Validity vs Information: Generalized Focused Moment Selection for GEL with Possibly Many Moments*

Di Tian
University of Pennsylvania

This Version: March 25, 2021

Abstract

More assumptions lead to more precision, bearing the risks of less credibility. To balance information and validity, I propose a moment selection criterion for GEL estimation in moment-based models, facing possibly misspecified moments and/or many moments. The criterion, generalized focused information criterion (GFIC), trades off bias and variance in the limit as in finite samples, rather than allow validity to dominate. In the setting of a growing number of moments, GEL generally has preferable performance, as shown in the simulation study.

Keywords: Moment selection, GEL estimation, Focused Information Criterion, Many Moments

1 Introduction

In econometrics, an incorrect model is not always unhelpful with regard to estimation. A simple example would be the inclusion of a possibly endogenous but relevant instrument, which introduces bias but compensates by decreasing the estimation variance. The question that whether the added instrument is worth the trouble leads to the idea of this paper. Selection of perfectly exogenous instruments is often difficult. This is especially true with an increasing number of proposed instruments. In this paper, I attempt to balance validity (bias) and information (variance) and derive in the framework that allows for many moments the selection criteria, which aim to minimize the asymptotic risk of the target parameter.

I focus on moment-based models, where model selection problem reduces to moment selection. Among moment-based estimation, generalized moment method (GMM) is most notable while there exist alternatives designed to improve upon GMM, including empirical likelihood (EL) estimator of Owen (1997), Qin and Lawless (1994) and Imbens (1997), the continuous updating estimator (CUE) of Hansen et al. (1996), and the exponential tilting (ET) estimator of Stutzer and Kitamura (1997) and Imbens et al. (1998). As shown by

*I thank Xu Cheng, Karun Adusumilli, and Francis DiTraglia for their many helpful comments and suggestions.

Smith (1997) and Newey and Smith (2004), all improved alternatives fall into the general structure of generalized empirical likelihood (GEL) estimation.

There are several advantages of GEL compared to GMM. First, GEL generally has lower higher order bias, especially when many moments are present like many instruments or when bias from estimation of the weighting matrix can be a serious problem like minimum distance estimation of panel data models. The simulation study in this paper corroborates this phenomenon. Second, some GEL estimators, especially EL, provides higher order variance efficiency which is quite noticeable in previous nonlinear Monte Carlo simulations. In this paper, I discuss moment selection criteria in the GEL estimator group.

Built on the trade-off between bias and variance, I apply a moment selection criterion – generalized focused information criterion (GFIC) – to GEL estimation with possibly many moments. GFIC does not select the correct model or the maximum set of correct moments but select a set of potentially mis-specified moments with the goal of minimizing asymptotic risk of a scalar target parameter. This means the selection problem is highly contingent on the parameter of interest and different targets could lead to different moments being the most useful.

I adopt a local asymptotic framework to ensure asymptotic bias and variance are comparable. Under this framework, all GEL estimators remain consistent despite the presence of locally mis-specified moments but a non-vanishing bias appear in the scaled asymptotic distribution. Therefore, GFIC deals with the bias-variance trade-off in the limit to approximate finite sample behavior. Among all potential moments, there are one "baseline" block that is generally acknowledged to be correct but most likely to be less informative and another block that is more controversial due to suspectable assumptions made but highly relevant to the target parameter at hand. For commonly used loss functions (L_p and linex), GFIC provides an asymptotically unbiased estimator of L_2 risk (AMSE) and linex risk, while only a plug-in estimator of L_1 risk (AMAE) is available. I continue to discuss a simulation procedure for valid confidence intervals of the target parameter proposed by DiTraglia (2016).

I exemplify this method in an empirically relevant example: a linear instrument setup where the researchers decide on the addition of strong instrumental variables that are potentially endogenous. To be exact, this setup incurs an instrument selection problem, a special case of which is choosing between ordinary least squares (OLS) and two-stage least squares (TSLS) estimators. According to the simulation study, GFIC performs well in terms of MSE and often outperforms traditional information criteria when sample size is small.

The generalized focused information criterion is originally inspired by the focused information criterion (FIC) of Claeskens and Hjort (2003), a model selection criterion for maximum likelihood estimation, and then extended to GMM estimation in DiTraglia (2016) and Chang and DiTraglia (2018). With similar drifting asymptotic framework, this paper, however, expands the horizon to a more general group of estimation - GEL estimation, and to the increasingly popular framework of the modern age which accommodates a growing number of moments. While the idea of asymptotics under local mis-specification dates back to Newey (1985), application to the general GEL group with possibly many moments is novel. Also, this paper discusses asymptotic risks derived from different loss functions rather than solely AMSE. Even though Claeskens et al. (2006) and Claeskens and Hjort (2008) briefly mentioned L_1 and linex loss, respectively, for FIC with respect to the maximum likelihood estimation, this paper is the first to study their theoretical properties in moment-based

models.

The line of literature on moment selection primarily covers consistent selection (See Andrews (1999) Andrews and Lu (2001) and Hong et al. (2003)). That is to select the maximum set of correctly specified moments consistently. And because of this consistency requirement, they discuss moment selection in a non-vanishing mis-specification setting instead of a local mis-specification one.

In contrast to consistent moment selection, another perspective is "efficient" moment selection. Here the goal is to select the set of moments which result in an "efficient" or "optimal" estimator in some sense. DiTraglia (2016) is the first to propose the use of GFIC for target parameter. Hall and Peixe (2003) and Cheng and Liao (2013) aim to avoid including redundant moment conditions after consistently eliminating invalid ones. Other proposals that consider minimizing MSE track high-order bias instead of first-order bias arising from local mis-specification, including Donald and Newey (2001), Donald et al. (2009), and Kuersteiner and Okui (2010).

GFIC is a *conservative* rather than consistent measure: the criterion does not converge in probability but in distribution. While consistency is generally a desirable trait, moment selection problem presents a complication: consistent and conservative measures have different merits, which proves unable to combine (Yang, 2005). GFIC is designed to mimic minimum-risk estimation, contrary to what consistent selection exhibits - unbounded minimax risk (Leeb and Pötscher, 2008).

Another strand of relevant literature pertains to the many moment framework. Newey and Windmeijer (2009) develop many weak moment asymptotics for GEL estimators with perfectly exogenous instruments. The approximated new limit improves the finite sample results. Chao and Swanson (2005) derive the linear case, and Han and Phillips (2006) consider GMM. Instead the many weak moment setup, Donald et al. (2003) focus on conditional moment restrictions which induce a many moment structure. Several recent papers have explored the testing of exogeneity violations. Berkowitz et al. (2012) provide a new resampling technique with the presence of local exogeneity violations. And Kolesr et al. (2015) analyze estimation in the context of many invalid instruments. This paper goes in a different direction. Instead of testing, I restrict my attention to the bias-variance trade-off when local exogeneity violation is present.

The remainder of the paper is organized as follows. Section 2 describes the drifting asymptotic framework and Section 3 derives the GFIC in general form and provides an inference method. Section 4 implements GFIC in a linear example, instrument selection, and displays simulation results. Proofs, computational details and supplementary material appear in the Appendix.

2 Assumptions and Asymptotic Framework

2.1 Local Mis-Specification

Let $f(\cdot,\cdot)$ be a m_n-vector valued moment functions of a random vector Z and an r-dimensional parameter vector θ, where the number of moments m_n might grow with the sample size n but the number of parameters r is fixed. For notational purposes, I define the following

quantities with i.i.d. data $\{Z_i\}_{i=1}^n$.

$$
\begin{aligned}
f_i(\theta) =& f(Z_i, \theta), \quad \widehat{f}(\theta) = \sum_i f_i(\theta)/n, \quad \bar{f}(\theta) = \mathbb{E}[f_i(\theta)] \\
\widehat{\Omega}(\theta) =& \sum_i f_i(\theta) f_i(\theta)'/n, \quad \Omega(\theta) = \mathbb{E}[f_i(\theta) f_i(\theta)'], \quad \Omega = \Omega(\theta_0) \\
\widehat{F}(\theta) =& \frac{1}{n} \sum_i \partial f_i(\theta)/\partial\theta, \quad F_i(\theta) = \partial f_i(\theta)/\partial\theta, \quad F(\theta) = \mathbb{E}[\partial f_i(\theta)/\partial\theta], \quad F = F(\theta_0)
\end{aligned}
$$

The function f is partitioned according to $f(\cdot,\cdot) = (g(\cdot,\cdot)', h(\cdot,\cdot)')'$, where $g(\cdot,\cdot)$ and $h(\cdot,\cdot)$ are p_n- and q_n-vectors of moment functions. Though all moment conditions could potentially be utilized to identify the true values of the parameters, the conditions associated with $g(\cdot,\cdot)$ are the "baseline" and assumed to be correctly specified, while those from $h(\cdot,\cdot)$ are potentially locally mis-specified. More precisely,

Assumption 2.1. $\mathbb{E}[f_i(\theta_0)] = \eta/n^\zeta$, *where* θ_0 *lies in the interior of a compact set* Θ, $\eta = [0'_{p_n}, \widetilde{\eta}'_{q_n}]'$ *is a* m_n*-vector and* $\zeta \geq 1/2$. *Define* $\delta(\theta) = \theta - \theta_0$. *Then there exist* $C > 0$ *and* $\widehat{D} = O_p(1)$*: i)* $||\delta|| \leq C||\bar{g}(\theta)||$ *for all* $\theta \in \Theta$*; ii) w.p.a.1,* $||\delta|| \leq C||\widehat{g}(\theta)|| + \widehat{D}$ *for all* $\theta \in \Theta$*; iii)* $\max_{l=1,\ldots,m} |\eta_l| \leq C$.

Above assumption is adopted to demonstrate local mis-specification, which is merely a tool to mimic finite-sample bias-variance trade-off in the asymptotic sense. All expectations are taken with respect to Z_i for sample size n, where I suppress the subscript n. For a fixed sample size, the additional moment condition h valued at true parameter θ_0 has expectation $\widetilde{\eta}/n^\zeta$, which is non-zero unless $\widetilde{\eta}$ is a zero-vector. This indicates locally it is mis-specified at a finite sample size and will introduce a comparable asymptotic bias. However, in the limit, the mis-specification vanishes, implying the convergence of moment expectations to 0. Assumption 2.1 i) and ii) are employed to ensure global identification of θ_0, as is standard in Newey and Windmeijer (2009).

2.2 Candidate GEL Estimators

A candidate GEL estimator $\widehat{\theta}_S$ only considers a subset S of all the moment conditions f in estimation and disregards the rest. Let v be the subset that only includes all the valid moments g and I only consider subsets which include g, i.e. $S \in \mathcal{S} = \{S : v \subset S\}$. This is a straightforward implementation because the valid moments only introduce more information, thus reducing estimation variance, but not any bias. Consequently, it is helpful to utilize all moments from the g block. This structure accommodates a wide range of problems. One example would be choosing between informative but biased IV and valid IV.

To ensure the GEL estimator is well-defined and unique, I assume $p_n \geq r$. Next, let Ξ_S be the $|S| \times m$ *moment selection matrix* corresponding to S. That is, Ξ_S is a matrix of ones and zeros arranged such that $\Xi_S f$ contains only the sample moment conditions used to estimate $\widehat{\theta}_S$. Define the corresponding moment function, covariance matrix and Jacobian matrix with respect to S as $f_S = \Xi_S f$, $\Omega_S = \Xi_S \Omega \Xi'_S$, and $F_S = \Xi_S F$. Thus, the GEL estimator of θ based on moment set S is given by

$$\widehat{\theta}_S = \underset{\theta\in\Theta}{\text{argmin}}\ \widehat{Q}_S(\theta) = \underset{\theta\in\Theta}{\text{argmin}}\ \sup_{\lambda\in\widehat{\Lambda}_S(\theta)}\ \widehat{P}_S(\theta,\lambda)$$
$$= \underset{\theta\in\Theta}{\text{argmin}}\ \sup_{\lambda\in\widehat{\Lambda}_S(\theta)} \sum_i \rho(\lambda'\Xi_S f_i(\theta))/n$$

where ρ is a concave function on an open neighborhood $\mathcal{V}$ around 0, and $\hat{\Lambda}_S(\theta) = \{\lambda : \lambda'\Xi_S f_i(\theta) \in \mathcal{V}, i = 1, ..., n\}$.

To guarantee a desirable performance of GEL estimation, the following smooth condition of ρ is required.

Assumption 2.2. *$\rho(u)$ is concave and three times continuously differentiable on $\mathcal{V}$. Without loss of generality, I normalize ρ so that we have $\rho'(0) = \rho''(0) = -1$.*

Above assumption is satisfied by the common GEL estimators, including CUE, where ρ is a quadratic function, EL, where $\rho = \ln(1-v)$, and ET, where $\rho = -e^v + 1$. Beside the conditions on ρ, we also need conditions on convergence of $\widehat{f}(\theta)$, as imposed in the following condition.

Assumption 2.3. *$m/n \to 0$. f is continuous in θ and there is $C > 0$ such that: i) $\sup_{\theta\in\Theta} \mathbb{E}[\{f_i(\theta)'f_i(\theta)\}^2]/n^3 \to 0$; ii) $\sup_{\theta\in\Theta} ||\widehat{\Omega}(\theta) - \Omega(\theta)|| \to_p 0$; iii) $1/C \leq \xi_{\min}(\Omega_v(\theta))$ and $1/C \leq \xi_{\min}(\Omega(\theta)) \leq \xi_{\max}(\Omega(\theta)) \leq C$ for all $\theta \in \Theta$; iv) $|a'[\Omega(\widetilde{\theta}) - \Omega(\theta)]b| \leq ||a||||b||||\widetilde{\theta} - \theta||$ for all $\widetilde{\theta}, \theta \in \Theta$ and $a, b \in R^m$; v) for every $\widetilde{C}$, there is C and $\widehat{D} = O_p(1)$ such that for all $\widetilde{\theta}, \theta \in \Theta$, $||\delta(\widetilde{\theta})|| \leq \widetilde{C}$, $||\delta(\theta)|| \leq \widetilde{C}$, we have $||\bar{f}(\widetilde{\theta}) - \bar{f}(\theta)|| \leq C||\widetilde{\theta} - \theta||$ and $||\widehat{f}(\widetilde{\theta}) - \widehat{f}(\theta)|| \leq \widehat{D}||\widetilde{\theta} - \theta||$; vi) $\{n\mathbb{E}[\sup_{\theta\in\Theta} ||f_i(\theta)||^\alpha]\}^{1/\alpha}\sqrt{m/n} \to 0$ for some $\alpha > 2$.*

These assumptions follow directly from the standard framework of Newey and Windmeijer (2009), which would lead to a rate requirement that is slightly stronger than $m^2/n \to 0$ under certain regularities. And as a direct result of above conditions, we have consistency of all candidate GEL estimators.

Theorem 2.1 (Consistency)**.** *Under Assumption 2.1 - 2.3, we have i) the GEL estimator, $\widehat{\theta}_S \to_p \theta_0$; ii) $||\widehat{f}_S(\widehat{\theta})|| = O_p(\sqrt{|S|/n})$; iii) $\widehat{\lambda}_S = argmax_{\lambda\in\widehat{\Lambda}_S(\widehat{\theta}_S)} \widehat{P}_S(\widehat{\theta}_S, \lambda)$ exists w.p.a.1 and $||\widehat{\lambda}_S|| = O_p(\sqrt{|S|/n})$.*

As we see from Theorem 2.1, of interest is the fact that *any* candidate GEL estimator $\widehat{\theta}_S$ is consistent for θ_0 under local mis-specification. Alternatively, local mis-specification does not affect consistency and will only show up after proper scaling as in Theorem 2.2. The remainder of Theorem 2.1 is standard GEL result needed for derivation of the limit distribution. But for asymptotic normality, further assumptions are needed to regulate the behavior of the derivatives.

Assumption 2.4. *$m^3/n \to 0$. f is twice continuously differentiable in a neighborhood $\mathcal{N}$ of θ_0 and there is $C > 0$ and $\widehat{D} = O_p(1)$ such that: i) $\xi_{\max}(\mathbb{E}[F_i(\theta_0)F_i(\theta_0)']) \leq C$; ii) if $\widetilde{\theta} \to_p \theta_0$, then $||\widehat{F}(\widetilde{\theta}) - \widehat{F}(\theta_0)|| \leq \widehat{D}||\widetilde{\theta} - \theta_0||$; iii) $\mathbb{E}[\sup_{\theta\in\mathcal{N}} ||f_i(\theta)||^4]\sqrt{m/n} \leq C$ and $\mathbb{E}[\sup_{\theta\in\mathcal{N}} |f_i^k(\theta)|^4] \leq C$ where f^k is the k-th element of f; vi) $F_S'\Omega_S^{-1}F_S \to U_S$, which is non-singular, and $n^{\frac{1}{2}-\zeta}F_S'\Omega_S^{-1}\Xi_S\eta \to \tau_S$, for all $S \in \mathcal{S}$.*

Above assumptions differ from those of Newey and Windmeijer (2009), because they consider a (near-) weakly identified framework, while I restrict my attention to the strong identification case only. However, both require similar rate condition where m^3 must grow slower than n. The additional drift term τ_S, which can be zero or non-zero, is exactly what governs the asymptotic bias introduced by using the potentially mis-specified moments in the h block. Note when $\zeta = 1/2$, a fixed q_n is needed to bound τ_S; and when $\zeta < 1/2$, we can allow q_n to grow together with m_n.

Under these regularity conditions, we can show that $\widehat{\theta}_S$ is asymptotically normal and that the variance estimator is consistent.

Theorem 2.2 (Asymptotic Normality)**.** *Under Assumption 2.1 - 2.4, we have*

$$\sqrt{n}(\widehat{\theta}_S - \theta_0) \to_d N(-U_S^{-1}\tau_S, U_S^{-1})$$

and $\widehat{U}_S = \widehat{F}_S(\widehat{\theta}_S)'\widehat{\Omega}_S^{-1}(\widehat{\theta}_S)\widehat{F}_S(\widehat{\theta}_S) \to_p U_S$.

The mis-specification parameter τ_S is introduced as the asymptotic bias in the limit distribution of Theorem 2.2 as long as h block moments are used. And this is precisely the reason why I employ local mis-specification framework to approximate finite-sample bias-variance trade-off we see in practice by an asymptotic one. The carefully designed local mis-specification rate and the drift term guarantee that bias and variance are comparable in the limit.

2.3 Identification

In all moment selection problems, an identifying assumption is required so that the true value of parameter θ_0 is clearly identifiable. One approach to identification is to assume that there exists a minimal set of at least r moment conditions known to be valid. This is the approach I follow here, as do Cheng and Liao (2013) and DiTraglia (2016).

Let $\widehat{\theta}_v$ denote the GEL estimator based solely on the moment conditions contained in the g block

$$\widehat{\theta}_v = \underset{\theta\in\Theta}{\text{argmin}} \sup_{\lambda\in\hat{\Lambda}_v(\theta)} \sum_i \rho(\lambda' g(Z_i,\theta))/n$$

where $\hat{\Lambda}_v(\theta) = \{\lambda : \lambda' g(Z_i,\theta) \in \mathcal{V}\}$. We call this the "valid estimator". From the fact Assumption 2.1, 2.2 and 2.4 are all satisfied by g, Theorem 2.2 immediately imply that the valid estimator shows no asymptotic bias.

Corollary 2.1 (Limit Distribution of Valid Estimator)**.** *Under Assumption 2.1 - 2.4, we have*

$$\sqrt{n}\left(\widehat{\theta}_v - \theta_0\right) \to_d N(0, U_v^{-1})$$

where $U_v = \lim F_v'\Omega_v^{-1}F_v$.

Corollary 2.1 highlights the difference between the two blocks of moment conditions. Whereas the g block contains widely accepted "baseline" assumptions about the world, the h block contains stronger assumptions. One would set aside these suspectable moment

conditions unless they bring strong information that outweighs the drawback. This is what GFIC intends to deal with: whether new information, model or theory is worth considering given its potential invalidity. In the statistical sense, the trade-off happens between bias (controversy) and variance (information).

Below, I discuss the GFIC moment selection criteria in a general case, which is designed to minimize loss of a target parameter. Then I will follow through with a simple but empirically relevant example, which is a linear example where researchers choose among instruments.

3 The Generalized Focused Information Criterion

3.1 The General Case

The GFIC is target-specific. Given a scalar target parameter of interest, GFIC chooses the set of moment conditions that provides a GEL estimator with optimal asymptotic risk. Denote this target parameter by μ, a real-valued, Z-almost continuous function of the parameter vector θ that is differentiable in a neighborhood of θ_0. Further, define the GEL estimator of μ based on $\widehat{\theta}_S$ by $\widehat{\mu}_S = \mu(\widehat{\theta}_S)$ and the true value of μ by $\mu_0 = \mu(\theta_0)$. Applying the Delta Method to Theorem 2.2 gives the asymptotic distribution of $\widehat{\mu}_S$.

Corollary 3.1 (Asymptotic Risk of Target)**.** *Under the hypotheses of Theorem 2.2,*

$$\sqrt{n}\left(\widehat{\mu}_S - \mu_0\right) \to_d N(\kappa_S'\tau_S, \kappa_S' U_S \kappa_S)$$

where $\kappa_S = -U_S^{-1}\nabla_\theta \mu(\theta_0)$*. AMAE* ($L_1$) *and AMSE* ($L_2$) *are respectively* [1]*:*

$$\begin{aligned} AMAE\left(\widehat{\mu}_S\right) =& 2\kappa_S'\tau_S\left(\Phi\left(\frac{\kappa_S'\tau_S}{\sqrt{\kappa_S' U_S \kappa_S}}\right) - \frac{1}{2}\right) + 2\sqrt{\kappa_S' U_S \kappa_S}\,\phi\left(\frac{\kappa_S'\tau_S}{\sqrt{\kappa_S' U_S \kappa_S}}\right) \\ AMSE\left(\widehat{\mu}_S\right) =& \kappa_S'(\tau_S\tau_S' + U_S)\kappa_S \end{aligned}$$

where Φ *and* ϕ *are cdf and pdf of a standard normal.*

Finally, we have the asymptotic linex risk, where the linex loss is $L(x_1, x_2) = \exp[c(x_1 - x_2)] - c(x_1 - x_2) - 1$*:*

$$ALL\left(\widehat{\mu}_S\right) = \exp\left(c^2\kappa_S' U_S \kappa_S/2 + c\kappa_S'\tau_S\right) - c\kappa_S'\tau_S - 1$$

where c *determines the degree of asymmetry and skewness and can be positive and negative.*

It is straightforward that the valid estimator $\widehat{\mu}_v$ of μ has zero asymptotic bias. The inclusion of h block moments translates the local mis-specification into the asymptotic bias of the candidate estimator $\widehat{\mu}_S$, whose scale is determined by the local mis-specification drift τ_S and the scale κ_S.

Using results in Corollary 3.1 for moment selection, GFIC is minimum-risk based. If the asymptotic risk is known, the optimal set of moments is the one minimizing the asymptotic

[1] See Appendix A for general L_p risks.

risk. Consequently, GFIC estimates the asymptotic risk and selects the set of moments which minimizes the estimated risk.

And estimating asymptotic risk requires estimators of the unknown quantities: θ_0, U_S, κ_S, and τ_S. Under local mis-specification, any GEL estimator $\widehat{\theta}_S$ is consistent for θ_0. One would consider the valid estimator $\widehat{\theta}_v$ as a natural choice, though other possibilities are also justified. Consistent estimators of U_S are described in Theorem 2.2. Recall that $\kappa_S = -U_S^{-1}\nabla_\theta \mu(\theta_0)$, which can be consistently estimated given consistent $\widehat{U}_S$ and $\widehat{\theta}_S$. The only remaining unknown is τ_S. Local mis-specification is essential for keeping asymptotic bias comparable to variance and from dominating the asymptotic risk. Unfortunately, it also prevents consistent estimation of the asymptotic bias and local mis-specification drift τ_S. Otherwise, one would prefer absorb the local mis-specification into the moment function. But Assumptions 2.1 and 2.4 allow us to construct an *asymptotically unbiased* estimator $\widehat{\tau}_S$ of τ_S by substituting $\widehat{\theta}_v$, the *asymptotically unbiased* estimator of θ_0 that uses only correctly specified moment conditions, into properly scaled $\widehat{f}_S(\theta)$, the sample analogue of the utilized moment conditions.

Theorem 3.1 (Asymptotic Distribution of $\widehat{\tau}_S$)**.** *Let* $\widehat{\tau}_S = n^{-3/2}\widehat{F}_S(\widehat{\theta}_v)'\widehat{\Omega}_S(\widehat{\theta}_v)^{-1}\widehat{f}_S(\widehat{\theta}_v)$ *where* $\widehat{\theta}_v$ *is the valid estimator, based only on the moment conditions contained in* g*. Then under Assumptions 2.1 - 2.4*

$$\widehat{\tau}_S \to_d N(\tau_S, U_S U_v^{-1} U_S - U_S)$$

Returning to Corollary 3.1, however, we see that it is not just τ_S that enters the expression for asymptotic risks. For AMAE, it is also concerned with Gaussian cumulative density function; for AMSE, it is $\tau_S\tau_S'$ which enters the formula; for linex risk, it is related to exponentials of τ_S. Although $\widehat{\tau}_S$ is an asymptotically unbiased estimator of τ_S, the limiting expectation of a non-linear function $\widehat{\tau}_S$ is asymptotically biased because $\widehat{\tau}_S$ has an asymptotic variance.

For AMSE and linex, it is possible to remove this bias.

Corollary 3.2 (Asymptotically Unbiased Estimators)**.** *Under Assumptions 2.1 - 2.4, let* $\widehat{\Upsilon}_S = \widehat{U}_S\widehat{U}_v^{-1}\widehat{U}_S - \widehat{U}_S$*. If* $\widehat{U}_S$ *and* $\widehat{\kappa_S}$ *are consistent estimators for* U_S *and* κ_S*, then* $\widehat{\tau}_S\widehat{\tau}_S' - \widehat{\Upsilon}_S$ *is an asymptotically unbiased estimator of* $\tau_S\tau_S'$ *and* $\exp(c\widehat{\kappa}_S'\widehat{\tau}_S - c^2\widehat{\kappa}_S'\widehat{\Upsilon}_S\widehat{\kappa}_S/2)$ *is an asymptotically unbiased estimator of* $\exp(c\kappa_S'\tau_S)$*.*

It follows that

$$\text{FMSC}_n(S) = \widehat{\kappa_S}'[\widehat{\tau}_S\widehat{\tau}_S' - \widehat{\Upsilon}_S + \widehat{U}_S]\widehat{\kappa_S} \tag{1}$$

$$\text{FMLC}_n(S) = \exp\left\{c^2\widehat{\kappa}_S'[\widehat{U}_S - \widehat{\Upsilon}_S]\widehat{\kappa}_S/2 + c\widehat{\kappa}_S'\widehat{\tau}_S\right\} - c\widehat{\kappa}_S'\widehat{\tau}_S - 1 \tag{2}$$

provides an asymptotically unbiased estimator of AMSE and asymptotic linex risk.

For other risks of consideration, we fail to identify the exact asymptotic bias because of the complexity of the transformation. This means that we cannot have an unbiased estimate for AMAE, and in this case a plug-in estimator with $\widehat{\tau}_S$ is often used:

$$\text{FMAC}_n(S) = 2\widehat{\kappa}_S'\widehat{\tau}_S\left(\Phi\left(\frac{\widehat{\kappa}_S'\widehat{\tau}_S}{\sqrt{\widehat{\kappa}_S'\widehat{U}_S\widehat{\kappa}_S}}\right) - \frac{1}{2}\right) + 2\sqrt{\widehat{\kappa}_S'\widehat{U}_S\widehat{\kappa}_S}\,\phi\left(\frac{\widehat{\kappa}_S'\widehat{\tau}_S}{\sqrt{\widehat{\kappa}_S'\widehat{U}_S\widehat{\kappa}_S}}\right) \tag{3}$$

For all generalized focused information criteria (GFIC), including FMAC, FMSC, FMLC and other loss function varieties, given a set $\mathcal{S}$ of candidate specifications, the GFIC selects the candidate S^* that *minimizes* the estimated asymptotic risk criterion, that is $S^*_{GFIC} = \arg\min_{S\in\mathcal{S}} \text{GFIC}_n(S)$.

As discussed above, local mis-specification prevents us to consistently estimate asymptotic risk.[2] The plug-in estimator based on $\widehat{\tau}_S$ is easy to implement, but always bears the risk of risk estimate "overshooting" or "undershooting" asymptotically depending on the loss function. Accordingly, it is natural to correct this bias as explained in Corollary 3.2. This is the same heuristic that constitutes AIC and TIC model selection criteria as well as more recent procedures such as those described in Claeskens and Hjort (2003) and Schorfheide (2005). Nevertheless, asymptotically unbiased estimators have their advantages but also disadvantages, with availability being one. For example, GFIC fails to derive an asymptotically unbiased risk estimator for AMAE because of the complicated transformation. The plug-in estimator, however, is always available.

3.2 Inference of μ

Corollary 3.1 characterizes the limit distribution of $\widehat{\mu}_S$, which unfortunately depends on a quantity τ_S which cannot be consistently estimated. This implies there is no readily available closed-form confidence interval for the estimator, and alternative measures are needed to conduct inference. Here in this paper, I employ a method from DiTraglia (2016) for valid confidence interval construction, which delivers a two-step conservative procedure for asymptotically valid confidence intervals.

Instead of a closed-form solution, this method employs a simulation-based approach. Returning to the limiting distribution in Corollary 3.1, we note $\widehat{\mu}_S \to_d \kappa'_S[\tau_S + M_S]$, where $M_S \sim N(0, U_S)$. So if all quantities and functions were known, the only randomness would come from a multivariate variable M_S, which can be simulated.

There are several remaining quantities to be estimated. First, we already have consistent estimators of θ_0, U_S and thus κ_S, all of which are also used in the calculation of GFIC. So everything else is consistent except $\widehat{\tau}_S$, which has a normal limit distribution as suggested by Theorem 3.1. To deal with $\widehat{\tau}_S$, I employ a two-stage procedure as suggested by DiTraglia (2016). First construct a $(1-\delta)\times 100\%$ confidence region $\mathcal{T}(\widehat{\tau}_S, \delta)$ for τ_S using Theorem 3.1. Then, for each $\tau^*_S \in \mathcal{T}(\widehat{\tau}_S, \delta)$, I simulate from the limiting distribution of $\widehat{\mu}_S$, defined in Corollary 3.1, to obtain a *collection* of $(1-\alpha)\times 100\%$ confidence intervals indexed by τ^*_S. Taking the lower and upper bounds of these intervals yields a *conservative* confidence interval for $\widehat{\mu}_S$. This interval has asymptotic coverage probability of *at least* $(1-\alpha-\delta)\times 100\%$. The precise algorithm is as follows.

1. For each $\tau^*_S \in \mathcal{T}(\widehat{\tau}_S, \delta)$

   (i) Generate J independent draws $M_{S,j} \sim N(0, \widehat{U}_S)$

   (ii) Set $\varphi_j(\tau^*_S) = \widehat{\kappa}'_S(\tau^*_S + M_{S,j})$

[2] This is not a defect of the GFIC: there is a fundamental trade-off between consistency and desirable risk properties.

(iii) Using the draws $\{\varphi_j(\tau_S^*)\}_{j=1}^J$, calculate $\widehat{a}(\tau_S^*)$ and $\widehat{b}(\tau_S^*)$ such that

$$P\left\{\widehat{a}(\tau_S^*) \leq \varphi(\tau_S^*) \leq \widehat{b}(\tau_S^*)\right\} = 1-\alpha$$

2. Set $\widehat{a}_{min}(\widehat{\tau}_S) = \min_{\tau_S^* \in \mathcal{T}(\widehat{\tau}_S,\delta)} \widehat{a}(\tau_S^*)$ and $\widehat{b}_{max}(\widehat{\tau}_S) = \max_{\tau_S^* \in \mathcal{T}(\widehat{\tau}_S,\delta)} \widehat{b}(\tau_S^*)$

3. The confidence interval for μ is $\text{CI}_{sim} = \left[\widehat{\mu}_S - \dfrac{\widehat{b}_{max}(\widehat{\tau}_S)}{\sqrt{n}}, \quad \widehat{\mu}_S - \dfrac{\widehat{a}_{min}(\widehat{\tau}_S)}{\sqrt{n}}\right]$

Under mild regularity conditions, with consistent estimates $\widehat{\theta}, \widehat{U}_S$ and $\widehat{\kappa}_S$, above confidence interval is valid through DiTraglia (2016) **Theorem 4.4**.

4 Simulations

4.1 Choosing Instrumental Variables Example

One straightforward application of the GFIC is choosing instruments in a general linear setting. Putting instrument selection in the framework of this paper, GFIC amounts to trading off endogeneity against instrument relevance. I now look at GEL estimators in an i.i.d. linear setting. Consider the following model:

$$y_i = x_i'\beta + \epsilon_i \tag{4}$$
$$x_i = \Pi_1' z_i^{(1)} + \Pi_2' z_i^{(2)} + v_i \tag{5}$$

where y is the outcome variable, x is an r-vector of regressors, some of which are endogenous, $z^{(1)}$ is a p-vector of instruments known to be exogenous, and $z^{(2)}$ is a q-vector of *potentially endogenous* instruments. The number of instruments p and q could be growing with the sample size n, resulting in the many instrument problem. The r-vector β, $p \times r$ matrix Π_1, and $q \times r$ matrix Π_2 are unknown parameters, with β being the main focus. The stacked system in matrix form is $y = X\beta + \epsilon$ and $X = Z\Pi + V$, where $Z = (Z_1, Z_2)$ and $\Pi = (\Pi_1', \Pi_2')'$.

In this example, the exogeneity of $z^{(1)}$ is associated with the valid g block moments while the strong instruments $z^{(2)}$ is associated with the h block moments where local misspecification arises from potential endogeneity. They would be a great source of information and variance reduction if we were confident about their exogeneity. Yet considering strong instruments indicate high relevance and correlation between $z^{(2)}$ and x, the exogeneity assumption on them seems dubious. In contrast, less informative instruments $z^{(1)}$ are exempt from this type of doubt and readily accepted to be valid. Therefore, the GFIC attempts to balance the introduced bias from using potentially endogenous instruments and the decreased variance from more informative and stronger instruments. To this end, we have our GEL estimators:

Theorem 4.1 (Choosing IVs Limit Distribution)**.** *Let (z_i, v_i, ϵ_i) be random variables such that $\mathbb{E}[z_i v_i] = 0$, $\mathbb{E}[z_i^{(1)}\epsilon_i] = 0$ and $\mathbb{E}[z_i^{(2)}\epsilon_i] = \eta/n^\zeta$ for all n. Let $\widehat{\beta}_S$ be a GEL estimator from*

a concave twice continuously differentiable function ρ using instruments $\Xi_S z_i$. Then, under Assumptions 2.1 - 2.4 [3]*,*

$$\sqrt{n}(\widehat{\beta}_S - \beta_0) \to_d N(-U_S^{-1}\tau_S, U_S^{-1})$$

where

$$U_S = \lim F_S'\Omega_S^{-1}F_S, \quad \tau_S = \lim n^{\frac{1}{2}-\zeta}F_S'\Omega_S^{-1}\Xi_S\eta$$

$F_S = \Xi_S F$, $\Omega_S = \Xi_S\Omega\Xi_S'$, $F = E[z_i z_i']\Pi$ *and* $\Omega = E[\epsilon_i^2 z_i z_i']$.

To implement the GFIC, say FMSC for this example, we simply need to specialize Equation 1. To simplify the notation, let

$$\Xi_1 = \left[\begin{array}{cc} I_p & 0_{p\times q}\end{array}\right], \quad \Xi_2 = \left[\begin{array}{cc} 0_{q\times p} & I_q\end{array}\right] \tag{6}$$

where $0_{m\times n}$ denotes an $m \times n$ matrix of zeros and I_m denotes the $m \times m$ identity matrix. Using this convention, $Z_1 = Z\Xi_1'$ and $Z_2 = Z\Xi_2'$. In this example the valid estimator $\widehat{\beta}_v$, defined in Corollary 2.1, is given by only using instruments included in Z_1. We estimate $\nabla_\beta\mu(\beta)$ with $\nabla_\beta\mu(\widehat{\beta}_v)$. Similarly,

$$\widehat{U}_S = (X'Z\Xi_S'(\Xi_S\widehat{\Omega}(\widehat{\beta}_v)\Xi_S')^{-1}\Xi_S Z'X)/n^2$$

consistently estimates U_S in the instrument selection example. Since Ξ_S is known, the only remaining quantities from Equation 1 are $\widehat{\tau}_S$. The following result specializes Theorem 3.1 to the present example.

Theorem 4.2. *Let* $\widehat{\tau}_S = n^{-3/2}X'Z\Xi_S'(\Xi_S\widehat{\Omega}(\widehat{\beta}_v)\Xi_S')^{-1}\Xi_S Z'(y - X\widehat{\beta}_v)$. *Under the conditions of Theorem 4.1 we have*

$$\widehat{\tau}_S \to_d N(\tau_S, U_S U_v^{-1} U_S - U_S)$$

where τ_S and U_S are defined in Theorem 4.1.

Directly from the theorem, an asymptotically unbiased estimator of $\tau_S\tau_S'$ can be constructed as $\widehat{\tau}_S\widehat{\tau}_S' - \widehat{U}_S\widehat{U}_v^{-1}\widehat{U}_S + \widehat{U}_S$.

Now every value in Equation 1 is properly estimated, and we can conduct the GFIC moment selection which minimizes the estimated AMSE. One thing to note is that, in practice, one might want to employ more robust estimates of the covariance matrix Ω which takes into consideration the local mis-specification framework. That is, to account for the finite-sample bias and allow more robustness, one could employ the centered heteroskedasticity-consistent estimator:

$$\widehat{\Omega}_S = \frac{1}{n}\sum_{i=1}^n u_i(\widehat{\beta}_S)^2 z_{iS}z_{iS}' - \left(\frac{1}{n}\sum_{i=1}^n u_i(\widehat{\beta}_S)z_{iS}\right)\left(\frac{1}{n}\sum_{i=1}^n u_i(\widehat{\beta}_S)z_{iS}'\right) \tag{7}$$

where $u_i(\beta) = y_i - x_i'\beta$, $\widehat{\beta}_S$ is the GEL estimator using instruments z_{iS}, where $z_{iS} = \Xi_S z_i$.

[3] Sufficient conditions for the linear case are similar to the ones in Newey and Windmeijer (2009).

4.2 Choosing Instrumental Variables Simulation

Following the setup in Section 4.1, I evaluate the performance of the FMSC (MSE) with the linear instrument selection example. For this section, I focus on the CUE (Continuous Updating Estimator) out of the GEL class. To be specific, if function ρ is quadratic, then GEL coincides with CUE. In order to better incorporate the many-instrument design, I employ a conditionally valid instrument, whose effect can be approximated using a spline or power series. As the number of splines or power series increases with sample size, we will have the same many-instrument structure discussed above. The simulation design is as follows:

$$\begin{aligned} y_i &= \theta_0 x_i + \epsilon_i, \quad \theta_0 = 0.5 && (8)\\ x_i &= 0.2\exp(z_i) + \gamma w_i + v_i && (9)\\ 0 &= \mathbb{E}(\epsilon_i | z_i) && (10) \end{aligned}$$

for $i = 1, 2, \ldots, n$. The econometrician is aware of the conditionally valid instrument z_i but unaware of how it comes into the formulation of x_i. Therefore, they could utilize a spline or power series $s_k(z_i),\ k = 1, \ldots, p_n$ as instruments. As for the potentially endogenous instrument w_i, the econometrician needs to determine whether or not to include it in the estimation.

To complete the data generating process, I let $(\epsilon_i, v_i, w_i, z_i)' \sim$ iid $N(0, \mathcal{V})$ with

$$\mathcal{V} = \begin{bmatrix} \mathcal{V}_1 & 0 \\ 0 & \mathcal{V}_2 \end{bmatrix}, \quad \mathcal{V}_1 = \begin{bmatrix} 1 & (0.5 - \gamma\rho) & \rho \\ (0.5 - \gamma\rho) & (1 - 0.04e^2 + 0.04e - \gamma^2) & 0 \\ \rho & 0 & 1 \end{bmatrix}, \quad \mathcal{V}_2 = 1 \qquad (11)$$

This setup keeps the variance of x fixed at one and the endogeneity of x, $Cor(x, \epsilon)$, fixed at 0.5 while allowing the relevance, $\gamma = Cor(x, w)$, and endogeneity, $\rho = Cor(w, \epsilon)$, of the instrument w to vary. Any function (I will use spline series in below simulation) of z will be valid and exogenous instruments due to the zero conditional expectation of ϵ. The additional instrument w is only relevant if $\gamma \neq 0$ and is only exogenous if $\rho = 0$. Since x has unit variance, with large enough p_n, the first-stage R-squared for this simulation design is approaching $1 - \sigma_v^2 \approx 0.18 + \gamma^2$ from below. Hence, when $\gamma = 0$, so that w is irrelevant, the first-stage R-squared is smaller than 18%. Increasing γ increases the R-squared of the first-stage. This design satisfies the sufficient conditions for Theorem 4.1.

The goal is to estimate the effect of x on y, in this case 0.5, with minimum MSE by choosing the optimal set of instruments: the *valid* estimator uses only $\{s_k(z)\}_{k=1}^p$ as instruments, while the *full* estimator uses $\{s_k(z)\}_{k=1}^p$ and w. The inclusion of $\{s_k(z)\}_{k=1}^p$ in both moment sets ensures that the fixed-sample MSE is well-defined. All estimators in this section are calculated via CUE and 5,000 simulation replications.

Figures 1-3 present respectively the relative RMSE values for different estimators when $p = 5$, $p = 10$, and $p = 15$. In each figure, I show the relative RMSE of the valid estimator, the full estimator, and the post-FMSC estimator for various combinations of γ, ρ, and n. The overall results are similar with a relatively small or large number of moments, which is consistent with the theoretical findings: FMSE works well in a many moment setup. For any

combination (γ, n) there is a positive value of ρ below which the full estimator yields a lower RMSE than the valid estimator. As the sample size increases, this cutoff becomes smaller; as γ increases, it becomes larger. The post-FMSC estimator represents a compromise between the two estimators over which the FMSC selects. When the RMSE of valid IV estimator is high, the FMSC behaves more like full IV estimator; and vice versa. Because the FMSC is itself a random variable, however, it sometimes makes moment selection mistakes. For this reason its RMSE curve cannot attain the lower envelope of the valid estimator or the full estimator. Nevertheless, the larger the RMSE difference between the valid estimator and the full estimator, the closer the FMSC comes to this lower envelope: costly mistakes are rare. The FMSC is specifically intended for situations in which an applied researcher fears that their "baseline" assumptions may be too weak and consequently considers adding one or more "controversial" assumptions. In this case, one believes that the exogenous instruments, spline or power series of z, are not particularly strong and ρ is small relative to n, and thus entertains the assumption that w is exogenous. It is precisely in these situations that the FMSC can provide large performance gains over valid estimator.

Another interesting finding is that sometimes the post-FMSC estimator outperforms both the valid and the full estimators in terms of RMSE, especially when n is small. This is because the valid estimator can behave erratically with a small sample size. As the first-stage R-squared never exceeds 18%, the explanatory power of the valid instruments is not particularly strong. So with an option of including a strong but potentially endogenous instrument, FMSC manages to yield a lower RMSE than both of the other two estimators.

I now compare the FMSC to some existing moment selection criteria in GEL estimation, mainly the adapted AIC (Akaike's Information Criterion) and BIC (Bayesian Information Criterion) criteria. Both take the form $2n\hat{Q}_S - (|S| - r)c_n$, where $\hat{Q}_S$ is the objective of GEL estimation under moment set S and c_n is the penalty harshness. $c_n = 2$ gives an analogue of AIC while $c_n = \log n$ gives an analogue of BIC. Just like in likelihood or standard GMM case, the BIC is consistent while the AIC, similar to FMSC, is conservative. Figures 4-6 show the relative RMSE values of the post-FMSC estimator alongside those of the post-AIC and BIC estimators, respectively for the cases where $p = 5$, $p = 10$, and $p = 15$. Again, similar patterns can be found with either a small or large number of moments. For small sample sizes, the AIC and BIC are quite erratic, which comes from the fact that the GEL objective can be very badly behaved in small samples. In this case, FMSC might even uniformly outperform the information criteria. As the sample size becomes larger, the classic tradeoff between consistent and conservative selection emerges. For the smallest values of ρ, the consistent criteria outperform the conservative criteria; for moderate values the situation is reversed. The consistent criteria, however, have the highest worst-case RMSE.

Finally, I compare the GEL (CUE) estimator with the standard efficient GMM estimator, both after the FMSC selection. In a standard case where the moment number is fixed, GEL and GMM should be asymptotically equivalent. However, just as Newey and Smith (2004) states, the GMM estimator has larger higher-order bias compared to the GEL estimator. This issue is particularly exacerbated with a large number of moments. Figure 7 confirms their results[4]. The bias of post-selection efficient GMM estimator is typically much larger than the CUE estimator, and the pattern is more apparent when $p = 15$. This larger bias is

[4]Results are similar for other combinations of n, γ and p, which are not shown here due to space limitations

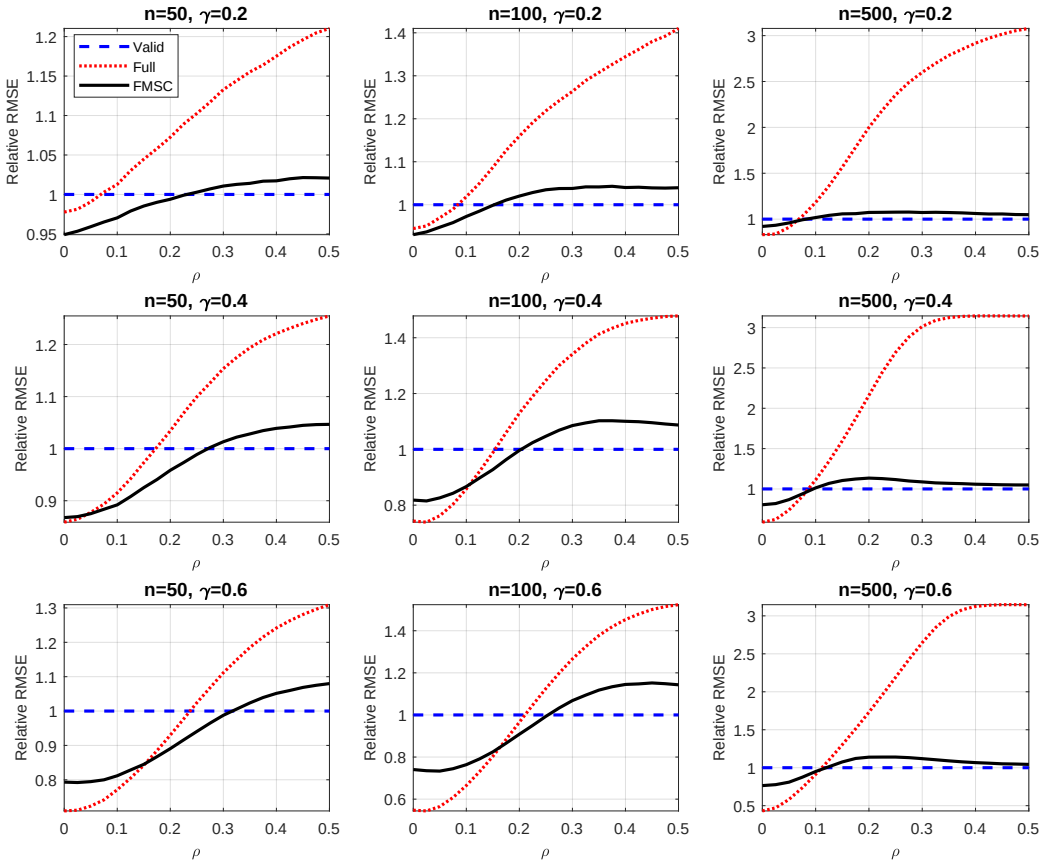

Figure 1: Relative RMSE values for the valid CUE estimator (baseline), where the valid moment number $p = 5$, the full CUE estimator, and the post-FMSC CUE estimator based on 5,000 simulation draws from the DGP given in Equations 8–11

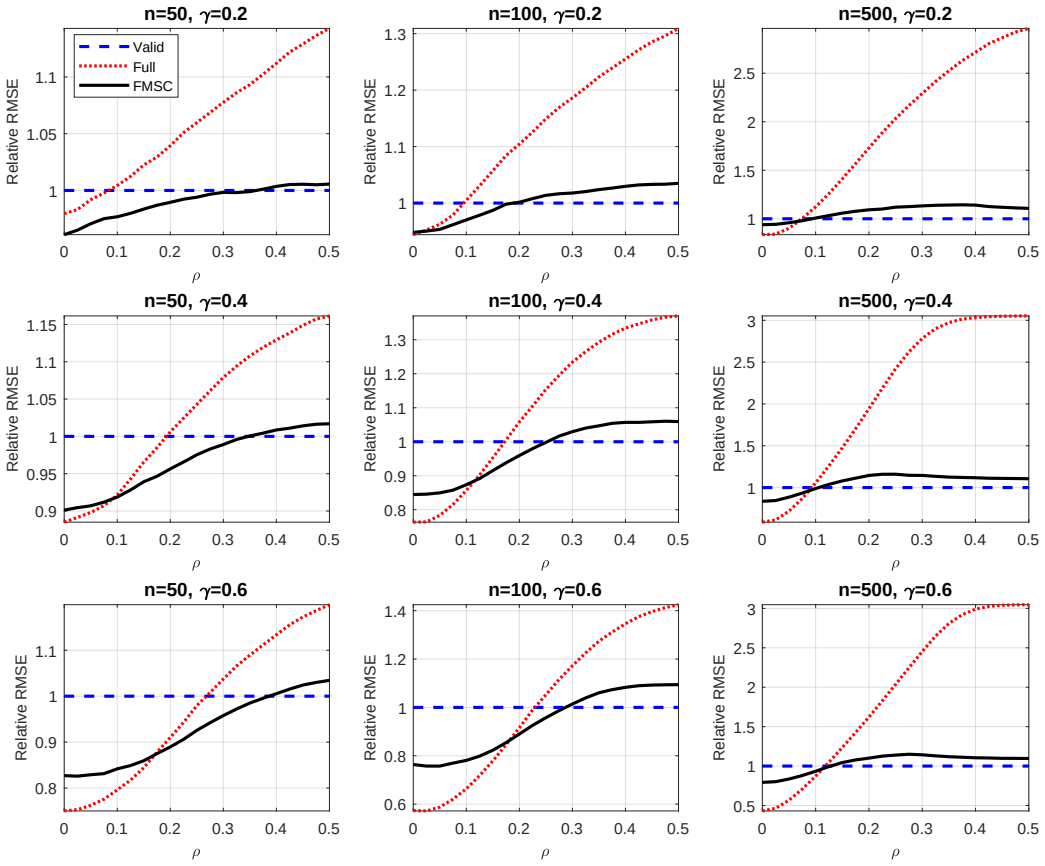

Figure 2: Relative RMSE values for the valid CUE estimator (baseline), where the valid moment number $p = 10$, the full CUE estimator, and the post-FMSC CUE estimator based on 5,000 simulation draws from the DGP given in Equations 8–11

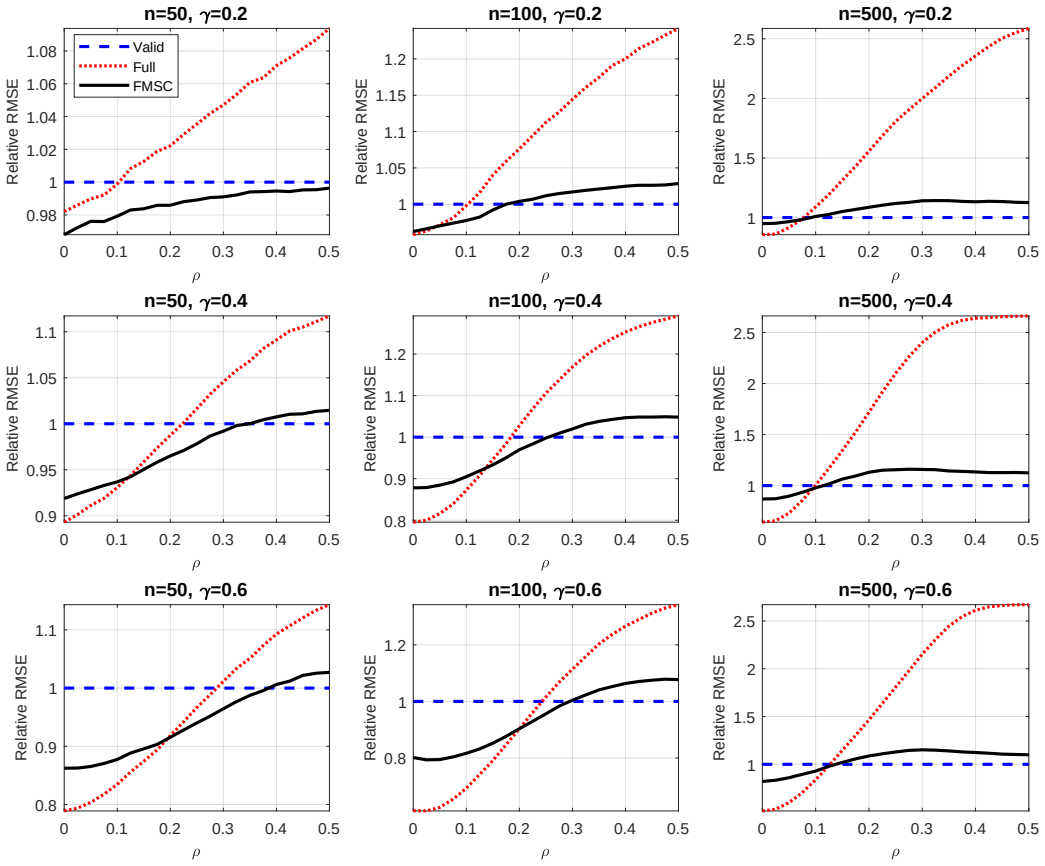

Figure 3: Relative RMSE values for the valid CUE estimator (baseline), where the valid moment number $p = 15$, the full CUE estimator, and the post-FMSC CUE estimator based on 5,000 simulation draws from the DGP given in Equations 8–11

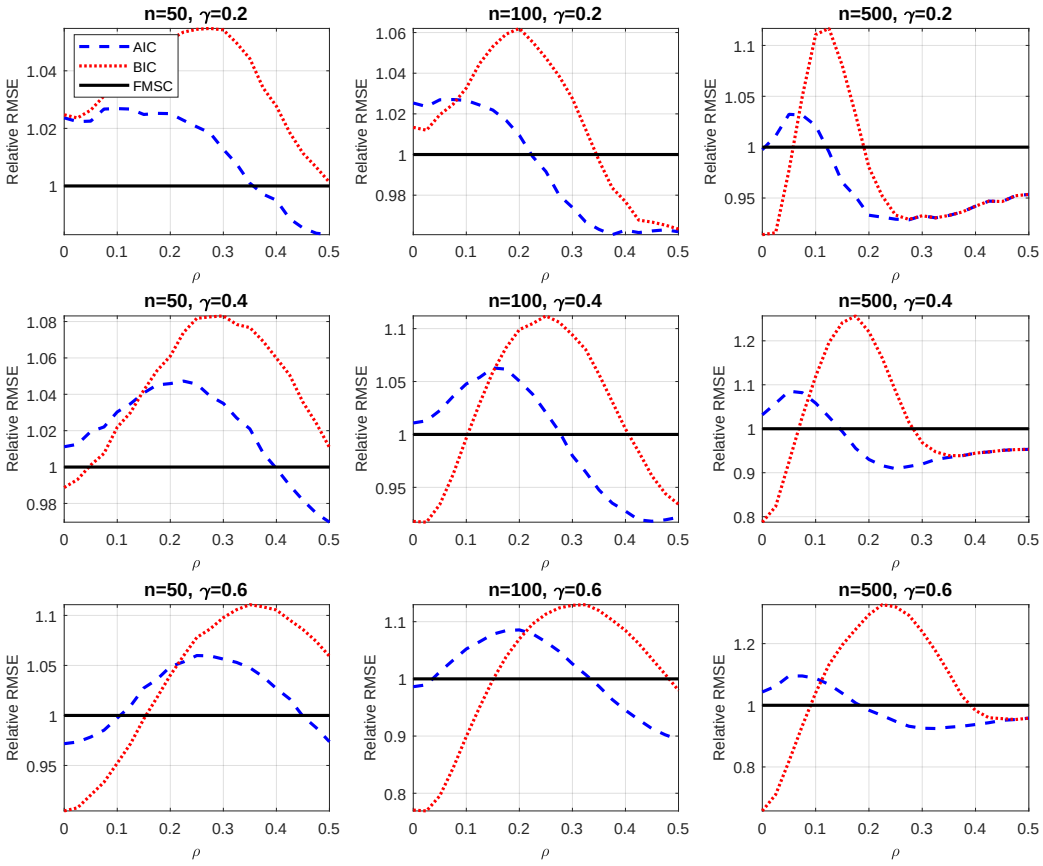

Figure 4: Relative RMSE values for the post-FMSC estimator and the AIC/BIC estimators based on 5,000 simulation draws from the DGP given in Equations 8–11, where the valid moment number $p = 5$

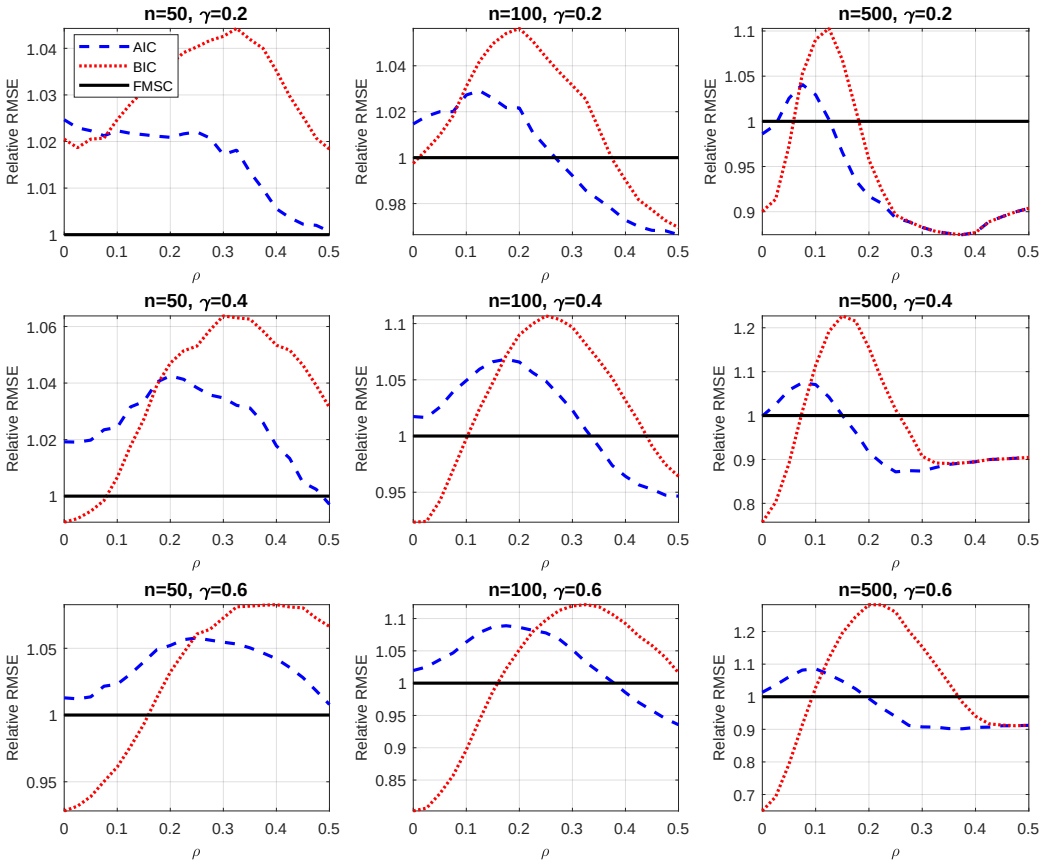

Figure 5: Relative RMSE values for the post-FMSC estimator and the AIC/BIC estimators based on 5,000 simulation draws from the DGP given in Equations 8–11, where the valid moment number $p = 10$

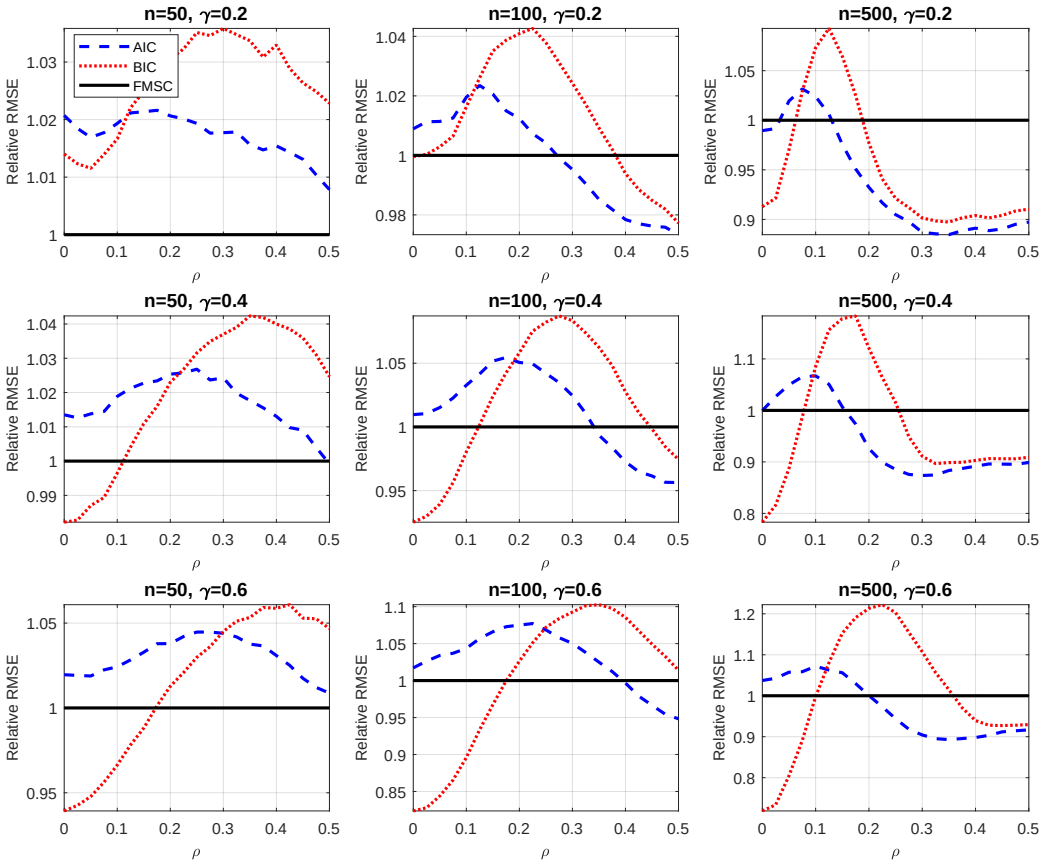

Figure 6: Relative RMSE values for the post-FMSC estimator and the AIC/BIC estimators based on 5,000 simulation draws from the DGP given in Equations 8–11, where the valid moment number $p = 15$

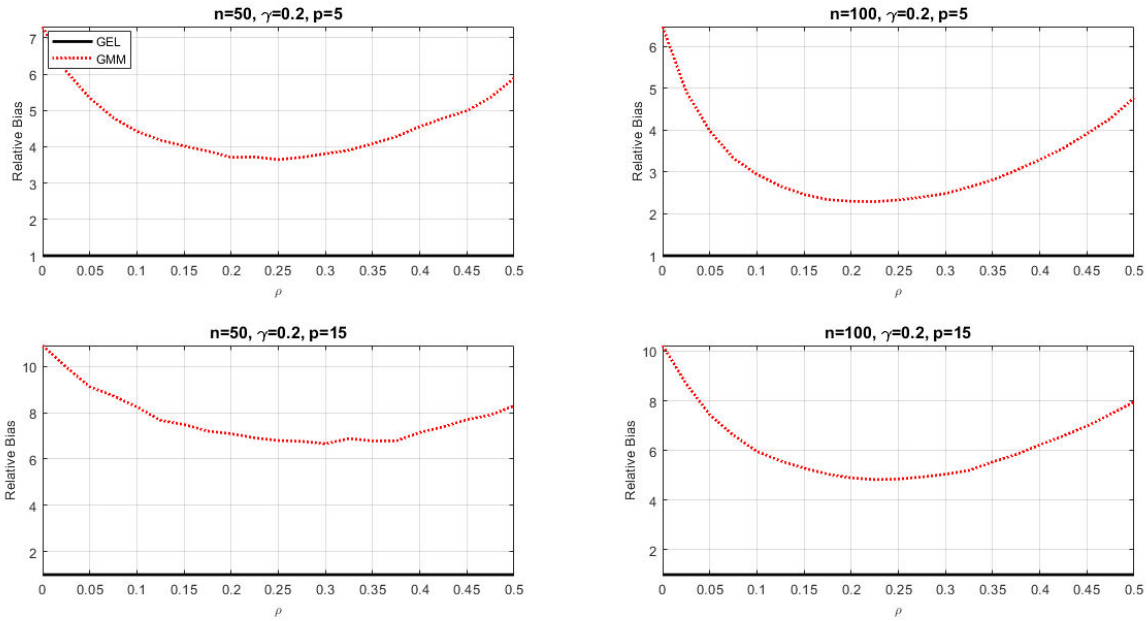

Figure 7: Relative bias values for the post-FMSC CUE estimator and the post-FMSC efficient GMM estimator based on 5,000 simulation draws from the DGP given in Equations 8–11

also carried into the larger MSE of the GMM estimator, which is not shown in the figures here.

5 Conclusion

This paper introduces the GFICs based on GEL estimators to choose moment conditions using asymptotic risks, which balances between the validity of the moments and extra information possibly provided. I allow the number of moments to grow with the sample size to incorporate the many moment structure. The criteria perform well in the simulation and it is suitable for a wide range of estimators classified in the GEL class, which can be an improvement upon the standard GMM estimator especially in terms of bias. This procedure remains feasible for problems of a realistic scale without the need for specialized computing resources and can yield sizeable benefits in empirically relevant settings, making them a valuable complement to existing methods. A potential extension could consider a dynamic panel setting and apply the GFICs in a way that simultaneous moments can be addressed. Other possibility includes an asymptotic framework which allows for weak identification as seen in Newey and Windmeijer (2009).

A Proofs

Let C stands for a generic positive constant that may be different in different uses. M, T, CS, and CLT will be the Markov inequality, Triangular inequality, Cauchy-Schwarz inequality, and the Lindeberg-Levy central limit theorem.

First I prove Assumptions 2.1, 2.3 and 2.4 are satisfied with g and f replaced by any f_S, $S \in \mathcal{S} = \{S : v \subset S\}$. Let $\Xi_{v,S}$ be the selection matrix so that $g = \Xi_{v,S} f_S$. Note $||\Xi_S|| \leq 1$ and $||\Xi_{v,S}|| \leq 1$.

For Assumption 2.1, $\mathbb{E}[f_{S,i}(\theta_0)] = \dfrac{\Xi_S \eta}{n^\zeta} = \dfrac{\eta_S}{n^\zeta}$. There exist C and $\widehat{D} = O_p(1)$ such that: i) $||\delta|| \leq C||\bar{g}(\theta)|| \leq C||\Xi_{v,S}||\ ||\bar{f}_S(\theta)|| \leq C||\bar{f}_S(\theta)||$; ii) w.p.a.1, $||\delta|| \leq C||\widehat{g}(\theta)|| + \widehat{D} \leq C||\widehat{f}_S(\theta)|| + \widehat{D}$; iii) $\max|\eta_{S,l}| \leq \max|\eta_l| \leq C$.

For Assumption 2.3, $|S|/n \to_p 0$. f_S is continuous in θ and there is $C > 0$ such that: i) $\sup_{\theta\in\Theta} \mathbb{E}[\{f_{S,i}(\theta)' f_{S,i}(\theta)\}^2]/n^3 \leq \sup_{\theta\in\Theta} \mathbb{E}[\{f_i(\theta)' f_i(\theta)\}^2]/n^3 \to_p 0$; ii) $\sup_{\theta\in\Theta} ||\widehat{\Omega}_S(\theta) - \Omega_S(\theta)|| \leq \sup_{\theta\in\Theta} ||\Xi_S||^2 ||\widehat{\Omega}(\theta) - \Omega(\theta)|| \to_p 0$; v) for every $\widetilde{C}$, there is C and $\widehat{D} = O_p(1)$ such that for all $\widetilde{\theta}, \theta \in \Theta$, $||\delta(\widetilde{\theta})|| \leq \widetilde{C}$, $||\delta(\theta)|| \leq \widetilde{C}$, we have $||\bar{f}_S(\widetilde{\theta}) - \bar{f}_S(\theta)|| \leq ||\bar{f}(\widetilde{\theta}) - \bar{f}(\theta)|| \leq C||\widetilde{\theta} - \theta||$ and $||\widehat{f}_S(\widetilde{\theta}) - \widehat{f}_S(\theta)|| \leq ||\widehat{f}(\widetilde{\theta}) - \widehat{f}(\theta)|| \leq \widehat{D}||\widetilde{\theta} - \theta||$; vi) $\{n\mathbb{E}[\sup_{\theta\in\Theta} ||f_{S,i}(\theta)||^\alpha]\}^{1/\alpha}\sqrt{|S|/n} \leq \{n\mathbb{E}[\sup_{\theta\in\Theta} ||f_i(\theta)||^\alpha]\}^{1/\alpha}\sqrt{m/n} \to 0$ for some $\alpha > 2$.

By Assumption 2.3 iii), $\Omega(\theta)$ is positive definite. Then for $\Omega_S(\theta)$, which is a principal submatrix of $\Omega(\theta)$, we have $1/C \leq \xi_{\min}(\Omega(\theta)) \leq \xi_{\min}(\Omega_S(\theta)) \leq \xi_{\max}(\Omega_S(\theta)) \leq \xi_{\max}(\Omega(\theta)) \leq C$.

Regarding Assumption 2.3 iv), for all $a, b \in R^{|S|}$, let $\widetilde{a} = \Xi_S' a$ and $\widetilde{b} = \Xi_S' b$. Then $||\widetilde{a}|| = ||a||$ and $||\widetilde{b}|| = ||b||$. $|a'[\Omega_S(\widetilde{\theta}) - \Omega_S(\theta)]b| = |\widetilde{a}'[\Omega(\widetilde{\theta}) - \Omega(\theta)]\widetilde{b}| \leq ||\widetilde{a}||\ ||\widetilde{b}||\ ||\widetilde{\theta} - \theta|| = ||a||\ ||b||\ ||\widetilde{\theta} - \theta||$.

For Assumption 2.4, $|S|^3/n \to 0$. f_S is twice continuously differentiable in a neighborhood $\mathcal{N}$ of θ_0 and there is $C > 0$ and $\widehat{D} = O_p(1)$ such that: i) $\xi_{\max}(\mathbb{E}[F_{S,i}(\theta_0) F_{S,i}(\theta_0)']) \leq \xi_{\max}(\mathbb{E}[F_i(\theta_0) F_i(\theta_0)']) \leq C$; ii) if $\widetilde{\theta} \to_p \theta_0$, then $||\widehat{F}_S(\widetilde{\theta}) - \widehat{F}_S(\theta_0)|| \leq ||\widehat{F}(\widetilde{\theta}) - \widehat{F}(\theta_0)|| \leq \widehat{D}||\widetilde{\theta} - \theta_0||$; iii) $\mathbb{E}[\sup_{\theta\in\mathcal{N}} ||f_{S,i}(\theta)||^4]\sqrt{|S|/n} \leq \mathbb{E}[\sup_{\theta\in\mathcal{N}} ||f_i(\theta)||^4]\sqrt{m/n} \leq C$.

Given above, I only need to show the proof of Theorem 2.1 and 2.2 for $\Xi_S = I_m$.

Lemma A.1. *Under Assumption 2.3, for any $\widehat{\Omega}_n = \widehat{\Omega}(\theta_n), \theta_n \in \Theta$, $1/C \leq \xi_{\min}(\widehat{\Omega}_n) \leq \xi_{\max}(\widehat{\Omega}_n) \leq C$ w.p.a.1.*

Proof.

$$||\widehat{\Omega}_n - \Omega(\theta_n)|| \leq \sup_{\theta\in\Theta} ||\widehat{\Omega}(\theta) - \Omega(\theta)|| \to_p 0$$

Then, by $|\xi(A) - \xi(B)| \leq ||A - B||$, where $\xi(A)$ denotes the minimum or maximum eigenvalue, it follows that $1/C \leq \xi_{\min}(\widehat{\Omega}_n) \leq \xi_{\max}(\widehat{\Omega}_n) \leq C$ w.p.a.1. □

Lemma A.2. *(X_i, Y_i) are i.i.d and $\dim(X_i) = \dim(Y_i) = m$. Let $\bar{X} = \sum_i X_i/n, \bar{Y} = \sum_i Y_i/n, \mu_X = \mathbb{E}[X_i]$, $\mu_Y = \mathbb{E}[Y_i]$, $\Sigma_{XX} = \mathbb{E}[X_i X_i'], \Sigma_{XY} = \mathbb{E}[X_i Y_i'], \Sigma_{YY} = \mathbb{E}[Y_i Y_i']$. If $\max\{\xi_{\max}(AA'), \xi_{\max}(A'A), \xi_{\max}(\Sigma_{XX}), \xi_{\max}(\Sigma_{YY})\} \leq C$, $m/a_n^2 \to 0$, $a_n/n \leq C$, $\mathbb{E}[(X_i'X_i)^2]/na_n^2 \to 0$, $\mathbb{E}[(Y_i'Y_i)^2]/na_n^2 \to 0$, $n\mu_X'\mu_X/a_n^2 \to 0$, $n\mu_Y'\mu_Y/a_n^2 \to 0$, then*

$$n\bar{X}'A\bar{Y}/a_n = tr(A\Sigma_{XY}')/a_n + n\mu_X' A\mu_Y/a_n + o_p(1)$$

Proof. Let $W_i = AY_i$. Then $\Sigma'_{XW} = A\Sigma'_{XY}$ and $\mu_W = A\mu_Y$.

$$\begin{aligned}
&\xi_{\max}(\mathbb{E}[W_i W_i']) = \xi_{\max}(A\Sigma_{YY}A') \leq C \\
&\mathbb{E}[(W_i'W_i)^2]/na_n^2 = \mathbb{E}[(Y_i'A'AY_i)^2]/na_n^2 \leq C\mathbb{E}[(Y_i'Y_i)^2]/na_n^2 \to 0
\end{aligned}$$

Thus the hypotheses and conclusion are satisfied with W in place of Y and $A = I$. Therefore, it suffices to show the result with $A = I$.

Note that:

$$\begin{aligned}
&\mathbb{E}[(X_i'Y_i)^2] \leq \mathbb{E}[(X_i'X_i)^2] + \mathbb{E}[(Y_i'Y_i)^2] \\
&\mathbb{E}[X_i'Y_jY_j'X_i] = \mathbb{E}[X_i'\Sigma_{YY}X_i] \leq C\mathbb{E}[X_i'X_i] = C\mathrm{tr}(\Sigma_{XX}) \leq Cm \\
&|\mathbb{E}[X_i'Y_jX_j'Y_i]| \leq C(\mathbb{E}[X_i'Y_jY_j'X_i] + \mathbb{E}[X_j'Y_iY_i'X_j]) \leq Cm
\end{aligned}$$

Now let $U_i = X_i - \mu_X$, $V_i = Y_i - \mu_Y$. Then $\xi_{\max}(\Sigma_{UU}) = \xi_{\max}(Var[X_i]) \leq \xi_{\max}(\Sigma_{XX})$ and $\mathbb{E}[(U_i'U_i)^2] \leq \mathbb{E}[(X_i'X_i)^2]$. So U, V also satisfy above inequalities.

Let $S_n = n\bar{X}'\bar{Y}/a_n, \widetilde{S}_n = n\bar{U}'\bar{V}/a_n$.

$$\begin{aligned}
S_n =& n(\bar{U} + \mu_X)'(\bar{V} + \mu_Y)/a_n \\
=& \widetilde{S}_n + n\bar{U}'\mu_Y/a_n + n\mu_X'\bar{V}/a_n + n\mu_X'\mu_Y/a_n
\end{aligned}$$

Note

$$\begin{aligned}
\mathbb{E}[\widetilde{S}_n^2]/n \leq& \mathbb{E}[(U_i'V_i)^2]/na_n^2 \leq (\mathbb{E}[(U_i'U_i)^2] + \mathbb{E}[(V_i'V_i)^2])/na_n^2 \to 0 \\
\mathbb{E}[\widetilde{S}_n^2] =& \mathbb{E}[\sum_{i,j,k,l}(U_i'V_jU_k'V_l)^2]/n^2a_n^2 \\
=& \mathbb{E}[(U_i'V_i)^2]/na_n^2 + (1 - \frac{1}{n})\{\mathbb{E}[\widetilde{S}_n]^2 + \mathbb{E}[U_i'V_jV_j'U_i]/a_n^2 + \mathbb{E}[U_i'V_jU_j'V_i]/a_n^2\} \\
=& \mathbb{E}[\widetilde{S}_n]^2 + o(1) \\
\mathbb{E}[(n\mu_X'\bar{V}/a_n)^2] =& n\mu_X'(\Sigma_{YY} - \mu_Y\mu_Y')\mu_X/a_n^2 \leq n\mu_X'\Sigma_{YY}\mu_X/a_n^2 \leq Cn\mu_X'\mu_X/a_n^2 \to 0
\end{aligned}$$

Then by M,

$$\begin{aligned}
S_n =& \widetilde{S}_n + n\mu_X'\mu_Y/a_n + o_p(1) = \mathrm{tr}(\Sigma'_{UV})/a_n + n\mu_X'\mu_Y/a_n + o_p(1) \\
=& \mathrm{tr}(\Sigma'_{XY})/a_n + \mathrm{tr}(-\mu_X\mu_Y'/a_n) + n\mu_X'\mu_Y/a_n + o_p(1) \\
=& \mathrm{tr}(\Sigma'_{XY})/a_n - a_n/n \cdot (n\mu_X'\mu_Y/a_n^2) + n\mu_X'\mu_Y/a_n + o_p(1) \\
=& \mathrm{tr}(\Sigma'_{XY})/a_n + n\mu_X'\mu_Y/a_n + o_p(1)
\end{aligned}$$

□

Lemma A.3. *Under Assumption 2.1 and 2.3, for any $C > 0$, $\sup_{\theta\in\Theta, ||\theta-\theta_0||\leq C} |\widehat{Q}^*(\theta) - Q(\theta)| \to_p 0$, where $\widehat{Q}^*(\theta) = \widehat{f}(\theta)'\widehat{\Omega}(\theta)^{-1}\widehat{f}(\theta)/2$, $Q(\theta) = \bar{f}(\theta)'\Omega(\theta)^{-1}\bar{f}(\theta)/2 + m/2n$.*

Proof. Let $\delta = \theta - \theta_0$. Note that $\mathbb{E}[||\widehat{f}(\theta_0)||^2] \leq \frac{\text{tr}(\Omega(\theta_0))}{n} + C\frac{\text{tr}(\eta\eta')}{n^{2\zeta}} \leq Cm/n$, so by M, $||\widehat{f}(\theta_0)|| = O_p(\sqrt{m/n})$. Then by Assumption 2.3 v) and T,

$$\sup_{||\delta||\leq C} ||\widehat{f}(\theta)|| \leq ||\widehat{f}(\theta_0)|| + \sup_{||\delta||\leq C} ||\widehat{f}(\theta) - \widehat{f}(\theta_0)|| = O_p(1)$$

Let $\widehat{a}(\theta) = \Omega(\theta)^{-1}\widehat{f}(\theta)$ and $\widetilde{Q}(\theta) = \widehat{f}(\theta)'\Omega(\theta)^{-1}\widehat{f}(\theta)/2$. Then $||\widehat{a}(\theta)|| \leq C||\widehat{f}(\theta)||$, so $\sup_{||\delta||\leq C} ||\widehat{a}(\theta)|| = O_p(1)$.
By T, Assumption 2.3, and Lemma A.1,

$$\begin{aligned}\sup_{||\delta||\leq C} |\widehat{Q}^*(\theta) - \widetilde{Q}(\theta)| &\leq \sup_{||\delta||\leq C} |\widehat{a}(\theta)'[\widehat{\Omega}(\theta) - \Omega(\theta)]\widehat{a}(\theta)| + \sup_{||\delta||\leq C} |\widehat{a}(\theta)'[\widehat{\Omega}(\theta) - \Omega(\theta)]\widehat{\Omega}(\theta)^{-1}[\widehat{\Omega}(\theta) - \Omega(\theta)]\widehat{a}(\theta)| \\ &\leq \sup_{||\delta||\leq C} ||\widehat{a}(\theta)||^2 \sup_{||\delta||\leq C} (||\widehat{\Omega}(\theta) - \Omega(\theta)|| + C||\widehat{\Omega}(\theta) - \Omega(\theta)||^2) \to_p 0\end{aligned}$$

Next let $a(\widetilde{\theta}, \theta) = \Omega(\theta)^{-1}\bar{f}(\widetilde{\theta})$ and $Q(\widetilde{\theta}, \theta) = \bar{f}(\widetilde{\theta})'\Omega(\theta)^{-1}\bar{f}(\widetilde{\theta})/2 + m/2n$. Then by Assumption 2.3 iii) and v), $\sup_{||\delta||\leq C, ||\widetilde{\delta}||\leq C} ||a(\widetilde{\theta}, \theta)|| \leq C$. By T, Assumption 2.3 iv) and v),

$$\begin{aligned}|Q(\widetilde{\theta}) - Q(\theta)| \leq& |Q(\widetilde{\theta}, \widetilde{\theta}) - Q(\widetilde{\theta}, \theta)| + |Q(\widetilde{\theta}, \theta) - Q(\theta, \theta)| \\ \leq& |a(\widetilde{\theta}, \widetilde{\theta})'[\Omega(\widetilde{\theta}) - \Omega(\theta)]a(\widetilde{\theta}, \theta)| + C(||\bar{f}(\widetilde{\theta}) - \bar{f}(\theta)||^2 + 2||\bar{f}(\theta)||||\bar{f}(\widetilde{\theta}) - \bar{f}(\theta)||) \\ \leq& C||\widetilde{\theta} - \theta||\end{aligned}$$

Therefore $Q(\theta)$ is equicontinuous on $||\delta|| \leq C$. Similarly, replacing $a(\widetilde{\theta}, \theta)$ and $Q(\widetilde{\theta}, \theta)$ with $\widehat{a}(\widetilde{\theta}, \theta) = \Omega(\theta)^{-1}\widehat{f}(\widetilde{\theta})$ and $\widetilde{Q}(\widetilde{\theta}, \theta) = \widehat{f}(\widetilde{\theta})'\Omega(\theta)^{-1}\widehat{f}(\widetilde{\theta})/2$, we have $|\widetilde{Q}(\widetilde{\theta}) - \widetilde{Q}(\theta)| \leq \widehat{D}(\widetilde{\theta} - \theta)$ for $\widehat{D} = O_p(1)$. So $\widetilde{Q}(\theta)$ is stochastic equicontinuous on $||\delta|| \leq C$.

Now, apply Lemma A.2 with $X_i = Y_i = f_i(\theta)$, $A = \Omega(\theta)^{-1}$ and $a_n = n$. By Assumption 2.3 iii), $\xi_{\max}(AA') = \xi_{\max}(A'A) = \xi_{\max}(\Omega(\theta)^{-2}) \leq C, \xi_{\max}(\Sigma_{XX}) = \xi_{\max}(\Sigma_{YY}) = \xi_{\max}(\Omega(\theta)^{-1}) \leq C$. By Assumption 2.3 i), $\mathbb{E}[(X_i'X_i)^2]/na_n^2 = \mathbb{E}[(Y_i'Y_i)^2]/na_n^2 = \mathbb{E}[\{f_i(\theta)'f_i(\theta)\}^2]/n^3 \to 0$. Next $Q(0) = \frac{\eta}{n^\zeta}'\Omega^{-1}\frac{\eta}{n^\zeta}/2 + m/2n \leq Cm/n$. By equicontinuity of $Q(\theta)$, $n\mu_X'\mu_X/a_n^2 = n\mu_Y'\mu_Y/a_n^2 \leq C\bar{f}(\theta)\Omega(\theta)^{-1}\bar{f}(\theta)/n = C(2Q(\theta) - m/n)/n \to 0$. Thus, the conditions of Lemma A.2 are satisfied. Note that $A\Sigma_{XY}' = I_m$, so by the conclusion of Lemma A.2

$$2\widetilde{Q}(\theta) = m/n + \bar{f}(\theta)\Omega(\theta)^{-1}\bar{f}(\theta) + o_p(1) = 2Q(\theta) + o_p(1)$$

Finally, given $\sup_{||\delta||\leq C} |\widehat{Q}^*(\theta) - \widetilde{Q}(\theta)| \to_p 0$, $\widetilde{Q}(\theta) = Q(\theta) + o_p(1)$ and stochastic equicontinuity of $\widetilde{Q}(\theta)$ and $Q(\theta)$, by Newey (1991, Theorem 2.1), we have $\sup_{||\delta||\leq C} |\widehat{Q}^*(\theta) - Q(\theta)| \to_p 0$. □

Lemma A.4. *Let $b_i = \sup_{\theta\in\Theta} ||f_i(\theta)||$. Then under Assumption 2.3 vi), for any $\Delta_n = o([n\mathbb{E}(b_i^\alpha)]^{-1/\alpha})$ and $\Lambda_n = \{\lambda : ||\lambda|| \leq \Delta_n\}$, we have $\max_{\theta\in\Theta, \lambda\in\Lambda_n, i} |\lambda' f_i(\theta)| \to_p 0$ and w.p.a.1 $\Lambda_n \subset \widehat{\Lambda}(\theta)$ for all $\theta \in \Theta$.*

Proof. By Assumption 2.3 vi) and M, $\max_i b_i = O_p([n\mathbb{E}(b_i^\alpha)]^{1/\alpha})$. Then

$$\max_{\theta\in\Theta, \lambda\in\Lambda_n, i} |\lambda' f_i(\theta)| \leq \Delta_n \max_i b_i \to_p 0$$

Given above conclusion, w.p.a.1, $\lambda' f_i(\theta) \in \mathcal{V}$ for all $\theta \in \Theta, \lambda \in \Lambda_n$ and i. □

Lemma A.5. *Under Assumption 2.3 and 2.2, for $\widetilde{\theta} \in \Theta$, if $||\widehat{f}(\widetilde{\theta})|| = O_p(\sqrt{m/n})$, then $\sup_{\lambda \in \widehat{\Lambda}(\widetilde{\theta})} \widehat{P}(\widetilde{\theta}, \lambda) \leq \rho(0) + O_p(m/n)$, $\widetilde{\lambda} = argmax_{\lambda \in \widehat{\Lambda}(\widetilde{\theta})} \widehat{P}(\widetilde{\theta}, \lambda)$ exists w.p.a.1 and $||\widetilde{\lambda}|| = O_p(\sqrt{m/n})$.*

Proof. Given Assumption 2.3 vi), choose Δ_n such that $\Delta_n = o([n\mathbb{E}(b_i^\alpha)]^{-1/\alpha})$ and $\sqrt{m/n} = o(\Delta_n)$. Then for $\Lambda_n = \{\lambda : ||\lambda|| \leq \Delta_n\}$, by Lemma A.4 and the fact that $\widehat{P}(\widetilde{\theta}, \lambda)$ is twice continuously differentiable on Λ_n, $\ddot{\lambda} = \text{argmax}_{\lambda \in \Lambda_n} \widehat{P}(\widetilde{\theta}, \lambda)$ exists w.p.a.1. Then by Lemma A.1, $\xi_{\min}(\widehat{\Omega}(\widetilde{\theta})) \geq C$ w.p.a.1. Furthermore, by Lemma A.4 and Assumption 2.2, for any $\dot{\lambda}$ on the line joining $\ddot{\lambda}$ and 0, w.p.a.1 $\max_i \rho''(\dot{\lambda}' f_i(\widetilde{\theta})) \leq -C$. By a Taylor expansion around $\lambda = 0$, w.p.a.1:

$$\begin{aligned}
\rho(0) =& \widehat{P}(\widetilde{\theta}, 0) \leq \widehat{P}(\widetilde{\theta}, \ddot{\lambda}) \\
=& \rho(0) - \ddot{\lambda}' \widehat{f}(\widetilde{\theta}) + \ddot{\lambda}' \left[\sum_i \rho''(\dot{\lambda}' f_i(\widetilde{\theta})) f_i(\widetilde{\theta}) f_i(\widetilde{\theta})' / n \right] \ddot{\lambda}/2 \\
\leq& \rho(0) + ||\ddot{\lambda}|| \cdot ||\widehat{f}(\widetilde{\theta})|| - C \ddot{\lambda}' \widehat{\Omega}(\widetilde{\theta}) \ddot{\lambda} \\
\leq& \rho(0) + ||\ddot{\lambda}|| \cdot ||\widehat{f}(\widetilde{\theta})|| - C||\ddot{\lambda}||^2
\end{aligned}$$

Rearranging the inequality, we have $||\ddot{\lambda}|| \leq ||\widehat{f}(\widetilde{\theta})||/C = O_p(\sqrt{m/n})$. So w.p.a.1 $||\ddot{\lambda}|| < \Delta_n$, i.e. $\ddot{\lambda} \in \text{int}(\Lambda_n)$. It then follows that $\partial \widehat{P}(\widetilde{\theta}, \ddot{\lambda})/\partial \lambda = 0$. Since $\Lambda_n \subset \widehat{\Lambda}(\widetilde{\theta})$, by concavity of $\widehat{P}(\widetilde{\theta}, \lambda)$ and convexity of $\widehat{\Lambda}(\widetilde{\theta})$, we have $\ddot{\lambda} = \widetilde{\lambda} = \text{argmax}_{\lambda \in \widehat{\Lambda}(\widetilde{\theta})} \widehat{P}(\widetilde{\theta}, \lambda)$ and $\widehat{P}(\widetilde{\theta}, \widetilde{\lambda}) = \max_{\lambda \in \widehat{\Lambda}(\widetilde{\theta})} \widehat{P}(\widetilde{\theta}, \lambda) \leq \rho(0) + ||\ddot{\lambda}|| \cdot ||\widehat{f}(\widetilde{\theta})|| - C||\ddot{\lambda}||^2 = \rho(0) + O_p(m/n)$. □

Lemma A.6. *Under Assumption 2.1 - 2.2, w.p.a.1, $\widehat{Q}(\widehat{\theta}) \leq \widehat{Q}(\theta_0) \leq \rho(0) + O_p(m/n)$.*

Proof. By the Assumption 2.1, 2.3 and M, the conditions for Lemma A.5 are satisfied with $\widetilde{\theta} = \theta_0$. So $\widehat{Q}(\theta_0) = \sup_{\lambda \in \widehat{\Lambda}(\theta_0)} \widehat{P}(\theta_0, \lambda) \leq \rho(0) + O_p(m/n)$. Then by definition of $\widehat{\theta}$, $\widehat{Q}(\widehat{\theta}) = \sup_{\lambda \in \widehat{\Lambda}(\widehat{\theta})} \widehat{P}(\widehat{\theta}, \lambda) \leq \sup_{\lambda \in \widehat{\Lambda}(\theta_0)} \widehat{P}(\theta_0, \lambda) = \widehat{Q}(\theta_0) \leq \rho(0) + O_p(m/n)$. □

Lemma A.7. *Under Assumption 2.1 - 2.2, $||\widehat{f}(\widehat{\theta})|| = O_p(\sqrt{m/n})$.*

Proof. Let $\widehat{f} = \widehat{f}(\widehat{\theta})$ and choose Δ_n such that $\Delta_n = o([n\mathbb{E}(b_i^\alpha)]^{-1/\alpha})$ and $\sqrt{m/n} = o(\Delta_n)$. $\Lambda_n = \{\lambda : ||\lambda|| \leq \Delta_n\}$. Let $\ddot{\lambda} = -\Delta_n \widehat{f}/||\widehat{f}||$, so that $\ddot{\lambda} \in \Lambda_n$ and $\ddot{\lambda}' \widehat{f} = -\Delta_n ||\widehat{f}||$. Then by Lemma A.1, $\xi_{\max}(\widehat{\Omega}(\widehat{\theta})) \leq C$ w.p.a.1. Furthermore, by Lemma A.4 and Assumption 2.2, for any $\dot{\lambda}$ on the line joining $\ddot{\lambda}$ and 0, w.p.a.1 $\max_i -\rho''(\dot{\lambda}' \widehat{f}) \leq C$. By a Taylor expansion around $\lambda = 0$, w.p.a.1:

$$\begin{aligned}
\widehat{P}(\widehat{\theta}, \ddot{\lambda}) =& \rho(0) - \ddot{\lambda}' \widehat{f} + \ddot{\lambda}' \left[\sum_i \rho''(\dot{\lambda}' f_i(\widehat{\theta})) f_i(\widehat{\theta}) f_i(\widehat{\theta})' / n \right] \ddot{\lambda}/2 \\
\geq& \rho(0) + \Delta_n ||\widehat{f}|| - C||\ddot{\lambda}||^2 \\
\geq& \rho(0) + \Delta_n ||\widehat{f}|| - C\Delta_n^2
\end{aligned}$$

Rearranging the inequality and applying Lemma A.6, we have $\Delta_n||\widehat{f}||-C\Delta_n^2 \leq O_p(m/n)$. Also, by the choice of Δ_n, $m/(n\Delta_n) = o(\sqrt{m/n}) = o(\Delta_n)$. So we have $||\widehat{f}|| \leq O_p(m/(n\Delta_n))+ C\Delta_n = O_p(\Delta_n)$.

Now for any $\epsilon_n \to 0$, consider $\bar{\lambda} = -\epsilon_n \widehat{f}$. Then $||\bar{\lambda}|| = o_p(\Delta_n)$, so that $\bar{\lambda} \in \Lambda_n \subset \widehat{\Lambda}(\widehat{\theta})$ w.p.a.1. It follows from above inequality and $\epsilon_n \to 0$ that when n is large enough,

$$\begin{aligned}\widehat{P}(\widehat{\theta}, \bar{\lambda}) \geq& \rho(0) - \bar{\lambda}'\widehat{f} - \frac{1}{2}||\bar{\lambda}||^2 \\ =& \rho(0) + ||\widehat{f}||^2(\epsilon_n^2 - \frac{1}{2}\epsilon_n^2) \\ >& \rho(0) + \frac{1}{2}||\widehat{f}||^2\epsilon_n^2\end{aligned}$$

Then by Lemma A.6, $||\widehat{f}||^2\epsilon_n = O_p(m/n)$. Since ϵ_n is any sequence converging to zero, we have $||\widehat{f}||^2 = O_p(m/n)$ □

Lemma A.8. *Under Assumption 2.1 - 2.2,* $\widehat{Q}^*(\widehat{\theta}) \leq \widehat{Q}^*(\theta_0) + o_p(m/n)$.

Proof. By Lemma A.7, we can apply Lemma A.5 with $\widetilde{\theta} = \widehat{\theta}$. So $\widehat{\lambda} = \text{argmax}_{\lambda \in \widehat{\Lambda}(\widehat{\theta})} \widehat{P}(\widehat{\theta}, \lambda)$ exists w.p.a.1 and $||\widehat{\lambda}|| = O_p(\sqrt{m/n})$. Because ρ is three times continuously differentiable and $\max_i |\widehat{\lambda}' f_i(\widehat{\theta})| = O_p(\sqrt{m/n})O_p([n\mathbb{E}(b_i^\alpha)]^{1/\alpha}) \to_p 0$, we can expand around $\lambda = 0$:

$$\begin{aligned}\widehat{Q}(\widehat{\theta}) =& \widehat{P}(\widehat{\theta}, \widehat{\lambda}) = \rho(0) - \widehat{\lambda}'\widehat{f} - \frac{1}{2}\widehat{\lambda}'\widehat{\Omega}(\widehat{\theta})\widehat{\lambda} + \widehat{r},\ \widehat{r} = \frac{1}{6}\sum_i \rho'''(\dot{\lambda}' f_i(\widehat{\theta}))(\widehat{\lambda}' f_i(\widehat{\theta}))^3/n \\ |\widehat{r}| \leq& C||\widehat{\lambda}||\ \max_i b_i\ \widehat{\lambda}'\widehat{\Omega}(\widehat{\theta})\widehat{\lambda} \leq o_p(1)||\widehat{\lambda}||^2 = o_p(m/n)\end{aligned}$$

Also $\widehat{\lambda}$ satisfy first-order condition $\sum_i \rho'(\widehat{\lambda}' f_i(\widehat{\theta}))f_i(\widehat{\theta})/n = 0$ w.p.a.1. By an expansion, for a $\dot{v}_i$ between 0 and $\widehat{\lambda}' f_i(\widehat{\theta})$, we have

$$\begin{aligned}0 =& -\widehat{f} - \widehat{\Omega}(\widehat{\theta})\widehat{\lambda} + \widehat{R},\ \widehat{R} = \frac{1}{2}\sum_i \rho'''(\dot{v}_i)(\widehat{\lambda}' f_i(\widehat{\theta}))^2 f_i(\widehat{\theta})/n \\ ||\widehat{R}|| \leq& C \max_i b_i\ \widehat{\lambda}'\widehat{\Omega}(\widehat{\theta})\widehat{\lambda} = O_p([n\mathbb{E}(b_i^\alpha)]^{1/\alpha} m/n) = o_p(\sqrt{m/n})\end{aligned}$$

Solving for $\widehat{\lambda} = \widehat{\Omega}(\widehat{\theta})^{-1}(-\widehat{f} + \widehat{R})$ and plugging into the expansion of $\widehat{Q}(\widehat{\theta})$

$$\widehat{Q}(\widehat{\theta}) = \rho(0) + \widehat{Q}^*(\widehat{\theta}) - \widehat{R}'\widehat{\Omega}(\widehat{\theta})^{-1}\widehat{R}/2 + \widehat{r} = \rho(0) + \widehat{Q}^*(\widehat{\theta}) + o_p(m/n)$$

An exactly analogous expansion, replacing $\widehat{\theta}$ with θ_0, gives

$$\widehat{Q}(\theta_0) = \rho(0) + \widehat{Q}^*(\theta_0) + o_p(m/n)$$

Then by definition of $\widehat{\theta}$

$$\widehat{Q}^*(\widehat{\theta}) = \widehat{Q}(\widehat{\theta}) - \rho(0) + o_p(m/n) \leq \widehat{Q}(\theta_0) - \rho(0) + o_p(m/n) = \widehat{Q}^*(\theta_0) + o_p(m/n)$$

□

Lemma A.9. *Under Assumption 2.1 - 2.2,* $||\widehat{\theta} - \theta_0|| = O_p(1)$.

Proof. By Lemma A.7, $||\widehat{f}(\widehat{\theta})|| = O_p(\sqrt{m/n})$, so by Assumption 2.1, w.p.a.1

$$||\widehat{\theta} - \theta_0|| \leq C||\widehat{f}(\widehat{\theta})|| + O_p(1) = O_p(1)$$

□

Lemma A.10. *Under Assumption 2.1 - 2.4,* $||\widehat{\theta} - \theta_0|| = O_p(\sqrt{m/n})$.

Proof. By Theorem 2.1, $\widehat{\theta} \to_p \theta_0$. By an expansion, $\widehat{f}(\widehat{\theta}) = \widehat{f}(\theta_0) + \widehat{F}(\widetilde{\theta})(\widehat{\theta} - \theta_0) = \widehat{f}(\theta_0) + \widetilde{F}(\widehat{\theta} - \theta_0)$. Let $\widehat{R}(\theta) = \widehat{f}(\theta)'\widehat{\Omega}(\theta)^{-1}\widehat{f}(\theta)$. Then,

$$\widehat{R}(\widehat{\theta}) = \widehat{R}(\theta_0) + 2\widehat{f}(\theta_0)'\widehat{\Omega}(\theta)^{-1}\widetilde{F}(\widehat{\theta} - \theta_0) + (\widehat{\theta} - \theta_0)'\widetilde{F}'\widehat{\Omega}(\theta)^{-1}\widetilde{F}(\widehat{\theta} - \theta_0)$$

By T and CS, for $\widehat{A} = [(\widehat{\theta} - \theta_0)'\widetilde{F}'\widehat{\Omega}(\theta)^{-1}\widetilde{F}(\widehat{\theta} - \theta_0)]^{1/2}$ and $\widehat{B} = [\widehat{R}(\widehat{\theta}) + \widehat{R}(\theta_0)]^{1/2}$,

$$\widehat{A}^2 \leq \widehat{B}^2 + \widehat{R}(\theta_0)^{1/2}\widehat{A} \leq \widehat{B}^2 + \widehat{B}\widehat{A} \leq \widehat{B}^2 + 2\widehat{B}\widehat{A}$$

Rearrange and we have $|\widehat{A} - \widehat{B}| \leq \sqrt{2}\widehat{B}$. By T, $|\widehat{A} - \widehat{B}| \geq \widehat{A} - \widehat{B}$, so that $\widehat{A} \leq (\sqrt{2}+1)\widehat{B} = C\widehat{B}$. By Lemma A.1, $\xi_{\max}(\widehat{\Omega}(\theta)^{-1}) \leq C$ w.p.a.1. Then by Lemma A.7 and T, $\widehat{B} = O_p(\sqrt{m/n})$.

Next by Assumption 2.3 and 2.4, $||F|| \leq C$. Then by Assumption 2.4,

$$\begin{aligned}
||\widehat{F}(\theta_0) - F||^2 =& O_p(\mathbb{E}[||F_i(\theta_0)||^2])/n = O_p(Cm\xi_{\max}(\mathbb{E}[F_i(\theta_0)F_i(\theta_0)']))/n = O_p(m/n) \\
||\widetilde{F}|| =& ||\widehat{F}(\theta_0)|| + o_p(1) = O_p(1) \\
||\widetilde{F} - F|| \leq& ||\widetilde{F} - \widehat{F}(\theta_0)|| + ||\widehat{F}(\theta_0) - F|| = o_p(1)
\end{aligned}$$

Also, we have

$$\begin{aligned}
||\widetilde{F}'\widehat{\Omega}(\widehat{\theta})^{-1}\widetilde{F} - F'\Omega^{-1}F|| \leq& ||\widetilde{F}'(\widehat{\Omega}(\widehat{\theta})^{-1} - \Omega(\widehat{\theta})^{-1})\widetilde{F}|| + ||\widetilde{F}'(\Omega(\widehat{\theta})^{-1} - \Omega^{-1})\widetilde{F}|| \\
&+ ||(\widetilde{F} - F)'\Omega^{-1}\widetilde{F}|| + ||F'\Omega^{-1}(\widetilde{F} - F)||
\end{aligned}$$

By Assumption 2.3 ii), $||\widehat{\Omega}(\widehat{\theta}) - \Omega(\widehat{\theta})|| \leq \sup_{\theta\in\Theta} ||\widehat{\Omega}(\theta) - \Omega(\theta)|| \to_p 0$. Then by the fact that eigenvalues of $\widehat{\Omega}(\widehat{\theta})$ and $\Omega(\widehat{\theta})$ are bounded w.p.a.1 and that $||\widetilde{F}|| = O_p(1)$, the first term goes to 0.

Now consider $\widetilde{F}^k$ to be the k-th column of $\widetilde{F}$, then $||\widetilde{F}^k|| = O_p(1)$. By Assumption 2.3 iv) and Lemma A.1, w.p.a.1

$$\begin{aligned}
||\widetilde{F}^{k\prime}(\Omega(\widehat{\theta})^{-1} - \Omega^{-1})\widetilde{F}^l|| =& ||\widetilde{F}^{k\prime}\Omega(\widehat{\theta})^{-1}(\Omega(\widehat{\theta}) - \Omega)\Omega^{-1}\widetilde{F}^l|| \\
\leq& C||\widetilde{F}^{k\prime}(\Omega(\widehat{\theta}) - \Omega)\widetilde{F}^l|| \\
\leq& C||\widetilde{F}^k||\ ||\widetilde{F}^l||\ ||\widehat{\theta} - \theta_0|| \to_p 0
\end{aligned}$$

For the third and forth terms, because $||\widetilde{F} - F|| = o_p(1)$, $||F|| \leq C$, $||\widetilde{F}|| = O_p(1)$ and $\xi_{\max}(\Omega^{-1}) \leq C$, they both go to 0. This gives $||\widetilde{F}'\widehat{\Omega}(\widehat{\theta})^{-1}\widetilde{F} - F'\Omega^{-1}F|| \to_p 0$, so by T,

$$\widetilde{F}'\widehat{\Omega}(\widehat{\theta})^{-1}\widetilde{F} \to_p U$$

Note U is non-singular, so that $\widehat{A}^2 \geq C||\widehat{\theta} - \theta_0||^2$ w.p.a.1. Thus,

$$C||\widehat{\theta} - \theta_0||^2 \leq \widehat{A}^2 \leq C\widehat{B}^2 = O_p(m/n)$$

. □

Lemma A.11. *Under Assumption 2.1 - 2.4, let* $\widetilde{\Omega} = -\sum_i \rho''(\widetilde{\lambda}' f_i(\bar{\theta}))f_i(\bar{\theta})f_i(\bar{\theta})'/n$, $\widetilde{F} = -\sum_i \rho'(\bar{\lambda}' f_i(\bar{\theta}))F_i(\bar{\theta})/n$, $\dot{F} = \widehat{F}(\dot{\theta})$, *where* $||\bar{\theta} - \theta_0|| = O_p(\sqrt{m/n})$, $||\dot{\theta} - \theta_0|| = O_p(\sqrt{m/n})$ *and* $||\widetilde{\lambda}|| = O_p(\sqrt{m/n})$, $||\bar{\lambda}|| = O_p(\sqrt{m/n})$. *Then we have* $||\widetilde{\Omega} - \Omega(\bar{\theta})|| \to_p 0$, $\widetilde{F}'\widetilde{\Omega}^{-1}\dot{F} \to_p U$ *and* $\sqrt{n}\widetilde{F}'\widetilde{\Omega}^{-1}\widehat{f}(\theta_0) \to_d N(\tau, U)$.

Proof. Note by Lemma A.10, $||\bar{\theta} - \theta_0|| = O_p(\sqrt{m/n})$. Also by Assumption 2.4, $\mathbb{E}[||F_i||^2] \leq Cm$, so by M, $||F_i||^2 = O_p(m)$. And $||F_i(\bar{\theta}) - F_i|| = O_p(m)||\bar{\theta} - \theta_0|| = O_p(m\sqrt{m/n})$, so $||F_i(\bar{\theta})|| = O_p(\sqrt{m})$. Then by CS, Assumption 2.3 and 2.2, w.p.a.1

$$\begin{aligned}
||\widetilde{F} - \widehat{F}(\bar{\theta})|| =& ||\sum_i [\rho'(\bar{\lambda}' f_i(\bar{\theta})) + 1]F_i(\bar{\theta})/n|| \\
\leq& C\sum_i |\bar{\lambda}' f_i(\bar{\theta})|||F_i(\bar{\theta})||/n \\
\leq& C\sqrt{\bar{\lambda}'\left(\frac{1}{n}\sum_i f_i(\bar{\theta})f_i(\bar{\theta})'\right)\bar{\lambda}}\sqrt{\sum_i ||F_i(\bar{\theta})||^2/n} \\
=& O_p(\sqrt{m/n})O_p(\sqrt{m}) \to_p 0
\end{aligned}$$

As the proof in Lemma A.10, $||\widetilde{F}|| = O_p(1)$ and by Assumption 2.4,

$$\begin{aligned}
||\widetilde{F} - F|| \leq& ||\widetilde{F} - \widehat{F}(\bar{\theta})|| + ||\widehat{F}(\bar{\theta}) - \widehat{F}(\theta_0)|| + ||\widehat{F}(\theta_0) - F|| \\
=& O_p(m/\sqrt{n}) + O_p(\sqrt{m/n}) + O_p(\sqrt{m/n}) \\
=& O_p(m/\sqrt{n}) \to_p 0
\end{aligned}$$

Similarly $||\dot{F}|| = O_p(1)$ and $||\dot{F} - F|| \to_p 0$.

Next note the k-th row l-th column element of $||\widehat{\Omega}(\bar{\theta}) - \Omega(\bar{\theta})||$ has the form $\frac{1}{n}\sum_i f_i^k(\bar{\theta})f_i^l(\bar{\theta}) - \mathbb{E}[f_i^k(\bar{\theta})f_i^l(\bar{\theta})]$. Because $\mathbb{V}[f_i^k(\bar{\theta})f_i^l(\bar{\theta})] \leq \sqrt{\mathbb{E}[|f_i^k(\bar{\theta})|^4]\mathbb{E}[|f_i^l(\bar{\theta})|^4]} \leq C$ w.p.a.1, then by CLT, each element of $||\widehat{\Omega}(\bar{\theta}) - \Omega(\bar{\theta})||$ is of order $O_p(1/\sqrt{n})$. So by Assumption 2.3 and 2.4, we have

$$
\begin{aligned}
||\widetilde{\Omega} - \widehat{\Omega}(\bar{\theta})|| \leq & || \sum_i [\rho''(\widetilde{\lambda}' f_i(\bar{\theta})) + 1] f_i(\bar{\theta}) f_i(\bar{\theta})'/n || \\
\leq & C \sum_i |\widetilde{\lambda}' f_i(\bar{\theta})| ||f_i(\bar{\theta}) f_i(\bar{\theta})'||/n \\
\leq & C \sqrt{\widetilde{\lambda}' \left(\frac{1}{n} \sum_i f_i(\bar{\theta}) f_i(\bar{\theta})' \right) \widetilde{\lambda}} \sqrt{\sum_i ||f_i(\bar{\theta})||^4/n} \\
= & O_p(\sqrt{m/n}) O_p(\mathbb{E}[\sup_{\theta \in \mathcal{N}} ||f_i(\theta)||^4])^{1/2} = O_p(m^{\frac{1}{4}}/n^{\frac{1}{4}}) \\
||\widetilde{\Omega} - \Omega(\bar{\theta})|| \leq & ||\widetilde{\Omega} - \widehat{\Omega}(\bar{\theta})|| + ||\widehat{\Omega}(\bar{\theta}) - \Omega(\bar{\theta})|| = O_p(m^{\frac{1}{4}}/n^{\frac{1}{4}}) + O_p(m/\sqrt{n}) \to_p 0
\end{aligned}
$$

So we have

$$
\begin{aligned}
||\widetilde{F}' \widetilde{\Omega}^{-1} \dot{F} - F' \Omega^{-1} F|| \leq & ||\widetilde{F}' (\widetilde{\Omega}^{-1} - \Omega(\bar{\theta})^{-1}) \dot{F}|| + ||\widetilde{F}' (\Omega(\bar{\theta})^{-1} - \Omega^{-1}) \dot{F}|| \\
& + ||(\widetilde{F} - F)' \Omega^{-1} \dot{F}|| + ||F' \Omega^{-1} (\dot{F} - F)||
\end{aligned}
$$

An analogue to the proof in Lemma A.10 gives the above going to 0. So by T,

$$
\widetilde{F}' \widetilde{\Omega}^{-1} \dot{F} \to_p U
$$

Next note $||\widetilde{\Omega}^{-1} \widehat{f}(\theta_0)|| \leq C ||\widehat{f}(\theta_0)|| = O_p(\sqrt{m/n})$, then

$$
\begin{aligned}
||\sqrt{n} \widetilde{F}' \widetilde{\Omega}^{-1} \widehat{f}(\theta_0) - \sqrt{n} F' \Omega^{-1} \widehat{f}(\theta_0)|| = & ||(\widetilde{F}' - F') \sqrt{n} \widetilde{\Omega}^{-1} \widehat{f}(\theta_0)|| + ||F' \widetilde{\Omega}^{-1} (\widetilde{\Omega} - \Omega(\bar{\theta})) \sqrt{n} \Omega(\bar{\theta})^{-1} \widehat{f}(\theta_0)|| \\
& + ||F' \Omega(\bar{\theta})^{-1} (\Omega(\bar{\theta}) - \Omega) \sqrt{n} \Omega^{-1} \widehat{f}(\theta_0)|| \\
= & O_p(m/\sqrt{n}\sqrt{m}) + O_p(m^{\frac{1}{4}}/n^{\frac{1}{4}}\sqrt{m}) + O_p(m/\sqrt{n}\sqrt{m}) \\
& + O_p(1) O_p(||\bar{\theta} - \theta_0||) O_p(\sqrt{m}) \\
= & O_p([m^3/n]^{\frac{1}{4}}) \to_p 0
\end{aligned}
$$

Now for any vector e such that $||e|| = 1$. Let $X_i = e' F' \Omega^{-1} f_i(\theta_0)/\sqrt{n}$.

$$
\begin{aligned}
& n\mathbb{E}[X_i] = e' F' \Omega^{-1} \sqrt{n} \mathbb{E}[f_i(\theta_0)] \to_p e' \tau \\
& n\mathbb{E}[X_i^2] = e' F' \Omega^{-1} \Omega \Omega^{-1} F e \to_p e' U e \\
& \sum_i (\mathbb{E}[X_i^2 | X_{i-1}, ..., X_1] - \mathbb{E}[X_i^2]) = 0 \\
& \sum_i \mathbb{E}[X_i^4] \leq \frac{1}{n} \mathbb{E}[||e' F' \Omega^{-1} f_i(\theta_0)||^4] \leq C \mathbb{E}(||f_i(\theta_0)||^4)/n \to_p 0
\end{aligned}
$$

By the martingale central limit theorem, $\sum_i X_i \to_d N(e'\tau, e'Ue)$. Then by the Cramer-Wold device, $\sqrt{n} F' \Omega^{-1} \widehat{f}(\theta_0) \to_d N(\tau, U)$. It follows that $\sqrt{n} \widetilde{F}' \widetilde{\Omega}^{-1} \widehat{f}(\theta_0) \to_d N(\tau, U)$.

□

Proof of Theorems 2.1. The second and third conclusions follow directly from Lemma A.7 and Lemma A.5.

Now consider any $\epsilon, \gamma > 0$. By Lemma A.3, A.8, A.9 and the fact that $Q(0) \leq Cm/n \to_p 0$, there is C such that for $\mathcal{A}_1 = \{\sup_{||\theta-\theta_0||\leq C} |\widehat{Q}^*(\theta) - Q(\theta)| < \gamma/4\}$, $\mathcal{A}_2 = \{\widehat{Q}^*(\widehat{\theta}) \leq \widehat{Q}^*(\theta_0) + \gamma/4\}$, $\mathcal{A}_3 = \{||\widehat{\theta} - \theta_0|| \leq C\}$, $\mathcal{A}_4 = \{Q(0) \leq \gamma/4\}$, we have $\Pr(\mathcal{A}_1 \cap \mathcal{A}_2 \cap \mathcal{A}_3 \cap \mathcal{A}_4) \geq 1 - \epsilon$ for all n large enough. And on $\mathcal{A} = \mathcal{A}_1 \cap \mathcal{A}_2 \cap \mathcal{A}_3 \cap \mathcal{A}_4$,

$$Q(\widehat{\theta}) \leq \widehat{Q}^*(\widehat{\theta}) + \gamma/4 \leq \widehat{Q}^*(\theta_0) + \gamma/2 \leq Q(\theta_0) + 3\gamma/4 \leq \gamma$$

Above implies that on $\mathcal{A}$, $\bar{f}(\widehat{\theta})'\Omega(\widehat{\theta})^{-1}\bar{f}(\widehat{\theta}) \leq 2\gamma$. Since ϵ, γ can be any positive constants, we have $\bar{f}(\widehat{\theta})'\Omega(\widehat{\theta})^{-1}\bar{f}(\widehat{\theta}) \to_p 0$. By Assumption 2.1 and 2.3 iii), w.p.a.1

$$\bar{f}(\widehat{\theta})'\Omega(\widehat{\theta})^{-1}\bar{f}(\widehat{\theta}) \geq C\bar{f}(\widehat{\theta})'\bar{f}(\widehat{\theta}) \geq C||\widehat{\theta} - \theta_0||^2$$

And this proves the first conclusion $\widehat{\theta} - \theta_0 \to_p 0$. □

Proof of Theorems 2.2. By Theorem 2.1, $\widehat{\lambda} = \text{argmax}_{\lambda \in \widehat{\Lambda}(\widehat{\theta})} \widehat{P}(\widehat{\theta}, \lambda)$ exists w.p.a.1 and $||\widehat{\lambda}|| = O_p(\sqrt{m/n})$. Choose $\Delta_n = o([n\mathbb{E}(b_i^\alpha)]^{-1/\alpha})$ and $\sqrt{m/n} = o(\Delta_n)$. Then $\widehat{\lambda} \in \Lambda_n = \{\lambda : ||\lambda|| \leq \Delta_n\}$ w.p.a.1, and by Lemma A.4, $\max_i |\widehat{\lambda}' f_i(\widehat{\theta})| \to_p 0$. Also, by consistency of $\widehat{\theta}$, $\widehat{\theta} \in \text{int}(\Theta)$ w.p.a.1. It follows that w.p.a.1 $\widehat{P}(\theta, \lambda)$ is twice continuously differentiable in a neighborhood of $(\widehat{\theta}, \widehat{\lambda})$. Then by first order condition $\partial\widehat{P}(\widehat{\theta}, \widehat{\lambda})/\partial\lambda = 0$ and the implicit function theorem, for all θ in the neighborhood of $\widehat{\theta}$, there is a continuously differentiable $\widehat{\lambda}(\theta)$, such that $\partial\widehat{P}(\theta, \widehat{\lambda}(\theta))/\partial\lambda = 0$. By concavity of $\widehat{P}$, we have $\widehat{Q}(\theta) = \widehat{P}(\theta, \widehat{\lambda}(\theta)) = \max_{\lambda \in \widehat{\Lambda}(\theta)} \widehat{P}(\theta, \lambda)$. Then the first-order conditions for $\widehat{\theta}$ and the envelope theorem give

$$0 = \partial\widehat{P}(\widehat{\theta}, \widehat{\lambda})/\partial\theta = -\widetilde{F}'\widehat{\lambda}, \quad \widetilde{F} = -\sum_i \rho'(\widehat{\lambda}' f_i(\widehat{\theta}))F_i(\widehat{\theta})/n$$

Expanding the first-order condition for $\widehat{\lambda}$ around $\lambda = 0$ gives

$$0 = -\widehat{f}(\widehat{\theta}) - \widetilde{\Omega}\widehat{\lambda} = 0, \quad \widetilde{\Omega} = -\sum_i \rho''(\widetilde{\lambda}' f_i(\widehat{\theta}))f_i(\widehat{\theta})f_i(\widehat{\theta})'/n$$

By Lemma A.10 and Theorem 2.1, $||\widehat{\theta} - \theta_0|| = O_p(\sqrt{m/n})$ and $||\widehat{\lambda}|| = O_p(\sqrt{m/n})$. Then by Lemma A.11, $\widetilde{\Omega}$ is non-singular w.p.a.1. So solving $\widehat{\lambda}$ in the FOC for $\widehat{\lambda}$ and plugging it in the FOC for $\widehat{\theta}$, we have

$$\widetilde{F}'\widetilde{\Omega}^{-1}\widehat{f}(\widehat{\theta}) = 0$$

Expanding $\widehat{f}$ around θ_0 gives, for a mean value $\dot{\theta}$ and $\dot{F} = \widehat{F}(\dot{\theta})$,

$$\widetilde{F}'\widetilde{\Omega}^{-1}\dot{F}(\widehat{\theta} - \theta_0) + \widetilde{F}'\widetilde{\Omega}^{-1}\widehat{f}(\theta_0) = 0$$

Again apply Lemma A.11 with with $\bar{\theta} = \widehat{\theta}$, $\dot{\theta} = \dot{\theta}$, $\widetilde{\lambda} = \widetilde{\lambda}$ and $\bar{\lambda} = \widehat{\lambda}$, we have $\widetilde{F}'\widetilde{\Omega}^{-1}\dot{F} \to_p U$, which is non-singular, and $\sqrt{n}\widetilde{F}'\widetilde{\Omega}^{-1}\widehat{f}(\theta_0) \to_d N(\tau, U)$. Then by Slutsky's Theorem,

$$\sqrt{n}(\widehat{\theta} - \theta_0) \to_d N(-U^{-1}\tau, U^{-1})$$

Finally, for the variance estimator $\widehat{U}$. Apply Lemma A.11 with $\bar{\theta} = \dot{\theta} = \widehat{\theta}$ and $\widetilde{\lambda} = \bar{\lambda} = 0$, we have $\widehat{U} \to_p U$. □

Proof of Corollary 2.1. Simply apply Theorem 2.2 to S_v. □

Proof of Corollary 3.1. Apply Delta method to asymptotic distribution of $\widehat{\theta}_S$ in Theorem 2.2 and we get that of $\widehat{\mu}_S$.

The asymptotic L_p risk functions of $\widehat{\mu}_S$ would be

$$\begin{aligned}
p \text{ odd: } ALP(p, \widehat{\mu}_S) =& \frac{1}{\sqrt{\pi}} \sum_{j=0}^{(p-1)/2} \binom{p}{2j} 2^j (\kappa_S' U_S \kappa_S)^j |\kappa_S' \tau_S|^{p-2j} \Gamma(j + \frac{1}{2}) \\
&+ \frac{1}{\sqrt{\pi}} \sum_{j=0}^{p} \binom{p}{j} 2^{j/2} (\kappa_S' U_S \kappa_S)^{j/2} (-|\kappa_S' \tau_S|)^{p-j} \Gamma(\frac{j+1}{2}, \frac{(\kappa_S' \tau_S)^2}{2\kappa_S' U_S \kappa_S}) \\
p \text{ even: } ALP(p, \widehat{\mu}_S) =& \frac{1}{\sqrt{\pi}} \sum_{j=0}^{p/2} \binom{p}{2j} 2^j (\kappa_S' U_S \kappa_S)^j (\kappa_S' \tau_S)^{p-2j} \Gamma(j + \frac{1}{2})
\end{aligned}$$

where $\Gamma(\cdot)$ is the gamma function and $\Gamma(a, x) = \int_x^\infty t^{a-1} e^{-t} dt$ is the incomplete gamma function.

We can prove above by looking at a univariate normal distribution $X \sim N(a, b^2)$:

$$\begin{aligned}
&\mathbb{E}[|X|^p] = \mathbb{E}[|a + bZ|^p] \qquad Z \sim N(0,1) \\
=& \frac{1}{\sqrt{2\pi}} \left(\int_{-\frac{a}{b}}^{\infty} (a + bz)^p e^{-\frac{z^2}{2}} dz + (-1)^p \int_{-\infty}^{-\frac{a}{b}} (a + bz)^p e^{-\frac{z^2}{2}} dz \right) \\
=& \frac{1}{\sqrt{2\pi}} \sum_{j=0}^{p} \binom{p}{j} b^j a^{p-j} \left(\int_{-\frac{a}{b}}^{\infty} z^j e^{-\frac{z^2}{2}} dz + (-1)^p \int_{-\infty}^{-\frac{a}{b}} z^j e^{-\frac{z^2}{2}} dz \right)
\end{aligned}$$

Then further simply the formula for p odd and even:

$$\begin{aligned}
p \text{ odd: } \mathbb{E}[|X|^p] =& \frac{1}{\sqrt{\pi}} \sum_{j=0}^{(p-1)/2} \binom{p}{2j} 2^j b^{2j} |a|^{p-2j} \Gamma(j + \frac{1}{2}) \\
&+ \frac{1}{\sqrt{\pi}} \sum_{j=0}^{p} \binom{p}{j} 2^{j/2} b^j (-|a|)^{p-j} \Gamma(\frac{j+1}{2}, \frac{a^2}{2b^2}) \\
p \text{ even: } \mathbb{E}[|X|^p] =& \frac{1}{\sqrt{\pi}} \sum_{j=0}^{p/2} \binom{p}{2j} 2^j b^{2j} a^{p-2j} \Gamma(j + \frac{1}{2})
\end{aligned}$$

And $ALP(p, \widehat{\mu}_S) = \mathbb{E}[|X|^p]$ with $a = \kappa_S'\tau_S$ and $b^2 = \kappa_S' U_S \kappa_S$. AMAE and AMSE are just special cases where $p = 1$ and $p = 2$.

Finally, for linex risk:

$$\begin{aligned} ALL(\widehat{\mu}_S) =& \mathbb{E}[e^{cX} - cX - 1] \\ =& \frac{1}{\sqrt{2\pi}} \int e^{ca+cbz} e^{-\frac{z^2}{2}} dz - ca - 1 \\ =& e^{c^2b^2/2+ca} - ca - 1 \end{aligned}$$

□

Proof of Theorem 3.1. First note by Assumption 2.3, Lemma A.7 and Theorem 2.2, w.p.a.1

$$\begin{aligned} ||\widehat{f}_S(\widehat{\theta}_v)|| \leq& ||\widehat{f}_S(\widehat{\theta}_S)|| + ||\widehat{f}_S(\widehat{\theta}_S) - \widehat{f}_S(\widehat{\theta}_v)|| \\ \leq& O_p(\sqrt{|S|/n}) + O_p(1)||\widehat{\theta}_S - \widehat{\theta}_v|| \\ \leq& O_p(\sqrt{|S|/n}) + O_p(1)[||\widehat{\theta}_S - \theta_0|| + ||\widehat{\theta}_v - \theta_0||] \\ =& O_p(\sqrt{|S|/n}) \end{aligned}$$

An analogue proof as in Lemma A.11 shows $||\widehat{\tau}_S - \sqrt{n}F_S'\Omega_S^{-1}\widehat{f}_S(\widehat{\theta}_v)|| \to_p 0$, where $F_S = F_S(\theta_0)$ and $\Omega_S = \Omega_S(\theta_0)$.

Next from the proof of Theorem 2.2, $\sqrt{n}(\widehat{\theta}_v - \theta_0) = -U_v^{-1}\sqrt{n}F_v'\Omega_v^{-1}\widehat{f}_v(\theta_0) + o_p(1) \to_d N(0, U_v^{-1})$. Also $||F_S|| \leq C$, $||\widehat{F}_S(\widetilde{\theta}) - F_S|| = o_p(1)$ when $\widetilde{\theta} \to_p \theta_0$, and eigenvalues of Ω_S are bounded.

By a mean-value expansion:

$$\begin{aligned} \widehat{f}_S(\widehat{\theta}_v) =& \widehat{f}_S(\theta_0) + \widehat{F}_S(\widetilde{\theta})\left(\widehat{\theta}_v - \theta_0\right) \\ \widehat{\tau}_S =& \sqrt{n}F_S'\Omega_S^{-1}\widehat{f}_S(\theta_0) - F_S'\Omega_S^{-1}\widehat{F}_S(\widetilde{\theta})U_v^{-1}\sqrt{n}F_v'\Omega_v^{-1}\widehat{f}_v(\theta_0) + o_p(1) \\ =& \sqrt{n}F_S'\Omega_S^{-1}\widehat{f}_S(\theta_0) - F_S'\Omega_S^{-1}F_S U_v^{-1}\sqrt{n}F_v'\Omega_v^{-1}\widehat{f}_v(\theta_0) + o_p(1) \end{aligned}$$

As in the proof of Lemma A.11, $\sqrt{n}F_S'\Omega_S^{-1}\widehat{f}_S(\theta_0) \to_d N(\tau_S, U_S)$, $F_S'\Omega_S^{-1}F_S \to U_S$ and $\sqrt{n}F_v'\Omega_v^{-1}\widehat{f}_v(\theta_0) \to_d N(0, U_v)$, then

$$\begin{aligned} &\mathbb{COV}\left(\sqrt{n}F_S'\Omega_S^{-1}\widehat{f}_S(\theta_0), -F_S'\Omega_S^{-1}F_S U_v^{-1}\sqrt{n}F_v'\Omega_v^{-1}\widehat{f}_v(\theta_0)\right) \\ =& -F_S'\Omega_S^{-1}F_S U_v^{-1}F_v'\Omega_v^{-1}\mathbb{COV}\left(\sqrt{n}\widehat{f}_v(\theta_0), \sqrt{n}\widehat{f}_S(\theta_0)\right)\Omega_S^{-1}F_S \\ =& -F_S'\Omega_S^{-1}F_S U_v^{-1}F_v'\Omega_v^{-1}\Xi_{v,S}\mathbb{COV}\left(\sqrt{n}\widehat{f}_S(\theta_0), \sqrt{n}\widehat{f}_S(\theta_0)\right)\Omega_S^{-1}F_S \\ \to_p& -U_S \end{aligned}$$

And by property of normal distribution, $\widehat{\tau}_S \to_d N(\tau_S, U_S U_v^{-1} U_S - U_S)$.

□

Proof of Corollary 3.2. If all $\widehat{U}_S$ are consistent, then $\widehat{\Upsilon}_S$ is consistent for $\Upsilon_S = U_S U_v^{-1} U_S - U_S$.

By Theorem 3.1 and the CMT, we have $\widehat{\tau}_S \widehat{\tau}_S' \to_d \widetilde{M}_S \widetilde{M}_S'$, where $\widetilde{M}_S \sim N(\tau_S, \Upsilon_S)$. Since $\mathbb{E}[\widetilde{M}_S \widetilde{M}_S'] = \Upsilon_S + \tau_S \tau_S'$, $\widehat{\tau}_S \widehat{\tau}_S{}' - \widehat{\Upsilon}_S$ is an asymptotically unbiased estimator of $\tau\tau'$.

For the second part, consider the variable $\widehat{V} = c\kappa_S' \widehat{\tau}_S \to_d N\left(c\kappa_S'\tau_S, c^2 \kappa_S' \Upsilon_S \kappa_S\right)$.

$$\begin{aligned}\mathbb{E}[e^{\widehat{V}}] =& \frac{1}{\sqrt{2\pi}} \int \exp\left((c^2 \kappa_S' \Upsilon_S \kappa_S)^{1/2} z + c\kappa_S'\tau_S\right) e^{-z^2/2} dz \\ =& \exp\left(c^2 \kappa_S' \Upsilon_S \kappa_S / 2 + c\kappa_S'\tau_S\right)\end{aligned}$$

Therefore, $\exp\left(c\widehat{\kappa}_S'\widehat{\tau}_S - c^2 \widehat{\kappa}_S' \widehat{\Upsilon}_S \widehat{\kappa}_S / 2\right)$ is an asymptotically unbiased estimator of $\exp\left(c\kappa_S'\tau_S\right)$.

□

Proof of Theorems 4.1, 4.2. The proof is standard, we only need to substitute the general moments with specific DGP induced moments and check for the assumptions. □

References

Andrews, D. W. K., May 1999. Consistent moment selection procedures for generalized methods of moments estimation. Econometrica 67 (3), 543–564.

Andrews, D. W. K., Lu, B., 2001. Consistent model and moment selection procedures for GMM estimation with application to dynamic panel data models. Journal of Econometrics 101, 123–164.

Berkowitz, D., Caner, M., Fang, Y., 2012. The validity of instruments revisited. Journal of Econometrics 166 (2), 255–266.

Chang, M., DiTraglia, F. J., 2018. A generalized focused information criterion for gmm. Journal of Applied Econometrics 33 (3), 378–397.

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

Cheng, X., Liao, Z., October 2013. Select the valid and relevant moments: An information-based LASSO for GMM with many moments, PIER Working Paper 13-062.
URL http://economics.sas.upenn.edu/system/files/13-062.pdf

Claeskens, G., Croux, C., Jo, 2006. Variable selection for logistic regression using a prediction-focused information criterion. Biometrics 62, 972–979.

Claeskens, G., Hjort, N. L., 2003. The focused information criterion. Journal of the American Statistical Association 98 (464), 900–945.

Claeskens, G., Hjort, N. L., 2008. Minimizing average risk in regression models. Econometric Theory 24, 493–527.

DiTraglia, F. J., 2016. Using invalid instruments on purpose: Focused moment selection and averaging for gmm. Journal of Econometrics 195 (2), 187–208.

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

Donald, S. G., Imbens, G. W., Newey, W. K., 2009. Choosing instrumental variables in conditional moment restriction models. Journal of Econometrics 152, 28–36.

Donald, S. G., Newey, W. K., September 2001. Choosing the number of instruments. Econometrica 69 (5), 1161–1191.

Hall, A. R., Peixe, F. P., 2003. A consistent method for the selection of relevant instruments in linear models. Econometric Reviews 22, 269–288.

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

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

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

Imbens, G. W., 1997. One-step estimators for over-identified generalized method of moments models. The Review of Economic Studies 64 (3), 359–383.

Imbens, G. W., Johnson, P., Spady, R. H., 1998. Information theoretic approaches to inference in moment condition models. Econometrica 66, 333–357.

Kolesr, M., Chetty, R., Friedman, J., Glaeser, E., 2015. Identification and inference with many invalid instruments. Journal of Business and Economic Statistics 33 (4), 474–484.

Kuersteiner, G., Okui, R., March 2010. Constructing optimal instruments by first-stage prediction averaging. Econometrica 78 (2), 679–718.

Leeb, H., Pötscher, B. M., 2008. Sparse estimators and the oracle property, or the return of Hodges' estimator. Journal of Econometrics 142, 201–211.

Newey, W. K., 1985. Genearlized method of moments specification testing. Journal of Econometrics 29, 229–256.

Newey, W. K., Smith, R. J., 2004. Higher order properties of gmm and generalized empirical likelihood estimators. Econometrica 72 (1), 219–255.

Newey, W. K., Windmeijer, F., 2009. Many weak moment asymptotics for generalized empirical likelihood estimators. Econometrica 77, 687–721.

Owen, A. B., 1997. Alternative semi-parametric likelihood approaches to generalized method of moments estimation. Biometrika 75 (2), 237–249.

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

Schorfheide, F., 2005. VAR forecasting under misspecification. Journal of Econometrics 128, 99–136.

Smith, R. J., 1997. Alternative semi-parametric likelihood approaches to generalized method of moments estimation. The Economic Journal 107 (441), 503–519.

Stutzer, M. J., Kitamura, Y., 1997. An information-theoretic alternative to generalized method of moments estimation. Econometrica 65 (4), 861–874.

Yang, Y., 2005. Can the strengths of AIC and BIC be shared? a conflict between model identification and regression estimation. Biometrika 92 (4), 937–950.