

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

[2]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 the μ of OneSe if one wants to avoid cross-validation.

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

$$\text{RMISE} = \sqrt{\frac{1}{M}\sum_{m=1}^{M}\left[\frac{1}{E}\sum_{e=1}^{E}\left(\widehat{F}_m(\beta_e) - F_0(\beta_e)\right)^2\right]}, \tag{12}$$

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

Table 1 summarizes the results of the Monte Carlo experiment. The first three columns report the sample size N, the number of grid points R, and the number of true support points S. The upper part of the table presents the measures on the accuracy of the estimated weights, 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.[3]

Table 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.
1000	25	17	0.067	0.04	0.034	0.035	0.017	0.014	56.05	67.95	0.808
1000	81	49	0.08	0.046	0.038	0.019	0.008	0.007	58.90	70.06	0.819
1000	289	161	0.088	0.057	0.045	0.006	0.004	0.003	54.87	71.20	0.822
10000	25	17	0.042	0.026	0.023	0.02	0.012	0.011	61.15	66.78	0.809
10000	81	49	0.05	0.031	0.027	0.015	0.008	0.007	59.31	69.16	0.818
10000	289	161	0.057	0.037	0.033	0.006	0.004	0.003	61.90	70.50	0.822

			Pos.			% True Pos.			% Sign		
N	R	S	FKRB	MSE	OneSe	FKRB	MSE	OneSe	FKRB	MSE	OneSe
1000	25	17	13.3	20.82	22.36	68.44	94.53	99.71	71.88	77.28	78.14
1000	81	49	15.47	49.58	54.67	27.02	82.12	90.4	53.1	77.65	81.38
1000	289	161	16.24	103.13	123.8	8.62	55.31	66.39	48.27	70.24	75.42
10000	25	17	17.17	19.39	19.73	90.32	98.12	99.53	86.16	87.86	88.46
10000	81	49	23.32	44.84	48.26	42.29	81.07	87.14	61.88	82.24	85.36
10000	289	161	24.88	97.39	105.84	13.53	55.07	59.94	50.76	71.96	74.46

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).

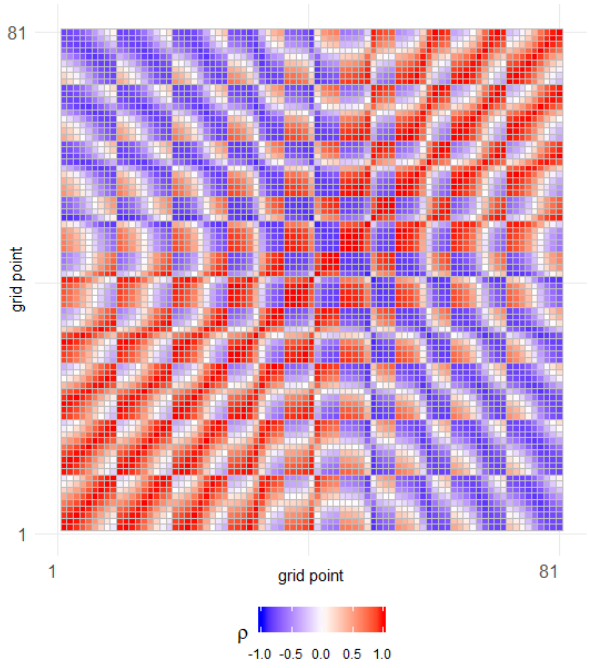
[3]In addition, we also considered the mean and median to summarize the absolute correlation among grid points. We focus on the third quantile since it best illustrates the strong correlation in this setup.

The results show that our generalized estimator outperforms the FKRB estimator for every combination of N and R, in particular when the tuning parameter μ is chosen based on the one-standard-error rule. With respect to the selection consistency, the generalized estimator recovers more true positive and sign consistent probability weights from the data 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 becomes significantly worse.

This is best illustrated by the small number of positive weights, which changes only slightly alongside the increasing complexity. In the extreme case of $R = 289$, the FKRB estimator estimates positive weights at no more than 16/25 of the grid points for $N = 1,000/10,000$ (in comparison to 124/106 for the generalized estimator).

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.

Figure 2: Correlation Matrix for $N = 10,000$ and $R = 81$



The plot of the correlation matrix in Figure 2 and the third quantile of the values of absolute correlation in Table 1 both illustrate that correlation among many grid points is strong.

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.

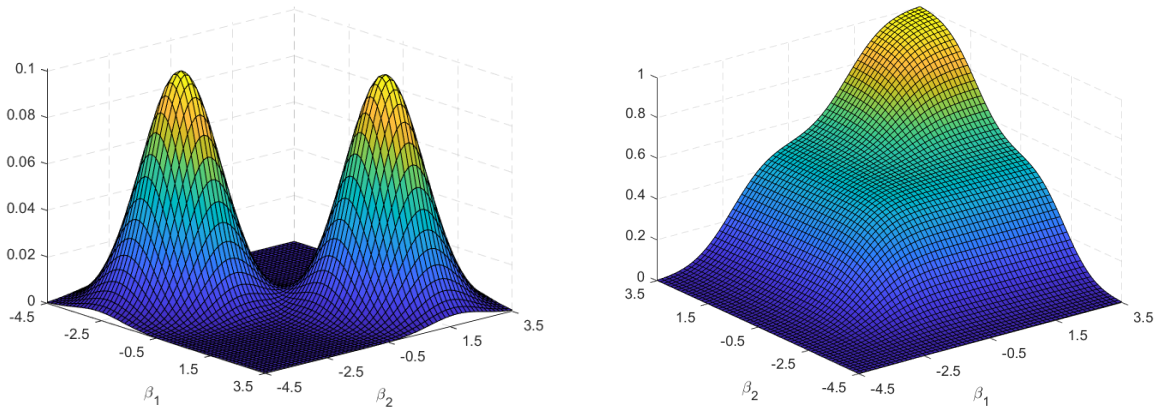
For the estimation, we use a fixed grid with points spread on $[-4.5, 3.5] \times [-4.5, 3.5]$. The fixed grid covers the support of the true distribution with probability close to one (0.993). We keep the correlation among grid points as low as possible and generate the grid points with a Halton sequence. 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 from 1,000 to 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 3 displays the bimodal joint density of the mixture of the two normals, and the right panel the joint distribution function.

Figure 3: True Density and Distribution Function of Mixture of two Normals



(a) PDF

(b) CDF

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

Table 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 predicted choices and the log-likelihood as criteria are reported

in Appendix A.

Table 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.
1000	25	17	0.085	0.072	0.055	9.65	12.94	17.78	21.2	74.01	0.823
1000	50	33	0.09	0.068	0.058	12.57	26.82	32.4	48.09	74	0.82
1000	100	67	0.095	0.07	0.061	13.65	47.09	54.85	58.6	74.37	0.822
1000	250	163	0.102	0.077	0.063	14.2	79.28	103.98	50.5	74.52	0.824
10000	25	17	0.063	0.061	0.057	11.65	12.57	14.89	18.3	73.94	0.823
10000	50	33	0.058	0.051	0.047	17.52	24.94	28.3	48.71	73.94	0.82
10000	100	67	0.06	0.048	0.043	19.87	39.59	46.7	51.63	74.04	0.823
10000	250	163	0.063	0.045	0.04	21.21	76.43	87.92	59.98	74.68	0.824

			% True Pos.			% Sign		
N	R	S	FKRB	MSE	OneSe	FKRB	MSE	OneSe
1000	25	17	49.06	66.35	88.68	60.12	70.48	81.48
1000	50	33	33.39	70.33	84.48	52.93	73.19	80.72
1000	100	67	18.01	63.46	74.1	43.48	70.95	77.44
1000	250	163	7.37	44.38	58.17	38.73	60.96	69.06
10000	25	17	57.94	63.35	77.24	64.2	67.86	77.48
10000	50	33	47.24	68.59	78.05	61.32	74.66	80.42
10000	100	67	26.57	54.72	64.84	48.74	66.73	73.19
10000	250	163	11.39	43.78	50.56	41.17	61.32	65.56

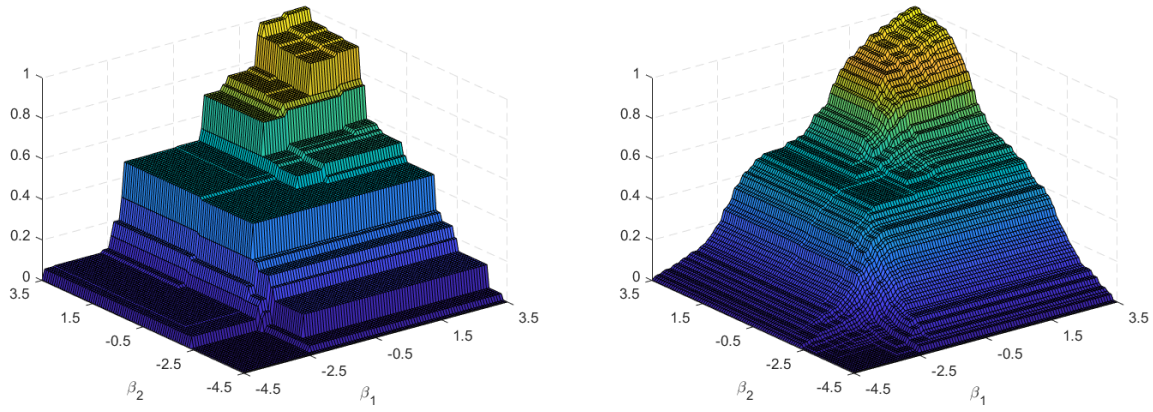
Note: The table reports the average summary statistics over all Monte Carlo replicates for the FKRB estimator (FKRB), and for our generalized estimator with tuning parameter μ from a 10-fold cross-validation and the MSE criterion (MSE) and the one-standard-error rule (OneSe).

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 probabil-

ity 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 4 plots an example of the joint distribution functions estimated with the FKRB estimator (Panel (a)) and our generalized estimator (Panel (b)). Figure 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 $R = 250$.

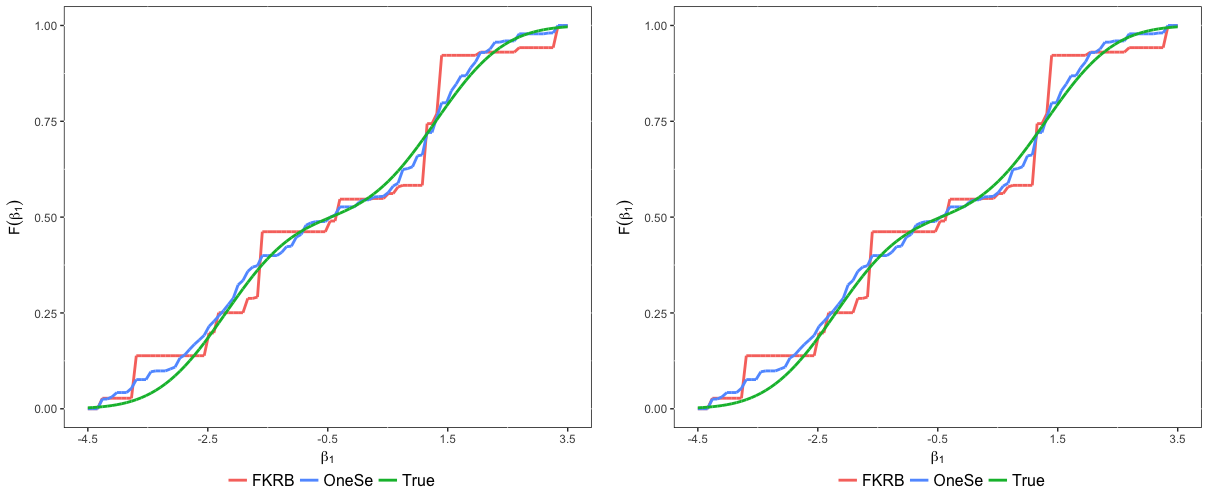
Figure 4: Estimated Joint Distribution Functions for $N = 10,000$ and $R = 250$



(a) FKRB

(b) Generalized with OneSe

Figure 5: True and Estimated Marginal Distribution Functions 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.

5 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 $3,593$ 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 (2019) who suggest a two-step estimator to estimate the model with fixed and random coefficients.[4] In the first step, all coefficients are estimated using a parametric 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 elastic net estimator. Houde and Myers (2019) justify the procedure with the argument that a mixed logit can recover the means of a distribution fairly well despite the incorrect assumptions on the random

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

coefficients' distribution. Thus, the fixed coefficients can be estimated consistently with the parametric approach. They illustrate this property in a Monte Carlo study.

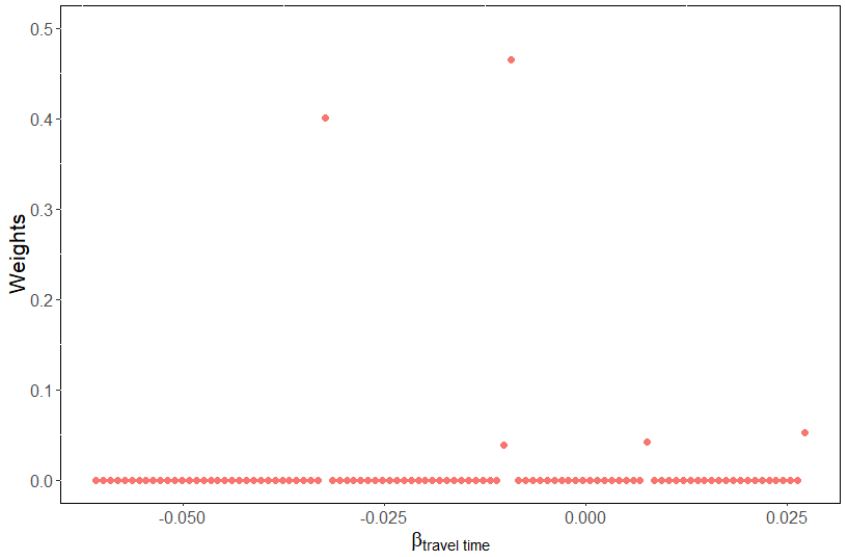
We center the grid of the random coefficient around the mean estimate of the travel coefficient from the first step[5] and add three standard deviations to each side. We estimate the second step with different numbers of grid points. The preferred specification uses $R = 100$ uniformly spread points 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 6 summarizes the mass and the distribution functions estimated with the FKRB and the ridge estimator.

The elastic net estimator results in a smooth mass function whereas the FKRB exhibits the LASSO behavior. The FKRB estimator only selects five out of 100 grid points whereas the elastic net estimator selects 75 grid points.[6] Furthermore, it can easily be seen that the estimated mass function obtained by the elastic net 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 elastic net to 10.1% which is lower than 12.6% for the mixed logit with normal distribution. The weighted mean of the coefficient of travel time for the FKRB estimator is -0.01593 and -0.01631 for the elastic net estimator. This is roughly the same as -0.01682, the mean coefficient obtained from the mixed logit model with normally distributed travel time coefficient.

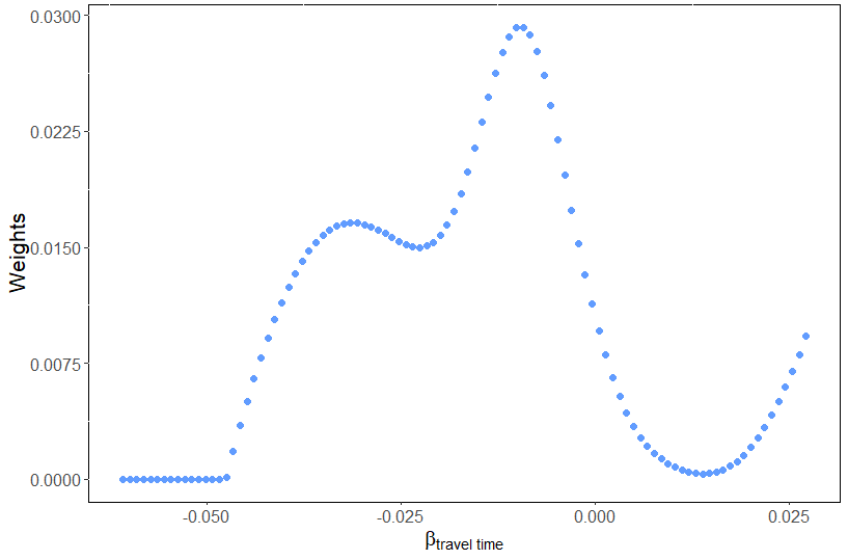
In addition to the estimated distributions, we report the mean (and median) over individuals' own- and cross-travel time elasticities for the FKRB estimator, the elastic net estimator and the semiparametric mixed logit with normal distribution in Appendix A. We also calculate the ratio between elasticities estimated with the FKRB estimator and the semiparametric estimator in comparison to the elasticities estimated with the elastic net estimator. The ratios show that the estimated elasticities are up to 1.8 times larger for the FKRB estimator and up to 4.5 times larger for the semiparametric estimator.

[5]The estimated coefficients of the first stage are provided in Appendix A.
[6]We again define a weight as positive if it is greater than 10^{-3}.

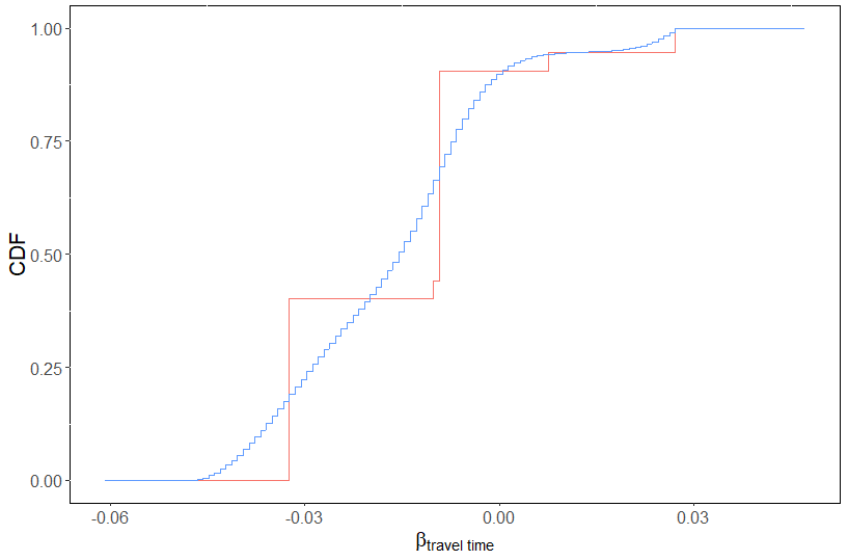
Figure 6: Estimated Distributions of Travel Time in Mode Canada Data with $R = 100$



(a) Mass Function for FKRB



(b) Mass Function for Elastic Net



(c) CDFs for FKRB (red) and Elastic Net (blue)

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 increasing computational speed and simplicity substantially. 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 elastic net estimator to a data set of travel choices made in the Toronto-Montréal corridor confirms the sparsity of the FKRB estimator. In contrast, the elastic net estimator selects substantially more grid points, resulting in a smooth distribution function. This illustrates the fact that the elastic net estimator is able to approximate continuous distribution functions.

References

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.

Blundell, W., Gowrisankaran, G., & Langer, A. (2018). *Escalation of scrutiny: The gains from dynamic enforcement of environmental regulations* (Tech. Rep.). National Bureau of Economic Research.

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

Croissant, Y. (2019). mlogit: Multinomial logit models [Computer software manual]. Retrieved from https://CRAN.R-project.org/package=mlogit (R package version 0.4-1)

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

El-Arini, K., Xu, M., Fox, E. B., & 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.

Fox, J. T., Kim, K., Ryan, S., & Bajari, P. (2011). A simple estimator for the distribution of random coefficients. *Quantitative Economics*, *2*(3), 381–418.

Fox, J. T., Kim, K., & Yang, C. (2016). A simple nonparametric approach to estimating the distribution of random coefficients in structural models. *Journal of Econometrics*, *195*(2), 236–254.

Gentle, J. E. (2007). *Matrix algebra: Theory, computations, and applications in statistics* (1st ed.). Springer Publishing Company, Incorporated.

Hastie, T., Tibshirani, R., & Friedman, J. (2009). *The elements of statistical learning: data mining, inference and prediction* (2nd ed.). Springer.

Hess, S., Bierlaire, M., & 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., & Kennard, R. W. (1970). Ridge regression: Biased estimation for nonorthogonal problems. *Technometrics*, *12*(1), 55–67.

Houde, S., & Myers, E. (2019, April). *Heterogeneous (mis-) perceptions of energy costs: Implications for measurement and policy design* (Working Paper No. 25722). Na-

tional Bureau of Economic Research.

Hu, Z., Follmann, D. A., & Miura, K. (2015). Vaccine design via nonnegative lasso-based variable selection. *Statistics in medicine*, *34*(10), 1791–1798.

Illanes, G., & Padi, M. (2019). *Competition, asymmetric information, and the annuity puzzle: Evidence from a government-run exchange in chile* (Tech. Rep.). Center for Retirement Research.

Jia, J., & Yu, B. (2010). On model selection consistency of the elastic net when p $\gg$ n. *Statistica Sinica*, *20*(2), 595–611.

Koppelman, F. S., & 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., sik Chan, K., Eichinger, B., & 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.

Marwick, K. P., & Koppelman, F. (1990). Proposals for analysis of the market demand for high speed rail in the quebec/ontario corridor. *Submitted to Ontario/Quebec Rapid Train Task Force*.

McFadden, D., & Train, K. (2000). Mixed mnl models for discrete response. *Journal of Applied Econometrics*, *15*(5), 447–470.

Nevo, A., Turner, J. L., & Williams, J. W. (2016). Usage-based pricing and demand for residential broadband. *Econometrica*, *84*(2), 411–443.

R Core Team. (2018). R: A language and environment for statistical computing [Computer software manual]. Vienna, Austria.

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

Slawski, M., & Hein, M. (2013). Non-negative least squares for high-dimensional linear models: Consistency and sparse recovery without regularization. *Electron. J. Statist.*, *7*, 3004–3056.

Takada, M., Suzuki, T., & 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. (2016). Mixed logit with a flexible mixing distribution. *Journal of Choice Modelling*, *19*, 40–53.

Wen, C.-H., & Koppelman, F. S. (2001). The generalized nested logit model. *Transportation Research Part B: Methodological*, *35*(7), 627–641.

Wu, L., & Yang, Y. (2014). Nonnegative elastic net and application in index tracking.

Applied Mathematics and Computation, *227*, 541–552.

Wu, L., Yang, Y., & Liu, H. (2014). Nonnegative-lasso and application in index tracking. *Computational Statistics and Data Analysis*, *70*, 116–126.

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

Zou, H., & 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.

Appendix

A Supplementary Tables

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

			RMISE					L_1					μ				ρ
N	R	S	FKRB	MSE	OneSe	LL	PredOut	FKRB	MSE	OneSe	LL	PredOut	MSE	OneSe	LL	PredOut	3rd Qu.
1000	25	17	0.067	0.04	0.034	0.055	0.045	0.035	0.017	0.014	0.027	0.021	56.05	67.95	13.14	35.03	0.808
1000	81	49	0.08	0.046	0.038	0.064	0.055	0.019	0.008	0.007	0.014	0.011	58.90	70.06	20.66	36.96	0.819
1000	289	161	0.088	0.057	0.045	0.068	0.06	0.006	0.004	0.003	0.005	0.004	54.87	71.20	30.71	35.75	0.822
10000	25	17	0.042	0.026	0.023	0.037	0.032	0.02	0.012	0.011	0.018	0.015	61.15	66.78	13.93	31.02	0.809
10000	81	49	0.05	0.031	0.027	0.044	0.037	0.015	0.008	0.007	0.013	0.011	59.31	69.16	13.85	30.90	0.818
10000	289	161	0.057	0.037	0.033	0.049	0.043	0.006	0.004	0.003	0.005	0.005	61.90	70.50	19.02	30.14	0.822

			Pos.					% True Pos.					% Sign				
N	R	S	FKRB	MSE	OneSe	LL	PredOut	FKRB	MSE	OneSe	LL	PredOut	FKRB	MSE	OneSe	LL	PredOut
1000	25	17	13.3	20.82	22.36	16.23	19.35	68.44	94.53	99.71	80.91	91.15	71.88	77.28	78.14	77.14	78.56
1000	81	49	15.47	49.58	54.67	31.07	40.88	27.02	82.12	90.4	53.58	69.17	53.1	77.65	81.38	65.97	72.72
1000	289	161	16.24	103.13	123.8	70.4	84.15	8.62	55.31	66.39	38.02	45.3	48.27	70.24	75.42	62.3	65.64
10000	25	17	17.17	19.39	19.73	17.81	18.64	90.32	98.12	99.53	92.94	96.35	86.16	87.86	88.46	87.16	88.48
10000	81	49	23.32	44.84	48.26	29.88	37.23	42.29	81.07	87.14	54.5	67.84	61.88	82.24	85.36	68.56	75.61
10000	289	161	24.88	97.39	105.84	53.06	69.47	13.53	55.07	59.94	29.93	39.3	50.76	71.96	74.46	59.28	64.04

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

			RMISE					Pos.					μ				ρ
N	R	S	FKRB	MSE	OneSe	LL	PredOut	FKRB	MSE	OneSe	LL	PredOut	MSE	OneSe	LL	PredOut	3rd Qu.
1000	25	17	0.085	0.072	0.055	0.081	0.066	9.65	12.94	17.78	10.38	14.61	21.2	74.01	2.69	34.44	0.823
1000	50	33	0.09	0.068	0.058	0.081	0.069	12.57	26.82	32.4	17.05	25.52	48.09	74	6.65	34.74	0.82
1000	100	67	0.095	0.07	0.061	0.084	0.075	13.65	47.09	54.85	24.59	36.9	58.6	74.37	11.37	29.57	0.822
1000	250	163	0.102	0.077	0.063	0.09	0.078	14.2	79.28	103.98	38.05	64.86	50.5	74.52	11.94	31.47	0.824
10000	25	17	0.063	0.061	0.057	0.063	0.06	11.65	12.57	14.89	11.78	13.41	18.3	73.94	1.2	29.19	0.823
10000	50	33	0.058	0.051	0.047	0.054	0.051	17.52	24.94	28.3	20.03	24.01	48.71	73.94	7.74	32.09	0.82
10000	100	67	0.06	0.048	0.043	0.054	0.048	19.87	39.59	46.7	27.61	35.56	51.63	74.04	10.91	32.55	0.823
10000	250	163	0.063	0.045	0.04	0.055	0.048	21.21	76.43	87.92	43.45	61.62	59.98	74.68	16.05	36.16	0.824

			% True Pos.					% Sign				
N	R	S	FKRB	MSE	OneSe	LL	PredOut	FKRB	MSE	OneSe	LL	PredOut
1000	25	17	49.06	66.35	88.68	53.09	74.59	60.12	70.48	81.48	62.66	75
1000	50	33	33.39	70.33	84.48	45.65	67.71	52.93	73.19	80.72	60.16	72.34
1000	100	67	18.01	63.46	74.1	33.56	50.03	43.48	70.95	77.44	53.38	63.15
1000	250	163	7.37	44.38	58.17	21.11	36.25	38.73	60.96	69.06	47.1	56.13
10000	25	17	57.94	63.35	77.24	58.68	68.21	64.2	67.86	77.48	64.68	71.12
10000	50	33	47.24	68.59	78.05	54.62	66.05	61.32	74.66	80.42	66.04	73.16
10000	100	67	26.57	54.72	64.84	37.76	49.11	48.74	66.73	73.19	55.98	63.24
10000	250	163	11.39	43.78	50.56	24.5	35.12	41.17	61.32	65.56	49.37	55.94

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

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

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

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

Elasticities estimated with FKRB:

	Car	Air	Train
Car	-0.8992 (-0.8444)	1.3982 (0.6692)	0.1164 (0.129)
Air	0.5895 (0.5943)	-1.2267 (-0.5079)	0.2049 (0.1589)
Train	-0.1622 (0.0346)	0.184 (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)	0.19 (0.1118)	-0.6285 (-0.7691)

Elasticities estimated semiparametrically:

	Car	Air	Train
Car	-1.3362 (-1.2584)	1.366 (0.9975)	0.6699 (0.6846)
Air	0.6194 (0.6093)	-1.3744 (-1.4473)	0.3076 (0.2281)
Train	0.2772 (0.1824)	0.3111 (0.1563)	-1.6449 (-1.7289)

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

Estimated Elasticities of FKRB divided by those of ENet:

	Car	Air	Train
Car	1.0728 (1.0922)	0.9929 (0.9813)	0.7908 (1.2783)
Air	1.1099 (1.1804)	0.975 (0.8905)	1.1605 (1.1864)
Train	1.8291 (0.9611)	0.9685 (1.2098)	1.068 (1.1521)

Semiparametrically estimated Elasticities divided by those of ENet:

	Car	Air	Train
Car	1.5941 (1.6277)	0.9701 (1.4627)	4.5492 (6.7854)
Air	1.1662 (1.2103)	1.0925 (2.5375)	1.7425 (1.7035)
Train	-3.1268 (5.0686)	1.6379 (1.398)	2.6173 (2.2478)

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 A.4 for (1) the FKRB estimator and elastic net estimator and (2) the semiparametric mixed logit with normal distribution and the elastic net estimator.

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:

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

The likelihood criterion given in Train (2008) is

$$LL(\beta^F, \beta^M) = \frac{1}{N} \sum_{i=1}^{N} \log\left(\sum_{r=1}^{R} \theta_r z_{i,y_i}^r\right) = \frac{1}{N} \sum_{i=1}^{N} \log\left(\sum_{r=1}^{R} \theta_r z_{i,y_i}(\beta^F, \beta_r^M)\right). \tag{B.2}$$

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

and being type r is

$$h_{i,r}(\pi) = \frac{\theta_r z_{i,y_i}(\beta^F, \beta_r^M)}{\sum_{r=1}^{R} \theta_r z_{i,y_i}(\beta^F, \beta_r^M)}. \tag{B.3}$$

Based on Equation (B.3) 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

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

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 (B.4) separately for β^F and θ. Since we use our generalized estimator given in Equation (9), we only maximize Equation (B.4) over β^F:

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

Plugging Equation (B.1) into Equation (B.5) gives

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

or equivalently

$$\{\beta^F\}^{t+1} = \arg\max_{\beta^F} \sum_{i=1}^{N}\sum_{j=1}^{J}\sum_{r=1}^{R} y_{i,j} h_{i,r}\left(\pi^t\right) \log\left(\frac{\exp\left(x_{i,j}^F \beta^F + x_{i,j}^M \beta_r^M\right)}{1 + \sum_{l=1}^{J} \exp\left(x_{i,l}^F \beta^F + x_{i,l}^M \beta_r^M\right)}\right). \tag{B.7}$$

This is is the formula of a weighted (standard) logit model where only the coefficients β^F are to be maximized and the coefficients β^M are treated as constants. The weights $h_{i,r}(\pi^t)$, calculated as given in Equation (B.3), 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 (9) 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 (9).

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 (B.7) .

7. Repeat steps 3 and 6 until convergence, using the updated coefficients $\pi_0 = \pi_1$, where $\theta_0 = \theta_1$ is updated in step 4.

8. Use these estimated weights $\widehat{\theta}$ to calculate the estimated distribution

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

C Proofs of Results in Section 3

Below, we provide the proofs of the results presented in Section 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

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

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 $\|A\|_2 = \max_{1\leq i\leq n} \sqrt{\psi_i(A^T A)}$ 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* $\gamma \geq 0$

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

Proof. Notice that

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

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

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

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 (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, its moment generating function satisfies

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

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

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

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

$$\begin{aligned}
\mathbb{P}\left(\left\|\frac{1}{NJ}\tilde{Z}^{T}\epsilon\right\|_{\infty}\geq\gamma\right) &\leq (R-1)J\max_{\substack{1\leq r\leq R-1\\1\leq j\leq J}}\mathbb{P}\left(\left|\frac{1}{N}\sum_{i=1}^{N}\tilde{z}_{i,j}^{r}\epsilon_{i,j}\right|\geq\gamma\right)\\
&\leq 2(R-1)J\exp\left(-\frac{N\gamma^2}{2}\right). \qquad \text{(C.3)}
\end{aligned}$$

□

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.

Lemma 2. *It holds that*

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

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

$$\frac{1}{\sqrt{NJ}} \tilde{Z}_S = ADM^T \tag{C.4}$$

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 (C.4) to rewrite

$$\begin{aligned} \left(\frac{1}{NJ} \tilde{Z}_S^T \tilde{Z}_S + \mu \mathrm{I}_S \right)^{-1} &= \left(MD^T A^T ADM^T + \mu \mathrm{I}_S \right)^{-1} = \left(MD^2 M^T + \mu MM^T \right)^{-1} \\ &= M \left(D^2 + \mu \mathrm{I}_S \right)^{-1} M^T \end{aligned} \tag{C.5}$$

Therefore,

$$\left\| \left(\frac{1}{NJ} \tilde{Z}_S^T \tilde{Z}_S + \mu \mathrm{I}_S \right)^{-1} \right\|_\infty = \left\| M \left(D^2 + \mu \mathrm{I}_S \right)^{-1} M^T \right\|_\infty \leq \sqrt{s} \left\| M \left(D^2 + \mu \mathrm{I}_S \right)^{-1} M^T \right\|_2 \tag{C.6}$$

$$\begin{aligned} &= \sqrt{s} \left\| \left(D^2 + \mu \mathrm{I}_S \right)^{-1} \right\|_2 = \sqrt{s} \max_{i \in S} \sqrt{\psi_i} \\ &= \sqrt{s} \max_{i \in S} \frac{1}{d_{ii}^2 + \mu} = \sqrt{s} \frac{1}{\min_{i \in S} d_{ii}^2 + \mu} = \sqrt{s} \frac{1}{\xi_{\min}^S(\mu)} \end{aligned}$$

where ψ_i denotes the eigenvalues of $\left(\left(D^2 + \mu \mathrm{I}_S \right)^{-1} \right)^T \left(D^2 + \mu \mathrm{I}_S \right)^{-1} = \left(D^2 + \mu \mathrm{I}_S \right)^{-2}$. 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 $1/(NJ)\tilde{Z}_S^T \tilde{Z}_S + \mu \mathrm{I}_S$ are defined as $\xi^S(\mu)$. $\xi_{\min}^S(\mu)$ corresponds to the minimal eigenvalue of the matrix. The first inequality in Equation (C.6) 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 (C.6) holds by the relation of singular values to eigenvalues. □

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

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

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

where

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

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

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

Proof. The Lagrangian of our adjusted estimator that follows from the transformed optimization problem in Equation (9) is

$$L(\theta) := \quad \frac{1}{2NJ} ||\tilde{y} - \tilde{Z}\theta|| + \lambda_n \left(\iota^T \theta - 1 \right) + \frac{1}{2} \mu \, \theta^T \theta - \nu^T \theta \tag{C.7}$$

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 (C.7) 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 (C.7) are satisfied. The KKT conditions are given by

$$-\frac{1}{NJ}\tilde{Z}^T\left(\tilde{y}-\tilde{Z}\hat{\theta}\right)+\lambda\iota+\mu\,\hat{\theta}-\nu=0, \tag{C.8}$$

$$\lambda\left(\iota^T\hat{\theta}-1\right)=0, \tag{C.9}$$

$$\nu_r\,\hat{\theta}_r=0, \tag{C.10}$$

$$\lambda\geq 0,\quad \nu_r\geq 0 \qquad\qquad \forall\quad r=1,\ldots,R-1. \tag{C.11}$$

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 (C.8) require

$$-\frac{1}{NJ}\tilde{Z}_S^T\left(\tilde{y}-\tilde{Z}_S\hat{\theta}_S\right)+\lambda\iota_S+\mu\hat{\theta}_S-\nu_S=0, \tag{C.12}$$

$$-\frac{1}{NJ}\tilde{Z}_{S^C}^T\left(\tilde{y}-\tilde{Z}_S\hat{\theta}_S\right)+\lambda\iota_{S^C}-\nu_{S^C}=0. \tag{C.13}$$

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

$$\nu_r=0 \qquad\qquad \forall\quad r\in S, \tag{C.14}$$

$$\nu_r\geq 0 \qquad\qquad \forall\quad r\notin S. \tag{C.15}$$

It follows from Condition (C.14) that Condition (C.12) simplifies to

$$-\frac{1}{NJ}\tilde{Z}_S^T\left(\tilde{y}-\tilde{Z}_S\hat{\theta}_S\right)+\lambda\iota_S+\mu\hat{\theta}_S=0. \tag{C.16}$$

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

$$-\frac{1}{NJ}\tilde{Z}_S^T\tilde{Z}_S\left(\theta_S^*-\hat{\theta}_S\right)-\frac{1}{NJ}\tilde{Z}_S^T\epsilon+\lambda\iota_S+\mu\hat{\theta}_S=0 \tag{C.17}$$

and

$$-\frac{1}{NJ}\tilde{Z}_{S^C}^T\tilde{Z}_S\left(\theta_S^*-\hat{\theta}_S\right)-\frac{1}{NJ}\tilde{Z}_{S^C}^T\epsilon+\lambda\iota_{S^C}-\nu_{S^C}=0. \tag{C.18}$$

Reformulating Condition (C.17) gives

$$\hat{\theta}_S=\underbrace{\left(\frac{1}{NJ}\tilde{Z}_S^T\tilde{Z}_S+\mu\mathrm{I}_S\right)^{-1}\left(\frac{1}{NJ}\tilde{Z}_S^T\epsilon}_{=:U}+\frac{1}{NJ}\tilde{Z}_S^T\tilde{Z}_S\theta_S^*-\lambda\iota_S\right)>0 \tag{C.19}$$

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

Plugging Equation (C.19) into Equation (C.18) and using Condition (C.15) yields

$$\underbrace{\frac{1}{NJ}\tilde{Z}_{S^C}^T\left[\tilde{Z}_S\left(\frac{1}{NJ}\tilde{Z}_S^T\tilde{Z}_S+\mu\mathrm{I}_S\right)^{-1}\left(\lambda\iota_S+\mu\theta_S^*-\frac{1}{NJ}\tilde{Z}_S^T\epsilon\right)+\epsilon\right]}_{=:V}\leq\lambda\iota_{S^C}. \tag{C.20}$$

U and V are defined in Equation (C.19) and Equation (C.20), 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 (C.20) 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

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

where $\rho:=\min_{i\in S}|g_i|$ with $g_i:=\left[\left(\frac{1}{NJ}\tilde{Z}_S^T\tilde{Z}_S+\mu\mathrm{I}_S\right)^{-1}\left(\frac{1}{NJ}\tilde{Z}_S^T\tilde{Z}_S\theta_S^*-\lambda\iota_S\right)\right]_i$.

Therefore, the event $\mathcal{M}(U)$ implies

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

where g_i, U_i and $\hat{\theta}_{S_i}$ denote the ith element of the respective vectors g, U and $\hat{\theta}_S$. The second last equality holds by definition of g_i and U_i (see Equation (C.19)) 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 (C.19) 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,*

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

Proof. V_j is sub-Gaussian with mean

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

Recall the Nonnegative Elastic Net Irrepresentable Condition (NEIC) is

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

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

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

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

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

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

□

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

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

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

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

since $\eta\lambda > 0$.

Lemma 5. *Suppose Assumption (1) holds. Let $\mathcal{M}^C(U)$ denote the complement of $\mathcal{M}(U)$. Then,*

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

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

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

In the next step Lemma 2 is applied again.

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

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

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 (C.7), it must hold that it minimizes (C.7), 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 (C.7), 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

$$\begin{aligned}&\frac{1}{2NJ}\left\|\tilde{Z}\left(\theta^* - \hat{\theta}\right)\right\|_2^2 + \lambda\left(\iota^T\hat{\theta} - 1\right) + \frac{\mu}{2}\hat{\theta}^T\hat{\theta} \leq \\ &\frac{1}{NJ}\epsilon^T\tilde{Z}\left(\hat{\theta} - \theta^*\right) + \lambda\left(\iota^T\theta^* - 1\right) + \frac{\mu}{2}\theta^{*T}\theta^*. \qquad \text{(C.21)}\end{aligned}$$

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

$$\begin{aligned}
\mathbb{P}\left(\left\| \frac{1}{NJ} \tilde{Z}^T \epsilon \right\|_\infty \geq \gamma \right) &\leq 2(R-1)J \exp\left(-N \left(\sqrt{\frac{2 \log\left(\frac{2(R-1)J}{\delta}\right)}{N}} \right)^2 \Big/ 2 \right) \\
&= 2(R-1)J \exp\left(\log\left(\left(\frac{2(R-1)J}{\delta} \right)^{-1} \right) \right) \\
&= \delta. \qquad \text{(C.22)}
\end{aligned}$$

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 (C.22). 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 (C.21) from above by

$$\frac{1}{2NJ} \left\| \tilde{Z} \left(\theta^* - \hat{\theta} \right) \right\|^2 + \lambda \left(\iota^T \hat{\theta} - 1 \right) + \frac{\mu}{2} \hat{\theta}^T \hat{\theta} \leq \gamma \left\| \hat{\theta} - \theta^* \right\|_1 + \lambda \left(\iota^T \theta^* - 1 \right) + \frac{\mu}{2} \theta^{*T} \theta^*. \quad \text{(C.23)}$$

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 (C.23) 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

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

$$\gamma\left\|\hat{\theta}-\theta^*\right\|_1+\lambda\left(\left\|\theta_S^*\right\|_1-1\right). \tag{C.24}$$

Note that the terms in (C.24) 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 $\sum_{r=1}^R\theta_r\leq 1$: (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 (C.24) simplifies to

$$\frac{1}{2NJ}\left\|\tilde{Z}\left(\theta^*-\hat{\theta}\right)\right\|_2^2+\frac{\mu}{2}\left(\hat{\theta}_S^T\hat{\theta}_S-\theta_S^{*T}\theta_S^*+\hat{\theta}_{S^C}^T\hat{\theta}_{S^C}\right)\leq\gamma\left\|\hat{\theta}-\theta^*\right\|_1. \tag{C.25}$$

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 (C.25) can be equivalently expressed as

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

$$\gamma\left\|\hat{\theta}-\theta^*\right\|_1+\frac{\mu}{2}\left(2\theta_S^{*T}\theta_S^*-2\theta_S^{*T}\hat{\theta}_S\right). \tag{C.26}$$

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

$$\theta_S^{*T}\left(\theta_S^*-\hat{\theta}_S\right)\leq\theta_S^{*T}\left|\hat{\theta}_S-\theta_S^*\right|\leq\left\|\theta_S^*\right\|_\infty\left\|\hat{\theta}_S-\theta_S^*\right\|_1\leq\sqrt{s}\left\|\theta_S^*\right\|_\infty\left\|\hat{\theta}_S-\theta_S^*\right\|_2 \tag{C.27}$$

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

Substituting Condition (C.27) back into the error bound in Equation (C.26) and using

the the fact that $||\hat{\theta} - \theta^*||_1 \leq \sqrt{(R-1)}\ ||\hat{\theta} - \theta^*||_2$, for $\gamma \leq k\lambda$, we can rewrite Equation (C.26) as

$$\frac{1}{2NJ}\left\|\tilde{Z}\left(\theta^* - \hat{\theta}\right)\right\|_2^2 + \frac{\mu}{2}\left\|\hat{\theta} - \theta^*\right\|_2^2 \leq k\lambda\sqrt{(R-1)}\left\|\hat{\theta} - \theta^*\right\|_2 + \mu\sqrt{s}\left\|\theta_S^*\right\|_\infty \left\|\hat{\theta}_S - \theta_S^*\right\|_2. \tag{C.28}$$

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 (C.28) can be summarized as

$$\frac{1}{2}\left(\hat{\theta} - \theta^*\right)^T \left[\frac{1}{NJ}\tilde{Z}^T\tilde{Z} + \mu\mathrm{I}\right]\left(\hat{\theta} - \theta^*\right) \leq \left(k\lambda\sqrt{(R-1)} + \mu\sqrt{s}\left\|\theta_S^*\right\|_\infty\right)\left\|\hat{\theta} - \theta^*\right\|_2. \tag{C.29}$$

Recall that $\xi_{\min}(\mu)$ defines the minimum eigenvalue of the real symmetric matrix $1/(NJ)\tilde{Z}^T\tilde{Z} + \mu\mathrm{I}$ over the set of vectors $\mathcal{H}$ (see Subsection (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 (C.29) 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:

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

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

□

Proof of Corollary 1.

By assumption, it holds that

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

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

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

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

which is equivalent to

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

□

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

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

□