Efficient two-step generalized empirical likelihood estimation and tests with martingale differences

Fei Jin[1] and Lung-fei Lee[*2]

[1]School of Economics, Fudan University, and Shanghai Institute of International Finance and Economics, Shanghai 200433, China

[2]Department of Economics, The Ohio State University, Columbus, OH 43210, USA

May 5, 2020

Abstract

This paper considers two-step generalized empirical likelihood (GEL) estimation and tests with martingale differences when there is a computationally simple $\sqrt{n}$-consistent estimator of nuisance parameters or the nuisance parameters can be eliminated with an estimating function of parameters of interest. As an initial estimate might have asymptotic impact on final estimates, we propose general $C(\alpha)$-type transformed moments to eliminate the impact, and use them in the GEL framework to construct estimation and tests robust to initial estimates. This two-step approach can save computational burden as the numbers of moments and parameters are reduced. A properly constructed two-step GEL (TGEL) estimator of parameters of interest is asymptotically as efficient as the corresponding joint GEL estimator. TGEL removes several higher order bias terms of a corresponding two-step generalized method of moments. Our moment functions at the true parameters are martingales, thus they cover some spatial and time series models. We investigate tests for parameter restrictions in the TGEL framework, which are locally as powerful as those in the joint GEL framework when the two-step estimator is efficient.

[*]Corresponding author. Tel.: +1 614 292 5508; fax: +1 614 292 3906. E-mail addresses: jin.fei@live.com (Fei Jin), lee.1777@osu.edu (Lung-fei Lee).

Keywords: Two-step estimation, empirical likelihood, GMM, martingale difference, higher order bias, efficiency, test

JEL classification: C12, C13, C14, C18, C51

1 Introduction

A two-step estimation method is often employed in empirical studies due to its computational simplicity. In this method, we obtain a computationally simple estimator of nuisance parameters in a first step, and then use it to derive an estimator of parameters of interest. This is helpful in computation, in particular when the number of nuisance parameters is large but their estimates can be easily obtained, as we only need to estimate a relatively small number of parameters in the second step. This can be the case if researchers include many covariates in the model or estimating equations are rather complex. Famous two-step estimators include those for the sample selection model (Heckman, 1976), and the linear expectation model (Barro, 1977). Properties of two-step estimators have been studied by Newey (1984) and Pagan (1984, 1986), among others. However, it is understood that a first-step estimate might have asymptotic impact on the variance of the second-step estimate, thus a properly constructed asymptotic variance of the final estimate is needed. Furthermore, two-step estimators might be less efficient than corresponding one-step ones.

This paper studies generalized empirical likelihood (GEL) estimation and tests of moment condition models when a $\sqrt{n}$-consistent estimator of nuisance parameters is available or the nuisance parameters can be eliminated by the method of elimination and substitution, which results in an estimating function of parameters of interest. The first contribution of our paper is the use of a $C(\alpha)$-type moment vector (Neyman, 1959) to eliminate the asymptotic impact of nuisance parameters in the GEL framework. The resulting two-step GEL (TGEL) estimator has an asymptotic distribution with a variance of the usual sandwich form, is computationally simple and can also be asymptotically as efficient as the original joint GEL estimator. The $C(\alpha)$-type moment vector is a transformed moment vector of the original one so that any $\sqrt{n}$-consistent estimator of nuisance parameters will not have an impact on the asymptotic distribution of a second-step estimate of parameters of interest. In forming the $C(\alpha)$-type moment vector, the number of moment conditions will be reduced by at least the number of nuisance parameters. Since the number of Lagrangian multipliers in GEL's saddle point characterization is equal to the number of moments, its computational burden is directly related to the number of moments. As a result, our two-step approach saves computational time for GEL in terms of the reduction in both the number of estimated parameters and that of Lagrangian parameters. Furthermore, we show that the TGEL estimator of parameters of interest is asymptotically as efficient as the joint GEL estimator if the number of moments reduced is equal to the number of nuisance parameters. We refer to the TGEL estimator in this case as the efficient TGEL (E-TGEL) estimator. If we are also interested in efficient estimation of "nuisance" parameters, then the E-TGEL estimate may be plugged into the joint GEL objective function in a subsequent step of estimation.

Our second contribution is the investigation of tests in the TGEL framework. In these tests, the nuisance parameter estimator can be any $\sqrt{n}$-consistent estimator, which may or may not relate to the original moments for estimation. The $C(\alpha)$-type moment vector at a $\sqrt{n}$-consistent nuisance parameter estimate behaves as if nuisance parameters were known, so various tests can be constructed with the TGEL objective function similarly as those in an ordinary GEL framework. Newey, Ramalho and Smith (2005) investigate a GEL estimator where estimated nuisance parameters are directly plugged into the GEL objective function. Their objective function will not provide asymptotically chi-squared distributed tests due to the asymptotic impact of the nuisance parameter estimator. Our GEL objective function with the $C(\alpha)$-type moment vector overcomes this problem. We investigate the GEL ratio tests, GEL score-type test, GEL Wald test, and GEL test with the generalized method of moments (GMM) gradient in the two-step estimation framework. As in Guggenberger and Smith (2005) and Smith (2011), a score-type test in the TGEL framework can be directly based on the derivative vector of the TGEL objective function. For a GMM gradient test, Lee and Yu (2012) and Dufour, Trognon and Tuvaandorj (2017) have investigated a $C(\alpha)$-type form with any $\sqrt{n}$-consistent restricted estimator. This $C(\alpha)$-type test can also be implemented in the GEL framework, which will only use its moments in formulating the test statistic as it internalizes its variance matrix. So a GEL test can be robust to unknown heteroskedasticity.[1] We show that tests in the TGEL framework can be locally as powerful as those in the joint GEL framework.

Our third contribution is the consideration of moment functions that are martingale arrays at the true parameters in a two-step estimation, so that they cover some time series autoregressive models (Chuang and Chan, 2002) and spatial autoregressive (SAR) models (Jin and Lee, 2019). For SAR models, linear and quadratic moment functions are basic ones, and at the true parameter values, they can be written as martingale arrays (See e.g., Kelejian and Prucha, 2001).

We also investigate the higher order bias of TGEL. One reason that GEL attracts much attention is that it can have smaller higher order asymptotic bias than that of the two-step optimal GMM (OGMM), as shown in Newey and Smith (2004) for random samples and in Anatolyev (2005) for stationarity time series models with non-i.i.d. data.[2] Finite sample Monte Carlo studies have reported that OGMM can have large bias (e.g., Altonji and Segal, 1996), and GEL performs better than OGMM for models with random samples (e.g., Hansen, Heaton and Yaron 1996; Imbens 1997; Ramalho 2002; Mittelhammer, Judge and Schoenberg 2005; Newey et al. 2005). The bias advantage of GEL carries over to TGEL, which can remove several higher order bias terms of a corresponding two-step GMM (TGMM) estimator. However, like a GEL estimator, a TGEL estimator might not have finite moments and the analysis does not necessarily imply exact finite sample properties of TGEL and TGMM. Hausman, Lewis, Menzel and Newey (2011) provide a theoretical analysis of bias issues of the continuous updating (CU) estimator, a member of the GEL estimator. Our Monte Carlo results show that two-step empirical likelihood (EL) and two-step exponential tilting (ET) estimators have smaller bias and dispersion than TGMM estimators in finite samples. Both EL and ET are GEL members.

Our TGEL is related to several approaches in the literature, but they are all about GMM

[1] On the other hand, for a GMM test to be robust to unknown heteroskedasticity, we need to be careful on using a proper variance for its moments, whose inverse is also the optimal weighting matrix in its GMM objective function.

[2] The empirical likelihood (EL) introduced in Owen (1991), as a member of GEL, has other advantages such as the Bartlett correctability of EL ratio tests and confidence intervals (e.g., DiCiccio et al., 1991; Corcoran, 1998; Chen and Cui, 2007), Bahadur efficiency (Otsu, 2010) and optimality in terms of large deviations of EL ratio tests (Kitamura, 2001) for i.i.d. random samples.

instead of GEL. First, it has some similarity to that in Crepon, Kramarz and Trognon (1997) for the GMM estimation. They require that a subset of empirical moment conditions has a unique solution of the nuisance parameter vector as a function of the parameters of interest. We consider both the case with an initial $\sqrt{n}$-consistent estimator of nuisance parameters and the case where the nuisance parameters can be replaced by an estimating function of parameters of interest. In our TGEL or E-TGEL, the initial consistent estimator of nuisance parameters or the estimating function need not necessarily be from a subset of originally proposed empirical moment conditions. So our result can be regarded as a generalization of Crepon et al. (1997) in a flexible extension to separate moments estimation in the first step and also martingale arrays. In another related two-stage GMM procedure proposed in Gouriéroux, Monfort and Renault (1996), the set of moments are partitioned and the set of parameters are unfolded to derive an asymptotically efficient estimator of all parameters. We note again that their initial consistent estimator is from a subset of the moment vector, but not any $\sqrt{n}$-consistent estimator. In the presence of a consistent estimator of nuisance parameters, Trognon and Gouriéroux (1990) propose a two-step estimation that can efficiently estimate all model parameters. They obtain an approximated objective function by a second order Taylor expansion of an original objective function at the initial consistent nuisance parameter estimate, thus the resulting estimator of nuisance parameters is just a second-round estimator of the Newton-Raphson method. We shall show that our E-TGEL estimator of parameters of interest is asymptotically equivalent to their estimator.[3] Frazier and Renault (2017) consider a general setting of two-step estimation where there are awkward occurrences of the parameters of interest. Their approach can be seen as a generalization of that in Trognon and Gouriéroux (1990). Song, Fan and Kalbfleisch (2005) propose the algorithm of maximization by parts for separable log-likelihood functions. It has been generalized to non-separable extremum estimation problems in Fan, Pastorello and Renault (2015). The algorithm is iterative and involves the tuning parameter of the number of iterations, which might need to be large.

There are also papers in the semiparametric framework for two-step estimation. Ackerberg, Chen, Hahn and Liao (2014) consider a particular model where nuisance functions are

[3] In a way, this equivalence provides an account to justify that our E-TGEL is asymptotically efficient for the estimation of the parameters of interest.

identified by conditional moment restrictions not involving parameters of interest, which is a subset of all moments used for estimation. They show that semiparametric two-step optimally weighted GMM estimators can achieve the efficiency bound. Thus, they have a special setting of moments of our E-TGEL to a specific semiparametric model. Chernozhukov, Chetverikov, Demirer, Duflo, Hansen, Newey and Robins (2018a) provide an excellent literature review on $C(\alpha)$-type orthogonalization and they show that it can be used to construct debiased/double machine learning estimators. Chernozhukov, Escanciano, Ichimura, Newey and Robins (2018b) give a general construction of debiased/locally robust/orthogonal moment functions for GMM, where a first step nonparametric estimation has no effect on the influence function.

This paper is organized as follows. Section 2 introduces the TGEL estimator and investigates its asymptotic distribution, efficiency and higher order asymptotic bias properties. Section 3 studies various tests for parameter restrictions. Monte Carlo results are reported in Section 4. Section 5 concludes. Lemmas and proofs of some theorems are provided in appendices, and more detailed proofs are in the online supplementary material associated with this article.

2 Two-step GEL

Suppose that a sample moment $g_n(\gamma) = \frac{1}{n}\sum_{i=1}^{n} g_{ni}(\gamma)$ is available, where n is the sample size, γ is a $k_\gamma \times 1$ parameter vector and $g_{ni}(\gamma)$ is $m_g \times 1$ with $m_g \geq k_\gamma$. The true parameter value γ_0 of γ is characterized by the equation

$$\mathbf{E}[g_n(\gamma_0)] = 0.$$

In general, $g_{ni}(\gamma)$ can depend on all n data observations. At the true value γ_0 of γ, $g_{ni}(\gamma_0)$'s are martingale differences (MD) with respect to an increasing sequence of σ-fields, but at other parameter values they may not be MDs. Denote $\gamma = (\alpha', \beta')'$, where α is a $k_\alpha \times 1$ subvector of nuisance parameters, and β is a $k_\beta \times 1$ subvector of parameters of interest such that $k_\gamma = k_\alpha + k_\beta$. We consider the estimation of β with the moment vector $g_n(\check{\alpha}, \beta)$ or $g_n(\dot{\alpha}_n(\beta), \beta)$, where $\check{\alpha}$ is any $\sqrt{n}$-consistent estimator of α and $\dot{\alpha}_n(\beta)$ is any estimating function of β such that its value at a consistent estimator of β is a $\sqrt{n}$-consistent estimator of α. An example of $\check{\alpha}$ is a GMM estimator derived from a moment vector $h_n(\alpha)$ that does not involve β, and an example of $\dot{\alpha}_n(\beta)$

is an estimating equation derived from $h_n(\alpha, \beta)$. Here $h_n(\alpha)$ and $h_n(\alpha, \beta)$ do not need to be subvectors of $g_n(\gamma)$. The interesting situation is the case where $\check{\alpha}$ and $\dot{\alpha}_n(\beta)$ are easy to derive, while moments for the joint estimation of both α and β are efficient but relatively complex. Some examples of $\check{\alpha}$ and $\dot{\alpha}_n(\beta)$ are given below.

Example 1. *Consider the following probit model with an endogenous regressor:*

$$\begin{aligned} &y_1^* = x_1'\kappa + y_2\tau + u, \quad y_2 = x_1'\alpha_1 + x_2'\alpha_2 + \epsilon, \\ &y_1 = 1 \textit{ if } y_1^* > 0, \textit{ and } y_1 = 0 \textit{ otherwise}, \\ &\left(\begin{smallmatrix} u \\ \epsilon \end{smallmatrix}\right)|x \sim N\Big(0, \Big(\begin{smallmatrix} \sigma_u^2 & \rho\sigma_u\sigma_\epsilon \\ \rho\sigma_u\sigma_\epsilon & \sigma_\epsilon^2 \end{smallmatrix}\Big)\Big), \end{aligned} \tag{2.1}$$

where x_1 and x_2 are exogenous variable vectors, $x = [x_1', x_2']'$, y_2 is an endogenous regressor, and κ, τ, α_1 and α_2 are parameters. While y_1^* is not observable, we can observe an indicator y_1, which takes value 1 *if y_1^* is positive, and takes value* 0 *otherwise. Since y_2 is generated by a linear model, a simple consistent estimator of $\alpha = [\alpha_1', \alpha_2']'$ is the OLS estimator, which relates to a moment vector $g_a(\alpha) = x(y_2 - x'\alpha)$ that only involves α. This estimator can be used to construct a TGEL estimator of parameters of interest, denoted by β, with some moment conditions. A TGEL estimator with a proper parameter normalization for identification is defined in Section 4.1. The whole moment vector may or may not include $g_a(\alpha)$. For example, the empirical moments consisting of scores are rather complex and do not include $g_a(\alpha)$ (See Rivers and Vuong, 1988).*

Example 2. *Consider the Box-Cox transformation model for a positive dependent variable y_i: $z_i(\beta) = x_i'\alpha + u_i$, where $z_i(\beta) = (y_i^\beta - 1)/\beta$ if $\beta \neq 0$, $z_i(\beta) = \log y_i$ if $\beta = 0$, x_i is a vector of exogenous variables and u_i is an error term with mean zero. Although the model is nonlinear in β, it is a linear regression model for given β, so an estimating function for α can be obtained by regressing $z_i(\beta)$ on x_i, which relates to the moment vector $x_i[z_i(\beta) - x_i'\alpha]$ that involves both α and β.*

Example 3. *Consider the following SAR model:*

$$Y_n = \beta W_n Y_n + X_n\alpha_1 + V_n, \tag{2.2}$$

where n is the sample size, Y_n is an $n \times 1$ vector of observations on the dependent variable, X_n is an $n \times k_x$ matrix of observations on k_x exogenous variables, W_n is an $n \times n$ spatial weights matrix with a zero diagonal, β is a scalar spatial dependence parameter, α_1 is a $k_x \times 1$ vector of coefficients, *and elements v_{ni}'s of $V_n = [v_{n1}, \dots, v_{nn}]'$ are i.i.d. with mean zero and variance σ^2. Model* (2.2) *can be estimated with moments linear in V_n and moments quadratic in V_n (Lee, 2007), where the latter ones are motivated from the quasi maximum likelihood estimation and Moran's I test for spatial dependence (Moran, 1950). Like the above Box-Cox transformation model, model* (2.2) *is a linear regression model $(I_n - \beta W_n)Y_n = X_n\alpha_1 + V_n$ for given β. Thus, natural estimating functions of α_1 and σ^2 can be, respectively,*

$$\dot{\alpha}_{1n}(\beta) = (X_n' X_n)^{-1} X_n' (I_n - \beta W_n) Y_n, \tag{2.3}$$

$$\dot{\sigma}_n^2(\beta) = \frac{1}{n} V_n'(\dot{\alpha}_{1n}(\beta), \beta) V_n(\dot{\alpha}_{1n}(\beta), \beta), \tag{2.4}$$

where $V_n(\alpha_1, \beta) = (I_n - \beta W_n)Y_n - X_n\alpha_1$. As in Jin and Lee (2019), model (2.2) *can be estimated with the moment vector:*

$$g_n(\gamma) = \frac{1}{n}[V_n'(\alpha_1,\beta)P_{1n}V_n(\alpha_1,\beta) - \sigma^2\operatorname{tr}(P_{1n}), \dots, V_n'(\alpha_1,\beta)P_{k_pn}V_n(\alpha_1,\beta) - \sigma^2\operatorname{tr}(P_{k_pn}), V_n'(\alpha_1,\beta)Q_n]', \tag{2.5}$$

where $\gamma = [\alpha', \beta]'$ with $\alpha = [\alpha_1', \sigma^2]'$, P_{jn}'s are $n \times n$ matrices, which are functions of W_n, and Q_n is an $n \times k_q$ IV matrix, which are functions of X_n and W_n. With $\dot{\alpha}_{1n}(\beta)$ and $\dot{\sigma}_n^2(\beta)$ substituting α_1 and σ^2 in (2.5)*, we can consider the two-step estimation of β.*

To eliminate possible asymptotic impact of $\check{\alpha}$ (or $\dot{\alpha}_n(\beta)$) on the estimation of β, we may use $C(\alpha)$-type moments as follows. Let $g_n(\gamma) = [g_{nb}'(\gamma), g_{na}'(\gamma)]'$, where $g_{nb}(\gamma)$ is $m_b \times 1$ and $g_{na}(\gamma)$ is $m_a \times 1$ such that $m_g = m_b + m_a$.[4] Consider the function $g_{nb}(\gamma) - \bar{C}_{1n}g_{na}(\gamma) = \bar{C}_n g_n(\gamma)$ of m_b moments, where $\bar{C}_{1n}$ is an $m_b \times m_a$ nonstochastic matrix with bounded elements and $\bar{C}_n = [I_{m_b}, -\bar{C}_{1n}]$ with I_{m_b} being the $m_b \times m_b$ identity matrix. We can show by the mean value theorem that $\sqrt{n}\bar{C}_n g_n(\check{\alpha}, \beta_0)$ has the same asymptotic distribution as that of $\sqrt{n}\bar{C}_n g_n(\gamma_0)$ if $\bar{C}_n\bar{G}_{n\alpha} = \bar{G}_{nb\alpha} - \bar{C}_{1n}\bar{G}_{na\alpha} = 0$ and $\check{\alpha} = \alpha_0 + O_p(n^{-1/2})$, where $\bar{G}_{n\alpha} = \mathbf{E}(\frac{\partial g_n(\gamma_0)}{\partial \alpha'})$, $\bar{G}_{nb\alpha} = \mathbf{E}(\frac{\partial g_{nb}(\gamma_0)}{\partial \alpha'})$, and $\bar{G}_{na\alpha} = \mathbf{E}(\frac{\partial g_{na}(\gamma_0)}{\partial \alpha'})$ is assumed to have full column rank.[5] Such a $\bar{C}_{1n}$ exists if $\text{rank}(\bar{\mathrm{G}}_{\mathrm{n}\alpha}) = \text{rank}(\bar{\mathrm{G}}_{\mathrm{na}\alpha})$. An example of $\bar{C}_{1n}$ is $\bar{G}_{nb\alpha}(\bar{G}_{na\alpha}'\bar{\Omega}_{naa}^{-1}\bar{G}_{na\alpha})^{-1}\bar{G}_{na\alpha}'\bar{\Omega}_{naa}^{-1}$, where $\bar{\Omega}_{naa} = n\mathbf{E}[g_{na}(\gamma_0)g_{na}'(\gamma_0)]$. There

[4] See Appendix A for a list of notations used in this paper.

[5] With $\bar{C}_n\bar{G}_{n\alpha} = 0$, by a first order Taylor expansion of $\sqrt{n}\bar{C}_n g_n(\check{\alpha}, \beta_0)$ at $\alpha = \alpha_0$, $\sqrt{n}\bar{C}_n g_n(\check{\alpha}, \beta_0)$ has the same asymptotic distribution as that of $\sqrt{n}\bar{C}_n g_n(\gamma_0)$ as long as $\check{\alpha} = \alpha_0 + o_p(n^{-1/4})$. Thus, it is possible to allow for the convergence rate of $\check{\alpha}$ to be slower than $\sqrt{n}$ but faster than $n^{1/4}$. We focus on the usual case with the $\sqrt{n}$-convergence rate of $\check{\alpha}$ in this paper.

are m_b moments in $\bar{C}_n g_n(\gamma)$, thus the number of moments is reduced by m_a from m_g. In the special case that $m_a = k_\alpha$, we have $\bar{C}_{1n} = \bar{G}_{nb\alpha}\bar{G}_{na\alpha}^{-1}$. For the following analyses, we assume that $m_a \geq k_\alpha$ and there exists a $\bar{C}_n$ such that $\bar{C}_n \bar{G}_{n\alpha} = 0$. However, the case $m_a = k_\alpha$ will be of special interest.

The transformation matrix $\bar{C}_n$ might involve unknown parameters α_0 and β_0. Let $C_n(\alpha,\beta)$ be a function of α and β such that $C_n(\alpha,\beta)$ at consistent estimators of α_0 and β_0 is a consistent estimator of $\lim_{n\to\infty}\bar{C}_n$. If expectations in $\bar{C}_n$ have closed forms, then we may simply let $C_n(\gamma) = \bar{C}_n(\gamma)$, where $\bar{C}_n(\gamma)$ is the matrix obtained by replacing γ_0 in $\bar{C}_n$ with γ; otherwise, relevant expectations can be estimated by corresponding sample averages. For example, when $m_a = k_\alpha$, $C_n(\gamma)$ can be $[I_{m_b}, -G_{nb\alpha}(\gamma)G_{na\alpha}^{-1}(\gamma)]$, where $G_{nb\alpha}(\gamma) = \frac{\partial g_{nb}(\gamma)}{\partial \alpha'}$ and $G_{na\alpha}(\gamma) = \frac{\partial g_{na}(\gamma)}{\partial \alpha'}$. We consider the following TGEL estimators[6]

$$\hat{\beta}_{\text{TGEL}} = \arg\min_{\beta\in\mathcal{B}} \max_{\mu\in\Lambda_{nd}(\check{\alpha},\beta)} \sum_{i=1}^{n} \rho\Big(\mu' d_{ni}(\check{\alpha},\beta)\Big), \tag{2.6}$$

and

$$\hat{\beta}_{\text{TGEL2}} = \arg\min_{\beta\in\mathcal{B}} \sup_{\mu\in\Lambda_{nd}(\dot{\alpha}_n(\beta),\beta)} \sum_{i=1}^{n} \rho\Big(\mu' d_{ni}(\dot{\alpha}_n(\beta),\beta)\Big), \tag{2.7}$$

where $d_{ni}(\gamma) = C_n(\gamma)g_{ni}(\gamma)$, $\Lambda_{nd}(\gamma) = \{\mu : \mu' d_{ni}(\gamma) \in \mathcal{V}, i = 1,\ldots,n\}$ for an open interval $\mathcal{V}$ containing 0, and $\rho(v)$ is a twice continuously differentiable concave function of a scalar v on $\mathcal{V}$. The TGEL estimators does not involve the estimation of any variance.[7] Denote $\rho_k(v) = \frac{\partial^k \rho(v)}{\partial v^k}$ for $k = 1,2$. We may let $\rho_1(0) = \rho_2(0) = -1$ without loss of generality, as long as $\rho_1(0) \neq 0$ and

[6] Alternatively, we may first derive a consistent but perhaps inefficient estimator $\check{\beta}$ of β_0 and use the moment $\hat{C}_n(\check{\alpha},\check{\beta})g_{ni}(\check{\alpha},\beta)$ instead of $C_n(\check{\alpha},\beta)g_{ni}(\check{\alpha},\beta)$ for estimation. Then an asymptotically equivalent TGEL estimator can be derived. Using $\hat{C}_n(\check{\alpha},\check{\beta})g_{ni}(\check{\alpha},\beta)$ involves an additional estimation step of β. These two moments also differ in terms of identification conditions. We focus on the TGEL estimator using $C_n(\check{\alpha},\beta)g_{ni}(\check{\alpha},\beta)$ in the main text, and investigate the other one in the supplementary material. The same comment applies to the case with $\dot{\alpha}_n(\beta)$ for $C_n(\dot{\alpha}_n(\beta),\beta)$.

[7] We may also consider optimal two-step GMM (TGMM) estimators, which require a consistent estimator of the variance of $\frac{1}{\sqrt{n}}\sum_{i=1}^{n} g_{ni}(\gamma_0)$ to construct an optimal weighting matrix. As $g_{ni}(\gamma_0)$'s are MDs, the variance of $\frac{1}{\sqrt{n}}\sum_{i=1}^{n} g_{ni}(\gamma_0)$ is $\bar{\Omega}_n = \frac{1}{n}\sum_{i=1}^{n}\mathbf{E}[g_{ni}(\gamma_0)g_{ni}'(\gamma_0)]$. With $\check{\alpha}$ and a consistent estimator $\check{\beta}$ of β_0, $\bar{\Omega}_n$ can be estimated as $\Omega_n(\check{\alpha},\check{\beta}) = \frac{1}{n}\sum_{i=1}^{n} g_{ni}(\check{\alpha},\check{\beta})g_{ni}'(\check{\alpha},\check{\beta})$. Denote $d_n(\gamma) = \hat{C}_n(\gamma)g_n(\gamma)$ and $\Omega_{nd}(\gamma) = \hat{C}_n(\gamma)\Omega_n(\gamma)\hat{C}_n'(\gamma)$. The proper optimal TGMM estimators of β to be compared with $\hat{\beta}_{\text{TGEL}}$ and $\hat{\beta}_{\text{TGEL2}}$ are, respectively, $\hat{\beta}_{\text{TGMM}} = \arg\min_{\beta\in\mathcal{B}} d_n'(\check{\alpha},\beta)\Omega_{nd}^{-1}(\check{\alpha},\check{\beta})d_n(\check{\alpha},\beta)$ and $\hat{\beta}_{\text{TGMM2}} = \arg\min_{\beta\in\mathcal{B}} d_n'(\dot{\alpha}_n(\beta),\beta)\Omega_{nd}^{-1}(\dot{\alpha}_n(\check{\beta}),\check{\beta})d_n(\dot{\alpha}_n(\beta),\beta)$.

$\rho_2(0) < 0$ (Newey and Smith, 2004). The class of GEL estimators include the EL estimator (Qin and Lawless, 1994; Smith, 1997), the ET estimator (Kitamura and Stutzer, 1997; Smith, 1997), and the CU estimator (Newey and Smith, 2004), which have, respectively, $\rho(v) = \ln(1-v)$ for $v < 1$, $\rho(v) = -e^v$ and $\rho(v) = -\frac{1}{2}(v+1)^2$. The TGEL estimators of the auxiliary vector μ corresponding to $\hat{\beta}_{\text{TGEL}}$ and $\hat{\beta}_{\text{TGEL2}}$ are, respectively, $\hat{\mu}_{\text{TGEL}} = \arg\max_{\mu\in\Lambda_{nd}(\check{\alpha},\hat{\beta}_{\text{TGEL}})} \sum_{i=1}^n \rho\Big(\mu' d_{ni}(\check{\alpha},\hat{\beta}_{\text{TGEL}})\Big)$, and

$$\hat{\mu}_{\text{TGEL2}} = \arg \max_{\mu\in\Lambda_{nd}(\dot{\alpha}_n(\hat{\beta}_{\text{TGEL2}}),\hat{\beta}_{\text{TGEL2}})} \sum_{i=1}^n \rho\Big(\mu' d_{ni}(\dot{\alpha}_n(\hat{\beta}_{\text{TGEL2}}),\hat{\beta}_{\text{TGEL2}})\Big).$$

A two-step GEL estimator has been considered in Newey et al. (2005) as

$$\arg\min_{\beta\in\mathcal{B}} \sup_{\lambda\in\Lambda_{ng}(\check{\alpha},\beta)} \sum_{i=1}^n \rho\Big(\lambda' g_{ni}(\check{\alpha},\beta)\Big),$$

where $\Lambda_{ng}(\alpha,\beta) = \{\lambda : \lambda' g_{ni}(\alpha,\beta) \in \mathcal{V}, i = 1,\ldots,n\}$. It is asymptotically equivalent to the GMM estimator $\arg\min_{\beta\in\mathcal{B}} g_n'(\check{\alpha},\beta)\Omega_n^{-1}(\tilde{\gamma})g_n(\check{\alpha},\beta)$, where $\Omega_n(\gamma) = \frac{1}{n}\sum_{i=1}^n g_{ni}(\gamma)g_{ni}'(\gamma)$ and $\tilde{\gamma}$ is an initial consistent estimator of γ, due to the self-normalization property of the GEL.[8] Since $\check{\alpha}$ is generally inefficient and has an impact on the asymptotic variance of estimates of β, the usual variance formula cannot be used and a correction is needed. An inefficient estimate $\check{\alpha}$ leads to the inefficiency of the second-step estimate of β. Also their two-step GEL objective function cannot be directly used to construct asymptotically pivotal tests either. As shown below, our proposed TGEL objective function and estimators are designed to overcome those issues.

For comparison purposes, we present also the ordinary GEL estimator, which is

$$\hat{\gamma}_{\text{GEL}} = \arg\min_{\gamma\in\Gamma} \max_{\lambda\in\Lambda_{ng}(\gamma)} \sum_{i=1}^n \rho\Big(\lambda' g_{ni}(\gamma)\Big), \tag{2.8}$$

where Γ is the parameter space of γ. The TGEL estimator (2.6) is computationally simpler than $\hat{\gamma}_{\text{GEL}}$ since there are fewer auxiliary parameters in μ than in λ and fewer parameters in β to be estimated than in γ.

[8] It can be shown as for the ordinary GEL that, with moments $g_{ni}(\check{\alpha},\beta)$ for $i = 1,\ldots,n$, the leading order term for the GEL estimator $\arg\min_{\beta\in\mathcal{B}} \sup_{\lambda\in\Lambda_{ng}(\check{\alpha},\beta)} \sum_{i=1}^n \rho\Big(\lambda' g_{ni}(\check{\alpha},\beta)\Big)$ is the same as that for a GMM estimator with the empirical moment $g_n(\check{\alpha},\beta)$ and the weighting matrix $[\frac{1}{n}\sum_{i=1}^n g_{ni}(\tilde{\gamma})g_{ni}'(\tilde{\gamma})]^{-1}$. So GEL employs $[\frac{1}{n}\sum_{i=1}^n g_{ni}(\tilde{\gamma})g_{ni}'(\tilde{\gamma})]^{-1}$ as the weighting internally, but that is not proper as $\check{\alpha}$ has an impact on the asymptotic variance of the moment $\sqrt{n}g_n(\check{\alpha},\beta_0)$.

Let $\bar{\Omega}_n(\gamma) = \frac{1}{n}\sum_{i=1}^n \mathbf{E}[g_{ni}(\gamma)g'_{ni}(\gamma)]$, $\bar{\Omega}_n = \bar{\Omega}_n(\gamma_0)$, $G_n(\gamma) = \frac{\partial g_n(\gamma)}{\partial \gamma'}$, $\bar{G}_n(\gamma) = \mathbf{E}[G_n(\gamma)]$, $\bar{G}_n = \bar{G}_n(\gamma_0)$ and $\bar{G}_{n\beta} = \mathbf{E}(\frac{\partial g_n(\gamma_0)}{\partial \beta'})$. The following regularity conditions are required for our asymptotic analysis on the consistency and asymptotic distributions of considered estimators.

Assumption 1. *(i)* $\mathbf{E}[g_n(\gamma_0)] = 0$ *and* $\gamma_0 \in \Gamma$ *is the unique solution to* $\lim_{n\to\infty} \mathbf{E}[g_n(\gamma)] = 0$*;*[9] *(ii)* Γ *is compact; (iii)* $g_{ni}(\gamma)$ *is continuous at each* $\gamma \in \Gamma$ *with probability one; (iv)* $\sup_{\gamma\in\Gamma} \|g_n(\gamma) - \mathbf{E}[g_n(\gamma)]\| = o_p(1)$ *and* $\mathbf{E}[g_n(\gamma)]$ *is continuous on* Γ *uniformly in n; (v)* $\sup_{\gamma\in\Gamma} \|g_{ni}(\gamma)\|^{\eta} \leq b_{ni}$ *for some* $\eta > 2$ *and* b_{ni} *with* $\frac{1}{n}\sum_{i=1}^n \mathbf{E}(b_{ni}) = O(1)$*; (vi)* $\sup_{\gamma\in\mathcal{N}} \|\Omega_n(\gamma) - \bar{\Omega}_n(\gamma)\| = o_p(1)$ *for a neighborhood* $\mathcal{N}$ *of* γ_0*,* $\bar{\Omega}_n(\gamma)$ *is continuous on* $\mathcal{N}$ *uniformly in n, and* $\lim_{n\to\infty} \bar{\Omega}_n$ *is nonsingular; (vii)* $\rho(v)$ *is concave on* $\mathcal{V}$*, twice continuously differentiable in a neighborhood of zero, and* $\rho_1(0) = \rho_2(0) = -1$*; (viii)* $\gamma_0 \in \mathrm{int}(\Gamma)$*; (ix)* $g_{ni}(\gamma)$ *is differentiable on* $\mathcal{N}$*,* $\sup_{\gamma\in\mathcal{N}} \|G_n(\gamma) - \bar{G}_n(\gamma)\| = o_p(1)$*,* $\bar{G}_n(\gamma)$ *is continuous on* $\mathcal{N}$ *uniformly in n, and* $\sup_{\gamma\in\mathcal{N}} \|\frac{\partial g_{ni}(\gamma)}{\partial \gamma'}\| \leq b_{ni}$ *for some* b_{ni} *with* $\frac{1}{n}\sum_{i=1}^n \mathbf{E}(b_{ni}) = O(1)$*; (x)* $\mathrm{rank}(\lim_{n\to\infty} \bar{G}_n) = k_\gamma$*; (xi)* $g_{ni}(\gamma_0)$*'s are MDs with respect to an increasing* σ*-field so that* $\sqrt{n} g_n(\gamma_0) \xrightarrow{d} N(0, \lim_{n\to\infty} \bar{\Omega}_n)$ *by a central limit theorem (CLT) for MD arrays.*

Assumption 2. *(i) There exists a nonstochastic* $m_b \times m_a$ *matrix* $\bar{C}_{1n}$ *such that* $\lim_{n\to\infty} \bar{C}_{1n}$ *exists and* $\bar{C}_n \bar{G}_{n\alpha} = 0$*, where* $\bar{C}_n = [I_{m_b}, -\bar{C}_{1n}]$ *and* $m_a \geq k_\alpha$*; (ii) for the case with* $\check{\alpha}$*,* $\sup_{\alpha\in\mathcal{N}_\alpha, \beta\in\mathcal{B}} \|C_n(\alpha,\beta) - \bar{C}_n(\alpha,\beta)\| = o_p(1)$ *and* $\bar{C}_n(\alpha,\beta)$ *is continuous on* $\mathcal{N}_\alpha \times \mathcal{B}$ *uniformly in n, where* $\mathcal{N}_\alpha$ *is a neighborhood of* α_0*; for the case with* $\dot{\alpha}_n(\beta)$*,* $\sup_{\gamma\in\Gamma} \|C_n(\gamma) - \bar{C}_n(\gamma)\| = o_p(1)$ *and* $\bar{C}_n(\gamma)$ *is continuous on* Γ *uniformly in n; (iii) for the case with* $\check{\alpha}$*,* $\check{\alpha} \in \mathcal{A}$*; for the case with* $\dot{\alpha}_n(\beta)$*,* $\dot{\alpha}_n(\beta)$ *is in the convex parameter space* $\mathcal{A}$ *of* α*; (iv) for the case with* $\check{\alpha}$*,* $\check{\alpha} = \alpha_0 + O_p(n^{-1/2})$*; for the case with* $\dot{\alpha}_n(\beta)$*,* $\dot{\alpha}_n(\beta_0) - \alpha_0 = O_p(n^{-1/2})$ *and there is some nonstochastic function* $\alpha_n(\beta)$ *of* β *such that* $\alpha_n(\beta) \in \mathcal{A}$ *for* $\beta \in \mathcal{B}$*,* $\sup_{\beta\in\mathcal{B}} \|\dot{\alpha}_n(\beta) - \alpha_n(\beta)\| = o_p(1)$*,* $\alpha_n(\beta)$ *is continuous uniformly in n and* $\lim_{n\to\infty} \alpha_n(\beta_0) = \alpha_0$*; (v)* β_0 *is the unique solution to* $\lim_{n\to\infty} \bar{C}_n(\alpha_0,\beta) \mathbf{E}[g_n(\alpha_0,\beta)] = 0$ *and* $\lim_{n\to\infty} \bar{C}_n(\alpha_n(\beta),\beta) \mathbf{E}[g_n(\alpha_n(\beta),\beta)] = 0$*; (vi) for the case with* $\check{\alpha}$*,* $C_n(\alpha,\beta)$ *is differentiable with respect to* β *on a neighborhood* $\mathcal{N}_\beta$ *of* β_0 *and* $\sup_{\gamma\in\mathcal{N}, 1\leq j\leq k_\beta} \|\frac{\partial C_n(\gamma)}{\partial \beta_j}\| = O_p(1)$*; for the case with* $\dot{\alpha}_n(\beta)$*,* $C_n(\gamma)$ *is differentiable on* $\mathcal{N}$*,* $\dot{\alpha}(\beta)$ *is differentiable on* $\mathcal{N}_\beta$*,* $\sup_{\gamma\in\mathcal{N}, 1\leq j\leq k_\gamma} \|\frac{\partial C_n(\gamma)}{\partial \gamma_j}\| = O_p(1)$ *and* $\sup_{\beta\in\mathcal{N}_\beta, 1\leq j\leq k_\beta} \|\frac{\partial \dot{\alpha}_n(\beta)}{\partial \beta_j}\| = O_p(1)$*; (vii)* $\mathrm{rank}(\lim_{n\to\infty} \bar{C}_n \bar{G}_{n\beta}) = k_\beta$*.*

Assumption 1 is for the ordinary GEL estimator, while the additional Assumption 2 is for

[9] The existence of the limit of $\mathbf{E}[g_n(\gamma)]$ is implicitly assumed in the expression $\lim_{n\to\infty} \mathbf{E}[g_n(\gamma)]$. This also applies to other expressions in the paper where limits are taken.

TGEL estimators. Conditions in Assumption 1 extend those in Newey and Smith (2004) for i.i.d data to allow for triangular arrays, where the moments at the true parameters are MD arrays. Since we have only assumed that $g_{ni}(\gamma_0)$'s at the true γ_0 are MDs, high level regularity conditions such as uniform convergence and continuity are imposed. For some specific models, more primitive conditions can be derived. For example, primitive conditions for GMM and GEL estimation of SAR models are given in Lee (2007) and Jin and Lee (2019). A CLT for MD arrays in Assumption 1(xi) can be found in, e.g., Gänsler and Stute (1977) and Hall and Heyde (1980). For SAR models, a CLT is derived in Kelejian and Prucha (2001) for linear-quadratic forms of disturbances.

Assumptions 2(ii)–(iv) list some basic conditions on $C_n(\gamma)$, $\check{\alpha}$ and $\dot{\alpha}_n(\beta)$. Assumption 2(ii) is needed for the consistency of the TGEL estimator. It may be verified by a proper law of large numbers (LLN) in specific models, since $C_n(\gamma)$ can be $\bar{C}_n(\gamma)$ or estimated by using sample averages for relevant expectations. Similarly, as $\dot{\alpha}_n(\beta)$ is a function of sample observations, an LLN might be applicable and it is plausible to assume some uniform convergence and continuity conditions on $\dot{\alpha}_n(\beta)$.[10] With $\check{\alpha}$, even if $\lim_{n\to\infty}\mathbf{E}[g_n(\alpha_0,\beta)]$ is uniquely zero at $\beta=\beta_0$, $\lim_{n\to\infty}\bar{C}_n(\alpha_0,\beta)\mathbf{E}[g_n(\alpha_0,\beta)]$ might not be so as the transformation reduces the number of moments. As an example, consider the estimation of the SAR model (2.2) with $g_n(\gamma)=[g_{nb}(\gamma),g_{na}(\gamma)]'$, where $g_{nb}(\gamma)=\frac{1}{n}V_n'(\alpha_1,\beta)W_nV_n(\alpha_1,\beta)$ and

$$g_{na}(\gamma)=\frac{1}{n}[V_n'(\alpha_1,\beta)V_n(\alpha_1,\beta)-n\sigma^2,V_n'(\alpha_1,\beta)Q_n]'$$

with Q_n being an $n\times k_q$ matrix for some $k_q\geq k_x$. In this special case, as $\mathbf{E}\frac{\partial g_{nb}(\gamma_0)}{\partial\alpha'}=0$, we may let $\bar{C}_{1n}=0$. Then $\lim_{n\to\infty}\bar{C}_n(\alpha_0,\beta)\mathbf{E}[g_n(\alpha_0,\beta)]=\lim_{n\to\infty}\mathbf{E}[g_{nb}(\alpha_0,\beta)]=\lim_{n\to\infty}\frac{\sigma_0^2}{n}(\beta_0-\beta)\mathrm{tr}[W_n(T_n+T_n')]+\lim_{n\to\infty}\frac{1}{n}(\beta_0-\beta)^2[(T_nX_n\alpha_{10})'W_n(T_nX_n\alpha_{10})+\sigma_0^2\,\mathrm{tr}(T_n'W_nT_n)]$, which can be zero at $\beta\neq\beta_0$, where $T_n=W_n(I_n-\beta_0W_n)^{-1}$. On the other hand, $\frac{1}{n}\mathbf{E}[g_n(\gamma)]$ can be uniquely zero at $\gamma=\gamma_0$. For example, when $\lim_{n\to\infty}\frac{1}{n}Q_n'[X_n,T_nX_n\alpha_{10}]$ has full column rank, $\lim_{n\to\infty}Q_n'V_n(\alpha_1,\beta)$ is uniquely

[10] As an example, consider the case that $g_{na}(\alpha,\beta)$ is $k_\alpha\times 1$ and $\dot{\alpha}_n(\beta)$ is the solution to $g_{na}(\alpha,\beta)=0$. If $\mathbf{E}[g_{na}(\alpha_0,\beta_0)]=0$, under regularity conditions, $\mathbf{E}[g_{na}(\alpha,\beta)]=0$ yields a solution $\alpha_n(\beta)$, which is a continuously differentiable function of β and satisfies $\alpha_n(\beta_0)=\alpha_0$. we can show that $\sup_{\beta\in\mathcal{B}}\|\dot{\alpha}_n(\beta)-\alpha_n(\beta)\|=o_p(1)$, and $\dot{\alpha}_n(\beta_0)=\alpha_0+O_p(n^{-1/2})$ by expanding $0=g_{na}(\dot{\alpha}_n(\beta_0),\beta_0)$ at (α_0,β_0). For the SAR model (2.2), with $\dot{\alpha}_n(\beta)=[\dot{\alpha}_{1n}'(\beta),\dot{\sigma}_n^2(\beta)]'$, where $\dot{\alpha}_{1n}(\beta)$ and $\dot{\sigma}_n^2(\beta)$ have explicit expressions in, respectively, (2.3) and (2.4), it is easy to show that Assumption 2(iv) is satisfied.

zero at $(\alpha_1,\beta)=(\alpha_{10},\beta_0)$, which implies that $\frac{1}{n}\mathbf{E}[g_n(\gamma)]$ is uniquely zero at $\gamma=\gamma_0$. In this example, the transformation of $g_n(\gamma)$ results in the loss of too much information so that only a moment quadratic in β is left, from which β_0 is not identified. We may move some linear moments from $g_{na}(\gamma)$ to $g_{nb}(\gamma)$ or add more quadratic moments to $g_{nb}(\gamma)$ to achieve the identification of β_0 when using transformed moments. It is also possible that the transformation of moments leads to weakened identification, which may result in worse finite sample performance, but it is beyond the scope of this paper. In the case that there is an identification issue, the penalization approach proposed in Frazier and Renault (2017) can be used.[11] For simplicity, we assume the identification uniqueness in Assumption $2(v)$. The same comment applies to the case with $\dot{\alpha}_n(\beta)$. Assumptions $2(vi)$–(vii) are needed for the $\sqrt{n}$-rate convergence of TGEL estimators.

Denote $\bar{D}_{n\beta}=\bar{C}_n\bar{G}_{n\beta}$ and $\bar{\Omega}_{nd}=\bar{C}_n\bar{\Omega}_n\bar{C}_n'$. Then we have the following theorem on the asymptotic properties of $\hat{\beta}_{\text{TGEL}}$ and $\hat{\beta}_{\text{TGEL2}}$.

Theorem 1. *Suppose that Assumptions 1–2 are satisfied.*

(i) $\sqrt{n}(\hat{\beta}_{\text{TGEL}}-\beta_0)\xrightarrow{d}N\Big(0,\lim_{n\to\infty}(\bar{D}'_{n\beta}\bar{\Omega}^{-1}_{nd}\bar{D}_{n\beta})^{-1}\Big)$ *and* $\sqrt{n}(\hat{\beta}_{\text{TGEL2}}-\beta_0)\xrightarrow{d}N\Big(0,\lim_{n\to\infty}(\bar{D}'_{n\beta}\bar{\Omega}^{-1}_{nd}\bar{D}_{n\beta})^{-1}\Big)$.

(ii) $2\Big[\sum_{i=1}^n\rho\Big(\hat{\mu}'_{\text{TGEL}}d_{ni}(\check{\alpha},\hat{\beta}_{\text{TGEL}})\Big)-n\rho(0)\Big]\xrightarrow{d}\chi^2(m_b-k_\beta)$ *and* $2[\sum_{i=1}^n\rho(\hat{\mu}'_{\text{TGEL2}}d_{ni}(\dot{\alpha}_n(\hat{\beta}_{\text{TGEL2}}),\hat{\beta}_{\text{TGEL2}}))-n\rho(0)]\xrightarrow{d}\chi^2(m_b-k_\beta)$.

(iii) $\hat{\beta}_{\text{TGEL}}$ *and* $\hat{\beta}_{\text{TGEL2}}$ *are generally less efficient relative to* $\hat{\beta}_{\text{GEL}}$*, where* $\hat{\beta}_{\text{GEL}}$ *is the joint GEL estimator of* β*, which is a subvector of* $\hat{\gamma}_{\text{GEL}}$ *in (2.8). But if* $m_a=k_\alpha$*, then* $\hat{\beta}_{\text{TGEL}}$ *and* $\hat{\beta}_{\text{TGEL2}}$ *(will be denoted as* $\hat{\beta}_{\text{E-TGEL}}$ *and* $\hat{\beta}_{\text{E-TGEL2}}$*) are asymptotically equivalent to* $\hat{\beta}_{\text{GEL}}$.

(iv) If $m_a=k_\alpha$ *and* $\mathbf{E}(\sup_{\alpha\in\mathcal{A},\beta\in\mathcal{N}_\beta}\|\frac{\partial g_n(\gamma)}{\partial\beta'}\|)<\infty$ *for the parameter space* $\mathcal{A}$ *of* α *and a neighborhood* $\mathcal{N}_\beta$ *of* β_0*, then* $\hat{\alpha}_{\text{E-TGEL}}$ *and* $\hat{\alpha}_{\text{E-TGEL2}}$*, where*

$$\hat{\alpha}_{\text{E-TGEL}}=\arg\min_{\alpha\in\mathcal{A}}\sup_{\lambda\in\Lambda_{ng}(\alpha,\hat{\beta}_{\text{E-TGEL}})}\sum_{i=1}^n\rho\Big(\lambda'g_{ni}(\alpha,\hat{\beta}_{\text{E-TGEL}})\Big)$$

[11] An alternative to the penalization approach in Frazier and Renault (2017) is to use Newton iterations for the TGEL objective function by starting from the initial consistent estimate $\tilde{\beta}$. The resulting estimate in each iteration will be consistent and the estimate sequence converges to a critical point of the TGEL objective function. As pointed out by Frazier and Renault (2017), the approach in Trognon and Gouriéroux (1990), which we shall describe later, may also have a similar identification issue.

and $\hat{\alpha}_{\text{E-TGEL2}} = \arg\min_{\alpha\in\mathcal{A}} \sup_{\lambda\in\Lambda_{ng}(\alpha,\hat{\beta}_{\text{E-TGEL2}})} \sum_{i=1}^{n} \rho(\lambda' g_{ni}(\alpha,\hat{\beta}_{\text{E-TGEL2}}))$, *are asymptotically equivalent to the joint estimate* $\hat{\alpha}_{\text{GEL}}$ *of* α *from* (2.8).

With the $C(\alpha)$-type formulation, the TGEL estimators have an asymptotically normal distribution with a limiting variance formed by a usual Jacobian matrix of given moments and a weighting matrix, and do not involve any asymptotic variance of initial estimates of nuisance parameters. The TGEL objective functions can provide overidentification tests in Theorem 1(ii). The $C(\alpha)$-type transformation reduces the number of moments for the estimation of β by m_a, where $m_a \geq k_\alpha$, in order to eliminate the asymptotic impact of the $k_\alpha \times 1$ dimensional estimate $\check{\alpha}$ or $\dot{\alpha}_n(\beta)$. The resulting TGEL estimates might not be efficient relative to the joint estimator $\hat{\beta}_{\text{GEL}}$. However, in the case that $m_a = k_\alpha$, $\hat{\beta}_{\text{TGEL}}$ and $\hat{\beta}_{\text{TGEL2}}$ do not lose asymptotic efficiency for the estimation of β, which is the parameter vector of interest. In the event that it is also desirable to have a relatively efficient estimate of the nuisance parameter vector α, then the efficient TGEL estimates $\hat{\beta}_{\text{E-TGEL}}$ and $\hat{\beta}_{\text{TGEL2}}$ may be plugged back into the original GEL objective functions to obtain second round estimates of α, which turn out to be asymptotically as efficient as the joint GEL estimate of α. For our results on GEL estimators in Theorem 1(iii)–(iv) with $m_a = k_\alpha$, the nuisance parameter vector α in $g_n(\alpha,\beta)$ can be replaced by any $\sqrt{n}$-consistent estimator $\check{\alpha}$ or any estimating function $\dot{\alpha}_n(\beta)$ satisfying regularity conditions, while in Crepon et al. (1997), α is replaced by $\hat{\alpha}_n(\beta)$, where $\hat{\alpha}_n(\beta)$ is the unique solution of $g_{na}(\alpha,\beta) = 0$ given β.

We note that our efficient TGEL estimators are asymptotically equivalent to a GMM estimator proposed in Trognon and Gouriéroux (1990). Consider the case with $\check{\alpha}$ as an example. Applying their method to the OGMM objective function $g_n'(\alpha,\beta)\Omega_n^{-1}(\check{\alpha},\check{\beta})g_n(\alpha,\beta)$, where $\check{\beta}$ is a consistent estimator of β_0, we derive the following objective function for a two-step estimator $[\alpha^*,\beta^*]$ of $[\alpha_0,\beta_0]$:

$$[g_n(\check{\alpha},\beta) + G_{n\alpha}(\check{\alpha},\beta)(\alpha-\check{\alpha})]'\Omega_n^{-1}(\check{\alpha},\check{\beta})[g_n(\check{\alpha},\beta) + G_{n\alpha}(\check{\alpha},\beta)(\alpha-\check{\alpha})], \tag{2.9}$$

where $G_{n\alpha}(\alpha,\beta) = \frac{\partial g_n(\alpha,\beta)}{\partial\alpha'}$.[12] This objective function is derived by a first order Taylor expansion of $g_n(\alpha,\beta)$ at $\alpha = \check{\alpha}$. For given β, the closed form solution of α is

$$\alpha^*(\beta) = \check{\alpha} - [G_{n\alpha}'(\check{\alpha},\beta)\Omega_n^{-1}(\check{\alpha},\check{\beta})G_{n\alpha}(\check{\alpha},\beta)]^{-1}G_{n\alpha}'(\check{\alpha},\beta)\Omega_n^{-1}(\check{\alpha},\check{\beta})g_n(\check{\alpha},\beta).$$

[12] $G_n(\check{\alpha},\beta)$ in (2.9) can be replaced by $G_n(\check{\alpha},\check{\beta})$ and an asymptotically equivalent estimator can be derived. See also Frazier and Renault (2017).

Substituting this expression into (2.9) yields the objective function for β^*:

$$g_n'(\check{\alpha},\beta)M_{n\alpha}(\beta)g_n(\check{\alpha},\beta), \tag{2.10}$$

where $M_{n\alpha}(\beta)=\Omega_n^{-1}(\check{\alpha},\check{\beta})-\Omega_n^{-1}(\check{\alpha},\check{\beta})G_{n\alpha}(\check{\alpha},\beta)[G_{n\alpha}'(\check{\alpha},\beta)\Omega_n^{-1}(\check{\alpha},\check{\beta})G_{n\alpha}(\check{\alpha},\beta)]^{-1}G_{n\alpha}'(\check{\alpha},\beta)\Omega_n^{-1}(\check{\alpha},\check{\beta})$. A two-step GMM (TGMM) estimator $\hat{\beta}_{\text{TGMM}}$, which is asymptotically equivalent to the TGEL estimator (see, e.g., Newey and Smith, 2004), has the objective function

$$g_n'(\check{\alpha},\beta)C_n'(\check{\alpha},\beta)[C_n(\check{\alpha},\check{\beta})\Omega_n(\check{\alpha},\check{\beta})C_n'(\check{\alpha},\check{\beta})]^{-1}C_n(\check{\alpha},\beta)g_n(\check{\alpha},\beta). \tag{2.11}$$

Assume that $C_n(\check{\alpha},\beta)G_{n\alpha}(\check{\alpha},\beta)=0$ for this comparison of our TGEL with the GMM in Trognon and Gouriéroux (1990). Then

$$[\Omega_n^{1/2}(\check{\alpha},\check{\beta})C_n'(\check{\alpha},\beta)]'\Omega_n^{-1/2}(\check{\alpha},\check{\beta})G_{n\alpha}(\check{\alpha},\beta)=C_n(\check{\alpha},\beta)G_{n\alpha}(\check{\alpha},\beta)=0$$

and $\operatorname{rank}([\Omega_n^{1/2}(\check{\alpha},\check{\beta})C_n'(\check{\alpha},\beta),\Omega_n^{-1/2}(\check{\alpha},\check{\beta})G_{n\alpha}(\check{\alpha},\beta)])=m_b+k_\alpha\le m_g$ as $m_a\ge k_\alpha$. Thus, $M_{n\alpha}(\beta)=\Omega_n^{-1/2}(\check{\alpha},\check{\beta})\cdot\Omega_n^{1/2}(\check{\alpha},\check{\beta})M_{n\alpha}(\beta)\Omega_n^{1/2}(\check{\alpha},\check{\beta})\cdot\Omega_n^{-1/2}(\check{\alpha},\check{\beta})\ge P_{n\alpha}(\beta)$, where

$$P_{n\alpha}(\beta)=\Omega_n^{-1/2}(\check{\alpha},\check{\beta})\cdot\Omega_n^{1/2}(\check{\alpha},\check{\beta})C_n'(\check{\alpha},\beta)[C_n(\check{\alpha},\beta)\Omega_n(\check{\alpha},\check{\beta})C_n'(\check{\alpha},\beta)]^{-1}C_n(\check{\alpha},\beta)\Omega_n^{1/2}(\check{\alpha},\check{\beta})\cdot\Omega_n^{-1/2}(\check{\alpha},\check{\beta}),$$

by the decomposition of projection in (3.25) of Ruud (2000), and

$$g_n'(\check{\alpha},\beta)M_{n\alpha}(\beta)g_n(\check{\alpha},\beta)\ge g_n'(\check{\alpha},\beta)P_{n\alpha}(\beta)g_n(\check{\alpha},\beta).$$

If $m_a=k_\alpha$, then $g_n'(\check{\alpha},\beta)M_{n\alpha}(\beta)g_n(\check{\alpha},\beta)=g_n'(\check{\alpha},\beta)P_{n\alpha}(\beta)g_n(\check{\alpha},\beta)$. The objective function in (2.11) differs from $g_n'(\check{\alpha},\beta)P_{n\alpha}(\beta)g_n(\check{\alpha},\beta)$ only in that the optimal weighting matrix does not involve unknown β, as the optimal GMM vs CU. Hence, the E-TGEL estimator $\hat{\beta}_{\text{E-TGEL}}$ is asymptotically equivalent to β^*. By plugging $\hat{\beta}_{\text{E-TGEL}}$ back into the original GEL objective function, we may also derive an estimator of α_0 that is asymptotically equivalent to α^*.

We next study higher order asymptotic biases of $\hat{\beta}_{\text{TGEL}}$ and $\hat{\beta}_{\text{TGEL2}}$ based on the Nagar-type expansion (Nagar, 1959) of an estimator $\hat{\beta}$:

$$\sqrt{n}(\hat{\beta}-\beta_0)=\psi_{n\beta}+n^{-1/2}\varphi_{n\beta}+O_p(n^{-1}),$$

where $\mathbf{E}(\psi_{n\beta})=0$, $\psi_{n\beta}=O_p(1)$ and $\varphi_{n\beta}=O_p(1)$. The higher order bias of $\hat{\beta}$ is computed as $\frac{1}{n}\mathbf{E}(\varphi_{n\beta})$. Newey and Smith (2004) show that the ordinary GEL can remove several bias terms of the ordinary feasible OGMM. For the TGEL and TGMM estimators, we expect $\check{\alpha}$, $\dot{\alpha}_n(\beta)$ and the estimation of $\bar{C}_n$ to result in additional higher order bias terms. To investigate this, we make the following assumption.

Assumption 3. *(i) For the case with $\check{\alpha}$, $\sqrt{n}(\check{\alpha} - \alpha_0) = \psi_{n\check{\alpha}} + O_p(n^{-1/2}) = O_p(1)$; for the case with $\dot{\alpha}_n(\beta)$, $\dot{\alpha}_n(\beta)$ is twice differentiable and $\alpha_n(\beta)$ is differentiable in a neighborhood $\mathcal{N}_\beta$ of β_0 such that $\frac{\partial \dot{\alpha}_n(\beta_0)}{\partial \beta'} - \frac{\partial \alpha_n(\beta_0)}{\partial \beta'} = O_p(n^{-1/2})$ and $\sup_{\beta \in \mathcal{B}} \|\nabla^2 \dot{\alpha}_n(\beta)\| = O_p(1)$, where ∇^j denotes a vector of all possible partial derivatives of order j; (ii) $\sqrt{n}[C_n(\gamma_0) - \bar{C}_n] = \psi_{nC} + O_p(n^{-1/2})$, $\mathbf{E}(\|\psi_{nC}\|^2) = O(1)$, $\nabla \bar{C}_n(\gamma)$ and $\nabla^2 C_n(\gamma)$ exist on $\mathcal{N}$, $\nabla C_n(\gamma_0) - \nabla \bar{C}_n(\gamma_0) = O_p(n^{-1/2})$ and $\sup_{\gamma \in \mathcal{N}} \|\nabla^2 C_n(\gamma)\| = O_p(1)$; (iii) for $0 \leq j \leq 4$ and all z, $\nabla^j g_{ni}(\gamma)$ exists on $\mathcal{N}$, $\sup_{\gamma \in \mathcal{N}} \|\nabla^j g_{ni}(\gamma)\| \leq b_{ni}$ for some b_{ni} with $\frac{1}{n}\sum_{i=1}^n \mathbf{E}(b_{ni}^5) = O(1)$, $\frac{1}{n}\sum_{i=1}^n \nabla^k r_{ni}(\gamma_0) - \frac{1}{n}\sum_{i=1}^n \mathbf{E}[\nabla^k r_{ni}(\gamma_0)] = O_p(n^{-1/2})$ for $k = 1, 2$ and $r_{ni}(\gamma_0) = g_{ni}(\gamma_0)$, $g_{ni}(\gamma_0)g'_{ni}(\gamma_0)$, or $g_{ni}(\gamma_0)g'_{ni}(\gamma_0)g_{ni}(\gamma_0)$; (iv) $\rho(v)$ is three times continuously differentiable with Lipschitz third derivative in a neighborhood of zero.*

In Assumptions 3(i)–(ii), $\psi_{n\check{\alpha}}$ and ψ_{nC} being respectively leading order terms of $\sqrt{n}(\check{\alpha} - \alpha_0)$ and $\sqrt{n}[C_n(\gamma_0) - \bar{C}_n]$ are involved to derive the higher order bias of $\hat{\beta}_{\text{TGEL}}$, where $\psi_{n\check{\alpha}}$ and ψ_{nC} may be correlated in general. If $C_n(\gamma)$ is equal to $\bar{C}_n(\gamma)$, which can be the case when expectations in $\bar{C}_n$ have closed forms, then $\psi_{nC} = 0$; otherwise, $\psi_{nC} \neq 0$. For example, when $m_b = k_\alpha$ so that $\bar{C}_n = [I_{m_b}, -\bar{G}_{nb\alpha}\bar{G}_{na\alpha}^{-1}]$ and we take $C_n(\gamma) = [I_{m_b}, -G_{nb\alpha}(\gamma)G_{na\alpha}^{-1}(\gamma)]$, then $\psi_{nC} = [0, -(G_{nb\alpha} - \bar{G}_{nb\alpha})\bar{G}_{na\alpha}^{-1} + \bar{G}_{nb\alpha}\bar{G}_{na\alpha}^{-1}(G_{na\alpha} - \bar{G}_{na\alpha})\bar{G}_{na\alpha}^{-1}]$, where $G_{nb\alpha} = G_{nb\alpha}(\gamma_0)$ and $G_{na\alpha} = G_{na\alpha}(\gamma_0)$. Other regularity conditions in Assumption 3 such as smoothness conditions on $\dot{\alpha}_n(\beta)$, $C_n(\gamma)$, $g_{ni}(\gamma)$ and $\rho(v)$ are needed since Nagar-type expansions are based on higher order Taylor expansions.

Let $g_{ni}(\gamma_0) = g_{ni}$, $g_n = g_n(\gamma_0)$, $G_{ni\beta} = \frac{\partial g_{ni}(\gamma_0)}{\partial \beta'}$, $G_{n\beta} = \frac{\partial g_n(\gamma_0)}{\partial \beta'}$, $\bar{G}_{n\beta}^{(j)} = \mathbf{E}(\frac{\partial^2 g_{ni}(\gamma_0)}{\partial \gamma_j \partial \beta'})$, $g_{ni}^{(j)} = \frac{\partial g_{ni}(\gamma_0)}{\partial \gamma_j}$, $\bar{C}_n^{(j)} = \frac{\partial \bar{C}_n(\gamma_0)}{\partial \gamma_j}$, $\alpha_n^{(j)} = \frac{\partial \alpha_n(\beta_0)}{\partial \beta_j}$, $\bar{\Sigma}_{nd} = (\bar{D}'_{n\beta}\bar{\Omega}_{nd}^{-1}\bar{D}_{n\beta})^{-1}$, $\bar{H}_{nd} = \bar{\Sigma}_{nd}\bar{D}'_{n\beta}\bar{\Omega}_{nd}^{-1}$, $\bar{P}_{nd} = \bar{\Omega}_{nd}^{-1} - \bar{\Omega}_{nd}^{-1}\bar{D}_{n\beta}\bar{\Sigma}_{nd}\bar{D}'_{n\beta}\bar{\Omega}_{nd}^{-1}$, $\rho_3(v) = \frac{d^3\rho(v)}{dv^3}$, ψ_j be the jth element of any vector ψ, and $e_{k_\beta j}$ be the jth unit column vector of dimension k_β. Denote $\psi_{n\beta} = -\bar{H}_{nd}\sqrt{n}\bar{C}_n g_n$ and $\psi_{n\mu} = -\bar{P}_{nd}\sqrt{n}\bar{C}_n g_n$, which are leading order terms of, respectively, $\sqrt{n}(\hat{\beta}_{\text{TGEL}} - \beta_0)$ and $\sqrt{n}\hat{\mu}_{\text{TGEL}}$. In addition, let the leading order terms of $\sqrt{n}[\dot{\alpha}_n(\hat{\beta}_{\text{TGEL}}) - \alpha_0]$, $\sqrt{n}[C_n(\check{\alpha}, \hat{\beta}_{\text{TGEL}}) - \bar{C}_n]$ and $\sqrt{n}[C_n(\dot{\alpha}_n(\hat{\beta}_{\text{TGEL}}), \hat{\beta}_{\text{TGEL}}) - \bar{C}_n]$ be, respectively, $\psi_{n\dot{\alpha}}$, $\psi_{n\check{C}}$ and $\psi_{n\dot{C}}$, whose explicit expressions are given in Appendix A.

Theorem 2. *Suppose that Assumptions 1–3 are satisfied.*

(i) The bias of $\hat{\beta}_{\mathrm{TGEL}}$ is

$$\mathrm{Bias}(\hat{\beta}_{\mathrm{TGEL}}) = [B^{I}_{nd} + (B^{\Omega}_{nd} + \frac{\rho_3(0)}{2}\tilde{B}^{\Omega}_{nd}) + (B^{G}_{nd} - \tilde{B}^{G}_{nd})] + [B^{C-\beta}_{nd} + B^{C-g}_{nd} + B^{C-\Omega}_{nd} + B^{C-G}_{nd} + B^{\check{\alpha}}_{nd}],$$

where $B^{I}_{nd} = \bar{H}_{nd}\,\mathbf{E}(\bar{C}_n G_{n\beta}\bar{H}_{nd}\bar{C}_n g_n) - \frac{1}{2n}\bar{H}_{nd}\sum_{j=1}^{k_\beta}\bar{C}_n\bar{G}^{(k_\alpha+j)}_{n\beta}\bar{\Sigma}_{nd}e_{k_\beta j}$, $B^{\Omega}_{nd} = \bar{H}_{nd}\,\mathbf{E}(\bar{C}_n\Omega_n\bar{C}'_n\bar{P}_{nd}\bar{C}_n g_n)$, $\tilde{B}^{\Omega}_{nd} = \frac{1}{n^2}\sum_{i=1}^{n}\bar{H}_{nd}\bar{C}_n\,\mathbf{E}(g_{ni}g'_{ni}\bar{C}'_n\bar{P}_{nd}\bar{C}_n g_{ni})$, $B^{G}_{nd} = -\bar{\Sigma}_{nd}\,\mathbf{E}(G'_{n\beta}\bar{C}'_n\bar{P}_{nd}\bar{C}_n g_n)$,

$$\tilde{B}^{G}_{nd} = -\frac{1}{n^2}\bar{\Sigma}_{nd}\sum_{i=1}^{n}\mathbf{E}(G'_{ni\beta}\bar{C}'_n\bar{P}_{nd}\bar{C}_n g_{ni}),$$

$B^{C-\beta}_{nd} = \frac{1}{n}\bar{\Sigma}_{nd}\Big[\mathrm{tr}[\bar{C}^{(k_\alpha+1)}_n\bar{G}_{n\alpha}\,\mathbf{E}(\psi_{n\check{\alpha}}\psi'_{n\mu})],\ldots,\mathrm{tr}[\bar{C}^{(k_\alpha+k_\beta)}_n\bar{G}_{n\alpha}\,\mathbf{E}(\psi_{n\check{\alpha}}\psi'_{n\mu})]\Big]'$,

$$B^{C-g}_{nd} = -\frac{1}{n}\bar{H}_{nd}\,\mathbf{E}(\sqrt{n}\psi_{n\check{C}}g_n + \psi_{n\check{C}}\bar{G}_{n\alpha}\psi_{n\check{\alpha}} + \psi_{n\check{C}}\bar{G}_{n\beta}\psi_{n\beta}),$$

$$B^{C-\Omega}_{nd} = -\frac{1}{n}\bar{H}_{nd}\,\mathbf{E}[(\psi_{nC}\bar{\Omega}_n\bar{C}'_n + \bar{C}_n\bar{\Omega}_n\psi'_{nC})\psi_{n\mu}] - \frac{1}{n}\bar{H}_{nd}\sum_{j=1}^{k_\alpha}(\bar{C}^{(j)}_n\bar{\Omega}_n\bar{C}'_n + \bar{C}_n\bar{\Omega}_n\bar{C}^{(j)'}_n)\,\mathbf{E}(\psi_{n\mu}\psi_{n\check{\alpha}j}),$$

$B^{C-G}_{nd} = \frac{1}{n}\,\mathbf{E}(\bar{\Sigma}_{nd}\bar{G}'_{n\beta}\psi'_{n\check{C}}\psi_{n\mu})$, *and*

$$B^{\check{\alpha}}_{nd} = \frac{1}{n}\,\mathbf{E}\Big[\bar{\Sigma}_{nd}\sum_{j=1}^{k_\alpha}\bar{G}^{(j)'}_{n\beta}\bar{C}'_n\psi_{n\mu}\psi_{n\check{\alpha}j} - \bar{H}_{nd}\sqrt{n}\bar{C}_n(G_{n\alpha} - \bar{G}_{n\alpha})\psi_{n\check{\alpha}} - \frac{1}{2}\bar{H}_{nd}\sum_{j=1}^{k_\alpha}\bar{C}_n\bar{G}^{(j)}_{n\alpha}\psi_{n\check{\alpha}}\psi_{n\check{\alpha}j}$$
$$- \bar{H}_{nd}\sum_{j=1}^{k_\alpha}\bar{C}_n\bar{G}^{(j)}_{n\beta}\psi_{n\beta}\psi_{n\check{\alpha}j} - \frac{1}{n}\bar{H}_{nd}\sum_{j=1}^{k_\alpha}\sum_{i=1}^{n}\bar{C}_n\,\mathbf{E}(g^{(j)}_{ni}g'_{ni} + g_{ni}g^{(j)'}_{ni})\bar{C}'_n\psi_{n\mu}\psi_{n\check{\alpha}j}\Big].$$

(ii) The bias of $\hat{\beta}_{\mathrm{TGEL2}}$ is

$$\mathrm{Bias}(\hat{\beta}_{\mathrm{TGEL2}}) = \widetilde{\mathrm{Bias}}(\hat{\beta}_{\mathrm{TGEL}}) + B^{C-\dot{\alpha}}_{nd} + (B^{\dot{\alpha}-G_\alpha}_{nd} - \tilde{B}^{\dot{\alpha}-G_\alpha}_{nd}) + B^{\dot{\alpha}-G_\alpha-C}_{nd},$$

where $B^{C-\dot{\alpha}}_{nd} = \frac{1}{n}\bar{\Sigma}_{nd}\Big[\mathrm{tr}[(\sum_{j=1}^{k_\alpha}\bar{C}^{(j)}_n\alpha^{(1)}_{nj})\bar{G}_{n\alpha}\,\mathbf{E}(\psi_{n\dot{\alpha}}\psi'_{n\mu})],\ldots,\mathrm{tr}[(\sum_{j=1}^{k_\alpha}\bar{C}^{(j)}_n\alpha^{(k_\beta)}_{nj})\bar{G}_{n\alpha}\,\mathbf{E}(\psi_{n\dot{\alpha}}\psi'_{n\mu})]\Big]'$, $B^{\dot{\alpha}-G_\alpha}_{nd} = -\bar{\Sigma}_{nd}\,\mathbf{E}(\alpha'_{n\beta}G'_{n\alpha}\bar{C}'_n\bar{P}_{nd}\bar{C}_n g_n)$, $\tilde{B}^{\dot{\alpha}-G_\alpha}_{nd} = -\frac{1}{n^2}\bar{\Sigma}_{nd}\sum_{i=1}^{n}\mathbf{E}(\alpha'_{n\beta}G'_{ni\alpha}\bar{C}'_n\bar{P}_{nd}\bar{C}_n g_{ni})$,

$$B^{\dot{\alpha}-G_\alpha-C}_{nd} = \frac{1}{n}\bar{\Sigma}_{nd}\alpha'_{n\beta}\,\mathbf{E}\Big[\Big(\sum_{j=1}^{k_\alpha}\bar{G}^{(j)}_{n\alpha}\bar{C}'_n\psi_{n\dot{\alpha}j} + \bar{G}'_{n\alpha}\psi'_{n\dot{C}}\Big)\psi_{n\mu}\Big],$$

and $\widetilde{\mathrm{Bias}}(\hat{\beta}_{\mathrm{TGEL}})$ *has the same form as that of* $\mathrm{Bias}(\hat{\beta}_{\mathrm{TGEL}})$ *in* (i) *except that* $\psi_{n\check{\alpha}}$ *in* $\mathrm{Bias}(\hat{\beta}_{\mathrm{TGEL}})$ *is replaced by* $\psi_{n\dot{\alpha}}$.

With the moment vector $\bar{C}_n g_n(\alpha_0, \beta)$, the bias terms B^{I}_{nd}, B^{Ω}_{nd} and B^{G}_{nd} have similar interpretations to corresponding ones for one-step estimators, respectively, the bias for a GMM estimator with the optimal linear combination $\bar{D}'_{n\beta}\bar{\Omega}^{-1}_{nd}\bar{C}_n g_n(\alpha_0, \beta)$, that from estimating the second moment matrix $\bar{\Omega}_n$ in $\bar{\Omega}_{nd}$ with the empirical variance $\frac{1}{n}\sum_{i=1}^{n} g_{ni}g'_{ni}$, and that from estimating $\bar{G}_{n\beta}$ in the gradient $\bar{C}_n\bar{G}_{n\beta}$. As $g_{ni}(\gamma)$'s are not i.i.d., $B^{G}_{nd} \neq \tilde{B}^{G}_{nd}$ and $B^{\Omega}_{nd} \neq \tilde{B}^{\Omega}_{nd}$ in general, and their differences depend on the strength of correlations across observations. If $g_{ni}(\gamma)$'s are i.i.d., then $B^{G}_{nd} - \tilde{B}^{G}_{nd}$ drops out from GEL's higher order bias. Furthermore, for EL, since $\rho_3(0) = -2$, the bias term $B^{\Omega}_{nd} + \frac{\rho_3(0)}{2}\tilde{B}^{\Omega}_{nd}$ also drops out in the i.i.d. case. The bias term $B^{C-\beta}_{nd} + B^{C-g}_{nd} + B^{C-\Omega}_{nd} + B^{C-G}_{nd} + B^{\check{\alpha}}_{nd}$ of $\hat{\beta}_{\text{TGEL}}$ arises since the ith transformed moment vector $d_{ni}(\check{\alpha}, \beta) = C_n(\check{\alpha}, \beta)g_{ni}(\check{\alpha}, \beta)$ involves $C_n(\alpha, \beta)$ and the initial estimate $\check{\alpha}$. The bias term $B^{C-\beta}_{nd}$arises from the derivatives $C_n^{(k_\alpha + j)}$ of $C_n(\check{\alpha}, \beta)$ with respective to β; B^{C-g}_{nd}, $B^{C-\Omega}_{nd}$ and B^{C-G}_{nd} arise from the estimation of $\bar{C}_n$ in, respectively, the moment vector $\bar{C}_n g_n(\alpha_0, \beta)$, the second moment matrix $\bar{C}_n\bar{\Omega}_n\bar{C}'_n$, and the gradient $\bar{C}_n\bar{G}_{n\beta}$; and $B^{\check{\alpha}}_{nd}$ arises from the initial estimate $\check{\alpha}$.

Compared with $\hat{\beta}_{\text{TGEL}}$, $\hat{\beta}_{\text{TGEL2}}$ has some additional bias terms $B^{C-\dot{\alpha}}_{nd}$, $B^{\dot{\alpha}-G_\alpha}_{nd}$, $\tilde{B}^{\dot{\alpha}-G_\alpha}_{nd}$ and $B^{\dot{\alpha}-G_\alpha-C}_{nd}$ due to the derivative of $\dot{\alpha}(\beta)$. The bias term $B^{C-\dot{\alpha}}_{nd}$ is the direct result of the derivative of $\dot{\alpha}(\beta)$, $B^{\dot{\alpha}-G_\alpha}_{nd}$ and $\tilde{B}^{\dot{\alpha}-G_\alpha}_{nd}$ are related to the correlation of g_n with estimated $\bar{G}_{n\alpha}$, and $B^{\dot{\alpha}-G_\alpha-C}_{nd}$ is related to the correlation of g_n with estimated C_n and also $\bar{G}_{n\alpha}$. With i.i.d. data, $\tilde{B}^{\dot{\alpha}-G_\alpha}_{nd} = B^{\dot{\alpha}-G_\alpha-C}_{nd}$, so $\tilde{B}^{\dot{\alpha}-G_\alpha}_{nd} - B^{\dot{\alpha}-G_\alpha-C}_{nd}$ drops out from $\text{Bias}(\hat{\beta}_{\text{TGEL2}})$.

In the special case that $g_{na}(\gamma) = g_{na}(\alpha)$ does not involve β and is $k_\alpha \times 1$, an initial consistent estimator $\check{\alpha}$ might be derived by solving $g_{na}(\alpha) = 0$. It follows that $C_n(\check{\alpha}, \beta)g_n(\check{\alpha}, \beta) = g_{nb}(\check{\alpha}, \beta)$ does not involve $C_n(\check{\alpha}, \beta)$. Furthermore, since $\mathbf{E}\,\frac{\partial g_{na}(\alpha_0)}{\partial \beta'} = 0$, $\bar{C}_n\bar{G}_{n\beta} = \mathbf{E}\,\frac{\partial g_{nb}(\gamma_0)}{\partial \beta'}$ does not involve $\bar{C}_{1n}$. Then $\hat{\beta}_{\text{TGEL}}$ will not have the bias terms B^{C-g}_{nd} and B^{C-G}_{nd}.[13]

Corollary 1. *If $g_{na}(\gamma) = g_{na}(\alpha)$ is $k_\alpha \times 1$, the unique solution $\check{\alpha}$ to $g_{na}(\alpha) = 0$ is a $\sqrt{n}$-consistent estimator of α_0 and $\check{\alpha}$ is used to derive $\hat{\beta}_{\text{TGEL}}$, then* $\text{Bias}(\hat{\beta}_{\text{TGEL}}) = [B^{I}_{nd} + (B^{\Omega}_{nd} + \frac{\rho_3(0)}{2}\tilde{B}^{\Omega}_{nd}) + (B^{G}_{nd} - \tilde{B}^{G}_{nd})] + [B^{C-\beta}_{nd} + B^{C-\Omega}_{nd} + B^{\check{\alpha}}_{nd}]$.

As in Newey and Smith (2004), Theorem 2 and Theorem D.1 in the supplementary mate-

[13] The bias term $B^{C-\Omega}_{nd}$ will still be present because $\bar{C}_{1n}$ appears in the second moment matrix $\bar{\Omega}_{nd}$, even though it does not appear in $\bar{C}_n g_n(\check{\alpha}, \beta) = g_{nb}(\check{\alpha}, \beta)$. For $\hat{\beta}_{\text{TGEL2}}$, we may consider the special case that $g_{na}(\gamma)$ is $k_\alpha \times 1$ and $\dot{\alpha}_n(\beta)$ is the unique solution to $g_{na}(\gamma) = 0$, where only the bias term B^{C-g}_{nd} disappears.

rial show that TGEL estimators have fewer bias sources than corresponding TGMM estimators. However, for general models, it is not clear whether TGEL estimators have smaller higher order biases than TGMM estimators or not, because the signs of higher order bias terms can be positive or negative and fewer bias terms do not necessarily mean smaller higher order bias. In addition, as pointed out in the introduction, a TGEL estimator might not have finite moments, so the analysis might not imply the exact finite sample bias of TGEL and TGMM. Criticisms and interpretations on using Nagar-type expansions for higher order bias analysis can be found in Rothenberg (1984) and references therein. To compare the higher order biases of GEL and GMM, Newey and Smith (2004) consider some special models including conditional moment restriction models and some minimum distance estimation models. For such models, GEL's bias does not increase with the number of moments while GMM's bias does. GEL automatically eliminates some bias terms due to the presence of unknown β in estimated $\bar{C}_n$. Except for these bias terms eliminated by GEL, the bias terms from the moment vector $\bar{C}_n g_n(\alpha_0, \beta)$ and TGMM's extra bias terms due to its two-step nature in forming an optimal weighting matrix, TGEL and TGMM estimators have the same higher order bias due to the estimation of $\bar{C}_n$ and the estimate $\check{\alpha}$ or $\dot{\alpha}_n(\beta)$. Thus TGEL estimates generally have bias advantages over TGMM estimates for those conditional moment restriction models and minimum distance estimation models. In our framework, $\check{\alpha}$ and $\dot{\alpha}_n(\beta)$ are arbitrary except for some regularity conditions, thus it is not easy to see the relation between the number of moments and the bias terms which are not from $\bar{C}_n g_n(\alpha_0, \beta)$.

3 Tests for parameter restrictions

In this section, we study tests for parameter restrictions with the TGEL estimator. We consider k_r general restrictions $r(\beta_0) = 0$ on the parameters of interests β for a $k_r \times 1$ vector of functions $r(\cdot)$ with $k_r < k_\beta$. The alternative hypothesis is $r(\beta_0) \neq 0$. For any $\sqrt{n}$-consistent estimator $\check{\alpha}$ of α, $d_{ni}(\check{\alpha}, \beta)$ plays the role of $g_{ni}(\gamma)$.[14]

[14] $\dot{\alpha}_n(\hat{\beta})$ at a $\sqrt{n}$-consistent estimator $\hat{\beta}$ of β_0 is a $\sqrt{n}$-consistent estimator of α_0 and the first order asymptotic analysis in the following text is the same, so we just use $\check{\alpha}$.

Let $\boldsymbol{\rho}_{nd}(\alpha,\beta,\mu)=\sum_{i=1}^{n}\rho\Big(\mu' d_{ni}(\alpha,\beta)\Big)$, $\hat{\beta}_{\text{RTGEL}}$ be the restricted TGEL estimator that solves

$$\min_{\beta\in\mathcal{B}}\ \sup_{\mu\in\Lambda_{nd}(\check{\alpha},\beta)} \boldsymbol{\rho}_{nd}(\check{\alpha},\beta,\mu), \quad \text{s.t.} \quad r(\beta)=0,$$

and $\hat{\mu}_{\text{RTGEL}}=\arg\max_{\mu\in\Lambda_{nd}(\check{\alpha},\hat{\beta}_{\text{RTGEL}})}\boldsymbol{\rho}_{nd}(\check{\alpha},\hat{\beta}_{\text{RTGEL}},\mu)$. The TGEL ratio test has the test statistic

$$\mathcal{R}_{\text{TGEL}}=2[\boldsymbol{\rho}_{nd}(\check{\alpha},\hat{\beta}_{\text{RTGEL}},\hat{\mu}_{\text{RTGEL}})-\boldsymbol{\rho}_{nd}(\check{\alpha},\hat{\beta}_{\text{TGEL}},\hat{\mu}_{\text{TGEL}})], \tag{3.1}$$

which follows the asymptotic distribution $\chi^2(k_r)$ under the null hypothesis $r(\beta_0)=0$. While $\mathcal{R}_{\text{TGEL}}$ requires both restricted and unrestricted estimates, it avoids the estimation of any variance and it has a form similar to the likelihood ratio test. The Wald test statistic with the TGEL estimate is

$$\mathcal{W}_{\text{TGEL}}=n\cdot r'(\hat{\beta}_{\text{TGEL}})[R(\hat{\beta}_{\text{TGEL}})\Sigma_{nd}(\check{\alpha},\hat{\beta}_{\text{TGEL}})R'(\hat{\beta}_{\text{TGEL}})]^{-1}r(\hat{\beta}_{\text{TGEL}}), \tag{3.2}$$

where $R(\beta)=\frac{\partial r(\beta)}{\partial\beta'}$ and $\Sigma_{nd}(\gamma)=[D'_{n\beta}(\gamma)\Omega_{nd}^{-1}(\gamma)D_{n\beta}(\gamma)]^{-1}$. Alternatively, we may consider a restricted GEL estimation and construct a test directly based on the GEL score $\frac{\partial\boldsymbol{\rho}_{nd}(\gamma,\mu)}{\partial\beta}$ evaluated at the restricted GEL estimate. Under the null that $r(\beta_0)=0$, the test statistic satisfies

$$\mathcal{S}_{\text{TGEL}}=\frac{1}{n}\frac{\partial\boldsymbol{\rho}_{nd}(\check{\alpha},\hat{\beta}_{\text{RTGEL}},\hat{\mu}_{\text{RTGEL}})}{\partial\beta'}\Sigma_{nd}(\check{\alpha},\hat{\beta}_{\text{RTGEL}})\frac{\partial\boldsymbol{\rho}_{nd}(\check{\alpha},\hat{\beta}_{\text{RTGEL}},\hat{\mu}_{\text{RTGEL}})}{\partial\beta}\xrightarrow{d}\chi^2(k_r). \tag{3.3}$$

This test only requires the restricted TGEL estimate.

A $C(\alpha)$-type gradient test and a corresponding GEL test can be applied with any $\sqrt{n}$-consistent restricted estimator $\check{\beta}_r$ such that $r(\check{\beta}_r)=0$. Let

$$\Psi_n(\alpha,\beta)=R(\check{\gamma}_r)\Sigma_{nd}(\check{\gamma}_r)D'_{n\beta}(\check{\gamma}_r)\Omega_{nd}^{-1}(\check{\gamma}_r)d_n(\alpha,\beta), \tag{3.4}$$

where $\check{\gamma}_r=(\check{\alpha},\check{\beta}_r')'$. Then $\sqrt{n}\Psi_n(\check{\alpha},\check{\beta}_r)$ is a $C(\alpha)$-type statistic such that it has the same asymptotic distribution as that of $\sqrt{n}\Psi_n(\check{\alpha},\beta_0)$ by the mean value theorem, and the same as that of $\sqrt{n}\Psi_n(\alpha_0,\beta_0)$ since $\sqrt{n}d_n(\check{\alpha},\beta_0)=\sqrt{n}d_n(\alpha_0,\beta_0)+o_p(1)$. Let $\Psi_{ni}(\alpha,\beta)$ be the vector derived by replacing $d_n(\alpha,\beta)$ in $\Psi_n(\alpha,\beta)$ with $d_{ni}(\alpha,\beta)$ so that $\Psi_n(\alpha,\beta)=\frac{1}{n}\sum_{i=1}^{n}\Psi_{ni}(\alpha,\beta)$. Then we have the following gradient test in the GEL framework:

$$\mathcal{G}_{\text{TGEL}}=2\Big[\sup_{\lambda\in\Lambda_{n\Psi}(\check{\alpha},\check{\beta}_r)}\sum_{i=1}^{n}\rho\Big(\lambda'\Psi_{ni}(\check{\alpha},\check{\beta}_r)\Big)-n\rho(0)\Big], \tag{3.5}$$

where $\Lambda_{n\Psi}(\alpha,\beta)=\{\lambda:\lambda'\Psi_{ni}(\alpha,\beta)\in\mathcal{V}, i=1,\ldots,n\}$. Note that $\frac{1}{n}\sum_{i=1}^{n}\Psi_{ni}(\check{\gamma}_r)\Psi'_{ni}(\check{\gamma}_r)$ is a consistent estimator of the limiting variance of $\sqrt{n}\Psi_n(\gamma_0)$ and its inverse is used internally by the

GEL as the optimal weighting matrix. An advantage of $\mathcal{G}_{\text{TGEL}}$ is its robustness to unknown heteroskedasticity, because $\Omega_n(\gamma_0) = \frac{1}{n}\sum_{i=1}^{n} g_{ni}(\gamma_0)g_{ni}'(\gamma_0)$ may capture unknown heteroskedasticity in $g_{ni}(\gamma_0)$ (Lee and Yu, 2012). This test statistic only requires the estimation of the auxiliary vector λ in the GEL framework, and no variance matrix needs to be estimated.[15]

We maintain Assumption 4 for asymptotic analysis on the above tests.

Assumption 4. *(i) The true β value in the data generating process is $\beta_n = \beta_0 + n^{-1/2}c$ for some constant vector c; (ii) $r(\beta)$ is continuously differentiable and $R = \frac{\partial r(\beta_0)}{\partial \beta'}$ has full row rank; (iii) $r(\check{\beta}_r) = 0$ and $\sqrt{n}(\check{\beta}_r - \beta_0) = O_p(1)$.*

With a continuously differentiable $r(\cdot)$, the Pitman drift in Assumption 4(i) implies a local violation of $r(\beta) = 0$.

Theorem 3. *Suppose that Assumptions 1–2 and 4 hold.*

(i) *$\mathcal{R}_{\text{TGEL}}$, $\mathcal{W}_{\text{TGEL}}$, $\mathcal{S}_{\text{TGEL}}$ and $\mathcal{G}_{\text{TGEL}}$ are all asymptotically equivalent with the asymptotic distribution $\chi^2(k_r, \lim_{n\to\infty} c'R'(R\bar{\Sigma}_{nd}R')^{-1}Rc)$, where $\chi^2(a_1, a_2)$ denotes a noncentral chi-squared distribution with a_1 degrees of freedom and the noncentrality parameter a_2.*

(ii) *$\lim_{n\to\infty} c'R'(R\bar{\Sigma}_{nd}R')^{-1}Rc \leq \lim_{n\to\infty} c'R'(R_\gamma\bar{\Sigma}_n R_\gamma')^{-1}Rc$, where $R_\gamma = \frac{\partial r(\beta_0)}{\partial \gamma'} = [0, R]$, $\bar{\Sigma}_n = (\bar{G}_n'\bar{\Omega}_n^{-1}\bar{G}_n)^{-1}$, $\lim_{n\to\infty} c'R'(R\bar{\Sigma}_{nd}R')^{-1}Rc$ is the noncentrality parameter in (i), and*

$$\lim_{n\to\infty} c'R'(R_\gamma\bar{\Sigma}_n R_\gamma')^{-1}Rc$$

is the noncentrality parameter for tests in the ordinary GEL framework in the supplementary material.

(iii) *If $m_a = k_\alpha$, then $\lim_{n\to\infty} c'R'(R\bar{\Sigma}_{nd}R')^{-1}Rc = \lim_{n\to\infty} c'R'(R_\gamma\bar{\Sigma}_n R_\gamma')^{-1}Rc$.*

The above theorem shows that tests in the TGEL framework are asymptotically equivalent under either the null or local alternative hypotheses. These tests are locally less powerful in general than those in the ordinary GEL framework, but they are locally as powerful as the latter ones when $m_a = k_\alpha$.

[15] In Jin and Lee (2019), this form of test is used to implement Moran's I test for spatial dependence, which can also be robust under unknown heteroskedasticity.

4 Monte Carlo

In this section, we conduct some Monte Carlo studies on finite sample performance of the two-step estimators and tests considered in this paper. We consider both the probit model (2.1) with an endogenous regressor and the SAR model (2.2).

4.1 Probit model with an endogenous regressor

We first consider estimation of model (2.1) with simple moments as those in Wilde (2008), where the initial consistent estimator of nuisance parameters for TGEL and TGMM is the solution to a subset of simple empirical moment conditions.[16]

The parameters for model (2.1) are not identifiable without a proper normalization. We may show that $\mathbf{E}(y_1|x) = \Phi((x_1'\kappa + x'\alpha\tau)/\sigma_s)$, where $\sigma_s^2 = \sigma_u^2 + \tau^2\sigma_\epsilon^2 + 2\rho\tau\sigma_u\sigma_\epsilon$, and $\Phi(\cdot)$ denotes the standard normal cumulative distribution function. Following Wilde (2008), with $\beta_1 = \kappa/\sigma_s$ and $\beta_2 = \tau/\sigma_s$, we have the simple moment vector $g(\gamma) = [g_b'(\gamma), g_a'(\gamma)]'$, where $g_b(\gamma) = x[y_1 - \Phi(x_1'\beta_1 + x'\alpha\beta_2)]$, $g_a(\gamma) = x(y_2 - x'\alpha)$, and $\gamma = (\alpha', \beta')'$ with $\beta = (\beta_1', \beta_2)'$. Given $g(\gamma)$, a convenient two-step estimation is to first derive the OLS estimate $\check{\alpha}$ by regressing y_2 on x, which is the solution to the empirical moment condition $\sum_{i=1}^n g_{ai}(\gamma) = 0$, where $g_{ai}(\gamma)$ denotes $g_a(\gamma)$ at the ith observation, and then use the transformed moment $C_n(\check{\alpha}, \beta)g(\check{\alpha}, \beta)$ to estimate β, where $C_n(\alpha, \beta) = [I_{m_b}, -(\sum_{i=1}^n \frac{\partial g_{bi}(\gamma)}{\partial \alpha'})(\sum_{i=1}^n \frac{\partial g_{ai}(\gamma)}{\partial \alpha'})^{-1}]$.

In our Monte Carlo experiments, x_1 contains 2 regressors of independent standard normal random variables and we set $\beta_{10} = \alpha_{10} = 0$. The two-step approaches have computational advantage, in particular, if x_2 contains many variables, so we let x_2 contain 5 or 20 regressors of independent standard normal random variables. In a linear regression model of y_1 with only an endogenous regressor, i.e., if y_1 in (2.1) were generated by $y_1 = y_2\tau + u$ instead, the first stage F, which is a measure of the IV strength (Staiger and Stock, 1997), is approximately $\frac{nR^2}{k_2(1-R^2)} + 1$, where $R^2 = \frac{\alpha_{20}'\alpha_{20}}{\alpha_{20}'\alpha_{20} + \sigma_\epsilon^2}$ and k_2 is the number of variables in x_2. $F > 10$ is usually regarded as the

[16] As an alternative asymptotically efficient estimation, in the supplementary material, we consider estimation with the score vector of the likelihood function, where the initial estimator of nuisance parameters is computationally simple but is not a solution derived from a subvector of the complex score vector.

case with strong IV (Staiger and Stock, 1997). We set R^2 to be 0.7 or 0.01 and set n to be 100 or 400, so that $R^2 = 0.7$ and $R^2 = 0.01$ correspond to, respectively, relatively strong and very weak IV cases. Elements of α_{20} are equal. The true values of σ_u and σ_ϵ are equal and are selected so that the true value of σ_s is equal to 1. Then $\beta_2 = \tau$ and $\sigma_\epsilon^2 = \frac{1}{1+\tau^2+2\rho\tau}$. The true value of τ is set to be 0, and the true ρ_0 is either 0, 0.2, 0.4, 0.6 or 0.8. Each exogenous variable in x_2 has the same coefficient. The number of Monte Carlo repetitions is 2000, and the nominal size for various tests is 0.05.

For GEL and TGEL estimators, we use a double optimization method. Both inner and outer optimizations use a quasi-Newton strategy with limited memory BFGS updating. For the inner optimization, the first order derivative of the objective function is provided and the starting value is a zero vector. For the outer optimization, the provided derivative is derived by the implicit function theorem. The starting value is the GMM estimate $\tilde{\beta} = \arg\min_\beta g_{nb}'(\check{\alpha}, \beta) g_{nb}(\check{\alpha}, \beta)$, where $\check{\alpha}$ is the OLS estimate by regressing y_2 on x_2 and the starting value for $\tilde{\beta}$ is a zero vector. More computational details are in the supplementary material.

For the performance of various estimators, we compute the following measures: median bias (MB), median absolute deviation (MAD), interdecile range (IDR), bias, standard deviation (SD), root mean squared error (RMSE), and tail probability (TP), which is the proportion of estimates with absolute values larger than $25 \times 90\% = 22.5$. We follow Guggenberger (2008) to use the number 25 for TP. The first three are robust measures of central tendency and dispersion. We consider the following two-step estimates: two-step ET (TET), two-step EL (TEL) and two-step GMM (TGMM), and compare them with the joint estimates ET, EL and (feasible optimal) GMM.[17] ET, EL and GMM use jointly all the moments in $g(\gamma)$.

4.1.1 Estimation results

The parameter β_2 for the endogenous regressor is often a parameter of interest, so we focus on the performance of various estimates of β_2. Figure 1 presents the estimation results when

[17] The CU estimator is often observed to possess multiple modes and thus generally considered to be less desirable than the EL and ET estimators (Hansen et al., 1996; Imbens et al., 1998). Our Monte Carlo results also show that CU has worse performance than ET and EL. For simplicity, we do not include results for CU in the main text but report them in the supplementary material.

$R^2 = 0.7$. Only the robust measures MB, MAD and IDR are reported, since all TPs are zero in this case and comparisons of robust measures are observed to be the same as those of the usual measures bias, SD and RMSE. All MBs are relatively small when x_2 contains 5 variables. The MB of GMM is large when x_2 contains 20 variables and $\rho_0 \neq 0$, but those of ET and EL are still relatively small. In terms of MAD and IDR, among ET, EL and GMM, EL performs the best and GMM performs the worst; among TET, TEL and TGMM, EL performs the best, and ET outperforms TGMM in most cases. The two-step estimates TET, TEL and TGMM have similar performance as corresponding one-step estimates in general. We also report TET and TEL estimates where there is no unknown β in the transformation matrix $\hat{C}_n$, which we denote by TET$_c$ and TEL$_c$. We observe that TET$_c$ and TEL$_c$ tend to have larger MB than corresponding TET and TEL, but they generally have smaller MAD and IDR.

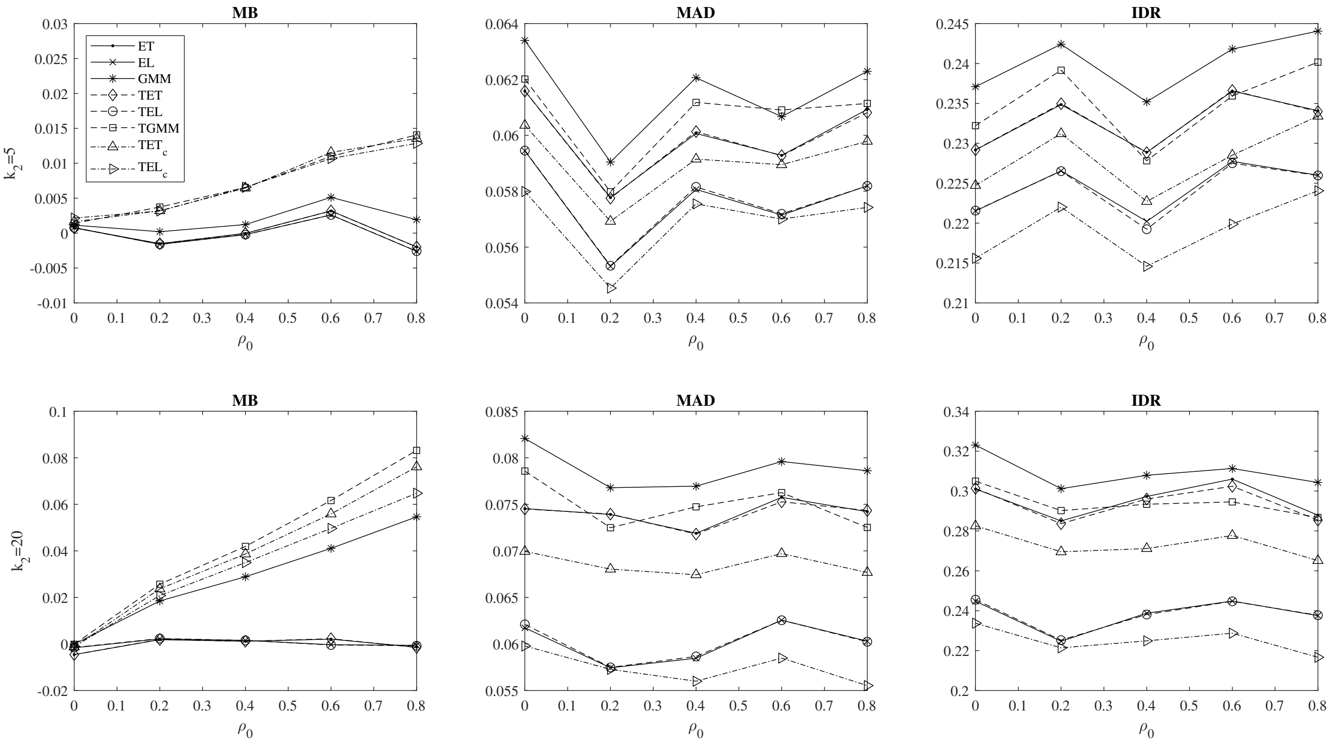

Figure 1: Performance of various estimates of β_2 in the probit model with $R^2 = 0.7$. All TPs are observed to be zero. TET$_c$ and TEL$_c$ are, respectively, TET and TEL estimators with $\hat{C}_n(\check{\alpha}, \bar{\beta})$. k_2 is the number of variables in x_2, the true value of τ is 0, and the sample size n is 100.

Figure 2 presents the estimation results when $R^2 = 0.01$. We observe very different results compared with those in Figure 1 when $R^2 = 0.7$. With $R^2 = 0.01$, ET, EL and GMM have nonzero TPs in all cases, while two-step estimates have zero TPs except TEL$_c$. EL has a smaller TP than that of ET, but larger than that of GMM. As a result, among ET, EL and GMM, ET has the largest

SD and RMSE, and GMM has the smallest SD and RMSE. Figure 2 omits the usual measures bias, SD and RMSE. In terms of robust measures MB, MAD and IDR, GMM has larger MB than those of ET and EL, but smaller MAD and IDR in some cases. ET has larger MB, MAD and IDR than those of EL. TEL and TEL have smaller MB than that of TGMM in some cases, but they generally have larger MAD and IDR than those of TGMM. While TET$_c$ and TEL$_c$ tend to have larger MB than that of corresponding TET and TEL, they have significantly smaller MAD and IDR. TEL$_c$ generally has the smallest MAD and IDR among two-step estimates.

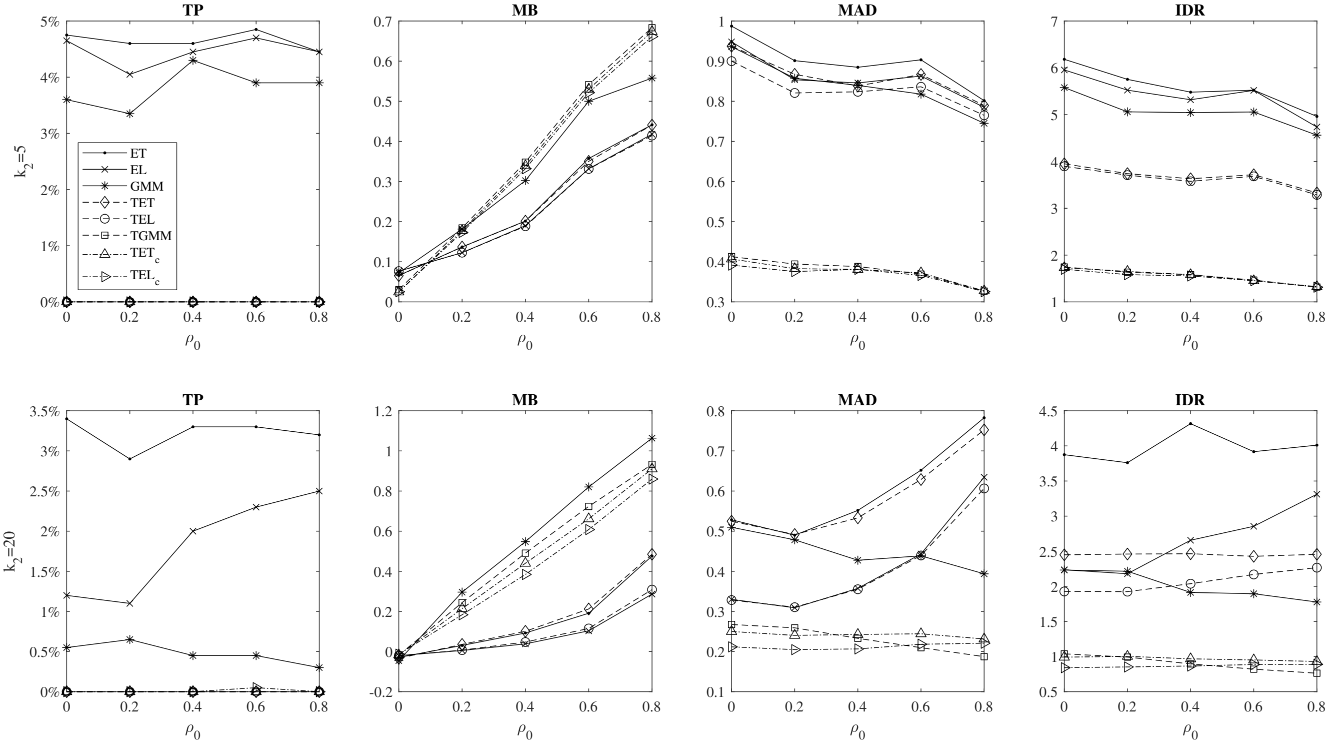

Figure 2: Performance of various estimates of β_2 in the probit model with $R^2 = 0.01$. TET$_c$ and TEL$_c$ are, respectively, TET and TEL estimators with $\hat{C}_n(\check{\alpha}, \tilde{\beta})$. k_2 is the number of variables in x_2, the true value of τ is 0, and the sample size n is 100.

To investigate the potential local minimum problem of GEL, we also conduct some Monte Carlo experiments where there is no x_1 in model (2.1), so that for two-step estimates, the unknown parameter $\beta = \beta_2$ is one-dimensional and we can do a grid search.[18] Following Guggenberger (2008), two-step estimates of β_2 are searched over the interval $[-25, 25]$ with a grid size 0.01. Figure 3 reports the results for the case with $R^2 = 0.7$. We observe similar patterns as those in the corresponding Figure 1, where grid search is not used. The results with grid search

[18] We thank an anonymous referee for this suggestion. For joint GMM and GEL estimates, we do not use grid search since the large number of unknown parameters makes grid search computationally demanding.

for the case with $R^2 = 0.01$ are similar to those in Figure 2. They are omitted to save space.

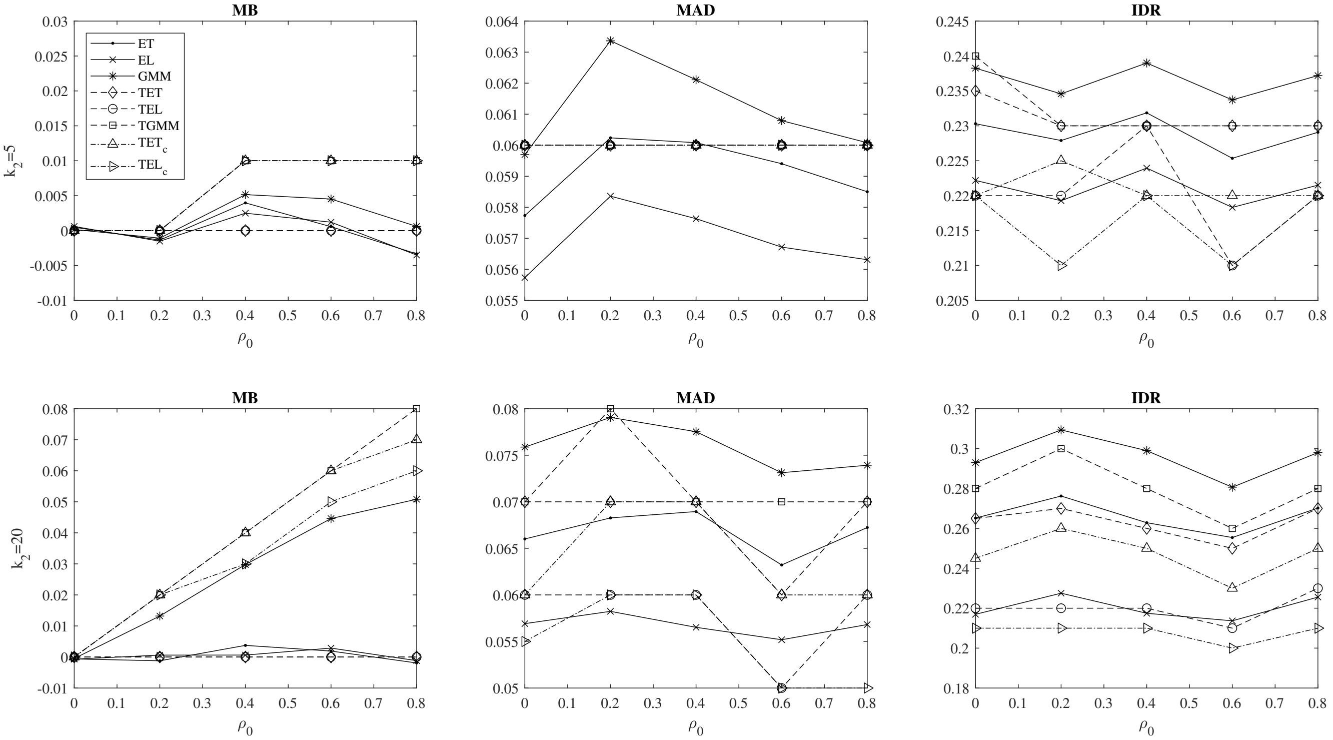

Figure 3: Performance of various estimates of β_2 with grid search in the probit model with $R^2 = 0.7$. All TPs are observed to be zero. TET$_c$ and TEL$_c$ are, respectively, TET and TEL estimators with $\hat{C}_n(\breve{\alpha},\tilde{\beta})$. k_2 is the number of variables in x_2, the true value of τ is 0, and the sample size n is 100.

From these results, we can see that ET and EL outperform GMM in cases with strong identification, but they may have a heavy tail problem in cases with weak identification. This is consistent with the Monte Carlo results in Guggenberger (2008). Note that in our theoretical analysis, we have assumed strong identification. In the weak identification case, the GEL estimators cease to be consistent and can have a nonstandard asymptotic distribution which is different from that of the ("optimal") GMM estimator (Stock and Wright, 2000; Guggenberger and Smith, 2005). Thus, we may observe the results in Figure 2.

Table 1 reports the computational time of various estimates where grid search is not used. GEL is computationally more intensive than GMM as expected from the saddle-point characterization of GEL. EL takes slightly more time to compute than that of ET. GEL takes about 5 to 20 times longer to compute than TGEL, and GMM takes about 3 to 10 times longer to compute than TGMM. TET$_c$ and TEL$_c$ take less time to compute than the corresponding TET and TEL in most cases. The computational time generally increases as the sample size n, k_2 and ρ_0 increase and as R^2 decreases.

Table 1: Computational time of various estimates for the probit model

		ET	EL	GMM	TET	TEL	TGMM	TET$_c$	TEL$_c$
$n = 100$	$k_2 = 5$, $R^2 = 0.7$, $\rho_0 = 0$	20.4	21.0	3.8	7.7	7.8	1.3	7.9	7.3
	$k_2 = 5$, $R^2 = 0.7$, $\rho_0 = 0.8$	37.2	40.0	4.9	7.8	8.4	1.6	7.0	7.7
	$k_2 = 5$, $R^2 = 0.01$, $\rho_0 = 0$	94.1	100.2	12.9	14.9	15.7	2.1	11.2	12.2
	$k_2 = 5$, $R^2 = 0.01$, $\rho_0 = 0.8$	109.2	124.1	13.0	15.4	16.7	2.1	11.4	12.4
	$k_2 = 20$, $R^2 = 0.7$, $\rho_0 = 0$	134.9	153.0	11.0	15.3	17.5	2.1	13.4	16.0
	$k_2 = 20$, $R^2 = 0.7$, $\rho_0 = 0.8$	234.9	277.7	11.8	15.5	18.7	2.2	13.3	16.1
	$k_2 = 20$, $R^2 = 0.01$, $\rho_0 = 0$	293.8	255.9	15.6	22.7	25.3	2.4	16.3	19.5
	$k_2 = 20$, $R^2 = 0.01$, $\rho_0 = 0.8$	417.1	446.3	17.0	26.7	31.7	2.4	18.8	23.7
$n = 400$	$k_2 = 5$, $R^2 = 0.7$, $\rho_0 = 0$	102.7	102.1	4.9	35.4	35.0	2.7	38.9	40.2
	$k_2 = 5$, $R^2 = 0.7$, $\rho_0 = 0.8$	245.0	245.4	6.9	36.0	36.2	2.8	39.9	40.4
	$k_2 = 5$, $R^2 = 0.01$, $\rho_0 = 0$	393.2	395.8	16.8	87.1	86.8	4.1	80.1	80.6
	$k_2 = 5$, $R^2 = 0.01$, $\rho_0 = 0.8$	619.4	624.4	18.8	88.0	90.9	3.9	80.9	82.9
	$k_2 = 20$, $R^2 = 0.7$, $\rho_0 = 0$	299.1	312.3	13.8	69.0	74.8	3.8	65.2	71.5
	$k_2 = 20$, $R^2 = 0.7$, $\rho_0 = 0.8$	625.0	651.9	16.9	71.6	80.0	3.7	64.9	71.4
	$k_2 = 20$, $R^2 = 0.01$, $\rho_0 = 0$	1301.2	1326.7	39.2	152.1	165.3	4.8	110.3	120.0
	$k_2 = 20$, $R^2 = 0.01$, $\rho_0 = 0.8$	1852.8	1814.0	43.1	158.7	177.2	4.8	120.1	136.2

(i) The reported numbers are the total time in seconds for computing each estimate 2000 times. The results are from Matlab on a desktop computer with Intel Core i7-8700 CPU and 16 gigabyte memory.

(ii) TET$_c$ and TEL$_c$ are, respectively, TET and TEL estimators with $\hat{C}_n(\check{\alpha}, \tilde{\beta})$.

(iii) k_2 is the number of variables in x_2. The true values of τ is 0.

4.1.2 Tests

To investigate the performance of various tests of $\beta_{20} = 0$, we set $R^2 = 0.7$ which corresponds to the case with relatively strong identification. Table 2 reports empirical sizes. With $k_2 = 5$, the size distortions of all tests are relatively small. With $k_2 = 20$, $\mathcal{W}_{\text{ET}}$, $\mathcal{W}_{\text{EL}}$, $\mathcal{W}_{\text{TET}}$, $\mathcal{W}_{\text{TEL}}$, $\mathcal{G}_{\text{TET}}$ and $\mathcal{G}_{\text{TEL}}$ have large size distortions for the small sample size $n = 100$, but their empirical sizes become much closer to the nominal 5% with $n = 400$. Tests in the two-step frameworks have similar empirical sizes as those of corresponding ones in the ordinary frameworks except for some cases with $k_2 = 20$, $n = 100$ and a large $\rho_0 = 0.8$.

Table 3 reports empirical powers of the tests when $n = 100$. The powers of all tests increase as τ_0 increases. With $\tau_0 = 0.4$, the powers are close to 1. For given ρ_0, τ_0 and k_2, different tests generally have similar powers.

Table 2: Empirical sizes of various tests for $\beta_{20} = 0$ in the probit model

	$k_2 = 5, n = 100$		$k_2 = 5, n = 400$		$k_2 = 20, n = 100$		$k_2 = 20, n = 400$	
	$\rho_0 = 0$	$\rho_0 = 0.8$	$\rho_0 = 0$	$\rho_0 = 0.8$	$\rho_0 = 0$	$\rho_0 = 0.8$	$\rho_0 = 0$	$\rho_0 = 0.8$
$\mathcal{R}_{\text{ET}}$	0.052	0.065	0.055	0.066	0.081	0.081	0.063	0.066
$\mathcal{R}_{\text{EL}}$	0.052	0.065	0.054	0.065	0.079	0.086	0.059	0.064
$\mathcal{W}_{\text{ET}}$	0.061	0.067	0.058	0.065	0.166	0.160	0.074	0.071
$\mathcal{W}_{\text{EL}}$	0.052	0.059	0.054	0.062	0.094	0.096	0.060	0.062
$\mathcal{S}_{\text{ET}}$	0.036	0.043	0.051	0.060	0.021	0.021	0.045	0.058
$\mathcal{S}_{\text{EL}}$	0.045	0.060	0.053	0.063	0.062	0.070	0.055	0.064
$\mathcal{G}_{\text{ET}}$	0.046	0.060	0.053	0.064	0.059	0.079	0.057	0.064
$\mathcal{G}_{\text{EL}}$	0.046	0.058	0.052	0.064	0.057	0.079	0.057	0.064
$\mathcal{R}_{\text{TET}}$	0.052	0.065	0.055	0.066	0.082	0.081	0.063	0.066
$\mathcal{R}_{\text{TEL}}$	0.052	0.065	0.054	0.065	0.080	0.084	0.059	0.064
$\mathcal{W}_{\text{TET}}$	0.060	0.068	0.058	0.066	0.161	0.163	0.073	0.071
$\mathcal{W}_{\text{TEL}}$	0.051	0.060	0.055	0.064	0.092	0.096	0.063	0.064
$\mathcal{S}_{\text{TET}}$	0.036	0.044	0.051	0.062	0.025	0.025	0.046	0.056
$\mathcal{S}_{\text{TEL}}$	0.046	0.058	0.052	0.063	0.065	0.061	0.054	0.063
$\mathcal{G}_{\text{TET}}$	0.049	0.057	0.055	0.063	0.063	0.131	0.057	0.068
$\mathcal{G}_{\text{TEL}}$	0.048	0.055	0.055	0.063	0.061	0.129	0.056	0.067

(i) k_2 is the number of variables in x_2. The nominal size is 5%.

(ii) $\mathcal{R}_{\text{ET}}$: ET ratio test; $\mathcal{R}_{\text{EL}}$: EL ratio test; $\mathcal{W}_{\text{ET}}$: ET Wald test; $\mathcal{W}_{\text{EL}}$: EL Wald test; $\mathcal{S}_{\text{ET}}$: score-type test in the ET framework; $\mathcal{S}_{\text{EL}}$: score-type test in the EL framework; $\mathcal{G}_{\text{ET}}$: ET gradient test; $\mathcal{G}_{\text{EL}}$: EL gradient test.

(iii) $\mathcal{R}_{\text{TET}}$: TET ratio test; $\mathcal{R}_{\text{TEL}}$: TEL ratio test; $\mathcal{W}_{\text{TET}}$: TET Wald test; $\mathcal{W}_{\text{TEL}}$: TEL Wald test; $\mathcal{S}_{\text{TET}}$: score-type test in the TET framework; $\mathcal{S}_{\text{TEL}}$: score-type test in the TEL framework; $\mathcal{G}_{\text{TET}}$: TET gradient test; $\mathcal{G}_{\text{TEL}}$: TEL gradient test.

Table 3: Empirical powers of various tests for $\beta_{20} = 0$ in the probit model

		$\rho_0 = 0$				$\rho_0 = 0.8$			
		$\tau_0 = 0.1$	$\tau_0 = 0.2$	$\tau_0 = 0.3$	$\tau_0 = 0.4$	$\tau_0 = 0.1$	$\tau_0 = 0.2$	$\tau_0 = 0.3$	$\tau_0 = 0.4$
$k_2 = 5$	$\mathcal{R}_{\text{ET}}$	0.245	0.668	0.914	0.992	0.208	0.546	0.807	0.941
	$\mathcal{R}_{\text{EL}}$	0.244	0.658	0.912	0.993	0.203	0.538	0.806	0.937
	$\mathcal{W}_{\text{ET}}$	0.265	0.687	0.929	0.996	0.272	0.619	0.864	0.960
	$\mathcal{W}_{\text{EL}}$	0.240	0.666	0.919	0.995	0.249	0.596	0.850	0.955
	$\mathcal{S}_{\text{ET}}$	0.193	0.580	0.874	0.984	0.160	0.469	0.754	0.908
	$\mathcal{S}_{\text{EL}}$	0.228	0.644	0.900	0.992	0.195	0.523	0.792	0.934
	$\mathcal{G}_{\text{ET}}$	0.233	0.654	0.910	0.993	0.213	0.551	0.814	0.943
	$\mathcal{G}_{\text{EL}}$	0.230	0.650	0.908	0.993	0.208	0.545	0.813	0.942
	$\mathcal{R}_{\text{TET}}$	0.245	0.668	0.914	0.992	0.208	0.546	0.807	0.941
	$\mathcal{R}_{\text{TEL}}$	0.244	0.658	0.912	0.993	0.203	0.538	0.806	0.937
	$\mathcal{W}_{\text{TET}}$	0.262	0.688	0.929	0.995	0.275	0.620	0.867	0.961
	$\mathcal{W}_{\text{TEL}}$	0.242	0.664	0.922	0.994	0.256	0.601	0.857	0.956
	$\mathcal{S}_{\text{TET}}$	0.184	0.561	0.848	0.969	0.110	0.353	0.642	0.826
	$\mathcal{S}_{\text{TEL}}$	0.223	0.625	0.886	0.985	0.144	0.422	0.718	0.879
	$\mathcal{G}_{\text{TET}}$	0.245	0.672	0.918	0.992	0.258	0.607	0.856	0.957
	$\mathcal{G}_{\text{TEL}}$	0.243	0.667	0.915	0.992	0.253	0.601	0.853	0.957
$k_2 = 20$	$\mathcal{R}_{\text{ET}}$	0.264	0.618	0.877	0.980	0.214	0.528	0.784	0.915
	$\mathcal{R}_{\text{EL}}$	0.275	0.638	0.892	0.984	0.224	0.546	0.804	0.926
	$\mathcal{W}_{\text{ET}}$	0.384	0.749	0.935	0.992	0.354	0.691	0.882	0.970
	$\mathcal{W}_{\text{EL}}$	0.308	0.668	0.905	0.989	0.282	0.625	0.840	0.953
	$\mathcal{S}_{\text{ET}}$	0.105	0.336	0.614	0.777	0.082	0.283	0.482	0.634
	$\mathcal{S}_{\text{EL}}$	0.243	0.589	0.874	0.982	0.204	0.512	0.786	0.915
	$\mathcal{G}_{\text{ET}}$	0.239	0.629	0.910	0.989	0.332	0.692	0.890	0.975
	$\mathcal{G}_{\text{EL}}$	0.235	0.626	0.909	0.989	0.331	0.690	0.888	0.975
	$\mathcal{R}_{\text{TET}}$	0.265	0.621	0.876	0.982	0.210	0.526	0.783	0.916
	$\mathcal{R}_{\text{TEL}}$	0.275	0.639	0.894	0.987	0.221	0.544	0.800	0.925
	$\mathcal{W}_{\text{TET}}$	0.378	0.744	0.938	0.988	0.361	0.698	0.887	0.969
	$\mathcal{W}_{\text{TEL}}$	0.299	0.667	0.908	0.990	0.291	0.636	0.849	0.959
	$\mathcal{S}_{\text{TET}}$	0.108	0.320	0.568	0.702	0.043	0.161	0.312	0.427
	$\mathcal{S}_{\text{TEL}}$	0.246	0.586	0.856	0.970	0.157	0.426	0.682	0.854
	$\mathcal{G}_{\text{TET}}$	0.255	0.677	0.931	0.982	0.462	0.814	0.946	0.981
	$\mathcal{G}_{\text{TEL}}$	0.254	0.677	0.931	0.982	0.460	0.810	0.945	0.981

(i) k_2 is the number of variables in x_2, the nominal size is 5%, and the sample size is 100.

(ii) $\mathcal{R}_{\text{ET}}$: ET ratio test; $\mathcal{R}_{\text{EL}}$: EL ratio test; $\mathcal{W}_{\text{ET}}$: ET Wald test; $\mathcal{W}_{\text{EL}}$: EL Wald test; $\mathcal{S}_{\text{ET}}$: score-type test in the ET framework; $\mathcal{S}_{\text{EL}}$: score-type test in the EL framework; $\mathcal{G}_{\text{ET}}$: ET gradient test; $\mathcal{G}_{\text{EL}}$: EL gradient test.

(iii) $\mathcal{R}_{\text{TET}}$: TET ratio test; $\mathcal{R}_{\text{TEL}}$: TEL ratio test; $\mathcal{W}_{\text{TET}}$: TET Wald test; $\mathcal{W}_{\text{TEL}}$: TEL Wald test; $\mathcal{S}_{\text{TET}}$: score-type test in the TET framework; $\mathcal{S}_{\text{TEL}}$: score-type test in the TEL framework; $\mathcal{G}_{\text{TET}}$: TET gradient test; $\mathcal{G}_{\text{TEL}}$: TEL gradient test.

4.2 SAR model

For the SAR model (2.2), let $V_n(\alpha_1,\beta) = [v_{n1}(\alpha_1,\beta),\dots,v_{nn}(\alpha_1,\beta)]'$, $P_{jn} = [p_{jn,rs}]$ and $Q_n = [Q_{n1},\dots,Q_{nn}]'$. Then $g_n(\theta)$ in (2.5) can be written as $g_n(\theta) = \frac{1}{n}\sum_{i=1}^n g_{ni}(\theta)$, where

$$g_{ni}(\theta) = \Big[\Big(v_{ni}^2(\alpha_1,\beta)-\sigma^2\Big)p_{1n,ii} + v_{ni}(\alpha_1,\beta)\sum_{j=1}^{i-1}(p_{1n,ij}+p_{1n,ji})v_{nj}(\alpha_1,\beta),$$

$$\dots,\Big(v_{ni}^2(\alpha_1,\beta)-\sigma^2\Big)p_{k_pn,ii} + v_{ni}(\alpha_1,\beta)\sum_{j=1}^{i-1}(p_{k_pn,ij}+p_{k_pn,ji})v_{nj}(\alpha_1,\beta), Q_{ni}'v_{ni}(\alpha_1,\beta)\Big]'.$$

Since $\mathbf{E}[g_{ni}(\theta_0)|v_{n1},\dots,v_{n,i-1}] = 0$ and $\mathbf{E}[g_{ni}(\theta_0)|v_{n1},\dots,v_{ni}] = g_{ni}(\theta_0)$ at the true value θ_0 of θ, $g_{ni}(\theta_0)$'s are martingale differences. Thus, we may consider GEL and TGEL estimators with $g_{ni}(\theta)$'s.

As one P_{jn} can be I_n and X_n is usually included in Q_n, we assume that $P_{k_pn} = I_n$ and $Q_n = [X_n, Q_{1n}]$, and rewrite the moment vector as $g_n(\theta) = [g_{nb}'(\theta), g_{na}'(\theta)]'$, where

$$g_{nb}(\theta) = \frac{1}{n}[V_n'(\alpha_1,\beta)P_{1n}V_n(\alpha_1,\beta)-\sigma^2\operatorname{tr}(P_{1n}),\dots,V_n'(\alpha_1,\beta)P_{k_p-1,n}V_n(\alpha_1,\beta)-\sigma^2\operatorname{tr}(P_{k_p-1,n}),V_n'(\alpha_1,\beta)Q_{1n}]'$$

and $g_{na}(\theta) = \frac{1}{n}[V_n'(\alpha_1,\beta)V_n(\alpha_1,\beta)-n\sigma^2, V_n'(\alpha_1,\beta)X_n]'$. TGEL estimators of β can be constructed with the $C(\alpha)$-moment $[I_{m_b}, -\bar{C}_{1n}]g_{ni}(\theta)$, where

$$\bar{C}_{1n} = \Big(\mathbf{E}\frac{\partial g_{nb}(\theta_0)}{\partial\alpha'}\Big)\Big(\mathbf{E}\frac{\partial g_{na}(\theta_0)}{\partial\alpha'}\Big)^{-1} = \begin{pmatrix} \frac{1}{n}\operatorname{tr}(P_{1n}) & 0 \\ \vdots & \vdots \\ \frac{1}{n}\operatorname{tr}(P_{k_{p-1},n}) & 0 \\ 0 & Q_{1n}'X_n(X_n'X_n)^{-1} \end{pmatrix}$$

does not involve unknown parameters.

In our Monte Carlo experiments, X_n contains 2 or 8 exogenous variables and each exogenous variable is randomly drawn from the standard normal distribution. The W_n is generated by the rook criterion and row-normalized to have row sums equal to one. We set the variance of v_{ni} to 1 and α_{10} is chosen such that $R^2 \equiv \operatorname{var}(X_n\alpha_{10})/[\operatorname{var}(X_n\alpha_{10})+1]$ is either 0.7 or 0.01. For estimation, we use two quadratic moments with $P_{1n} = W_n$ and $P_{2n} = I_n$, and the IV matrix is $[X_n, W_nX_n]$.

Figures 4–5 report the MBs, MADs and IDRs of various estimates of β.[19] In terms of MB, GEL and TGEL estimators perform better than GMM and TGMM estimators, especially when

[19] Note that the spatial dependence parameter β is often the parameter of interest in practice. The parameter

$\beta_0 \neq 0$ and $k_x = 8$. In terms of MAD and IDR, GMM, ET and EL have similar performance. For two-step estimates, we observe that TET and TEL perform better than TGMM in terms of MAD and IDR.

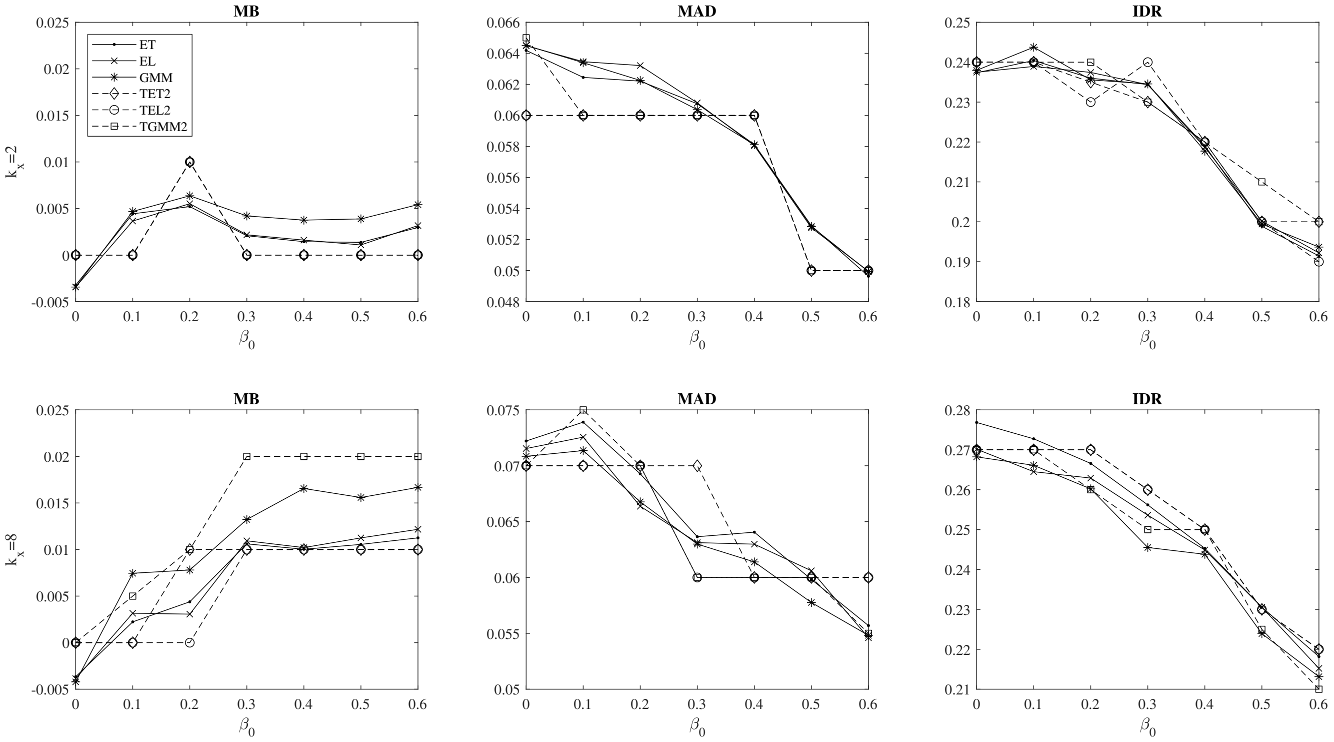

Figure 4: Performance of various estimates of β in the SAR model with $R^2 = 0.7$. The sample size n is 100.

5 Conclusion

This paper considers the TGEL estimation of parameters of interest via moment functions, which are martingale differences at the true parameter values, when there is a $\sqrt{n}$-consistent estimator of nuisance parameters or the nuisance parameters can be eliminated by an estimating function of parameters of interest. We propose to employ a $C(\alpha)$-type moment vector derived from proper linear combinations of the original moments. Such a two-step approach can elim-

space of β is $(1/\lambda_{\min}, 1/\lambda_{\max})$, where $\lambda_{\min}$ and $\lambda_{\max}$ are, respectively, the minimum negative and maximum positive eigenvalues of W_n. As our W_n is row-normalized, $\lambda_{\max} = 1$. For convenience, two-step estimates of β are searched over the interval $[-0.99, 0.99]$ with a grid size of 0.01 and other estimates do not use grid search but also use this parameter space for consistency, so we do not report TPs. The comparisons of the usual measures bias, SD and RMSE for various estimates are similar to those for robust measures, so only robust measures are reported.

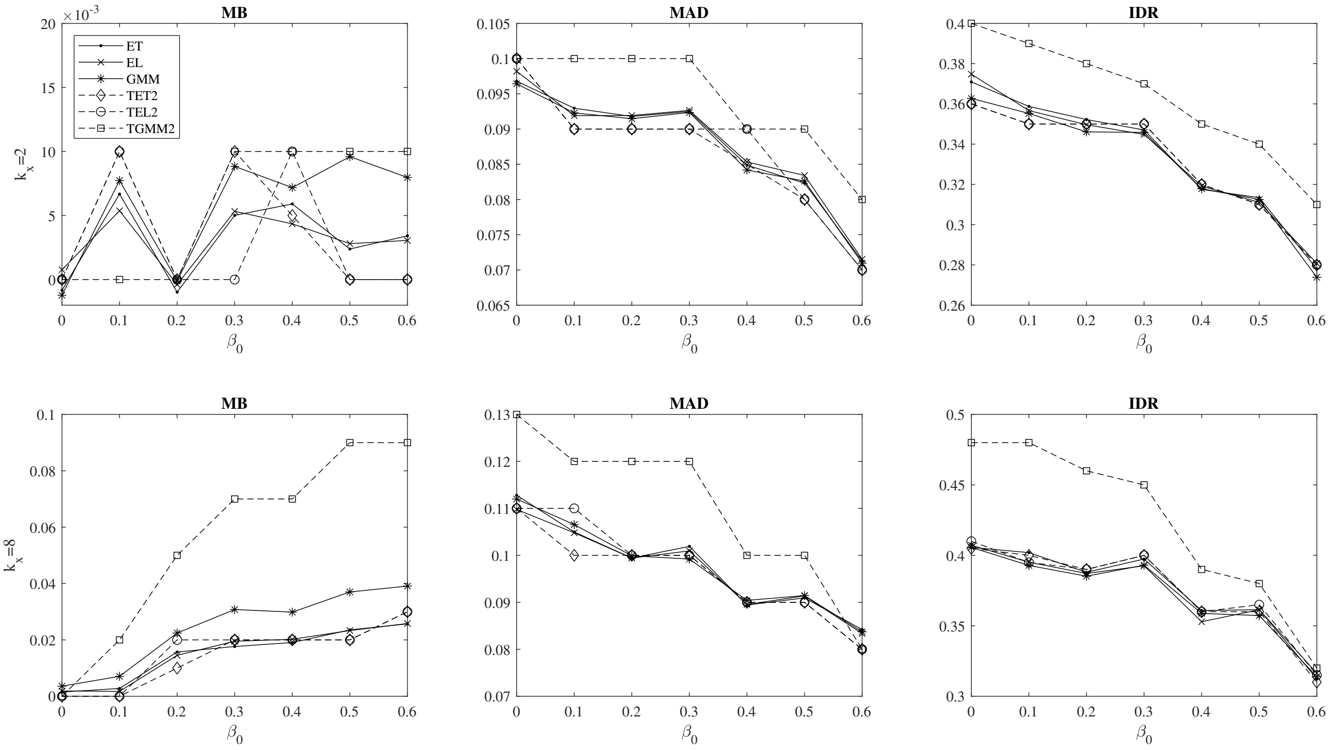

Figure 5: Performance of various estimates of β in the SAR model with $R^2 = 0.01$. The sample size n is 100.

inate the asymptotic impact of the nuisance parameter estimate, so that confidence intervals and various asymptotically pivotal tests can still be constructed with the TGEL objective function. Meanwhile, a TGEL approach can save computational time relative to the GEL approach due to its reduction in the number of moments and the number of parameters to be estimated. We show that TGEL does not lead to efficiency loss if the linearly combined $C(\alpha)$-type moment vector only reduces the number of moments by the number of nuisance parameters. The TGEL approach has a higher order bias advantage over its corresponding TGMM. In addition to the reduction in bias terms as for the ordinary GEL, TGEL does not have a bias term of TGMM which results from using an estimated feasible optimal weighting. Furthermore, we investigate various tests for parameter restrictions in both the ordinary and two-step GEL and GMM frameworks. Tests in the two frameworks can have equal local power.[20]

In a recent paper, Cattaneo et al. (2018) show that a first order bias emerges when the number of included covariates in the first step of a two-step GMM estimation procedure is

[20] Our two-step approaches are very useful in deriving simple and efficient estimators for some models. In Jin et al. (2018), TGMM is applied to dynamic short panel data models and is shown to generate closed-form root estimators of the dynamic parameter that are asymptotically as efficient as quasi maximum likelihood estimators.

large, rending standard inference procedures invalid. Our analysis is based on a fixed number of nuisance parameters. It is of interest to study in future research how to extend our analysis to such a situation, where a bias correction might be needed in addition to the use of a $C(\alpha)$-type formulation. Our Monte Carlo results show that two-step EL tests perform similarly to ordinary EL tests, which may have size distortions in small samples. More accurate inference for EL can be conducted by applying the Bartlett correction. It is of interest to study the Bartlett correctability of our two-step EL.[21]

Acknowledgements

We are grateful to Co-Editor Patrik Guggenberger and three anonymous referees for their valuable suggestions and comments. Some of issues on testing of hypotheses using a GMM approach with an initial consistent estimate in a two-step estimation framework were posed by Professor Junsen Zhang of the Chinese University of Hong Kong. We appreciate Junsen for bringing some of the testing issues addressed in this paper to our attention. This paper has been presented at SMU and in the Third Guangzhou Econometrics Workshop at the Sun Yat-Sen University, Guangzhou, China. We are grateful to Zhentao Shi, Liangjun Su, Zhenlin Yang, Jun Yu and other seminar participants for helpful comments. The first author gratefully acknowledges the financial support from the National Natural Science Foundation of China (No. 71973030, No. 71833004) and Program for Innovative Research Team of Shanghai University of Finance and Economics.

SUPPLEMENTARY MATERIAL

Fei Jin and Lung-fei Lee (2020): Supplement to "Efficient two-step generalized empirical likelihood estimation and tests with martingale differences", Econometric Theory Supplementary Material. To view, please visit:

[21] See Chen (1994), Lazar and Mykland (1999) and Chen and Cui (2006) for some studies on the Bartlett correctability in the presence of nuisance parameters.

REFERENCES

Ackerberg, D., X. Chen, J. Hahn and Z. Liao (2014) Asymptotic efficiency of semiparametric two-step GMM. *Review of Economic Studies* 81(3), 919–943.

Altonji, J. G. and L. M. Segal (1996) Small-sample bias in GMM estimation of covariance structures. *Journal of Business & Economic Statistics* 14(3), 353–366.

Anatolyev, S. (2005) GMM, GEL, serial correlation, and asymptotic bias. *Econometrica* 73(3), 983–1002.

Barro, R. J. (1977) Unanticipated money growth and unemployment in the United States. *American Economic Review* 67(2), 101–115.

Cattaneo, M. D., M. Jansson and X. Ma (2018) Two-step estimation and inference with possible many included covariates. *Review of Economic Studies* 86(3), 1095–1122.

Chen, S. X. (1994) Empirical likelihood confidence intervals for linear regression coefficients. *Journal of Multivariate Analysis* 49(1), 24–40.

Chen, S. X. and H. Cui (2006) On Bartlett correction of empirical likelihood in the presence of nuisance parameters. *Biometrika* 93(1), 215–220.

Chen, S. X. and H. Cui (2007) On the second-order properties of empirical likelihood with moment restrictions. *Journal of Econometrics* 141, 492–516.

Chernozhukov, V., D. Chetverikov, M. Demirer, E. Duflo, C. Hansen, W. Newey and J. Robins (2018a) Double/debiased machine learning for treatment and structural parameters. *Econometrics Journal* 21(1), C1–C68.

Chernozhukov, V., J. C. Escanciano, H. Ichimura, W. K. Newey and J. R. Robins (2018b) Locally robust semiparametric estimation. Working paper.

Chuang, C.-S. and N. H. Chan (2002) Empirical likelihood for autoregressive models, with applications to unstable time series. *Statistical Sinica* 12, 387–407.

Corcoran, S. A. (1998) Bartlett adjustment of empirical discrepancy statistics. *Biometrika* 85(4), 967–972.

Crepon, B., F. Kramarz and A. Trognon (1997) Parameters of interest, nuisance parameters and orthogonality conditions An application to autoregressive error component models. *Journal of Econometrics* 82(1), 135–156.

DiCiccio, T., P. Hall and J. Romano (1991) Empirical likelihood is Bartlett-correctable. *Annals of Statistics* 19(2), 1053–1061.

Dufour, J.-M., A. Trognon and P. Tuvaandorj (2017) Invariant tests based on M-estimators, estimating functions, and the generalized method of moments. *Econometric Reviews* 36(1–3), 182–204.

Fan, Y., S. Pastorello and E. Renault (2015) Maximization by parts in extremum estimation. *Econometrics Journal* 18(2), 147–17.

Frazier, D. T. and E. Renault (2017) Efficient two-step estimation via targeting. *Journal of Econometrics* 201, 212–227.

Gänsler, P. and W. Stute (1977) *Wahrscheinlichkeitstheorie*. Springer Verlag, New York.

Gouriéroux, C., A. Monfort and E. Renault (1996) Two-stage generalized moment method with applications to regressions with heteroscedasticity of unknown form. *Journal of Statistical Planning and Inference* 50(1), 37–63.

Guggenberger, P. (2008) Finite sample evidence suggesting a heavy tail problem of the generalized empirical likelihood estimator. *Econometric Reviews* 26(4–6), 526–541.

Guggenberger, P. and R. J. Smith (2005) Generalized empirical likelihood estimation and tests under partial, weak and strong identification. *Econometric Theory* 21, 667–709.

Hall, P. and C. Heyde (1980) *Martingale limit theory and its applications*. Academic Press, New York.

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

Hausman, J., R. Lewis, K. Menzel and W. Newey (2011) Properties of the CUE estimator and a modification with moments. *Journal of Econometrics* 165(1), 45–57.

Heckman, J. J. (1976) The common structure of statistical models of truncation, sample selection, and limited dependent variables and a simple estimator for such models. *Annals of Economic and Social Measurement* 5, 475–492.

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

Imbens, G. W., R. H. Spady and P. Johnson (1998) Information theoretic approaches to inference in moment condition models. *Econometrica* 66(2), 333–357.

Jin, F. and L.-F. Lee (2019) GEL estimation and tests of spatial autoregressive models. *Journal of Econometrics* 208(2), 585–612.

Jin, F., L.-F. Lee and J. Yu (2018) Sequential and efficient GMM estimation of dynamic short panel data models. Working paper, Fudan University.

Kelejian, H. H. and I. R. Prucha (2001) On the asymptotic distribution of the Moran I test statistic with applications. *Journal of Econometrics* 104, 219–257.

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

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

Lazar, N. A. and P. A. Mykland (1999) Empirical likelihood in the presence of nuisance parameters. *Biometrika* 86(1), 203–211.

Lee, L.-F. (2007) GMM and 2SLS estimation of mixed regressive, spatial autoregressive models. *Journal of Econometrics* 137, 489–514.

Lee, L.-F. and J. Yu (2012) The $C(\alpha)$-type gradient test for spatial dependence in spatial autoregressive models. *Letters in Spatial and Resource Sciences* 5(3), 119–135.

Mittelhammer, R. C., G. G. Judge and R. Schoenberg (2005) Empirical evidence concerning the finite sample performance of EL-type structural equation estimation and inference methods. In D. W. K. Andrews and J. H. Stock (eds.), *Identification and inference for econometric models: Essays in honor of Thomas Rothenberg*, Cambridge University Press, 282–305.

Moran, P. A. P. (1950) Notes on continuous stochastic phenomena. *Biometrika* 35, 255–260.

Nagar, A. L. (1959) The bias and moment matrix of the general k-class estimators of the parameters in simultaneous equations. *Econometrica* 27, 575–595.

Newey, W. K. (1984) A method of moments interpretation of sequential estimators. *Economics Letters* 14, 201–206.

Newey, W. K., J. J. Ramalho and R. J. Smith (2005) Asymptotic bias for GMM and GEL estimators with estimated nuisance parameters. In D. W. K. Andrews and J. H. Stock (eds.), *Identification and inference for econometric models: Essays in honor of Thomas Rothenberg*, Cambridge University Press, 245–281.

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

Neyman, J. (1959) Optimal asymptotic tests of composite statistical hypotheses. In U. Grenander (ed.), *Probability and Statistics, the Harald Cramer Volume*, Wiley, New York.

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

Owen, A. (1991) Empirical likelihood for linear models. *Annals of Statistics* 19, 1725–1747.

Pagan, A. (1984) Econometric issues in the analysis of regressions with generated regressors. *International Economic Review* 25(1), 221–247.

Pagan, A. (1986) Two stage and related estimators and their applications. *Review of Economic Studies* 53(4), 517–538.

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

Ramalho, J. J. (2002) *Alternative estimation methods and specification tests for moment condition models*. Ph.D. thesis, University of Bristol.

Rivers, D. and Q. H. Vuong (1988) Limited information estimators and exogeneity tests for simultaneous probit models. *Journal of Econometrics* 39(3), 347–366.

Rothenberg, T. J. (1984) Approximating the distributions of econometric estimators and test statistics. In Z. Griliches and M. D. Intriligator (eds.), *Handbook of Econometrics*, vol. II, Elsevier Science Publishers BV, 881–935.

Ruud, P. A. (2000) *An Introduction to Classical Econometric Theory*. Oxford University Press.

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

Smith, R. J. (2011) GEL criteria for moment condition models. *Econometric Theory* 27, 1192–1235.

Song, P. X.-K., Y. Fan and J. D. Kalbfleisch (2005) Maximization by parts in likelihood inference. *Journal of the American Statistical Association* 100(472), 1145–1158.

Staiger, D. and J. H. Stock (1997) Instrumental variables regression with weak instruments. *Econometrica* 65, 557–586.

Stock, J. H. and J. H. Wright (2000) GMM with weak identification. *Econometrica* 68(5), 1055–1096.

Trognon, A. and C. Gouriéroux (1990) A note on the efficiency of two-step estimation methods. In P. Champsaur (ed.), *Essays in Honor of Edmond Malinvaud 3*, MIT Press, 232–248.

Wilde, J. (2008) A note on GMM estimation of probit models with endogenous regressors. *Statistical Paper* 49, 471–484.

Appendix A List of notations

$\gamma = (\alpha', \beta')'$, γ is $k_\gamma \times 1$, α is $k_\alpha \times 1$, β is $k_\beta \times 1$.

$g_n(\gamma) = \frac{1}{n}\sum_{i=1}^n g_{ni}(\gamma)$, $\bar{g}_n(\gamma) = \mathbf{E}[g_n(\gamma)]$, $g_n(\gamma) = (g'_{nb}(\gamma), g'_{na}(\gamma))'$, $g_n(\gamma)$ is $m_g \times 1$, $g_{nb}(\gamma)$ is $m_b \times 1$, $g_{na}(\gamma)$ is $m_a \times 1$. $g_{ni} = g_{ni}(\gamma_0)$, $g_n = g_n(\gamma_0)$, $g_{nb} = g_{nb}(\gamma_0)$, γ_j is the jth element of γ, $g_{ni}^{(j)} = \frac{\partial g_{ni}(\gamma_0)}{\partial \gamma_j}$, $G_{ni\alpha}(\gamma) = \frac{\partial g_{ni}(\gamma)}{\partial \alpha'}$, $G_{ni\alpha} = G_{ni\alpha}(\gamma_0)$, $G_{n\alpha}(\gamma) = \frac{1}{n}\sum_{i=1}^n G_{ni\alpha}(\gamma)$, $\bar{G}_{n\alpha} = \mathbf{E}[G_{n\alpha}(\gamma_0)]$, $G_{ni\alpha}^{(j)}(\gamma) = \frac{\partial G_{ni\alpha}(\gamma)}{\partial \gamma_j}$, $G_{ni\alpha}^{(j)} = G_{ni\alpha}^{(j)}(\gamma_0)$, $G_{n\alpha}^{(j)} = \frac{1}{n}\sum_{i=1}^n G_{ni\alpha}^{(j)}$, $\bar{G}_{n\alpha}^{(j)} = \mathbf{E}[G_{n\alpha}^{(j)}(\gamma_0)]$.

$G_{ni\beta} = \frac{\partial g_{ni}(\gamma_0)}{\partial \beta'}$, $G_{n\beta} = \frac{\partial g_n(\gamma_0)}{\partial \beta'}$, $\bar{G}_{n\beta} = \mathbf{E}(\frac{\partial g(\gamma_0)}{\partial \beta'})$, $\bar{G}_{n\beta}^{(j)} = \mathbf{E}(\frac{\partial^2 g_n(\gamma_0)}{\partial \gamma_j \partial \beta'})$, $G_n(\gamma) = \frac{\partial g_n(\gamma)}{\partial \gamma'}$, $\bar{C}_n^{(j)} = \frac{\partial \bar{C}_n(\gamma_0)}{\partial \gamma_j}$, $\alpha_n^{(j)} = \frac{\partial \alpha_n(\beta_0)}{\partial \beta_j}$.

$\Omega_n(\gamma) = \frac{1}{n}\sum_{i=1}^n g_{ni}(\gamma) g'_{ni}(\gamma)$, $\bar{\Omega}_n = \mathbf{E}[\Omega_n(\gamma_0)]$, $\bar{\Omega}_{naa} = \mathbf{E}[n g_{na}(\gamma_0) g'_{na}(\gamma_0)]$, $\bar{G}_{nb\beta} = \mathbf{E}(\frac{\partial g_{nb}(\gamma_0)}{\partial \beta'})$, $\bar{G}_{nb\alpha} = \mathbf{E}(\frac{\partial g_{nb}(\gamma_0)}{\partial \alpha'})$.

$d_{ni}(\gamma) = \hat{C}_n g_{ni}(\gamma)$, $d_n(\gamma) = \hat{C}_n g_n(\gamma)$, $\Omega_{nd}(\gamma) = \hat{C}_n \Omega_n(\gamma)\hat{C}_n'$, $\bar{\Omega}_{nd} = \bar{C}_n\bar{\Omega}_n\bar{C}_n'$.

$\bar{D}_{n\beta} = \bar{C}_n\bar{G}_{n\beta}$, $\bar{\Sigma}_{nd} = (\bar{D}_{n\beta}'\bar{\Omega}_{nd}^{-1}\bar{D}_{n\beta})^{-1}$, $\bar{H}_{nd} = \bar{\Sigma}_{nd}\bar{D}_{n\beta}'\bar{\Omega}_{nd}^{-1}$, $\bar{P}_{nd} = \bar{\Omega}_{nd}^{-1} - \bar{\Omega}_{nd}^{-1}\bar{D}_{n\beta}\bar{\Sigma}_{nd}\bar{D}_{n\beta}'\bar{\Omega}_{nd}^{-1}$.

$\psi_{n\mu} = -\bar{P}_{nd}\sqrt{n}\bar{C}_n g_n$, $\psi_{n\beta} = -\bar{H}_{nd}\sqrt{n}\bar{C}_n g_n$, $\psi_{n\dot{\alpha}} = \sqrt{n}(\dot{\alpha}_n(\beta_0)-\alpha_0)+\alpha_{n\beta}\psi_{n\beta}$, $\psi_{n\check{C}} = \psi_{nC}+\sum_{j=1}^{k_\alpha}\bar{C}_n^{(j)}\psi_{n\check{\alpha}j}+\sum_{j=1}^{k_\beta}\bar{C}_n^{(k_\alpha+j)}\psi_{n\beta j}$, and $\psi_{n\dot{C}} = \psi_{nC}+\sum_{j=1}^{k_\alpha}\bar{C}_n^{(j)}\psi_{n\dot{\alpha}j}+\sum_{j=1}^{k_\beta}\bar{C}_n^{(k_\alpha+j)}\psi_{n\beta j}$.

$e_{k_\beta j}$ is the jth unit column vector of dimension k_β.

Appendix B Proofs

In this section, MVT will denote the mean value theorem.

Proof of Theorem 1. We first prove the consistency of $\hat{\beta}_{\text{TGEL}}$. It is shown in the proof of Lemma H.1(i) that $\sup_{\beta\in\mathcal{B}}\|g_n(\check{\alpha},\beta) - \bar{g}_n(\alpha_0,\beta)\| = o_p(1)$ and $\sup_{\beta\in\mathcal{B}}\|\bar{g}_n(\alpha_0,\beta)\| = O(1)$, where $\bar{g}_n(\gamma) = \mathbf{E}[g_n(\gamma)]$. We may similarly prove that $\sup_{\beta\in\mathcal{B}}\|C_n(\check{\alpha},\beta)-\bar{C}_n(\alpha_0,\beta)\| = o_p(1)$ and $\sup_{\beta\in\beta}\|\bar{C}_n(\alpha_0,\beta)\| = O(1)$. It follows that

$$
\begin{aligned}
&\sup_{\beta\in\mathcal{B}}\|C_n(\check{\alpha},\beta)g_n(\check{\alpha},\beta) - \bar{C}_n(\alpha_0,\beta)\bar{g}_n(\alpha_0,\beta)\|\\
&= \sup_{\beta\in\mathcal{B}}\|[C_n(\check{\alpha},\beta) - \bar{C}_n(\alpha_0,\beta)][g_n(\check{\alpha},\beta) - \bar{g}_n(\alpha_0,\beta)] + \bar{C}_n(\alpha_0,\beta)[g_n(\check{\alpha},\beta) - \bar{g}_n(\alpha_0,\beta)]\\
&\qquad + [C_n(\check{\alpha},\beta) - \bar{C}_n(\alpha_0,\beta)]\bar{g}_n(\alpha_0,\beta)\|\\
&\le \sup_{\beta\in\mathcal{B}}\|C_n(\check{\alpha},\beta) - \bar{C}_n(\alpha_0,\beta)\|\sup_{\beta\in\mathcal{B}}\|g_n(\check{\alpha},\beta) - \bar{g}_n(\alpha_0,\beta)\| + \sup_{\beta\in\mathcal{B}}\|\bar{C}_n(\alpha_0,\beta)\|\sup_{\beta\in\mathcal{B}}\|g_n(\check{\alpha},\beta) - \bar{g}_n(\alpha_0,\beta)\|\\
&\quad + \sup_{\beta\in\mathcal{B}}\|C_n(\check{\alpha},\beta) - \bar{C}_n(\alpha_0,\beta)\|\sup_{\beta\in\mathcal{B}}\|\bar{g}_n(\alpha_0,\beta)\|\\
&= o_p(1).
\end{aligned}
$$

By Lemma H.4(ii), $C_n(\check{\alpha},\hat{\beta}_{\text{TGEL}})g_n(\check{\alpha},\hat{\beta}_{\text{TGEL}}) = O_p(n^{-1/2})$. Then

$$
\begin{aligned}
&\|\bar{C}_n(\alpha_0,\hat{\beta}_{\text{TGEL}})\bar{g}_n(\alpha_0,\hat{\beta}_{\text{TGEL}})\|\\
&\quad \le \|\bar{C}_n(\alpha_0,\hat{\beta}_{\text{TGEL}})\bar{g}_n(\alpha_0,\hat{\beta}_{\text{TGEL}}) - C_n(\check{\alpha},\hat{\beta}_{\text{TGEL}})g_n(\check{\alpha},\hat{\beta}_{\text{TGEL}})\| + \|C_n(\check{\alpha},\hat{\beta}_{\text{TGEL}})g_n(\check{\alpha},\hat{\beta}_{\text{TGEL}})\| = o_p(1).
\end{aligned}
$$

Since $\bar{C}_n(\alpha_0,\beta)\bar{g}_n(\alpha_0,\beta)$ is uniquely zero at $\beta = \beta_0$ for large enough n and it is uniformly equicontinuous, $\|\bar{C}_n(\alpha_0,\beta)\bar{g}_n(\alpha_0,\beta)\|$ must be bounded away from zero outside of any neighborhood of β_0 for large enough n. Hence $\hat{\beta}_{\text{TGEL}}$ must be inside any neighborhood of β_0 with probability approaching one (w.p.a.1), i.e., $\hat{\beta}_{\text{TGEL}} = \beta_0 + o_p(1)$. As $C_n(\check{\alpha}, \hat{\beta}_{\text{TGEL}})g_n(\check{\alpha}, \hat{\beta}_{\text{TGEL}}) = O_p(n^{-1/2})$, Lemma H.3 holds for $\bar{\beta} = \hat{\beta}_{\text{TGEL}}$ and $h_{ni}(\beta) = C_n(\check{\alpha}, \beta)g_{ni}(\check{\alpha}, \beta)$. Therefore,

$$\hat{\mu}_{\text{TGEL}} = \arg \max_{\mu \in \Lambda_n(\check{\alpha}, \hat{\beta}_{\text{TGEL}})} \sum_{i=1}^{n} \rho\Big(\mu' C_n(\check{\alpha}, \hat{\beta}_{\text{TGEL}}) g_{ni}(\check{\alpha}, \hat{\beta}_{\text{TGEL}})\Big)$$

exists w.p.a.1, and $\hat{\mu}_{\text{TGEL}} = O_p(n^{-1/2})$.

We next investigate the asymptotic distribution of $\hat{\theta}_{\text{TGEL}}$. Let $v_{ni}(\alpha, \theta) = \mu' C_n(\alpha, \beta) g_{ni}(\alpha, \beta)$, $C_n^{(j)}(\alpha, \beta) = \frac{\partial C_n(\alpha,\beta)}{\partial \gamma_j}$, where γ_j is the jth element of $\gamma = (\alpha', \beta')'$, $q_{ni}(\alpha, \theta) = [\frac{\partial v_{ni}(\alpha,\theta)}{\partial \beta'}, \frac{\partial v_{ni}(\alpha,\theta)}{\partial \mu'}]'$, $h_{ni}(\alpha, \theta) = \begin{pmatrix} G'_{ni\beta}(\alpha,\beta) C'_n(\alpha,\beta)\mu \\ C_n(\alpha,\beta) g_{ni}(\alpha,\beta) \end{pmatrix}$, and $\hat{\Delta}_{ni}(\alpha, \theta) = [\mu' C_n^{(k_\alpha+1)}(\alpha, \beta) g_{ni}(\alpha, \beta), \ldots, \mu' C_n^{(k_\alpha+k_\beta)}(\alpha, \beta) g_{ni}(\alpha, \beta)]'$. Then $q_{ni}(\alpha, \theta) = h_{ni}(\alpha, \theta) + (\begin{smallmatrix} \hat{\Delta}_{ni}(\alpha,\theta) \\ 0 \end{smallmatrix})$, where $\hat{\Delta}_{ni}(\alpha, \theta)$ is due to the unknown β in $C_n(\check{\alpha}, \beta)$. The first order condition of $\hat{\theta}_{\text{TGEL}}$ is $\sum_{i=1}^n \rho_1(v_{ni}(\check{\alpha}, \hat{\theta}_{\text{TGEL}})) q_{ni}(\check{\alpha}, \hat{\theta}_{\text{TGEL}}) = 0$. Applying the MVT to this condition at $\theta = \theta_0$ yields

$$\begin{aligned} 0 &= \sum_{i=1}^{n} \rho_1(0) q_{ni}(\check{\alpha}, \theta_0) + \sum_{i=1}^{n} \Big[\rho_1(v_{ni}(\check{\alpha}, \ddot{\theta})) \frac{\partial q_{ni}(\check{\alpha}, \ddot{\theta})}{\partial \theta'} + \rho_2(v_{ni}(\check{\alpha}, \ddot{\theta})) q_{ni}(\check{\alpha}, \ddot{\theta}) q'_{ni}(\check{\alpha}, \ddot{\theta})\Big](\hat{\theta}_{\text{TGEL}} - \theta_0) \\ &= \sum_{i=1}^{n} \rho_1(0) h_{ni}(\check{\alpha}, \theta_0) + \sum_{i=1}^{n} \Big[\rho_1(v_{ni}(\check{\alpha}, \ddot{\theta})) \Big(\frac{\partial h_{ni}(\check{\alpha}, \ddot{\theta})}{\partial \theta'} + \begin{pmatrix} \frac{\partial \hat{\Delta}_{ni}(\check{\alpha},\ddot{\theta})}{\partial \theta'} \\ 0 \end{pmatrix}\Big) \\ &\quad + \rho_2(v_{ni}(\check{\alpha}, \ddot{\theta})) q_{ni}(\check{\alpha}, \ddot{\theta}) q'_{ni}(\check{\alpha}, \ddot{\theta})\Big](\hat{\theta}_{\text{TGEL}} - \theta_0), \end{aligned} \tag{B.1}$$

where $\ddot{\theta} = (\ddot{\beta}', \ddot{\mu}')'$ lies between θ_0 and $\hat{\theta}_{\text{TGEL}}$,

$$\frac{\partial \hat{\Delta}_{ni}(\alpha, \theta)}{\partial \theta_j} = \Big[\mu' \frac{\partial [C_n^{(k_\alpha+1)}(\alpha, \beta) g_{ni}(\alpha, \beta)]}{\partial \beta_j}, \ldots, \mu' \frac{\partial [C_n^{(k_\alpha+k_\beta)}(\alpha, \beta) g_{ni}(\alpha, \beta)]}{\partial \beta_j}\Big]'$$

for $1 \leq j \leq k_\beta$, and $\frac{\partial \hat{\Delta}_{ni}(\alpha,\theta)}{\partial \theta_j} = [e'_{m_b, j-k_\beta} C_n^{(k_\alpha+1)}(\alpha, \beta) g_{ni}(\alpha, \beta), \ldots, e'_{m_b, j-k_\beta} C_n^{(k_\alpha+k_\beta)}(\alpha, \beta) g_{ni}(\alpha, \beta)]'$ for $k_\beta + 1 \leq j \leq k_\theta$. With $\hat{\mu}_{\text{TGEL}} = O_p(n^{-1/2})$ and the consistency of $\hat{\beta}_{\text{TGEL}}$, $\max_{1 \leq i \leq n} |v_{ni}(\check{\alpha}, \ddot{\theta})| = o_p(1)$ by Lemma H.2. Then $\frac{1}{n}\sum_{i=1}^n \rho_1(v_{ni}(\check{\alpha}, \ddot{\theta})) g_{ni}(\check{\alpha}, \ddot{\beta}) = -g_n(\check{\alpha}, \ddot{\beta}) + o_p(1) = -\bar{g}_n(\alpha_0, \ddot{\beta}) + o_p(1) = -\bar{g}_n(\alpha_0, \beta_0) + o_p(1) = o_p(1)$. It follows that $\frac{1}{n}\sum_{i=1}^n \rho_1(v_{ni}(\check{\alpha}, \ddot{\theta})) \frac{\partial \hat{\Delta}_{ni}(\check{\alpha},\ddot{\theta})}{\partial \theta_j} = o_p(1)$ for $k_\beta + 1 \leq j \leq k_\theta$. With $\hat{\mu}_{\text{TGEL}} = O_p(n^{-1/2})$, similarly, we have $\frac{1}{n}\sum_{i=1}^n \rho_1(v_{ni}(\check{\alpha}, \ddot{\theta})) \frac{\partial \hat{\Delta}_{ni}(\check{\alpha},\ddot{\theta})}{\partial \theta_j} = o_p(1)$ for $1 \leq j \leq k_\beta$,

$$\frac{1}{n} \sum_{i=1}^{n} \rho_2(v_{ni}(\check{\alpha}, \ddot{\theta})) h_{ni}(\check{\alpha}, \ddot{\theta}) \hat{\Delta}'_{ni}(\check{\alpha}, \ddot{\theta}) = o_p(1)$$

and $\frac{1}{n}\sum_{i=1}^n \rho_2(v_{ni}(\check{\alpha}, \ddot{\theta})) \hat{\Delta}_{ni}(\check{\alpha}, \ddot{\theta}) \hat{\Delta}'_{ni}(\check{\alpha}, \ddot{\theta}) = o_p(1)$. Then, by (B.1), the term $\hat{\Delta}_{ni}(\alpha, \theta)$ for the derivative $\frac{\partial v_{ni}(\alpha,\theta)}{\partial \beta}$ has no impact on the asymptotic distribution of $\hat{\theta}_{\text{TGEL}}$. Then, as in the proof of Theorem 3.2 in Newey and Smith (2004), we may derive from (B.1) that

$$\sqrt{n}(\hat{\theta}_{\text{TGEL}} - \theta_0) = -\begin{pmatrix} \bar{H}_{nd} \\ \bar{P}_{nd} \end{pmatrix} \bar{C}_n \sqrt{n} g_n(\check{\alpha}, \beta_0) + o_p(1). \tag{B.2}$$

Applying the MVT to $\bar{C}_n \sqrt{n} g_n(\check{\alpha}, \beta_0)$ yields $\bar{C}_n \sqrt{n} g_n(\check{\alpha}, \beta_0) = \bar{C}_n \sqrt{n} g_n(\gamma_0) + \bar{C}_n G_{n\alpha}(\ddot{\alpha}, \beta_0)\sqrt{n}(\check{\alpha} - \alpha_0)$, where $\ddot{\alpha}$ lies between α_0 and $\check{\alpha}$. Since $\sqrt{n}(\check{\alpha} - \alpha_0) = O_p(1)$, $G_{n\alpha}(\ddot{\alpha}, \beta_0) = \bar{G}_{n\alpha} + o_p(1)$ under Assumption 1(ix). Using $\bar{C}_n \bar{G}_{n\alpha} = 0$, (B.2) becomes

$$\sqrt{n}(\hat{\theta}_{\text{TGEL}} - \theta_0) = -\begin{pmatrix} \bar{H}_{nd} \\ \bar{P}_{nd} \end{pmatrix} \bar{C}_n \frac{1}{\sqrt{n}} \sum_{i=1}^{n} g_{ni}(\gamma_0) + o_p(1). \tag{B.3}$$

With the asymptotic distribution of $\sqrt{n} g_n(\gamma_0)$ in Assumption 1(xi),

$$\sqrt{n}(\hat{\theta}_{\text{TGEL}} - \theta_0) \xrightarrow{d} N\Big(0, \lim_{n\to\infty} \Big(\begin{smallmatrix} \bar{\Sigma}_{nd} & 0 \\ 0 & \bar{P}_{nd} \end{smallmatrix}\Big)\Big).$$

For $\hat{\beta}_{\text{TGEL2}}$, $\sup_{\beta\in\mathcal{B}} \|C_n(\dot{\alpha}_n(\beta), \beta) - \bar{C}_n(\alpha_n(\beta), \beta)\| = o_p(1)$ and $\sup_{\beta\in\mathcal{B}} \|\bar{C}_n(\alpha_n(\beta), \beta)\| = O(1)$ by arguments similar to those for $\sup_{\beta\in\mathcal{B}} \|g_n(\dot{\alpha}_n(\beta), \beta) - \bar{g}_n(\alpha_n(\beta), \beta)\| = o_p(1)$ and $\sup_{\beta\in\mathcal{B}} \|\bar{g}_n(\alpha_n(\beta), \beta)\| = O(1)$ in the proof of Lemma H.1(i). It follows that

$$\sup_{\beta\in\mathcal{B}} \|C_n(\dot{\alpha}_n(\beta), \beta) g_n(\dot{\alpha}_n(\beta), \beta) - \bar{C}_n(\alpha_n(\beta), \beta)\bar{g}_n(\alpha_n(\beta), \beta)\| = o_p(1).$$

Since $\lim_{n\to\infty} \bar{C}_n(\alpha_n(\beta), \beta)\bar{g}_n(\alpha_n(\beta), \beta)$ is uniquely zero at $\beta = \beta_0$, $\bar{C}_n(\alpha_n(\beta), \beta)\bar{g}_n(\alpha_n(\beta), \beta)$ must be bounded away from zero outside of any neighborhood of β_0. Hence $\hat{\beta}_{\text{TGEL2}}$ must be inside any neighborhood of β_0 w.p.a.1, i.e., $\hat{\beta}_{\text{TGEL2}} = \beta_0 + o_p(1)$. By Lemma H.3, $\hat{\mu}_{\text{TGEL2}} = O_p(n^{-1/2})$. For the asymptotic distribution of $\hat{\theta}_{\text{TGEL2}}$, compared with that of $\hat{\theta}_{\text{TGEL}}$, we need to take into account the additional derivative term due to the unknown β in $\dot{\alpha}(\beta)$. This additional term does not affect the asymptotic distribution of $\hat{\theta}_{\text{TGEL2}}$, as the derivative term due to the unknown β in $C_n(\check{\alpha}, \beta)$ for the asymptotic distribution of $\hat{\theta}_{\text{TGEL}}$. Then we may similarly show that $\hat{\theta}_{\text{TGEL2}}$ has the same asymptotic distribution as that of $\hat{\theta}_{\text{TGEL}}$.

(ii) For (ii)–(iv), we only prove the results for $\hat{\beta}_{\text{TGEL}}$, since those for $\hat{\beta}_{\text{TGEL2}}$ can be similarly proved. As $\rho(0) = \frac{1}{n}\sum_{i=1}^{n} \rho(0 \cdot d_{ni}(\check{\alpha}, \hat{\beta}_{\text{TGEL}}))$, by a first order Taylor expansion of $\frac{1}{n}\sum_{i=1}^{n} \rho(0 \cdot d_{ni}(\check{\alpha}, \hat{\beta}_{\text{TGEL}}))$ at $\hat{\mu}_{\text{TGEL}}$ and using the first order condition of $\hat{\mu}_{\text{TGEL}}$,

$$\rho(0) = \frac{1}{n}\sum_{i=1}^{n} \rho\big(\hat{\mu}'_{\text{TGEL}} d_{ni}(\check{\alpha}, \hat{\beta}_{\text{TGEL}})\big) + \frac{1}{2n}\sum_{i=1}^{n} \rho_2\big(\ddot{\mu}' d_{ni}(\check{\alpha}, \hat{\beta}_{\text{TGEL}})\big)\hat{\mu}'_{\text{TGEL}} d_{ni}(\check{\alpha}, \hat{\beta}_{\text{TGEL}}) d'_{ni}(\check{\alpha}, \hat{\beta}_{\text{TGEL}})\hat{\mu}_{\text{TGEL}},$$

where $\ddot{\mu}$ lies between 0 and $\hat{\mu}_{\text{TGEL}}$. Denote $\bar{M}_{nd} = I_{m_b} - \bar{\Omega}_{nd}^{-1/2}\bar{D}_{n\beta}(\bar{D}'_{n\beta}\bar{\Omega}_{nd}^{-1}\bar{D}_{n\beta})^{-1}\bar{D}'_{n\beta}\bar{\Omega}_{nd}^{-1/2}$. Then,

$$
\begin{aligned}
&2n\Big[\frac{1}{n}\sum_{i=1}^{n}\rho\Big(\hat{\mu}'_{\text{TGEL}}d_{ni}(\check{\alpha},\hat{\beta}_{\text{TGEL}})\Big) - \rho(0)\Big]\\
&= -\sqrt{n}\hat{\mu}'_{\text{TGEL}}\frac{1}{n}\sum_{i=1}^{n}\rho_2\Big(\ddot{\mu}'_n d_{ni}(\check{\alpha},\hat{\beta}_{\text{TGEL}})\Big)d_{ni}(\check{\alpha},\hat{\beta}_{\text{TGEL}})d'_{ni}(\check{\alpha},\hat{\beta}_{\text{TGEL}})\sqrt{n}\hat{\mu}_{\text{TGEL}}\\
&= \sqrt{n}\hat{\mu}'_{\text{TGEL}}\frac{1}{n}\sum_{i=1}^{n}d_{ni}(\check{\alpha},\hat{\beta}_{\text{TGEL}})d'_{ni}(\check{\alpha},\hat{\beta}_{\text{TGEL}})\sqrt{n}\hat{\mu}_{\text{TGEL}} + o_p(1) \qquad \text{(B.4)}\\
&= [\bar{C}_n\sqrt{n}g_n(\gamma_0)]'\bar{P}_{nd}\bar{C}_n\sqrt{n}g_n(\gamma_0) + o_p(1)\\
&= [\bar{\Omega}_{nd}^{-1/2}\bar{C}_n\sqrt{n}g_n(\gamma_0)]'\bar{M}_{nd}\bar{\Omega}_{nd}^{-1/2}\bar{C}_n\sqrt{n}g_n(\gamma_0) + o_p(1)\\
&\xrightarrow{d} \chi^2(m_b - k_\beta),
\end{aligned}
$$

where the third equality uses the properties $\frac{1}{n}\sum_{i=1}^{n} d_{ni}(\check{\alpha},\hat{\beta}_{\text{TGEL}})d'_{ni}(\check{\alpha},\hat{\beta}_{\text{TGEL}}) = \bar{\Omega}_{nd} + o_p(1)$, (B.3) and $\bar{P}_{nd}\bar{\Omega}_{nd}\bar{P}_{nd} = \bar{P}_{nd}$; and the asymptotic distribution follows because $\bar{\Omega}_{nd}^{-1/2}\bar{C}_n\sqrt{n}g_n(\gamma_0)$ is asymptotically multivariate standard normal and $\bar{M}_{nd}$ is a projection matrix with rank $(m_b - k_\beta)$.

(iii) For $\hat{\gamma}_{\text{GEL}} = (\hat{\alpha}'_{\text{GEL}}, \hat{\beta}'_{\text{GEL}})'$, by (A.8) in Newey and Smith (2004),

$$\sqrt{n}(\hat{\gamma}_{\text{GEL}} - \gamma_0) = -(\bar{G}'_n\bar{\Omega}_n^{-1}\bar{G}_n)^{-1}\bar{G}'_n\bar{\Omega}_n^{-1}\sqrt{n}g_n(\gamma_0) + o_p(1).$$

As $\bar{G}_n = [\bar{G}_{n\alpha}, \bar{G}_{n\beta}]$, by the block matrix inverse formula,

$$
\begin{aligned}
\sqrt{n}(\hat{\beta}_{\text{GEL}} - \beta_0) = &-\Big\{\bar{G}'_{n\beta}[\bar{\Omega}_n^{-1} - \bar{\Omega}_n^{-1}\bar{G}_{n\alpha}(\bar{G}'_{n\alpha}\bar{\Omega}_n^{-1}\bar{G}_{n\alpha})^{-1}\bar{G}'_{n\alpha}\bar{\Omega}_n^{-1}]\bar{G}_{n\beta}\Big\}^{-1}\\
&\cdot \bar{G}'_{n\beta}[\bar{\Omega}_n^{-1} - \bar{\Omega}_n^{-1}\bar{G}_{n\alpha}(\bar{G}'_{n\alpha}\bar{\Omega}_n^{-1}\bar{G}_{n\alpha})^{-1}\bar{G}'_{n\alpha}\bar{\Omega}_n^{-1}]\sqrt{n}g_n(\gamma_0) + o_p(1),
\end{aligned}
\qquad \text{(B.5)}
$$

where $\bar{\Omega}_n^{-1} - \bar{\Omega}_n^{-1}\bar{G}_{n\alpha}(\bar{G}'_{n\alpha}\bar{\Omega}_n^{-1}\bar{G}_{n\alpha})^{-1}\bar{G}'_{n\alpha}\bar{\Omega}_n^{-1} = \bar{\Omega}_n^{-1/2}[I_{m_g} - \bar{\Omega}_n^{-1/2}\bar{G}_{n\alpha}(\bar{G}'_{n\alpha}\bar{\Omega}_n^{-1}\bar{G}_{n\alpha})^{-1}\bar{G}'_{n\alpha}\bar{\Omega}_n^{-1/2}]\bar{\Omega}_n^{-1/2}$ with $m_g = m_b + m_a$. It follows that the asymptotic variance of $\hat{\beta}_{\text{GEL}}$ is

$$\lim_{n\to\infty}\Big\{\bar{G}'_{n\beta}\bar{\Omega}_n^{-1/2}[I_{m_g} - \bar{\Omega}_n^{-1/2}\bar{G}_{n\alpha}(\bar{G}'_{n\alpha}\bar{\Omega}_n^{-1}\bar{G}_{n\alpha})^{-1}\bar{G}'_{n\alpha}\bar{\Omega}_n^{-1/2}]\bar{\Omega}_n^{-1/2}\bar{G}_{n\beta}\Big\}^{-1}.$$

On the other hand, the asymptotic variance of $\hat{\beta}_{\text{TGEL}}$ is

$$\lim_{n\to\infty}[\bar{G}'_{n\beta}\bar{C}'_n\bar{\Omega}_{nd}^{-1}\bar{C}_n\bar{G}_{n\beta}]^{-1} = \lim_{n\to\infty}[\bar{G}'_{n\beta}\bar{\Omega}_n^{-1/2}\cdot\bar{\Omega}_n^{1/2}\bar{C}'_n\bar{\Omega}_{nd}^{-1}\bar{C}_n\bar{\Omega}_n^{1/2}\cdot\bar{\Omega}_n^{-1/2}\bar{G}_{n\beta}]^{-1}.$$

Note that $(\bar{\Omega}_n^{1/2}\bar{C}'_n)'\bar{\Omega}_n^{-1/2}\bar{G}_{n\alpha} = 0$, and $\bar{\Omega}_n^{1/2}\bar{C}'_n$ and $\bar{\Omega}_n^{-1/2}\bar{G}_{n\alpha}$ both have full column rank. Thus, for the $m_g \times (m_b + k_\alpha)$ matrix $E = [\bar{\Omega}_n^{1/2}\bar{C}'_n, \bar{\Omega}_n^{-1/2}\bar{G}_{n\alpha}]$,

$$E(E'E)^{-1}E' = \bar{\Omega}_n^{1/2}\bar{C}'_n\bar{\Omega}_{nd}^{-1}\bar{C}_n\bar{\Omega}_n^{1/2} + \bar{\Omega}_n^{-1/2}\bar{G}_{n\alpha}(\bar{G}'_{n\alpha}\bar{\Omega}_n^{-1}\bar{G}_{n\alpha})^{-1}\bar{G}'_{n\alpha}\bar{\Omega}_n^{-1/2},$$

by Exercise (3.17) on pp. 71–72 of Ruud (2000). Therefore,

$$I_{m_g} - \bar{\Omega}_n^{-1/2}\bar{G}_{n\alpha}(\bar{G}'_{n\alpha}\bar{\Omega}_n^{-1}\bar{G}_{n\alpha})^{-1}\bar{G}'_{n\alpha}\bar{\Omega}_n^{-1/2} - \bar{\Omega}_n^{1/2}\bar{C}'_n\bar{\Omega}_{nd}^{-1}\bar{C}_n\bar{\Omega}_n^{1/2} = I_{m_g} - E(E'E)^{-1}E'$$

is nonnegative definite, but will be positive definite if $m_b + k_\alpha < m_g$. Thus, $\hat{\beta}_{\text{TGEL2}}$ is generally less efficient relative to $\hat{\beta}_{\text{GEL}}$.

If $m_a = k_\alpha$, then $E(E'E)^{-1}E' = I_{m_g}$. Thus,

$$I_{m_g} - \bar{\Omega}_n^{-1/2}\bar{G}_{n\alpha}(\bar{G}'_{n\alpha}\bar{\Omega}_n^{-1}\bar{G}_{n\alpha})^{-1}\bar{G}'_{n\alpha}\bar{\Omega}_n^{-1/2} = \bar{\Omega}_n^{1/2}\bar{C}'_n\bar{\Omega}_{nd}^{-1}\bar{C}_n\bar{\Omega}_n^{1/2},$$

and $\hat{\beta}_{\text{TGEL}}$ has the same asymptotic variance as that of $\hat{\beta}_{\text{GEL}}$.

(iv) For $\hat{\alpha}_{\text{TGEL}}$, the consistency can be similarly proved to that for $\hat{\beta}_{\text{TGEL}}$. Then an equation for $\hat{\alpha}_{\text{TGEL}}$ as that for $\hat{\beta}_{\text{TGEL}}$ in (B.2) is $\sqrt{n}(\hat{\alpha}_{\text{TGEL}} - \alpha_0) = -(\bar{G}'_{n\alpha}\bar{\Omega}_n^{-1}\bar{G}_{n\alpha})^{-1}\bar{G}'_{n\alpha}\bar{\Omega}_n^{-1}\sqrt{n}g_n(\alpha_0, \hat{\beta}_{\text{TGEL}}) + o_p(1)$. By the MVT,

$$\begin{aligned}\sqrt{n}(\hat{\alpha}_{\text{TGEL}} - \alpha_0) &= -(\bar{G}'_{n\alpha}\bar{\Omega}_n^{-1}\bar{G}_{n\alpha})^{-1}\bar{G}'_{n\alpha}\bar{\Omega}_n^{-1}[\sqrt{n}g_n(\alpha_0, \beta_0) + \bar{G}_{n\beta}\sqrt{n}(\hat{\beta}_{\text{TGEL}} - \beta_0)] + o_p(1)\\ &= -(\bar{G}'_{n\alpha}\bar{\Omega}_n^{-1}\bar{G}_{n\alpha})^{-1}\bar{G}'_{n\alpha}\bar{\Omega}_n^{-1}[\sqrt{n}g_n(\alpha_0, \beta_0) + \bar{G}_{n\beta}\sqrt{n}(\hat{\beta}_{\text{GEL}} - \beta_0)] + o_p(1),\end{aligned}$$

where the second equality follows because $\sqrt{n}(\hat{\beta}_{\text{TGEL}} - \beta_0) = \sqrt{n}(\hat{\beta}_{\text{GEL}} - \beta_0) + o_p(1)$ when $m_a = k_\alpha$. This equation is the same as that we can obtain from the first order condition for $\hat{\alpha}_{\text{GEL}}$. Hence, $\sqrt{n}(\hat{\alpha}_{\text{TGEL}} - \alpha_0) = \sqrt{n}(\hat{\alpha}_{\text{GEL}} - \alpha_0) + o_p(1)$. ■

Proof of Corollary 1. As $\bar{C}_n = [I_{m_b}, -\bar{C}_{1n}]$, let $\psi_{n\check{C}} = [0_{m_b\times m_b}, -\psi_{nC_1}]$. With $\mathbf{E}\,\frac{\partial g_{na}(\alpha_0)}{\partial\beta'} = 0$, $\psi_{n\check{C}}\bar{G}_{n\beta} = [0_{m_b\times m_b}, -\psi_{nC_1}]\begin{pmatrix}\mathbf{E}\,\frac{\partial g_{nb}(\gamma_0)}{\partial\beta'}\\ 0\end{pmatrix} = 0$. Thus, $B_{nd}^{C-G} = 0$. By a first order Taylor expansion, $0 = g_{na}(\check{\alpha}) = g_{na}(\alpha_0) + \bar{G}_{na\alpha}(\check{\alpha} - \alpha_0) + O_p(n^{-1}) = g_{na}(\alpha_0) + n^{-1/2}\bar{G}_{na\alpha}\psi_{n\check{\alpha}} + O_p(n^{-1})$, where $\bar{G}_{na\alpha} = \mathbf{E}\,\frac{\partial g_{na}(\alpha_0)}{\partial\alpha'}$. Thus, $-\frac{1}{n}\bar{H}_{nd}\psi_{n\check{C}}(\sqrt{n}g_n + \bar{G}_{n\alpha}\psi_{n\check{\alpha}}) = -\frac{1}{n}\bar{H}_{nd}\psi_{nC_1}[\sqrt{n}g_{na}(\alpha_0) + \bar{G}_{na\alpha}\psi_{n\check{\alpha}}] = O_p(n^{-3/2})$. Hence the higher bias of order $O(n^{-1})$ for $\hat{\beta}_{\text{TGEL}}$ does not contain B_{nd}^{C-g}. ■

Proof of Theorem 3. (i) To derive the asymptotic distribution of $\mathcal{R}_{\text{TGEL}}$, we use the results

$$\sqrt{n}(\hat{\beta}_{\text{rTGEL}} - \beta_0) = -[\bar{\Sigma}_{nd} - \bar{\Sigma}_{nd}R'(R\bar{\Sigma}_{nd}R')^{-1}R\bar{\Sigma}_{nd}]\bar{D}'_{n\beta}\bar{\Omega}_{nd}^{-1}\sqrt{n}d_n(\check{\alpha}, \beta_0) + o_p(1). \tag{B.6}$$

and

$$\sqrt{n}\hat{\mu}_{\text{rTGEL}} = [-\bar{\Omega}_{nd}^{-1} + \bar{\Omega}_{nd}^{-1}\bar{D}_{n\beta}\bar{\Sigma}_{nd}\bar{D}'_{n\beta}\bar{\Omega}_{nd}^{-1} - \bar{\Omega}_{nd}^{-1}\bar{D}_{n\beta}\bar{\Sigma}_{nd}R'(R\bar{\Sigma}_{nd}R')^{-1}R\bar{\Sigma}_{nd}\bar{D}'_{n\beta}\bar{\Omega}_{nd}^{-1}]\sqrt{n}d_n(\check{\alpha}, \beta_0) + o_p(1), \tag{B.7}$$

which are derived in Section I.5 of the supplementary material. With (B.7), as in (B.4), we have

$$\begin{aligned}&2n\Big[\frac{1}{n}\sum_{i=1}^{n}\rho\Big(\hat{\mu}'_{\text{RTGEL}}d_{ni}(\check{\alpha},\hat{\beta}_{\text{RTGEL}})\Big)-\rho(0)\Big]\\ &=\sqrt{n}d_n'(\gamma_0)[\bar{\Omega}_{nd}^{-1}-\bar{\Omega}_{nd}^{-1}\bar{D}_{n\beta}\bar{\Sigma}_{nd}\bar{D}'_{n\beta}\bar{\Omega}_{nd}^{-1}+\bar{\Omega}_{nd}^{-1}\bar{D}_{n\beta}\bar{\Sigma}_{nd}R'(R\bar{\Sigma}_{nd}R')^{-1}R\bar{\Sigma}_{nd}\bar{D}'_{n\beta}\bar{\Omega}_{nd}^{-1}]\sqrt{n}d_n(\gamma_0)+o_p(1).\end{aligned} \tag{B.8}$$

By (B.4) and (B.8),

$$\mathcal{R}_{\text{TGEL}}=\sqrt{n}d_n'(\alpha_0,\beta_0)\bar{\Omega}_{nd}^{-1}\bar{D}_{n\beta}\bar{\Sigma}_{nd}R'(R\bar{\Sigma}_{nd}R')^{-1}R\bar{\Sigma}_{nd}\bar{D}'_{n\beta}\bar{\Omega}_{nd}^{-1}\sqrt{n}d_n(\alpha_0,\beta_0)+o_p(1).$$

By the MVT,

$$\begin{aligned}\sqrt{n}g_n(\alpha_0,\beta_0)&=\sqrt{n}g_n(\alpha_0,\beta_n)+G_{n\beta}(\alpha_0,\ddot{\beta}_n)\sqrt{n}(\beta_0-\beta_n)\\ &=\sqrt{n}g_n(\alpha_0,\beta_n)-\bar{G}_{n\beta}c+o_p(1)\xrightarrow{d}N\Big(-\lim_{n\to\infty}\bar{G}_{n\beta}c,\lim_{n\to\infty}\bar{\Omega}_n\Big),\end{aligned}$$

where $\ddot{\beta}_n$ lies between β_0 and β_n. Hence, $\mathcal{R}_{\text{TGEL}}\xrightarrow{d}\chi^2(k_r,\lim_{n\to\infty}c'R'(R\bar{\Sigma}_{nd}R')^{-1}Rc)$.

For $\mathcal{W}_{\text{TGEL}}$, by the MVT, $\sqrt{n}r(\hat{\beta}_{\text{TGEL}})=R(\ddot{\beta})\sqrt{n}(\hat{\beta}_{\text{TGEL}}-\beta_0)=-R\bar{\Sigma}_{nd}\bar{D}'_{n\beta}\bar{\Omega}_{nd}^{-1}\sqrt{n}d_n(\gamma_0)+o_p(1)$, where the second equality follows by (B.3). It follows that $\mathcal{W}_{\text{TGEL}}=\mathcal{R}_{\text{TGEL}}+o_p(1)$.

For $\mathcal{S}_{\text{TGEL}}$,

$$\begin{aligned}\frac{1}{\sqrt{n}}\frac{\partial}{\partial\beta}\boldsymbol{\rho}_{nd}(\check{\alpha},\hat{\beta}_{\text{RTGEL}},\hat{\mu}_{\text{RTGEL}})&=\frac{1}{n}\sum_{i=1}^{n}\rho_1\Big(\hat{\mu}'_{\text{RTGEL}}d_{ni}(\check{\alpha},\hat{\beta}_{\text{RTGEL}})\Big)\frac{\partial d'_{ni}(\check{\alpha},\hat{\beta}_{\text{RTGEL}})}{\partial\beta}\sqrt{n}\hat{\mu}_{\text{RTGEL}}\\ &=-\bar{D}'_{n\beta}\sqrt{n}\hat{\mu}_{\text{RTGEL}}+o_p(1)\\ &=R'(R\bar{\Sigma}_{nd}R')^{-1}R\bar{\Sigma}_{nd}\bar{D}'_{n\beta}\bar{\Omega}_{nd}^{-1}\sqrt{n}d_n(\check{\alpha},\beta_0)+o_p(1).\end{aligned}$$

It follows that

$$\mathcal{S}_{\text{TGEL}}=\sqrt{n}d_n'(\alpha_0,\beta_0)\bar{\Omega}_{nd}^{-1}\bar{D}_{n\beta}\bar{\Sigma}_{nd}R'(R\bar{\Sigma}_{nd}R')^{-1}R\bar{\Sigma}_{nd}\bar{D}'_{n\beta}\bar{\Omega}_{nd}^{-1}\sqrt{n}d_n(\alpha_0,\beta_0)+o_p(1)=\mathcal{R}_{\text{TGEL}}+o_p(1).$$

For $\mathcal{G}_{\text{TGEL}}$, as in Lemma H.2, $\check{\lambda}_r=\arg\max_{\lambda\in\Lambda_{n\Psi}(\check{\alpha},\check{\beta}_r)}\sum_{i=1}^{n}\rho\Big(\lambda'\Psi_{ni}(\check{\alpha},\check{\beta}_r)\Big)$ exists w.p.a.1, and the first order condition $\sum_{i=1}^{n}\rho_1\Big(\check{\lambda}_r'\Psi_{ni}(\check{\alpha},\check{\beta}_r)\Big)\Psi_{ni}(\check{\alpha},\check{\beta}_r)=0$ holds. Applying the MVT to this first order condition at $\lambda=0$, we have

$$0=\sum_{i=1}^{n}\rho_1(0)\Psi_{ni}(\check{\alpha},\check{\beta}_r)+\sum_{i=1}^{n}\rho_2\Big(\ddot{\lambda}'\Psi_{ni}(\check{\alpha},\check{\beta}_r)\Big)\Psi_{ni}(\check{\alpha},\check{\beta}_r)\Psi'_{ni}(\check{\alpha},\check{\beta}_r)\check{\lambda}_r,$$

where $\ddot{\lambda}$ lies between 0 and $\check{\lambda}_r$. Then,

$$\begin{aligned}\sqrt{n}\check{\lambda}_r &= \left[\frac{1}{n}\sum_{i=1}^{n}\rho_2\Big(\ddot{\lambda}'\Psi_{ni}(\check{\alpha},\check{\beta}_r)\Big)\Psi_{ni}(\check{\alpha},\check{\beta}_r)\Psi_{ni}'(\check{\alpha},\check{\beta}_r)\right]^{-1}\frac{1}{\sqrt{n}}\sum_{i=1}^{n}\Psi_{ni}(\check{\alpha},\check{\beta}_r)\\ &= -(R\bar{\Sigma}_{nd}R')^{-1}R\bar{\Sigma}_{nd}\bar{D}_{n\beta}'\bar{\Omega}_{nd}^{-1}\sqrt{n}d_n(\gamma_0)+o_p(1).\end{aligned}$$

Hence, by an expansion as that in (B.4),

$$\mathcal{G}_{\text{TGEL}} = \sqrt{n}\check{\lambda}_r'\frac{1}{n}\sum_{i=1}^{n}\Psi_{ni}(\check{\alpha},\check{\beta}_r)\Psi_{ni}'(\check{\alpha},\check{\beta}_r)\sqrt{n}\check{\lambda}_r + o_p(1) = \mathcal{R}_{\text{TGEL}} + o_p(1). \tag{B.9}$$

(ii) Note that $\bar{G}_n = [\bar{G}_{n\alpha}, \bar{G}_{n\beta}]$, then by the block matrix inverse formula, $R_\gamma\bar{\Sigma}_nR_\gamma' = R[\bar{G}_{n\beta}'\bar{\Omega}_n^{-1}\bar{G}_{n\beta} - \bar{G}_{n\beta}'\bar{\Omega}_n^{-1}\bar{G}_{n\alpha}(\bar{G}_{n\alpha}'\bar{\Omega}_n^{-1}\bar{G}_{n\alpha})^{-1}\bar{G}_{n\alpha}'\bar{\Omega}_n^{-1}\bar{G}_{n\beta}]^{-1}R' \leq R[\bar{G}_{n\beta}'\bar{C}_n'(\bar{C}_n\bar{\Omega}_n\bar{C}_n')^{-1}\bar{C}_n\bar{G}_{n\beta}]^{-1}R' = R\bar{\Sigma}_{nd}R'$, where the inequality has used $I_{k_g} - \bar{\Omega}_n^{-1/2}\bar{G}_{n\alpha}(\bar{G}_{n\alpha}'\bar{\Omega}_n^{-1}\bar{G}_{n\alpha})^{-1}\bar{G}_{n\alpha}'\bar{\Omega}_n^{-1/2} \geq \bar{\Omega}_n^{1/2}\bar{C}_n'(\bar{C}_n\bar{\Omega}_n\bar{C}_n')^{-1}\bar{C}_n\bar{\Omega}_n^{1/2}$ when $m_a \geq k_\alpha$, which is shown in the proof of Theorem 1. Thus, $c'R'(R_\gamma\bar{\Sigma}_nR_\gamma')^{-1}Rc \leq c'R'(R\bar{\Sigma}_{nd}R')^{-1}Rc$.

(iii) When $m_a = k_\alpha$, $I_{k_g} - \bar{\Omega}_n^{-1/2}\bar{G}_{n\alpha}(\bar{G}_{n\alpha}'\bar{\Omega}_n^{-1}\bar{G}_{n\alpha})^{-1}\bar{G}_{n\alpha}'\bar{\Omega}_n^{-1/2} = \bar{\Omega}_n^{1/2}\bar{C}_n'(\bar{C}_n\bar{\Omega}_n\bar{C}_n')^{-1}\bar{C}_n\bar{\Omega}_n^{1/2}$ and $c'R'(R_\gamma\bar{\Sigma}_nR_\gamma')^{-1}Rc = c'R'(R\bar{\Sigma}_{nd}R')^{-1}Rc$. ■