# Reduced-Dimension Control Regression

version of: 25. 5. 2006

John W. Galbraith and Victoria Zinde-Walsh

Department of Economics
McGill University
855 Sherbrooke St. West
Montreal, Quebec H3A 2T7 CANADA

*Abstract*

A model to investigate the relationship between one variable and another usually requires controls for numerous other effects which are not constant across the sample; where the model omits some elements of the true process, estimates of parameters of interest will typically be inconsistent. Here we investigate conditions under which, with a set of potential controls which is large (possibly infinite), orthogonal transformations of a subset of potential controls can nonetheless be used in a parsimonious regression involving a reduced number of orthogonal components (the 'reduced-dimension control regression'), to produce consistent (and asymptotically normal, given further restrictions) estimates of a parameter of interest, in a general setting. We examine selection of the particular orthogonal directions, using a new 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 good finite-sample performance of the method.

AMS 2000 MSC: primary 62J05; secondary 62M10.

Key words: Dimension reduction, linear regression, orthogonal components

We thank Frank Diebold, Denis Pelletier and Aman Ullah for valuable comments, 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.

## 1. Introduction

One of the primary uses of the regression model is to allow statistical controls to substitute for the experimental controls which are not available in typical problems in the social sciences. In many situations, however, the number of potential controls is too large to permit inclusion of each control as a separate regressor, and it is difficult to choose which elements of a set of possible regressors to include. 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, and in this paper we suggest an approach to the problem via dimension reduction.

Consider the linear process $y = c + X\beta + Z\gamma + \epsilon$, where our interest is in $\beta$, but the subset of regressors $Z$ enters non-trivially in that a model which omits elements of this subset will yield inconsistent estimates of $\beta$ if there is asymptotic correlation between the omitted elements and $X$. Normally, empirical researchers try to find $\beta$ and $\gamma$ (despite the fact that $\beta$ is the item of interest). However, if $\gamma$ is of high dimension (given $N$), this may be inefficient or infeasible. An alternative (principal components regression, e.g. Kendall 1957, McCallum 1970, Farebrother 1972, Greenberg 1975; factor analysis, e.g. Chamberlain 1983, Forni and Lippi 2001; methods related to central mean subspace estimation, e.g. Li 1991, Cook and Weisberg 1991, Zhu and Fang 1996, Cook and Li 2002) is to abandon these interpretable parameters, and fit the full conditional mean using a dimension reduction method. One does not obtain $\beta$ but may get very good fitted values or forecasts (Stock and Watson 2002a,b).

Another alternative, which the present study pursues, is to concentrate on $\beta$ alone, and to abandon estimation of immediately-interpretable parameters $\gamma$, while continuing to use the information in $Z$. Because cases in which $Z$ is of high dimension are those in which it is most useful to use dimension reduction methods such as those to be examined here, we treat $Z$ as being of unknown and possibly non-finite dimension. We then propose a method directed at providing good regression estimates of $\beta$, rather than at uncovering the entire process; we describe this as a control regression, i.e. a regression designed to allow inference on a small set of parameters while controlling for variation in others. We examine the choice of regression directions with which to exploit information from the controls. When these directions are chosen appropriately, the biases associated with omitted controls can be eliminated asymptotically and reduced to low levels in finite samples; the consequences of lack of full knowledge of the true process can therefore be limited.

From the theoretical viewpoint, a key element of the present study is the fact that the process is treated as having unknown and potentially unbounded dimension. Thus 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.

More specifically, our asymptotic results examine the possibility of consistent estimation of $\beta$ when the number of included directions from the space spanned by $Z$ grows with sample size. We impose the condition that the $L_1$ norm of the vector $\gamma$ be bounded; then consistency (and asymptotic normality, with further conditions) holds for the estimates of $\beta$ in finite models whose dimension increases with sample size. We then design a criterion for selection of directions in the regressor space and demonstrate that with this criterion, dimension reduction in the space of included regressors is achievable. The criterion 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 $\beta$ of directions tends to decline with diminishing value of the criterion. We then 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 $\beta$, 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 associated conditions, and asymptotic results. The main results of the paper are in Section 3, which describes the dimension reduction methods used to obtain statistical controls. We orthogonalize the regressors, and show that our criterion for selecting among the orthogonalized directions allows consistency 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 performance of the methods, and section 5 a brief empirical illustration. Proofs are collected in Appendix 1.

## 2. Processes, notation and preliminary results

### 2.1 *Processes and notation*

We assume that the observed data are associated with realizations of a multi-dimensional (possibly infinite-dimensional) stationary random process. 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 $\ell$, 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, $X_i' = (w_{2i}, \ldots w_{m+1,i}) = (x_{1i}, \ldots x_{m,i})$ a set of $m$ conditioning variables of interest, and $Z_i' = [z_{1i}, z_{2i}, \ldots] = [w_{m+2,i}, w_{m+3,i}, \ldots]$ 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.

*Assumption 1 (A1)*

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

. *(ii)* $W_{\ell_1}, W_{\ell_2}$ *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* $\Sigma_W$ *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 $< w_{\ell_1 i}, w_{\ell_2 j} >= \text{cov}(w_{\ell_1 i} w_{\ell_2 j})$, and there is a Wold representation for each $W_\ell$. Equation (2.1) gives the conditional expectation function, in which $\beta$ 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 $\varepsilon_i = y_i - (c + X_i'\beta + Z_i'\gamma)$, $i = 1, \ldots, N$; 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), $\Sigma_W$ 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 $\text{cov}(w_{\ell_1 i} w_{\ell_2 j}) = \gamma_{\ell_1 \ell_2}(|i-j|)$

if $i = j$, zero otherwise; and (ii) the case in which some $W_\ell$ are lagged values of others, so that for example $Z_i'$ could include $y_{i-h}$ and elements of $Z_{i-h}'$, $h > 0$. 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*

Assumption 1 implies a bound on the sum of squared coefficients on the $Z_i$, and a martingale difference sequence (m.d.s.) structure on the error $\{\varepsilon_i\}$.

**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.

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; $\gamma(k)$ and $\gamma(k{+}1,\infty)$ are the corresponding coefficient vectors.

The following additional assumption is sufficient for the influence of the tail part of the process to go to zero:

*Assumption 2 (A2)*

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

**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. 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. The assumptions cover the case in which $L$ is finite (so that A2 is trivially satisfied) but not known, and where an upper bound on $L$ is not known. Other cases in which A2 can be verified arise where $Z\gamma$ represents an infinite expansion of a multivariate function. In a time series context, suppose that $y, X$ are linear processes and each $Z_\ell$, $\ell = 1, \ldots \overline{m}$ is a stationary and invertible ARMA process; $\overline{m}$ is finite, but an unbounded number of lags of each $Z_\ell$ 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 $\sum_{i=1}^{m} \sum_{\nu=1}^{\infty} |\gamma_{i\nu}| < \infty$, so that A2 holds. If in addition $y$ is an ARMA process, then $\sum_{i=1}^{m} \sum_{\nu=k}^{\infty} |\gamma_{i\nu}| = O(exp(-k))$.

2.3 *Consistency and asymptotic normality of the OLS estimator*

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 a consistent estimator of $\beta$.

Proof: See Appendix 1.

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 $f(k) = \sum_{i=1}^{\infty} |\gamma_{k+i}|$.

Thus, depending on the rate of decay in the coefficients $\gamma_\ell$, we 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, and also applies in the time series example above, with stationary ARMA processes.

For the following theorem we define $\tilde{k}$ $(0 \leq \tilde{k} \leq k)$ as the number of excluded sample points (for example, those lost to lags).

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

Proof: See Appendix 1.

The weighting matrix $H = G_k^{-1}V_kG_k^{-1}$ can be estimated consistently by $\hat{\sigma}_\varepsilon^2\left(\frac{X'M_kX}{N-\tilde{k}}\right)$, where $\hat{\sigma}_\varepsilon^2 = (N-\tilde{k})^{-1}u'u$ can be shown to be a consistent estimator of $\sigma_\varepsilon^2$, 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 $(\hat{\beta}_k - \beta_0)'\hat{H}(\hat{\beta}_k - \beta_0)\xrightarrow{D}\chi_m^2$.

**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). 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 $\kappa$ ($\kappa \leq K$) 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 partition the parameter vector conformably. We treat the case in which $K < N$, i.e. fewer potential explanatory series than data points. and 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.[1] Next define a selection matrix $\Pi_{K\times\kappa}$ such that $C(\kappa) = C(K)\Pi$ is a $K \times \kappa$ matrix which

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

contains $\kappa$ of the $K$ eigenvectors: $\kappa$ will be the number of control regressors included in the model (we discuss the choice of $\kappa$ below).[2]

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 $\beta$ 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 $S(K, K)$. We consider the OLS estimator with $\kappa$ orthogonalized regressors, 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 $S(\kappa, K) = [S_1, \ldots, S_\kappa]$.

Different methods of selection of the elements (columns) of $S(\kappa, K)$ from those of $S(K, K)$ are of course possible. If $\Pi$ selects the eigenvectors corresponding with the $\kappa$ largest eigenvalues, then $S(\kappa, K)$ contains the first $\kappa$ principal

[2]Each column of $\Pi$ 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 $\Pi_{ij'} = 0\ \forall j' \neq j$.

components of $Z(K)$. An alternative, where we take $X$ to be a single vector ($m = 1$), is to choose the $\kappa$ 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 $X$.

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 $\beta$ 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 $\beta$ in one regression, or component-by-component in a sequence of $m$ separate control regressions for each individual $\beta_i$. 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 are the $\kappa$ most important regressors to include from the set $S(K,K))$, we provide 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 $\kappa$ of the regressors have been selected, and consider the impact of adding $S_\nu$ to this set. Given $X$ and the vector $\beta$ of parameters of interest, $S_\nu$ has more impact the larger is the change in the estimate of $\beta$ :

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

where $\hat{\beta}_\kappa$ is the vector of regression coefficients obtained when $S(\kappa, K)$ is the matrix of $\kappa$ initially-included orthogonalized regressors, and $\hat{\beta}_{\kappa,\nu}$ is the estimate on a set of orthogonalized regressors which also includes $S_\nu$ 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 consider the following choices:

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

We examine $\Delta_{\kappa,\nu}$ and show that it can be decomposed into two random functions, $\psi_1$ and $\psi_2$, one of which ($\psi_2$) 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 $\kappa$ 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 $S_\nu$ on the regressors in $E_\kappa$. 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

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

As $N \to \infty$, $\psi_1$ becomes a good approximation to $\Delta_{\kappa,\nu}$; therefore our selection criteria will exploit $\psi_1$.

**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 $\kappa$ and any choice of selected regressors $S(\kappa, K)$ and $S_\nu$. 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 $\theta_\nu$ is unknown, 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 $S_\nu$ on $e_{(1)}, \ldots, e_{(\kappa)}$ and is equivalent to the selection criterion $R^2_\nu(\kappa)\hat{\lambda}^2_\nu$, where $R^2_\nu(\kappa)$ 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 $E_\kappa$ in the general case of $m$ effects of interest. Denote the correlation between $x$ and $S_j$, $j = 1, \ldots \kappa$, by $\rho_j$. 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

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

respectively. Since the denominator does not depend on $S_\nu$, either of these criteria reduce further 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 $S_\nu$ (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 $S_{\nu_1}$ is ordered before $S_{\nu_2}$, i.e. $\nu_1 < \nu_2$, if

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

We now show in Theorem 5 that, with the selection rule (3.12) for the ordering of the $K$ orthogonalized regressors, consistent estimation can be based on a subset of $\kappa < K$ of these regressors, where $K$ satisfies the conditions of Theorem 2. 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 $\Omega$ satisfy 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.

Note that selecting $\kappa = K - O(f(K))K^{\frac{1}{2}}$ ensures that $\sup_\Omega |\hat{\beta}_\kappa - \beta_0|$ converges at the same rate as $|\hat{\beta}_K - \beta_0|$. Thus we have shown that uniformly over $\Omega$, even in the most unfavorable cases, consistency obtains for a reduced dimension $\kappa$.

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 Akaike information criterion for choice of $\kappa$, as well as providing information on the finite-sample performance of the methods.

We close this section with a summary of the method of implementation of reduced-dimension control regression in the models (2.3)–(3.2):

- 1. From the $(N \times K)$ matrix $Z(K)$ of data on the controls, compute the eigenvalues $(\hat{\lambda}_i)$ and corresponding eigenvectors of the moment matrix $Z(K)'Z(K)$.
- 2. Order the eigenvectors by the product $\hat{\lambda}_i |\hat{\rho}_i|$, where $\hat{\rho}_i$ is the sample correlation between the eigenvector and $x$.
- 3. Select a large $\overline{\kappa}$ (e.g. 12, 20) and for each $\kappa = 1, \ldots, \overline{\kappa}$, form the matrix of eigenvectors $C(\kappa)$ from the first $\kappa$ eigenvectors as ordered in step 2, and then the $N \times \kappa$ matrix $S(\kappa, K) = Z(K)C(\kappa)$ of orthogonal components. For each $\kappa$, use the corresponding $S(\kappa, K)$ to estimate the regression model, and compute the AIC.[3]

[3]The AIC is defined as $\ln \hat{\sigma}^2_{[\kappa]} + \frac{2h}{N}$, where $\hat{\sigma}^2_{[\kappa]}$ is the estimated residual variance

- 4. Select the value of $\kappa$ that yields the lowest AIC. (If this occurs at or near the upper bound $\overline{\kappa}$, increase $\overline{\kappa}$ and compute further estimates to ensure a global minimum AIC). The regression with this value of $\kappa$ is the chosen control regression and the values of $\hat{c}$ and $\hat{\beta}$ are the final estimates.

**4. Finite-sample evaluation of bias and RMSE reduction**

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). We note therefore that these simulations have the unusual property that, while parameters governing the random number generation and other elements of structure are chosen by the investigator, the results are largely determined by the randomly-generated parameters. By averaging across many such parameter combinations, we expect more representative results than would be possible through investigation of a few selected cases.

There are two classes of cases treated here. In the first, all observable potential explanatory factors have at least some degree of relevance to the DGP, so that as $N \to \infty$ all should be selected into the model. In the second class, there exist two orthogonal groups of regressors; a second group is added which has no explanatory power, and its elements are therefore irrelevant as statistical controls.

The overall DGP for the first set of simulations is as follows: $N = 200$ 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 $\kappa$ 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

---

for the regression model with $\kappa$ columns in the matrix of orthogonal components, $S_i.$, $h$ is the total number of regressors ($= \kappa + 2$ here) and $N$ is sample size.

such that the $h$th 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 $\alpha$ 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 $\Gamma$ 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 $\kappa$.

The effect of the decay parameter is to vary the average relative importance of effects within the set of $Z$'s: with $\alpha$ near unity, each of the $Z_i$'s has an expected absolute coefficient near 1, and as $\alpha$ falls, the importance of coefficients other than the first few is reduced correspondingly. For relatively low values of $\alpha$, 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 $\alpha^j$ 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 factors. By contrast, for $\alpha$ 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 the factors 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 $\alpha$ such as 0.8 or 0.9. We therefore emphasize these values in the experiments.

The simulations corresponding with the second class of cases mentioned above uses a structure similar to i)–vi), but with a matrix of additional orthogonal variables available as potential explanatory factors. Elements i) and iv) above are modified, to:

- i′) $\dim Z = 20,\ \dim Z_2 = 20; \dim[Z : Z_2] = K = 40;\ \kappa = 1, \ldots, 20$
- iv′) $Z = Z_0\Gamma,\ \{Z_0\}_{ij} \sim N(0,1);\ Z_2 = Z_{20}\Gamma,\ \{Z_{20}\}_{ij} \sim N(0,1)$ :

that is, of the potential explanatory factors, only half are now in fact relevant to the DGP. The generation of $y_i, x_i$ as in v-vi above remains constant in that only elements of $Z$ are relevant.

Estimation of $\beta$ (=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 $\kappa$ eigenvalues) and by the product of eigenvalue and absolute value of the correlation with $x$ (3.12), as described above (labelled 'alternative eigenvector selection' in the figures). The selection of regressors from $Z$ is included for comparison: this represents a mechanical implementation of a standard regression method. The selection of regressors is determined by choosing ten random combinations of $\kappa$ 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 $\kappa$ 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 $\kappa$ are recorded in Figures 1a-d and 2a-d.[4] 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 $\beta$) 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. Figure 1 records the 'Class 1' cases given by i)-vi) above, and Figure 2 the 'Class 2' cases given by substituting i') and iv').

Clearly, even a small number of terms in the auxiliary model produces a substantial bias and RMSE reduction, and bias is typically very close to zero with approximately eight terms. The effect of bias clearly dominates the RMSE; increase in variance with $\kappa$ is small (that is, the RMSE does begin to increase for large $\kappa$, but the effect is so small as to be hard to detect in the figures).

In the first class of cases, 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 $\kappa$ are clearly sub-optimal in general, and that in the region of interest the orthogonalized regressor methods are clearly superior on both criteria.

In the second class of cases (where some of the potential orthogonalized re-

[4]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, $\alpha = 0.95$.

gressors are irrelevant), the overall pattern of results is very similar; however the magnitude of the gains produced by the selection method (3.12) which considers correlation with $x$ is larger, reflecting the greater gains available from focusing on a subset of potential regressors. With respect to RMSE, the rule (3.12) dominates all others. With respect to bias alone, as in the first class of cases, the relative performance of selection of untransformed regressors is better, although the alternative eigenvector method again comes to dominate for $\kappa$ sufficiently large ($> 6$).

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), 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.[5] [6]

Further experiments investigated cases specific to time series: one set of experiments analogous to example 2.3.3 above, and one in which lags of $y$ enter the process and are either included as part of the set of potential explanatory factors $Z$, or are treated specially and used as separate regressors. Results are not recorded as they follow similar patterns to those in Figures 1 and 2; in particular, separation of the dynamic effects (lagged $y'$s) from the other $Z'$s produced little difference in the results.

## 5. An illustrative application

Although one expects the effect of an increased interest rate (cost of capital)

---

[5]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.

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

on production of a capital good such as housing to be negative, the simple regression of housing starts on short-term real interest rates (i.e., those most directly subject to influence by monetary authorities) may show a positive estimated impact of the real interest rate. Using a long-term real interest rate (a rate which may be of particular interest to house purchasers), the point estimate is negative in our data, but not significant at conventional levels. Of course, housing starts are likely to be affected by a large number of other economic factors which affect the household's ability to purchase, sense of prosperity, etc., many of which will be correlated with interest rates. Controlling adequately for this large number of potential factors might be expected to produce substantially improved estimates of the true effect of real interest rates on housing starts.

Here we use a sample of US data from 1959:01 through 1999:12 ($N = 492$), with estimation on the smaller sample beginning in 1961:1 ($N - \tilde{k} = 468$) to allow for lags. The dependent variable is the total number of housing starts, seasonally adjusted, farm and non-farm; the parameter of interest is the effect of an interest rate on this quantity. The data set $Z$ of potentially explanatory factors includes not only the interest rate data to be described below, but also 65 additional series containing measures of macroeconomic quantities related to industrial production, personal income, price and wage indices, real sales and consumption, and employment and hours. Lags of these series or transformations are also used, for a total of 195 potential explanatory series. It is from this group of 195 series that eigenvectors will be extracted for statistical controls.

We use three real interest rate variables to investigate this relation: the Federal Funds rate, three-month Treasury bill rate, and five-year treasury bill rate. In each case we compute an ex-post real rate from the annualized data by subtracting the previous twelve months' inflation rate, and investigate the impact of these rates on the conditional expectation of housing starts. In initial investigation of lag orders, we find that only the first lag of the real rate is typically significant at conventional levels, and that the coefficients of the concurrent and first lagged real rates are approximately equal and opposite. In the results reported here, we therefore report the coefficient on the difference, $\Delta r_i = r_i - r_{i-1}$. Table 1 reports the results of the univariate regression $h_i = c_0 + \beta \Delta r_i + \epsilon_i$ and the regression with orthogonal component controls, with number of controls chosen by AIC and selection by the alternative criterion (3.12), for each of the three choices of interest rate. Given this criterion and the AIC for selection of $\kappa$, model selection is automatic (the numbers of components selected are 3, 5 and 7 for the three orthogonal component regressions).

**Table 1**

**Regressions of housing starts on change in real interest rate**[7]

$N - \tilde{k} = 468$

| | Univariate | | | Orthogonalized controls[8] | | |
|---|---|---|---|---|---|---|
| Parameter | $\Delta r_{(1)}$ | $\Delta r_{(2)}$ | $\Delta r_{(1)}$ | $\Delta r_{(1)}$ | $\Delta r_{(2)}$ | $\Delta r_{(1)}$ |
| $\hat{c_0}$ | 1515 | 1516 | 1516 | 1424 | 1408 | 1427 |
| $(t-)$ | (101) | (101) | (101) | (79.4) | (76.5) | (79.7) |
| $\hat{\beta}$ | 45.7 | 23.1 | -43.8 | -20.3 | -45.6 | -79.0 |
| $(t-)$ | (2.02) | (0.87) | (-1.29) | (-0.96) | (-1.88) | (-2.58) |

All point estimates of effects are negative in the model with orthogonalized controls. With the long-term rate that one would expect to have the greatest effect on the housing market, the effect is significant at conventional levels; the three-month rate as well shows substantial evidence against the null of no effect.

## 6. 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. That is, a traditional difficulty in classical principal components regression concerns the interpretation of the coefficients, but this difficulty does not arise here because of our separation of the effect of interest from the set of data from which eigenvectors are extracted.

We show that consistent estimation of an effect of interest is possible under fairly general circumstances that do not require 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

[7] $\Delta r_{(1)}$ = Federal funds rate; $\Delta r_{(2)}$ = 3-month Treasury bill rate; $\Delta r_{(3)}$ = 5-year Treasury bill rate.

[8] Note that coefficients $\hat{\gamma}$ on the orthogonalized components are not reported.

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, these methods appear to have substantial utility.

## 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 $W_\ell$ 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 $\mu_{\nu_i}$ 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

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

(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 $E|\varepsilon_i| \leq (E(\varepsilon_i^2))^{\frac{1}{2}}$ by Jensen's inequality. Up to the constant, $\varepsilon_i = A'W,\ A = (1, -\beta, -\gamma)'$. Therefore $\varepsilon_t^2 = E(A'WW'A) \leq \|A\|\|\Sigma_W\| \leq \overline{\lambda}(\Sigma_W)$. Since $\|A\|$ is finite by part (i) of the Lemma, it follows that $E|\varepsilon_i| < \infty$. ∎

**Proof of Theorem 1.**

Consider

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

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 $\beta$, 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[9]

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

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

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

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

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

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

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

and since $\varepsilon_i$ 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 $\ell'$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

$$|b_\ell| \le \left(\frac{1}{N-\tilde{k}}X'M_kX\right)^{\frac{1}{2}}\left(\frac{1}{N-\tilde{k}}\sum_{j=k}^{N}(\sum_{i=1}^{\infty}Z_{k+i+j}\gamma_{k+i+j})^2\right)^{\frac{1}{2}}$$

$$\le O_p(1)\left(\frac{1}{N-\tilde{k}}\sum_{j=k}^{N}(\sum_{i=1}^{\infty}Z_{k+i+j}\gamma_{k+i+j})^2\right)^{\frac{1}{2}} = o_p(1),$$

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 $N\to\infty,\ k\to\infty,\ k^{-1}N\to\infty$, 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 $\ell$'th component $(N-k)^{\frac{1}{2}}\ b_\ell$, where by (A2.8), $|b_\ell|\le O_p(1)\sum_{i=1}^{\infty}|\gamma_{k+i}|$. 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 $\varepsilon$ is independent of $(X,Z)$ then $G_k^{-1}V_kG_k^{-1} = \sigma_\varepsilon^2E\left(\frac{X'M_kX}{N-\tilde{k}}\right)$. ■

**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 $S_\nu$, 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 $\kappa$ included orthogonal regressors. Consider a regression of $S_\nu$ 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 $M_\kappa, S_\nu$, $Pr(A_\kappa(\nu)'A_\kappa(\nu) < \overline{A}) \to 1$ 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 $X_\ell$ 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 $X_\ell$ 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 $E_\ell$, $E(E_\ell^2) = E(E_\ell - \alpha_{\overline{\mu}_\ell}(X_\ell)\overline{\mu}_\ell)^2 + \alpha_{\overline{\mu}_\ell}(X_\ell)^2$. 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

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

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

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

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

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

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 $\Omega$,

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

Rewrite $(X'M_kX)^{-1}X'S(\kappa+1,K)\theta(\kappa+1,K) =$

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

Note that $|\theta_\nu| \le \|\gamma\|$ 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 $M_\kappa$

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

Thus

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

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 $\varepsilon$ 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.

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.

Cook, R.D. and B. Li (2002) Dimension reduction for conditional mean in regression. *Annals of Statistics* 30, 455-474.

Cook, R.D. and S. Weisberg (1991) Discussion of 'Sliced inverse regression for dimension reduction'. *Journal of the American Statistical Association* 86, 328–332.

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.

Greenberg, E. (1975) Minimum variance properties of principal component regression. *Journal of the American Statistical Association* 70, 184-197.

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.

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

Li, K.C. (1991) Sliced inverse regression for dimension reduction (with discussion) *Journal of the American Statistical Association* 86, 316-342.

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

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.

Xia, Y., Tong, H., Li, W.K., and L.-X Zhu (2002) An adaptive estimation of dimension reduction space. *Journal of the Royal Statistical Society Ser. B* 64, 363-410.

Zhu, X.L and K.-T. Fang (1996) Asymptotics for kernel estimate of sliced inverse regression. *Annals of Statistics* 24, 1053-1068.