中国科学技术大学
University of Science and Technology of China

Review of sparse reduced-rank regression model

报告人：尚尔淦

数学科学学院
概率统计系

2021 年 10 月 29 日

大纲

A graph G=(V,E), V: set of vertices,E: set of edges;XY is adjacent if there's an edge joining them:$X \sim Y$
Suppose V represents a set of random variables having jointly distribution P.
The graph gives a visual way of understanding the joint distribution of the entire set of random variables. The absence of an edge between 2 vertices has a special meaning: **the corresponding random variables are conditionally independent given other variables**.
We assume that the observation has a multivariate Gaussian distribution with mean μ and covariance matrix Σ. Then all conditional distributions are also Gaussian. Σ^{-1} contains partial covariance:**covariance between i and j conditioned on all other variables**

Our major work is to determine which features of the system are conditionally independent i.e. **Estimating the inverse covariance matrix** Σ^{-1}

We can put forward a regularization parameter λ that controls the sparsity of the graph; A new approach to model selection based on **model stability**.

We start with a large regularization which corresponds to an empty and hence highly stable graph. We gradually reduce the amount of regularization until there's a small but acceptable amount of variability of the graph across subsamples.

Estimating a undirected graph

$X = (X(x), \cdots, X(p))'$ is the random vector with distribution P, G=(V,E) with vertices $V = \{X(1), \cdots, X(p)\}$. We use E to denote the adjacency matrix and edges.

Our aim is to infer E from i.i.d observed data $X_1, \cdots, X_n$ where $X_i = (X_i(1), \cdots, X_i(p))'$

Suppose P is Gaussian with mean vector μ and covariance matrix Σ, let $\Omega = \Sigma^{-1}$

$X = (X(x), \cdots, X(p))'$ is the random vector with distribution P, G=(V,E) with vertices $V = \{X(1), \cdots, X(p)\}$. We use E to denote the adjacency matrix and edges.

Our aim is to infer E from i.i.d observed data $X_1, \cdots, X_n$ where $X_i = (X_i(1), \cdots, X_i(p))'$

Suppose P is Gaussian with mean vector μ and covariance matrix Σ, let $\Omega = \Sigma^{-1}$

$$\Omega_{jk} = 0 \iff \textit{no} \quad \textit{edge} \quad \textit{corresponding} \quad \textit{to} \quad X(j) \quad \textit{and} \quad X(k)$$

$X = (X(x), \cdots, X(p))'$ is the random vector with distribution P, G=(V,E) with vertices $V = \{X(1), \cdots, X(p)\}$. We use E to denote the adjacency matrix and edges.
Our aim is to infer E from i.i.d observed data $X_1, \cdots, X_n$ where $X_i = (X_i(1), \cdots, X_i(p))'$
Suppose P is Gaussian with mean vector μ and covariance matrix Σ, let $\Omega = \Sigma^{-1}$

$$\Omega_{jk} = 0 \iff no \quad edge \quad corresponding \quad to \quad X(j) \quad and \quad X(k)$$

We come to estimate the sparsity pattern of Ω,if ignoring the constants, the log-likelihood:

$$\ell(\Omega) = log|\Omega| - trace(\hat{\Sigma}\Omega)$$

$\hat{\Sigma}$:the sample covariance matrix

$$\hat{\Sigma} = \frac{1}{N}\sum_{i=1}^{N}(x_i - \bar{x})(x_i - \bar{x})'$$

$\hat{\Omega}(\lambda) = arg\min_{\Omega>0}(-\ell(\Omega) + \lambda||\Omega||_1)$ where $||\Omega||_1 = \sum_{j,k}|\Omega_{jk}|$ is the ℓ_1 norm

Therefore we get the estimator

$$\hat{G}(\lambda) = (V, \hat{E}(\lambda))$$

Obviously, larger values of λ tend to yield sparse graphs, we define $\Lambda = \frac{1}{\lambda}$ so that small Λ corresponds to sparse graph. We get **Grid of regularization parameters** $G_n = \{\lambda_1 \cdots, \Lambda_K\}$, to choose one $\hat{\Lambda} \in G_n$ such that the true graph E is contained in $\hat{E}(\hat{\lambda})$ with high probability.

Let $d(\lambda)$ denote the degree of freedom of the corresponding Gaussian model, we have these existing criterion

$$AIC: \quad \hat{\Lambda} = arg \min_{\lambda \in G_n}(-2\ell(\Omega(\Lambda)) + 2d(\Lambda))$$

$$BIC: \quad \hat{\Lambda} = arg \min_{\lambda \in G_n}(-2\ell(\Omega(\Lambda)) + d(\Lambda) \cdot logn)$$

A common practice is to calculate $d(\Lambda) = m(\Lambda)(m(\Lambda) - 1)/2 + p$,where $m(\Lambda)$ denotes the number of nonzero elements of $\hat{\Omega}(\Lambda)$

When $\Lambda = 0$, the graph is empty. As we increase Λ, the variability of graph increases and the stability decreases. Now we put forward the concrete rule for choosing Λ

When $\Lambda = 0$, the graph is empty. As we increase Λ, the variability of graph increases and the stability decreases. Now we put forward the concrete rule for choosing Λ

Let $b = b(n)$ be the parameter such that 1<b(n)<n, we draw N random subsamples $S_1, \cdots, S_N$ from $X_1, \cdots, X_n$ each of size b. We choose a large number N of subsamples at random.**Each subsample is drawn without replacement**

When $\Lambda = 0$, the graph is empty. As we increase Λ, the variability of graph increases and the stability decreases. Now we put forward the concrete rule for choosing Λ

Let $b = b(n)$ be the parameter such that 1<b(n)<n, we draw N random subsamples $S_1, \cdots, S_N$ from $X_1, \cdots, X_n$ each of size b. We choose a large number N of subsamples at random.**Each subsample is drawn without replacement**

For each $\Lambda \in G_n$, we construct a graph using glasso for each subsample. This results in N estimated edge matrices $\hat{E}_1^b(\Lambda), \cdots, \hat{E}_N^b(\Lambda)$

When $\Lambda = 0$, the graph is empty. As we increase Λ, the variability of graph increases and the stability decreases. Now we put forward the concrete rule for choosing Λ

Let $b = b(n)$ be the parameter such that 1<b(n)<n, we draw N random subsamples $S_1, \cdots, S_N$ from $X_1, \cdots, X_n$ each of size b. We choose a large number N of subsamples at random.**Each subsample is drawn without replacement**

For each $\Lambda \in G_n$, we construct a graph using glasso for each subsample. This results in N estimated edge matrices $\hat{E}_1^b(\Lambda), \cdots, \hat{E}_N^b(\Lambda)$

Now we focus on one edge (s,t) and one value of Λ. Let $\psi^\lambda(\cdot)$ denotes the glasso algorithm with Λ

For any subsample S_j, let $\psi_{st}^{\Lambda}(S_j) = 1$ if the algorithm puts an edge between s and t, otherwise $\psi_{st}^{\Lambda}(S_j) = 0$

For any subsample S_j, let $\psi_{st}^{\Lambda}(S_j) = 1$ if the algorithm puts an edge between s and t, otherwise $\psi_{st}^{\Lambda}(S_j) = 0$

Define

$$\theta_{st}^{b}(\Lambda) = P(\psi_{st}^{\Lambda}(X_1, \cdots, X_b) = 1)$$

For any subsample S_j, let $\psi_{st}^{\Lambda}(S_j) = 1$ if the algorithm puts an edge between s and t, otherwise $\psi_{st}^{\Lambda}(S_j) = 0$

Define

$$\theta_{st}^{b}(\Lambda) = P(\psi_{st}^{\Lambda}(X_1, \cdots, X_b) = 1)$$

To estimate $\theta_{st}^{b}(\Lambda)$, define $\hat{\theta}_{st}^{b}(\Lambda) = \frac{1}{N}\sum_{j=1}^{N} \psi_{st}^{\Lambda}(S_j)$

For any subsample S_j, let $\psi_{st}^{\Lambda}(S_j) = 1$ if the algorithm puts an edge between s and t, otherwise $\psi_{st}^{\Lambda}(S_j) = 0$

Define

$$\theta_{st}^{b}(\Lambda) = P(\psi_{st}^{\Lambda}(X_1, \cdots, X_b) = 1)$$

To estimate $\theta_{st}^{b}(\Lambda)$, define $\hat{\theta}_{st}^{b}(\Lambda) = \frac{1}{N}\sum_{j=1}^{N} \psi_{st}^{\Lambda}(S_j)$

Now define the parameter

$$\xi_{st}^{b}(\Lambda) = 2\theta_{st}^{b}(\Lambda)(1 - \theta_{st}^{b}(\Lambda))$$

For any subsample S_j, let $\psi_{st}^{\Lambda}(S_j) = 1$ if the algorithm puts an edge between s and t, otherwise $\psi_{st}^{\Lambda}(S_j) = 0$
Define

$$\theta_{st}^{b}(\Lambda) = P(\psi_{st}^{\Lambda}(X_1, \cdots, X_b) = 1)$$

To estimate $\theta_{st}^{b}(\Lambda)$, define $\hat{\theta}_{st}^{b}(\Lambda) = \frac{1}{N}\sum_{j=1}^{N} \psi_{st}^{\Lambda}(S_j)$
Now define the parameter

$$\xi_{st}^{b}(\Lambda) = 2\theta_{st}^{b}(\Lambda)(1 - \theta_{st}^{b}(\Lambda))$$

let $\hat{\xi}_{st}^{b}(\Lambda) = 2\hat{\theta}_{st}^{b}(\Lambda)(1 - \hat{\theta}_{st}^{b}(\Lambda))$ be its estimate.
We can regard $\xi_{st}^{b}(\Lambda)$ as being twice the variance of the Bernoulli indicator of the edge (s,t) and **as a measure of instability of the edge across subsample with** $0 \leq \xi_{st}^{b}(\Lambda) \leq \frac{1}{2}$

Define the total instability by averaging over all edges

$$\hat{D}_b(\Lambda) = \frac{\sum_{s<t} \hat{\xi}_{st}^b}{C_p^2}$$

Define the total instability by averaging over all edges

$$\hat{D}_b(\Lambda) = \frac{\sum_{s<t} \hat{\xi}_{st}^{b}}{C_p^2}$$

Then define

$$\bar{D}_b(\Lambda) = \sup_{0 \le t \le \Lambda} \hat{D}_b(t) \quad \hat{\Lambda}_s = sup\{\Lambda : \bar{D}_b(\Lambda) \le \beta\}$$

β is the threshold

1 Stability approach to regularization selection

2 sparse reduced-rank regression for simultaneous
- introduction and penalty
- numerical solution and tuning
- asymptotic analysis

3 sparse RRR for simultaneous rank and variable

4 joint variable and rank selection for parsimonious

5 Dimension reduction and coefficient estimation

6 Reduced rank stochastic regression

7 wavelet-based sparse reduce-rank regression

The dimension reduction aspect of multivariate regression is taken care of by reduced-rank regression(RRR)
The variable selction aspect is addressed by adding a penalty to the least squares fitting criterion to enforce the sparsity of reduced-rank coefficient matrix.

The dimension reduction aspect of multivariate regression is taken care of by reduced-rank regression(RRR)

The variable selction aspect is addressed by adding a penalty to the least squares fitting criterion to enforce the sparsity of reduced-rank coefficient matrix.

The model is

$$Y = XC + E$$

with n observations. Taking advantage of possible interrelationships between response variables is to impose a constraint on the rank of C:$rank(C) = r \leq min(p, q)$

Then $C = BA'$, where $B : p \times r, A : q \times r$, and

$$Y = (XB)A' + E$$

XB is of reduced dimension with only r components, which can be interpreted as unobservable latent factors. By solving the optimization problem

$$\min_{C:rank(C)=r} ||Y - XC||^2$$

Denote $S_{xx} = \frac{1}{n}X'X, S_{xy} = \frac{1}{n}X'Y, S_{yx} = \frac{1}{n}Y'X$, we have the solution

$$\hat{A}^{(r)} = V \quad \hat{B}^{(r)} = S_{xx}^{-1}S_{xy}V$$

where $V = (v_1, \cdots, v_r)$ and v_j is the eigen vector of $S_{yx}S_{xx}^{-1}S_{xy}$
The solution satisfies that: $A'A = I_r, B'S_{xx}B$ being diagnonal.

SRRR through penalized least squares

Exclude the redundant predictors when some predictor variables are not useful for prediction $\Longleftrightarrow$ set as zero an entire row of B.

Exclude the redundant predictors when some predictor variables are not useful for prediction $\Longleftrightarrow$ set as zero an entire row of B. We Consider the following optimization problem

$$\min_{A,B} ||Y - XBA'||^2 + \sum_{i=1}^{p} \lambda_i ||B^i|| \quad such \quad that \quad A'A = I$$

where B^i denotes the ith row of B. This is a penalized regression with a grouped lasso penalty.

Exclude the redundant predictors when some predictor variables are not useful for prediction $\Longleftrightarrow$ set as zero an entire row of B. We Consider the following optimization problem

$$\min_{A,B} ||Y - XBA'||^2 + \sum_{i=1}^{p} \lambda_i ||B^i|| \quad such \quad that \quad A'A = I$$

where B^i denotes the ith row of B. This is a penalized regression with a grouped lasso penalty.

Lemma 2.1

The solution to the optimization problem is unique up to an orthogonal matrix.

Proof 2.1

Let $(\hat{A}, \hat{B})$ *is a solution and* Q *is an orthogonal matrix. Let* $\tilde{A} = \hat{A}Q, \tilde{B} = \hat{B}Q$*, then* $\tilde{B}\tilde{A}' = \hat{B}\hat{A}'$ *and* $||\tilde{B}^i|| = ||\hat{B}^i||$ *via* $QQ' = I$*. As a result,* $(\tilde{A}, \tilde{B})$ *is also a solution*

Moreover, if $(\tilde{A}, \tilde{B})$ *and* $(\hat{A}, \hat{B})$ *are both the solution, considering* $rank(\hat{A}) = rank(\tilde{A}) = r$*, then there's a non-singular matrix* Q *of* $r \times r$ *such that* $\tilde{A} = \hat{A}Q$*, we reach the conclusion that* Q *is orthogonal because* $I_r = \tilde{A}'\tilde{A} = Q'\hat{A}'\hat{A}Q = Q'Q$

Finally, $||\tilde{B}^i|| = ||\tilde{B}^i Q'||$ *and* $||Y - XBA'|| = ||Y - XBQ'(AQ')'||$*, we know that* $(\hat{A}, \tilde{B})$ *is the solution. Then* $\hat{B} = \tilde{B}Q'$

Definition 2.1

If the entire row j of B is zero, then the predictor variable X_j is called a nonactive variable, otherwise it is called an active variable.

Definition 2.1

If the entire row j of B is zero, then the predictor variable X_j is called a nonactive variable, otherwise it is called an active variable.

Lemma 2.2

The set of active variables obtained by the optimization problem is uniquely determined.

For fixed B, the optimization problem is equivalent with

$$\min_{A} ||Y - XBA'|| \quad such \quad that \quad A'A = I$$

which is an orthogobal Procrustes Problem(Gower&Dijksterhuis 2004), and the solution is $\hat{A} = UV'$, where U, V is the SVD of $Y'XB$,i.e.$Y'XB = UDV'$

For fixed B, the optimization problem is equivalent with

$$\min_{A} ||Y - XBA'|| \quad such \quad that \quad A'A = I$$

which is an orthogobal Procrustes Problem(Gower&Dijksterhuis 2004), and the solution is $\hat{A} = UV'$, where U, V is the SVD of $Y'XB$,i.e.$Y'XB = UDV'$

Then with the fixed A, considering the column vector of A is orthogonal, we can let $(A, A^{\perp})$ be an orthogonal matrix. Then

$$||Y-XBA'||^2 = ||(Y-XBA')(A, A^{\perp})||^2 = ||YA-XB||^2 + ||YA^{\perp}||^2$$

For fixed B, the optimization problem is equivalent with

$$\min_A ||Y - XBA'|| \quad such \quad that \quad A'A = I$$

which is an orthogobal Procrustes Problem(Gower&Dijksterhuis 2004), and the solution is $\hat{A} = UV'$, where U, V is the SVD of $Y'XB$,i.e.$Y'XB = UDV'$
Then with the fixed A, considering the column vector of A is orthogonal, we can let $(A, A^{\perp})$ be an orthogonal matrix. Then

$$||Y-XBA'||^2 = ||(Y-XBA')(A, A^{\perp})||^2 = ||YA-XB||^2 + ||YA^{\perp}||^2$$

The optimization problem is equivalent with, for A fixed, $\min_B ||YA - XB||^2 + \sum_{i=1}^{p} \lambda_i ||B^i||$

The subgradient quations about B^ℓ is defined as follows

$$2X'_\ell(XB - YA) + \lambda_\ell s_\ell \quad \forall \ell = 1, 2, \cdots, p$$

where $s_\ell = \frac{B^\ell}{||B^\ell||}, ||B^\ell|| \neq 0$

The subgradient quations about B^ℓ is defined as follows

$$2X_\ell'(XB - YA) + \lambda_\ell s_\ell \quad \forall \ell = 1, 2, \cdots, p$$

where $s_\ell = \frac{B^\ell}{||B^\ell||}, ||B^\ell|| \neq 0$

If $||B^\ell|| = 0$, the equation becomes

$$2X_\ell'(\sum_{k\neq\ell}^{p} X_k B^k - YA) + \lambda_\ell s_\ell = 0$$

, then $s_\ell = -\frac{2}{\lambda_\ell} X_\ell'(\sum_{k\neq\ell}^{p} X_k B^k - YA) := \frac{2}{\lambda_\ell} X_\ell' R_\ell$

resulting in

$$B^\ell = (X'_\ell X_\ell + \frac{\lambda_\ell}{2||B^\ell||})^{-1} X'_\ell R_\ell$$

. Noting that the RHS involves $||B^\ell||$, we let $c = ||B^\ell||$ and solve the equation and plug in $c = \frac{||X'_\ell R_\ell|| - \frac{1}{2}\lambda_\ell}{||X_\ell||^2}$ we get the final solution

$$B^\ell = \frac{1}{X'_\ell X_\ell}(1 - \frac{\lambda_\ell}{2||X'_\ell R_\ell})_+ X'_\ell R_\ell$$

which is a vector version of the soft-thresholding rule.

Noting that the truth that $\min_C \frac{1}{2}(cx^2 + \frac{1}{c}) = |x|$, then the problem is equivalent

$$f = ||YA - XB||^2 + \sum_{i=1}^{p} \frac{\lambda_i}{2}(\mu_i||B^i||^2 + \frac{1}{\mu_i}) \quad \textit{jointly} \quad \textit{over} \quad B \quad \textit{and} \quad \mu_i$$

Noting that the truth that $\min_C \frac{1}{2}(cx^2 + \frac{1}{c}) = |x|$, then the problem is equivalent

$$f = ||YA - XB||^2 + \sum_{i=1}^{p} \frac{\lambda_i}{2}(\mu_i ||B^i||^2 + \frac{1}{\mu_i}) \quad \textit{jointly over } B \textit{ and } \mu_i$$

For fixed B, the solution of μ_i is that $\mu_i = \frac{1}{||B^i||}, i = 1, \cdots, p$

For fixed μ_i, we have $\frac{\partial f}{\partial B^i} = -2X_i'(YA - XB) + \lambda_i \mu_i B^i = 0$

Noting that the truth that $\min_C \frac{1}{2}(cx^2 + \frac{1}{c}) = |x|$, then the problem is equivalent

$$f = ||YA - XB||^2 + \sum_{i=1}^{p} \frac{\lambda_i}{2}(\mu_i ||B^i||^2 + \frac{1}{\mu_i}) \quad jointly \quad over \quad B \quad and \quad \mu_i$$

For fixed B, the solution of μ_i is that $\mu_i = \frac{1}{||B^i||}, i = 1, \cdots, p$
For fixed μ_i, we have $\frac{\partial f}{\partial B^i} = -2X_i'(YA - XB) + \lambda_i \mu_i B^i = 0$
As a result, the final solution is

$$B = \{X'X + \frac{1}{2}diag(\lambda_i \mu_i, \cdots, \lambda_p \mu_p)\}^{-1} X'YA$$

We give the following algorithm of iteration

Input:X, Y, λ

Output:A, B

while(A, B are not convergent)

for fixed B, we get the solution of A by SVD

while(B is not convergent)

for every ℓ solve B^{ℓ} and check whether B is convergent

Assumptions

Assuming p, q are fixed with n going to infinity

1 :There's a positive definite matrix Σ such that $\frac{XX'}{n} \to \Sigma(n \to \infty)$

2 :The first p_0 variables are important and the rest are irrelevant i.e. $||C_*^i|| > 0(i \leq p_0)$ and $||C_*^i|| = 0(i > p_0)$ where C_*^i is the ith row of C^*,which is the rank-r coefficient matrix used to generate data in the model

Theorem 2.1 (consistency of parameter estimation)

suppose $\frac{\lambda_i}{\sqrt{n}} = \frac{\lambda_{n,i}}{\sqrt{n}} \to 0, \forall i \leq p_0$ *Then*

- *There is a local minimizer* $\hat{C}$ *that is* $\sqrt{n}$*-consistent in estimating* C_*
- $\hat{C} = \hat{U}\hat{D}\hat{V}'$ *is SVD,* $\hat{C} = \hat{U}\hat{D}\hat{V}$ *are* $\sqrt{n}$*-consistent in estimating* U_*, D_*, V_*

Theorem 2.1 (consistency of parameter estimation)

suppose $\frac{\lambda_i}{\sqrt{n}} = \frac{\lambda_{n,i}}{\sqrt{n}} \to 0, \forall i \leq p_0$ *Then*

- *There is a local minimizer* $\hat{C}$ *that is* $\sqrt{n}$*-consistent in estimating* C_*
- $\hat{C} = \hat{U}\hat{D}\hat{V}'$ *is SVD,* $\hat{C} = \hat{U}\hat{D}\hat{V}$ *are* $\sqrt{n}$*-consistent in estimating* U_*, D_*, V_*

Theorem 2.2 (concistency of variable selection)

if $\frac{\lambda_{n,i}}{\sqrt{n}} \to 0$ *for* $i \leq p_0$ *and* $\frac{\lambda_{n,i}}{\sqrt{n}} \to \infty (i > p_0)$, *then*

$P(\hat{C}^i = 0) = P(\hat{U}^i = 0) \to 1 \quad i > p_0$

$$Y = XC + E \quad C \in \mathbb{R}^{p \times q} \quad rankC \leq r$$

Using the polar decomposition, we have $C = U\tilde{V}'$,resulting in $Y = XU\tilde{V}' + E$

$$Y = XC + E \quad C \in \mathbb{R}^{p\times q} \quad rankC \leq r$$

Using the polar decomposition, we have $C = U\tilde{V}'$,resulting in
$Y = XU\tilde{V}' + E$
This equation is related to a factor analysis model:XU can be regarded as a common factor matrix and $\tilde{V}$ can be regarded as a loading matrix. Furthermore, if we assume
$\mathbb{E}[x_i] = 0, cov[x_i] = \Gamma_i$, then $cov[U'X] = U'\Gamma U = I_r$

University of Science and Technology of China

$$Y = XC + E \quad C \in \mathbb{R}^{p\times q} \quad rankC \leq r$$

Using the polar decomposition, we have $C = U\tilde{V}'$,resulting in $Y = XU\tilde{V}' + E$

This equation is related to a factor analysis model:XU can be regarded as a common factor matrix and $\tilde{V}$ can be regarded as a loading matrix. Furthermore, if we assume $\mathbb{E}[x_i] = 0, cov[x_i] = \Gamma_i$, then $cov[U'X] = U'\Gamma U = I_r$

We have the SFAR model(sequential co-sparse factor regression):

$$Y = XUDV' \quad such \quad that \quad U'\Gamma U = I_r \quad V'V = I_r$$

with the coefficient matrix is $C = UDV'$

We have the optimization problem

$$\min_{U,D,V} \frac{1}{2}||Y - XUDV'||^2 + \lambda_1 \sum_{i=1}^{p} \sum_{j=1}^{r} w_{ij}^{(u)} |u_{ij}| + \lambda_2 \sum_{i=1}^{q} \sum_{j=1}^{r} w_{ij}^{(v)} |v_{ij}|$$

$$\textit{such} \quad \textit{that} \quad U' \frac{X'X}{n} U = I_r \quad V'V = I_r$$

where $w_{ij}^{(u)}, w_{ij}^{(v)}$ is called adaptive weights with positive values

The minimization problem for the kth latent factor is given by

$$\min_{dk,\boldsymbol{u_k},\boldsymbol{v_k}} \frac{1}{2}||Y_k - d_k X \boldsymbol{u_k} \boldsymbol{v_k}^T||^2 + \sum_{i=1}^{p} w_{ki}^{(u)}|u_{ki}| + \sum_{i=1}^{q} w_{ki}^{(w)}|v_{ki}|$$

$$such \quad that \quad d_k \geq 0 \quad \boldsymbol{u_k}^T X' X \boldsymbol{u_k} = n \quad \boldsymbol{v_k} \boldsymbol{v_k} = 1$$

where $Y_k = Y - \sum_{j=1}^{k-1} d_j X \boldsymbol{u_j} \boldsymbol{v_j}^T$(**SeCURE algorithm Mishra(2017)**)

Definition 3.1

- $St(r,q) := \{V \in \mathbb{R}^{q\times r} | V'V = I_r\} (q \geq r)$ *called Stiefel manifold*
- $GSt(r,p) := \{U \in \mathbb{R}^{p\times r} | U'GU = I_r\} (p \geq r)$, $G \in \mathbb{R}^{p\times p}$ *is definite positive, called generalized Stiefel manifold.*

In this paper, we use $G = \frac{X'X}{n}$ The optimization is equivalent with

$$\min_{U\in GSt(r,p), D\in\mathbb{R}^{r\times r}, V\in St(r,q)} \frac{1}{2}||Y - XUDV'||^2 + n\lambda_1 \sum_{i=1}^{p}\sum_{j=1}^{r} w_{ij}^{(u)} |u_{ij}| +$$

$$n\lambda_2 \sum_{i=1}^{q}\sum_{j=1}^{r} w_{ij}^{(v)} |v_{ij}|$$

We propose the following minimization problem

$$\min_{U\in GSt(r,p),D\in\mathbb{R}^{r\times r},V\in St(r,q)} \frac{1}{2}||Y-XUDV'||^2+n\lambda_1\sum_{i=1}^{p}\sum_{j=1}^{r}w_{ij}^{(u)}|u_{ij}|+$$

$$n\alpha\lambda_2\sum_{i=1}^{q}\sum_{j=1}^{r}w_{ij}^{(v)}|v_{ij}|+n\sqrt{q}(1-\alpha)\lambda_2\sum_{i=1}^{r}w_i^{(d)}I(\mathbf{v_i}\neq 0)$$

$w_i^{(d)}$ is an adaptive weight with a positive value proposed by Zou(2006) and the group selection in the fourth term plays the role of the rank selection of the coefficient matrix C.

$$Y = XA + E \quad r = rankA \quad q = rankX$$

Let J denotes the index set of the nonzero rows of A. Only $r(n + |J| - r)$ free parameters need to be estimated by SVD, where $|J| = \#J$
We can reduce X of rank q to an $m \times q$ matrix with q independent columns and always assume that $|J| \leq q$. Using penalized least squares methods, removing predictor X_j from the model is equivalent with setting the jth row in A to zero.

JRRS is short for the single-stage joint rank and row selection estimator.We can modify the rank selection criterion(RSC) as

$$\hat{B} = arg\min_{B}\{||Y - XB||^2 + pen(B)\} \qquad (1)$$

where

$pen(B) = c\sigma^2 r(B)(2n + log(2e)|J(B)| + |J(B)| log\frac{ep}{|J(B)|})$,and we call the equation (1) JRRS1

JRRS is short for the single-stage joint rank and row selection estimator.We can modify the rank selection criterion(RSC) as

$$\hat{B} = arg\min_{B}\{||Y - XB||^2 + pen(B)\} \qquad (1)$$

where

$pen(B) = c\sigma^2 r(B)(2n + log(2e)|J(B)| + |J(B)|log\frac{ep}{|J(B)|})$,and we call the equation (1) JRRS1

Theorem 4.1

The single-stage JRRS estimator $\hat{B}$ using pen(B) with $c = 12^3$ satisfies

$$\mathbb{E}[||XA - X\hat{B}||^2] \leq 10||XA - XB||^2 + 8pen(B) + 768n\sigma^2 e^{-\frac{n}{2}} \quad \forall r(B) \geq 1$$

In particular if $r(A) >\geq 1$

$$\mathbb{E}[||XA - X\hat{B}||^2] \lesssim \sigma^2 r(A)(n + |J(A)|log\frac{p}{|J(A)|})$$

Theorem 4.2

For any collection of (random) nonzero matrices $B_1, B_2, \cdots,$
the single-stage JRRS estimator
$\tilde{B} = arg\min_{B_j}(||Y - XB_j||^2 + pen(B_j))$ *with* $c = 12$, *satisfies*

$$\mathbb{E}[||XA - X\tilde{B}||^2] \leq inf\{10\mathbb{E}[||XA - XB_j||^2] + 8\mathbb{E}[pen(B_j)]\} + 768n\sigma^2 e^{-\frac{n}{2}}$$

We propose a convex relaxation for the pen(B) with $||B||_{2,1} = \sum_{j=1}^{p} ||b_j||_2$ rows b_j of B

$$\hat{B}_k = arg \min_{r(B) \leq k} \{||Y - XB||^2 + 2\lambda ||B||_{2,1}\} \tag{2}$$

The expression (2)is called **rank constrained group lasso(RCGL)**(here we have two parameters nedded to be tuned k and λ)

We propose a convex relaxation for the pen(B) with $||B||_{2,1} = \sum_{j=1}^{p} ||b_j||_2$ rows b_j of B

$$\hat{B}_k = arg \min_{r(B)\leq k} \{||Y - XB||^2 + 2\lambda||B||_{2,1}\} \qquad (2)$$

The expression (2)is called **rank constrained group lasso(RCGL)**(here we have two parameters nedded to be tuned k and λ)

Assumption A:We say $\Sigma \in \mathbb{R}^{p\times p}$ satisfies condition $A(I, \delta_I)$ for an index set $I \subset \{1, \cdots, p\}$ and $\delta_I > 0 \iff tr(M'\Sigma M) \geq \delta_I \sum_{i\in I} ||m_i||_2^2$ for all $p \times n$ matrices M(with rows m_i)satisfying $\sum_{i\in I} ||m_i||_2 \leq 2\sum_{i\in I^c} ||m_i||_2$(which is ℓ_2 norm) In our paper, $\Sigma = \frac{X'X}{m}$,and we have the following remark.

Note

- *the constant 2 can be replaced by any constant>1*
- *A sufficient condition is : there exists a diagonal matrix D with* $D_{jj} = \delta_I$ *for* $j \in I$ *and* D_{jj} *otherwise such that* $\Sigma - D > 0$

Let $\lambda_1(\Sigma)$ denotes the largest eigenvalue of Σ and set the parameter $\lambda = c\sigma\sqrt{\lambda_1(\Sigma) kmlog(ep)} (c > 0)$

Theorem 4.3

Let $\hat{B}_k$ be the global minimizer corresponding λ above with c large enough. Then

$$\mathbb{E}[||X\hat{B}_K - XA||^2] \lesssim ||XB - XA||^2 + k\sigma^2(n + (1 + \frac{\lambda_1(\Sigma)}{\delta_{J(B)}})|J(B)|log(p))$$

$\forall B \in \mathbb{R}^{p\times n}$ *with* $1 \leq r(B) \leq k$ *provided* Σ *satisfies Assumption* $A(J(B), \delta_{J(B)})$

method 1(RSC→RCGL)

- Use RSC to select $k = \hat{r}$(review the last report) as the number of singular values of PY that exceed $\sigma(\sqrt{2n} + \sqrt{2q}), P = X(X'X)^{-1}X'$
- Compute the rank constrained GLASSO estimator $\hat{B}_k$ with $k = \hat{r}$ to obtain the final estimator $\hat{B}^{(1)} = \hat{B}r$

- Use RSC to select $k = \hat{r}$(review the last report) as the number of singular values of PY that exceed $\sigma(\sqrt{2n} + \sqrt{2q}), P = X(X'X)^{-1}X'$
- Compute the rank constrained GLASSO estimator $\hat{B}_k$ with $k = \hat{r}$ to obtain the final estimator $\hat{B}^{(1)} = \hat{B}r$

This two-step estimator adapts to both rank and row sparsity under 2 additional mild restrictions

- Ass C_1: $d_r(XA) > 2\sqrt{2}\sigma(\sqrt{n} + \sqrt{q})$ (RSC)
- Ass C_2: $log(||XA||_F) \leq (\sqrt{2} - 1)^2 \frac{n+q}{4}$

- Use RSC to select $k = \hat{r}$(review the last report) as the number of singular values of PY that exceed $\sigma(\sqrt{2n} + \sqrt{2q}), P = X(X'X)^{-1}X'$
- Compute the rank constrained GLASSO estimator $\hat{B}_k$ with $k = \hat{r}$ to obtain the final estimator $\hat{B}^{(1)} = \hat{B}r$

This two-step estimator adapts to both rank and row sparsity under 2 additional mild restrictions

- Ass C_1: $d_r(XA) > 2\sqrt{2}\sigma(\sqrt{n} + \sqrt{q})$ (RSC)
- Ass C_2: $log(||XA||_F) \leq (\sqrt{2} - 1)^2 \frac{n+q}{4}$

And we have the following property.

Theorem 4.4
Let Σ satisfy $A(J, \delta_J)$ with $J = J(A) \neq \Phi$, let $\frac{\lambda_1(\Sigma)}{\delta_J}$ be bounded and let C_1, C_2 hold.

$$\mathbb{E}[||X\hat{B}^{(1)} - XA||^2] \lesssim nr + |J|r \cdot log(p)$$

The practical choice of the threshold $2\sigma(\sqrt{n} + \sqrt{q})$ can be done by CV

- pre-specify a grid Λ of values for λ and use (2) to construct $\mathcal{B} = \{\hat{B_{k,\lambda}} : \lambda \in \Lambda\}$
- Compute $\hat{B^{(2)}} = \arg\min_{B\in\mathcal{B}}(||Y - XB||^2 + pen(B))$

- pre-specify a grid Λ of values for λ and use (2) to construct $\mathcal{B} = \{\hat{B}_{k,\lambda} : \lambda \in \Lambda\}$
- Compute $\hat{B}^{(2)} = \arg\min_{B \in \mathcal{B}}(||Y - XB||^2 + pen(B))$

Theorem 4.5

Provided Σ satisfies condition $A(J, \delta_J)$ with $J = J(A) \neq \Phi$, $\lambda_1(\Sigma)/\delta_J$ is bounded, and Λ contains λ for $c >> 1$

$$\mathbb{E}[||X\hat{B}^{(2)} - XA||^2] \lesssim nr + |J|log(p)r$$

and $\hat{B}^{(2)}$ has the same rate as $\hat{B}^{(1)}$

- Select the predictors via GLASSO
- Based only on the selected predictors, use RSC to construct an adaptive estimator of reduced rank

For minimizing $F(B;\lambda) = \frac{1}{2}||Y - XB||^2 + \lambda||B||_{2,1}$ over all $p \times n$ matrices B of rank less than or equal to k. By using the polar decomposition $B = SV'$, where V is orthogonal and S is semi-positive definite.

$$(\hat{S}, \hat{V}) := F(S, V; \lambda) = arg \min_{S \in \mathbb{R}^{p \times k}, V \in O^{n \times k}} \frac{1}{2}||Y - XCV'||^2 + \lambda||S||_{2.1} \tag{3}$$

We propse the following iterative optimization precedure.

Given $1 \leq k \leq m \wedge p \wedge n, \lambda \geq 0, V_{k,\lambda}^{(0)} \in O^{n\times k}$(first k columns of $I_{n\times n}$)

$j \leftarrow 0$,converged$\leftarrow$ FALSE while not converged do:

(a). $S_{k,\lambda}^{(j+1)} \leftarrow arg\min_{S\in\mathbb{R}^{p\times k}} \frac{1}{2}||YV_{k,\lambda}^{(j)} - XS||^2 + \lambda||S||_{2,1}$

(b). $W \leftarrow Y'XS_{k,\lambda}^{(j+1)}, W \in \mathbb{R}^{n\times k}$,Using SVD $W = U_w D_w V'w$

(c). $V_{k,\lambda}^{(j+1)} \leftarrow U_w V'_w$

(d). $B_{k,\lambda}^{(j+1)} \leftarrow S_{k,\lambda}^{(j+1)}(V_{k,\lambda}^{(j+1)})'$

(e). $converged \leftarrow |F(B_{k,\lambda}^{(j+1)};\lambda) - F(B_{k,\lambda}^{(j)};\lambda)| < \epsilon$

(f). $j \leftarrow j+1$

end while and diliver $\hat{B_{k,\lambda}} = B_{k,\lambda}^{(j+1)}, \hat{S_{k,\lambda}} = S_{k,\lambda}^{(j+1)}, \hat{V_{k,\lambda}} = V_{k,\lambda}^{(j+1)}$

Note

We run the algorithm to obtain a solution path, for each (k, λ) in a 2-dimensional grid or a grid of λ with k determined by RSC.

From the solution path, we get a series of candidate estimates. Then the single stage JRRS or other tuning criterion can be used to select the optimal estimate.

Note

We run the algorithm to obtain a solution path, for each (k, λ) *in a 2-dimensional grid or a grid of* λ *with* k *determined by RSC.*

From the solution path, we get a series of candidate estimates. Then the single stage JRRS or other tuning criterion can be used to select the optimal estimate.

Step (a) needs to solve a GLASSO optimization problem.

$$Y = XB + E \to Y = F\Omega + E$$

where $B = \Gamma\Omega, F = X\Gamma, \Gamma \in \mathbb{R}^{p\times r}$ for $r \leq min(p, q)$. The $\Omega \in \mathbb{R}^{r\times q}$ is called factor loading matrix. The columns of $F, F_j(j = 1, \cdots, r)$ represent the so-called factors

$$Y_j = XB_j + E_j$$

$$Y = XB + E \rightarrow Y = F\Omega + E$$

where $B = \Gamma\Omega, F = X\Gamma, \Gamma \in \mathbb{R}^{p\times r}$ for $r \leq min(p, q)$. The $\Omega \in \mathbb{R}^{r\times q}$ is called factor loading matrix. The columns of $F, F_j (j = 1, \cdots, r)$ represent the so-called factors

$$Y_j = XB_j + E_j$$

(j represents the columns of a matrix) The basic idea of dimension reduction is that the regression coefficient $B_1, \cdots, B_q$ actually come from a linear space $\mathcal{B}$ of dimension lower than p.

$$Y = XB + E \to Y = F\Omega + E$$

where $B = \Gamma\Omega, F = X\Gamma, \Gamma \in \mathbb{R}^{p\times r}$ for $r \le min(p, q)$. The $\Omega \in \mathbb{R}^{r\times q}$ is called factor loading matrix. The columns of $F, F_j(j = 1, \cdots, r)$ represent the so-called factors

$$Y_j = XB_j + E_j$$

(j represents the columns of a matrix) The basic idea of dimension reduction is that the regression coefficient $B_1, \cdots, B_q$ actually come from a linear space $\mathcal{B}$ of dimension lower than p.

As a result, we have a set of basis elements $\{\eta_1, \cdots, \eta_p\}$ for $\mathbb{R}^p$ and a subset $\mathcal{A} \subset \{1, \cdots, p\}$ such that $\mathcal{B} \subset span\{\eta_i : i \in \mathcal{A}\}$

Now we have the model as follows:

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

where $F = F_1, \cdots, F_p), F_i = X\eta_i, \Gamma = (\eta_1, \cdots, \eta_p), B = \Gamma\Omega = (\eta_1, \cdots, \eta_p)\Omega$.

Now we have the model as follows:

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

where $F = F_1, \cdots, F_p), F_i = X\eta_i, \Gamma = (\eta_1, \cdots, \eta_p), B = \Gamma\Omega = (\eta_1, \cdots, \eta_p)\Omega$.A family of estimates for this can be obtained by

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

where ω_i is the ith row of Ω, W is a weight matrix with common choices Σ^{-1} or I(which is corresponding to Frobenius norm).

Now we have the model as follows:

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

where $F = F_1, \cdots, F_p), F_i = X\eta_i, \Gamma = (\eta_1, \cdots, \eta_p), B = \Gamma\Omega = (\eta_1, \cdots, \eta_p)\Omega$.A family of estimates for this can be obtained by

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

where ω_i is the ith row of Ω, W is a weight matrix with common choices Σ^{-1} or I(which is corresponding to Frobenius norm). Here we assume $W = I$.

$$Y = X\eta_1\omega_1 + \cdots + X\eta_p\omega_p$$

The ith factor will be included if and only if ω_i is non-zero.

$$Y = X\eta_1\omega_1 + \cdots + X\eta_p\omega_p$$

The ith factor will be included if and only if ω_i is non-zero.We choose $\alpha = 2$ and we need to obtain ηs first.

$$Y = X\eta_1\omega_1 + \cdots + X\eta_p\omega_p$$

The ith factor will be included if and only if ω_i is non-zero.We choose $\alpha = 2$ and we need to obtain ηs first.We choose ηs to be the eigenvectors of BB', because this set of basis contains the basis of $\mathcal{B}$.

$$Y = X\eta_1\omega_1 + \cdots + X\eta_p\omega_p$$

The ith factor will be included if and only if ω_i is non-zero.We choose $\alpha = 2$ and we need to obtain ηs first.We choose ηs to be the eigenvectors of BB', because this set of basis contains the basis of $\mathcal{B}$. We can understand this by the following truth: $B = UDV'$ is the SVD, then $BB' = UD^2U'$ we choose $(\eta_1, \cdots, \eta_p) = U$ span the column space of B
Then $\Omega = DV', \omega = \sigma_i v_i$,where v_i is the ith column of V,and $||\omega_i|| = \sigma_i$.

$$Y = X\eta_1\omega_1 + \cdots + X\eta_p\omega_p$$

The ith factor will be included if and only if ω_i is non-zero.We choose $\alpha = 2$ and we need to obtain ηs first.We choose ηs to be the eigenvectors of BB', because this set of basis contains the basis of $\mathcal{B}$. We can understand this by the following truth: $B = UDV'$ is the SVD, then $BB' = UD^2U'$ we choose $(\eta_1, \cdots, \eta_p) = U$ span the column space of B
Then $\Omega = DV', \omega = \sigma_i v_i$,where v_i is the ith column of V,and $||\omega_i|| = \sigma_i$. Finally we have the optimization problem:

$$min(tr(Y - XB)(Y - XB)') \quad subject \quad to \quad \sum_{i=1}^{min(p,q)} \sigma_i \leq t \quad (4)$$

The last term is known as Ky Fan norm for B. There's is no restriction of B because once the estimation B is available, the basis ηs can be obtained as its left singular vectors U. Therefore, we can also compute the factors $F_i = X\eta_i$ and loading $\Omega = DV$.

The last term is known as Ky Fan norm for B. There's is no restriction of B because once the estimation B is available, the basis ηs can be obtained as its left singular vectors U.
Therefore, we can also compute the factors $F_i = X\eta_i$ and loading $\Omega = DV$.
For the tuning α cases, we get the optimization problem $tr(Y - XB)(Y - XB)'$ subject to $(\sum_i \sigma_i^{\alpha})^{\frac{1}{\alpha}} \leq t$ and RRR is another special case of expression with $\alpha = 0^+$

Lemma 5.1

Let $\hat{U}^{LS}\hat{D}^{LS}\hat{V}^{LS}$ be the SVD of the least squares estimate $\hat{B}^{LS}$. Then, under the orthogonal design where $X'X = nI$, the minimizer of (4) is $\hat{B} = \hat{U}^{LS}\hat{D}(\hat{V}^{LS})'$, $\hat{D}_{ii} = max(\hat{D}^{LS}_{ii}, 0)$(singular values are shrunk), and $\lambda \geq 0$ is a constant such that $\sum_i \hat{D}_{ii} = min(t, \sum_i \hat{D}^{LS}_{ii})$

Lemma 5.1

Let $\hat{U}^{LS}\hat{D}^{LS}\hat{V}^{LS}$ *be the SVD of the least squares estimate* $\hat{B}^{LS}$*. Then, under the orthogonal design where* $X'X = nI$*, the minimizer of (4) is* $\hat{B} = \hat{U}^{LS}\hat{D}(\hat{V}^{LS})'$, $\hat{D}_{ii} = max(\hat{D}_{ii}^{LS}, 0)$*(singular values are shrunk), and* $\lambda \geq 0$ *is a constant such that* $\sum_i \hat{D}_{ii} = min(t, \sum_i \hat{D}_{ii}^{LS})$

Note

In fact, the λ *arisen is the result of Lagrange multipilication*

Lemma 5.1

Let $\hat{U}^{LS}\hat{D}^{LS}\hat{V}^{LS}$ be the SVD of the least squares estimate $\hat{B}^{LS}$. Then, under the orthogonal design where $X'X = nI$, the minimizer of (4) is $\hat{B} = \hat{U}^{LS}\hat{D}(\hat{V}^{LS})'$, $\hat{D}_{ii} = max(\hat{D}^{LS}_{ii}, 0)$(singular values are shrunk), and $\lambda \geq 0$ is a constant such that $\sum_i \hat{D}_{ii} = min(t, \sum_i \hat{D}^{LS}_{ii})$

Note

In fact, the λ arisen is the result of Lagrange multipilication

Lemma 5.2

Suppose that $max(p, q) = o(n)$, under the orthogonal design, if $\lambda \to 0$ in such a fashion that $\frac{max(p,q)}{n} = o(\lambda^2)$. Then $|\sigma_i(\hat{B}) - \sigma_i(B)| \to 0$ with probability if $\sigma(B) > 0$ and $P(\sigma(\hat{B}) = 0) \to 1$ if $\sigma(B) = 0$

We develop a GCV type of statistic for determining t.We give a lagrange form:

$$Q_n(B) = \frac{1}{2} tr(Y - XB)(Y - XB)' + n\lambda \sum_{i=1}^{p \wedge q} \sigma_i(B) \tag{5}$$

The following lemma explicitly describes the relationship between t and λ

We develop a GCV type of statistic for determining t.We give a lagrange form:

$$Q_n(B) = \frac{1}{2} tr(Y - XB)(Y - XB)' + n\lambda \sum_{i=1}^{p \wedge q} \sigma_i(B) \qquad (5)$$

The following lemma explicitly describes the relationship between t and λ

Lemma 5.3

write $\hat{d}_i = \hat{D}_{ii}$ for $i = 1, \cdots, p \wedge q$. For any $t \leq \sum_i \hat{d}_i$, the minimizer of equation (5) coincides with the minimizer of (4) if

$$n\lambda = \frac{1}{\#(\hat{d}_i > 0)} \sum_{\hat{d}_i > 0} (\tilde{X}_i' \tilde{Y}_i - \tilde{X}_i' \tilde{X}_i \hat{d}_i) \qquad (6)$$

$\tilde{Y}_i$ is the ith column of $\tilde{Y} = Y\hat{U}$ and $\tilde{X}_i$ is the ith column of $\tilde{X} = X\hat{V}$

$\tilde{Y}_i$ is the ith column of $\tilde{Y} = Y\hat{U}$ and $\tilde{X}_i$ is the ith column of $\tilde{X} = X\hat{V}$

Note

We can transform $\sum_{i=1}^{p\wedge q} \sigma_i(B)$ *as follows:*

$$\sum_{i=1}^{p\wedge q} \sigma_i(B) = \sum_{i=1}^{p\wedge q} \hat{D}_{ii} = \sum_{i=1}^{p} \sigma_i(\hat{B}K\hat{B}') = tr(\hat{B}K\hat{B}')$$

where $K = \sum_{\hat{D}_{ii}>0} \frac{1}{\hat{D}_{ii}} \hat{v}_i \hat{v}_i'$.

$\tilde{Y}_i$ is the ith column of $\tilde{Y} = Y\hat{U}$ and $\tilde{X}_i$ is the ith column of $\tilde{X} = X\hat{V}$

Note

We can transform $\sum_{i=1}^{p\wedge q} \sigma_i(B)$ as follows:

$$\sum_{i=1}^{p\wedge q} \sigma_i(B) = \sum_{i=1}^{p\wedge q} \hat{D}_{ii} = \sum_{i=1}^{p} \sigma_i(\hat{B}K\hat{B}') = tr(\hat{B}K\hat{B}')$$

where $K = \sum_{\hat{D}_{ii}>0} \frac{1}{\hat{D}_{ii}} \hat{v}_i\hat{v}'_i$.Considering that $\hat{B}K\hat{B}' = \sum_{\hat{D}_{ii}>0} \frac{1}{D_{ii}} \hat{B}\hat{v}_i(\hat{B}\hat{v}_i)' = \sum_i \frac{1}{D_{ii}} \sigma_i u_i \sigma_i u'_i = \sum_i \sigma_i u_i u'_i$ Using $u_i \perp u_j$,we get the eigenvalue of $\hat{B}K\hat{B}'$ is σ_i. However, $(\hat{B}K\hat{B}')(\hat{B}K\hat{B}')' = \sum_i \sigma_i^2 u_i u'_i$. As a result, the singular value of $\hat{B}K\hat{B}'$ is σ_i

Then we can transform the Lagrange form into

$$\frac{1}{2}tr(Y - XB)(Y - XB)' + n\lambda tr(BKB') \tag{7}$$

Since $\hat{B}$ is the minimizer of (7),it can be expressed as
$\hat{B} = (X'X + 2n\lambda K)^{-1}X'Y$

Then we can transform the Lagrange form into

$$\frac{1}{2}tr(Y - XB)(Y - XB)' + n\lambda tr(BKB') \tag{7}$$

Since $\hat{B}$ is the minimizer of (7),it can be expressed as
$\hat{B} = (X'X + 2n\lambda K)^{-1}X'Y$
We can define the hat matrix for expression (6) as
$H = X(X'X + 2n\lambda K)^{-1}X'$ and the degree of freedom as
$df(t) = qtrH$

Then we can transform the Lagrange form into

$$\frac{1}{2}tr(Y - XB)(Y - XB)' + n\lambda tr(BKB') \tag{7}$$

Since $\hat{B}$ is the minimizer of (7),it can be expressed as
$\hat{B} = (X'X + 2n\lambda K)^{-1}X'Y$
We can define the hat matrix for expression (6) as
$H = X(X'X + 2n\lambda K)^{-1}X'$ and the degree of freedom as
$df(t) = qtrH$
The GCV score is given by $GCV(t) = \frac{tr(Y - X\hat{B})(Y - X\hat{B})'}{qp - df(t)}$. We choose a tuning parameter by minimizing GCV(t).

To sum up:

Step 1: for each candidate t-value

(a). compute the minimizer of (4) (denote the solution $\hat{B}(t)$)

(b). evaluate λ by using (6)

(c). compute the GCV score

Step 2: denote t^* the minimizer of the GCV score. Return $\hat{B}(t^*)$ as the estimator of B

$$Y = XC + E \quad rankC = r^*$$

We have SVD:

$$C = UDV' = \sum_{k=1}^{r^*} d_k \boldsymbol{u_k} \boldsymbol{v_k'} := \sum_{k=1}^{r^*} C_k$$

where $U = (\boldsymbol{u_1}, \cdots, \boldsymbol{u_{r^*}})$, $V = (\boldsymbol{v_1}, \cdots, \boldsymbol{v_{r^*}})$, $C_k = d_k \boldsymbol{u_k} \boldsymbol{v_k'}$
C_k is the layer k unit rank matrix of C.

$$Y = XC + E \quad rankC = r^*$$

We have SVD:

$$C = UDV' = \sum_{k=1}^{r^*} d_k \boldsymbol{u_k} \boldsymbol{v_k'} := \sum_{k=1}^{r^*} C_k$$

where $U = (\boldsymbol{u_1}, \cdots, \boldsymbol{u_{r^*}}), V = (\boldsymbol{v_1}, \cdots, \boldsymbol{v_{r^*}}), C_k = d_k \boldsymbol{u_k} \boldsymbol{v_k'}$
C_k is the layer k unit rank matrix of C.Here all the singular values are assumed to be distinct so that this SVD is unique because in practice, the singular values rarely conincide.

We propse to estimate C by minimizing the following objective function with respect to $(d_k, \boldsymbol{u_k}, \boldsymbol{v}_{\boldsymbol{k}}')$ for $k = 1, \cdots, r^*$

$$\frac{1}{2}||Y - X\sum_{k=1}^{r^*} d_k \boldsymbol{u_k} \boldsymbol{v}_{\boldsymbol{k}}'||^2 + \sum_{k=1}^{r^*} Pe(\lambda_k, (d_k, u_k, \boldsymbol{v}_{\boldsymbol{k}}'))$$

We propse to estimate C by minimizing the following objective function with respect to $(d_k, \boldsymbol{u_k}, \boldsymbol{v'_k})$ for $k = 1, \cdots, r^*$

$$\frac{1}{2}||Y - X\sum_{k=1}^{r^*} d_k \boldsymbol{u_k} \boldsymbol{v'_k}||^2 + \sum_{k=1}^{r^*} Pe(\lambda_k, (d_k, u_k, \boldsymbol{v'_k}))$$

We consider

$$Pe = \lambda_k \sum_{i=1}^{p}\sum_{j=1}^{q} w_{ijk}|d_k u_{ik} v_{jk}| = \lambda_k(w_k^{(d)} d_k)(\sum_{i=1}^{p} w_{ik}^{(u)}|u_{ik})(\sum_{j=1}^{q} w_{jk}^{(v)}|v_{jk}|) \tag{8}$$

We propse to estimate C by minimizing the following objective function with respect to $(d_k, \boldsymbol{u_k}, \boldsymbol{v_k'})$ for $k = 1, \cdots, r^*$

$$\frac{1}{2}||Y - X\sum_{k=1}^{r^*} d_k \boldsymbol{u_k}\boldsymbol{v_k'}||^2 + \sum_{k=1}^{r^*} Pe(\lambda_k, (d_k, u_k, \boldsymbol{v_k'}))$$

We consider

$$Pe = \lambda_k \sum_{i=1}^{p}\sum_{j=1}^{q} w_{ijk}|d_k u_{ik} v_{jk}| = \lambda_k (w_k^{(d)} d_k)(\sum_{i=1}^{p} w_{ik}^{(u)}|u_{ik})(\sum_{j=1}^{q} w_{jk}^{(v)}|v_{jk}|) \tag{8}$$

where $w_{ijk} = w_k^{(d)} w_{ik}^{(u)} w_{jk}^{(v)}$ are data-driven weights to be done below. It can be viewed as penalizing each of the singular vectors comprising the SVD layer.

The weights can be chosen as

$$w^{(d)} = |\tilde{d}|^{-\gamma}$$

$$w^{(u)} = (w_1^{(u)}, \cdots, w_P^{(u)})' = |\tilde{u}|^{-\gamma}$$
$$w^{(v)} = (w_1^{(v)}, \cdots, w_q^{(q)})' = |\tilde{v}|^{-\gamma}$$

where γ is a prespecified non-negative parameter and $|\cdot|^{(-\gamma)}$ is defined **componentwise** for the enclosed vector (and we use $\gamma = 2$).

The weights can be chosen as

$$w^{(d)} = |\tilde{d}|^{-\gamma}$$

$$w^{(u)} = (w_1^{(u)}, \cdots, w_P^{(u)})' = |\tilde{u}|^{-\gamma}$$
$$w^{(v)} = (w_1^{(v)}, \cdots, w_q^{(q)})' = |\tilde{v}|^{-\gamma}$$

where γ is a prespecified non-negative parameter and $|\cdot|^{(-\gamma)}$ is defined **componentwise** for the enclosed vector (and we use $\gamma = 2$).When $r^* = 1$, the problem is with respect to $(d, \boldsymbol{u}, \boldsymbol{v})$:

$$\frac{1}{2}||Y - dX\boldsymbol{u}\boldsymbol{v}'||^2 + \lambda \sum_{i=1}^{q} \sum_{j=1}^{q} w_{ij}|du_i v_j|$$

For fixed $\boldsymbol{u}$ the problem with respect to $(d, \boldsymbol{v})$ becomes:

$$\frac{1}{2}||y - X^{(v)}\check{v}||^2 + \lambda^{(v)} \sum_{j=1}^{q} |\check{v}_j| \tag{9}$$

where
$\check{v} = diag(dw^{(v)})v, y = vec(Y), X^{(v)} = diag(w^{(v)})^{-1} \otimes (Xu)$ and
$\lambda^{(v)} = \lambda w^{(d)}(\sum_{i=1}^{p} w_i^{(u)}|u_i|)$

For fixed $\boldsymbol{u}$ the problem with respect to $(d, \boldsymbol{v})$ becomes:

$$\frac{1}{2}||y - X^{(v)}\check{v}||^2 + \lambda^{(v)}\sum_{j=1}^{q}|\check{v}_j| \tag{9}$$

where
$\check{v} = diag(dw^{(v)})v, y = vec(Y), X^{(v)} = diag(w^{(v)})^{-1} \otimes (Xu)$ and
$\lambda^{(v)} = \lambda w^{(d)}(\sum_{i=1}^{p} w_i^{(u)}|u_i|)$
This model can be recognized as a lasso regression with respect to $\check{v}$

In contrast, for fixed v, the problem with respect to $(d, \boldsymbol{u})$ becomes

$$\frac{1}{2}||y - X^{(u)}\check{u}||^2 + \lambda^{(u)} \sum_{i=1}^{p} |\check{u}_i| \tag{10}$$

where $\check{u} = diag(dw^{(u)})u, X^{(u)} = v \otimes Xdiag(w^{(u)})^{-1}, \lambda^{(u)} = \lambda w^{(d)}(\sum_{j=1}^{q} w_j^{(v)}|v_j|)$. Again this is a lasso regression problem with respect to $\check{u}$

Denote $(\hat{d}^{(\lambda)}, \hat{u}^{(\lambda)}, \hat{v}^{(\lambda)})$ as the fitted value of $(d, \boldsymbol{u}, \boldsymbol{v})$ with the regularization parameter being λ. Define BIC as:

$$BIC(\lambda) = log(SSE(\lambda)) + \frac{log(nq)}{nq} df(\lambda)$$

where $SSE(\lambda) = ||Y - \hat{d}^{(\lambda)} X \hat{u}^{(\lambda)} \hat{v}^{(\lambda)}||^2$, $\hat{df}(\lambda) = \sum_{i=1}^{p} I(\hat{u_i^{(\lambda)}} \neq 0) + \sum_{j=1}^{q} I(\hat{v_j^{(\lambda)}} \neq 0) - 1$

- Exclusive extraction algorithm(EEA)

This idea is to seek a $\hat{C}$ with sparse SVD structure near some initial consistent estimator, e.g. the least squares reduced rank regression estimator $\tilde{C}$ whose SVD is given by $\sum_{k=1}^{r^*} \tilde{d}_k \tilde{u}_k \tilde{v}_k' = \sum_{k=1}^{r^*} \tilde{C}_k$

- Exclusive extraction algorithm(EEA)

This idea is to seek a $\hat{C}$ with sparse SVD structure near some initial consistent estimator, e.g. the least squares reduced rank regression estimator $\tilde{C}$ whose SVD is given by
$\sum_{k=1}^{r^*} \tilde{d}_k \tilde{u}_k \tilde{v}_k = \sum_{k=1}^{r^*} \tilde{C}_k$
The EEA is as follows:

(a). for each $k \in \{1, \cdots, r^*\}$

 (1). construct the adaptive weights
 $w_K^{(d)} = |\tilde{d}_k|^{-\gamma}, w_k^{(u)} = |\tilde{u}_k|^{-\gamma}$ and $w_k^{(v)} = |\tilde{v}_k|^{-\gamma}$

 (2). construct the exclusive layer $Y_k = Y - X(\tilde{C} - \tilde{C}_k)$

 (3). find $(\hat{d}_k, \hat{u}_k, \hat{v}_k)$ by performing the sparse unit rank regression of Y_k on X with λ_k chosen by BIC

(b). The final estimator C is given by $\hat{C} = \sum_{k=1}^{r^*} \hat{d}_k \hat{u}_k \hat{v}_k$

Note

We can also have the method done iteratively called the iterative exclusive extraction algorithm e.g.

$$\frac{||C^{(\hat{i+1})} - \hat{C^{(i)}}||}{||\hat{C^{(i)}}}< \epsilon$$

for example $\epsilon = 10^{-6}$

$$\tilde{Y} = DWV' + \tilde{N} \tag{11}$$

where D is an $n \times n$ orthogonal matrix(using Schmidt orthogonalization), V is $p \times r$ unknown orthogonal matrix, $W \in \mathbb{R}^{n \times r}$

$$\tilde{Y} = DWV' + \tilde{N} \tag{11}$$

where D is an $n \times n$ orthogonal matrix(using Schmidt orthogonalization), V is $p \times r$ unknown orthogonal matrix, $W \in \mathbb{R}^{n \times r}$

The optimization problem is as follows:

$$J(V, W) = \frac{1}{2}||\tilde{Y} - DWV'||^2 + \sum_i \lambda_i ||w_{(i)}||_1 \tag{12}$$

$$(\hat{V}, \hat{W}) = arg\min_{V,W} J(V, W) \quad such \quad that \quad V'V = I_r \tag{13}$$

(1). W-step:Given a fixed V, the problem can be written as

$$arg\min_{W} \frac{1}{2}||B - W||^2 + \sum_i \lambda_i ||w_{(i)}||_1 \tag{14}$$

where $B = D'\tilde{Y}V$, and the solution is

$$\hat{w}_{ji} = max(0, |b_{ji}| - \lambda_i)\frac{b_{ij}}{|b_{ij}|} \tag{15}$$

(1). W-step:Given a fixed V, the problem can be written as

$$\arg\min_{W} \frac{1}{2}||B-W||^2 + \sum_i \lambda_i ||w_{(i)}||_1 \tag{14}$$

where $B = D'\tilde{Y}V$, and the solution is

$$\hat{w}_{ji} = max(0, |b_{ji}| - \lambda_i)\frac{b_{ij}}{|b_{ij}|} \tag{15}$$

(2). V-step: Given a fixed W

$$\arg\min_{V} ||\tilde{Y} - DWV'||^2 \quad such \quad that \quad V'V = I_r$$

which has a solution given by $\hat{V} = QG'$ where Q and G are computed using $M = \hat{W}'D'\tilde{Y} = Q\Sigma G'$

Using $\tilde{Y} = U_{\tilde{Y}} S_{\tilde{Y}} V'_{\tilde{Y}}$, and setting $V_0 = V_{\tilde{Y}}$, we can proceed the algorithm above.

Noting that (14) is separable and can be written as

$$arg \min_{w_{(i)}} \frac{1}{2} ||b_{(i)} - w_{(i)}||^2 + \lambda_i ||w_{(i)}||_1$$

Using $\tilde{Y} = U_{\tilde{Y}} S_{\tilde{Y}} V'_{\tilde{Y}}$, and setting $V_0 = V_{\tilde{Y}}$, we can proceed the algorithm above.
Noting that (14) is separable and can be written as

$$arg\min_{w_{(i)}} \frac{1}{2}||b_{(i)} - w_{(i)}||^2 + \lambda_i ||w_{(i)}||_1$$

The **SURE criterion** is given by

$$SURE(\lambda_i) = ||\hat{w_{(i)}} - b_{(i)}||^2 - n + 2n_i$$

where $n_i = \#\{j : |\hat{w}_{ji}| \geq \lambda_i\}$

The rank trace plot

The algorithm is from the journal:**Discovering genetic associations with high-dimensionality neuroimaging phenotypes a sparse reduced-rank regression approach**

The rank trace plot

The algorithm is from the journal:**Discovering genetic associations with high-dimensionality neuroimaging phenotypes a sparse reduced-rank regression approach**

The full rank coefficient matrix $\hat{C}_{(R)}$ and the estimated residual covariance matrix $\hat{S}_{\epsilon\epsilon(Y)} = (Y - X\hat{C}_{(r)})'(Y - X\hat{C}_{(r)})$.

The algorithm is from the journal:**Discovering genetic associations with high-dimensionality neuroimaging phenotypes a sparse reduced-rank regression approach**

The full rank coefficient matrix $\hat{C}_{(R)}$ and the estimated residual covariance matrix $\hat{S}_{\epsilon\epsilon(Y)} = (Y - X\hat{C}_{(r)})'(Y - X\hat{C}_{(r)})$.The rank trace plotting is obtained by plotting, for all values of r in a range from 0 to R, the following 2 quantities:

$$\Delta\hat{C}_{(r)} = \frac{||\hat{C}_{(R)} - \hat{C}_{(r)}||}{||\hat{C}_{(R)} - \hat{C}_{(0)}||} \qquad \Delta\hat{S}_{\epsilon\epsilon(r)} = \frac{||\hat{S}_{\epsilon\epsilon(R)} - \hat{S}_{\epsilon\epsilon(r)}||}{||\hat{S}_{\epsilon\epsilon(R)} - \hat{S}_{\epsilon\epsilon(0)}||}$$

The rank trace plot

The algorithm is from the journal:**Discovering genetic associations with high-dimensionality neuroimaging phenotypes a sparse reduced-rank regression approach**

The full rank coefficient matrix $\hat{C_{(R)}}$ and the estimated residual covariance matrix $\hat{S_{\epsilon\epsilon(Y)}} = (Y - X\hat{C_{(r)}})'(Y - X\hat{C_{(r)}})$.The rank trace plotting is obtained by plotting, for all values of r in a range from 0 to R, the following 2 quantities:

$$\Delta\hat{C_{(r)}} = \frac{||\hat{C_{(R)}} - \hat{C_{(r)}}||}{||\hat{C_{(R)}} - \hat{C_{(0)}}||} \qquad \Delta\hat{S_{\epsilon\epsilon(r)}} = \frac{||\hat{S_{\epsilon\epsilon(R)}} - \hat{S_{\epsilon\epsilon(r)}}||}{||\hat{S_{\epsilon\epsilon(R)}} - \hat{S_{\epsilon\epsilon(0)}}||}$$

As r varies from 0 to R in both X and Y axes, both coefficients take values in $[0, 1]$. As more ranks are added, starting at the top-right corner with $r = 0$, the curve moves towards the origin of the plot. When a further rank addition doesn't produce a **significant reduction** the plot indicates an "optimal" rank R^* which has been found.

谢谢!