Dimension Reduction and Coefficient Estimation in the Multivariate Linear Regression

Ming Yuan[a,1], Ali Ekici[a], Zhaosong Lu[b] and Renato Monteiro[a]

[a]*School of Industrial and Systems Engineering, Georgia Institute of Technology*

[b]*Department of Mathematical Sciences, Carnegie Mellon University*

(March 10, 2006)

Abstract

In this paper, 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 proposed method can be formulated as a novel penalized least squares estimate. The penalty we employed is the coefficient matrix's Ky Fan norm. Such penalty encourages the sparsity among singular values and at the same time gives shrinkage coefficient estimates, thus conducts dimension reduction and coefficient estimation simultaneously in the multivariate linear model. We also propose a GCV type criterion for the selection of the tuning parameter in the penalized least squares. Simulations and an application in financial econometrics demonstrate the competitive performance of the new method. An extension to the nonparametric factor model is also discussed.

Key words: penalized likelihood, Ky Fan norm, conic programming, dimension reduction, group variable selection.

1 Introduction

Multivariate linear regressions are routinely used in chemometrics, econometrics, financial engineering, psychometrics and many other areas of applications to model the predictive relation-

[1] Address for correspondence: Ming Yuan, School of Industrial and Systems Engineering, Georgia Institute of Technology, 755 Ferst Drive NW, Atlanta, GA 30332-0205 (E-mail: myuan@isye.gatech.edu).

ships of multiple related responses on a set of predictors. In the 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 a $n \times q$ matrix, $X = (\mathbf{x}_1, \ldots, \mathbf{x}_n)'$ is a $n \times p$ matrix, B is a $p \times q$ coefficient matrix, $E = (\mathbf{e}_1, \ldots, \mathbf{e}_n)'$ is the regression noise, and $\mathbf{e}'$s are independently sampled from $\mathcal{N}(0, \Sigma)$. Throughout the paper, we center each input variable so that there is no intercept in (1) and also scale each input variable so that the observed standard deviation is one.

The standard approach to estimating the coefficient matrix B is by means of the 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 sub-optimally 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 method is the linear factor regression in which the response Y is regressed against a small number of linearly transformed predictors, 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 $r \leq \min\{p, q\}$, and Ω is a $r \times q$ matrix. The columns of F, $F_j (j = 1, \ldots, r)$, represent the so-called factors. Clearly (2) is an alternative representation of (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 the least squares for (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 (Wood, 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 (2). For smaller numbers of factors, 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 (2) through either hypothesis testing or cross validation. The coefficient matrix is typically estimated based on the selected number of factors. 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 upon the least squares. Variable selection and ridge regression are among the most popular ones. Both types of method are most often studied in the special case of (1) when $q = 1$, which amounts to the 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, Vannucci and Fearn, 1998; Brown, Fearn and Vannucci, 1999; Brown, Vannucci and Fearn, 2002; Turlach, Venables and Wright, 2005; Lutz and Bühlmann, 2005). Variable selection is most powerful when there are many redundant predictors common to all responses, which can be unrealistic in many applications of the multivariate linear model (Reinsel and Velu, 1998). Ridge regression, on the other hand, oftentimes can not offer easily interpretable models because it does not perform dimension reduction and all elements of the estimated coefficient matrix are typically nonzero.

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

The rest of the paper is organized as follows. The proposed methodology is introduced in the next section. An algorithm for solving the optimization problem in our formulation is relegated to the Appendix. Our algorithm takes advantage of the recent advance in convex optimization by deriving an equivalent second order cone program of the optimization problem in our formulation, which is readily solvable using the standard software. We consider the special case of orthogonal design in Section 3 to better understand the new estimate. A GCV type statistic is introduced in Section 4 to choose the optimal tuning parameter for the proposed method. Simulations and a real world example are given in Sections 5 and 6 to illustrate the methodology. The proposed method can also be extended to the nonparametric 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 Y_j, B_j and E_j the jth columns of Y, B and E respectively. From (1), the jth response can be modelled by

$$Y_j = XB_j + E_j, \tag{3}$$

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 $\{\eta_1, \ldots, \eta_p\}$ 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 $\text{span}\{\cdot\}$ stands for the linear space spanned by a set of vectors. Both variable selection and linear factor model (2) can be formulated in this framework. In variable selection η's 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 $F_i = X\eta_i$ 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 η's as factor estimation and the identification of $\mathcal{A}$ as factor selection. In this paper, we propose a procedure that imposes less restrictions than variable selection and linear factor regression by allowing both η's and $\mathcal{A}$ to be completely determined by the data.

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

$$Y = F\Omega + E, \tag{4}$$

where Ω is a $p \times q$ matrix such that $(\eta_1, \ldots, \eta_p)\Omega = B$. Now factor selection for (1) can be cast as a variable selection problem for (4). As pointed out by Turlach, Venables and Wright (2005), a family of estimate for this purpose can be obtained by

$$\min \operatorname{trace}\left\{(Y - F\Omega)' W (Y - F\Omega)\right\} \qquad \text{subject to} \qquad \sum_{i=1}^{p} ||\omega_i||_\alpha \leq t, \tag{5}$$

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

$$||\omega_i||_\alpha = \left(\Omega_{i1}^{\alpha} + \ldots + \Omega_{iq}^{\alpha}\right)^{1/\alpha}. \tag{6}$$

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 (5) reduces to the popular Lasso (Tibshirani, 1996) when $q = 1$. Similar to the Lasso, if t is appropriately chosen, minimizing (5) yields a shrinkage estimate that is sparse in the sense that some of the ω_i's will be set to zero. Consequently, the ith factor will be included in the final estimate if and only if ω_i is nonzero. Therefore, factor selection and coefficient estimation are done simultaneously. Two most obvious choices of 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, Venables and Wright (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 (5) to be invariant to any orthogonal transformation of the responses, which can be useful in many practical situations.

In order to use (5), we need to obtain η's first. Similar to variable selection, factor selection is most powerful if all responses can be predicted by a small subset of common factors. Ideally, one wants $\{\eta_1, \ldots, \eta_p\}$ to contain a set of basis of $\mathcal{B}$ to allow the sparsest representation of B in the factor space. This is typically not the case for the existing linear factor regression methods. For example, in the 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 η's to be the eigenvectors of BB'. Clearly this set of basis contains basis that span $\mathcal{B}$. 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 $||\omega_i||_2 = \sigma_i(B)$. Therefore, (5) with $\alpha = 2$ gives

$$\min \operatorname{trace}\left\{(Y - XB)'(Y - XB)\right\} \qquad \text{subject to} \qquad \sum_{i=1}^{\min\{p,q\}} \sigma_i(B) \leq t, \tag{7}$$

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 η's is required in (7) and we shall use the minimizer of (7) as our final estimate of B. An efficient algorithm for minimizing (7) is presented in the Appendix. The penalty we employed in (7) encourages the sparsity among the singular values of B and at the same time gives shrinkage estimates for U and V, thus conducts dimension reduction and estimation simultaneously in the multivariate linear model.

The proposed estimate defined as the minimizer of (7) is closely connected with several other popular methods. In particular, (7), reduced rank regression and ridge regression can all be viewed as the minimizer of

$$\operatorname{trace}\left\{(Y - XB)'(Y - XB)\right\} \qquad \text{subject to} \qquad \left(\sum_i \sigma_i^\alpha(B)\right)^{1/\alpha} \leq t \tag{8}$$

with difference choices of α.

The ridge regression defined as the minimizer of

$$||Y - XB||^2 + \lambda \operatorname{trace}(B'B) \tag{9}$$

corresponds to $\alpha = 2$ because $\operatorname{trace}(B'B) = \sum \sigma_i^2(B)$. It is well known that the ridge regression provides shrinkage estimate that often outperforms the least squares. The proposed estimate, corresponding to $\alpha = 1$, enjoys the 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 (7) can now be expressed as

$$\min \operatorname{trace}\left\{(Y - XB)'(Y - XB)\right\} \qquad \text{subject to} \qquad (B'B)^{1/2} \leq t, \tag{10}$$

which is nothing else but the usual ridge regression.

The reduced rank regression is another special case of (8) with $\alpha = 0^+$. Both (7) and the reduced rank regression set some of the singular values of B to zero and lead to estimate 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 overfit or highly correlated settings, the reduced rank regression may suffer in such situations as well. In contrast, the new method gives shrinkage estimate that overcomes this problem.

3 Orthogonal Design

To further understand the statistical properties of the proposed method, we consider the special case of orthogonal design. The following lemma gives an explicit expression to the minimizer of (7) in this situation.

Lemma 1 *Let $\widehat{U}^{\mathrm{LS}}\widehat{D}^{\mathrm{LS}}\widehat{V}^{\mathrm{LS}}$ be the singular value decomposition of the least squares estimate $\widehat{B}^{\mathrm{LS}}$. Then under the orthogonal design where $X'X = nI$, the minimizer of (7) is $\widehat{B} = \widehat{U}^{\mathrm{LS}}\widehat{D}\left(\widehat{V}^{\mathrm{LS}}\right)'$, where $\widehat{D}_{ij} = 0$ if $i \neq j$, $\widehat{D}_{ii} = \max\{\widehat{D}^{\mathrm{LS}}_{ii} - \lambda, 0\}$ and $\lambda \geq 0$ is a constant such that $\sum_i \widehat{D}_{ii} = \min\{t, \sum \widehat{D}^{\mathrm{LS}}_{ii}\}$.*

Proof. (7) can be equivalently written in a Lagrange form:

$$Q_n(B) = \frac{1}{2}\mathrm{trace}\left\{(Y - XB)'(Y - XB)\right\} + n\lambda \sum_{i=1}^{\min\{p,q\}} \sigma_i(B), \tag{11}$$

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

$$\begin{aligned} & \mathrm{trace}\left\{(Y - XB)'(Y - XB)\right\} \\ = \; & \mathrm{trace}\left((Y - X\widehat{B}^{\mathrm{LS}})'(Y - X\widehat{B}^{\mathrm{LS}})\right) + \mathrm{trace}\left\{\left(\widehat{B}^{\mathrm{LS}} - B\right)' X'X \left(\widehat{B}^{\mathrm{LS}} - B\right)\right\} \\ = \; & \mathrm{trace}\left((Y - X\widehat{B}^{\mathrm{LS}})'(Y - X\widehat{B}^{\mathrm{LS}})\right) + n\mathrm{trace}\left\{\left(\widehat{B}^{\mathrm{LS}} - B\right)'\left(\widehat{B}^{\mathrm{LS}} - B\right)\right\} \end{aligned} \tag{12}$$

Together with the fact that $\mathrm{trace}\{B'B\} = \sum_i \sigma_i^2(B)$, (11) equals to

$$\frac{1}{2}\sum_{i=1}^{q}\sigma_i^2(B) - \mathrm{trace}\left\{B'\widehat{B}^{\mathrm{LS}}\right\} + \lambda\sum_{i=1}^{q}\sigma_i(B), \tag{13}$$

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

$$\mathrm{trace}\left\{B'\widehat{B}^{\mathrm{LS}}\right\} \leq \sum \sigma_i(B)\widehat{D}^{\mathrm{LS}}_{ii} \tag{14}$$

Therefore,

$$Q_n(B) \geq \frac{1}{2}\sum_{i=1}^{q}\sigma_i^2(B) - \sum \sigma_i(B)\widehat{D}^{\mathrm{LS}}_{ii} + \lambda\sum_{i=1}^{q}\sigma_i(B) \tag{15}$$

Note that $\sigma_i(B) \geq 0$. The right hand side of (15) is minimized at

$$\sigma_i(B) = \max\{\widehat{D}_{ii}^{\mathrm{LS}} - \lambda, 0\}, \qquad i = 1, \ldots, q. \tag{16}$$

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

This closed-form minimizer of (7) allows for a better understanding of our estimate. Specifically, the following lemma indicates that we can always find an appropriate tuning parameter such that the nonzero singular values of B are consistently estimated and the rest are set to zero with probability one.

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

Proof. Note that

$$\widehat{B}^{\mathrm{LS}} = (X'X)^{-1}X'Y = X'(XB + E)/n = B + X'E/n. \tag{17}$$

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

$$\sigma_1\left(X'E/n\right) \leq \sigma_1\left(X'E\Sigma^{-1/2}/n\right)\sigma_1\left(\Sigma^{1/2}\right) \sim \sigma_1^{1/2}\left(\Sigma\right)\frac{\sqrt{p} + \sqrt{q}}{\sqrt{n}}. \tag{18}$$

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

$$\left|\sigma_i(B) - \sigma_i\left(\widehat{B}^{\mathrm{LS}}\right)\right| \leq \sigma_1\left(X'E/n\right) = O_p\left(\frac{\sqrt{p} + \sqrt{q}}{\sqrt{n}}\right). \tag{19}$$

Therefore, if λ goes to zero at a slower rate than the right hand side of (19), the proposed estimate can provide consistent estimate of the nonzero singular values of B, and at the same time shrink the rest of the singular values to zero. ▮

Lemma 1 also indicates that the singular values of the proposed method are shrunk in a similar fashion as the Lasso under orthogonal designs. The Lasso has proved highly successful in a number of studies, particularly in the case when the predictors are correlated and p is large relatively to the sample size. In Section 5, we show that the proposed estimate 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 (7). One common method used in practice is the cross validation, which of course can be computationally demanding in large scale problems. In this section, we develop a generalized cross validation (GCV; Golub, Heath and Wahba, 1979) type statistic for determining t.

We first characterize the equivalence between (7) and its Lagrange form, (11), since it is easier to work with (11) in deriving our GCV type statistic. Denote $\widehat{B}$ the minimizer of (7) and $\widehat{U}\widehat{D}\widehat{V}'$ its singular value decomposition. Note that (7) is equivalent to (11) and we can always find a λ such that $\widehat{B}$ is also the minimizer of (11). The following lemma explicitly describes the relationship between t and λ.

Lemma 3 *Write $\widehat{d}_i = \widehat{D}_{ii}$ for $i = 1, \ldots, \min\{p, q\}$. For any $t \leq \sum_i \widehat{d}_i$, the minimizer of (11) coincides with the minimizer of (7), $\widehat{B}$, if*

$$n\lambda = \frac{1}{\mathrm{card}\{\widehat{d}_i > 0\}} \sum_{\widehat{d}_i > 0} \left(\tilde{X}_i'\tilde{Y}_i - \tilde{X}_i'\tilde{X}_i\widehat{d}_i \right) \tag{20}$$

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

Proof. Note that

$$\begin{aligned} \sum_{i=1}^{\min\{p,q\}} \sigma_i(\widehat{B}) &= \sum_{i=1}^{\min\{p,q\}} \widehat{D}_{ii} \\ &= \sum_{i=1}^{p} \sigma_i(\widehat{B}K\widehat{B}') \\ &= \mathrm{trace}(\widehat{B}K\widehat{B}'), \end{aligned} \tag{21}$$

where

$$K = \sum_{\widehat{D}_{ii} > 0} \frac{1}{\widehat{D}_{ii}} \widehat{V}_i \widehat{V}_i', \tag{22}$$

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

$$\frac{1}{2}\mathrm{trace}\left\{ (Y - XB)'(Y - XB) \right\} + n\lambda \mathrm{trace}(BKB'). \tag{23}$$

From (23), $\widehat{d}$ is the minimizer of

$$\frac{1}{2} \sum_{i=1}^{\min\{p,q\}} \left(\tilde{Y}_i - \tilde{X}_i d_i \right)^2 + n\lambda \sum_{i=1}^{\min\{p,q\}} d_i, \tag{24}$$

subject to the constraint that $d_i \geq 0$. The first order optimality condition for (24) yields

$$n\lambda = \tilde{X}_i'\tilde{Y}_i - \tilde{X}_i'\tilde{X}_i\widehat{d}_i, \tag{25}$$

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

Since $\widehat{B}$ is the minimizer of (23), it can be expressed as $\widehat{B} = \left(X'X + 2n\lambda K\right)^{-1} X'Y$. Neglecting the fact that W also depends on $\widehat{B}$, we can define the hat matrix for (23) as $X\left(X'X + 2n\lambda K\right)^{-1} X'$ and the degrees of freedom as

$$\mathrm{df}(t) = q \times \mathrm{trace}\left(X\left(X'X + 2n\lambda K\right)^{-1} X'\right). \tag{26}$$

Now the GCV score is given by

$$\mathrm{GCV}(t) = \frac{\mathrm{trace}\left\{\left(Y - X\widehat{B}'\right)'\left(Y - X\widehat{B}'\right)\right\}}{qp - \mathrm{df}(t)}, \tag{27}$$

and we choose a tuning parameter by minimizing GCV(t).

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

```
(1) For each candidate $t$ value
    (a) compute the minimizer of (7), denote the solution $\widehat{B}(t)$
    (b) evaluate $\lambda$ using (20)
    (c) compute the GCV score (27)
(2) Denote $t^*$ the minimizer of the GCV score obtained in Step (1).
    Return $\widehat{B}(t^*)$ 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 we compared include

(i) (`FES`) The proposed method for factor estimation and selection with the tuning parameter selected by GCV;

(ii) (OLS) The ordinary least square estimate $(X'X)^{-1}X'Y$;

(iii) (CW) The Curd and Whey with GCV procedure developed by Breiman and Friedman (1997);

(iv) (RRR) Reduced-rank regression with the rank selected by ten fold cross-validation;

(v) (PLS) Two-block partial least squares (Wold, 1975) with the number of components selected by ten fold cross-validation;

(vi) (PCR) Principal components regression (Massy, 1965) with the number of components selected by ten fold cross-validation;

(vii) (RR) Ridge regression with the tuning parameter selected by ten fold cross-validation;

(iix) (CAIC) Forward selection using the corrected AIC proposed by Bedrick and Tsai (1994). The corrected AIC for a specific submodel of (1) is defined as

$$n \ln |\widehat{\Sigma}| + \frac{n(n+k)q}{n-k-q-1} + nq \ln 2\pi, \tag{28}$$

where k is the number of predictors included in the submodel and $\widehat{\Sigma}$ 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 $\widehat{B}$ is given by

$$ME(\widehat{B}) = \left(\widehat{B} - B\right)' V \left(\widehat{B} - B\right), \tag{29}$$

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

We consider the following four models.

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 a 8×8 random matrix whose elements are independently sampled from $\mathcal{N}(0, 1)$, 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 $n = 20$.

II. Same as (I) except that the singular values are $\sigma_1 = \ldots = \sigma_8 = 0.85$.

III. Same set-up as before, but with singular values $(5, 0, 0, 0, 0, 0, 0, 0)$.

IV. This is a larger problem with $p = 20$ predictors and $q = 20$ responses. Random coefficient matrix is generated in the same fashion as before with the first ten singular values being one and last ten singular values being zero. $\mathbf{x}$ and $\mathbf{y}$ are generated as the previous examples. The sample size is set to be $n = 50$.

For each of these models, two hundred datasets were simulated. Table 1 gives the means and standard errors (in parentheses) over the 200 simulated datasets. To gain further insight into the comparison, we also provide pairwise prediction accuracy comparison between `FES` and the other methods for Model I-IV (except for `CAIC`) in Figure 1.

	FES	OLS	CW	RRR	PLS	PCR	RR	CAIC
Model 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)
Model 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)
Model 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)
Model 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)

Table 1: Comparisons on the Simulated Datasets

We use these examples to examine the relative merits of the methods in different scenarios. Model I has a moderate number of moderate size singular values, the first two rows of Table 1 and the first column of Figure 1 indicate that `FES` enjoys the best prediction accuracy followed by the ridge regression and the Curd and Whey. Model II represents a different scenario with a full rank coefficient matrix. The table indicates that the ridge regression performs the best, with `FES` being the only other method to improve the ordinary least square by more than 50%. Model III has a small number of large singular values. In this case, `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. The ridge regression performs the best, with `FES` being a close second. Note that `CAIC` performs poorly for all four models because there are no redundant variables in general. In all examples, `FES` demonstrates competitive performance when compared with the other six methods. The ridge regression does a little better than the proposed method for Models II and IV, but worse for the other two models. It

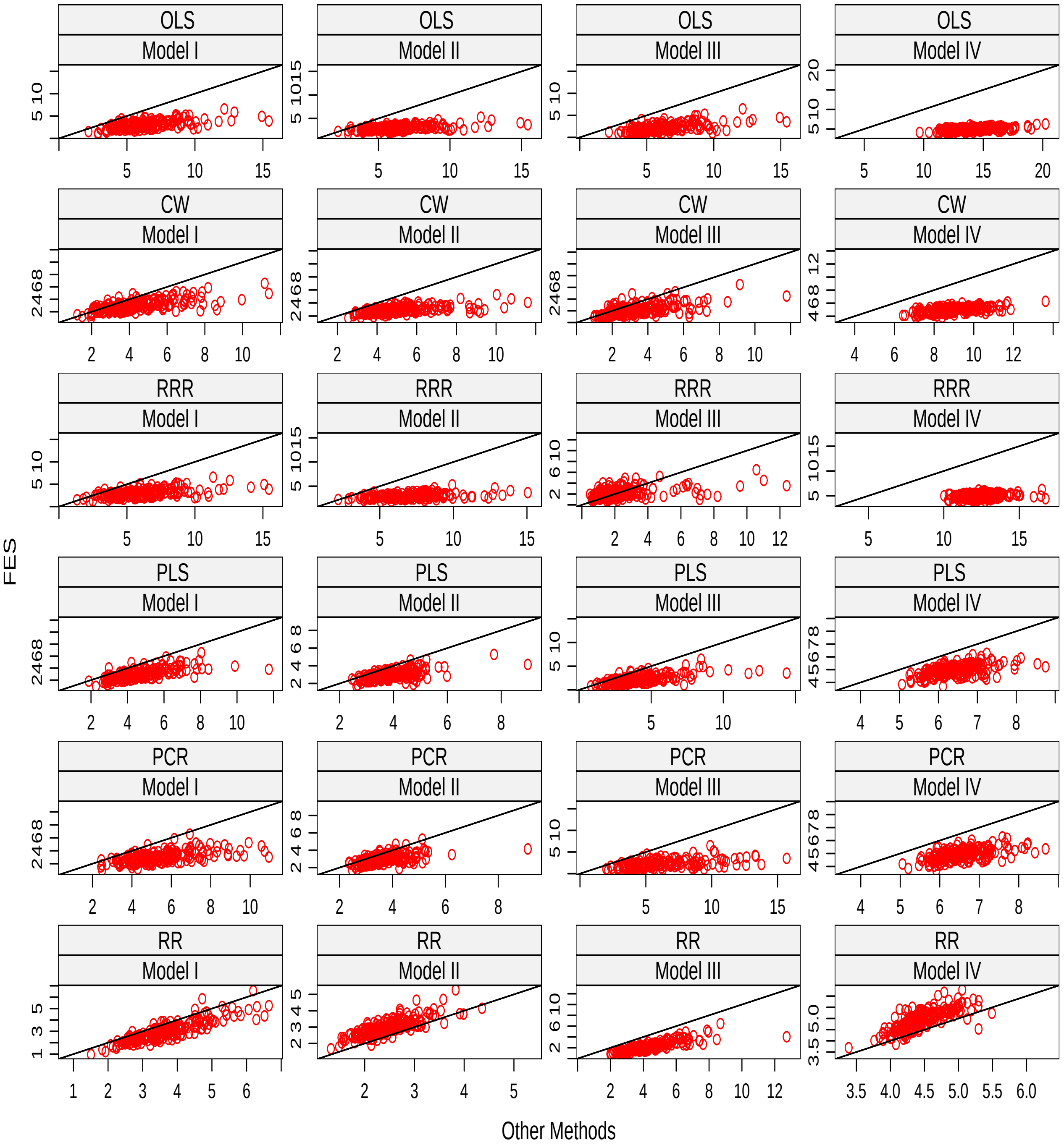

Figure 1: Pairwise Model Error Comparison between FES and Other Methods

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

6 Application

To demonstrate the utility of the proposed method, we now consider a real example in financial econometrics. 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 based on their historical performance. Vector autoregressive (VAR) models are often used for this purpose (Reinsel, 1997). Let $\mathbf{y}_t$ be the vector of returns at time t. The VAR model with order one is given by

$$\mathbf{y}_t = \mathbf{y}_{t-1} B + E. \tag{30}$$

Clearly (30) is a special case of the multivariate linear model. Accurate estimate of B in (30) leads to good forecasts which, in turn, can serve as instruments for efficient portfolio allocation and revealing arbitrage opportunities. Also important is the identification of the factors in (30), which can help construct benchmark portfolios or diversify investments.

To illustrate our method, we applied (30) to the stock prices in 2004 of the ten largest American companies ranked by Fortune magazine on the basis of their 2003 revenue. We 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 (30) 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. The left panel of Figure 2 gives the trajectory of the singular values of our proposed method as the Ky Fan norm of B increases and the right panel depicts the GCV curve. The grey line in the left panel corresponds to the Ky Fan norm selected by GCV.

In this example, GCV retains four nonzero singular values for B. The loadings of the four factors are given in Table 2.

It is of great interest to understand the meaning of these four factors. Note that from (30), the factors summarize the asset return history in predicting the future returns. The classical investment theory indicates that the market index is a good summary of the asset prices and

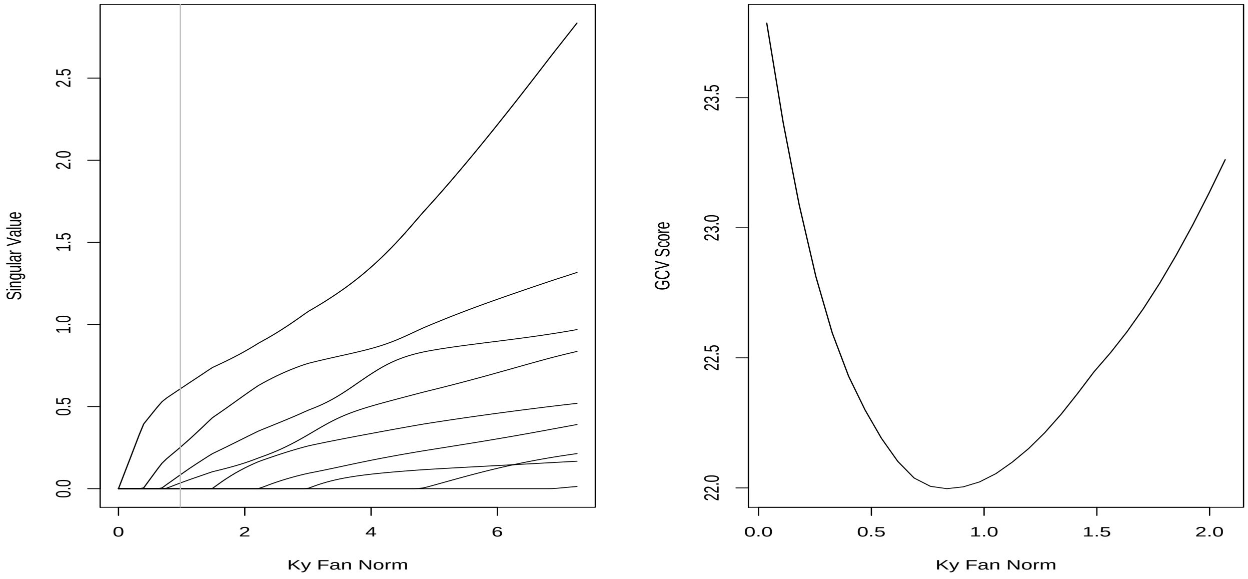

Figure 2: Solution Paths for the Stock Example

Factor	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

Table 2: Factor Loadings for the Stock Example

should lie in the factor space. To approximate the market index, we picked the S&P 500 and NASDAQ indices. To check if their returns approximately fall into the factor space estimated from the stock data, we constructed their best linear approximations in the estimated four dimensional factor space. The log returns of S&P500 and NASDAQ indices in the year of 2004 together with their approximations are given in Figure 3. Both approximations track the actual log return processes fairly well. This exercise confirms that the factors revealed by our method are indeed meaningful and should provide insight into further studies of the dynamics of the financial market.

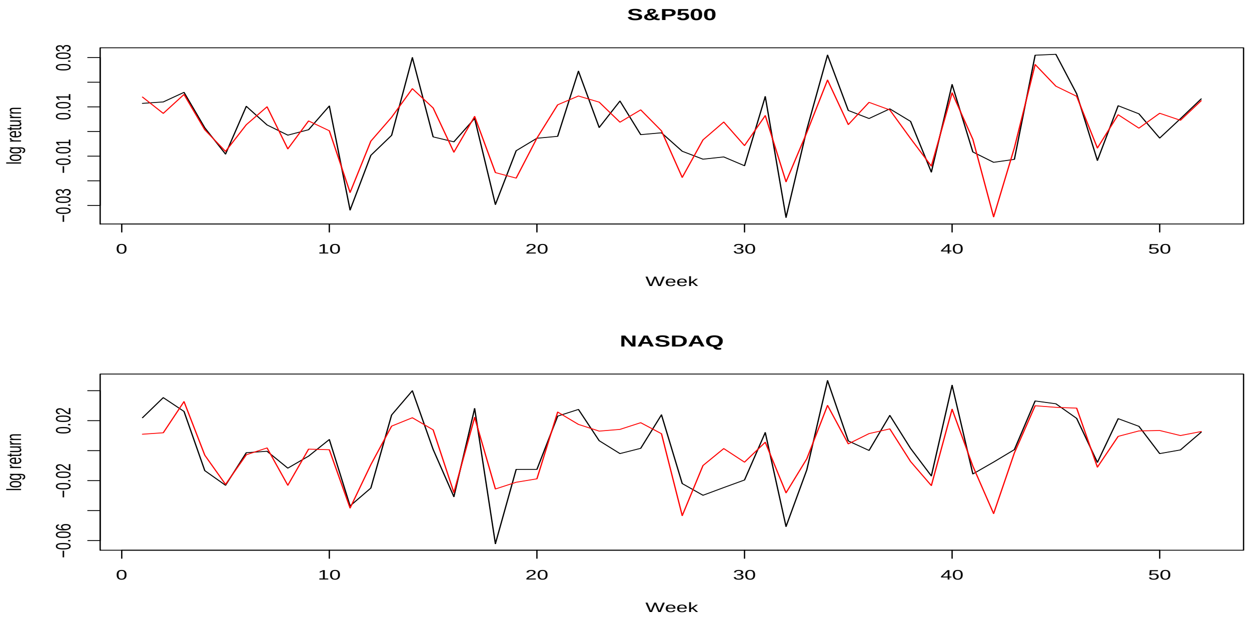

Figure 3: S&P500 and NASDAQ Indices (black lines) Together with Their Approximations in the Factor Space (red lines)

To further compare the proposed method 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 proposed method clearly provides better prediction than the other methods in this example.

	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

Table 3: Out-of-sample Mean Squared Error (×0.001)

7 Nonparametric Factor Model

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

$$y_j = g_j(\mathbf{x}) + e_j, \qquad j = 1, \ldots, q, \tag{31}$$

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

$$g_j(x) = \omega_{1j} f_1(x) + \omega_{2j} f_2(x) + \ldots + \omega_{rj} f_r(x), \tag{32}$$

where $\Omega = (\omega_{kj})_{r\times q}$ is unknown. To estimate g_j's, we model the nonparametric factors using regression splines:

$$f_k(x) = \beta_{1k} x + \ldots + \beta_{sk} x^s + \sum_{m=1}^{M} \beta_{m+s,k} \left(x - \kappa_m\right)_+^s, \qquad k = 1, \ldots, r, \tag{33}$$

where $s \geq 1$ is an integer, $(u)_+^s = u^s I(u \geq 0)$, 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 (31) can be rewritten as

$$\mathbf{y} = \mathbf{z}B + \mathbf{e}, \tag{34}$$

which is in the form of the multivariate linear regression (1). In the 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 nonlinear 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 (Eliers 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 nonparametric factor model can be defined as the solution of

$$\min_B \operatorname{trace}\left\{(Y - ZB)'(Y - ZB)\right\} \qquad \text{subject to} \qquad \sum_{k=1}^{r}\left(\sum_{i=1}^{M+s}\beta_{ik}^2\right)^{1/2} \leq t, \tag{35}$$

where $Y = (\mathbf{y}_1, \ldots, \mathbf{y}_n)'$ and $Z = (\mathbf{z}_1, \ldots, \mathbf{z}_n)'$. For identifiability purpose, we further assume that $A'A = \Omega\Omega' = I_r$. Then (35) is equivalent to

$$\min_B \operatorname{trace}\left\{(Y - ZB)'(Y - ZB)\right\} \qquad \text{subject to} \qquad \sum_i \sigma_i(B) \leq t, \tag{36}$$

which is of the same form as (7) and can also be solved using the algorithm provided in the appendix.

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's are written as

$$g_j(\mathbf{x}) = g_{j1}(x_1) + \ldots + g_{jp}(x_p), \qquad j = 1, \ldots, q, \tag{37}$$

where $g_{j1}, \ldots, g_{jp}$ are univariate functions. Consider a nonparametric factor model for each component on the right hand side of (37):

$$g_{ji}(x_i) = \omega_{1j}^{(i)} f_{i1}(x_i) + \omega_{2j}^{(i)} f_{i2}(x_i) + \ldots + \omega_{r_ij}^{(i)} f_{ir_i}(x_i), \tag{38}$$

where

$$f_{ik}(x_i) = \beta_{1k}^{(i)} x_i + \ldots + \beta_{sk}^{(i)} x_i^s + \sum_{m=1}^{M} \beta_{m+s,k}^{(i)}\left(x_i - \kappa_m^{(i)}\right)_+^s, \qquad k = 1, \ldots, r_i. \tag{39}$$

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

$$\mathbf{y} = \mathbf{z}_1 B_1 + \ldots + \mathbf{z}_p B_p + \mathbf{e}. \tag{40}$$

Similar to (36), we define our estimate of B_i's as the solution of

$$\min_{B_1,\ldots,B_p} \operatorname{trace}\left\{(Y-ZB)'(Y-ZB)\right\} \qquad \text{subject to} \qquad \sum_j \sigma_j(B_i) \le t_i, j = i, \ldots, p, \tag{41}$$

where $Z = (Z_1, \ldots, Z_p)$ and $B = (B_1', \ldots, B_p')'$. Using the algorithm presented in the appendix, (41) can be solved in an iterative fashion.

```
(1) Initialize $B_i = 0$, $i = 1, \ldots, p$.
(2) For $i = 1$ to $p$
    (i) Compute $Y^* = Y - Z_1B_1 - \ldots - Z_{i-1}B_{i-1} - Z_{i+1}B_{i+1} - \ldots - Z_pB_p$.
    (ii) Update $B_i$ by minimizing trace$\{(Y^* - Z_iB_i)'(Y^* - Z_iB_i)\}$ subject to $\sum_j \sigma_j(B_i) \le t_i$.
(3) Repeat (2) until $B$ does not change.
```

To illustrate, we re-analyze the biochemical data from Smith et al. (1962). The data contain chemical measurements on several characteristics of 33 individual samples of man urine specimens. There are five response variables, pigment creatinine, concentrations of phosphate, phosphorous, creatinine, and choline. The goal of the analysis is to relate these responses to three predictors, weight of the subject, volume and specific gravity. Reinsel and Velu (1998) postulated a multivariate linear model to analyze the data. A nonparametric extension such as (38) could be more powerful if nonlinear effects of the predictors are suspected. To this end, we model the effect of each of the predictor by (33) with $s = 2$ and $M = 5$. The 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(ii) in each iteration using the GCV criterion developed before. Figure 4 shows the estimated effect of each predictor on each response. Clear departure from 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. Based on this formulation, we proposed a new method

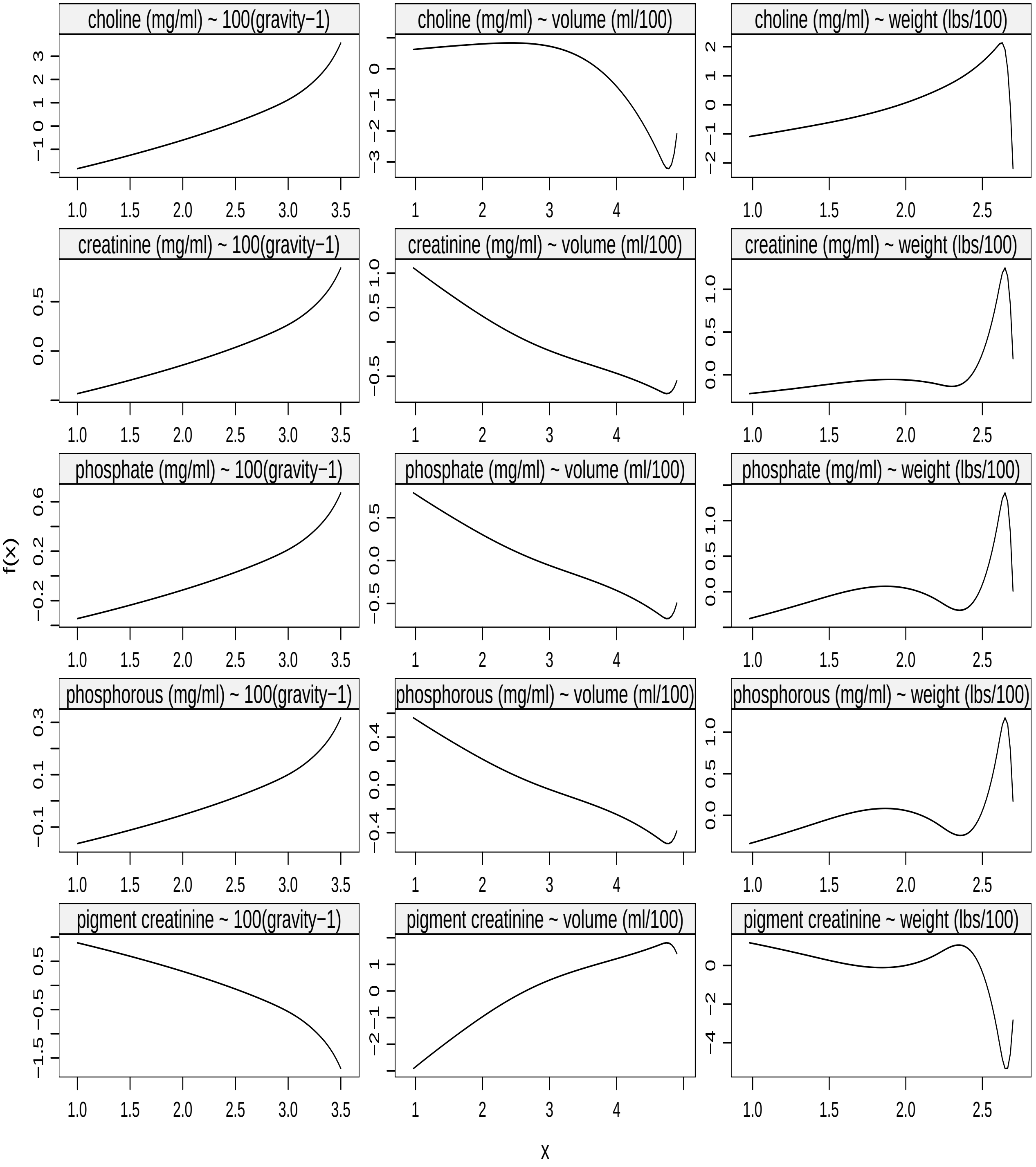

Figure 4: Fitted Components for Biochemistry Data

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 nonparametric 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. On the other hand, ridge regression often enjoys superior prediction accuracy because it leads to shrinkage estimate, but does not provide easily interpretable models. Our method combines and retains the advantages of both approaches. Formulated as a penalized least squares estimate, the proposed method gives shrinkage estimate with reduced ranks. We demonstrate by numerical examples that the proposed method enjoys competitive performance when compared with other popular methods.

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

References

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

[2] Anderson, T. (2003), *An Introduction to Multivariate Statistical Analysis (3rd edition)*, Wiley, New York.

[3] Bakin, S. (1999) Adaptive regression and model selection in data mining problems, *unpublished PhD thesis*, Australian National University.

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

[5] Ben-Tal, A. and Nemirovski, A. (2001), *Lectures on Modern Convex Optimization: Analysis, Algorithms, Engineering Applications*, SIAM, Philadelphia.

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

[7] Breiman, L. and Friedman, J. (1997), Predicting multivariate responses in multiple linear regression, *J. R. Statist. Soc. B*, **59**, 3-54.

[8] Brooks, R. and Stone, M. (1994), Joint continuum regression for multiple predictands, *J. Amer. Statist. Assoc.*, **89**, 1374-1377.

[9] 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.

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

[11] Brown, P., Vannucci, M. and Fearn, T. (2002), Bayesian model averaging with selection of regressors, *J. R. Statist. Soc. B*, **64**, 519-536.

[12] Eliers, P. and Marx, B. (1996), Flexible smoothing with B-splines and penalties (with discussion), *Statist. Sci.*, **11**, 89-121.

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

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

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

[16] Hastie, T. and Tibshirani, R. (1990), Generalized Additive Models, Chapman & Hall, London.

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

[18] Hotelling, H. (1935), The most predictable criterion, *J. Edu. Psych.*, **26**, 139-142.

[19] Hotelling, H. (1936), Relations between two sets of variables, *Biometrika*, **28**, 321-377.

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

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

[22] Lutz, R. and Bühlmann P. (2005), Boosting for high-multivariate responses in high-dimensional linear regression, to appear in *Statist. Sinica*.

[23] Massey, W. (1965), Principal components regression with exporatory statistical research, *J. Amer. Statist. Assoc.*, **60**, 234-246.

[24] Reinsel, G. (1997), *Elements of Multivariate Time Series Analysis (2nd edition)*, Springer-Verlag, New York.

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

[26] Ruppert, D. and Carroll, R. (1997), Penalized regression splines, *Technical Report*.

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

[28] Tibshirani, R. (1996), Regression shrinkage and selection via the lasso, *J. Royal. Statist. Soc. B.*, **58**, 267-288.

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

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

[31] 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), Academic Press, New York.

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

Appendix – Algorithm for Solving (7)

To solve (7), we take advantage of the recent advance in convex optimization. We show that (7) is equivalent to a second order cone program and can be solved using standard solvers such as SDPT3 (Tütüncü, Toh and Todd, 2003).

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

$$\mathcal{L}^m = \left\{ \mathbf{x} = (x_1, \ldots, x_m) \in \mathcal{R}^m : x_1 \geq \sqrt{x_2^2 + \ldots + x_m^2} \right\}. \tag{42}$$

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

$$\text{svec}(X) = (X_{11}, \sqrt{2}X_{21}, X_{22}, \ldots, \sqrt{2}X_{n1}, \ldots, \sqrt{2}X_{n,n-1}, X_{nn})'. \tag{43}$$

SDPT3 can solve problems of the following form:

$$\begin{aligned} \min_{X_j^s, \mathbf{x}_i^q, \mathbf{x}^l} \quad & \sum_{j=1}^{n_s} \text{trace}(C_j^s X_j^s) + \sum_{i=1}^{n_q} (\mathbf{c}_i^q)' \mathbf{x}_i^q + (\mathbf{c}^l)'\mathbf{x}^l \\ \text{such that} \quad & \sum_{j=1}^{n_s} (A_j^s)' \text{svec}(X_j^s) + \sum_{i=1}^{n_q} (A_i^q)' \mathbf{x}_i^q + (A^l)'\mathbf{x}^l = b, \\ & X_j^s \succeq 0 \quad \forall j, \mathbf{x}_i^q \in \mathcal{L}^{q_i} \quad \forall i, \mathbf{x}^l \in \mathcal{R}_+^{n_l}. \end{aligned} \tag{44}$$

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 a n_l dimensional vector, and the dimensions of matrices A's and vector b are clear from the context.

Next we show that (7) can be equivalently written in the form of (44). Similar to (12), the objective function of (7) can be rewritten as

$$\text{trace}\left((B - \widehat{B}^{\text{LS}})'X'X(B - \widehat{B}^{\text{LS}})\right) = \text{trace}(C'C) \tag{45}$$

up to a constant free of B where $C = (B - \widehat{B}^{\text{LS}})Q\Lambda^{1/2}$ and $Q\Lambda Q'$ is the eigenvalue decomposition of $X'X$. By the definition of the second order cone, (7) can be equivalently written as

$$\min_{M,C,B} \quad M$$

$$\text{such that} \qquad (M, C_{11}, \ldots, C_{1p}, C_{21}, \ldots, C_{qp})' \in \mathcal{L}^{pq+1}$$

$$\sum_{i=1}^{q} \sigma_i(B) \le t, \qquad C = (B - \widehat{B}^{\mathrm{LS}})Q\Lambda^{1/2} \tag{46}$$

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

$$\sum_{i=1}^{\min\{p,q\}} \mu_i(A) \le t \tag{47}$$

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

$$A = \begin{pmatrix} 0 & (BQ)' \\ (BQ) & 0 \end{pmatrix}. \tag{48}$$

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

$$qs + \mathrm{trace}(Z) \ \le \ t \tag{49}$$

$$Z - \begin{pmatrix} 0 & (C\Lambda^{-1/2} + \widehat{B}^{\mathrm{LS}}Q)' \\ (C\Lambda^{-1/2} + \widehat{B}^{\mathrm{LS}}Q) & 0 \end{pmatrix} + sI \ \succeq \ 0 \tag{50}$$

$$Z \ \succeq \ 0 \tag{51}$$

Now, (7) is equivalent to

$$\begin{aligned}
\min_{M, C, \mathbf{s}, Z_1, Z_2} \quad & M \\
\text{subject to} \quad & q(s_1 - s_2) + \mathrm{trace}(Z_1) + s_3 = t \\
& Z_2 - Z_1 + \begin{pmatrix} 0 & (C\Lambda^{-1/2})' \\ (C\Lambda^{-1/2}) & 0 \end{pmatrix} - (s_1 - s_2)I \\
& \qquad = \begin{pmatrix} 0 & -(\widehat{B}^{\mathrm{LS}}Q)' \\ -(\widehat{B}^{\mathrm{LS}}Q) & 0 \end{pmatrix} \\
& Z_1, Z_2 \succeq 0 \\
& (M, C_{11}, \ldots, C_{1p}, C_{21}, \ldots, C_{qp})' \in \mathcal{L}^{pq+1} \\
& \mathbf{s} \in \mathcal{R}^3_+
\end{aligned} \tag{52}$$

which is readily computable using SDPT3.