Econometric Estimation with High-Dimensional Moment Equalities

Zhentao Shi*

(Job Market Paper)

November 19, 2013

Abstract

Structural models involving moment conditions are in widespread practical use, and commonly include many moments to capture the stylized facts in large datasets. The number of moments m can explode with the sample size n. The growing complexity of the model challenges familiar estimators like Generalized Method of Moments (GMM) or Empirical Likelihood (EL), whose asymptotic normality demands that n dominate m^3. In the extreme case $m > n$, the two estimators break down even numerically, as the weighting matrix is non-invertible in the criterion function of two-step GMM, and the constraints are infeasible in the primal problem of EL.

We consider a structural model in which the number of moments is not limited by the sample size, and where the econometric problem is to estimate and perform inference on a finite-dimensional parameter. We develop a novel two-step estimation procedure. We call the first step the Relaxed Empirical Likelihood (REL), which relaxes the moment constraints of the primal problem of EL. While EL requires that all moment constraints equal zero, REL tolerates a small violation specified by the user as a tuning parameter. The tuning parameter controls, as it shrinks to zero asymptotically, the maximal approximation error of the sample means of the moment functions. Under a high-dimensional asymptotic framework, we derive the consistency of REL and its rate of convergence. As the relaxation introduces first-order bias that slows the rate, the

*Department of Economics, Yale University. Email: zhentao.shi@yale.edu. I am deeply indebted to my advisor Peter Phillips for his continual inspiration, guidance, encouragement and support. I am grateful to Donald Andrews and Yuichi Kitamura for their advice. I thank Timothy Armstrong, Andrew Barron, Marianne Bruins, Xiaohong Chen, Xu Cheng, Timothy Christensen, Samuel Kortum, Zhipeng Liao, Chu-An Liu, Taisuke Otsu, Byoung Gun Park, Tabri Rami, Stefan Schneeberger, Xiaoxia Shi and Liangjun Su for comments. All remaining errors are mine. Preliminary results were presented in the Inaugural China Meeting of the Econometric Society in June 2013.

second step selects a small subset of moments in a computationally efficient manner to correct the bias of REL. The algorithm adds one moment in each iteration—the one that maximizes an information criterion evaluated at REL conditional on the moments chosen in the preceding iterations. We establish asymptotic normality and efficiency of bias-corrected REL. To the best of our knowledge, this paper provides the first asymptotically normally distributed estimator in such an environment.

The new estimator is shown to have favorable finite sample properties in simulations. Estimating an international trade model with the massive China Customs Database and Annual Survey of Manufacturing Firms, our empirical application investigates the heterogeneity and efficiency of Chinese exporting firms. In comparison with empirical works on their French counterparts, we find the Chinese firms are of lower cost efficiency.

1 Introduction

The unprecedented accessibility of large micro-datasets at the individual agent-level opens up opportunities to investigate a multitude of new economic problems as well as old problems from new perspectives. Respecting a long tradition of building parsimonious, often nonlinear, structural models to characterize complex economic phenomena, empirical econometric analysis typically concentrates on a few key parameters that bear economic meaning. Structural models involving moment conditions are widely used in many areas of econometrics. While moment restrictions arise naturally from the economic context, theory usually gives little guidance about the choice of the moment restrictions. In modern empirical work it is nothing extraordinary to estimate with hundreds or even thousands of moments. The fundamental challenge in such empirical work is to develop a theory of estimation and inference that allows for high-dimensionality in the moment conditions relative to the sample size.

Several of latest empirical applications take advantage of large datasets and large models. For instance, Altonji et al. (2013) examine a joint model of earnings, employment, job changes, wage rates, and work hours over a career with a full specification of 2429 moments. Eaton et al. (2011) explore the sales of French manufacturing firms in 113 destination countries with 1360 moments. Han et al. (2005) investigate the cost efficiency of the Spanish saving banks in a time-varying coefficient model with 872 moments. All these empirical examples estimate a finite-dimensional parameter of interest in structural models with many nonlinear moments. The underlying models take the following form. A "true" parameter β_0 satisfies the unconditional moment restrictions

$$\mathbb{E}\left[g\left(Z_i, \beta\right)\right] = 0_m, \quad (i = 1, \ldots, n),$$

where $\{Z_i\}_{i=1}^n$ is the observed data, $\beta_0 \in \mathcal{B} \subset \mathbb{R}^D$ is finite-dimensional, g is an $\mathbb{R}^m$-valued non-degenerate moment function, and 0_m is an $m \times 1$ vector of zeros.

It is known from the literature that the relative magnitude of m and n shapes the asymptotic properties of generalized method of moments (GMM) and empirical likelihood (EL) (Koenker and Machado, 1999; Donald et al., 2003). Under the usual assumptions, consistency demands $m = o(n)$ and $\sqrt{n}$-asymptotic normality demands $m^3 = o(n)$. Though the sample sizes in these cited examples can run to thousands or even millions, valid asymptotic statistical inference may require many more observations still.

In this paper, we consider a nonlinear structural model in which m can be much larger than n, and the econometric problem is to estimate and perform inference on a finite-dimensional parameter of interest. The phrase *high-dimensional* in the title refers specifically to the $m > n$ case.

Han and Phillips (2006) explain the difficulties encountered by GMM in the large m case,

considering the simplest equally-weighted GMM (EW-GMM). They find that the consistency of EW-GMM to the true parameter depends on the strength of the main signal, which contains information about the true parameter in the population, relative to the signal variability, which arises from the sample approximation to the population means of the moment functions. Despite a small contribution from each moment component, the signal variability induced by a large number of moments accumulates in the quadratic-form of the GMM criterion function and can overwhelm the main signal. In later work Newey and Windmeijer (2009) provide a consistent variance estimator of the generalized empirical likelihood (GEL) estimator under many weak moments. Estimation with high-dimensional moments is not considered in these two papers.

Nonlinear models include the linear instrumental variable (IV) model as a special case. High-dimensional IVs are considered by Bai and Ng (2009, 2010) and Belloni et al. (2012). In their setting a large number of IVs all meet the orthogonality condition (zero correlation with the structural error), while only a small handful of the IVs satisfy the relevance condition (correlation with the endogenous regressors) but the identities of the relevant IVs are unknown. Under a sparsity assumption, Belloni et al. (2012) use Lasso as the selector, and then plug in the Lasso-predicted endogenous regressors or use the post-Lasso two-stage least squares (2SLS) to efficiently estimate the structural parameter. They apply the moderate deviation theory of self-normalized sums (Jing et al., 2003) to high-dimensional regression, which relaxes distributional assumptions on the error term. Bai and Ng (2009) utilize boosting to select the IVs in the reduced-forms, and Bai and Ng (2010) develop a factor IV estimator to achieve optimality. However, in the nonlinear structural model no counterpart exists for the first-stage reduced-form equation, so these methods become inapplicable.

We address the problem in a different approach via a novel two-step procedure. In the first step, we propose an estimator called *relaxed empirical likelihood* (REL). REL tolerates a small violation of the equality moment constraints in the primal problem of the standard EL. The magnitude of the tolerance is specified by the user as a tuning parameter, which controls the maximal approximation error of the means of the moment functions. Under weak regularity conditions, REL is consistent in high-dimensional asymptotics.

Relaxation of the moment equalities introduces first order bias, however, so that the rate of convergence of REL is slower than $n^{-1/2}$. We therefore recommend a second-step refinement to correct the bias of REL. The effectiveness of the bias correction hinges on the computational feasibility of the moment selection and the quality of the selected moments. We propose a boosting-type greedy algorithm that constructs an increasing sequence of selected moments by adding in each iteration only one moment, namely the one that maximizes an information criterion given the moments selected in the preceding iterations. The

computational burden is trivial,[1] while the selection is powerful. Under certain regularity conditions, the bias-corrected version (BC-REL) follows an asymptotic normal distribution. Furthermore, the estimation can recover the total quantity of information that is finite and concentrates in a small subset of moments. To the best of our knowledge, this paper is the first to establish asymptotic normality in this high-dimensional nonlinear model.

Similar to the high-dimensional IV literature, we assume that the identities of the strong moments are unknown. While if some strong moments are known *a priori*, we can use these moments in either GMM or EL to get a consistent estimator in the first step, and then apply the second step selection to improve efficiency.

Our methodology follows a vast literature. The use of many linear IVs originated in Angrist (1990), which motivated intensive econometric research, for example Hahn (2002), Chao and Swanson (2005) and Chao et al. (2011), to name a few. Bai and Ng (2010) and Belloni et al. (2012) resolve the problem of $m > n$ in the linear IV setting. Gautier and Tsybakov (2013) propose a new IV estimator based on the Dantzig selector (Candes and Tao, 2007) in a more general setting that allows for a high-dimensional parameter in the linear structural equation; their focus is different from the other two papers. Fan and Liao (2011) deal with nonlinear moments and the selection of a high-dimensional parameter in a structural equation. They assume the true coefficients of the endogenous regressors are zeros, an assumption that limits economic applicability. In another line of research, Carrasco and Florens (2000, 2014) develop GMM theory for many moments or a continuum of moments with g in a Hilbert space. The Hilbert space setting can be restrictive when the moments are generated from detailed observations in large datasets, for example a large number of mutually orthogonal non-degenerate IVs.

Our first step estimation REL is built on EL (Owen, 1988; Qin and Lawless, 1994). Kitamura (1997) establishes EL theory for weakly dependent data. In the model with conditional moment restrictions, Kitamura et al. (2004) and Donald et al. (2003) provide local and global approximations, respectively, to achieve asymptotic efficiency. As an information-theoretic alternative to GMM, Kitamura (2001), Kitamura and Stutzer (1997), Newey and Smith (2003) and Otsu (2010) find theoretical advantages of EL. Latest developments of EL to cope with an infinite-dimensional parameter include Otsu (2007) and Lahiri and Mukhopadhyay (2012).

In terms of estimation with shrinkage methods, in particular Lasso (Tibshirani, 1996), Caner (2009) and Caner and Zhang (2013) introduce a Lasso-type penalty to GMM for

[1] The conventional moment selection procedures evaluate an information criterion for each candidate model. Even for a moderate m, the number of candidate models can be far beyond the current computational capability. For example, selecting 10 out of a total 200 moments amounts to $\binom{200}{10} = 2.245 \times 10^{16}$ combinations, a job that will take 710 years to complete on a 1,000-CPU cluster if each core processes on average 1,000 models per second. In dramatic contrast, our boosting-type selection takes less than 2 seconds on a single such CPU to accomplish the task.

variable estimation. Belloni et al. (2010, 2011), Belloni and Chernozhukov (2011) and Belloni et al. (2011) contribute to various aspects of estimation and inference in high-dimensional econometric and statistical models.

Regarding the selection step, Andrews and Lu (2001) propose several information criteria in GMM, and Hong et al. (2003) give the EL counterpart. The problem of redundant moments is discussed by Breusch et al. (1999), and a selection criterion is developed by Hall and Peixe (2003) and Hall et al. (2007). Liao (2013) and Cheng and Liao (2012) use a Lasso-type penalty to select possibly misspecified as well as redundant moments in one step.

Up to now no moment selection procedure has been shown to work in the high-dimensional context. Moreover, redundancy of moments is defined in the literature relative to a fixed set of moments. In this paper as we select an increasing sequence of moments, the reference point of redundancy is expanding. This approach helps to rule out more redundant moments—those that become redundant when other moments are added.

The rest of the paper is organized as follows. Section 2 introduces the idea of REL, proves its consistency and derives its rate of convergence. Section 3 constructs BC-REL, with emphasis on moment selection. We prove the asymptotic normality of BC-REL with or without a sparsity condition. Section 4 provides two simulation examples to check the finite-sample performance of REL and BC-REL in comparison with GMM. Section 5 applies the new procedure in an international trade model with the China Custom Database and Annual Survey of Manufacturing Firms. Section 6 concludes the paper.

2 First Step: Relaxed Empirical Likelihood

2.1 Why Familiar Estimators Break Down

The use of a large number of moments is not always advantageous in structural estimation, as we now explain. GMM (Hansen, 1982) solves the unconstrained M-estimation problem

$$\min_{\beta\in\mathcal{B}} \mathbb{E}_n\left[g_i\left(\beta\right)\right]' W_{\text{GMM}} \mathbb{E}_n\left[g_i\left(\beta\right)\right],$$

where $\mathbb{E}_n\left[\cdot\right] = \frac{1}{n}\sum_{i=1}^n \cdot$ is the empirical mean, and $g_i\left(\beta\right)$ represents $g\left(Z_i, \beta\right)$. The weighting matrix W_{GMM} may depend on the sample and the parameter β. For example, the two-step GMM or optimally-weighted GMM (OW-GMM) (Hansen et al., 1996) uses

$$W_{\text{GMM}}\left(\beta\right) = \left(\mathbb{E}_n\left[\left(g_i\left(\beta\right) - \mathbb{E}_n\left[g_i\left(\beta\right)\right]\right)\left(g_i\left(\beta\right) - \mathbb{E}_n\left[g_i\left(\beta\right)\right]\right)'\right]\right)^{-1}$$

evaluated at a preliminary consistent estimator of β_0, while the continuous updating estimator (CUE) (Hansen et al., 1996) determines the weighting $W_{\text{GMM}}(\beta)$ together with $\mathbb{E}_n[g_i(\beta)]$ in the optimization. EL, an alternative to GMM, solves the constrained problem

$$\max_{\beta\in\mathcal{B},p\in\Delta_n}\sum_{i=1}^{n}\log p_i \quad \text{subject to } \sum_{i=1}^{n}p_i g_i(\beta)=0_m \tag{1}$$

where $\Delta_n=\{p\in[0,1]^n:\sum_{i=1}^n p_i=1\}$ is the n-dimensional simplex. We call

$$\left\{\beta\in\mathcal{B}:p\in\Delta_n,\sum_{i=1}^{n}p_i g_i(\beta)=0_m\right\}$$

the feasible set, and when it is empty we say the primal problem (1) is *infeasible*.

The inconvenience of these two estimators is immediately apparent in the high-dimensional context where m may surpass n. Neither the covariance matrix of OW-GMM or CUE is invertible, as the rank of the $m\times m$ sample covariance matrix is at most n. A mechanical remedy such as a pseudo-inverse does not resolve the difficulty. In particular, if we use the Moore-Penrose pseudo-inverse as the weighting matrix in a linear IV model when $m>n$, then 2SLS reduces to OLS in the structural equation estimation. The estimator is then inconsistent if endogeneity is present.

The non-invertibility of the empirical covariance matrix is a natural consequence of the quadratic-form GMM criterion function. The covariance matrix is supposed to weight the moments, but the scarce data, relative to the magnitude of the model, is insufficient to handle so many moments. Even if we are not interested in efficiency but only concerned about consistency, EW-GMM fails to converge to the true parameter in general, as explained by Han and Phillips (2006).

EL does not survive $m>n$ either, for a different reason. Given a trial value β, the restriction $\sum_{i=1}^n p_i g_i(\beta)=0_m$ is an m-equation system to be solved by n parameters p_i's. Under the simplex restriction, n free parameters cannot in general simultaneously solve a linear m-equation system. The constraints are too severe to make (1) feasible.

Numerical and statistical problems remain when $n>m$ but m/n is non-trivial. The weighting matrix of GMM is numerically unstable, and the primal problem of EL, with non-zero probability, is infeasible even at $\beta=\beta_0$. Statistically, as mentioned in the introduction, consistency demands $n/m\to\infty$ and asymptotic normality requires $n/m^3\to\infty$.

The anatomy of GMM and EL hints at a potential solution. In GMM we can replace the quadratic form criterion function by a norm in which the approximation errors do not accumulate, for example the sup-norm. This solution is alluded to, in the special case of the linear IV model, by Belloni et al. (2012, Section 4.2) as a complement to their primary

Lasso-related IV estimator. With EL, we can relax the excessively severe restrictions in a controlled manner such that the feasible set is non-empty with high probability when the sample size is large. In the rest of this section we investigate the latter solution, and we will explain the difference between the two solutions at the end of this section after we have displayed and discussed the assumptions.

2.2 Relaxation

The idea of relaxing EL is simple. For each $j \leq m$, let $g_{ij}(\beta)$ be the j-th element of $g(Z_i, \beta)$. We replace for each $j \leq m$ the conventional moment restriction $\sum_{i=1}^n p_i g_{ij}(\beta) = 0$ of EL by $\left|\sum_{i=1}^n p_i g_{ij}(\beta)\right| / \widehat{\sigma}_j(\beta) \leq \tau$, where $\tau \geq 0$ is a tuning parameter and $\widehat{\sigma}_j(\beta)$ is the sample standard deviation of $\{g_{ij}(\beta)\}_{i=1}^n$. Standardization by $\widehat{\sigma}_j(\beta)$ makes the estimation invariant to a scale change of g. Denote $h_{ij}(\beta) := g_{ij}(\beta) / \widehat{\sigma}_j(\beta)$. Formally, the REL estimator is defined as

$$\widehat{\beta} := \max_{\beta \in \mathcal{B}} \ell_n^\tau(\beta) \tag{2}$$

where

$$\ell_n^\tau(\beta) := \max_{\{p_i\}_{i=1}^n \in \mathcal{P}^\tau(\beta)} \frac{1}{n} \sum_{i=1}^n \log p_i + \log n \tag{3}$$

and

$$\mathcal{P}^\tau(\beta) := \left\{ p \in \Delta_n : \max_{j \leq m} \left| \sum_{i=1}^n p_i h_{ij}(\beta) \right| \leq \tau \right\}. \tag{4}$$

We call the optimization in (2) the *outer loop*, and the optimization in (3) the *inner loop*. For a trial value $\beta \in \mathcal{B}$, if $\mathcal{P}^\tau(\beta) \neq \emptyset$ we denote the solution of the inner loop as $\widehat{p}(\beta)$; otherwise we set $\ell_n^\tau(\beta) = -\infty$.

The technique of relaxing the sup-norm in (4) can be traced back to Candes and Tao (2007) in the linear regression model and Gautier and Tsybakov (2013) in the linear IV model. Geometrically, $\{p \in \Delta_n : \sum_{i=1}^n p_i h_i(\beta) = 0_m\}$ is an n-dimensional convex hull that includes all discrete probability measures that satisfy the constraints of EL under a given β. When m is larger than n, the volume of the convex hull is too small to capture the origin of the m-dimensional coordinate system even at $\beta = \beta_0$; in other words $\{p \in \Delta_n : \sum_{i=1}^n p_i h_i(\beta_0) = 0_m\} = \emptyset$. The relaxation covers the convex hull with a layer whose thickness is controlled by τ. When the model is correctly specified, the particular discrete probability measure $\{p_i = 1/n\}_{i=1}^n$ makes $\max_{j \leq m} \left|\sum_{i=1}^n p_i h_i(\beta_0)\right|$ close to 0. Under regularity conditions, if τ is sufficiently large the extended convex hull captures the origin with high probability.

Despite the high-dimensionality, convexity in $\{p_i\}_{i=1}^n$ makes the inner loop computationally feasible. With the help of modern algorithms, we can solve standard convex opti-

mization problems efficiently for thousands, or even millions of parameters. On the other hand, the outer loop is non-convex in general. When the moment functions are continuously differentiable with respect to β in $\mathcal{B}$, as we will assume, the outer loop is smooth and low-dimensional. A Newton-type optimization routine then suffices.

In terms of the programming of the inner loop, it is easier to work with the primal problem in (3);[2] in terms of asymptotic analysis, it is more convenient to manipulate an unconstrained problem. We formulate the unconstrained problem via Lagrangian. Let $\|\cdot\|_1$ be the l_1-norm of a vector, i.e. $\|\gamma\|_1 = \sum_{j=1}^{m} |\gamma_j|$. Straightforward calculation gives the Lagrangian corresponding to (3),

$$\mathcal{L}_n^{\tau}(\beta, p, \gamma_A, \gamma) = \frac{1}{n}\sum_{i=1}^{n} \log p_i + \log n + \gamma_A \left(1 - \sum_{i=1}^{n} p_i\right) + \gamma' \sum_{i=1}^{n} p_i h_i(\beta) - \tau \|\gamma\|_1 \qquad (5)$$

where a scalar γ_A and an $m \times 1$ vector γ are two Lagrangian multipliers associated with the simplex constraint and the moment constraints, respectively.

The only difference between the Lagrangian (5) of REL and that of EL is the penalty term $\tau \|\gamma\|_1$. The l_1-norm penalty stems from the duality to the sup-norm in the primal problem. The connection between the sup-norm in (3) and the l_1-norm in (5) resembles the connection between the Dantzig selector (Candes and Tao, 2007) and Lasso (Tibshirani, 1996).

Applied to a linear regression with many covariates, Lasso will coerce the small coefficients to zero; so will REL in view of its dual form (5), which penalizes the sum of the absolute values of the Lagrangian multipliers. This phenomenon has an interesting economic interpretation. Lagrange multipliers are interpreted in economics as the shadow prices of resource constraints. For each trial value $\beta \in \mathcal{B}$, the l_1-penalization will eliminate the "cheap moments" in terms of their shadow prices, that is, the moments associated with small Lagrange multipliers. However, unlike the linear regression model where Lasso can be used as a selector of the significant variables, the significance of the moments in our model vary with β. If β lies in a neighborhood that shrinks fast enough to β_0, all moments satisfy $\mathbb{E}[h_{ij}(\beta)] \approx 0$, so we cannot tell which moment is relevant. That is why we should not interpret the moments associated with $\widehat{\gamma}_j\left(\widehat{\beta}\right) \neq 0$ as the relevant moments. REL is *not* a moment selection procedure in this model. The task of moment selection is left to the second step refinement.

While the Lagrangian involves $D + n + m + 1$ parameters, the dimensionality can be reduced in optimization to the parameter set (β, γ) of dimension $D + m$.

[2] In the numerical implementation of REL in this paper, we always feed the solver with the primal problem even when m is smaller than n. A modern solver of convex optimization such as `MOSEK` can automatically transform it to its dual if the dual is computationally more efficient than the primal. Appendix C gives the details about the programming of the inner loop.

Proposition 1. *When the primal problem in (3) is feasible, the solution to the saddle point problem*

$$\max_{\beta\in\mathcal{B}}\min_{\gamma\in\mathbb{R}^m} L_n^\tau(\beta,\gamma) \tag{6}$$

where $L_n^\tau(\beta,\gamma) := -\mathbb{E}_n\left[\log\left(1+\gamma' h_i(\beta) - \tau\|\gamma\|_1\right)\right]$ *is the same as the* β *and* γ *components of the solution to the penalized Lagrangian (5). Let* $\mathbb{E}_{\widehat{p}}[\cdot]$ *be the empirical measure under the estimated probability* $\widehat{p}_i := \widehat{p}_i\left(\widehat{\beta}\right) = \left(n\left(1+\widehat{\gamma}' h_i\left(\widehat{\beta}\right) - \tau\|\widehat{\gamma}\|_1\right)\right)^{-1}$*; that is,* $\mathbb{E}_{\widehat{p}}[x_i] = \sum_{i=1}^n \widehat{p}_i x_i$*. We further have*

$$\begin{aligned} sign\left(\widehat{\gamma}_j\right)\left(\mathbb{E}_{\widehat{p}}\left[h_{ij}\left(\widehat{\beta}\right)\right]\right) &= \tau,\ \forall j\in\{j:\widehat{\gamma}_j\neq 0\} \\ \left|\mathbb{E}_{\widehat{p}}\left[h_{ij}\left(\widehat{\beta}\right)\right]\right| &\leq \tau,\ \forall j\in\{j:\widehat{\gamma}_j = 0\}. \end{aligned}$$

If we compare the max-min problem of REL with that of EL, it is as if we replace the moment function g_{ij} in EL by $h_{ij}-\tau$ or $h_{ij}+\tau$, depending on the sign of $\mathbb{E}_{\widehat{p}}\left[h_{ij}\left(\widehat{\beta}\right)\right]$. The penalty term in (5) is carried into the parenthesis in the expression of $L_n^\tau(\beta,\gamma)$. This concise form holds for the l_1-penalization, but not for other penalty schemes in general. This max-min formulation is particularly useful in the development of the asymptotic theory.

2.3 Consistency

We view the model as a triangular array of models indexed by n and m. In the limiting statement, we explicitly pass $n\to\infty$ while m is understood as $m=m_n\to\infty$ implicitly.

The first set of assumptions restricts the nature of the triangular array of models. To clarify the development of the asymptotic theory, some assumptions are stronger than necessary.

Assumption 1. *(a)* $m = o\left(\exp\left(n^{1/9}\right)\right)$.

(b) The data $Z_1,\ldots,Z_n$ *are independently and identically distributed (i.i.d.) for all* $n\in\mathbb{N}$.

(c) $\mathcal{B}\subset\mathbb{R}^D$ *is a compact finite-dimensional parameter space.*

Assumption 1(a) highlights the high-dimensional asymptotic framework. Textbook asymptotics typically pass only n to infinity, not m; the theory of sieve semiparametric estimation (Chen, 2007) allows $m\to\infty$ but much more slowly than n. In contrast to these asymptotics, we permit m to pass to infinity much faster than n. Consistency and asymptotic normality established in this $m/n\to\infty$ framework certainly carries over to the $m/n\to 0$ case. One may notice that this rate is slower than $m = o\left(\exp\left(n^{1/3}\right)\right)$ as in Belloni et al. (2012) for the linear IV model. This slower rate reflects the difficulty in handling the large

weighting matrix in the second step, but is still flexible enough for most economic applications. Condition (b) assumes that the sample is generated from random sampling. No technical difficulty hinders extension from the i.i.d. framework as the law of large numbers and central limit theory we use accommodates the independent, non-identically distributed (i.n.i.d.) framework. The compactness of the parameter space in Condition (c) is commonly imposed in the theory of extremum estimation of a finite-dimensional parameter.

Next we assume some regularity conditions on the moment functions. Throughout this paper, a *universal constant* is a fixed (invariant of n) positive finite constant. Let $\|\cdot\|_2$ be the L_2-norm of a vector. Let $\sigma_j(\beta)$ be the population standard deviation of $g_j(Z,\beta)$.

Assumption 2. *For all $n \in \mathbb{N}$, there exists a universal constant C such that*

(a) $\mathbb{E}[g_{ij}(\beta)]$ *is continuously differentiable in β for all $\beta \in \mathcal{B}$ and $j \leq m$, and there exists a measurable function $B_{nj} : \mathcal{Z} \mapsto \mathbb{R}^+$ such that*

$$\sup_{\beta_1,\beta_2 \in \mathcal{B}} \frac{|g_j(Z,\beta_1) - g_j(Z,\beta_2)|}{\|\beta_1 - \beta_2\|_2} \leq B_{nj}(Z)$$

and $\max_{j\leq m} \mathbb{E}\left[B_{nj}^2(Z)\right] < C$.

(b) $\sup_{\beta\in\mathcal{B},j\leq m} \mathbb{E}\left[g_{ij}^6(\beta)\right] < C$.

(c) $\inf_{\beta\in\mathcal{B},j\leq m} \sigma_j^2(\beta) > 1/C$.

(d) $\sup_{\beta\in\mathcal{B},j\leq m} \left|\mathbb{E}_n\left[g_{ij}^4(\beta)\right]\right| = O_{\mathrm{p}}(1)$.

Assumption 2(a) is a sufficient condition for stochastic equicontinuity. Continuity of $\mathbb{E}[g_{ij}(\beta)]$ suffices for consistency, but we will need differentiability for the second step refinement. Condition (b) implies the finiteness of the lower order moments. It unifies several assumptions on the finiteness of some integer moments into one condition. Condition (c) requires that the population variance of the moment function is bounded away from 0, that is, each moment function $g_j(Z,\beta)$ is non-degenerate. This assumption can be relaxed to allow the lower bound to approach 0 as n increases, at the price of a more sophisticated expression of the rate of convergence. Condition (d) is a high-level condition to control the maximal approximation error of the sample variances to the population variances.

The basic logic of the M-estimation is to use sample means to mimic population means. The above two assumptions ensure the following uniform convergence of the mean and variance.

Lemma 1. *Under Assumption 1 and 2, we have*

$$\sup_{\beta\in\mathcal{B},j\leq m}\left|\left(\mathbb{E}_n-\mathbb{E}\right)\left[g_{ij}\left(\beta\right)\right]\right| = O_{\mathrm{p}}\left(\sqrt{\left(\log m\right)/n}\right) \tag{7}$$

$$\sup_{\beta\in\mathcal{B},j\leq m}\left|\widehat{\sigma}_j^2\left(\beta\right)-\sigma_j^2\left(\beta\right)\right| = O_{\mathrm{p}}\left(\sqrt{\left(\log m\right)/n}\right). \tag{8}$$

We know $\sup_{\beta\in\mathcal{B}}\left|\left(\mathbb{E}_n-\mathbb{E}\right)\left[g_{ij}\left(\beta\right)\right]\right|=O_{\mathrm{p}}\left(n^{-1/2}\right)$ holds for a fixed j under those assumptions. The large-dimensionality of m slows the rate of convergence by a factor $O\left(\sqrt{\log m}\right)$. This result became available recently thanks to the advance in the probability theory of moderate deviation of self-normalized sums (Jing et al., 2003), and the theory is applied in econometrics to handle high-dimensional variables by Belloni et al. (2012).

To recover the true parameter from the sample, we have to assume identification. Before we state this assumption, we introduce some notation. Let $\phi_{\min}\left(\cdot\right)$ be the minimal eigenvalue of a matrix. Let $G_{ij}\left(\beta\right)=\partial g_{ij}\left(\beta\right)/\partial\beta'$ be the derivative of $g_{ij}\left(\beta\right)$ with respect to β ($1\times D$ vector), and $G_i\left(\beta\right):=\left(G_{i1}'\left(\beta\right),\ldots,G_{im}'\left(\beta\right)\right)'$ stack $\left(G_{ij}\left(\beta\right)\right)_{j=1}^m$ into a $m\times D$ matrix. Let S be a generic subset of $\{1,\ldots,m\}$ to index submatrices or subvectors. For example, $G_{iS}\left(\beta\right)=\left(\left(G_{ij}'\left(\beta\right)\right)_{j\in S}\right)'$ is a $\#\left\{S\right\}\times D$ matrix, where $\#\left\{\cdot\right\}$ denotes the cardinality of a set. Let $\mathcal{N}_\varepsilon\left(\beta_0\right):=\left\{\beta\in\mathcal{B}:\left\|\beta-\beta_0\right\|_2<\varepsilon\right\}$ be an open neighborhood of radius ε around β_0.

Assumption 3. *For all $n\in\mathbb{N}$,*

(a) There exists a unique $\beta_0\in int\left(\mathcal{B}\right)$ such that $\mathbb{E}\left[g_i\left(\beta_0\right)\right]=0_m$.

(b) For an arbitrarily small fixed $\varepsilon>0$, there exists a strong identification index set $S^*=S_n^*\left(\beta_0\right)$ such that $\#\left\{S^*\right\}=D$ and

$$\inf_{\beta\in\mathcal{N}_\varepsilon\left(\beta_0\right)}\phi_{\min}\left(\mathbb{E}\left[G_{iS^*}\left(\beta\right)\right]'\mathbb{E}\left[G_{iS^*}\left(\beta\right)\right]\right)\geq\eta^\dagger>0,$$

where $\eta^\dagger$ is a universal constant.

Assumption 3(a) assumes that the model is correctly specified, and Condition (b) assumes strong identification by a finite set S^*, which may depend on n and β_0. The constant $\eta^\dagger$ bounds away from zero the smallest singular value of $\mathbb{E}\left[G_{iS^*}\left(\beta_0\right)\right]$. The existence of the strong identification set rules out semi-strong identification, weak identification or identification failure in the population. Strong identification is not necessary for consistency, but simplifies the asymptotic development and rate of convergence.

When S^* is known, we do not need REL. Instead, we can simply use the moments in S^* and apply GMM or EL under the usual conditions to achieve consistency. REL handles the

challenging case when S^* is unknown (as in the first simulation example discussed later in Section 4), or S^* depends on β_0 (as in the second simulation example.) In the first case, we cannot randomly pick a small subset of moments, as those moments may weakly identify or fail to identify the parameter. In the second case, we cannot use different sets in the criterion function for different values of β—such a criterion function would be highly discontinuous and this severely affects the numerical and statistical properties of the estimator.

Under the assumption of correct model specification, if the tuning parameter τ shrinks to zero slowly enough, signal variability is controlled and the true value β_0 falls into the feasible set with high probability. Uniform convergence from Lemma 1 and the identification assumption then suggest that $\widehat{\beta} \xrightarrow{\mathrm{p}} \beta_0$ via standard arguments.

Theorem 1. *Suppose Assumptions 1, 2 and 3 hold. There exists a convergent sequence* (a_n) *such that* $\lim_{n\to\infty} a_n \in (0,\infty)$ *and if we specify* $\tau = a_n\sqrt{(\log m)/n}$*, then*

$$\widehat{\beta} - \beta_0 = O_{\mathrm{p}}\left(\sqrt{(\log m)/n}\right).$$

Remark 1. The sequence (a_n) determines the behavior of $\widehat{\beta}$. Because $\max_{j\leq m} |\mathbb{E}_n[h_{ij}(\beta_0)]| = O_{\mathrm{p}}\left(\sqrt{(\log m)/n}\right)$, if $\lim_{n\to\infty} a_n$ is sufficiently large, with probability approaching one (w.p.a.1.) we have $\max_{j\leq m} |\mathbb{E}_n[h_{ij}(\beta_0)]| < \tau$ as $n \to \infty$, and $\|\widehat{\gamma}(\beta_0)\|_1 = 0$ accordingly. On the other hand, if $a_n \to 0$ in the limit the situation is similar to the standard EL and the primal problem would be infeasible w.p.a.1., in which case $\|\widehat{\gamma}(\beta)\|_1 = \infty$ for all $\beta \in \mathcal{B}$. Consistency of course fails in this case.

Theorem 1 gives the nearly-optimal rate $O_{\mathrm{p}}\left(\sqrt{(\log m)/n}\right)$, where the $\sqrt{\log m}$ factor accounts for not knowing the identities of the strong moments. This reduction in the rate of convergence is a moderate price in many economic applications provided m is not too large. The tuning parameter τ is the key to balance the main signal and signal variability. For consistency it is necessary to shrink τ to zero asymptotically, but the rate must be slow to control the signal variability. This slowly shrinking τ is the source of the slower-than-$\sqrt{n}$ convergence rate of the REL estimator.[3]

Similar results hold for other members of the generalized empirical likelihood (GEL) family. The relaxation can be applied to either EL, exponential tilting or CUE, and the stated consistency and the given rate of convergence follow. We demonstrate the idea of relaxation via EL as it is convenient and the most well-known. The relaxed CUE is particularly interesting. The standard CUE can be viewed as a member of the GMM family due to its quadratic-like criterion function. Besides its favorable asymptotic properties inherited from GEL, the optimization searches only over the low-dimensional parameter

[3] We will discuss in Section 4 the choice of τ in the implementation.

space, as the Lagrangian multipliers can be solved in closed-form in the criterion of CUE. However, this computational advantage is lost after relaxation, since the CUE version of (5) is also non-differentiable in γ.

We close this section with a discussion of the difference between the (generalized) sup-score estimator and REL. The (generalized) sup-score estimator, in our setting, is

$$\widehat{\beta}_{\text{ss}} := \arg\min_{\beta\in\mathcal{B}} \max_{j\leq m} \left|\mathbb{E}_n\left[h_{ij}\left(\beta\right)\right]\right|$$

Under the same conditions, both the sup-score estimator and REL are consistent and converge to β_0 at rate $O_{\text{p}}\left(\sqrt{\left(\log m\right)/n}\right)$. They are closely connected. In particular, if we use $\widehat{\tau}_{\text{ss}} = \max_{j\leq m}\left|\mathbb{E}_n\left[h_{ij}\left(\widehat{\beta}_{\text{ss}}\right)\right]\right|$ as the data-driven tuning parameter for REL and $\widehat{\beta}_{\text{ss}}$ is unique, then the REL estimate is exactly the sup-score estimate, because under $\widehat{\tau}_{\text{ss}}$ only $\widehat{\beta} = \widehat{\beta}_{\text{ss}}$ makes $\{\widehat{p}_i\left(\beta\right) = 1/n\}_{i=1}^n$ hold, which achieves the maximal possible value of the empirical log-likelihood function. Any sequence of tuning parameters $c_n\widehat{\tau}_{\text{ss}}$ with $c_n < 1$ and $c_n \to 1$ is asymptotically valid; the smaller is c_n, the less biased is the estimator, at the price of a larger variance. The key distinction is that, inherited from their primitives, the former uses violation of the moment restrictions as the criterion while the latter separates the criterion and the restriction. It is easier to implant a tuning parameter into REL, either prescribed or data driven, to balance the bias and variance. In fact, REL and the sup-score estimators are complementary procedures. The simplex algorithm remains fast over a low-dimensional parameter space despite its non-differentiability, so the sup-score estimator is computationally faster than REL and can be run from multiple initial values to secure a global optimizer, whereas REL can be computationally expensive to do so over the inner loop and the outer loop. In the simulation examples we will use the former to locate a single starting value for the latter, and the latter will improve the accuracy of the former, at least in these examples, even if we maintain a generic prescribed tuning parameter.

3 Second Step: Bias-Correction

In Section 2, we established the consistency and the nearly-optimal rate of convergence of REL. The first-order bias caused by moment condition relaxation leads to the slower-than-$\sqrt{n}$ rate of convergence, but the relaxation is essential to control the signal variability. In this section, we refine REL via bias-correction. We show that under certain regularity conditions, BC-REL is asymptotically normal; if the true model also satisfies a sparsity condition, BC-REL achieves its lowest variance.

3.1 Infeasible Bias-Corrected REL

We lay out the idea of BC-REL in an infeasible version, which overviews the building blocks to construct the asymptotically normal estimator. Throughout this section, we maintain Assumptions 1, 2 and 3, so that $\widehat{\beta} - \beta_0 = O_{\mathrm{p}}\left(\sqrt{(\log m)/n}\right)$. In contrast to the equality constraints in the primal problem of EL, the REL estimator satisfies the inequalities $\left|\mathbb{E}_{\widehat{p}}\left[h_{ij}\left(\widehat{\beta}\right)\right]\right| \leq \tau$ for all $j \leq m$, as in Proposition 1. Our objective in bias correction is to adjust REL to offset the effect of these nonzero moment conditions. To explain the process, suppose we can differentiate $\mathbb{E}_{\widehat{p}}\left[h_{ij}(\beta)\right]$ with respect to β for all $j \leq m$ and $n \in \mathbb{N}$. Let $H_{ij}(\beta) := \partial h_{ij}(\beta)/\partial\beta$ ($1 \times D$ vector). We take a Taylor expansion in the j-th moment

$$\mathbb{E}_{\widehat{p}}\left[h_{ij}\left(\widehat{\beta}\right)\right] = \mathbb{E}_{\widehat{p}}\left[h_{ij}(\beta_0)\right] + \mathbb{E}_{\widehat{p}}\left[H_{ij}\left(\dot{\beta}\right)\right]\left(\widehat{\beta} - \beta_0\right),$$

where $\dot{\beta}$ is on the line segment joining $\widehat{\beta}$ and β_0. Rearranging the equation and multiplying $\sqrt{n}$ on both sides, we have m conditions of the form

$$\mathbb{E}_{\widehat{p}}\left[H_{ij}\left(\dot{\beta}\right)\right]\sqrt{n}\left(\widehat{\beta} - \beta_0\right) - \sqrt{n}\mathbb{E}_{\widehat{p}}\left[h_{ij}\left(\widehat{\beta}\right)\right] = -\sqrt{n}\mathbb{E}_{\widehat{p}}\left[h_{ij}(\beta_0)\right], \quad j \leq m. \tag{9}$$

We seek to use these conditions to isolate the impact of the bias on the estimator $\widehat{\beta}$ implied by the second term on the left-hand side of (9). This process can be achieved by a weighted least squares regression, but only a small subset of moments can be used, as the signal variability would again accumulate had we used all moments. The question then turns on the selection and weighting of the moment conditions.

Since $\dot{\beta}$ is unknown, we use $\mathbb{E}_{\widehat{p}}\left[H_{ij}\left(\widehat{\beta}\right)\right]$ to approximate $\mathbb{E}_{\widehat{p}}\left[H_{ij}\left(\dot{\beta}\right)\right]$. Then, if we can extract $\#\{S\}$ rows from the $m \times D$ matrix $\mathbb{E}_{\widehat{p}}\left[H_i\left(\widehat{\beta}\right)\right]$ and stack them into a $\#\{S\} \times D$ full rank matrix $\mathbb{E}_{\widehat{p}}\left[H_{iS}\left(\widehat{\beta}\right)\right]$, a $D \times D$ Hermitian matrix $\mathbb{E}_{\widehat{p}}\left[H_{iS}\left(\widehat{\beta}\right)\right]' W \mathbb{E}_{\widehat{p}}\left[H_{iS}\left(\widehat{\beta}\right)\right]$ can be constructed, where W is a compatible full rank weighting matrix.

For the moment, we assume that we have an *oracle* about a sequence of sets that makes the Hermitian matrix invertible with high probability for a large n. Such a sequence leads to a mechanism for bias-correction. This mechanism is *infeasible* because such an oracle is unavailable in practice. We resolve in Section 3.3 the problem of moment selection in the absence of an oracle.

We will encounter a multitude of vectors and matrices indexed by S, a subset of $\{1, \ldots, m\}$. We address the notation first. Let

$$V^{\widehat{p}}_{ijj'}(\beta) := \left(h_{ij}(\beta) - \mathbb{E}_{\widehat{p}}\left[h_{ij}(\beta)\right]\right)\left(h_{ij'}(\beta) - \mathbb{E}_{\widehat{p}}\left[h_{ij'}(\beta)\right]\right),$$

with the superscript $\widehat{p}$ meaning that the sample mean is calculated under the $\mathbb{E}_{\widehat{p}}[\cdot]$ empirical

measure. Let the sample sub-covariance matrix be

$$\mathbb{E}_{\widehat{p}}\left[V_{iSS}^{\widehat{p}}(\beta)\right]=\left(\mathbb{E}_{\widehat{p}}\left[V_{ijj'}^{\widehat{p}}(\beta)\right]\right)_{j,j'\in S},$$

the corresponding weighting matrix be

$$W_{S}^{\widehat{p}}(\beta):=\left(\mathbb{E}_{\widehat{p}}\left[V_{iSS}^{\widehat{p}}(\beta)\right]\right)^{-}$$

where $(\cdot)^{-}$ is the Moorse-Penrose pseudo-inverse, and a sandwich-form matrix be

$$\Psi_{S}^{\widehat{p}}(\beta):=\mathbb{E}_{\widehat{p}}\left[H_{iS}(\beta)\right]'W_{S}^{\widehat{p}}(\beta)\,\mathbb{E}_{\widehat{p}}\left[H_{iS}(\beta)\right].$$

Next, define $h_{ij}^{\sigma}(\beta):=g_{ij}(\beta)/\sigma_{j}(\beta)$. The only difference between $h_{ij}(\beta)$ and $h_{ij}^{\sigma}(\beta)$ is that the former is standardized by the sample standard deviation $\widehat{\sigma}_{j}(\beta)$, whereas the latter is standardized by the population deviation $\sigma_{j}(\beta)$. Similarly, let

$$\begin{aligned}V_{ijj'}^{0,\sigma}(\beta)&:=\left(h_{ij}^{\sigma}(\beta)-\mathbb{E}\left[h_{ij}^{\sigma}(\beta)\right]\right)\left(h_{ij'}^{\sigma}(\beta)-\mathbb{E}\left[h_{ij'}^{\sigma}(\beta)\right]\right)\\W_{S}^{0,\sigma}(\beta)&:=\left(\mathbb{E}\left[V_{iSS}^{0,\sigma}(\beta)\right]\right)^{-}\\\Psi_{S}^{0,\sigma}(\beta)&:=\mathbb{E}\left[H_{iS}^{\sigma}(\beta)\right]'W_{S}^{0,\sigma}(\beta)\,\mathbb{E}\left[H_{iS}^{\sigma}(\beta)\right],\end{aligned}$$

where $H_{ij}^{\sigma}(\beta)=G_{ij}(\beta)/\sigma_{j}(\beta)$. The superscript "0" emphasizes their association with the population, and "σ" indicates the non-randomness in the scale-normalization. The quantities $V_{ijj'}^{0,\sigma}(\beta)$, $W_{S}^{0,\sigma}(\beta)$ and $\Psi_{S}^{0,\sigma}(\beta)$ are the population counterparts of $V_{ijj'}^{\widehat{p}}(\beta)$, $W_{S}^{\widehat{p}}(\beta)$ and $\Psi_{S}^{\widehat{p}}(\beta)$, respectively.

The inverse of the $D\times D$ matrix $\Psi_{S}^{0,\sigma}(\beta_{0})$ in this triangular array asymptotics resembles the semiparametric efficiency bound for a model with moments $\{\mathbb{E}\left[g_{ij}(\beta_{0})\right]=0\}_{j\in S}$. We use $\Psi_{S}^{\widehat{p}}\left(\widehat{\beta}\right)$, its empirical counterpart, as the base of the information criterion for moment selection, and this matrix also appears in the bias-correction formula below. To estimate $\Psi_{S}^{\widehat{p}}\left(\widehat{\beta}\right)$ from the sample, we need to compute $\mathbb{E}_{\widehat{p}}\left[H_{iS}\left(\widehat{\beta}\right)\right]$ and $W_{S}^{\widehat{p}}\left(\widehat{\beta}\right)$.

When we take the inverse of the sub-covariance matrix $\mathbb{E}_{\widehat{p}}\left[V_{iSS}^{\widehat{p}}\left(\widehat{\beta}\right)\right]$, the approximation error of $\mathbb{E}_{\widehat{p}}\left[V_{iSS}^{\widehat{p}}\left(\widehat{\beta}\right)\right]$ to $\mathbb{E}\left[V_{iSS}^{0,\sigma}(\beta_{0})\right]$ will be amplified if the minimal eigenvalue of $\mathbb{E}\left[V_{iSS}^{0,\sigma}(\beta_{0})\right]$ is too small. In the extreme case when $\#\{S\}>n$, the empirical sub-covariance matrix is always rank deficient. To stabilize the inverse, we must control the cardinality of S and restrict the minimal eigenvalue of $\mathbb{E}\left[V_{iSS}^{0,\sigma}(\beta_{0})\right]$.

Let $\mathcal{S}_{K}:=\{S\subset\{1,\ldots,m\}:\#\{S\}\leq K\}$ be the class of subsets of $\{1,\ldots,m\}$ with at

most K members. We restrict to

$$\bar{m} := \left\lfloor (n/\log m)^{1/4} \right\rfloor$$

as the maximal number of moments we consider for each of the D components of the parameter β, where $\lfloor a \rfloor$ denotes the smallest integer greater or equal to $a \in \mathbb{R}$. The parameter β has D components, so we can choose at most $D\bar{m}$ moments, which is of much smaller order than n. After controlling the cardinality, we impose the following assumption, a counterpart of the *restricted minimal eigenvalue* (Bickel et al., 2009).

Assumption 4. *For all $n \in \mathbb{N}$, there exists $\eta^* = \eta_n^* > 0$ such that*

$$\min_{S \in \mathcal{S}_{(2 \vee D)\bar{m}}} \phi_{\min}\left(\mathbb{E}\left[V_{iSS}^{0,\sigma}(\beta_0)\right]\right) \geq \eta^*$$

and $\eta^* / \left[(\log m)/n\right]^{1/8} \to \infty$.

Remark 2. Due to the scale normalization, the diagonal elements of $\mathbb{E}\left[V_{iSS}^{0,\sigma}(\beta_0)\right]$ are all equal to 1, so that $\eta^* \leq 1$. When $D = 1$, the restricted minimal eigenvalue is imposed on sets of cardinality $2\bar{m}$, rather than $\bar{m}$, to give leeway to accommodate redundancy in the selection step. When $D \geq 2$, the bias-correction formula below requires this restricted minimal eigenvalue condition being satisfied on the sets of cardinality $D\bar{m}$.

Denote

$$\begin{aligned}
\xi_S^{\widehat{p}}\left(\widehat{\beta}, \beta_0\right) &:= \mathbb{E}_{\widehat{p}}\left[H_{iS}\left(\widehat{\beta}\right)\right]' W_S^{\widehat{p}}\left(\widehat{\beta}\right) \mathbb{E}_{\widehat{p}}\left[h_{iS}(\beta_0)\right] \\
\xi_S^{\widehat{p}}\left(\widehat{\beta}\right) &:= \mathbb{E}_{\widehat{p}}\left[H_{iS}\left(\widehat{\beta}\right)\right]' W_S^{\widehat{p}}\left(\widehat{\beta}\right) \mathbb{E}_{\widehat{p}}\left[h_{iS}\left(\widehat{\beta}\right)\right] \\
\mathbb{B}_S^{\widehat{p}}\left(\widehat{\beta}\right) &:= \left(\Psi_S^{\widehat{p}}\left(\widehat{\beta}\right)\right)^{-1} \xi_S^{\widehat{p}}\left(\widehat{\beta}\right).
\end{aligned}$$

The $D \times \#\{S\}$ matrix $\mathbb{E}_{\widehat{p}}\left[H_{iS}\left(\widehat{\beta}\right)\right]' W_S^{\widehat{p}}\left(\widehat{\beta}\right)$ gives weight to the $\#\{S\}$ moments $\mathbb{E}_{\widehat{p}}\left[h_{iS}\left(\widehat{\beta}\right)\right]$. Evaluated at $\widehat{\beta}$, the $D \times 1$ vector $\xi_S^{\widehat{p}}\left(\widehat{\beta}\right)$ is a weighted combination of the biased moment estimate $\mathbb{E}_{\widehat{p}}\left[h_{iS}\left(\widehat{\beta}\right)\right]$. Pre-multiplying $\xi_S^{\widehat{p}}\left(\widehat{\beta}\right)$ by $\left(\Psi_S^{\widehat{p}}\left(\widehat{\beta}\right)\right)^{-1}$ transforms the bias in the non-zero moment conditions to the bias in the parameter estimator. $\mathbb{B}_S^{\widehat{p}}\left(\widehat{\beta}\right)$ is the estimated bias of $\widehat{\beta}$, and its formula resembles the weighted least squares estimator. The bias-corrected estimator thus follows. On a sequence of sets $(S_{\text{ifs},n})_{n \in \mathbb{N}}$, where the subscript "ifs" indicates infeasibility, holds the asymptotic normality of the bias-corrected estimator provided that the conditions in the statement are satisfied.

Proposition 2. *Suppose* $(S_{\text{ifs}} = S_{\text{ifs},n} \in \mathcal{S}_{D\bar{m}})_{n\in\mathbb{N}}$ *is a sequence of sets such that*

$$\Psi^{\widehat{p}}_{S_{\text{ifs}}}\left(\widehat{\beta}\right) - \Psi^{0,\sigma}_{S_{\text{ifs}}}\left(\beta_0\right) \xrightarrow{\text{p}} 0_{D\times D} \tag{10}$$

and

$$\left(\Psi^{\widehat{p}}_{S_{\text{ifs}}}\left(\widehat{\beta}\right)\right)^{-1/2}\sqrt{n}\xi^{\widehat{p}}_{S_{\text{ifs}}}\left(\widehat{\beta},\beta_0\right) \Rightarrow N\left(0, I_D\right), \tag{11}$$

where I_D *is the* $D \times D$ *identity matrix, then*

$$\left(\Psi^{\widehat{p}}_{S_{\text{ifs}}}\left(\widehat{\beta}\right)\right)^{1/2}\sqrt{n}\left(\tilde{\beta}_{S_{\text{ifs}}} - \beta_0\right) \Rightarrow N\left(0, I_D\right)$$

where $\tilde{\beta}_S := \widehat{\beta} - \mathbb{B}^{\widehat{p}}_S\left(\widehat{\beta}\right)$ *is the bias-corrected estimator for a given set* S*.*

We establish (10) and (11), two high-level conditions, in Section 3.2, and develop a data-driven mechanism for moment selection in Section 3.3.

3.2 Convergence

In this subsection, we discuss the sufficient conditions for (10) and (11). Denote $\|\cdot\|_\infty$ as the "max norm" such that $\|A\|_\infty = \max_{j,j'}\left|a_{jj'}\right|$. Let $v^0_{jj'}$ be the (j,j')-th entry of the $m \times m$ covariance matrix $\mathbb{E}\left[V^{0,\sigma}_i\left(\beta_0\right)\right]$. By the scale normalization, for all $j, j' \leq m$, a diagonal entry $v^0_{jj} = 1$, and an off-diagonal entry $\left|v^0_{jj'}\right| \leq 1$, $j \neq j'$, equals the correlation coefficient of the j and j'-th moments.

Assumption 5. *For some universal constant* C*,*

(a) $\sup_{n\in\mathbb{N}}\sup_{j\leq m}\sum_{j'=1}^m\left|v^0_{jj'}\right| < C$.

(b) $\sup_{n\in\mathbb{N}}\sup_{\beta\in\mathcal{N}_\varepsilon(\beta_0)}\left\|\sum_{j=1}^m\mathbb{E}\left[G_{ij}\left(\beta\right)\right]'\mathbb{E}\left[G_{ij}\left(\beta\right)\right]\right\|_\infty < C$.

(c) $\sup_{\beta\in\mathcal{N}_\varepsilon(\beta_0)}\left\|\mathbb{E}_n\left[G^2_i\left(\beta\right)\right]\right\|_\infty = O_{\text{p}}\left(1\right)$.

(d) $\sup_{\beta\in\mathcal{N}_\varepsilon(\beta_0)}\left\|\left(\mathbb{E}_n - \mathbb{E}\right)\left[G_i\left(\beta\right)\right]\right\|_\infty = O_{\text{p}}\left(\sqrt{\left(\log m\right)/n}\right)$.

Assumption 5(a) restricts the maximal row-wise sum in the (standardized) covariance matrix evaluated at β_0. This is an assumption analogous to a bounded long-run variance in time series. Although time series often have a natural ordering of the autocorrelations, here we do not require an ordering of the moments. Condition (b) restricts the (equally weighted) main signal in the Jacobian to be finite. It rules out the possibility of super-efficiency, which can potentially deliver even stronger results, say faster-than-$\sqrt{n}$ rate of convergence. Conditions (c) and (d) are high-level assumptions. The rate of convergence

$\sqrt{(\log m)/n}$ in Condition (d) simplifies the explicit expressions of the rates in the results below. It turns out a slight strengthening of Assumption 2(a) suffices for Condition (d) (See Appendix A.12). Assumption 5 leads to the following lemma.

Lemma 2. *Under Assumptions 1, 2, 3, 4 and 5, we have*

$$\begin{aligned}
\max_{j\leq m}\left|\mathbb{E}_{\widehat{p}}\left[h_{ij}\left(\widehat{\beta}\right)\right]-\mathbb{E}\left[h_{ij}^{\sigma}\left(\beta_0\right)\right]\right| &= O_{\mathrm{p}}\left(\sqrt{(\log m)/n}\right)\\
\max_{j,j'\leq m}\left|\mathbb{E}_{\widehat{p}}\left[V_{ijj'}^{\widehat{p}}\left(\widehat{\beta}\right)\right]-\mathbb{E}\left[V_{ijj'}^{0,\sigma}\left(\beta_0\right)\right]\right| &= O_{\mathrm{p}}\left(\sqrt{(\log m)/n}\right)\\
\max_{j\leq m,d\leq D}\left|\mathbb{E}_{\widehat{p}}\left[H_{ijd}\left(\widehat{\beta}\right)\right]-\mathbb{E}\left[H_{ijd}^{\sigma}\left(\beta_0\right)\right]\right| &= O_{\mathrm{p}}\left(\sqrt{(\log m)/n}\right).
\end{aligned}$$

The above lemma gives the uniform rate of convergence of various quantities of interest, which are the building blocks of the following convergence result.

Theorem 2. *Suppose Assumptions 1, 2, 3, 4 and 5 hold, then*

(a) $\sup_{S_n\in\mathcal{S}_{(2\vee D)\bar{m}}}\left\|\Psi_{S_n}^{\widehat{p}}\left(\widehat{\beta}\right)-\Psi_{S_n}^{0,\sigma}\left(\beta_0\right)\right\|_{\infty}=O_{\mathrm{p}}\left(\frac{\bar{m}}{\eta^{*2}}\sqrt{\frac{\log m}{n}}\right)$.

(b) Let $\widetilde{m}=\widetilde{m}_n\in\mathbb{N}$. *On every sequence of sets* $\left(S_n\in\mathcal{S}_{(2\vee D)\widetilde{m}}\right)_{n\in\mathbb{N}}$ *such that*

$$\liminf_{n\to\infty}\phi_{\min}\left(\Psi_{S_n}^{0,\sigma}\right)>0 \quad \textit{and} \quad \limsup_{n\to\infty}\left[\widetilde{m}/\left(\eta^{*}\bar{m}\right)\right]\leq 1,$$

we have

$$\left(\Psi_{S_n}^{\widehat{p}}\left(\widehat{\beta}\right)\right)^{-1/2}\sqrt{n}\xi_{S_n}^{\widehat{p}}\left(\widehat{\beta},\beta_0\right)\Rightarrow N\left(0,I_D\right).$$

The rate in Theorem 2(a) is easily interpretable. $\bar{m}$ comes from the sandwich form of $\Psi_S^{0,\sigma}(\beta)$, which involves a weighted sum of at most $(2\vee D)\bar{m}$ terms; η^{*2} emerges in the sample approximation to $W_{S_n}^{0,\sigma}(\beta_0)$, and finally $\sqrt{(\log m)/n}$ follows by Lemma 2 as the uniform rate of convergence. The cardinality of the sets in Theorem 2(a) is reduced to $\widetilde{m}=O\left(\eta^{*}\bar{m}\right)$ to compensate for the scaling factor $\sqrt{n}$. The weak convergence follows by the Liapunov central limit theorem.

3.3 Moment Selection

As explained in Section 3.1, we plan to select a small subset of moments for (9). The conventional method of moment or model selection relies on the computation of an information criterion for each candidate. When m is moderate and $\bar{m}$ is non-trivial, the number of possible sets $\binom{m}{\bar{m}}$ is far beyond current computational feasibility. In this section we seek a computationally feasible approach, and establish the relevant econometric theory.

The procedure we propose selects a set of moments for each of the D components of β via an iterative algorithm. Let $\Psi_S^{\widehat{p},d}(\beta)$ be the d-th diagonal element of $\Psi_S^{\widehat{p}}(\beta)$. For a fixed d, we choose in the first iteration the moment that maximizes $\Psi_S^{\widehat{p},d}\left(\widehat{\beta}\right)$ for all $S \in \mathcal{S}_1$, which is the information evaluated for a single moment. In every following iteration, we combine the chosen moments with a new one according to a criterion that reflects the effect of the chosen set. For $S, T \in \mathcal{S}_{\bar{m}}$ with $S \supset T$, define

$$\Delta\Psi_{S|T}^{\widehat{p}}(\beta) := \Psi_S^{\widehat{p}}(\beta) - \Psi_T^{\widehat{p}}(\beta).$$

This matrix is positive-semidefinite (Breusch et al., 1999, Lemma 1). $\Delta\Psi_{S|T}^{\widehat{p},d}(\beta)$ can be viewed as the increment of information for β_d from T to S. Denote $\widehat{S}_{r-1}^d$ as the selected set in the first $r-1$ iterations. In the r-th iteration, we add one new moment that maximizes $\Delta\Psi_{\left\{\widehat{S}_{r-1},j\right\}|\widehat{S}_{r-1}^d}^{\widehat{p},d}\left(\widehat{\beta}\right)$. We call this algorithm the *boosting-type greedy algorithm*, where "greedy" indicates that it collects the "best looking" moment in each iteration. The idea is inspired by the componentwise boosting approach (Bühlmann and Yu, 2003; Bühlmann, 2006), but we apply the principle of boosting here in a completely different context that allows for nonlinear moment conditions. The algorithm works as follows.

Algorithm

1.1 Set $r = 1$. For a fixed $d \in \{1, \ldots, D\}$, calculate $\Psi_S^{\widehat{p},d}\left(\widehat{\beta}\right)$ for each $S \in \mathcal{S}_1$. Save $\widehat{S}_1^d = \arg\max_{S \in \mathcal{S}_1} \Psi_S^{\widehat{p},d}\left(\widehat{\beta}\right)$.

1.2 Update r to $r+1$. Calculate $\Psi_S^{\widehat{p},d}\left(\widehat{\beta}\right)$ for each $S \in \widehat{\mathcal{S}}_r$ where

$$\widehat{\mathcal{S}}_r = \widehat{\mathcal{S}}_r\left(\widehat{S}_{r-1}^d\right) := \left\{S \in \mathcal{S}_r : \#\{S\} = r, S \supset \widehat{S}_{r-1}^d\right\}$$

for $r \geq 2$. Save $\widehat{S}_r^d = \arg\max_{S \in \widehat{\mathcal{S}}_r} \Delta\Psi_{S|\widehat{S}_{r-1}^d}^{\widehat{p},d}\left(\widehat{\beta}\right)$.

1.3 Iterate Step 1.2 until $r > \widehat{m}$, where $\widehat{m} \in \mathbb{N}$ is a tuning parameter specified by the user.

2 Repeat 1.1–1.3 for each $d \in \{1, \ldots, D\}$. The overall selected set is $\widehat{S}_{\widehat{m}} := \bigcup_{d=1}^D \widehat{S}_{\widehat{m}}^d$.

The tuning parameter $\widehat{m}$ controls the number of selected moments.[4] In each iteration, the Algorithm compares at most m models, so it evaluates at most $Dm\widehat{m}$ times until it gets $\widehat{S}_{\widehat{m}}$, which is much smaller than $\binom{m}{\widehat{m}}$ for a small D and non-trivial $\widehat{m}$. This greedy approach may select redundant moments, as some moments might contribute little after new moments

[4]We can choose different $\widehat{m}$ for different components of β though, in practice using the same $\widehat{m}$ is more convenient.

are added. In this model including redundant moments does not undermine the first-order asymptotics provided that $\widehat{m}$ is controlled.

Under the following sparsity condition, this Algorithm delivers strong results. Let $\Psi^m(\beta_0) := \Psi^{0,\sigma}_{\{1,\ldots,m\}}(\beta_0)$ be the full information matrix, the information in all the m moments, and its d-th diagonal element $\Psi^{m,d}(\beta_0)$ be the information for the d-th parameter. Let $m^* = m^*_n \in \mathbb{N}$, and $\Psi^{m^*,d}(\beta_0) := \sup_{S\in\mathcal{S}_{m^*}} \Psi^{0,\sigma,d}_S(\beta_0)$ be the maximal information contained in an index set of m^* members. Denote $\Delta_d(m, m^*) := \left(\Psi^{m,d} - \Psi^{m^*,d}\right)(\beta_0)$ as the gap between the full information and the maximal information in $S \in \mathcal{S}_{m^*}$.

Assumption 6 (Sparsity). *For some $d \leq D$,*

(a) $\limsup_{n\to\infty} \Psi^{m,d}(\beta_0) < \infty$.

(b) $\Delta_d(m, m^*) \to 0$ *for some sequence* (m^*) *such that*

$$m^* \to \infty \quad \text{and} \quad m^* / \left[\eta^{*2} (n/\log m)^{1/4}\right] \to 0.$$

The full information $\Psi^{m,d}(\beta_0)$ is a desirable target to pursue, as its inverse is the smallest possible asymptotic variance of BC-REL as in Proposition 2. The boundedness in Assumption 6(a) is necessary for this target to be approached from below as the selected moments accumulate. Condition (b) presumes that almost all the information in $\Psi^{m,d}(\beta_0)$ is concentrated in an m^*-member set, with m^* of much smaller order than n.

We establish a key inequality that justifies the greedy algorithm in the population. It implies a lower bound of progress in each iteration when we choose the new moment. Denote $\varphi^*_{\widehat{m}} := \max_{S\in\mathcal{S}_{\widehat{m}}} \phi_{\max}\left(\mathbb{E}\left[V^{0,\sigma}_{iSS}(\beta_0)\right]\right)$, where $\phi_{\max}(\cdot)$ is the maximal eigenvalue of a matrix.

Lemma 3. *Under Assumption 4, for any index set $S \in \mathcal{S}_{\widehat{m}}$, we have*

$$\max_{j\leq m} \Delta\Psi^{0,\sigma,d}_{\{S,j\}|S}(\beta_0) \geq \frac{1}{m^*}\frac{\eta^*}{\varphi^*_{\widehat{m}}} \times \left(\Psi^{m^*,d} - \Psi^{0,\sigma,d}_S\right)(\beta_0). \tag{12}$$

On the left-hand side of (12) is the maximal amount of information we can progress conditioning on S; on the right-hand side the second factor is the "information gap"[5] between the best m^*-member set and the set S, and the first factor indicates that an iteration narrows the information gap by at least a fraction of $\eta^*/\left(m^*\varphi^*_{\widehat{m}}\right)$. The factor m^* is unavoidable as $\Psi^{m^*,d}(\beta_0)$ contains m^* moments. The extra factor reflects the difficulty to work with an ill-conditioned covariance matrix.

[5]When we talk about the information gap on the right-hand side, we presume it is non-negative. This quantity might be negative when $\widehat{m} > m^*$, in which case the inequality trivially holds, since the left-hand side is always non-negative,

Under the sparsity assumption and the regularity conditions, the Algorithm selects a set that approaches the full information. Let "$\asymp$" be the symbol for asymptotic order equivalence; that is, $a_n \asymp b_n$ means there exists a universal constant C such that $1/C \leq \liminf_{n\to\infty} [(a_n/b_n) \wedge (b_n/a_n)] \leq \limsup_{n\to\infty} [(a_n/b_n) \vee (b_n/a_n)] \leq C$.

Theorem 3. *Suppose that Assumptions 1, 2, 3, 4 and 5 hold, and further that Assumption 6 holds for some $d \leq D$. If we choose $\widehat{m}$ such that*

$$\widehat{m} \asymp \eta^* (n/\log m)^{1/4} \quad \text{and} \quad \limsup_{n\to\infty} (\widehat{m}/\bar{m}) \leq 1,$$

then

$$\Psi^{m,d}(\beta_0) - \Psi^{0,\sigma,d}_{\widehat{S}^d_{\widehat{m}}}(\beta_0) \xrightarrow{\text{p}} 0.$$

Remark 3. Here are the heuristics for Theorem 3. Suppose we could run the boosting-type algorithm in the population, where we have no sampling error. Under Assumption 5(a), the Gershgorin circle theorem implies

$$\max_{j\leq m} \Delta\Psi^{0,\sigma,d}_{\{S,j\}|S}(\beta_0) \geq \frac{\eta^*}{Cm^*} \times \left(\Psi^{m^*,d} - \Psi^{0,\sigma,d}_S\right)(\beta_0),$$

so that the gap is narrowed by a factor of $(1 - \eta^*/(Cm^*))$. Denote $\widetilde{S}^d_r$, a counterpart of $\widehat{S}^d_r$, as the selected set in the population boosting up to the r-th iteration. After r iterations we are left with

$$\Psi^{m^*,d} - \Psi^{0,\sigma,d}_{\widetilde{S}^d_r} \leq \left(1 - \frac{\eta^*}{Cm^*}\right)^r \Psi^{m^*,d}$$

Since

$$\frac{\eta^* \widehat{m}}{Cm^*} \asymp \frac{\eta^{*2} (n/\log m)^{1/4}}{m^*} \to \infty$$

by Assumption 6(b), together with Assumption 6(a) we have

$$\lim_{n\to\infty} \left(\Psi^{m^*,d} - \Psi^{0,\sigma,d}_{\widetilde{S}^d_{\widehat{m}}}\right) \leq \lim_{n\to\infty} \left(1 - \frac{\eta^*}{Cm^*}\right)^{\widehat{m}} \Psi^{m^*,d} \leq \exp(-\infty) \limsup_{n\to\infty} \Psi^{m,d}(\beta_0) = 0;$$

As $\Delta_d(m, m^*)$ is assumed to be negligible in the limit, the information gap will shrink to zero. In the sample we must take the randomness into account. We further develop in the population a *weakly* greedy version of the boosting-type selection via weakening the greatest increment by a constant factor $\alpha \in (0, 1)$. That is, the weakened version proceeds with any set $\check{S}^d \in \widehat{\mathcal{S}}^d_r$ such that

$$\Delta\Psi^{0,\sigma,d}_{\check{S}^d|\widetilde{S}^d_{r-1}}(\beta_0) \geq (1-\alpha) \max_{S\in\widetilde{\mathcal{S}}^d_r} \Delta\Psi^{0,\sigma,d}_{S|\widetilde{S}^d_{r-1}}(\beta_0).$$

On the other hand, we implement the (fully) greedy algorithm in the sample. Under the

regularity conditions, as $n \to \infty$ the sample version will make larger progress than the weakened population version w.p.a.1. in all the $\widehat{m}$ iterations. Because $\eta^* \widehat{m} / (Cm^*) \asymp (1-\alpha)\, \eta^* \widehat{m} / (Cm^*)$, in the limit the weakly greedy algorithm in the population fills the information gap, so does the greedy sample version, w.p.a.1.

After we repeat the selection for each component of β, we plug $\widehat{S}_{\widehat{m}} = \bigcup_{d=1}^{D} \widehat{S}^d_{\widehat{m}}$ into the formula for bias correction. The feasible BC-REL is

$$\tilde{\beta}_{\widehat{S}_{\widehat{m}}} := \widehat{\beta} - \mathbb{B}^{\widehat{p}}_{\widehat{S}_{\widehat{m}}}.$$

The following result is an implication of Proposition 2 as the conditions in the infeasible version have been established and the sequence of sets is now data-driven.

Theorem 4. *Suppose the assumptions in Theorem 3 hold. Suppose further that Assumption 6 holds for all $d \leq D$, then*

(a) $\left(\Psi^{\widehat{p}}_{\widehat{S}_{\widehat{m}}}\left(\widehat{\beta}\right)\right)^{1/2} \sqrt{n}\left(\tilde{\beta}_{\widehat{S}_{\widehat{m}}} - \beta_0\right) \Rightarrow N\left(0, I_D\right);$

(b) $\Psi^{\widehat{p}}_{\widehat{S}_{\widehat{m}}}\left(\widehat{\beta}\right) - \Psi^{m}\left(\beta_0\right) \xrightarrow{\mathrm{p}} 0_{D\times D}.$

In Theorem 4 as well as Corollary 1 below, asymptotic normality holds invariant to the random choice set $\widehat{S}_{\widehat{m}}$. This result can be used for hypothesis testing or confidence interval construction.

The boosting-type algorithm can garner information even if Assumption 6(b) fails, although in this case the ultimate target $\Psi^{m,d}(\beta_0)$ becomes unattainable as the information is spread into a large number of moments, each of which may contain a tiny amount. Let D^* be an arbitrary fixed finite integer with $D^* \geq D$. The set selected by the Algorithm performs no worse than the best D^*-member set of moments in terms of lifting the diagonal elements of $\Psi^{0,\sigma}_S$. The bias-corrected estimator is still asymptotically normally distributed under a strengthened identification condition.

Similar to $\Psi^{m^*,d}(\beta_0)$, denote $\Psi^{D^*,d}(\beta_0) := \sup_{S\in\mathcal{S}_{D^*}} \Psi^{0,\sigma,d}_S(\beta_0)$. We state the result with no sparsity assumption in the following corollary.

Corollary 1. *Suppose Assumptions 1, 2, 3, 4 and 5 hold.*

(a) If we choose $\widehat{m}$ such that

$$\eta^* \widehat{m} \to \infty, \quad \widehat{m} = O\left(\eta^* \bar{m}\right) \quad \textit{and} \quad \limsup_{n\to\infty} \left[\widehat{m}/\bar{m}\right] \leq 1,$$

then for each $d \leq D$,

$$0 \vee \left(\Psi^{D^*,d} - \Psi^{0,\sigma,d}_{\widehat{S}^d_{\widehat{m}}}\right)(\beta_0) \xrightarrow{\mathrm{p}} 0.$$

(b) For some arbitrarily small fixed constant $\alpha \in (0,1)$, let

$$\mathcal{S}^*_{D\widehat{m}}(\alpha) := \left\{ S \in \mathcal{S}_{D\widehat{m}} : \Psi_S^{0,\sigma,d}(\beta_0) \geq (1-\alpha)\, \Psi_{S^*}^{0,\sigma,d} \text{ for all } d \leq D \right\}.$$

Recall S^* is the strong identification set in Assumption 3(b). If

$$\liminf_{n\to\infty} \inf_{S \in \mathcal{S}^*_{D\widehat{m}}(\alpha)} \phi_{\min}\left(\Psi_S^{0,\sigma}(\beta_0)\right) > 0, \tag{13}$$

then

$$\left(\Psi^{\widehat{p}}_{\widehat{S}_{\widehat{m}}}\left(\widehat{\beta}\right)\right)^{1/2} \sqrt{n}\left(\tilde{\beta}_{\widehat{S}_{\widehat{m}}} - \beta_0\right) \Rightarrow N(0, I_D).$$

Under Assumption 3(b), as the D moments in S^* are linearly independent for all $n \in \mathbb{N}$, the minimal eigenvalue $\phi_{\min}\left(\Psi_{S^*}^{0,\sigma}\right)$ is bounded away from 0, so is each diagonal element of $\Psi_{S^*}^{0,\sigma}$, because the positive-semidefiniteness of $\Psi_{S^*}^{0,\sigma}$ implies $\min_{d\leq D} \Psi_{S^*}^{0,\sigma,d} \geq \phi_{\min}\left(\Psi_{S^*}^{0,\sigma}\right)$. Since α is arbitrarily small, the class of sets $\mathcal{S}^*_{D\widehat{m}}(\alpha)$ contains those index sets whose identification power of each single component of β is at least almost as strong as that of S^*.

When (13) is satisfied, $\phi_{\min}\left(\Psi_{\widehat{S}_{\widehat{m}}}^{0,\sigma}(\beta_0)\right) > 0$ w.p.a.1. as $n \to \infty$ because $\widehat{S}_{\widehat{m}} \in \mathcal{S}^*_{D\widehat{m}}(\alpha)$ w.p.a.1 by Corollary 1(a). In the meantime, even if some of those selected moments contain no information at all, the magnitude of the signal variability is controlled to the order induced by the non-informative moments

$$O_{\mathrm{p}}\left(\frac{\widehat{m}}{\eta^{*2}}\sqrt{\frac{\log m}{n}}\right) = O_{\mathrm{p}}\left(\frac{1}{\eta^*}\left(\frac{\log m}{n}\right)^{1/4}\right) = o_{\mathrm{p}}\left(\left(\frac{\log m}{n}\right)^{1/8}\right),$$

which is dominated by any non-diminishing signal.

The additional assumption in (13) is a sufficient condition for asymptotic normality with no sparsity. It strengthens Assumption 3(b). Generally speaking, identification holds if and only if $\mathbb{E}[G_{iS}(\beta_0)]$ is of full rank. When $\mathbb{E}[G_{iS}(\beta_0)]$ is rank deficient, so is the sandwich-form information matrix $\Psi_S^{0,\sigma}$. Identification may fail for two reasons. First, it fails when the entries in a column of $\mathbb{E}[G_{iS}(\beta_0)]$ are all zeros, in which case the corresponding diagonal element of $\Psi_S^{0,\sigma}$ is zero as well. In our context this is eliminated w.p.a.1. by Conclusion (a). Second, even if $\min_{d\leq D} \Psi_S^{0,\sigma,d}(\beta_0) > 0$, identification may still break down because the columns of $\mathbb{E}[G_{iS}(\beta_0)]$ are linearly dependent. The additional assumption rules out such rank deficiency in the class of sets $\mathcal{S}^*_{D\widehat{m}}(\alpha)$. If (13) fail, there is a pathological set on which the D columns of the Jacobian turn out to be linearly dependent even if each column well identifies a component of the parameter. The greedy algorithm is blind to this type of identification failure, as it does not check the dependence among the components of the parameter. When it is computationally infeasible to evaluate an information criterion for

all $\binom{m}{D}$ combinations, we are not aware of any moment selection method that can avoid this type of identification failure.

4 Simulation

In this section we conduct two simulation exercises to check the finite-sample performance of REL and BC-REL in comparison with the sup-score estimator, EW-GMM and OW-GMM. Section 5 includes a third example tailored to the empirical application.

REL and BC-REL face the choice of the tuning parameters. Tuning parameter selection is a long-standing statistical problem and of great practical relevance. A full development of the optimal selection is beyond the scope of the current paper, and here we use tuning parameters that respect the correct order. In particular, we use $\tau = 0.5\sqrt{(\log m)/n}$ and $\widehat{m} = \left\lfloor (n/\log m)^{1/5} \right\rfloor$ in all the cases. Moreover, to ensure the numerical stability of the inverse of the sample covariance matrix in finite samples, we consider the candidate sets $S \in \left\{ S \in \mathcal{S}_{\widehat{m}} : \phi_{\min}\left(\mathbb{E}_{\widehat{p}}\left[V^{\widehat{p}}_{iSS}\left(\widehat{\beta}\right)\right]\right) \geq \eta \right\}$, where η is a generic small number. Here we set $\eta = 0.02$, but in the asymptotic theory this lower bound is not needed as it is embedded into the choice of $\widehat{m}$.

EW-GMM uses the $m \times m$ identity matrix as the weighting matrix, and the EW-GMM point estimate serves as the preliminary estimator of OW-GMM to construct the "optimal" weighting matrix (We use the Moore-Penrose pseudo inverse as default when the sample covariance matrix is rank deficient).

Multiple local optima of the criterion functions may exist on the parameter space. To increase the chance of capturing the global extremum, we try 121 initial points evenly distributed on the compact parameter space for the sup-score estimator. Since the sup-score estimator is consistent, it is close to the true parameter with high probability in large sample. We therefore use it as the initial value for the optimization of REL and EW-GMM. The OW-GMM uses the the estimate of EW-GMM as the initial value. We run 500 replications for each data generating process (DGP).

4.1 Simulation 1: Linear IV with Sparse Instrumentation

Following Angrist (1990) and Angrist and Krueger (1991), applied economists have often used many IVs in the hope of better identifying the effect of endogenous variables. It was later extended to the estimation of models of conditional moment restrictions, and then high-dimensional IVs. Though nonlinear models are the focus of this paper, the linear IV model is of continuing practical importance so we apply our method in this classic example.

The data is generated as follows. The structural equation is

$$e_i^{(0)} = y_i - (x_{i1}, x_{i2})\,\beta$$

where $e_i^{(0)}$ is the structural error and (x_{i1}, x_{i2}) are two endogenous variables. The true reduced-form equations for the endogenous variables are $x_{i1} = 0.5z_{i1} + 0.5z_{i2} + e_i^{(1)}$ and $x_{i2} = 0.5z_{i3} + 0.5z_{i4} + e_i^{(2)}$, respectively, where $\left(e_i^{(1)}, e_i^{(2)}\right)$ are the reduced-form errors. Each endogenous variable is supported by two relevant IVs. We assume that the econometrician happens to have a large number of IVs $(z_{ij})_{j=1}^m$ orthogonal to the structural error but does not know which ones are relevant. In the DGP we generate $(z_{ij})_{j=1}^m \sim$ i.i.d.$N(0,1)$ and

$$\begin{pmatrix} e_i^{(0)} \\ e_i^{(1)} \\ e_i^{(2)} \end{pmatrix} \sim N\left(\begin{pmatrix} 0 \\ 0 \\ 0 \end{pmatrix}, 0.25 \begin{pmatrix} 1 & \rho & \rho \\ \rho & 1 & 0 \\ \rho & 0 & 1 \end{pmatrix}\right).$$

The correlation ρ between the structural error and the reduced-form errors is the source of endogeneity.

In the simulation we try combinations of dimensionalities and strength of endogeneity, namely $n = 120$ or 240, $m = 85$ or 160, and $\rho = 0.3, 0.6$ or 0.8. The true structural parameter is $\beta_0 = (1,1)'$. Due to the symmetry of x_{i1} and x_{i2}, we report only the estimation of β_1 in Figure 1. In the graph the vertical black line represents the true value, and the shaded region is the kernel density of BC-REL. The curves show the kernel density of the other four estimators.

We observe from the graph that BC-REL, REL and the sup-score estimator are centered around the true value. BC-REL is the best concentrated among the three, and REL is the second. EW-GMM estimators is even more concentrated than BC-REL, but it is miscentered; the bias of OW-GMM is even worse. The concentration and wrong centering are anticipated, as in the linear IV model when $m > n$ the OW-GMM is the OLS estimator of the structural equation.

We use the asymptotic normality of BC-REL for a statistical test. The test statistic is simply $\sqrt{n}\left(\Psi^{\widehat{p}}_{\widehat{S}_{\widehat{m}}}\right)^{1/2}\left(\tilde{\beta}_{\widehat{S}_{\widehat{m}}} - \beta_{\text{Hypo}}\right)$, where β_{Hypo} is some hypothesized value of the parameter. The critical value is the quantile of the standard normal distribution for the desirable test size. We list in Table 1 empirical sizes for the two-sided test with a nominal size of 5%. Some distortion of the nominal size is apparent, although the distortion is substantially reduced when $n = 240$ and $m = 80$.

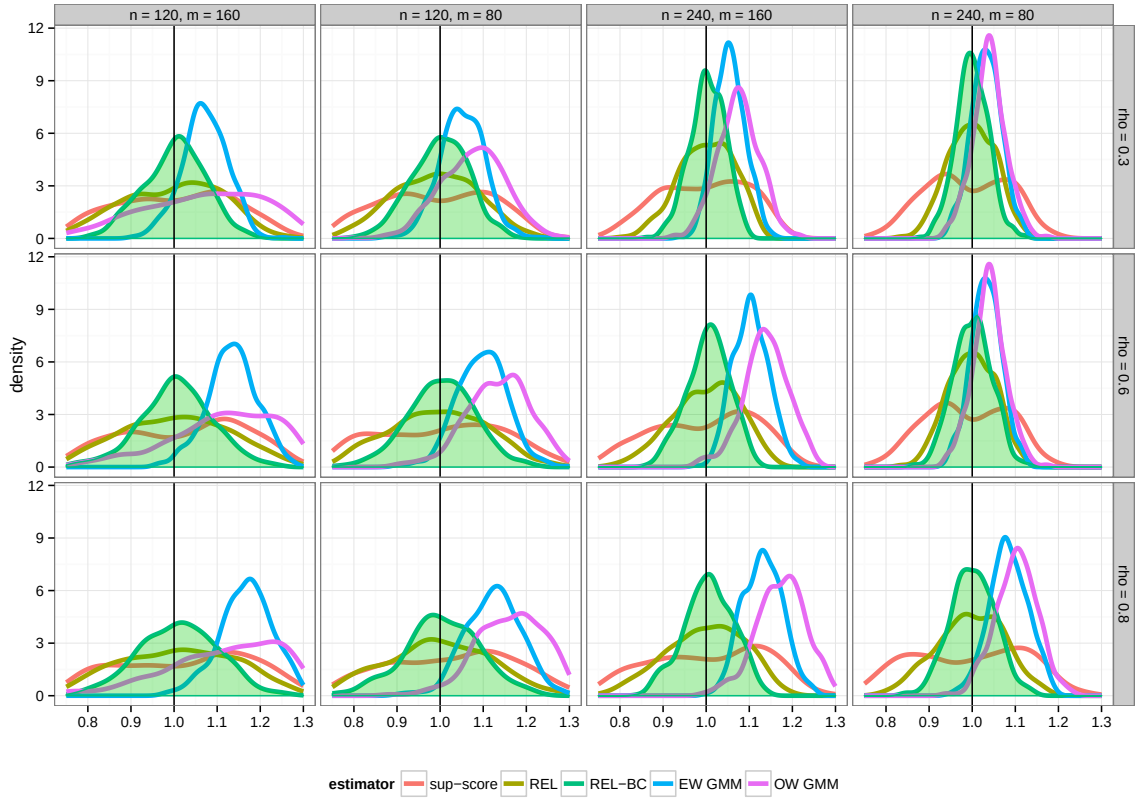


Figure 1: Kernel Density of Estimates in Simulation 1

4.2 Simulation 2: Time-Varying Individual Heterogeneity

The second simulation exercise estimates a panel data model with time-varying individual heterogeneity. The DGP we use is a modification of Example 17 of Han and Phillips (2006), motivated by Han et al. (2005). The structural model is

$$y_{ij} = \lambda_j(\beta_1)\,\alpha_i + \beta_2 + e_{ij},$$

where α_i is an individual-specific heterogeneity that might be correlated with e_{ij}. The functions $\lambda_j(\cdot)$, $j = 0, 1 \ldots, m$, capture the time-varying feature of the original form of the model.[6] In this example we specify $\lambda_j(\beta_1) = 1 + \frac{j}{m}\exp\left(-\frac{1}{3}(\beta_1 m - j)^2\right)$. In this model $\beta_1 m$ is the location of the peak of the time-varying effect, and $\mathbb{E}[\alpha_i] + \beta_2$ is the mean of the initial observation of the i-th time series. The econometrician knows the functional form up to the two parameters $(\beta_1, \beta_2) \in [0.2, 1.01] \times [-5, 1]$.

[6] To keep the notation consistent with the rest of the paper, here we keep using $j \in \{1, \ldots, m\}$, instead of $t = 1, \ldots, T$, to index a time period.

Table 1: Empirical Size of Asymptotic Test in Simulation 1

$(n,,m)$	(120,80)	(120,160)	(240,80)	(240,160)
$\rho = 0.3$	0.224	0.238	0.094	0.154
$\rho = 0.6$	0.180	0.240	0.084	0.146
$\rho = 0.8$	0.194	0.228	0.086	0.172

Eliminating the fixed effect α_i from $y_{i0} = \alpha_i + \beta_2 + e_{i0}$ produces m valid moments

$$\begin{aligned} 0 &= \mathbb{E}\left[e_{ij} - \lambda_j\left(\beta_1\right) e_{i0}\right] \\ &= \mathbb{E}\left[\left(y_{ij} - \lambda_j\left(\beta_1\right)\alpha_i - \beta_2\right) - \lambda_j\left(\beta_1\right)\left(y_{i0} - \beta_2 - \alpha_i\right)\right] \\ &= \mathbb{E}\left[g_{it}\left(\beta\right)\right] \end{aligned}$$

where

$$g_{ij}\left(\beta\right) = y_{ij} - \lambda_j\left(\beta_1\right) y_{i0} - \left(1 - \lambda_j\left(\beta_1\right)\right)\beta_2$$

for $j = 1, \ldots, m$. The derivatives are

$$\begin{aligned} \frac{\partial}{\partial \beta_1} g_{ij}\left(\beta\right) &= \frac{2}{3} j\left(\beta_1 m - j\right)\left(y_{i0} - \beta_2\right)\exp\left(-\frac{1}{3}\left(\beta_1 m - j\right)^2\right) \\ \frac{\partial}{\partial \beta_2} g_{ij}\left(\beta\right) &= \lambda_j\left(\beta_1\right) - 1. \end{aligned}$$

Since $\exp\left(-\frac{1}{3}\left(\beta m - j\right)^2\right)$ decays exponentially fast when j goes away from $\beta_1 m$, only those moments whose indexes are close to $\beta_1 m$ can identify β_1 and β_2. In addition, the identification of β_1 requires $\mathbb{E}\left[y_{i0}\right] \neq \beta_2$. In our DGP, we generate $\alpha_i \sim N\left(1,1\right)$ and $\left(e_{i0}, \ldots, e_{im}\right) \sim$ i.i.d.$N\left(0,1\right)$. The true value of the parameter is set as $\beta_1^0 = 0.9$ and $\beta_2^0 = -2$. As the relevance of the moments varies with the true parameter β_1, to get consistency we have to include all moments in the estimation.

We try $n = 120$ or 240 and $m = 80$ or 160. In this nonlinear optimization, it happens with small probability that the algorithms do not converge, a numerical issue also reported in Han and Phillips (2006). In our case, throughout the 2000 replications, REL, EW-GMM and OW-GMM fail 8 times, 5 times and 10 times, respectively. We remove in total 19 failures (multiple estimators fail in 4 replications). As β_1 is the key parameter that dates the peak of the time-varying heterogeneity, we summarize in Table 2 the bias, variance and mean-squared error (MSE) of the estimation of β_1. It turns out that the bias only takes a tiny proportion of the MSE, so we focus on the comparison of the MSE. REL outperforms the sup-score estimator in all cases, and BC-REL refines REL. Interestingly, in three cases EW-GMM and OW-GMM are more accurate than REL and EW-GMM even holds a small margin over BC-REL in terms of MSE, but these methods perform poorly when $n = 120$

and $m = 160$. With no theoretical guarantee of consistency, the behavior of the GMM estimators is difficult to predict, whereas REL and BC-REL deliver robust outcomes in all DGPs, including the third simulation example in Section 5. In addition, as shown in Table 3 when $n = 240$ the empirical test sizes are close to the nominal size 5% in the two-sided asymptotic test based on the asymptotic normality of BC-REL.

Table 2: Bias, Variance and MSE in Simulation 2

m		$n = 120$			$n = 240$		
		bias$\times 10^5$	var.$\times 10^5$	MSE$\times 10^5$	bias$\times 10^5$	var.$\times 10^5$	MSE$\times 10^5$
80	sup-score	-67.9737	11.6949	11.7411	25.4501	5.4718	5.4783
	REL	-29.4867	3.8937	3.9024	-23.8245	1.0923	1.0980
	BC-REL	-3.0333	0.6219	0.6197	3.4848	0.2408	0.2409
	EW-GMM	6.8262	0.3979	0.3984	-3.2911	0.1953	0.1954
	OW-GMM	11.9971	1.4076	1.4090	-4.4326	0.3016	0.3018
160	sup-score	-4.7627	2.3536	2.3539	1.2425	1.4103	1.4103
	REL	-18.7364	0.6466	0.6501	-4.9872	0.3178	0.3180
	BC-REL	0.5077	0.1509	0.1509	-1.2610	0.0601	0.0601
	EW-GMM	36.4705	2.8803	2.8936	-1.1573	0.0505	0.0506
	OW-GMM	57.2958	5.5382	5.5711	1.0853	0.1460	0.1460

Table 3: Empirical Size of Asymptotic Test in Simulation 2

$(n,,m)$	(120,80)	(120,160)	(240,80)	(240,160)
	0.116	0.094	0.060	0.056

5 Empirical Application

The world's biggest exporter China contributed 10.4% to the world export volume in 2011, and China's trade-to-GDP ratio was 53.1% during 2009–2011.[7] With its overall weight in the international trade network and its economic dependence on exports, China as well as its trading partners can be profoundly affected by a trade policy change or technological innovation. To evaluate *ex ante* the consequence of, for example, a lower quota, tariff, or transportation cost, economists conduct counterfactual experiments in structural models. In the past, empirical studies in international trade were implemented mainly in aggregate cross-sectional or panel datasets at the country-level or state-level. While Eaton et al. (2011) (EKK, henceforth) use a large firm-level dataset, they integrate the empirical model with microeconomic theory. They develop a structural model featured in manufacturers' cost efficiency and heterogeneity, and estimate the unknown parameters from the data.

[7] World Trade Organization, 2013. `http://stat.wto.org/CountryProfile/WSDBCountryPFView.aspx?Country=CN&`

Given these parameters, the structural model can be used to simulate likely results in some counterfactual environments. In the rest of this section, we will adopt EKK's theoretical framework,[8] and estimate by REL and BC-REL the parameters for Chinese exporting firms.

5.1 Data

EKK's estimation demands both domestic and international sales of each firm on each market, where a market is defined by a country. No existing single dataset contains such sale information, so the requisite statistics are obtained by merging datasets.

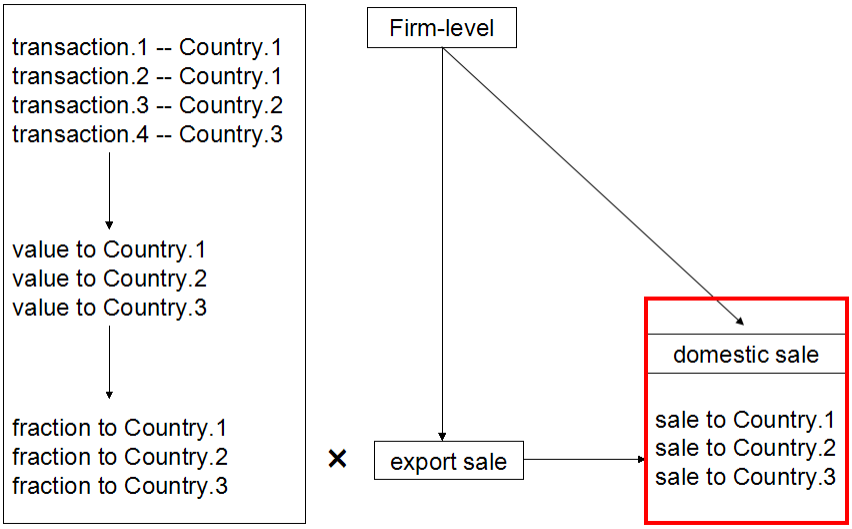


Figure 2: Data-Merge Workflow

We work with two administrative databases, the China Annual Survey of Manufacturing Firms, from China's National Bureau of Statistics, and the China Customs Database, from China's General Administration of Customs. The datasets we obtain are both of a short, wide panel form. The firm-level data spans 2004–2007 and the customs data spans 2000–2006, with three overlapping years. As the economic model is static, we use the 2006 observations as the sample, and the 2005 data to calculate the aggregate market condition, a set of parameters needed to complete the empirical model.

We illustrate in Figure 2 the workflow of the data-merge. We can find in the former dataset the total sale and export sale, and we then take the difference of these two variables as the domestic sale (the middle section of the figure). The latter dataset records each transaction processed by China's customs, with the identity of the exporter, the trading

[8] For detail, we refer interested readers to EKK's paper. We cite only some essential building blocks in the main text and Appendix D.

value, quantity, and the destination country. We compute the weight of a destination country for that firm in terms of the transaction value (the left section). Whenever we identify firms in the two datasets with the same zip code and telephone number, we merge them into one, and multiply the firm's total exporting value by the weight of a country as its sale to that country, as shown by the (red) rectangle in the right section. Since the two datasets adopt different code systems of firm identities, we lose a substantial fraction of observations in the merge. Both the cross-sections contain 65535 unique observations, but we are left with about 10% of the firms after the merge.[9] In particular, we have 6317 merged firms in 2005, and 6754 in 2006.

Suppose we are interested in estimating β through m features of the theoretical prediction $\{\theta_j(\beta)\}_{j=1}^m$. As the complex economic model does not admit an explicit form for $\theta_j(\beta)$, we simulate the behavior of n^{a} artificial firms and take the mean to approximate $\theta_j(\beta)$, that is,

$$\theta_j(\beta) = \int \vartheta_j(\beta|\alpha_i, \mathbf{v}) \operatorname{pdf}(\alpha_i) \, \mathrm{d}\alpha_i \approx \widehat{\theta}_{n^{\mathrm{a}},j}(\beta) = \frac{1}{n^{\mathrm{a}}} \sum_{i=1}^{n^{\mathrm{a}}} \vartheta_j(\beta|\alpha_i, \mathbf{v}), \tag{14}$$

where the conditioning variable α_i is a vector of firm-specific idiosyncratic shocks and $\mathbf{v}$ is a vector of the market condition that we view as fixed and known *ex ante*. $\vartheta_j(\cdot|\alpha_i, \mathbf{v})$ is an explicit function described in Appendix D. The discrepancy between the simulated quantity $\widehat{\theta}_{n^{\mathrm{a}},j}(\beta)$ and the exact value $\theta_j(\beta)$ is arbitrarily small as $n^{\mathrm{a}} \to \infty$. In the simulation we take a large n^{a} and the simulation approximation error is ignored.

The theory of this paper covers M-estimation with differentiable moment functions. We formulate the moments as $g_{ij}(\beta) = y_{ij} - \theta_j(\beta)$ to adapt to M-estimation, where y_{ij} is the j-th feature of the i-th real firm. Correct specification implies $\mathbb{E}[g_{ij}(\beta_0)] = 0$.

5.2 Estimation with Simulated Data

Unlike the two simulation examples in Section 4, EKK's model is highly nonlinear and has no closed-form expressions for the moments let alone the derivatives. We rely on simulated moments and numerical differentiation. Before proceeding to the real data application, we run a simulation to check the quality of estimation. The econometric model involves 5 parameters, which we denote as $\beta = (\beta_1, \ldots, \beta_5)$. The first component $\beta_1 := \beta_{11}/(\beta_{12} - 1)$, where β_{11} is a parameter that measures the distribution of the firms' efficiency and β_{12} is the elasticity of the buyer's utility function. The larger is β_{11}, a larger proportion of firms concentrates on the low efficiency side. β_1 can be identified in the model but not β_{11} and

[9] Yu (2011) describes in detail the datasets and matching, and we follow his handling. Dai et al. (2011, Table 1) provide comparable summary statistics of the merged and unmerged firms, which alleviates the concern of non-random selection in the merging. Not ideal though, at present these are the best accessible Chinese administrative datasets about manufacturers and international trade.

β_{12}. The second component β_2 is a parameter indicating the cost of reaching a fraction of potential buyers. The larger is its value, the less costly it is for the exporter. β_3, β_4 and β_5 are the variance of the demand shock, the variance of the entry cost shock, and the correlation coefficient of the two shocks, respectively.

Evaluating the simulated function takes more computation time than that in Section 4, in which we have explicit forms. We therefore only estimate one unknown parameter here. We simulate the data with the "true" parameter $\beta_0 = (5, 1, 0.3, 0.3, -0.5)$, and suppose that the parameters are known except β_1. We specify the compact parameter space of β_1 as $[4, 6]$.

We try $n = 100, 200, 400$ and $m = 80, 120$. The number of artificial firms is $n^{\mathrm{a}} = 5n$, and the market condition is generated according to Appendix D. Once the parameters are provided, we simulate an artificial firm's entry decision and the sale value on each market. We match the sample mean to the (simulated) population mean for each destination country. The mean sale is a natural quantity of interest, and it avoids the non-differentiability of the quantile-based indicator functions used by EKK. In particular, y_{ij} is the i-th firm's sale on the j-th market, and $\theta_j(\beta)$ is the simulated average sale over the artificial firms on that market.

Figure 3 plots the kernel density of 500 replications of the estimates. When the sample size is small, the sup-score, REL, EW-GMM and OW-GMM all exhibit non-trivial bias, among which REL is the least biased. The sup-score and REL gradually concentrate around the true value as n increases, but EW-GMM and OW-GMM remain wrongly centered with no sign of concentration. BC-REL corrects the bias and is the most concentrated in all cases. Table 4 shows again the empirical size of the two-sided asymptotic test. Size distortion is observed when $n = 100$, while the empirical size is close to the nominal size 5% when $n = 400$.

Table 4: Empirical Size of Asymptotic Test in Simulation 3

$(n, , m)$	(100,60)	(100,120)	(200,60)	(200,120)	(400,60)	(400,120)
	0.268	0.396	0.134	0.166	0.074	0.074

5.3 Estimation with Real Data

The simulation provides favorable evidence for the finite-sample behavior of REL and BC-REL. Now we apply this two-step procedure to the real data. The difference between the simulation example and the real data application is small: (i) y_{ij} is replaced by the real sale of the i-th firm to the j-th country in 2006; (ii) The market condition $\mathbf{v}$ is computed from the 2005 real data and is viewed as non-random. This treatment requires a pre-selection of the countries, which is described in Appendix D. We keep 127 countries in the estimation,

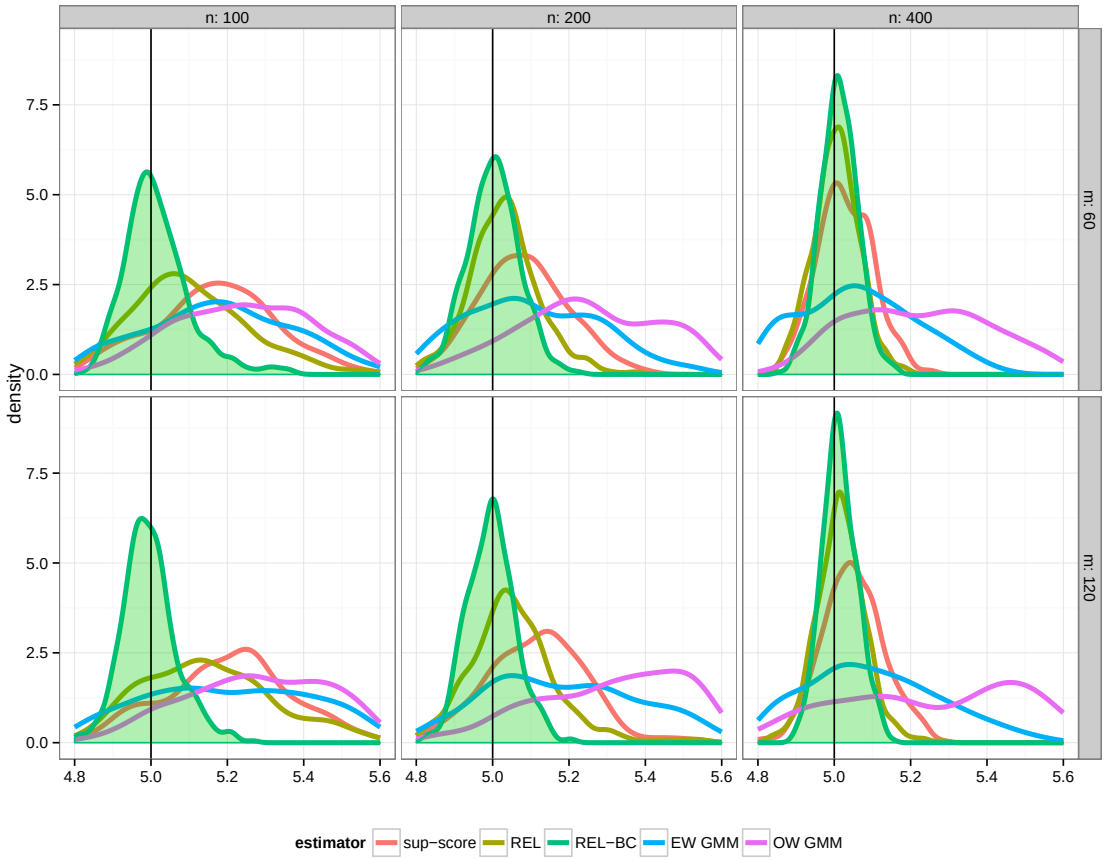


Figure 3: Kernel Density of Estimates in Simulation 3

including the home country China. We simulate 33770 artificial firms (5 times the sample size). Given an initial point, the REL algorithm through the inner loop and the outer loop is affordable although the numerical searching evaluates a big matrix for hundreds or thousands of times, but repeating REL for many initial values can be time-consuming. To get a single sensible initial point, we run only the inner loop for 4000 i.i.d. randomly generated trial values uniformly distributed on a compact parameter space of the 5-dimensional cube

$$\mathcal{B} = [1.5, 10] \times [0.5, 5] \times [0.1, 5] \times [0.1, 5] \times [-0.9, 0.9]\,.$$

This calculation provides a rough idea of the shape of the profile log-likelihood function. It turns out that only on 406 points the inner problem has a solution, meaning that about 90% of the volume of the parameter space is far from the true value. The random trial value that maximizes the profile log-likelihood among the 406 points is displayed in the second row of Table 5. Starting from this initial value, REL's numerical searching terminates at the point

shown in the third row of Table 5.

Table 5: Point and Interval Estimates

	β_1	β_2	β_3	β_4	β_5
estimates from EKK	2.46	0.91	1.69	0.34	-0.65
initial value	6.5439	3.5599	1.9383	0.5503	-0.4138
REL	6.5026	3.5355	2.3174	0.5507	-0.3110
BC-REL	6.4038	6.2533	2.1608	0.5470	-0.3786
standard deviation	0.1912	1.4266	0.0576	0.0060	0.0875
lower bound of 95% CI	6.0290	3.4571	2.0479	0.5352	-0.5501
upper bound of 95% CI	6.7785	9.0494	2.2737	0.5588	-0.2071

Before we move on to BC-REL, we check whether this handling of the initial value can secure the global optimum with high probability. According to the explicit form of ϑ, the shape of the profile log-likelihood function is crucially determined by β_1 and β_2. Since the surface has 5 dimensions, we fix $(\beta_2, \beta_3, \beta_4) = (2.3174, 0.5507, -0.3110)$ and plot the values of the profile log-likelihood in a grid system over the parameter space of (β_1, β_2). The left subgraph of Figure 4 is the 3D image of the function and the right subgraph shows the contour. The interval of β_1 is narrowed to $[4.5, 10]$, because no solution exists for values in $[1.5, 4.5]$.

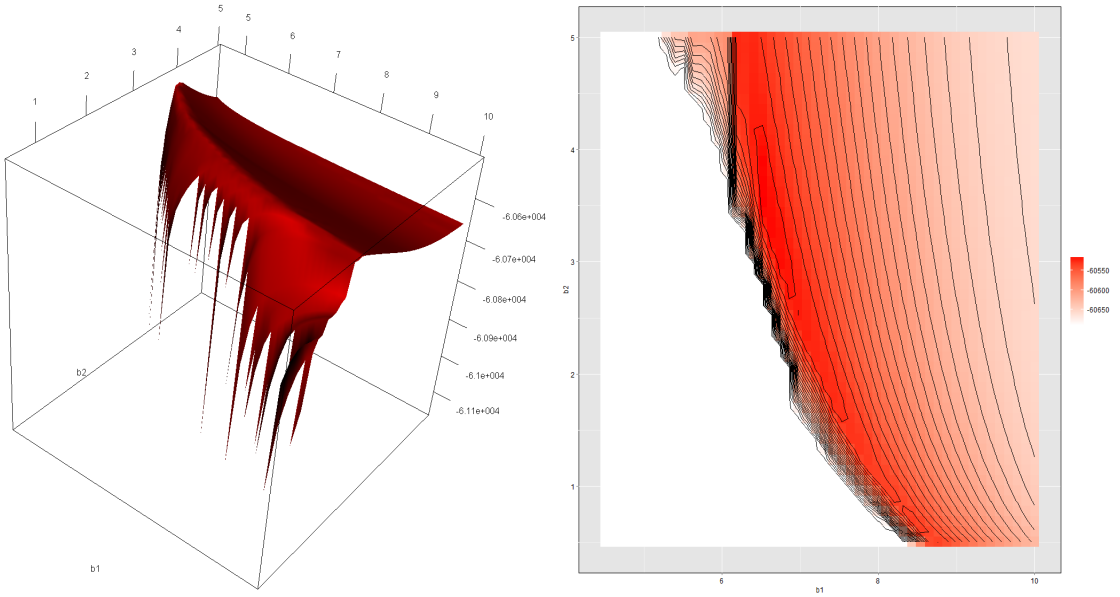
A clear ridge sits near the center of the 3D graph. Left to the ridge is a steep cliff that dramatically falls to minus infinity—REL has no solution there. A gentle downhill lies to the right of the ridge. In the right subgraph the black curves are the contour lines, and the darker the region is, the larger the function value is. The estimate $\left(\widehat{\beta}_1, \widehat{\beta}_2\right)$ is surrounded by the highest closed contour curve. Though other local maxima exist, say, the small ellipse near $(\beta_1, \beta_2) = (8.5, 0.7)$, the shape of the surface is smooth and the global maximum is easy to locate.

In the moment selection of BC-REL, we iterate

$$\widehat{m} = \left\lfloor (n/\log m)^{1/5} \right\rfloor = \left\lfloor (6754/\log 126)^{1/5} \right\rfloor = 5$$

times for each parameter. As the parameter has 5 components, we collect in total 25 moments, among which 14 are unique. The 14 moments correspond to the following countries, ranked by the total exporting volume from high to low: USA, Japan, Australia, Brazil, Iran, Philippine, Argentina, Peru, Kuwait, Algeria, Costa Rica, Yugoslavia,[10] Cambodia, Ethiopia. They are located on the world map of Figure 5. The first three countries are China's first, second and eighth biggest trading partner in terms of the exporting volume. As explained in Section 3, the selection algorithm does not seek the most parsimonious

[10]Despite the political upheaval, Yugoslavia is an artifact used in the database.



(a) 3D plot (b) Contour Plot

Figure 4: Value of Profile Log-Likelihood

set, and may have included some less relevant countries, although asymptotically this does not undermine the properties of the estimator. The BC-REL point estimates, standard deviations, lower bounds and upper bounds of the 95% asymptotic confidence intervals are tabulated in Table 5, respectively.

Amongst the 5 parameters, 4 point estimates of REL fall into the 95% confidence interval of BC-REL. What draws our attention is the BC-REL estimate $\tilde{\beta}_2$, whose point estimate 6.2533 lies out of the conjectured parameter space [0.5,5]. The relatively large variance associated with β_2 is also observed in EKK's estimation, and it may indicate a potential identification problem for this parameter.

Although we use different datasets, different moments and different estimation method, the interpretation of the parameters is the same as EKK's. EKK use 1986 French data, and get the point estimates listed in the first row of Table 5. Our first and second parameter estimates contrast significantly with theirs. Since the model can identify β_1 but not β_{11} and β_{12} separately, we assume the buyers of French goods and those of Chinese goods share the same utility function, so that β_{12} is the same and the difference of β_1 is caused by β_{11}, which measures the distribution of production efficiency. If we take EKK's estimates as the benchmark, the Chinese firms are more concentrated on the low efficiency side. China mainly exports labor-intensive goods, and the cost efficiency is low in labor-intensive industries. β_2 measures the cost of reaching foreign buyers. In the two decades between 1986 and 2006 the communication technology advanced, transportation cost dropped, and trade barriers

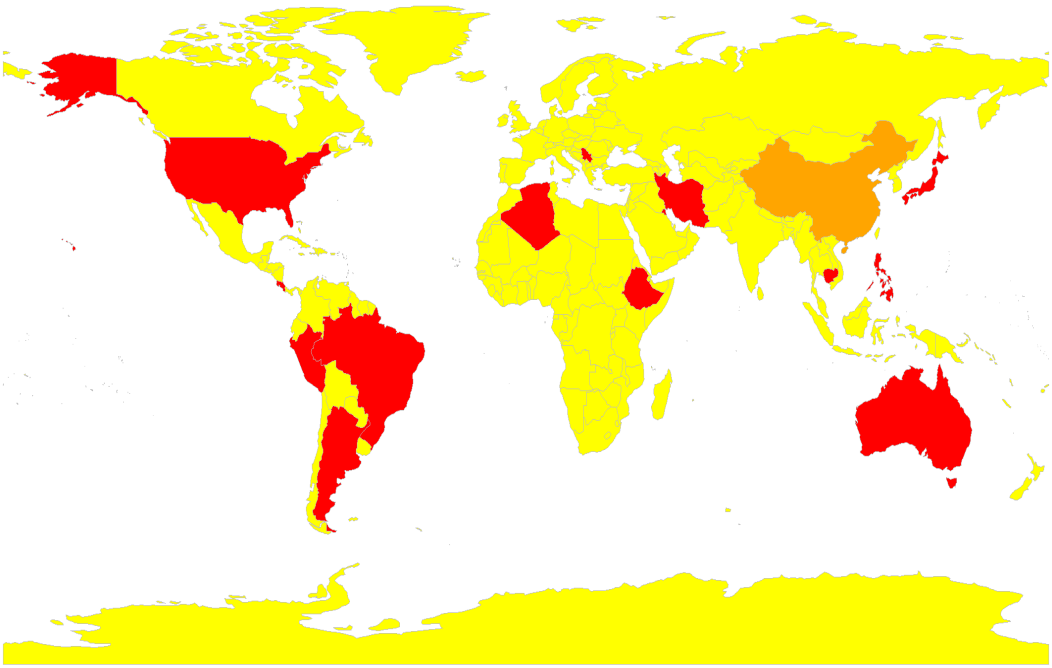


Figure 5: Countries Associated with Selected Moments

lowered: they contributed to the reduction of the transaction cost of international trade. The scale of the estimates of the other three parameters are comparable to EKK's.

6 Conclusion

The two-step procedure proposed in this paper, to the best of our knowledge, contributes the first asymptotically normally distributed estimator in a structural model involving high-dimensional moment equalities. The two steps are specifically constructed to adapt to high-dimensionality. In the first step, we keep all the moments in the estimation while controlling them through relaxation. Under regularity conditions, the uniform convergence of the empirical processes guarantees the consistency of REL with a nearly-optimal rate. The bias caused by the relaxation calls for correction. To achieve asymptotic normality, it is essential to select a small subset of moments. Under the assumption of the existence of a small set of moments that identify the parameter, the boosting-type greedy algorithm progresses in each iteration with the total number of selected moments under control. We plug the selected moments into the bias-correction formula, and the feasible BC-REL converges in distribution to a jointly normal random vector.

Several directions of extension can be carried out under this framework. (i) We can use other consistent estimators with a rate of convergence $O_{\mathrm{p}}\left(\sqrt{(\log m)/n}\right)$ as first-step

preliminary estimators. (ii) Similar to Belloni et al. (2012), it is possible to use the selected moments to re-estimate the parameter by GMM or EL. We can call this the *post-selection* GMM or EL, and we conjecture they are first-order asymptotically equivalent to BC-REL.

The present paper suggests research in the following areas. (i) A more detailed comparison between REL and the sup-score estimator would help understand the optimal rate of the tuning parameter τ. (ii) Including many moments can amplify the risk of misspecification, raising the importance of tests for misspecification and robust estimation when misspecification is detected.

References

Altonji, J. G., J. Anthony A. Smith, and I. Vidangos (2013). Modeling earnings dynamics. *Econometrica 81*, 1395–1454.

Andrews, D. and B. Lu (2001). Consistent model and moment selection procedures for gmm estimation with application to dynamic panel data models. *Journal of Econometrics 101*(1), 123–164.

Angrist, J. and A. Krueger (1991). Does compulsory school attendance affect schooling and earnings? *Quarterly Journal of Economics 106*(4), 979–1014.

Angrist, J. D. (1990). Lifetime earnings and the vietnam era draft lottery: evidence from social security administrative records. *The American Economic Review 80*, 313–336.

Bai, J. and S. Ng (2009). Boosting diffusion indices. *Journal of Applied Econometrics 24*(4), 607–629.

Bai, J. and S. Ng (2010). Instrumental variable estimation in a data rich environment. *Econometric Theory 26*(6), 1577.

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

Belloni, A. and V. Chernozhukov (2011). l1-penalized quantile regression in high-dimensional sparse models. *The Annals of Statistics 39*(1), 82–130.

Belloni, A., V. Chernozhukov, and C. Hansen (2010). LASSO Methods for Gaussian Instrumental Variables Models. Arxiv preprint arXiv:1012.1297.

Belloni, A., V. Chernozhukov, and C. Hansen (2011). Estimation of treatment effects with high-dimensional controls. *arXiv preprint arXiv:1201.0224*.

Belloni, A., V. Chernozhukov, and L. Wang (2011). Square-root lasso: pivotal recovery of sparse signals via conic programming. *Biometrika 98*(4), 791–806.

Bickel, P., Y. Ritov, and A. Tsybakov (2009). Simultaneous analysis of Lasso and Dantzig selector. *Annals of statistics 37*(4), 1705–1732.

Breusch, T., H. Qian, P. Schmidt, and D. Wyhowski (1999). Redundancy of moment conditions. *Journal of econometrics 91*(1), 89–111.

Bühlmann, P. (2006). Boosting for high-dimensional linear models. *The Annals of Statistics 34*(2), 559–583.

Bühlmann, P. and S. van de Geer (2011). *Statistics for High-Dimensional Data: Methods, Theory and Applications*. Springer.

Bühlmann, P. and B. Yu (2003). Boosting with the l 2 loss: regression and classification. *Journal of the American Statistical Association 98*(462), 324–339.

Candes, E. and T. Tao (2007). The Dantzig selector: Statistical estimation when p is much larger than n. *Annals of Statistics 35*(6), 2313–2351.

Caner, M. (2009). Lasso-type GMM estimator. *Econometric Theory 25*(01), 270–290.

Caner, M. and H. Zhang (2013). Adaptive elastic net gmm with diverging number of moments. *Journal of Business and Economics Statistics*, forthcoming.

Carrasco, M. and J.-P. Florens (2000). Generalization of gmm to a continuum of moment conditions. *Econometric Theory 16*(06), 797–834.

Carrasco, M. and J.-P. Florens (2014). On the asymptotic efficiency of gmm. *Econometric Theory*.

Chao, J. and N. Swanson (2005). Consistent estimation with a large number of weak instruments. *Econometrica 73*(5), 1673–1692.

Chao, J. C., N. R. Swanson, J. A. Hausman, W. K. Newey, and T. Woutersen (2011). Asymptotic distribution of jive in a heteroskedastic iv regression with many instruments. *Econometric Theory 28*(1), 42.

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

Cheng, X. and Z. Liao (2012). Select the valid and relevant moments: A one-step procedure for gmm with many moments.

Dai, M., M. Maitra, and M. Yu (2011). Unexceptional exporter performance in china? role of processing trade. *SSRN)*, http://papers.ssrn.comsol3papers.cfm?abstract_id=1963652.

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

Eaton, J., S. Kortum, and F. Kramarz (2011). An anatomy of international trade: Evidence from french firms. *Econometrica 79*, 1453–1498.

Fan, J. and Y. Liao (2011). Ultra high dimensional variable selection with endogenous covariates. Technical report, Working Paper.

Gautier, E. and A. Tsybakov (2013). Pivotal uniform inference in high-dimensional regression with random design in wide classes of models via linear programming. Technical report, CREST, ENSAE.

Hahn, J. (2002). Optimal inference with many instruments. *Econometric Theory 18*(1), 140–168.

Hall, A. R., A. Inoue, K. Jana, and C. Shin (2007). Information in generalized method of moments estimation and entropy-based moment selection. *Journal of Econometrics 138*(2), 488–512.

Hall, A. R. and F. P. Peixe (2003). A consistent method for the selection of relevant instruments. *Econometric Reviews 22*(3), 269–287.

Han, C., L. Orea, and P. Schmidt (2005). Estimation of a panel data model with parametric temporal variation in individual effects. *Journal of Econometrics 126*(2), 241–267.

Han, C. and P. Phillips (2006). Gmm with many moment conditions. *Econometrica 74*(1), 147–192.

Hansen, L. (1982). Large sample properties of generalized method of moments estimators. *Econometrica: Journal of the Econometric Society 50*, 1029–1054.

Hansen, L., J. Heaton, and A. Yaron (1996). Finite-sample properties of some alternative gmm estimators. *Journal of Business & Economic Statistics 14*(3), 262–280.

Hong, H., B. Preston, and M. Shum (2003). Generalized empirical likelihood-based model selection criteria for moment condition models. *Econometric Theory 19*(6), 923–943.

Horn, R. A. and C. R. Johnson (1985). *Matrix analysis*. Cambridge university press.

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

Kitamura, Y. (1997). Empirical likelihood methods with weakly dependent processes. *The Annals of Statistics 25*(5), 2084–2102.

Kitamura, Y. (2001). Asymptotic optimality of empirical likelihood for testing moment restrictions. *Econometrica 69*(6), 1661–1672.

Kitamura, Y. (2006). Empirical likelihood methods in econometrics: Theory and practice.

Kitamura, Y. and M. Stutzer (1997). An information-theoretic alternative to generalized method of moments estimation. *Econometrica: Journal of the Econometric Society 65*, 861–874.

Kitamura, Y., G. Tripathi, and H. Ahn (2004). Empirical likelihood-based inference in conditional moment restriction models. *Econometrica 72*(6), 1667–1714.

Koenker, R. and J. Machado (1999). Gmm inference when the number of moment conditions is large. *Journal of Econometrics 93*(2), 327–344.

Lahiri, S. and S. Mukhopadhyay (2012). A penalized empirical likelihood method in high dimensions. *The Annals of Statistics 40*, 2511–2540.

Liao, Z. (2013). Adaptive GMM Shrinkage Estimation with Consistent Moment Selection. *Econometric Theory*, Forthcoming.

Newey, W. K. and R. J. Smith (2003). Higher order properties of gmm and generalized empirical likelihood estimators. *Econometrica 72*(1), 219–255.

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

Otsu, T. (2007). Penalized empirical likelihood estimation of semiparametric models. *Journal of Multivariate Analysis 98*(10), 1923–1954.

Otsu, T. (2010). On bahadur efficiency of empirical likelihood. *Journal of Econometrics 157*(2), 248–256.

Owen, A. B. (1988). Empirical likelihood ratio confidence intervals for a single functional. *Biometrika 75*(2), 237–249.

Owen, A. B. (2001). *Empirical likelihood*, Volume 92. Chapman & Hall/CRC.

Peña, V., T. Lai, and Q. Shao (2009). *Self-normalized processes: Limit theory and Statistical Applications*. Springer.

Qin, J. and J. Lawless (1994). Empirical likelihood and general estimating equations. *The Annals of Statistics 22*, 300–325.

Tibshirani, R. (1996). Regression shrinkage and selection via the lasso. *Journal of the Royal Statistical Society. Series B (Methodological) 58*, 267–288.

Yu, M. (2011). Processing trade, tarff reductions, and firm productivity: Evidence from chinese products. *SSRN*, http:papers.ssrn.comsol3papers.cfm?abstract_id=1734720.

A Proofs

A.1 Proof of Proposition 1

First we fixed an arbitrary $\beta \in \mathcal{B}$. Since we search for $\{(p_i)_{i=1}^n, \gamma_A, \gamma\}$ to optimize (5), the derivatives and the subderivatives give the necessary condition for optimization

$$\begin{aligned}
\widehat{p}_i(\beta)\left(\widehat{\gamma}_A + \widehat{\gamma}' h_i(\beta)\right) &= 1/n, && (15)\\
\sum_{i=1}^n \widehat{p}_i(\beta) &= 1, && (16)\\
\operatorname{sign}(\widehat{\gamma}_j)\left(\sum_{i=1}^n \widehat{p}_i(\beta) h_{ij}(\beta)\right) &= \tau,\ \forall j \in \{j : \widehat{\gamma}_j \neq 0\}, && (17)\\
\left|\sum_{i=1}^n \widehat{p}_i(\beta) h_{ij}(\beta)\right| &\leq \tau,\ \forall j \in \{j : \widehat{\gamma}_j = 0\}, && (18)
\end{aligned}$$

where (17) and (18) follow by the Karush-Kuhn-Tucker condition. Evaluating (17) and (18) at $\beta = \widehat{\beta}$, we have the second conclusion.

Sum (15) across $i = 1, \ldots, n$,

$$1 = \sum_{i=1}^n \widehat{p}_i(\beta)\left(\widehat{\gamma}_A + \widehat{\gamma}' h_i(\beta)\right) = \widehat{\gamma}_A \sum_{i=1}^n \widehat{p}_i(\beta) + \widehat{\gamma}' \sum_{i=1}^n \widehat{p}_i(\beta) h_i(\beta) = \widehat{\gamma}_A + \widehat{\gamma}' \sum_{i=1}^n \widehat{p}_i(\beta) h_i(\beta).$$

Rearrange the equation we have $\widehat{\gamma}_A = 1 - \widehat{\gamma}' \sum_{i=1}^n \widehat{p}_i(\beta) h_i(\beta)$. If we can show

$$\widehat{\gamma}' \sum_{i=1}^n \widehat{p}_i(\beta) h_i(\beta) = \tau \|\widehat{\gamma}\|_1, \tag{19}$$

then we can explicitly solve $\widehat{\gamma}_A = 1 - \tau \|\widehat{\gamma}\|_1$. In view of (15), we can then solve $\widehat{p}_i(\beta)$ as

$$\widehat{p}_i(\beta) = \left[n\left(1 + \widehat{\gamma}' h_i(\beta) - \tau \|\widehat{\gamma}\|_1\right)\right]^{-1}. \tag{20}$$

Substitute out $p_i = \widehat{p}_i(\beta)$ into the penalized Lagrangian (5). We transform (5) to (6) since the first term of (5) is

$$\begin{aligned}
L_n^\tau &= -\mathbb{E}_n\left[\log\left(n\left(1 + \widehat{\gamma}' h_i(\beta) - \tau \|\widehat{\gamma}\|_1\right)\right)\right] + \log n\\
&= -\mathbb{E}_n\left[\log\left(1 + \widehat{\gamma}' h_i(\beta) - \tau \|\widehat{\gamma}\|_1\right)\right],
\end{aligned}$$

the second term is 0 and the (19) cancels out the third and the fourth terms.

Now we verify (19). Decompose

$$\begin{aligned}
& \widehat{\gamma}' \sum_{i=1}^{n} \widehat{p}_i(\beta) h_i(\beta) - \tau \|\widehat{\gamma}\|_1 \\
= \; & \widehat{\gamma}' \sum_{i=1}^{n} \widehat{p}_i(\beta) h_i(\beta) - \tau \widehat{\gamma}' \text{sign}(\widehat{\gamma}) \\
= \; & \sum_{\{j:\widehat{\gamma}_j=0\}} \widehat{\gamma}_j \left(\sum_{i=1}^{n} \widehat{p}_i(\beta) h_{ij}(\beta) - \tau \text{sign}(\widehat{\gamma}_j) \right) + \sum_{\{j:\widehat{\gamma}_j\neq 0\}} \widehat{\gamma}_j \left(\sum_{i=1}^{n} \widehat{p}_i(\beta) h_{ij}(\beta) - \tau \text{sign}(\widehat{\gamma}_j) \right) \\
= \; & \sum_{\{j:\widehat{\gamma}_j\neq 0\}} \widehat{\gamma}_j \left(\sum_{i=1}^{n} \widehat{p}_i(\beta) h_{ij}(\beta) - \tau \text{sign}(\widehat{\gamma}_j) \right) \\
= \; & \sum_{\{j:\widehat{\gamma}_j\neq 0\}} \widehat{\gamma}_j \left(\sum_{i=1}^{n} \widehat{p}_i(\beta) h_{ij}(\beta) - \text{sign}^2(\widehat{\gamma}_j) \sum_{i=1}^{n} \widehat{p}_i(\beta) h_{ij}(\beta) \right) = 0
\end{aligned}$$

where (17) justifies the fourth equality.

We have shown for a β such that the primal problem is feasible, maximizing (5) by $\{(p_i)_{i=1}^n, \gamma_A, \gamma\}$ and minimizing (6) by γ are equivalent. The equivalence holds at $\beta = \widehat{\beta}$. □

A.2 Proof of Lemma 1

We first work with a fixed $\beta \in \mathcal{B}$. Under Assumption 1(a)(b) and 2(b)(c), Lemma 7 tells us

$$\max_{j\leq m} \left| \frac{(\mathbb{E}_n - \mathbb{E})[g_{ij}(\beta)]}{\widehat{\sigma}_j(\beta)} \right| = O_{\text{p}}\left(\sqrt{\frac{\log m}{n}} \right).$$

To get rid of the denominator term $\widehat{\sigma}_j(\beta)$ on the left-hand side of the above equation,

$$\begin{aligned}
\max_{j\leq m} |(\mathbb{E}_n - \mathbb{E})[g_{ij}(\beta)]| &\leq \max_{j\leq m} \left| \frac{(\mathbb{E}_n - \mathbb{E})[g_{ij}(\beta)]}{\widehat{\sigma}_j(\beta)} \right| \max_{j\leq m} \widehat{\sigma}_j(\beta) \\
&\leq O_{\text{p}}\left(\sqrt{(\log m)/n} \right) \max_{j\leq m} \widehat{\sigma}_j(\beta). \qquad (21)
\end{aligned}$$

By Assumption 2(d),

$$\max_{j\leq m} \widehat{\sigma}_j^2(\beta) \leq \max_{j\leq m} \left[\mathbb{E}_n \left[g_{ij}^2(\beta) \right] \right] \leq \max_{j\leq m} \left[\sqrt{\mathbb{E}_n \left[g_{ij}^4(\beta) \right]} \right] = O_{\text{p}}(1). \qquad (22)$$

Under Assumption 2(b)(d) we invoke Lemma 8 to obtain

$$\max_{j\leq m} \left| (\mathbb{E}_n - \mathbb{E}) \left[g_{ij}^2(\beta) \right] \right| = o_{\text{p}}\left(\sqrt{(\log m)/n} \right) \qquad (23)$$

and furthermore

$$\begin{aligned}
& \max_{j\leq m}\left|\widehat{\sigma}_j^2(\beta)-\sigma_j^2(\beta)\right| \\
= & \max_{j\leq m}\left|\mathbb{E}_n\left[g_{ij}^2(\beta)\right]-\mathbb{E}\left[g_{ij}^2(\beta)\right]+\left(\mathbb{E}_n\left[g_{ij}(\beta)\right]\right)^2-\left(\mathbb{E}\left[g_{ij}(\beta)\right]\right)^2\right| \\
= & \max_{j\leq m}\left|\mathbb{E}_n\left[g_{ij}^2(\beta)\right]-\mathbb{E}\left[g_{ij}^2(\beta)\right]\right|+\max_{j\leq m}\left|\left(\mathbb{E}_n\left[g_{ij}(\beta)\right]\right)^2-\left(\mathbb{E}\left[g_{ij}(\beta)\right]\right)^2\right| \\
= & O_{\mathrm{p}}\left(\sqrt{(\log m)/n}\right).
\end{aligned}$$

Assumption 2(a) carries over the pointwise consistency for a fixed β to the uniform convergence over $\beta\in\mathcal{B}$.

A.3 Proof of Theorem 1

Given (8) in Lemma 1 and Assumption 2(c), for some sufficiently large finite C the event

$$E_1:=\left\{1/C<\inf_{\beta\in\mathcal{B},j\leq m}\widehat{\sigma}_j(\beta)\leq\sup_{\beta\in\mathcal{B},j\leq m}\widehat{\sigma}_j(\beta)<C\right\}$$

occurs w.p.a.1. as $n\to\infty$. Moreover, Lemma 1 implies

$$\sup_{\beta\in\mathcal{B},j\leq m}\left|\left(\mathbb{E}_n\left[h_{ij}(\beta)\right]-\frac{\mathbb{E}\left[g_{ij}(\beta)\right]}{\sigma_j(\beta)}\right)\right|=O_{\mathrm{p}}\left(\sqrt{(\log m)/n}\right).$$

The tuning parameter is $\tau=a_n\sqrt{(\log m)/n}$. As $\lim_{n\to\infty}a_n$ is finite but sufficiently large, the event

$$E_2=\left\{\sup_{\beta\in\mathcal{B},j\leq m}\left|\mathbb{E}_n\left[h_{ij}(\beta)\right]-\frac{\mathbb{E}\left[g_{ij}(\beta)\right]}{\sigma_j(\beta)}\right|<\tau\right\}.$$

occurs w.p.a.1. as $n\to\infty$. From now on we condition on E_1 and E_2.

Because $\mathbb{E}\left[g_{ij}(\beta_0)\right]=0$, under E_2 we have $\max_{j\leq m}\left|\mathbb{E}_n\left[h_{ij}(\beta_0)\right]\right|<\tau$. As a consequence, the estimated probability $\widehat{p}_i(\beta_0)=1/n$ for all $i\leq n$, so that $\ell_n^\tau(\beta_0)=0$ achieves the maximal possible value of the criterion function. Since $\widehat{\beta}$ maximizes $\ell_n^\tau\left(\widehat{\beta}\right)$, we also have $\ell_n^\tau\left(\widehat{\beta}\right)=0$ as well as

$$\widehat{p}_i\left(\widehat{\beta}\right)=1/n \tag{24}$$

Under Assumption 3(a), for any $\beta\in\mathcal{B}\backslash\mathcal{N}(\beta_0,\varepsilon)$ and $n\in\mathbb{N}$ we have at least one moment j that violates $\mathbb{E}\left[g_{ij}(\beta)\right]=0$, so that there exists a fixed $\bar{\delta}=\bar{\delta}(\varepsilon)>0$ such that

$$\inf_{\beta\in\mathcal{B}\backslash\mathcal{N}(\beta_0,\varepsilon)}\max_{j\leq m}\left|\mathbb{E}\left[g_{ij}(\beta)\right]\right|\geq\bar{\delta}$$

for all $n \in \mathbb{N}$. Conditional on E_2,

$$\begin{aligned} & \liminf_{n\to\infty} \inf_{\beta\in\mathcal{B}\setminus\mathcal{N}(\beta_0,\varepsilon)} \max_{j\le m} \left|\mathbb{E}_n\left[h_{ij}\left(\beta\right)\right]\right| \\ = & \liminf_{n\to\infty} \left(\inf_{\beta\in\mathcal{B}\setminus\mathcal{N}(\beta_0,\varepsilon)} \max_{j\le m} \left|\frac{\mathbb{E}\left[g_{ij}\left(\beta\right)\right]}{\sigma_j\left(\beta\right)}\right| - \sup_{\beta\in\mathcal{B},j\le m} \left|\mathbb{E}_n\left[h_{ij}\left(\beta\right)\right] - \frac{\mathbb{E}\left[g_{ij}\left(\beta\right)\right]}{\sigma_j\left(\beta\right)}\right| \right) \\ \ge & \bar{\delta}/C - \lim_{n\to\infty} \tau > 0. \end{aligned}$$

Because E_1 and E_2 both occur w.p.a.1., the above inequality gives

$$\mathbb{P}\left(\sup_{\beta\in\mathcal{B}\setminus\mathcal{N}(\beta_0,\varepsilon)} \left\{\left\{\widehat{p}_i\left(\beta\right) = 1/n\right\}_{i=1}^n\right\} \right) \to 0.$$

That is, had $\widehat{\beta}$ stayed out of the neighborhood $\mathcal{N}\left(\beta_0,\varepsilon\right)$, the event $\left\{\widehat{p}_i\left(\widehat{\beta}\right) = 1/n\right\}_{i=1}^n$ would happen with arbitrarily small probability as $n\to\infty$. Equivalently, this means $\widehat{\beta}\in\mathcal{N}\left(\beta_0,\varepsilon\right)$ w.p.a.1. We have the consistency of $\widehat{\beta}$.

Next, to obtain the rate of convergence we consider a shrinking neighborhood $\mathcal{N}\left(\beta_0, b\tau\right)$ for some constant $b \in \left(2C\sqrt{D/\eta^\dagger}, \infty\right)$. When n is large enough, $\mathcal{N}\left(\beta_0,\varepsilon\right)\setminus\mathcal{N}\left(\beta_0, b\tau\right) \neq \emptyset$. Again we condition on the events E_1 and E_2. As discussed in the consistency part, we have $\max_{j\le m}\left|\mathbb{E}_n\left[h_{ij}\left(\widehat{\beta}\right)\right]\right| < \tau$ under these events, and this of course implies

$$\max_{j\in S^*}\left|\mathbb{E}_n\left[h_{ij}\left(\widehat{\beta}\right)\right]\right| < \tau.$$

Take a Taylor expansion for each of these moment conditions, we have

$$\mathbb{E}\left[g_{iS^*}\left(\widehat{\beta}\right)\right] = \mathbb{E}\left[g_{iS^*}\left(\widehat{\beta}\right)\right] - \mathbb{E}\left[g_{iS^*}\left(\beta_0\right)\right] = \mathbb{E}\left[G_{iS^*}\left(\dot{\beta}\right)\right]\left(\widehat{\beta}-\beta_0\right),$$

where $\dot{\beta}$ is on the line segment joining $\widehat{\beta}$ and β_0 ($\dot{\beta}$ can vary across equations).

Suppose $\widehat{\beta}\in\mathcal{N}\left(\beta_0,\varepsilon\right)\setminus\mathcal{N}\left(\beta_0, b\tau\right)$. Under Assumption 3(b)

$$\begin{aligned} \sum_{j\in S^*}\left(\mathbb{E}\left[g_{ij}\left(\widehat{\beta}\right)\right]\right)^2 &= \mathbb{E}\left[g_{iS^*}\left(\widehat{\beta}\right)\right]'\mathbb{E}\left[g_{iS^*}\left(\widehat{\beta}\right)\right] \\ &= \left(\widehat{\beta}-\beta_0\right)'\mathbb{E}\left[G_{iS^*}\left(\dot{\beta}\right)\right]'\mathbb{E}\left[G_{iS^*}\left(\dot{\beta}\right)\right]\left(\widehat{\beta}-\beta_0\right) \\ &\ge \phi_{\min}\left(\mathbb{E}\left[G_{iS^*}\left(\dot{\beta}\right)\right]'\mathbb{E}\left[G_{iS^*}\left(\dot{\beta}\right)\right]\right)\left(\widehat{\beta}-\beta_0\right)'\left(\widehat{\beta}-\beta_0\right) \\ &\ge \eta^\dagger b^2\tau^2. \end{aligned}$$

and it gives

$$\max_{j\in S^*}\left|\mathbb{E}\left[g_{ij}\left(\widehat{\beta}\right)\right]\right| \geq b\tau\sqrt{\eta^\dagger/D}$$

as $\#\{S^*\} = D$. Under E_1 and E_2, the corresponding sample quantity is

$$\begin{aligned}
& \max_{j\in S^*}\left|\mathbb{E}_n\left[h_{ij}\left(\widehat{\beta}\right)\right]\right| \\
\geq\ & \left(\max_{j\in S^*}\left|\mathbb{E}\left[g_{ij}\left(\widehat{\beta}\right)\right]\right|/\sigma_j\left(\widehat{\beta}\right) - \max_{j\in S^*}\left|\mathbb{E}_n\left[h_{ij}\left(\widehat{\beta}\right)\right] - \mathbb{E}\left[g_{ij}\left(\widehat{\beta}\right)\right]/\sigma_j\left(\widehat{\beta}\right)\right|\right) \\
\geq\ & \left(\max_{j\in S^*}\left|\mathbb{E}\left[g_{ij}\left(\widehat{\beta}\right)\right]\right|/\sigma_j\left(\widehat{\beta}\right) - \tau\right) \\
\geq\ & \frac{b\tau}{C}\sqrt{\frac{\eta^\dagger}{D}} - \tau > \tau.
\end{aligned}$$

where the second inequality holds due to the event E_2, and the last inequality follows by $b > 2C\sqrt{D/\eta^\dagger}$. The above inequality violates $\max_{j\in S^*}\left|\mathbb{E}_n\left[h_{ij}\left(\widehat{\beta}\right)\right]\right| < \tau$ as well as (24). This argument indicates $\widehat{\beta} \in \mathcal{N}(\beta_0, b\tau)$ under events E_1 and E_2, which is equivalent to $\left\|\widehat{\beta} - \beta_0\right\|_2 = O_{\mathrm{p}}\left(\sqrt{(\log m)/n}\right)$.

The rate $O_{\mathrm{p}}\left(\sqrt{(\log m)/n}\right)$ is a lower bound that REL can achieve. If all other moments are irrelevant, meaning $\mathbb{E}[G_{ij}(\beta_0)] = 0$ for all $j \in \{1,\ldots,m\}\setminus S^*$, then this is the sharp bound. If we have other relevant moments, the rate might be sharper.

A.4 Proof of Proposition 2

We suppress the subscript "ifs" of S_{ifs}. According to (9), for a set $S = S_n \subset \{1,\ldots,m\}$ with $\#\{S\} \geq D$, we collect these $\#\{S\}$ equations and stack them as ($\dot{\beta}$ does not have to be the same in different equations.)

$$\mathbb{E}_{\widehat{p}}\left[H_{iS}\left(\dot{\beta}\right)\right]\sqrt{n}\left(\widehat{\beta} - \beta_0\right) - \sqrt{n}\mathbb{E}_{\widehat{p}}\left[h_{iS}\left(\widehat{\beta}\right)\right] = -\sqrt{n}\mathbb{E}_{\widehat{p}}\left[h_{iS}(\beta_0)\right]$$

Pre-multiplying both sides by a $D\times\#\{S\}$ matrix $\mathbb{E}_{\widehat{p}}\left[H_{iS}\left(\widehat{\beta}\right)\right]' W_S^{\widehat{p}}\left(\widehat{\beta}\right)$, we transform the $\#\{S\}$-equation system into a D-equation system

$$\Psi_S^{\widehat{p}}\left(\widehat{\beta},\dot{\beta}\right)\times\sqrt{n}\left(\widehat{\beta} - \beta_0\right) - \sqrt{n}\xi_S^{\widehat{p}}\left(\widehat{\beta}\right) = -\sqrt{n}\xi_S^{\widehat{p}}\left(\widehat{\beta},\beta_0\right).$$

where $\Psi_S^{\widehat{p}}\left(\widehat{\beta},\dot{\beta}\right) := \mathbb{E}_{\widehat{p}}\left[H_{iS}\left(\widehat{\beta}\right)\right]' W_S^{\widehat{p}}\left(\widehat{\beta}\right)\mathbb{E}_{\widehat{p}}\left[H_{iS}\left(\dot{\beta}\right)\right]$. If $\Psi_S^{\widehat{p}}\left(\widehat{\beta},\dot{\beta}\right)$ is invertible, we pre-multiply its inverse to both sides of the above equation,

$$\sqrt{n}\left[\widehat{\beta} - \left(\Psi_S^{\widehat{p}}\left(\widehat{\beta},\dot{\beta}\right)\right)^{-1}\xi_S^{\widehat{p}}\left(\widehat{\beta}\right) - \beta_0\right] = -\left(\Psi_S^{\widehat{p}}\left(\widehat{\beta},\dot{\beta}\right)\right)^{-1}\sqrt{n}\xi_S^{\widehat{p}}\left(\widehat{\beta},\beta_0\right). \tag{25}$$

The fact $\widehat{\beta} \xrightarrow{\mathrm{p}} \beta_0$ squeezes $\dot{\beta} \xrightarrow{\mathrm{p}} \beta_0$ for all $j \leq m$ (again, since $\dot{\beta}$ can be different across equations), so $\Psi_S^{\widehat{p}}\left(\widehat{\beta}, \dot{\beta}\right) - \Psi_S^{\widehat{p}}\left(\widehat{\beta}\right) \xrightarrow{\mathrm{p}} 0_{D\times D}$. Together with the assumption $\Psi_S^{\widehat{p}}\left(\widehat{\beta}\right) - \Psi_S^0\left(\beta_0\right) \xrightarrow{\mathrm{p}} 0_{D\times D}$, we have

$$\Psi_S^{\widehat{p}}\left(\widehat{\beta}, \dot{\beta}\right) - \Psi_S^{0,\sigma}\left(\beta_0\right) \xrightarrow{\mathrm{p}} 0_{D\times D}. \tag{26}$$

As we assume $\Psi_S^0\left(\beta_0\right)$ is invertible in the proposition, $\Psi_S^{\widehat{p}}\left(\widehat{\beta}, \dot{\beta}\right)$ is invertible w.p.a.1. The conclusion follows as

$$\begin{aligned}
& \left(\Psi_S^{\widehat{p}}\left(\widehat{\beta}\right)\right)^{1/2}\sqrt{n}\left(\tilde{\beta}_S - \beta_0\right) \\
= & \left(\Psi_S^{\widehat{p}}\left(\widehat{\beta}\right)\right)^{1/2}\sqrt{n}\left(\widehat{\beta}_S - \left(\Psi_S^{\widehat{p}}\left(\widehat{\beta}\right)\right)^{-1}\xi_S^{\widehat{p}}\left(\widehat{\beta}\right) - \beta_0\right) \\
= & \left(\Psi_S^{\widehat{p}}\left(\widehat{\beta}, \dot{\beta}\right)\right)^{1/2}\sqrt{n}\left(\widehat{\beta}_S - \left(\Psi_S^{\widehat{p}}\left(\widehat{\beta}, \dot{\beta}\right)\right)^{-1}\xi_S^{\widehat{p}}\left(\widehat{\beta}\right) - \beta_0\right) + o_{\mathrm{p}}\left(1\right) \\
= & -\left(\Psi_S^{\widehat{p}}\left(\widehat{\beta}, \dot{\beta}\right)\right)^{-1/2}\sqrt{n}\xi_S^{\widehat{p}}\left(\widehat{\beta}, \beta_0\right) + o_{\mathrm{p}}\left(1\right) \\
= & -\left(\Psi_S^{\widehat{p}}\left(\widehat{\beta}\right)\right)^{-1/2}\sqrt{n}\xi_S^{\widehat{p}}\left(\widehat{\beta}, \beta_0\right) + o_{\mathrm{p}}\left(1\right) \\
\Rightarrow & N\left(0, I_D\right) + o_{\mathrm{p}}\left(1\right).
\end{aligned}$$

where the first equality holds by the definition of $\tilde{\beta}_S$, the third equality by (25), and the last equality by (11). □

A.5 Proof of Lemma 2

Lemma 4. *Under Assumptions 1, 2, 3, 4 and 5, we have*

$$\max_{j\leq m}\left|\mathbb{E}_{\widehat{p}}\left[h_{ij}\left(\widehat{\beta}\right)\right] - \mathbb{E}\left[h_{ij}^{\sigma}\left(\beta_0\right)\right]\right| = O_{\mathrm{p}}\left(\sqrt{\frac{\log m}{n}}\right) \tag{27}$$

$$\max_{j,j'\leq m}\left|\mathbb{E}_{\widehat{p}}\left[h_{ij}\left(\widehat{\beta}\right)h_{ij'}\left(\widehat{\beta}\right)\right] - \mathbb{E}\left[h_{ij}^{\sigma}\left(\beta_0\right)h_{ij'}^{\sigma}\left(\beta_0\right)\right]\right| = O_{\mathrm{p}}\left(\sqrt{\frac{\log m}{n}}\right). \tag{28}$$

Proof. With the choice of τ stated in Theorem 1, w.p.a.1. as $n \to \infty$ we have $\{\widehat{p}_i = 1/n\}_{i=1}^n$ so that $\mathbb{E}_n = \mathbb{E}_{\widehat{p}}$. We conditional on this event from now on.

Since $f_1\left(x\right) = 1/\sqrt{x}$ and $f_2\left(x\right) = 1/x$ are both uniformly continuous in $[1/C, C]$ for

some finite C, Equation (8) in Lemma 1 implies

$$\max_{j\leq m}\left|\frac{1}{\widehat{\sigma}_j(\beta)}-\frac{1}{\sigma_j(\beta)}\right| = O_{\mathrm{p}}\left(\sqrt{(\log m)/n}\right) \tag{29}$$

$$\max_{j\leq m}\left|\frac{1}{\widehat{\sigma}_j^2(\beta)}-\frac{1}{\sigma_j^2(\beta)}\right| = O_{\mathrm{p}}\left(\sqrt{(\log m)/n}\right). \tag{30}$$

By the triangle inequality,

$$\begin{aligned}
&\max_{j\leq m}\left|\mathbb{E}_n\left[h_{ij}\left(\widehat{\beta}\right)\right]-\mathbb{E}\left[h_{ij}^{\sigma}(\beta_0)\right]\right| \\
\leq\ &\max_{j\leq m}\left|\mathbb{E}_n\left[h_{ij}\left(\widehat{\beta}\right)\right]-\mathbb{E}_n\left[h_{ij}^{\sigma}\left(\widehat{\beta}\right)\right]\right|+\max_{j\leq m}\left|\mathbb{E}_n\left[h_{ij}^{\sigma}\left(\widehat{\beta}\right)\right]-\mathbb{E}\left[h_{ij}^{\sigma}(\beta_0)\right]\right| \\
=:\ &T_1+T_2.
\end{aligned}$$

The first term

$$\begin{aligned}
T_1 &= \max_{j\leq m}\left|\mathbb{E}_n\left[h_{ij}\left(\widehat{\beta}\right)\right]-\mathbb{E}_n\left[h_{ij}^{\sigma}\left(\widehat{\beta}\right)\right]\right| \\
&= \max_{j\leq m}\left|\frac{\mathbb{E}_n\left[g_{ij}\left(\widehat{\beta}\right)\right]}{\widehat{\sigma}_j\left(\widehat{\beta}\right)}-\frac{\mathbb{E}_n\left[g_{ij}\left(\widehat{\beta}\right)\right]}{\sigma_j\left(\widehat{\beta}\right)}\right| \\
&\leq \max_{j\leq m}\left|\widehat{\sigma}_j^{-1}\left(\widehat{\beta}\right)-\sigma_j^{-1}\left(\widehat{\beta}\right)\right|\max_{j\leq m}\left|\mathbb{E}_n\left[g_{ij}\left(\widehat{\beta}\right)\right]\right| \\
&= O_{\mathrm{p}}\left(\sqrt{(\log m)/n}\right)O_{\mathrm{p}}(1)=O_{\mathrm{p}}\left(\sqrt{(\log m)/n}\right).
\end{aligned}$$

For the second term, Assumption 2(c) and 5(b) imply uniform continuity of $\max_{j\leq m}\sigma_j(\beta)$ in β, so that

$$\max_{j\leq m}\left|\sigma_j^{-1}\left(\widehat{\beta}\right)-\sigma_j^{-1}(\beta_0)\right|\xrightarrow{\mathrm{p}}0.$$

Also, use (7) and Assumption 5(b),

$$\begin{aligned}
&\max_{j\leq m}\left|\mathbb{E}_n\left[g_{ij}\left(\widehat{\beta}\right)\right]-\mathbb{E}\left[g_{ij}(\beta_0)\right]\right| \\
\leq\ &\max_{j\leq m}\left|\mathbb{E}_n\left[g_{ij}\left(\widehat{\beta}\right)\right]-\mathbb{E}\left[g_{ij}\left(\widehat{\beta}\right)\right]\right|+\max_{j\leq m}\left|\mathbb{E}\left[g_{ij}\left(\widehat{\beta}\right)\right]-\mathbb{E}\left[g_{ij}(\beta_0)\right]\right| \\
=\ &O_{\mathrm{p}}\left(\sqrt{(\log m)/n}\right)+\max_{j\leq m}\left|\mathbb{E}\left[G_{ij}\left(\dot{\beta}\right)\right]\right|\left(\widehat{\beta}-\beta_0\right) \\
=\ &O_{\mathrm{p}}\left(\sqrt{(\log m)/n}\right)+O(1)\,O_{\mathrm{p}}\left(\sqrt{(\log m)/n}\right)=O_{\mathrm{p}}\left(\sqrt{(\log m)/n}\right).
\end{aligned}$$

Therefore

$$\begin{aligned}
T_2 &= \max_{j\leq m}\left|\frac{\mathbb{E}_n\left[g_{ij}\left(\widehat{\beta}\right)\right]}{\sigma_j\left(\widehat{\beta}\right)}-\frac{\mathbb{E}\left[g_{ij}\left(\beta_0\right)\right]}{\sigma_j\left(\beta_0\right)}\right| = \max_{j\leq m}\left|\frac{\mathbb{E}_n\left[g_{ij}\left(\widehat{\beta}\right)\right]}{\sigma_j\left(\widehat{\beta}\right)}-0\right| \\
&\leq \max_{j\leq m}\left|\frac{\mathbb{E}_n\left[g_{ij}\left(\widehat{\beta}\right)\right]}{\sigma_j\left(\widehat{\beta}\right)}-\frac{\mathbb{E}_n\left[g_{ij}\left(\widehat{\beta}\right)\right]}{\sigma_j\left(\beta_0\right)}\right|+\max_{j\leq m}\left|\frac{\mathbb{E}_n\left[g_{ij}\left(\widehat{\beta}\right)\right]}{\sigma_j\left(\beta_0\right)}\right| \\
&= \max_{j\leq m}\left|\sigma_j^{-1}\left(\widehat{\beta}\right)-\sigma^{-1}\left(\beta_0\right)\right|\max_{j\leq m}\left|\mathbb{E}_n\left[g_{ij}\left(\widehat{\beta}\right)\right]\right|+\sigma_j^{-1}\left(\beta_0\right)\max_{j\leq m}\left|\mathbb{E}_n\left[g_{ij}\left(\widehat{\beta}\right)\right]\right| \\
&= O_{\mathrm{p}}\left(\sqrt{\left(\log m\right)/n}\right)O_{\mathrm{p}}\left(\sqrt{\left(\log m\right)/n}\right)+O\left(1\right)O_{\mathrm{p}}\left(\sqrt{\left(\log m\right)/n}\right) \\
&= O_{\mathrm{p}}\left(\sqrt{\left(\log m\right)/n}\right).
\end{aligned}$$

Following essential the same steps with h_{ij} being replaced by $h_{ij}h_{ij'}$ and h_{ij}^{σ} being replaced by $h_{ij}^{\sigma}h_{ij'}^{\sigma}$, we can prove (28). □

Proof of Lemma 2 (27) gives the first conclusion in Lemma 2. (27) and (28) imply the second conclusion in Lemma 2. We are left with the third conclusion. Since D is fixed and finite, without loss of generality we set $D=1$. (If $D\geq 1$, the proof is repeating the steps for each $d=1,\ldots,D$.) In the following derivation, we omit d in the index of the Jacobians H_{ijd} and G_{ijd}; we use H_{ij} and G_{ij} as scalars. Similar to the proof of Lemma 4, we condition on the event $\mathbb{E}_n=\mathbb{E}_{\widehat{p}}$, which occurs w.p.a.1. as $n\to\infty$.

$$\begin{aligned}
&\max_{j\leq m}\left|\mathbb{E}_n\left[H_{ij}\left(\widehat{\beta}\right)\right]-\mathbb{E}\left[H_{ij}^{\sigma}\left(\beta_0\right)\right]\right| \\
\leq\ &\max_{j\leq m}\left|\mathbb{E}_n\left[H_{ij}^{\sigma}\left(\widehat{\beta}\right)\right]-\mathbb{E}\left[H_{ij}^{\sigma}\left(\widehat{\beta}\right)\right]\right|+\max_{j\leq m}\left|\mathbb{E}_n\left[H_{ij}\left(\widehat{\beta}\right)-H_{ij}^{\sigma}\left(\widehat{\beta}\right)\right]\right| \\
&+\max_{j\leq m}\left|\mathbb{E}\left[H_{ij}^{\sigma}\left(\widehat{\beta}\right)\right]-\mathbb{E}\left[H_{ij}^{\sigma}\left(\beta_0\right)\right]\right| \\
=:\ &T_1+T_2+T_3.
\end{aligned}$$

Under Assumption 5(d), the first term

$$\begin{aligned}
T_1 &= \max_{j\leq m}\left|\frac{1}{\sigma_j\left(\widehat{\beta}\right)}\left(\mathbb{E}_n-\mathbb{E}\right)\left[G_{ij}\left(\widehat{\beta}\right)\right]\right| \\
&\leq \max_{j\leq m}\sigma_j^{-1}\left(\widehat{\beta}\right)\max_{j\leq m}\left|\left(\mathbb{E}_n-\mathbb{E}\right)\left[G_{ij}\left(\widehat{\beta}\right)\right]\right| \\
&= O_{\mathrm{p}}\left(1\right)O_{\mathrm{p}}\left(\sqrt{\left(\log m\right)/n}\right)=O_{\mathrm{p}}\left(\sqrt{\left(\log m\right)/n}\right).
\end{aligned}$$

The explicit form of $H_{ij}(\beta)$ is

$$H_{ij}(\beta) = \frac{G_{ij}(\beta)}{\widehat{\sigma}_j(\beta)} - \frac{g_{ij}(\beta)}{2\widehat{\sigma}_j^3(\beta)} \mathrm{cov}_n\left(g_{ij}(\beta), G_{ij}(\beta)\right);$$

therefore

$$\begin{aligned}
T_2 &= \max_{j\le m}\left|\mathbb{E}_n\left[H_{ij}\left(\widehat{\beta}\right) - H_{ij}^{\sigma}\left(\widehat{\beta}\right)\right]\right| \\
&= \max_{j\le m}\left|\mathbb{E}_n\left[\frac{G_{ij}\left(\widehat{\beta}\right)}{\widehat{\sigma}_j\left(\widehat{\beta}\right)} - \frac{g_{ij}\left(\widehat{\beta}\right)}{2\widehat{\sigma}_j^3\left(\widehat{\beta}\right)}\mathrm{cov}_n\left(g_{ij}\left(\widehat{\beta}\right), G_{ij}\left(\widehat{\beta}\right)\right) - \frac{G_{ij}\left(\widehat{\beta}\right)}{\sigma_j\left(\widehat{\beta}\right)}\right]\right| \\
&= \max_{j\le m}\left|\mathbb{E}_n\left[\left(\widehat{\sigma}_j^{-1}\left(\widehat{\beta}\right) - \sigma_j^{-1}\left(\widehat{\beta}\right)\right) G_{ij}\left(\widehat{\beta}\right) - \frac{g_{ij}\left(\widehat{\beta}\right)}{2\sigma_j^3\left(\widehat{\beta}\right)}\frac{\sigma_j^3\left(\widehat{\beta}\right)}{\widehat{\sigma}_j^3\left(\widehat{\beta}\right)}\mathrm{cov}_n\left(g_{ij}\left(\widehat{\beta}\right), G_{ij}\left(\widehat{\beta}\right)\right)\right]\right| \\
&\le \max_{j\le m}\left(\widehat{\sigma}_j^{-1}\left(\widehat{\beta}\right) - \sigma_j^{-1}\left(\widehat{\beta}\right)\right)\max_{j\le m}\left|\mathbb{E}_n\left[G_{ij}\left(\widehat{\beta}\right)\right]\right| \\
&\quad + \max_{j\le m}\left|\mathbb{E}_n\left[\frac{g_{ij}\left(\widehat{\beta}\right)}{2\sigma_j^3\left(\widehat{\beta}\right)}\right]\right|\max_{j\le m}\left(\frac{\sigma_j^3\left(\widehat{\beta}\right)}{\widehat{\sigma}_j^3\left(\widehat{\beta}\right)}\right)\max_{j\le m}\mathrm{cov}_n\left(g_{ij}\left(\widehat{\beta}\right), G_{ij}\left(\widehat{\beta}\right)\right) \\
&=: T_{21} + T_{22}.
\end{aligned}$$

Since

$$\begin{aligned}
\max_{j\le m}\left|\mathbb{E}_n\left[G_{ij}\left(\widehat{\beta}\right)\right]\right| &= \max_{j\le m}\left|\left(\mathbb{E}_n - \mathbb{E}\right)\left[G_{ij}\left(\widehat{\beta}\right)\right]\right| + \max_{j\le m}\left|\mathbb{E}\left[G_{ij}\left(\widehat{\beta}\right)\right]\right| \\
&\le O_{\mathrm{p}}\left(\sqrt{(\log m)/n}\right) + O_{\mathrm{p}}(1) = O_{\mathrm{p}}(1)
\end{aligned}$$

where in the inequality the first term follows by Assumption 5(d), and the last second term is an implication of Assumption 5(b) and the consistency of $\widehat{\beta}$. Combining it with (29), we have $T_{21} = O_{\mathrm{p}}\left(\sqrt{(\log m)/n}\right)$.

For the term T_{22},

$$\max_{j\le m}\left|\mathbb{E}_n\left[\frac{g_{ij}\left(\widehat{\beta}\right)}{2\sigma_j^3\left(\widehat{\beta}\right)}\right]\right| = O_{\mathrm{p}}\left(\sqrt{\frac{\log m}{n}}\right)$$

and by (8) $\max_{j\le m}\left(\frac{\sigma_j^3(\widehat{\beta})}{\widehat{\sigma}_j^3(\widehat{\beta})}\right) = O_{\mathrm{p}}(1)$. Moreover, under Assumption 5(c)(d) we have

$$\begin{aligned}
&\max_{j\le m}\mathrm{cov}_n\left(g_{ij}\left(\widehat{\beta}\right), G_{ij}\left(\widehat{\beta}\right)\right) \\
\le\ &\max_{j\le m}\left(\mathbb{E}_n\left[g_{ij}^2\left(\widehat{\beta}\right)\right]\right)^{1/2}\max_{j\le m}\left(\mathbb{E}_n\left[G_{ij}^2\left(\widehat{\beta}\right)\right]\right)^{1/2} = O_{\mathrm{p}}(1)\, O_{\mathrm{p}}(1).
\end{aligned}$$

We conclude $T_{22} = O_{\mathrm{p}}\left(\sqrt{(\log m)/n}\right)$.

Regarding the third term T_3,

$$\begin{aligned} T_3 &= \max_{j\leq m}\left|\frac{\mathbb{E}\left[G_{ij}\left(\widehat{\beta}\right)\right]}{\sigma_j\left(\widehat{\beta}\right)} - \frac{\mathbb{E}\left[G_{ij}\left(\beta_0\right)\right]}{\sigma_j\left(\beta_0\right)}\right| \\ &\leq \max_{j\leq m}\left|\sigma_j^{-1}\left(\widehat{\beta}\right) - \sigma_j^{-1}\left(\beta_0\right)\right| \max_{j\leq m}\left|\mathbb{E}\left[G_{ij}\left(\widehat{\beta}\right)\right]\right| \\ &\quad + \max_{j\leq m}\left[\sigma_j^{-1}\left(\beta_0\right)\right] \max_{j\leq m}\left|\mathbb{E}\left[G_{ij}\left(\widehat{\beta}\right) - G_{ij}\left(\beta_0\right)\right]\right| \\ &= O_{\mathrm{p}}\left(\sqrt{\frac{\log m}{n}}\right) O_{\mathrm{p}}\left(1\right) + O\left(1\right) O_{\mathrm{p}}\left(\sqrt{\frac{\log m}{n}}\right) = O_{\mathrm{p}}\left(\sqrt{\frac{\log m}{n}}\right) \end{aligned}$$

where the second term of the second equality holds under Assumption 2(a). We complete the proof. □

A.6 Proof of Theorem 2(a)

To simplify notation, we denote $\widehat{J}_S := \mathbb{E}_{\widehat{p}}\left[H_{iS}\left(\widehat{\beta}\right)\right]$ ($\#\{S\}\times D$ matrix), $J_S^0 := \mathbb{E}\left[H_{iS}^{\sigma}\left(\beta_0\right)\right]$, $\widehat{W}_S := W_S^{\widehat{p}}\left(\widehat{\beta}\right)$ ($\#\{S\} \times \#\{S\}$ matrix), $W_S^0 := W_S^{0,\sigma}\left(\beta_0\right)$, $\widehat{V}_S := \mathbb{E}_{\widehat{p}}\left[V_{iSS}^{\widehat{p}}\left(\widehat{\beta}\right)\right]$ and $V_S^0 := \mathbb{E}\left[V_{iSS}^{0,\sigma}\left(\beta_0\right)\right]$. For the above symbols, the set is taken as $\{1,\ldots,m\}$ if the subscript S is dropped. The approximation error to the sandwich information matrix can be decomposed as

$$\begin{aligned} & \Psi_S^{\widehat{p}}\left(\widehat{\beta}\right) - \Psi_S^{0,\sigma}\left(\beta_0\right) \\ = \;& \widehat{J}_S'\left(W_S^0 + \varepsilon_{W,S}\right)\widehat{J}_S - \Psi_S^{0,\sigma}\left(\beta_0\right) \\ = \;& \widehat{J}_S' W_S^0 \widehat{J}_S + \widehat{J}_S'\varepsilon_{W,S}\widehat{J}_S - \Psi_S^{0,\sigma}\left(\beta_0\right) \\ = \;& \left(J_S^0 + \nu_{J,S}\right)' W_S^0\left(J_S^0 + \nu_{J,S}\right) + \widehat{J}_S'\varepsilon_{W,S}\widehat{J}_S - \Psi_S^{0,\sigma}\left(\beta_0\right) \\ = \;& J_S^{0\prime} W_S^0 \nu_{J,S} + \nu_{J,S}' W_S^0 J_S^0 + \nu_{J,S}' W_S^0 \nu_{J,S} + \widehat{J}_S'\varepsilon_{W,S}\widehat{J}_S \\ =: \;& T_1 + T_2 + T_3 + T_4 \end{aligned}$$

where $\nu_{J,S} := \widehat{J}_S - J_S^0$ and $\varepsilon_{W,S} := \widehat{W}_S - W_S^0$.

A representative element of $\nu_{J,S}$ is $\nu_{J,jd} = \widehat{J}_{jd} - J_{jd}^0$ for some $j \in S$ and $d \in \{1,\ldots,D\}$.

Let $\nu_{J,Sd}$ and J^0_{Sd} be the d-th column of $\nu_{J,S}$ and J^0_S, respectively. For any $d, d' \leq D$,

$$\begin{aligned} |T_{1dd'}| = |T_{2d'd}| &= \left|\nu'_{J,Sd} W^0_S J^0_{Sd'}\right| \\ &\leq \phi_{\max}\left(W^0_S\right) O\left(\bar{m}\right) \max_{j \in S} |\nu_{J,jd}| \left|J^0_{jd'}\right| \\ &\leq O\left(\frac{\bar{m}}{\eta^*}\right) \|\nu_J\|_\infty C \\ &= O_{\mathrm{p}}\left(\frac{\bar{m}}{\eta^*}\sqrt{\frac{\log m}{n}}\right). \end{aligned} \tag{31}$$

where the first inequality follows by (57), the second inequality by Assumption 4 and 5(b), and the last equality by Lemma 2. By the Cauchy-Schwarz inequality,

$$\begin{aligned} |T_{3dd'}| &= \left|\nu'_{J,Sd} W^0_S \nu_{J,Sd'}\right| \\ &\leq \left(\left|\nu'_{J,Sd} W^0_S \nu_{J,Sd}\right| \left|\nu'_{J,Sd'} W^0_S \nu_{J,Sd'}\right|\right)^{1/2} \\ &\leq \max_{d \leq D} \left|\nu'_{J,Sd} W^0_S \nu_{J,Sd}\right| \\ &\leq \phi_{\max}\left(W^0_S\right) \max_{d \leq D} \left(\nu'_{J,Sd} \nu_{J,Sd}\right) \\ &\leq O\left(\frac{\bar{m}}{\eta^*}\right) \|\nu_J\|^2_\infty = O_{\mathrm{p}}\left(\frac{\bar{m}}{\eta^*}\frac{\log m}{n}\right). \end{aligned} \tag{32}$$

The last term T_4 is complicated due to the presence of the inverse of a random matrix that becomes infinite-dimensional as $n \to \infty$. Since Assumption 5(b) and Lemma 2 give

$$\begin{aligned} \sum_{j \in S} \widehat{J}'_{jd} \widehat{J}_{jd} &= \sum_{j \in S} \left(J^0_{jd} + \nu_{J,jd}\right)' \left(J^0_{jd} + \nu_{J,jd}\right) \\ &\leq 2\left(\sum_{j \in S} J^{0\prime}_{jd} J^0_{jd} + \sum_{j \in S} \nu'_{J,jd} \nu_{J,jd}\right) \\ &= C + O\left(\bar{m}\right) O_{\mathrm{p}}\left(\left(\log m\right)/n\right) = O_{\mathrm{p}}\left(1\right), \end{aligned}$$

using the Cauchy-Schwarz inequality and (57) we have

$$\begin{aligned} |T_{4dd'}| &= \left|\widehat{J}'_{Sd} \varepsilon_{W,S} \widehat{J}_{Sd'}\right| \\ &\leq \max_{d \leq D} \widehat{J}'_{Sd} \varepsilon_{W,S} \widehat{J}_{Sd} \\ &\leq \phi_{\max}\left(\varepsilon_{W,S}\right) \max_{d \leq D} \sum_{j \in S} \widehat{J}'_{jd} \widehat{J}_{jd} \\ &= \phi_{\max}\left(\varepsilon_{W,S}\right) O_{\mathrm{p}}\left(1\right). \end{aligned} \tag{33}$$

We need some extra work to handle $\varepsilon_{W,S}$. Let $\phi_{\mathrm{spec}}\left(\cdot\right)$ be the spectral norm as defined in

Section B.2. The spectral norm of a real, symmetric matrix equals its maximal eigenvalue. We first verify

$$\begin{aligned}\phi_{\text{spec}}\left(W_S^0\nu_{V,S}\right) &\leq \phi_{\text{spec}}\left(\left(V_S^0\right)^{-1}\right)\phi_{\text{spec}}\left(\nu_{V,S}\right) = \phi_{\max}\left(\left(V_S^0\right)^{-1}\right)\phi_{\max}\left(\nu_{V,S}\right) \\ &\leq \frac{1}{\eta^*}\phi_{\max}\left(\nu_{V,S}\right) \leq O\left(\frac{\bar{m}}{\eta^*}\right)\left\|\nu_{V,S}\right\|_\infty \\ &= O_{\text{p}}\left(\frac{\bar{m}}{\eta^*}\sqrt{\frac{\log m}{n}}\right) \end{aligned} \tag{34}$$

where the first inequality follows by the submultiplicativity of the spectral norm as in (55), the second inequality by Assumption 4, the third inequality by (56) and the last equality by Lemma 2. Since

$$\varepsilon_{W,S} = \left(\widehat{V}_S\right)^{-1} - \left(V_S^0\right)^{-1} = \left(V_S^0 + \nu_{V,S}\right)^{-1} - \left(V_S^0\right)^{-1}$$

where $\nu_{V,S} := \widehat{V}_S - V_S^0$, given (34) we can invoke (58) to get

$$\begin{aligned}\varepsilon_{W,S} &= \left(I_{\#\{S\}} + \sum_{k=1}^{\infty}\left(W_S^0\nu_{V,S}\right)^k\right)\left(V_S^0\right)^{-1} - \left(V_S^0\right)^{-1} \\ &= \left[\sum_{k=1}^{\infty}\left(W_S^0\nu_{V,S}\right)^k\right]W_S^0. \end{aligned} \tag{35}$$

With the series representation (35), we bound the maximal eigenvalue of $\varepsilon_{W,S}$ by

$$\begin{aligned}\phi_{\max}\left(\varepsilon_{W,S}\right) &= \phi_{\text{spec}}\left(\varepsilon_{W,S}\right) = \phi_{\text{spec}}\left(\left[\sum_{k=1}^{\infty}\left(W_S^0\nu_{V,S}\right)^k\right]W_S^0\right) \\ &\leq \phi_{\text{spec}}\left(W_S^0\right)\phi_{\text{spec}}\left(\sum_{k=1}^{\infty}\left(W_S^0\nu_{V,S}\right)^k\right) \leq \phi_{\max}\left(W_S^0\right)\sum_{k=1}^{\infty}\phi_{\text{spec}}\left(\left(W_S^0\nu_{V,S}\right)^k\right) \\ &= \phi_{\max}\left(W_S^0\right)\frac{\phi_{\text{spec}}\left(W_S^0\nu_{V,S}\right)}{1-\phi_{\text{spec}}\left(W_S^0\nu_{V,S}\right)} \leq \frac{1}{\eta^*}\frac{O_{\text{p}}\left(\frac{\bar{m}}{\eta^*}\sqrt{\frac{\log m}{n}}\right)}{1-O_{\text{p}}\left(\frac{\bar{m}}{\eta^*}\sqrt{\frac{\log m}{n}}\right)} \\ &= O_{\text{p}}\left(\frac{\bar{m}}{\eta^{*2}}\sqrt{\frac{\log m}{n}}\right). \end{aligned} \tag{36}$$

where the first inequality follows by the submultiplicativity in (55), the second inequality by the triangle inequality (54), and the last inequality by (34). Combining (33) and (36) we have $|T_{4dd'}| = O_{\text{p}}\left(\frac{\bar{m}}{\eta^{*2}}\sqrt{\frac{\log m}{n}}\right)$.

Collect the rates of the terms $T_{1dd'}, T_{2dd'}$, $T_{3dd'}$ and $T_{4dd'}$. The probability of the events for all $S \in \mathcal{S}_{(2\vee D)\bar{m}}$ is controlled uniformly by Lemma 2. Since D is finite, we complete the proof.

A.7 Proof of Theorem 2(b)

Step 1: Let λ be any constant vector such that $\lambda \in \mathbb{R}^D$ and $\lambda'\lambda = 1$. We keep using the simplified notation defined at the beginning of Section A.6. Further simplify $\widehat{\Psi}_S := \Psi_S^{\widehat{\rho}}\left(\widehat{\beta}\right)$, $\Psi_S^0 := \Psi_S^{0,\sigma}(\beta_0)$, and $\xi_{iS}^0 := J_S^{0\prime} W_S^0 h_{iS}^\sigma(\beta_0)$. Define

$$Y_{iS}(\lambda) := n^{-1/2}\lambda'\left(\Psi_S^0\right)^{-1/2}\xi_{iS}^0.$$

Step 1: We verify the conditions of the Liapunov central limit theorem to establish

$$\sum_{i=1}^n Y_{iS}(\lambda) \Rightarrow N(0,1). \tag{37}$$

Since $\mathbb{E}\left[h_{ij}^\sigma(\beta_0)\right] = \mathbb{E}\left[g_{ij}(\beta_0)\right]/\sigma_j(\beta_0) = 0$ for all $j \leq m$, we have the mean

$$\mathbb{E}\left[Y_{iS}(\lambda)\right] = n^{-1/2}\lambda'\left(\Psi_S^0\right)^{-1/2}\mathbb{E}\left[\xi_{iS}^0\right] = 0,$$

the variance

$$\begin{aligned}
\operatorname{var}\left(Y_{iS}(\lambda)\right) &= \frac{1}{n}\lambda'\left(\Psi_S^0\right)^{-1/2} J_S^{0\prime} W_S^0 \mathbb{E}\left[h_{iS}^\sigma(\beta_0)\, h_{iS}^\sigma(\beta_0)'\right] W_S^0 J_S^0 \left(\Psi_S^0\right)^{-1/2}\lambda \\
&= \frac{1}{n}\lambda'\left(\Psi_S^0\right)^{-1/2} J_S^{0\prime} W_S^0 J_S^0 \left(\Psi_S^0\right)^{-1/2}\lambda \\
&= \frac{1}{n}\lambda' I_D \lambda = \frac{1}{n}.
\end{aligned}$$

Under Assumption 2(b), the Liapunov condition

$$\sum_{i=1}^n \mathbb{E}\left[\left|Y_{iS}(\lambda)\right|^6\right] \leq \left\|\lambda' n^{-1/2}\left(\Psi_S^0\right)^{-1/2} J_S^{0\prime} W_S^0\right\|_1^6 \max_{j\in S}\sum_{i=1}^n \mathbb{E}\left[\left|h_{iS}^\sigma(\beta_0)\right|^6\right].$$

By the Cauchy-Schwarz inequality, the first factor

$$\begin{aligned}
& \left\|\lambda' n^{-1/2}\left(\Psi_S^0\right)^{-1/2} J_S^{0\prime} W_S^0\right\|_1^6 \\
\leq\ & \left(\sqrt{\widetilde{m}}\left\|n^{-1/2}\lambda'\left(\Psi_S^0\right)^{-1/2} J_S^{0\prime} W_S^0\right\|_2\right)^6 \\
\leq\ & \left[\sqrt{\widetilde{m}}\phi_{\max}^{1/2}\left(W_S^0\right)\right]^6 \left\|n^{-1/2}\lambda'\left(\Psi_S^0\right)^{-1/2} J_S^{0\prime}\left(W_S^0\right)^{1/2}\right\|_2^6 \\
=\ & \left[\sqrt{\widetilde{m}}\phi_{\max}^{1/2}\left(W_S^0\right)\right]^6 \left[\operatorname{var}\left(Y_{iS}\left(\lambda\right)\right)\right]^3 \leq \left(\frac{\widetilde{m}}{\eta^*}\right)^3 \times \frac{1}{n^3}.
\end{aligned}$$

Assumption 2(b) implies the second factor $\max_{j\leq m}\sum_{i=1}^n \mathbb{E}\left[\left\|h_{ij}^{\sigma}\left(\beta_0\right)\right\|_\infty^6\right] \leq nC$. As

$$\widetilde{m}/\eta^* = \bar{m} = O\left(\left(n/\log m\right)^{1/4}\right),$$

we thus have

$$\sum_{i=1}^n \mathbb{E}\left[\left|Y_{iS}\left(\lambda\right)\right|^6\right] = O\left(\frac{\left(n/\log m\right)^{3/4}}{n^2}\right) = o\left(1\right).$$

We apply the Liapunov central limit theorem to conclude (37).

Step 2: Since $\mathbb{E}_n = \mathbb{E}_{\widehat{p}}$ w.p.a.1. as $n\to\infty$, we have

$$\begin{aligned}
\sum_{i=1}^n n\widehat{p}_i Y_{iS}\left(\lambda\right) &= \mathbb{E}_{\widehat{p}}\left[nY_{iS}\left(\lambda\right)\right] = \mathbb{E}_n\left[nY_{iS}\left(\lambda\right)\right] + o_{\mathrm{p}}\left(1\right) \\
&= \sum_{i=1}^n Y_{iS}\left(\lambda\right) + o_{\mathrm{p}}\left(1\right) \Rightarrow N\left(0,1\right).
\end{aligned}$$

Step 3: Let $\mathbf{1}_S$ be a $\#\left\{S\right\}\times 1$ vector of ones. We will verify

$$\begin{aligned}
o_{\mathrm{p}}\left(1\right)\mathbf{1}_D &= \sqrt{n}\xi_S^{\widehat{p}}\left(\widehat{\beta},\beta_0\right) - J_S^{0\prime} W_S^0 \sqrt{n}\mathbb{E}_{\widehat{p}}\left[h_{iS}^{\sigma}\left(\beta_0\right)\right] \\
&= \left(\widehat{J}_S'\widehat{W}_S - J_S^{0\prime} W_S^0\right)\sqrt{n}\mathbb{E}_{\widehat{p}}\left[h_{iS}^{\sigma}\left(\beta_0\right)\right] \\
&= \left(\widehat{J}_S'\varepsilon_{W,S} + \nu_{J,S}' W_S^0 + \nu_{J,S}'\varepsilon_{W,S}\right)\sqrt{n}\mathbb{E}_{\widehat{p}}\left[h_{iS}^{\sigma}\left(\beta_0\right)\right] \\
&=: T_1 + T_2 + T_3. \qquad (38)
\end{aligned}$$

T_1, T_2 and T_3 are $D \times 1$ vectors. For any $d \leq D$,

$$\begin{aligned}
T_{1d} &= \widehat{J}'_{Sd}\varepsilon_{W,S}\sqrt{n}\mathbb{E}_{\widehat{p}}\left[h^{\sigma}_{iS}\left(\beta_0\right)\right] \\
&\leq \sqrt{n}\left(\widehat{J}'_{Sd}\varepsilon_{W,S}\mathbb{E}_{\widehat{p}}\left[h^{\sigma}_{iS}\left(\beta_0\right)\right]\mathbb{E}_{\widehat{p}}\left[h^{\sigma}_{iS}\left(\beta_0\right)\right]'\varepsilon_{W,S}\widehat{J}_{Sd}\right)^{1/2} \\
&\leq \sqrt{n}\phi^{1/2}_{\max}\left(\varepsilon_{W,S}\mathbb{E}_{\widehat{p}}\left[h^{\sigma}_{iS}\left(\beta_0\right)\right]\mathbb{E}_{\widehat{p}}\left[h^{\sigma}_{iS}\left(\beta_0\right)\right]'\varepsilon_{W,S}\right)\left(\widehat{J}'_{Sd}\widehat{J}_{Sd}\right)^{1/2} \\
&= \sqrt{n}\phi^{1/2}_{\max}\left(\varepsilon_{W,S}\mathbb{E}_{\widehat{p}}\left[h^{\sigma}_{iS}\left(\beta_0\right)\right]\mathbb{E}_{\widehat{p}}\left[h^{\sigma}_{iS}\left(\beta_0\right)\right]'\varepsilon_{W,S}\right)O_{\mathrm{p}}\left(1\right) \\
&\leq \sqrt{n}\phi_{\max}\left(\varepsilon_{W,S}\right)\phi^{1/2}_{\max}\left(\mathbb{E}_{\widehat{p}}\left[h^{\sigma}_{iS}\left(\beta_0\right)\right]\mathbb{E}_{\widehat{p}}\left[h^{\sigma}_{iS}\left(\beta_0\right)\right]'\right)O_{\mathrm{p}}\left(1\right) \\
&\leq \sqrt{n}O_{\mathrm{p}}\left(\frac{\widetilde{m}}{\eta^{*2}}\sqrt{\frac{\log m}{n}}\right)\phi^{1/2}_{\max}\left(\mathbb{E}_{\widehat{p}}\left[h^{\sigma}_{iS}\left(\beta_0\right)\right]\mathbb{E}_{\widehat{p}}\left[h^{\sigma}_{iS}\left(\beta_0\right)\right]'\right)O_{\mathrm{p}}\left(1\right) \\
&= \sqrt{n}O_{\mathrm{p}}\left(\frac{\widetilde{m}}{\eta^{*2}}\sqrt{\frac{\log m}{n}}\right)O_{\mathrm{p}}\left(\widetilde{m}^{1/2}\sqrt{\frac{\log m}{n}}\right)O_{\mathrm{p}}\left(1\right) \\
&= O_{\mathrm{p}}\left(\frac{\widetilde{m}^{3/2}}{\eta^{*2}}\sqrt{\frac{\log^2 m}{n}}\right),
\end{aligned}$$

where the second equality follows as in the derivation of (33), the fourth inequality by (36). Assumption 1(a) gives $\log m = o\left(n^{1/9}\right)$, so that

$$\frac{\widetilde{m}^{3/2}}{\eta^{*2}}\sqrt{\frac{\log^2 m}{n}} = O\left(\left(\frac{n}{\log m}\right)^{7/16}\right)\frac{\log m}{n^{1/2}} = O\left(\frac{(\log m)^{9/16}}{n^{1/16}}\right) = o\left(1\right),$$

and then $T_{1d} = o_{\mathrm{p}}\left(1\right)$.

For the second term, (57) gives

$$\begin{aligned}
T_{2d} &= \nu'_{J,Sd}W^0_S\left(\beta_0\right)\sqrt{n}\mathbb{E}_{\widehat{p}}\left[h^{\sigma}_{iS}\left(\beta_0\right)\right] \\
&\leq \phi_{\max}\left(W^0_S\left(\beta_0\right)\right)O\left(\widetilde{m}\right)\left\|\nu_J\right\|_{\infty}\left\|\sqrt{n}\mathbb{E}_{\widehat{p}}\left[h^{\sigma}_{iS}\left(\beta_0\right)\right]\right\|_{\infty} \\
&= O\left(\frac{\widetilde{m}}{\eta^*}\right)O_{\mathrm{p}}\left(\sqrt{\frac{\log m}{n}}\right)\sqrt{n}\left[\left\|\mathbb{E}_n\left[h^{\sigma}_{iS}\left(\beta_0\right)\right]\right\|_{\infty} + O_{\mathrm{p}}\left(\sqrt{\frac{\log m}{n}}\right)\right] \\
&= O\left(\frac{\widetilde{m}}{\eta^*}\right)O_{\mathrm{p}}\left(\sqrt{\frac{\log m}{n}}\right)O_{\mathrm{p}}\left(\sqrt{\log m}\right) = o_{\mathrm{p}}\left(1\right)
\end{aligned}$$

Moreover, T_{3d} is of smaller order than T_{1d} and T_{2d}. Since D is finite, we have verified (38).

Step 4: Because

$$\widehat{\Psi}_S = \Psi^0_S + O_{\mathrm{p}}\left(\frac{\widetilde{m}}{\eta^{*2}}\sqrt{\frac{\log m}{n}}\right)\mathbf{1}_D\mathbf{1}'_D = \Psi^0_S\left(1 + o_{\mathrm{p}}\left(1\right)\mathbf{1}_D\mathbf{1}'_D\right); \tag{39}$$

we have

$$\begin{aligned}
& \lambda'\left(\widehat{\Psi}_S\right)^{-1/2}\sqrt{n}\xi_S^{\widehat{p}}\left(\widehat{\beta},\beta_0\right) \\
= \; & \lambda'\left(\widehat{\Psi}_S\right)^{-1/2}\left(J_S^{0\prime}W_S^0\sqrt{n}\mathbb{E}_{\widehat{p}}\left[h_{iS}^{\sigma}\left(\beta_0\right)\right]+o_{\mathrm{p}}\left(1\right)\mathbf{1}_D\right) \\
= \; & \lambda'\left(\Psi_S^0\right)^{-1/2}J_S^{0\prime}W_S^0\sqrt{n}\mathbb{E}_{\widehat{p}}\left[h_{iS}^{\sigma}\left(\beta_0\right)\right]+o_{\mathrm{p}}\left(1\right) \\
= \; & \sum_{i=1}^{n}n\widehat{p}_iY_{iS}\left(\lambda\right)+o_{\mathrm{p}}\left(1\right)\Rightarrow N\left(0,1\right)
\end{aligned}$$

where the first equality follows by Step 3, the second equality by (39), and the third equality by Step 2. The conclusion follows by the using of the Cramér-Wold device. □

A.8 Proof of Lemma 3

Let S and T be two generic sets such that $S,T\in\mathcal{S}_{\widehat{m}}$, $S\supset T$ and $\#\left\{S\backslash T\right\}\geq 1$. Let $\left\{h_{i\left(S^{\perp}|T\right)}\left(\beta\right)\right\}_{i=1}^{n}$ be the residuals of $\left\{h_{iS}\left(\beta\right)\right\}_{i=1}^{n}$ after being projected onto the linear space spanned by $\left\{h_{iT}\left(\beta\right)\right\}_{i=1}^{n}$, and define

$$\begin{aligned}
H_{i(S|T)}\left(\beta\right) &= \frac{\partial}{\partial\beta}h_{i\left(S^{\perp}|T\right)}\left(\beta\right) \\
W_{S|T}^{\widehat{p}}\left(\beta\right) &= \left\{\mathbb{E}_{\widehat{p}}\left[\left(h_{i\left(S^{\perp}|T\right)}\left(\beta\right)-\mathbb{E}_{\widehat{p}}\left[h_{i\left(S^{\perp}|T\right)}\left(\beta\right)\right]\right)'\left(h_{i\left(S^{\perp}|T\right)}\left(\beta\right)-\mathbb{E}_{\widehat{p}}\left[h_{i\left(S^{\perp}|T\right)}\left(\beta\right)\right]\right)\right]\right\}^{-}.
\end{aligned}$$

We can show (See Breusch et al. (1999, Lemma 1))

$$\Delta\Psi_{S|T}^{\widehat{p}}\left(\beta\right) = \mathbb{E}_{\widehat{p}}\left[H_{i(S|T)}\left(\beta\right)\right]'W_{S|T}^{\widehat{p}}\left(\beta\right)\mathbb{E}_{\widehat{p}}\left[H_{i(S|T)}\left(\beta\right)\right]. \tag{40}$$

By construction, $\Delta\Psi_{S|T}^{\widehat{p}}\left(\beta\right)$ is positive-semidefinite and a diagonal element $\Delta\Psi_{S|T}^{\widehat{p},d}\left(\beta\right)$ is non-negative.

From now on we work in the population. We suppress the dependence of $h_S^{\sigma}\left(Z,\beta_0\right)$ on Z and β_0 and focus on the sets. Recall

$$\Psi_S^{0,\sigma}:=\mathbb{E}\left[\frac{\partial h_S^{\sigma}}{\partial\beta}\right]'W_S^{0,\sigma}\mathbb{E}\left[\frac{\partial h_S^{\sigma}}{\partial\beta}\right]$$

is the sandwich form matrix that measures the quantity of information in h_S^{σ}. Similar to the sample version (40), the increment in the population is

$$\Delta\Psi_{S|T}^{0,\sigma}=\mathbb{E}\left[H_{(S|T)}^{\sigma}\right]'W_{S|T}^{0,\sigma}\mathbb{E}\left[H_{(S|T)}^{\sigma}\right]$$

where $\Delta\Psi_{S|T}^{0,\sigma}$, $H_{(S|T)}^{\sigma}$ and $W_{S|T}^{0,\sigma}$ are the population counterparts of $\Delta\Psi_{S|T}^{\widehat{p}}$, $H_{i(S|T)}$ and $W_{S|T}^{\widehat{p}}$,

respectively. Denote $U_{ST} = \mathbb{E}\left[V_{iST}^{0,\sigma}\right]$. The covariance matrix can be partitioned as

$$U = \begin{bmatrix} U_{(S\setminus T)(S\setminus T)} & U_{(S\setminus T)T} \\ U_{T(S\setminus T)} & U_{TT} \end{bmatrix},$$

so the upper-left block of $W_S^{0,\sigma}$ is $W_{S|T}^{0,\sigma} = \left(U_{(S\setminus T)(S\setminus T)} - U_{(S\setminus T)T}U_{TT}^{-1}U_{T(S\setminus T)}\right)^{-1}$.

Since we use the d-th diagonal term as the information criterion for the d-th component of β, without loss of generality we assume $D = 1$ so we can suppress the superscript d. Thus $\Delta\Psi_{S|T}^{0,\sigma}$ is a scalar and $\mathbb{E}\left[H_{(S|T)}^{\sigma}\right]$ is a vector. Denote $S^{\#}$ as the best m^*-member set; that is, $\Psi^{m^*,\sigma} = \Psi_{S^{\#}}^{0,\sigma}$. In view of Lemma 10, if we replace a, W_{11}, and W in the second inequality of (59) by $\mathbb{E}\left[H_{(S^{\#}|T)}^{\sigma}\right]$, $W_{S^{\#}|T}^{0,\sigma}$ and $W_{S^{\#}\cup T}^{0,\sigma}$, we have

$$\begin{aligned} \Delta\Psi_{S^{\#}|T}^{0,\sigma} &\leq \left(\phi_{\min}\left(U_{S^{\#}\cup T}\right)\right)^{-1} \sum_{j\in S^{\#}\setminus T} \left(\mathbb{E}\left[H_{(\{T,j\}|T)}^{\sigma}\right]\right)^2 \\ &\leq m^* \left(\phi_{\max}\left(W_{S^{\#}\cup T}^{0,\sigma}\right)\right) \max_{j\in S^{\#}\setminus T} \left(\mathbb{E}\left[H_{(\{T,j\}|T)}^{\sigma}\right]^2\right) \\ &\leq m^* \left(\phi_{\max}\left(W_{S^{\#}\cup T}^{0,\sigma}\right)\right) \max_{j\leq m} \left(\mathbb{E}\left[H_{(\{T,j\}|T)}^{\sigma}\right]^2\right). \end{aligned} \tag{41}$$

Similarly, replace a, W_{11}, and W in the first inequality of (59) by $\mathbb{E}\left[H_{(\{T,j\}|T)}^{\sigma}\right]$, $W_{\{T,j\}|T}^{0,\sigma}$ and $W_{\{T,j\}}^{0,\sigma}$, and multiply both sides by $\phi_{\max}\left(U_{\{T,j\}}\right)$,

$$\mathbb{E}\left[H_{(\{T,j\}|T)}^{\sigma}\right]^2 \leq \phi_{\max}\left(U_{\{T,j\}}\right)\Delta\Psi_{\{T,j\}|T}^{0,\sigma} \leq \sup_{S\in\mathcal{S}_{\widehat{m}}} \left(\phi_{\min}\left(W_S^{0,\sigma}\right)\right)^{-1} \Delta\Psi_{\{T,j\}|T}^{0,\sigma}.$$

Substitute the above inequality into (41), and note $\phi_{\max}\left(W_{S^{\#}\cup T}^{0,\sigma}\right) = \phi_{\min}\left(U_{S^{\#}\cup T}\right) \geq \eta^*$ as $S^{\#}\cup T \in \mathcal{S}_{2\bar{m}}$,

$$\Delta\Psi_{S^{\#}|T}^{0,\sigma} \leq m^* \frac{\phi_{\max}\left(W_{S^{\#}\cup T}^{0,\sigma}\right)}{\sup_{S\in\mathcal{S}_{\widehat{m}}} \phi_{\min}\left(W_S^{0,\sigma}\right)} \max_{j\leq m} \Delta\Psi_{\{T,j\}|T}^{0,\sigma} \leq m^* \frac{\varphi_{\widehat{m}}^*}{\eta^*} \max_{j\leq m} \Delta\Psi_{\{T,j\}|T}^{0,\sigma}$$

Replacing T by S as in the statement of the lemma, we complete the proof. □

A.9 Proof of Theorem 3

In this proof we work with a fixed $d \in \{1, \ldots, D\}$. Without loss of generality, we assume $D = 1$ so that we drop d in all related symbols for notational conciseness. We follow the basic strategy of the proof of prediction consistency of the componentwise boosting in a linear

regression under fixed design (Bühlmann and van de Geer, 2011, Section 12.8.2), whereas our building blocks are very different from theirs.

We first work in the population. For any index set S, define $\mathcal{R}\Psi_S^{0,\sigma}(\beta_0) := \Psi^m - \Psi_S^{0,\sigma}(\beta_0)$.

Definition 1. For a fixed constant $\alpha \in (0,1)$, we call $\mathbb{S}_n(\alpha) = (S_r)_{r=1}^{\widehat{m}}$ an *α-weakly greedy sequence* if it satisfies $S_r \subset S_{r+1}$, $\#\{S_{r+1} \backslash S_r\} = 1$ for every r, and

$$\mathcal{R}\Psi_{S_{r+1}}^{0,\sigma}(\beta_0) \leq \mathcal{R}\Psi_{S_r}^{0,\sigma}(\beta_0) - (1-\alpha)\max_{j\leq m}\Delta\Psi_{\{S_r,j\}|S_r}^{0,\sigma}(\beta_0). \tag{42}$$

Each sequence $\mathbb{S}_n(\alpha)$ contains $\widehat{m}$ ordered elements $S_1 \subset S_2 \subset \cdots \subset S_{\widehat{m}}$. Conditional on a set S_r, in the $(r+1)$-th iteration on any of such α-weakly greedy sequences, the information gap $\mathcal{R}\Psi_{S_r}^{0,\sigma}(\beta_0)$ is narrowed by $(1-\alpha)\max_{j\leq m}\Delta\Psi_{\{S_r,j\}|S_r}^{0,\sigma}(\beta_0)$. The following lemma shows that under the sparsity condition the weakly greedy mechanism eventually exhausts all information in the limit.

Lemma 5. *Under the assumptions in Theorem 3, for any sequence in $\mathbb{S}_n(\alpha)$ that satisfies the definition, we have $\mathcal{R}\Psi_{S_{\widehat{m}}}^{0}(\beta_0) \to 0$ as $n \to \infty$.*

Proof. By the definition of $\Delta(m, m^*)$ and Lemma 3, for any $r \in \{0, \ldots, \widehat{m}\}$ and any set S_r we have

$$\mathcal{R}\Psi_{S_r}^{0,\sigma}(\beta_0) - \Delta(m,m^*) = \Psi^{m^*}(\beta_0) - \Psi_{S_r}^{0,\sigma}(\beta_0) \leq \frac{\eta^*}{m^*\varphi_{\widehat{m}}^*}\max_{j\leq m}\Delta\Psi_{\{S_r,j\}|S_r}^{0,\sigma}(\beta_0). \tag{43}$$

Use (43) to substitute out $\max_{j\leq m}\Delta\Psi_{\{S_r,j\}|S_r}^{0,\sigma}(\beta_0)$ on the right-hand side of (42),

$$\begin{aligned}\mathcal{R}\Psi_{S_{r+1}}^{0,\sigma}(\beta_0) &\leq \mathcal{R}\Psi_{S_r}^{0,\sigma}(\beta_0) - (1-\alpha)\frac{\eta^*}{m^*\varphi_{\widehat{m}}^*}\left(\mathcal{R}\Psi_{S_r}^{0,\sigma}(\beta_0) - \Delta(m,m^*)\right) \\ &= b_n\mathcal{R}\Psi_{S_r}^{0,\sigma}(\beta_0) + \tilde{\Delta}(m,m^*)\end{aligned} \tag{44}$$

where

$$\begin{aligned} b_n &:= 1 - (1-\alpha)\eta^*/(m^*\varphi_{\widehat{m}}^*) \\ \tilde{\Delta}(m,m^*) &:= (1-\alpha)\Delta(m,m^*)\eta^*/(m^*\varphi_{\widehat{m}}^*) = (1-b_n)\Delta(m,m^*). \end{aligned}$$

Starting from $r = \widehat{m} - 1$, we iterate the recursive inequality (44) backward,

$$\begin{aligned}
\mathcal{R}\Psi^{0,\sigma}_{S_{\widehat{m}}}(\beta_0) &\leq b_n \mathcal{R}\Psi^{0,\sigma}_{S_{\widehat{m}-1}}(\beta_0) + \tilde{\Delta}(m^*) \\
&\leq b_n \left[b_n \mathcal{R}\Psi^{0,\sigma}_{S_{\widehat{m}-2}}(\beta_0) + \tilde{\Delta}(m^*) \right] + \tilde{\Delta}(m^*) \\
&= b_n^2 \mathcal{R}\Psi^{0,\sigma}_{S_{\widehat{m}-2}}(\beta_0) + \tilde{\Delta}(m, m^*)[1 + b_n] \\
&\leq \cdots \\
&\leq b_n^{\widehat{m}} \mathcal{R}\Psi^{0,\sigma}_{\emptyset}(\beta_0) + \tilde{\Delta}(m, m^*) \sum_{r=0}^{\widehat{m}-1} b_n^r \\
&\leq b_n^{\widehat{m}} \mathcal{R}\Psi^{0,\sigma}_{\emptyset}(\beta_0) + \tilde{\Delta}(m, m^*) \sum_{r=0}^{\infty} b_n^r \\
&= b_n^{\widehat{m}} \Psi^m + \tilde{\Delta}(m, m^*)(1 - b_n)^{-1} \\
&= b_n^{\widehat{m}} \Psi^m + \Delta(m, m^*).
\end{aligned}$$

We check the limits of the terms on the right-hand side. Under Assumption 5(a), by the Gershgorin circle theorem, the maximal eigenvalue $\varphi^*_{\widehat{m}}$ is bounded by

$$\varphi^*_{\widehat{m}} \leq \max_{S \in \mathcal{S}_{\widehat{m}}} \max_{j \in S} \sum_{j' \in S} \left| v^0_{jj'} \right| \leq C.$$

Under Assumption 6(b) $m^* = o\left(\eta^{*2} (n/\log m)^{1/4}\right)$, we have

$$\frac{\varphi^*_{\widehat{m}}}{\eta^*} \frac{m^*}{\widehat{m}} \leq \frac{C}{\eta^*} o\left(\eta^{*2} \left(\frac{n}{\log m} \right)^{1/4} \right) O\left(\frac{1}{\eta^*} \left(\frac{n}{\log m} \right)^{-1/4} \right) = o(1),$$

so that

$$\lim_{n \to \infty} b_n^{\widehat{m}} = \lim_{n \to \infty} \left(1 - (1 - \alpha) \frac{\eta^*}{m^* \varphi^*_{\widehat{m}}} \right)^{\widehat{m}} = 0.$$

Since $\Delta(m, m^*) \to 0$ under Assumption 6, we conclude that $\mathcal{R}\Psi^{0,\sigma}_{\beta_0}(S_{\widehat{m}}) \to 0$. □

Up to now we have been calculating in the population. We link the α-weakly greedy sequence in population to the sequence generated by the greedy algorithm in sample. Let

$$\widehat{j} = \widehat{j}(S) := \arg\max_{j \leq m} \Delta\Psi^{\widehat{p}}_{\{j,S\}|S}\left(\widehat{\beta}\right)$$

be the selected moment in the sample in the j-th iteration conditional on a set S. Let

$$\delta_{\widehat{m}} := \sup_{S\in\mathcal{S}_{\widehat{m}},j\leq m} 2\left|\Psi^{\widehat{p}}_{\{S,j\}|S}\left(\widehat{\beta}\right)-\Psi^{0,\sigma}_{\{S,j\}|S}\left(\beta_0\right)\right|.$$

According to Theorem 2,

$$\begin{aligned}\delta_{\widehat{m}} &\leq \max_{S\in\mathcal{S}_{\widehat{m}},j\leq m} 2\left(\left|\Psi^{\widehat{p}}_{\{S,j\}}\left(\widehat{\beta}\right)-\Psi^{0,\sigma}_{\{S,j\}}\left(\beta_0\right)\right|+\left|\Psi^{\widehat{p}}_{S}\left(\widehat{\beta}\right)-\Psi^{0,\sigma}_{S}\left(\beta_0\right)\right|\right)\\ &= O_{\mathrm{p}}\left(\frac{\widehat{m}}{\eta^{*2}}\sqrt{\left(\log m\right)/n}\right). \qquad (45)\end{aligned}$$

Lemma 6. *For a fixed $0<\alpha<1$ and some $S\in S_{\widehat{m}}$, if the event*

$$\left\{\max_{j\leq m}\Delta\Psi^{0,\sigma}_{\{j,S\}|S}\left(\beta_0\right)>\alpha^{-1}\delta_{\widehat{m}}\right\}$$

holds, then

$$\Delta\Psi^{0,\sigma}_{\{\widehat{j},S\}|S}\left(\beta_0\right)\geq(1-\alpha)\max_{j\leq m}\Delta\Psi^{0,\sigma}_{\{j,S\}|S}\left(\beta_0\right). \qquad (46)$$

Proof. By the triangle inequality, on any set $S\in\mathcal{S}_{\widehat{m}}$ we have

$$\begin{aligned}\Delta\Psi^{0,\sigma}_{\{\widehat{j},S\}|S}\left(\beta_0\right) &\geq \Delta\Psi^{\widehat{p}}_{\{\widehat{j},S\}|S}\left(\widehat{\beta}\right)-\left|\Delta\Psi^{\widehat{p}}_{\{\widehat{j},S\}|S}\left(\widehat{\beta}\right)-\Delta\Psi^{0,\sigma}_{\{\widehat{j},S\}|S}\left(\beta_0\right)\right|\\ &\geq \Delta\Psi^{\widehat{p}}_{\{\widehat{j},S\}|S}\left(\widehat{\beta}\right)-\frac{1}{2}\delta_{\widehat{m}}\\ &= \max_{j\leq m}\Delta\Psi^{\widehat{p}}_{\{j,S\}|S}\left(\widehat{\beta}\right)-\frac{1}{2}\delta_{\widehat{m}}\\ &\geq \max_{j\leq m}\left(\Delta\Psi^{0,\sigma}_{\{j,S\}|S}\left(\beta_0\right)-\left|\Delta\Psi^{0,\sigma}_{\{j,S\}|S}\left(\beta_0\right)-\Delta\Psi^{\widehat{p}}_{\{j,S\}|S}\left(\widehat{\beta}\right)\right|\right)-\frac{1}{2}\delta_{\widehat{m}}\\ &\geq \max_{j\leq m}\Delta\Psi^{0,\sigma}_{\{j,S\}|S}\left(\beta_0\right)-\delta_{\widehat{m}}\\ &\geq (1-\alpha)\max_{j\leq m}\Delta\Psi^{0,\sigma}_{\{j,S\}|S}\left(\beta_0\right).\end{aligned}$$

The equality follows by the definition of $\widehat{j}$, and the last line is the consequence of the event $\left\{\max_{j\leq m}\Delta\Psi^{0,\sigma}_{\{j,S\}|S}\left(\beta_0\right)>\alpha^{-1}\delta_{\widehat{m}}\right\}$. □

Denote $\widehat{S}_r$ as the selected set in the first r iterations. For any sequence $\widehat{\mathbb{S}}=\left(\widehat{S}_r\right)_{r=1}^{\widehat{m}}$ generated from the Algorithm in sample, either one of the two cases occurs.

- Case 1: $\max_{j\leq m}\Delta\Psi^{0,\sigma}_{\{\widehat{S}_{\widehat{m}},j\}|\widehat{S}_{\widehat{m}}}\left(\beta_0\right)>\alpha^{-1}\delta_{\widehat{m}}$;

- Case 2: $\max_{j\leq m}\Delta\Psi^{0,\sigma}_{\left\{\widehat{S}_{\widehat{m}},j\right\}|\widehat{S}_{\widehat{m}}}\left(\beta_0\right)\leq\alpha^{-1}\delta_{\widehat{m}}$.

In Case 1, by (46) we have

$$\mathcal{R}\Psi^{0,\sigma}_{\widehat{S}_{r+1}}\left(\beta_0\right)=\mathcal{R}\Psi^{0,\sigma}_{\widehat{S}_r}\left(\beta_0\right)-\Delta\Psi^{0,\sigma}_{\left\{\widehat{S}_r,\widehat{j}\right\}|\widehat{S}_r}\left(\beta_0\right)\leq\mathcal{R}\Psi^{0,\sigma}_{\widehat{S}_r}\left(\beta_0\right)-\left(1-\alpha\right)\max_{j\leq m}\Delta\Psi^{0,\sigma}_{\left\{\widehat{S}_r,j\right\}|\widehat{S}_r}\left(\beta_0\right)$$

for every $r\leq\widehat{m}$. The sequence $\widehat{\mathbb{S}}$ satisfies the definition of the α-weakly greedy sequence. By Lemma 5, we have $\mathcal{R}\Psi^{0}_{\widehat{S}_{\widehat{m}}}\left(\beta_0\right)\to 0$ w.p.a.1.

In Case 2, by Assumption 6(b) the information is concentrated in a set of cardinality m^*. Even if none of the components in the "best set" is selected, because the cardinality is m^* and the maximal information in a single unselected moment is smaller than $\alpha^{-1}\delta_{\widehat{m}}$, we have

$$\begin{aligned}
\mathcal{R}\Psi^{0,\sigma}_{\widehat{S}_{\widehat{m}}}\left(\beta_0\right) &= \Psi^{m}-\Psi^{0,\sigma}_{\widehat{S}_{\widehat{m}}}\left(\beta_0\right)\\
&= \Delta\left(m,m^*\right)+\left(\Psi^{m^*}-\Psi^{0}_{\widehat{S}_{\widehat{m}}}\left(\beta_0\right)\right)\\
&\leq \Delta\left(m,m^*\right)+\sup_{S\in S_{m^*}}\left(\Psi^{0}_{\widehat{S}_{\widehat{m}}\cup S}\left(\beta_0\right)-\Psi^{0}_{\widehat{S}_{\widehat{m}}}\left(\beta_0\right)\right)\\
&\leq \Delta\left(m,m^*\right)+m^*\cdot\alpha^{-1}\delta_{\widehat{m}}\\
&= o\left(1\right)+\alpha^{-1}o\left(\eta^{*2}\left(\frac{n}{\log m}\right)^{1/4}\right)O_{\mathrm{p}}\left(\frac{\widehat{m}}{\eta^{*2}}\sqrt{\frac{\log m}{n}}\right)\\
&= o\left(1\right)+\alpha^{-1}o\left(\eta^{*2}\left(\frac{n}{\log m}\right)^{1/4}\right)O_{\mathrm{p}}\left(\frac{\eta^*\left(n/\log m\right)^{1/4}}{\eta^{*2}}\sqrt{\frac{\log m}{n}}\right)\\
&= o\left(1\right)+o_{\mathrm{p}}\left(\eta^*\right)=o_{\mathrm{p}}\left(1\right).
\end{aligned}$$

We complete the proof.

A.10 Proof of Theorem 4

Since $\Psi^{m}\left(\beta_0\right)$ is the full information, the information gap (in matrix form) $\Psi^{m}\left(\beta_0\right)-\Psi^{0,\sigma}_{S}\left(\beta_0\right)$ is positive-semidefinite for all $S\subset\{1,\ldots,m\}$. As we impose Assumption 6 for all $d\leq D$, Theorem 3 implies that the diagonal elements of $\Psi^{m}\left(\beta_0\right)-\Psi^{0,\sigma}_{\widehat{S}_{\widehat{m}}}\left(\beta_0\right)$ shrink to zero w.p.a.1, so that

$$\begin{aligned}
\left\|\Psi^{m}\left(\beta_0\right)-\Psi^{0,\sigma}_{\widehat{S}_{\widehat{m}}}\left(\beta_0\right)\right\|_{\infty} &\leq \mathbf{1}_D'\left[\Psi^{m}\left(\beta_0\right)-\Psi^{0,\sigma}_{\widehat{S}_{\widehat{m}}}\left(\beta_0\right)\right]\mathbf{1}_D\\
&\leq \phi_{\max}\left(\Psi^{m}\left(\beta_0\right)-\Psi^{0,\sigma}_{\widehat{S}_{\widehat{m}}}\left(\beta_0\right)\right)\mathbf{1}_D'\mathbf{1}_D\\
&\leq D\times\operatorname{tr}\left(\Psi^{m}\left(\beta_0\right)-\Psi^{0,\sigma}_{\widehat{S}_{\widehat{m}}}\left(\beta_0\right)\right)\xrightarrow{\mathrm{p}}0.
\end{aligned}$$

By Assumption 3(b), for all $n \in \mathbb{N}$, there is a set S^* such that

$$\begin{aligned}
\phi_{\min}\left(\Psi^{0,\sigma}_{S^*}(\beta_0)\right) &= \phi_{\min}\left(\mathbb{E}\left[H^{0,\sigma}_{iS^*}(\beta_0)\right]' W^{0,\sigma}_{S^*}(\beta_0)\,\mathbb{E}\left[H^{0,\sigma}_{iS^*}(\beta_0)\right]\right) \\
&\geq \phi_{\min}\left(W^{0,\sigma}_{S^*}(\beta_0)\right)\phi_{\min}\left(\mathbb{E}\left[H^{0,\sigma}_{iS^*}(\beta_0)\right]'\mathbb{E}\left[H^{0,\sigma}_{iS^*}(\beta_0)\right]\right) \\
&\geq \frac{\phi_{\min}\left(W^{0,\sigma}_{S^*}(\beta_0)\right)}{\max_{j\leq m}\sigma_j^2(\beta_0)}\phi_{\min}\left(\mathbb{E}\left[G^{0,\sigma}_{iS^*}(\beta_0)\right]'\mathbb{E}\left[G^{0,\sigma}_{iS^*}(\beta_0)\right]\right) \\
&\geq \frac{\eta^\dagger}{\phi_{\max}\left(\mathbb{E}\left[V^{0,\sigma}_i(\beta_0)\right]\right)\max_{j\leq m}\sigma_j^2(\beta_0)} \geq \eta^\dagger/C \qquad (47)
\end{aligned}$$

where the last inequality follows by Assumption 2(c) and 5(a). $\Psi^{0,\sigma}_{S^*}(\beta_0)$ is positive definite, so as $\Psi^m(\beta_0)$. Now that the invertibility of $\Psi^m(\beta_0)$ is satisfied, and $\left\|\Psi^m(\beta_0) - \Psi^{0,\sigma}_{\widehat{S}_{\widehat{m}}}(\beta_0)\right\|_\infty \xrightarrow{\text{p}} 0$, the asymptotic normality follows by Proposition 2.

A.11 Proof of Corollary 1

The proof of this corollary is similar and simpler than that of Theorem 3 as we use the finite D^* to replace m^*. D^* is finite while the m^* diverges to infinity. For any $r \in \{0,\ldots,\widehat{m}\}$ and any set S_r by (12) we have

$$\max_{j\leq m}\Delta\Psi^{0,\sigma}_{\{S_r,j\}|S_r}(\beta_0) \geq \frac{\eta^*}{D^*C}\left[\Psi^{D^*}(\beta_0) - \Psi^{0,\sigma}_{S_r}(\beta_0)\right]. \qquad (48)$$

On those α-weakly greedy sequences, parallel to the recursive equality (44) if $\Psi^{D^*} - \Psi^{0,\sigma}_{\widetilde{S}_{r+1}}(\beta_0) \geq 0$, we have

$$\begin{aligned}
\Psi^{D^*} - \Psi^{0,\sigma}_{\widetilde{S}_{r+1}}(\beta_0) &\leq \left(\Psi^{D^*} - \Psi^{0,\sigma}_{\widetilde{S}_r}(\beta_0)\right) - (1-\alpha)\frac{\eta^*}{D^*\varphi^*_{\widehat{m}}}\left(\Psi^{D^*} - \Psi^{0,\sigma}_{\widetilde{S}_r}(\beta_0)\right) \\
&= b_n\left(\Psi^{D^*} - \Psi^{0,\sigma}_{\widetilde{S}_r}(\beta_0)\right)
\end{aligned}$$

where $b_n := 1 - (1-\alpha)\eta^*/(D^*C)$. Starting from $r = \widehat{m} - 1$, we iterate the recursive inequality backward,

$$\Psi^{D^*} - \Psi^0_{\widetilde{S}_{\widehat{m}}}(\beta_0) \leq b_n\left(\Psi^{D^*} - \Psi^{0,\sigma}_{\widetilde{S}_{\widehat{m}}}(\beta_0)\right) \leq \cdots \leq b_n^{\widehat{m}}\Psi^{D^*}(\beta_0).$$

Since $\frac{\eta^*\widehat{m}}{D^*C} \to \infty$, we conclude that $\Psi^{D^*} - \Psi^0_{\beta_0}\left(\widetilde{S}_{\widehat{m}}\right) \to 0$.

According to the specified rate of $\widehat{m}$,

$$\sup_{S\in\mathcal{S}_{D\widehat{m}}}\left\|\Psi_S^{\widehat{p}}\left(\widehat{\beta}\right)-\Psi_S^{0,\sigma}\left(\beta_0\right)\right\|_\infty=O_{\mathrm{p}}\left(\frac{\widehat{m}}{\eta^{*2}}\sqrt{\frac{\log m}{n}}\right)=o_{\mathrm{p}}\left(1\right).$$

Following the same steps after Lemma 6 as in the proof of Theorem 3, we get the first conclusion.

For the second conclusion, because

$$\Psi^{D^*,d}\geq\Psi_{S^*}^{0,\sigma,d}$$

for all $d\leq D$, the first conclusion $0\vee\left(\Psi^{D^*,d}-\Psi_{\widehat{S}_{\widehat{m}}}^{0,\sigma,d}\right)\xrightarrow{\mathrm{p}}0$ implies that $\widehat{S}_{\widehat{m}}\in\mathcal{S}_{D\widehat{m}}^*\left(\alpha\right)$ w.p.a.1. Under the additional Assumption (13), we have $\phi_{\min}\left(\Psi_{\widehat{S}_{\widehat{m}}}^{0,\sigma}\right)>0$ being bounded away from zero w.p.a.1. The second conclusion follows as a direct implication of Proposition 2 and Theorem 2.

A.12 Sufficient Condition for Assumption 5(d)

Assumption 5(b) implies

$$\sup_{n\in\mathbb{N}}\sup_{\beta\in\mathcal{N}_\varepsilon(\beta_0)}\left\|\mathbb{E}\left[G_i^2\left(\beta\right)\right]\right\|_\infty\leq C.$$

It is suffice to strengthen Assumption 2(a).

Assumption 7. *For all $n\in\mathbb{N}$, we have $\mathbb{E}\left[g_{ij}\left(\beta\right)\right]$ is continuously differentiable in β for all $\beta\in\mathcal{B}$ and $j\leq m$, and there exists measurable functions $\tilde{B}_{nj}:\mathcal{Z}\mapsto\mathbb{R}^+$ such that*

$$\sup_{\beta_1,\beta_2\in\mathcal{B},d\leq D}\frac{\left|g_j\left(Z,\beta_1\right)-g_j\left(Z,\beta_2\right)\right|}{\left\|\beta_1-\beta_2\right\|_2}\leq\tilde{B}_{nj}\left(Z\right)$$

and $\max_{j\leq m}\mathbb{E}\left[\tilde{B}_{nj}^3\left(Z\right)\right]<C$.

This assumption implies $\max_{j\leq m}\mathbb{E}\left[\sup_{\beta\in\mathcal{N}_\varepsilon(\beta_0),d\leq D}\left|G_{ij}^3\left(\beta\right)\right|\right]\leq C$. For a fixed $\beta\in$

$\mathcal{N}_\varepsilon(\beta_0)$ and $d \in \{1, \ldots, D\}$, we invoke Lemma 7 to get

$$\begin{aligned}
& \max_{j \leq m} \left| \left(\mathbb{E}_n - \mathbb{E} \right) \left[G_{ijd}(\beta) \right] \right| \\
= & \max_{j \leq m} \left| \frac{\left(\mathbb{E}_n - \mathbb{E} \right) \left[G_{ijd}(\beta) \right]}{\sqrt{\operatorname{var}_n \left(G_{ijd}(\beta) \right)}} \sqrt{\operatorname{var}_n \left(G_{ijd}(\beta) \right)} \right| \\
\leq & \max_{j \leq m} \left| \frac{\left(\mathbb{E}_n - \mathbb{E} \right) \left[G_{ijd}(\beta) \right]}{\sqrt{\operatorname{var}_n \left(G_{ijd}(\beta) \right)}} \right| \left(\max_{j \leq m} \left[\mathbb{E}_n \left(G_{ijd}^2(\beta) \right) \right] \right)^{1/2} \\
= & \max_{j \leq m} \left| \frac{\left(\mathbb{E}_n - \mathbb{E} \right) \left[G_{ijd}(\beta) \right]}{\sqrt{\operatorname{var}_n \left(G_{ijd}(\beta) \right)}} \right| O_{\mathrm{p}}(1) \\
= & O_{\mathrm{p}} \left(\sqrt{(\log m)/n} \right).
\end{aligned}$$

This rate carries over to $\beta \in \mathcal{B}$ and $d \leq D$ uniformly by the continuity of $G_{ijd}(\beta)$ under the assumption.

B Generic Lemmas

B.1 Moderate Deviation Theory of Self-Normalized Sums

Jing et al. (2003) establish the moderate deviation theory of self-normalized sums. Belloni et al. (2012) innovatively use it to derive the maximal inequality for high-dimensional random vectors. In this section, we will use the moderate deviation theory in our setting. We follow the proof strategy of Belloni et al. (2012, p.2413, Step 2 and 3).

Let $\{X_1, \ldots, X_n\}$ be an i.i.d. sample. X_1 is an m-dimensional random vector with $\mathbb{E}[X_1] = 0_m$. The following lemma establishes the rate of convergence of the maximum of the self-normalized sums.

Lemma 7. *For all $n \in \mathbb{N}$ suppose there exists a universal constant C such that $1/C \leq \mathbb{E}\left[X_{1j}^2\right] \leq C$ for all $j \leq m$ and $\max_{j \leq m} \mathbb{E}\left[|X_{1j}|^3\right] < C$. If $\log m = o\left(n^{1/3}\right)$, then*

$$\max_{j \leq m} \left| \mathbb{E}_n[X_{ij}] / \sqrt{\mathbb{E}_n\left[X_{ij}^2\right]} \right| = O_{\mathrm{p}} \left(\sqrt{(\log m)/n} \right).$$

Proof. By the Bonferroni bound,

$$\mathbb{P}\left(\max_{j\leq m}\frac{\mathbb{E}_n\left[X_{ij}\right]}{\sqrt{\mathbb{E}_n\left[X_{ij}^2\right]}}\geq\sqrt{\frac{2}{n}\log\left(\frac{4m}{\varepsilon}\right)}\right)\leq m\max_{j\leq m}\mathbb{P}\left(\frac{\mathbb{E}_n\left[X_{ij}\right]}{\sqrt{\mathbb{E}_n\left[X_{ij}^2\right]}}\geq\sqrt{\frac{2}{n}\log\left(\frac{4m}{\varepsilon}\right)}\right) \tag{49}$$

Denote

$$d_n:=\min_{j\leq m}\left[\left(n\mathbb{E}\left[X_{1j}^2\right]\right)^{1/2}/\left(n\mathbb{E}\left[\left|X_{1j}\right|^3\right]\right)^{1/3}\right].$$

By the assumption, $\mathbb{E}\left[X_{1j}^2\right]=O\left(1\right)$ and $\mathbb{E}\left[\left|X_{1j}\right|^3\right]=O\left(1\right)$ for all $j\leq m$, so that $d_n\asymp n^{1/6}$. As $\log m=o\left(n^{1/3}\right)$, obviously $\sqrt{2\log\left(4m/\varepsilon\right)}=o\left(d_n\right)$. This allows us to apply Theorem 7.4 of Peña et al. (2009). For every $j\leq m$,

$$\begin{aligned}&\mathbb{P}\left(\sqrt{n}\mathbb{E}_n\left[X_{ij}\right]/\sqrt{\mathbb{E}_n\left[X_{ij}^2\right]}\geq\sqrt{2\log\left(4m/\varepsilon\right)}\right)\\ \leq\ &\left(1-\Phi\left(\sqrt{2\log\left(4m/\varepsilon\right)}\right)\right)\left(1+o\left(1\right)\right).\end{aligned} \tag{50}$$

By the fact $1-\Phi\left(t\right)\leq 2\exp\left(-t^2/2\right)/\left(1+t\right)$ for $t\geq 0$,

$$1-\Phi\left(\sqrt{2\log\left(\frac{4m}{\varepsilon}\right)}\right)\leq\frac{\varepsilon}{2m}\left(1+\sqrt{2\log\left(\frac{4m}{\varepsilon}\right)}\right)^{-1} \tag{51}$$

Combine (49), (50) and (51), we have

$$\begin{aligned}&\mathbb{P}\left(\max_{j\leq m}\sqrt{n}\frac{\mathbb{E}_n\left[X_{ij}\right]}{\sqrt{\mathbb{E}_n\left[X_{ij}^2\right]}}\geq\sqrt{2\log\left(\frac{4m}{\varepsilon}\right)}\right)\\ \leq\ &\frac{\varepsilon}{2}\left(1+\sqrt{2\log\left(\frac{4m}{\varepsilon}\right)}\right)^{-1}\left(1+o\left(1\right)\right)=o\left(\varepsilon\right)\end{aligned}$$

The same argument leads to

$$\mathbb{P}\left(\min_{j\leq m}\sqrt{n}\frac{\mathbb{E}_n\left[X_{ij}\right]}{\sqrt{\mathbb{E}_n\left[X_{ij}^2\right]}}\leq-\sqrt{2\log\left(\frac{4m}{\varepsilon}\right)}\right)\leq o\left(\varepsilon\right).$$

We complete the proof. □

The next lemma establishes the convergence of the maximum discrepancy of the entries of the $m \times m$ covariance matrix.

Lemma 8. *For all $n \in \mathbb{N}$ suppose there exists a universal constant C such that $1/C \leq var\left[X_{1j}^2\right] \leq C$ for all m and $\max_{j \leq m} \mathbb{E}\left[X_{1j}^6\right] < C$. If $\kappa_n^{(4)} := \max_{j \leq m} \mathbb{E}_n\left[X_{ij}^4\right] = O_{\mathrm{p}}(1)$ and $\log m = o\left(n^{1/3}\right)$, then*

$$\max_{j \leq j' \leq m} \left|\mathbb{E}_n\left[X_{ij}X_{ij'}\right] - \mathbb{E}\left[X_{ij}X_{ij'}\right]\right| = O_{\mathrm{p}}\left(\sqrt{(\log m)/n}\right).$$

Proof. The steps in this proof is similar to those of Lemma 7. Let $v_{i,jj'} = X_{ij}X_{ij'} - \mathbb{E}\left[X_{ij}X_{ij'}\right]$. By the Cauchy-Schwarz inequality,

$$\left(\mathbb{E}_n\left[X_{ij}^2X_{ij'}^2\right]\right)^{1/2} \leq \left(\mathbb{E}_n\left[X_{ij}^4\right]\mathbb{E}_n\left[X_{ij'}^4\right]\right)^{1/4} \leq \sqrt{\kappa_n^{(4)}} = O_{\mathrm{p}}(1),$$

we have

$$\begin{aligned}
& \mathbb{P}\left(\max_{j \leq j' \leq m} \mathbb{E}_n\left[v_{i,jj'}\right] \geq \sqrt{\frac{4}{n}\log\left(\frac{2m}{\varepsilon}\right)}\sqrt{\kappa_n^{(4)}}\right) \\
\leq & \mathbb{P}\left(\max_{j \leq j' \leq m} \frac{\mathbb{E}_n\left[v_{i,jj'}\right]}{\sqrt{\mathbb{E}_n\left[X_{ij}^2X_{ij'}^2\right]}} \geq \sqrt{\frac{4}{n}\log\left(\frac{2m}{\varepsilon}\right)}\right) \\
\leq & \mathbb{P}\left(\max_{j \leq j' \leq m} \frac{\mathbb{E}_n\left[v_{i,jj'}\right]}{\sqrt{\mathbb{E}_n\left[v_{i,jj'}^2\right]}} \geq \sqrt{\frac{4}{n}\log\left(\frac{2m}{\varepsilon}\right)}\right) \\
\leq & \frac{m(m+1)}{2}\max_{j \leq j' \leq m} \mathbb{P}\left(\sqrt{n}\frac{\mathbb{E}_n\left[v_{i,jj'}\right]}{\sqrt{\mathbb{E}_n\left[v_{ij}^2\right]}} \geq \sqrt{4\log\left(\frac{2m}{\varepsilon}\right)}\right), \qquad (52)
\end{aligned}$$

where the last inequality follows by the Bonferroni bound. Denote

$$d_n := \min_{j \leq m}\left[\left(n\mathbb{E}\left[v_{1jj'}^2\right]\right)^{1/2} / \left(n\mathbb{E}\left[\left|v_{1jj'}\right|^3\right]\right)^{1/3}\right].$$

By the assumption $\mathbb{E}\left[v_{1jj}^2\right] = O(1)$ and $\mathbb{E}\left[\left|v_{1jj'}\right|^3\right] = O(1)$ for all $j \leq j' \leq m$, we have $d_n \asymp n^{1/6}$. As $\log m = o\left(n^{1/3}\right)$, obviously $\sqrt{2\log\left(\frac{4m^2}{\varepsilon^2}\right)} = \sqrt{4\log\left(\frac{2m}{\varepsilon}\right)} = o(d_n)$. This

allows us to apply Theorem 7.4 of Peña et al. (2009). Similar to as (50) and (51), for every $j \leq j' \leq m$ we have

$$\begin{aligned}
& \mathbb{P}\left(\sqrt{n}\mathbb{E}_n\left[v_{i,jj'}\right]/\sqrt{\mathbb{E}_n\left[v_{i,ij}^2\right]} \geq \sqrt{2\log\left(4m^2/\varepsilon^2\right)}\right) \\
\leq & \left(1-\Phi\left(\sqrt{2\log\left(4m^2/\varepsilon^2\right)}\right)\right)(1+o(1)) \\
= & \left(1-\Phi\left(\sqrt{4\log\left(\frac{2m}{\varepsilon}\right)}\right)\right)(1+o(1)) \\
\leq & \frac{\varepsilon^2}{2m^2}\left(1+\sqrt{4\log\left(\frac{2m}{\varepsilon}\right)}\right)^{-1}(1+o(1)). \qquad (53)
\end{aligned}$$

(52) and (53) give

$$\begin{aligned}
& \mathbb{P}\left(\max_{1\leq j\leq j'\leq m}\sqrt{n}\mathbb{E}_n\left[v_{i,jj'}\right] \geq \sqrt{4\log\left(2m/\varepsilon\right)}\sqrt{\kappa_n^{(4)}}\right) \\
\leq & \frac{m(m+1)}{4m^2}\varepsilon^2\left(1+\sqrt{4\log\left(\frac{2m}{\varepsilon}\right)}\right)^{-1}(1+o(1)) = o\left(\varepsilon^2\right),
\end{aligned}$$

and similarly

$$\begin{aligned}
& \mathbb{P}\left(\max_{1\leq j\leq j'\leq m}\sqrt{n}\mathbb{E}_n\left[v_{i,jj'}\right] \geq -\sqrt{4\log\left(2m/\varepsilon\right)}\sqrt{\kappa_n^{(4)}}\right) \\
\leq & \frac{m(m+1)}{4m^2}\varepsilon^2\left(1+\sqrt{4\log\left(\frac{2m}{\varepsilon}\right)}\right)^{-1}(1+o(1)) = o\left(\varepsilon^2\right).
\end{aligned}$$

Since $\sqrt{4\log\left(\frac{2m}{\varepsilon}\right)/n} = O\left(\sqrt{(\log m)/n}\right)$ and $\kappa_n^{(4)} = O_{\mathrm{p}}(1)$ by assumption, we complete the proof. □

B.2 Useful Lemmas on Matrices

The *spectral norm* of a matrix A is defined as $\phi_{\mathrm{spec}}(A) := \sqrt{\phi_{\max}(A'A)}$, where the maximal eigenvalue for a $k \times k$ square matrix M is $\phi_{\max}(M) := \sup_{\{c\in\mathbb{R}^k : c'c=1\}} |c'Mc|$. The spectral norm is a matrix norm that satisfies

$$\phi_{\mathrm{spec}}(A+B) \leq \phi_{\mathrm{spec}}(A) + \phi_{\mathrm{spec}}(B) \quad \text{(triangle inequality)} \qquad (54)$$

$$\phi_{\mathrm{spec}}(AB) \leq \phi_{\mathrm{spec}}(A)\,\phi_{\mathrm{spec}}(B) \quad \text{(submultiplicativity)}. \qquad (55)$$

See Horn and Johnson (1985, pp.290–296) for the properties of matrix norms. If M is symmetric, then by definition $\phi_{\text{spec}}(M) = \phi_{\max}(M)$.

We further prove the following results.

Lemma 9. *If A is a $k \times k$ matrix, then*

$$\phi_{\max}(A) \leq k \|A\|_\infty \tag{56}$$

If M is a $k \times k$ symmetric matrix, then for any $k \times 1$ vectors x and y,

$$|xMy| \leq k \|x\|_\infty \|y\|_\infty \phi_{\max}(M). \tag{57}$$

If A and $A - B$ are invertible, and $\phi_{\text{spec}}\left(A^{-1}B\right) < 1$, then

$$(A-B)^{-1} = \left(I_k + \sum_{l=1}^{\infty} \left(A^{-1}B\right)^l\right) A^{-1}. \tag{58}$$

Proof. (56) is a direct implication of the Gershgorin circle theorem. Since M is symmetric,

$$\begin{aligned} \left|x'My\right| &= \left|x'M^{1/2}M^{1/2}y\right| \leq \sqrt{x'Mx}\sqrt{y'My} \leq \phi_{\max}(M)\sqrt{x'x}\sqrt{y'y} \\ &\leq \phi_{\max}(M)\sqrt{k\|x\|_\infty^2}\sqrt{k\|y\|_\infty^2} = k\|x\|_\infty\|y\|_\infty\phi_{\max}(M). \end{aligned}$$

where the first inequality follows by the Cauchy-Schwarz inequality. We have (57). (58) follows by

$$(A-B)^{-1} = \left(A\left(I_k - A^{-1}B\right)\right)^{-1} = \left(I_k - \left(A^{-1}B\right)\right)^{-1} A^{-1} = \left(I_k + \sum_{l=1}^{\infty}\left(A^{-1}B\right)^l\right)A^{-1}$$

where the last line is the series representation of the inverse given $\phi_{\text{spec}}\left(A^{-1}B\right) < 1$. □

We prove one more result. If V is a $k \times k$ symmetric positive-definite matrix, and we partition it as

$$V = \begin{bmatrix} \underset{k_1\times k_1}{V_{11}} & \underset{k_1\times(k-k_1)}{V_{12}} \\ \underset{(k-k_1)\times k_1}{V_{21}} & \underset{(k-k_1)\times(k-k_1)}{V_{22}} \end{bmatrix}.$$

Let $W = V^{-1}$, and $W_{11} = \left(V_{11} - V_{12}V_{22}^{-1}V_{21}\right)^{-1}$ is the $k_1 \times k_1$ upper-left block of W.

Lemma 10. *For any $a \in \mathbb{R}^k$, we have*

$$\left(\phi_{\max}(V)\right)^{-1} a'a \leq a'W_{11}a \leq \left(\phi_{\min}(V)\right)^{-1} a'a \tag{59}$$

Proof. Since W_{11} is a submatrix of W, we have

$$\phi_{\min}(W) \leq \phi_{\min}(W_{11}) \leq \phi_{\max}(W_{11}) \leq \phi_{\max}(W),$$

so that

$$a'W_{11}a \leq \phi_{\max}(W_{11})\, a'a \leq \phi_{\max}(W)\, a'a = (\phi_{\min}(V))^{-1}\, a'a.$$

Similarly,

$$a'W_{11}a \geq \phi_{\min}(W_{11})\, a'a \geq \phi_{\min}(W)\, a'a = (\phi_{\max}(V))^{-1}\, a'a.$$

C Numerical Implementation of REL

The literature suggests implementing the EL optimization through the inner loop and outer loop (Owen, 2001; Kitamura, 2006). Much care has to be executed in the programming of the high-dimensional inner loop. After experimenting with several solvers, we find `MOSEK`, a commercial solver specialized in convex programming, efficiently returns reliable results. We use the `MOSEK Matlab toolbox`[11] in the simulations and the empirical example under its free academic license. In the box below is the chuck of `Matlab` code that formulates the problem.

```
% the criterion function
prob.opr = repmat('log', [n 1]);
prob.opri = zeros(n, 1);
prob.oprj = (1:n)';
prob.oprf = ones(n, 1);
prob.oprg = ones(n, 1);
% the constraints
prob.c = sparse( zeros(n, 1) );
prob.a = [ ones(1,n); h'] ; % data
prob.blc = [ 1; -tau*ones(m, 1) ]; % moment lower bound
prob.buc = [ 1; tau*ones(m, 1) ]; % moment upper bound
prob.blx = sparse( zeros(n, 1) ); % lower bound of pi's
prob.bux = ones(n, 1); % upper bound of pi's
```

Under a fixed trial value β, the probability mass $p = \{p_i\}_{i=1}^n$ is the parameter to be optimized. The first five lines tell the solver the criterion function is $\sum_{i=1}^n \log p_i$. The next six lines specify the constraints. `prob.a` corresponds to the data matrix in the center of (60) below, and `prob.blc` and `prob.buc` are associated with the lower bound and the upper

[11] `http://docs.mosek.com/7.0/toolbox.pdf`

bound of each restriction.

$$\begin{pmatrix} 1 \\ -\tau \\ \vdots \\ -\tau \end{pmatrix} \leq \begin{pmatrix} 1 & \cdots & 1 \\ h_{11}(\beta) & \cdots & h_{n1}(\beta) \\ \vdots & \ddots & \vdots \\ h_{m1}(\beta) & \cdots & h_{nm}(\beta) \end{pmatrix} \begin{pmatrix} p_1 \\ \vdots \\ p_n \end{pmatrix} \leq \begin{pmatrix} 1 \\ \tau \\ \vdots \\ \tau \end{pmatrix} \tag{60}$$

`prob.blx` and `prob.bux` specify $0 \leq p_i \leq 1$ for every i. We feed all the ingredients of the optimization problem into the following command

```
res = mskscopt(prob.opr, prob.opri, prob.oprj, prob.oprf,...
      prob.oprg, prob.c, prob.a, prob.blc, prob.buc, prob.blx, prob.bux,...
      [], 'maximize' );
```

It is recommended to check the problem status and solution status after execution. If the solution status is `'OPTIMAL'`, we are safe to collect the results; otherwise, say if the constraints are infeasible, we can assign `Inf` to the profile log-likelihood function.

```
if strcmp( res.sol.itr.solsta, 'OPTIMAL')
     L_hat = -res.sol.itr.pobjval; % the optimal value of the function
     gam = sparse( res.sol.itr.y/n ); % collect the Lagrangian multipliers
     gam(1) = []; % remove the 1st element associated with sum(pi) == 1
else
     L_hat = Inf; % specify Inf, or a large value to indicate infeasibility
end
```

D Implementation of Empirical Application

We first briefly explain the simulated moments in Section 5. Details can be found in Eaton et al. (2011, Section 4). Index the markets by $l = 0, 1, \ldots, L$, in which $l = 0$ represents the home country. The market condition is $\mathbf{v} = \left(v_l = \left(N_l^{(v)}, \bar{X}_l^{(v)}\right)\right)_{i=1}^{L}$, where $N_l^{(v)}$ is the probability of entry and $\bar{X}_l^{(v)}$ is the average sale on the l-th market. We take $\mathbf{v}$ as known and non-random. The following two steps simulate the artificial firm i's entry and sale to the country l, which allow us to calculate various of moments.

- Step 1. For a firm i, generate $\left(a_{il}^{(1)}, a_{il}^{(2)}\right) \sim$ i.i.d. Normal $(0, 1)$.
- Step 2. Let

$$\begin{aligned} \kappa_1 &= \exp\left(\frac{1}{2}\beta_1^2\beta_4^2\right) \\ \kappa_2 &= \left[\frac{\beta_1}{\beta_1 - 1} - \frac{\beta_1}{\beta_1 + \beta_2 - 1}\right] \exp\left\{\frac{1}{2}\left[\beta_3 + 2\beta_3\beta_4\beta_5(\beta_1 - 1) + \beta_4(\beta_1 - 1)^2\right]\right\}. \end{aligned}$$

Compute the following quantities one by one for each i and l.

$$\begin{aligned}
\alpha_{il}^{(1)} &= \exp\left(\beta_3\left(1-\beta_5^2\right)^{1/2} a_{il}^{(1)} + \beta_3\beta_5 a_{il}^{(2)}\right) \\
\alpha_{il}^{(2)} &= \exp\left(\beta_4 \alpha_{il}^{(2)}\right) \\
\bar{u}_{il} &= \left(\alpha_{ij}^{(2)}\right)^{\beta_1} N_l/\kappa_1 \\
\bar{u}_l &= \bar{u}_{i,l=0} \wedge \max_{l=1,\ldots,m^a} \bar{u}_{il} \\
u_l &= v_l \bar{u}_l \\
\delta_{il} &= \mathbf{1}\left\{u_l \leq \bar{u}_{il}\right\} \\
X_{il} &= \delta_{il}\left(\alpha_{il}^{(1)}/\alpha_{il}^{(2)}\right)\left(1-u_l/\bar{u}_{il}\right)^{\beta_2/\beta_1}\left(u_l/\bar{u}_{il}\right)^{-1/\beta_1}\bar{X}_l\kappa_1/\kappa_2.
\end{aligned}$$

The last two quantities δ_{il} and X_{il} are the entry and sale, respectively.

EKK create the majority of their moments by estimating several quantiles of the sales in a market, and use these estimated quantiles to categorize the firms into bins, so that the moments are the average of indicator functions. To apply the second step bias-correction, we avoid non-differentiable functions. We match the mean sales in each destination country. The variation of the entry probabilities over the L countries identifies the 5 parameters of interest.

In the simulation exercise in Section 5.2, we vary

$$N_l^{(v)} = \begin{cases} 1, & \text{if } l = 0; \\ 1.5, & \text{if } l = 1,\ldots,6; \\ 0.3, & \text{if } l > 6. \end{cases}$$

while keep $\bar{X}_l^{(v)} = 1$ for all $0 \leq l \leq L$.[12] Under the true parameter value we specify, we simulate n firms as the "real data"; then in the estimation we generate another $5n$ firms as the "artificial data", and we match the mean sales of the artificial data to the "real data".

The real data application differs from the simulation in (i) the 2006 sample is used as the real data, (ii) the market condition $\mathbf{v}$ is calculated from the 2005 sample. As we ignore the randomness in $\mathbf{v}$, we select the countries that in 2005 at least 30 firms enter. In addition, to ensure the comparability of the 2005 and 2006 mean sales, we eliminate three countries in which the average sale in 2005 is more than 150 times larger than that in 2006. These large changes may indicate dramatic shift of the market environment in those countries, a phenomenon that EKK's theory does not intend to explain. After the elimination, we are left with 127 countries, including China.

[12] In the simulation exercise where we know the values of the parameters, we do not require the simulated mean sale being close to $\bar{X}_l^{(v)}$.