This is to certify that the
dissertation entitled

Three Essays on Generalized Method of Moments

presented by

Artem B. Prokhorov

has been accepted towards fulfillment
of the requirements for the

Ph.D. degree in **Economics**

Peter Schmidt

Major Professor's Signature

4/28/06

Date

THREE ESSAYS ON GENERALIZED METHOD OF MOMENTS

By

Artem B. Prokhorov

A DISSERTATION

Submitted to
Michigan State University
in partial fulfillment of the requirements
for the degree of

DOCTOR OF PHILOSOPHY

Department of Economics

2006

ABSTRACT

THREE ESSAYS ON GENERALIZED METHOD OF MOMENTS

BY

ARTEM B. PROKHOROV

Generalized Method of Moments (GMM) is a powerful estimation method based on orthogonality conditions known to hold in the population of interest. GMM is sufficiently general to incorporate most of the extremum and minimum distance estimators in econometrics including (Q)MLE, M-estimator, weighted and nonlinear LS. By taking advantage of GMM's universality, my thesis seeks to contribute to three areas of (micro)econometric research: modelling processes with missing observations (e.g., attrition and self-selection in panel data sets, counterfactual outcomes for treatment and control groups), modelling likelihood using copulas (e.g., PROBIT, LOGIT, selectivity models), and modelling covariance structures (e.g., LISREL, fixed effects, factor analysis).

The first essay, "GMM Redundancy Results for General Missing Data Problem," considers alternative GMM estimators of a parameter vector that enters into one set of moment equations along with another vector that also enters into an additional set of moment conditions and may be known. Alternative estimators are ranked in terms of relative efficiency, and conditions for no efficiency gains are derived. The results are applied to a general missing data problem. Conditions for the counterintuitive result of the missing data literature that estimating selection probabilities is better than knowing them arise naturally in the general problem. Efficiency gains from using both weighted and unweighted moment equations under exogenous sampling are considered.

The second essay, "Robustness, Redundancy, and Validity of Copulas in Like-

lihood Models," considers likelihood-based estimation of multivariate models, in which only marginal distributions are correctly specified. The unknown joint distribution is modelled with a copula function, which may be misspecified. In a GMM framework, we study robustness and efficiency of resulting estimators, propose improvements to existing estimators and discuss tests of copula validity. It is shown that radially symmetric copulas are robust against misspecification in problems about sample means if the true joint density is also radially symmetric. Efficiency results suggest that knowledge of the true copula is redundant if and only if the covariance matrix for relevant moment conditions is singular. A simple simulation supports the theoretical result about robustness of the Frank, Farlie-Gumbel-Morgenstern and Ali-Mikhail-Haq copula families.

The third essay, "Modelling Covariance Structures: First and Second Order Asymptotics," considers estimation of covariance structure models by quasi maximum likelihood (QMLE), generalized method of moments (GMM) and empirical likelihood (EL). A general condition is derived under which the GMM (and EL) estimators do not dominate normal QMLE in terms of first-order efficiency. The condition is formulated in terms of the fourth order moments of the true distribution. The second-order asymptotic bias of QMLE is derived and a formal proof is presented of the intuitive result that, under normality, this bias is the same as that of EL.

To *the memory of my father*

ACKNOWLEDGEMENTS

This dissertation raises more questions than it answers. But the questions it raises and answers would not have been answered (perhaps not even raised) if it had not been for the interaction with my *Doctorvater*, University Distinguished Professor Peter Schmidt. Professor Schmidt's wisdom and creativity, his ability to put things in perspective and prioritize, his talent of succinct statements and lively examples, his patience, approachability and close interaction with his students, his generous support of their endeavors and his understanding of their concerns – all this makes him a great mentor and this dissertation worth reading. I have found immensely enriching both my TA-ing for Professor Schmidt and attending the fabulous TA appreciation dinners Professors Peter Schmidt and Christine Amsler sponsor each semester for their TAs.

It is impossible to overestimate the support I have received from the other members of my dissertation committee. Professor Jeffrey Wooldridge has been prompt and thorough in reading drafted parts of the dissertation as they appeared and providing feedback. He also generously supported my travel to Australia to present the results of the first essay. Professor Richard Baillie gave me valuable insights into the world of financial econometrics during my RA-ship for him. Professor Hira Koul has helped add statistical rigor to the dissertation and to my graduate training in econometrics.

MSU graduate travel grants enabled me to attend the following meetings, where parts of this dissertation were presented: the 2006 AEA/ASSA meetings in Boston, the 33rd Annual Australian Conference of Economists in Sydney (October 2004), the 5th Villa Mondragone Workshop on Economic Theory and Econometrics in Rome (July 2005) and the 2004 Empirical Research Summer School on Experimental Economics and Econometrics in Mannheim.

The final stages of the research were supported by a Dissertation Completion Fellowship from the Graduate School of MSU.

Emma Iglesias of MSU, Ivana Komunjer of UCSD and Rustam Ibragimov of Harvard provided helpful discussions of some of the results. So did the participants of the above mentioned conferences and of the econometrics seminars at Michigan State, Bates White LLC, Concordia, Florida State, New South Wales, Massey, Emory University and Central Michigan University.

Finally and most importantly, Irina Agafonova is the person who made this all worthwhile.

I am very grateful to these people.

Table of Contents

List of Tables

List of Figures

Essay 1

GMM Redundancy Results for General Missing Data Problem

1.1 Introduction

There are many models that can be formulated as two sets of moment conditions with two parameter vectors one of which enters in only one of these sets and the other in both. For example, Newey (1984) shows that multi-step estimators that employ estimates of an additional parameter vector in estimation of the primary parameter vector of interest can be represented in such a generalized method of moment (GMM) framework with exact identification of the parameters. *Generated regressors* models of Pagan (1984), *latent variable* models of Zellner (1970) and Goldberger (1972) and many others are two-step cases of this formulation. However, the primary focus of and the motivation for this essay are the missing data

(or selectivity) models.

Selectivity models deal with samples in which some observations are omitted (we call such samples "selected"). The missing data problem arises when using selected samples in an estimation procedure results in a biased estimator. For example, if we were to conduct a survey of young mothers to study the effect of mother's smoking on the weight of the newborn, the survey would typically have missing data due to non-response. It is likely that non-response is associated with heavy smoking and poor birth weight. If the missing data were ignored the effect of smoking would be underestimated. In such cases it is common to construct a probabilistic model for the missing data generating process (we call this model a "selection model") and then to appropriately adjust the primary model of interest for the effect of selection into the sample.

This paper is motivated by a puzzle in the selectivity literature. Consider the setting of a GMM problem is which we have a set of moment conditions, with some parameters θ_1 (the "parameters of interest"), and these moment conditions hold in the unselected sample. However, we also have a selection mechanism such that the moment conditions do not hold in the selected sample. Under certain assumptions given below (typically referred to as "ignorability" or "selection on observables"), weighting the original moment conditions by the inverse of the probability of selection yields a modified set of moment conditions that do hold in the selected sample. We will follow Wooldridge (2002b, 2005) in calling the estimator based on these weighted moment conditions the "inverse probability weighting" (IPW) estimator.

Unless the probability of selection is known for each selected observation, imple-

mentation of the IPW estimator will require a model that permits the estimation of the probability of selection. Let θ_2 be the parameters (the "selection parameters") in the moment conditions derived from this model. Typically these moment conditions will be based on the score function from the likelihood function for the selection process. A two-step IPW procedure can be considered, in which the first step is the estimation of θ_2 from the selection model, and the second step is the estimation of θ_1 by GMM on the weighted moment conditions, where the weighting is done using the estimated probabilities of selection.

In this setting, the puzzle is that it is better to estimate the selection probabilities than to use the true selection probabilities, even if the latter are known. In other words, in terms of the augmented model described above, we get a better estimator of θ_1 when we use the estimated θ_2 in the second step than if we used the true θ_2. This phenomenon has been discussed by Wooldridge (1999, 2001, 2002b, 2005), and it has also been noted in a number of previous works, including Rosenbaum (1987); Imbens (1992); Robins and Rotnitzky (1995); Crepon et al. (1997), and Hirano et al. (2003). This is puzzling because knowledge of θ_2, if properly exploited, cannot be harmful.

To resolve this puzzle, we follow Newey and McFadden (1994) in setting up an augmented set of moment conditions, where the first subset are the weighted original moment conditions, which now contain both θ_1 and θ_2, and the second subset are the moment conditions from the selection model, which contain only θ_2. We show that the second set of moment conditions is useful (non-redundant), even when θ_2 is known. This is true because the second set of moment conditions is correlated with the first set in the selected sample (even though it is not in the full sample). So the inefficiency of the estimator based on known θ_2 and the first set of moment

conditions only is due to its failure to exploit the information in the second set of moment conditions; whereas, when θ_2 is not known, there is no choice but to include the second set of moment conditions.

This raises the question of whether, when θ_2 is known, we can improve on the two-step estimator (which uses estimated θ_2 in the second step) by using a GMM estimator based on both sets of moment conditions, but where only θ_1 is estimated. After all, this GMM estimator cannot be worse than the two-step estimator of θ_1. The answer to this question is a bit complicated. In the case that the original GMM problem (the one that contains the parameter of interest) is overidentified, the two-step estimator is dominated by a one-step estimator that estimates θ_1 and θ_2 jointly in the augmented GMM model. However, we show that, in the augmented GMM model, knowledge of θ_2 is redundant (does not improve the precision of estimation of θ_1). So, while it can never hurt to know more, if that knowledge is used properly, in this case it does not help either.

The result just quoted is given in Section 1.3 of the paper. In Section 1.2, we set the stage by giving a number of results on efficiency and redundancy of estimation in a general GMM setting, when one set of moment conditions depends on θ_1 and θ_2, while a second set of moment conditions depends only on θ_2. Some of these results are original and interesting in their own right. We consider "m-redundancy", which is redundancy of moment conditions in the sense of Breusch et al. (1999), and we also consider "p-redundancy", which is redundancy of the knowledge of some of the parameters for estimation of the other parameters. One of our results gives an interesting connection between these two concepts: the first set of moment conditions with θ_1 known is m-redundant for estimation of θ_2 if and only if knowledge of θ_2 is p-redundant for estimation of θ_1. This is in fact the key

result in establishing our subsequent results for the selectivity model.

In Section 1.3 we also consider the selectivity model under a stronger "exogeneity of selection" assumption under which both the unweighted moment conditions and the weighted moment conditions hold in the selected population. Wooldridge (2001) has shown that in this circumstance it is better to use the unweighted moment conditions than the weighted moment conditions. However, this does not rule out the possibility that it would be better to use both. We show that in this circumstance the weighted moment conditions are m-redundant for estimation of θ_1, so that using both sets is no better than using just the unweighted moment conditions. Thus when we do not have to weight for reasons of consistency, we also do not have to weight for reasons of efficiency.

GMM is sufficiently general to accommodate most of the extremum and minimum distance estimators in econometrics (see, e.g., Newey and McFadden, 1994, p.2118). The arguments we present can be applied, for example, to (Q)MLE, M-estimation, WLS, and NLS. They also extend to the asymptotic equivalents of GMM such as empirical likelihood and exponential tilting estimators. Our results relate to the treatment effect estimation literature (e.g., Rosenbaum and Rubin, 1983; Hirano et al., 2003; Heckman et al., 1998; Hahn, 1998), to stratified-sampling literature (e.g., Manski and Lerman, 1977; Manski and McFadden, 1981; Cosslett, 1981a,b; Imbens, 1992; Tripathi, 2003) and other similarly-structured problems (e.g., Hellerstein and Imbens, 1999; Nevo, 2002, 2003; Imbens, 1992; Crepon et al., 1997).

1.2 Efficiency and redundancy results for the general estimation problem

1.2.1 Preliminaries

Consider a family of distributions $\{P_\theta, \theta \in \Theta = \Theta_1 \times \Theta_2 \subset \mathbb{R}^{p_1} \times \mathbb{R}^{p_2}, \Theta \text{ compact}\}$, a random vector $W^* \in \mathcal{W}^* \subset \mathbb{R}^{\dim(W^*)}$ from $P_{\theta_o}, \theta_o \in \Theta$, and a real valued, measurable function $h : \mathcal{W}^* \times \Theta \to \mathbb{R}^m$ such that

$$\mathbb{E}_{\theta_o}[h(W^*, \theta)] = 0, \quad \text{if and only if } \theta = \theta_o. \tag{1.1}$$

The expectation is with respect to the distribution of W^* indexed by θ_o. In the sequel we suppress the subscript.

Let $||\cdot||$ denote the Euclidean norm, $N(\theta, \delta) \subset \Theta$ denote an open $p_1 + p_2$-ball of radius δ with center at θ, $\nabla_\theta h(\cdot, \theta)$ denote the $m \times (p_1 + p_2)$ Jacobian of $h(\cdot, \theta)$ with respect to θ, and "w.p.1" stand for "with probability one".

Assumption 1.2.1 *Assume that the moment function in (1.1) satisfies the following conditions:*

(i) $\theta_o \in int(\Theta)$;

(ii) $h(W^*, \theta)$ *is continuous at each* $\theta \in \Theta$, *w.p.1;*

(iii) $h(W^*, \theta)$ *is (once) continuously differentiable on* $N(\theta_o, \delta)$, *for some* $\delta > 0$, *w.p.1;*

(iv) $\mathbb{E}\{\sup_{\theta\in\Theta} ||h(W^*,\theta)||^2\} < \infty$;

(v) $\mathbb{E}\{\sup_{\theta\in N(\theta_o,\delta)} ||\nabla_\theta h(W^*,\theta)||\} < \infty$, *for some* $\delta > 0$;

(vi) $\mathbb{E}[\nabla_\theta h(W^*,\theta)|_{\theta=\theta_o}]$ *is of full column rank.*

For simplicity, we assume here that W_i^*, $i = 1,\ldots,N$, are i.i.d. draws from P_{θ_o}.

The generalized method of moments (GMM) estimator of θ_o is the solution to the following minimization problem

$$\min_{\theta\in\Theta} \bar{h}(\theta)'\mathbb{W}\bar{h}(\theta), \tag{1.2}$$

where

$$\bar{h}(\theta) = \frac{1}{N}\sum_{i=1}^{N} h(W_i^*,\theta)$$

is the sample analogue of the population moment condition which is zero at θ_o, and $\mathbb{W}$ is a positive semi-definite weighting matrix (see, e.g., Hansen, 1982). In the GMM framework, the choice of the weighting matrix may depend on θ_o. In such cases, a preliminary consistent estimate of θ_o is used to construct an estimate of $\mathbb{W}$ used in the above definition of the GMM estimator. We will comment on this point again later.

Theorem 1.2.1 *(see, e.g., Newey and McFadden, 1994, Theorems 2.6 and 3.4) Under Assumption 1.2.1, the GMM estimator of* θ_o *is consistent and asymptotically normal (CAN).*

Proofs: See the Appendix for proofs of all theorems and corollaries. □

1.2.2 The general estimation problem

Let $\theta = (\theta_1', \theta_2')'$ and

$$h(W^*; \theta) = \begin{pmatrix} h_1(W^*; \theta_1, \theta_2) \\ h_2(W^*; \theta_2) \end{pmatrix},$$

where $\theta_1 \in \Theta_1$, $\theta_2 \in \Theta_2$, and $h_1(\cdot)$ and $h_2(\cdot)$ are m_1- and m_2-vectors of known functions ($m = m_1 + m_2$). Then if we suppress W^* we can write (1.1) as

$$\begin{aligned} &\text{(A)} \quad \mathbb{E}[h_1(\theta_{o1}, \theta_{o2})] = 0, \\ &\text{(B)} \quad \mathbb{E}[h_2(\theta_{o2})] = 0. \end{aligned} \tag{1.3}$$

We consider the general case of overidentification, i.e., $m_1 \geq p_1$ and $m_2 \geq p_2$.

The optimal weighting matrix for GMM will be the inverse of the following covariance matrix or its components:

$$C = \mathbb{V}[h(\theta_o)] = \begin{bmatrix} C_{11} & C_{12} \\ C_{21} & C_{22} \end{bmatrix}, \tag{1.4}$$

where variance is with respect to P_{θ_o} as before. Note that C is a function of θ_o and is generally unknown. In defining alternative GMM estimators and deriving their asymptotic variance matrices, we will behave as if we knew θ_o and thus knew C. In practice if we wish to use C in the weighting matrix of the GMM estimator we would typically first obtain an estimate of C based on a preliminary consistent estimate of θ_o. Such a preliminary estimate of θ_o can be the GMM estimator that uses the identity matrix for weighting.

We assume that C is finite and nonsingular so its inverse exists. Let

$$C^{-1} = \begin{bmatrix} C^{11} & C^{12} \\ C^{21} & C^{22} \end{bmatrix}.$$

Define the $(m_1 + m_2) \times (p_1 + p_2)$ matrix of expected derivatives

$$D = \mathbb{E}\frac{\partial h(\theta)}{\partial \theta'}\Big|_{\theta=\theta_o} = \begin{bmatrix} D_{11} & D_{12} \\ 0 & D_{22} \end{bmatrix}. \tag{1.5}$$

We assume that D_{11} and D_{22} are of full column rank so that h_2 alone identifies θ_2 and h_1 identifies θ_1 given θ_2.

Similar to C, D depends on θ_o. In deriving the GMM asymptotic variance matrices, we will treat D as know. Consistent estimates of D (and C) can be obtained using consistent estimates of θ_o in practice.

We now define four different GMM estimators that differ in which moment conditions are used and/or whether θ_2 is treated as known. For each of these estimators we treat C as known. We will comment on this point once again in the next subsection.

Definition 1.2.1 *Call the estimator of θ that minimizes (1.2) with the optimal weighting matrix $\mathbb{W} = C^{-1}$ the* ONE-STEP *estimator.*

This is the usual GMM estimator that uses both moment conditions (1.3A) and (1.3B) jointly to estimate θ_{o1} and θ_{o2}.

Definition 1.2.2 *Call the estimator of θ obtained in the following two step procedure the* TWO-STEP *estimator: (i) the estimator $\hat{\theta}_2$ is obtained by minimizing (1.2), where $h(\theta) = h_2(\theta_2)$ and $\mathbb{W} = C_{22}^{-1}$; (ii) the estimator of θ_1 is obtained by minimizing (1.2), where $h(\theta) = h_1(\theta_1, \theta_2)$, $\mathbb{W} = C_{11}^{-1}$, and $\theta_2 = \hat{\theta}_2$ is treated as known.*

This estimator uses the orthogonality condition (1.3B) first to obtain a consistent estimator of the unknown parameter subvector θ_{o2} and then uses the moment condition (1.3A) to obtain the estimator of θ_{o1}. Estimators considered in Wooldridge (2003), Newey (1984), Newey and McFadden (1994, pp. 2176-2184) and many others are TWO-STEP estimators with $m_1 = p_1$, $m_2 = p_2$.

Definition 1.2.3 *Call the estimator of θ_1 obtained by minimizing (1.2), where $h(\theta) = h_1(\theta_1, \theta_2)$, $\mathbb{W} = C_{11}^{-1}$, and θ_2 is treated as known, the* KNOW-θ_2 *estimator.*

Here, equation (B) in (1.3) is ignored. However, the results of Section 1.3 of the paper all derive from understanding that (B) is potentially informative even though θ_{2o} is known because it imposes additional restrictions on the population.

Definition 1.2.4 *Call the estimator of θ_1 obtained by minimizing (1.2), where $h(\cdot)$ contains both $h_1(\cdot)$ and $h_2(\cdot)$, $\mathbb{W} = C^{-1}$, and θ_2 is treated as known the* KNOW-θ_2-JOINT *estimator.*

This is the augmented GMM estimator of θ_{o1} of the form considered in Qian and Schmidt (1999). Here, the information in (1.3B) is kept even though θ_{o2} is assumed known.

Under Assumption 1.2.1 all four estimators are CAN.

Theorem 1.2.2 *Let* $\mathbb{V}_{\text{ONE-STEP}}$, $\mathbb{V}_{\text{TWO-STEP}}$, $\mathbb{V}_{\text{KNOW-}\theta_2}$, *and* $\mathbb{V}_{\text{KNOW-}\theta_2\text{-JOINT}}$ *denote the asymptotic variance of the* ONE-STEP, TWO-STEP, KNOW-θ_2, *and* KNOW-θ_2-JOINT *estimators, respectively. Then,*

$$\mathbb{V}_{\text{ONE-STEP}} = (D'C^{-1}D)^{-1}, \tag{1.6}$$

$$\mathbb{V}_{\text{TWO-STEP}} = BCB', \tag{1.7}$$

$$\mathbb{V}_{\text{KNOW-}\theta_2} = (D_{11}'C_{11}^{-1}D_{11})^{-1}, \tag{1.8}$$

$$\mathbb{V}_{\text{KNOW-}\theta_2\text{-JOINT}} = (D_{11}'C^{11}D_{11})^{-1}, \tag{1.9}$$

where B *is defined in equation (1.31) of the Appendix.*

In the above expressions, we use the standard notation that "the asymptotic variance of $\hat{\theta}$ is $\mathbb{V}$" means "$\sqrt{N}(\hat{\theta} - \theta_o)$ converges in distribution to $N(\mathbf{0}, \mathbb{V})$."

1.2.3 Efficiency and redundancy results

We can now state several asymptotic relative efficiency results (noting that a known parameter is always more efficient than its estimator).

Theorem 1.2.3 *For the estimators defined in Definitions 1.2.1-1.2.4 with asymptotic variances given in (1.6)-(1.9), respectively, the following statements hold:*

1. KNOW-θ_2-JOINT *is no less asymptotically efficient than* KNOW-θ_2.

2. KNOW-θ_2-JOINT *is no less asymptotically efficient than* ONE-STEP.

3. ONE-STEP *in no less asymptotically efficient than* TWO-STEP.

4. *If* $C_{12} = 0$ *then* KNOW-θ_2-JOINT *and* KNOW-θ_2 *are equally asymptotically efficient [M-redundancy].*

5. *If* $D_{12} = 0$ *then* TWO-STEP *and* KNOW-θ_2 *are equally asymptotically efficient for* θ_1.

6. *If* $C_{12} = 0$ *and* $D_{12} = 0$ *then* ONE-STEP, TWO-STEP, KNOW-θ_2-JOINT *and* KNOW-θ_2 *are all equally asymptotically efficient for* θ_1, ONE-STEP *and* TWO-STEP *are equally asymptotically efficient for* θ_2, *too [M/P-redundancy].*

7. *If* $m_1 = p_1$ *then* ONE-STEP *of* θ_2 *and* TWO-STEP *of* θ_2 *are equal.*

8. *If* $m_1 = p_1$ *and* $m_2 = p_2$ *then the* ONE-STEP *and* TWO-STEP *estimates are equal (for both* θ_1 *and* θ_2*).*

9. *If* $m_1 = p_1$ *and* $C_{12} = 0$ *then the* ONE-STEP *and* TWO-STEP *estimates are equally efficient (for both* θ_1 *and* θ_2*).*

10. *If* $D_{12} = C_{12}C_{22}^{-1}D_{22}$ *then* KNOW-θ_2-JOINT *and* ONE-STEP *are equally asymptotically efficient for* θ_1 *[P-redundancy].*

11. *If* $D_{12} = C_{12}C_{22}^{-1}D_{22}$ *then* ONE-STEP, TWO-STEP *and* KNOW-θ_2-JOINT *are no less asymptotically efficient for* θ_1 *than* KNOW-θ_2.

As noted above, we have defined our estimators as depending on known C. In practice, C is replaced by an initial consistent estimate. This has no effect on the asymptotic variance of the estimates and so it does not affect our efficiency

comparisons. For Statements 7 and 8, which do not involve asymptotic arguments, we would need to require that the same initial consistent estimate is used.

Statements 1-3 state the obvious fact that KNOW-θ_2-JOINT dominates KNOW-θ_2, ONE-STEP and TWO-STEP. The known value of θ_{o2} is at least as efficient as any estimate of θ_{o2}, and the KNOW-θ_2-JOINT estimate of θ_{o1} is the efficient GMM estimate of θ_{o1} based on the full set of available moment conditions.

Statement 4 is essentially the result of Qian and Schmidt (1999). With θ_{o2} known, the second set of moment conditions contains no unknown parameters, and Qian and Schmidt show that using these conditions in addition to the first set of moment conditions improves efficiency except in the special case that $C_{12} = 0$. We call this type of redundancy the *knowledge-of-moment* redundancy (M-redundancy). Also, if we combine Statements 2-3 and 4, we have the corollary that if $C_{12} = 0$, KNOW-θ_2 is at least as efficient as ONE-STEP and TWO-STEP.

Statement 5 is essentially the result of Newey and McFadden (1994, pp. 2179-2180) for the condition under which first stage estimation of a nuisance parameter (θ_{o2}) does not affect the asymptotic variance of the second stage estimate of the parameter of interest (θ_{o1}). See also Wooldridge (2002a, pp. 353-356). However, our version treats the overidentified case as well.

Statement 6 combines the conditions of Statements 4 and 5. Therefore the equal efficiency of TWO-STEP, KNOW-θ_2 and KNOW-θ_2-JOINT follows from those statements. The fact that ONE-STEP is also equally efficient is an additional result. This statement provides conditions for redundancy of both the knowledge of θ_{o2} and of the extra moment conditions in (B) for estimating θ_{o1} (M/P-redundancy). One

case when the conditions hold is when θ_{o2} does not enter (A) and the two moment conditions are uncorrelated. This statement can also be viewed as a special case of Theorem 7 of Breusch et al. (1999) that deals with partial redundancy of moment conditions.

Statement 7 is the GMM separability result of Ahn and Schmidt (1995) that says that the GMM estimate of θ_2 is unaffected if an equal number of parameters and moment conditions is added, because the additional conditions only determine θ_1 in terms of θ_2. Further, it can be shown (see the Appendix of Ahn and Schmidt, 1995) that if D_{11} is nonsingular (which is true since D_{11} is of full column rank) the ONE-STEP estimator of θ_{o1} is expressed in terms of the ONE-STEP estimator of θ_{o2} using the equation $\bar{h}_1(\hat{\theta}_1, \hat{\theta}_2) = C_{12}C_{22}^{-1}\bar{h}_2(\hat{\theta}_2)$. Thus, ONE-STEP for θ_{o1} is derived from the same equation as TWO-STEP for θ_{o1} as long as $\bar{h}_2(\hat{\theta}_2) = 0$ (which holds under exact identification of θ_2) or C_{12} is zero asymptotically. The former condition implies equivalence of the estimators (Statement 8); the latter implies their equal efficiency asymptotically (Statement 9).

Statements 10 and 11 are novel and interesting. They discuss implications of the condition that $D_{12} = C_{12}C_{22}^{-1}D_{22}$. This is the condition for redundancy of h_1 given h_2, for estimation of θ_{o2} when θ_{o1} is known (see Breusch et al., 1999, p. 94), which is an m-redundancy result. Under this condition, Statement 10 says that KNOW-θ_2-JOINT and ONE-STEP are equally efficient. This means that knowledge of θ_{o2} does not help efficiency of estimation of θ_{o1} (from the set of all moment conditions) under this condition, which is a p-redundancy result. This link between m-redundancy and p-redundancy (the first set of moment conditions with θ_{o1} known is m-redundant for estimation of θ_{o2} if and only if knowledge of θ_{o2} is p-redundant for estimation of θ_{o1}) is quite interesting and (so far as we know)

original.

Under the same condition, Statement 11 says that KNOW-θ_2 is dominated by the other three estimators. This is because knowledge of θ_{o2} is not useful, and the KNOW-θ_2 estimator fails to use the second set of moment conditions, which is useful unless $C_{12} = 0$. Note, however, that although the TWO-STEP estimator dominates the KNOW-θ_2 estimator under this condition, the TWO-STEP estimator is still not as efficient as the ONE-STEP or KNOW-θ_2-JOINT estimators unless $m_1 = p_1$ (the first equation is exactly identified for θ_1, given θ_2).

This condition is also important because it implies that conservative inference can be made using the asymptotic standard errors obtained from exactly identified estimations that neglect the first step (Statement 11).

The condition of Statements 10 and 11 will often hold when $h_2(\theta_2)$ is the score of a log-likelihood function that depends on θ_2 but not θ_1. In this case the estimate of θ_{o2} based on h_2 will be efficient, and another moment condition based on $h_1(\theta_1, \theta_2)$ with θ_{o1} known should be m-redundant. More precisely, the generalized information equality (GIME) implies that the expectation of the derivative of h_1 (with respect to θ_2), evaluated at θ_o, equals minus its covariance with the score so that $D_{12} = -C_{12}$, and the usual information equality implies that $D_{22} = -C_{22}$, so that $D_{12} = C_{12}C_{22}^{-1}D_{22}$ holds. Indeed this is exactly what occurs in the selectivity model of the next section.

Example 1.2.1 A sufficient condition for Statements 6, 10, and 11 to hold is that $h_1(\theta_1, \theta_2) = \nabla_{\theta_1} \ln f(w^*|\theta_1, \theta_2)$ and $h_2(\theta_2) = \nabla_{\theta_2} \ln f(w^*|\theta_1, \theta_2)$, where $f(w^*|\theta_1, \theta_2)$ is the density of W^*. Then, the asymptotic variance matrix of the

estimator of θ_{o2} can be equivalently written as C_{22}^{-1} and as C^{22}. This implies that the information matrix for θ_1 and θ_2 is block diagonal, i.e. $D_{12} = -C_{12} = 0$. Thus by Theorem 1.2.3 we can claim more than Statements 10 and 11 in this case: it does not make any difference for the efficiency of the estimate of θ_1 whether θ_2 is estimated or known, and in fact all four estimators are equally efficient (Statement 6). □

We now apply these results to the missing data problem.

1.3 Application to missing data problem

1.3.1 The population problem

Consider again a random vector $W^* \in \mathcal{W}^*$ from the distribution P_{θ_o}, $\theta_o \in \Theta = \Theta_1 \times \Theta_2 \subset \mathbb{R}^{p_1} \times \mathbb{R}^{p_2}, \Theta$ compact. Let W^* contain random vector $W \in \mathcal{W} \subset \mathbb{R}^{\dim(W)}$. Consider a real valued measurable function $g : \mathcal{W} \times \Theta_1 \to \mathbb{R}^{m_1} (m_1 \geq p_1)$ such that

$$\mathbb{E}[g(W, \theta_1)] = 0, \text{ if and only if } \theta_1 = \theta_{o1}. \tag{1.10}$$

As before, expectation is with respect to P_{θ_o}. Assume that the moment function in (1.10) satisfies Assumption 1.2.1.

We are interested in estimating θ_{o1}. The parameter θ_{o1} usually describes some feature of the distribution of W such as the conditional mean, the conditional

variance, the conditional quantiles, etc. The vector W is often partitioned into $(X, Y) \in \mathcal{X} \times \mathcal{Y}$ and $\mathbb{E}(Y|x)$ is often the feature of interest (see Example 1.3.1).

Example 1.3.1 Consider the M-estimation of the parameter θ_{o1} in a general nonlinear least squares model for $\mathbb{E}(Y|x) = m(x, \theta_{o1})$. This is one of the examples considered in Wooldridge (2003). We assume that the model is correctly specified. Let the identifying moment functions be the first order conditions for optimization of $q(x, y; \theta_1) = (y - m(x, \theta_1))^2$. Then, $W = (X, Y)$, $m_1 = p_1$, and $g(W, \theta_1) = -(Y - m(X, \theta_1))[\nabla_{\theta_1} m(X, \theta_1)]'$. Note that a stronger condition than (1.10) holds in this case, namely $\mathbb{E}[g(W, \theta_{o1})|x] = 0$. □

Example 1.3.2 Consider the maximum likelihood estimation of a LOGIT model where Y is a binary outcome variable and X is a vector of regressors and the conditional probability $p(y|x, \theta_{o1})$ is modelled as $\mathrm{G}(x'\theta_{o1})^y \cdot (1-\mathrm{G}(x'\theta_{o1}))^{1-y}$, where $\mathrm{G}(\cdot)$ is the logistic *cdf*. Likelihood equations can be used to construct the GMM estimator based on the expectation of the score function $\mathbb{E}\left[\nabla_{\theta_1} \ln f(X, Y; \theta_1)\Big|_{\theta=\theta_{o1}}\right] = 0$, where $f(x, y; \theta_{o1})$ is the joint density of X and Y. If the distribution of X does not depend on θ_1, then $f(x, y; \theta_1) = p(y|x, \theta_1) f(x)$, where $f(x)$ is the unknown *pdf* of X. Then, the identifying moment condition can be rewritten as $\mathbb{E}\left\{\nabla_{\theta_1}[\ln p(Y|x; \theta_1) + \ln f(X)]\Big|_{\theta_1=\theta_{o1}}\right\} = \mathbb{E}\left[\nabla_{\theta_1} \ln p(Y|x; \theta_1)\Big|_{\theta_1=\theta_{o1}}\right]$ and the ML estimation is equivalent to the conditional ML estimation. For this example, $W = (X, Y)$, $m_1 = p_1$, and $g(W, \theta_1) = \frac{X'(Y-\mathrm{G}(X'\theta_1))}{\mathrm{G}(X'\theta_1)(1-\mathrm{G}(X'\theta_1))} \cdot \mathrm{g}(X'\theta_1)$, where $\mathrm{g}(\cdot)$ is the logistic *pdf*. □

Example 1.3.3 Consider estimation of the population averages μ_0 and μ_1 under control and treatment. Suppose a random sample is available of each unit's

outcome under both control and treatment. Let $Y(0)$ denote the outcome under control; $Y(1)$ under treatment. The identifying moment restriction for each group is $\mathbb{E}(Y(t) - \mu_{ot}) = 0, t = 0, 1$. So for this example $W = Y(t)$, $m_1 = p_1 = 1$, and $g(W, \theta) = Y(t) - \mu_t, t = 0, 1$. We can also consider the average treatment effect $\tau = \mu_1 - \mu_0$. □

The above model (1.10) holds in the entire (unselected) population. Now we consider the selected population defined by a random variable $S \in \{0, 1\}$ such that W is observed if and only if $S = 1$. We assume that the probability of selection depends on some additional variables Z, where $Z \in \mathcal{Z} \subset \mathbb{R}^{\dim(Z)}$ is always observed. Some or all of Z may be in W; that is, some of W may always be observed, but all of W is observed only when $S = 1$. Define

$$\mathrm{P}(z, \theta_{o2}) = \mathrm{P}(S = 1|z), \tag{1.11}$$

where $\mathrm{P}(z, \theta_2)$ is a parametric model for the probability of selection and is known up to the parameter vector $\theta_2 \in \Theta_2 \subset \mathbb{R}^{p_2}$. Again, in many problems, the joint density of $\{S, Z\}$ can be written as the product $\mathrm{P}(s|z, \theta_2)r(z)$, where $r(z)$ is the *pdf* of Z.

Assume $\{S, Z\}$ is a subvector of W^* from P_{θ_o}. Suppose there exists a real valued measurable function $u : \{0, 1\} \times \mathcal{Z} \times \Theta_2 \to \mathbb{R}^{m_2}$ $(m_2 \geq p_2)$ such that

$$\mathbb{E}\, u(S, Z; \theta_2) = 0, \text{ if and only if } \theta_2 = \theta_{o2}. \tag{1.12}$$

(The expectation is with respect to P_{θ_o}.) Assume that (1.12) satisfies Assumption 1.2.1. We call moment condition (1.12) the "selection moment condition".

Examples 1.3.4-1.3.6 in the next section show how (1.12) can be obtained from (1.11).

The GMM estimator based on (1.10), but with missing data, in effect makes the empirical moments $\frac{1}{N}\sum_{i=1}^{N} S_i g(W_i, \theta_1)$ close to zero. These empirical moments are the random sample analogues of the population moments of the form

$$\mathbb{E}[S\, g(W, \theta_1)] = 0. \tag{1.13}$$

We call these moment conditions the "unweighted selected population moments" to emphasize that they hold in the selected rather than the target population and to distinguish them from the weighted selected population moments that we will define shortly. The selectivity problem is that the unweighted selected population moment conditions (1.13) may not hold at θ_{o1}; more precisely, the value θ_{o1} that solves (1.10) may not solve (1.13).

We also consider the "weighted selected population moments" that weight the moment function in (1.13) by the inverse of the selection probability (see, e.g., Horvitz and Thompson, 1952):

$$\mathbb{E}\left[\frac{S}{\mathrm{P}(Z, \theta_2)} g(W, \theta_1)\right] = 0. \tag{1.14}$$

The weighted selected population moments also may not hold. Indeed, it is intuitively clear that whether (1.13) or (1.14) hold must depend on what is assumed about the relationship of the selection mechanism and W.

1.3.2 Motivation and definitions

We follow Wooldridge (2002b, 2005) in making the following "ignorability" (or "selection on observables") assumption.

Assumption 1.3.1 (*ignorability of selection*) $\mathrm{P}(S = 1|w, z) = \mathrm{P}(S = 1|z) = \mathrm{P}(z, \theta_{o2})$.

Assumption 1.3.1 says that, conditional on Z, S and W are independent. This is commonly written as $S \perp W \mid Z$. In some cases, ignorability is true by construction. An example would be the case that Z is an indicator of stratum, and selection is random within stratum. In other cases it is a substantial behavioral assumption.

As Wooldridge notes, this assumption does not imply that the unweighted selected population moment conditions (1.13) hold at θ_{o1}. This can be seen as follows:

$$\begin{aligned} \mathbb{E}S \cdot g(W, \theta_{o1}) &= \mathbb{E}\mathbb{E}[S \cdot g(W, \theta_{o1})|z], \text{ using LIE} \\ &= \mathbb{E}\mathbb{E}(S|z)\mathbb{E}[g(W, \theta_{o1})|z], \text{ using ignorability} \\ &= \mathbb{E}\mathrm{P}(Z, \theta_{o2})\mathbb{E}[g(W, \theta_{o1})|z], \end{aligned} \tag{1.15}$$

(where LIE means law of iterated expectations), and our assumptions do not in general imply that $\mathbb{E}[g(W, \theta_{o1})|z] = 0$. However, the weighted selected moment

conditions (1.14) do hold at θ_o, since

$$\begin{aligned} \mathbb{E}\frac{S}{\mathrm{P}(Z,\theta_{o2})}g(W,\theta_{o1}) &= \mathbb{E}\mathbb{E}[\frac{S}{\mathrm{P}(Z,\theta_{o2})}g(W,\theta_{o1})|z] \\ &= \mathbb{E}\frac{1}{\mathrm{P}(Z,\theta_{o2})}\mathbb{E}(S|z)\mathbb{E}[g(W,\theta_{o1})|z] \\ &= \mathbb{E}\mathbb{E}[g(W,\theta_{o1})|z] \\ &= \mathbb{E}g(W,\theta_{o1}) = 0. \end{aligned} \tag{1.16}$$

The simplest assumption under which the unweighted moment condition (1.13) holds in the selected sample is the following.

Assumption 1.3.2 $\mathrm{P}(S = 1|w) = \mathrm{P}(S = 1)$. *That is, S is independent of W.*

This assumption is easy to understand and clearly implies that (1.13) holds, since S is independent of $g(W, \theta_1)$. This condition is sometimes referred to as "missing completely at random" (see, e.g., Little and Rubin, 2002) but we will not use this terminology further, since there seems to be some inconsistency in the literature in the use of these words.

It should be noted that this assumption is neither stronger nor weaker than the assumption of ignorability (Assumption 1.3.1). That is, "S independent of W" does not imply, and is not implied by, "S independent of W conditional on Z". It is perhaps intuitive that the first condition is stronger than the second, but in fact that intuition is not correct.[1]

[1]The intuition referred to here is based on the fact that, for general Y, X_1, X_2, $\mathbb{E}(Y|x_1, x_2) = 0$ does imply that $\mathbb{E}(Y|x_1) = 0$ by the law of iterated expectations. But there is no comparable law for conditional independence.

The simplest assumption under which both the unweighted and the weighted moment conditions hold is the following.

Assumption 1.3.3 *(independence of selection)* (S, Z) *is independent of* W.

This assumption is also easy to understand, but it would appear to be too strong to apply in practical cases.

We now consider an exogeneity condition that is weaker than 1.3.3 and which does imply that both the weighted and unweighted moment conditions hold.

Assumption 1.3.4 *(exogeneity of selection)*

(i) Assumption 1.3.1 (ignorability of selection) holds.

(ii) $\mathbb{E}g(W, \theta_{o1})|z = 0$.

This is essentially the same definition of exogeneity as in Wooldridge (2005).

Under Assumption 1.3.4, selection is both ignorable and exogenous with respect to the primary problem of interest. For example, if $W = (Y, X)$ and $\mathcal{Z} \subseteq \mathcal{X}$, then having X in the conditioning set in the original problem is sufficient for the assumption to hold. If selection is based on covariates other than X, i.e. $\mathcal{X} \subseteq \mathcal{Z}$, then $g(Y, X; \theta_1)$ has to be uncorrelated with any function of $X^- \in \mathcal{Z} \setminus \mathcal{X}$ given X.

We now show that under Assumption 1.3.4, both the weighted and unweighted moment conditions hold. We first state without proof the following basic result.

Lemma 1.3.1 *Suppose Assumption 1.3.1 holds. Then* $f(w|z,s) = f(w|z)$.

(Here $f(\cdot)$ is generic notation for probability density.) Then it is easy to see that the following result is true.

Theorem 1.3.1 *Suppose Assumption 1.3.4 (exogeneity) holds. Then*

$$\mathbb{E}g(W,\theta_{o1})|z,s=0. \tag{1.17}$$

This is a much simpler and stronger result than Wooldridge obtained. It immediately implies that any function of Z and S is uncorrelated with $g(W,\theta_{o1})$, and therefore that the unweighted moment condition (1.13) and the weighted moment condition (1.14) both hold in the selected sample. In fact, this is true whether or not the weights are correct (in the sense that they do in fact represent $P(S=1|z)$). All that is required is that the weights be a function of Z and S.

We conclude that under *ignorable* selection,

$$\mathbb{E}\begin{bmatrix} \frac{S}{\mathrm{P}(Z,\theta_{o2})}g(W,\theta_{o1}) \\ u(S,Z;\theta_{o2}) \end{bmatrix} = 0 \tag{1.18}$$

and, under *independent* or *exogenous* selection,

$$\mathbb{E}\begin{bmatrix} Sg(W,\theta_{o1}) \\ u(S,Z;\theta_{o2}) \end{bmatrix} = 0, \tag{1.19}$$

$$\mathbb{E}\left[\begin{array}{c} \frac{S}{\mathrm{P}(Z,\theta_{o2})} g(W,\theta_{o1}) \\ u(S,Z;\theta_{o2}) \end{array}\right] = 0, \tag{1.20}$$

$$\mathbb{E}\left[\begin{array}{c} \left[\begin{array}{c} Sg(W,\theta_{o1}) \\ \frac{S}{\mathrm{P}(Z,\theta_{o2})} g(W,\theta_{o1}) \end{array}\right] \\ u(S,Z;\theta_{o2}) \end{array}\right] = 0. \tag{1.21}$$

Example 1.3.4 Suppose that sampling in Example 1.3.1 is nonrandom and the selection mechanism can be modelled as a PROBIT. Then, $\mathrm{P}(Z,\theta_2) = \Phi(Z'\theta_2)$, where $\Phi(\cdot)$ is standard normal *cdf*. Then, the selection moment conditions for this problem contain the likelihood equations for the log-likelihood $l(\theta_2|s,z) \equiv s\ln\Phi(z'\theta_2) + (1-s)\ln(1-\Phi(z'\theta_2))$. Thus, $m_2 = p_2$ and

$$u(S,Z;\theta_2) = \frac{Z'(S-\Phi(Z'\theta_2))}{\Phi(Z'\theta_2)\,(1-\Phi(Z'\theta_2))} \cdot \phi(Z'\theta_2),$$

where $\phi(\cdot)$ is the standard normal *pdf*. Note that we not only have $\mathbb{E}u(S,Z;\theta_{o2}) = 0$ but also $\mathbb{E}[u(S,Z;\theta_{o2})|z] = 0$. Under the *ignorability of selection* assumption, we can use the moment condition $\mathbf{E}\frac{S}{\Phi(Z'\theta_{o2})} g(X,Y;\theta_{o1}) = 0$, where $g(\cdot)$ is defined in Example 1.3.1. □

Example 1.3.5 (Variable Probability Sampling) Suppose that sampling in Example 1.3.2 is stratified. Let the sample space $\mathcal{W}$ be partitioned into J nonempty and disjoint strata $\mathcal{W}_1, \mathcal{W}_2, \ldots, \mathcal{W}_J$. If an observation lies in stratum $\mathcal{W}_j$, it is retained with probability P_{oj} that is usually known. So, the selection predictor Z can be defined by vector $(Z_1,\ldots,Z_J)'$ with $Z_j = \mathbb{I}\{W \in \mathcal{W}_j\}$, $j = 1,2,\ldots,J$, where $\mathbb{I}\{\cdot\}$ is the indicator function, and the probability model $\mathrm{P}(z,\theta_{o2}) = \sum_{j=1}^{J} \mathrm{P}_{oj} z_j$, where $\theta_{o2} = (\mathrm{P}_{o1},\ldots,\mathrm{P}_{oJ})'$. The *ignorability* assumption is satisfied by design. We

have $\mathrm{P}(s|z,\theta_2) = \prod_{j=1}^{J}\left[\mathrm{P}_j^s \cdot (1-\mathrm{P}_j)^{1-s}\right]^{z_j}$ in (1.11). Hence, the selection moment function for this problem contains the likelihood equation of the log-likelihood function $l(\theta_2|s,z) \equiv \sum_{j=1}^{J} z_j[s\ln \mathrm{P}_j + (1-s)\ln(1-\mathrm{P}_j)]$. Thus, $m_2 = p_2 = J$ and $u(S,Z;\theta_2) = \left(\frac{Z_1(S-\mathrm{P}_1)}{\mathrm{P}_1(1-\mathrm{P}_1)}, \ldots, \frac{Z_J(S-\mathrm{P}_J)}{\mathrm{P}_J(1-\mathrm{P}_J)}\right)'$. The weighted selected population moment condition contains $\frac{S}{\sum_{j=1}^{J}\mathrm{P}_j Z_j} \cdot g(X,Y;\theta_1)$, where $g(\cdot)$ is defined in Example 1.3.2. If, in addition, stratification is based on exogenous variables, i.e. *exogeneity of selection* assumption holds, the unweighted moment conditions (1.13) can also be used. □

Example 1.3.6 (Average Treatment Effect) Suppose that the sample in Example 1.3.3 is not entirely observed. Instead we observe $Y(0)$ only for the units that are in the control group and $Y(1)$ only for those that are in the treatment group. Understandably, the counterfactual data are missing. If Z are treatment predictors, the selection model for the treatment group is $\mathrm{P}(S=1|z) = P(z;\theta_{o2})$ and for the control group $\mathrm{P}(S=0|z) = 1 - P(z;\theta_{o2})$, where $\mathrm{P}(z;\theta_{o2})$ is the probability of receiving treatment. The *ignorability of selection* assumption implies in this case that $\mathrm{P}(S=1|y(0),z) = \mathrm{P}(S=1|z)$ and $\mathrm{P}(S=0|y(1),z) = \mathrm{P}(S=0|z)$. The selection moment condition for this example is the same as in Example 1.3.4. The weighted population moment condition will contain $\frac{S}{\mathrm{P}(Z;\theta_{o2})}(Y(1)-\mu_{o1})$ for the treatment group and $\frac{1-S}{1-\mathrm{P}(Z;\theta_{o2})}(Y(0)-\mu_{o0})$ for the control group. The average treatment effect can be identified using $\frac{S}{\mathrm{P}(Z;\theta_{o2})}Y(1) - \frac{1-S}{1-\mathrm{P}(Z;\theta_{o2})}Y(0) - \tau_o$. □

1.3.3 Relative efficiency results under ignorable selection

First consider estimation based on (1.18), under ignorable selection. Following the notation of Section 1.2 we write the weighted selected population moment

condition as $\mathbb{E}h_1(W^*, \theta_{o1}, \theta_{o2}) = 0$, where W^* contains W, S and Z, and where

$$h_1(W^*, \theta_{o1}, \theta_{o2}) = \frac{S}{\mathrm{P}(Z, \theta_{o2})} g(W, \theta_{o1}). \tag{1.22}$$

Wooldridge (2005) discusses estimation based on (1.22), for the exactly identified case. He compares the estimator of θ_{o1} when θ_{o2} is known to the estimator of θ_{o1} when θ_{o2} is replaced by some consistent estimate $\hat{\theta}_2$. In order to analyze this or other related issues, we have to say something about how θ_{o2} is estimated. In general terms, it is estimated by GMM based on a moment condition $\mathbb{E}h_2(S, Z; \theta_{o2}) = 0$, which puts the analysis into the framework of Section 1.2. However, following Wooldridge, we make the specific assumption that θ_{o2} is estimated by MLE based on the model $\mathrm{P}(s = 1|z) = \mathrm{P}(z, \theta_{o2})$. That is, $h_2(S, Z; \theta_{o2})$ is the score function corresponding to the likelihood for this model. Specifically,

$$h_2(S, Z; \theta_{o2}) = u(S, Z; \theta_{o2}) = \nabla_{\theta_2} \mathrm{P}(Z, \theta_2)|_{\theta_2 = \theta_{o2}} \frac{S - \mathrm{P}(Z, \theta_{o2})}{\mathrm{P}(Z, \theta_{o2})[1 - \mathrm{P}(Z, \theta_{o2})]}. \tag{1.23}$$

Under these assumptions, we have the puzzle referred to in the Introduction; namely, the TWO-STEP estimator of θ_{o1} that uses $\hat{\theta}_2$ in (1.22) is better than the KNOW-θ_2 estimator that uses the true value of θ_{o2} in (1.22). We will verify that this result holds also in the case that (1.22) is overidentified, and also provide our explanation of the puzzle, using the results of Section 1.2. To apply these results

we need to do some calculations involving the following:

$$\begin{aligned} C_{12} &= \mathbb{E}h_1(W^*, \theta_o)h_2(S, Z, \theta_{o2})', \\ C_{22} &= \mathbb{E}h_2(S, Z, \theta_{o2})h_2(S, Z, \theta_{o2})', \\ D_{12} &= \mathbb{E}\,\nabla_{\theta_2} h_1(W^*, \theta_1, \theta_2)\Big|_{\theta=\theta_o}, \\ D_{22} &= \mathbb{E}\,\nabla_{\theta_2} h_2(S, Z, \theta_2)\Big|_{\theta_2=\theta_{o2}}. \end{aligned} \tag{1.24}$$

Theorem 1.3.2 *Under the ignorability of selection assumption,*

(a) $C_{12} = \mathbb{E}\frac{g(W,\theta_{o1})}{P(Z,\theta_{o2})}\,\nabla_{\theta_2} P(Z, \theta_2)\Big|_{\theta_2=\theta_{o2}}$, *which is (in general) not equal to zero;*

(b) $D_{12} = -C_{12}$, $D_{22} = -C_{22}$, *and so* $D_{12} = C_{12}C_{22}^{-1}D_{22}$.

To understand Theorem 1.3.2, note first that in the unselected population, $C_{12}^* \equiv \mathbb{E}g(W, \theta_{o1}) \cdot h_2(S, Z, \theta_{o2})' = 0$. That is, the original moment condition $g(W, \theta_{o1})$ is uncorrelated with the score function $h_2(S, Z, \theta_{o2})$ by the generalized information equality. However, in the selected sample, $C_{12} \neq 0$. That is, $h_1(W^*, \theta_{o1}, \theta_{o2})$ and $h_2(S, Z, \theta_{o2})$ are correlated. This correlation makes $h_2(S, Z, \theta_{o2})$ relevant for estimation of θ_{o1} even if θ_{o2} is known, and the inefficiency of the KNOW-θ_2 estimator is due to its failure to capture the information in the moment condition based on $h_2(S, Z, \theta_{o2})$.

Although we do not pursue this point, it would appear that the inefficiency of the KNOW-θ_2 estimator (at least relative to the KNOW-θ_2-JOINT estimator) would hold even if $h_2(S, Z, \theta_2)$ were not a score function. It depends only on $C_{12} \neq 0$, not on the particular form of C_{12}.

Part (b) of Theorem 1.3.2 gives a number of information equalities which do depend on $h_2(S, Z, \theta_2)$ being a score function. They establish that $D_{12} = C_{12}C_{22}^{-1}D_{22}$, which is the condition for Statements 10 and 11 of Theorem 1.2.3. Statement 11 of Theorem 1.2.3 says that the KNOW-θ_2 estimator is inefficient relative to the ONE-STEP, TWO-STEP and KNOW-θ_2-JOINT estimators. This extends the previously-cited (but, we hope, no longer puzzling!) result, namely that KNOW-θ_2 is inefficient relative to TWO-STEP, to a larger set of other estimators, and also to the case that the GMM problem for the parameters of interest is overidentified.

Statement 10 of Theorem 1.2.3 says further that θ_{o2} is p-redundant, so that the ONE-STEP and KNOW-θ_2-JOINT estimators are equally efficient. So long as one includes the score function $h_2(S, Z, \theta_{o2})$ in the estimation problem, it does not matter (in terms of efficiency of estimation of θ_{o1}) whether θ_{o2} is known or not.

Another note is that, although the TWO-STEP estimator is better than the KNOW-θ_2 estimator, it is not necessarily efficient. In the exactly identified case, it is efficient because it equals the ONE-STEP estimator (Statement 7 of Theorem 1.2.3), but in the overidentified case it is generally less efficient than the KNOW-θ_2-JOINT and ONE-STEP estimators.

Example 1.3.7 Continuing Example 1.3.4 under *ignorable* selection with the ML estimate of θ_{o2}, $\mathbb{E}[S \cdot u(S, z; \theta_{o2})'|z]$ can be written as

$$\mathbb{E}\left[S \cdot \frac{(S - \mathrm{P}(Z, \theta_{o2}))}{\mathrm{P}(Z, \theta_{o2})\,(1 - \mathrm{P}(Z, \theta_{o2}))} \cdot \nabla_{\theta_2} P(Z, \theta_2)\Big|_{\theta_2 = \theta_{o2}} \Big| z\right] =$$

$$= \frac{[\mathbb{E}(S^2|z) - \mathbb{E}(S|z) \cdot \mathrm{P}(z, \theta_{o2})]}{\mathrm{P}(z, \theta_{o2})\,(1 - \mathrm{P}(z, \theta_{o2}))} \cdot \nabla_{\theta_2} P(z, \theta_2)\Big|_{\theta_2 = \theta_{o2}} = \nabla_{\theta_2} P(z, \theta_2)\Big|_{\theta_2 = \theta_{o2}},$$

where the second equality follows because $\mathbb{E}(S^2|z) = \mathbb{E}(S|z)$ and $\mathbb{E}(S|z) = \mathrm{P}(z, \theta_2)$. This is non-zero. Also, $\mathbb{E}[g(W;\theta_1)|z] \neq 0$ unless there is also *exogeneity*. Thus, C_{12}, which can be expressed by the law of iterated expectations as

$$\mathbb{E}\left\{\frac{1}{P(Z,\theta_{o2})}\mathbb{E}[g(W,\theta_{o1})|z]\mathbb{E}[Su(S,z;\theta_{o2})'|z]\right\},$$

is generally non-zero. In fact, $C_{12} = \mathbb{E}\left[\frac{g(W,\theta_{o1})}{\mathrm{P}(Z,\theta_{o2})} \cdot \nabla_{\theta_2} P(Z,\theta_2)\Big|_{\theta_2=\theta_{o2}}\right]$. We cannot therefore claim m-redundancy under *ignorability of selection*: using orthogonality conditions from the selection process helps in estimating θ_{o1} even if the weighting probabilities are known. However, we can claim p-redundancy by Theorem 1.3.2: using known selection probabilities with the additional moment conditions for selection is as efficient as estimating the probabilities in a one-step or two-step procedure. Each of the three alternatives is equally preferred to only using the original problem with known probabilities. □

Example 1.3.8 Continuing Example 1.3.5 under *ignorability* with the ML estimates of P_{oj}, $\mathbb{E}[S \cdot u(S,z;\theta_{o2})'|z]$ contains elements of the form $\frac{z_j(\mathbb{E}(S^2|z)-\mathbb{E}(S|z)\cdot \mathrm{P}_{oj})}{\mathrm{P}_{oj}\left(1-\mathrm{P}_{oj}\right)}$, $j = 1,\ldots,J$. Since $\mathbb{E}(S^2|z) = \mathbb{E}(S|z) = \sum_{j=1}^{J} \mathrm{P}_{oj} z_j$, the elements can be written as $\frac{z_j \sum_{j=1}^{J} \mathrm{P}_{oj} z_j}{\mathrm{P}_{oj}}$, $j = 1,\ldots,J$. Thus, C_{12}, which can be expressed by the law of iterated expectations as $\mathbb{E}\left\{\frac{1}{\sum_{j=1}^{J} \mathrm{P}_{oj} Z_j} \cdot \mathbb{E}[g(W,\theta_{o1})|z] \cdot \mathbb{E}[Su(S,z;\theta_{o2})'|z]\right\}$, can be simplified to $\mathbb{E}\left[\frac{g(W,\theta_{o1})\mathbb{I}\{W\in\mathcal{W}_1\}}{\mathrm{P}_{o1}},\ldots,\frac{g(W,\theta_{o1})\mathbb{I}\{W\in\mathcal{W}_J\}}{\mathrm{P}_{oJ}}\right]$. This is nonzero, unless there is also the *exogeneity* or *independence* assumption. Similarly to Example 1.3.7, under *ignorable* selection, using selection moment conditions increases precision of estimating θ_{o1}. Also, if knowledge of selection probabilities is available it provides for the same precision of $\hat{\theta}_1$ as the one-step or two-step procedures as long as all $m_1 + m_2$ moment conditions are used. □

Example 1.3.9 Continuing Example 1.3.6, with ML estimates of treatment probabilities $\mathrm{P}(z;\hat{\theta}_2)$ from PROBIT, the correlation matrix between the moment conditions that identify μ_{o1} and the likelihood equations from PROBIT for θ_{o2} is $\mathbb{E}\left[\frac{S(Y(1)-\mu_{o1})}{\mathrm{P}(Z;\theta_{o2})} \times \frac{(S-\mathrm{P}(Z;\theta_{o2}))\nabla_{\theta_2}\mathrm{P}(Z;\theta_2)\big|_{\theta_2=\theta_{o2}}}{(1-\mathrm{P}(Z;\theta_{o2}))}\right]$. This, under *ignorability*, can be rewritten as $-\mathbb{E}\left[\frac{(Y(1)-\mu_{o1})\nabla_{\theta_2}\mathrm{P}(Z;\theta_2)|_{\theta_2=\theta_{o2}}}{\mathrm{P}(Z;\theta_{o2})}\right]$ which is non-zero unless $Y(1)\perp Z$ and equal to minus the expected derivative of the weighted moment equation for μ_{o1} with respect to θ_2. A similar argument is valid for estimating μ_{o0} and, consequently, τ_o. Hence, in average treatment effect estimation, m-redundancy cannot be claimed: knowledge about the treatment assignment process should be included into the estimation. There is p-redundancy, however: it does not matter asymptotically whether the parameters of the assignment process are known or estimated as long as all available moments are used. □

1.3.4 Relative efficiency results under exogenous selection

Consider now estimation based on (1.19)-(1.21), under exogenous selection.

Wooldridge (2002b, Theorem 5.2) shows, under the *exogenous selection* assumption, that the IPW M-estimator that uses known selection probabilities is as efficient as a two-step estimator that employs initial ML estimates of the selection probabilities. The results of Section 1.2 allow to restate this result for other estimators and for the cases of overidentification in the primary problem of interest.

Using definitions (1.22)-(1.24), it is easy to verify that, for the GMM estimator based on (1.20), the following is true.

Theorem 1.3.3 *Under the exogeneity of selection assumption:*

(a) $C_{12} = 0$;

(b) $D_{12} = 0$.

So, by Theorem 1.2.3(6), we have m-redundancy of the selection moment condition and p-redundancy of θ_{o2}. ONE-STEP, TWO-STEP, KNOW-θ_2 and KNOW-θ_2-JOINT estimators of θ_{o1} are equally efficient asymptotically.

Wooldridge (2005, Theorem 4.3) shows, under exogeneity and the further assumption that the original moment conditions satisfy the conditional information matrix equality, that the estimator based on the unweighted moment conditions is more efficient than the estimator based on the weighted moment conditions. This is fine as far as it goes, but it does not rule out the possibility that using both could be more efficient than using either. Our next result does rule out this possibility.

Theorem 1.3.4 *Suppose Assumption 1.3.4 holds. Then the optimal moment conditions in the selected population are the same as in the unselected population.*

To see why this result is true, note that the optimal moment conditions in the unselected population are the following:

$$\mathbb{E}D(Z)'C(Z)^{-1}g(W,\theta_{o1}) = 0, \tag{1.25}$$

where $D(z) = \mathbb{E}\ \nabla_{\theta_1} g(W,\theta_1)\Big|_{\theta_1=\theta_{o1}} \Big| z$ and $C(z) = \mathbb{E}g(W,\theta_{o1})g(W,\theta_{o1})'|z$. The optimal moment conditions in the selected population are:

$$\mathbb{E}D(Z,S=1)'C(Z,S=1)^{-1}Sg(W,\theta_{o1}) = 0, \tag{1.26}$$

where $D(z,S=1) = \mathbb{E}\left\{ \nabla_{\theta_1} g(w,\theta_1)\Big|_{\theta_1=\theta_{o1}} \Big| z, S=1 \right\}$ and $C(z,S=1) = \mathbb{E}\left\{g(W,\theta_{o1})g(W,\theta_{o1})'|z,S=1\right\}$. But $D(z,S=1) = D(z)$ by the ignorability assumption, and similarly $C(z,S=1) = C(z)$.

An implication of this result is that the weighted moment conditions are m-redundant for the estimation of θ_{o1}. More precisely, assuming that weighting was not part of the efficient estimation problem in the unselected population, it also plays no role in the efficient problem in the selected population. Thus in this circumstance we do not have to weight for reasons of consistency, and we also do not have to weight for reasons of efficiency.

Theorem 1.3.4 is a useful result, but it falls short of being the final word on efficiency. The question is whether the moment conditions in equation (1.17) capture all of the information in the exogeneity assumption. The first part of the exogeneity assumption is that $\mathbb{E}g(W,\theta_{o1})|z = 0$, and the efficient GMM estimator under this conditional moment restriction (with full observability) is well understood. The second part of the exogeneity assumption is the ignorability condition, and Theorem 1.3.1 shows that this makes the original conditional moment restriction valid in the selected sample as well. More precisely, we then have $\mathbb{E}g(W,\theta_{o1})|z,s=0$ and Theorem 1.3.4 gives the form of the efficient estimator under this conditional moment restriction. However, what is not clear is whether all of the information in the ignorability condition is captured by the extension of the original moment

conditions to the selected population.

We defined $\mathrm{P}(z, \theta_{o2}) = \mathbb{E}(S|z)$, so that $\mathbb{E}[S - \mathrm{P}(z, \theta_{o2})]|z = 0$. However, under ignorability, we have the stronger condition that $\mathbb{E}[S - \mathrm{P}(z, \theta_{o2})]|z, w = 0$. The score function for estimation of θ_{o2}, as given in (1.23) above, will not be useful for estimation of θ_{o1}, because it is a function of Z and S only, and we have already used the optimal functions of Z and S in (1.26) above. The question is whether the fact that $\mathbb{E}[S - \mathrm{P}(z, \theta_{o2})]|w = 0$ adds anything. This question is complicated by the fact that W is only observed when $S = 1$. If no part of W (other than Z, if Z is a subset of W) is always observed, we do not see any way to make use of the condition that $\mathbb{E}[S - \mathrm{P}(z, \theta_{o2})]|w = 0$. However, now suppose that some subset of W is always observed. Let W_o be the part of W which (i) is always observed, and (ii) is not part of Z. Then we can consider moment conditions of the form

$$\mathbb{E}k(W_o)[S - \mathrm{P}(z, \theta_{o2})] = 0. \tag{1.27}$$

These moment conditions are not useful for estimation of θ_{o2}, but they may be useful for estimation of θ_{o1}, if they are correlated with the original moment conditions. It is easy to see that they are not correlated with the unselected original moment conditions:

$$\begin{aligned} \mathbb{E}g(W, \theta_{o1})\, k(W_o)' \left[S - \mathrm{P}(z, \theta_{o2})\right] &= \mathbb{E}\,\mathbb{E}[S - \mathrm{P}(z, \theta_{o2})]|z\, \mathbb{E}g(W, \theta_{o1})k(W_o)'|z \\ &= 0. \end{aligned}$$

However, they *are* correlated with the selected original moment conditions:

$$\begin{aligned}\mathbb{E}Sg(W,\theta_{o1})\,k(W_o)'\,[S-\mathrm{P}(Z,\theta_{o2})] &= \mathbb{E}\,\mathbb{E}S[S-\mathrm{P}(z,\theta_{o2})]|z\,\mathbb{E}g(W,\theta_{o1})k(W_o)'|z \\ &= \mathbb{E}\mathrm{P}(Z,\theta_{o2})[1-\mathrm{P}(Z,\theta_{o2})]\,\mathbb{E}g(W,\theta_{o1})k(W_o)'|z \\ &\neq 0.\end{aligned}$$

Thus the moment conditions in (1.27) may possibly be useful in estimation of θ_{o1}. We leave further exploration of this point for future work.

1.4 Concluding remarks

We summarize relative efficiency results for four alternative GMM estimators of a parameter vector that enters into one set of moment conditions along with another vector that also enters into an additional set of conditions and may be known.

We provide formal statements and proofs of efficiency claims and spell out conditions under which some knowledge may be redundant. If the two sets of moment conditions are uncorrelated and the expected derivative of the first set with respect to the additional parameter vector is zero, both the additional moment conditions and the knowledge of the additional parameters are redundant. These are the strongest sufficient conditions we consider. The weaker condition of moment uncorrelatedness is sufficient for redundancy of extra moment conditions when the additional parameters are known and for equal efficiency of the multi-step and one-step estimators under exact identification of the original set of moment conditions. The condition of zero expected derivative of the original set of moments with respect to the additional parameter vector turns out to be sufficient for no influence

of the first step estimation over the second step standard errors in very general settings. We provide a sufficient condition for equal relative efficiency of the estimator that treats additional parameters as known using the full set of moment conditions and the estimator that involves estimating both parameter vectors.

We apply these results to a general missing data problem after showing that the weighted and unweighted GMM estimators on the selected sample preserve desired asymptotic properties under reasonable assumptions. We explain the counterintuitive result that estimating selection probabilities dominates using known probabilities if this knowledge is available. It turns out that this is an outcome of ignoring the moment conditions that characterize the selection process. Interestingly, however, a proper use of such knowledge along with known selection probabilities turns out to be as good as estimating the probabilities using the same moment conditions. Redundancy of the parameter knowledge applies. We show that this redundancy result is driven by two factors: the ignorability assumption on selection and the use of the score function in estimation of the selection probabilities. The ignorability condition says that the first-stage score function for the conditional likelihood $f(s|z)$ is in fact the score function for the conditional likelihood $f(s|z,w)$ and thus GCIME can be applied producing the sufficient condition for parameter knowledge redundancy.

When selection is based on exogenous variables with respect to a correctly specified feature of conditional distribution, any function of the exogenous variable can be used as a weight in the weighted GMM estimation. This implies two interesting results. First, the weighted GMM estimation on the selected sample is robust to selection model misspecification. Second, using both weighted and unweighted moment conditions dominates using only one of them unless the original

moment function incorporates the optimal weights in the first place. No efficiency improvements are possible in that case.

Besides the examples we give, the following specific missing data problems can be studied in the framework of Section 1.2: using auxiliary data to estimate probabilities of selection (see Hellerstein and Imbens, 1999; Nevo, 2002, 2003), weighting by nonparametric estimates of propensity scores in estimation of average treatment effects (see Hirano et al., 2003), estimating weights for choice-based samples in pseudo-MLE settings (see Manski and Lerman, 1977; Manski and McFadden, 1981; Cosslett, 1981a,b; Imbens, 1992), EL and GMM estimation for stratified samples with possibly known sampling or population frequencies (see Tripathi, 2003).

Bibliography

AHN, S. AND P. SCHMIDT (1995): "A separability result for GMM estimation, with applications to GLS prediction and conditional moment tests," *Econometric Reviews*, 14, 19–34.

BREUSCH, T., H. QIAN, P. SCHMIDT, AND D. WYHOWSKI (1999): "Redundancy of moment conditions," *Journal of Econometrics*, 91, 89–111.

COSSLETT, S. R. (1981a): "Efficient estimation of discrete-choice models," in *Structural Analysis of Discrete Data and Econometric Applications*, ed. by C. F. Manski and D. L. McFadden, Cambridge: The MIT Press, 51–111.

——— (1981b): "Maximum likelihood estimator for choice-based samples," *Econometrica*, 49, 1289–1316.

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, 135–156.

GOLDBERGER, A. (1972): "Maximum likelihood estimation of regressions containing unobservable independent variables," *International Economic Review*, 13, 1–15.

HAHN, J. (1998): "On the role of the propensity score in efficient semiparametric estimation of average treatment effects," *Econometica*, 66, 315–331.

HANSEN, L. (1982): "Large sample properties of generalized method of moments estimators," *Econometrica*, 50, 1029–1054.

HECKMAN, J. J., H. ICHIMURA, AND P. TODD (1998): "Matching as an econometric evaluation estimator," *The Review of Economic Studies*, 65, 261–294.

HELLERSTEIN, J. K. AND G. W. IMBENS (1999): "Imposing moment restrictions from auxiliary data by weighting," *The Review of Economics and Statistics*, 81, 1–14.

HIRANO, K., G. IMBENS, AND G. RIDDER (2003): "Efficient estimation of average treatment effects using the estimated propensity score," *Econometrica*, 71, 1161–1189.

HORVITZ, D. AND D. THOMPSON (1952): "A generalization of sampling without replacement from a finite universe," *Journal of the American Statistical Association*, 47, 663–685.

IMBENS, G. W. (1992): "An efficient method of moments estimator for discrete choice models with choice-based sampling," *Econometrica*, 60, 1187–1214.

LITTLE, R. J. A. AND D. B. RUBIN (2002): *Statistical analysis with missing data*, Wiley series in probability and statistics, Wiley-Interscience, 2 ed.

MANSKI, C. F. AND S. R. LERMAN (1977): "The estimation of choice probabilities from choice based samples," *Econometrica*, 45, 1977–1988.

MANSKI, C. F. AND D. L. MCFADDEN (1981): "Alternative estimators and sample designs for discrete choice analysis," in *Structural Analysis of Discrete Data with Econometric Applications*, ed. by C. F. Manski and D. L. McFadden, The MIT Press, 2–50.

NEVO, A. (2002): "Sample selection and information-theoretic alternatives to GMM," *Journal of Econometrics*, 107, 149–157.

——— (2003): "Using weights to adjust for sample selection when auxiliary information is available," *Journal of Business and Economic Statistics*, 21, 43–53.

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

NEWEY, W. AND D. MCFADDEN (1994): "Large sample estimation and hypothesis testing," in *Handbook of Econometrics*, ed. by R. Engle and D. McFadden, vol. IV, 2113–2241.

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

QIAN, H. AND P. SCHMIDT (1999): "Improved instrumental variables and generalized method of moments estimators," *Journal of Econometrics*, 91, 145–169.

ROBINS, J. M. AND A. ROTNITZKY (1995): "Semiparametric efficiency in multivariate regression models with missing data," *Journal of the American Statistical Association*, 90, 122–129.

ROSENBAUM, P. R. (1987): "Model-based direct adjustment," *Journal of American Statistical Association*, 82, 387–394.

ROSENBAUM, P. R. AND D. B. RUBIN (1983): "The central role of the propensity score in observational studies for causal effects," *Biometrika*, 70, 41–55.

TRIPATHI, G. (2003): "GMM and empirical likelihood with stratified data," *Working Paper, University of Wisconsin.*

WOOLDRIDGE, J. (1999): "Asymptotic properties of weighted M-estimators for variable probability samples," *Econometrica*, 67, 1385–1406.

——— (2001): "Asymptotic properties of weighted M-estimators for standard stratified samples," *Econometric Theory*, 17, 451–470.

——— (2002a): *Econometric analysis of cross section and panel data*, Cambridge, Mass.: MIT Press.

——— (2002b): "Inverse probability weighted M-estimators for sample selection, attrition and stratification," *Portuguese Economic Journal*, 1, 117–139.

——— (2003): "Inverse probability weighted estimation for general missing data problems," *Working Paper, Michigan State University*, www.msu.edu/~ec/faculty/wooldridge/current%20research/wght2r6.pdf.

——— (2005): "Inverse probability weighted estimation for general missing data problems," *Working Paper, Michigan State University*, www.msu.edu/~ec/faculty/wooldridge/current%20research/wght2r7.pdf.

ZELLNER, A. (1970): "Estimation of regression relationships containing unobservable independent variables," *International Economic Review*, 11, 441–454.

Appendix: Proofs

PROOF OF THEOREM 1.2.1: Proofs are given, e.g., in Theorems 2.6 and 3.4 of Newey and McFadden (1994). Also, see Hansen (1982). Condition (i) is the identification assumption. Conditions (ii) and (iv) are needed for consistency, conditions (iii)-(v) are needed for asymptotic normality, while conditions (iv) and (v) ensure that the objective function in (1.2) and its first derivative, respectively, converge uniformly to their population analogues and condition (vi) provides for invertibility of a part of the mean-value expansion. Some of the conditions can be relaxed at the expense of complicating proofs. □

PROOF OF THEOREM 1.2.2: Equations (1.6), (1.8), and (1.9) follow from the standard asymptotic variance derivation for the GMM estimation using the optimal weighting matrix (see, e.g., p. 2148 of Newey and McFadden, 1994; Hansen, 1982, Theorems 3.1 and 3.2).

Equation (1.7) is obtained similarly but we separately expand the first order conditions corresponding to (A) and (B).

The TWO-STEP estimator of θ_{o2} minimizes $\bar{h}_2(\theta_2)' C_{22}^{-1} \bar{h}_2(\theta_2)$. The first order conditions that the estimator solves are $D_{22}' C_{22}^{-1} \bar{h}_2(\hat{\theta}_2) = 0$. Expanding around θ_2 gives

$$\hat{\theta}_2 - \theta_{o2} = -(D_{22}' C_{22}^{-1} D_{22})^{-1} D_{22}' C_{22}^{-1} \bar{h}_2(\theta_{o2}) + o_p(N^{-1/2}). \tag{1.28}$$

The TWO-STEP estimator of θ_{o1} minimizes $\bar{h}_1(\theta_1, \hat{\theta}_2)' C_{22}^{-1} \bar{h}_1(\theta_1, \hat{\theta}_2)$. The first order conditions that the estimator solves are $D_{11}' C_{11}^{-1} \bar{h}_1(\hat{\theta}_1, \hat{\theta}_2) = 0$. Expanding around θ_{o1} and using (1.28) gives

$$\begin{aligned} \hat{\theta}_1 - \theta_{o1} &= -(D_{11}' C_{11}^{-1} D_{11})^{-1} D_{11}' C_{11}^{-1} \bar{h}_1(\theta_{o1}, \theta_{o2}) + \\ &+ (D_{11}' C_{11}^{-1} D_{11})^{-1} D_{11}' C_{11}^{-1} D_{12} (D_{22}' C_{22}^{-1} D_{22})^{-1} D_{22}' C_{22}^{-1} \bar{h}_2(\theta_{o2}) \\ &+ o_p(N^{-1/2}). \end{aligned} \tag{1.29}$$

On multiplying by $\sqrt{N}$ and combining (1.28)-(1.29), we get

$$\mathbb{V}_{\text{TWO-STEP}} = BCB', \tag{1.30}$$

where C is defined in (1.4) and

$$B = \begin{pmatrix} B_{11} & B_{12} \\ 0 & B_{22} \end{pmatrix} \tag{1.31}$$

with

$$\begin{aligned} B_{11} &= -(D_{11}' C_{11}^{-1} D_{11})^{-1} D_{11}' C_{11}^{-1}, \\ B_{12} &= (D_{11}' C_{11}^{-1} D_{11})^{-1} D_{11}' C_{11}^{-1} D_{12} (D_{22}' C_{22}^{-1} D_{22})^{-1} D_{22}' C_{22}^{-1}, \\ B_{22} &= -(D_{22}' C_{22}^{-1} D_{22})^{-1} D_{22}' C_{22}^{-1}. \end{aligned} \tag{1.32}$$

□

PROOF OF THEOREM 1.2.3: Statements 1 and 4 are proved on p. 148 of Qian and Schmidt (1999) where it is shown that there is no gain in efficiency if and only

if $D_{11}C_{11}^{-1}C_{12} = 0$. When the original problem is exactly identified ($m_1 = p_1$) and D_{11} is non-singular (by assumption), this is true if and only if $C_{12} = 0$. If the original problem is overidentified ($m_1 > p_1$) then the condition $C_{12} = 0$ is sufficient for no gain in efficiency.

To prove Statement 3 first note that $BD = -I$, where I is the $p_1 + p_2$ dimensional identity matrix. Then,

$$\begin{aligned} \mathbb{V}_{\text{TWO-STEP}} - \mathbb{V}_{\text{ONE-STEP}} &= BCB' - (D'C^{-1}D)^{-1} \\ &= BCB' - BD(D'C^{-1}D)^{-1}D'B' \\ &= BC^{\frac{1}{2}}[I - C^{-\frac{1}{2}}D(D'C^{-\frac{1}{2}}C^{-\frac{1}{2}}D)^{-1}D'C^{-\frac{1}{2}}]C^{\frac{1}{2}}B. \end{aligned} \tag{1.33}$$

The matrix is brackets is the projection orthogonal to $C^{-1/2}D$, which is positive semidefinite.

$\mathbb{V}_{\text{ONE-STEP}}$ for θ_1 is of the form $(D_{11}'C^{11}D_{11} - M_{12}M_{22}^{-1}M_{21})^{-1}$, where $M_{12} = M_{21}' = D_{11}'C^{11}D_{12} + D_{11}'C^{12}D_{22}$ and M_{22} is the lower right p_2-block of $D'C^{-1}D$, which is positive semidefinite. Hence, the inverse of $\mathbb{V}_{\text{ONE-STEP}}$ for θ_1 minus $\mathbb{V}^{-1}_{\text{KNOW-}\theta_2\text{-JOINT}}$ is negative semidefinite. Thus, $\mathbb{V}_{\text{ONE-STEP}}$ of θ_1 is no smaller than $\mathbb{V}_{\text{KNOW-}\theta_2\text{-JOINT}}$ in positive definite sense, which proves Statement 2.

Further, KNOW-θ_2-JOINT and ONE-STEP are equally efficient if $M_{12} = 0$ but $M_{12} = D_{11}'[C^{11}D_{12} + C^{12}D_{22}]$. This fact along with the fact that $C_{12}C_{22}^{-1} = -(C^{11})^{-1}C^{12}$ implies that if $D_{12} = C_{12}C_{22}^{-1}D_{22}$ than $M_{12} = 0$ which proves Statement 10.

Statement 11 can be proved in two parts. First, since $M_{12} = 0$ the inverse of

$\mathbb{V}_{\text{ONE-STEP}}$ for θ_1 is simply $D'_{11}C^{11}D_{11}$ which is generally greater than $\mathbb{V}^{-1}_{\text{KNOW-}\theta_2} = D'_{11}C_{11}^{-1}D_{11}$ in the positive definite sense since $C^{11} - C_{11}^{-1}$ is positive semidefinite. This along with Statement 10 implies that ONE-STEP and KNOW-θ_2-JOINT are no less efficient for θ_1 than KNOW-θ_2. Second, to prove that TWO-STEP is no less efficient for θ_1 than KNOW-θ_2 note that, by (1.30)-(1.32), $\mathbb{V}_{\text{TWO-STEP}}$ for θ_1 is equal to $B_{11}C_{11}B'_{11} + B_{12}C_{21}B'_{11} + B_{11}C_{12}B'_{12} + B_{12}C_{22}B'_{12}$. Also note that $B_{11}C_{11}B'_{11} = (D'_{11}C_{11}^{-1}D_{11})^{-1}$ and that, under $D_{12} = C_{12}C_{22}^{-1}D_{22}$, the symmetric positive semidefinite matrices $-B_{12}C_{21}B'_{11}$ and $-B_{11}C_{12}B'_{12}$ are equal to $B_{12}C_{22}B'_{12}$. $\mathbb{V}_{\text{TWO-STEP}}$ for θ_1 reduces therefore to $\mathbb{V}_{\text{KNOW-}\theta_2}$ minus a positive semidefinite matrix, which completes the second part of the proof.

Statements 7-9 follow from Theorem 1 of Ahn and Schmidt (1995) and subsequent discussion (pp. 21-22).

Statement 5 holds since if $D_{12} = 0$ then (1.7) reduces to $(D'_{11}C_{11}^{-1}D_{11})^{-1}$, which is equal to (1.8).

Statement 6 follows from Statements 4 and 10 and a trivial comparison of variances in (1.7) and (1.6) under given conditions. □

PROOF OF THEOREM 1.3.1: Follows trivially from Lemma 1.3.1 and part (ii) of Assumption 1.3.4. □

PROOF OF THEOREM 1.3.2: (a) First, note that, by ignorability and (1.23), $\mathbb{E}[S \cdot h_2(S, z; \theta_{o2})'|z]$ can be written as $\mathbb{E}[S \cdot \frac{(S - \mathrm{P}(z,\theta_{o2}))}{\mathrm{P}(z,\theta_{o2})(1-\mathrm{P}(z,\theta_{o2}))} \cdot \nabla_{\theta_2} P(z, \theta_2)|_{\theta_2=\theta_{o2}} |z] =$

$\frac{[\mathbb{E}(S^2|z)-\mathbb{E}(S|z)\cdot \mathrm{P}(z,\theta_{o2})]}{\mathrm{P}(z,\theta_{o2})(1-\mathrm{P}(z,\theta_{o2}))} \cdot \nabla_{\theta_2} P(z,\theta_2)|_{\theta_2=\theta_{o2}} = \nabla_{\theta_2} P(z,\theta_2)|_{\theta_2=\theta_{o2}}$, since $\mathbb{E}(S^2|z) = \mathbb{E}(S|z)$ and $\mathbb{E}(S|z) = \mathrm{P}(z,\theta_{o2})$. This is nonzero in general. Second, $\mathbb{E}[g(W;\theta_{o1})|z] \neq 0$ in general. Finally,

$$\begin{aligned} C_{12} &= \mathbb{E}h_1(W^*,\theta_{o1},\theta_{o2})h_2(S,Z,\theta_{o2})' \\ &= \mathbb{E}\{\tfrac{1}{P(Z,\theta_{o2})}\mathbb{E}[g(W,\theta_{o1})|z]\mathbb{E}[Sh_2(S,z;\theta_{o2})'|z]\}, \quad \text{by ignorability} \quad (1.34) \\ &= \mathbb{E}[\tfrac{g(W,\theta_{o1})}{\mathrm{P}(Z,\theta_{o2})} \cdot \nabla_{\theta_2} P(Z,\theta_2)|_{\theta_2=\theta_{o2}}], \quad \text{by LIE} \end{aligned}$$

which is generally non-zero.

(b) Follows by (generalized) information equality, where $h_2(\cdot)$ is the score, D_{22} is the expected Hessian, C_{22} is the expected outer product of the score, D_{12} is the expected derivative of h_1 with respect to θ_2 evaluated at θ_{o2} and C_{12} is the covariance of h_1 with the score. One may also write

$$\begin{aligned} D_{12} &= \mathbb{E}\{\nabla_{\theta_2} \tfrac{S}{P(Z,\theta_2)}\Big|_{\theta_2=\theta_{o2}} g(W;\theta_{o1})\}, \quad \text{by (1.22)} \\ &= -\mathbb{E}[\tfrac{S}{P(Z,\theta_{o2})^2} \cdot g(W;\theta_{o1}) \cdot \nabla_{\theta_2} P(Z,\theta_2)|_{\theta_2=\theta_{o2}}] \\ &= -\mathbb{E}[\tfrac{\mathbb{E}(S|z)\mathbb{E}(g(W;\theta_{o1})|z)}{P(Z,\theta_{o2})^2}\nabla_{\theta_2} P(Z,\theta_2)|_{\theta_2=\theta_{o2}}], \quad \text{by LIE} \\ &= -\mathbb{E}[\tfrac{g(W;\theta_{o1})}{P(Z,\theta_{o2})}\nabla_{\theta_2} P(Z,\theta_2)|_{\theta_2=\theta_{o2}}], \quad \text{as } \mathbb{E}(S|z) = \mathrm{P}(z,\theta_{o2}) \\ &= -C_{12} \quad \text{by (1.34)} \end{aligned}$$

□

PROOF OF THEOREM 1.3.3: (a) By LIE and exogeneity,

$$\begin{aligned} C_{12} &= \mathbb{E}\left[\frac{S}{P(Z,\theta_{o2})} g(Y,X;\theta_{o1}) \cdot u(S,Z;\theta_{o2})'\right] \\ &= \mathbb{E}\mathbb{E}\left[\frac{S}{P(z,\theta_{o2})} g(Y,X;\theta_{o1}) \cdot u(S,z;\theta_{o2})'|z\right] \\ &= \mathbb{E}\left\{\mathbb{E}[g(W,\theta_{o1})|z]\mathbb{E}\left[\frac{S}{P(z,\theta_{o2})} \cdot u(S,z;\theta_{o2})'|z\right]\right\} \\ &= 0. \end{aligned}$$

(b) By LIE and exogeneity,

$$\begin{aligned} D_{12} &= \mathbb{E}\left\{\nabla_{\theta_2} \left.\frac{S}{P(Z,\theta_2)}\right|_{\theta_2=\theta_{o2}} g(Y,X;\theta_{o1})\right\} \\ &= \mathbb{E}\left\{\mathbb{E}\left[g(Y,X;\theta_{o1})|z\right] \nabla_{\theta_2}\left[\left.\frac{S}{P(Z,\theta_2)}\right|_{\theta_2=\theta_{o2}}\right]\right\} \\ &= 0. \end{aligned}$$

□

Essay 2

Robustness, Redundancy, and Validity of Copulas in Likelihood Models

2.1 Introduction

In multivariate economic models, one is often ready to assume marginal distributions but is reluctant to impose a joint distribution. For example, in a panel setting, economists often use a specific likelihood for each cross section separately (e.g., PROBIT or LOGIT) but avoid modelling the joint distribution of the cross-sections over time. Similarly, in selectivity models, it is often desired to allow for unrestricted dependence between the disturbances in the primary and the selection models, each of which has a well-defined likelihood.

The usual way to handle the indeterminacy of the joint distribution is to assume independence of the marginal distributions and employ quasi-MLE or to assume joint normality and employ pseudo-MLE (e.g. White, 1982; Gourieroux et al., 1984). In certain cases these approaches result in a consistent estimation while a "sandwich" covariance matrix may be used for valid inference.

However, these approaches suffer from major weaknesses. First, there are important cases when using a pseudo-likelihood does not result in consistent estimates. Green (2002, Section 17.9) and Wooldridge (2002, Chapter 13) discuss such cases. Second, as we show below, there are estimators that dominate traditional QMLE under non-independence.

The copula approach used here allows to replace normality or independence with an alternative assumption about the joint distribution. Clearly such a replacement is only warranted if the new distribution possesses some useful properties such as ease of computation, robustness to misspecification, and improved efficiency. Arguably, copulas (or at least some of their families) may have such properties in certain econometric models. The copula approach also incorporates multivariate normality and independence as special cases.

The copula approach is relatively new to econometrics. A note by Lee (1983) appears to be the earliest application of this approach in econometrics. Copulas have recently received a lot of attention in finance literature. They are used to model dependence in financial time series (e.g., Patton, 2001; Breymann et al., 2003) and in risk management applications (e.g., Embrechts et al., 2003, 2002). Bouyé et al. (2000) provide an extensive discussion of prospects for copula in finance. Use of copula in other subfields of econometrics still appears rather limited.

Smith (2003) incorporates a copula in selectivity models and provides applications to labor supply and duration of hospitalization; Cameron et al. (2004) use a copula to develop a bivariate count data model with an application to the number of doctor visits.

We start by presenting some basics on copulas. This is done in Section 2.2. Section 2.3 introduces the GMM representation of the likelihood-based models used in the sequel. We show that imposing a joint distribution amounts to adding moment conditions.

Imposing moment conditions makes consistency of the resultant estimator conditional on the moment validity. Moreover, there are infinitely many alternative multivariate distributions that can be used. Section 2.4 shows that estimation of means remains robust against copula misspecification as long as the used copula and the true joint density share a symmetry property. A simple simulation employs most commonly used copula families to study their robustness properties.

It is well known that additional moment conditions cannot reduce asymptotic efficiency if properly used. However, sometimes the additional moments do not help even if properly used, i.e. are redundant in the sense of Breusch et al. (1999). In Section 2.5 we develop conditions for such redundancy.

Section 2.6 proposes tests of copula validity that can help deciding on the copula. Section 2.7 concludes.

2.2 Preliminaries

Definition 2.2.1 *(Nelsen, 1999, p.40) An M-dimensional* ***copula*** *is a function* $C : [0,1]^M \to [0,1]$ *that has the following properties:*

i. $C(u_1, \ldots, u_{m-1}, 0, u_{m+1}, \ldots, u_M) = 0$, $m = 2, \ldots, M-1$.

ii. $C(1, \ldots, 1, u_m, 1, \ldots, 1) = u_m$, $m = 1, \ldots, M$.

iii. *C is M-increasing: for every M-box* $B = [a_1, b_1] \times [a_2, b_2] \times \ldots \times [a_M, b_M]$, *whose* 2^M *vertices* $(c_1, \ldots, c_M)$ *are in* $[0,1]^M$, *the C-volume of B, defined by*

$$V_C(B) \equiv \sum_{i_1=1}^{2} \ldots \sum_{i_M=1}^{2} (-1)^{i_1+\ldots+i_M} C(c_{1i_1}, \ldots, c_{Mi_M}),$$

where $c_{j1} = a_j$ *and* $c_{j2} = b_j$ *for all* $j \in \{1, \ldots, M\}$, *satisfies*

$$V_C(B) \geq 0.$$

Property (iii) implies for $M = 2$ that $C(a_1, a_2) - C(a_1, b_2) - C(b_1, a_2) + C(b_1, b_2) \geq 0$ for any vectors $(a_1, a_2), (b_1, b_2) \in [0,1]^2$ such that $a_m \leq b_m$, $m = 1, 2$, i.e. $C(a, b)$ is non-decreasing in (a, b).

It follows from the definition that an M-dimensional copula C is an M-dimensional cdf whose M marginals are uniform on $[0,1]$. One may also note that for any M-dimensional copula C, $M \geq 3$, each m-marginal of C, $2 \leq m < M$, is an m-dimensional copula.

The following well-known theorem establishes existence of such a function for any joint distribution function of random variables. We restate it without proof.

Theorem 2.2.1 *(Sklar, 1959, p.229-230) Let H be an M-dimensional distribution function with marginals $F_1, \ldots, F_M$. Then there exists an M-dimensional copula C such that for all $x_m \in \mathbb{R}$, $m = 1, \ldots, M$*

$$H(x_1, \ldots, x_M) = C(F_1(x_1), \ldots, F_M(x_M)). \tag{2.1}$$

If $F_1, \ldots, F_M$ are continuous, then C is unique. Conversely, if C is an M-dimensional copula and $F_1, \ldots, F_M$ are distribution functions, then the function H in (2.1) is an M-dimensional distribution function with marginals $F_1, \ldots, F_M$.

Thus, a copula is a multivariate distribution function that connects two or more marginal distributions to exactly form the joint distribution. A copula thus completely parameterizes the entire dependence structure between two or more random variables. It is important to note that a given joint distribution function H defines a unique set of marginal distribution functions F_m, $m = 1, \ldots, M$, whereas given marginal distributions do not determine a unique joint distribution (and the implied copula).

To connect copulas to likelihood-based models, let h and c be the derivatives of the distribution functions H and C, respectively; let f_m be the derivatives of the

marginal distribution functions F_m, $m = 1, \ldots, M$. Then,

$$\begin{aligned} h(x_1, \ldots, x_M) &= \frac{\partial^M H(x_1, \ldots, x_M)}{\partial x_1 \ldots \partial x_M} \\ &= \frac{\partial^M C(F_1(x_1), \ldots, F_M(x_M))}{\partial x_1 \ldots \partial x_M} \\ &= \left. \frac{\partial^M C(u_1, \ldots, u_M)}{\partial u_1 \ldots \partial u_M} \right|_{u_m = F_m(x_m), m=1,\ldots,M} \prod_{m=1}^{M} f_m(x_m) \\ &= c(F_1(x_1), \ldots, F_M(x_M)) \prod_{m=1}^{M} f_m(x_m), \end{aligned}$$

i.e., the joint density is the product of the copula density and the marginal densities.

In what follows we restrict our attention to the bivariate case ($M = 2$). We let the marginal densities f_1 and f_2 be functions of an unknown parameter vector $\theta \in \mathbb{R}^p$ and the copula density c and the joint density h be functions of an additional parameter vector $\rho \in \mathbb{R}^q$. Then

$$\ln h(x_1, x_2; \theta, \rho) = \ln c(F_1(x_1; \theta), F_2(x_2; \theta); \rho) + \ln f_1(x_1; \theta) + \ln f_2(x_2; \theta). \quad (2.2)$$

Note that ρ parameterizes the entire dependence between the two random variables. See Appendix A for selected copula families used in this paper.

For our discussion of copula misspecification, we let K denote some copula other than the true copula C and we let k denote the corresponding copula density function.

2.3 The GMM representation

MLE assumes a complete and correctly specified joint likelihood in (2.2). For the purposes of this paper, quasi-MLE (QMLE) assumes correctly specified marginal distributions and maintains their independence and thus only uses the last two terms in (2.2). In panel settings, what we call QMLE is often referred to as the *partial* likelihood method (see Wooldridge, 2002, Section 13.8). Pseudo-MLE (PMLE) assumes an incorrect joint distribution and thus uses an incorrectly specified copula term in (2.2). The (correct) copula term in (2.2) is therefore what distinguishes MLE from QMLE (and PMLE).

It is well known that likelihood-based models can be represented as GMM models based on likelihood equations (see Godambe, 1960, 1976). The expected value of the score function for the correctly specified joint log-likelihood (2.2) is zero at the true value of parameters. Furthermore, if the marginal densities are correctly specified, the same is true for the marginal log-likelihoods.

Let θ_o and ρ_o denote the true values of the parameters θ and ρ, respectively. Assume that the following four moment conditions hold if and only if $\theta = \theta_o$ and $\rho = \rho_o$:

$$\begin{aligned}
&\mathbb{E}\tfrac{\partial}{\partial\theta}\ln f_1(X_1;\theta_o) = 0, && \text{(A)}\\
&\mathbb{E}\tfrac{\partial}{\partial\theta}\ln f_2(X_2;\theta_o) = 0, && \text{(B)}\\
&\mathbb{E}\tfrac{\partial}{\partial\theta}\ln c(F_1(X_1;\theta_o), F_2(X_2;\theta_o);\rho_o) = 0, && \text{(C)}\\
&\mathbb{E}\tfrac{\partial}{\partial\rho}\ln c(F_1(X_1;\theta_o), F_2(X_2;\theta_o);\rho_o) = 0. && \text{(D)}
\end{aligned} \tag{2.3}$$

We call moment conditions (A) and (B) the "marginal moments" and (C) and (D) the "true copula moments". Note that as stated in (2.3), the GMM problem is overidentified: it involves $p+q$ parameters and $3p+q$ moment conditions.

Here we will assume that the marginal distributions are correctly specified but the copula function may not be. If the copula is incorrectly specified, then copula moments (C-D) may not hold at (θ_o, ρ_o). They may however hold at (θ_o, ρ_o^k), where $\rho_o^k \neq \rho_o$. If they do we will say that the copula is "robust". In this case, we will replace (C) and (D) in (2.3) with

$$\begin{aligned} &\mathbb{E}\tfrac{\partial}{\partial\theta} \ln k(F_1(X_1;\theta_o), F_2(X_2;\theta_o); \rho_o^k) = 0, \qquad (\mathrm{C}') \\ &\mathbb{E}\tfrac{\partial}{\partial\rho} \ln k(F_1(X_1;\theta_o), F_2(X_2;\theta_o); \rho_o^k) = 0. \qquad (\mathrm{D}') \end{aligned} \tag{2.4}$$

We call (C′-D′) the "misspecified copula moments". For the sense in which a parametric model for a distribution is correctly specified see, for example, Wooldridge (1994, p. 2672).

Our primary focus is estimation of θ_o. GMM is an appropriate framework for our analysis because it allows studying robustness and efficiency of various likelihood-based estimators of θ_o (MLE, QMLE, PMLE) by considering misspecification and redundancy of copula moments.

Specifically, consider MLE versus QMLE in terms of efficiency. The MLE procedure for (θ_o, ρ_o) maximizes the joint likelihood in (2.2). This is equivalent to the optimal GMM estimation based on the expectation of the score of the joint likelihood, i.e.,

$$\begin{bmatrix} \mathbb{E}\tfrac{\partial}{\partial\theta} \{\ln f_1(X_1;\theta_o) + \ln f_2(X_2;\theta_o) + \ln c(F_1(X_1;\theta_o), F_2(X_2;\theta_o); \rho_o)\} \\ \mathbb{E}\tfrac{\partial}{\partial\rho} \ln c(F_1(X_1;\theta_o), F_2(X_2;\theta_o); \rho_o) \end{bmatrix} = 0. \tag{2.5}$$

We show in Section 2.5.1 that this is equivalent to the optimal GMM based on (2.3).

At the same time, the QMLE procedure for θ_o maximizes the joint likelihood assuming independence of marginal distributions. This is equivalent to the optimal GMM based on

$$\mathbb{E}\frac{\partial}{\partial\theta}\ln f_1(X_1;\theta_o)+\mathbb{E}\frac{\partial}{\partial\theta}\ln f_2(X_2;\theta_o)=0. \tag{2.6}$$

We will show in Section 2.5.1 that the optimal GMM based on (2.6) is no more efficient than the optimal GMM estimator based on

$$\begin{bmatrix} \mathbb{E}\frac{\partial}{\partial\theta}\ln f_1(X_1;\theta_o) \\ \mathbb{E}\frac{\partial}{\partial\theta}\ln f_2(X_2;\theta_o) \end{bmatrix}=0, \tag{2.7}$$

which we will call the Improved QMLE (IQMLE).

Thus MLE and (I)QMLE are equally efficient only if the extra copula moments in (2.3) do not help improve efficiency of estimation of θ_o. In GMM literature this is known as partial redundancy of copula moment conditions given the marginal moment conditions in estimation of θ_o (see Breusch et al., 1999, Section 4).

Similarly, the PMLE procedure for (θ_o,ρ_o^k) maximizes the joint likelihood in (2.2) with a misspecified copula. This is equivalent to GMM based on

$$\begin{bmatrix} \mathbb{E}\frac{\partial}{\partial\theta}\left\{\ln f_1(X_1;\theta_o)+\ln f_2(X_2;\theta_o)+\ln k(F_1(X_1;\theta_o),F_2(X_2;\theta_o);\rho_o^k)\right\} \\ \mathbb{E}\frac{\partial}{\partial\rho}\ln k(F_1(X_1;\theta_o),F_2(X_2;\theta_o);\rho_o^k) \end{bmatrix}=0. \tag{2.8}$$

Since we assume correct specification of the marginal distributions, the moment conditions in (2.8) do not hold if and only if the moment conditions in (2.4) do not hold, i.e. if and only if the copula moments are not robust to misspecification.

Finally, for robust misspecified copula moments, we will show in Section 2.5.2 that PMLE is dominated by the optimal GMM estimator using (2.3A-B)-(2.4C′-D′), which we will call the Improved PMLE (IPMLE). Thus the question of relative efficiency of IQMLE versus IPMLE for θ_o is that of partial redundancy of the misspecified copula moments.

2.4 Robustness of copula terms

Redundancy applies to valid moment conditions. We therefore first discuss robustness of copula terms to misspecification. We seek to characterize an incorrect copula K, for which the copula moments in (2.4) hold in the population.

2.4.1 A theoretical result

Let X_1 and X_2 be random variables with joint distribution function H, marginal distribution functions F_1 and F_2, respectively, and copula C. Let (μ_1, μ_2) be a point in $\mathbb{R}^2$.

Definition 2.4.1 (X_1, X_2) *is* ***radially symmetric (RS) about*** (μ_1, μ_2) *if*

$$H(\mu_1 + x_1, \mu_2 + x_2) = 1 - F_1(\mu_1 - x_1) - F_2(\mu_2 - x_2) + H(\mu_1 - x_1, \mu_2 - x_2), \quad (2.9)$$

for all (x_1, x_2) *in* $\{\mathbb{R}^2 \bigcup \{\pm\infty\}\}$.

Essentially, RS requires that any two points equally distant from (μ_1, μ_2) that lie on the same line identify tail segments under the joint density function that have equal volume. It is clear from (2.9) that the true joint density $h(x_1, x_2)$ satisfies $h(\mu_1 + x_1, \mu_2 + x_2) = h(\mu_1 - x_1, \mu_2 - x_2)$ under RS. Moreover, if x_1 or x_2 in (2.9) is taken to be equal ∞, it follows that $F_i(\mu_i + x_i) = 1 - F_i(\mu_i - x_i)$, or $Prob(X_i - \mu_i \leq x_i) = Prob(\mu_i - X_i \leq x_i)$, i.e. X_1 and X_2 are **marginally symmetric (MS) about** (μ_1, μ_2). RS is therefore a stronger symmetry concept than the usual (univariate) symmetry of random variables. It is however weaker than joint symmetry, which holds when $h(\mu_1+x_1, \mu_2+x_2) = h(\mu_1+x_1, \mu_2-x_2) = h(\mu_1 - x_1, \mu_2 + x_2) = h(\mu_1 - x_1, \mu_2 - x_2)$ (see Nelsen, 1993, for details). Many commonly used distributions are RS. For example, bivariate Normal, bivariate Student-t, bivariate Cauchy and other elliptically contoured distributions are RS. For a discussion of the elliptically contoured family of distributions, see Mardia et al. (1979, Section 2.7.2).

Now consider some copula $K \neq C$.

Definition 2.4.2 *A copula K is* ***radially symmetric (RS)*** *if*

$$K(1-u, 1-v) = 1 - u - v + K(u, v), \text{ for all } (u, v) \text{ in } \mathbb{I}^2. \tag{2.10}$$

Radial symmetry of copulas requires of the copula function what radial symmetry of random variables requires of the joint density function. Eq.(2.10) suggests that for the rectangles $[0, u] \times [0, v]$ and $[1-u, 1] \times [1-v, 1]$, the volume under the copula density function is the same for any (u, v).

It can be shown that (marginally) symmetric random variables X_1 and X_2 are radially symmetric if and only if C satisfies (2.10)(see Nelsen, 1999, p.33). So if (X_1, X_2) is RS then (2.10) holds for the true copula C. However, (2.10) may hold for many other RS copulas.

It is sometimes easier to verify radial symmetry of a copula function K by checking whether the copula density k satisfies the equation $k(1-v, 1-u) = k(v, u), \forall u, v$. For example, for FGM family it is easier to verify that the density function satisfies this condition than to verify that the copula function satisfies (2.10). In contrast, for other families in Appendix A it is easier to check (2.10). Using one of the methods, one can establish that the independence, FGM, Normal, Plackett, and Frank families are RS, while the Logistic, AMH, Joe, Clayton and Gumbel families are not. Interestingly, Frank (1979) shows that the only Archimedean copula family (see Appendix A for the definition) that satisfies (2.10) is the Frank family. Joe, AMH, Clayton and Gumbel are all Archimedean copulas that are not RS.

Theorem 2.4.1 *If (X_1, X_2) are RS about (μ_1, μ_2) then*

$$\mathbb{E}\frac{\partial}{\partial \mu} \ln k(F_1(\mu_1 + X_1), F_2(\mu_2 + X_2), \rho) = 0, \forall \rho \in \mathbb{R}^q,$$

where k is any RS copula density.

Proof: See Appendix B for all proofs that are not given in the main text.

By Theorem 2.4.1, the misspecified copula moment condition in (C′) can be used to consistently estimate the symmetry point (μ_1, μ_2) as long as the copula function and the true joint density share the property of radial symmetry.

Note that the theorem does not state anything about moment condition (D′) in (2.4) and the true copula dependence parameter ρ. Generally, under the regularity conditions, (D′) will hold in the population for some value of (θ, ρ) but not necessarily for (θ_o, ρ_o). However, (C′) holds under the conditions of the theorem no matter what value of ρ is used in (C′).

2.4.2 An illustrative simulation

To illustrate the result of the theorem and to study the behavior of both the misspecified copula moments (C′) and (D′) in finite samples, this section presents results of a simple simulation concerning a sample mean problem.

For copula K, define the sample analogues of misspecified copula moments (C′) and (D′) in (2.4)

$$\bar{\delta}^{\theta}(\theta, \rho) \equiv T^{-1} \sum_{t=1}^{T} \frac{\partial}{\partial \theta} \ln k(F_1(X_{1t}; \theta), F_2(X_{2t}; \theta); \rho) \tag{2.11}$$

and

$$\bar{\delta}^{\rho}(\theta, \rho) \equiv T^{-1} \sum_{t=1}^{T} \frac{\partial}{\partial \rho} \ln k(F_1(X_{1t}; \theta), F_2(X_{2t}; \theta); \rho). \tag{2.12}$$

Clearly, if $K = C$ then $\bar{\delta}(\theta_o, \rho_o) \rightarrow_p 0$ since (C) and (D) hold in population. Moreover, by WLLN, for any misspecified copula for which (2.4) holds, $\bar{\delta}(\theta_o, \rho_o^k) \rightarrow_p 0$. However, for non-robust copulas, the probability limit may be non-zero.

In order to be able to compare copulas we define a common measure of dependence. There are very many such measures (see Nelsen, 1999, Section 5). We pick one

that has a simple copula representation.

Definition 2.4.3 *For any two continuous random variables U and V whose copula is K,* ***Kendall's τ measure of concordance*** *is given by*

$$\tau = 4 \int \int_{\mathbb{I}^2} K(u, v; \rho) dK(u, v; \rho) - 1. \tag{2.13}$$

It follows from (2.13) that

$$\tau = 4 \int \int_{\mathbb{I}^2} K(u, v; \rho) k(u, v; \rho) du dv - 1 = 4\mathbb{E}K(U, V; \rho) - 1. \tag{2.14}$$

For two random variables, Kendall's τ can be viewed as the probability that "large" ("small") values of one are associated with "large" ("small") values of the other (the probability of concordance) minus the probability that "large" ("small") values of one are associated with "small" ("large") values of the other (the probability of discordance). Importantly, various copulas cover unequal ranges of dependence as measured by Kendall's τ (see Appendix A). We therefore control for τ in all one-parameter copulas.

In the simulation, we use the fact (see, e.g., Kendall, 1949) that for the Normal copula with Normal margins, Pearson's correlation coefficient ρ is related to τ:

$$\rho = \sin \frac{\pi}{2} \tau. \tag{2.15}$$

This allows us to derive the value of Kendall's τ that corresponds to the true value of Pearson's correlation coefficient ρ employed in simulating the joint Normal distribution.

We employ the following procedure:

Step 1. Generate T realizations of $\begin{pmatrix} X_1 \\ X_2 \end{pmatrix} \sim N\left(\begin{bmatrix} m \\ m \end{bmatrix}, \begin{bmatrix} 1 & r \\ r & 1 \end{bmatrix}\right)$ by

- generating $Z \sim N\left(\begin{bmatrix} 0 \\ 0 \end{bmatrix}, \begin{bmatrix} 1 & 0 \\ 0 & 1 \end{bmatrix}\right)$;
- using the Cholesky decomposition

$$\begin{pmatrix} X_1 \\ X_2 \end{pmatrix} = m + \begin{bmatrix} 1 & 0 \\ r & \sqrt{1-r^2} \end{bmatrix} Z;$$

Step 2. For each realization t, calculate

- $u_{it}(\mu) = \Phi(X_{it} - \mu), i = 1, 2$, where $\Phi(\cdot)$ is the Standard Normal c.d.f.;
- $k_t(\mu, \rho) \equiv k(u_{1t}(\mu), u_{2t}(\mu); \rho)$;
- $\delta_t^{\mu}(\mu, \rho) \equiv \frac{\partial}{\partial \mu} \ln k_t(\mu, \rho)$ and $\delta_t^{\rho}(\mu, \rho) \equiv \frac{\partial}{\partial \rho} \ln k_t(\mu, \rho)$;

Step 3. Calculate sample averages

$$\bar{\delta}^{\cdot}(\mu, \rho) \equiv \sum_{t=1}^{T} \delta_t^{\cdot}(\mu, \rho).$$

Step 4. Plot the resultant functions $\bar{\delta}^{\mu}(\mu, \rho)$ and $\bar{\delta}^{\rho}(\mu, \rho)$ over a relevant range of μ and ρ.

Step 5. Evaluate the sample means $\bar{\delta}^{\mu}$ and $\bar{\delta}^{\rho}$ and the sample standard errors

$$se(\bar{\delta}^{\mu}) = s^{\mu}/\sqrt{T} se(\bar{\delta}^{\rho}) = s^{\rho}/\sqrt{T}$$

Table 2.1: The true values for Kendall's τ and ρ used in simulation

Copula	ρ_o^k	τ_o
Independence	–	0
Logistic	–	1/3
Farlie-Gumbel-Morgenstern (FGM)	0.872880	0.193973
Joe	1.426845	0.193973
Ali-Mikhail-Haq (AMH)	0.697058	0.193973
Clayton	0.481321	0.193973
Gumbel	1.240654	0.193973
Frank	1.801160	0.193973
Normal with Normal margins	0.3	0.193973

at the true parameter values $\mu = m_o$ and $\rho = \rho_o^k$, where

$$s^{\cdot} = \sqrt{\frac{\sum_{t=1}^{T}(\delta_t^{\cdot}(m_o, \rho_o^k) - \bar{\delta}^{\cdot}(m_o, \rho_o^k))^2}{T-1}}.$$

The true parameter values in Step 1 are $m_o = 0$ and $r_o = 0.3$. We use (2.15) to calculate the true τ and then we use (2.14) to derive the value of ρ corresponding to the true value of τ for each copula. We consider the independence, Logistic, Farlie-Gumbel-Morgenstern, Joe, Ali-Mikhail-Haq, Clayton, Gumbel, Frank and Normal copulas. For some of these copulas it is possible to obtain an analytical solution for ρ in terms of τ using (2.14) (see Appendix A), otherwise we use numerical methods to approximate the true value of ρ with desired accuracy. Note that the independence, Farlie-Gumbel-Morgenstern, Frank and Normal families are radially symmetric.

Table 2.1 contains the true values of τ and ρ for the considered families of copulas. We choose $r_o = 0.3$ because it corresponds to a value of τ within the coverage of all the one-parameter copula families we consider. Note that the two no-parameter

copulas, independence and Logistic, imply dependence measures that are different from the true.

Figures 2.1 through 2.8 of Appendix C contain the plots of $\bar{\delta}^{\mu}(\mu, \rho)$ and $\bar{\delta}^{\rho}(\mu, \rho)$ obtained in Step 4. The sample size used for the plots is 200. According to Figure 2.1, the independence copula is robust: the copula term is identically zero even though the marginal terms are not independent. The copula term for the Logistic copula is zero for a value of μ around 0.33.

Figures 2.2–2.8 illustrate how the one-parameter copulas compare in terms of robustness. Note that all the surfaces appear to intersect the zero plane at around the true values of the parameters, which suggests general robustness. As we show below, however, one cannot accept the hypothesis of zero $\bar{\delta}$ for all copula families. The benchmark for comparisons is the Normal copula – Figure 2.7.

Interestingly, the sample analogue of the Normal copula moment (C) is close to zero at the true value of μ for any value of ρ and at $\rho = 0$ for any value of μ – panel (6a). The FGM, AMH and Frank families display a similar feature – panels (1a), (3a) and (7a). Clearly, when $\rho = 0$, these four families of copulas reduce to the independence copula, which is known to be robust. When $\rho \neq 0$, $\bar{\delta}^{\mu}$ is still close to zero at the true m_o. This observation suggests robustness of the FGM, AMH and Frank families. With these copulas, one can use the copula moment (C) with any assumed ρ and obtain a consistent estimate of μ. The other families do not exhibit this advantage.

Of course, the FGM and Frank families of copulas are RS. The observed robustness of these families is clearly a consequence of the theoretical result in the previous

section. However, the AMH family is not RS. Why is the AMH copula robust? To answer this question, write the AMH copula as an infinite sum of a geometric sequence

$$\frac{uv}{1-\rho(1-u)(1-v)} = uv\sum_{k=0}^{\infty}[\rho(1-u)(1-v)]^k. \tag{2.16}$$

The FGM copula is then the first-order approximation to the AMH family, which explains similar robustness.

To test the features illustrated on the figures, in Step 5 we calculate $\bar{\delta}^{\mu}$ and $\bar{\delta}^{\rho}$ at the true parameter values $\mu = m_o = 0$ and $\rho = \rho_o$ and evaluate standard errors for these averages. Table 2.2 shows these values along with the estimated Pearson's correlation coefficient $\hat{r}_o$ as sample size grows from 200 to 30,000. The ratio of the sample average to the standard error in parenthesis is a test statistic. Under $H_o : \hat{\delta}^{\cdot} = 0$, it is asymptotically standard Normal.

The table entries for the Logistic copula are significantly different from zero. This copula is not RS and it implies a different measure of dependence ($\tau = 1/3$). This suggests running the same simulation with common $\tau = 1/3$ for all copulas. However, this value falls outside the coverage range for several one-parameter copula families (see Appendix A), making a general comparison infeasible.

As expected, the entries for the Normal copula are insignificantly different from zero for all sample sizes. For the two RS copula families, FGM and Frank, one cannot reject the null either. The AMH family is fairly robust, too. For the Joe, Clayton and Gumbel families, the sample averages are significantly different from zero for at least one sample size which confirms the observation that these non-RS copulas are not robust in this setting.

Table 2.2: Relative robustness measures for selected copulas, their standard errors, and estimated Pearson's correlation coefficient $\hat{r}_o$ for three sample sizes

	T=200		T=3,000		T=30,000	
	$\bar{\delta}^{\mu}(m_o, \rho_o)$	$\bar{\delta}^{\rho}(m_o, \rho_o)$	$\bar{\delta}^{\mu}(m_o, \rho_o)$	$\bar{\delta}^{\rho}(m_o, \rho_o)$	$\bar{\delta}^{\mu}(m_o, \rho_o)$	$\bar{\delta}^{\rho}(m_o, \rho_o)$
Independence	0	–	0	–	0	–
Logistic	-0.15036	–	-0.17491	–	-0.16269	–
	(0.05027)		(0.01143)		(0.00365)	
Farlie-Gumbel-Morgenstern$^{\sharp}$	0.00243	-0.00271	-0.00684	-0.01072	-0.00399	-0.00403
	(0.02990)	(0.03026)	(0.00748)	(0.00747)	(0.00238)	(0.00237)
Joc	0.07855	-0.19118	0.03835	-0.10287	0.03922	-0.09700
	(0.03015)	(0.05752)	(0.00857)	(0.01450)	(0.00260)	(0.00452)
Ali-Mikhail-Haq	0.02062	-0.00678	-0.00308	-0.01755	0.00218	-0.01101
	(0.02964)	(0.04951)	(0.00709)	(0.01218)	(0.00226)	(0.00384)
Clayton	-0.00061	-0.08578	-0.02670	-0.08060	-0.01942	-0.06285
	(0.03309)	(0.06315)	(0.00780)	(0.01404)	(0.00248)	(0.00418)
Gumbel	0.04436	-0.15441	0.01174	-0.05754	0.01579	-0.04967
	(0.02682)	(0.08081)	(0.00739)	(0.02055)	(0.00223)	(0.00640)
Frank$^{\sharp}$	-0.00039	-0.00254	-0.00453	-0.00145	-0.00390	0.00063
	(0.02786)	(0.11547)	(0.00674)	(0.00297)	(0.00216)	(0.00932)
Normal$^{\sharp}$	**0.00984**	**-0.06024**	**-0.00599**	**-0.00481**	**-0.00348**	**0.00194**
	(0.02762)	(0.08817)	(0.00684)	(0.02151)	(0.00214)	(0.00666)
$\hat{r}_o$	0.3181		0.3042		0.3005	

Notes: $^{\sharp}$ denotes copulas for which $\bar{\delta}^{\mu}$ and $\bar{\delta}^{\rho}$ are insignificantly different from zero at the 5% level for every sample size.

Among the one-parameter copula families, several entries in the table stand out. First, the Frank family sample averages are at least as close to zero as the Normal benchmark for all sample sizes. Second, the FGM family sample averages are closer to zero for $T = 200$ than the Normal family average. For the other sample sizes, sample averages for the two families are comparable. Third, the AMH family also performs well in the sense of the sample averages being insignificantly different from zero. In particular, $\bar{\delta}^{\mu}$ for this family is not significantly different from zero for all sample sizes. Finally, the Clayton family averages are close to zero for the smaller sample size but not for the larger.

In the previous section, it was noted that (D′) does not generally have to hold in the population for RS copulas. An interesting observation from Table 2.2 is that sample analogues of (D′) are insignificantly different from zero for RS copulas and significantly different from zero for others. This does not follow from Theorem 2.4.1.

2.5 Redundancy of copula terms

We now turn to the question of redundancy of copula moments. We assume that we either have the true copula moments (2.3C-D) or the robust misspecified copula moments (2.4C′-D′) that hold at the true value of θ. We would like to study conditions under which using valid copula moments (either the true or misspecified ones) does not result in efficiency gains in estimation of θ.

2.5.1 Redundancy with correct copula

We first prove a lemma that reveals the structure of the variance and derivative matrices of the moment functions in (2.3). Recall that correct specification of the copula is assumed in (2.3).

Lemma 2.5.1 *Denote the covariance matrix of the moment functions in (2.3) by* $\mathbf{C}$*, their expected derivative matrix with respect to* (θ, ρ) *by* $\mathbf{D}$*. Then,*

$$\mathbf{C} = \left[\begin{array}{cc|cc} \mathbf{A} & \mathbf{G} & -\mathbf{G} & \mathbf{0} \\ \mathbf{G}' & \mathbf{B} & -\mathbf{G}' & \mathbf{0} \\ \hline -\mathbf{G}' & -\mathbf{G} & \mathbf{J} & \mathbf{E} \\ \mathbf{0} & \mathbf{0} & \mathbf{E}' & \mathbf{F} \end{array}\right] \tag{2.17}$$

and

$$\mathbf{D} = \left[\begin{array}{c|c} -\mathbf{A} & \mathbf{0} \\ -\mathbf{B} & \mathbf{0} \\ \hline \mathbf{G}+\mathbf{G}'-\mathbf{J} & -\mathbf{E} \\ -\mathbf{E}' & -\mathbf{F} \end{array}\right], \tag{2.18}$$

where $\mathbf{A}, \mathbf{B}, \mathbf{E}, \mathbf{F}, \mathbf{G}, \mathbf{J}$ *are matrix-functions of* (θ, ρ) *defined in Appendix B.*

Several important observations immediately follow from the lemma. First, (A) and (B) are uncorrelated with (C) if and only if (A) and (B) are uncorrelated with each other ($\mathbf{G} = 0$). Second, the optimal GMM based on (2.3) is identical to the ML estimation in (2.5), as claimed in Section 2.3. To see this explicitly, note that the optimal GMM on (2.3) does not change if (2.3) is pre-multiplied by a matrix

$\mathbf{W}$ such that $\mathbf{W} = \mathbf{D}'\mathbf{C}^{-1}$, if $\mathbf{C}$ is nonsingular. But, by Lemma 2.5.1,

$$\mathbf{W} = \mathbf{D}'\mathbf{C}^{-1} = -\begin{bmatrix} \mathbb{I} & \mathbb{I} & \mathbb{I} & \mathbf{0} \\ \mathbf{0} & \mathbf{0} & \mathbf{0} & \mathbb{I} \end{bmatrix} \mathbf{C}\mathbf{C}^{-1} = -\begin{bmatrix} \mathbb{I} & \mathbb{I} & \mathbb{I} & \mathbf{0} \\ \mathbf{0} & \mathbf{0} & \mathbf{0} & \mathbb{I} \end{bmatrix},$$

where $\mathbb{I}$ denotes the identity matrix of the relevant dimension. Clearly, this reproduces the MLE first order conditions (2.5). Not surprisingly, estimators that use the same first order conditions yield the same asymptotic variance matrices. In particular, for non-singular $\mathbf{C}$, the asymptotic variance matrix of the optimal GMM estimator of (θ, ρ) based on (2.3) can be written as

$$\mathbb{V}_{\text{GMM}} = (\mathbf{D}'\mathbf{C}^{-1}\mathbf{D})^{-1}. \tag{2.19}$$

(We use the standard notation according to which "$\mathbb{V}$ is the asymptotic variance of an estimator $\hat{\theta}$" means that "$\sqrt{N}(\hat{\theta} - \theta_o)$ converges in distribution to $N(\mathbf{0}, \mathbf{V})$." It is implicit that $\mathbf{D}$ and $\mathbf{C}$ in the asymptotic variance formulas are evaluated at the true values θ_o and ρ_o.) By Lemma 2.5.1, this is identical to the asymptotic variance matrix of the MLE estimator of (θ, ρ)

$$\mathbb{V}_{\text{MLE}} = -\left(\begin{bmatrix} \mathbb{I} & \mathbb{I} & \mathbb{I} & \mathbf{0} \\ \mathbf{0} & \mathbf{0} & \mathbf{0} & \mathbb{I} \end{bmatrix} \mathbf{D}\right)^{-1} = \left(\begin{bmatrix} \mathbb{I} & \mathbb{I} & \mathbb{I} & \mathbf{0} \\ \mathbf{0} & \mathbf{0} & \mathbf{0} & \mathbb{I} \end{bmatrix} \mathbf{C} \begin{bmatrix} \mathbb{I} & \mathbf{0} \\ \mathbb{I} & \mathbf{0} \\ \mathbb{I} & \mathbf{0} \\ \mathbf{0} & \mathbb{I} \end{bmatrix}\right)^{-1}. \tag{2.20}$$

In contrast to $\mathbb{V}_{\text{GMM}}$, $\mathbb{V}_{\text{MLE}}$ is defined even if $\mathbf{C}$ is singular. In fact the last representation in (2.20) involves the outer-product-of-the-score form of the information matrix, while the one before the last involves the expected-Hessian form of the information matrix. Both are non-singular under regularity conditions.

By a similar argument, it follows from Lemma 2.5.1 that the marginal moments (2.7) are *not* equivalent to the QMLE first order conditions (2.6). To see this explicitly, partition $\mathbf{C}$ and $\mathbf{D}$ as follows:

$$\mathbf{C} = \begin{bmatrix} \mathbf{C_{11}} & \mathbf{C_{12}} \\ \mathbf{C_{21}} & \mathbf{C_{22}} \end{bmatrix}, \quad \mathbf{D} = \begin{bmatrix} \mathbf{D_{11}} & \mathbf{0} \\ \mathbf{D_{21}} & \mathbf{D_{22}} \end{bmatrix}, \tag{2.21}$$

where $\mathbf{C_{11}}, \mathbf{C_{12}}, \mathbf{C_{21}}, \mathbf{C_{22}}, \mathbf{D_{11}}, \mathbf{D_{21}}, \mathbf{D_{22}}$ correspond to the blocks separated by the dotted lines in (2.17-2.18). The optimal GMM based on (2.7) does not change if the moment conditions (2.7) are pre-multiplied by a matrix $\mathbf{W_{11}}$ such that $\mathbf{W_{11}} = \mathbf{D_{11}}'\mathbf{C_{11}^{-1}}$, if $\mathbf{C_{11}}$ is nonsingular. Now, using Lemma 2.5.1,

$$\mathbf{W_{11}} = \mathbf{D_{11}}'\mathbf{C_{11}^{-1}} = -\begin{bmatrix} \mathbb{I} & \mathbb{I} \end{bmatrix} - \begin{bmatrix} -\mathbf{G}' & -\mathbf{G} \end{bmatrix} \mathbf{C_{11}}^{-1}.$$

The last term is what distinguishes the optimal GMM based on the stacked marginal moments (2.7) from summation (2.6) employed by QMLE. Call the GMM estimator based on (2.7), the *Improved QML* estimator (IQMLE).

Schmidt (2004) shows that correlation between marginal scores used in the optimal weighting matrix results in efficiency gains over summation and that there are interesting cases when the two estimation methods are equally efficient. A trivial such case is when there is no correlation between the marginal scores, i.e. $\mathbf{G} = 0$. We provide a formal statement and a proof of this relative efficiency result in the following theorem. The logic of the proof will be used again when we compare PMLE and IPMLE.

Theorem 2.5.1 *(Schmidt, 2004) Let* $\mathbb{V}_{\text{IQMLE}}$ *and* $\mathbb{V}_{\text{QMLE}}$ *denote the asymptotic variance matrices of the IQMLE and QMLE of* θ_o*, respectively. Then,* $\mathbb{V}_{\text{QMLE}} -$

$\mathbb{V}_{\text{IQMLE}}$ *is positive semi-definite.*

Proof. Define $\mathbb{A} = [\mathbb{I} \quad \mathbb{I}]$. Then, (2.6) can be rewritten as (2.7) pre-multiplied by $\mathbb{A}$. Correspondingly, the variance matrix of the moment functions in (2.6) can be expressed as $\mathbb{A}\mathbf{C_{11}}\mathbb{A}'$, where $\mathbf{C_{11}}$ is the variance matrix for the moment functions in (2.7), defined in (2.21). Similarly, the expected derivative matrix for the moment conditions in (2.6) can be expressed in terms of the relevant matrix for (2.7) as $\mathbb{A}\mathbf{D_{11}}$.

Then,

$$\mathbb{V}_{\text{QMLE}} = [(\mathbb{A}\mathbf{D_{11}})'(\mathbb{A}\mathbf{C_{11}}\mathbb{A}')^{-1}(\mathbb{A}\mathbf{D_{11}})]^{-1}, \tag{2.22}$$

while

$$\mathbb{V}_{\text{IQMLE}} = [\mathbf{D_{11}}'\mathbf{C_{11}^{-1}}\mathbf{D_{11}}]^{-1}. \tag{2.23}$$

But $\mathbb{V}_{\text{QMLE}} - \mathbb{V}_{\text{IQMLE}}$ is positive semi-definite (PSD) if and only if $\mathbb{V}_{\text{IQMLE}}^{-1} - \mathbb{V}_{\text{QMLE}}^{-1} = \mathbf{D_{11}}'\mathbf{C_{11}^{-1}}\mathbf{D_{11}} - \mathbf{D_{11}}'\mathbb{A}'(\mathbb{A}\mathbf{C_{11}}\mathbb{A}')^{-1}\mathbb{A}\mathbf{D_{11}}$ is PSD. The last expression can be rewritten as $\mathbf{D_{11}}'\mathbf{C_{11}^{-1/2}}[\mathbb{I} - \mathbf{C_{11}^{1/2}}\mathbb{A}'(\mathbb{A}\mathbf{C_{11}^{1/2}}\mathbf{C_{11}^{1/2}}\mathbb{A}')^{-1}\mathbb{A}\mathbf{C_{11}^{1/2}}]\mathbf{C_{11}^{-1/2}}\mathbf{D_{11}}$. This is PSD because the matrix in brackets is the PSD projection matrix orthogonal to $\mathbf{C_{11}^{1/2}}\mathbb{A}'$. □

Conditions under which the copula moments do not help in terms of efficiency for θ can be derived by comparing $\mathbb{V}_{\text{IQMLE}}$ with the upper left $p \times p$ block of $\mathbb{V}_{\text{MLE}}$. When $\mathbf{C}$ is non-singular, the comparisons can be equivalently made to the upper left $p \times p$ block of $\mathbb{V}_{\text{GMM}}$.

Breusch et al. (1999) (henceforth, BQSW) developed a very useful toolbox for analyzing redundancy of a set of moment conditions given another set of moment

conditions. However, their analysis assumes nonsingular $\mathbf{C}$. For this reason, we do not employ their results here but compare $\mathbb{V}_{\mathrm{IQMLE}}$ with the relevant block of $\mathbb{V}_{\mathrm{MLE}}$ directly.

Theorem 2.5.2 $\mathbb{V}_{\mathrm{MLE}}$ *for* θ *and* $\mathbb{V}_{\mathrm{IQMLE}}$ *are equal if and only if*

$$\mathbf{J} - \mathbf{C}_{\mathbf{21}}^{\theta}\mathbf{C}_{\mathbf{11}}^{-\mathbf{1}}\mathbf{C}_{\mathbf{12}}^{\theta} - \mathbf{E}\mathbf{F}^{-\mathbf{1}}\mathbf{E} = \mathbf{0}, \tag{2.24}$$

where $\mathbf{C}_{\mathbf{21}}^{\theta} = \mathbf{C}_{\mathbf{12}}^{\theta\,\prime} = [-\mathbf{G}' \quad -\mathbf{G}]$.

The cumbersome expression in (2.24) has a simple interpretation in terms of singularity of $\mathbf{C}$. It states that the linear projection of moment condition (C) on moment conditions (A), (B) and (D) is uncorrelated with moment condition (C). More specifically, (2.24) can be rewritten as follows

$$\mathbb{E}\left\{\left(\frac{\partial}{\partial\theta}\ln c - \mathbf{\Omega_{21}}\mathbf{\Omega_{11}^{-1}}\begin{bmatrix}\frac{\partial}{\partial\theta}\ln f_1\\ \frac{\partial}{\partial\theta}\ln f_2\\ \frac{\partial}{\partial\rho}\ln c\end{bmatrix}\right)\frac{\partial}{\partial\theta'}\ln c\right\} = \mathbf{0},$$

where

$$\mathbf{\Omega_{21}} = [-\mathbf{G}' \quad -\mathbf{G} \quad \mathbf{E}], \quad \mathbf{\Omega_{11}} = \begin{bmatrix}\mathbf{A} & \mathbf{G} & \mathbf{0}\\ \mathbf{G}' & \mathbf{B} & \mathbf{0}\\ \mathbf{0} & \mathbf{0} & \mathbf{F}\end{bmatrix}$$

and the arguments of the moment functions have been suppressed for brevity. In other words, (C) has to be a linear combination of (A), (B) and (D) for the copula information to be redundant in terms of asymptotic efficiency of estimation of θ. Thus $\mathbf{C}$ has to be singular.

Since $\mathbb{V}_{\mathrm{MLE}} = \mathbb{V}_{\mathrm{GMM}}$ for non-singular $\mathbf{C}$, and $\mathbb{V}_{\mathrm{IQMLE}}$ is equal to $\mathbb{V}_{\mathrm{MLE}}$ for θ if and only if $\mathbf{C}$ is singular, thus equality of $\mathbb{V}_{\mathrm{IQMLE}}$ and $\mathbb{V}_{\mathrm{GMM}}$ for θ is impossible unless (C) is a linear combination of (A), (B) and (D).

Corollary 2.5.1 *If* (C) *is a linear combination of* (A) *and* (B) *with* ρ *known then*

1. $\mathbf{E} = \mathbf{0}$;
2. $\mathbf{J} - \mathbf{C}_{21}^{\theta}\mathbf{C}_{11}^{-1}\mathbf{C}_{12}^{\theta} = \mathbf{0}$;
3. *IQMLE is efficient.*

We therefore have two cases when the copula knowledge in (C) and (D) is redundant given the knowledge of the marginals in (A) and (B). One case is when the copula moment (C) is a linear combination of (A) and (B). The other case is when (C) is not a linear combination of (A) and (B) but is a linear combination of (A), (B) and (D). In both cases, $\mathbf{C}$ is singular.

Examples at the end of this section illustrate how one can apply the redundancy results in practice.

2.5.2 Redundancy with misspecified copula

Now suppose incorrect but zero-mean copula terms in (2.4C′) and (2.4D′) are used in estimation. When is such knowledge redundant in terms of efficient estimation of θ?

Lemma 2.5.2 *Denote the covariance matrix of the moment functions in (2.3) that employ the copula moments (2.4C′) and (2.4D′) instead of (2.3C) and (2.3D), respectively, by* $\mathbf{C^k}$*, their expected derivative matrix with respect to* (θ, ρ^k) *by* $\mathbf{D^k}$*. Then,*

$$\mathbf{C^k} = \left[\begin{array}{cc|cc} \mathbf{A} & \mathbf{G} & -\mathbf{K} & -\mathbf{P} \\ \mathbf{G'} & \mathbf{B} & -\mathbf{L'} & -\mathbf{Q'} \\ \hline -\mathbf{K'} & -\mathbf{L} & \mathbf{N} & \mathbf{V} \\ -\mathbf{P'} & -\mathbf{Q} & \mathbf{V'} & \mathbf{W} \end{array}\right]$$

and

$$\mathbf{D^k} = \left[\begin{array}{c|c} -\mathbf{A} & \mathbf{0} \\ -\mathbf{B} & \mathbf{0} \\ \hline \mathbf{K'} + \mathbf{L} - \mathbf{M} & -\mathbf{S} \\ -\mathbf{S'} & -\mathbf{T} \end{array}\right],$$

where $\mathbf{A}, \mathbf{B}, \mathbf{G}$ *are as in Lemma 2.5.1,* $\mathbf{K}, \mathbf{L}, \mathbf{M}, \mathbf{N}, \mathbf{P}, \mathbf{Q}, \mathbf{S}, \mathbf{T}, \mathbf{V}, \mathbf{W}$ *are matrix-functions of* (θ, ρ^k) *defined in Appendix B.*

Lemma 2.5.2 can be used to make the following observation. The optimal GMM estimator using (2.3A-B)-(2.4C′-D′) is *not* identical to the PML estimator. This is in contrast with Lemma 2.5.1, in which MLE coincided with GMM using (2.3A-D) because we had knowledge of the correct copula. More specifically, the optimal GMM estimator based on (2.3A-B)-(2.4C′-D′) is unchanged if (2.3A-B)-(2.4C′-D′) are pre-multiplied by matrix $\mathbf{W^k} = \mathbf{D^{k'}}(\mathbf{C^k})^{-1}$ if $\mathbf{C^k}$ is non-singular. Using Lemma 2.5.2, it can be shown that

$$\mathbf{D^{k'}}(\mathbf{C^k})^{-1} = -\left[\begin{array}{cccc} \mathbb{I} & \mathbb{I} & \mathbb{I} & \mathbf{0} \\ \mathbf{0} & \mathbf{0} & \mathbf{0} & \mathbb{I} \end{array}\right] + \mathbf{Z}(\mathbf{C^k})^{-1},$$

where $\mathbf{Z}$ contains $\mathbf{G}' - \mathbf{K}'$, $\mathbf{G} - \mathbf{L}$, $\mathbf{N} - \mathbf{M}'$, $\mathbf{P}'$, $\mathbf{Q}$, $\mathbf{V}' - \mathbf{S}'$, $\mathbf{W} - \mathbf{T}'$. Clearly, Lemma 2.5.2 becomes Lemma 2.5.1 if $k = c$. In this case, $\mathbf{Z} = 0$, $\mathbf{W^k} = \mathbf{W}$, the optimal weighting retrieves (2.5), and PMLE is equivalent to MLE.

For $k \neq c$, correlation patterns impossible in Lemma 2.5.1 now provide potential efficiency gains over PMLE. We call the GMM estimator using (2.3A-B)-(2.4C'-D') the *Improved PML* estimator (IPMLE).

Theorem 2.5.3 *Let* $\mathbb{V}_{\text{IPMLE}}$ *and* $\mathbb{V}_{\text{PMLE}}$ *denote the asymptotic variance matrices of the IPMLE and PMLE of* (θ_o, ρ_o^k)*, respectively. Then,* $\mathbb{V}_{\text{PMLE}} - \mathbb{V}_{\text{IPMLE}}$ *is positive semi-definite.*

Proof. Define

$$\mathbb{A} = \begin{bmatrix} \mathbb{I} & \mathbb{I} & \mathbb{I} & \mathbf{0} \\ \mathbf{0} & \mathbf{0} & \mathbf{0} & \mathbb{I} \end{bmatrix}.$$

Then, (2.8) can be rewritten as (2.3) pre-multiplied by $\mathbb{A}$. Correspondingly, the variance matrix of the moment functions in (2.8) can be expressed as $\mathbb{A}\mathbf{C}_{\mathbf{11}}^{\mathbf{k}}\mathbb{A}'$. Similarly, the expected derivative matrix for the moment conditions in (2.8) can be expressed as $\mathbb{A}\mathbf{D}^{\mathbf{k}}$.

Then,

$$\mathbb{V}_{\text{PMLE}} = [(\mathbb{A}\mathbf{D}^{\mathbf{k}})'(\mathbb{A}\mathbf{C}^{\mathbf{k}}\mathbb{A}')^{-1}(\mathbb{A}\mathbf{D}^{\mathbf{k}})]^{-1}, \tag{2.25}$$

while

$$\mathbb{V}_{\text{IPMLE}} = [\mathbf{D}^{\mathbf{k}'}(\mathbf{C}^{\mathbf{k}})^{-1}\mathbf{D}^{\mathbf{k}}]^{-1}. \tag{2.26}$$

$\mathbb{V}_{\text{PMLE}} - \mathbb{V}_{\text{IPMLE}}$ is PSD if and only if $\mathbb{V}_{\text{IPMLE}}^{-1} - \mathbb{V}_{\text{PMLE}}^{-1} = \mathbf{D^{k\prime}C^{k}}^{-1}\mathbf{D^{k}} - \mathbf{D^{k\prime}A'(AC^{k}A')}^{-1}\mathbf{AD^{k}}$ is PSD. Rewrite the last expression as

$$\mathbf{D^{k\prime}}(\mathbf{C^{k}})^{-1/2}[\mathbb{I} - (\mathbf{C^{k}})^{1/2}\mathbb{A}'(\mathbb{A}(\mathbf{C^{k}})^{1/2}(\mathbf{C^{k}})^{1/2}\mathbb{A}')^{-1}\mathbb{A}(\mathbf{C^{k}})^{1/2}](\mathbf{C^{k}})^{-1/2}\mathbf{D^{k}}.$$

This is PSD because the matrix in brackets is the PSD projection matrix orthogonal to $(\mathbf{C^{k}})^{1/2}\mathbb{A}'$. □

Clearly, (I)PMLE does not improve precision of estimation of θ over IQMLE if and only if the upper left $p \times p$ block of $\mathbb{V}_{\text{(I)PMLE}}$ is equal to $\mathbb{V}_{\text{IQMLE}}$. We focus on $\mathbb{V}_{\text{IPMLE}}$ because by Theorem 2.5.4, if IPMLE does not improve over precision of IQMLE for θ, then neither does PMLE. $\mathbb{V}_{\text{IPMLE}}$ is only defined when $\mathbf{C^{k}}$ is non-singular, thus we can apply the redundancy toolbox of BQSW.

Theorem 2.5.4 $\mathbb{V}_{\text{IPMLE}}$ *for* θ *and* $\mathbb{V}_{\text{IQMLE}}$ *are equal if and only if*

$$\mathbf{M} - \mathbf{C_{21}^{\theta k}C_{11}^{-1}C_{12}^{\theta}} - \mathbf{ST}^{-1}(\mathbf{R} - \mathbf{C_{21}^{\rho k}C_{11}^{-1}C_{12}^{\theta}}) = \mathbf{0}, \tag{2.27}$$

where $\mathbf{C_{21}^{\theta k}} = [-\mathbf{K}' \quad -\mathbf{L}]$, $\mathbf{C_{21}^{\rho k}} = [-\mathbf{P}' \quad -\mathbf{Q}]$.

In (2.27), $\mathbf{M} - \mathbf{C_{21}^{\theta k}C_{11}^{-1}C_{12}^{\theta}}$ and $\mathbf{R} - \mathbf{C_{21}^{\rho k}C_{11}^{-1}C_{12}^{\theta}}$ can be viewed as covariance matrices between copula moments (C′-D′) and the error in the linear projection of

the true copula moment (C) on the marginal moments (A-B). More explicitly,

$$\mathbf{M} - \mathbf{C}_{\mathbf{21}}^{\boldsymbol{\theta}\mathbf{k}}\mathbf{C}_{\mathbf{11}}^{-\mathbf{1}}\mathbf{C}_{\mathbf{12}}^{\boldsymbol{\theta}} = \mathbb{E}\left\{\frac{\partial}{\partial\theta}\ln k\left(\frac{\partial}{\partial\theta}\ln c - \mathbf{C}_{\mathbf{21}}\mathbf{C}_{\mathbf{11}}^{-\mathbf{1}}\begin{bmatrix}\frac{\partial}{\partial\theta}\ln f_1\\ \frac{\partial}{\partial\theta}\ln f_2\end{bmatrix}\right)'\right\}, \quad (2.28)$$

$$\mathbf{R} - \mathbf{C}_{\mathbf{21}}^{\boldsymbol{\rho}\mathbf{k}}\mathbf{C}_{\mathbf{11}}^{-\mathbf{1}}\mathbf{C}_{\mathbf{12}}^{\boldsymbol{\theta}} = \mathbb{E}\left\{\frac{\partial}{\partial\rho}\ln k\left(\frac{\partial}{\partial\theta}\ln c - \mathbf{C}_{\mathbf{21}}\mathbf{C}_{\mathbf{11}}^{-\mathbf{1}}\begin{bmatrix}\frac{\partial}{\partial\theta}\ln f_1\\ \frac{\partial}{\partial\theta}\ln f_2\end{bmatrix}\right)'\right\}.$$

Clearly, when both of these matrices are zero, (2.27) holds for any $\mathbf{S}$. Also, if only (2.28) is zero and $\mathbf{S} = \mathbf{0}$, (2.27) holds for any $\mathbf{R}$ and $\mathbf{C}_{\mathbf{21}}^{\boldsymbol{\rho}\mathbf{k}}$.

Corollary 2.5.2 *If (C) is a linear combination of (A) and (B) with ρ known then*

1. $\mathbf{M} - \mathbf{C}_{\mathbf{21}}^{\boldsymbol{\theta}\mathbf{k}}\mathbf{C}_{\mathbf{11}}^{-\mathbf{1}}\mathbf{C}_{\mathbf{12}}^{\boldsymbol{\theta}} = \mathbf{0}$;
2. $\mathbf{R} - \mathbf{C}_{\mathbf{21}}^{\boldsymbol{\rho}\mathbf{k}}\mathbf{C}_{\mathbf{11}}^{-\mathbf{1}}\mathbf{C}_{\mathbf{12}}^{\boldsymbol{\theta}} = \mathbf{0}$;
3. *IQMLE and IPMLE for θ are equally efficient.*

By the corollary, knowledge about robust but misspecified copulas is redundant in estimation of θ given (A) and (B) when the true copula moment (C) is not informative given (A) and (B).

2.5.3 Examples

The following four examples illustrate how the redundancy results can be used in practice. The first three examples show problems where the copula moment

conditions are redundant and thus IQMLE is efficient. The last example considers a situation when copula moment conditions are not redundant in general and IQMLE is generally inefficient.

Bivariate Normal with common mean. Assume Normal marginal densities with $\sigma_1^2 = \sigma_2^2 = 1$ and $\mu_1 = \mu_2 = \mu$

$$f_1(x_1;\mu) = \frac{1}{\sqrt{2\pi}} e^{-\frac{(x_1-\mu)^2}{2}},$$

$$f_2(x_2;\mu) = \frac{1}{\sqrt{2\pi}} e^{-\frac{(x_2-\mu)^2}{2}}.$$

Let the true joint density be Normal, i.e.,

$$h(x_1, x_2; \mu, \rho) = \frac{1}{2\pi\sqrt{1-\rho^2}} e^{-\frac{(x_1-\mu)^2+(x_2-\mu)^2-2\rho(x_1-\mu)(x_2-\mu)}{2(1-\rho^2)}}.$$

Then, the implied copula is the Normal copula

$$c(F_1(x_1;\mu), F_2(x_2;\mu); \rho) = \frac{1}{\sqrt{1-\rho^2}} e^{-\frac{\rho\left(\rho(x_1-\mu)^2+\rho(x_2-\mu)^2-2(x_1-\mu)(x_2-\mu)\right)}{2(1-\rho^2)}},$$

where ρ is the copula dependence parameter (Pearson's correlation coefficient). Thus we have the setup of our simulation in Section 2.4.2.

The relevant moment conditions are

$$\mathbb{E}\{X_1 - \mu\} = 0, \tag{A}$$

$$\mathbb{E}\{X_2 - \mu\} = 0, \tag{B}$$

$$\mathbb{E}\left\{-\frac{((X_1-\mu)+(X_2-\mu))\rho}{\rho+1}\right\} = 0, \tag{C}$$

$$\mathbb{E}\left\{-\frac{\rho(X_1^2+X_2^2)+\mu(1-\rho)^2(X_1+X_2)-\left(1+\rho^2\right)X_1X_2+\rho(\rho^2-1)-\mu^2(1-\rho)^2}{(\rho-1)^2(\rho+1)^2}\right\} = 0. \tag{D}$$

(C) is clearly a linear combination of (A) and (B) for known ρ. By Corollary 2.5.1, the true copula moments are redundant for estimation of μ. Furthermore, by Corollary 2.5.2, any valid misspecified copula moments do not help improve precision of estimation over IQMLE of μ. IQMLE of μ is efficient.

Section 2.4.2 provided evidence of robustness of independence, FGM, AMH and Frank copula families. None of them would allow to improve efficiency over IQMLE of μ.

Note that using the Normal moment generating function, one can show that

$$\mathbf{C} = \begin{bmatrix} 1 & \rho & -\rho & 0 \\ \rho & 1 & -\rho & 0 \\ -\rho & -\rho & \frac{2\rho^2}{1+\rho} & 0 \\ 0 & 0 & 0 & \frac{1+\rho^2}{(\rho-1)^2(\rho+1)^2} \end{bmatrix}, \quad \mathbf{D} = \begin{bmatrix} -1 & 0 \\ -1 & 0 \\ \frac{2\rho}{1+\rho} & 0 \\ 0 & -\frac{1+\rho^2}{(\rho-1)^2(\rho+1)^2} \end{bmatrix},$$

where $\det(\mathbf{C}) = 0$, and

$$\mathbb{V}_{\text{MLE}} = \mathbb{V}_{\text{IQMLE}} = \frac{1+\rho}{2}.$$

Bivariate Normal regression. Let $\mathbf{y} = \mathbf{x}\beta + \epsilon$, where $\mathbf{y} = (y_1, y_2)', \mathbf{x} = (x_1, x_2)'$. Suppose $\mathbf{x}$ is non-random. Let $\epsilon = (\epsilon_1, \epsilon_2)' \sim \mathbb{N}(\mathbf{0}, \mathbf{\Sigma})$, where

$$\mathbf{\Sigma} = \begin{bmatrix} \sigma_1^2 & \rho \\ \rho & \sigma_2^2 \end{bmatrix}$$

and σ_1^2, σ_2^2 are known but ρ is not.

Then,

$$f_1(y_1; x_1, \beta) = \frac{1}{\sqrt{2\pi\sigma_1^2}} e^{-\frac{(y_1 - x_1\beta)^2}{2\sigma_1^2}},$$

$$f_2(y_2; x_2, \beta) = \frac{1}{\sqrt{2\pi\sigma_2^2}} e^{-\frac{(y_2 - x_2\beta)^2}{2\sigma_2^2}},$$

$$h(\mathbf{y}; \mathbf{x}, \beta, \rho) = \frac{1}{2\pi\sqrt{|\mathbf{\Sigma}|}} e^{-\frac{1}{2}(\mathbf{y} - \mathbf{x}\beta)'\mathbf{\Sigma}^{-1}(\mathbf{y} - \mathbf{x}\beta)}.$$

Then, the implied copula is Normal,

$$\begin{aligned} c(F_1(y_1; x_1, \beta), F_2(y_2; x_2, \beta); \rho) &= \frac{\sqrt{\sigma_1^2\sigma_2^2}}{\sigma_1^2\sigma_2^2 - \rho^2} e^{-\frac{\epsilon_1}{2}\left(\frac{\epsilon_1\sigma_2^2 - \epsilon_2\rho}{\sigma_1^2\sigma_2^2 - \rho^2}\right) - \frac{\epsilon_2}{2}\left(\frac{\epsilon_2\sigma_1^2 - \epsilon_1\rho}{\sigma_1^2\sigma_2^2 - \rho^2}\right)} \\ &\quad \times e^{\frac{1}{2}\left(\frac{\epsilon_1^2}{2\sigma_1^2} + \frac{\epsilon_2^2}{2\sigma_2^2}\right)}, \end{aligned}$$

where $\epsilon_i = y_i - x_i\beta$, $i = 1, 2$.

The relevant moment conditions are

$$\mathbb{E}\left\{\frac{x_1\epsilon_1}{\sigma_1^2}\right\} = 0, \qquad (A)$$

$$\mathbb{E}\left\{\frac{x_2\epsilon_2}{\sigma_2^2}\right\} = 0, \qquad (B)$$

$$\mathbb{E}\left\{-\frac{\rho(\sigma_1^2\sigma_2^2x_1\epsilon_2+\sigma_1^2\sigma_2^2x_2\epsilon_1-\sigma_1^2\rho x_2\epsilon_2-\sigma_2^2\rho x_1\epsilon_1)}{\sigma_1^2\sigma_2^2(\sigma_1^2\sigma_2^2-\rho^2)}\right\} = 0, \qquad (C)$$

$$\mathbb{E}\left\{\frac{\sigma_1^2\sigma_2^2\epsilon_1\epsilon_2+\rho^2\epsilon_1\epsilon_2-\sigma_2^2\rho\epsilon_1^2-\sigma_1^2\rho\epsilon_2^2+\rho\sigma_1^2\sigma_2^2-\rho^3}{(\sigma_1^2\sigma_2^2-\rho^2)^2}\right\} = 0. \qquad (D)$$

Again, (C) is a linear combination of (A) and (B). The use of (C) and (D) or any other zero mean copula terms does not help estimate β more precisely than IQMLE.

The covariance and expected derivative matrices are

$$\mathbf{C} = \begin{bmatrix} \frac{x_1^2}{\sigma_1^2} & \frac{\rho x_1x_2}{\sigma_1^2\sigma_2^2} & -\frac{\rho x_1x_2}{\sigma_1^2\sigma_2^2} & 0 \\ \frac{\rho x_1x_2}{\sigma_1^2\sigma_2^2} & \frac{x_2^2}{\sigma_2^2} & -\frac{\rho x_1x_2}{\sigma_1^2\sigma_2^2} & 0 \\ -\frac{\rho x_1x_2}{\sigma_1^2\sigma_2^2} & -\frac{\rho x_1x_2}{\sigma_1^2\sigma_2^2} & \frac{\left(x_2^2\sigma_1^2+x_1^2\sigma_2^2-2\rho x_1x_2\right)\rho^2}{\sigma_1^2\sigma_2^2\left(\sigma_1^2\sigma_2^2-\rho^2\right)} & 0 \\ 0 & 0 & 0 & \frac{\sigma_1^2\sigma_2^2+\rho^2}{\left(\sigma_1^2\sigma_2^2-\rho^2\right)^2} \end{bmatrix},$$

$$\mathbf{D} = \begin{bmatrix} -\frac{x_1^2}{\sigma_1^2} & 0 \\ -\frac{x_2^2}{\sigma_2^2} & 0 \\ \frac{\rho\left(2x_1x_2\sigma_1^2\sigma_2^2-\rho x_1^2\sigma_2^2-\rho\sigma_1^2x_2^2\right)}{\left(\sigma_1^2\sigma_2^2-\rho^2\right)\sigma_1^2\sigma_2^2} & 0 \\ 0 & -\frac{\sigma_1^2\sigma_2^2+\rho^2}{\left(\sigma_1^2\sigma_2^2-\rho^2\right)^2} \end{bmatrix}.$$

$\mathbf{C}$ is singular. $\mathbb{V}_{\text{MLE}} = \mathbb{V}_{\text{IQMLE}} = \frac{\sigma_1^2\sigma_2^2-\rho^2}{x_2^2\sigma_1^2+x_1^2\sigma_2^2-2\rho x_1 x_2}$, which is also the variance of the GLS estimator of β, $(\mathbf{x'\Sigma^{-1}x})^{-1}$.

Bivariate Normal with common variance. Assume Normal marginal densities with $\sigma_1^2 = \sigma_2^2 = \sigma^2$ and $\mu_1 = \mu_2 = 0$

$$f_1(x_1;\sigma) = \frac{1}{\sqrt{2\pi\sigma^2}} e^{-\frac{x_1^2}{2\sigma^2}},$$

$$f_2(x_2;\sigma) = \frac{1}{\sqrt{2\pi\sigma^2}} e^{-\frac{x_2^2}{2\sigma^2}}.$$

Again, let the true joint distribution be Normal, i.e.,

$$h(x_1, x_2; \sigma, \rho) = \frac{1}{2\pi\sqrt{\sigma^4-\rho^2}} e^{-\frac{x_1^2\sigma^2-2x_1x_2\rho+x_2^2\sigma^2}{2(\sigma^4-\rho^2)}}.$$

Then, the implied copula is Normal,

$$c(F_1(x_1;\sigma), F_2(x_2;\sigma);\rho) = \frac{\sigma^2}{\sqrt{\sigma^4-\rho^2}} e^{-\frac{\rho\left(x_1^2\rho+x_2^2\rho-2\sigma^2x_1x_2\right)}{2\left(\sigma^4-\rho^2\right)\sigma^2}}.$$

The relevant moment conditions are

$$\mathbb{E}\left\{\frac{X_1{}^2-\sigma^2}{2\sigma^4}\right\} = 0, \qquad (A)$$

$$\mathbb{E}\left\{\frac{X_2{}^2-\sigma^2}{2\sigma^4}\right\} = 0, \qquad (B)$$

$$\mathbb{E}\left\{-\frac{\left(\left(3\rho\sigma^4-\rho^3\right)(X_1^2+X_2^2)-4\sigma^6X_1X_2-2\sigma^2\rho(\sigma^4-\rho^2)\right)\rho}{2\left(\sigma^2-\rho\right)^2\left(\sigma^2+\rho\right)^2\sigma^2}\right\} = 0, \qquad (C)$$

$$\mathbb{E}\left\{-\frac{\rho\sigma^2(X_1^2+X_2^2)-(\rho^2+\sigma^4)X_1X_2-\rho(\sigma^4-\rho^2)}{\left(\sigma^2+\rho\right)^2\left(\sigma^2-\rho\right)^2}\right\} = 0. \qquad (D)$$

(C) is not a linear combination of (A) and (B). However, (C) is a linear combination of (A), (B), and (D). Indeed, after some algebra, the moment function in (C) can be written as a weighted sum of the moment functions in (A), (B) and (D) with the weight

$$\frac{\rho^2\sigma^2}{(\sigma^4+\rho^2)}$$

on (A) and (B) and

$$\frac{2\rho\sigma^4}{(\rho^2+\sigma^4)}$$

on (D). Thus (2.24) holds and $\mathbf{C}$ is singular.

$$\mathbf{C} = \begin{bmatrix} \frac{1}{2\sigma^4} & \frac{\rho^2}{2\sigma^8} & -\frac{\rho^2}{2\sigma^8} & 0 \\ \frac{\rho^2}{2\sigma^8} & \frac{1}{2\sigma^4} & -\frac{\rho^2}{2\sigma^8} & 0 \\ -\frac{\rho^2}{2\sigma^8} & -\frac{\rho^2}{2\sigma^8} & \frac{\rho^2(4\sigma^8-3\sigma^4\rho^2+\rho^4)}{\sigma^8(\sigma^2-\rho)^2(\sigma^2+\rho)^2} & -\frac{2\sigma^2\rho}{(\sigma^2-\rho)^2(\sigma^2+\rho)^2} \\ 0 & 0 & -\frac{2\sigma^2\rho}{(\sigma^2-\rho)^2(\sigma^2+\rho)^2} & \frac{\sigma^4+\rho^2}{(\sigma^2-\rho)^2(\sigma^2+\rho)^2} \end{bmatrix},$$

$$\mathbf{D} = \begin{bmatrix} -\frac{1}{2\sigma^4} & 0 \\ -\frac{1}{2\sigma^4} & 0 \\ \frac{\rho^2(\rho^2-3\sigma^4)}{\sigma^4(\sigma^2-\rho)^2(\sigma^2+\rho)^2} & \frac{2\sigma^2\rho}{(\sigma^2-\rho)^2(\sigma^2+\rho)^2} \\ \frac{2\sigma^2\rho}{(\sigma^2-\rho)^2(\sigma^2+\rho)^2} & -\frac{\sigma^4+\rho^2}{(\sigma^2-\rho)^2(\sigma^2+\rho)^2} \end{bmatrix}.$$

By Theorem 2.5.2, IQMLE of σ^2 is efficient, in fact $\mathbb{V}_{\text{MLE}} = \mathbb{V}_{\text{IQMLE}} = \sigma^4 + \rho^2$.

Farlie-Gumbel-Morgenstern copula with general marginals. For $i = 1, 2$ denote the marginal p.d.f.'s and c.d.f.'s by

$$f_i \equiv f_i(x_i; \theta)$$

and

$$F_i \equiv F_i(x_i;\theta) = \int_{-\infty}^{x_i} f_i(z;\theta)dz,$$

respectively.

Assume the FGM copula. Then

$$c(u,v;\rho) = 1 + \rho - 2\rho u - 2\rho v + 4\rho uv.$$

Our moment conditions are now

$$\begin{aligned}
\mathbb{E}\left\{\frac{1}{f_1}\frac{\partial f_1}{\partial\theta}\right\} &= 0, && (A)\\
\mathbb{E}\left\{\frac{1}{f_2}\frac{\partial f_2}{\partial\theta}\right\} &= 0, && (B)\\
\mathbb{E}\left\{\frac{2\rho f_1+2\rho f_2-4\rho f_1F_2-4\rho F_2F_1}{1+\rho-2\rho F_1-2\rho F_2+4\rho F_1F_2}\right\} &= 0, && (C)\\
\mathbb{E}\left\{\frac{1-F_1-F_2+4F_1F_2}{1+\rho-2\rho F_1-2\rho F_2+4\rho F_1F_2}\right\} &= 0. && (D)
\end{aligned}$$

In general, (C) is *not* a linear combination of (A), (B) or (A), (B) and (D). So the copula based terms are not redundant in general and IQMLE is generally inefficient.

2.6 Validity of copula terms

Suppose we are ready to assume the correctness of the marginal distributions (the marginal moments in (2.3)) but are doubtful about the correctness of the joint distribution (the copula moments in (2.3)). One may test the validity of a copula

by testing the validity of the moment restrictions (C) and (D) in (2.3). There are at least two ways to do that.

2.6.1 Theoretical results

It was noted earlier that the moment conditions in (2.3) are usually overidentified. There are at least as many marginal moments as marginal parameters (or more if the marginal distributions share parameters), plus there are as many copula moments as there are parameters in total. Since the parameters are overidentified, the moment conditions in (2.3) imply restrictions. Consequently, if the model that led to the moment conditions is incorrect (i.e., the assumed joint distribution is wrong) then at least some of the moment conditions will be systematically violated in the sample. This suggests the possibility for testing copula validity by a test of the overidentifying restrictions (see, e.g., Hansen, 1982; Newey and West, 1987).

We will need more notation. For $m = 1, 2$ and $i = 1, \ldots, N$, denote $f_{mi}(\theta) = f_m(X_{1i}; \theta)$, $c_i(\theta, \rho) = c(F_1(X_{1i}; \theta), F_2(X_{2i}; \theta); \rho)$,

$$\psi_i(\theta, \rho) = \begin{bmatrix} \frac{\partial}{\partial \theta} \ln f_{1i}(\theta) \\ \frac{\partial}{\partial \theta} \ln f_{2i}(\theta) \\ \frac{\partial}{\partial \theta} \ln c_i(\theta, \rho) \\ \frac{\partial}{\partial \rho} \ln c_i(\theta, \rho) \end{bmatrix}, \quad g_i(\theta) = \begin{bmatrix} \frac{\partial}{\partial \theta} \ln f_{1i}(\theta) \\ \frac{\partial}{\partial \rho} \ln f_{2i}(\theta) \end{bmatrix},$$

$$r_i(\theta, \rho) = \begin{bmatrix} \frac{\partial}{\partial \theta} \ln c_i(\theta, \rho) \\ \frac{\partial}{\partial \rho} \ln c_i(\theta, \rho) \end{bmatrix}.$$

Note that ψ_i is a (3p+q)-vector. Let

$$\bar{\psi}(\theta,\rho) \equiv \frac{1}{N}\sum_{i=1}^{N}\psi_i(\theta,\rho), \quad \bar{g}(\theta) \equiv \frac{1}{N}\sum_{i=1}^{N}g_i(\theta), \quad \bar{r}(\theta,\rho) \equiv \frac{1}{N}\sum_{i=1}^{N}r_i(\theta,\rho).$$

Following our previous notation, let

$$\begin{aligned}
\mathbf{C_o} &\equiv \mathbb{E}\psi(\theta_o,\rho_o)\psi(\theta_o,\rho_o)', \\
\mathbf{C_{11}^o} &\equiv \mathbb{E}g(\theta_o)g(\theta_o)', \\
\mathbf{C_{22}^o} &\equiv \mathbb{E}r(\theta_o,\rho_o)r(\theta_o,\rho_o)', \\
\mathbf{C_{12}^o} = \mathbf{C_{21}^o}' &\equiv \mathbb{E}g(\theta_o)r(\theta_o,\rho_o)', \\
\mathbf{D_o} &\equiv \mathbb{E}\frac{\partial}{\partial(\theta',\rho')}\psi(\theta_o), \\
\mathbf{D_{11}^o} &\equiv \mathbb{E}\frac{\partial}{\partial\theta'}g(\theta_o), \\
\mathbf{D_{21}^o} &\equiv \mathbb{E}\frac{\partial}{\partial\theta'}r(\theta_o,\rho_o), \\
\mathbf{D_{22}^o} &\equiv \mathbb{E}\frac{\partial}{\partial\rho'}r(\theta_o,\rho_o),
\end{aligned}$$

where expectations are with respect to the joint density $h(x_1,x_2)$.

Proposition 2.6.1 *Let $(\breve{\theta},\breve{\rho})$ denote the optimal GMM estimate of (θ,ρ) based on (2.3). Then*

$$N\bar{\psi}(\breve{\theta},\breve{\rho})'\mathbf{C_o^{-1}}\bar{\psi}(\breve{\theta},\breve{\rho}) \overset{a}{\sim} \chi^2_{2p}. \tag{2.29}$$

This test is a specification test which, given that the marginal distributions are correct, should capture copula misspecification. A consistent estimator of $\mathbf{C_o}$ such as

$$\breve{\mathbf{C}}_\mathbf{o} = \frac{1}{N}\sum_{i=1}^{N}\psi_i(\breve{\theta},\breve{\rho})\psi_i(\breve{\theta},\breve{\rho})'$$

is usually used in (2.29). It is however important to note that the statistic in (2.29) can be used only if $\mathbf{C}$ in non-singular, i.e. if copula terms are not redundant.

The second way to test copula validity we propose is based on a two step procedure.

Proposition 2.6.2 *Let $\hat{\theta}$ be the optimal GMM estimate based on $\mathbb{E}g(\theta) = 0$. Let $\hat{\rho}$ be obtained by minimizing $\bar{r}(\hat{\theta}, \rho)'\mathbf{B_o^{-1}}\bar{r}(\hat{\theta}, \rho)$, where*

$$\begin{aligned}\mathbf{B_o} = \mathbf{C_{22}^o} \; & -\mathbf{D_{21}^o}(\mathbf{D_{11}^o}\mathbf{C_{11}^o}^{-1}\mathbf{D_{11}^o})^{-1}\mathbf{D_{11}^o}'\mathbf{C_{11}^o}^{-1}\mathbf{C_{12}^o} \\ & -\mathbf{C_{21}^o}\mathbf{C_{11}^o}^{-1}\mathbf{D_{11}^o}(\mathbf{D_{11}^o}\mathbf{C_{11}^o}^{-1}\mathbf{D_{11}^o})^{-1}\mathbf{D_{21}^o}' \\ & +\mathbf{D_{21}^o}(\mathbf{D_{11}^o}\mathbf{C_{11}^o}^{-1}\mathbf{D_{11}^o})^{-1}\mathbf{D_{21}^o}'.\end{aligned}$$

Then,

$$N\bar{r}(\hat{\theta}, \hat{\rho})'\mathbf{B_o^{-1}}\bar{r}(\hat{\theta}, \hat{\rho}) \overset{a}{\sim} \chi_p^2. \tag{2.30}$$

Similarly to Proposition 2.6.1, consistent estimates of the elements of $\mathbf{C_o}$ and $\mathbf{D_o}$ will be used in practice for calculating the test statistic in (2.30).

2.7 Concluding remarks

We have proposed considering likelihood-based models in a GMM setting, in which knowledge about the joint distribution can be represented as copula moment conditions and efficiency and robustness of estimators can be assessed in terms of redundancy and robustness of the copula moments.

In considering copula robustness, all of the copula families that we compared to the normal benchmark except the Frank family are not comprehensive, i.e., they do not cover all possible values of the dependence measure τ. This makes such copula families relevant for modelling only certain degrees of dependence to which our robustness comparisons would apply.

For the Frank and Normal families, $\tau \in (-1, 1)$, so they are comprehensive. Given the simulation results, the Frank copula appears as useful in modelling any degree of dependence as the Normal family. It would be desirable to make comparisons with other comprehensive copulas such as the Plackett family. Similarly, comparisons of the Logistic copula to copulas with the same coverage should reveal its relative robustness.

The behavior of the AMH family of copulas was quite similar to that of the FGM family in our simulation. This was due to the small value of the dependence parameter ρ. The first order approximation in (2.16) is in this case quite accurate. It may not be so for larger ρ.

Finally, our results on copula robustness are problem-specific. For example, they are generally inapplicable to problems involving higher moments of a distribution. In similar simulations with problems other then sample-mean problems, radially symmetric copulas may not be robust to misspecification, but it should still be possible to compare robustness properties of copula families since the true copula is known.

Bibliography

BOUYÉ, E., V. DURRLEMAN, A. NIKEGHBALI, G. RIBOULET, AND T. RONCALLI (2000): "Copulas for Finance: A Reading Guide and Some Applications," *Crédit Lyonnais Working Paper*, gro.creditlyonnais.fr/content/wp/copula-survey.pdf.

BREUSCH, T., H. QIAN, P. SCHMIDT, AND D. WYHOWSKI (1999): "Redundancy of moment conditions," *Journal of Econometrics*, 91, 89–111.

BREYMANN, W., A. DIAS, AND P. EMBRECHTS (2003): "Dependence structures for multivariate high-frequency data in finance," *Quantitative Finance*, 3, 1–14, http://www.iop.org/EJ/abstract/1469-7688/3/1/301/.

CAMERON, A. C., T. LI, P. K. TRIVEDI, AND D. M. ZIMMER (2004): "Modelling the differences in counted outcomes using bivariate copula models with application to mismeasured counts," *Econometrics Journal*, 7, 566–84.

EMBRECHTS, P., A. HÖING, AND A. JURI (2003): "Using copulae to bound the Value-at-Risk for functions of dependent risks," *Finance and Stochastics*, 7, 145 167.

EMBRECHTS, P., A. MCNEIL, AND D. STRAUMANN (2002): "Correlation and dependence in risk management: properties and pitfalls," in *Risk Management: Value at Risk and Beyond*, ed. by M. Dempster, Cambridge: Cambridge University Press, 176223.

FRANK, M. (1979): "On the simultaneous associativity of $F(x, y)$ and $x + y - F(x, y)$," *Aequationes Mathematicae*, 19, 194–226.

GODAMBE, V. P. (1960): "An optimum property of regular maximum likelihood estimation," *The Annals of Mathematical Statistics*, 31, 1208–1211.

——— (1976): "Conditional likelihood and unconditional optimum estimating equations," *Biometrika*, 63, 277–284.

GOURIEROUX, C., A. MONFORT, AND A. TROGNON (1984): "Pseudo maximum likelihood methods: theory," *Econometrica*, 52, 681–700.

GREEN, W. (2002): *Econometric Analysis*, Prentice Hall.

HANSEN, L. (1982): "Large sample properties of generalized method of moments estimators," *Econometrica*, 50, 1029–1054.

KENDALL, M. G. (1949): "Rank and product-moment correlation," *Biometrika*, 36, 177–193.

LEE, L.-F. (1983): "Generalized econometric models with selectivity," *Econometrica*, 51, 507–512.

MARDIA, K., J. KENT, AND J. BIBBY (1979): *Multivariate Analysis*, Probability and Mathematical Statistics, London: Academic Press.

NELSEN, R. B. (1993): "Some concepts of bivariate symmetry," *Journal of Nonparametric Statistics*, 3, 95–101.

——— (1999): *An Introduction to Copulas*, vol. 139 of *Lecture Notes in Statistics*, Springer.

NEWEY, W. AND K. WEST (1987): "Hypothesis testing with efficient method of moments estimation," *International Economic Review*, 28, 777–787.

PATTON, A. (2001): "Modelling time-varying exchange rate dependence using the conditional copula," *UC San Diego Department of Economics Discussion Paper 2001-09.*

SCHMIDT, P. (2004): "Likelihood-based estimation in a panel setting," *MSU Working Paper.*

SKLAR, A. (1959): "Fonctions de répartition à n dimensions et leurs marges," *Publications de l'Institut de Statistique de l'Université de Paris*, 8, 229–231.

SMITH, M. D. (2003): "Modelling sample selection using Archimedian copulas," *Econometrics Journal*, 6, 99–123.

WHITE, H. (1982): "Maximum likelihood estimation of misspecified models," *Econometrica*, 50, 1–26.

WOOLDRIDGE, J. (1994): "Estimation and inference for dependent processes," in *Handbook of Econometrics*, ed. by R. Engle and D. McFadden, vol. IV.

——— (2002): *Econometric analysis of cross section and panel data*, Cambridge, Mass.: MIT Press.

Appendix A: Selected copula families

1. Independence copula:

$$C(u,v) = uv$$
$$c(u,v) = 1$$
$$\tau = 0$$

2. Logistic copula:

$$C(u,v) = \frac{uv}{u+v-uv}$$
$$c(u,v) = \frac{2uv}{(u+v-uv)^3}$$
$$\tau = 1/3$$

3. Farlie-Gumbel-Morgenstern family:

$$C(u,v,\rho) = uv(1+\rho(1-u)(1-v))$$
$$c(u,v,\rho) = 1+\rho-2\rho u-2\rho v+4\rho uv$$
$$\rho \in [-1;1]$$
$$\tau = 2\rho/9 \in [-2/9, 2/9]$$

4. Joe family*:

$$C(u,v,\rho) = 1 - ((1-u)^{\rho} + (1-v)^{\rho} - (1-u)^{\rho}(1-v)^{\rho})^{1/\rho}$$

$$\rho \in [1,\infty)$$

$$\varphi(t) = -\log(1-(1-t)^{\rho})$$

$$\tau \in [0,1)$$

5. Ali-Mikhail-Haq family*:

$$C(u,v,\rho) = \frac{uv}{1-\rho(1-u)(1-v)}$$

$$\rho \in [-1,1)$$

$$\varphi(t) = \log\frac{1-\rho(1-t)}{t}$$

$$\tau \in [-0.182, 1/3)$$

6. Clayton family*:

$$C(u,v,\rho) = \begin{cases} uv, & \rho = 0 \\ (u^{-\rho} + v^{-\rho} - 1)^{-1/\rho}, & \rho \neq 0 \end{cases}$$

$$\rho \in [0,\infty)$$

$$\varphi(t) = \frac{1}{\rho}(t^{-\rho} - 1)$$

$$\tau = \frac{\rho}{\rho+2} \in [0,1)$$

7. Gumbel family*:

$$C(u,v,\rho) = \exp\left[-((-\ln u)^{\rho} + (-\ln v)^{\rho})^{1/\rho}\right]$$

$$\rho \in [1,\infty)$$

$$\varphi(t) = (-\log t)^{\rho}$$

$$\tau = \frac{\rho - 1}{\rho} \in [0,1)$$

8. Frank family*:

$$C(u,v,\rho) = \begin{cases} uv, & \rho = 0 \\ -\frac{1}{\rho}\ln\left[1 + \frac{(e^{-\rho u}-1)(e^{-\rho v}-1)}{e^{-\rho}-1}\right], & \rho \neq 0 \end{cases}$$

$$\rho \in (-\infty,\infty)$$

$$\varphi(t) = -\ln\frac{e^{-\rho t}-1}{e^{-\rho}-1}$$

$$\tau \in (-1,1)$$

9. Plackett family:

$$C(u,v,\rho) = \begin{cases} uv, & \rho = 1 \\ \frac{\left(1+(u+v)(\rho-1)-\sqrt{(1+(u+v)(\rho-1))^2-4uv\rho(\rho-1)}\right)}{2(\rho-1)}, & \rho \neq 1 \end{cases}$$

$$\rho \in (0,\infty)$$

$$\tau \in (-1,1)$$

10. Normal family:

$$C(u, v, \rho) = \Phi_2(\Phi^{-1}(u), \Phi^{-1}(v); \rho)$$
$$\rho \in (-1, 1)$$
$$\tau = \frac{2}{\pi} \arcsin \rho \in (-1, 1)$$

Note: * denotes Archimedean copulas, i.e. copulas generated as

$$C(u, v) = \varphi^{-1}(\varphi(u) + \varphi(v)),$$

where $\varphi : \mathbb{I} \to [0, \infty]$, continuous, $\varphi'(t) < 0$ and $\varphi''(t) > 0$ $\forall t \in (0, 1)$ is called the generator function. It can be shown (see, e.g., Nelsen, 1999, p.130) that for Archimedean copulas, Kendall's

$$\tau = 1 + 4 \int_0^1 \frac{\varphi(t)}{\varphi'(t)} dt.$$

Appendix B: Proofs

PROOF OF THEOREM 2.4.1:

We show that $\mathbb{E}\frac{\partial}{\partial\mu}\ln k(F_1(\mu_1+X_1),F_2(\mu_2+X_2);\rho^k)=0$, where $\mu=(\mu_1,\mu_2)'$, holds for any RS K.

By the chain rule, $\frac{\partial}{\partial\mu}\ln k(F_1(\mu_1+x_1),F_2(\mu_2+x_2);\rho^k)$ contains terms of the form

$$\frac{1}{k(F_1(\mu_1+x_1),F_2(\mu_2+x_2);\rho^k)}\times\frac{\partial k(F_1(\mu_1+x_1),F_2(\mu_2+x_2);\rho^k)}{\partial F_i(\mu_i+x_i)}\times f_i(\mu_i+x_i), \tag{2.31}$$

$i=1,2$.

Due to MS of (X_1,X_2) and RS of K, $f_i(\mu_i+x_i)=f_i(\mu_i-x_i)$ and $k(F_1(\mu_1+x_1),F_2(\mu_2+x_2))=k(1-F_1(\mu_1+x_1),1-F_2(\mu_2+x_2))=k(F_1(\mu_1-x_1),F_2(\mu_2-x_2))$. So the first term in (2.31) is the same whether evaluated at (x_1,x_2) or $(-x_1,-x_2)$. Similarly, the last term is the same whether evaluated at x_i or $-x_i$.

Furthermore,

$$\begin{aligned}\frac{\partial k(F_1(\mu_1+x_1),F_2(\mu_2+x_2);\rho^k)}{\partial F_i(\mu_i+x_i)} &= \frac{\partial k(1-F_1(\mu_1+x_1),1-F_2(\mu_2+x_2);\rho^k)}{\partial(1-F_i(\mu_i-x_i))}\\ &= -\frac{\partial k(F_1(\mu_1-x_1),F_2(\mu_2-x_2);\rho^k)}{\partial F_i(\mu_i-x_i)}.\end{aligned}$$

Thus, $\frac{\partial}{\partial\mu}\ln k(F_1(\mu_1+x_1),F_2(\mu_2+x_2);\rho^k)=-\frac{\partial}{\partial\mu}\ln k(F_1(\mu_1-x_1),F_2(\mu_2-x_2);\rho^k)$.

Denote $g(x_1,x_2)\equiv\frac{\partial}{\partial\mu}\ln k(F_1(\mu_1+x_1),F_2(\mu_2+x_2);\rho^k)\cdot h(\mu_1+x_1,\mu_2+x_2;\rho)$. From the above, it follows with RS that $g(-x_1,-x_2)=-g(x_1,x_2)$.

We thus have

$$\begin{aligned}
\mathbb{E}\frac{\partial}{\partial\mu}\ln k(F_1(\mu_1+X_1),F_2(\mu_2+X_2);\rho^k) &= \int_{-\infty}^{\infty}\int_{-\infty}^{\infty} g(x_1,x_2)dx_1dx_2 \\
&= \int_{-\infty}^{0}\int_{-\infty}^{0} g(x_1,x_2)dx_1dx_2 \\
&+ \int_{-\infty}^{0}\int_{0}^{\infty} g(x_1,x_2)dx_1dx_2 \\
&+ \int_{0}^{\infty}\int_{0}^{\infty} g(x_1,x_2)dx_1dx_2 \\
&+ \int_{0}^{\infty}\int_{-\infty}^{0} g(x_1,x_2)dx_1dx_2 \\
&= \int_{0}^{\infty}\int_{0}^{\infty} g(-x_1,-x_2)dx_1dx_2 \\
&+ \int_{0}^{\infty}\int_{-\infty}^{0} g(-x_1,-x_2)dx_1dx_2 \\
&+ \int_{0}^{\infty}\int_{0}^{\infty} g(x_1,x_2)dx_1dx_2 \\
&+ \int_{0}^{\infty}\int_{-\infty}^{0} g(x_1,x_2)dx_1dx_2 \\
&= 0.
\end{aligned}$$

PROOF OF LEMMA 2.5.1: By the information matrix equality (IME),

$$\mathbf{A} \equiv \mathbb{E}\left\{\frac{\partial}{\partial\theta}\ln f_1(X_1;\theta)\frac{\partial}{\partial\theta'}\ln f_1(X_1;\theta)\right\} = -\mathbb{E}\frac{\partial^2}{\partial\theta\partial\theta'}\ln f_1(X_1;\theta). \tag{2.32}$$

Similar for $\mathbf{B},\mathbf{F}$.

By the generalized IME (GIME),

$$\begin{aligned}
\mathbf{E} &\equiv \mathbb{E}\left\{\frac{\partial}{\partial\theta}\ln c(F_1(X_1;\theta),F_2(X_2;\theta);\rho)\frac{\partial}{\partial\rho'}\ln c(F_1(X_1;\theta),F_2(X_2;\theta);\rho)\right\} \\
&= -\mathbb{E}\frac{\partial^2}{\partial\theta\partial\rho'}\ln c(F_1(X_1;\theta),F_2(X_2;\theta);\rho)
\end{aligned} \tag{2.33}$$

and, for $i = 1, 2$,

$$\mathbb{E}\left\{\frac{\partial}{\partial\theta}\ln f_i(X_i;\theta)\frac{\partial}{\partial\rho'}\ln c(F_1(X_1;\theta),F_1(X_2;\theta);\rho)\right\} = -\mathbb{E}\frac{\partial^2}{\partial\theta\partial\rho'}\ln f_i(X_i;\theta) = \mathbf{0}.$$

Also by GIME and (2.2),

$$\mathbb{E}\left\{\frac{\partial}{\partial\theta}\ln f_i(X_i;\theta)\frac{\partial}{\partial\theta'}\left[\ln f_1(X_1;\theta)+\ln f_2(X_2;\theta)+\ln c(.,.;\rho)\right]\right\} = \\ = -\mathbb{E}\frac{\partial^2}{\partial\theta\partial\theta'}\ln f_i(X_i;\theta)$$

for $i = 1, 2$, which, along with (2.32), implies that

$$\begin{aligned}\mathbf{G} &\equiv \mathbb{E}\left\{\frac{\partial}{\partial\theta}\ln f_1(X_1;\theta)\frac{\partial}{\partial\theta'}\ln f_2(X_2;\theta)\right\} \\ &= -\mathbb{E}\left\{\frac{\partial}{\partial\theta}\ln f_1(X_1;\theta)\frac{\partial}{\partial\theta'}\ln c(F_1(X_1;\theta),F_1(X_2;\theta);\rho)\right\}\end{aligned}$$

and

$$\mathbb{E}\left\{\frac{\partial}{\partial\theta}\ln f_2(X_2;\theta)\frac{\partial}{\partial\theta'}\ln f_1(X_1;\theta)\right\} = \\ = -\mathbb{E}\left\{\frac{\partial}{\partial\theta}\ln f_2(X_2;\theta)\frac{\partial}{\partial\theta'}\ln c(F_1(X_1;\theta),F_1(X_2;\theta);\rho)\right\} = \mathbf{G}'. \qquad (2.34)$$

Finally, by GIME and (2.2),

$$\mathbb{E}\left\{\frac{\partial}{\partial\theta}\ln c(F_1(X_1;\theta),F_1(X_2;\theta);\rho)\frac{\partial}{\partial\theta'}\times\right. \\ \left.\times\left[\ln f_1(X_1;\theta)+\ln f_2(X_2;\theta)+\ln c(F_1(X_1;\theta),F_1(X_2;\theta);\rho)\right]\right\} = \\ = -\mathbb{E}\frac{\partial^2}{\partial\theta\partial\theta'}\ln c(F_1(X_1;\theta),F_1(X_2;\theta);\rho).$$

With $\mathbf{G}$ as defined above and

$$\mathbf{J} \equiv \mathbb{E}\left\{\frac{\partial}{\partial\theta}\ln c(F_1(X_1;\theta), F_1(X_2;\theta);\rho)\frac{\partial}{\partial\theta'}\ln c(F_1(X_1;\theta), F_1(X_2;\theta);\rho)\right\}$$

this implies that

$$\mathbb{E}\frac{\partial^2}{\partial\theta\partial\theta'}\ln c(F_1(X_1;\theta), F_1(X_2;\theta);\rho) = \mathbf{G} + \mathbf{G}' - \mathbf{J}.$$

PROOF OF THEOREM 2.5.1: See text.

PROOF OF THEOREM 2.5.2: By (2.20) and (2.23),

$$\mathbb{V}_{\text{MLE}} = \begin{bmatrix} \mathbf{A}+\mathbf{B}+\mathbf{J}-\mathbf{G}-\mathbf{G}' & \mathbf{E}' \\ \mathbf{E} & \mathbf{F} \end{bmatrix}^{-1},$$

$$\mathbb{V}_{\text{IQMLE}} = \left(\begin{bmatrix} -\mathbf{A} & -\mathbf{B} \end{bmatrix} \begin{bmatrix} \mathbf{A} & \mathbf{G} \\ \mathbf{G}' & \mathbf{B} \end{bmatrix}^{-1} \begin{bmatrix} -\mathbf{A} \\ -\mathbf{B} \end{bmatrix} \right)^{-1}. \tag{2.35}$$

Using partitioned inverse formulas, the upper left $p \times p$ block of $\mathbb{V}_{\text{MLE}}$ can be written as $\mathbf{\Sigma}^{-1}$, where $\mathbf{\Sigma} = \mathbf{A}+\mathbf{B}+\mathbf{J}-\mathbf{G}-\mathbf{G}'-\mathbf{E}'\mathbf{F}^{-1}\mathbf{E}$.

Also,

$$\begin{aligned}
\mathbb{V}_{\text{IQMLE}}^{-1} &= \left(\begin{bmatrix} -\mathbf{G}' & -\mathbf{G} \end{bmatrix} + \begin{bmatrix} \mathbb{I} & \mathbb{I} \end{bmatrix} \begin{bmatrix} \mathbf{A} & \mathbf{G} \\ \mathbf{G}' & \mathbf{B} \end{bmatrix} \right) \begin{bmatrix} \mathbf{A} & \mathbf{G} \\ \mathbf{G}' & \mathbf{B} \end{bmatrix}^{-1} \times \\
&\quad \times \left(\begin{bmatrix} \mathbf{A} & \mathbf{G} \\ \mathbf{G}' & \mathbf{B} \end{bmatrix} \begin{bmatrix} \mathbb{I} \\ \mathbb{I} \end{bmatrix} + \begin{bmatrix} -\mathbf{G} \\ -\mathbf{G}' \end{bmatrix} \right) \qquad (2.36) \\
&= \begin{bmatrix} -\mathbf{G}' & -\mathbf{G} \end{bmatrix} \begin{bmatrix} \mathbf{A} & \mathbf{G} \\ \mathbf{G}' & \mathbf{B} \end{bmatrix}^{-1} \begin{bmatrix} -\mathbf{G} \\ -\mathbf{G}' \end{bmatrix} \\
&\quad -\mathbf{G}' - \mathbf{G} + \mathbf{A} + \mathbf{G} + \mathbf{G}' + \mathbf{B} - \mathbf{G} - \mathbf{G}'. \qquad (2.37)
\end{aligned}$$

Thus, $\mathbb{V}_{\text{IQMLE}}^{-1} = \boldsymbol{\Sigma}$ if and only if

$$\mathbf{J} - \mathbf{E}'\mathbf{F}^{-1}\mathbf{E} = \begin{bmatrix} -\mathbf{G}' & -\mathbf{G} \end{bmatrix} \begin{bmatrix} \mathbf{A} & \mathbf{G} \\ \mathbf{G}' & \mathbf{B} \end{bmatrix}^{-1} \begin{bmatrix} -\mathbf{G} \\ -\mathbf{G}' \end{bmatrix}.$$

PROOF OF COROLLARY 2.5.1:

1. If (C) is a linear combination of (A) and (B) then covariances between moment functions in (C) and (D) are linear combinations of covariances between (D) and (A-B), which are all zero by Lemma 2.5.1.

2. Rewrite $\mathbf{J} - \mathbf{C}_{21}^{\theta}\mathbf{C}_{11}^{-1}\mathbf{C}_{12}^{\theta}$ as

$$\mathbb{E}\left\{ \left(\frac{\partial}{\partial \theta} \ln c - \mathbf{C}_{21}^{\theta}\mathbf{C}_{11}^{-1} \begin{bmatrix} \frac{\partial}{\partial \theta} \ln f_1 \\ \frac{\partial}{\partial \theta} \ln f_2 \end{bmatrix} \right) \frac{\partial}{\partial \theta'} \ln c \right\}.$$

This is identically zero because, due to linearity of (C) in (A-B),

$$\frac{\partial}{\partial\theta}\ln c - \mathbf{C}_{21}^{\theta}\mathbf{C}_{11}^{-1}\begin{bmatrix}\frac{\partial}{\partial\theta}\ln f_1\\ \frac{\partial}{\partial\theta}\ln f_2\end{bmatrix} = \mathbf{0}.$$

3. By Theorem 2.5.2.

PROOF OF LEMMA 2.5.2: By construction, blocks $\mathbf{A}, \mathbf{B}, \mathbf{G}$ of matrices $\mathbf{C^k}$ and $\mathbf{D^k}$ are the same as in Lemma 2.5.1. However, GIME does not apply now.

$$\mathbf{V} \equiv \mathbb{E}\left\{\frac{\partial}{\partial\theta}\ln k(F_1(X_1;\theta),F_2(X_2;\theta);\rho^k)\frac{\partial}{\partial(\rho)'}\ln k(F_1(X_1;\theta),F_2(X_2;\theta);\rho^k)\right\} \neq$$
$$\neq -\mathbb{E}\left\{\frac{\partial^2}{\partial\theta\partial\rho'}\ln k(F_1(X_1;\theta),F_2(X_2;\theta);\rho^k)\right\} \equiv -\mathbf{S}.$$

$$\begin{aligned}-\mathbf{P} &\equiv \mathbb{E}\left\{\frac{\partial}{\partial\theta}\ln f_1(X_1;\theta)\frac{\partial}{\partial\rho'}\ln k(F_1(X_1;\theta),F_1(X_2;\theta);\rho^k)\right\}\\ &\neq -\mathbb{E}\frac{\partial^2}{\partial\theta\partial\rho'}\ln f_1(X_1;\theta) = \mathbf{0}\end{aligned}$$

and

$$\begin{aligned}-\mathbf{Q}' &\equiv \mathbb{E}\left\{\frac{\partial}{\partial\theta}\ln f_2(X_2;\theta)\frac{\partial}{\partial\rho'}\ln k(F_1(X_1;\theta),F_1(X_2;\theta);\rho^k)\right\}\\ &\neq -\mathbb{E}\frac{\partial^2}{\partial\theta\partial\rho'}\ln f_2(X_2;\theta) = \mathbf{0}.\end{aligned}$$

$$\begin{aligned}\mathbf{G} &\equiv \mathbb{E}\left\{\frac{\partial}{\partial\theta}\ln f_1(X_1;\theta)\frac{\partial}{\partial\theta'}\ln f_2(X_2;\theta)\right\} \neq\\ &\neq -\mathbb{E}\left\{\frac{\partial}{\partial\theta}\ln f_1(X_1;\theta)\frac{\partial}{\partial\theta'}\ln k(F_1(X_1;\theta),F_1(X_2;\theta);\rho^k)\right\} \equiv \mathbf{K}\end{aligned}$$

and

$$\mathbb{E}\left\{\frac{\partial}{\partial\theta}\ln f_2(X_2;\theta)\frac{\partial}{\partial\theta'}\ln f_1(X_1;\theta)\right\} \neq$$

$$\neq -\mathbb{E}\left\{\frac{\partial}{\partial\theta}\ln f_2(X_2;\theta)\frac{\partial}{\partial\theta'}\ln k(F_1(X_1;\theta),F_1(X_2;\theta);\rho^k)\right\} \equiv \mathbf{L}'.$$

However, by GIME and (2.2),

$$\begin{aligned}\mathbb{E}\frac{\partial^2}{\partial\theta\partial\theta'}\ln k(F_1(X_1;\theta),F_1(X_2;\theta);\rho^k) &= -\mathbb{E}\left\{\frac{\partial}{\partial\theta}\ln k(F_1(X_1;\theta),F_1(X_2;\theta);\rho^k)\times\right.\\ &\quad\times\left[\frac{\partial}{\partial\theta'}\ln f_1(X_1;\theta)+\right.\\ &\qquad+\frac{\partial}{\partial\theta'}\ln f_2(X_2;\theta)+\\ &\qquad\left.\left.+\frac{\partial}{\partial\theta'}\ln c(F_1(X_1;\theta),F_1(X_2;\theta);\rho)\right]\right\}\\ &\equiv \mathbf{K}'+\mathbf{L}-\mathbf{M}, \qquad (2.38)\end{aligned}$$

and

$$\begin{aligned}\mathbb{E}\frac{\partial^2}{\partial\rho\partial\theta'}\ln k(F_1(X_1;\theta),F_1(X_2;\theta);\rho^k) &= -\mathbb{E}\left\{\frac{\partial}{\partial\rho}\ln k(F_1(X_1;\theta),F_1(X_2;\theta);\rho^k)\right.\\ &\quad\times\left[\frac{\partial}{\partial\theta'}\ln f_1(X_1;\theta)+\right.\\ &\qquad+\frac{\partial}{\partial\theta'}\ln f_2(X_2;\theta)+\\ &\qquad\left.\left.+\frac{\partial}{\partial\theta'}\ln c(F_1(X_1;\theta),F_1(X_2;\theta);\rho)\right]\right\}\\ &\equiv \mathbf{P}'+\mathbf{Q}-\mathbf{R}, \qquad (2.39)\end{aligned}$$

$$\begin{aligned}-\mathbf{T} &\equiv \mathbb{E}\frac{\partial^2}{\partial\rho\partial\rho'}\ln k(F_1(X_1;\theta),F_1(X_2;\theta);\rho^k)\\ &= -\mathbb{E}\left\{\frac{\partial}{\partial\rho}\ln k(F_1(X_1;\theta),F_1(X_2;\theta);\rho^k)\frac{\partial}{\partial\rho'}\ln c(F_1(X_1;\theta),F_1(X_2;\theta);\rho)\right\},\end{aligned}$$

and

$$\begin{aligned} -\mathbf{S} &\equiv \mathbb{E}\frac{\partial^2}{\partial\theta\partial\rho'}\ln k(F_1(X_1;\theta),F_1(X_2;\theta);\rho^k) \\ &= -\mathbb{E}\left\{\frac{\partial}{\partial\theta}\ln k(F_1(X_1;\theta),F_1(X_2;\theta);\rho^k)\frac{\partial}{\partial\rho'}\ln c(F_1(X_1;\theta),F_1(X_2;\theta);\rho)\right\}. \end{aligned}$$

Also,

$$\mathbf{N} \equiv \mathbb{E}\left\{\frac{\partial}{\partial\theta}\ln k(F_1(X_1;\theta),F_1(X_2;\theta);\rho^k)\frac{\partial}{\partial\theta'}\ln k(F_1(X_1;\theta),F_1(X_2;\theta);\rho^k)\right\} \neq \mathbf{M}$$

and

$$\mathbf{W} \equiv \mathbb{E}\left\{\frac{\partial}{\partial\rho}\ln k(F_1(X_1;\theta),F_1(X_2;\theta);\rho)\frac{\partial}{\partial\rho'}\ln k(F_1(X_1;\theta),F_1(X_2;\theta);\rho^k)\right\} \neq \mathbf{T}.$$

Finally, by the well known algebraic property of cross-partial derivatives,

$$\mathbf{S} = -\mathbf{P} - \mathbf{Q}' + \mathbf{R}'.$$

PROOF OF THEOREM 2.5.3: See main text.

PROOF OF THEOREM 2.5.4: By Theorem 8(C) of Breusch et al. (1999), (C′-D′) are redundant for θ given (A-B) if and only if

$$\begin{bmatrix} \mathbf{K}'+\mathbf{L}-\mathbf{M} \\ \mathbf{P}'+\mathbf{Q}-\mathbf{R} \end{bmatrix} - \begin{bmatrix} -\mathbf{K}' & -\mathbf{L} \\ -\mathbf{P}' & -\mathbf{Q} \end{bmatrix} \mathbf{C}_{11}^{-1}\mathbf{D}_{11} = \begin{bmatrix} -\mathbf{S} \\ -\mathbf{T} \end{bmatrix} \mathbb{B},$$

for some matrix $\mathbb{B} : q \times p$.

This is equivalent to

$$-\mathbf{M} - [-\mathbf{K}' \quad -\mathbf{L}]\mathbf{C}_{\mathbf{11}}^{-\mathbf{1}} \begin{bmatrix} \mathbf{G} \\ \mathbf{G}' \end{bmatrix} = -\mathbf{S}\mathbb{B},$$

$$-\mathbf{R} - [-\mathbf{P}' \quad -\mathbf{Q}]\mathbf{C}_{\mathbf{11}}^{-\mathbf{1}} \begin{bmatrix} \mathbf{G} \\ \mathbf{G}' \end{bmatrix} = -\mathbf{T}\mathbb{B}.$$

$\mathbf{T}$ is symmetric and invertible, so we can substitute $\mathbb{B}$ from the latter equation into the former to obtain

$$\mathbf{M} - [-\mathbf{K}' \quad -\mathbf{L}]\mathbf{C}_{\mathbf{11}}^{-\mathbf{1}}\mathbf{C}_{\mathbf{12}}^{\boldsymbol{\theta}} = \mathbf{S}\mathbf{T}^{-\mathbf{1}}(\mathbf{R} - [-\mathbf{P}' \quad -\mathbf{Q}]\mathbf{C}_{\mathbf{11}}^{-\mathbf{1}}\mathbf{C}_{\mathbf{12}}^{\boldsymbol{\theta}}),$$

which completes the proof.

PROOF OF COROLLARY 2.5.2:

1. By (2.28), $\mathbf{M} - \mathbf{C}_{\mathbf{21}}^{\boldsymbol{\theta}\mathbf{k}}\mathbf{C}_{\mathbf{11}}^{-\mathbf{1}}\mathbf{C}_{\mathbf{12}}^{\boldsymbol{\theta}}$ is identically zero under linearity of (C) in (A-B).

2. As in 1.

3. By Theorem 2.5.4.

PROOF OF PROPOSITION 2.6.1: See proof of Lemma 4.2 of Hansen (1982).

PROOF OF PROPOSITION 2.6.2: First note that, by standard optimal GMM results, $\hat{\theta}$ satisfies

$$\sqrt{N}(\hat{\theta} - \theta_o) = -(\mathbf{D_{11}^o}'\mathbf{C_{11}^o}^{-1}\mathbf{D_{11}^o})^{-1}\mathbf{D_{11}^o}'\mathbf{C_{11}^o}^{-1}\sqrt{N}\bar{g}(\theta_o) + o_p(1). \quad (2.40)$$

The first order condition for $\hat{\rho}$ can equivalently be written as

$$\begin{aligned} [\frac{\partial}{\partial \rho'}\bar{r}(\hat{\theta}, \hat{\rho})]'\mathbf{B_o}^{-1}\bar{r}(\hat{\theta}, \hat{\rho}) &= 0, \\ \mathbf{D_{22}^o}'\mathbf{B_o}^{-1}\sqrt{N}\bar{r}(\hat{\theta}, \hat{\rho}) &= o_p(1). \end{aligned} \quad (2.41)$$

Now, by the mean-value theorem, we have

$$\sqrt{N}\bar{r}(\hat{\theta}, \hat{\rho}) = \sqrt{N}\bar{r}(\theta_o, \rho_o) + \mathbf{D_{21}^o}\sqrt{N}(\hat{\theta} - \theta_o) + \mathbf{D_{22}^o}\sqrt{N}(\hat{\rho} - \rho_o) + o_p(1). \quad (2.42)$$

Substituting (2.40) into (2.42), pre-multiplying by $\mathbf{D_{22}^o}'\mathbf{B_o}^{-1}$, and solving for $\sqrt{N}(\hat{\rho} - \rho_o)$ using (2.41) yields

$$\begin{aligned} \sqrt{N}(\hat{\rho} - \rho_o) &= -(\mathbf{D_{22}^o}'\mathbf{B_o^{-1}}\mathbf{D_{22}^o})^{-1}\mathbf{D_{22}^o}'\mathbf{B_o^{-1}}\sqrt{N}\bar{r}(\theta_o, \rho_o) \\ &\quad +(\mathbf{D_{22}^o}'\mathbf{B_o^{-1}}\mathbf{D_{22}^o})^{-1}\mathbf{D_{22}^o}'\mathbf{B_o^{-1}}\mathbf{D_{21}^o} \times \\ &\qquad \times(\mathbf{D_{11}^o}'\mathbf{C_{11}^o}^{-1}\mathbf{D_{11}^o})^{-1}\mathbf{D_{11}^o}'\mathbf{C_{11}^o}^{-1}\sqrt{N}\bar{g}(\theta_o) \\ &\quad +o_p(1). \end{aligned} \quad (2.43)$$

Substituting (2.43) and (2.40) into (2.42) and simplifying results in

$$\sqrt{N}\bar{r}(\hat{\theta}, \hat{\rho}) = \mathbf{R_o}\sqrt{N}\bar{\phi}(\theta_o, \rho_o) + o_p(1), \quad (2.44)$$

where

$$\begin{aligned}\mathbf{R_o} &= \mathbb{I} - \mathbf{D_{22}^o}(\mathbf{D_{22}^o}'\mathbf{B_o^{-1}}\mathbf{D_{22}^o})^{-1}\mathbf{D_{22}^o}'\mathbf{B_o^{-1}}, \\ \bar{\phi}(\theta_o, \rho_o) &= \bar{r}(\theta_o, \rho_o) - \mathbf{D_{21}^o}(\mathbf{D_{11}^o}'\mathbf{C_{11}^o}^{-1}\mathbf{D_{11}^o})^{-1}\mathbf{D_{11}^o}'\mathbf{C_{11}^o}^{-1}\bar{g}(\theta_o).\end{aligned}$$

Note that $\sqrt{N}\bar{\phi}(\theta_o, \rho_o) \sim \mathbb{N}(\mathbf{0}, \mathbf{B_o})$, and thus $\mathbf{B_o^{-1/2}}\sqrt{N}\bar{\phi}(\theta_o, \rho_o) \sim \mathbb{N}(\mathbf{0}, \mathbb{I})$. Also, note that $\mathbf{R_o'B_o^{-1}R_o} = \mathbf{B_o^{-\frac{1}{2}}}[\mathbf{I} - \mathbf{B_o^{-\frac{1}{2}}}\mathbf{D_{22}^o}(\mathbf{D_{22}^o}'\mathbf{B_o^{-1}}\mathbf{D_{22}^o})^{-1}\mathbf{D_{22}^o}'\mathbf{B_o^{-\frac{1}{2}}}]\mathbf{B_o^{-\frac{1}{2}}}$.

Thus, the test statistic in (2.30) can be written as

$$N\bar{h}(\hat{\theta}, \hat{\rho})'\mathbf{B_o^{-1}}\bar{h}(\hat{\theta}, \hat{\rho}), \tag{2.45}$$

i.e. as a quadratic form in standard normals with the coefficient matrix

$$\mathbb{P} = \mathbf{I}_{p+q} - \mathbf{B_o^{-1/2}}\mathbf{D_{22}^o}(\mathbf{D_{22}^o}'\mathbf{B_o^{-1}}\mathbf{D_{22}^o})^{-1}\mathbf{D_{22}^o}'\mathbf{B_o^{-1/2}}. \tag{2.46}$$

This matrix is idempotent: it is the projection matrix orthogonal to $\mathbf{B_o^{-\frac{1}{2}}}\mathbf{D_{22}^o}$. The χ^2-test in (2.30) follows immediately because $tr(\mathbb{P}) = p + q - rank(D_{22}^o) = p$.

Appendix C: Plots of simulated sample moments

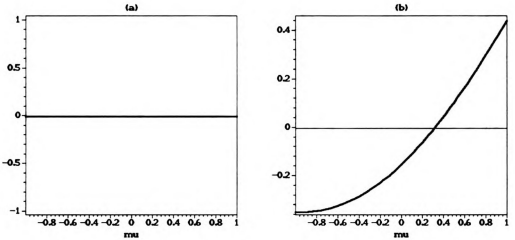

Figure 2.1: $\bar{\delta}^{\mu}(\mu)$ for no-parameter copulas: (a) Independence copula; (b) Logistic copula.

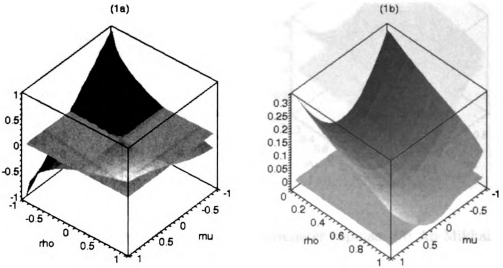

Figure 2.2: $\bar{\delta}^{\mu}(\mu,\rho)$ and $\bar{\delta}^{\rho}(\mu,\rho)$ for one-parameter copulas: (1) Farlie-Gumbel-Morgenstern.

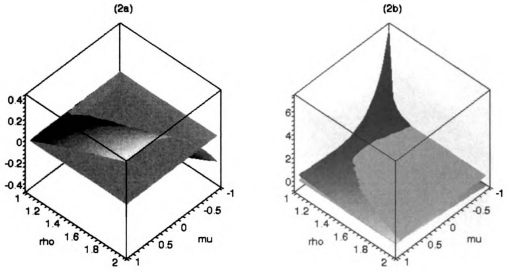

Figure 2.3: $\bar{\delta}^{\mu}(\mu, \rho)$ and $\bar{\delta}^{\rho}(\mu, \rho)$ for one-parameter copulas: (2) Joe.

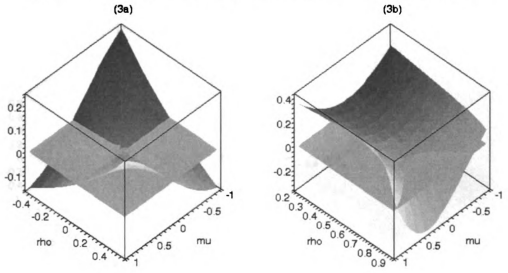

Figure 2.4: $\bar{\delta}^{\mu}(\mu, \rho)$ and $\bar{\delta}^{\rho}(\mu, \rho)$ for one-parameter copulas: (3) Ali-Mikhail-Haq.

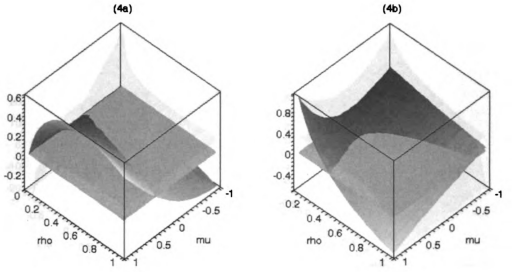

Figure 2.5: $\bar{\delta}^{\mu}(\mu, \rho)$ and $\bar{\delta}^{\rho}(\mu, \rho)$ for one-parameter copulas: (4) Clayton.

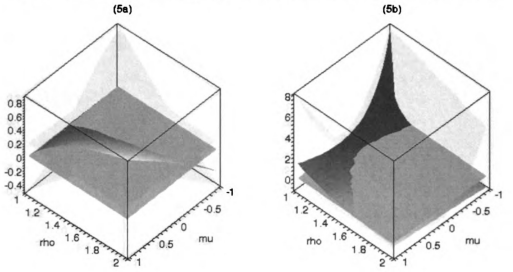

Figure 2.6: $\bar{\delta}^{\mu}(\mu, \rho)$ and $\bar{\delta}^{\rho}(\mu, \rho)$ for one-parameter copulas: (5) Gumbel.

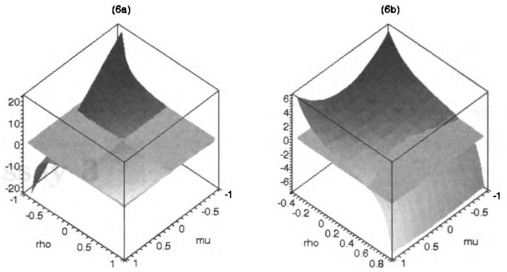

Figure 2.7: $\bar{\delta}^{\mu}(\mu,\rho)$ and $\bar{\delta}^{\rho}(\mu,\rho)$ for one-parameter copulas: (6) Normal.

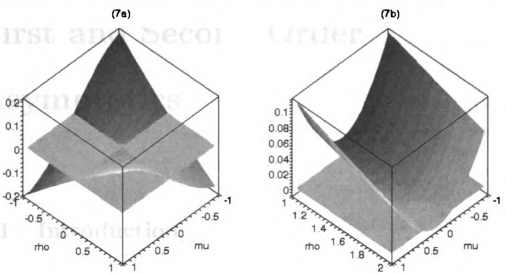

Figure 2.8: $\bar{\delta}^{\mu}(\mu,\rho)$ and $\bar{\delta}^{\rho}(\mu,\rho)$ for one-parameter copulas: (7) Frank.

Essay 3

Modelling Covariance Structures: First and Second Order Asymptotics

3.1 Introduction

This paper considers estimation of covariance structure models, i.e. models formulated in terms of the second moments of the data. One situation when such models arise is when there are some variables that are unobserved but whose presence in the model introduces a particular pattern of correlation between observed variables (e.g., linear structural relationship (LISREL) models, multiple indicators multiple causes (MIMIC) models, factor analysis (FA) and random effect (RE) models).

Traditional estimation methods for such models are based on the assumption of multivariate normality (see, e.g., Jöreskog, 1970; Jöreskog and Goldberger, 1975; Jöreskog and Sörbom, 1977). Because even moments of normally distributed data are function of the second moment, no improvements can be made by using higher order moments of the data. The maximum likelihood estimator (MLE) is efficient under normality.

If the data are not normal, MLE is still consistent. However, the MLE standard errors are wrong and consequently inference may be incorrect. An obvious way to make inference robust to non-normality is to adjust standard errors using the "sandwich" form of the variance matrix. The exact form of the variance matrix for normal quasi-MLE of covariance structures can be found, e.g., in Chamberlain (1984, p. 1295). Although obvious, the QMLE improvement to covariance structure modelling does not seem to be widely implemented. For example, the popular software packages used for covariance structure models do not seem to do that (see, e.g., Jöreskog and Sörbom, 1996).

It is well known that the efficient generalized method of moments estimator (GMM) optimally uses the information available in the moment conditions ("efficient" here means "first-order efficient"). For covariance structures, this means that GMM makes efficient use of the restrictions on the second moments whether or not the data are in fact normal. Similarly, the family of empirical likelihood estimators that are first order asymptotic equivalents of GMM, possess the same property. It is therefore intuitive that GMM of covariance structure should be no worse than normal QMLE asymptotically. This intuition has been noted in Chamberlain (1982, 1984); Ahn and Schmidt (1995) and other papers. The common argument is that the GMM estimator attains the lower bound for the asymptotic variance

matrix for estimators that use the second moment restrictions.

One of this paper's contributions is that it provides a formal comparison between the estimators to the first order in terms of the relevant variance matrices. It presents an explicit condition for equal relative efficiency of GMM and normal QMLE. The condition is expressed in terms of the fourth moments of the data and normality is shown to be one case when the condition holds. Such a representation provides a clear form of the efficiency gain and identifies a family of distributions for which normal QMLE and optimal GMM are equally efficient. This result is given in Section 3.3.

Section 3.2 describes the general model and the estimators. The linear interdependent structural relationship (LISREL) model is a special case of the general model. We describe this widely used model in Subsection 3.2.2.

Section 3.4 considers second order asymptotics. It is well known that the GMM estimator has a second order bias that contains more terms than that of the EL estimator. Newey et al. (2003); Newey and Smith (2004) derive the relevant bias expressions. The extra bias terms in GMM come from the estimation of the optimal weighting matrix and the derivative matrix that are both parts of the GMM first order conditions. It is unknown how the two estimators (GMM and EL) compare to normal QMLE in terms of the second order bias.

Intuitively, the answer to the question of second order asymptotic comparisons is clear. If the true distribution of the data is discrete then MLE and EL are identical. Furthermore, the bias expression for EL does not depend on discreteness. So if the assumed distribution (normal, in the case of Gaussian QMLE) turns out to

be correct, we should have the same bias for EL as for (Q)MLE. However it is still interesting to have an explicit form of the QMLE bias so that comparisons with other distributions can conceivably be made. Note that equal first order efficiency of QMLE and GMM (EL) may not be attainable for other distributions. We first derive the second order bias of normal QMLE expressed in terms of higher moments of the true distribution and then show formally that it is in fact the same as EL if the data are normal.

3.2 Preliminaries

3.2.1 Setup and assumptions

Consider a family of distributions $\{P_{\boldsymbol{\theta}}, \boldsymbol{\theta} \in \Theta \subset \mathbb{R}^p, \Theta \text{ compact}\}$ and a random vector $\mathbf{Z} \in \mathcal{Z} \subset \mathbb{R}^q$ from $P_{\boldsymbol{\theta}_o}, \boldsymbol{\theta}_o \in \Theta$, such that $\mathbb{E}\mathbf{Z} = 0$, $\mathbb{E}\{||\mathbf{Z}||^4\} < \infty$ and

$$\mathbb{E}\left[\mathbf{Z}\mathbf{Z}'\right] = \boldsymbol{\Sigma}(\boldsymbol{\theta}), \text{ if and only if } \boldsymbol{\theta} = \boldsymbol{\theta}_o. \tag{3.1}$$

Expectation is with respect to $P_{\boldsymbol{\theta}_o}$. The measurable real-valued matrix function $\boldsymbol{\Sigma}(\boldsymbol{\theta})$ comes from any structural model such as a factor analysis (FA) model, a random effects (RE) model, a simultaneous equations model (SEM), a conditional expectation model. For example, Ahn and Schmidt (1995) show how this setup arises in a dynamic panel setting.

For a random sample $(\mathbf{Z}_1, \ldots, \mathbf{Z}_N)$, where $\mathbf{Z}_i$ is measured as deviations from the

mean, denote

$$\mathbf{S}_i \equiv \mathbf{Z}_i \mathbf{Z}_i' \tag{3.2}$$

and

$$\mathbf{S} \equiv \frac{1}{N} \sum_{i=1}^{N} \mathbf{S}_i. \tag{3.3}$$

The problem is to estimate $\boldsymbol{\theta}_o$ given the random sample $(\mathbf{Z}_1, \ldots, \mathbf{Z}_N)$.

It is easy to see that since we assumed that the fourth moments exist, then $\mathbf{S}$ satisfies a central limit theorem. Thus we can write

$$vec(\mathbf{S}) \rightarrow N(vec(\boldsymbol{\Sigma}(\boldsymbol{\theta}_o)), \Delta(\boldsymbol{\theta}_o)),$$

where

$$\Delta(\boldsymbol{\theta}) = \mathbb{V}(vec(\mathbf{S}_i)) = \mathbb{E} vec(\mathbf{S_i}) vec(\mathbf{S_i})' - vec(\boldsymbol{\Sigma}(\boldsymbol{\theta})) vec(\boldsymbol{\Sigma}(\boldsymbol{\theta}))' \tag{3.4}$$

and *vec* denotes vertical vectorization of a matrix. To save space we will often omit the argument of matrix-functions and write $\boldsymbol{\Sigma}$ instead of $\boldsymbol{\Sigma}(\boldsymbol{\theta})$, $\boldsymbol{\Sigma}_o$ instead of $\boldsymbol{\Sigma}(\boldsymbol{\theta}_o)$, Δ_o instead of $\Delta(\boldsymbol{\theta}_o)$, etc.

It is well known (see, e.g., Magnus and Neudecker, 1988, p. 253) that for the multivariate normal distribution we have

$$\Delta_o = (\boldsymbol{\Sigma}_o \otimes \boldsymbol{\Sigma}_o)(\mathbb{I}_{q^2} + \Pi_{q^2}) = (\mathbb{I}_{q^2} + \Pi_{q^2})(\boldsymbol{\Sigma}_o \otimes \boldsymbol{\Sigma}_o), \tag{3.5}$$

where $\otimes$ denotes the Kronecker product, $\mathbb{I}_k$ is the identity matrix of dimension k, Π_{m^2} is the *commutation matrix*, i.e. such an $m^2 \times m^2$-matrix that $\boldsymbol{\Pi}_{m^2} vec(\mathbf{A}) =$

$vec(\mathbf{A}')$, for any $m \times m$ matrix $\mathbf{A}$.[1] Thus the fourth moments of the multivariate normal distribution are expressed in terms of the second moments.

We will also need certain smoothness conditions on $\mathbf{\Sigma}$. Such conditions combined with the above restrictions on $\mathbf{Z}$ are summarized in the following assumption.

Assumption 3.2.1 *(i)* $\mathbf{Z}_i \in \mathcal{Z} \subset \mathbb{R}^q, i = 1, \ldots, N$ *are iid from a distribution* $P_{\boldsymbol{\theta}_o}, \boldsymbol{\theta}_o \in \Theta \subset \mathbb{R}^p, \Theta$ *compact;*
(ii) $\mathbb{E}\mathbf{Z} = 0$, $\mathbb{E}\{||\mathbf{Z}||^4\} < \infty$ *and* $\mathbb{E}\left[\mathbf{Z}\mathbf{Z}'\right] = \mathbf{\Sigma}(\boldsymbol{\theta})$ *if and only if* $\boldsymbol{\theta} = \boldsymbol{\theta}_o$*;*
(iii) $\boldsymbol{\theta}_o \in int(\Theta)$ *and* $p \leq \frac{1}{2}q(q+1)$*;*
(iv) $vec(\mathbf{\Sigma})$ *is continuous at each* $\boldsymbol{\theta} \in \Theta$*;*
(v) $vec(\mathbf{\Sigma})$ *is three times continuously differentiable on a neighborhood* $\mathcal{N}$ *of* $\boldsymbol{\theta}_o$*.*

The following example shows that the simple setup in (3.1)-(3.3) can be used to represent fairly complex covariance structures.

3.2.2 An example

Consider the following *Linear Interdependent Structural Relationship* (LISREL) model pioneered by Karl Jöreskog (see, e.g., Jöreskog and Sörbom, 1977, p.287-

[1] One important property of the commutation matrix which also gave it its name is that it allows to interchange the two matrices in a Kronecker product while reversing the order of multiplication as in (3.5).

288)

$$\mathbf{Y} = \boldsymbol{\Lambda}_y \boldsymbol{\eta} + \boldsymbol{\epsilon}, \tag{3.6}$$

$$\mathbf{X} = \boldsymbol{\Lambda}_x \boldsymbol{\xi} + \boldsymbol{\delta}, \tag{3.7}$$

$$\mathbf{B}\boldsymbol{\eta} = \boldsymbol{\Gamma}\boldsymbol{\xi} + \boldsymbol{\gamma}, \tag{3.8}$$

where $\mathbf{Y}$ and $\mathbf{X}$ ($\dim \mathbf{X} + \dim \mathbf{Y} = q$) are measured as deviations from their means, $\boldsymbol{\eta}$ and $\boldsymbol{\xi}$ are common factors and $\boldsymbol{\epsilon}$ and $\boldsymbol{\delta}$ are unique factors such that $\mathbb{E}(\boldsymbol{\eta}) = \mathbf{0}$, $\mathbb{E}(\boldsymbol{\xi}) = \mathbf{0}$, $\mathbb{E}(\boldsymbol{\epsilon}) = \mathbf{0}$, $\mathbb{E}(\boldsymbol{\delta}) = \mathbf{0}$, $\mathbb{E}(\boldsymbol{\eta}\boldsymbol{\epsilon}') = \mathbf{0}$, $\mathbb{E}(\boldsymbol{\xi}\boldsymbol{\delta}') = \mathbf{0}$, $\mathbb{E}(\boldsymbol{\epsilon}\boldsymbol{\epsilon}') = \boldsymbol{\Theta}_\epsilon^2$, $\mathbb{E}(\boldsymbol{\delta}\boldsymbol{\delta}') = \boldsymbol{\Theta}_\delta^2$, $\mathbb{E}(\boldsymbol{\epsilon}\boldsymbol{\delta}') = \mathbf{0}$, where $\boldsymbol{\Theta}_\delta^2$ and $\boldsymbol{\Theta}_\epsilon^2$ are diagonal matrices.

Then, after some algebra, the covariance matrix of the observed variables $(\mathbf{Y}', \mathbf{X}')'$ can be written as follows:

$$\boldsymbol{\Sigma} = \begin{pmatrix} \boldsymbol{\Lambda}_y \boldsymbol{\Omega}_{\eta\eta} \boldsymbol{\Lambda}_y' + \boldsymbol{\Theta}_\epsilon^2 & \boldsymbol{\Lambda}_y \boldsymbol{\Omega}_{\eta\xi} \boldsymbol{\Lambda}_x' \\ \boldsymbol{\Lambda}_x \boldsymbol{\Omega}_{\xi\eta} \boldsymbol{\Lambda}_y' & \boldsymbol{\Lambda}_x \boldsymbol{\Omega}_{\xi\xi} \boldsymbol{\Lambda}_x' + \boldsymbol{\Theta}_\delta^2 \end{pmatrix}, \tag{3.9}$$

where

$$\begin{pmatrix} \boldsymbol{\Omega}_{\eta\eta} & \boldsymbol{\Omega}_{\eta\xi} \\ \boldsymbol{\Omega}_{\xi\eta} & \boldsymbol{\Omega}_{\xi\xi} \end{pmatrix} = \begin{pmatrix} \mathbf{B}^{-1}\boldsymbol{\Gamma}\boldsymbol{\Phi}\boldsymbol{\Gamma}'\mathbf{B}'^{-1} + \mathbf{B}^{-1}\boldsymbol{\Psi}\mathbf{B}'^{-1} & \mathbf{B}^{-1}\boldsymbol{\Gamma}\boldsymbol{\Phi} \\ \boldsymbol{\Phi}\boldsymbol{\Gamma}'\mathbf{B}'^{-1} & \boldsymbol{\Phi} \end{pmatrix}$$

and $\boldsymbol{\Phi} = \mathbb{E}(\boldsymbol{\xi}\boldsymbol{\xi}')$, $\boldsymbol{\Psi} = \mathbb{E}(\boldsymbol{\gamma}\boldsymbol{\gamma}')$.

If we let $\boldsymbol{\theta}$ denote the vector of all distinct parameters in $\boldsymbol{\Lambda}_y, \boldsymbol{\Lambda}_x, \mathbf{B}, \boldsymbol{\Gamma}, \boldsymbol{\Phi}, \boldsymbol{\Psi}, \boldsymbol{\Theta}_\epsilon^2, \boldsymbol{\Theta}_\delta^2$ and let $\mathbf{Z} = (\mathbf{Y}', \mathbf{X}')'$ we will obtain the setup of Section 3.2.1.

By imposing appropriate restrictions, the LISREL model reduces to many well-known models (see, e.g., Aigner et al., 1984). For example, equation (3.9) reduces

to a FA model if one imposes sufficient restrictions to retain only the upper-left block in the form $\mathbf{\Gamma\Phi\Gamma'+\Psi}$. From (3.6)-(3.8), SEM can be obtained by restricting $\mathbf{B=I}$, $\mathbf{\Theta}_\delta^2=\mathbf{\Theta}_\epsilon^2=\mathbf{0}$. To obtain a model for the conditional expectation of $\mathbf{Y}|\mathbf{X}$, one can restrict $\mathbf{\Lambda}_x$ to $\mathbb{I}$, $\mathbf{\Theta}_\delta^2$ to $\mathbf{0}$, and $\mathbf{\Phi}$ to the sample covariance matrix of $\mathbf{X}$. See Jöreskog (1970) for other special cases.

A well known special case of LISREL known as the multiple indicators multiple causes model (MIMIC) is obtained from (3.6)-(3.8) by setting $\mathbf{\Lambda}_x=\mathbb{I}$, $\mathbf{B}=\mathbb{I}$ and $\mathbf{\Theta}_\delta^2=\mathbf{0}$ (see, e.g., Jöreskog and Goldberger, 1975).

3.2.3 Estimators

3.2.3.1 Normal (Q)MLE

The normal QML estimator is

$$\hat{\boldsymbol{\theta}}_{\text{QMLE}}=\arg\max_{\theta\in\Theta}\sum_{i=1}^{N}\ln f(\mathbf{Z}_i,\boldsymbol{\theta}), \tag{3.10}$$

where

$$f(\mathbf{Z}_i,\boldsymbol{\theta})=\frac{1}{(2\pi)^{\frac{p}{2}}\sqrt{|\mathbf{\Sigma}|}}e^{-\frac{1}{2}\mathbf{Z}_i'\mathbf{\Sigma}^{-1}\mathbf{Z}_i}.$$

It is easy to see that the problem in (3.10) can be equivalently written as

$$\hat{\boldsymbol{\theta}}_{\text{QMLE}}=\arg\min_{\theta\in\Theta}F_{\text{MLE}}(\boldsymbol{\theta}),$$

where

$$F_{\text{MLE}}(\boldsymbol{\theta}) = \log|\boldsymbol{\Sigma}| + \text{tr}(\mathbf{S}\boldsymbol{\Sigma}^{-1}). \tag{3.11}$$

Thus QMLE amounts to finding the value of $\boldsymbol{\theta}$ that minimize distance (3.11) between the sample covariance matrix $\mathbf{S}$ and the covariance matrix $\boldsymbol{\Sigma}$ imposed by the model.

It is a standard result (see, e.g., Chamberlain, 1984, p. 1289) that, under Assumption 3.2.1, the normal QMLE of $\boldsymbol{\theta}_o$ is consistent and asymptotically normal.

3.2.3.2 GMM

The optimal GMM estimator of $\boldsymbol{\theta}_o$ is based on the distinct elements of (3.1), i.e. on the moment conditions

$$\mathbb{E}[\mathbf{m}(\mathbf{Z}_i; \boldsymbol{\theta}_o)] = \mathbf{0}, \tag{3.12}$$

where $\mathbf{m}(\mathbf{Z}_i; \boldsymbol{\theta}) = vech(\mathbf{S}_i) - vech(\boldsymbol{\Sigma})$ and $vech$ denotes vertical vectorization of the lower triangle of a matrix. Thus $\mathbf{m}$ is a $\frac{1}{2}q(q+1)$-vector.

The optimal GMM estimator of $\boldsymbol{\theta}_o$ is obtained as the solution to the following problem:

$$\hat{\boldsymbol{\theta}}_{\text{GMM}} = \arg\min_{\theta \in \Theta} F_{\text{GMM}}(\boldsymbol{\theta}),$$

where

$$F_{\text{GMM}}(\boldsymbol{\theta}) = \mathbf{m}_N(\boldsymbol{\theta})' \mathbf{W} \mathbf{m}_N(\boldsymbol{\theta}), \tag{3.13}$$

$$\begin{aligned}\mathbf{m}_N(\boldsymbol{\theta}) &= \frac{1}{N}\sum_{i=1}^{N}\mathbf{m}(\mathbf{z}_i;\boldsymbol{\theta}) \\ &= vech(\mathbf{S}) - vech(\boldsymbol{\Sigma}),\end{aligned}$$

and $\mathbf{W}$ is the appropriate (optimal) weighting matrix.

The optimal weighting matrix is

$$\mathbf{W}_o = \{\mathbb{E}[\mathbf{m}(\boldsymbol{Z}_i;\boldsymbol{\theta}_o)\mathbf{m}(\boldsymbol{Z}_i;\boldsymbol{\theta}_o)']\}^{-1}. \tag{3.14}$$

But in (3.13), one would typically use the following consistent estimator of $\mathbf{W}_o$ based on a preliminary consistent estimate of $\boldsymbol{\theta}_o$

$$\widehat{\mathbf{W}} = \left\{\frac{1}{N}\sum_{t=1}^{N}[\mathbf{m}(\boldsymbol{z}_i;\hat{\boldsymbol{\theta}})\mathbf{m}(\boldsymbol{z}_i;\hat{\boldsymbol{\theta}})']\right\}^{-1}.$$

Note that there is a connection between $\mathbf{W}$ in (3.14) and $\boldsymbol{\Delta}$ in (3.4). To show the connection we need to define matrices that transform *vech* into *vec* and vice versa. Magnus and Neudecker (1988, p. 49) show that, for a symmetric $k \times k$ matrix $\mathbf{A}$ there exists a unique $k^2 \times \frac{k(k+1)}{2}$ *duplication* matrix $\mathbf{H}_k$ such that $\mathbf{H}_k\, vech(\mathbf{A}) = vec(\mathbf{A})$. Thus $\mathbf{H}_k$ transforms *vech* into *vec*, while the Moore-Penrose inverse of $\mathbf{H}_k$, $\bar{\mathbf{H}}_k = (\mathbf{H}_k'\mathbf{H}_k)^{-1}\mathbf{H}_k'$, transforms *vec* into *vech*. Matrices $\mathbf{H}_k$ and $\bar{\mathbf{H}}_k$ have the following properties:

(i) $\bar{\mathbf{H}}_k\,\mathbf{H}_k = \mathbb{I}_{\frac{k(k+1)}{2}}$;

(ii) $\boldsymbol{\Pi}_{k^2}\,\mathbf{H}_k = \mathbf{H}_k$, where $\boldsymbol{\Pi}_{k^2}$ is the commutation matrix defined above;

(iii) $\mathbf{H}_k\,\bar{\mathbf{H}}_k = \frac{1}{2}(\mathbb{I}_{k^2} + \boldsymbol{\Pi}_{k^2})$;

(iv) $(\mathbb{I}_{k^2} + \mathbf{\Pi}_{k^2})\,\mathbf{H}_k = 2\,\mathbf{H}_k$ and $\bar{\mathbf{H}}_k\,(\mathbb{I}_{k^2} + \mathbf{\Pi}_{k^2}) = 2\,\bar{\mathbf{H}}_k$.

Thus, omitting the dimensionality subscript, we can write $\mathbf{\Delta} = \mathbb{V}[vec(\mathbf{S}_i)] = \mathbb{V}[\mathbf{H}\,vech(\mathbf{S}_i)] = \mathbf{H}\mathbb{V}[vech(\mathbf{S}_i)]\mathbf{H}'$. But $\mathbb{V}[vech(\mathbf{S}_i)] = \mathbb{E}[\mathbf{m}(\boldsymbol{Z}_i;\boldsymbol{\theta})\mathbf{m}(\boldsymbol{Z}_i;\boldsymbol{\theta})']$. We can therefore write the optimal weighting matrix in (3.14) as $[\bar{\mathbf{H}}\mathbf{\Delta}_o\bar{\mathbf{H}}']^{-1}$.

It is easy to verify that, under Assumption 3.2.1 and with $\widehat{\mathbf{W}} \xrightarrow{p} \mathbf{W}$, the standard conditions for consistency and asymptotically normality of the GMM estimator of $\boldsymbol{\theta}_o$ hold (see, e.g., Newey and McFadden, 1994, Theorems 2.6 and 3.4).

3.2.3.3 EL

The EL estimator of $\boldsymbol{\theta}_o$ is obtained as follows:

$$\hat{\boldsymbol{\theta}}_{\mathrm{EL}} = \arg\max_{\theta\in\Theta} \sum_{i=1}^{N} \ln \pi_i$$

subject to

$$\sum_{i=1}^{N} \pi_i \mathbf{m}(\mathbf{z}_i;\boldsymbol{\theta}) = 0$$

and

$$\sum_{i=1}^{N} \pi_i = 1$$

It can also be shown that Assumption 3.2.1 is sufficiently strong to satisfy the conditions for consistency and asymptotic normality of $\hat{\boldsymbol{\theta}}_{\mathrm{EL}}$ (see, e.g., Kitamura, 1997; Owen, 2001).

3.3 First order analysis

3.3.1 The first order conditions

Let $\mathbf{G}(\boldsymbol{\theta})$ denote the *Jacobian* matrix of the moment functions in (3.12). Then

$$\mathbf{G} \equiv \mathbf{G}(\boldsymbol{\theta}) = \frac{\partial \mathbf{m}(\mathbf{z}_i, \boldsymbol{\theta})}{\partial \boldsymbol{\theta}'} = \frac{\partial vech(\boldsymbol{\Sigma})}{\partial \boldsymbol{\theta}'}.$$

The following lemmas are used in derivation of the main results of the paper. They are well known and thus given without proof (see, e.g., Chamberlain, 1984; Hansen, 1982; Qin and Lawless, 1994, for some relevant proofs).

Lemma 3.3.1 *Under Assumption 3.2.1, the first order condition for* $\hat{\boldsymbol{\theta}}_{\text{QMLE}}$ *is*

$$\mathbf{G}'\mathbf{H}'(\boldsymbol{\Sigma} \otimes \boldsymbol{\Sigma})^{-1}\mathbf{H}\,[vech(\boldsymbol{S}) - vech(\boldsymbol{\Sigma})] = 0. \tag{3.15}$$

Lemma 3.3.2 *Under Assumption 3.2.1, the first order condition for* $\hat{\boldsymbol{\theta}}_{\text{GMM}}$ *is*

$$\mathbf{G}'\widehat{\mathbf{W}}^{-1}[vech(\boldsymbol{S}) - vech(\boldsymbol{\Sigma})] = 0. \tag{3.16}$$

Lemma 3.3.3 *Under Assumption 3.2.1, the first order condition for* $\hat{\boldsymbol{\theta}}_{\text{EL}}$ *is*

$$\mathbf{G}'\left[\sum_{i=1}^{N} \pi_i \mathbf{m}_i \mathbf{m}_i'\right]^{-1} [vech(\boldsymbol{S}) - vech(\boldsymbol{\Sigma})] = 0, \tag{3.17}$$

where $\mathbf{m}_i = \mathbf{m}(\mathbf{Z}_i; \boldsymbol{\theta})$.

In Section 3.4, we will use an alternative way of writing the first order conditions that circumvents the need to operate with the inverse. Define $\boldsymbol{\lambda} = -[\boldsymbol{\Sigma}(\boldsymbol{\theta}) \otimes \boldsymbol{\Sigma}(\boldsymbol{\theta})]^{-1}\mathbf{H}\,\mathbf{m}_N(\boldsymbol{\theta})$. Then the QMLE first order condition can be written as

$$\mathbf{s}_N(\boldsymbol{\beta}) = -\begin{bmatrix} \mathbf{G}(\boldsymbol{\theta})'\mathbf{H}'\boldsymbol{\lambda} \\ \mathbf{H}\,\mathbf{m}_N(\boldsymbol{\theta}) + [\boldsymbol{\Sigma}(\boldsymbol{\theta}) \otimes \boldsymbol{\Sigma}(\boldsymbol{\theta})]\boldsymbol{\lambda} \end{bmatrix} = 0$$

and we now have a $p+q^2$-vector of parameters $\boldsymbol{\beta} = (\boldsymbol{\theta}', \boldsymbol{\lambda}')'$. A similar representation of the GMM and EL first order conditions was used, for example, by Newey and Smith (2004).

It is clear from (3.15)-(3.17) that the only thing that distinguishes the three estimators is the way in which the empirical moments $\mathbf{m}_N(\boldsymbol{\theta})$ are weighted. One way to compare the first order variances of GMM and QMLE is to note that $\hat{\boldsymbol{\theta}}_{\text{QMLE}}$ comes from the GMM problem that employs a suboptimal weighting matrix $\mathbf{H}'(\boldsymbol{\Sigma} \otimes \boldsymbol{\Sigma})^{-1}\mathbf{H}$ and is therefore inferior to $\hat{\boldsymbol{\theta}}_{\text{GMM}}$ in terms of first-order asymptotic relative efficiency. However, that argument cannot be used to derive the equal efficiency condition.

3.3.2 Relative efficiency to the first order

Theorem 3.3.1 *Suppose Assumption 3.2.1 holds. Let* $\mathbb{V}$ *denote the first order asymptotic variance matrix of the relevant estimator, i.e.* $\mathbb{V} = Avar[N^{\frac{1}{2}}(\hat{\boldsymbol{\theta}} - \boldsymbol{\theta}_o)]$.

Then,

$$\begin{aligned}
\mathbb{V}_{\text{QMLE}} &= [\mathbf{G}_o'\mathbf{H}'(\boldsymbol{\Sigma}_o \otimes \boldsymbol{\Sigma}_o)^{-1}\mathbf{H}\mathbf{G}_o]^{-1} \\
&\quad \times \mathbf{G}_o'\mathbf{H}'(\boldsymbol{\Sigma}_o \otimes \boldsymbol{\Sigma}_o)^{-1}\boldsymbol{\Delta}_o(\boldsymbol{\Sigma}_o \otimes \boldsymbol{\Sigma}_o)^{-1}\mathbf{H}\mathbf{G}_o \qquad (3.18) \\
&\quad \times [\mathbf{G}_o'\mathbf{H}'(\boldsymbol{\Sigma}_o \otimes \boldsymbol{\Sigma}_o)^{-1}\mathbf{H}\mathbf{G}_o]^{-1}, \\
\mathbb{V}_{\text{GMM}} &= \mathbb{V}_{\text{EL}} = [\mathbf{G}_o'(\bar{\mathbf{H}}\boldsymbol{\Delta}_o\bar{\mathbf{H}}')^{-1}\mathbf{G}_o]^{-1}. \qquad (3.19)
\end{aligned}$$

Proof. See Chamberlain (1984, p. 1295) for derivation of (3.18); see Hansen (1982, p. 1048) and Qin and Lawless (1994, p. 306) for derivation of (3.19). □

If the data are multivariate normal then it is easy to show that the above expressions for variance are the same. More specifically, on using properties of the duplication matrix and equation (3.5) without the dimensionality subscript, the following simplifications apply:

$$\begin{aligned}
\mathbf{H}'(\boldsymbol{\Sigma} \otimes \boldsymbol{\Sigma})^{-1}\boldsymbol{\Delta}(\boldsymbol{\Sigma} \otimes \boldsymbol{\Sigma})^{-1}\mathbf{H} &= \mathbf{H}'(\boldsymbol{\Sigma} \otimes \boldsymbol{\Sigma})^{-1}(\mathbb{I} + \Pi)(\boldsymbol{\Sigma} \otimes \boldsymbol{\Sigma})(\boldsymbol{\Sigma} \otimes \boldsymbol{\Sigma})^{-1}\mathbf{H} \\
&= \mathbf{H}'(\boldsymbol{\Sigma} \otimes \boldsymbol{\Sigma})^{-1}(\mathbb{I} + \Pi)\mathbf{H} \\
&= 2\,\mathbf{H}'(\boldsymbol{\Sigma} \otimes \boldsymbol{\Sigma})^{-1}\mathbf{H},
\end{aligned}$$

$$\begin{aligned}
\bar{\mathbf{H}}\boldsymbol{\Delta}\bar{\mathbf{H}}' &= \bar{\mathbf{H}}(\mathbb{I} + \Pi)(\boldsymbol{\Sigma} \otimes \boldsymbol{\Sigma})\bar{\mathbf{H}}' \\
&= 2\,\bar{\mathbf{H}}(\boldsymbol{\Sigma} \otimes \boldsymbol{\Sigma})\bar{\mathbf{H}}'.
\end{aligned}$$

To see that $[\bar{\mathbf{H}}'(\boldsymbol{\Sigma}\otimes\boldsymbol{\Sigma})\bar{\mathbf{H}}]^{-1}$ is equal to $\mathbf{H}'(\boldsymbol{\Sigma}\otimes\boldsymbol{\Sigma})^{-1}\mathbf{H}$, note that

$$\begin{aligned}
\mathbf{H}'(\boldsymbol{\Sigma}\otimes\boldsymbol{\Sigma})^{-1}\mathbf{H}\,\bar{\mathbf{H}}(\boldsymbol{\Sigma}\otimes\boldsymbol{\Sigma})\bar{\mathbf{H}}' &= \frac{1}{2}\mathbf{H}'(\boldsymbol{\Sigma}\otimes\boldsymbol{\Sigma})^{-1}(\mathbb{I}+\Pi)(\boldsymbol{\Sigma}\otimes\boldsymbol{\Sigma})\bar{\mathbf{H}}' \\
&= \frac{1}{2}\mathbf{H}'(\boldsymbol{\Sigma}\otimes\boldsymbol{\Sigma})^{-1}(\boldsymbol{\Sigma}\otimes\boldsymbol{\Sigma})(\mathbb{I}+\Pi)\bar{\mathbf{H}}' \\
&= \frac{1}{2}\mathbf{H}'(\mathbb{I}+\Pi)\bar{\mathbf{H}}' \\
&= \frac{1}{2}2\mathbf{H}'\bar{\mathbf{H}}' \\
&= \mathbb{I}.
\end{aligned}$$

Equation (3.19) of Theorem 3.3.1 states that the asymptotic variance matrices of optimal GMM and EL are equal, i.e. the two estimators of $\boldsymbol{\theta}_o$ are asymptotically equivalent to the first order. It is not immediately clear from only the form of (3.18) that QMLE is dominated by the other two estimators.

The main first-order asymptotic result of this paper is stated in the next theorem.

Theorem 3.3.2 *Let Assumption 3.2.1 hold. The estimators* $\hat{\boldsymbol{\theta}}_{\mathrm{GMM}}$ *and* $\hat{\boldsymbol{\theta}}_{\mathrm{EL}}$ *are no less asymptotically efficient to the first order than* $\hat{\boldsymbol{\theta}}_{\mathrm{QMLE}}$. *Equal first-order efficiency occurs under the following equivalent conditions:*

(i) $\mathbf{G}_o$ *is in the column space of* $\bar{\mathbf{H}}\boldsymbol{\Delta}_o(\boldsymbol{\Sigma}_o\otimes\boldsymbol{\Sigma}_o)^{-1}\mathbf{H}\mathbf{G}_o$;

(ii) There exists a $\frac{q(q+1)}{2}\times\frac{q(q+1)}{2}$ *matrix* $\mathbb{D}$ *such that*

$$\mathbf{G}_o = \bar{\mathbf{H}}\boldsymbol{\Delta}_o(\boldsymbol{\Sigma}_o\otimes\boldsymbol{\Sigma}_o)^{-1}\mathbf{H}\mathbf{G}_o\mathbb{D}.$$

Proof. $\mathbb{V}_{\text{QMLE}} - \mathbb{V}_{\text{GMM}}$ is positive semidefinite (PSD) if and only if $\mathbb{V}_{\text{GMM}}^{-1} - \mathbb{V}_{\text{QMLE}}^{-1}$ is PSD. Denote $\bar{\mathbf{H}}\boldsymbol{\Delta}_o\bar{\mathbf{H}}'$ by $\mathbb{C}$ and $\mathbf{H}'(\boldsymbol{\Sigma}_o \otimes \boldsymbol{\Sigma}_o)^{-1}\mathbf{H}$ by $\mathbb{A}$. We have

$$\begin{aligned}\mathbb{V}_{\text{GMM}}^{-1} - \mathbb{V}_{\text{QMLE}}^{-1} &= \mathbf{G}_o'\mathbb{C}^{-1}\mathbf{G}_o - \mathbf{G}_o'\mathbb{A}\mathbf{G}_o[\mathbf{G}_o'\mathbb{A}\mathbb{C}\mathbb{A}\mathbf{G}_o]^{-1}\mathbf{G}_o'\mathbb{A}\mathbf{G}_o \\ &= \mathbf{G}_o'\mathbb{C}^{-\frac{1}{2}}[\mathbb{I} - \mathbb{C}^{\frac{1}{2}}\mathbb{A}\mathbf{G}_o[\mathbf{G}_o'\mathbb{A}\mathbb{C}^{\frac{1}{2}}\mathbb{C}^{\frac{1}{2}}\mathbb{A}\mathbf{G}_o]^{-1}\mathbf{G}_o'\mathbb{A}\mathbb{C}^{\frac{1}{2}}]\mathbb{C}^{-\frac{1}{2}}\mathbf{G}_o.\end{aligned}$$

This is PSD because the middle part is the idempotent projection matrix onto $\mathbb{C}^{1/2}\mathbb{A}\mathbf{G}_o$. This proves the first part of the theorem.

The difference is zero if and only if $\mathbb{C}^{-1/2}\mathbf{G}_o$ is in the column space spanned by $\mathbb{C}^{1/2}\mathbb{A}\mathbf{G}_o$, or equivalently, $\mathbf{G}_o$ is in the column space of $\mathbb{C}\mathbb{A}\mathbf{G}_o$. Note that

$$\begin{aligned}\mathbb{C}\mathbb{A}\mathbf{G}_o &= \bar{\mathbf{H}}\boldsymbol{\Delta}_o\bar{\mathbf{H}}'\mathbf{H}'(\boldsymbol{\Sigma}_o \otimes \boldsymbol{\Sigma}_o)^{-1}\mathbf{H}\mathbf{G}_o \\ &= \bar{\mathbf{H}}\boldsymbol{\Delta}_o\frac{1}{2}(\mathbb{I} + \boldsymbol{\Pi})(\boldsymbol{\Sigma}_o \otimes \boldsymbol{\Sigma}_o)^{-1}\mathbf{H}\mathbf{G}_o \\ &= \bar{\mathbf{H}}\boldsymbol{\Delta}_o(\boldsymbol{\Sigma}_o \otimes \boldsymbol{\Sigma}_o)^{-1}\frac{1}{2}(\mathbb{I} + \boldsymbol{\Pi})\mathbf{H}\mathbf{G}_o \\ &= \bar{\mathbf{H}}\boldsymbol{\Delta}_o(\boldsymbol{\Sigma}_o \otimes \boldsymbol{\Sigma}_o)^{-1}\frac{1}{2}2\mathbf{H}\mathbf{G}_o \\ &= \bar{\mathbf{H}}\boldsymbol{\Delta}_o(\boldsymbol{\Sigma}_o \otimes \boldsymbol{\Sigma}_o)^{-1}\mathbf{H}\mathbf{G}_o.\end{aligned}$$

This proves both (i) and (ii). □

Theorem 3.3.2 is novel in that it states the first order efficiency properties of QMLE, GMM and EL explicitly in terms of the fourth moments $\boldsymbol{\Delta}$ of the distribution. It is clear from the theorem, that GMM and EL dominate QMLE because they make efficient use of the second moment information without imposing restrictions on the fourth moments. Ahn and Schmidt (1995, Appendix 2) showed that the GMM estimator of covariance structures reaches the semiparametric efficiency bound of

Newey (1990). Theorem 3.3.2 provides an explicit expression for the gain attained by GMM over QMLE.

Not surprisingly, the conditions of Theorem 3.3.2 hold for the multivariate normal distribution. Using (3.5), one can write

$$\bar{\mathbf{H}}\boldsymbol{\Delta}_o(\boldsymbol{\Sigma}_o \otimes \boldsymbol{\Sigma}_o)^{-1}\mathbf{H}\mathbf{G}_o = \bar{\mathbf{H}}(\mathbb{I} - \boldsymbol{\Pi})\mathbf{H}\mathbf{G}_o = 2\,\bar{\mathbf{H}}\mathbf{H}\mathbf{G}_o = 2\,\mathbf{G}_o.$$

So condition (ii) trivially holds. However, there may conceivably exist other distributions that satisfy the equal first order asymptotic efficiency conditions of Theorem 3.3.2. We leave further exploration of this point for future work.

3.4 Second order analysis

3.4.1 Stochastic expansions to the second order

Higher order stochastic expansions are based on the Taylor approximation of the first order conditions at the true value. The expansions have the following form

$$\sqrt{N}(\hat{\boldsymbol{\beta}} - \boldsymbol{\beta}_o) = \boldsymbol{\mu} + N^{-\frac{1}{2}}\boldsymbol{\tau} + O_p(N^{-1}), \tag{3.20}$$

where $\boldsymbol{\mu}$ and $\boldsymbol{\tau}$ are $O_p(1)$ random vectors.

It is well known that the first order bias can be obtained by taking the expectation of the first term. Since QMLE, GMM, and EL are $\sqrt{N}$ consistent, their first order bias is zero. Similarly, the first order variances can be obtained as the expectation

of the outer product of first term. The second order bias is based on the expectation of the first two terms in (3.20). Alternatively, the second order bias can be obtained using the Edgeworth approximation to the distribution as in Rothenberg (1984) and McCullagh (1987).

General expressions for $\boldsymbol{\mu}$ and $\boldsymbol{\tau}$ of extremum and minimum distance estimators with many examples can be found in Rilstone et al. (1996); Bao and Ullah (2003); Ullah (2004); Kim (2005). Specialized expressions for the GMM and (generalized) EL can be found in Newey et al. (2003) and Newey and Smith (2004).

Derivation of higher order stochastic expansions involves higher order derivatives of the objective functions. Rilstone et al. (1996) use a recursive definition of derivatives which is useful in general settings. In our derivation we follow Newey and Smith (2004) in using the usual definition because we do not go to the order higher than two and because we wish to compare the QMLE bias to the GMM and EL bias expressions they derive.

Define

$$
\begin{aligned}
\mathbf{s}_i(\boldsymbol{\beta}) &= -\begin{bmatrix} \mathbf{G}'\mathbf{H}'\boldsymbol{\lambda} \\ \mathbf{H}\,\mathbf{m}_i + (\boldsymbol{\Sigma}\otimes\boldsymbol{\Sigma})\boldsymbol{\lambda} \end{bmatrix}, \\
\mathbf{M}_j &= \mathbb{E}\frac{\partial^2 \mathbf{s}_i(\boldsymbol{\beta}_o)}{\partial\boldsymbol{\beta}'\partial\boldsymbol{\beta}_j}, \quad \text{where } \boldsymbol{\beta}_o = (\boldsymbol{\theta}_o', \mathbf{0}')', \\
\mathbf{R} &= [\mathbf{G}'\mathbf{H}'(\boldsymbol{\Sigma}\otimes\boldsymbol{\Sigma})^{-1}\mathbf{H}\mathbf{G}]^{-1}, \\
\mathbf{Q} &= \mathbf{R}\mathbf{G}'\mathbf{H}'(\boldsymbol{\Sigma}\otimes\boldsymbol{\Sigma})^{-1}, \\
\mathbf{P} &= (\boldsymbol{\Sigma}\otimes\boldsymbol{\Sigma})^{-1} - (\boldsymbol{\Sigma}\otimes\boldsymbol{\Sigma})^{-1}\mathbf{H}\mathbf{G}\mathbf{Q}.
\end{aligned}
$$

Theorem 3.4.1 *Under Assumption 3.2.1, the estimator* $\boldsymbol{\beta}_{\mathrm{QMLE}}$ *satisfies (3.20) with*

$$\begin{aligned} \boldsymbol{\mu} &= \begin{bmatrix} \mathbf{Q}_o \\ \mathbf{P}_o \end{bmatrix} \mathbf{H} \frac{1}{\sqrt{N}} \sum_{i=1}^{N} [vech(\mathbf{S}_i) - vech(\boldsymbol{\Sigma}_o)], \\ \boldsymbol{\tau} &= 1/2 \begin{bmatrix} -\mathbf{R}_o & \mathbf{Q}_o \\ \mathbf{Q}_o' & \mathbf{P}_o \end{bmatrix} \sum_{j=1}^{p+q^2} \boldsymbol{\mu}_j \mathbf{M}_j \boldsymbol{\mu}, \end{aligned} \tag{3.21}$$

where $\boldsymbol{\mu}_j$ *is the* j*-th element of* $\boldsymbol{\mu}$*.*

Proof. See Appendix B. □

Note that $\mathbb{E}\boldsymbol{\mu} = 0$ and the first order variance of $\boldsymbol{\beta}_{\mathrm{QMLE}}$ based on (3.21) can be written as

$$\begin{aligned} \mathbb{E}\boldsymbol{\mu}\boldsymbol{\mu}' &= \begin{bmatrix} \mathbf{Q}_o \\ \mathbf{P}_o \end{bmatrix} \mathbf{H}\, \mathbb{E}[\boldsymbol{m}(\boldsymbol{Z}_i; \boldsymbol{\theta}_o)\boldsymbol{m}(\boldsymbol{Z}_i; \boldsymbol{\theta}_o)']\mathbf{H}' \begin{bmatrix} \mathbf{Q}_o \\ \mathbf{P}_o \end{bmatrix}' \\ &= \begin{bmatrix} \mathbf{Q}_o\boldsymbol{\Delta}_o\mathbf{Q}_o' & \mathbf{Q}_o\boldsymbol{\Delta}_o\mathbf{P}_o' \\ \mathbf{P}_o\boldsymbol{\Delta}_o\mathbf{Q}_o' & \mathbf{P}_o\boldsymbol{\Delta}_o\mathbf{P}_o' \end{bmatrix}, \end{aligned} \tag{3.22}$$

where the upper left $p \times p$ block of (3.22) represents the first order asymptotic variance of $\boldsymbol{\theta}_{\mathrm{QMLE}}$ in (3.18).

Interestingly, the matrix in (3.22) is not in general block diagonal unlike the EL and GMM analogues (see, e.g., Qin and Lawless, 1994, Theorem 1). However, in

the case of multivariate normality, the blocks of (3.22) can be simplified as follows

$$\begin{aligned} \mathbf{Q}\boldsymbol{\Delta}\mathbf{Q}' &= 2\mathbf{R}, \\ \mathbf{Q}\boldsymbol{\Delta}\mathbf{P}' &= \mathbf{0}, \\ \mathbf{P}\boldsymbol{\Delta}\mathbf{P}' &= (\mathbb{I} + \boldsymbol{\Pi})\mathbf{P}. \end{aligned} \tag{3.23}$$

And thus $\hat{\boldsymbol{\theta}}_{\mathrm{QMLE}}$ and $\hat{\boldsymbol{\lambda}}_{\mathrm{QMLE}}$ are in this case asymptotically uncorrelated.

3.4.2 Second order bias of QMLE

Let $\mathbb{B}$ denote the second order bias of the relevant estimator. Using (3.20), the bias can be written in terms of the expected value of $\boldsymbol{\tau}$ as

$$\mathbb{B} = \mathbb{E}\boldsymbol{\tau}/N.$$

Thus, an explicit form of the QMLE bias contains $\mathbb{E}\boldsymbol{\mu}_j\mathbf{M}_j\boldsymbol{\mu}$, $j = 1, \ldots, p + q^2$. But $\mathbf{M}_j$ can be written as

$$\begin{aligned} \mathbf{M}_j &= -\mathbb{E}\,\frac{\partial^2}{\partial\boldsymbol{\beta}'\partial\boldsymbol{\beta}_j}\begin{bmatrix} \mathbf{G}'\mathbf{H}'\boldsymbol{\lambda} \\ \mathbf{H}\,\mathbf{m}_i + (\boldsymbol{\Sigma}\otimes\boldsymbol{\Sigma})\boldsymbol{\lambda} \end{bmatrix}\Bigg|_{\boldsymbol{\theta}=\boldsymbol{\theta}_o,\boldsymbol{\lambda}=0} \\ &= \begin{cases} -\begin{bmatrix} \mathbf{0} & \mathbf{G}_o^{j\prime}\mathbf{H}' \\ \mathbf{H}\mathbf{G}_o^j & \boldsymbol{\Omega}_o^j \end{bmatrix}, & j = 1, \ldots, p \\ -\begin{bmatrix} \mathbf{G}_{oj} & \mathbf{0} \\ \boldsymbol{\Omega}_{oj} & \mathbf{0} \end{bmatrix}, & j = 1, \ldots, q^2 \end{cases} \end{aligned}$$

where $\mathbf{G}_o^j = \frac{\partial}{\partial \theta_j}\mathbf{G}_o$, $\mathbf{G}_{oj} = \frac{\partial}{\partial \boldsymbol{\theta}}[\mathbf{G}_o'\mathbf{H}']_j$, $\boldsymbol{\Omega}_o^j = \frac{\partial}{\partial \theta_j}(\boldsymbol{\Sigma}_o \otimes \boldsymbol{\Sigma}_o)$ and $\boldsymbol{\Omega}_{oj} = \frac{\partial}{\partial \boldsymbol{\theta}}[\boldsymbol{\Sigma}_o \otimes \boldsymbol{\Sigma}_o]_j$. Therefore $\mathbf{M}_j$ is non-random and we can write

$$\mathbb{E}\boldsymbol{\mu}_j\mathbf{M}_j\boldsymbol{\mu} = \begin{cases} -\begin{bmatrix} \mathbf{0} & \mathbf{G}_o^{j\prime}\mathbf{H}' \\ \mathbf{H}\mathbf{G}_o^j & \boldsymbol{\Omega}_o^j \end{bmatrix} \mathbb{E}\boldsymbol{\mu}\boldsymbol{\mu}'\mathbf{e}_j, & j = 1, \ldots, p \\ -\begin{bmatrix} \mathbf{G}_{oj} & \mathbf{0} \\ \boldsymbol{\Omega}_{oj} & \mathbf{0} \end{bmatrix} \mathbb{E}\boldsymbol{\mu}\boldsymbol{\mu}'\mathbf{e}_{p+j}, & j = 1, \ldots, q^2, \end{cases} \tag{3.24}$$

where $\mathbf{e}_k$ is a $p+q^2$-vector of zeros with the k-th element equal to 1. Substituting (3.22) into (3.24) and simplifying yields the result of the following theorem.

Theorem 3.4.2 *Under Assumption 3.2.1, the second order bias of* $\boldsymbol{\beta}_{\text{QMLE}}$ *can be written as follows*

$$\begin{aligned} \mathbb{B}_{\text{QMLE}} = -\frac{1}{2N}\begin{bmatrix} -\mathbf{R}_o & \mathbf{Q}_o \\ \mathbf{Q}_o' & \mathbf{P}_o \end{bmatrix} \Bigg\{ \sum_{j=1}^{p} \begin{bmatrix} \mathbf{0} & \mathbf{G}_o^{j\prime}\mathbf{H}' \\ \mathbf{H}\mathbf{G}_o^j & \boldsymbol{\Omega}_o^j \end{bmatrix} \begin{bmatrix} \mathbf{Q}_o\boldsymbol{\Delta}_o\mathbf{Q}_o' \\ \mathbf{P}_o\boldsymbol{\Delta}_o\mathbf{Q}_o' \end{bmatrix} \mathbf{e}_j \\ + \sum_{j=1}^{q^2} \begin{bmatrix} \mathbf{G}_{oj} \\ \boldsymbol{\Omega}_{oj} \end{bmatrix} \mathbf{Q}_o\boldsymbol{\Delta}_o\mathbf{P}_o'\mathbf{e}_j \Bigg\}, \end{aligned} \tag{3.25}$$

where $\mathbf{e}_k$ *is the zero vector of relevant dimension in which the k-th element is 1.*

McCullagh (1987) and Linton (1997) give expressions for the second order bias of QMLE in terms of cumulants; we use the higher-moment representation to enable comparison with second order biases derived in Newey and Smith (2004).

Based on (3.23), the following simplification applies in the multivariate normal case:

$$\begin{aligned}
\mathbb{B}_{\text{QMLE}} &= -\frac{1}{2N}\begin{bmatrix} -\mathbf{R}_o & \mathbf{Q}_o \\ \mathbf{Q}_o' & \mathbf{P}_o \end{bmatrix}\left\{\sum_{j=1}^{p}\begin{bmatrix} \mathbf{0} & \mathbf{G}_o^{j\prime}\mathbf{H}' \\ \mathbf{H}\mathbf{G}_o^{j} & \boldsymbol{\Omega}_o^{j} \end{bmatrix}\begin{bmatrix} 2\mathbf{R}_o \\ \mathbf{0} \end{bmatrix}\mathbf{e}_j\right\} \\
&= -\frac{1}{2N}\begin{bmatrix} -\mathbf{R}_o & \mathbf{Q}_o \\ \mathbf{Q}_o' & \mathbf{P}_o \end{bmatrix}\begin{bmatrix} \mathbf{0} \\ 2\sum_{j=1}^{p}\mathbf{H}\mathbf{G}_o^{j}\mathbf{R}_o\mathbf{e}_j \end{bmatrix} \\
&= -\frac{1}{N}\begin{bmatrix} \mathbf{Q}_o\mathbf{H}\sum_{j=1}^{p}\mathbf{G}_o^{j}\mathbf{R}_o\mathbf{e}_j \\ \mathbf{P}_o\mathbf{H}\sum_{j=1}^{p}\mathbf{G}_o^{j}\mathbf{R}_o\mathbf{e}_j \end{bmatrix}. \qquad (3.26)
\end{aligned}$$

3.4.3 Comparison to GMM and EL

Newey and Smith's (2004, Theorems 4.1 and 4.6) second order biases of GMM and EL of $\boldsymbol{\theta}_o$ are, in our notation,

$$\begin{aligned}
\mathbb{B}_{\text{EL}}(\boldsymbol{\theta}) &= -\frac{1}{2N}\mathbf{Q}_o^{\text{EL}}\sum_{j=1}^{p}\mathbf{G}_o^{j}\mathbf{R}_o^{\text{EL}}\mathbf{e}_j, \qquad (3.27) \\
\mathbb{B}_{\text{GMM}}(\boldsymbol{\theta}) &= \mathbb{B}_{\text{EL}} + \frac{1}{N}\mathbf{Q}_o^{\text{EL}}\mathbf{U}_o,
\end{aligned}$$

where

$$\begin{aligned}
\mathbf{Q}^{\text{EL}} &= \mathbf{R}^{\text{EL}}\mathbf{G}'[\mathbb{E}\mathbf{m}_i\mathbf{m}_i'], \\
\mathbf{R}^{\text{EL}} &= (\mathbf{G}'[\mathbb{E}\mathbf{m}_i\mathbf{m}_i']^{-1}\mathbf{G})^{-1}, \\
\mathbf{U} &= \mathbb{E}[\mathbf{m}_i\mathbf{m}_i'\mathbf{P}\mathbf{m}_i].
\end{aligned}$$

It is not clear how these compare to $\mathbb{B}_{\text{QMLE}}(\boldsymbol{\theta})$ in general. However, when $\mathbf{Z}$ is

multivariate normal, it is easy to show that the upper block of $\mathbb{B}_{\mathrm{QMLE}}$ is equal to (3.27) since, under normality,

$$\begin{aligned}
\mathbf{R}^{\mathrm{EL}} &= \{\mathbf{G}'[2\bar{\mathbf{H}}(\boldsymbol{\Sigma}\otimes\boldsymbol{\Sigma})\bar{\mathbf{H}}']^{-1}\mathbf{G}\}^{-1} \\
&= 2[\mathbf{G}'\mathbf{H}'(\boldsymbol{\Sigma}\otimes\boldsymbol{\Sigma})^{-1}\mathbf{H}\mathbf{G}]^{-1} \\
&= 2\mathbf{R}, \\
\mathbf{Q}^{\mathrm{EL}} &= \mathbf{R}^{\mathrm{EL}}\mathbf{G}'[2\bar{\mathbf{H}}(\boldsymbol{\Sigma}\otimes\boldsymbol{\Sigma})\bar{\mathbf{H}}']^{-1} \\
&= \mathbf{R}\mathbf{G}'\mathbf{H}'(\boldsymbol{\Sigma}\otimes\boldsymbol{\Sigma})^{-1}\mathbf{H} \\
&= \mathbf{Q}\mathbf{H}.
\end{aligned}$$

3.5 Concluding remarks

The paper examined estimation methods available for covariance structure models in terms of their first and second order asymptotic properties. The results suggest the following strategy in estimating models of covariance structure.

First, if we have large samples so that the first-order asymptotic results can be applied we should prefer GMM or EL to quasi-MLE. Due to increased computational difficulty of EL, the GMM estimator would be preferable. If efficiency is not an issue and we are ready to sacrifice efficiency for a simpler and yet consistent estimation technique we may prefer the traditional normal QMLE approach.

Second, if we have small samples, EL would be the preferred method of estimation. If the data is normal, normal QMLE will have the same second-order bias as EL. The bias can be estimated using (3.27) and the bias-adjusted estimator can be

constructed. If the data are not normal and we still use the QMLE, construction of the bias-adjusted estimator may be more complicated but is still possible using (3.25).

Interesting related questions are how different are the alternative estimates in applications and whether the equal efficiency and equal bias results can be shown for other distributions.

Bibliography

AHN, S. C. AND P. SCHMIDT (1995): "Efficient estimation of models for dynamic panel data," *Journal of Econometrics*, 68, 5–27.

AIGNER, D. J., C. HSIAO, A. KAPTEYN, AND T. WANSBEEK (1984): "Latent variable models in econometrics," in *Handbook of Econometrics*, ed. by Z. Griliches and M. D. Intriligator, vol. II, 1323–1393.

BAO, Y. AND A. ULLAH (2003): "The Second-Order Bias and Mean Squared Error of Estimators in Time Series Models," *Working Paper, University of California - Riverside*, http://www.economics.ucr.edu/papers/papers03/03-08.pdf.

CHAMBERLAIN, G. (1982): "Multivariate regression models for panel data," *Journal of Econometrics*, 18, 5–46.

——— (1984): "Panel data," in *Handbook of Econometrics*, ed. by Z. Griliches and M. D. Intriligator, vol. II, 1248–1313.

HANSEN, L. (1982): "Large sample properties of generalized method of moments estimators," *Econometrica*, 50, 1029–1054.

JÖRESKOG, K. G. (1970): "A general method for analysis of covariance structures," *Biometrika*, 57, 239–251.

JÖRESKOG, K. G. AND A. S. GOLDBERGER (1975): "Estimation of a model with multiple indicators and multiple causes of a single latent variable," *Journal of American Statistical Association*, 70, 631–939.

JÖRESKOG, K. G. AND D. SÖRBOM (1977): "Statistical models and methods for analysis of longitudinal data," in *Latent Variables in Socio-Economic Models*, ed. by D. a. A.Goldberger, Amsterdam: North-Holland Publishing Company, Contributions to Economic Analysis, 285–325.

——— (1996): *LISREL 8 User's Reference Guide.*, SSI Scientific Software.

KIM, K.-I. (2005): "Higher order bias correcting moment equation for M-estimation," *UCLA Papers.*

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

LINTON, O. (1997): "An asymptotic expansion in the GARCH(1,1) model," *Econometric Theory*, 13, 558.

MAGNUS, J. R. AND H. NEUDECKER (1988): *Matrix differential calculus with applications in statistics and econometrics*, Wiley Series in Probability and Statistics, Chichester: John Wiley and Sons Ltd.

MCCULLAGH, P. (1987): *Tensor Methods in Statistics*, Monographs on Statistics and Applied Probability, London: Chapman and Hall.

NEWEY, W. (1990): "Semiparametric efficiency bounds," *Journal of Applied Econometrics*, 5, 99–135.

NEWEY, W. AND D. MCFADDEN (1994): "Large sample estimation and hypothesis testing," in *Handbook of Econometrics*, ed. by R. Engle and D. McFadden, vol. IV, 2113–2241.

NEWEY, W. K., J. S. RAMALHO, AND R. J. SMITH (2003): "Asymptotic bias for GMM and GEL estimators with estimated nuisance parameters," *CEMMAP working paper CWP05/03.*

NEWEY, W. K. AND R. J. SMITH (2004): "Higher order properties of GMM and Generalized Empirical Likelihood estimators," *Econometrica*, 72, 219–255.

OWEN, A. B. (2001): *Empirical likelihood*, Monographs on statistics and applied probability; 92, Boca Raton, Fla. : Chapman and Hall.

QIN, J. AND J. LAWLESS (1994): "Empirical likelihood and general estimating equations," *The Annals of Statistics*, 22, 300–325.

RILSTONE, P., V. SRIVASTAVA, AND A. ULLAH (1996): "The second-order bias and mean squared error of nonlinear estimators," *Journal of Econometrics*, 75, 369–395.

ROTHENBERG, T. (1984): "Approximating the distributions of econometric estimators and test statictics," in *Handbook of Econometrics*, ed. by Z. Griliches and M. D. Intriligator, vol. II, 881–935.

ULLAH, A. (2004): *Finite Sample Econometrics*, Advanced Texts in Econometrics, Oxford University Press.

Appendix: Proofs

PROOF OF THEOREM 3.4.1: Let $\bar{\mathbf{M}}(\boldsymbol{\beta}) = \frac{1}{N}\sum_{i=1}^{N}\frac{\partial \mathbf{s}_i(\boldsymbol{\beta})}{\partial \boldsymbol{\beta}'}$, $\mathbf{M}(\boldsymbol{\beta}) = \mathbb{E}\frac{\partial \mathbf{s}_i(\boldsymbol{\beta})}{\partial \boldsymbol{\beta}'}$, $\bar{\mathbf{M}}_j(\boldsymbol{\beta}) = \frac{1}{N}\sum_{i=1}^{N}\frac{\partial^2 \mathbf{s}_i(\boldsymbol{\beta})}{\partial \boldsymbol{\beta}'\partial \boldsymbol{\beta}_j}$ and $\bar{\boldsymbol{\beta}}$ be between $\hat{\boldsymbol{\beta}}$ and $\boldsymbol{\beta}_o$. By the second-order Taylor expansion of (3.15) around $\boldsymbol{\beta}_o$, we have

$$\begin{aligned}
\mathbf{s}_N(\hat{\boldsymbol{\beta}}) &= \mathbf{0} \\
&= \mathbf{s}_N(\boldsymbol{\beta}_o) + \bar{\mathbf{M}}(\boldsymbol{\beta}_o)(\hat{\boldsymbol{\beta}} - \boldsymbol{\beta}_o) + \frac{1}{2}\sum_{j=1}^{p+q^2}(\hat{\boldsymbol{\beta}}_j - \boldsymbol{\beta}_{oj})\bar{\mathbf{M}}_j(\bar{\boldsymbol{\beta}})(\hat{\boldsymbol{\beta}} - \boldsymbol{\beta}_o) \\
&= \mathbf{s}_N(\boldsymbol{\beta}_o) + \mathbf{M}(\boldsymbol{\beta}_o)(\hat{\boldsymbol{\beta}} - \boldsymbol{\beta}_o) + [\bar{\mathbf{M}}(\boldsymbol{\beta}_o) - \mathbf{M}(\boldsymbol{\beta}_o)](\hat{\boldsymbol{\beta}} - \boldsymbol{\beta}_o) \\
&\quad + \frac{1}{2}\sum_{j=1}^{p+q^2}(\hat{\boldsymbol{\beta}}_j - \boldsymbol{\beta}_{oj})\mathbf{M}_j(\boldsymbol{\beta}_o)(\hat{\boldsymbol{\beta}} - \boldsymbol{\beta}_o) + \\
&\quad + \frac{1}{2}\sum_{j=1}^{p+q^2}(\hat{\boldsymbol{\beta}}_j - \boldsymbol{\beta}_{oj})[\bar{\mathbf{M}}_j(\bar{\boldsymbol{\beta}}) - \mathbf{M}_j(\boldsymbol{\beta}_o)](\hat{\boldsymbol{\beta}} - \boldsymbol{\beta}_o).
\end{aligned}$$

Note that $\bar{\mathbf{M}}(\boldsymbol{\beta}_o) = \mathbf{M}(\boldsymbol{\beta}_o)$ so that the third term in the last equation is zero. Also note that the last term is $O_p(N^{-3/2})$.

Assume that $\bar{\mathbf{M}}(\boldsymbol{\beta}_o)$ is not singular. Then,

$$\begin{aligned}
\hat{\boldsymbol{\beta}} - \boldsymbol{\beta}_o &= -[\mathbf{M}(\boldsymbol{\beta}_o)]^{-1}\mathbf{s}_N(\boldsymbol{\beta}_o) \\
&\quad -\frac{1}{2}[\mathbf{M}(\boldsymbol{\beta}_o)]^{-1}\sum_{j=1}^{p+q^2}(\hat{\boldsymbol{\beta}}_j - \boldsymbol{\beta}_{oj})\mathbf{M}_j(\boldsymbol{\beta}_o)(\hat{\boldsymbol{\beta}} - \boldsymbol{\beta}_o) \\
&\quad +O_p(N^{-3/2}). \qquad (3.28)
\end{aligned}$$

But $\mathbf{M}(\boldsymbol{\beta}_o) = -\begin{bmatrix} \mathbf{0} & \mathbf{G}_o'\mathbf{H}' \\ \mathbf{H}\mathbf{G}_o & \boldsymbol{\Sigma}_o \otimes \boldsymbol{\Sigma}_o \end{bmatrix}$, $\mathbf{s}_N(\boldsymbol{\beta}_o) = -\begin{bmatrix} \mathbf{0} \\ \mathbf{H}\,\mathbf{m}_N(\boldsymbol{\theta}_o) \end{bmatrix}$ and the second term is $O_p(N^{-1})$. We thus have

$$\begin{aligned} \hat{\boldsymbol{\beta}} - \boldsymbol{\beta}_o &= \frac{1}{\sqrt{N}} \begin{bmatrix} \mathbf{Q}_o \\ \mathbf{P}_o \end{bmatrix} \mathbf{H} \frac{1}{\sqrt{N}} \sum_{i=1}^{N} [vech(\mathbf{S}_i) - vech(\boldsymbol{\Sigma}_o)] + O_p(N^{-1}) \\ &= \frac{1}{\sqrt{N}} \boldsymbol{\mu} + O_p(N^{-1}). \end{aligned} \tag{3.29}$$

Substituting (3.29) into (3.28), multiplying by $\sqrt{N}$ and collecting terms of the same order yields the result.

Department of Economics

Michigan State University

East Lansing, MI 48824

Email address: prohorov@msu.edu