J. R. Statist. Soc. B (2007)
69, *Part* 3, *pp.* 329–346

Dimension reduction and coefficient estimation in multivariate linear regression

Ming Yuan and Ali Ekici,

Georgia Institute of Technology, Atlanta, USA

Zhaosong Lu

Carnegie Mellon University, Pittsburgh, USA

and Renato Monteiro

Georgia Institute of Technology, Atlanta, USA

[Received March 2006. Revised November 2006]

Summary. We introduce a general formulation for dimension reduction and coefficient estimation in the multivariate linear model. We argue that many of the existing methods that are commonly used in practice can be formulated in this framework and have various restrictions. We continue to propose a new method that is more flexible and more generally applicable. The method proposed can be formulated as a novel penalized least squares estimate. The penalty that we employ is the coefficient matrix's Ky Fan norm. Such a penalty encourages the sparsity among singular values and at the same time gives shrinkage coefficient estimates and thus conducts dimension reduction and coefficient estimation simultaneously in the multivariate linear model. We also propose a generalized cross-validation type of criterion for the selection of the tuning parameter in the penalized least squares. Simulations and an application in financial econometrics demonstrate competitive performance of the new method. An extension to the non-parametric factor model is also discussed.

Keywords: Conic programming; Dimension reduction; Group variable selection; Ky Fan norm; Penalized likelihood

1. Introduction

Multivariate linear regressions are routinely used in chemometrics, econometrics, financial engineering, psychometrics and many other areas of applications to model the predictive relationships of multiple related responses on a set of predictors. In general multivariate linear regression, we have n observations on q responses $\mathbf{y} = (y_1, \ldots, y_q)'$ and p explanatory variables $\mathbf{x} = (x_1, \ldots, x_p)'$, and

$$Y = XB + E \tag{1}$$

where $Y = (\mathbf{y}_1, \ldots, \mathbf{y}_n)'$ is an $n \times q$ matrix, $X = (\mathbf{x}_1, \ldots, \mathbf{x}_n)'$ is an $n \times p$ matrix, B is a $p \times q$ coefficient matrix, $E = (\mathbf{e}_1, \ldots, \mathbf{e}_n)'$ is the regression noise and the $\mathbf{e}$s are independently sampled from $\mathcal{N}(0, \Sigma)$. Throughout the paper, we centre each input variable so that there is no intercept in equation (1) and also scale each input variable so that the observed standard deviation is 1.

Address for correspondence: Ming Yuan, School of Industrial and Systems Engineering, Georgia Institute of Technology, 755 Ferst Drive North West, Atlanta, GA 30332-0205, USA.
E-mail: myuan@isye.gatech.edu

 1369–7412/07/69329

The standard approach to estimating the coefficient matrix B is by means of ordinary least squares or maximum likelihood estimation methods (Anderson, 2003). The resulting estimates are equivalent to regressing each response on the explanatory variables separately. Clearly such estimates may perform suboptimally since they do not utilize the information that the responses are related. It is also well known that this type of estimate performs poorly in the presence of highly correlated explanatory variables or when p is relatively large.

A large number of methods have been proposed to overcome these problems. Most of these methods are based on dimension reduction. A particularly attractive family of methods is linear factor regression in which the response Y is regressed against a small number of linearly transformed predictors, which are often referred to as factors. These methods can be expressed in the following way:

$$Y = F\Omega + E, \tag{2}$$

where $F = X\Gamma$, Γ is a $p \times r$ matrix for some $r \leqslant \min(p, q)$ and Ω is an $r \times q$ matrix. The columns of F, $F_j (j = 1, \ldots, r)$, represent the so-called factors. Clearly equation (2) is an alternative representation of equation (1) in that $B = \Gamma\Omega$, and the dimension of the estimation problem reduces as r decreases. Estimation in the linear factor regression most often proceeds in two steps: the factors, or equivalently Γ, are first estimated and then Ω is estimated by least squares for equation (2). Many popular methods including canonical correlation (Hotelling, 1935, 1936), reduced rank (Anderson, 1951; Izenman, 1975; Reinsel and Velu, 1998), principal components (Massy, 1965), partial least squares (Wold, 1975) and joint continuum regression (Brooks and Stone, 1994) among others can all be formulated in the form of linear factor regression. They differ in the way in which the factors are determined.

It is obviously of great importance to be able to determine the number of factors, r, for equation (2). For smaller numbers of factors, a more accurate estimate is expected since there are fewer free parameters. But too few factors may not be sufficient to describe the predictive relationships. In all of the aforementioned methods, the number of factors, r, is chosen in a separate step from the estimation of equation (2) through either hypothesis testing or cross-validation. The coefficient matrix is typically estimated on the basis of the number of factors selected. Because of its discrete nature, this type of procedure can be very unstable in the sense of Breiman (1996): small changes in the data can result in very different estimates.

There are also other approaches to improve on the least squares method. Variable selection and ridge regression are among the most popular. Both types of method are most often studied in the special case of equation (1) when $q = 1$, which amounts to classical linear regression. In recent years, considerable effort has also been devoted to the more general situations with $q > 1$ (Frank and Friedman, 1993; Bedrick and Tsai, 1994; Fujikoshi and Satoh, 1997; Brown *et al.*, 1998, 1999, 2002; Turlach *et al.*, 2005; Lutz and Bühlmann, 2006). Variable selection is most powerful when there are many redundant predictors that are common to all responses, which can be unrealistic in many applications of the multivariate linear model (Reinsel and Velu, 1998). Ridge regression, however, oftentimes cannot offer easily interpretable models because it does not perform dimension reduction and all elements of the estimated coefficient matrix are typically non-zero.

In this paper, we propose a new technique for estimating the coefficient matrix that combines and retains the advantages of the existing methods. To achieve parsimonious models with enhanced interpretability, we introduce a formulation which is similar to but more general than the linear factor regression (2). Instead of estimating the coefficient matrix in multiple steps, as would be done in traditional linear factor regression methods, we simultaneously choose the number of factors, determine the factors and estimate the factor loadings Ω. Similarly to

ridge regression, the method proposed can be formulated as a penalized least squares estimate. The penalty that we employ is the coefficient matrix's Ky Fan norm defined as the sum of its singular values. Such a penalty encourages sparsity among singular values and at the same time gives shrinkage coefficient estimates and thus conducts dimension reduction and estimation simultaneously in the multivariate linear model.

The rest of the paper is organized as follows. The methodology proposed is introduced in the next section. An algorithm for solving the optimization problem in our formulation is relegated to Appendix A. Our algorithm takes advantage of recent advances in convex optimization by deriving an equivalent second-order cone program of the optimization problem in our formulation, which is readily solvable by using standard software. We consider the special case of orthogonal design in Section 3 to understand the new estimate better. A generalized cross-validation (GCV) type of statistic is introduced in Section 4 to choose the optimal tuning parameter for the method proposed. Simulations and a real data example are given in Sections 5 and 6 to illustrate the methodology. The method can also be extended to the non-parametric situation. In particular, we consider an extension to the vector additive model in Section 7. We conclude with some discussions in Section 8.

2. Factor estimation and selection

Denote by Y_j, B_j and E_j the jth columns of Y, B and E respectively. From equation (1), the jth response can be modelled by

$$Y_j = XB_j + E_j,$$

where $B_j \in R^p$, $j = 1, \ldots, q$. The basic idea of dimension reduction is that the regression coefficients $B_1, B_2, \ldots, B_q$ actually come from a linear space $\mathcal{B}$ of dimension lower than p. A general dimension reduction approach consists of two main ingredients: a set of basis elements $\{\eta_1, \ldots, \eta_p\}$ for R^p, and a subset $\mathcal{A}$ of $\{1, \ldots, p\}$ such that $\mathcal{B} \subseteq \text{span}\{\eta_i : i \in \mathcal{A}\}$ where span$\{\cdot\}$ stands for the linear space spanned by a set of vectors. Both variable selection and the linear factor model (2) can be formulated in this framework. In variable selection ηs are known, i.e. $\eta_i = e_i$, where e_i is the ith column of I_p; and we want to estimate $\mathcal{A}$. In the case of linear factor regression, the ith factor is given by $F_i = X\eta_i$ where ηs are estimated in a separate step, and $\mathcal{A}$ takes the form $\{1, 2, \ldots, r\}$ where r is to be estimated. Because of this connection, we shall refer to the estimation of ηs as factor estimation and the identification of $\mathcal{A}$ as factor selection. In this paper, we propose a procedure that imposes fewer restrictions than variable selection and linear factor regression by allowing both ηs and $\mathcal{A}$ to be estimated simultaneously.

To develop ideas, we start with factor selection and assume that $\{\eta_1, \ldots, \eta_p\}$ are known up to a permutation. With a slight abuse of notation, write $F = (F_1, \ldots, F_p)$ where $F_i = X\eta_i$; then

$$Y = F\Omega + E, \tag{3}$$

where Ω is a $p \times q$ matrix such that $(\eta_1, \ldots, \eta_p)\Omega = B$. Now factor selection for equation (1) can be cast as a variable selection problem for equation (3). As pointed out by Turlach *et al.* (2005), a family of estimates for this can be obtained by

$$\min[\text{tr}\{(Y - F\Omega)W(Y - F\Omega)'\}] \qquad \text{subject to } \sum_{i=1}^{p} \|\omega_i\|_\alpha \leqslant t, \tag{4}$$

where W is a weight matrix, ω_i is the ith row of Ω, $t \geqslant 0$ is a regularization parameter and $\|\cdot\|_\alpha$ is the l_α-norm for some $\alpha \geqslant 1$, i.e.

$$\|\omega_i\|_\alpha = (\Omega_{i1}^\alpha + \ldots + \Omega_{iq}^\alpha)^{1/\alpha}.$$

Common choices of the weight matrix include Σ^{-1} and I. To fix ideas, in the rest of the paper we shall assume that $W = I$.

It is clear that expression (4) reduces to the popular lasso (Tibshirani, 1996) when $q = 1$. Similarly to the lasso, if t is appropriately chosen, minimizing expression (4) yields a shrinkage estimate that is sparse in the sense that some of the ω_is will be set to 0. Consequently, the ith factor will be included in the final estimate if and only if ω_i is non-zero. Therefore, factor selection and coefficient estimation are done simultaneously. Two most obvious choices for α are $\alpha = 2$ and $\alpha = \infty$. The former has been studied by Bakin (1999) and Yuan and Lin (2006) whereas the latter has been discussed in Turlach *et al.* (2005). In this paper, we shall choose $\alpha = 2$. The advantage of this choice in the current setting will become clear in our later discussion. $\alpha = 2$ is appealing also because it allows the estimate from expression (4) to be invariant to any orthogonal transformation of the responses, which can be useful in many practical situations.

To use expression (4), we need to obtain ηs first. Similarly to variable selection, factor selection is most powerful if all responses can be predicted by a small subset of common factors. Ideally, we want $\{\eta_1, \ldots, \eta_p\}$ to contain a set of basis of $\mathcal{B}$ to allow the sparsest representation of B in the factor space. This is typically not so for the existing linear factor regression methods. For example, in principal components regression, the factors are chosen to be the principal components of the predictors, which may not necessarily contain the basis of $\mathcal{B}$. In our method, we choose ηs to be the eigenvectors of BB'. Clearly this set of basis contains the basis of $\mathcal{B}$. Interestingly, we can proceed even without actually estimating the factors if this choice is to be made in conjunction with $\alpha = 2$. To elaborate on this, write $U = (\eta_1, \ldots, \eta_p)$. The singular value decomposition of B can be expressed as $B = UDV'$ for some $q \times q$ orthonormal matrix V and a $p \times q$ matrix D such that $D_{ij} = 0$ for any $i \neq j$ and $D_{ii} = \sigma_i(B)$ where $\sigma_i(\cdot)$ represents the ith largest singular value of a matrix. Now $\Omega = DV'$ and $\omega_i = \sigma_i(B)V_i$ where V_i is the ith column of V, which implies that $\|\omega_i\|_2 = \sigma_i(B)$. Therefore, expression (4) with $\alpha = 2$ gives

$$\min[\mathrm{tr}\{(Y - XB)(Y - XB)'\}] \qquad \text{subject to} \sum_{i=1}^{\min(p,q)} \sigma_i(B) \leqslant t, \tag{5}$$

where $\sum_{i=1}^{\min(p,q)} \sigma_i(B)$ is known as the Ky Fan (p or q) norm of B. Clearly no knowledge of ηs is required in expression (5) and we shall use the minimizer of expression (5) as our final estimate of B. In Appendix A, we show that expression (5) is equivalent to a conic program and can be computed efficiently. The penalty that we employed in expression (5) encourages the sparsity among the singular values of B and at the same time gives shrinkage estimates for U and V; it thus conducts dimension reduction and estimation simultaneously in the multivariate linear model. Once the estimate of B is available, the basis ηs can be obtained as its left singular vectors U. Therefore, we can also compute the factors $F_i = X\eta_i$, as well as the factor loadings $\Omega = DV$.

The proposed estimate defined as the minimizer of expression (5) is closely connected with several other popular methods. In particular, expression (5), reduced rank regression and ridge regression can all be viewed as the minimizer of

$$\mathrm{tr}\{(Y - XB)(Y - XB)'\} \qquad \text{subject to} \left\{ \sum_i \sigma_i^\alpha(B) \right\}^{1/\alpha} \leqslant t \tag{6}$$

with difference choices of α.

Ridge regression defined as the minimizer of

$$\mathrm{tr}\{(Y - XB)(Y - XB)'\} + \lambda\, \mathrm{tr}(B'B)$$

corresponds to $\alpha = 2$ because $\mathrm{tr}(B'B) = \Sigma\, \sigma_i^2(B)$. It is well known that ridge regression provides a shrinkage estimate that often outperforms least squares. The estimate proposed, corresponding to $\alpha = 1$, enjoys a similar shrinkage property. To illustrate, consider the special case when there is only one response. In this case, $\sigma_1(B) = (B'B)^{1/2}$, and therefore expression (5) can now be expressed as

$$\min[\mathrm{tr}\{(Y - XB)(Y - XB)'\}] \qquad \text{subject to } (B'B)^{1/2} \leqslant t,$$

which is nothing other than the usual ridge regression.

Reduced rank regression is another special case of expression (6) with $\alpha = 0^+$. Both expression (5) and reduced rank regression set some of the singular values of B to 0 and lead to estimates with reduced ranks. Compared with reduced rank regression, the new method shrinks the singular values smoothly and is more stable. Note that the reduced rank regression estimate differs from the least squares estimate only in its singular values (Reinsel and Velu, 1998). Since the least squares estimate behaves poorly in overfitted or highly correlated settings, reduced rank regression may suffer in such situations as well. In contrast, the new method gives a shrinkage estimate that overcomes this problem.

3. Orthogonal design

To understand further the statistical properties of the method proposed, we consider the special case of orthogonal design. The following lemma gives an explicit expression for the minimizer of expression (5) in this situation.

Lemma 1. Let $\hat{U}^{\mathrm{LS}} \hat{D}^{\mathrm{LS}} \hat{V}^{\mathrm{LS}}$ be the singular value decomposition of the least squares estimate $\hat{B}^{\mathrm{LS}}$. Then, under the orthogonal design where $X'X = nI$, the minimizer of expression (5) is

$$\hat{B} = \hat{U}^{\mathrm{LS}} \hat{D} (\hat{V}^{\mathrm{LS}})',$$

where $\hat{D}_{ij} = 0$ if $i \neq j$, $\hat{D}_{ii} = \max(\hat{D}_{ii}^{\mathrm{LS}} - \lambda, 0)$ and $\lambda \geqslant 0$ is a constant such that $\Sigma_i\, \hat{D}_{ii} = \min(t, \Sigma \hat{D}_{ii}^{\mathrm{LS}})$.

Proof. Expression (5) can be equivalently written in a Lagrange form:

$$Q_n(B) = \frac{1}{2}\, \mathrm{tr}\{(Y - XB)(Y - XB)'\} + n\lambda \sum_{i=1}^{\min(p,q)} \sigma_i(B), \tag{7}$$

for some $\lambda > 0$. Simple algebra yields

$$\begin{aligned} \mathrm{tr}\{(Y - XB)(Y - XB)'\} &= \mathrm{tr}\{(Y - XB)'(Y - XB)\} \\ &= \mathrm{tr}\{(Y - X\hat{B}^{\mathrm{LS}})'(Y - X\hat{B}^{\mathrm{LS}})\} + \mathrm{tr}\{(\hat{B}^{\mathrm{LS}} - B)'X'X(\hat{B}^{\mathrm{LS}} - B)\} \\ &= \mathrm{tr}\{(Y - X\hat{B}^{\mathrm{LS}})'(Y - X\hat{B}^{\mathrm{LS}})\} + n\, \mathrm{tr}\{(\hat{B}^{\mathrm{LS}} - B)'(\hat{B}^{\mathrm{LS}} - B)\}. \end{aligned} \tag{8}$$

Together with the fact that $\mathrm{tr}(B'B) = \Sigma_i\, \sigma_i^2(B)$, equation (7) equals

$$\frac{1}{2} \sum_{i=1}^{q} \sigma_i^2(B) - \mathrm{tr}(B'\hat{B}^{\mathrm{LS}}) + \lambda \sum_{i=1}^{q} \sigma_i(B),$$

up to constants not depending on B. Now an application of von Neumann's trace inequality yields

$$\mathrm{tr}(B'\hat{B}^{\mathrm{LS}}) \leqslant \sum \sigma_i(B) \hat{D}_{ii}^{\mathrm{LS}}$$

Therefore,

$$Q_n(B) \geqslant \frac{1}{2} \sum_{i=1}^{q} \sigma_i^2(B) - \sum \sigma_i(B) \hat{D}_{ii}^{\mathrm{LS}} + \lambda \sum_{i=1}^{q} \sigma_i(B). \tag{9}$$

Note that $\sigma_i(B) \geqslant 0$. The right-hand side of inequality (9) is minimized at

$$\sigma_i(B) = \max(\hat{D}_{ii}^{\mathrm{LS}} - \lambda, 0), \qquad i = 1, \ldots, q.$$

The proof is now completed by noting that $\hat{B}$ achieves the lower bound for Q_n. □

This closed form minimizer of expression (5) allows a better understanding of our estimate. Specifically, the following lemma indicates that we can always find an appropriate tuning parameter such that the non-zero singular values of B are consistently estimated and the rest are set to 0 with probability 1.

Lemma 2. Suppose that $\max(p, q) = o(n)$. Under the orthogonal design, if $\lambda \to 0$ in such a fashion that $\max(p, q)/n = o(\lambda^2)$, then $|\sigma_i(\hat{B}) - \sigma_i(B)| \to_{\mathrm{p}} 0$ if $\sigma_i(B) > 0$ and $P\{\sigma_i(\hat{B}) = 0\} \to 1$ if $\sigma_i(B) = 0$.

Proof. Note that

$$\begin{aligned} \hat{B}^{\mathrm{LS}} &= (X'X)^{-1} X'Y \\ &= X'(XB + E)/n \\ &= B + X'E/n. \end{aligned}$$

Since $X'X = nI$ and the rows of E are independent observations from $\mathcal{N}(0, \Sigma)$, each entry of $X'E\Sigma^{-1/2}/\sqrt{n}$ follows $\mathcal{N}(0, 1)$ and is independent of each other. Applying the result from Johnstone (2001), we have

$$\sigma_1(X'E\Sigma^{-1/2}/n) \sim (\sqrt{p} + \sqrt{q})/\sqrt{n}.$$

Therefore,

$$\sigma_1\left(\frac{X'E}{n}\right) \leqslant \sigma_1\left(\frac{X'E\Sigma^{-1/2}}{n}\right) \sigma_1(\Sigma^{1/2}) \sim \sigma_1^{1/2}(\Sigma) \frac{\sqrt{p} + \sqrt{q}}{\sqrt{n}}.$$

Now an application of theorem 3.3.16 of Horn and Johnson (1991) yields

$$\begin{aligned} |\sigma_i(B) - \sigma_i(\hat{B}^{\mathrm{LS}})| &\leqslant \sigma_1\left(\frac{X'E}{n}\right) \\ &= O_p\left(\frac{\sqrt{p} + \sqrt{q}}{\sqrt{n}}\right). \end{aligned} \tag{10}$$

Therefore, if $\lambda \to 0$ at a slower rate than the right-hand side of equation (10), the proposed estimate can provide consistent estimates of the non-zero singular values of B and at the same time shrink the rest of the singular values to 0. □

Lemma 1 also indicates that the singular values of the method proposed are shrunk in a similar fashion to the lasso under orthogonal designs. The lasso has proved highly successful in various studies, particularly when the predictors are correlated and p is large relatively to the sample size. In Section 5, we show that the estimate proposed is very successful in similar situations as well.

4. Tuning

Like any other regularization method, it is important to be able to choose a good tuning parameter t in expression (5). One common method that is used in practice is cross-validation, which of course can be computationally demanding in large scale problems. In this section, we develop a GCV (Golub *et al.*, 1979) type of statistic for determining t.

We first characterize the equivalence between expression (5) and its Lagrange form (7), since it is easier to work with equation (7) in deriving our GCV-type statistic. Denote $\hat{B}$ the minimizer of expression (5) and $\hat{U}\hat{D}\hat{V}'$ its singular value decomposition. Note that expression (5) is equivalent to equation (7) and we can always find a λ such that $\hat{B}$ is also the minimizer of equation (7). The following lemma explicitly describes the relationship between t and λ.

Lemma 3. Write $\hat{d}_i = \hat{D}_{ii}$ for $i = 1, \ldots, \min(p, q)$. For any $t \leqslant \Sigma_i \hat{d}_i$, the minimizer of equation (7) coincides with the minimizer of expression (5), $\hat{B}$, if

$$n\lambda = \frac{1}{\text{card}(\hat{d}_i > 0)} \sum_{\hat{d}_i > 0} (\tilde{X}_i' \tilde{Y}_i - \tilde{X}_i' \tilde{X}_i \hat{d}_i) \tag{11}$$

where $\text{card}(\cdot)$ stands for the cardinality of a set, $\tilde{Y}_i$ is the ith column of $\tilde{Y} = Y\hat{U}$ and $\tilde{X}_i$ is the ith column of $\tilde{X} = X\hat{V}$.

Proof. Note that

$$\begin{aligned} \sum_{i=1}^{\min(p,q)} \sigma_i(\hat{B}) &= \sum_{i=1}^{\min(p,q)} \hat{D}_{ii} \\ &= \sum_{i=1}^{p} \sigma_i(\hat{B}K\hat{B}') \\ &= \text{tr}(\hat{B}K\hat{B}'), \end{aligned}$$

where

$$K = \sum_{\hat{D}_{ii} > 0} \frac{1}{\hat{D}_{ii}} \hat{V}_i \hat{V}_i',$$

and $\hat{V}_i$ is the ith column of V. Therefore, $\hat{B}$ is also the minimizer of

$$\tfrac{1}{2} \text{tr}\{(Y - XB)(Y - XB)'\} + n\lambda \, \text{tr}(BKB'). \tag{12}$$

From expression (12), $\hat{d}$ is the minimizer of

$$\frac{1}{2} \sum_{i=1}^{\min(p,q)} (\tilde{Y}_i - \tilde{X}_i d_i)^2 + n\lambda \sum_{i=1}^{\min(p,q)} d_i, \tag{13}$$

subject to the constraint that $d_i \geqslant 0$. The first-order optimality condition for expression (13) yields

$$n\lambda = \tilde{X}_i' \tilde{Y}_i - \tilde{X}_i' \tilde{X}_i \hat{d}_i,$$

for any $\hat{d}_i > 0$. The proof is now completed by taking an average of the above expression over all is such that $\hat{d}_i > 0$. □

Since $\hat{B}$ is the minimizer of expression (12), it can be expressed as

$$\hat{B} = (X'X + 2n\lambda K)^{-1} X'Y.$$

Neglecting the fact that W also depends on $\hat{B}$, we can define the hat matrix for expression (12) as $X(X'X+2n\lambda K)^{-1}X'$ and the degrees of freedom as

$$\mathrm{df}(t) = q\,\mathrm{tr}\{X(X'X+2n\lambda K)^{-1}X'\}.$$

Now the GCV score is given by

$$\mathrm{GCV}(t) = \frac{\mathrm{tr}\{(Y-X\hat{B})(Y-X\hat{B})'\}}{qp-\mathrm{df}(t)}, \tag{14}$$

and we choose a tuning parameter by minimizing $\mathrm{GCV}(t)$.

Summing up, an implementation of our estimate which chooses the tuning parameter automatically is as follows.

Step 1: for each candidate t-value

(a) compute the minimizer of expression (5) (denote the solution $\hat{B}(t)$),
(b) evaluate λ by using equation (11) and
(c) compute the GCV score (14).

Step 2: denote t^* the minimizer of the GCV score that is obtained in step 1. Return $\hat{B}(t^*)$ as the estimate of B.

5. Simulation

In this section, we compare the finite sample performance of the proposed estimate with several other popular approaches for multivariate linear regression. The methods that we compared include the following:

(a) FES, the method proposed for factor estimation and selection with the tuning parameter selected by GCV;
(b) OLS, the ordinary least square estimate $(X'X)^{-1}X'Y$;
(c) CW, the curd and whey with GCV procedure that was developed by Breiman and Friedman (1997);
(d) RRR, reduced rank regression with the rank selected by tenfold cross-validation;
(e) PLS, two-block partial least squares (Wold, 1975) with the number of components selected by tenfold cross-validation;
(f) PCR, principal components regression (Massy, 1965) with the number of components selected by tenfold cross-validation;
(g) RR, ridge regression with the tuning parameter selected by tenfold cross-validation;
(h) CAIC, forward selection using the corrected Akaike information criterion that was proposed by Bedrick and Tsai (1994). The corrected Akaike information criterion for a specific submodel of model (1) is defined as

$$n\,\ln|\hat{\Sigma}| + \frac{n(n+k)q}{n-k-q-1} + nq\,\ln(2\pi),$$

where k is the number of predictors included in the submodel and $\hat{\Sigma}$ is the maximum likelihood estimate of Σ under the submodel.

We compare these methods in terms of the model error. The model error of an estimate $\hat{B}$ is given by

$$\mathrm{ME}(\hat{B}) = (\hat{B}-B)'V(\hat{B}-B),$$

where $V = E(X'X)$ is the population covariance matrix of X.

Table 1. Comparisons on the simulated data sets

Model	*Results for the following methods:*							
	FES	*OLS*	*CW*	*RRR*	*PLS*	*PCR*	*RR*	*CAIC*
I	3.02	6.31	4.47	6.14	4.72	5.46	3.72	11.6
	(0.06)	(0.15)	(0.12)	(0.16)	(0.10)	(0.12)	(0.07)	(0.20)
II	2.97	6.31	5.20	6.97	3.95	3.70	2.46	5.40
	(0.04)	(0.15)	(0.11)	(0.15)	(0.06)	(0.05)	(0.04)	(0.03)
III	2.20	6.31	3.49	2.42	4.15	6.01	4.36	15.6
	(0.06)	(0.15)	(0.11)	(0.13)	(0.14)	(0.18)	(0.10)	(0.51)
IV	4.95	14.23	8.91	12.45	6.45	6.57	4.47	9.65
	(0.03)	(0.13)	(0.08)	(0.08)	(0.04)	(0.04)	(0.02)	(0.04)

We consider the following four models.

(a) For model I we consider an example with $p = q = 8$. A random 8×8 matrix with singular values $(3, 2, 1.5, 0, 0, 0, 0, 0)$ was first generated as the true coefficient matrix. This is done as follows. We first simulated an 8×8 random matrix whose elements are independently sampled from $\mathcal{N}(0, 1)$, and then replace its singular values with $(3, 2, 1.5, 0, 0, 0, 0, 0)$. Predictor $\mathbf{x}$ is generated from a multivariate normal distribution with correlation between x_i and x_j being $0.5^{|i-j|}$. Finally, $\mathbf{y}$ is generated from $\mathcal{N}(\mathbf{x}B, I)$. The sample size for this example is $n = 20$.

(b) Model II is the same as model I except that the singular values are $\sigma_1 = \ldots = \sigma_8 = 0.85$.

(c) Model III is the same set-up as before, but with singular values $(5, 0, 0, 0, 0, 0, 0, 0)$.

(d) Model IV is a larger problem with $p = 20$ predictors and $q = 20$ responses. A random-coefficient matrix is generated in the same fashion as before with the first 10 singular values being 1 and last 10 singular values being 0. $\mathbf{x}$ and $\mathbf{y}$ are generated as in the previous examples. The sample size is set to be $n = 50$.

For each of these models, 200 data sets were simulated. Table 1 gives the means and standard errors (in parentheses) over the 200 simulated data sets. To gain further insight into the comparison, we also provide a pairwise prediction accuracy comparison between method FES and the other methods for models I–IV (except for method CAIC) in Fig. 1.

We use these examples to examine the relative merits of the methods in various scenarios. Model I has a moderate number of moderate-size singular values; the first two rows of Table 1 and Fig. 1(a) indicate that method FES enjoys the best prediction accuracy followed by ridge regression and the curd and whey method. Model II represents a different scenario with a full rank coefficient matrix. Table 1 indicates that ridge regression performs the best, with method FES being the only other method to improve on ordinary least squares by more than 50%. Model III has a small number of large singular values. In this case, method FES does the best. Also as expected, reduced rank regression performs relatively well but with considerably more variation. The last example is a bigger model with higher dimension. Ridge regression performs the best, with method FES being a close second. Note that method CAIC performs poorly for all four models because there are no redundant variables in general. In all the examples, method FES demonstrates competitive performance when compared with the other six methods. Ridge regression does a little better than the proposed method for models II and IV, but worse for the

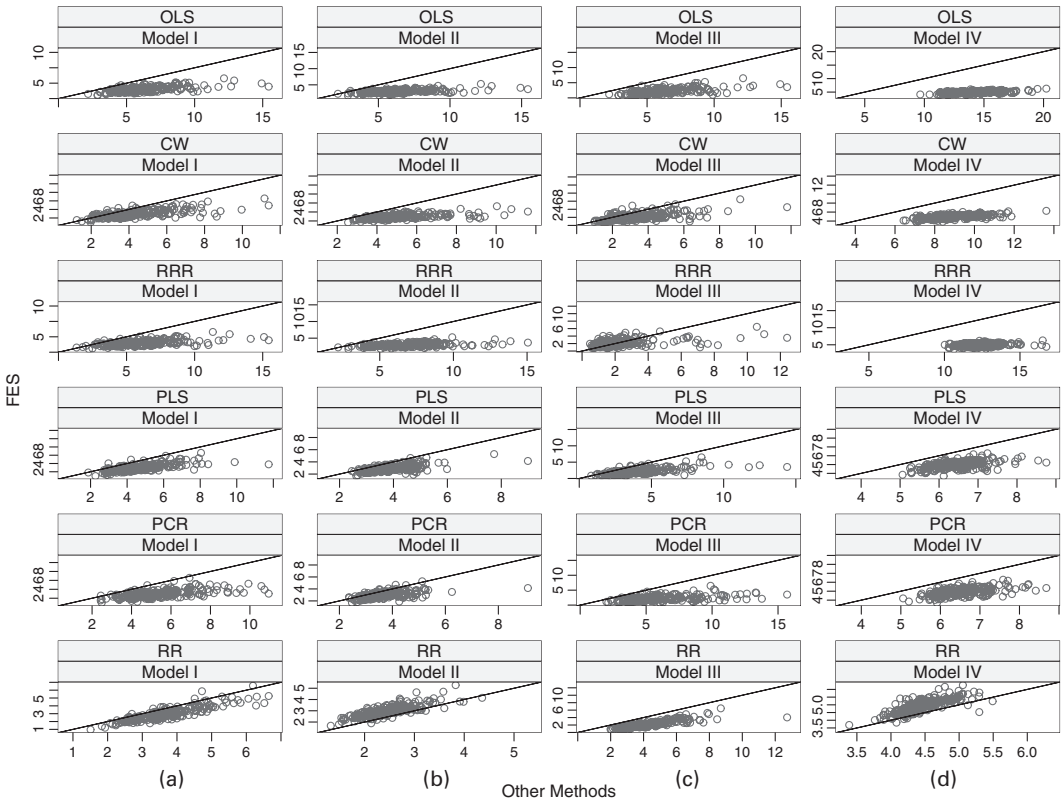

Fig. 1. Pairwise model error comparison between model FES and the other methods

other two models. It is worth pointing out that, when compared with ridge regression, method FES also has the further advantage of producing interpretable models.

6. Application

To demonstrate the utility of the method proposed, we now consider a real example in financial econometrics. The multivariate linear model has a wide range of applications in finance, because portfolios, one of the main objects in financial studies, are typically generated from vector-valued processes. A particularly important task in financial econometrics is to predict the future returns of assets on the basis of their historical performance. Vector autoregressive models are often used for this (Reinsel, 1997). Let $\mathbf{y}_t$ be the vector of returns at time t. The vector autoregressive model with order 1 is given by

$$\mathbf{y}_t = \mathbf{y}_{t-1} B + E. \tag{15}$$

Clearly model (15) is a special case of the multivariate linear model. Accurate estimation of B in model (15) leads to good forecasts which, in turn, can serve as instruments for efficient portfolio allocation and revealing opportunities for arbitrage. Also important is the identification of the factors in model (15), which can to help construct bench-mark portfolios or to diversify investments.

To illustrate our method, we applied model (15) to the stock prices in 2004 of the 10 largest American companies ranked by *Fortune* magazine on the basis of their 2003 revenue. We

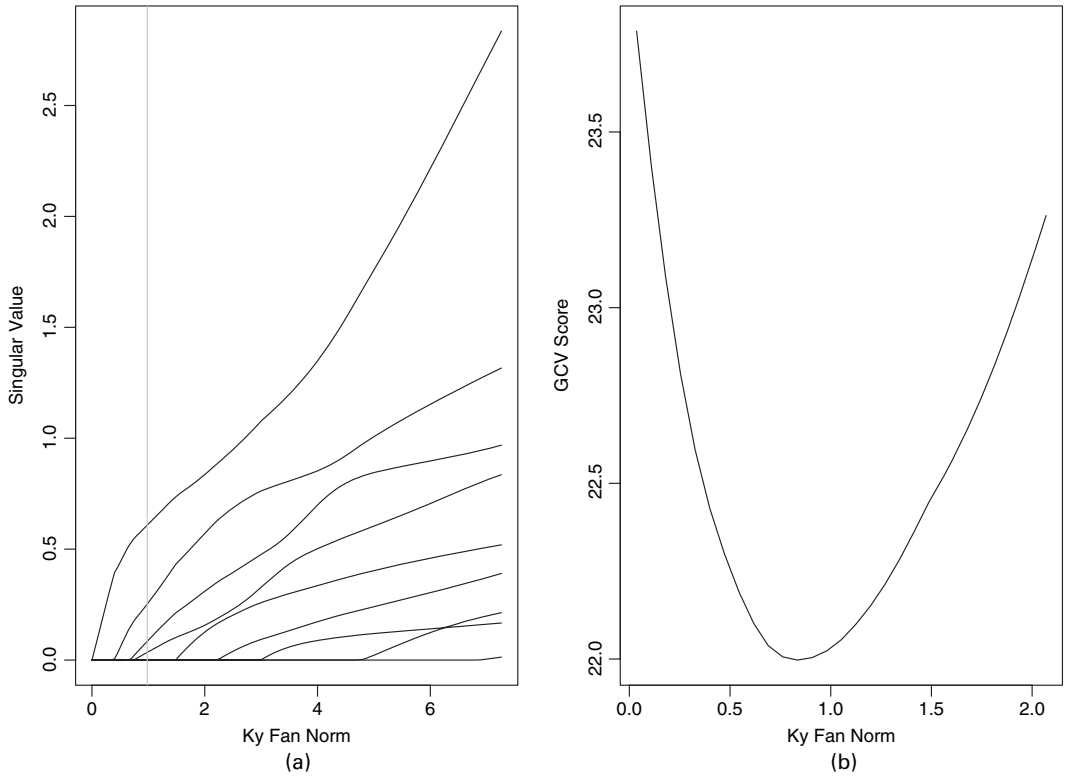

Fig. 2. Solution paths for the stocks example

excluded Chevron in the analysis because its stock price dropped nearly a half in the 38th week of the year, which indicates the non-stationarity of its return process. We fit model (15) to the weekly log-returns of the stocks for the first half of the year and use the data from the second half of the year to evaluate the predictive performance.

We first apply the proposed factor estimation and selection method on the training data. Fig. 2(a) gives the trajectory of the singular values of our proposed method as the Ky Fan norm of B increases and Fig. 2(b) depicts the GCV curve. The vertical line in Fig. 2(a) corresponds to the Ky Fan norm that was selected by GCV.

In this example, GCV retains four non-zero singular values for B. Recall that the factors are estimated by the left singular vectors of the regression coefficient matrix estimate. The corresponding loadings of the four selected factors are given in Table 2.

It is of great interest to understand the meaning of these four factors. From expression (15), the factors summarize the asset return history in predicting the future returns. Classical investment theory indicates that the market index is a good summary of the asset prices and should lie in the factor space. To approximate the market index, we picked the Standard and Poors index S&P500 and the Nasdaq Stock Market index NASDAQ. To check whether their returns approximately fall into the factor space that is estimated from the stock data, we constructed their linear projections in the estimated four-dimensional factor space. The log-returns of S&P500 and NASDAQ in the year of 2004 together with their approximations are given in Fig. 3. Both approximations track the actual log-return processes fairly well. This exercise confirms that the

Table 2. Factor loadings for the stocks example

Company	*Loadings for the following factors:*			
	1	*2*	*3*	*4*
Walmart	−0.47	−0.42	−0.30	0.19
Exxon	0.20	−0.68	0.07	−0.40
GM	0.05	0.19	−0.61	−0.31
Ford	0.18	0.22	−0.42	−0.13
GE	−0.35	0.13	−0.03	−0.44
ConocoPhillips	0.42	0.04	0.05	−0.52
Citigroup	−0.45	0.13	−0.26	−0.17
IBM	−0.24	0.43	0.49	−0.21
AIG	−0.38	−0.22	0.22	−0.39

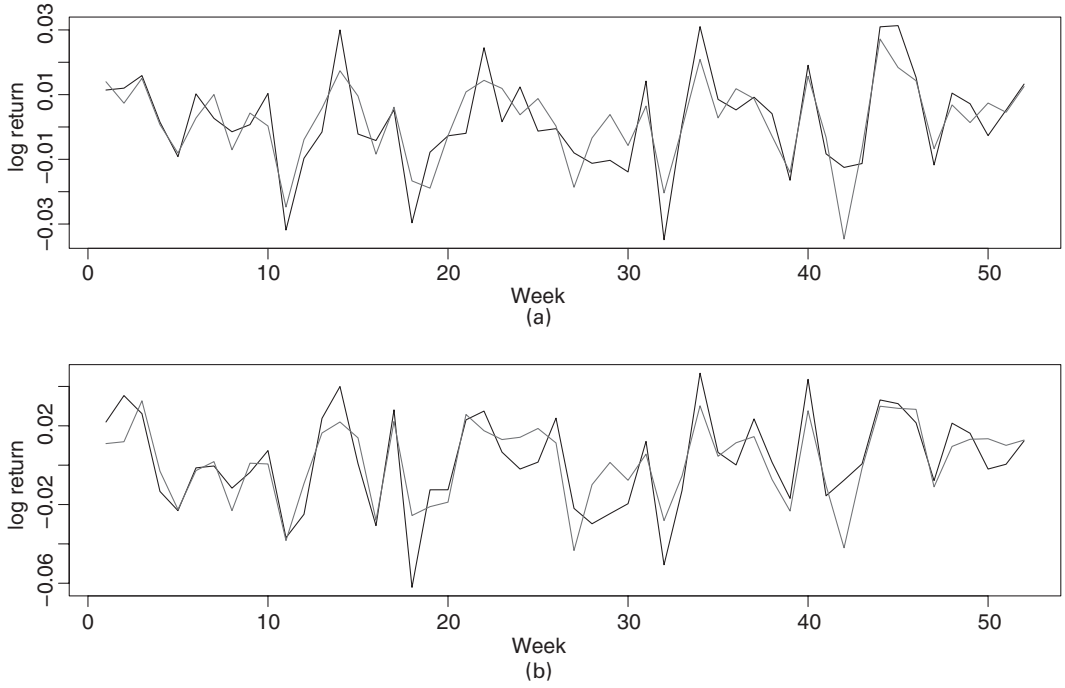

Fig. 3. (a) S&P500 and (b) NASDAQ indices (———) together with their approximations in the factor space (· · · · · · ·)

factors that are revealed by our method are indeed meaningful and should provide insight into further studies of the dynamics of the financial market.

To compare further the method proposed with the other methods from the last section, we compare their prediction errors on the data from the second half of the year. For each individual stock, we reported the averaged predictive squared error of the forecast. We also reported the prediction error averaged over all nine stocks. The prediction performances are summarized in Table 3. The method proposed clearly provides better prediction than the other methods in this example.

Table 3. Out-of-sample mean-squared error

Company	*Mean-squared errors (× 0.001) for the following methods:*							
	FES	*OLS*	*CW*	*RRR*	*PLS*	*PCR*	*RR*	*CAIC*
Walmart	0.40	0.98	0.69	0.50	0.44	0.44	0.43	0.42
Exxon	0.29	0.39	0.37	0.32	0.33	0.32	0.32	0.30
GM	0.62	1.68	1.29	1.53	0.68	0.69	0.62	0.67
Ford	0.69	2.15	1.31	2.22	0.65	0.77	0.68	0.74
GE	0.41	0.58	0.45	0.49	0.44	0.45	0.42	0.44
ConocoPhillips	0.79	0.98	1.63	0.79	0.83	0.79	0.79	0.79
Citigroup	0.59	0.65	0.63	0.66	0.60	0.65	0.58	0.61
IBM	0.51	0.62	0.58	0.54	0.62	0.49	0.49	0.48
AIG	1.74	1.93	1.86	1.86	1.81	1.92	1.81	1.80
Average	0.67	1.11	0.98	0.99	0.71	0.72	0.68	0.70

7. Non-parametric factor model

The method proposed can also be extended to vector non-parametric regression models where the jth response is related to predictor $\mathbf{x}$ through the regression equation

$$y_j = g_j(\mathbf{x}) + e_j, \qquad j = 1, \ldots, q, \tag{16}$$

where the g_js are unknown smooth functions to be estimated. We begin with the case when the predictor is univariate. A non-parametric extension of the linear factor model is to assume that g_j can be expressed as a linear combination of a small number of non-parametric factors f_k, $k = 1, \ldots, r$:

$$g_j(x) = \omega_{1j}\, f_1(x) + \omega_{2j}\, f_2(x) + \ldots + \omega_{rj}\, f_r(x),$$

where $\Omega = (\omega_{kj})_{r \times q}$ is unknown. To estimate the g_js, we model the non-parametric factors by using regression splines:

$$f_k(x) = \beta_{1k} x + \ldots + \beta_{sk} x^s + \sum_{m=1}^{M} \beta_{m+s,\,k} (x - \kappa_m)_+^s, \qquad k = 1, \ldots, r, \tag{17}$$

where $s \geqslant 1$ is an integer, $(u)_+^s = u^s\, I(u \geqslant 0)$ and $\kappa_1 < \ldots < \kappa_M$ are fixed knots. Write $\mathbf{z} = (x, \ldots, x^s, (x - \kappa_1)_+^s, \ldots, (x - \kappa_M)_+^s)$, $A = (\beta_{ik})_{(M+s) \times r}$ and $B = A\Omega$. Then equation (16) can be rewritten as

$$\mathbf{y} = \mathbf{z}B + \mathbf{e},$$

which is in the form of the multivariate linear regression (1). In traditional regression spline approaches, the choice of the knots $\kappa_1, \ldots, \kappa_M$ is crucial since too many knots result in overfitting whereas too few knots may not be able to capture the non-linear structure. Sophisticated knot selection procedures are often employed to ensure good performance of the regression spline estimate. A method that often enjoys better performance and is simpler to implement is the so-called penalized regression spline method (Eilers and Marx, 1996; Ruppert and Carroll, 1997) where a large number of knots are included and overfitting is avoided through shrinkage. Adopting this idea, an estimate of the non-parametric factor model can be defined as the solution of

$$\min_B[\mathrm{tr}\{(Y-ZB)(Y-ZB)'\}] \qquad \text{subject to } \sum_{k=1}^{r}\left(\sum_{i=1}^{M+s}\beta_{ik}^2\right)^{1/2} \leqslant t, \tag{18}$$

where $Y=(\mathbf{y}_1,\ldots,\mathbf{y}_n)'$ and $Z=(\mathbf{z}_1,\ldots,\mathbf{z}_n)'$. For identifiability, we further assume that $A'A=\Omega\Omega'=I_r$. Then expression (18) is equivalent to

$$\min_B[\mathrm{tr}\{(Y-ZB)(Y-ZB)'\}] \qquad \text{subject to } \sum_i \sigma_i(B) \leqslant t, \tag{19}$$

which is of the same form as expression (5) and can also be solved by using the algorithm that is provided in Appendix A.

In most practical situations, the predictors are multivariate. To alleviate the 'curse of dimensionality', additive models (Hastie and Tibshirani, 1990) are commonly used where multivariate functions g_j are written as

$$g_j(\mathbf{x})=g_{j1}(x_1)+\ldots+g_{jp}(x_p), \qquad j=1,\ldots,q. \tag{20}$$

Here $g_{j1},\ldots,g_{jp}$ are univariate functions. Consider a non-parametric factor model for each component on the right-hand side of equation (20):

$$g_{ji}(x_i)=\omega_{1j}^{(i)}\, f_{i1}(x_i)+\omega_{2j}^{(i)}\, f_{i2}(x_i)+\ldots+\omega_{r_ij}^{(i)}\, f_{ir_i}(x_i), \tag{21}$$

where

$$f_{ik}(x_i)=\beta_{1k}^{(i)}x_i+\ldots+\beta_{sk}^{(i)}x_i^s+\sum_{m=1}^{M}\beta_{m+s,k}^{(i)}(x_i-\kappa_m^{(i)})_+^s, \qquad k=1,\ldots,r_i.$$

Denote $\mathbf{z}_i=(x_i,\ldots,x_i^s,(x_i-\kappa_1^{(i)})_+^s,\ldots,(x_i-\kappa_M^{(i)})_+^s)$, $A_i=(\beta_{jk}^{(i)})_{(M+s)\times r}$, $\Omega_i=(\omega_{jk}^{(i)})$ and $B_i=A_i\Omega_i$. Then a non-parametric factor model for multivariate predictors can be given as

$$\mathbf{y}=\mathbf{z}_1B_1+\ldots+\mathbf{z}_pB_p+\mathbf{e}.$$

Similarly to expression (19), we define our estimate of the B_is as the solution of

$$\min_{B_1,\ldots,B_p}[\mathrm{tr}\{(Y-ZB)(Y-ZB)'\}] \qquad \text{subject to } \sum_j \sigma_j(B_i)\leqslant t_i,\ j=i,\ldots,p, \tag{22}$$

where $Z=(Z_1,\ldots,Z_p)$ and $B=(B_1',\ldots,B_p')'$. Using the algorithm that is presented in Appendix A, expression (22) can be solved in an iterative fashion.

Step 1: initialize $B_i=0$, $i=1,\ldots,p$.
Step 2: for $i=1,\ldots,p$,

(a) compute $Y^*=Y-Z_1B_1-\ldots-Z_{i-1}B_{i-1}-Z_{i+1}B_{i+1}-\ldots-Z_pB_p$ and
(b) update B_i by minimizing $\mathrm{tr}\{(Y^*-Z_iB_i)(Y^*-Z_iB_i)'\}$ subject to $\Sigma_j\,\sigma_j(B_i)\leqslant t_i$.

Step 3: repeat step 2 until B does not change.

To illustrate, we reanalyse the biochemical data from Smith *et al.* (1962). The data contain chemical measurements on several characteristics of 33 individual samples of men's urine specimens. There are five response variables: pigment creatinine, concentrations of phosphate, phosphorus, creatinine and choline. The goal of the analysis is to relate these responses to three predictors: the weight of the subject, volume and specific gravity. Reinsel and Velu (1998) postulated a multivariate linear model to analyse the data. A non-parametric extension such as equation (21) could be more powerful if non-linear effects of the predictors are suspected. For this, we model the effect of each of the predictors by equation (17) with $s=2$ and $M=5$. The

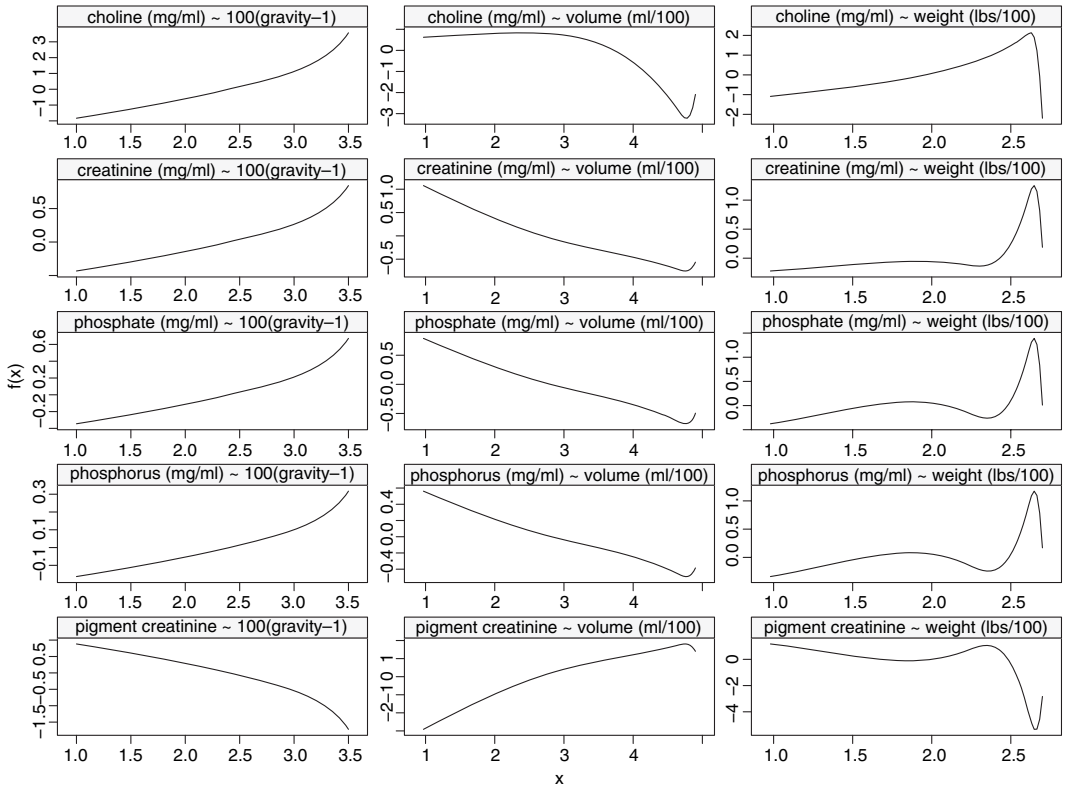

Fig. 4. Fitted components for the biochemistry data

knots are chosen to be equally spaced quantiles of the corresponding covariate. A practical issue in using this method is the choice of tuning parameters $t_1, \ldots, t_p$. We adopted a strategy that is commonly used in smoothing spline models when there are multiple tuning parameters: t_i is tuned at step 2(b) in each iteration by using the GCV criterion that was developed before. Fig. 4 shows the estimated effect of each predictor on each response. Clear departure from a linear assumption can be observed.

8. Discussion

In this paper, we introduced a general formulation for dimension reduction and coefficient estimation in the multivariate linear model. On the basis of this formulation, we proposed a new method for shrinkage and dimension reduction. The method has connection with many existing methods but has been demonstrated to enjoy considerably better performance. We also extended the method to a non-parametric model for predicting multiple responses. The implementation of our method takes advantage of recent advances in convex optimization.

Linear factor regression reduces the dimensionality of the estimating problem and often leads to models with enhanced interpretability. However, it can be unstable because of the discrete nature of selecting the number of factors. Also, the factors are often constructed in an *ad hoc* fashion and may not allow sufficient dimension reduction. In contrast, ridge regression often enjoys superior prediction accuracy because it leads to a shrinkage estimate, but it does not provide easily interpretable models. Our method combines and retains the advantages of

both approaches. Formulated as a penalized least squares estimate, the method proposed gives a shrinkage estimate with reduced ranks. We demonstrated by numerical examples that the method proposed enjoys competitive performance when compared with other popular methods.

The penalty that we employed is the coefficient matrix's Ky Fan norm, which shares some similar characteristics with the absolute value constraints that are used by the lasso in the special case of orthogonal designs as illustrated in Section 3. Such similarity and the encouraging results that were reported here suggest that this penalty may prove useful in other statistical problems where a matrix of high dimension is to be estimated.

Acknowledgement

Yuan's research was supported in part by National Science Foundation grant DMS-0624841.

Appendix A: Algorithm for solving expression (5)

To solve expression (5), we take advantage of the recent advance in convex optimization. We show that expression (5) is equivalent to a second-order cone program and can be solved by using standard solvers such as SDPT3 (Tütüncü *et al.*, 2003).

Let us first introduce some notation. Denote by $\mathcal{L}^m$ the m-dimensional second-order cone:

$$\mathcal{L}^m = \{\mathbf{x} = (x_1, \ldots, x_m) \in \mathcal{R}^m : x_1 \geqslant \sqrt{(x_2^2 + \ldots + x_m^2)}\}.$$

Write $\mathcal{R}_+^m = \{\mathbf{x} = (x_1, \ldots, x_m)' : x_i \geqslant 0, i = 1, \ldots, m\}$ and $X \succeq 0$ to indicate that the symmetric matrix X is positive semidefinite. Also for an $n \times n$ symmetric matrix X, define the vectorization operator svec as

$$\text{svec}(X) = (X_{11}, X_{21}\sqrt{2}, X_{22}, \ldots, X_{n1}\sqrt{2}, \ldots, X_{n,n-1}\sqrt{2}, X_{nn})'.$$

SDPT3 can solve problems of the form

$$\min_{X_j^s, \mathbf{x}_i^q, \mathbf{x}^l} \left\{ \sum_{j=1}^{n_s} \text{tr}(C_j^s X_j^s) + \sum_{i=1}^{n_q} (\mathbf{c}_i^q)' \mathbf{x}_i^q + (\mathbf{c}^l)' \mathbf{x}^l \right\}$$

such that

$$\sum_{j=1}^{n_s} (A_j^s)' \text{svec}(X_j^s) + \sum_{i=1}^{n_q} (A_i^q)' \mathbf{x}_i^q + (A^l)' \mathbf{x}^l = b, \qquad X_j^s \succeq 0 \quad \forall j, \mathbf{x}_i^q \in \mathcal{L}^{q_i} \quad \forall i, \mathbf{x}^l \in \mathcal{R}_+^{n_l}, \tag{23}$$

where C_j^s is a symmetric matrix of the same dimension as X_j^s, $\mathbf{c}_i^q$ is a q_i-dimensional vector, $\mathbf{c}^l$ is an n_l-dimensional vector and the dimensions of matrices A and vector b are clear from the context.

Next we show that expression (5) can be equivalently written in the form of problem (23). Similarly to equation (8), the objective function of expression (5) can be rewritten as

$$\text{tr}\{(B - \hat{B}^{\text{LS}})' X' X (B - \hat{B}^{\text{LS}})\} = \text{tr}(C'C)$$

up to a constant free of B where $C = \Lambda^{1/2} Q (B - \hat{B}^{\text{LS}})$ and $Q' \Lambda Q$ is the eigenvalue decomposition of $X'X$. By the definition of the second-order cone, expression (5) can be equivalently written as

$$\min_{M,C,B} (M)$$

such that

$$(M, C_{11}, \ldots, C_{1q}, C_{21}, \ldots, C_{pq})' \in \mathcal{L}^{pq+1}, \qquad \sum_{i=1}^{q} \sigma_i(B) \leqslant t, \quad C = \Lambda^{1/2} Q (B - \hat{B}^{\text{LS}}).$$

Using the Schur complement lemma (Ben-Tal and Nemirovski, 2001), the constraint $\Sigma\, \sigma_i(B) = \Sigma\, \sigma_i(QB) \leqslant t$ is equivalent to

$$\sum_{i=1}^{\min(p,q)} \mu_i(A) \leqslant t$$

where $\mu_i(A)$ is the ith eigenvalue of A and

$$A = \begin{pmatrix} 0 & (QB)' \\ (QB) & 0 \end{pmatrix}.$$

Together with formula (4.2.2) of Ben-Tal and Nemirovski (2001), page 147, this constraint is also equivalent to

$$qs + \mathrm{tr}(Z) \leqslant t,$$

$$Z - \begin{pmatrix} 0 & (\Lambda^{-1/2}C + Q\hat{B}^{\mathrm{LS}})' \\ (\Lambda^{-1/2}C + Q\hat{B}^{\mathrm{LS}}) & 0 \end{pmatrix} + sI \succeq 0,$$

$$Z \succeq 0.$$

Now, expression (5) is equivalent to

$$\min_{M, C, \mathbf{s}, Z_1, Z_2} (M)$$

subject to

$$q(s_1 - s_2) + \mathrm{tr}(Z_1) + s_3 = t,$$

$$Z_2 - Z_1 + \begin{pmatrix} 0 & (\Lambda^{-1/2}C)' \\ (\Lambda^{-1/2}C) & 0 \end{pmatrix} - (s_1 - s_2)I = \begin{pmatrix} 0 & -(Q\hat{B}^{\mathrm{LS}})' \\ -(Q\hat{B}^{\mathrm{LS}}) & 0 \end{pmatrix},$$

$$Z_1, Z_2 \succeq 0,$$

$$(M, C_{11}, \ldots, C_{1q}, C_{21}, \ldots, C_{pq})' \in \mathcal{L}^{pq+1},$$

$$\mathbf{s} \in \mathcal{R}^3_+,$$

which is readily computable by using SDPT3.

When $W \neq I$, the objective function becomes

$$\mathrm{tr}\{(Y - XB)W(Y - XB)'\}.$$

A slight modification needs to be made to the above derivation. In this case, the objective function can be rewritten as

$$\mathrm{tr}\{(B - \hat{B}^{\mathrm{LS}})'X'X(B - \hat{B}^{\mathrm{LS}})W\}$$

up to a constant that is free of B. Let $P\Delta P'$ be the eigenvalue decomposition of W and define

$$C = \Lambda^{1/2}Q(B - \hat{B}^{\mathrm{LS}})P\Delta^{1/2}.$$

Similar arguments lead to the following equivalent conic program:

$$\min_{M, C, \mathbf{s}, Z_1, Z_2} (M)$$

subject to

$$q(s_1 - s_2) + \mathrm{tr}(Z_1) + s_3 = t,$$

$$Z_2 - Z_1 + \begin{pmatrix} 0 & (\Lambda^{-1/2}C\Delta^{-1/2})' \\ (\Lambda^{-1/2}C\Delta^{-1/2}) & 0 \end{pmatrix} - (s_1 - s_2)I = \begin{pmatrix} 0 & -(Q\hat{B}^{\mathrm{LS}}P)' \\ -(Q\hat{B}^{\mathrm{LS}}P) & 0 \end{pmatrix},$$

$$Z_1, Z_2 \succeq 0,$$

$$(M, C_{11}, \ldots, C_{1q}, C_{21}, \ldots, C_{pq})' \in \mathcal{L}^{pq+1},$$

$$\mathbf{s} \in \mathcal{R}^3_+.$$

References

Anderson, T. (1951) Estimating linear restrictions on regression coefficients for multivariate normal distributions. *Ann. Math. Statist.*, **22**, 327–351.

Anderson, T. (2003) *An Introduction to Multivariate Statistical Analysis*, 3rd edn. New York: Wiley.

Bakin, S. (1999) Adaptive regression and model selection in data mining problems. *PhD Thesis*. Australian National University, Canberra.

Bedrick, E. and Tsai, C. (1994) Model selection for multivariate regression in small samples. *Biometrics*, **50**, 226–231.

Ben-Tal, A. and Nemirovski, A. (2001) *Lectures on Modern Convex Optimization: Analysis, Algorithms, Engineering Applications*. Philadelphia: Society for Industrial and Applied Mathematics.

Breiman, L. (1996) Heuristics of instability and stabilization in model selection. *Ann. Statist.*, **24**, 2350–2383.

Breiman, L. and Friedman, J. H. (1997) Predicting multivariate responses in multiple linear regression (with discussion). *J. R. Statist. Soc.* B, **59**, 3–54.

Brooks, R. and Stone, M. (1994) Joint continuum regression for multiple predictands. *J. Am. Statist. Ass.*, **89**, 1374–1377.

Brown, P., Fearn, T. and Vannucci, M. (1999) The choice of variables in multivariate regression: a non-conjugate Bayesian decision theory approach. *Biometrika*, **86**, 635–648.

Brown, P. J., Vannucci, M. and Fearn, T. (1998) Multivariate Bayesian variable selection and prediction. *J. R. Statist. Soc.* B, **60**, 627–641.

Brown, P. J., Vannucci, M. and Fearn, T. (2002) Bayes model averaging with selection of regressors. *J. R. Statist. Soc.* B, **64**, 519–536.

Eilers, P. and Marx, B. (1996) Flexible smoothing with B-splines and penalties (with discussion). *Statist. Sci.*, **11**, 89–121.

Frank, I. and Friedman, J. (1993) A statistical view of some chemometrics regression tools (with discussion). *Technometrics*, **35**, 109–148.

Fujikoshi, Y. and Satoh, K. (1997) Modified AIC and C_p in multivariate linear regression. *Biometrika*, **84**, 707–716.

Golub, G., Heath, M. and Wahba, G. (1979) Generalized cross validation as a method for choosing a good ridge parameter. *Technometrics*, **21**, 215–224.

Hastie, T. and Tibshirani, R. (1990) *Generalized Additive Models*. London: Chapman and Hall.

Horn, R. and Johnson, C. (1991) *Topics in Matrix Analysis*. Cambridge: Cambridge University Press.

Hotelling, H. (1935) The most predictable criterion. *J. Educ. Psychol.*, **26**, 139–142.

Hotelling, H. (1936) Relations between two sets of variables. *Biometrika*, **28**, 321–377.

Izenman, A. (1975) Reduced-rank regression for the multivariate linear model. *J. Multiv. Anal.*, **5**, 248–264.

Johnstone, I. (2001) On the distribution of the largest eigenvalue in principal components analysis. *Ann. Statist.*, **29**, 295–327.

Lutz, R. and Bühlmann, P. (2006) Boosting for high-multivariate responses in high-dimensional linear regression. *Statist. Sin.*, **16**, 471–494.

Massy, W. (1965) Principal components regression with exploratory statistical research. *J. Am. Statist. Ass.*, **60**, 234–246.

Reinsel, G. (1997) *Elements of Multivariate Time Series Analysis*, 2nd edn. New York: Springer.

Reinsel, G. and Velu, R. (1998) *Multivariate Reduced-rank Regression: Theory and Applications*. New York: Springer.

Ruppert, D. and Carroll, R. (1997) Penalized regression splines. *Technical Report*. Cornell University, Ithaca.

Smith, H., Gnanadesikan, R. and Hughes, J. (1962) Multivariate analysis of variance (ANOVA). *Biometrics*, **18**, 22–41.

Tibshirani, R. (1996) Regression shrinkage and selection via the lasso. *J. R. Statist. Soc.* B, **58**, 267–288.

Turlach, B., Venables, W. and Wright, S. (2005) Simultaneous variable selection. *Technometrics*, **47**, 349–363.

Tütüncü, R., Toh, K. and Todd, M. (2003) Solving semidefinite-quadratic-linear programs using SDPT3. *Math. Progrmmng*, **95**, 189–217.

Wold, H. (1975) Soft modeling by latent variables: the nonlinear iterative partial least squares approach. In *Perspectives in Probability and Statistics: Papers in Honour of M. S. Bartlett* (ed. J. Gani). New York: Academic Press.

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