Dimension Reduction and Variable Selection for High-dimensional Multivariate Linear Regression

Prof. N. Balakrishnan

McMaster University, Hamilton, Canada
bala@mcmaster.ca

In Collaboration with

Wenxing Guo (McMaster University, Hamilton, Canada)
Mengjie Bian (McMaster University, Hamilton, Canada)
Mu He (McMaster University, Hamilton, Canada)

Outline

Outline

Introduction

In this part, consider a multivariate linear regression has n observations with r responses and p predictors, and can be expressed as

$$\boldsymbol{Y} = \boldsymbol{X}\boldsymbol{B} + \boldsymbol{\varepsilon}, \tag{1}$$

where $\boldsymbol{Y} \in \mathbb{R}^{n \times r}$ denotes a multivariate response matrix,
$\boldsymbol{X} \in \mathbb{R}^{n \times p}$ represents a matrix of predictors,
$\boldsymbol{B} \in \mathbb{R}^{p \times r}$ is the regression coefficient matrix,
$\boldsymbol{\varepsilon} \in \mathbb{R}^{n \times r}$ is an error matrix with its entries ε_{ij}'s being independent of each other with mean zero and variance σ_{ij}^2.

Introduction

If we assume a low rank structure on the coefficient matrix $\boldsymbol{B}$ such that $\text{rank}(\boldsymbol{B}) = d < min(p, r)$, then the coefficient matrix $\boldsymbol{B}$ can be decomposed as $\boldsymbol{B} = \boldsymbol{AC}$, where $\boldsymbol{A}$ is a $p \times d$ matrix and $\boldsymbol{C}$ is an $d \times r$ matrix.
The model in (1) is called the reduced rank regression model.
By assuming this low-rank structure of the coefficient matrix, we effectively reduce the number of parameters to be estimated from pr to $(p + r)d$.
This is a quite reasonable assumption in many multivariate regression problems, which can be interpreted as follows: the r responses are related to the p predictors only through d effective linear factors.

Introduction

If some column vectors of a predictor matrix $\boldsymbol{X}$ are nearly linearly dependent, the situation known as multicollinearity, the performance of the reduced rank estimator would not be satisfactory when the predictor variables are highly correlated or the ratio of signal to noise is small.

Outline

Main results

To overcome this problem, in this work, we propose three reduced rank estimators with a nuclear norm penalty in multivariate linear regression model in terms of single random projection, averaged random projection and principal component analysis, respectively.
In our model, the number of parameters p and r can be either less than the observed value n or greater than n.
Moreover, the entry ε_{ij} in error matrix can have different variance σ_{ij}^2.

Main results

In this work, we develop a two-step method.
First, a low-rank matrix is utilized to approximate the data matrix, and we next perform reduced rank regression in terms of nuclear norm penalty.
For the first low rank matrix approximation method, we use a low-rank matrix approximation of $\boldsymbol{X}$ to be

$$\tilde{\boldsymbol{X}} = \boldsymbol{Q}\boldsymbol{Q}^T\boldsymbol{X}, \tag{2}$$

where $\boldsymbol{X}\boldsymbol{S} = \boldsymbol{Q}\boldsymbol{R}$, $\boldsymbol{S} \in \mathbb{R}^{p\times k}$ is a random matrix, $\boldsymbol{Q} \in \mathbb{R}^{n\times k}$ is a matrix with k orthonormal columns, and $\boldsymbol{R} \in \mathbb{R}^{k\times k}$ is an upper triangular matrix with positive diagonal elements.

Main results

For a data matrix $\boldsymbol{X}$, let $\boldsymbol{\Gamma}\boldsymbol{\Lambda}\boldsymbol{P}^T$ be the SVD of $\boldsymbol{X}$. Then, the top k principal components $\boldsymbol{X}_k$ can be extracted from $\boldsymbol{X}$, by setting $\boldsymbol{X}_1 = \boldsymbol{X}\boldsymbol{P}_1$, where $\boldsymbol{P}_1 \in \mathbb{R}^{p\times k}$ denotes the top k right singular vectors of $\boldsymbol{X}$.
For the second low rank matrix approximation method, we obtain an estimator of $\boldsymbol{X}$ by using the top k right singular vectors of $\boldsymbol{X}$, as

$$\hat{\boldsymbol{X}} = \boldsymbol{X}\boldsymbol{P}_1\boldsymbol{P}_1^T. \tag{3}$$

Main results

A reduced rank estimator with nuclear norm penalty, based on random projection, can then be derived by minimizing the penalized least squares criterion

$$\frac{1}{2} \parallel \boldsymbol{Y} - \tilde{\boldsymbol{X}}\boldsymbol{B} \parallel_F^2 + \mu ||\tilde{\boldsymbol{X}}\boldsymbol{B}||_*, \tag{4}$$

where μ a tuning parameter.

Proposition 1.1. *Let* $\tilde{\boldsymbol{X}}$ *equal* $\boldsymbol{Q}\boldsymbol{Q}^T\boldsymbol{X}$ *and* $\boldsymbol{P}_{\tilde{X}}\boldsymbol{Y}$ *have a singular value decomposition as* $\tilde{\boldsymbol{U}}\tilde{\boldsymbol{D}}\tilde{\boldsymbol{V}}^T$. *Then, a minimizer of (4) is*

$$\tilde{\boldsymbol{B}} = \boldsymbol{X}^+\boldsymbol{Q}\boldsymbol{Q}^T\boldsymbol{Y}\tilde{\boldsymbol{V}}\tilde{\boldsymbol{D}}^+\tilde{\boldsymbol{D}}_\mu\tilde{\boldsymbol{V}}^T, \tag{5}$$

where $\tilde{\boldsymbol{D}}_\mu = diag[\{\lambda_i(\boldsymbol{P}_{\tilde{X}}\boldsymbol{Y}) - \mu\}_+,\ i = 1, \cdots, n \wedge r]$.

Main results

Proposition 1.2. *Let* $\hat{\boldsymbol{X}} = \boldsymbol{X}\boldsymbol{P}_1\boldsymbol{P}_1^T$ *and* $\boldsymbol{P}_{\hat{\boldsymbol{X}}}\boldsymbol{Y}$ *have a singular value decomposition to be* $\hat{\boldsymbol{U}}\hat{\boldsymbol{D}}\hat{\boldsymbol{V}}^T$. *Then,*

$$\hat{\boldsymbol{B}} = \boldsymbol{P}_1\boldsymbol{P}_1^T\boldsymbol{X}^+\boldsymbol{Y}\hat{\boldsymbol{V}}\hat{\boldsymbol{D}}^+\hat{\boldsymbol{D}}_\mu\hat{\boldsymbol{V}}^T, \tag{6}$$

where $\hat{\boldsymbol{D}}_\mu = diag[\{\lambda_i(\boldsymbol{P}_{\hat{\boldsymbol{X}}}\boldsymbol{Y}) - \mu\}_+,\ i = 1, \cdots, n \wedge r]$.

Main results

The rank of the coefficient matrix $\boldsymbol{B}$, denoted by r_0, can be regarded as the number of effective linear combinations of predictor variables relating to response variables. In practice, we need to estimate the rank of $\boldsymbol{B}$. Here, we develop a method of rank estimation of $\boldsymbol{B}$ that is expressed as

$$\tilde{r} = \max\left\{ i : \ \lambda_i(\boldsymbol{P}_{\tilde{X}} \boldsymbol{Y}) > \frac{k\mu}{\eta r_{\mathrm{X}}} \right\}, \tag{7}$$

where k, r_{X} and η represent the number of columns of random projection matrix $\boldsymbol{S}$, rank of predictor matrix $\boldsymbol{X}$ and a tuning parameter, respectively. In practice, we can get the values of k, η and μ by cross-validation.

Main results

Theorem 1.3. *Suppose the entries of* $\varepsilon \in \mathbb{R}^{n\times r}$ *are independent of each other and* $\varepsilon_{ij} \sim N(0, \sigma_{ij}^2)$.
Also, for any $\theta > 0$, *let* $\mu = \eta r_X(1+\theta)\sqrt{2V(\boldsymbol{P}_{\tilde{X}}\varepsilon)\log(n+r)}/(k\delta)$ *and* $\lambda_{r_0}(\tilde{\boldsymbol{X}}\boldsymbol{B}) > \dfrac{2k\mu}{\eta r_X}$.
Then, we have

$$P(\tilde{r} \neq r_0) \longrightarrow 0 \text{ as } n + r \longrightarrow \infty.$$

Main results

Theorem 1.4.

$$E[\| \boldsymbol{XB} - \tilde{\boldsymbol{X}}\tilde{\boldsymbol{B}} \|_F] \leq \left(1 - \frac{k}{n}\right)^{1/2} \Big[\sum_{i=1}^{q} \lambda_i^2(\boldsymbol{X})\Big]^{1/2} \| \boldsymbol{B} \|_F + 2\sqrt{2r_0}\left\{\Big[\frac{k}{n}\sum_{i=1}^{n}\sum_{j=1}^{r}\sigma_{ij}^2\Big]^{1/2} + \mu\right\},$$

where k represents the number of columns of the random projection matrix $\boldsymbol{S}$.

Corollary 1.5.

$$E[\| \boldsymbol{XB} - \hat{\boldsymbol{X}}\hat{\boldsymbol{B}} \|_F] \leq 2\sqrt{2r_0}\left\{\Big[\sum_{i=1}^{n}\Big(\sum_{j=1}^{r}\sigma_{ij}^2\Big)\Big(\sum_{h=1}^{k}\gamma_{ih}^2\Big)\Big]^{1/2} + \mu\right\} + \Big[\sum_{i=k+1}^{q} \lambda_i^2(\boldsymbol{X})\Big]^{1/2} \| \boldsymbol{B} \|_F,$$

where k represents the number of principal components used.

Outline

Simulations

Existing reduced rank estimators:
ANR, RRR and RAN denote adaptive nuclear norm penalized estimator, rank penalized estimator and robustified adaptive nuclear norm penalized estimator, respectively.

Proposed reduced rank estimators:
PNR, SNR and MSN represent nuclear norm penalized estimator with principal component analysis, single random projection and averaged random projection, respectively.
MRE represents the median rank estimate and correct rank recovery percentage.

Simulations

Table 1.1 Comparisons of different methods based on 100 simulation runs with n=30, p=40, r=40, $r_0 = 10$. The median rank estimate and correct rank recovery percentage.

	ANR	RRR	RAN	PNR	SNR	MSN
				S_1		
$\rho = 0.1$	9, 22%	6, 2%	9, 22%	10, 60%	10, 57%	10, 62%
$\rho = 0.5$	8, 17%	6, 2%	8, 20%	10, 57%	10, 65%	10, 69%
$\rho = 0.9$	6, 2%	4, 0%	6, 3%	11, 33%	10, 59%	10, 63%
				S_2		
$\rho = 0.1$	2, 0%	1, 0%	2, 0%	10, 51%	10, 53%	10, 58%
$\rho = 0.5$	2, 0%	2, 0%	1, 0%	11, 5%	12, 2%	11, 6%
$\rho = 0.9$	1, 0%	1, 0%	2, 0%	4, 0%	5, 0%	5, 0%
				S_3		
$\rho = 0.1$	1, 0%	1, 0%	1, 0%	10, 100%	10, 100%	10, 100%
$\rho = 0.5$	2, 0%	1, 0%	2, 0%	8, 0%	12, 3%	11, 5%
$\rho = 0.9$	1, 0%	1, 0%	1, 0%	3, 0%	4, 0%	4, 0%

S_1, S_2 and S_3 denote Scenario 1, Scenario 2 and Scenario 3, respectively.

Simulations

Table 1.2 Comparisons of MSE ($\boldsymbol{B}$) and PM ($\boldsymbol{XB}$) with n=30, p=40, r=40.

		ANR	RRR	RAN	PNR	SNR	MSN
				S_1			
$\rho = 0.1$	MSE ($\boldsymbol{B}$)	0.413	0.415	0.413	0.397	0.397	0.395
	PM ($\boldsymbol{XB}$)	0.152	0.196	0.152	0.153	0.153	0.151
$\rho = 0.5$	MSE ($\boldsymbol{B}$)	0.159	0.161	0.159	0.143	0.143	0.141
	PM ($\boldsymbol{XB}$)	0.149	0.192	0.149	0.147	0.147	0.145
$\rho = 0.9$	MSE ($\boldsymbol{B}$)	0.052	0.059	0.051	0.043	0.040	0.036
	PM ($\boldsymbol{XB}$)	0.114	0.146	0.114	0.120	0.130	0.120
				S_2			
$\rho = 0.1$	MSE ($\boldsymbol{B}$)	0.017	0.032	0.013	0.012	0.012	0.011
	PM ($\boldsymbol{XB}$)	0.346	0.589	0.345	0.390	0.427	0.397
$\rho = 0.5$	MSE ($\boldsymbol{B}$)	0.024	0.046	0.014	0.007	0.008	0.006
	PM ($\boldsymbol{XB}$)	0.366	0.472	0.364	0.420	0.523	0.480
$\rho = 0.9$	MSE ($\boldsymbol{B}$)	0.101	0.126	0.016	0.002	0.003	0.002
	PM ($\boldsymbol{XB}$)	0.397	0.480	0.395	0.347	0.569	0.470
				S_3			
$\rho = 0.1$	MSE ($\boldsymbol{B}$)	0.024	0.148	0.017	0.015	0.016	0.014
	PM ($\boldsymbol{XB}$)	0.773	2.671	0.723	0.676	0.685	0.641
$\rho = 0.5$	MSE ($\boldsymbol{B}$)	0.043	0.140	0.016	0.011	0.013	0.010
	PM ($\boldsymbol{XB}$)	1.037	1.903	1.020	0.917	1.076	1.007
$\rho = 0.9$	MSE ($\boldsymbol{B}$)	0.298	0.336	0.016	0.003	0.005	0.003
	PM ($\boldsymbol{XB}$)	1.212	1.285	1.210	1.010	1.368	1.184

Outline

Real data

A breast cancer dataset contains 89 samples comprising gene expression measurements and comparative genomic hybridization measurements, and these data are available in the R package PMA.
Previous studies have demonstrated that certain types of cancer are characterized by abnormal DNA copy-number changes.
It will, therefore, be of interest to identify the relationship between DNA copy numbers and RNA expression levels.
In this case, we consider chromosome 18, where $p = 294$, $r = 51$ and $n = 89$. We centered and scaled both predictor matrix $\boldsymbol{X}$ and response matrix $\boldsymbol{Y}$.

Real data

For comparison of prediction accuracy, a prediction mean squared error (PMSE) is defined as

$$\text{PMSE} = \| \boldsymbol{Y}_t - \boldsymbol{X}_t \hat{\boldsymbol{B}} \|_F^2 / (n_t r), \tag{8}$$

where $(\boldsymbol{Y}_t, \boldsymbol{X}_t)$ represents the test dataset and $\hat{\boldsymbol{B}}$ represents the estimator of $\boldsymbol{B}$ corresponding to each method.
In addition, we randomly split the data into a training set of size 70 and a test set of size 19.
The training dataset is used to achieve the estimation in the model, and then the test dataset is used to evaluate the prediction performance of estimators. All the tuning parameters were selected by ten-fold cross-validation.

Real data

Table 1.3 Prediction comparisons based on data split at random 100 times.

	ANR	RRR	RAN	PNR	SNR	MSN
Rank	21	14	21	26	26	25
	(6.7)	(22.1)	(5.8)	(0.0)	(0.0)	(0.0)
PMSE	0.782	0.798	0.693	0.602	0.620	0.582
	(0.013)	(0.012)	(0.012)	(0.014)	(0.015)	(0.008)

Rank and PMSE represent the estimated rank and the prediction mean squared error, respectively. The numbers in parentheses are the corresponding standard deviations.

Real data

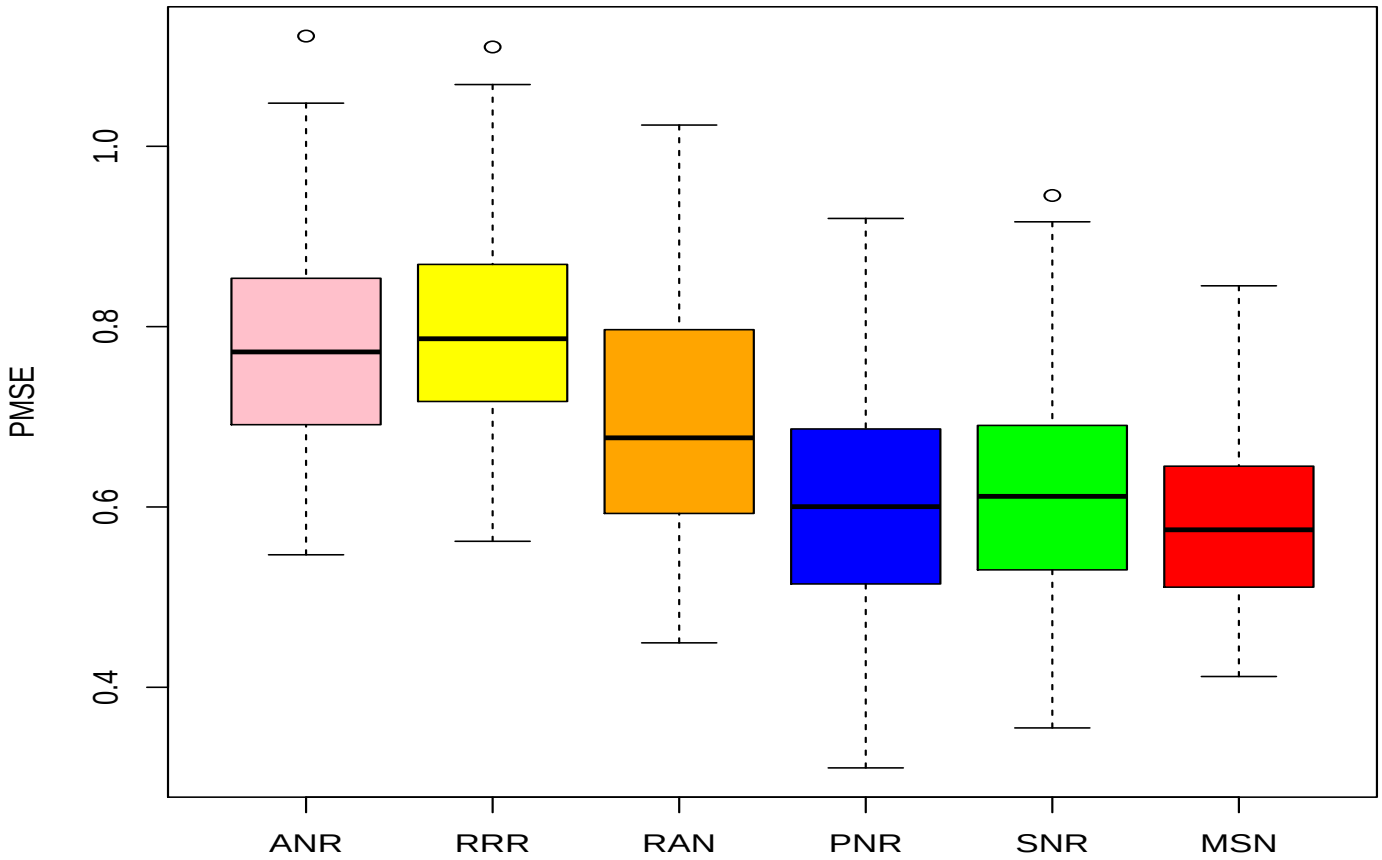

Figure 1.1: The distribution of PMSE based on data split at random 100 times.

Outline

Introduction

In this part, consider a multivariate linear regression model:

$$Y = \beta X + \varepsilon, \tag{9}$$

where $\beta \in \mathbb{R}^{r \times p}$ is the regression coefficient matrix. $X \in \mathbb{R}^p$ is a predictor vector, $Y \in \mathbb{R}^r$ is a response vector, and $\varepsilon \in \mathbb{R}^r$ is an error vector with $\text{cov}(\varepsilon)=\mathbf{\Sigma}$.
Envelope method is developed for the multivariate linear model to reduce the dimensions of response variables or predictor variables.
The key point of the envelope method is to seek the smallest reducing subspace ξ of $\mathbf{\Sigma}$ that contain span(β) such that

$$\boldsymbol{Q}_\xi Y|X \sim \boldsymbol{Q}_\xi Y, \ \text{cov}(\boldsymbol{Q}_\xi Y, \boldsymbol{P}_\xi Y|X) = 0. \tag{10}$$

Introduction

In the following, we consider the matrix form of the multivariate linear regression model:

$$\boldsymbol{Y} = \beta \boldsymbol{X} + \varepsilon, \tag{11}$$

where $\beta \in \mathbb{R}^{r \times p}$ is the regression coefficient matrix. $\boldsymbol{X} \in \mathbb{R}^{p \times n}$, $\boldsymbol{Y} \in \mathbb{R}^{r \times n}$, and $\varepsilon \in \mathbb{R}^{r \times n}$.
The envelope-based reduced rank regression in the multivariate linear model can be expressed as

$$\Gamma^T \boldsymbol{Y} = \eta \boldsymbol{B} \boldsymbol{X} + \Gamma^T \varepsilon,$$

where $\beta = \boldsymbol{A}\boldsymbol{B} = \Gamma\gamma = \Gamma\eta\boldsymbol{B}$.
Using singular value decomposition, we have $\eta\boldsymbol{B} = \boldsymbol{U}\boldsymbol{D}\boldsymbol{V}^T$, where $\boldsymbol{U} \in \mathbb{R}^{u \times d}$ and $\boldsymbol{V} \in \mathbb{R}^{p \times d}$ are rank-d matrices with orthogonal columns, respectively.

Outline

Main Results

Let $\boldsymbol{F} = \boldsymbol{D}\boldsymbol{V}^T$. Suppose $\boldsymbol{\Gamma}$ and $\boldsymbol{U}$ are given, we then minimize the following objective function:

$$\begin{aligned} Q(\boldsymbol{F}) &= \frac{1}{2n}\|\boldsymbol{U}^T\boldsymbol{\Gamma}^T\boldsymbol{Y} - \boldsymbol{F}\boldsymbol{X}\|_F^2 + \lambda_n \sum_{j=1}^{p} \omega_j(\|F_j\|_2) \\ &= \frac{1}{2n}\mathrm{tr}((\boldsymbol{U}^T\boldsymbol{\Gamma}^T\boldsymbol{Y} - \boldsymbol{F}\boldsymbol{X})^T(\boldsymbol{U}^T\boldsymbol{\Gamma}^T\boldsymbol{Y} - \boldsymbol{F}\boldsymbol{X})) + \lambda_n \sum_{j=1}^{p} \omega_j(\|F_j\|_2) \end{aligned} \tag{12}$$

where, F_j denotes the jth column of $\boldsymbol{F}$, $\beta = \boldsymbol{\Gamma}\boldsymbol{U}\boldsymbol{F}$, and $\|\cdot\|_2$ denotes the standard Euclidean norm on R_n. $\lambda_n \sum_{j=1}^{p} \omega_j(\|F_j\|_2)$ is adaptive group LASSO penalty function.

Main Results

We prove the proposed estimator has the oracle property when the sample size n increases, when r and p remain fixed. The following regularity conditions are assumed:

(A1) There exists a positive definite matrix $\boldsymbol{M}$ such that $\boldsymbol{X}\boldsymbol{X}^T/n \to \boldsymbol{M}$, as $n \to \infty$;

(A2) There exists a positive constant C_1 such that $\omega_j \leq C_1$, for all $j \in \mathcal{A}$.

Theorem 2.1. *Under regularity conditions (A1) and (A2), assume* $\hat{\boldsymbol{\Gamma}}$ *and* $\hat{\boldsymbol{U}}$ *are* $\sqrt{n}$*-consistent estimators of* $\boldsymbol{\Gamma}$ *and* $\boldsymbol{U}$*, respectively. If* $\sqrt{n}\lambda_n \to 0$ *as* $n \to \infty$*, then there exists a local minimizer* $\hat{\boldsymbol{F}}$ *of* $Q(\boldsymbol{F})$ *such that* $\hat{\boldsymbol{\beta}}$ *is a* $\sqrt{n}$*-consistent estimator of* $\boldsymbol{\beta}$*, and that this* $\hat{\boldsymbol{\beta}}$ *must satisfy*

(a) Sparsity: $P(\hat{\boldsymbol{\beta}}_{\mathcal{A}^c} = 0) \to 1$,

(b) Asymptotic normality: $\sqrt{n}(vec(\hat{\boldsymbol{\beta}}_{\mathcal{A}}) - vec(\boldsymbol{\beta}_{\mathcal{A}})) \xrightarrow{D} N(0, \boldsymbol{\Sigma}_{\boldsymbol{\beta}_{\mathcal{A}}})$, *where* $\boldsymbol{\Sigma}_{\boldsymbol{\beta}_{\mathcal{A}}}$ *is the upper-left* $pq \times pq$ *block of* $\boldsymbol{\Sigma}_{RE}$.

Main Results

We prove the consistency of the model selection when p and r are allowed to increase with the sample size n increases.

Theorem 2.2. *Under some regularity conditions, provided* $\lambda_n = o(n^{-(1-c_2+c_1)/2})$ *and* $r_n = o(n^{c_2/2})$, *the model selection of (12) is consistent; that is,* $P(\{j : \|\hat{\beta}_j\|_2 \neq 0\} = \mathcal{A}) \to 1$.

where $\mathcal{A} = \{j : \|\beta_j\|_2 \neq 0, j = 1, \ldots, p\}$

Outline

Simulations

The prediction mean squared error (PMSE) defined as

$$\text{PMSE} = E||\hat{\boldsymbol{\beta}}\boldsymbol{X} - \boldsymbol{Y}||^2/nr.$$

We compare prediction accuracy of all the methods in terms of the PMSE. We also compare accuracy of variable selection of these methods in terms of the average correct ratio (ACR) between the number of correct selection and the total number of relevant variables, which measures the ability of selecting relevant variables.

Simulations

OLS, ENV and ENRRR denote the ordinary least-squares estimator, standard envelope estimator and envelope-based reduced rank regression estimator, respectively.
Further, SPLS denotes the sparse partial least-squares estimator, SRRR and aSRRR denote sparse reduced rank regression estimator with group LASSO penalty and adaptive group LASSO penalty, respectively.
ENSRRR and aENSRRR denote envelope-based sparse reduced rank regression estimator with group LASSO penalty and adaptive group LASSO penalty, respectively.

Simulations

Table 2.1 Prediction comparisons based on 50 simulation runs with p and r being smaller than n

PMSE	n=200, $u = 10$, $s = 20$, $d = 5$			
	p=30, r=20	p=50, r=30	p=100, r=50	p=50, r=100
OLS	1.661	2.331	4.142	3.074
ENV	1.474	1.843	2.331	2.365
ENRRR	1.440	1.780	2.163	2.341
SPLS	1.841	2.148	2.245	2.476
SRRR	1.465	1.821	2.238	2.402
aSRRR	1.459	1.798	2.129	2.367
ENSRRR	1.438	1.775	2.168	2.347
aENSRRR	1.436	1.771	2.127	2.335

Simulations

Table 2.2 Variable selection comparisons based on 50 simulation runs with p and r being smaller than n

ACR	n=200, $u = 10$, $s = 20$, $d = 5$			
	p=30, r=20	p=50, r=30	p=100, r=50	p=50, r=100
SPLS	0.79	0.65	0.94	0.82
SRRR	0.67	0.60	0.20	0.40
aSRRR	1.00	0.99	0.96	0.98
ENSRRR	0.69	0.65	0.30	0.55
aENSRRR	1.00	0.99	0.97	0.99

Simulations

Table 2.3 Prediction comparisons based on 50 simulation runs with p and r being greater than n

PMSE	$n = 60, p = 70, r = 70, d = 5$		$n = 100, p = 150, r = 150, d = 5$	
	$u = 10, s = 20$	$u = 20, s = 40$	$u = 10, s = 20$	$u = 20, s = 40$
SPLS	3.324	5.435	2.472	3.090
SRRR	2.687	4.915	2.525	3.173
aSRRR	2.658	4.055	2.506	3.044
ENSRRR	2.524	4.764	2.392	2.950
aENSRRR	2.446	3.882	2.381	2.930

Simulations

Table 2.4 Variable selection comparisons based on 50 simulation runs with p and r being greater than n

ACR	$n = 60, p = 70, r = 70$		$n = 100, p = 150, r = 150$	
	$u = 10, s = 20$	$u = 20, s = 40$	$u = 10, s = 20$	$u = 20, s = 40$
SPLS	0.83	0.68	0.96	0.76
SRRR	0.77	0.76	0.93	0.68
aSRRR	0.97	0.97	0.97	0.86
ENSRRR	0.82	0.80	0.95	0.70
aENSRRR	0.98	0.98	0.97	0.88

Outline

Real data

A yeast cell cycle data set can be obtained from the R package spls.
The response matrix $\boldsymbol{Y}$ consists of 542 cell-cycle-regulated genes.
The cell cycle was measured by taking RNA levels on genes at 18 time points using the α-factor arrest method.
The 542×106 predictor matrix $\boldsymbol{X}$ contains the binding information of the target genes for a total of 106 transition factors (TFs).

Real data

In this case, we have $n = 360$, $r = 18$ and $p = 106$ in the training dataset.
We centered and scaled both predictor matrix $\boldsymbol{X}$ and response matrix $\boldsymbol{Y}$.
By using five-fold cross-validation, we selected the number of factors $d = 4$ for SRRR, aSRRR, ENSRRR and aENSRRR.
Similarly, the dimension of the envelope, u, was selected to be 6 by five-fold CV.

Real data

Table 2.5 Prediction comparisons based on data split at random 100 times.

	OLS	ENV	ENRRR	SPLS	SRRR	aSRRR	ENSRRR	aENSRRR
PMSE	0.534	0.512	0.490	0.416	0.466	0.415	0.401	0.399

Table 2.6 Variable selection comparisons based on data split at random 100 times.

	SPLS	SRRR	aSRRR	ENSRRR	aENSRRR
MNSP	30	77	64	76	64
RNSP	[19, 53]	[45, 89]	[46, 80]	[47, 88]	[48, 81]

MNSP and RNSP denote median and range of the numbers of selected predictors in the 100 splits, respectively.

References

This presentation is based on the following two papers:

1. Guo, W., Balakrishnan, N. and Bian, M., Reduced rank regression with matrix projections for high-dimensional multivariate linear regression model. Under Review.
2. Guo, W., Balakrishnan, N. and He, M., Envelope-based sparse reduced rank regression for high-dimensional multivariate linear regression model. Under Review.

Thanks for your attention!