2011s-57

Partially Dimension-Reduced Regressions with Potentially Infinite-Dimensional Processes

John W. Galbraith, Victoria Zinde-Walsh

Série Scientifique
Scientific Series

Montréal
Septembre 2011

CIRANO
Allier savoir et décision

Centre interuniversitaire de recherche en analyse des organisations

CIRANO

Le CIRANO est un organisme sans but lucratif constitué en vertu de la Loi des compagnies du Québec. Le financement de son infrastructure et de ses activités de recherche provient des cotisations de ses organisations-membres, d'une subvention d'infrastructure du Ministère du Développement économique et régional et de la Recherche, de même que des subventions et mandats obtenus par ses équipes de recherche.

CIRANO is a private non-profit organization incorporated under the Québec Companies Act. Its infrastructure and research activities are funded through fees paid by member organizations, an infrastructure grant from the Ministère du Développement économique et régional et de la Recherche, and grants and research mandates obtained by its research teams.

Les partenaires du CIRANO

Partenaire majeur
Ministère du Développement économique,
de l'Innovation et de l'Exportation

Partenaires corporatifs
Autorité des marchés financiers
Banque de développement du Canada
Banque du Canada
Banque Laurentienne du Canada
Banque Nationale du Canada
Banque Royale du Canada
Banque Scotia
Bell Canada
BMO Groupe financier
Caisse de dépôt et placement du Québec
CSST
Fédération des caisses Desjardins du Québec
Financière Sun Life, Québec
Gaz Métro
Hydro-Québec
Industrie Canada
Investissements PSP
Ministère des Finances du Québec
Power Corporation du Canada
Rio Tinto Alcan
State Street Global Advisors
Transat A.T.
Ville de Montréal

Partenaires universitaires
École Polytechnique de Montréal
HEC Montréal
McGill University
Université Concordia
Université de Montréal
Université de Sherbrooke
Université du Québec
Université du Québec à Montréal
Université Laval

Le CIRANO collabore avec de nombreux centres et chaires de recherche universitaires dont on peut consulter la liste sur son site web.

Les cahiers de la série scientifique (CS) visent à rendre accessibles des résultats de recherche effectuée au CIRANO afin de susciter échanges et commentaires. Ces cahiers sont écrits dans le style des publications scientifiques. Les idées et les opinions émises sont sous l'unique responsabilité des auteurs et ne représentent pas nécessairement les positions du CIRANO ou de ses partenaires.
This paper presents research carried out at CIRANO and aims at encouraging discussion and comment. The observations and viewpoints expressed are the sole responsibility of the authors. They do not necessarily represent positions of CIRANO or its partners.

ISSN 1198-8177

Partenaire financier
Développement
économique, Innovation
et Exportation
Québec

Partially Dimension-Reduced Regressions with Potentially Infinite-Dimensional Processes [*]

John W. Galbraith[†], *Victoria Zinde-Walsh*[‡]

Résumé/Abstract

Regression models sometimes contain a linear parametric part and a part obtained by reducing the dimension of a larger set of data. This paper considers properties of estimates of the interpretable parameters of the model, in a general setting in which a potentially unbounded set of other variables may be relevant, and where the number of included factors or components representing these variables can also grow without bound as sample size increases. We show that consistent (and asymptotically normal, given further restrictions) estimation of a parameter of interest is possible in this setting. We examine selection of the particular orthogonal directions, using a criterion which takes into account both the magnitude of the eigenvalue and the correlation of the eigenvector with the variable of interest. Simulation experiments show that an implementation of this method may have good finite-sample performance.

Mots clés *:* Dimension reduction, eigenvector, infinite-dimensional process, orthogonalized regressors

[*] We have received valuable comments from many colleagues and from participants in conferences of the Econometric Society and CIRANO. We thank the Fonds québécois de la recherche sur la société et la culture (FQRSC) and the Social Sciences and Humanities Research Council of Canada (SSHRC) for financial support of this research, and the Centre Interuniversitaire de recherche en analyse des organisations (CIRANO) and Centre Interuniversitaire de recherche en économie quantitative (CIREQ) for research facilities.

[†] Department of Economics, McGill University, 855 Sherbrooke St. West, Montreal, Quebec, H3A 2T7 Canada, john.galbraith@mcgill.ca

[‡] Department of Economics, McGill University, 855 Sherbrooke St. West, Montreal, Quebec, H3A 2T7 Canada.

1. Introduction

Regression models sometimes contain a linear parametric part and a part obtained by reducing the dimension of a larger set of data; for example, factor-augmented regression is widely used for forecasting (see for example Stock and Watson 2002a,b). In many cases the fits or forecasts, rather than parameter values, are the primary objects of interest. In other contexts the values of a few interpretable parameters may be the focus of attention, while other variables or factors are included as statistical controls for other effects which, if omitted, would lead to poor inference on the parameters of interest (e.g. Magnus and Durbin 1999).

In a linear context, we can represent this situation as

$$y = c + X\beta + Z\gamma + \epsilon, \tag{1.1}$$

where our interest is in β, but where Z has elements which may be correlated with elements of X, and where Z may be of potentially large, even infinite, dimension. The researcher typically chooses a subset of Z to include as controls in the regression; however, the researcher will often be unable to identify the most important elements of Z to include, so that information remains in the neglected elements. It is well known that unless a selection of regressors is sufficient for all information in the set of controls relative to the parameter of interest, estimates from the resulting model are strictly inconsistent, and the omitted variable bias may be substantial. The very widespread use of the linear regression model where the true process is not known *a priori* makes this problem one of great practical importance.

This paper examines the asymptotics of regression to obtain information from a large set of control regressors, where there is a relatively small number of parameters of interest. It is closely related to two existing literatures: one in which parameters of a model are separated into two classes based on *a priori* considerations of importance to the investigator, and another in which the order of a process is allowed to be unbounded in order to establish asymptotic results without finite-order restrictions. We will treat a generic problem related to the first literature, and for generality will use an infinite-dimensional setting of the type seen in the second.

A number of different econometric models, including factor-augmented models and semi-parametric (partially linear) models, make a distinction between a small number of parameters of particular interest and others that are necessary as controls. In semi-parametric models (e.g. Robinson 1988), a dependent variable is allowed to be a general function of a small number of regressors, and to depend linearly on a further set. The distinction is made in cases in which the number of regressors is too large for a nonparametric regression on the full set, and nonparametric modelling of a subset is deemed to be of particular interest; other regressors are nonetheless retained as controls with linear effects. Although the present paper deals with a purely linear context (a restriction which can be relaxed), we treat cases where the set of potential explanatory variables may be large enough that including all of them would be impossible or impractical because of efficiency loss.

Another context in which a distinction is made between classes of parameters, more closely related to the models examined here, is that of factor-augmented regression models. Here (see in particular Bai and Ng 2006), a finite factor structure is usually assumed. A factor model expresses an $m-$ dimensional array in terms of a smaller number q of factors, that is, $z_{it} = \Lambda F_t + \epsilon_{it},\ i = 1, \dots m,\ t = 1, \dots \dots n$, where F_t is a $q \times 1$ set of mutually orthogonal factors, and where ϵ_{it} is usually taken to be orthogonal to $\epsilon_{js}, i \neq j,\ s \neq t$, although this assumption is relaxed in, e.g., Chamberlain (1983), Forni and Lippi (2001). The q factors are chosen to provide a parsimonious characterization of Z, so that much of the information in Z may be extracted. Chamberlain and Rothschild (1983) treat the finite factor structure as an approximation, as an alternative to allowing the number of factors to increase without bound, but do not pursue the latter. While these models have typically been used in forecasting applications, some authors have used the extracted factors to provide controls for estimation of a small number of parameters of interest; see for example Magnus et al. (2009).

There is also an increasingly substantial literature which treats processes of interest as being of potentially unbounded order. It is well known that many processes can be represented as infinite-order linear series; in time series econometric contexts, for example, these series may have an AR(∞), MA(∞) or ARCH(∞) form. Finite-order models used in practice can be treated as truncations of such processes, and there are asymptotic results on consistent LS estimation of such truncations, with model order increasing at an appropriate rate, dating back to the classic work of Berk (1974). These results have been extended to multivariate cases including that of vector autoregressions. A more recent literature extends results of this type to quantile estimation; see Belloni and Chernozhukov (2009), Zernov et al. (2009). The general setting has the advantage that it allows for the case, of practical importance in econometrics, in which models are truncations of more general models that we would prefer to estimate if sample sizes and data sets permitted. Since in practice we are restricted by limitations on sample size and observed series, it is interesting to investigate the consequences of these limitations and of gradually increasing model orders as the constraints are loosened.

The present study draws on both literatures. In extracting information from a large set of explanatory series, we consider computations similar to those employed in factor-augmented models; in particular, we derive some smaller number of regression directions from eigenvectors of a matrix of potential regressors. Principal components are one example of such an extracted set of regression directions, although we suggest a modified method. Dimension reduction methods are applied to a part of the model; the resulting regressors are used as statistical controls for another part in which interpretable parameters are required. We give conditions under which this allows consistent estimation of these interpretable parameters. That is, in the model (1.1) we use a parsimonious characterization of Z via a smaller number of components. However, effects of Z are not the focus of our interest, and we do not require a finite dimension for Z nor any finite bound for the number of orthogonal components derived from Z which can be used as controls. We are able to prove consistency of estimation of an effect of interest β under more general circumstances than in existing literature, and because we do not assume a finite factor structure for the control regressors, the results allow the researcher to define parameters of interest without the constraint that the remaining regressors meet a finite-factor condition.

The fact that the process is treated as having unknown and potentially unbounded dimension is a key technical element of this study. Allowing for unbounded dimension of the process requires more elaborate methods of proof; however we are able to show that, applied with an appropriate algorithm for augmenting the set of regressors as $N \to \infty$, linear regression is consistent for parameters of interest in a more general context than has previously been established. Specifically, the asymptotic results examine the possibility of consistent estimation of β when the number of included directions from the space spanned by Z grows with sample size. We establish asymptotic theory for the estimates in finite models whose dimension increases with sample size. We then consider a new criterion for selection of directions which orders orthogonal directions in the space spanned by Z by magnitude of the product of the eigenvalue and correlation with X; this implies that the importance for estimation of β of directions tends to decline with diminishing value of the criterion. We show that the dimension of the regressor space can be reduced by excluding Z's with the lowest values of the criterion, without affecting consistency of estimates of β, and that there is a uniform upper bound for a given sample size on the number of directions that need be included.

In section 2 we provide a formal definition of the problem and conditions for asymptotic results describing consistent and asymptotically normal estimation of a finite part of an infinite-dimensional process. The other main results of the paper are in Section 3, which describes orthogonalization of the regressors and shows that a criterion for selecting among the orthogonalized directions allows consistent estimation of β in a model of reduced dimension. Together the results of these sections establish consistency of the regression method for the parameter of interest, using the regressor selection algorithm to augment model order at a controlled rate, in a general problem of unknown order. Section 4 provides simulation evidence on the finite-sample behaviour in such models. Proofs are given in the Appendix.

2. Processes, notation and preliminary results

We begin with a schematic outline of the class of methods to be considered, for the linear model (1.1):

$$y = c + X\beta + Z\gamma + \epsilon$$

with a potentially large number of regressors contained in Z and a (typically smaller) set of variables of interest X, with a corresponding vector of parameters of interest β. A partially dimension-reduced regression method allows extraction of information from a matrix of potential regressors which may be too large to allow individual inclusion of each. The following is one implementation of such a method:

- 1. From the matrix of data Z, compute the eigenvalues and corresponding eigenvectors of the moment matrix $Z'Z$.

- 2. Order the eigenvectors by the product of the eigenvalue and the correlation between the eigenvector and the regressor of interest, x (note that if we ordered by eigenvalue alone, we would be extracting principal components in the next step).
- 3. For a given number κ of components, compute corresponding orthogonal regressors by the product of the matrix of κ eigenvectors and the original matrix of potential regressors. Repeat for a range of values of κ and select the value of κ that yields the lowest Akaike Information Criterion (AIC).
- 4. The final regression model uses the regressor of interest x augmented by κ orthogonal components as control regressors.

Alternative implementations might use a different information criterion, or model averaging to combine components; the focus in the present study is not on comparison of these possible devices, but on consistent estimation via model order which increases at a controlled rate.

As an example, consider a macroeconomic data set comprising several hundred variables but a smaller number, e.g. 200, time series observations. We are interested in estimating the effect on percentage change in industrial production, y, of a change in a short term interest rate, x. Industrial production is potentially affected by a large number of the measured variables, but we cannot include hundreds of these quantities, and lags, in a regression; nonetheless omission of the relevant quantities will lead to inconsistent estimates of the effect of x on y. By retaining the interest rate x and reducing the dimension of the set of other variables appropriately, we can nonetheless obtain good estimates of the effect of interest. As well, the estimates are not dependent on the usual judgments as to which other variables to include in a regression.

Below we derive properties of the general procedure and show that it allows consistent estimation of interpretable effects of interest without *a priori* knowledge of which regression directions to include. There are two classes of technical problem to solve. First, because we do not require Z to be of finite dimension, we must establish consistency of the regression procedure with a truncation of Z that increases in order with sample size (of course, there is a substantial simplification if one is willing to assume finite dimension for the true process). Second, we will apply dimension reduction methods to the finite truncation, and establish properties of the full procedure. The remainder of section 2 addresses the first point, and section 3 the second point.

In the following sections we establish a more precise notation as well as proving the necessary theorems.

2.1 *Processes and notation*

We will first specify conditions on the process and define the parameter of interest in a linear conditional expectation function. It is important to note that the conditions are specified in a general form that allows either finite- or infinite-dimensional processes. In particular, we assume that the observed data are associated with realizations of a multi-dimensional (possibly infinite-dimensional) stationary random process, i.e. the assumptions below allow the number of random variables in the process, indexed by ℓ, to be unbounded:

$\ell = 1, \dots \infty$. A finite dimension, with $\ell = 1, \dots L < \infty$, is of course also covered by the theorems below.

Consider a set of real-valued random variables $W = \{w_{\ell i}\}_{\ell=1, i=-\infty}^{\infty\ \ \infty}$ such that $W_\ell = \{w_{\ell i}\}_{i=-\infty}^{\infty}$ is a stochastic process for each ℓ, where i indexes the N observations. For each observation $W_{\cdot i} = \{w_{\ell i}\}_{\ell=1}^{\infty}$ on the set of random variables, define random vectors (partitions) $W_i' = (y_i; X_i'; Z_i')$, where $y_i = w_{1i}$ represents a dependent variable, $X_i' = (w_{2i}, \dots w_{m+1,i}) = (x_{1i}, \dots x_{m,i})$ a set of m conditioning variables of interest, and $Z_i' = [z_{1i}, z_{2i}, \dots] = [w_{m+2,i}, w_{m+3,i}, \dots]$ a vector of additional conditioning variables; Z_i' may or may not be of finite dimension. This data generation process is assumed to satisfy the following conditions, a number of which can evidently be weakened.

Assumption 1 (A1)

. *(i) W_ℓ is a stationary stochastic process for every ℓ :*
$E(w_{\ell i}) = \mu_\ell,\ cov(w_{\ell i} w_{\ell j}) = \phi_{\ell\ell}(|i-j|)$;

. *(ii) W_{ℓ_1}, W_{ℓ_2} are co-stationary: $cov(w_{\ell_1 i} w_{\ell_2 j}) = \phi_{\ell_1 \ell_2}(|i-j|)$;*

. *(iii) There exists an increasing sequence of $\sigma-$ fields $\{\mathcal{F}_i\}_{-\infty}^{\infty}$ such that X_i', Z_i' are measurable with respect to $\mathcal{F}_i$ and*

$$E(y_i|\mathcal{F}_i) = c + X_i'\beta + Z_i'\gamma; \tag{2.1}$$

. *(iv) The lowest and highest eigenvalues, $\underline{\lambda}(\Sigma_W)$ and $\overline{\lambda}(\Sigma_W)$, of the covariance matrix Σ_W of $W_{\cdot i}$ (or of any subset) are such that*

$$0 < \underline{\zeta} < \underline{\lambda}(\Sigma_W) < \overline{\lambda}(\Sigma_W) < \overline{\zeta} < \infty.$$

. *(v)* $\sup_{1 \le \ell \le \infty} E(w_{\ell i}^4) < \infty$.

From Assumption 1 (i),(ii), the $w_{\ell i}$ span a separable Hilbert space $\mathcal{H}$ with the scalar product given by $< w_{\ell_1 i}, w_{\ell_2 j} >=$ cov$(w_{\ell_1 i} w_{\ell_2 j})$, and there is a Wold representation for each W_ℓ. Equation (2.1) gives the conditional expectation function, in which β is the key object of interest. Note also that part (iii) of A1 implies that

$$(c, \beta, \gamma) = \arg\min_{\tilde{c}, \tilde{\beta}, \tilde{\gamma}} E(y_i - \tilde{c} - X_i'\tilde{\beta} - Z_i'\tilde{\gamma})^2.$$

Define $\varepsilon_i = y_i - (c + X_i'\beta + Z_i'\gamma),\ i = 1, \dots, N$; then $E(\varepsilon_i) = E(\varepsilon_i|\mathcal{F}_i) = 0$. Part (iv) of A1 implies that none of the regressors (in X or Z) is in the span of the others, that the inverse of the covariance matrix has a bounded norm, and that the corresponding coefficient is therefore identified. As well, for any non-stochastic $\overline{m} \times \infty$ matrix A of rank $\overline{m}$, $E(AWW'A') = A\Sigma A'$ is of rank $\overline{m}$, since by (iv), Σ_W is invertible. Note also that (v) implies $\sup E(w_{\ell i}^2) < (\sup E(w_{\ell i}^4))^{\frac{1}{2}} < \infty$ by Jensen's inequality.

Two special cases of the above structure are (i) $\mathcal{F}_i = \mathcal{F}_j = \mathcal{F}$ and all observations $W_{\cdot i}$ are independent, in which case $\operatorname{cov}(w_{\ell_1 i} w_{\ell_2 j}) = \gamma_{\ell_1 \ell_2}(|i-j|)$ if $i = j$, zero otherwise; and (ii) the case in which some W_ℓ are lagged values of others, so that for example Z_i' could include y_{i-h} and elements of Z_{i-h}', for several lags $h,\ h > 0$. The former case may be an adequate characterization of cross-sectional contexts, whereas the latter may arise in time series models.

2.2 *Preliminary results*

We now show that the conditions of Assumption 1 imply a bound on the sum of squared coefficients on the Z_i, and that the error $\{\varepsilon_i\}$ is a martingale difference sequence (m.d.s.) with respect to the sequence $\mathcal{F}_i$.

Lemma 1. If A1 is satisfied, then

- (i) $\sum_\ell (\gamma_\ell^2) < \infty$
- (ii)$\{\varepsilon_i, \mathcal{F}_i\}$ is a m.d.s.

Proof: see Appendix 1.

The parameter of interest in this problem is β, which can be estimated consistently without controlling for Z only if $X \perp Z$ or $\gamma = 0$. If the dimension of Z_i' is small, then estimation of the linear relationship $y_i = c + X_i\beta + Z_i'\gamma + \varepsilon_i$ is straightforward. In many problems of interest, however, $\dim(Z_i')$ is large relative to the available sample size (so that incorporating all elements of Z_i' is not practical). In such cases, the investigator is normally forced to choose a subset of the Z's, losing information in the other Z's which is orthogonal to the space spanned by this subset. The alternative examined here is to use methods that allow us to summarize the information in a large number of Z's with a smaller number of constructed regressors, which are used in place of the Z's to control for their effects.

A crucial question can now be formulated. For any k, the model (2.1) can be represented in the form

$$y = c + X\beta + Z(k)\gamma(k) + Z(k+1,\infty)\gamma(k+1,\infty) + \varepsilon_i, \tag{2.2}$$

where $Z(k)$ contains k regressors from Z and $Z(k+1,\infty)$ those remaining, possibly infinite in number. When can we obtain consistent estimates and asymptotically valid inference while ignoring some parts of Z; that is, when by increasing k appropriately as $N \rightarrow \infty$ does the contribution of $Z(k+1,\infty)$ decrease to zero?

Consider an additional assumption:[1]

Assumption 2 (A2)

$$\sum_\ell |\gamma_\ell| < \infty.$$

[1] This assumption implies that the regression coefficients on Z are themselves bounded in ℓ_1; this differs from the approach taken in the Lasso (see e.g. Tibshirani 1996, eq. 1), where an ℓ_1 penalty term is added to the LS criterion, with the effect of eliminating variables entirely from the specification.

Theorem 1 then provides an answer to the question.

Theorem 1. Under A1 and A2, as $k \to \infty$, $E(Z(k+1,\infty)\gamma(k+1,\infty))^2 \to 0$, and $Z(k+1,\infty)\gamma(k+1,\infty)\xrightarrow{p}0$.
Proof: see Appendix 1.

Note that A2 is sufficient, but not necessary. Moreover there are important examples that can easily be shown to satisfy A2, some of which we will describe below.

Assumptions A1 and A2 are much weaker than those required for consistent estimation in the standard context of regression with control variables, which embody not only a finite model but also that any omitted elements of the process are uncorrelated with the variable of interest. Nonetheless it is instructive to consider specific classes of cases for which these higher level assumptions are known to be satisfied.

A particularly important, because widely applicable, case is that in which L is finite but such that its magnitude cannot be established *a priori*, or the fact that L is finite may be unknown. In this case the condition that $\sum_{\ell=1}^{L} |\gamma_\ell| < \infty$ is trivially satisfied, and for large enough k, all relevant regression directions will be included. Thus, even with finite L, the assumptions remain substantially weaker than those normally imposed in using a regression model to control for nuisance effects: we are not required to know L (nor whether it is finite or not), although there is of course a sacrifice in efficiency relative to the case where L is known. Another easily verifiable example involves non-finite L, where Z is an expansion. Consider IID data related by $y = c + X\beta + Z\gamma + \epsilon$, where $Z\gamma$ represents an expansion such as a polynomial expansion of a function of a finite number of variables. For example, let $Z\gamma = g(z) = (1-qz)^{-1} = 1 + qz + q^2z^2 + \ldots$, with $|q| < 1$; in this simple example, if the powers of z satisfy A1 (e.g. if z is a bounded variable: $|z| < 1$), it follows that A2 holds. Finally, consider an unbounded number of lags in a time series process. Let y, X be linear processes and let each Z_ℓ, $\ell = 1, \ldots \overline{m}$ be a stationary and invertible ARMA process. Although $\overline{m}$ is finite, an unbounded number of lags of each Z_ℓ may be included in the model, leading to unbounded L. The process is $y = c + \sum_{i=1}^{m} \beta_i X_{it} + \sum_{i=1}^{m} \sum_{\nu=1}^{\infty} \gamma_{i\nu} Z_{i,t-\nu} + \epsilon_t$. It follows that $\sum_{i=1}^{m} \sum_{\nu=1}^{\infty} |\gamma_{i\nu}| < \infty$, i.e. A2 holds.

In practice, our assumptions allow us to treat cases of arbitrarily large numbers of potential regressors, or cases in which the number of potential regressors can increase without bound as measurements permit; an increasing number of data sets involving such large numbers of potential regressors are available. The well known National Longitudinal Survey of Youth (NLSY) data, for example, contain various component surveys, typically comprising hundreds of responses to questions.[2] Macroeconomic data sets of high dimension are also common; Stock and Watson (2002b), for example, give a detailed description of a data set containing 215 series, apart from lags. However, the purpose of that study is forecasting using extracted factors; macroeconomic studies to estimate parameters of interest almost invariably involve prior selection by the investigator of a much smaller set of regressors, rather than the use of dimension reduction methods.

[2] See http://www.bls.gov/nls .

2.3 *Consistency and asymptotic normality*

We now provide theorems demonstrating consistency and asymptotic normality of the estimate of the parameter of interest when only a finite part of the regressor space spanned by $Z(k)$ is included, and that of $Z(k+1,\infty)$ is ignored, as $k \to \infty$. Theorem 2 establishes consistency as $k \to \infty$ for estimates in the finite truncation of (2.2),

$$y = c + X\beta + Z(k)\gamma(k) + \epsilon_i, \tag{2.3}$$

which omits $Z(k+1,\infty)$. Let $M_{Z(k)}$ denote projection orthogonally to $Z(k)$: $M_{Z(k)} = I - Z(k)[Z(k)'Z(k)]^{-1}Z(k)'$.

Theorem 2. Suppose that A1 and A2 hold. Then if $N \to \infty, k \to \infty$, and $kN^{-\frac{1}{2}} \to 0$, the OLS estimator $\hat{\beta}_k = (X'M_{Z(k)}X)^{-1}X'M_{Z(k)}y$ in (2.3) is consistent: $\hat{\beta}_k \xrightarrow{p} \beta$.

Proof: See Appendix 1.[3]

Below, after introducing dimension reduction by a criterion derived in Section 3, we provide a further consistency result for the reduced-dimension regression model (Theorem 5).

It follows from the proof of Theorem 2 that $(\hat{\beta}_k - \beta) = O_p(f(k)) + O_p((N - \tilde{k})^{-\frac{1}{2}})$, where $f(k) = \sum_{i=1}^{\infty} |\gamma_{k+i}|$; for finite L, $(\hat{\beta}_k - \beta) = O_p(N^{-\frac{1}{2}})$.

Thus, depending on the rate of decay in the coefficients γ_ℓ, we may get either standard parametric or slower convergence rates. In particular, if $\gamma_\ell = O(\ell^{-\nu})$ with $\nu > 1$ (polynomial rate of decay), then we have that $\sum_{k+1}^{\infty} |\gamma_\ell| = O(k^{-\nu+1})$, which even for $k \simeq N^{\frac{1}{2}}$ provides the rate

$$(\hat{\beta}_k - \beta) = O_p(N^{-\frac{1-\nu}{2}}),$$

which is always slower than the parametric rate. If by contrast $\gamma_\ell = O(\alpha^\ell)$ with $\alpha < 1$ (exponential rate of decay), then $\sum_{\ell=k+1}^{\infty} \gamma_\ell = O(\alpha^{k+1})$ and for any $k = O(N^\gamma),\ \gamma < \frac{1}{2}$, the polynomial power dominates and a parametric rate obtains: that is, $(\hat{\beta}_k - \beta) = O_p(N^{-\frac{1}{2}})$. The latter is the usual case when the number of regressors is assumed to be finite, and is the case examined in, for example, the principal components literature, where exponential decay follows trivially from the fact that for large enough $\ell > L,\ \gamma_\ell = 0$.

It is also possible to show that the estimator is asymptotically normal, and so may be used for inference on β in the usual way. Theorem 3 gives this result in the case where $f(k)$, the absolute sum of coefficients on the omitted remainder terms, goes to zero sufficiently quickly. For finite L, the condition on $f(k)$ is always satisfied.

For the following theorem we define $\tilde{k}$ as the number of sample points lost to lags.

[3] In fact it can be seen from the proof that a stronger result, that $\hat{\beta}_k$ converges to β in the L_2 norm, also holds.

Theorem 3. Suppose that A1 and A2 hold, that $k \to \infty$ as $N \to \infty$, $kN^{-\frac{1}{2}} \to 0$, and also that k can be chosen such that $f(k) = o(N^{-\frac{1}{2}})$. Then

$$(N - \tilde{k})^{\frac{1}{2}} V_k^{-\frac{1}{2}} G_k(\hat{\beta}_k - \beta) \xrightarrow{D} N(0, I_m),$$

where $G_k = E\left(\frac{1}{N-\tilde{k}} X' M_k X\right)$ and $V_k = E(\frac{1}{N-\tilde{k}} X' M_k \varepsilon \varepsilon' M_k X)$. If ε is independent of (X, Z) then $G_k^{-1} V_k G_k^{-1} = \sigma_\varepsilon^2 E\left(\frac{X' M_k X}{N-\tilde{k}}\right)$.

Proof: See Appendix 1.

The weighting matrix $H = G_k^{-1} V_k G_k^{-1}$ can be estimated consistently by $\hat{\sigma}_\varepsilon^2 \left(\frac{X' M_k X}{N-\tilde{k}}\right)$, where $\hat{\sigma}_\varepsilon^2 = (N - \tilde{k})^{-1} u'u$ can be shown to be a consistent estimator of σ_ε^2, where u is the residual vector from the regression (2.3) on X and $Z(k)$. An operational test of $H_0 : \beta = \beta_0$ is then given by $(\hat{\beta}_k - \beta_0)' \hat{H} (\hat{\beta}_k - \beta_0) \xrightarrow{D} \chi_m^2$.

We have established in this section that we can obtain consistent and asymptotically normal estimates of the parameter of interest from a sequence of finite models of the potentially infinite-dimensional process, given appropriate conditions, especially on rates of growth of the number of regressors.

3. Dimension reduction and estimation of parameters of interest

We now turn to dimension reduction for the finite part $Z(k)$ in the model (2.3), to exploit as much information as possible, where k is too large to allow estimation of all k parameters. It will now be necessary to distinguish two column dimensions related to Z: therefore rather than using k, we will use K for the full column dimension of Z, and κ ($\kappa \leq K$) for the column dimension of a set of included components, which are linear transformations of Z. With this distinction, we will establish properties of a distance measure useful in selecting particular controls on a finite sample; we also show that the selection rule that is derived ensures consistency of the estimator based on $\kappa (< K)$ components.

3.1 *Estimation by regression on orthogonal components*

Given a finite sample of size N, we use models of finite dimension despite the possibly-infinite dimension of the vector Z_i' which enters the true process. Where Z is not of finite column dimension, we treat a finite number K of included elements of Z, such that K may increase with N: i.e. K is the number of data series used as potential controls. Define $Z(K)$ and $Z(K+1, \infty)$ as the included and excluded parts of Z respectively, and use the partition $Z_i' = [Z_i(K)' : Z_i(K+1, \infty)']$ (the corresponding parameter vector γ, with individual element γ_j, is partitioned conformably so that $\gamma' = [\gamma(K)' : \gamma(K+1, \infty)']$).

We treat here the case in which $K \leq N$, i.e. fewer included potential explanatory series than data points. We first compute sets of orthogonalized vectors which span the same space as $Z(K)$.

Define a $K \times K$ matrix $C(K)$ such that $C(K)'Z(K)'Z(K)C(K) = \overline{\Lambda}$, where $\overline{\Lambda}$ is the $K \times K$ matrix with the K eigenvalues $(\lambda_\ell, \ell = 1, \ldots K)$ of $Z(K)'Z(K)$ on the main diagonal, zeroes elsewhere. That is, the columns of $C(K)$ contain the K eigenvectors of $Z(K)'Z(K)$, and $C(K)'C(K) = C(K)C(K)' = I$; $C(K)$ is therefore a random matrix, which depends on the sample.[4] Next define a selection matrix $\Pi_{K \times \kappa}$ such that $C(\kappa) = C(K)\Pi$ is a $K \times \kappa$ matrix which contains κ of the K eigenvectors: κ will be the number of control regressors included in the model (we discuss the choice of κ below).[5]

Finally define the auxiliary model regressors $S(\kappa, K)_{N \times \kappa} = Z(K)C(\kappa)$ and also $S(K, K)_{N \times K} = Z(K)C(K)$, which uses the full set of eigenvectors; $S(K, K)$ contains all principal components of, and spans the same space as, $Z(K)$. From the representation (2.1) of the process, we can write

$$\begin{aligned}
y_i &= c + X_i'\beta + Z_i'\gamma + \varepsilon_i \\
&= c + X_i'\beta + Z_i'(K)\gamma(K) + Z_i(K+1, \infty)'\gamma(K+1, \infty) + \varepsilon_i \\
&= c + X_i'\beta + S_i(K, K)'\delta(K) + Z_i(K+1, \infty)'\gamma(K+1, \infty) + \varepsilon_i \\
&= c + X_i'\beta + S_i(\kappa, K)'\delta(\kappa) + S_i(K-\kappa, K)'\delta(K-\kappa) \\
&\qquad + Z_i(K+1, \infty)'\gamma(K+1, \infty) + \varepsilon_i, \\
&\equiv c + X_i'\beta + S_i(\kappa, K)'\delta(\kappa) + R_i'(\kappa, \infty)\theta(\kappa, \infty) + \varepsilon_i,
\end{aligned} \tag{3.1}$$

where $S(K-\kappa, K)$ is the $N \times (K-\kappa)$ matrix containing the $(K-\kappa)$ columns of $S(K, K)$ not present in $S(\kappa, K)$, and R collects all of the conditioning variables $R \equiv [S(K-\kappa, K) : Z(K+1, \infty)]$ not present in $S(\kappa, K)$. Note that $S(K, K)\delta(K) = Z(K)\gamma(K)$ so that $C(K)\delta(K) = \gamma(K)$, and that $\theta'(\kappa, \infty)$ is defined as the vector $[\delta'(K-\kappa) : \gamma'(K+1, \infty)]$. Note also that $S(K, K)$ is a sample-dependent transformation, so that only the first two lines of (3.1) characterize the process itself.

Estimation of β is based on the auxiliary model–a reduction of the data generation process–

$$y_i = c + X_i'\beta^* + S_i(\kappa, K)'\delta^* + e_i, \tag{3.2}$$

which uses the subset $S(\kappa, K)$ of the available orthogonalized regressors contained in $S(K, K)$. If the estimation method is OLS, we refer to the estimator based on (3.2) as OLS(κ, K), indicating that κ orthogonalized regressors are used from the set of K explanatory variables (i.e. $\hat{\beta}(\kappa) = (X'M_\kappa X)^{-1} X'M_\kappa y$, where projection orthogonally to $S(\kappa, K)$ is defined by $M_\kappa = I - S(\kappa, K)(S(\kappa, K)'S(\kappa, K))^{-1}S(\kappa, K)'$ with $S(\kappa, K) = [S_1, \ldots, S_\kappa]$.)

[4] For simplicity of notation this dependence is not explicitly indicated.

[5] Each column of Π will have one element equal to one, all others zero, with no repeated columns; i.e. if $\Pi_{ij} = 1$, then $\Pi_{i'j} = 0 \ \forall i' \neq i$, and $\Pi_{ij'} = 0 \ \forall j' \neq j$.

Different methods of selection of the elements (columns) of $S(\kappa, K)$ from those of $S(K, K)$ are of course possible. If Π selects the eigenvectors corresponding with the κ largest eigenvalues, then $S(\kappa, K)$ contains the first κ principal components of $Z(K)$. An alternative, where we take X to be a single vector ($m = 1$), is to choose the κ eigenvectors of $Z(K)'Z(K)$ corresponding with the largest values of $\{\lambda_\ell \cdot \text{ corr}(S(K, K)_\ell, X)\}$; that is, large eigenvalues are given more weight if they correspond with eigenvectors that are highly correlated with X. In that case we choose $C(\kappa)$ as the matrix containing the set of κ eigenvectors with the highest values of this criterion, rather than the eigenvectors corresponding to the κ largest eigenvalues. We will examine these selection methods below.

The auxiliary model (3.2) has the advantage that it does not require knowledge of which elements of $Z(K)$ are important in explaining y, and therefore facilitates approximately unbiased estimation of β where the data generation process is unknown (as when some elements of γ are zero), and the column dimension of Z or of $Z(K)$ is too large to allow estimation of parameters on the full set of potential explanatory factors, given the available sample. This is a generic form of problem occurring in both cross-sectional and time series applications.

3.2 *Selection of orthogonalized regressors on a finite sample*

We now state a theorem on the selection of the orthogonalized regressors used in the model (3.2); that is, an ordering principle for the columns of S. We begin with a general treatment for $m \geq 1$, and then consider the case where we target one parameter of interest in each control regression, so that $m = 1$ and the remaining regressors in X are added to $Z(K)$. Any subvector of β can be estimated from (3.2), using the corresponding submatrix of X, as long as the excluded components of X are included in $Z(K)$ to be orthogonalized; in this way the information in components of X not directly included as regressors is retained through the orthogonalized regressors S. We may estimate the $m \times 1$ vector β in one regression, or component-by-component in a sequence of m separate control regressions for each individual β_i. In finite samples, the latter may be preferable, as it allows us to focus on selection of controls that are optimal for each individual coefficient.

In order to judge which orthogonalized regressors to include in a regression of given order (that is, in order to choose the κ most important from the set $S(K, K)$), we need measures of the impact of the addition of a particular orthogonalized regressor $S_\nu \in (S_1, \ldots S_K)$ on the coefficients of interest. Assume that κ of the regressors have been selected, and consider the impact of adding S_ν to this set. Given X and the vector β of parameters of interest, S_ν has more impact the larger is the change in the estimate of β :

$$\Delta_{\kappa,\nu} = (\hat{\beta}_{\kappa,\nu} - \beta) - (\hat{\beta}_\kappa - \beta), \tag{3.3}$$

where $\hat{\beta}_\kappa$ is the vector of regression coefficients obtained when $S(\kappa, K)$ is the matrix of κ initially-included orthogonalized regressors, and $\hat{\beta}_{\kappa,\nu}$ is the estimate on a set of orthogonalized regressors which also includes S_ν as well as $S_1 \ldots, S_\kappa$. To evaluate this change, consider a weighted distance measure,

$$d = \Delta'_{\kappa,\nu} D \Delta_{\kappa,\nu}, \tag{3.4}$$

where D is a symmetric non-negative definite matrix. Note that we want the criterion to be invariant to a change in scale of one or more X's, and D must be chosen accordingly. We might consider the following choices:

$$\begin{aligned} (i)\ d_0 &= \|\Delta_{\kappa,\nu}\|^2 \qquad \text{for } D = I \\ (ii)\ d_1 &= \Delta'_{\kappa,\nu} D_1 \Delta_{\kappa,\nu} \text{ for } D_1 = N^{-1}X'X \\ (iii)\ d_2 &= \Delta'_{\kappa,\nu} D_2 \Delta_{\kappa,\nu} \text{ for } D_2 = N^{-1}X'M_\kappa X. \end{aligned} \tag{3.5}$$

Since d_0 is not invariant to scale changes in X, we will consider only d_1 and d_2.

Now consider the use of the distance measure (3.4) to develop an ordering of the set of orthogonal component regressors.

We first examine $\Delta_{\kappa,\nu}$ and show that it can be decomposed into two parts, one of which has a probability limit of zero as $K, N \to \infty$ and $KN^{-\frac{1}{2}} \to 0$, then use this fact to construct selection criteria based on (3.4). Define $e_{(i)} = M_\kappa X_i$, the vector of residuals from regression of X_i on the κ orthogonalized regressors included in $S(\kappa, K)$, and also the $N \times m$ matrix $E_\kappa = (e_{(1)}, \ldots, e_{(m)})$, and $\hat{A}_\kappa(\nu) = \hat{\lambda}_\nu^{-1}(E'_\kappa E_\kappa)^{-1}E'_\kappa S_\nu$, where $\hat{\lambda}_\nu$ is the estimated eigenvalue. Denote by $\hat{\zeta}_\kappa(\nu)$ the coefficient in the OLS regression of S_ν on the regressors in E_κ. We will decompose the distance measure into two random functions ψ_1 and ψ_2 which also depend on unknown parameters of the process; we will then show that we can concentrate attention on one of the components. Define

$$\psi_1(\hat{\lambda}_\nu, \hat{A}_\kappa(\nu)) \equiv \hat{\zeta}_\kappa(\nu)\theta_\nu = -(X'M_\kappa X)^{-\frac{1}{2}}\hat{\lambda}_\nu \theta_\nu \hat{A}_\kappa(\nu) \tag{3.6}$$

and

$$\begin{aligned} \psi_2(\hat{A}_\kappa(\nu)) = (X'M_\kappa X)^{-\frac{1}{2}}&[I - \hat{A}_\kappa(\nu)\hat{A}_\kappa(\nu)']^{-1} \\ &\cdot \hat{A}_\kappa(\nu)[\hat{\lambda}_\nu^{-1}S'_\nu Z(\kappa+1,\infty)\gamma(\kappa+1,\infty) + \hat{\lambda}_\nu^{-1}\hat{S}'_\nu \varepsilon]. \end{aligned} \tag{3.7}$$

We will see that as $N \to \infty$, ψ_1 will become a good approximation to $\Delta_{\kappa,\nu}$; therefore our selection criteria will exploit ψ_1.

Theorem 4. Let the conditions A1 and A2 hold. Then

$$\Delta_{\kappa,\nu} = \psi_1(\hat{\lambda}_\nu, \hat{A}_\kappa(\nu)) + \psi_2(\hat{A}_\kappa(\nu)), \tag{3.8}$$

and as $K, N \to \infty$ and $KN^{-\frac{1}{2}} \to 0$,

$$\Delta_{\kappa,\nu} - \psi_1(\hat{\lambda}_\nu, \hat{A}_\kappa(\nu)) = \psi_2(\hat{A}_\kappa(\nu)) \xrightarrow{p} 0, \tag{3.9}$$

uniformly over κ and any choice of selected regressors $S(\kappa, K)$ and S_ν. Finally

$$d_1 - \theta_\nu^2 \hat{\zeta}_\kappa(\nu)' X' X \hat{\zeta}_\kappa(\nu) \xrightarrow{p} 0 \text{ and } d_2 - \theta_\nu^2 \hat{\zeta}_\kappa(\nu)' X' M_\kappa X \hat{\zeta}_\kappa(\nu) \xrightarrow{p} 0.$$

Proof. See Appendix 1.

Since θ_ν is unknown,[6] a selection criterion will have to abstract from this parameter, and therefore will reduce to

$$\overline{d}_1 = \hat{\zeta}_\kappa(\nu)' X' X \hat{\zeta}_\kappa(\nu) \tag{3.10}$$

for d_1 (3.5(ii)), or

$$\overline{d}_2 = \hat{\zeta}_\kappa(\nu)' X' M_\kappa X \hat{\zeta}_\kappa(\nu) \tag{3.11}$$

for d_2 (3.5(iii)). Note that (3.11) gives the formula for the regression sum of squares in the regression of S_ν on $e_{(1)}, \ldots, e_{(\kappa)}$ and is equivalent to the selection criterion $R_\nu^2(\kappa)\hat{\lambda}_\nu^2$, where $R_\nu^2(\kappa)$ is the R^2 from that regression.

These criteria simplify considerably when we deal with one parameter of interest (for example because we may use a separate control regression for each of several such parameters), so that $m = 1$. For this case we will write x and e for the $N \times 1$ vectors denoted X and E_κ in the general case of m effects of interest. Denote the correlation between x and S_j, $j = 1, \ldots \kappa$, by ρ_j. Recall also that $M_\kappa S_\nu = S_\nu$ and $e = M_\kappa x$, so that $e' S_\nu = x' S_\nu$, and

$$e'e = x' M_\kappa x = x'x - \sum_{j=1}^{\kappa} \frac{(x' S_j)^2}{\lambda_j^2} = \|x\| \left(1 - \sum_{j=1}^{\kappa} \hat{\rho}_j^2 \right).$$

Therefore

$$\hat{\zeta}(\nu) = \frac{\hat{\rho}_\nu \hat{\lambda}_\nu}{\|x\|(1 - \sum_{j=1}^{\kappa} \hat{\rho}_j^2)},$$

and the criteria (3.10) and (3.11) become

$$\overline{d}_1 = \frac{\hat{\rho}_\nu^2 \hat{\lambda}_\nu^2}{(1 - \sum_{j=1}^{\kappa} \hat{\rho}_j^2)^2} \tag{3.10$'$}$$

and

$$\overline{d}_2 = \frac{\hat{\rho}_\nu^2 \hat{\lambda}_\nu^2}{\|x\|(1 - \sum_{j=1}^{\kappa} \rho_j^2)} \tag{3.11$'$}$$

[6] Recall that by (3.1) ff., θ_ν is the coefficient on S_ν. The same approach as for estimation of β could in principle be applied to consistent estimation of θ_ν; the consistent estimator would then permit the use of $\hat{\zeta}_\kappa(\nu)\theta_\nu$ in (3.6) as a criterion function.

respectively. Since the denominator does not depend on S_ν, either of these criteria reduce to selection by $\hat{\rho}_\nu^2\hat{\lambda}_\nu^2$, or equivalently, $|\hat{\rho}_\nu|\hat{\lambda}_\nu$, the product of the eigenvalue and the absolute value of the correlation between x and the potential regressor S_ν (by contrast, selection by principal components uses $\hat{\lambda}_\nu$ alone). By these rules, then, the set $\{S_\nu\}_{\nu=1}^K$ is ordered such that S_{ν_1} is ordered before S_{ν_2}, i.e. $\nu_1 < \nu_2$, if

$$|\hat{\rho}_{\nu_1}|\hat{\lambda}_{\nu_1} > |\hat{\rho}_{\nu_2}|\hat{\lambda}_{\nu_2}, \tag{3.12}$$

that is, ordering is by product of eigenvalue and correlation. Related ideas have been suggested by Joliffe (1982) and Sun (1995) in prediction contexts.

Theorem 2 required that $k = K$ for consistency. With the selection rule (3.12) for the ordering of the K orthogonalized regressors, we show in Theorem 5 that consistent estimation can be based on a subset of $\kappa << K$ of these. Consider $\Omega \equiv \Omega(\{\mu_l\}, \{\phi_{l_1 l_2}\}, \{c, \beta, \gamma\})$, the set of all processes satisfying A1 with the same parameters and bounds.

Theorem 5. Let processes from the set Ω satisfy the conditions A1 and A2, with components selected by the rule (3.12). Then as $N \to \infty, K \to \infty,\ KN^{\frac{1}{2}} \to 0$ and $\kappa > K - o(K^{\frac{1}{2}})$,

$$\sup_{\Omega} |\hat{\beta}_\kappa - \beta_0| \xrightarrow{p} 0.$$

Proof. See Appendix 1.

Thus we have shown that uniformly over Ω, even in the most unfavorable cases, consistency obtains for a reduced dimension κ.

The criteria just described give an ordering for the eigenvectors and therefore the orthogonal components, but do not describe the 'stopping rule', or number of regressors to include. For this purpose, conditional on the ordering just defined, information criteria may be used. The finite-sample simulations in the next section suggest the use of the AIC for choice of κ, as well as providing information on the finite-sample performance of the methods. However, various other alternatives including model averaging may be used to select or combine components.

4. Finite-sample evaluation of bias and RMSE

Although the primary focus of this paper is asymptotic behaviour of estimators of interpretable parameters, it also is interesting to consider the finite-sample characteristics of estimation in the context given above. The results indicate that, in addition to desirable asymptotic properties, simple selection of orthogonalized components as controls can produce low finite-sample RMSE. Further reductions may be available through model averaging; see for example Magnus et al. (2009), who derive MSE results in a model with fixed regressors and Gaussian errors.

The process (2.1) covers a wide class of cases, both time series and cross-sectional, and does not restrict the number of factors or their relative importance. It is therefore difficult to specify a small number of representative parameter configurations for finite-sample evaluation. Rather than specifying a few examples, we instead use randomly selected sets of coefficients to parameterize both the relation between y and Z and the correlation between X and Z. We report results which are averages across these sets of randomly-selected data generation processes (as well, of course, as being averages across repeated experiments on each randomly-chosen DGP). By averaging across many such parameter combinations, we expect more representative results than would be possible through investigation of a few selected cases.

All observable potential explanatory variables in these simulations have at least some degree of relevance to the DGP, so that as $N \to \infty$ all should be selected into the model, but the relevance may be very small. The DGP for the simulations is: $N = 200$ and

- i) $\dim Z = L = K = 40;\ \kappa = 1, \ldots, 20$
- ii) $\gamma_j = 5\alpha^j \eta_j,\ j = 1, \ldots, 40,\ \eta_j \sim IN(0,1),\ \alpha \in (0.5, 1.0)$
- iii) $\Gamma'\Gamma = \mathrm{cov}(v)$ (v is defined below)
- iv) $Z = Z_0\Gamma,\ \{Z_0\}_{ij} \sim N(0,1)$
- v) $x_i = Z_i'\mu + e_{1,i},\ \mu_j \sim N(0,1),\ e_{1,i} \sim N(0,1),$
- vi) $y_i = x_i + Z_i'\gamma + e_{2,i},\ e_{2,i} \sim IN(0,1)$

where κ is the column dimension of $S(\kappa, K)$ (i.e. the number of controls), K is the column dimension of Z, and v is a set of K random series defined recursively such that the hth series is a linear combination of series 1 to $h-1$, with random weights. The set v allows us to create random correlation structures in Z, so that results are not specific to particular patterns of correlation. That is, randomly selected coefficients are chosen, and a decay parameter α is applied (step ii). The correlation structure of Z is defined for each parameterization by a recursive form, from which a correlation matrix is obtained from a random realization of the structure, and is applied to a raw matrix of white-noise entries using the Cholesky decomposition Γ of the correlation matrix (steps iii-iv). These steps ii-iv are repeated 200 times; for each of these 200 cases, 1000 replications of steps v-vi are computed, in each of which x is defined as a linear combination with weights randomly drawn from N(0,1) of the Z's, plus white noise (step v), and y is obtained from each of these explanatory factors (step vi). On each of these replications, the methods are applied for each value of κ.

The effect of the decay parameter is to vary the average relative importance of effects within the set of Z's: with α near unity, each of the Z_i's has an expected absolute coefficient near 1, and as α falls, the importance of coefficients other than the first few is reduced correspondingly. For relatively low values of α, e.g. near 0.5, only a few of the 40 explanatory factors have any substantial weight: the draw from $N(0,1)$ is scaled by α^j for factor j, so that factors beyond five or six are very likely to have coefficients near zero. In these cases the DGP is close to a process with only a small number of relevant explanatory factors. By contrast, for α near 1, the set of coefficients on the Z_i is close to a set of independent mean-zero random variables, and there is some tendency for cancellation to occur among variables projecting onto x_i. Realistic problems in which there is a substantial number of explanatory factors, but in which a few tend to dominate, may be best represented by moderate values of α such as 0.8 or 0.9. We emphasize these values in the experiments.

Estimation of β (=1) is carried out for each class of case by several methods: using (3.2), as well as by the univariate model, and finally by selection of un-orthogonalized regressors from Z. In using (3.2), we select the orthogonalized regressors both by principal components (largest κ eigenvalues) and by the product of eigenvalue and absolute value of the correlation with x (3.12) (labelled 'alternative eigenvector selection' in the figures). The selection of regressors from Z is included for comparison: the selection of regressors is determined by choosing ten random combinations of κ of the explanatory series, which are compared by minimum sum of squared residuals (equivalent to standard information criteria here, since the number of parameters is equal to κ in each case). The best-fitting combination is taken and compared with the dimension reduction methods.

For each of the 200 randomly-selected parameterizations and for each class of case described above, 1000 replications are drawn for each parameterization and the results for each value of κ are recorded in Figures 1, a-d.[7] These figures record the absolute biases and root mean squared errors in models of the form (3.2), for each of the orthogonal-regressor selection methods, and for comparison also record the fixed absolute bias (relative to β) in the univariate model $y_i = X_i\beta^\bullet + e_i^\bullet$, which uses no information in Z, so that $\text{plim}\hat{\beta}^\bullet = \beta + \gamma(Z'Z)^{-1}Z'X$. Each of the means is taken across both sets of parameterizations and replications of the experiment.

Clearly, augmenting the model with even a small number of terms produces a substantial bias and RMSE reduction. The effect of bias clearly dominates the RMSE; increase in variance with κ is small (that is, the RMSE does begin to increase for large κ, but the effect is so small as to be hard to detect in the figures). Selection of orthogonalized regressors by the product $\lambda \cdot \text{corr}(X, S_\ell)$ produces small but consistent reductions in bias and RMSE relative to selection by largest eigenvalue. With respect to selection of untransformed regressors, both standard principal components and the alternative selection method dominate regressor selection with respect to RMSE, although for $\kappa = 1, 2$ all selection methods are approximately equivalent at a decay parameter of 0.80. With respect to the bias component, however, regressor selection is better up to lag 3 for each value of the decay parameter, and thereafter the orthogonalized regressor methods are preferable. Note of course that very low values of κ are clearly sub-optimal in general, and that in the region of interest the orthogonalized regressor methods are clearly superior on both criteria.

Finally, we note the performance of some standard information criteria in selection of a number of orthogonalized regressors. The RMSE results, showing a strong asymmetry between deviations of the chosen order below and above the optimum, suggest that criteria that yield relatively generous parameterizations will perform relatively well on this problem. In comparing the criteria of Akaike (1974) (and the Hurvich-Tsai 1989 modification),

[7] In each case, panel a records the absolute bias for $\alpha = 0.80$; panel b: absolute bias, $\alpha = 0.95$; panel c: RMSE, $\alpha = 0.80$; panel d: RMSE, $\alpha = 0.95$.

Hannan-Quinn (1979), and Schwarz (1978), we find that all of the criteria tend to perform well in selecting an appropriate number of control regressors–in part a consequence of the tendency to flatness of the function in the region of the optimum, indicating that small deviations from the optimum have very low cost. Nonetheless, the relatively generous AIC performs particularly well, a result of the fact that a given degree of over-parameterization is in general less costly than the same degree of under-parameterization.[8] [9]

5. Concluding remarks

When a model is designed for the purpose of providing statistical controls for estimation of a small set of effects of interest, regressor selection can be adapted to this specific purpose. In particular, control regressors need not correspond with individually identifiable data series: they can instead be selected using eigenvectors of the moment matrix of available data so as to provide the greatest effect for a given number of regressors. A traditional difficulty in classical principal components regression concerns the interpretation of the coefficients, but this difficulty does not arise in models of this type because of the separation of the effect of interest from the set of data from which eigenvectors are extracted.

This paper establishes properties of models used in this way, under much more general conditions than have been used in the previous literature. We show that consistent estimation of an effect of interest is possible without requiring the existence of finite orders for the number of relevant controls nor for the number of eigenvectors used to extract information from them. We also show that selection of eigenvectors by the principal component method can be effective in this context, but that alternative selection methods designed for the problem at hand, in particular by taking account of the correlation between an eigenvector and the variable of interest, can produce better results.

Given the increasing availability of large numbers of data series and the applicability of these methods in both cross-sectional and time series contexts, and given as well the difficulty involved in specifying a regression model by selecting an appropriate subset from a large number of regressors, methods in this class appear to have substantial utility. The use of further devices for using the information, for example through model averaging, are of course also possible.

[8] The criteria were examined for a variety of sample sizes in addition to the case with sample size 200 recorded in the figures. For illustration, however, in the $N = 200$ case the mean selected orders were: AIC, 11.4; AIC-Hurvich/Tsai, 10.8; Hannan-Quinn, 9.5; Schwarz, 8.8.

[9] A natural alternative to the use of information criteria would be to compute the coefficients of interest for various values of κ, and to select a value of κ at which the estimated values of the coefficient become stable for small changes in κ. We find that the AIC tends to be successful in making such a choice.

Figure 1a:
Estimated bias by number of eigenvectors, k
T=200, N=40, decay parameter = 0.80

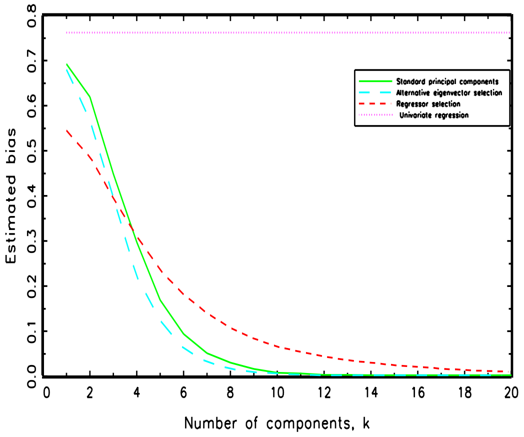


Figure 1b:
Estimated bias by number of eigenvectors, k
T=200, N=40, decay parameter = 0.95

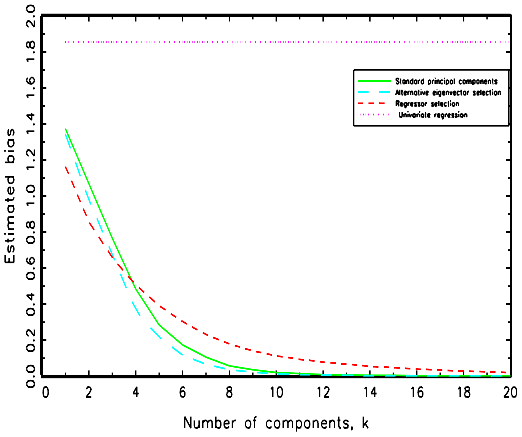


Figure 1c:
Estimated RMSE by number of eigenvectors, k
T=200, N=40, decay parameter = 0.80

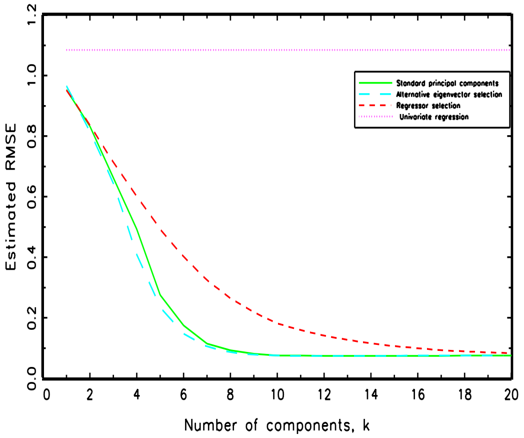
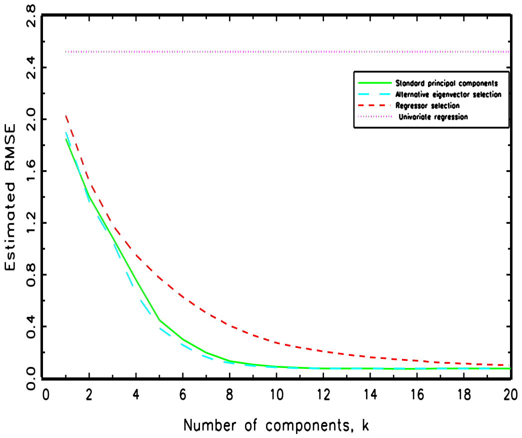


Figure 1d:
Estimated RMSE by number of eigenvectors, k
T=200, N=40, decay parameter = 0.95



Appendix 1
Proofs of Lemma 1 and Theorems 1–4

Proof of Lemma 1.

(i) Let $\{\mu_\nu\}_{\nu=1}^{\infty}$ be an orthonormal basis for the Hilbert space $\mathcal{H}$ such that the Wold decomposition for each W_ℓ is expressed in this basis. Then we can express $y_i - c,\ X_{\ell i}, \ell = 1, \dots m$, and $Z_{\ell i}, \ell = 1, \dots \infty$ in this basis and write

$$E(y_i - c|\mathcal{F}_i) = \sum_{\nu=1}^{\infty} a_{\nu_i}(y)\mu_{\nu_i} = \sum_{\ell=1}^{m} \beta_\ell \sum_{\nu=1}^{\infty} a_\nu(X_{\ell i})\mu_{\nu_i} + \sum_{\ell=1}^{\infty} \gamma_\ell \sum_{\nu=1}^{\infty} a_\nu(Z_{\ell i})\mu_{\nu_\ell},$$

where the μ_{ν_i} are measurable with respect to $\mathcal{F}_i$. Then
$a_\nu(y) = \sum_{\ell=1}^{m} \beta_\ell a_\nu(X) + \sum_{\ell=1}^{\infty} \gamma_\ell a_\nu(Z)$. By stationarity of the process, $\sum_{\nu=1}^{\infty}(a_\nu(y))^2 < \infty$ and $\sum_{\nu=1}^{\infty}(a_\nu(Z))^2 < \infty$. Therefore $\sum_{\nu=1}^{\infty} \left(\sum_{\ell=1}^{\infty} \gamma_\ell a_\nu(Z_\ell)\right)^2 < \infty$; since $\sum_{\nu=1}^{\infty} \left(\sum_{\ell=1}^{\infty} \gamma_\ell a_\nu(Z_\ell)\right)^2 = \gamma' \Sigma_Z \gamma = \|\Sigma_Z^{\frac{1}{2}} \gamma\|$, we have $\|\Sigma_Z^{\frac{1}{2}}\gamma\|^2 < \infty$. By A1 (iv), $\underline{\lambda}(\Sigma_Z) > \underline{\zeta}$. Then $\|\Sigma_Z^{-\frac{1}{2}}\| < \underline{\zeta}^{-\frac{1}{2}}$ and

$$\|\gamma\| \le \|\Sigma_Z^{-\frac{1}{2}} \Sigma_Z^{\frac{1}{2}} \gamma\| \le \|\Sigma_Z^{-\frac{1}{2}}\| \|\Sigma_Z^{\frac{1}{2}}\gamma\| < \infty.$$

(ii) Since $E(\varepsilon_i|\mathcal{F}_i) = 0$ from (2.1), to show this we need only verify that $E|\varepsilon_i|$ is finite. Now $E|\varepsilon_i| \le (E(\varepsilon_i^2))^{\frac{1}{2}}$ by Jensen's inequality. Up to the constant, $\varepsilon_i = A'W,\ A = (1, -\beta, -\gamma)'$. Therefore $\varepsilon_t^2 = E(A'WW'A) \le \|A\| \|\Sigma_W\| \le \overline{\lambda}(\Sigma_W)$. Since $\|A\|$ is finite by part (i) of the Lemma, it follows that $E|\varepsilon_i| < \infty$. ∎

Proof of Theorem 1.

Consider

$$E(Z(k+1,\infty)\gamma(k+1,\infty))^2 = E(\sum_{\ell=1}^{\infty} Z_{k+\ell}\gamma_{k+\ell})^2 \le \sup_\ell E(Z_{k+\ell})^2 (\sum_{\ell=1}^{\infty} |\gamma_{k+\ell}|)^2.$$

Here, $\sup_\ell E(Z_{k+\ell})^2$ is bounded by A1(v), and $\sum_{\ell=1}^{\infty} |\gamma_{k+\ell}| \to 0$ as $k \to \infty$ since by A2, $\sum_{\ell=1}^{\infty} |\gamma_\ell| < \infty$. Thus $E(Z(k+1,\infty)\gamma(k+1,\infty))^2 \to 0$ and by Chebyshev's inequality, $Z(k+1,\infty)\gamma(k+1,\infty) \xrightarrow{p} 0$. ∎

Proof of Theorem 2.

To avoid treating the constant we assume without loss of generality that all variables are expressed in deviations from the mean. Using the OLS estimator of β, we have

$$\hat{\beta}_k - \beta = (X' M_k X)^{-1} X' M_k (Z(k+1,\infty)\gamma(k+1,\infty) + \varepsilon), \qquad (A2.1)$$

where $M_k = I - Z(k)(Z(k)'Z(k))^{-1}Z(k)'$. From Hannan (1960) it follows that under Assumption A1 (i-iii, v, vi), for any $\delta_1 > 0$ and for large enough N,

$$\sup_{\ell_1,\ell_2}(N-\tilde{k})E\left(\frac{1}{N-\tilde{k}}\sum_{i=\tilde{k}}^{N} W_{\ell_1,i}W_{\ell_2,i+\xi} - \phi_{\ell_1,\ell_2}(|\xi|)\right)^2 < \delta_1,$$

and so as $N \to \infty,\ k \to \infty$, and $kN^{-1} \to 0$,

$$(N-\tilde{k})^{-1}\sum_{i=\tilde{k}}^{N} W_{\ell_1,i}W_{\ell_2,i+\xi} - \phi_{\ell_1,\ell_2}(|\xi|) = O_p(N-\tilde{k})^{-\frac{1}{2}}.$$

Therefore

$$\frac{1}{N-\tilde{k}}Z(k)'Z(k) - E(\frac{1}{N-\tilde{k}}Z(k)'Z(k)) = O_p(N-\tilde{k})^{-\frac{1}{2}}, \qquad (A2.2)$$

$$\frac{1}{N-\tilde{k}}X'Z(k) - E(\frac{1}{N-\tilde{k}}X'Z(k)) = O_p(N-\tilde{k})^{-\frac{1}{2}}, \qquad (A2.3)$$

uniformly as $N \to \infty,\ k \to \infty$, and $kN^{-1} \to 0$. From Assumption A1(iv) it follows that $E(\frac{1}{N-\tilde{k}}Z(k)'Z(k))$ is invertible, that its inverse has a finite norm, and from Berk (1974, Lemma 3), for $k^2N^{-1} \to 0$ it is straightforward to show that[10]

$$\left\|\left(\frac{1}{N-\tilde{k}}Z(k)'Z(k)\right)^{-1} - \left[E\left(\frac{1}{N-\tilde{k}}Z(k)'Z(k)\right)\right]^{-1}\right\| = o_p(1). \qquad (A2.4)$$

Thus, substituting from (A2.2–A2.4), we have

$$\|\frac{1}{N-\tilde{k}}X'M_kX - G_k\| = o_p(1), \qquad (A2.5)$$

where $G_k = E(\frac{1}{N-\tilde{k}}X'M_kX) =$

$$E\left[(\frac{1}{N-\tilde{k}})^{\frac{1}{2}}X'[I - (\frac{1}{N-\tilde{k}})^{\frac{1}{2}}Z(k)Q_k(\frac{1}{N-\tilde{k}})^{\frac{1}{2}}Z(k)'](\frac{1}{N-\tilde{k}})^{\frac{1}{2}}X\right],$$

with $Q_k = (E(\frac{1}{N-\tilde{k}}Z(k)'Z(k)))^{-1}$.

[10] The notation $\|.\|$ refers to either the vector or matrix norm in the Euclidean vector space.

Since by Assumption A1(iv) X cannot belong to the space spanned by the $Z's$, the eigenvalues of G_k are bounded away from zero independently of k; it is straightforward to show that

$$\left\| \left(\frac{1}{N-\tilde{k}} X' M_k X \right)^{-1} - G_k^{-1} \right\| \xrightarrow{p} 0. \qquad (A2.6)$$

Next consider $\frac{1}{N-\tilde{k}} X' M_k (R_k \theta + \varepsilon)$. For $\frac{1}{N-\tilde{k}} X' M_k \varepsilon$, write

$$\frac{1}{N-\tilde{k}} X'\varepsilon - \left(\frac{1}{N-\tilde{k}} X' Z(k) \right) \left(\frac{1}{N-\tilde{k}} Z(k)' Z(k) \right)^{-1} \left(\frac{1}{N-\tilde{k}} \right) Z(k)'\varepsilon.$$

For $\frac{1}{N-\tilde{k}} X'\varepsilon$, by Hannan (1960) we have

$$\left\| \frac{1}{N-\tilde{k}} X'\varepsilon - E(\frac{1}{N-\tilde{k}} X'\varepsilon) \right\| = O_p((N-\tilde{k})^{-\frac{1}{2}},$$

and since ε_i is a martingale difference sequence with respect to $\mathcal{F}_i$, $E(\frac{1}{N-\tilde{k}} X'\varepsilon) = 0$ and so $\frac{1}{N-\tilde{k}} X'\varepsilon = O_p((N-\tilde{k})^{-\frac{1}{2}})$. Exactly the same considerations provide $\frac{1}{N-\tilde{k}} Z(k)'\varepsilon = O_p((N-\tilde{k})^{-\frac{1}{2}})$. By (A2.3) and (A2.4),

$$\left(\frac{1}{N-\tilde{k}} X' Z(k) \right) \left(\frac{1}{N-\tilde{k}} Z(k)' Z(k) \right)^{-1} = O_p(1),$$

and we obtain that $\frac{1}{N-\tilde{k}} X' M_k \varepsilon = O_p((N-\tilde{k})^{-\frac{1}{2}})$.

Finally, $\frac{1}{N-\tilde{k}} X' M_k (Z(k+1,\infty)\gamma(k+1,\infty))$ is an $m \times 1$ vector with ℓ'th component

$$b_\ell = \frac{1}{N-\tilde{k}} \sum_{j=1}^{N} (X' M_k)_{\ell j} \cdot \sum_{i=1}^{\infty} Z_{k+i+1} \gamma_{k+i+1}.$$

Then

$$\begin{aligned} |b_\ell| &\leq \left(\frac{1}{N-\tilde{k}} X' M_k X \right)^{\frac{1}{2}} \left(\frac{1}{N-\tilde{k}} \sum_{j=k}^{N} (\sum_{i=1}^{\infty} Z_{k+i+j} \gamma_{k+i+j})^2 \right)^{\frac{1}{2}} \\ &\leq O_p(1) \left(\frac{1}{N-\tilde{k}} \sum_{j=k}^{N} (\sum_{i=1}^{\infty} Z_{k+i+j} \gamma_{k+i+j})^2 \right)^{\frac{1}{2}} = o_p(1), \end{aligned}$$

where the second inequality follows from (A2.5) and the last result by Theorem 1.

It follows that $\hat{\beta}_k - \beta = O_p(1) \cdot \sum_{i=1}^{\infty} |\theta_{k+i}| + O_p((N-k)^{-\frac{1}{2}})$, and Theorem 2 follows. ∎

Proof of Theorem 3.

From (A2.1) we can write

$$(N-\tilde{k})^{\frac{1}{2}}(\hat{\beta}_k - \beta) = \left(\frac{X'M_kX}{N-\tilde{k}}\right)^{-1} (N-\tilde{k})^{-\frac{1}{2}} X'M_k(Z(k+1,\infty)\gamma(k+1,\infty)+\varepsilon).$$

By (A2.6) this is

$$[G_k^{-1} + o_p(1)]\ [(N-\tilde{k})^{-\frac{1}{2}} X'M_kZ(k+1,\infty)\gamma(k+1,\infty) + (N-\tilde{k})^{-\frac{1}{2}} X'M_k\varepsilon].$$

Since $\{\varepsilon_i, \mathcal{F}_\rangle\}$ is a martingale difference (m.d.) sequence, the moment conditions (v) imply that the m.d. central limit theorem applies to the m.d. array, and as $N \rightarrow \infty$, $k \rightarrow \infty$, $k^{-1}N \rightarrow \infty$, for $V_k = E(\frac{1}{N-\tilde{k}} X'M_k\varepsilon\varepsilon'M_kX)$, we have

$$(N-\tilde{k})^{\frac{1}{2}} V_k^{-\frac{1}{2}} X'M_k\varepsilon \xrightarrow{D} N(0, I_m).$$

Recall that $(N-\tilde{k})^{-\frac{1}{2}} X'M_kZ(k+1,\infty)\gamma(k+1,\infty)$ is an $m \times 1$ vector with ℓ'th component $(N - k)^{\frac{1}{2}}\ b_\ell$, where by (A2.8), $|b_\ell| \leq O_p(1) \sum_{i=1}^{\infty} |\gamma_{k+i}|$. By the conditions of Theorem 3, $\sum_{i=1}^{\infty} |\gamma_{k+i}| = o((N-\tilde{k})^{\frac{1}{2}})$. Therefore

$$(N-\tilde{k})^{\frac{1}{2}} V_k^{-\frac{1}{2}} G_k(\hat{\beta}_k - \beta) \xrightarrow{D} N(0, I_m).$$

If ε is independent of (X, Z) then $G_k^{-1} V_k G_k^{-1} = \sigma_\varepsilon^2 E\left(\frac{X'M_kX}{N-\tilde{k}}\right)$. ∎

Proof of Theorem 4.

Consider (A2.1) and the last line of (3.1), to write

$$\hat{\beta}_\kappa - \beta = (X'M_\kappa X)^{-1} X'M_\kappa(R(\kappa,\infty)\theta(\kappa,\infty) + \epsilon); \qquad (A4.1)$$

$$\hat{\beta}_{\kappa,\nu} - \beta = (X'M_{\kappa,\nu} X)^{-1} X'M_{\kappa,\nu}(R(\kappa,\infty)\theta(\kappa,\infty) - S_\nu\theta_\nu + \epsilon). \qquad (A4.2)$$

Here we define $P_{\kappa,\nu}$ as the projection onto the space spanned by $S(\kappa, K)$ and S_ν, and $M_{\kappa,\nu} = I - P_{\kappa,\nu}$, we note that $M_\kappa S_\nu = P_{\kappa,\nu} S_\nu = S_\nu$, $P_{\kappa,\nu}X = P_\kappa X + \hat{\lambda}_\nu^{-2} S_\nu S_\nu' X$, and also $M_{\kappa,\nu}X = M_\kappa X - \hat{\lambda}_\nu^{-2} S_\nu S_\nu' X$; further, $X'M_{\kappa,\nu}X = X'M_\kappa X - \hat{\lambda}_\nu^{-2} X'S_\nu S_\nu' X$.

For$(X'M_{\kappa,\nu}X)^{-1}$ we can write

$$(X'M_{\kappa,\nu}X)^{-1} = (X'M_\kappa X)^{-\frac{1}{2}}[I-D]^{-1}(X'M_\kappa X)^{-\frac{1}{2}}, \qquad (A4.3)$$

where $D = \hat{\lambda}_\nu^{-2}(X'M_{\kappa,\nu}X)^{-\frac{1}{2}}X'M_\kappa S_\nu S'_\nu M_\kappa X(X'M_\kappa X)^{-\frac{1}{2}}$.

Next consider the $m \times 1$ vector

$$\hat{A}_\kappa(\nu) = \hat{\lambda}_\nu^{-1}(X'M_\kappa X)^{-\frac{1}{2}}X'M_\kappa S_\nu, \qquad (A4.4)$$

and the matrix $D = \hat{A}_\kappa(\nu)\hat{A}_\kappa(\nu)'$. Recall that the matrix norm for this matrix is $\|\hat{A}_\kappa(\nu)\hat{A}_\kappa(\nu)'\| = \sup_{\|x\|=1} x'\hat{A}_\kappa(\nu)\hat{A}_\kappa(\nu)'x = \hat{A}'_\kappa(\nu)\hat{A}_\kappa(\nu)$.

For each X_i, $E_i = M_\kappa X_i$ is the vector of residuals from regressing X_i on the κ included orthogonal regressors. Consider a regression of S_ν on E; the R^2 in that regression is

$$1 \geq R^2 = \hat{\lambda}_\nu^{-2}(S'_\nu E(E'E)^{-1}E'S_\nu) = \hat{A}'_\kappa(\nu)\hat{A}_\kappa(\nu). \qquad (A4.5)$$

We next show that there exists $\overline{A} < 1$ such that for any M_κ, S_ν, $Pr(A_\kappa(\nu)'A_\kappa(\nu) < \overline{A}) \to 1$ as $N, K \to \infty$ under the conditions of Theorem 2.

Consider the Wold decomposition of X, Z expressed in the orthonormal basis of $\mathcal{H}$, $\{\mu_\nu\}_{\nu=1}^\infty$. By A1(iv) there exists some $\overline{\mu}_\ell$ such that for X_ℓ the coefficient on $\overline{\mu}_\ell$, $\alpha_{\overline{\mu}_\ell}(X_\ell)$, is non-zero, but for any Z_j, $\alpha_{\overline{\mu}_\ell}(Z_j) = 0$. Then for any projection M of X_ℓ orthogonally to any subset of $\{Z_\ell\}$, $MX_\ell = E_\ell$, the coefficient on $\overline{\mu}_\ell$ is $\alpha_{\overline{\mu}_\ell}(X_\ell)$. For any transformation CZ, where $CC' = I$, of Z, the corresponding coefficient is zero. Then for E_ℓ, $E(E_\ell^2) = E(E_\ell - \alpha_{\overline{\mu}_\ell}(X_\ell)\overline{\mu}_\ell)^2 + \alpha_{\overline{\mu}_\ell}(X_\ell)^2$. Under the conditions of Theorem 2, and by methods similar to the proof of Theorem 2, convergence of sample moments to population moments follows. Then $Pr(\hat{A}'_\kappa(\nu)\hat{A}_\kappa(\nu) < \overline{A}) \to 1$ for $\overline{A} = 1 - \min_{1\leq\ell\leq m}\left(\frac{\alpha_{\overline{\mu}_\ell}(X_\ell)^2}{var(X_\ell)}\right)$.

It follows that

$$(I - \hat{A}_\kappa(\nu)\hat{A}'_\kappa(\nu))^{-1} = I + \hat{A}_\kappa(\nu)\hat{A}'_\kappa(\nu) + \cdots + (\hat{A}_\kappa(\nu)\hat{A}'_\kappa(\nu))^n + \cdots \qquad (A4.6)$$

is a valid expansion; note that

$$(I - \hat{A}_\kappa(\nu)\hat{A}'_\kappa(\nu))^{-1} = I + \hat{A}_\kappa(\nu)(I - \hat{A}'_\kappa(\nu)\hat{A}_\kappa(\nu))^{-1}\hat{A}'_\kappa(\nu). \qquad (A4.7)$$

Express the right-hand side of (A4.3) via $\hat{A}_{\kappa+1}$ from (A4.4); by applying (A4.7) we can verify that (A4.3) can be written as

$$\begin{aligned}&(X'M_{\kappa,\nu}X)^{-1}\\ &= (X'M_\kappa X)^{-\frac{1}{2}}[I + \hat{A}_\kappa(\nu)(I - \hat{A}'_\kappa(\nu)\hat{A}_\kappa(\nu))^{-1}\hat{A}'_\kappa(\nu))](X'M_\kappa X)^{-\frac{1}{2}} \qquad (A4.8)\\ &= (X'M_\kappa X)^{-1} + \Omega_{\kappa,\nu},\end{aligned}$$

where $\Omega_{\kappa,\nu} = (X'M_\kappa X)^{-\frac{1}{2}}\hat{A}_\kappa(\nu)(I - \hat{A}'_\kappa(\nu)\hat{A}_\kappa(\nu))^{-1}\hat{A}'_\kappa(\nu)(X'M_\kappa X)^{-\frac{1}{2}}$.

Next define $v = R(\kappa,\infty)\theta(\kappa,\infty) + \epsilon$. Then

$$\hat{\beta}_{\kappa,\nu} - \beta = [(X'M_\kappa X)^{-1} + \Omega_{\kappa,\nu}][X'M_\kappa - \hat{\lambda}_\nu^{-2}X'S_\nu S'_\nu][v - S_\nu\theta_\nu]. \qquad (A4.9)$$

From (A4.1),(A4.4), (A4.8) and (3.1), (A4.9) becomes

$$\begin{aligned}\hat{\beta}_{\kappa,\nu} - \beta = \hat{\beta}_\kappa - \beta - (X'M_\kappa X)^{-\frac{1}{2}}\hat{\lambda}_\nu\theta_\nu\hat{A}_\kappa(\nu) + (X'M_\kappa X)^{-\frac{1}{2}} \\ \cdot [I - \hat{A}_\kappa(\nu)\hat{A}'_\kappa(\nu)]^{-1}\hat{A}_\kappa(\nu)\hat{\lambda}_\nu^{-1}[S'_\nu Z(K+1,\infty)\gamma(K+1,\infty) + S'_\nu\epsilon],\end{aligned} \qquad (A4.10)$$

and (3.8) follows. In $\psi_2(\hat{A}_\kappa(\nu))$ of (3.7), the factor
$(X'M_\kappa X)^{-\frac{1}{2}}[I - \hat{A}_\kappa(\nu)\hat{A}'_\kappa(\nu)]^{-1}\hat{A}_\kappa(\nu) = N^{-\frac{1}{2}}(\frac{X'M_\kappa X}{N})^{-\frac{1}{2}}[I - \hat{A}_\kappa(\nu)\hat{A}'_\kappa(\nu)]^{-1}\hat{A}_\kappa(\nu)$
$= O_p(N^{-\frac{1}{2}})$, since $\|\hat{A}_\kappa(\nu)\| < \overline{A}$ with probability arbitrarily close to 1 (for large enough N), and

$$|N^{-\frac{1}{2}}\frac{S_\nu}{\hat{\lambda}_\nu}Z(K+1,\infty)\gamma(K+1,\infty)| \leq O_p(N^{-\frac{1}{2}})|Z(K+1,\infty)\gamma(K+1,\infty)|,$$

which goes to zero in probability by Theorem 1. As well, for any ν_i and ν_j, $E(S_{\nu_i}\epsilon_i) = 0$ and $\text{cov}(S_{\nu_i}\epsilon_i, S_{\nu_j}\epsilon_j) = 0$ since ϵ_i is a m.d. sequence. Therefore

$$\sup_\nu P(|N^{-1}\sum_{i=1}^N S_{\nu_i}\epsilon_i| > \epsilon) \leq \frac{\sup_\ell E(W_\ell^2)}{N\epsilon}.$$

Thus (3.9) of Theorem 4 follows. Recall that $\psi_1(\hat{\lambda}_\nu, \hat{A}_\kappa(\nu)) = (E'_\kappa E_\kappa)^{-1}E'_\kappa S_\nu\theta_\nu = \hat{\zeta}_\kappa(\nu)\theta_\nu$; the rest of the theorem then follows. ■

Proof of Theorem 5.

To simplify the proof consider $m = 1$. All that we need to show in addition to the result of Theorem 2 is that uniformly over all processes in Ω,

$$\left|(X'M_k X)^{-1}X'S(\kappa+1,K)\theta(\kappa+1,K)\right| \to_p 0. \qquad (A5.1)$$

Rewrite

$$(X'M_k X)^{-1}X'S(\kappa+1,K)\theta(\kappa+1,K) = \left(\frac{X'M_k X}{N}\right)^{-1}\left(\frac{X'X}{N}\right)^{\frac{1}{2}}\sum_{\nu=\kappa+1}^K \hat{\rho}_\nu\frac{\hat{\lambda}_\nu}{\sqrt{N}}\theta_\nu.$$

Note that $|\theta_\nu| \leq \|\gamma\|$ which by Lemma 1 is bounded. Using the Wold decomposition similarly to the proof of Theorem 4 we show that convergence of sample moments to population moments and A1(iv) imply that for some constant C_1 independently of M_κ

$$\Pr\left\{\sup_{M_\kappa}\left(\frac{X'M_kX}{N}\right)^{-1}\left(\frac{X'X}{N}\right)^{\frac{1}{2}} > C_1\right\} \to 0.$$

Thus

$$\Pr\left\{\left(\frac{X'M_kX}{N}\right)^{-1}\left(\frac{X'X}{N}\right)^{\frac{1}{2}}\sum_{\nu=\kappa+1}^{K}\hat{\rho}_\nu\frac{\hat{\lambda}_\nu}{\sqrt{N}}\theta_\nu < C_1\|\gamma\|\sum_{\nu=\kappa+1}^{K}|\hat{\rho}_\nu|\frac{\hat{\lambda}_\nu}{\sqrt{N}}\right\} \to 1. \qquad (A5.2)$$

Consider a vector $(|\hat{\rho}_1|\frac{\hat{\lambda}_1}{\sqrt{N}}, ..., |\hat{\rho}_K|\frac{\hat{\lambda}_K}{\sqrt{N}})'$; its norm is the same as that of $(\hat{\rho}_1\frac{\hat{\lambda}_1}{\sqrt{N}}, ..., \hat{\rho}_K\frac{\hat{\lambda}_K}{\sqrt{N}})'$. By convergence of sample moments and boundedness of the matrix norm of the covariance matrix, for some constant C_2 and any K

$$\Pr\left\{\sum_{\nu=1}^{K}\left(\hat{\rho}_\nu\frac{\hat{\lambda}_\nu}{\sqrt{N}}\right)^2 > C_2\right\} \to 0.$$

Consider now a set

$$\Xi = \{x = (x_1, ..., x_K)' \in R^K : \|x\| = C; x_{\nu_1} \geq x_{\nu_2} \text{ for } \nu_1 < \nu_2\}$$

and solve

$$\max_{x\in\Xi}\sum_{\nu=\kappa+1}^{K}|x_\nu|.$$

It is easy to see that the solution is x with all components equal to $\frac{C}{\sqrt{K}}$; thus the maximized value is $C\frac{K-\kappa}{\sqrt{K}}$. As $K \to \infty$ the maximum goes to zero if $\kappa = K - o(\sqrt{K})$.

Then for any ε if $\kappa = K - o(\sqrt{K})$

$$\Pr\left\{\sum_{\nu=\kappa+1}^{K}|\hat{\rho}_\nu|\frac{\hat{\lambda}_\nu}{\sqrt{N}} > \varepsilon\right\} \to 0$$

always, and combined with (A5.2) the result of Theorem 5 follows. ■

References

Akaike, H. (1974) A new look at the statistical model identification. *IEEE Transactions on Automatic Control,* AC-19, 716-723.

Bai, J. and S. Ng (2002) Determining the number of factors in approximate factor models. *Econometrica* 70, 191-221.

Bai, J. and S. Ng (2006) Confidence intervals for diffusion index forecasts and inference for factor-augmented regressions. *Econometrica* 74, 1133-1150.

Banerjee, A. and M. Marcellino (2008) Factor-augmented error-correction models. Working paper, EUI.

Belloni, A. and V. Chernozhukov (2009) ℓ_1-Penalized quantile regression in high- dimensional sparse models. Working paper, MIT.

Berk, K.N. (1974) Consistent Autoregressive Spectral Estimates. *Annals of Statistics* 2, 489-502.

Chamberlain, G. (1983) Funds, factors and diversification in arbitrage pricing models. *Econometrica* 51, 1281-1304.

Chamberlain, G. and M. Rothschild (1983) Arbitrage, factor structure and mean-variance analysis in large asset markets. *Econometrica* 51, 1305-1324.

Farebrother, R.W. (1972) Principal components estimators and minimum mean squared error criteria in regression analysis. *Review of Economics and Statistics* 54, 332-336.

Forni, M. and M. Lippi (2001) The generalized dynamic factor model: representation theory. *Econometric Theory* 17, 1113-1141.

Geweke, J. (1977) The dynamic factor analysis of economic time series. In Aigner, D.J. and A.S. Goldberger, eds., *Latent Variables in Socio-Economic Models,* North-Holland, Amsterdam, 365-383.

Groen, J.J.J. and G. Kapetanios (2009) Model selection criteria for factor-augmented regressions. Staff report 363, Federal Reserve Bank of New York.

Hannan, E.J. and B.G. Quinn (1979) The determination of the order of an autoregression. *Journal of the Royal Statistical Society, Ser. B*, 41, 190-195.

Hurvich, C.M. and C. Tsai (1989) Regression and time series model selection in small samples. *Biometrika* 76, 297-307.

Joliffe, I.T. (1982) A note on the use of principal components in regression. *Applied Statistics* 31, 300-303.

Kendall, M.G. (1957) *A Course in Multivariate Analysis.* Charles Griffin & Co., London.

Magnus, J.R. and J. Durbin (1999) Estimation of regression coefficients of interest when other regression coefficients are of no interest. *Econometrica* 67, 639-643.

Magnus, J.R., O. Powell and P. Prüfer (2009) A comparison of two model averaging techniques with an application to growth empirics. *Journal of Econometrics* , in press.

McCallum, B.T. (1970) Artificial orthogonalization in regression analysis. *Review of Economics and Statistics* 52, 110-113.

Robinson, P.M. (1988) Root-N-consistent semiparametric regression. *Econometrica* 56, 931-954.

Sargent, T.J. and C.A. Sims (1977) Business cycle modelling without pretending to have too much *a priori* economic theory. In Sims, C.A., ed., *New Methods in Business Cycle Research,* Federal Reserve Bank of Minneapolis, Minneapolis, 45-109.

Schwarz, G. (1978) Estimating the dimension of a model. *Annals of Statistics* 6, 461-464.

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

Stock, J.H. and M.W. Watson (2002b) Macroeconomic forecasting using diffusion indexes. *Journal of Business and Economic Statistics* 20, 147-162.

Stone, M. and R.J. Brooks (1990) Continuum regression: cross-validated sequentially constructed prediction embracing ordinary least squares, partial least squares and principal components regression. *Journal of the Royal Statistical Society, Ser. B*52, 237-269.

Sun, J. (1995) A correlation principal component regression analysis of NIR data. *Journal of Chemometrics* 9, 21-29.

Tibshirani, R. (1996) Regression shrinkage and selection via the Lasso. *Journal of the Royal Statistical Society, Ser. B*58, 267-288.

Zernov, S., V. Zinde-Walsh and J.W. Galbraith (2009) Asymptotics for Estimation of Quantile Regressions with Truncated Infinite-Dimensional Processes. *Journal of Multivariate Analysis* 100, 497-508.