Statistica Sinica **26** (2016), 1411-1426
doi:http://dx.doi.org/10.5705/ss.2014.211t

MULTIVARIATE STOCHASTIC REGRESSION IN TIME SERIES MODELING

Tze Leung Lai and Ka Wai Tsang

Stanford University

Abstract: This paper begins with a brief review of multivariate time series analysis, covering canonical correlation analysis and scalar components of vector ARMA models, pioneered by Tiao and his collaborators, and vector ARMAX models in linear systems theory. It then presents a fast stepwise regression procedure that includes parsimonious variable selection followed by rank selection in stochastic regression models. The procedure overcomes a long-standing difficulty with parameter estimation in these models, the dauntingly large number of parameters in the matrix of regression coefficients relative to the sample size n. Recent attempts to address this difficulty have used group lasso and hard thresholding of small singular values to take advantage of coefficient and rank sparsity. However, the underlying theory is based on non-random or independent regressors, whereas the procedure and its underlying theory developed herein are applicable to stochastic regressors in multivariate time series models.

Key words and phrases: Multivariate stochastic regression, orthogonal greedy algorithm, rank selection, sparsity, time series.

1. Introduction

Multivariate time series analysis is one of Professor Tiao's major areas of research, to which he has made many seminal contributions. According to Peña and Tsay (2011, p.420), he "got interested in multiple time series actually through canonical correlation analysis" of multivariate data in economics, since "economic data are mostly time series and people are obviously interested in dynamic relationships" among the multiple time series. His work on canonical correlation analysis started with the question of how autocorrelations or nonstationarity of economic time series data would affect traditional principal component analysis that assumes i.i.d. observations. "Now one way is to do principal component analysis and another one is to think about transformation and try to explain the relationship with the past," leading to the canonical analysis (Peña and Tsay (2011, p.417)). Whereas principal component analysis is often used as a dimension-reduction technique that attempts to approximate a p-dimensional random vector by r linear combinations (factors), with r considerably smaller

than p, canonical correlation analysis is used by Tsay and Tiao (1985) for identification of VARMA (vector autoregressive moving average) models, and by Tiao and Tsay (1989) to determine the orders of the matrix polynomials in a VARMA process "and, more importantly to reveal the unobserved underlying structure of a given vector process" through scalar components of a linear transformation of the original process.

In this paper we apply canonical correlation analysis, or more precisely, the closely related methodology of reduced rank regression to estimate the coefficient matrix $\mathbf{B} = (B_{ji})_{1\le j\le p, 1\le i\le q}$ in the stochastic regression model

$$\mathbf{y}_t = \mathbf{B}^T\mathbf{x}_t + \boldsymbol{\epsilon}_t, \quad t = 1, \ldots, n, \tag{1.1}$$

where $\mathbf{y}_t = (y_{t1}, \ldots, y_{tq})^T$ is the observed output vector at time t, $\mathbf{x}_t = (x_{t1}, \ldots, x_{tp})^T$ is assumed to be $\mathcal{F}_{t-1}$-measurable and $\boldsymbol{\epsilon}_t = (\epsilon_{t1}, \ldots, \epsilon_{tq})^T$ represents an unobservable random error vector that is assumed to form a martingale difference sequence. In particular, this includes the vector autoregressive VAR(1) model with $\mathbf{x}_t = \mathbf{y}_{t-1}$. With this choice of $\mathbf{x}_t$, Box and Tiao (1977, pp.355-357) introduce a predictability measure of $\mathbf{y}_t$ by $\mathbf{x}_t$ via a linear transformation $\mathbf{m}^T\mathbf{y}_t$. In particular, for univariate y_t, they define the predictability of y_t from $\mathbf{x}_t$ in (1.1) by the squared correlation coefficient $\mathrm{Var}(\mathbf{B}^T\mathbf{x}_t)/\mathrm{Var}(y_t)$. Using $\mathbf{m}^T\mathbf{y}_t$ to transform multivariate $\mathbf{y}_t$ to the univariate case leads them to consider how $\mathbf{m}$ should be chosen, and the choice that maximizes the predictability corresponds to the first canonical pair $(\boldsymbol{\alpha}^T\mathbf{x}_t, \mathbf{m}^T\mathbf{y}_t)$ in canonical correlation analysis; see Reinsel and Velu (1998, Sec. 2.3 and 2.4) and Lai and Xing (2008, Sec. 9.1 and 9.2), where the relationship between canonical correlations and reduced rank regression is also described. For general VAR(k) models, we can also express them, or their linear transformations (such as differenced versions to handle unit-root nonstationarity), as stochastic regression models (1.1). However, since our interest is in reduced rank regression, such representation may not be useful as the rank constraints reflecting the underlying dynamics should be placed on certain linear combinations of the parameter matrices; see Lai and Tsang (2014).

Besides applications to economics, multivariate stochastic regression models (1.1) are also important in control engineering, in which they take the form of MIMO (multiple input, multiple output) systems, also called multivariate ARX (autoregressive with exogenous inputs) models:

$$\mathbf{y}_t + \mathbf{A}_1\mathbf{y}_{t-1} + \cdots + \mathbf{A}_k\mathbf{y}_{t-k} = \mathbf{B}_d\mathbf{u}_{t-d} + \cdots + \mathbf{B}_h\mathbf{u}_{t-h} + \boldsymbol{\epsilon}_t, \tag{1.2}$$

in which $d \ge 1$ represents the delay; see Goodwin and Sin (1984), Caines (1988), and Hannan and Deistler (1988). The MIMO system again contains a large number of parameters and a model selection to come up with a parsimonious model

has attracted much attention in this area. Model selection in the engineering literature has primarily focused on determining the order (k, h) of the model; see Huang and Guo (1990) for a review. One can clearly reduce the number of parameters even more substantially by putting rank constraints on the $\mathbf{A}_i$ and $\mathbf{B}_j$ and making use of certain sparsity features of these matrices.

For the multivariate regression model (1.1), Bunea, She, and Wegkamp (2011, 2012) recently studied estimation of $\mathbf{B}$ under a low rank constraint and sparsity (in the sense of many zero entries) of $\mathbf{B}$, when $\mathbf{x}_t$ are nonrandom and ϵ_{ti} are i.i.d. sub-Gaussian. Letting $\mathbf{X} = (\mathbf{x}_1, \ldots, \mathbf{x}_n)^T \in \mathbb{R}^{n\times p}$ and $\mathbf{Y} = (\mathbf{y}_1, \ldots, \mathbf{y}_n)^T \in \mathbb{R}^{n\times q}$, they proposed a rank selection criterion (RSC) to estimate the rank of $\mathbf{B}$ by counting the number of singular values of the matrix $\mathbf{X}\widehat{\mathbf{B}}^{OLS}$ that exceed some threshold $H > 0$ that depends on q, $\min(p, n)$, and the common variance σ^2 of the ϵ_{ti}, where $(\mathbf{X}^T\mathbf{X})^-$ is the Moore-Penrose inverse of $\mathbf{X}^T\mathbf{X}$ and $\widehat{\mathbf{B}}^{OLS} = (\mathbf{X}^T\mathbf{X})^-\mathbf{X}^T\mathbf{Y}$ is the least squares estimator. They showed the RSC estimator $\hat{r}$ to be consistent even when either p or q, or both, may grow faster than n. Making use of $\hat{r}$, they proposed to estimate $\mathbf{B}$ by

$$\widehat{\mathbf{B}}^{RCGL} = \arg \min_{\mathrm{r}(\mathbf{B})\le\hat{r}} \{\|\mathbf{Y} - \mathbf{X}\mathbf{B}\|_F^2 + \lambda\|\mathbf{B}\|_{2,1}\}, \tag{1.3}$$

where $\mathrm{r}(\mathbf{B})$ denotes the rank of $\mathbf{B}$, $\|\mathbf{A}\|_F$ is the Frobenius norm $\big(\mathrm{Tr}(\mathbf{A}^T\mathbf{A})\big)^{1/2}$, and $\|\mathbf{B}\|_{2,1}$ is the sum of the Euclidean norms of the rows of $\mathbf{B}$. Assuming $\log p = o(n)$ and additional conditions on $\mathbf{X}^T\mathbf{X}/n$, they showed that their rank-constrained group lasso (RCGL) estimator $\widehat{\mathbf{B}}^{RCGL}$ is consistent in the sense that $\|(\mathbf{X}\widehat{\mathbf{B}}^{RCGL} - \mathbf{X}\mathbf{B})/(nq)\|_F \xrightarrow{p} 0$. These results rely heavily on the fact that the $\mathbf{x}_t$ are nonrandom or, more generally, that the $\mathbf{x}_t$ are independent of $\boldsymbol{\epsilon}_t$ so that we can condition on $\mathbf{x}_t$. Their argument cannot be extended to stochastic regressors $\mathbf{x}_t$ that depend on the information set $\mathcal{F}_{t-1}$ involving past observations. Moreover, the assumption of i.i.d. sub-Gaussian ϵ_{tj} is also overly restrictive in applications to economics and engineering.

The theory for lasso and its variants, including group lasso on which their results are based, has been developed only for nonrandom $\mathbf{x}_t$ and it is difficult to extend the theory to stochastic regressors. Ing and Lai (2011) have proposed an alternative approach, called the *orthogonal greedy algorithm* (OGA), by using forward stepwise regression in conjunction with a high-dimensional information criterion for sequential variable selection. They have shown that its asymptotic properties are comparable to lasso and that it also performs favorably in simulation studies. More importantly, because the procedure involves stepwise least squares regression instead of convex optimization for penalized least squares with the L_1-penalty, the theoretical analysis can be extended to stochastic regression models, as has recently been shown by Ing and Lai (2014). After extending OGA

in Section 2.2 to multivariate stochastic regression models, leading to group OGA as an alternative to group lasso, we develop in Section 2.3 an information criterion for rank selection in multivariate stochastic regression models, providing an alternative to RSC whose theory requires nonrandom regressors and i.i.d. sub-Gaussian ϵ_{tj} with estimable common variance. Section 2.4 presents an integrated asymptotic theory of group OGA followed by the new rank selection procedure in multivariate stochastic regression models satisfying coefficient and rank sparsity, as defined in Section 2.1. Simulation studies illustrating the finite-sample performance of the proposed method are given in Section 3. Section 4 concludes with further discussion and applications.

2. A Parsimonious Approach to Multivariate Stochastic Regression

2.1. Coefficient and rank sparsity: 2-stage parsimonious procedure

Although the apparent number of parameters in multivariate regression (1.1) may be daunting when p and q are large, we may not need a correspondingly large sample size n if the coefficient matrix $\mathbf{B}$ is sparse so that it has a relatively small effective number of parameters. Bunea, She, and Wegkamp (2012) have noted that two types of sparsity are often assumed in multivariate linear regression, which they call *rank sparse* and *row sparse* regression models. Rank sparsity means that $\mathbf{B}$ can be approximated by a low-rank matrix. If the rank of $\mathbf{B}$ is r, then by the singular value decomposition, we only need to estimate $r(p+q-r)$ (which can be much smaller than pq) parameters of $\mathbf{B}$. Row sparsity, here also called *coefficient sparsity*, means that only a small number of rows of $\mathbf{B}$ (or columns of $\mathbf{B}^T$ in (1.1)) are non-zero, that is, few predictors in $\mathbf{x}_t$ are relevant to the output vector (i.e., have non-zero regression coefficients for some components of $\mathbf{y}_t$). Assuming both coefficient sparsity and rank sparsity for $\mathbf{B}$, we propose a 2-stage procedure to estimate $\mathbf{B}$ in the multivariate stochastic regression model (1.1).

1. Use the group orthogonal greedy algorithm described in Section 2.2 to select a subset S of the p predictors, and calculate from $\mathbf{Y}$ and $\mathbf{X}_S$ the least squares estimator $\widehat{\mathbf{B}}^{GOGA}$, where $\mathbf{X}_j$ denotes the jth column vector of $\mathbf{X}$ and $\mathbf{X}_J$ denotes the submatrix of $\mathbf{X}$ consisting of the column vectors $\mathbf{X}_j$, $j \in J \subset \{1,\ldots,p\}$.
2. Apply the information criterion in Section 2.3 to obtain the rank estimator $\hat{r}$ of $\mathbf{B}$. Then carry out reduced rank regression (Lai and Xing (2008, p.204)) to obtain the final estimator $\widehat{\mathbf{B}} = \widehat{\mathbf{B}}^{GOGA}\mathbf{P}_{\hat{r}}$, where $\mathbf{P}_{\hat{r}} = \mathbf{V}_{\hat{r}}\mathbf{V}_{\hat{r}}^T$ and $\mathbf{V}_{\hat{r}} = [\mathbf{v}_1,\ldots,\mathbf{v}_{\hat{r}}]$ consists of the right singular vectors of $\widehat{\mathbf{Y}} = \mathbf{X}_S\widehat{\mathbf{B}}^{GOGA}$.

In contrast to the two-stage procedure of Bunea, She, and Wegkamp (2011, 2012) described in Section 1, our method selects a subset of relevant variables before estimating the rank of $\mathbf{B}$. This circumvents the difficulties of their RSC in the case $p > n$ and unknown σ. Another difference is that we use the group orthogonal greedy algorithm instead of their group lasso and also a different criterion for rank selection, for which the theory can be extended to stochastic regression models.

2.2. Group orthogonal greedy algorithm

In this section we develop a modification of the orthogonal greedy algorithm (OGA) introduced by Ing and Lai (2011, 2014). This modification, called GOGA (group orthogonal greedy algorithm), is similar to the modification of lasso by group lasso (Yuan and Lin (2006)). Let $\widehat{I}_k$ be the set of indices of the input variables selected by GOGA after k iterations. The QR decomposition (Lai and Xing (2008, p.6)) of $\mathbf{X}_{\widehat{I}_k}$ can be used to implement GOGA as follows. First initialize with $\mathbf{U}^0 = [\mathbf{U}_v^0]_{v=1}^q = \mathbf{Y}$, $\widehat{I}_0 = \emptyset$, and empty matrices $\mathbf{Q}_0$ and $\mathbf{R}_0$. For $k = 1$ to m do:

1. choose $\hat{i}_k = \arg\min_{1\le i\le p} \left(\min_{\boldsymbol{\beta}\in\mathbb{R}^q} ||\mathbf{U}^{k-1} - \mathbf{X}_i\boldsymbol{\beta}^T||_F^2\right)$;
2. update $\widehat{I}_k = \widehat{I}_{k-1} \bigcup \{\hat{i}_k\}$ and compute the QR decomposition
$$\mathbf{X}_{\widehat{I}_k} = [\mathbf{X}_{\widehat{I}_{k-1}}\ \mathbf{X}_{\hat{i}_k}] = \left[\mathbf{Q}_{k-1}\ \mathbf{q}_k\right] \begin{bmatrix} \mathbf{R}_{k-1} & \vdots \\ 0\cdots 0 & r_k \end{bmatrix} = \mathbf{Q}_k\mathbf{R}_k;$$
3. update $\mathbf{U}^k = \mathbf{U}^{k-1} - \mathbf{q}_k\boldsymbol{\beta}_k^T$, where $\boldsymbol{\beta}_k^T = \mathbf{q}_k^T\mathbf{U}^{k-1}$;
4. End for, with $\hat{i}_k$th row of $\widehat{\mathbf{B}} \in \mathbb{R}^{p\times q}$ equal to the kth row of $\mathbf{R}_m^{-1}[\boldsymbol{\beta}_1\cdots\boldsymbol{\beta}_m]^T$ and the other rows equal to $\mathbf{0}^T$.

Here $\mathbf{R}_m^{-1}[\boldsymbol{\beta}_1\cdots\boldsymbol{\beta}_m]^T$ can be computed by backward substitution without calculating the inverse of the upper triangular matrix $\mathbf{R}_m$, and the QR decomposition is used to implement forward stepwise regression, instead of sequentially orthogonalizing the input variables as in Section 2.2 of Ing and Lai (2011). Therefore at every stage GOGA chooses the input variable that yields the largest reduction in the squared Frobenius norm of the residual matrix, hence the adjective "greedy" in its name.

Let $m = K_n$ be a prescribed upper bound on the number of GOGA iterations. The convergence theory in Section 2.4 suggests terminating the GOGA iterations after $K_n = O(\{n/\log(p_nq_n)\}^{1/2})$ steps or, equivalently, after K_n input variables have been included in the regression model. As in Ing and Lai (2011) for the case $q = 1$, we can further reduce the number of input variables along the GOGA path by using the "high-dimensional information criterion"

$$\mathrm{HDIC}(k) = n \log\left((nq_n)^{-1}\|\mathbf{U}^k\|_F^2\right) + kw_n \log(p_n q_n), \tag{2.1}$$

in which different criteria correspond to different choices of w_n. Here $\mathbf{U}^k$, and hence $\mathrm{HDIC}(k)$, can be readily computed at the kth GOGA iteration, and therefore selection of $\hat{k}_n = \arg\min_{1\le k\le K_n} \mathrm{HDIC}(k)$ along the GOGA path involves little additional computational cost. In particular, for $w_n = \log n$, $q_n\mathrm{HDIC}(k)$ corresponds to HDBIC, and $q_n\mathrm{HDIC}(k)$ with w_n equal to a constant c corresponds to HDAIC, since there are q_n univariate regressions in (2.1) and the number of parameters in $\mathbf{B}$, when there are k input variables, is kq_n. Thus, by using (2.1) in conjunction with the GOGA iterations, the first stage of the two-stage procedure in Section 2.1 ends with a selected set S consisting of $\hat{k}_n$ (instead of K_n) input variables.

2.3. Information criterion for rank selection

As noted in Section 2.1, we can use reduced rank regression to compute

$$\mathbf{B}(h) = \arg\min_{\mathbf{B}\in\mathbb{R}^{\hat{k}_n\times q}, \mathrm{r}(\mathbf{B})\le h} \|\mathbf{Y} - \mathbf{X}_S\mathbf{B}\|_F^2 \tag{2.2}$$

after obtaining the selected subset of $\hat{k}_n$ predictors by GOGA. Lai and Tsang (2014) propose to use the information criterion

$$\mathrm{IC}(h) = nq \log \hat{\sigma}^2(h) + hc(n+q)\log\left(\frac{nq}{n+q}\right) \tag{2.3}$$

to choose the rank $\hat{r} = \arg\min_h \mathrm{IC}(h)$, where $\hat{\sigma}^2(h) = \|\mathbf{Y} - \mathbf{X}_S\mathbf{B}(h)\|_F^2/(nq)$, and have shown $\hat{r}$ to be consistent for any choice of c in (2.3) under rank sparsity. They also propose a data-dependent choice to improve finite-sample performance. A natural modification of cross-validation for time series data is the accumulated predictive error (APE) criterion introduced by Rissanen (1986); see also Wei (1992). The idea is to choose c from a grid of values to minimize

$$\mathrm{APE}(c) = \sum_{t=m_0+1}^{n} \|\mathbf{y}_t - \mathbf{B}_{t-1}^T(\hat{r}(c))\mathbf{x}_t^S\|^2, \tag{2.4}$$

where the transpose of $\mathbf{x}_t^S$ is the tth row of $\mathbf{X}_S$, $\mathbf{B}_s(h)$ is the rank-h estimate based on $\{(\mathbf{x}_i^S, \mathbf{y}_i) : i \le s\}$, and m_0 is the starting sample size for which $\mathbf{B}_{m_0}(\hat{r}(c))$ is uniquely defined for all c belonging to the grid.

2.4. Asymptotic theory of 2-stage parsimonious procedure

We give an integrated asymptotic theory for GOGA and the rank selection criterion (2.3) in the proposed 2-stage procedure under assumptions (C1)−(C3) on the stochastic regression model (1.1) in which $p = p_n$ and $q = q_n$. Let

$$v_j = n^{-1}\mathbf{X}_j^T\mathbf{X}_j, \quad \mathbf{\Gamma}(J) = n^{-1}\mathbf{X}_J^T\mathbf{X}_J, \quad \mathbf{D}_J = \mathrm{diag}(\|\mathbf{X}_j\|, j\in J) \tag{2.5}$$

for $J \subset \{1, \ldots, p_n\}$, where we use $\|\mathbf{x}\|$ to denote the Euclidean norm $(\sum_i x_i^2)^{1/2}$, $\|\mathbf{x}\|_1$ to denote the ℓ_1-norm $\sum_i |x_i|$, and $\lambda_{\min}(\cdot)$ to denote the minimum eigenvalue of a symmetric matrix.

(C1) $$\max_{1 \le \#(J) \le K_n} \lambda_{\min}^{-1}(\mathbf{\Gamma}(J)) = O_p(1),$$

$$\max_{1 \le \#(J) \le K_n, i \notin J} \frac{\|\mathbf{\Gamma}^{-1}(J)\mathbf{D}_J^{-1}\mathbf{X}_J^T\mathbf{X}_i\|_1}{\|\mathbf{X}_i\|} = O_p(1).$$

(C2) $\sup_{n \ge 1} \max_{1 \le i \le q_n} \sum_{j=1}^{p_n} |B_{ji}| \sqrt{v_j} = O_p(1)$, $\log(p_n q_n) = o(n)$.

(C3) $\boldsymbol{\epsilon}_t$ is independent of $\mathcal{F}_{t-1}$ and $\sup_{n \ge 1} \max_{1 \le i \le q_n, 1 \le t \le n} E\{\exp(\theta \epsilon_{ti})\} < \infty$ for $|\theta| \le \theta_0$.

First consider the case $q = 1$ and i.i.d. $\mathbf{x}_t$, for which $\mathbf{B}$ is a $p \times 1$ vector (therefore no low-rank approximation is needed) and simple OGA can be used. Ing and Lai (2011) have assumed (C3) together with 'population versions' of (C1) and (C2) that replace v_j, $\mathbf{\Gamma}(J)$, $\mathbf{D}_J^{-1}\mathbf{X}_J$ and $\mathbf{X}_i/\|\mathbf{X}_i\|$ by their population values, together with the additional condition

$$\limsup_{n \to \infty} \max_{1 \le j \le p} E\{\exp(\delta u_{tj}^2)\} < \infty, \text{ for some } \delta > 0, \tag{2.6}$$

where $u_{tj} = x_{tj}/\sigma_j$ and $\sigma_j^2 = \mathrm{Var}(x_{tj})$, recalling that $x_{1j}, \ldots, x_{nj}$ are assumed to be i.i.d. for every j. Basically they use (2.6) to prove the 'sample versions' (C1) and (C2) and also to analyze the conditional mean squared prediction error $\mathrm{CPE} = E\{(\hat{\mathbf{B}}^T\mathbf{x} - \mathbf{B}^T\mathbf{x})^2 | y_1, \mathbf{x}_1, \ldots, y_n, \mathbf{x}_n\}$. They show that for OGA that terminates after m iterations, the corresponding CPE that we denote by CPE_m, satisfies

$$\max_{1 \le m \le K_n} \frac{\mathrm{CPE}_m}{(m^{-1} + n^{-1} m \log p_n)} = O_p(1). \tag{2.7}$$

In (2.7), m^{-1} represents the order of the squared bias in using only m input variables, chosen along the OGA path, to enter the regression model with p_n input variables, and $n^{-1} m \log p_n$ represents the order of the squared bias. We can interpret $O(n^{-1})$ as the variance per regression coefficient and $O(\log p_n)$ as the variance inflation factor due to the data-dependent choice of the input variables; see Ing and Lai (2011, p.1478). As mentioned, the convergence rate result in (2.7) suggests choosing the proper trade-off between the squared bias and the variance along the OGA path by using the high-dimensional information criterion (2.1).

To extend the theory of OGA and HDIC to stochastic regression models, Ing and Lai (2014) work directly with assumptions (C1) and (C2), as assumption of the type (2.6) no longer generates exponential bounds for tail probabilities when

$\mathbf{x}_t$ are not independent, making it difficult to use moment generating functions of their sums. They use exponential or moment bounds for self-normalized martingales to extend the theory of OGA and HDIC to stochastic regressors $\mathbf{x}_t$ when the $\boldsymbol{\epsilon}_t$ satisfy (C3). In lieu of the conditional mean squared prediction error, they consider the empirical squared error and show that it satisfies an analog of (2.7):

$$\max_{1\le m\le K_n} n^{-1}\sum_{t=1}^{m}\frac{(\hat{\mathbf{B}}_m^T\mathbf{x}_t-\mathbf{B}^T\mathbf{x}_t)^2}{(m^{-1}+n^{-1}m\log p_n)}=O_p(1), \tag{2.8}$$

where $\hat{\mathbf{B}}_m$ denotes the OGA estimate after m iterations (i.e., based on m input variables included sequentially along the OGA path). To further extend to multivariate stochastic regression models, we include $\max_{1\le i\le q_n}$ in (C2) and (C3). Since $\mathbf{B}$ in (1.1) is now a $p_n\times q_n$ matrix instead of a $p_n\times 1$ vector, the empirical squared error is $(nq_n)^{-1}\|\mathbf{X}\hat{\mathbf{B}}-\mathbf{X}\mathbf{B}\|_F^2$, since multivariate regression basically involves q_n univariate multiple regressions, one for each output variable. The risk inflation factor is now $\log(p_nq_n)$, and this explains the condition $\log(p_nq_n)=o(n)$ in (C2) as an extension of the case $q_n=1$ considered by Ing and Lai (2011, 2014). The arguments used by Ing and Lai (2014) to prove (2.8) can be extended to show that for the GOGA estimate $\hat{\mathbf{B}}_m$ after m iterations,

$$\max_{1\le m\le K_n}(nq_n)^{-1}\frac{\|\mathbf{X}\hat{\mathbf{B}}_m-\mathbf{X}\mathbf{B}\|_F^2}{(m^{-1}+n^{-1}m\log(p_nq_n))}=O_p(1). \tag{2.9}$$

Making use of (2.9), Lai and Tsang (2014) have shown HDBIC to be consistent for strongly sparse models, for which there exists $0\le\gamma<1$ such that $n^\gamma=o\left(\{n/\log(p_nq_n)\}^{1/2}\right)$ and

$$\liminf_{n\to\infty} n^\gamma \max_{1\le i\le q_n}\left\{\min_{1\le j\le p_n, B_{ji}\ne 0}|B_{ji}|\sqrt{v_j}\right\}>0$$

in probability. This is an extension of the strong sparsity condition of Ing and Lai (2011) for the case $q_n=1$ and i.i.d. $\mathbf{x}_t$. Even if the strong sparsity does not hold, they show that under additional regularity conditions, GOGA with HDAIC can provide an asymptotically optimal set of input variables to yield a consistent rank estimator $\hat{r}$ via the information criterion (2.3), thereby establishing the asymptotic efficiency of the proposed 2-stage estimator of $\mathbf{B}$ under coefficient and rank sparsity.

3. Simulation Studies

3.1. Comparison of proposed procedure with RCGL

In this section, we report on simulation studies of the performance of the 2-stage parsimonious procedure, first in the case of i.i.d. $\mathbf{x}_t$ and then in the case of a VAR(1) model for the $\mathbf{x}_t$. For the i.i.d. case, we used the same setting as

that considered by Bunea, She, and Wegkamp (2012), in which the rows of the design matrix $\mathbf{X} \in \mathbb{R}^{n\times p}$ are i.i.d. and generated from a multivariate normal distribution $\mathrm{N}(\mathbf{0}, \mathbf{\Sigma})$, with $\Sigma_{jk} = \rho^{|j-k|}$ and $\rho = 0.5$. We took $n = 250, q = 10$ and considered $p = 100$ $(< n)$ and $p = 300$ $(> n)$. The first $p_0 = 15$ rows of the coefficient matrix $\mathbf{B}$ are "significantly nonzero", of the form $\mathbf{B}_0\mathbf{B}_1$ with $\mathbf{B}_0 \in \mathbb{R}^{p_0\times r}$, $\mathbf{B}_1 \in \mathbb{R}^{r\times q}$, $r = 3$ or 8, such that all entries in $\mathbf{B}_0$ and $\mathbf{B}_1$ are i.i.d. with the same distribution as $Z + 0.1 \cdot \mathrm{sign}(Z)$, where $Z \sim N(0, 1)$. The remaining $p - p_0$ rows of $\mathbf{B}$ are zero rows. The random errors ϵ_{tj} are standard normal. A simulation study of $\hat{\mathbf{B}}^{RCGL}$ in a similar setting has been carried out by Bunea, She, and Wegkamp (2012), whose rank selection criterion (RSC) estimates r by counting the number of singular values of $\mathbf{X}\hat{\mathbf{B}}^{OLS}$ that are greater than $\sigma\sqrt{2(q + \mathrm{r}(\mathbf{X}))}$, where $\mathrm{r}(\mathbf{X})$ is the rank of $\mathbf{X}$ and $\sigma^2 = 1$ is the common variance of ϵ_{tj}, that we assume to be known in applying RSC, following Bunea, She, and Wegkamp (2012) who need this assumption for the case $p > n$. We used $\lambda = C\sigma\sqrt{\lambda_1(\mathbf{X}^T\mathbf{X}/n)\hat{r}n(1 + \log(p))}$, suggested in Theorem 3 in Bunea, She, and Wegkamp (2012), for the regularization parameter for the lasso penalty. Here $\hat{r}$ is the rank estimate by RSC, $\lambda_1(\mathbf{X}^T\mathbf{X}/n)$ is the largest eigenvalue of $\mathbf{X}^T\mathbf{X}/n$, and C is a constant selected by cross-validation ranging from 0.1 to 10.

For each setting, a training set of n observations $(\mathbf{x}_t, \mathbf{y}_t)$ and an independent data set of size n to be used as test set were generated. For GOGA, we chose $K_n = \lfloor 10(n/(\log p_n q_n))^{1/2} \rfloor$ and selected the number of predictors between 1 and K_n by the information criterion (2.1) with $w_n \log(p_n q_n) = \log n$. Since $p_n q_n$ has the same order as n, this reduces (2.1) to the usual BIC. Table 1 gives the prediction accuracy measured by the mean squared error

$$\mathrm{MSE} = \frac{1}{nq}\|\mathbf{X}_{test}\mathbf{B} - \mathbf{X}_{test}\hat{\mathbf{B}}\|_F^2 \tag{3.1}$$

using the test data $\mathbf{X}_{test}$, and by the mean squared in-sample error

$$\mathrm{MSIE} = \frac{1}{nq}\|\mathbf{X}\mathbf{B} - \mathbf{X}\hat{\mathbf{B}}\|_F^2 \tag{3.2}$$

using the training data $\mathbf{X}$. Here MSIE is the same as the empirical squared error in Section 2.4. In addition, Table 1 gives the mean number of relevant predictors selected (denoted by 'correct'), the mean number of predictors selected (denoted by 'total'), and the average estimated rank (denoted by $\hat{r}$). Each result in Table 1 is based on 1,000 simulations. The table shows that RCGL and our proposed two-stage parsimonious method, abbreviated by 2SP, perform well in selecting all relevant variables and the rank of $\mathbf{B}$. However, RCGL, which is based on group lasso, selects other irrelevant predictors and gives biased estimates because of the penalty term. This explains why RCGL has larger MSE and MSIE than 2SP.

Table 1. Performance comparison between RCGL and 2SP for iid regressors, with $p = 100, 300$; $r = 3, 8$; $n = 250$, $p_0 = 15$ and $q = 10$. The standard deviation of each value is given in parentheses; 'correct' denotes the mean number of relevant predictors selected by a method and 'total' denotes the total mean number of predictors selected.

n	p	q	r	Method	MSE	MSIE	$\hat{r}$	correct	total
250	100	10	3	RCGL	0.06(0.01)	0.06(0.01)	3 (0)	15(0)	67.9(7.1)
				2SP	0.03(0.01)	0.03(0)	3 (0)	15(0)	15 (0)
			8	RCGL	0.14(0.02)	0.12(0.01)	7.9(0.3)	15(0)	73.6(4.2)
				2SP	0.07(0.04)	0.06(0.04)	7.9(0.4)	15(0)	15 (0)
250	300	10	3	RCGL	0.14(0.03)	0.12(0.02)	3 (0)	15(0)	35.7(4.8)
				2SP	0.03(0.01)	0.03(0.00)	3 (0)	15(0)	15 (0)
			8	RCGL	0.20(0.04)	0.16(0.03)	7.7(0.4)	15(0)	121.7(10.1)
				2SP	0.07(0.06)	0.06(0.05)	7.9(0.4)	15(0)	15 (0)

We next consider the performance of 2SP and RCGL in the stochastic regression model (1.1), in which

$$\mathbf{x}_{t+1} = \mathbf{A}\mathbf{x}_t + \mathbf{w}_{t+1} \tag{3.3}$$

is the VAR(1) model, with a random coefficient matrix $\mathbf{A}$ that satisfies some stability restriction to ensure the stationarity of $\mathbf{x}_t$, e.g., $\|\mathbf{A}\| < 0.9$, where $\|\mathbf{A}\|$ denotes the spectral norm $(\lambda_{\max}(\mathbf{A}^T\mathbf{A}))^{1/2}$. In particular, if we choose the distribution of $\mathbf{A}$ to be that of i.i.d. standard normal entries conditional on $\|\mathbf{A}\| < 0.9$, then $\mathbf{A}$ can be generated by rejection sampling that rejects a simulated sample of i.i.d. standard normal variables if they comprise a matrix whose spectral norm is ≥ 0.9. However, this rejection sampling scheme has high rejection rate and is very inefficient for high-dimensional $\mathbf{A}$. We address this difficulty by using an alternative distribution for $\mathbf{A}$ that can be generated in the following way. Again we start by generating a matrix $\mathbf{A}^0$ with i.i.d. standard normal entries. Let $\lambda_1^0, \ldots, \lambda_p^0$ be the eigenvalues of $\mathbf{A}^0$ and $\mathbf{V} = [\mathbf{v}_1^0, \ldots, \mathbf{v}_p^0]$ be the matrix consisting of the corresponding normalized eigenvectors. Define $\tilde{\lambda}_i = \lambda_i^0$ if $|\lambda_i^0| \leq 1$ and $\tilde{\lambda}_i = 1/\lambda_i^0$ otherwise. Let $\widetilde{\mathbf{A}} = (\tilde{a}_{ij})_{1\leq i,j\leq p} = \mathbf{V}\mathrm{diag}(\tilde{\lambda}_1^0, \ldots, \tilde{\lambda}_p^0)\mathbf{V}^-$ and $\mathbf{A}^1 = (\mathrm{Re}(\tilde{a}_{ij}))_{1\leq i,j\leq p}$, where $\mathbf{V}^-$ is the Moore-Penrose generalized inverse of $\mathbf{V}$. To ensure stationary we choose the distribution of $\mathbf{A}$ to be the conditional distribution of $\mathbf{A}^1$ given that $\max_{1\leq i\leq p} |\lambda_i^1| \leq 0.9$. Hence $\mathbf{A}$ can be generated by rejection sampling as before, but applied to $\mathbf{A}^1$, and the rejection rate is much lower than that applied to $\mathbf{A}^0$.

The coefficient matrix $\mathbf{B}$ in (1.1) was constructed in the same way as in the i.i.d. setting. We chose $n = 250, q = 10$, $p = 100$ or 300 as before. The random errors w_{ti} in $\mathbf{w}_t$ were assumed to be i.i.d. standard normal. The acceptance rate of the rejection sampling procedure (described in the preceding paragraph)

Table 2. Performance comparison between RCGL and 2SP for stochastic regressors, with $p = 100, 300$; $r = 3, 8$; $n = 250$, $p_0 = 15$ and $q = 10$. The standard deviation of each value is given in parentheses; 'correct' denotes the mean number of relevant predictors selected by a method and 'total' denotes the total mean number of predictors selected.

n	p	q	r	Method	MSE	MSIE	$\hat{r}$	correct	total
250	100	10	3	RCGL	0.59(1.44)	0.47(1.13)	3 (0)	15(0)	28.4(7.3)
				2SP	0.03(0.01)	0.03(0.01)	3 (0)	15(0)	15.9(1.8)
			8	RCGL	3.58(10.7)	3.07(9.8)	8 (0.2)	15(0)	22.1(5.7)
				2SP	0.07(0.03)	0.06(0.03)	7.9(0.3)	15(0)	15.9(1.8)
250	300	10	3	RCGL	2.35(5.5)	1.8 (4.2)	3 (0)	15(0.2)	34.9(13.6)
				2SP	0.04(0.17)	0.03(0.08)	3 (0)	15(0.1)	16.8(3.8)
			8	RCGL	10.8 (19.67)	9.17(17.2)	7.9(0.3)	15(0.1)	29 (13)
				2SP	0.09(0.61)	0.07(0.31)	7.9(0.2)	15(0.1)	16.8(4.1)

Table 3. Number of $\hat{r} \neq r$ in 10,000 simulations for RSC and IC.

n	p	q	r	Method	I.I.D.	VAR(1)
250	100	10	3	RSC	0	0
				IC	0	0
			8	RSC	31	14
				IC	61	32
250	300	10	3	RSC	0	0
				IC	0	0
			8	RSC	108	26
				IC	92	49

used to generate $\mathbf{A}$ was 0.72 in this simulation study. The prediction accuracy measured by MSE and MSIE and other performance measures are given in Table 2, each result of which is based on 1,000 simulations. The table shows that the performance of 2SP for stochastic regressors is similar to its performance for i.i.d. regressors, but the performance of RCGL is much worse. On the other hand, RSC still performs well in stochastic regression models, as shown by Table 3. The table, each result of which is based on 10,000 simulations, compares the performance of the rank estimate in Section 2.3 using IC with that of RSC that assumes $\sigma = 1$ to be known, and shows that both rank estimates perform well for both i.i.d. and stochastic regressors. In the next section we explain why this is the case by using the SVD (singular value decomposition) of $\mathbf{X}_S\hat{\mathbf{B}}^{GOGA}$.

3.2. Extensions to conditionally heteroscedastic errors

A stylized feature for many economic and financial time series is conditionally heteroscedastic errors; see Chapter 6 of Lai and Xing (2008). As noted by Lai and Tsang (2014), (C3) can be greatly relaxed if moment bounds are used instead

Table 4. Performance comparison between RCGL and 2SP for stochastic regressors with heteroscedastic errors, in the case $p = 100, 300$; $r = 3, 8$; $n = 250$, $p_0 = 15$ and $q = 10$. The standard deviation of each value is given in parentheses; 'correct' denotes the mean number of relevant predictors selected by a method and 'total' denotes the total mean number of predictors selected.

n	p	q	r	Method	MSE	MSIE	$\hat{r}$	correct	total
250	100	10	3	RCGL	1.15(3.29)	0.96(2.82)	3.07(0.26)	15 (0.2)	28 (9.2)
				2SP	0.08(0.07)	0.07(0.07)	3 (0)	15 (0.1)	15.9(2.1)
			8	RCGL	5.39(9.7)	4.68(8.76)	7.9 (0.3)	15 (0.1)	21.4(6)
				2SP	0.18(0.2)	0.17(0.17)	7.9 (0.4)	15 (0)	16 (2)
250	300	10	3	RCGL	3.64(6.73)	2.82(4.99)	3.1 (0.5)	14.9(0.3)	34.9(27.2)
				2SP	0.09(0.23)	0.08(0.14)	3 (0)	15 (0.2)	16.8(3.8)
			8	RCGL	15.1 (21.8)	13.1 (19.6)	7.7 (0.5)	14.9(0.4)	28.4(20.9)
				2SP	0.18(0.16)	0.17(0.15)	7.9 (0.4)	15 (0)	16.5(3.1)

of exponential bounds; in this case, we only require the random errors to be martingale differences satisfying certain assumptions for the asymptotic theory of 2SP. In particular, for random errors ϵ_{tj} that are generated by GARCH(1,1) model

$$\epsilon_{tj} = \sigma_{tj} z_{tj}, \quad \sigma_{tj}^2 = \omega + a\sigma_{t-1,j}^2 + b\epsilon_{t-1,j}^2, \tag{3.4}$$

for $1 \le j \le q$, where z_{tj} are i.i.d. $N(0,1)$, we have carried out a simulation study of the performance of 2SP, assuming that $\omega \sim \text{Unif}(0,1)$ and $a = b = 0.4$. The long-run variance of the GARCH errors is $\omega/(1-a-b)$; see Lai and Xing (2008, p.147). In case $p < n$, this can be estimated by $\hat{\sigma}^2 = \|\mathbf{Y} - \mathbf{X}\widehat{\mathbf{B}}^{OLS}\|_F/(q(n-p))$, as suggested by Engle and Mezrich (1996). The variance estimate $\hat{\sigma}^2$ is used by Bunea, She, and Wegkamp (2011) to implement RCGL when the assumed common variance σ^2 of ϵ_{tj} is unknown. For $p > n$, since this variance estimate is not applicable, we replace σ^2 by $\omega/(1-a-b)$ to implement RCGL for comparison with 2SP in the GARCH model (3.4) for the random errors. The prediction accuracy measured by MSE and MSIE and other performance measures similar to those in Tables 1 and 2 are given in Table 4, each result of which is based on 1,000 simulations. The table shows that both 2SP and RCGL can still choose all the relevant predictors and select the rank close to the actual rank, but that 2SP has much smaller MSE and MSIE than RCGL.

4. Discussion and Time Series Applications

There has been much recent interest in low-rank estimators of high-dimensional matrices in a variety of applications, ranging from matrix completion problems in web-based personalized recommendation systems to medical imaging and remote sensing; see Candès and Plan (2010), Negahban and Wainwright (2011),

Rohde and Tsybakov (2011), and the references therein. Penalized least squares are typically used, with penalty proportional to the nuclear norm of the parameter matrix, under coefficient and rank sparsity. Whereas these papers consider univariate outputs y_i whose means are $\mathrm{Tr}(\mathbf{B}\mathbf{x}_i^T)$ ($\mathbf{x}_i$ here can be a matrix) for an unknown parameter matrix, Bunea, She, and Wegkamp (2011, 2012) consider the case of multivariate $\mathbf{y}_i$, which is much closer in spirit to the multivariate time series models considered herein. However, these recent developments are restricted to nonrandom or independent regressors $\mathbf{x}_t$, and does not apply to the more general stochastic regressors commonly encountered in multivariate time series models.

We introduce a new approach to multivariate stochastic regression that takes advantage of coefficient and rank sparsity. Instead of using the rank-constrained group lasso, as in Bunea, She, and Wegkamp (2012), we use the group orthogonal greedy algorithm (GOGA) to enter input variables sequentially up to a stopping time and then to choose along the GOGA path the set of input variables with which we perform reduced rank regression, with the rank determined by the information criterion (2.3). Although (2.3) appears to be completely different from the RSC that uses the threshold $\sigma\sqrt{2(q+\mathrm{r}(\mathbf{X}))}$ to cut off smaller singular values of $\mathbf{X}\hat{\mathbf{B}}^{OLS}$ (see the first paragraph of Section 3), it is in fact operationally similar to RSC. To see its connection to RSC, rewrite (2.3) as

$$\mathrm{IC}(h) = nq\log\left(\mathrm{Tr}(\mathbf{Y}^T\mathbf{Y}) - \sum_{i=1}^{h}\mu_i^2\right) + hc(n+q)\log\left(\frac{nq}{n+q}\right),$$

in which the μ_i are singular values of $\mathbf{X}_S\hat{\mathbf{B}}^{GOGA}$. This implies that accepting a higher rank that is associated with a small μ_i only diminishes the first summand of $\mathrm{IC}(h)$ slightly but increases the second summand by $c(n+q)\log(nq/(n+q))$. Hence we can regard the criterion (2.3) as if there is a threshold below which the singular values of $\mathbf{X}_S\hat{\mathbf{B}}^{GOGA}$ are set to 0, similar to RSC which uses $\mathbf{X}\hat{\mathbf{B}}^{OLS}$ instead of $\mathbf{X}_S\hat{\mathbf{B}}^{GOGA}$.

The information criterion (2.3) also has some resemblance to, and in fact was inspired by, that introduced by Bai and Ng (2002) for determining the number of factors in $\mathbf{X}$. Stock and Watson (1999, 2002) have introduced factor models for $\mathbf{X}$ to forecast $\mathbf{Y}$. The basic idea is that although $\mathbf{x}_t$ (assumed to be stationary) may be high-dimensional, the covariance matrix may have a few dominant factors (principal components associated with the largest eigenvalues), leading Bai and Ng to develop an information criterion to estimate the number of factors. This information criterion was subsequently widely used in macroeconomic studies; see e.g., Bernanke, Bovian, and Eliasz (2005). Since the goal is to forecast $\mathbf{y}_t$, it seems more direct to estimate the number of nonzero singular values of the matrix

B instead. The advantage of our procedure over the dynamic factor modeling approach of Stock and Watson in macroeconomic studies are discussed in Lai and Tsang (2014). This advantage was already observed by Professor Tiao when he started working in multiple time series modeling. He commented further on the transformations associated with canonical analysis that he found to reveal more than one would expect from principal component analysis: "when you have all these nonstationary things that move in tandem, maybe there are only one or two underlying components that explain all the growth" (Peña and Tsay (2011, p.417)). His conjectural foresight subsequently materialized in the development of cointegration in econometrics via reduced rank regression or canonical analysis by Johansen (1988, 1991) and Reinsel and Ahn (1992); see Reinsel and Velu (1998, Sec. 5.2, 5.3 and 5.6).

Noting that VARMA models "contain a dauntingly large number" of parameters, Tiao and Box (1981, p.805) suggested that "often models of rather low order can provide adequate approximation" and that "methods of seeking simplification, for example, principal component analysis or canonical analysis (see Box and Tiao (1977)), will often prove effective." Our approach has added the ideas of coefficient sparsity and rank sparsity to come up with estimable models in more general high-dimensional multivariate stochastic regression. In addition, Tiao and Box (1981, p.815) suggested that "in modeling as well as analysis of vector time series one often finds it useful to perform various eigenvalue and eigenvector analysis". We have followed their suggestion and used advances in numerical linear algebra to compute more general singular values and singular vectors of $\mathbf{X}\widehat{\mathbf{B}}^{OLS}$ in implementing the RCGL procedure of Bunea, She, and Wegkamp (2012) in Section 3. For the two-stage parsimonious procedure proposed herein, we use the simpler QR decomposition to implement GOGA in the first stage and an information criterion to estimate the rank in the second stage.

Acknowledgement

Lai's research is supported by the National Science Foundation grant DMS 1407828.

References

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

Bernanke, B. S., Bovian, J., and Eliasz, P. (2005). Measuring the effects of monetary policy: a factor-augmented vector autoregressive (FAVAR) approach. *Quart. J. Econom.* **120**, 387-422.

Box, G. E. P. and Tiao, G. C. (1977). A canonical analysis of multiple time series. *Biometrika* **64**, 355-365.

Bunea, F., She, Y. and Wegkamp, M. H. (2011). Optimal selection of reduced rank estimators of high-dimensional matrices. *Ann. Statist.* **39**, 1282-1309.

Bunea, F., She, Y. and Wegkamp, M. H. (2012). Joint variable and rank selection for parsimonious estimation of high dimensional matrices. *Ann. Statist.* **40**, 2359-2388.

Caines, P. E. (1988). *Linear Stochastic Systems*. Wiley, New York.

Candès, E. J. and Plan, Y. (2010). Tight oracle bounds for low-rank matrix recovery from a minimal number of random measurements. *IEEE Trans. Inform. Theory* **57**, 2342-2359.

Engle, R. and Mezrich, J. (1996). GARCH for groups. *Risk* **9**, 36-40.

Goodwin, G. C. and Sin, K. S. (1984). *Adaptive Filtering, Prediction and Control*. Prentice Hall, Englewood Cliffs, N.J.

Hannan, E. J. and Deistler, M. (1988). *The Statistical Theory of Linear Systems*. Wiley, New York.

Huang, D. and Guo, L. (1990). Estimation of nonstationary ARMAX models based on the Hannan-Rissanen method. *Ann. Statist.* **18**, 1729-1756.

Ing, C. K. and Lai, T. L. (2011). A stepwise regression method and consistent model selection for high-dimensional sparse linear models. *Statist. Sinica* **21**, 1473-1513.

Ing, C. K. and Lai, T. L. (2014). An efficient pathwise variable selection criterion in weakly sparse ARX models. Tech. Report, Dept. Statistics, Stanford Univ.

Johansen, S. (1988). Statistical analysis of cointegration vectors. *J. Econ. Dynamics & Control* **12**, 231-254.

Johansen, S. (1991). Estimation and hypothesis testing of cointegration vectors in Gaussian vector autoregression models. *Econometrica* **59**, 1551-1580.

Lai, T. L. and Tsang, K. W. (2014). A new approach to macroeconomic time series modeling and its applications. Tech. Report, Dept. Statistics, Stanford Univ.

Lai, T. L. and Xing, H. (2008). *Statistical Models and Methods for Financial Markets*. Springer, New York.

Negahban, S. and Wainwright, M. J. (2011). Estimation of (near) low-rank matrices with noise and high-dimensional scaling. *Ann. Statist.* **39**, 1069-1097.

Peña, D. and Tsay, R. S. (2011). A conversation with George C. Tiao. *Statist. Sci.* **25**, 408-428.

Reinsel, G. C. and Ahn, S. K. (1992). Vector autoregressive models with unit roots and reduced rank structure: Estimation, likelihood ratio test, and forecasting. *J. Time Series Anal.* **13**, 353-375.

Reinsel, G. C. and Velu, R. P. (1998). *Multivariate Reduced-Rank Regression: Theory and Applications*. Springer.

Rissanen, J. (1986). A predictive least squares principle. *IMA J. Math. Control Inform.* **3**, 211-222.

Rohde, A. and Tsybakov, A. B. (2011). Estimation of high-dimensional low-rank matrices. *Ann. Statist.* **39**, 887-930.

Stock, J. H. and Watson, M. W. (1999). Forecasting inflation. *J. Monetary Econ.* **44**, 293-335.

Stock, J. H. and Watson, M. W. (2002). Macroeconomic forecasting using diffusion indexes. *J. Bus. Econ. Statist.* **20**, 147-162.

Tiao, G. C. and Box, G. E. P. (1981). Modeling multiple time series with applications. *J. Amer. Statist. Assoc.* **76**, 802-816.

Tiao, G. C. and Tsay, R. S. (1989). Model specification in multivariate time series (with discussion). *J. Roy. Statist. Soc. B* **51**, 157-213.

Tsay, R. S. and Tiao, G. C. (1985). Use of canonical analysis in time series model identification. *Biometrika* **72**, 299-315.

Wei, C. Z. (1992). On predictive least squares principles. *Ann. Statist.* **20**, 1-42.

Yuan, M. and Lin, Y. (2006). Model selection and estimation in regression with grouped variables. *J. Roy. Statist. Soc. Ser. B* **68**, 49-67.

Department of Statistics, Stanford University, Stanford, CA 94305-4065, U.S.A.

E-mail: lait@stanford.edu

Institute for Computational & Mathematical Engineering, Stanford University, Stanford, CA 94305-4065, U.S.A.

E-mail: ktsang@stanford.edu

(Received July 2014; accepted February 2015)