ARCHIVES
MASSACHUSETTS INSTITUTE OF TECHNOLOGY
JUN 09 2015
LIBRARIES

High-Dimensional Econometrics and Model Selection

by

Ye Luo

B.S., Massachusetts Institute of Technology (2010)

Submitted to the Department of Economics
in partial fulfillment of the requirements for the degree of

Doctor of Philosophy in Economics

at the

MASSACHUSETTS INSTITUTE OF TECHNOLOGY

June 2015

©2015 Ye Luo, All Rights Reserved

The author hereby grants to MIT permission to reproduce and to distribute publicly paper and electronic copies of this thesis document in whole or in part in any medium now known or hereafter created.

Author ..... Signature redacted .....
Department of Economics
May 15, 2015

Certified by .. Signature redacted .....
Victor Chernozhukov
Professor of Economics
Thesis Supervisor

Certified by ... Signature redacted .....
Jerry Hausman
Professor of Economics
Thesis Supervisor

Accepted by .... Signature redacted .....
Ricardo Caballero
Ford International Professor of Economics
Chairman, Department Committee on Graduate Theses

High-Dimensional Econometrics and Model Selection

by

Ye Luo

Submitted to the Department of Economics
on May 15, 2015, in partial fulfillment of the
requirements for the degree of
Doctor of Philosophy in Economics

Abstract

This dissertation consists of three chapters. Chapter 1 proposes a new method to solve the many moment problem: in Generalized Method of Moments (GMM), when the number of moment conditions is comparable to or larger than the sample size, the traditional methods lead to biased estimators. We propose a LASSO based selection procedure in order to choose the informative moments and then, using the selected moments, conduct optimal GMM. My method can significantly reduce the bias of the optimal GMM estimator while retaining most of the information in the full set of moments. We establish theoretical asymptotics of the LASSO and post-LASSO estimators. The formulation of LASSO is a convex optimization problem and thus the computational cost is low compared to all existing alternative moment selection procedures. We propose penalty terms using data-driven methods, of which the calculation is carried out by a non-trivial adaptive algorithm.

In Chapter 2, we consider partially identified models with many inequalities. Under such circumstances, existing inference procedures may break down asymptotically and are computationally difficult to conduct. We first propose a combinatorial method to select the informative inequalities in the Core Determining Class problem, in which a large set of linear inequalities are generated from a bipartite graph. Our method selects the set of irredudant inequalities and outperforms all existing methods in shrinking the number of inequalities and computational speed. We further consider a more general problem with many linear inequalities. We propose an inequality selection method similar to the Dantzig selector. We establish theoretical results of such a selection method under our sparsity assumptions.

Chapter 3 proposes an innovative way of reporting results in empirical analysis of economic data. Instead of reporting the Average Partial Effect, we propose to report multiple effects sorted in increasing order, as an alternative and more complete summary measure of the heterogeneity in the model. We established asymptotics and inference

for such a procedure via functional delta method. Numerical examples and an empirical application to female labor supply using data from the 1980 U.S. Census illustrate the performance of our methods in finite samples.

Thesis Supervisor: Victor Chernozhukov
Title: Professor of Economics

Thesis Supervisor: Jerry Hausman
Title: Professor of Economics

Acknowledgments

I am very grateful to my advisors Victor Chernozhukov and Jerry Hausman for their guidance, support and advice for this dissertation. Conversations with Victor Chernozhukov and Jerry Hausman helped to inspire my research ideas and form a general view on econometrics. Victor Chernozhukov taught me high-dimensional econometrics and especially the LASSO method, which plays an essential part in this dissertation. I am also in deep debt to my thesis committee members Anna Mikusheva and Whitney Newey for their criticisms to help me learn practical econometrics and the way of writing a paper in economics.

Chapter 2 is a joint work with my colleague Hai Wang, a Ph.d Candidate in MIT Operation Research Center. Chapter 3 is a joint work with Victor Chernozhukov and Ivan Fernandez-val. Other project collaborators include Denis Chetverikov, Wiston Wei Dou, Jerry Hausman, Christopher Palmer and Martin Spindler. I thank them for their patience and enjoyed working together very much. They are really wonderful co-workers and good friends. I thank Alexandar Belloni for providing an inspiring idea to apply penalties to the Farkas Lemma in Chapter 2. I thank Alfred Galichon for providing very useful comments and suggestions in improving Chapter 2.

I am in debt to Chunrong Ai, Isaiah Andrews, Joshua Angrist, Cynthia Barnhart, Andrew Chesher, Bruce Hansen, Anna Mikusheva, Whitney Newey, Jack Porter, Xiaoxia Shi, Norman Swanson and participants in MIT Economics seminar for many helpful discussions on the chapters of my dissertations.

After all, I am most grateful to my wife Lixin and my parents for supporting me in my research. I also thank my uncle Jianhui Luo and my brother Sihang Liu for helping me in my personal life during my ph.d.

Contents

List of Figures

List of Tables

Chapter 1

Selecting Informative Moments via LASSO

1.1 Introduction

The optimal two-step GMM estimator has been widely used in economic applications. It is quite common to have an application with a large number of moment restrictions that can be used for estimation and inference. For example, a conditional moment restriction provides an infinite number of potential unconditional moments by allowing the use of different functions of the conditioning variable as instruments. However, applying an efficient GMM estimator to many moment conditions typically results in biased estimators and poor accuracy of confidence sets. Hansen, Hausman, Newey (2006) shows that the presence of many valid instrumental variables (IV later) may improve efficiency, but the inference procedure becomes inaccurate due to second order bias. If the number of moments exceeds the sample size, then an efficient GMM estimator does not exist at all. The problem with many moments arises from the efficient GMM's need for the optimal weighting matrix, which is an inverse of a large dimensional random matrix. The ill-posedeness of the inversion problem leads to poor performance of the optimal GMM estimator. However, simply throwing out over-identified moments is undesirable due to efficiency losses.

The main goal of this paper is to improve the GMM procedure by selecting informative moment conditions from a large set of available moments. The selection procedure proposed in this paper utilizes the basic spirit of regression-LASSO, using the L_1 penalty to find a nearly optimal combination of moment conditions. The goal is to select moments without loss of asymptotic efficiency but that will guarantee the accurate coverage property of post selection inferences.

The main assumption needed to ensure the validity of the suggested procedure is approximate sparsity. The exact sparsity assumption means that all but a relatively small (though increasing with the sample size) number of moments is absolutely uninformative about the parameter we are trying to estimate. Approximate sparsity weakens this condition by allowing all moments to have some information about the parameter of interest but, in fact, the majority of moments has so little informational content that no loss of asymptotic efficiency occurs from not using those moments. The number and identity of truly informative moments are unknown, but we need to impose bounds on the growth rate of the number of informative moments requiring that it be much smaller than the sample size.

The LASSO method proposed in this paper could be viewed as a complementary method to the traditional methods for the many moments problem. We provide a description of the convergence rate properties of such a selection mechanism. As we prove, under the approximate sparsity assumption together with other technical conditions, the LASSO-based estimators are asymptotically efficient. Our estimators have much less second order bias for valid inference when compared to the optimal GMM estimator and different versions of bias-corrected estimators. Our method also has low computational cost and is easy to implement in practice as an optimization problem with a globally convex function and L_1 penalty.

One of main challenges we face in this paper is the selection of the appropriate penalty terms that would guarantee the efficient performance of the LASSO-selection procedure. We derive theoretical penalty terms that guarantee the asymptotic behavior of post-LASSO estimators and develop a feasible version of those penalties. We adopt a modest deviation theory of self-normalized vectors to construct data-driven penalty terms which is based on the relatively novel results stated in De La Puna, Lai and Shao

(2008) and Jing, Shao and Wang (2003). Our method also requires the use of an adaptive penalty. Similar procedures are considered in Zou (2006), Huang, Ma and Zhang (2008) and Buehlmann, Van der Geer and Zhou (2011). We propose computationally tractable iterative algorithms that implement the LASSO method proposed in this paper.

In Monte-Carlo examples, we compare the performance of our method to that of the traditional GMM and CUE when the number of moments is comparable to the sample size. We also present the performance of our method when the number of moments is larger than the sample size. We show that in both situations, the LASSO based estimators are more efficient than both GMM and CUE and result in less bias as well.

The paper closest in flavor of this paper is Belloni, Chen, Chernozhukov and Hansen (2012) (later BCCH), which considers a linear IV model with many instruments and selects the informative instruments via a LASSO selection procedure applied in the first stage regression. This paper relies heavily on an approximate sparsity assumption which means that the large set of available instruments contains only a few truly informative ones. The linear structure of the optimal instruments in the first step regression is important in their analysis, while our method does not rely on it. This paper can be considered a direct generalization of traditional regression-LASSO and the optimal IV method proposed in BCCH.

The performance of our procedure is derived based on many results in the LASSO literature and related fields. For the theoretical performance of LASSO, see, for example, Bickel, Ritov and Tsybakov (2009), Belloni and Chernozhukov (2012), Tibshirani (1996), Zhang and Huang (2008). For performance of post-LASSO, see also Belloni and Chernozhukov (2012).

There are also alternative approaches to selecting informative moments. Donald, Imbens and Newey (2008) considers choosing the optimal set of moments via minimizing asymptotic mean-squared-error criteria. Their method is convenient to judge which of two sets of moments is better in terms of smaller asymptotic MSE, however, it is not computationally feasible for selecting the "best model" from a large set of potential sets of moment conditions. Shi (2013) considers a novel "relaxed empirical likelihood" estimator that the number of moments are allowed to increase with speed $O(\exp(n^{\frac{1}{9}}))$, where n is

the sample size. Our method relaxes this constraint to $O(\exp(n^{\frac{1}{3}}))$ when the number of truly informative moments increases slowly enough along with other technical conditions. The estimator proposed in Shi (2013) is also more difficult to compute compared to our method.

As alternatives to the selection approaches in the work mentioned above, there are many methods for correcting second order bias of two-stage GMM with many moments. In the instrumental variables setting, it is well known that LIML and Fuller estimators are robust to the many IV problem. Under GMM, LIML-like estimators, such as CUE proposed in Hansen (1996) and GEL proposed in Imbens (2002), are also robust to the many moments problem. However, the validity of these estimators holds only under the assumption that the number of moment conditions grows at a fractional polynomial rate of the sample size. In contrast to that, our approach allows the number of moment conditions to exceed the sample size. Other studies such as Chao and Swanson (2005), Han and Phillips (2006), Hansen, Hausman and Newey (2008), Chao, Swanson, Hansen, Newey and Woutersen (2012) propose bias-correction methods for many IV problems in different settings.

Another approach to the many moments problem is to acknowledge that the usual variance of GMM estimators seems to be small and produces low coverage in practice. Bekker (1994) proposes a standard deviation robust to the many IV case when the distribution of the residual is normal. Newey and Windjimejer (2009) proposes a variance robust to many moments for the GEL estimator. But again the working assumption of these papers is that the number of moments is growing at most at a fractional polynomial rate of the sample size.

We outline this paper as follows: Section 1.1 introduces the basic settings. Section 1.2 proposes a LASSO method for selecting the informative moments. Section 1.3 presents high level assumptions and theoretical results of the LASSO and post-LASSO estimators. Section 1.4 discusses the validity of preliminary high-level conditions as stated in Section 1.3. Section 1.5 includes Monte-Carlo examples to illustrate the performance of our selection procedure. Section 1.6 concludes the paper.

In the paper we will use the following notations: Let $||\cdot||_2$ be the (Euclidean) L^2

norm of any real vector with any length. Similarly, let $||\cdot||_1$ and $||\cdot||_\infty$ be the L_1 norm and L_∞ norm of a real vector. Let $||\cdot||_0$ be the L_0 norm of a real vector, i.e., the number of non-zero components of the vector.

1.2 Setting

Let us begin with a set of moment conditions

$$\mathbb{E}[g_j(Z, \beta_0)] = 0, j = 1, .., m \tag{1.2.1}$$

that holds uniquely for the true d-dimensional parameter β_0 which lies in the interior of the compact parameter space $\mathcal{D}$. In this paper we treat the dimension d as fixed. Assume we have data Z_i, $i = 1, 2, ..., n$ consisting of independent observations. Let $g(Z, \beta) = (g_1(Z, \beta), g_2(Z, \beta), ..., g_m(Z, \beta))'$.

The main interest of this paper is to explore a situation that arises when the number of moment conditions m is large or may even exceed the sample size n. We will allow the number of moments m_n to increase with n, but we drop the index in order to simplify the notation. The setting with many moment conditions often arise in applications and is very important in empirical practice. Below are the two such examples.

Example 1 (conditional moment restrictions) *Suppose the model is described by conditional moment restriction* $\mathbb{E}[g(x, \beta_0)|z] = 0$, *where* β_0 *is the true parameter. Then the following set of unconditional moments holds:* $E[g(x, \beta_0)f(z)] = 0$, *where* $f(z)$ *can be any set of transformations of* z *such as polynomials, triangular series, splines and so on. In principle, there are an infinitely many number of moment conditions that can be formed. Newey (1989) discusses the optimal moment conditions under in this setting.*

Example 2 (panel data) *Suppose* $\mathbb{E}[y_{i,t}|y_{i,t-1}, ..., y_{i,0}] = \alpha_i + \beta_0 y_{i,t-1}$, $1 \leqslant i \leqslant n, 1 \leqslant t \leqslant T$. *Denote* $\Delta y_{i,t} = y_{i,t} - y_{i,t-1}$. *One can form the following moment conditions for*

any transformation $f(\cdot)$*:*

$$\begin{aligned} \mathbb{E}[\Delta y_{i,t} - \beta_0(\Delta y_{i,t-1})f(y_{i,s})] &= 0, \qquad 1 \leqslant s \leqslant t-2, \\ \mathbb{E}[(y_{i,T} - \beta_0 y_{i,T-1})f(\Delta y_{i,s} - \beta_0 \Delta y_{i,s-1})] &= 0, \quad 1 \leqslant s \leqslant T-1. \end{aligned}$$

Denote $\mathbb{E}_n$ as the empirical average operator. Let $\hat{W}$ be a $m \times m$ semi-positive definite matrix. The GMM estimator is defined as:

$$\hat{\beta}_{GMM} := argmin_\beta \mathbb{E}_n[g(Z,\beta)]' \hat{W} \mathbb{E}_n[g(Z,\beta)].$$

The two-step efficient GMM is the typical method used to obtain efficient estimates within a GMM framework. In two-step efficient GMM, the critical step is to consistently estimate the variance-covariance matrix of the residual $\Omega_0 := \mathbb{E}[g(Z,\beta_0)g(Z,\beta_0)']$. We can estimate the Ω_0 by the following plugged in estimator[1]:

$$\hat{\Omega} := \mathbb{E}_n[g(Z,\tilde{\beta})g(Z,\tilde{\beta})'],$$

where $\tilde{\beta}$ is a preliminary consistent estimator of β_0. If $\hat{\Omega}$ is a consistent positive definite estimator of Ω_0, then the two-step GMM estimator, $\hat{\beta}_{TGMM}$, can be defined as:

$$\hat{\beta}_{TGMM} := argmin_\beta \mathbb{E}_n[g(Z,\beta)]' \hat{\Omega}^{-1} \mathbb{E}_n[g(Z,\beta)].$$

In general the preliminary estimator $\tilde{\beta}$ must be consistent but does not need to be $\sqrt{n}$ consistent. We can obtain the preliminary estimator using the first d moment conditions by setting $\mathbb{E}_n[g_j(Z,\tilde{\beta})] = 0$, $1 \leqslant j \leqslant d$. Or similarly, one is free to select a set of moment conditions (containing at least d moments) which the researcher thinks is important. Throughout the paper, the following general assumption on a preliminary estimator $\tilde{\beta}$ will be made:

Assumption C.1 (Convergence of $\tilde{\beta}$) *There exists an priori estimator* $\tilde{\beta}$ *of* β_0 *and*

[1]In this paper I consider i.i.d. data. For serially correlated data, the $\hat{\Omega}$ can be estimated by the Newey-West estimator, which is semi-positive definite. The logic presented in this paper can be carried over to serially correlated data with more careful attention to detail. I leave the case of serially correlated data as a topic for future research.

a constant $\frac{1}{2} \geqslant \rho > 0$, *such that*

$$||\tilde{\beta} - \beta_0||_2 = O_p\left(n^{-\rho}\right). \tag{1.2.2}$$

The traditional two-step efficient GMM typically has large bias when the number of moments m is large compared to the sample size n. Newey, Donald and Imbens (2008) provides a decomposition of the asymptotic second order bias, which can grow with the number of moments. The main source of such bias arises from poor accuracy of the estimation of $\hat{\Omega}$ when the size of this matrix grows. If m grows fast enough, then estimator $\hat{\Omega}$ may even be inconsistent. The high level of uncertainty in the estimation of Ω_0 causes the instability of the inverse matrix, $\hat{\Omega}^{-1}$, due to the "ill-posedness" problem, as the smallest eigenvalue of $\hat{\Omega}$ can be very close to 0. If one has more moment conditions than available observations ($m > n$), the two-step efficient GMM is not well defined since $\hat{\Omega}$ is not invertible. Thus the main challenge to the behavior of the efficient two-step GMM comes from the estimation of the optimal weighting matrix Ω_0^{-1}.

This paper examines at the problem from a different perspective. Rather than estimating and inverting $\hat{\Omega}$, we are searching for an optimal linear combination of moments that would be the most informative about the parameter β, or equivalently, the optimal combination matrix suggested in Hansen (1982). Hansen (1982) shows that if m is fixed, the $m \times d$ optimal combination matrix $\Omega_0^{-1}G_0(\beta_0)$ can generate an efficient estimator of β_0 by estimating the just identified system of equations:

$$G_0(\beta_0)'\Omega_0^{-1}\mathbb{E}_n[g(Z, \hat{\beta}_C)] = 0, \tag{1.2.3}$$

where $G_0(\beta) := \mathbb{E}[\frac{\partial g(Z,\beta)}{\partial \beta}]$ is the gradient matrix of $\mathbb{E}[g(Z, \beta)]$. Since the above equation is asymptotically equivalent to the first order condition of the two-step efficient GMM for fixed m and growing n, the estimator $\hat{\beta}_C$ is efficient and first order equivalent to the optimal two-step GMM estimator $\hat{\beta}_{TGMM}$. One way to interpret this result is that the two-step efficient GMM procedure tries to find the optimal combination of moment conditions.

In general, estimating the $m \times d$ optimal combination matrix $\Omega_0^{-1}G_0(\beta_0)$ is easier and

more accurate than estimating the optimal weighting matrix Ω_0^{-1}, especially when m is large. The number of elements in the optimal combination matrix remains very large to allow for effective estimation. In this paper we make an assumption on the approximate sparsity of such a matrix, which means that a small number of moment conditions (with unknown indices) contains most of the information about the unknown parameter β contained in the full set of moments. The number of very informative moments, s_n, is unknown and may increase with the sample size but much more slowly than the total number of moments. The sparsity assumption is stated and discussed in detail in Section 1.3.

Given the sparsity assumption on the optimal combination matrix, the main task solved by the paper consists of selecting the informative moments. This task is best performed by employing a special form of the LASSO estimation for the optimal combination matrix that has been adapted to the presence of a poorly invertible covariance matrix. Previously, the LASSO method has been applied to the selection of informative instruments in instrumental variable regression with many potential instruments by BCCH. This paper generalizes this selection idea to a non-linear GMM setting with many moment conditions.

The assumption below allows us to linearize the set of moment conditions even when the number of moments is large. The linear approximation of generally non-linear moments is an important preliminary step in our analysis.

Assumption C.2 *[Regularity conditions on g and G] Suppose the domain of β is a compact set $\Theta \subset \mathbb{R}^d$ and the true parameter β_0 lies in the interior of Θ. There exist an absolute constant K and constants $K_{M,n}$, $K_{G,n}$ and $K_{B,n}$ depending on n only, such that with probability converging to 1 the following statements hold:*

(1) There exists a positive measurable function $K_M(Z)$ which does not depend on n such that for any β and β' in Θ, $\max_{1 \leqslant j \leqslant m} |g_j(Z,\beta) - g_j(Z,\beta')| \leqslant K_M(Z)\|\beta - \beta'\|_2$. $\mathbb{E}[K_M(Z)] \leqslant K$ and $\max_{1 \leqslant i \leqslant n} K_M(Z_i) \leqslant K_{M,n}$;

(2) There exists a positive measurable function $K_M(Z)$ which does not depend on n such that for any β and β' in Θ, $\max_{1 \leqslant j \leqslant m} \|\frac{\partial g_j}{\partial \beta}(Z,\beta) - \frac{\partial g_j}{\partial \beta}(Z,\beta')\|_2 \leqslant K_G(Z)\|\beta - \beta'\|_2$. $\mathbb{E}[K_G(Z)] \leqslant K$ and $\max_{1 \leqslant i \leqslant n} K_G(Z) \leqslant K_{G,n}$;

(3) $\max_{1\leqslant j\leqslant m}\mathbb{E}_n[|g_j(Z,\beta_0)|^2]\leqslant K$, $\max_{1\leqslant j\leqslant m}\mathbb{E}[|g_j(Z,\beta_0)|^3]\leqslant K$.

(4) $\max_{1\leqslant j\leqslant m,1\leqslant i\leqslant n}|g_j(Z_i,\beta_0)|\leqslant K_{B,n}$.

(5) $G_0(\beta_0)'\Omega_0^{-1}\mathbb{E}[g(Z,\beta)]=0$ *holds uniquely for* $\beta=\beta_0$ *in* Θ. *For any* $\xi>0$, *there exists* $\eta>0$ *which does not depend on* n *and* m *such that for any* β *with* $||\beta-\beta_0||_2>\xi$, $||G_0(\beta_0)'\Omega_0^{-1}\mathbb{E}[g(Z,\beta)]||_2>\eta$.

Assumption C.2 puts restrictions on the smoothness of the moment conditions. Constants $K_{G,n}$, $K_{M,n}$ and $K_{B,n}$ typically increase with n as the number of moment conditions is growing. The constraints on the speed with which they increase is stated in the Section 1.5. Under the conditions described in the Section 5, statement (3) of Assumption C.2 is implied by statement (4) if $K_{B,n}$ grows slowly enough. Statement (5) guarantees identification and consistency of the GMM estimator.

In addition, we assume here that the information in the full set of moment conditions is limited and in particular the super-consistent estimators of β are ruled out. This assumption below also rules out the weak identification problem. This assumption is generally true for conditional moment restriction settings.

Assumption C.3 (Limited Information) *Assume the maximal and minimum eigenvalues of*
$G_0(\beta_0)'\Omega_0^{-1}G_0(\beta_0)$ *are bounded away from below and above by absolute constants.*

1.3 LASSO and Sparsity

1.3.1 Formulation of LASSO estimation

The main task of this paper is to estimate the optimal combination matrix $\Omega_0^{-1}G_0(\beta_0)$. Let I_d be the identity matrix of dimension $d\times d$ and let e_l be the l^{th} column of I_d, $1\leqslant l\leqslant d$. To estimate the $m\times d$ optimal combination matrix $\Omega_0^{-1}G_0(\beta_0)$, it suffices to estimate

$\Omega_0^{-1}G_0(\beta_0)e_l$ for all $1 \leqslant l \leqslant d$. Let us fix the vector v and estimate $\lambda^*(v) := \Omega_0^{-1}G_0(\beta_0)v$, or we simply write λ^* for notational convenience when there is no confusion.

Let us define the estimator $\hat{\lambda}$ for λ^* as the solution to the following minimization problem $\mathcal{P}$:

$$\mathcal{P} : \min_{\lambda} \frac{1}{2}\lambda'\hat{\Omega}\lambda - \lambda'\hat{G}(\tilde{\beta})v + \sum_{j=1}^{m} \frac{t}{n}|\lambda_j\gamma_j|, \tag{1.3.1}$$

where $\hat{G}(\beta) := \mathbb{E}_n[\frac{\partial g}{\partial \beta}(Z,\beta)]$, $\gamma_j > 0$ is the moment-specific penalty loading for the j^{th} moment condition, $1 \leqslant j \leqslant m$, and $t > 0$ is the uniform penalty loading.[2]

The problem $\mathcal{P}$ consists of two components: the objective function $\hat{Q}(\lambda) := \frac{1}{2}\lambda'\hat{\Omega}\lambda - \hat{G}(\tilde{\beta})v$ and the penalty $\sum_{j=1}^{m} \frac{t}{n}|\lambda_j\gamma_j|$. If $\hat{\Omega}$ is invertible, the minimizer of the objective function $\hat{Q}(\lambda)$ alone is $\hat{\Omega}^{-1}\hat{G}(\tilde{\beta})v$, which can serve as a good estimator of λ^* when the number of moments is fixed. The penalty terms t and γ_j, $1 \leqslant j \leqslant m$ should be chosen in such a way that small coefficients in λ^* shrink to 0, and large coefficients remain non-zero. Thus, the solution to the minimization problem with the appropriate penalty, $\hat{\lambda}$, has non-zero coefficients only for the moments which contain significant information on the unknown parameter β_0. Hence, $\mathcal{P}$ can be interpreted as a moment selection mechanism.

The minimization problem $\mathcal{P}$ also has a computational advantage compared to other methods such as the moment selection mechanism proposed in Donald, Imbens and Newey (2008). The objective function $\hat{Q}(\lambda)$ is convex, and the penalty function $\sum_{j=1}^{m} \frac{t}{n}|\lambda_j\gamma_j|$ is strictly convex. Thus, the solution to the problem $\mathcal{P}$ is unique. The minimization

---

[2] *Economists may have primitive information (which could come from either economic models or intuition) that a subset of moments should always be included in a GMM model. In practice we can assume that there are two sets of moment conditions. The first set, the baseline group, contains moment conditions with indices $1, 2, ..., B$. The second set, the additional group, contains moment conditions with indices $B+1, ..., m$. The baseline group is assumed to be economically important, and therefore, this group of moment conditions always needs to be considered. To avoid excluding any moment conditions in the baseline group, the selection mechanism $\mathcal{P}$ can be modified as follows:*

$$\mathcal{P}_1 : \min_{\lambda} \frac{1}{2}\lambda'\hat{\Omega}\lambda - \lambda'\hat{G}v + \sum_{j=B+1}^{m} \frac{t}{n}|\lambda_j\gamma_j|.$$

In this paper, we focus on the analysis of $\mathcal{P}$. All results for $\mathcal{P}$ can be carried over to $\mathcal{P}_1$ under exactly the same conditions.

procedure can be performed with any convex minimization algorithms like the Shooting algorithm, for example. These algorithms typically converge in $O(m\log(m))$ time, compared to $O(2^m)$ as proposed in Donald, Imbens and Newey (2008).

The selection procedure $\mathcal{P}$ described in equation (1.3.1) is a generalization of the first stage IV selection procedure proposed in BCCH for homoskedastic models.

Example 3 (Many IV) *Assume we observe data from a linear IV model:*

$$Y = X\beta + W\gamma + U,$$

$$X = Z\Pi + V,$$

with d-dimensional regressor X, m-dimensional instruments Z, and homoskedastic error term U. BCCH (2012) considers the following LASSO approach applied to the first stage regression:

$$\min_{\Pi_l} \mathbb{E}_n[(X_l - Z\Pi_l)^2] + \sum_{j=1}^{m} \frac{2t}{n}|\Pi_{lj}\tilde{\gamma}_{lj}|. \tag{1.3.2}$$

In the above equation, t is the uniform penalty and $\tilde{\gamma}_{lj}$ is the moment specific penalty for the endogenous variable X_l, $1 \leqslant l \leqslant d$.

If we rewrite this within the GMM framework, the moment conditions are $\mathbb{E}[Z'(Y - X\beta)] = 0$. Consequently, $\hat{G}(\tilde{\beta}) = \mathbb{E}_n Z'X$, $G_0(\beta) := \mathbb{E}[Z'X]$, and $\Omega_0 = \mathbb{E}[Z'UU'Z] = \sigma_u^2\mathbb{E}[Z'Z]$. Let $\hat{\Omega} := \hat{\sigma}_u^2\mathbb{E}_n[Z'Z]$, where $\hat{\sigma}_u^2 > 0$ is an estimator of σ_u^2. Then, the selection mechanism $\mathcal{P}$ for $\lambda^*(e_l)$ can be written as:

$$\min_{\lambda} \frac{1}{2}\hat{\sigma}_u^2\lambda'\mathbb{E}_n[Z'Z]\lambda - \lambda'\mathbb{E}_n[Z'X_l] + \sum_{j=1}^{m} \frac{t}{n}|\lambda_j\gamma_j|. \tag{1.3.3}$$

If $\tilde{\gamma}_{jl} = \hat{\sigma}_u^2\gamma_j$ then optimization problems (1.3.2) and (1.3.3) are equivalent, in particular, $\Pi_l = \hat{\sigma}_u^2\lambda$.

Furthermore, the formulation of $\mathcal{P}$ also includes the OLS regression with LASSO penalties:

Example 4 (Regression LASSO) *Suppose we have the OLS equation* $Y = X\beta + \epsilon$. *The regression LASSO is:*

$$\min_{\beta} \mathbb{E}_n[(Y - X\beta)^2] + \frac{2t}{n}||\beta||_1. \tag{1.3.4}$$

If $\gamma_j = 1$ *for all* j, *problem* $\mathcal{P}$ *is identical to the regression LASSO in equation (1.3.4), since* $\hat{\Omega} = \mathbb{E}_n[X'X]$ *and* $\hat{G}(\tilde{\beta})v := \mathbb{E}_n[XY]$.

In this paper we investigate the performance of two estimators, the LASSO estimator $\hat{\beta}_L$ and the post-LASSO estimator $\hat{\beta}_{PL}$, which are defined below. Let $\hat{\lambda}(l)$ be the solution of the optimization problem $\mathcal{P}$ for $\lambda^*(e_l)$, $\hat{T}_l$ be the set of indices of non-zero components of $\hat{\lambda}(l)$, $\hat{T} = \cup_{l=1}^{d} \hat{T}_l$ and $g_{\hat{T}}(Z, \beta)$ be the vector containing only moments with indices in $\hat{T}$. Define

$$\hat{\beta}_L = \text{argmin}_{\beta \in \mathcal{D}} \sum_{1 \leqslant l \leqslant d} (\hat{\lambda}(l)' \mathbb{E}_n[g(Z, \beta)])^2,$$

$$\hat{\beta}_{PL} = \text{argmin}_{\beta \in \mathcal{D}} \mathbb{E}_n[g_{\hat{T}}(Z, \beta)]' \hat{\Omega}_{\hat{T}}^{-1} \mathbb{E}_n[g_{\hat{T}}(Z, \beta)],$$

where $\hat{\Omega}_{\hat{T}} = \mathbb{E}_n[g_{\hat{T}}(Z, \tilde{\beta}) g_{\hat{T}}(Z, \tilde{\beta})']$.

When the informative moment conditions are rare among a full set of moments, the LASSO estimator $\hat{\beta}_L$ and the post-LASSO estimator $\hat{\beta}_{PL}$ are expected to perform well under the sparsity assumptions proposed in the next subsection. These two estimators are less biased compared to two-step efficient GMM simply because much significantly fewer moments are used in the second step of the estimation procedure, and these estimators are nearly efficient since the most informative moments are preserved by the selection mechanism.

1.3.2 Sparsity Assumption

The LASSO approach performs extremely well under certain sparsity assumptions on the high-dimensional parameter, as shown in Belloni and Chernozhukov (2011a), (2011b), Belloni, Chernozhukov and Hansen (2012), BCCH (2012) and Bickel et al. (2008). Sim-

ilarly, we propose the following approximate sparsity assumption which adapts specifically to the analysis under a GMM framework. We begin with an exact sparsity assumption which rarely holds in practice but provides additional theoretical properties to the selection procedure. Then we show how this assumption may be weakened. Let us now fix $v \in \mathbb{R}^d$, $||v||_2 = 1$, and consider the combination vector $\lambda^* := \Omega_0^{-1} G_0(\beta_0) v$.

Assumption C.4 (Sparse Combination Matrix) *Denote $s_n := ||\lambda^*||_0$ to be the number of non-zero components of λ^*.*

(1) $s_n = o(n)$;

(2) there exists a generic constant K_1 such that $||\lambda^*||_1 \leqslant K_1$.

The exact sparsity assumption imposes the restriction that most of the elements of λ^* must be zero, though their indices are unknown. The number of non-zero coefficients, s_n, is also assumed to be unknown to the researcher. Our results will typically impose rate restrictions on s_n allowing it to increase, but not too quickly. If the exact sparsity condition holds, then by choosing the correct penalties t and γ, the selection procedure $\mathcal{P}$ will possess the oracle property, i.e., the identity of non-zero coefficients in λ is recovered with probability going to 1. Such an oracle property has been discussed previously in the LASSO literature under exact sparsity conditions, for example, Bunea et al. (2007) Zou (2006). We discuss the oracle property of $\mathcal{P}$ in Section 1.4.

Typically, the exact sparsity assumption is too strong to be relevant to most applications, so a much weaker assumption is used to achieve the main results about the good performance of LASSO and post-LASSO estimators.

Assumption C.5 (Approximate Sparse Combination Matrix) *Suppose there exist absolute positive constants K_λ^u, K_λ^l and K_r and a non-stochastic $m \times 1$ vector $\tilde{\lambda}$ such that:*

(1) $||\tilde{\lambda}||_0 = s_n = o(n)$.

(2) $K_\lambda^l \leqslant ||\tilde{\lambda}||_1 \leqslant K_\lambda^u$.

(3) $||\tilde{\lambda} - \lambda^*||_1 = o\left(\sqrt{\frac{\log(m \vee n)}{n}}\right)$.

Assumption C.5 applied d times to $v = e_1, e_2, ..., e_d$ implies that the optimal combination matrix $\Omega_0^{-1}G_0(\beta_0)$ can be approximated by a matrix with only a few non-zero components. In statement (3), the quality of the approximation is measured by a bound on L_1 distance between λ^* and $\tilde{\lambda}$. The true vector λ^* may not even have zero coefficients at all, but its elements should shrink quickly enough, as we described in the example below.

Example 5 *Let $\lambda^*_{(j)}$ be the j^{th} largest (in absolute value) component of λ^*. Assume that the absolute values of all components of λ^* are different, and $|\lambda^*_{(j)}| = O(j^{-q})$, where $q > \frac{3}{2}$. Assume also that $\max_{1\leqslant j\leqslant m}\mathbb{E}[||\frac{\partial g_j}{\partial \beta}(Z_i,\beta_0)||_2] \leqslant K_0$, with K_0 being an absolute constant. Then λ^* is approximately sparse with $s_n = \lceil n^{\frac{1}{2q-2}}\rceil$. The approximating vector $\tilde{\lambda}$ can be chosen as*

$$\tilde{\lambda}_j = \lambda^*_j \cdot \mathbb{I}\left\{\lambda^*_j \geqslant \lambda^*_{(s_n)}\right\}.$$

The approximate sparsity assumption C.5 implies no super consistency of the assumption C.3 under mild regularity conditions.

Lemma 1 *Suppose the assumption C.5 holds for $v = e_l$ for $l = 1, ..., d$. Assume there exists an absolute constant K_0 such that $\max_{1\leqslant j\leqslant m}\mathbb{E}[||\frac{\partial g_j}{\partial \beta}(Z_i,\beta_0)||_2] \leqslant K_0$. Then if m grows at rate $m = O(\exp(n))$, the full set of moments does not have superconsistency, i.e., the maximal eigenvalue of $G_0(\beta_0)'\Omega_0^{-1}G_0(\beta_0)$ is bounded from above.*

1.4 Main Results

In this section, we establish our main results by employing three high level assumptions that are often used in LASSO analysis. In the next section we discuss what primitive assumptions imply the validity of these high level conditions.

The inversion of matrix $\hat{\Omega}$ may be an undesirable estimator (if it exists) of Ω_0^{-1}, however, the inversion of diagonal submatrices with size $s \times s$ of $\hat{\Omega}$ can be stable when s is small compared to n. Recall that for any $\delta \in \mathbb{R}^m$, $||\delta||_0$ is the number of non-zero

components of δ. For a semi-positive definite matrix M, we define κ and ϕ as lower and upper bounds of eigenvalues of all diagonal submatrices of size, at most, $s \times s$:

Definition 1.4.1 *For any positive real number $s \geqslant 1$ and a $m \times m$ semi-positive definite matrix M, define $\kappa(s, M)$ and $\phi(s, M)$ as:*

$$\kappa(s, M) := \min_{\delta \in \mathbb{R}^m, ||\delta||_0 \leqslant s, \delta \neq 0} \frac{\delta' M \delta}{||\delta||^2}, \tag{1.4.1}$$

$$\phi(s, M) := \max_{\delta \in \mathbb{R}^m, ||\delta||_0 \leqslant s, \delta \neq 0} \frac{\delta' M \delta}{||\delta||^2}. \tag{1.4.2}$$

Assumption C.6 (Eigenvalues of sub-matrices) *There exist constants $0 \leqslant \kappa_1 \leqslant \kappa_2$ such that with probability increasing to one as the sample size grows we have*

$$\kappa_1 \leqslant \kappa(log(n)s_n, \hat{\Omega}) \leqslant \phi(log(n)s_n, \hat{\Omega}) \leqslant \kappa_2.$$

The high level assumption C.6 allows us to robustly invert any diagonal, square, sub-matrix of $\hat{\Omega}$ of size at most $O(s_n)$ that grows more slowly than n. This will be the key assumption that will guarantee the good asymptotic behavior of LASSO and post-LASSO estimators. The validity of Assumption C.6 essentially depends on the growing speed of sequence $K_{B,n}$ as defined in Assumption C.2 and on the accuracy of the preliminary estimator $\tilde{\beta}$. Section 1.5.1 discusses the primitive conditions necessary for Assumption C.6 to hold.

Recall that the moment selection problem $\mathcal{P}$ as stated in (3.1.3.1) consists of two components: the objective function $\frac{1}{2}\lambda'\hat{\Omega}\lambda - \lambda'\hat{G}(\tilde{\beta})v$ and the penalty term $\sum_{j=1}^{m} \frac{t}{n}|\lambda_j \gamma_j|$. Let us define the score function $\hat{S}(\lambda)$ as the derivative of the objective function, i.e.,

$$\hat{S}(\lambda) = \hat{\Omega}\lambda - \hat{G}(\tilde{\beta})v.$$

The second high-level assumption needed to analyze the performance of LASSO is that the score function evaluated at $\tilde{\lambda}$ is dominated by the vector of the penalties.

Assumption C.7 (Dominance of Penalty) *For a given sequence of positive numbers α_n converging to zero we have*

$$\mathrm{P}\left(\max_{1\leqslant j\leqslant m}\left|\frac{\hat{S}(\tilde{\lambda})_j}{\gamma_j}\right| < \frac{t}{n}\right) \geqslant 1-\alpha_n, \tag{1.4.3}$$

where $\hat{S}(\tilde{\lambda})_j$ is the j^{th} entry of $\hat{S}(\tilde{\lambda})$.

Assumption C.7 guarantees that the penalty is harsh enough and thus a relatively small number of moments will be chosen by the LASSO selection procedure. The validity of Assumption C.7 is guaranteed by the proper choice of penalties t and γ_j. The choice of these penalties is discussed in Sections 1.5.2 and 1.5.3, where a feasible procedure for choosing penalties is put forward. Let ϵ be an absolute positive constant all throughout the remainder of this paper.[3] For a given vector v, the traditional choice of t is $t = (1+\epsilon)\sqrt{n\Phi^{-1}(1-\frac{2m}{\alpha_n})}$. When $t = (1+\epsilon)\sqrt{n\Phi^{-1}(1-\frac{2md}{\alpha_n})}$, the dominance condition stated in Assumption C.7 can hold uniformly for $v = e_1, .., e_d$ with probability at least $1-\alpha_n$, which means that we can put an additional supremum over $l = 1, ..., d$ inside the probability in equation (1.4.3).

Assumption C.8 (Bounded Penalty) *There exist absolute, positive constants a and b such that $a \leqslant \min_{1\leqslant j\leqslant m}\gamma_j \leqslant \max_{1\leqslant j\leqslant m}\gamma_j \leqslant b$ with probability increasing to one.*

With the three high level assumptions C.6-C.8 stated above, we are now able to derive the main results on the performance of the LASSO estimator $\hat{\beta}_L$ and the post-LASSO estimator $\hat{\beta}_{PL}$. For any $\delta \in \mathbb{R}^m$, define the semi-norm $||\delta||_{2,n} := \delta'\hat{\Omega}\delta$.

Theorem 1 (LASSO estimator of β) *Consider optimization problems $\mathcal{P}$ for $v = e_l, l = 1, .., d$. Let $\hat{\lambda}(l)$ be the solution to the problem of estimating $\tilde{\lambda}(l)$, which is a sparse approximation of $\lambda^*(l) = \Omega_0^{-1}G_0(\beta_0)e_l$. Suppose Assumptions C.1-C.3 and C.5-C.8 hold for all $v = e_1, ..., e_d$ with $t = (1+\epsilon)\sqrt{n\Phi^{-1}(1-\frac{2md}{\alpha_n})}$, $\alpha_n \to 0$ and $m\alpha_n \to \infty$. Additionally we require that Assumption C.7 holds uniformly for $v = e_1, e_2, ..., e_d$.*

[3] The traditional recommendation (e.g. BCCH (2012)) of the value ϵ is 0.1.

Then there exists an absolute constant K_λ and a sequence $\epsilon_n \to 0$ such that with probability at least $1 - \alpha - \epsilon_n$, the following statements are true:

$$\max_{l \leqslant l \leqslant d} ||\hat{\lambda}(l) - \tilde{\lambda}(l)||_{2,n} \leqslant K_\lambda \sqrt{\frac{s_n \log(\frac{md}{\alpha_n})}{n}}, \tag{1.4.4}$$

and

$$\max_{l \leqslant l \leqslant d} ||\hat{\lambda}(l) - \tilde{\lambda}(l)||_1 \leqslant K_\lambda \sqrt{\frac{s_n^2 \log(\frac{md}{\alpha_n})}{n}}. \tag{1.4.5}$$

If, in addition, we have $s_n^2(K_{M,n} \vee K_{B,n})^2 \log(m) = o(n)$, where $K_{M,n}$ and $K_{B,n}$ are constants as defined in Assumption C.2, then the LASSO estimator $\hat{\beta}_L$ has the following rate :

$$||\hat{\beta}_L - \beta_0||_2 = O_p\left(\frac{1}{\sqrt{n}} \vee \frac{s_n \log(m)}{n}\right). \tag{1.4.6}$$

Furthermore, if $s_n^2 \log(m)^2 = o(n)$, the estimator $\hat{\beta}_L$ is asymptotically normal:

$$\sqrt{n}(\hat{\beta}_L - \beta_0) \to_d N(0, (G_0(\beta_0)' \Omega_0^{-1} G_0(\beta_0))^{-1}). \tag{1.4.7}$$

Theorem 1 considers the case of approximate sparsity, that is, when only several (s_n) moment conditions are truly informative, while at the same time many coefficients in the optimal combination matrix may be non-zero. The LASSO selection procedure tries to estimate approximate combination vectors $\tilde{\lambda}(l)$ for $l = 1, ..., d$ rather than optimal combinations $\lambda^*(l)$. Equations (1.4.4) and (1.4.5) state the accuracy with which this estimation occurs in L_1 metric and the semi-norm $|| \cdot ||_{2,n}$ correspondingly. Since the dimensionality of vectors, m, is increasing to infinity, these metrics are different. The two terms on the right hand side of statement (1.4.6) provide rates for the LASSO estimator. The first of them corresponds to the variance, while the second relates to the bias. The bias of the LASSO estimator arises from the inversion of matrices of size $O(s_n)$ and correlation between $\hat{\lambda}(l)$ and the residual $\mathbb{E}_n[g(Z_i, \beta_0)]$. Statement (1.4.7) is

obtained if the bias term is stochastically dominated by the uncertainty of the LASSO estimator. It is important to notice that according to statement (1.4.7) the LASSO estimator is asymptotically efficient, even though we have effectively eliminated nearly uninformative moments.

Set $\hat{T} = \cup_{l=1}^{d} \hat{T}_l$ serves as a moment selector, where $\hat{T}_l$ is the set of indices of non-zero components of $\hat{\lambda}(l)$. Denote $T_{0,l}$ as the set of indices of non-zero components of $\tilde{\lambda}(l)$ and $T_0 := \cup_{l=1}^{d} T_{0,l}$. The post-LASSO estimator simply deletes moment conditions with indices outside $\hat{T}$ and performs the two-step efficient GMM on selected moments only. Lemma 2 below shows that $\hat{T}$ is a suitable estimator of the set T_0 under similar regularity conditions as stated in Theorem 1. In particular $\hat{T}$ has $O_p(s_n)$ elements and, if all the non-zero elements of $\tilde{\lambda}$ are large enough, then we are able to uncover all those elements asymptotically.

Lemma 2 (Selector $\hat{T}$) *Suppose Assumptions C.1-C.3 and C.5- C.8 hold for all $v = e_1, ..., e_d$ with $t = (1+\epsilon)\sqrt{n\Phi^{-1}(1-\frac{2md}{\alpha_n})}$, $\alpha_n \to 0$ and $m\alpha_n \to \infty$. Additionally we require that Assumption C.7 holds uniformly for $v = e_1, e_2, ..., e_d$. If $\frac{s_n^2 \log(m)}{n} \to 0$, then there exists a sequence $\epsilon_n \to 0$ such that with probility at least $1 - \alpha_n - \epsilon_n$,*

(1) $|\hat{T}| = O(s_n)$.

(2) $\frac{\sqrt{n} \min_{1 \leqslant l \leqslant d, j \in T_{0,l}} |\tilde{\lambda}(l)_j|}{s_n^2 \log(m)} \to \infty$, then $T_0 \subset \hat{T}$.

Theorem 2 (post-LASSO estimator of β) *Suppose Assumptions C.1-C.3 and C.5-C.8 hold for all $v = e_1, ..., e_d$ with $t = (1+\epsilon)\sqrt{n\Phi^{-1}(1-\frac{2md}{\alpha_n})}$, $\alpha_n \to 0$ and $m\alpha_n \to \infty$. Additionally we require that Assumption C.7 holds uniformly for $v = e_1, e_2, ..., e_d$. If $\frac{s_n^2 \log(m)}{n}(K_{G,n} \vee K_{M,n})^2 \to 0$, then there exists a sequence $\epsilon_n \to 0$ such that with probability at least $1 - \alpha - \epsilon_n$,*

$$||\hat{\beta}_{PL} - \beta_0||_2 = O_p\left(\frac{1}{\sqrt{n}} \vee \frac{s_n \log(m)}{n}\right).$$

Furthermore, if $s_n^2 \log(m)^2 (K_{M,n} \vee K_{B,n})^2 = o(n)$, the estimator $\hat{\beta}_{PL}$ is asymptotically normal:

$$\sqrt{n}(\hat{\beta}_{PL} - \beta_0) \to_d N(0, (G'\Omega_0^{-1}G)^{-1}).$$

The rates and asymptotic properties of the post-LASSO estimator stated in Theorem 2 are identical to those of the LASSO estimator stated in Theorem 1. However, it is reasonable to expect that the post-LASSO estimator has better finite sample performance. If the exact sparsity assumption C.4 is true, then we obtain a stronger result often referred as the Oracle Property, which means that the post-LASSO estimator obeys asymptotics as if the true model were being used for estimation. Therefore, the post-LASSO estimator may achieve asymptotic normality under weaker restrictions on the rate of s_n.

Corollary 1 (Oracle Property under Exact Sparsity) *Suppose Assumptions C.1-C.4 and C.6-C.8 hold for all* $v = e_1, ..., e_d$ *with* $t = (1+\epsilon)\sqrt{n\Phi^{-1}(1-\frac{2md}{\alpha_n})}$, $\alpha_n \to 0$ *and* $m\alpha_n \to \infty$. *We additionally require that Assumption C.7 holds uniformly for* $v = e_1, e_2, ..., e_d$. *If* $\frac{\sqrt{n}\min_{1\leqslant l\leqslant d, j\in T_{0,l}}|\tilde{\lambda}(l)_j|}{s_n^2\log(m)} \to \infty$ *and* $s_n^2\log(m)(K_{M,n}\vee K_{B,n})^2 = o(n)$, *then the post Selection estimator* $\hat{\beta}_{PL}$ *is asymptotically normal:*

$$\sqrt{n}(\hat{\beta}_{PL} - \beta_0) \to_d N(0, (G'\Omega_0^{-1}G)^{-1}).$$

1.5 Primitive Conditions for Assumptions C.6-C.8

1.5.1 Primitive Conditions for Assumption C.6

Assumption C.6 places restrictions on eigenvalues of any diagonal submatrices of $\hat{\Omega} = \mathbb{E}_n\left[g(Z_i,\tilde{\beta})g(Z_i,\tilde{\beta})'\right]$ of size at most $s_n\log(n)$. The validity of this assumption hinges heavily on two statements. First, we need a similar property to hold for the empirical covariance matrix $\hat{\Omega}_0 = \mathbb{E}_n\left[g(Z_i,\beta_0)g(Z_i,\beta_0)'\right]$ evaluated at the true β_0 rather than the preliminary estimated $\tilde{\beta}$. Second, we need the difference between $\hat{\Omega}$ and $\hat{\Omega}_0$ to be small enough.

Assumption C.9 (Eigenvalues of submatrices of $\hat{\Omega}_0$) *There exist positive constants* $\kappa_{1,0} \leqslant \kappa_{2,0}$ *such that with probability increasing to one,* $\kappa_{1,0} \leqslant \kappa(s_n\log(n), \hat{\Omega}_0) \leqslant \phi(s_n\log(n), \hat{\Omega}_0) \leqslant \kappa_{2,0}$.

Assumptions similar to Assumption 9 are common in the LASSO literature. For example, in an OLS-LASSO with model $Y = X\beta + u$, $\hat{\Omega}_0 = \mathbb{E}_n[X'X]$. Tibshirani (1990) makes the assumption that eigenvalues of diagonal sub-matrices of $\mathbb{E}_n[X'X]$ has rate similar to that stated in Assumption C.9. In IV-LASSO with model $Y = X\beta + u$ and $X = Z\Pi + v$, $\hat{\Omega}_0 = \mathbb{E}_n[Z'Z]$. BCCH makes the assumption that $\mathbb{E}_n[Z'Z]$ satisfies Assumption C.9. Belloni and Chernozhukov (2011) constructs preliminary conditions for Assumption C.9 under a Gaussian assumption on $g(Z_i, \beta_0)$. BCCH proves Assumption C.9 by imposing conditions on the speed of growth of $K_{B,n}$, a constant defined in Assumption C.2. We combine the facts stated in the above two papers into Lemma 3:

Lemma 3 (Sufficient condition for Assumption C.9) *Suppose there exist positive absolute constants a_1 and a_2 such that*

$$a_1 \leqslant \kappa(s_n \log(n), \Omega_0) \leqslant \phi(s_n \log(n), \Omega_0) \leqslant a_2.$$

(1) If $K_{B,n} s_n \log^2 n \log^2(s_n \log n) \log(m \vee n) = o_p(n)$, then Assumption C.9 holds.

(2) Suppose that $g(Z_i, \beta_0)$ are i.i.d. Gaussian random vectors with mean 0, $1 \leqslant i \leqslant n$. Then if $s_n \log(n) \log(m) = o(n)$, Assumption C.9 holds.

Lemma 4 (Primitive conditions for Assumption C.6) *Suppose conditions C.1 and C.2 hold. If $\frac{s_n \log(n) K_{M,n}^2}{n^{2\rho}} \to 0$ as n goes to infinity, then, $\phi(s_n \log(n), \hat{\Omega} - \hat{\Omega}_0) \to_p 0$. If in addition to that Assumption C.9 holds, then Assumption C.6 holds as well.*

1.5.2 Bounds on Score Function

In this subsection we discuss about primitive conditions to satisfy Assumptions C.7 and C.8. Assumption C.7 requires that the penalty terms γ_j, $j = 1, 2, ..., m$ be large enough.

Assumption C.10 (Bounds on Higher Order Moments) *Assume there exist absolute constants C_1 and C_2 such that*

(1) $\max_{1\leqslant j\leqslant m}\mathbb{E}\left[g_j(Z_i,\beta_0)^6\right]\leqslant C_1$,

(2) $\mathbb{E}\left[\left\|\frac{\partial g_j}{\partial\beta}(Z_i,\beta_0)\right\|_2^3\right]\leqslant C_2$.

For a given fixed $v\in\mathbb{R}^d$, we establish bounds on the score function $\hat{S}(\tilde{\lambda})$. Lemma 5 below suggest one potential choice of penalties which, however, is not feasible in practice.

Lemma 5 (Ideal choice of penalty levels) *Suppose Assumptions C.1, C.2, C.5 and C.10 hold. Let* $t_0=\sqrt{n\Phi^{-1}(1-\frac{4m}{\alpha_n})}$. *Assume that* $\alpha_n\to 0$ *and* $m\alpha_n\to\infty$. *Assume that* $\frac{\log(m)}{n^{1/3}}\to 0$. *Let*

$$\gamma_j:=(1+c)(K_{G,n}+2K_{M,n}||\tilde{\lambda}||_1)\frac{\sqrt{n}||\tilde{\beta}-\beta_0||_2}{\Phi^{-1}\left(1-\frac{\alpha_n}{4m}\right)} \tag{1.5.1}$$

$$+\left\{\mathbb{E}_n\left(\sum_{k=1}^m g_k(Z_i,\beta_0)g_j(Z_i,\beta_0)\tilde{\lambda}_k\right)^2-\left[\mathbb{E}_n\left(\sum_{k=1}^m g_k(Z_i,\beta_0)g_j(Z_i,\beta_0)\tilde{\lambda}_k\right)\right]^2\right\}^{\frac{1}{2}}$$

$$+\left\{\mathbb{E}_n\left(\frac{\partial g_j}{\partial\beta}(Z_i,\beta_0)v\right)^2-\left[\mathbb{E}_n\frac{\partial g_j}{\partial\beta}(Z_i,\beta_0)v\right]^2\right\}^{\frac{1}{2}}+\frac{\sqrt{n}\left|(\Omega_0(\tilde{\lambda}-\lambda^*))_j\right|}{\Phi^{-1}\left(1-\frac{\alpha_n}{4m}\right)},$$

where $1>c>0$ *is an arbitrarily small absolute constant.*

Then there exists a sequence $\epsilon_n\to 0$ *such that*

$$\mathrm{P}\left(\sqrt{n}\max_{1\leqslant j\leqslant m}\left|\frac{\hat{S}_j(\tilde{\lambda})}{\gamma_j}\right|\leqslant\frac{t_0}{n}\right)\geqslant 1-\alpha_n-\epsilon_n. \tag{1.5.2}$$

The penalty γ_j suggested in Lemma 5 is not feasible in practice since we do not have any knowledge about β_0 or $\tilde{\lambda}$ or about any bounding constants. These penalty terms are also complicated to compute. In many situations we can choose seemingly simpler penalties than the ones suggested in equation (1.5.1).

Corollary 2 (Asymptotic penalty) *Assume that all conditions of Lemma 5 hold and, in addition,* $\frac{(K_{G,n}\vee K_{M,n})^2}{n^{2\rho-1}\log(m)}\to 0$. *Define the following two sets of penalties:*

(1) Refined asymptotic penalty γ_j^R, $1 \leqslant j \leqslant m$:

$$\gamma_j^R := \left\{ \mathbb{E}_n \left(\sum_{k=1}^{m} g_k(Z_i, \beta_0) g_j(Z_i, \beta_0) \tilde{\lambda}_k \right)^2 - \left[\mathbb{E}_n \left(\sum_{k=1}^{m} g_k(Z_i, \beta_0) g_j(Z_i, \beta_0) \tilde{\lambda}_k \right) \right]^2 \right\}^{\frac{1}{2}} + \left\{ \mathbb{E}_n \left(\frac{\partial g_j}{\partial \beta}(Z_i, \beta_0) v \right)^2 - \left[\mathbb{E}_n \frac{\partial g_j}{\partial \beta}(Z_i, \beta_0) v \right]^2 \right\}^{\frac{1}{2}}. \tag{1.5.3}$$

(2) Coarse asymptotic penalty γ_j^C, $1 \leqslant j \leqslant m$:

$$\gamma_j^C = ||\tilde{\lambda}||_1 \left\{ \max_{k \in T_0} \left(\mathbb{E}_n g_k(Z_i, \beta_0)^4 - [\mathbb{E}_n g_k(Z_i, \beta_0)^2]^2 \right) \right\}^{\frac{1}{2}} + \left\{ \mathbb{E}_n \left(\frac{\partial g_j}{\partial \beta}(Z_i, \beta_0) v \right)^2 - \left[\mathbb{E}_n \frac{\partial g_j}{\partial \beta}(Z_i, \beta_0) v \right]^2 \right\}^{\frac{1}{2}}. \tag{1.5.4}$$

If both γ^R *and* γ^C *satisfy Assumption C.8, then there exists a sequence* $\epsilon_n \to 0$ *such that statement (1.5.2) holds with* $\gamma_j = \gamma_j^R$ *or* $\gamma_j = \gamma_j^C$ *for all* $1 \leqslant j \leqslant m$.

The vector of refined penalties γ^R is sharper than the vector of coarse penalties γ^C, that is, for all realizations we have $\gamma_j^R \leqslant \gamma_j^C$. In simulations we find that the refined penalties result in estimators with smaller mean squared errors in finite samples, however, in practice, the refined penalties are more difficult to construct than the coarse penalties. Both γ^R and γ^C depend on the target of estimation, $\tilde{\lambda}$. These penalties are still infeasible but can be estimated once we have a consistent estimator $\hat{\lambda}$ such that $||\hat{\lambda} - \tilde{\lambda}||_1 \to_p 0$. We discuss how to construct feasible penalties in Section 1.5.3.

Theorems 1 and 2 require that Assumption C.7 be satisfied uniformly for the set of vectors $e_1, ..., e_d$. The statement below shows how this can be achieved by adjusting the common penalty term t.

Lemma 6 (Uniform Dominance of Penalty) *Suppose* $\gamma_{j,l}$ *is the penalty term for* $v = e_l$ *and moment* j, $1 \leqslant l \leqslant d$ *and* $1 \leqslant j \leqslant m$. *Suppose Assumption C.7 holds*

for each $l \in \{1, 2, ..., d\}$ *with* $t = t_0$ *and the set of penalties* $\{\gamma_{j,l}\}_{j=1}^m$. *Then by setting* $t = \sqrt{n\Phi^{-1}(1 - \frac{4md}{\alpha_n})}$, *Assumption C.7 holds uniformly for* $v = e_1, ..., e_d$.

Assumption C.8 is a general assumption which can only be examined for specific choices of penalties. In what follows, we consider the necessary conditions for γ^R and γ^C to satisfy C.8.

Assumption C.11 (Bounds on Empirical Higher Order Moments) *There exist absolute positive constants* K_g^u, K_G^l *and* K_G^u *such that:*

(1)

$$\max_{1 \leqslant j \leqslant m} \mathbb{E}_n[g_j(Z_i, \beta_0)^4] \leqslant K_g^u;$$

(2)

$$K_G^l \leqslant \min_{1 \leqslant j \leqslant m} \left\{ \mathbb{E}_n[(\frac{\partial g_j}{\partial \beta}(Z_i, \beta_0)v)^2] - [\mathbb{E}_n \frac{\partial g_j}{\partial \beta}(Z_i, \beta_0)v]^2 \right\}$$
$$\leqslant \max_{1 \leqslant j \leqslant m} \left\{ \mathbb{E}_n[(\frac{\partial g_j}{\partial \beta}(Z_i, \beta_0)v)^2] - [\mathbb{E}_n \frac{\partial g_j}{\partial \beta}(Z_i, \beta_0)v]^2 \right\} \leqslant K_G^u.$$

In BCCH (2012), the authors impose a condition that is stronger than (1), in particular, they assume that $\mathbb{E}_n[g_j(Z_i, \beta_0)^8]$ is bounded, which is not required here.

Lemma 7 *If Assumptions C.5 and C.11 hold, then Assumption C.8 holds for* γ^R *and* γ^C.

1.5.3 Feasible Penalties

We are tempted to use the suggested γ^R or γ^C penalties, however, they are not feasible. There are two obstacles to their use. The first is that we need to know the true theoretical parameter β_0, the other is that $\tilde{\lambda}$ is unknown. We can and will substitute the unknown β_0 with the preliminary estimator $\tilde{\beta}$. Finding a substitute for $\tilde{\lambda}$ is a more delicate task.

Assume we have a preliminary guess of $\tilde{\lambda}$, which we denote $\breve{\lambda}$. If $||\breve{\lambda} - \tilde{\lambda}||_1 \to 0$, we can construct a feasible version of penalty terms γ_j^R and γ_j^C.

Definition 1.5.1 (Empirical Penalty) *(1) Refined empirical penalty $\hat{\gamma}_j^R$, $1 \leqslant j \leqslant m$, is defined as:*

$$\hat{\gamma}_j^R := \left\{ \mathbb{E}_n(\sum_{1 \leqslant k \leqslant m} g_k(Z_i, \tilde{\beta}) g_j(Z_i, \tilde{\beta}) \breve{\lambda}_k)^2 - [\mathbb{E}_n \sum_{1 \leqslant k \leqslant m} g_k(Z_i, \tilde{\beta}) g_j(Z_i, \tilde{\beta}) \breve{\lambda}_k)]^2 \right\}^{\frac{1}{2}} \tag{1.5.5}$$

$$+ \left\{ \mathbb{E}_n (G(Z_i, \tilde{\beta}) v)_j^2 - [\mathbb{E}_n (G(Z_i, \tilde{\beta}) v)_j]^2 \right\}^{\frac{1}{2}}.$$

(2) Coarse empirical penalty $\hat{\gamma}_j^C$, $1 \leqslant j \leqslant m$, is defined as:

$$\hat{\gamma}_j^C = ||\breve{\lambda}||_1 \left\{ \max_{1 \leqslant k \leqslant m} \left(\mathbb{E}_n g_k(Z_i, \tilde{\beta})^4 - [\mathbb{E}_n \hat{g}_k(Z_i, \tilde{\beta})^2]^2 \right) \right\}^{\frac{1}{2}} \tag{1.5.6}$$

$$+ \left\{ \mathbb{E}_n |(G(Z_i, \tilde{\beta}) v)_j|^2 - [\mathbb{E}_n (G(Z_i, \tilde{\beta}) v)_j]^2 \right\}^{\frac{1}{2}}.$$

The empirical penalties $\hat{\gamma}^R$ and $\hat{\gamma}^C$ should be close and work similarly to the asymptotic penalties γ^R and γ^C. The features needed for the empirical penalties to perform well are summarized in the Lemma below.

Lemma 8 (Consistency of empirical penalties) *Suppose Assumptions C.1, C.2, C.5 hold. Assume Assumption C.8 hold for theoretical penalty γ^R (or γ^C) as defined in equation (1.5.5) or (1.5.6) correspondingly. If a preliminary guess is such that $||\breve{\lambda} - \tilde{\lambda}||_1 \to_p 0$ and $(K_{M,n} \vee K_{G,n}) n^{-\rho} \to_p 0$, then the empirical penalty terms $\hat{\gamma}_j^R$ (or $\hat{\gamma}_j^C$), $1 \leqslant j \leqslant m$, satisfy the following condition. There are two non-random sequences u_n and l_n converging to one such that with probability increasing to one*

$$l_n \leqslant |\frac{\hat{\gamma}_j}{\gamma_j}| \leqslant u_n \quad \textit{for all} \quad 1 \leqslant j \leqslant m.$$

A result similar to that of Theorems 1 and 2 can be derived for empirical penalties $\hat{\gamma}^R$ (or $\hat{\gamma}^C$).

Theorem 3 *Suppose Assumptions C.1-C.3, C.5 and C.9-C.11 hold for all* $v = e_1, ..., e_d$ *with* $t = (1+\epsilon)\sqrt{n\Phi^{-1}(1-\frac{2md}{\alpha_n})}$, $\alpha_n \to 0$ *and* $m\alpha_n \to \infty$. *Let the penalty terms be* $\hat{\gamma}_j^R$ *(or* $\hat{\gamma}_j^C$*), which are based on preliminary guess* $\breve{\lambda}$. *Suppose the following growing conditions hold with probability increasing to one:*

(1) $\frac{\log(m)^3}{n} \to 0$, $\frac{s_n^2 \log(m)(K_{B,n} \vee K_{M,n})^2}{n} \to 0$.

(2) $\frac{s_n \log(n) K_{M,n}^2}{n^{2\rho}} \to 0$, $\frac{(K_{G,n} \vee K_{M,n})^2}{n^{2\rho-1}\log(m)} \to 0$.

(3) $||\breve{\lambda} - \tilde{\lambda}||_1 \to 0$.

Then, there exists a sequence $\epsilon_n \to 0$ *such that with probability at least* $1 - \alpha - \epsilon_n$,

$$||\hat{\beta}_L - \beta_0||_2 = O\left(\frac{1}{\sqrt{n}} \vee \frac{s_n \log(m)}{n}\right), \tag{1.5.7}$$

$$||\hat{\beta}_{PL} - \beta_0||_2 = O\left(\frac{1}{\sqrt{n}} \vee \frac{s_n \log(m)}{n}\right). \tag{1.5.8}$$

Furthermore, if $s_n^2 \log(m)^2 = o(n)$, *the estimator* $\hat{\beta}_L$ *and* $\hat{\beta}_{PL}$ *are asymptotically normal:*

$$\sqrt{n}(\hat{\beta}_L - \beta_0) \to_d N(0, (G'\Omega_0^{-1}G)^{-1}). \tag{1.5.9}$$

$$\sqrt{n}(\hat{\beta}_{PL} - \beta_0) \to_d N(0, (G'\Omega_0^{-1}G)^{-1}). \tag{1.5.10}$$

To come up with an accurate guess $\breve{\lambda}$ of $\tilde{\lambda}$ one could employ an adaptive procedure that uses estimators of $\tilde{\lambda}$ obtained as a result of LASSO selection. Given any $\breve{\lambda}$, define a function Π that maps $\breve{\lambda} \in \mathbb{R}^m$ to the solution of problem $\mathcal{P}$ with empirical penalties $\hat{\gamma}_j^R$ (or $\hat{\gamma}_j^C$) which have been constructed based on $\breve{\lambda}$. For the true unknown value $\tilde{\lambda}$, we know that by Theorem 1, under certain regularity conditions, $||\Pi(\tilde{\lambda}) - \tilde{\lambda}||_1 = O_p\left(\sqrt{\frac{s_n^2 \log(m)}{n}}\right)$. So $\tilde{\lambda}$ is "nearly" a fixed point of Π when $\frac{s_n^2 \log(m)}{n} \to 0$. *Vice versa*, when Π only has one unique fixed point, this fixed point should lie very close to $\tilde{\lambda}$. Thus, to implement moment selection procedure $\mathcal{P}$, we can iteratively update the empirical penalty terms $\hat{\gamma}_j^R$ (or $\hat{\gamma}_j^C$), $1 \leqslant j \leqslant m$. Consider a sequence of $m \times 1$ vectors $\lambda^{(p)}$, and $p = 0, 1, ...$ $\lambda^{(p)}$ will be the solution of $\mathcal{P}$ with penalties computed based on $\lambda^{(0)}$, $\lambda^{(1)}$,...,$\lambda^{(p-1)}$. First, we

need to discuss an algorithm that converges globally for coarse penalty $\hat{\gamma}^C$.

Algorithm 1 (Binomial Search) *For coarse penalty $\hat{\gamma}^C$, let $\lambda_j^{(0)} = 0$ and $\lambda_j^{(1)} = \xi$ for all $j = 1, 2, ..., m$, where ξ is a positive number. Denote $x_0 = ||\lambda^{(0)}||_1$ and $x_1 = ||\lambda^{(1)}||_1$. Let η be a small number representing the precision level of our algorithm.*[4]

Notice that $\hat{\gamma}_j^C$ only depends on $||\breve{\lambda}||_1$. We can consider a mapping Π_1 which maps a non-negative real number w to an $m \times 1$ vector γ^C by plugging $||\lambda||_1 = w$ into equation (1.5.6).

Start of Algorithm 1:

While $(|x_0 - x_1| > \eta)$

Let $x_2 = \frac{x_1 + x_0}{2}$;

If $||\Pi_1(x_2)||_1 > x_2$, *then* $x_0 = x_2$;

else $x_1 = x_2$.

end

Termination of Algorithm 1.

Lemma 9 (Convergence of Algorithm 1) *For penalty terms $\hat{\gamma}^C$, if the value ξ is large enough, x_1 and x_2 in Algorithm 1 converge to a non-negative fixed point $x^C \in \mathbb{R}$. The fixed point x^C is unique.*[5]

Lemma 10 (Property of Fix Point using $\hat{\gamma}^C$) *Suppose we apply coarse empirical penalties and Algorithm 1 to perform LASSO in problem $\mathcal{P}$. Define λ^C as the solution of problem $\mathcal{P}$, given the penalty set as $\Pi_1(x^C)$. Suppose Assumptions C.1-C.3, C.5 and C.9-C.11 hold for all $v = e_1, ..., e_d$ with $t = (1+\epsilon)\sqrt{n\Phi^{-1}(1 - \frac{2md}{\alpha_n})}$, $\alpha_n \to 0$ and $m\alpha_n \to \infty$. Suppose the following growth conditions hold with probability increasing to one:*

[4] We recommend the use of $\lambda^{(0)} = 0$ and $\eta = 0.0001$.

[5] If we use a shooting algorithm to compute the minimization problem $\mathcal{P}$, the operational time of Algorithm 1 is $O(|\log(\frac{\xi}{\eta}) m \log(m)|)$.

(1) $\frac{\log(m)^3}{n} \to 0$, $\frac{s_n^2 \log(m)(K_{B,n} \vee K_{M,n})^2}{n} \to 0$.

(2) $\frac{s_n \log(n) K_{M,n}^2}{n^{2\rho}} \to 0$, $\frac{(K_{G,n} \vee K_{M,n})^2}{n^{2\rho-1}\log(m)} \to 0$.

Then the LASSO estimator $\hat{\beta}_L$ and the post-LASSO estimator $\hat{\beta}_{PL}$ based on λ^C satisfy statements (1.5.7) and (1.5.8). Furthermore, if $s_n^2 \log(m)^2 = o(\sqrt{n})$, $\hat{\beta}_L$ and $\hat{\beta}_{PL}$ satisfy (1.5.9) and (1.5.10).

For penalty $\hat{\gamma}^R$, we propose the application of Algorithm 2 to iteratively estimate $\hat{\gamma}^R$ and $\tilde{\lambda}$. This algorithm is usually required to perform with a latency $\geqslant 2$, since a naive iterative algorithm often diverges in practice. This algorithm can serve for a general adaptive penalization procedure when penalties are computed based on the target of estimation. The fixed point of Algorithm 2, λ^R, has superior finite sample performance compared to λ^C, since penalties γ^R are sharper than penalties γ^C. We illustrate this point with Monte-Carlo examples in Section 1.6.

Algorithm 2 (Adaptive Algorithm with Latency) *Let $w \geqslant 1$ be the length of the latency. Let $B \geqslant w$ be the length of the incubational period. Let $\lambda^{(0)}$ be the initial value of λ. Let η be the tolerance level.*

Start of Algorithm 2:

While$(p \geqslant 0)$

If $(p < B)$ *(incubational period)*

(1a) Compute penalty $\hat{\gamma}^R$ using $\lambda^{(0)}$.

(2a) Compute the optimization problem $\mathcal{P}$ with the penalty term γ.

(3a) $\lambda^{(p)}$ is set to be the optimizer of the problem $\mathcal{P}$ using penalty γ obtained in (1a).

(4a) $p = p + 1$.

else (converging period)

(1b) Update penalty term $\hat{\gamma}^R$ using $\lambda = \bar{\lambda}^{(p)}$, where $\bar{\lambda}^{(p)} := \frac{\sum_{q=p-w}^{p-1} \lambda^{(q)}}{w}$.

(2b) Compute the optimization problem $\mathcal{P}$ with the penalty term $\hat{\gamma}^R$ and obtain optimizer $\lambda^{(p)}$.

(3b) If $\sum_{q=1}^{w} ||\lambda^{(p-q+1)} - \frac{\sum_{q'=p-w}^{p} \lambda^{(q')}}{w+1}||_1 < \eta$: *Terminate.*

(4b) $p = p + 1$.

end

end

Termination of Algorithm 2.

Lemma 11 (Property of Fixed Point using λ^R) *Suppose we make use of refined empirical penalties and Algorithm 2 to perform LASSO in problem $\mathcal{P}$. Suppose the initial value $\lambda^{(0)}$ satisfies $||\lambda^{(0)} - \tilde{\lambda}||_1 \to 0$. Assume that the sequence $\lambda^{(0)}, \lambda^{(1)}, ...$ converges to a fixed point λ^R.*[6] *Suppose Assumptions C.1-C.3, C.5 and C.9-C.11 hold for all $v = e_1, ..., e_d$ with $t = (1+\epsilon)\sqrt{n\Phi^{-1}(1 - \frac{2md}{\alpha_n})}$, $\alpha_n \to 0$ and $m\alpha_n \to \infty$. Further, suppose the following growth conditions hold with probability increasing to one:*

(1) $\frac{\log(m)^3}{n} \to 0$, $\frac{s_n^2 \log(m)(K_{B,n} \vee K_{M,n})^2}{n} \to 0$.

(2) $\frac{s_n \log(n) K_{M,n}^2}{n^{2\rho}} \to 0$, $\frac{(K_{G,n} \vee K_{M,n})^2}{n^{2\rho-1}\log(m)} \to 0$.

Then the LASSO $\hat{\beta}_L$ and post-LASSO $\hat{\beta}_{PL}$ estimators based on λ^R satisfy (1.5.7) and (1.5.8). Furthermore, if $s_n^2 \log(m)^2 = o(\sqrt{n})$, $\hat{\beta}_L$ and $\hat{\beta}_{PL}$ satisfy (1.5.9) and (1.5.10).

1.6 Simulation

Han and Phillips (2006) introduces several interesting economic examples with many moment conditions that are non-linear in the parameter of interest. Similar to Shi (2013), we consider a Monte-Carlo experiment based on time-varying individual heterogeneity models (example 17 as described in Han and Phillips (2006)).

[6] λ^C can serve as an initial value for Algorithm 2 when we use the refined penalties to perform LASSO. Although it is still unknown whether or not Algorithm 1 converges in all cases, in our simulations the convergence criterion of Algorithm 2 is always satisfied within less than 100 iterations when k is set to 3.

Assume that we observe i.i.d. data (x_i, y_i), $i = 1, 2, ..., n$, where x_i is an $(m+1) \times d_x$ matrix and y_i is an $m \times 1$ vector. Suppose that for any $1 \leqslant j \leqslant m$ and $1 \leqslant i \leqslant n$, we model $y_{i,j}$ as:

$$y_{i,j} = f_j(\beta_1)\alpha_i + x_{i,j}\beta_2 + \epsilon_{i,j},$$

where $f_j(\cdot)$ are a known function that depends on j, α_i are some unknown individual heterogeneity, and $\epsilon_{i,j}$ are random i.i.d. errors with mean 0. The key parameter of interest, β_1, is a scalar while β_2 is a $d \times 1$ vector $\in \mathbb{R}^d$. The variation of $f_j(\beta_1)$ captures how does the effect of individual heterogeneity change across period j, $0 \leqslant j \leqslant m$. Thus, in this model, it is important to estimate β_1 well. For any $1 \leqslant j \leqslant m$, we can consider the following moment conditions based on first difference strategy:

$$\mathbb{E}[y_{i,j}f_{j-1}(\beta_1) - y_{i,j-1}f_j(\beta_1) - (f_{j-1}(\beta_1)x_{i,j} - f_j(\beta_1)x_{i,j-1})\beta_2] = \mathbb{E}[f_{j-1}(\beta_1)\epsilon_{i,j} - f_j(\beta_1)\epsilon_{i,j-1}] = 0. \tag{1.6.1}$$

Equation (1.6.1) holds for every $j = 1, 2, ..., m-1$, that is to say we can form a moment condition for each $1 \leqslant j \leqslant m-1$. In addition, we consider to add a moment condition based on long-difference:

$$\mathbb{E}[y_{i,m}f_0(\beta_1) - y_{i,0}f_m(\beta_1) - (f_0(\beta_1)x_{i,j} - f_m(\beta_1)x_{i,0})\beta_2] = \mathbb{E}[f_0(\beta_1)\epsilon_{i,m} - f_m(\beta_1)\epsilon_{i,0}] = 0. \tag{1.6.2}$$

Thus, there are m moment conditions in total. Let $\beta = (\beta_1, \beta_2')$ be a vector in $\mathbb{R}^d$ and β_0 be the true parameter of β. The moment conditions implied by equation (1.6.1) can be written as follows:

$$g_j(y_i, \beta) := y_{i,j}f_{j-1}(\beta_1) - y_{i,j-1}f_j(\beta_1) - (f_{j-1}(\beta_1)x_{i,j} - f_j(\beta_1)x_{i,j-1})\beta_2. \tag{1.6.3}$$

And we have:

$$\frac{\partial g_j}{\partial \beta}(y_i, \beta) = (\frac{\partial f_{j-1}}{\partial \beta_1}(\beta_1) \cdot (y_{i,j} - x_{i,j}\beta_2) - \frac{\partial f_j}{\partial \beta_1}(\beta_1) \cdot (y_{i,j-1} - x_{i,j-1}\beta_2), \tag{1.6.4}$$

$$-(f_{j-1}(\beta_1)x_{i,j} - f_j(\beta_1)x_{i,j-1})).$$

Similarly, we can write down empirical moment conditions for equation (1.6.2). Given

a consistent estimator $\tilde{\beta}$ of β_0, it is easy to build the basic constructing blocks in moment selection procedure $\mathcal{P}$:

$$\hat{\Omega} = \mathbb{E}_n[g_i(y_i, \tilde{\beta})g_i(y_i, \tilde{\beta})'],$$

$$\hat{G}(\tilde{\beta})v = \mathbb{E}_n[\frac{\partial g_j}{\partial \beta}(y_i, \tilde{\beta})v],$$

where $v = e_l$, $l = 1, 2, ..., d$.[7]

Shi (2013) considers a design that the covariates x_{ij} only consists of a constant. Also in Shi (2013), the "strength" of the moment conditions, i.e., $||\mathbb{E}[\frac{\partial g_j}{\partial \beta}(y_i, \beta_0)]||_2$ decays exponentially (after being sorted in a decreasing order) due to the formulation of $f_j(\cdot)$, $1 \leqslant j \leqslant m$. Similar to Shi (2013), we perform our procedure on a design with a covariate being the constant term. However, we only place a restriction that the quantities $||\mathbb{E}[\frac{\partial g_j}{\partial \beta}(y_i, \beta_0)]||_2$ decay in polynomial speed (after being sorted in decreasing order.) Unlike example 2 in Shi (2013), in our design the optimal GMM estimator is severely biased, while the LASSO and post-LASSO estimators are much less biased.

Example 6 (Approximate Sparse Design 1) *Assume that for* $0 \leqslant j \leqslant m-1$, $f_j(\beta_1) = \frac{1}{1+\beta_1 j^a}$, *where* $a = 1$. *So* $f_j(\beta_1)$ *decays with polynomial speed. For* $j = m$, *let* $f_m(\beta_1) = \beta_1$ *such that* m^{th} *moment is informative about the true parameter* β_0. *When* m *is large, the last moment is computationally difficult to be detected and thereby being used in the estimation. Let* $\alpha_i \sim N(1,1)$ *be i.i.d across individual* i. *In addition, let the constant be the only covariate in (1.6.1) and the true parameter* $\beta_0 := (\beta_{10}, \beta_{20}) = (0.6, -2)$. *Assume that the domain of the parameter* β *is* $[0.1, 1] \times [-1, 5]$. *Let* $\tilde{\beta}$ *be the efficient GMM estimator estimated from the first five moment conditions.*

Therefore, for any $1 \leqslant j \leqslant m-1$, $g_j(z_i, \beta_0) = f_{j-1}(\beta_{10})\epsilon_{i,j} - f_j(\beta_{10})\epsilon_i$. *So for any* $1 \leqslant j \leqslant m-1$, $Var(g_j(z_i, \beta_0)) = \left(\frac{1}{1+\beta_{10}j^a}\right)^2 + \left(\frac{1}{1+\beta_{10}(j-1)^a}\right)^2 \geqslant \frac{1}{\beta_{10}^2 j^{2a}}$. *Hence, we divide the moment condition* g_j *by* j^{-a} *in order to normalize these moment conditions, i.e., for* $j = 0, 1, ..., m-1$,

$$\tilde{g}_j(z_i, \beta) = g_j(z_i, \beta) \cdot j^a.$$

We don't normalize the m^{th} *moment condition because it has variance bounded away from* 0 *and from above. So* $\tilde{g}_m(z_i, \beta) = g_m(z_i, \beta)$.

[7] In practice, certain normalization may be needed. We discuss this in the specific examples below.

The partial derivative of $\tilde{g}(z_i, \beta)$ *can be written as follows:* $\frac{\partial \tilde{g}_j}{\partial \beta}(z_i, \beta)$

$= j^a \cdot \Big(\frac{\partial f_{j-1}}{\partial \beta_1}(\beta_1) \cdot (y_{i,j} - x_{i,j}\beta_2) - \frac{\partial f_j}{\partial \beta_1}(\beta_1) \cdot (y_{i,j-1} - x_{i,j-1}\beta_2), -(f_{j-1}(\beta_1)x_{i,j} - f_j(\beta_1)x_{i,j-1}) \Big)$.

And the expected gradient of the j^{th} *moment is:*

$$\tilde{G}_{0j}(\beta_0) = \mathbb{E}[\frac{\partial \tilde{g}_j}{\partial \beta}(z_i, \beta_0)] = j^a \cdot \Big(\frac{\partial f_{j-1}}{\partial \beta}(\beta_{10}) f_j(\beta_{10}) - \frac{\partial f_j}{\partial \beta}(\beta_{10}) f_{j-1}(\beta_{10}), f_j(\beta_1) - f_{j-1}(\beta_1) \Big),$$

for $0 \leqslant j \leqslant m-1$. *It is easy to see that when* $a = 1$,

(1) $j^a \{ \frac{\partial f_{j-1}}{\partial \beta}(\beta_{10}) f_j(\beta_{10}) - \frac{\partial f_j}{\partial \beta}(\beta_{10}) f_{j-1}(\beta_{10}) \} = \frac{j^a \beta_{10}}{(1+j^a\beta_{10})^2(1+(j-1)^a\beta_{10})^2}$;

(2) $j^a(f_j(\beta_{10}) - f_{j-1}(\beta_{10})) = -\frac{j^a \beta_{10}}{(1+j^a\beta_{10})(1+(j-1)^a\beta_{10})}$.

So $||\tilde{G}_{0j}||_2$ *is decaying with polynomial speed* $O(j^{-a})$. *We perform the selection procedure* $\mathcal{P}$ *with* $\tilde{g}$. *The initial value of GMM estimation procedures is all set at* $(0.5, -1.5)$. *We demonstrate the performance of our selection procedure about the key parameter* β_1 *in Tables 1.1, 1.2 and Figures 1.1-1.3. In Table 1.1, it is clear that* $\hat{\beta}_L$ *and* $\hat{\beta}_{PL}$ *are more efficient compared to efficient GMM and equally weighted GMM (EW-GMM later). In addition, the bias of LASSO and the post-LASSO estimators are much smaller. CUE has small bias when* $n = 400$ *but it is more dispersed because of its heavy tails, as such phenomenon is discussed in Hausman et al. (2007). In Table 1.2, we can see that the average number of moments selected when* $n = 400$ *is larger than that when* $n = 200$. *When sample size is smaller, the penalties are larger and thus less moments are selected via LASSO. In our example, Algorithm 2 runs iteratively for less than 15 times on average before it hits the stopping criteria, as described in the second row of Table 1.2*[8]. *Figure 1 illustrates the frequencies of moment conditions selected by the* $\hat{T}$ *as described in Lemma 2. As we can observe in Figure 1.1, the moments picked by the selector* $\hat{T}$ *includes the first a few ones and the last one, which is expected to be a strongly informative moment condition and can be hardly picked by traditional methods such as AIC and BIC procedures proposed in Andrews (1999). In addition, our procedure does not rely on perfect selection: the* 4^{th} *moment condition is only used in* 65% *of the time when* $n = 200$ *and* 25% *of the time when* $n = 400$. *Figure 2-3 confirm asymptotic normality*

[8]For GMM, when $m = 240 > n = 200$, the traditional GMM estimator is not well defined. Instead of using $\hat{\Omega}^{-1}$, I use $(\hat{\Omega} + \frac{I_m}{\sqrt{n}})^{-1}$.

m=240	n=400				n=200			
	Bias	$\sqrt{n}\hat{SE}$	MSE	Rej. Rate	Bias	$\sqrt{n}\hat{SE}$	MSE	Rej. Rate
LASSO	0.0039	0.904	0.0023	0.042	0.0078	0.923	0.0043	0.062
post-LASSO	0.0065	0.875	0.0020	0.048	0.0045	0.922	0.0043	0.068
GMM	0.149	1.164	0.0255	0.770	0.175	1.332	0.0396	0.538
EW-GMM	0.350	6.051	0.2145	1.000	0.402	1.265	0.170	1.000
CUE	0.0067	2.206	0.110	0.830	NA	NA	NA	NA
Eff. Bound	NA	0.756	NA	NA	NA	0.756	NA	NA

Table 1.1: Comparison of $\hat{\beta}_L$ and $\hat{\beta}_{PL}$ on the key parameter β_1.

m=240	n=400		n=200	
	LASSO	CUE	LASSO	CUE
Average number of moments selected	5.76	NA	4.79	NA
Average number of iteration	14.2	NA	12.3	NA
Average running time per instance	15.56sec	162.12 sec	13.52sec	116.75sec

Table 1.2: More details on performance of LASSO and post-LASSO.

of $\hat{\beta}_L$ and $\hat{\beta}_{PL}$.

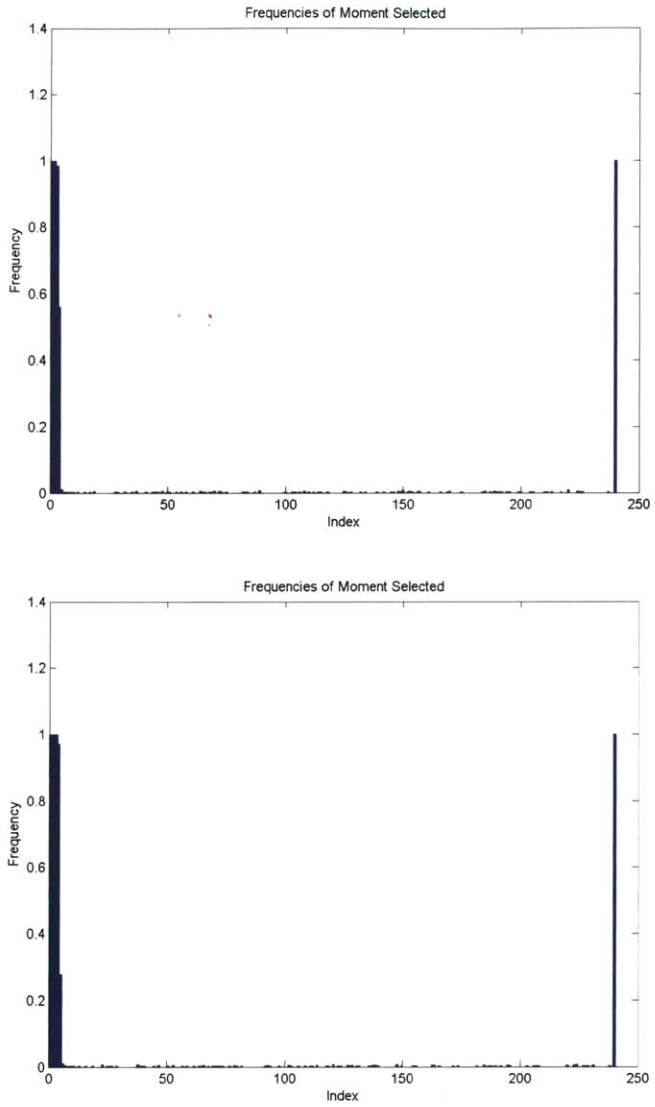

Figure 1-1: Frequencies of Moments Selected: Top: $n = 200$; Bottom: $n = 400$

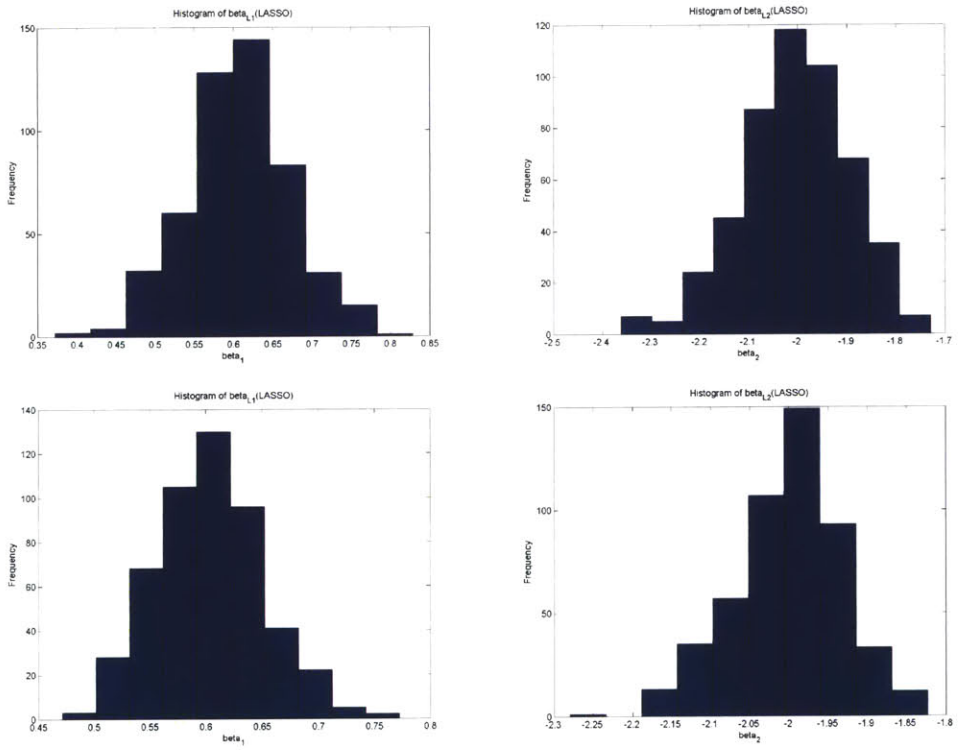

Figure 1-2: Distribution of $\hat{\beta}_{1L}$ and $\hat{\beta}_{2L}$. Top: $n = 200$; Bottom: $n = 400$.

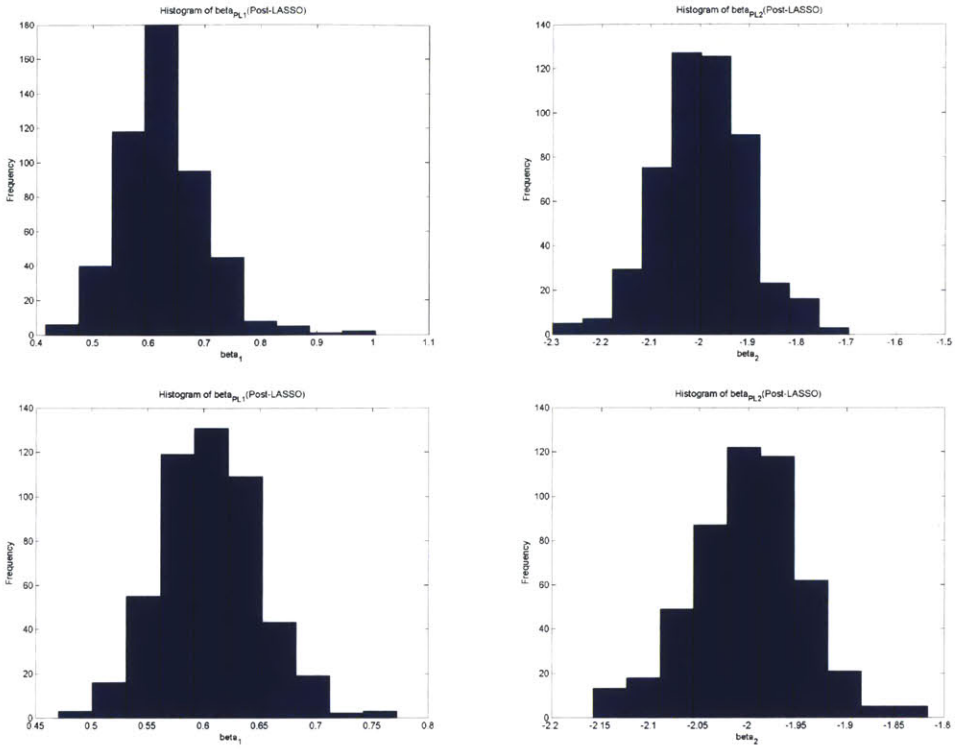

Figure 1-3: Distribution of $\hat{\beta}_{1PL}$ and $\hat{\beta}_{2PL}$. Top: $n = 200$; Bottom: $n = 400$.

The next Monte-Carlo experiment is based on example 3 of Hausman, Lewis, Menzel and Newey (2007). This example considers a Generalized-IV model when the "second stage regression" contains non-linear functions of the parameter β.

Example 7 (Approximate Sparse Design 2) *Consider the following setting:*

$$y = \exp(x\beta_0) + \epsilon,$$

$$x = z\Pi + v.$$

We slightly modify the assumptions in Hausman, Lewis, Menzel and Newey (2007): First, x is two dimensional, not one; Second, the number of instruments, m, is larger than the sample size, n. The true beta $\beta_0 = (0, 0)$. Let $m = 500$ and $n = 200$. $v \sim N(0, 1)$, $z_1, ..., z_m \sim N(0, 1)$ and they are independent, $\Pi(1, j) = 1/j^a + 1/(m + 1 - j)^a$,

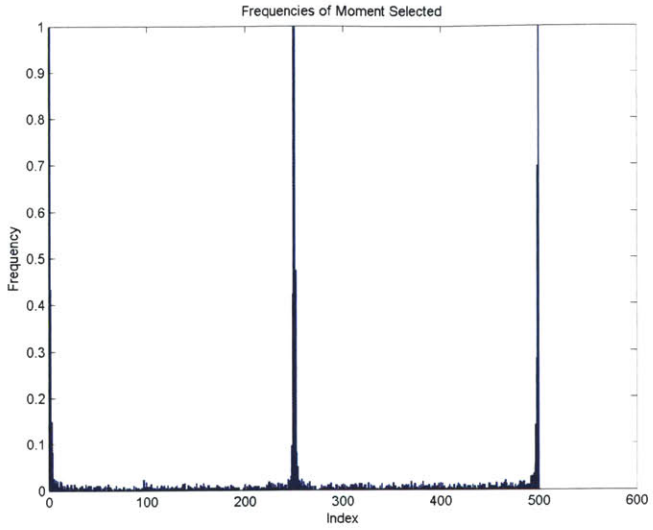

Figure 1-4: Frequencies of Moments Selected: Approximate Sparse Design 2

$\Pi(2,j) = 1/(m/2+1/2-j)^a$, *and* $\epsilon = \rho w_1 + \sqrt{1-\rho^2}\{\phi w_2 \cdot (\sum_{j=1}^m z_j/j) + \sqrt{1-\phi^2} w_3\}$, *where* $a = 1$ *and* w_1, w_2, w_3 *are independent standard normal random variables. The preliminary estimate is set as the equally weighted GMM (EWGMM) estimator.*

In this setting, traditional GMM and CUE estimators do not exist. We compare our estimator with the performance of EWGMM estimator in Table 1.3. In Table 1.3, though EWGMM estimator seems to be quite close to the true parameter, but its bias to too large and lead to incorrect inference. Such bias will lead to even worse rejection rate when n increases. Also in Table 1.3, our LASSO and post-LASSO estimators are nearly efficient compared to the efficiency bound $(G_0(\beta_0)\Omega_0^{-1}G_0(\beta_0))^{-1}$, *although on average 10.3 out of 500 moments are selected. The empirical variance of EW. The asymptotic test of LASSO and post-LASSO estimators have the size which are slightly larger than* 0.05. *This is perhaps due to the randomness of the moment selector* $\hat{T}$. *In Table 1.1, we also see the similar phenomenon in LASSO and post-LASSO estimators when* $n = 200$. *Again, similar to the previous Monte-Carlo example, post-LASSO is slightly better than LASSO in MSE. Figure 4 presents the frequencies of moments which are picked by the LASSO selector. According to our design, we should expect three groups clustered around* 1, 250, *and* 500. *We can observe exactly the same phenomenon in Figure 1.4.*

	Bias	$\sqrt{n}\hat{Var}$
LASSO	$(0.0123, -0.0005)$	$\begin{pmatrix} 0.582 & 0.011 \\ 0.011 & 0.147 \end{pmatrix}$
post-LASSO	$(0.0166, 0.0034)$	$\begin{pmatrix} 0.524 & 0.006 \\ 0.006 & 0.145 \end{pmatrix}$
EW-GMM	$(0.0559, 0.0174)$	$\begin{pmatrix} 0.392 & -0.016 \\ -0.016 & 0.137 \end{pmatrix}$
Eff. Bound	NA	$\begin{pmatrix} 0.477 & -0.060 \\ -0.060 & 0.123 \end{pmatrix}$

	MSE	Rej. Rate
LASSO	$(3.11, 0.75) \times 10^{-3}$	0.060
post-LASSO	$(2.92, 0.73) \times 10^{-3}$	0.060
EW-GMM	$(5.09, 0.98) \times 10^{-3}$	0.248
Eff. Bound	NA	NA

Average # of Moments Chosen	10.3

Table 1.3: $m = 500, n = 200$, Comparison of $\hat{\beta}_L$ and $\hat{\beta}_{PL}$ with other GMM estimators.

1.7 Conclusion

This paper applies the LASSO method to solve the many moments problem. Instead of implementing traditional optimal GMM with the full set of moments, we consider selecting the informative moments before conducting the traditional optimal GMM procedure. Since the optimal GMM estimator can be obtained via the optimal combination matrix $G_0(\beta_0)'\Omega_0^{-1}$, we formulate a quadrative objective function with LASSO type penalties to estimate the rows in the optimal combination matrix (d rows in total). This method has several advantages compare to the traditional optimal GMM or GEL when the number of moments is comparable to the sample size or even much larger than the sample size. When approximate sparisty holds, first of all, our method can substantially reduce any second order bias simply because most of the informative moments are dropped by the selection procedure; second, our method is computationally tractable compared to other moment selection procedures such as those proposed in Donald, Imbens and Newey (2008).

Theoretically, we establish the asymptotic bounds of the LASSO estimator of the optimal combination matrix under L_1 distance and semi-norm $||\cdot||_2$. Based on these bounds, we are able to prove consistency and establish bounds for the LASSO based GMM estimator $\hat{\beta}_L$ and post-LASSO based GMM-estimator $\hat{\beta}_{PL}$. Furthermore, when the number of truly informative moments, s_n (as defined in the approximate sparsity assumption), grows with speed $\frac{s_n^2 \log(m)^2}{n} \to 0$, we can prove that together with a set of high-level conditions, both $\hat{\beta}_L$ and $\hat{\beta}_{PL}$ are asymptotically normal and nearly efficient. These high-level assumptions are common in the LASSO literature. We establish primitive conditions for the high-level assumptions such that the validity of these assumptions mainly relies on a set of growth conditions for the parameters $K_{G,n}, K_{M,n}, K_{B,n}$ (which characterizes the behavior of the tail of the residuals and the smoothness of the moment conditions) and ρ (which characterizes the accuracy of the preliminary estimator). All these results are novel in dealing with non-linearity when the LASSO method is applied.

In addition to these theoretical results, we propose a set of feasible and valid penalties to implement the LASSO procedure. Due to the complexity of our problem, our penalty terms depend on the target of estimation, which is one of the main challenges we en-

counter. Such a challenge does not arise when traditional LASSO is applied to OLS and 2SLS. We propose adaptive algorithms to solve this difficulty that are computationally tractable. Our algorithm 1 converges globally, which guarantees the performance of the algorithm with fast speed. Our algorithm 2 is more general and works well in Monte-Carlo experiments, though the theoretical convergence speed is not yet known. We prove that the convergence points in our algorithms satisfy the same properties as if we were using the penalties constructed from the true parameter. The excellent performance of these adaptive algorithms is demonstrated in Monte-Carlo experiments.

Bibliography

[1] Andrews, D. W. (1999). Consistent moment selection procedures for generalized method of moments estimation. Econometrica, 67(3), 543-563.

[2] Bekker, P.A., Alternative Approximations to the Distributions of Instrumental Variables Estimators, Econometrica 63, 657-681, 1994.

[3] Belloni, A., and Chernozhukov, V. (2010). Post-l_1-penalized estimators in high-dimensional linear regression models (No. CWP13/10). cemmap working paper.

[4] (a) Belloni, A., and Chernozhukov, V. (2011). High dimensional sparse econometric models: An introduction (pp. 121-156). Springer Berlin Heidelberg.

(b) Belloni, A., and Chernozhukov, V. (2011). l_1-penalized quantile regression in high-dimensional sparse models. The Annals of Statistics, 39(1), 82-130.

[5] Belloni, A., Chernozhukov, V., and Hansen, C. (2011). Inference for high-dimensional sparse econometric models. arXiv preprint arXiv:1201.0220.

[6] Belloni, A., Chen, D., Chernozhukov, V., and Hansen, C. (2012). Sparse models and methods for optimal instruments with an application to eminent domain. Econometrica, 80(6), 2369-2429.

[7] Bickel, P. J., Ritov, Y. A., and Tsybakov, A. B. (2009). Simultaneous analysis of Lasso and Dantzig selector. The Annals of Statistics, 1705-1732.

[8] Bunea, F., Tsybakov, A., and Wegkamp, M. (2007). Sparsity oracle inequalities for the Lasso. Electronic Journal of Statistics, 1, 169-194.

[9] Candes, E., and Tao, T. (2007). The Dantzig selector: Statistical estimation when p is much larger than n. The Annals of Statistics, 2313-2351.

[10] Puna, V. H., Lai, T. L., and Shao, Q. M. (2008). Self-normalized processes: Limit theory and Statistical Applications. Springer.

[11] Donald, S. G., Imbens, G. W., and Newey, W. K. (2003). Empirical likelihood estimation and consistent tests with conditional moment restrictions. Journal of Econometrics, 117(1), 55-93.

[12] Donald, S. G., Imbens, G., and Newey, W. (2008). Choosing the number of moments in conditional moment restriction models.

[13] Hansen., L, Large Sample Properties of Generalized Method of Moments Estimators, Econometrica, 1982, vol. 50, issue 4, pages 1029-54.

[14] Han, C., and Phillips, P. C. (2006). GMM with many moment conditions. Econometrica, 74(1), 147-192.

[15] Hansen, C., Hausman, J., and Newey, W. (2008). Estimation with many instrumental variables. Journal of Business and Economic Statistics, 26(4).

[16] Hausman, J., Menzel, K., Lewis, R., and Newey, W. (2007). A reduced bias GMM-like estimator with reduced estimator dispersion (No. CWP24/07). cemmap working paper, Centre for Microdata Methods and Practice.

[17] Huang, J., Ma, S., and Zhang, C. H. (2008). Adaptive Lasso for sparse high-dimensional regression models. Statistica Sinica, 18(4), 1603.

[18] Jing, B. Y., Shao, Q. M., and Wang, Q. (2003). Self-normalized Cramĺęr-type large deviations for independent random variables. The Annals of probability, 31(4), 2167-2215.

[19] Newey, W. K., and Smith, R. J. (2004). Higher order properties of GMM and generalized empirical likelihood estimators. Econometrica, 72(1), 219-255.

[20] Newey, W. K., and Windmeijer, F. (2009). Generalized method of moments with many weak moment conditions. Econometrica, 77(3), 687-719.

[21] Rosenbaum, M., and Tsybakov, A. B. (2010). Sparse recovery under matrix uncertainty. The Annals of Statistics, 38(5), 2620-2651

[22] Shi, Z. (2013). Econometric Estimation with High-Dimensional Moment Equalities. In Meeting of the Econometric Society in June (p. 1).

[23] Tchuente, G., and Carrasco, M. (2013). Regularized LIML for many instruments (No. 2013s-20). CIRANO.

[24] Tibshirani, R. (1996). Regression shrinkage and selection via the lasso. Journal of the Royal Statistical Society. Series B (Methodological), 267-288.

[25] van de Geer, S., Buhlmann, P., and Zhou, S. (2010). The adaptive and the thresholded Lasso for potentially misspecified models. arXiv preprint arXiv:1001.5176.

[26] Zhang, C. H., and Huang, J. (2008). The sparsity and bias of the Lasso selection in high-dimensional linear regression. The Annals of Statistics, 1567-1594.

[27] Zhao, P., and Yu, B. (2006). On model selection consistency of Lasso. The Journal of Machine Learning Research, 7, 2541-2563.

[28] Zou, H. (2006). The adaptive lasso and its oracle properties. Journal of the American statistical association, 101(476), 1418-1429.

Chapter 2

Core Determining Class: Construction, Approximation, and Inference

In this modern era of "Big Data", it is important to draw information and create values from the fast expanding datasets. We call the observable data as outcomes, and the unobservable sources which lead to the outcomes as events. In many situations the relations between events and outcomes are indeterministic, i.e., a single event may lead to different outcomes. Such relations can be characterized by a bipartite graph $G = (\mathcal{U}, \mathcal{Y}, \varphi)$, where $\mathcal{U}$ is a set of unobservable events, $\mathcal{Y}$ is a set of observed outcomes, and φ is a correspondence mapping from $\mathcal{U}$ to $\mathcal{Y}$ such that $\varphi(u) \subset \mathcal{Y}$ is the set of all possible outcomes that could be led by event $u \in \mathcal{U}$. In this paper, we consider estimating the probability measure on $\mathcal{U}$ given observations on $\mathcal{Y}$. One application is to infer individual player's private information given the observations of players' strategies when there exist multiple equilibria; another application is to infer demand/customer characteristics given the purchase histories and sales data.

The feasible set of probability measure on $\mathcal{U}$ is defined by a set of linear inequality constraints. In general, the number of inequality constraints could grow exponentially with $|\mathcal{U}|$. Such many inequalities may lead to two problems for performing inference on the measure on $\mathcal{U}$: (1) traditional inference procedures such as those described in Chernozhukov, Hong and Tamer (2007) (later CHT) may fail; (2) traditional inference

procedures are computationally intractable when $|\mathcal{U}|$ is large.

If we can dramatically reduce the number of inequalities defining the feasible set of probability measure on $\mathcal{U}$, we are able to perform valid inference with much less computational cost. Notice that there may exist many redundant or nearly redundant inequalities, we aim to select the informative ones from the full set of inequalities:

(1) We propose a method to construct the set of irredundant inequalities for the bipartite graph when data noise is not taken into consideration. Such set is referred as Core Determining Class described in Galichon and Henry (2011). We prove that the inequalities selected are independent from the probability measure observed on $\mathcal{Y}$ under certain mild conditions.

(2) For a general problem of linear inequalities selection under noise, we propose a selection procedure similar to the Dantzig-selector described in Candes and Tao (2007). We prove that the selection procedure has good statistical properties under some sparse assumptions.

(3) We apply the selection procedure to construct the set of irredundant inequalities for the bipartite graph with data noise. We prove that the selection procedure has better statistical properties compared to that applied to the general problem due to the structure of the graph.

(4) We demonstrate the good performance of our selection procedure through several sets of Monte-Carlo experiments: first, the inference based on the selection procedure has desired size; second, it has strong power against local alternatives; third, it is relatively computationally efficient.

The closest researches to our topic are Galichon and Henry (2006, 2011) and Chesher and Rosen (2012). Galichon and Henry (2011) proposes the Core Determining Class problem, i.e., finding the minimum set of inequalities to describe the feasible region of probability measure on $\mathcal{U}$. Chesher and Rosen (2012) provides an inequality selection algorithm, but may still contain some redundant inequalities in the selected set. Andrews and Soares(2013) proposes moment inequality selection procedure using criterions such as BIC.

There are many studies on performing inference of sets. CHT (2007) proposes general inference procedure with moment inequality constraints. Romano and Shaikh (2010) provides improvements for CHT (2007). Beresteanu, Molchanov and Molinari (2011) uses random set theory to perform inference with convex inequality restrictions. Andrews and Shi (2013) construct inference based on conditional moment inequalities. For related empirical studies, see Tamer and Manski (2002), Bajari, Benkard and Levin (2004), Bajari, Hong and Ryan (2010) and etc..

There is also a wide literature on detection and elimination of redundant constraints when data noise is not taken into consideration. For example, Telgen (1983) develops two methods to identify redundant constraints and implicit equalities. Caron, McDonald and Ponic (1989) presents a degenerate extreme point strategy which classifies linear constraints as either redundant or necessary. Paulraj, Chellappan and Natesan (2010) proposes a heuristic approach using an intercept matrix to identify redundant constraints.

We organize the paper as follows: Section 2.1 introduces the model and basic assumptions through out the entire paper. Section 2.2 studies the Core Determining Class from the structure of the bipartite graph and provides a method to construct the exact Core Determining Class when data noise is not taken into consideration. Section 2.3 proposes a general linear inequalities selection procedure under noisy data with the definition of sparse assumptions. Section 2.4 discusses the additional technical assumptions and proves main theorems of the statistical properties of the selection procedure, with application to the Core Determining Class. Section 2.5 implements our selection procedure in a large bipartite graph through Monte-Carlo experiments and illustrates its performance. Section 2.6 concludes the paper.

2.1 Core Determining Class

Given a bipartite graph $G = (\mathcal{U}, \mathcal{Y}, \varphi)$, suppose $\mathcal{U}$ is a set vertices representing events, and $\mathcal{Y}$ is a set of vertices representing outcomes. Suppose an event $u \in \mathcal{U}$ leads to a set of possible outcomes $\varphi(u) \subset \mathcal{Y}$, where $\varphi(u)$ is a set of vertices in $\mathcal{Y}$. For any

set $A \subset \mathcal{U}$, $\varphi(A) := \cup_{u \in A} \varphi(u)$. Therefore, $\varphi : 2^{\mathcal{U}} \mapsto 2^{\mathcal{Y}}$ is a correspondence mapping between $\mathcal{U}$ and $\mathcal{Y}$. The inverse of φ, denoted as φ^{-1} is defined as $\varphi^{-1} : 2^{\mathcal{Y}} \mapsto 2^{\mathcal{U}}$, $\varphi^{-1}(B) = \{u \in \mathcal{U} | \varphi(u) \cap B \neq \emptyset\}$, $\forall B \subset \mathcal{Y}$.

Let v be the probability measure on $\mathcal{U}$. Let $\mu_{n,0}$ be the true measure on $\mathcal{Y}$ which could change with the model. Let $\hat{\mu}_n$ be the measure observed in a sample set of outcomes $\mathcal{Y}$. Denote $d_1 = |\mathcal{U}|$ and $d_2 = |\mathcal{Y}|$. For a graph $G = (\mathcal{U}, \mathcal{Y}, \varphi)$, say G is connected if $\forall A_1, A_2 \subset G$ and $A_1 \cup A_2 = G$, it holds that $\varphi(A_1) \cap \varphi(A_2) \neq \emptyset$.

Assumption C.12 (Non-Degeneracy of G, $\mu_{n,0}$ and $\hat{\mu}_n$) *(1) Assume G is connected. We say G is non-degenerate if G is connected.*

(2) For the probability measure $\mu = \mu_{n,0}$ or $\hat{\mu}_n$, assume that for any $y \in \mathcal{Y}$, $\mu(y) > 0$. We say that μ is non-degenerate if $\mu(y) > 0$ for any $y \in \mathcal{Y}$.

We assume that Assumption C.12 holds through out the paper.

The parameter of interest in this paper is the $d_1 \times 1$ vector v, which is the probability measure which generates the events $u \in \mathcal{U}$. In general we are unable to obtain a point estimation of v unless additional information is provided. Instead, we can obtain inequality bounds on v given the bipartite graph $G = (\mathcal{U}, \mathcal{Y}, \varphi)$ and the measure μ on $\mathcal{Y}$. More specifically, for any set of events $A \subset \mathcal{U}$, the outcome should fall into the set $\varphi(A)$. Thus, for any $A \subset \mathcal{U}$, we can obtain the inequality $v(A) := \sum_{u \in A} v(u) \leqslant \mu(\varphi(A)) := \sum_{y \in \varphi(A)} \mu(y)$.

The Artstein's theorem stated in Artstein (1983) presents that all information of v in the biparte graph model $G = (\mathcal{U}, \mathcal{Y}, \varphi)$ is characterized by the set of constraints described below:

Lemma 12 (Artstein's Theorem) *The following set of inequalities/equalities contains sharp information on v:*

1. *For any $A \subset \mathcal{U}$,*

$$v(A) := \sum_{u \in A} v(u) \leqslant \mu(\varphi(A)),$$

where $\mu(\varphi(A)) := \sum_{y \in \varphi(A)} \mu(y)$;

2. $\sum_{u \in \mathcal{U}} v(u) = 1$.

Our model, denoted as $\mathcal{P}_G$, is presented below:

Definition 2.1.1 (Model $\mathcal{P}_G$) *Find the set of all feasible probability measure v on $\mathcal{U}$ such that:*

(1) For any $A \subset \mathcal{U}$, $v(A) \leqslant \mu(\varphi(A))$;

(2) $\sum_{u \in \mathcal{U}} v(u) = 1$.

Comment 2.1.1 *The non-degeneracy assumption prevents the problem $\mathcal{P}_G$ from decomposition, i.e., we can not decompose graph G into G_1 and G_2 and proceed with problem $\mathcal{P}_{G_1}$ and $\mathcal{P}_{G_2}$. Otherwise the problem can be simplified by looking at G_1 and G_2 separately.*

In general, the set of inequality constraints stated in Definition 2.1.1 contains redundant inequalities. Define the minimum model T_0 of $\mathcal{P}_G$ as the set of linear constraints stated in (1) such that T_0 together with the equality (2) has the minimum number of constraints which generate the same set of feasible measure as $\mathcal{P}_G$. In other words, T_0 consists of all irredundant constraints in $\mathcal{P}_G$. If the number of irredundant constraints in T_0 is much less than $2^{d_1} - 1$ stated in Definition 2.1.1, then it is more accurate and computational efficient to conduct inference on the Core Determining Class using T_0. Galichon and Henry (2011) proposes the concept "Core Determining Class" as follows

Definition 2.1.2 (Core Determining Class problem) *The Core Determining Class problem is the problem of finding all binding constraints in model $\mathcal{P}_G$. The Core Determining Class is any collection of subsets of $\mathcal{U}$ that contains the sharp information, i.e., the corresponding inequalities includes all binding inequalities. The exact Core Determining Class is defined as the set of subsets of $\mathcal{U}$ which corresponds to the irredundant inequalities in model T_0.*[1]*:*

[1]The definition of Core-Determining Class in Galichon and Henry (2006) is slightly different from ours. Galichon and Henry (2006) defines Core-Determining Class as any set that contains all the binding inequalities. In this paper, we refer "exact Core-Determining Class" as the set of binding inequalities, i.e., the smallest set (in cardinality) which characterizes the identified set of parameter of interest.

Comment 2.1.2 *In many cases there may exist a parametric model for v, denoted as $v_i = F_i(\theta)$. The function F_i can be non-linear. The inference on θ can be generally difficult if the number of inequalities about v is large. Therefore, we can find the truly binding inequalities about v, we would perform estimation and inference on θ much faster.*

We provide an example on the model $\mathcal{P}_G$.

Example 8 (Two players entry game) *Suppose there are two firms in a market. The cost for firm 1 and firm 2 is $c + r_1$ and $c + r_2$ respectively, where c is a constant, r_1 and r_2 are random shocks which are observable only by the corresponding firm.*

The two firms face a total demand $D = a_1 - a_2 p$. If they are both in the market, they will play a Cournot Nash equilibrium. If there is only one player, then this player will reach a monopolist's equilibrium. If the costs are too large for both players that even a monopolist is unprofitable, then there will be no player in the market. Therefore, there are 4 possible equilibria: $(0,0)$, $(1,0)$, $(0,1)$, and $(1,1)$:

(1) if $\frac{a_1}{a_2} - c \geqslant 2/3r_1 - 1/3r_2$ and $\frac{a_1}{a_2} - c \geqslant 2/3r_2 - 1/3r_1$, then the equilibrium is $(1,1)$;

(2) if $\frac{a_1}{a_2} - c < 2/3r_1 - 1/3r_2$ and $\frac{a_1}{a_2} - c \geqslant 2/3r_2 - 1/3r_1$, then the equilibrium is $(0,1)$;

(3) if $\frac{a_1}{a_2} - c \geqslant 2/3r_1 - 1/3r_2$ and $\frac{a_1}{a_2} - c < 2/3r_2 - 1/3r_1$, then the equilibrium is $(1,0)$;

else if $\frac{a_1}{a_2} - c < 2/3r_1 - 1/3r_2$ and $\frac{a_1}{a_2} - c < 2/3r_2 - 1/3r_1$:

(4) if $c + r_1 \leqslant \frac{a_1}{a_2}$ and $c + r_2 \leqslant \frac{a_1}{a_2}$, then there are two equilibria:$(1,0)$ and $(0,1)$;

(5) if $c + r_1 \leqslant \frac{a_1}{a_2}$ and $c + r_2 > \frac{a_1}{a_2}$, then the equilibrium is: $(1,0)$;

(6) if $c + r_1 > \frac{a_1}{a_2}$ and $c + r_2 \leqslant \frac{a_1}{a_2}$, then the equilibrium is: $(0,1)$;

(7) if $c + r_1 > \frac{a_1}{a_2}$ and $c + r_2 > \frac{a_1}{a_2}$, then the equilibrium is: $(0,0)$.

Let $\mathcal{U} = \{u_1, u_2, u_3, u_4, u_7\}$, where u_i is the event representing case (i), with the exceptions that u_2 represents (2) and (6), and u_3 represents (3) and (5). Let $Y :=$

$\{y_1, y_2, y_3, y_4\}$, *where* $y_1 = (1,1)$, $y_2 = (0,1)$, $y_3 = (1,0)$, *and* $y_4 = (0,0)$. *So* $d_1 = |\mathcal{U}| = 5$ *and* $d_2 = |\mathcal{Y}| = 4$. *The correspondence mapping* φ *between* $\mathcal{U}$ *and* $\mathcal{Y}$ *is:*

$\varphi(u_1) = \{y_1\}$, $\varphi(u_2) = \{y_2\}$, $\varphi(u_3) = \{y_3\}$, $\varphi(u_4) = \{y_2, y_3\}$, *and* $\varphi(u_7) = \{y_4\}$.

The correspondence mapping for Example 8 is illustrated in Figure 2-1.

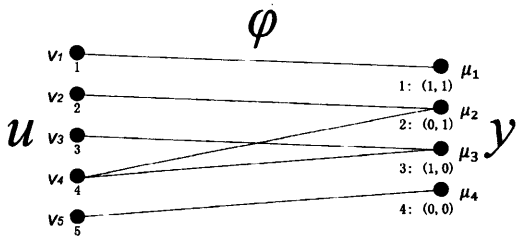

Figure 2-1: Correspondence Mapping for Example 8

Given the probability measure μ *on* $\mathcal{Y}$, *the bounds of the probability measure* v *on* $\mathcal{U}$ *is given by the inequalities stated in the Artstein's theorem. According to the Artstein's theorem statement (1), there are* $2^5 - 2 = 30$ *inequalities. In fact, it is obvious that the Core-Determining Class in this example consist of only* 5 *sets (inequalities):* $\{u_1\}$, $\{u_2\}$, $\{u_3\}$, $\{u_2, u_3, u_4\}$ *and* $\{u_5\}$.

In reality, the true probability measure $\mu_{n,0}$ on the outcome set $\mathcal{Y}$ is unobservable. Instead, given the data, we could observe the empirical measure $\hat{\mu}_n$ on $\mathcal{Y}$. Due to uncertainty of the data, we would like to solve a relaxed problem $\mathcal{P}'_G$, whose solution set covers the solution set of the true model $\mathcal{P}_G$ with probability approaching 1 as the data sample size n approaching infinity. This relaxed problem $\mathcal{P}'_G$ provides conservative inference for model $\mathcal{P}_G$.

Definition 2.1.3 (Model $\mathcal{P}'_G$) *For a small* λ, *find the set of all feasible probability measure* v *on* $\mathcal{U}$ *such that:*

(1) For any $A \subset \mathcal{U}$, $v(A) := \sum_{u \in A} v(u) \leqslant \hat{\mu}_n(\varphi(A)) + \lambda$;

(2) $\sum_{u \in \mathcal{U}} v(u) = 1$.

Ideally λ should converge to 0 when $n \to \infty$. The dimensionality of the problem, $|\mathcal{U}|$, and the number of inequalities in $\mathcal{P}'_G$, should affect the tuning parameter λ. In fact, λ should be chosen properly such that: (1) the feasible set of v found in model $\mathcal{P}'_G$ covers the feasible set of v found in model $\mathcal{P}_G$ with probability approaching 1, so $\mathcal{P}'_G$ provides inference on $\mathcal{P}_G$; and (2) λ is not be too large to exaggerate the feasible set of v found in model $\mathcal{P}'_G$. We will discuss the choice of λ in Section 2.3.

According to the Artstein's theorem, model $\mathcal{P}_G$ contains $2^{d_1} - 2$ inequalities. It is a very large number when d_1 is large and even grows with n in some contexts. The numerous inequalities lead to both computational difficulties and undesirable statistical properties. In fact, some or even most of the inequalities stated in the Artstein's theorem may be redundant. Galichon and Henry (2011) analyzes the monotonic structure of the graph G and claims that there are at most $2d_1 - 2$ sets in the Core Determining Class under a special structure. Chesher and Rosen (2012) provides an algorithm which could get rid of some, but not necessarily all redundant inequalities. In Section 2.2, we fully characterize the Core Determining Class by the exploring the combinatorial structure of the bipartite graph G. We prove that the Core Determining Class only rely on the structure of G under the non-degeneracy assumption of μ. The results are novel compared to existing studies. We also propose a fast algorithm in Section 2.2 to compute the exact Core Determining Class when data noise is not taken into consideration.

In addition, besides those redundant inequalities, many of the binding inequalities could be "nearly" redundant, meaning that although they are informative in Model $\mathcal{P}_\mathcal{G}$ with empirical $\hat{\mu}$, they could be "implied" by other inequalities in Model $\mathcal{P}'_G$ with a small relaxation λ. Therefore, it may be possible to use a smaller number of inequalities, i.e., a "small" model, to approximate the full one. Such a small model will enjoy better statistical properties compared to the full model, i.e., it will be less sensitive to modeling errors. We propose an general inequality selection procedure similar to the Dantzig Selector in Section 2.2.

2.2 Exact Core Determining Class

In this section, we present our discovery of the combinatorial structure of the Core Determining Class, along with a fast algorithm to generate the Core Determining Class. In Galichon and Henry (2011), whether an inequality $v(A) \leqslant \mu(\varphi(A))$ is in the Core Determining Class is examined by numerical computation using the probability measure μ.

In fact, given the correspondence mapping φ of the bipartite graph $G = (\mathcal{U}, \mathcal{Y}, \varphi)$, we can identify the redundant inequalities without any observations of the outcomes in $\mathcal{Y}$. For example, for $A_1 \in \mathcal{U}$ and $A_2 \in \mathcal{U}$, if $A_1 \cap A_2 = \emptyset$ and $\varphi(A_1) \cap \varphi(A_2) = \emptyset$, then the two inequalities, $v(A_1) \leqslant \mu(\varphi(A_1))$ and $v(A_2) \leqslant \mu(\varphi(A_2))$ can generate the inequality $v(A_1 \cup A_2) = v(A_1) + v(A_2) \leqslant \mu(\varphi(A_1)) + \mu(\varphi(A_2)) = \mu(\varphi(A_1) \cup \varphi(A_2)) = \mu(\varphi(A_1 \cup A_2))$, which is exactly the inequality corresponding to $A = A_1 \cup A_2$. In another word, the inequality $v(A) \leqslant \mu(A)$ is redundant given $v(A_1) \leqslant \mu(\varphi(A_1))$ and $v(A_2) \leqslant \mu(\varphi(A_2))$. Also, if $u \notin A$ satisfies $\varphi(\{u\}) \subset \varphi(A)$, then the inequality $v(\{u\} \cup A) \leqslant \mu(\varphi(\{u\} \cup A))$ will imply a redundant inequality $v(A) \leqslant \mu(\varphi(A))$.

In this section, we propose a combinatorial method to generate the exact Core Determining Class. We prove that, in theory, if the probability measure μ is non-degenerate, our method excludes all redundant inequalities in the model $\mathcal{P}_G$ regardless the values of μ. That is to say, the Core Determining Class can be exactly constructed with the method and the Core Determining Class is independent from μ.

Definition 2.2.1 (Set $\mathcal{S}_u$) *$\mathcal{S} \subset 2^{\mathcal{U}}$ is the collection of all non-empty subsets $A \subset \mathcal{U}$ and $A \neq \mathcal{U}$, such that*

$$v^M(A) > \mu(\varphi(A)),$$

where $v^M(A) := \max\{v(A) | v(A') \leqslant \mu(\varphi(A')), \forall A' \subset \mathcal{U}, A' \neq A\}$.

Set $\mathcal{S}_u$ is defined with probability measure μ. The inequality generated by any $A \in \mathcal{S}_u$ is informative: it is irredundant given other inequalities described in statement (1) of the Artstein's theorem. Essentially, $\mathcal{S}_u$ identifies the irreducible inequalities for Model $\mathcal{P}_G$ when the critical equality $\sum_{u \in \mathcal{U}} v(u) = 1$ is not taken into consideration.

Definition 2.2.2 (Set $\mathcal{S}'_u$) *$\mathcal{S}' \subset 2^{\mathcal{U}}$ is the collection of all non-empty subsets $A \subset \mathcal{U}$ and $A \neq \mathcal{U}$, such that:*

(1)A is self-connected, i.e., $\forall A_1, A_2 \subset A$ such that $A_1, A_2 \neq \emptyset$ and $A_1 \cup A_2 = A$, it holds that $\varphi(A_1) \cap \varphi(A_2) \neq \emptyset$;

(2)There exists no $u \in \mathcal{U}$, such that $u \notin A$ and $\varphi(u) \subset \varphi(A)$.

Lemma 13 *If μ is non-degenerate, the collection of subsets defined in Definition 2.2.1 and Definition 2.2.2 are identical. $\mathcal{S}_u = \mathcal{S}'_u$.*

$\mathcal{S}_u$ and $\mathcal{S}'_u$ describe the irreducible inequalities in $\mathcal{P}_G$ if the equality $\sum_{u\in\mathcal{U}} v(u) = 1$ is not taken into consideration. Theorem 5 of Chesher and Rosen (2012) proposes a subset of inequalities with property (1) stated in Definition 2.2.2. This subset contains the set of all binding inequalities, which is Core Determining. Lemma 13 shows that with an additional property (2) in Definition 2.2.2, we can find all binding inequalities without considering the equality: $\sum_{u\in\mathcal{U}} v(u) = 1$. In fact, adding this equality can further substantially reduce the number of inequalities and it is impossible to find the minimum set of inequalities in $\mathcal{P}_G$ without the key equation $\sum_{u\in\mathcal{U}} v(u) = 1$. To find the minimum binding set of inequalities, i.e., the exact Core-Determining Class, we look at the problem $\mathcal{P}_G$ from the opposite direction: consider the inequalities from $\mathcal{Y}$ to $\mathcal{U}$. For any non-degenerate probability measure $\widetilde{v}$ on $\mathcal{U}$, we define $\mathcal{S}_y$ and $\mathcal{S}'_y$, which are collection of subsets of $\mathcal{Y}$ and similar to $\mathcal{S}_u$ and $\mathcal{S}'_u$.

Definition 2.2.3 (Set $\mathcal{S}_y$) *Given a non-degenerate probability measure $\widetilde{v}$ on $\mathcal{U}$, $\mathcal{S}_y \subset 2^{\mathcal{Y}}$ is the collection of all subsets $B \subset \mathcal{Y}$ and $B \neq \mathcal{Y}$, such that*

$$\mu^M(B) > \widetilde{v}(\varphi^{-1}(B)),$$

where $\mu^M(B) := \max\{\widetilde{\mu}(B) | \widetilde{\mu}(B') \leqslant \widetilde{v}(\varphi^{-1}(B')), \forall B' \subset \mathcal{Y}, B' \neq B\}$, where $\widetilde{\mu}$ is a probability measure on $\mathcal{Y}$.

Definition 2.2.4 (Set $\mathcal{S}'_y$) *$\mathcal{S}'_y \subset 2^{\mathcal{Y}}$ is the collection of all subsets $B \subset \mathcal{Y}$ and $B \neq \mathcal{Y}$, such that:*

(1) B is self-connected, i.e., $\forall B_1, B_2 \subset B$, such that $B_1, B_2 \neq \emptyset$ and $B_1 \cup B_2 = B$, it holds that $\varphi^{-1}(B_1) \cap \varphi^{-1}(B_2) \neq \emptyset$;

(2) There exists no $y \in \mathcal{Y}$, such that $y \notin B$ and $\varphi^{-1}(y) \subset \varphi^{-1}(B)$.

The Lemma below presents result similar to Lemma 13.

Lemma 14 $\mathcal{S}_y = \mathcal{S}'_y$.

Definition 2.2.5 (Set $\mathcal{S}_y^{-1}$) *Set $\mathcal{S}_y^{-1}$ is the collection of $A \subset \mathcal{U}$ and $A \neq \mathcal{U}$ such that there exists $B \subset \mathcal{S}'_y$ that $A = \varphi^{-1}(B)^c$.*

Below we give a numerical definition of the exact Core Determining Class using linear programming:

Definition 2.2.6 (Set $\mathcal{S}^*$) *The Core Determining Class $\mathcal{S}^*$ is the collection of all subsets $A \subset \mathcal{U}$ and $A \neq \mathcal{U}$, such that*

$$v^{M^*}(A) > \mu(\varphi(A)),$$

where $v^{M^*}(A) := \max\{v(A) | v(A') \leqslant \mu(\varphi(A')), \forall A' \subset \mathcal{U}, A' \neq A; v(\mathcal{U}) = 1\}$.

In the definition above, the equality $v(\mathcal{U}) = 1$ is considered. $\mathcal{S}^*$ are subsets in $\mathcal{U}$ corresponding to irreducible inequalities under $v(\mathcal{U}) = 1$. The theorem below characterizes the Core Determining Class $\mathcal{S}^*$:

Theorem 4 *The Core Determining Class is characterized by the following equation:*

$$\mathcal{S}^* = \mathcal{S}_u \cap \mathcal{S}_y^{-1}$$

Notice that both $\mathcal{S}_u$ and $\mathcal{S}_y^{-1}$ are defined via combinatorial rules, the Core Determining Class is independent from μ if μ is non-degenerate.

In Example 9, we show that considering only S_u may not able to substantially reduce the number of inequalities, where $S_u \cap S_y^{-1}$ can be a very small set in cardinality.

Example 9 *Consider set* $\mathcal{U} = \{u_1, ..., u_{d_1}\}$ *and set* $\mathcal{Y} = \{y_1, ..., y_{d_1+1}\}$. φ *is the correspondence mapping between* $\mathcal{U}$ *and* $\mathcal{Y}$ *such that:*

$$\varphi(u_j) = \{y_j, y_{d_1+1}\}$$

for all $1 \leqslant j \leqslant d_1$.

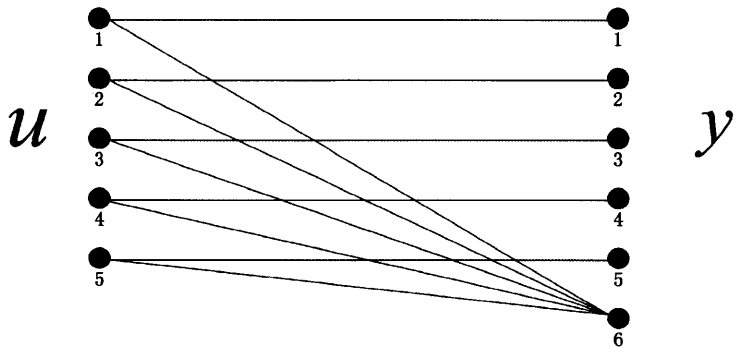

Figure 2-2: Correspondence Mapping of Example 9

Considering only S_u *would obtain* $\mathcal{S} = 2^{\mathcal{U}} - \{\emptyset, \mathcal{U}\}$, *which consists of* $2^{d_1} - 2$ *subsets and essentially make no selection of inequalities. The Core Determining Class* $\mathcal{S}^*$ *constructed in our approach is* $\{\mathcal{U} - u_j | 1 \leqslant j \leqslant d_1\}$. *It is obvious that this is the minimum number of subsets carrying full information for model* $\mathcal{P}_G$. *The Core Determining Class* $\mathcal{S}^*$ *contains* d_1 *inequalities, which is much less than* $2^{d_1} - 2$ *inequalities selected by Theorem 5 of Chesher and Rosen (2012).*

We utilize the combinatorial structure revealed in Definition 2.2.2 and Definition 2.2.4 to construct $\mathcal{S}'_u$ and $\mathcal{S}'_y$: algorithm 1 computes $\mathcal{S}'_u$ and a similar algorithm computes $\mathcal{S}'_y$ and $\mathcal{S}_y^{-1}$. Then we obtain the Core Determining Class $\mathcal{S}^* = \mathcal{S}_u \cap \mathcal{S}_y^{-1}$.

The complexity of the algorithm is $o(2^{\max(d_u, d_y)} \cdot d_1^2 \cdot d_2^2)$, where

d_u is defined as

$$d_u := \max_A |A|$$

$$s.t. A \subset \mathcal{U}$$

$$\varphi(A) = \mathcal{Y}$$

$$\varphi(A/u) \subsetneq \mathcal{Y}, \forall u \in A$$

d_y is defined as

$$d_y := \max_B |B|$$

$$s.t. B \subset \mathcal{Y}$$

$$\varphi^{-1}(B) = \mathcal{U}$$

$$\varphi^{-1}(B/y) \subsetneq \mathcal{U}, \forall y \in B$$

Under the assumption of non-degenerate G and μ, in a bipartite graph with practical application, d_u and d_y is much smaller than d_1 and d_2 respectively, so the algorithm is fast in practice.

Algorithm 3 *Input: Bipartite Graph* $G = (\mathcal{U}, \mathcal{Y}, \varphi)$

Output: Set S'_u

Initiation: $S'_u = \{\emptyset\}$

for $i = 0$ *to* $|\mathcal{U}| - 1$ *do*

Identify additional $A' \in S'_u$ *as union of* $u \in \mathcal{U}$ *with* $|A| = i$

for each $A \in S'_u$ *with* $|A| = i$ *do*

for each $u \notin A$ *that* $\varphi(u) \cap \varphi(A) \neq \emptyset$ *do*

$A' = A \cup \{u\}$

if $\varphi(A') < 1$ *then*

for each $u' \notin A'$ *do*

if $\varphi(u') \subset \varphi(A')$ *then* $A' = A' \cup u'$

end

end

end

end

Termination: $S'_u = S'_u - \{\emptyset\}$

2.3 A general selection procedure and sparse assumption

Essentially, the objective of model $\mathcal{P}_G$ is to obtain a feasible set of v given observation $\hat{\mu}$, i.e., to obtain $\hat{Q} := \{v | v(A) \leqslant \hat{\mu}_n(\varphi(A)), \forall A \subset \mathcal{U};\ \sum_{u \in \mathcal{U}} v(u) = 1\}$. In Section 3 we explore the structure of the bipartite graph G to obtain the set of irreducible inequalities to define $\hat{Q}$. In this section, we propose a procedure for a general problem of linear inequalities selection under data noise. This procedure chooses the set of linear inequalities with sharp information as $n \to \infty$. It can identify the inequalities which are binding but "close" to redundant, so to further reduce the number of inequalities in $\hat{Q}$. The procedure can be applied to general linear inequality selection problems, including Core Determining Class problem allowing mixed strategy as defined in Galichon and Henry (2011).

2.3.1 General Selection Procedure

Problem $\mathcal{P}$ can be interpreted as computing the feasible region of a collection of linear inequality constraints. It could be generalized as computing the feasible region of

$$Q := \{v | Mv \leqslant b, v \geqslant 0\},$$

where M is a $m \times d_1$ matrix, v is $d_1 \times 1$ vector, and b is a $m \times 1$ vector.

In many situations the number of inequalities, m, is too large to effectively conduct any known estimation and inference procedure such as CHT inference. For example, there are $m = 2^{d_1}$ inequalities in the Core Determining Class problem without any

inequalities selection procedures.[2] There are two reasons that we do not use the entire set of inequalities: first, there could be many redundant inequalities which are not informative at all; second, when m and d_1 are growing, there could be many inequalities which are nearly redundant, compared to the size of noise in the data.

Notice that the random noise of b, which comes from $\hat{\mu}_n - \mu_{n,0}$, is ignored in Section 3 when data noise is not taken into consideration. In this section, we develop a procedure to select informative inequalities in a general Q considering the random noise on b.

For any subset $\mathcal{I}$ of $\{1, 2, ..., m\}$, denote $M_{\mathcal{I}}$ as the matrix comprised of the rows indexed by $\mathcal{I}$ in matrix M. Similarly, denote $b_{\mathcal{I}}$ as the subvector of b indexed by $\mathcal{I}$. By the Farkas Lemma, for a general matrix M and a vector b, if the set of constraints indexed by $\mathcal{I}$ can imply all other constraints, i.e., the set $Q_{\mathcal{I}} := \{v | M_{\mathcal{I}} v \leqslant b_{\mathcal{I}}, v \geqslant 0\}$ equals Q, then there must exist a non-negative $m \times m$ matrix Π such that:

$$(1)\ \Pi M \geqslant M,$$

$$(2)\ \Pi b \leqslant b,$$

$$(3)\ \Pi_{lj} = 0,$$

for any $1 \leqslant l \leqslant m$ and $j \notin \mathcal{I}$.

For any $j \in \{1, 2, ..., m\}$, denote M_j as the j^{th} row of M, denote Π_j as the j^{th} row of Π and Π^k as the k^{th} column of Π.

The coefficient matrix Π described above can serve as a signal of the importance of each inequality. If all the coefficients of the k^{th} inequality, Π^k, are zero or close to 0, then this inequality is not very informative to v. Inspired by the Farkas Lemma, we propose the following selection procedure which slightly relaxes the constraints on Π:

Problem $\hat{\mathcal{R}}$:

$$\min_{\Pi} \sum_{k=1}^{m} g(\Pi^k)$$

[2]We could view $v(\mathcal{U}) = 1$ as two inequalities: $v(\mathcal{U}) \leqslant 1$ and $v(\mathcal{U}) \geqslant 1$.

subject to:

$$(1)\Pi M \geqslant M, \Pi \geqslant 0,$$

$$(2)\Pi\hat{b} \leqslant \hat{b} + \Lambda_n,$$

where observed $\hat{b}$ is a $m \times 1$ vector which converges to b as the data sample size n goes to ∞, and $\Lambda_{n,m} = (\lambda_{n,m}, \lambda_{n,m}, ..., \lambda_{n,m})'$ in which $\lambda_{n,m}$ is a relaxing parameter measuring the maximum error allowed for each inequality.[3]

We choose the objective function $g(\cdot)$ such that it measures the importance of the constraints. One choice is $g(\Pi^k) = sign(\Sigma_{1\leqslant j\leqslant m}\Pi_{jk})$. With this function $g(\cdot)$, the selection procedure $\hat{\mathcal{R}}$ is essentially a binary integer programming to select the minimum number of inequalities. We call such a procedure the "L^0 selector".

The L^0 selector is extremely difficult to implement when m is large. However, many studies on LASSO and the Dantzig Selector show that some L^1 objective functions could enjoy nice statistical properties in model selection and low computational costs. Below we propose a feasible L^1 objective function $g(\cdot)$:

$$g(\Pi^k) = \max_{1\leqslant j\leqslant m} \Pi_{jk}, \tag{2.3.1}$$

where Π_{jk} is the $(j,k)^{th}$ entry of Π.

With the above choice of $g(\cdot)$, the formulation of the problem $\hat{\mathcal{R}}$ is rewritten as:

Problem $\hat{\mathcal{R}}$:

$$\min_{\Pi} \sum_{k=1}^{m} \max_{1\leqslant j\leqslant m} \Pi_{jk}$$

subject to:

$$(1)\ \Pi M \geqslant M, \Pi \geqslant 0,$$

$$(2)\ |(\Pi\hat{b} - \hat{b})_+|_{l^\infty} \leqslant \lambda.^4$$

[3]The Vector $\Lambda_{n,m}$ can also be chosen to be specific to each inequality. For the inequality which we believe to be too important to be ruled out, we could set the corresponding $\lambda_{n,m}$ in $\Lambda_{n,m}$ to 0.

[4] The formulation of the problem $\hat{\mathcal{R}}$ is similar to the Dantzig Selector described in Candes and Tao (2005). The main difference is that the Dantzig Selector has two-sided constraints, which shrink the feasible solution to a point, while our problem has one-sided constraints, which consign the feasible solution to a convex set. The benefit of this formulation is that it turns an integer programming problem

2.3.2 Sparse Assumption of the Problem $\hat{\mathcal{R}}$

Sparse assumptions play the essential role in the analysis of some L^1 penalization procedures, such as LASSO and the Dantzig Selector. In this subsection, we define a sparse assumption for the Problem $\hat{\mathcal{R}}$.

For any $1 \leqslant j \leqslant m$, define separation of inequality j as:

$$c_j := \max_{v \in Q_j} M_j v - b_j,$$

where

$$Q_j := \{v | M_i v \leqslant b_i, \forall i \neq j; v \geqslant 0\}$$

c_j measure the maximal separation of the j^{th} inequality for all points in Q_j. If $c_j > 0$, the j^{th} inequality is irredundant, otherwise the j^{th} inequality is redundant. Let T_0 be the set of indices j with $c_j > 0$ to denote the set of irredundant inequalities. Since c_j characterizes the information carried by the j^{th} inequality, we define a sparse assumption using c_j.

Definition 2.3.1 (Exact Sparse) *Recall that T_0 is the subset of $\{1, 2, ..., m\}$ denoting all irredundant inequalities. Let $\widetilde{\Pi}^*$ be the solution of the following problem:*

Problem $\mathcal{R}$:

$$\min_{\Pi} \sum_{k \in T_0} g(\Pi^k)$$

subject to:

$$(a) \Pi M \geqslant M, \Pi \geqslant 0,$$

$$(b) \Pi b \leqslant b,$$

$$(c) \Pi^k = 0 \text{ if } k \notin T_0.$$

For any $m \times m$ matrix Π, denote $g(\Pi) := (g(\Pi^1), ..., g(\Pi^m))'$, which is a $m \times 1$ vector.

(minimize L^0 norm) into a linear programming problem (minimize L^1 norm).

The exact sparse assumption of $\mathcal{P}$ is defined below:

There exists absolute positive constants K^u, r and K, and an absolute constant $c_{g,n}$ which may depend on n, such that:

(1) $s_0 := |T_0| = o(n \wedge m)$, which may increase at slow rate as m and n increases;

(2) The sum of coefficients needed to construct each inequality is bounded from the above:

$\max_{1 \leqslant j \leqslant m} ||\widetilde{\Pi}_j^*||_{L^1} \leqslant K d_1^r$;[5]

(3) $\max_{1 \leqslant j \leqslant m} g(\widetilde{\Pi}^{*j}) \leqslant K^u$;

(4) $\min_{j \in T_0} c_j \geqslant c_{g,n}$;

Exact sparse assumption assumes that the binding constraints are informative. Therefore, we are able to distinguish these constraints when the noise is small enough. Denote set $\mathcal{I}^*$ as the set of non-zero components in $g(\widetilde{\Pi}^*)$. In general set $\mathcal{I}^*$ is not necessarily the same as the minimum set of constraints, T_0. However, in the special case of Core Determining Class problem with bipartite graph, we show that $\mathcal{I}^* = T_0$. That is to say, the L^1 selector recovers the Core Determining Class when λ is set to be 0. We expect the set $\mathcal{I}^*$ should not be too large compared to T_0. We show that in the next section, similar to Candes and Tao (2005), the number of non-zero components has a order $O(s_0)$ with probability approaching 1 by employing a cutoff $0 < \eta < 1$ to $g(\Pi_0)$. For a q dimensional vector $\widetilde{b}$ and scalar λ, define $\widetilde{b} + \lambda := (\widetilde{b}_1 + \lambda, ..., \widetilde{b}_q + \lambda)$. Through out the paper, assume that M is a fixed matrix. Define $\mathscr{F} := \{\widetilde{Q}_{\mathcal{I}}(\widetilde{b}) | \widetilde{Q}_{\mathcal{I}}(\widetilde{b}) = \{v | M_{\mathcal{I}} v \leqslant \widetilde{b}_{\mathcal{I}}\}, \widetilde{b} \in \mathbb{R}^m$ and $\mathcal{I} \subset \{1, 2, ..., m\}\}$ as a collection of sets which takes the formulation $\{v | M_{\mathcal{I}} v \leqslant \widetilde{b}_{\mathcal{I}}\}$ for some $\widetilde{b} \in \mathbb{R}^m$ and $\mathcal{I} \subset \{1, 2, ..., m\}$. Define the operation $\oplus$ which maps a set $\widetilde{Q}_{\mathcal{I}}(\widetilde{b}) \in \mathscr{F}$ and a real number λ into another set $\widetilde{Q}_{\mathcal{I}}(\widetilde{b}) \oplus \lambda := \widetilde{Q}_{\mathcal{I}}(\widetilde{b} + \lambda) = \{v | M_{\mathcal{I}} v \leqslant b_{\mathcal{I}} + \lambda\}$. In the rest of the paper, let $\widetilde{Q}_{\mathcal{I}} \oplus \lambda$ be the abbreviation of $\widetilde{Q}_{\mathcal{I}}(\widetilde{b}) \oplus \lambda$ if there is no confusion.

By analogy with the exact sparse assumption, we propose a more feasible approximate sparse assumption:

[5] If $||M_i||_{l^2}$ is normalized to be 1, then in general $r = \frac{1}{2}$. In the Core-Determining Class problem, we prove that $K = 1$ and $r = 1$.

Definition 2.3.2 (Approximate Sparse) *Suppose we can order the separations* $c_1, ..., c_m$ *into* $c_{(1)} \geqslant c_{(2)} \geqslant ... \geqslant c_{(m)}$ *and suppose there exists a positive integer* s^* *such that:*

(1) $s^* = o(n \wedge m)$.

Let T^* *be the set of indices of the inequalities with the first* s^* *largest separations. Suppose* K *and* r *are absolute positive constants. Let* $\sigma^2 := \max_{1 \leqslant j \leqslant m} Var(b_{ij})^2$. *Let* $\widetilde{\Pi}^*$ *be the solution of the following problem:*

Problem $\mathcal{R}$:

$$\min_{\Pi} \sum_{k \in T^*} g(\Pi^k)$$

subject to:

$$(a)\Pi M \geqslant M, \Pi \geqslant 0,$$

$$(b)\Pi b \leqslant b + K d_1^r \sigma \sqrt{\frac{\log(s^*)}{n}},$$

$$(c)\Pi^k = 0 \text{ if } k \notin T^*.$$

Then, it holds that:

(2) $Q_{T^*} \subset Q \oplus K d_1^r \sqrt{\frac{\log(s^*)}{n}}$;

(3) There exists an absolute constant K^u *such that* $\max_{1 \leqslant j \leqslant m} g(\widetilde{\Pi}^{*j}) \leqslant K^u$;

(4)$\max_{1 \leqslant j \leqslant m} ||\widetilde{\Pi}_j^*||_1 \leqslant K d_1^r$.

In the approximate sparse assumption, we allow $c_j > 0$ for all $1 \leqslant j \leqslant m$. So in the worst case, $g(\widetilde{\Pi}^{*j}) > 0$ for all $j \in \{1, 2, ..., m\}$. The approximate sparse assumption assumes that there is a small set T^* indicating a feasible region similar to Q while the size of T^* is much smaller than m.

2.4 Properties of the Selection procedure $\hat{\mathcal{R}}$ with Application in the Core Determining Class Problem

2.4.1 General Properties

In this subsection, we analyze the property of the selection procedure $\hat{\mathcal{R}}$ and the choice of the relaxation parameter $\lambda_{n,m}$. We impose high level assumptions on $\hat{b}$ and $\lambda_{n,m}$ and then discuss a set of sufficient conditions for the assumption.

Assumption C.13 (Dominance of λ) *Suppose we have data $B_1, B_2, ..., B_n$ with dimension $m \times 1$ such that $b = \mathbb{E}[B_i]$, $1 \leqslant i \leqslant n$. Suppose in practice we use $\hat{b} := \mathbb{E}_n[B_i]$ to estimate b. Suppose that with probability at least $1 - \alpha$,*

(1) $\max_{1 \leqslant j \leqslant m} |\hat{b}_j - b_j| \leqslant \lambda_{n,m}$;

(2) $\lambda_{n,m} \to 0$.

In the Assumption C.13, we require that the choice of relaxation parameter $\lambda_{n,m}$ should dominate the maximal discrepancy between $\hat{b}_j$ and b_j for all $j \in \{1, 2, ..., m\}$. In additional, $\lambda_{n,m}$ should be converging to 0 as sample size increases to guarantee consistency.

Given $\lambda_{n,m}$, suppose that the solution to $\hat{\mathcal{R}}$ is $\hat{\Pi}$. Denote $\hat{g}_k := \max_{1 \leqslant j \leqslant k} \hat{\Pi}_{jk}$, for all $1 \leqslant k \leqslant m$. Define $\hat{\mathcal{I}} := \{k | \hat{g}_k \neq 0\}$ as the set selected by the procedure $\mathcal{R}$. We consider the post-selection estimator $\hat{Q}_{\hat{\mathcal{I}}} := \{v | M_{\hat{\mathcal{I}}} v \leqslant \hat{b}_{\hat{\mathcal{I}}}\}$ as the feasible set defined by the inequalities with indices in $\hat{\mathcal{I}}$.

Lemma 15 *If Assumption C.13 holds, then with probability $\geqslant 1 - \alpha$, $\hat{Q}_{\hat{\mathcal{I}}}$ satisfies:*

(1) $Q \subset \hat{Q}_{\hat{\mathcal{I}}} \oplus \lambda_{n,m}$.

(2) $\hat{Q}_{\hat{\mathcal{I}}} \subset Q \oplus 2\lambda_{n,m}$.

Lemma 15 shows that Q and $\hat{Q}$ are very close to each other. Therefore, $\lambda_{n,m}$ should be at least as large as the $(1-\alpha)$ quantile of the random variable

$$r_{n,m} := \max_{1 \leqslant j \leqslant m} |\hat{b}_j - b_j|.$$

Chernozhukov, Chetverikov and Kato (2013) (CCK later) shows that the distribution of $\sqrt{n} r_{n,m}$ can be well approximated by the distribution of the maxima of a Gaussian vector under certain conditions under $\frac{(\log m)^7}{n} \to 0$ along with other mild regularity conditions. The calculation can be easily performed via Gaussian Multiplier bootstrap. A weaker bound (but still relatively sharp in many cases) of the $(1-\alpha)$ quantile of $r_{n,m}$ could be obtained using modest deviation theory of self-normalized vectors described in De La Puna (2009), which requires $\frac{(\log m)^{(2+\delta)}}{n} \to 0$ where $\delta > 0$.

Assumption C.14 (Regularity Conditions) *(1) The data b_i is i.i.d..*[6]

(2) There exists an absolute constant $C > 0$ such that

$$\max_{1 \leqslant i \leqslant n, 1 \leqslant j \leqslant m} |b_{ij}| \leqslant C.$$

(3) There exist absolute positive constants c_1 such that

$$\min_{1 \leqslant j \leqslant m} \mathbb{E}[b_{ij}^2] \geqslant c_1.$$

The statement (2) in Assumption C.14 holds for the Core Determining Class problem with the constant $C = 1$. Statement (3) may not be true in the Core Determining Class problem when the dimension d_1 grows. However, the problem can be fixed by multiplying $\sqrt{d_1}$ to b_{ij} when we make the assumption that $\frac{c}{d_1} \leqslant v(u_i) \leqslant \frac{c'}{d_1}$ for some absolute positive constants c and c'. We use Assumption C.14 to derive properties for general selection procedure $\hat{\mathcal{R}}$. In Section 2.5.2, we apply $\hat{\mathcal{R}}$ to the Core Determining Class problem without Assumption C.14.

[6]The i.i.d. assumption can be extended to the i.n.i.d. assumption as Lemma 5 and Lemma 6 both allow i.n.i.d data with small modifications in the statement.

Under Assumption C.14, we are able to obtain the following two Lemmas on the choice of relaxation parameter λ. These two Lemmas are based on results stated in De La Puna (2009).

Lemma 16 (Choosing λ using Multiplier Bootstrap) *Let* $r_{n,d}^G := \max_{1\leqslant j\leqslant m} \frac{\sum_{1\leqslant i\leqslant n} B_{ij}e_{ij}}{n}$, *where* e_{ij} *are independent standard normal random variables. Suppose Assumption C.14 holds and* $\frac{\log(m\vee n)^7}{n} \to 0$, *then the* $1-\alpha$ *quantile of* $\sqrt{n}r_{n,d}^G$ *is a consistent estimator of the* $1-\alpha$ *quantile of* $\sqrt{n}r_{n,d}$.

Lemma 17 (Choosing λ using Modest Deviation Theory of Self-Normalized Vectors) *Denote* $\hat{\sigma}^2 := \max_{1\leqslant j\leqslant m}\{\mathbb{E}_n[B_{ij}^2] - [\mathbb{E}_n B_{ij}]^2\}$. *Let* $\lambda_{n,m} := \frac{C\hat{\sigma}\Phi^{-1}(1-\frac{\alpha}{2m})}{\sqrt{n}}$ *for some constant* $C > 1$. *Suppose Assumption C.14 holds and* $\frac{(\log m)^{2+\delta}}{n} \to 0$ *for some* $\delta > 0$, *then as* $n \to \infty$, *with probability at least* $1-\alpha$,

$$\max_{1\leqslant j\leqslant m} |\hat{b}_j - b_j| \leqslant \lambda_{n,m}$$

Next we discuss the performance of the L^1 Selector $\hat{g}$ under the sparse assumptions.

Theorem 5 (Recovery of Informative Inequalities under ES Assumption) *Suppose Assumptions C.13, C.14 and the exact sparse assumption hold. Recall that* c_j *is the maximal separation of the* j^{th} *inequality and* $c_{g,n} \leqslant c_j$ *for all* $j \in T_0$. *Let* $0 < \eta < 1$ *be an absolute constant. Assume that* m, n, s_0, d_1 *and* $c_{g,n}$ *obeys key growing conditions:*

$$\frac{(d_1^{2r}\log(s_0)) \vee \log(m)}{nc_{g,n}^2} \to 0.$$

Consider the following two step procedure:

(a) Step 1: Set $\lambda_S := (1+\epsilon)Kd_1^r\hat{\sigma}\sqrt{\frac{\log(\frac{4s_0}{\alpha})}{n}} + \lambda_{n,m}$, *with* $\epsilon > 0$ *be an absolute constant. Let* $\hat{g}_S$ *be the solution of* $\hat{\mathcal{R}}$ *with* $\lambda = \lambda_S$. *Let* $\hat{\mathcal{I}}_S := \{j|\hat{g}_{S,j} \neq 0\}$. *Set* $\lambda_{n,m}$ *to be chosen according to Lemma 16 or Lemma 17.*

(b) Step 2: With probability $\geqslant 1-\alpha$, *the set* $\hat{\mathcal{I}}_{S,\eta} := \{j|\hat{g}_S \geqslant \eta\}$ *has the following properties:*

(1) There exists an absolute constant C_T such that $||\hat{\mathcal{I}}_\eta||_0 \leqslant \frac{C_T s_0}{\eta}$;

(2) $\hat{\mathcal{I}}_\eta \supset T_0$;

(3) $\hat{Q}_{\hat{\mathcal{I}}_\eta} \subset Q \oplus \lambda_{n,m}$;

(4) $Q \subset \hat{Q}_{\hat{\mathcal{I}}_\eta} \oplus \lambda_{n,m}$.

In Theorem 5, we consider a two step procedure. First, we select the inequalities use a larger relaxation parameter λ_S. Such relaxation can significantly reduce the number of inequalities. However, as soon as λ_S converges to 0 fast enough, the informative inequalities will be preserved. The cutoff strategy additionally throws away some nearly redundant inequalities which were not detected in the first step selection. The set of those inequalities survive the two step procedure has good properties: (1) it has the same size compare to minimum set of inequalities, T_0, up to a constant multiplier; (2) it contains T_0 with probability increasing to one; (3) $\hat{Q}_{\mathcal{I}_\eta}$ is close to the true feasible region, $\{v|Mv \leqslant b\}$ with error up to $O(\lambda_{n,m})$.

Comment 2.4.1 *The constant K can be computed via M. If $||M_j||_2 = 1$ for all j and $M_{ij} > 0$ for all i, j, then $K \leqslant 1$ and $r = \frac{1}{2}$. In practice s_0 is unknown, so we recommend to use n for s_0 as starting value and then iterate a few times. We recommend to use $\epsilon = 0.1$ in practice.*

Theorem 6 (Recovery of Informative Inequalities under AS Assumption) *Suppose Assumptions C.13, C.14 and the approximate sparse assumption hold. Let $0 < \eta < 1$ be an absolute constant. Assume that m, n, s_0, d_1 and $c_{g,n}$ obeys key growing condition:*

$$\frac{(d_1^{2r} \log(s^*)) \vee \log(m)}{n c_{g,n}^2} \to 0.$$

Consider the following estimation procedure:

(a) Step 1: Set $\lambda_S := 2(1+\epsilon) K d_1^r \hat{\sigma} \sqrt{\frac{\log(\frac{4s^*}{\alpha})}{n}} + \lambda_{n,m}$, with $\epsilon > 0$ be an absolute constant. Let $\hat{g}_S$ be the solution of $\hat{\mathcal{R}}$ with $\lambda = \lambda_S$. Let $\hat{\mathcal{I}}_S := \{j|\hat{g}_{S,j} \neq 0\}$. Set $\lambda_{n,m}$ to be chosen according to Lemma 16 or Lemma 17.

(b) Step 2: With probability $\geqslant 1-\alpha$, *the set* $\hat{\mathcal{I}}_{S,\eta} := \{j|\hat{g}_S \geqslant \eta\}$ *has the following properties:*

(1) There exists an absolute constant C_T *such that* $||\hat{\mathcal{I}}_\eta||_0 \leqslant \frac{C_T s^*}{\eta}$;

(2) $\hat{Q}_{\hat{\mathcal{I}}_\eta} \subset Q \oplus \frac{\lambda_{n,m}+\lambda_S}{2}$;

(3) $Q \subset \hat{Q}_{\hat{\mathcal{I}}_\eta} \oplus \lambda_{n,m}$.

Again, in practice we can set $s^* = n$ as starting value and then iterate for a few times. If the approximate sparse assumption holds instead of the exact sparse assumption, the estimation procedure suffers from additional estimation error with size λ_S, which depends on the unknown parameter s^*.

2.4.2 Application in Estimating Measure v in Core Determining Class problem

To find the Core Determining Class given a bipartite graph $G = (\mathcal{U}, \mathcal{Y}, \varphi)$, we can use the method proposed in Section 3 to eliminate all the redundant inequalities and find exact solution when data noise is not taken into consideration. We can also use the L_1 selector proposed in Section 2.3.1 to find an approximate solution to the Core Determining Class problem. In addition, we can consider a hybrid method: first, we find the exact solution according to the method described in Section 2.1, and then apply the selection procedure presented in Section 2.3.1 using the inequalities selected from the previous step. The hybrid method may speed up the selection procedure significantly. In this subsection, we discuss the general selection procedure first, and then briefly discuss the hybrid method.

In the Core Determining Class problem, the equality $v(\mathcal{U}) = 1$ is never redundant. Therefore, we let the $(m-1)^{th}$ and m^{th} inequalities be $v(\mathcal{U}) \geqslant 1$ and $v(\mathcal{U}) \leqslant 1$ among the total m inequalities. Since there is no reason to drop the last two inequalities, we define problems $\mathcal{R}^C$ and $\hat{\mathcal{R}}^C$:

Problem $\mathcal{R}^C$:

$$\min_{\Pi} \sum_{k=1}^{m-2} \max_{1 \leqslant j \leqslant m-2} \Pi_{jk},$$

subject to:

$$(1)\Pi M \geqslant M, \Pi \geqslant 0,$$

$$(2)\Pi b \leqslant b,$$

and Problem $\hat{\mathcal{R}}^C$:

$$\min_{\Pi} \sum_{k=1}^{m-2} \max_{1 \leqslant j \leqslant m} \Pi_{jk},$$

subject to:

$$(1)\Pi M \geqslant M, \Pi \geqslant 0,$$

$$(2)\Pi \hat{b} \leqslant \hat{b} + \Lambda,$$

where $\Lambda := (\lambda_{n,m}, ..., \lambda_{n,m}, 0, 0)$ with $\lambda_{n,m}$ left to be chosen.

Let Π_0 be the solution to $\mathcal{R}^C$ and $\hat{\Pi}$ be the solution to $\hat{\mathcal{R}}^C$. First, we prove an important result specific to the Core Determining Class.

Lemma 18 (Perfect Recovery of the Minimum Model T_0) *If $\hat{\mu}_n$ is non-degenerate and $\lambda_{n,m} = 0$, then:*

(1) The L^0 norm of $\hat{g}$, $||\hat{g}||_0$, satisfies $||\hat{g}||_0 = s_0$;

(2) $\max_{1 \leqslant j \leqslant m-2} ||\hat{\Pi}_j||_1 \leqslant d_1$;

(3) $\max_{1 \leqslant j \leqslant m-2} \hat{g}(\hat{\Pi}^j) \leqslant 1$;

(4) The set of indices with non-zero entries of $\hat{g}$ satisfies:

$$\hat{\mathcal{I}} := \{k | \hat{g}_k \neq 0\} = T_0.$$

As a special case, $\mathcal{I}^* = T_0$.

Lemma 18 indicates that under the exact sparse assumption, the recovery of model T_0

could be done simply by looking at the non-zero entries of the solution of the problem $\hat{\mathcal{R}}$. Due to the special property presented in Lemma 18, we show that the relaxation parameter λ_S in Theorem 6 can be much tighter. Therefore, the selection procedure would require much less number of observations in order to achieve good performance.

Definition 2.4.1 (Approximate Sparse On Core Determining Class) *Suppose we can sort the separations $c_1, ..., c_m$ as $c_(1) \geqslant c_{(2)} \geqslant ... \geqslant c_{(m)}$ and there exists a positive integer s^*, such that:*

(1) $s^* = o(n \wedge m)$.

Let T^* be the set of indices of the inequalities with the first s^* largest separations. Suppose K and r are absolute positive constants. Let $\sigma^2 := \max_{1 \leqslant j \leqslant m} Var(B_i)^2$. Let $\widetilde{\Pi}^*$ be the solution of the following problem:

Problem $\mathcal{R}$:

$$\min_{\Pi} \sum_{k \in T^*} g(\Pi^k),$$

subject to:

$$(a)\Pi M \geqslant M, \Pi \geqslant 0,$$

$$(b)\Pi b \leqslant b + \sigma \sqrt{\frac{\log(s^*)}{n}},$$

$$(c)\Pi^k = 0 \text{ if } k \notin T^*.$$

(2) $Q_{T^*} \subset Q \oplus K d_1^r \sqrt{\frac{\log(s^*)}{n}}$.

(3) There exists an absolute constant K^u such that $\max_{1 \leqslant j \leqslant m} g(\Pi^{*j}) \leqslant K^u$.

Define $\hat{\sigma}^2 := \max_{1 \leqslant j \leqslant m} \mathbb{E}_n[B_i^2] - [\mathbb{E}_n B_i]^2$

Lemma 19 (Recovery of Informative Inequalities under Core Determining Class) *Suppose Assumptions C.13, C.14 and the exact sparse assumption hold. Suppose G and $\hat{\mu}_n$ are non-degenerate. Recall that c_j is the maximal separation of the j^{th} inequality and $c_{g,n} \leqslant c_j$ for all $j \in T^*$. Let $0 < \eta < 1$ be an absolute constant. Set*

$\lambda_S^C := (1+\epsilon)\hat{\sigma}\sqrt{\frac{\log(\frac{4s^*}{\alpha})}{n}}$, *where* $\epsilon > 0$ *is a constant. Assume that* s^* *and* $c_{g,n}$ *satisfy the key growing condition:*

$$\frac{\log(s^*)}{nc_{g,n}^2} \to 0.$$

Assume that with probability going to 1, *the empirical measure* $\hat{\mu}_n$ *obeys:*

$$\max_{1 \leqslant l \leqslant d_2} \frac{|\hat{\mu}_n(l) - \mu(l)|}{\mu(l)} \to 0.$$

Let $\hat{\Pi}$ *be the solution of* $\mathcal{R}^C$ *with* $\lambda_{m,n} = \lambda_S^C$. *Let* $\hat{g}_{S,k} := \max_{1 \leqslant j \leqslant m} \hat{\Pi}_{jk}$.

Then, with probability $\geqslant 1 - \alpha$, *the set* $\hat{\mathcal{I}}_{S,\eta} := \{j | \hat{g}_{S,k} \geqslant \eta\}$ *has the following properties:*

(1) There exists an absolute constant C_T *such that* $||\hat{\mathcal{I}}_\eta||_0 \leqslant \frac{C_T s^*}{\eta}$;

(2) $\hat{Q}_{\hat{\mathcal{I}}_\eta} \subset Q \oplus 2\lambda_S^C$;

(3) $Q \subset \hat{Q}_{\hat{\mathcal{I}}_\eta} \oplus \lambda_S^C$.

The value s^* *can be obtained iteratively, by setting* $s^* = n$ *as the initial value.*

The key assumption $\max_{1 \leqslant l \leqslant d_2} \frac{|\hat{\mu}_n(l) - \mu(l)|}{\mu(l)} \to 0$ mainly relies on the growing rate of d_2. When $\mu(l) = O(\frac{1}{d_2})$ and $\frac{d_2^3}{n} \to 0$, the assumption $\max_{1 \leqslant l \leqslant d_2} \frac{|\hat{\mu}_n(l) - \mu(l)|}{\mu(l)} \to 0$ holds. Lemma 19 obtains stronger results compare to Theorem 6 due to the structure of the bipartite graph.

It is natural to consider a hybrid estimation strategy combining the combinatorial method in Section 2.1 and the selection procedure in Section 2.3. There are a few points that we would like to make about the hybrid method:

(1) When s_0 is small, the hybrid method performs similarly to the combinatorial method only.

(2) When s_0 is large, there may be significant gains from the hybrid method in terms of computational speed compared to the selection procedure only, and significant inequality reduction compared to the combinatorial method only.

We illustrate these points in the Monte-Carlo experiments in the next section.

2.5 Monte-Carlo Experiments

Consider a simple setting in which many marginal firms are facing a volatile market. Let u be a random variable representing the cost of a firm. Let $\theta \in \{H, L\}$ be the private information of the firm which we do not observe. Let y be the action of the firm based on the information θ and cost u. Assuming the objective of firm is to maximize profit $\pi(y, u, \theta)$, they might adopt different actions when facing $\theta = H$ or $\theta = L$.

Suppose action y is the price set by the firm. We consider a simple case of decision making problem by the firms. Given observations of a sequence of decisions, we are interested in learning the distribution of the costs of these firms.

Assume that the profit function is

$$\pi(y, u, H) = (y - u)(C - y),$$

$$\pi(y, u, L) = (y - u)(C/2 - y),$$

where C is a constant.

If the firm consider any price $y^* \in \{y|\pi(y, u, \theta) \geqslant \max_y \pi(y, u, \theta) - w, a_1 \leqslant y \leqslant a_2\}$, where w is a constant for robust price control and a_1, a_2 are bounds on y, then $\varphi(u) := \{y|\pi(y, u, \theta) \geqslant \max_y \pi(y, u, \theta) - w, \theta \in \{H, L\}, a_1 \leqslant y \leqslant a_2\}$ is the correspondence mapping from the set of cost (event) $\mathcal{U}$ to the set of price (outcome) $\mathcal{Y}$.

We can only observe $\hat{y}$, the empirical measures on price $\mathcal{Y}$. The objective is to find an approximate feasible set of probability measure on cost $\mathcal{U}$. Assume that u is i.i.d. across observations.

Example 10 (Monte-Carlo Experiment 1) *Set $\mathcal{U}$, $\mathcal{Y}$, C and w as follows:*

$C = 4$, $\mathcal{U} = [0, 3]$, $\mathcal{Y} = [1, 3.5]$, $w = 0.01$.

So $\varphi(u) = [(1.9 + u/2), (2.1 + u/2) \wedge 3.5] \cup [(0.9 + u/2 \vee 1), (1.1 + u/2)]$.

To estimate the probability measure on cost $\mathcal{U}$, we discretize the continuous set of cost (event) $\mathcal{U}$ and price (outcome) $\mathcal{Y}$. Let $d_1 = 15$ and $d_2 = 25$ be the number of discretized segments of cost and price, respectively. Then $u_i = ((i-1)/5, i/5)$ and $y_j = ((j-1)/10+1, j/10+1)$ for $i = 1, 2, ..., d_1$ and $j = 1, 2, ..., d_2$. The correspondence mapping φ_d from the discretized set $\mathcal{U}_d = \{u_i | i = 1, 2, ..., 15\}$ to the discretized set $\mathcal{Y}_d = \{y_j | j = 1, 2, ..., 25\}$ is generated by:

$$\varphi_d(u_i) = \{y_j | y_j \cap \varphi(u_i) \neq \emptyset\}$$

Therefore, $\varphi(u_1) = \{y_1, y_2, y_{10}, y_{11}, y_{12}\}$, $\varphi(u_{15}) = \{y_{14}, y_{15}, y_{16}, y_{24}, y_{25}\}$. For any $2 \leqslant i \leqslant 14$, $\varphi(u_i) = \{y_{i-1}, y_i, y_{i+1}, y_{i+9}, y_{i+10}, y_{i+11}\}$. Figure 2-5 illustrates the correspondence mapping for Example 10.

Suppose μ, the true probability measure on $\mathcal{Y}$, follows the formula $\mu(j) \propto \max(1, |j-13|^{1.5})$ for $1 \leqslant j \leqslant 25$. Suppose the sample size n (the number of observed y) is 2000 and 500 and the sample y is randomly drawn according to measure μ. Let $\hat{\mu}_n$ be the empirical measure (observed frequency) of y. Let $\hat{\sigma}^2 = \max_{1 \leqslant j \leqslant 25} \hat{\mu}_n(j) - \hat{\mu}_n(j)^2$ and $\alpha = 0.05$. According to Lemma 17, the penalty term $\lambda_{n,m} = 1.05 \min\{\hat{\sigma}\sqrt{\frac{\log(m/\alpha)}{n}}, \frac{1}{2}\frac{\log(m/\alpha)}{n}\}$, where m is the number of inequalities selected in the algorithm describe in Section 2.1.

Problem $\hat{\mathcal{R}}$ is implemented to further select the inequalities. The results of a set of Monte-Carlo experiments with 100 repetitions are presented in Table 2.1. For each instance, we apply a cut-off value η to the optimal L^1 coefficient $g(\Pi_i)$: select an inequality if the corresponding $g(\Pi_i) \geqslant \eta$ and discard it otherwise. We present the average, maximum and minimum number of selected inequalities with cut-off value $\eta = 0, 0.1$ and 0.2.

A critical concept concerning the selection performance is "coverage": in one instance, the feasible set (of the probability measure) on $\mathcal{U}$ corresponding to the true μ is subset of the feasible set (of the probability measure) on $\mathcal{U}$ defined by the selected inequalities with empirical measure $\hat{\mu}_n$. We present the "frequency of coverage" corresponding to different cut-off value η. As η increases, the procedure selects fewer inequalities, so the approximate feasible set will become larger, which is more likely to "cover" the true feasible set and produce a larger "frequency of coverage". In the numerical experiments,

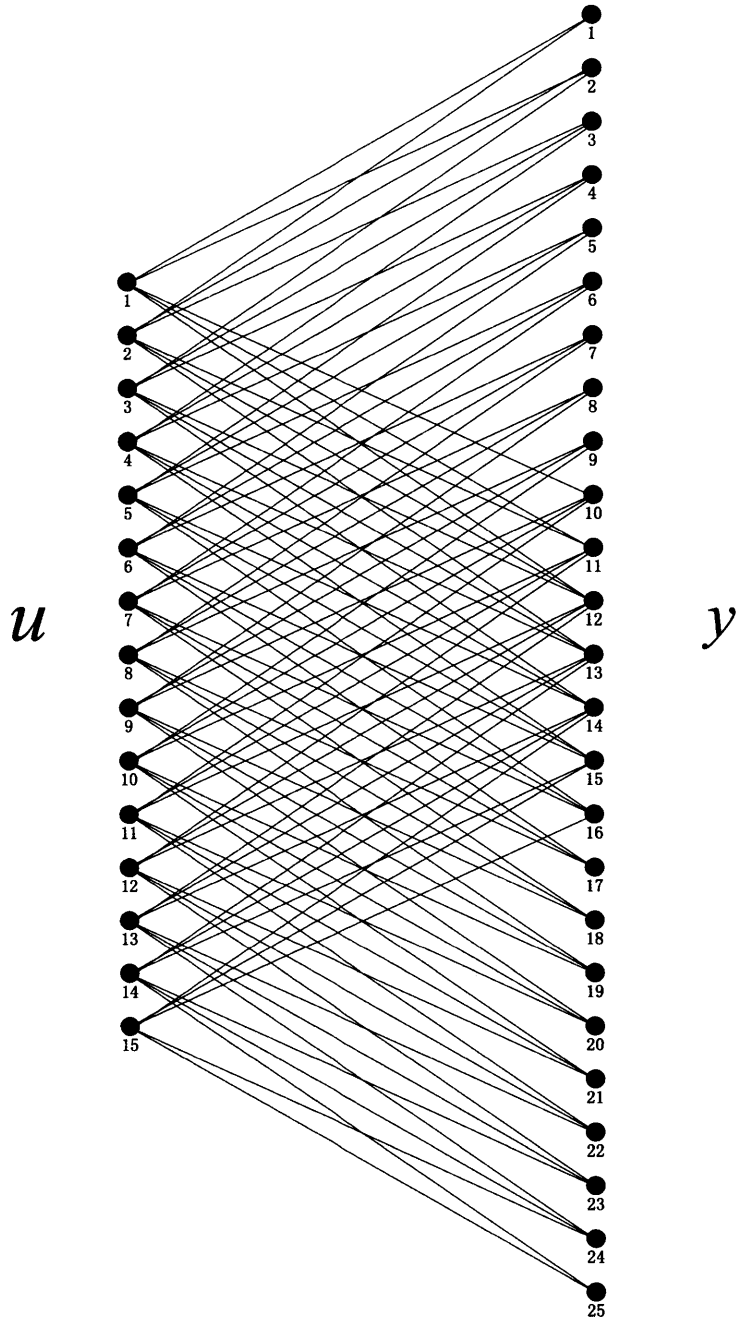

Figure 2-3: Correspondence Mapping for Example 10

the "frequency of coverage" is greater than 95% when the cut-off value $\eta = 0$ (essentially

Number of experiments (M)	100	
Number of events $\times$ number of outcomes ($d_1 \times d_2$)	15×25	
Number of inequalities in true model	471	
Conservative bound of acceptance rate ($1-\alpha$)	0.95	
Sample size (n)	500	2000
Average λ	0.0710	0.0355
Frequency of Coverage ($\eta = 0$)	97%	99%
Avg. number of inequalities selected ($\eta = 0$)	184.66	187.42
Max. number of inequalities selected ($\eta = 0$)	241	234
Min. number of inequalities selected ($\eta = 0$)	145	92
Frequency of Coverage ($\eta = 0.1$)	99%	100%
Avg. number of inequalities selected ($\eta = 0.1$)	32.59	86.02
Max. number of inequalities selected ($\eta = 0.1$)	43	145
Min. number of inequalities selected ($\eta = 0.1$)	27	27
Frequency of Coverage ($\eta = 0.2$)	99%	100%
Avg. number of inequalities selected ($\eta = 0.2$)	26.73	56.69
Max. number of inequalities selected ($\eta = 0.2$)	28	108
Min. number of inequalities selected ($\eta = 0.2$)	24	27
Running time (sec/instance)	87	146

Table 2.1: Results of Monte-Carlo Experiments on Example 10

the case of no cut-off). It agrees with the parameter selection $\alpha = 0.05$ (type 1 error) in the formula of the penalty term λ described in Lemma 17.

We compare the inequalities selection of the integer programming L^0 procedure with the linear programming L^1 procedure. Figure 2.4 illustrates the comparisons with respect to the magnitude of the L^1 selector coefficient $g(\Pi_i)$ in the optimal solution of problem $\hat{\mathcal{R}}$. Figure 2.5 illustrates the comparisons with respect to the separation of each inequality, which is

$$c(A) := \max\{v(A) - \mu(\varphi(A)) | v(A') \leqslant \mu(\varphi(A')), \forall A' \subset \mathcal{S}_u^*, A' \neq A\}$$

Table 2.2 presents the detailed selection results. It can be seen that the L^0 selector (the model to select minimum number of inequalities) is recovered by the L^1 selector to a large extent, while the L^1 selector enjoys extremely high computational advantage. Generally, inequalities selected by the L^0 selector have comparatively large L^1 coefficients $g(\Pi_i)$,

Number of inequalities selected in L^0	79
Number of inequalities selected in L^1	211
Number of inequalities that L^0 model selected in L^1, $\eta = 0$	79
Number of inequalities that L^0 model selected in L^1, $\eta = 0.05$	78
Number of inequalities that L^0 model selected in L^1, $\eta = 0.10$	78
Number of inequalities that L^0 model selected in L^1, $\eta = 0.15$	77
Number of inequalities that L^0 model selected in L^1, $\eta = 0.20$	72
Running time of L^0 model (min)	2195
Running time of L^1 model (min)	1.45

Table 2.2: Comparisons of L^0 and L^1

which makes it easy to be selected by the L^1 selector under a reasonable cut-off value η. In addition, the L^1 selector is able to successfully differentiate inequalities with close separation values but opposite L^0 coefficients.

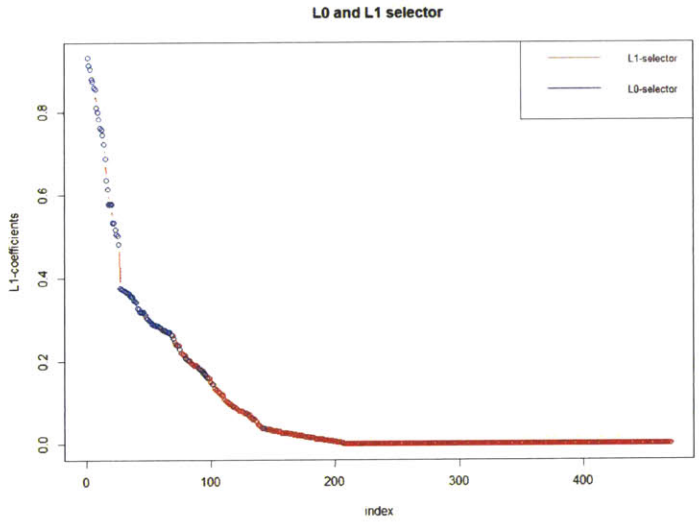

Figure 2-4: L^0 versus L^1: with respect to L^1 Coefficient

We project our L^1 estimator compared to L^0 and the true feasible set onto v_1, v_2, v_3, a three-dimension subspace. Figure 2.6 compares the performance of L^0 and L^1 selectors: the L^1 selector with $\eta = 0.1$ is slightly more conservative than the L^0 selector. Figure 2.7 shows that the L^1 selector covers the true feasible set by a small margin.

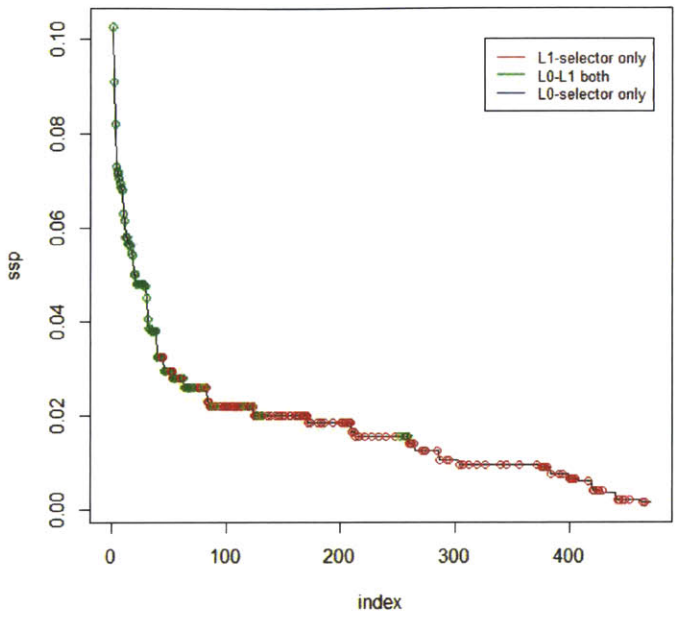

Figure 2-5: L^0 versus L^1: with respect to Inequality Separation

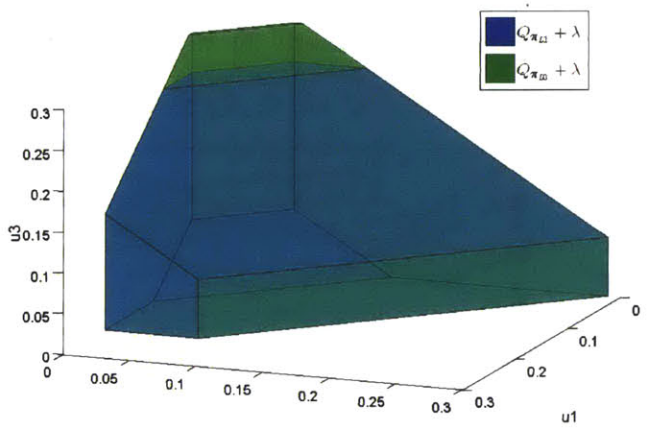

Figure 2-6: L^0 versus L^1: Projection onto v_1, v_2, v_3.

We also demonstrate through a smaller example the sharpness of the relaxing parameter λ. If (1) the empirical measure $\hat{\mu}_n$ is largely mis-specified from the true measure μ, and (2) the true feasible set (of the probability measure) on $\mathcal{U}$ is still a subset of

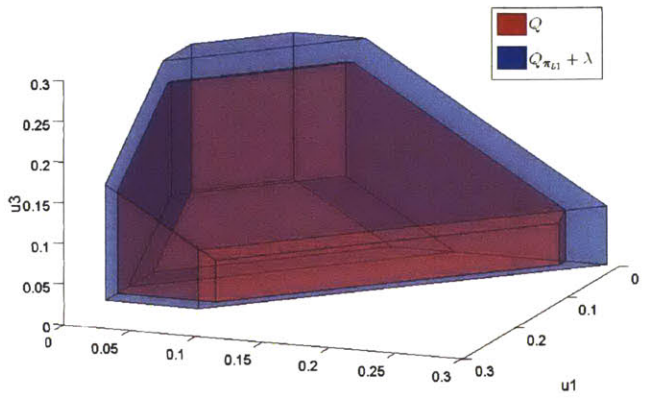

Figure 2-7: L^1 versus True Feasible Set: Projection onto v_1, v_2, v_3.

the approximate feasible set (of the probability measure) on $\mathcal{U}$ obtained in the selection procedure, then a type 2 error occurs. The λ implied by α limits the magnitude of the type 1 error, and a sharp λ will also limit the occurrences of type 2 errors at the same time. In another set of Monte-Carlo experiments below, we examine the type 2 error in the case that the empirical measure $\hat{\mu}_n$ is locally mis-specified.

Example 11 (Monte-Carlo Experiment 2) *Figure 2-8 is the correspondence mapping for an example with size of* 7×7.

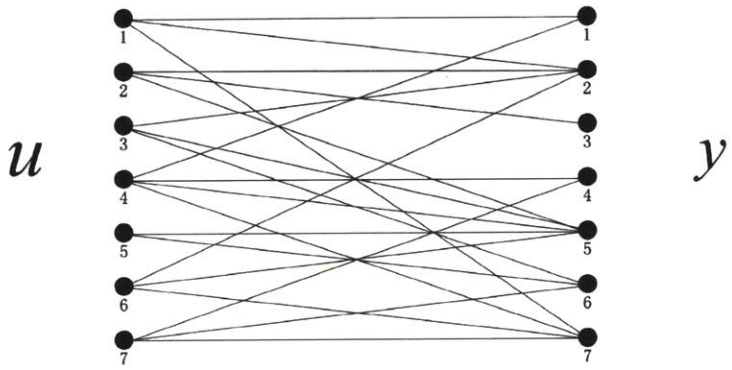

Figure 2-8: Correspondence Mapping for Example 11

Assuming the true probability measure μ *on* $\mathcal{Y}$ *is* $(0.1, 0.25, 0.2, 0.1, 0.1, 0.2, 0.05)$, *we perturb* $\mu_{n,0}$ *with* $\gamma(a_1, a_2, ..., a_7)/\sqrt{n}$, *where* a_i *is randomly and uniformly drawn from*

$M = 10000, n = 200$	$\gamma = 1$	$\gamma = 2$	$\gamma = 3$	$\gamma = 4$	$\gamma = 5$
Type 1 error, $\eta = 0$	3.64%	3.73%	3.84%	3.85%	4.21%
Type 1 error, $\eta = 0.1$	3.56%	3.55%	3.71%	3.77%	4.07%
Type 2 error, $\eta = 0$	29.65%	11.18%	6.98%	5.86%	4.58%
Type 2 error, $\eta = 0.1$	29.88%	11.35%	7.08%	6.04%	4.73%

Table 2.3: Type 1 and Type 2 Errors

$\{-1, 1\}$, $1 \leqslant i \leqslant 7$. *So the empirical measure* $\hat{\mu} \propto \mu + \gamma(a_1, a_2, ..., a_7)/\sqrt{n}$ *in the case of mis-specified perturbation.*

We run 10000 instances for each setting of perturbation γ and cut-off value η. Table 2.3 presents the type 1 and type 2 error for each setting. The results show that, while the type 1 error is less than 0.05 as designed, the type 2 error is also relatively small, which means the approximate feasible set of probability measure on $\mathcal{U}$ does not over exaggerate the true feasible set. λ is sharp and the model has strong power against local alternatives.

2.6 Conclusion

In this paper we consider estimating the probability measure on the unobservable events given observations on the outcomes. We try to select the set of minimum number of inequalities, which is called the Core Determining Class, to describe the feasible set of target probability measure. We propose a procedure to construct the exact Core Determining Class when data noise are not taken into consideration. We prove that, if there is no degeneracy, the Core Determining Class only depends on the structure of the bipartite Graph, not the probability measure μ on the outcomes.

For a general problem of linear inequalities selection under noise, we propose a selection procedure similar to the Dantzig selector. A formulation is proposed to identify the importance of each inequality in a feasible set defined by many inequalities constraints. We describe the exact sparse assumptions and approximate sparse assumptions, which are are similar to the traditional sparse assumptions in a linear regression environment.

We prove that the selection procedure has good statistical properties under the sparse assumptions.

We apply the selection procedure to the Core Determining Class problem and develop a hybrid selection method combined with a combinatorial algorithm. We prove that the hybrid selection procedure has better statistical properties due to the structure of the graph.

We demonstrate the good performance of our selection procedure through several set of Monte-Carlo experiments. First, the inference based on the selection procedure has desired size; second, it has power against local alternatives; third, it is relatively computationally efficient.

2.7 Reference

Ackerberg, D., Lanier Benkard, C., Berry, S., Pakes, A. (2007). Econometric tools for analyzing market outcomes. Handbook of econometrics, 6, 4171-4276.

Andrews, D. W., Shi, X. (2013). Inference based on conditional moment inequalities. Econometrica, 81(2), 609-666.

Andrews, D. W., Soares, G. (2010). Inference for parameters defined by moment inequalities using generalized moment selection. Econometrica, 78(1), 119-157.

Artstein, Z. (1983). Distributions of random sets and random selections. Israel Journal of Mathematics, 46(4), 313-324.

Bajari, P., Benkard, C. L., Levin, J. (2007). Estimating dynamic models of imperfect competition. Econometrica, 75(5), 1331-1370.

Bajari, P., Hong, H., Ryan, S. P. (2010). Identification and estimation of a discrete game of complete information. Econometrica, 78(5), 1529-1568.

Belloni, A., Freund, R. M. (2008). On the symmetry function of a convex set. Mathematical Programming, 111(1-2), 57-93.

Beresteanu, A., Molchanov, I., Molinari, F. (2011). Sharp identification regions in models with convex moment predictions. Econometrica, 79(6), 1785-1821.

Candes, E., Tao, T. (2007). The Dantzig selector: Statistical estimation when p is much larger than n. The Annals of Statistics, 2313-2351.

Caron, R. J., McDonald, J. F., Ponic, C. M. (1989). A degenerate extreme point strategy for the classification of linear constraints as redundant or necessary. Journal of Optimization Theory and Applications, 62(2), 225-237.

Chernozhukov, V., D., Chetverikov and K., Kato(2014). Central Limit Theory and multiplier bootstrap when p is much larger than n. The Annals of Statistics.

Chesher, A., Rosen, A. (2012). Simultaneous equations models for discrete outcomes: coherence, completeness, and identification, cemmap working paper CWP21/12.

Eaves, B. C., Freund, R. M. (1982). Optimal scaling of balls and polyhedra. Mathematical Programming, 23(1), 138-147.

Galichon, A., Henry, M. (2006). Inference in incomplete models. Available at SSRN 886907.

Galichon, A., Henry, M. (2011). Set identification in models with multiple equilibria. The Review of Economic Studies, 78(4), 1264-1298.

Jovanovic, B. (1989). Observable implications of models with multiple equilibria. Econometrica: Journal of the Econometric Society, 1431-1437.

Manski, C. F., Tamer, E. (2002). Inference on regressions with interval data on a regressor or outcome. Econometrica, 70(2), 519-546.

Paulraj, S., Chellappan, C., Natesan, T. R. (2006). A heuristic approach for identification of redundant constraints in linear programming models.International Journal

of Computer Mathematics, 83(8-9), 675-683.

Puna, V., T, Lai and Qi, Shao(2009). Self-Normalized Processes: Limit Theory and Statistical Applications. Springer.

Romano, J. P., Shaikh, A. M. (2010). Inference for the identified set in partially identified econometric models. Econometrica, 78(1), 169-211.

Telgen, J. (1983). Identifying redundant constraints and implicit equalities in systems of linear constraints. Management Science, 29(10), 1209-1222.

Chapter 3

Summarizing Partial Effects beyond Averages

In nonlinear models the effects of interest are often functions of the parameters and data. For example, the conditional choice probabilities and the marginal or partial effects derived from them in the probit model $\mathrm{P}(Y = 1|X) = \Phi(X'\beta)$ depend on the probit coefficient β and covariates X, where Φ is the standard normal distribution. A common empirical practice is to report the average effect as a single summary measure of the heterogeneity in the effects (e.g., Wooldridge (2010, Chap 2)). In this paper we propose to complement this measure by reporting multiple effects sorted in increasing order and indexed by a ranking index with respect to the distribution of the covariates in the part of the population of interest. These sorted effects correspond to quantiles of the effects and provide a more complete representation of the heterogeneity of the model. In the probit model, for example, the values of $\Phi(X'\beta)$ sorted in increasing order with respect to the distribution of X correspond to the quantiles of the conditional choice probabilities over the entire population. We name these effects as sorted predictive effects (SPE). We use predictive effects (PE) instead of treatment, partial or marginal effects because we do not take a stand on whether the source model is descriptive or structural.

Let X denote a covariate vector, $\Delta(X)$ denote the PE of interest, which might be

parametrically or nonparametrically specified, $\mu(X)$ denote the probability measure of X in the part of the population of interest, and $\mathcal{X}$ denote the support of μ. The SPE are obtained by sorting the multivariate function $x \mapsto \Delta(x)$ in increasing order with respect to μ. We show in this paper that this multivariate sorting operator is Hadamard differentiable with respect to the PE function Δ and the probability measure μ at the regular values of $x \mapsto \Delta(x)$ on $\mathcal{X}$. This result allows us to derive the large sample properties of the empirical SPE, which replace Δ and μ by sample analogs, using the delta method. In particular, we derive a functional central limit theorem and a bootstrap functional central limit theorem for the empirical SPE. The main requirement of these theorems is that the empirical Δ and μ also satisfy functional central limit theorems, which hold for most estimators used in empirical economics under general sampling conditions. We use the properties of the empirical SPE to construct confidence sets for the SPE that hold uniformly over quantile indexes.

We illustrate the results of the paper with numerical simulations, and an empirical application to the effect of fertility on female labor supply following Angrist and Evans (1998) and Angrist (2001). The numerical simulations show that the large sample properties provide a good approximation to the behavior of the empirical SPE for sample sizes that are relevant for practice. The empirical application uses U.S 1980 Census data to show that there is substantial heterogeneity in the effect of fertility on both the extensive and the intensive margins of the labor supply of married women. Thus, among women with at least two children, the negative effect of having a third child on the probability of labor force participation ranges between 10% and 17%, and the negative effect on the number of weeks worked ranges between 0 and 19 weeks for working women. We also found important heterogeneity within subpopulations defined by the fertility third child indicator, education and husband's income.

Related literature: We extend the analysis of Chernozhukov, Fernandez-Val and Galichon (2010) for the univariate rearrangement (sorting) operator to the multivariate case. The multivariate case requires different techniques than the univariate case because the topological structure of the set of regular values of the PE function $x \mapsto \Delta(x)$ becomes more complex with the dimension of x. In particular, we show that under some

regularity conditions this set is a $(d_x - 1)$-manifold in $\mathbb{R}^{d_x}$, where d_x is the dimension of X. The manifold of dimension 0 when $d_x = 1$ is a finite number of points. Sasaki (2014) used a similar topological analysis to characterize the properties of derivatives of conditional quantiles in nonseparable models.

Organization of the paper: In section 3.1 we discuss the quantities of interest in nonlinear models with examples, and introduce the SPE. In section 3.2 we characterize the analytical properties of the multivariate sorting operator. In section 3.3 we derive the properties of the empirical SPE in large samples and show how to use these properties to make inference on the SPE uniformly over quantile indexes. In section 3.4 we discuss how to incorporate discrete covariates in the PE. In section 3.5 we provide numerical simulation and empirical results. We give a summary of the main results and conclude in Section 3.6. We gather the proofs of the main results in the Appendix.

Notation: For a random variable X, $\mathcal{X}$ denotes the part of the support of X of interest, $\mu(x)$ denotes the probability measure of X over $\mathcal{X}$, and $\widehat{\mu}(x)$ denotes the empirical probability of X over $\mathcal{X}$. We denote the expectation with respect to the measure $\widetilde{\mu}$ by $\mathbb{E}_{\widetilde{\mu}}$. We denote the PE as $\Delta(x)$ and the empirical PE as $\widehat{\Delta}(x)$. We denote $\nabla\Delta(x) := \partial\Delta(x)/\partial x$, the gradient of $x \mapsto \Delta(x)$. We also use $a \wedge b$ to denote the minimum of a and b. For a vector $v = (v_1, \ldots, v_{d_v}) \in \mathbb{R}^{d_v}$, $\|v\|$ denotes the Euclidian norm of v, that is $\|v\| = \sqrt{v'v}$, where the superscript $'$ denotes transpose. For a non-negative integer r and an open set $\mathcal{K}$, the class $\mathcal{C}^r$ on $\mathcal{K}$ includes the set of r times continuously differentiable real valued functions on $\mathcal{K}$. The symbol $\rightsquigarrow$ denotes weak convergence (convergence in distribution), and $\to_{\mathrm{P}}$ denotes convergence in probability.

3.1 Sorted Effects in Nonlinear Models

We discuss the objects of interest in nonlinear models and introduce the sorted effects.

3.1.1 Effects of Interest in Nonlinear Models

We consider a general nonlinear model characterized by a predictive function $g(X)$, where X is a d_x-vector of covariates that can be discrete or continuous, and g can be parametrically or nonparametrically specified. The vector X might include unobservable components such as unobserved individual heterogeneity or control variables. The function g usually arises from a model for a response variable Y, which can be discrete or continuous. We call the function g predictive because the underlying model can be either descriptive or structural. For example, in a mean regression model, $g(X) = \mathbb{E}[Y|X]$ corresponds to the expectation function of Y conditional on X; in a binary choice model, $g(X) = \mathrm{P}[Y = 1|X]$ corresponds to the choice probability of $Y = 1$ conditional on X; in a quantile regression model, $g(X) = \mathbb{Q}_Y[\tau|X]$ corresponds to the τ-quantile function of Y conditional on X; and in a structural model, $Y = g(X) + \varepsilon$ where $\mathbb{E}[\varepsilon|Z] = 0$ and Z is a vector of instrumental variables, g corresponds to a structural mean function of Y conditional on X. In the conditional quantile model the function g should be indexed by the quantile index τ, but we omit this dependence to lighten the notation.

Let $X = (T, W)$, where T is the covariate or treatment of interest, and W is a vector of control variables. We are interested in the effects of changes in the variable T on the function g. These effects are usually called partial effects, marginal effects, or treatment effects. We name them as predictive effects (PE) instead to emphasize that the function g might or might not have a structural or causal interpretation. If T is discrete, the PE is

$$\Delta(x) = \Delta(t, w) = g(t_1, w) - g(t_0, w) \tag{3.1.1}$$

where t_1 and t_0 are two values of T that might depend on t (e.g., $t_0 = t$ and $t_1 = t + 1$). This PE measures the effect of changing T from t_0 to t_1 holding W constant at w. If T is continuous, the PE is

$$\Delta(x) = \Delta(t, w) = \partial_t g(t, w), \tag{3.1.2}$$

where ∂_t denotes $\partial/\partial t$, the partial derivative with respect to t. This PE measures the effect of a marginal change of T from the level t holding W constant at w.

We consider the following examples in the empirical application of Section 3.5.

Example 12 (Probit model) *Let Y be a binary response variable such as a female labor force participation indicator, and X be a vector of covariates related to Y. The structural function of the probit model is*

$$g(X) = \mathrm{P}(Y = 1|X) = \Phi(P(X)'\beta),$$

where $P(X)$ is a vector of known transformations of X, β is a parameter vector, and Φ is the distribution of the standard normal. If T is a binary variable such as a an indicator for having 3 children and W is a vector of women characteristics, the PE

$$\Delta(x) = \Phi(P(1, w)'\beta) - \Phi(P(0, w)'\beta)$$

measures the effect of having a third child on the probability of participation for a woman with characteristics $W = w$, i.e. the effect on the extensive margin of the labor supply.

Example 13 (Tobit model) *Let Y be a nonnegative response variable such as female labor supply, and X a vector of covariates related to Y. The structural function of the tobit type I model is*

$$g(X) = \mathbb{E}[Y|X, Y > 0] = P(X)'\beta + \sigma\lambda(P(X)'\beta/\sigma),$$

where $P(X)$ is a vector of known transformations of X, (β, σ) is a parameter vector, Φ is the distribution of the standard normal, and $\lambda(z) = \partial_z\Phi(z)/\Phi(z)$ is the inverse Mills ratio. If T is a binary variable such as a an indicator for having 3 children and W is a vector of women characteristics, the PE

$$\Delta(x) = [P(1, w)'\beta + \sigma\lambda(P(1, w)'\beta/\sigma)] - [P(0, w)'\beta + \sigma\lambda(P(0, w)'\beta/\sigma)]$$

measures the effect of having a third child on the labor supply for a woman with characteristics $W = w$ who is working, i.e. the effect on the intensive margin of labor supply.

3.1.2 Sorted Effects

In Examples 1–2, the PE $\Delta(x)$ is a function of x and therefore can be different for each women. To summarize this effect in a single measure, a common practice in empirical economics is to average the PE

$$\mathbb{E}_\mu[\Delta(X)] = \int \Delta(x)\mu(x),$$

where μ is the distribution of X in the part of the population of interest. For example, when μ is the distribution on the entire population we obtain the average predictive effect (APE); whereas if μ is the distribution on a group characterized by X taking values in some specified set, we obtain a conditional average predictive effect (CAPE). Averaging, however, masks most the heterogeneity in the PE allowed by a nonlinear model.

We propose reporting multiple values of the PE sorted in increasing order and indexed by a ranking index $u \in [0, 1]$ with respect to the population of interest. These sorted effects provide a more complete representation of the heterogeneity in the PE than the average effects.

Definition 3.1.1 (u-SPE) *The u-sorted partial effect with respect to μ is*

$$\Delta^*_\mu(u) := \inf_{\delta \in \mathbb{R}}\{F_{\Delta,\mu}(\delta) \geqslant u\}, \quad u \in [0, 1],$$

where $F_{\Delta,\mu}$ denotes the distribution function of $\Delta(X)$ with respect to the probability measure μ for X.

The u-SPE is the u^{th}-quantile of the PE $\Delta(X)$ when X is distributed according to μ. As for the average effect, μ can be chosen to select the part of the population of interest. For example, if T is a treatment indicator and the PE is defined as in (3.1.2) with $t_0 = 0$ and $t_1 = 1$, the u-SPE corresponds to the u-quantile of the PE distribution when μ is the distribution of X in the entire population, or to the u-quantile of the PE distribution of the treated when μ is the distribution of X conditional on $T = 1$.

By considering $\Delta^*_\mu(u)$ at multiple indexes u, we obtain a one-dimensional represen-

tation of the heterogeneity of the PE. Accordingly, our object of interest is the SPE-function

$$\{u \mapsto \Delta_\mu^*(u) : u \in \mathcal{U}\}, \quad \mathcal{U} \subset (0,1),$$

where $\mathcal{U}$ is the set of indexes of interest. For example, in the empirical application of Section 3.5 we find substantial heterogeneity in the SPE-function of fertility on the extensive and intensive margins of the female labor supply, which is missed by traditional empirical analysis that only reports average effects.

3.2 Sorting Multivariate Functions: Analytical Properties

To analyze the analytical properties of the SPE-function, it is convenient to treat the PE as a multivariate real-valued function $\Delta : \mathcal{X} \to \mathbb{R}$, where $\mathcal{X} \subseteq \mathbb{R}^{d_x}$. Let μ be a probability measure on $\mathcal{X}$. The distribution of Δ with respect to μ is the function $F_{\Delta,\mu} : \mathbb{R} \to [0,1]$ with

$$F_{\Delta,\mu}(\delta) = \mu(\Delta \leqslant \delta) = \int_{\mathcal{X}} 1\{\Delta(x) \leqslant \delta\}\mu(x). \tag{3.2.1}$$

The SPE-function is

$$\Delta_\mu^* : \mathcal{U} \subset [0,1] \to \mathbb{R},$$

defined at each point as the left-inverse function of $F_{\Delta,\mu}$, i.e.,

$$\Delta_\mu^*(u) := F_{\Delta,\mu}^{\leftarrow}(u) = \inf_{\delta \in \mathbb{R}} \{F_{\Delta,\mu}(\delta) \geqslant u\}. \tag{3.2.2}$$

From this functional perspective, $u \mapsto \Delta_\mu^*(u)$ is the result of applying to $x \mapsto \Delta(x)$ an operator that sorts values in increasing order weighted by μ. In this section we characterize some analytical properties of the distribution function $\delta \mapsto F_{\Delta,\mu}(\delta)$ and the sorted function $u \mapsto \Delta_\mu^*(u)$, and derive the functional derivatives of $F_{\Delta,\mu}$ and Δ_μ^* with respect to Δ and μ.

3.2.1 Background on Differential Geometry

We recall some definitions from differential geometry that are used in the analysis. For a continuously differentiable function $\Delta : \mathcal{X} \to \mathbb{R}$ defined on an open set $\mathcal{X} \subseteq \mathbb{R}^{d_x}$, $x \in \mathcal{X}$ is a *critical point* of Δ, if

$$\nabla\Delta(x) = 0, \tag{3.2.3}$$

where $\nabla\Delta(x)$ is the gradient of $\Delta(x)$, otherwise x is a *regular point* of Δ. A value δ is a *critical value* of Δ on $\mathcal{X}$, if the set $\{x \in \mathcal{X} : \Delta(x) = \delta\}$ contains one critical point. Otherwise δ is a *regular value* of Δ on $\mathcal{X}$. In the multi-dimensional space, $d_x > 1$, a function Δ can have continuums of critical points. For example, the function $\Delta(x_1, x_2) = \cos(x_1^2 + x_2^2)$ has infinitely many critical points on the circle $x_1^2 + x_2^2 = k\pi$ for every non-negative integer k.

We recall now several core concepts related to manifolds from Spivak (1965) and Munkres (1990).

Definition 3.2.1 (Manifold) *Let d_k, d_x and r be positive integers such that $d_x \geqslant d_k$. Suppose that $\mathcal{M}$ is a subspace of $\mathbb{R}^{d_x}$ that satisfies the following property: for each point $m \in \mathcal{M}$, there is a set $\mathcal{V}$ containing m that is open in $\mathcal{M}$, a set $\mathcal{K}$ that is open in $\mathbb{R}^{d_k}$, and a continuous map $\alpha_m : \mathcal{K} \to \mathcal{V}$ carrying $\mathcal{K}$ onto $\mathcal{V}$ in a one-to-one fashion, such that: (1) α_m is of class $\mathcal{C}^r$ on $\mathcal{K}$, (2) $\alpha_m^{-1} : \mathcal{V} \to \mathcal{K}$ is continuous, (3) the Jacobian matrix of α_m, $D\alpha_m(k)$, has rank d_k for each $k \in \mathcal{K}$. Then $\mathcal{M}$ is called a d_k-manifold without boundary in $\mathbb{R}^{d_x}$ of class $\mathcal{C}^r$. The map α_m is called a coordinate patch on $\mathcal{M}$ about m. A set of coordinate patches that covers $\mathcal{M}$ is called an atlas.*

Definition 3.2.2 (Connected Branch) *For any subset $\mathcal{M}$ of a topological space, if any two points m_1 and m_2 can not be connected via path in $\mathcal{M}$, then we say that m_1 and m_2 are not connected. Otherwise, we say that m_1 and m_2 are connected. We say that $\mathcal{V} \subset \mathcal{M}$ is a connected branch of $\mathcal{M}$ if all points of $\mathcal{V}$ are connected to each other and do not connect to any points in $\mathcal{M} - \mathcal{V}$.*

Definition 3.2.3 (Volume) *For a $d_x \times d_k$ matrix $A = (x_1, x_2, ..., x_{d_k})$ with $x_i \in \mathbb{R}^{d_x}$,*

$1 \leqslant i \leqslant d_k \leqslant d_x$, *we define* $Vol(A) = \sqrt{\det(A'A)}$, *which is the volume of the parallelepiped spanned by* $x_1, x_2, ..., x_{d_k}$.

The volume measures the amount of mass of a d_k-dimensional parallelepiped in $\mathbb{R}^{d_k}$. This concept is essential for integration on the manifold, which we recall here.

Definition 3.2.4 (Integration on a parametrized manifold) *Let* $\mathcal{K}$ *be open in* $\mathbb{R}^{d_k}$, *and let* $\alpha : \mathcal{K} \to \mathbb{R}^{d_x}$ *be of class* $\mathcal{C}^r$ *on* $\mathcal{K}$, $r \geqslant 1$. *The set* $\mathcal{M} = \alpha(\mathcal{K})$ *together with the map* α *constitute a parametrized* d_k*-manifold without boundary in* $\mathbb{R}^{d_x}$ *of class* $\mathcal{C}^r$. *Let* g *be a real-valued continuous function defined at each point of* $\mathcal{M}$. *We define the integral of* g *over* $\mathcal{M}$ *with respect to volume by*

$$\int_{\mathcal{M}} g(m) dVol := \int_{\mathcal{K}} (g \circ \alpha)(k)\, Vol(D\alpha(k)), \tag{3.2.4}$$

provided that the integral exists. Here $D\alpha(k)$ *is the Jacobian matrix of the mapping* $k \mapsto \alpha(k)$, *and* $Vol(D\alpha(k))$ *is the volume of matrix* $D\alpha(k)$ *as defined in Definition 3.2.3.*

The above definition coincides with the usual interpretation of integration. The integral can be extended to manifolds that do not admit a global parametrization α using the notion of partition of unity. This partition is a set of smooth local functions defined in a neighborhood of the manifold. The following Lemma shows the existence of the partition to unity and is proven in Theorem 3.11 in Munkres (1990).

Lemma 20 (Partition to Unity) *Let* $\mathcal{M} \subseteq \mathbb{R}^{d_x}$ *and let* ϑ *be an open cover of* $\mathcal{M}$. *Then, there is a collection* $\mathcal{P} = \{p_i \in \mathcal{C}^\infty : i \in \mathcal{I}\}$, *where* p_i *is defined on an open set containing* $\mathcal{M}$ *for all* $i \in \mathcal{I}$, *with the following properties: (1) For each* $m \in \mathcal{M}$ *and* $i \in \mathcal{I}$, $0 \leqslant p_i(m) \leqslant 1$. *(2) For each* $m \in \mathcal{M}$ *there is an open set* $\mathcal{V}$ *containing* m *such that all but finitely many* $p_i \in \mathcal{P}$ *are* 0 *on* $\mathcal{V}$. *(3) For each* $m \in \mathcal{M}$, $\sum_{p_i \in \mathcal{P}} p_i(m) = 1$. *(4) For each* $p_i \in \mathcal{P}$ *there is an open set* $\mathcal{U} \in \vartheta$, *such that* $supp(p_i) \subset \mathcal{U}$.

Now we are ready to recall the definition of integration on a manifold.

Definition 3.2.5 (Integration on a manifold with partition of unity) *Let* $\vartheta := \{\vartheta_j : j \in J\}$ *be an open cover of a* d_k*-manifold without boundary* $\mathcal{M}$ *in* $\mathbb{R}^{d_x}$ *of class* $\mathcal{C}^r$, $r \geqslant 1$. *Suppose there is an coordinate patch* $\alpha_j : \mathcal{V}_j \subset \mathbb{R}^{d_k} \to \vartheta_j$, *one-to-one and of class* $\mathcal{C}^r$ *on* $\mathcal{V}_j$ *for each* $j \in J$. *Denote* $\mathcal{K}_j = \alpha_j^{-1}(\mathcal{M} \cap \vartheta_j)$. *Then for a differentiable function* g *defined on an open set that contains* $\mathcal{M}$, *we define the integral of* g *over* $\mathcal{M}$ *with respect to volume by:*

$$\int_{\mathcal{M}} g(m) dVol := \sum_{j \in J} \sum_{i \in \mathcal{I}} \int_{\mathcal{K}_j} [(p_i g) \circ \alpha_j](k) \, Vol(D\alpha_j(k)), \tag{3.2.5}$$

where $\{p_i \in \mathcal{C}^\infty : i \in \mathcal{I}\}$ *is a partition to unity that satisfies the conditions of Lemma 1. Theorem 25.4 in Munkres (1990) shows that the integral is well defined and does not depend on the choice of cover and partition to unity.*

3.2.2 Basic Analytical Properties of Sorted Functions

Recall that the main functions in the analysis are the PE function $\Delta(x)$ and the probability measure $\mu(x)$. We make the following technical assumptions about these functions:

S.1. The domain of Δ of interest, $\mathcal{X}$, is open and its closure $\bar{\mathcal{X}}$ is compact. There exists an open set $B(\mathcal{X})$ containing $\bar{\mathcal{X}}$ such that $x \mapsto \Delta(x)$ is $\mathcal{C}^1$ on $B(\mathcal{X})$ and $x \mapsto \mu(x)$ is $\mathcal{C}^0$ on $B(\mathcal{X})$.

S.2. For any regular value δ of Δ on $\bar{\mathcal{X}}$, $\mathcal{M}_\Delta(\delta) := \{x \in B(\mathcal{X}) : \Delta(x) = \delta\}$ has a finite number of connected branches.

Comment 3.2.1 (Continuous X**)** *S.1 assumes that* μ *is continuous on an open set that contains* $\mathcal{X}$. *This restricts* X *to include only continuos components. We differ the treatment of discrete components to Section 3.4.*

Comment 3.2.2 (Properties of $\mathcal{M}_\Delta(\delta)$**)** *Lemma 28 in the Appendix shows that S.1 and S.2 imply that* $\mathcal{M}_\Delta(\delta)$ *is a* $(d_x - 1)$*-manifold without boundary in* $\mathbb{R}^{d_x}$ *of class* $\mathcal{C}^1$, *for any* δ *that is a regular value of* $x \mapsto \Delta(x)$ *on* $\bar{\mathcal{X}}$.

The following lemma establishes the properties of the distribution function $\delta \mapsto F_{\Delta,\mu}(\delta)$ and the SPE-function $u \mapsto \Delta^*_\mu(u)$. Define $\mathcal{D}^*$ as the set of regular values of $x \mapsto \Delta(x)$ on $\bar{\mathcal{X}}$.

Lemma 21 (Basic Properties of $F_{\Delta,\mu}$ and Δ^*_μ) *Under conditions S.1 and S.2:*

1. For any $\delta \in \mathcal{D}^*$, the derivative of $F_{\Delta,\mu}(\delta)$ with respect to δ, denoted as $f_{\Delta,\mu}(\delta)$, can be expressed as:

$$f_{\Delta,\mu}(\delta) := \partial_\delta F_{\Delta,\mu}(\delta) = \int_{\mathcal{M}_\Delta(\delta)} \frac{\mu(x)}{\|\nabla\Delta(x)\|} dVol. \tag{3.2.6}$$

This integral is well-defined because $\mathcal{M}_\Delta(\delta) \cap \bar{\mathcal{X}}$ is a compact set, and the gradient $x \mapsto \nabla\Delta(x)$ is finite and continuous, and bounded away from 0 on $\mathcal{M}_\Delta(\delta) \cap \bar{\mathcal{X}}$.

2. For any $u \in \{\tilde{u} : \Delta^*_\mu(\tilde{u}) \in \mathcal{D}^*\}$, $\Delta^*_\mu(u)$ has derivative with respect to u:

$$\partial_u \Delta^*_\mu(u) = \frac{1}{f_{\Delta,\mu}(\Delta^*_\mu(u))}. \tag{3.2.7}$$

Comment 3.2.3 (Properties of μ at the boundary of $\mathcal{X}$) *S.1 imposes that the probability measure $x \mapsto \mu(x)$ is continuous and vanishes at the boundary of $\mathcal{X}$. To understand the importance of this condition, we consider the following example with dimension $d_x = 2$:*

$$\Delta(x) = sin(x_1^2 + x_2^2)$$

on $\mathcal{X} = \{(x_1, x_2) : x_1^2 + x_2^2 < \pi/6\}$, and $x \mapsto \mu(x)$ is uniform on $\mathcal{X}$. It is easy to see that $\delta = 1/2$ is a regular value of Δ on $\bar{\mathcal{X}}$. However, $F_{\Delta,\mu}(\delta) = \int_{\mathcal{X}} 1\{\Delta(x) \leqslant \delta\} d\mu$ is not differentiable at $\delta = 1/2$. The right derivative $\lim_{\eta \to 0^+}[F_{\Delta,\mu}(\delta+\eta) - F_{\Delta,\mu}(\delta)]/\eta = 0$, whereas the left derivative $\lim_{\delta \to 0^-}[F_{\Delta,\mu}(\delta+\eta) - F_{\Delta,\mu}(\delta)]/\eta = \sqrt{2\pi^3/3}$. Lemma 21 does not apply because $x \mapsto \mu(x)$ is not continuous on any open set $B(\mathcal{X}) \supset \bar{\mathcal{X}}$. Technically, we can replace the continuity of $x \mapsto \mu(x)$ at the boundary of $\mathcal{X}$ by the weaker condition

$$\int_{\mathcal{M}_\Delta(\delta)} 1\{x \in \partial\mathcal{X}\}\mu(x) dVol = 0, \tag{3.2.8}$$

where $\partial\mathcal{X}$ denotes the boundary of $\mathcal{X}$.

Comment 3.2.4 (Case $d_x = 1$) *The derivatives of Lemma 21 coincide with the expressions for the derivatives of rerrangement-related functions in Proposition 1 of Chernozhukov, Fernandez-Val, and Galichon (2010) when $d_x = 1$. In this case, the manifold $\mathcal{M}_\Delta(\delta)$ has dimension 0, i.e. it is a set of finite number of points.*

Comment 3.2.5 (Derivatives over $\delta \in \Delta(\mathcal{X})$) *Lemma 21 states that $\delta \mapsto F_{\Delta,\mu}(\delta)$ $(u \mapsto \Delta^*_\mu(u))$ is $\mathcal{C}^1$ on any compact set of $\mathcal{D}^*$ (the Δ^*_μ pre-image of $\mathcal{D}^*$). $\mathcal{D}^* = \Delta(\mathcal{X}) := \{\Delta(x) : x \in \mathcal{X}\}$ when the map $x \mapsto \Delta(x)$ does not have critical points on $\mathcal{X}$. For example, this holds when the PE $\Delta(x)$ is strictly monotonic in one of the components of x, say the first component x_1 when $x = (x_1, x_{-1})$. In this case, the derivative of $\delta \mapsto F_{\Delta,\mu}(\delta)$ can be expressed as the Lebesgue integral*

$$f_{\Delta,\mu}(\delta) = \int_{\mathcal{X}_{-1}} f_{\Delta(X)|X_{-1}}(\delta|x_{-1})\mu_{-1}(x_{-1}),$$

where μ_{-1} is the probability measure of X_{-1}, $\mathcal{X}_{-1}$ is the support of μ_{-1}, $f_{\Delta(X)|X_{-1}}(\delta|x_{-1}) = \mu_{1|-1}(\Delta^{-1}(\delta, x_{-1})|x_{-1})$, $\mu_{1|-1}$ is the probability measure of X_1 conditional on X_{-1}, and $\delta \mapsto \Delta^{-1}(\delta, x_{-1})$ is the inverse function of $x_1 \mapsto \Delta(x_1, x_{-1})$. Chernozhukov, Fernandez-Val, Hoderlein, Holzmann, and Newey (2014) use a similar condition to identify quantile derivatives in nonseparable panel models.

3.2.3 Functional Derivatives of Sorting-Related Operators

We consider the properties of the distribution function and the SPE-function as functional operators $(\Delta, \mu) \mapsto F_{\Delta,\mu}$ and $(\Delta, \mu) \mapsto \Delta^*_\mu$. We show that these operators are Hadamard differentiable with respect to (Δ, μ). These results are crucial to derive the limiting distribution of the empirical versions of $F_{\Delta,\mu}$ and Δ^*_μ in Section 3.3.

We first recall the definition of Hadamard differentiability from van der Vaart and Wellner (1996).

Definition 3.2.6 (Hadamard Derivative) *Suppose the linear spaces $\mathbb{F}$ and $\mathbb{G}$ are*

equipped with the norms $\|\cdot\|_{\mathbb{F}}$ *and* $\|\cdot\|_{\mathbb{G}}$. *A map* $\phi : \mathbb{F}_\phi \subseteq \mathbb{F} \to \mathbb{G}$ *is called Hadamard-differentiable at* $f \in \mathbb{F}_\phi$ *tangentially to* $\mathbb{F}_0 \subseteq \mathbb{F}$ *if there is a continuous linear map* $\partial_f \phi : \mathbb{F}_0 \to \mathbb{G}$ *such that*

$$\frac{\phi(f + t_n h_n) - \phi(f)}{t_n} \to \partial_f \phi[h], \quad n \to \infty, \tag{3.2.9}$$

for all converging real sequences $t_n \to 0$ *and* $\|h_n - h\|_{\mathbb{F}} \to 0$ *such that* $f + t_n h_n \in \mathbb{F}_\phi$ *for every* n, *and* $h \in \mathbb{F}_0$.

Hadamard differentiability of $F_{\Delta,\mu}$ and Δ^*_μ with respect to Δ

We first show differentiability with respect to the PE:

Lemma 22 (Hadamard differentiability of $\Delta \mapsto F_{\Delta,\mu}$ and $\Delta \mapsto \Delta^*_\mu$) *Let* $\mathbb{F}$ *denote the family of all continuously differentiable functions on* $B(\mathcal{X})$ *equipped with sup-norm, and* $\mathbb{F}_0$ *denote a set of uniformly bounded continuously differentiable functions on* $B(\mathcal{X})$ *equipped with sup-norm. Suppose that S.1-S.2 hold. Then:*

(a) For any $\delta \in \mathcal{D}^*$, *the map* $F_{\Delta,\mu}(\delta) : \mathbb{F} \to \mathbb{R}$ *is Hadamard-differentiable at* Δ *tangentially to* $\mathbb{F}_0$, *with derivative defined by*

$$G \mapsto \partial_\Delta F_{\Delta,\mu}(\delta)[G] := -\int_{\mathcal{M}_\Delta(\delta)} \frac{G(x)\mu(x)}{\|\nabla \Delta(x)\|} dVol,$$

as a map from $\mathbb{F}_0$ *to* $\mathbb{R}$.

(b) For any $u \in \{\tilde{u} : \Delta^*_\mu(\tilde{u}) \in \mathcal{D}^*\}$, *the map* $\Delta^*_\mu(u) : \mathbb{F} \to \mathbb{R}$ *is Hadamard-differentiable at* Δ *tangentially to* $\mathbb{F}_0$, *with derivative*

$$G \mapsto \partial_\Delta \Delta^*_\mu(u)[G] := -\frac{\partial_\Delta F_{\Delta,\mu}(\Delta^*_\mu(u))[G]}{f_{\Delta,\mu}(\Delta^*_\mu(u))},$$

as a map from $\mathbb{F}_0$ *to* $\mathbb{R}$.

Hadamard differentiability of $F_{\Delta,\mu}$ and Δ^*_μ with respect to μ

To show differentiability with respect to the measure μ, it is convenient to identify μ with an operator $g \mapsto \int_{\mathcal{X}} g(x)\mu(x)$ mapping $\mathbb{F}_M$ to $\mathbb{R}$, where $\mathbb{F}_M$ is a subset of all L^1-integrable functions on $B(\mathcal{X})$ uniformly bounded by 1 in terms of absolute value. Define $\mathbb{H}_0$ as the set of all bounded linear operators on $\mathbb{F}_M$ with the following norm $L^{*\infty}$:

$$\|H\|_{L^{*\infty}} = \sup_{f \in \mathbb{F}_M, f \neq 0} |H(f)|.$$

and define the corresponding distance between two operators H_1 and H_2 in $\mathbb{H}_0$ as $\|H_1 - H_2\|_{L^{*\infty}} = \sup_{f \in \mathbb{F}_M, f \neq 0} |H_1(f) - H_2(f)|$. In this setting, $\mu \in \mathbb{H}_0$.

Lemma 23 (Hadamard Differentiability of $\mu \mapsto F_{\Delta,\mu}$ and $\mu \mapsto \Delta^*_\mu(u)$) *Suppose that S.1-S.2 hold. Then,*

(a) For any $\delta \in \mathcal{D}^*$, the map $F_{\Delta,\mu}(\delta) : \mathbb{H}_0 \to \mathbb{R}$ is Hadamard differentiable at μ tangentially to $\mathbb{H}_0$, with derivative defined by

$$\partial_\mu F_{\Delta,\mu}(\delta)[H] := H(1(\Delta \leqslant \delta)), \tag{3.2.10}$$

as a map from $\mathbb{H}_0$ to $\mathbb{R}$.

(b) For any $u \in \{\tilde{u} : \Delta^*_\mu(\tilde{u}) \in \mathcal{D}^*\}$, the map $\Delta^*_\mu(u) : \mathbb{H}_0 \to \mathbb{R}$ is Hadamard differentiable at μ tangentially to $\mathbb{H}_0$, with the derivative map defined by

$$H \mapsto \partial_\mu \Delta^*_\mu(u)[H] := -\frac{H(1(\Delta \leqslant \delta))}{f_{\Delta,\mu}(\Delta^*_\mu(u))}, \tag{3.2.11}$$

as a map from $\mathbb{H}_0$ to $\mathbb{R}$.

Hadamard differentiability of $F_{\Delta,\mu}$ and Δ^*_μ with respect to (Δ, μ)

We combine the results of the previous two subsections using the following assumption:

S.3. Let $\mathbb{F}_0$ denote the space of functions of Lemma 22. There exists a class of functions $\mathbb{F}_1$ defined on $B(\mathcal{X})$ such that:

(1) $\mathbb{F}_0 \subseteq \mathbb{F}_1$.

(2) For any $\delta \in \mathcal{D}^*$ and $H \in \mathbb{H}_0$, H is continuous on $\mathbb{F}_{\mathcal{M}_\Delta}(\delta) := \{1(\widetilde{\Delta} \leqslant \delta) : \widetilde{\Delta} \in \mathbb{F}_1\}$. Assuming $\mathbb{F}_{\mathcal{M}_\Delta}(\delta)$ is $\mu-$ Donsker.

Let $\mathbb{D} := \mathbb{F}_1 \times \mathbb{H}_0$ and $\mathbb{D}_0 := \mathbb{F}_0 \times \mathbb{H}_0$. The following Lemma gives the main result of this Section.

Lemma 24 (Hadamard differentiability of $(\Delta, \mu) \mapsto F_{\Delta,\mu}$ and $(\Delta, \mu) \mapsto \Delta^*_\mu$) *Suppose that S.1-S.3 hold.*

(a) For any $\delta \in \mathcal{D}^*$, the map $F_{\Delta,\mu}(\delta) : \mathbb{D} \to \mathbb{R}$ is Hadamard differentiable at (Δ, μ) tangentially to $\mathbb{D}_0$, with derivative defined by

$$(G, H) \mapsto \partial_{\Delta,\mu} F_{\Delta,\mu}(\delta)[G, H] := -\int_{\mathcal{M}_\Delta(\delta)} \frac{G(x)\mu(x)}{\|\nabla\Delta(x)\|} dVol + H(1(\Delta \leqslant y)),$$

as a map from $\mathbb{D}_0$ to $\mathbb{R}$.

(b) For any $u \in \{\tilde{u} : \Delta^*_\mu(\tilde{u}) \in \mathcal{D}^*\}$, the map $\Delta^*_\mu(u) : \mathbb{D} \to \mathbb{R}$ is Hadamard differentiable at (Δ, μ) tangentially to $\mathbb{D}_0$, with derivative map defined by

$$(G, H) \mapsto \partial_{\Delta,\mu}\Delta^*_\mu(u)[G, H] := -\frac{\partial_{\Delta,\mu} F_{\Delta,\mu}(\Delta^*_\mu(u))[G, H]}{f_{\Delta,\mu}(\Delta^*_\mu(u))}.$$

3.3 Asymptotic Theory for Empirical SPE

3.3.1 Empirical SPE

In practice, we replace the PE Δ and the probability measure μ by sample analogs to construct plug-in estimators of the SPE. Let $\widehat{\Delta}(x)$ and $\widehat{\mu}(x)$ be estimators of $\Delta(x)$ and $\mu(x)$ obtained from a sample of size n. The estimator of Δ^*_μ is

$$\widehat{\Delta^*_\mu}(u) := \hat{\Delta}^*_{\hat{\mu}}(u) = \inf_{\delta \in \mathbb{R}} \{F_{\hat{\Delta},\hat{\mu}}(\delta) \geqslant u\},$$

where $F_{\hat{\Delta},\hat{\mu}}(\delta) = \mathbb{E}_{\widehat{\mu}}[1\{\widehat{\Delta}(X) \leqslant \delta\}] =: \widehat{F_{\Delta,\mu}}(\delta)$.

Example 1 (Probit model, cont.) Given $\{(Y_i, X_i) : 1 \leqslant i \leqslant n\}$, a sample of (Y, X), the estimator of the PE is

$$\hat{\Delta}(x) = \Phi\left(P(1,w)'\hat{\beta}\right) - \Phi\left(P(0,w)'\hat{\beta}\right),$$

where $\hat{\beta}$ is the maximum likelihood estimator (MLE) of β,

$$\hat{\beta} \in \arg\max_{b \in \mathbb{R}^{d_p}} \sum_{i=1}^{n} [Y_i \log \Phi(P(X_i)'b) + (1 - Y_i) \log \Phi(-P(X_i)'b)],$$

for $d_p = \dim P(X)$. If μ is the distribution of X in a part of the population defined by $X \in \mathcal{X}$, for some set $\mathcal{X}$ with positive measure, we can estimate it by the empirical distribution in $\mathcal{X}$

$$\hat{\mu}(x) = \sum_{i=1}^{n} 1\{X_i \in \mathcal{X}\} 1\{X_i \leqslant x\} / \sum_{i=1}^{n} 1\{X_i \in \mathcal{X}\}.$$

Example 2 (Tobit model, cont.) Given $\{(Y_i, X_i) : 1 \leqslant i \leqslant n\}$, a sample of (Y, X), the estimator of the PE is

$$\hat{\Delta}(x) = \left[P(1,w)'\hat{\beta} + \hat{\sigma}\lambda(P(1,w)'\hat{\beta}/\hat{\sigma})\right] - \left[P(0,w)'\hat{\beta} + \hat{\sigma}\lambda(P(0,w)'\hat{\beta}/\hat{\sigma})\right]$$

where $(\hat{\beta}, \hat{\sigma})$ is the MLE of (β, σ),

$$(\hat{\beta}, \hat{\sigma}) \in \arg\max_{b \in \mathbb{R}^{d_p}, s \in \mathbb{R}_+} \sum_{i=1}^{n} [1\{Y_i = 0\} \log \Phi(-P(X_i)'b/s) + 1\{Y_i > 0\} \log\{s^{-1}\Phi'((Y_i - P(X_i)'b)/s)\}],$$

for $d_p = \dim P(X)$. As in Example 1, we can estimate the measure μ using the corresponding empirical distribution.

We use the Hadamard differentiability of the sorting-related operators and the delta method to derive functional central limit theorems for $\delta \mapsto \widehat{F_{\Delta,\mu}}(\delta)$ and $u \mapsto \widehat{\Delta^*_\mu}(u)$ over regions that exclude the critical values of $x \mapsto \Delta(x)$ on $\bar{\mathcal{X}}$. To describe this results, let $\ell^\infty(\mathcal{V})$ denote the set of bounded and measurable functions $g : \mathcal{V} \mapsto \mathbb{R}$.

We consider 3 different cases depending on whether the PE and the probability

measure are treated as known or unknown.

3.3.2 Case 1: Δ unknown, μ known

We first discuss the properties of $F_{\hat{\Delta},\mu}$ and $\hat{\Delta}^*_\mu$, the estimators of $F_{\Delta,\mu}$ and Δ^*_μ when μ is treated as known. We make the following assumptions about the estimator of the PE:

S.4. The estimator $\hat{\Delta}$ of Δ obeys a functional central limit theorem, namely,

$$a_n(\hat{\Delta} - \Delta) \rightsquigarrow G_\infty \text{ in } \ell^\infty(B(\mathcal{X})),$$

where a_n is a sequence such that $a_n \to \infty$ as $n \to \infty$, and G_∞ is a tight process that has uniformly continuous sample paths over $\mathcal{B}(\mathcal{X})$ a.s. [CAN WE REPLACE $B(\mathcal{X})$ BY $\mathcal{X}$ HERE?]

S.5. The gradient $x \mapsto \nabla\hat{\Delta}(x)$ exists, is continuous at each $x \in B(\mathcal{X})$, and

$$\sup_{x \in B(\mathcal{X})} \|\nabla\hat{\Delta}(x) - \nabla\Delta(x)\| \to_{\mathrm{P}} 0.$$

The following result is a corollary of Lemma 22:

Proposition 1 (FCLT for $F_{\hat{\Delta},\mu}$ and $\hat{\Delta}^*_\mu$) *Under S.1, S.2, S.4, and S.5, as $n \to \infty$:*

(a) In $\ell^\infty(\mathcal{V})$, where $\mathcal{V}$ is any compact subset of $\mathcal{D}^*$,

$$a_n(F_{\hat{\Delta},\mu}(\delta) - F_{\Delta,\mu}(\delta)) \rightsquigarrow \partial_\Delta F_{\Delta,\mu}(\delta)[G_\infty] = -\int_{\mathcal{M}_\Delta(\delta)} \frac{G_\infty(x)\mu(x)}{\|\nabla\Delta(x)\|} dVol =: T_\infty(\delta), \quad (3.3.1)$$

as a stochastic process indexed by $\delta \in \mathcal{V}$.

(b) In $\ell^\infty(\mathcal{U}_\mathcal{V})$, with $\mathcal{U}_\mathcal{V} = \{u \in \mathcal{U} : \Delta^*_\mu(u) \in \mathcal{V}\}$,

$$a_n(\hat{\Delta}^*_\mu(u) - \Delta^*_\mu(u)) \rightsquigarrow \partial_\Delta \Delta^*_\mu(u)[G_\infty] = -\frac{T_\infty(\Delta^*_\mu(u))}{f_{\Delta,\mu}(\Delta^*_\mu(u))}, \quad (3.3.2)$$

as a stochastic process indexed by $u \in \mathcal{U}_\mathcal{V}$.

Comment 3.3.1 *Replacing the expressions of* $T_\infty(\delta)$ *and* $f_{\Delta,\mu}(\delta)$ *in the last limit,*

$$\partial_\Delta \Delta^*_\mu(u)[G_\infty] = \frac{\int_{\mathcal{M}_\Delta(\Delta^*_\mu(u))} \frac{G_\infty(x)\mu(x)}{\|\nabla\Delta_\mu\|} dVol}{\int_{\mathcal{M}_\Delta(\Delta^*_\mu(u))} \frac{\mu(x)}{\|\nabla\Delta(x)\|} dVol}.$$

The limit process is therefore the average of the process $G_\infty(x)$ *on* $\mathcal{M}_\Delta(\Delta^*_\mu(u))$ *with respect to the density*

$$\frac{\frac{\mu(x)}{\|\nabla\Delta(x)\|}}{\int_{\mathcal{M}_\Delta(\Delta^*_\mu(u))} \frac{\mu(x)}{\|\nabla\Delta(x)\|} dVol}. \tag{3.3.3}$$

3.3.3 Case 2: Δ known, μ unknown

We consider the properties of $F_{\Delta,\hat{\mu}}$ and $\Delta^*_{\hat{\mu}}$, the estimators of $F_{\Delta,\mu}$ and Δ^*_μ when Δ is treated as known. We make the following assumptions about $\hat{\mu}$, the estimator of the measure μ:

S.6. The function $x \mapsto \hat{\mu}(x)$ is a measure over $\mathcal{X}$ obeying in $\ell^\infty(\mathcal{V})$,

$$\int_{\mathcal{X}} 1\{\Delta(x) \leqslant \delta\} b_n(\hat{\mu}(x) - \mu(x)) \rightsquigarrow H_\infty(\delta), \tag{3.3.4}$$

as an stochastic process indexed by $\delta \in \mathcal{V}$, where $\mathcal{V}$ is any compact subset of $\mathcal{D}^*$, and b_n is a sequence such that $b_n \to \infty$ as $n \to \infty$.

This assumption is satisfied by most of the estimators used in practice. For example, when $\hat{\mu}$ is the empirical measure for the entire population from a random sample, $b_n = \sqrt{n}$ and $H_\infty(\delta) := B_\mu(1\{\Delta \leqslant \delta\})$, where B_μ is a μ-Brownian Bridge, i.e. a Gaussian process with zero mean and covariance function $(g_1, g_2) \mapsto \int g_1 g_2 d\mu - \int g_1 d\mu \int g_2 d\mu$.

The following result is a corollary of Lemma 23:

Proposition 2 (FCLT for $F_{\Delta,\hat{\mu}}$ and $\Delta^*_{\hat{\mu}}$) *Under S.1,S.2, S.3, and S.6, as* $n \to \infty$*:*

(a) In $\ell^\infty(\mathcal{V})$*, where* $\mathcal{V}$ *is any compact subset of* $\mathcal{D}^*$*,*

$$b_n(F_{\Delta,\hat{\mu}}(\delta) - F_{\Delta,\mu}(\delta)) \rightsquigarrow \partial_\mu F_{\Delta,\mu}(\delta)[H_\infty] = H_\infty(\delta),$$

as a stochastic process indexed by $\delta \in \mathcal{V}$.

(b) In $\ell^\infty(\mathcal{U}_\mathcal{V})$, *with* $\mathcal{U}_\mathcal{V} = \{u \in \mathcal{U} : \Delta^*_\mu(u) \in \mathcal{V}\}$,

$$b_n(\Delta^*_{\hat{\mu}}(u) - \Delta^*_\mu(u)) \rightsquigarrow \partial_\mu \Delta^*_\mu(u)[H_\infty] = -\frac{H_\infty(\Delta^*_\mu(u))}{f_{\Delta,\mu}(\Delta^*_\mu(u))},$$

as a stochastic process indexed by $u \in \mathcal{U}_\mathcal{V}$.

3.3.4 Case 3: Δ unknown, μ unknown

We combine the results of Propositions 1 and 2 to deal with the most empirically relevant case where both the PE and the probability measure are estimated. Let $r_n := a_n \wedge b_n$, the slowest of the rates of convergence of $\widehat{\Delta}$ and $\widehat{\mu}$. Then, $r_n/a_n \to s_\Delta \in [0,1]$ and $r_n/b_n \to s_\mu \in [0,1]$, where $s_\Delta = 0$ when $b_n = o(a_n)$ and $s_\mu = 0$ when $a_n = o(b_n)$.

The following result is a corollary of Lemma 24.

Theorem 7 (FCLT for $F_{\hat{\Delta},\hat{\mu}}$ and $\hat{\Delta}^*_{\hat{\mu}}$) *Suppose that S.1-S.6 hold,* $\hat{\Delta} \in \mathbb{F}_1$ *with probability approaching 1, and* $\mathbb{F}_{\mathcal{M}_\Delta(\delta)}$ *is* μ*-Donsker for any* $\delta \in \mathcal{D}^*$. *Then, as* $n \to \infty$,

(a) In $\ell^\infty(\mathcal{V})$, *where* $\mathcal{V}$ *is any compact subset of* $\mathcal{D}^*$,

$$r_n(F_{\hat{\Delta},\hat{\mu}}(\delta) - F_{\Delta,\mu}(\delta)) \rightsquigarrow \partial_{\Delta,\mu} F_{\Delta,\mu}(\delta)[s_\Delta G_\infty, s_\mu H_\infty] = s_\Delta T_\infty(\delta) + s_\mu H_\infty(\delta), \quad (3.3.5)$$

as a stochastic process indexed by $\delta \in \mathcal{V}$, *where* $T_\infty(\delta)$ *is defined in Proposition 1.*

(b) In $\ell^\infty(\mathcal{U}_\mathcal{V})$, *with* $\mathcal{U}_\mathcal{V} = \{u \in \mathcal{U} : \Delta^*_\mu(u) \in \mathcal{V}\}$,

$$r_n(\hat{\Delta}^*_{\hat{\mu}}(u) - \Delta^*_\mu(u)) \rightsquigarrow \partial_{\Delta,\mu}\Delta^*_\mu(u)[s_\Delta G_\infty, s_\mu H_\infty] = -\frac{s_\Delta T_\infty(\Delta^*_\mu(u)) + s_\mu H_\infty(\Delta^*_\mu(u))}{f_{\Delta,\mu}(\Delta^*_\mu(u))} =: Z_\infty(u), \quad (3.3.6)$$

as a stochastic process indexed by $u \in \mathcal{U}_\mathcal{V}$.

Comment 3.3.2 (Known Δ or known μ) *Proposition 1 can be seen a special case of Theorem 7 with* $r_n = a_n$ *and* $s_\mu = 0$. *Similarly, Proposition 2 can be seen a special case*

of Theorem 7 with $r_n = b_n$ and $s_\Delta = 0$. Accordingly, we shall not distinguish between these cases in the rest of paper.

Comment 3.3.3 (Critical Values of Δ) *Theorem 7 applies to regular values $\delta \in \mathcal{D}^*$. If $\Delta^*_\mu(u)$ is in a neighborhood of a critical value, the finite-sample distribution of $r_n(\hat{\Delta}^*_{\hat{\mu}}(u) - \Delta^*_\mu(u))$ can be very different from the asymptotic distribution $Z_\infty(u)$. We illustrate this point in Section 3.5 through numerical simulations. The limit distribution of $r_n(\hat{\Delta}^*_{\hat{\mu}}(u) - \Delta^*_\mu(u))$ local to a critical value is an interesting problem for further investigation.*

Comment 3.3.4 (Distribution of Z_∞) *The limit Z_∞ is usually a Gaussian process with zero mean and a covariance function that simplifies because $T_\infty(\delta)$ and $H_\infty(\delta)$ are independent. This is the case, for example, when Δ is some characteristic of the conditional distribution of an outcome variable Y given X, which is estimated by MLE, OLS, GMM or quantile regression methods on a random sample of (Y, X), and the measure μ is estimated by the empirical distribution of X on the random sample.*

Comment 3.3.5 (Donsker condition) *The assumption that $\mathbb{F}_{\mathcal{M}_\Delta(\delta)}$ is μ-Donsker holds under standard conditions if the PE is parametrically estimated as in Examples 12 and 13. It also holds for many semiparametric and nonparametric estimators such as least squares, GMM, quantile regression, local kernel regression, and global series regression under appropriate conditions.*

3.3.5 Inference on SPE

We can construct asymptotically valid confidence interval for the SPE using the functional central limit theorems for $\widehat{\Delta^*_\mu}$. We consider pointwise intervals that cover the SPE at a specified value of u, and uniform bands that cover the SPE-function simultaneously over a region of values of u. The next two results are corollaries of Theorem 7.

Corollary 3 (Pointwise Inference on SPE) *Under the assumptions of Theorem 7, for any* $u \in \{\tilde{u} \in \mathcal{U} : \Delta^*_\mu(\tilde{u}) \in \mathcal{D}^*\}$ *and* $0 < \alpha < 1$,

$$\mathrm{P}\left\{\Delta^*_\mu(u) \in \left[\widehat{\Delta^*_\mu}(u) - Z_{\infty,1-\alpha}(u)/r_n, \widehat{\Delta^*_\mu}(u) + Z_{\infty,1-\alpha}(u)/r_n\right]\right\} \to 1-\alpha,$$

where $Z_{\infty,\alpha}(u)$ *is the* α*-quantile of* $|Z_\infty(u)|$ *for the random variable* $Z_\infty(u)$ *defined in Theorem 7.*

Corollary 4 (Uniform Inference on SPE-function) *Let* $\mathcal{V}$ *be any compact set of* $\mathcal{D}^*$ *and* $\mathcal{U}_\mathcal{V} = \{u \in \mathcal{U} : \Delta^*(u) \in \mathcal{V}\}$. *Under the assumptions of Theorem 7, for any* $0 < \alpha < 1$,

$$\mathrm{P}\left\{\Delta^*_\mu(u) \in \left[\widehat{\Delta^*_\mu}(u) - Z_{\infty,1-\alpha}(\mathcal{U}_\mathcal{V})/r_n, \widehat{\Delta^*_\mu}(u) + Z_{\infty,1-\alpha}(\mathcal{U}_\mathcal{V})/r_n\right] : u \in \mathcal{U}_\mathcal{V}\right\} \to 1-\alpha,$$

where $Z_{\infty,\alpha}(\mathcal{U}_\mathcal{V})$ *is the* α*-quantile of* $Z_\infty(\mathcal{U}_\mathcal{V}) := \sup_{u\in\mathcal{U}_\mathcal{V}} |Z_\infty(u)|$ *for the process* $U \mapsto Z_\infty(u)$ *defined in Theorem 7.*

Comment 3.3.6 *Note that Corollary 3 is a special case of Corollary 4 when the set* $\mathcal{U}_\mathcal{V}$ *is a singleton, so we shall not consider separately pointwise inference in the rest of the paper.*

3.3.6 Bootstrap Inference

The critical value $Z_{\infty,1-\alpha}(\mathcal{U}_\mathcal{V})$ to construct the confidence band of Corollary 4 can be hard to obtain in practice. In principle one can simulate the process $Z_\infty(\mathcal{U}_\mathcal{V})$, but it might be difficult to numerically locate and parametrize the manifold $\mathcal{M}_\Delta(\delta)$, and to evaluate integrals on $\mathcal{M}_\Delta(\delta)$. This creates a real challenge to implement our inference methods. To deal with this challenge we propose using exchangeable bootstrap to compute critical values (Praestgaard and Wellner (1993) and van der Vaart and Wellner (1996)), instead of simulation. We show that the bootstrap law is consistent to approximate the distribution of the limit process Z_∞ of Theorem 7.

We start describing the algorithm to obtain the bootstrap law of $Z_\infty(\mathcal{U}_\mathcal{V})$. Let

$(\omega_1, \ldots, \omega_n)$ denote the bootstrap weights, which are nonnegative random variables independent from the data. For example, $(\omega_1, \ldots, \omega_n)$ is multinomial vector with dimension n and probabilities $(1/n, \ldots, 1/n)$ in the empirical bootstrap.

Algorithm 4 (Bootstrap law of $Z_\infty(\mathcal{U}_V)$) *1. Draw a realization of the bootstrap weights $(\omega_1, \ldots, \omega_n)$.*

2. *For each $u \in \mathcal{U}_V$, compute $\widetilde{\Delta}^*_\mu(u) = \tilde{\Delta}^*_{\tilde{\mu}}(u)$, a bootstrap draw of $\widehat{\Delta}^*_\mu(u) = \hat{\Delta}^*_{\hat{\mu}}(u)$, where $\tilde{\Delta}$ and $\tilde{\mu}$ are the bootstrap versions of $\hat{\Delta}$ and $\hat{\mu}$ that use $(\omega_1, \ldots, \omega_n)$ as sampling weights in the computation of the estimators.*

3. *Repeat steps (1)-(2) B times, where B is a large number. For example $B = 500$.*

4. *Use the empirical distribution of $\sup_{u \in \mathcal{U}_V} r_n |\widetilde{\Delta}^*_\mu(u) - \widehat{\Delta}^*_\mu(u)|$ across the S repetitions to approximate the bootstrap law of $Z_\infty(\mathcal{U}_V) = \sup_{u \in \mathcal{U}_V} |Z_\infty(u)|$.*

To state the bootstrap validity result formally, we follow the notation and definitions in van der Vaart and Wellner (1996). Let D_n denote the data vector and let $\mathrm{B}_n = (\omega_1, \ldots, \omega_n)$ be the vector of bootstrap weights. Consider a random element $\widetilde{Z}_n = Z_n(\mathrm{D}_n, \mathrm{B}_n)$ in a normed space $\mathbb{D}$. We say that the bootstrap law of $\widetilde{Z}_n$ consistently estimates the law of some tight random element Z_∞ and write $\widetilde{Z}_n \rightsquigarrow_{\mathrm{P}} Z_\infty$ if

$$\sup_{h \in \mathrm{BL}_1(\mathbb{D})} |\mathbb{E}_{\mathrm{B}_n} h(\widetilde{Z}_n) - \mathbb{E}_{\mathrm{P}} h(Z_\infty)| \to_{\mathrm{P}} 0,$$

where $\mathrm{BL}_1(\mathbb{D})$ denotes the space of functions with Lipschitz norm at most 1; $\mathbb{E}_{\mathrm{B}_n}$ denotes the conditional expectation with respect to B_n given the data D_n; $\mathbb{E}_{\mathrm{P}}$ denotes the conditional expectation with respect to P, the distribution of the data D_n; and $\to_{\mathrm{P}}$ denotes convergence in (outer) probability.

We assume that the bootstrap weights satisfy:

S.7. $\mathrm{B}_n = (\omega_1, \ldots, \omega_n)$ is an exchangeable, nonnegative random vector, which is independent of the data D_n, such that for some $\epsilon > 0$,

$$\sup_n \mathrm{E}[\omega_i^{2+\epsilon}] < \infty, \quad n^{-1} \sum_{i=1}^n (\omega_i - \bar{\omega})^2 \to_{\mathrm{P}} 1, \quad \bar{\omega} \to_{\mathrm{P}} 1,$$

where $\bar{\omega} = n^{-1} \sum_{i=1}^{n} \omega_i$.[1]

The next result is a consequence of the functional delta method for the exchangeable bootstrap.

Theorem 8 (Bootstrap FCLT for $\widehat{\Delta}_{\mu}^{*}$) *Let $\mathcal{V}$ be any compact set of $\mathcal{D}^*$ and $\mathcal{U}_{\mathcal{V}} = \{u \in \mathcal{U} : \Delta^*(u) \in \mathcal{V}\}$. Suppose that the assumptions of Theorem 7 and S.7 hold, $a_n(\tilde{\Delta} - \hat{\Delta}) \rightsquigarrow_{\mathrm{P}} G_\infty$ in $\ell^\infty(B(\mathcal{X}))$, and $\int_{\mathcal{X}} 1\{\Delta(x) \leqslant \delta\} b_n(\tilde{\mu}(x) - \hat{\mu}(x)) \rightsquigarrow_{\mathrm{P}} H_\infty(\delta)$ in $\ell^\infty(\mathcal{V})$. Then,*

$$r_n(\widetilde{\Delta_{\mu}^{*}} - \widehat{\Delta_{\mu}^{*}}) \rightsquigarrow_{\mathrm{P}} Z_\infty \text{ in } \ell^\infty(\mathcal{U}_{\mathcal{V}}).$$

3.4 Discrete variables

We consider the case where the covariate X includes discrete components. Without lost of generality we assume that the first component of X is discrete and the rest are continuous. Accordingly, we do the partition $X = (D, C)$. Let $\mathcal{X}_{c|d}$ denote the support of C conditional on $D = d$, $\mathcal{X}_d$ denote the support of D, $\mu_{c|d}$ the probability measure of C conditional on $D = d$, and $\mu_d(d) = \mathrm{P}(D = d)$. We continue denoting by $d_x = \dim(X)$ and by $\mathcal{D}^*$ the set of regular values of Δ on $\mathcal{X} := \cup_{d \in \mathcal{X}_d} \{d\} \times \mathcal{X}_{c|d}$.

We adjust $S.1$–$S.6$ to hold conditionally at each value of the discrete covariate.

$S.1'$. For any $d \in \mathcal{X}_d$, the set $\mathcal{X}_{c|d}$ is open and its closure $\bar{\mathcal{X}}_{c|d}$ is compact. There exists an open set $B(\mathcal{X}_{c|d})$ containing $\bar{\mathcal{X}}_{c|d}$ such that $c \mapsto \Delta(d, c)$ is $\mathcal{C}^1$ on $B(\mathcal{X}_{c|d})$ and $c \mapsto \mu_{c|d}(c)$ is $\mathcal{C}^0$ on $B(\mathcal{X}_{c|d})$.

$S.2'$. For any $d \in \mathcal{X}_d$ and any regular value δ of Δ on $\bar{\mathcal{X}}_{c|d}$, $\mathcal{M}_{\Delta|d}(\delta) := \{c \in B(\mathcal{X}_{c|d}) : \Delta(d, c) = \delta\}$ is either a $(d_x - 2)-$ manifold without boundary on $\mathbb{R}^{d_x - 1}$ of class $\mathcal{C}^1$ with finite number of connected branches or an empty set. $\mathcal{M}_{\Delta}(\delta) := \cup_{d \in \mathcal{X}_d} \mathcal{M}_{\Delta|d}(\delta)$ is also a $(d_x - 2)-$ manifold without boundary on $\mathbb{R}^{d_x - 1}$ of class $\mathcal{C}^1$ or an empty set.

$S.3'$. Let $\mathcal{B}(\mathcal{X}) := \cup_{d \in \mathcal{X}_d} \{d\} \times B(\mathcal{X}_{c|d})$, $\mathbb{F}_0$ denote a set of uniformly bounded con-

[1] A sequence of random variables $\omega_1, \omega_2, ..., \omega_n$ is exchangeable if for any finite permutation σ of the indices $1, 2, ..., n$ the joint distribution of the permuted sequence $\omega_{\sigma(1)}, \omega_{\sigma(2)}, ..., \omega_{\sigma(n)}$ is the same as the joint distribution of the original sequence.

tinuously differentiable functions on $B(\mathcal{X})$ equipped with sup-norm, and $\mathbb{F}_M$ denote the set of all L^1-integrable functions on $B(\mathcal{X})$ uniformly bounded by 1 in terms of absolute value. Suppose there exists a class of functions $\mathbb{F}_1$ defined on $\mathcal{B}(\mathcal{X})$ such that:

(1) $\mathbb{F}_0 \subseteq \mathbb{F}_1$.

(2) Given $\delta \in \mathcal{D}^*$, for any $H \in \mathbb{H}_0$, H is continuous on $\mathbb{F}_{\mathcal{M}_\Delta}(\delta) := \{1(\tilde{\Delta} \leqslant \delta) : \tilde{\Delta} \in \mathbb{F}_1\}$, where $\mathbb{H}_0$ is the set of all bounded linear operators on $\mathbb{F}_M$ equipped with the norm $L^{*\infty}$: $|H|_{L^{*\infty}} = \sup_{f \in \mathbb{F}_M, f \neq 0} |H(f)|$.

S.4'. The estimator $\hat{\Delta}$ of Δ obeys a functional central limit theorem, namely,

$$a_n(\hat{\Delta} - \Delta) \rightsquigarrow G_\infty \text{ in } \ell^\infty(\mathcal{B}(\mathcal{X})),$$

where a_n is a sequence such that $a_n \to \infty$ as $n \to \infty$, and G_∞ is a tight process that has uniformly continuous sample paths over $\mathcal{B}(\mathcal{X})$ a.s.

S.5'. For any $d \in \mathcal{X}_d$, the gradient with respect to c, $c \mapsto \nabla_c \hat{\Delta}(d, c)$, exists, is continuous at each $c \in \mathcal{B}(\mathcal{X}_{c|d})$, and

$$\sup_{(d,c) \in B(\mathcal{X})} \|\nabla_c \hat{\Delta}(d, c) - \nabla_c \Delta(d, c)\| \to_{\mathrm{P}} 0.$$

S.6'. The functions $d \mapsto \hat{\mu}(d)$ and $c \mapsto \hat{\mu}_{c|d}(c)$ are measures over $\mathcal{X}_d$ and $\mathcal{X}_{c|d}$ obeying in $\ell^\infty(\mathcal{V})$,

$$b_n \left[\sum_{d \in \mathcal{X}_d} \hat{\mu}_d(d) \int_{\mathcal{X}_{c|d}} 1\{\Delta(d, c) \leqslant \delta\} \hat{\mu}_{c|d}(c) - \sum_{d \in \mathcal{X}_d} \mu_d(d) \int_{\mathcal{X}_{c|d}} 1\{\Delta(d, c) \leqslant \delta\} \mu_{c|d}(c) \right] \rightsquigarrow H_\infty(\delta),$$

as an stochastic process indexed by $\delta \in \mathcal{V}$, where $\mathcal{V}$ is any compact subset of $\mathcal{D}^*$, b_n is a sequence such that $b_n \to \infty$ as $n \to \infty$, and

$$H_\infty(\delta) = \sum_{d \in \mathcal{X}_d} H_{d,\infty}(d) \int_{\mathcal{X}_{c|d}} 1\{\Delta(d, c) \leqslant \delta\} \mu_{c|d}(c) + \sum_{d \in \mathcal{X}_d} \mu(d) H_{c|d,\infty}(1\{\Delta \leqslant \delta\}),$$

where $H_{d,\infty}$ and $H_{c|d,\infty}$ are random elements.

Comment 3.4.1 (S.6') *The estimator of the measure μ is $\widehat{\mu} := \hat{\mu}_d \times \hat{\mu}_{c|d}$. For example, when $\widehat{\mu}$ is the empirical measure for the entire population from a random sample, $b_n = \sqrt{n}$, $H_{d,\infty}$ is a multivariate normal distribution with zero mean and covariance matrix with typical (i,j)-element equal to $1(i = j)\mu_d(i) - \mu_d(i)\mu_d(j)$, and $H_{c|d,\infty}$ is a $\mu_{c|d}$-Brownian Bridge.*

The next lemma generalizes Lemmas 21 and 24 to the case where X include discrete components.

Lemma 25 (Properties of $F_{\Delta,\mu}$ and Δ^*_μ with discrete X) *Suppose that S.1' and S.2' hold. Then, $\delta \mapsto F_{\Delta,\mu}(\delta)$ is differentiable at any $\delta \in \mathcal{D}^*$, with derivative function $f_{\Delta,\mu}(\delta)$ defined as:*

$$f_{\Delta,\mu}(\delta) := \partial_\delta F_{\Delta,\mu}(\delta) = \sum_{d \in \mathcal{X}_d} \mu_d(d) \int_{\mathcal{M}_{\Delta|d}(\delta)} \frac{\mu_{c|d}(c)}{\|\nabla_c \Delta(d,c)\|} dVol.$$

Define $\mathbb{D} := \mathbb{F}_1 \times \mathbb{H}_0$ and $\mathbb{D}_0 := \mathbb{F}_0 \times \mathbb{H}_0$. Under S.1'-S.3',

(1) For any $\delta \in \mathcal{D}^*$, the map $F_{\Delta,\mu}(\delta) : \mathbb{D} \to \mathbb{R}$ is Hadamard differentiable at (Δ, μ) tangentially to $\mathbb{D}_0$, with derivative defined by:

$$\begin{aligned}(G,H) \mapsto \partial_{\Delta,\mu} F_{\Delta,\mu}(\delta)[G,H] := & - \sum_{d \in \mathcal{X}_d} \mu_d(d) \int_{\mathcal{M}_{\Delta|d}(\delta)} \frac{G(d,c)\mu_{c|d}(c)}{\|\nabla_c \Delta(d,c)\|} dVol \\ & + \sum_{d \in \mathcal{X}_d} H_d(d) \int_{\mathcal{X}_{c|d}} 1\{\Delta(d,c) \leqslant \delta\} \mu_{c|d}(c) \\ & + \sum_{d \in \mathcal{X}_d} \mu_d(d) H_{c|d}(1\{\Delta \leqslant \delta\}),\end{aligned}$$

as a map from $\mathbb{D}_0$ to $\mathbb{R}$, where $H := (H_d, H_{c|d})$ is defined as a bounded functional operator which maps $g \in L^2(\mathcal{X})$ to:

$$H(g) := \sum_{d \in \mathcal{X}_d} H_d(d) \int_{\mathcal{X}_{c|d}} g(d,c)\mu_{c|d}(c) + \sum_{d \in \mathcal{X}_d} \mu_d(d) H_{c|d}(g(d,c)),$$

where $d \mapsto H_c(d)$ is a linear function and $g \mapsto H_{c|d}(g)$ is a bounded linear operator.

(2) For any $u \in \{\tilde{u} : \Delta_\mu^*(\tilde{u}) \in \mathcal{D}^*\}$, *the map* $\Delta_\mu^*(u) : \mathbb{D} \to \mathbb{R}$ *is Hadamard differentiable at* (Δ, μ) *tangentially to* $\mathbb{D}_0$, *with derivative defined by:*

$$(G, H) \mapsto \partial_{\Delta,\mu}\Delta_\mu^*(u)[G, H] := -\frac{\partial F_{\Delta,\mu}(\Delta_\mu^*(u))[G, H]}{f_{\Delta,\mu}(\Delta_\mu^*(u))}.$$

as a map from $\mathbb{D}_0$ *to* $\mathbb{R}$.

We are now ready to derive a functional central limit theorem for the estimator of the SPE-function. As in Theorem 7, let $r_n := a_n \wedge b_n$, the slowest of the rates of convergence of $\widehat{\Delta}$ and $\widehat{\mu}$, where $r_n/a_n \to s_\Delta \in [0, 1]$ and $r_n/b_n \to s_\mu \in [0, 1]$.

Theorem 9 (FCLT for $\widehat{\Delta_\mu^*}(u)$ with discrete X) *Suppose that S.1'-S.6' hold,* $\hat{\Delta} \in \mathbb{F}_1$ *with probability approaching 1, and* $\mathbb{F}_{\mathcal{M}_\Delta(\delta)}$ *is* μ*-Donsker for any* $\delta \in \mathcal{D}^*$. *In* $\ell^\infty(\mathcal{U}_\mathcal{V})$, *with* $\mathcal{U}_\mathcal{V} = \{u \in \mathcal{U} : \Delta_\mu^*(u) \in \mathcal{V}\}$,

$$r_n(\hat{\Delta}_{\hat{\mu}}^*(u) - \Delta_\mu^*(u)) \rightsquigarrow \partial_{\Delta,\mu}\Delta_\mu^*(u)[s_\Delta G_\infty, s_\mu H_\infty], \tag{3.4.1}$$

as a stochastic process indexed by $u \in \mathcal{U}_\mathcal{V}$.

Comment 3.4.2 (Bootstrap FCLT for $\widehat{\Delta_\mu^*}(u)$ with discrete X) *The exchangeable bootstrap law is consistent to approximate the distribution of the limit process in* (3.4.1) *by the same argument as in Theorem 8, replacing S.1-S.6 by S.1'-S.6'. Accordingly, we do not repeat the statement here.*

3.5 Numerical Examples

3.5.1 Monte-Carlo Simulations

We evaluate the accuracy of the asymptotic approximations to the distribution of the empirical SPE in small samples using numerical simulations. In particular, we compare pointwise 95% confidence intervals for the SPE based on the asymptotic and exact distributions of the empirical SPE. The exact distribution is approximated numerically by

simulation. The asymptotic distribution is obtained analytically from the CLT of Theorem 7, and approximated by bootstrap using Theorem 8. We consider two simulation designs where the limit process in Theorem 7 has a convenient closed-form analytical expression. The designs differ on whether the PE-function $x \mapsto \Delta(x)$ has critical points or not. We hold fix the values of the covariate vector X in all the calculations, and accordingly we treat the measure μ as known. For the bootstrap inference, we use empirical bootstrap with $B = 3,000$ repetitions. All the results are based on $S = 3,000$ Monte Carlo simulations with a sample size $n = 1,000$.

Example 14 (No critical points) *We consider the PE-function*

$$\Delta(x) = x_1 + x_2, \quad x = (x_1, x_2),$$

with covariate vector X uniformly distributed in $\mathcal{X} = (-1,1) \times (-1,1)$. The corresponding SPE is

$$\Delta^*_\mu(u) = 2(\sqrt{2u} - 1)1(u \leqslant 1/2) + 2(1 - \sqrt{2(1-u)})1(u > 1/2),$$

where $1(\cdot)$ denotes the indicator function, and we use that $\Delta(X)$ has a triangular distribution with parameters $(-2, 0, 2)$. Figure 3-1 plots $x \mapsto \Delta(x)$ on $\mathcal{X}$, and $u \mapsto \Delta^*_\mu(u)$ on $(0,1)$. Here we see that $x \mapsto \Delta(x)$ does not have critical values, and that $u \mapsto \Delta^*_\mu(u)$ is a smooth function.

To obtain an analytical expression of the limit $Z_\infty(u)$ of Theorem 7, we make the following assumption on the distribution of the estimator of the PE:

$$\sqrt{n}(\hat{\Delta}(x) - \Delta(x)) \sim G_\infty(x) = \Delta(x)Z,$$

where Z is a standard normal random variable independent of (X_1, X_2). This assumption is analytically convenient because after some calculations we find that

$$Z_\infty(u) \sim -\Delta^*_\mu(u)Z,$$

so that $\hat{\Delta}^*_\mu(u) \overset{a}{\sim} N(\Delta^*_\mu(u), \Delta^*_\mu(u)^2/n)$, where $\overset{a}{\sim}$ denotes asymptotic approximation to the

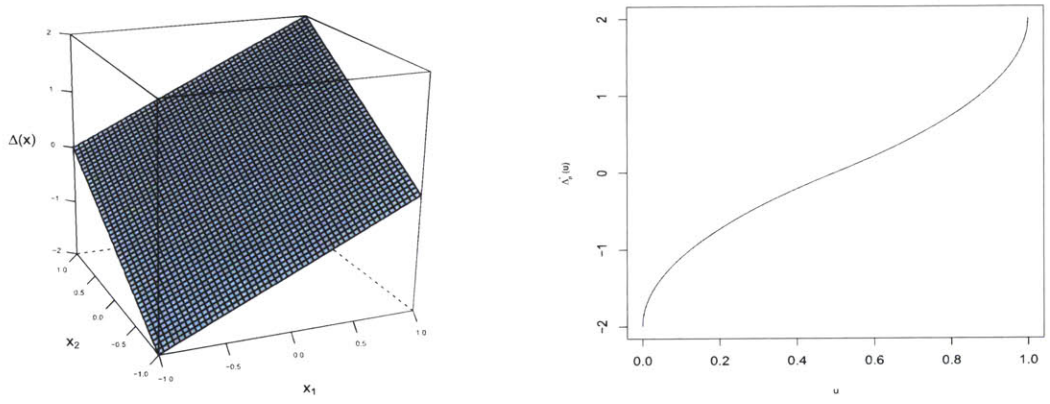

Figure 3-1: PE-function and SPE-function in Design 1. Left: PE function $x \mapsto \Delta(x)$. Right: SPE function $u \mapsto \Delta^*_\mu(u)$.

Table 3.1: Monte-Carlo example 14, $n = 1000$, Monte-Carlo rounds = 3000,bootstrap rounds=3000.

quantile	$u = 0.1$	$u = 0.2$	$u = 0.3$	$u = 0.4$	$u = 0.5$
theoretical st.d	0.0444	0.0410	0.0383	0.0362	0.0350
monte-carlo st.d	0.0441	0.0408	0.0383	0.0359	0.0360
rej. rate(T)	0.050	0.057	0.050	0.051	0.062
rej. rate(B)	0.054	0.037	0.085	0.070	0.044

quantile	$u = 0.6$	$u = 0.7$	$u = 0.8$	$u = 0.9$
theoretical st.d	0.0377	0.0411	0.0456	0.0529
monte-carlo st.d	0.0381	0.0417	0.0467	0.0536
rej. rate(T)	0.053	0.058	0.054	0.054
rej. rate(B)	0.049	0.046	0.062	0.050

exact distribution.

Table 3.1 compares the standard deviation of the empirical SPE in samples of size $n = 1,000$ *with the asymptotic standard deviation* $|\Delta^*_\mu(u)|/\sqrt{n}$ *at the quantile indexes* $u \in \{0.1, 0.2, \ldots, 0.9\}$. *The asymptotic approximation is very close to the exact standard deviation. We also find that 95% confidence intervals constructed using the asymptotic approximation,* $\hat{\Delta}^*_\mu(u) \pm 1.96|\Delta^*_\mu(u)|/\sqrt{n}$, *have coverage probabilities close to the nominal level at all quantiles. These asymptotic confidence intervals are not feasible in general, because* $\Delta^*_\mu(u)$ *is unknown or more generally it is not possible to characterize analytically the distribution of* $Z_\infty(u)$. *In practice we propose to approximate this distribution by bootstrap. The table also shows that the empirical coverages of bootstrap 95% confidence intervals are close to their nominal levels at all quantiles. In this case, empirical bootstrap is equivalent to redraw the realizations of the empirical PE* $\hat{\Delta}(x)$ *with replacement. Figure 3-2 presents confidence bands of estimated sort-curve using theoretical, monte-carlo and bootstrapped bands. These bands are close to each other with indistinguishable differences.*

Example 15 (Critical points) *We consider the PE-function*

$$\Delta(x) = x^3 - 3x,$$

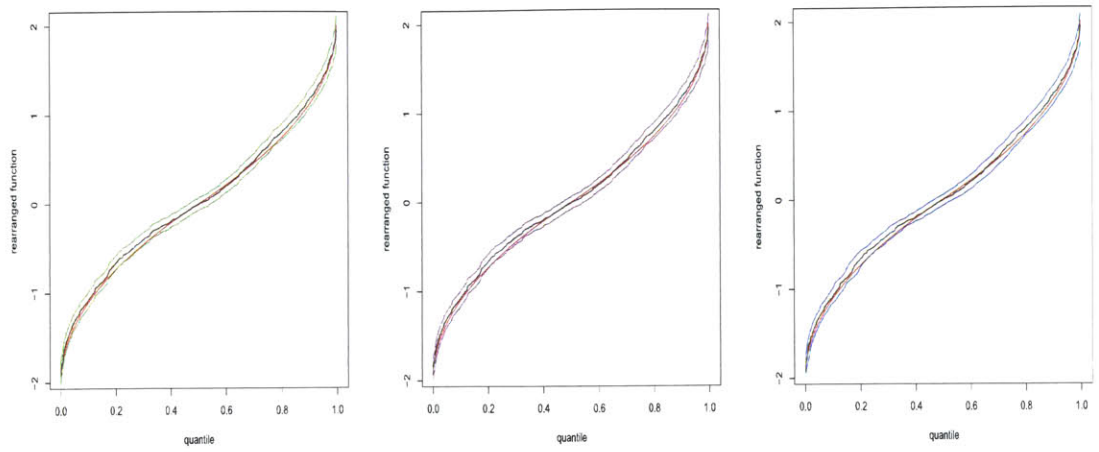

Figure 3-2: Confidence bands for SPE in Design 1. Left: Asymptotic bands. Center: Simulation finite-sample Bounds. Right: Bootstrap bands.

with covariate X uniformly distributed in $\mathcal{X} = (-3, 3)$. Figure 3-3 plots $x \mapsto \Delta(x)$ on $\mathcal{X}$, and $u \mapsto \Delta^*_\mu(u)$ on $(0, 1)$.[2] *Here we see that $x \mapsto \Delta(x)$ has two critical points at $x = -1$ and $x = 1$ with corresponding critical values at $\delta = 2$ and $\delta = -2$. The SPE-function $u \mapsto \Delta^*_\mu(u)$ has two kinks at $u = 1/6$ and $u = 5/6$, the Δ^*_μ pre-images of the critical values.*

Fig. 3-3 suggests the convergence of the empirical SPE is going to be irregular at the kinks, and might be slow in neighborhoods around the kinks. To show more evidence about these convergence issues, we compare the exact and asymptotic distribution of the empirical SPE. To obtain these distributions, we make a convenient assumption on the distribution of the estimator of the PE:

$$\sqrt{n}(\hat{\Delta}(x) - \Delta(x)) \sim G_\infty(x) = (x/2)^2 Z,$$

where Z is a standard normal random variable independent of X.

Table 3.2 compares the standard deviation of the empirical SPE in samples of size $n = 1,000$ with the asymptotic standard deviation at the quantile indexes $u \in \{0.1, 0.2, \ldots, 0.9\}$.

[2] We obtain $u \mapsto \Delta^*_\mu(u)$ analytically using the characterization of Chernozhukov, Fernandez-Val, and Galichon (2010) for the univariate case.

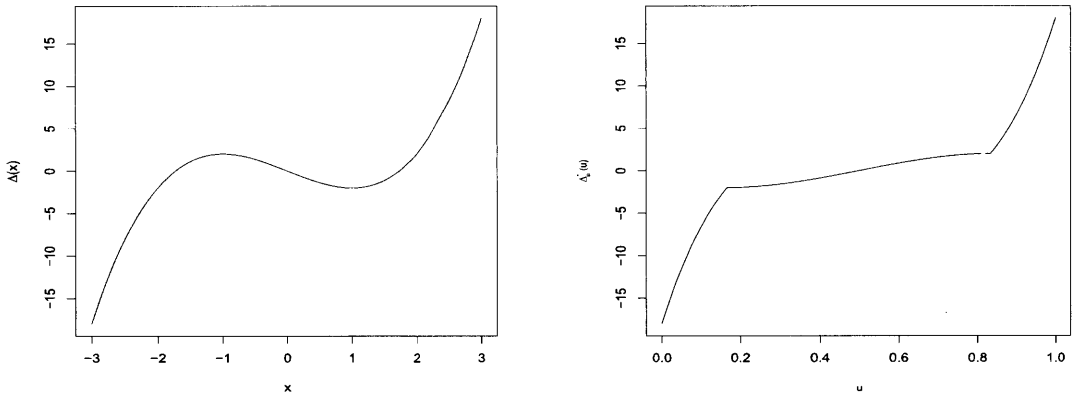

Figure 3-3: PE-function and SPE-function in Design 2. Left: PE function $x \mapsto \Delta(x)$. Right: SPE function $u \mapsto \Delta^*_\mu(u)$.

Table 3.2: Monte-Carlo example 15, $n = 1000$, Monte-Carlo rounds = 3000,bootstrap rounds=3000.

quantile	$u = 0.1$	$u = 0.2$	$u = 0.3$	$u = 0.4$	$u = 0.5$
theoretical st.d	0.820	0.0268	0.0852	0.129	0.144
monte-carlo st.d	0.799	0.0336	0.0847	0.123	0.140
rej. rate(T)	0.0440	0.0677	0.0497	0.0393	0.0490
rej. rate(B)	0.036	0.081	0.068	0.043	0.068

quantile	$u = 0.6$	$u = 0.7$	$u = 0.8$	$u = 0.9$
theoretical st.d	0.130	0.0860	0.0275	0.818
monte-carlo st.d	0.128	0.0872	0.0304	0.822
rej. rate(T)	0.0517	0.0587	0.0797	0.0577
rej. rate(B)	0.081	0.056	0.011	0.11

The asymptotic approximation is close to the exact standard deviation with the largest differences occurring at $u = 0.2$ *and* $u = 0.8$*, the quantiles that are closer to the kink points at* $1/6$ *and* $5/6$*. We also find that pointwise 95% confidence intervals constructed using the asymptotic distribution and empirical bootstrap have coverage probabilities close to the nominal level, with the largest distortions occurring at the quantiles* $u = 0.2$ *and* $u = 0.8$*. Interestingly, while the asymptotic approximation undercovers the SPE at the quantiles close to the kink points, the bootstrap approximation is conservative.*

3.5.2 Empirical Example: Women Labor Supply and the Number of Children

In this subsection, we follow Angrist and Evans (1998) and Angrist (2001) to examine how women labor supply is affected by the number of children. We employ 1980 Census Public Use Micro Samples (PUMS). Following the estimation strategies in the above two papers, we consider apply sorting technique discussed in this paper. We provide graphs on Sorted partial effects on number of children on women labor supply verse traditional mean average partial effect. The basic model consists information about whether a

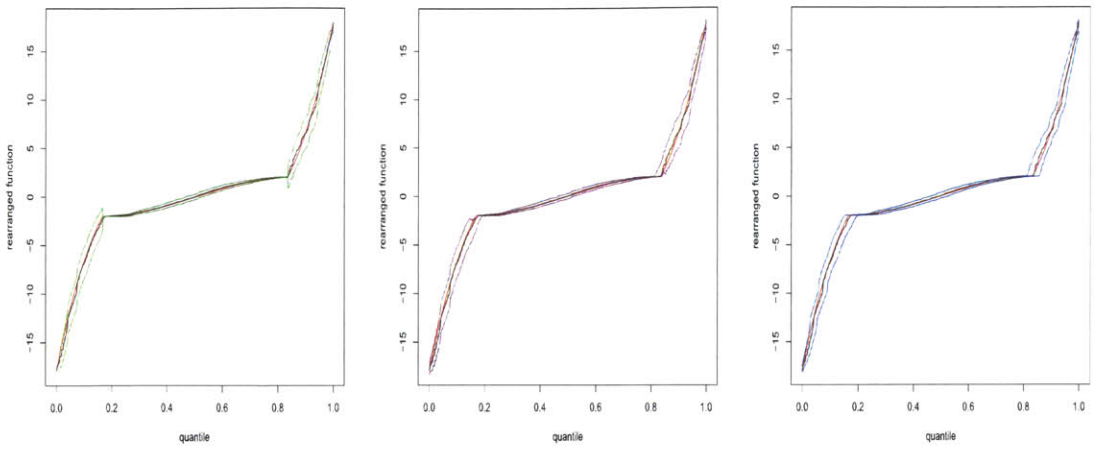

Figure 3-4: Confidence bands for SPE in Design 2. Left: Asymptotic bands. Center: Simulation finite-sample Bounds. Right: Bootstrap bands.

female is working, how many weeks she work per year, whether she has three or more kids, age of first birth, gender of the first and the second child, family income, father income, as well as a set of demographic dummies.

Angrist (2001) considers different models estimating the partial effect of having more than two children on women labor based on different variables. We replicate the results and apply sorting to probit and tobit models. In addition, we also add interaction terms to make the function form more flexible.

More specifically, consider the following model:

$$workedm_i = 1(\alpha_1 D_i + \beta_1 X_i + \gamma_1 X_i D_i + \epsilon_{1i} \geqslant 0),$$

$$weeksm_i = \alpha_2 D_i + \beta_2 X_i + \gamma_2 X_i D_i + \epsilon_{2i} | workedm_i \geqslant 0.$$

$workedm_i$ indicates for employment status, $weeksm_i$ indicates for the number of working weeks per year, and D_i indicates for having at least 3 kids.

We use probit model to estimate the effect of D_i on $E[Pr(workedm_i = 1|D_i, X_i)]$, i.e., $E[Pr(workedm_i = 1|D_i = 1, X_i)] - E[Pr(workedm_i = 1|D_i = 0, X_i)]$. We use tobit

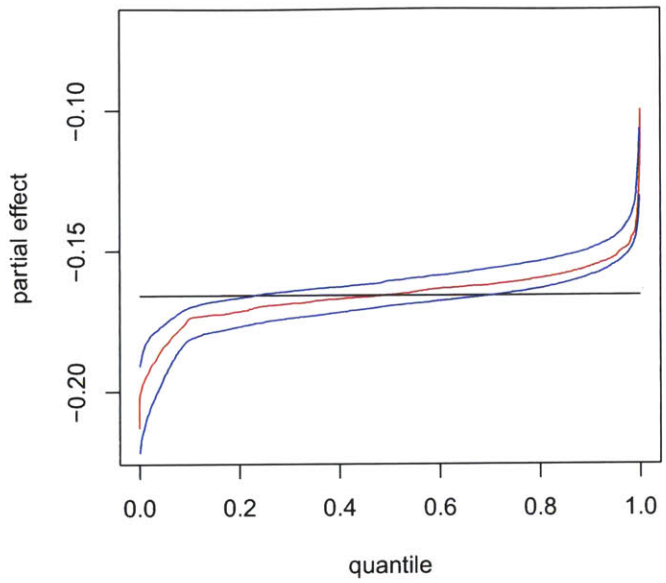

Figure 3-5: The probability of working changing from having less than 3 children to at least 3 children ($Pr(E_i|workedm_i = 1, X_i) - Pr(E_i|workedm_i = 0, X_i)$): Black-APE, Red-Sorted curve of partial effects, Blue-confidence bands for Rearranged curve. Top graph: Basic probit model. Bottom graph: Probit model with interaction terms of D_i and X_i. Blue bands are 95% pointwise confidence bands.

model to estimate the effect of D_i on $E[weeksm_i|D_i = 1, X_i]$, i.e., $E[weeksm_i|D_i = 1, X_i] - E[weeksm_i|D_i = 0, X_i]$. The X_i vector includes other covariates. The results are presented below in Figures 3-5 and 3-6.

We can observe that there exists heterogeneity of individual partial effects in both probit and tobit models. In Figure 3-5, there exists 55% of the population with effects higher than APE. The confidence bands shows that there exists 35% of the population with partial effect significantly larger than APE. In the lower tail, there exists 10% of the population much that are much more likely to drop the work force when the number of children increases to be $\geqslant 3$. In Figure 3-6, similar patterns can be seen that the lower tail of the population tended to reduce more work hours compared to the rest of the population.

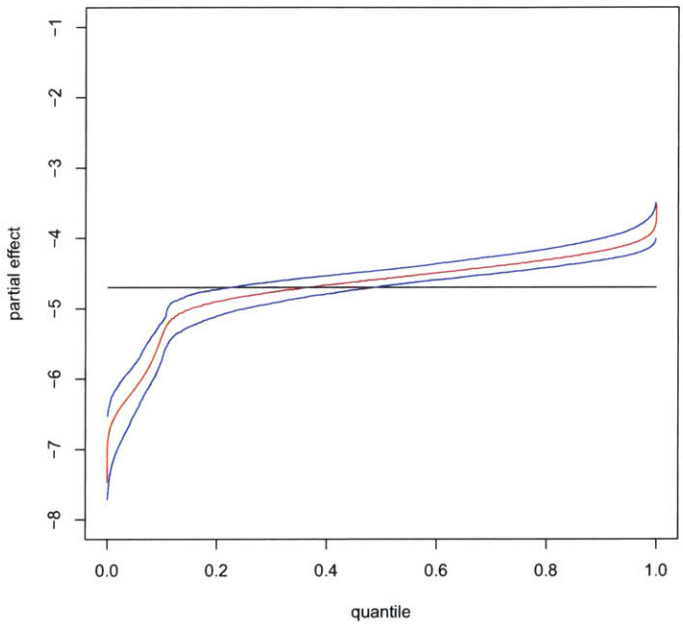

Figure 3-6: The number of hours of working per week changing from having more than 2 children to less than or equal to 2 children ($E[weeksm_i|D_i = 1, X_i] - E[weeksm_i|D_i = 0, X_i]$):Black-APE, Red-Sorted curve of partial effects, Blue-confidence bands for Rearranged curve. Left: Basic tobit model. Right: Tobit model with interaction terms of D_i and X_i. Blue bands are 95% pointwise confidence bands.

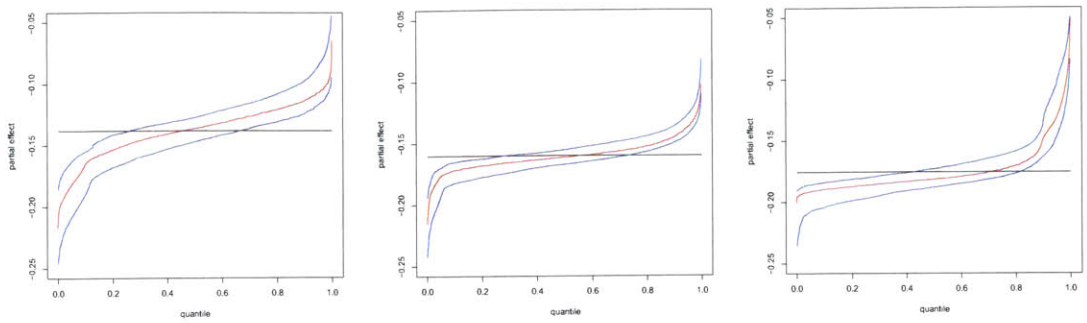

Figure 3-7: The probability of working changing from having less than 3 children to at least 3 children ($Pr(workedm_i|D_i = 1, X_i) - Pr(workedm_i|D_i = 0, X_i)$): Probit model with interaction terms. Left:Women's education less than high school. Middle: Women's education equals to high school. Right: Women's education above high school. Blue bands are 95% pointwise confidence bands.

The results in the above figures draw a similar, but more detailed picture as Angrist and Evans (1998). In table 9 of Angrist and Evans (1998), they consider to evaluate APEs of the subgroups based on woman's education level. Low education group consists women with less than high school education, median education group consists women with high school education, and high education group consists women with more than high school education, e.g., college and above. Thus, we further apply our technique to investigate these subgroups. We find an decreasing level of APE, similar to Angrist and Evans (1998). However, we find more heterogeneity at quantiles of the partial effects by comparing the three different groups. In Figures 3-7, thought extreme tails are similar, the APE is pushed down primarily by the population in mid-quantiles. In Figures 3-8, The drop of working hours in upper tails is primarily between low education and median education groups, while the lower tail is primarily changed between median education and high education groups.

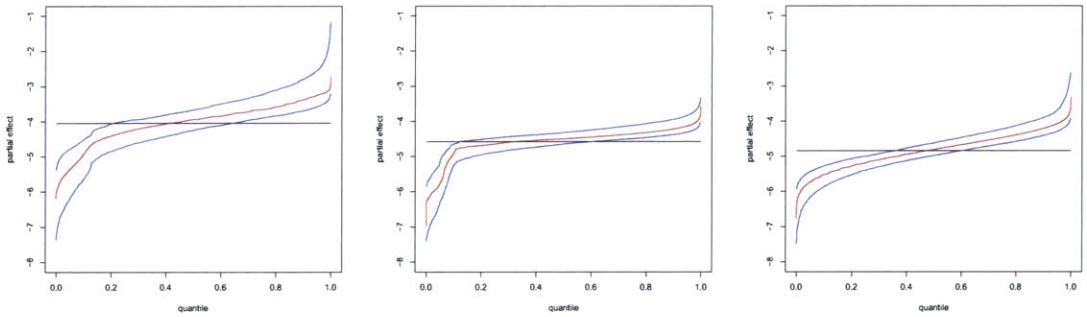

Figure 3-8: The number of hours of working per week changing from having more than 2 children to less than or equal to 2 children ($E[weeksm_i|D_i = 1, X_i] - E[weeksm_i|D_i = 0, X_i]$), conditional on subgroups:Tobit model with interaction terms. Left:Women's education less than high school. Middle: Women's education equals to high school. Right: Women's education above high school. Blue bands are 95% pointwise confidence bands.

3.6 Conclusion

The sorting of partial effects is a new method to report empirical results as well as a new equipment to analyze casual effect in each quantile in the entire population, compare to the traditional mean-variance analysis. This paper develops the large sample asymptotic property of the general multi-dimensional rearrangment operator. We establishes asymptotical property of the sorting operator via functional delta method. The validity of the theorems relies only on a set of weak regularity assumptions on the behavior of the estimated function Δ, the distribution of explanatory variables x, and the shape of the boundary of the domain $\mathcal{X}$. These assumptions are reasonable for most cases in econometric analysis. Furthermore, we provide theorems for inference. Although the numerical calculation of the confidence bands can be difficult, we prove and demonstrate that bootstrap confidence bands work as well as the ideal confidence bands. We provide simulation and empirical results to illustrate how our technique can be applied in practice and why it may be better than the traditional way of reporting empirical results.

Bibliography

[1] Alberto Abadie, Joshua Angrist and Guido Imbens, 2002. "Instrumental Variables Estimates of the Effect of Subsidized Training on the Quantiles of Trainee Earnings," Econometrica, Econometric Society, vol. 70(1), pp. 91-117.

[2] Angrist, J. and William Evans, 1998, "Children and their parents's labor supply: Evidence from exogenous variation in Family Size", AER, vol. 88, No. 3, pp. 450-477.

[3] Angrist, J., 2001, "Estimation of Limited Dependent Variable Models With Dummy Endogenous Regressors: Simple Strategies for Empirical Practice".

[4] Angrist, J., Chernozhukov, V. and Fernandez-Val, I. (2006), Quantile Regression under Misspecification, with an Application to the U.S. Wage Structure. Econometrica, 74: 539-563.

[5] Joel G. Broida and S. Gill Williamson (1989) A Comprehensive Introduction to Linear Algebra, chapter 4.6, Cauchy-Binet theorem, pp 208-14, Addison-Wesley ISBN 0-201-50065-5.

[6] Victor Chernozhukov, Ivan Fernandez-Val, Alfred Galichon, "Quantile and Probability Curve Without Cross", 2006.

[7] Victor Chernozhukov, Ivan Fernandez-Val, Alfred Galichon, "Improving Point and Interval Estimates of Monotone Functions by rearrangement", 2007.

[8] V. Chernozhukov, I. Fernandez-Val and A. Galichon, 2009. "Improving point and interval estimators of monotone functions by rearrangement," Biometrika, Oxford University Press for Biometrika Trust, vol. 96(3), pages 559-575.

[9] Victor Chernozhukov, Ivan Fernandez-Val and Alfred Galichon, 2010 "Rearranging Edgeworth-Cornish-Fisher expansions," Economic Theory, Centre for Microdata Methods and Practice, Institute for Fiscal Studies.

[10] Roger Koenker, Quantile Regression, Econometric Society Monographs, 2005.

[11] James Munkres, "Analysis on Manifolds", Westview Press, 1990.

[12] Micheal Spivak, "Calculus on Manifold", MIT press, 1990, pages 109-134.

[13] Aad W.van der Vaart and Jon a.Wellner, "Weak convergence and Empirical Processes", Springer, 2000.

Appendix A

Proofs

A.1 Proofs in Chapter 1

A.1.1 Proofs in Section 1.4

Proof of Lemma 1

By assumption, for any $v \in \mathbb{R}^d$, $||v||_2 = 1$, $||G_0(\beta_0)v||_\infty \leqslant K_0$. If $m = O(\exp(n))$, by statement (2) and (4) of C.5, $||\lambda^*||_1 \leqslant K_\lambda^u + o(\sqrt{\frac{\log(m)}{n}})$ is bounded from the above.

Thus, $vG_0(\beta_0)'\Omega_0^{-1}G_0(\beta_0)v = (\lambda^*)'G_0(\beta_0)v \leqslant ||\lambda^*||_1||G_0(\beta_0)v||_\infty \leqslant K_0||\lambda^*||_1$ is bounded from the above. That is to say, the maximal eigenvalue of $G_0(\beta_0)'\Omega_0^{-1}G_0(\beta_0)$ is bounded from the above.

Proof of Theorem 1

To prove Theorem 1, I follow the strategy in BCCH (2012). The proof of this theorem is divided into three steps. The first step provides proof for (1.4.4) and (1.4.5). The second step provides proof for consistency of $\hat{\beta}_L$. The third step proves (1.4.6) and (1.4.7).

Step 1: In this step, we establish bounds for $\hat{\lambda}(l) - \widetilde{\lambda}(l)$, $1 \leqslant l \leqslant m$.

For any vector $x \in \mathbb{R}^m$, define norm $||\cdot||_{1,n}$ of x as $||x||_{1,n} := \sum_{1 \leqslant j \leqslant m} |x_j||\gamma_j|$, and define semi-norm $||\cdot||_{2,n}$ of x as $x'\hat{\Omega}x$. For any $c > 0$ and set $T \subset \{1, 2, ..., m\}$, define the following quantities:

$$\kappa_c^2(T) := \min_{||\delta_{T^c}||_1 \leqslant c||\delta_T||_1, \delta \neq 0} \frac{s||\delta||_{2,n}^2}{||\delta_T||_1^2}, \tag{A.1.1}$$

$$\kappa_{c,n}^2(T) := \min_{||\delta_{T^c}||_{1,n} \leqslant c||\delta_T||_1, \delta \neq 0} \frac{s||\delta||_{2,n}^2}{||\delta_T||_{1,n}^2}, \tag{A.1.2}$$

where $s = |T|$.

Lemma 3 of Bickel, Ritov and Tsybakov (2009) proves that bounds on $\kappa(s, \hat{\Omega})$ imply a lower bound for $\kappa_c(T)$ and $\kappa_{c,n}(T)$. More specifically, for any positive integer s_1, Bickel, Ritov and Tsybakov (2009) shows that $\kappa_c(T)^2 \geqslant \kappa(s_1, \hat{\Omega})(1 - \bar{\epsilon}\sqrt{\frac{s\ \phi(s_1, \hat{\Omega})}{s_1\ \kappa(s_1, \hat{\Omega})}})$. For $s \leqslant s_n$ and $s_1 = s\log(n)$, the Assumption C.6 implies that

$$\kappa_c(T)^2 \geqslant \kappa_1 \left(1 - \bar{\epsilon}\sqrt{\frac{\kappa_2}{\log(n)\kappa_1}}\right),$$

which is bounded from below away from 0 as n approaches infinity. So $\kappa_{c,n}(T)^2 \geqslant \frac{a^2}{b^2}\kappa_c(T)$ $\geqslant \frac{a^2}{b^2}\kappa_1 \left(1 - \bar{\epsilon}\sqrt{\frac{\kappa_2}{\log(n)\kappa_1}}\right)$, which verifies that $\kappa_{c,n}(T)^2$ is bounded from below if $s \leqslant s_n$.

Let $\delta := \hat{\lambda} - \widetilde{\lambda}$. For $\widetilde{\lambda}$, let T be the set of indices of non-zero components of $\widetilde{\lambda}$. Below I establish non-asymptotic bounds for the solution to $\mathcal{P}$.

Lemma 26 (Bounds for LASSO Selector λ) *Given $v \in \mathbb{R}^d$, suppose $\widetilde{\lambda}$ is the sparse vector to be estimated and $\hat{\lambda}$ is the solution to the convex optimization problem $\mathcal{P}$. Let T be the set of indices of non-zero components in $\widetilde{\lambda}$. Assume that conditions C.1-C.3 and C.6-C.8 hold. Denote $\bar{\epsilon} = \frac{\epsilon+2}{\epsilon}$ Then, with probability at lease $1 - \alpha_n$,*

$$||\delta||_{1,n} \leqslant \bar{\epsilon}||\delta_T||_{1,n}, \tag{A.1.3}$$

and

$$||\delta||_{2,n} \leqslant \sqrt{s}\frac{2t_0(2+\epsilon)}{n\kappa_{\bar{\epsilon},n}(T)}. \tag{A.1.4}$$

Proof of Lemma 26

By definition, $\hat{\lambda}$ is the minimizer of the problem $\mathcal{P}$. Therefore, $\frac{t}{n}(||\delta_T||_{1,n} - ||\delta_{T^c}||_{1,n}) \geqslant \hat{Q}(\hat{\lambda}) - \hat{Q}(\widetilde{\lambda})$.

Meanwhile, $\hat{Q}(\hat{\lambda}) - \hat{Q}(\widetilde{\lambda})$ can be decomposed as:

$$\hat{Q}(\hat{\lambda}) - \hat{Q}(\widetilde{\lambda}) = \frac{1}{2}\delta'\hat{\Omega}\delta - \delta'\hat{\Omega}\widetilde{\lambda} + v'\hat{G}'\delta \geqslant \frac{1}{2}\delta'\hat{\Omega}\delta - ||\hat{S}(\widetilde{\lambda})||_\infty||\delta||_1. \tag{A.1.5}$$

By Assumption C.7, $\mathrm{P}(\frac{t_0}{n} \geqslant \max_{1\leqslant j\leqslant m}|\frac{\hat{S}_j(\widetilde{\lambda})}{\gamma_j}|) \geqslant 1 - \alpha_n$.

By inequality (A.1.5), $\frac{t}{n}||\widetilde{\lambda}||_{1,n} - \frac{t}{n}||\lambda||_{1,n} \geqslant |\hat{Q}(\hat{\lambda}) - \hat{Q}(\widetilde{\lambda})| \geqslant ||\hat{S}||_\infty||\delta||_1 \geqslant -\frac{t_0}{n}||\delta||_{1,n}$. By setting $t = (1+\epsilon)t_0$, we know that $(1+\epsilon)||\delta_T||_{1,n} - (1+\epsilon)||\delta_{T^c}|| \geqslant -||\delta_T||_{1,n} - ||\delta_{T^c}||_{1,n}$. Thus,

$$||\delta_{T^c}||_{1,n} \leqslant \bar{\epsilon}||\delta_T||_{1,n}.$$

Restarting with (A.1.5), again $\frac{t}{n}(||\delta_T||_{1,n} - ||\delta_{T^c}||_{1,n}) \geqslant \frac{1}{2}||\delta||_{2,n}^2 - \frac{t_0}{n}||\delta_T||_{1,n} - \frac{t_0}{n}||\delta_{T^c}||_{1,n}$.

Therefore, $\frac{1}{2}||\delta||_{2,n}^2 \leqslant \frac{t_0(2+\epsilon)}{n}||\delta_T||_{1,n} - \frac{t\epsilon}{n}||\delta_{T^c}||_{1,n} \leqslant \frac{t_0(2+\epsilon)\sqrt{s}}{n\kappa_{\bar{\epsilon},n}(T)}||\delta||_{2,n}$. Thus,

$$||\delta||_{2,n} \leqslant \frac{2t_0(2+\epsilon)\sqrt{s}}{n\kappa_{\bar{\epsilon},n}(T)}.$$

In the Lemma 26, I establish bounds for $\hat{\lambda} - \widetilde{\lambda}$ given a vector $v \in \mathbb{R}^d$. To obtain a just identified system of moment conditions, we can consider repeating the selection procedure (1.4.4) for $v = e_1, e_2, ..., e_d$. Define $\delta(l) := \hat{\lambda}(l) - \widetilde{\lambda}(l)$, $T_{0,l} = \{j|\widetilde{\lambda}(l)_j \neq 0\}$ and $T_0 := \cup_{1\leqslant l\leqslant d}T_{0,l}$.

For any $1 \leqslant l \leqslant d$, by Lemma 26, $||\delta(l)||_{2,n} \leqslant \frac{2t_0(2+\epsilon)\sqrt{s_n}}{n\kappa_{\bar{\epsilon},n}(T_0)}$. Also, $t_0 = \Phi^{-1}(1 - \frac{\alpha_n}{4md}) \leqslant$

$\sqrt{\log(\frac{4md}{\alpha_n})}$. Combining these inequalities, we immediately obtain (1.4.5):

$$\max_{l\leqslant l\leqslant d}||\delta(l)||_{2,n} \leqslant K'_\lambda\sqrt{\frac{s_n\log(\frac{4md}{\alpha_n})}{n}},$$

where $K'_\lambda := \frac{2(2+\epsilon)}{\kappa_{\bar{\epsilon},n}(T_0)}$.

For (1.4.4), by conclusions in Lemma 26, $||\delta_{T^c}(l)||_{1,n} \leqslant \bar{\epsilon}||\delta_T(l)||_{1,n}$. Therefore $||\delta(l)||_1 \leqslant \frac{b}{a}||\delta(l)||_{1,n} \leqslant \frac{b}{a}\frac{\sqrt{s_n}}{\kappa_{\bar{\epsilon},n}}||\delta(l)||_{2,n}$. The inequality (1.4.5) holds with constant $K_\lambda = \frac{b}{a\kappa_{\bar{\epsilon},n}}K'_\lambda$.

Step 2: In this step, we prove consistency of $\hat{\beta}_L$.

Let $\hat{\Lambda} = (\hat{\lambda}(1), \hat{\lambda}(2), ..., \hat{\lambda}(d))$, $\widetilde{\Lambda} = (\widetilde{\lambda}(1), \widetilde{\lambda}(2), ..., \widetilde{\lambda}(d))$ and $\Delta = (\delta(1)', \delta(2)', ..., \delta(d))$. By definition $\Delta = \hat{\Lambda} - \widetilde{\Lambda}$.

The GMM estimator $\hat{\beta}_L$ has the following property:

$$\hat{\Lambda}'\mathbb{E}_n[g(Z_i, \hat{\beta}_L)] = 0. \tag{A.1.6}$$

We prove that $\hat{\beta}_L$ is consistent. Let

$$\hat{q}_n(\beta) = \sum_{l=1}^{d}(\hat{\lambda}(l)'\mathbb{E}_n[g(Z_i, \beta)])^2,$$

$$q_n(\beta) = \sum_{l=1}^{d}(\widetilde{\lambda}(l)'\mathbb{E}_n[g(Z_i, \beta)])^2,$$

$$q_{n,0}(\beta) = \sum_{l=1}^{d}(\widetilde{\lambda}(l)'\mathbb{E}[g(Z_i, \beta)])^2,$$

and

$$q^*_n(\beta) = \sum_{l=1}^{d}(\lambda^*(l)'\mathbb{E}[g(Z_i, \beta)])^2$$

Consider the following decomposition:

$$\hat{q}_n(\beta) - q_{n,0}(\beta) = \hat{q}_n(\beta) - q_n(\beta) + q_n(\beta) - q_{n,0}(\beta). \tag{A.1.7}$$

For the first term $\hat{q}_n(\beta) - q_n(\beta)$ of (A.1.7) can be bounded as follows:

$$|\hat{q}_n(\beta) - q_n(\beta)| \leqslant |\sum_{l=1}^{d} \mathbb{E}_n[\{(\hat{\lambda}(l) - \widetilde{\lambda}(l))' g(Z_i, \beta)\}^2]| \tag{A.1.8}$$

$+2|\sum_{l=1}^{d} \mathbb{E}_n[\{(\hat{\lambda}(l) - \widetilde{\lambda}(l))' g(Z_i, \beta)\}\{\widetilde{\lambda}(l) g(Z_i, \beta)\}]|$.

In (A.1.8), the important component $|(\hat{\lambda}(l) - \widetilde{\lambda}(l))g(Z_i, \beta)|$ goes to 0 since $\|\hat{\lambda}(l) - \widetilde{\lambda}(l)\|_1 \to 0$ fast enough. More specifically, $|(\hat{\lambda}(l) - \widetilde{\lambda}(l))' g(Z_i, \beta)| \leqslant |(\hat{\lambda}(l) - \widetilde{\lambda}(l))' g(Z_i, \beta_0)| + K_{M,n}\|(\hat{\lambda}(l) - \widetilde{\lambda}(l))\|_1 \cdot \|\beta - \beta_0\|_2$. By Holder's inequality, $|(\hat{\lambda}(l) - \widetilde{\lambda}(l))' g(Z_i, \beta_0)| \leqslant \|\hat{\lambda}(l) - \widetilde{\lambda}(l)\|_1 \max_{1 \leqslant j \leqslant m} |g_j(Z_i, \beta_0)| = O_p\left(\sqrt{\frac{s_n^2 \log(m)}{n}} K_{B,n}\right)$. And $K_{M,n}\|(\hat{\lambda}(l) - \widetilde{\lambda}(l))'\|_1 \cdot \|\beta - \beta_0\|_2 \leqslant K_{M,n} d(\Theta)$, where $d(\Theta)$ is the diameter of Θ which is a finite constant. So $K_{M,n}\|(\hat{\lambda}(l) - \widetilde{\lambda}(l))'\|_1 \cdot \|\beta - \beta_0\|_2 = O_p(\sqrt{\frac{s_n^2 \log(m)}{n}} K_{M,n})$.

Therefore, $|(\hat{\lambda}(l) - \widetilde{\lambda}(l))' g(Z_i, \beta)| = O_p\left((K_{B,n} \vee K_{M,n})\sqrt{\frac{s_n^2 \log(m)}{n}}\right)$.

Using the bounds obtained above for $|(\hat{\lambda}(l) - \widetilde{\lambda}(l)) g(Z_i, \beta)|$, in (A.1.8), the fist component can be bounded by:

$$|\sum_{l=1}^{d} \mathbb{E}_n[\{(\hat{\lambda}(l) - \widetilde{\lambda}(l))' g(Z_i, \beta)\}^2]| = O_p\left((K_{B,n} \vee K_{M,n})^2 \frac{s_n^2 \log(m)}{n}\right);$$

The second component can be bounded by:

$$2|\sum_{l=1}^{d} \mathbb{E}_n[\{(\hat{\lambda}(l) - \widetilde{\lambda}(l))' g(Z_i, \beta)\}\{\widetilde{\lambda}(l)' g(Z_i, \beta)\}]|$$

$$\leqslant 2|\sum_{l=1}^{d} |\max_{1 \leqslant i \leqslant n} |(\hat{\lambda}(l) - \widetilde{\lambda}(l))' g(Z_i, \beta)| \mathbb{E}_n[|\widetilde{\lambda}(l)' g(Z_i, \beta)|]|,$$

where $\mathbb{E}_n[|\widetilde{\lambda}(l)' g(Z_i, \beta)|] \leqslant \mathbb{E}_n[|\widetilde{\lambda}(l)' g(Z_i, \beta_0)|] + \|\widetilde{\lambda}(l)\|_1 d(\Theta) K_{M,n}$.

By statement (4) of Assumption C.2,

$$\mathbb{E}_n[|\widetilde{\lambda}(l)'g(Z_i,\beta_0)|] = \sum_{1\leqslant j\leqslant m} |\widetilde{\lambda}(l))_j|\cdot|\mathbb{E}_n[|g_j(Z_i,\beta_0)|] \leqslant \sum_{1\leqslant j\leqslant m} |\widetilde{\lambda}(l))_j||\mathbb{E}_n[|g_j(Z_i,\beta_0)|^2] \leqslant K\|\widetilde{\lambda}(l)\|_1.$$

Therefore,

$$2|\sum_{l=1}^{d} |\max_{1\leqslant i\leqslant n} |(\hat{\lambda}(l)-\widetilde{\lambda}(l))'g(Z_i,\beta)|\mathbb{E}_n[|\widetilde{\lambda}(l)'g(Z_i,\beta)|]| = O_p\left(K_{M,n}\sqrt{\frac{s_n^2\log(m)}{n}}\right).$$

Combining the bounds obtained above,

$$\hat{q}_n(\beta) - q_n(\beta) = O_p\left((K_{B,n}\vee K_{M,n})\sqrt{\frac{s_n^2\log(m)}{n}}\right),$$

for any $\beta\in\Theta$.

For the second component $q_n(\beta)-q_{n,0}(\beta)$ in (A.1.7), we need to apply the ULLN for arrays. The statement (1) of Assumption C.2 implies that for any β and $\beta'\in\Theta$, $|\widetilde{\lambda}'g(Z,\beta)-\widetilde{\lambda}'g(Z,\beta')|\leqslant\|\widetilde{\lambda}\|_1K(Z)||\beta-\beta'||_2$, where $\mathbb{E}[\|\widetilde{\lambda}\|_1K(Z)]\leqslant KK_\lambda^u<\infty$. So by ULLN for arrays,

$$\max_{1\leqslant l\leqslant d}|\mathbb{E}_n[\widetilde{\lambda}g(Z,\beta)]-\mathbb{E}[\widetilde{\lambda}g(Z,\beta)]|\to_p 0$$

uniformly for any $\beta\in\Theta$.

Thus, $|\hat{q}_n(\beta)-q_{n,0}(\beta)|\to_p 0$ uniformly for $\beta\in\Theta$. So by construction $q_{n,0}(\hat{\beta}_L)\to_p 0$, since $\hat{q}_n(\hat{\beta}_L)=0$ and $(K_{B,n}\vee K_{M,n})\sqrt{\frac{s_n^2\log(m)}{n}}\to_p 0$. In addition, by Assumption C.5, $\|\widetilde{\lambda}(l)-\lambda^*(l)\|_1=o(\sqrt{\frac{\log(m)}{n}})$, for all $1\leqslant l\leqslant d$. Hence,

$$|\widetilde{\lambda}(l)'\mathbb{E}[g(Z,\beta)]-\lambda^*(l)\mathbb{E}[g(Z,\beta)]| = |(\widetilde{\lambda}(l)-\lambda^*(l))'\{\mathbb{E}[g(Z,\beta)]-\mathbb{E}[g(Z,\beta_0)]\}|$$

$$\leqslant ||\widetilde{\lambda}(l)-\lambda^*(l)||_1 d(\Theta)\mathbb{E}[K_M(Z)].$$

So by assumption that $\frac{\log(m)^3}{n}\to 0$, $q_n^*(\beta)-q_{n,0}(\beta)\to 0$ uniformly in $\beta\in\Theta$. Hence, $q_n^*(\hat{\beta}_L)\to_p 0$. It follows immediately that together with statement (5) of Assumption C.2, $\|\hat{\beta}_L-\beta_0\|_2\to_p 0$.

Step 3: In this step, we prove asymptotic properties of $\hat{\beta}_L$.

By Assumption C.2, the local expansion can be expanded as:

$$-\hat{\Lambda}'\mathbb{E}_n[g(Z_i,\beta_0)] = \hat{\Lambda}'\{\mathbb{E}_n[g(Z_i,\hat{\beta}_L)] - \mathbb{E}_n[g(Z_i,\beta_0)]\} \tag{A.1.9}$$

On the left hand side of (A.1.9),

$$\hat{\Lambda}'\mathbb{E}_n[g(Z_i,\beta_0)] = \widetilde{\Lambda}'\mathbb{E}_n[g(Z_i,\beta_0)] + (\hat{\Lambda} - \widetilde{\Lambda})'\mathbb{E}_n[g(Z_i,\beta_0)]. \tag{A.1.10}$$

By Assumption C.2 and C.4, the first component of (A.1.10) consists the mean of a $d \times 1$ random vector with bounded variance. So by the array Linderberg Feller Central limit theorem,

$$\sqrt{n}\widetilde{\Lambda}'\mathbb{E}_n[g(Z_i,\beta_0)] \to_d N(0, \widetilde{\Lambda}'\Omega_0\widetilde{\Lambda}),$$

where $\widetilde{\Lambda}'\Omega_0\widetilde{\Lambda} = (G_0(\beta_0)'\Omega_0^{-1}G_0(\beta_0))^{-1} + o(\sqrt{\frac{\log(m)}{n}})$, which is nearly efficient.

The second component of (A.1.10) consists the bias comming from the correlation of the estimated optimal combination matrix $\widetilde{\Lambda}$ and $\mathbb{E}_n[g(Z_i,\beta_0)]$. By (1.4.5), for all $1 \leqslant l \leqslant d$, with probability increasing to one, $||\hat{\lambda}(l) - \widetilde{\lambda}(l)||_1 \leqslant K'_\lambda\sqrt{\frac{s_n^2\log(4m/\alpha_n)}{n}}$.

To obtain an upper bound for $\max_{1\leqslant j\leqslant m}\mathbb{E}_n[g(Z_i,\beta_0)]$, we need to use the moderate-deviation theory of self-normalized vectors. For detailed theory and bounds, we refer to Shao and Zhou (2003) and De La Pena et. al.(2009). If statement (3) of Assumption C.2 holds, Lemma 5 of BCCH provides a useful result to bound $\max_{1\leqslant j\leqslant m}\mathbb{E}_n[g(Z_i,\beta_0)]$.

Lemma 27 *Suppose that for each $1 \leqslant j \leqslant m$, $R_j := \frac{\sum_{1\leqslant j\leqslant m}U_j}{\sqrt{\sum_{1\leqslant j\leqslant m}U_j^2}}$, where U_{ij} are independent random variables across i with mean 0. If $\mathbb{E}[|U_{ij}|^3] \precsim 1$, then with $\frac{\log(m)}{n^{1/3}} \to 0$, there exists a sequence $l_n \to \infty$ such that for any α small enough,*

$$\mathrm{P}(\max_{1\leqslant j\leqslant m}|R_j| \leqslant \Phi^{-1}(1-\frac{\alpha}{2m})) \geqslant 1 - \alpha(1+\frac{A}{l_n}), \tag{A.1.11}$$

where A is an absolute constant.

Hence, apply Lemma 27 to our problem, with probability increasing to one,

$$|\max_{1\leqslant j\leqslant m}\mathbb{E}_n[g_j(Z_i,\beta_0)]|\leqslant\frac{\Phi^{-1}(1-\frac{2m}{\alpha_n})}{n}\sqrt{\max_{1\leqslant j\leqslant m}\mathbb{E}_n[|g_j(Z_i,\beta_0)|^2]}=O(\sqrt{\frac{\log(m)}{n}}).$$

Therefore, the second component of (A.1.10) can be bounded by:

$$|(\hat{\lambda}(l)-\widetilde{\lambda}(l))'\mathbb{E}_n[g_j(Z_i,\beta_0)]|\leqslant||(\hat{\lambda}(l)-\widetilde{\lambda}(l))||_1\max_{1\leqslant j\leqslant m}|\mathbb{E}_n[g_j(Z_i,\beta_0)]|$$

$$\leqslant_p C\sqrt{\frac{s_n^2\log(m)}{n}}\sqrt{\frac{\log(m)}{n}}=C\frac{s_n\log(m)}{n},$$

with C being some generic constant.

On the right hand side of (A.1.9), we have the following decomposition:

$$\hat{\Lambda}\{\mathbb{E}_n[g(Z,\hat{\beta}_L)]-\mathbb{E}_n[g(Z,\beta_0)]\}=\{\mathbb{E}_n[\widetilde{\Lambda}g(Z,\hat{\beta}_L)]-\mathbb{E}_n[\widetilde{\Lambda}g(Z,\beta_0)]\} \tag{A.1.12}$$

$$+(\hat{\Lambda}-\widetilde{\Lambda})\{\mathbb{E}_n[g(Z,\hat{\beta}_L)]-\mathbb{E}_n[g(Z,\beta_0)]\}.$$

For the first component of (A.1.12), $\{\mathbb{E}_n[\widetilde{\Lambda}g(Z,\hat{\beta}_L)]-\mathbb{E}_n[\widetilde{\Lambda}g(Z,\beta_0)]\}=\mathbb{E}_n[\widetilde{\Lambda}\frac{\partial g}{\partial\beta}(Z,\beta^*)](\hat{\beta}_L-\beta_0)$ for some $\beta^*=\beta_0+O(||\hat{\beta}_L-\beta_0||_2)$.

We consider apply ULLN to $\frac{\partial g}{\partial\beta}(Z,\beta^*)$. Let $\mathcal{F}_n:=\{\frac{\partial g}{\partial\beta}(Z,\beta)-\frac{\partial g}{\partial\beta}(Z,\beta_0)|\beta\in\Theta\}$ be a class of functions indicated by elements in Θ. So $|\frac{\partial g}{\partial\beta}(Z,\beta)-\frac{\partial g}{\partial\beta}(Z,\beta_0)|\leqslant K_G(Z)d(\Theta)K_\lambda^u$. By statement (1) of Assumption C.2, $\mathbb{E}[K_G(Z)]\leqslant K$, so the ULLN for arrays holds:

$$\lim_{n\to\infty}\sup_{f\in\mathcal{F}_n}||\mathbb{E}_n[f(Z)]-\mathbb{E}[f(Z)]||_2\to 0\text{ almost surely.}$$

Therefore, $\{\mathbb{E}_n[\widetilde{\Lambda}'g(Z,\hat{\beta}_L)]-\mathbb{E}_n[\widetilde{\Lambda}'g(Z,\beta_0)]\}=(\widetilde{\Lambda}'G_0(\beta_0)+o_p(1))(\hat{\beta}_L-\beta_0)$.

For the second component of (A.1.12), $(\hat{\Lambda}-\widetilde{\Lambda})\{\mathbb{E}_n[g(Z,\hat{\beta}_L)]-\mathbb{E}_n[g(Z,\beta_0)]\}\leqslant\max_{1\leqslant l\leqslant d}||\hat{\lambda}(l)-\widetilde{\lambda}(l)||_1K_{M,n}||\hat{\beta}_L-\beta_0||_2=O_p(K_{M,n}\sqrt{\frac{s_n^2\log(m)}{n}})$.

By assumption, $K_{M,n}\sqrt{\frac{s_n^2\log(m)}{n}}\to_p 0$, so the right hand side of (A.1.9) can be written as:

$$(\widetilde{\Lambda}'G_0(\beta_0)+o_p(1))(\hat{\beta}_L-\beta_0).$$

It is also easy to verify that by statement (3) of Assumption C.5, $\widetilde{\Lambda}'G_0(\beta_0) = G_0(\beta_0)\Omega_0^{-1}G_0(\beta_0) + o(\sqrt{\frac{\log(m)}{n}})$.

Combining the asymptotic approximation of two sides in (A.1.9), we get:

$$||\hat{\beta}_L - \beta_0||_2 = O_p\left(\frac{1}{\sqrt{n}} \vee \frac{s_0 \log(m)}{n}\right).$$

In addition, if $\frac{s_0^2 \log(m)^2}{n} = o(1)$, then $\frac{s_0 \log(m)}{n} = o(\frac{1}{\sqrt{n}})$. Therefore,

$$\sqrt{n}(\hat{\beta}_L - \beta_0) \to_d N(0, V_n), \tag{A.1.13}$$

with $V_n - (G_0(\beta_0)'\Omega_0^{-1}G_0(\beta_0))^{-1} \to 0$.

Proof of Lemma 2

Here we need a set of results similar to Lemma 7-Lemma 9 stated in BCCH. For any $l = 1, 2, ..., d$, we define $\hat{m} = |\hat{T}_l \setminus T_l|$, where T_l is the set of indices of non-zero components in $\widetilde{\lambda}(l)$. We need to prove that $\hat{m} = O_p(s_n)$. In fact we only need to prove that $\hat{m} = O_p(|T_l|)$.

Notice that the first order condition gives:

$$\hat{S}(\hat{\lambda}(l))_j = sign(\hat{\lambda})_j \gamma_j \frac{t}{n},$$

for any $j \in \hat{T}_l \setminus T_l$. For simplicity, for two vectors A and B with same length, we let $\frac{A}{B} := (\frac{A_j}{B_j})_{j \geqslant 1}$.

Therefore,

$$(1+\epsilon)\frac{t_0}{n}\sqrt{\hat{m}_l} = ||\frac{\hat{S}(\hat{\lambda})}{\gamma}_{\hat{T}_l \setminus T_l}||_2 \leqslant ||\frac{\hat{S}(\widetilde{\lambda}(l))}{\gamma}_{\hat{T}_l \setminus T_l}||_2 + ||\frac{\hat{\Omega}(\hat{\lambda}(l) - \widetilde{\lambda}(l))}{\gamma}_{\hat{T}_l \setminus T_l}||_2. \tag{A.1.14}$$

The first component of the above expression is bounded by:

$$||\frac{\hat{S}(\widetilde{\lambda}(l))}{\gamma}_{\hat{T}_l \setminus T_l}||_2 \leqslant \sqrt{\hat{m}_l}\frac{t_0}{n},$$

with probability at least $1 - \alpha_n - \epsilon_n$.

For the second term,

$$||\frac{\hat{\Omega}(\hat{\lambda}(l) - \widetilde{\lambda}(l))}{\gamma}_{\hat{T}_l \setminus T_l}||_2 \leqslant \frac{1}{a}||\{\hat{\Omega}(\hat{\lambda}(l) - \widetilde{\lambda}(l)\}_{\hat{T}_l \setminus T_l}||_2.$$

Define $UT_l = T_l \cup T_l$.

For any $j \in T_l \setminus T_l$, $\hat{\Omega}(\hat{\lambda}(l) - \widetilde{\lambda}(l)_j := \sum_{j \in \hat{UT}_l} \hat{\Omega}_{jk}(\hat{\lambda}(l)_k - \widetilde{\lambda}(l)_k)$. Therefore, $\{\hat{\Omega}(\hat{\lambda}(l) - \widetilde{\lambda}(l))\}_{\hat{T}_l \setminus T_l} = \{\hat{\Omega}_{UT_l}(\hat{\lambda}(l)_{UT_l} - \widetilde{\lambda}(l)_{UT_l})\}_{\hat{T}_l \setminus T_l}$, which only depends on the subset UT_l with size at most $\hat{m}_l + s_n$.

Next, we need a simple fact from Lemma 9 of BCCH. Namely, the function $\phi(s, M)$ is sub-additive, i.e., $\phi(s_1, M) + \phi(s_2, M) \geqslant \phi(s_1 + s_2, M)$, for s_1 and s_2 being positive integers. This implies that for any real number $q \geqslant 1$ and positive integer s, $\phi(\lceil sq, M \rceil) \leqslant \lceil q \rceil \phi(s, M)$.

$$||\{\hat{\Omega}(\hat{\lambda}(l) - \widetilde{\lambda}(l))\}_{\hat{T}_l \setminus T_l}||_2 \leqslant ||\{\hat{\Omega}_{UT_l}(\hat{\lambda}(l)_{UT_l} - \widetilde{\lambda}(l)_{UT_l})\}||_2 = ||\{\hat{\Omega}^{\frac{1}{2}}_{UT_l}(\hat{\lambda}(l)_{UT_l} - \widetilde{\lambda}(l)_{UT_l})||_{2,n}.$$

We know that $||\{\hat{\Omega}^{\frac{1}{2}}_{UT_l}(\hat{\lambda}(l)_{UT_l} - \widetilde{\lambda}(l)_{UT_l})||_2 = ||\hat{\lambda}(l)_{UT_l} - \widetilde{\lambda}(l)_{UT_l}||_{2,n} \leqslant K_\lambda \frac{\sqrt{s_n}t}{n}$ with probability at least $1 - \alpha_n - \epsilon_n$ (in fact this result holds at the same events when $||\frac{\hat{S}(\widetilde{\lambda}(l))}{\gamma}_{\hat{T}_l \setminus T_l}||_2 \leqslant \sqrt{\hat{m}_l}\frac{t_0}{n}$, therefore we don't need to correct the size of the probability to let both situation hold simultaneously). Therefore,

$$||\{\hat{\Omega}(\hat{\lambda}(l) - \widetilde{\lambda}(l))\}_{\hat{T}_l \setminus T_l}||_2 \leqslant ||\{\hat{\Omega}_{UT_l}(\hat{\lambda}(l)_{UT_l} - \widetilde{\lambda}(l)_{UT_l})\}||_2 \leqslant \frac{\sqrt{s_n}K_\lambda t}{an}\sqrt{\phi(s_n + \hat{m}_l, \hat{\Omega})}.$$

Combining the bounds we obtained above, we get the following bounds as we review

(A.1.14):

$$\epsilon \frac{t_0}{n} \sqrt{\hat{m}_l} \leqslant \frac{\sqrt{s_n} K_\lambda t}{an} \sqrt{\phi(s_n + \hat{m}_l, \hat{\Omega})}.$$

Consequently, $\hat{m}_l \leqslant s_n (\frac{K_\lambda(1+\epsilon)}{a\epsilon})^2 \sqrt{\phi(s_n + \hat{m}_l, \hat{\Omega})}$. Denote $C_0 := (\frac{K_\lambda(1+\epsilon)}{a\epsilon})^2 \kappa_2$. Let C^* be an absolute constant $> 2C_0$. Let $p^* = C^* s_n$. Therefore, $\phi(\lceil p^* \rceil, \hat{\Omega}) \leqslant \kappa_2$ as $n \to \infty$, as we assume that $\phi(s_n \log(n), \hat{\Omega}) \leqslant \kappa_2$. Therefore, $\hat{m}_l \leqslant C_0 s_n \phi(\lceil p^* \rceil, \hat{\Omega}) \lceil \frac{s_n + \hat{m}_l}{C^* s_n} \rceil \leqslant C_0 s_n \kappa_2 (1 + \frac{s_n + \hat{m}_l}{C^* s_n}) \leqslant C_0 s_n \kappa_2 + \frac{s_n + \hat{m}_l}{2}$. Hence, $\hat{m}_l \leqslant (2C_0 \kappa_2 + 1) s_n$.

Step (2) is derived based on results stated in (1.4.5). It is quite obvious that $\hat{\lambda}(l)_j$ is non-zero if $j \in T_l$ when the assumption $\frac{\sqrt{n} \min_{1 \leqslant l \leqslant d, j \in T_{0,l}} |\widetilde{\lambda}(l)_j|}{s_n^2 \log(m)} \to \infty$ holds.

Proof of Theorem 2

Based on Lemma 2, we have that $|\hat{T}| \leqslant C s_n$ for some absolute constant C with probability at least $1 - \alpha_n - \epsilon_n$. So $\hat{\Omega}_{\hat{T}}$ has eigenvalues bounded from above and away from 0.

Consider $\hat{\lambda}_{\hat{T}}(l) := \hat{\Omega}_{\hat{T}}^{-1} \hat{G}(\widetilde{\beta}) e_l$. Therefore $\hat{Q}(\hat{\lambda}_{\hat{T}}) \leqslant \hat{Q}(\hat{\lambda}(l))$ by construction. Denote $\delta_{\hat{T}}(l) := \hat{\lambda}_{\hat{T}}(l) - \hat{\lambda}(l)$. Therefore,

$$\frac{1}{2} \delta_{\hat{T}}(l)' \hat{\Omega} \delta_{\hat{T}}(l) \leqslant -\delta_{\hat{T}}(l)' \hat{S}(\hat{\lambda}(l)) = -\delta_{\hat{T}}(l)' \hat{S}(\widetilde{\lambda}_l) - \delta_{\hat{T}}(l)' \hat{\Omega}(\hat{\lambda}(l) - \widetilde{\lambda}(l)).$$

By the fact that $||\delta_{\hat{T}}||_0 \leqslant C s_n$ and Assumption C.7, for the first component $-\delta_{\hat{T}}(l)' \hat{S}(\widetilde{\lambda}_l)$, we have

$$| - \delta_{\hat{T}}(l)' \hat{S}(\widetilde{\lambda}_l)| \leqslant \frac{t}{n} ||\delta_{\hat{T}}(l)||_{1,n} \leqslant \frac{bt\sqrt{Cs_n}}{n\kappa_1} ||\delta_{\hat{T}}(l)||_{2,n}.$$

For the second component, $\delta_{\hat{T}}(l)' \hat{\Omega}(\hat{\lambda}(l) - \widetilde{\lambda}(l))$, we have

$$|\delta_{\hat{T}}(l)' \hat{\Omega}(\hat{\lambda}(l) - \widetilde{\lambda}(l))| \leqslant ||\delta_{\hat{T}}(l)||_{2,n} ||\hat{\lambda}(l) - \widetilde{\lambda}(l)||_{2,n} \leqslant K_\lambda \frac{t\sqrt{s_n}}{n} ||\delta_{\hat{T}}(l)||_{2,n}.$$

The two bounds above implies that $||\delta_{\hat{T}}(l)||_{2,n} \leqslant \hat{K}(l) \frac{t\sqrt{s_n}}{n}$, where $\hat{K}(l) := 2(\frac{b\sqrt{C}}{\kappa_1} + K_\lambda)$ is an absolute constant. Therefore, it follows that $||\hat{\lambda}_{\hat{T}}(l) - \widetilde{\lambda}(l)||_{2,n} \leqslant (\hat{K}(l) + K_\lambda) \frac{t\sqrt{s_n}}{n}$.

It follows that

$$||\hat{\lambda}_{\hat{T}}(l) - \widetilde{\lambda}(l)||_2 \leqslant \frac{||\hat{\lambda}_{\hat{T}}(l) - \widetilde{\lambda}(l)||_{2,n}}{\kappa(\hat{m}_l + s_n, \hat{\Omega})} \leqslant \frac{(\hat{K}(l) + K_\lambda)}{\kappa_1} \frac{t\sqrt{s_n}}{n},$$

and by Cauchy-Schwarz inequality

$$||\hat{\lambda}_{\hat{T}}(l) - \widetilde{\lambda}(l)||_1 \leqslant \sqrt{\hat{m}_l + s_n}||\hat{\lambda}_{\hat{T}}(l) - \widetilde{\lambda}(l)||_2 \leqslant \frac{(\hat{K}(l) + K_\lambda)\sqrt{C+1}}{\kappa_1} \frac{ts_n}{n}.$$

So we have established similar bounds for $\hat{\lambda}_{\hat{T}}(l) - \widetilde{\lambda}(l)$ as we did for $\hat{\lambda}(l) - \widetilde{\lambda}(l)$ in Theorem 1.

Since all the arguments in step 2 and step 3 in Theorem 1 can be carried over to this theorem, based on $\hat{\lambda}_{\hat{T}}(l) - \widetilde{\lambda}(l)$ instead of $\hat{\lambda}(l) - \widetilde{\lambda}(l)$, so the conclusion stated in Theorem 2 holds due to similar argument illustrated in the proof of Theorem 1.

Proof of Corollary 1

This results is based on the fact that $T_0 \subset \hat{T}$ and $|\hat{T}| = O(s_n)$ with probability at least $1 - \alpha_n - \epsilon_n$. The derivation is similar to Step 2 and 3 in Theorem 1 except that we do not pay additional $\log(m)$ is the bounds stated in (1.4.6). I abbreviate the proof here.

A.1.2 Proofs in Section 1.5

Proof of Lemma 4

Denote $w_i = g(Z_i, \widetilde{\beta}) - g(Z_i, \beta_0)$. By Assumption C.2, $\max_{1 \leqslant j \leqslant m} |w_{ij}| \leqslant K_{G,n}||\widetilde{\beta} - \beta||_2$.

Consider the sparse eigenvalue of $\hat{\Omega} = \mathbb{E}_n[g(Z_i, \widetilde{\beta})g(Z_i, \widetilde{\beta})']$. We compare $\hat{\Omega}$ with $\hat{\Omega}_0 = \mathbb{E}_n[g(Z_i, \beta_0)g(Z_i, \beta_0)']$. For any $||\delta||_2 = 1$ and $||\delta||_0 \leqslant s_n \log(n)$,

$$\delta'\hat{\Omega}\delta - \delta'\hat{\Omega}_0\delta = 2\mathbb{E}_n[(\delta' w_i)(g(Z_i, \beta_0)'\delta)] + \mathbb{E}_n[(\delta' w_i)^2]. \tag{A.1.15}$$

By Cauchy-Schwarz inequality,

$$|\mathbb{E}_n[(\delta' w_i)(g(Z_i,\beta_0)'\delta)]| \leqslant \mathbb{E}_n[(w_i'\delta)^2]^{\frac{1}{2}}\mathbb{E}_n[(g(Z_i,\beta_0)'\delta)^2]^{\frac{1}{2}} \leqslant \phi(s_n\log(n),\hat{\Omega}_0)^{\frac{1}{2}}\mathbb{E}_n[(w_i'\delta)^2]^{\frac{1}{2}}.$$

By Holder-inequality with the conditions $||\delta||_0 \leqslant s_n\log(n)$ and $s_n\log(n)K_{M,n}^2 = o_p(n^{2\rho})$,

$\mathbb{E}_n[(w_i'\delta)^2] \leqslant \mathbb{E}_n[(||w_i||_\infty||\delta||_1)^2] \leqslant$

$\mathbb{E}_n[K_{M,n}^2||\widetilde{\beta}-\beta_0||_2^2 s_n\log(n)||\delta||_2^2] = K_M^2 s_n\log(n)o_p(n^{-2\rho}) = o_p(1).$

Therefore,

$$\phi(log(s)s_n,\hat{\Omega}-\hat{\Omega}_0) = \max_{||\delta||_2=1,||\delta||_0\leqslant s_n\log(n)}|\delta'\hat{\Omega}\delta-\delta'\hat{\Omega}_0\delta|$$

$$\leqslant 2(\frac{K_M^2 s_n\log(n)}{n^{2\rho}})^{\frac{1}{2}}\kappa_{1,0}+\frac{K_M^2 s_n\log(n)}{n^{2\rho}} \to_p 0.$$

If in addition Assumption C.9 holds that: $\kappa_1 \leqslant \kappa(Cs,\hat{\Omega}_0) \leqslant \phi(Cs,\hat{\Omega}_0) \leqslant \kappa_2$, for any δ on the unit sphere and $||\delta_0|| \leqslant Cs_0$,

$$\delta'\hat{\Omega}\delta = \delta'\hat{\Omega}_0\delta + (\delta'\hat{\Omega}\delta-\delta'\hat{\Omega}_0\delta) = \delta'\hat{\Omega}_0\delta + o_p(1).$$

Let c be a small absolute constant such that $c < \frac{\kappa_{1,0}}{2}$. So with probability going to one, $0 < \kappa_{1,0}-c \leqslant \kappa(s_n\log(n),\hat{\Omega}) \leqslant \phi(s_n\log(n),\hat{\Omega}) \leqslant \kappa_{2,0}+c$.

Proof of Lemma 5

The score of objective function $\hat{Q}$ at $\widetilde{\lambda}$ can be expanded as: $\hat{S}(\widetilde{\lambda}) = (\hat{\Omega}-\Omega_0)\widetilde{\lambda}-(\hat{G}(\widetilde{\beta})-G_0(\beta_0))+\Omega_0(\widetilde{\lambda}-\lambda^*)$. Let T be the set of indices of non-zero components of $\widetilde{\lambda}$. Therefore, for each $j \in \{1,2,...,m\}$,

$$S_j = \sum_{k\in T}\widetilde{\lambda}_k(\mathbb{E}_n[g_j(Z_i,\widetilde{\beta})g_k(Z_i,\widetilde{\beta})]-E[g_j(Z_i,\beta_0)g_k(Z_i,\beta_0)])-(\mathbb{E}_n[\frac{\partial g_j}{\partial\beta}(Z_i,\widetilde{\beta})v] \quad \text{(A.1.16)}$$

$$-E[\frac{\partial g_j}{\partial \beta}(Z_i, \beta_0)v]) + \sum_{1 \leqslant k \leqslant m} \Omega_{0,jk}(\widetilde{\lambda}_k - \lambda_k^*).$$

There are three components in the above equation A.1.16):

(1)$\sum_{k \in T} \widetilde{\lambda}_k(\mathbb{E}_n[g_j(Z_i, \widetilde{\beta})g_k(Z_i, \widetilde{\beta})] - E[g_j(Z_i, \beta_0)g_k(Z_i, \beta_0)])$;

(2) $\mathbb{E}_n[\frac{\partial g_j}{\partial \beta}(Z_i, \widetilde{\beta})v] - E[\frac{\partial g_j}{\partial \beta}(Z_i, \beta_0)v]$;

(3) $\sum_{1 \leqslant k \leqslant m} \Omega_{0,jk}(\widetilde{\lambda}_k - \lambda_k^*)$.

Step 1: Consider the upper bound of the first component in equation (A.1.16). Obviously,

$$||(\hat{\Omega} - \Omega_0)\widetilde{\lambda}||_\infty \leqslant ||(\hat{\Omega} - \hat{\Omega}_0)\widetilde{\lambda}||_\infty + ||(\hat{\Omega}_0 - \Omega_0)\widetilde{\lambda}||_\infty. \tag{A.1.17}$$

The first part of the decomposition (A.1.17) is $||(\hat{\Omega} - \hat{\Omega}_0)\widetilde{\lambda}||_\infty$.

Similar to the calculation in Lemma 4, we could prove that:

$||(\hat{\Omega} - \hat{\Omega}_0)\widetilde{\lambda}||_\infty$ is bounded by:

$$||(\hat{\Omega} - \hat{\Omega}_0)\widetilde{\lambda}||_\infty \leqslant ||(\hat{\Omega} - \hat{\Omega}_0)||_\infty ||\widetilde{\lambda}||_1 \leqslant_p (1+\epsilon)2K_{M,n}||\tilde{\beta} - \beta_0||_2||\widetilde{\lambda}||_1,$$

where $1 > c > 0$ be a small absolute constant and $||(\hat{\Omega} - \hat{\Omega}_0)||_\infty := \max_{1 \leqslant j \leqslant m, 1 \leqslant k \leqslant m} |\hat{\Omega}_{jk} - \hat{\Omega}_{0,jk}|$.

The second part of the decomposition (A.1.17) is $||(\hat{\Omega}_0 - \Omega_0)\widetilde{\lambda}||_\infty$. To obtain an upper bound for this term, again we need to use the moderate-deviation theory from Lemma 13.

By assumption $\{(g(Z_i, \beta_0)g(Z_i, \beta_0)' - \Omega_0)\widetilde{\lambda}\}_j$ are independent random variables across i with mean 0. It is easy to show that

$$\mathbb{E}[|\{(g(Z_i, \beta_0)g(Z_i, \beta_0)' - \Omega_0)\widetilde{\lambda}\}_j|^3] \leqslant \max_{1 \leqslant j \leqslant m} \mathbb{E}[|g_j(Z_i, \beta_0)^6|]||\widetilde{\lambda}||_1^3 \leqslant K_g^u(K_\lambda^u)^3 < \infty.$$

Thus, for n large enough, there exists a term $q_n := \frac{A}{l_n} \to 0$ such that $Pr(|\{(\hat{\Omega}_0 - \hat{\Omega}_0)\widetilde{\lambda}\}_j| < \mathbb{E}_n[((g(Z_i, \beta_0)g(Z_i, \beta_0) - \Omega_0)'\widetilde{\lambda})_j^2]\Phi^{-1}(1 - \frac{\alpha}{4m}), \text{for all} 1 \leqslant j \leqslant m) \geqslant 1 - \alpha(1 + q_n)$.

However, $E_n[(\{(g(Z_i,\beta_0)g(Z_i,\beta_0)' - \Omega_0)\widetilde{\lambda}\}_j)^2]$ can not be used in practice to bound $|\{(\hat{\Omega}_0 - \hat{\Omega}_0)\widetilde{\lambda}\}_j|$ because Ω_0 is unknown. Instead, we establish bounds using the empirical variance $E_n[(\{(g(Z_i,\beta_0)g(Z_i,\beta_0)' - \hat{\Omega}_0)\widetilde{\lambda}\}_j)^2]$. Let U_{ij} be $\{(g(Z_i,\beta_0)g(Z_i,\beta_0)' - \Omega_0)\widetilde{\lambda}\}_j$. Denote $\bar{U} = \mathbb{E}_n[U]$. We apply a modified version of Lemma 13 to bound $\max_{1\leqslant j\leqslant m}\bar{U}_j$. Let $z_j = \frac{\bar{U}}{\{E_n[(\{(g(Z_i,\beta_0)g(Z_i,\beta_0)'-\Omega_0)\widetilde{\lambda}\}_j)^2]\}^{\frac{1}{2}}}$ be the infeasible $Z-$ statistic. Lemma 13 establishes bounds for the $Z-$ statistic.

Consider the t-statistic $t_j := \sqrt{n}\frac{\bar{U}_j}{su_j}$, where $su_j := Se(U_j) := \{\frac{1}{n-1}\sum_{1\leqslant j\leqslant m}(U_j - \bar{U})^2\}^{\frac{1}{2}}$. Notice that there is a simple relationship between t_j and z_j:

$$t_j := z_j\sqrt{\frac{n-1}{n-z_j^2}}.$$

Thus, $\mathrm{P}(t_j \geqslant x) = \mathrm{P}(z_j \geqslant x(\frac{n}{n+x^2-1})^{\frac{1}{2}})$ for any x. Let $x = \Phi^{-1}(1-\frac{\alpha}{4m})$, so by assumption $x^3 = O(\log(m)^3) = o(n)$. Therefore, $\mathrm{P}(t_j \geqslant x) = \mathrm{P}(z_j \geqslant x(\frac{n}{n+x^2-1})^{\frac{1}{2}}) \geqslant (1-\eta)\mathrm{P}(z_j \geqslant \Phi^{-1}(1-\frac{\alpha}{4m})) \geqslant 1 - \frac{\alpha}{2} - \eta$ as $n \to \infty$ for any small number $\eta > 0$.

Now replace U_{ij} with $\{(g(Z_i,\beta_0)g(Z_i,\beta_0)' - \Omega_0)\widetilde{\lambda}\}_j$. So

$$\frac{n}{n-1}su_j^2 = \mathbb{E}_n[(\sum_{k=1}^{m} g_k(Z_i,\beta_0)g_j(Z_i,\beta_0)\widetilde{\lambda}_k)^2] - [\mathbb{E}_n(\sum_{k=1}^{m} g_k(Z_i,\beta_0)g_j(Z_i,\beta_0)\widetilde{\lambda}_k)]^2.$$

Therefore, $Pr(\frac{\sqrt{n}|((\hat{\Omega}_0-\Omega_0)\widetilde{\lambda})_j|}{\{\mathbb{E}_n[(\sum_{k=1}^{m} g_k(Z_i,\beta_0)g_j(Z_i,\beta_0)\widetilde{\lambda}_k)^2]-\mathbb{E}_n[\sum_{k=1}^{m} g_k(Z_i,\beta_0)g_j(Z_i,\beta_0)\widetilde{\lambda}_k]^2\}^{\frac{1}{2}}} \leqslant \Phi^{-1}(1-\frac{\alpha}{4m})$, for all $1 \leqslant j \leqslant m) \geqslant 1 - \frac{\alpha_n}{2} + o(1)$.

(2) The second component in (A.1.16) can be decomposed as:

$$\mathbb{E}_n[\frac{\partial g_j}{\partial\beta}(Z_i,\widetilde{\beta})v] - \mathbb{E}[\frac{\partial g_j}{\partial\beta}(Z_i,\beta_0)v] = (\mathbb{E}_n[(\frac{\partial g_j}{\partial\beta}(Z_i,\widetilde{\beta})v)] - \mathbb{E}_n[(\frac{\partial g_j}{\partial\beta}(Z_i,\beta_0)v)]) \quad \text{(A.1.18)}$$

$$+(\mathbb{E}_n[(\frac{\partial g_j}{\partial\beta}(Z_i,\beta_0)v)] - \mathbb{E}[(\frac{\partial g_j}{\partial\beta}(Z_i,\beta_0)v)]).$$

By Lemma 3, the first part of (A.1.18) is bounded by $K_G||\tilde{\beta}-\beta_0||_2$. The second part of (A.1.18) can be bounded by $\mathbb{E}_n[|\frac{\partial g_j}{\partial\beta}(Z_i,\beta_0)|^2] - \mathbb{E}_n[\frac{\partial g_j}{\partial\beta}(Z_i,\beta_0)]^2$ using the same strategy as

described in the previous component. Thus, $Pr(\sqrt{n}\frac{|\mathbb{E}_n[\frac{\partial g_j}{\partial \beta}(Z_i,\beta_0)v]-\mathbb{E}[\frac{\partial g_j}{\partial \beta}(Z_i,\beta_0)v]|}{\{\mathbb{E}_n[(\frac{\partial g_j}{\partial \beta}(Z_i,\beta_0)v)^2]-[\mathbb{E}_n(\frac{\partial g_j}{\partial \beta}(Z_i,\beta_0)v)]^2\}^{\frac{1}{2}}} \leqslant \Phi^{-1}(1-\frac{\alpha}{4m}), \text{for all } 1 \leqslant j \leqslant m) \geqslant 1-\frac{\alpha_n}{2}+o(1)$.

(3) The third component $\Omega_0(\tilde{\lambda}-\lambda^*)$ is non-stochastic constant. It equals to 0 if λ^* obeys exact sparse assumption C.4.

The penalty γ_j proposed in Lemma 5 is a combination of the upper bounds described in steps (a), (b) and (c). Hence,

$$\mathrm{P}(\max_{1\leqslant j\leqslant m}\left|\frac{\hat{S}_j(\widetilde{\lambda})}{\gamma_j}\right| \leqslant \Phi^{-1}(1-\frac{\alpha_n}{4m})) \geqslant 1-\alpha_n+o(1). \tag{A.1.19}$$

Proof of Corollary 2

By Assumption C.5, we know that $||\widetilde{\lambda}-\lambda^*||_1 = o_p(\sqrt{\frac{m}{n}})$. Therefore, $\|\Omega_0(\tilde{\lambda}-\lambda^*)\| \to_p 0$. If $\frac{(K_{G,n}\vee K_{M,n})^2}{n^{2\rho-1}\log(m)} \to 0$, the first term stated in 1.5.1 would also converge to 0. Hence, if assumption C.8 holds, the penalties γ_j^R is uniformly converging to the infeasible penalties γ_j stated in Lemma 5, for all $1 \leqslant j \leqslant m$.

Hence, there exists a sequence $\epsilon_n' \to 0$ such that with probability at least $1-\alpha_n-\epsilon_n-\epsilon_n'$,

$$\max_{1\leqslant j\leqslant m}\left|\frac{\hat{S}_j(\widetilde{\lambda})}{\gamma_j}\right| \leqslant \frac{t_0}{n}.$$

Proof of Lemma 6

Lemma 13 allows us to bound the maxima of a vector with length m of empirical averages of mean zero i.i.d data. For $l=1,2,...,d$, let $\gamma(l)$ be the vector of penalties for $v=e_l$, and $\hat{S}^l(\lambda)$ be the score function of $\frac{1}{2}\lambda'\hat{\Omega}\lambda-\lambda'\hat{G}(\widetilde{\beta})e_l$.

To bound the maxima of $\left|\frac{\hat{S}^l(\widetilde{\lambda})_j}{\gamma_j(l)}\right|$ for all $1 \leqslant j \leqslant m$ and $1 \leqslant l \leqslant d$, we can simply consider the $1 \times md$ vector $S^* := (S^1(\widetilde{\lambda})', S^2(\widetilde{\lambda})', ..., S^d(\widetilde{\lambda})')$. So we can apply Lemma 13 to S^* and the uniform upper bound for $\left|\frac{\hat{S}^l(\widetilde{\lambda})_j}{\gamma_j(l)}\right|$ will be $\sqrt{n\Phi^{-1}(1-\frac{4md}{\alpha_n})}$, with probability at least $1-\alpha_n-o(1)$.

Proof of Lemma 7

The proof of this Lemma is quite straight forward. Statement (2) in Assumption C.11 guarantees that both γ_j^R and γ_j^C are bounded from below by K_G^l, $1 \leqslant j \leqslant m$. Assumption C.5 and statement (1) in Assumption C.11 guarantees that both γ_j^R and γ_j^C are bounded from the above.

Proof of Lemma 8

(1) For the refined penalty γ_j^R, denote

$$w_j^g := \{\mathbb{E}_n[(\sum_{1\leqslant k\leqslant m} g_k(Z_i,\widetilde{\beta})g_j(Z_i,\widetilde{\beta})\hat{\lambda}_k)^2] - [\mathbb{E}_n(\sum_{1\leqslant k\leqslant m} g_k(Z_i,\widetilde{\beta})g_j(Z_i,\widetilde{\beta})\hat{\lambda}_k)]^2\}$$

$$-\{\mathbb{E}_n[(\sum_{k\in T_0} g_k(Z_i,\beta_0)g_j(Z_i,\beta_0)\widetilde{\lambda}_k)^2] - [\mathbb{E}_n(\sum_{k\in T_0} g_k(Z_i,\beta_0)g_j(Z_i,\beta_0)\widetilde{\lambda}_k)]^2\},$$

and

$$w_j^G := \{\mathbb{E}_n[(\frac{\partial g_j}{\partial\beta}(Z_i,\widetilde{\beta})v)^2] - [\mathbb{E}_n\frac{\partial g_j}{\partial\beta}(Z_i,\widetilde{\beta})v]^2\} - \{\mathbb{E}_n[\frac{\partial g_j}{\partial\beta}(Z_i,\beta_0)v^2] - \mathbb{E}_n[\frac{\partial g_j}{\partial\beta}(Z_i,\beta_0)v]^2\}.$$

It is easy to rewrite w_j^g and w_j^G as:

$$w_j^g = \{\mathbb{E}_n[(\sum_{1\leqslant k\leqslant m} g_k(Z_i,\widetilde{\beta})g_j(Z_i,\widetilde{\beta})\hat{\lambda}_k)^2] - \mathbb{E}_n[(\sum_{k\in T_0} g_k(Z_i,\beta_0)g_j(Z_i,\beta_0)\widetilde{\lambda}_k)^2]\}$$

$$-\{[\mathbb{E}_n(\sum_{1\leqslant k\leqslant m} \hat{g}_k(Z_i,\beta_0)\hat{g}_j(Z_i,\beta_0)\hat{\lambda}_k)]^2 - [\mathbb{E}_n(\sum_{k\in T_0} g_k(Z_i,\beta_0)g_j(Z_i,\beta_0)\widetilde{\lambda}_k)]^2\},$$

and

$$w_j^G = \{\mathbb{E}_n[(\frac{\partial g_j}{\partial\beta}(Z_i,\widetilde{\beta})v)^2] - \mathbb{E}_n[(\frac{\partial g_j}{\partial\beta}(Z_i,\beta_0)v)^2]\} - \{[\mathbb{E}_n\frac{\partial g_j}{\partial\beta}(Z_i,\widetilde{\beta})v]^2 - [\mathbb{E}_n\frac{\partial g_j}{\partial\beta}(Z_i,\beta_0)v]^2\}.$$

First,

$$\{\mathbb{E}_n[(\sum_{1\leqslant k\leqslant m} g_k(Z_i,\widetilde{\beta})g_j(Z_i,\widetilde{\beta})\hat{\lambda}_k)^2]-\mathbb{E}_n[(\sum_{k\in T_0} g_k(Z_i,\beta_0)g_j(Z_i,\beta_0)\widetilde{\lambda}_k)^2]\}$$

$$=\mathbb{E}_n[\{\sum_{1\leqslant k\leqslant m} g_k(Z_i,\widetilde{\beta})g_j(Z_i,\widetilde{\beta})\hat{\lambda}_k-g_k(Z_i,\beta_0)g_j(Z_i,\beta_0)\widetilde{\lambda}_k\}^2]$$

$$-2\mathbb{E}_n[\{\sum_{1\leqslant k\leqslant m} g_k(Z_i,\widetilde{\beta})g_j(Z_i,\widetilde{\beta})\hat{\lambda}_k-g_k(Z_i,\beta_0)g_j(Z_i,\beta_0)\widetilde{\lambda}_k\}^2]\{\sum_{1\leqslant k'\leqslant m} g_{k'}(Z_i,\beta_0)g_j(Z_i,\beta_0)\widetilde{\lambda}_{k'}]\}$$

$$\leqslant\mathbb{E}_n[\{\sum_{1\leqslant k\leqslant m} g_k(Z_i,\widetilde{\beta})g_j(Z_i,\widetilde{\beta})\hat{\lambda}_k-g_k(Z_i,\beta_0)g_j(Z_i,\beta_0)\widetilde{\lambda}_k\}^2]$$

$$+2\mathbb{E}_n[\{\sum_{1\leqslant k\leqslant m} g_k(Z_i,\widetilde{\beta})g_j(Z_i,\widetilde{\beta})\hat{\lambda}_k-g_k(Z_i,\beta_0)g_j(Z_i,\beta_0)\widetilde{\lambda}_k\}^2]^{\frac{1}{2}}\mathbb{E}_n[\{\sum_{1\leqslant k'\leqslant m} g_{k'}(Z_i,\beta_0)g_j(Z_i,\beta_0)\widetilde{\lambda}_{k'}\}^2].$$

The key component $\sum_{1\leqslant k\leqslant m}\{g_k(Z_i,\widetilde{\beta})\hat{g}_j(Z_i,\widetilde{\beta})\hat{\lambda}_k-g_k(Z_i,\beta_0)g_j(Z_i,\beta_0)\widetilde{\lambda}_k\}$ is bounded by:

$$\sum_{1\leqslant k\leqslant m}\left|\{g_k(Z_i,\widetilde{\beta})g_j(Z_i,\widetilde{\beta})\hat{\lambda}_k-g_k(Z_i,\beta_0)g_j(Z_i,\beta_0)\widetilde{\lambda}_k\}\right|\leqslant\left|\sum_{1\leqslant k\leqslant m} g_k(Z_i,\beta_0)g_j(Z_i,\beta_0)(\hat{\lambda}_k-\widetilde{\lambda}_k)\right|$$

$$+\left|\sum_{1\leqslant k\leqslant m}\{g_k(Z_i,\widetilde{\beta})g_j(Z_i,\widetilde{\beta})-g_k(Z_i,\beta_0)g_j(Z_i,\beta_0)\}\widetilde{\lambda}_k\right|$$

$$\leqslant\sum_{1\leqslant k\leqslant m}|g_k(Z_i,\beta_0)g_j(Z_i,\beta_0)|\,|\hat{\lambda}_k-\widetilde{\lambda}_k|+(2K_{M,n}||\widetilde{\beta}-\beta_0||_2+K_{M,n}^2||\widetilde{\beta}-\beta_0||_2^2)||\widetilde{\lambda}||_1^2.$$

By assumption $K_{M,n}||\widetilde{\beta}-\beta_0||_2\to_p 0$, so $K_{M,n}^2||\widetilde{\beta}-\beta_0||_2^2||\widetilde{\lambda}||_1^2$ is smaller than $K_{M,n}||\widetilde{\beta}-\beta_0||_2$ as $n\to\infty$ with probability going to one. So there exists a small $1>c>0$ such that $2K_{M,n}||\widetilde{\beta}-\beta_0||_2+K_{M,n}^2||\widetilde{\beta}-\beta_0||_2^2\leqslant(2+c)K_{M,n}||\widetilde{\beta}-\beta_0||_2$ with probability going to one.

Therefore, $\mathbb{E}_n[\{\sum_{1\leqslant k\leqslant m} g_k(Z_i,\beta_0)g_j(Z_i,\beta_0)\hat{\lambda}_k-g_k(Z_i,\beta_0)g_j(Z_i,\beta_0)\widetilde{\lambda}_k\}^2]$

$\leqslant\sum_{1\leqslant k,k'\leqslant m}\mathbb{E}_n[|\hat{\lambda}_k-\widetilde{\lambda}_k||\hat{\lambda}_{k'}-\widetilde{\lambda}_{k'}||g_k(Z_i,\beta_0)g_{k'}(Z_i,\beta_0)g_j(Z_i,\beta_0)^2|]$

$\leqslant\max_{1\leqslant j\leqslant m}\mathbb{E}_n[g_j(Z_i,\beta_0)^4]\sum_{1\leqslant k,k'\leqslant m}|\hat{\lambda}_k-\widetilde{\lambda}_k||\hat{\lambda}_{k'}-\widetilde{\lambda}_{k'}|\leqslant K_g^u||\hat{\lambda}-\widetilde{\lambda}||_1^2.$

Similarly, $\mathbb{E}_n[(\sum_{1\leqslant k'\leqslant m} g_{k'}(Z_i,\beta_0)g_j(Z_i,\beta_0)\widetilde{\lambda}_{k'})^2] \leqslant K_g^u||\widetilde{\lambda}||_1^2$.

So by the fact that $(x+y)^2 \leqslant 2x^2+2y^2$, we get: $\mathbb{E}_n[\{\sum_{1\leqslant k\leqslant m} g_k(Z_i,\widetilde{\beta})g_j(Z_i,\widetilde{\beta})\hat{\lambda}_k - g_k(Z_i,\beta_0)g_j(Z_i,\beta_0)\widetilde{\lambda}_k\}^2] \leqslant$

$2\mathbb{E}_n[\{\sum_{1\leqslant k\leqslant m} g_k(Z_i,\beta_0)g_j(Z_i,\beta_0)\hat{\lambda}_k - g_k(Z_i,\beta_0)g_j(Z_i,\beta_0)\widetilde{\lambda}_k\}^2]$

$+\mathbb{E}_n[\{\sum_{1\leqslant k\leqslant m}\{g_k(Z_i,\widetilde{\beta})g_j(Z_i,\widetilde{\beta}) - g_k(Z_i,\beta_0)g_j(Z_i,\beta_0)\}\widetilde{\lambda}_k\}^2]$

$\leqslant K_g^u||\hat{\lambda}-\widetilde{\lambda}||_1^2 + 2(2+c)^2K_{M,n}^2||\widetilde{\beta}-\beta_0||_2^2||\widetilde{\lambda}||_1^2$.

By assumption, $K_{M,n}||\widetilde{\beta}-\beta_0||_2 \to_p 0$, $||\hat{\lambda}-\widetilde{\lambda}||_1 \to_p 0$, $||\widetilde{\lambda}||_1 \leqslant K_\lambda^u$, therefore

$$w_j^g = \mathbb{E}_n[(\sum_{1\leqslant k\leqslant m} g_k(Z_i,\widetilde{\beta})g_j(Z_i,\widetilde{\beta})\hat{\lambda}_k)^2] - \mathbb{E}_n[(\sum_{k\in T_0} g_k(Z_i,\beta_0)g_j(Z_i,\beta_0)\widetilde{\lambda}_k)^2] \to_p 0.$$

Second,

$$w_j^g = \{\mathbb{E}_n[(\sum_{1\leqslant k\leqslant m} g_k(Z_i,\widetilde{\beta})g_j(Z_i,\widetilde{\beta})\hat{\lambda}_k)^2] - \mathbb{E}_n[(\sum_{k\in T_0} g_k(Z_i,\beta_0)g_j(Z_i,\beta_0)\widetilde{\lambda}_k)^2]\}$$

$$-\{[\mathbb{E}_n(\sum_{1\leqslant k\leqslant m} \hat{g}_k(Z_i,\beta_0)\hat{g}_j(Z_i,\beta_0)\hat{\lambda}_k)]^2 - [\mathbb{E}_n(\sum_{k\in T_0} g_k(Z_i,\beta_0)g_j(Z_i,\beta_0)\widetilde{\lambda}_k)]^2\},$$

and

$$w_j^G = \{\mathbb{E}_n[(\frac{\partial g_j}{\partial\beta}(Z_i,\widetilde{\beta})v)^2] - \mathbb{E}_n[(\frac{\partial g_j}{\partial\beta}(Z_i,\beta_0)v)^2]\} - \{[\mathbb{E}_n\frac{\partial g_j}{\partial\beta}(Z_i,\widetilde{\beta})v]^2 - [\mathbb{E}_n\frac{\partial g_j}{\partial\beta}(Z_i,\beta_0)v]^2\}.$$

Statement (1) of Assumption C.2 implies that $\max_{i,j}|(\frac{\partial g_j}{\partial\beta}(Z_i,\beta_0)v)_j| \leqslant K_{M,n}$. Therefore,

$$\mathbb{E}_n[(\frac{\partial g_j}{\partial\beta}(Z_i,\widetilde{\beta})v)^2] - \mathbb{E}_n[(\frac{\partial g_j}{\partial\beta}(Z_i,\beta_0)v)^2] = \mathbb{E}_n[(\frac{\partial g_j}{\partial\beta}(Z_i,\widetilde{\beta})v - \frac{\partial g_j}{\partial\beta}(Z_i,\beta_0)v)^2]$$

$$+2\mathbb{E}_n[\{\frac{\partial g_j}{\partial\beta}(Z_i,\widetilde{\beta})v - \frac{\partial g_j}{\partial\beta}(Z_i,\beta_0)v\}\frac{\partial g_j}{\partial\beta}(Z_i,\beta_0)v].$$

$$\leqslant K_G^2||\widetilde{\beta}-\beta_0||_2^2 + 2\mathbb{E}_n[K_G(Z)]K_{M,n}||\widetilde{\beta}-\beta_0||_2.$$

And similarly,

$$\{[\mathbb{E}_n \frac{\partial g_j}{\partial \beta}(Z_i, \widetilde{\beta})v]^2 - [\mathbb{E}_n \frac{\partial g_j}{\partial \beta}(Z_i, \beta_0)v]^2| \leqslant K_G^2 ||\widetilde{\beta} - \beta_0||_2^2 + \mathbb{E}_n[K_G(Z)]K_{M,n}||\widetilde{\beta} - \beta_0||_2.$$

By the LLN for arrays, $\mathbb{E}_n[K_G(Z)] \to_p \mathbb{E}[K_G(Z)] \leqslant K$. Therefore, by assumption $K_{G,n}||\widetilde{\beta} - \beta_0||_2 \to_p 0$ and $K_{M,n}||\widetilde{\beta} - \beta_0||_2 \to_p 0$. Thus, $w_j^G \to_p 0$.

Notice the upper bounds that we derived for w_j^g and w_j^G are uniform in j, therefore w_j^g and w_j^G converges to 0 w.p. $\to 1$ uniformly. Hence $\hat{\gamma}_j^R \to \gamma_j^R$ uniformly, i.e., there exists a constant $K_{\gamma,n} \to 0$ such that $\max_{1 \leqslant j \leqslant m} |\hat{\gamma}_j^R - \gamma_j^R| \leqslant K_{\gamma,n}$.

By assumption **C.7**, γ_j^R is bounded from above and below. Suppose $a \leqslant \gamma_j^R \leqslant b$ for all $1 \leqslant j \leqslant m$. Therefore, $1 - \frac{K_{\gamma,n}}{a} \leqslant \frac{\hat{\gamma}_j^R}{\gamma_j^R} \leqslant 1 + \frac{K_{\gamma,n}}{a}$.

(2) The proof is similar to the argument presented in (1). I abbreviate the proof here.

Proof of Theorem 3

In fact it is suffice to show that a similar version of Assumption C.7 holds, since the Proof of Theorem 1 and Theorem 2 only relies on Assumption C.7 for the penalty terms together with other technical conditions. The high level Assumptions C.6 are verified by the Assumptions C.1-C.3, C.5, C.9-C.11 and the growing conditions (1) and (2)stated in Theorem 3.

By Lemma 8, the $\hat{\gamma}^R$ (or $\hat{\gamma}^C$) satisfies $\frac{\hat{\gamma}_j^R}{\gamma_j^R} \to 1$ uniformly for $j = 1, 2, ..., m$. From now on, we only mention γ^R since all logics for γ^R can be carried over to γ^C. So Assumption C.8 also holds as n approaches infinity.

Also by Lemma 8, there exists a sequence of positive numbers ϵ_n' converging to 0 such that with probability at least $\geqslant 1 - \epsilon_n'$, $\frac{1}{1+\frac{\epsilon}{3}} \leqslant \frac{\hat{\gamma}^R}{\gamma_j^R} \leqslant 1 + \frac{\epsilon}{3}$, where we know that ϵ is a small but fixed positive number. Therefore, since we know that with probability at least $1 - \alpha_n - \epsilon_n$, we have:

$$\max_{1 \leqslant j \leqslant m} \left| \frac{|\hat{S}(\widetilde{\lambda})_j|}{\gamma_j^R} \right| \leqslant \frac{t_0}{n}.$$

Then, with probability at least $1 - \alpha_n - \epsilon_n - \epsilon'_n$,

$$\max_{1 \leqslant j \leqslant m} \left| \frac{\hat{S}(\widetilde{\lambda})_j}{\hat{\gamma}_j^R} \right| \leqslant \frac{t_0}{n}(1 + \frac{\epsilon}{3}).$$

Denote $t'_0 = \frac{t_0}{n}(1 + \frac{\epsilon}{3})$. So for $\epsilon < 0.2$, we know that $t > t'_0(1 + \frac{\epsilon}{2})$. Then all the derivations in Theorem 1 and Theorem 2 follows with the constants constructed with $\frac{\epsilon}{2}$, not ϵ.

Proof of Lemma 9

Suppose there are λ^1 and λ^2 such that $||\lambda^1||_1 \leqslant ||\lambda^2||_1$. Let $\hat{\lambda}^1 = \Pi(\lambda^1)$ and $\hat{\lambda}^2 = \Pi(\lambda^2)$ to be the solutions of the problem $\mathcal{P}$ with penalties terms constructed by the parameters λ^1 and λ^2.

Denote $q = \{\max_{1 \leqslant j \leqslant m}(\mathbb{E}_n g_j(Z_i, \beta_0)^4 - [\mathbb{E}_n g_j(Z_i, \beta_0)^2]^2)\}^{\frac{1}{2}}$. q is a positive real number bounded from the above. So $\gamma_j^C(\lambda) = q||\lambda||_1 + \left\{\mathbb{E}_n|(G(Z_i, \widetilde{\beta})v)_j|^2 - [\mathbb{E}_n(G(Z_i, \widetilde{\beta})v)_j]^2\right\}^{\frac{1}{2}}$.

Consider the objective function $\hat{Q}(\lambda) = \frac{1}{2}\lambda'\hat{\Omega}\lambda - \lambda'\hat{G}(\widetilde{\beta})v$.

By definition, $\hat{Q}(\lambda^1) + q||\lambda^1||_1||\hat{\lambda}^1||_1 + \left\{\mathbb{E}_n|(G(Z_i, \widetilde{\beta})v)_j|^2 - [\mathbb{E}_n(G(Z_i, \widetilde{\beta})v)_j]^2\right\}^{\frac{1}{2}} ||\hat{\lambda}^1||_1 \leqslant \hat{Q}(\lambda^2) + q||\lambda^1||_1||\hat{\lambda}^2||_1 + \left\{\mathbb{E}_n|(G(Z_i, \widetilde{\beta})v)_j|^2 - [\mathbb{E}_n(G(Z_i, \widetilde{\beta})v)_j]^2\right\}^{\frac{1}{2}} ||\hat{\lambda}^2||_1$,

and $\hat{Q}(\lambda^2) + q||\lambda^2||_1||\hat{\lambda}^2||_1 + \left\{\mathbb{E}_n|(G(Z_i, \widetilde{\beta})v)_j|^2 - [\mathbb{E}_n(G(Z_i, \widetilde{\beta})v)_j]^2\right\}^{\frac{1}{2}} ||\hat{\lambda}^2||_1 \leqslant \hat{Q}(\lambda^1) + q||\lambda^2||_1||\hat{\lambda}^1||_1 + \left\{\mathbb{E}_n|(G(Z_i, \widetilde{\beta})v)_j|^2 - [\mathbb{E}_n(G(Z_i, \widetilde{\beta})v)_j]^2\right\}^{\frac{1}{2}} ||\hat{\lambda}^1||_1$.

Adding the above two inequalities together, we obtain the following inequality:

$$||\lambda^1||_1||\hat{\lambda}^1||_1 + ||\lambda^2||_1||\hat{\lambda}^2||_1 \leqslant ||\lambda^2||_1||\hat{\lambda}^1||_1 + ||\lambda^1||_1||\hat{\lambda}^2||_1.$$

Since we assume $||\lambda^1||_1 \leqslant ||\lambda^2||_1$, it is easy to see that $||\hat{\lambda}^1||_1 \geqslant ||\hat{\lambda}^2||_1$. That is to say, the function $\Pi_2 : x \mapsto ||\Pi_1(x)||_1$ is a non-negative, decreasing function mapping from $[0, \infty)$ to $[0, \infty)$. Therefore, Π_2 only has one unique fixed point $\in [0, +\infty)$, that is x^C.

It is suffice to choose $\xi > \frac{\max_{1\leqslant j\leqslant m}\left\{\mathbb{E}_n|(G(Z_i,\widetilde{\beta})v)_j|^2-[\mathbb{E}_n(G(Z_i,\widetilde{\beta})v)_j]^2\right\}^{\frac{1}{2}}}{q}$, since 0 would be the solution to $\mathcal{P}$ under the penalty terms with $||\breve{\lambda}||_1 = \xi$ in (1.5.6). So the fixed point must lies between $\Pi_2(x_0)$ and $\Pi_2(x_1)$. The binomial search algorithm finds the fixed point within logarithm time of $\frac{\xi}{\eta}$.

A.1.3 Proof of Lemma 10 and Lemma 11

The proofs of Lemma 10 and Lemma 11 are based on Theorem 3. By the properties of $\hat{\gamma}^C$ and $\hat{\gamma}^R$, it is easy to verify that $\lambda^C(l)$ and $\lambda^R(l)$ are consistent estimators of $\widetilde{\lambda}(l)$ in L_1 norm, for all $1 \leqslant l \leqslant d$:

(1) For the empirical coarse penalty $\hat{\gamma}^C$, since we know that by Theorem 3, $||\Pi(\widetilde{\lambda}) - \widetilde{\lambda}||_1 = O_p(\sqrt{\frac{s_n^2\log(m)}{n}})$ is converging to 0, then $\Pi_2(||\widetilde{\lambda}||_1) - ||\widetilde{\lambda}||_1 \to_p 0$. By Lemma 9, we know that Π_2 is a decreasing function.

The fixed point x^C must satisfy that $\Pi_2(||\widetilde{\lambda}||_1) \leqslant x^C \leqslant ||\widetilde{\lambda}||_1$ or $||\widetilde{\lambda}||_1 \leqslant x^C \leqslant \Pi_2(||\widetilde{\lambda}||_1)$. In both cases we have $x^C - ||\widetilde{\lambda}||_1 \to_p 0$. Then it follows that $\hat{\gamma}_j$ defined by $||\breve{\lambda}||_1 = x^C$ are valid for all $1 \leqslant j \leqslant m$, i.e., they satisfy the results stated in Lemma 8. If follows that $||\lambda^C - \widetilde{\lambda}||_1 \to_p 0$.

(2) For the empirical refined penalty $\hat{\gamma}^R$, since the initial value is consistent in L_1 distance, we would have that for any $p \geqslant 0$, $||\lambda^{(p)} - \widetilde{\lambda}||_1 \to_p 0$, followed by the results stated in Theorem 3. Therefore, $||\lambda^R - \widetilde{\lambda}||_1 \to 0$.

Therefore, the corresponding penalties using λ^C and λ^R are valid, i.e., the results stated in Lemma 8 holds. Then the logic of Theorem 3 leads all results stated in Lemma 10 and Lemma 11.

A.2 Proofs in Chapter 2

A.2.1 Proofs in Section 2.1

Proof of Lemma 13

For any $A \notin \mathcal{S}_u'$, suppose (1) $\exists A_1, A_2 \subset A, A_1, A_2 \neq \emptyset, A_1 \cup A_2 = A$, such that $\varphi(A_1) \cap \varphi(A_2) = \emptyset$; or (2)$\exists u \in \mathcal{U}$, such that $u \notin A$, and $\varphi(u) \subset \varphi(A)$.

If (1) is true, $v^M(A) = v^M(A_1 \cup A_2) = v(A_1) + v(A_2) \leqslant \mu(\varphi(A_1)) + \mu(\varphi(A_2)) = \mu(\varphi(A_1) \cup \varphi(A_2)) = \mu(\varphi(A_1 \cup A_2)) = \mu(\varphi(A))$, so $A \notin \mathcal{S}_u$.

If (2) is true, $v^M(A) \leqslant v(A \cup \{u\}) \leqslant \mu(\varphi(A \cup \{u\})) = \mu(\varphi(A))$, so $A \notin \mathcal{S}_u$.

Therefore, by definition 2.2.2, $\mathcal{S}_u \subset \mathcal{S}_u'$.

$\forall A \notin \mathcal{S}_u$, assuming elements in $\mathcal{S}_u$ are denoted as A_i, $1 \leqslant i \leqslant |\mathcal{S}_u|$. For simplicity of notations, we can consider A_i as a vector in $\{0,1\}^{d_1}$. By definition, $\exists \pi \geqslant 0, s.t. (1) \sum_{i=1}^r \pi_i A_i \geqslant A, (2) \sum_{i=1}^r \pi_i \mu(\varphi(A_i)) \leqslant \mu(\varphi(A))$, where $r := |S_u|$. Without loss of generality, assume $\pi_i > 0$, $i = 1, 2, ..., r$, otherwise we would simply omit the A_i which corresponds to $\pi_i = 0$ in the sum above. Such an assumption does not affect our analysis below.

Since $\sum \pi_i A_i \geqslant A$, so

$\sum_{i=1}^r \pi_i 1(A_i \cap \varphi^{-1}(y) \neq \emptyset) \geqslant 1(A \cap \varphi^{-1}(y) \neq \emptyset)$, for any $y \in \mathcal{Y}$.

By Galichon and Henry (2011), μ is sub-modular. Therefore,

$\sum \pi_i \mu(\varphi(A_i)) = \sum_{y \in \mathcal{Y}} \sum_{i=1}^r \pi_i \mu(y) 1(A_i \cap \varphi^{-1}(y) \neq \emptyset) \geqslant \sum_{y \in \mathcal{Y}} \mu(y) 1(A \cap \varphi^{-1}(y)) = \mu(\varphi(A))$.

But we know that $\sum \pi_i \mu(\varphi(A_i)) \leqslant \mu(\varphi(A))$, by construction. Hence the inequality above holds as an equality, i.e., for any $y \in \mathcal{Y}$, $\sum_{i=1}^r \pi_i 1(A_i \cap \varphi^{-1}(y) \neq) = 1(A \cap \varphi^{-1}(y))$.

But we know that $\sum_{i=1}^r \pi_i A_i \geqslant A$. Therefore, for any $y \in \mathcal{Y}$, $\varphi^{-1}(y) \cap A \subset A_i$ or $\varphi^{-1}(y) \cap A \cap A_i = \emptyset$ for all i.

We prove the above argument by contradiction. Assuming that there exists a $y \in \mathcal{Y}$

and $1 \leqslant i \leqslant r$ such that $\varphi^{-1}(y) \cap A \cap A_i \neq$, and $\varphi^{-1}(y) \cap A \subsetneq A_i$. Therefore, there exists $u \neq u'$ such that $u, u' \in \varphi^{-1}$, $u \in A \cap A_i'$, $u' \in A$ but $u' \notin A_i$. Thus,

$\sum_{i=1}^{r} \pi_i A_i 1(A_i \cap \varphi^{-1}(y) \neq) = \pi_i + \sum_{j \neq i} \pi_j A_j 1(A_j \cap \varphi^{-1}(y) \geqslant \pi_i + \sum_{j \neq i} \pi_j 1(u' \in A_j) = \pi_i + \sum_{j=1}^{r} \pi_j 1(u' \in A_j) \geqslant \pi_i + 1 > 1 = 1(A \cap \varphi^{-1}(y))$, contradiction!

Thus, for any $y \in \mathcal{Y}$, $\varphi^{-1}(y) \cap A \subset A_i$ or $\varphi^{-1}(y) \cap A \cap A_i = \emptyset$ for all i.

The above statement immediately implies the following conclusion:

If A is self-connected, then for any A_i, either $A_i \cap A = \emptyset$ or $A_i \cap A = A$. By the equality, for any A_i, there exists no $y \in \varphi(A_i)$ such that $y \notin \varphi(A)$. So $\varphi(A_i) = \varphi(A)$. Since $A \neq A_i$, so there exists $u \in \mathcal{U}$ such that $\varphi(u) \subset \varphi(A)$, but $u \notin A$. So $A \notin \mathcal{S}_u'$.

Otherwise A is not self-connected, so $A \notin \mathcal{S}_u'$. Therefore, in both cases, $A \notin \mathcal{S}_u'$. This means that $\mathcal{S}_u \supset \mathcal{S}_u'$. Combining with the result that $\mathcal{S}_u' \supset \mathcal{S}_u$, $\mathcal{S}_u = \mathcal{S}_u'$.

Proof of Lemma 14 is similar to that of Lemma 13.

Proof of Theorem 4

$\mathcal{S}^*$ is the minimum set of inequalities which contains all information if condition $v(\mathcal{U}) = 1$ holds. Therefore, $\mathcal{S}^* \subset \mathcal{S}_u$, $\mathcal{S}^* \subset \mathcal{S}_y^{-1}$. So $\mathcal{S}^* \subset \mathcal{S}_u \cap \mathcal{S}_y^{-1}$.

For any $A \in \mathcal{S}_u \cap \mathcal{S}_y^{-1}$, suppose $A \notin \mathcal{S}^*$. So there exists $\pi_i > 0$ and $A_i \in \mathcal{S}^*$, $1 \leqslant i \leqslant r$, and $\pi_0 \geqslant 0$, such that:

(1)$\sum_{1 \leqslant i} \pi_i A_i - \pi_0 \geqslant A$.

(2)$\sum_{1 \leqslant i} \pi_i \mu(\varphi(A_i)) - \pi_0 \geqslant \mu(\varphi(A))$.

By the similar argument of Lemma 13, all the inequalities in (2) must holds as an equality. Again, for any $y \in \mathcal{Y}$, either $\varphi^{-1}(y) \cap A$ is a subset of A_i, or it does not intersect with A_i. Since $A \in \mathcal{S}_u$ is connected, so for any A_i, either $A_i \supset A$ or $A_i \cap A = \emptyset$.

Since there exists B such that $\varphi^{-1}(B) = A^c$, then $\varphi^{-1}(\varphi(A)^c) = A^c$. Without loss of generality, let $B = \varphi(A)^c$. Since the graph is connected, so it must be that $\varphi(u) \cap \varphi(A) \neq \emptyset$, for some $u \in A^c$. Since $\pi_0 > 0$, then there must exist a set A_{i_0} such

that $u \in A_{i_0}$. So $A_i \supset A$, since $\varphi(u) \cap \varphi(A) \neq \emptyset$. Also, for any $y \in B$, it is also required that $\varphi^{-1}(b) \subset A_i$ or $\varphi^{-1}(b) \cap A_i = \emptyset$.

However, the set B is self-connected! Therefore for any A_i, $B \subset A_i$ or $B \cap A_i = \emptyset$. Hence, $A_{i_0} = \mathcal{U}$, which contradicts with the definition of $\mathcal{S}^*$.

Therefore, $\mathcal{S}^* = \mathcal{S}_u \cap \mathcal{S}_y^{-1}$.

A.2.2 Proofs in Section 2.3

Proof of Lemma 14

The selected set $\hat{\mathcal{I}}$ implies all relaxed inequalities $M_j v \leqslant \hat{b}_j + \lambda_{n,m}$. Therefore, $\hat{Q}_{\hat{\mathcal{I}}} \subset \hat{Q} \oplus \lambda_{n,m}$. By assumption 2, $\max_{1 \leqslant j \leqslant m} |\hat{b} - b| \leqslant \lambda_{n,m}$ with probability $1 - \alpha$, so $Q \subset \hat{Q} \oplus \lambda_{n,m}$ and $\hat{Q} \subset Q \oplus \lambda_{n,m}$ with probability $1 - \alpha$. Hence, $Q \subset \hat{Q}_{\hat{\mathcal{I}}} \oplus \lambda_{n,m}$ and $\hat{Q}_{\hat{\mathcal{I}}} \subset \hat{Q} \oplus \lambda_{n,m} \subset \hat{Q} \oplus 2\lambda_{n,m}$ with probability $1 - \alpha$.

Proof of Theorem 5

Consider $\widetilde{\Pi}^*$ defined in Definition 2.3.1. For every $1 \leqslant j \leqslant m$, $|\widetilde{\Pi}_j^*(\hat{b}-b)| \leqslant ||\widetilde{\Pi}_j^*||_1 \max_{j \in T_0} |\hat{b}-b| \leqslant K d_1^r \hat{\sigma} \sqrt{\frac{\log(\frac{2s_0}{\alpha})}{n}}$ with probability at least $1 - \alpha$. Therefore, it is easy to see that $\widetilde{\Pi}_0$ is a feasible solution to the problem $\hat{\mathcal{R}}$ with probability at least $1 - \alpha$. Now let's focus on the event when $\widetilde{\Pi}_0$ is a feasible solution of $\hat{\mathcal{R}}$.

Let $\hat{\Pi}$ be the solution to the problem $\hat{\mathcal{R}}$. So

$$||g(\hat{\Pi})||_1 \leqslant ||g(\widetilde{\Pi}^*)||_1 \leqslant s_0 K^u.$$

So $\hat{\mathcal{I}}_{S,\eta} \leqslant \frac{s_0 K^u}{\eta}$.

For any $j \in T_0$, let v_j be the point such that the maximal separation is realized while

other inequalities hold for v. Therefore, by construction

$$\hat{\Pi}(Mv_j - \hat{b}) \geqslant Mv_j - \hat{b} - \lambda_S. \tag{A.2.1}$$

We have $Mv_j \geqslant b + c_{g,n}$, and $Mv_{j'} - \hat{b} \leqslant 0$ for all $j' \neq j$. So the j^{th} inequality of (8.2) indicates that

$$\hat{\Pi}_{jj}(c_{g,n} - \hat{b}_j + b_j) \geqslant c_{g,n} - \lambda_S - \hat{b}_j + b_j.$$

So $\hat{\Pi}_{jj} \geqslant \frac{c_{g,n} - \lambda_S - (\hat{b}_j - b_j)}{c_{g,n} - (\hat{b}_j - b_j)} \geqslant \frac{c_{g,n} - \lambda_S - \lambda_{n,m}}{c_{g,n} - \lambda_{n,m}}$.

The growing condition $\frac{d_1^{2r} \log(s_0) \vee \log(m)}{n c_{g,n}^2} \to 0$ guarantees that $\hat{\Pi}_{jj} > \eta$ for any $\eta < 1$, as $n \to \infty$. Therefore $j \in \hat{\mathcal{I}}$ and $j \in \hat{\mathcal{I}}_\eta$. Thus, $\hat{I}_\eta \supset T_0$.

Since we know that $T_0 \subset \hat{I}_\eta$, so $\hat{Q}_{\hat{\mathcal{I}}_\eta} \subset \hat{Q}_{T_0} \subset Q \oplus \lambda_{n,m}$. By construction $Q \subset \hat{Q} \oplus \lambda_{n,m} \subset \hat{Q}_{\hat{\mathcal{I}}_\eta} \oplus \lambda_{n,m}$.

Proof of Theorem 6

Consider Π^* defined in Definition 2.3.2. For every $1 \leqslant j \leqslant m$, $|\Pi_j^*(\hat{b} - b)| \leqslant ||\Pi_j^*||_1 \max_{j \in T_0} |\hat{b} - b| \leqslant K d_1^r \hat{\sigma} \sqrt{\frac{\log(\frac{2s_0}{\alpha})}{n}}$ with probability at least $1 - \alpha$. Therefore, it is easy to see that Π^* is a feasible solution to the problem $\hat{\mathcal{R}}$ with probability at least $1 - \alpha$. Now let's focus on the event when Π^* is a feasible solution of $\hat{\mathcal{R}}$.

Let $\hat{\Pi}$ be the solution to the problem $\hat{\mathcal{R}}$. So

$$||g(\hat{\Pi})||_1 \leqslant ||g(\Pi^*)||_1 \leqslant s^* K^u.$$

So $\hat{\mathcal{I}}_{S,\eta} \leqslant \frac{s^* K^u}{\eta}$.

For any $j \in T^*$, let v_j be the point such that the maximal separation is realized while other inequalities hold for v. Therefore, by construction

$$\hat{\Pi}(Mv_j - \hat{b}) \geqslant Mv_j - \hat{b} - \lambda_S. \tag{A.2.2}$$

We have $Mv_j \geqslant b + c_{g,n}$, and $Mv_{j'} - \hat{b} \leqslant 0$ for all $j' \neq j$. So the j^{th} inequality of (8.2)

indicates that

$$\hat{\Pi}_{jj}(c_{g,n} - \hat{b}_j + b_j) \geqslant c_{g,n} - \lambda_S - \hat{b}_j + b_j.$$

So $\hat{\Pi}_{jj} \geqslant \frac{c_{g,n}-\lambda_S-(\hat{b}_j-b_j)}{c_{g,n}-(\hat{b}_j-b_j)} \geqslant \frac{c_{g,n}-\lambda_S-\lambda_{n,m}}{c_{g,n}-\lambda_{n,m}}$.

The growing condition $\frac{d_1^{2r}\log(s_0)\vee\log(m)}{nc_{g,n}^2} \to 0$ guarantees that $\hat{\Pi}_{jj} > \eta$ for any $\eta < 1$, as $n \to \infty$. Therefore $j \in \hat{\mathcal{I}}$ and $j \in \hat{\mathcal{I}}_\eta$. Thus, $\hat{I}_\eta \supset T^*$.

Since we know that $T^* \subset \hat{I}_\eta$, so $\hat{Q}_{\hat{\mathcal{I}}_\eta} \subset \hat{Q}_{T^*} \subset Q \oplus \frac{\lambda_S+\lambda_{n,m}}{2}$. By construction $Q \subset \hat{Q} \oplus \lambda_{n,m} \subset \hat{Q}_{\hat{\mathcal{I}}_\eta} \oplus \lambda_{n,m}$.

Proof of Lemma 18

Let Π be a feasible solution of the problem $\mathcal{R}$:

$$\min_{\Pi} \sum_{k=1}^{m} \max_{1 \leqslant j \leqslant j} (\Pi_{jk}),$$

subject to:

$$(1)\Pi M \geqslant M, \Pi \geqslant 0,$$

$$(2)\Pi b \leqslant b.$$

$$\Pi_{ij} = 0, \text{ if j } \notin T_0.$$

Any feasible solution of this above problem is that $\Pi_{ii} = 1$, for all $i \in T_0$, and $\Pi_{ij} = 0$, for all $i \neq j$. Hence, the optimal value of the objective function is s_0.

In our case, except for the p^{th} row of M, every row satisfies: $M_i \in \{0,1\}^d$. Again, for the problem $\mathcal{R}$, any optimal solution must satisfy $\Pi_{ii} = 1$, for any $i \in T_0$. Therefore the value of the objective function is at least s_0.

Meanwhile, for any $i \notin T_0$, by definition, there exists $\alpha_j \geqslant 0$, for any $j \neq i$, $j \in T_0$

and $\alpha_p \geqslant 0$ such that:

$\sum_{j \in T_0} \alpha_j M_j - \alpha_p \geqslant M_i$,

and $\sum_{j\in T_0}\alpha_j b_j - \alpha_p \leqslant b_i$.

Without loss of generality, we could assume that $\alpha_1 \geqslant \alpha_2 \geqslant ... \geqslant \alpha_r > 0 = \alpha_{r+1} = ... = \alpha_{p-1}$. Next we prove that there must be a feasible vector of α_i such that $\alpha_1 \leqslant 1$. Then we could conclude that the minimum value of the objective function in problem $\mathcal{R}$ is s_0, and the optimal solution exactly recovers the true model. Denote the set A correspond to M_j, and A_i correspond to M_i. Without loss of generality, assume that

By Galichon and Henry (2011), $\mu(\varphi(A))$ is a sub-modular.

$b_j = \mu(\phi(A))$, therefore $\sum_{1\leqslant i\leqslant r}\alpha_i b_j - \alpha_p = \sum_{1\leqslant i\leqslant r}\alpha_i\mu(\varphi(A_i)) - \alpha_p\mu(\varphi(\mathcal{U})) \geqslant \mu(\varphi(\sum_{1\leqslant i\leqslant r}\alpha_i A_i - \alpha_p)) \geqslant \mu(A) = b_j$. Therefore the above equality holds as an equality. If $\alpha_1 > 1$, then $\alpha_p > 0$. So for any $u \notin A_i$, there must be $j \in T_0$ such that $u \in M_j$.

So for any $y \in \varphi(A)$, either $\phi^{-1}(y) \cap A_i \cap A = \emptyset$ or $(\phi^{-1}(y) \cap A) \subset A_i$. Similarly, for any $y \notin \varphi(A)$, $\phi^{-1}(y) \cap A_i = \emptyset$ or $\phi^{-1}(y) \subset A_i$.

(1) A is connected. Let A' be $\{u|\varphi(u) \subset A\}$. So A' implies A. We only need to prove that A' can be constructed via $\sum_{1\leqslant i\leqslant r}\alpha_i A_i - \alpha_p\mathcal{U}$.

(2) A is connected and there is no $u \notin A$ such that $\varphi(u) \subset \varphi(A)$. Therefore $A \subset \mathcal{S}_u$. Hence $B := \varphi(A)^c$ is not connected. Let $B_1, ..., B_r$ as all the disconnected branches of B. Let $C_k = \varphi(B_k)$, for any $1 \leqslant k \leqslant r$. So $\cup_{k=1}^r C_k = A^c$, $C_{k_1} \cap C_{k_2} = \emptyset$, for any $k_1 \neq k_2$. So each C_k is connected with A.

Denote $C^k = \{u|u \in A^c, u \notin C_k\}$.

So $A\cup C^1$, $A\cup C^2$,..., $A\cup C^r$ are sets in $\mathcal{S}_u$. It is also sets in $\mathcal{S}_y^{-1}$ since $C_k = (A\cup C^k)^c$ is connected. Therefore, All these sets are in $\mathcal{S}^*$. And Let $\alpha_i = 1$, $\alpha_p = r - 1$, we could reconstruct the inequality indicated by A. And since $r \geqslant 2$, so all the coefficients $\alpha_k \leqslant 1$.

(3) A is not connected. Let $A_1, ..., A_w$ be the connected branches. Let $B = \varphi(A^c)$. Without loss of generality, similar to step (1), we could assume that each $A_i \in \mathcal{S}_u$, $1 \leqslant i \leqslant w$. Assume $B_1, ..., B_k$ is the connected branches of B. Let $C_i = \varphi^{-1}(B_i)$, $1 \leqslant i \leqslant k$. Therefore, $C_{i_1} \cap C_{i_2} = \emptyset$, for any $i_1 \neq i_2$. $C_i \cap A \neq \emptyset$, for any i. Therefore C_i, $1 \leqslant i \leqslant k$ and A_j, $1 \leqslant j \leqslant w$ form a bipartite-graph G_0. For every A_i, let $AC_1, ..., AC_{i_r}$ to be the connect branches of $G_0 - \{A_i\}$. Since the entire graph is connected, so AC_i

is connected with A_i, $1 \leqslant i \leqslant i_r$. Let $AC^i := \{u | u \notin AC_i\}$. So AC^i is a set in $\mathcal{S}_u \cap \mathcal{S}_y^{-1} = \mathcal{S}^*$. Therefore, the set A_i could be constructed by $\sum_{k=1}^{i_r} AC_k - (i_r - 1)\mathcal{U}$.

If for some set AC_k appears in the different i, let AC be such as set such that it appears in $1 \leqslant i \leqslant J$, $J \geqslant 2$. Hence, $A_1, A_2, ..., A_J \subset AC$. Without loss of generality, suppose $C_1, ..., C_q \subset AC$, $q \geqslant 1$, and $C_{q+1}, ..., C_k \cap AC = \emptyset$. For any $1 \leqslant i \leqslant J$, $AC - A_i$ is a connected branch in $G_0 - A_i$, which means that $C_1, ..., C_q$ does not connected with $A - AC$,and $C_{q+1}, ..., C_k$ does not connect with $AC - A_i$. If $J \geqslant 2$, $C_{q+1}, ..., C_k$ does not connect with $AC - A_1$ and $AC - A_2$. But $AC - A_1 \cup AC - A_2 = A$. So $C_{q+1}, ..., C_k$ does not connect with AC. And $C_1, ..., C_q$ does not connect with AC. So AC and A are not connected! Hence, each AC_k can near appear twice in constructing A_i, $1 \leqslant i \leqslant k$. Therefore there exists one way to construct A from $\mathcal{S}^*$ such that all the coefficients $\pi_{ij} \leqslant 1$, for $1 \leqslant j \leqslant p - 2$.

Hence, the optimal solution of the problem $\mathcal{R}$ is s_0. And $\mathcal{I}^* = T_0$.

Proof of Lemma 19

The proof is similar to that in Theorem 6. However, this Lemma achieves better rates because the structure of the Core Determining Class is special. For any $\Pi \geqslant 0$ such that $\Pi M \geqslant M$, as we show in the proof of Lemma 18, the residual $\Pi b - b$ can be rewritten as a sum $\sum_{1 \leqslant l \leqslant d_2} a_l \mu(l)$ where $a_l > 0$ for all $1 \leqslant l \leqslant d_2$. Therefore, when replacing μ with $\hat{\mu}_n$, the residual $\Pi b - b$ and $\Pi \hat{b} - \hat{b}$ are very close. By the assumption that $\max_{1 \leqslant l \leqslant d_2} \frac{|\mu(l) - \hat{\mu}_n(l)|}{\mu} \to 0$, $\Pi b - b = \Pi \hat{b} - \hat{b}(1 + o_p(1))$. Therefore with probability $\geqslant 1 - \alpha$, $\Pi^* \hat{b} - \hat{b} = (\Pi^* b - b)(1 + o_p(1)) \leqslant \lambda_S$. So Π^* is a feasible solution to $\hat{\mathcal{R}}$ with probability $\geqslant 1 - \alpha$. The rest of the derivation follows the proof of Theorem 6.

A.3 Proofs in Chapter 3

A.3.1 Proofs in Section 3.2.1-3.2.2

Lemma 28 *If δ is a regular value of Δ on $\mathcal{X}$, then $M_\Delta(\delta)$ is a $d_x - 1$ manifold in $\mathbb{R}^{d_x}$ of class $\mathcal{C}^1$.*

Proof of Lemma 28

By implicit function theorem, for every regular point $x \in M_\Delta(\delta)$, there exists an open neighborhood N_x of x and an open set $V \subset \mathbb{R}^{d_x}$ such that there exists an one to one and $\mathcal{C}^1$ mapping: $g : N_x \to V$, $g(x_1, x_2, ..., x_{d_x}) \to (x_1, x_2, ..., x_{d_x-1}, y)$, where $y := f^{-1}(\delta | x_1, ..., x_{d_x-1})$ is unique on N_x. Thus, by definition, $M_\Delta(\delta)$ is a $d_x - 1$ manifold in $\mathbb{R}^{d_x}$ of class $\mathcal{C}^1$.

Proof of Lemma 20.

We refer this proof to theorem 3.11 in (Spivak 1965).

A.3.2 Proofs in Section 3.2.3

We need the following Lemma in order to prove the results in Lemma 21.

Lemma 29 *For a compact set Ω in a metric space $\mathcal{D}$, suppose there is an open cover $\theta_i, i \in I$ of Ω. Then exists a finite sub-cover of Ω, and there exists a $\eta > 0$, such that for every point in Ω, the η-ball around it is contained in the finite sub-cover.*

Proof of Lemma 29

Since Ω is a compact set in a metric space (with metric $|\cdot|$), then any open cover $\theta_i, i \in I$ of Ω has a finite subset θ_i', $i = 1, 2, ..., m$ which covers Ω.

Let $\Theta = \cup_{i=1}^{m} \theta_i$.

We prove the statement of this Lemma by contradiction.

Suppose for any $i > 0$, there exists some point x_i in the metric space $\mathcal{D}$ such that $d(x_i, \Omega) := \inf_{v \in \Omega} |x_i - v| < \frac{1}{i}$ and $x_i \notin \Theta$. Then there exists $v_i \in \Omega$ such that $d(x_i, \Omega) = d(x_i, v_i) < \frac{1}{i}$, by compactness of Ω.

$\{v_i, i \geqslant 1\}$ must have a limit point, say v_0. $v_0 \in \Omega$ by compactness of Ω.

So $d(x_i, v_0) \to 0$ as $i \to \infty$. But Θ is an open cover of V. Therefore, there must be a open ball $B(v_0)$ around v_0 such that $B(v_0) \subset \Theta$. This is a contradiction with $d(x_i, v_0) \to 0$, since $x_i \notin \Theta$. Therefore, There must be an η such that the $\eta-$ ball around Ω is covered by Θ.

Proof of Lemma 21

For statement (2) of this Lemma, it follows directly from the inverse function theorem.

The proof of (1) is divided into the following 3 steps.

Step 1 *For any regular value δ of the function Δ in $B(\mathcal{X})$, by Lemma 28, the set $M_\Delta(\delta)$ is a $(n-1)$ manifold. For any set $S \subset B(\mathcal{X})$, denote $I(S) := \Delta(S)$ as the image of Δ on a set S. It is easy to see that the set of critical values of Δ is closed in $I(S)$ for any compact set $S \subset B(\mathcal{X})$. And therefore the set of regular values of Δ on the domain $B(\mathcal{X})$ must be open in $I(B(\mathcal{X}))$.*

For any point $x \in \mathbb{R}^{d_x}$, define $B_\epsilon(x) = \{x' | ||x - x'||_2 \leqslant \epsilon\}$ as the $\epsilon-$ ball around x.

By Assumption S.1, there is a bounded open set $C(\mathcal{X})$ such that $B(\mathcal{X}) \supset int(C(\mathcal{X})) \supset \mathcal{X}^\epsilon$, where $\epsilon > 0$ is a fixed real number and $\mathcal{X}^\epsilon = \cup_{x \in \mathcal{X}} B_\epsilon(x)$. For any given regular value δ of Δ in $I(\mathcal{X})$, there must exist a neighborhood $U_\eta(\delta) = [y - \eta, y + \eta] \subset I(B(\mathcal{X}))$ of δ such that $U_\eta(\delta)$ contains no critical values of Δ on the domain $B(\mathcal{X})$. Denote $M_\Delta(\delta)$ as the union of $M(y')$ where $y' \in int(U_\eta(\delta))$. So $M_\Delta(\delta)$ is a bounded open set in $B(\mathcal{X})$.

Denote $\bar{M}_\Delta(\delta)$ as the closure of $M_\Delta(\delta)$ in $B(\mathcal{X})$. So $\bar{M}_\Delta(\delta)$ is the union of $M_\Delta(\delta')$ where $\delta' \in U_\eta(\delta)$.

Since $\Delta(x)$ is a C^1 function of $B(\mathcal{X})$, for any closed subset $D(X)$ of $B(\mathcal{X})$ and every $x \in \bar{M}_\Delta(\delta) \cap D(X)$, there exists a constant $C > 0$ such that $||\nabla\Delta(x)|| > c$ and $max_{1\leqslant i\leqslant d_x}||\frac{\partial\Delta(x)}{\partial x_i}|| < C$. From now on, we use $D(X) = C(\mathcal{X})$ with $C(\mathcal{X})$ described from the above paragraph.

For any set S and $\epsilon > 0$, we define the notation $\widetilde{S}_\epsilon$ as the union of all ϵ open balls centered at some point in S. By assumption S.3, there exists a closed set $C(\mathcal{X}) \subset B(\mathcal{X})$ such that $int(C(\mathcal{X})) \supset \bar{\mathcal{X}}$ and $\mu(x)$ in continuous at $C(\mathcal{X})$. Therefore there exists a constant $\zeta > 0$ and a closed set $C_1(X)$ such that $C(\mathcal{X}) \supset \widetilde{C_1(X)}_\zeta$, and $int(C_1(X)) \supset \widetilde{\mathcal{X}}_\zeta$.

Step 2 *In this step, we establish a finite cover of $M_\Delta(\delta)$.*

By step 1, for each $x \in M_\Delta(\delta) \cap D(X)$, since $||\nabla\Delta(x)|| > c$, there exists $1 \leqslant i \leqslant n$ such that $|\frac{\partial\Delta(x)}{\partial x_i}| > \frac{c}{\sqrt{d_x}} := c_1 > 0$. Since Δ is a C^1 function, there exists a open box $\theta(x) = X_1 \times X_2 \times ... \times X_{d_x} \subset C(\mathcal{X})$ with longest length $\leqslant \frac{\zeta}{\sqrt{k}}$ centered at point $x \in C_1(X)$ with X_j to be an interval (a_j, b_j) such that for every $x' \in \theta(x)$, $|\frac{\partial f}{\partial x_i}(x')| > \frac{c_1}{2}$. By continuity, the partial derivative $\frac{\partial\Delta}{\partial x_i}(x')$ should be greater than $\frac{c_1}{2}$ or less than $-\frac{c_1}{2}$ for all $x' \in \theta(x)$. WLOG, let's assume that $i = k$ and $\frac{\partial\Delta}{\partial x_i}(x') > \frac{c_1}{2}$ for all $x' \in \theta(x)$.

Let $C' = \frac{kC}{c_1} > 0$. Consider an open box $\theta(x)' = X_1' \times X_2' \times ... \times X_{d_x-1}' \times X_{d_x}' \subset \theta(x)$ with X_i' centered at x_i with interval length $2a$ for any $1 \leqslant i \leqslant d_x - 1$, and X_{d_x}' is centered at x_{d_x} with length $2C'a$. So for any given $x'_{-d_x} \in X_1' \times X_2' \times ... \times X_{d_x-1}'$, the value set $\{f(x') : x'_{d_x} \in X_{d_x}'\}$ will contain the interval $[\delta - aC, \delta + aC]$. For any $\delta' \in [\delta - aC, \delta + aC]$ and $x'_{-n} \in X_1' \times X_2' \times ... \times X_{d_x-1}'$, there is a unique x'_{d_x} such that $f(x'_{-d_x}, x'_{d_x}) = \delta'$.

Consider all such boxes $\theta(x)'$ for every $x \in C_1(X) \cap \bar{M}_\Delta(\delta)$. So $\theta(x)' \subset C(\mathcal{X})$. Since $C_1(X) \cap \bar{M}_\Delta(\delta)$ is a compact set, there will be a finite open sub-cover $\theta_1, \theta_2, ..., \theta_m$ that covers $C_1(X) \cap M_\Delta(\delta)$. So by lemma 5 there exists $\rho > 0$ such that $\{x' : |x' - x| < \rho, x \in M_\Delta(\delta) \cap C_1(X)\} \subset \cup_{i=1}^m \theta_i$. Since $||\nabla\Delta||$ is bounded below, so there exists δ_0 such that for any $\delta' < \delta_0$, $M_{\delta'}(y) \cap \mathcal{X} \subset \cup_{i=1}^m \theta_i$. We can simply assume that $\delta < \delta_0$ for simplicity, otherwise we can re-pick a small δ' at the beginning instead.

For simplicity of notation, we say that $M_\Delta(\delta)$ intersects θ_i at x_k axis in the above analysis. We know that for each θ_i, $M_\Delta(\delta)$ should intersect θ_i at x_j axis, for some $j \in \{1, 2, ..., d_x\}$.

Step 3 *In this step, we apply partition of unity to the open cover we construct in the last step.*

By Lemma 1, for every finite open cover $\{\theta_i\}_{i=1}^m$ *of a manifold* $M_\Delta(\delta) \cap B(\mathcal{X})$*, we can find a set of* C^∞ *partition of unity* $p_j(x)$, $1 \leqslant j \leqslant J$, *such that: (1)*$\sum_{1 \leqslant j \leqslant J} p_j(x) = 1$,

(2)$supp(p_j) \subset \theta_i$, *and*

(3)$p_j(x) \in [0,1]$.

Our main goal is to compute the following quantity, of which the limit is the derivative of $F_{\Delta,\mu}(\delta)$ *as* δ *goes to zero:*

$$\int_{M_\Delta(\delta)} \mu(x)dx = \int_{M_\Delta(\delta)\cap(\cup_{i=1}^m \theta_i)} \mu(x) \sum_{j=1}^{J} p_j(x)dx = \sum_{1 \leqslant i \leqslant m, 1 \leqslant j \leqslant J} \int_{\theta_i \cap M_\Delta(\delta)} p_j(x)\mu(x)dx. \tag{A.3.1}$$

This equation holds because any $x \in M_\Delta(\delta)$ *outside the* $\cup_{i=1}^m \theta_i$ *has 0 probability density on it.*

In each θ_i, *WLOG, suppose* $M_\Delta(\delta)$ *intersects* $\theta_i = X_{i1} \times X_{i2} \times ... \times X_{id_x}$ *at* x_{d_x} *axis.*

Since Δ *is non-singular in* θ_i, *we define the inverse function*

$$g : X_{i1} \times X_{i2} \times ... \times X_{i(d_x-1)} \times (\delta - \eta, \delta + \eta) \to X_{d_x}, \tag{A.3.2}$$

such that $f(x_1, ..., x_{d_x-1}, g(x_1, ..., x_{d_x-1}, \delta')) = \delta'$ *for all* $(x_1, ..., x_{d_x-1}, \delta') \in X_{i1} \times X_{i2} \times ... \times X_{i(d_x-1)} \times (\delta - \eta, \delta + \eta)$.

Define the one-to-one mapping ψ_k *as:*

$$\psi_k : X_1 \times X_2 \times ... \times X_{d_x-1} \times (\delta_0 - \delta, \delta_0 + \delta) \to X_1 \times X_2 \times ... \times X_{d_x-1} \times X_{d_x}, \tag{A.3.3}$$

$$\psi_n(x_1, x_2, ..., x_{d_x-1}, \delta) = (x_1, x_2, ..., x_{d_x-1}, g(x_1, x_2, ..., x_{d_x-1}, \delta)). \tag{A.3.4}$$

In the equation above, by continuity of ψ_n, Δ *and* p_j, *for each contribution in the sum:*

$$\int_{\theta_i \cap M_\Delta(\delta)} p_j(x)\mu(x)dx = \int_{X_{i1}\times X_{i2}\times ...\times X_{id_x-1}\times[\delta-\eta,\delta+\eta]} p_j \circ \psi_{d_x} \cdot \mu \circ \psi_{d_x} |det(D_{\psi_{d_x}})| dy dx_{-d_x}$$

$$= \int_{X_{i1}\times X_{i2}\times ...\times X_{id_x-1}} \int_{[\delta-\eta,\delta+\eta]} p_j \circ \psi_{d_x} \cdot \mu \circ \psi_{d_x} \frac{1}{|\frac{\partial f}{\partial x_{d_x}} \circ \psi_{d_x}|} dy dx_{-d_x}$$

$$= \int_{X_{i1}\times X_{i2}\times ...\times X_{id_x-1}} p_j \circ \psi_{d_x} \cdot \mu \circ \psi_{d_x} \frac{2\eta}{|\frac{\partial f}{\partial x_{d_x}}|} dx_{-d_x} + o(\delta)$$

Note that in the latest expression, the last component of ψ_{d_x} is fixed to be δ_0 without being specified for simplicity. In the later of the proof, this notation will be sustained.

If we write the main part of the above equation as the integration on a manifold, which is

$$\int_{X_{i1}\times X_{i2}\times ...\times X_{id_x-1}} p_j \circ \psi_{d_x} \cdot \mu \circ \psi_{d_x} \frac{2\eta}{|\frac{\partial f}{\partial X_{d_x}} \circ \psi_{d_x}|} dx_{-d_x} = 2\eta \int_{\theta_i \cap M_\Delta(\delta)} p_j \mu(x) \frac{1}{\| \nabla \Delta \|} dvol, \tag{A.3.5}$$

and sum up over i and j, we have the following equation:

$$\int_{M_\Delta(\delta)} \mu(x) dvol = \int_{M_\Delta(\delta)} \mu(x) \sum_{j=1}^{m} p_j(x) dvol = 2\delta \int_{M_\Delta(\delta)} \mu(x) \frac{1}{\| \nabla f \|} dvol + o(\eta). \tag{A.3.6}$$

The equation above is tricky: The mapping $\alpha : X_{i1} \times X_{i2} \times ... \times X_{id_x-1} \to X_{i1} \times X_{i2} \times ... \times X_{id_x-1} \times X_{id_x}$ such that $\alpha(x_1, ..., x_{d_x-1}) = (x_1, ..., x_{d_x-1}, g(x_1, ..., x_{d_x-1}))$ has Jacobian matrix

$$D\alpha^{tr} = \begin{bmatrix} 1 & 0 & ... & 0 & \frac{\partial g}{\partial x_1} \\ 0 & 1 & ... & 0 & \frac{\partial g}{\partial x_2} \\ ... & ... & ... & ... & ... \\ 0 & ... & ... & 1 & \frac{\partial g}{\partial x_{d_x-1}} \end{bmatrix}$$

The volume element is $Vol(D\alpha) = \sqrt{det(D\alpha^{tr} D\alpha)}$ is difficult to compute directly.

By Cauchy-Binet formula, $Vol(D\alpha) = \sqrt{\sum_{i=1}^{k-1}(\frac{\partial g}{\partial x_i})^2 + 1} = \frac{||\nabla\Delta||}{|\frac{\partial f}{\partial x_{d_x}}|}$.

So the left hand side of equation (A.3.6) is

$$\int_{X_1 \times X_2 \times \ldots \times X_{d_x - 1}} p_j \circ \psi_{d_x} \mu \circ \psi_{d_x} \frac{2\eta}{|\frac{\partial \Delta}{\partial x_{d_x}} \circ \psi_{d_x}|} dx_{-d_x} \tag{A.3.7}$$

$$= \int_{X_1 \times X_2 \times \ldots \times X_{d_x - 1}} p_j \circ \psi_{d_x} \mu \circ \psi_{d_x} \frac{2\eta}{||\nabla\Delta \circ \psi_{d_x}||} Vol(D\alpha).$$

Hence, by definition of integration on manifold in Definition 3.4, equation (A.3.6) holds.

Then combine with the property that $\sum_{1 \leqslant j \leqslant J} p_j(x) = 1$, *we have* $\frac{F_{\Delta,\mu}(\delta+\eta) - F_{\Delta,\mu}(\delta-\eta)}{2\eta} = \int_{M_\Delta(\delta)} \mu(x) \frac{1}{\|\nabla\Delta\|} dvol + o(1)$.

We can consider the one-side derivatives and obtain the following similar results:

$$\frac{F_{\Delta,\mu}(\delta+\eta) - F_{\Delta,\mu}(\delta)}{\delta} = \int_{M_\Delta(\delta)} \mu(x) \frac{1}{\| \nabla\Delta \|} dvol + o(1), \tag{A.3.8}$$

and

$$\frac{F_{\Delta,\mu}(\delta) - F_{\Delta,\mu}(\delta-\eta)}{\delta} = \int_{M_\Delta(\delta)} \mu(x) \frac{1}{\| \nabla\Delta \|} dvol + o(1). \tag{A.3.9}$$

That said, from the above two equations, $F_{\Delta,\mu}(\delta)$ *is differentiable at regular value of* δ *of* Δ *in* $I(\mathcal{X})$ *with derivative*

$$f_{\Delta,\mu}(\delta) = \int_{M_\Delta(\delta)} \mu(x) \frac{1}{\| \nabla\Delta \|} dvol. \tag{A.3.10}$$

End of proof.

Proof of Lemma 22

Part (a):

For any $h_n \in \mathcal{F} \to h \in \mathcal{F}_0$ with L^∞ norm, and $t_n \to 0$ as n approaches infinity, we consider the quantity below:

$\frac{F_{\Delta+t_n h_n,\mu}(\delta)-F_{\Delta,\mu}(\delta)}{t_n}$.

By assumption, the function $h \in \mathcal{F}_M$ is bounded on $C(\mathcal{X})$ and uniformly continuous. So h_n is uniformly bounded for $n \geqslant N$, since $h_n \to h$ under L^∞ norm. Let $C = sup_{x \in C(\mathcal{X}), n \geqslant N} h_n(x)$.

For any δ_0 which is a regular value of Δ on the domain $\mathcal{X}$, we consider a procedure similar to theorem 1. Suppose we have a rectangle cover $\theta_i \subset C(\mathcal{X})$ of $M_\delta(\delta_0)$ and a partition of unity $p_j(x)$ on the cover sets, $1 \leqslant i \leqslant m, 1 \leqslant j \leqslant J$. By lemma 9, there exists $\eta > 0$ such that the set $B_\eta(M_\Delta(\delta_0)) := \{x | dist(x, M_\Delta(\delta_0) \cap \mathcal{X}) < \eta\} \subset \cup_{1 \leqslant i \leqslant m} \theta_i$.

Therefore, given a η small enough, there exist N large enough such that $sup_{x \in C(\mathcal{X}), n \geqslant N} t_n |h_n - \eta| < \eta$. For any fixed positive number ζ, the following fact holds for any $x \in B_\eta(M(\delta_0)) \cap \mathcal{X}$, as n is large enough,

$$1\{\Delta(x) + t_n h_n(x) \leqslant \delta_0\} \leqslant 1\{\Delta(x) + t_n(h(x) - \zeta) \leqslant \delta_0\}. \tag{A.3.11}$$

And outside the set $B_\Lambda(M(\delta_0)) \cap \mathcal{X}$, $1\{\Delta(x) + t_n h_n(x) \leqslant \delta\}$ agrees with $1\{\Delta(x) \leqslant \delta\}$. Therefore,

$$\frac{\int_X 1\{\Delta(x) + t_n h_n(x) \leqslant \delta_0\} - 1\{\Delta(x) + t_n(h(x) - \zeta) \leqslant \delta_0\}\mu(x)dx}{t_n} \leqslant$$

$$\frac{\int_{B_\eta M(\delta_0)} 1\{\Delta(x) + t_n(h(x) - \zeta) \leqslant \delta_0\} - 1\{\Delta(x) \leqslant \delta_0\}\mu(x)dx}{t_n}$$

$$= \int_{B_\eta M(\delta_0)} \frac{1\{\Delta(x) \leqslant \delta_0 - t_n(h(x) - \zeta)\} - 1\{\Delta(x) \leqslant \delta_0\}\mu(x)}{t_n} dx.$$

Below, for simplicity, we use t for notation t_n.

And hence

$$\int_{B_\eta M_\Delta(\delta_0)} \frac{1\{\Delta(x) \leqslant \delta_0 - t(h(x)-\zeta)\} - 1\{\Delta(x) \leqslant \delta_0\}\mu(x)}{t} dx$$

$$= \lim_{t\to 0} \int_{\cup\theta_i} \frac{1\{\delta_0 \leqslant \Delta(x) \leqslant \delta_0 - t(h(x)-\zeta)\}\mu(x)}{t} dx$$

$$= \lim_{t\to 0} \sum_{i,j} \int_{\theta_i} p_j(x) \frac{1\{\delta_0 \leqslant \Delta(x) \leqslant \delta_0 - t(h(x)-\zeta)\}\mu(x)}{t} dx. \tag{A.3.12}$$

Note that the function $1\{\delta_0 \leqslant \Delta(x) \leqslant \delta_0 - t(h(x)-\zeta)\}$ equals -1 for all x such that $\delta_0 \geqslant \Delta(x) \geqslant \delta_0 - t(h(x)-\zeta)$ (if $\zeta < 0$).

Suppose $\theta_i = X_1 \times X_2 \times ... \times X_{d_x}$ intersects $B_\eta M(\delta_0)$ only at hyper-planes parallel to $X_{d_x} = 0$. And under parametrization

$$\psi_n : X_1 \times X_2 \times ... \times X_{d_x-1} \times [\delta_0 - \eta, \delta_0 + \eta]$$

$$\psi_n(x_1, ..., x_{d_x-1}, \delta) = (x_1, x_2, ..., x_{d_x-1}, g(x_1, x_2, ..., x_{d_x-1}, \delta))$$

,

where $g(x_1, x_2, ..., x_{d_x-1}, \delta)$ is the implicit function derived from equation $\Delta(x) = \delta$.

Therefore, by continuity of $h(x)$,

$$\int_{\theta_i} p_j \frac{1\{\delta_0 \leqslant \Delta(x) \leqslant y - t(h(x)-\zeta)\}\mu(x)}{t}$$

$$= \int_{X_1\times X_2\times ...\times X_{d_x-1}} \int_{[\delta_0-\eta,\delta_0+\eta]} p_j \circ \psi_{d_x} \times \frac{1\{\delta_0 \leqslant \Delta(x) \leqslant \delta_0 - t(h(x)-\zeta)\}\mu \circ \psi_{d_x}}{t|\frac{\partial\Delta}{\partial X_{d_x}} \circ \psi_{d_x}|} d\delta_0 dx_{-d_x}$$

$$= \int_{X_1 \times X_2 \times \ldots \times X_{d_x - 1}} p_j \circ \psi_{d_x} \mu \circ \psi_{d_x} \frac{-t(h(x) - \zeta)}{t|\frac{\partial \Delta}{\partial x_n} \circ \psi_{d_x}|} d\delta_0 dx_{-d_x} + o(t)$$

$$= - \int_{\theta_i \cap M(\delta_0)} p_j(x) \mu(x) \frac{h(x) - \zeta}{\| \nabla \Delta \|} dvol + o(t)$$

Since the number m and J are fixed for any $n \geqslant N$ and $|h(x) - \delta|$ bounded by constant C, $\sum_j p_j(x) = 1$ and $p_j(x) \geqslant 0$, we can exchange the sum and limit in (9.37). So (9.37) is equivalent to the following inequality:

$$\int_{B_\eta M(\delta_0)} - \frac{1\{\Delta(x) \leqslant \delta_0 - t(h(x) - \zeta)\} - 1\{\Delta(x) \leqslant \delta_0\}\mu(x)}{t} dx$$

$$= \int_{M_\Delta(\delta)} \frac{\mu(x)(h(x) - \delta)}{|\nabla \Delta(x)|} dvol + o(\delta) \qquad \text{(A.3.13)}$$

Let $\zeta \to 0$, so $\lim_{n \to \infty} \frac{F_{\Delta + t_n h_n, \mu}(\delta) - F_{\Delta, \mu}(\delta)}{t_n} \leqslant - \int_{M_\Delta(\delta)} \frac{\mu(x) h(x)}{|\nabla \Delta(x)|} dvol.$

If we bound the above limit from the other direction use similar approaches as above, we get:

$\lim_{n \to \infty} \frac{F_{\Delta + t_n h_n, \mu}(\delta) - F_{\Delta, \mu}(\delta)}{t_n} \geqslant - \int_{M_\Delta(\delta)} \frac{\mu(x) h(x)}{|\nabla \Delta(x)|} dvol.$

Combine the above two inequalities, we conclude that $F_{\Delta, \mu}(\delta)$ is Hadamard-differentiable at Δ tangentially to $\mathcal{F}_M$.

Part b:

By the inverse mapping Lemma 3.9.20 in (Van der Vaart 2000), we know that $\Delta^*_\mu(u)$ has Hadamard derivative at Δ tangentially to $\mathcal{F}_M$. The derivative maps $h \in \mathcal{F}_M$ to

$$- \frac{\partial F_{\Delta, \mu}(h; \delta)}{f_{\Delta, \mu}(\delta)}$$

Proof of Lemma 23

Part a:

Let $\mathbb{H}_0$ to be the set of bounded linear operator on the measurable functional space $\mathcal{F}_M$ with $L^{*\infty}$ norm. Suppose the operators $H_n \to H \in \mathbb{H}_0$, i.e.,

$$H_n \to H \Longleftrightarrow sup_{f\in\mathcal{F}_M, f\neq 0}|(H_n - H)f| \to 0.$$

Suppose $t_n \to 0$, and let $Q_n = (\mu + t_n H_n)$. So $F_{\Delta,Q_n}(\delta) - F_{\Delta,\mu}(\delta) = (Q_n - \mu)1\{\Delta(x) \leqslant \delta\} = t_n H_n(1\{\Delta(x) \leqslant \delta\}) = t_n(H(1\{\Delta(x) \leqslant \delta\}) + t_n(H_n - H)(1\{\Delta(x) \leqslant \delta\}))$.

Therefore, by assumption that $H_n \to H$ under $L^{*\infty}$ norm, $\frac{F_{\Delta,Q_n}(\delta) - F_{\Delta,\mu}(\delta)}{t_n} = H(1\{\Delta(x) \leqslant \delta\}) + o(1)$.

Hence $F_{\Delta,\mu}(\delta)$ has Hadamard derivative $\partial_\mu F_{\Delta,\mu}(H;\delta) = H(1\{\Delta(x) \leqslant \delta\})$ at μ tangentially to $\mathbb{H}_0$.

Part b:

By the inverse mapping Lemma 3.9.20 in (Van der Vaart 2000) and the conclusion of part (a), at any regular value $\delta = \Delta^*_\mu(u)$ of Δ on domain $\mathcal{X}$, $\Delta^*_\mu(u)$ is Hadamard differentiable at μ tangentially to $\mathbb{H}_0$. The Hadamard derivative reads as the following:

$$\partial_\mu \Delta^*(H; u, \mu) = \frac{H(1(\Delta(x) \leqslant \delta))}{f_{\Delta,\mu}(u)}.$$

Proof of Lemma 24

For statement (a), Consider $(h_{n,n}) \to (h, H) \in \mathbb{D}_0$ and $t_n \to 0$. For statement (b), it follows by results in (a) and the inverse function theorem.

Let $f_n = f + t_n h_n$ and $\widetilde{\mu}_n = \mu + t_n H_n$. $F_{\Delta,\widetilde{\mu}_n}(\delta) - F_{\Delta,\mu}(\delta) = (F_{\Delta,\widetilde{\mu}_n}(\delta) - F_{\Delta_n,\widetilde{\mu}_n}(\delta)) + (F_{\Delta_n,\widetilde{\mu}_n}(\delta) - F_{\Delta,\mu}(\delta))$.

By Lemma 22, $(F_{\Delta_n,\mu}(\delta) - F_{\Delta,\mu}(\delta)) = t_n(-\int_{M_\Delta(\delta)} \frac{h(x)\mu(x)}{||\nabla\Delta||} dVol) + o(t_n)$.

Since $|\Delta_n - \Delta|_\infty \to 0$ and $\mathcal{X}$ is compact, $|1\{\Delta_n(x) \leqslant \delta\} - 1\{\Delta(x) \leqslant \delta\}|_1 \to 0$. Therefore, $H_n(1\{\Delta_n(x) \leqslant \delta\} - 1\{\Delta(x) \leqslant \delta\}) \to 0$ since H_n, $n \geqslant 1$, are uniformly continuous operators.

$(F_{\Delta,\tilde{\mu}_n}(\delta) - F_{\Delta,\mu}(\delta) := t_n(H_n(1\{f_n(x) \leqslant \delta\}) - H_n(1\{\Delta(x) \leqslant y\}) + H_n(1\{\Delta(x) \leqslant \delta\})) = o(t_n) + t_n(H(1\{f_n(x) \leqslant \delta\}) - H(1\{\Delta(x) \leqslant \delta\})) + t_n H_n(1\{\Delta(x) \leqslant y\})) = o(t_n) + t_n H_n(1\{\Delta(x) \leqslant \delta\})$.

By Lemma 23, $t_n H_n(1\{\Delta(x) \leqslant \delta\})) = o(t_n) + t_n H(1\{\Delta(x) \leqslant \delta\}))$. Therefore, $\frac{F_{\Delta,\tilde{\mu}_n}(\delta) - F_{\Delta,\mu}(\delta)}{t_n} \to -\int_{M_\Delta(\delta)} \frac{h(x)\mu(x)}{||\nabla\Delta||} dvolol + H(1\{\Delta(x) \leqslant \delta\})$.

A.3.3 Proofs in Section 3.3

Below we recall Theorem 3.9.4 of Var der Vaart (2000).

Lemma 30 *Let $\mathcal{F}$ and $\mathcal{G}$ to be metrizable topological vector spaces. Let $g : \mathcal{F}_g \subset \mathcal{F} \to \mathcal{G}$ be a Hadamard differentiable mapping at $f \in \mathcal{F}$. Let $X_n : \Omega_n \to \mathcal{F}_g$ be maps with $r_n(X_n - X_0) \to J$ for some $r_n \to \infty$, where J is separable and takes value in $\mathcal{F}$. Then $r_n(g(X_n) - g(X_0)) \rightsquigarrow g_f(J)$, where g_f is the Hadamard derivative of g with respect to Δ.*

Proof of Proposition 1

By Lemma 22, we know that $F_{\Delta,\mu}(\delta)$ and $\Delta^*_\mu(u)$ are Hadamard differentiable with respect to Δ at regular value δ and $u = F_{\Delta,\mu}(\delta)$, tangentially to $\mathcal{F}_0$. By assumption S.5, we know that $a_n(\hat{\Delta} - \Delta) \to G(x)$. The following convergence laws are directly deduced from Lemma 11, by setting $\mathcal{F}_g = \mathcal{F}$, $X_0 = \Delta$, $X_n = \hat{\Delta}$:

$$a_n(F_{\hat{\Delta},\mu}(\delta) - F_{\Delta,\mu}(\delta)) \rightsquigarrow -\int_{M_\Delta(\delta)} \frac{G(x)\mu(x)}{\|\nabla\Delta\|} dvol$$

and

$$a_n(\hat{\Delta}^*_\mu(u) - \Delta^*_\mu(u)) \rightsquigarrow \frac{\int_{M_\Delta(\delta)} \frac{G(x)\mu(x)}{\|\nabla f\|} dvol}{f_{\Delta,\mu}(u)}$$

Proof of Proposition 2

(a) It follows by direct calculation that: $\sqrt{n}(F_{\Delta,\mu_n}(\delta)-F_{\Delta,\mu}(\delta)) = \int 1\{\Delta(x) \leqslant \delta\}\sqrt{n}(d(\hat{\mu}(x)-\mu(x)) \rightsquigarrow H_\delta$.

(b) The conclusion is followed by the inverse mapping Theorem 3.9.20 of Van der Vaart (2000).

Proof of Theorem 7

If $\hat{\Delta} \in \mathcal{F}_1$, and $F_{M_\Delta(\delta)}$ is $\mu-$ Donsker, then $\sqrt{n}(\mu_n - \mu)$ is uniformly bounded on $F_{M_\Delta(\delta)}$ with probability going to 1.

By Lemma 21, $\Lambda(y; f.\mu)$ and $\Delta^*_\mu(u)$ are Hadamard differentiable with respect to (f, μ) tangentially to $\mathbb{D}_0 := \mathcal{F}_1 \times \mathbb{H}_0$, where $\sqrt{n}(\mu_n - \mu) \in \mathbb{H}_0$ with probability going to 1.

Denote $b_n = \sqrt{n}$.

(a) If $a_n = o(b_n)$, then let $X_n = (\hat{\Delta}, \hat{\mu})$ and $X = (f, \mu)$. So $a_n(X_n - X) \rightsquigarrow (G(x), 0)$.

Therefore, by Lemma 21 and Propositions 1-2,

$$a_n(\hat{\Delta}^*_{\hat{\mu}}(u) - \Delta^*_\mu(u)) \to_d \frac{\int_{M_\Delta(\delta)} \frac{G(x)\mu(x)}{\|\nabla\Delta\|} dvolol}{f_{\Delta,\mu}(u)}.$$

(b) If $a_n = b_n$, then let $X_n = (\hat{\Delta}, \hat{\mu})$ and $X = (f, \mu)$. So $a_n(X_n - X) \rightsquigarrow (G(x), -H_\delta)$.

Therefore, Lemma 21 and Propositions 1-2,

$$a_n(\hat{\Delta}^*_{\hat{\mu}}(u) - \Delta^*_\mu(u)) \to_d \frac{\int_{M_\Delta(\delta)} \frac{G(x)\mu(x)}{\|\nabla\Delta\|} dvol - H(1\{\Delta(x) \leqslant y\})}{f_{\Delta,\mu}(u)}.$$

(c) If $b_n = o(a_n)$, then let $X_n = (\hat{\Delta}, \hat{\mu})$ and $X = (f, \mu)$. So $b_n(X_n - X) \rightsquigarrow (0, -H_\delta)$.

Therefore, by Lemma 21 and Propositions 1-2,

$$b_n(\hat{\Delta}^*_{\hat{\mu}}(u) - \Delta^*_{\mu}(u)) \to_p \frac{-H(1\{\Delta(x) \leqslant y\})}{f_{\Delta,\mu}(u)}.$$

The special case for $\hat{\mu} = \mu_n$ holds correspondingly.

The proofs of Corollary 3 and 4 follow directly from the results in Theorem 7.

Proof of Theorem 8

We prove Theorem 8 to validify inference by bootstrap. We recall Theorem 3.9.11 of Van der Vaart (2000).

Lemma 31 (Delta-method for bootstrap in probability) *Let $\mathcal{H}$ be the operator space on $\mathcal{F} = BL_1(\mathbb{R}^k)$, where BL_1 denotes the set of Lipshitz functions of order 1. Let $g : \mathbb{D}_\phi \subset \mathbb{D} \mapsto \mathbb{E}$ to be a Hadamard-differentiable mapping at $\mathbb{P}$ tangentially to $\mathbb{D}_0$. Let $\mathbb{P}_n$ be a random element such that $\sqrt{n}(\mathbb{P}_n - \mathbb{P}) \rightsquigarrow \mathbb{G}$. Let $\hat{\mathbb{P}}_n$ be in random elements in $\mathbb{D}$ such that $\sqrt{n}(\hat{\mathbb{P}}_n - \mathbb{P}_n) \rightsquigarrow \mathbb{G}$. I.e.,*

$sup_{h \in BL^1(\mathbb{R}^k)} |Eh(\sqrt{n}(\hat{P}_n - P_n)) - Eh(\mathbb{G})|$, and $E[h(\sqrt{n}(\hat{P}_n - P_n))|^* - E|h(\sqrt{n}(\hat{P}_n - P_n))]_* \to 0$.

Furthermore, if $\sqrt{n}(\mathbb{P}_n - \mathbb{P}) \rightsquigarrow \mathbb{G}$, then

$sup_{h \in BL^1(\mathbb{R}^k)} |Eh(\sqrt{n}(g(\hat{P}_n) - g(P_n))) - Eh(g_\mu(\mathbb{G}))| \to 0$, and $E[h(\sqrt{n}(g(\hat{P}_n) - g(P_n)))]^* - E[h(\sqrt{n}(g(\hat{P}_n) - g(P_n)))]_* \to 0$ hold in outer probability.

Proof of Theorem 8

For i.i.d data, the $\sqrt{n}(\hat{\mu}_n - \mu_n)$ weakly converges to a $\mu-$ Brownian Bridge. We also know that, by construction, $\sqrt{n}(\hat{\Delta}_k - \hat{\Delta})$ converges to $G(x)$.

$$\sqrt{n}((\mu_n^j, \hat{\Delta}_j) - (\mu_n, \hat{\Delta})) \rightsquigarrow (\mathbb{B}, G(\cdot)).$$

Therefore, the condition in Lemma 11 is satisfied. By Lemma 7, we know that $\Delta^*_\mu(u)$ is Hardmard differentiable with respect to (μ, f).

At regular value δ of $\Delta(x)$ on domain $\mathcal{X}$, by assumption, $\sqrt{n}((\mu_n^j, \hat{\Delta}_j) - (\mu_n, \hat{f})) \rightsquigarrow (\mathbb{B}, G(\cdot))$ and $\sqrt{n}((\mu_n, \hat{\Delta}) - (\mu, \hat{\Delta})) \rightsquigarrow (\mathbb{B}, G(\cdot))$. Let $\mathbb{P}_n = (\hat{\Delta}, \mu_n)$, $\hat{\mathbb{P}}_n = (\hat{f}_j, \mu_n^j)$, and $\mathbb{P} = (f, \mu)$. Then by Lemma 12, the conclusion holds.

A.3.4 Proofs in Section 3.4

Proof of Lemma 25

(a) First of all, the result of Lemma 21 holds for every fixed value of $d \in \mathcal{D}$. Thus,

$$F_{\Delta,\mu}(\delta) = \sum_{d \in \mathcal{D}} \mu(d) \int_{z \in \mathcal{Z}|d} 1(f(d,x) \leqslant \delta) \mu(d,x) dz,$$

and

$$f_{\Delta,\mu}(\delta) = \sum_{d \in \lceil} \mu(d) \int_{M_\Delta(\delta)_d} \frac{\mu(z|d)}{||\nabla_x f(d,x)||} d\text{Vol}.$$

(b) Similar to Lemma 6, the Hardmard derivative of $F_{\Delta,\mu}(\delta)$ with respect to (f, μ) can be calculated for given $D \in \mathcal{D}$. Given D, the Hardmard derivative of $F_{\Delta,\mu}(\delta)$ evaluated at (h, H) is:

$$-\int_{M_\Delta(\delta)_D} \mu(D) \frac{\mu(z|D)}{||\nabla_x \Delta(d,x)||} d\text{Vol} + \int_{x \in \mathcal{X}_d} H_2(x|d) 1(f(\cdot, d) \leqslant \delta).$$

The Hardmard derivative of $F_{\Delta,\mu}$ evaluated at (h, H) comprises one addition term of H_1, unconditional on d. This term equals:

$$\sum_{d \in \mathcal{D}} H_1(d) \mu(1(f(d,\cdot) \leqslant \delta)).$$

Hence, the conclusion in (b) follows.

(c) It follows directly from the inverse mapping Theorem 3.9.20 of Van der Vaart (2000).

Proof of Theorem 5

Similar to the proof of Theorem 7, this theorem follows directly from Hardmard differentiability of $\Delta_\mu^*(u)$ stated in Lemma 21 with respect to (f, μ) and Lemma 24.