UNIVERSITY OF CAMBRIDGE

Cambridge Working Papers in Economics

IV Estimation of Panels with Factor Residuals

Donald Robertson and Vasilis Sarafidis

CWPE 1321

IV Estimation of Panels with Factor Residuals

Donald Robertson and Vasilis Sarafidis*

This paper proposes a new instrumental variables approach for consistent and asymptotically efficient estimation of panel data models with weakly exogenous or endogenous regressors and residuals generated by a multi-factor error structure. In this case, the standard dynamic panel estimators fail to provide consistent estimates of the parameters. The novelty of our approach is that we introduce new parameters to represent the unobserved covariances between the instruments and the factor component of the residual; these parameters are estimable when N is large. Some important estimation and identification issues are studied in detail. The finite sample performance of the proposed estimators is investigated using simulated data. The results show that the method produces reliable estimates of the parameters over several parametrisations.

KEYWORDS: Generalised Method of Moments, Dynamic Panel Data, Factor Residuals.

JEL Classification: C23, C26.

---

*University of Cambridge, and Monash University and the University of Sydney. We gratefully acknowledge financial support from the Research Unit of the Faculty of Economics and Business, University of Sydney. An earlier version of this paper circulated as Robertson, Sarafidis and Symons. We are particularly grateful to Jim Symons for his enormous contribution to this work. We would also like to thank two anonymous referees for providing us with several constructive comments and suggestions, and seminar participants at the ESWC 2010 and SETA 2011. All remaining errors are ours.

1 Introduction

This paper develops a new approach that is based on instrumental variables for consistent and asymptotically efficient estimation of panel data models with errors generated by a multi-factor structure. The factor structure is an attractive framework as it permits general forms of unobserved heterogeneity that may otherwise contaminate estimation and statistical inference. Factor residuals can be motivated in several ways, depending on the application in mind. In macroeconometric panels, the factors may be thought of as economy wide shocks that affect all individuals, albeit with different intensities; essentially, this allows cross sections to inhabit a common environment, to which they may respond differently. In microeconometric panels, the factor structure may capture different sources of unobserved individual-specific heterogeneity, the impact of which varies intertemporally in an arbitrary way. For instance, in studies of production functions, the factor loadings may capture distinct components of firm-specific technical efficiency, which varies through time. In models of earnings determination, the factor loadings may reflect an individual's several different unobserved skills, while the factors represent the industry wide price of these skills, which is not necessarily constant over time (see also the detailed discussions in Ahn, Lee and Schmidt (2001, 2010) and Bai (2009)). Systematic changes in tastes is another plausible example. In some circumstances such variables could be measured and directly included in the model, but often the details of measurement might be difficult, contentious and, in any case, outside the focus of the analysis.[1] In such cases it is inviting to allow the model residual to be composed of one or more unspecified factors, themselves to be estimated. One can interpret such a procedure as allowing some degree of cross-sectional dependence in the model residuals. An overview of the current literature on panel data models with error cross-sectional dependence is provided by Sarafidis and Wansbeek (2012).

Consider the simplest case of a one regressor, one factor model in the standard form

$$y_{it} = \phi x_{it} + \lambda_i f_t + \varepsilon_{it} \quad t = 1, ..., T \; i = 1, ..., N. \tag{1.1}$$

In some cases the values of f_t, or λ_i, are assumed to be known, such as when fitting

[1] For example, how does one measure monetary shocks? Does one look at interest rates or monetary aggregates? Which monetary aggregates? How does one handle financial innovation?

an error components structure, or polynomial time trends, but here we shall treat the f's as vectors of parameters to be estimated. In this case, one might fit this model by nonlinear least squares based on principal components analysis; see e.g. Bai (2009). Pesaran (2006) suggests the alternative of augmenting the regression model by the cross-sectional averages of the variables, y_{it} and x_{it}, which will span the unobserved factors for large N. Both these methods require that the set of regressors is strictly exogenous with respect to the idiosyncratic error component, ε_{it}, and N, T are both large. Sarafidis and Yamagata (2013) propose an IV estimator in a model with a lagged dependent variable and strictly exogenous covariates. The procedure involves defactoring the exogenous covariates using principal components analysis, which then are used to form a set of instruments. The proposed estimator is valid for large N and T. In the present paper we focus on the case where N is large, T fixed and some (or all) of the regressors are either weakly exogenous, or endogenous with respect to ε_{it}. This scenario is empirically relevant in many economic applications. For example, our framework allows models with lags of the dependent variable on the right hand side, as in partial adjustment models for labour supply, Euler equations for household consumption, and empirical growth models. In these models the coefficient of the lagged dependent variable captures inertia, habit formation and costs of adjustment and therefore it has structural significance (see e.g. Arellano, 2003, Ch. 7). Furthermore, since underlying economic behaviour is intrinsically dynamic, past residual errors might influence the current value of explanatory variables even when lagged dependent variables are not directly present in the model, leading to weak exogeneity. For instance, in panels of observations on economies, expectational errors are likely to work through the whole economy over time, and it is natural to expect that a given variable is often not immune from this process (see e.g. Sarafidis and Wansbeek, 2012). Finally, our framework also permits endogenous regressors, due to (say) errors of measurement, omitted variables and/or simultaneity. As a result, it possesses an appealing generality.

When unobserved heterogeneity is subject to an error components structure, a popular method to estimate models with weakly exogenous, or endogenous regressors is the Generalised Method of Moments (GMM), analysed in the dynamic panel data context by Arellano and Bond (1991), Ahn and Schmidt (1995), Arellano and Bover (1995), Blundell and Bond (1998) and others. Among the many economic applications where GMM has been used include estimation of (i) production functions and technological spillovers (e.g. Blundell and Bond, 2000), (ii) the demand for money (e.g. Bover and Watson, 2005), (iii) the responsiveness of labor supply to wages (e.g. Ziliak, 1997), (iv) the structure and profitability of the banking sector (e.g. Tregenna, 2009) and the empirical growth literature (e.g. Presbitero, 2008). In all these applications the set of regressors includes weakly exogenous variables, the cross-sectional dimension is fairly large and T is relatively small.[2]

However, as shown by Sarafidis and Robertson (2009) and Sarafidis, Yamagata and Robertson (2009), these procedures fail to provide consistent estimates of the parameters when the errors are generated by a multi-factor structure because the moment conditions they utilise are invalidated. Panel data models with a single factor structure and a small number of time series observations have been studied by Holtz-Eakin, Newey and Rosen (1988), Ahn, Lee and Schmidt (2001) and Nauges and Thomas (2003). All these studies utilise some form of quasi-differencing that eliminates the factor component from the residuals. More recently, in a seminal paper Ahn, Lee and Schmidt (2010) develop a GMM estimator that allows for multiple factors using multi-quasi-differencing.

In this paper we develop a new instrumental variables approach; instead of eliminating the factors using some form of quasi-differencing, our methodology introduces new parameters that represent the unobserved covariances between the instruments and the factor component of the residual. The proposed estimator is more efficient than the existing quasi-differencing type GMM estimators and attains the semi-parametric efficiency bound discussed by Newey (1990).

The basic intuition behind our method is as follows. Assume in the above model we have some variable (instrument) z_{it} for which the moment condition $E(z_{it}\varepsilon_{it}) = 0$ holds true. This implies that

$$E(z_{it}y_{it}) = \phi E(z_{it}x_{it}) + g_t f_t, \tag{1.2}$$

where $g_t = E(z_{it}\lambda_i)$. We treat the$g$'s as parameters to be estimated. Replacing the expectations with their sample moments, one has T such orthogonality conditions and $2T + 1$ parameters to be estimated (ϕ and the g's and f's): too many to be identified. However, if all lags of z_{it} are instruments, the number of orthogonality conditions is expanded to $T(T + 1)/2$, while the number of parameters remains the same; one has now more moment conditions than parameters for $T > 3$, so identification becomes feasible. We shall call estimators in this class Factor In-

[2]In particular, Blundell and Bond (2000) use a panel of 509 R&D performing US manufacturing companies, Bover and Watson (2005) use data on 5,649 firms operating in Spain, Ziliak (1997) surveys 534 individuals, Tregenna (2007) considers 644 banking institutions, while Presbitero (2008) utilises data from 144 countries. T ranges from 5 to 27 in these applications.

strumental Variables (FIV) estimators. FIV estimators have been introduced by Robertson and Symons (2007); the present treatment greatly improves and extends that paper. FIV estimators have the traditional attraction of method of moments estimators in that they exploit only the orthogonality conditions implied by the structure of the model, which in fact may be the implication of an underlying economic theory, and make no use of subsidiary assumptions such as homoskedasticity or other assumed distributional properties of the error process. The method is general in the sense that all that is required is the existence of some instrument z_{it} with orthogonality conditions at sufficiently many periods other than t to identify the model parameters.

In most practical circumstances the instrument set will include lags of the dependent and independent variables of the model. In this case, a set of linear restrictions can be demonstrated to hold among the parameters (ϕ and the g's and f's) of the model, leading to greater estimation efficiency. We call this estimator FIVR (restricted FIV) in contrast to the estimator obtained when these restrictions are not imposed, FIVU (unrestricted FIV). FIVR is asymptotically efficient in the class of estimators that make use of second moment information. Using simulated data we show that these extra restrictions can be important and FIVR largely outperforms FIVU in terms of RMSE.

2 Stochastic Framework

We consider the following model:

$$\mathbf{x}_{it}'\boldsymbol{\beta} = \boldsymbol{\lambda}_i'\mathbf{f}_t + \varepsilon_{it}, \quad i = 1, ..., N, \ t = 1, ..., T, \tag{2.1}$$

where $\mathbf{x}_{it} = (y_{it}, x_{1it}, x_{2it}, \ldots, x_{q-1,it})'$ is a $q \times 1$ vector containing the (endogenous and exogenous) observed variables. The $q \times 1$ vector $\boldsymbol{\beta}$ is assumed to be a function of r free parameters $\boldsymbol{\phi}$:

$$\boldsymbol{\beta} = \boldsymbol{\beta}(\boldsymbol{\phi}).$$

In the work below we shall usually take $\boldsymbol{\beta} = (1, -\boldsymbol{\phi}')'$ where $\boldsymbol{\phi}$ is a $(q-1) \times 1$ vector of parameters. $\boldsymbol{\lambda}_i$ is a stochastic $n \times 1$ vector of factor loadings and $\mathbf{f}_t$ is an $n \times 1$ vector of factors which are treated as time-specific parameters; ε_{it} is a purely idiosyncratic disturbance.[3] The sampling structure is that we have N sufficiently

[3]We shall treat n as known. The results presented below are not affected when the number of factors is unknown and is estimated consistently. A formal proof for this argument is provided

independent draws, indexed by i, from the population. The following assumption is made:

ASSUMPTION 1. Existence of instruments. We assume potential instruments are given by a vector $\mathbf{w}_i$ of dimension d; these instruments may correspond to the variables of the model or be extraneous variables. In each period t, c_t instruments are available, expressed in vector form as follows:

$$\mathbf{w}_{it} = S_t \mathbf{w}_i, \tag{2.2}$$

for which the condition $E(\mathbf{w}_{it}\varepsilon_{it}) = \mathbf{0}$ holds.

Here S_t is the selector matrix of 0's and 1's that picks out from all potential instruments in $\mathbf{w}_i$, those that are valid at date t. The matrix S_t has dimension $c_t \times d$ where c_t is the number of orthogonality conditions associated with ε_{it}. The instruments available depend on the structure of the model. Thus, for example, in a model with a single explanatory variable, $\mathbf{w}_i$ could consist of all values of this variable, from $t = 1$ to $t = T$. If the variable was strictly exogenous with respect to ε_{it} then S_t would be the identity matrix I_T at each t. If the variable was only weakly exogenous then the selector matrix for each t would pick out values dated t and earlier, provided that $E(\varepsilon_{is}\varepsilon_{it}) = 0,\ s \neq t$. Mixed cases can occur naturally, such as when the covariates consist of (say) a weakly exogenous and an endogenous variable. In this case, $\mathbf{w}_i$ is a $2T \times 1$ vector and the selector matrix will pick out current and lagged values of the weakly exogenous variable, as well as the appropriate dates for the endogenous variable.

The model (2.1) can be stacked over t to take the form

$$X_i\boldsymbol{\beta} = (I_T \otimes \boldsymbol{\lambda}_i')\mathbf{f} + \boldsymbol{\varepsilon}_i, \tag{2.3}$$

where $X_i = [\mathbf{x}_{i1}, ..., \mathbf{x}_{iT}]'$, $\mathbf{f} = vec(F')$, $F = [\mathbf{f}_1, ..., \mathbf{f}_T]'$, $\boldsymbol{\varepsilon}_i = (\boldsymbol{\varepsilon}_{i1}, ..., \boldsymbol{\varepsilon}_{iT})'$.

by Bai (2003, footnote 5). A consistent estimate of the number of factors in this context can be obtained using a sequential method based on Sargan's overidentifying restrictions test statistic. The intuition is that when the number of factors fitted is smaller than the true value, Sargan's statistic will reject the null hypothesis for N sufficiently large. Alternatively, one can estimate the number of factors consistently using an information based criterion. Ahn, Lee and Schmidt (2010) provide specific details and proofs for both methods. See also Sarafidis and Yamagata (2010) for further discussion.

The corresponding instrument matrix Z_i is defined by

$$Z_i' = \begin{bmatrix} \mathbf{w}_{i1} & 0 & .. & 0 \\ 0 & \mathbf{w}_{i2} & & 0 \\ \vdots & & \ddots & \\ 0 & 0 & .. & \mathbf{w}_{iT} \end{bmatrix}, \tag{2.4}$$

such that

$$E(Z_i'\boldsymbol{\varepsilon}_i) = \mathbf{0}, \tag{2.5}$$

where Z_i' is $c \times T$ and $c = \sum_{t=1}^{t=T} c_t$ denotes the total number of moment conditions. In view of (2.2), the matrix of instruments can be written as

$$Z_i' = S(I_T \otimes \mathbf{w}_i), \tag{2.6}$$

where

$$S = \begin{bmatrix} S_1 & 0 & .. & 0 \\ 0 & S_2 & .. & 0 \\ \vdots & \vdots & \vdots & \vdots \\ 0 & 0 & .. & S_T \end{bmatrix}. \tag{2.7}$$

The matrix S has dimension $c \times Td$. The vector of orthogonality conditions we use to estimate the model parameters is

$$E[Z_i'X_i\boldsymbol{\beta} - Z_i'(I_T \otimes \boldsymbol{\lambda}_i')\mathbf{f} - Z_i'\boldsymbol{\varepsilon}_i] = \mathbf{0}, \tag{2.8}$$

which, by use of (2.5) and (2.6), can be written as follows:

$$M\boldsymbol{\beta} - S(I_T \otimes G)\mathbf{f} = \mathbf{0}, \tag{2.9}$$

where $M = E(Z_i'X_i)$ and $G = E(\mathbf{w}_i\boldsymbol{\lambda}_i')$. Matrices M and G have dimensions $c \times q$ and $d \times n$, respectively. Alternative forms of the second term in (2.9) are

$$S(I_T \otimes G)\mathbf{f} = S\text{vec}(GF') = S(F \otimes I_d)\mathbf{g}, \tag{2.10}$$

where $\mathbf{g} = \text{vec}(G)$. A compact expression of the orthogonality conditions is thus

$$M\boldsymbol{\beta} - S\text{vec}(GF') = \mathbf{0}. \tag{2.11}$$

When the instruments consist of current and all lagged values: the canonical case As an example, consider the case where all instruments available can be naturally arranged in a $T \times p$ matrix V_i of T observations on p variables (so that $\mathbf{w}_i = \text{vec}(V_i)$), and ε_{it} is orthogonal to the block of potential instruments from $s = 1$ to $s = t$, i.e. the orthogonality conditions are given by

$$E(\mathbf{z}_{is}\varepsilon_{it}) = \mathbf{0},\ t = 1, ..., T;\ s = 1, ..., t, \tag{2.12}$$

where $\mathbf{z}'_{is}$ is the s^{th} row of V_i. This can be viewed as a canonical case in the sense that there exists a collection of contemporaneous instruments and their lagged values; it arises, for example, when all variables in the model are weakly exogenous, such as in the AR(1) dynamic panel data model with factor residuals (in which case $p = 1$). Define $M_{st} = E(\mathbf{z}_{is}\mathbf{x}'_{it})$ and $\mathrm{G}_s = E(\mathbf{z}_{is}\boldsymbol{\lambda}'_i)$, which have dimensions $p \times q$ and $p \times n$, respectively. The orthogonality conditions are given by

$$M_{st}\boldsymbol{\beta} - G_s\mathbf{f}_t = \mathbf{0},\ \ t = 1, ..., T;\ \ s = 1, ..., t. \tag{2.13}$$

These conditions can be stacked as

$$\begin{bmatrix} M_{11}\boldsymbol{\beta} \\ M_{12}\boldsymbol{\beta} \\ M_{22}\boldsymbol{\beta} \\ \vdots \\ M_{1T}\boldsymbol{\beta} \\ M_{2T}\boldsymbol{\beta} \\ \vdots \\ M_{TT}\boldsymbol{\beta} \end{bmatrix} - \begin{bmatrix} G_1\mathbf{f}_1 \\ G_1\mathbf{f}_2 \\ G_2\mathbf{f}_2 \\ \vdots \\ G_1\mathbf{f}_T \\ G_2\mathbf{f}_T \\ \vdots \\ G_T\mathbf{f}_T \end{bmatrix} = \mathbf{0}. \tag{2.14}$$

More succinctly, this is

$$M\boldsymbol{\beta} - \text{vech}(GF') = \mathbf{0}, \tag{2.15}$$

where M is the stacked M_{st} terms and the vech operator is understood to act on $p \times 1$ submatrices. Let $\widetilde{S}_T$ be the selector matrix of 0's and 1's that turns vec into vech (acting on $T \times T$ matrices). Then

$$M\boldsymbol{\beta} - \text{vech}(GF') = M\boldsymbol{\beta} - (\widetilde{S}_T \otimes I_p)\text{vec}(GF') = \mathbf{0}, \tag{2.16}$$

which is of the form of (2.11), with the selector matrix S given by $S = \widetilde{S}_T \otimes I_p$.

3 The unrestricted estimator FIVU

Define the following moment function:

$$\boldsymbol{\psi}(\boldsymbol{\theta}, Z_i'X_i) = Z_i'X_i\boldsymbol{\beta}(\boldsymbol{\phi}) - S\text{vec}(GF'), \tag{3.1}$$

where $\boldsymbol{\theta} = (\boldsymbol{\phi}', \boldsymbol{g}', \mathbf{f}')'$. Then by construction $E(\boldsymbol{\psi}(\boldsymbol{\theta})) = \mathbf{0}$ at the true value $\boldsymbol{\theta}_0$. Our aim is to estimate $\boldsymbol{\theta}_0$ by minimising $\boldsymbol{\psi}(\boldsymbol{\theta}, \widehat{M})'C\boldsymbol{\psi}(\boldsymbol{\theta}, \widehat{M})$ where $\widehat{M} = \sum_{i=1}^{N} Z_i'X_i/N$ is the matrix of sample moments and C is a given fixed matrix. As it stands, the model is not identified because

$$M\boldsymbol{\beta} - S\text{vec}(GF') = M\boldsymbol{\beta} - S\text{vec}(GUU^{-1}F'), \tag{3.2}$$

for any $n \times n$ invertible U. This particular indeterminancy is typically eliminated by requiring an $n \times n$ submatrix of F' to be the identity matrix. However, it turns out that this identity restriction on a submatrix of F is not in general sufficient for full identification so that further restrictions may be required. The required restrictions will vary depending upon the specification of the model. In what follows we provide sufficient conditions for identification of the full parameter vector $\boldsymbol{\theta}$ and illustrate with an example.

Let Ω be the full set of possible parameter vectors.

ASSUMPTION 2. $\boldsymbol{\theta}_0$ belongs to the interior of $\Theta_r \subseteq \Omega$ where Θ_r is obtainable by restrictions on the G, F components of the vectors in Ω, together with some possible further restrictions excluding a closed set. We assume $\boldsymbol{\theta}_0$ is identified on Θ_r in the sense that $E(\boldsymbol{\psi}(\boldsymbol{\theta})) = \mathbf{0}$ for $\boldsymbol{\theta} \in \Theta_r$ implies $\boldsymbol{\theta} = \boldsymbol{\theta}_0$.

Let

$$\Gamma = E\left(\frac{\partial \boldsymbol{\psi}}{\partial \boldsymbol{\theta}_r'}(\boldsymbol{\theta}_0)\right), \tag{3.3}$$

and

$$\Delta = E\left(\boldsymbol{\psi}(\boldsymbol{\theta}_0)\boldsymbol{\psi}(\boldsymbol{\theta}_0)'\right), \tag{3.4}$$

where $\boldsymbol{\theta}_r$ consists of the free parameters in a restricted $\boldsymbol{\theta}$.

ASSUMPTION 3. Γ and Δ exist and are uniformly positive definite.

ASSUMPTION 4. $E|z_{it\ell}^2|^{1+\delta} < \varpi < \infty$ and $E|x_{it\ell}^2|^{1+\delta} < \varpi < \infty$ for some $\delta > 0$ and for all i, t, ℓ, and the function $\boldsymbol{\beta}(.)$ is twice continuously differentiable.

As an example, consider the AR(1) one factor model:

$$y_{it} = \phi y_{it-1} + \lambda_i f_t + \varepsilon_{it}.$$

The moment conditions are derived from taking expectations in

$$y_{is} y_{it} = \phi y_{is} y_{it-1} + y_{is} \lambda_i f_t + y_{is} \varepsilon_{it}, \;\; s = 0, \ldots, t-1; \;\; t = 1, \ldots, T.$$

The full parameter vector $\boldsymbol{\theta} = (\phi, g_1, \ldots, g_T, f_1, \ldots, f_T) \in \Omega$ is not identified at $\boldsymbol{\theta}_0$. One identifying restriction is simply a rescaling of $\mathbf{g}$ and $\mathbf{f}$, obtained by setting one entry in $\mathbf{f}$ equal to 1, e.g. $f_T = 1$. In this model one column of the matrix Γ consists of zeros except for a single entry that equals g_1, so the full rank condition for Γ requires as well that $g_1 \neq 0$. Thus we may take $\Theta = \{\boldsymbol{\theta} = (\phi, g_1, \ldots, g_T, f_1, \ldots, f_T);$ $g_1, \; f_T \neq 0\} \subset \Omega$ and $\Theta_r = \{\boldsymbol{\theta} = (\phi, g_1, \ldots, g_T, f_1, \ldots, f_T); \; g_1 \neq 0, \; f_T = 1\} \subset \Theta$.

Assumption 2 implies that the model is identified. Some plausible models will not fit this framework. For instance if the set of instruments is not correlated with the factor loadings (as it would occur in an uncorrelated random effects formulation), then all the g's will be zero, which implies there can be no restrictions that would allow identification of the f's. However, this case is trivial as pointed out by Ahn, Lee and Schmidt (2010), because the structural parameter vector ϕ can be straightforwardly estimated by OLS.

Note that the positive definiteness assumption for Γ itself implies that $\boldsymbol{\theta}_0$ is locally identified. The above set of assumptions is sufficient to make an appeal to standard GMM theory in order to derive the asymptotic properties of FIVU. In our context the result is given in the following proposition:

Proposition 1. Distributional result for FIVU. *Let Θ_c be a compact subset of Θ_r containing $\boldsymbol{\theta}_0$ in its interior and let*

$$\widehat{\boldsymbol{\theta}}(\Theta_c) = \arg \min_{\boldsymbol{\theta} \in \Theta_c} \boldsymbol{\psi}(\boldsymbol{\theta}, \widehat{M})' C \boldsymbol{\psi}(\boldsymbol{\theta}, \widehat{M}), \tag{3.5}$$

recalling that $\widehat{M} = \sum_{i=1}^{N} Z_i' X_i / N$ and C is a given fixed positive definite matrix. Then $\widehat{\boldsymbol{\theta}}$ converges in probability to $\boldsymbol{\theta}_0$ and

$$\sqrt{N}(\widehat{\boldsymbol{\theta}} - \boldsymbol{\theta}_0) \xrightarrow{d} N(\mathbf{0}, (\Gamma' C \Gamma)^{-1} (\Gamma' C \Delta C \Gamma)(\Gamma' C \Gamma)^{-1}). \tag{3.6}$$

Proof. This is straightforward enough; see e.g. Newey and McFadden (1994) for further details.[4] $\square$

If C is chosen as Δ^{-1} the covariance matrix of the asymptotic distribution is $(\Gamma'\Delta^{-1}\Gamma)^{-1}$, in which case the estimator has certain optimality properties. These distributional results hold as well if the unobserved Δ is replaced by a consistent estimate.

Appendix II establishes a general identification scheme for FIVU under a multi-factor structure. In many circumstances, the full parameter vector is not the object of interest and one is interested only in estimates of $\boldsymbol{\phi}$. In this case we show below that it is not essential to impose identifying restrictions on the factors in FIVU estimation as the value of $\boldsymbol{\phi}$ obtained by unrestricted estimation (over Ω) will coincide with the restricted estimate (over Θ_c) under one further assumption:

ASSUMPTION 5 There exists an open set Θ where $\Omega \supseteq \Theta \supseteq \Theta_r$ with Θ dense in Ω such that for all $\boldsymbol{\theta} = (\boldsymbol{\phi}', \mathbf{g}', \mathbf{f}')' \in \Theta$

$$S\text{vec}(GF') = S\text{vec}(G_r F_r'), \tag{3.7}$$

for some $(\boldsymbol{\phi}', \mathbf{g}_r', \mathbf{f}_r')' \in \Theta_r$. Assume as well that $\boldsymbol{\psi}(\boldsymbol{\theta}_r, M)' C \boldsymbol{\psi}(\boldsymbol{\theta}_r, M)$, $\boldsymbol{\theta}_r \in \Theta_r$, is bounded away from zero outside some given compact set.

Theorem 2. EQUIVALENCE OF UNRESTRICTED AND RESTRICTED ESTIMATION. *Under Assumptions 1-5 $\widehat{\boldsymbol{\phi}}(\Omega) \to \widehat{\boldsymbol{\phi}}(\Theta_c)$ in probability, where Θ_c is defined in the distributional result for FIVU. Define $\boldsymbol{\nu} = (\mathbf{g}', \mathbf{f}')'$ and $\boldsymbol{\nu}_r$ as the subvector of free parameters in $\boldsymbol{\nu}$. If, moreover, at the true values of $\boldsymbol{\nu}$ and $\boldsymbol{\nu}_r$*

$$Span \frac{\partial Svec(GF')}{\partial \boldsymbol{\nu}'} = Span \frac{\partial Svec(G_r F_r')}{\partial \boldsymbol{\nu}_r'}, \tag{3.8}$$

then the covariance matrix of $\widehat{\boldsymbol{\phi}}(\Omega)$ obtained using the generalised inverse of $(\partial\boldsymbol{\psi}/\partial\boldsymbol{\theta}')' C_N \partial\boldsymbol{\psi}/\partial\boldsymbol{\theta}'$ coincides with the covariance matrix of $\widehat{\boldsymbol{\phi}}(\Theta_r)$ inferred from the inverse of $(\partial\boldsymbol{\psi}/\partial\boldsymbol{\theta}_r')' C_N \partial\boldsymbol{\psi}/\partial\boldsymbol{\theta}_r'$.

Proof. See Appendix I. $\square$

[4] It is easy to see that our assumptions imply the assumptions employed by Newey-McFadden, except perhaps for their assumption of dominance, i.e. the norm of the moment function is dominated by a function of $\hat{M}$ of finite expectation. In fact this follows easily in our case from compactness and the existence of second moments.

Theorem 2 shows that the distribution of $\hat{\boldsymbol{\phi}}$ obtained from estimation subject to a set of restrictions on G and F coincides with that obtained from optimisation without imposing these restrictions. If the restrictions constitute a set of identifying restrictions, Proposition 1 tells us that the distribution of $\hat{\boldsymbol{\phi}}$ over the restricted parameter space, and hence in this case the distribution of $\hat{\boldsymbol{\phi}}$ without restrictions, is that given by Newey and McFadden (1994). Essentially, the spanning condition (3.8) ensures that the submatrix of the covariance matrix of $\hat{\boldsymbol{\theta}}$ corresponding to the parameters of interest has not been altered by the restrictions imposed on G and F. In principle, for any proposed model one would need to write down identifying restrictions and check whether $S\text{vec}(GF')$ equals $S\text{vec}(G_rF_r')$ and also whether the spanning condition (3.8) holds, in which case estimation could proceed simply over the unrestricted parameter space.

In the AR(1) one factor example discussed above the free parameters $\boldsymbol{\nu}_r$ consist of $(g_1, \ldots, g_T, f_1, \ldots, f_T)$ with f_T removed. Fixing f_T removes $\partial\boldsymbol{\psi}/\partial f_T$ from $\partial\boldsymbol{\psi}/\partial\boldsymbol{\theta}'$; the spanning condition requires that such deletion does not change the linear space spanned by the columns of $\partial\boldsymbol{\psi}/\partial\boldsymbol{\nu}'$. In Appendix II we demonstrate that Assumptions 1-5 and condition (3.8) are satisfied under the identification scheme proposed for the AR(1) one factor model, so that FIVU can be implemented for this model with unrestricted optimisation.

Estimation for FIVU

The FIVU estimator is straightforward to obtain. Let B be the Choleski matrix of C. Then the objective function has the form

$$Q_B(\boldsymbol{\theta}, \widehat{M}) = \left\| B\boldsymbol{\psi}(\boldsymbol{\theta}, \widehat{M}) \right\|^2 = \left\| [\widehat{M}\boldsymbol{\beta} - S\text{vec}(GF')] \right\|^2. \tag{3.9}$$

When $\boldsymbol{\beta}$ is a linear function of the parameters $\boldsymbol{\phi}$, then, if either G or F is held fixed, the expression $B[\widehat{M}\boldsymbol{\beta}(\boldsymbol{\phi}) - S\text{vec}(GF')]$ is a linear function of the remaining parameters, and the conditional minimum of (3.9) may be found by a one pass least squares procedure. One may then seek a joint minimum by iteration over G and F. This appears to work well in practice. In Appendix III we obtain first and second derivatives for the RHS in (3.9), so Gauss-Newton procedures are also available.

The condition (2.11) takes a particularly simple form when $f_t \equiv 1$ for all t, as in the one way error components model. In this case one has

$$S\text{vec}(GF') = S(\boldsymbol{\iota}_T \otimes I_d)\mathbf{g}. \tag{3.10}$$

Therefore using (3.9), we obtain

$$BM\boldsymbol{\beta} - BS(\boldsymbol{\iota}_T \otimes I_d)\boldsymbol{g} = \mathbf{0}, \tag{3.11}$$

which can be interpreted as a classical regression when M is replaced by its sample counterpart. When $\boldsymbol{\beta}$ is a linear function of $\boldsymbol{\phi}$, FIVU may be obtained by a one pass least squares estimate of (3.11).

Quasi-differencing

An alternative approach to FIVU is obtained by multi-quasi-differencing, which removes the factor component in (2.11). This is achieved by constructing a matrix $D = D(F)$ such that $D(F)S\text{vec}(GF') = \mathbf{0}$. The orthogonality conditions then become

$$D(F)M\boldsymbol{\beta} = \mathbf{0}. \tag{3.12}$$

Quasi-differencing is the method employed by Holtz-Eakin, Newey and Rosen (1988), Ahn, Lee and Schmidt (2001) and Nauges and Thomas (2003) for the one factor case, and Ahn, Lee and Schmidt (2010) for the multi-factor case, as well as Arellano and Bond (1991) (*mutatis mutandis).* In general, this approach eliminates the factor component from the error at the same cost in moment conditions. As shown in Appendix I, such transformations of moment conditions produce estimators of the same asymptotic efficiency as working with the untransformed moment conditions. This result is summarised in the following theorem:

Theorem 3. ASYMPTOTIC EQUIVALENCE RESULT. *Under Assumptions 1-4 FIVU in model (2.1) is asymptotically equivalent to a Generalised Method of Moments estimator based on quasi-differencing and upon constructing $D(F)$.*

Proof. See Appendix I. □

To see this intuitively, consider without loss of generality an AR(1) model with a single factor structure ($n = 1$):

$$y_{it} = \phi y_{it-1} + \lambda_i f_t + \varepsilon_{it}, \qquad t = 1, ..., T. \tag{3.13}$$

FIVU does not rely on any form of differencing and as such it will estimate $1+2T$ parameters (ϕ, T g's and T f's) using $T(T+1)/2$ moment conditions. The quasi-differencing procedure proposed by Holtz-Eakin, Newey and Rosen (1988) and adopted by Nauges and Thomas (2003) transforms the model as

$$\begin{aligned} y_{it} - r_t y_{it-1} &= \phi(y_{it-1} - r_t y_{it-2}) + \lambda_i(f_t - r_t f_{t-1}) + (\varepsilon_{it} - r_t\varepsilon_{it-1}) \\ &= \phi(y_{it-1} - r_t y_{it-2}) + (\varepsilon_{it} - r_t\varepsilon_{it-1}),\ t = 2, \dots T, \end{aligned} \tag{3.14}$$

where $r_t = f_t/f_{t-1}$. This requires estimating $1 + T$ parameters (ϕ and T f's) using $T(T-1)/2$ moment conditions. The procedure proposed by Ahn, Lee and Schmidt (2001, 2010) involves normalising the value of the factor over the last period to unity and transforming the model as follows:

$$\begin{aligned} y_{it} - \widetilde{f}_t y_{iT} &= \phi(y_{it-1} - \widetilde{f}_t y_{iT-1}) + \lambda_i(\widetilde{f}_t - \widetilde{f}_t\widetilde{f}_T) + (\varepsilon_{it} - \widetilde{f}_t\varepsilon_{iT}) \\ &= \phi(y_{it-1} - \widetilde{f}_t y_{iT-1}) + (\varepsilon_{it} - \widetilde{f}_t\varepsilon_{iT}), \end{aligned} \tag{3.15}$$

where $\widetilde{f}_t$ is the normalised value of f_t such that $\widetilde{f}_t{=}f_t\widetilde{f}_T/f_T$ and $\widetilde{f}_T{=}1$. Again, this requires estimating $1 + T$ parameters using $T(T-1)/2$ moment conditions. The net difference between the number of moment conditions and parameters across all these methods is the same. Hence the resulting estimators are asymptoticallly equivalent.

Remark. Computational procedures based on some quasi-differencing transformations could run into problems if some of the factor values are close enough to zero. For example, strictly speaking r_t exists only if $f_t{\neq}0$ for all t. Similarly, the normalisation $\widetilde{f}_t{=}f_t\widetilde{f}_T/f_T$ requires $f_T{\neq}0$. On the other hand, our approach is not restricted by a particular identication scheme. Of course, we do need to assume the model is identified to invoke general GMM results. However, as Theorem 2 makes clear any identification will suffice for this purpose. In fact, if one is only interested in estimating the structural parameters ϕ it is not even necessary to impose identifying restrictions on the factors.

Notice that it is also possible to construct a matrix $D = D(G)$ to eliminate the g terms. To see how this can be achieved, assume a single factor and consider the column vector $S\text{vec}(\mathbf{gf}')$, consisting of scalar terms of the form $g_s f_t$. Consider the following operations on $S\text{vec}(\mathbf{gf}')$:

1. Transform $S\text{vec}(\mathbf{gf}')$ so that all coefficients of terms in the scalar g_1 are unity.
2. Choose one of the g_1 terms and use it to difference away the rest.
3. Eliminate the (single) remaining term in g_1.

One now repeats these operations for the remaining g's. The key point is that all these operations can be accomplished by left multiplication on $S\text{vec}(\mathbf{gf}')$ by matrices of the form $D(G)$. Where there is more than one factor, $\text{vec}(GF')$ consists of sums of terms of the form $\text{vec}(\mathbf{gf}')$. Since the above operations preserve the structure of these terms, the operations may be applied sequentially to the later terms to eliminate them in their turn. Similarly as before, this approach eliminates dn parameters (the g's) at the same cost in moment conditions and so there is no asymptotic efficiency gain/loss over the aforementioned methods.

4 Parameter restrictions: the FIVR estimator

When elements of $\mathbf{x}_{it}$ occur as instruments, model (2.1) implies restrictions on G, the imposition of which will lead to greater efficiency. These restrictions require:

Assumption 6 $E(\boldsymbol{\lambda}_i \varepsilon_{it}) = \mathbf{0}$, for all i and t.

The extra restrictions can be obtained by pre-multiplying (2.1) by $\boldsymbol{\lambda}_i$ and taking expectations, which yields

$$E(\boldsymbol{\lambda}_i \mathbf{x}_{it}')\boldsymbol{\beta} = \Omega_\Lambda \mathbf{f}_t, \quad t = 1, ..., T, \tag{4.1}$$

for N large, where $\Omega_\Lambda = E(\boldsymbol{\lambda}_i \boldsymbol{\lambda}_i')$. The key point is that, when the instrument set includes elements of $\mathbf{x}_{it}$, the entries in $E(\boldsymbol{\lambda}_i \mathbf{x}_{it}')$ include terms in various of the g's so that the LHS of (4.1) is a linear function of the ensemble vector $\mathbf{g}$. Some examples will illustrate.

Example 1. One lagged dependent variable and a single factor The model is

$$y_{it} = \phi y_{it-1} + \lambda_i f_t + \varepsilon_{it}. \tag{4.2}$$

Here $\mathbf{x}_{it} = (y_{it}, y_{it-1})'$, $\boldsymbol{\beta} = (1, -\phi)'$, $z_{it} = y_{it-1}$, $g_s = E(y_{is-1}\lambda_i)$. The linear restrictions (4.1) take the form

$$g_{s+1} = \phi g_s + \sigma^2 f_s, \tag{4.3}$$

where $\sigma^2 = E(\lambda_i^2)$, which can be written in matrix form as

$$\begin{bmatrix} -\phi & 1 & 0 & .. & 0 \\ 0 & -\phi & & : & 0 \\ : & : & : & 1 & : \\ 0 & 0 & .. & -\phi & 1 \end{bmatrix} \begin{bmatrix} g_1 \\ g_2 \\ : \\ g_{T+1} \end{bmatrix} = \sigma^2 f. \tag{4.4}$$

Notice the appearance of the "out-of-sample" term g_{T+1}, which we regard as a constant to be estimated.[5] Section this matrix equation into the form

$$\begin{bmatrix} H & \mathbf{e}_T \end{bmatrix} \begin{bmatrix} \mathbf{g} \\ g_{T+1} \end{bmatrix} = \sigma^2 \mathbf{f}, \tag{4.5}$$

where $\mathbf{e}_T$ is the $T \times 1$ dimensional column vector with 1 in the T^{th} position. The restriction has the form

$$H\mathbf{g} = \sigma^2 \mathbf{f} + \delta \mathbf{e}_T \qquad (\delta \in \mathbb{R}). \tag{4.6}$$

We shall call $H = H(\boldsymbol{\beta})$ the structure matrix; it is specific to the particular model considered.

Example 2. One lagged dependent variable and two factors. In this case $\mathbf{g}_s = E(y_{is-1}\boldsymbol{\lambda}_i')$ is a 1×2 row vector and the restrictions have the form $\mathbf{g}_{s+1}' = \phi \mathbf{g}_s' + \Omega_\Lambda \mathbf{f}_s$. The matrix of restrictions is as in Example 1 except that $\mathbf{g}$ is replaced by $\text{vec}(G')$ and $\boldsymbol{\delta} \in \mathbb{R}^2$. Therefore, we have

$$(H \otimes I_2) P_{T,2} \mathbf{g} = (I_T \otimes \Omega_\Lambda)\mathbf{f} + U\boldsymbol{\delta}, \tag{4.7}$$

where $\mathbf{g}$ is a $2T \times 1$ vector and U is the $2T \times 2$ matrix with columns one and two being $\mathbf{e}_{2T-1}$ and $\mathbf{e}_{2T}$ respectively, and $P_{m,n}$ is the permutation matrix such that $P_{m,n}\text{vec}(A) = \text{vec}(A')$ for $m \times n$ matrices A.

Example 3. One lagged dependent variable, one weakly exogenous variable and one factor. The model is

$$y_{it} = \phi y_{it-1} + \alpha x_{it} + \lambda_i f_t + \varepsilon_{it}. \tag{4.8}$$

In this case the instrument vector is $\mathbf{z}_{it} = (y_{it-1}, x_{it})'$. Note the g's are two-dimensional:

$$\mathbf{g}_s = (\; g_s^1, \quad g_s^2 \;)' = E((\; y_{is-1}\lambda_i, \quad x_{is}\lambda_i \;)'). \tag{4.9}$$

The restrictions can be written $\mathrm{g}_{s+1}^1 = \phi \mathrm{g}_s^1 + \alpha \mathrm{g}_s^2 + \sigma^2 f_s$. In matrix form we have

[5] Strictly speaking, the value of g_{T+1} is *defined* by the restriction it appears in (4.3). We adopt this convention so as to have a neat formula for the full vector f.

$$
\begin{bmatrix} -\phi & -\alpha & 1 & 0 & 0 & .. & 0 \\ 0 & 0 & -\phi & -\alpha & 1 & .. & 0 \\ : & : & : & : & : & : & : \\ 0 & 0 & 0 & .. & -\phi & -\alpha & 1 \end{bmatrix} \begin{bmatrix} g_1^1 \\ g_1^2 \\ : \\ g_T^1 \\ g_T^2 \\ g_{T+1}^1 \end{bmatrix} = \sigma^2 \mathbf{f}, \qquad (4.10)
$$

which can be written more generally as

$$
H\mathbf{g} = \sigma^2 \mathbf{f} + \delta \mathbf{e}_{T,} \qquad \delta \in \mathbb{R}. \qquad (4.11)
$$

where the structure matrix H is now $T \times 2T$.

One can obtain a transformation of (4.11) that is useful when $f_t \equiv 1$ for all t. Since H will in general have a null space of dimension T, (4.11) determines $\mathbf{g}$ only up to T free parameters. Section H into $T \times T$ submatrices so that $H = \begin{bmatrix} H_1 & H_2 \end{bmatrix}$ and section $\mathbf{g}$ conformably as $\mathbf{g} = (\mathbf{g}_1', \boldsymbol{\zeta}')'$. Then the general solution to (4.11) is given by

$$
\mathbf{g}_1 = H_1^{-1}(\mathbf{f} + \delta \mathbf{e}_T - H_2 \boldsymbol{\zeta}), \qquad (4.12)
$$

where $\boldsymbol{\zeta} \in \mathbb{R}^T$ is a free vector of parameters. One can now substitute for $\mathbf{g}$ in (3.11). For a given value of $\boldsymbol{\beta}$, the only unknowns are the parameters δ and $\boldsymbol{\zeta}$, which can be estimated by OLS. The RSS from this regression is the minimand of (3.9). Thus, this procedure effects a concentration $RSS = RSS(\boldsymbol{\beta})$. Finding estimates of the structural parameters is reduced to minimising this function.

Example 4. Two lagged dependent variables and one factor. The model is

$$
y_{it} = \phi_1 y_{it-1} + \phi_2 y_{it-2} + \lambda_i f_t + \varepsilon_{it}. \qquad (4.13)
$$

In this case $\mathbf{w}_i = (y_{i0}, \ldots, y_{iT-1})'$, $z_{it} = y_{it-1}$ and the matrix of restrictions takes the form

$$
\begin{bmatrix} -\phi_2 & -\phi_1 & 1 & 0 & \cdots & 0 \\ 0 & -\phi_2 & -\phi_1 & 1 & \ldots & \vdots \\ 0 & 0 & 0 & \ddots & \ddots & 0 \\ \vdots & \vdots & \vdots & -\phi_2 & -\phi_1 & 1 \end{bmatrix} \begin{bmatrix} g_0 \\ g_1 \\ g_2 \\ : \\ g_T \\ g_{T+1} \end{bmatrix} = \sigma^2 \mathbf{f}. \qquad (4.14)
$$

This is partitioned conformably into

$$\begin{bmatrix} -\phi_2\mathbf{e}_1 & H & \mathbf{e}_T \end{bmatrix} \begin{bmatrix} g_0 \\ \mathbf{g} \\ g_{T+1} \end{bmatrix} = \sigma^2\mathbf{f}, \tag{4.15}$$

with solution

$$H\mathbf{g} = \sigma^2\mathbf{f} + \begin{bmatrix} \mathbf{e}_1 & \mathbf{e}_T \end{bmatrix} \boldsymbol{\delta} \qquad (\boldsymbol{\delta} \in \mathbb{R}^2). \tag{4.16}$$

We turn to the general case. The restrictions take the form

$$H(\boldsymbol{\beta})P_{d,n}\mathbf{g} = (I_T \otimes \Omega_\Lambda)\mathbf{f} + U\boldsymbol{\delta}, \tag{4.17}$$

where $H(\boldsymbol{\beta})$ is an $nT \times nd$ matrix that depends on the structure of the model, as in the above examples, U is a matrix of $\mathbf{e}$ elementary column vectors and $\boldsymbol{\delta}$ is a vector of free parameters corresponding to the "out-of-sample" observations in the above examples. The FIVR estimator (restricted FIV estimator) chooses $\boldsymbol{\theta}$ to minimise (3.9) subject to (4.17). FIVR will in general have fewer parameters to estimate than FIVU and as such it will be more efficient.

The term $H(\boldsymbol{\beta})$ is a linear function of $\boldsymbol{\beta}$ and one has

$$H(\boldsymbol{\beta}) = \sum_{i=1}^{q} K_i\boldsymbol{\beta}_i = K(\boldsymbol{\beta} \otimes I_{nd}), \tag{4.18}$$

where $K = \begin{bmatrix} K_1 & \dots & K_q \end{bmatrix}$. Note that the K_i are given fixed $nT \times nd$ matrices depending on the structure of the model. Then $H(\boldsymbol{\beta})P_{d,n}\mathbf{g} = K(I_q \otimes P_{d,n}\mathbf{g})\boldsymbol{\beta}$ and one can write the restrictions in the form

$$K(I_q \otimes P_{d,n}\mathbf{g})\boldsymbol{\beta} = (I_T \otimes \Omega_\Lambda)\mathbf{f} + U\boldsymbol{\delta}. \tag{4.19}$$

With no restrictions on F (such as $f_t = 1$ for all t, as in the one-way error components model), the model can be reparameterised so as to have $\Omega_\Lambda = I_n$. If there are specific restrictions on some elements of F, (4.19) still holds and can be used to eliminate $\mathbf{f}$ in the objective function at the cost of introducing some parameters in Ω_Λ corresponding to the restricted factors.

Identification and Estimation for FIVR One does not need to develop a separate theory of identification for FIVR; this can be inferred from the FIVU results.

If Assumptions 1-5 hold, and given the equivalence of restricted and unrestricted estimation, then the FIVU estimator may be obtained by minimising the criterion function over the whole parameter space. FIVR minimises the criterion over a closed neighbourhood of $\boldsymbol{\theta}_0$ and this implies straightforwardly that the FIVR estimator likewise has probability limit $\boldsymbol{\theta}_0$. Since FIVR is obtained by expressing some of the nuisance parameters in terms of the remaining parameters, its covariance matrix may be obtained from the FIVU matrix by application of the appropriate Jacobian (calculated in Appendix III). Of course, FIVR will be identified in cases where FIVU is not, since FIVR estimates fewer parameters.

The standard method of solving a minimisation problem subject to an exact constraint is to use the constraint to solve out for some of the choice variables and substitute into the minimand. For the general case we have

$$\mathbf{f} = K(I_q \otimes P_{d,n}\mathbf{g})\boldsymbol{\beta} - U\boldsymbol{\delta}. \tag{4.20}$$

Then one can minimise (3.9) over $(\boldsymbol{\beta}(\boldsymbol{\phi}), \mathbf{g}, \boldsymbol{\delta})$, having substituted for $\mathbf{f}$ from (4.20). In practice one can use a constrained nonlinear optimisation procedure to find the minimum. Formulae for the derivatives are given in Appendix III.

The FIVR estimator effects a more parsimonious parametrisation of the nuisance parameters $\mathbf{g}$, which leads to more efficient GMM estimation of the parameters of interest. Thus FIVR is strictly superior to FIVU and since FIVU is itself asymptotically equivalent to quasi-differencing methods it is superior to these as well. This is summarised in the following theorem:

Theorem 4. Distribution result for FIVR. *Under Assumptions 1-4, 6 and model (2.1) FIVR is asymptotically more efficient than FIVU. Furthermore, it is the efficient estimator in the class of estimators that make use of second moment information.*

Proof. See Appendix I. □

Remark. When $n = 1$ and $f_t = 1$ for $t = 1, ..., T$, the set of linear restrictions (4.3) becomes

$$g_{s+1} = \phi g_s + \sigma^2. \tag{4.21}$$

In this case, FIVR utilises the same set of orthogonality conditions as FIVU, $T(T+1)/2$ in total, but estimates only three parameters, namely ϕ, g_1 and σ^2. Therefore, FIVR makes efficient use of second moment information and intuitively we should expect that it is asymptotically equivalent to the GMM estimator proposed by Ahn and Schmidt (1995). Under stationary initial conditions there is an extra restriction in that $g_1 = \sigma^2/(1-\phi)$. In this case the number of estimable parameters decreases by one and a version of FIVR that uses this extra restriction is asymptotically equivalent to the system GMM estimator proposed by Arellano and Bover (1995) and Blundell and Bond (1998).

5 Finite Sample Performance

In this section we investigate the performance of the GMM estimators proposed in this paper in finite samples. Our focus is on the signal-to-noise ratio of the model, the proportion of the variance of the total error component that is due to the factor component and the degree of persistence in the model.

Design

The data generating process is given by[6]

$$\begin{aligned} y_{it} &= \alpha y_{it-1} + \beta x_{it} + u_{it}; \\ & u_{it} = \boldsymbol{\lambda}_i'\mathbf{f}_t + \varepsilon_{it} = \sum_{j=1}^{n} \lambda_i^j f_t^j + \varepsilon_{it}, \end{aligned} \tag{5.1}$$

and

$$\begin{aligned} x_{it} &= \rho x_{it-1} + \boldsymbol{\gamma}_i'\mathbf{f}_t + \upsilon_{it} = \rho x_{it-1} + \sum_{j=1}^{n} \gamma_i^j f_t^j + \upsilon_{it}; \\ & \upsilon_{it} = \nu_{it} + \pi\varepsilon_{it-1}, \end{aligned} \tag{5.2}$$

where $\varepsilon_{it} \sim i.i.d.N\left(0, c_1\sigma_{\varepsilon_i}^2\right)$, with $\sigma_{\varepsilon_i}^2 \sim i.i.d.U[0,2]$, $\nu_{it} \sim i.i.d.N\left(0,\sigma_\nu^2\right)$, $\lambda_i^j \sim i.i.d.N\left(0, c_2\sigma_{\lambda_i}^2\right)$ with $\sigma_{\lambda_i}^2 \sim i.i.d.U[0,2]$ and $f_t^j \sim i.i.d.N\left(0,1\right)$ for all j, such that $E\left(c_1\sigma_{\varepsilon_i}^2\right) = c_1 > 0$ and $E\left(c_2\sigma_{\lambda_i}^2\right) = c_2 > 0$. Thus, our design allows for substantial cross-sectional heteroskedasticity in the idiosyncratic error and the factor loadings.

The zero mean assumption of the factor variates and the idiosyncratic error component is not restrictive since in practice one can remove the non zero mean for a multi-factor structure by adding individual- and time-specific effects. In particular, one can always reparameterise the error term $u_{it} = \boldsymbol{\lambda}_i'\mathbf{f}_t + \varepsilon_{it} = \eta_i + \tau_t + (\boldsymbol{\lambda}_i - \bar{\boldsymbol{\lambda}})'(\mathbf{f}_t - \bar{\mathbf{f}}) + \varepsilon_{it}$, where $\eta_i = \boldsymbol{\lambda}_i'\bar{\mathbf{f}}$ and $\tau_t = \bar{\boldsymbol{\lambda}}\mathbf{f}_t$. Similarly, adding a global intercept will remove the non zero mean of ε_{it}.

[6] In a previous version of our paper we investigate the performance of our estimators based on a pure AR(1) panel model. That version is available on line at http://mpra.ub.uni-muenchen.de/26166/ .

The factor loadings of the x and y processes are correlated such that

$$\gamma_i^j = \varrho_{\gamma\lambda}\lambda_i^j + \left(1 - \varrho_{\gamma\lambda}^2\right)^{1/2} \varpi_i^j, \ \ \varpi_i^j \sim i.i.d.N\left(0, c_2\sigma_{\lambda_i}^2\right) \ \ \forall j. \tag{5.3}$$

y_{it} can be expressed recursively as follows:

$$\begin{aligned}
y_{it} &= \beta\sum_{s=0}^{\infty}\alpha^s x_{it-s} + \boldsymbol{\lambda}_i'\sum_{s=0}^{\infty}\alpha^s \mathbf{f}_{t-s} + \sum_{s=0}^{\infty}\alpha^s\varepsilon_{it-s} \\
&= \beta\sum_{s=0}^{\infty}\alpha^s(\boldsymbol{\gamma}_i'\sum_{\tau=0}^{\infty}\rho^\tau \mathbf{f}_{t-s-\tau} + \sum_{\tau=0}^{\infty}\rho^\tau \upsilon_{it-s-\tau}) + \boldsymbol{\lambda}_i'\sum_{s=0}^{\infty}\alpha^s \mathbf{f}_{t-s} + \sum_{s=0}^{\infty}\alpha^s\varepsilon_{it-s} \\
&= \beta\boldsymbol{\gamma}_i'\sum_{s=0}^{\infty}\alpha^s\sum_{\tau=0}^{\infty}\rho^\tau \mathbf{f}_{t-s-\tau} + \beta\sum_{s=0}^{\infty}\alpha^s\sum_{\tau=0}^{\infty}\rho^\tau \upsilon_{it-s-\tau} + \boldsymbol{\lambda}_i'\sum_{s=0}^{\infty}\alpha^s \mathbf{f}_{t-s} + \sum_{s=0}^{\infty}\alpha^s\varepsilon_{it-s},
\end{aligned} \tag{5.4}$$

since

$$x_{it} = \boldsymbol{\gamma}_i'\sum_{\tau=0}^{\infty}\rho^\tau \mathbf{f}_{t-\tau} + \sum_{\tau=0}^{\infty}\rho^\tau \upsilon_{it-\tau}. \tag{5.5}$$

As described in Kiviet (1995) and Bun and Kiviet (2006), the variances of ν_{it} and $\boldsymbol{\lambda}_i$ are major determinants of the relative strength of the signal-to-noise ratio and the error components, respectively. Noticing that on average

$$\mathrm{var}(\upsilon_{it}) = \sigma_\upsilon^2 = \sigma_\nu^2 + \pi^2 c_1, \tag{5.6}$$

the average variance of the signal of the model, conditionally on $\boldsymbol{\lambda}_i'\mathbf{f}_t$ and $\boldsymbol{\gamma}_i'\mathbf{f}_t$, is given by

$$\begin{aligned}
\sigma_s^2 &= \mathrm{var}(y_{it}|\boldsymbol{\lambda}_i'\mathbf{f}_t, \boldsymbol{\gamma}_i'\mathbf{f}_t) - \mathrm{var}(\varepsilon_{it}) \\
&= \mathrm{var}(\beta\sum_{s=0}^{\infty}\alpha^s\sum_{\tau=0}^{\infty}\rho^\tau \upsilon_{it-s-\tau}) + \sum_{s=0}^{\infty}\alpha^s\varepsilon_{it-s} \\
&+ 2\mathrm{cov}(\beta\sum_{s=0}^{\infty}\alpha^s\sum_{\tau=0}^{\infty}\rho^\tau \upsilon_{it-s-\tau}, \sum_{s=0}^{\infty}\alpha^s\varepsilon_{it-s}) - \mathrm{var}(\varepsilon_{it}) \\
&= \frac{\beta^2}{(1-\alpha^2)(1-\rho^2)}\sigma_\nu^2 + \frac{\beta^2\pi^2}{(1-\alpha^2)(1-\rho^2)}c_1 + \frac{1}{(1-\alpha^2)}c_1 \\
&+ \frac{2\beta\alpha\pi}{(1-\alpha\rho)(1-\alpha^2)}c_1 - c_1 \\
&= \frac{\beta^2}{(1-\alpha^2)(1-\rho^2)}\sigma_\nu^2 + \frac{\beta^2\pi^2 + (1-\alpha\rho)(1-\rho^2) + 2\beta\alpha\pi(1-\rho^2)}{(1-\alpha^2)(1-\rho^2)(1-\alpha\rho)} - c_1.
\end{aligned} \tag{5.7}$$

The signal-to-noise ratio is defined as

$$SNR \equiv \frac{\sigma_s^2 - c_1}{c_1}. \tag{5.8}$$

We normalise $c_1 = 1$, which implies that SNR depends on the value of σ_ν^2 only, as far as the variance parameters are concerned. Hence, we set σ_ν^2 such that SNR is controlled across experiments. In particular, solving for σ_ν^2 yields

$$\sigma_\nu^2 = \left(SNR + 1 - \frac{\beta^2\pi^2 + (1-\alpha\rho)(1-\rho^2) + 2\beta\alpha\pi(1-\rho^2)}{(1-\alpha^2)(1-\rho^2)(1-\alpha\rho)} \right) \frac{(1-\alpha^2)(1-\rho^2)}{\beta^2}. \tag{5.9}$$

Recalling that $E\left(c_2\sigma_{\lambda_i}^2\right) = c_2$, the value of c_2 is determined according to the average proportion of the variance of the total error, u_{it}, that is due to the factor component, $\boldsymbol{\lambda}_i'\mathbf{f}_t$. It is easy to show that this ratio equals

$$F_\lambda = nc_2(c_2+1)^{-1}.$$

Thus, for example, $F_\lambda = 0.2$ means that 20% of the variance of the total error is due to the unobserved factors; thus, the factor component has small influence in this case. Solving for c_2 yields

$$c_2 = \frac{nF_\lambda}{1-F_\lambda}.$$

We specify $T = 10$, $\varrho_{\gamma\lambda} = 0.5$, $\pi = 0.2$, $N \in \{100, 400\}$, $\rho \in \{0.5, 0.95\}$, $\alpha \in \{0.2, 0.8\}$, $F_\lambda \in \{0.2, 0.8\}$, $SNR \in \{3, 9\}$, $n = 1, 2$, giving rise to 64 different experiments. $\rho = 0.95$ allows us to examine the case where the covariate is close to a unit root process. $\alpha = 0.8$ implies that the y process is highly persistent and receives relatively small influence from x. The SNR values are based on previous literature (e.g. Bun and Kiviet, 2006). 2,000 replications are performed in all cases.

Results

The results are reported in Tables 1-4. We distinguish between one step and two step GMM estimators; $FIVU_j$ ($FIVR_j$) refers to the j step FIVU (FIVR) estimator, $j = 1, 2$. One step estimators make use of the identity matrix as a weighting matrix. Two step estimators make use of the optimal weighting matrix, computed using estimates of the parameters obtained from the first stage. For the one factor case all estimators make use the two most recent available instruments for both y and x. This means there are 35 moment conditions available, while there are 29 parameters for FIVU and 22 for FIVR. For the two factor case, all estimators make use of the four (three) most recent available instruments for y (x). This means there are 56 moment conditions utilised, 54 parameters for FIVU and 40 for FIVR. For FIVU minima are found by an iterative least squares procedure, as described in Section 3; for FIVR we use a constrained nonlinear optimisation algorithm based on Matlab's fmincon function. Convergence is deemed to have occurred when the modulus of the gradient vector is less than 10^{-5}. To obtain initial values for the factors we investigate a grid of values for ρ and for each one we estimate $\mathbf{f}$ using the first n principal components of the resulting residual $\tilde{e}_{it} = x_{it} - \tilde{\rho} x_{it-1}$; we pick $\mathbf{f}$ corresponding to the value of $\tilde{\rho}$ that minimises the criterion function. Notice that identifying restrictions on the factor parameters are not imposed.

For comparison, we also use two versions of the two step GMM estimators proposed by Arellano and Bond (1991), hereafter DIF, and Blundell and Bond (1998), hereafter SYS. Two step refers to the fact that the estimators make use of the optimal weighting matrix in each case. Although these estimators are not consistent under cross-sectional dependence generated by a multi-factor error structure, it is useful to examine their performance under such circumstances given their popularity and the fact that cross-sectional dependence is a highly likely empirical scenario. DIF_a and SYS_a make use of the two most recent available instruments for both y and x with respect to the equations in first differences, while DIF_b and SYS_b make use of the four most recent available instruments with respect to the equations in first differences. The SYS estimators use, in addition, Δy_{it-1} as an instrument for y_{it-1} in the model in levels, $t = 3, ..., T$. Thus, DIF_a, DIF_b, SYS_a and SYS_b utilise 31, 55, 47, and 71 moment conditions respectively, quantities that are well below the size of N.

The results are reported using the following format: average, (standard deviation), [RMSE], {size} of the z-statistic for the structural parameters of the model and |size| of the overidentifying restrictions test statistic. Nominal size is set equal to 5%.[7]

It is clear that FIVU and FIVR perform well under all circumstances. Naturally, their performance improves when the signal-to-noise ratio increases. The same holds as F_λ increases, for $\alpha = .5$, especially when x is highly persistent. Bias for FIVU and FIVR is negligible in all experiments. FIVR has lower standard

[7] For DIF and SYS, since the moment conditions are invalid under a factor structure, the entries in | | reflect power, as opposed to size.

deviation than FIVU and therefore it performs better in terms of RMSE, often by a substantial margin. The difference in the performance of the two estimators with regards to RMSE appears to become larger with higher values of ρ and α when the factor component has a relatively small contribution in the variance of the total error (i.e. $F_\lambda = .2$). For example, for $SNR = 9$ the ratio of the standard deviation of the estimated autoregressive parameter for $FIVR_2$ over the standard deviation of $FIVU_2$ is roughly about 77% when $\alpha = .5$ and $\rho = .5$ and decreases to around 73% for $\alpha = .8$ and $\rho = .95$. Gains in terms of dispersion and RMSE obtained using FIVR appear to be smaller for β compared to α. As expected two step estimators, which are asymptotically efficient, outperform their one step counterparts, especially when x is highly persistent. All estimators perform well in terms of the empirical size of the z-statistic for the structural parameters of the model. The overidentifying restrictions test statistic is valid only for the optimal (two step) GMM estimators and in this case there are only small size distortions.

The performance of DIF and SYS is generally poor and highly sensitive to the design. As expected, bias is smaller when $F_\lambda = 0.2$ relative to $F_\lambda = 0.8$. Even in the former case however, bias can be very large, especially when $\rho = 0.95$ and/or $\alpha = 0.8$. There also appear to be large size distortions for the z-statistic, especially when bias is large, in which case the null hypothesis is rarely not rejected. The power of the overidentifying restrictions test statistic depends crucially on the number of instruments used. For DIF_a power is high, particularly when $F_\lambda = 0.8$. In contrast, the power of SYS_b is close to zero, even in those cases where the degree of misspecification is huge. Practically what this means is that provided the number of moment conditions used is large enough, it is most likely that one fails to reject the validity of the model based on SYS, even if the model is not well specified and the estimator performs poorly.

Similar conclusions apply for the two factor model in that FIVU and FIVR perform well in all experiments. Compared to the one factor case, the dispersion of FIVU increases slightly, while FIVR appears to remain largely unaffected. The performance of the estimators improves for $N = 400$ and, as expected, their standard deviation decreases roughly at the rate of $N^{1/2}$. To save space we do not report these results.

6 Concluding Remarks

The Generalised Method of Moments is a popular approach for estimating dynamic panel data models with large N and T fixed. This approach has the advantage that, compared to maximum likelihood, requires much weaker assumptions about the initial conditions of the data generating process, and avoids full specification of the serial correlation and heteroskedasticity properties of the error, or indeed any other distributional assumptions. On the other hand, under a multi-factor error structure these estimators are inconsistent as the moment conditions they utilise are invalid. In this paper we develop a new GMM type approach for consistent and asymptotically efficient estimation of panel data models with factor residuals. One novelty of our approach is that we do not use quasi-differencing to remove the factor structure - rather, we introduce new parameters to represent the unobserved covariances between the instruments and the factor component of the residual. We develop estimators that are asymptotically efficient and appear to behave well in small samples under a wide range of parametrisations.

References

[1] Ahn, S.C., Schmidt, P., 1995. Efficient Estimations of Models for Dynamic Panel Data. Journal of Econometrics 68, 5-28.

[2] Ahn, S.C., Lee, Y.H., Schmidt, P., 2001. GMM estimation of linear panel data models with time-varying individual effects. Journal of Econometrics 101, 219-255.

[3] Ahn, S.C., Lee, Y.H., Schmidt, P., 2010. Panel Data Models with Multiple Time-Varying Individual Effects. Mimeo, Arizona State University.

[4] Arellano, M., 2003, Panel Data Econometrics. Oxford University Press, Oxford.

[5] Arellano, M., Bond, S.R., 1991. Some Specification tests for Panel Data: Monte Carlo Evidence and an Application to Employment Equations. Review of Economic Studies 58, 277-298.

[6] Arellano, M., Bover, O., 1995. Another Look at the Instrumental Variable Estimation of Error-Component Models. Journal of Econometrics 68, 29-51.

[7] Bai, J., 2003. Inferential Theory for Factor Models of Large Dimensions. Econometrica 71, 135-171.

[8] Bai, J., 2009. Panel Data Models with Interactive Fixed Effects. Econometrica 77, 1229-1279.

[9] Blundell, R., Bond, S.R., 1998. Initial Conditions and Moment Restrictions in Dynamic Panel Data Models. Journal of Econometrics 87, 115-143.

[10] Blundell, R., Bond, S.R., 2000. GMM Estimation with Persistent Data: An application to Production Functions. Econometric Reviews 19, 321-340.

[11] Bun, M., Kiviet, J., 2006. The effects of dynamic feedbacks on LS and MM estimator accuracy in panel data models. Journal of Econometrics 132, 409-444.

[12] Bover, O., Watson, N., 2005. Are there economies of scale in the demand for money by firms? Some panel data estimates. Journal of Monetary Economics 52, 1569-1589.

[13] Chamberlain, G., 1987. Asymptotic Efficiency in Estimation with Conditional Moment Restrictions. Journal of Econometrics 34, 305-334.

[14] Hansen, L.P., 1982. Large Sample Properties of Generalized Method of Moments Estimators. Econometrica 50, 1029-1054.

[15] Hansen, C.B., 2007. Asymptotic Properties of a Robuast Variance Matrix Estimator for Panel Data when T is Large. Journal of Econometrics 141, 597-620.

[16] Holtz-Eakin D., Newey, W., Rosen, H., 1988. Estimating Vector Autoregressions with Panel Data. Econometrica 56, 1371-1395.

[17] Kiviet, J., 1995. On bias, inconsistency and efficiency of various estimators in dynamic panel data models. Journal of Econometrics 68, 53-78.

[18] Nauges, C., Thomas, A., 2003. Consistent estimation of dynamic panel data models with time-varying individual effects. Annales d'Economie et de Statistique 70, 54-75.

[19] Newey, W.K., 1990. Semiparametric Efficiency Bound. Journal of Applied Econometrics 5, 99-136.

[20] Newey, W.K., McFadden, D., 1994. Large Sample Estimation and Hypothesis Testing. Handbook of Econometrics, Vol 4.

[21] Pesaran, M.H., 2006. Estimation and Inference in Large Heterogeneous panels with a Multifactor Error Structure. Econometrica 74, 967-1012.

[22] Presbitero, A.F., 2008. The Debt-Growth Nexus in Poor Countries: A Reassessment. *Economics: The Open-Access, Open-Assessment E-Journal*, Vol. 2, 2008-30. http://dx.doi.org/10.5018/economics-ejournal.ja.2008-30.

[23] Robertson, D., Symons, J., 2007. Estimating Cross-Sectionally Dependent Panels with Weak Exogeneity. Paper presented at SETA, Hong Kong, 2007.

[24] Sarafidis, V., Robertson, D., 2009. On the Impact of Cross-Sectional Dependence in Short Dynamic Panel Estimation. The Econometrics Journal 12, 62-81.

[25] Sarafidis, V., Yamagata, T., 2013. Instrumental Variable Estimation of Dynamic Linear Panel Models with Defactored Regressors under Cross-Sectional Dependence. Mimeo, Monash University.

[26] Sarafidis, V., Yamagata, T., Robertson, D., 2009. A Test of Error Cross Section Dependence for a Linear Dynamic Panel Model with Regressors. Journal of Econometrics 148, 149-161.

[27] Sarafidis, V., Wansbeek, T., 2012. Cross-sectional Dependence in Panel Data Analysis. Econometric Reviews 31, 483-531.

[28] Tregenna, F., 2009. The fat years: the structure and profitability of the US banking sector in the pre-crisis period. Cambridge Journal of Economics 33, 609-632.

[29] Ziliak, J.P., 1997. Efficient Estimation with Panel Data When Instruments Are Predetermined: An Empirical Comparison of Moment-Condition Estimators. Journal of Business and Economics Statistics 15, 419-431.

Appendix I: Proofs of Theorems

Theorem 2

Proof. Assumption 5 guarantees that $\widehat{\boldsymbol{\phi}}(\Theta) = \widehat{\boldsymbol{\phi}}(\Theta_r)$. According to the boundedness assumption, we may choose Θ_c such that the objective function is bounded away from zero outside of this set. Since the minimising value over this set converges to true $\boldsymbol{\theta}$ in probability, it follows that, for N sufficiently large, $\widehat{\boldsymbol{\phi}}(\Theta_c) =$

$\widehat{\phi}(\Theta_r)$ with arbitrarily high probability. The result that $\widehat{\phi}(\Omega) \to \widehat{\phi}(\Theta_c)$ now follows from the density of Θ in Ω.[8] The result for the covariance matrices follows from the following observation. Let X and Y be matrices with the same number of rows. Then the submatrix in the north west corner of the inverse or generalised inverse of $\begin{bmatrix} X & Y \end{bmatrix}' \begin{bmatrix} X & Y \end{bmatrix}$, which is of dimension that of $X'X$, is $(X'M_Y X)^{-1}$, where M_Y is the projection that removes Y, i.e. $M_Y = I - Y(Y'Y)^{-1}Y'$. This follows from the partitioned inverse formula. Thus the covariance matrix of the parameters of interest is obtained by removing from Γ the linear space spanned by the columns corresponding to the nuisance variables; two sets of nuisance variables generating the same span will yield the same covariance matrix. □

Theorem 3

Proof. Assume we have an k-dimensional moment function

$$\boldsymbol{\psi} = \begin{bmatrix} \psi_1(\mathbf{m}, \boldsymbol{\theta}) \\ \vdots \\ \psi_k(\mathbf{m}, \boldsymbol{\theta}) \end{bmatrix}, \tag{6.1}$$

where $\mathbf{m}$ is a collection of moments and $\boldsymbol{\theta}$ is a parameter vector. Consider the usual GMM estimator of the true value based on $\boldsymbol{\psi}$. This has asymptotic variance

$$var(\widehat{\boldsymbol{\theta}}) = (\Gamma'\Delta^{-1}\Gamma)^{-1}, \tag{6.2}$$

where

$$\Gamma = E\left[\frac{\partial \boldsymbol{\psi}}{\partial \boldsymbol{\theta}'}\right]; \qquad \Delta = E(\boldsymbol{\psi}\boldsymbol{\psi}'), \tag{6.3}$$

both evaluated at the true value $\boldsymbol{\theta}_0$. Assume Γ and Δ have full rank and let $\boldsymbol{\theta} = (\boldsymbol{\varphi}', \boldsymbol{\nu}')'$ be a decomposition of the parameter space into two subsets. $\boldsymbol{\varphi}$ is a vector that includes the parameters of interest (and possibly some nuisance parameters) and the vector $\boldsymbol{\nu}$ contains the remaining nuisance parameters. Let

$$Q = E\left[\frac{\partial \boldsymbol{\psi}}{\partial \boldsymbol{\varphi}'}\right]; \qquad R = E\left[\frac{\partial \boldsymbol{\psi}}{\partial \boldsymbol{\nu}'}\right], \tag{6.4}$$

[8] "Dense subset" means that one can find something in the subset arbitrarily close to any element in the superset. For example the set of invertible square matrices is dense in the set of all square matrices, because one can find an invertible matrix arbitrarily close to a given singular matrix. In our context, certain arguments concerning identification will not go through if certain submatrices of F and G are singular. For example in the AR(1), one factor case, we require $g_1 \neq 0$. Density allows us to assume away $g_1 = 0$ and thus obtain identification.

so that $\Gamma = \begin{bmatrix} Q & R \end{bmatrix}$. Since Γ is of full rank, so too are Q and R. Assume that, for some $\ell \times k$ matrix $D(\boldsymbol{\varphi})$ of full rank $\ell \leq k$

$$D(\boldsymbol{\varphi})\boldsymbol{\psi}(\boldsymbol{\varphi}, \boldsymbol{\nu}) = \bar{\boldsymbol{\psi}}(\boldsymbol{\varphi}), \qquad \text{for all } \boldsymbol{\varphi}, \boldsymbol{\nu}, \tag{6.5}$$

i.e. D represents a set of transformations that eliminates the nuisance parameters $\boldsymbol{\nu}$ at the cost of some loss of moment conditions. Then $\bar{\boldsymbol{\psi}}$ is a moment function and inference about $\boldsymbol{\varphi}$ may be based on it. One has the asymptotic variance matrix

$$var(\bar{\boldsymbol{\varphi}}) = (\bar{\Gamma}'\bar{\Delta}^{-1}\bar{\Gamma})^{-1}, \tag{6.6}$$

where $\bar{\Gamma} = E(\partial\bar{\boldsymbol{\psi}}(\mathbf{m}, \boldsymbol{\theta}_0)/\partial\boldsymbol{\varphi}')$ and $\bar{\Delta} = E(\bar{\boldsymbol{\psi}}\bar{\boldsymbol{\psi}}')$. Differentiating (6.5) with respect to $\boldsymbol{\varphi}$ and using the fact that $E(\boldsymbol{\psi}(\mathbf{m}, \boldsymbol{\theta}_0)) = \mathbf{0}$ one has

$$DQ = \bar{\Gamma}. \tag{6.7}$$

Differentiating (6.5) with respect to $\boldsymbol{\nu}$ one has

$$DR = \mathbf{0}, \tag{6.8}$$

where, in both cases, D is evaluated at $\boldsymbol{\theta}_0$. One has as well that

$$\bar{\Delta} = D\Delta D'. \tag{6.9}$$

The asymptotic covariance matrix of $\bar{\boldsymbol{\varphi}}$ is now

$$var(\bar{\boldsymbol{\varphi}}) = [Q'D'(D\Delta D')^{-1}DQ]^{-1}. \tag{6.10}$$

Make the transformations $D_\Delta = D\Delta^{1/2}$, $\Gamma_\Delta = \Delta^{-1/2}\Gamma = \begin{bmatrix} Q_\Delta & R_\Delta \end{bmatrix}$. Then, using results for partitioned inverses, one finds

$$var(\hat{\boldsymbol{\varphi}}) = (Q'_\Delta(I_M - P_{R_\Delta})Q_\Delta)^{-1}, \tag{6.11}$$

where $P_{R_\Delta} = R_\Delta(R'_\Delta R_\Delta)^{-1}R'_\Delta$. One also has

$$var(\bar{\boldsymbol{\varphi}}) = (Q'_\Delta P_{D_\Delta} Q_\Delta)^{-1}, \tag{6.12}$$

where $P_{D_\Delta} = D'_\Delta(D_\Delta D'_\Delta)^{-1}D_\Delta$. Then $var(\bar{\boldsymbol{\varphi}}) > var(\hat{\boldsymbol{\varphi}})$ (as positive matrices) if and only if

$$Q'_{\Delta}(I_M - P_{R_\Delta} - P_{D_\Delta})Q_\Delta > 0. \tag{6.13}$$

Now condition (6.8) implies that the matrices inside the brackets are orthogonal projections so the sandwich matrix is a projection of rank $k - \ell - \dim(R)$. There are thus no losses in efficiency from eliminating the $\boldsymbol{\nu}$ parameters in this way if $\dim(R) = k - \ell$, i.e. the number of eliminated parameters is equal to the number of lost moment conditions. $\square$

Remark. In the case where $f_t \equiv 1$ for all t with linear $\boldsymbol{\beta}$ the moment conditions are linear of the form

$$\mathbf{m} + Q\boldsymbol{\phi} + R\boldsymbol{\xi} = \mathbf{0}, \tag{6.14}$$

where vector $\mathbf{m}$ and matrices Q and R consist of observable moments. The parameters $\boldsymbol{\xi}$ are here the g's from the development in the text. The first-differenced GMM estimator proposed by Arellano and Bond (2001) introduces a differencing matrix of full rank to eliminate R:

$$D\mathbf{m} + DQ\boldsymbol{\phi} = \mathbf{0}. \tag{6.15}$$

Both forms give rise to GMM estimates of the parameters of interest $\boldsymbol{\phi}$ by a one pass regression, given estimates of the error variance-covariance matrix. Let Ω_1 and Ω_2 be such estimates for (6.14) and (6.15) respectively. Call these estimates *compatible* if $\Omega_2 = D\Omega_1 D'$. One might form compatible estimates by first developing an estimate of the covariance matrix for (6.14) and then adjusting it appropriately for (6.15). The following is true:

Proposition. *GMM estimates based on (6.14) and (6.15) are arithmetically equal if they employ compatible estimates of the error variance-covariance matrix.*

To prove this one shows

$$Q'\Omega^{-1/2}(I - P)_{\Omega^{-1/2}R}\Omega^{-1/2}Q = QD'(D\Omega D')^{-1}DQ, \tag{6.16}$$

for any conformable full rank symmetric Ω. This is will be so if $(I - P)_{\Omega^{-1/2}R} = P_{\Omega^{1/2}D}$. It is easy to see that $P_{\Omega^{-1/2}R}P_{\Omega^{1/2}D} = 0$, so that the projections are orthogonal. Consideration of ranks now delivers the result.

In our context, this result shows the first differenced GMM of the error components model is precisely the FIVU estimator, given compatible covariance matrix estimates. In practice, AB estimates and FIVU estimates need not be the same as first step estimates of the structural parameters may differ when the two equations are considered in isolation. In this case, equality is only asymptotic.

Theorem 4.

Proof. Let

$$\boldsymbol{\nu} = \boldsymbol{\nu}(\boldsymbol{\phi}, \boldsymbol{\tau}), \tag{6.17}$$

where $\boldsymbol{\nu}$ is defined above and $\boldsymbol{\tau}$ is a vector of nuisance parameters which has lower dimension than $\boldsymbol{\nu}$. We assume $\boldsymbol{\nu}(.)$ is linear in $\boldsymbol{\tau}$, i.e. $\boldsymbol{\nu}(\boldsymbol{\phi}, \boldsymbol{\tau}) = \mathbf{v}(\boldsymbol{\phi})\boldsymbol{\tau}$, though the argument to be presented would go through under the assumption of sufficient differentiability at the true value. We consider the estimator $\bar{\boldsymbol{\phi}}$ based on the moment conditions in terms of $\boldsymbol{\phi}, \boldsymbol{\tau}$. One has $\overline{\Gamma} = \begin{bmatrix} Q + RJ & RV \end{bmatrix}$ where $J = \partial \boldsymbol{\nu}(\boldsymbol{\phi}, \boldsymbol{\tau})/\partial \boldsymbol{\phi}'$ so, as in (6.11)

$$var(\overline{\boldsymbol{\xi}}) = [(Q + RJ)'_{\Delta}(I_M - P_{(RV)_{\Delta}})(Q + RJ)_{\Delta})]^{-1}. \tag{6.18}$$

Since $(I_M - P_{R_{\Delta}})((Q + RJ)_{\Delta}) = (I_M - P_{R_{\Delta}})Q$ and $P_{R_{\Delta}} > P_{(RV)_{\Delta}}$, one sees from (6.11) that

$$var(\hat{\boldsymbol{\phi}}) \geq var(\bar{\boldsymbol{\phi}}) \tag{6.19}$$

with equality if and only if $(P_{R_{\Delta}} - P_{(RV)_{\Delta}})(Q + RJ)_{\Delta} = 0$. Since in general there is no particular reason for this equality to hold, it follows that a more parsimonious parametrisation of the nuisance parameters will typically deliver a more efficient estimator of the parameters of interest.[9] □

It is also straightforward to prove that FIVR is efficient in the class of estimators that make use of second moment information, based on an argument similar to that provided by Ahn and Schmidt (1995, section 4). Therefore this proof is omitted. In summary, FIVR reaches the semi-parametric efficiency bound discussed by Newey (1990) using standard results of Chamberlain (1987). Thus, FIVR is asymptotically efficient relative to a QML estimator, but the estimators are equally efficient under normality.

[9] The condition will hold if $J = 0$ and $Q'_{\Delta}R_{\Delta} = 0$. This will be so when the reparametrisation can be accomplished independently of $\boldsymbol{\phi}$ and the GMM estimates of the parameters of interest are independent of the estimates of the nuisance parameters.

Appendix II: Identification for FIVU

We focus on the canonical case, where the set of instruments consists of current and lagged values of the variables. Extension to the general case is straightforward. The moment conditions take the form (2.15), i.e. $M\boldsymbol{\beta} - \text{vech}(GF') = \mathbf{0}$. The problem is to impose conditions on vech(GF') so that the values of G and F can be uniquely inferred from knowledge of vech(GF'), at the same time ensuring that the original vech(GF') can be obtained from the restricted G and F. Consider the upper triangular elements of the product GF':

$$\begin{bmatrix} G_1\mathbf{f}_1 & G_1\mathbf{f}_2 & \dots & G_1\mathbf{f}_T \\ & G_2\mathbf{f}_2 & \dots & G_2\mathbf{f}_T \\ & & \ddots & \vdots \\ & & & G_T\mathbf{f}_T \end{bmatrix}. \tag{6.20}$$

One can impose the restriction that the last n columns of F' be I_n. We assume $n \leq (T+1)/2$, so that an $n \times n$ block of terms exists above the main diagonal in (6.20). If this is done, all G_s, for s =1,... $T-n+1$, may be inferred from the values of the terms in (6.20). When $s > T-n+1$ this is no longer so, as such terms as $G_{T-n+2}\mathbf{f}_{T-n+1}$ are not observed. In this case we impose the restrictions that the last $s-T+n-1$ columns of G_s are zero. This enables the unique inference of all the G_s in (6.20) i.e. the full G matrix. Consider now the problem of inferring $\mathbf{f}_t$ when $t \leq T-n$. The matrix

$$\widetilde{G}_t\mathbf{f}_t = \begin{bmatrix} G_1 \\ \vdots \\ G_t \end{bmatrix} \mathbf{f}_t$$

is observed. The number of rows of $\widetilde{G}_t$ is pt. When $pt \geq n$ we impose the restriction that the null space of $\widetilde{G}_t$ be zero, the full rank assumption on $\widetilde{G}_t$. When $pt < n$ (which need not occur), we set the last $n - pt$ entries of $\widetilde{G}_t$ to zero and impose the condition that the appropriately truncated submatrix of $\widetilde{G}_t$ be of full rank. This establishes the identification of G and F. The scheme has the following characteristics:

1. The last n columns of F' form I_n.

2. There are additional zero restrictions on G and F.

3. There is a collection of full rank conditions on submatrices of G.

Let Θ_r be the collection of parameters such that 1-3 hold and Θ be the collection such that both 3 holds and the matrix formed from the last n columns of F' is of full rank. The following facts are straightforward to show:

Properties of the identification scheme.

Assume $n \leq (T+1)/2$.

1. *With $\boldsymbol{\phi}$ held fixed, any $\boldsymbol{\theta} \in \Theta_r$ is identified from the moment conditions.*
2. *For any $\boldsymbol{\theta} \in \Theta$, $\boldsymbol{\psi}(\boldsymbol{\theta}) = \boldsymbol{\psi}(\boldsymbol{\theta}_r)$ for some $\boldsymbol{\theta}_r \in \Theta_r$. Θ is dense in the unrestricted parameter set Ω.*
3. *$E(\partial\boldsymbol{\psi}/\partial\boldsymbol{\nu}_r')$ is of full rank where $\boldsymbol{\nu}_r$ is the vector of free parameters in restricted G, F.*
4. *For any $\boldsymbol{\theta} \in \Theta$, $\boldsymbol{\psi}(\boldsymbol{\theta}) = \boldsymbol{\psi}(\boldsymbol{\theta}_r)$ for some $\boldsymbol{\theta}_r \in \Theta_r$.*
5. *The spanning condition* (3.8) *holds.*

These results establish all of Assumption 5 in the canonical case except the boundedness condition for $\boldsymbol{\theta} \in \Theta_r$. To see this, assume $\boldsymbol{\phi}$ is restricted to a compact set. Then

$$\|B(M\boldsymbol{\beta}(\boldsymbol{\phi}) - \mathrm{vech}(GF')\| \geq \left| \|G\| \left\| B\mathrm{vech}(\bar{G}F') \right\| - \|BM\boldsymbol{\beta}(\boldsymbol{\phi})\| \right|,$$

where $\|G\|$ is the Hilbert-Schmidt norm of G and $\left\|\bar{G}\right\|=1$, where $\bar{G} = G/\left\|\bar{G}\right\|$. The second term can be made arbitrarily large by choice of $\|G\|$ provided $\left\|B\mathrm{vech}(\bar{G}F')\right\|$ can be bounded away from zero. Now $\left\|B\mathrm{vech}(\bar{G}F')\right\| \geq b\left\|\mathrm{vech}(\bar{G}F')\right\|$ where b is the smallest eigenvalue of B.[10] The identification restrictions on G are such that each element of the matrix either appears as a separate term in $\mathrm{vech}(\bar{G}F')$ or is zero. This implies $\left\|\mathrm{vech}(\bar{G}F')\right\| \geq \left\|\bar{G}\right\| = 1$, thus delivering the result.

These conditions suffice to identify the factors; it remains to consider identification for the full vector $\boldsymbol{\theta}$. We shall give a condition for the *one factor* case. We examine when $\Gamma = E(\partial\boldsymbol{\psi}/\partial\boldsymbol{\theta}_r')$ is of full rank, assuming linear $\boldsymbol{\beta}(.)$. Local identification will follow from the full rank of Γ. Write the moment condition (2.14) in terms of upper-triangular matrices

[10] This argument is facilitated by the assumption that B is the symmetric square root of the weight matrix C rather than the Choleski matrix.

$$\begin{bmatrix} \mathrm{M}_{11}\boldsymbol{\beta} & M_{12}\boldsymbol{\beta} & \dots & \mathrm{M}_{1T}\boldsymbol{\beta} \\ & \mathrm{M}_{22}\boldsymbol{\beta} & \dots & \mathrm{M}_{2T}\boldsymbol{\beta} \\ & & \ddots & \vdots \\ & & & \mathrm{M}_{TT}\boldsymbol{\beta} \end{bmatrix} - \begin{bmatrix} \mathbf{g}_1 f_1 & \mathbf{g}_1 f_2 & \dots & \mathbf{g}_1 f_T \\ & \mathbf{g}_2 f_2 & \dots & \mathbf{g}_2 f_T \\ & & \ddots & \vdots \\ & & & \mathbf{g}_T f_T \end{bmatrix} = \mathbf{0}. \tag{6.21}$$

The identification restriction is here that $f_T = 1$ and $\mathbf{g}_T \neq 0$, the latter being the full rank condition on submatrices of G. If this is so, and given that the full rank of $\partial\boldsymbol{\psi}/\partial\boldsymbol{\nu}_r'$ is established, Γ can fail to have full rank only if

$$\mathrm{vech}(M^{\dagger}(I_T \otimes \boldsymbol{\phi}^*)) = \frac{\partial \mathrm{vech}(\mathbf{gf}')}{\partial \mathbf{g}'}\mathbf{g}^* + \frac{\partial \mathrm{vech}(\mathbf{gf}')}{\partial \mathbf{f}'}\mathbf{f}^* \tag{6.22}$$

for some non-zero $(\boldsymbol{\phi}^{*\prime}, \mathbf{g}^{*\prime}, \mathbf{f}^{*\prime})'$, where $M^{\dagger}$ is the $Tp \times (q-1)T$ matrix comprised of the $p \times q$ matrices M_{st} with their first columns removed. In this expression $f_T^* = 0$ since the identification procedure has removed the last column of $\partial\boldsymbol{\psi}/\partial\mathbf{f}'$. Making use of (2.10), this can be written as

$$\mathrm{vech}(M^{\dagger}(I_T \otimes \boldsymbol{\phi}^*)) = \mathrm{vech}(\boldsymbol{g}^*\mathbf{f}') + \mathrm{vech}(\mathbf{gf}^{*\prime}), \tag{6.23}$$

such that the term on the left hand side is $T^2 p \times 1$. One can give a condition under which this relationship cannot hold, and thus Γ calculated for the unrestricted elements of $\boldsymbol{\theta}$ must be of full rank. Assume $T \geq 3$. For the 2×2 submatrix M of terms from the north east of $M^{\dagger}$ one finds

$$M(I_2 \otimes \boldsymbol{\phi}^*) = \mathbf{g}^*\mathbf{f}' + \mathbf{gf}^{*\prime}, \tag{6.24}$$

where the terms on the right now each consist of two elements of the original vectors on the right of (6.23), dated $1, 2$ for both $\mathbf{g}$ vectors and $T-1, T$ for the $\mathbf{f}$ vectors. Exploiting the conditions $f_T = 1$, $f_T^* = 0$, one can show that $(M^{(1)} - f_{T-1}M^{(2)})\boldsymbol{\phi}^* = f_{T-1}^*\mathbf{g}$ where $M^{(1)}$ and $M^{(2)}$ are the first and second blocks of $q-1$ columns of M, respectively. Thus Γ being not of full rank implies that the subvector $\mathbf{g} \in Span(M^{(1)} - f_{T-1}M^{(2)})$ i.e the $2p \times 1$ vector $\mathbf{g}$ is a linear combination of the $q-1$ columns of $M^{(1)} - f_{T-1}M^{(2)}$. Thus:

Identification in the canonical case with one factor *Assume $T \geq 3$. Then Γ has full rank in the case of one factor and linear $\boldsymbol{\beta}(.)$ if* $\mathbf{g}_1 \neq 0$, $f_T = 1$

and

$$\begin{bmatrix} \mathbf{g}_1 \\ \mathbf{g}_2 \end{bmatrix} \notin Span(M^{(1)} - f_{T-1}M^{(2)}) \tag{6.25}$$

at the true values of the parameters.

As a specific example of the canonical case, consider a single lagged dependent variable with this (and its lags) as the instrument and assume $0 < |\phi| < 1$. The model is

$$y_{it} = \phi y_{it-1} + \lambda_i f_t + \varepsilon_{it}. \tag{6.26}$$

If one assumes that the observed data are generated by a process beginning in the distant past, this can be solved as

$$\begin{aligned} \mathrm{y}_{it} &= \lambda_i (I - \phi L)^{-1} f_t + (I - \phi L)^{-1} \varepsilon_{it} && (6.27) \\ &= \lambda_i f_t^c + \eta_{it}, && (6.28) \end{aligned}$$

where the $f_t^c = (I - \phi L)^{-1} f_t$ are redefined factors and η_{it} is a stationary AR(1) (if the ε_{it} are homoskedastic). If we assume λ_i and ε_{it} are independent, it follows that

$$m_{st}^{\dagger} = E(y_{is-1} y_{it}) = \sigma_\lambda^2 f_t^c f_{s-1}^c + \sigma_\eta^2 \phi^{|t-s+1|}, \;\; s = 1, \ldots, t; \; t = 1, \ldots, T. \tag{6.29}$$

One has as well that

$$g_s = E(\lambda_i y_{is-1}) = \sigma_\lambda^2 f_{s-1}^c. \tag{6.30}$$

Using these formulae, one can show Γ has full rank unless

$$\begin{bmatrix} f_0^c \\ f_1^c \end{bmatrix} \propto \begin{bmatrix} \phi \\ 1 \end{bmatrix}. \tag{6.31}$$

If this condition is false the structural parameter of the AR(1) model is identified.

There is a somewhat more complicated version of (6.25) for the multi-factor case. If this condition is satisfied then Assumptions 1-5 can be taken to hold (save for Δ being full rank) and hence the distributional result; since the spanning condition has been demonstrated, the equivalence of restricted and unrestricted estimation may be invoked in the canonical case. One caveat is that the condition (6.25) is not in terms of primitive parameters (i.e. those giving a complete description of the stochastic process generating the data) so it is possible in principle that the condition is in fact vacuous. We have shown this is not the case for the AR(1).

Appendix III: Derivatives

We shall derive the gradient function and the Hessian for a number of FIV models. The notation will be as follows. If $A(\boldsymbol{\theta})$ is a (column) vector-valued function of $\boldsymbol{\theta}$ then $D_{\boldsymbol{\theta}}A(\boldsymbol{\theta}) = \partial A/\partial \boldsymbol{\theta}'$. If A is a matrix then $D_{\boldsymbol{\theta}}A(\boldsymbol{\theta}) = \partial \text{vec}(A)/\partial \boldsymbol{\theta}'$. The chain rule takes the form $D_{\boldsymbol{\theta}}(A(B(\boldsymbol{\theta}))) = D_{\text{vec}B}(A(B))D_{\boldsymbol{\theta}}B$. The product rule is

$$D_{\boldsymbol{\theta}}(A(\boldsymbol{\theta})B(\boldsymbol{\theta})) = (B' \otimes I_m)D_{\boldsymbol{\theta}}A + (I_q \otimes A)D_{\boldsymbol{\theta}}B, \tag{6.32}$$

where A is $m \times p$ and B is $p \times q$. The gradient vector is defined as $\nabla_{\boldsymbol{\theta}}A = (D_{\boldsymbol{\theta}}A)'$.

FIVU gradient vector

In this case the minimand is

$$Q_B = \boldsymbol{\psi}'B'B\boldsymbol{\psi}, \tag{6.33}$$

where

$$\boldsymbol{\psi} = \widehat{M}\boldsymbol{\beta} - Svec(GF'). \tag{6.34}$$

This is optimised with respect to $\boldsymbol{\theta} = (\boldsymbol{\phi}', \mathbf{f}', \mathbf{g}')'$. One has $D_{\boldsymbol{\theta}}Q_B = 2\boldsymbol{\psi}'B'BD_{\boldsymbol{\theta}}\boldsymbol{\psi}$ and, using (2.10)

$$D_{\boldsymbol{\theta}}\boldsymbol{\psi} = \left[\begin{array}{ccc} (\widehat{M}D_{\boldsymbol{\phi}}\boldsymbol{\beta} & -S(I_T \otimes G) & -S(F \otimes I_d) \end{array} \right]. \tag{6.35}$$

The gradient vector is then calculated as

$$\nabla Q_B = 2(D_{\boldsymbol{\theta}}\boldsymbol{\psi})'B'B\boldsymbol{\psi}. \tag{6.36}$$

FIVR gradient vector

As a general principle, the derivatives of the restricted models can be obtained from the FIVU derivatives by use of appropriate Jacobian matrices. Assume the restrictions effect a reparametrisation $\boldsymbol{\theta} = \boldsymbol{\theta}(\boldsymbol{\xi})$ and let $J_{\boldsymbol{\xi}}(\boldsymbol{\theta}) = D_{\boldsymbol{\xi}}\boldsymbol{\theta}$ be the Jacobian. Then

$$(\nabla_R Q_B(\boldsymbol{\xi}))' = \partial Q_B/\partial \boldsymbol{\xi}' = \partial Q_B/\partial \boldsymbol{\theta}' J_{\boldsymbol{\xi}}(\boldsymbol{\theta}) = (\nabla_U Q_B)' J_{\boldsymbol{\xi}}(\boldsymbol{\theta}). \tag{6.37}$$

The FIVR minimisation is in terms of the $\boldsymbol{\xi}$ vector consisting of $\boldsymbol{\phi}, \mathbf{g}, \boldsymbol{\delta}$ where $\mathbf{f} = HP_{d,n}\mathbf{g} - U\boldsymbol{\delta}$. The Jacobian matrix is given by

$$J = \begin{bmatrix} I_r & 0_{r\times nd} & 0_{r\times u} \\ K(I_q \otimes P_{d,n}\mathbf{g})D_{\boldsymbol{\phi}}\boldsymbol{\beta} & H(\boldsymbol{\beta})P_{d,n} & -U \\ 0_{nd\times r} & I_{nd} & 0_{nd\times u} \end{bmatrix}. \tag{6.38}$$

Second derivatives

Write $Q_B = \mathbf{u}'\mathbf{u}$ where $\mathbf{u} = B\boldsymbol{\psi}$. For any parameter vector $\boldsymbol{\theta}$ one has

$$\nabla Q_B = 2\frac{\partial \mathbf{u}'}{\partial \boldsymbol{\theta}}\mathbf{u}, \tag{6.39}$$

so

$$\begin{aligned} D^2_{\boldsymbol{\theta}} Q_B &= D_{\boldsymbol{\theta}}\nabla Q_B & (6.40)\\ &= 2D_{\boldsymbol{\theta}}[\frac{\partial \mathbf{u}'}{\partial \boldsymbol{\theta}}\mathbf{u}] & (6.41)\\ &= 2[(\mathbf{u}' \otimes I_{\dim \boldsymbol{\theta}})D_{\boldsymbol{\theta}}(\frac{\partial \mathbf{u}'}{\partial \boldsymbol{\theta}}) + (D_{\boldsymbol{\theta}}\mathbf{u})'(D_{\boldsymbol{\theta}}\mathbf{u}). & (6.42) \end{aligned}$$

Denote the first term within the brackets $\mathbf{v}(\boldsymbol{\theta})$. One can show that

$$\mathbf{v} = \sum_{i=1}^{\dim \mathbf{u}} u_i D^2_{\boldsymbol{\theta}} u_i, \tag{6.43}$$

where $u_i = B\psi_i$. For both FIVU and FIVR the $\mathbf{u}$ vector is linear in the stochastic term $\widehat{M}\boldsymbol{\beta}$ (when $\boldsymbol{\beta}$ is a linear function of $\boldsymbol{\phi}$) so the second derivatives are non-stochastic functions of $\boldsymbol{\theta}$. Since the $\mathbf{u}$ vector is zero in expectation at the true $\boldsymbol{\theta}_0$ in Method of Moments models we have that, evaluated at $\boldsymbol{\theta}_0$,

$$E(D^2_{\boldsymbol{\theta}} Q_B) = E((D_{\boldsymbol{\theta}}\mathbf{u})'(D_{\boldsymbol{\theta}}\mathbf{u})), \tag{6.44}$$

which suggests that the non-negative matrix $(D_{\boldsymbol{\theta}}\mathbf{u})'(D_{\boldsymbol{\theta}}\mathbf{u})$ may give a good approximation to the Hessian close to convergence.

FIVU second derivatives in the canonical case.

For the FIVU residual vector $\boldsymbol{\psi}$, write $\boldsymbol{\psi}^* = B'B\boldsymbol{\psi}$ and section it into $p\times 1$ submatrices so that $\boldsymbol{\psi}^* = (\boldsymbol{\psi}^{*\prime}_1, ..., \boldsymbol{\psi}^{*\prime}_{T(T+1)/2})'$. Create a $T \times T$ upper semi-triangular matrix V^*, with dimensions $pT \times T$, from these submatrices so that $\text{vech}(V^*) = \boldsymbol{\psi}^{*}$. Then one can show that

$$V(\boldsymbol{\theta}) = \begin{bmatrix} 0_{r\times r} & 0_{r\times nT} & 0_{r\times npT} \\ 0_{nT\times r} & 0_{nT\times nT} & I_n \otimes V^{*\prime} \\ 0_{npT} & I_n \otimes V^{*} & 0_{npT\times npT} \end{bmatrix}. \tag{6.45}$$

The Hessian for FIVU is thus

$$D^2_{\boldsymbol{\theta}} Q_B = V + (D_{\boldsymbol{\theta}}\mathbf{u})'(D_{\boldsymbol{\theta}}\mathbf{u}). \tag{6.46}$$

It is easy to see that the eigenvalues of V are $\pm\sqrt{\mu_i}, i = 1, ..., nT$ (plus zero), where the μ_i are the eigenvalues of $V^{*\prime}V^{*}$. Thus the positivity of the Hessian is not assured in (6.46). In fact, observe that the second term is independent of $\boldsymbol{\phi}$ (see (6.35)), whereas the first term is not. If one imagines a scale increase in $\boldsymbol{\phi}$ then eventually the first term will grow as the square of the expansion factor and the resulting Hessian will have saddlepoints. This shows that an original bad approximation to $\boldsymbol{\phi}$ will lead to problems with algorithms based on the unmodified Hessian.

Concentrations.

For FIVU one has

$$\mathbf{u} = B\boldsymbol{\psi} = B(\widehat{M}\boldsymbol{\beta} - S\mathrm{vec}(GF')). \tag{6.47}$$

By use of (2.10) one has

$$\mathbf{u} = B\begin{bmatrix} \widehat{M} & -S(I_T\otimes G) \end{bmatrix}\begin{bmatrix} \boldsymbol{\beta} \\ \mathbf{f} \end{bmatrix} = B\begin{bmatrix} \widehat{M} & -S(F\otimes I_d) \end{bmatrix}\begin{bmatrix} \boldsymbol{\beta} \\ \mathbf{g} \end{bmatrix}. \tag{6.48}$$

These relationships imply that, given F one can minimise the criterion function by a one pass linear regression, and similarly for G. Iterating these procedures will produce a declining sequence of values of the criterion which usually in practice converges to a local minimum. As a general rule in FIVU estimation we use these concentrations as they are much swifter than line-search methods based on the Hessian. No such concentrations are available for FIVR as, after substituting out for $\mathbf{f}$, the resulting residual vector $\mathbf{u}$ is quadratic in $\mathbf{g}$, so there we are forced to rely on Hessian methods.

Table 1: *Monte Carlo results,* $\rho = .5$

SNR	F_λ	$FIVU_1$	$FIVU_2$	$FIVR_1$	$FIVR_2$	DIF_a	DIF_b	SYS_a	SYS_b
					$\alpha = .5$				
		.497	.499	.499	.499	.481	.476	.491	.490
		(.034)	(.029)	(.023)	(.021)	(.044)	(.040)	(.035)	(.033)
3	.2	[.034]	[.029]	[.023]	[.021]	[.048]	[.047]	[.036]	[.035]
		{.067}	{.060}	{.056}	{.053}	{.146}	{.182}	{.142}	{.191}
		|.382|	|.031|	|.989|	|.039|	|.279|	|.039|	|.102|	|.002|
		.501	.500	.500	.499	.363	.362	.429	.425
		(.032)	(.030)	(.025)	(.021)	(.139)	(.135)	(.119)	(.119)
3	.8	[.032]	[.030]	[.025]	[.021]	[.195]	[.193]	[.139]	[.141]
		{.055}	{.068}	{.059}	{.057}	{.681}	{.775}	{.677}	{.746}
		|.960|	|.064|	|.960|	|.041|	|.982|	|.545|	|.545|	|.037|
		.499	.500	.500	.500	.493	.491	.497	.497
		(.020)	(.017)	(.014)	(.013)	(.024)	(.021)	(.018)	(.017)
9	.2	[.020]	[.017]	[.014]	[.013]	[.025]	[.023]	[.018]	[.017]
		{.059}	{.048}	{.056}	{.055}	{.082}	{.104}	{.088}	{.118}
		|.995|	|.048|	|.997|	|.041|	|.216|	|.039|	|.099|	|.002|
		.500	.500	.500	.500	.416	.414	.465	.461
		(.020)	(.017)	(.012)	(.011)	(.107)	(.101)	(.088)	(.086)
9	.8	[.020]	[.017]	[.012]	[.011]	[.136]	[.133]	[.094]	[.095]
		{.053}	{.058}	{.056}	{.057}	{.643}	{.732}	{.662}	{.720}
		|.998|	|.075|	|1.00|	|.043|	|.964|	|.466|	|.718|	|.040|
					$\beta = .5$				
		.500	.500	.500	.498	.512	.515	.515	.518
3	.2	(.027)	(.025)	(.022)	(.020)	(.031)	(.026)	(.024)	(.021)
		[.027]	[.025]	[.022]	[.021]	[.097]	[.030]	[.028]	[.028]
		{.034}	{.053}	{.041}	{.039}	{.033}	{.109}	{.137}	{.200}
		.499	.499	.501	.499	.647	.674	.661	.673
3	.8	(.026)	(.024)	(.026)	(.021)	(.086)	(.076)	(.066)	(.067)
		[.026]	[.025]	[.026]	[.021]	[.171]	[.187]	[.174]	[.186]
		{.053}	{.067}	{.062}	{.057}	{.717}	{.868}	{.909}	{.957}
		.500	.500	.500	.499	.503	.504	.504	.505
9	.2	(.015)	(.014)	(.012)	(.011)	(.015)	(.013)	(.011)	(.010)
		[.015]	[.014]	[.012]	[.011]	[.015]	[.014]	[.012]	[.012]
		{.032}	{.052}	{.042}	{.060}	{.073}	{.062}	{.090}	{.101}
		.499	.499	.500	.499	.552	.570	567	.576
9	.8	(.014)	(.013)	(.012)	(.011)	(.052)	(.046)	(.037)	(.038)
		[.014]	[.013]	[.012]	[.011]	[.074]	[.083]	[.078]	[.085]
		{.052}	{.057}	{.046}	{.058}	{.487}	{.677}	{.776}	{.867}

$N = 100$; $T = 10$; $n = 1$; 2,000 replications.

Table 2: *Monte Carlo results,* $\rho = .95$

SNR	F_λ	$FIVU_1$	$FIVU_2$	$FIVR_1$	$FIVR_2$	DIF_a	DIF_b	SYS_a	SYS_b
					$\alpha = .5$				
		.491	.495	.496	.497	.430	.412	.437	.443
		(.061)	(.055)	(.052)	(.040)	(.063)	(.059)	(.059)	(.055)
3	.2	[.062]	[.055]	[.052]	[.040]	[.094]	[.106]	[.086]	[.080]
		{.056}	{.051}	{.059}	{.043}	{.282}	{.542}	{.322}	{.434}
		$\vert.249\vert$	$\vert.035\vert$	$\vert.824\vert$	$\vert.046\vert$	$\vert.771\vert$	$\vert.000\vert$	$\vert.895\vert$	$\vert.000\vert$
		.500	.497	.496	.497	.189	.208	.290	.300
		(.059)	(.052)	(.063)	(.046)	(.132)	(.136)	(.152)	(.153)
3	.8	[.059]	[.052]	[.063]	[.046]	[.338]	[.322]	[.322]	[.251]
		{.046}	{.067}	{.034}	{.039}	{.922}	{.930}	{.814}	{.819}
		$\vert.722\vert$	$\vert.064\vert$	$\vert.993\vert$	$\vert.043\vert$	$\vert.991\vert$	$\vert.607\vert$	$\vert.839\vert$	$\vert.023\vert$
		.495	.499	.496	.499	.468	.462	.471	.470
		(.046)	(.038)	(.046)	(.030)	(.044)	(.041)	(.036)	(.036)
9	.2	[.046]	[.038]	[.046]	[.030]	[.054]	[.056]	[.047]	[.047]
		{.055}	{.052}	{.045}	{.040}	{.133}	{.231}	{.184}	{.257}
		$\vert.702\vert$	$\vert.036\vert$	$\vert.985\vert$	$\vert.052\vert$	$\vert.267\vert$	$\vert.042\vert$	$\vert.117\vert$	$\vert.003\vert$
		.500	.498	.498	.499	.256	.262	.329	.335
		(.047)	(.039)	(.049)	(.032)	(.117)	(.120)	(.132)	(.135)
9	.8	[.047]	[.039]	[.049]	[.032]	[.270]	[.267]	[.217]	[.214]
		{.052}	{.068}	{.042}	{.046}	{.893}	{.908}	{.794}	{.803}
		$\vert.929\vert$	$\vert.055\vert$	$\vert.999\vert$	$\vert.042\vert$	$\vert.991\vert$	$\vert.635\vert$	$\vert.847\vert$	$\vert.046\vert$
					$\beta = .5$				
		.508	.503	.505	.501	.734	.684	.575	.572
3	.2	(.056)	(.052)	(.052)	(.042)	(.192)	(.118)	(.053)	(.053)
		[.057]	[.052]	[.053]	[.042]	[.302]	[.218]	[.092]	[.087]
		{.038}	{.053}	{.058}	{.043}	{.355}	{.524}	{.382}	{.534}
		.501	.502	.505	.502	.903	.879	.752	.748
3	.8	(.048)	(.043)	(.057)	(.041)	(.141)	(.129)	(.128)	(.126)
		[.048]	[.043]	[.057]	[.041]	[.427]	[.400]	[.289]	[.278]
		{.050}	{.065}	{.039}	{.044}	{.946}	{.970}	{.867}	{.881}
		.503	.500	.503	.500	.586	.579	.529	.530
9	.2	(.038)	(.032)	(.042)	(.028)	(.147)	(.111)	(.029)	(.028)
		[.038]	[.032]	[.042]	[.028]	[.171]	[.136]	[.041]	[.041]
		{.044}	{.044}	{.040}	{.046}	{.133}	{.193}	{.212}	{.287}
		.500	.500	.502	.500	.859	.841	.684	.683
9	.8	(.037)	(.031)	(.042)	(.027)	(.149)	(.127)	(.097)	(.099)
		[.037]	[.031]	[.042]	[.027]	[.389]	[.364]	[.208]	[.208]
		{.051}	{.064}	{.040}	{.041}	{.901}	{.961}	{.844}	{.866}

$N = 100$; $T = 10$; $n = 1$; 2,000 replications.

Table 3: *Monte Carlo results, $\rho = .5$*

<table>
<tr><th>SNR</th><th>F_λ</th><th>$FIVU_1$</th><th>$FIVU_2$</th><th>$FIVR_1$</th><th>$FIVR_2$</th><th>DIF_a</th><th>DIF_b</th><th>SYS_a</th><th>SYS_b</th></tr>
<tr><td colspan="10">$\alpha = .8$</td></tr>
<tr><td></td><td></td><td>.796</td><td>.797</td><td>.798</td><td>.798</td><td>.482</td><td>.476</td><td>.492</td><td>.491</td></tr>
<tr><td></td><td></td><td>(.044)</td><td>(.033)</td><td>(.023)</td><td>(.022)</td><td>(.042)</td><td>(.039)</td><td>(.033)</td><td>(.032)</td></tr>
<tr><td>3</td><td>.2</td><td>[.044]</td><td>[.033]</td><td>[.023]</td><td>[.022]</td><td>[.321]</td><td>[.325]</td><td>[.310]</td><td>[.032]</td></tr>
<tr><td></td><td></td><td>{.062}</td><td>{.051}</td><td>{.058}</td><td>{.057}</td><td>{1.00}</td><td>{1.00}</td><td>{1.00}</td><td>{1.00}</td></tr>
<tr><td></td><td></td><td>|.826|</td><td>|.035|</td><td>|.988|</td><td>|.039|</td><td>|.272|</td><td>|.039|</td><td>|.103|</td><td>|.003|</td></tr>
<tr><td></td><td></td><td>.799</td><td>.798</td><td>.797</td><td>.798</td><td>.367</td><td>.366</td><td>.432</td><td>.429</td></tr>
<tr><td></td><td></td><td>(.047)</td><td>(.041)</td><td>(.031)</td><td>(.027)</td><td>(.138)</td><td>(.133)</td><td>(.117)</td><td>(.118)</td></tr>
<tr><td>3</td><td>.8</td><td>[.047]</td><td>[.041]</td><td>[.031]</td><td>[.021]</td><td>[.454]</td><td>[.454]</td><td>[.386]</td><td>[.390]</td></tr>
<tr><td></td><td></td><td>{.084}</td><td>{.072}</td><td>{.061}</td><td>{.053}</td><td>{.999}</td><td>{.995}</td><td>{.999}</td><td>{.998}</td></tr>
<tr><td></td><td></td><td>|.961|</td><td>|.059|</td><td>|.998|</td><td>|.036|</td><td>|.981|</td><td>|.537|</td><td>|.781|</td><td>|.039|</td></tr>
<tr><td></td><td></td><td>.799</td><td>.799</td><td>.799</td><td>.799</td><td>.497</td><td>.496</td><td>.499</td><td>.499</td></tr>
<tr><td></td><td></td><td>(.020)</td><td>(.016)</td><td>(.014)</td><td>(.012)</td><td>(.015)</td><td>(.013)</td><td>(.010)</td><td>(.010)</td></tr>
<tr><td>9</td><td>.2</td><td>[.020]</td><td>[.016]</td><td>[.014]</td><td>[.012]</td><td>[.303]</td><td>[.304]</td><td>[.301]</td><td>[.010]</td></tr>
<tr><td></td><td></td><td>{.055}</td><td>{.056}</td><td>{.053}</td><td>{.057}</td><td>{1.000}</td><td>{1.00}</td><td>{1.00}</td><td>{1.00}</td></tr>
<tr><td></td><td></td><td>|.999|</td><td>|.051|</td><td>|.998|</td><td>|.041|</td><td>|.196|</td><td>|.038|</td><td>|.089|</td><td>|.001|</td></tr>
<tr><td></td><td></td><td>.800</td><td>.799</td><td>.800</td><td>.799</td><td>.454</td><td>.451</td><td>.484</td><td>.481</td></tr>
<tr><td></td><td></td><td>(.025)</td><td>(.019)</td><td>(.011)</td><td>(.010)</td><td>(.068)</td><td>(.066)</td><td>(.055)</td><td>(.053)</td></tr>
<tr><td>9</td><td>.8</td><td>[.025]</td><td>[.019]</td><td>[.011]</td><td>[.010]</td><td>[.353]</td><td>[.356]</td><td>[.321]</td><td>[.324]</td></tr>
<tr><td></td><td></td><td>{.076}</td><td>{.038}</td><td>{.061}</td><td>{.046}</td><td>{1.00}</td><td>{1.00}</td><td>{1.00}</td><td>{1.00}</td></tr>
<tr><td></td><td></td><td>|1.00|</td><td>|.082|</td><td>|1.00|</td><td>|.057|</td><td>|.946|</td><td>|.404|</td><td>|.658|</td><td>|.026|</td></tr>
<tr><td colspan="10">$\beta = .2$</td></tr>
<tr><td></td><td></td><td>.199</td><td>.199</td><td>.200</td><td>.199</td><td>.511</td><td>.513</td><td>.513</td><td>.516</td></tr>
<tr><td>3</td><td>.2</td><td>(.024)</td><td>(.022)</td><td>(.019)</td><td>(.018)</td><td>(.028)</td><td>(.025)</td><td>(.022)</td><td>(.020)</td></tr>
<tr><td></td><td></td><td>[.024]</td><td>[.022]</td><td>[.019]</td><td>[.018]</td><td>[.312]</td><td>[.314]</td><td>[.314]</td><td>[.317]</td></tr>
<tr><td></td><td></td><td>{.041}</td><td>{.057}</td><td>{.061}</td><td>{.049}</td><td>{1.00}</td><td>{1.00}</td><td>{1.00}</td><td>{1.00}</td></tr>
<tr><td></td><td></td><td>.200</td><td>.200</td><td>.202</td><td>.200</td><td>.637</td><td>.660</td><td>.651</td><td>.663</td></tr>
<tr><td>3</td><td>.8</td><td>(.022)</td><td>(.021)</td><td>(.023)</td><td>(.019)</td><td>(.083)</td><td>(.074)</td><td>(.064)</td><td>(.064)</td></tr>
<tr><td></td><td></td><td>[.022]</td><td>[.021]</td><td>[.023]</td><td>[.019]</td><td>[.444]</td><td>[.466]</td><td>[.456]</td><td>[.468]</td></tr>
<tr><td></td><td></td><td>{.051}</td><td>{.074}</td><td>{.039}</td><td>{.043}</td><td>{.998}</td><td>{1.00}</td><td>{1.00}</td><td>{1.00}</td></tr>
<tr><td></td><td></td><td>.200</td><td>.200</td><td>.200</td><td>.200</td><td>.501</td><td>.501</td><td>.502</td><td>.502</td></tr>
<tr><td>9</td><td>.2</td><td>(.010)</td><td>(.008)</td><td>(.007)</td><td>(.007)</td><td>(.009)</td><td>(.008)</td><td>(.007)</td><td>(.006)</td></tr>
<tr><td></td><td></td><td>[.010]</td><td>[.008]</td><td>[.007]</td><td>[.007]</td><td>[.301]</td><td>[.302]</td><td>[.302]</td><td>[.302]</td></tr>
<tr><td></td><td></td><td>{.036}</td><td>{.048}</td><td>{.047}</td><td>{.041}</td><td>{1.00}</td><td>{1.00}</td><td>{1.00}</td><td>{1.00}</td></tr>
<tr><td></td><td></td><td>.200</td><td>.200</td><td>.200</td><td>.200</td><td>.522</td><td>.531</td><td>.531</td><td>.536</td></tr>
<tr><td>9</td><td>.8</td><td>(.009)</td><td>(.008)</td><td>(.007)</td><td>(.007)</td><td>(.030)</td><td>(.027)</td><td>(.021)</td><td>(.022)</td></tr>
<tr><td></td><td></td><td>[.009]</td><td>[.008]</td><td>[.006]</td><td>[.007]</td><td>[.323]</td><td>[.332]</td><td>[.331]</td><td>[.336]</td></tr>
<tr><td></td><td></td><td>{.060}</td><td>{.069}</td><td>{.066}</td><td>{.061}</td><td>{1.00}</td><td>{1.00}</td><td>{1.00}</td><td>{1.00}</td></tr>
</table>

$N = 100$; $T = 10$; $n = 1$; 2,000 replications.

Table 4: *Monte Carlo results*, $\rho = .95$

SNR	F_λ	$FIVU_1$	$FIVU_2$	$FIVR_1$	$FIVR_2$	DIF_a	DIF_b	SYS_a	SYS_b
					$\alpha = .8$				
		.793	.794	.796	.795	.404	.399	.417	.427
		(.051)	(.044)	(.037)	(.032)	(.074)	(.067)	(.069)	(.064)
3	.2	[.052]	[.045]	[.037]	[.033]	[.403]	[.407]	[.389]	[.379]
		{.046}	{.055}	{.043}	{.041}	{1.00}	{1.00}	{1.00}	{1.00}
		|.149|	|.038|	|.756|	|.036|	|.374|	|.060|	|.157|	|.001|
		.797	.796	.795	.795	.173	.194	.281	.291
		(.055)	(.044)	(.049)	(.038)	(.137)	(.141)	(.157)	(.157)
3	.8	[.055]	[.045]	[.049]	[.039]	[.642]	[.623]	[.543]	[.533]
		{.071}	{.087}	{.055}	{.046}	{1.00}	{1.00}	{.995}	{.996}
		|.061|	|.057|	|.995|	|.049|	|.990|	|.558|	|.813|	|.023|
		.797	.798	.799	.799	.484	.480	.486	.448
		(.038)	(.030)	(.030)	(.022)	(.032)	(.029)	(.024)	(.024)
9	.2	[.038]	[.030]	[.030]	[.022]	[.318]	[.322]	[.315]	[.317]
		{.051}	{.055}	{.039}	{.047}	{1.00}	{1.00}	{1.00}	{1.00}
		|.907|	|.039|	|.999|	|.045|	|.222|	|.041|	|.099|	|.002|
		.799	.798	.800	.800	.331	.324	.373	.373
		(.046)	(.032)	(.033)	(.025)	(.097)	(.100)	(.109)	(.112)
9	.8	[.046]	[.033]	[.033]	[.025]	[.479]	[.486]	[.441]	[.441]
		{.092}	{.068}	{.051}	{.044}	{1.00}	{1.00}	{.999}	{.999}
		|.986|	|.061|	|1.00|	|.043|	|.975|	|.549|	|.055|	|.055|
					$\beta = .2$				
		.208	.203	.202	.201	.809	.767	.616	.609
3	.2	(.052)	(.051)	(.052)	(.044)	(.202)	(.163)	(.073)	(.066)
		[.052]	[.051]	[.052]	[.044]	[.641]	[.590]	[.423]	[.414]
		{.046}	{.061}	{.041}	{.044}	{.921}	{.963}	{1.00}	{1.00}
		.202	.202	.203	.203	.905	.885	.774	.771
3	.8	(.032)	(.027)	(.034)	(.026)	(.137)	(.129)	(.140)	(.138)
		[.032]	[.027]	[.035]	[.027]	[.719]	[.697]	[.591]	[.588]
		{.065}	{.079}	{.042}	{.045}	{1.00}	{1.00}	{1.00}	{1.00}
		.201	.200	.200	.200	.533	.534	.513	.515
9	.2	(.020)	(.017)	(.019)	(.015)	(.105)	(.075)	(.018)	(.018)
		[.020]	[.017]	[.019]	[.015]	[.349]	[.018]	[.314]	[.316]
		{.051}	{.055}	{.059}	{.051}	{.926}	{.989}	{1.00}	{1.00}
		.201	.201	.201	.200	.795	.785	.627	.630
9	.8	(.022)	(.016)	(.019)	(.014)	(.157)	(.125)	(.074)	(.077)
		[.022]	[.016]	[.019]	[.014]	[.615]	[.598]	[.433]	[.437]
		{.078}	{.070}	{.046}	{.041}	{.994}	{1.00}	{1.00}	{1.00}

$N = 100$; $T = 10$; $n = 1$; 2,000 replications.