Dimension Reduction and Coefficient Estimation in Multivariate Linear Regression

Ganchao Wei

November 17, 2021

Overview

1. Introduction
2. Factor Estimation and Selection
3. Orthogonal Design
4. Tuning
5. Simulation
6. Application
7. Non-parametric Factor Model

Introduction

Multivariate linear regression:

$$Y = XB + E$$

, where $Y \in \mathbb{R}^{n \times q}$, $X \in \mathbb{R}^{n \times p}$ and $B \in \mathbb{R}^{p \times q}$ To reduce dimension, rewrite the model as the **linear factor regression**:

$$Y = XB + E = X\Gamma\Omega + E = F\Omega + E$$

, where $\Gamma \in \mathbb{R}^{p \times r}$ for some $r \leq \min(p, q)$. The columns of F represent the so-called factors.

In this paper, they propose a method simultaneously (1) choose the number of factors, (2) determine the factors and (3) estimate the factor loading Ω.

Factor Estimation and Selection

Let $\{\eta_1, \ldots, \eta_p\}$ be a set of basis for $\mathbb{R}^p$. If columns of B come from a linear space $\mathcal{B} \subseteq span\{\eta_i : i \in \mathcal{A} \subset \{1, \ldots, p\}\}$, then we can do dimension reduction.

Assume $\{\eta_1, \ldots, \eta_p\}$ are known, let $F = (F_1, \ldots, F_p)$, where $F_i = X\eta_i$, then

$$Y = F\Omega + E$$

,where $\Omega \in \mathbb{R}^{p \times q}$, s.t., $\{\eta_1, \ldots, \eta_p\}\Omega = B$. Then the factor selection problem can be reformed as:

$$\min\{tr\{(Y - F\Omega)W(Y - F\Omega)'\}\} \text{ subject to} \sum_{i=1}^{p} ||\omega_i||_\alpha \leq t$$

, where W is a weight matrix (assume $W = I$ in this paper), ω_i is the ith row of Ω, $t \geq 0$ is a regularization parameter and $|| \cdot ||_\alpha$ is the l_α-norm. (They choose $\alpha = 2$, since the optimization problem is invariant to orthogonal transformation of the response)

Factor Estimation and Selection

In this paper, they let $\{\eta_i\}$ be the eigenvectors of BB'.

Denote the SVD of B as $B = UDV'$, where $V \in \mathbb{R}^{p \times p}$. Then columns of U form $\{\eta_i\}$. Further, let $D_{ii} = \sigma_i(B)$ be the ith largest singular value. Then $\Omega = DV'$ and $\omega_i = \sigma_i(B)V_i$. Clearly, $||\omega_i||_2 = \sigma_i(B)$ and then the previous objective function can be rewritten as:

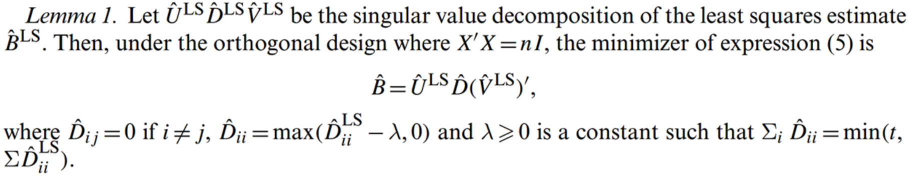
$$\min\{tr\{(Y - XB)(Y - XB)'\}\} \text{ subject to } \sum_{i=1}^{\min(p,q)} \sigma_i(B) \leq t$$

, where $\sum_{i=1}^{\min(p,q)} \sigma_i(B)$ is the Ky Fan norm of B. This is equivalent to a conic program and can be computed efficiently.

The proposed method is closely related to other popular methods, such as reduced rank regression (RRR) and ridge regression.

Orthogonal Design

To understand further the statistical properties of the method, consider the special case of orthogonal design.



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

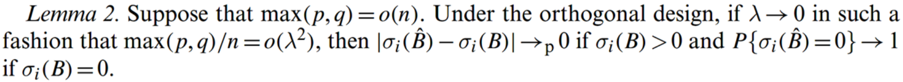
$$\hat{B} = \hat{U}^{\mathrm{LS}}\hat{D}(\hat{V}^{\mathrm{LS}})',$$

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 $\Sigma_i \hat{D}_{ii} = \min(t, \Sigma \hat{D}_{ii}^{\mathrm{LS}})$.

Lemma 1 gives an explicit expression for the minimizer.

Orthogonal Design

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 w.p.1.



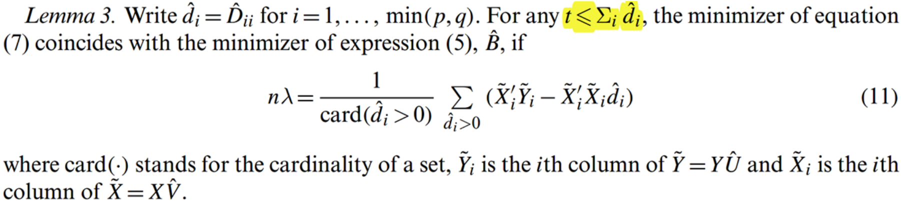
Lemma 2. Suppose that $\max(p,q)=o(n)$. Under the orthogonal design, if $\lambda \to 0$ in such a fashion that $\max(p,q)/n=o(\lambda^2)$, then $|\sigma_i(\hat{B})-\sigma_i(B)| \to_{\mathrm{p}} 0$ if $\sigma_i(B)>0$ and $P\{\sigma_i(\hat{B})=0\} \to 1$ if $\sigma_i(B)=0$.

Tuning

Tuning parameter: t

Can choose by CV, but is cumbersome. Here, they use GCV type of statistic to determine t.

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

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

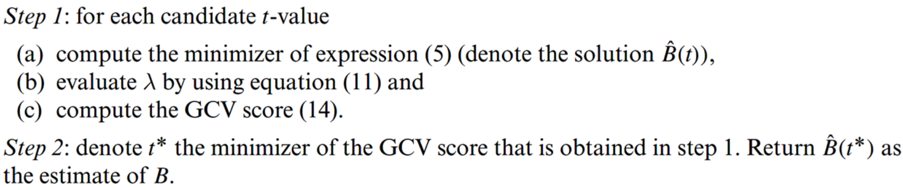
where card$(\cdot)$ stands for the cardinality of a set, $\tilde{Y}_i$ is the ith column of $\tilde{Y} = Y\hat{U}$ and $\tilde{X}_i$ is the ith column of $\tilde{X} = X\hat{V}$.

Tuning

The minimized B can be expressed as $\hat{B} = (X'X + 2n\lambda K)^{-1}X'Y$ and the GCV score is given by

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

In summary:



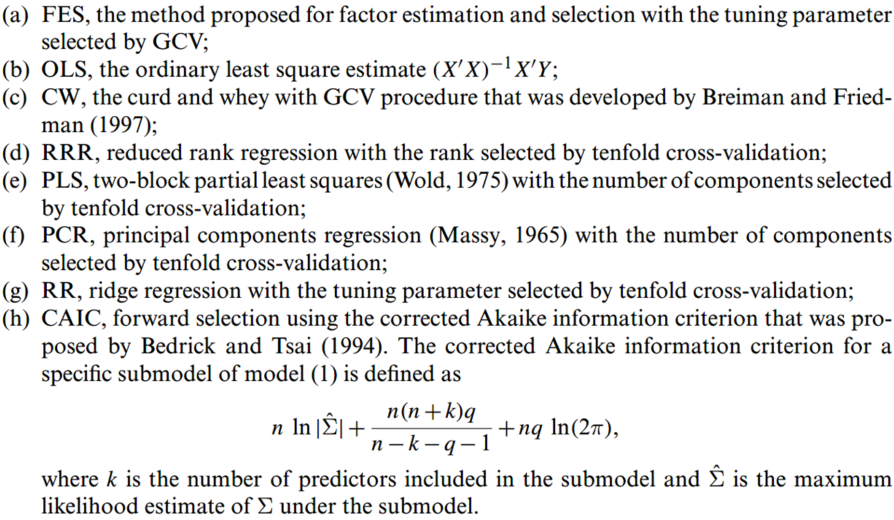
Step 1: for each candidate t-value

(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.

Simulation

The following methods are compared:



(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;

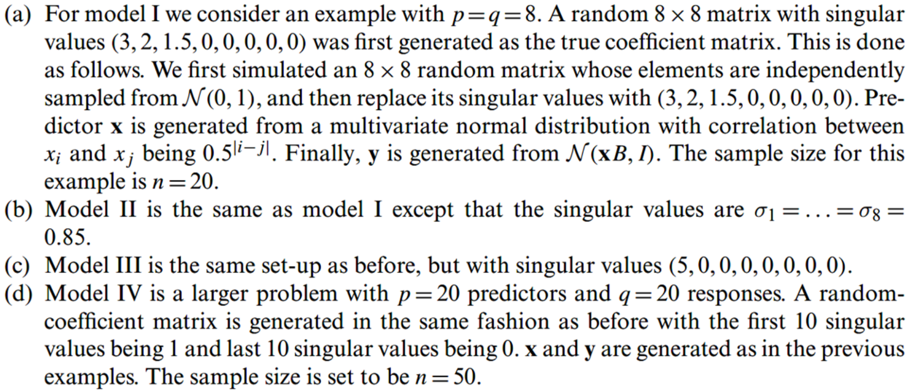
(h) CAIC, forward selection using the corrected Akaike information criterion that was proposed by Bedrick and Tsai (1994). The corrected Akaike information criterion for a specific submodel of model (1) is defined as

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

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

Simulation

Comparison is based on the model error $ME(\hat{B}) = (\hat{B} - B)'V(\hat{B} - B)$, where $V = E(X'X)$ is the population covariance.
Consider the following 4 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 $n = 20$.

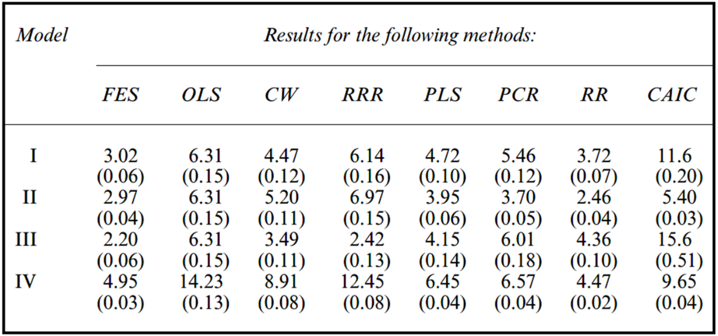
(b) Model II is the same as model I except that the singular values are $\sigma_1 = \ldots = \sigma_8 = 0.85$.

(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 $n = 50$.

Simulation

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)

Simulation

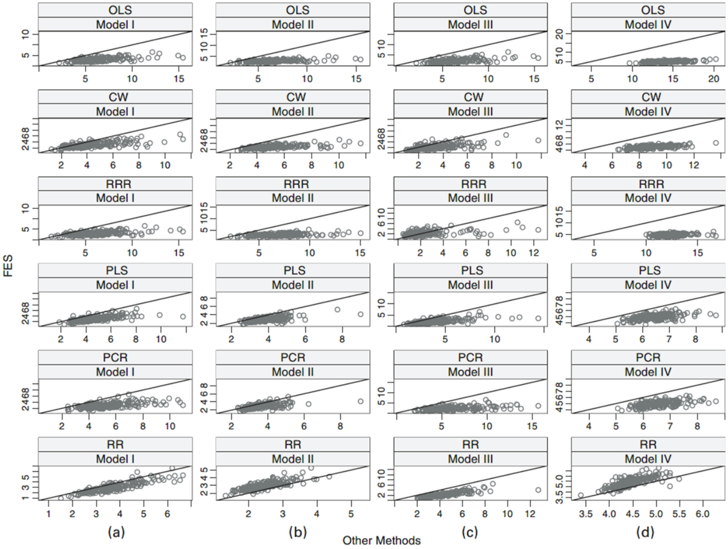


Fig. 1. Pairwise model error comparison between model FES and the other methods

Application

The financial data: let y_t be the vector of return at time t. The AR(1) model is given by $y_t = y_{t-1}B + E$

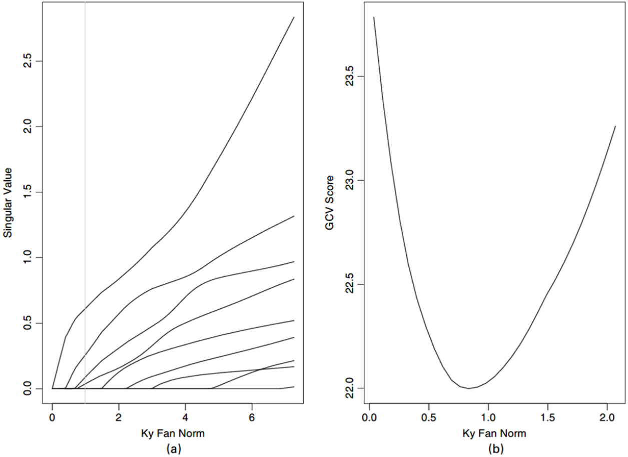


Fig. 2. Solution paths for the stocks example

Application

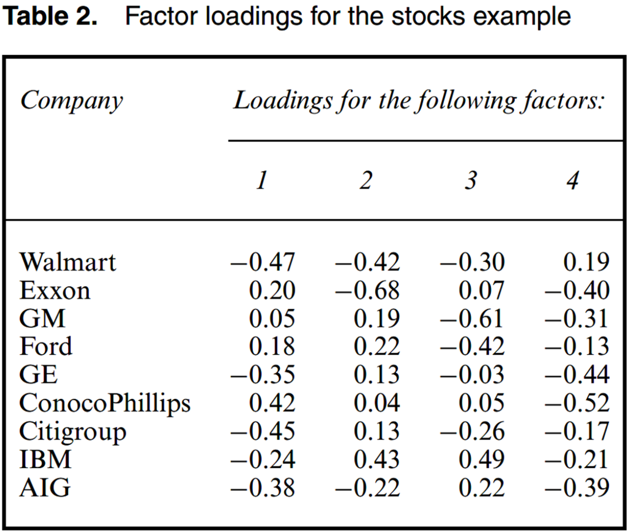


Table 2. Factor loadings for the stocks example

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

Application

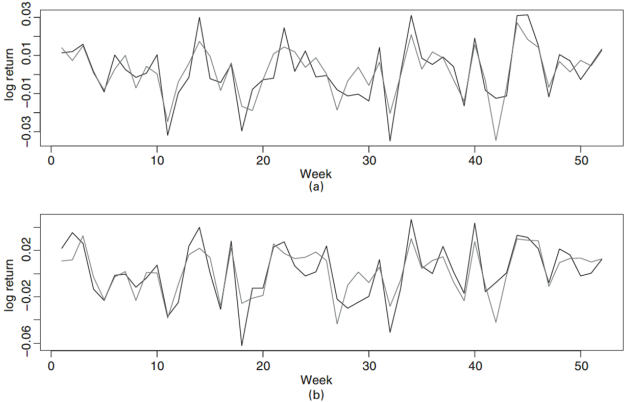


Fig. 3. (a) S&P500 and (b) NASDAQ indices (———) together with their approximations in the factor space (·······)

Application

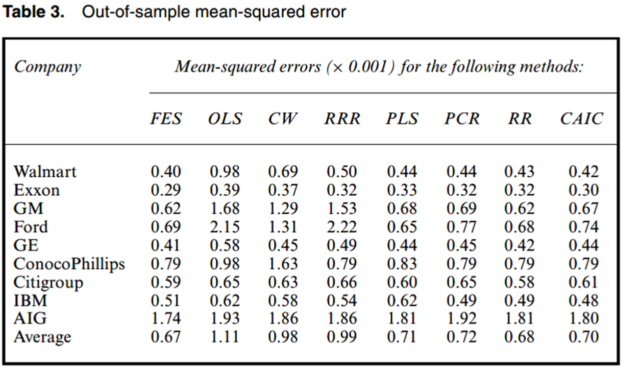


Table 3. Out-of-sample mean-squared error

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

Non-parametric actor model

By using the penalized spline regression with additive model, we can easily handle the vector non-linear (non-parametric) regression models.

This idea is implemented to reanalyse the biochemical data ($n = 33$).

- **5 responses**: pigment creatinine, concentrations of phosphate, phosphorus, creatinine and choline.
- **3 predictors**: the weight of the subject, volume and specific gravity.

Non-parametric actor model

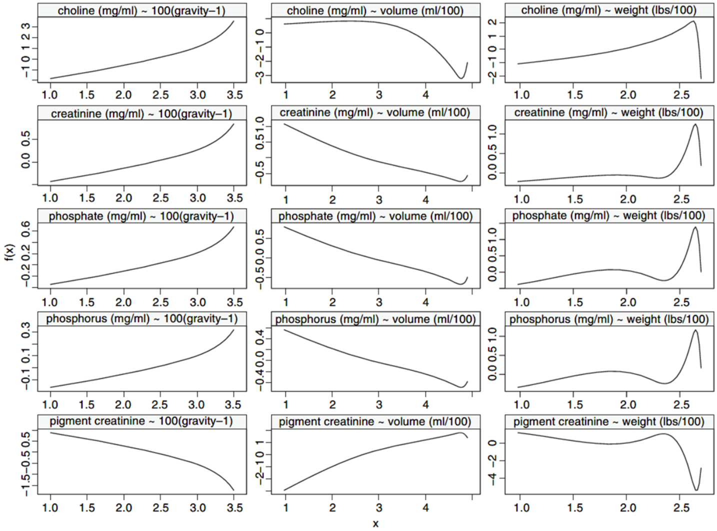


Fig. 4. Fitted components for the biochemistry data