The Stata Journal

The *Stata Journal* publishes reviewed papers together with shorter notes or comments, regular columns, book reviews, and other material of interest to Stata users. Examples of the types of papers include 1) expository papers that link the use of Stata commands or programs to associated principles, such as those that will serve as tutorials for users first encountering a new field of statistics or a major new technique; 2) papers that go "beyond the Stata manual" in explaining key features or uses of Stata that are of interest to intermediate or advanced users of Stata; 3) papers that discuss new commands or Stata programs of interest either to a wide spectrum of users (e.g., in data management or graphics) or to some large segment of Stata users (e.g., in survey statistics, survival analysis, panel analysis, or limited dependent variable modeling); 4) papers analyzing the statistical properties of new or existing estimators and tests in Stata; 5) papers that could be of interest or usefulness to researchers, especially in fields that are of practical importance but are not often included in texts or other journals, such as the use of Stata in managing datasets, especially large datasets, with advice from hard-won experience; and 6) papers of interest to those who teach, including Stata with topics such as extended examples of techniques and interpretation of results, simulations of statistical concepts, and overviews of subject areas.

The *Stata Journal* is indexed and abstracted by *CompuMath Citation Index*, *Current Contents/Social and Behavioral Sciences*, *RePEc: Research Papers in Economics*, *Science Citation Index Expanded* (also known as *SciSearch*), *Scopus*, and *Social Sciences Citation Index*.

For more information on the *Stata Journal*, including information for authors, see the webpage

http://www.stata-journal.com

Subscriptions are available from StataCorp, 4905 Lakeway Drive, College Station, Texas 77845, telephone 979-696-4600 or 800-STATA-PC, fax 979-696-4601, or online at

http://www.stata.com/bookstore/sj.html

Subscription rates listed below include both a printed and an electronic copy unless otherwise mentioned.

U.S. and Canada		Elsewhere	
Printed & electronic		**Printed & electronic**	
1-year subscription	$115	1-year subscription	$145
2-year subscription	$210	2-year subscription	$270
3-year subscription	$285	3-year subscription	$375
1-year student subscription	$ 85	1-year student subscription	$115
1-year institutional subscription	$345	1-year institutional subscription	$375
2-year institutional subscription	$625	2-year institutional subscription	$685
3-year institutional subscription	$875	3-year institutional subscription	$965
Electronic only		**Electronic only**	
1-year subscription	$ 85	1-year subscription	$ 85
2-year subscription	$155	2-year subscription	$155
3-year subscription	$215	3-year subscription	$215
1-year student subscription	$ 55	1-year student subscription	$ 55

Back issues of the *Stata Journal* may be ordered online at

http://www.stata.com/bookstore/sjj.html

Individual articles three or more years old may be accessed online without charge. More recent articles may be ordered online.

http://www.stata-journal.com/archives.html

The *Stata Journal* is published quarterly by the Stata Press, College Station, Texas, USA.

Address changes should be sent to the *Stata Journal*, StataCorp, 4905 Lakeway Drive, College Station, TX 77845, USA, or emailed to sj@stata.com.

The *Stata Journal* (ISSN 1536-867X) is a publication of Stata Press. Stata, Stata, Stata Press, Mata, Mata, and NetCourse are registered trademarks of StataCorp LP.

The Stata Journal (2015)
15, Number 4, pp. 1075–1097

Implementing a strategy to reduce the instrument count in panel GMM

Maria Elena Bontempi
Department of Economics
University of Bologna
Bologna, Italy
mariaelena.bontempi@unibo.it

Irene Mammi
Department of Economics
University of Bologna
Bologna, Italy
irene.mammi@unibo.it

Abstract. The problem of instrument proliferation and its consequences—overfitting of the endogenous explanatory variables, biased instrumental-variables and generalized method of moments estimators, and weakening of the power of the overidentification tests—are well known. This article introduces a statistical method to reduce the instrument count. Principal component analysis is applied on the instrument matrix, and the principal-component analysis scores are used as instruments for the panel generalized method of moments estimation. This strategy is implemented through the new command `pca2`.

Keywords: st0414, pca2, proliferation of instruments, principal component analysis, panel data, generalized method of moments

1 Introduction

The generalized method of moments (GMM) estimator, in the formulations of Arellano and Bond (1991), Arellano and Bover (1995), and Blundell and Bond (1998), has gained a leading role among the dynamic panel-data (DPD) estimators, mainly because of its flexibility and of the few assumptions it requires about the data-generating process. In addition, the availability of lags of the endogenous variables provides many instrumental variables (IVs) directly exploitable for GMM estimation. However, the estimation of DPD models by GMM with many instruments has its own drawbacks. In his seminal work, Sargan (1958) stressed that in the context of IV estimation, the marginal improvements from an increase in the number of instruments beyond three are generally small, whereas they can negatively affect the consistency of the estimates and the reliability of the specification tests. Since then, the potential distortions in parameter estimates when the instrument count gets larger have been extensively investigated in the literature (Kiviet [1995], Anderson and Sørenson [1996], Ziliak [1997], among others). In particular, instrument proliferation is intrinsic in GMM estimation of DPD models when all the lags of the endogenous explanatory variables are exploited, as the number of moment conditions increases with T and with the dimension of the vector of endogenous regressors. While, in principle, the availability of a wider set of conditions should improve efficiency (Dagenais and Dagenais 1997), the bias due to overfitting is quite severe as the number of moment conditions expands, which outweighs the gains in efficiency (Bekker 1994; Newey and Smith 2004; Ziliak 1997). Such tradeoff between bias and efficiency is exacerbated by the weak instrument problem (Bound, Jaeger, and Baker

1995, Staiger and Stock 1997) and by the correlation between the sample moments and the estimated optimal weighting matrix: sampling errors are magnified in the weighting matrix (Altonji and Segal 1996). Poor estimates of the variance–covariance matrix of the moments lower the power of the specification tests such as the Sargan/Hansen test for overidentifying restrictions, which suffers from a severe underrejection problem (Sargan 1958; Anderson and Sørenson 1996; Bowsher 2002).

Overall, such evidence supports the importance of properly addressing instrument proliferation, although this problem is often overlooked in empirical analyses; indeed, strategies to reduce the instrument count such as lag truncation and collapse (Roodman 2009a) are used only seldom in empirical applications. In addition to these two operational strategies, already implemented in Stata, the selection of correct or optimal instruments from a large set of potential candidates has received attention in a broader, more theoretical perspective. This latter stream of literature has developed statistically grounded methods for consistently selecting the GMM conditions and has investigated the statistical properties of the estimators exploiting the resulting sets of moments. Relevant contributions in this area include the information criteria methods and downward and upward testing procedures of Andrews (1999) and Andrews and Lu (2001), the Lasso-type instrument selection of Caner (2009) and Belloni et al. (2012), and the GMM shrinkage methods of Liao (2013). A recent contribution of Caner, Maasoumi, and Riquelme (2014) provides an extensive overview and a simulation-based comparison of moment-selection approaches.

Our aim in this article is to tackle the issue of instrument proliferation by providing a statistically grounded and directly implementable procedure that reduces the instrument count. We advocate the use of principal components analysis (PCA) of the instrument matrix as a way to shrink the available instruments into a set of linear combinations of the original variables (the scores of the PCA). The weights used in such orthogonal combinations follow from the main features of the data and reflect the contribution of each variable to the total observed variability. We label this strategy "principal components instrumental variables reduction" (PCIVR).[1]

PCIVR exploits the same tool as that found in Doran and Schmidt (2006), who propose an eigenvalue-eigenvector decomposition of the GMM weighting matrix to reduce its dimension; however, our method implies the drop of linear combinations of the instruments rather than of linear combinations of the moment conditions. Moreover, while the strategy proposed by Doran and Schmidt (2006) is applicable to all overidentified GMM problems that require a weighting matrix in the estimation procedure, our procedure directly addresses the instrument matrix. Therefore, it can be applied to any IV estimation problem with many IVs.

The `pca2` command directly implements PCIVR by means of an ado-file.[2] Thanks to its flexibility, the `pca2` procedure adds useful features to the Stata command `pca`. It is straightforwardly applicable in Stata to any type of dataset (cross-section, time

1. A first sketch of a PCA-based reduction of GMM-style IVs can be found in Mehrhoff (2009).
2. The `pca` option in the user-written `xtabond2` (Roodman 2009b) command provides a first application of PCA on GMM-style IVs within Stata.

series, and panel), and in a single command line, it automates, through specific options, alternative ways to extract the principal components and to select those to be retained for the computation of the PCA scores. More specifically, `pca2` first allows GMM-style instruments for one or more variables to be generated, either in levels or in first-differences (or both), introducing the possibility of defining a specific lag structure for each variable. Then, the procedure extracts the principal components either separately from each instrument set or from the matrix that jointly includes all the instrumental variables. Finally, it retains a certain number of principal components according to alternative criteria specified by the user and predicts the corresponding scores, adding them to the dataset as new variables to be used as instruments.

The procedure presents several distinguishing features when compared to the existing strategies to reduce the instrument count developed in the literature, and it extends the set of tools available to the researcher for the purpose.

First, `pca2` acts as a complementary tool with respect to lag truncation and collapse, which impose a priori restrictions not tailored on the data. The main difference between these tools and `pca2` is that our approach provides a flexible statistical rule for the selection of nonredundant instruments that adjusts to the empirical problem at hand and reflects the specific features of the data. Lag truncation assumes that the relevant information is conveyed only by the most recent (usually one or two) available lags of the endogenous variables, while the collapsing of the instrument matrix assumes specific dynamics in the data. Because such assumptions cannot be tested a priori, to identify potential critical aspects related to the issue at hand, we recommend comparing the GMM estimates obtained with lag truncation and collapsing with those provided by PCIVR, which has the advantage of exploiting information from the whole set of instruments to select the lags that contribute to total variability the most.

Second, if compared with the other approaches mentioned at the beginning of this section, `pca2` becomes especially attractive when the large number of potential instruments makes moment selection procedures such as those proposed by Andrews (1999) and Andrews and Lu (2001) potentially cumbersome to implement. In fact, the high number of candidate subsets renders the identification of the correct orthogonality conditions based on the J-statistic (Liao 2013) computationally intensive. At the same time, it may be sometimes challenging to apply the GMM shrinkage estimators proposed by Liao (2013): the problem at hand may not provide strong enough prior beliefs to split the IVs in the two distinct sets that allow separation of "credibly" valid and potentially invalid moment conditions.

Third, a broader perspective suggests that a unified framework can be conceived to address the problem of instrument as well as regressor or predictor proliferation using alternative selection algorithms (for example, least-angle regression, forward stagewise regression, and lasso estimation) that involve a prescreening of the variables of interest to extract the subset of those that are correlated the most with the target variable. In the Stata framework, the user-written `lars.ado` command by Mander (2006) implements these alternative algorithms for the selection of a subset of targeted variables. More specifically, in an instrumental variable estimation context, the `lars` command

can implement the approach of Belloni et al. (2012), which estimates optimal IVs in linear models with many valid instruments by selecting those that convey the strongest information about the target (endogenous) variables. This prescreening approach can be seen as the first step of a sequential strategy that combines targeted IVs and a subsequent PCA on the selected instruments. By following this approach, the researcher can improve the overall efficiency of instrument-reduction procedures and move a step toward the identification of better IVs. This is for instance in line with the findings of Bai and Ng (2008) in the forecasting context. They show that the extraction of principal components from a set including fewer but more informative predictors leads to the selection of better ones than those obtained by applying PCA to the original set. In this perspective, the use of `pca2` can be seen as a complementary tool with respect to the `lars` procedure and can be applied once the IVs have already been prescreened.

The rest of the article is organized as follows. Section 2 summarizes the main methodological underpinnings of the strategy we present: section 2.1 reviews the GMM estimation of DPD models, and section 2.2 describes the extraction of principal components from a matrix of instruments. Section 3 details the syntax of `pca2` and its options and also provides some empirical examples. Section 4 carries out a guided example of robustness analysis in the context of published research on the determinants of the discretionary fiscal policy in the Euro-area countries.

2 The methodological framework

2.1 GMM estimation of DPD models

Consider the general two-way error component DPD model

$$y_{it} = \alpha y_{it-1} + \boldsymbol{\beta}'\mathbf{x}_{i,t} + \phi_t + \upsilon_{it} \ , \ \upsilon_{it} = \eta_i + \varepsilon_{it} \tag{1}$$

where $i = 1, \ldots, N$, $t = 1, \ldots, T$, $\mathbf{x}$ is an m-dimensional vector of potentially endogenous or predetermined regressors, the ϕ_t are the time effects, the η_i are the individual effects, and ε_{it} is a zero-mean idiosyncratic error, allowed to be heteroskedastic but not serially correlated. The standard assumptions are $E(\eta_i) = E(\varepsilon_{it}) = E(\eta_i \varepsilon_{it}) = 0$ and predetermined initial conditions $E(y_{i1}\varepsilon_{it}) = 0$.

The Arellano–Bond and Arellano–Bover/Blundell–Bond estimators are linear GMM estimators for the model in (1) in first-differences (DIF GMM) or in levels (LEV GMM) or both (SYS GMM); the instrument matrix $\mathbf{Z}$ includes the lagged values of the endogenous variables. The columns of $\mathbf{Z}$ correspond, respectively, to two different sets of meaningful moment conditions.

The Arellano–Bond DIF GMM estimator exploits the following moment conditions for the (1) in first-differences,

$$E\{(\mathbf{Z_i}^{\text{dif}})'\Delta\upsilon_i\} = E\{(\mathbf{Z_{i,t-l}}^{\text{dif}})'\Delta\upsilon_{it}\} = 0 \quad \text{for} \quad t \geq 3, l \geq 2 \tag{2}$$

where l denotes the lag depth.

The Blundell–Bond SYS GMM estimator also exploits the additional nonredundant orthogonality conditions for the (1) in levels:[3]

$$E\{(\mathbf{Z_i}^{\text{lev}})'\upsilon_i\} = E\{(\mathbf{Z_{is}}^{\text{lev}})'\upsilon_{iT}\} = 0 \text{ for } s = 2, \ldots, T-1 \tag{3}$$

Because DPD GMM uses lags of the explanatory variables as IVs, according to Han and Phillips (2006, 149), "the phenomenon of moment condition proliferation is far from being a theoretical construct and arises in a natural way in many empirical econometric settings". The dimension of the GMM-type instrument matrix grows as the number of time periods and endogenous regressors expands.

2.2 Extracting principal components from the matrix of instruments

The adoption of PCA or factor analysis to extract a small number of factors from a large set of variables has become popular in macroeconometrics, forecasting being the main field of application. Stock and Watson (2002) prove consistency of the factors as the number of original variables gets sufficiently large so that the principal components are estimated precisely enough to be used instead of the original variables in subsequent regressions. Kloek and Mennes (1960) and Amemiya (1966) first propose the use of principal components in the IV estimation. Important recent contributions, among the others, are Kapetanios and Marcellino (2010), Groen and Kapetanios (2009), and Bai and Ng (2010).[4]

The issue of instrument proliferation can be addressed by extracting the principal components from the instrument matrix $\mathbf{Z}$. The aim of PCIVR is to reexpress the information conveyed by highly correlated variables in terms of a set of optimal orthogonal linear combinations of the original variables and then to retain a smaller number of them.

In detail, with $\mathbf{Z}$ defined as the general p-columns GMM-style instrument matrix, we extract p eigenvalues $\lambda_1, \lambda_2, \ldots, \lambda_p \geq 0$ from the correlation or covariance matrix of $\mathbf{Z}$, ordered from the largest to the smallest, and derive the corresponding eigenvectors (principal components) $\mathbf{u}_1, \mathbf{u}_2, \ldots, \mathbf{u}_p$. Our new instruments will be the scores from PCA that are defined as

$$\mathbf{s}_k = \mathbf{Z}\mathbf{u}_k \text{ for } k = 1, 2, \ldots, p$$

If we write $\mathbf{Z} = (\mathbf{z}_1 \ldots \mathbf{z}_j \ldots \mathbf{z}_p)$ with $\mathbf{z}_j$ being the jth column of the instrument matrix, the score $\mathbf{s}_k$ corresponding to the kth component can be rewritten as

$$\mathbf{s}_k = u_{k1}\mathbf{z}_1 + \cdots + u_{kj}\mathbf{z}_j + \cdots + u_{kp}\mathbf{z}_p$$

where u_{kj} is the jth element of the principal component $\mathbf{u}_k$. With the matrix of PCA loadings defined as $\mathbf{V} = (\mathbf{u}_1 \ldots \mathbf{u}_k \ldots \mathbf{u}_p)$ and the matrix of PCA scores defined as $\mathbf{S}$, we

3. The LEV GMM estimation considers, for each endogenous variable, time period, and lag distance, all the available lags for the equation in levels because they are nonredundant.
4. A review of the literature on factor-IV and factor-GMM estimations is in the introduction of Kapetanios and Marcellino (2010).

obtain $\mathbf{S} = \mathbf{ZV}$. Instead of the moment conditions in (2), we will therefore exploit the following restrictions in GMM DIF:

$$E\{(\mathbf{S}^{\text{dif}})'\Delta\boldsymbol{v}\} = E\{(\mathbf{Z}^{\text{dif}}\mathbf{V})'\Delta\boldsymbol{v}\} = 0 \tag{4}$$

Similarly, in the GMM SYS, we will also exploit the following additional orthogonality conditions instead of those in (3):

$$E\{(\mathbf{S}^{\text{lev}})'\boldsymbol{v}\} = E\{(\mathbf{Z}^{\text{lev}}\mathbf{V})'\boldsymbol{v}\} = 0 \tag{5}$$

Because the aim of the PCIVR is the reduction of the dimension of the instrument matrix, a criterion to select the scores to be retained has to be adopted. The idea is retaining only $(m+1) \leq q < p$ principal components, where m is the number of endogenous regressors other than the lagged dependent variable; thus only the q corresponding score vectors will form the new transformed instrument matrix in both (4) and (5).

One possibility is to retain the q principal components corresponding to eigenvalues above the average of the eigenvalues ("average criterion"); alternatively, one may keep those accounting for a given percentage of the variance of the data, generally 70% to 90% ("variance criterion").

The number of moment restrictions resulting from the PCIVR depends on the nature of the data at hand. If $q < (m+1)$, the equation of interest is not identified. For instance, this can happen when the variables are highly persistent (near unit-root processes): in this case, the PCA is driven by spurious trends, and very few principal components are retained.

3 The pca2 command

3.1 Syntax

The user-written command `pca2` implements the PCIVR procedure presented above: in a unique step, it extracts the principal components from the variables in *varlist* according to the preferences specified through its options; then it computes the scores corresponding to the principal components retained on the basis of the selection criterion chosen by the researcher. These scores can be used in any IV/GMM estimation command in Stata in place of the original IVs.

The extraction of principal components through the `pca2` command exploits the Stata `pca` command. Its innovative feature consists of augmenting the `pca` command with specific options for the creation of GMM-style IVs, for the selection of principal components, and for the computation of the scores.

The syntax of pca2 is

pca2 *varlist* [*if*] [*in*] [, nt(*timevar*|*panelvar timevar*) variance(#) avg covariance prefix(*string*) see gmmliv(#|# #) gmmdiv(#|# #) lagsl(*varlistl*, ll(#|# #)) lagsd(*varlistd*, ll(#|##)) togvar togld retain]

Time-series and panel data must be tsset before using pca2. See [TS] **tsset** for more information. pca2 does not allow time-series operators in *varlist*; to use lags of the variables in *varlist*, you must generate them using Stata time-series operators before applying the pca2 command; see help tsvarlist.

3.2 Options

nt(*timevar*|*panelvar timevar*) is required in time series and panel data to create GMM-style instruments and to apply PCA on them. If this option is omitted, the dataset is treated as a cross-section, and all the observations are pooled.

variance(#) allows you to apply the variance criterion (default criterion); that is, only those principal components that account for at least the chosen percentage of the variability in the original data are retained for the computation of the scores. The number defining the percentage must be an integer greater than 0 and lower or equal to 100. The default is variance(90).

avg selects the principal components to be kept for score computation according to the average criterion; that is, only those eigenvectors whose corresponding eigenvalues are above the average of the eigenvalues are retained. Note that when the avg option is chosen, pca2 also computes the scores according to the default 90% variance criterion and saves both of them in the dataset: the scores obtained according to the two criteria can thus be compared.

covariance performs PCA of the covariance matrix. The default is to perform PCA on the correlation matrix; see help pca.

prefix(*string*) specifies the prefix for the name of the scores generated by the pca2 command corresponding to the retained principal components. For example, if you write prefix(sys), you will obtain _sys_varscore* and _sys_avgscore*. This option is particularly useful when the pca2 command is repeated many times on the same dataset to create different scores from different instrument sets, eventually according to different criteria. The default prefix is _BM, which retains the scores with labels such as _BM_varscore* and _BM_avgscore*.

see asks Stata to display the outcome of the PCA.

gmmliv(#|# #) generates the GMM-style instruments in levels (for the equations in first-differences) for all the variables included in the pca2 *varlist*. If only one argument is specified (for example, gmmliv(k)), all the available lags from $t-k$ back to

the initial observation for each variable in *varlist* of the `pca2` command are used. If two arguments are specified (for example, `gmmliv(`*k1 k2*`)` with $k1 \leq k2$), the lags from $t - k1$ to $t - k2$ are considered. The PCA is performed on all the specified GMM-style lags in levels of each variable taken separately. If the `togvar` option (full description below) is also added, the PCA is performed on all the generated GMM-style lags in levels of all the variables in *varlist* considered together. With this option, the lag structure is the same for each variable.

`gmmdiv(`*#* | *# #*`)` generates the GMM-style instruments in first-differences (for the equations in levels) for all the variables included in the `pca2` *varlist*. If only one argument is specified (for example, `gmmdiv(`*k*`)`), all the available lags from $t - k$ back to the initial observation for each variable in *varlist* of the `pca2` command are used. If two arguments are specified (for example, `gmmdiv(`*k1 k2*`)` with $k1 \leq k2$), the lags from $t - k1$ to $t - k2$ are considered. The PCA is performed on all the specified GMM-style lags in first-differences of each variable taken separately. If the `togvar` option (full description below) is also added, the PCA is performed on all the generated GMM-style lags in first-differences of all the variables in *varlist* considered together. With this option, the lag structure is the same for each variable.

`lagsl(`*varlistl*`, ll(`*#* | *# #*`))` generates the GMM-style instruments in levels for a specific *varlistl*. It is a more flexible alternative to the `gmmliv()` option because it allows for a different lag structure of each variable. The `lagsl()` option may be used more than once: different lag structures may thus be defined for the variables in each *varlistl*. The `ll()` suboption specifies the lag structure of the variables in each *varlistl*: if only one argument is specified (for example, `ll(`*k*`)`, all the available lags from $t - k$ back to the initial observation for each variable in *varlistl* are used. If two arguments are specified (for example, `ll(`*k1 k2*`)` with $k1 \leq k2$), the lags from $t - k1$ to $t - k2$ are considered. The PCA is performed on all the specified GMM-style lags in levels of each variable taken separately. If the `togvar` option (full description below) is also added, the PCA is performed on all the generated GMM-style lags in levels of all the variables in *varlistl* considered together. `lagsl()` cannot be used with the `gmmliv()` option, while it is allowed with either the `lagsd()` option or the `gmmdiv()` option. When `lagsl()` is used alone or with `lagsd()`, the number of variables in both *varlistl* and *varlistd* must be at least equal to the number of variables in *varlist* of the `pca2` command. The `lagsl()` option can have fewer variables than those included in *varlist* of the `pca2` command only when associated with `gmmdiv()`.

`lagsd(`*varlistd*`, ll(`*#* | *# #*`))` generates the GMM-style instruments in first-differences for a specific *varlistd*. It is a more flexible alternative to the `gmmdiv()` option because it allows for a different lag structure of each variable. The `lagsd()` option may be used more than once: different lag structures may thus be defined for the variables in each *varlistd*. The `ll()` suboption specifies the lag structure of the variables in each *varlistd*: if only one argument is specified (for example, `ll(`*k*`)`), all the available lags from $t - k$ back to the initial observation for each variable in *varlistd* are used. If two arguments are specified (for example, `ll(`*k1 k2*`)` with $k1 \leq k2$), the lags from $t - k1$ to $t - k2$ are considered. The PCA is performed on all the specified GMM-style lags in first-differences of each variable taken separately. If the `togvar` option

(full description below) is also added, the PCA is performed on all the generated GMM-style lags in first-differences of all the variables in *varlistd* considered together. `lagsd()` cannot be used with the `gmmdiv()` option, while it is allowed with either the `lagsl()` option or the `gmmliv()` option. When `lagsd()` is used alone or with `lagsl()`, the number of variables in both *varlistl* and *varlistd* must be at least equal to the number of variables in *varlist* of the `pca2` command. The `lagsd()` option can have fewer variables than those included in *varlist* of the `pca2` command only when associated with `gmmliv()`.

`togvar` specifies that the PCA be performed on the matrix that includes all the variables in *varlist* and not on each variable separately. For example, the syntax `pca2 x z, togvar` implies that the PCA is performed jointly on the variables `x` and `z`. This option must be specified to apply the PCA to GMM-style lags of more than one variable taken together instead of the lags of each variable taken separately. For example, `pca2 x z, gmmliv(2) togvar` implies that the principal components are extracted from the matrix that includes all the available lags for the variables `x` and `z` in levels from $t-2$, $t-3$, and so on.

`togld` specifies that once instruments in levels and first-differences are generated, the PCA is applied to the matrix that includes all of these instruments together for each variable in *varlist* of `pca2`. If the `togld` option is used with the `togvar` option, the principal components are extracted from the matrix that includes all the lags in first-differences and in levels of all the variables in *varlist*.

`retain` adds the generated GMM-style IVs as new variables to the dataset. These IVs are named `_GMML`*varnamePERIODlag* and `_GMMD`*varnamePERIODlag*; for example, `_GMMLn1978L2` stands for the $t-2$ observation in levels for the variable n in the year $t = 1978$.

3.3 The use of the pca2 command: An example

We illustrate the `pca2` command through an empirical example based on `abdata.dta` used in Arellano and Bond (1991) and Blundell and Bond (1998).

We fit the Blundell and Bond (1998) model, a simple autoregressive distributed lags model of labor demand,

$$n_{it} = \alpha n_{it-1} + \beta_0 w_{it} + \beta_1 w_{it-1} + \gamma_0 k_{it} + \gamma_1 k_{it-1} + \eta_i + \phi_t + \nu_{it} \tag{6}$$

where n_{it}, w_{it}, and k_{it} are the log of employment, the log of the real product wage, and the log of the capital stock in firm i in year t, respectively. The sample is an unbalanced panel of 140 UK-listed manufacturing companies with between 7 and 9 annual observations over the period 1976–1984.

First, we replicate the original DIF GMM results in column 3 of table 4 in Blundell and Bond (1998); then, we fit the same model by DIF GMM estimates, exploiting the set of IVs resulting from the PCA.

To do so, we run the command

```
pca2 n w k, nt(id year) gmmliv(2) retain avg
```

This syntax generates the GMM-style instruments in levels for each of the `n`, `w`, and `k` variables. These variables are labeled `_GMML*` and will be used as instruments for (6) in first-differences. In this case, the `gmmliv(2)` option specifies the same lag structure for all the variables, and the IVs are generated from lag $t-2$ up to the last lag available. By specifying the `retain` option, we add the `_GMML*` instruments as new variables in the dataset.

Then, we separately extract the principal components for each variable from its own lags. Next, because we use the `avg` option, we retain the principal components according to both selection criteria (that is, the default variance criterion and the average criterion). Finally, we save the corresponding scores in the dataset as new variables labeled `_BM_var*` and `_BM_avg*`, where `var` and `avg` refer to the selection criterion.

As shown below, the output of the `pca2` command reports information about the lag structure of the GMM-style IVs and summary statistics for the extraction of the principal components.

```
. use http://fmwww.bc.edu/ec-p/data/macro/abdata.dta
(Layard & Nickell, Unemployment in Britain, Economica 53, 1986 from Ox dist)
. xtset id year
  (output omitted)
. quietly tabulate year, generate(tauyear)
. pca2 n w k, nt(id year) gmml(2) retain avg
General description of the dataset
       panel variable:  id (unbalanced)
        time variable:  year, 1976 to 1984
                delta:  1 unit
The prefix is: _BM_
You are creating GMM-style IVs in levels for a panel
_____ variable: n _____
Lag selection in GMML(): from t-2 to the last available lag
_____ variable: w _____
Lag selection in GMML(): from t-2 to the last available lag
_____ variable: k _____
Lag selection in GMML(): from t-2 to the last available lag
_____ PCA LEV VAR BY VAR: n
You are applying PCA to GMM-style LEV lags of one or more than one variable,
keeping the variables separated with the same lags structure
 _________ Some information about PCA of IV in levels for n __________
Trace of the matrix:
> 28
By default percentage of selected variability to be explained:
> 90%
Percentage of variance explained by the variability criterion:
> 92.943733%
Number of retained scores according to the variability criterion:
> 8
Percentage of variance explained by the average criterion:
> 86.399506%
Number of retained scores according to the average criterion:
> 6
```

```
_____ PCA LEV VAR BY VAR: w
You are applying PCA to GMM-style LEV lags of one or more than one variable,
keeping the variables separated with the same lags structure
 _________ Some information about PCA of IV in levels for w __________
Trace of the matrix:
> 28
Percentage of variance explained by the variability criterion:
> 90.305677%
Number of retained scores according to the variability criterion:
> 7
Percentage of variance explained by the average criterion:
> 87.588082%
Number of retained scores according to the average criterion:
> 6
_____ PCA LEV VAR BY VAR: k
You are applying PCA to GMM-style LEV lags of one or more than one variable,
keeping the variables separated with the same lags structure
 _________ Some information about PCA of IV in levels for k __________
Trace of the matrix:
> 28
Percentage of variance explained by the variability criterion:
> 90.223503%
Number of retained scores according to the variability criterion:
> 7
Percentage of variance explained by the average criterion:
> 86.652737%
Number of retained scores according to the average criterion:
> 6
```

To get the original DIF GMM estimates for the model in (6), we can use the user-written `xtabond2` command (see Roodman [2009b]) with its native syntax:

```
. quietly xtabond2 n l.n l(0/1).(w k) tauyear3-tauyear9,
> ivstyle(tauyear3-tauyear9, equation(diff))
> gmmstyle(n, laglimits(2 .) equation(diff))
> gmmstyle(w, laglimits(2 .) equation(diff))
> gmmstyle(k, laglimits(2 .) equation(diff))
> h(2) noleveleq robust nodiffsargan
```

However, to illustrate how to exploit the instrumental variables obtained through the `pca2` command (in this case, the `_GMML*` IVs just added to the dataset), we can reproduce the same estimates by typing the following, where the new variables are used as traditional IVs through the option `ivstyle()`. The two commands generate the same output.

```
. xtabond2 n l.n l(0/1).(w k) tauyear3-tauyear9,
> ivstyle(tauyear3-tauyear9, equation(diff))
> ivstyle(_GMML_n_*, equation(diff) pass)
> ivstyle(_GMML_w_*, equation(diff) pass) ivstyle(_GMML_k_*, equation(diff) pass)
> h(2) noleveleq robust nodiffsargan
Favoring space over speed. To switch, type or click on mata: mata set matafavor
> speed, perm.
Dynamic panel-data estimation, one-step difference GMM
------------------------------------------------------------------------------
Group variable: id                              Number of obs      =       751
Time variable : year                            Number of groups   =       140
Number of instruments = 91                      Obs per group: min =         5
Wald chi2(12) =   1163.33                                      avg =      5.36
Prob > chi2   =     0.000                                      max =         7
------------------------------------------------------------------------------
                    |               Robust
                  n |      Coef.   Std. Err.      z    P>|z|     [95% Conf. Interval]
--------------------+---------------------------------------------------------------
                  n |
                L1. |   .7074701   .0841788     8.40   0.000     .5424827    .8724576
                    |
                  w |
                --. |  -.7087965    .117102    -6.05   0.000    -.9383122   -.4792809
                L1. |   .5000149   .1113282     4.49   0.000     .2818157    .7182141
                    |
                  k |
                --. |   .4659776    .101044     4.61   0.000      .267935    .6640203
                L1. |  -.2151309   .0858525    -2.51   0.012    -.3833987   -.0468631
                    |
           tauyear3 |   .0057636   .0166077     0.35   0.729    -.0267868     .038314
           tauyear4 |   .0136366   .0193748     0.70   0.482    -.0243374    .0516106
           tauyear5 |  -.0071557   .0213479    -0.34   0.737    -.0489969    .0346855
           tauyear6 |  -.0340692   .0264327    -1.29   0.197    -.0858763    .0177379
           tauyear7 |  -.0059175   .0272325    -0.22   0.828    -.0592922    .0474573
           tauyear8 |   .0187213   .0288529     0.65   0.516    -.0378294     .075272
           tauyear9 |   .0352279   .0331578     1.06   0.288    -.0297603    .1002161
------------------------------------------------------------------------------
Instruments for first differences equation
  Standard
    _GMML_k_1978L2 _GMML_k_1979L2 _GMML_k_1979L3 _GMML_k_1980L2 _GMML_k_1980L3
  (output omitted)
    _GMML_n_1984L6 _GMML_n_1984L7 _GMML_n_1984L8
    D.(tauyear3 tauyear4 tauyear5 tauyear6 tauyear7 tauyear8 tauyear9)
------------------------------------------------------------------------------
Arellano-Bond test for AR(1) in first differences: z =  -5.60  Pr > z =  0.000
Arellano-Bond test for AR(2) in first differences: z =  -0.14  Pr > z =  0.891
------------------------------------------------------------------------------
Sargan test of overid. restrictions: chi2(79)   = 125.19  Prob > chi2 =  0.001
  (Not robust, but not weakened by many instruments.)
Hansen test of overid. restrictions: chi2(79)   =  88.80  Prob > chi2 =  0.211
  (Robust, but weakened by many instruments.)
```

We now fit the model in (6) by DIF GMM on the set of instruments that results from PCIVR. The `pca2` command run above saves the PCA scores `_BM_varscoreDIF*` and `_BM_avgscoreDIF*` as new variables in the dataset. Therefore, we can get the estimates on the new set of instruments by using, for example, the variables `_BM_var*` in `xtabond2` as new instruments instead of the standard ones as follows:

```
. xtabond2 n l.n l(0/1).(w k) tauyear3-tauyear9,
> ivstyle(tauyear3-tauyear9, equation(diff))
> ivstyle(_BM_var*n*, equation(diff) pass)
> ivstyle(_BM_var*w*, equation(diff) pass)
> ivstyle(_BM_var*k*, equation(diff) pass)
> h(2) noleveleq robust nodiffsargan
Favoring space over speed. To switch, type or click on mata: mata set matafavor
> speed, perm.
Dynamic panel-data estimation, one-step difference GMM
------------------------------------------------------------------------------
Group variable: id                              Number of obs      =       751
Time variable : year                            Number of groups   =       140
Number of instruments = 29                      Obs per group: min =         5
Wald chi2(12) =   1146.02                                      avg =      5.36
Prob > chi2   =     0.000                                      max =         7
------------------------------------------------------------------------------
                    |               Robust
                  n |      Coef.   Std. Err.      z    P>|z|     [95% Conf. Interval]
--------------------+---------------------------------------------------------------
                  n |
                L1. |   .8021886   .1255146     6.39   0.000     .5561845    1.048193
                    |
                  w |
                --. |  -.8621674   .2094745    -4.12   0.000     -1.27273   -.4516048
                L1. |   .2224614   .2941419     0.76   0.449    -.3540461     .798969
                    |
                  k |
                --. |   .5783907   .2253891     2.57   0.010     .1366362    1.020145
                L1. |  -.4108413   .1947894    -2.11   0.035    -.7926216    -.029061
                    |
           tauyear3 |  -.0202252   .0272124    -0.74   0.457    -.0735604      .03311
           tauyear4 |  -.0114123   .0355594    -0.32   0.748    -.0811074    .0582829
           tauyear5 |  -.0209936   .0374262    -0.56   0.575    -.0943475    .0523603
           tauyear6 |   -.034543    .049461    -0.70   0.485    -.1314848    .0623988
           tauyear7 |   .0148526   .0524715     0.28   0.777    -.0879897    .1176949
           tauyear8 |   .0556274   .0447092     1.24   0.213     -.032001    .1432558
           tauyear9 |   .0688565   .0555122     1.24   0.215    -.0399454    .1776584
------------------------------------------------------------------------------
Instruments for first differences equation
  Standard
    _BM_varscoreLEVkN1 _BM_varscoreLEVkN2 _BM_varscoreLEVkN3
  (output omitted)
    _BM_varscoreLEVnN7 _BM_varscoreLEVnN8
    D.(tauyear3 tauyear4 tauyear5 tauyear6 tauyear7 tauyear8 tauyear9)
------------------------------------------------------------------------------
Arellano-Bond test for AR(1) in first differences: z =  -3.41  Pr > z =  0.001
Arellano-Bond test for AR(2) in first differences: z =  -0.61  Pr > z =  0.544
------------------------------------------------------------------------------
Sargan test of overid. restrictions: chi2(17)   =  32.49  Prob > chi2 =  0.013
  (Not robust, but not weakened by many instruments.)
Hansen test of overid. restrictions: chi2(17)   =  23.43  Prob > chi2 =  0.136
  (Robust, but weakened by many instruments.)
```

Because the aim of the PCIVR is the reduction in the instrument count, as expected, the Hansen test has 79 degrees of freedom in the standard DIF GMM estimates, while they fall to 17 when the scores relative to the principal components are extracted according to the variance criterion.

So far, we have focused on the syntax for the GMM-style instruments and PCA scores in the DIF GMM estimation. In addition, we can use `pca2` to create IVs and PCA scores to be used in SYS GMM; we can thus replicate the results in column 4 of table 4 in Blundell and Bond (1998) and get SYS GMM estimates with the PCA scores as instruments.

The syntax

```
. use http://fmwww.bc.edu/ec-p/data/macro/abdata.dta, clear
(Layard & Nickell, Unemployment in Britain, Economica 53, 1986 from Ox dist)
. xtset id year
  (output omitted)
. quietly tabulate year, generate(tauyear)
. pca2 n w k, nt(id year) gmml(2) gmmd(1 1) retain avg
  (output omitted)
```

creates both the instruments in levels from $t-2$ up to the last lag available and first-differences for the first available lag. The IVs in levels (that is, the `_GMML*` variables) and the instruments in first-differences (that is, the `_GMMD*` variables) are included in the dataset as new variables.[5] The PCA is run on the instruments in first-differences and on the instruments in levels for each variable separately; the scores relative to the retained principal components (`_BM_varscoreDIF*` and `_BM_avgscoreDIF*`, `_BM_varscoreLEV*` and `_BM_avgscoreLEV*`) are also added to the dataset.

5. When the `pca2` command is run more than once, the researcher can exploit the `prefix()` option to define the names of the new scores to be added in the dataset, thus also maintaining in the dataset the ones created previously. Of course, it is not possible to save GMM-style IVs already present in the dataset: when the researcher needs to regenerate and store the same IVs again, those created previously have to be canceled by typing, for example, `drop _GMM*`.

Following the same line of reasoning, we can get the standard SYS GMM estimates by using `xtabond2` with the `_GMMD*` and `_GMML*` variables as instruments through the following command:

```
. xtabond2 n l.n l(0/1).(w k) tauyear3-tauyear9,
> ivstyle(tauyear3-tauyear9, equation(both))
> ivstyle(_GMML_n_*, equation(diff) pass)
> ivstyle(_GMML_w_*, equation(diff) pass)
> ivstyle(_GMML_k_*, equation(diff) pass)
> ivstyle(_GMMD_n_*L1, equation(lev)) ivstyle(_GMMD_w_*L1, equation(lev))
> ivstyle(_GMMD_k_*L1, equation(lev)) h(1) robust nodiffsargan
Favoring space over speed. To switch, type or click on mata: mata set matafavor
> speed, perm.
Dynamic panel-data estimation, one-step system GMM
```

Group variable: id			Number of obs	= 891
Time variable : year			Number of groups	= 140
Number of instruments = 113			Obs per group: min	= 6
Wald chi2(12) = 4147.85			avg	= 6.36
Prob > chi2 = 0.000			max	= 8

n	Coef.	Robust Std. Err.	z	P>|z|	[95% Conf.	Interval]
n						
L1.	.8108394	.0579982	13.98	0.000	.6971649	.9245138
w						
--.	-.7945394	.0971517	-8.18	0.000	-.9849532	-.6041257
L1.	.55012	.151645	3.63	0.000	.2529012	.8473388
k						
--.	.4285055	.0763361	5.61	0.000	.2788895	.5781215
L1.	-.2802184	.0776689	-3.61	0.000	-.4324466	-.1279903
tauyear3	.0077488	.0200664	0.39	0.699	-.0315806	.0470781
tauyear4	.020829	.0236973	0.88	0.379	-.025617	.0672749
tauyear5	-.0002589	.0252166	-0.01	0.992	-.0496826	.0491648
tauyear6	-.0271456	.02961	-0.92	0.359	-.0851801	.030889
tauyear7	.0012306	.026954	0.05	0.964	-.0515983	.0540596
tauyear8	.014436	.0254967	0.57	0.571	-.0355367	.0644087
tauyear9	.0003278	.0307739	0.01	0.992	-.059988	.0606436
_cons	1.006162	.430149	2.34	0.019	.1630853	1.849238

```
Instruments for first differences equation
  Standard
    _GMML_k_1978L2 _GMML_k_1979L2 _GMML_k_1979L3 _GMML_k_1980L2 _GMML_k_1980L3
  (output omitted)
    D.(tauyear3 tauyear4 tauyear5 tauyear6 tauyear7 tauyear8 tauyear9)
Instruments for levels equation
  Standard
    _GMMD_k_1978L1 _GMMD_k_1979L1 _GMMD_k_1980L1 _GMMD_k_1981L1 _GMMD_k_1982L1
  (output omitted)
    tauyear3 tauyear4 tauyear5 tauyear6 tauyear7 tauyear8 tauyear9
    _cons
```

```
Arellano-Bond test for AR(1) in first differences: z =  -6.49  Pr > z =  0.000
Arellano-Bond test for AR(2) in first differences: z =  -0.08  Pr > z =  0.934

Sargan test of overid. restrictions: chi2(100)  = 113.34  Prob > chi2 =  0.171
  (Not robust, but not weakened by many instruments.)
Hansen test of overid. restrictions: chi2(100)  = 115.73  Prob > chi2 =  0.135
  (Robust, but weakened by many instruments.)
```

Similarly, we can get the SYS GMM estimates with the set of PCA scores from PCIVR (_BM_varscoreDIF* and _BM_avgscoreDIF*, _BM_varscoreLEV* and _BM_avgscoreLEV*) as follows:

```
. xtabond2 n l.n l(0/1).(w k) tauyear3-tauyear9,
> ivstyle(_BM_varscoreLEVn*, equation(diff) pass)
> ivstyle(_BM_varscoreLEVw*, equation(diff) pass)
> ivstyle(_BM_varscoreLEVk*, equation(diff) pass)
> ivstyle(_BM_varscoreDIFn*, equation(lev) pass)
> ivstyle(_BM_varscoreDIFw*, equation(lev) pass)
> ivstyle(_BM_varscoreDIFk*, equation(lev) pass)
> ivstyle(tauyear3-tauyear9, equation(both)) h(1) robust nodiffsargan
Favoring space over speed. To switch, type or click on mata: mata set matafavor
> speed, perm.
Dynamic panel-data estimation, one-step system GMM
------------------------------------------------------------------------------
Group variable: id                              Number of obs      =       891
Time variable : year                            Number of groups   =       140
Number of instruments = 51                      Obs per group: min =         6
Wald chi2(12) =   5587.27                                      avg =      6.36
Prob > chi2   =     0.000                                      max =         8
------------------------------------------------------------------------------
             |               Robust
           n |      Coef.   Std. Err.      z    P>|z|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
           n |
          L1.|   .9016193   .0477017    18.90   0.000     .8081257     .995113
             |
           w |
          --.|   -.742429   .1542546    -4.81   0.000    -1.044763   -.4400956
          L1.|   .4643432   .1950932     2.38   0.017     .0819675    .8467189
             |
           k |
          --.|     .53362    .096368     5.54   0.000     .3447423    .7224978
          L1.|  -.4411184   .1025934    -4.30   0.000    -.6421978    -.240039
             |
    tauyear3 |  -.0025501   .0226948    -0.11   0.911    -.0470312     .041931
    tauyear4 |   .0129266   .0272024     0.48   0.635    -.0403891    .0662423
    tauyear5 |   .0004112   .0272325     0.02   0.988    -.0529634    .0537858
    tauyear6 |  -.0197377   .0340792    -0.58   0.562    -.0865317    .0470564
    tauyear7 |   .0179079   .0346922     0.52   0.606    -.0500877    .0859034
    tauyear8 |   .0328657   .0278118     1.18   0.237    -.0216443    .0873758
    tauyear9 |      .0287   .0339306     0.85   0.398    -.0378027    .0952028
       _cons |   .9899051   .3951924     2.50   0.012     .2153422    1.764468
------------------------------------------------------------------------------
```

```
Instruments for first differences equation
  Standard
    D.(tauyear3 tauyear4 tauyear5 tauyear6 tauyear7 tauyear8 tauyear9)
  (output omitted)
    _BM_varscoreLEVnN7 _BM_varscoreLEVnN8
Instruments for levels equation
  Standard
    tauyear3 tauyear4 tauyear5 tauyear6 tauyear7 tauyear8 tauyear9
  (output omitted)
    _cons
------------------------------------------------------------------------------
Arellano-Bond test for AR(1) in first differences: z =  -5.56  Pr > z =  0.000
Arellano-Bond test for AR(2) in first differences: z =  -0.27  Pr > z =  0.785
------------------------------------------------------------------------------
Sargan test of overid. restrictions: chi2(38)   =  57.54  Prob > chi2 =  0.022
  (Not robust, but not weakened by many instruments.)
Hansen test of overid. restrictions: chi2(38)   =  57.60  Prob > chi2 =  0.022
  (Robust, but weakened by many instruments.)
```

Even in this case, we observe a drop in the degrees of freedom of the Hansen test that fall from 100 in the standard SYS GMM estimates to 38 when the scores relative to the principal components are extracted according to the variance criterion.

To further clarify the syntax and the options, we provide additional examples for the creation of instrumental variables and scores.

The syntax

```
pca2 n w k, nt(id year) lagsl(n, ll(2)) lagsl(w k, ll(3))
```

creates PCA scores according to the variance criterion (90%) for each variable taken separately: the principal components are extracted both from the set of instruments in levels for `n`, which includes lags from $t-2$ up to the last lag available, and from the two sets of instruments in levels, respectively, for `w` and `k`, which include lags from $t-3$ up to the last lag available.

The syntax

```
pca2 n w k, nt(id year) gmmdiv(2) lagsl(n w k, ll(2 3))
```

applies the PCA on the sets of instruments in first-differences from $t-2$ up to the last lag available for each variable taken separately and on the set of instruments in levels from $t-2$ to $t-3$ for each variable.

The syntax

```
pca2 n w k, nt(id year) lagsd(n w k, ll(2)) gmmliv(2) togvar togld ///
  variance(80) avg
```

runs the PCA on the set of instruments that includes the lags of interest both in levels and in first-differences of all the variables taken together. The principal components are retained according to both the average and the variance (80%) criteria.

It is worth noticing that the syntax

```
pca2 n w k
```

pools all the observations and runs the PCA on the three-column matrix of `n`, `w`, and `k`. It retains the principal components according to the variance criterion. The difference with respect to the syntax `pca n w k` is that `pca2` also selects the principal components to be retained and computes the corresponding scores without the need of additional command lines.

4 pca2 at work: An application to the estimation of a fiscal policy rule

In this section, we illustrate more in detail the empirical implications and the operational advantages of the proposed procedure by applying it to the estimation of fiscal policy rules, as discussed in the article by Golinelli and Momigliano (2009)—GM henceforth. The authors assess the robustness of the estimates of a fiscal policy rule on a panel of 11 Eurozone countries over the post-Maastricht period (that is, 1994–2008) by using alternative model specifications and exploiting data from different sources (European Commission, International Monetary Fund, and Organisation for Economic Cooperation and Development [OECD]) and data vintages (latest available and real-time data). Of main interest here, GM run a number of alternative regressions to estimate the parameters of the rule by SYS GMM and provide extensive motivations for their choice, which, in this case, comes out as the most appropriate, in line with well-established indications in the literature.[6] However, we have stressed in previous sections that when the cross-sectional dimension is smaller than the time dimension, there is the risk of getting biased estimates in case of a high number of overidentifying restrictions. Because GM's dataset spans over $N = 11$ and $T = 15$, their analysis lacks robustness checks with respect to the number of orthogonality conditions exploited by the SYS GMM.

In this section, we estimate the discretionary policy rule reported in GM (2009, 45),

$$\Delta\text{CAPB}_{it} = \mu_i + \tau_t + \beta_1\text{GAP}_{i,t-1} + \beta_2\text{CAPB}_{i,t-1} + \beta_3\text{DEBT}_{i,t-1} + \varepsilon_{it} \qquad (7)$$

where the dependent variable is the change in the cyclically adjusted primary balance on potential gross domestic product measured with the latest available data (that is, the best measure over time of the fiscal policy stance). The explanatory variables are the output gap (GAP), which accounts for the economic cycle, the cyclically adjusted primary balance (CAPB), and the debt (DEBT) as ratios on potential gross domestic product, the last two capturing the fiscal initial conditions. The explanatory variables are specified in $t-1$ and measured with real-time data (that is, the information available at the time when the fiscal policy is set). Finally, μ_i are country fixed effects, τ_t are time fixed effects, and ε_{it} are random policy shocks assumed to be independent and identically distributed.

6. The empirical analysis exploits the user-written command `xtabond2`.

Table 1 reports estimates for (7) under alternative specifications of the overidentifying restrictions exploited by SYS GMM. First, we fit the model, including all the available lags (column 1); then, we reduce the instrument set by lag-truncation (columns 2 and 3) and collapse (column 4); finally, we exploit as new IVs the scores from PCIVR on the full set of instruments (column 5), on the truncated set (column 6), and on all lags of all the endogenous variables considered together (column 7).

In particular, the estimates reported in the first column follow the same approach as in GM and are obtained by instrumenting all the explanatory variables with all the available lags as in GM (2009, table 3, column 5, "OECD-HP").[7] These outcomes are in line with those of GM: overall, the authors interpret their evidence as indicating that the fiscal initial conditions do affect policy choices, while the counter cyclicality of fiscal policies comes out to be only slightly significant. It is worth noticing that in column 1, the number of overidentifying restrictions (152) is very close to the number of observations (165) and that the p-value of the Sargan test is almost 0.7: in light of Sargan's caveats and of the discussion of the previous sections, we argue that the outcome of the test for overidentifying restrictions may be weakened by the high instrument count with respect to the number of observations.

To assess the robustness of GM's findings, columns 2 through 7 of table 1 report estimates where the number of overidentifying restrictions is reduced by adopting alternative approaches.

Table 1. Estimation of a fiscal policy rule

Dependent variable: ΔCAPB								
Variable		(1) all lags	(2) lags 2–3	(3) lag 2	(4) collapse	(5) PCIVR all	(6) PCIVR lags 2–3	(7) PCIVR tog
L.CAPB	coeff	−0.317	−0.306	−0.303	−0.382	−0.318	−0.306	−0.334
	sd	0.058	0.06	0.064	0.089	0.062	0.067	0.08
	t	−5.43	−5.12	−4.72	−4.3	−5.13	−4.6	−4.15
L.DEBT	coeff	0.017	0.016	0.014	0.017	0.014	0.014	0.021
	sd	0.004	0.005	0.005	0.017	0.005	0.006	0.009
	t	3.91	3.41	2.76	0.96	2.73	2.58	2.38
L.GAP	coeff	0.146	0.131	0.128	0.238	0.16	0.097	0.21
	sd	0.096	0.099	0.107	0.137	0.1	0.114	0.146
	t	1.52	1.32	1.2	1.74	1.6	0.85	1.44
Constant	coeff	−0.757	−0.635	−0.532	−0.703	−0.496	−0.556	−1.03
	sd	0.454	0.46	0.485	1.221	0.478	0.512	0.694
	t	−1.67	−1.38	−1.1	−0.58	−1.04	−1.09	−1.48
$N \times T$		165	165	165	165	165	165	165
N		11	11	11	11	11	11	11
T		15	15	15	15	15	15	15
Sargan test:								
degrees of freedom		152	132	87	48	117	83	46
p-value		0.6928	0.6246	0.1612	0.2276	0.2691	0.2048	0.1451

7. These estimates do not perfectly match those in GM, because here the DEBT is not considered as strictly exogenous.

More precisely, in column 2, the lag depth of the instruments is truncated to include only the lags for $t-2$ and $t-3$; in column 3, only the lags for $t-2$ are considered. Finally, the estimates in column 4 exploit a collapsed instrument matrix. Not surprisingly, the p-values of the Sargan test decrease with the instrument count, but this drop is also associated with substantial changes in the overall picture that emerges from the original GM estimates and from those in column 1 of table 1. Strikingly, the two most frequently used approaches to reduce the instrument count contrast the GM findings in opposite directions. Indeed, the lag-truncation leads to estimates that strengthen the evidence of a-cyclicality of the fiscal policy because the GAP parameter is not significant; the collapse of the instrument matrix does not give significant coefficient estimates for the DEBT, while the GAP parameter reaches a significance level of 10% and appears somewhat supportive of the counter cyclicality of fiscal policies.

Such mixed evidence substantiates our concerns over the importance of introducing more compelling robustness checks in the cases when the instrument count is high and needs to be reduced.

To investigate the issue further, we see that the last three columns in table 1 report results for the estimating equation when the IVs count is reduced using the PCIVR strategy, as implemented through the `pca2` command. In column 5, the SYS GMM estimator exploits as IVs the PCA scores relative to the principal components of the matrix that includes all the available lags, retained according to the average criterion. This strategy is less parsimonious than the collapsing in terms of number of moment restrictions, and it provides results in line with those obtained on the whole instrument set. In column 6, the principal components to be retained according to the average criterion for the computation of the scores are extracted from the instruments matrix that includes only the lags relative to $t-2$ and $t-3$. Note also that the number of degrees of freedom of the Sargan test is larger than that obtained from the collapsing. To reduce the instrument count to a number in line with that of the collapse, in the estimates of column 7, we extract the principal components from both the instrument matrix that includes all the lags in levels of all the variables taken together and from that with all the IVs in first-differences.[8]

Overall, the empirical exercise performed in this section conveys important indications. First, we see that the estimates on the sets of instruments obtained through the `pca2` command are in line with the findings of columns 1–3, providing both a lower instrument count and a higher reliability of the Sargan test, which is consistently characterized by a lower p-value. The results corroborate the idea that the PCIVR, being a purely statistical way to tackle the issue of the excess of IVs, has the advantage of doing that without imposing heavy (and somewhat arbitrary) restrictions on the data structure. This feature emerges in particular from column 7, whose outcome closely mirrors the one in column 1 but is now obtained with a number of overidentifying restrictions that is only one-third the number of the latter.

8. This is done by specifying the `togvar` option for the `pca2` command. The data and the do-file with the commands to replicate table 1 are provided as complementary material.

Finally, with respect to the policy implications of the GM study, we have shown that estimates based on collapsed instruments would have changed the view over the determinants of the policy rules because the stock of debt is found as nonsignificant in contrast with the significant coefficient across all the other specifications. Thanks to the newly implemented `pca2` procedure, we have shown that this shift is not directly driven by the reduction in the number of IVs (also carried out by lag-truncation and PCIVR); rather, it is due to the restrictions imposed on the instrument matrix.

5 Summary

This article introduces a new command for creating GMM-style instruments for dynamic panel-data models, for running principal component analysis on these instruments, and for obtaining the PCA scores to be used as new instruments in GMM estimation. The command `pca2` adds important features to the Stata command `pca`: in particular, it allows for the selection of principal components to be retained according to alternative criteria and for the extraction of principal components from different sets of GMM-style instruments at the same time.

6 Acknowledgments

The authors are especially indebted to Roberto Golinelli and Matteo Lippi Bruni for their support and their suggestions. They also thank Roberto Golinelli and Sandro Momigliano for providing the data used in the empirical application and an anonymous reviewer for insightful comments.

7 References

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

Amemiya, T. 1966. On the use of principal components of independent variables in two-stage least-squares estimation. *International Economic Review* 7: 283–303.

Anderson, T. G., and B. E. Sørenson. 1996. GMM estimation of a stochastic volatility model: A Monte Carlo study. *Journal of Business and Economic Statistics* 14: 328–352.

Andrews, D. W. K. 1999. Consistent moment selection procedures for generalized method of moments estimation. *Econometrica* 67: 543–564.

Andrews, D. W. K., and B. Lu. 2001. Consistent model and moment selection procedures for GMM estimation with application to dynamic panel data models. *Journal of Econometrics* 101: 123–164.

Arellano, M., and S. Bond. 1991. Some tests of specification for panel data: Monte Carlo evidence and an application to employment equations. *Review of Economic Studies* 58: 277–297.

Arellano, M., and O. Bover. 1995. Another look at the instrumental variable estimation of error-components models. *Journal of Econometrics* 68: 29–51.

Bai, J., and S. Ng. 2008. Forecasting economic time series using targeted predictors. *Journal of Econometrics* 146: 304–317.

———. 2010. Instrumental variable estimation in a data rich environment. *Econometric Theory* 26: 1577–1606.

Bekker, P. A. 1994. Alternative approximations to the distributions of instrumental variable estimators. *Econometrica* 62: 657–681.

Belloni, A., D. Chen, V. Chernozhukov, and C. Hansen. 2012. Sparse models and methods for optimal instruments with an application to eminent domain. *Econometrica* 80: 2369–2429.

Blundell, R., and S. Bond. 1998. Initial conditions and moment restrictions in dynamic panel data models. *Journal of Econometrics* 87: 115–143.

Bound, J., D. A. Jaeger, and R. M. Baker. 1995. Problems with instrumental variables estimation when the correlation between the instruments and the endogeneous explanatory variable is weak. *Journal of the American Statistical Association* 90: 443–450.

Bowsher, C. G. 2002. On testing overidentifying restrictions in dynamic panel data models. *Economics Letters* 77: 211–220.

Caner, M. 2009. Lasso-type GMM estimator. *Econometric Theory* 25: 270–290.

Caner, M., E. Maasoumi, and J. A. Riquelme. 2014. Moment and IV selection approaches: A comparative simulation study.
http://econ.ohio-state.edu/caner/r3minvalid.pdf.

Dagenais, M. G., and D. L. Dagenais. 1997. Higher moment estimators for linear regression models with errors in the variables. *Journal of Econometrics* 76: 193–221.

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

Golinelli, R., and S. Momigliano. 2009. The cyclical reaction of fiscal policies in the Euro area: The role of modelling choices and data vintages. *Fiscal Studies* 30: 39–72.

Groen, J. J. J., and G. Kapetanios. 2009. Parsimonious estimation with many instruments. Staff Report 386, Federal Reserve Bank of New York.

Han, C., and P. C. B. Phillips. 2006. GMM with many moment conditions. *Econometrica* 74: 147–192.

Kapetanios, G., and M. Marcellino. 2010. Factor-GMM estimation with large sets of possibly weak instruments. *Computational Statistics and Data Analysis* 54: 2655–2675.

Kiviet, J. F. 1995. On bias, inconsistency, and efficiency of various estimators in dynamic panel data models. *Journal of Econometrics* 68: 53–78.

Kloek, T., and L. B. M. Mennes. 1960. Simultaneous equations estimation based on principal components of predetermined variables. *Econometrica* 28: 45–61.

Liao, Z. 2013. Adaptive GMM shrinkage estimation with consistent moment selection. *Econometric Theory* 29: 857–904.

Mander, A. 2006. lars: Stata module to perform least angle regression. Statistical Software Components S456860, Department of Economics, Boston College. https://ideas.repec.org/c/boc/bocode/s456860.html.

Mehrhoff, J. 2009. A solution to the problem of too many instruments in dynamic panel data GMM. Discussion Paper No. 1/2009, Deutsche Bundesbank.

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

Roodman, D. 2009a. A note on the theme of too many instruments. *Oxford Bulletin of Economics and Statistics* 71: 135–158.

———. 2009b. How to do xtabond2: An introduction to difference and system GMM in Stata. *Stata Journal* 9: 86–136.

Sargan, J. D. 1958. The estimation of economic relationships using instrumental variables. *Econometrica* 26: 393–415.

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

Stock, J. H., and M. W. Watson. 2002. Forecasting using principal components from a large number of predictors. *Journal of the American Statistical Association* 97: 1167–1179.

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* 14: 419–431.

About the authors

Maria Elena Bontempi is an associate professor in the Department of Economics at the University of Bologna, Italy.

Irene Mammi is a postdoctoral research fellow in the Department of Economics at the University of Bologna, Italy.