Title
Regularization Methods for Canonical Correlation Analysis, Rank Correlation Matrices and Renyi Correlation Matrices

Author
Xu, Ying

Publication Date
2011

Peer reviewed|Thesis/dissertation

Regularization Methods for Canonical Correlation Analysis, Rank Correlation Matrices and Renyi Correlation Matrices

by

Ying Xu

A dissertation submitted in partial satisfaction of the
requirements for the degree of
Doctor of Philosophy

in

Statistics

in the

Graduate Division

of the

University of California, Berkeley

Committee in charge:

Professor Peter J. Bickel, Chair
Professor Haiyan Huang
Professor Laurent EL Ghaoui

Fall 2011

Regularization Methods for Canonical Correlation Analysis, Rank Correlation Matrices and Renyi Correlation Matrices

Abstract

Regularization Methods for Canonical Correlation Analysis, Rank Correlation Matrices and Renyi Correlation Matrices

by

Ying Xu

Doctor of Philosophy in Statistics

University of California, Berkeley

Professor Peter J. Bickel, Chair

In multivariate analysis, canonical correlation analysis is a method that enable us to gain insight into the relationships between the two sets of variables. It determines linear combinations of variables of each type with maximal correlation between the two linear combinations. However, in high dimensional data analysis, insufficient sample size may lead to computational problems, inconsistent estimates of parameters. In Chapter 1, three new methods of regularization are presented to improve the traditional CCA estimator in high dimensional settings. Theoretical results have been derived and the methods are evaluated using simulated data.

While the linear methods are successful in many circumstances, it certainly has some limitations, especially in cases where strong nonlinear dependencies exist. In Chapter 2, I investigate some other measures of dependence, including the rank correlation and its extensions, which can capture some non-linear relationship between variables. Finally the Renyi correlation is considered in Chapter 3. I also complement my analysis with simulations that demonstrate the theoretical results.

Key words and phrases. High dimension, canonical correlation analysis, banding, thresholding, tapering, l_1 regularized, rank correlation, M-correlation, Renyi correlation, convergence rate.

Contents

Contents **i**

List of Figures **ii**

List of Tables **iii**

1 Canonical Correlation Analysis in High dimension **1**
1.1 Introduction 1
1.2 Problem Set-up 4
1.3 Banding Method 5
1.4 Thresholding Method 8
1.5 l_1 Regularized CCA 10
1.6 Simulation 14

2 Rank Correlation Matrices and Extensions **19**
2.1 Introduction 19
2.2 Measure of Dependence 19
2.3 Spearman's Rank Correlation 20
2.4 Extending to Multivariate Cases 21
2.5 Regularization methods in High-dimensional Cases 22
2.6 Extension of Rank Correlation 34
2.7 Theoretical Results on M-correlation Matrices 35
2.8 Apply the Results to CCA Problems 42
2.9 Simulation 43

3 Renyi Correlation **49**
3.1 Introduction 49
3.2 Discussion of: Brownian distance covariance 49
3.3 Extension of canonical correlation analysis to non-linear cases 52

Bibliography **56**

List of Figures

1.1 Correlation matrices and cross-correlation matrix structure of Example 1 15
1.2 Correlation matrices and cross-correlation matrix structure of Example 2 17

3.1 Relationship between X and Y and between $\hat{f}$ and $\hat{g}$ 52

List of Tables

1.1 Results for CCA example 1 16
1.2 Results for CCA example 2 18

2.1 Results for banded rank correlation matrices 44
2.2 Results for thresholded rank correlation matrices 44
2.3 Results for banded M-correlation matrices 45
2.4 Results for thresholded M-correlation matrices 46
2.5 Results for banded normal score transformed correlation matrices 46
2.6 Results for thresholded normal score transformed correlation matrices 47
2.7 Results for the mixture model: the sample rank correlation matrix 47
2.8 Results for the mixture model: Banding 48
2.9 Results for the mixture model: Thresholding 48

3.1 Results of $K - L$ correlation 52

Acknowledgments

I would like to thank everybody who made this thesis possible and contributed to its development.

First of all I would like to thank my thesis advisor, Professor Peter Bickel for his tremendous help in developing this thesis. Without his expertise, involvement, sound advice and constant support this work would not have been possible.

I would like to thank my thesis committee members, Professor Haiyan Huang and Professor Laurent EL Ghaoui for their helpful comments, stimulating discussions and questions .

I am also grateful to Professor Bin Yu and Professor Ching-Shui Cheng for taking time out of the busy schedule to help.

I would like to thank fellow students and friends at Berkeley.

Last but not least, I would like to thank my parents and my husband, Luqiao Liu for their patience and support throughout my studies.

Chapter 1

Canonical Correlation Analysis in High dimension

1.1 Introduction

Canonical correlation analysis (CCA) [18] is a method of correlating linear relationships between two multidimensional variables. By using this method, we can usually gain insight into the relationships between the two sets of variables.

The goal of this method is to find basis vectors for two sets of variables such that the correlations between the projections of the variables onto these basis vectors are mutually maximized.

Given two zero-mean random variables $X \in \mathbb{R}^{p_1}$ and $Y \in \mathbb{R}^{p_2}$, CCA finds pairs of directions w_x and w_y that maximize the correlation between the projections $u = w_x'X$ and $v = w_y'Y$ (in the context of CCA, the projections u and v are also referred to as canonical variates). More formally, CCA maximizes the function:

$$\rho = \frac{Cov(u, v)}{\sqrt{Var(u)Var(v)}} = \frac{\mathbb{E}(w_x'XY'w_y)}{\sqrt{\mathbb{E}(w_x'XX'w_x)\mathbb{E}(w_y'YY'w_y)}}$$

CCA has certain maximal properties that are very similar to the principle component analysis (PCA), which have been studied intensively. However, the biggest difference between these two methods lies in that CCA considers the relationship between two sets of variables, while PCA focuses on the interrelationship within only one set of variables,

CCA can be utilized for many purposes. One of the main applications of CCA is to integrate multiple data sets that are related to the same subject. For example, in [36], the

authors utilizes CCA to integrate multiple fMRI data sets in the context of predictive fMRI data analysis.

Another main application of CCA is to reduce the dimensionality of the data by extracting a small (compared to the superficial dimensionality of the data) number of linear features, thus alleviating subsequent computations. Previously, PCA has been extensively utilized to achieve this. However, one has to bear in mind that the goal of PCA is to minimize the reconstruction error; and in particular, PCA-features might not be well suited for regression tasks. Consider a mapping $\phi : X \to Y$, there is no reason to believe that the features extracted by PCA on the variable x will reflect the functional relation between x and y in any way. Even worse, it is possible that information vital to establishing this relation is discarded when projecting the original data onto the PCA-feature space. CCA, along with partial least squares (PLS) and multivariate linear regression (MLR) may be better suited for regression tasks since it considers the relation between explanatory variables and response variables. CCA, in particular, has some very attractive properties, for example, it is invariant w.r.t. affine transformations, and thus scaling of the input variables.

CCA can be used for purposes far beyond those listed above. It's a prominent method whenever we need to establish a relation between two sets of measurements, such as learning a semantic representation between web images and their associated text [17], studying on the association of gene expression with multiple phenotypic or genotypic measures [31], etc.

Standard solutions to CCA

The CCA problem can be formulated as a optimization problem,

$$\begin{aligned} &\max_{w_x, w_y} \mathbb{E}(w_x' X Y' w_y) \\ &\text{s.t. } \mathbb{E}(w_x' X X' w_x) = \mathbb{E}(w_y' Y Y' w_y) = 1 \end{aligned} \tag{1.1}$$

This optimization problem then can be converted into a standard eigen-problem. We adapted the results from [17] and [26] (chapter 10).

- Corresponding Lagrangian of (1.1) is

$$L(\lambda, w_x, w_y) = w_x' \Sigma_{xy} w_y - \frac{\lambda_x}{2}(w_x' \Sigma_{xx} w_x - 1) - \frac{\lambda_y}{2}(w_y' \Sigma_{yy} w_y - 1)$$

- Taking derivatives in respect to w_x and w_y, and let them equal 0,

$$\Sigma_{xy} w_y - \lambda_x \Sigma_{xx} w_x = 0 \qquad \Sigma_{yx} w_x - \lambda_y \Sigma_{yy} w_y = 0$$

- $$\Rightarrow \; w_x' \Sigma_{xy} w_y - \lambda_x w_x' \Sigma_{xx} w_x = w_y' \Sigma_{yx} w_x - \lambda_y w_y' \Sigma_{yy} w_y = 0$$

  together with constrains,

$$\Rightarrow \; \lambda_x = \lambda_y$$

- Assume Σ_{yy} is invertible, we have $w_y = \Sigma_{yy}^{-1}\Sigma_{yx}w_x/\lambda$. Finally, we can see that the solution of (1.1) is to equivalent to $\Sigma_{xy}\Sigma_{yy}^{-1}\Sigma_{yx}w_x = \lambda^2\Sigma_{xx}w_x$, which is a generalized eigen-problem.

If Σ_{xx} is also invertible, the vectors w_x and w_y are the eigenvectors corresponding to the largest eigenvalue of the matrices

$$\Sigma_{xx}^{-1}\Sigma_{xy}\Sigma_{yy}^{-1}\Sigma_{yx}, \qquad \Sigma_{yy}^{-1}\Sigma_{yx}\Sigma_{xx}^{-1}\Sigma_{xy}. \tag{1.2}$$

Both of the above matrices have the same positive eigenvalues and the largest one equals the first canonical correlation.

In practice, one typically computes the sample covariance matrix out of the data set $(x_1, y_1), \cdots, (x_n, y_n) \in \mathbb{R}^{p_1} \times \mathbb{R}^{p_2}$. Computing eigenvectors/values of the empirical counterparts of the matrices results then immediately in estimates of the canonical variates and correlations.

Potential problems in the high-dimensional setting

Nowadays, people face some serious problems when dealing with the high dimensional data sets using traditional methods. Major advances in network and measurement technologies allow people to collect, store, analyze, and transport massive amounts of data. Applications in various domains often lead to very high dimensional data, for example, gene arrays, fMRI, image processing, various kinds of spectroscopy, climate studies, $\cdots$. In many situations, the data dimension p is comparable to or even larger than the sample size n.

In the meantime, recent advances in random matrix theory allowed an in-depth theoretical study on the traditional estimator - the sample (empirical) covariance matrix. And it is shown that without regularization the traditional estimator performs poorly in high dimensions. The empirical eigenvectors and eigenvalues are inconsistent in terms of estimating the corresponding population quantities, if $p/n \to c$, $0 < c \leq \infty$ (Wigner (1955), Wachter (1978), Johnstone (2001), Johnstone & Lu (2006), Paul (2005), Bair et al. (2006), Tracy and Widom (1996) and etc.).

It is generally believed that it is almost impossible to give estimation in the high-dimensional settings without additional structural constraints. The sparsity assumption is one of most popular remedies for this situation. The notion of sparsity is often referred to as that only a few variables have large effects on the quantities that people are interested in, while most of the others are negligible. This is often a reasonable assumption in many applications and is now widespread in high-dimensional statistical inference. These structural constraints not only make estimation feasible, but also may enhance the interpretability of the estimators.

Sparse CCA has also been proposed in recent studies. Parkhomenko et al. (2007) [30], Waaijenborg et al. (2008)[44], Parkhomenko et al. (2009) [31], Le Cao et al. (2008) [23], and Witten et al. (2009)[45] and many others have proposed methods for penalized CCA and utilized the methods in various areas. Although their regularization methods and algorithms may differ from each other, they share one thing in common, which is that they all put the sparsity assumption on the coefficients of canonical variates, i.e. w_x and w_y.

This assumption is certainly reasonable in many applications, but one can not tell from the data itself whether the assumption can be applicable or not. Instead, we propose to add sparsity assumptions directly onto the covariance matrices and the cross-covariance matrices, which can be easily checked from the data.

In the following sections, we'll present three methods of regularization to improve the traditional CCA estimator in high dimensional settings.

1.2 Problem Set-up

We will assume throughout this paper that we observe $X_1, \cdots, X_n \in \mathbb{R}^{p_1}$ i.i.d. random variables with mean 0 and covariance matrix Σ_{11}; as well as $Y_1, \cdots, Y_n \in \mathbb{R}^{p_2}$ i.i.d. random variables with mean 0 and covariance matrix Σ_{22}. And we write

$$X_i = (X_{i1}, \cdots, X_{ip_1})^\top, \quad Y_i = (Y_{i1}, \cdots, Y_{ip_2})^\top.$$

Furthermore, $\mathbb{E}(X_i^\top Y_i) = \Sigma_{12}$.

It's well known that the usual MLE of $\Sigma_{\cdot\cdot}$'s are

$$\begin{aligned}
\hat{\Sigma}_{11} &= \frac{1}{n}\sum_{i=1}^{n}(X_i - \bar{X})(X_i - \bar{X})^\top \\
\hat{\Sigma}_{22} &= \frac{1}{n}\sum_{i=1}^{n}(Y_i - \bar{Y})(Y_i - \bar{Y})^\top \\
\hat{\Sigma}_{12} &= \frac{1}{n}\sum_{i=1}^{n}(X_i - \bar{X})(Y_i - \bar{Y})^\top
\end{aligned}$$

Notations:

- Denote $\Omega_i = \Sigma_{ii}^{-1}$, $i = 1, 2$ to be the true concentration matrix.
- For any matrix $M = [m_{ij}]$, write $M^+ = \mathrm{diag}(M)$ for a diagonal matrix with the same diagonal as M, and $M^- = M - M^+$.

- $|\cdot|_1$ for the l_1 norm of a vector or matrix vectorized
- Define the operator norm of a $n \times p$ matrix M as

$$\|M\| = sup\{\|Mx\|_2 : x \in \mathbb{R}^p,\ \|x\|_2 = 1\},$$

  which for symmetric matrices reduces to $\|M\| = \max_i |\lambda_i(M)|$.
- Denote $p = p_1 + p_2$.

1.3 Banding Method

First, we consider the situations where there exits a natural metric on the index set. For instance, we can expect that large $|i - j|$ implies near independence or conditional (given the intervening indexes) independence of $X_{\cdot i}$ and $X_{\cdot j}$. This is a reasonable assumption in many applications, such as climatology and spectroscopy, and it will be especially plausible for the studies involving the time series.

Banding the empirical covariance matrices of high-dimension has been well studied. In this section, we will adapt the results of Bickel and Levina [4] and extend it to the CCA problem.

Preliminary

Let's restate the banding operator in [4] first. For any matrix $M = [m_{ij}]_{p\times p}$, and any $0 \le k < p$, define,

$$B_k(M) = [m_{ij}1(|i - j| \le k)].$$

As shown in [4], we'll also define a banded-approximatable matrices class. Since the CCA problem involves the cross-covariance matrices, which in general are not symmetric any more, we need to modify the original definition in [4] slightly:

$$\begin{aligned}\mathcal{U}(\epsilon_0, \alpha, c) = \ \{\ \ & \Sigma : \max_j \sum_i \{|\sigma_{ij}| :\ |i - j| > k\} \le ck^{-\alpha} \text{ for all } k > 0; \\ & \text{and } 0 < \epsilon_0 \le \lambda_{\min}^{1/2}(\Sigma^\top\Sigma) \le \lambda_{\max}^{1/2}(\Sigma^\top\Sigma) \le 1/\epsilon_0\}.\end{aligned}$$

According to [4], by either banding the sample covariance matrix or estimating a banded version of the inverse population covariance matrix, we can obtain estimates which are consistent at various rates in the operator norm as long as $\log p/n \to 0$.

Main results

This regularization method contains two steps:

- Banding the sample covariance matrices and cross-covariance matrices;
- Plug-in the banded sample covariance matrices and cross-covariance matrices into (1.2) and solve for the eigenvectors and eigenvalues.

In the following, we'll show that the regularized estimator of (1.2) is consistent under the operator norm, under the suitable assumptions.

Theorem 1.3.1. *Suppose that X and Y are Gaussian distributed; $\mathcal{U}(\epsilon_0, \alpha, c)$ is defined above. Then, if $k_i \asymp (n^{-1}\log p_i)^{-1/(2(\alpha_i+1))}$ $(i = 1, 2, 3)$, where $p_3 := \min(p_1, p_2)$,*

$$\begin{aligned}&\|B_{k_1}(\hat{\Sigma}_{11})^{-1}B_{k_3}(\hat{\Sigma}_{12})B_{k_2}(\hat{\Sigma}_{22})^{-1}B_{k_3}(\hat{\Sigma}_{21}) - \Omega_1\Sigma_{12}\Omega_2\Sigma_{21}\| \\ =&O_P\Big((\frac{\log p_m}{n})^{\alpha/(2(\alpha+1))}\Big)\end{aligned}$$

uniformly on $\Sigma_{11} \in \mathcal{U}(\epsilon_1, \alpha_1, c_1)$, $\Sigma_{12} \in \mathcal{U}(\epsilon_3, \alpha_3, c_3)$, $\Sigma_{22} \in \mathcal{U}(\epsilon_2, \alpha_2, c_2)$. Here, $\alpha := \max_i \alpha_i$, $p_m := \max_i p_i$.

Proof. :

According to [4] Theorem 1, under the assumptions given, we have

$$\begin{aligned}\|B_{k_1}(\hat{\Sigma}_{11})^{-1} - \Omega_1\| &= O_P\Big((\frac{\log p_1}{n})^{\alpha_1/(2(\alpha_1+1))}\Big); \\ \|B_{k_2}(\hat{\Sigma}_{22})^{-1} - \Omega_2\| &= O_P\Big((\frac{\log p_2}{n})^{\alpha_2/(2(\alpha_2+1))}\Big).\end{aligned}$$

For Σ_{12}, one can not apply their results directly, but from the proof of their theorem, one can easily go through each step for the cross-covariance matrices. Therefore,

$$\|B_{k_3}(\hat{\Sigma}_{12}) - \Sigma_{12}\| = O_P\Big((\frac{\log p_3}{n})^{\alpha_3/(2(\alpha_3+1))}\Big).$$

Thus, we have

$$\begin{aligned}
&\|B_{k_1}(\hat{\Sigma}_{11})^{-1}B_{k_3}(\hat{\Sigma}_{12})B_{k_2}(\hat{\Sigma}_{22})^{-1}B_{k_3}(\hat{\Sigma}_{21}) - \Omega_1\Sigma_{12}\Omega_2\Sigma_{21}\| \\
\leq&\|B_{k_1}(\hat{\Sigma}_{11})^{-1}B_{k_3}(\hat{\Sigma}_{12})B_{k_2}(\hat{\Sigma}_{22})^{-1}B_{k_3}(\hat{\Sigma}_{21}) - B_{k_1}(\hat{\Sigma}_{11})^{-1}B_{k_3}(\hat{\Sigma}_{12})\Omega_2\Sigma_{21}\| \\
&\quad + \|B_{k_1}(\hat{\Sigma}_{11})^{-1}B_{k_3}(\hat{\Sigma}_{12})\Omega_2\Sigma_{21} - \Omega_1\Sigma_{12}\Omega_2\Sigma_{21}\| \\
\leq&\|B_{k_1}(\hat{\Sigma}_{11})^{-1}B_{k_3}(\hat{\Sigma}_{12})\|\left(\|B_{k_2}(\hat{\Sigma}_{22})^{-1} - \Omega_2\|\|B_{k_3}(\hat{\Sigma}_{12})^\top\| + \|\Omega_2\|\|B_{k_3}(\hat{\Sigma}_{12})^\top - \Sigma_{12}^\top\|\right) \\
&\quad + \|\Omega_2\Sigma_{12}^\top\|\left(\|B_{k_1}(\hat{\Sigma}_{11})^{-1} - \Omega_1\|\|B_{k_3}(\hat{\Sigma_{12}})\| + \|\Omega_1\|\|B_{k_3}(\hat{\Sigma_{12}}) - \Sigma_{12}\|\right) \\
=&\|B_{k_1}(\hat{\Sigma}_{11})^{-1}\|\|B_{k_3}(\hat{\Sigma}_{12})\|^2\|B_{k_2}(\hat{\Sigma}_{22})^{-1} - \Omega_2\| \\
&\quad + \left(\|B_{k_1}(\hat{\Sigma}_{11})^{-1}\|\|B_{k_3}(\hat{\Sigma}_{12})\| + \|\Omega_1\|\|\Sigma_{12}\|\right)\|\Omega_2\|\|B_{k_3}(\hat{\Sigma}_{12}) - \Sigma_{12}\| \\
&\quad + \|B_{k_3}(\hat{\Sigma}_{12})\|\|\Sigma_{12}\|\|\Omega_2\|\|B_{k_1}(\hat{\Sigma}_{11})^{-1} - \Omega_1\| \\
=&O_P(\epsilon_1^{-1}\epsilon_3^{-2})O_P\Big((\frac{\log p_2}{n})^{\alpha_2/(2(\alpha_2+1))}\Big) + O_P(\epsilon_1^{-1}\epsilon_2^{-1}\epsilon_3^{-1})O_P\Big((\frac{\log p_3}{n})^{\alpha_3/(2(\alpha_3+1))}\Big) \\
&\quad + O_P(\epsilon_2^{-1}\epsilon_3^{-2})O_P\Big((\frac{\log p_1}{n})^{\alpha_1/(2(\alpha_1+1))}\Big)
\end{aligned}$$

□

Remarks:

1. The Gaussian assumption may be weakened, which has been pointed out in [4]. It can be replaced by the following. Suppose $Z_i := \binom{X_i}{Y_i} \in \mathbb{R}^{p_1+p_2}$ are i.i.d., $Z_{1j} \sim F_j$, where F_j is the c.d.f. of Z_{1j}, and $G_j(t) = F_j(\sqrt{t}) - F_j(-\sqrt{t})$; Then for Theorem 1.3.1 to hold it suffices to assume that

$$\max_{1\leq j\leq p_1}\int_0^\infty \exp(\lambda t)dG_j(t) < \infty, \quad \text{for } 0 < |\lambda| < \lambda_0.$$

   for some λ_0.

2. Cai et.al. [7] points out that the convergence rate obtained by the banding method can be improved by a very similar regularized estimator , called the tapering estimator. If we use the tapering estimator for $\hat{Sigma}_{ij}$, then under the same assumption of Theorem 1.3.1, it can obtain the rate of

$$O_P(\max\{n^{-\frac{\alpha}{2\alpha+1}} + \sqrt{\frac{\log p_m}{n}}, \sqrt{\frac{p_m}{n}}\}).$$

1.4 Thresholding Method

Although the banding method can be applied in many situations, it will not be applicable if the labels are meaningless as in microarrays genomics, where the coordinates simply label genes. To deal with this, here we propose using the second method, of Bickel-Levina [3] and El Karoui [12], the thresholding method, which can solve such difficulties in certain conditions.

We'll use direct thresholding method for $\hat{\Sigma_{12}}$ as in [3] and the SPICE method as in [35] for $\hat{\Sigma_{11}}^{-1}$ and $\hat{\Sigma_{22}}^{-1}$.

$\hat{\Sigma_{12}}$

According to Bickel and Levina (2008) [3], We define the thresholding operator by

$$T_s(M) = [m_{ij}1(|m_{ij}| \geq s)]$$

which we refer to as M thresholded at s.

Next, we define a uniformity class of covariance matrices

$$\mathcal{U}_\tau(q, c_0(p), M) = \big\{\Sigma : \sigma_{ii} \leq M, \sum_{j=1}^{p} |\sigma_{ij}|^q \leq c_o(p) \text{ for all } i\big\},$$

for $0 \leq q < 1$. And, if $q = 0$,

$$\mathcal{U}_\tau(0, c_0(p), M) = \big\{\Sigma : \sigma_{ii} \leq M, \sum_{j=1}^{p} 1(\sigma_{ij} \neq 0) \leq c_o(p) \text{ for all } i\big\}$$

defines a class of sparse matrices.

Then, uniformly on $\mathcal{U}_\tau(q, c_0(p), M)$, for sufficiently large M', if $t_n = M'\sqrt{\frac{\log p}{n}}$, and $\frac{\log p}{n} = o(1)$, then

$$\|T_{t_n}(\hat{\Sigma}) - \Sigma\| = O_P\left(c_0(p)(\frac{\log p}{n})^{\frac{1-q}{2}}\right)$$

$\Longrightarrow$

$$\|T_{t_n}(\hat{\Sigma_{12}}) - \Sigma_{12}\| \leq \|T_{t_n}(\hat{\Sigma}) - \Sigma\| = O_P\left(c_0(p)(\frac{\log p}{n})^{\frac{1-q}{2}}\right) \tag{1.3}$$

Further more, we know from their results that $\|\Sigma_{12}\|$ and $\|T_{t_n}(\hat{\Sigma_{12}})\|$ are bounded, $O_P(c_o(p))$.

$\hat{\Sigma_{11}}^{-1}$ and $\hat{\Sigma_{22}}^{-1}$

Based on SPICE method by A.J. Rothman et al. (2008)[35], we make additional assumptions on Ω_1 and Ω_2.

A1: Let the set $S_k = \{(i,j) : \Omega_{kij} \neq 0, i \neq j\}$, $k = 1, 2$. Then $\text{card}(S_k) \leq s_k$.

A2: $\phi_{min}(\Sigma_{ii}) \geq \underline{k} > 0$, $i = 1, 2$.

Since we already assume that $\Sigma \in \mathcal{U}_\tau(q, c_0(p), M)$, $\phi_{max}(\Sigma_{ii})$ $(i = 1, 2)$ are bounded without additional assumptions.

And we use the correlation matrix rather than the covariance matrix. Let W_i be the diagonal matrix of true standard deviations; $\Gamma_i = W_i^{-1}\Sigma_{ii}W_i^{-1}$, $i = 1, 2$.

Let $K_i = \Gamma_i^{-1}$. Define a SPICE estimate of K_i by

$$\hat{K}_{\lambda_i} = \text{argmim}_{K \succ 0}\{\text{tr}(K\hat{\Gamma}_i) - \log|K| + \lambda_i|K^-|_1\}$$

Then we can define a modified correlation-based estimator of the concentration matrix by

$$\tilde{\Omega}_{\lambda_i} = \hat{W_i}^{-1}\hat{K}_{\lambda_i}\hat{W_i}^{-1}, \quad i = 1, 2$$

Under A1, A2, uniformly on $\mathcal{U}_\tau(q, c_0(p), M)$, if $\lambda_i \asymp \sqrt{\frac{\log p_i}{n}}$, then

$$\|\tilde{\Omega_{\lambda_i}} - \Omega_i\| = O_P\left(\sqrt{\frac{(s_i+1)\log p_i}{n}}\right), \quad i = 1, 2 \tag{1.4}$$

Main Results

Theorem 1.4.1. *Suppose X and Y follow Gaussian distributions. Under A1, A2, uniformly on $\Sigma \in \mathcal{U}_\tau(q, c_0(p), M)$, for sufficiently large M', if $t_n = M'\sqrt{\frac{\log p}{n}}$, $\lambda_i \asymp \sqrt{\frac{\log p_i}{n}}$, and $\frac{\log p}{n} = o(1)$, then*

$$\begin{aligned}&\|\tilde{\Omega_{\lambda_1}}T_{t_n}(\hat{\Sigma_{12}})\tilde{\Omega_{\lambda_2}}T_{t_n}(\hat{\Sigma_{21}}) - \Omega_1\Sigma_{12}\Omega_2\Sigma_{21}\| \\ =&O_P(c_0(p)^2)O_P\left(\sqrt{\frac{(s_m+1)\log p_m}{n}}\right),\end{aligned}$$

where $s_m = \max\{s_1, s_2\}, p_m = \max\{p_1, p_2\}$.

Proof. :

$$
\begin{aligned}
&\|\tilde{\Omega_{\lambda_1}} T_{t_n}(\hat{\Sigma_{12}})\tilde{\Omega_{\lambda_2}} T_{t_n}(\hat{\Sigma_{21}}) - \Omega_1\Sigma_{12}\Omega_2\Sigma_{21}\| \\
\leq &\|\tilde{\Omega_{\lambda_1}} T_{t_n}(\hat{\Sigma_{12}})\tilde{\Omega_{\lambda_2}} T_{t_n}(\hat{\Sigma_{21}}) - \tilde{\Omega_{\lambda_1}} T_{t_n}(\hat{\Sigma_{12}})\Omega_2\Sigma_{21}\| \\
&\quad + \|\tilde{\Omega_{\lambda_1}} T_{t_n}(\hat{\Sigma_{12}})\tilde{\Omega_{\lambda_2}} T_{t_n}(\hat{\Sigma_{21}}) - \Omega_1\Sigma_{12}\Omega_2\Sigma_{21}\| \\
\leq &\|\tilde{\Omega_{\lambda_1}} T_{t_n}(\hat{\Sigma_{12}})\| \left(\|\tilde{\Omega_{\lambda_2}} - \Omega_2\|\|T_{t_n}(\hat{\Sigma_{12}})^\top\| + \|\Omega_2\|\|T_{t_n}(\hat{\Sigma_{12}})^\top - \Sigma_{12}^\top\|\right) \\
&\quad + \|\Omega_2\Sigma_{12}^\top\| \left(\|\tilde{\Omega_{\lambda_1}} - \Omega_1\|\|T_{t_n}(\hat{\Sigma_{12}})\| + \|\Omega_1\|\|T_{t_n}(\hat{\Sigma_{12}}) - \Sigma_{12}\|\right) \\
= &\|\tilde{\Omega_{\lambda_1}}\|\|T_{t_n}(\hat{\Sigma_{12}})\|^2\|\tilde{\Omega_{\lambda_2}} - \Omega_2\| \\
&\quad + \left(\|\tilde{\Omega_{\lambda_1}}\|\|T_{t_n}(\hat{\Sigma_{12}})\| + \|\Omega_1\|\|\Sigma_{12}\|\right)\|\Omega_2\|\|T_{t_n}(\hat{\Sigma_{12}}) - \Sigma_{12}\| \\
&\quad + \|T_{t_n}(\hat{\Sigma_{12}})\|\|\Sigma_{12}\|\|\Omega_2\|\|\tilde{\Omega_{\lambda_1}} - \Omega_1\| \\
= &O_P(c_0(p)^2)O_P\left(\sqrt{\frac{(s_1+1)\log p_1}{n}}\right) + O_P(c_0(p))O_P\left(c_0(p)(\frac{\log p}{n})^{\frac{1-q}{2}}\right) \\
&\quad + O_P(c_0(p)^2)O_P\left(\sqrt{\frac{(s_2+1)\log p_2}{n}}\right)
\end{aligned}
$$

□

1.5 l_1 Regularized CCA

In the first two approaches, the matrices are considered separately. Thus, the sparsity assumptions are required for each matrix. In this section, we'll consider a different approach to the CCA problem.

Notice that the products $\Sigma_{11}^{-1}\Sigma_{12}$ and $\Sigma_{22}^{-1}\Sigma_{21}$ are related to the linear regression problem Y on X and X on Y respectively. So it is natural to ask: can we use the regularization methods of the linear regression to regularize the products of the matrices directly?

More precisely, let's consider the following problems. Denote $A^* = \Sigma_{11}^{-1}\Sigma_{12}$ and $B^* = \Sigma_{22}^{-1}\Sigma_{21}$. Consider two multiple responses linear models:

$$Y = XA + \epsilon$$

and

$$X = YB + \eta,$$

where $X \in \mathbb{R}^{p_1}$ and $Y \in \mathbb{R}^{p_2}$ are random variables we considered in the original CCA problem. Then, theoretically, $A^* = argmin_A E\|Y - XA\|^2$ and $B^* = argmin_A E\|X - YB\|^2$.

Therefore, the CCA problem can be solved in two steps:

- solve the optimization problems for the linear models;
- solve the eigenvalue problem for the matrix $\hat{A}^*\hat{B}^*$.

Previous work

For low dimensional cases, the ordinary least square (OLS) estimator $\hat{A}_{ols} = \hat{\Sigma}_{11}^{-1}\hat{\Sigma}_{12}$ would be a good estimator of A^*. However, as pointed out earlier, this wouldn't be the case in high dimensional settings. Recently, quite a lot of attentions has been given to regularization methods in high dimensional linear regression. Among the popular regularization methods, a great amount of work has focused on the behavior of l_1-based relaxations. And most of the work focus on the single response linear regression model. A variety of practical algorithms have been proposed and studied, including basis pursuit[10], the Lasso [40], and the Dantzig selector [8]. Various authors have obtained convergence rates for different error metrics, including l_2-error [27], prediction loss [21] [43], as well as model selection consistency [28] . In addition, a range of sparsity assumptions have been analyzed, including the case of hard sparsity or soft sparsity assumptions, based on imposing a certain decay rate on the ordered entries of coefficients.In the following, we'll apply the related results from the linear regression to our CCA problem.

Existing works on l_1 penalized CCA include the double barrelled lasso which is based on a convex least squares approach [16], and CCA as a sparse solution to the generalized eigenvalue problem [37] which is based on constraining the cardinality of the solution to the generalized eigenvalue problem to obtain a sparse solution. Another recent solution is by Witten et al. (2009)[45], which uses a penalized matrix decomposition framework to compute a sparse solution of CCA. However, all the above methods are imposing the sparse condition on the coefficients of canonical variates, which are different from ours assumptions.

In the CCA problem, what we're interested most is the l_2 loss of the estimator of the coefficients. The restricted eigenvalue (RE) conditions introduced by Bickel et al. (2009) [5] are among the weakest and hence the most general conditions in literature imposed on the

Gram matrix in order to guarantee nice statistical properties for the Lasso and the Dantzig selector. Moreover, under this condition, they derived bounds on l_2 and l_r prediction loss, where $1 \leq r \leq 2$, for estimating the parameters of both the Lasso and the Dantzig selector in both linear regression and nonparametric regression models.

We'll consider the following linear model in this subsection,

$$y = X\beta^* + w \tag{1.5}$$

where X is an $n \times p$ matrix and each row follows i.i.d. $N(0, \Sigma)$ and w is independent of X, $\sim N(0, \sigma^2 I_n)$. Then the lasso estimator of β^* is defined by

$$\hat{\beta} = argmin_{\beta \in \mathbb{R}^p} \{\frac{1}{n}|y - X\beta|_2^2 + 2\lambda|\beta|_1\}.$$

We adapt the notation v_{J_0} to always represent the subvector of $v \in \mathbb{R}^p$ confined to J_0, which corresponds to the locations of the s largest coefficients of v in absolute value.

Now, we are ready to introduce the Restricted Eigenvalue assumption that is formalized in Bickel et al. (2009) .

Assumption 5.1. (Restricted Eigenvalue assumption $RE(s, k_0, X)$ (Bickel et al., 2009)) For some integer $1 \leq s \leq p$ and a positive number k_0, the following holds:

$$\frac{1}{K(s, k_0, X)} := \min_{J_0 \subseteq \{1, \cdots, p\}, |J_0| \leq s} \quad \min_{v \neq 0, \|v_{J_0^c}\|_1 \leq k_0 \|v_{J_0}\|_1} \frac{\|Xv\|_2}{\sqrt{n}\|v_{J_0}\|_2} > 0. \tag{1.6}$$

However, the above RE condition cannot apply to our problem directly, since what we have is a random design matrix X, not fixed. In the paper of Zhou (2009) [47], a detailed discussion of RE conditions on random design matrices had been given, which we're going to apply to the CCA problem.

Assumption 5.2.(Restricted eigenvalue assumption $RE(s, k_0, \Sigma)$) Suppose $\Sigma_{jj} = 1, j = 1, \cdots, p$, and for some integer $1 \leq s \leq p$, and a positive number k_0, the following condition holds:

$$\frac{1}{K(s, k_0, \Sigma)} := \min_{J_0 \subseteq \{1, \cdots, p\}, |J_0| \leq s} \quad \min_{v \neq 0, \|v_{J_0^c}\|_1 \leq k_0 \|v_{J_0}\|_1} \frac{\|\Sigma^{1/2} v\|_2}{\|v_{J_0}\|_2} > 0. \tag{1.7}$$

Zhou shows that if Σ satisfies the RE condition, X will satisfy the RE condition with overwhelming probability, given n that is sufficiently large.

Define

$$\sqrt{\rho_{\min}(m)} := \min_{\|t\|_2=1, |supp(t)|\leq m} \|\Sigma^{1/2}t\|_2,$$

$$\sqrt{\rho_{\max}(m)} := \max_{\|t\|_2=1, |supp(t)|\leq m} \|\Sigma^{1/2}t\|_2.$$

Summarizing the results in the paper [47], we have that

Theorem 1.5.1. *Set $1 \leq n \leq p$. Let $s < p/2$. Consider the linear model (1.5) with random design X , and Σ satisfies (1.7). Assume $\rho_{\min}(2s) > 0$. Let $\hat{\beta}$ be an optimal solution to the Lasso with $\lambda_n = O_p(\sqrt{\log p/n}) = 0(1)$. Suppose that n satisfies for $C = 3(2+k_0)K(s,k_0,\Sigma)\sqrt{\rho_{\max}(s)}$,*

$$n > c' \max(C^2 s \log(5ep/s), 9 \log p).$$

Then with probability at least $1 - 2\exp(-c\theta^2 n)$, we have

$$\|\hat{\beta} - \beta\|_2 \leq 8cK^2(s,3,\Sigma)\lambda_n\sqrt{s} \tag{1.8}$$

CCA

We return to the CCA problem. We propose the following procedure,

- Let $Y_{(i)}$ denote the ith column of Y, $i = 1, \cdots, p_2$; and $A^*_{(i)}$ be the ith column vector of A^*.
  - Decompose the multiple response linear model into p_2 individual single linear models. Consider
    $$Y_{(i)} = XA^*_{(i)} + \epsilon_{(i)}, \qquad i = 1, \cdots, p_2.$$
  - Find the lasso estimator of $A^*_{(i)}$, denoting it as $\hat{A}^*_{(i)}$.
- Let $X_{(i)}$ denote the ith column of X, $i = 1, \cdots, p_1$; and $B^*_{(i)}$ be the ith column vector of B^*.
  - Decompose the multiple response linear model into p_1 individual single linear models. Consider
    $$X_{(i)} = YB^*_{(i)} + \eta_{(i)}, \qquad i = 1, \cdots, p_1.$$
  - Find the lasso estimator of $B^*_{(i)}$, denoting it as $\hat{B}^*_{(i)}$.
- Solve the eigen-problem for $\hat{A}^*\hat{B}^*$.

Then if Σ_{11}, Σ_{22} satisfies the RE conditions in theorem 5.1. , we have that

$$\|\hat{A}^* - A^*\|_F^2 = O_p\left(\frac{\log p_1}{n}(\sum_{i=1}^{p_2} s_i K^4(s_i, 3, \Sigma_{11}))\right)$$

and

$$\|\hat{B}^* - B^*\|_F^2 = O_p\left(\frac{\log p_2}{n}(\sum_{i=1}^{p_1} t_i K^4(t_i, 3, \Sigma_{22}))\right)$$

Let $s := \max_i s_i$ and $t := \max_i t_i$. Then,

$$\sum_{i=1}^{p_2} s_i K^4(s_i, 3, \Sigma_{11}) \leq s p_2 K^4(s, 3, \Sigma_{11}), \quad \sum_{i=1}^{p_1} t_i K^4(t_i, 3, \Sigma_{22}) \leq t p_1 K^4(t, 3, \Sigma_{22}),$$

Therefore,

$$\|\hat{A}^*\hat{B}^* - A^*B^*\|_F^2 = O_p\left(\frac{s p_2 \log p_1}{n} K^4(s, 3, \Sigma_{11}) + \frac{t p_1 \log p_2}{n} K^4(t, 3, \Sigma_{22})\right)$$

If in fact some of the s_i's and t_i's are just 0, then we may have that $\sum_{i=1}^{p_2} s_i << p_2$ and $\sum_{i=1}^{p_1} t_i << p_1$, which will give a better rate.

1.6 Simulation

The dependency between two sets of variables can be modeled using latent variables. Using this idea for simulation of CCA model can be found in [31]. Suppose there exist latent variables U that affect both a subset of observed variables in X and a subset of observed variables in Y. Formally, we have that, for $X, \epsilon \in \mathbb{R}^{p_1}$, $Y, \eta \in \mathbb{R}^{p_2}$, $U \in \mathbb{R}^w$, $A \in \mathbb{R}^{p_1} \times \mathbb{R}^w$ and $B \in \mathbb{R}^{p_2} \times \mathbb{R}^w$,

$$X = AU + \epsilon,$$
$$Y = BU + \eta.$$

where U, ϵ and η are mutually independent.

Let $U \sim N(o, \Sigma)$, $\epsilon \sim N(0, \sigma_1^2 I)$ and $\eta \sim N(0, \sigma_2^2 I)$. Then we have that,

$$\Sigma_{11} = A\Sigma A^\top + \sigma_1^2 I$$
$$\Sigma_{22} = B\Sigma B^\top + \sigma_2^2 I$$
$$\Sigma_{12} = A\Sigma B^\top$$

For simplicity, we'll set $\Sigma = I$ in the following simulations. By imposing different structures to A and B, and setting different values to σ_1^2 and σ_2^2, the model covers a lot of situations.

Example 1

Let A, B be banded matrices with bandwidths k_1 and k_2 respectively. Then Σ_{11}, Σ_{22} and Σ_{12} as defined above are all banded matrices with bandwidths $2k_1$, $2k_2$, and $k_1 + k_2$. The Figure 1.1 illustrates this setting . To be specific, we set A and B be

$$a_{ij} = \rho_1^{|i-j|} I_{\{|i-j| \le k_1\}}, \quad i = 1, \cdots, p_1; \ j = 1, \cdots, w.$$
$$b_{ij} = \rho_2^{|i-j|} I_{\{|i-j| \le k_2\}}, \quad i = 1, \cdots, p_2; \ j = 1, \cdots, w.$$

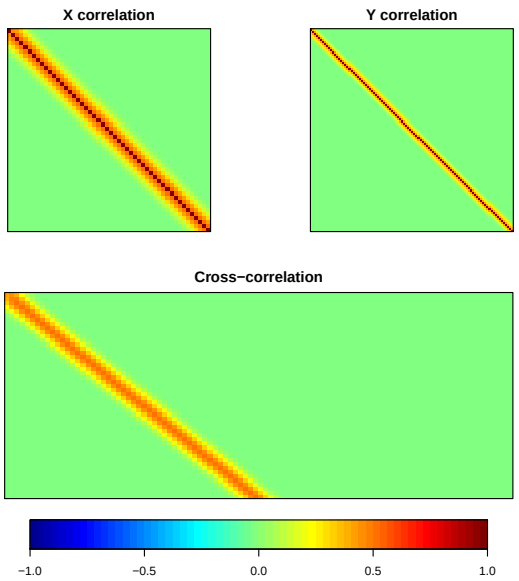

Figure 1.1: Correlation matrices and cross-correlation matrix structure of Example 1

The results provided are under the settings of $p_1 = 50$, $p_2 = 100$, $n = 100$, $k_1 = 3, k_2 = 2, \rho_1 = 0.7, \rho_2 = 0.6$, $\sigma_1 = \sigma_2 = 2$. The true canonical correlation is 0.84. With 100 replications, we have the following results,

Mehtod	Estimated canonical correlation	Frobenius norm	$\lvert\hat{\lambda}_{max} - \lambda_{max}\rvert$
Sample	1.00(0.00)	6.15(0.00)	0.16(0.00)
Banding	0.79(0.15)	2.88(0.12)	0.08(0.13)
Lasso	0.73(0.14)	2.96(0.16)	0.13(0.12)

Table 1.1: Results for CCA example 1

We can see that without regularization, the empirical estimator performs very badly under this setting. And since the underlying structure of covariance matrices of example 1 are banded, the banding method gets the best performance.

Example 2

In this example, we adopt the simplest case under this model scheme, which is that there is only one latent variable. Then A and B become vectors. And assume that $U \sim N(0,1)$, $\epsilon \sim N(0,I)$, $\eta \sim N(0,I)$. In this case, we have

$$\Sigma_{11} = AA^\top + I \;\; \Sigma_{22} = BB^\top + I \;\; \Sigma_{12} = AB^\top$$

Then CCA has a simple solution,

$$\begin{aligned} w_x &= \lambda_{max}(AA^\top) \\ w_y &= \lambda_{max}(BB^\top) \\ cor(w_x^\top X, w_y^\top Y) &= \frac{A^\top AB^\top B}{(1+A^\top A)(1+B^\top B)} \end{aligned}$$

By varying the norm of vector A and B, the canonical correlation can obtain any value in $(0,1)$.

To demonstrate our methods, we further set A and B be sparse. Let the first s_1 components of A be non-zero, and the first s_2 components of B be non-zero. Then, Σ_{11}, Σ_{12} and Σ_{22} are all banded matrices. Thus the banding and thresholding methods can be applied.

Also we have that

$$X = \frac{1}{B'B}AB'Y + (\epsilon - \frac{1}{B'B}AB'\eta)$$
$$Y = \frac{1}{A'A}BA'X + (\eta - \frac{1}{A'A}BA'\epsilon)$$

where both AB' and BA' are sparse matrices. Therefore, the lasso method can also be utilized here.

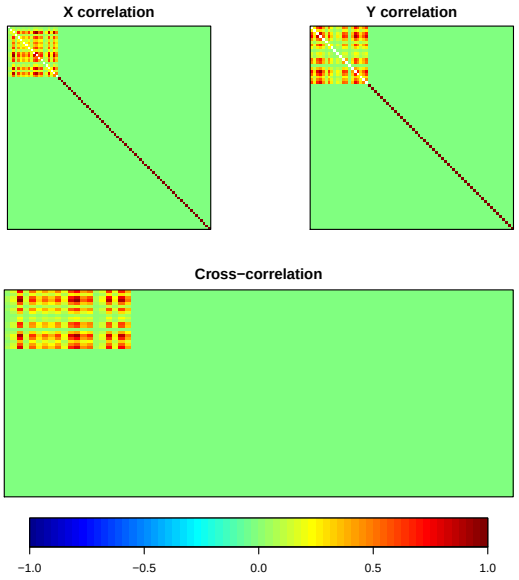

Figure 1.2: Correlation matrices and cross-correlation matrix structure of Example 2

The results provided are under the settings of $p_1 = 80$, $p_2 = 70$, $n = 100$ and $s_1 = s_2 = 20$. The first 20 components of A and B are generated from $U(0, 1/3)$. The true canonical correlation is 0.68 . With 100 replications, the results are shown in Table 1.2.

As expected, without regularization the empirical estimator performs very badly under this setting. The Lasso method performs slightly better in this case, since the model construction in this example is slightly in favor of Lasso.

Mehtod	Estimated canonical correlation	Frobenius norm	$\lvert\hat{\lambda}_{max} - \lambda_{max}\rvert$
Sample	1.00(0.00)	10.20(0.00)	0.32(0.00)
Banding	0.54(0.16)	0.75(0.19)	0.16(0.08)
Lasso	0.59(0.14)	0.46(0.17)	0.12(0.07)

Table 1.2: Results for CCA example 2

Chapter 2

Rank Correlation Matrices and Extensions

2.1 Introduction

In the precious chapter, we focus on the CCA method, which are using the covariance matrices and cross-covariance matrix to describe the linear relationship between the two sets of random variables. Despite its successful application in many circumstances, this can be a modeling limitation, in cases where strong nonlinear dependencies exist. In this chapter, we'll study some other measures of dependence, including the rank correlation and its extensions, which enable us to capture some non-linear relationship between variables.

2.2 Measure of Dependence

The study of dependence plays an important role in statistics. Over the past years, people have developed sophisticated theories in this field. The theory and history is rich, and easily forms a separate volume, e.g. Doruet, Mari and Kotz (2001) [11].

The most widely used and understood measure in the study of dependence is the product-moment correlation coefficient, which was first invented by Francis Galton in 1885, and then refined into its modern form by Karl Pearson in 1895 [33]. The product-moment correlation is a measure of linear dependence between random variables. It attains the maximum magnitude of 1 if and only if a liner relationship exits between random variables. Also the definition of the product-moment correlation implies that it is defined only when the variances of the random variables are finite. Thus, it is not an appropriate measure for very heavy-tailed distributions.

In the statistical literature, there are many alternative approaches, for example, Spearman's rank correlation, Kendall's rank correlation, the distance correlation of Szekely et.al.

(2007) [39], etc. Here we first focus on Spearman's rank correlation, which was introduced by Spearman (1904) and some generalization.

As a different way to describe the dependence of random variables, in its population version Spearman's rank correlation is actually nothing but the linear correlation between the transformed version of the original random variables by their cumulative distribution functions, which always exist.

Rank correlation is also a robust way to describe the dependence structure among the variables. Usually, severe outliers can dramatically change the sample covariance estimator, but the sample rank correlation matrix is much less affected. We also notice that, to estimate the covariance matrix, one requires the random variable to have at least second moments, while one does not need any moments if the rank correlation is used.

2.3 Spearman's Rank Correlation

Given two random variables X and Y, the classic Pearson's correlation coefficient between them is defined as $\rho_{X,Y} = \frac{\mathrm{Cov}(X,Y)}{\sqrt{\mathrm{Var}(X)\mathrm{Var}(Y)}}$; while the Spearman's rank correlation is defined as $r(X,Y) = \frac{\mathrm{Cov}(F(X),G(Y))}{\sqrt{\mathrm{Var}(F(X))\mathrm{Var}(G(Y))}}$, where F and G are the cumulative distribution functions of X and Y respectively. It can be seen from the above equations that the population version of Spearman's rank correlation is just the classic Pearson's correlation between $F(X)$ and $G(Y)$. In the following, we'll present some useful facts about the Spearman's rank correlation, which one can find in detail in the 3rd Chapter of the book Uncertainty analysis with high dimensional dependence modelling [22].

Proposition. If X is a random variable with a continuous invertible cumulative distribution function F, then $F(X)$ has the uniform distribution on $[0,1]$, denoted by $U[0,1]$.

Thus, the rank correlation is independent of marginal distributions in the continuous case and the formula can be simplified as

$$r(X,Y) = \frac{\mathbb{E}[F(X)G(Y)] - (1/2)^2}{(1/12)} = 12\mathbb{E}[F(X)G(Y)] - 3 \tag{2.1}$$

Therefore, Spearman's correlation coefficient is often described as being "nonparametric". For continuous random variables, their exact sampling distribution can be obtained without requiring knowledge of marginal distributions of the data coordinates..

Further, the rank correlation is invariant under non-linear strictly increasing transformations.

Proposition. If $H : R \to R$ is a strictly increasing function, then,

$$r(X, Y) = r(H(X), Y).$$

A perfect Spearman correlation can be obtained when X and Y are related by any monotonic function. And in contrast, the Pearson correlation only gives a perfect value when X and Y are related by a linear function.

In the Gaussian case, the Spearman's rank correlation r is an increasing function of the ordinary correlation coefficient ρ,

$$\rho = 2\sin(\frac{\pi}{6}r).$$

And $r = 0 \Leftrightarrow \rho = 0$, $r = \pm 1 \Leftrightarrow \rho = \pm 1$.

In practice, given n pairs of i.i.d. samples $\{x_i, y_i\}$, denote $R(x_i)$ to be the rank of x_i among n samples, similar for $R(y_i)$. Then as illustrated in many textbooks, e.g. [11], the empirical Spearman's rank correlation can be computed as following,

$$\hat{r} = \frac{n(\sum R(x_i)R(y_i)) - (\sum R(x_i))(\sum R(y_i))}{\sqrt{n(\sum R(x_i)^2) - (\sum R(x_i))^2}\ \sqrt{n(\sum R(y_i)^2) - (\sum R(y_i))^2}}.$$

For the purpose of theory development, we'd like to introduce following notation,

$$P(f) := \mathbb{E}f(X),\ P_n(f) := \frac{1}{n}\sum_{i=1}^{n} f(X_i),\ \forall f.$$

Then under the assumptions that X and Y have continuous marginal c.d.f. F and G, we can rewrite $\hat{\rho}$ as

$$\begin{aligned}\hat{\rho} =& \frac{P_n(\hat{F}\hat{G}) - P_n(\hat{F})P_n(\hat{G})}{\sqrt{P_n(\hat{F}^2) - (P_n(\hat{F}))^2}\ \sqrt{P_n(\hat{G}^2) - (P_n(\hat{G}))^2}} \\ =& [P_n(\hat{F}\hat{G}) - (\frac{n+1}{2n})^2]/(\frac{n^2-1}{12n^2}) \\ =& \frac{12n^2}{n^2-1}P_n(\hat{F}\hat{G}) - \frac{3(n+1)}{n-1}\end{aligned}$$

where $\hat{F}$ and $\hat{G}$ are empirical c.d.f. ($\hat{F}(x) = \frac{1}{n}\sum_{i=1}^{n} 1_{[x_i,\infty)}(x)$).

2.4 Extending to Multivariate Cases

Spearman's rank correlation often appears as the measurement between only two variables. We'll first extend the rank correlation idea to multi-variable cases. The correlations between

random vectors X are collectively represented by the correlation matrix. The $(i,j)th$ element of the matrix represents the correlation between components X_i and X_j. In parallel, we define the rank correlation matrix of the random vector X, where the $(i,j)th$ element of the matrix represents the rank correlation between components X_i and X_j.

In the multi-dimensional settings, we consider a random sample $\{X_1, X_2, \cdots, X_n\}$ i.i.d., $X_i = (X_{i1}, \cdots, X_{ip})^T \in \mathbb{R}^p$, and assume it has continuous marginal cumulative distribution function (c.d.f.) $F_{1,} \cdots, F_p$.

Using $R = \big(r_{ij}\big)$ denote the rank correlation matrix, which is $\in \mathbb{R}^{p\times p}$, with

$$r_{ij} = 12P(F_iF_j) - 3.$$

Let $\hat{R} = \big(\hat{r}_{ij}\big)$ denote the empirical rank correlation matrix, $\in \mathbb{R}^{p\times p}$, with

$$\hat{r}_{ij} = \frac{12n^2}{n^2-1} P_n(\hat{F}_i\hat{F}_j) - \frac{3(n+1)}{n-1},$$

where $\hat{F}_j(x) = \frac{1}{n}\sum_{i=1}^n 1_{[X_{ij},\infty)}(x)$, which is the empirical c.d.f.'s.

In addition to the above, we'd like to introduce more matrix notations which will appear in the following subsections :

$$R^0 := \Big(P(F_iF_j)\Big)_{i,j},$$
$$\tilde{R^0} := \Big(P_n(F_iF_j)\Big)_{i,j},$$
$$\hat{R^0} := \Big(P_n(\hat{F}_i\hat{F}_j)\Big)_{i,j}.$$

Simply notice that,

$$\hat{R} - R = 12(\hat{R^0} - R^0) - \frac{6}{n-1}11^T + \frac{12}{n^2-1}\hat{R^0}, \tag{2.2}$$

where $1 = (1, \cdots, 1)^T \in \mathbb{R}^p$.

2.5 Regularization methods in High-dimensional Cases

In many situations, the data dimension p is comparable to or larger than the sample size n. Developments in random matrix theory, Jonestone [19], Johnstone & Lu [20] , Paul [32], Bair et al. (2006), Tracy and Widom [41] etc., have made it clear that the sample (empirical) covariance matrix may have a very poor performance in high dimension problems. Therefore

it's obvious that the M-correlation matrices will have the same difficulty when the dimension is high.

Many methods of regularization have been proposed for covariance matrices. Ledoit and Wolf [24], replaced the sample covariance with its linear combination with the identity matrix. Furrer and Bengtsson [15], Bickel and Levina proposed banding [4] and thresholding [3] methods. Yuan and Lin[46], Banerjee et al. [2] and Friedman et al. [14] developed different algorithms for estimators based on regularized maximum likelihood using an l_1 constraint on the entries of sample covariance matrices.

In this section we'll present three regularized version of $\hat{R}$ when the dimension of p is large and we will show also that all three versions will give consistent results.

The results of the convergence rate of estimators are under the matrix L_2 norm, also called the operator norm,

$$\|M\| := \sup\{\|Mx\| : \|x\| = 1\} = \sqrt{\lambda_{\max}(M^\top M)}$$

which for symmetric matrices reduces to $\|M\| = \max_i |\lambda_i(M)|$.

In addition to the operator norm, we also use the matrix infinity norm, defined as

$$\|M\|_\infty := \max_{i,j} |m_{ij}|;$$

and the matrix $(1,1)$ norm as

$$\|M\|_{(1,1)} := \max_j \sum_i |m_{ij}|.$$

Preliminaries

Before introducing our regularized estimators of rank correlation matrices, we'll give out two basic, but useful lemmas first.

Lemma 2.5.1. $\mathbb{P}(\ |P_n(F_iF_j) - P(F_iF_j)|\ \geq t) \leq 2\exp(-\frac{nt^2}{8})$

Proof. Notice that $\{F_i(X_{li})F_j(X_{lj}) - P(F_iF_j)\}$, $l = 1, \cdots, n$ are i.i.d. random varibles with mean 0.

Moreover, $|F(X_{li})F_j(X_{lj}) - P(F_iF_j)| \leq 2$.

So we can use the Hoeffding's inequality and the desired result follows. □

Lemma 2.5.2.

$$\|\tilde{R}^0 - R^0\|_\infty = O_P(\sqrt{\frac{\log p}{n}})$$

Proof. Using the results in Lemma 2.5.1,

$$\begin{aligned}
\mathbb{P}(\|\tilde{R}^0 - R^0\|_\infty \geq t) &= 1 - \mathbb{P}(\max_{i,j} |P_n(F_iF_j) - P(F_iF_j)| < t) \\
&= 1 - \prod_{i,j} \mathbb{P}(|P_n(F_iF_j) - P(F_iF_j)| < t) \\
&\leq 1 - \prod_{i,j} (1 - 2\exp(-\frac{nt^2}{8})) \\
&\leq 2p^2 \exp(-\frac{nt^2}{8})
\end{aligned}$$

Let $t = O(\sqrt{\frac{\log p}{n}})$, we have

$$\|\tilde{R}^0 - R^0\|_\infty = O_P(\sqrt{\frac{\log p}{n}}) \tag{2.3}$$

□

Further, in order to get the consistency results, we also need to know the convergence rate of the empirical c.d.f.'s. There is a well known inequality in the literature of asymptotic statistics.

Dvoretzky-Kiefer-Wolfowitz inequality: Given a natural number n, let $X_1, X_2, \cdots, X_n$ be real-valued, independent and identically-distributed random variables with distribution function F. Let $\hat{F}_n$ denote the associated empirical distribution function defined by $\hat{F}_n(x) = \frac{1}{n}\sum_{i=1}^n 1_{[X_i,\infty)}(x)$. Then, $\forall\ t > 0$

$$\mathbb{P}(\sup_{x\in\mathcal{R}} |\hat{F}_n(x) - F(x)| \geq t) \leq 2\exp(-2nt^2)$$

Hence, the convergence rate of the difference between the empirical distribution function and the population c.d.f is

$$\|\hat{F}_n - F\|_\infty = O_P(1/\sqrt{n})$$

Now, we can use this result to bound the difference between $\hat{R}^0$ and $\tilde{R}^0$ and get some intuitions for the final results.

$$\mathbb{P}(|P_n(\hat{F}_i\hat{F}_j) - P_n(F_iF_j)| \geq t)$$

$$=\mathbb{P}(\Big|\frac{1}{n}\sum_{l=1}^{n}\hat{F}_i(X_{li})\hat{F}_j(X_{lj}) - \frac{1}{n}\sum_{l=1}^{n}F_i(X_{li})F_j(X_{lj})\Big| \geq t)$$

$$\leq\mathbb{P}(\frac{1}{n}\sum_{l=1}^{n}\big|\hat{F}_i(X_{li}) - F_i(X_{li})\big|\hat{F}_j(X_{lj}) \geq \frac{t}{2}) + \mathbb{P}(\frac{1}{n}\sum_{l=1}^{n}\big|\hat{F}_j(X_{lj}) - F_j(X_{lj})\big|F_i(X_{li}) \geq \frac{t}{2})$$

$$\leq\mathbb{P}(\|\hat{F}_i - F_i\|_\infty\frac{1}{n}\sum_{l=1}^{n}\hat{F}_j(X_{lj}) \geq \frac{t}{2}) + \mathbb{P}(\|\hat{F}_j - F_j\|_\infty\frac{1}{n}\sum_{l=1}^{n}F_i(X_{li}) \geq \frac{t}{2})$$

$$\leq\mathbb{P}(\|\hat{F}_i - F_i\|_\infty\frac{n(n+1)}{2n^2} \geq \frac{t}{2}) + \mathbb{P}(\|\hat{F}_j - F_j\|_\infty\frac{1}{n}\sum_{l=1}^{n}\big[\hat{F}_i(X_{li}) + |\hat{F}_i(X_{li}) - F_i(X_{li})|\big] \geq \frac{t}{2})$$

$$\leq 2\exp(-2n^3t^2/(n+1)^2) + \mathbb{P}(\|\hat{F}_j - F_j\|_\infty\frac{n(n+1)}{2n^2} \geq \frac{t}{2}) + \mathbb{P}(\|\hat{F}_i - F_i\|_\infty\|\hat{F}_j - F_j\|_\infty \geq \frac{t}{2})$$

$$\leq 4\exp(-2n^3t^2/(n+1)^2) + (2\exp(-2nt/2))^2$$

Based on this inequality, with arguments similar to those used for (2.3), we have

$$\|\hat{R}^0 - \tilde{R}^0\|_\infty = O_P(\sqrt{\frac{\log p}{n}}) \tag{2.4}$$

Combining equations (2.3) and (2.4), desired result follows:

$$\|\hat{R}^0 - R^0\|_\infty \leq \|\hat{R}^0 - \tilde{R}^0\|_\infty + \|\tilde{R}^0 - R^0\|_\infty = O_P(\sqrt{\frac{\log p}{n}}) \tag{2.5}$$

Banding the rank correlation matrix

As introduced in chapter 1, banding, which is well studied by Bickel and Levina [4], is proved to be as an effective method of regularization in high dimensional settings. In this subsection, we'll apply this method to the rank correlation matrix and derive the consistency results for it.

First, we define a similar matrices class as of [4]

$$\mathcal{U}(\alpha, c) = \{R : \max_j \sum_i \{|r_{ij}| : \ |i-j| > k\} \leq ck^{-\alpha} \text{ for all } k > 0; \ |r_{ij}| \leq 1\}.$$

We'll show that banded sample rank correlation matrices give the consistent estimates under the operator norm in this class of matrices.

Restate the banding operator in [4] first. For any matrix $M = [m_{ij}]_{p\times p}$, and any $0 \leq k < p$, define,

$$B_k(M) = [m_{ij} 1(|i-j| \leq k)].$$

Theorem 2.5.1. *If* $k \asymp (n^{-1}\log p)^{-1/(2(\alpha+1))}$,

$$\|B_k(\hat{R}) - R\| = O_P\Big((\frac{\log p}{n})^{\alpha/(2(\alpha+1))}\Big)$$

uniformly on $R \in \mathcal{U}(\alpha, c)$.

Proof. Notice that,

$$\|B_k(\hat{R}) - R\| \leq \|B_k(\hat{R}) - B_k(R)\| + \|B_k(R) - R\| \tag{2.6}$$

Thus, we can deal the two parts separately. The second part on the right hand side has a simple bound:

$$\|B_k(R) - R\| \leq \|B_k(R) - R\|_{(1,1)} \leq ck^{-\alpha} \tag{2.7}$$

While for the first part on the right hand side, we have the following argument.

By the equation (2.2),

$$\|B_k(\hat{R}) - B_k(R)\| \leq 12\|B_k(\hat{R}^0) - B_k(R^0)\| + \frac{6}{n-1}\|B_k(11^T)\| + \frac{12}{n^2-1}\|B_k(\hat{R}^0)\| \tag{2.8}$$

The second term on the right hand side is easy to bound:

$$\frac{6}{n-1}\|B_k(11^T)\| = \frac{6}{n-1}(2k+1) = O_P(\frac{k}{n})$$

To deal with the first two terms, we have to use some some basic facts on norms of matrices, and the results in the subsection above, equation (2.5), it's easy to see that,

$$\|B_k(\hat{R^0}) - B_k(R^0)\| = O_P(k\|B_k(\hat{R^0}) - B_k(R^0)\|_\infty) = O_P(k\sqrt{\frac{\log p}{n}}).$$

and

$$\|B_k(\hat{R^0})\| = O_P(k\|B_k(\hat{R})\|_\infty) = O_P(k)$$

Back to the inequality (2.8), we have

$$\begin{aligned}\|B_k(\hat{R}) - B_k(R)\| &\leq 12O_P(k\sqrt{\frac{\log p}{n}}) + O_P(\frac{k}{n}) + O_P(\frac{k}{n^2-1}) \\ &= O_P(k\sqrt{\frac{\log p}{n}})\end{aligned}$$

Finally, combining the results from (2.7), if set $k \asymp (n^{-1}\log p)^{-1/(2(\alpha+1))}$, the desired result follows. □

Tapering

A recent paper by Cai et.al. [7] proposed another kind of regularized estimator for the class of banded-approximatable matrices defined in [4] , which they call the tapering estimator and pointed out that it is the rate optimal over the banded-approximatable matrices. So in this section, we'll investigate the tapering estimators for rank correlation matrices, which will also yield the optimal rate of convergence over matrices class $\mathcal{U}(\alpha, c)$ as defined above.

Parallel to what's defined in [7], for a given integer $k \in [1, p]$, the tapering estimator of sample rank correlation matrix with parameter k is defined as following,

$$W_k(\hat{R}) := \left(w_{ij}\hat{r}_{ij}\right)_{p\times p}$$

where

$$w_{ij} = k^{-1}\{(2k - |i-j|)_+ - (k - |i-j|)_+\}.$$

We'll show as in Cai et.al. [7] for sample covariance matrix, a tapering estimator for sample rank correlation matrix with the optimal choice of k also attains the optimal rate of convergence under operator norm. Moreover, the optimal rate of convergence for estimating the rank correlation matrix under the operator norm is the same as for the covariance matrix estimation, which is showed in the following theorem.

Theorem 2.5.2. *Suppose $p \leq \exp(\gamma n)$ for some constant $\gamma > 0$. The minimax risk of estimating the rank correlation matrix R over the class $\mathcal{U}(\alpha, c)$ under the operator norm satisfies*

$$\inf_{\hat{R}} \sup_{\mathcal{U}(\alpha,c)} \mathbb{E}\|\hat{R} - R\|^2 \asymp \min\left\{ n^{-\frac{2\alpha}{2\alpha+1}} + \frac{\log p}{n}, \frac{p}{n} \right\}$$

We'll derive a minimax upper bound and a lower bound respectively in the following subsections.

Upper bound

The minimax upper bound is obtained by constructing a special class of tapering estimators.

Notice that $\tilde{R}^0$ follows the subgaussian distribution, so the theorem in [7] can be directly applied. So we only need to deal with the difference between $\tilde{R}^0$ and $\hat{R}^0$, which will be almost the same as in banding section.

Lower bound

We shall show that the following minimax lower bound holds.

Theorem 2.5.3. *Under the same assumptions as described in Theorem 2.5.2, the minimax risk satisfies*

$$\inf_{\hat{R}} \sup_{\mathcal{U}(\alpha,c)} \mathbb{E}\|\hat{R} - R\|^2 \geq c' n^{-\frac{2\alpha}{2\alpha+1}} + c' \frac{\log p}{n}.$$

The basic strategy underlying the proof is similar as to [7], to carefully construct a finite collection of multivariate normal distributions and calculate the total variation affinity between pairs of probability measures in the collection.

First, we'll show that any correlation matrix can be a rank correlation matrix, which will enable us to use part of the proof of [7].

As for rank correlation matrices ,they are just all correlation matrices of variables having $U(-0.5, 0.5)$ marginal (in continuous case), and let $\mathcal{M}$ denote the set of all correlation matrices of this type. Since uniform marginal is preserved for $aF_1 + (1-a)F_2$, where F_i is the underlying probability distribution of $M_1, M_2 \in \mathcal{M}$, it follows that $aM_1 + (1-a)M_2 \in \mathcal{M}$, which means that $\mathcal{M}$ is convex.

Now consider matrices of the form $[e_i e_j]/12$. This is covariance matrix of Ue where $U \sim U(-0.5, 0.5)$ and $e = (e_1, \cdots, e_p)^\top$. Therefore all covariance matrices of the form $\sum(a_k[e_i^k e_j^k])$ with $\sum(a_k) = 1$ are possible. By spectral theorem any covariance matrix can be written as c times such a matrix for some c. Therefore all correlation matrices are possible.

Set $k = n^{1/(2\alpha+1)}$ and $a = k^{-(\alpha+1)}$. Define the same matrices classes $\mathcal{F}_{11}$ as in [7],

$$\mathcal{F}_{11} = \big\{R(\theta) : I_p + \tau a \sum_{m=1}^{k} \theta_m B(m,k), \qquad \theta = (\theta_m) \in \{0,1\}^k\big\}$$

where $0 < \tau < 2^{-\alpha-1}c$ and $p \times p$ matrix $B(m,k) = (b_{ij})_{p\times p}$ with,

$$b_{ij} = I\{i = m \text{ and } m+1 \le j \le 2k, \text{ or } j = m \text{ and } m+1 \le i \le 2k\}.$$

It's easy to check that $\mathcal{F}_{11} \subseteq \mathcal{U}(\alpha, c)$.

And since any matrix in $\mathcal{F}_{11}$ is a correlation matrix, thus can be a rank correlation matrix. It immediately follows the proof in $\mathcal{F}_{11}$,

$$\inf_{\hat{R}} \sup_{\mathcal{F}_{11}} \mathbb{E}\|\hat{R} - R\|^2 \ge c' n^{-\frac{2\alpha}{2\alpha+1}}. \tag{2.9}$$

However, the ways of the construction the second matrix class is quite different. In paper of Cai et.al., it's constructed as

$$\{\Sigma_m : \Sigma_m = I + (\sqrt{\frac{\tau}{n}\log p_1} I_{\{i=j=m\}})_{p\times p}\},$$

which cannot be a correlation matrix. For any correlation matrix, the diagonal entries are always 1, thus, cannot make any changes. Therefore, a different construction should be used.

In addition to $\mathcal{F}_{11}$, we define

$$\mathcal{F}_{12} = \Big\{R_m : R_m = A - B(m),\ 0 \le m \le p_1, m \text{ odd}\Big\}$$

where $p_1 = \min\{p, e^n\}$ and the $p \times p$ matrix $A = (a_{ij})_{p\times p}$ with,

$$a_{ii} = 1, \quad a_{i,i+1} = a_{i+1,i} = 0.5 \text{ for } i \text{ is odd};$$

and define

$$B(m) = \Big(\sqrt{\frac{\tau}{n}\log p_1} I\{i = m, j = m+1\}; I\{i = m+1, j = m\}\Big).$$

We'll show that

$$\inf_{\hat{R}} \sup_{\mathcal{F}_{12}} \mathbb{E}\|\hat{R} - R\|^2 \ge c' \frac{\log p}{n} \tag{2.10}$$

for some constant c'. Combining with 2.9, the conclusion of Theorem 2.5.3 follows.

We now apply Le Cam's method to derive the lower bound (2.10).

First construct a parameter set. For $1 \leq m \leq p_1$ and m odd, let $R(m)$ be a block-diagonal matrix of block-size 2 with the $(m+1)/2$th block be

$$\begin{bmatrix} 1 & 0.5 - \sqrt{\frac{\tau}{n}\log p_1} \\ 0.5 - \sqrt{\frac{\tau}{n}\log p_1} & 1 \end{bmatrix}$$

and the rest of the blocks be $\begin{bmatrix} 1 & 0.5 \\ 0.5 & 1 \end{bmatrix}$. And let $R_0 = A$ as defined above.

Clearly $R(m)$ is positive definite, thus could define a multivariate normal distribution with $R(m)$ be its correlation matrix.

We restate the Le Cam's Lemma here to make our proof more complete.

Lemma 2.5.3. *Let T be an estimator of θ based on an observation from a distribution in the collection $\{P_\theta, \theta \in L = \{\theta_0, \cdots, \theta_{p_1}\}\}$, then*

$$sup_\theta \mathbb{E}L(T,\theta) \geq \frac{1}{2} r \min \|P_{\theta_0} \wedge \bar{P}\|.$$

First, we need to construct a parameter set. Let $X_1, \cdots, X_n \sim N(0, R(m))$, and denote the joint density by f_m.

Let $\theta_m = R(m)$ and L be the squared operator norm.

$$\begin{aligned} r(\theta_0, \theta_m) &:= inf_t[\|t-\theta_0\|^2 + \|t-\theta_m\|^2] \\ &= \|\frac{\theta_0+\theta_m}{2} - \theta_0\|^2 + \|\frac{\theta_0+\theta_m}{2} - \theta_m\|^2 \\ &= \frac{1}{2}\tau\frac{\log p_1}{n}. \end{aligned}$$

Thus, $r_{min} = inf_m r(\theta_0, \theta_m) = \frac{1}{2}\tau\frac{\log p_1}{n}$.

To apply the Le Cam's method, we also need to show that there exists a constant $c' > 0$ such that

$$\|\mathbb{P}_{\theta_0} \wedge \bar{\mathbb{P}}\| \geq c'.$$

where $\bar{\mathbb{P}} = \frac{1}{p_2}\sum_{l=1}^{p_2} \mathbb{P}_{\theta_{2l-1}}$. Here $p_2 = \lfloor \frac{p_1}{2} \rfloor$.

Shown by [7] , we only need to prove

$$\int \left(\frac{1}{p_2}\sum_{l=1}^{p_2} f_m\right)^2 / f_0 d\mu \to 1.$$

Denote $0.5 - \sqrt{\frac{\tau}{n}\log p_1}$ by b. Then,

$$\int \frac{f_m^2}{f_0}d\mu = \frac{(\sqrt{2\pi(1-b^2)})^{-2n}}{(\sqrt{2\pi(1-0.5^2)})^{-n}} \prod_{i=1}^{n} \int \exp\left[-\frac{2}{2}(x_m^i \; x_{m+1}^i)\begin{pmatrix} 1 & b \\ b & 1 \end{pmatrix}^{-1}\begin{pmatrix} x_m^i \\ x_{m+1}^i \end{pmatrix}\right]$$
$$\Big/ \exp\left[-\frac{1}{2}(x_m^i \; x_{m+1}^i)\begin{pmatrix} 1 & .5 \\ .5 & 1 \end{pmatrix}^{-1}\begin{pmatrix} x_m^i \\ x_{m+1}^i \end{pmatrix}\right] dx_m^i dx_{m+1}^i$$

$$= \left(\frac{(1-b^2)^2}{1-0.5^2}\right)^{-n/2} \left(\frac{(b^2-2b+3/2)^2 - (b^2-5b-1)^2/4}{(1-0.5^2)^2(1-b^2)^2}\right)^{-n/2}$$

$$= c_1[c_2(1-\tau\frac{\log p_1}{n}) + o(\frac{\log p_1}{n})]^{-n/2}$$

Thus,

$$\int \left(\frac{1}{p_2}\sum_{l=1}^{p_2} f_m\right)^2 / f_0 d\mu - 1$$

$$= \frac{1}{p_2^2}\sum_{l=1}^{p_2}\left(\int \frac{f_m^2}{f_0}d\mu - 1\right)$$

$$\leq \exp\left[-\log p_2 - \frac{n}{2}\log(1-\tau\frac{\log p_1}{n})\right] - \frac{1}{p_1} \to 0.$$

Thresholding

Parallel to the method of thresholding for the covariance matrices in the paper of Bickel and Levina [3], we also consider thresholding the rank correlation matrices.

Similarly, the results are uniform over families of matrices which satisfy a fairly natural notion of sparsity:

$$\mathcal{U}(q, c(p)) = \{R : \sum_{j=1}^{p} |r_{ij}|^q \leq c(p) \;\; \text{for all } i; \; |r_{ij}| \leq 1\}.$$

Theorem 2.5.4. *Uniformly on $\mathcal{U}(q, c(p))$, for sufficiently large M', if*

$$t_n = M'\sqrt{\frac{\log p}{n}},$$

and $\frac{\log p}{n} = o(1)$*, then*

$$\|T_{t_n}(\hat{R}) - R\| = O_P\Big(c(p)\big(\frac{\log p}{n}\big)^{\frac{1-q}{2}}\Big).$$

Proof.

$$\|R - T_t(\hat{R})\| \leq \|R - T_t(R)\| + \|T_t(R) - T_t(\hat{R})\|$$

With this inequality, we can deal the two parts separately. The first part on the right hand side has a simple bound:

$$\|R - T_t(R)\| \leq \max_i \sum_{j=1}^{p} |r_{ij}| I_{(|r_{ij}| \leq t)} \leq t^{1-q} c(p)$$

Let's focus on the second part. Notice the following inequality,

$$\|T_t(R) - T_t(\hat{R})\| \leq \|T_t(R) - T_t(\tilde{R})\| + \|T_t(\tilde{R}) - T_t(\hat{R})\|$$

Consider $\|T_t(R) - T_t(\tilde{R})\|$ first.

$$\begin{aligned}
\|T_t(R) - T_t(\tilde{R})\| \leq & \max_i \sum_{j=1}^{p} |\tilde{r}_{ij}| \\
\leq & \max_i \sum_{j=1}^{p} |\tilde{r}_{ij}| I_{(|\tilde{r}_{ij}| \geq t, |r_{ij}| < t)} && \dots (A) \\
& + \max_i \sum_{j=1}^{p} |r_{ij}| I_{(|\tilde{r}_{ij}| < t, |r_{ij}| \geq t)} && \dots (B) \\
& + \max_i \sum_{j=1}^{p} |\tilde{r}_{ij} - r_{ij}| I_{(|\tilde{r}_{ij}| \geq t, |r_{ij}| \geq t)} && \dots (C)
\end{aligned}$$

$$\begin{aligned}
(A) \leq & \max_i \sum_{j=1}^{p} |\tilde{r}_{ij} - r_{ij}| I_{(|\tilde{r}_{ij}| \geq t, |r_{ij}| < t)} + \max_i \sum_{j=1}^{p} |r_{ij}| I_{(|r_{ij}| < t)} \\
\leq & O_P\Big(c(p) t^{-q} \sqrt{\frac{\log p}{n}}\Big) + t^{1-q} c(p)
\end{aligned}$$

$$(C) \leq \max_{i,j} |\tilde{r}_{ij} - r_{ij}| \max_i \sum_{j=1}^{p} |r_{ij}|^q t^{-q} = O_p\left(c(p) t^{-q} \sqrt{\frac{\log p}{n}}\right)$$

For (B), we have,

$$\begin{aligned}(B) \leq & \max_i \sum_{j=1}^{p} (|\tilde{r}_{ij} - r_{ij}| + |\tilde{r}_{ij}|) I_{(|\tilde{r}_{ij}|<t, |r_{ij}|\geq t)} \\ \leq & \max_{i,j} |\tilde{r}_{ij} - r_{ij}| \sum_{j=1}^{p} I_{(|r_{ij}|\geq t)} + t \max_i \sum_{j=1}^{p} I_{(|r_{ij}|\geq t)} \\ = & O_P\Big(c(p) t^{-q} \sqrt{\frac{\log p}{n}} + t^{1-q} c(p)\Big)\end{aligned}$$

Now it's turn to deal with $\|T_t(\tilde{R}) - T_t(\hat{R})\|$.

$$\begin{aligned}\|T_t(\tilde{R}) - T_t(\hat{R})\| \leq & \max_i \sum_{j=1}^{p} |\hat{r}_{ij}| I_{(|\hat{r}_{ij}|\geq t, |r_{ij}|<t)} && \dots (D) \\ & + \max_i \sum_{j=1}^{p} |\tilde{r}_{ij}| I_{(|\hat{r}_{ij}|<t, |\tilde{r}_{ij}|\geq t)} && \dots (E) \\ & + \max_i \sum_{j=1}^{p} |\tilde{r}_{ij} - \hat{r}_{ij}| I_{(|\hat{r}_{ij}|\geq t, |\tilde{r}_{ij}|\geq t)} && \dots (F)\end{aligned}$$

Look at (F) first.

$$(F) \leq c\sqrt{\frac{\log p}{n}} \max_i \sum_{j=1}^{p} I_{(|\tilde{r}_{ij}|\geq t, |\hat{r}_{ij}|\geq t)} \leq c\sqrt{\frac{\log p}{n}} \max_i \sum_{j=1}^{p} I_{(|\tilde{r}_{ij}|\geq t)}$$

Notice that,

$$\begin{aligned}I_{(|\tilde{r}_{ij}|\geq t)} &= I_{(|\tilde{r}_{ij}|\geq t, |r_{ij} - \tilde{r}_{ij}|\leq \epsilon)} + I_{(|\tilde{r}_{ij}|\geq t, |r_{ij} - \tilde{r}_{ij}|> \epsilon)} \\ &\leq I_{(|r_{ij}|\geq t-\epsilon)} + I_{(|r_{ij} - \tilde{r}_{ij}|>\epsilon)}\end{aligned}$$

Define $N_i(\epsilon) := \sum_{j=1}^{p} I_{(|r_{ij} - \tilde{r}_{ij}|>\epsilon)}$.

$$\mathbb{P}(\max_i N_i(\epsilon) > 0) = \mathbb{P}(\max_{i,j} |r_{ij} - \tilde{r}_{ij}| > \epsilon) \leq 2p^2 \exp(-n\epsilon^2/8)$$

Thus, as long as $2\log p - n\epsilon^2/8 \to -\infty$,

$$\begin{aligned}(F) \leq & c\sqrt{\frac{\log p}{n}} \max_i N_i(\epsilon) + c\sqrt{\frac{\log p}{n}} c(p)(t-\epsilon)^{-q} \\ = & O_P\Big(c(p) t^{-q} \sqrt{\frac{\log p}{n}}\Big)\end{aligned}$$

For (D),

$$\begin{aligned}(D) &\leq \max_i \sum_{j=1}^{p} (|\tilde{r}_{ij} - \hat{r}_{ij}| + |\tilde{r}_{ij}|) I_{(|\tilde{r}_{ij}|<t, |\hat{r}_{ij}|\geq t)} \\ &\leq \max_{i,j} |\tilde{r}_{ij} - r_{ij}| \sum_{j=1}^{p} I_{(|\hat{r}_{ij}|\geq t)} + t \max_i \sum_{j=1}^{p} I_{(|\hat{r}_{ij}|\geq t)}\end{aligned}$$

$$\begin{aligned}I_{(|\hat{r}_{ij}|\geq t)} &= I_{(|\hat{r}_{ij}|\geq t, |\hat{r}_{ij}-\tilde{r}_{ij}|\leq \epsilon)} + I_{(|\hat{r}_{ij}|\geq t, |\hat{r}_{ij}-\tilde{r}_{ij}|> \epsilon)} \\ &\leq I_{(|\tilde{r}_{ij}|\geq t-\epsilon)} + I_{(|\hat{r}_{ij}-\tilde{r}_{ij}|>\epsilon)}\end{aligned}$$

Thus, using the results from (F), we have

$$(D) = O_P\Big(c(p)t^{-q}\sqrt{\frac{\log p}{n}} + t^{1-q}c(p)\Big)$$

□

2.6 Extension of Rank Correlation

Normal score transform

Some interpolation and simulation methods assume the input data to be normally distributed, such as kriging method [29] in the field of geostatistics , which is a technique to interpolate the value of a random field (e.g., the elevation, z, of the landscape as a function of the geographic location) at an unobserved location from observations of its value at nearby locations. The normal score transformation is designed to transform the data set so that it marginally closely resembles a standard normal distribution. It achieves this by ranking the values in the data set from lowest to highest and matching these ranks to equivalent ranks generated from a normal distribution. Steps in the transformation are as follows: the data set is sorted and ranked; an equivalent rank from a standard normal distribution is found for each rank from the data set; and the normal distribution values associated with these ranks make up the transformed data set. The ranking process can be done using the frequency distribution or the cumulative distribution of the data set. And studying the dependence structure of the transformed random variables is a very important step for many applications, and this also makes it a direct extension of the Spearman's rank correlation.

Formally, for random variable X with c.d.f F, after the normal score transformation it becomes $\Phi^{-1}(F(X))$. Thus, the normal score transformed correlation between X and Y is defined as, $cor(\Phi^{-1}(F(X)), \Phi^{-1}(G(Y)))$.

In the high dimensional scenario, Liu, Lafferty and Wasserman (2009) [25] studied a semiparametric Gaussian copula model, which they called "nonparanormal". Under their model, the correlations between coordinates are in fact the normal score transformed correlation. Consistency results are obtained for the differences between sample normal score transformed correlation and the true normal score transformed correlation matrix under the matrix l_∞ norm, and also for the l_1 regularized sample concentration covariance matrices.

M-Correlation

Here we extend the rank correlation to a more general monotone transformed version, which we call M-correlation.

Definition: We are given a strictly monotone increasing function $\psi : [0,1] \to R$, which is square-integrable. Then, define the M-correlation between X and Y as

$$r_m(X,Y) = cor(\psi(F(X)), \psi(G(Y)))$$

where F and G are the cumulative distribution functions of X and Y respectively.

It's easy to see that , when $\psi = Identity$, the M-correlation is just the rank correlation itself. And when $\psi = \Phi^{-1}$, the M-correlation is the normal score transformed correlation.

For simplicity, without loss of generosity, we can always assume that $\int_0^1 \psi(u)du = 0$ and $\int_0^1 \psi^2(u)du = 1$. Then the population M-correlation becomes

$$r_m(X,Y) = \mathbb{E}[\psi(F(X))\psi(G(Y))]$$

which holds because of the assumption on ψ , and that $F(X)$ and $G(Y)$ are uniformly distributed on $[0,1]$.

2.7 Theoretical Results on M-correlation Matrices

In this section, we'll extend the results for the rank correlation matrices to M-correlation matrices.

Notations

We'll denote the given monotone transformed function by ψ, and assume it's a strictly increasing function from $[0,1] \to R$ and $\int_0^1 \psi(u)du = 0$, $\int_0^1 \psi^2(u)du = 1$. Then, we know that ψ is differentiable almost everywhere, i.e. the non-differentiable set has Lebesgue measure

0. In order to simplify the proofs, we can assume that ψ is differentiable on $(0,1)$.

The correlations between random vectors X are collectively represented by the correlation matrix. The $(i,j)th$ element of the matrix represents the correlation between components X_i and X_j. In parallel, we define the M-covariance matrix of the random vector X, where the $(i,j)th$ element of the matrix represents the M-covariance between components X_i and X_j.

Using the same notation $R = \big(r_{ij}\big)$ for the M-correlation matrix, which belongs to $\mathbb{R}^{p\times p}$, with

$$r_{ij} = P\big(\psi(F_i)\psi(F_j)\big).$$

Let $\hat{R} = \big(\hat{r}_{ij}\big)$ denote the empirical M-correlation matrix, $\in \mathbb{R}^{p\times p}$, with

$$\hat{r}_{ij} = P_n\big(\psi(\hat{F}_i)\psi(\hat{F}_j)\big) - P_n\big(\psi(\hat{F}_i)\big)P_n\big(\psi(\hat{F}_j)\big),$$

where $\hat{F}_j(x) = \frac{1}{n}\sum_{i=1}^{n} 1_{[X_{ij},\infty)}(x)$, which is the empirical c.d.f.'s.

For ψ which is bounded on $[0,1]$, the above empirical M-correlation is well-defined. However, if it's unbounded, the value of $|\psi(\hat{F})|$ can go to infinity. Tsukahara (2005) [42] suggested using $\frac{n}{n+1}\hat{F}_j$ instead of $\hat{F}_j$. However, as pointed out by [25], in the high dimensional scenario, unless ψ is regularized at the end points there could be problems with convergence of the empirical M-correlation. As in [25], we need to use a truncated estimator:

$$\tilde{F}_j(x) = \begin{cases} \delta_n & \text{if } \hat{F}_j(x) < \delta_n \\ \hat{F}_j(x) & \text{if } \delta_n \le \hat{F}_j(x) \le \delta_n \\ 1-\delta_n & \text{if } \hat{F}_j(x) > 1-\delta_n \end{cases} \tag{2.11}$$

where δ_n will be determined later in the theorem.

In the remainder of this section, we'll use $\tilde{F}$ instead of $\hat{F}$ to formulate our estimators. Denote $\tilde{R} = \big(\tilde{r}_{ij}\big)$, with

$$\tilde{r}_{ij} = P_n\big(\psi(\tilde{F}_i)\psi(\tilde{F}_j)\big) - P_n\big(\psi(\tilde{F}_i)\big)P_n\big(\psi(\tilde{F}_j)\big),$$

In addition, we will also define the matrices

$$\tilde{R}^0 := \Big(P_n\big(\psi(\tilde{F}_i)\psi(\tilde{F}_j)\big)\Big)_{i,j}.$$
$$\tilde{R}^1 := \big(P_n\big(\psi(F_i)\psi(F_j)\big)\big)_{i,j}.$$

Results on $\tilde{R}$

Before considering any regularization methods, we'll first present some theoretical results about $\tilde{R}$ in this subsection.

As discussed earlier, the boundary points might cause trouble, we'd like to first get rid of them by using the truncation.

Define $U_{ij} := F_j(X_{ij})$. For each $j \in \{1, \cdots, p\}$, $U_{1j}, \cdots, U_{nj}$ are i.i.d. $U[0, 1]$.

Define the event

$$\mathcal{A}_n := \{\delta_n \leq U_{ij} \leq 1 - \delta_n,\ i = 1, \cdots, n,\ j = 1, \cdots, p\}.$$

Then

$$\begin{aligned}\mathbb{P}(\mathcal{A}_n^c) &\leq \sum_{j=1}^{p} 2\mathbb{P}(\max_{i=1,\cdots,n} U_{ij} > 1 - \delta_n) \\ &= 2p[1 - (1 - \delta_n)^n].\end{aligned}$$

Choose $\delta_n = 1 - (1 - p^{-\gamma})^{1/n}$, for $\gamma > 1$, then the above probability tends to 0.

And

$$\mathbb{P}(*) \leq \mathbb{P}(*|\mathcal{A}_n) + \mathbb{P}(\mathcal{A}_n^c)$$

Thus, we only need to carry out our analysis on the event $\mathcal{A}_n$.

Since ψ is a monotone function, the maximum is obtained on the boundary. Denote

$$M_n := \max\{|\psi(\delta_n)|, |\psi(1 - \delta_n)|\} = \max_{\delta_n \leq t \leq 1-\delta_n} |\psi(t)|.$$

Let

$$B_n := \sup_{\delta_n \leq t \leq 1-\delta_n} |\psi'(t)|.$$

Theorem 2.7.1. *If* $C_n := \max(B_n, M_n) = o((n/\log p)^{1/4})$, *then*

$$\|\tilde{R} - R\|_\infty = O_P(M_n C_n \sqrt{\frac{\log p}{n}})$$

As a first step we'll give the big picture of the proof.

According to the triangular inequality, we have

$$\|\tilde{R} - R\|_\infty \leq \|\tilde{R} - \tilde{R}^0\|_\infty + \|\tilde{R}^0 - R\|_\infty$$

Notice that $\tilde{R} - \tilde{R}^0$ with element $P_n\big(\psi(\tilde{F}_i)\big)P_n\big(\psi(\tilde{F}_j)\big)$ is of higher order. Thus, we only need to consider the second term on the right hand side of the above inequality.

Using the triangular inequality again, we have

$$\|\tilde{R}^0 - R\|_\infty \leq \|\tilde{R}^0 - \tilde{R}^1\|_\infty + \|\tilde{R}^1 - R\|_\infty.$$

In the following, we'll treat the two terms on the right hand side separately.

•$\|\tilde{R}^1 - R\|_\infty$

Lemma 2.7.1. $\mathbb{P}\big(\ |P_n\big(\psi(F_i)\psi(F_j)\big) - P\big(\psi(F_i)\psi(F_j)\big)|\ \geq t|\mathcal{A}_n\big) \leq 2\exp(-\frac{nt^2}{2(M_n^2+1)^2})$

Proof. Notice that $\{Y_l := \psi\big(F_i(X_{li})\big)\psi\big(F_j(X_{lj})\big) - P\big(\psi(F_i)\psi(F_j)\big)\}$, $l = 1, \cdots, n$ are i.i.d. random variables with mean 0.

Also

$$|P\big(\psi(F_i)\psi(F_j)| \leq \sqrt{P(\psi^2(F_i))P(\psi^2(F_j))} = 1$$

Thus, conditioning on $\mathcal{A}_n$, $|Y_l|$ is bounded by $M_n^2 + 1$.

By Hoeffding's inequality, we have

$$\mathbb{P}(|\bar{Y}_l| \geq t|\mathcal{A}_n) \leq 2\exp(-nt^2/2(M_n^2+1)^2).$$

□

Lemma 2.7.2. *If* $M_n = o((n/\log p)^{1/4})$,

$$\|\tilde{R}^1 - R\|_\infty = O_P(\sqrt{M_n^2\frac{\log p}{n}})$$

Proof. Using the results in Lemma 2.7.1,

$$\begin{aligned}
&\mathbb{P}(\|\tilde{R}^1 - R\|_\infty \geq t|\mathcal{A}_n) \\
&= 1 - \mathbb{P}(\max_{i,j} |P_n\big(\psi(F_i)\psi(F_j)\big) - P\big(\psi(F_i)\psi(F_j)\big)| < t|\mathcal{A}_n) \\
&= 1 - \prod_{i,j} \mathbb{P}(|P_n\big(\psi(F_i)\psi(F_j)\big) - P\big(\psi(F_i)\psi(F_j)\big)| < t|\mathcal{A}_n) \\
&\leq 1 - \prod_{i,j} (1 - 2\exp(-\frac{nt^2}{8})) \\
&\leq 2p^2 \exp(-\frac{nt^2}{2(M_n^2+1)^2})
\end{aligned}$$

Let $t = O(M_n^2\sqrt{\frac{\log p}{n}})$, we have

$$\|\tilde{R}^1 - R\|_\infty = O_P(M_n^2\sqrt{\frac{\log p}{n}}) \tag{2.12}$$

□

Remark: From this lemma, we see that to make the estimator consistent, ψ cannot grow too fast when approaching the boundary points.

•$\|\tilde{R}^0 - \tilde{R}^1\|_\infty$

First, notice that,

$$\mathbb{P}\left(|P_n(\psi(\tilde{F}_i)\psi(\tilde{F}_j)) - P_n(\psi(F_i)\psi(F_j))| \geq t \ |\mathcal{A}_n\right) \tag{2.13}$$

$$\leq \mathbb{P}\left(|P_n\Big((\psi(\tilde{F}_i) - \psi(F_i))(\psi(\tilde{F}_j) - \psi(F_j))\Big)| \geq \frac{t}{3} \ |\mathcal{A}_n\right) \tag{2.14}$$

$$+ 2\mathbb{P}\left(|P_n\Big((\psi(\tilde{F}_i) - \psi(F_i))(\psi(F_j))\Big)| \geq \frac{t}{3} \ |\mathcal{A}_n\right) \tag{2.15}$$

Further, $|P_n\Big((\psi(\tilde{F}_i) - \psi(F_i))(\psi(\tilde{F}_j) - \psi(F_j))\Big)|$ is of higher order than $|P_n\Big((\psi(\tilde{F}_i) - \psi(F_i))(\psi(F_j))\Big)|$. So we only need to analyze the second term on the right hand of the above inequality.

Further, by Dvoretzky-Kiefer-Wolfowitz inequality , for $\tilde{F}_n$, we have

$$\begin{aligned}
&\mathbb{P}(\sup_{x\in\mathcal{R}}|\tilde{F}_n(x)-F(x)|\geq t)\\
\leq&\mathbb{P}(\sup_{x\in\mathcal{R}}|\tilde{F}_n(x)-\hat{F}(x)|\geq\frac{t}{2})+\mathbb{P}(\sup_{x\in\mathcal{R}}|\hat{F}_n(x)-F(x)|\geq\frac{t}{2})\\
\leq&I_{\{\delta_n\geq\frac{t}{2}\}}+2\exp(-2nt^2)
\end{aligned}$$

Now, we're ready to analyze the probability in (2.13).

Lemma 2.7.3. *If $B_n<\infty$, then, for $t>2\delta_n$,*

$$\mathbb{P}\left(|P_n\Big((\psi(\tilde{F}_i)-\psi(F_i))(\psi(F_j))\Big)|\geq\frac{t}{3}\ |\mathcal{A}_n\right)\leq\exp(-\frac{2nt^2}{9M_n^2B_n^2}) \tag{2.16}$$

Proof. Let $S_n:=\{x:\delta_n\leq F_i(x)\leq 1-\delta_n\}$.

$$\begin{aligned}
&\mathbb{P}\left(|P_n\Big((\psi(\tilde{F}_i)-\psi(F_i))(\psi(F_j))\Big)|\geq\frac{t}{3}\ |\mathcal{A}_n\right)\\
\leq&\mathbb{P}\left(\sup_{S_n}|\psi(\tilde{F}_i(x))-\psi(F_i(x))|\ |P_n(\psi(F_j))|\geq\frac{t}{3}\ |\mathcal{A}_n\right)
\end{aligned}$$

By the mean value theorem, $\exists\ s\in[\delta_n,1-\delta_n]$, s.t.

$$\psi(\tilde{F}_n(x))-\psi(F(x))=\psi'(s)(\tilde{F}_n(x))-F(x))$$

Hence,

$$\sup_{S_n}|\psi(\tilde{F}_i(x))-\psi(F_i(x))|\leq\sup_{\{\delta_n\leq s\leq 1-\delta_n\}}|\psi'(s)|\sup_{S_n}|\tilde{F}_i(x)-F_i(x)|$$

Also on $\mathcal{A}_n$, $\psi(F_j)$ is bounded by M_n, and so is $P_n(\psi(F_j))$.

Therefore,

$$\begin{aligned}
&\mathbb{P}\left(|P_n\Big((\psi(\tilde{F}_i)-\psi(F_i))(\psi(F_j))\Big)|\geq\frac{t}{3}\ |\mathcal{A}_n\right)\\
\leq&\mathbb{P}\left(\sup_{S_n}|\tilde{F}_i(x)-F_i(x)|B_nM_n\geq\frac{t}{3}\ |\mathcal{A}_n\right)\\
\leq&\exp(-\frac{2nt^2}{9M_n^2B_n^2})
\end{aligned}$$

□

Corollary 2.7.1. *If* $M_n B_n = o((n/\log p)^{1/2})$, *then*

$$\|\tilde{R}^0 - \tilde{R}^1\|_\infty = O_P(M_n B_n \sqrt{\frac{\log p}{n}})$$

Proof. Using the same argument as in the proof of lemma 2.7.2, and the results of lemma 2.7.3, the desired result follows. □

Remark: Here another assumption for ψ has been imposed, which is that it should not change too dramatically on $[\delta_n, 1-\delta_n]$, i.e. $|\psi'|$ cannot grow too fast.

The result in Theorem 2.7.1 directly follows from Lemma 2.7.2 and corollary 2.7.1.

Banding the M-correlation matrix

As for the rank correlation matrices, we'll show that banded sample M-correlation matrices also give consistent estimates under the operator norm in the matrix class of

$$\mathcal{U}(\alpha, c) = \{R : \max_j \sum_i \{|r_{ij}| : \ |i-j| > k\} \leq ck^{-\alpha} \text{ for all } k > 0\}$$

Theorem 2.7.2. *If* $k \asymp (M_n^2 C_n^2 n^{-1} \log p)^{-1/(2(\alpha+1))}$,

$$\|B_k(\tilde{R}) - R\| = O_P\Big(\big(M_n^2 C_n^2 \frac{\log p}{n}\big)^{\alpha/(2(\alpha+1))}\Big)$$

uniformly on $R \in \mathcal{U}(\alpha, c)$.

Proof. Notice that,

$$\|B_k(\tilde{R}) - R\| \leq \|B_k(\tilde{R}) - B_k(R)\| + \|B_k(R) - R\| \tag{2.17}$$

Thus, we can deal the two parts separately.

The second part on the right hand side has a simple bound:

$$\|B_k(R) - R\| \leq \|B_k(R) - R\|_{(1,1)} \leq ck^{-\alpha} \tag{2.18}$$

To deal with the first part, we have to use some basic facts on norms of matrices, and the results of theorem 2.7.1. After doing this, we have,

$$\|B_k(\tilde{R}) - B_k(R)\| = O_P(k\|B_k(\tilde{R}) - B_k(R)\|_\infty) = O_P(kM_n C_n \sqrt{\frac{\log p}{n}}).$$

Combining these results, and setting $k \asymp (M_n^2 C_n^2 n^{-1} \log p)^{-1/(2(\alpha+1))}$, we will get the desired result. □

Tapering

Given a general ψ, it's hard to construct a finite collection of distributions whose M-correlation matrices have desired structures. Here, we give the following conjecture for the minimax result without proof.

Conjecture 2.7.1. *Suppose $p \leq \exp(\gamma n)$ for some constant $\gamma > 0$. The minimax risk of estimating the M-correlation matrix R over the class $\mathcal{U}(\alpha, c)$ under the operator norm satisfies*

$$\inf_{\hat{R}} \sup_{\mathcal{U}(\alpha,c)} \mathbb{E}\|\hat{R} - R\|^2 \asymp \min\left\{ n^{-\frac{2\alpha}{2\alpha+1}} + M_n^2 C_n^2 \frac{\log p}{n}, \frac{p}{n} \right\}$$

Thresholding

Similarly, the results are uniform over families of M-correlation matrices which satisfy a fairly natural notion of sparsity:

$$\mathcal{U}(q, c(p)) = \{R : \sum_{j=1}^{p} |r_{ij}|^q \leq c(p) \ \text{ for all } i\}.$$

Theorem 2.7.3. *Uniformly on $\mathcal{U}(q, c(p))$, for sufficiently large M', if*

$$t_n = M' M_n C_n \sqrt{\frac{\log p}{n}},$$

and $M_n C_n \sqrt{\frac{\log p}{n}} = o(1)$, then

$$\|T_{t_n}(\hat{R}) - R\| = O_P\Big(c(p)\big(M_n^2 C_n^2 \frac{\log p}{n}\big)^{\frac{1-q}{2}}\Big).$$

Proof: The proof procedure will be very similar to the proof thresholded rank correlation matrices. So we omit the details here.

2.8 Apply the Results to CCA Problems

As for the CCA problem, if the original data sets are not from normal distribution, or there exits strong non-linear effect among random variables, we propose to use the normal score transformation before doing CCA. Then the solution can be obtained by replacing the covariance and cross-covariance matrices in 1.2 by the corresponding normal score transformed correlation matrices. Then combining the results from chapter 1 and chapter 2, we can also derived the consistency results for this generalized problem.

2.9 Simulation

Choice of the tuning parameters

Foe the banding method, we provide the rate of the bandwidth k to guarantee convergence; and for the thresholding method, the rate of the threshold t is also provided. However, how to choose the tuning parameters remains a question in practice. Here we'll give some practical guidance for selecting them. The procedure is similar as in [4] and [3].

Random split the original sample into two parts and use the sample M-correlation matrix of one sample as the "target" to choose the best tuning parameter for the other sample. Let n_1, $n_2 = n - n_1$ be the two sample sizes for the random split, and let $\hat{R}_1^v$, $\hat{R}_2^v$ be the two sample M-correlation matrices from the hvth split, for h$v = 1, \cdots, N$.

Minimize the estimated risk

$$\hat{k} = argmin_k \frac{1}{N} \sum_{v=1}^{N} \|B_k(\hat{R}_1^v) - \hat{R}_2^v\|_{(1,1)}$$

$$\hat{t} = argmin_t \frac{1}{N} \sum_{v=1}^{N} \|T_t(\hat{R}_1^v) - \hat{R}_2^v\|_F$$

Rank correlation

First, notice that the sample rank correlation shares the same property as of the population rank correlation, which is that it's invariant under the monotone transformations. Thus, the simulation results are independent of the choice of marginal distributions of random variables. Thus, we can choose the simplest way to generate the data. The procedure to generate the data with a certain rank correlation structure is as follows,

- let R be the rank correlation matrix with the designed structure and compute $\Sigma = (\frac{6}{\pi} arcsin(\frac{r_{ij}}{2}))$;
- get $X_1, \cdots, X_n$ i.i.d. from $N(0, \Sigma)$.

Then we have that the rank correlation matrix of X is R.

The true rank matrix R is constructed as

$$r_{ij} = \rho^{|i-j|} I_{\{|i-j| \le k\}}, \quad i = 1, \cdots, p; \ j = 1, \cdots, p.$$

Thus the true bandwidth of the rank correlation matrix is k.

We consider three values of $p = 30, 100, 200$ and the sample size is fixed at $n = 100$. And set $\rho = 0.5$, $k = 2$, $N = 50$.

First, we'll present the banding results in Table 2.1. Here we use four different evaluation methods to measure the performance, the matrix operator norm, the matrix 1-norm, the Frobenius norm and the absolute value between the largest eigenvalue, $|\hat{\lambda}_{max} - \lambda_{max}|$. The first three measurements are different kinds of matrix norms of $\hat{R} - R$; while the last one assesses how accurate each of the estimators would be in estimating the first principal component. The estimated bandwidths are also shown in the table. The table reports average losses and standard deviations of the above measures over 100 replications.

p	Estimated bandwidth	Operator norm	Matrix 1-norm	Frobenius norm	$\lvert\hat{\lambda}_{max} - \lambda_{max}\rvert$
30	1.96(0.20)	1.02(0.06)	0.59(0.11)	1.05(0.22)	0.08(0.10)
100	1.97(0.17)	1.13(0.05)	0.68(0.10)	1.90(0.34)	0.09(0.08)
200	1.99(0.10)	1.22(0.03)	0.74(0.09)	2.69(0.28)	0.11 (0.06)

Table 2.1: Results for banded rank correlation matrices

Next, we also give out the results for the thresholding method in the following table.

p	Optimal threshold	Operator norm	Matrix 1-norm	Frobenius norm	$\lvert\hat{\lambda}_{max} - \lambda_{max}\rvert$
30	0.35(0.04)	1.08(0.08)	1.15(0.26)	2.19(0.36)	0.27(0.15)
100	0.48(0.02)	1.22(0.07)	1.50(0.01)	6.35(0.56)	0.47(0.11)
200	0.51(0.01)	1.23(0.03)	1.50(0.00)	9.88(0.34)	0.49(0.09)

Table 2.2: Results for thresholded rank correlation matrices

M-correlation

For the general M-correlation, it's not easy to generate the data which has the desired M-correlation structure. Here we'll give one specific example for a particular choice of ψ.

Example: $\psi(u) = -2\log(1-u)$, which is the inverse c.d.f of $\chi^2(2)$ distribution.

Generate the date:
- get $Z_1, \cdots, Z_{p+1}$ i.i.d. from $N(0,1)$;
- let $X_j := Z_j^2 + Z_{j+1}^2$, then $X_j \sim \chi^2(2), j = 1, \cdots, p$;
- repeat n times .

It's easy to check that $(X_1, \cdots, X_p)$ has marginal $\chi^2(2)$ distribution and the correlation matrix is a banded matrix with bandwidth 1 ($cor(X_i, X_j) = 0.5 I_{|i-j|=1}$).

Estimate:
- compute $\tilde{F}_j(X_{ij})$;
- compute the sample correlation matrix of $(\psi(\tilde{(F)}_j(X_{ij}))$;
- find the bandwidth by cross-validation;
- band the sample M-correlation matrix .

We consider three values of $p = 30, 100, 200$ and the sample size is fixed at $n = 100$.

Table 2.3 and Table 2.4 show the results for the banding method and thresholding method respectively. The table reports average losses and standard deviations of four measures over 100 replications.

p	Estimated bandwidth	Operator norm	Matrix 1-norm	Frobenius norm	$\lvert\hat{\lambda}_{max} - \lambda_{max}\rvert$
30	1.04(0.20)	0.32(0.06)	0.41(0.07)	0.85(0.12)	0.04(0.03)
100	1.16(0.42)	0.37(0.07)	0.49(0.07)	1.58(0.12)	0.07(0.03)
200	1.26(0.54)	0.42(0.08)	0.80(0.08)	3.00(0.13)	0.31 (0.07)

Table 2.3: Results for banded M-correlation matrices

Normal score transformed correlation

Notice that $(\Phi^{-1}(F_1(X_1)), \cdots, \Phi^{-1}(F_p(X_p)))$ follows a multivariate Gaussian distribution. Also, the same reason as showing for the rank correlation simulation, it's irrelevant of choices of the marginal distributions of samples. Thus, we have the following procedure to generate the data.

- let R be the normal score transformed correlation matrix with the designed structure ;

p	Optimal threshold	Operator norm	Matrix 1-norm	Frobenius norm	$\lvert\hat{\lambda}_{max} - \lambda_{max}\rvert$
30	0.43(0.04)	0.66(0.07)	0.85(0.20)	1.94(0.49)	0.05(0.04)
100	0.55(0.03)	0.92(0.04)	1.00(0.00)	5.96(0.42)	0.09(0.06)
200	0.57(0.01)	0.96(0.03)	1.00(0.00)	8.73(0.27)	0.08 (0.05)

Table 2.4: Results for thresholded M-correlation matrices

- get $X_1, \cdots, X_n$ i.i.d. from $N(0, R)$;

In this way, the normal score transformed correlation matrix of X is R .

The true normal score transformed correlation matrix R is constructed as

$$r_{ij} = \rho^{|i-j|} I_{\{|i-j| \leq k\}}, \quad i = 1, \cdots, p; \; j = 1, \cdots, p.$$

with $k = 2$ and $\rho = 0.5$.

Consider three values of $p = 30, 100, 200$ and the sample size is fixed at $n = 100$. Table 2.5 and Table 2.6 show the results for the banding method and thresholding method respectively. The table reports average losses and standard deviations of four measures over 100 replications.

p	Estimated bandwidth	Operator norm	Matrix 1-norm	Frobenius norm	$\lvert\hat{\lambda}_{max} - \lambda_{max}\rvert$
30	1.65(0.52)	0.50(0.12)	0.70(0.15)	1.41(0.47)	0.22(0.21)
100	1.85(0.57)	0.53(0.09)	0.81(0.13)	2.46(0.76)	0.20(0.15)
200	1.84(0.69)	0.54(0.09)	0.85(0.14)	3.68(1.13)	0.24 (0.15)

Table 2.5: Results for banded normal score transformed correlation matrices

The mixture model

Here, we consider that $X_1, \cdots, X_n$ i.i.d. follow a Gaussian mixture model

$$(1 - \epsilon)N(0, \Sigma) + \epsilon N(0, I),$$

p	Optimal threshold	Operator norm	Matrix 1-norm	Frobenius norm	$\lvert\hat{\lambda}_{max} - \lambda_{max}\rvert$
30	0.37(0.05)	0.84(0.12)	1.25(0.22)	2.38(0.39)	0.27(0.17)
100	0.5(0.18)	1.36(0.06)	1.50(0.01)	6.57(0.38)	0.46(0.10)
200	0.52(0.01)	1.41(0.05)	1.50(0.01)	9.73(0.22)	0.44(0.11)

Table 2.6: Results for thresholded normal score transformed correlation matrices

i.e. with probability $(1-\epsilon)$ it samples form $N(0,\Sigma)$ distribution, and with probability ϵ it samples form $N(0,I)$ distribution.

Let R be a $p \times p$ matrix, defined as

$$r_{ij} = \rho^{|i-j|}.$$

And let $\sigma_{ij} = \frac{6}{\pi} arcsin(r_{ij})$. Then it's easy to check that $\Sigma := (\sigma_{ij})$ is a non-negative definite matrix, which makes it a correlation matrix. Thus, a random variable following $N(0,\Sigma)$ distribution has the rank correlation matrix as R.

In consequence, the marginal distribution of X_i is $N(0,1)$. Then after some computation, we can deduce that the rank correlation of the mixture model is just

$$(1-\epsilon)r_{ij} \quad \text{if } i \neq j, \quad \text{and 1 if } i = j.$$

In the tables below, the results are obtained under the setting of $\epsilon = 0.5$, $n = p = 100$ with 100 replicates.

ρ	Operator norm	Matrix 1-norm	Frobenius norm	$\lvert\hat{\lambda}_{max} - \lambda_{max}\rvert$
0.5	3.44(0.15)	10.12(0.33)	11.18(0.11)	2.85(0.22)
0.6	3.53(0.19)	10.20(0.37)	11.17(0.12)	2.77(0.27)
0.8	3.79(0.41)	10.46(0.41)	11.15(0.19)	2.43(0.57)

Table 2.7: Results for the mixture model: the sample rank correlation matrix

ρ	Estimated bandwidth	Operator norm	Matrix 1-norm	Frobenius norm	$\lvert\hat{\lambda}_{max} - \lambda_{max}\rvert$
0.5	1.09(0.75)	0.90(0.02)	1.43(0.07)	5.73(0.36)	0.25(0.17)
0.6	1.65(0.80)	0.99(0.03)	1.77(0.12)	5.91(0.37)	0.26(0.21)
0.8	4.31(1.53)	1.41(0.19)	3.34(0.26)	6.92(0.52)	0.96 (0.56)

Table 2.8: Results for the mixture model: Banding

ρ	Optimal threshold	Operator norm	Matrix 1-norm	Frobenius norm	$\lvert\hat{\lambda}_{max} - \lambda_{max}\rvert$
0.5	0.58(0.01)	0.83(0.01)	1.50(0.00)	6.43(0.00)	0.49(0.00)
0.6	0.58(0.02)	0.98(0.02)	2.00(0.00)	7.35(0.00)	0.95(0.15)
0.8	0.55(0.05)	2.14(0.16)	4.50(0.02)	10.42(0.26)	2.81(0.52)

Table 2.9: Results for the mixture model: Thresholding

We have also carried out additional simulations for other combinations of sample sizes and dimensions. The behavior of the banding and Thresholding estimators are similar.

We notice some interesting facts from the results,

- both banding and thresholding methods outperform the sample rank correlation matrices ;
- when banding is given the correct order of variables, it performs better than thresholding, since it is taking advantage of the underlying structure;
- the same model requires relatively more regularization in higher dimensions.

Chapter 3

Renyi Correlation

3.1 Introduction

In this chapter, we further discuss other measures of independence. The section 3.2 is based on a discussion paper of Brownian distance covariance [38] by Peter Bickel and I. Then in section 3.3, we used the idea of the Renyi correlation to generalize the CCA problem. It includes the basic framework and some future directions.

3.2 Discussion of: Brownian distance covariance

Szekely and Rizzo [38] present a new interesting measure of correlation, brownian distance correlation. The idea of using $\int |\phi_n(u,v) - \phi_n^{(1)}(u)\phi_n^{(2)}(v)|^2 d\mu(u,v)$, where ϕ_n, $\phi_n^{(1)}$, $\phi_n^{(2)}$ are the empirical characteristic functions of a sample $(X_i,\ Y_i)\, i = 1, \cdots, n$ of independent copies of X and Y is not so novel. A.Feuerverger considered such measures in a series of papers [13] . Aiyou Chen and Peter Bickel have actually analyzed such a measure for estimation in [9] in connection with ICA.

However, the choice of $\mu(\cdot,\cdot)$ which makes the measure scale free, the extension to $X \in \mathbb{R}^p$, $Y \in \mathbb{R}^q$ and its identification with the Brownian distance covariance is new, surprising and interesting. There are three other measures available, for general p, q

1. The canonical correlation ρ between X and Y.
2. The rank correlation r (for $p = q = 1$) and its canonical correlation generalization.
3. The Renyi correlation R.

All vanish along with the Brownian distance (BD) correlation in the case of independence and all are scale free. The Brownian distance and Renyi covariance are the only ones which vanish iff X and Y are independent.

However, the three classical measures also give a characterization of total dependence. If $|\rho| = 1$, X and Y must be linearly related; if $|r| = 1$, Y must be a monotone function of X and if $R = 1$, then either there exist non trivial functions f and g such that $\mathbb{P}(f(X) = g(Y)) = 1$ or at least there is a sequence of such non trivial functions f_n, g_n of variance 1 such that $\mathbb{E}(f_n(X) - g_n(Y))^2 \to 0$.

In this respect, by Theorem 4 of Szekely and Rizzo, for the common $p = q = 1$ case, BD correlation doesn't differ from Pearson correlation.

Szekely and Rizzo illustrate several possible applications of distance covariance by examples. Although we found the examples varied and interesting and the computation of p values for the BD covariance effective, we are not convinced that the comparison with the rank and Pearson correlations is quite fair.

Intuitively, the closer the form of observed dependence is to that exhibited for the extremal value of the statistic, the more power one should expect. In their example 1, the data are from the NIST Statistical Reference Data sets, where Y is a distinctly non monotone function of X plus noise, a situation where we would expect the rank correlation to be weak. And similarly the other examples correspond to non-linear relationships between X and Y in which we would expect the Pearson correlation to perform badly. In general for goodness of fit, it is important to have statistics with power in directions which are plausible departures, Bickel and Ritov [6].

An alternative being studied here in the context of high dimensional data is the empirical Renyi correlation.

Let $f_1, f_2, \cdots$ be an orthonormal basis of $L_2(P_X)$ and $g_1, g_2, \cdots$ an orthonormal basis of $L_2(P_Y)$ where $L_2(P_X)$ is the Hilbert space of function f such that $\mathbb{E}f^2(X) < \infty$ and similarly for $L_2(P_Y)$.

Let the (K, L) approximate Renyi correlation be defined as,

$$\max\{\text{corr}(\sum_{k=1}^{K} \alpha_k f_k(X), \sum_{l=1}^{L} \beta_l g_l(Y))\}$$

where corr is Pearson correlation.

This is seen to be the canonical correlation of $\underline{f}(X)$ and $\underline{g}(Y)$ where $\underline{f} \equiv (f_1, \cdots, f_K)^T$, $\underline{g} \equiv (g_1, \cdots, g_L)^T$, and is easily calculated as a generalized eigenvalue problem. The empirical (K, L) correlation is just the solution of the corresponding empirical problem where the variance covariance matrices $Var\ \underline{f}(X) \equiv \mathbb{E}[\underline{f_c}(X)\underline{f_c}^T(X)]$ where $\underline{f_c}(X) \equiv \underline{f}(X) - \mathbb{E}\underline{f}(X)$, $Var\ \underline{g}(Y)$ and $Cov(\underline{f}(X),\ \underline{g}(Y))$ are replaced by their empirical counterparts. For $K, L \to \infty$, the (K, L) correlation tends to the Renyi correlation,

$$R \equiv \max\{\text{corr}(f(X), g(Y)) : f \in L_2(P_X),\ g \in L_2(P_Y)\}$$

For the empirical (K, L) correlation, K and L have to be chosen in a data determined way, although evidently each K, L pair provides a test statistic. An even more important choice is that of the f_k and g_l (which need not be orthonormal but need only have a linear span dense in their corresponding Hilbert spaces).

We compare the performance of these test statistics in the first of the Szekely-Rizzo example in the next section.

Comparison on Data Example

Here we'll investigate the performance of the standard ACE estimate of the Renyi correlation and a version of (K, L) correlation in the first of the Szekely-Rizzo examples.

Recall that the proposed nonlinear model is

$$y = \frac{\beta_1}{\beta_2} \exp\Big\{\frac{-(x-\beta_3)^2}{2\beta_2^2}\Big\} + \epsilon.$$

Breiman and Friedman (1985) provided an algorithm, known as alternating conditional expectations (ACE), for estimating the transformations f_0, g_0 and R itself.

The estimated Renyi correlation is very close to 1 (0.9992669) in this case, as expected since Y is a function of X plus some noise. The plot below shows the original relationship between X and Y on the left and the relationship between the estimated transformations $\hat{f}$ and $\hat{g}$ on the right.

Having computed $\hat{R}$, the estimate of R, we compute its significance under the null hypothesis of independence using the permutation distribution just as Szekely and Rizzo did. The p-value is ≤ 0.001, which is extremely small as it should be.

Next, we compute the empirical (K, L) correlation. For this case, we chose, as an orthonormal basis with respect to Lebesgue measure, one defined by the Hermite polynomials defined as $H_n(x) = (-1)^n e^{x^2/2} \frac{d^n}{dx^n} e^{-x^2/2}$, for both X and Y. We take $f_k(\cdot) = g_k(\cdot) = e^{-\frac{x^2}{4}} H_k(\cdot)$.

The following table gives the computation results of different combination of K and L. As before, the p-value is computed by a permutation test, based on 999 replicates.

The value, not surprisingly is close to $\hat{R}$, for $K = L = 5$.

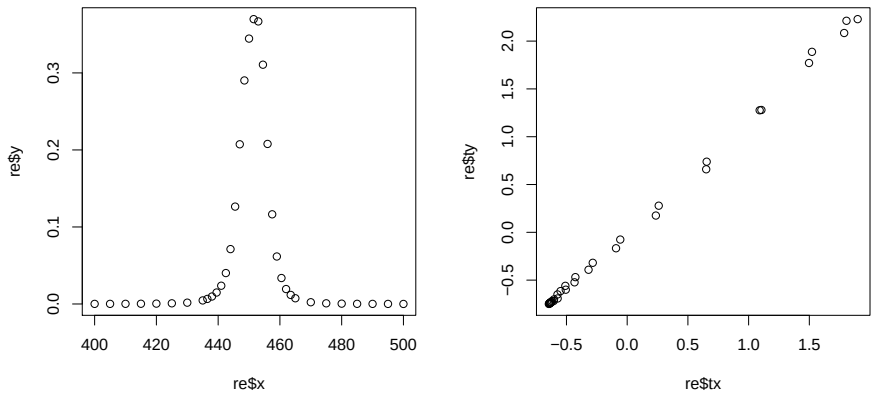

Figure 3.1: Relationship between X and Y and between $\hat{f}$ and $\hat{g}$

	K=2, L=2	K=3, L=4	K=5, L=5
Estimated (K,L) correlation	0.8160803	0.9170764	0.977163
P-value	0.002	0.002	≤ 0.001

Table 3.1: Results of $K - L$ correlation

3.3 Extension of canonical correlation analysis to non-linear cases

Recall from chapter 1, canonical correlation analysis (CCA) is a method of correlating linear relationships between two multidimensional variables. Although it performs well in many cases, linearity certainly has many constrains. Inspired by Renyi correlation, we propose a non-linear version of CCA.

Our Problem (Population version):

Consider two sets of random variables, $X = (X_1, X_2, \cdots, X_p)^\top$ and $Y = (Y_1, Y_2, \cdots, Y_q)^\top$. We are interested in seeking the optimal transformations for each random variable, say $f_1, \cdots, f_p$ and $g_1, \cdots, g_q$, such that correlation between $\sum_{i=1}^{p} f_i(X_i)$ and $\sum_{j=1}^{q} g_j(Y_j)$ is maximized.

Without loss of generality, we assume that $\mathbb{E}f_i(X_i) = 0$, $\mathbb{E}f_i^2(X_i) = 1$ $(i = 1, \cdots, p)$, and $\mathbb{E}g_j(Y_j) = 0$, $\mathbb{E}g_j^2(Y_j) = 1$ $(j = 1, \cdots, q)$.

Thus,

$$\Sigma_x^f := \Big[\mathbb{E}f_i(X_i)f_j(X_j)\Big]_{p\times p} = I_{p\times p}, \qquad \Sigma_y^g := \Big[\mathbb{E}g_i(Y_i)g_j(Y_j)\Big]_{q\times q} = I_{q\times q}$$

Also define the cross-correlation matrix as

$$\Sigma_{xy}^{fg} := \Big[\mathbb{E}f_i(X_i)g_j(Y_j)\Big]_{p\times q}.$$

So this can be formed as an optimization problem,

$$\max_{f,g}\left\{\max_{a,b}\frac{a^\top\Sigma_{xy}^{fg}b}{\sqrt{a^\top\Sigma_x^f a\ b^\top\Sigma_y^g b}}\right\} \tag{3.1}$$

under the constrains that

$$\mathbb{E}f_i(X_i)=0, \mathbb{E}f_i^2(X_i)=1\ (i=1,\cdots,p);\ \mathbb{E}g_j(Y_j)=0, \mathbb{E}g_j^2(Y_j)=1\ (j=1,\cdots,q)$$

Notice that the inner maximization problem is just an ordinary CCA problem for $(f_1(X_1),\cdots,f_p(X_p))$ and $(g_1(Y_1),\cdots,g_q(Y_q))$.

Remarks: Notice that we cannot use the pair-wised Renyi correlation to solve the optimization problem of 3.1.

Problem solving

Directly optimizing over the functional space is hard, here we propose to use (K,L) correlation instead, which is an approximation of Renyi's correlation.

For random variables U, $V\in\mathbb{R}^1$, let $H_1,H_2,\cdots$ be an orthonormal basis of $L_2(P_U)$ and $Q_1,Q_2,\cdots$ an orthonormal basis of $L_2(P_V)$ where $L_2(P_U)$ is the Hilbert space of function f such that $\mathbb{E}H^2(U)<\infty$ and similarly for $L_2(P_V)$.

Let the (K,L) approximate Renyi correlation be defined as,

$$\max_{\alpha,\beta}\{\text{corr}(\sum_{k=1}^{K}\alpha_k H_k(U),\sum_{l=1}^{L}\beta_l Q_l(V))\}.$$

Back to our cases, where X and Y are both multidimensional variables, similarly let $H_1^i,H_2^i,\cdots$ be an orthonormal basis of $L_2(P_{X_i})$ and $Q_1^j,Q_2^j,\cdots$ an orthonormal basis of $L_2(P_{Y_j})$. Thus varying over functions of f_i,g_j can be approximated by varying their coefficients of the projections onto the corresponding basis.

Using vector notations,

$$H^i(X_i) := (H_1^i(X_i), \cdots, H_{K_i}^i(X_i))^\top,$$

$$Q^j(Y_j) := (Q_1^j(Y_j), \cdots, Q_{L_i}^j(Y_j))^\top.$$

$$\alpha^i := (\alpha_1^i, \cdots, \alpha_{K_i}^i)^\top$$

$$\beta^j := (\beta_1^j, \cdots, \beta_{L_i}^j)^\top$$

And define matrices as

$$\Sigma_{HQ} := \begin{bmatrix} \mathbb{E}H^1(X_1)Q^1(Y_1)^\top & \cdots & \mathbb{E}H^1(X_1)Q^q(Y_q)^\top \\ \vdots & \ddots & \vdots \\ \mathbb{E}H^p(X_p)Q^1(Y_1)^\top & \cdots & \mathbb{E}H^p(X_p)Q^q(Y_q)^\top \end{bmatrix}$$

$$\Sigma_H := \begin{bmatrix} \mathbb{E}H^1(X_1)H^1(X_1)^\top & \cdots & \mathbb{E}H^1(X_1)H^p(X_p)^\top \\ \vdots & \ddots & \vdots \\ \mathbb{E}H^p(X_p)H^1(X_1)^\top & \cdots & \mathbb{E}H^p(X_p)X^p(X_p)^\top \end{bmatrix}$$

$$\Sigma_Q := \begin{bmatrix} \mathbb{E}Q^1(Y_1)Q^1(Y_1)^\top & \cdots & \mathbb{E}Q^1(Y_1)Q^q(Y_q)^\top \\ \vdots & \ddots & \vdots \\ \mathbb{E}Q^q(Y_q)Q^1(Y_1)^\top & \cdots & \mathbb{E}Q^q(Y_q)Q^q(Y_q)^\top \end{bmatrix}$$

$$A := \begin{bmatrix} \alpha^1 & & 0 \\ & \ddots & \\ 0 & & \alpha^p \end{bmatrix}$$

$$B := \begin{bmatrix} \beta^1 & & 0 \\ & \ddots & \\ 0 & & \beta^q \end{bmatrix}$$

Then the optimization problem of (1) can be approximated by

$$\max_{A,B} \left\{ \max_{a,b} \frac{a^\top (A^\top \Sigma_{HQ} B) b}{\sqrt{a^\top (A^\top \Sigma_H A) a \; b^\top (B^\top \Sigma_Q B) b}} \right\}$$

which is equivalent to

$$\max_{\tilde{a},\tilde{b}} \frac{\tilde{a}^\top \Sigma_{HQ} \tilde{b}}{\sqrt{\tilde{a}^\top \Sigma_H \tilde{a} \; \tilde{b}^\top \Sigma_Q \tilde{b}}} \tag{3.2}$$

where $\tilde{a} \in \mathbb{R}^{\sum_{i=1}^{p} K_i}$ and $\tilde{b} \in \mathbb{R}^{\sum_{j=1}^{q} L_j}$.

Interestingly, (2) has exactly the same form as the standard CCA problem.

Sample version

Let $X^{(1)}, \cdots, X^{(n)}$ be i.i.d. samples of X and $Y^{(1)}, \cdots, Y^{(n)}$ be i.i.d. samples of Y.

Look at the sub-matrix of Σ_{HQ} first. Define $M_{ij} := \mathbb{E}(H^i(X_i)Q^j(Y_j)^\top)$, which is a $K_i \times L_j$ matrix. The natural estimation of M_{ij} would be

$$\hat{M}_{ij} = \left(\frac{1}{n} \sum_{t=1}^{n} H^i(X_i^{(t)}) Q^j(Y_j^{(t)})^\top \right)$$

When K_i, L_j fixed, n goes to infinity, $\hat{M}_{ij}$ will converge to M_{ij} under operator norm.

Future directions

In this section, I present an extension of canonical correlation analysis by utilizing an approximate method of Renyi correlation. It leaves us a lot of open questions for the future work. Even in the finite dimension case, it's worth of discussion on how to choose the base functions. When n, p and q all go to infinity, regularization methods should be used to obtain the consistency results. Then the question comes up, what kind of regularization methods is suitable for the problem and what can we expect for the convergency rate. I'd like to propose the group lasso [34] [1]. Naturally in our problem, $\tilde{a}$ and $\tilde{b}$ fall into p and q groups respectively. When imposing the sparse assumptions, the group effects play an important role, which makes the group lasso method promising.

Bibliography

[1] F.R. Bach. "Consistency of the group Lasso and multiple kernel learning". In: *The Journal of Machine Learning Research* 9 (2008), pp. 1179–1225.

[2] O. Banerjee and L. El Ghaoui. "First-Order Methods for Sparse Covariance Selection". In: *SIAM Journal on Matrix Analysis and Applications* 30 (2008), p. 56.

[3] P.J. Bickel and E. Levina. "Covariance regularization by thresholding". In: *The Annals of Statistics* 36.6 (2008), pp. 2577–2604.

[4] P.J. Bickel and E. Levina. "Regularized estimation of large covariance matrices". In: *Annals of Statistics* 36.1 (2008), pp. 199–227.

[5] P.J. Bickel, Y. Ritov, and A.B. Tsybakov. "Simultaneous analysis of Lasso and Dantzig selector". In: *The Annals of Statistics* 37.4 (2009), pp. 1705–1732.

[6] P.J. Bickel and T.M.S. Ya'acov Ritov. "Mathematics¿ Statistics Title: Tailor-made tests for goodness of fit to semiparametric hypotheses". In: *Journal reference: Annals of Statistics* 34.2 (2006), pp. 721–741.

[7] T. Cai, C.H. Zhang, and H. Zhou. "Optimal rates of convergence for covariance matrix estimation". In: *Unpublished manuscript* (2008).

[8] E. Candes and T. Tao. "The Dantzig selector: Statistical estimation when p is much larger than n". In: *The Annals of Statistics* 35.6 (2007), pp. 2313–2351.

[9] A. Chen and PJ Bickel. "Consistent independent component analysis and prewhitening". In: *IEEE Transactions on Signal Processing* 53.10 Part 1 (2005), pp. 3625–3632.

[10] S. Chen and D. Donoho. "Basis pursuit". In: *Signals, Systems and Computers, 1994. 1994 Conference Record of the Twenty-Eighth Asilomar Conference on*. Vol. 1. IEEE. 1994, pp. 41–44.

[11] M.D. Doruet and S. Kotz. *Correlation and Dependence*. London: Imperial College Press, 2001.

[12] N. El Karoui. "Operator norm consistent estimation of large-dimensional sparse covariance matrices". In: *The Annals of Statistics* 36.6 (2008), pp. 2717–2756.

[13] A. Feuerverger and R.A. Mureika. "The empirical characteristic function and its applications". In: *The Annals of Statistics* (1977), pp. 88–97.

[14] J. Friedman, T. Hastie, and R. Tibshirani. "Sparse inverse covariance estimation with the graphical lasso". In: *Biostatistics* 9.3 (2008), p. 432.

[15] R. Furrer and T. Bengtsson. "Estimation of high-dimensional prior and posterior covariance matrices in Kalman filter variants". In: *Journal of Multivariate Analysis* 98.2 (2007), pp. 227–255.

[16] D. Hardoon and J. Shawe-Taylor. "The double-barrelled lasso". In: (2008).

[17] D.R. Hardoon, S. Szedmak, and J. Shawe-Taylor. "Canonical correlation analysis: An overview with application to learning methods". In: *Neural Computation* 16.12 (2004), pp. 2639–2664.

[18] H. Hotelling. "Relations between two sets of variates". In: *Biometrika* 28.3/4 (1936), pp. 321–377.

[19] I.M. Johnstone. "On the distribution of the largest eigenvalue in principal components analysis". In: *Annals of Statistics* 29.2 (2001), pp. 295–327.

[20] I.M. Johnstone and A.Y. Lu. "Sparse principal components analysis". In: *J. Amer. Statist. Assoc* (2007).

[21] V. Koltchinskii. "Sparsity in penalized empirical risk minimization". In: *Annales de l'Institut Henri Poincaré, Probabilités et Statistiques*. Vol. 45. 1. Institut Henri Poincaré. 2009, pp. 7–57.

[22] D. Kurowicka and R. Cooke. *Uncertainty analysis with high dimensional dependence modelling*. Wiley series in probability and statistics. Wiley, 2006. ISBN: 9780470863060. URL: `http://books.google.com/books?id=dRVGNYU7RskC`.

[23] K.A. Lê Cao et al. "A sparse PLS for variable selection when integrating omics data". In: *Statistical applications in genetics and molecular biology* 7.1 (2008), p. 35.

[24] O. Ledoit and M. Wolf. "A well-conditioned estimator for large-dimensional covariance matrices". In: *Journal of Multivariate Analysis* 88.2 (2004), pp. 365–411.

[25] H. Liu, J. Lafferty, and L. Wasserman. "The Nonparanormal: Semiparametric Estimation of High Dimensional Undirected Graphs". In: *Journal of Machine Learning Research* 10 (2009), pp. 1–37.

[26] K.V. Mardia, J.T. Kent, and J.M. Bibby. *Multivariate Analysis*. Academic Press, 1980.

[27] N. Meinshausen and P. Bühlmann. "High-dimensional graphs and variable selection with the lasso". In: *The Annals of Statistics* 34.3 (2006), pp. 1436–1462.

[28] N. Meinshausen and B. Yu. "Lasso-type recovery of sparse representations for high-dimensional data". In: *The Annals of Statistics* 37.1 (2009), pp. 246–270.

[29] M.A. Oliver and R. Webster. "Kriging: a method of interpolation for geographical information systems". In: *International Journal of Geographical Information System* 4.3 (1990), pp. 313–332.

[30] E. Parkhomenko, D. Tritchler, and J. Beyene. "Genome-wide sparse canonical correlation of gene expression with genotypes". In: *BMC proceedings*. Vol. 1. Suppl 1. BioMed Central Ltd. 2007, S119.

[31] E. Parkhomenko, D. Tritchler, and J. Beyene. "Sparse canonical correlation analysis with application to genomic data integration". In: *Statistical Applications in Genetics and Molecular Biology* 8.1 (2009), p. 1.

[32] D. Paul. "Asymptotics of sample eigenstructure for a large dimensional spiked covariance model". In: *Statistica Sinica* 17.4 (2007), p. 1617.

[33] K. Pearson. "Notes on the history of correlation". In: *Biometrika* 13.1 (1920), pp. 25–45.

[34] V. Roth and B. Fischer. "The group-lasso for generalized linear models: uniqueness of solutions and efficient algorithms". In: *Proceedings of the International Conference on Machine Learning (ICML)*. Citeseer. 2008.

[35] A.J. Rothman et al. "Sparse permutation invariant covariance estimation". In: *Electronic Journal of Statistics* 2 (2008), pp. 494–515.

[36] I. Rustandi, M.A. Just, and T.M. Mitchell. "Integrating multiple-study multiple-subject fMRI datasets using canonical correlation analysis". In: *Proceedings of the MICCAI 2009 Workshop*. Citeseer. 2009.

[37] B. Sriperumbudur, D. Torres, and G. Lanckriet. "The sparse eigenvalue problem". In: (2009).

[38] G.J. Szekely and M.L. Rizzo. "Brownian distance covariance". In: *The annals of applied statistics* 3.4 (2009), pp. 1236–1265.

[39] G.J. Székely, M.L. Rizzo, and N.K. Bakirov. "Measuring and testing dependence by correlation of distances". In: *The Annals of Statistics* 35.6 (2007), pp. 2769–2794.

[40] R. Tibshirani. "Regression shrinkage and selection via the lasso". In: *Journal of the Royal Statistical Society. Series B (Methodological)* (1996), pp. 267–288.

[41] C.A. Tracy and H. Widom. "On orthogonal and symplectic matrix ensembles". In: *Communications in Mathematical Physics* 177.3 (1996), pp. 727–754.

[42] H. Tsukahara. "Semiparametric estimation in copula models". In: *Canadian Journal of Statistics* 33.3 (2005), pp. 357–375.

[43] S.A. Van De Geer. "High-dimensional generalized linear models and the lasso". In: *The Annals of Statistics* 36.2 (2008), pp. 614–645.

[44] S. Waaijenborg and A.H. Zwinderman. "Correlating multiple SNPs and multiple disease phenotypes: penalized non-linear canonical correlation analysis". In: *Bioinformatics* 25.21 (2009), p. 2764.

[45] D.M. Witten and R.J. Tibshirani. "Extensions of sparse canonical correlation analysis with applications to genomic data". In: *Statistical applications in genetics and molecular biology* 8.1 (2009), p. 28.

[46] M. Yuan and Y. Lin. "Model election and estimation in the Gaussian graphical model". In: *Biometrika* (2007).

[47] S. Zhou. "Restricted eigenvalue conditions on subgaussian random matrices". In: *Arxiv preprint arXiv:0912.4045* (2009).