CS 229 Final Project
Reduced Rank Regression

Name : Ka Wai Tsang
SID : 005589301

1. Introduction

Given m observations of the predictors $X_i \in \mathbb{R}^p$ and the corresponding responses $Y_i \in \mathbb{R}^n$, let $Y = [Y_1, Y_2, \ldots, Y_m]^T$ and $X = [X_1, X_2, \ldots, X_m]^T$. Suppose X and Y are related by

$$Y = XA + E \tag{1.1}$$

where A is an unknown $p \times n$ matrix of coefficients and E is an unobserved $m \times n$ random noise matrix with independent mean zero and variance σ^2. We want to find an estimate $\hat{A}$ such that $||Y - X\hat{A}||$ is small. If we use standard least square estimation directly to estimate A in (1.1) without adding any constraints, then it is just the same as regressing each response on the predictors separately. In this way, we actually ignore the possibility that the responses may be correlated among themselves. Besides, when there are many attributes (p is large) and many different kinds of responses (n is large), the number of unknowns can be larger than the sample size m. We may then need much more effort to collect more samples to increase m or the least square method simply cannot be applied.
To address this problem, one popular way to handle it is reduced rank regression. Let $r(M)$ be the rank of a matrix M. If we expect $r(A) = r < \min(p, n)$ or A can be well approximated by a low rank matrix, we can write A as a product of two matrices with rank r, see [1]. That is $A = B_r C_r$, $B_r \in \mathbb{R}^{p \times r}$ and $C_r \in \mathbb{R}^{r \times n}$ which have total $r \times (n + p)$ unknowns needed to be estimated. It can be much less than m if $r(A)$ is very small. The model (1.1) then become

$$Y = (XB_r)C_r + E$$

It can be interpreted as instead of p attributes, Y actually only depends on r factors. Each factor is a linear combination of the attributes. In another words, this model says that the attributes are correlated which is often the case in many real situations.
It is quite often that too many attributes are considered when we build a model. If we believe that some attributes are actually not important to determine a response or we are only interested to those really important attributes, then we would like A to be row sparse. Let $J(M)$ be the index set of the non-zero rows of a matrix M and $|J(M)|$ is the corresponding cardinality. We would like to have $|J(A)|$ (or $|J(\hat{A})|$) small so that we can only consider those attributes in $J(A)$ in future prediction. A common way to make the estimate $\hat{A}$ sparse is adding appropriate norm penalty, such as zero-norm or 1-norm.
In this project, I am going to introduce three reduced rank regression methods, see [2], which can give an estimate $\hat{A}$ with $r(\hat{A})$ and $|J(\hat{A})|$ small.

2. Method 1

Step 1. We first find an estimate $\hat{A}_1$ such that

$$\hat{A}_1 = \arg \min_{B \in \mathbb{R}^{p \times n}} \{||Y - XB||_F^2 + \mu r(B)\} \tag{2.1}$$

with $\mu = 2\sigma^2(\sqrt{n} + \sqrt{p})^2$, see [3]. In this step, we add a penalty term $\mu r(B)$, so we would expect $\hat{A}_1$ is a low rank matrix. Let $\hat{k} = r(\hat{A}_1)$.

Step 2. Use the $\hat{k}$ computed before, find $\hat{A}$ such that

$$\hat{A} = \arg \min_{B \in \mathbb{R}^{p \times n}, r(B) \leq k} \{\frac{1}{2}||Y - XB||_F^2 + \lambda ||B||_{2,1}\} \tag{2.2}$$

where $||B||_{2,1} = \sum_{i=1}^{p} ||b_i^T||_2$, b_i^T is the i-th row of B. Here $k = \hat{k}$, λ is a regularization parameter that can be estimated by cross validation method. Notice that if the constraint $r(B) \leq k$ is removed, then $\hat{A}$ is a group Lasso estimator. Lasso method will output sparsity in the estimation.

3. Method 2

Step 1. For each pair $(k, \lambda), for 1 \leq k \leq p$ and a range of values λ, we find an estimate $\hat{A}_{k,\lambda}$ such that it is the minimizer of (2.2).

Step 2. Among the $\hat{A}_{k,\lambda}$ computed above, choose the one $\hat{A}$ such that

$$\hat{A} = \arg \min_{\hat{A}_{k,\lambda}} \{||Y - X\hat{A}_{k,\lambda}||_F^2 + 2\sigma^2(2n + |J(\hat{A}_{k,\lambda}|)|r(\hat{A}_{k,\lambda})|)\} \tag{3.1}$$

This method is called selective reduced rank regression introduced in [2]. We observe that the penalty term here penalizes both $r(\hat{A}_{k,\lambda})$ and $|J(\hat{A}_{k,\lambda})|$. It is an penalty designed for selecting estimators with the best bias-variance trade-off relative to a list of possible candidates.

4. Method 3

Step 1. We first find an estimate $\hat{A}_1$ such that

$$\hat{A}_1 = \arg \min_{B \in \mathbb{R}^{p \times n}} \{\frac{1}{2}||Y - XB||_F^2 + \lambda ||B||_{2,1}\} \tag{4.1}$$

The result is an group lasso estimator and so will be a sparse matrix. Let $\hat{X}$ be the predictor matrix that only contains the attribute columns selected here, that is the columns that the corresponding rows in $\hat{A}_1$ are non-zero.

Step 2. Find $\hat{A}$ such that

$$\hat{A} = \arg \min_{B \in \mathbb{R}^{q \times n}} \{||Y - \hat{X}B||_F^2 + \mu r(B)\} \tag{4.2}$$

with $q = |J(\hat{A}_1)|$,$\mu = 2\sigma^2(\sqrt{n} + \sqrt{q})^2$ as before.

5. Simulation

We first make up some examples to see if these methods work. The setting I use here is the similar as that in [2] so that I can compare the results. We construct the matrix of dependent variables X with rows i.i.d. from a multivariate normal distribution MVN(0,Σ), with $\Sigma_{jk} = \rho^{|j-k|}$, $\rho > 0$, $1 \leq j, k \leq p$. Set the coefficient matrix

$$A = \begin{bmatrix} bB_0B_1 \\ 0 \end{bmatrix}$$

where $b > 0$, B_0 is a $J \times r$ matrix and B_1 is a $r \times n$ matrix. All entries in B_0 and B_1 are i.i.d. $N(0, 1)$. Therefore, almost for sure $J = J(A)$ and $r = r(A)$. Generate $E \in \mathbb{R}^{m \times n}$ such that $E_{ij} \sim N(0, 1)$. Then set $Y = XA + E$.

To evaluate how good an approximation $\hat{A}$ is, for the same setting (m, $|J|$, p, n, r, rho and b) we repeat the methods 50 times and then we look at: 1. mean square error (MSE) $||XA - X\hat{A}||_F^2/(mn)$ using test data at each run; 2. the mean number of predictors ($|\hat{J}|$) and mean rank estimate ($\hat{R}$); and maybe the most important 3. how many correct(or wrong) predictors were selected in comparison to the correct coefficient matrix A. It is measured by the missing rate ($M = |J - \hat{J}|/|J|$) and false chosen ($FC = |\hat{J} - J|/(p - |J|)$) rate. A good approximation should have low M and FA.
For the setting ($m = 30$, $|J| = 15$, $p = 100$, $n10$, $r = 2$, $\rho = 0.1$ and $b = 2$), the simulation results showed that the performances of these methods are close to each other. It is not surprising because as the main components of these methods actually are similar. All these methods successfully estimated the rank of the correct coefficient matrix and selected most (around 70%) of actual related features

	MSE	$\hat{J}$	$\hat{R}$	M	FC
Method 1	42.0	22.2	2	0.30	0.14
Method 2	38.7	21.9	2	0.27	0.13
Method 3	44.3	22.6	2	0.29	0.14

6. Vector Autoregressive Model

To apply these methods in real applications, I have collected the interest rate swaps for 1-, 2-, 3-, 4-, 5-, 7-, 10-, 30-year from Jan 2009 to Sep 2011 [4]. My problem here is to estimate the future values of the swaps using the past values. Consider a vector autoregressive (VAR) model

$$\begin{aligned} y_t &= \alpha_0 + y_{t-1}A_1 + \cdots + y_{t-d}A_d + \epsilon_t \\ &= (1\ y_{t-1}\ y_{t-2} \dots y_{t-d}) \begin{pmatrix} \alpha_0 \\ A_1 \\ A_2 \\ \vdots \\ A_d \end{pmatrix} \end{aligned}$$

where $y_t \in \mathbb{R}^n$ contains n swaps' value at day t, $A_i \in \mathbb{R}^{n\times n}$ are coefficient matrices and $\epsilon_t \in \mathbb{R}^n$ are i.i.d. with mean 0 and covariance matrix Σ. Suppose we use $m + d$ days to approximate A_i, then we have $Y = XA + E$. $Y \in \mathbb{R}^{m\times n}$, $X \in \mathbb{R}^{m\times p}$ and $A \in \mathbb{R}^{p\times n}$ where $p = 1 + d \times n$. The i-th row of Y is y_{t-i} and the corresponding row of X is $(1\ y_{t-i-1} \dots y_{t-i-d})$. Notice that we have the term σ^2 in the above methods. For the case the covariance matrix $\Sigma = \sigma I$, [3] suggests an unbiased estimator

$$S^2 = ||Y - PY||_F^2/(mn - pn) \tag{6.1}$$

where P is the orthogonal projection matrix on the column space of X.
In order to satisfy the assumptions of the VAR model, I first differentiate the times series in the time dimension. By augmented Dickey-Fuller test, no unit-root is present in each time series with 95% confidence level.
However, the performance of the methods in this application is not good. It may be because :

(1) The above methods tend to generate an estimate with more zero rows which means the corresponding columns should not be selected in the model. In autoregressive

model, people expect the values today depend more on the values right before today than the days further before. Before the experiment, I thought most of the zero rows would appear at the bottom of the estimate matrix $\hat{A}$ which are corresponding to the dates farthest away from the current day t so that the methods can tell me how many days (d) should be used to predict the future values. However, the result showed that the number of zero rows at the top of $\hat{A}$ is roughly the same as that at the bottom. I think the problem is that if we want to find out how many days should be used in the VAR model, then all the swaps' value at the same day should be consider as one group, either all the swaps in this day should be chosen or none of them.

(2) Obviously the values of the interest rate swaps with different maturities are correlated. The assumption that the variance matrix $E = \sigma I$ may be too strong for this application and so (6.1) is not a good estimate.

To make the methods gave me something make sense, I only considered $AR(1)$ model, that is $d = 1$. Therefore, selecting variables is now nothing about selecting approximate number of regression days in the model, but selecting the interest rate swaps that contain most of the information about the swaps' values tomorrow.

The procedure of this experiment is as follow:

(1) Divide the data set into one training set (Jan 2009 to Dec 2010) and one test set (Jan 2011 to Sep 2011).
(2) Use the training set to compute $\hat{A}$ by the above methods.
(3) For t in test set, compute prediction $\hat{y}_t = y_{t-1}\hat{A}$ and error term $|y_t^{(i)} - \hat{y}_t^{(i)}|$, where $y_t = (y_t^{(1)}, y_t^{(2)}, \ldots, y_t^{(n)})$ is a vector of actual differentiated interest rate swaps at day t.

The table below shows the results of those three methods. "mean$(|\hat{y}_t|)_i$" is the mean of the estimated differentiated interest rates swaps using method i. From the table, we can see that the results are very poor. I have modified the methods such as fixing the matrix to be full rank and increasing d to see if these help improve the results. It helps a little bit but still cannot give reasonable estimates.

	1-yr	2-yr	3-yr	4-yr	5-yr	7-yr	10-yr	30-yr
mean$(\lvert y_t \rvert)$	0.009	0.020	0.032	0.053	0.060	0.063	0.068	0.066
mean$(\lvert \hat{y}_t \rvert)_1$	0.001	0.004	0.006	0.009	0.009	0.008	0.007	0.006
mean$(\lvert \hat{y}_t \rvert)_2$	0.001	0.004	0.006	0.009	0.009	0.009	0.007	0.007
mean$(\lvert \hat{y}_t \rvert)_3$	0.001	0.006	0.008	0.011	0.012	0.010	0.008	0.007

The reason may be the methods are based on the assumption that the actual coefficient matrix is of low rank and sparse. Clearly, as swaps are correlated, the matrix should be of low rank. However the matrix probably not sparse in this case.

7. Future works

To improve the methods so that they can be applied on more application, I think there are at least 2 things we can do. First, we can introduce grouping into the algorithm. Sometimes

the features are naturally grouped, like in the VAR model the features are group by dates. If our question is which group should be chosen in the regression model, then treating each feature individually should not be a good approach. To achieve this, instead of adding penalty $||B||_{2,1} = \sum_{i=1}^{p} ||b_i^T||_2$, maybe we can use $\sum_I ||B_I^T||_F$, where B_I^T is a sub-matrix of B formed by the row vector in a group with indices in I.

Besides, we may also try to improve the estimate (6.1) or use another variance estimate when the assumption the variance matrix $E = \sigma I$ is not likely to be true.

References

[1] G. C. Reinsel and R. P. Velu, "Multivariate Reduced-Rank Regression: Theory and Applications", Lecture Notes in Statistics 136, Springer, New York, 1998.

[2] F. Bunea, Y. She and M. Wegkamp, "Joint variable and rank selection for parsimonious estimation of high dimensional matrices", 2011.

[3] F. Bunea, Y. She and M. Wegkamp, "Optimal selection of reduced rank estimators of high-dimensional matrices", Annals of Statistics, 39, 1282-1309, 2011.

[4] http://www.gsb.stanford.edu/jacksonlibrary/