Demand Estimation with High-Dimensional Consumer Demographics

Tianqi Li*

Link to the latest version

Abstract

Random-coefficient multinomial logit models are widely applied to study discrete choices in economics. By assuming random coefficients for each individual, the models can account for unobserved individual heterogeneity and suggest more realistic substitution patterns, compared to standard logit models. In this paper, I find that random coefficients become undetectable (i.e., estimated variances are zero) even if they exist, as many observed individual covariates are incorporated. Having zero estimates of variances not only yields bias in estimating other parameters but also raises the concern of parameters on boundary. To address these issues, I propose l_1-regularized maximum likelihood estimation for simultaneous covariate selection, and develop a debiased machine learning estimator to correct regularization bias while accounting for parameter constraints, such as non-negativity of variance. I derive non-asymptotic probability bounds for the regularized estimator and limiting distributions for the debiased estimator. Finally, I validate the estimators with thorough Monte Carlo simulations, and illustrate the impacts of high-dimensional covariates in an application to soft-drink markets in North Carolina.

Keywords: high dimensions, debiased machine learning, boundary inference, demand estimation

*Department of Economics, University of North Carolina at Chapel Hill. Email: tianqili@ad.unc.edu

1 Introduction

The random-coefficient multinomial logit (RC-Logit) model is a highly flexible model that can approximate any discrete choice model derived from random utility maximization under mild regularity conditions (McFadden and Train, 2000; Train, 2002). Compared to standard logit estimation of choice models, the RC-Logit model relaxes the assumption of the independence of irrelevant alternatives (IIA), allows different coefficients for each individual, and suggests more realistic substitution patterns[1]. As a consequence, the RC-Logit model and its variants have become popular across fields including health economics (Ho, 2006; Hall et al., 2006; Ericson and Starc, 2012; Hole and Kolstad, 2012), transportation economics (Hess et al., 2005; Train and Winston, 2007; Léon and Miguel, 2017) and industrial organization.

In industrial organization literature, one well-known example is the differentiated product demand model developed by Berry et al. (1995, BLP1995 henceforth). In their model, the indirect utility

$$U_{ij} = X_j \otimes \beta_i - \alpha P_j + \xi_j + \varepsilon_{ij} \tag{1.1}$$

of consumer i purchasing product j depends on exogenous product characteristics X_j, price P_j and an unobserved term ξ_j to researchers that is correlated to the price (e.g., the quality of a product). The random coefficient β_i can be decomposed as $\beta_i = \beta^x + L_i\Pi + \nu_i\Sigma$, where L_i is the observed individual demographics, and ν_i captures the remaining unobserved individual heterogeneity, which is assumed to follow a known distribution such as the standard normal distribution. Particularly, given the endogeneity of price

[1]The IIA assumption implies that the ratio of choice probabilities between any two options remains unchanged if an option is added or removed from the choice set. As an example, in the classic red bus-blue bus problem under a standard logit model, suppose that the market shares of trains and red buses are each 50%. Then, introducing blue buses, which are identical to red buses except for color, will decrease the shares of trains and red buses to 33%. The RC-Logit model can address this issue by capturing the unobserved individual preferences on transportation options.

due to ξ_j, instrumental variables are required to estimate α in the generalized method of moments or two-stage least squares.

Numerous studies are based on the framework in BLP1995, for example, Nevo (2000a,b); Berry et al. (2004a); Goolsbee and Petrin (2004); Dubois et al. (2018); Conlon and Gortmaker (2020, 2023). Through this framework, researchers can estimate parameters on preference (i.e., α, β^x, Π and Σ in Eq.(1.1)), calculate price elasticities, and analyze welfare effects (e.g., Gowrisankaran et al., 2015). Notably, BLP1995 allows researchers to estimate demand using only macro-level data such as market shares and demographic distributions. However, incorporating micro-statistics, such as consumers' second choices and the average number of kids in a household that purchases a minivan, may significantly reduce variances of the estimators and tighten substitution patterns (Petrin, 2002; Berry et al., 2004a; Nurski and Verboven, 2016). See Conlon and Gortmaker (2023, Table 1) for a comprehensive summary of papers using micro BLP1995 estimators.

With advancements in collecting information (e.g., in-home scanners and online surveys), researchers may have access to individual-level data consisting of choices and characteristics. For instance, the consumer panel dataset from NielsonIQ records household-level purchases, and its complementary dataset provides demographic variables such as income, gender, race, education as well as occupation of household members. Taking advantages of individual-level data, researchers can directly[2] model individuals' choices without drawing covariates from distributions[3], and therefore, incorporate covariates as many as they need. Including many covariates is not feasible in BLP1995 due to the curse of dimensionality in drawing covariates[4]. This is not an issue given individual-level data, however, selecting potential covariates into models can be challenging when there are hundreds of covariates in a dataset.

[2] "Directly" means that the choice of each individual is observed.

[3] Market shares can be constructed based on individual choices if the data are representative of the population in markets. See discussions in Conlon and Gortmaker (2023) for other sampling strategies in surveys.

[4] If researchers only know the (joint) distributions of covariates, then they need sufficiently large numbers of draws to approximate the distributions well when the number of covariates is large.

Typically, the selection procedure is according to the researchers' expertise, such that the model is expected to be correctly specified or at least include the most important covariates to mitigate omitted variable bias. Nonetheless, it is incorrect to include too many covariates which makes the model high-dimensional. Traditional methods such as simulated maximum likelihood (SML) estimation and simulated method of moments (SMM) are consistent but may be biased, especially when the sample size is small relative to the number of covariates. In this context, the estimated hessian matrix can be singular or ill-conditioned, leading to invalid standard errors. Although adding regularization to the objective function can select variables simultaneously (and potentially address the singularity issue), it will also introduce regularization bias to the estimators and the bias is less understood in RC-Logit models (cf. high-dimensional linear models).

To conduct valid statistical inference, researchers must first debias their estimators. However, inference for RC-Logit models is non-standard due to constraints on parameters related to random coefficients, for instance, standard deviations must be non-negative. Intuitively, since random coefficients capture unobserved heterogeneity, they naturally become degenerate and approach zero when most of the heterogeneity is adequately explained by covariates. When the true parameters lie on (or near) the boundary of the parameter space (e.g., $\theta_0 = 0$ in $\Theta = [0, \infty)$), the estimators are not asymptotically normal in general. Moreover, incorporating many individual covariates can lead to overfitting and misleading zero estimates for the standard deviations, as demonstrated in my simulation studies.

In this paper, I offer solutions to these challenges by integrating the literature of high-dimensional inference and inference on the boundary. For estimation, I recommend using the l_1-regularized maximum likelihood estimation (RMLE) approach to select high-dimensional parameters, which remains effective even in the presence of multi-collinearity and endogeneity. Since the l_1-penalty is non-differentiable at zero, algorithms such as BFGS and Newton-Raphson may fail. To address this, I propose a proximal gradient

descent algorithm that accommodates box constraints. Note that the contraction mapping in BLP1995 can be implemented together with the algorithm but with extra cost. For inference, I build on Li (2024) and develop a constrained debiased machine learning (CDML) estimator, which is constructed from the first-stage RMLE. To prevent potential overfitting in the RMLE, I implement a K-fold cross-fitting procedure. As an M-estimator for low-dimensional target parameters, the CDML estimator solves a Neyman orthogonal score function, subject to constraints such as the non-negativity. I also propose a quasi-likelihood ratio (QLR) test for hypothesis testing whose critical values can be computed through simulations.

The first contribution of this paper is the derivation of non-asymptotic probability bounds for RMLE. To the best of my knowledge, this is the first paper that provides non-asymptotic results for RC-Logit models with high-dimensional covariates. It allows the number of covariates to increase with the sample size, reflecting the finite-sample-large-dimension setting, and accounts for the growing number of alternatives, which is crucial for consistently recovering parameters under endogeneity. Under mild assumptions, I derive probability bounds for the estimation errors in RMLE. Regardless of the boundary issue, these bounds indicates a slower convergence rate than the existing high-dimensional literature due to the increasing number of alternatives.

The second contribution is to supplement the high-dimensional inference literature. I prove the root-n consistency for the CDML estimator and derive its asymptotic distribution, where the score function is nonlinear in parameters. Specifically, the distribution is multivariate Gaussian if the true value of the target parameter is an interior point within its parameter space, while it is a projection of the multivariate Gaussian onto a polytope when the true value is on the boundary of its parameter space. Given the orthogonal structure of the score function, I show that whether the nuisance parameter is on the parameter space or not does not impact the asymptotic distribution.

The third contribution is the novel idea of conducting inference when a high-dimensional

parameter potentially lie on the boundary. By expanding the squared Euclidean norm of the Neyman orthogonal score function at the true parameter, it suffices to study the asymptotic properties of the CDML estimator through a quadratic function, conditioned on the first-stage estimation. Given the smoothness of the soft-max function in RC-Logit models, techniques from the literature of boundary inference can be widely applied, as long as the quality of the first-stage estimation (typically requirement on convergence rates) is good enough. As a consequence, my framework can be extended to scenarios when the true parameter is a sequence converging to the boundary, where a uniformly valid test becomes necessary.

I illustrate the impacts of high dimensions and parameters on the boundary through comprehensive Monte Carlo simulations. In these simulations, I design an indirect utility model that interacts product characteristics with individual characteristics, and estimate the model by SML, RMLE and CDML under varying sample sizes, dimensions and number of products. As expected, these methods perform similarly when the dimension is small. As the dimension increases, MLE suffers from overfitting with biased estimates and large variances. RMLE is also biased but with the smallest variances among three methods. Both MLE and RMLE fail to detect random coefficients when the dimension is large. In contrast, CDML provides valid confidence intervals, especially in the presence of boundary issues. Additionally, I compare the own-price and cross-price elasticities calculated based on the estimates from these methods. When the model includes many covariates, MLE produces a range of estimated elasticities that is nearly twice as large as the oracle model, making it less informative. RMLE has a consistent bias towards zero but the range is much tighter than MLE. CDML offers a more balanced trade-off between bias and variance.

Furthermore, to discover the impact and possibly degenerate random coefficients in real-world data, I estimate the demand for soft-drink markets in North Carolina and calculate the own- and cross-price elasticities based on the estimates. In addition to the consumer panel data and the retailer scanner data provided by NielsonIQ, I incorporate

household-level information from two detailed surveys that investigate households' scientific knowledge, shopping preferences and health conditions. My results indicate that the estimates of both parameters and elasticities are affected by the inclusion of this additional information. In baseline models, all three methods (MLE, RMLE and CDML) yield similar estimates, as expected. However, as the number of parameters increases, the estimates become quite different, emphasizing the importance of careful variable selection and the need for debiasing in high-dimensional settings.

Related literature Numerous literature has explored estimation and inference in the presence of many parameters (e.g., Breiman and Freedman, 1983; Chamberlain and Imbens, 2004; Belloni et al., 2014; Ning and Liu, 2017; Chernozhukov et al., 2018; Cattaneo et al., 2019). There are four papers to which this paper is closely related. Ning and Liu (2017) consider an inference framework for l_1-penalized M-estimation based on decorrelated score test statistics. Their framework can be applied to SML and SMM, however, they only verify their high-level assumptions for the objective functions that are quadratic in parameters, such as generalized linear models. Thus, it is unclear if their assumptions hold for RC-Logit models. Chernozhukov et al. (2018) introduce the Neyman orthogonal score function for general likelihood functions and propose double/debiased machine learning (DML) estimators, which are asymptotically normal. My CDML estimator is an extension to the DML estimator that allows for parameters on the boundary. Horowitz and Nesheim (2021) use penalized maximum likelihood estimation with adaptive LASSO for variable selection and show that their estimator is oracle efficient. Although adaptive LASSO can select non-zero parameters with probability one, inference after imperfect variable selection can be misleading (Belloni et al., 2014). Gillen et al. (2019) propose the BLP-2LASSO method that selects variables for multiple times to mitigate the imperfection, however, their first-step selection ignores random coefficients and they do not have a theoretical proof. None of these papers consider parameters on the boundary, which is

a natural concern with random coefficients.

There is a strand of literature that addresses the issue of parameters on the boundary, especially in the context of random coefficient models. For linear random coefficient models, Hildreth and Houck (1968) suggest a restricted least squares estimator for the variance of random coefficients, while Hsiao (1975) recommends methods such as generalized least squares and maximum likelihood estimation. Breusch and Pagan (1979) develop a Lagrangian multiplier test for random coefficients, building on the work of Aitchison and Silvey (1958). More recently, Ketz (2018, 2019) formally study the inference when (some of) the variances of random coefficients are zero or close to zero in RC-Logit models. However, his results rely on the assumption of fixed dimensions and large sample sizes. Lesellier et al. (2023) test and relax the distributional assumption on random coefficients. My inference framework builds upon the techniques developed by Andrews (1999, 2001), and my QLR test can be extended to achieve uniform validity using the techniques in Fan and Shi (2023).

Note that RC-Logit models are a special case of generalized linear mixed model (GLMM) associated with categorical distribution and logit link function. The literature on GLMMs is extensive, see the review in Tuerlinckx et al. (2006). There are algorithms for solving GLMMs such as expectation-maximization algorithm (Dempster et al., 1977), Laplace approximation (Raudenbush et al., 2000), and penalized quasi-likelihood approximation (Breslow and Clayton, 1993). Groll and Tutz (2014) and Schelldorfer et al. (2014) propose the use of LASSO in GLMMs for variable selection. Hypothesis testing for random effects in GLMMs has also been studied, for example, Self and Liang (1987); Stram and Lee (1994); Verbeke and Molenberghs (2003). However, none of these papers are tailored to RC-Logit models.

Structure of the paper In Section 2, I briefly review simulated maximum likelihood estimation and simulated method of moments as traditional approaches, and then in-

troduce regularized maximum likelihood estimation and its properties in the absence of endogeneity. In Section 3, I adapt the regularized approach for endogenous price in a BLP-style model. In Section 4, I introduce constrained debiased machine learning estimation and prove its asymptotic properties. In Section 5, I conduct Monte Carlo simulations to illustrate the effects of high dimensions and parameters on the boundary. In Section 6, I apply these approaches to estimate the demand in soft-drink markets in North Carolina. Finally, in Section 7, I summarize this paper.

Notations I use the following notations in this paper. For a p-dimensional vector $x \in \mathbb{R}^p$, $d_x := \dim(x)$ is the dimension of x, $\|x\|_1 := \sum_{j=1}^p |x_j|$ is the l_1-norm, $\|x\|_2 := \sqrt{\sum_{j=1}^p x_j^2}$ is the Euclidean norm, and $\|x\|_\infty := \max_{j=1,\dots,p} |x_j|$ is the sup-norm. For a p-by-q matrix $A \in \mathbb{R}^{p\times q}$, $\|A\|_1 := \sum_{i=1}^p \sum_{j=1}^q |A_{ij}|$ and $\|A\|_\infty := \max_{i=1,\dots,p} \max_{j=1,\dots,q} |A_{ij}|$ are induced by vector norms. $\|A\|_F := \sqrt{\sum_{i=1}^p \sum_{j=1}^q |A_{ij}|^2}$ is the Frobenius norm. For both vectors and matrices, the inequalities (i.e., $\leq$ and $\geq$) and notations $O(\cdot), o(), O_P(\cdot)$ and $o_P(\cdot)$ are element-wise. For a function $f(x, y)$, I use $D_x f := \frac{\partial}{\partial x} f$ and $D_x^2 f := \frac{\partial^2}{\partial x \partial x'} f$ to simplify notations without ambiguity. For more details about the derivatives, see Section F.1 in Appendix.

2 High-Dimensional RC-Logit Model

I begin with a brief review of traditional methods for the estimation and inference of random-coefficient logit models, which serves as a foundation for understanding the challenges posed by high dimensions and hence motivates the regularized models. Since the dimension of parameters diverges asymptotically, I establish non-asymptotic probability bounds for the proposed regularized estimator. Moreover, some algorithms such as Newton, BFGS and gradient descent may fail due to the non-smooth penalty on parameters, so I suggest using the proximal gradient descent algorithm instead. For clarity

and intuition, I assume that all data are exogenous in this section, with the discussion on addressing endogeneity deferred to Section 3.

Suppose that individual $i = 1, \ldots, n$ makes decisions among an outside alternative $j = 0$ and J inside alternatives $j = 1, \ldots, J$. The binary outcome variable Y_{ij} is equal to one if individual i derives the highest indirect utility U_{ij} from alternative j, where the indirect utility has a linear specification $U_{ij} = X_{ij}'\Pi + Q_{ij}'\Sigma_i + \varepsilon_{ij}$. Here X_{ij} and Q_{ij} are independent and identically distributed (i.i.d.) covariates. The term Σ_i represents the random coefficient capturing the individual heterogeneity, whose distribution may be either modeled parametrically or estimated non-parametrically. For the purpose of this paper, I assume that $\Sigma_i = \Sigma \odot \nu_i$ where Σ is a vector of standard deviations and the taste $\nu_i \sim i.i.d.\ N(0, I_{d_Q})$ with dimension d_Q is unobservable to researchers. A nature restriction on $\Sigma = (\Sigma_1, \ldots, \Sigma_{d_Q})'$ is that its components $\Sigma_1, \ldots, \Sigma_{d_Q}$ are non-negative. When the idiosyncratic errors ε_{ij}'s follow a Type I extreme value distribution (also known as the Gumbel distribution), the individual choice probability, integrated over ν_i, is given by

$$s_{ij} := Pr\left(Y_{ij} = 1 \mid X_i, Q_i; \Pi, \Sigma\right) = \int \frac{\exp(X_{ij}'\Pi + (Q_{ij} \odot \nu_i)'\Sigma)}{1 + \sum_{k=1}^{J} \exp(X_{ik}'\Pi + (Q_{ik} \odot \nu_i)'\Sigma)} \phi(\nu_i) d\nu_i \qquad (2.1)$$

where $\phi(\cdot)$ is the probability density function of $N(0, I_{d_Q})$, and $U_{i0} = 0$ is normalized for identification purposes.

In this section, the objective is to estimate the parameter $\theta := (\Pi', \Sigma')' \in \Theta \subset \mathbb{R}^{d_X} \times [0, \infty)^{d_Q}$ when Π is high-dimensional but sparse. Since the dimension d_X may increase with the sample size n, it is appropriate to assume a triangular array of data $(Y_{ij,n}, X_{ij,n}, Q_{ij,n}) \sim \mathcal{P}_n$ with the parameter Π_n. To ease notation, I omit the subscript n unless necessary for clarity or when it could lead to ambiguity. Inference for θ, which may lie on the boundary of the parameter space Θ, will be addressed in Section 4.

2.1 Simulated Maximum Likelihood and Method of Moments

When the model is correctly specified, the maximum likelihood estimator (MLE) is asymptotically normal and efficient under mild conditions, for example, the true parameter θ_0 is an interior point in a compact parameter space Θ (Newey and McFadden, 1994). For the RC-Logit model, the log-likelihood function is expressed as $L_{nJ}(\theta) := \sum_{i=1}^{n}\sum_{j=0}^{J} Y_{ij}\log s_{ij}(\theta)$. Although the integral in $s_{ij}(\theta)$ does not have a closed-form solution, it can be numerically approximated using Monte Carlo integration or Gauss-Hermite quadrature (see Appendix D). Let $\hat{s}_{ijB}(\theta) := B^{-1}\sum_{b=1}^{B} s_{ij}(\theta, \nu_{ib})$ denote the approximation of $s_{ij}(\theta)$ using B drawn nodes $\nu_{ib}'s$. Plugging in $\hat{s}_{ijB}$'s, the simulated log-likelihood function is defined as $L_{nJB}(\theta) := \sum_{i=1}^{n}\sum_{j=0}^{J} Y_{ij}\log \hat{s}_{ijB}(\theta)$, and the simulated maximum likelihood (SML) estimator $\hat{\theta}^{SML} := \arg\max_{\theta\in\Theta} L_{nJB}(\theta)$ is obtained by maximizing the function. It is important to note that $\mathbb{E}_\nu[\log \hat{s}_{ijB}(\theta)] \neq \log s_{ij}(\theta)$ even if $\mathbb{E}_\nu[\hat{s}_{ijB}(\theta)] = s_{ij}(\theta)$, suggesting sufficient draws are required to mitigate the approximation error $\log \hat{s}_{ijB}(\theta) - \log s_{ij}(\theta)$. As is shown in Train (2002, pp.255), when the dimension d_θ is fixed, it suffices to have $B \to \infty$ to achieve consistency, and $Bn^{-1/2} \to \infty$ to asymptotically ignore the approximation error as well as achieve the same limiting distribution as in the MLE. Namely, $\sqrt{n}(\hat{\theta}^{SML} - \theta_0) \to_d N(0, (-\mathbb{E}[H_{nJ}(\theta_0)])^{-1})$ as $n \to \infty$, where the hessian matrix $\mathbb{E}[H_{nJ}(\theta_0)] := \mathbb{E}[\frac{d^2}{d\theta d\theta'} L_{nJ}(\theta)]|_{\theta=\theta_0}$ should be negative definite.

Instead of using the MLE score function $\frac{d}{d\theta'} L_{nJ}(\theta)$ as the instrument variables Z_{ij}, the simulated method of moments (SMM) estimator $\hat{\theta}^{SMM} \in \Theta$ is the solution to $n^{-1}\sum_{i=1}^{n}\sum_{j=0}^{J}(Y_{ij} - \hat{s}_{ijB}(\theta))Z_{ij} = 0$ (Train, 2002, pp.276). It sacrifices efficiency but only requires $B \to \infty$ because the choice probability $s_{ij}(\theta)$ enters linearly. It can be shown that, by the central limit theorem,

$$\sqrt{n}(\hat{\theta}^{SMM} - \theta_0) \to_d \left(\mathbb{E}\left[\sum_{j=1}^{J} \frac{d}{d\theta} s_{ij}(\theta_0) Z_{ij}\right]\right)^{-1} N\left(0, Var\left[\sum_{j=1}^{J}(Y_{ij} - s_{ij}(\theta_0))Z_{ij}\right]\right), \quad n, B \to \infty$$

and the asymptotic variance is different from $(-\mathbb{E}[H_{nJ}(\theta_0)])^{-1}$ without ideal instruments.

When the dimension $d_\theta = d_X + d_Q$ is comparable to or greater than the sample size n, there are four concerns on the estimators $\hat{\theta}^{SML}$ and $\hat{\theta}^{SMM}$. First, both estimators may be biased under finite sample sizes, even if $d_\theta < n$ is fixed. Traditional proofs relying on the law of large numbers or the central limit theorem require either $d_\theta < n$ is fixed or $d_\theta = o(n)$. Second, the sample analogs used to construct test statistics or confidence intervals, such as $\frac{d^2}{d\theta d\theta'} L_{nJ}(\hat{\theta}^{SML})$ and $\frac{d}{d\theta'} n^{-1} \sum_i \sum_j s_{ij}(\hat{\theta}^{SMM}) Z_{ij}$, may be ill-conditioned. Moreover, they may not be consistently estimated given the bias from the estimators. Third, both methods are prone to overfitting when the dimension d_X is large, potentially masking the presence of the random coefficients with zero estimates of β^r, while the predicted market shares may still match the data very well. Finally, in practice the data covariates X_{ij} and Q_{ij} may exhibit multi-collinearity that can inflate variances, especially when dealing with dummy variables.

2.2 Regularized Maximum Likelihood Estimation

A natural solution to addressing high-dimensionality and overfitting is to perform variable selection according to their importance, by adding a penalty term into the objective function. Given that the number of drawn nodes B is selected by researchers, I simplify the analysis by assuming that $s_{ij}(\theta)$ is known (or can be approximated arbitrarily well) for all $\theta \in \Theta$. The simplification eliminates the need to account for the approximation errors and allows us to focus on the regularization without loss of generality. I suggest the regularized maximum likelihood estimator (RMLE) as follows:

$$\hat{\theta}^{RMLE} := \arg\min_{\theta \in \Theta} -L_{nJ}(\theta) + P_{\lambda_n}(\theta) \tag{2.2}$$

where $P_{\lambda_n}(\theta) \geq 0$ is a known penalty function with a pre-determined tuning parameter $\lambda_n > 0$. The literature has intensively studied various penalties and criteria for selecting the

tuning parameter. For instance, in ridge regression (Hoerl and Kennard, 1970) the penalty is $P_{\lambda_n} = \lambda_n \|\theta\|_2^2 = \lambda_n \sum_{d=1}^{d_\theta} \theta_d^2$, and in the least absolute shrinkage and selection operator (LASSO, Tibshirani, 1996), it is $P_{\lambda_n} = \lambda_n \|\theta\|_1 = \lambda_n \sum_{d=1}^{d_\theta} |\theta_d|$. The tuning parameter λ_n can be chosen using information criteria or K-fold cross-validations (see Appendix C. There are many variants of LASSOs designed for different contexts. For example, the adaptive LASSO (Zou, 2006; Horowitz and Nesheim, 2021), the generalized LASSO (Tibshirani and Taylor, 2011) and the (sparse) group LASSO (Yuan and Lin, 2006; Meier et al., 2008; Babii et al., 2022).

In this paper, I focus on the LASSO and let $P_{\lambda_n}(\theta) = \lambda_n \|\theta\|_1$, which has demonstrated its strength in consistent estimation (e.g., Bickel et al., 2009) and variable selection (e.g., **?**). Note that $\hat{\theta}^{RMLE}$ in Eq.(2.2) is a regularized M-estimator (Negahban et al., 2012). The other *decomposable* regularizers such as the weighted LASSO and the group LASSO should also work effectively. Consistent with much of the LASSO literature, I allow the dimension $d_\theta \to \infty$ sub-exponentially fast as $n \to \infty$, but need to assume that the true parameters are sparse. Specifically, let $S_n := \{d = 1, \ldots, d_\theta \mid \theta_{0,d} \neq 0\}$, and its cardinality $\mathbf{s}_n := |S_n|$ is known as the *exact sparsity* of θ_0. I assume that $\mathbf{s}_n$ is much smaller than n. The following Assumption 1-3 suffice to derive an probability bound on $\|\hat{\theta}^{RMLE} - \theta_0\|_1$ and $\|\hat{\theta}^{RMLE} - \theta_0\|_2$.

Assumption 1 (Score Condition)**.** *Assume that the true parameter* $\theta_0 = (\Pi_0', \Sigma_0')' \in \Theta \subset \mathbb{R}^{d_X} \times [0, \infty)^{d_Q}$ *is the unique maximizer of* $\mathbb{E}\left[\sum_{j=1}^J Y_{ij} \log s_{ij}(\theta_0)\right]$ *such that* $\frac{d}{d\theta}\mathbb{E}[L_{nJ}(\theta_0)] = \mathbb{E}\left[\frac{d}{d\theta}\sum_{j=1}^J Y_{ij} \log s_{ij}(\theta_0)\right] = 0$.

Assumption 1 posits the interchangeability of the derivative $\frac{d}{d\theta}$ and the expectation $\mathbb{E}[\cdot]$ over the data. Additionally, it assumes that the true parameter $\theta_0 \in \Theta$ can be identified through the population score function. These conditions are standard and mild in the context of RC-Logit models. Importantly, the assumption does not exclude the possibility that some true parameters may lie on the boundary of Θ.

Assumption 2 (Local Convexity). *Suppose that $-L_{nJ}(\theta)$ is locally convex in a neighborhood of $\theta_0 \in \Theta$. In addition, with probability at least $1 - a_n$ for some constants $a_n \to 0$, the first-order approximation error*

$$-L_{nJ}(\theta_0 + \Delta) + L_{nJ}(\theta_0) + \left(\frac{d}{d\theta} L_{nJ}(\theta_0)\right)' \Delta \geq n\kappa_L \|\Delta\|_2^2 \quad \textit{for all} \quad \Delta \in \mathbb{C}$$

where $\mathbb{C} := \{\Delta \in \mathbb{R}^{d_\theta} \mid \sum_{j \in S_n^c} |\Delta_j| \leq 3 \sum_{j \in S_n} |\Delta_j|\}$ is a convex cone and $\kappa_L > 0$ is a universal constant.

Assumption 2, known as the *restricted strong convexity* condition, assumes sufficient curvature in the neighborhood of θ_0, particularly in the convex cone $\mathbb{C}$ where the lasso estimation error resides (Lemma 1 in Negahban et al., 2012). Intuitively, a larger curvature κ_L implies a steeper gradient around θ_0, and the estimation error should decrease in κ_L. It can be regarded as the non-linear version of the restricted eigenvalue condition in the LASSO literature (Bickel et al., 2009). In general, $-L_{nJ}(\theta)$ is neither globally concave nor convex due to the integral and the soft-max function. In fact, it can be even completely flat in certain directions at $\theta_0 \in \Theta$ when $d_\theta > n$ (Figure 3 in Negahban et al., 2012). Since the neighborhood is unknown, in practice, researchers often try multiple starting points or restrict the parameter space to ensure the numerical convergence of algorithms.

Assumption 3 (Bounded Data and Shares). *The data covariates $Y_i = (Y_{i0}, \ldots, Y_{iJ})'$, $X_i = (X_{i1}', \ldots, X_{iJ}')'$ and $Q_i = (Q_{i1}', \ldots, Q_{iJ}')'$ are i.i.d. random vectors across $i = 1, \ldots, n$. In addition, assume that $X_i \in [-C_{data}, C_{data}]^{d_X}$, $Q_i \in [-C_{data}, C_{data}]^{d_Q}$ and $\min_{j=0,\ldots,J} s_{ij}(\theta_0) \geq C_s J^{-1} > 0$ for some finite and universal constants $C_{data}, C_s > 0$.*

To derive the probability bounds for $\hat{\theta}^{RMLE} - \theta_0$, another sufficient condition is that the score vector concentrates in the sup-norm with high probability, formally, $\|n^{-1} \frac{d}{d\theta} L_{nJ}(\theta_0)\|_\infty \leq \rho_n$ with probability approaching one. Notice that $\frac{d}{d\theta} \log s_{ij}(\theta_0) = s_{ij}^{-1}(\theta_0) \frac{d}{d\theta} s_{ij}(\theta_0)$, and the choice probability $s_{ij}(\theta_0)$ can be arbitrarily close to zero if the number of alternatives J is large enough, suggesting that the derivative is unbounded. When there is no endogeneity,

a small-J-and-large-n setting may suffice as θ_0 can be directly[5] estimated. However, in the presence of endogeneity, a large J is necessary as will be shown in Section 3. By Lemma 1, Assumption 3 is a sufficient condition to derive the rate ρ_n. I assume that all choice probabilities converge to zero at the same rate $O(J^{-1})$ as $J \to \infty$ (cf., Berry et al., 2004b).

Lemma 1. *Given Assumption 1, for any* $c > 4C_s^{-1}C_{data}$,

$$\left\| \frac{1}{n} \frac{d}{d\theta} L_{nJ}(\theta_0) \right\|_\infty \leq c\rho_n$$

with probability greater than $1-b_n = 1-2\exp\left((1-\frac{c^2C_s^2}{8C_{data}^2})\log d_\theta\right)$, *and the rate* $\rho_n = J\sqrt{n^{-1}\log d_\theta}$. *This implies* $\|n^{-1}\frac{d}{d\theta}L_{nJ}(\theta_0)\|_\infty = O_P(J\sqrt{n^{-1}\log d_\theta})$.

The proof of Lemma 1 is derived by leveraging the bound $\|\frac{d}{d\theta}s_{ij}(\theta_0)\|_\infty \leq 2C_{data}$ and then applying the McDiarmid's inequality. However, since both C_{data} and C_s^{-1} can be large, the sup-norm of the score vector may also become large unless the sample size n is sufficiently large. The following Theorem 1 shows that the rate $\rho_n = J\sqrt{n^{-1}\log d_\theta}$ also serves as the rate for the estimation errors.

Theorem 1 (l_1/l_2-Error Bounds)**.** *Suppose that Assumption 1-3 hold. If* $\lambda_n \geq 2n\rho_n$, *then with probability at least* $1 - a_n - b_n$,

$$\|\hat{\theta}^{RMLE} - \theta_0\|_2 \leq \frac{3\sqrt{\mathbf{s}_n}\lambda_n}{n\kappa_L} \quad \textit{and} \quad \|\hat{\theta}^{RMLE} - \theta_0\|_1 \leq \frac{12\mathbf{s}_n\lambda_n}{n\kappa_L}$$

where $\mathbf{s}_n$ *is the number of non-zero elements in the vector* θ_0.

The proof of Theorem 1 refers to the Corollary 1 in Negahban et al. (2012) but adapts to a random design. When the number of alternatives $J = O(1)$ is stochastically bounded, the rate $J\sqrt{\mathbf{s}_n n^{-1}\log d_\theta}$ from Lemma 1 aligns with the common rate $\sqrt{\mathbf{s}_n n^{-1}\log d_\theta}$ for linear models in the LASSO literature. The rate suggests that the dimension d_θ can

[5]What I meant directly here is that we can estimate θ_0 simultaneously through MLE without using contraction mapping or IV regressions.

grow at most sub-exponentially with the sample size, i.e., $O(\exp(n^r))$ for some constant $r \in (0,1)$. When $J \to \infty$, the rate slows down due to the diverging term $s_{ij}^{-1}(\theta_0) = O(J)$ in the score vector. Although the probability bounds in the theorem are not tight, the regularized estimator $\hat{\theta}^{RMLE}$ is expected to have a non-parametric rate of convergence when the dimension $d_\theta \to \infty$ is too high or the number of alternatives $J \to \infty$ is too large. As a consequence, inference based on $\sqrt{n}(\hat{\theta}^{RMLE} - \theta_0)$ can be improper without further adjustments (Armstrong et al., 2023).

2.3 Algorithm

Because of the integration over the random coefficients, the RC-Logit models are intensive and unstable in computation, which limits their application to a small number of covariates. Although the integral in Eq.(2.1) has no closed-form solution, its approximation has been developed in the literature, trading off accuracy and computational efficiency. For instance, researchers can approximate the integral using Laplace approximation, Gauss-Hermite quadrature, or quasi Monte Carlo integration with low-discrepancy sequences. Specifically, the Gauss-Hermite quadrature is precise and efficient in computation, yet it suffers from the curse of dimensionality of the random coefficients. Conversely, the Monte Carlo integration is widely implemented when there are many potentially related random coefficients, but it requires more simulation draws to achieve sufficient accuracy. Table D.1 summarizes these methods, and a more detailed review can be found in Tuerlinckx et al. (2006) and Conlon and Gortmaker (2020).

Orthogonal to the numerical integration, another challenge arises from the non-differentiability of the l_1-penalty $P_{\lambda_n}(\theta) = \sum_{d=1}^{d_\theta} \lambda_n |\theta_d|$ at the origin. That is, $\partial|\theta|/\partial\theta = 1$ if $x > 0$ and $= -1$ if $x < 0$, but it is undefined at $x = 0$. This is fine when the true value θ_0 is strictly away from 0. However, such non-zero assumption contradicts the sparsity condition in high-dimensional settings and fails when the variances lie on the boundary. In my numerical experiments, popular algorithms such as BFGS and (conjugate) gradient

descent still tend to converge when the dimension d_θ is small. Nevertheless, their minimizers often contain tiny non-zeros ($\leq 10^{-4}$) instead of exact zeros. When d_θ is large, most algorithms are prone to get trapped in local minima, stopping too early or oscillating between points.

In Algorithm 1, I propose the proximal gradient descent method to solve $\hat{\theta}^{RMLE} \in \Theta$ with box constraints $\Theta = \otimes_{d=1}^{d_\theta}[\theta_d^{lb}, \theta_d^{ub}]$ and numerical integration $\hat{s}_{ijB}(\theta) = B^{-1}\sum_{b=1}^{B} s_{ij}(\theta, \nu_{ib})$.

Algorithm 1 Proximal Gradient Descent Algorithm for RC-Logit

1. Choose an initial $\theta^{(0)} = (\theta_1^{(0)}, \ldots, \theta_{d_\theta}^{(0)})' \in \Theta$, an initial step-size $t_{\text{init}} > 0$, a shrinking rate $\rho \in (0,1)$, and a penalty $\lambda_n \geq 0$;
2. For $h = 1, 2, \ldots,$ start with $t = t_{\text{init}}$
   (a) Calculate the proximal mapping $\theta^{(h)} := Prox^{box}(\theta^{(h-1)}, \lambda_n, t)$, where the d-th coordinate is defined as

$$\left[Prox^{box}(\theta, \lambda_n, t)\right]_d := \begin{cases} \theta_d^{ub} & \text{if } \varphi(\theta, d, t) \in [\theta_{ub}^k, \infty) \\ \varphi(\theta, d, t) - t\lambda_n & \text{if } \varphi(\theta, d, t) \in [t\lambda_n, \theta_{ub}^k) \\ 0 & \text{if } \varphi(\theta, d, t) \in [-t\lambda_n, t\lambda_n) \\ \varphi(\theta, d, t) + t\lambda_n & \text{if } \varphi(\theta, d, t) \in [\theta_d^{lb}, -t\lambda_n) \\ \theta_d^{lb} & \text{otherwise} \end{cases} \tag{2.3}$$

   and $\varphi(\theta, d, t) := \theta_d + t\frac{\partial}{\partial \theta_d} L_{nJB}(\theta)$;
   (b) Verify the criterion of line search

$$\begin{aligned} -L_{nJB}\left(\theta^{(h)}\right) \leq -L_{nJB}\left(\theta^{(h-1)}\right) - \frac{d}{d\theta} L_{nJB}(\theta^{(h-1)})(\theta^{(h)} - \theta^{(h-1)}) \\ + \frac{1}{2t}\|\theta^{(h)} - \theta^{(h-1)}\|^2 \end{aligned} \tag{2.4}$$

   (c) If (b) fails, shrink $t \leftarrow \rho t$ and go back to (a);
3. Repeat Step 2 until convergence, and $\hat{\theta}^{RMLE}$ is the final $\theta^{(h)}$.

Since the unpenalized objective function $L_{nJB}(\theta)$ is smooth, by the Taylor's expansion,

$$L_{nJB}(\theta) \approx L_{nJB}(\theta_0) + \frac{d}{d\theta} L_{nJB}(\theta_0)(\theta - \theta_0) - \frac{1}{2t}\|\theta - \theta_0\|_2^2 \tag{2.5}$$

for some $t \geq 0$. Some algebra shows that $\hat{\theta}^{RMLE}$ can be approximated[6] by the solution to $\min_{\theta \in \Theta} \frac{1}{2t}\|\theta_0 + t\frac{d}{d\theta} L_{nJB}(\theta_0) - \theta\|_2^2 + \lambda_n \|\theta\|_1$, whose sub-gradient optimality condition implies the proximal mapping $Prox^{box}(\theta_0, \lambda_n, t)$ in Eq.(2.3). As a type of *soft-thresholding operators*, the proximal mapping truncates large values and forces smaller ones to be zero, depending on the penalty size λ_n, the step size t, and the imposed box constraints. While the gradient $-\frac{d}{d\theta} L_{nJB}(\theta)$ indicates the direction of the steepest descent, selecting an improper t can still lead to slow convergence or even divergence. In the algorithm, t is chosen as the largest t^* such that $-L_{nJB}(\cdot)$ decreases after the update $\theta \leftarrow Prox^{box}(\theta, \lambda_n, t)$, through a procedure known as *backtracking line search*. Lemma 2 shows that Algorithm 1 can achieve a convergence rate of $O(h^{-1})$, which is typical for gradient descent algorithms.

Lemma 2. *Let $\{\theta^{(h)} : h = 0, 1, \dots\}$ be a sequence of updates in Algorithm 1 such that $L_{nJB}(\cdot)$ is concave at $\theta^{(h)}$, and θ^* be the unique minimum of $-L_{nJB}(\theta) + \lambda_n\|\theta\|_1$. Then,*

$$L_{nJB}(\theta^{(h)}) + \lambda_n\|\theta^{(h)}\|_1 - L_{nJB}(\theta^*) - \lambda_n\|\theta^*\|_1 \leq \frac{J\sqrt{n \log d_\theta}\|\theta^{(0)} - \theta^*\|_2^2}{\rho h}$$

where the right-hand side goes to zero as $h \to \infty$.

The proof of Lemma 2 follows a standard approach by verifying the Lipschitz condition for the sup-norm $\|\frac{d}{d\theta} L_{nJB}(\theta)\|_\infty$ for every $\theta \in \Theta$ and finite n. The result is intuitive: the optimization becomes more challenging as the number of parameters increases and the approximation in Eq.(2.5) is worsen. The concavity assumption can be relaxed, for example, using the techniques in Li and Lin (2015). As with many l_1-regularized problems, the algorithm can be improved using accelerated proximal gradient (e.g., Beck and Teboulle,

[6]For some smooth function $g(\theta)$, let G be its gradient at θ_0. The second-order Taylor's expansion at θ_0 is $g(\theta_0) + G'(\theta - \theta_0) + \frac{1}{2t}(\theta - \theta_0)'(\theta - \theta_0)$ by assuming the hessian is $\frac{1}{t}I$. Clearly, the first term is a constant. Now we expand $\frac{1}{2t}\|\theta_0 - tG - \theta\|^2$, which is $G'(\theta - \theta_0) + \frac{1}{2t}(\theta - \theta_0)'(\theta - \theta_0) + \frac{t}{2}G'G$, and the last term is also a constant.

2009) and/or (block) coordinate descent techniques (e.g., Friedman et al., 2007; Beck and Tetruashvili, 2013), achieving faster convergence rates or having excellent performance in practice (e.g., *glmnet* package in R). However, we do not apply these methods here for two reasons. First, compared to the accelerated methods, the proximal gradient descent is more stable with simulation errors and performs adequately in my context. Second, while the gradient has an explicit form, it is still costly to calculate, especially it must be updated for each coordinate in coordinate gradient descent methods. Similarly, I do not consider proximal newton methods. The complexity of calculating a full gradient vector and a full hessian matrix is $O(d_\theta n J^2 S)$ and $O(d_\theta^2 n J^3 S)$, respectively. In practice, it is strongly recommended to provide the algorithm solver with the analytical gradient (and hessian), which can be found in Appendix F.1.

Finally, it is worth noting that the RC-Logit model in Eq.(2.1) can be viewed as a special case of generalized linear mixed models (GLMMs) with categorical outcomes and Gaussian random coefficients. In addition to the aforementioned techniques, the penalized[7] quasi-likelihood method (PQL, Breslow and Clayton, 1993) is computationally efficient and widely implemented in GLMMs. This method takes advantages of the concavity of the log-likelihood function in the exponential family. As high-dimensional extensions of PQL, Groll and Tutz (2014) and Schelldorfer et al. (2014) suggest using LASSO to select variables. However, PQL relies on the Laplace approximation and has been found to be biased when the variance is large or the mean is small (Bolker et al., 2009). It is still unclear whether PQL can approximate the soft-max function nicely, which I leave for future research.

[7]It is called "penalized" because there is a quadratic term $b'D^{-1}b$ of the random coefficients b (but not the variance covariance matrix D) in the approximated likelihood function, which solves b as an intermediate parameter and then estimate its variance.

3 Modeling Endogeneity

In this section, I consider a specification of RC-Logit model which is of interest in the literature of industrial organizations. Extending the single-market model in Section 2, I now assume that there are $t = 1, \ldots, T$ markets. In each market t, the individual $i \in \mathcal{I}_t$ chooses the alternative $j \in \mathcal{J}_t$ if the indirect utility $U_{ijt} > U_{ikt}$ for $k \in \mathcal{J}_t$, where

$$
\begin{aligned}
U_{ijt} &= \delta_{jt} + \mu_{ijt} + \varepsilon_{ijt}, \quad \varepsilon_{ijt} \sim i.i.d. \text{ Type I EV} \\
\delta_{jt} &= \bar{\beta} + X_{jt}'\beta^x + P_{jt}\alpha + \xi_{jt} \\
\mu_{ijt} &= (X_{jt} \otimes L_i)'\Pi + (X_{jt} \odot \nu_i)'\Sigma, \quad \nu_i \sim i.i.d.\ N(0, I_{d_Q})
\end{aligned}
$$

The term δ_{jt} represents the choice-specific utility from alternative j in market t, which is a linear function of observable characteristics X_{jt}, price P_{jt} and unobserved quality ξ_{jt} (i.e., unobservable to researchers). I assume that X_{jt} is independent of ξ_{jt} while P_{jt} and ξ_{jt} are correlated. The term μ_{ijt} captures the individual-level utility, which depends on observable individual characteristics L_i and unobservable individual tastes ν_i. For the identification purpose, let δ_{0t} and μ_{i0t} be equal to zero so $U_{i0t} = \varepsilon_{i0t}$. The individual choice probability can be derived as

$$
\begin{aligned}
s_{ijt}(\delta_t, \Pi, \Sigma) &:= Pr\left(Y_{ijt} = 1 \mid X_t, L_i; \delta_t, \Pi, \Sigma\right) \\
&= \int \frac{\exp\left(\delta_{jt} + (X_{jt} \otimes L_i)'\Pi + (X_{jt} \odot \nu_i)'\Sigma\right)}{1 + \sum_{k \in \mathcal{J}_t} \exp\left(\delta_{kt} + (X_{kt} \otimes L_i)'\Pi + (X_{kt} \odot \nu_i)'\Sigma\right)} \phi(\nu_i) d\nu_i
\end{aligned}
$$

where $\delta_t = (\delta_{1t}, \ldots, \delta_{J_t t})'$ and $X_t = (X_{1t}', \ldots, X_{J_t t}')'$.

The goal is to consistently estimate $\theta := (\bar{\beta}, \beta^{x'}, \alpha; \Pi', \Sigma')'$, where Π is a high-dimensional parameter. As is discussed in Section 2, the challenges associated with high-dimensional Π apply here as well, such as the difficulty in detecting Σ due to overfitting. To illustrate this, first consider an experimental setting that X_{jt} and P_{jt} are freely manipulated by the researcher, and there is no endogeneity $\xi_{jt} = 0$. In this case, the identification of

α is straightforward: the researcher can vary the price and compare the market shares before and after the change. β^x can be similarly identified since X_{jt} is common to all individuals in market t, while the identifying power is impaired by the high dimensions. Ideally, Π can be identified by comparing the decisions made by different groups of individuals (grouped by their demographics L_i) when they face varying characteristics X_{jt}. The standard deviation Σ can be identified if the researcher can observe different choices made by individuals with the same demographics L_i. But when the dimension of L_i is large and the number of individuals n_t is relatively small, it becomes difficult to group individuals with similar $L_i's$. When there exists endogeneity such that $\xi_{jt} \neq 0$, P_{jt} and ξ_{jt} change together so α is not identifiable without instrument variables. However, δ_{jt} can be uniquely determined by matching the market shares and the predicted shares for any given $(\Pi', \Sigma')'$ according to BLP1995. In this case, the overfitting in Π and Σ may introduce bias in recovering δ_{jt}, and hence bias in estimating $\bar{\beta}$, β^x and α.

When individual-level data are available, the δ_{jt}'s can be estimated as parameters using SML or SMM (Goolsbee and Petrin, 2004; Train and Winston, 2007), such that the predicted market share $\hat{s}_{jt} = |\mathcal{I}_t|^{-1} \sum_{i \in \mathcal{I}_t} s_{ijt}(\hat{\delta}, \hat{\Pi}, \hat{\Sigma})$ matches the observed market share s_{jt}^{data} for each alternative j and market t. If the total number of products $J := \sum_{t=1}^{T} |\mathcal{J}_t|$ across T markets is small relatively to the total number of individuals $N := \sum_{t=1}^{T} |\mathcal{I}_t|$, that is, $J = o(N)$, then we can simultaneously estimate δ, Π and Σ using RMLE as follows:

$$
\begin{aligned}
(\hat{\delta}^{RMLE}, \hat{\Pi}^{RMLE}, \hat{\Sigma}^{RMLE}) &:= \underset{\Sigma \geq 0, \Pi, \delta}{\arg\min} -L_{NJT}(\theta) + P_{\lambda_n}(\theta) \\
&= \underset{\Sigma \geq 0, \Pi, \delta}{\arg\min} - \sum_{t=1}^{T} \sum_{i \in \mathcal{I}_t} \sum_{j \in \mathcal{J}_t} Y_{ijt} \log s_{ijt}(\delta_{\cdot t}, \Pi, \Sigma) + \lambda_n \|(\Pi', \Sigma')'\|_1
\end{aligned} \tag{3.1}
$$

Note that the parameter δ is not penalized because it is generally non-sparse. The properties of RMLE have been discussed in the previous section. The following Assumption 4 and 5 are adapted from Assumption 1-3 to account for multiple markets.

Assumption 4. *Suppose that* $\theta_0 = (\delta_0', \Pi_0', \Sigma_0')' \in \mathbb{R}^{d_\theta}$ *is a unique maximizer of* $\mathbb{E}\left[L_{NJT}(\theta_0)\right] = \mathbb{E}\left[\sum_{t=1}^{T}\sum_{i\in\mathcal{I}_t}\sum_{j\in\mathcal{J}_t} Y_{ijt}\log s_{ijt}(\theta_0)\right]$ *and the score condition*

$$\frac{d}{d\theta}\mathbb{E}\left[L_{NJT}(\theta_0)\right] = \mathbb{E}\left[\sum_{t=1}^{T}\sum_{i\in\mathcal{I}_t}\sum_{j\in\mathcal{J}_t} Y_{ijt}s_{ijt}^{-1}(\theta_0)\frac{d}{d\theta}s_{ijt}(\theta_0)\right] = 0$$

holds. In addition, for every market $t = 1, \ldots, T$,

1. the random vectors $L_i \in [-C_{data}, C_{data}]^{d_L}$ and $X_{jt} \in [-C_{data}, C_{data}]^{d_X}$ have bounded support for all $i \in \mathcal{I}_t$ and $j \in \mathcal{J}_t$;

2. L_i's are i.i.d. across $i \in \cup_t \mathcal{I}_t$;

3. let $J_t := |\mathcal{J}_t|$ and $\min_{j\in\mathcal{J}_t} s_{ijt}(\theta_0) \geq C_s J_t^{-1} > 0$ for a universal constant $C_s \in (0,1)$

Assumption 5. *Suppose that* $-L_{NJT}(\theta)$ *is locally convex in a neighborhood of* $\theta_0 \in \Theta$. *In addition, with probability at least* $1 - a_n'$, *the first-order approximation error*

$$-L_{NJT}(\theta_0+\Delta) + L_{NJT}(\theta_0) + \left(\frac{d}{d\theta}L_{NJT}(\theta_0)\right)'\Delta \geq N\kappa_L\|\Delta\|_2^2 \quad \textit{for all} \quad \Delta \in \mathbb{C}$$

where $N = \sum_{t=1}^{T}|\mathcal{I}_t|$ *is the total number of individuals,* $\mathbb{C} := \{\Delta \in \mathbb{R}^{d_\theta} \mid \sum_{j\in S_n^c}|\Delta_j| \leq 3\sum_{j\in S_n}|\Delta_j|\}$ *is a convex cone and* $\kappa_L > 0$ *is a universal constant.*

Having multiple markets can increase variation in the characteristics X_{jt} and the price P_{jt}, which aids the estimation of $(\bar{\beta}, \beta^{x'}, \alpha)'$. Moreover, the sup-norm of the score $\|\frac{d}{d\theta}L_{NJT}(\theta_0)\|_\infty$ is of the order $\max_{t=1,\ldots T}\max_{j\in\mathcal{J}_t} s_{ijt}^{-1}(\theta_0) = O(\max_t J_t)$ rather than $O(\sum_t J_t)$. These two rates are equivalent when $T < \infty$ is fixed (i.e., a few large markets, each with many alternatives), but the former is smaller when $J < \infty$ is fixed (i.e., many small markets, each with a few alternatives). As a corollary of Lemma 1 and Theorem 1, Corollary 1 derives the rate of the sup-norm as well as the l_1- and l_2-norm of the estimation error.

Corollary 1. *Suppose that Assumption 4 holds. Then, for any* $c > 4C_s^{-1}C_{data}^2$*, with probability greater than* $1 - b_n' = 1 - 2\exp\left((1 - \frac{c^2C_s^2}{8C_{data}^4})\log d_\theta\right)$*,*

$$\left\| \frac{1}{N}\frac{d}{d\theta}L_{NJT}(\theta_0) \right\|_\infty \leq c\rho_N$$

where $\rho_N = \max_t J_t\sqrt{N^{-1}\log d_\theta}$*. This implies* $\|N^{-1}\frac{d}{d\theta}L_{NJT}(\theta_0)\|_\infty = O_P(\max_t J_t\sqrt{N^{-1}\log d_\theta})$*. In addition, given Assumption 5 and choose* $\lambda_N \geq 2N\rho_N$*, then*

$$\|\hat{\theta}^{RMLE} - \theta_0\|_2 \leq \frac{3\sqrt{\mathbf{s}_N}\lambda_N}{N\kappa_L} \quad \textit{and} \quad \|\hat{\theta}^{RMLE} - \theta_0\|_1 \leq \frac{12\mathbf{s}_N\lambda_N}{N\kappa_L}$$

with probability at least $1 - a_n' - b_n'$*, where* $\mathbf{s}_N$ *is the number of non-zero elements in* θ_0*.*

To illustrate the benefit of having multiple markets, consider the scenario that the dimension is low (i.e., $\mathbf{s}_N = d_\theta = O(1)$) and the markets are similar (i.e., $J_t = O(J)$ for every t). In this scenario, the l_2-norm is shrinking when $J_t = o(N^{1/2})$ by Corollary 1, which is a weaker condition compared to $J_t = o(T^{-1}N^{1/2})$ by Theorem 1.

Algorithm 1 can be applied to Eq.(3.1) without extra costs. Notably, dividing a huge market with $N = nT$ individuals and $J = jT$ alternatives into T smaller markets with $|\mathcal{I}_t| = n$ and $|\mathcal{J}_t| = j$ can significantly reduce computational complexity. In this configuration, the complexity of calculating the gradient vector $\frac{d}{d\theta}L_{NJT}(\theta)$ decreases from $O(T^3d_\theta nj^2)$ to $O(Td_\theta nj^2)$, and the complexity of calculating the hessian matrix $\frac{d^2}{d\theta d\theta'}L_{NJT}(\theta)$ decreases from $O(T^4d_\theta^2nJ^3)$ to $O(Td_\theta^2nj^3)$.

The contraction mapping in BLP1995 can also assist in estimating $\hat{\theta}^{RMLE}$, but it comes with additional costs. Their paper demonstrates that, for any pair of parameters Π and Σ, the parameter $\delta_{jt} = \delta_{jt}(\Pi, \Sigma)$ can be obtained by solving the equation $s_{jt}^{data} = s_{jt}(\delta(\Pi, \Sigma), \Pi, \Sigma)$ with an implicit function $\delta(\cdot)$. Their proof holds as long as Π and Σ are finite-dimensional, even if the dimensions are growing. As a result, another version of RMLE is given by

$$(\hat{\Pi}^{RMLE}, \hat{\Sigma}^{RMLE}) = \underset{\Sigma \geq 0, \Pi}{\arg\min} - \sum_{t=1}^{T} \sum_{i \in \mathcal{I}_t} \sum_{j \in \mathcal{J}_t} Y_{ijt} \log s_{ijt}(\delta_{\cdot t}(\Pi, \Sigma), \Pi, \Sigma) + \lambda_n \|(\Pi', \Sigma')'\|_1 \tag{3.2}$$

where $\delta_{jt} = \delta_{jt}(\Pi, \Sigma)$ is solved via contraction mapping. Compared with Eq.(3.1), the sparsity assumption on $\theta^{NL} := (\Pi', \Sigma')'$ (known as the nonlinear parameters) is more plausible since δ is not treated as a parameter. However, since δ is a function of Π and Σ, the gradient $\frac{d}{d\theta^{NL}} s_{ijt}(\delta_{\cdot t}(\theta^{NL}), \theta^{NL}) = \frac{d}{d\theta^{NL}} s_{ijt}(\delta_{\cdot t}, \theta^{NL}) + \frac{d}{d\delta_{\cdot t}} s_{ijt}(\delta_{\cdot t}, \theta^{NL}) \frac{d}{d\theta^{NL}} \delta_{\cdot t}(\theta^{NL})$ needs to be computed by the chain rule. This requires additionally solving the derivative $\frac{d}{d\theta^{NL}} \delta_{\cdot t}(\theta^{NL})$ by the implicit function theorem:

$$\frac{d}{d\theta^{NL}} \delta_{\cdot t} = - \left(\frac{d}{d\delta_{\cdot t}} s_{jt}(\delta_{\cdot t}, \theta^{NL}) \right)^{-1} \frac{d}{d\theta^{NL}} s_{jt}(\delta_{\cdot t}, \theta^{NL})$$

which brings extra complexity to both the proof and the computation. Since the contraction mapping requires high precision and the numerical gradient of $s_{ijt}(\theta^{NL})$ is not feasible when θ^{NL} is high-dimensional, it is unclear whether solving Eq.(3.2) is more efficient than solving Eq.(3.1).

Once the choice-specific utilities δ_{jt}'s are estimated by $\hat{\delta}_{jt}$'s, the parameters $(\bar{\beta}, \beta^{x'}, \alpha)'$ can be identified using instrument variables $Z = (Z_{jt})_{j \in \mathcal{J}_t, t=1,\ldots,T}$ such that $\mathbb{E}[\xi_{jt} \mid Z_{jt}] = 0$ (Train, 2009). The literature has extensively studied the choice and construction of instrument variables as well as the efficient estimation of parameters (see discussions in Nevo, 2000b; Conlon and Gortmaker, 2020). For instance, the two-stage least squares (2SLS) estimator is given by

$$(\hat{\bar{\beta}}, \hat{\beta}^{x'}, \hat{\alpha})' := \left(\sum_{t=1}^{T} \sum_{j \in \mathcal{J}_t} Z_{jt} (1, X_{jt}', P_{jt}) \right)^{-1} \sum_{t=1}^{T} \sum_{j \in \mathcal{J}_t} Z_{jt} \hat{\delta}_{jt} \tag{3.3}$$

where the number of instrument variables is equal to $2 + d_X$. Eq.(3.3) offers insights into

how the bias $\hat{\delta} - \delta := (\hat{\delta}_{jt} - \delta_{jt})_{j\in\mathcal{J}_t, t=1,\dots,T}$ affects the estimation. Since $\delta_{jt} = \bar{\beta} + X'_{jt}\beta^x + P_{jt}\alpha + \xi_{jt}$, then

$$\begin{pmatrix} \hat{\bar{\beta}} \\ \hat{\beta}^x \\ \hat{\alpha} \end{pmatrix} - \begin{pmatrix} \bar{\beta} \\ \beta^x \\ \alpha \end{pmatrix} = \left(\sum_{t=1}^{T}\sum_{j\in\mathcal{J}_t} Z_{jt}(1, X'_{jt}, P_{jt})\right)^{-1}\left(\sum_{t=1}^{T}\sum_{j\in\mathcal{J}_t} Z_{jt}\xi_{jt}\right) + \left(\sum_{t=1}^{T}\sum_{j\in\mathcal{J}_t} Z_{jt}(1, X'_{jt}, P_{jt})\right)^{-1}\left(\sum_{t=1}^{T}\sum_{j\in\mathcal{J}_t} Z_{jt}(\hat{\delta}_{jt} - \delta_{jt})\right), \tag{3.4}$$

suggesting that the 2SLS estimators are $\sqrt{J}$-consistent if $J^{-1/2}Z(\hat{\delta} - \delta) = o_P(1)$.

A key interest in the literature is the estimation of own- and cross-price elasticities, which are defined as below:

$$\frac{\partial s_{jt}}{\partial p_{kt}}\frac{p_{kt}}{s_{jt}} = \begin{cases} \dfrac{p_{jt}}{s_{jt}}\int \alpha s_{ijt}(1 - s_{ijt}) f_\mu(\mu_{i\cdot t} \mid \Pi, \Sigma) d\mu_{i\cdot t} & \text{if } j = k \\ -\dfrac{p_{kt}}{s_{kt}}\int \alpha s_{ikt} s_{ijt} f_\mu(\mu_{i\cdot t} \mid \Pi, \Sigma) d\mu_{i\cdot t} & \text{otherwise} \end{cases}$$

Given the individual-level data (with equal weights in my setting) and the estimated parameters, the elasticities given a price p_{kt} can be approximated by the sample average over individuals:

$$\frac{\partial s_{jt}(\hat{\theta})}{\partial p_{kt}}\frac{p_{kt}}{s_{jt}(\hat{\theta})} \approx \begin{cases} \dfrac{p_{jt}}{\sum_{i\in\mathcal{I}_t}\int \tilde{s}_{ijt}\phi(\nu_i)d\nu_i}\sum_{i\in\mathcal{I}_t}\int \alpha\tilde{s}_{ijt}\left(1 - \tilde{s}_{ijt}\right)\phi(\nu_i)d\nu_i & \text{if } j = k \\ -\dfrac{p_{kt}}{\sum_{i\in\mathcal{I}_t}\int \tilde{s}_{ikt}\phi(\nu_i)d\nu_i}\sum_{i\in\mathcal{I}_t}\int \alpha\tilde{s}_{ikt}\tilde{s}_{ijt}\phi(\nu_i)d\nu_i & \text{otherwise} \end{cases} \tag{3.5}$$

where $\tilde{s}_{ijt} = \dfrac{\exp\left(\hat{\delta}_{jt} + (X_{jt}\otimes L_i)'\hat{\Pi} + (X_{jt}\odot\nu_i)'\hat{\Sigma}\right)}{1 + \sum_{k\in\mathcal{J}_t}\exp\left(\hat{\delta}_{kt} + (X_{kt}\otimes L_i)'\hat{\Pi} + (X_{kt}\odot\nu_i)'\hat{\Sigma}\right)}$ is the predicted individual choice probability conditioning on the taste ν_i.

4 Constrained Debiased Machine Learning

Recall that in Section 2 and 3, I discuss the challenges in estimating θ when the parameter Π is high-dimensional, and prove the consistency of my regularized estimator $\hat{\theta}^{RMLE}$ under mild assumptions. In this section, I partition the parameter $\theta := (\gamma', \eta_N')' \in \Theta$ into a low-dimensional vector $\gamma \in \Gamma$, which represents the target parameter, and a high-dimensional[8] vector $\eta_N \in \mathcal{N}_N$, which represents the nuisance parameter. I assume that the parameter space Θ can be decomposed as $\Gamma \times \mathcal{N}_N$, where Γ and $\mathcal{N}_N$ are compact sets for any finite sample size N. Then, I conduct inference for the target parameter γ by constructing a $\sqrt{N}$-consistent estimator $\hat{\gamma}$, allowing both $\gamma \in \Gamma$ and $\eta_N \in \mathcal{N}_N$ to potentially lie on the boundary of their parameter spaces. As an example, let $\gamma := (\Pi_1, \Sigma')'$ consist of the first element of Π (e.g., the effect of income) and the whole Σ (e.g., the heterogeneity on soft-drink flavors). In this case, $\Gamma := [-C, C] \times [0, C)^{d_\Sigma}$ for some large number $C > 0$. Theoretically, my results also apply to the case $\gamma := (\delta', \Pi_1, \Sigma')' \in \Gamma := [-C, C]^{d_\delta + 1} \times [0, C)^{d_\Sigma}$, when δ is low-dimensional.

The idea is to construct a Neyman orthogonal score function based on the score function of the MLE, which can partial out the first-order bias in the estimation of nuisance parameters (Ning and Liu, 2017; Chernozhukov et al., 2018; Kennedy, 2023). Li (2024) proposes a framework called constrained debiased machine learning (CDML), which extends the debiased machine learning (Chernozhukov et al., 2018) to allow for constraints on parameters. Let $L_{NJ}(\gamma, \eta_N) := \sum_{i=1}^{N} \sum_{j=0}^{J} Y_{ij} \log s_{ij}(\gamma, \eta_N)$ denote the log-likelihood function where $s_{ij}(\gamma, \eta_N)$ is the choice probability integrated over ν_i. The generalization to multiple markets is straightforward so $T = 1$ is assumed for simplicity. Then, the Neyman orthogonal score function for $L_{NJ}(\gamma, \eta_N)$ is given by

$$M_{NJ}(\gamma; \eta_N, \mu_N) := \frac{1}{N} \frac{\partial}{\partial \gamma} L_{NJ}(\gamma, \eta_N) - \frac{1}{N} \mu_N \frac{\partial}{\partial \eta_N} L_{NJ}(\gamma; \eta_N) \tag{4.1}$$

[8] Although θ is also high-dimensional, I only add subscript N to η (and μ later) to highlight their dimensions are growing as $N \to \infty$.

where μ_N is a d_γ-by-d_{η_N} de-correlation matrix that will be introduced later. The score condition of MLE $\mathbb{E}[\frac{d}{d\theta}L_{NJ}(\gamma_0, \eta_{N,0})] = 0$, for example, in Assumptions 1 or 4, implies $\mathbb{E}[M_{NJ}(\gamma_0; \eta_{N,0}, \mu)] = 0$ for any μ. Thus, a naive DML estimator $\hat{\gamma}^{DML}$, conditioning on the first-stage machine learning (e.g., LASSO) estimators $\hat{\eta}_N$ and $\hat{\mu}_N$, can be defined as the solution to $M_{NJ}(\gamma; \hat{\eta}_N, \hat{\mu}_N) = 0$. It is ideal if $\hat{\eta}_N \approx \eta_{N,0}$ and $\hat{\mu}_N \approx \mu_{N,0}$, but the estimator tend to overfit the noise due to over-parameterization, which may cause additional bias. To mitigate the risk of overfitting in $\hat{\eta}_N$ and $\hat{\mu}_N$, it is recommended[9] to implement the technique known as *K-fold cross-fitting* from the machine learning literature. Formally, the procedure is introduced as below.

For some natural number $K \in \mathbb{N}$, consider a partition $\mathbf{I}_1, \ldots, \mathbf{I}_K$ of the indices $\{1, \ldots, N\}$ and define $\mathbf{I}_{-k} := \{1, \ldots, N\} \backslash \mathbf{I}_k$. For simplicity, assume that the size of $\mathbf{I}_1, \ldots, \mathbf{I}_K$ are equal to $\lfloor N/K \rfloor$. Let $L_{NJ}^{(k)}(\theta) := \sum_{i \in \mathbf{I}_k} \sum_{j=0}^{J} Y_{ij} \log s_{ij}(\gamma, \eta_N)$ denote the testing-data log-likelihood associated with the individuals indexed by $\mathbf{I}_k$ and $L_{NJ}^{(-k)}(\theta) := \sum_{i \in \mathbf{I}_{-k}} \sum_{j=0}^{J} Y_{ij} \log s_{ij}(\gamma, \eta_N)$ denote the training-data log-likelihood from individuals indexed by $\mathbf{I}_{-k}$. Then, for each $k = 1, \ldots, K$:

1. Solve the first-stage regularized estimator $\hat{\theta}_k^{RMLE}$ based on the training data

$$\hat{\theta}_k^{RMLE} = (\hat{\gamma}_k^{RMLE}, \hat{\eta}_{N,k}^{RMLE}) := \arg\min_{\theta \in \Theta} L_{NJ}^{(-k)}(\theta) + \frac{K-1}{K} \lambda_N \|\theta\|_1$$

2. Solve the d_γ-by-d_{η_N} de-correlation matrix $\hat{\mu}_{N,k}^{RMLE}$ as a dantzig selector (Candes and Tao, 2007)

$$\hat{\mu}_{N,k}^{RMLE} = \arg\min \|\mu_k\|_1 \quad \text{such that} \quad \|(\hat{J}_{\eta\eta}^{(-k)})(\hat{J}_{\gamma\eta}^{(-k)} - \mu_k \hat{J}_{\eta\eta}^{(-k)})'\|_\infty \leq a_N \tag{4.2}$$

where $a_N \to 0$ is a tuning parameter and the hessian matrix $\frac{d}{d\theta} L_{NJ}^{(-k)}(\hat{\theta}^{RMLE})$ is

[9] It is not necessary to implement cross-fitting using LASSO in high-dimensional linear models, for example, see Ning and Liu (2017) and Li (2024). However, cross-fitting may be useful in non-linear models.

partitioned into four blocks corresponding to the partition of θ, denoted as

$$\begin{pmatrix} \hat{J}_{\gamma\gamma}^{(-k)} & \hat{J}_{\gamma\eta}^{(-k)} \\ \hat{J}_{\eta\gamma}^{(-k)} & \hat{J}_{\eta\eta}^{(-k)} \end{pmatrix} := \hat{J}^{(-k)} := \frac{d}{d\theta} L_{NJ}^{(-k)}(\hat{\theta}_k^{RMLE})$$

If $\hat{J}_{\eta\eta}$ is invertible and not ill-conditioned, then $\hat{\mu}_{N,k}^{RMLE} := (\hat{J}_{\eta\eta}^{(-k)})^{-1}\hat{J}_{\gamma\eta}^{(-k)}$.

The larger $K \in \mathbb{N}$ is, the more individuals are included in each training set $\mathbf{I}_{-k}$. In practice, K is typically chosen to be 5 or 10, using 80% or 90% of data to estimate LASSO for each k, respectively. To understand the selector in Eq.(4.2), let $\hat{J}_{\gamma\eta,1}^{(-k)}$ and $\mu_{k,1}$ be the first rows of $\hat{J}_{\gamma\eta}^{(-k)}$ and μ_k, respectively. We can always write $\hat{J}_{\gamma\eta,1}^{(-k)} = \mu_{k,1}\hat{J}_{\eta\eta}^{(-k)} + \epsilon_1$ as a linear projection with the 1-by-d_{η_N} projection error ϵ_1, and the constraint implies $\|\hat{J}_{\eta\eta}^{(-k)}\epsilon_1\|_\infty \leq a_N$. When the dimension d_{η_N} is small, we can estimate $\mu_{k,1}$ by the least squares. When d_{η_N} is large, there is no unique exact solution to $\hat{J}_{\gamma\eta,1}^{(-k)} = \mu_{k,1}\hat{J}_{\eta\eta}^{(-k)}$ as $\hat{J}_{\eta\eta}^{(-k)}$ can be singular, and Eq.(4.2) finds the approximate solution with the minimal l_1-norm. In this section, the true value $\mu_{N,0}$ is defined as $\mathbb{E}[\frac{\partial^2}{\partial\eta\partial\eta'}L_{NJ}(\theta_0)]^{-1}\mathbb{E}[\frac{\partial^2}{\partial\gamma\partial\eta'}L_{NJ}(\theta_0)]$ by assuming the first expectation is non-singular. Chernozhukov et al. (2018) also shows that Eq.(4.1) satisfies the Neyman near-orthogonal condition if $\mu_{N,0} := \arg\min \|\mu\|_1$ such that

$$\left\| \mathbb{E}\left[\frac{\partial^2}{\partial\eta\partial\eta'}L_{NJ}(\theta_0)\right]\left(\mathbb{E}\left[\frac{\partial^2}{\partial\gamma\partial\eta'}L_{NJ}(\theta_0)\right] - \mu\mathbb{E}\left[\frac{\partial^2}{\partial\eta\partial\eta'}L_{NJ}(\theta_0)\right]\right)' \right\|_\infty \leq a_N$$

According to Li (2024), the CDML estimator $\hat{\gamma}^{CDML}$ given the first-stage estimates $\{\hat{\eta}_{N,k}^{RMLE}, \hat{\mu}_{N,k}^{RMLE}\}_{k=1}^K$, is defined as

$$\begin{aligned} \hat{\gamma}^{CDML} &:= \arg\min_{\gamma\in\Gamma} M_{NJK}(\gamma; \hat{\eta}_N^{RMLE}, \hat{\mu}_N^{RMLE})' M_{NJK}(\gamma; \hat{\eta}_N^{RMLE}, \hat{\mu}_N^{RMLE}) \\ &:= \arg\min_{\gamma\in\Gamma} \left\| \frac{1}{K}\sum_{k=1}^K \frac{1}{N/K}\sum_{i\in\mathbf{I}_k} m_i(\gamma; \hat{\eta}_{N,k}^{RMLE}, \hat{\mu}_{N,k}^{RMLE}) \right\|_2^2 \end{aligned} \tag{4.3}$$

where $m_i(\gamma;\eta,\mu) := \sum_{j=0}^J Y_{ij}\left(\frac{\partial}{\partial\gamma}\log s_{ij}(\gamma,\eta) - \mu\frac{\partial}{\partial\eta}\log s_{ij}(\gamma,\eta)\right)$ is the orthogonal score

function for individual i (cf. Eq.(3.5.) in Chernozhukov et al., 2018). To shorten notation, let $M_N(\gamma) := M_{NJK}(\gamma; \hat{\eta}_N^{RMLE}, \hat{\mu}_N^{RMLE})$. In fact, $\hat{\gamma}^{CDML}$ is a method of moment estimator while the solution to $M_N(\gamma) = 0$ may be infeasible due to the constraints in Γ. Hopefully, the objective function is well-defined even for $\gamma \notin \Gamma$ such that the derivative $\frac{d}{d\gamma}\|M_N(\gamma)\|_2^2$ at $\gamma_0 \in \Gamma$ can be taken from any directions, which makes the problem easier to analyze (cf. Andrews, 1999; Ketz, 2018). For the rest of this section, I will prove that (i) $\hat{\gamma}^{CDML} \to_P \gamma_0$ is consistent; (ii) $\sqrt{N}(\hat{\gamma}^{CDML} - \gamma_0) \to_d N(0, V)$ if γ_0 is an interior point of Γ; and (iii) $\sqrt{N}(\hat{\gamma}^{CDML} - \gamma_0)$ converges to the projection of a multivariate normal distribution onto a polytope if γ_0 is a boundary point of Γ.

Consider the second-order Taylor's expansion of $\|M_{NJK}(\gamma)\|_2^2$ at γ_0. Some algebraic works show that

$$\begin{aligned}\|M_N(\gamma)\|_2^2 = \|M_N(\gamma_0)\|_2^2 - \frac{1}{2} D_\gamma \|M_N(\gamma_0)\|_2^2 \left[D_\gamma^2 \|M_N(\gamma_0)\|_2^2\right]^{-1} D_{\gamma'} \|M_N(\gamma_0)\|_2^2 \\ - \frac{1}{2N} q_N(\sqrt{N}(\gamma - \gamma_0)) + R_N(\gamma, \gamma_0)\end{aligned} \tag{4.4}$$

where $R_N(\gamma, \gamma_0)$ is the remainder term and

$$\begin{aligned} D_{\gamma'} \|M_N(\gamma)\|_2^2 :=& 2 \left[D_{\gamma'} M_N(\gamma)\right] M_N(\gamma) \\ D_\gamma^2 \|M_N(\gamma)\|_2^2 :=& 2 \left[D_{\gamma'} M_N(\gamma)\right] D_\gamma M_N(\gamma) + 2 \sum_{j=1}^{d_\gamma} M_{N,j}(\gamma) D_\gamma^2 M_{N,j}(\gamma) \\ q_N(x) :=& -\left(x + \left[D_\gamma^2 \|M_N(\gamma_0)\|_2^2\right]^{-1} N^{1/2} D_{\gamma'} \|M_N(\gamma_0)\|_2^2\right)' D_\gamma^2 \|M_N(\gamma_0)\|_2^2 \\ & \left(x + \left[D_\gamma^2 \|M_N(\gamma_0)\|_2^2\right]^{-1} N^{1/2} D_{\gamma'} \|M_N(\gamma_0)\|_2^2\right) \end{aligned} \tag{4.5}$$

Here $M_{N,j}$ is the j-th coordinate of the vector M_N. On the right-hand side of Eq.(4.4), only the quadratic function $q_N(\cdot)$ and the remainder $R_N(\gamma, \gamma_0)$ depend on γ. If the remainder is sufficiently small in the neighborhood of γ_0 and the estimator $\hat{\gamma}^{CDML} \to_p \gamma_0$ is consistent, it is equivalent to study the asymptotic properties of $-\frac{1}{2N} q_N(\sqrt{N}(\gamma - \gamma_0))$ and $\|M_N(\gamma)\|_2^2$. According to Andrews (1999, Lemma 1), the following Assumption 6 guarantees that

$R_N(\gamma, \gamma_0)$ is asymptotic negligible, which is mild as $M_N(\gamma)$ is a smooth function of γ.

Assumption 6. *For any* $c_N \to 0$,

$$\sup_{\gamma \in \Theta_\gamma(c_N)} \left\| D_\gamma^2 \|M_N(\gamma)\|_2^2 - D_\gamma^2 \|M_N(\gamma_0)\|_2^2 \right\|_F = o_P(1)$$

where $\Theta_\gamma(c_N) := \{\gamma \in \Theta_\gamma : \|\gamma - \gamma_0\|_2 \leq c_N\}$.

Given Assumption 6, it suffices to show the convergence of the vector $D_{\gamma'} \|M_N(\gamma)\|_2^2$ and the matrix $D_\gamma^2 \|M_N(\gamma)\|_2^2$ by the definitions in Eq.(4.5). Under some regularity conditions, it is expected that $M_{NJK}(\gamma_0; \eta_{N,0}, \mu_{N,0}) \to_P 0$ and $D_{\gamma'} M_{NJK}(\gamma_0; \eta_{N,0}, \mu_{N,0}) \to_P \Omega_M$ for some symmetric and non-singular matrix Ω_M by the law of large numbers, and $N^{1/2} M_{NJK}(\gamma_0; \eta_{N,0}, \mu_{N,0}) \to_d N(0, \Sigma_M)$ for some variance-covariance matrix Σ_M by the central limit theorem. The term $D_\gamma^2 M_{NJK,j}(\gamma_0; \eta_{N,0}, \mu_{N,0})$ is tricky, however, it is asymptotically negligible as long as its rate is slower than $\sqrt{N}$. Then, by the Slutsky's theorem,

$$\begin{aligned} N^{1/2} D_{\gamma'} \|M_N(\gamma_0)\|_2^2 &= 2 \left[D_{\gamma'} M_{NJK}(\gamma_0; \eta_{N,0}, \mu_{N,0}) \right] N^{1/2} M_{NJK}(\gamma_0; \eta_{N,0}, \mu_{N,0}) \\ &\to_d 2\Omega_M N(0, \Sigma_M) \\ D_\gamma^2 \|M_N(\gamma_0)\|_2^2 &= 2 \left[D_{\gamma'} M_{NJK}(\gamma_0; \eta_{N,0}, \mu_{N,0}) \right] D_{\gamma'} M_{NJK}(\gamma_0; \eta_{N,0}, \mu_{N,0}) + o_P(1) \\ &\to_P 2\Omega_M \Omega_M' \end{aligned} \tag{4.6}$$

as $N \to \infty$. Although $\hat{\eta}_N$ is shown to be "consistent" in the previous sections, its rate is slower than $\sqrt{N}$ due to the high dimensions, so that $\sqrt{N}(\hat{\eta}_N - \eta_{N,0})$ diverges. We need to additionally shows that $\sqrt{N} M_{NJK}(\gamma_0; \hat{\eta}_N^{RMLE}, \hat{\mu}_N^{RMLE})$ is a good approximation of $\sqrt{N} M_{NJK}(\gamma_0; \eta_{N,0}, \mu_{N,0})$ (and also their partial derivatives with respect to γ) when $\hat{\eta}_{N,0}$ is close to $\eta_{N,0}$. Formally, it suffices to prove that

$$\begin{aligned} \sqrt{N} \|M_{NJK}(\gamma_0; \hat{\eta}_N^{RMLE}, \hat{\mu}_N^{RMLE}) - M_{NJK}(\gamma_0; \eta_{N,0}, \mu_{N,0})\|_2 &= o_P(1) \\ \|D_{\gamma'} M_{NJK}(\gamma_0; \hat{\eta}_N^{RMLE}, \hat{\mu}_N^{RMLE}) - D_{\gamma'} M_{NJK}(\gamma_0; \eta_{N,0}, \mu_{N,0})\|_F &= o_P(1) \end{aligned} \tag{4.7}$$

In Theorem 2, I will show that both Eq.(4.6) and Eq.(4.7) hold given the following Assumptions 7 and 8.

Assumption 7 (Assumptions for Approximations). *Let* $\{\Delta_N, r_{N,\eta}, r_{N,\mu}\}_{N=1}^{\infty}$ *be sequences of constants such that* $\Delta_N, r_{\eta,N}, r_{\mu,N} \to 0$ *as* $N \to \infty$. *Suppose that the following conditions hold:*

1. The data covariates are all bounded by a universal constant $C_{data} > 0$ that does not depend on N;

2. With probability greater than $1-\Delta_N$, the nuisance estimators $\hat{\eta}_{N,k} \in T_N^{\eta}$ and $\hat{\mu}_{N,k} \in T_N^{\mu}$ belong to some nuisance realization sets $T_N^{\eta} := \{\eta : \|\eta-\eta_{N,0}\|_1 \vee \|\eta-\eta_{N,0}\|_2 \leq r_{\eta,N}\} \subset \mathcal{N}_N$ and $T_N^{\mu} := \{\mu : \|\mu - \mu_{N,0}\|_1 \leq r_{\mu,N}\} \subset \mathbb{R}^{d_\gamma \times d_{\eta_N}}$, respectively;

3. For any $\eta_N \in T_N^{\eta}$, $\min_{j=0,\ldots,J} s_{ij}(\gamma_0, \eta_N) \geq C_s J^{-1}$;

4. $\mu_{N,0}$ is a sparse matrix such that $\|\mu_{N,0}\|_1 = \mathbf{s}_{\mu,N}$ for a sequence of constants $\mathbf{s}_{\mu,N} = o(\sqrt{N})$.

Conditions 1 and 3 in Assumption 7, adapted from Assumption 3, are sufficient to bound the sup-norm $\|\frac{d}{d\theta}L_{NJ}(\theta_0)\|_\infty$. Combined with Condition 4, $\|M_N(\gamma_0)\|_\infty$ is also bounded. I refer to the proofs on K-fold cross-fitting in Chernozhukov et al. (2018), and derive the rate for $M_{NJK}(\gamma; \hat{\eta}_N^{RMLE}, \hat{\mu}_N^{RMLE}) - M_{NJK}(\gamma_0; \eta_{N,0}, \mu_{N,0})$ by assuming Condition 2. The rates $r_{\eta,N}$ and $r_{\mu,N}$ rely on the machine learning method applied in the first-stage estimation. For example, in the case of LASSO, the rate $r_{\eta,N} = J\mathbf{s}_{\mu,N}\sqrt{N^{-1}\log d_\theta}$ is given by Theorem 1.

Assumption 8 (Assumptions for LLN and CLT). $N^{-1/2}(1 + \mathbf{s}_{\mu,N})J = o(1)$ *and*

$$\lim_{N\to\infty} Var(m_i(\gamma_0; \eta_{N,0}, \mu_{N,0})) = \Sigma_M$$

for some positive definite matrix Σ_M.

Assumption 8 is a regularity condition for the Lindeberg-Feller central limit theorem.

Theorem 2 (Approximation and Convergence)**.** *Under Assumption 7,*

$$\begin{aligned}
&\sqrt{N}\|M_{NJK}(\gamma_0;\hat{\eta}_N^{RMLE},\hat{\mu}_N^{RMLE}) - M_{NJK}(\gamma_0;\eta_{N,0},\mu_{N,0})\|_2 \\
&\qquad = O_P\left(J^2(1+\mathbf{s}_{\mu,N})r_{\eta,N} + Jr_{\mu,N} + \sqrt{N}J^3(1+\mathbf{s}_{\mu,N})r_{\eta,N} + \sqrt{N}J^2 r_{\mu,N}r_{\eta,N}\right) \\
&\|D_{\gamma'}M_{NJK}(\gamma_0;\hat{\eta}_N^{RMLE},\hat{\mu}_N^{RMLE}) - D_{\gamma'}M_{NJK}(\gamma_0;\eta_{N,0},\mu_{N,0})\|_F \\
&\qquad = O_P\left(J^3(1+\mathbf{s}_{\mu,N})r_{\eta,N} + J^2 r_{\mu,N} + J^3 r_{\eta,N}r_{\mu,N}\right) \\
&\|D_\gamma^2 M_{NJK,j}(\gamma_0;\eta_{N,0},\mu_{N,0})\|_F = O_P(J^3(1+\mathbf{s}_{\mu,N})) \quad \textit{for any } \gamma = 1,\ldots,d_\gamma
\end{aligned}$$

Furthermore, if Assumption 8 holds, then

$$N^{1/2}M_{NJK}(\gamma_0;\eta_{N,0},\mu_{N,0}) \to_d N(0,\Sigma_M)$$
$$D_{\gamma'}M_{NJK}(\gamma_0;\eta_{N,0},\mu_{N,0}) \to_p \Omega_M$$

where

$$\Omega_M := \mathbb{E}\left[\sum_{j=0}^{J}\Upsilon_{ij}\left(\frac{\partial^2}{\partial\gamma\partial\gamma'}\ln s_{ij}(\gamma_0,\eta_{N,0}) - \mu_{N,0}\frac{\partial^2}{\partial\eta\partial\gamma'}\ln s_{ij}(\gamma_0,\eta_{N,0})\right)\right]$$

Theorem 2 establishes both the rates of approximation and the limiting distributions. It imposes restrictions on the sparsity $\mathbf{s}_{\mu,N}$, the number of alternatives J, and the quality of machine learning algorithms reflected in $r_{\eta,N}$ and $r_{\mu,N}$. In the case of LASSO, if we assume that the sparsity of parameters satisfies $\mathbf{s}_{\eta,N} \vee \mathbf{s}_{\mu,N} = O(1)$, then the qualify of

approximation is guaranteed when

$$\begin{aligned} r_{\eta,N} &= o\left(\frac{1}{J^3} \wedge \frac{1}{N^{1/4}J^{3/2}}\right) \\ r_{\mu,N} &= o\left(\frac{1}{J^2}\right) \\ r_{\eta,N} r_{\mu,N} &= o\left(\frac{1}{N^{1/2}J^2}\right) \\ \frac{J^3}{\sqrt{N}} &= o(1) \end{aligned} \tag{4.8}$$

Furthermore, if we assume that J is small compared to N (e.g., $J = \log N$), the conditions in Eq.(4.8) holds if $r_{\eta,N}$ and $r_{\mu,N}$ are smaller than $N^{-1/4}$, which is achievable for various machine learning algorithms and nonparametric methods.

Assumption 9. *Suppose that Γ is a non-empty compact subset of $\mathbb{R}^{d_\gamma}$, $\gamma_0 \in \Gamma$ is the unique solution to $\mathbb{E}[M_{NJK}(\gamma; \eta_{N,0}, \mu_{N,0})] = 0$ and $\sup_{\gamma\in\Gamma} \max_j s_{ij}^{-1}(\gamma, \eta_{N,0}) = O(J \log N)$ almost surely. In addition, assume $\mathbf{s}_{\mu,N} = O(1)$ and the rates in Eq.(4.8).*

Assumption 9 provides sufficient conditions for the uniform law of large numbers (ULLN) to prove $\hat{\gamma}^{CDML} \to_P \gamma_0$. In addition to the conditions for identification and approximation[10], I assume that $\sup_{\gamma\in\Gamma} \max_j s_{ij}^{-1}(\gamma, \eta_{N,0})$ may diverge (slightly faster than the rate at $\gamma = \gamma_0$) but the divergence is not too fast. As a crucial condition in the ULLN,

$$\sup_{\gamma\in\Gamma} \left| \|M_{NJK}(\gamma; \hat{\eta}_N^{RMLE}, \hat{\mu}_N^{RMLE})\|_2^2 - \|\mathbb{E}[M_{NJK}(\gamma; \eta_{N,0}, \mu_{N,0})]\|_2^2 \right| = o_P(1) \tag{4.9}$$

may fail if $s_{ij}^{-1}(\gamma, \eta_{N,0})$ can be arbitrarily large for fixed J. Since the term $s_{ij}^{-1}(\gamma, \eta_{N,0})$ comes from the MLE score function $\frac{d}{d\theta} \ln s_{ij}(\gamma, \eta_{N,0})$, and the score can be regarded as the efficient instrument (see Section 2), 9 can be interpreted as a boundedness condition on the instrument. For example, the Assumption A4 in Berry et al. (2004b) require the instrument

[10]The rate on $\mathbf{s}_{\mu,N}$ can be relaxed, as long as the approximation in Theorem 2 is good. I assume $\mathbf{s}_{\mu,N} = O(1)$ for simplicity.

variables to be $O_P(\sqrt{J})$, which is equivalent to $\sup_{\gamma\in\Gamma}\max_j s_{ij}^{-1}(\gamma,\eta_{N,0}) = (1+\mathbf{s}_{\mu,N})^{-1}\sqrt{J}$ in my setting.

Theorem 3 below proves the consistency of $\hat{\gamma}^{CDML}$, and then Theorem 4 derives the limiting distribution for $\sqrt{N}(\hat{\gamma}^{CDML}-\gamma_0)$.

Theorem 3 (Consistency)**.** *Given Assumption 7 and 9, $\hat{\gamma}^{CDML}\to_P\gamma_0$ as $N\to\infty$.*

Theorem 4 (Limiting Distribution)**.** *Suppose that Assumptions 6, 7, 8 and 9 hold. Moreover, assume that $\Omega_M\Omega_M'$ is positive-(semi)definite. Then, as $N\to\infty$,*

$$\sqrt{N}(\hat{\gamma}^{CDML}-\gamma_0)\to_d\tilde{\gamma}:=\underset{\xi\in\Gamma(\gamma_0)}{\arg\min}\|\xi+N(0,V_M)\|_{\Omega_M}^2$$

where $\Gamma(\gamma_0)$ is a convex cone locally equal[11] *to $\Gamma-\gamma_0$, $V_M:=(\Omega_M\Omega_M')^{-1}\Omega_M\Sigma_M\Omega_M'(\Omega_M\Omega_M')^{-1}$, and the (semi-)norm $\|\gamma\|_{\Omega_M}:=\sqrt{\gamma'\Omega_M\Omega_M'\gamma}$. Moreover,*

$$\begin{aligned}\mathcal{M}_N(\gamma_0):=&N\left(\|M_{NJK}(\hat{\gamma}^{CDML};\hat{\eta}_N^{RMLE},\hat{\mu}_N^{RMLE})\|_2^2-\|M_{NJK}(\gamma_0;\hat{\eta}_N^{RMLE},\hat{\mu}_N^{RMLE})\|_2^2\right)\\ \to_d&-\tilde{\gamma}'\Omega_M\Omega_M'\tilde{\gamma}\end{aligned}$$

When γ_0 is an interior point of Γ, $\tilde{\gamma}$ follows the distribution $N(0,V_M)$ which aligns with the DML literature. The variance-covariance matrix V_M consists of two components: Σ_M and Ω_M. Here, Σ_M is the population variance-covariance matrix for $m_i(\gamma_0;\eta_{N,0},\mu_{N,0})$, and hence, a simple estimator is the variance-covariance matrix averaged over K:

$$\begin{aligned}\hat{\Sigma}_M:=&\frac{1}{K}\sum_{k=1}^K\frac{1}{N/K}\sum_{i\in\mathcal{I}_k}m_i(\hat{\gamma}^{CDML};\hat{\eta}_{N,k}^{RMLE},\hat{\mu}_{N,k}^{RMLE})m_i'(\hat{\gamma}^{CDML};\hat{\eta}_N^{RMLE},\hat{\mu}_{N,k}^{RMLE})\\ &-M_{NJK}(\hat{\gamma}^{CDML};\hat{\eta}_N^{RMLE},\hat{\mu}_N^{RMLE})M_{NJK}'(\hat{\gamma}^{CDML};\hat{\eta}_N^{RMLE},\hat{\mu}_N^{RMLE})\end{aligned}$$

Recall the definition of Ω_M in Theorem 2, then the estimator $\hat{\Omega}_M$ can be constructed based

[11] If a vector b is in a parameter space B, then $B-b=\{x-b\mid x\in B\}$ is the shifted parameter space, and $0\in B-b$. We say a convex cone A is locally equal to $B-b$ if $A\cap Ball(0,\epsilon)=B\cap Ball(0,\epsilon)$ for some $\epsilon>0$.

on $\hat{J}$:

$$\hat{\Omega}_M = \frac{1}{K}\sum_{k=1}^{K} \hat{J}_{\gamma\gamma}^{(-k)} - \hat{\mu}_{N,k}^{RMLE}\hat{J}_{\eta\gamma}^{(-k)}$$

Note that $\hat{\Omega}_M$ is not necessarily symmetric, especially under cross-fittings. Although the distribution of $-\tilde{\gamma}'\Omega_M\Omega_M'\tilde{\gamma}$ does not have an explicit form for general $\Gamma(\gamma_0) \subset \mathbb{R}^{d_\gamma}$, its quantile $C(\alpha, \gamma_0)$ can be simulated given the estimators $\hat{\Omega}_M$ and $\hat{\Sigma}_M$. In most scenarios, $\Gamma(\gamma_0)$ can be written as the product space of $(-\infty, \infty)$, $(-\infty, 0]$ and $[0, \infty)$, so $\tilde{\gamma}$ is the projection onto a convex set and can be easily solved numerically by convex cone quadratic programming. Depending on the setting, sometimes $\tilde{\gamma}$ may have a closed-form solution. See the following two examples.

Example 1. If the standard deviation is the only target parameter, then $\gamma = \Sigma \in [0, \infty)$. As a result, the convex cone $\Gamma(\gamma_0) = (-\infty, \infty)$ if the true value $\Sigma_0 > 0$, and $\Gamma(\gamma_0) = [0, \infty)$ if the true value $\Sigma_0 = 0$. As a result, $\tilde{\gamma} \sim_d N(0, V_M)$ in the former case while $\tilde{\gamma} \sim_d \max\{0, -N(0, V_M)\}$ in the latter case.

Example 2. Suppose that there is only one random coefficient and its true value is on the boundary. Let $\gamma = (\alpha', \Sigma)' = (\alpha', 0)'$, the projection $\tilde{\gamma}$ is explicit by verifying the Karush-Kuhn-Tucker (KKT) condition. Suppose that the projection has the form

$$\tilde{\gamma} = \underset{\xi_\Sigma \geq 0}{\arg\min} \begin{pmatrix} \xi_\alpha + b_\alpha \\ \xi_\Sigma + b_\Sigma \end{pmatrix}' \begin{pmatrix} A & B \\ C & D \end{pmatrix} \begin{pmatrix} \xi_\alpha + b_\alpha \\ \xi_\Sigma + b_\Sigma \end{pmatrix}$$

where $(b_\alpha', b_\Sigma)'$ follows the multivariate normal distribution $N(0, V_M)$. Then, it can be derived that

$$(\tilde{\gamma}_\alpha, \tilde{\gamma}_\Sigma) = \begin{cases} (-\frac{1}{2}A^{-1}(C' + B)b_\Sigma - b_\alpha, 0) & \text{if } \left(2D - \frac{1}{2}(C + B')A^{-1}(C' + B)\right) b_\Sigma > 0 \\ (-b_\alpha, -b_\Sigma) & \text{otherwise} \end{cases}$$

where $\tilde{\gamma}_\Sigma$ is a mixture of a degenerate distribution at zero and a normal distribution, and

$\tilde{\gamma}_\alpha$ is a mixture of normal distributions.

In addition, Theorem 4 implies that $\mathcal{M}_n(\gamma_0)$ can be regarded as a quasi-likelihood ratio statistic, whose critical values can be obtained from simulations. I introduce the procedure in Algorithm 2.

Algorithm 2 QLR Test with Simulated Critical Values

1. Given a null hypothesis $H_0 : \gamma_0 = b$, solve the convex cone $\Gamma(b)$;
2. Estimate the CDML variance $\hat{V}_M$ and the matrix $\hat{\Omega}_M$;
3. Draw a sequence of vectors $x_1, \ldots, x_B \sim i.i.d.\ N(0, I_{d_\gamma}))$, then solve
$$\tilde{\gamma}_i = \arg\min_{\xi \in \Gamma(b)} \left\| \xi + \hat{V}_M^{1/2} x_i \right\|_{\hat{\Omega}_M}^2 \quad \text{for} \quad i = 1, \ldots, B$$
4. Solve $\hat{C}_i(b) := -\tilde{\gamma}_i' \hat{\Omega}_M \hat{\Omega}_M' \tilde{\gamma}_i$ for $i = 1, \ldots, B$;
5. The α-quantile of $C_i(b)$, denoted by $\hat{C}_{n,B}(\alpha, b)$, is the critical value for the QLR test at the level α.

5 Simulation

5.1 Settings

In this section, I illustrate the effects of high dimensions and parameters on the boundary by Monte Carlo simulations. Consider the following model of indirect utility for individual i choosing alternative j in market t:

$$\begin{aligned} U_{ijt} &= \delta_{jt} + \mu_{ijt} + \varepsilon_{ijt} \\ \delta_{jt} &= \bar{\beta} + X_{jt}' \beta^x + P_{jt} \alpha = -2 + 3X_{jt} - 2P_{jt} \\ \mu_{ijt} &= (X_{jt} \otimes L_i)' \Pi + (X_{jt} \odot \nu_i)' \Sigma = (1, 2, 3, 0, \ldots, 0)(X_{jt} \otimes L_i) + X_{jt} \nu_i \Sigma \end{aligned}$$

where $\delta_{0t} = \mu_{i0t} = 0$ for every $t = 1, \ldots, T$. The exogenous variable X_{jt} is a scalar that follows i.i.d. $N(0,1)$ truncated by the interval $[-3,3]$. There is only one random coefficient with standard deviation Σ which is equal to either 0 (on the boundary) or 1 (off the boundary). Many questions are true-or-false or categorical in survey data, and researchers construct dummy variables carefully to avoid of multi-collinearity. To mimic the scenario, I partition the d_L-dimensional random vector L_i into a vector $L_{i,normal}$ of length $\lfloor d_L/10 \rfloor$, which follows $N(0, \Sigma_L)$ truncated by -3 and 3 with spike-identity design matrix

$$\Sigma_Q = \begin{pmatrix} 1 & & 0.2 \\ & \ddots & \\ 0.2 & & 1 \end{pmatrix}$$

and another vector $L_{i,bernoulli}$ of length $d_L - \lfloor d_L/10 \rfloor$, where each element follows Bernoulli(0.5). The error term $\varepsilon_{ij} \sim i.i.d.$ Gumbel$(0,1)$ has mean 0.577 and variance 1.645. Since the high dimensions and parameters on the boundary enter similarly regardless of endogeneity, I assume that the price $P_{jt} \sim i.i.d.\ U[0,1]$ is exogenous for simplicity. To account for endogeneity[12], we can consider the design in Lu et al. (2023) and let $P_{jt} = Q_{jt} + \xi_{jt} + \epsilon_{jt}$, where $Q_{jt} \sim N(0,1)$ is the exogenous cost shifter, $\xi_{jt} \sim N(0, 0.5^2)$ is the unobserved quality, and $\epsilon_{jt} \sim N(0, 0.1^2)$ is the cost shock. I leave this for future research.

I consider $T = 10$ markets where each market has $n_t \in \{100, 200\}$ consumers making decisions among $J_t \in \{3, 5\}$ alternatives and one outside option ($j = 0$). Then, the total number of consumers $N = \sum_t n_t \in \{1000, 2000\}$, and the total number of alternatives $J = \sum_t J_t \in \{40, 60\}$ including the outside options. I choose the dimensions $d_L \in \{10, 200, 300\}$. The expected values of market shares are $(0.4174, 0.1961, 0.1913, 0.1953)$ for $J_t = 3$ and $(0.2722, 0.1465, 0.1396, 0.1504, 0.1442, 0.1471)$ for $J_t = 5$. The shares are decreasing as J_t increases and are at the same magnitude. Given the single random coefficient, Gauss-Hermite quadrature is more precise and efficient in computation than quasi Monte Carlo

[12] I am still running simulations for this case.

integration (see Appendix D). I choose $B = 100$ nodes to mitigate the error from numerical integration (cf. Train, 2000; Conlon and Gortmaker, 2020). In simulations, I find that larger B is necessary for CDML to be numerically stable because CDML is built upon the gradient containing $s_{ij}^{-1}(\cdot)$.

I compare four methods in 500 simulations: the traditional MLE (denoted as MLE later), the regularized MLE (RMLE) in Eq.(2.2), the CDML estimator (CDML) in Eq.(4.3) with the sample analog $\hat{\mu}^{RMLE}$ and the CDML estimator (CDML-Dantzig) with the dantzig selector $\hat{\mu}^{RMLE}$ in Eq.(4.2). Both MLE and RMLE estimate all parameters $\theta = (\bar{\beta}, \beta^x, \alpha, \Pi', \Sigma)'$ simultaneously, while CDML and CDML-Dantzig (CDMLs) only estimate the target parameter $\gamma = (\bar{\beta}, \beta^x, \alpha, \Pi_1, \Sigma)'$. The first-stage estimates in CDMLs are obtained through 10-fold cross-fitting to utilize more samples in each fold. The choice of the tuning parameter λ_N in RMLE and CDMLs is based on 10-fold cross-validations (see Appendix C).

5.2 Estimation

Table 5.1 summarizes the statistics for the estimators when the true parameter $\Sigma = 1 \in [0, \infty)$ is off the boundary. The first column displays the total dimensions and the ratios d_θ/N in brackets, and the second column indicates the estimation methods. The next five columns report the average biases across 500 simulations for five parameters, respective, while the last five columns present the square root of the mean square errors (RMSE). Starting with the upper panel, where the sample size $N = 1,000$ is small, all four methods perform similarly well when the dimension $d_\theta = 14$ is low. RMLE slightly outperforms the others in estimating β^x, as indicated by the smallest RMSE. However, as the dimension increases to $d_\theta = 204$, all methods are impaired and struggle to estimate Σ correctly, particularly MLE, which constantly estimates Σ as "zero". Although MLE has the smallest biases in estimating $\bar{\beta}$ and β^x, it exhibits large variances reflected in the RMSE. In contrast, RMLE estimates are biased but have smaller variances. CDMLs have moderate biases and

Table 5.1: Bias and RMSE of Target Parameter Estimates (Off Boundary)

d_θ	Method	Bias					RMSE				
		β_0	β_x	α	π_1	σ_x	β_0	β_x	α	π_1	σ_x
						$N = 1000$					
5	MLE	0.006	0.013	-0.014	0.009	-0.001	0.145	0.142	0.210	0.109	0.173
	RMLE	0.008	-0.027	0.005	-0.048	-0.042	0.145	0.138	0.207	0.112	0.172
	CDML	0.006	0.012	-0.014	0.009	-0.002	0.145	0.142	0.210	0.108	0.172
100	MLE	0.005	-0.002	-0.001	0.005	-0.277	0.148	0.845	0.201	0.112	0.336
	RMLE	0.011	-0.106	0.052	-0.157	-0.118	0.147	0.180	0.205	0.185	0.204
	CDML	-0.002	0.159	-0.084	0.095	0.289	0.167	0.659	0.283	0.218	0.561
200	MLE	0.023	-0.089	0.002	-0.003	-0.932	0.160	1.353	0.201	0.117	1.015
	RMLE	0.015	-0.119	0.046	-0.181	-0.131	0.154	0.186	0.197	0.202	0.215
	CDML	0.006	0.105	-0.079	0.060	0.200	0.165	0.575	0.258	0.161	0.434
300	MLE	0.037	0.109	-0.094	0.127	-1.407	0.163	2.049	0.242	0.193	1.409
	RMLE	0.019	-0.133	0.052	-0.190	-0.160	0.145	0.194	0.204	0.209	0.237
	CDML	0.005	0.088	-0.075	0.069	0.193	0.157	0.533	0.269	0.183	0.459
						$N = 2000$					
5	MLE	0.002	0.002	0.008	0.000	-0.008	0.109	0.103	0.145	0.073	0.115
	RMLE	0.002	-0.026	0.022	-0.040	-0.037	0.108	0.103	0.145	0.081	0.119
	CDML	0.002	0.002	0.007	0.000	-0.009	0.109	0.103	0.145	0.073	0.115
100	MLE	0.002	0.017	0.002	0.001	-0.135	0.097	0.580	0.140	0.075	0.182
	RMLE	0.006	-0.076	0.040	-0.110	-0.091	0.096	0.134	0.143	0.131	0.147
	CDML	-0.007	0.129	-0.054	0.061	0.187	0.113	0.548	0.195	0.141	0.378
200	MLE	0.004	0.023	0.006	-0.001	-0.301	0.111	0.906	0.146	0.075	0.332
	RMLE	0.007	-0.086	0.048	-0.127	-0.105	0.109	0.131	0.150	0.144	0.156
	CDML	-0.004	0.128	-0.061	0.064	0.195	0.125	0.600	0.215	0.151	0.416
300	MLE	0.005	-0.086	0.013	0.000	-0.542	0.106	1.095	0.143	0.080	0.591
	RMLE	0.005	-0.092	0.050	-0.132	-0.116	0.103	0.138	0.148	0.148	0.171
	CDML	-0.006	0.079	-0.055	0.065	0.186	0.118	0.557	0.199	0.159	0.416

Notes: In the first column d_θ, the natural numbers show the total dimension of the parameter $\theta = (\vec{\beta}, \beta^x, \alpha, \Pi', \Sigma)'$ and the brackets report the dimension-sample ratio d_θ/N. The second column reports four methods of estimation in our simulation study: the maximum likelihood estimation (MLE), the regularized MLE (RMLE), the constrained debiased machine learning (CDML) and the CDML with Dantzig selector (CDML-D). Columns 3 to 7 indicate the average biases $\frac{1}{S}\sum_{s=1}^{S}(\hat{\theta}_j^{(s)} - \theta_{j,0})$ of five estimators over $S = 500$ Monte Carlo simulations. The last five columns report the square root of mean square errors (RMSE) $\sqrt{\frac{1}{S}\sum_{s=1}^{S}(\hat{\theta}_j^{(s)} - \theta_{j,0})^2}$ for the estimators.

Table 5.2: Bias and RMSE of Target Parameter Estimates ($J = 30$, On Boundary)

Notes: In the first column d_θ, the natural numbers show the total dimension of the parameter $\theta = (\bar{\beta}, \beta^x, \alpha, \Pi', \Sigma)'$ and the brackets report the dimension-sample ratio d_θ/N. The second column reports four methods of estimation in our simulation study: the maximum likelihood estimation (MLE), the regularized MLE (RMLE), the constrained debiased machine learning (CDML) and the CDML with Dantzig selector (CDML-D). Columns 3 to 7 indicate the average biases $\frac{1}{S}\sum_{s=1}^{S}(\hat{\theta}_j^{(s)} - \theta_{j,0})$ of five estimators over $S = 500$ Monte Carlo simulations. The last five columns report the square root of mean square errors (RMSE) $\sqrt{\frac{1}{S}\sum_{s=1}^{S}(\hat{\theta}_j^{(s)} - \theta_{j,0})^2}$ for the estimators.

RMSEs across parameters, dominating MLE and RMLE. When $d_\theta = 304$, MLE estimates are strongly biased except for $\bar{\beta}$. RMLE and CDMLs, however, are robust to the increasing dimensions, and yield similar results to those for $d_\theta = 204$. The lower panel, where the sample size is increased to $N = 2,000$, reveals similar conclusions. All biases and RMSEs decreases as expected, suggesting convergence of the estimators. MLE is now able to detect non-zero Σ when $d_\theta = 204$, but it is still outperformed by RMLE and CDMLs. Table 5.1 also demonstrates that, even in the low-dimensional settings $d_\theta = 14$, sufficiently large sample sizes are necessary to achieve precise estimation.

Similar to Table 5.1, Table 5.2 reports the biases and RMSEs when the parameter $\Sigma = 0 \in [0, \infty)$ is on the boundary. Interestingly, inability to detect the random coefficient appears to be an advantage in this case. When the dimension $d_\theta = 14$ is low, the estimators from four methods are comparable. When $d_\theta = 204$, MLE and RMLE estimate Σ with great precision. However, both methods are severely biased in estimating other parameters except for the intercept $\bar{\beta}$. In contrast, CDMLs show some bias in estimating Σ but outperforms MLE and RMLE on the other parameters, and its performance improves further when $d_\theta = 304$. I also examine cases with more alternatives $J = 5$, and summarize the corresponding results in Table A.1 and A.2 for the readers' references.

Table 5.3: Coverage and Length of 95% Confidence Intervals (Off Boundary)

d_θ	Method	95% CI Coverage (%)					95% CI Length (Median)				
		β_0	β_x	α	π_1	σ_x	β_0	β_x	α	π_1	σ_x
						$N = 1000$					
5	MLE	96.0	94.8	94.2	94.0	93.2	0.570	0.538	0.761	0.404	0.318
	RMLE	96.0	93.8	94.4	91.0	93.0	0.566	0.524	0.754	0.387	0.311
	CDML	96.0	94.8	94.2	93.6	93.0	0.569	0.538	0.758	0.404	0.317
100	MLE	93.8	92.8	95.0	94.2	59.6	0.565	3.285	0.748	0.405	0.383
	RMLE	94.4	100.0	94	66.2	98.0	0.568	3.404	0.777	0.404	0.479
	CDML	94.6	91.0	96.6	98.0	97.2	0.596	1.938	0.901	0.624	0.734
200	MLE	92.4	90.2	93.8	90.8	11.0	0.559	4.333	0.751	0.387	0.426
	RMLE	91.6	100.0	89	51.6	99.8	0.547	4.722	0.704	0.326	1.040
	CDML	93.4	93.4	96.2	97.2	97.6	0.587	1.622	0.876	0.559	0.587
300	MLE	92.6	85.2	89.2	73.4	0.6	0.572	5.918	0.772	0.418	0.284
	RMLE	95.0	100.0	91.6	50.0	100.0	0.558	6.266	0.706	0.390	3.750
	CDML	95.0	90.8	95.6	98.0	97.2	0.593	1.573	0.878	0.587	0.626
						$N = 2000$					
5	MLE	94.0	95.0	94.0	95.0	95.4	0.399	0.383	0.540	0.286	0.223
	RMLE	94.2	92.6	93.6	88.8	94.4	0.398	0.376	0.537	0.276	0.219
	CDML	93.8	94.6	94.2	94.8	95.6	0.399	0.383	0.541	0.286	0.223
100	MLE	96.2	95.0	95.4	93.4	77.0	0.400	2.223	0.545	0.285	0.224
	RMLE	96.2	100.0	94	63.8	93.8	0.398	2.214	0.542	0.274	0.256
	CDML	97.0	92.4	97.4	98.2	99.2	0.418	1.579	0.633	0.409	0.486
200	MLE	93.6	92.0	94.0	93.0	30.4	0.399	3.120	0.538	0.284	0.227
	RMLE	95.0	100.0	94	59.8	97.6	0.402	3.212	0.559	0.290	0.336
	CDML	95.4	92.4	97.0	98.8	99.4	0.418	1.579	0.636	0.414	0.478
300	MLE	93.8	90.6	94.0	91.8	3.6	0.401	3.799	0.539	0.285	0.246
	RMLE	94.6	100.0	94.0	75.4	100.0	0.407	4.106	0.588	0.334	0.554
	CDML	94.6	91.8	96.2	98.2	97.4	0.418	1.423	0.630	0.409	0.457

Notes: In the first column d_θ, the natural numbers show the total dimension of the parameter $\theta = (\vec{\beta}, \beta^x, \alpha, \Pi', \Sigma)'$ and the brackets report the ratio d_θ/N. The second column reports four methods of estimation in our simulation study: the maximum likelihood estimation (MLE), the regularized MLE (RMLE), the constrained debiased machine learning (CDML) and the CDML with Dantzig selector (CDML-D). The 95% confidence intervals (CI) are constructed by the estimators plus and minus 1.96 multiplied by the estimated standard errors. The percentages are the counts that the true parameter is covered by the estimated CI divided by 500 (total number of simulations). The length of 95% CI is calculated as 3.92 multiplied by the estimated standard errors. To mitigate the influence of extreme estimates due to the small sample sizes, the median length is reported instead of the mean.

Table 5.4: Coverage and Length of 95% Confidence Intervals ($J = 30$, On Boundary)

Notes: In the first column d_θ, the natural numbers show the total dimension of the parameter $\theta = (\bar{\beta}, \beta^x, \alpha, \Pi', \Sigma)'$ and the brackets report the ratio d_θ/N. The second column reports four methods of estimation in our simulation study: the maximum likelihood estimation (MLE), the regularized MLE (RMLE), the constrained debiased machine learning (CDML) and the CDML with Dantzig selector (CDML-D). The 95% confidence intervals (CI) are constructed by the estimators plus and minus 1.96 multiplied by the estimated standard errors except for Σ, which is on the boundary. The 95% CI for Σ is constructed as $[\hat{\Sigma} - 1.64\widehat{SE(\Sigma)}, \hat{\Sigma}]$, similar to a one-sided t-test. The percentages are the counts that the true parameter is covered by the estimated CI divided by 500 (total number of simulations). The length of 95% CI is calculated as 3.92 multiplied by the estimated standard errors, except for Σ which is $1.64\widehat{SE(\Sigma)}$. To mitigate the influence of extreme estimates due to the small sample sizes, the median length is reported instead of the mean.

5.3 Inference

Statistical inference for the target parameters is also of interest. Table 5.3 presents the frequencies of coverage and the median[13] lengths of the 95% confidence intervals (CI) for the target parameters across four methods when $J = 3$ and $\Sigma = 1 \in [0, \infty)$. Note that I calculate the standard errors for RMLE in the same way as those in MLE, as the penalty term is shrinking after dividing by N, in order to show the consequence of misusing the standard errors. In the small-sample and small-dimensional setting with $N = 1,000$ and $d_\theta = 14$, four methods yield similar CI coverages and lengths, except for the parameter Σ. CDMLs show about 30% lower coverages for Σ with about 30% reduction in the length of CIs because they tend to under-estimate the standard errors for $\hat{\Sigma}$ (cf. Table 5.1), and the trend persists in large-dimensional and large-sample settings. When $d_\theta = 204$, inference based on RMLE is invalid due to the regularization bias in the estimation. MLE exhibits valid coverages for only β^x and α, but fails completely to cover Σ. CDMLs have good coverages for all parameters except Σ, which still dominates MLE and RMLE. In the high-dimensional case with $d_\theta = 304$, CDMLs surpass both MLE and RMLE and demonstrate their robustness. When the sample size increases to $N = 2,000$, MLE is greatly improved but still unable to cover Σ.

[13] I report the median to mitigate the influence of the extreme estimates, which can arise in small sample sizes.

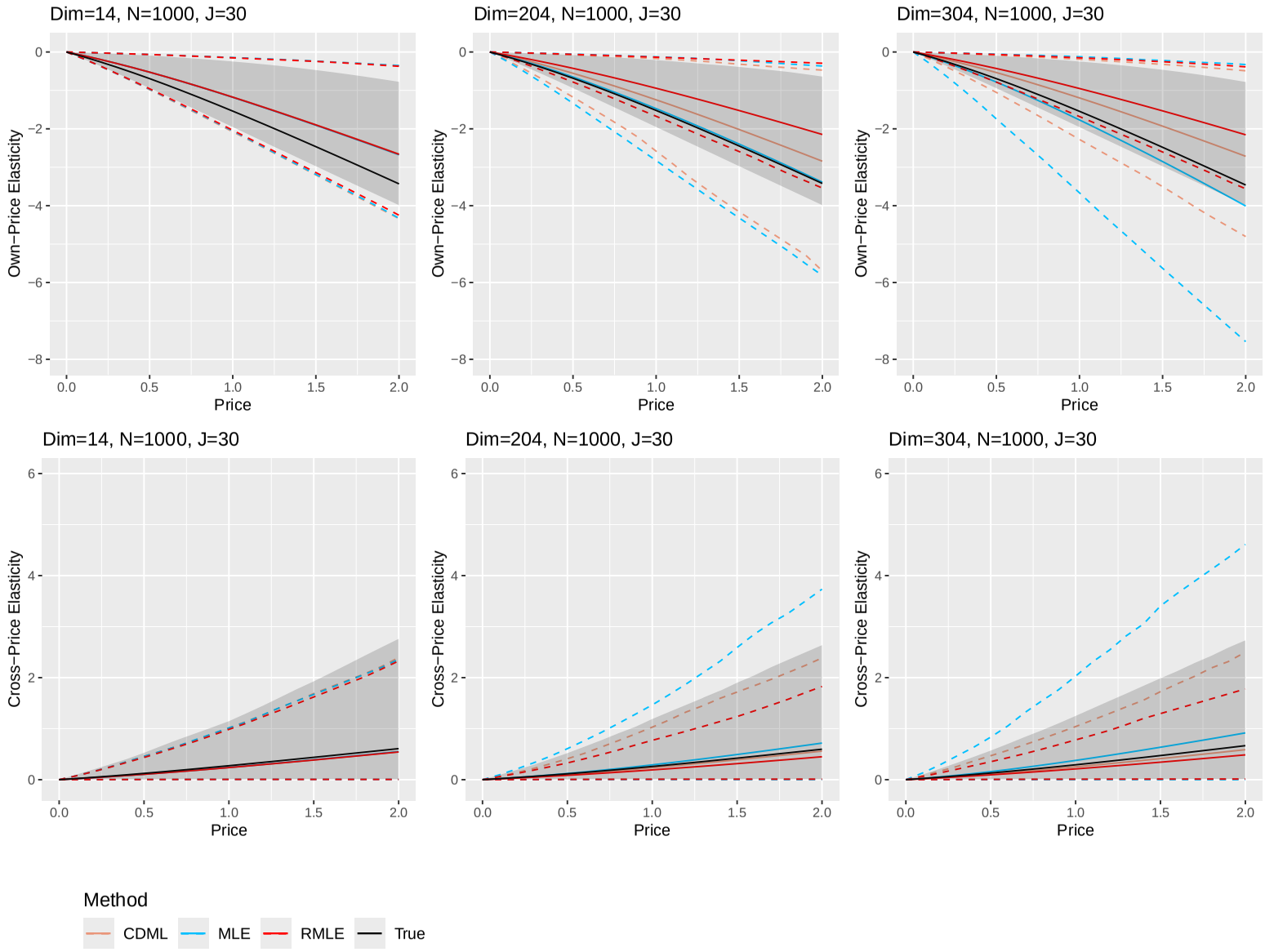
Table 5.4 reports the coverage rates and lengths when $\Sigma = 0$ is on the boundary and $J = 3$. Notably, the 95% CI for Σ is constructed based on the critical region of a one-sided t-test ($H_0 : \Sigma = 0$ versus $H_1 : \Sigma > 0$) at the 5% significance level, so the length of CI is 1.64 times the estimated standard error. The results show that CDMLs provide around 95% coverages for Σ when $d_\theta = 204$ and $d_\theta = 304$, whereas both MLE and RMLE have 100% coverage due to coincidence. Instead of adjusting the standard errors due to the boundary (see Example 2 in Section 4), the CIs for the unconstrained parameters are constructed as if they follow an asymptotic normal distribution. The table illustrates that both MLE and RMLE suffer from the increasing dimensions to which CDMLs are robust. The results for the cases where $J = 5$ are provided in Table A.3 and A.4.

5.4 Price Elasticities

Finally, I calculate the own-price and cross-price elasticities defined in Eq.(3.5), referring to the procedure in Lesellier et al. (2023): simulate 500 datasets, estimate the coefficients for four methods, solve the own-price elasticity for the alternative $j^* = 1$ in the market $t = 1$ and the cross-price elasticity for the alternative $j = 2$ with respect to the price p_{j^*}. I calculate the elasticities at an equally spaced grid of p_{j^*} from $[-2, 2]$ for the 500 datasets, and then summarize the mean, the 2.5% and 97.5% quantiles of the elasticities over the simulations in Figures 5.1, A.1, A.2 and A.3.

Figure 5.1 illustrates the price elasticities when $N = 1,000$, $J = 30$ and $\Sigma = 1$ (off the boundary). The upper three panels display the own-price elasticities and the lower three panels show the cross-price elasticities, where the dimension d_θ is 14, 204 and 304, respectively. The black solid lines represent the average elasticities given the true parameter θ_0, with the shaded areas indicating the 95% confidence intervals. The other solid lines correspond to the average estimated elasticities for the four methods, and the dashed lines represent the 2.5% and 97.5% quantiles for each method. In the low-dimensional case with $d_\theta = 14$ (i.e., the first column), all four methods provide similar means and quan-

Figure 5.1: Price Elasticities ($N = 1000$, $J = 30$, Off Boundary)



Notes: The upper three panels display the own-price elasticities and the lower three panels show the cross-price elasticities, where the dimension d_θ is 14, 204 and 304, respectively. The black solid lines represent the average elasticities given the true parameter θ_0 over 500 simulations, with the shaded areas indicating the 95% confidence intervals. The other solid lines correspond to the average estimated elasticities for the four methods (MLE, RMLE, CDML and CDML-Dantzig), and the dashed lines represent the 2.5% and 97.5% quantiles for each method.

tiles for the estimated elasticities. The own-price elasticities are overestimated with the means greater than the true value, and the cross-price elasticities are estimated precisely. When d_θ increases to 204, differences between the methods become more pronounced. For the own-price elasticities, MLE has the closest mean to the true value, but with the largest variance. RMLE has the smallest variance but suffers from a significant upward bias. CDMLs also exhibit larger variances than the true value, though CDML-Dantzig performs slightly better than CDML in this regard. For the cross-price elasticity, CDMLs outperform MLE and RMLE, where MLE shows too much variance and RMLE shows too little. In the high dimension with $d_\theta = 304$, MLE becomes less informative for both elasticities with its CIs nearly double the length of the true CIs. RMLE is robust to the increasing dimension but continues to be biased towards zero. CDMLs remain robust to the dimension, showing less bias than RMLE and smaller variances than MLE. Figures A.1, A.2 and A.3 suggest similar conclusions when the large sample size $N = 2,000$ and/or $\Sigma = 0$ is on the boundary.

6 Application

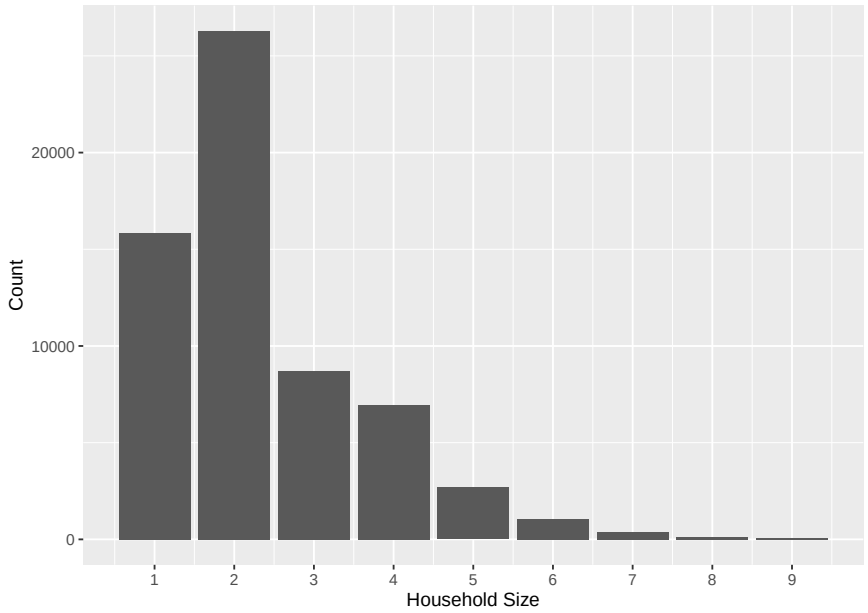
To demonstrate the effect of high-dimensional individual covariates and possibly diluted random coefficients, I estimate the demand for soft drinks in North Carolina in the year 2011 and calculate the own- and cross-price elasticities based on the estimates, using the micro-level datasets provided by NielsonIQ.

6.1 Data

Comprehensive marketing datasets, such as the NielsenIQ *Consumer Panel* dataset and the *Retailer Scanner* dataset, are accessible upon requests through the Kilts Center at the Chicago Booth School of Business. The consumer panels consist of representative[14] and

[14]The active panelists are projectable to the total United States using household projection factors, according to the Kilts Center website.

Figure 6.1: Household Sizes in the 2011 Consumer Panels



qualified households (40,000-60,000 every year) that continuously report their personal purchases using in-home scanners. As an example, in the panel year 2011, there were 62,092 households recorded in the consumer panel whose family sizes are depicted in Figure 6.1, where

approximately 25.5% of households comprise only one member and 42.3% of households comprise two. The retailer scanner data include weekly pricing, volume and store environments from over 90 participating retail chains with 35,000-50,000 participating stores. These two datasets provide researchers with valuable information, for example, (i) household demographic variables (e.g., income, family size, age, composition and living conditions), (ii) geographic variables related to households and stores (zip codes, region, etc.) and (iii) product characteristics (e.g., description, brand, pack size, flavor for food products). Moreover, each household is assigned a unique household ID (HHID) and each product is identified by a universal product code (UPC), enabling researchers to track the information on what products were purchased, when, and where.

In addition to the *Consumer Panel*, NielsonIQ offers two complementary survey datasets:

Annual Ailments, Health, and Wellness Survey (Ailment Survey) and *Custom Panel View Surveys (Custom Survey).* The ailment survey, started in 2011 and is updated annually, provides detailed information on panelists' health conditions covering heartburn, muscle pain, diabetes, cancer, heart disease, obesity, and other ailments. All household members aged thirteen and older are asked to report their experiences with thirty-four different ailments. For each selected ailment, the respondent is required to provide details on the timeline of diagnosis, the treatment (e.g., medications and/or exercises), and their most recent experience with the condition. The survey also includes the enrollment of health insurance and related health activities. The custom survey is available in the panel year 2008 and 2011. Each household member participating in the survey is asked about the age, gender, highest educational level (and their major if applicable), current employment, (sub-)occupation, categories of products the member has purchased or used, scientific questions corresponding to the categories (e.g., ingredient), attitude towards price, brand (store brands versus national brands) and quality as well as the role within the household.

Since all the four datasets are available only in 2011, I focus on the panel year 2011 to include individual information as much as possible. Importantly, both the ailment survey and the custom survey collect responses at the individual level via questionnaires, so they cannot be directly merged with the consumer panels. I provide my strategy of merging in Appendix B.

Although the datasets cover most states in the U.S., focusing on a single state not only simplifies the computation but also the choice of the instrumental variables. Following the literature (Allcott et al., 2019; Conlon and Gortmaker, 2023), I use contemporaneous prices in the states that are adjacent to the North Carolina (i.e., Virginia, Tennessee, and South Carolina) as a Hausman (1996)-type instrument for prices. Moreover, restricting the analysis to a small region helps define the markets clearly and ensures that most inside options are available to all consumers. In the 2011 Consumer Panel, 2,246 households were from North Carolina while 1,179 of them completed both surveys with full information. It

is possible that some households never consider purchasing soft drinks, and hence, are not in the soft-drink market. Therefore, I further restrict the sample to 1,136 households who had purchased soft drinks at least once in 2011 and regard them as in-market consumers.

By matching the HHIDs with the UPCs, 1,693 varieties of soft drinks (including diet drinks but excluding juice and fruit drinks) were sold during approximate 230,000 visits to NielsonIQ retailers. These soft drinks differed in brands, sizes, flavors and containers. However, it is unrealistic to assume that households have full access to and rationally compare all these numerous products before selecting the best one. In addition, Assumption 9 may fail in this context due to the tiny choice probabilities. To address these issues and provide reliable results, I calculate the market shares and retain only the top twenty brands by sales. Specifically, I classify different sizes of the same product (e.g., Regular Coke in 24 cans and 6 cans) as the same brand, while different varieties (e.g., Regular Coke, Diet Coke, and Caffeine Free Coke) are considered distinct brands. The remaining brands are grouped as outside options. The market shares s_{jt} are calculated quarterly for each quarter $t = 1, \ldots, 4$ of the panel year 2011. For each t, s_{jt} is equal to the total ounces of brand j purchased divided by the total ounces purchased in the market. There are eighty inside alternatives over the year with four outside options. The minimum market share $\min_{j,t} s_{jt}$ was approximately 1%, and the maximum market share $\max_{j,t} s_{jt}$ was about 13%. I also find that the top brands are widely recognized, and believe that they were fully accessible to the in-market consumers.

6.2 Model

The following model is used to analyze the markets of soft drinks. For each market (quarter) $t = 1, \ldots, 4$ and the market-specific twenty alternatives $j = 1, \ldots, 20$, the indirect

utility of a household $i = 1, \ldots, n_t$ in the market is defined as

$$
\begin{aligned}
U_{ijt} &= \delta_{jt} + \mu_{ijt} + \varepsilon_{ijt}, \quad \varepsilon_{ijt} \sim i.i.d. \text{ Type I EV} \\
\delta_{jt} &= \bar{\beta} + \text{Diet}_{jt}\beta^x + P_{jt}\alpha + \xi_{jt} \\
\mu_{ijt} &= (\text{Diet}_{jt} \otimes L_i)'\Pi + (\text{Diet}_{jt} \odot v_i)\Sigma, \quad v_i \sim i.i.d.\ N(0,1)
\end{aligned}
$$

where $\text{Diet}_{jt} = 1$ if the soft drink j in market t is diet and $\text{Diet}_{jt} = 0$ otherwise. P_{jt} is the price per ounce. I normalize the outside options $U_{i0t} = 0$ and introduce plentiful household-level characteristics in L_i. There are two settings for L_i representing varying levels of household detail:

- L_i^{Base} is a vector of 8 variables related to households' income, age, size, and race;
- L_i^{Sci}, in addition to L_i^{Base}, includes 674 variables related to the households' composition, scientific knowledge of daily products, and preferences on prices and brands.

For each setting, I estimate six models and compare the coefficients. I first assume exogenous price (so $\xi_{jt} = 0$) and estimate all parameters simultaneously with MLE, RMLE and CDML[15]. Next, I assume endogeneity (so $\xi_{jt} \neq 0$), estimate the mean utility δ_{jt} as well as the nonlinear parameters Π and Σ. Then I use two-stage least squares (2SLS) to recover the linear parameters $\bar{\beta}$, β^x and α. Particularly, given that $(\delta_{jt})_{j,t}$ has a large dimension $d_\delta = 80$, in CDML I only debias the coefficients for the household income (denoted as Π_1) and the standard deviation Σ of the diet dummy. Note that the estimated linear parameters are slightly different by RMLE and CDML in this case because of the 4-fold cross-fitting. Also, the price elasticities can be different due to the random coefficient. Given the single random coefficient, I use a Gauss-Hermite quadrature with $S = 100$ nodes to balance the accuracy of approximation and the cost of computation. The choice of tuning parameter λ_N is based on a 4-fold cross-validation over markets.

[15]CDML debiases the target parameters $\bar{\beta}$, β^x, α, Σ and the effect of household income Π_1, with a 4-fold cross-fitting.

Table 6.1: Estimated Coefficients for Baseline Models

Notes: Models in columns (1)-(3) assume that the prices of soft drinks are exogenous, while models in columns (4)-(6) account for the price endogeneity. The intercept, the coefficients for the diet dummy and price are directly estimated in column (1)-(3) and are recovered by two-stage least squares (2SLS) in columns (4)-(6). The row RC.Diet represents the estimated standard deviation for the random coefficient of the diet dummy. Standard errors of estimates are listed in parentheses. The standard errors of RMLE are infeasible except for those in Column (5) that are derived from 2SLS. The asterisks *, ** and *** represent the significant level at 10%, 5% and 1%. One-sided t-tests $H_0 : \Sigma = 0$ versus $H_1 : \Sigma > 0$ are conducted for the standard deviations.

6.3 Result

Table 6.1 presents the estimated target parameters and corresponding standard errors (only for MLE, CDML and 2SLS) from the baseline models using $L_i = L_i^{Base}$. The standard errors of RMLE are infeasible except for those in Column (5) that are derived from 2SLS. Columns (1)-(3) report estimates from models that assumes exogenous prices, while columns (4)-(6) account for price endogeneity and recover the intercept, the coefficients for the diet dummy and price by 2SLS. In columns (1) and (3), all coefficients are statistically significant at least at the 10% level. Particularly, the coefficient of price has a p-value below 0.1%, suggesting a strong price effect on soft-drink demand. The standard deviations of the random coefficients are significantly positive, which provide evidence of heterogeneous preferences for diet drinks among households. Columns (4)-(6) tell a different story and imply the presence of price endogeneity. The intercepts and the coefficients of price remain negative but become statistically insignificant. In fact, most coefficients in columns (4) and (5) are insignificant. CDML in column (6) reveals a significantly positive coefficient for the interaction term, suggesting that higher-income households may place more value on health-related attributes when purchasing soft drinks.

Incorporating information on scientific knowledge and shopping preferences of households, Table 6.2 presents the results when $L_i = L_i^{Sci}$. The estimates for the intercept and the price effect in columns (1)-(3), ignoring the price endogeneity, closely resemble those in columns (1)-(3) in Table 6.1. However, the coefficients for the diet dummy and the in-

Table 6.2: Estimated Coefficients for Models Including Scientific Knowledge and Shopping Preferences

Notes: Models in columns (1)-(3) assume that the prices of soft drinks are exogenous, while models in columns (4)-(6) account for the price endogeneity. The intercept, the coefficients for the diet dummy and price are directly estimated in column (1)-(3) and are recovered by two-stage least squares (2SLS) in columns (4)-(6). The row RC.Diet represents the estimated standard deviation for the random coefficient of the diet dummy. Standard errors of estimates are listed in parentheses (or in brackets if the estimated variance-covariance matrix is singular). The standard errors of RMLE are infeasible except for those in Column (5) that are derived from 2SLS. The asterisks *, ** and *** represent the significant level at 10%, 5% and 1%. One-sided t-tests $H_0 : \Sigma = 0$ versus $H_1 : \Sigma > 0$ are conducted for the standard deviations.

teraction term become smaller in magnitude compared to the basic models. Particularly, in columns (1) and (2), MLE and RMLE indicate no remaining unobserved heterogeneity given the tiny estimates. In contrast, CDML in column (3) suggests the presence of unobserved heterogeneity, as the estimated standard deviation is significantly positive. When accounting for endogeneity, MLE in column (4) fails to detect unobserved heterogeneity. Both RMLE in column (5) and CDML in column (6) detect the heterogeneity and report larger price effects than MLE.

Finally, I compare the own- and cross-price elasticities for the top 1 and top 2 selling brands in the year 2011, where the top 1 seller holds the largest market share among 21 alternatives and the top 2 seller follows. Tables 6.3 and 6.4 report the estimated price elasticities for six models. According to columns (1)-(3) in both tables, I observe that the magnitude of own-price elasticities generally increases after introducing additional covariates, while the cross-price elasticities decreases. As an example, the cross-price elasticity of the outside option with respect to the price of the top 1 brand declines from 0.090 to 0.056 in column (1). In columns (4)-(6), after controlling for endogeneity, the magnitude of both elasticities increases as the dimension increases. In Table 6.3, the estimates from MLE in column (4) and CDML in column (6) are quite similar. However, in Table 6.4, the CDML estimates in column (6) are around 40% larger than the maximum likelihood estimates in column (4), which are different from the regularized estimates in

Table 6.3: Price Elasticities for Baseline Models

Notes: Models in columns (1)-(3) assume that the prices of soft drinks are exogenous, while models in columns (4)-(6) account for the price endogeneity. Both own-price and cross-price elasticities are calculated based on estimated parameters. In Panel A, Top1 and Top2 represent the best seller brand and the second-best seller brand in the year 2011. In Panel B, A-B means the cross-price elasticity of alternative A with respect to the alternative B.

Table 6.4: Price Elasticities for Baseline Models

Notes: Models in columns (1)-(3) assume that the prices of soft drinks are exogenous, while models in columns (4)-(6) account for the price endogeneity. Both own-price and cross-price elasticities are calculated based on estimated parameters. In Panel A, Top1 and Top2 represent the best seller brand and the second-best seller brand in the year 2011. In Panel B, A-B means the cross-price elasticity of alternative A with respect to the alternative B.

column (5) as well. This underscores the importance of variable selection and debiasing when we includes high-dimensional covariates in the model, since failing to address these issues can lead to incorrect substitution patterns.

7 Conclusion

In this paper, I propose a framework of estimation and inference for random-coefficient logit models in the presence of high-dimensional individual-level covariates, bridging the literature of high-dimensional inference and inference on the boundary. For estimation, I suggest an l_1-regularized maximum likelihood estimation (RMLE) approach to select the high-dimensional covariates, which remains effective even in the presence of multi-collinearity and endogeneity. To address the non-differentiability of l_1-penalty, I propose a proximal gradient descent algorithm that accommodates box constraints. Under mild assumptions, I derive non-asymptotic probability bounds for the estimation errors in RMLE, indicating a slower convergence rate than the existing high-dimensional literature due to the increasing number of alternatives. For inference, I implement a procedure called K-fold

cross-fitting, and construct a constrained debiased machine learning (CDML) estimator based on the first-stage RMLE. I prove the root-n consistency for the CDML estimator and derive its asymptotic distribution, which is multivariate Gaussian if there is no boundary issue, and otherwise, is a projection of the multivariate Gaussian onto a polytope. I also propose a quasi-likelihood ratio (QLR) test for hypothesis testings. The performance of RMLE and CDML is illustrated in comprehensive Monte Carlo simulations. Finally, using micro-level data provided by NielsonIQ, I apply these approaches to analyze the soft-drink market in North Carolina.

References

Aitchison, J., and S. D. Silvey 1958."Maximum-Likelihood Estimation of Parameters Subject to Restraints." *The Annals of Mathematical Statistics*. 29 (3): 813–828, 10.1214/aoms/1177706538, Publisher: Institute of Mathematical Statistics.

Allcott, Hunt, Rebecca Diamond, Jean-Pierre Dubé, Jessie Handbury, Ilya Rahkovsky, and Molly Schnell 2019."Food Deserts and the Causes of Nutritional Inequality." *Q. J. Econ.*. 134 (4): 1793–1844, 10.1093/qje/qjz015.

Andrews, Donald W. K. 1999."Estimation When a Parameter is on a Boundary." *Econometrica*. 67 (6): 1341–1383, https://www.jstor.org/stable/2999564, Wiley, Econometric Society.

——— 2001."Testing When a Parameter is on the Boundary of the Maintained Hypothesis." *Econometrica*. 69 (3): 683–734, 10.1111/1468-0262.00210.

Armstrong, Timothy B, Michal Kolesár, and Soonwoo Kwon 2023."Bias-Aware Inference in Regularized Regression Models." *Working Paper*.

Babii, Andrii, Eric Ghysels, and Jonas Striaukas 2022."Machine Learning Time Series Regressions With an Application to Nowcasting." *Journal of Business & Economic Statistics*. 40 (3): 1094–1106, 10.1080/07350015.2021.1899933.

Beck, Amir, and Marc Teboulle 2009."A Fast Iterative Shrinkage-Thresholding Algorithm for Linear Inverse Problems." *SIAM Journal on Imaging Sciences*. 2 (1): 183–202, 10.1137/080716542.

Beck, Amir, and Luba Tetruashvili 2013."On the Convergence of Block Coordinate Descent Type Methods." *SIAM Journal on Optimization*. 23 (4): 2037–2060, 10.1137/120887679.

Belloni, Alexandre, Victor Chernozhukov, and Christian Hansen 2014."Inference on Treatment Effects after Selection among High-Dimensional Controls." *Rev. Econ. Stat.*. 81 (2): 608–650, 10.1093/restud/rdt044.

Berry, Steve, Oliver B. Linton, and Ariel Pakes 2004b."Limit Theorems for Estimating the Parameters of Differentiated Product Demand Systems." *Rev. Econ. Stat.*. 71 (3): 613–654, 10.1111/j.1467-937X.2004.00298.x.

Berry, Steven, James Levinsohn, and Ariel Pakes 1995."Automobile Prices in Market Equilibrium." *Econometrica*. 63 (4): 841–890, 10.2307/2171802, Wiley, Econometric Society.

——— 2004a."Differentiated Products Demand Systems from a Combination of Micro and Macro Data: The New Car Market." *J. Polit. Econ.*. 112 (1): 68–105, 10.1086/379939, Publisher: The University of Chicago Press.

Bickel, Peter J., Ya'acov Ritov, and Alexandre B. Tsybakov 2009."Simultaneous analysis of Lasso and Dantzig selector." *The Annals of Statistics*. 37 (4): 1705–1732, 10.1214/08-AOS620, Publisher: Institute of Mathematical Statistics.

Bilodeau, Blair, Yanbo Tang, and Alex Stringer 2023."On the tightness of the Laplace approximation for statistical inference." *Statistics & Probability Letters*. 198, 109839, 10.1016/j.spl.2023.109839.

Bolker, Benjamin M., Mollie E. Brooks, Connie J. Clark, Shane W. Geange, John R. Poulsen, M. Henry H. Stevens, and Jada-Simone S. White 2009."Generalized linear mixed models: a practical guide for ecology and evolution." *Trends in Ecology & Evolution*. 24 (3): 127–135, 10.1016/j.tree.2008.10.008.

Breiman, L., and D. Freedman 1983."How Many Variables Should be Entered in a Regression Equation?" *J. Am. Stat. Assoc.*. 78 (381): 131–136, 10.2307/2287119, American Statistical Association, Taylor & Francis, Ltd.

Breslow, N. E., and D. G. Clayton 1993."Approximate Inference in Generalized Linear Mixed Models." *J. Am. Stat. Assoc.*. 88 (421): 9–25, 10.2307/2290687, American Statistical Association, Taylor & Francis, Ltd.

Breusch, T. S., and A. R. Pagan 1979."A Simple Test for Heteroscedasticity and Random Coefficient Variation." *Econometrica*. 47 (5): 1287–1294, 10.2307/1911963, Wiley, Econometric Society.

Candes, Emmanuel, and Terence Tao 2007."The Dantzig selector: Statistical estimation when p is much larger than n." *The Annals of Statistics*. 35 (6): 2313–2351, 10.1214/009053606000001523, Publisher: Institute of Mathematical Statistics.

Cattaneo, Matias D, Michael Jansson, and Xinwei Ma 2019."Two-Step Estimation and Inference with Possibly Many Included Covariates." *Rev. Econ. Stat.*. 86 (3): 1095–1122, 10.1093/restud/rdy053.

Chamberlain, Gary, and Guido Imbens 2004."Random Effects Estimators with many Instrumental Variables." *Econometrica*. 72 (1): 295–306, 10.1111/j.1468-0262.2004.00485.x.

Chernozhukov, Victor, Denis Chetverikov, Mert Demirer, Esther Duflo, Christian Hansen, Whitney Newey, and James Robins 2018."Double/debiased machine learning for treatment and structural parameters." *The Econometrics Journal*. 21 (1): C1–C68, 10.1111/ectj.12097.

Conlon, Christopher, and Jeff Gortmaker 2020."Best practices for differentiated products demand estimation with PyBLP." *Rand. J. Econ.*. 51 (4): 1108–1161, 10.1111/1756-2171.12352.

——— 2023."Incorporating Micro Data into Differentiated Products Demand Estimation with PyBLP." *Working Paper*.

Dempster, A. P., N. M. Laird, and D. B. Rubin 1977."Maximum Likelihood from Incomplete Data Via the *EM* Algorithm." *Journal of the Royal Statistical Society Series B: Statistical Methodology*. 39 (1): 1–22, 10.1111/j.2517-6161.1977.tb01600.x.

Dubois, Pierre, Rachel Griffith, and Martin O'Connell 2018."The Effects of Banning Advertising in Junk Food Markets." *Rev. Econ. Stat.*. 85 (1): 396–436, 10.1093/restud/rdx025.

Ericson, Keith Marzilli, and Amanda Starc 2012."Heuristics and Heterogeneity in Health Insurance Exchanges: Evidence from the Massachusetts Connector." *Am. Econ. Rev.*. 102 (3): 493–497, 10.1257/aer.102.3.493.

Fan, Yanqin, and Xuetao Shi 2023."Wald, QLR, and score tests when parameters are subject to linear inequality constraints." *J. Econometrics*. 235 (2): 2005–2026, 10.1016/j.jeconom.2023.02.009.

Friedman, Jerome, Trevor Hastie, Holger Höfling, and Robert Tibshirani 2007."Pathwise coordinate optimization." *The Annals of Applied Statistics*. 1 (2): 302–332, 10.1214/07-AOAS131, Publisher: Institute of Mathematical Statistics.

Gillen, Benjamin J, Sergio Montero, Hyungsik Roger Moon, and Matthew Shum 2019."BLP-2LASSO for aggregate discrete choice models with rich covariates." *The Econometrics Journal*. 22 (3): 262–281, 10.1093/ectj/utz010.

Goolsbee, Austan, and Amil Petrin 2004."The Consumer Gains from Direct Broadcast Satellites and the Competition with Cable TV." *Econometrica*. 72 (2): 351–381, https://www.jstor.org/stable/3598906, Wiley, Econometric Society.

Gowrisankaran, Gautam, Aviv Nevo, and Robert Town 2015."Mergers When Prices Are Negotiated: Evidence from the Hospital Industry." *Am. Econ. Rev.*. 105 (1): 172–203, 10.1257/aer.20130223.

Groll, Andreas, and Gerhard Tutz 2014."Variable selection for generalized linear mixed models by L1-penalized estimation." *Statistics and Computing*. 24 (2): 137–154, 10.1007/s11222-012-9359-z.

Hall, Jane, Denzil G. Fiebig, Madeleine T. King, Ishrat Hossain, and Jordan J. Louviere 2006."What influences participation in genetic carrier testing?: Results from a discrete choice experiment." *J. Health Econ.*. 25 (3): 520–537, 10.1016/j.jhealeco.2005.09.002.

Hausman, Jerry A. 1996."Valuation of New Goods under Perfect and Imperfect Competition." In *The Economics of New Goods*: University of Chicago Press, 207–248, https://www.nber.org/books-and-chapters/economics-new-goods/valuation-new-goods-under-perfect-and-imperfect-competition.

Hess, Stephane, Michel Bierlaire, and John W. Polak 2005."Estimation of value of travel-time savings using mixed logit models." *Transportation Research Part A: Policy and Practice*. 39 (2): 221–236, 10.1016/j.tra.2004.09.007.

Hildreth, Clifford, and James P. Houck 1968."Some Estimators for a Linear Model With Random Coefficients." *J. Am. Stat. Assoc.*. 63 (322): 584–595, 10.1080/01621459.1968.11009277.

Ho, Katherine 2006."The welfare effects of restricted hospital choice in the US medical care market." *J. Appl. Econom.*. 21 (7): 1039–1079, 10.1002/jae.896.

Hoerl, Arthur E., and Robert W. Kennard 1970."Ridge Regression: Biased Estimation for Nonorthogonal Problems." *Technometrics*. 12 (1): 55–67, 10.2307/1267351, Taylor & Francis, Ltd., American Statistical Association, American Society for Quality.

Hole, Arne Risa, and Julie Riise Kolstad 2012."Mixed logit estimation of willingness to pay distributions: a comparison of models in preference and WTP space using data from a health-related choice experiment." *Empirical Economics*. 42 (2): 445–469, 10.1007/s00181-011-0500-1.

Horowitz, Joel L., and Lars Nesheim 2021."Using penalized likelihood to select parameters in a random coefficients multinomial logit model." *J. Econometrics*. 222 (1, Part A): 44–55, 10.1016/j.jeconom.2019.11.008.

Hsiao, Cheng 1975."Some Estimation Methods for a Random Coefficient Model." *Econometrica*. 43 (2): 305–325, 10.2307/1913588, Wiley, Econometric Society.

Kennedy, Edward H. 2023."Semiparametric doubly robust targeted double machine learning: a review." January, 10.48550/arXiv.2203.06469, arXiv:2203.06469 [stat].

Ketz, Philipp 2018."Subvector inference when the true parameter vector may be near or at the boundary." *J. Econometrics*. 207 (2): 285–306, 10.1016/j.jeconom.2018.08.003.

——— 2019."On asymptotic size distortions in the random coefficients logit model." *J. Econometrics*. 212 (2): 413–432, 10.1016/j.jeconom.2019.02.008.

Léon, Gianmarco, and Edward Miguel 2017."Risky Transportation Choices and the Value of a Statistical Life." *Am. Econ. J. Appl. Econ.*. 9 (1): 202–228, 10.1257/app.20160140.

Lesellier, Max, Hippolyte Boucher, and Gokce Gokkoca 2023."Testing and Relaxing Distributional Assumptions on Random Coefficients in Demand Models." *Working Paper*.

Li, Huan, and Zhouchen Lin 2015."Accelerated Proximal Gradient Methods for Nonconvex Programming." In *Advances in Neural Information Processing Systems* 28: Curran Associates, Inc., https://papers.nips.cc/paper_files/paper/2015/hash/f7664060cc52bc6f3d620bcedc94a4b6-Abstract.html.

Li, Tianqi 2024."High-dimensional Inference when the True Parameter is on the Boundary."

Liu, Qing, and Donald A. Pierce 1994."A note on Gauss-Hermite quadrature." *Biometrika*. 81 (3): 624–629, 10.1093/biomet/81.3.624.

Lu, Zhentong, Xiaoxia Shi, and Jing Tao 2023."Semi-nonparametric estimation of random coefficients logit model for aggregate demand." *J. Econometrics*. 235 (2): 2245–2265, 10.1016/j.jeconom.2022.10.011.

McFadden, Daniel, and Kenneth Train 2000."Mixed MNL models for discrete response." *J. Appl. Econom.*. 15 (5): 447–470, 10.1002/1099-1255(200009/10)15:5<447::AID-JAE570>3.0.CO;2-1.

Meier, Lukas, Sara Van De Geer, and Peter Bühlmann 2008."The group lasso for logistic regression." *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*. 70 (1): 53–71, 10.1111/j.1467-9868.2007.00627.x.

Negahban, Sahand N., Pradeep Ravikumar, Martin J. Wainwright, and Bin Yu 2012."A Unified Framework for High-Dimensional Analysis of M-Estimators with Decomposable Regularizers." *Statistical Science*. 27 (4): 538–557, 10.1214/12-STS400, Publisher: Institute of Mathematical Statistics.

Nevo, Aviv 2000a."Mergers with Differentiated Products: The Case of the Ready-to-Eat Cereal Industry." *Rand. J. Econ.*. 31 (3): 395–421, 10.2307/2600994, RAND Corporation, Wiley.

——— 2000b."A Practitioner's Guide to Estimation of Random-Coefficients Logit Models of Demand." *Journal of Economics & Management Strategy*. 9 (4): 513–548, 10.1111/j.1430-9134.2000.00513.x.

Newey, Whitney K., and Daniel McFadden 1994."Chapter 36 Large sample estimation and hypothesis testing." In *Handbook of Econometrics* 4: Elsevier, 2111–2245, 10.1016/S1573-4412(05)80005-4.

Ning, Yang, and Han Liu 2017."A general theory of hypothesis tests and confidence regions for sparse high dimensional models." *The Annals of Statistics*. 45 (1): 158–195, 10.1214/16-AOS1448, Publisher: Institute of Mathematical Statistics.

Nurski, Laura, and Frank Verboven 2016."Exclusive Dealing as a Barrier to Entry? Evidence from Automobiles." *Rev. Econ. Stat.*. 83 (3): 1156–1188, 10.1093/restud/rdw002.

Petrin, Amil 2002."Quantifying the Benefits of New Products: The Case of the Minivan." *J. Polit. Econ.*. 110 (4): 705–729, 10.1086/340779, Publisher: The University of Chicago Press.

Raudenbush, Stephen W., Meng-Li Yang, and Matheos Yosef 2000."Maximum Likelihood for Generalized Linear Models with Nested Random Effects via High-Order, Multivariate Laplace Approximation." *Journal of Computational and Graphical Statistics*. 9 (1): 141–157, 10.2307/1390617, American Statistical Association, Taylor & Francis, Ltd., Institute of Mathematical Statistics, Interface Foundation of America.

Schelldorfer, Jürg, Lukas Meier, and Peter Bühlmann 2014."GLMMLasso: An Algorithm for High-Dimensional Generalized Linear Mixed Models Using l_1-Penalization." *Journal of Computational and Graphical Statistics*. 23 (2): 460–477, https://www.jstor.org/stable/43305738, American Statistical Association, Taylor & Francis, Ltd., Institute of Mathematical Statistics, Interface Foundation of America.

Self, Steven G., and Kung-Yee Liang 1987."Asymptotic Properties of Maximum Likelihood Estimators and Likelihood Ratio Tests Under Nonstandard Conditions." *J. Am. Stat. Assoc.*. 82 (398): 605–610, 10.2307/2289471, American Statistical Association, Taylor & Francis, Ltd.

Sen, Bodhisattva 2022."A Gentle Introduction to Empirical Process Theory and Applications." *Working Paper*.

Stram, Daniel O., and Jae Won Lee 1994."Variance Components Testing in the Longitudinal Mixed Effects Model." *Biometrics*. 50 (4): , 1171, 10.2307/2533455.

Tibshirani, Robert 1996."Regression Shrinkage and Selection via the Lasso." *Journal of the Royal Statistical Society. Series B (Methodological)*. 58 (1): 267–288, https://www.jstor.org/stable/2346178, Royal Statistical Society, Oxford University Press.

Tibshirani, Ryan J., and Jonathan Taylor 2011."The solution path of the generalized lasso." *The Annals of Statistics*. 39 (3): 1335–1371, 10.1214/11-AOS878, Publisher: Institute of Mathematical Statistics.

Train, Kenneth 2000."Halton Sequences for Mixed Logit." *Department of Economics, Working Paper Series*. https://ideas.repec.org//p/cdl/econwp/qt6zs694tp.html, Number: qt6zs694tp Publisher: Department of Economics, Institute for Business and Economic Research, UC Berkeley.

——— 2002. *Discrete Choice Methods with Simulation*.

Train, Kenneth E. 2009."Endogeneity." In *Discrete Choice Methods with Simulation*, Cambridge: Cambridge University Press, 2nd edition 315–346, 10.1017/CBO9780511805271.013.

Train, Kenneth E., and Clifford Winston 2007."Vehicle Choice Behavior and the Declining Market Share of U.s. Automakers." *Int. Econ. Rev.*. 48 (4): 1469–1496, 10.1111/j.1468-2354.2007.00471.x.

Tuerlinckx, Francis, Frank Rijmen, Geert Verbeke, and Paul De Boeck 2006."Statistical inference in generalized linear mixed models: A review." *British Journal of Mathematical and Statistical Psychology*. 59 (2): 225–255, 10.1348/000711005X79857.

Van Der Vaart, Aad W., and Jon A. Wellner 1996. *Weak Convergence and Empirical Processes*. , Springer Series in Statistics, New York, NY: Springer, 10.1007/978-1-4757-2545-2.

Verbeke, Geert, and Geert Molenberghs 2003."The Use of Score Tests for Inference on Variance Components." *Biometrics*. 59 (2): 254–262, 10.1111/1541-0420.00032.

Vershynin, Roman 2018. *High-Dimensional Probability: An Introduction with Applications in Data Science*. , Cambridge Series in Statistical and Probabilistic Mathematics, Cambridge: Cambridge University Press, 10.1017/9781108231596.

Yuan, Ming, and Yi Lin 2006."Model Selection and Estimation in Regression with Grouped Variables." *Journal of the Royal Statistical Society Series B: Statistical Methodology*. 68 (1): 49–67, 10.1111/j.1467-9868.2005.00532.x.

Zou, Hui 2006."The Adaptive Lasso and Its Oracle Properties." *J. Am. Stat. Assoc.*. 101 (476): 1418–1429, 10.1198/016214506000000735.

A Figures and Tables

Table A.1: Bias and RMSE of Target Parameter Estimates ($J = 50$, Off Boundary)

Notes: In the first column d_θ, the natural numbers show the total dimension of the parameter $\theta = (\vec{\beta}, \beta^x, \alpha, \Pi', \Sigma)'$ and the brackets report the dimension-sample ratio d_θ/N. The second column reports four methods of estimation in our simulation study: the maximum likelihood estimation (MLE), the regularized MLE (RMLE), the constrained debiased machine learning (CDML) and the CDML with Dantzig selector (CDML-D). Columns 3 to 7 indicate the average biases $\frac{1}{S}\sum_{s=1}^{S}(\hat{\theta}_j^{(s)} - \theta_{j,0})$ of five estimators over $S = 500$ Monte Carlo simulations. The last five columns report the square root of mean square errors (RMSE) $\sqrt{\frac{1}{S}\sum_{s=1}^{S}(\hat{\theta}_j^{(s)} - \theta_{j,0})^2}$ for the estimators.

Table A.2: Bias and RMSE of Target Parameter Estimates ($J = 50$, On Boundary)

Notes: In the first column d_θ, the natural numbers show the total dimension of the parameter $\theta = (\vec{\beta}, \beta^x, \alpha, \Pi', \Sigma)'$ and the brackets report the dimension-sample ratio d_θ/N. The second column reports four methods of estimation in our simulation study: the maximum likelihood estimation (MLE), the regularized MLE (RMLE), the constrained debiased machine learning (CDML) and the CDML with Dantzig selector (CDML-D). Columns 3 to 7 indicate the average biases $\frac{1}{S}\sum_{s=1}^{S}(\hat{\theta}_j^{(s)} - \theta_{j,0})$ of five estimators over $S = 500$ Monte Carlo simulations. The last five columns report the square root of mean square errors (RMSE) $\sqrt{\frac{1}{S}\sum_{s=1}^{S}(\hat{\theta}_j^{(s)} - \theta_{j,0})^2}$ for the estimators.

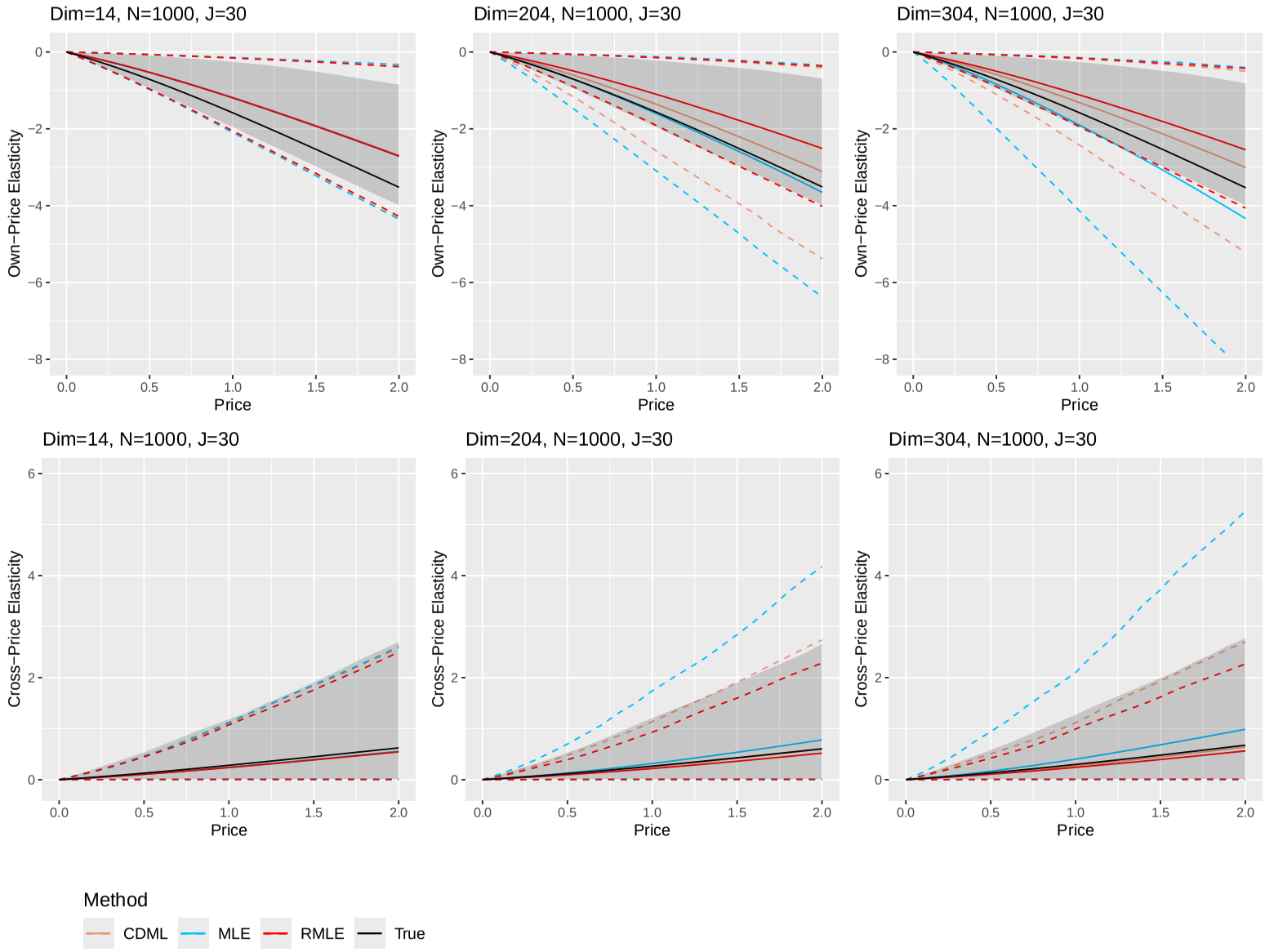
Table A.3: Coverage and Length of 95% Confidence Intervals ($J = 50$, Off Boundary)

Notes: In the first column d_θ, the natural numbers show the total dimension of the parameter $\theta = (\vec{\beta}, \beta^x, \alpha, \Pi', \Sigma)'$ and the brackets report the ratio d_θ/N. The second column reports four methods of estimation in our simulation study: the maximum likelihood estimation (MLE), the regularized MLE (RMLE), the constrained debiased machine learning (CDML) and the CDML with Dantzig selector (CDML-D). The 95% confidence intervals (CI) are constructed by the estimators plus and minus 1.96 multiplied by the estimated standard errors. The percentages are the counts that the true parameter is covered by the estimated CI divided by 500 (total number of simulations). The length of 95% CI is calculated as 3.92 multiplied by the estimated standard errors. To mitigate the influence of extreme estimates due to the small sample sizes, the median length is reported instead of the mean.

Table A.4: Coverage and Length of 95% Confidence Intervals ($J = 50$, On Boundary)

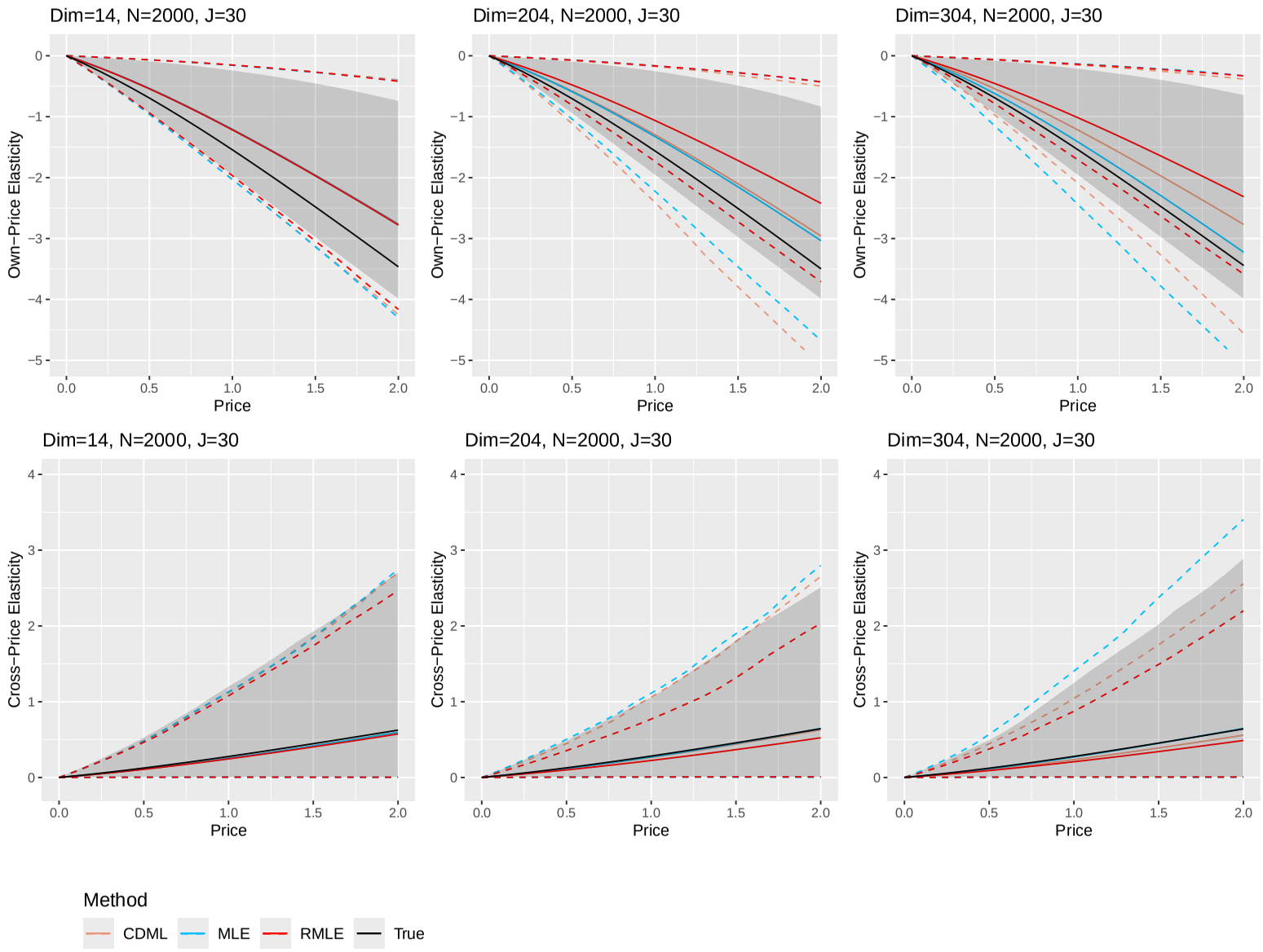
Notes:In the first column d_θ, the natural numbers show the total dimension of the parameter $\theta = (\bar{\beta}, \beta^x, \alpha, \Pi', \Sigma)'$ and the brackets report the ratio d_θ/N. The second column reports four methods of estimation in our simulation study: the maximum likelihood estimation (MLE), the regularized MLE (RMLE), the constrained debiased machine learning (CDML) and the CDML with Dantzig selector (CDML-D). The 95% confidence intervals (CI) are constructed by the estimators plus and minus 1.96 multiplied by the estimated standard errors except for Σ, which is on the boundary. The 95% CI for Σ is constructed as $[\hat{\sigma}_x - 1.64\widehat{SE(\Sigma)}, \hat{\Sigma}]$, similar to a one-sided t-test. The percentages are the counts that the true parameter is covered by the estimated CI divided by 500 (total number of simulations). The length of 95% CI is calculated as 3.92 multiplied by the estimated standard errors, except for Σ which is $1.64\widehat{SE(\Sigma)}$. To mitigate the influence of extreme estimates due to the small sample sizes, the median length is reported instead of the mean.

Figure A.1: Price Elasticities ($N = 1000$, $J = 3$, On Boundary)



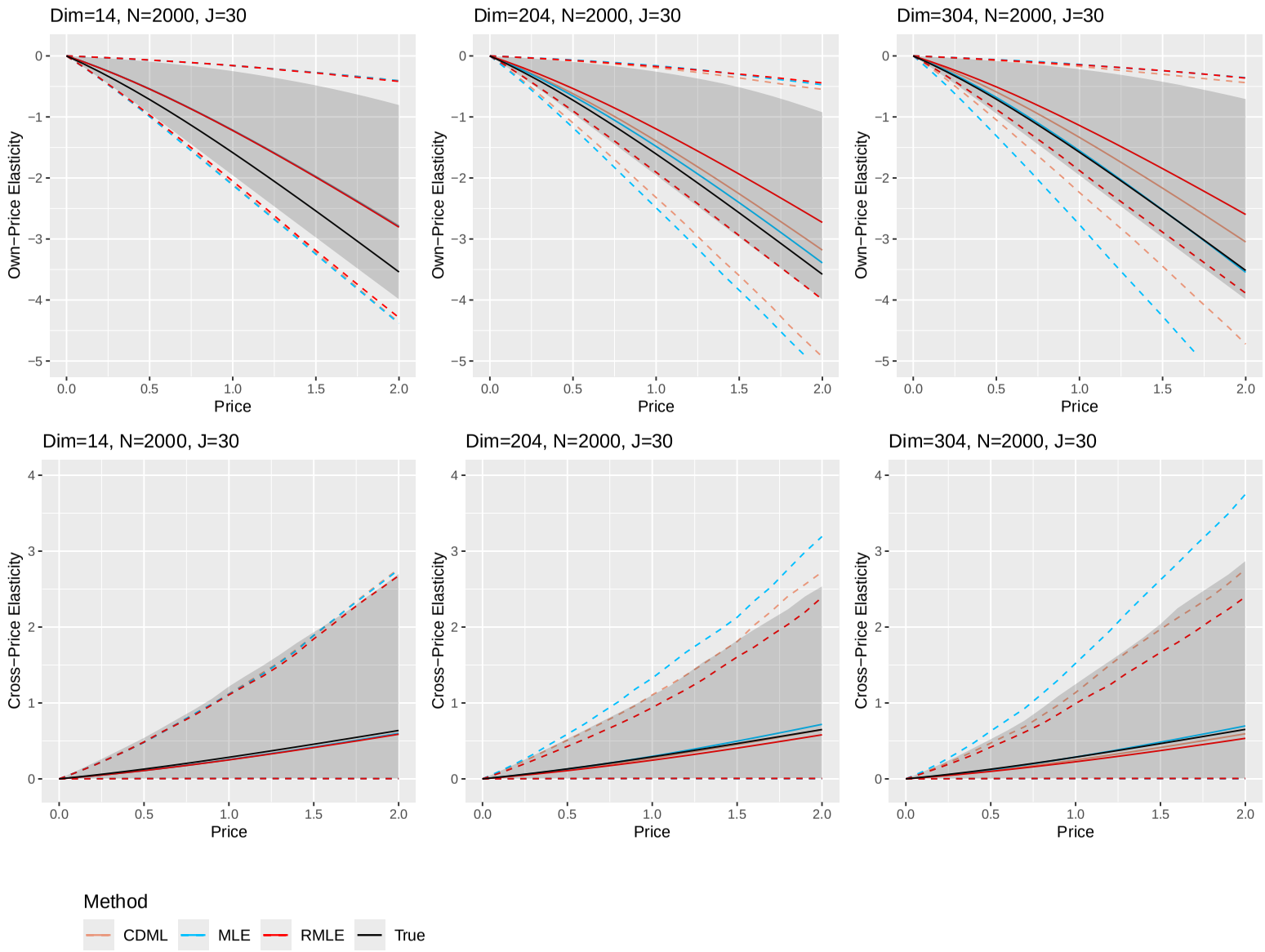
Notes: The upper three panels display the own-price elasticities and the lower three panels show the cross-price elasticities, where the dimension d_θ is 14, 204 and 304, respectively. The black solid lines represent the average elasticities given the true parameter θ_0 over 500 simulations, with the shaded areas indicating the 95% confidence intervals. The other solid lines correspond to the average estimated elasticities for the four methods (MLE, RMLE, CDML and CDML-Dantzig), and the dashed lines represent the 2.5% and 97.5% quantiles for each method.

Figure A.2: Price Elasticities ($N = 2000$, $J = 30$, Off Boundary)



Notes: The upper three panels display the own-price elasticities and the lower three panels show the cross-price elasticities, where the dimension d_θ is 14, 204 and 304, respectively. The black solid lines represent the average elasticities given the true parameter θ_0 over 500 simulations, with the shaded areas indicating the 95% confidence intervals. The other solid lines correspond to the average estimated elasticities for the four methods (MLE, RMLE, CDML and CDML-Dantzig), and the dashed lines represent the 2.5% and 97.5% quantiles for each method.

Figure A.3: Price Elasticities ($N = 2000$, $J = 30$, On Boundary)



Notes: The upper three panels display the own-price elasticities and the lower three panels show the cross-price elasticities, where the dimension d_θ is 14, 204 and 304, respectively. The black solid lines represent the average elasticities given the true parameter θ_0 over 500 simulations, with the shaded areas indicating the 95% confidence intervals. The other solid lines correspond to the average estimated elasticities for the four methods (MLE, RMLE, CDML and CDML-Dantzig), and the
dashed lines represent the 2.5% and 97.5% quantiles for each method.

B Dataset Construction

In the 2011 Custom Survey, there are 80,205 household members. Within each household, only one individual can be designated as the primary shopper. Given the likelihood that household members share information and benefit from collective knowledge, for questions related to knowledge (e.g., know-or-not questions), a household is treated as "knowing" if at least one member answers "know". As an example, consider a household of three members. If two members answer that they "know" brand A of a headache reliever, then the dummy $know.reliever.A$ is set to true. A continuing example is about the ingredient of the reliever A. If one member chooses Aspirin and the other chooses Ibuprofen, but the true answer is Aspirin, then the variable $reliever.A.correct$ is equal to 0.5. For attitude-related questions, I calculate the average responses across members within each household. For all other questions, I pick the decision of the primary shopper. Following this strategy, the survey data is reduced to 56,258 households with 404 variables.

There were 109,036 household members participating in the 2011 Ailment Survey. Intuitively, households with more members experiencing ailments would likely have higher demands for health-related products. To capture this, I count the number of household members who reported ailments within a household for every ailment. After generating dummy variables for each ailment and aggregating the data at household-level, I obtain a dataset consisting of 62,002 households and 1,203 variables.

C K-fold Cross-validation

Choice of tuning parameter (i.e., the penalty λ_N) is crucial to the regularized maximum likelihood method. Larger penalty can better capture the sparsity of the true parameter, however, introduce more regularization bias in the estimation. In the machine learning literature, the procedure called K-fold cross-validation is (K-CV) widely applied to determine tuning parameters. I borrow the idea of K-CV and implement the following procedure in my simulations and empirical application.

K-CV for Regularized Maximum Likelihood Estimation

1. For each $t = 1, \ldots, T$, consider a random partition of $\mathcal{I}_t := \{1, \ldots, n_t\}$ into K different segments $\mathcal{I}_{1,t}, \ldots, \mathcal{I}_{K,t}$;
2. Let $Test_k := \cup_{t=1}^{T} \mathcal{I}_{k,t}$ be the collection of the indices of individuals in all T markets for the segment $k = 1, \ldots, K$, and let $Train_k := \cup_{t=1}^{T} \mathcal{I}_t \backslash Test_k$ be the indices of the rest of individuals;
3. Let $\Lambda := \{\lambda_1, \ldots, \lambda_M\}$ be a sequence of penalty parameters;
4. For each $m = 1, \ldots, M$ and $k = 1, \ldots, K$,
   (a) Solve the regularized likelihood estimator $\hat{\theta}^{RMLE}_{k,m} = \hat{\theta}^{RMLE}(Train_k, \lambda_m)$ based on the individuals indexed by the training set $Train_k$ and the penalty $\lambda_m > 0$;
   (b) Calculate the unpenalized log-likelihood $L_{NJT}(Test_k, \hat{\theta}^{RMLE}_{k,m})$ based on the individuals indexed by the testing set $Test_k$ and the estimator $\hat{\theta}^{RMLE}_{k,m}$;
5. The searched optimal tuning parameter λ_{opt} minimizes the loss, which is negative testing log-likelihood,

$$\lambda_{opt} = \arg\min_{m=1,\ldots,M} - \sum_{k=1}^{K} L_{NJT}(Test_k, \hat{\theta}^{RMLE}_{k,m})$$

Choice of Λ My choice of Λ is according to Lemma 1. Although the corollary indicates $\rho_N = \max_t J_t \sqrt{N^{-1} \log d_\theta}$ and $\lambda_N \geq 2N\rho_N$, the penalty λ_N could be overwhelming whenever the sample size N is not large enough. However, the rate $N\rho_N$ may be informative. I consider 20 different penalties, which are constants multiplied by $N\rho_N$, when I do the $K = 10$ folds cross-validation:

$$\Lambda = (1, 2, 3, 4, 5, 6, 7, 8, 9) \otimes (10^{-3}, 10^{-2}, 10^{-1}) \times N\rho_N$$

I only penalize the high-dimensional and potentially sparse parameter Π, presuming that the other parameters are known to be non-sparse and important.

Computation The K-CV procedure, solving $\hat{\theta}^{RMLE}_{k,m}$ for $K \times |\Lambda|$ times, is costly in computation, especially for mixed models because of the numerical integration. There are some techniques in the LASSO literature such as least angle regression algorithm *(Efron et al., 2004)* and cyclical coordinate descent algorithm *(Friedman, Hastie and Tibshirani, 2010)* that can accelerate the solution of the whole regularization path. However, they may not work well either because of the nonlinearity or the high cost of solving the gradient. Hopefully, each penalty and fold can be solved independently so it is ideal to implement parallel computing. I evaluated 10-fold CV for each case on a cluster with 72 cores and 216 gigabytes memory, and it took at most 7 hours to finish the whole procedure (4 combinations of n and J) for one case.

Results Given datasets (one for each case) generated according to Table C.1, I implement the K-CV procedure and display the testing log-likelihood for various n_t and J_t in Figure C.1-C.2. The baseline models in C.1 imply that adding penalty is unnecessary for low-dimensional models, although small penalty seems to be harmless and slightly improve the out-of-sample prediction. In contrast, Figures C.2-C.2 demonstrate the importance of penalizing many parameters. Most curves have a U-shape with tight confidence intervals. On one hand, when there is no (or small) penalty, the models with high-dimensional parameters are likely to overfit the data. On the other hand, when there is too much penalty, most parameters are zero so the models predict the same choice probability for each product and consumer. The K-CV procedure should be ideally implemented for every dataset generated in the simulation, but the computation is too costly. I summarize the following Table C.1 for references. Since the log-likelihood curves are flat in the neighborhood of the optimum, when I run the 500 Monte Carlo simulations,

1. in two baseline cases (the low-dimensional models), CDML nuisance estimates are from MLE but not RMLE;

2. in the high-dimensional models (Case 1-4), we simply choose the c^* for RMLE (and hence CDML) from Table C.1.

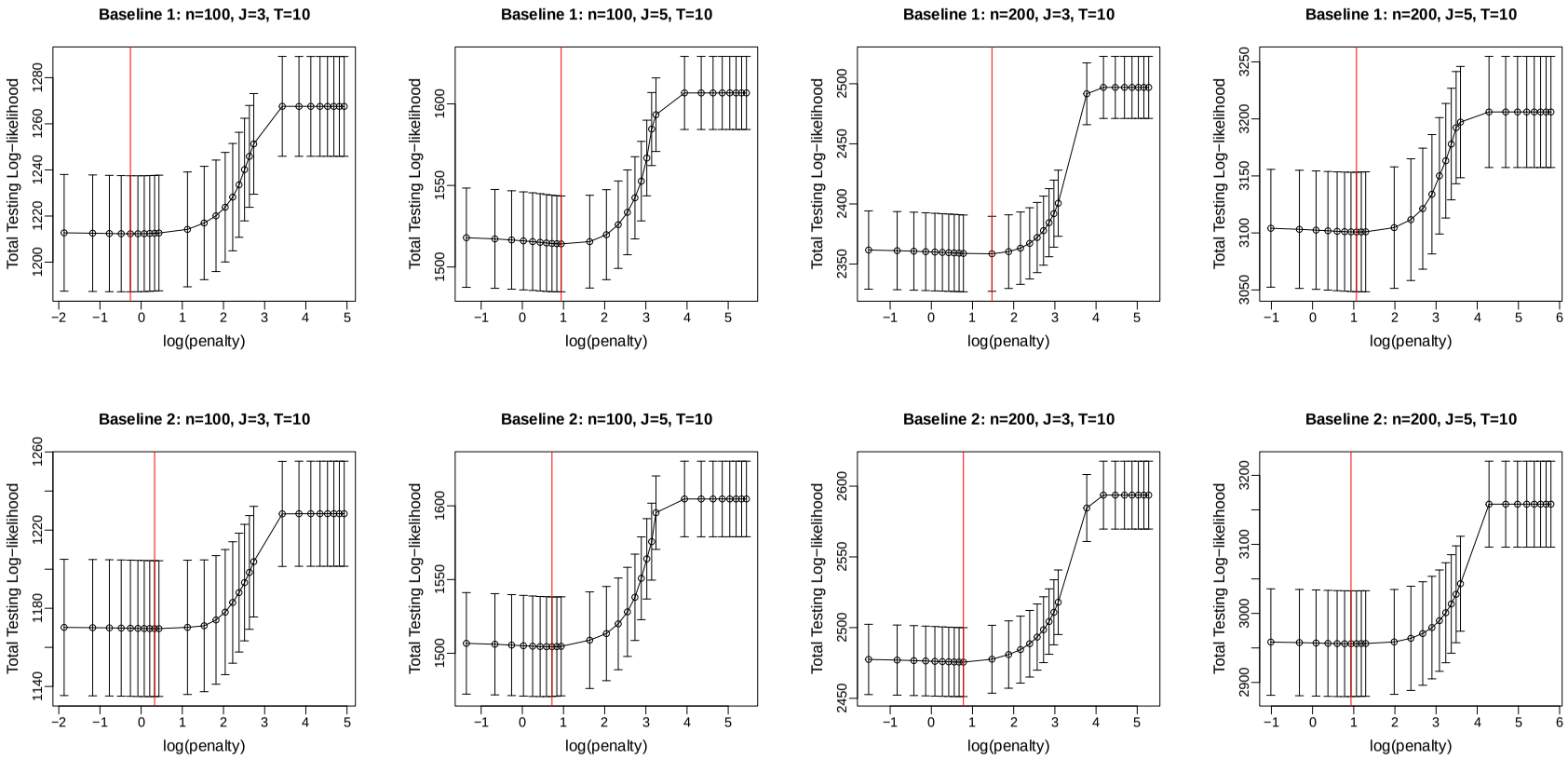
I also plot the paths of both penalized coefficients Π (solid lines) and unpenalized coefficients $\bar{\beta}, \beta^x, \alpha, \Sigma$ (dashed lines) in Figure C.4-C.5. The paths are expected: the penalized coefficients shrink to zero as the penalty increases while the unpenalized coefficients vary slightly and do not shrink, except for the standard deviation. The behaviors of the standard

Table C.1: Optimal Penalty Multipliers from 10-fold CV

Cases			Optimal Multipliers c^* in $\lambda_N = c^* N \rho_N$			
	d_Π	Σ	$n_t = 100, J_t = 3$	$n_t = 100, J_t = 5$	$n_t = 200, J_t = 3$	$n_t = 200, J_t = 5$
Baseline 1	10	1	0.003	0.01	0.01	0.01
Baseline 2	10	0	0.01	0.01	0.02	0.008
1	200	1	0.05	0.03	0.05	0.03
2	200	0	0.03	0.02	0.04	0.03
3	400	1	0.05	0.03	0.04	0.03
4	400	0	0.03	0.02	0.05	0.03

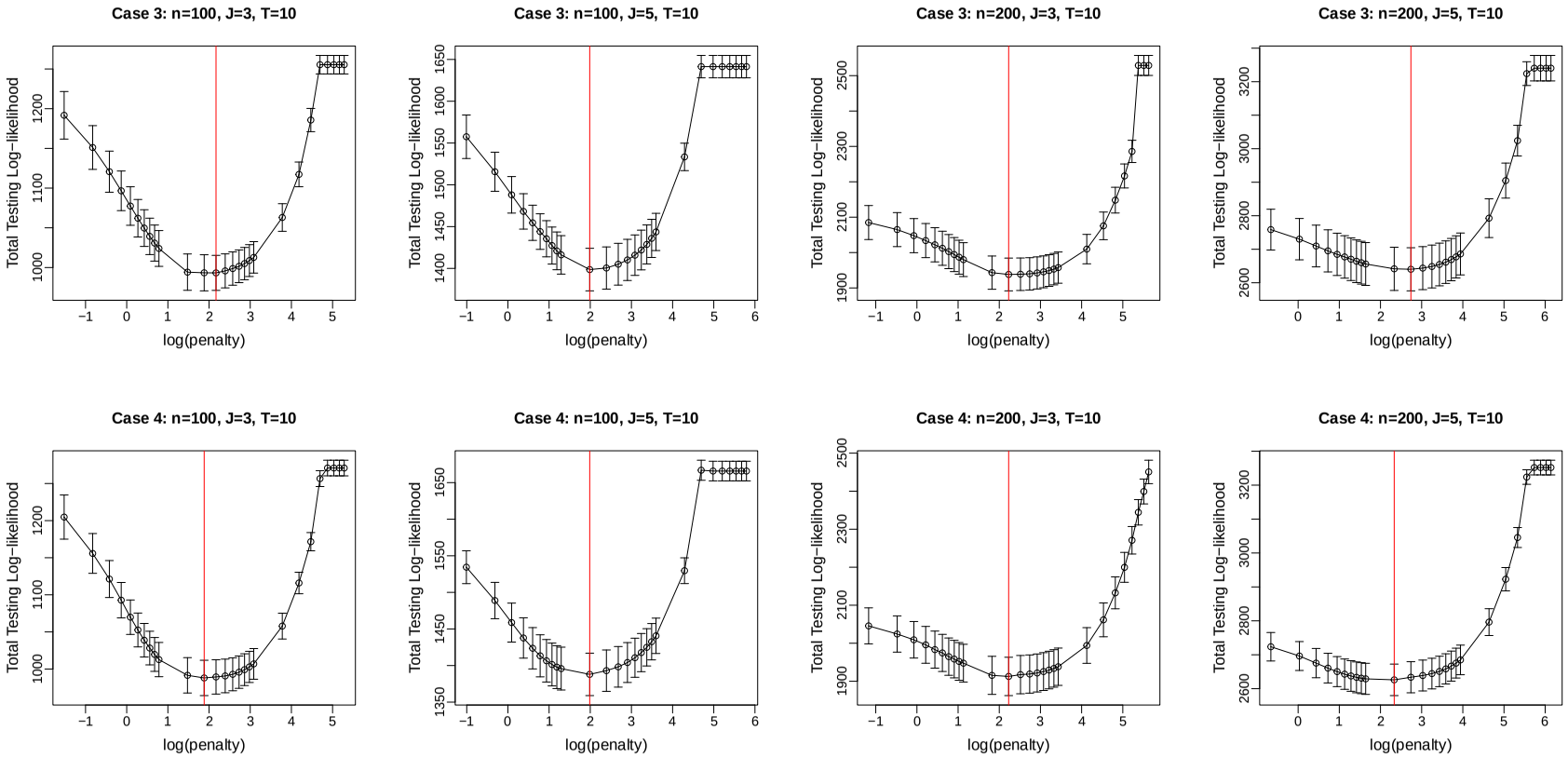
deviation are interesting. In the baseline models whose dimensions are relatively low, the standard deviations Σ can be estimated properly under zero or small penalty. In Case 1-4 whose dimensions are relatively high, the effect of random coefficients is dispersed when the penalty is small and many non-zero coefficients, because the covariates $X_{jt} \otimes L_i$ and $X_{jt} \odot \nu_i$ share the same product characteristic X_{jt} and the sample size is not enough to unveil $Cov(L_i, \nu_i) = 0$. When the penalty is moderate such that the sparsity is recovered, the standard deviation can then be estimated properly. It is worth noting that, when the penalty is too large, the estimation can be wrong again because all penalized coefficients are forced to be zero and the standard deviation will be over-estimated (see the baseline 2, Case 2 and 4).

Figure C.1: Path of Negative Testing Log-likelihood in 10-fold CV (Baselines)



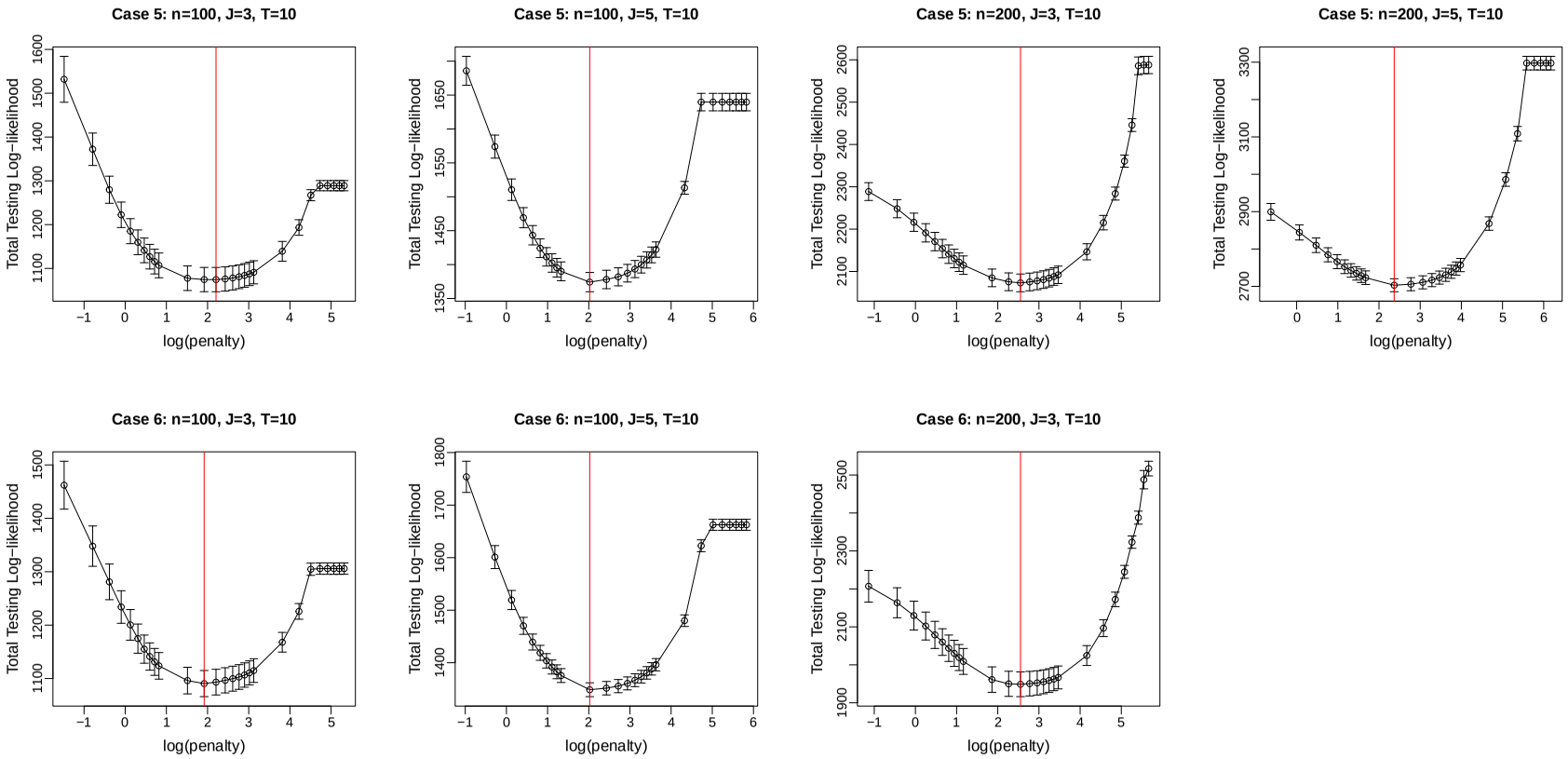
Notes: These figures summarize the log-likelihood under the testing samples and estimated parameters under different penalties. In each title, $n_t \in \{100, 200\}$ is the number of consumers in a market, $J_t \in \{3, 5\}$ is the number of inside options (products) and $T = 10$ is the number of markets. The horizontal axes indicate the natural logarithm of the penalty, and the vertical axes are the summation of the negative testing log-likelihood over five folds of testing samples. The hollow dots and the bars represent the values and the confidence intervals (1.96 times standard deviation), respectively. The red solid line shows the optimal penalty that minimizes the total testing log-likelihood.

Figure C.2: Path of Negative Testing Log-likelihood in 10-fold CV (Case 1 and 2)



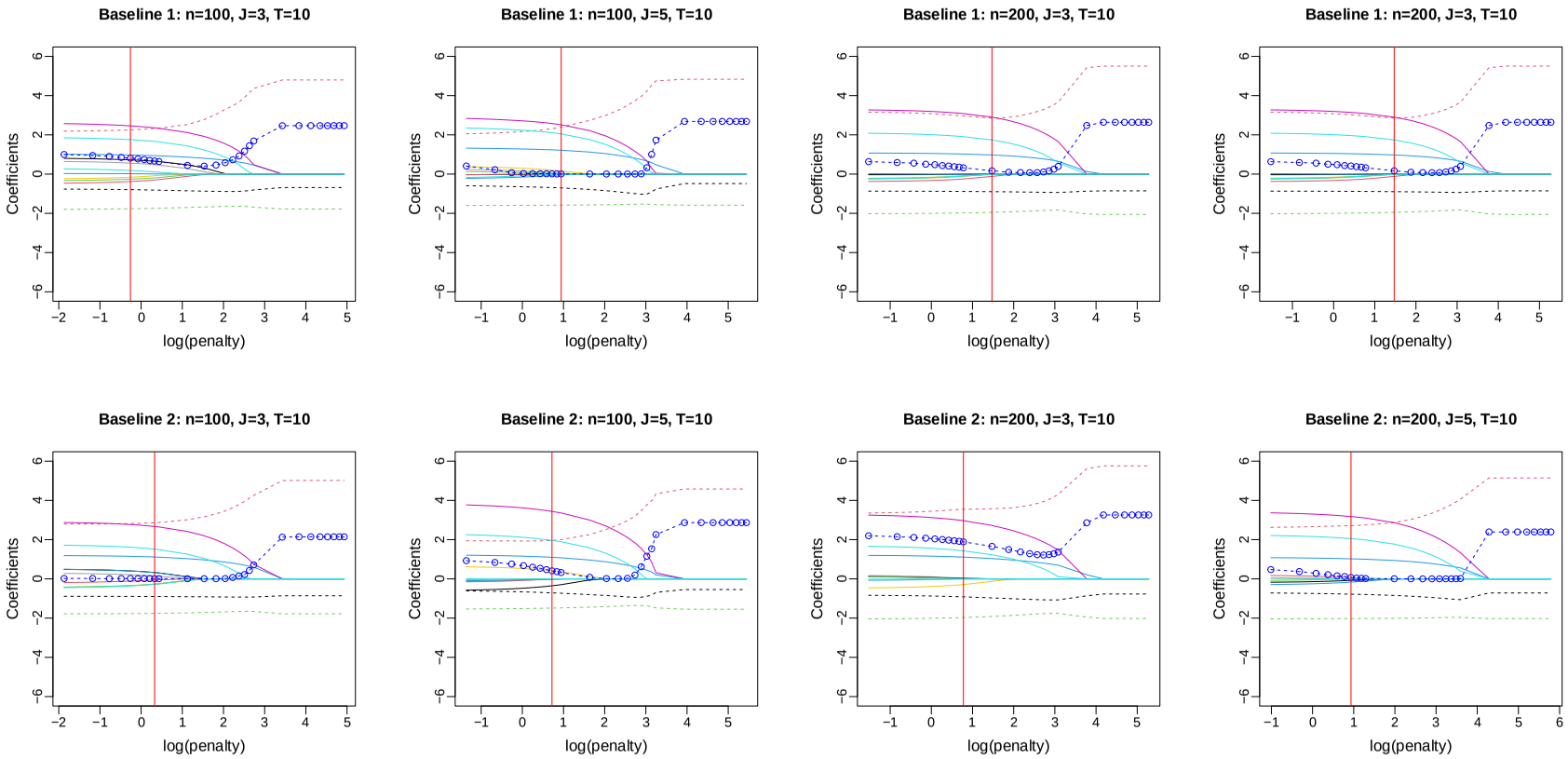
Notes: These figures summarize the log-likelihood under the testing samples and estimated parameters under different penalties. In each title, $n_t \in \{100, 200\}$ is the number of consumers in a market, $J_t \in \{3, 5\}$ is the number of inside options (products) and $T = 10$ is the number of markets. The horizontal axes indicate the natural logarithm of the penalty, and the vertical axes are the summation of the negative testing log-likelihood over five folds of testing samples. The hollow dots and the bars represent the values and the confidence intervals (1.96 times standard deviation), respectively. The red solid line shows the optimal penalty that minimizes the total testing log-likelihood.

Figure C.3: Path of Negative Testing Log-likelihood in 10-fold CV (Case 3 and 4)



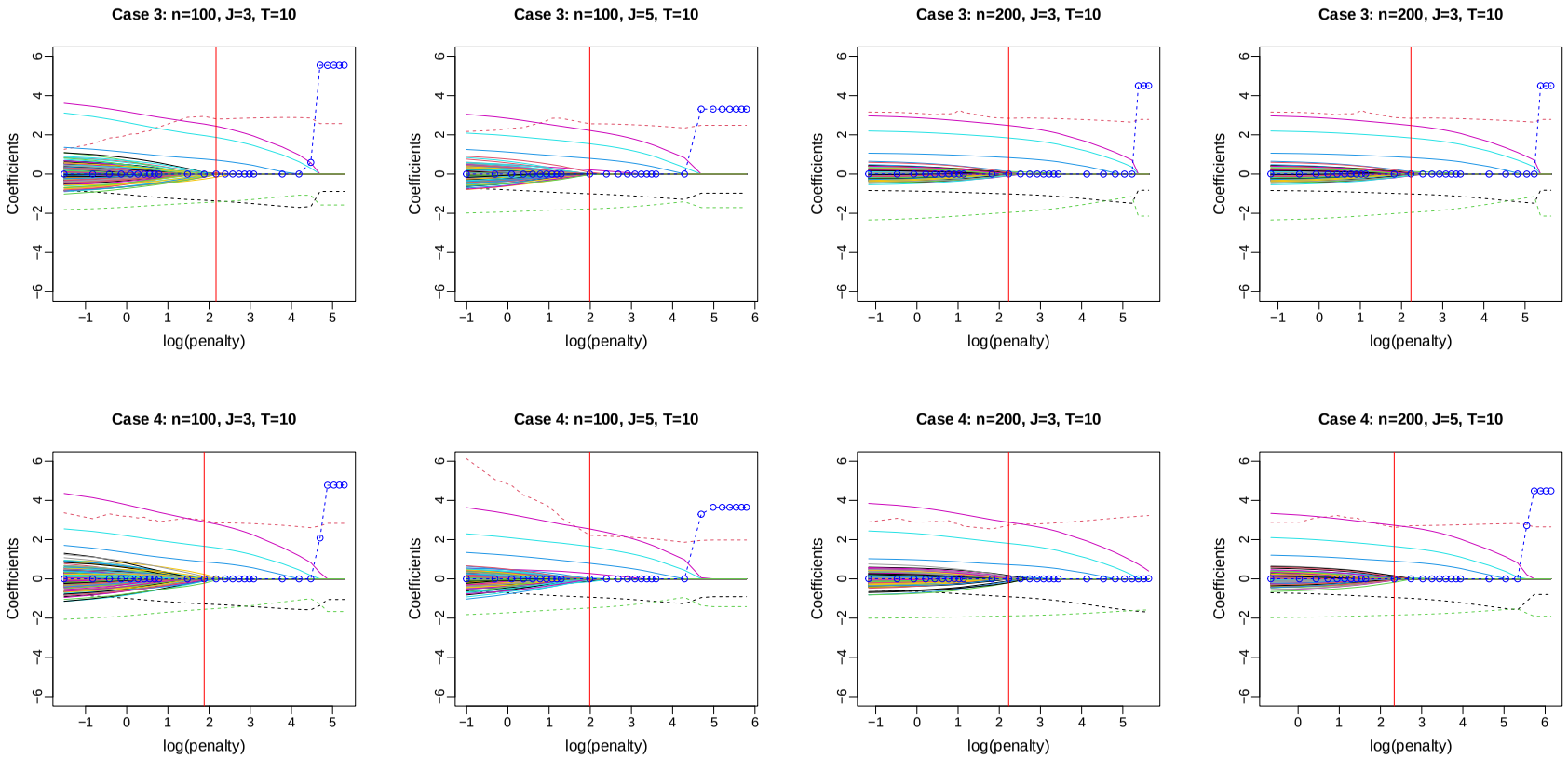
Notes: These figures summarize the log-likelihood under the testing samples and estimated parameters under different penalties. In each title, $n_t \in \{100, 200\}$ is the number of consumers in a market, $J_t \in \{3, 5\}$ is the number of inside options (products) and $T = 10$ is the number of markets. The horizontal axes indicate the natural logarithm of the penalty, and the vertical axes are the summation of the negative testing log-likelihood over five folds of testing samples. The hollow dots and the bars represent the values and the confidence intervals (1.96 times standard deviation), respectively. The red solid line shows the optimal penalty that minimizes the total testing log-likelihood.

Figure C.4: Paths of Estimated Coefficients in 10-fold CV (Baselines)



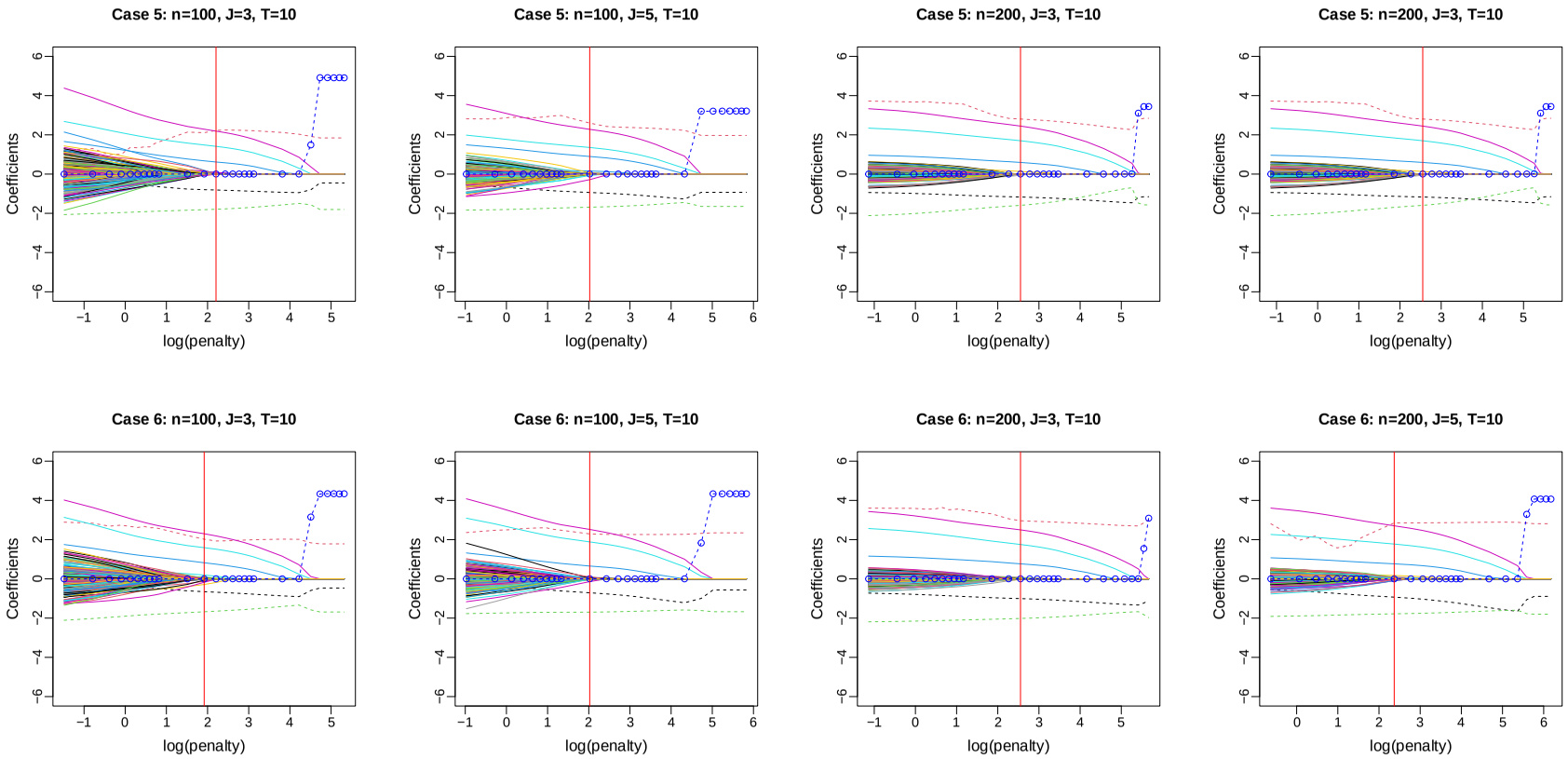
Notes: These figures summarize the path of all coefficients under different penalties. In each title, $n_t \in \{100, 200\}$ is the number of consumers in a market, $J_t \in \{3, 5\}$ is the number of inside options (products) and $T = 10$ is the number of markets. The horizontal axes indicate the natural logarithm of the penalty, and the vertical axes indicate the average estimated coefficients over five folds of testing samples. The penalized coefficients π are in solid line while the unpenalized coefficients are in dashed line (black for β_0, pink for β^x, green for α). The blue dashed line with hollow dots represent the standard deviation Σ of the random coefficient. The red solid line shows the optimal penalty that minimizes the total testing log-likelihood.

Figure C.5: Path of Negative Testing Log-likelihood in 10-fold CV (Case 3 and 4)



Notes: These figures summarize the path of all coefficients under different penalties. In each title, $n_t \in \{100, 200\}$ is the number of consumers in a market, $J_t \in \{3, 5\}$ is the number of inside options (products) and $T = 10$ is the number of markets. The horizontal axes indicate the natural logarithm of the penalty, and the vertical axes indicate the average estimated coefficients over five folds of testing samples. The penalized coefficients π are in solid line while the unpenalized coefficients are in dashed line (black for β_0, pink for β^x, green for α). The blue dashed line with hollow dots represent the standard deviation Σ of the random coefficient. The red solid line shows the optimal penalty that minimizes the total testing log-likelihood.

Figure C.6: Path of Negative Testing Log-likelihood in 10-fold CV (Case 5 and 6)



Notes: These figures summarize the path of all coefficients under different penalties. In each title, $n_t \in \{100, 200\}$ is the number of consumers in a market, $J_t \in \{3, 5\}$ is the number of inside options (products) and $T = 10$ is the number of markets. The horizontal axes indicate the natural logarithm of the penalty, and the vertical axes indicate the average estimated coefficients over five folds of testing samples. The penalized coefficients π are in solid line while the unpenalized coefficients are in dashed line (black for β_0, pink for β^x, green for α). The blue dashed line with hollow dots represent the standard deviation Σ of the random coefficient. The red solid line shows the optimal penalty that minimizes the total testing log-likelihood.

Table D.1: Integral Approximation, Accuracy and Number of Drawn Nodes

Method	**1-dim Approx.**	**Nodes in d-dim**
Laplace	$\int e^{-Su(t)}dt \approx (2\pi)^{1/2}\lvert Bu''(\hat{t})\rvert^{-1/2}\exp(-Bu(\hat{t}))$ relative error: $O(B^{-1})$	1
Gauss-Hermite	$\int u(t)e^{-t^2}dt \approx \sum_{b=1}^{B} w_b u(t_b)$ exact up to $(2B-1)$-th polynomials	B^d
Monte Carlo (MC)	$\int u(t)F(dt) \approx B^{-1}\sum_{b=1}^{B} u(t_b)$ with $t_b \sim i.i.d.\ F(t)$ absolute error: $O(B^{-1/2})$	B
Quasi MC	$\int u(t)F(dt) \approx B^{-1}\sum_{b=1}^{B} u(t_b)$ with $t_b := F^{-1}(h_b)$ absolute error: $O(B^{-1}\log B)$	B

Notes: This table summarizes the approximation of integrals in a single dimension by different approaches and the corresponding approximation errors in $K \in \mathbb{N}$ dimensions. In the Laplace approximation, $\hat{t} = \arg\min_t u(t)$ and $u''(\cdot)$ is the second-order derivative of $u(\cdot)$, see Bilodeau et al. (2023). In Gauss-Hermite quadrature, $w_b = 2^{b-1}b!\sqrt{\pi}b^{-2}[H_{b-1}(t_b)]^{-2}$ where $H_b(t)$ is the physicist's Hermite polynomial, see Liu and Pierce (1994) for discussions. If $I(u)$ is the integral and $\hat{I}(u)$ is the approximation, then the *relative* error denotes $\lvert I(u)/\hat{I}(u) - 1\rvert$ and the *absolute* error denotes $\lvert I(u) - \hat{I}(u)\rvert$.

D Numerical Integration

D.1 Gauss-Hermite Quadrature

Given the smoothness of the soft-max function, Gauss-Hermite quadrature can be implemented to approximate the following integral with change of variable:

$$s_{ij}(\theta) := \int_{-\infty}^{\infty} \frac{\exp(\omega_{ij}(\nu,\theta))}{1+\sum_{j'}\exp(\omega_{ij'}(\nu,\theta))}\frac{1}{\sqrt{2\pi}}\exp(-\nu^2/2)d\nu$$

where $\omega_{ij}(\nu,\theta) = \delta_j + (X_j \otimes L_i)'\Pi + (X_j \odot \nu)'\Sigma$. Let $m := \nu/\sqrt{2}$, then

$$s_{ij}(\theta) = \int_{-\infty}^{\infty} \frac{\exp(\omega_{ij}(\sqrt{2}m))}{1+\sum_{j'}\exp(\omega_{ij'}(\sqrt{2}m))}\frac{1}{\sqrt{\pi}}\exp(-m^2)dm \approx \frac{1}{\sqrt{\pi}}\sum_{b=1}^{B} w_b \frac{\exp(\omega_{ij}(\sqrt{2}m_b))}{1+\sum_{j'}\exp(\omega_{ij'}(\sqrt{2}m_b))}$$

which exactly corresponds to the Gauss-Hermite Quadrature. The node m_b and the weight m_b are given by the physicists' version of Hermite polynomial. This approximation is exact for $(2B-1)$-th order of polynomials. As an example, when $B = 5$, we have

$$w_1 = 0.0199, w_2 = 0.3936, w_3 = 0.9453, w_4 = 0.3936, w_5 = 0.0199$$

$$m_1 = -2.020, m_2 = -0.959, m_3 = 0, m_4 = 0.959, m_5 = 2.020$$

which is exact for 9-th order of polynomials. Similarly,

$$\frac{d}{d\theta}s_{ij}(\theta) = \int_{-\infty}^{\infty}\sum_{k=1}^{J}\psi'_k(\omega_i(v_i,\theta),j)\begin{pmatrix} Q_{ik}\odot v_i \\ X_{ik}\end{pmatrix}\frac{1}{\sqrt{2\pi}}\exp(-v_i^2/2)dv$$
$$\approx \frac{1}{\sqrt{\pi}}\sum_{b=1}^{B}\sum_{k=1}^{J} w_b\psi'_k(\omega_i(\sqrt{2}m_b,\theta),j)\begin{pmatrix} Q_{ik}\odot\sqrt{2}m_b \\ X_{ik}\end{pmatrix}$$

$$\frac{d^2}{d\theta d\theta'}s_{ij}(\theta) = \int_{-\infty}^{\infty}\sum_{k=1}^{J}\sum_{l=1}^{J}\psi''_l(\omega_i,j,k)\begin{pmatrix} (Q_{il}\odot v_i)(Q_{ik}\odot v_i)' & (Q_{il}\odot v_i)X'_{ik} \\ X_{il}(Q_{ik}\odot v_i)' & X_{il}X'_{ik}\end{pmatrix}$$
$$\approx \frac{1}{\sqrt{\pi}}\sum_{b=1}^{B}\sum_{k=1}^{J}\sum_{l=1}^{J} w_b\psi''_l(\omega_i(\sqrt{2}m_b,\theta),j,k)\begin{pmatrix} (Q_{il}\odot\sqrt{2}m_b)(Q_{ik}\odot\sqrt{2}m_b)' & (Q_{il}\odot\sqrt{2}m_b)X'_{ik} \\ X_{il}(Q_{ik}\odot\sqrt{2}m_b)' & X_{il}X'_{ik}\end{pmatrix}$$

where $\psi'_k(\omega_i,j) = \psi_j(\omega_i)\left(I[k=j]-\psi_k(\omega_i)\right)$ and $\psi''_l(\omega_i,j,k) = \psi'_l(\omega_i,j)I[k=j]-\psi'_l(\omega_i,j)\psi_k(\omega_i)-\psi'_l(\omega_i,k)\psi_j(\omega_i)$.

D.2 Quasi-Monte Carlo Integration

Train (2000, also see his references) discusses the random draws in mixed logit models based on Halton sequences. These draws based on deterministic sequences can achieve better symmetry and smaller variance[16] in simulation errors, especially when the dimension is relatively large. It is better to under such Halton sequence in an example. A Halton sequence for number 3 (must be a prime number) is constructed as follows:

- First, divide $(0,1)$ into 3 parts and get nodes $1/3, 2/3$;
- Next, divide each parts into 3 parts and get nodes from each parts sequentially $1/9$, $4/9, 7/9, 2/9, 5/9, 8/9$;
- Next, divide each parts into 3 parts ...

Then the Halton sequence is

$$(h_1(3), h_2(3), h_3(3), \dots) := (\frac{1}{3}, \frac{2}{3}, \frac{1}{9}, \frac{4}{9}, \frac{7}{9}, \frac{2}{9}, \frac{5}{9}, \frac{8}{9}, \dots)$$

and similar procedure can be done with other prime numbers. To draw a sequence of random variables $x_1, \dots, x_n$ from $F(\cdot)$ according to the Halton sequence $\{h_i(3)\}_{i=1}^n$, simply let $x_i := F^{-1}(h_i(3))$. To draw random vectors, just consider multiple Halton sequences such as $\{(h_i(2), h_i(3), h_i(5), \dots)\}_{i=1}^n$.

[16] Train (2000) finds that the variance over draws in the simulated probability for an observation is half as large with 100 Halton draws than 1,000 random draws.

Train (1999) suggests the following steps to simulate the random part $\theta_i^{(1)}, \ldots, \theta_i^{(B)} \in \mathbb{R}^Q$ for $i = 1, \ldots, n$:

1. Calculate a Halton sequence of length $(Bn + 10)$ using Q prime numbers $k_1, \ldots, k_Q$:

$$(h_i)_{i=1}^{Bn+10} := (h_i(k_1), h_i(k_2), \ldots, h_i(k_Q))_{i=1}^{Bn+10}$$

   where the first 10 entries are discarded to mitigate potential correlation;

2. For each $i = 1, \ldots, n$, $l = 1, \ldots, B$ and $q = 1, \ldots, Q$, calculate

$$\tilde{\theta}_{i,q}^{(l)} = F^{-1}(h_{10+(i-1)B+l}(k_q))$$

   where F^{-1} can be the inverse CDF of $N(0, 1)$

3. For each $i = 1, \ldots, n$ and $l = 1, \ldots, B$, calculate

$$(\theta_{i,1}^{(l)}, \ldots, \theta_{i,Q}^{(l)}) = \Sigma^{1/2}(\tilde{\theta}_{i,1}^{(l)}, \ldots, \tilde{\theta}_{i,Q}^{(l)})$$

Remark 1. Both *cmmixlogit* in STATA17 and the MATLAB code by Train (2006) use low-discrepancy sequences such as Halton sequences and Hammersley sets.

E Proof of Main Results

E.1 Proof of Lemma 1

Recall that

$$\frac{d}{d\theta}L_{nJ}(\theta_0) = \sum_{i=1}^{n}\sum_{j=0}^{J}\Upsilon_{ij}\frac{1}{s_{ij}(\theta_0)}\frac{d}{d\theta}s_{ij}(\theta_0)$$

where $s_{ij}(\theta_0) = \int \psi_j(\omega_i(\theta_0))\Phi(d\nu_i)$, and $\Phi(\cdot)$ is the CDF of the multivariate normal distribution $N(0, I_{d_\nu})$. Given that $\sum_{j=0}^{J}\Upsilon_{ij} = 1$ and the assumption $\min_j s_{ij}(\theta_0) \geq C_J J^{-1} > 0$,

$$\frac{d}{d\theta}L_{nJ}(\theta_0) \leq C_J^{-1} J \sum_{i=1}^{n}\max_j \frac{d}{d\theta}s_{ij}(\theta_0)$$

where the inequality here is element-wise. It suffices to derive the rate on $\frac{d}{d\theta}s_{ij}(\theta_0)$. Later, we will show that for any i and j,

$$\left|\frac{d}{d\theta}s_{ij}(\theta_0)\right| \leq 2C_{data}\iota'_{d_\theta} \quad \text{almost surely} \tag{E.1}$$

where ι_{d_θ} is a vector of ones with length d_θ. Now let $\mathcal{W}_i := \sum_{j=0}^{J}\Upsilon_{ij}s_{ij}(\theta_0)^{-1}\frac{d}{d\theta}s_{ij}(\theta_0)$ so $\frac{d}{d\theta}L_{nJ}(\theta_0) = \sum_{i=1}^{n}\mathcal{W}_i$. By Assumption 1, $\mathcal{W}_i$ is i.i.d. across $i = 1, \ldots, n$. By the union bound of probability

$$\begin{aligned}\mathbb{P}\left(\left\|\sum_{i=1}^{n}\mathcal{W}_i\right\|_\infty \geq \rho_n\right) = \mathbb{P}\left(\bigcup_{k=1}^{d_\theta}\left|\sum_{i=1}^{n}\mathcal{W}_{ik}\right| \geq \rho_n\right) &\leq \sum_{k=1}^{d_\theta}\mathbb{P}\left(\left|\sum_{i=1}^{n}\mathcal{W}_{ik}\right| \geq \rho_n\right) \\ &\leq d_\theta \max_k \mathbb{P}\left(\left|\sum_{i=1}^{n}\mathcal{W}_{ik}\right| \geq \rho_n\right)\end{aligned}$$

We will derive the bound by using the McDiarmid's inequality (see Lemma 5 in Appendix). Let $\mathcal{Z} := f(\mathcal{W}_1, \ldots, \mathcal{W}_n) := \sum_{i=1}^{n}\mathcal{W}_i$ so $\mathbb{E}[\mathcal{Z}] = 0 \in \mathbb{R}^{d_\theta}$ by Assumption 1. Without loss of generality, consider the first coordinate $k = 1$ so we can temporarily drop the subscript k. Now we verify the critical block, which is called the bounded difference property:

$$\sup_{w_1,\ldots w_n, w_{i'}} |f(w_1, \ldots, w_n) - f(w_1, \ldots, w_{i-1}, w_{i'}, w_{i+1}, \ldots, w_n)| \leq c_i, \quad 1 \leq i \leq n$$

for some constants $c_1, \ldots, c_n \geq 0$. In our design, $c_1, \ldots, c_n = 4(J+1)\max\{C_{data}, C_{data}\sqrt{2/\pi}\}$

because,

$$
\begin{aligned}
&|f(w_1,\ldots,w_n) - f(w_1,\ldots,w_{i-1},w_{i'},w_{i+1},\ldots,w_n)| \\
&\qquad = \sum_{j=0}^{J} y_{ij} s_{ij}(\theta_0)^{-1} \frac{d}{d\theta_1} s_{ij}(\theta_0) - \sum_{j=0}^{J} y_{i'j} s_{i'j}(\theta_0)^{-1} \frac{d}{d\theta_1} s_{i'j}(\theta_0) \\
&\qquad \leq 2 \max_i s_{ij}(\theta_0)^{-1} \times \max_i \left| \sum_{j=0}^{J} y_{ij} \frac{d}{d\theta_1} s_{ij}(\theta_0) \right| \\
&\qquad \leq 4 C_J^{-1} J C_{data}
\end{aligned}
$$

where the first inequality comes from the triangle inequality, and the second inequality comes from Assumption 3 and Eq.(E.1). Let $C := 4C_J^{-1}C_{data}$. Note that the bounded difference property holds for all k. As a result, by the McDiarmid's inequality,

$$
\mathbb{P}\left(|\mathcal{Z}_k - \mathbb{E}[\mathcal{Z}_k]| > c\right) = \mathbb{P}\left(\left|\sum_{i=1}^{n} \mathcal{W}_{ik}\right| > c\right) \leq 2\exp\left(-\frac{2c^2}{nJ^2C^2}\right)
$$

and hence

$$
\mathbb{P}\left(\rho_n^{-1}\left\|\frac{1}{n}\frac{d}{d\theta}L_{nJ}(\theta_0)\right\|_\infty \geq c\right) \leq 2\exp\left(-\frac{2n^2\rho_n^2 c^2}{nJ^2C^2} + \log d_\theta\right)
$$

Let $\rho_n = J\sqrt{n^{-1}\log d_\theta}$, then the right-hand side $2\exp((1-2c^2/C^2)\log d_\theta)$ can be arbitrarily small (as $d_\theta \to \infty$) whenever $c > C$. Therefore, we conclude that

$$
\left\|\frac{1}{n}\frac{d}{d\theta}L_{nJ}(\theta_0)\right\|_\infty = O_P(\rho_n) = O_P\left(J\sqrt{n^{-1}\log d_\theta}\right)
$$

Finally, we show Eq.(E.1). To save notation, we write $\psi_{ij} := \psi_j(\omega_i(\theta_0))$ without confusions. By the chain rule and the exchangeability,

$$
\frac{d}{d\theta}s_{ij}(\theta_0) = \int \frac{d}{d\theta}\psi_{ij}\Phi(d\nu_i) = \int \sum_{k=1}^{J} \frac{\partial\psi_{ij}}{\partial\omega_{ik}}\frac{d\omega_{ik}}{d\theta}\Phi(d\nu_i)
$$

By the property of soft-max function, $\sum_{j=0}^{J}\psi_{ij} = 1$ and

$$
\sum_{k=1}^{J}\frac{\partial\psi_{ij}}{\partial\omega_{ik}} = \sum_{k=1}^{J}\left(I[k=j]-\psi_{ij}\right)\psi_{ik} = \psi_{ij} - \sum_{k=1}^{J}\psi_{ij}\psi_{ik} = \psi_{ij}(1-\psi_{i0}) \in [0,\psi_{ij}]
$$

for all ν_i. Note that $\omega_{ik} = X_{ik}'\beta + (Q_{ik} \odot \nu_i)'\sigma$ and X_{ik} is independent of ν_i, similarly

$$\int \sum_{k=1}^{J} \frac{\partial \psi_{ij}}{\partial \omega_{ik}} \frac{d\omega_{ik}}{d\beta} \Phi(d\nu_i) = \int \psi_{ij} X_{ij} - \sum_{k=1}^{J} \psi_{ij}\psi_{ik} X_{ik} \Phi(d\nu_i) = X_{ij} s_{ij}(\theta_0) - \sum_{k=1}^{J} X_{ik} \int \psi_{ij}\psi_{ik} \Phi(d\nu_i)$$

For the first term, each coordinate $|X_{ijl} s_{ij}(\theta_0)| \leq C_{data}$ is bounded almost surely given the assumption $|X_{ijl}| \leq C_{data}$ almost surely. For the second term, notice that $\psi_{ij} \in [0,1]$ and $\sum_{j=1}^{J} \psi_{ij} = 1 - \psi_{i0} \leq 1$ for all ν_i, so

$$\sum_{k=1}^{J} X_{ik} \int \psi_{ij}\psi_{ik} \Phi(d\nu_i) \geq -C_{data} \int \psi_{ij}(1 - \psi_{i0}) \Phi(d\nu_i) \geq -C_{data}$$

Thus,

$$\left| \int \sum_{k=1}^{J} \frac{\partial \psi_{ij}}{\partial \omega_{ik}} \frac{d\omega_{ik}}{d\beta} \Phi(d\nu_i) \right| \leq 2 C_{data} \iota_{d_X}$$

where ι_{d_X} is a vector of ones with length d_X. It is slightly different for the derivative with respect to σ:

$$\int \sum_{k=1}^{J} \frac{\partial \psi_{ij}}{\partial \omega_{ik}} \frac{d\omega_{ik}}{d\beta} \Phi(d\nu_i) = \int \psi_{ij} Q_{ij} \odot \nu_i \Phi(d\nu_i) - \sum_{k=1}^{J} \int \psi_{ij}\psi_{ik} Q_{ik} \odot \nu_i \Phi(d\nu_i)$$

For the first term, since Q_{ij} is independent of ν_i, we can factor Q_{ij} out and rewrite it as $Q_{ij} \odot \int \psi_{ij}\nu_i \Phi(d\nu_i)$. Since $\psi_{ij} \in [0,1]$, then for $l = 1, \ldots, d_Q$,

$$\begin{aligned} \int \psi_{ij} Q_{ijl} \nu_{il} \Phi(d\nu_{il}) &= \int_{\nu_{il} \leq 0} \psi_{ij} Q_{ijl} \nu_{il} \Phi(d\nu_{il}) + \int_{\nu_{il} \geq 0} \psi_{ij} Q_{ijl} \nu_{il} \Phi(d\nu_{il}) \\ &\leq 2 \int_{\nu_{il} \geq 0} Q_{ijl} \nu_{il} \Phi(d\nu_{il}) = C_{data} \sqrt{\frac{2}{\pi}} \end{aligned}$$

where the last equality is from the mean of half-normal distribution[17] and the assumption

[17] If $Y = |X|$ and $|X| \sim N(0, \sigma^2)$, then Y follows a half-normal distribution with the density function

$$f(y) = \frac{\sqrt{2}}{\sigma\sqrt{\pi}} \exp\left(-\frac{y^2}{2\sigma^2}\right), \quad y \geq 0$$

and mean $\mathbb{E}[Y] = \frac{\sigma\sqrt{2}}{\sqrt{\pi}}$. In our case, $\sigma = 1$ and $\int_{\nu_{il} \geq 0} \nu_{il} \Phi(d\nu_{il}) = \frac{1}{2}\mathbb{E}[Y] = 1/\sqrt{2\pi}$.

that $|Q_{ijl}| \leq C_{data}$ almost surely. For the second term, we do similar decomposition

$$\int \psi_{ij}\psi_{ik}Q_{il}\nu_{il}\Phi(d\nu_{il}) = \int_{\nu_{il}\leq 0} \psi_{ij}\psi_{ik}Q_{il}\nu_{il}\Phi(d\nu_{il}) + \int_{\nu_{il}\geq 0} \psi_{ij}\psi_{ik}Q_{il}\nu_{il}\Phi(d\nu_{il})$$

and then

$$\int \psi_{ij}\psi_{ik}Q_{il}\nu_{il}\Phi(d\nu_{il}) \geq 2C_{data}\int_{\nu_{il}\leq 0} \psi_{ij}\psi_{ik}\nu_{il}\Phi(d\nu_{il})$$

Again, as $\sum_{k=1}^{J}\psi_{ik} = 1 - \psi_{i0}$ and $\psi_{ij} \in [0,1]$,

$$\sum_{k=1}^{J}\int \psi_{ij}\psi_{ik}Q_{il}\nu_{il}\Phi(d\nu_{il}) \geq 2C_{data}\int_{\nu_{il}\leq 0} \psi_{ij}(1-\psi_{i0})\nu_{il}\Phi(d\nu_{il}) \geq -C_{data}\sqrt{\frac{2}{\pi}}$$

Therefore,

$$\int \sum_{k=1}^{J}\frac{\partial\psi_{ij}}{\partial\omega_{ik}}\frac{d\omega_{ik}}{d\beta}\Phi(d\nu_i) \leq 2C_{data}\sqrt{\frac{2}{\pi}}\iota_{d_Q}$$

Combine with the previous bound and the proof is finished.

E.2 Proof of Theorem 1

We first verify the conditions for Theorem 1 in Negahban et al. (2012). Their condition (G1) is naturally satisfied with LASSO penalty and $\tau_L = 0$ as a result. Their condition (G2) also holds: $L_{nJ}(\theta)$ is a smooth function of θ and we assume conditions in Assumption 2. Since the dual norm of l_1-norm $\|\cdot\|_1$ is the l_∞-norm, $\mathcal{R}^*(\nabla L_{nJ}(\theta_0))$ is exactly $\|\frac{d}{d\theta}L_{nJ}(\theta_0)\|_\infty$, and Lemma 1 shows the rate to be $n\rho_n$.

Now we go through their proof of Theorem 1 in the supplementary material. Define

$$\mathcal{F}(\Delta) := L_{nJ}(\theta_0+\Delta) - L_{nJ}(\theta_0) + \lambda_n(\|\theta_0+\Delta\|_1 - \|\theta_0\|)$$

which is the difference between the penalized likelihood at $\theta_0+\Delta$ and θ_0. The authors first proves their Lemma 3: *if* $\lambda_n \geq 2\mathcal{R}^*(\nabla L_{nJ}(\theta_0))$ *and* L_{nJ} *is convex, then* $L_{nJ}(\theta_0+\Delta) - L_{nJ}(\theta_0) \geq -\frac{\lambda_n}{2}(\|\Delta_S\|_1 + \|\Delta_{S^c}\|_1)$. The proof is standard and the same when we consider a random design and sufficiently assume $L_{nJ}(\theta)$ is convex in a neighborhood of θ_0 which contains $\theta_0+\Delta$ (especially $\hat{\theta}^{RMLE}$). The statement now becomes: on the event $\mathcal{E}(\lambda_n) := \{\lambda_n \geq 2\mathcal{R}^*(\nabla L_{nJ}(\theta_0))\}$ (with probability greater than $1-a_n$), we have the inequality holds.

Since $\|\cdot\|_1$ is convex everywhere and $L_{nJ}(\cdot)$ is convex at $\theta_0 + t\Delta$ for all $t \in (0,1)$, $\mathcal{F}(\cdot)$ is also convex at $t\Delta$. Such convexity can be used to prove their Lemma 4: *if* $\mathcal{F}(\Delta) > 0$ *for all*

$\Delta \in \mathbb{C} \cap \{\|\Delta\| = \zeta\}$, *then* $\|\hat{\Delta}\| = \|\hat{\theta}^{RMLE} - \theta_0\| \leq \zeta$. A contrapositive statement is proved: *if there exists* $\hat{\Delta}$ *such that* $\|\hat{\Delta}\| > \zeta$, *then there is some* $\tilde{\Delta} \in \mathbb{C} \cap \{\|\Delta\| = \zeta\}$ *such that* $\mathcal{F}(\tilde{\Delta}) \leq 0$. Since $\|\hat{\Delta}\| > \zeta$ and $\mathbb{C}$ is a cone, then there must be some $t^* \in (0, 1)$ such that $\|t^*\hat{\Delta}\| = \zeta$ on the line joining $\hat{\Delta}$ and 0. By Jensen's inequality and $\mathcal{F}(0) = 0$,

$$\mathcal{F}(t^*\hat{\Delta} + (1 - t^*)0) \leq t^*\mathcal{F}(\hat{\Delta}) + (1 - t^*)\mathcal{F}(0) = t^*\mathcal{F}(\hat{\Delta}) \leq 0$$

The last inequality comes from the fact that $\hat{\theta}^{RMLE}$ is the minimum point so $\mathcal{F}(\hat{\Delta}) \leq 0$. Then, simply define $\tilde{\Delta} = t^*\hat{\Delta}$ suggesting $\mathcal{F}(\tilde{\Delta}) \leq 0$ and hence the Lemma 4 is proved.

The rest of the proof is then straightforward. By (G1) and (G2), with probability greater than $1 - a_n$,

$$\mathcal{F}(\Delta) \geq \left(\frac{d}{d\theta}L_{nJ}(\theta_0)\right)' \Delta + n\kappa_L\|\Delta\|_2^2 + \lambda_n(\|\theta_0 + \Delta\|_1 - \|\theta_0\|_1)$$

for all $\Delta \in \mathbb{C}$. Using the Hölder's inequality, with probability greater than $1 - b_n$,

$$\left|\left(\frac{d}{d\theta}L_{nJ}(\theta_0)\right)' \Delta\right| \leq \left\|\frac{d}{d\theta}L_{nJ}(\theta_0)\right\|_\infty \|\Delta\|_1 \leq \frac{1}{2}\lambda_n\|\Delta\|_1$$

for $\lambda_n \geq 2n\rho_n$. Then, with probability greater than $1 - a_n - b_n$,

$$\mathcal{F}(\Delta) \geq n\kappa_L\|\Delta\|_2^2 + \lambda_n(\|\theta_0 + \Delta\|_1 - \|\theta_0\|_1) - \frac{1}{2}\lambda_n\|\Delta\|_1$$

Some algebra shows that $\|\theta_0 + \Delta\|_1 - \|\theta_0\| \geq \|\Delta_{S^c}\|_1 - \|\Delta_S\|_1 - 2\|\theta_{0,S^c}\|_1 = \|\Delta_{S^c}\|_1 - \|\Delta_S\|_1$. Plug in this inequality and we obtain

$$\mathcal{F}(\Delta) \geq n\kappa_L\|\Delta\|_2^2 - \frac{3}{2}\lambda_n\|\Delta_S\|_1 \geq n\kappa_L\|\Delta\|_2^2 - \frac{3}{2}\lambda_n\sqrt{s_n}\|\Delta\|_2$$

According to the Lemma 4, strictly positive right-hand side implies $\|\hat{\Delta}\|_2 \leq \zeta$, which is true as long as $\|\Delta\|_2 \geq \frac{3\sqrt{s_n}\lambda_n}{2n\kappa_L}$ as $\kappa_L > 0$. Thus, we can let $\zeta = \frac{3\sqrt{s_n}\lambda_n}{n\kappa_L}$, which generate the first error bound in our lemma. The second error bound is then by the Hölder's inequality $\|\hat{\Delta}\|_1 \leq \sqrt{s_n}\|\hat{\Delta}\|_2$.

E.3 Proof of Corollary 1

The proof is almost the same except that we need to derive the concentration inequality for $\|N^{-1}\frac{d}{d\theta}L_{NJT}(\theta_0)\|_\infty$. Again, by the definition,

$$\frac{d}{d\theta}L_{NJT}(\theta_0) = \sum_{t=1}^{T}\sum_{i\in\mathcal{I}_t}\sum_{j\in\mathcal{J}_t} Y_{ijt}\frac{1}{s_{ijt}(\theta_0)}\frac{d}{d\theta}s_{ijt}(\theta_0)$$

Given that $\sum_{j\in\mathcal{J}_t} Y_{ijt} = 1$ and $\min_{j\in\mathcal{J}_t} s_{ijt}(\theta_0) \geq C_J J_t^{-1} > 0$ from Assumption 4,

$$\sum_{j\in\mathcal{J}_t} Y_{ijt}\frac{1}{s_{ijt}(\theta_0)}\frac{d}{d\theta}s_{ijt}(\theta_0) \leq C_J^{-1}J_t^{-1}\max_{j\in\mathcal{J}_t}\frac{d}{d\theta}s_{ijt}(\theta_0)$$

where the inequality here is element-wise, and hence

$$\frac{d}{d\theta}L_{NJT}(\theta_0) \leq \sum_{t=1}^{T}\sum_{i\in\mathcal{I}_t} C_J^{-1}J_t^{-1}\max_{j\in\mathcal{J}_t}\frac{d}{d\theta}s_{ijt}(\theta_0)$$

It suffices to derive the bound on $\frac{d}{d\theta}s_{ijt}(\theta_0)$ for each i, j, t. Consider $\theta = (\delta', \beta^{o'}, \beta^{u'})'$ for generality. Some algebra shows that

$$\frac{d}{d\theta}s_{ijt}(\theta_0) = \int\sum_{k=1}^{J}\frac{\partial\psi_{ij}}{\partial\omega_{ik}}\frac{d\omega_{ik}}{d\theta}\Phi(d\nu_i) \quad \text{where} \quad \frac{d\omega_{ik}}{d\theta} = \begin{pmatrix} e_k \\ X_{jt}\otimes L_i \\ X_{jt}\odot\nu_i \end{pmatrix}$$

Here $e_k = (0,\ldots,0,1,0,\ldots,0)'$ is a vector of length J whose k-th entry is equal to one and the others are all zero. By Assumption 4, X_{jt} and L_i are all bounded random vectors. Therefore, we can prove in the exact same way as Lemma 1 and show that for any i, j, t,

$$\left|\frac{d}{d\theta}s_{ijt}(\theta_0)\right| \leq 2\left(\iota'_{d_\delta}, C_{data}^2\iota'_{d_X d_L}, C_{data}\iota'_{d_X}\right)' \quad \text{almost surely}$$

Let $\mathcal{W}_{it} := C_J^{-1}J_t^{-1}\max_{j\in\mathcal{J}_t}\frac{d}{d\theta}s_{ijt}(\theta_0)$ and $\mathcal{W}_{it,k}$ be the k-the entry. The union bound inequality implies

$$\mathbb{P}\left(\left\|\sum_{t=1}^{T}\sum_{i\in\mathcal{I}_t}\mathcal{W}_{it}\right\|_\infty \geq \rho_N\right) \leq d_\theta\max_k\mathbb{P}\left(\left|\sum_{i=1}^{n}\mathcal{W}_{it,k}\right| \geq \rho_N\right) \quad \text{for} \quad \rho_N > 0$$

The bounded difference property is satisfied with constant

$$C := 4C_J^{-1} \times \max_t J_t \times \max\{1, C_{data}^2, C_{data}\}$$

By the McDiarmid's inequality again, for any $c > 0$ and $\rho_N > 0$,

$$\mathbb{P}\left(\rho_N^{-1}\left\|\frac{1}{N}\frac{d}{d\theta}L_{NJT}(\theta_0)\right\|_\infty \geq c\right) \leq 2\exp\left(-\frac{2N^2\rho_N^2 c^2}{N(\max_t J_t)^2 C^2} + \log d_\theta\right)$$

and hence,

$$\left\|\frac{1}{N}\frac{d}{d\theta}L_{NJT}(\theta_0)\right\|_\infty = O_P\left(\max_t J_t\sqrt{N^{-1}\log d_\theta}\right)$$

The error bounds for the regularized estimator are direct results from Lemma 1.

E.4 Proof of Theorem 2

For clarity, the proof is separated into eight steps. The outline is given in Step 1 where we assume all rates are known. In Step 2-3, we derive the rate for $M_{NJK}(\gamma_0; \hat{\eta}, \hat{\mu}) - M_{NJK}(\gamma_0; \eta_0, \mu_0)$ by using empirical process notations and the cross-fitting technique. In Step 4-5, we derive the rate for $D_{\gamma'}M_{NJK}(\gamma_0; \hat{\eta}, \hat{\mu}) - D_{\gamma'}M_{NJK}(\gamma_0; \eta_0, \mu_0)$ which does not rely on the cross-fitting. In Step 6, we derive the rate for $D_\gamma^2 M_{NJ,l}(\gamma_0)$. Step 7 is the central limit theorem for $N^{1/2}M_{NJK}(\gamma_0; \eta_{N0}, \mu_{N0})$ and Step 8 is the weak law of large numbers for $D_{\gamma'}M_{NJK}(\gamma_0; \eta_0, \mu_0)$, proved by Hoeffding's inequality.

Step 1 By the definition in Eq.(4.1),

$$\begin{aligned}
M_{NJK}(\gamma_0; \hat{\eta}, \hat{\mu}) - M_{NJK}(\gamma_0; \eta_0, \mu_0) &= \frac{1}{K}\sum_{k=1}^K \frac{1}{N/K}\sum_{i\in\mathcal{I}_k} m_i(\gamma_0; \hat{\eta}_k, \hat{\mu}_k) - m_i(\gamma_0; \eta_0, \mu_0) \\
&= \frac{1}{K}\sum_{k=1}^K \frac{1}{N/K}\left(\frac{\partial}{\partial\gamma'}L_{NJ}^{(k)}(\gamma_0; \hat{\eta}_k) - \frac{\partial}{\partial\gamma'}L_{NJ}^{(k)}(\gamma_0; \eta_0)\right. \\
&\quad \left. + \mu_0\frac{\partial}{\partial\eta'}L_{NJ}^{(k)}(\gamma_0; \eta_0) - \hat{\mu}_k\frac{\partial}{\partial\eta'}L_{NJ}^{(k)}(\gamma_0; \hat{\eta}_k)\right)
\end{aligned}$$

$$D_{\gamma'}M_{NJK}(\gamma_0;\hat{\eta},\hat{\mu}) - D_{\gamma'}M_{NJK}(\gamma_0;\eta_0,\mu_0) = \frac{1}{K}\sum_{k=1}^{K}\frac{1}{N/K}\sum_{i\in\mathcal{I}_k}\frac{\partial}{\partial\gamma'}m_i(\gamma_0;\hat{\eta}_k,\hat{\mu}_k) - \frac{\partial}{\partial\gamma'}m_i(\gamma_0;\eta_0,\mu_0)$$
$$= \frac{1}{K}\sum_{k=1}^{K}\frac{1}{N/K}\left(\frac{\partial^2}{\partial\gamma\partial\gamma'}L_{NJ}^{(k)}(\gamma_0;\hat{\eta}_k) - \frac{\partial^2}{\partial\gamma\partial\gamma'}L_{NJ}^{(k)}(\gamma_0;\eta_0)\right.$$
$$\left. + \mu_0\frac{\partial^2}{\partial\eta\partial\gamma'}L_{NJ}^{(k)}(\gamma_0;\eta_0) - \hat{\mu}_k\frac{\partial}{\partial\eta\partial\gamma'}L_{NJ}^{(k)}(\gamma_0;\hat{\eta}_k)\right)$$

and

$$D_\gamma^2 M_{NJ,j}(\gamma_0) = \frac{1}{K}\sum_{k=1}^{K}\frac{1}{N/K}\sum_{i\in\mathcal{I}_k}\frac{\partial^2}{\partial\gamma\partial\gamma'}m_{ij}(\gamma_0;\eta_0,\mu_0)$$
$$= \frac{1}{K}\sum_{k=1}^{K}\frac{1}{N/K}\frac{\partial^2}{\partial\gamma\partial\gamma'}\frac{\partial}{\partial\gamma_j}L_{NJ}^{(k)}(\gamma_0;\eta_0) - \frac{\partial^2}{\partial\gamma\partial\gamma'}\mu_{0,j}\frac{\partial}{\partial\eta'}L_{NJ}^{(k)}(\gamma_0;\eta_0)$$

Taking the advantage of cross-fitting, we consider the decomposition in *Chernozhukov et al. (2018, C55)* and use our Lemma 6.

We first derive the rate for $\sqrt{N}\|M_{NJK}(\gamma_0;\hat{\eta},\hat{\mu}) - M_{NJK}(\gamma_0;\eta_0,\mu_0)\|_2$. Let $\mathbb{P}_{n,k}$ be the empirical measure over the subsamples indexed by $\mathcal{I}_k$. Thus,

$$M_{NJK}(\gamma_0;\hat{\eta},\hat{\mu}) - M_{NJK}(\gamma_0;\eta_0,\mu_0) = \frac{1}{K}\sum_{k=1}^{K}\mathbb{P}_{n,k}\left(m_i(\gamma_0;\hat{\eta}_k,\hat{\mu}_k) - m_i(\gamma_0;\eta_0,\mu_0)\right)$$

Also define the empirical process as follows: let $W_k = \{Y_{ij}, X_{ij}, Q_{ij} : i \in \mathcal{I}_k, j \in \mathcal{J}\}$ and $W_{-k} = \{Y_{ij}, X_{ij}, Q_{ij} : i \in \mathcal{I}_{-k}, j \in \mathcal{J}\}$ (or $W_k = \{Y_{ij}, X_j, L_{ij}, Q_j : i \in \mathcal{I}_k, j \in \mathcal{J}\}$ and $W_{-k} = \{Y_{ij}, X_j, L_{ij}, Q_j : i \in \mathcal{I}_{-k}, j \in \mathcal{J}\}$ in the BLP setting).

$$\mathbb{G}_{n,k}m_i = \sqrt{N/K}\left(\mathbb{P}_{n,k}m_i - \mathbb{E}\left[m_i \mid W_{-k}\right]\right)$$

Then, $\mathbb{P}_{n,k}\left(m_i(\gamma_0;\hat{\eta}_k,\hat{\mu}_k) - m_i(\gamma_0;\eta_0,\mu_0)\right) = (N/K)^{-1/2}(A_{1,k} + A_{2,k})$ where

$$A_{1,k} := \mathbb{G}_{n,k}\left(m_i(\gamma_0;\hat{\eta}_k,\hat{\mu}_k) - m_i(\gamma_0;\eta_0,\mu_0)\right)$$
$$A_{2,k} := \sqrt{N/K}\left(\mathbb{E}\left[m_i(\gamma_0;\hat{\eta}_k,\hat{\mu}_k) \mid W_{-k}\right] - \mathbb{E}\left[m_i(\gamma_0;\eta_0,\mu_0)\right]\right)$$

Later, in Step 2, we will show that

$$\|A_{1,k}\|_2 = O_P(J^2(1+s_\mu)r_{N,\eta} + Jr_{N,\mu})$$

In Step 3, we will show that

$$\|A_{2,k}\|_2 = \sqrt{N/K} O_P(J^3(1+s_\mu)r_{N,\eta}^2 + J^2 r_{N,\mu} r_{N,\eta})$$

As $K \in \mathbb{N}$ is finite and fixed, combine these two rates and we obtain

$$\begin{aligned}\sqrt{N}\|M_{NJK}(\gamma_0;\hat{\eta},\hat{\mu}) - M_{NJK}(\gamma_0;\eta_0,\mu_0)\|_2 &= \|A_{1,k}\|_2 + \|A_{2,k}\|_2 \\ &= O_P\left(J^2(1+s_\mu)r_{N,\eta} + J r_{N,\mu} + \sqrt{N}J^3(1+s_\mu)r_{N,\eta}^2 + \sqrt{N}J^2 r_{N,\mu} r_{N,\eta}\right)\end{aligned}$$

Next, we derive the rate for $\|D_{\gamma'} M_{NJK}(\gamma_0;\hat{\eta},\hat{\mu}) - D_{\gamma'} M_{NJK}(\gamma_0;\eta_0,\mu_0)\|_F$. This part is relatively simple to prove regardless of cross-fitting technique. Let

$$A_{3,k} := \frac{\partial^2}{\partial\gamma\partial\gamma'} L_{NJ}^{(k)}(\gamma_0;\hat{\eta}_k) - \frac{\partial^2}{\partial\gamma\partial\gamma'} L_{NJ}^{(k)}(\gamma_0;\eta_0) \quad \text{and} \quad A_{4,k} := \mu_0 \frac{\partial^2}{\partial\eta\partial\gamma'} L_{NJ}^{(k)}(\gamma_0;\eta_0) - \hat{\mu}_k \frac{\partial}{\partial\eta\partial\gamma'} L_{NJ}^{(k)}(\gamma_0;\hat{\eta}_k)$$

In the following Step 4, we will prove that

$$\frac{1}{N/K}\|A_{3,k}\|_F = O_P\left(J^3 r_{N,\eta}\right)$$

and finally in Step 5,

$$\frac{1}{N/K}\|A_{4,k}\|_F = O_P\left(s_\mu J^3 r_{N,\eta} + J^2 r_{N,\mu} + J^3 r_{N,\mu} r_{N,\eta}\right)$$

and therefore,

$$\|D_{\gamma'} M_{NJK}(\gamma_0;\hat{\eta},\hat{\mu}) - D_{\gamma'} M_{NJK}(\gamma_0;\eta_0,\mu_0)\|_F = O_P\left((1+s_\mu)J^3 r_{N,\eta} + J^2 r_{N,\mu} + J^3 r_{N,\eta} r_{N,\mu}\right)$$

Finally, we derive the rate of $D_\gamma^2 M_{NJ,j}(\gamma_0)$ in Step 6, which is shown as $O_P((1+s_\mu)J^3)$.

As a result, to have $\sqrt{N}\|M_{NJK}(\gamma_0;\hat{\eta},\hat{\mu}) - M_{NJK}(\gamma_0;\eta_0,\mu_0)\|_2 = o_P(1)$, we need

$$J^2(1+s_\mu)r_{N,\eta} = o_P(1), \quad J r_{N,\mu} = o_P(1), \quad \sqrt{N}J^3(1+s_\mu)r_{N,\eta}^2 = o_P(1) \quad \text{and} \quad \sqrt{N}J^2 r_{N,\mu} r_{N,\eta} = o_P(1)$$

To have $\|D_{\gamma'} M_{NJK}(\gamma_0;\hat{\eta},\hat{\mu}) - D_{\gamma'} M_{NJK}(\gamma_0;\eta_0,\mu_0)\|_F = o_P(1)$, we need

$$(1+s_\mu)J^3 r_{N,\eta} = o_P(1), \quad J^2 r_{N,\mu} = o_P(1) \quad \text{and} \quad J^3 r_{N,\eta} r_{N,\mu} = o_P(1)$$

To have $D_\gamma^2 M_{NJ,j}(\gamma_0) = o_P(N^{1/2})$, we need

$$N^{-1/2}(1 + s_\mu)J^3 = o_P(1)$$

In Step 7 and 8, we will prove that

$$\sqrt{N} M_{NJK}(\gamma_0; \eta_0, \mu_0) \to_d M := N(0, \Sigma_M) \quad \text{and} \quad D_{\gamma'} M_{NJK}(\gamma_0; \eta_0, \mu_0) \to_p \Omega_M$$

Remark 2. Intuitively, we can also do the following decomposition for $M_{NJK}(\gamma_0; \hat{\eta}, \hat{\mu}) - M_{NJK}(\gamma_0; \eta_0, \mu_0)$: for each $k = 1, \ldots, K$, let

$$\begin{aligned} I_{1,k} &:= \frac{\partial}{\partial \gamma'} L_{NJ}^{(k)}(\gamma_0; \hat{\eta}_k) - \frac{\partial}{\partial \gamma'} L_{NJ}^{(k)}(\gamma_0; \eta_0) \\ I_{2,k} &:= \mu_0 \frac{\partial}{\partial \eta'} L_{NJ}^{(k)}(\gamma_0; \eta_0) - \hat{\mu}_k \frac{\partial}{\partial \eta'} L_{NJ}^{(k)}(\gamma_0; \hat{\eta}_k) \end{aligned}$$

By the triangle inequality of the l_2-norm and the Frobenius norm,

$$\|M_{NJK}(\gamma_0; \hat{\eta}, \hat{\mu}) - M_{NJK}(\gamma_0; \eta_0, \mu_0)\|_2 \leq \frac{1}{K} \sum_{k=1}^{K} \frac{1}{N/K} \left(\|I_{1,k}\|_2 + \|I_{2,k}\|_2 \right)$$

so it suffices to derive bounds for $I_{1,k}$ and $I_{2,k}$, e.g., using Taylor's expansion at $\eta = \eta_0$. However, it is necessary to have $I_{1,k} \vee I_{2,k} = o_P(N^{-1/2})$ which is very strong. The proof could be quite difficult because such decomposition may not use the fact that $I_{1,k} - I_{2,k}$ is the different of score functions whose mean will concentrate at zero.

Step 2 We derive the rate of $\|A_{1,k}\|_2$ in this part. According to Lemma 6,

$$\begin{aligned} \mathbb{E}[\|A_{1,k}\|_2^2 \mid W_{-k}] &\leq \mathbb{E}[\|m_i(\gamma_0; \hat{\eta}_k, \hat{\mu}_k) - m_i(\gamma_0; \eta_0, \mu_0)\|_2^2 \mid W_{-k}] \\ &\leq \sup_{\eta \in T_N^\eta, \mu \in T_N^\mu} \mathbb{E}[\|m_i(\gamma_0; \eta, \mu) - m_i(\gamma_0; \eta_0, \mu_0)\|_2^2 \mid W_{-k}] \\ &= \sup_{\eta \in T_N^\eta, \mu \in T_N^\mu} \mathbb{E}[\|m_i(\gamma_0; \eta, \mu) - m_i(\gamma_0; \eta_0, \mu_0)\|_2^2] \end{aligned}$$

Also, by the Lemma 6.1 in *Chernozhukov et al. (2018)*, $\mathbb{E}[\|A_{1,k}\|_2^2 \mid W_{-k}] \leq r_{1,k}^2$ implies $\|A_{1,k}\|_2 = O_P(r_{1,k})$. Thus, it suffices to provide the rate of $\|m_i(\gamma_0; \eta, \mu) - m_i(\gamma_0; \eta_0, \mu_0)\|_2$.

By the definition of the log-likelihood function,

$$m_i(\gamma_0;\eta,\mu) - m_i(\gamma_0;\eta_0,\mu_0) = \sum_{j=0}^{J} Y_{ij}\bigg(\frac{\partial}{\partial\gamma'}\ln s_{ij}(\gamma_0,\eta) - \mu\frac{\partial}{\partial\eta'}\ln s_{ij}(\gamma_0,\eta) - \frac{\partial}{\partial\gamma'}\ln s_{ij}(\gamma_0,\eta_0) + \mu_0\frac{\partial}{\partial\eta'}\ln s_{ij}(\gamma_0,\eta_0)\bigg)$$

which is a d_γ-by-1 vector. To avoid of the tensor notation, we consider its l-th coordinate without loss of generality. Let

$$B_{1,k,l} := \frac{\partial}{\partial\gamma_l}\ln s_{ij}(\gamma_0,\eta) - \frac{\partial}{\partial\gamma_l}\ln s_{ij}(\gamma_0,\eta_0)$$
$$B_{2,k,l} := \mu_{0,l}\frac{\partial}{\partial\eta'}\ln s_{ij}(\gamma_0,\eta_0) - \mu_l\frac{\partial}{\partial\eta'}\ln s_{ij}(\gamma_0,\eta)$$

We first derive the rate for $B_{1,k,l}$. Consider the Taylor's expansion of $\frac{\partial}{\partial\gamma_l}\ln s_{ij}(\gamma_0,\eta)$ (which is at least two-times continuously differentiable everywhere) at $\eta = \eta_0$:

$$\frac{\partial}{\partial\gamma_l}\ln s_{ij}(\gamma_0,\eta) = \frac{\partial}{\partial\gamma_l}\ln s_{ij}(\gamma_0,\eta_0) + (\eta-\eta_0)'\frac{\partial^2}{\partial\eta'\partial\gamma_l}\ln s_{ij}(\gamma_0,\eta_0) + o(\|\eta-\eta_0\|_2)$$

where the remainder is by the Taylor's approximation theorem. According to Lemma 3, the second term on the right-hand side is bounded almost surely, then by the triangle inequality,

$$\left\|\frac{\partial^2}{\partial\eta'\partial\gamma_l}\ln s_{ij}(\gamma_0,\eta_0)\right\|_\infty = \max_j s_{ij}^{-2}(\gamma_0,\eta_0)\left\|\frac{\partial}{\partial\eta'}s_{ij}(\gamma_0,\eta_0)\right\|_\infty + \max_j s_{ij}^{-1}(\gamma_0,\eta_0)\left\|\frac{\partial}{\partial\eta'}\frac{\partial}{\partial\gamma_l}s_{ij}(\gamma_0,\eta_0)\right\|_\infty$$

where the first term is bounded by $2C_J^{-2}J^2C_{Data}$ and the second term is bounded by $6C_J^{-1}JC_{Data}^2$. Thus, with probability one,

$$\left\|\frac{\partial^2}{\partial\eta'\partial\gamma_l}\ln s_{ij}(\gamma_0,\eta_0)\right\|_\infty \le 2C_J^{-2}J^2C_{Data} + 6C_J^{-1}JC_{Data}^2 = O(J^2)$$

which is also true for $\partial^2/\partial\eta\partial\eta'$. Then, by the Hölder's inequality, almost surely we have

$$|B_{1,k,l}| = \left|\frac{\partial}{\partial\gamma_l}\ln s_{ij}(\gamma_0,\eta) - \frac{\partial}{\partial\gamma_l}\ln s_{ij}(\gamma_0,\eta_0)\right| \le \|\eta-\eta_0\|_1 \times \left\|\frac{\partial^2}{\partial\eta'\partial\gamma_l}\ln s_{ij}(\gamma_0,\eta_0)\right\|_\infty + o(\|\eta-\eta_0\|_2)$$
$$= O\left(J^2\|\eta-\eta_0\|_1\right)$$

It is worth noting that this rate holds for not only all l's but also

$$\frac{\partial}{\partial \eta'} \ln s_{ij}(\gamma_0, \eta) - \frac{\partial}{\partial \eta'} \ln s_{ij}(\gamma_0, \eta_0)$$

Next, we derive the rate for $B_{2,k,l}$. Notice that[18]

$$\begin{aligned} B_{2,k,l} = \mu_{0,l} \frac{\partial}{\partial \eta'} \ln s_{ij}(\gamma_0, \eta_0) - \mu_l \frac{\partial}{\partial \eta'} \ln s_{ij}(\gamma_0, \eta) = \mu_{0,l} \left(\frac{\partial}{\partial \eta'} \ln s_{ij}(\gamma_0, \eta_0) - \frac{\partial}{\partial \eta'} \ln s_{ij}(\gamma_0, \eta) \right) \\ + (\mu_{0,l} - \mu_l) \left(\frac{\partial}{\partial \eta'} \ln s_{ij}(\gamma_0, \eta) - \frac{\partial}{\partial \eta'} \ln s_{ij}(\gamma_0, \eta_0) \right) \\ + (\mu_0 - \mu_l) \frac{\partial}{\partial \eta'} \ln s_{ij}(\gamma_0, \eta_0) \end{aligned}$$

By the triangle inequality, the Hölder's inequality and the sub-multiplicativity of Frobenius norm,

$$\begin{aligned} |B_{2,k,l}| \leq \|\mu_{0,l}\|_1 \times \left\| \frac{\partial}{\partial \eta'} \ln s_{ij}(\gamma_0, \eta_0) - \frac{\partial}{\partial \eta'} \ln s_{ij}(\gamma_0, \eta) \right\|_\infty \\ + \|\mu_{0,l} - \mu_l\|_1 \times \left\| \frac{\partial}{\partial \eta'} \ln s_{ij}(\gamma_0, \eta_0) - \frac{\partial}{\partial \eta'} \ln s_{ij}(\gamma_0, \eta) \right\|_\infty \\ + \|\mu_{0,l} - \mu_l\|_1 \times \left\| \frac{\partial}{\partial \eta'} \ln s_{ij}(\gamma_0, \eta_0) \right\|_\infty \end{aligned}$$

Recall that we just proved

$$\left\| \frac{\partial}{\partial \eta'} \ln s_{ij}(\gamma_0, \eta_0) - \frac{\partial}{\partial \eta'} \ln s_{ij}(\gamma_0, \eta) \right\|_\infty = O\left(J^2 \|\eta - \eta_0\|_1 \right)$$

almost surely. In addition, by Lemma 3 again, for any $l = 1, \ldots, d_\eta$,

$$\left| \frac{\partial}{\partial \eta_l} \ln s_{ij}(\gamma_0, \eta_0) \right| \leq \max_j s_{ij}^{-1}(\gamma_0, \eta_0) \times \left| \frac{\partial}{\partial \eta_l} s_{ij}(\gamma_0, \eta_0) \right| \leq 2C_J^{-1} J C_{Data} = O(J)$$

almost surely, so

$$|B_{2,k,l}| \leq O\left(J^2 (\|\mu_{0,l}\|_1 + \|\mu_{0,l} - \mu_l\|_1) \|\eta - \eta_0\|_1 + J \|\mu_{0,l} - \mu_l\|_1 \right)$$

[18] We use the fact that $AB - CD = A(B - D) + (A - C)(D - B) + (A - C)B$.

Therefore, with probability one,

$$\|m_i(\gamma_0;\eta,\mu) - m_i(\gamma_0;\eta_0,\mu_0)\|_2 \leq O\left(J^2(1+\|\mu_0\|_1+\|\mu_0-\mu\|_1)\|\eta-\eta_0\|_1 + J\|\mu_0-\mu\|_1\right)$$

where $\|\mu_0\|_1 = \sum_{l=1}^{d_\gamma}\|\mu_{0,l}\|_1$. Given the definition of T_N^η and T_N^μ and that μ_0 is sparse, we know $\|\mu_0\|_1 = O(s_\mu)$, $\|\eta-\eta_0\|_1 \leq r_{N,\eta}$ and $\|\mu-\mu_0\|_1 \leq r_{N,\mu}$, so

$$\|A_{1,k}\|_2 = O_P(J^2(1+s_\mu)r_{N,\eta} + Jr_{N,\mu})$$

Step 3 We derive the rate of $\|A_{2,k}\|_2$ in this part. Let

$$f_k(r) := \mathbb{E}\left[m_i(\gamma_0;\eta_0 + r(\hat\eta_k-\eta_0), \mu_0 + r(\hat\mu_k-\mu_0)) \mid W_{-k}\right] - \mathbb{E}\left[m_i(\gamma_0;\eta_0,\mu_0)\right], \quad r\in[0,1]$$

so $A_{2,k} = \sqrt{N/K}f_k(1)$ and $f_k(1) = f_k(0) + f_k'(0) + f_k''(r)/2$ for some $\tilde r \in (0,1)$ by the Taylor's expansion. Here $f_k(0) = 0$ because the score has zero expectation, and $f_k'(0) = 0$ because of the Neyman orthogonality. The third term

$$\begin{aligned} f_k''(r) = \mathbb{E}\Bigg[\sum_j \Upsilon_{ij}\frac{\partial^2}{\partial r^2}\Bigg(&\frac{\partial}{\partial\gamma'}\ln s_{ij}(\gamma_0,\eta_0+r(\hat\eta_k-\eta_0)) \\ &- (\mu_0+r(\hat\mu_k-\mu_0))\frac{\partial}{\partial\eta'}\ln s_{ij}(\gamma_0,\eta_0+r(\hat\eta_k-\eta_0))\Bigg) \mid W_{-k}\Bigg] \end{aligned}$$

Since $\hat\eta_k$ and $\hat\mu_k$ are known conditioning on W_{-k}, we temporarily use η and μ without ambiguity. To avoid the tensor notation, consider a single $l = 1,\dots,d_\gamma$ and let $\hat s_{ij} := s_{ij}(\gamma_0,\eta_0+r(\eta-\eta_0))$ to save notations:

$$\begin{aligned} \frac{\partial^2}{\partial r^2}\frac{\partial}{\partial\gamma_l}\ln\hat s_{ij} = {}& 2\hat s_{ij}^{-3}\left(\frac{\partial}{\partial\eta}\hat s_{ij}\cdot(\eta-\eta_0)\right)^2\frac{\partial}{\partial\gamma_l}\hat s_{ij} - \hat s_{ij}^{-2}(\eta-\eta_0)'\left(\frac{\partial^2}{\partial\eta\partial\eta'}\hat s_{ij}\right)(\eta-\eta_0)\frac{\partial}{\partial\gamma_l}\hat s_{ij} \\ &- \hat s_{ij}^{-2}\left(\frac{\partial}{\partial\eta}\hat s_{ij}\right)(\eta-\eta_0)\left(\frac{\partial^2}{\partial\eta\partial\gamma_l}\hat s_{ij}\right)(\eta-\eta_0) + \hat s_{ij}^{-1}(\eta-\eta_0)'\frac{\partial^3}{\partial\eta\partial\eta'\partial\gamma_l}\hat s_{ij}(\eta-\eta_0) \end{aligned}$$

By Assumption 7, $\hat s_{ij}^{-1} \leq C_J^{-1}J$ for all $\eta\in T_N^\eta$ and j. In addition, the first-, second- and the third-order partial derivatives are all bounded[19] almost surely for all $l = 1,\dots,d_\gamma$, then

[19] In fact, for all $\frac{\partial}{\partial\theta}s_{ij}(\cdot)$, $\frac{\partial^2}{\partial\theta\partial\theta'}s_{ij}(\cdot)$ and $\frac{\partial^3}{\partial\theta_k\partial\theta_l\partial\theta_m}s_{ij}(\cdot)$ because the covariates all have finite supports, see our Lemma 3.

for every $l = 1, \ldots, d_\gamma$,

$$\frac{\partial^2}{\partial r^2}\frac{\partial}{\partial \gamma_l}\ln \hat{s}_{ij} = O(J^3\|\eta - \eta_0\|_2^2), \quad a.s.$$

This rate also holds for $\frac{\partial^2}{\partial r^2}\frac{\partial}{\partial \eta_l}\ln \hat{s}_{ij}$ for similar reasons. It is the same logic to derive the other term. Let $\tilde{\mu}_{k,l} = \mu_{0,l} + r(\mu_l - \mu_{0,l})$, then

$$\frac{\partial^2}{\partial r^2}\tilde{\mu}_{k,l}\frac{\partial}{\partial \eta'}\ln \hat{s}_{ij} = 2(\mu_l - \mu_{0,l})\frac{\partial^2}{\partial r \partial \eta'}\ln \hat{s}_{ij} + \tilde{\mu}_{k,l}\frac{\partial^2}{\partial r^2}\frac{\partial}{\partial \eta}\ln \hat{s}_{ij}$$

where, with probability one,

$$\frac{\partial^2}{\partial r \partial \eta'}\ln \hat{s}_{ij} = -\hat{s}_{ij}^2\left(\frac{\partial \hat{s}_{ij}}{\partial \eta}\right)(\eta - \eta_0)\frac{\partial \hat{s}_{ij}}{\partial \eta} + \hat{s}_{ij}^{-1}\left(\frac{\partial^2}{\partial \eta \partial \eta'}\hat{s}_{ij}\right)(\eta - \eta_0) = O(J^2\|\eta - \eta_0\|_1)$$

$$\left|\tilde{\mu}_{k,l}\frac{\partial^2}{\partial r^2}\frac{\partial}{\partial \eta}\ln \hat{s}_{ij}\right| \leq \|\tilde{\mu}_{k,l}\|_1 \times \left\|\frac{\partial^2}{\partial r^2}\frac{\partial}{\partial \eta}\ln \hat{s}_{ij}\right\|_\infty = O(J^3\|\tilde{\mu}_{k,l}\|_1\|\eta - \eta_0\|_2^2)$$

Therefore, for every $l = 1, \ldots, d_\gamma$,

$$\mathbb{E}\left[\frac{\partial^2}{\partial r^2}\frac{\partial}{\partial \gamma_l}\ln \hat{s}_{ij} - \frac{\partial^2}{\partial r^2}\tilde{\mu}_{k,l}\frac{\partial}{\partial \eta'}\ln \hat{s}_{ij}\right] = O\left(J^3\left(1 + \|\tilde{\mu}_{k,l}\|_1\right)\|\eta - \eta_0\|_2^2 + J^2\|\mu - \mu_{0,l}\|_1\|\eta - \eta_0\|_1\right)$$

and hence,

$$\begin{aligned}
A_{2,k} \leq \sqrt{N/K} \sup_{r\in(0,1),\eta\in T_N^\eta,\mu\in T_N^\mu} \mathbb{E}\Bigg\{ \sum_{j=0}^{J} Y_{ij}\Bigg(&\frac{\partial^2}{\partial r^2}\frac{\partial}{\partial \gamma}\ln s_{ij}(\gamma_0, \eta_0 + r(\eta - \eta_0)) \\
&- (\mu_0 + r(\mu - \mu_0))\frac{\partial^2}{\partial r^2}\frac{\partial}{\partial \eta}\ln s_{ij}(\gamma_0, \eta_0 + r(\eta - \eta_0))\Bigg)\Bigg\} \\
\leq \sqrt{N/K} \sup_{r\in(0,1),\eta\in T_N^\eta,\mu\in T_N^\mu} O\left(J^3\left(1 + \|\mu_0\|_1 + r\|\mu - \mu_0\|_1\right)\|\eta - \eta_0\|_2^2 + J^2\|\mu - \mu_0\|_1\|\eta - \eta_0\|_1\right)
\end{aligned}$$

The inequality is element-wise. Since $T_N^\eta = \{\eta : \|\eta - \eta_0\|_1 \vee \|\eta - \eta_0\|_2 \leq r_{N,\eta}\}$ and $T_N^\mu = \{\mu : \|\mu - \mu_0\|_1 \leq r_{N,\mu}\}$ with $r_{N,\eta}, r_{N,\mu} \to 0$, then

$$A_{2,k} = \sqrt{N/K} \times O_P(J^3(1 + s_\mu)r_{N,\eta}^2 + J^2 r_{N,\mu} r_{N,\eta})$$

Step 4 In this part, we derive the bound on $A_{3,k}$. Since

$$A_{3,k} = \frac{\partial^2}{\partial\gamma\partial\gamma'} L_{NJ}^{(k)}(\gamma_0;\hat{\eta}_k) - \frac{\partial^2}{\partial\gamma\partial\gamma'} L_{NJ}^{(k)}(\gamma_0;\eta_0) = \sum_{i\in\mathcal{I}_k}\sum_{j=0}^{J} Y_{ij}\left(\frac{\partial^2}{\partial\gamma\partial\gamma'}\ln s_{ij}(\gamma_0,\hat{\eta}_k) - \frac{\partial^2}{\partial\gamma\partial\gamma'}\ln s_{ij}(\gamma_0,\eta_0)\right)$$

and γ has fixed dimensions, without loss of generality, consider an entry of the matrix. For each $l, m = 1,\ldots,d_\gamma$, consider the Taylor's expansion of $\frac{\partial^2}{\partial\gamma_l\partial\gamma_m}\ln s_{ij}(\gamma_0,\eta)$ at $\eta = \eta_0$:

$$\frac{\partial^2}{\partial\gamma_l\partial\gamma_m}\ln s_{ij}(\gamma_0,\eta) = \frac{\partial^2}{\partial\gamma_l\partial\gamma_m}\ln s_{ij}(\gamma_0,\eta_0) + (\eta-\eta_0)'\frac{\partial}{\partial\eta'}\frac{\partial^2}{\partial\gamma_l\partial\gamma_m}\ln s_{ij}(\gamma_0,\eta_0) + o(\|\eta-\eta_0\|_2)$$

Thus, it is sufficient to study the third-order partial derivatives. According to Lemma 3,

$$\begin{aligned}\frac{\partial}{\partial\eta'}\frac{\partial^2}{\partial\gamma_l\partial\gamma_m}\ln s_{ij}(\gamma_0,\eta) = {} & 2s_{ij}^{-3}(\gamma_0,\eta)\frac{\partial s_{ij}(\gamma_0,\eta)}{\partial\eta'}\frac{\partial s_{ij}(\gamma_0,\eta)}{\partial\gamma_l}\frac{\partial s_{ij}(\gamma_0,\eta)}{\partial\gamma_m} \\ & - s_{ij}^{-2}(\gamma_0,\eta)\times\left(\frac{\partial^2 s_{ij}(\gamma_0,\eta)}{\partial\eta'\partial\gamma_l}\frac{\partial s_{ij}(\gamma_0,\eta)}{\partial\gamma_m} + \frac{\partial^2 s_{ij}(\gamma_0,\eta)}{\partial\eta'\partial\gamma_m}\frac{\partial s_{ij}(\gamma_0,\eta)}{\partial\gamma_l}\right. \\ & \left. + \frac{\partial^2 s_{ij}(\gamma_0,\eta)}{\partial\gamma_m\partial\gamma_l}\frac{\partial s_{ij}(\gamma_0,\eta)}{\partial\eta'}\right) \\ & + s_{ij}^{-1}(\gamma_0,\eta)\frac{\partial^3 s_{ij}(\gamma_0,\eta)}{\partial\eta'\partial\gamma_l\partial\gamma_m}\end{aligned}$$

and all the partial derivatives are bounded almost surely: for any $l, m = 1,\ldots,d_\gamma$,

$$\begin{aligned}\frac{\partial s_{ij}(\gamma_0,\eta)}{\partial\eta'}\frac{\partial s_{ij}(\gamma_0,\eta)}{\partial\gamma_l}\frac{\partial s_{ij}(\gamma_0,\eta)}{\partial\gamma_m} &\le 8C_{Data}^3 \\ \frac{\partial^2 s_{ij}(\gamma_0,\eta)}{\partial\eta'\partial\gamma_l}\frac{\partial s_{ij}(\gamma_0,\eta)}{\partial\gamma_m}, \frac{\partial^2 s_{ij}(\gamma_0,\eta)}{\partial\eta'\partial\gamma_m}\frac{\partial s_{ij}(\gamma_0,\eta)}{\partial\gamma_l}, \frac{\partial^2 s_{ij}(\gamma_0,\eta)}{\partial\gamma_m\partial\gamma_l}\frac{\partial s_{ij}(\gamma_0,\eta)}{\partial\eta'} &\le 12C_{Data}^3 \\ \frac{\partial^3 s_{ij}(\gamma_0,\eta)}{\partial\eta'\partial\gamma_l\partial\gamma_m} &\le 21C_{Data}^3\end{aligned}$$

where the inequalities are element-wise. Therefore,

$$\left\|\frac{\partial}{\partial\eta'}\frac{\partial^2}{\partial\gamma_l\partial\gamma_m}\ln s_{ij}(\gamma_0,\eta)\right\|_\infty \le \left(16C_J^{-3}J^3 + 36C_J^{-2}J^2 + 21C_J^{-1}J\right)C_{Data}^3 = O_P(J^3)$$

and by the Hölder's inequality,

$$\left|\frac{\partial^2}{\partial\gamma_l\partial\gamma_m}\ln s_{ij}(\gamma_0,\eta) - \frac{\partial^2}{\partial\gamma_l\partial\gamma_m}\ln s_{ij}(\gamma_0,\eta_0)\right| \le O_P(J^3\|\eta-\eta_0\|_1)$$

Hence, by the triangle inequality,

$$\|A_{3,k}\|_F \leq \sum_{i\in\mathcal{I}_k}\sum_{j=0}^{J} Y_{ij} d_\gamma^2 O_P(J^3\|\hat{\eta}_k - \eta_0\|_1) = O_P\left(\frac{NJ^3 r_{N,\eta}}{K}\right)$$

Step 5 We derive the bound for $A_{4,k}$, which is a combination of our Step 2 and 4. The same plus-and-minus technique can be applied such that

$$\begin{aligned}
A_{4,k} = \mu_0 \frac{\partial^2}{\partial\eta\partial\gamma'} L_{NJ}^{(k)}(\gamma_0;\eta_0) - \hat{\mu}_k \frac{\partial}{\partial\eta\partial\gamma'} L_{NJ}^{(k)}(\gamma_0;\hat{\eta}_k) \\
&= \mu_0 \left(\frac{\partial^2}{\partial\eta\partial\gamma'} L_{NJ}^{(k)}(\gamma_0;\eta_0) - \frac{\partial}{\partial\eta\partial\gamma'} L_{NJ}^{(k)}(\gamma_0;\hat{\eta}_k)\right) \\
&\quad + (\mu_0 - \hat{\mu}_k)\left(\frac{\partial}{\partial\eta\partial\gamma'} L_{NJ}^{(k)}(\gamma_0;\hat{\eta}_k) - \frac{\partial^2}{\partial\eta\partial\gamma'} L_{NJ}^{(k)}(\gamma_0;\eta_0)\right) \\
&\quad + (\mu_0 - \hat{\mu}_k)\frac{\partial^2}{\partial\eta\partial\gamma'} L_{NJ}^{(k)}(\gamma_0;\eta_0)
\end{aligned}$$

Without loss of generality, for $l = 1,\ldots,d_\gamma$, let

$$\begin{aligned}
B_{4,k,l}^{(1)} &:= \mu_{0,l}\left(\frac{\partial^2}{\partial\eta\partial\gamma'} L_{NJ}^{(k)}(\gamma_0;\eta_0) - \frac{\partial}{\partial\eta\partial\gamma'} L_{NJ}^{(k)}(\gamma_0;\hat{\eta}_k)\right) \\
B_{4,k,l}^{(2)} &:= (\mu_{0,l} - \hat{\mu}_{k,l})\left(\frac{\partial}{\partial\eta\partial\gamma'} L_{NJ}^{(k)}(\gamma_0;\hat{\eta}_k) - \frac{\partial^2}{\partial\eta\partial\gamma'} L_{NJ}^{(k)}(\gamma_0;\eta_0)\right) \\
B_{4,k,l}^{(3)} &:= (\mu_{0,l} - \hat{\mu}_{k,l})\frac{\partial^2}{\partial\eta\partial\gamma'} L_{NJ}^{(k)}(\gamma_0;\eta_0)
\end{aligned}$$

Similar to Step 4,

$$\left|\frac{\partial^2}{\partial\eta_q\partial\gamma_m}\ln s_{ij}(\gamma_0,\eta) - \frac{\partial^2}{\partial\eta_q\partial\gamma_m}\ln s_{ij}(\gamma_0,\eta_0)\right| \leq O_P(J^3\|\eta-\eta_0\|_1) \quad \text{for all} \quad q = 1,\ldots,d_\eta; m = 1,\ldots,d_\gamma$$

so

$$\begin{aligned}
\|B_{4,k,l}^{(1)}\|_1 + \|B_{4,k,l}^{(2)}\|_1 &\leq (\|\mu_{0,l}\|_1 + \|\mu_{0,l} - \hat{\mu}_{k,l}\|_1)\left\|\frac{\partial^2}{\partial\eta\partial\gamma'} L_{NJ}^{(k)}(\gamma_0;\eta_0) - \frac{\partial}{\partial\eta\partial\gamma'} L_{NJ}^{(k)}(\gamma_0;\hat{\eta}_k)\right\|_\infty \\
&= O_P\left(\frac{N}{K}(s_\mu + \|\mu_{0,l} - \hat{\mu}_{k,l}\|_1)J^3\|\hat{\eta}_k - \eta_0\|_1\right)
\end{aligned}$$

Note that the rate for $\frac{\partial^2}{\partial\eta'\partial\gamma_l}\ln s_{ij}(\gamma_0,\eta_0)$ is $O_P(J^2)$ for all l, so

$$\begin{aligned}\left\|\frac{\partial^2}{\partial\eta\partial\gamma'}L_{NJ}^{(k)}(\gamma_0;\eta_0)\right\|_\infty &\leq \sum_{i\in\mathcal{I}_k}\sum_{j=0}^{J}Y_{ij}\left\|\frac{\partial^2}{\partial\eta\partial\gamma'}\ln s_{ij}(\gamma_0,\eta_0)\right\|_\infty \\ &= O_P(NJ^2/K)\end{aligned}$$

and

$$\|B_{4,k,l}^{(3)}\|_1 = O_P\left(\frac{NJ^2\|\mu_{0,l}-\hat{\mu}_{k,l}\|_1}{K}\right)$$

Combine the rates for $\|B_{4,k,l}^{(1)}\|_1$, $\|B_{4,k,l}^{(2)}\|_1$ as well as $\|B_{4,k,l}^{(3)}\|_1$, then

$$A_{4,k,l} = O_P\left(\frac{N(s_\mu+\|\mu_{0,l}-\hat{\mu}_{k,l}\|_1)J^3\|\hat{\eta}_k-\eta_0\|_1 + NJ^2\|\mu_{0,l}-\hat{\mu}_{k,l}\|_1}{K}\right)$$

as d_γ is fixed and finite. Hence,

$$\|A_{4,k}\|_F = \|A_{4,k,l}\|_1 = (N/K)O_P\left(J^3(s_\mu+r_{N,\mu})r_{N,\eta}+J^2r_{N,\mu}\right)$$

Step 6 We study the d_γ-by-d_γ matrix $D_\gamma^2 M_{NJ,q}(\gamma_0)$ which includes the third-order derivatives of $\ln s_{ij}(\gamma_0,\eta_0)$. For each $l,m=1,\ldots,d_\gamma$,

$$\begin{aligned}\frac{\partial^2}{\partial\gamma_l\partial\gamma_m}M_{NJ,q}(\gamma_0) =& \frac{1}{K}\sum_{k=1}^{K}\frac{1}{N/K}\frac{\partial^2}{\partial\gamma_l\partial\gamma_m}\frac{\partial}{\partial\gamma_q}L_{NJ}^{(k)}(\gamma_0;\eta_0) - \frac{\partial^2}{\partial\gamma_l\partial\gamma_m}\mu_{0,q}\frac{\partial}{\partial\eta'}L_{NJ}^{(k)}(\gamma_0;\eta_0) \\ =& \frac{1}{K}\sum_{k=1}^{K}\frac{1}{N/K}\sum_{i\in\mathcal{I}_k}\sum_{j=0}^{J}Y_{ij}\left(\frac{\partial^3}{\partial\gamma_l\partial\gamma_m\partial\gamma_q}\ln s_{ij}(\gamma_0;\eta_0) - \mu_{0,q}\frac{\partial^2}{\partial\gamma_l\partial\gamma_m}\frac{\partial}{\partial\eta'}\ln s_{ij}(\gamma_0;\eta_0)\right)\end{aligned}$$

We have discussed the rate for $\frac{\partial^3}{\partial\gamma_l\partial\gamma_m\partial\eta'}\ln s_{ij}(\gamma_0;\eta_0)$ in Step 4, which is $O_P(J^3)$, and it is identical for $\frac{\partial^3}{\partial\theta_l\partial\theta_m\partial\theta_k}\ln s_{ij}(\gamma_0;\eta_0)$. See Lemma 3. Therefore, by the triangle inequality

and Hölder's inequality,

$$\begin{aligned}\left|\frac{\partial^2}{\partial\gamma_l\partial\gamma_m}M_{NJ,q}(\gamma_0)\right| &\leq \sum_{j=0}^{J}\Upsilon_{ij}\left(\left|\frac{\partial^3}{\partial\gamma_l\partial\gamma_m\partial\gamma_q}\ln s_{ij}(\gamma_0;\eta_0)\right| + \|\mu_{0,q}\|_1 \times \left\|\frac{\partial^2}{\partial\gamma_l\partial\gamma_m}\frac{\partial}{\partial\eta'}\ln s_{ij}(\gamma_0;\eta_0)\right\|_\infty\right)\\ &\leq \sum_{j=0}^{J}\Upsilon_{ij}(O_P(J^3)+O(s_\mu J^3))\\ &= O_P\left((1+s_\mu)J^3\right)\end{aligned}$$

Therefore, $D_\gamma^2 M_{NJ,q}(\gamma_0) = O_P((1+s_\mu)J^3)$.

Step 7 In this part, we prove the asymptotic normality for

$$N^{1/2}M_{NJK}(\gamma_0;\eta_0,\mu_0) = \frac{1}{\sqrt{N}}\sum_{i=1}^{N}\sum_{j=0}^{J}\Upsilon_{ij}s_{ij}^{-1}(\gamma_0,\eta_0)\left(\frac{\partial}{\partial\gamma'}s_{ij}(\gamma_0,\eta_0)-\mu_0\frac{\partial}{\partial\eta'}s_{ij}(\gamma_0,\eta_0)\right)$$

Since we allow $J\to\infty$ as $N\to\infty$, we apply the Lindeberg-Feller central limit theorem (e.g., Proposition 2.27 in Vaart, 1998). Note that $Var(m_i(\gamma_0;\eta_0,\mu_0))$ may vary across N as $\eta_0=\eta_{N0}$ and $\mu_0=\mu_{N0}$ actually depends on N, even if $J<\infty$ is fixed. We temporarily highlight the subscript N for clarity. The Lindeberg's condition that needs to be verified in our setting is

$$\lim_{N\to\infty}\mathbb{E}\left[\|m_{Ni}(\gamma_0;\eta_{N0},\mu_{N0})\|_2^2 I[\|m_{Ni}(\gamma_0;\eta_{N0},\mu_{N0})\|_2\geq\epsilon\sqrt{N})\right]=0 \quad \text{for every } \epsilon>0$$

For any $l,q=1,\ldots,d_\gamma$,

$$\begin{aligned}\mathbb{E}[m_{Ni,l}(\gamma_0;\eta_{N0},\mu_{N0})m_{Ni,q}(\gamma_0;\eta_{N0},\mu_{N0})] = \mathbb{E}\Bigg\{&\sum_{j=0}^{J_N}\sum_{k=0}^{J_N}\Upsilon_{ij}\Upsilon_{ik}s_{ij}^{-1}(\gamma_0,\eta_{N0})s_{ik}^{-1}(\gamma_0,\eta_{N0})\\ &\left(\frac{\partial}{\partial\gamma_l}s_{ij}(\gamma_0,\eta_{N0})-\mu_{N0,l}\frac{\partial}{\partial\eta'}s_{ij}(\gamma_0,\eta_{N0})\right)\\ &\left(\frac{\partial}{\partial\gamma_q}s_{ik}(\gamma_0,\eta_{N0})-\mu_{N0,q}\frac{\partial}{\partial\eta'}s_{ik}(\gamma_0,\eta_{N0})\right)\Bigg\}\end{aligned}$$

The integrand above is bounded by $C_J^{-2}J_N^2(1+s_{\mu,N})^2 4C_{Data}^2$ almost surely because (i)

$\sum_{j=0}^{J} Y_{ij} = 1$, (ii) $\max_j s_{ij}^{-1}(\gamma_0, \eta_{N0}) \leq C_J^{-1} J_N$ and (iii)

$$\max_{j=1,\ldots,J_N} \left| \frac{\partial}{\partial \gamma_l} s_{ij}(\gamma_0, \eta_{N0}) - \mu_{N0,l} \frac{\partial}{\partial \eta'} s_{ij}(\gamma_0, \eta_{N0}) \right| \leq 2C_{Data} + \|\mu_{N0,l}\|_1 \left\| \frac{\partial}{\partial \eta'} s_{ij}(\gamma_0, \eta_{N0}) \right\|_\infty \leq 2(1+s_{\mu,N}) C_{Data}$$

Although the integrand diverges as $N \to \infty$, its rate is known as $J_N^2(1 + s_{\mu,N})^2$. As long as $J_N(1 + s_{\mu,N}) = o(\sqrt{N})$, for any $\epsilon > 0$, there exists large enough $N_\epsilon \in \mathbb{N}$ such that

$$N^{-1/2} \|m_{Ni}(\gamma_0; \eta_{N0}, \mu_{N0})\|_2 \leq \epsilon \quad \forall N > N_\epsilon$$

with probability one. Therefore, the Lindeberg's condition holds when $N^{-1/2} J_N(1+s_{\mu,N}) = o(1)$. By the Lindeberg-Feller central limit theorem, if $Var(m_{Ni}(\gamma_0; \eta_{N0}, \mu_{N0})) \to \Sigma_M$, then

$$N^{1/2} M_{NJK}(\gamma_0; \eta_{N0}, \mu_{N0}) \to_d N(0, \Sigma_M)$$

Step 8 In this part, we prove the convergence in probability for $D_{\gamma'} M_{NJK}(\gamma_0; \eta_0, \mu_0)$, which is

$$\hat{\Omega}_M := \frac{1}{N} \sum_{i=1}^{N} \sum_{j=0}^{J} Y_{ij} \left(\frac{\partial^2}{\partial \gamma \partial \gamma'} \ln s_{ij}(\gamma_0, \eta_0) - \mu_0 \frac{\partial^2}{\partial \eta \partial \gamma'} \ln s_{ij}(\gamma_0, \eta_0) \right)$$

We want to prove that $\|\hat{\Omega}_M - \Omega_M\|_F = o_P(1)$, where

$$\Omega_M = \mathbb{E} \left[\sum_{j=0}^{J} Y_{ij} \left(\frac{\partial^2}{\partial \gamma \partial \gamma'} \ln s_{ij}(\gamma_0, \eta_0) - \mu_0 \frac{\partial^2}{\partial \eta \partial \gamma'} \ln s_{ij}(\gamma_0, \eta_0) \right) \right]$$

is the population analog. Since d_γ is fixed and finite, an element-wise proof is sufficient. For any $m, l = 1, \ldots, d_\gamma$, let $\hat{\Omega}_{M,ml}$ and $\Omega_{M,ml}$ be the (m, l)-th entry of $\hat{\Omega}_M$ and Ω_M, respectively. Obviously, $\hat{\Omega}_{M,ml} - \Omega_{M,ml}$ is mean zero. Since

$$\left\| \frac{\partial^2}{\partial \eta' \partial \gamma_l} \ln s_{ij}(\gamma_0, \eta_0) \right\|_\infty \leq 2C_J^{-2} J^2 C_{Data} + 6C_J^{-1} J C_{Data}^2 \leq C' J^2$$

for some constant $C' > 0$ almost surely, where the first inequality is derived in Step 2 in the Proof of Lemma , then

$$\begin{aligned}
&\left|\sum_{j=0}^{J} Y_{ij}\left(\frac{\partial^2}{\partial\gamma_m\partial\gamma_l}\ln s_{ij}(\gamma_0,\eta_0) - \mu_{0,m}\frac{\partial^2}{\partial\eta\partial\gamma_l}\ln s_{ij}(\gamma_0,\eta_0)\right)\right| \\
\leq &\sum_{j=0}^{J} Y_{ij}\left|\frac{\partial^2}{\partial\gamma_m\partial\gamma_l}\ln s_{ij}(\gamma_0,\eta_0) - \mu_{0,m}\frac{\partial^2}{\partial\eta\partial\gamma_l}\ln s_{ij}(\gamma_0,\eta_0)\right| \\
\leq &\sum_{j=0}^{J} Y_{ij}\left(\left|\frac{\partial^2}{\partial\gamma_m\partial\gamma_l}\ln s_{ij}(\gamma_0,\eta_0)\right| + \|\mu_{0,m}\|_1 \times \left\|\frac{\partial^2}{\partial\eta\partial\gamma_l}\ln s_{ij}(\gamma_0,\eta_0)\right\|_\infty\right) \\
\leq &\sum_{j=0}^{J} Y_{ij}(1+s_\mu)C'J^2 \\
= &(1+s_\mu)C'J^2, \quad \text{almost surely}
\end{aligned}$$

where the first and second inequalities are by the triangle inequality and the Hölder's inequality, and the last equality is by $\sum_{j=0}^{J} Y_{ij} = 1$. Then, by the Hoeffding's inequality for bounded random variables, for any $t > 0$,

$$\mathbb{P}\left(N|\hat{\Omega}_{M,ml} - \Omega_{M,ml}| \geq t\right) \leq 2\exp\left(-\frac{t^2}{N(1+s_\mu)C'J^2}\right)$$

suggesting that, for any $\epsilon > 0$,

$$\mathbb{P}\left(\frac{|\hat{\Omega}_{M,ml} - \Omega_{M,ml}|}{\sqrt{\frac{C'(1+s_\mu)J^2}{N}}} \geq t\right) \leq 2\exp\left(-t^2\right) < \epsilon$$

by choosing $t > \sqrt{\log(2/\epsilon)}$. Thus, $|\hat{\Omega}_{M,ml} - \Omega_{M,ml}| = O_P(\sqrt{N^{-1}C'(1+s_\mu)J^2})$ and converges to zero in probability as long as $N^{-1}(1+s_\mu)J^2 = o(1)$, which is weaker than the condition $N^{-1}(1+s_\mu)^2J^2 = o(1)$ required by the central limit theorem.

E.5 Proof of Theorem 3

The proof is in the same logic as the Step 2, 3 and 8 in the proof of Theorem 2. It suffices to verify the first condition

$$\sup_{\gamma\in\Gamma}\left|\|M_{NJK}(\gamma;\hat{\eta}_k,\hat{\mu}_k)\|_2^2-\|\mathbb{E}[M_{NJK}(\gamma;\eta_0,\mu_0)]\|_2^2\right|\to_p 0$$

Some algebra shows that

$$\begin{aligned}\|M_{NJK}(\gamma;\hat{\eta}_k,\hat{\mu}_k)\|_2^2-&\|\mathbb{E}[M_{NJK}(\gamma;\eta_0,\mu_0)]\|_2^2\\ &=\|M_{NJK}(\gamma;\hat{\eta}_k,\hat{\mu}_k)-\mathbb{E}[M_{NJK}(\gamma;\eta_0,\mu_0)]\|_2^2\\ &\quad+2\left(M_{NJK}(\gamma;\hat{\eta}_k,\hat{\mu}_k)-\mathbb{E}[M_{NJK}(\gamma;\eta_0,\mu_0)]\right)'\mathbb{E}[M_{NJK}(\gamma;\eta_0,\mu_0)]\end{aligned}$$

By the triangle inequality,

$$\begin{aligned}\|M_{NJK}(\gamma;\hat{\eta}_k,\hat{\mu}_k)-\mathbb{E}[M_{NJK}(\gamma;\eta_0,\mu_0)]\|_2 &\le \|M_{NJK}(\gamma;\hat{\eta}_k,\hat{\mu}_k)-M_{NJK}(\gamma;\eta_0,\mu_0)\|_2\\ &\quad+\|M_{NJK}(\gamma;\eta_0,\mu_0)-\mathbb{E}[M_{NJK}(\gamma;\eta_0,\mu_0)]\|_2\end{aligned}$$

For the first term $r_1(\gamma):=\|M_{NJK}(\gamma;\hat{\eta}_k,\hat{\mu}_k)-M_{NJK}(\gamma;\eta_0,\mu_0)\|_2$, by using the same empirical process techniques, we need to derive the bounds (uniformly over Γ) on

$$\begin{aligned}A_{1,k}(\gamma)&:=\mathbb{G}_{n,k}\left(m_i(\gamma;\hat{\eta}_k,\hat{\mu}_k)-m_i(\gamma;\eta_0,\mu_0)\right)\\ A_{2,k}(\gamma)&:=\sqrt{N/K}\left(\mathbb{E}\left[m_i(\gamma;\hat{\eta}_k,\hat{\mu}_k)\mid W_{-k}\right]-\mathbb{E}\left[m_i(\gamma;\eta_0,\mu_0)\right]\right)\end{aligned}$$

which will be shown in Step 1 and 2, respectively. The second term $r_2(\gamma):=\|M_{NJK}(\gamma;\eta_0,\mu_0)-\mathbb{E}[M_{NJK}(\gamma;\eta_0,\mu_0)]\|_2$ is more difficult although it does not depend on $\hat{\eta}_k$ and $\hat{\mu}_k$. We will be derived the bound through the supremum of the empirical process in Step 3. Finally, we need to derive the rate for $r_3(\gamma):=\|\mathbb{E}[M_{NJK}(\gamma;\eta_0,\mu_0)]\|_2$ in Step 4. We will show that

$$\begin{aligned}\sup_{\gamma\in\Gamma}r_1(\gamma)&\le N^{-1/2}\left(\sup_{\gamma\in\Gamma}\|A_{1,k}(\gamma)\|_2+\sup_{\gamma\in\Gamma}\|A_{2,k}(\gamma)\|_2\right)\\ &=O_P\left(\frac{\rho_J^2(1+s_\mu)r_{N,\eta}+\rho_J r_{N,\mu}}{\sqrt{N}}+\rho_J^3(1+s_\mu)r_{N,\eta}^2+\rho_J^2 r_{N,\mu}r_{N,\eta}\right)\\ \sup_{\gamma\in\Gamma}r_2(\gamma)&=O_P\left(\frac{\rho_J^2(1+s_\mu)}{\sqrt{N}}\right)\\ \sup_{\gamma\in\Gamma}r_3(\gamma)&=O_P(\rho_J(1+s_\mu))\end{aligned}$$

Then, by the Hölder's inequality,

$$\sup_{\gamma\in\Gamma}\left|\|M_{NJK}(\gamma;\hat{\eta}_k,\hat{\mu}_k)\|_2^2-\|\mathbb{E}[M_{NJK}(\gamma;\eta_0,\mu_0)]\|_2^2\right|\leq \sup_{\gamma\in\Gamma}(r_1(\gamma)+r_2(\gamma))^2 + \sup_{\gamma\in\Gamma}(r_1(\gamma)+r_2(\gamma))\times\sup_{\gamma\in\Gamma}r_3(\gamma)$$

Given that $\rho_J = J\log N$ and $s_\mu = O(1)$, we have $\rho_J^2 = J^2(\log N)^2$ and the rates are smaller to those in Theorem 2 (because $N^{-p}\log N\to 0$ for any $p>0$). As long as the rates in Eq.(4.8) are satisfied, we have $\sup_{\gamma\in\Gamma}r_1(\gamma)+\sup_{\gamma\in\Gamma}r_2(\gamma)=o_P(1)$, so the proof is done as

$$\sup_{\gamma\in\Gamma}\left|\|M_{NJK}(\gamma;\hat{\eta}_k,\hat{\mu}_k)\|_2^2-\|\mathbb{E}[M_{NJK}(\gamma;\eta_0,\mu_0)]\|_2^2\right|\leq \sup_{\gamma\in\Gamma}r_1(\gamma)+\sup_{\gamma\in\Gamma}r_2(\gamma)=o_P(1)$$

Step 1 To show $A_{1,k}(\gamma)$, we need to work on

$$B_{1,k,l}(\gamma) := \frac{\partial}{\partial\gamma_l}\ln s_{ij}(\gamma,\eta)-\frac{\partial}{\partial\gamma_l}\ln s_{ij}(\gamma,\eta_0)$$
$$B_{2,k,l}(\gamma) := \mu_{0,l}\frac{\partial}{\partial\eta'}\ln s_{ij}(\gamma,\eta_0)-\mu_l\frac{\partial}{\partial\eta'}\ln s_{ij}(\gamma,\eta)$$

for some $k=1,\ldots,K$ and $l=1,\ldots,d_\gamma$. By the Taylor's expansion, Lemma 3 and Assumption 9,

$$\begin{aligned}\left\|\frac{\partial^2}{\partial\eta'\partial\gamma_l}\ln s_{ij}(\gamma,\eta_0)\right\|_\infty &= \max_j s_{ij}^{-2}(\gamma,\eta_0)\left\|\frac{\partial}{\partial\eta'}s_{ij}(\gamma,\eta_0)\right\|_\infty+\max_j s_{ij}^{-1}(\gamma,\eta_0)\left\|\frac{\partial}{\partial\eta'}\frac{\partial}{\partial\gamma_l}s_{ij}(\gamma,\eta_0)\right\|_\infty\\ &\leq \sup_{\gamma\in\Gamma}\max_j s_{ij}^{-2}(\gamma,\eta_0)O(C_{Data})+\sup_{\gamma\in\Gamma}\max_j s_{ij}^{-1}(\gamma,\eta_0)O(C_{Data}^2)\\ &= O(\rho_J^2)\quad\text{almost surely,}\end{aligned}$$

Then, we can similarly derive that $\sup_{\gamma\in\Gamma}|B_{1,k,l}(\gamma)| = O\left(\rho_J^2\|\eta-\eta_0\|_1\right)$ and

$$\sup_{\gamma\in\Gamma}|B_{2,k,l}(\gamma)| = O\left(\rho_J^2(\|\mu_{0,l}\|_1+\|\mu_{0,l}-\mu_l\|_1)\|\eta-\eta_0\|_1+\rho_J\|\mu_{0,l}-\mu_l\|_1\right)$$

almost surely. Hence,

$$\sup_{\gamma\in\Gamma}\|A_{1,k}(\gamma)\|_2 = O_P(\rho_J^2(1+s_\mu)r_{N,\eta}+\rho_J r_{N,\mu})$$

Step 2 Let

$$f_k(r) := \mathbb{E}\left[m_i(\gamma; \eta_0 + r(\hat{\eta}_k - \eta_0), \mu_0 + r(\hat{\mu}_k - \mu_0)) \mid W_{-k}\right] - \mathbb{E}\left[m_i(\gamma; \eta_0, \mu_0)\right], \quad r \in [0,1]$$

so $A_{2,k} = \sqrt{N/K} f_k(1)$ and $f_k(1) = f_k''(r)/2$ for some $\tilde{r} \in (0,1)$ by the Taylor's expansion. Here $\tilde{r}$ may depend on $\gamma \in \Gamma$. The second-order derivative

$$\begin{aligned} f_k''(r) = \mathbb{E}\Bigg[\sum_j Y_{ij} \frac{\partial^2}{\partial r^2} \Bigg(& \frac{\partial}{\partial \gamma'} \ln s_{ij}(\gamma, \eta_0 + r(\hat{\eta}_k - \eta_0)) \\ & - (\mu_0 + r(\hat{\mu}_k - \mu_0)) \frac{\partial}{\partial \eta'} \ln s_{ij}(\gamma, \eta_0 + r(\hat{\eta}_k - \eta_0)) \Bigg) \mid W_{-k} \Bigg] \end{aligned}$$

Now let $\tilde{s_{ij}} := s_{ij}(\gamma, \eta_0 + r(\eta - \eta_0))$. Since for any $l = 1, \ldots, d_\gamma$,

$$\begin{aligned} \frac{\partial^2}{\partial r^2} \frac{\partial}{\partial \gamma_l} \ln \tilde{s}_{ij} = {} & 2\tilde{s}_{ij}^{-3} \left(\frac{\partial}{\partial \eta} \tilde{s}_{ij} \cdot (\eta - \eta_0) \right)^2 \frac{\partial}{\partial \gamma_l} \tilde{s}_{ij} - \tilde{s}_{ij}^{-2} (\eta - \eta_0)' \left(\frac{\partial^2}{\partial \eta \partial \eta'} \tilde{s}_{ij} \right) (\eta - \eta_0) \frac{\partial}{\partial \gamma_l} \tilde{s}_{ij} \\ & - \tilde{s}_{ij}^{-2} \left(\frac{\partial}{\partial \eta} \tilde{s}_{ij} \right) (\eta - \eta_0) \left(\frac{\partial^2}{\partial \eta \partial \gamma_l} \tilde{s}_{ij} \right) (\eta - \eta_0) + \tilde{s}_{ij}^{-1} (\eta - \eta_0)' \frac{\partial^3}{\partial \eta \partial \eta' \partial \gamma_l} \tilde{s}_{ij} (\eta - \eta_0) \end{aligned}$$

then, by Lemma 3, the Hölder's inequality, the triangle inequality and Assumption 9,

$$\sup_{\gamma \in \Gamma} \frac{\partial^2}{\partial r^2} \frac{\partial}{\partial \gamma_l} \ln \tilde{s}_{ij} = O(\rho_J^3 \|\eta - \eta_0\|_2^2), \quad a.s.$$

Let $\tilde{\mu}_{k,l} = \mu_{0,l} + r(\mu_l - \mu_{0,l})$, since

$$\frac{\partial^2}{\partial r^2} \tilde{\mu}_{k,l} \frac{\partial}{\partial \eta'} \ln \tilde{s}_{ij} = 2(\mu_l - \mu_{0,l}) \frac{\partial^2}{\partial r \partial \eta'} \ln \tilde{s}_{ij} + \tilde{\mu}_{k,l} \frac{\partial^2}{\partial r^2} \frac{\partial}{\partial \eta} \ln \tilde{s}_{ij}$$

and

$$\left| \tilde{\mu}_{k,l} \frac{\partial^2}{\partial r^2} \frac{\partial}{\partial \eta} \ln \tilde{s}_{ij} \right| \leq \|\tilde{\mu}_{k,l}\|_1 \times \sup_{\gamma \in \Gamma} \left\| \frac{\partial^2}{\partial r^2} \frac{\partial}{\partial \eta} \ln \tilde{s}_{ij} \right\|_\infty \leq O(\rho_J^3 \|\tilde{\mu}_{k,l}\|_1 \|\eta - \eta_0\|_2^2) \quad a.s.,$$

then, by taking supremum over Γ on both sides,

$$\sup_{\gamma \in \Gamma} \mathbb{E}\left[\frac{\partial^2}{\partial r^2} \frac{\partial}{\partial \gamma_l} \ln \hat{s}_{ij} - \frac{\partial^2}{\partial r^2} \tilde{\mu}_{k,l} \frac{\partial}{\partial \eta'} \ln \hat{s}_{ij} \right] = O\left(\rho_J^3 \left(1 + \|\tilde{\mu}_{k,l}\|_1\right) \|\eta - \eta_0\|_2^2 + \rho_J^2 \|\mu - \mu_{0,l}\|_1 \|\eta - \eta_0\|_1 \right)$$

Therefore,

$$\sup_{\gamma\in\Gamma}\|A_{2,k}\| = \sqrt{N/K}\times O_P(\rho_J^3(1+s_\mu)r_{N,\eta}^2 + \rho_J^2 r_{N,\mu}r_{N,\eta})$$

Step 3 Recall that

$$M_{NJK}(\gamma;\eta_0,\mu_0) - \mathbb{E}[M_{NJK}(\gamma;\eta_0,\mu_0)] = \frac{1}{N}\sum_{i=1}^N m_i(\gamma;\eta_0,\mu_0) - \mathbb{E}[m_i(\gamma;\eta_0,\mu_0)]$$

It suffices to consider

$$\sup_{\gamma\in\Gamma}\left|\frac{1}{N}\sum_{i=1}^N m_{il}(\gamma;\eta_0,\mu_0) - \mathbb{E}[m_{il}(\gamma;\eta_0,\mu_0)]\right|$$

for any $l = 1,\dots,d_\gamma$.

We introduce the empirical process notations. Without loss of generality, let $f_\gamma(W_i) := m_{i1}(\gamma;\eta_0,\mu_0)$ which is a mapping from $\mathbb{R}^d$ to $\mathbb{R}$ and $d = O(J+d_\theta)$ is the dimension[20] of the data vector W_i. Also let $\mathcal{F} := \{f_\gamma(\cdot) \mid \gamma\in\Gamma\}$ and $\|\mathbb{P}_N - P\|_\mathcal{F} := \sup_{f\in\mathcal{F}}|N^{-1}\sum_{i=1}^N f(W_i) - \mathbb{E}[f(W_i)]|$. Notice that

$$\|\mathbb{P}_N - P\|_\mathcal{F} = \|\mathbb{P}_N - P\|_\Gamma := \sup_{\gamma\in\Gamma}|N^{-1}\sum_{i=1}^N f_\gamma(W_i) - \mathbb{E}[f_\gamma(W_i)]|$$

Then, by the Markov's inequality, for any $t>0$,

$$\mathbb{P}(\|\mathbb{P}_N - P\|_\Gamma > t) \le t^{-1}N^{-1/2}\mathbb{E}[\sqrt{N}\|\mathbb{P}_N - P\|_\Gamma]$$

Thus, to show the convergence in probability, it suffices to discuss the convergence of the supremum of the empirical process. Since $f_\gamma(W_i)$ is a smooth function in γ, by the Taylor's expansion and the Hölder's inequality

$$\begin{aligned}
f_{\gamma_1}(W_i) - f_{\gamma_2}(W_i) &= m_{i1}(\gamma_1;\eta_0,\mu_0) - m_{i1}(\gamma_2;\eta_0,\mu_0)\\
&= (\gamma_1-\gamma_2)'\frac{\partial}{\partial\gamma}m_{i1}(\gamma_2;\eta_0,\mu_0) + o(\|\gamma_1-\gamma_2\|_2)\\
&\le \|\gamma_1-\gamma_2\|_2 \times \left\|\sum_{j=0}^J Y_{ij}\frac{\partial^2}{\partial\gamma'\partial\gamma_1}\ln s_{ij}(\gamma_2,\eta_0) - \mu_{0,1}\frac{\partial^2}{\partial\gamma\partial\eta'}\ln s_{ij}(\gamma_2,\eta_0)\right\|_2
\end{aligned}$$

[20] $d = J + d_X + d_Q$ in the exogenous case and $d = 2J + d_X(d_L + d_Q)$ in the endogenous case.

In Step 1, we show that for every j, l and $\gamma \in \Gamma$,

$$\left\| \frac{\partial^2}{\partial \eta' \partial \gamma_l} \ln s_{ij}(\gamma, \eta_0) \right\|_\infty = O(\rho_J^2) \quad \text{almost surely}$$

The rate is the same for $\frac{\partial^2}{\partial \gamma' \partial \gamma_1} \ln s_{ij}(\gamma, \eta_0)$. Thus, for some large constant $C > 0$ (only depends on C_{Data})

$$\left\| \sum_{j=0}^{J} Y_{ij} \frac{\partial^2}{\partial \gamma' \partial \gamma_1} \ln s_{ij}(\gamma_2, \eta_0) - \mu_{0,1} \frac{\partial^2}{\partial \gamma \partial \eta'} \ln s_{ij}(\gamma_2, \eta_0) \right\|_2 \leq C(1 + s_\mu)\rho_J^2, \quad a.s.$$

and $|f_{\gamma_1}(W_i) - f_{\gamma_2}(W_i)| \leq \|\gamma_1 - \gamma_2\|_2 C(1 + s_\mu)\rho_J^2$. In other words, $f_\gamma(W_i)$ is Lipschitz in the index parameter γ with respect to the Euclidean distance $\|\cdot\|_2$. Moreover, for some constant $c_1 > 0$ (only depends on C_{Data}),

$$\begin{aligned} f_\gamma(W_i) &:= \sum_{j=0}^{J} Y_{ij} s_{ij}^{-1}(\gamma, \eta_0) \left(\frac{\partial}{\partial \gamma_l} s_{ij}(\gamma, \eta_0) - \mu_{0,l} \frac{\partial}{\partial \eta'} s_{ij}(\gamma, \eta_0) \right) \\ &\leq \sup_{\gamma \in \Gamma} \max_j s_{ij}^{-1}(\gamma, \eta_0) \times \sup_{\gamma \in \Gamma} \left| \frac{\partial}{\partial \gamma_l} s_{ij}(\gamma, \eta_0) - \mu_{0,l} \frac{\partial}{\partial \eta'} s_{ij}(\gamma, \eta_0) \right| \\ &\leq c_1 \rho_J (1 + s_\mu) \end{aligned}$$

where the third line is by the triangle inequality, the Hölder's inequality and Lemma 3. Thus, $F_1(w) := c_1 \rho_J (1 + s_\mu)$ is the envelope function. As $\rho_J \to \infty$, define the constant function $F(w) := C(1 + s_\mu)\rho_J^2 \geq F_1(w)$ with some constant $C > 0$. By Theorem 2.7.11 in Van Der Vaart and Wellner (1996), the bracketing number

$$N_{[]}(2\epsilon \|F\|_{P,2}, \mathcal{F}, \|\cdot\|_{P,2}) \leq N(\epsilon, \Gamma, \|\cdot\|_2)$$

is bounded by the covering number associated with the L_2-norm $\|X\|_{P,2} = (\int X^2 dP)^{1/2}$ and the Euclidean norm $\|\cdot\|_2$. Given the constant function $F(w)$, $\|F\|_{P,2} = C(1 + s_\mu)\rho_J^2 (\int 1 dP)^{1/2} = C(1 + s_\mu)\rho_J^2$. Since $\Gamma \subset \mathbb{R}^{d_\gamma}$ is a bounded subset with fixed dimensions, then

$$c_2 \epsilon^{-1} \leq N(\epsilon, \Gamma, \|\cdot\|_2) \leq c_3 \epsilon^{-1}$$

for some constant $0 < c_2 < 1 < c_3 < \infty$ depending on the volume of Γ, according to

Lemma 2.7 in Sen (2022). Therefore, for any $\epsilon > 0$,

$$\log N_{[]}(\epsilon, \mathcal{F}, \|\cdot\|_{P,2}) \leq \log 2c_3\|F\|_{P,2}\epsilon^{-1}$$

By the maximal inequality with bracketing (Theorem 4.12, Sen, 2022),

$$\mathbb{E}[\sqrt{N}\|\mathbb{P}_N - P\|_\Gamma] \leq c_4 J_{[]}(\|F\|_{P,2}, \mathcal{F} \cup \{0\}, \|\cdot\|_{P,2})$$

where the bracketing integral

$$\begin{aligned}
J_{[]}(\|F\|_{P,2}, \mathcal{F} \cup \{0\}, L_2(P)) &:= \int_0^{\|F\|_{P,2}} \sqrt{\log N_{[]}(\xi, \mathcal{F} \cup \{0\}, \|\cdot\|_{P,2})}d\xi \\
&\leq \int_0^{\|F\|_{P,2}} \sqrt{\log 2c_3\|F\|_{P,2}\xi^{-1}}d\xi \\
&= 2c_3\|F\|_{P,2}\int_0^{\frac{1}{2c_3}} \sqrt{\log t^{-1}}dt \\
&\leq c_5\|F\|_{P,2}
\end{aligned}$$

for some constant $c_4, c_5 > 0$. The third line is by letting $\xi = 2c_3\|F\|_{P,2}t$ (so $d\xi = 2c_3\|F\|_{P,2}dt$) and the fourth line is because $\int_0^1 \sqrt{\log t^{-1}}dt$ is a converged[21] integral (cf. Dudley's inequality, Ch.8, Vershynin, 2018). Then,

$$\mathbb{E}[\|\mathbb{P}_N - P\|_\Gamma] \leq N^{-1/2}c_4c_5\|F\|_{P,2} = O\left(\frac{(1+s_\mu)\rho_J^2}{\sqrt{N}}\right)$$

and hence,

$$\sup_{\gamma\in\Gamma}\left|\frac{1}{N}\sum_{i=1}^N m_{i1}(\gamma;\eta_0,\mu_0) - \mathbb{E}[m_{i1}(\gamma;\eta_0,\mu_0)]\right| = O_P\left(\frac{(1+s_\mu)\rho_J^2}{\sqrt{N}}\right)$$

Repeat the procedure for each coordinate $l = 1, \ldots, d_\gamma$ and obtain

$$\sup_{\gamma\in\Gamma}\|M_{NJK}(\gamma;\eta_0,\mu_0) - \mathbb{E}[M_{NJK}(\gamma;\eta_0,\mu_0)]\|_2 = O_P\left(\frac{(1+s_\mu)\rho_J^2}{\sqrt{N}}\right)$$

[21] It is known that $\lim_{x\to\infty}\ln(x)/x^p = 0$ for any $p > 0$. Equivalently, $\lim_{y\to 0}\frac{\ln(1/y)}{1/y^p} = 0$. Although $\lim_{y\to 0} 1/y^p = \infty$, $\int_0^1 1/y^p dx$ converges (so it is finite) when $p < 1$, suggesting $\int_0^1 \log\frac{1}{y}dy$ also converges. Hence, $\int_0^1 \sqrt{\log\frac{1}{y}}dy$ converges.

Remark 3. By the definition,By Assumption 9, $\sup_{\gamma\in\Gamma}\max_j s_{ij}^{-1}(\gamma,\eta_0) = O(\rho_J)$ with probability one. By ,

$$\left|\frac{\partial}{\partial\gamma_l}s_{ij}(\gamma,\eta_0) - \mu_{0,l}\frac{\partial}{\partial\eta'}s_{ij}(\gamma,\eta_0)\right| \le (1+s_\mu)C_{Data} \quad a.s.$$

Combining with the rate ρ_J, we have the score

$$|m_{il}(\gamma;\eta_0,\mu_0)| \le C\rho_J(1+s_\mu) \quad a.s.$$

bounded for every $\gamma\in\Gamma$ and some constant $C>0$. By the Hoeffding's inequality, we can prove that for each l,

$$\mathbb{P}\left(\left|\frac{1}{N}\sum_{i=1}^N m_{il}(\gamma;\eta_0,\mu_0) - \mathbb{E}[m_{il}(\gamma;\eta_0,\mu_0)]\right| \ge t\right) \le 2\exp\left(-\frac{2Nt^2}{C^2\rho_J^2(1+s_\mu)^2}\right)$$

so $\|M_{NJK}(\gamma;\eta_0,\mu_0) - \mathbb{E}[M_{NJK}(\gamma;\eta_0,\mu_0)]\|_2 = O_P(\rho_J(1+s_\mu)/\sqrt{N})$ for every $\gamma\in\Gamma$. Although it has been proved to be true, we cannot simply take supremum on both sides and then claim the supremum is also at the rate $\rho_J(1+s_\mu)/\sqrt{N}$. The supremum is not necessarily mean zero if $N^{-1}\sum_{i=1}^N m_{il}(\gamma;\eta_0,\mu_0) - \mathbb{E}[m_{il}(\gamma;\eta_0,\mu_0)]$ is mean zero.

Step 4 Finally, we discuss the rate $r_3(\gamma) := \|\mathbb{E}[M_{NJK}(\gamma;\eta_0,\mu_0)]\|_2 = \|\mathbb{E}[m_i(\gamma;\eta_0,\mu_0)]\|_2$. We have shown in the remark above that for each l,

$$\sup_{\gamma\in\Gamma}|m_{il}(\gamma;\eta_0,\mu_0)| \le C\rho_J(1+s_\mu) \quad a.s.$$

Then,

$$\begin{aligned}
\sup_{\gamma\in\Gamma} r_3^2(\gamma) &= \sup_{\gamma\in\Gamma}\sum_{l=1}^{d_\gamma}(\mathbb{E}[m_{il}(\gamma;\eta_0,\mu_0)])^2 \\
&\le \sum_{l=1}^{d_\gamma}\left(\sup_{\gamma\in\Gamma}\mathbb{E}[m_{il}(\gamma;\eta_0,\mu_0)]\right)^2 \\
&\le \sum_{l=1}^{d_\gamma}\left(\mathbb{E}\sup_{\gamma\in\Gamma} m_{il}(\gamma;\eta_0,\mu_0)\right)^2 \\
&\le O(\rho_J^2(1+s_\mu)^2)
\end{aligned}$$

where the second inequality is because the expectation of supremum is greater than the supremum of expectation. Therefore, $\sup_{\gamma\in\Gamma} r_3(\gamma) = O(\rho_J(1+s_\mu))$.

E.6 Proof of Theorem 4

Our proof is a corollary of Andrews (1999, Theorem 3). Assumption 6 suffices for his Assumption 2^{2*} given that (i) the log-likelihood function is well-defined and continuously differentiable over the whole real vector space, and (ii) Θ is either $\mathbb{R}^{d_\gamma}$ or the product of $\mathbb{R}$'s and $[0, \infty)$. It is worth noting that his objective function $l_N(\theta)$ (we use N instead of T) is aggregated and he requires $N^{-1} l_N(\theta) \to_p l(\theta)$ for some $l(\theta)$ uniformly, but we have $M_{NJK}(\theta)$ divided by N. Hence, to plug in his result, $l_N(\theta) = N\|M_{NJK}(\theta)\|_2^2$ so $N^{-1} l_N(\theta) = \|M_{NJK}(\theta)\|_2^2$, which converges to $\|M(\theta)\|_2^2$ by the continuous mapping theory and

$$\frac{1}{N/K} \frac{\partial}{\partial \gamma} L_{NJ}^{(k)}(\gamma; \hat{\eta}_k^{RMLE}) \to_p \mathbb{E}\left[\frac{\partial}{\partial \gamma} L(\gamma; \eta_0)\right] \quad \text{for any} \quad k = 1, \ldots, K$$

Assumption 7 and 8 are enough for his Assumption 3 that we proved

$$N^{1/2} D_{\gamma'} \|M_{NJK}(\gamma_0; \hat{\eta}, \hat{\mu})\|_2^2 \to_d 2\Omega_M N(0, \Sigma_M)$$
$$D_{\gamma'}^2 \|M_{NJK}(\gamma_0; \hat{\eta}, \hat{\mu})\|_2^2 \to_P 2\Omega_M \Omega_M'$$

Assumption 9 and Theorem 3 satisfy his Assumption 1. His Assumption 4 holds according to his Theorem 1. His Assumption 5 and 6 are automatically satisfied given the structure of Θ and the common convergence rate $\sqrt{N}$. Then, by his Theorem 3,

$$\sqrt{N}(\hat{\gamma}^{CDML} - \gamma_0) \to_d \tilde{\gamma} = \underset{\xi \in \Gamma(\gamma_0)}{\arg\min} \left(\xi + N(0, V_M)\right)' \Omega_M \Omega_M' \left(\xi + N(0, V_M)\right)$$

where $V_M = (\Omega_M \Omega_M')^{-1} \Omega_M \Sigma_M \Omega_M' (\Omega_M \Omega_M')^{-1} = \Omega_M^{-1} \Sigma_M \Omega_M^{-1}$. One important point to emphasize is the quadratic approximation. In his Theorem 3, the third result says that $l_N(\hat{\theta}) - l_N(\theta_0) \to_d \frac{1}{2} \hat{\lambda}' \mathcal{I} \hat{\lambda}$. The quadratic approximation in our case is

$$\|M_{NJK}(\hat{\gamma}^{CDML}; \hat{\eta}, \hat{\mu})\|_2^2 - \|M_{NJK}(\gamma_0; \hat{\eta}, \hat{\mu})\|_2^2 \approx -\frac{1}{2N} \tilde{\gamma}' D_\gamma^2 \|M_{NJK}(\hat{\gamma}^{CDML}; \hat{\eta}, \hat{\mu})\|_2^2 \tilde{\gamma}$$

and hence

$$N\left(\|M_{NJK}(\hat{\gamma}^{CDML}; \hat{\eta}, \hat{\mu})\|_2^2 - \|M_{NJK}(\gamma_0; \hat{\eta}, \hat{\mu})\|_2^2\right) \to_d -\tilde{\gamma}' \Omega_M \Omega_M' \tilde{\gamma}$$

The details of derivations are provided in Appendix F.4 and F.5.

F Auxiliary Results

F.1 Derivatives of the Log-likelihood

To be clear, for the k-by-1 vector $x = (x_1, \ldots, x_d)'$ and the function $y = (y_1(x), \ldots, y_p(x))'$, we use the following notations

$$D_x y_1(x) = \frac{d}{dx} y_1(x) = \begin{pmatrix} \frac{\partial}{\partial x_1} y_1(x) & \cdots & \frac{\partial}{\partial x_d} y_1(x) \end{pmatrix}$$

$$D_x y(x) = \frac{d}{dx} y(x) = \begin{pmatrix} \frac{\partial}{\partial x_1} y_1(x) & \cdots & \frac{\partial}{\partial x_d} y_1(x) \\ \vdots & \ddots & \vdots \\ \frac{\partial}{\partial x_1} y_p(x) & \cdots & \frac{\partial}{\partial x_d} y_p(x) \end{pmatrix}$$

$$D_{x'} y_1(x) = (D_x y_1(x))', \quad D_{x'} y(x) = (D_x y(x))'$$

$$D_x^2 y_1(x) = \frac{d^2}{dx dx'} y_1(x) = \begin{pmatrix} \frac{\partial^2}{\partial x_1^2} y_1(x) & \cdots & \frac{\partial^2}{\partial x_d \partial x_1} y_1(x) \\ \vdots & \ddots & \vdots \\ \frac{\partial^2}{\partial x_1 \partial x_d} y_1(x) & \cdots & \frac{\partial^2}{\partial x_d^2} y_1(x) \end{pmatrix}$$

Note that $D_x y_1(x)$ is 1-by-d and $D_{x'} y_1(x)$ is d-by-1. For clarity, we suppress the subscripts such as N, J and T. We use different notations for dimensions but they are unambiguous.

we plug in the logit probability for the function $L(\gamma; \eta)$. Recall that $L(\theta) := \sum_t \sum_i \sum_j Y_{ijt} \log s_{ijt}(\theta)$,

then

$$
\begin{aligned}
D_\theta L(\theta) &= \sum_t \sum_i \sum_j Y_{ijt} s_{ijt}^{-1}(\theta) D_{\theta'} s_{ijt}(\theta) \\
D_\theta^2 L(\theta) &= \sum_t \sum_i \sum_j Y_{ijt} \left[-s_{ijt}^{-2}(\theta) D_{\theta'} s_{ijt}(\theta) D_\theta s_{ijt}(\theta) + s_{ijt}^{-1}(\theta) D_\theta^2 s_{ijt}(\theta) \right] \\
\frac{\partial^3}{\partial\theta_k \partial\theta_l \partial\theta_m} L(\theta) &= \sum_t \sum_i \sum_j Y_{ijt} \Bigg\{ 2 s_{ijt}^{-3}(\theta) \frac{\partial}{\partial\theta_k} s_{ijt}(\theta) \frac{\partial}{\partial\theta_l} s_{ijt}(\theta) \frac{\partial}{\partial\theta_m} s_{ijt}(\theta) \\
&\quad - s_{ijt}^{-2}(\theta) D_\theta^2 s_{ijt}(\theta) + s_{ijt}^{-1}(\theta) D_\theta^3 s_{ijt}(\theta) \\
&\quad - s_{ijt}^{-2}(\theta) \Bigg[\frac{\partial^2}{\partial\theta_k \partial\theta_l} s_{ijt}(\theta) \frac{\partial}{\partial\theta_m} s_{ijt}(\theta) + \frac{\partial^2}{\partial\theta_k \partial\theta_m} s_{ijt}(\theta) \frac{\partial}{\partial\theta_l} s_{ijt}(\theta) \\
&\quad - \frac{\partial^2}{\partial\theta_l \partial\theta_m} s_{ijt}(\theta) \frac{\partial}{\partial\theta_k} s_{ijt}(\theta) \Bigg] \\
&\quad + s_{ijt}^{-1}(\theta) \frac{\partial^3}{\partial\theta_k \partial\theta_l \partial\theta_m} s_{ijt}(\theta) \Bigg\} \quad \forall\, k, l, m
\end{aligned}
$$

It suffices to derive the first, second and third partial derivatives for $s_{ijt}(\theta)$. With the interchangeability of the integral and the derivative,

$$
\begin{aligned}
s_{ijt}(\theta) &= \int \psi_j(\omega_{i\cdot t}) \phi(\nu_i) d\nu_i \\
\frac{\partial}{\partial\theta_k} s_{ijt}(\theta) &= \int \frac{\partial}{\partial\theta_k} \psi_j(\omega_{i\cdot t}) \phi(\nu_i) d\nu_i \\
\frac{\partial}{\partial\theta_k \partial\theta_l} s_{ijt}(\theta) &= \int \frac{\partial}{\partial\theta_k \partial\theta_l} \psi_j(\omega_{i\cdot t}) \phi(\nu_i) d\nu_i \\
\frac{\partial}{\partial\theta_k \partial\theta_l \partial\theta_m} s_{ijt}(\theta) &= \int \frac{\partial}{\partial\theta_k \partial\theta_l \partial\theta_m} \psi_j(\omega_{i\cdot t}) \phi(\nu_i) d\nu_i
\end{aligned}
$$

where $\omega_{i\cdot t} = (\omega_{i1t}, \ldots, \omega_{iJ_t t})'$ and the logit probability (also known as the soft-max function) for the product j is

$$
\psi_j(\omega_{i\cdot t}) = \frac{\exp(\omega_{ijt})}{1 + \sum_{j'} \exp(\omega_{ij't})}
$$

We temporarily drop the subscript i, t and the dot without loss of generality. By the chain rule,

$$
\frac{\partial}{\partial\theta_k} \psi_j(\omega) = \sum_q \frac{\partial}{\partial\omega_q} \psi_j(\omega) \frac{\partial}{\partial\theta_k} \omega_q
$$

Hopefully, ω_q is a linear function of θ_k, so $\frac{\partial}{\partial\theta_k}\omega_q$ is no longer a function of θ_k which helps

simplify the second and third-order of derivatives:

$$\frac{\partial^2}{\partial\theta_l\partial\theta_k}\psi_j(\omega) = \sum_r\sum_q \frac{\partial^2}{\partial\omega_r\partial\omega_q}\psi_j(\omega)\frac{\partial}{\partial\theta_l}\omega_r\frac{\partial}{\partial\theta_k}\omega_q$$
$$\frac{\partial^3}{\partial\theta_m\partial\theta_l\partial\theta_k}\psi_j(\omega) = \sum_t\sum_r\sum_q \frac{\partial^3}{\partial\omega_t\partial\omega_r\partial\omega_q}\psi_j(\omega)\frac{\partial}{\partial\theta_m}\omega_t\frac{\partial}{\partial\theta_l}\omega_r\frac{\partial}{\partial\theta_k}\omega_q$$

There are some nice properties of the soft-max function: it is continuously differentiable for all real vectors ω and bounded:

$$\frac{\partial}{\partial\omega_j}\psi_j(\omega) = \frac{\exp(\omega_j)}{1+\sum_{j'}\exp(\omega_{j'})} - \frac{\exp(\omega_j)\exp(\omega_j)}{(1+\sum_{j'}\exp(\omega_{j'}))^2} = \psi_j(\omega) - \psi_j^2(\omega) \in (0,1)$$
$$\frac{\partial}{\partial\omega_q}\psi_j(\omega) = -\frac{\exp(\omega_j)\exp(\omega_q)}{(1+\sum_{j'}\exp(\omega_{j'}))^2} = -\psi_j(\omega)\psi_q(\omega) \in (-1,0) \quad \text{for} \quad q \neq j$$

Equivalently, $\frac{\partial}{\partial\omega_q}\psi_j(\omega) = \psi_q(\omega)[I[q=j] - \psi_j(\omega)]$ and hence,

$$\begin{aligned}\frac{\partial^2}{\partial\omega_r\partial\omega_q}\psi_j(\omega) &= \left(I[q=j]-\psi_j(\omega)\right)\frac{\partial}{\partial\omega_r}\psi_q(\omega) - \psi_q(\omega)\frac{\partial}{\partial\omega_r}\psi_j(\omega)\\ &= \left(I[q=j]-\psi_j(\omega)\right)\psi_r(\omega)[I[r=q]-\psi_q(\omega)] - \psi_q(\omega)\psi_r(\omega)[I[r=j]-\psi_j(\omega)]\\ &= \psi_r(\omega)I[r=q=j] - \psi_r(\omega)\psi_q(\omega)\left(I[q=j]+I[r=j]\right) - \psi_r(\omega)\psi_j(\omega)I[r=q]\\ &\quad + 2\psi_r(\omega)\psi_q(\omega)\psi_j(\omega)\end{aligned}$$

which is still bounded (e.g., by $[-5,5]$) for any ω. Similarly,

$$\begin{aligned}\frac{\partial^3}{\partial\omega_r\partial\omega_q\partial\omega_t}\psi_j(\omega) &= \psi_t(\omega)[I[r=t]-\psi_r(\omega)]I[r=q=j]\\ &\quad + (I[q=j]+I[r=j])\psi_t(\omega)\left(I[r=t]-\psi_r(\omega)+I[q=t]-\psi_q(\omega)\right)\\ &\quad + I[r=q]\psi_t(\omega)\left(I[r=t]-\psi_r(\omega)+I[j=t]-\psi_j(\omega)\right)\\ &\quad + 2\psi_r(\omega)\left[I[r=t]+I[q=t]+I[j=t]-\psi_r(\omega)-\psi_q(\omega)-\psi_j(\omega)\right]\end{aligned}$$

which is also bounded.

Second, we prove the following important lemma which helps derive the rates of the derivatives.

Lemma 3. *Suppose that $\omega_{ij} = W_{ij}\theta_1 + Z_{ij}\nu_i\theta_2$ where the random variables $|W_{ij}| \leq C$ and $|Z_{ij}| \leq C$ for some constant $0 < C < \infty$ have bounded supports. Additionally, assume that*

(W_{ij}, Z_{ij}) is independent of ν_i. Then, for any $k, l, m = 1, 2$,

$$\frac{\partial}{\partial\theta_k}s_{ij}(\theta) \le 2C, \quad \frac{\partial^2}{\partial\theta_l\partial\theta_k}s_{ij}(\theta) \le 6C^2 \quad \textit{and} \quad \frac{\partial^3}{\partial\theta_m\partial\theta_l\partial\theta_k}s_{ij}(\theta) \le 21C^3$$

with probability one, where $\omega_i = (\omega_{i1}, \ldots, \omega_{iJ})$.

The proof is based on the fact that $\psi_j(\omega_i) \in [0, 1]$ and $\sum_{j=1}^{J}\psi_j(\omega_i) = 1 - \psi_0(\omega_i)$ for all ω_i and j. Since ω_{ij} is linear in $\theta_1$and θ_2, then $\partial\omega_{ij}/\partial\theta_1 = W_{ij}$ and $\partial\omega_{ij}/\partial\theta_2 = Z_{ij}\nu_i$. Then,

$$\begin{aligned}
\frac{\partial}{\partial\theta_k}\psi_j(\omega_i) &= \sum_{q=1}^{J}\psi_q(\omega_i)[I[q = j] - \psi_j(\omega_i)]\frac{\partial}{\partial\theta_k}\omega_{iq} \\
&= \left(\psi_j(\omega_i) - \psi_j^2(\omega_i)\right)\frac{\partial}{\partial\theta_k}\omega_{ij} - \sum_{q\neq j}\psi_q(\omega_i)\psi_j(\omega_i)\frac{\partial}{\partial\theta_k}\omega_{iq} \\
&= \left(\psi_j(\omega_i) - \psi_j^2(\omega_i)\right)\frac{\partial}{\partial\theta_k}\omega_{ij} - (1 - \psi_0(\omega_i) - \psi_j(\omega_i))\psi_j(\omega_i)\frac{\partial}{\partial\theta_k}\omega_{iq}
\end{aligned}$$

We take the expectation over ν_i on both sides:

$$\begin{aligned}
\frac{\partial}{\partial\theta_k}s_{ij}(\theta) = &\int\left(\psi_j(\omega_i) - \psi_j^2(\omega_i)\right)\frac{\partial}{\partial\theta_k}\omega_{ij}\phi(\nu_i)d\nu_i \\
&- \int(1 - \psi_0(\omega_i) - \psi_j(\omega_i))\psi_j(\omega_i)\frac{\partial}{\partial\theta_k}\omega_{iq}\phi(\nu_i)d\nu_i
\end{aligned}$$

For the first term,

- when $k = 1$, since $0 < \psi_q(\omega_i)\psi_j(\omega_i) < 1$ for any q and ω_i,

$$\int\left(\psi_j(\omega_i) - \psi_j^2(\omega_i)\right)W_{ij}\phi(\nu_i)d\nu_i = W_{ij}\int\left(\psi_j(\omega_i) - \psi_j^2(\omega_i)\right)\phi(\nu_i)d\nu_i \le C\int\phi(\nu_i)d\nu_i = C$$

- when $k = 2$,

$$\begin{aligned}
\int\left(\psi_j(\omega_i) - \psi_j^2(\omega_i)\right)Z_{ij}\nu_i\phi(\nu_i)d\nu_i &= Z_{ij}\int\left(\psi_j(\omega_i) - \psi_j^2(\omega_i)\right)\nu_i\phi(\nu_i)d\nu_i \\
&\le -C\int_{\{\nu_i<0\}}\nu_i\phi(\nu_i)d\nu_i + C\int_{\{\nu_i>0\}}\nu_i\phi(\nu_i)d\nu_i \\
&\le 2C\int_{\{\nu_i>0\}}\nu_i\phi(\nu_i)d\nu_i \\
&= \sqrt{2/\pi}C \le C
\end{aligned}$$

where we use the $(0,1)$ bound again in the second line, and the last line is by[22] the Gamma function $\Gamma(1) = 1$:

$$\int_0^\infty \nu_i (2\pi)^{-1/2} e^{-\nu_i^2/2} d\nu_i = (2\pi)^{-1/2} \int_0^\infty \sqrt{2t} e^{-t} \frac{\sqrt{2}}{2\sqrt{t}} dt = (2\pi)^{-1/2} \Gamma(1)$$

The second term can be shown similarly using the fact that $(1 - \psi_0(\omega_i) - \psi_j(\omega_i))\psi_j(\omega_i)$ is between $[0,1]$. Therefore,

$$\left| \frac{\partial}{\partial \theta_k} s_{ij}(\theta) \right| \leq 2C$$

Next,

$$\begin{aligned}
\frac{\partial^2}{\partial\theta_l \partial\theta_k} \psi_j(\omega_i) &= \sum_r \sum_q \frac{\partial^2}{\partial\omega_r \partial\omega_q} \psi_j(\omega_i) \frac{\partial}{\partial\theta_l} \omega_{ir} \frac{\partial}{\partial\theta_k} \omega_{iq} \\
&= \sum_r \sum_q \Big(\psi_r(\omega_i) I[r = q = j] - \psi_r(\omega_i)\psi_q(\omega_i) \left(I[q = j] + I[r = j] \right) \\
&\quad - \psi_r(\omega_i)\psi_j(\omega_i) I[r = q] \Big) \frac{\partial}{\partial\theta_l} \omega_{ir} \frac{\partial}{\partial\theta_k} \omega_{iq} \\
&\quad + \sum_r \sum_q 2\psi_r(\omega_i)\psi_q(\omega_i)\psi_j(\omega_i) \frac{\partial}{\partial\theta_l} \omega_{ir} \frac{\partial}{\partial\theta_k} \omega_{iq} \\
&= \psi_j(\omega_i) \frac{\partial}{\partial\theta_l} \omega_{ij} \frac{\partial}{\partial\theta_k} \omega_{ij} - \sum_r \psi_r(\omega_i)\psi_j(\omega_i) \frac{\partial}{\partial\theta_l} \omega_{ir} \frac{\partial}{\partial\theta_k} \omega_{ij} \\
&\quad - \sum_q \psi_j(\omega_i)\psi_q(\omega_i) \frac{\partial}{\partial\theta_l} \omega_{ij} \frac{\partial}{\partial\theta_k} \omega_{iq} - \sum_r \psi_r(\omega_i)\psi_j(\omega_i) \frac{\partial}{\partial\theta_l} \omega_{ir} \frac{\partial}{\partial\theta_k} \omega_{ir} \\
&\quad + \sum_r \sum_q 2\psi_r(\omega_i)\psi_q(\omega_i)\psi_j(\omega_i) \frac{\partial}{\partial\theta_l} \omega_{ir} \frac{\partial}{\partial\theta_k} \omega_{iq}
\end{aligned}$$

[22] Another way to think about this is through the mean of the half-normal distribution. If $Y = |X|$ and $|X| \sim N(0, \sigma^2)$, then Y follows a half-normal distribution with the density function

$$f(y) = \frac{\sqrt{2}}{\sigma\sqrt{\pi}} \exp\left(-\frac{y^2}{2\sigma^2} \right), \quad y \geq 0$$

and mean $\mathbb{E}[Y] = \frac{\sigma\sqrt{2}}{\sqrt{\pi}}$. In our case, $\sigma = 1$ and $\int_{\nu_i \geq 0} \nu_i \Phi(d\nu_i) = \frac{1}{2}\mathbb{E}[Y] = 1/\sqrt{2\pi}$.

Recall that there are four combinations:

$$\frac{\partial}{\partial\theta_l}\omega_{ir}\frac{\partial}{\partial\theta_k}\omega_{iq} = \begin{cases} W_{ir}W_{iq} & \text{if } l=1,k=1 \\ W_{ir}Z_{iq}\nu_i & \text{if } l=1,k=2 \\ W_{iq}Z_{ir}\nu_i & \text{if } l=2,k=1 \\ Z_{iq}Z_{ir}\nu_i^2 & \text{if } l=2,k=2 \end{cases}$$

The first three cases are essentially the same as previous case so the upper bounds are just C^2: for any function $g(\nu_i) \in [0,1]$,

$$\int g(\nu_i)\frac{\partial}{\partial\theta_l}\omega_{ij}\frac{\partial}{\partial\theta_k}\omega_{ij}\phi(\nu_i)d\nu_i$$
$$= \begin{cases} \int g(\nu_i)W_{ir}W_{iq}\phi(\nu_i)d\nu_i \le C^2\int\phi(\nu_i)d\nu_i = C^2 & \text{if } l=1,k=1 \\ \int g(\nu_i)W_{ir}Z_{iq}\nu_i\phi(\nu_i)d\nu_i \le C^2 2\int_0^\infty \nu_i\phi(\nu_i)d\nu_i = \sqrt{2/\pi}C^2 & \text{if } l=1,k=2 \end{cases}$$

The last one seems different but has the same upper bound because $\nu_i^2 \ge 0$, and hence,

$$\int_{-\infty}^{\infty} Z_{iq}Z_{ir}\nu_i^2\phi(\nu_i)d\nu_i \le C^2 Var(\nu_i) = C^2$$

Thus,

$$\left|\frac{\partial^2}{\partial\theta_l\partial\theta_k}s_{ij}(\theta)\right| \le 6C^2$$

Finally, for the third-order derivative, there are eight combinations

$$\frac{\partial}{\partial\theta_m}\omega_{it}\frac{\partial}{\partial\theta_l}\omega_{ir}\frac{\partial}{\partial\theta_k}\omega_{iq} = \begin{cases} W_{ir}W_{iq}W_{it} & \text{if } l=1,k=1,m=1 \\ W_{ir}W_{iq}Z_{it}\nu_i & \text{if } l=1,k=1,m=2 \\ W_{ir}Z_{iq}W_{it}\nu_i & \text{if } l=1,k=2,m=1 \\ W_{ir}Z_{iq}Z_{it}\nu_i^2 & \text{if } l=1,k=2,m=2 \\ Z_{ir}W_{iq}W_{it}\nu_i & \text{if } l=2,k=1,m=1 \\ Z_{ir}W_{iq}Z_{it}\nu_i^2 & \text{if } l=2,k=1,m=2 \\ Z_{ir}Z_{iq}W_{it}\nu_i^2 & \text{if } l=2,k=2,m=1 \\ Z_{ir}Z_{iq}Z_{it}\nu_i^3 & \text{if } l=2,k=2,m=2 \end{cases}$$

It suffices to verify that $\int_{-\infty}^{\infty} Z_{ir}Z_{iq}Z_{it}\nu_i^3\phi(\nu_i)d\nu_i$ is bounded. Letting $x = \nu_i^2/2$ (so $\nu_i = \sqrt{2x}$

and $dv_i/dx = 1/\sqrt{2x}$)

$$(2\pi)^{-1}\int_0^\infty v_i^3 e^{-v_i^2/2}dv_i = \pi^{-1}\int_0^\infty xe^{-x}dx = \pi^{-1}\Gamma(2) = \pi^{-1}$$

Hence,

$$\int_{-\infty}^\infty Z_{ir}Z_{iq}Z_{it}v_i^3\phi(v_i)dv_i \leq 2C^3\int_0^\infty v_i^3\phi(v_i)dv_i \leq C^3$$

Some algebra shows that

$$\left|\frac{\partial}{\partial\theta_k\partial\theta_l\partial\theta_m}s_{ijt}(\theta)\right| \leq 21C^3$$

as there are 21 terms that are between 0 and 1 in the summation.

The generalization to vectors θ_1 and θ_2 are straightforward as long as $Var(v_i)$ is diagonal.

F.2 Implicit Differentiation in BLP Models

In this section, we derive the derivatives while using contraction mapping to obtain $\delta = \delta(\theta) := \delta(\beta^o, \beta^u)$. By the chain-rule,

$$\frac{d}{d\theta}L_{NJT}(\theta) = \sum_t\sum_i\sum_j Y_{ijt}s_{ijt}^{-1}(\theta)\frac{d}{d\theta}s_{ijt}(\delta(\theta),\theta)$$

Since $s_{ijt} = \int\psi_j(\omega_{i\cdot t})\phi(v_i)dv_i$, interchange the integral and the partial derivative and then $\frac{ds_{ijt}}{d\theta} = \int\frac{d}{d\theta}\psi_j(\omega_{i.t})\phi(v_i)dv_i$. According to the previous section, $\frac{\partial}{\partial\beta^o}\psi_j(\omega_{i.t}) = \sum_q\frac{\partial}{\partial\omega_{iqt}}\psi_j(\omega_{i.t})\frac{\partial}{\partial\beta^o}\omega_{iqt}$ but now $\frac{\partial\omega_q}{\partial\beta^o} = \frac{\partial\delta_{qt}}{\partial\beta^o} + \frac{\partial\mu_{iqt}}{\partial\beta^o}$. Similarly, $\frac{\partial\omega_q}{\partial\beta^u} = \frac{\partial\delta_{qt}}{\partial\beta^u} + \frac{\partial\mu_{iqt}}{\partial\beta^u}$. By the implicit differentiation,

$$\frac{\partial\delta_{qt}(\beta^o,\beta^u)}{\partial\beta^o} = -\left(\frac{\partial s_{qt}(\delta_{\cdot t},\beta^o,\beta^u)}{\partial\delta_{qt}}\right)^{-1}\frac{\partial s_{qt}(\delta_{\cdot t},\beta^o,\beta^u)}{\partial\beta^o}$$

$$\frac{\partial\delta_{qt}(\beta^o,\beta^u)}{\partial\beta^u} = -\left(\frac{\partial s_{qt}(\delta_{\cdot t},\beta^o,\beta^u)}{\partial\delta_{qt}}\right)^{-1}\frac{\partial s_{qt}(\delta_{\cdot t},\beta^o,\beta^u)}{\partial\beta^u}$$

where

$$\begin{aligned}\frac{\partial s_{qt}(\delta_{\cdot t}, \beta^o, \beta^u)}{\partial \delta_{qt}} &= \int \frac{\exp(\delta_{qt} + \mu_{iqt})}{1 + \sum_{j'} \exp(\delta_{j't} + \mu_{ijt})} \left(1 - \frac{\exp(\delta_{qt} + \mu_{iqt})}{1 + \sum_{j'} \exp(\delta_{j't} + \mu_{ijt})}\right) d\mathcal{P}_{\mu_{i\cdot t}} \\ &= \int \psi_q(\omega_{i\cdot t}) \left(1 - \psi_q(\omega_{i\cdot t})\right) d\mathcal{P}_{\mu_{i\cdot t}} \\ &= \frac{1}{n} \sum_i \int \psi_q(\omega_{i\cdot t}(\nu_i)) \left(1 - \psi_q(\omega_{i\cdot t}(\nu_i))\right) \phi(\nu_i) d\nu_i\end{aligned}$$

is uniformly bounded[23] between 0 and 1, and

$$\frac{\partial s_{qt}(\delta_{\cdot t}, \beta^o, \beta^u)}{\partial \beta^o} = \frac{1}{n} \sum_i \int \sum_q \psi_q(\omega_{i\cdot t}) \left(1 - \psi_q(\omega_{i\cdot t})\right) \frac{\partial \omega_{iqt}}{\partial \beta^o} \phi(\nu_i) d\nu_i$$

Here $\mathcal{P}_{\mu_{i\cdot t}}$ is the probability measure of $\mu_{i\cdot t}$. Combine the results and we obtain

$$\frac{d}{d\theta} s_{ijt}(\delta(\theta), \theta) = \int \sum_{q=1}^{J} \left(1\{q = j\} - \psi_q(\omega_{i\cdot t})\right) \psi_j(\omega_{i\cdot t}) \left[\frac{d\delta_{qt}(\beta^o, \beta^u)}{d\theta} + \frac{d\omega_{iqt}}{d\theta}\right] \phi(\nu_i) d\nu_i$$

To calculate $\frac{d}{d\theta} s_{ijt}(\delta(\theta), \theta)$, it suffices to first calculate the partial derivatives with respect to β^o and β^u as if δ is given as data, then calculate $\frac{d}{d\theta} s_{ijt}$ plugging in the partial derivatives.

F.3 Derivatives of the SMM objective function

By the definition, the objective function is $\|M_n(\theta)\|_2^2 = \|n^{-1} \sum_{i=1}^n \sum_{j=0}^J (Y_{ij} - \hat{s}_{ij}(\theta)) Z_{ij}\|_2^2$. The first-order derivative is

$$D_\theta \|M_n(\theta)\|_2^2 = 2 \left[D_{\theta'} M_n(\theta)\right] M_n(\theta)$$

where the d_θ-by-d_θ matrix

$$D_{\theta'} M_n(\theta) = -\frac{1}{n} \sum_{i=1}^n \sum_{j=0}^J Z_{ij} D_\theta \hat{s}_{ij}(\theta) = -\frac{1}{n} \sum_{i=1}^n \sum_{j=0}^J Z_{ij} \frac{1}{B} \sum_{b=1}^B s_{ijb}(\theta) \sum_{k=1}^J \left(I\{k = j\} - s_{ikb}(\theta)\right) \frac{d\omega_{ikb}(\theta)}{d\theta}$$

[23] Its inverse, however, has an increasing upper bound that grows at the rate of $O(J)$ almost surely (cf. the previously derivatives with respect to θ).

The second-order derivative is

$$D_\theta^2 \|M_n(\theta)\|_2^2 = 2\left[D_{\gamma'} M(\gamma)\right] D_\gamma M(\gamma) + \sum_{j=1}^{d} 2M_j(\gamma) D_\gamma^2 M_j(\gamma)$$

F.4 Quadratic Approximation (QA)

We revisit the quadratic approximation in Andrews (1999) for references. To show that

$$\begin{aligned}
&f(\gamma_0) + D_\gamma f(\gamma_0)(\gamma - \gamma_0) + \frac{1}{2}(\gamma - \gamma_0)' D_\gamma^2 f(\gamma_0)(\gamma - \gamma_0) + R(\gamma, \gamma_0) \\
&= f(\gamma_0) - \frac{1}{2} D_\gamma f(\gamma_0) \left[D_\gamma^2 f(\gamma_0)'\right]^{-1} D_{\gamma'} f(\gamma_0) + \frac{1}{2N} q(\sqrt{N}(\gamma - \gamma_0)) + R(\gamma, \gamma_0)
\end{aligned}$$

with

$$q(x) := \left(x + \left[D_\gamma^2 f(\gamma_0)\right]^{-1} N^{1/2} D_{\gamma'} f(\gamma_0)\right)' D_\gamma^2 f(\gamma_0) \left(x + \left[D_\gamma^2 f(\gamma_0)\right]^{-1} N^{1/2} D_{\gamma'} f(\gamma_0)\right)$$

Firstly, expand the quadratic term $q(x)$ and use the fact that $D_\gamma^2 f(\gamma_0)$ is symmetric,

$$\begin{aligned}
q(x) =& x' D_\gamma^2 f(\gamma_0) x \\
&+ N^{1/2} D_\gamma f(\gamma_0) \left[D_\gamma^2 f(\gamma_0)\right]^{-1} D_\gamma^2 f(\gamma_0) \left[D_\gamma^2 f(\gamma_0)\right]^{-1} N^{1/2} D_{\gamma'} f(\gamma_0) \\
&+ x' D_\gamma^2 f(\gamma_0) \left[D_\gamma^2 f(\gamma_0)\right]^{-1} N^{1/2} D_{\gamma'} f(\gamma_0) \\
&+ N^{1/2} D_\gamma f(\gamma_0) \left[D_\gamma^2 f(\gamma_0)\right]^{-1} D_\gamma^2 f(\gamma_0) x \\
=& x' D_\gamma^2 f(\gamma_0) x + N D_\gamma f(\gamma_0) \left[D_\gamma^2 f(\gamma_0)\right]^{-1} D_{\gamma'} f(\gamma_0) \\
&+ 2N^{1/2} D_\gamma f(\gamma_0) x
\end{aligned}$$

Plugging in $x = \sqrt{N}(\gamma - \gamma_0)$, then

$$\frac{1}{2N} q(\sqrt{N}(\gamma - \gamma_0)) = \frac{1}{2}(\gamma - \gamma_0)' D_\gamma^2 f(\gamma_0)(\gamma - \gamma_0) + \frac{1}{2} D_\gamma f(\gamma_0) \left[D_\gamma^2 f(\gamma_0)\right]^{-1} D_{\gamma'} f(\gamma_0) + D_\gamma f(\gamma_0)(\gamma - \gamma_0)$$

Rearrange the equation and the proof is done.

Another interesting result is about the difference:

$$f(\gamma) - f(\gamma_0) = -\frac{1}{2} D_\gamma f(\gamma_0) \left[D_\gamma^2 f(\gamma_0)'\right]^{-1} D_{\gamma'} f(\gamma_0) + \frac{1}{2N} q(\sqrt{N}(\gamma - \gamma_0)) + R(\gamma, \gamma_0)$$

For simplicity, assume that $R(\gamma, \gamma_0) = 0$ for all γ. To solve $\inf_{\gamma \in \Theta} f(\gamma) - f(\gamma_0)$, or equivalently, $\inf_{\gamma \in \Theta} f(\gamma)$, it is identical to solve $\inf_{\gamma \in \Theta} (2N)^{-1} q(\sqrt{N}(\gamma - \gamma_0))$. Since our Θ is the

product of $\mathbb{R}$'s and $[0,\infty)$'s, then the shifted parameter space $\Theta-\theta_0$ is either the product of only $\mathbb{R}$'s (if σ_j's are all positive numbers) or the product of $\mathbb{R}$'s and $[0,\infty)$'s (if at least one $\sigma_j=0$). This suggests $\sqrt{N}(\Theta-\theta_0)=\Theta-\theta_0$, so $\inf_{\gamma\in\Theta}(2N)^{-1}q(\sqrt{N}(\gamma-\gamma_0))=\inf_{\gamma\in\Theta-\theta_0}(2N)^{-1}q(\gamma)$. Note that $q(\cdot)$ is a quadratic function. If $D^2_\gamma f(\gamma_0)$ is positive (semi-)definite, then $\sqrt{x'D^2_\gamma f(\gamma_0)x}$ induced a (semi-)norm $\|\cdot\|$ and

$$\inf_{\gamma\in\Theta-\theta_0} q^{1/2}(\lambda)=\inf_{\gamma\in\Theta-\theta_0}\|\gamma+\left[D^2_\gamma f(\gamma_0)\right]^{-1}N^{1/2}D_{\gamma'}f(\gamma_0)\|$$

The solution, denoted by $\tilde{\gamma}$, is the projection of $\left[D^2_\gamma f(\gamma_0)\right]^{-1}N^{1/2}D_{\gamma'}f(\gamma_0)$ onto $\Theta-\theta_0$. Andrews (1999) claims the orthogonal property that $\tilde{\gamma}'D^2_\gamma f(\gamma_0)(\tilde{\gamma}+\left[D^2_\gamma f(\gamma_0)\right]^{-1}N^{1/2}D_{\gamma'}f(\gamma_0))=0$. To see this, if $\Theta-\theta_0=[0,\infty)$ and $D^2_\gamma f(\gamma_0)=1$ so that $\|\cdot\|$ is the Euclidean norm, then $\tilde{\gamma}=0$ if $N^{1/2}D_{\gamma'}f(\gamma_0)\geq 0$ and $\tilde{\gamma}=-N^{1/2}D_{\gamma'}f(\gamma_0)$ if $N^{1/2}D_{\gamma'}f(\gamma_0)<0$. Hence, $\tilde{\gamma}'D^2_\gamma f(\gamma_0)\tilde{\gamma}=-\tilde{\gamma}'N^{1/2}D_{\gamma'}f(\gamma_0)$. Therefore, if $\hat{\gamma}$ minimizes $f(\gamma)$ over Θ, then plug in the formula of $q(x)$ and

$$\begin{aligned}
f(\hat{\gamma})-f(\gamma_0)&=-\frac{1}{2}D_\gamma f(\gamma_0)\left[D^2_\gamma f(\gamma_0)'\right]^{-1}D_{\gamma'}f(\gamma_0)+\frac{1}{2N}q(\tilde{\gamma})\\
&=\frac{1}{2N}\left(\tilde{\gamma}'D^2_\gamma f(\gamma_0)\tilde{\gamma}+2N^{1/2}D_\gamma f(\gamma_0)\tilde{\gamma}\right)\\
&=\frac{1}{2N}\left(\tilde{\gamma}'D^2_\gamma f(\gamma_0)\tilde{\gamma}-2\tilde{\gamma}'D^2_\gamma f(\gamma_0)\tilde{\gamma}\right)\\
&=-\frac{1}{2N}\tilde{\gamma}'D^2_\gamma f(\gamma_0)\tilde{\gamma}
\end{aligned}$$

which is the 3(c) in Theorem 3 in Andrews (1999).

F.5 QA of the CDML Loss

Next, we plug in the CDML objective function $f(\gamma)=\|M(\gamma)\|_2^2$ which is the squared l_2-norm of a d-by-1 vector. By the definition, $\|M(\gamma)\|_2^2=\sum_{j=1}^d M_j^2(\gamma)$, then some algebra shows

$$\begin{aligned}
D_\gamma\|M(\gamma)\|_2^2&=\sum_{j=1}^d D_\gamma M_j^2(\gamma)=\sum_{j=1}^d 2M_j(\gamma)D_{\gamma'}M_j(\gamma)\\
&=2\begin{pmatrix}D_{\gamma_1}M_1(\gamma)&\cdots&D_{\gamma_1}M_d(\gamma)\\ \vdots&\ddots&\vdots\\ D_{\gamma_d}M_1(\gamma)&\cdots&D_{\gamma_d}M_d(\gamma)\end{pmatrix}\begin{pmatrix}M_1(\gamma)\\ \vdots\\ M_d(\gamma)\end{pmatrix}\\
&=2\left[D_{\gamma'}M(\gamma)\right]M(\gamma)
\end{aligned}$$

$$
\begin{aligned}
D_\gamma^2 \|M(\gamma)\|_2^2 &= \sum_{j=1}^{d} 2D_\gamma \begin{pmatrix} M_j(\gamma) D_{\gamma_1} M_j(\gamma) \\ \vdots \\ M_j(\gamma) D_{\gamma_d} M_j(\gamma) \end{pmatrix} \\
&= \sum_{j=1}^{d} 2 \begin{pmatrix} D_{\gamma_1}\left[M_j(\gamma) D_{\gamma_1} M_j(\gamma)\right] & \cdots & D_{\gamma_d}\left[M_j(\gamma) D_{\gamma_1} M_j(\gamma)\right] \\ \vdots & \ddots & \vdots \\ D_{\gamma_1}\left[M_j(\gamma) D_{\gamma_d} M_j(\gamma)\right] & \vdots & D_{\gamma_d}\left[M_j(\gamma) D_{\gamma_d} M_j(\gamma)\right] \end{pmatrix} \\
&= \sum_{j=1}^{d} 2 \begin{pmatrix} D_{11} & \cdots & D_{d1} \\ \vdots & \ddots & \vdots \\ D_{1d} & \cdots & D_{dd} \end{pmatrix} \\
&\quad \text{where} \quad D_{kl} := \left[D_{\gamma_k} M_j(\gamma)\right] D_{\gamma_l} M_j(\gamma) + M_j(\gamma) \left[D_{\gamma_k} D_{\gamma_l} M_j(\gamma)\right] \\
&= \sum_{j=1}^{d} 2 \left[D_{\gamma'} M_j(\gamma)\right] D_\gamma M_j(\gamma) + 2 M_j(\gamma) D_\gamma^2 M_j(\gamma) \\
&= 2 \left[D_{\gamma'} M(\gamma)\right] D_\gamma M(\gamma) + \sum_{j=1}^{d} 2 M_j(\gamma) D_\gamma^2 M_j(\gamma)
\end{aligned}
$$

and if the remainder term is 0, by the previous subsection,

$$
\|M(\hat{\gamma})\|_2^2 - \|M(\gamma)\|_2^2 = -\frac{1}{2N} \tilde{\gamma}' D_\gamma^2 \|M(\gamma)\|_2^2 \tilde{\gamma}
$$

where $\hat{\gamma} = \arg\min_{\gamma \in \Theta} \|M(\hat{\gamma})\|_2^2$ and

$$
\begin{aligned}
\tilde{\gamma} = \underset{x \in \Theta - \theta_0}{\arg\min} \Big(& x + \left[D_\gamma^2 \|M(\gamma_0)\|_2^2\right]^{-1} N^{1/2} D_\gamma \|M(\gamma_0)\|_2^2 \Big)' \\
& D_\gamma^2 \|M(\gamma_0)\|_2^2 \Big(x + \left[D_\gamma^2 \|M(\gamma_0)\|_2^2\right]^{-1} N^{1/2} D_\gamma \|M(\gamma_0)\|_2^2 \Big)
\end{aligned}
$$

We further plug in the score function. Given $M(\gamma) = D_{\gamma'} L(\gamma; \eta) - \mu D_{\eta'} L(\gamma; \eta)$, obviously $M_j(\gamma) = D_{\gamma_j} L(\gamma; \eta) - \mu_j D_\eta L(\gamma; \eta)$ and $D_{\gamma_k} M_j(\gamma) = D_{\gamma_k} D_{\gamma_j} L(\gamma; \eta) - \mu_j D_{\gamma_k} D_\eta L(\gamma; \eta)$, where μ_j is the j-th row of μ . Then,

$$
\begin{aligned}
D_{\gamma'} M(\gamma) &= \begin{pmatrix} D_{\gamma_1} D_{\gamma_1} L(\gamma; \eta) - \mu_1 D_{\gamma_1} D_\eta L(\gamma; \eta) & \cdots & D_{\gamma_1} D_{\gamma_d} L(\gamma; \eta) - \mu_d D_{\gamma_1} D_\eta L(\gamma; \eta) \\ \vdots & \ddots & \vdots \\ D_{\gamma_d} D_{\gamma_1} L(\gamma; \eta) - \mu_1 D_{\gamma_d} D_\eta L(\gamma; \eta) & \cdots & D_{\gamma_1} D_{\gamma_1} L(\gamma; \eta) - \mu_d D_{\gamma_d} D_\eta L(\gamma; \eta) \end{pmatrix} \\
&= D_\gamma^2 L(\gamma; \eta) - D_{\gamma\eta} L(\gamma; \eta) \mu' \\
D_\gamma^2 M_j(\gamma) &= D_\gamma^2 D_{\gamma_j} L(\gamma; \eta) - D_\gamma^2 \mu_j D_\eta L(\gamma; \eta)
\end{aligned}
$$

Note that $D_{\gamma'}M(\gamma)$ is not symmetric even at the true value $(\gamma, \eta, \mu) = (\gamma_0, \eta_0, \mu_0)$. For the second equation, the first term, as the third-order derivative, is simpler:

$$D_\gamma^2 D_{\gamma_j} L(\gamma;\eta) = \begin{pmatrix} D_{11j} & \cdots & D_{d1j} \\ \vdots & \ddots & \vdots \\ D_{1dj} & \cdots & D_{ddj} \end{pmatrix} \quad \text{where} \quad D_{klj} := \frac{\partial^3}{\partial\gamma_k \partial\gamma_l \partial\gamma_j} L(\gamma;\eta)$$

The second term is more tricky as μ_j is a 1-by-p vector and $\mu_j D_\eta L(\gamma;\eta) = \sum_{q=1}^p \mu_{jq} D_{\eta_q} L(\gamma;\eta)$, hence,

$$\begin{aligned} D_\gamma^2 \mu_j D_\eta L(\gamma;\eta) &= \sum_{q=1}^p \mu_{jq} D_\gamma^2 D_{\eta_q} L(\gamma;\eta) \\ &= \sum_{q=1}^p \mu_{jq} \begin{pmatrix} L_{11q} & \cdots & L_{d1q} \\ \vdots & \ddots & \vdots \\ L_{1dq} & \cdots & L_{ddq} \end{pmatrix} \quad \text{where} \quad L_{klq} := \frac{\partial^3}{\partial\gamma_k \partial\gamma_l \partial\eta_q} L(\gamma;\eta) \end{aligned}$$

F.6 QA of the Generalized CDML Loss

The generalized CDML loss function contains a GMM-type weighting matrix $f(\gamma) = M(\gamma)'WM(\gamma) = \sum_{j=1}^d \sum_{k=1}^d M_j(\gamma) M_k(\gamma) W_{jk}$, where W is assumed to be symmetric.

F.7 Non-negative Quadratic Programming

Consider the following quadratic programming problem with a non-negative constraint on the scalar x_2:

$$\min_{x_1 \in \mathbb{R}^d, x_2 \geq 0} \begin{pmatrix} x_1' + b_1', x_2' + b_2' \end{pmatrix} \begin{pmatrix} A & C \\ B & D \end{pmatrix} \begin{pmatrix} x_1 + b_1 \\ x_2 + b_2 \end{pmatrix}$$

Here $A \in \mathbb{R}^{d \times d}$ is assumed symmetric and positive definite, $D > 0$, B is a d-by-1 vector and C is a 1-by-d vector. Expand the quadratic function and we obtain the objective function

$$(x_1 + b_1)'A(x_1 + b_1) + (x_2 + b_2)C(x_1 + b_1) + (x_1 + b_1)'B(x_2 + b_2) + (x_2 + b_2)^2 D$$

The associated Lagrangian is

$$L(x, \lambda) = (x_1 + b_1)'A(x_1 + b_1) + (x_2 + b_2)(C + B')(x_1 + b_1) + (x_2 + b_2)^2 D - \lambda x_2, \quad \lambda \geq 0$$

Then, the first-order conditions are

$$\nabla_{x_1} L(x, \lambda) = 2A'(x_1 + b_1) + (C' + B)(x_2 + b_2) = 0$$
$$\nabla_{x_2} L(x, \lambda) = (C + B')(x_1 + b_1) + 2D(x_2 + b_2) - \lambda = 0$$

and the complement slackness condition is

$$\lambda \geq 0 \quad \text{and} \quad \lambda x_2 = 0$$

When $x_2^* = 0$,

$$x_1^* = -\frac{1}{2}A^{-1}(C' + B)b_2 - b_1 \quad \text{and} \quad \lambda^* = \left(2D - \frac{1}{2}(C + B')A^{-1}(C' + B)\right) b_2$$

where $\lambda^* > 0$ needs to be verified. When $\lambda^* \leq 0$,

$$x_1^* = -b_1 \quad \text{and} \quad x_2^* = -b_2$$

F.8 Inequalities

Lemma 4. *For any $A \in \mathbb{R}^{n \times m}$ and $b \in \mathbb{R}^m$, the following inequality holds*

$$\|Ab\|_2 \leq \|A\|_F \|b\|_2$$

where $\|A\|_F = \sqrt{\sum_{i=1}^n \sum_{j=1}^m A_{ij}^2}$ is the Frobenius norm and $\|b\|_2 = \sqrt{\sum_{i=1}^m b_i^2}$ is the Euclidean norm.

Proof of Lemma 4. It suffices to show that $\|Ab\|_2^2 \leq \|A\|_F^2 \|b\|_2^2$. Let A_{i*} be the i-th row of A, then

$$\|Ab\|_2^2 = \sum_{i=1}^n (A_{i*}b)^2 = \sum_{i=1}^n \left(\sum_{j=1}^m A_{ij} b_j\right)^2 \leq \sum_{i=1}^n \left(\sum_{j=1}^m A_{ij}^2\right)\left(\sum_{j=1}^m b_j^2\right) = \left(\sum_{j=1}^m b_j^2\right)\left(\sum_{i=1}^n \sum_{j=1}^m A_{ij}^2\right)$$
$$= \|b\|_2^2 \|A\|_F^2$$

where we apply Cauchy-Schwarz inequality $(\sum_{j=1}^m A_{ij} b_j)^2 \leq (\sum_{j=1}^m A_{ij}^2)(\sum_{j=1}^m b_j^2)$ to complete the proof. □

Lemma 5 (McDiarmid's Inequality)**.** *Suppose that $X_1, \ldots, X_n \in \mathcal{X}$ are independent random vectors and $Z = f(X_1, \ldots, X_n)$ is a random variable where f has the bounded difference property:*

there exists some non-negative constants $c_1, \ldots, c_n$ *such that*

$$\mathbb{E}|f(X_1, \ldots, X_n) - f(X_1, \ldots, X_{i-1}, X_i', X_{i+1}, \ldots, X_n)| \leq c_i, \quad 1 \leq i \leq n$$

Then,

$$\mathbb{P}(|Z - \mathbb{E}[Z]| > t) \leq 2\exp\left(-\frac{2t^2}{\sum_{i=1}^n c_i^2}\right) \quad \textit{for all } t > 0$$

Lemma 6 (Kennedy, 2023)**.** *Let* $f(W_i, \theta)$ *be a d-dimensional vector-valued function of data* W_i *and a parameter* $\theta \in \Theta$*. Let* $\hat{\theta} \in \Theta$ *be the estimator from a sample* $W^N = (W_{n+1}, \ldots, W_N)$*,* $\mathbb{P}$ *denote the measure conditional on* W^N*, and* $\mathbb{P}_n$ *denote the empirical measure over i.i.d. samples* $W^n = (W_1, \ldots, W_n)$*, which is independent of* W^N*. Then,*

1. If $d = 1$, then

$$\mathbb{G}_n(f(\hat{\theta}) - f(\theta)) := \sqrt{n}(\mathbb{P}_n - \mathbb{P})\left(f(\hat{\theta}) - f(\theta)\right) = O_{\mathbb{P}}\left(\frac{\mathbb{P}[(f(\hat{\theta}) - f(\theta))^2]}{\sqrt{n}}\right)$$

2. For any $d \in \mathbb{N}$,

$$\mathbb{E}_{W^n}\left[\|\mathbb{G}_n(f(\hat{\theta}))\| \mid W^N\right] \leq \mathbb{E}_{W^n}\left[\|f(\hat{\theta}) - f(\theta)\|_2^2 \mid W^N\right]$$

Proof of Lemma 6. The key idea in the proof is that $\hat{\theta}$ can be taken as a constant conditional on W^N. The first result is directly from *Kennedy et al. (2020)* and we excerpt the proof here for readers' references. Notice that the empirical process term $\mathbb{G}_n(f(\hat{\theta})) = n^{1/2}\left[\mathbb{P}_n f(\hat{\theta}) - \mathbb{P} f(\hat{\theta})\right]$ has mean zero conditional on W^N since

$$\mathbb{E}_{W^n}\left[\mathbb{P}_n(f(\hat{\theta}) - f(\theta)) \mid W^N\right] = \mathbb{E}_{W^n}\left[f(\hat{\theta}) - f(\theta) \mid W^N\right] = \mathbb{P}(f(\hat{\theta}) - f(\theta))$$

By the i.i.d. assumption, the conditional variance is

$$\begin{aligned}
Var\left(\mathbb{G}_n(f(\hat{\theta}) - f(\theta)) \mid W^N\right) &= Var\left(n^{1/2}\mathbb{P}_n(f(\hat{\theta}) - f(\theta)) \mid W^N\right) \\
&= n^{-1} Var\left(f(\hat{\theta}) - f(\theta) \mid W^N\right) \\
&= n^{-1}\left\{\mathbb{E}_{W_n}\left[(f(\hat{\theta}) - f(\theta))^2 \mid W^N\right] - \left(\mathbb{E}_{W_n}\left[f(\hat{\theta}) - f(\theta) \mid W^N\right]\right)^2\right\} \\
&\leq n^{-1}\mathbb{P}(f(\hat{\theta}) - f(\theta))^2
\end{aligned}$$

Then, by the law of iterated expectation and Chebyshev's inequality,

$$\begin{aligned} Pr\left(\frac{\mathbb{G}_n(f(\hat{\theta})-f(\theta))}{\sqrt{n^{-1}\mathbb{P}(f(\hat{\theta})-f(\theta))^2}} \geq t\right) &= \mathbb{E}\left[Pr\left(\frac{\mathbb{G}_n(f(\hat{\theta})-f(\theta))}{\sqrt{n^{-1}\mathbb{P}(f(\hat{\theta})-f(\theta))^2}} \geq t \middle| W^N\right)\right] \\ &\leq \mathbb{E}\left[\frac{1}{t^2}\frac{Var\left(\mathbb{G}_n(f(\hat{\theta})-f(\theta)) \mid W^N\right)}{n^{-1}\mathbb{P}(f(\hat{\theta})-f(\theta))^2}\right] \\ &\leq t^{-2} \end{aligned}$$

For any given $\epsilon > 0$, one can pick $t = \epsilon^{-1/2}$ and the proof is done.

The second result is mentioned in Chernozhukov et al. (2018, C56) but we have the inequality. The proof is quite similar. The conditional variance is now a d-by-d matrix:

$$\begin{aligned} \mathbb{E}_{W_n}\left[n^{-1}\mathbb{G}_n(f(\hat{\theta})-f(\theta))\mathbb{G}_n(f(\hat{\theta})-f(\theta))' \mid W^N\right] &= Var\left(n^{-1/2}\mathbb{G}_n(f(\hat{\theta})-f(\theta)) \mid W^N\right) \\ &= n^{-1}Var\left(f(\hat{\theta})-f(\theta) \mid W^N\right) \\ &= n^{-1}\mathbb{P}\left\{\left(f(\hat{\theta})-f(\theta)\right)\left(f(\hat{\theta})-f(\theta)\right)'\right\} \\ &\quad - n^{-1}\left\{\mathbb{P}\left(f(\hat{\theta})-f(\theta)\right)\right\}\left\{\mathbb{P}\left(f(\hat{\theta})-f(\theta)\right)\right\}' \end{aligned}$$

where the first line is because of the zero conditional mean. Drop the n^{-1} and take trace on both sides, and we obtain

$$\mathbb{E}_{W_n}\left[\|\mathbb{G}_n(f(\hat{\theta})-f(\theta))\|_2^2 \mid W^N\right] = \mathbb{E}_{W_n}\left[\|f(\hat{\theta})-f(\theta)\|_2^2 \mid W^N\right] - \|\mathbb{E}_{W_n}[f(\hat{\theta})-f(\theta) \mid W^N]\|_2^2$$

Then, the proof is done as the second term on the right-hand side is non-negative. □

Lemma 7 (Hoeffding's Inequality for Bounded Random Variables)**.** *Let $W_1, \ldots, W_N$ be independent random variables. Assume that $W_i \in [m_i, W_i]$ for every $i = 1, \ldots, N$. Then, for any $t > 0$,*

$$\mathbb{P}\left(\frac{1}{N}\left|\sum_{i=1}^{N} X_i - \mathbb{E}X_i\right| \geq t\right) \leq 2\exp\left(-\frac{2N^2t^2}{\sum_{i=1}^{N}(M_i - m_i)^2}\right)$$

Proof of Lemma 7. See Theorem 2.2.6 in Vershynin (2018). □