LMU LUDWIG-MAXIMILIANS-UNIVERSITÄT MÜNCHEN

INSTITUT FÜR STATISTIK

Hans Schneeweiss

The linear GMM model with singular covariance matrix due to the elimination of a nuisance parameter

Technical Report Number 165, 2014
Department of Statistics
University of Munich

http://www.stat.uni-muenchen.de

The linear GMM model with singular covariance matrix due to the elimination of a nuisance parameter

Hans Schneeweiss

Abstract

When in a linear GMM model nuisance parameters are eliminated by multiplying the moment conditions by a projection matrix, the covariance matrix of the model, the inverse of which is typically used to construct an efficient GMM estimator, turns out to be singular and thus cannot be inverted. However, one can show that the generalized inverse can be used instead to produce an efficient estimator. Various other matrices in place of the projection matrix do the same job, i.e., they eliminate the nuisance parameters. The relations between those matrices with respect to the efficiency of the resulting estimators are investigated.

Key Words: generalized method of moments, orthogonal projection, nuisance parameter, singular covariance matrix, weighting matrix, generalized inverse, panel data model.

1 Introduction

A Generalized Method of Moments (GMM) model is essentially a vector valued function $\psi(v,\gamma)$ of an observable random data vector v and an unknown parameter vector γ such that the so-called moment conditions $\mathbb{E}\psi(v,\gamma_0)=0$ hold for a unique value γ_0 of γ, the "true" parameter value, Hansen (1982), Hall (2005). The moment conditions can be used to construct an estimator of γ on the basis of an i.i.d. sample v_n, $n=1,\ldots,N$, of the data vector v by minimizing the quadratic form $Q:=\sum_1^N \psi(v_n,\gamma)V\psi(v_n,\gamma)^\top$ with respect to γ, where V is some weighting matrix. We obtain an efficient estimator if V is chosen to be the inverse of an

estimate of the covariance matrix of ψ, i.e., of $\Omega := \mathbb{E}\psi(v,\gamma_0)\psi(v,\gamma_0)^\top$, which is assumed to be nonsingular.

Here we focus on a linear GMM model, where ψ is a linear function in γ. In particular we suppose that the statistician is interested in estimating only a sub-vector β of $\gamma = (\alpha^\top, \beta^\top)^\top$, and the complementary sub-vector α is a nuisance parameter. The function ψ then takes the form $\psi(a,A,B,\alpha,\beta) = a - A\alpha - B\beta$, where a, A, and B are observable data matrices. This scenario is common in the context of a linear model, where the intercept term is often of minor interest. It also turns up in the GMM case. A typical example is a panel data model, where the individual effects are nuisance parameters. In addition one may have measurement errors in a panel data model, where the measurement error variance is treated as a nuisance parameter, Wansbeek (2001), Xiao *et al.* (2007), Schneeweiss *et al.*(2014).

One way to deal with this problem is to eliminate the nuisance parameter by multiplying the function ψ from the left by some matrix $K^\top$, often taken to be a projection matrix, such that $K^\top \mathbb{E}A = 0$. We then work with the new moment conditions $\mathbb{E}K^\top\psi(a,B,\beta) = 0$. The corresponding covariance matrix of $K^\top\psi$ is $K^\top\Omega K$. Quite often, in particular if K is a projection matrix, this matrix is singular and so its inverse cannot be used as an optimal weighting matrix. It turns out, however, that the generalized inverse can be used as a weighting matrix and this will lead to an efficient estimator of β; in the context of panel data models and instrumental variables see Biørn and Klette (1998) and White (1986). When the covariance matrix of ψ is singular for some other reason we can still use the generalized inverse, but this will not necessarily result in an efficient estimator. The paper investigates these propositions and clarifies some of the relations between various choices of the matrix K.

Similar investigations can be found in Xiao *et al.* (2007) and Xiao *et al.* (2010), although these papers are mainly interested in panel data models with measurement errors and do not discuss the singularity problem in its general setting, as is done here. Dorana and Schmid (2006) also use the generalized inverse of the covariance matrix as a weighting matrix, however after having modified the covariance matrix by reducing it to some of its principal components. They do this to improve the small sample properties of the GMM estimator, which is not the focus of the present paper.

It may be noted that the results presented in this paper have their analogues in linear regression models, e.g., Rao *et al.* (2008), Seber and Lee (2006), Puntanen *et al.* (2013), see also White (1986). But their are two main differences. Fist, in a linear model, $y = X\beta + u$, say, the residual vector u is independent of or at least uncorrelated with the data matrix X, whereas $a - A\alpha - B\beta$ need not be

independent of the data (A,B). Secondly, the number of rows N, say, of (y,X) increases with the sample size N and indeed (y,X) *is* the sample, whereas (a,A,B) has a fixed number of rows and there is a sample (a_n,A_n,B_n), $n=1\ldots,N$, of data to be used in the estimation procedure. Consequently, the results on the efficiency of estimators in the linear model are concerned with finite sample properties while in the linear GMM model they are of an asymptotic nature.

In the next section the linear GMM model and its corresponding GMM estimators are recapitulated. Section 3 shows how to estimate the parameter of interest in the presence of a nuisance parameter by estimating the complete parameter vector. In Section 4 the same is done by first eliminating the nuisance parameter, and various possibilities of doing so are discussed and compared to each other. Section 5 deals with the important special case of a fixed matrix A. The more general case of a singular covariance matrix at the outset is studied in Section 6. Some concluding remarks are found in Section 7.

2 The linear GMM model

The linear GMM model consists of an i.i.d. sample (a_n,C_n) of observable $(q\times 1)$ random vectors a_n and $(q\times c)$ random matrices C_n, $n=1,\cdots,N$, and an unknown $(c\times 1)$ parameter vector γ such that for some $\gamma=\gamma_0$ the following q *moment conditions* hold

$$\mathbb{E}(a_n - C_n\gamma_0) = 0. \tag{1}$$

We use a bar to denote averages over n, e.g., $\bar{a} = \frac{1}{N}\sum_{n=1}^N a_n$ and $\bar{C} = \frac{1}{N}\sum_{n=1}^N C_n$, and we use a tilde to denote expectations, e.g., $\tilde{a} = \mathbb{E}a_n$ and $\tilde{C} = \mathbb{E}C_n$.

The matrix $\tilde{C}$ is assumed to have full column rank thereby guaranteeing the uniqueness of γ_0 (*identifiability condition* for γ). In particular this implies that $q \geq c$. There is a sample analogue to the identifiability condition, viz., that $\bar{C}$ has full column rank. When the population identifiability condition is satisfied, its finite sample analogue will also be satisfied, at least with high probability and for sufficiently large N. We will therefore always tacitly assume the finite sample analogue to be valid, too, whenever the population condition is assumed to be valid. This remark also applies to other sample analogues in the sequel.

In addition to the moment conditions it is assumed that $a_n - C_n\gamma_0$ has a nonsingular covariance matrix

$$\Omega := \mathbb{E}(a_n - C_n\gamma_0)(a_n - C_n\gamma_0)^\top.$$

The objective is to estimate γ_0 with the help of the sample (a_n,C_n), $n=1,\cdots,N$. Let me briefly state the well-known basic facts of this estimation problem.

A General Method of Moments (GMM) estimator of γ_0 is found by applying a weighted least squares approach to (1) and is given by

$$\hat{\gamma}_V = (\bar{C}^\top \hat{V} \bar{C})^{-1} \bar{C}^\top \hat{V} \bar{a}, \tag{2}$$

where $\hat{V}$ is a $c \times c$ positive semi-definite weighting matrix to be chosen such that $\text{plim}(\hat{V}) = V$ exists and $V\tilde{C}$ has full column rank (*admissibility condition* for the weighting matrix). $\hat{\gamma}_V$ is consistent and asymptotically normal. The simplest choice of $\hat{V}$ is $\hat{V} = I$, but an optimal choice is $\hat{V} = \hat{\Omega}^{-1}$, where

$$\hat{\Omega} = \overline{(a - C\hat{\gamma}_1)(a - C\hat{\gamma}_1)^\top} \tag{3}$$

and $\hat{\gamma}_1$ is a provisional (first step) estimate of γ found from (2) with, e.g., $\hat{V} = I$. The estimator with weighting matrix $\hat{\Omega}^{-1}$,

$$\hat{\gamma}_{opt} = (\bar{C}^\top \hat{\Omega}^{-1} \bar{C})^{-1} \bar{C}^\top \hat{\Omega}^{-1} \bar{a}, \tag{4}$$

is optimal in the sense that it is (asymptotically) efficient in the class of estimators $\hat{\gamma}_V$. For an estimator of this class the asymptotic variance is given by

$$\text{Avar}(\hat{\gamma}_V) = \frac{1}{N} (\tilde{C}^\top V \tilde{C})^{-1} \tilde{C}^\top V \Omega V \tilde{C} (\tilde{C}^\top V \tilde{C})^{-1},$$

whereas for the efficient GMM estimator it is given by

$$\text{Avar}(\hat{\gamma}_{opt}) = \frac{1}{N} (\tilde{C}^\top \Omega^{-1} \tilde{C})^{-1}. \tag{5}$$

That $\hat{\gamma}_{opt}$ is at least as efficient as $\hat{\gamma}_V$ means that $\text{Avar}(\hat{\gamma}_V) \geq \text{Avar}(\hat{\gamma}_{opt})$ in the sense that $\text{Avar}(\hat{\gamma}_V) - \text{Avar}(\hat{\gamma}_{opt})$ is positive semi-definite, and this can be shown in the present case, e.g., Hall (2005). $\hat{\gamma}_{opt}$ corresponds to the GLS estimator in linear regression analysis.

So much for the basics of traditional GMM theory as far as it is restricted to linear models. It may be noted that the i.i.d. assumption can be generalized by assuming stationarity instead, but we stick to the simpler i.i.d. case.

3 Estimating a sub-vector

Now suppose we are only interested in estimating a $p \times 1$ sub-vector β of γ. Let $\gamma = (\alpha^\top, \beta^\top)^\top$, and partition C_n correspondingly as $C_n = (A_n, B_n)$, then the moment conditions (1) become

$$\mathbb{E}(a_n - A_n \alpha_0 - B_n \beta_0) = 0. \tag{6}$$

β is the parameter of interest and α is regarded as a nuisance parameter.

The following very simple example may serve to illustrate this kind of model.

Example 1: Let y_n and x_n be observable i.i.d. random variables, $n = 1,\ldots,N$, and α and β unknown parameters. Assume the following linear relation to hold:

$$y_n = \alpha + x_n\beta + u_n, \quad \mathbb{E}u_n = 0. \tag{7}$$

(For simplicity we refrain from denoting the true parameter value by the subscript 0). For some reason (e.g., missing variables that are correlated with x_n, the equation is part of a multi-equation system, errors in the variables) the unobservable random variable u_n is not independent of x_n. But some q instrumental variables z_{in}, $i = 1,\ldots,q$, are available, which by definition are independent of all the u_n but correlated with x_n. Let $z_n = (z_{1n},\ldots,z_{qn})^\top$ and multiply (7) by z_n. We get

$$z_n y_n = z_n\alpha + z_n x_n\beta + z_n u_n,$$

which corresponds to (6) with $a_n = z_n y_n$, $A_n = z_n$, and $B_n = z_n x_n$. The covariance matrix is

$$\Omega = \sigma_u^2 \mathbb{E} z_n z_n^\top.$$

Of course, we can still estimate the complete parameter vector γ and then select the sub-vector $\hat{\beta}_V$ from the estimate $\hat{\gamma}_V$. In doing so we find the efficient GMM estimator $\hat{\beta}_{opt}$, using $\hat{\Omega}^{-1}$ as optimal weighting matrix, by solving the following equations system

$$\begin{pmatrix} \bar{A}^\top\hat{\Omega}^{-1}\bar{A} & \bar{A}^\top\hat{\Omega}^{-1}\bar{B} \\ \bar{B}^\top\hat{\Omega}^{-1}\bar{A} & \bar{B}^\top\hat{\Omega}^{-1}\bar{B} \end{pmatrix} \begin{pmatrix} \hat{\alpha}_{opt} \\ \hat{\beta}_{opt} \end{pmatrix} = \begin{pmatrix} \bar{A}^\top\hat{\Omega}^{-1}\bar{a} \\ \bar{B}^\top\hat{\Omega}^{-1}\bar{a} \end{pmatrix},$$

which results in

$$\hat{\beta}_{opt} = (\bar{B}^\top\hat{\Omega}^{-\frac{1}{2}}\hat{P}_\Omega\hat{\Omega}^{-\frac{1}{2}}\bar{B})^{-1}\bar{B}^\top\hat{\Omega}^{-\frac{1}{2}}\hat{P}_\Omega\hat{\Omega}^{-\frac{1}{2}}\bar{a} \tag{8}$$

with the orthogonal projection matrix

$$\hat{P}_\Omega = I - \hat{\Omega}^{-\frac{1}{2}}\bar{A}(\bar{A}^\top\hat{\Omega}^{-1}\bar{A})^{-1}\bar{A}^\top\hat{\Omega}^{-\frac{1}{2}}. \tag{9}$$

(The same result exists in asymptotic regression analysis, Putanen *et al.* (2013), 10.34 (c)). The asymptotic variance of $\hat{\beta}_{opt}$ is given by

$$\mathrm{Avar}(\hat{\beta}_{opt}) = \frac{1}{N}(\tilde{B}^\top\Omega^{-\frac{1}{2}}P_\Omega\Omega^{-\frac{1}{2}}\tilde{B})^{-1}, \tag{10}$$

where

$$P_\Omega = I - \Omega^{-\frac{1}{2}}\tilde{A}(\tilde{A}^\top\Omega^{-1}\tilde{A})^{-1}\tilde{A}^\top\Omega^{-\frac{1}{2}}.$$

(To verify (10) note that

$$\hat{\beta}_{opt} - \beta_0 = (\bar{B}^\top\hat{\Omega}^{-\frac{1}{2}}\hat{P}_\Omega\hat{\Omega}^{-\frac{1}{2}}\bar{B})^{-1}\bar{B}^\top\hat{\Omega}^{-\frac{1}{2}}\hat{P}_\Omega\hat{\Omega}^{-\frac{1}{2}}(\bar{a} - \bar{B}\beta_0).$$

Because $\hat{P}_\Omega\hat{\Omega}^{-\frac{1}{2}}\bar{A} = 0$, we can replace the term $\bar{a} - \bar{B}\beta_0$ with $\bar{a} - \bar{A}\alpha_0 - \bar{B}\beta_0 = \bar{a} - \bar{C}\gamma_0$. Equation (10) then easily follows.)

4 Correcting for nuisance parameter

Another way of estimating the parameter of interest β is to first eliminate the nuisance parameter α together with its data matrix A and then to estimate β from the remaining moment conditions. In the linear model this is usually done by applying an orthogonal projection matrix. However, one can also use any matrix that nullifies A. We will construct an efficient estimator of β using such a matrix and will then study its relation to orthogonal projection matrices as well as to the estimator introduced in Section 3.

Let $\hat{K}$ be an observable $q \times k$ matrix (depending on N) such that $\hat{K}^\top\bar{A} = 0$. We say that $\hat{K}$ eliminates $\bar{A}$. Assume $\hat{K}$ to converge in probability: $\text{plim}(\hat{K}) = K$. Then $K^\top\tilde{A} = 0$ (i.e., K eliminates $\tilde{A}$). Let us assume that, at least for large N, $\hat{K}$ and K have the same rank. A case in point might be that $\hat{K}$ is a constant matrix and $\hat{K} = K$, see Section 5 below. If we multiply the moment conditions (6) by $K^\top$ from the left, we get rid of the nuisance parameter α and obtain new moment conditions as follows:

$$\mathbb{E}[K^\top(a_n - B_n\beta_0)] = 0, \tag{11}$$

Even if $\tilde{B}$ (together with $\tilde{A}$) identifies β, this need not be so for $K^\top\tilde{B}$. In order to be able to identify β from the new moment conditions (11) we must adopt the further condition that $K^\top\tilde{B}$ has full column rank (*identifiability condition* for β *given* K). (Note that $K^\top\tilde{B}$ has k rows instead of the original q rows of $\tilde{B}$).

We can use (11) to construct GMM estimators of β. However, as K is generally unknown, we have to replace it by its estimate $\hat{K}$. Let $\hat{V}$ be a $k \times k$ weighting matrix , which converges in probability to some positive semi-definite matrix V (not necessarily the same V as in Section 2), for which we assume that $VK^\top\tilde{B}$ has full column rank (*admissibility condition* for the weighting matrix). The corresponding GMM estimator is then given by

$$\hat{\beta}_{KV} = (\bar{B}^\top\hat{K}\hat{V}\hat{K}^\top\bar{B})^{-1}\bar{B}^\top\hat{K}\hat{V}\hat{K}^\top\bar{a}, \tag{12}$$

As, due to $\hat{K}^\top \bar{A} = 0$,

$$\hat{\beta}_{KV} - \beta_0 = (\bar{B}^\top \hat{K}\hat{V}\hat{K}^\top \bar{B})^{-1}\bar{B}^\top \hat{K}\hat{V}\hat{K}^\top(\bar{a} - \bar{A}\alpha_0 - \bar{B}\beta_0)$$

and $\sqrt{N}(\bar{a} - \bar{A}\alpha_0 - \bar{B}\beta_0) \xrightarrow{d} \mathrm{N}(0, \Omega)$, it is clear that $\hat{\beta}_{KV}$ is consistent and asymptotically normal with asymptotic variance

$$\mathrm{Avar}(\hat{\beta}_{KV}) = \frac{1}{N}(\tilde{B}^\top KVK^\top \tilde{B})^{-1}\tilde{B}^\top KVK^\top \Omega KVK^\top \tilde{B}(\tilde{B}^\top KVK^\top \tilde{B})^{-1}.$$

In looking for an optimal GMM estimator within the class of estimators $\hat{\beta}_{KV}$ with an admissible weighting matrix $\hat{V}$, we might think of choosing for the weighting matrix the inverse of Ω (or rather of its estimate $\hat{\Omega}$) as in Section 2. But this turns out not to be optimal. Instead one should try to use the inverse of $K^\top \Omega K$ (or of its estimate). But if $\mathrm{rank}(K) < k$, which may well be possible, in particular if K is a projection matrix, $K^\top \Omega K$ (just as its estimate) will be singular and cannot be inverted. However, we can always use the (Moore-Penrose) generalized inverse $(\hat{K}^\top \hat{\Omega}\hat{K})^+$ instead. We are thus led to the optimal estimator

$$\hat{\beta}_K = [\bar{B}^\top \hat{K}(\hat{K}^\top \hat{\Omega}\hat{K})^+ \hat{K}^\top \bar{B}]^{-1}\bar{B}^\top \hat{K}(\hat{K}^\top \hat{\Omega}\hat{K})^+ \hat{K}^\top \bar{a}, \tag{13}$$

which has asymptotic variance

$$\mathrm{Avar}(\hat{\beta}_K) = \frac{1}{N}[\tilde{B}^\top K(K^\top \Omega K)^+ K^\top \tilde{B}]^{-1}. \tag{14}$$

Note that, due to the identifiability assumption given K (i.e., $K^\top \tilde{B}$ has full column rank), the matrix in brackets $\tilde{B}^\top K(K^\top \Omega K)^+ K^\top \tilde{B}$ is nonsingular (i.e., $(K^\top \Omega K)^+$ is an admissible weighting matrix) and so its inverse exists. This follows from the subsequent lemma:

Lemma 1 *For any $(q \times p)$ matrix B, $(q \times k)$ matrix K, and positive definite $(q \times q)$ matrix Ω*

$$\mathrm{rank}[B^\top K(K^\top \Omega K)^+ K^\top B] = \mathrm{rank}(K^\top B).$$

Proof: First note that $B^\top K(K^\top \Omega K)^+ K^\top B = B_0^\top P_0 B_0$, where $B_0 := \Omega^{-\frac{1}{2}}B$, $P_0 := K_0(K_0^\top K_0)^+ K_0^\top$, and $K_0 := \Omega^{\frac{1}{2}}K$. P_0 is an orthogonal projection matrix. Therefore the rank of the matrix $B^\top K(K^\top \Omega K)^+ K^\top B$ equals

$$\mathrm{rank}(B_0^\top P_0 B_0) = \mathrm{rank}(P_0 B_0) = \mathrm{rank}(K_0^\top B_0) = \mathrm{rank}(K^\top B),$$

where the middle equality follows from

$$\mathrm{rank}(K_0^\top B_0) \geq \mathrm{rank}(P_0 B_0) \geq \mathrm{rank}(K_0^\top P_0 B_0) = \mathrm{rank}(K_0^\top B_0). \quad \diamond$$

The optimality of $\hat{\beta}_K$ is asserted in the following theorem:

Theorem 1 *For a given elimination matrix $\hat{K}$ such that $K^\top \tilde{B}$ has full column rank, p, the estimator $\hat{\beta}_K$ given by* (13) *is efficient within the class of estimators $\hat{\beta}_{KV}$ given by* (12)*.*

Proof: The proof is a slight modification of the corresponding efficiency proof for $\hat{\gamma}_{opt}$, e.g., Harris and Mátyás (1999), Arellano (2003). Let

$$D^\top = (\tilde{B}^\top K V K^\top \tilde{B})^{-1} \tilde{B}^\top K V K^\top \Omega^{\frac{1}{2}}$$
$$F^\top = \tilde{B}^\top K (K^\top \Omega K)^+ K^\top \Omega^{\frac{1}{2}}.$$

Then $D^\top D = N \operatorname{Avar}(\hat{\beta}_{KV})$ and $(F^\top F)^{-1} = N \operatorname{Avar}(\hat{\beta}_K)$. Because of the identity

$$K^\top \Omega K (K^\top \Omega K)^+ K^\top = K^\top$$

it follows that $D^\top F = I$, and so, with $G := F(F^\top F)^{-1} F^\top$,

$$N\left(\operatorname{Avar}(\hat{\beta}_{KV}) - \operatorname{Avar}(\hat{\beta}_K)\right) = D^\top D - (F^\top F)^{-1} = D^\top (I - G) D \geq 0. \quad \diamond$$

In Example 1, the efficient estimator (13) is

$$\hat{\beta}_K = [\overline{zx}^\top \hat{K} (\hat{K}^\top \hat{\Omega} \hat{K})^+ \hat{K}^\top \overline{zx}]^{-1} \overline{zx}^\top \hat{K} (\hat{K}^\top \hat{\Omega} \hat{K})^+ \hat{K}^\top \overline{zy}.$$

The optimal estimator $\hat{\beta}_K$ depends, of course, on the choice of the matrix $\hat{K}$. Apart from the requirement that $\hat{K}^\top \bar{A} = 0$ we are free to choose $\hat{K}$. It turns out, however, that $\hat{\beta}_K$ depends only on the column space of $\hat{K}$. (Note that the property of eliminating $\bar{A}$ also depends only on the column space of $\hat{K}$). If we denote the column space of a matrix by $\{\cdot\}$, we can state the following theorem.

Theorem 2 *Let $\hat{K}_i$, $i = 1,2$, be two elimination matrices of dimensions $q \times k_i$ with* $\operatorname{rank}(\hat{K}_i) \geq p$*, then the optimal estimators $\hat{\beta}_{K_1}$ and $\hat{\beta}_{K_2}$ given by* (13) *with K_i in place of K are equal for all $\bar{a}$ and all $\bar{B}$ with* $\operatorname{rank}(\hat{K}_i^\top \bar{B}) = p$ *if, and only if, $\{\hat{K}_1\} = \{\hat{K}_2\}$.*

Note that the condition $\operatorname{rank}(\hat{K}_i^\top \bar{B}) = p$ is necessary and sufficient for the existence of $\hat{\beta}_{K_i}$, see Lemma 1.

The proof of the theorem is based on the following lemma:

Lemma 2 *Let K_i, $i = 1,2$, be two $(q \times k_i)$ matrices with* $\operatorname{rank}(K_i) \geq p$*. Let $\mathfrak{B} = \{b \in \mathbb{R}^q \,|\, K_i^\top b \neq 0,\ i = 1,2\}$. Then*

1. *For any $b \in \mathfrak{B}$ there are $p-1$ further $b_j \in \mathfrak{B}$, $j = 2,\ldots,p$ such that the matrix $B := (b, b_2, \ldots, b_p)$ satisfies* $\operatorname{rank}(K_i^\top B) = p$, $i = 1,2$.

2. *The set $\mathfrak{B}$ contains q vectors $b_1, \ldots, b_q$ which form a basis for $\mathbb{R}^q$.*

Proof:

1. Define the projection matrices $P_i := K_i(K_i^\top K_i)^+ K_i^\top$, $i = 1,2$. Then $\mathfrak{B} = \{b \,|\, P_i b \neq 0,\ i = 1,2\}$. For any given $b \in \mathfrak{B}$ select $p-1$ vectors $b_2, \ldots, b_p \in \{P_1\}$ such that the tupel $(P_1 b, b_2, \ldots, b_p)$ is linearly independent. This can be done because $\{P_1\}$ has dimension $\geq p$. For each b_j choose a sufficiently small neighborhood $\mathfrak{U}_j$ so that whatever $b'_j \in \mathfrak{U}_j$ are chosen the tupel $(P_1 b, P_1 b'_2, \ldots, P_1 b'_p)$ remains linearly independent and thus the matrices $B' := (b, b'_2, \ldots, b'_p)$ satisfy $\operatorname{rank}(P_1 B') = p$. Finally choose $b'_2, \ldots, b'_p$ in such a way that also $\operatorname{rank}(P_2 B') = p$. As by Lemma 1 $\operatorname{rank}(P_i B) = \operatorname{rank}(K_i^\top B)$ for any $(q \times p)$ matrix B, this proves part 1 of the lemma.

2. Part 2 is a consequence of the fact that $\mathfrak{B}$ is the set-theoretic complement in $\mathbb{R}^q$ of the union $\mathfrak{N}(P_1) \cup \mathfrak{N}(P_2)$ of the two null-spaces $\mathfrak{N}(P_i)$ of the projections P_i and both $\mathfrak{N}(P_i)$ are linear subspaces of $\mathbb{R}^q$ of dimension less than q, see also Xiao *et al.* (2010), Lemma 1 for a similar proposition. $\diamond$

Proof of Theorem 2: The equality $\hat{\beta}_{K_1} = \hat{\beta}_{K_2}$ is equivalent to

$$(\hat{B}^\top \hat{P}_1 \hat{B})^{-1} \hat{B}^\top \hat{P}_1 \hat{a} = (\hat{B}^\top \hat{P}_2 \hat{B})^{-1} \hat{B}^\top \hat{P}_2 \hat{a} \tag{15}$$

with $\hat{P}_i := \hat{\Omega}^{\frac{1}{2}} \hat{K}_i (\hat{K}_i^\top \hat{\Omega} \hat{K}_i)^+ \hat{K}_i^\top \hat{\Omega}^{\frac{1}{2}}$, $\hat{B} := \hat{\Omega}^{-\frac{1}{2}} \bar{B}$, and $\hat{a} := \hat{\Omega}^{-\frac{1}{2}} \bar{a}$. The two orthogonal projection matrices $\hat{P}_1$ and $\hat{P}_2$ are equal iff they have identical image spaces, i.e., iff $\{\hat{\Omega}^{\frac{1}{2}} \hat{K}_1\} = \{\hat{\Omega}^{\frac{1}{2}} \hat{K}_2\}$, which is equivalent to $\{\hat{K}_1\} = \{\hat{K}_2\}$. Therefore $\{\hat{K}_1\} = \{\hat{K}_2\}$ implies (15) and thus $\hat{\beta}_{K_1} = \hat{\beta}_{K_2}$.

Conversely suppose (15) is true for all $\hat{a}$ and all $\hat{B} = \hat{\Omega}^{-\frac{1}{2}} \bar{B}$ such that $\operatorname{rank}(\hat{K}_i^\top \bar{B}) = p$. We need to prove that $\hat{P}_1 = \hat{P}_2$. Now if (15) is true for all $\hat{a}$, then

$$(\hat{B}^\top \hat{P}_1 \hat{B})^{-1} \hat{B}^\top \hat{P}_1 = (\hat{B}^\top \hat{P}_2 \hat{B})^{-1} \hat{B}^\top \hat{P}_2,$$

which implies

$$\hat{P}_1 \hat{B} = \hat{P}_2 \hat{B} \hat{C} \tag{16}$$

with $\hat{C} := (\hat{B}^\top \hat{P}_2 \hat{B})^{-1} \hat{B}^\top \hat{P}_1 \hat{B}$. Multiplying (16) by $\hat{B}^\top$ yields $\hat{B}^\top \hat{P}_1 \hat{B} = \hat{B}^\top \hat{P}_2 \hat{B} \hat{C}$ and squaring (16) yields $\hat{B}^\top \hat{P}_1 \hat{B} = \hat{C}^\top \hat{B}^\top \hat{P}_2 \hat{B} \hat{C}$. The last two equations imply $\hat{C} = I$ and thus (16) reduces to $\hat{P}_1 \hat{B} = \hat{P}_2 \hat{B}$, or equivalently,

$$\hat{P}_1 \hat{\Omega}^{-\frac{1}{2}} \bar{B} = \hat{P}_2 \hat{\Omega}^{-\frac{1}{2}} \bar{B} \tag{17}$$

for all $\bar{B}$ such that $\text{rank}(\hat{K}_i^\top \bar{B}) = p$. Now consider the set $\mathfrak{B}$ of Lemma 2 with $K_i = \hat{K}_i$. For any $b \in \mathfrak{B}$ construct B according to Lemma 2, part 1, and call it $\bar{B}$. Then, by Lemma 2, $\text{rank}(\hat{K}_i^\top \bar{B}) = p$ and therefore (17) holds for this $\bar{B}$. As b is a column of $\bar{B}$, (17) implies $\hat{P}_1 \hat{\Omega}^{-\frac{1}{2}} b = \hat{P}_2 \hat{\Omega}^{-\frac{1}{2}} b$ for any $b \in \mathfrak{B}$. As by Lemma 2, part 2, one can select a basis for $\mathbb{R}^q$ out of the set $\mathfrak{B}$, it follows that $\hat{P}_1 = \hat{P}_2$ and thus $\{\hat{K}_1\} = \{\hat{K}_2\}$. $\quad \diamond$

Theorem 2 has some interesting consequences:

1. Let $P_K = \hat{K}(\hat{K}^\top \hat{K})^+ \hat{K}^\top$ be the orthogonal projection matrix derived from $\hat{K}$. Then $\{P_K\} = \{\hat{K}\}$ and thus $\hat{\beta}_{P_K} = \hat{\beta}_K$. Thus we can always resort to orthogonal projection matrices if we want to eliminate $\bar{A}$. A typical projection matrix to this purpose is

$$\hat{P} := I - \bar{A}(\bar{A}^\top \bar{A})^{-1} \bar{A}^\top. \tag{18}$$

   Its limit is $P := I - \tilde{A}(\tilde{A}^\top \tilde{A})^{-1} \tilde{A}^\top$. One can show that, due to the identifiability condition of the original model (1), $\text{rank}(P\tilde{B}) = \text{rank}(\tilde{B}) = p$ so that the identifiability condition given P is satisfied.

   Taking a more general elimination matrix $\hat{K}$ does not improve the efficiency of $\hat{\beta}_K$.

2. On the other hand, we need not take a projection matrix to eliminate $\bar{A}$. We can use other elimination matrices $\hat{K}$, thereby enhancing flexibility, even though we cannot increase efficiency. For example, suppose $A_n = \bar{A} = \iota$, ι being a $q \times 1$ vector consisting of ones. Then we can use the $q \times q$ projection matrix $P = I - \frac{1}{q}\iota\iota^\top$ to eliminate ι, but we may just as well use the $q \times (q-1)$ differencing matrix

$$\hat{K} = \Delta := \begin{pmatrix} 1 & 0 & \cdots & 0 \\ -1 & 1 & \cdots & 0 \\ \vdots & \vdots & \ddots & \vdots \\ 0 & 0 & \cdots & 1 \\ 0 & 0 & \cdots & -1 \end{pmatrix},$$

   which has the same column space as P but in contrast to P has full column rank, in fact, $P_\Delta := \Delta(\Delta^\top \Delta)^{-1} \Delta^\top = P$. Alternatively, we may simply delete the last column of P, which can formally be achieved by multiplying P from the right by the $q \times (q-1)$ matrix $J = (I, 0)^\top$ and use $\hat{K} = PJ$, see Schneeweiss *et al.*(2014) for a more complex example. The advantage of these alternative procedures is that in both cases the resulting matrix $\hat{K}$ has full column rank and the generalized inverse $(\hat{K}^\top \hat{\Omega} \hat{K})^+$ in (13) becomes a simple inverse.

3. More generally, any given $\hat{K}$ can be replaced with any matrix $\hat{K}_0$ the columns of which form a basis of $\{\hat{K}\}$, so that $\hat{\beta}_K = \hat{\beta}_{K_0}$ and $\hat{K}_0^\top \hat{\Omega} \hat{K}_0$ is invertible and its inverse can be used as an optimal weighting matrix. A matrix $\hat{K}_0$ can be found by a singular value decomposition or a rank factorization of $\hat{K}$.

When the column spaces of $\hat{K}_1$ and $\hat{K}_2$ differ, the two corresponding optimal estimators also differ, at least for some $\bar{B}$. Intuitively it seems clear that if $\{\hat{K}_1\} \supset \{\hat{K}_2\}$, then $\hat{\beta}_{K_1}$ is more efficient than $\hat{\beta}_{K_2}$. Actually, when we compare the asymptotic variances of the two estimators we do not compare the matrices $\hat{K}_1$ and $\hat{K}_2$ but rather their limits K_1 and K_2 and it is *their* column spaces which are relevant for the efficiency comparison. Before we state the corresponding theorem let us note that $\{K_1\} \supseteq \{K_2\}$ if, and only if, $K_2 = K_1 S$ with some $k_1 \times k_2$ matrix S. Clearly $K_1^\top \tilde{A} = 0$ implies $K_2^\top \tilde{A} = 0$, and if $K_2^\top \tilde{B}$ has full column rank so also has $K_1^\top \tilde{B}$.

Theorem 3 *Let $\hat{K}_1$ and $\hat{K}_2$ be two matrices that eliminate $\bar{A}$ and let K_1 and K_2 be their respective limits. Assume that* $\operatorname{rank}(K_i) \geq p$ *and* $\operatorname{rank}(K_i^\top \tilde{B}) = p$. *Then $\hat{\beta}_{K_1}$ is at least as efficient as $\hat{\beta}_{K_2}$ if, and only if, $\{K_1\} \supseteq \{K_2\}$.*

Proof: According to (14), $\hat{\beta}_{K_1}$ is at least as efficient as $\hat{\beta}_{K_2}$ iff

$$[\tilde{B}^\top K_1 (K_1^\top \Omega K_1)^+ K_1^\top \tilde{B}]^{-1} \leq [\tilde{B}^\top K_2 (K_2^\top \Omega K_2)^+ K_2^\top \tilde{B}]^{-1}, \tag{19}$$

Similarly as in the proof of Theorem 2 one can show with the help of Lemma 2 that (19) is true for all identifying $\tilde{B}$ given K_i, $i = 1, 2$, iff

$$\Omega^{\frac{1}{2}} K_1 (K_1^\top \Omega K_1)^+ K_1^\top \Omega^{\frac{1}{2}} \geq \Omega^{\frac{1}{2}} K_2 (K_2^\top \Omega K_2)^+ K_2^\top \Omega^{\frac{1}{2}}. \tag{20}$$

Now these last two matrices are orthogonal projection matrices, which we may denote by P_1 and P_2, respectively, so that (20) reads $P_1 \geq P_2$. This is equivalent to $\{\Omega^{\frac{1}{2}} K_1\} \supseteq \{\Omega^{\frac{1}{2}} K_2\}$, which again is equivalent to $\{K_1\} \supseteq \{K_2\}$. $\diamond$

As a consequence of Theorem 3, a matrix K_A that eliminates $\bar{A}$ and for which $\{K_A\}$ contains the column space of any other matrix $\hat{K}$ that eliminates $\bar{A}$ is most efficient among the class of GMM estimators that use any matrix $\hat{K}$ and an optimal weighting matrix. Such a K_A exists. It is given by the above mentioned orthogonal projection matrix (18) with its limit

$$P := I - \tilde{A}(\tilde{A}^\top \tilde{A})^{-1} \tilde{A}^\top = I - P_A,$$

or by any equivalent matrix K_A. (Indeed, for any K, $K^\top \tilde{A} = 0$ implies $P_K P_A = 0$ and thus $P_K P = P_K$, which implies $\{K\} \subseteq \{P\}$). Thus

$$\hat{\beta}_P = [\bar{B}^\top \hat{P} (\hat{P} \hat{\Omega} \hat{P})^+ \hat{P} \bar{B}]^{-1} \bar{B}^\top \hat{P} (\hat{P} \hat{\Omega} \hat{P})^+ \hat{P} \bar{a}$$

is the most efficient estimator. Its asymptotic variance is

$$\mathrm{Avar}(\hat{\beta}_P) = \frac{1}{N}[\tilde{B}^\top P(P\Omega P)^+ P\tilde{B}]^{-1}.$$

As might have been expected from linear regression theory, this most efficient estimator is the same as the optimal estimator of Section 3, see Putanen *et al.* (2013), 10.51.

Theorem 4 *The two estimators $\hat{\beta}_{opt}$ and $\hat{\beta}_P$ are the same if they use the same estimate $\hat{\Omega}$ of the covariance matrix Ω.*

Proof: We need to show that

$$\hat{\Omega}^{-\frac{1}{2}}\hat{P}_\Omega\hat{\Omega}^{-\frac{1}{2}} = \hat{P}(\hat{P}\hat{\Omega}\hat{P})^+\hat{P},$$

where $\hat{P}_\Omega$ is given by (9). Equivalently, we have to show that

$$\hat{P}_\Omega = \hat{\Omega}^{\frac{1}{2}}\hat{P}(\hat{P}\hat{\Omega}\hat{P})^+\hat{P}\hat{\Omega}^{\frac{1}{2}} =: Q.$$

Obviously Q is an orthogonal projection matrix, and so in order to prove the equality of $\hat{P}_\Omega$ and Q we need only show that both matrices have the same nullspace. Now the nullspace of $\hat{P}_\Omega$ is $\mathfrak{N}(\hat{P}_\Omega) = \{\hat{\Omega}^{-\frac{1}{2}}\bar{A}\}$, and since $Q\hat{\Omega}^{-\frac{1}{2}}\bar{A} = 0$, we have $\mathfrak{N}(\hat{P}_\Omega) \subseteq \mathfrak{N}(Q)$. Conversely let $x \in \mathfrak{N}(Q)$, then $Qx = 0$, which implies

$$0 = \hat{P}\hat{\Omega}\hat{P}(\hat{P}\hat{\Omega}\hat{P})^+\hat{P}\hat{\Omega}^{\frac{1}{2}}x = \hat{P}\hat{\Omega}^{\frac{1}{2}}x,$$

which means that $\hat{\Omega}^{\frac{1}{2}}x \in \mathfrak{N}(\hat{P})$. But $\mathfrak{N}(\hat{P}) = \{\bar{A}\}$ and so $x \in \{\hat{\Omega}^{-\frac{1}{2}}\bar{A}\} = \mathfrak{N}(\hat{P}_\Omega)$, so that $\mathfrak{N}(Q) \subseteq \mathfrak{N}(\hat{P}_\Omega)$. $\diamond$

5 Special case: A fixed

An important special case of the GMM model (6) arises when the matrix A_n is a fixed (i.e., non-stochastic) known matrix A independent of n. In this case, $\bar{A} = \tilde{A} = A$ and $\hat{P}_A = P_A = A(A^\top A)^{-1}A^\top$. More generally, we may have a fixed projection matrix P (or a corresponding fixed matrix K) such that $PA_n = 0$ for all n. In either case, things simplify greatly, the most important simplification being that now $P\Omega P = PWP$, where $W = \mathbb{E}(a_n - B_n\beta)$. This implies that in computing the optimal weight matrix we need only have a (preliminary) estimate of β and not an additional one of α, because now we can use

$$\hat{W} = \overline{(a - B\hat{\beta}_1)(a - B\hat{\beta}_1)^\top}$$

instead of $\hat{\Omega}$.

Another important aspect of this case is that the nuisance parameter, in an extension of the model, can now be an unknown random parameter. In this case the estimation of the whole model, as in Section 3, does not work but the elimination procedure still works. The following example illustrates this point.

Example 2: Consider a linear panel data model

$$y_n = \iota\alpha_n + x_n\beta + u_n, \quad \mathbb{E}u_n = 0, \tag{21}$$

where y_n, x_n, and u_n are $T \times 1$ i.i.d. random vectors , $n = 1,\ldots,N$, the former two observable, and ι is a $T \times 1$ vector of ones. Again, for similar reasons as in Example 1, u_n may not be independent of x_n, but m instrumental variables z_{in}, $i = 1,\ldots,m$, may be available. Let $z_n = (z_{1n},\ldots,z_{mn})^\top$, then we can set up the following system of equations

$$z_n \otimes y_n = (z_n \otimes \iota)\alpha_n + (z_n \otimes x_n)\beta + z_n \otimes u_n,$$

from which the moment conditions

$$\mathbb{E}[z_n \otimes y_n - (z_n \otimes \iota)\alpha_n - (z_n \otimes x_n)\beta] = 0$$

follow, which correspond to (6) with $a_n = z_n \otimes y_n$, $A_n = z_n \otimes \iota = (I_m \otimes \iota)z_n$, $B_n = z_n \otimes x_n$, and $q = mT$, however, with one notable difference. This time α_n is not a fixed parameter but rather a varying or random parameter, as the case may be, and possibly not independent of x_n or z_n. With the fixed projection matrix $P_\otimes := I_m \otimes (I - \frac{1}{T}\iota\iota^\top)$ we have $P_\otimes A_n = 0$. Instead of $P_\otimes$ we may also use $I_m \otimes \Delta^\top$ and $(I_m \otimes J^\top)P_\otimes$, where here Δ and J are $T \times (T-1)$, see the second remark after Theorem 2.

The following is a more elaborate example.

Example 3: Wansbeek (2001) studies a panel data model similar to the one in Example 2 but with measurement errors v in the x-variables:

$$\begin{aligned} y_n &= \iota_T\alpha_n + \xi_n\beta + \varepsilon_n \\ x_n &= \xi_n + v_n \end{aligned}$$

v_n, ε_n and ξ_n are supposed to be independent. In this case the (error prone) x_{tn} can be taken as instrumental variables. He ends up with the following moment conditions

$$\mathbb{E}\{M_R(I_T \otimes A_T)[x_n \otimes (y_n - x_n\beta)]\} = 0, \tag{22}$$

where $A_T = I_T - \frac{1}{T}\iota_T\iota_T^\top$, $M_R = I_{T^2} - R(R^\top R)^+R^\top$, $R = (I_T \otimes A_T)R_0$, and R_0 is a known $T^2 \times m$ matrix describing the structure of the measurement error variances and covariances, see also Xiao *et al.* (2010). Now $M := M_R(I_T \otimes A_T)$ is a fixed projection matrix eliminating both $I_T \otimes \iota$ and the measurement error structure matrix R_0. So (22) corresponds to (11) with $a_n = x_n \otimes y_n$, $B_n = x_n \otimes x_n$, and $K = M$. The optimal weighting matrix in this case is

$$V_{opt} = (MWM)^+,$$

where $W = \mathbb{E}\{[x_n \otimes (y_n - x_n\beta)][x_n \otimes (y_n - x_n\beta)]^\top\}$, and the efficient estimator of β according to (13) is

$$\hat{\beta}_M = \left[\overline{(x\otimes x)}^\top M(M\hat{W}M)^+M\overline{(x\otimes x)}\right]^{-1}\overline{(x\otimes x)}^\top M(M\hat{W}M)^+M\overline{(x\otimes y)},$$

where

$$\hat{W} = \overline{[x\otimes(y-x\hat{\beta}_1)][x\otimes(y-x\hat{\beta}_1)]^\top}.$$

According to (14) it has asymptotic variance

$$\mathrm{Avar}(\hat{\beta}_M) = \frac{1}{N}\{\mathbb{E}(x_n\otimes x_n)^\top M(MWM)^+M\mathbb{E}(x_n\otimes x_n)\}^{-1}.$$

6 Singular covariance matrix

The difficulty we encountered in finding an optimal weighting matrix originated from the singularity of the covariance matrix of the moment conditions after these were corrected for nuisance parameters. This difficulty, however, was easily overcome by using a generalized inverse instead of the more common ordinary inverse of the covariance matrix as weighting matrix. The question is whether the same strategy can be applied when the covariance matrix Ω and its estimate $\hat{\Omega}$ happen to be singular at the outset for whatever reason. Thus suppose that in the original model (1) of Section 2 the covariance matrix Ω is singular even without multiplying (1) by some eliminating matrix $K^\top$. For simplicity let us also suppose that Ω is known so that we need not estimate it. Then a valid estimator of γ can be constructed by replacing $\hat{\Omega}^{-1}$ in (4) by Ω^+. Thus let

$$\hat{\gamma}_* = (\bar{C}^\top\Omega^+\bar{C})^{-1}\bar{C}^\top\Omega^+\bar{a} \tag{23}$$

be our new estimator of γ in this case. Its asymptotic variance is given by

$$\mathrm{Avar}(\hat{\gamma}_*) = \frac{1}{N}(\tilde{C}^\top\Omega^+\tilde{C})^{-1},$$

assuming that $\tilde{C}^\top \Omega^+ \tilde{C}$ is nonsingular. However, this estimator is not necessarily optimal in the class of estimators (2). The following counterexample demonstrates this. For $q = 2$ and $c = 1$, let

$$C_n = \begin{pmatrix} c_{1n} \\ c_{2n} \end{pmatrix}, \quad a_n = \begin{pmatrix} a_{1n} \\ a_{2n} \end{pmatrix}, \quad \Omega = \begin{pmatrix} \omega_{11} & 0 \\ 0 & 0 \end{pmatrix}, \quad \tilde{c}_1 \neq 0, \quad \tilde{c}_2 \neq 0, \quad \omega_{11} \neq 0.$$

Then $\Omega^+ = \begin{pmatrix} \omega_{11}^{-1} & 0 \\ 0 & 0 \end{pmatrix}$ and, assuming $\bar{c}_1 \neq 0$,

$$\hat{\gamma}_* = (\omega_{11}^{-1} \bar{c}_1^2)^{-1} \omega_{11}^{-1} \bar{c}_1 \bar{a}_1 = \frac{\bar{a}_1}{\bar{c}_1},$$

which has asymptotic variance

$$\text{Avar}(\hat{\gamma}_*) = \frac{1}{N} \frac{\omega_{11}}{\tilde{c}_1^2}.$$

But the alternative estimator

$$\hat{\gamma}_{**} = \frac{\bar{a}_2}{\bar{c}_2}$$

of γ has asymptotic variance 0, because $a_{2n} = c_{2n}\gamma$ holds exactly due to $\omega_{22} = 0$. Thus $\hat{\gamma}_*$, which was constructed with weighting matrix Ω^+, is not optimal.

But when is the weighting matrix Ω^+ optimal? The next theorem gives an answer. (For an analogous result in linear regression analysis see Putanen *et al.* (2013), 10.20).

Theorem 5 *Suppose that in the linear GMM model* (1) $\{\Omega\} \supseteq \{\tilde{C}\}$ *and* $\tilde{C}^\top \Omega^+ \tilde{C}$ *is nonsingular, then the estimator* $\hat{\gamma}_*$ *of* (23) *is efficient in the class of estimators* $\hat{\gamma}_V$ *of* (2)*.*

Proof: First note that $\{\Omega\} \supseteq \{\tilde{C}\}$is equivalent to the relation $\Omega H = \tilde{C}$ with some $(q \times c)$ matrix H. Now let

$$\begin{aligned} D^\top &:= (\tilde{C}^\top V \tilde{C})^{-1} \tilde{C}^\top V \Omega^{\frac{1}{2}} \\ F^\top &:= \tilde{C}^\top \Omega^+ \Omega^{\frac{1}{2}}. \end{aligned}$$

Then

$$\begin{aligned} D^\top F &= (\tilde{C}^\top V \tilde{C})^{-1} \tilde{C}^\top V \Omega \Omega^+ \tilde{C} \\ &= (\tilde{C}^\top V \tilde{C})^{-1} \tilde{C}^\top V \Omega \Omega^+ \Omega H \\ &= (\tilde{C}^\top V \tilde{C})^{-1} \tilde{C}^\top V \Omega H \\ &= (\tilde{C}^\top V \tilde{C})^{-1} \tilde{C}^\top V \tilde{C} = I \end{aligned}$$

Therefore, with $G = F(F^\top F)^{-1}F^\top$,

$$N\,[\text{Avar}(\hat{\gamma}_V) - \text{Avar}(\hat{\gamma}_*)] = D^\top D - (F^\top F)^{-1} = D^\top (I - G)D \geq 0. \quad \diamond$$

We may note that there is a strong connection of this section with Section 4. Theorem 1 can be seen to follow from Theorem 5. Let K be the matrix defined in Section 4 which eliminates $\tilde{A}$ and let Ω be the nonsingular covariance matrix of Section 4. Define $\Omega_0 := K^\top \Omega K$, $\tilde{a}_0 := K^\top \tilde{a}$, and $\tilde{B}_0 := K^\top \tilde{B}$. Then the model of Theorem 1 is $\tilde{a}_0 - \tilde{B}_0\beta_0 = 0$, which corresponds to the model of Theorem 5 if we let Ω_0, $\tilde{a}_0$, $\tilde{B}_0$, β_0 correspond to Ω, $\tilde{a}$, $\tilde{C}$, γ_0, respectively. We see that the main condition of Theorem 5 is satisfied because, for any $k \times 1$ vector x, $\Omega_0 x = 0$ implies $Kx = 0$ and thus $\tilde{B}_0^\top x = 0$, i.e., $\mathfrak{N}(\Omega_0) \subseteq \mathfrak{N}(\tilde{B}_0^\top)$ or, equivalently, $\{\Omega_0\} \supseteq \{\tilde{B}_0\}$. Also the identifiability condition of Theorem 1, i.e., $\text{rank}(K^\top \tilde{B}) = p$ implies that $\tilde{B}_0^\top \Omega_0^+ \tilde{B}_0$ is nonsingular. So we can apply Theorem 5, from which follows that an estimate of $\Omega_0^+ = (K^\top \Omega K)^+$ is the optimal weighting matrix for the GMM estimator of β in Theorem 1.

7 Conclusion

The paper discusses the linear GMM model with a singular covariance matrix. To construct an efficient GMM estimator one needs to have an optimal weighting matrix. Typically one would take the inverse of the covariance matrix, but when the latter is singular, its generalized (Moore-Penrose) inverse has to be taken instead. In the last section of the paper a criterion is developed that makes sure that this, indeed, produces an efficient estimator. The column space of the covariance matrix should contain the column space of the mean of the data matrix. This criterion is satisfied when in a linear GMM model the moment conditions are transformed by multiplying them with some matrix, with the purpose of eliminating nuisance parameters together with the corresponding sub-matrix of the data matrix. There are various ways to select such a "purging" matrix and their relations to each other are discussed.

Acknowledgment: Thanks go to Shalabh for pointing out some important literature.

8 Bibliography

Arellano M. (2003): *Panel Data Econometrics*. Oxford University Press, Oxford.

Biørn E. and Klette T.J. (1998): Panel data with errors-in-variables: Essential and redundant orthogonality conditions in GMM-estimation. *Economics Letters* **59**, 275-282.

Dorana H. and Schmidt P. (2006): GMM estimators with improved finite sample properties using principal components of the weighting matrix, with an application to the dynamic panel data model. *Journal of Econometrics* **133**, 387-409.

Hall A. R. (2005): *Generalized Method of Moments*. Oxford University Press, Oxford.

Harris D. and Mátyás L. (1999): Introduction to the generalized method of moments estimation. Chapter 1 in: László Mátyás (ed.): *Generalized Method of Moments Estimation*. Cambridge University Press, Cambridge.

Hansen L. P. (1982): Large sample properties of Generalized Method of Moments estimators. *Econometrica* **50**, 1029-54.

Putanen S., Styan G., and Isotalo J. (2013): *Formulas Useful for Linear Regression Analysis and Related Matrix Theory.* Springer, New York.

Rao C. R., Toutenburg H., Shalabh, Heumann C. (2008): *Linear Models and Generalizations: Least Squares and Alternatives*. 3^{rd} ed. Springer, Heidelberg.

Schneeweiss H., Ronning G., and Schmid M. (2014): Panel model with multiplicative measurement errors. In: Beran Jan, Feng Yuanhua, Hebbel Hartmut (eds): *Empirical Economic and Financial Research - Theory, Methods and Practice. Festschrift in honour of Prof. Siegfried Heiler*. Series: Advanced Studies in Theoretical and Applied Econometrics. Springer, 119 - 139.

Seber G. A. and Lee A. J. (2006): *Linear Regression Analysis*. Wiley, New York.

White, H. (1986): Instrumental variables analogs of generalized least squares estimators. *Advances in Statistical Analysis and Statistical Computing: Theory and Applications* **1**, 173 - 227.

Xiao Zhiguo, Shao Jun, Xu Ruifeng, and Palta Mari (2007): Efficiency of GMM estimation in panel data models with measurement error. *Sankhya: The Indian Journal of Statistics* **69**, 101-118.

Xiao Zhiguo, Shao Jun, and Palta Mari (2010): Instrumental variable and GMM estimation for panel data with measurement error. *Statistica Sinica* **20**, 1725-1741.