IV Estimation of Panels with Factor Residuals

Robertson, Donald; Sarafidis, Vasilis and Symons, James
University of Cambridge, University of Sydney, UCLA

25. October 2010

IV Estimation of Panels with Factor Residuals

Donald Robertson, Vasilis Sarafidis and James Symons*

October 25, 2010

Abstract

This paper considers panel data regression 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. We propose a new estimation approach, based on instrumental variables, which retains the traditional attractive features of method of moments estimators. One 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 typically 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 various parametrizations and is robust to large values of the autoregressive parameter and/or the variance of the factor loadings.

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

JEL Classification: C23, C26.

*University of Cambridge, University of Sydney and University College London (emeritus) respectively. We gratefully acknowledge financial support from the Research Unit of the Faculty of Economics and Business, University of Sydney. An earlier version of the paper was presented at the SETA 2007. We also thank seminar participants at the ESWC 2010, Sydney University and Monash University.

1 Introduction

This paper develops a new approach, 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. There are several ways factor residuals can come about, 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 technical efficiency of a given firm that vary through time. In models of earnings determination, the factor loadings may reflect several different unobserved skills of an individual, with the factors representing the economy-wide price of these skills, which is not necessarily constant over time. Systematic changes in tastes (*weltanshaung)* 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.[2]

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

$$x_{1it} = \phi x_{2it} + \lambda_i f_t + \varepsilon_{it} \quad t = 1, ..., T \; i = 1, ..., N. \tag{1.1}$$

In some cases the values of f_t, or λ_i, are assumed to be known, such as when fitting i-specific, or t-specific intercepts (fixed-effects), or polynomial time-trends, but here we shall treat the f_t as vectors of parameters to be estimated. In this case, one can fit this model by non-linear least-squares, based on principal components analysis;

[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?

[2] An overview of the current literature on panel data models with error cross-sectional dependence is provided by Sarafidis and Wansbeek (2010).

see e.g. Bai (2009). Pesaran (2006) suggests the alternative of augmenting the regression model by the cross-sectional averages of the variables, x_{1it} and x_{2it}, which will span the unobserved factors for large N. Both these methods require that the set of regressors is strongly exogenous with respect to the idiosyncratic error component, ε_{it}, and N, T are both large. In the present paper we focus on the case where N is large, T fixed and the model includes regressors that are not strongly exogenous. This is an empirically relevant scenario in many applied circumstances. 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 has structural significance. 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. Finally, our framework also permits models with endogenous regressors, due to errors of measurement and/or simultaneity, and so it possesses an appealing generality.

When the values of f_t, or λ_i, are known, as in the fixed effects specification, a popular strategy to estimate models with weakly exogenous, or endogenous regressors has been to use 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 while T is relatively small.[3]

[3]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)

However, as shown by Sarafidis and Robertson (2009) and Sarafidis, Yamagata and Robertson (2009), all these procedures fail to provide consistent estimates of the parameters when the errors are generated by a multi-factor structure. Intuitively, this is because the standard moment conditions used are invalidated in this case. 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 from the residuals. More recently, Ahn, Lee and Schmidt (2006) in a seminal paper develop a GMM estimator that allows for multiple factors using multi-quasi-differencing. In this paper we develop an instrumental variables approach that does not involve quasi-differencing and is, in general, more efficient than the existing quasi-differencing-type GMM estimators because it exploits extra restrictions implied by the model.

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

$$E(z_{it}x_{1it}) = \phi E(z_{it}x_{2it}) + g_t f_t, \tag{1.2}$$

where $g_t = E(z_{it}\lambda_i)$. We treat the gs as parameters to be estimated. Replacing the $E(.)$ terms by their sample moments, one has T such orthogonality conditions and $2T+1$ parameters to be estimated (ϕ and the gs and fs): 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 conditions than parameters for $T > 3$, so one can hope for unique estimates. We shall call estimators in this class Factor Instrumental 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 MoM estimators in that they exploit only the orthogonality conditions, which may in fact 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

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.

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 introduced g parameters.

In most practical circumstances the instrument set will include lags of the dependent and independent variables of the model. In this case, a number of linear restrictions can be demonstrated to hold among the parameters (ϕ and the gs and fs) of the model. Greater efficiency can be obtained if these are imposed in estimation. We call this the FIVR (restricted FIV) in contrast to the estimator obtained when these restrictions are not imposed, FIVU (unrestricted FIV).

We remark that our approach is valid under the fixed-effects framework as well. In this case, FIVU is asymptotically equivalent to the GMM estimator proposed by Arellano and Bond (1991) and FIVR is asymptotically equivalent to the GMM estimator proposed by Ahn and Schmidt (1995) for the fixed effects case. Furthermore, within FIVR it is straighforward to impose mean-stationarity on the initial conditions, in which case FIVR is asymptotically equivalent to the system GMM estimator (Arellano and Bover, 1995 and Blundell and Bond, 1998). Thus, FIV estimators offer a unifying treatment of existing dynamic panel estimators.

2 Stochastic Framework

We assume we have a population of vectors Y_i of common dimension which obey

$$x_{it}^T\beta = \lambda_i^T f_t + \varepsilon_{it,} \qquad t = 1, ..., T, \tag{2.1}$$

where variables subscripted i are formed from subvectors of Y_i. The q-vector β is assumed to be a function of r free parameters ϕ:

$$\beta = \beta(\phi).$$

In the work below we shall usually take $\beta = (1, -\phi^T)^T$ where ϕ is an r-vector of parameters We assume an n-factor model i.e. λ_i is a stochastic $n \times 1$ vector (the factor loadings) and f_t is an $n \times 1$ vector of parameters (the factors) at time t; ε_{it} is a purely idiosyncratic disturbance.[4] The sampling structure is that we have N

[4]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 by

sufficiently independent draws, indexed by i, from the population of Y_i. The following assumptions are made:

ASSUMPTION 1. Existence of instruments. We assume potential instruments are given by a vector 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:

$$W_{it} = S_t W_i, \tag{2.2}$$

for which $E(W_{it}\varepsilon_{it}) = 0$. Here S_t is the selector matrix of 0s and 1s that picks out from all potential instruments W_i those that apply 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}. Thus, for example, in the event that the model has a single strongly exogenous independent variable, W_i would consist of all values of this variable, from $t = 1$ to $t = T$ and S_t would be the identity matrix I_T. In the event that the independent variable were only weakly exogenous, then the selector matrix would pick out values dated t and earlier. Mixed cases can occur naturally, such as when the covariates consist of (say) a weakly exogenous variable and an endogenous variable. In this case, W_i is a $2T \times 1$ matrix and the selector matrix will pick out current and lagged values of the weakly exogenous variable, as well as the appropriate dates for the second variable.

PROPOSITION. *Assume* $E(\varepsilon_{it}\lambda_i^T) = 0$. *If* x_{it} *is weakly exogenous and all lags of* x_{it} *belong to the instrument set then* $E(\varepsilon_{is}\varepsilon_{it}) = 0$, $s \neq t$.

The point of the proposition is that in these circumstances the orthogonality of the disturbances is guaranteed by Assumption 1, so that they do not add to moment conditions beyond those implied by this assumption.

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

$$X_i\beta = (I_T \otimes \lambda_i^T)f + \varepsilon_i, \tag{2.3}$$

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 (2006) provide specific details and proofs for both methods. See also Sarafidis and Yamagata (2010) for a discussion.

where $f = vecF^T$, $F^T = [f_1, ..., f_T]$. The corresponding instrument matrix Z_i is defined by

$$Z_i^T = \begin{bmatrix} W_{i1} & 0 & .. & 0 \\ 0 & W_{i2} & & 0 \\ : & & \ddots & \\ 0 & 0 & .. & W_{iT} \end{bmatrix}, \tag{2.4}$$

where one has $E(Z_i^T \varepsilon_i) = 0$, where ε_i denotes the column vector of ε_{it}. Note that Z_i^T is $c \times T$, where $c = \sum_{t=1}^{t=T} c_t$ is the total number of orthogonality conditions. From (2.2) we have

$$Z_i^T = S(I_T \otimes W_i), \tag{2.5}$$

where

$$S = \begin{bmatrix} S_1 & 0 & .. & 0 \\ 0 & S_2 & .. & 0 \\ : & : & : & : \\ 0 & 0 & .. & S_T \end{bmatrix}. \tag{2.6}$$

The matrix S has dimension $c \times Td$. The orthogonality condition for the instruments is now

$$E(Z_i^T X_i \beta - Z_i^T (I_T \otimes \lambda_i^T) f) = 0. \tag{2.7}$$

By use of (2.5) this can be written as

$$M\beta - S(I_T \otimes G)f = 0, \tag{2.8}$$

where $M = E(Z_i^T X_i)$ and $G = E(W_i \lambda_i^T)$. Matrices M and G have dimensions $c \times q$ and $d \times n$ respectively. Some alternative forms of the second term in (2.8) are

$$S(I_T \otimes G)f = Svec(GF^T) = S(F \otimes I_d)g, \tag{2.9}$$

where $g = vecG$. A compact expression of the orthogonality conditions is thus

$$M\beta - Svec(GF^T) = 0. \tag{2.10}$$

When the instruments consist of current and all lagged values: the canonical case As an example, consider when the instrument matrix V_i is naturally pre-

sented as a $T \times p$ matrix of T observations on p variables (so that $W_i = vecV_i$) and ε_{it} is orthogonal to the block of potential instruments from $s = 1$ to $s = t$, i.e. the orthogonality conditions are

$$E(z_{is}\varepsilon_{it}) = 0 \; t = 1, ..., T; s = 1, ..., t, \tag{2.11}$$

where z_{is}^T 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 the independent variables in the model are weakly exogenous, such as the frequently used AR(1) dynamic panel data model with factor residuals. Define $m_{st} = E(z_{is}x_{it}^T)$ and $g_s = E(z_{is}\lambda_i^T)$. The orthogonality conditions are then

$$m_{st}\beta - g_s f_t = 0, \qquad t = 1, ..., T; s = 1, ..., t. \tag{2.12}$$

These conditions can be stacked as

$$\begin{bmatrix} m_{11}\beta \\ m_{12}\beta \\ m_{22}\beta \\ \vdots \\ m_{1T}\beta \\ m_{2T}\beta \\ \vdots \\ m_{TT}\beta \end{bmatrix} - \begin{bmatrix} g_1 f_1 \\ g_1 f_2 \\ g_2 f_2 \\ \vdots \\ g_1 f_T \\ g_2 f_T \\ \vdots \\ g_T f_T \end{bmatrix} = 0. \tag{2.13}$$

More succinctly, this is

$$M\beta - vech(GF^T) = 0, \tag{2.14}$$

where M is the stacked m_{st} terms and the $vech$ operator is understood to act on $p \times 1$ submatrices. Let S_d be the selector matrix of 0s and 1s that turns vec into $vech$ (acting on $d \times d$ matrices). Then

$$M\beta - vech(GF^T) = M\beta - (S_T \otimes I_p)vec(GF^T) = 0, \tag{2.15}$$

which is of the form of (2.10), with the selector matrix S given by $S = S_T \otimes I_p$.

3 The unrestricted estimator FIVU

Define a moment function by

$$\Psi(\theta, Z_i^T X_i) = Z_i^T X_i \beta(\phi) - Svec(GF^T), \tag{3.1}$$

where $\theta = (\phi^T, g^T, f^T)^T$. Then by construction $E(\Psi(\theta)) = 0$ at the true value θ_0. Our aim is to estimate θ_0 by minimising $\Psi(\theta, \widehat{M})^T C \Psi(\theta, \widehat{M})$ where $\widehat{M} = \sum_{i=1}^N Z_i^T X_i / N$ is the matrix of empirical moments and C is a given fixed matrix. As it stands, the model is not identified since

$$M\beta - Svec(GF^T) = M\beta - Svec(GUU^{-1}F^T), \tag{3.2}$$

for any $n \times n$ invertible U. One possible set of restrictions is to require some $n \times n$ submatrix of F^T to be the identity matrix. It turns out that the identity restriction on a submatrix of F is not in general sufficient for full identification; further restrictions are required. In what follows, we provide sufficient conditions for identification of the full parameter vector θ and we establish some primitive conditions for the nuisance parameters, g, and f, as well as the full parameter vector θ in Appendix II.

Let Ω be the full set of possible parameter vectors.

ASSUMPTION 2. We assume that θ_0 belongs to the interior[5] of $\Theta_r \subseteq \Omega$ where Θ_r is obtainable by $0, 1$ restrictions on the G, F components of the vectors in Ω, together with some possible further restrictions excluding a closed set. We assume θ_0 is identified on Θ_r in the sense that $E(\Psi(\theta)) = 0$ for $\theta \in \Theta_r$ implies $\theta = \theta_0$.

Let

$$\Gamma = E\left(\frac{\partial \Psi}{\partial \theta_r^T}(\theta_0)\right), \tag{3.3}$$

and

$$\Delta = E\left(\Psi(\theta_0)\Psi(\theta_0)^T\right), \tag{3.4}$$

where θ_r consists of the free parameters in a restricted θ.

ASSUMPTION 3 We assume both Γ and Δ exist and are of full rank.

[5]The interior is defined in the relative topology induced on Θ_r by the natural topology on Ω.

ASSUMPTION 4 We assume that the elements of Z_i and X_i have finite moments up to order two, and that the function $\beta(.)$ is twice continuously differentiable.

Note that the full rank condition for Γ itself implies that θ_0 is locally identified[6]. 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:

Theorem 1. DISTRIBUTION RESULT FOR FIVU. *Let Θ_c be a compact subset of Θ_r containing θ_0 in its interior and let*

$$\widehat{\theta}(\Theta_c) = \arg\min_{\theta\in\Theta_c} \Psi(\theta, \widehat{M})^T C \Psi(\theta, \widehat{M}), \tag{3.5}$$

where $\widehat{M} = \sum_{i=1}^{N} Z_i^T X_i/N$ and C is a given fixed positive-definite matrix. Then $\widehat{\theta}$ converges in probability to θ_0 and

$$\sqrt{N}(\widehat{\theta} - \theta_0) \xrightarrow{d} N(0, (\Gamma^T C \Gamma)^{-1}(\Gamma^T C \Delta C \Gamma)(\Gamma^T C \Gamma)^{-1}). \tag{3.6}$$

Proof. This is well-known; see e.g. Newey and McFadden (1994) for further details.[7] □

If C is chosen as Δ^{-1} the covariance matrix of the asymptotic distribution is $(\Gamma^T \Delta^{-1} \Gamma)^{-1}$, in which case the estimator has certain optimality properties. These distributional results hold as well if the unobserved Δ is replaced by an estimate based on the Hansen (1982) two-step procedure. We shall call the estimator with the Hansen version of Δ the GMM *unrestricted factor instrumental variables estimator* FIVU (GMM). If instead C is chosen as the identity matrix, so that $\Psi^T\Psi$ is minimised, we call the estimator *minimum-distance* FIVU, denoted FIVU (MD).

Appendix II establishes an identification scheme for FIVU. As a practical matter, if one is interested only in estimates of ϕ, it turns out that it is not essential to impose identifying restrictions on the factors in estimation with FIVU as the value of

[6] This requires the moment function to be twice continuously differentiable, hence Assumption 4.

[7] 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.

ϕ obtained by unrestricted estimation will coincide with the restricted estimate under one further assumption:

ASSUMPTION 5 Assume there exists an open set Θ where $\Omega \supseteq \Theta \supseteq \Theta_r$ where Θ is dense in Ω such that for all $\theta = (\phi^T, g^T, f^T)^T \in \Theta$

$$SvecGF^T = SvecG_rF_r^T \tag{3.7}$$

for some $(\phi^T, g_r^T, f_r^T)^T \in \Theta_r$. Assume as well that $\Psi(\theta_r, M)^T C \Psi(\theta_r, M)$, $\theta_r \in \Theta_r$, is bounded away from zero outside some given compact set.

Theorem 2. EQUIVALENCE OF UNRESTRICTED AND RESTRICTED ESTIMATION. *Under Assumptions 1-5* $\widehat{\phi}(\Omega) \to \widehat{\phi}(\Theta_c)$ *in probability. If, moreover,*

$$Span\frac{\partial SvecGF^T}{\partial \nu^T} = Span\frac{\partial SvecG_rF_r^T}{\partial \nu_r^T}, \tag{3.8}$$

where $\nu = (g^T, f^T)^T$ *and* ν_r *is the subvector of free parameters, then the covariance matrix of* $\widehat{\phi}(\Omega)$ *inferred from the generalised inverse of* $(\partial\Psi/\partial\theta^T)^T C \partial\Psi/\partial\theta^T$ *coincides with the covariance matrix of* $\widehat{\phi}(\Theta_r)$ *inferred from the inverse of* $(\partial\Psi/\partial\theta_r^T)^T C \partial\Psi/\partial\theta_r^T$.

Proof. See Appendix I. □

To see the point of this result, consider a one-factor model with the identification restriction $f_T = 1$, obtainable by re-scaling g and f. It turns out the full-rank condition for Γ requires as well $g_1 \neq 0$. Thus we take $\Theta = \{\theta = (\phi^T, g^T, f^T)^T;\ g_1,\ f_T \neq 0\}$ and $\Theta_r = \{\theta = (\phi^T, g^T, f^T)^T;\ g_1 \neq 0,\ f_T = 1\}$. The free parameters ν_r consist of θ with f_T removed. Fixing f_T removes $\partial\Psi/\partial f_T$ from $\partial\Psi/\partial\theta$; the spanning condition requires that such a deletion does not change the linear space spanned by the columns of $\partial\Psi/\partial\nu^T$.

In Appendix II we demonstrate that Assumptions 1-5 are satisfied under the identification scheme. We show as well that the pre-conditions for the equivalence of restricted and unrestricted estimation hold. We provide sufficient conditions for the identification of the AR(1) model in the multi-factor case.

Estimation for FIVU

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

$$Q_B(\theta, \widehat{M}) = \left\| B\Psi(\theta, \widehat{M}) \right\|^2 = \left\| B[\widehat{M}\beta - Svec(GF^T)] \right\|^2. \tag{3.9}$$

When β is a linear function of the parameters ϕ, then, if either G or F is held fixed, the expression $B[\widehat{M}\beta(\phi) - Svec(GF^T)]$ 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 (2.10) takes a particularly simple form when f_t is the fixed-effects factor, $f_t \equiv 1$ for all t. In this case one has

$$Svec(GF^T) = S(\iota_T \otimes I_d)g. \tag{3.10}$$

Therefore using (3.9), we obtain

$$BM\beta - BS(\iota_T \otimes I_d)g = 0, \tag{3.11}$$

which can be interpreted as a classical regression when M is replaced by its empirical counterpart. When β is a linear function of ϕ, a FIVU estimate 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 from the right of (2.10). This is achieved by constructing a matrix $D = D(F)$ such that $D(F)Svec(GF^T) = 0$. The orthogonality conditions then become

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

To see how this is achieved, assume a single factor and consider the column vector $Svec(gf^T)$, consisting of scalar terms of the form $g_s f_t$. Consider the following operations on $Svec(gf^T)$:

1. Transform 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 $Svec(gf^T)$ by matrices of the form $D(F)$. Where there is more than one factor, $vec(GF^T)$ consists of sums of terms of the form $vec(gf^T)$. 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.

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 (2006) for the multi-factor case, as well as Arellano and Bond (1991) (*mutatis mutandis).* In general, this approach eliminates dn parameters (the gs) 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.*

Proof. See Appendix I. □

Remark. In the case of fixed-effects, simple first-differencing suffices to remove the g terms. A one-pass OLS estimate of β (if a linear function of ϕ) can be obtained from (3.12), just as for FIVU. This is the standard first-differenced GMM estimator proposed by Arellano and Bond (1991). It turns out the GMM versions of the estimators are arithmetically the same provided corresponding estimates of the weighting matrices are employed, i.e. the optimal weighting matrix is obtained from the FIVU version by the $D\,.\,D^T$ transformation. This is discussed more fully in the appendix.

4 Parameter restrictions: the FIVR estimator

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

ASSUMPTION 6 $E(\lambda_i \varepsilon_{it} = 0)$ for all i and t.

To obtain the extra restrictions, multiply (2.1) through by λ_i and take expectations:

$$E(\lambda_i x_{it}^T)\beta = \Omega_\Lambda f_t \quad t = 1, ..., T, \tag{4.1}$$

where $\Omega_\Lambda = E(\lambda_i \lambda_i^T)$. The key point is that, when the instrument set includes elements of the x_{it}, the terms in $E(\lambda_i x_{it}^T)$ include terms in various of the g_s so that the LHS of (4.1) is a linear function of the ensemble vector 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 $x_{it}^T = (y_{it}, y_{it-1})$, $\beta^T = (1, -\phi)$, $z_{it} = y_{it-1}$, $g_s = E(y_{is-1}\lambda_i)$. 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 a matrix 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.[8] Section this matrix equation into the form

[8] 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.

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

where e_T is the T-dimensional column vector with 1 in the T^{th} position. The restriction has the form

$$Hg = \sigma^2 f + \delta e_T \qquad (\delta \in \mathbb{R}). \tag{4.6}$$

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

Example 2. One lagged dependent variable and two factors. In this case $g_s = E(y_{is-1}\lambda_i^T)$ is a 1×2 row vector and the restrictions have the form $g_{s+1}^T = \phi g_s^T + \Omega_\Lambda f_s$. The matrix of restrictions is as in Example 1 except that g is replaced by $vecG^T$ and $\delta \in \mathbb{R}^2$. Therefore, we have

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

where U is the $2T \times 2$ matrix with columns one and two being e_{2T-1} and e_{2T} respectively, and $P_{m,n}$ is the permutation matrix such that $P_{m,n}vecA = vecA^T$ for $m \times n$ matrices Z.

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

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

In this case the instrument vector is $z_{it}^T = (y_{it-1}, r_{it})$. Note the g_s are two-dimensional:

$$g_s^T = \begin{bmatrix} g_s^1 & g_s^2 \end{bmatrix} = E(\begin{bmatrix} y_{is-1}\lambda_i & r_{is}\lambda_i \end{bmatrix}). \tag{4.9}$$

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

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

which can be written more generally as

$$Hg = \sigma^2 f + \delta e_{2T,} \qquad \delta \in \mathbb{R}. \tag{4.11}$$

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

One can obtain a transformation of (4.11) useful when f is known to be fixed-effects. Since H will in general have a null-space of dimension T, (4.11) determines 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 g conformably as $g = [g_1^T, \zeta^T]$. Then the general solution to (4.11) is given by

$$g_1 = H_1^{-1}(f + \delta e_{2T} - H_2\zeta), \tag{4.12}$$

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

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

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

In this case the matrix of restrictions takes the form

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

This is partitioned conformably into

$$\begin{bmatrix} -\phi_2 e_1 & H & e_T \end{bmatrix} \begin{bmatrix} g_0 \\ g \\ g_{T+1} \end{bmatrix} = \sigma^2 f, \tag{4.15}$$

with solution

$$Hg = \sigma^2 f + \begin{bmatrix} e_1 & e_T \end{bmatrix} \delta \qquad (\delta \in \mathbb{R}^2). \tag{4.16}$$

We turn to the general case. Assume there are no restrictions on F such as fixed effects. With F unrestricted, it may be reparametrised in (2.1) so as to have $\Omega_\lambda = I_n$. In general the family of restrictions given by (4.1) takes the form

$$H(\beta) P_{d,n} g = f + U\delta. \tag{4.17}$$

Here $H(\beta)$ is the $nT \times nd$ structure matrix as considered in the above examples, U is a matrix of e elementary column vectors and δ is a vector of free parameters corresponding to the "out-of-sample" observations in the above examples. Both H and U depend on the structure of the model. The FIVR estimator (restricted FIV estimator) chooses θ 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 $H(\beta)$ is a linear function of β and one has

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

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

can write the restrictions in the form

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

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 of parameter space. FIVR minimises the criterion over a closed neighbourhood of θ_0 and this implies straightforwardly that the FIVR estimator likewise has probability limit θ_0. Since FIVR is obtained by a change of variables, 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 a restricted set of parameters. For the AR(1) case there are $(n^2 - n)/2$ redundancies among the factor terms for FIVR. For FIVU in contrast there are $2n^2 - n$ redundancies. Thus for $n = 1$, there are no redundancies among the factor terms for FIVR, but one redundancy for FIVU.

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 f we have

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

Then one can minimise (3.9) over $(\beta(\phi), g, \delta)$, having substituted for f from (4.20). In practice we use a Gauss-Newton 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 g, which should lead to more efficient GMM estimators of the parameters of interest. Thus FIVR is strictly superior to FIVU and since FIVU is itself 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 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). Although not pursued in this paper, this extra restriction is clearly testable.

5 Finite Sample Performance

In this section we investigate the performance of FIVU and FIVR using an AR(1) with one- and two-factor residuals. For comparison, we also include in the experiments the GMM estimators developed by Arellano and Bond (1991) and Ahn and Schmidt (1995), denoted as AB and AS respectively. These estimators are not designed to handle the general factor structure but given their popularity it is of practical interest to see how far they go in resolving the problem. For FIVU minima are found by an iterative OLS procedure, as described in the text; for FIVR we use Gauss-Newton. Initial values for FIVU are specified as $i.i.d.N(0,1)$ for the factor variables and $i.i.d.U(0,1)$ for the AR(1) parameter. Convergence is deemed to have occurred when the modulus of the gradient vector is less than 0.001. We re-initialise starting values when the algorithm is perceived to be travelling slowly across the surface of the objective function; we have found that it is usually better to start afresh than to try to struggle through difficult terrain. Our procedures have occasionally found local minima, especially for FIVU. To tackle this issue we re-initialise the starting conditions 5 times and we pick up the one that minimises the criterion function. For

FIVR we investigate a grid of values for ϕ and for each of these we estimate f using the first n principal components of $x_{it}^T\beta$; we then obtain an initial estimate of g by minimising the criterion function for the value of f and ϕ obtained before. We pick up the value of ϕ that minimises the criterion function. Notice that identifying restrictions on the factor parameters are not imposed.

The factor variates and the idiosyncratic residual, ε_{it}, are all i.i.d. normally distributed with mean zero. This is not restrictive since in practice one can remove the non-zero mean for a n-factor structure by adding individual- and time-specific effects. In particular, one can always reparameterise the error term $u_{it} = \lambda_i^T f_t + \varepsilon_{it} = \eta_i + \tau_t + (\lambda_i - \bar{\lambda})^T(f_t - \bar{f}) + \varepsilon_{it}$, where $\eta_i = \lambda_i^T \bar{f}$ and $\tau_t = \bar{\lambda} f_t$. Similarly, adding a global intercept will remove the non-zero mean of ε_{it}. The variance of both f_t and ε_{it} is standardised to unity. Again, this is not restrictive because $\lambda_i^T(\sigma f_t) = (\lambda_i^T\sigma)f_t$ for any scalar σ and so changing the variance of λ_i has the same effect as changing the variance of f_t. The variance of the factor loadings is determined according to the ratio of the variance of the reduced form of the dependent variable, $y_{it} = \lambda_i^T(1-\phi L)^{-1} f_t + (1-\phi L)^{-1}\varepsilon_{it}$, that is due to factor noise, $\lambda_i^T f_t$, over the total noise. It is easy to show that this ratio equals $F_1 = \sigma_\lambda^2(\sigma_\lambda^2+1)^{-1}$. We report results for $F_1 \in \{0.2, 0.5, 0.8\}$. Thus, for example, $F_1 = 0.2$ means that 20% of the variance of the total error is due to factor noise, and so on. We specify $N = 200$ and $T = 10$ and we choose the autoregressive parameter such that $\phi \in \{0.1, 0.3, 0.5, 0.7, 0.9\}$. Instruments are the lagged dependent variable and its lags. We report the average and the median (from 1000 repetitions) of the parameter on the lagged dependent variable. As a measure of dispersion we report the standard deviation (in brackets beneath the mean, denoted *stdev*) as well as the radius of the interval centred on the median containing precisely 75% of the observations, divided by 1.15 (in brackets beneath the median). This latter statistic, which we shall call the *quasi-standard deviation* (denoted *qstdev*), is an estimate of the population standard deviation if the distribution is normal, with the advantage that it is more robust to the occurence of infrequent outliers. Study of these outliers indicates that they are in large part associated with multiple minima of the moment function, and can be made to disappear for different starting conditions for the minimisation procedure. Table 1 reports some simulation results for FIVU and FIVR.[9]

[9]To save space, results for $\phi = 0.3$ and $\phi = 0.7$ are not reported here. They are available from the authors upon request.

		FIVU MD		FIVU GMM		FIVR MD		FIVR GMM	
ϕ	F_1	$Mean$ $(stdev)$	$Median$ $(qstdev)$	$Mean$ $(stdev)$	$Median$ $(qstdev)$	$Mean$ $(stdev)$	$Median$ $(qstdev)$	$Mean$ $(stdev)$	$Median$ $(qstdev)$
0.1	0.2	.098 (0.040)	.098 (0.034)	.100 (0.043)	.099 (0.043)	.100 (0.028)	.100 (0.026)	.098 (0.031)	.098 (0.031)
	0.5	.098 (0.057)	.098 (0.036)	.100 (0.043)	.098 (0.044)	.099 (0.028)	.100 (0.026)	.101 (0.027)	.102 (0.026)
	0.8	.100 (0.112)	.101 (0.035)	.098 (0.045)	.098 (0.040)	.101 (0.025)	.101 (0.025)	.101 (0.023)	.100 (0.023)
0.5	0.2	.497 (0.042)	.498 (0.032)	.498 (0.043)	.498 (0.040)	.498 (0.027)	.498 (0.025)	.499 (0.027)	.500 (0.024)
	0.5	.497 (0.039)	.497 (0.032)	.499 (0.036)	.499 (0.035)	.499 (0.024)	.499 (0.023)	.501 (0.024)	.501 (0.023)
	0.8	.496 (0.095)	.499 (0.032)	.505 (0.073)	.502 (0.038)	.499 (0.024)	.498 (0.024)	.501 (0.027)	.501 (0.026)
0.9	0.2	.891 (0.055)	.898 (0.019)	.898 (0.027)	.897 (0.018)	.895 (0.024)	.896 (0.015)	.899 (0.017)	.900 (0.014)
	0.5	.899 (0.034)	.900 (0.018)	.898 (0.031)	.900 (0.018)	.900 (0.021)	.899 (0.014)	.900 (0.016)	.901 (0.013)
	0.8	.893 (0.066)	.894 (0.020)	.896 (0.054)	.900 (0.018)	.899 (0.028)	.898 (0.014)	.900 (0.019)	.901 (0.013)

$N = 200$; $T = 10$; $f_t \sim i.i.d.N(0,1)$; $\varepsilon_{it} \sim i.i.d.N(0,1)$; 1,000 replications.

Table 1: *Monte Carlo results for a panel AR(1)*

It is clear that the bias of the estimators is negligible, while their dispersion is small across the whole range of values for ϕ and F_1. As expected, FIVR outperfoms FIVU in terms of bias and RMSE, in all circumstances. FIVU becomes slightly more diffuse as F_1 grows larger while FIVR appears to be robust to different values of F_1. The fact that for FIVU the difference between stdev and qstdev increases with higher values of F_1 is consistent with increased frequency of multiple minima at these values. In contrast, there is very little difference between the stdev and qstdev values for FIVR. The GMM version of the estimators does better, in general, especially for FIVR, although the gains appear to be small. Preliminary results show that the gains in efficiency become more subtantial for higher values of N.[10]

The following pictures illustrate the performance of the estimators FIVU and FIVR relative to AB and AS. It is apparent that AB and AS exhibit large biases, which increase with the value of ϕ, even when the factor component constitutes a small proportion of total noise, i.e. $F_1 = 0.2$. Thus, FIVU and FIVR completely outperform AB and AS in terms of bias and RMSE, although RMSE for AS appears

[10]The results are available upon request.

to be more stable at different values of ϕ. The level of superiority of FIVU and FIVR increases as the fraction of total noise that is due to the factor component rises to 80%. In this case the RMSE of FIVU is at most one third of that for AB and AS while the RMSE of FIVR is at most one fifth of that for AS.

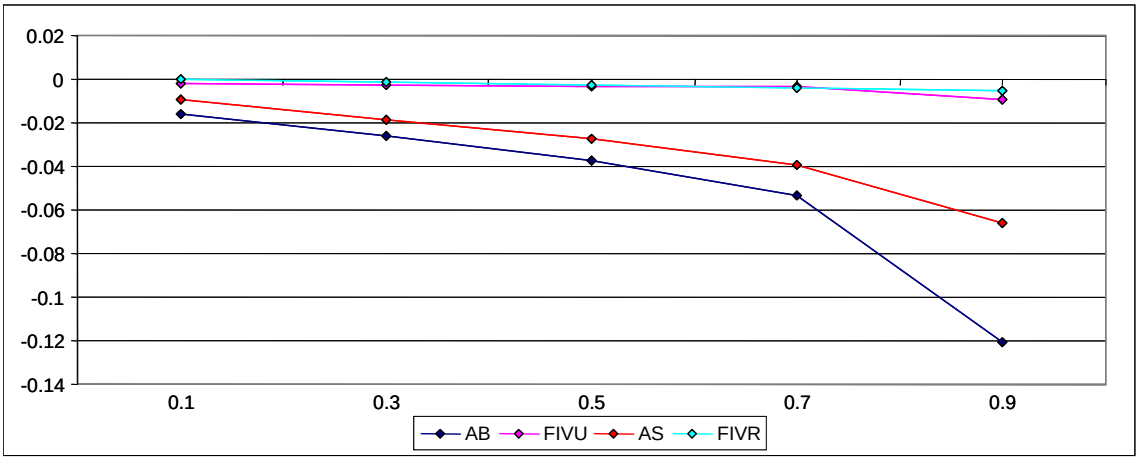

Figure 5.1: Bias, $F_1 = 0.2$

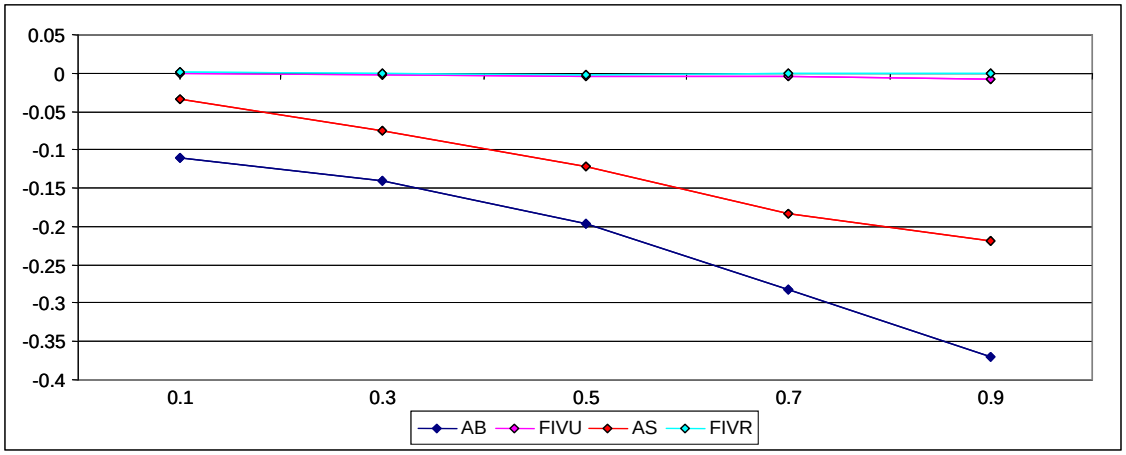

Figure 5.2: Bias, $F_1 = 0.8$

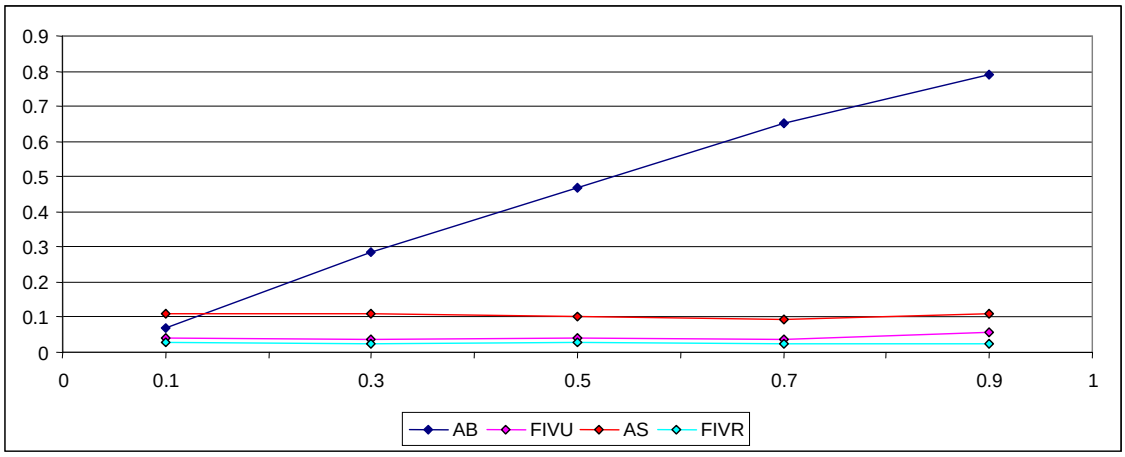

Figure 5.3: RMSE, $F_1 = 0.2$

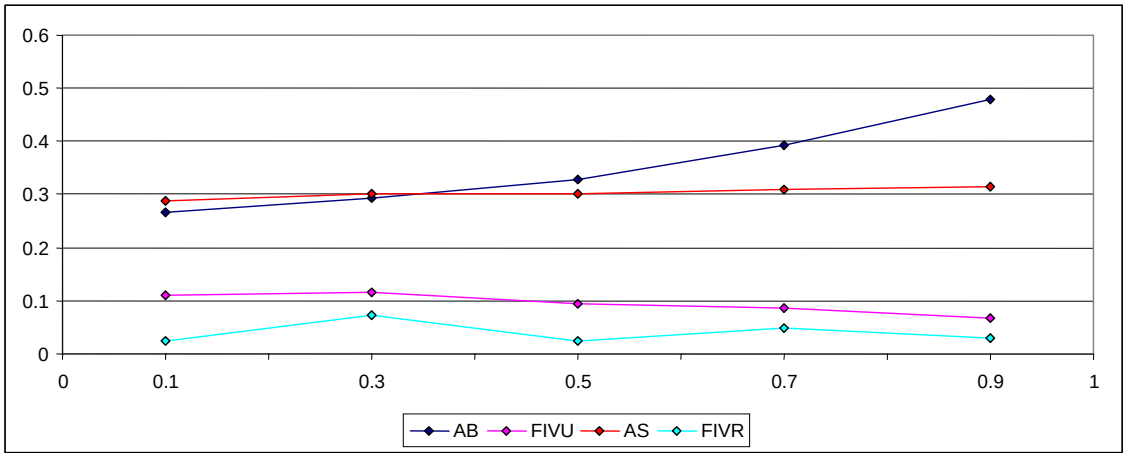

Figure 5.4: RMSE, $F_1 = 0.8$

The following table presents results for a two-factor residual and $F_1 = 0.8$. Similar conclusions are reached for $F_1 = 0.2$ and $F_1 = 0.5$. We can see that the estimators FIVU and FIVR perform well in terms of both bias and RMSE. Compared to the one-factor residual case, the dispersion of FIVU slightly increases, while FIVR appears to do well in all circumstances.

	FIVU MD		FIVU GMM		FIVR MD		FIVR GMM	
ϕ	*Mean* (*stdev*)	*Median* (*qstdev*)	*Mean* (*stdev*)	*Median* (*qstdev*)	*Mean* (*stdev*)	*Median* (*qstdev*)	*Mean* (*stdev*)	*Median* (*qstdev*)
0.1	.098 (0.057)	.098 (0.063)	.097 (0.061)	.098 (0.065)	.102 (0.028)	.101 (0.030)	.100 (0.029)	.100 (0.030)
0.5	.493 (0.070)	.498 (0.050)	.498 (0.071)	.497 (0.049)	.499 (0.029)	.499 (0.027)	.501 (0.024)	.500 (0.022)
0.9	.886 (0.142)	.895 (0.034)	.889 (0.139)	.897 (0.031)	.896 (0.032)	.899 (0.029)	.901 (0.023)	.900 (0.021)

$N = 200$; $T = 10$; $f_t \sim i.i.d.N(0,1)$; $\varepsilon_{it} \sim i.i.d.N(0,1)$; $F_1 = 0.8$; 1,000 replications.

Table 2: *Monte Carlo results for a panel AR(1) with a two-factor residual*

6 Concluding Remarks

The Generalised Method of Moments is a standard 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 cross-sectional dependence these estimators are inconsistent as the moment conditions they utilise are false. In this paper we develop a new GMM-type approach for consistent and asymptotically efficient estimation of panel data models with multi-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 more efficient than the existing quasi-differencing methods and behave well under a wide range of parametrisations, including a large value of the autoregressive parameter.

In a companion paper we apply our method to an autoregressive process with multi-factor residuals and individual fixed effects in order to estimate a model of investment rates for a large panel of firms operating in the US. Using simulated data we show that the proposed estimators perform well, unless the cross-sectional dimension is small.

Appendix I: Proofs of Theorems

Proof of Theorem 2

Assumption 5 guarantees that $\widehat{\phi}(\Theta) = \widehat{\phi}(\Theta_r)$. 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 argmin over this set converges to true θ in probability, it follows that, for N sufficiently large, $\widehat{\phi}(\Theta_c) = \widehat{\phi}(\Theta_r)$ with arbitrarily high probability. The result that $\widehat{\phi}(\Omega) \to \widehat{\phi}(\Theta_c)$ now follows from the density of Θ in Ω.[11] 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 sub-matrix in the NW corner of the inverse or generalised inverse of $\begin{bmatrix} X & Y \end{bmatrix}^T \begin{bmatrix} X & Y \end{bmatrix}$, which is of dimension that of X^TX, is $(X^TM_YX)^{-1}$, where M_Y is the projection that removes Y, i.e. $M_Y = I - Y(Y^TY)^{-1}Y^T$. 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. QED

Proof of Theorem 3

Assume we have an M-dimensional moment function

$$\Psi = \begin{bmatrix} \psi_1(m, \theta) \\ \vdots \\ \psi_M(m, \theta) \end{bmatrix}, \tag{6.1}$$

where m is a collection of moments and θ is a parameter vector. Consider the usual GMM estimator of the true value based on Ψ. This has asymptotic variance

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

[11] "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 sub-matrices 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.

where

$$\Gamma = E\left[\frac{\partial \Psi}{\partial \theta^T}\right] \qquad \Delta = E(\Psi\Psi^T), \tag{6.3}$$

(both evaluated at the true value θ_0. Assume Γ and Δ have full rank and let $\theta = (\phi^T, \nu^T)^T$ be a decomposition of the parameter space into two subsets. The variables ϕ are the parameters of interest and the ν are nuisance parameters. Let

$$Q = \frac{\partial \Psi}{\partial \phi^T} \qquad R = \frac{\partial \Psi}{\partial \nu^T} \tag{6.4}$$

so that $\Gamma = \left[\begin{array}{cc} Q & R \end{array}\right]$. Since Γ is of full rank, so too are Q and R. Assume that, for some $L \times M$ matrix $D(\phi)$ of full rank $L \leq M$

$$D(\phi)\Psi(\phi, \nu) = \bar{\Psi}(\phi) \qquad \text{for all } \phi, \nu, \tag{6.5}$$

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

$$var(\bar{\phi}) = (\bar{\Gamma}^T \bar{\Delta}^{-1} \bar{\Gamma})^{-1}, \tag{6.6}$$

where $\bar{\Gamma} = E(\partial \bar{\Psi}(m, \theta_0)/\partial \phi^T)$ and $\bar{\Delta} = E(\bar{\Psi}\bar{\Psi}^T)$. Differentiating (6.5) with respect to ϕ and using the fact that $E(\Psi(m, \theta_0)) = 0$ one has

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

Differentiating (6.5) with respect to ν one has

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

where, in both cases, D is evaluated at θ_0. One has as well that

$$\bar{\Delta} = D\Delta D^T. \tag{6.9}$$

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

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

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

$$var(\hat{\phi}) = (Q_\Delta^T(I_M - P_{R_\Delta})Q_\Delta)^{-1}, \tag{6.11}$$

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

$$var(\bar{\phi}) = (Q_\Delta^T P_{D_\Delta} Q_\Delta)^{-1}, \tag{6.12}$$

where $P_{D_\Delta} = D_\Delta^T(D_\Delta D_\Delta^T)^{-1}D_\Delta$. Then $var(\bar{\phi}) > var(\hat{\phi})$ (as positive matrices) if and only if

$$Q_\Delta^T(I_M - P_{R_\Delta} - P_{D_\Delta})Q_\Delta > 0. \tag{6.13}$$

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

Remark. In the case of fixed effects with linear β the moment conditions are linear of the form

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

where vector m and matrices Q and R consist of observable moments. The parameters ξ are here the gs from the development in the text. The first-differenced GMM estimator proposed by Arellano and Bond introduces a differencing matrix of full rank to eliminate R:

$$Dm + DQ\phi = 0. \tag{6.15}$$

Both forms give rise to GMM estimates of the parameters of interest ϕ by a one-pass regression, given estimates of the error-covariance-matrices. Let Ω_1 and Ω_2 be such estimates for (6.14) and (6.15) respectively. Call these estimates *compatible* if $\Omega_2 = D\Omega_1 D^T$. 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-covariance matrices.*

To prove this one shows

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

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 fixed effects model is precisely the FIVU estimator, given compatible covariance matrix estimates. In practice, AB estimates and FIVU estimates need not be the same as initial minimum-distance estimates of the structural parameters may differ when the two equations are considered in isolation. In this case, equality is only asymptotic.

Proof of Theorem 4.

Let

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

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

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

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

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

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.[12] QED

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.

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 (2.14), $M\beta - vech(GF^T) = 0$. The problem is to impose conditions on $vechGF^T$so that the values of G and F can be uniquely inferred from knowledge of $vechGF^T$, at the same ensuring that the original $vechGF^T$can be obtained from restricted G and F. Consider the representation of $vechGF^T$ as an upper-triangular matrix:

$$vechGF^T \equiv \begin{bmatrix} g_1f_1 & g_1f_2 & \cdots & g_1f_T \\ & g_2f_2 & \cdots & g_2f_T \\ & & \ddots & \vdots \\ & & & g_Tf_T \end{bmatrix}. \tag{6.20}$$

One can impose the restriction that the last n columns of F^T 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 $g_{T-n+2}f_{T-n+1}$ 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 f_t when

[12]The condition will hold if $J=0$ and $Q_\Delta^T R_\Delta = 0$. This will be so when the reparametrisation can be accomplished independently of ϕ and the GMM estimates of the parameters of interest are independent of the estimates of the nuisance parameters.

$t \leq T - n$. The matrix

$$G_t f_t = \begin{bmatrix} g_1 \\ \vdots \\ g_t \end{bmatrix} f_t$$

is observed. The number of rows of G_t is pt. When $pt \geq n$ we impose the restriction that the null space of G_t be zero, the full-rank assumption on G_t. When $pt < n$ (which need not occur), we set the last $n - pt$ entries of G_t to zero and impose the condition that the appropriately truncated sub-matrix of 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^Tform I_n.
2. There are additional zero restrictions on G and F.
3. There is a collection of full-rank conditions on sub-matrices 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^T is of full rank. The following facts are straightforward to show:

PROPERTIES OF THE IDENTIFICATION SCHEME.

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

1. *With* ϕ *held fixed, any* $\theta \in \Theta_r$ *is identified from the moment conditions.*
2. *For any* $\theta \in \Theta$, $\Psi(\theta) = \Psi(\theta_r)$ *for some* $\theta_r \in \Theta_r$.Θ *is dense in the unrestricted parameter set* Ω.
3. $\partial\Psi/\partial\nu_r$*has full rank where* ν_r *is the vector of free parameters in restricted* G, F.
4. *For any* $\theta \in \Theta$, $\Psi(\theta) = \Psi(\theta_r)$ *for some* $\theta_r \in \Theta_r$.
5. *The spanning condition* (3.8) *holds.*

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

$$\left\| B(M\beta(\phi) - vechGF^T \right\| \geq \left| \|G\| \left\| Bvech\bar{G}F^T \right\| - \|BM\beta(\phi)\| \right|,$$

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 $\left\|Bvech\bar{G}F\right\|$can be bounded away from zero. Now $\left\|Bvech\bar{G}F\right\| \geq b\left\|vech\bar{G}F\right\|$ where b is the smallest eigenvalue of B[13]. The identification restrictions on G are such that each element of the matrix either appears as a separate term in $vech\bar{G}F$ or is zero. This implies $\left\|vech\bar{G}F\right\| \geq \left\|\bar{G}\right\| = 1$, thus delivering the result.

These conditions suffice to identify the factors; it remains to consider identification for the full vector θ. We shall give a condition for the one-factor case. We examine when $\Gamma = \partial\Psi/\partial\theta_r^T$ is of full-rank, assuming linear $\beta(.)$. Local identification will follow from the full-rank of Γ. Write the moment condition (2.13) in terms of upper-triangular matrices

$$\begin{bmatrix} m_{11}\beta & m_{12}\beta & \dots & m_{1T}\beta \\ & m_{22}\beta & \dots & m_{2T}\beta \\ & & \ddots & \vdots \\ & & & m_{TT}\beta \end{bmatrix} - \begin{bmatrix} g_1f_1 & g_1f_2 & \dots & g_1f_T \\ & g_2f_2 & \dots & g_2f_T \\ & & \ddots & \vdots \\ & & & g_Tf_T \end{bmatrix} = 0. \tag{6.21}$$

The identification restriction is here that $f_T = 1$ and $g_T \neq 0$, the latter being the full-rank condition on sub-matrices of G. If this is so, and given that the full rank of $\partial\Psi/\partial\nu_r^T$is established, Γ can fail to have full rank only if

$$vechM^{\dagger}(I_T \otimes \phi^*) = \frac{\partial vechgf^T}{\partial g^T}g^* + \frac{\partial vechgf^T}{\partial f^T}f^* \tag{6.22}$$

for some non-zero (ϕ, g^*, f^*), where $M^{\dagger}$ is the matrix comprised of the m_{st} with their first columns removed. In this expression $f_T^* = 0$ since the identification procedure has removed the last column of $\partial\Psi/\partial f^T$. Making use of (2.9), this can be written

$$vechM^{\dagger}(I_T \otimes \phi^*) = vechg^*f^T + vechgf^{*T}. \tag{6.23}$$

One can give a condition under which this relationship cannot hold, and thus Γ calculated for the unrestricted elements of θ must be of full rank. Assume $T \geq 3$. For

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

the 2×2 sub-matrix m of terms from the North-East of $M^{\dagger}$ one finds

$$m(I_2 \otimes \phi^*) = g^* f^T + g f^{*T}, \tag{6.24}$$

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 g vectors and $T-1, T$ for the f vectors. Exploiting the conditions $f_T = 1$, $f_T^* = 0$, one can show that $(m^{(1)} - f_{T-1} m^{(2)})\phi^* = f_{T-1}^* g$ where $m^{(1)}$ and $m^{(2)}$ are the first and second blocks of $r = q - 1$ columns of m, respectively. Thus Γ being not of full-rank implies that the sub-vector $g \in Span(m^{(1)} - f_{T-1} m^{(2)})$ i.e the $2p \times 1$ vector g is a linear combination of the r 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* $\beta(.)$ *if* $g_1 \neq 0$, $f_T = 1$ *and*

$$\begin{bmatrix} g_1 \\ g_2 \end{bmatrix} \notin Span(m^{(1)} - f_{T-1} m^{(2)}) \tag{6.25}$$

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

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

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

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

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

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

One has as well that

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

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 $A(\theta)$ is a (column) vector-valued function of θ then $D_\theta A(\theta) = \partial A/\partial\theta^T$. If A is a matrix then $D_\theta A(\theta) = \partial vecA/\partial\theta^T$. The chain rule takes the form $D_\theta(A(B(\theta))) = D_{vecB}(A(B))D_\theta B$. The product rule is

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

where A is $m \times p$ and B is $p \times q$. The gradient vector is defined as $\nabla_\theta A = (D_\theta A)^T$.

FIVU gradient vector

In this case the minimand is

$$Q_B = \Psi^T B^T B \Psi, \tag{6.33}$$

where

$$\Psi = \widehat{M}\beta - SvecGF^T. \tag{6.34}$$

This is optimised with respect to $\theta = (\phi^T, f^T g^T)^T$. One has $D_\theta Q_B = 2\Psi^T B^T B D_\theta \Psi$ and, using (2.9)

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

The gradient vector is then calculated as

$$\nabla Q_B = 2(D_\theta \Psi)^T B^T B \Psi. \tag{6.36}$$

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 re-parametrisation $\theta = \theta(\xi)$ and let $J_\xi(\theta) = D_\xi \theta$ be the Jacobian. Then

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

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

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

FIVR when one factor is fixed effects.

It is sometimes of interest to specify that one of the factors (say the first) is fixed effects. If this is done then the re-parametrisation of f so as to have $\Omega_\Gamma = I_p$ can no longer be achieved: the most that can be done is to have $\Omega_\Gamma = \sigma^2 I_p$ for a scale term σ^2. In this case, the optimisation is now with respect to $\phi, \sigma^2, f^0, \delta, \zeta$ where f^0 stands for the unrestricted factor terms.

Second derivatives

Write $Q_B = u^T u$ where $u = B\Psi$. For any parameter vector θ one has

$$\nabla Q_B = 2 \frac{\partial u^T}{\partial \theta} u, \tag{6.39}$$

so

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

Denote the first term within the brackets $V(\theta)$. One can show that

$$V = \sum_{i=1}^{\dim u} u_i D_\theta^2 u_i. \tag{6.43}$$

For both FIVU and FIVR the u vector is linear in the stochastic term $\widehat{M}\beta$ (when β is a linear function of ϕ) so the second derivatives are non-stochastic functions of θ. Since the u vector is zero in expectation at the true θ_0 in MoM models we have that, evaluated at θ_0,

$$E(D_\theta^2 Q_B) = E((D_\theta u)^T (D_\theta u)), \tag{6.44}$$

which suggests that the non-negative matrix $(D_\theta u)^T (D_\theta u)$ may give a good approximation to the Hessian close to convergence.

FIVU second derivatives in the canonical case.

For the FIVU residual vector Ψ, write $\Psi^* = B^T B \Psi$ and section it into $p \times 1$ sub-matrices so that $\Psi^* = (\Psi_1^{*T}, ..., \Psi_{T(T+1)/2}^{*T})^T$. Create a $T \times T$ upper semi-triangular matrix V^* from these sub-matrices so that $vech V^* = \Psi^{*\cdot}$ (Note that V^* is a $pT \times T$ matrix of scalars.) Then one can show

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

The Hessian for FIVU is thus

$$D_\theta^2 Q_B = V + (D_\theta u)^T (D_\theta u). \tag{6.46}$$

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 $V^{*T} V^*$. Thus the positivity of the Hessian is not assured

in (6.46). In fact, observe that the second term is independent of ϕ (see (6.35)), whereas the first term is not. If one imagines a scale increase in ϕ 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 ϕ will lead to problems with algorithms based on the unmodified Hessian.

Concentrations.

For FIVU one has

$$u = B\Psi = B(\widehat{M}\beta - SvecGF^T). \tag{6.47}$$

By use of (2.9) one has

$$u = B \left[\begin{array}{cc} \widehat{M} & -S(I_T \otimes G) \end{array} \right] \left[\begin{array}{c} \beta \\ f \end{array} \right] = B \left[\begin{array}{cc} \widehat{M} & -S(F \otimes I_d) \end{array} \right] \left[\begin{array}{c} \beta \\ g \end{array} \right]. \tag{6.48}$$

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 f, the resulting residual vector u is quadratic in g, so there we are forced to rely on Hessian methods.

References

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

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

[3] Ahn, S. C., Y. H. Lee, and P. Schmidt (2006): "Panel Data Models with Multiple Time-Varying Individual Effects," Mimeo, Arizona State University.

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

[5] Arellano, M., and O. Bover (1995): "Another Look at the Instrumental Variable Estimation of Error-Component Models," *Journal of Econometrics*, 68, 29-51.

[6] Bai, J. (2003): "Inferential Theory for Factor Models of Large Dimensions," *Econometrica*, 71, 135-171.

[7] Bai, J. (2009): "Panel Data Models with Interactive Fixed Effects," *Econometrica*, 77, 1229-1279.

[8] Blundell, R., and S. Bond (1998): "Initial Conditions and Moment Restrictions in Dynamic Panel Data Models," *Journal of Econometrics*, 87, 115-143.

[9] Blundell, R., and S. Bond (2000): "GMM Estimation with Persistent Data: An application to Production Functions," *Econometric Reviews*, 19, 321-340.

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

[11] Chamberlain, G. (1987): "Asymptotic Efficiency in Estimation with Conditional Moment Restrictions," *Journal of Econometrics*, 34, 305-334.

[12] Hansen, L. P. (1982): "Large Sample Properties of Generalized Method of Moments Estimators," *Econometrica*, 50, 1029-1054.

[13] Holtz-Eakin D, W. Newey H. and Rosen (1988): "Estimating Vector Autoregressions with Panel Data," *Econometrica*, 56, 1371-1395.

[14] Nauges, C., and A. Thomas (2003): "Consistent estimation of dynamic panel data models with time-varying individual effects," *Annales d'Économie et de Statistique*, 70, p. 54-75.

[15] Newey, W. K. (1990): "Semiparametric Efficiency Bound," *Journal of Applied Econometrics*, 5, 99-136.

[16] Newey, W. K., and D. McFadden (1994): "Large Sample Estimation and Hypothesis Testing," *Handbook of Econometrics*, Vol 4.

[17] Pesaran, M. H. (2006): "Estimation and Inference in Large Heterogeneous panels with a Multifactor Error Structure," *Econometrica.*

[18] 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. doi:10.5018/economics-ejournal.ja.2008-30, http://dx.doi.org/10.5018/economics-ejournal.ja.2008-30.

[19] Robertson, D., and J. Symons (2s007): "Estimating Cross-Sectionally Dependent Panels with Weak Exogeneity," Paper presented at SETA, Hong Kong, 2007.

[20] Sarafidis, V., and D. Robertson (2009): "On the Impact of Cross-sectional Dependence in Short Dynamic Panel Estimation," *The Econometrics Journal*, 12(1), 62-81.

[21] Sarafidis, V. and T. Yamagata (2010): "Instrumental Variable Estimation of Dynamic Linear Panel Models with Defactored Regressors under Cross-sectional Dependence," Mimeo, The University of Sydney.

[22] Sarafidis, V., T. Yamagata, and D. Robertson (2009): "A Test of Error Cross Section Dependence for a Linear Dynamic Panel Model with Regressors," *Journal of Econometrics*, 148(2), 149-161.

[23] Sarafidis, V. and T. Wansbeek (2010): "Cross-sectional Dependence in Panel Data Analysis," Mimeo, The University of Sydney.

[24] 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.

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