Contributions to Effect Size Analysis with Large Scale Data

by

Ming-Chi Hsu

A dissertation submitted in partial fulfillment
of the requirements for the degree of
Doctor of Philosophy
(Statistics)
in The University of Michigan
2014

Doctoral Committee:

Professor Kerby A. Shedden, Chair
Professor Hui Jiang
Professor Naisyin Wang
Professor Ji Zhu

ACKNOWLEDGEMENTS

I am sincerely grateful to my Ph.D. advisor, Prof. Kerby Shedden, for the support and guidance throughout my Ph.D. study. I would also like to thank Prof. Naisyin Wang for her guidance beyond research.

My special thanks go to Prof. Hui Jiang and Prof. Ji Zhu for participating in my dissertation committees in this busiest time.

I am truly indebted to my friends, Toshiya Hoshikawa, Yanming Li, and Xi Xia for their helps and encouragement. In addition, it has been such a pleasure working with people from CSCAR.

TABLE OF CONTENTS

LIST OF FIGURES

Figure

LIST OF TABLES

Table

CHAPTER I

Introduction

1.1 Introduction

Large and complex data are common to the modern life. Adapting existing statistical methods to large/complex data face challenges. These data sets are mines of information, statisticians are now developing new statistical techniques to explore information from them. This dissertation contributes statistical methods that can be used to explore such challenging types of data sets. I will present the results from three projects concerning assessing dissimilarity, functional summaries, and correlation paths. The first project proposes a method to assess the dissimilarity among several effect sizes in a regression analysis. The second and third projects explore the dependence information in a set of variables.

In chapter II, we propose a measure to quantify the degree to which the effects of risk factors differ form each other. Our measure is a nonstandard quantity. The naive plug-in estimate can be used to derive a point estimate. However, the performance deteriorates as the number of predictors grows or as the magnitudes of the effect sizes become similar. Alternative estimates like bootstrap bias corrected estimate and a Bayesian estimate are considered.

When the parameter of interest is not a smooth function of the data, analytic approaches like the delta method cannot be applied. The profile likelihood method

remains well-defined, but computation is difficult. We therefore develop a stochastic search algorithm to approximate the profile likelihood function, and then use the approximation to build a confidence interval. Our simulation results show that the Bayesian estimate outperforms the plug-in and bootstrap estimates, and the coverage rates of the profile likelihood confidence intervals are good as the parameter of interest is not fall close to the boundary value and the information from data is appropriate. Our algorithm can reduce the computation complexity. We apply the procedures to National Health and Nutrition Examination Survey (NHANES) from 2011 to 2012.

In chapter III, we propose a functional summaries to reveal dependence structure in multivariate data. When analyzing high dimensional data, merely calculating the mean and the standard deviation of each component fails to identify many relationships among the variables. Sample correlation and covariance matrices aim to capture the covariance structure in full, but are often too large to be directly interpreted. It is desirable to summarize such a p by p covariance structure in an accessible and easily visualized form. Some existing methods like the effective variance and effective dependence use univariate summaries that allow us to compare different data sets. The "corrgram" is a graphical tool used to display the magnitudes of the data and reorder variables in the correlation matrix such that similar variables are positioned adjacently. However, it can be difficult to extract useful information from a corrgram when the dimension is high.

We propose to summarize the covariance structure in a way that treats the variables anonymously. The proposed functional summaries allow us to visualize the differences in the covariance structures between two data sets, even when they have different dimensions. Our summaries emphasize the degree by which each variable is predictable from the others, with a special focus on the number of variables required to predict another variable. We apply our functional summaries to two gene expression data sets, one consists of 108 normal heart tissue samples from the Cleveland

Clinic Kaufman Center, and the other consists of 734 whole-blood RNA samples from the Estonian Biobank.

In chapter IV, we propose a projection-based approach for exploring conditional correlations. To explore the dependencies among a set of variables, many existing methods use either Pearson correlation coefficients (marginal correlation) or the partial correlation coefficients (the conditional correlation between two variables after removing effects that are due to other variables). There are many other correlation coefficients that can be defined through conditioning. We propose a graphical tool that enables us to explore the change in dependencies from marginal correlations to partial correlations. This path is built via adding information gradually to reach the partial correlation.

The proposed projection-based approach can be applied to another type of conditional correlation matrix - the conditional correlation matrix conditioned on a linear statistic of the data. We can explore the change in correlation matrices when the values of these linear statistics are varied. We apply this approach to a gene expression data set containing 108 normal heart tissue from the Cleveland Clinic Kaufman Center.

CHAPTER II

Assessing the Dissimilarity among Several Effect Sizes in a Regression Analysis

2.1 Introduction

Consider the common statistical problem of modeling the relationship between a response variable Y and its associated predictors (or features) $X_1, X_2, \cdots, X_p$, based on a sample of size n. In regression analysis, we estimate the coefficients that capture the strength of the relationship between each predictor and the response conditioned on the other predictors.

In some settings, the goal is to evaluate the effect sizes of several variables that are strongly believed to have nonzero effects. For example, in a study of health outcomes, we might be interested in the contributions of several risk factors. Specifically, we might wish to quantify the degree to which the effects of risk factors differ form each other. This can be measured as the ratio of the maximum magnitude to the minimum magnitude among the effects, i.e.

$$g(\beta) = \frac{\max_i |\beta_i|}{\min_i |\beta_i|} \vee K, \tag{2.1}$$

with K being a tuning parameter. Point estimation and confidence intervals for this nonstandard quantity are the subject to this chapter.

There exists a naive plug-in estimate $\hat{g} = g(\hat{\beta})$ where $\hat{\beta}$ are the estimated param-

eter values. However, to estimate $g(\beta)$ precisely is challenging since $g(\beta)$ involves estimation of extreme values and also involves a ratio. The performance of the naive plug-in estimate deteriorates when the power to differentiate some effects from zero is small. The denominator can fall close to zero, in which case the ratio goes up. In addition, the plug-in estimate performs poorly when some effect magnitudes are close to the two extreme effect magnitudes. Taking the ratio of extreme values will make the bias upward. The problem is increasingly severe as p grows. Clipping the estimate using the tuning parameter in (2.1) substantially resolves this issue.

When the ultimate goal is to estimate g precisely, we use mean squared error (MSE) or root mean squared error (RMSE) to judge the performance. The MSE can be partitioned into the sum of the squared bias and the variance of estimator. To achieve good estimation performance, both of these quantities need to be as small as possible. However, there is a trade-off between bias and variance. The decrease in bias generally causes an increase in variance and vise-versa. Therefore, a bias correction procedure might or might not be preferred to a naive estimate.

There is no obvious procedure to build an interval estimate for g. Unlike the point estimation setting where we can use the natural plug-in estimate of g, the information for interval estimation for each coefficient β_j cannot be employed to build an interval estimate for $g(\beta)$. In addition, standard analytic techniques such as linearization do not apply here. This is because g is not smooth.

The question studied here is distinct from well-studied classical questions involving ratio estimation or extreme values. When a linear regression model is used, the ratio estimation is related to calibration problems, the Fieller-Creasy problem (cf. Fieller, 1954; Creasy, 1954), slope-ratio assay, parallel-line assay, and bioequivalence. By using an orthogonal parameterization, Ghosh et al. (2003) present a Bayesian analysis using objective priors. Ghosh et al. (2006) showed that approaches based on the profile likelihood and modifications can result a interval that is infinitely wide. Bebu

et al. (2009) use the generalized confidence interval and shown good performance of coverage probabilities and their procedure can be implemented for slope-ratio assays, and parallel-line assays under a probit model. The general ratio estimation does not involve the extreme values. To our setting, both the numerator and the denominator of g are extreme values. In addition, extreme values in our setting are different from the setting of extreme value theory (cf. Gumbel, 2004) where the goal is to assess the behavior on a sequence of samples.

In section 2, we discuss procedures for point estimation. In addition to the plug-in estimate, alternative approaches like bootstrap bias correction and the approximate Bayesian computation (ABC) are discussed. In section 3, we build interval estimates for $g(\beta)$. In section 4, we use simulation studies to discuss the properties. In section 5, we apply our procedures to a study of health risk factors.

2.2 Point Estimation

The naive estimate for g is the plug-in estimate $\hat{g} = g(\hat{\beta})$ where $\hat{\beta}$ are the estimated coefficient values from the multiple regression model. This estimator for $g(\beta)$ can be quite biased. Reducing the bias has the potential to decrease the MSE, which is our goal. Since the target of interest g involves extreme values, g is not differentiable. Techniques based on the delta method (cf. Oehlert, 1992) therefore cannot be used here.

2.2.1 Bias Correction

The Bootstrap method (Efron, 1979) can be used to do the bias correction, although it is unclear if this is theoretically supported in the case of extreme values. For ith bootstrap sample of size n, $i = 1, \cdots, B$, we let $\tilde{\beta}^{(i)}$ be the estimate for β and estimate $g(\beta)$ using the plug-in estimate $g(\tilde{\beta}^{(i)})$. The bootstrap based approximation

to bias is

$$\widehat{Bias} = \frac{1}{B}\sum_{i=1}^{B} g(\tilde{\beta}^{(i)}) - g(\hat{\beta}).$$

The bootstrap bias corrected estimator for $g(\beta)$ is $g(\hat{\beta}) - \widehat{Bias}$ which is equivalent to

$$2g(\hat{\beta}) - \frac{1}{B}\sum_{i=1}^{B} g(\tilde{\beta}^{(i)}).$$

This procedure is widely used for bias correction of smooth estimators. We apply it here to the parameter of interest in which g involves a more complex form.

The bootstrap bias correction generally increases the variance even though it may reduce the bias. The behavior in terms of RMSE will depend on the trade-off between bias and variance.

2.2.2 Approximate Bayesian Computation

2.2.2.1 Estimating the Dissimilarity among Effect Sizes

In this section, a method called approximate Bayesian computation (ABC) (Rubin (1984), Diggle and Gratton (1984), Tavare et al. (1997), and Sunnaker et al. (2013)) is used to construct the estimate. ABC can be used to improve the performance of an arbitrary estimator $g(\hat{\beta})$ of the parameter $\theta \equiv g(\beta)$. The basic idea is to replace $g(\hat{\beta})$ with

$$Q[\theta \big| g(\hat{\beta}) = g(\hat{\beta}_{obs})], \tag{2.2}$$

where a prior is specified for the unknown parameter θ, $\hat{\beta}_{obs}$ is the observed regression estimate from (Y, X), and Q is a numerical summary of the distribution (e.g. mean, median, or mode). If $g(\hat{\beta})$ is a sufficient statistic for θ, this is equivalent to $Q[\theta|Y, X]$, the standard Bayesian estimation (cf. Gelman et al., 2004).

The specification of a prior distribution is described as follows. The empirical

Bayes method is used to obtain the prior distribution for θ. The plug-in estimate of the sampling distribution

$$N_{p+1}(\hat{\beta}, s^2(X^T X)^{-1}), \tag{2.3}$$

is our prior distribution for β and this implies the prior distribution for θ. This specified prior distribution in (2.3) should place non-negligible density around the true value of θ.

2.2.2.2 Implementation Details

To approximate the posterior $g|g(\hat{\beta})$ in (2.2), we generate samples $\tilde{D} = (\tilde{Y}, X)$ from the model

$$\tilde{Y} = X\beta_{pr} + \sigma_\tau e,$$

where β_{pr} is drawn from the prior distribution, X are the observed covariates, e is standard Gaussian noise, and σ_τ indicates the variability of error terms in the model which equals to the maximum likelihood estimator (MLE) of the error standard deviation using the observed data.

For each combination of β_{pr} and $\tilde{D}$, we have an ordered pair $(\grave{g}, \tilde{g})$, where $\grave{g} \equiv g(\beta_{pr})$ and $\tilde{g}$ is the plug-in estimate from regressing $\tilde{Y}$ on X. If the generated $\tilde{g}$ values are close to the $g(\hat{\beta}_{obs})$, the corresponding $\grave{g}$ values are used to approximate the posterior distribution. In other words, we use

$$\{\grave{g} \big| |\tilde{g} - g(\hat{\beta}_{obs})| \leq \delta\} \tag{2.4}$$

with a small δ value as an empirical counterpart to the distribution of $g|g(\hat{\beta}) = g(\hat{\beta}_{obs})$.

In our study, we specify Q as the median function, and use the empirical median of the set $\{\grave{g}||\tilde{g} - g(\hat{\beta}_{obs})| \leq \delta\}$ as an estimator for $g(\beta)$.

2.3 Interval Estimation

In this section, we construct the interval estimate for $g(\beta)$. Using standard techniques, confidence intervals for smooth functions of the parameters can be easily built. However, since g is not differentiable, we cannot apply analytic approaches like the delta method in the present setting.

The parameter of interest is a function of the parameters, Venzon and Moolgavkar (1988) used the profile likelihood to construct a confidence interval for parameter of interest. The idea is to invert a series of likelihood ratio tests to obtain a confidence interval for the parameter of interest.

Let $\{(y_i, x_i), i = 1, \cdots, n\}$ be the observed variables with density function $f(x, y; \beta)$. The corresponding log-likelihood based on $\{(y_i, x_i), i = 1, \cdots, n\}$ is

$$l(\beta) = \sum_{i=1}^{n} \log f(y_i, x_i; \beta) \equiv \log f(Y, X; \beta). \tag{2.5}$$

Since we are interested in $g(\beta) = \frac{\max_i |\beta_i|}{\min_i |\beta_i|} \vee K$, we can profile g to obtain a log-likelihood function

$$l^*(g) = \sup_{\beta: g(\beta)=g} \sum_{i=1}^{n} logf(y_i, x_i; \beta) \equiv \sup_{\beta: g(\beta)=g} logf(Y, X; \beta). \tag{2.6}$$

From the invariance property of the MLE (Casella and Berger, 2001), the MLE for g in (2.6) will be $g(\hat{\beta})$ with $\hat{\beta}$ being the MLE from (2.5). Although this is not a regular setting, motivated by statistical theory for regular likelihood ratio tests (LRT), the interval for $g(\beta)$ can be built using

$$2(l^*(\hat{g}) - l^*(g)). \tag{2.7}$$

In a regular setting, the LRT $2(l^*(\hat{g}) - l^*(g))$ in (2.7) is approximately χ^2_1 distributed

when the sample size n is large. The 95% confidence interval thus consists of all the values of g for which

$$\{g \big| l^*(g) \geq l^*(\hat{g}) - 1.92\}.$$

Here we use the χ^2_1 reference distribution, although it may not be asymptotically correct. We explore the implications of this choice using simulation.

The value of the profile log-likelihood of l^* at $g = \hat{g}$ in (2.7) is known. However, the nonlinear constraint makes the computation of $l^*(g) = \sup_{g(\beta)=g} \log f(Y, X; \beta)$ at $g \neq \hat{g}$ difficult. Therefore, we use a stochastic search algorithm (Spall, 2003) to approximate $l^*(g)$.

The basic idea is to generate many β values and use the local maximum $\max\{l(\beta)|g(\beta) \approx g\}$ as the approximation for the value of l^* at g. As an illustration, we sample 100 observations from $Y = 2X_1 + X_2 + \epsilon$ with $X_1 \sim N(0,1)$, $X_2 \sim N(0,1)$, and $\epsilon \sim N(0,1)$. In Figure 2.1, the red line is the true log-likelihood function and the dots are the likelihood values evaluated at sampled β values. In this example, the distribution used to draw β is $N_{p+1}(\hat{\beta}, s^2(X^TX)^{-1})$ which is same as the prior distribution used in ABC. We know that all likelihood values from the sampled β are below the true likelihood function. As long as we can draw sufficiently many β, the pointwise local maximum can be used to approximate $l^*(g)$ in (2.6).

The distribution $N_{p+1}(\hat{\beta}, s^2(X^TX)^{-1})$ tends to oversample values of g that are close to the plug-in estimate, and therefore well inside the confidence interval. We can improve the algorithm by modifying the distribution used to sample β. A random convex combination of $N_{p+1}(\hat{\beta}, s^2(X^TX)^{-1})$ and a point mass at β_c chosen such that $g(\beta_c) = 1$ can enable us to obtain more g values that are around the lower bound of the interval. That is, we set sampled β equal to

$$\lambda * \beta_c + (1 - \lambda) * \beta_{sd}, \tag{2.8}$$

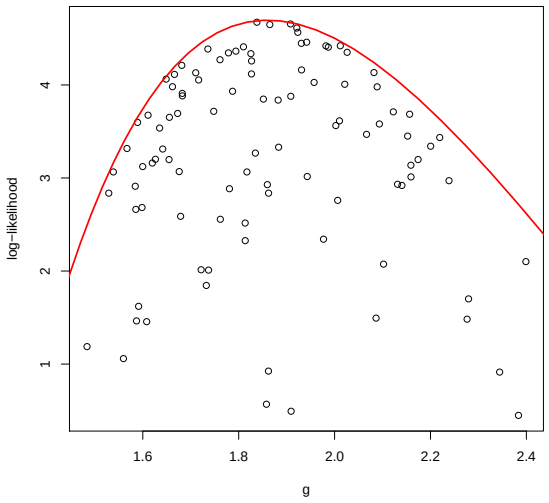


Figure 2.1: An illustration showing how the profile likelihood function is approximated using a stochastic search algorithm.

where β_c is derived from the constrained likelihood under $g(\beta_c) = 1$, the distribution of β_{sd} is $N_{p+1}(\hat{\beta}, s^2(X^TX)^{-1})$, and λ follows a uniform distribution in $(0, 1)$.

Another way to improve the performance of the algorithm is to precede in two stages. We first draw β from a given distribution, e.g. the random convex combination discussed above. Specifically, let $\check{\beta} = (\check{\beta}_0, \check{\beta}_1, \cdots, \check{\beta}_p)^T$ be the sampled β and we split it into two components, the intercept component $\check{\beta}_{(1)} = (\check{\beta}_0, 0, \cdots, 0)^T$ and the slope component $\check{\beta}_{(-1)} = (0, \check{\beta}_1, \cdots, \check{\beta}_p)^T$. Next, we maximize $l(a\check{\beta}_{(1)} + b\check{\beta}_{(-1)})$ over $a, b \in \mathbf{R}$, which is a least squares problem that yields $\tilde{\beta}$. Note that $l(\tilde{\beta}) \geq l(\beta)$ and $g(\tilde{\beta}) = g(\beta)$. This algorithm yields points that are closer to the profile likelihood function. By doing this, we use the slope component of $\check{\beta}$ as a direction constraint and find the corresponding constrained maximum log-likelihood. Figure 2.2 is the contour plot of log-likelihood for the example used in Figure 2.1. The log-likelihood is optimized for any given (β_1, β_2). We draw a line with $g(\beta) = 2$ and the labeled point is the local maximum. For a sampled β with $g(\beta) = 2$, its log-likelihood $l(\beta)$ must be less than or equal to the log-likelihood at the labeled point. In Figure 2.3, we displays the scatterplot of log-likelihood using either $l(\beta)$ or $l(\tilde{\beta})$. The improvement

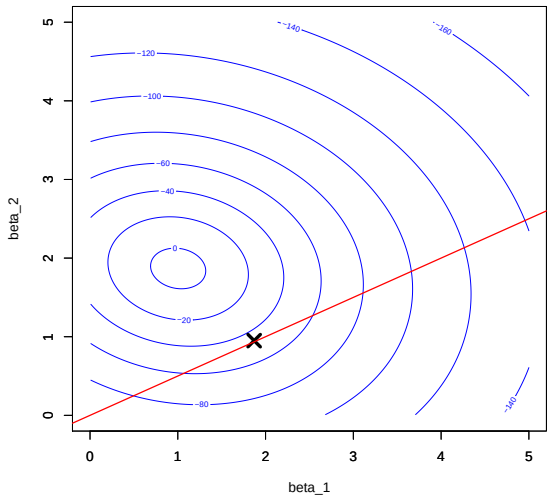


Figure 2.2: An illustrative example showing how searching direction improves the computation.

via using $l(\tilde{\beta})$ to replace $l(\beta)$ is pretty good here. While p increases, we still see the improvement but the difference between $l(\beta)$ and $l(\tilde{\beta})$ diminishes. In addition, the computation complexity is reduced.

2.4 Simulation Study

In this section, we first explore the behavior of our Bayesian estimate and then use simulation studies to demonstrate the performance of our procedures on point estimation and interval estimation.

2.4.1 Study I: Behavior of ABC

In this section, we show that the ABC procedure using an empirical Bayes approach to set the prior provide a shrinkage estimator, and the procedures using a prior that is not data-dependent approximate the plug-in estimate.

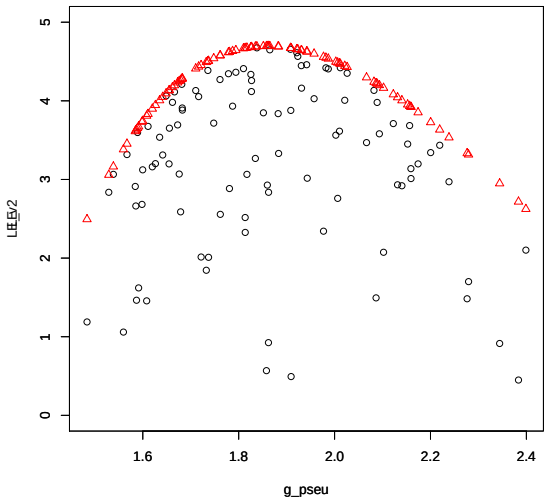


Figure 2.3: Comparing likelihood value between $l(\beta)$ (black circle) and $l(\tilde{\beta})$ (red triangle).

We generate n samples from the linear model

$$Y = X\beta_{pr} + \epsilon, \tag{2.9}$$

where $X \sim N_p(0, I_p)$, $\epsilon \sim N(0, \sigma^2)$, $\beta_{pr} \sim N(\beta, \sigma^2(X^TX)^{-1})$, and $\sigma^2 = \frac{1-r^2}{r^2}|\beta|_2^2$. We let $r^2 = 0.8$, $p \in \{3, 5, 7\}$, and the components of β are equally spaced from 1 to $g(\beta)$ with $g(\beta) \in \{2, 3, 4, 6\}$. For example, if g is 2 and p is 5, β will be $(1, 1.25, 1.5, 1.75, 2)^T$. The range for the sample size n is {100, 200, 400, 800}. For each scenario, 100,000 replicates are generated, and the local estimate is used to find the posterior median Q.

Figure 2.4 displays the relationship between the plug-in estimate and the estimate from ABC procedures under $p = 3$ and $n = 100, 800$. When the sampling distribution of $\hat{\beta}$ is used to set the prior, this empirical Bayes structure results an estimate that shrinks toward $g(\beta)$. With this informative prior, the Bayes estimate is preferred. As n increases, the range for $g(\hat{\beta})$ will be narrower. This is because the variability of $g(\hat{\beta})$ decreases as sample size increases. The results for other scenarios are similar.

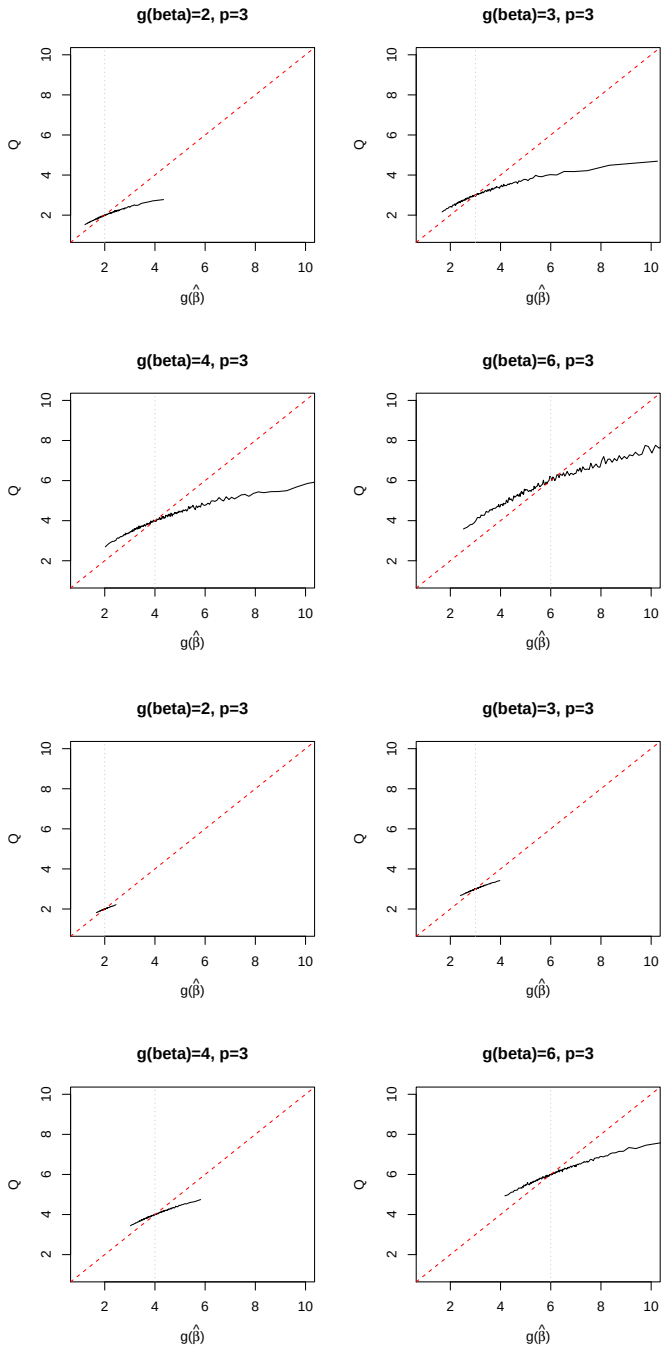


Figure 2.4: The relationship between plug-in estimate and estimate from ABC procedures. The upper two rows are with sample size 100, and the others are with sample size 800.

Next, we let the distribution of β_{pr} be more diffuse to demonstrate the behavior of ABC procedure under a prior that is not data-dependent. We let the prior distribution for β_{pr} be

$$N_p(\beta, k\frac{\sigma^2}{n}I_p)$$

with $k \in \{2, 3, 4, 5\}$. Figure 2.5 displays the results with $k = 2$, 5 and $n = 100$. As k increases, the prior is more diffuse. The prior information becomes less informative. We find that the difference between the plug-in estimate and the Bayes estimate will become small as k increases. We have similar conclusions for other scenarios.

2.4.2 Study II: Performance Evaluation

We generate n samples from the linear model

$$Y = X\beta + \epsilon, \tag{2.10}$$

where $X \sim N_p(0, \Sigma)$ and $\epsilon \sim N(0, \sigma^2)$. We let $p \in \{3, 5, 7\}$, and the coefficient vector β has equally spaced values ranging form 1 to g. The covariance structure Σ can be either $\Sigma_{i,j} = 0$, $\Sigma_{i,j} = 0.3^{|i-j|}$, or $\Sigma_{i,j} = 0.6^{|i-j|}$ which are denoted as $AR(0)$, $AR(0.3)$, and $AR(0.6)$, respectively. The value of σ^2 is controlled by specifying $r^2 \in \{.2, .4, .6, .8, .9, .95\}$ and the relationship between σ^2 and r^2 for given Σ and β is $\sigma^2 = \frac{1-r^2}{r^2}\beta^T\Sigma\beta$.

Since the covariance structure for the sampling distribution of $\hat{\beta}$ is $\frac{1-r^2}{r^2n}(\beta^T\Sigma\beta/(X^TX/n))^{-1}$, some population structures can yield identical sampling distributions of $\hat{\beta}$. Table 2.1 shows the choices of (n, r^2) that have the same sampling distribution provided that X^TX/n are the same. Therefore, the sample size n is fixed at 100 for the simulation studies.

For the ABC procedures, we generate 100,000 prior β values for each (Y, X) and use the closest 1% to build the posterior distribution to estimate $g(\beta)$. To build the

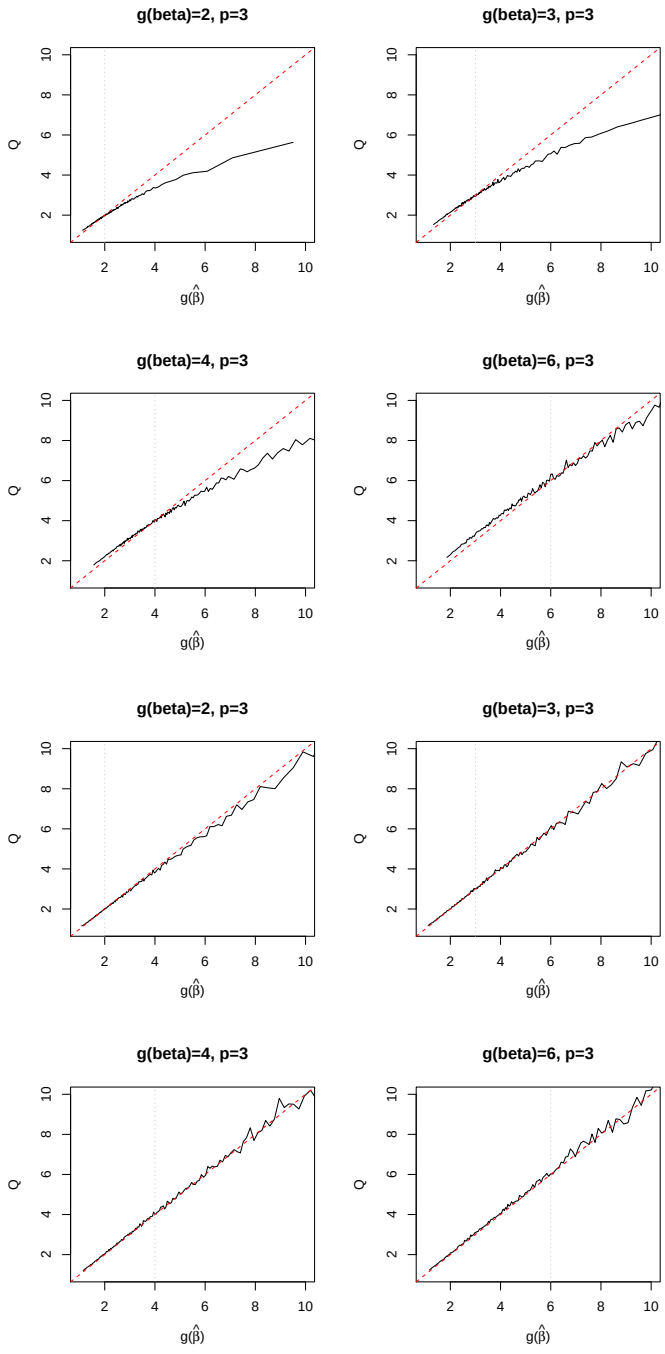


Figure 2.5: The relationship between the plug-in estimate and the estimate from ABC procedures with more diffuse prior. The upper two rows are with $k = 2$, and the others are with $k = 5$.

	r^2				
r^2_{100}	0.1	0.2	0.3	0.4	0.5
0.2	225	100	54	38	25
0.4	600	267	156	100	67
0.6	1350	600	350	225	150
0.8	3600	1600	934	600	400
0.9	8100	3600	2100	1350	900
0.95	17100	7600	4434	2850	1900

Table 2.1: Population structures yielding identical sampling distributions for $\hat{\beta}$. A sample size of $n = 100$ with $r^2 = r^2_{100}$ is equivalent to the sample sizes given in the table for the five specified values of r^2. The sampling distributions are equivalent when $X^T X/n$ is equal in the two populations.

profile interval, we also sample 100,000 β values.

To begin with, we demonstrate the performance of three procedures using different tuning parameter K. Figure 2.6 displays the difference in RMSE and bias under different K. The values inside the parenthesis indicate the value of K used, and no parenthesis indicates $K = \infty$. It shows that the plug-in estimate (PI) and the bootstrap bias corrected estimate (BS) are sensitive to K, and the Bayesian posterior median (PM) is less affected or unaffected by K. When r^2 increases, the influence from K diminishes. When the information from the data is low, clipping the estimate using tuning parameter reduces the bias. Through the end of this chapter, we let $K = 20$.

Figure 2.7 to Figure 2.15 display the performance for point estimation. Each figure contains 6 graphs that plot the RMSE and bias for three procedures under a specific g, and different figures represent different choices of p and Σ. The solid line is used to represent the RMSE over 400 replicates, and the dashed line shows the bias. We use black color to indicate the results for plug-in estimate, red color for bootstrap bias corrected estimate, and blue color for posterior median. In some figures (e.g. $6th$ graph in Figure 2.7), the bias and the RMSE for the plug-in estimate are not monotone decreasing function of r^2 which is counterintuitive. The reason is because the proportion of minimum magnitude $\min_i |\beta|_i$, the denominator of g, that

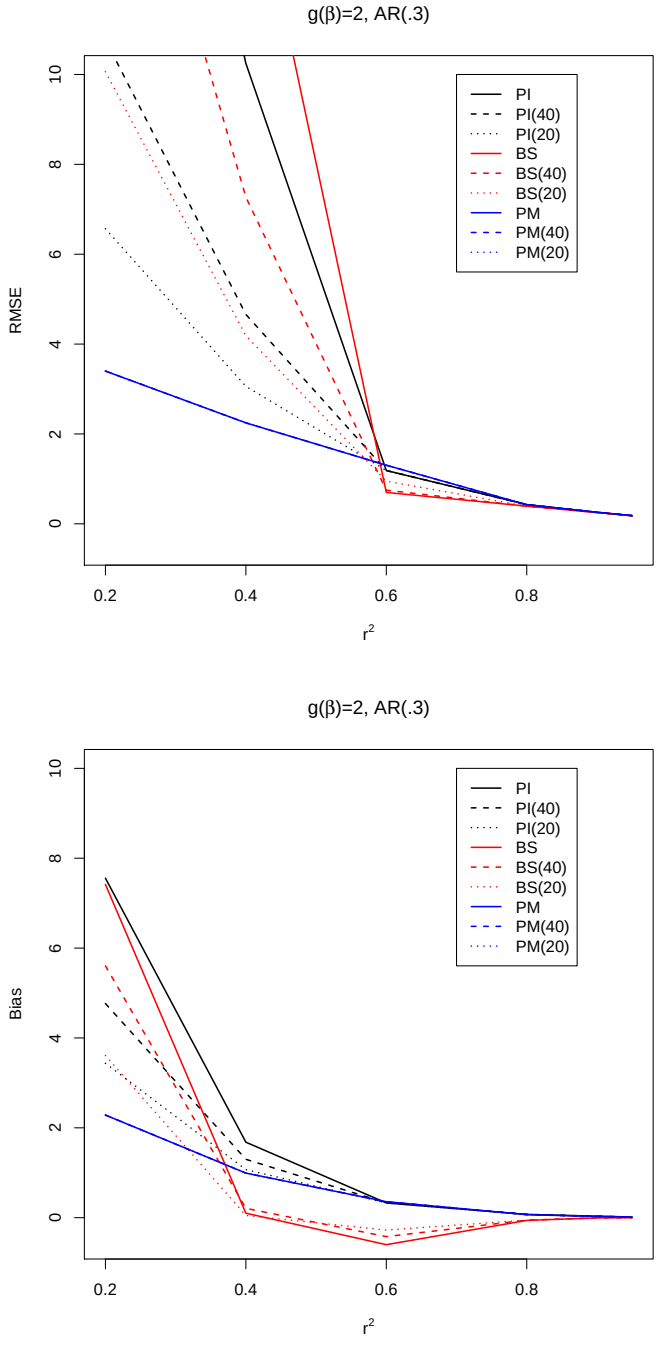


Figure 2.6: The RMSE and bias under different clipping values given $g(\beta) = 3$ with $\Sigma_{ij} = .3^{|i-j|}$. The value inside the parentheses indicates the clipped value.

are around zero will increase first and then decrease as r^2 increases.

Both the plug-in estimate and Bayesian posterior median are biased toward positive values. For the bootstrap bias corrected estimate, they are positively biased for $g(\beta) = 1$ and can be negatively biased for other $g(\beta)$ values. The bootstrap bias corrected estimate is generally less bias than the other two approaches. However, the trade-off between bias and standard deviation makes its RMSE larger than the others, which makes the bootstrap bias corrected estimate less favorable.

All the estimates have their bias move toward 0 eventually as r^2 increases. When r^2 is small, the RMSE can be larger than $g(\beta)$, which makes the estimate become less useful. While r^2 is large, the difference in RMSE among three procedures becomes small. Since the goal is to estimate g precisely, we use RMSE as a criterion to judge the performance. The Bayesian posterior median is preferred since its RMSE is generally better than the other two approaches.

Next, we compare the performance on interval estimation via using χ^2_1 as reference distribution. Figure 2.16 displays the coverage rate when $p = 3$. When $g = 1$, the coverage rates for the 95% confidence intervals are underestimated. The coverage rates for the other g values will move toward 95% as r^2 increases.

Figure 2.17 displays the coverage rate when $p = 5$. The coverage rates for $g = 1$ are underestimated and the values are lower than those with $p = 3$. When $g(\beta) = 1.5$, the coverage rates are underestimates for small r^2 and the coverage rates will move toward 95% as r^2 increases. With large $g(\beta)$, the coverage rates tend to be overestimated with smaller r^2, but will move toward 95% eventually.

The results for $p = 7$ are in Figure 2.18. While $g = 1$, the coverage rates are far smaller than those with $p = 5$. When g is small but $g \neq 1$, the corresponding coverage rates are underestimated for small r^2, and will increase and move toward 95% as r^2 increases. When g is larger, the coverage rate will overestimate at small r^2, and move toward 95% as r^2 increases. That is, the coverage rate will move toward

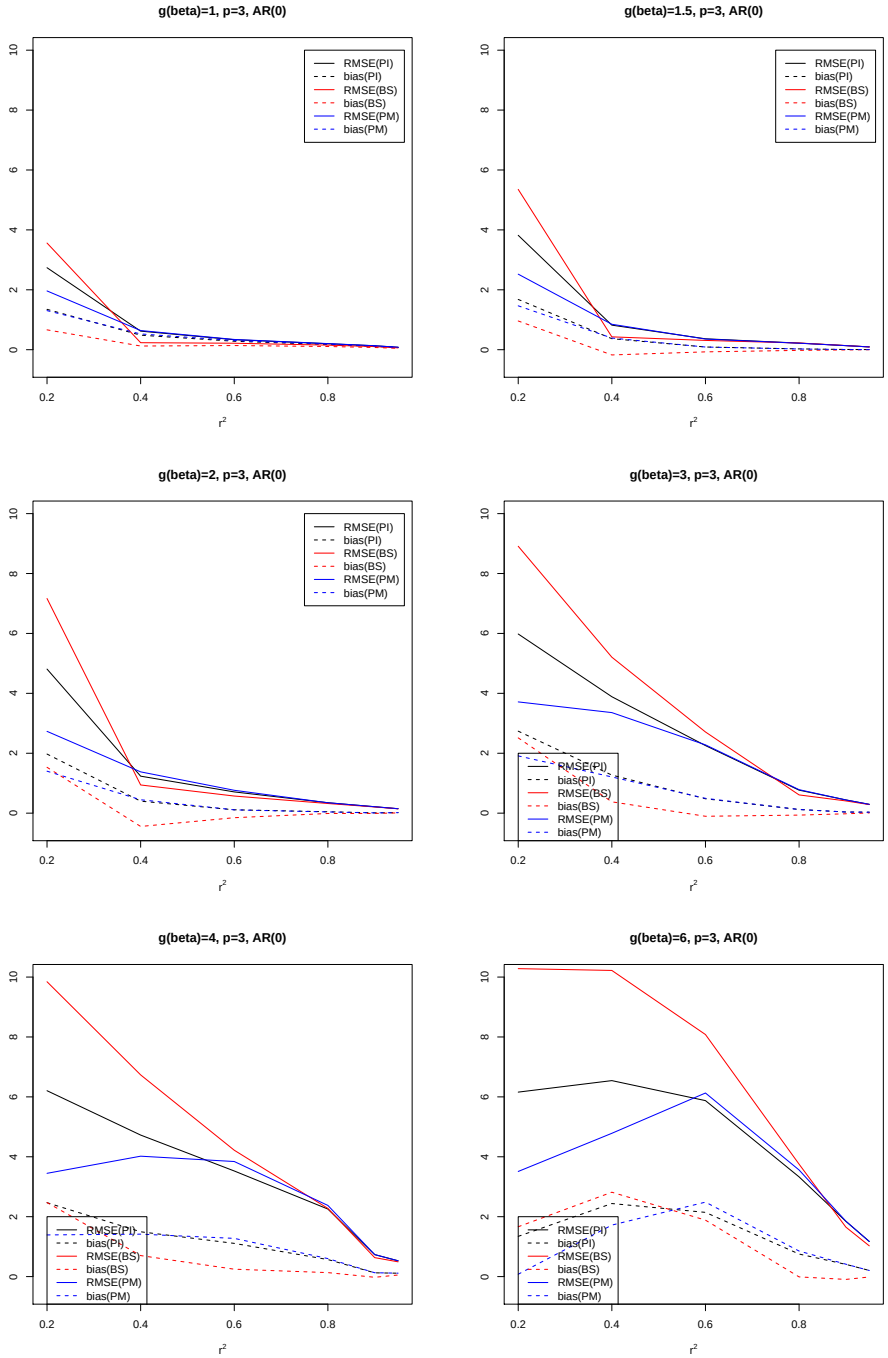


Figure 2.7: The RMSE and the bias under $p = 3$ and $\Sigma_{ij} = 1_{[i=j]}$

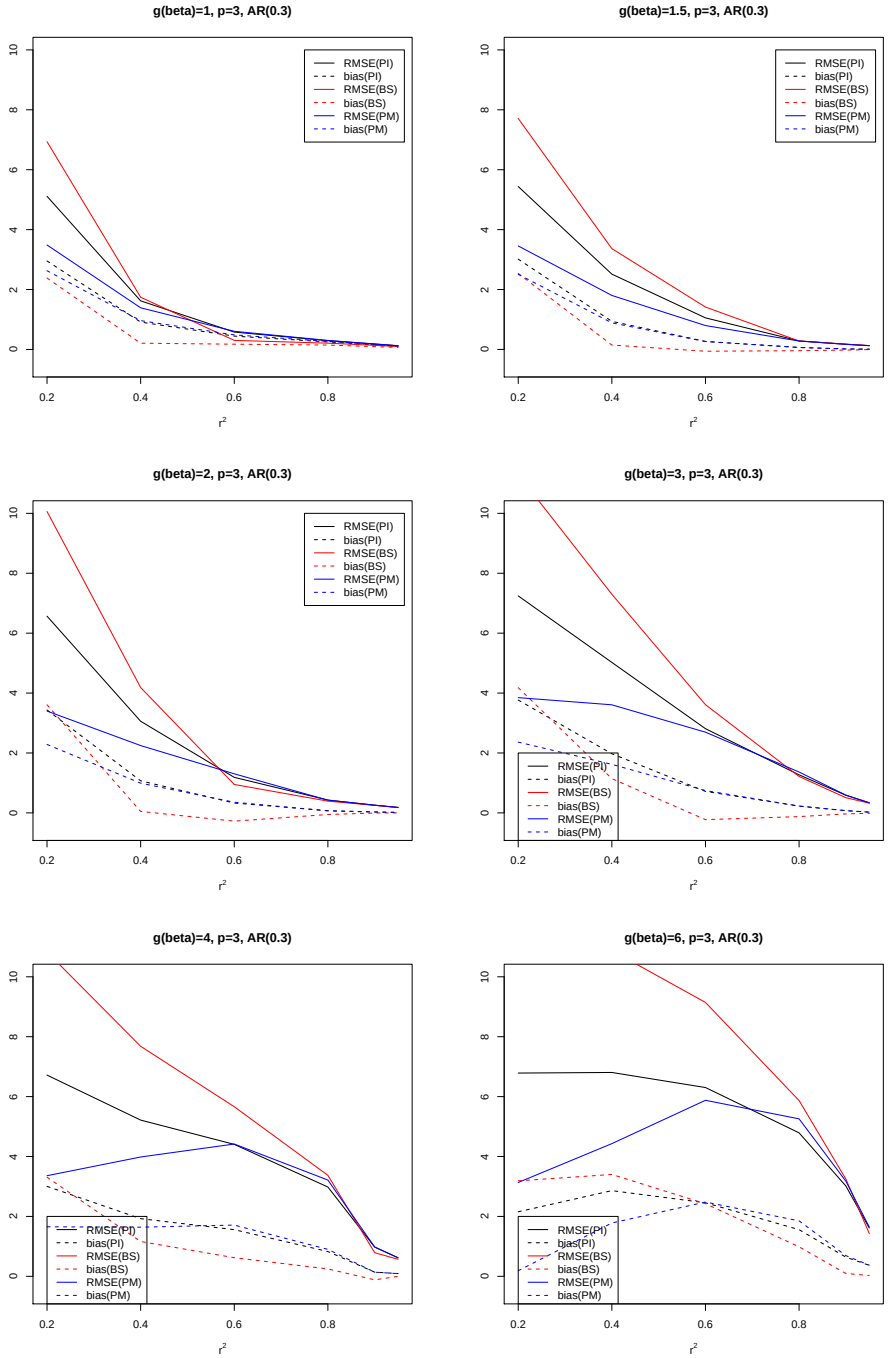


Figure 2.8: The RMSE and the bias under $p = 3$ and $\Sigma_{ij} = .3^{|i-j|}$

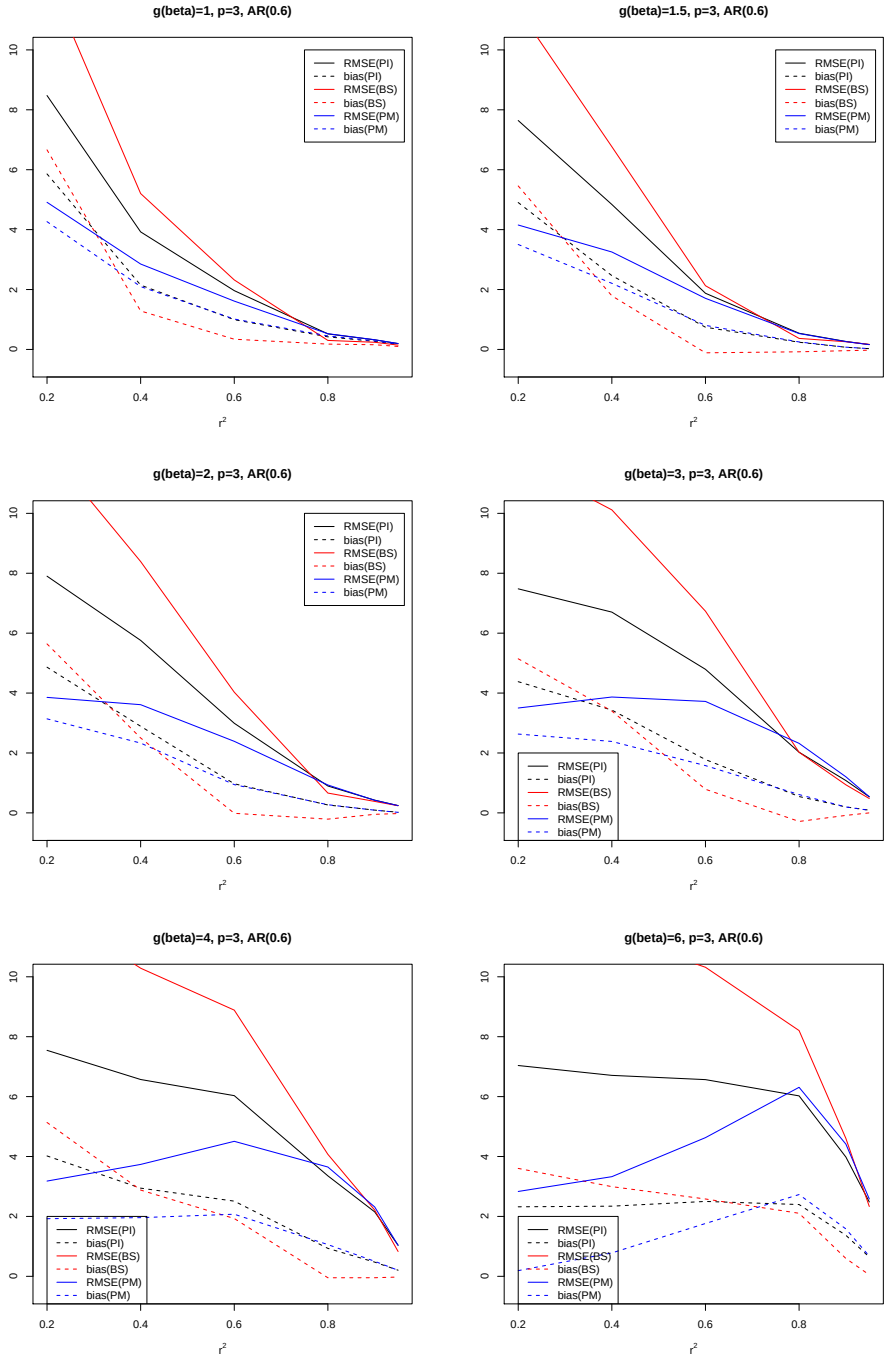


Figure 2.9: The RMSE and the bias under $p = 3$ and $\Sigma_{ij} = .6^{|i-j|}$

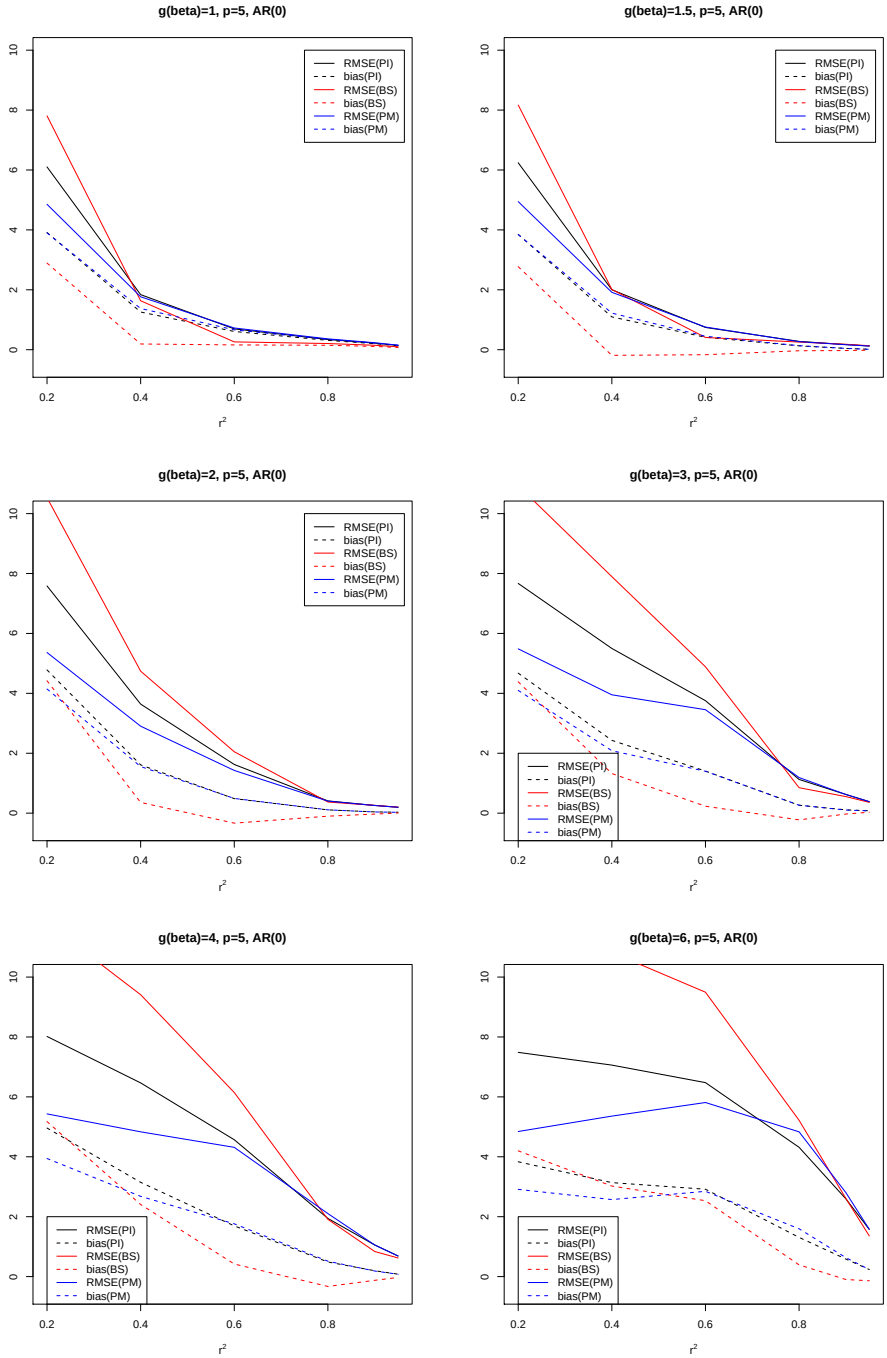


Figure 2.10: The RMSE and the bias under $p = 5$ and $\Sigma_{ij} = 1_{[i=j]}$

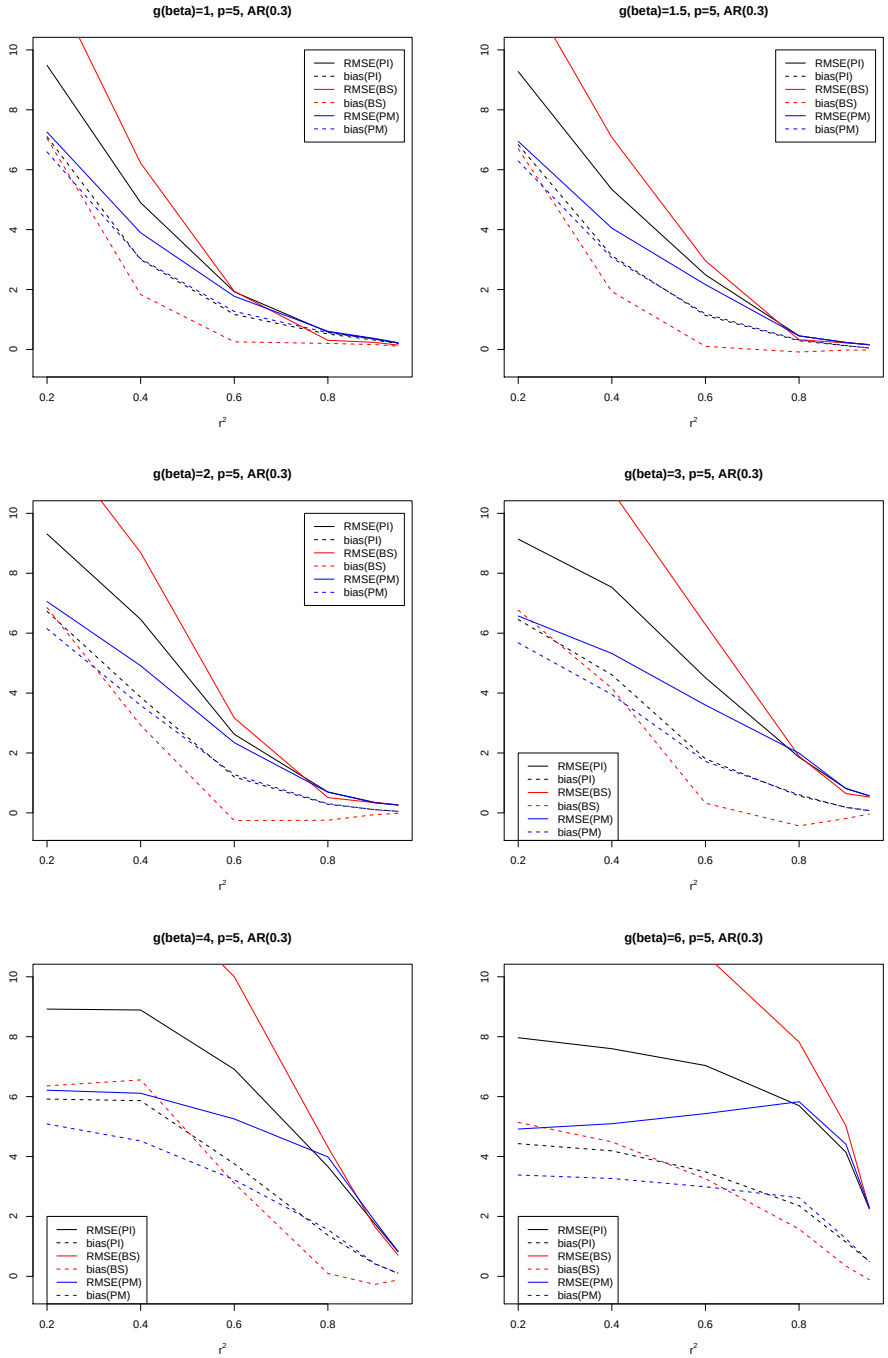


Figure 2.11: The RMSE and the bias under $p = 5$ and $\Sigma_{ij} = .3^{|i-j|}$

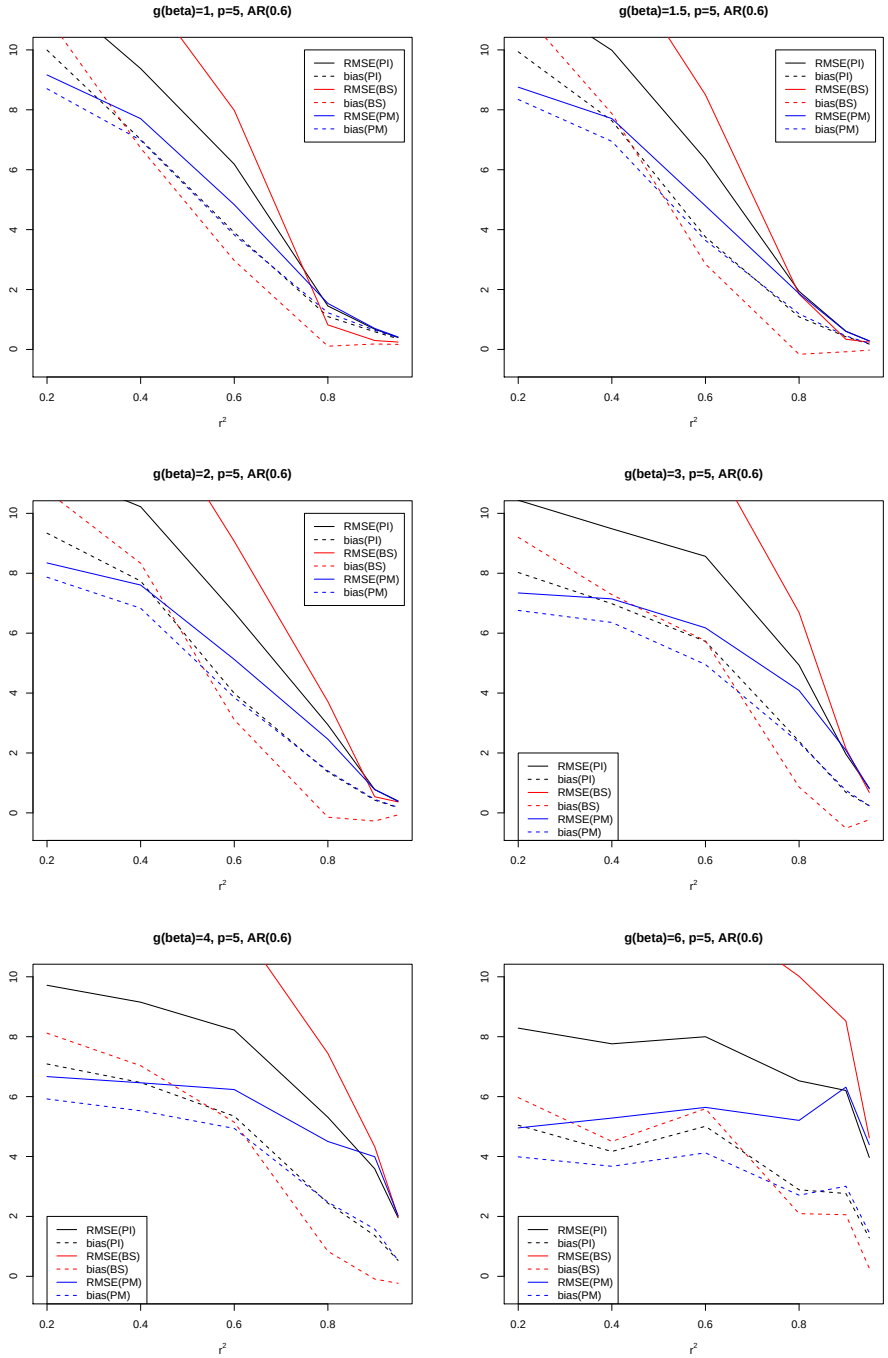


Figure 2.12: The RMSE and the bias under $p = 5$ and $\Sigma_{ij} = .6^{|i-j|}$

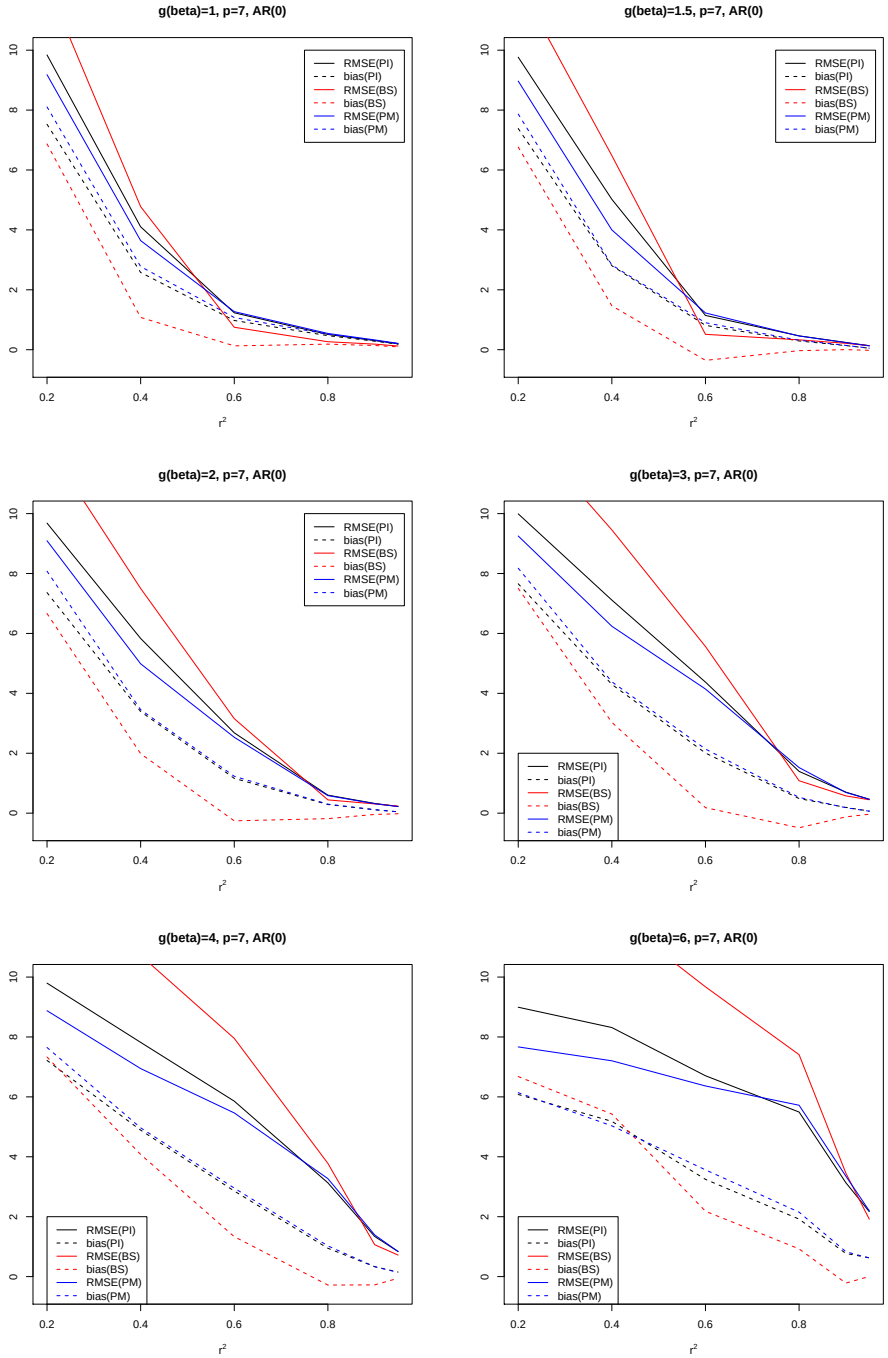


Figure 2.13: The RMSE and the bias under $p = 7$ and $\Sigma_{ij} = 1_{[i=j]}$

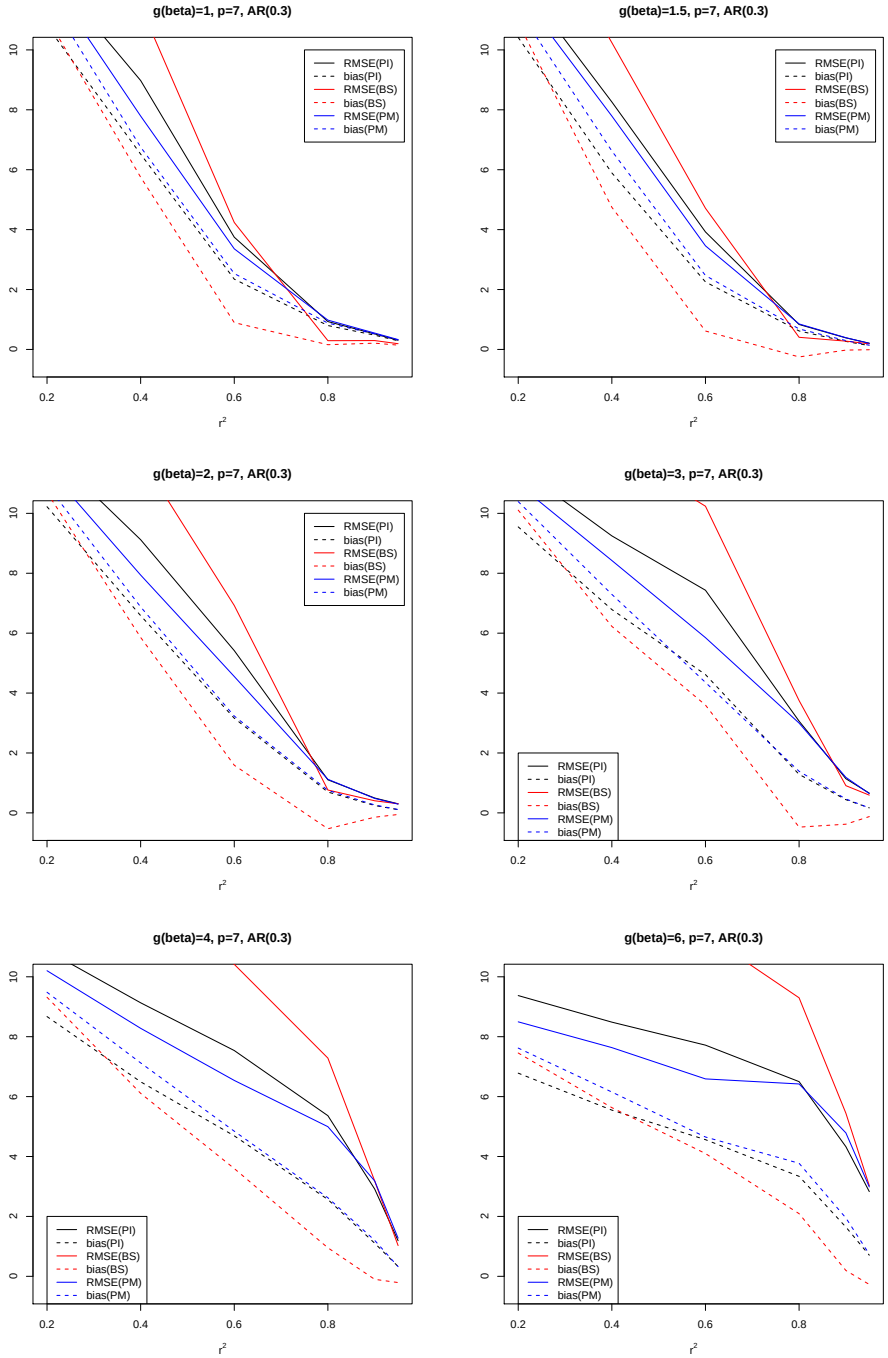


Figure 2.14: The RMSE and the bias under $p = 7$ and $\Sigma_{ij} = .3^{|i-j|}$

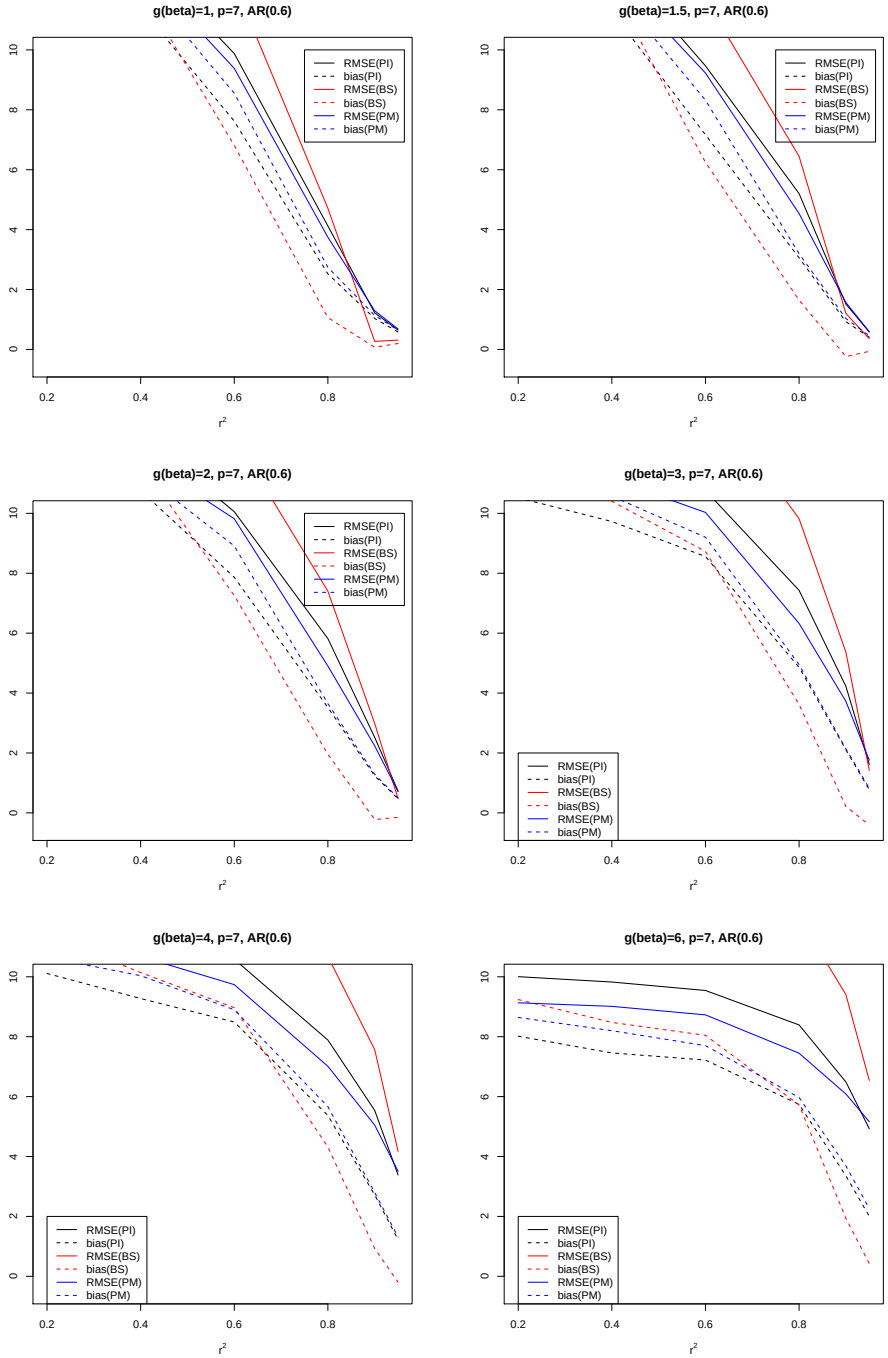


Figure 2.15: The RMSE and the bias under $p = 7$ and $\Sigma_{ij} = .6^{|i-j|}$

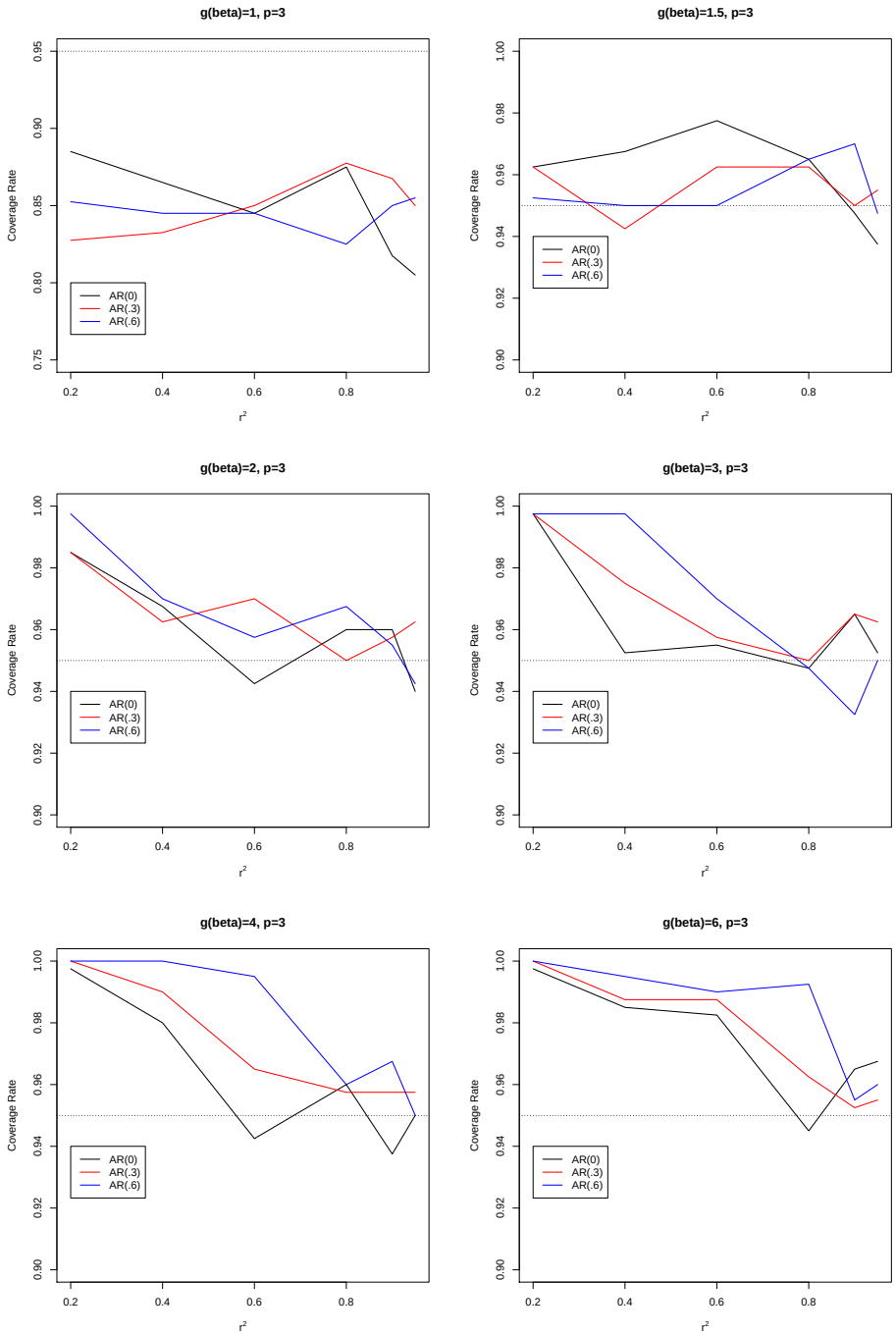


Figure 2.16: The coverage rate for the interval estimation under $p = 3$.

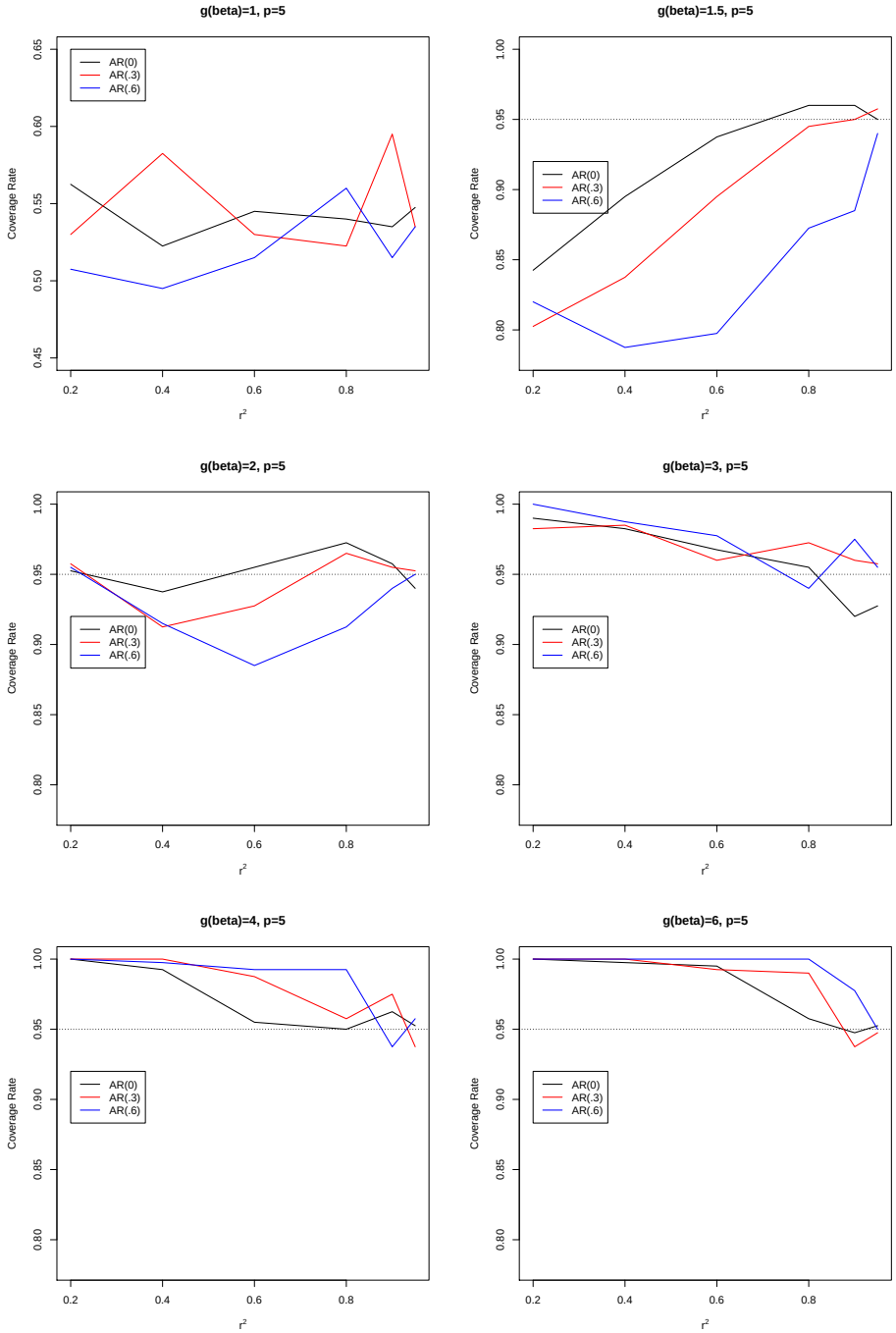


Figure 2.17: The coverage rate for the interval estimation under $p = 5$.

95% as r^2 increases provided that g is not too close to 1.

To explore why the coverage for large g with small r^2 always overestimates, we examine the length of the profile confidence interval. Since the upper bound can be large, we set an upper bound of 50 for the profile confidence intervals. We find that for a given r^2 the average length or the median length of confidence interval under a larger g tend to be higher. For example, when $r^2 = .8$ and $\Sigma = I_5$, the median and the mean length of the profile intervals under $g = 6$ are 36.88 and 29.05, respectively. Under $g = 1.5$, the corresponding median and average lengths are 1.105 and 2.207, respectively. As r^2 increases, the lengths decrease. When r^2 is smaller, it is more likely to observe a profile interval with upper bound exceeds 50.

When $g \approx 1$, we found that the performance from the confidence interval is not good. If the goal is to apply hypothesis testing on g, the likelihood ratio test (LRT) can be carried out easily for $\mathrm{H}_0 : g(\beta) = 1$. For this LRT, the degree of freedom will be $p - 1$. We apply the LRT to different choices of p, Σ, and r^2, and the results for 54 (3*3*6) coverage rates are ranging from 0.919 to 0.964 which are pretty good. However, when the null value for $g(\beta)$ is not 1, the computation of the maximum likelihood value under the null hypothesis is difficult.

2.5 Application to a Study of Health Risk Factors

In this section, we apply our procedures to the National Health and Nutrition Examination Survey (NHANES) from 2011 to 2012. The response of interest is the mean arterial pressure (MAP) and the predictors are age, gender, and body mass index (BMI). We apply our procedures to estimate the dissimilarity among standardized effect sizes of age, gender, and BMI. Since the unusual blood pressures for young people are more likely caused by diseases or other factors, we exclude subjects who are less than 18 years old. The sample size for this study is 4938.

We fit a linear regression model, and the coefficients for standardized age, gender,

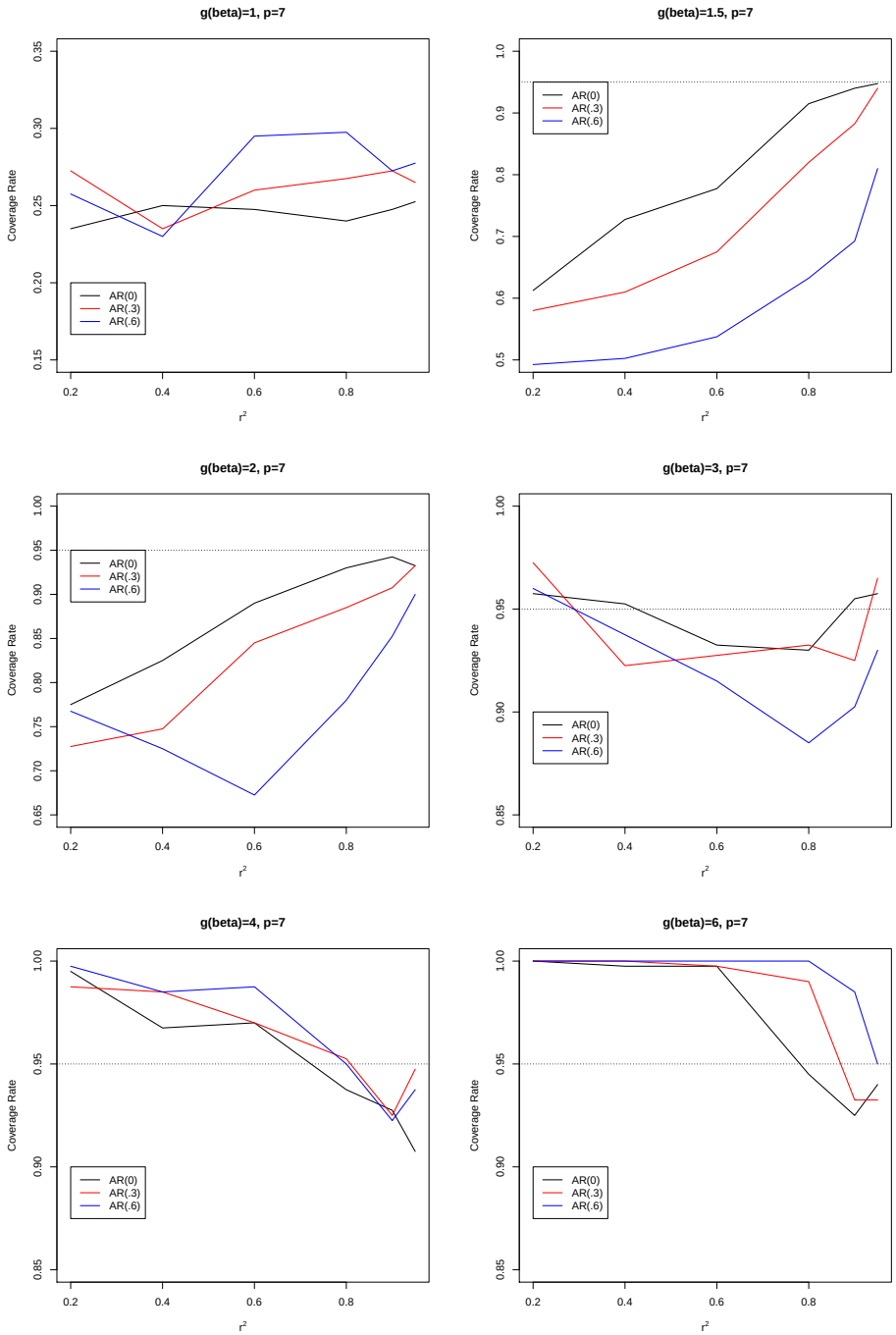


Figure 2.18: The coverage rate for the interval estimation under $p = 7$.

and BMI are 5.78, -1.81, 1.71, and the $\hat{r}^2$ is 0.19. The plug-in estimate for dissimilarity is 3.37. Since the sample size is 4938 and the $\hat{r}^2$ is 0.19, Table 2.1 yields a sampling distribution that is roughly equivalent to a sample size of $n = 100$ with r^2_{100} that is between 0.9 and 0.95.

Based on the simulation results presented in Section 2.4, we expect the difference among plug-in estimate, bootstrap bias corrected estimate, and the Bayesian estimate with all data used to be small. To further explore the behavior of our approach in this type of data, we apply procedures to random subsamples. We let the subsample size $n^* \in \{100, 200, 400, 600, 1000, 2000\}$. For each subsample size n^*, we take 50 subsamples and apply our procedures. The average values of $\hat{r}^2$, point estimates, and the average length of confidence interval are recorded.

Table 2.2 presents the results of our procedures using MAP as the response variable. The values inside parentheses are the bootstrap estimate of standard errors. The values for plug-in estimate, bootstrap bias corrected estimate, and the Bayesian estimate using all data are 3.37, 3.18, and 3.40 respectively, which are similar. The length of the confidence interval using all data is 1.57. From simulation studies, we show that the estimates from the three methods are similar when the information is higher.

Next, we describe the results based on subsamples. From our simulation studies, we know that the bootstrap bias corrected estimate has the least bias among the three methods, and the other two methods are biased upward. In this example, we find that the bootstrap bias corrected estimate on average over subsamples is generally smaller than the other two methods which is consistent with the simulation study. We also note that the bootstrap bias corrected estimate approximates the full data value of the plug-in estimate with less data than the other methods. However, even with $n \approx 5000$ the width of confidence interval is 1.57, indicating considerable uncertainty about the population value of g remains. Thus, we cannot exclude the possibility

n	$\hat{r}^2$	PI	BS	PM	CI width
100	0.21	8.26(6.31)	7.18(7.90)	7.76(4.93)	48.39
200	0.20	6.03(5.18)	4.67(6.18)	5.90(4.43)	38.41
400	0.19	4.72(3.08)	3.08(3.35)	4.82(2.82)	27.73
600	0.20	4.63(2.52)	3.39(2.69)	4.72(2.82)	18.37
1000	0.20	4.16(1.40)	3.33(1.03)	4.21(1.45)	9.86
2000	0.19	3.75(0.68)	3.37(0.59)	3.79(0.67)	3.71
4938	0.19	3.37	3.18	3.40	1.57

Table 2.2: Results for different subsample sizes via using MAP as response.

that the larger subsample estimates given by the Bayesian posterior median are more accurate.

We find that the average length of the confidence intervals decreases as the sample size increases. When the information from the data is low, except for a large upper bound some intervals have their lower bounds reach the boundary value 1. The corresponding coverage rate can be higher than expected. By comparing the width of confidence intervals with subsample size 1000 or 2000, even though a few of the intervals with subsample size 1000 are truncated, the length of confidence interval decreases more than a factor of $\sqrt{2}$ which is faster than the general rule that the width is inversely proportional to the square root of sample size.

2.6 Discussion

Since the parameter of interest g involves estimation of extreme values and also involves a ratio, the performance of the plug-in estimate might be poor given that the information from the data is low. We showed that the bootstrap bias corrected estimate can have lower bias than other methods, but the trade-off between bias and variance makes it less favorable. We use the ABC procedure to provide a shrinkage estimator and show that it outperforms the plug-in estimate, provided that the information level is low. When the information level is high, its performance is comparable to the plug-in estimate.

To build a confidence interval, we use a stochastic search algorithm to approximate the profile likelihood, and we propose two ways to improve the algorithm. The simulation studies show that the coverage rate is reasonable when g is not too close to 1 and the information level is appropriate. When g is close to 1, the coverage rates are low and deteriorate as p increases. Even in the simple setting where one is considering the ratio between two regression coefficients that have been pre-specified, previous research has shown that the profile confidence interval may be infinitely wide. Thus, we should not be surprised to see such a wide interval in our setting. In application to NHANES data, the values for plug-in estimate, bootstrap bias corrected estimate, and the Bayesian estimate using all data are 3.37, 3.18, and 3.40 respectively. The length of the confidence interval using all data is 1.57, and we do not exclude the possibility that the larger subsample estimates given by the Bayesian posterior median are more accurate.

CHAPTER III

Functional Summaries of Covariance Structures

3.1 Introduction

When analyzing high dimensional data, it is often desirable to summarize the covariance structure in an accessible and easily visualized form. A complete description of a covariance structure is generally impossible to represent in a compact way. A major simplification results if we summarize the covariance structure in a way that treats the variables anonymously. Such a summary is unchanged if the variables are permuted. By treating variables anonymously we can compare data sets with different numbers of variables. The average squared correlation coefficient between pairs of variables $\sum_{i<j} \mathrm{Cor}^2(X_i, X_j)/\binom{p}{2}$ is of this form. Our goal in this chapter is to develop new summaries of this type.

The covariance matrix can be seen as a type of summary. However, this is a large $p \times p$ object that cannot be summarized or visualized easily if the dimension is large. In order to reduce the dimension, it is common to apply an orthogonal transformation to rotate the data such that the principal axes align with the coordinate frame. This essentially amounts to linearly transforming the data to the coordinates defined by the principal components. Since this transformation removes all correlations, it converts the original $p \times p$ covariance matrix into a p dimensional vector of variance components. In other words, we focus only on the length of the principal axes rather than on their orientations relative to the original covariate axes. Doing this reduces

the size of summary from $O(p^2)$ elements to p elements. However in some applications, the information that is lost in this reduction may be important.

In this chapter, we present a new framework for constructing summaries of covariance structures. The summaries should reflect interpretable patterns in the data, and as noted above should satisfy certain invariances, such as being unaffected by relabeling of the variables. Our summaries emphasize the degree by which each variable is predictable from the others, with a special focus on the number of variables required to predict another variable. The proposed functional summaries allow us to visualize the differences in the covariance structures between two data sets, even when they have different dimensions.

This chapter is organized as follows. In section 3.2, we review some correlation summaries proposed in previous research. These include both scalar and functional summaries. In section 3.3, we propose a new type of functional summary for covariance structures. In section 3.4, we illustrate the proposed functional summaries using artificial populations with known structure. To focus on the population behavior, we use large sample size. In section 3.5, we apply the functional summaries to two genomics data sets. In section 3.6, we use simulations to show that the functional summaries have power to distinguish correlation structures. In section 3.7, we discuss approaches for bias correction. In section 3.8, we consider an alternative approach for functional summaries using ridge regression.

3.2 Literature Review

In this section, we review several existing summaries for covariance structures. Let $X \in R^p$ be a random vector with covariance matrix Σ and correlation matrix R. To summarize the covariance structure, it is sometimes useful to use a scalar summary.

A summary that takes all pairwise correlation coefficients into consideration is the generalized variance (Wilks, 1932). This is defined as the determinant of the

covariance matrix $|\Sigma|$. The generalized variance takes on a small value when there are strong correlations among the variables. This is related to the fact that when correlations among the variables are very strong, the covariance matrix Σ is nearly singular. The total variance (Seber, 1984) is the trace of the covariance matrix $\sum_{i=1}^{p} \Sigma_{ii}$ that can be used as a measure for overall dispersion. When comparing two data sets of same dimension, these two measures can be used. However, they do not allow us to make comparisons between data sets of different dimension. This is because the generalized variance is a measure of the hypervolume that the distribution occupies in the space and the total variance tends to increases as p increases. The effective variance $|\Sigma|^{1/p}$ and the effective dependence $1-|R|^{1/p}$ introduced and studied by Pena and Rodriguez (2003) are descriptive summaries of covariance matrices that allow us to make comparison between data sets of different dimensions.

The heatmap of a correlation or covariance matrix, a graphical representation, can be used to gain a quick overview of pairwise correlation or covariance relationships. However, the heatmap is not invariant under permuting variables. The "corrgram" proposed by Friendly (2002) extends the idea of a correlation or covariance heatmap. It displays not only the correlations but also reorders the variables in the correlation matrix such that similar variables are positioned adjacently.

Graphical models (Dempster (1972), and Edwards (2000)) use graphs to represent multivariate dependence. Each node in the graph represents a random variable. The edges in the graph represent the dependencies between variables. When presenting the relationships among variables as a graph (network), the degree distribution (Dorogovtsev and Mendes, 2002) that counts the number of edges for each variable can be treated as a summary for covariance structure. The degree of a node is the number of edges incident to the node, and the degree distribution is the probability distribution of these degrees over the whole network.

The spectrum (the sorted eigenvalues) of a covariance matrix is another type of

functional summary. This functional summary represents the length of the principal axes. In principal component analysis (PCA), we use a linear transformation to represent the variation in a collection of vectors. The scree plot of the spectrum displays the variability explained by each of the principal components. The spectrum can be viewed as a functional summary to express the extent to which the data have an approximate low dimensional structure. However, rotating the variables makes it difficult to interpret the results in terms of the original variables.

Other useful summaries could be the quantile or cumulative distribution function of all pairwise correlations. These summaries are explicitly pairwise, and ignore the joint dependence between three or more variables. However, we have better information on how variables are pairwisely associated. In addition, the average of squared correlation coefficients $\sum_{i<j} \mathrm{Cor}^2(X_i, X_j)/\binom{p}{2}$ can also be treated as a scalar summary. It determine the degree of pairwise linear relationships since R captures only the linear dependence among variables. The scalar summaries mentioned above are all invariant under relabeling.

3.3 Proposed Functional Summaries

In this section, we propose a new type of functional summary to describe the degree to which each variable is predictable by the others. We place special focus on whether one variable can be predicted from the others using either a simple or a complex model. For example, a variable may be easily predicted using just one other variable in the data set. This is a special type of strong dependence and is usually easy to detect. A more difficult situation is when a variable is predicted by a combination of other variables in the data set but not by any single variable. Finally, a variable may be impossible to predict from the remaining variables in any way, in which it is independent of them.

For variable X_j, the ith row of correlation matrix with diagonal term removed

shows the bivariate associations of X_j with each other variable. We use the maximum value of squared correlation coefficients called $r^2_{j,1}$ to describe how X_j is predicted by a single other variable. The sequence $(r^2_{1,1}, \cdots, r^2_{p,1})$ describes how variables are explained by a single other variable. As a measure for dependence among variables, we use the quantiles of $(r^2_{1,1}, \cdots, r^2_{p,1})$ denoted by F_1 as a summary for dependence on how variables are predicted by one other variable. If p is not big, we can use the sorted sequence $(r^2_{(1),1}, \ldots, r^2_{(p),1})$ instead. It allows us to see how predictability changes. For example, we can use it to see how many of them are strongly/weakly predicted by a single other variable. To compare the data sets with different dimensions, we use the quantile summary F_1 to describe the change in dependence structure.

With the correlation matrix, we can derive a dependence summary that describes how variables are explained by a simple model with a single variable. To extend the summary of predictability to more than one other variable, we consider the predictability of X_j from a specific subset of k other variables $\{X_q : q \in Q_j\}$ with Q_j being a subset of $\{1, 2 \cdots, j-1, j+1, \cdots, p\}$ such that $|Q_j| = k$. We use the squared correlation coefficient between X_j and $E[X_j|X_q, q \in Q_j]$,

$$Cor^2(X_j, E[X_j|X_q, q \in Q_j]), \tag{3.1}$$

as a dependence measure between X_j and specific $\{X_q : q \in Q_j\}$. Since $E[X_j|X_q, q \in Q_j]$ is generally unknown, we can estimate it using regression techniques.

To measure the dependence on how X_j is predicted by k other variables, we could go through all possible subsets and use the maximum value as dependence measure. For convenience, we can rewrite the proposed measure as

$$\tilde{r}^2_{j,k} = \max_{\theta, |\theta|_0 = k} \mathrm{Cor}^2(X_j, X^T_{(-j)}\theta), \tag{3.2}$$

where $X_{(-j)} = X/X_j$ and $|\theta|_0$ be the 0-norm (the number of nonzero entries in θ).

The non-zero θ values point out the variables used. As a measure for dependence among k other variables, we use the quantile or sorted sequence of $(\tilde{r}^2_{1,k}, \cdots, \tilde{r}^2_{p,k})$ to describe the predictability from k other variables.

Calculation of (3.2) requires checking $\binom{p-1}{k}$ possible subsets of $X_{(-j)}$. To go through all possible subset exhaustively requires heavy computations if p is high. We consider an alternative for (3.2) that allows the computation to be substantially reduced. Instead of whole subsets computations, variable selection based on convex optimization is applied to determine a subset of $X_{(-j)}$ to approximate (3.2).

We use least absolute shrinkage and selection operator (LASSO) (Tibshirani, 1996) to select specific k variables. The LASSO is applied to regress X_j on $X_{(-j)}$ to select variables. Variables that have non-zero regression coefficients are "selected" by the LASSO algorithm. We then regress X_j on the k selected variables, $X^k_{(-j)}$, to find the squared correlation coefficient. For X_j, we use the measure

$$r^2_{j,k} = \max_{\theta} \mathrm{Cor}^2(X_j, {X^k_{(-j)}}^T \theta) \tag{3.3}$$

as a measure of dependence to describe how X_j is predicted by k other variables, where $X^k_{(-j)}$ are the selected k variables from $X_{(-j)}$. By varying X_j, we derive a summary F_k that describes the degree by which each variable is predicted from k others.

3.4 Illustrations

In this section, we apply functional summaries using simulated examples as illustrations for descriptive measures. To first focus on the population behavior, functional summaries with large sample size n = 100,000 are used. We use functional summaries that describe how variables are predicted from up to five other variables; i.e. $F_1, \cdots, F_5$. We use a graph to display the correlations between variables. This

is different from a traditional graphical model that uses a graph to represent the inverse covariance matrix. To begin with, we let $p = 20$ and the edge indicates the correlation coefficient is 0.4. If no edge exists between 2 nodes, these two variables are independent. The data are sampled from multivariate Gaussian distribution with mean 0 with $diag(\Sigma) = I_p$. We consider the functional summaries for the following four scenarios:

1. All variables are independent.

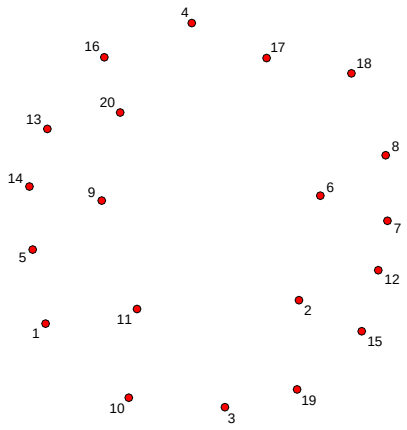
2. Variables with consecutive indices are connected by an edge. The correlation structure is Toeplitz matrix with two bands.

3. Edges exist among X_1 and all other variables and no edge exists among other pairs.

4. Generate random graph according to the Erdos-Renyi model (Erdos and Renyi, 1959). The number of edges is the same as scenario 2 and 3.

Figure 3.1 displays independent structure (scenario 1) and the corresponding functional summaries. Since no edge exists between any two nodes, all variables are independent, we expect $r^2_{j,k}$ in (3.3) used to derive F_k to be zero. The corresponding functional summaries F_k, $j = 1 \cdots, 5$, are zero vectors. The large-sample functional summaries support these expectations.

In the second scenario (Figure 3.2), we use a chain structure such that adjacent variables are connected by an edge. This means that except for variables X_1 and X_2, all other variables are correlated with exactly 2 other variables and that two variables are independent. For example, X_2 is correlated with X_1 and X_3 and the correlation between X_1 and X_3 is 0. Since all variables have at least one dependent neighbors, we then expect F_1 to be a non-zero constant function. The variable of F_2 is derived from $(r^2_{2,2}, \cdots, r^2_{p,2})$. Since $X_2, \cdots, X_{19}$ are correlated with two uncorrelated (independent)



(a) Scenario 1

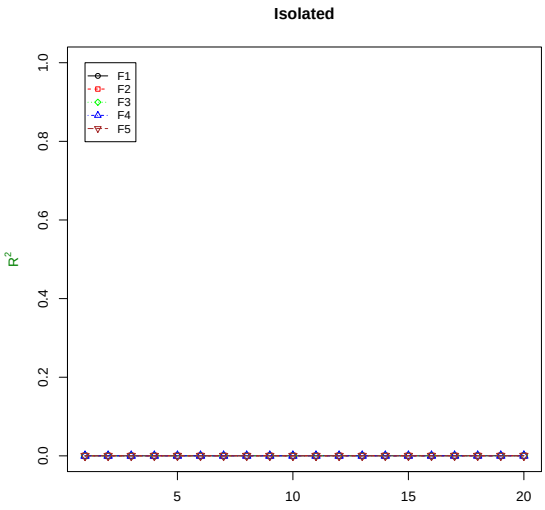
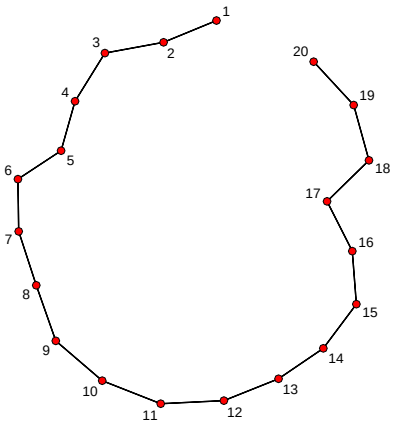


Figure 3.1: Functional summaries for the independent case. The upper graph indicates the pairwise relationship which indicates all variables are independent. The lower graph shows the functional summaries for the independent case.

variables. We expect $r^2_{j,2} = 2r^2_{j,1}$, $j = 2, \ldots, 19$. People may expect that $r^2_{1,2} = r^2_{1,1}$, since X_1 has only one dependent variables. However, the chain structure will create the indirect association between X_1 and X_j, $j \geq 3$. These indirect associations will result in minor increases from $r^2_{1,1}$. The same situation can apply to $r^2_{20,2}$. The graphs of F_2 will have the form of step function. Comparing with F_1, the first two values of F_2 have minor changes and the others are doubled. Due to thees indirect associations, a minor shift exists between F_3 and F_2. The large sample functional summaries support these expectation.

In simulation scenario 3, we fix the number of edges and make one variable that is correlated with all others and all others only have one dependent variable X_1. We expect F_1 to be a constant non-zero function since all variables have at least one dependent neighbor. Let $F_k(j)$ be the jth value of F_k. The last value of F_2, $F_2(20)$, will be $2F_1(20)$, since variable X_1 has more than one dependent variables that are independent of each other. The indirect associations cause the minor shift from $F_1(l)$ to $F_2(l), l \neq 20$. When comparing F_3 and F_2, there still exist minor shift except for the $F_3(20)$. The difference from $F_3(20)$ to $F_2(20)$ is equal to difference from $F_2(20)$ to $F_1(20)$. This is because variable X_1 have several correlated variables that are independent of each other. By looking at higher order of $F_k(20)$, we expect a constant shift from previous $F_{k-1}(20)$. The large sample functional summaries in Figure 3.3 support these expectations. If we remove variable X_1, all other variables are then independent and the functional summaries will degenerate to 0 values. The removal of variables that have high degree of dependence may result in dramatic change in functional summaries. To compare the function before and after removing a data point, we can use the quantile version of F_k to visualize the change. From scenario 2 and 3, we know that shift in F_k caused by indirect association is relatively small.



(a) Scenario 2

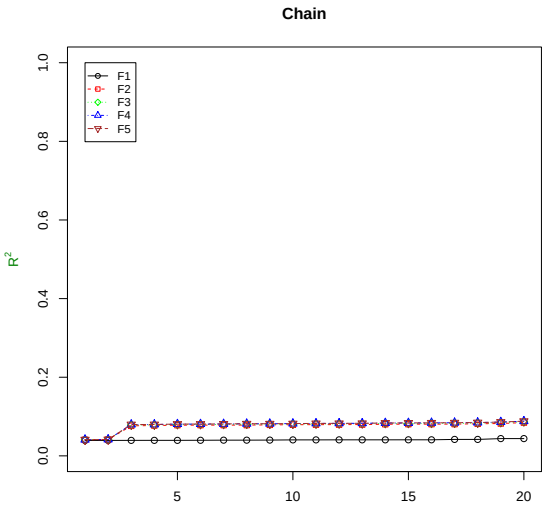
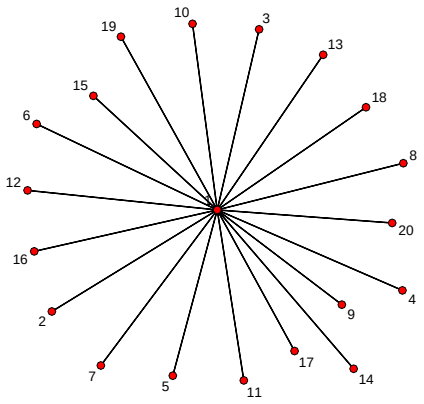


Figure 3.2: Functional summaries for the dependent case (scenario 2). The upper graph indicates the chain relationship such that each variable is dependent with the adjacent variables. The lower graph shows the functional summaries.



(a) Scenario 3

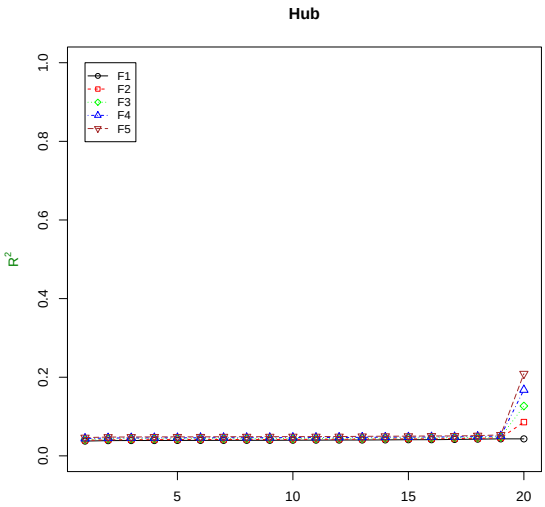


Figure 3.3: Functional summaries for the dependent case (scenario 3). The upper graph indicates the dependence relationship. The lower graph shows the functional summaries for the independent case.

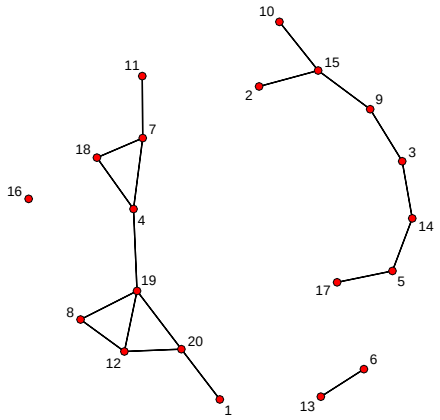
In scenario 4, we randomly assign edges with fixed number of edges. The first graph of Figure 3.4 displays the dependence structure among Xs. The dependence structure can be divided into four independent blocks. One block has only one isolated point, another has associations with only two other variables and the remaining two blocks have more complex dependence structure. In addition to minor effects from indirect relationship, variables that are associated with several other variables might have the chance to contribute to the functional summary. We expect the functional summaries to be a more complex form. The functional summaries in Figure 3.4 show that one variable may be isolated from others and 7 variables may have one dependent variables and others have at least 2 dependent variables.

From scenarios discussed above, the change in F_ks may relate to the number of edges each variable has, which is the degree distribution. Here, we provide an example to claim that the proposed function can capture some information that the degree distribution cannot given the correlations are either zero or a nonzero constant value. The upper two graphs in Figure 3.5 display two correlation structures that have same degree distribution, but the lower two graphs Figure 3.5 show that they have different functional summaries.

In these illustrative examples, the functional summaries characterize some dependence patterns. In the next section, we apply functional summaries to two gene data sets to detect some patterns and claim that the results are reproducible.

3.5 Data Analysis

In this section, we apply the functional summaries to two genomics data sets. The first data is the gene expressions in heart tissue from the left ventricular free wall of organ donors with no diagnosed heart disease. Heart tissue was collected by the Cleveland Clinic Kaufman Center for Heart Failure human heart tissue bank ($n = 108$) between August 1993 - May 2005. There are 33297 gene expressions.



(a) Scenario 4

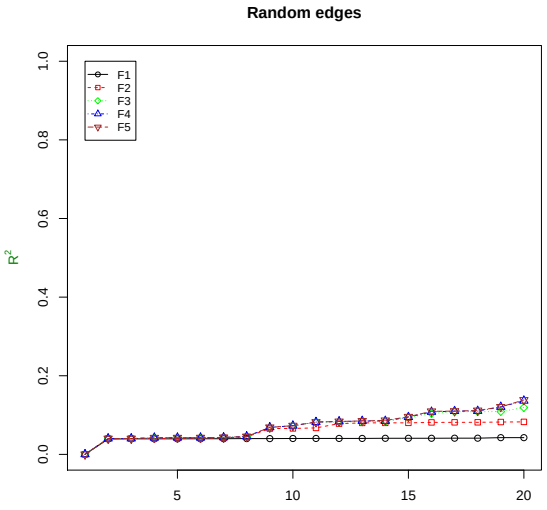


Figure 3.4: Functional summaries for artificial dependence structure.

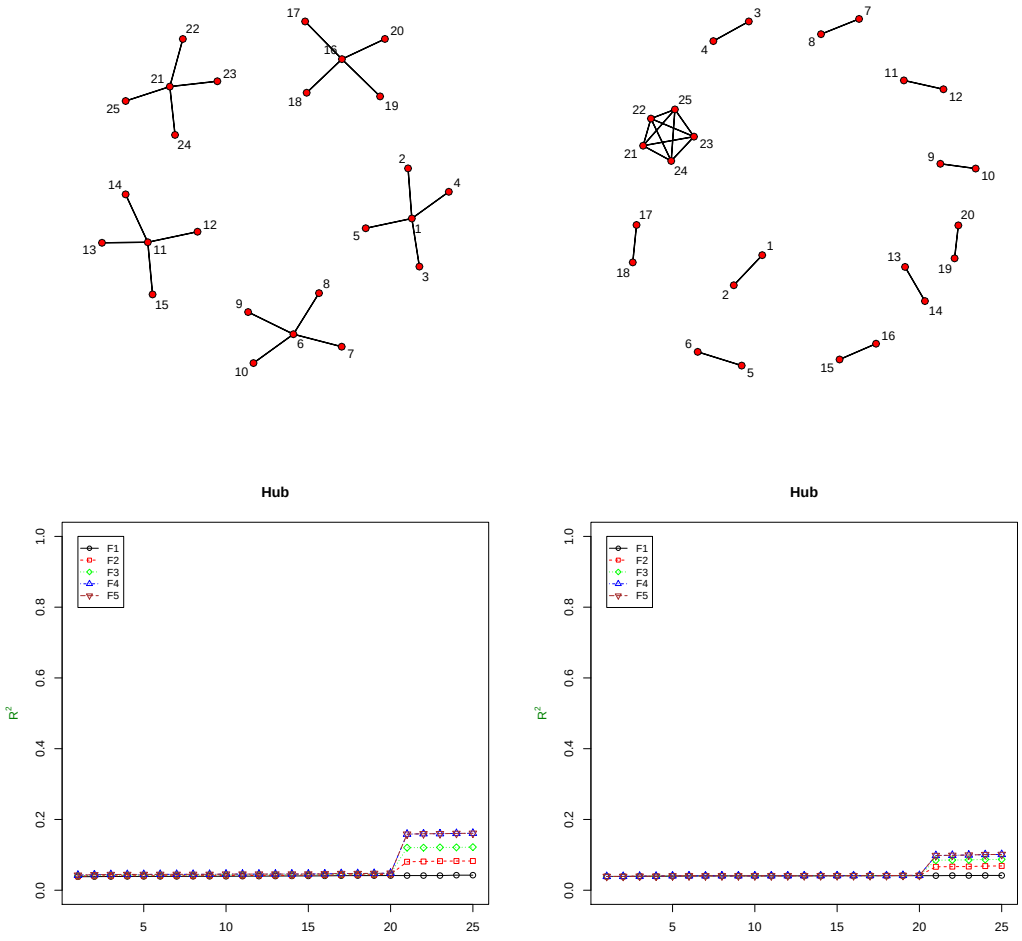


Figure 3.5: Two structures that have same degree distributions but different functional summaries

We apply functional summaries to gene sets in the Molecular Signatures Database (MSigDB) (Subramanian et al. (2005)). Here, we consider seven classes of gene sets:

- C1 Positional gene sets for each human chromosome and cytogenetic band.
- C2 Curated gene sets from online pathway databases, publications in PubMed, and knowledge of domain experts.
- C3 Motif gene sets based on conserved cis-regulatory motifs from a comparative analysis of the human, mouse, rat, and dog genomes.
- C4 Computational gene sets defined by mining large collections of cancer-oriented microarray data.
- C5 GO gene sets consist of genes annotated by the same GO terms.
- C6 Oncogenic signatures defined directly from microarray gene expression data from cancer gene perturbations.
- C7 Immunologic signatures defined directly from microarray gene expression data from immunologic studies.

For each gene set in a class of gene sets, we match the gene id and apply the functional summary to the matched genes. The number of matched genes might vary, and we only use the gene sets such that the number of matched genes are greater than or equal to 10. To compare gene sets with different dimensions, we apply the quantile version of functional summary. One of our interests here is to show that the proposed functional summaries can capture different dependence patterns. In addition, we also claim that the proposed functional summaries are reproducible.

To use functional summaries to capture different dependence patterns, we calculate $F_1, \cdots, F_5$ for each gene set and convert functional summaries as a vector via vectorize the functional summaries. We then project them to lower dimensions and

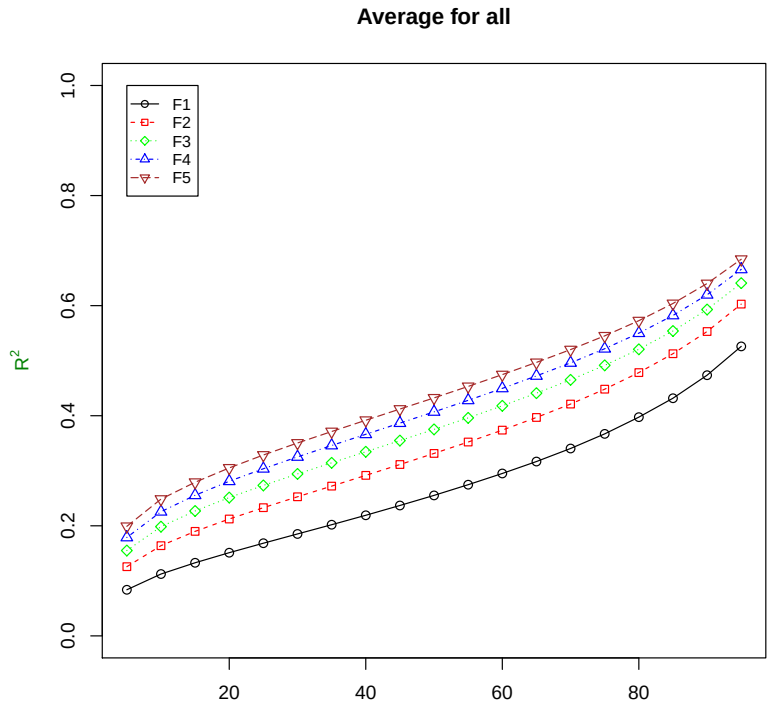


Figure 3.6: Average of functional summaries among all gene sets. The average curves are smooth.

examine some points. For example, to examine whether some gene sets have distinct overall summaries, we can stack the F_ks as a vector and apply principal component analysis (PCA) to project them to lower dimensions. We then take the gene sets that have unusual behavior in the projected space, and plot the corresponding functional summaries. We can also calculate the difference $F_2 - F_1$ and apply same the procedure to check whether some gene sets have distinct behavior on $F_2 - F_1$. To see how F_1 changes, we can converts F_ks to the consecutive difference in F_1 and apply PCA to detect some gene sets.

We first plot the average of functional summaries among all gene sets. The average summaries displayed in Figure 3.6 show that the shift will decrease as k increases and the curves for each F_ks are smooth. We list some functional summaries that have extreme values when projecting them to low dimension space.

To visualize the overall change, we stack the F_ks as a vector and project them to a low dimension space. We apply PCA to the stacked summaries. Figure 3.7 displays the projected score for first two PCs and we label the 6 points that have extreme projected values on each axis. The first PC accounts for 88% of the variability and the second PC accounts for 8% of the variability. The loadings for the first PC are the negative weighted average of the F_ks with more weight placed on the tail F_ks and the loadings for second PC are the weighted difference between the tail F_ks and the first half F_ks. Points A, B and F have small values in the first PC. Point C has the largest value on the first PC score. Points D and E have larger values on the second PC score. We expect that the behavior on C is different from A, B and F since the first PC accounts for 88% of the variability. The graphs of A, B and F in Figure 3.8 show that at least half of the variables are moderately or highly correlated with one other variable, and graph C show that variables are weakly correlated. Even though A and B have similar projected values, graph A and B show similar patterns except for the early quantiles. Graphs D and E show a big difference between the early quantiles and the tail quantiles.

Next, we apply the PCA on $F_2 - F_1$ to find functional summaries that have distinct patterns from F_1 and F_2. Figure 3.9 shows the projected scores for first two PCs and Figure 3.10 plots functional summaries for the labeled points in Figure 3.9. The loadings for the first PC are the negative weighted values of $F_2 - F_1$ which accounts for 44% of the variability and the loadings for the second PC are the weighted difference between the tail half $F_2 - F_1$ and the first half $F_2 - F_1$ which accounts for 14% of the variability. Graph D in Figure 3.10 shows that the difference between F_1 and F_2 is minimized, since the first PC is from the negative weighted average of $F_2 - F_1$. The other 5 graphs showed difference exists between F_1 and F_2. Since graph A has the largest first PC score, large differences exist between F_1 to F_2.

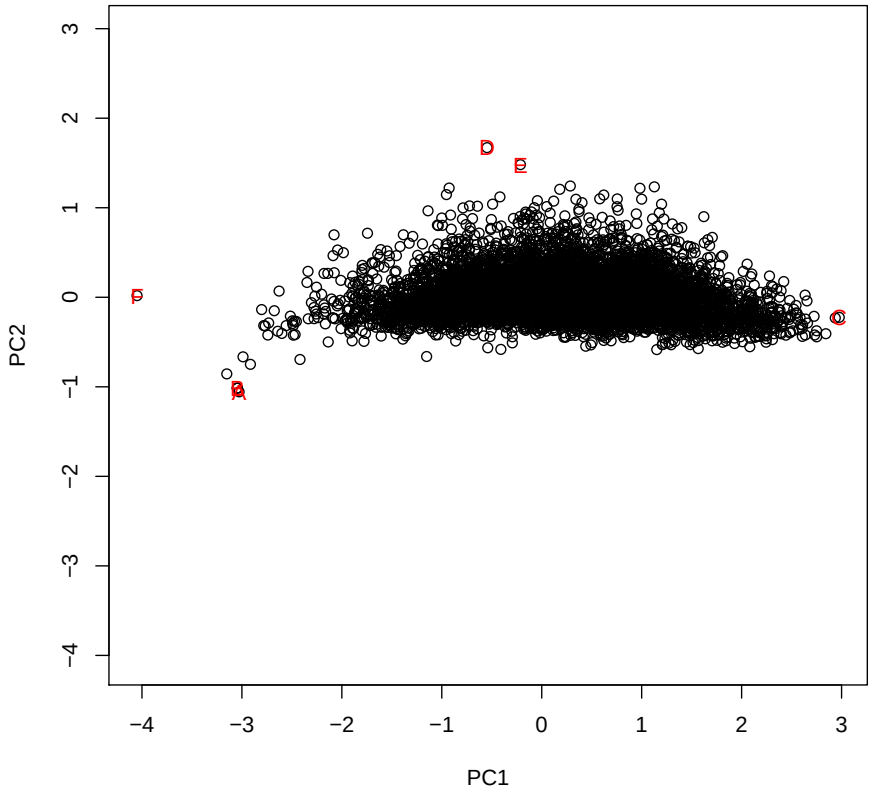


Figure 3.7: Projection of PC scores for stacked F_ks.

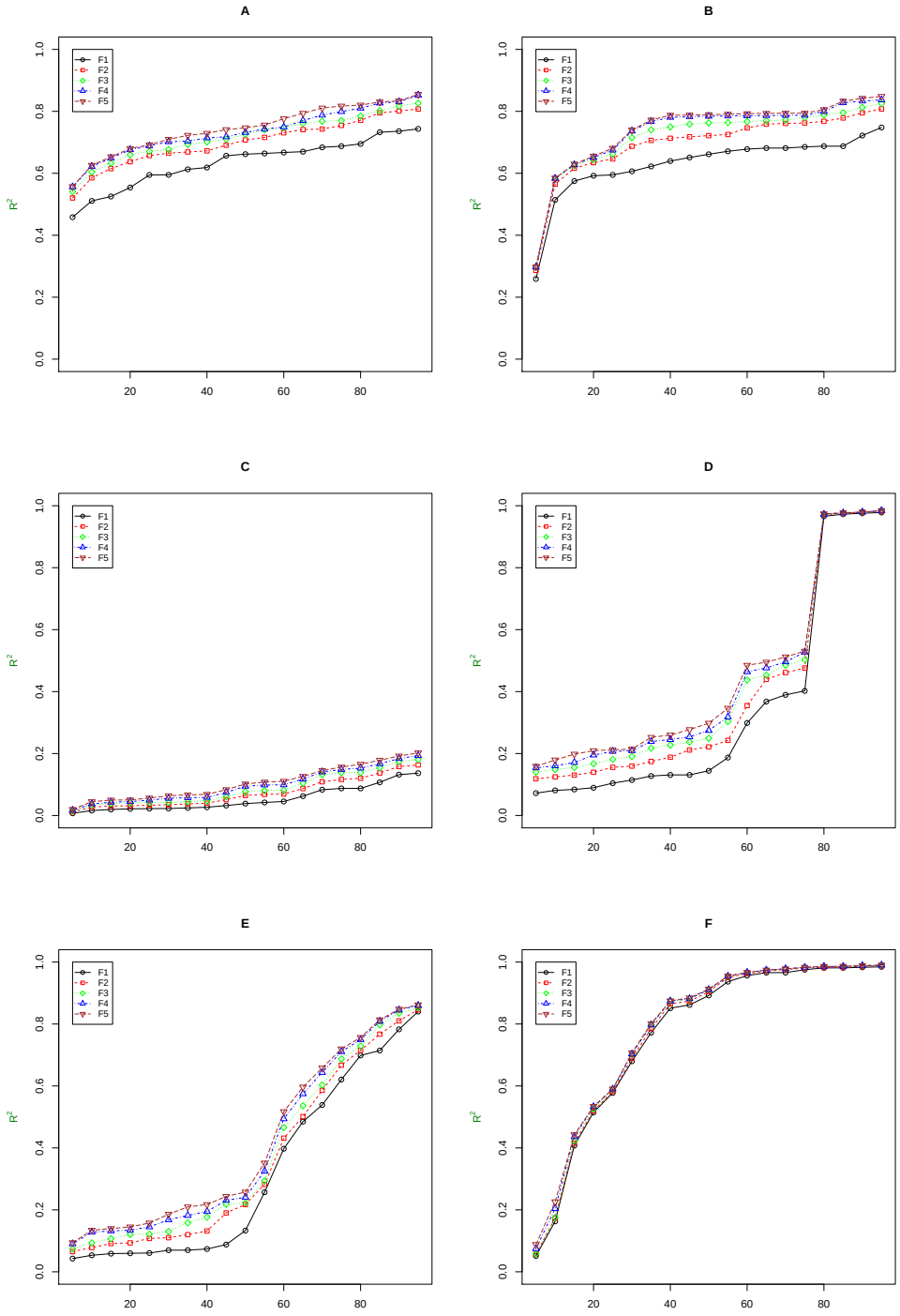


Figure 3.8: Functional summaries for six points that are labeled in Figure 3.7. Distinct summaries are captured via examining those that have extreme projected values.

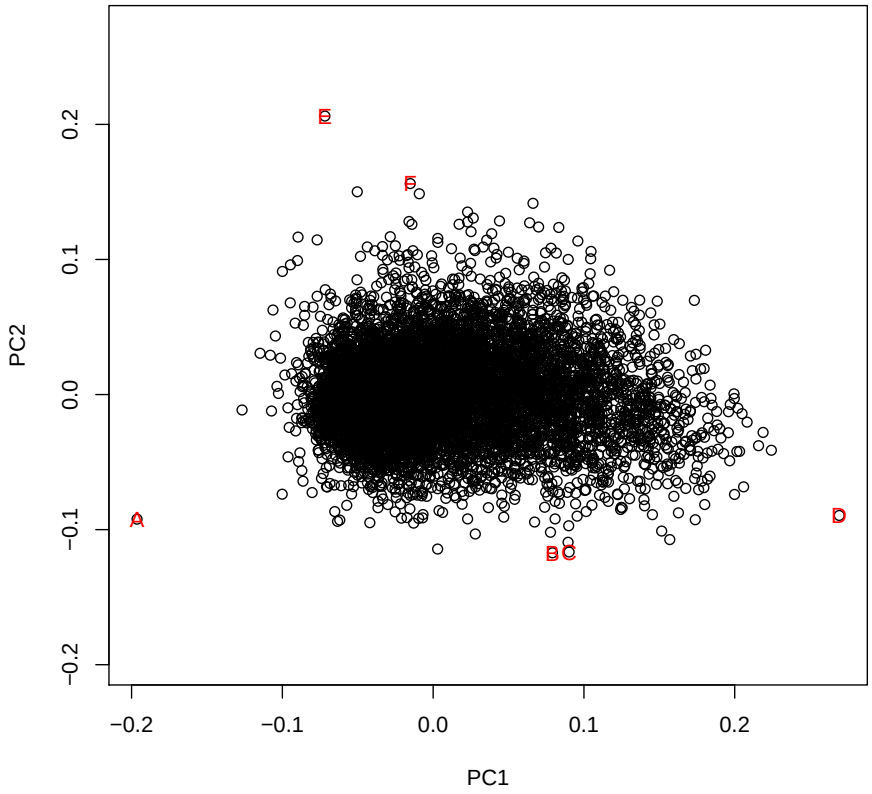


Figure 3.9: Projection of PC scores for $F_2 - F_1$.

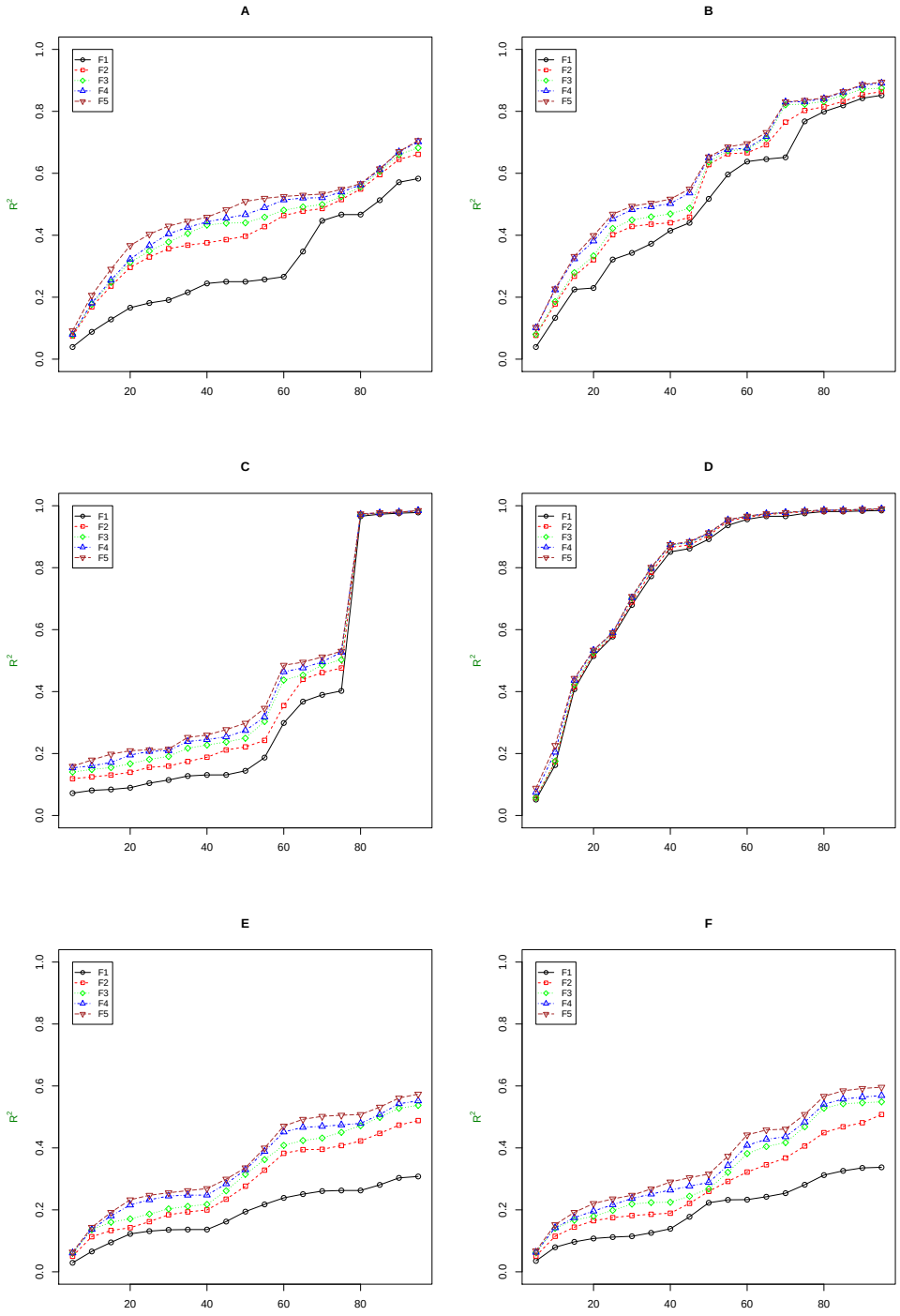


Figure 3.10: Functional summaries for six points that labeled in Figure 3.9. Distinct summaries are captured via examining those that have extreme projected values.

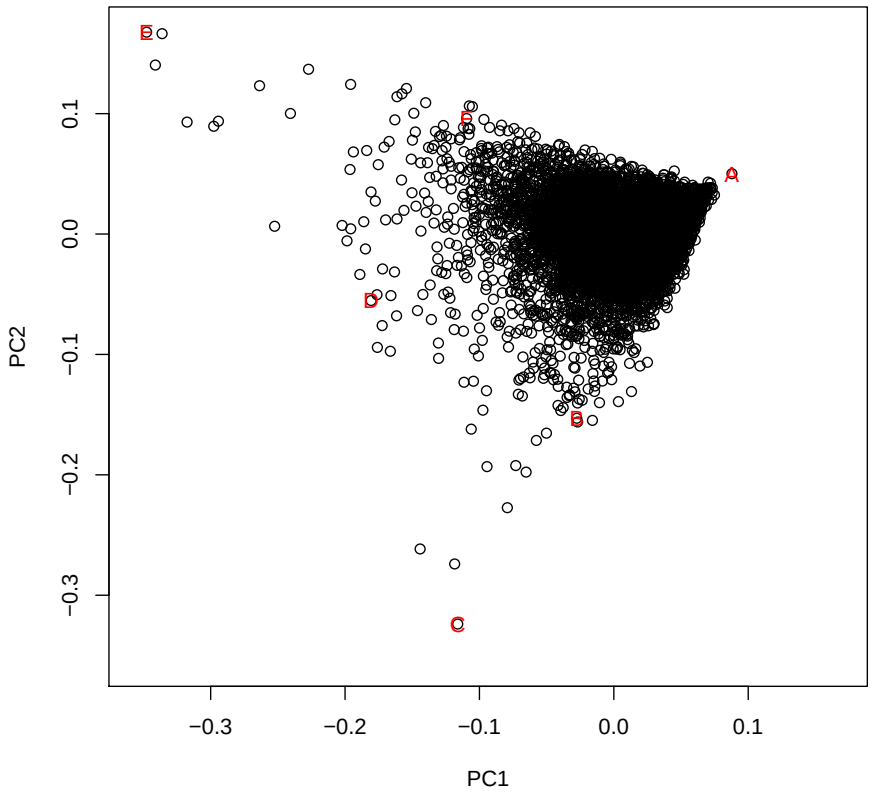


Figure 3.11: Projection of PC scores for consecutive difference of F_1

To capture the change within F_1, we calculate the consecutive differences and project them to lower dimensions. Figure 3.11 presents the projected scores and some points are labeled such that the corresponding functional summaries are in Figure 3.12. The loadings for the first two PCs are roughly the linear combination of tail consecutive difference of F_1, and they account for 24% and 14% of the variability. Since the consecutive difference is a nonnegative vector, the projected scores showed cone shape in 3.12. When we apply the PCA on consecutive difference of F_1 and the loadings have more weight on tail consecutive difference of F_1, we expect distinct patterns on tail F_1. All graphs in Figure show distinct tail behavior in F_1.

Next, we use another gene data set that has larger sample size to confirm that the proposed functional summaries are reproducible. That is, researchers should

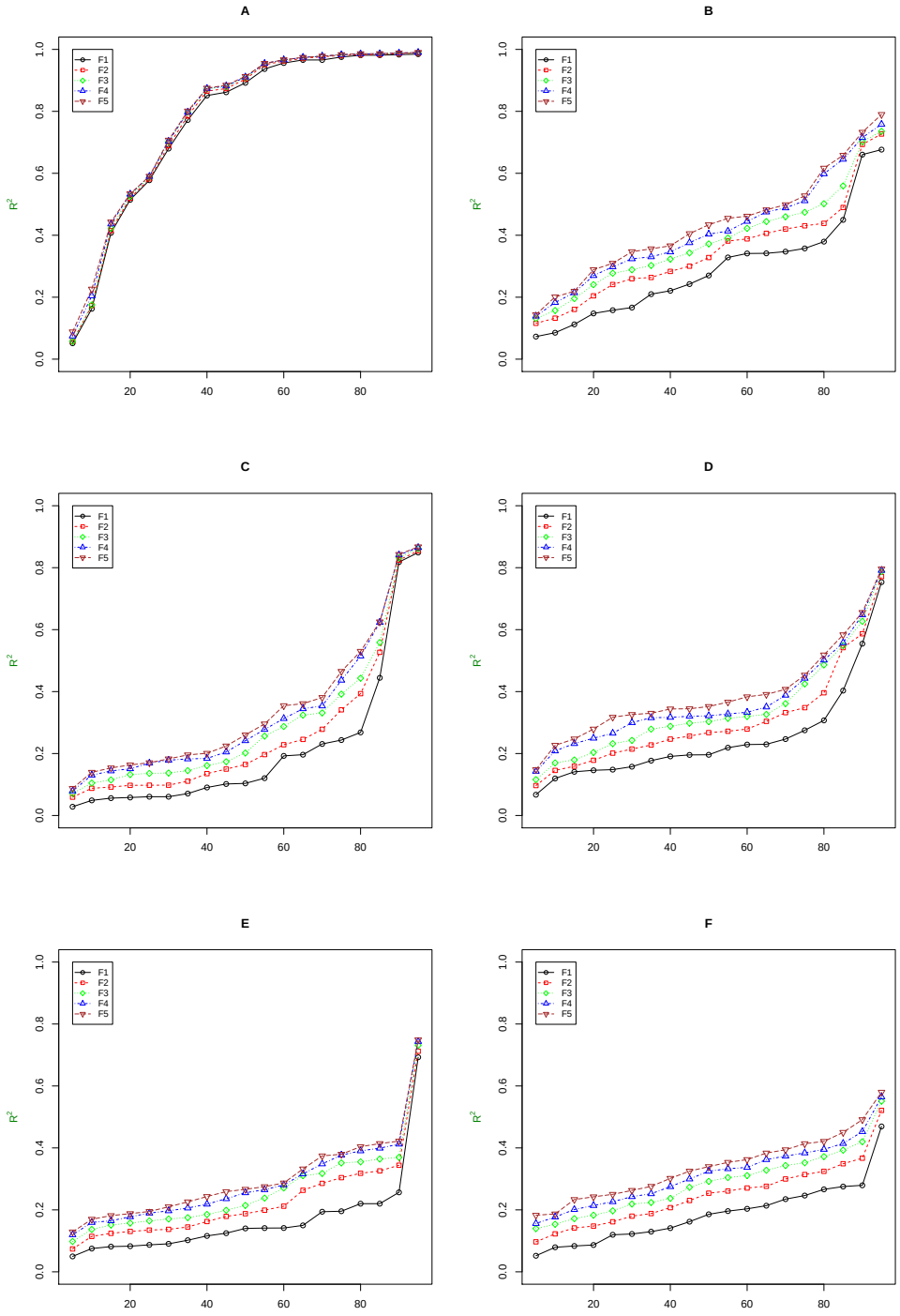


Figure 3.12: Functional summaries for the 6 points labeled in Figure 3.11. Distinct tail F_1 patterns are captured.

be able to replicate the analysis on equivalent data, and obtain similar summaries. There are 734 whole-blood RNA samples from the Estonian Biobank were profiled to find molecular mechanisms behind complex human diseases. We then take 2 non-overlapping subsamples from the gene data set and apply the proposed functional summaries to each gene set. We use gene sets from gene class C1 only. To see whether the functional summaries from the same gene set behave similarly or not, we first stack the functional summaries $F_1, \cdots, F_5$ as a vector and merge the summaries from two subsamples. The principal component analysis is then applied to the merged data set to check reproducibility. If the proposed method is reproducible, the functional summaries will be similar for each gene set even under different subsamples and hence similar on the corresponding principal component scores. We treat the projected scores from the same each gene set but under different subsample as a pair. The scatter plot of those pairs are used to check the reproducibility. We plot the projected scores for the first four principal components. We apply the procedure to subsample of size 25, 50, 100, 200, and 300 under gene class C1 and the results are quite well with sample size 100. Figure 3.13 displays the projected scores and the scatter plots from the first PC is quite linear but other PCs are not given $n = 25$. With $n = 100$, the plots in Figure 3.14 shows that the scatterplots for first three PCs are quite linear which indicates that the proposed method is reproducible.

Another way to visually check the reproducibility is to apply the proposed functional summaries on each split data and apply the dimension reduction procedures to project them to lower dimension and compare the patterns for each split data. We employ the principal component analysis again. We plot the first two PCs, find 4 extreme points in the first scatterplot, and also label the same gene sets on the second split data. Figure 3.15 displays the results for $n \in \{25, 50, 100, 200\}$. The results show that the position for the first PC scores is roughly the same, but the position for the 2nd PC score might vary. When the sample size is 100 or large, the

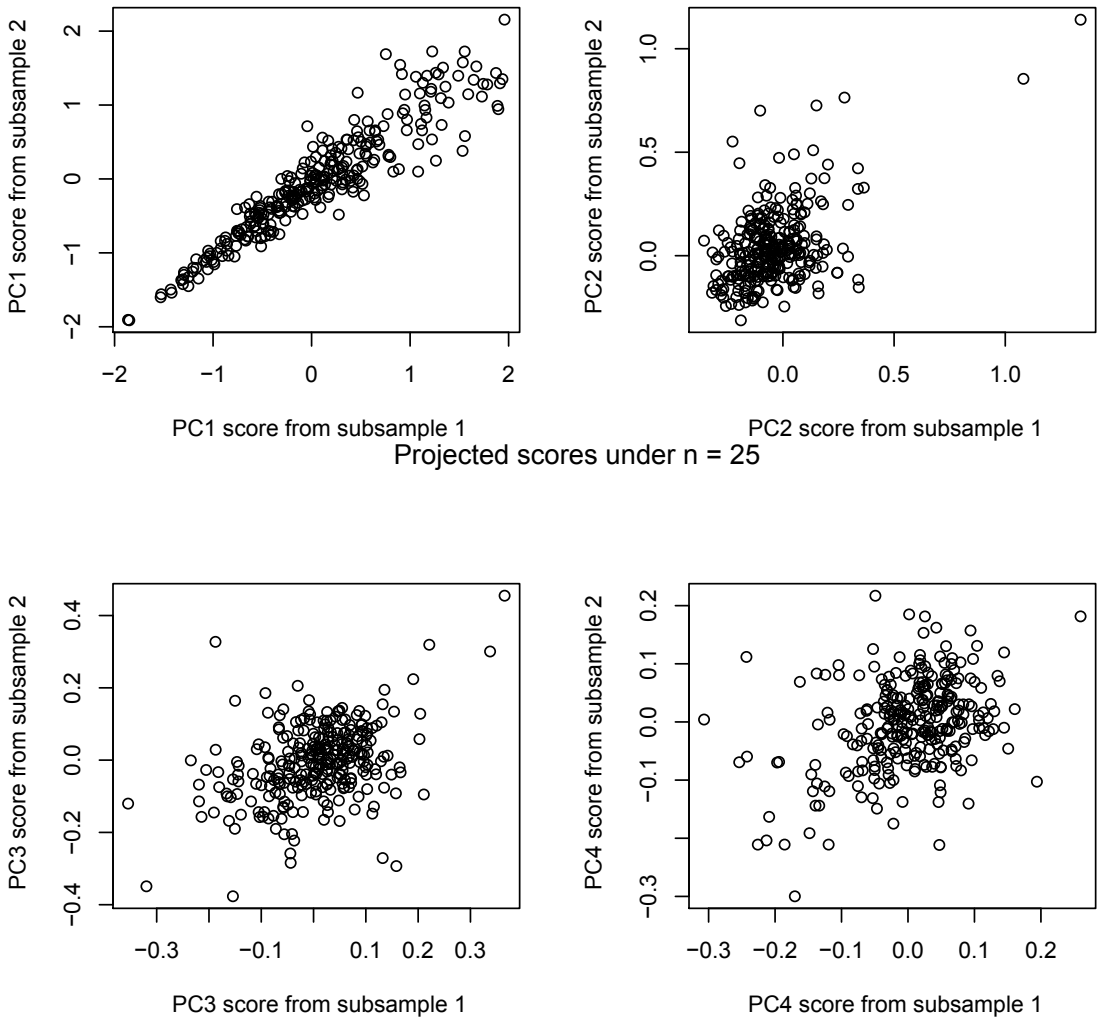


Figure 3.13: Check of reproducibility with n = 25. The projected scores for the first two PCs are strongly linearly associated which indicates that the proposed method is somewhat reproducible. The first two PCs account for 94 (87+7) percent of the variability.

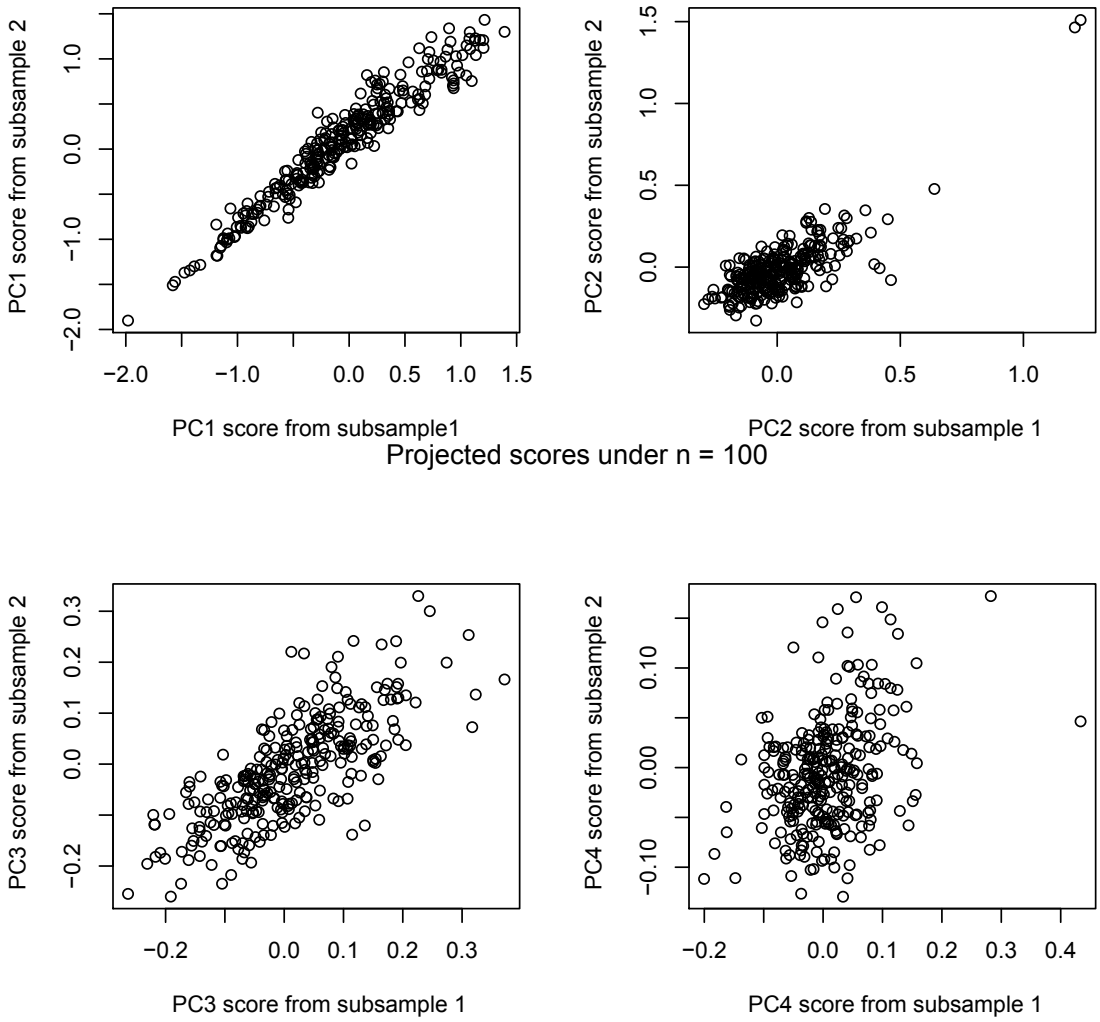


Figure 3.14: Check of reproducibility with n=100. The projected scores for the first 3 PCs are strongly linearly associated which indicate the proposed method is reproducible.

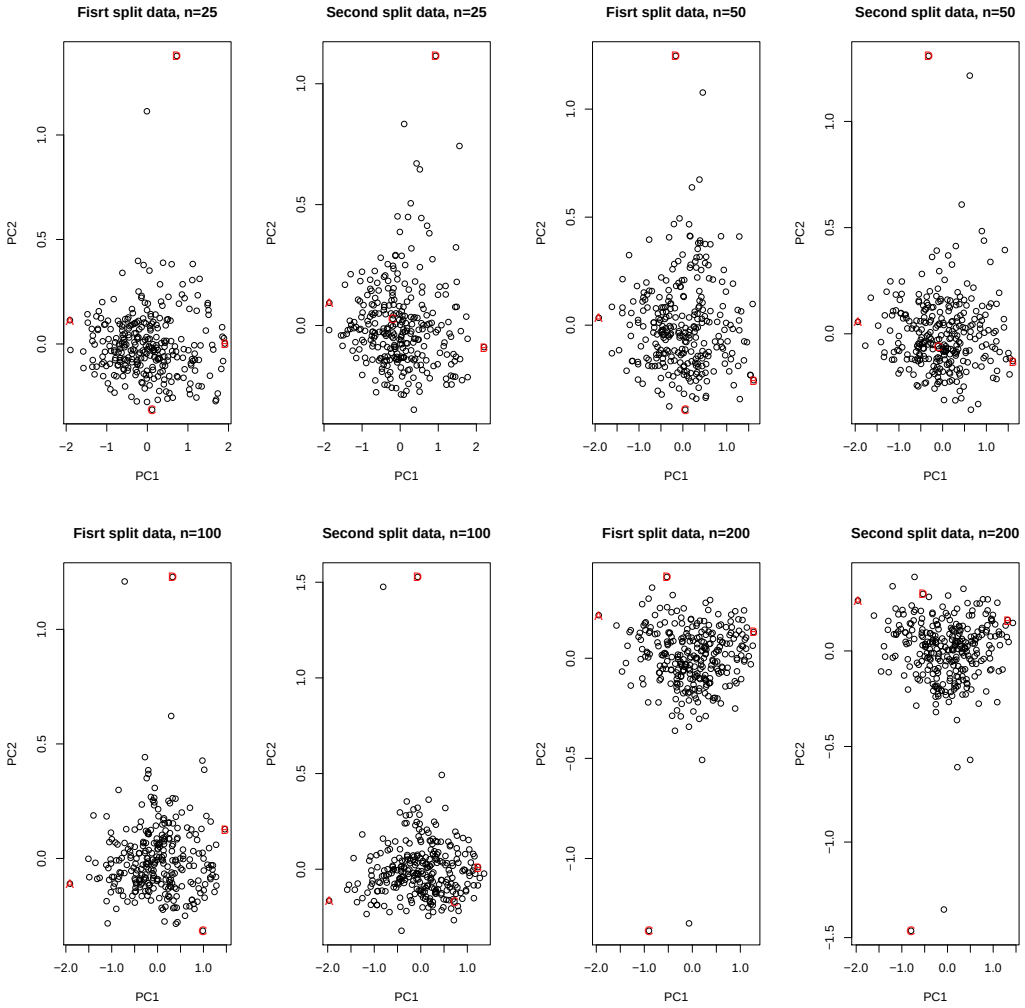


Figure 3.15: Check of reproducibility. Project functional summaries on lower dimensional space and label the extreme value in the first split data set. In the second split data, we label the gene sets that are labeled in the first split data.

relative positions are similar for both split data. With appropriate sample size, the results are reproducible.

3.6 Simulation Study

In this section, we use a simulation study to show that the proposed functional summaries have power to distinguish different structures. We consider a sparse covariance matrix with $p = 21$ which can be split into 7 independent blocks. There

are 3 variables within each block and the dependence structure within each block is exchangeable structure with correlation ρ. The sample size can be 100, 200, or 400 and ρ can be 0 , 0.3, or 0.6.

Figure 3.16 and 3.18 display the average curves for 400 Monte Carlo samples and the corresponding low dimension projections. The low dimension projections are calculated via applying PCA to a new response that treats F_ks as a new vector. The first PC accounts for 99% of variability. Even with $n = 100$, the average curves can be used to distinguish three ρ values.

To measure how two arbitrary structure differ, we borrow Kullback-Leibler divergence (Kullback and Leibler, 1951). We fit a bivariate normal distribution to the first two PCs scores for each structure and calculate the average Kullback-Leibler divergence between two structures since Kullback-Leibler divergence is asymmetric. When n is 100, the distance between $\rho = 0$ and $\rho = 0.3$ is 21.20, the distance between $\rho = 0.3$ and $\rho = 0.6$ is 100.34, and the distance between $\rho = 0$ and $\rho = 0.6$ is 513.61. When n is 200, the corresponding distance will be 143.27, 205.73, and 2517.14, respectively. When n is 400, the corresponding distance will be 685.17, 427.40, and 9911.08, respectively. With greater sample size, the distance becomes larger and the summaries also support these.

3.7 Bias Correction

The r-squared values are used to build the functional summaries and the r-squared values tends to overestimate the strength of association. Suppose that the true (population) correlation is zero, you will not get a sample r-squared that is zero. We call this the model fitting bias. The sorting used in the functional summary will also make the estimate biased. These two possible sources of bias will deteriorate the finite sample functional summary. We then consider several approaches to do the bias correction.

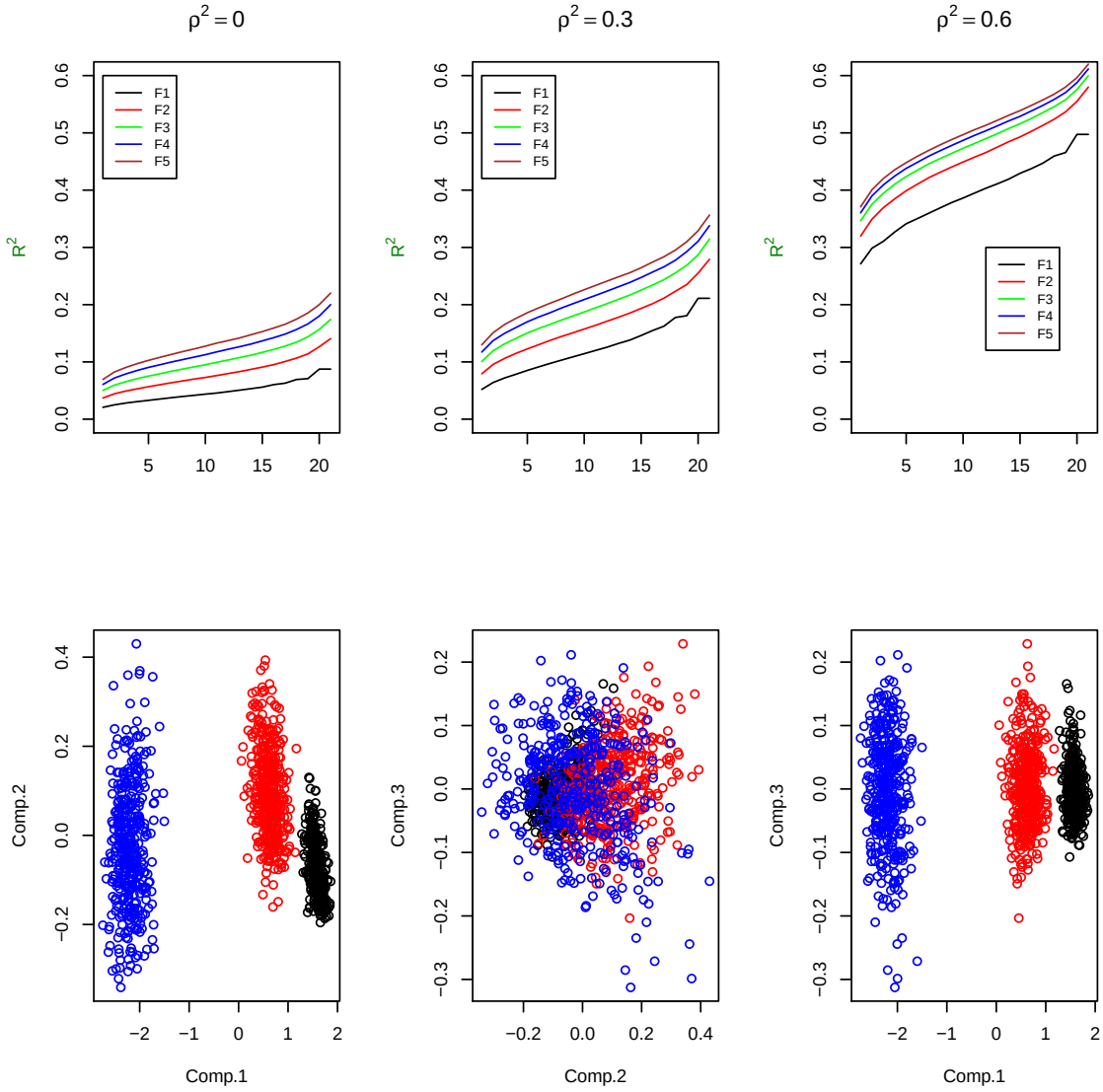


Figure 3.16: Simulation study with n=100. The upper row represents the average curve for 400 Monte Carlo samples. The lower row represents the lower dimension projections. Black points are projected scores given $\rho = 0$, the red points are projected scores given $\rho = 0.3$, and blue points are projected scores given $\rho = 0.6$.

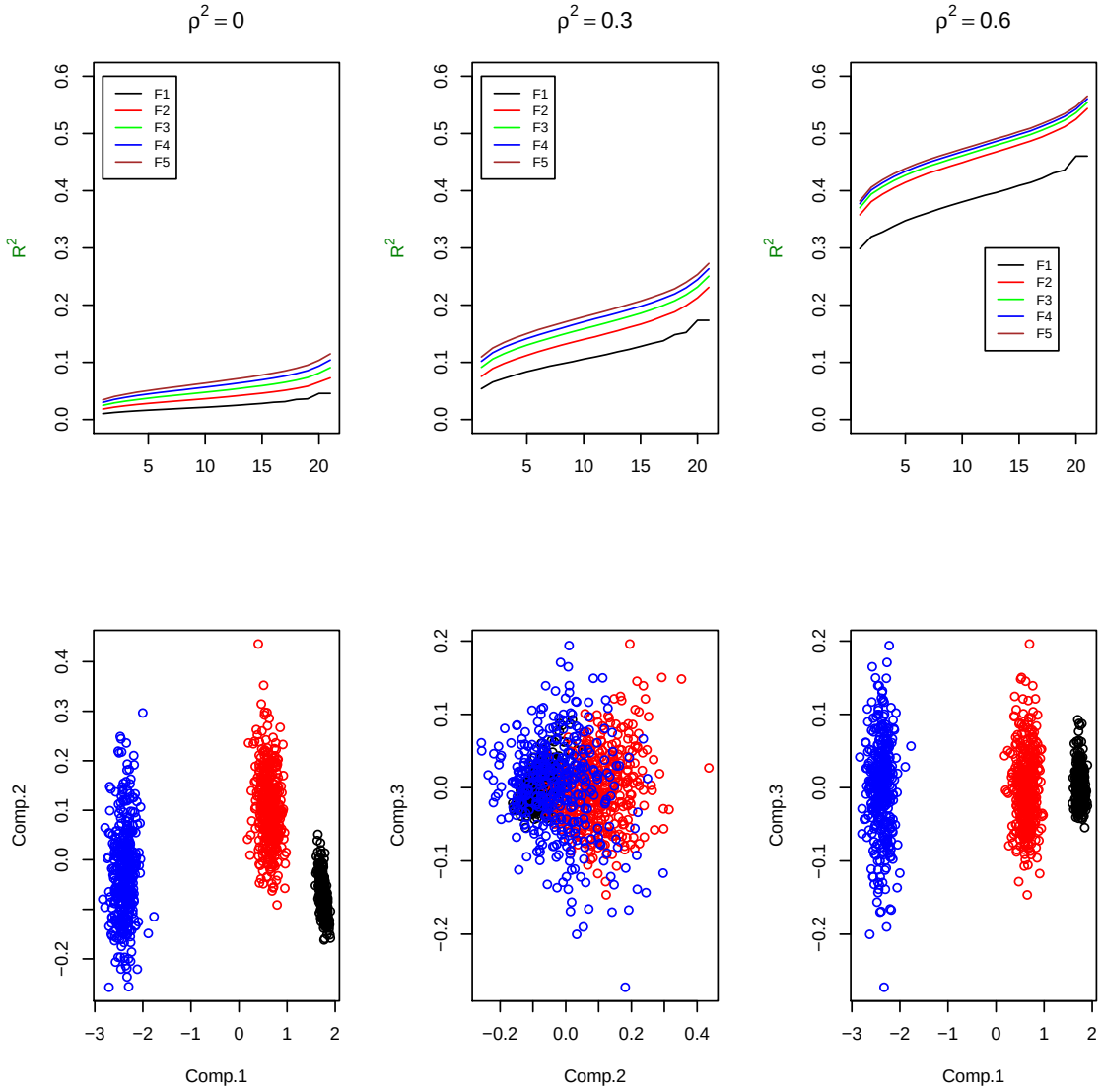


Figure 3.17: Stimulation study with n=200. The upper row represents the average curve for 400 Monte Carlo samples. The lower row represents the lower dimension projections. The black points are the projected scores given $\rho = 0$, the red points are the projected scores given $\rho = 0.3$, and blue points are the projected scores given $\rho = 0.6$.

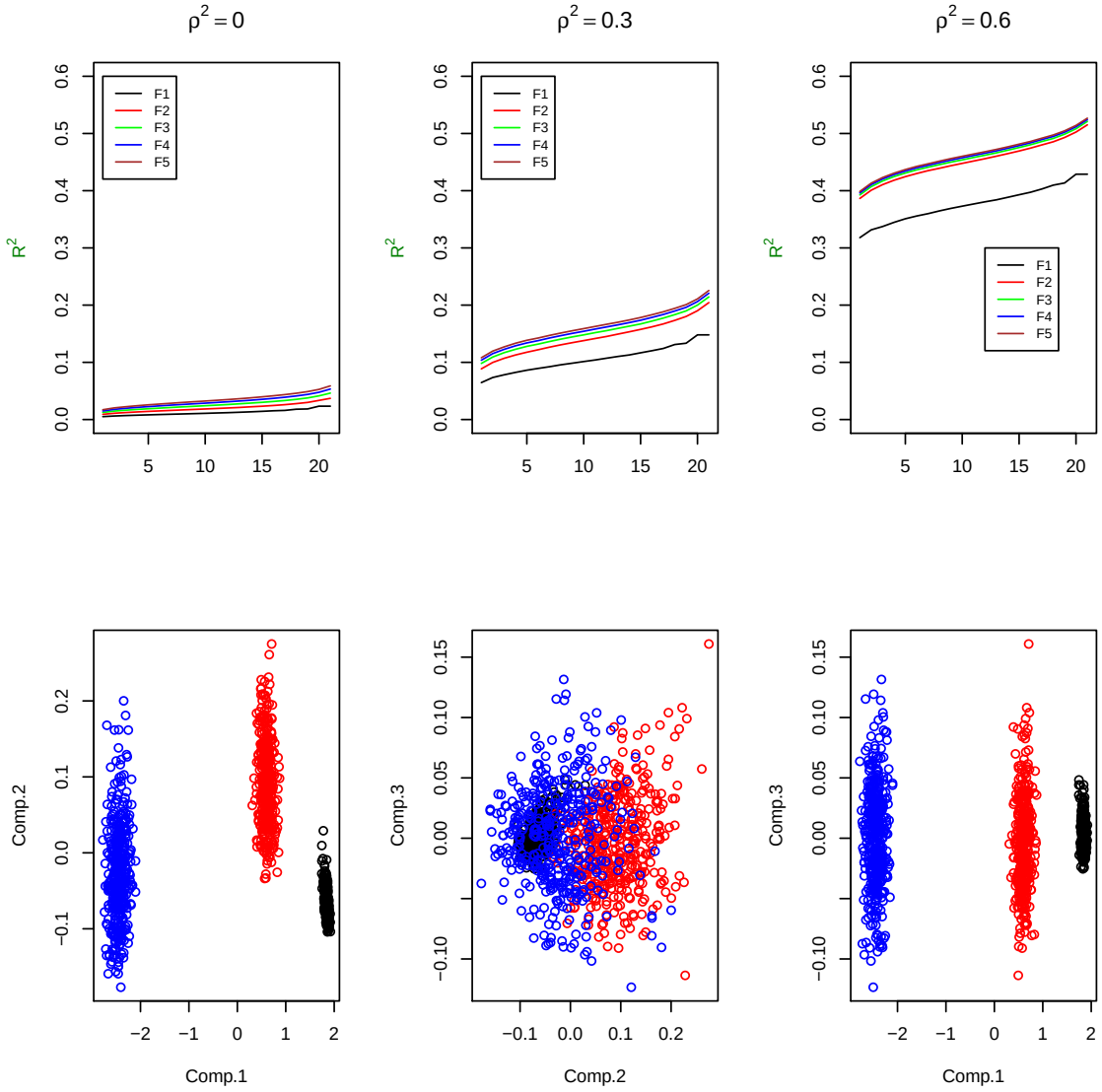


Figure 3.18: Stimulation study with n=400. The upper row represents the average curve for 400 Monte Carlo samples. The lower row represents the lower dimension projections. The black points are the projected scores given $\rho = 0$, the red points are the projected scores given $\rho = 0.3$, and blue points are the projected scores given $\rho = 0.6$.

The increase in the number of regressors will make the r-squared non decreasing, even if we add noise variables to the model. To remove the bias that is due to the model fitting bias, the adjusted r-squared or the predicted r-squared is considered. For the predicted r-squared, we use LASSO to select variables. We then systematically remove each observation from the data set, estimate the regression equation, and determine the prediction for removed observation. The predicted r-squared is the r-squared between response and prediction.

Another approach for bias correction is to remove the sorting bias. While the sample size increases, the estimate for F_{kj}, the jth element in F_k, is less biased. We consider determine a trend connect sample size n and F_{kj} and apply a simulation and extrapolation (SIMEX) approach. We take bootstrap samples with different size n^* to find the estimate for F_{kj}, call $\hat{F}_{kj}^{b,n^*}$, where b indicates bth bootstrap sample, $b = 1, \cdots, B$. Since the limiting distribution of sample r-squared is normally distributed,

$$\sqrt{n}(\hat{r}^2 - \rho^2) \to \mathrm{N}(0, \sigma_\rho^2),$$

and $\hat{F}_{kj}$ is the jth order statistics over finite terms, we consider fit a line $a + b/\sqrt{n}$ and use the limiting value a as a bias corrected estimate. The least squares method is used to derive the estimate. Note that the F_{kj}s are adjusted pointwisely.

We use the simulation study to demonstrate the performance for these three approaches. The average of mean integrated squared errors (MISE) over first five functional summaries is used to measure the performance for bias correction. We use the same blockwise structures. In the third approach, we let $n^* \in \{.4n, .5n, \cdots, n\}$ and set B to be 50 for each n^*.

Table 3.1 displays the average of mean integrated squared errors (MISE) for $F_1, \cdots, F_5$. The values inside the parenthesis are the corresponding squared bias. All bias reduction approaches show smaller average MISE and squared bias. When

Table 3.1: The average of mean integrated squared errors (MISE) for $F_1, \cdots, F_5$ and the value inside parenthesis is the squared bias. (Each value is multiplied by 100.)

r	n	R^2	Adjusted R^2	predicted R^2	SIMEX
0.0	100	1.116(1.095)	0.641(0.620)	0.254(0.237)	0.125(0.093)
	200	0.298(0.293)	0.163(0.157)	0.064(0.060)	0.054(0.046)
	400	0.079(0.078)	0.042(0.041)	0.017(0.015)	0.019(0.016)
0.3	100	0.811(0.749)	0.520(0.458)	0.378(0.315)	0.438(0.358)
	200	0.304(0.268)	0.232(0.196)	0.207(0.171)	0.197(0.151)
	400	0.129(0.113)	0.110(0.094)	0.102(0.085)	0.077(0.056)
0.6	100	0.636(0.545)	0.575(0.485)	0.613(0.515)	0.467(0.352)
	200	0.305(0.259)	0.281(0.237)	0.284(0.238)	0.167(0.115)
	400	0.144(0.122)	0.137(0.112)	0.137(0.112)	0.074(0.050)

ρ is zero, the SIMEX approach has slightly negative adjusted values. However, the functional summaries are supposed to be nonnegative. If we truncate the negative values to zero, the average MISE and squared bias will be almost zero for SIMEX approach. From Table 3.1, the SIMEX approach is generally preferred and the predicted r-squared is the next best performing method.

3.8 Alternative Approach Based on Ridge Regression

The LASSO is used in the proposed functional summaries. LASSO is a regularized version of least squares that use the L_1 regularization. One might consider using another well-know regularized least square method, ridge regression (Hoerl, 1962). Both regularized least squares methods shrink the parameter estimates. Since the LASSO will reduce the parameter estimates to zero as the penalty increases, we use it to select important features to derive the functional summary.

The ridge approach does not shrink the parameter estimates to zero. If we use a small penalty, we incorporate more structures to the model. People might con-

sider using ridge regression with effective degree of freedom controlled to derive the functional summary. The r-squared value from ridge regression to regress X_j on X_{-j} with k effective degree of freedom controlled is use to replace $r^2_{j,k}$ in F_k. However, the control of effective degree of freedom in integral values doest not clearly reflect the change in correlation structure. Figure 3.19 displays the functional summaries via either ridge regression or LASSO. The difference from F_k to F_{k+1} under ridge regression is relative small compared to these from LASSO. The control of integer values in effective degree of freedom do not clearly reflect the change in dependence structure. Actually, the change in r-squared value is small if we set the effective degrees of freedom to be integer values. Figure 3.19 displays the functional summaries for blockwise structures. The summaries with effective degrees of freedom controlled do not reflect the change in dependence structure when more structures are included. The purpose of ridge regression is used as a remedy for multicollinearity. It does effect the r-squared values much. Instead, we consider using coefficient of determination,

$$1 - \frac{MSE^k_{X_j}}{Var(X_j)},$$

to replace the term $r^2_{j,k}$ in F_k, where $MSE^k_{X_j}$ is the MSE of ridge regression of X_j on $X_{(-j)}$ with k effective degree of freedom controlled.

Figure 3.20 displays the functional summaries via ridge alternative with $n = 400$. The first row plots the first five summaries for blockwise structure with ρ being 0, 0.3, and 0.6, respectively. These are the average curves over 400 Monte Carlo samples. We can figure out that they look distinct, especially in higher effective degree of freedom. In the second row, the projected scores from principal component analysis that merge results from three structures are displayed. We apply PCA to the vector $(F_1, \cdots .F_5)$. The colors are used to label the scores from the same structure. The first components accounts for 99 percentage of variability and can be used to

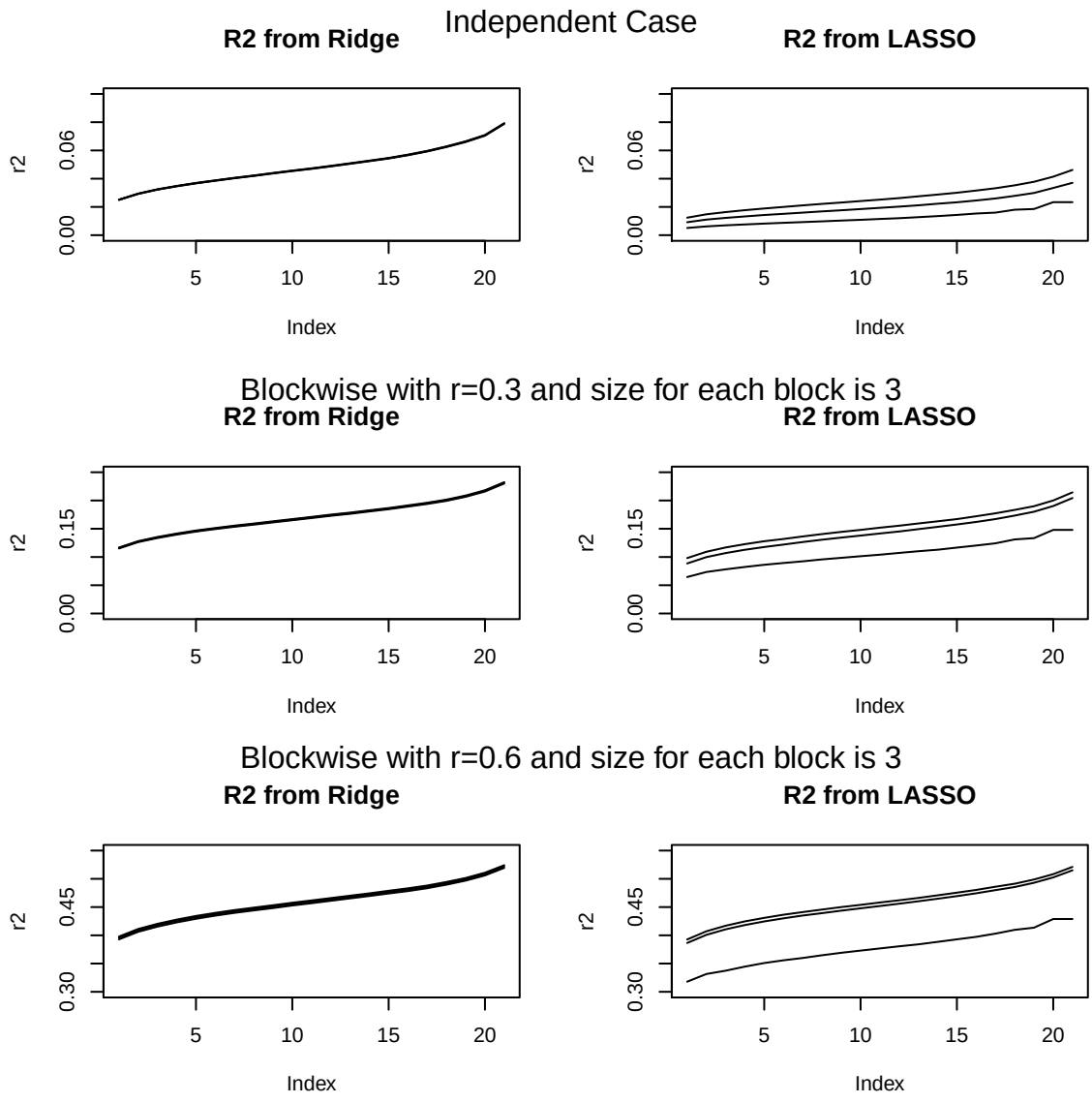


Figure 3.19: Comparing proposed functional summaries and summaries that derived from ridge squared correlations.

identify distinct structures. Again, we use the average Kullback-Leibler divergence as a measure of difference between the two structures. When n is 100, the distance between $\rho = 0$ and $\rho = 0.3$ is 14.88, the distance between $\rho = 0.3$ and $\rho = 0.6$ is 83.78, and the distance between $\rho = 0$ and $\rho = 0.6$ is 338.55. When n is 200, the corresponding distance will be 81.37, 192.73, and 1373.54, respectively. When n is 400, the corresponding distance will be 332.64, 502.64, and 5349.84, respectively. With more sample size, the distance becomes larger and the summaries also support these. However, the proposed functional summaries exist higher Kullback-Leibler divergence than ridge alternative.

3.9 Discussion

The proposed functional summaries can be used to describe the dependence structure among variables. Our summaries emphasize the degree by which each variable is predictable from the others, with a special focus on the number of variables required to predict another variable. We have shown that the proposed summaries have power to distinct different covariance structures. It allows us to compare structures with different dimensions. The bias correction is also provided.

In the simulation study, the sparse covariance structures are used. The functional summaries via LASSO have higher average Kullback-Leibler divergences than ridge alternative. When applied to gene expression dataset, we identify gene sets that have distinct dependence structures and confirm that the proposed functional summaries are reproducible.

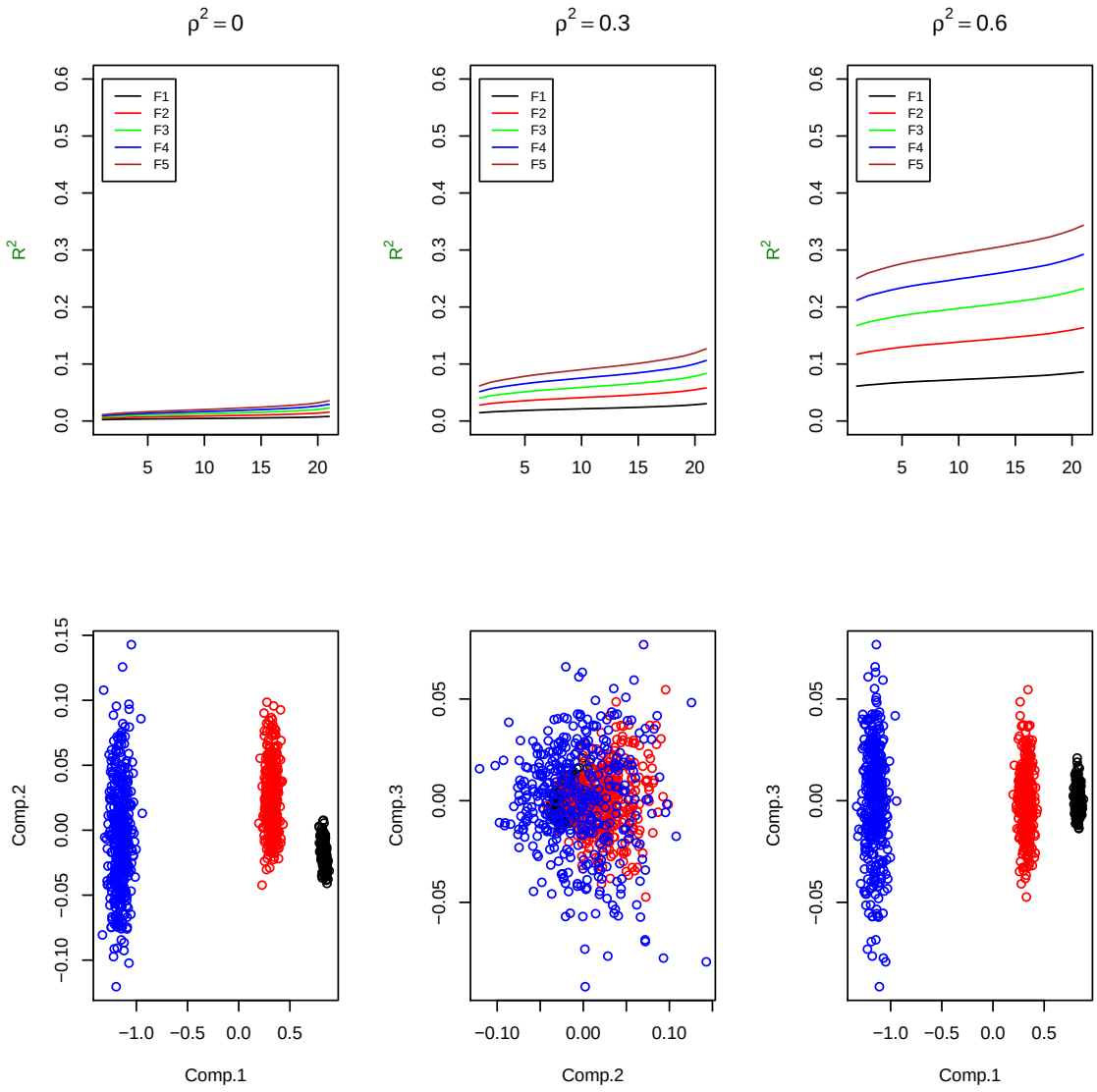


Figure 3.20: Functional summaries via ridge alternative with n be 400

CHAPTER IV

A Projection Based Approach for Exploring Conditional Correlation Paths

4.1 Introduction

To explore the dependencies among a set of variables, many existing methods start with the Pearson correlation coefficient (marginal correlation). It is a scale free measure that describes the degree to which two variables are related. It can be used to assess a direct relationship between two variables. A positive correlation indicates two variables vary together positively. In a negative correlation, two variables vary oppositely. This measure ignores effects from other variables. This may result in that the strong correlation between variables that might due to some indirect interactions or regulations by other variables. The marginal correlation does not allow us to elucidate such complicated relationships.

People then explore the dependencies via using partial correlation, which is the conditional correlation between two variables after removing effects that are due to all other variables, to overcome some disadvantages in marginal correlation. It determines the degree of dependency between two variables if influences from other variables are removed. When we take previously ignored variables into consideration, a well known statistical phenomenon called Simpson's paradox might occur. It is observed when the relationship between two variables is reversed after adjusting for other variables, or the relationship between two variables differs within subgroups

compared to that observed for the aggregated data.

The marginal correlation and the partial correlation are two example of correlation coefficients conditioned on linear statistics of the data, one condition on nothing (empty set) and the other conditions on all remaining variables. Falling between these extremes is the vast set of correlation coefficients conditioned on arbitrary linear statistics of the data. Techniques for exploring such huge space have only minimally developed. In this chapter, we propose a projection-based approach for exploring conditional correlations.

Two types of conditional correlations are considered. One is the expected value for conditional correlation and the other is to treat conditional correlation as a random variable such that the values will depend on the conditioned value. Both the marginal correlation and partial correlation mentioned above are of the first type. We propose a graphical tool that enable us to explore the change in dependencies from marginal correlations to partial correlations. This path is built via adding information from others gradually to reach partial correlations. This projection-based approach can also be applied to second type of conditional correlations. Note that the proposed approach can be applied to both correlation and covariance matrices.

The rest of of this chapter is organized as follows. In section 4.2, we describe the correlation paths for two types of conditional correlations. In section 4.3, we use simulation study as illustration. In section 4.4, we apply the correlation paths to the real data set.

4.2 Correlation Paths

Let $\mathbf{X} = (X_1, \cdots, X_p)^T$ denote the variables of interest. To infer the structure among a set of variables, one simple method is to compute the pairwise marginal

correlations,

$$\rho_{ij} = Cor(X_i, X_j), i, j = 1, \cdots, p. \tag{4.1}$$

The marginal correlation matrix describes the degree of relationships between a set of variables. For example, gene-gene interactions play an important role in biological processes. The study of correlated gene expression data is important because genes that are strongly correlated might have similar functions. Since the marginal correlation matrix is a p by p object, people devise approaches to summarize characteristics from the data set. These often come with visual methods.

The biplot introduced by Gabriel (1932) is an enhanced scatterplot that uses both points and vectors to represent data. Friedman and Tukey (1974) introduced projection pursuit to describe the process of finding interesting linear projections. Huber (1985) tries to find the most"interesting" possible projections via using a search algorithm that optimizes some fixed criterion of "interestingness". Targeted projection pursuit (Faith et al. (2006)) allows the user to explore the space of projections by manipulating data points directly in an interactive scatterplot. The "corrgram" proposed by Friendly (2002) displays not only the correlation magnitudes but also reorders the variables such that similar variables are positioned adjacently.

Since the high correlation between gene pairs might due to some indirect interactions or regulations by common genes. This naive correlation does not allow us to elucidate the complicated relationships. Instead, people use partial correlation, the conditional correlation between two variables after removing effects that are due to other variables, to determinate the dependency among variables. For a chosen pair

(X_i, X_j), the partial correlation coefficient is defined as

$$\begin{aligned}\rho_{ij\cdot V\backslash\{i,j\}} &= E[Cor(X_i, X_j|\mathbf{X}\backslash\{X_i, X_j\})] \\ &= E[(X_i - E[X_i|\mathbf{X}\backslash\{X_i, X_j\}])(X_j - E[X_j|\mathbf{X}\backslash\{X_i, X_j\}])]. \end{aligned} \tag{4.2}$$

In graphical Gaussian models (Dempster (1972), and Edwards (2000)), the precision matrix (the inverse of covariance matrix) $\mathbf{P}$ is used as a measure of conditional dependence of any two variables that are related to the partial correlation coefficients via

$$\rho_{ij\cdot V\backslash\{i,j\}} = -\frac{\mathbf{P}_{ij}}{\sqrt{\mathbf{P}_{ii}\mathbf{P}_{jj}}}.$$

Both $\{\rho_{ij}, 1 \leq i < j \leq p\}$ and $\{\rho_{ij\cdot V\backslash\{i,j\}}, 1 \leq i < j \leq p\}$ are expected values of conditional correlation coefficients, they capture the dependences among a set a variables, and each set contains $O(p^2)$ elements. Between these two extremes is the vast set of correlation coefficients conditioned on arbitrary linear statistics of the data. We provide a projection-based graphical tool to enable us connect these two sets.

Another type of conditional correlation matrix is to treat conditional correlation correlations as a random object that the values may vary when we vary the conditioned value. For example, for a given linear statistics, $\theta^T\mathbf{X}$ with $\theta = [\theta_{ij}]$, the conditional correlation matrix is defined as

$$Cor(\mathbf{X}|\theta^T\mathbf{X}). \tag{4.3}$$

The conditional correlation matrix in (4.3) is a p by p random object, the proposed graphical tool can also be applied to describe the change in dependencies.

4.2.1 Correlation Paths for Expected Values of Conditional Correlation Matrix

In this section, we build a correlation path to connect the marginal correlations $\{\rho_{ij}, 1 \leq i < j \leq p\}$ and partial correlations $\{\rho_{ij \cdot V \setminus \{i,j\}}, 1 \leq i < j \leq p\}$ and use it to describe the change in associations when more information is used to condition on. Let S^0 denote the marginal correlation matrix and S^* with $S^*_{ij} = \rho_{ij \cdot V \setminus \{i,j\}}$ being the partial correlation between X_i and X_j after removing the effects from the remaining variables. They describe the dependence at two extreme levels of conditioning. We explore the change between them by adding information gradually to condition on.

One might aim to find a sequence of projection matrices θ_ℓ, $\ell \in 1, \cdots, L$, of increasing rank such that $E[Cor(X|\theta_L^T \mathbf{X})]$ contains the same information as S^*. However, the off-diagonal terms of S^* are conditioned on different subsets of variables. No such sequence of θ_ℓs exist. Instead, we build a sequence of matrices that describe the dependence at certain levels of conditioning to connect them.

To build a path to link ρ_{ij} and $\rho_{ij \cdot V \setminus \{i,j\}}$ for a chosen pair (X_i, X_j), we start with marginal correlations and add variables one by one gradually to condition on to reach $\rho_{ij \cdot V \setminus \{i,j\}}$. There are $(p-2)!$ possibility to connect these two correlations. There are more possibilities to connect S^0 and S^*. Instead, we let the data determine its path automatically. We employ the principal component analysis to determine or order of conditioning for each chosen pair and hence build a sequence of matrices to connect S^0 and S^*. Let PC_k^{ij} be the first k principal components from $\mathbf{X} \setminus \{X_i, X_j\}$ and PC_0^{ij} be an empty set. We use $E[Cor(X_i, X_j|PC_k^{i,j})], k = 0, \cdots, p-2$ to connect marginal correlation and partial correlation. When k=p-2, $E[Cor(X_i, X_j|PC_k^{i,j})]$ is equivalent to $\rho_{ij \cdot V \setminus \{i,j\}}$ in (4.2). This is because the span of $PC_{p-2}^{i,j}$ is the span of $\mathbf{X} \setminus \{X_i, X_j\}$.

We can further set S^k be a p by p matrix with $S^k_{i,j} = E[Cor(X_i, X_j|PC_k^{i,j})]$, $k = 0, \cdots, p-2$. The S^ks describes the dependence structure at certain level of conditioning. When k increases, each element of S^k is conditioned on more informa-

tion from the remaining variables. Note that $S^{p-2} = S^*$ and S^k, $k \geq 1$, are not a conditional correlation matrix of the form $E[Cor(X|g(X)]$.

To explore the change in dependencies among a sequence of p by p objects, we compress them to lower dimensions to enable visualizations. We vectorize each S^k and apply the dimension reduction technique to compress them to lower dimensions. The principal component analysis is used as a dimension reduction technique. We project each S^k into 2 coordinates. Dots are then connected with a line and called it the correlation path. This enable us to visualize the change in dependence structure among a set of variables. When a short path is observed, the cumulative difference in dependence structure among consequent S^ks is relative small compare to a long path. When a path is loop around, we expect the difference between two arbitrary S^ks to be small and hence the difference from the marginal correlations to partial correlations is small. When a straight line is observed, the change in the difference of consequent S^ks moves toward a fixed direction, but the increments may not be constant. One application of correlation paths is to screen many data sets, to quickly see which sets of variables have correlation structures that differ from the others.

4.2.1.1 Correlation Path From Sample Covariance Matrix

In this section, we show that each step of conditional correlations, S^k, used to derive the correlation path can be derived from the sample covariance matrix, provided that the regression model is used to remove the effects from others. That is, the entire correlation path can be derived from the sample covariance matrix. Without loss of generality, let $\mathbf{x} = (x_1, \cdots, x_p) \in \mathbf{R^{n \times p}}$ be the observed data with empirical means are zero. First, the marginal correlation matrix can be derived directly from the sample covariance matrix. The next step is to show that each conditional correlation among arbitrary chosen pair can be derived from the sample covariance matrix.

To ease notation, we let $W = (x_i, x_j) \in \mathbf{R^{n \times 2}}$ be the chosen variables and $Z =$

$\mathbf{x}\backslash\{x_i, x_j\} \in \mathbf{R^{n\times(p-2)}}$ represents the remaining variables , $1 \leq i < j \leq p$. The eigen decomposition of $\widehat{Cov(Z)}$ is $P\Lambda P^T$ with $\Lambda = diag(\lambda_1, \cdots, \lambda_p)$. If we let P_k be the first k columns of P, ZP_k will be the first k principal components. The conditional correlation among W after removing the effects from the first k principal components of Z can be derived from the covariance of residuals of ZP_k regressed on W. The residuals are derived as follows:

$$\hat{e} = \{I - ZP_k((ZP_k)^T ZP_k)^{-1}(ZP_k)^T\}W.$$

We then have

$$\begin{aligned}
\widehat{Cov(e)} &= \frac{1}{n}W^T\{I - ZP_k((ZP_k)^T ZP_k)^{-1}(ZP_k)^T\}W \\
&= \frac{1}{n}W^T W - \frac{1}{n}\{W^T ZP_k\}\{n * diag(\lambda_1, \cdots, \lambda_k)\}^{-1}\{(ZP_k)^T W\} \\
&= \frac{1}{n}W^T W - \frac{1}{n}\{W^T ZP_k\}\{diag(\lambda_1, \cdots, \lambda_k)\}^{-1}\frac{1}{n}\{(ZP_k)^T W\}.
\end{aligned}$$

Using sample covariance matrix to simplify, we have

$$\widehat{Cov(e)} = \widehat{Cov(W)} - \{P_k^T \widehat{Cov(Z, W)}\}^T * diag(\lambda_1^{-1}, \cdots, \lambda_k^{-1}) * \{P_k^T \widehat{Cov(Z, W)}\}. \tag{4.4}$$

We then have the conditional correlation among W after removing effects from first k principal components of Z. Since $\widehat{Cov(W)}$, $\widehat{Cov(Z, W)}$, P_k, and $(\lambda_1, \cdots, \lambda_k)$ can be derived from sample covariance matrix, each element of S^k can be derived from the sample covariance matrix.

4.2.2 Correlation Paths For Conditional Correlation Matrix Conditioned On Linear Statistics

Another type of conditional correlations matrix is to treat the conditional correlations as a random object. Conditioned on linear statistics $\theta^T\mathbf{X}$, the conditional correlation matrix in (4.3) is a p by p object that varies over the range of $\theta^T\mathbf{X}$. This object can be estimated by using local correlation estimates. To get preliminary information about the conditional dependence, we use the projection procedures described in the previous section to project a sequence of conditional correlation matrices to lower dimensions to enable visualization.

4.3 Simulation Study

In this section, we use simulation studies as illustrations and start with the correlation path for $S^0, \cdots, S^{p-2}$. The correlation path that treats correlation matrix as a random object will be discussed later. We start with three scenarios on correlation matrix: exchangeable, blockwise exchangeable, and autoregressive. In the blockwise exchangeable structure, we let m be the number of variables within each block if p is a multiple of m. If not, all blocks except for the last block have m variables, and the last block has $p\%m$ variables with $\%$ being a modulo operator. For each structure, we use r as parameter to describe association among variables. In exchangeable structure EX(r), the correlations for arbitrary chosen pair is r. In blockwise-exchangeable structure B-EX(r), the correlation is r if variables are in the same block and 0 otherwise. In the autoregressive structure AR(r), the correlation between X_i and X_j is $r^{|i-j|}$.

4.3.1 Simulation Results for Exchangeable Structure

Figure 4.1 displays the results for exchangeable structure EX(r) with different n and p. We let the non-zero values of r be 0.2 and 0.4 and sample 20 replicates from each structure. In addition, 20 replicates from independent normal are added to the graph. Within each graph, row one through three contain results for sample sizes 50, 100, and 200, respectively. The columns represent different values of p. Column one with sample size 10 and column two with sample size 20. Each line represents a replicate and different colors indicate different dependencies. The thick dot is the initial point of the correlation path which is the projected value for the vectorized version of the marginal correlations S^0.

We find that correlation paths for EX(0.4) generally have the longest trajectory on the correlation path while comparing to EX(0.2) and independent Gaussian. The difference between first two steps of correlation path in EX(0.4) is the largest and the difference between the remaining consequent steps are small. The correlation path for EX(0.2) has the similar forms. For further investigate, we can examine the pairwise correlations from the marginal correlations to the partial correlations. Figure 4.2 plots all pairwise correlations given $n = 200$, $r = 0.4$, and $p = 20$. Each line represents a chosen pair and describes the change from the marginal correlation to the partial correlation. The difference from the marginal correlations to the first conditional correlations is the largest for each pair and the changes for remaining conditional correlations were relative small. Results for different r and p.

Since the structure is known, we can derive the conditional correlations. Assuming that the data is from a multivariate Gaussian distribution, we can derive the conditional correlation between chosen pair $W = (X_i, X_j)$ conditioned on first k principal components of $Z = \mathbf{X}/Y$ via calculate the conditional covariance,

$$Cov(W) - \{P_k^T Cov(Z, W)\}^T \{P_k^T Cov(Z) P_k\}^{-1} \{P_k^T Cov(Z, W)\}, \tag{4.5}$$

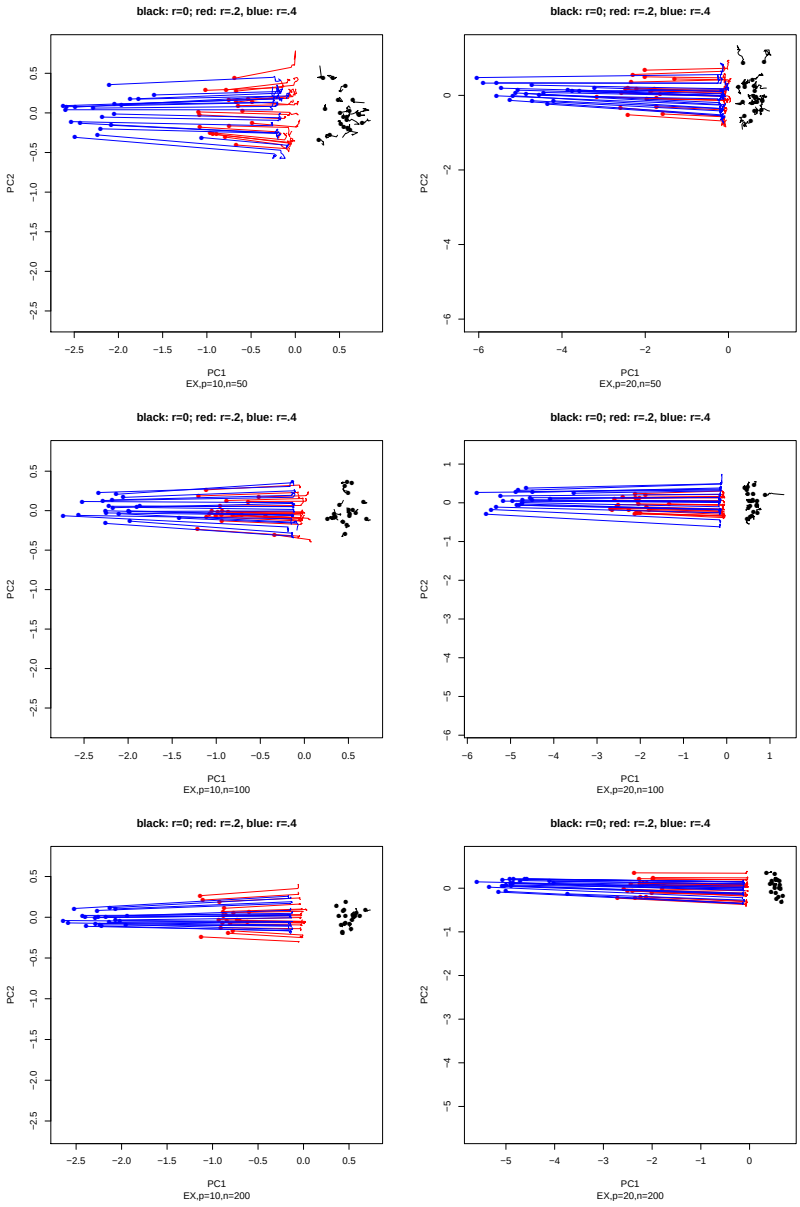


Figure 4.1: Correlation path under exchangeable structure. While r is larger, it has linger trajectory. The difference from the projected value of S_0 to the projected value of S_1 is the largest.

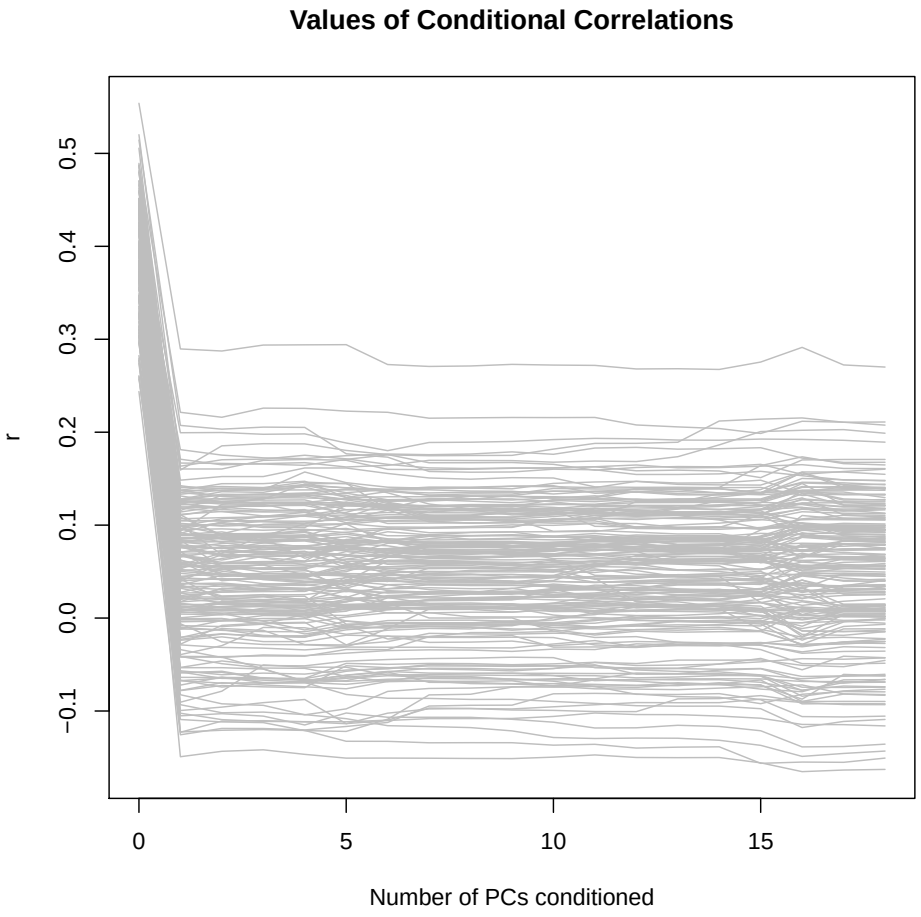


Figure 4.2: Pairwise correlations given $r = 0.4$, and $p = 20$ with sample size 200.

where P_k is the first k eigenvectors from $Cov(Z)$. The derivation of (4.5) is straightforward.

In exchangeable structure, the dependence structure for Z is still exchangeable. The first eigenvalue and eigenvector are $1+(p-3)r$, $\frac{1}{\sqrt{p-2}}(1,\cdots,1)^T$, respectively. Since $Cov(Z,W)$ is a $(p-2)$ by 2 matrix with constant value r, all eigenvectors except for the first one times $Cov(Z,Y)$ will be a zero vector. This means that the conditional correlation conditioned on 2 or more principal components is equivalent to the conditional correlation conditioned on the first principal component. The conditional covariance conditioned on the first k principal components is

$$\begin{pmatrix} 1-\frac{(p-2)r^2}{1+(p-3)r} & r-\frac{(p-2)r^2}{1+(p-3)r} \\ r-\frac{(p-2)r^2}{1+(p-3)r} & 1-\frac{(p-2)r^2}{1+(p-3)r} \end{pmatrix},$$

$k \geq 1$. Except for the marginal correlation, the conditional correlations conditioned on first k principal components of $Z = \mathbf{X}/Y, k \geq 1$, have a constant value

$$\frac{r-r^2}{1+(p-3)r-(p-2)r^2}.$$

4.3.2 Simulation Results for Blockwise-Exchangeable Structure

Figure 4.3 displays the results for blockwise-exchangeable structure B-EX(r) . The organization of graphs is the same as the organization in Figure 4.1 except using blockwise-exchangeable structure to replace exchangeable structure. When p is 10, the sample size required to separate them well is 200. When p is 20, the sample size required to distinguish them well is 100.

The results show that under appropriate sample size, the correlation path for each structure preserves certain forms. The structure B-EX(0.4) had the longest trajectory. We examine all pairwise correlations for one replicate in Figure 4.4. The sample size is 200, p is 20, and r is 0.4. With the blockwise-exchangeable structure, we find

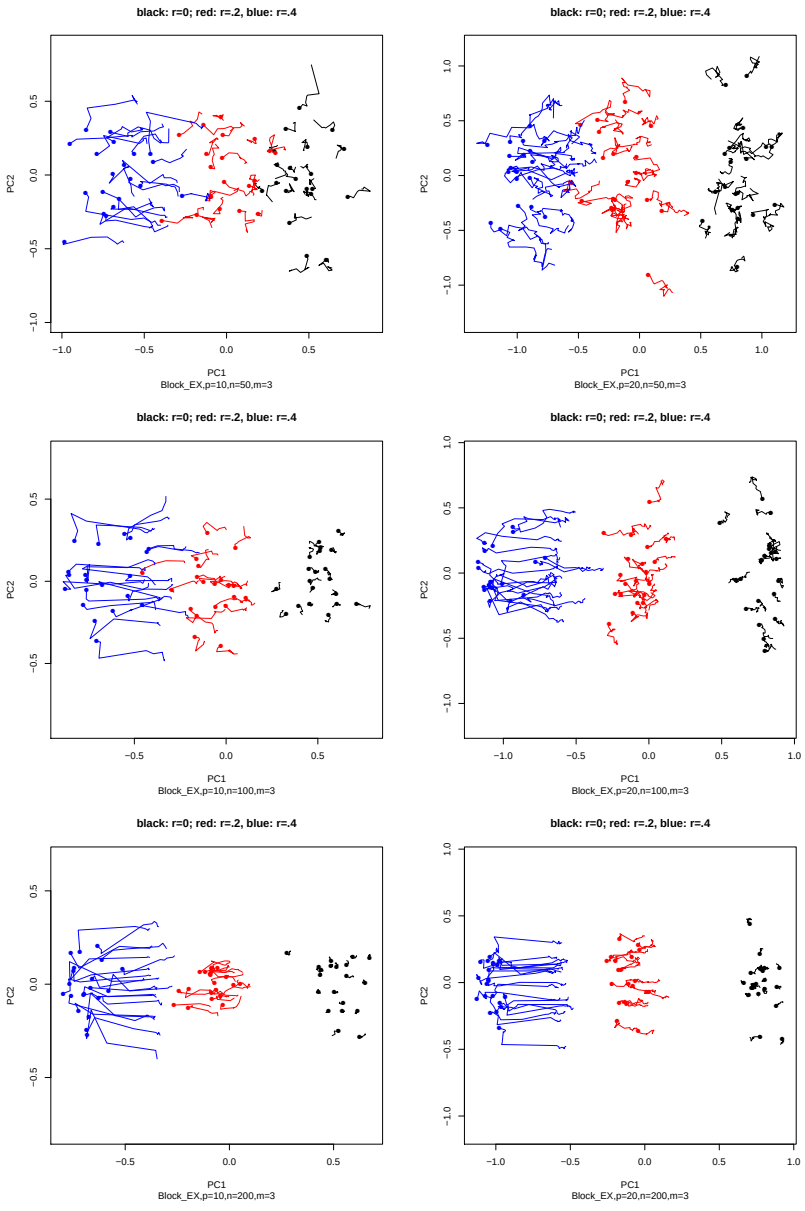


Figure 4.3: Correlation path under blockwise-exchangeable structure.

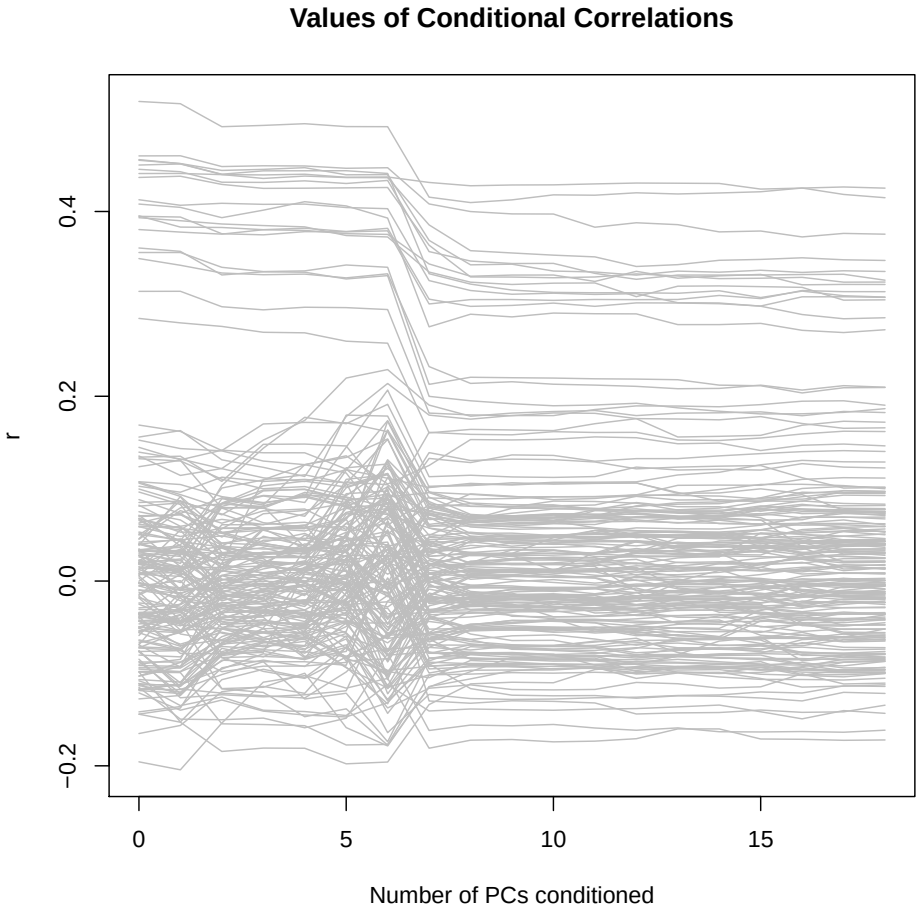


Figure 4.4: Pairwise correlations given $r = 0.4$ and $p = 20$ with sample size 200 under blockwise-exchangeable structure.

that most of them have correlations around zero and some of them have nonzero correlations. For those that have non-zero marginal correlations, the conditional correlation would stay constant for a while and then drop by certain level and them remain constant. This is because those curves indicate two variables in the same block and variables from other blocks form PCs that have larger variability. For those that had marginal correlations around zero, the variability in first couple conditional correlations are quite irregular, but the variability will remain stable eventually.

Remark. While deriving the theoretical conditional correlations for blockwise-exchangeable structure, it depends on whether the chosen pair are from the same

block or not and the size of each block. When two variables are from different blocks, the principal components from the remaining variables will be dependent with at most one of them. The conditional correlations then stay unchanged. However, the fluctuation from sample conditional correlations might exist. When two variables are from the same block, only one variable left given $m = 3$ and that variable will form a principal component itself. Other blocks will form different principal components and some of them have larger variance than the principal component from the chosen block. The corresponding conditional correlation will stay the same for a while, decrease at certain point, and then remain constant.

4.3.3 Simulation Results for Autoregressive Structure

Figure 4.5 represents results under different p and n. Three structures can be separated given the sample size is at least 100. Three structures have different initial points on the correlation path since the marginal correlation matrix are different. When r decreases, the marginal correlation matrix becomes more sparse and hence the shorter trajectory of correlation path. The correlation path for AR(0.4) generally has longer trajectory than others.

Figure 4.6 plots one realization of all pairwise correlations under $r = 0.4$ and $p = 20$. Results indicate that some of them have marginal correlations that are away from zero and their conditional correlations decrease gradually and end with nonzero partial correlations. For those that have marginal correlations around zero, their conditional correlations move toward zero.

We put the correlation paths from three structures together. Since the AR and B-EX have many marginal correlations that are zero or close to zero. We expect that the initial value of the correlation paths will stay relative close when comparing all three structures. In addition, the projected values for S^1 will drop a lot in the EX structure. The finite sample replicates in Figure 4.7 verify these expectations.

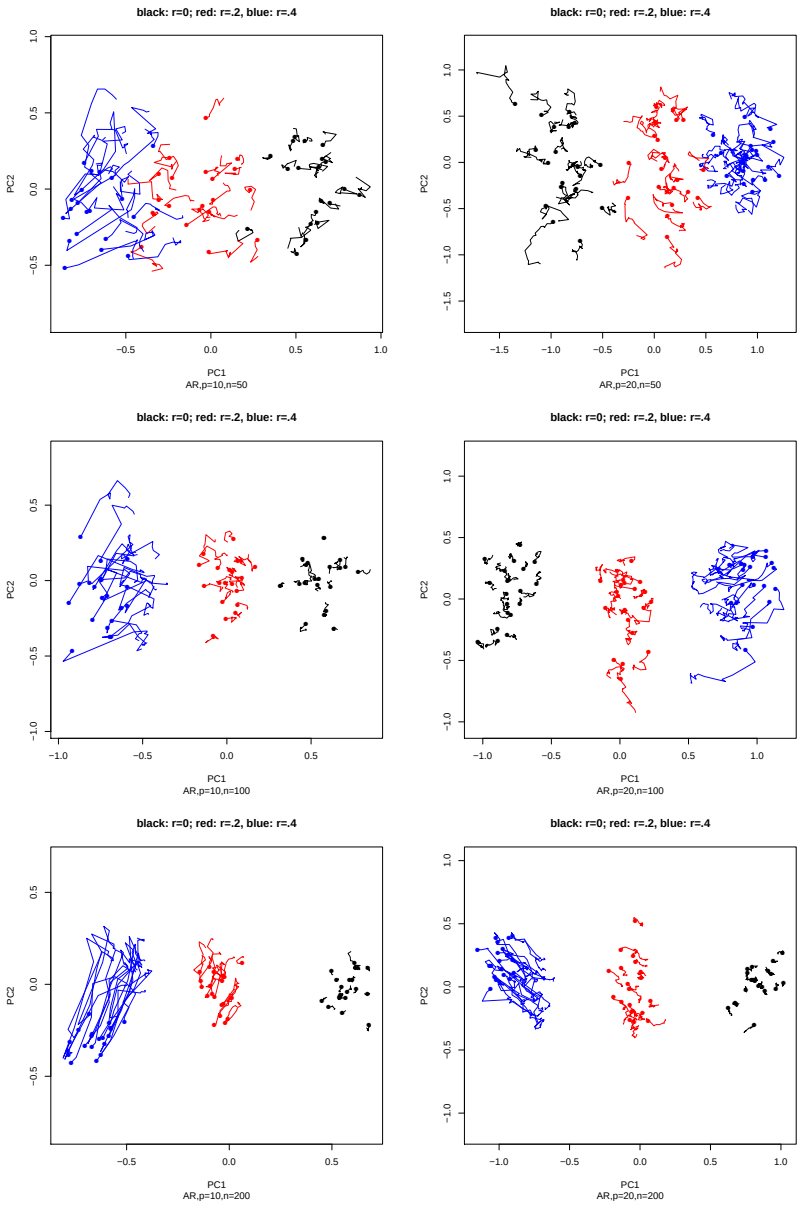


Figure 4.5: Correlation path for AR structure.

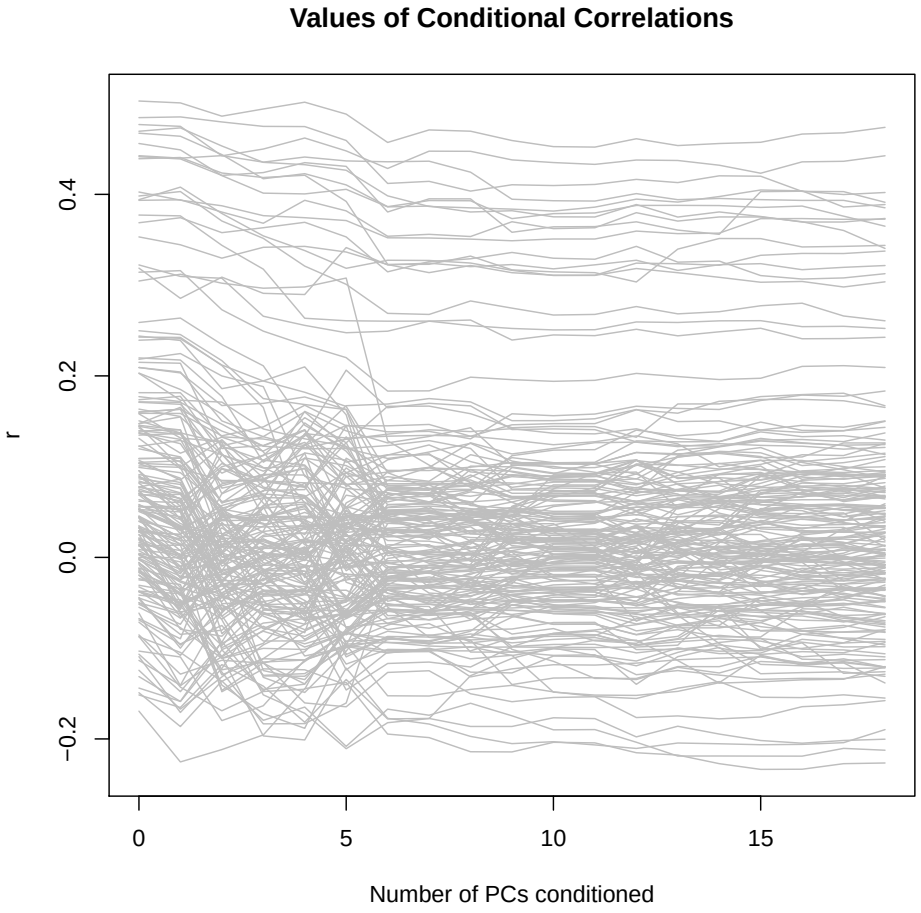


Figure 4.6: Pairwise correlations given $r = 0.4$ and $p = 20$ with sample size 200 under autoregressive structure.

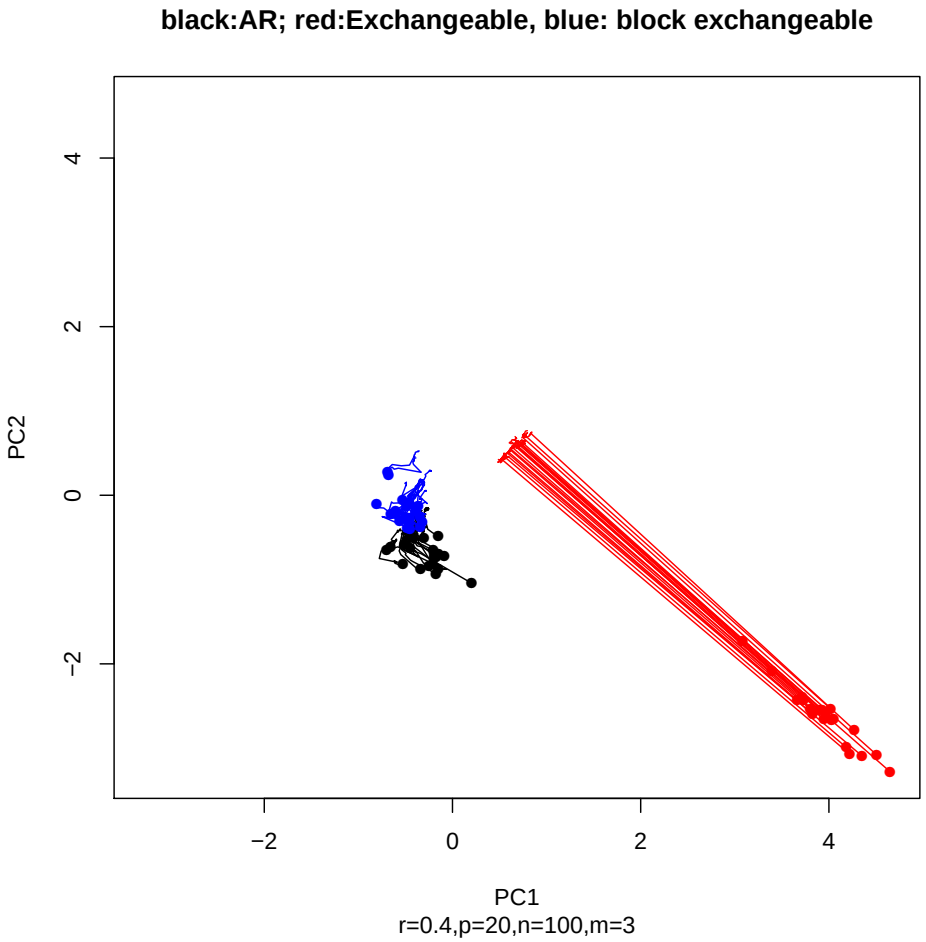


Figure 4.7: Correlation paths from three structures with fixed $r = 0.4$.

4.3.4 Factor Structures

Next, we let the marginal covariance matrix that satisfies the following form

$$D + ABA', \tag{4.6}$$

where D is a diagonal matrix, A be a rank q orthogonal matrix and B be a q by q matrix with $q << p$. In the simulation below, we set $p = 20$ and $q = 2$ and B be

$$\begin{pmatrix} 1 & r \\ r & 1 \end{pmatrix} \tag{4.7}$$

with different r values.

The first graph in Figure 4.8 displays the correlation paths for r being -0.5, 0, or 0.5. Since the rank for ABA' is 2, the rank for submatrix of the ABA' is still 2. For a chosen pair, the first two principal components from remaining variables tends to explain more variability than others and other principal components are independent of chosen pair. The first 2 steps of correlation paths then vary a lot. Different r values have different marginal correlation structures and hence different initial points on the correlation path. The graphs in the second row of Figure 4.8 display conditional correlations for all pairs. This graph ignores the indices of the chosen pairs, we found that the two graphs have similar behavior. If we take indices into consideration, they behave differently. In the third row of Figure 4.8, we take 10 indices pairs from each structure and draw the change from marginal to partial correlations. Same colors indicate same indices and graphs show that they behave differently.

4.3.5 Sample from Similar Structures

In this section, we use simulation study to claim that the correlation path can be used to distinguish different structures given the sample size is appropriate. We use

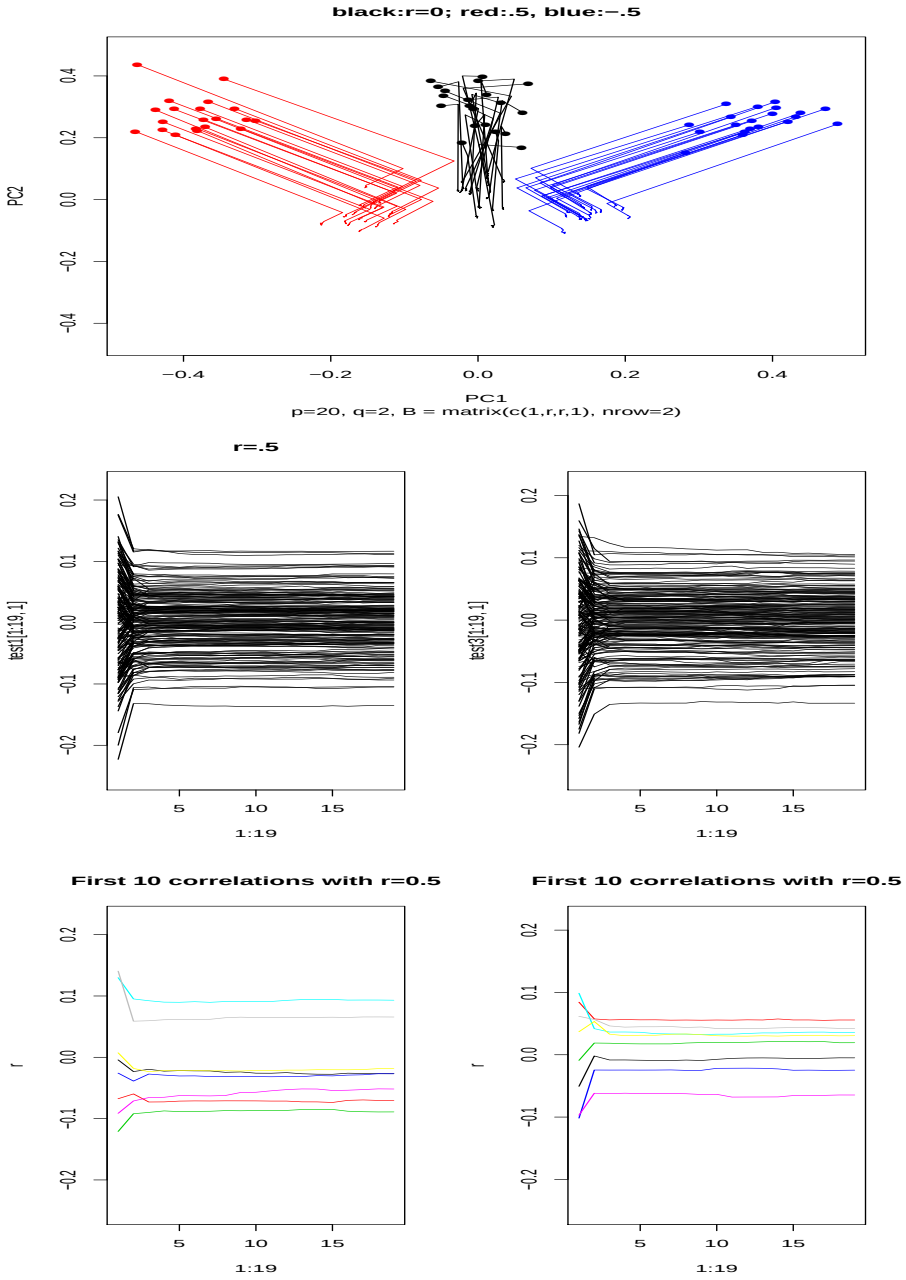


Figure 4.8: Correlation paths from factor structure with $q = 2$. The first row represents correlation paths, the second row plots pairwise correlations for r being -.5 and .5, and the third row plots 10 indices pairs.

the scaled Wishart distribution $\frac{1}{m}W_p(V, m)$, the Wishart distribution scaled by its degree of freedom, to generate similar structures. We treat the samples from

$$\frac{1}{m}W_p(AR(0.2), m)$$

as given covariance structures and sample data using multivariate Gaussian with those structures. As m becomes larger, the sampled structures would be more similar. In Figure 4.3.5, we consider 3 m values. In the top-left, the m is 100 with sample size 100. In the top-right, the m is 200 with sample size 100. In the lower-right, the m is fixed at 500. The lower-left with sample size 100 and lower-right with sample size 50. The results show that the correlation paths for samples from the same structure tend to cluster together. When m is larger, the underlying structures were be similar, and exist overlaps among correlation paths. Large sample size are then required to distinguish them.

4.3.6 Simulation Study for Correlation Paths Conditioned on Linear statistics

In this section, we use simulation study to show that correlation paths conditioned on linear statistics of the data. We apply the singular value decomposition on the sample covariance to get the orthogonal matrix and each column of orthogonal matrix is treated as a projection direction θ. These directions are equivalent to the loadings in the principal component analysis on the sample data. In application, the directions of θ may depends on researchers' interests. The local estimate with Epanechnikov kernel is used to estimate the conditional correlations conditioned on l-th quantile of $\theta^T\mathbf{X}$, $l \in \{5, 10, \cdots, 95\}$, provided that there are at least 30 observations for each quantile. Each conditional correlation matrix is vectorized, projected to lower dimension, and the dots with same θs are linked to create the correlation path.

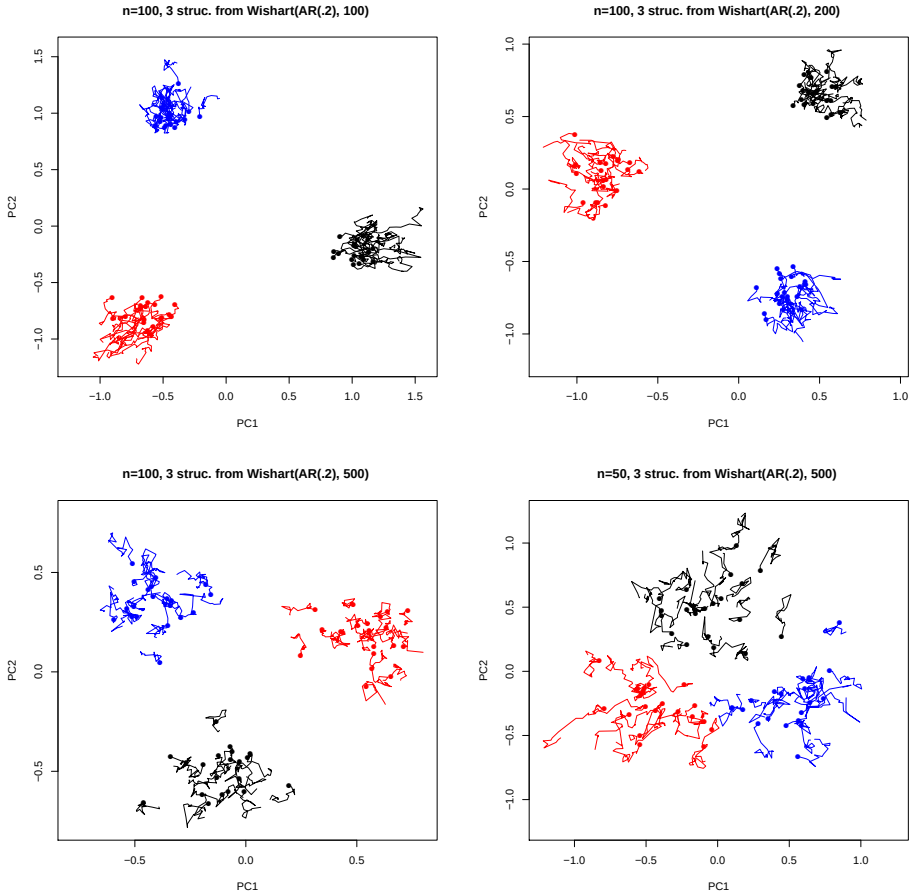


Figure 4.9: Generate data from similar structures and claim that the correlation paths can be used to distinguish them when the sample size is appropriate.

If the data is sampled from the multivariate normal distribution, the conditional correlation matrix conditioned on a linear statistic $\theta^T\mathbf{X}$ would be a constant matrix for any $\theta^T\mathbf{X} = u$. The conditional covariance conditioned on a linear statistic $\theta^T\mathbf{X}$ is known to be

$$\Sigma - (\theta^T\Sigma)^T\{\theta^T\Sigma\theta\}^{-1}(\theta^T\Sigma) \tag{4.8}$$

That is, the analytical correlation path under a given direction will degenerate to a single point.

Figure 4.10 plot the correlation paths under different sample sizes and population structures with $p = 20$. For the correlation paths, we plot the first five and last five correlation paths. The solid line is used to indicate the first five correlation paths and dashed is for the last five paths. The black lines are the correlation path that conditioned on the first or last direction, the red lines are conditioned on the second or second to last direction, the green lines are for the third or third to last direction, the blue lines are for the fourth or fourth to last direction, and the gray lines are for the fifth or fifth to last direction. The dot labels are analytical projected values conditioned on first five directions, the triangular labels are for last five directions, and the color is used to indicate the order of directions.

As the sample size becomes larger, the distance from the sample correlation path to the analytical value decreases and some of the correlation paths are less tangled with others. The trajectories of the correlation paths became shorter. The correlation path that conditioned on a direction that have larger variability have inclination to move further away from others. These segregations are more clear when the sample size is larger. In Figure 4.11, we increase the sample size to 10,000 and the corresponding correlation paths tend to degenerate to a single point and each path is close to the analytical projected value.

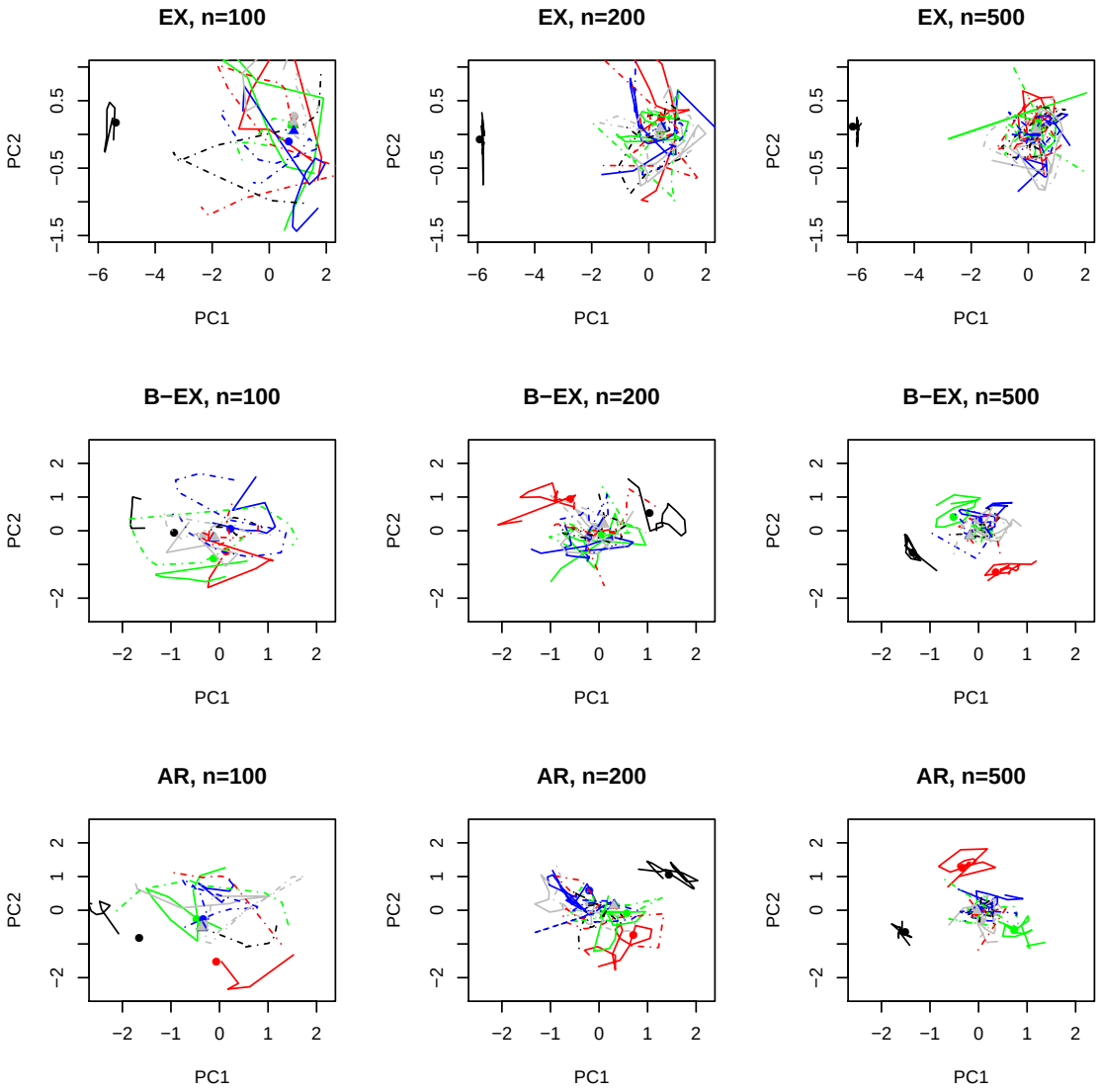


Figure 4.10: Correlation path under different structure with different sample size. While sample size is larger, the trajectories of correlation path becomes shorter and the analytical projected values are close to the correlation paths.

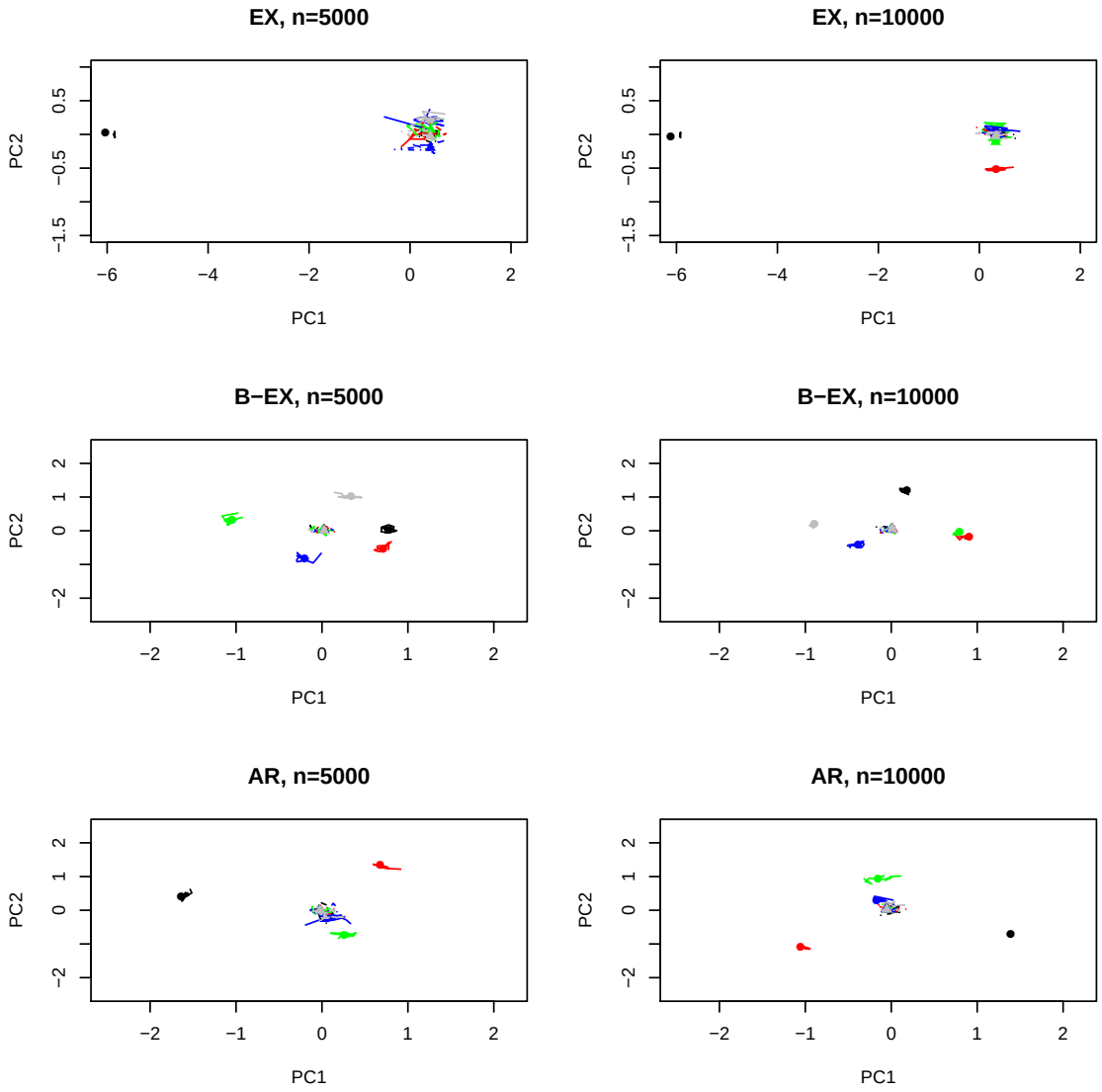


Figure 4.11: Correlation paths under different sample size. While sample size is larger, the length of correlation path becomes shorter and the analytical projected values are close to the correlation paths.

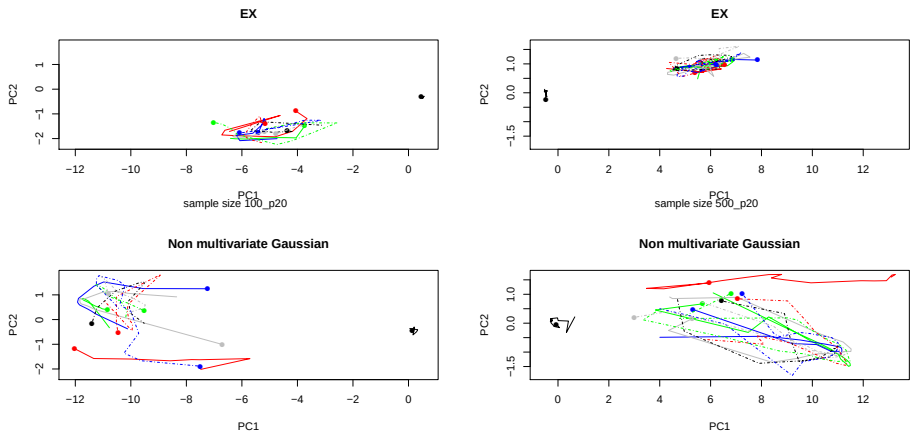


Figure 4.12: Correlation paths for Gaussian and Non-Gaussian structures.

Next, we apply the correlation paths to the non multivariate normal data. To begin with, we sample data from the multivariate normal $N(0, \Sigma)$ and then replace some of them to achieve non-Gaussianity by the following criterion:

- If X_p is less than 0, replace X_j by $X_p + 0.2 * N(0, 1), j = 1, \cdots, p - 1$.

In Figure 4.12, we sample from EX(.4) and non-Gaussian structure. The non-Gaussian structure is derived from a EX(.4) structure. As the sample size increases, the correlation paths for non-Gaussian data did not degenerate. As a byproduct to diagnosis Gaussianity, the large sample size is required.

4.4 Application to Normal Heart Tissue

In this section, we apply the correlation path to the gene expressions in the heart tissue from the left ventricular free wall of organ donors with no diagnosed heart disease. Heart tissue was collected by the Cleveland Clinic Kaufman Center for Heart Failure human heart tissue bank ($n = 108$) between August 1993 - May 2005. There are 33297 gene expressions. We consider apply the correlation paths for gene sets in the Molecular Signatures Database (MSigDB). We use the gene set class C1: Positional gene sets for each human chromosome and cytogenetic band. Gene sets

are corresponding to each human chromosome and each cytogenetic band that has at least one gene.

4.4.1 Correlation Paths for $S^0, \cdots, S^{p-2}$

To begin with, we apply the correlation paths to visualize the change from marginal correlations to partial correlations. The first step is to apply correlation paths on S_ks. To compare gene sets, there exist a limitation. The number of match genes for different gene sets generally are different. We have to narrow down the number of genes to have same number of genes. For example, we use gene sets with number of matched genes ranging from 10 to 20, and randomly select 10 variables from each gene set to build the correlation paths for all matched gene sets.

Figure 4.13 represents the correlation paths for those gene sets that the number of genes used is 10. There are 47 correlation paths (gene sets) in the graph. This figure can be used to visualize what is similar and what is different. After we condition on more other information, most of the correlation paths have the tendency to move toward the origin. The conditional correlations for different gene sets are more similar than the marginal correlations. Different gene sets have different paths and we label some of them and examine them.

Figure 4.14 plots all pairwise movements from marginal correlation to partial correlation for the labeled correlation paths in Figure 4.13. Each line represents a chosen pair and describes the changes from marginal correlation to the partial correlations as we add more informations from other variables to condition on. Since all four labeled correlation paths have different initial and end points, the corresponding behaviors on marginal correlations and partial correlations are different.

The red correlation path in Figure 4.13 has the longest trajectory and its first step changed a lot and so is its second step. The corresponding pairwise correlations (top-left in Figure 4.14) reflects this situation. Most of the correlation continued to

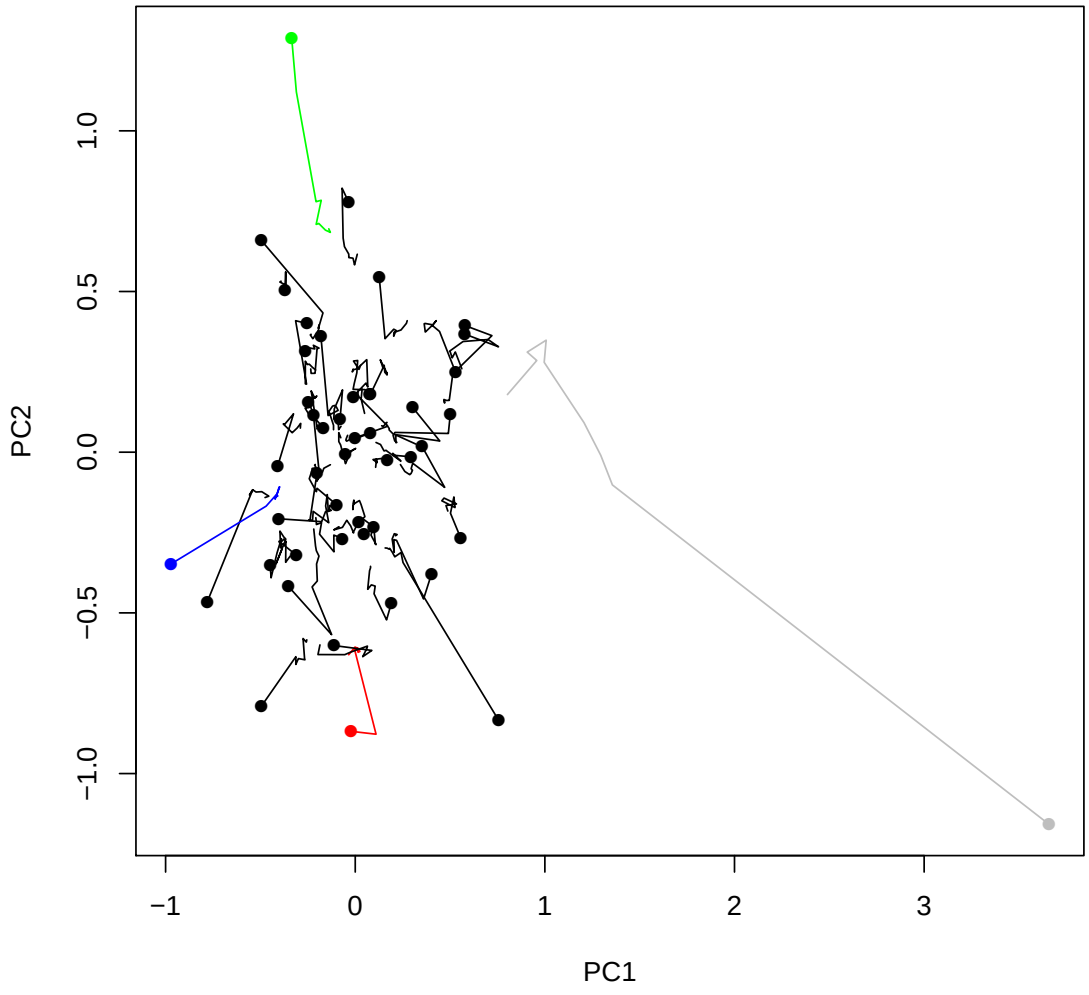


Figure 4.13: Correlation paths for gene sets with $p = 10$.

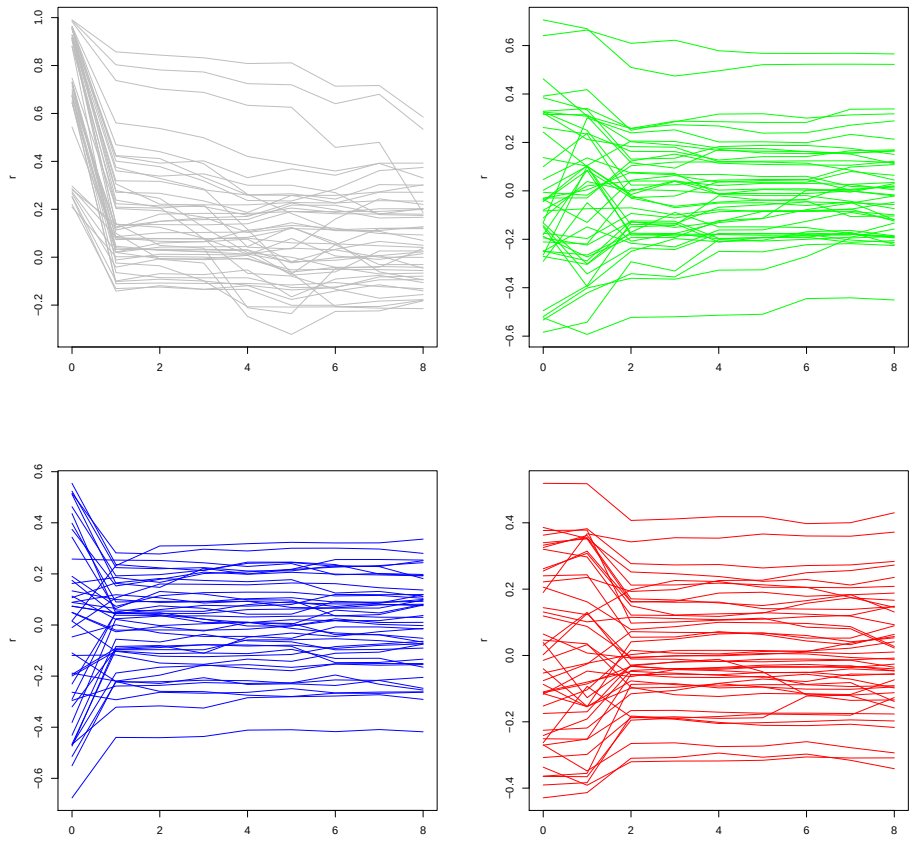


Figure 4.14: All pairwise conditional correlation for 4 selected correlation paths.

drop as we add more information to condition on. Some of them are reverted and the corresponding correlation path reflect this.

The green correlation path in Figure 4.13 has its difference from the second step to the third step was the largest. While examining the pairwise correlations (top-right in Figure 4.14), the correlations tend to diminish eventually. But the first two conditional correlations vary a lot. Since some of the first conditional correlations move away from zero and most of the second conditional correlations move toward zero, we expect the second step of correlation paths moved more toward zero than the first step.

The blue correlation path in Figure 4.13 has its first step changed a lot and remainings have very small movement. The pairwise correlations (bottom-left in Figure 4.14) show that most of the first conditional correlation move toward zero and the fluctuations for remainings are quite small.

The first two steps of gray correlation path in Figure 4.13 change a lot and the remaining steps have little movements. The pairwise correlations (bottom-right in Figure 4.14) showed most of the first conditional correlation stay the same or don't move toward zero, but most of the second conditional correlations move toward zero. The fluctuations for remainings paths are quite small.

Next, we use the gene sets with the number of matched genes ranging from 20 to 30 and randomly select 20 genes from each set. Figure 4.15 represents the correlation paths for those gene sets that the number of genes used is 20. There are 46 correlation paths (gene sets) in the graph. Since there are more variables within each gene set, there are more pairwise correlations. The correlation path can be used to visualize what is similar and what is different. After we condition on more information, some of the correlation paths are still away from the origin. Different gene sets have different paths and we label some of them and examine them.

Figure 4.16 displays pairwise correlations for selected correlation paths. Since p

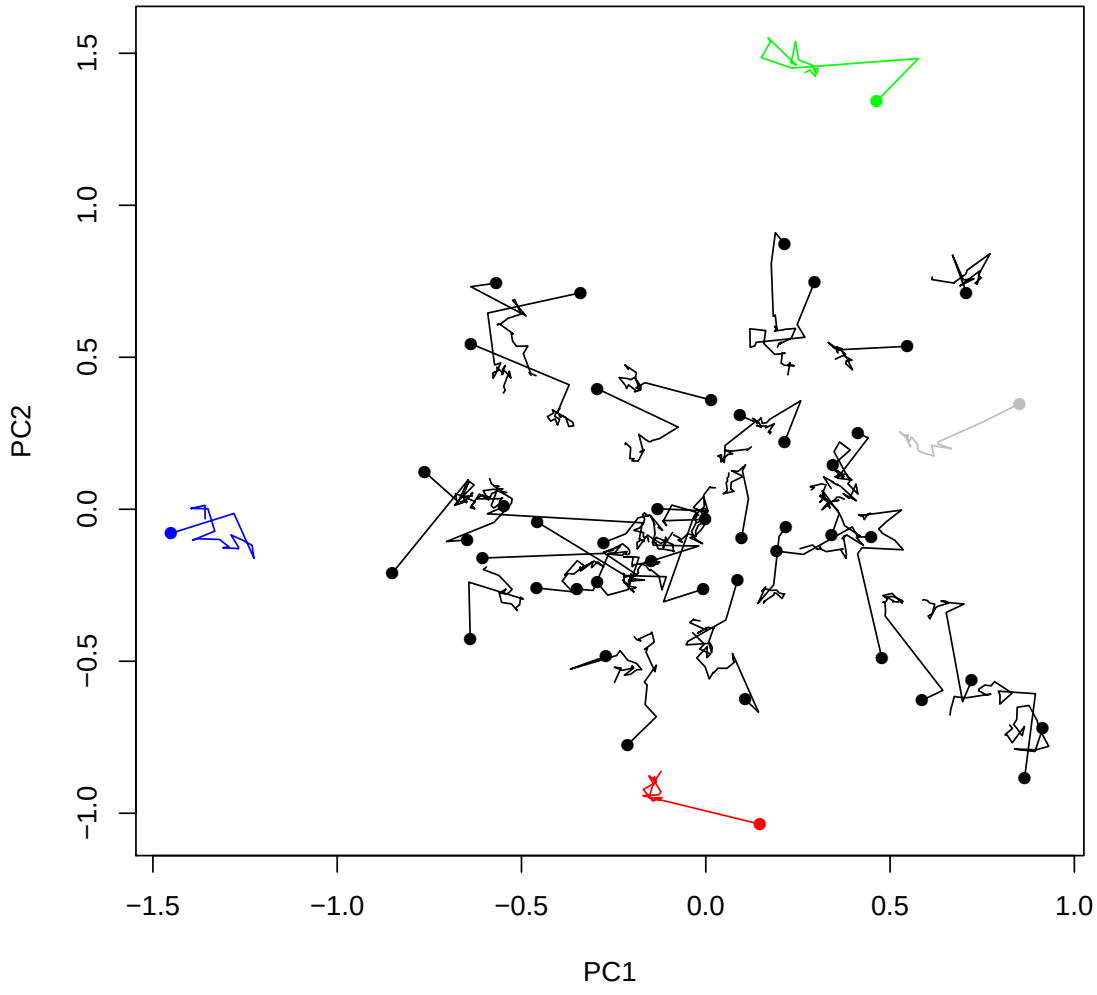


Figure 4.15: Correlation paths for gene sets with $p = 20$.

is doubled, each plot have about 4 times the pairwise correlations than $p = 10$. Four graphs had different overall patterns. Most of them have their correlations shrunk eventually. Some of them have their correlation diminished, but reverted at the end.

While the blue correlation path in Figure 4.15 does not move toward zero. It stays around its projected value for S_0, the corresponding pairwise correlations (bottom-left in Figure 4.14) show that most of them don't move toward zero. The same situation can be applied to the green correlation path. However, its first two steps varied more than others, the corresponding pairwise correlations (top-right in Figure 4.14) reflected these.

The pairwise correlations (top-left in Figure 4.14) for red correlation path in Figure 4.15 show that most of the first two conditional correlations changed a lot. The pairwise correlations (bottom-right in Figure 4.14) for red correlation path in Figure 4.15 show that most of the first conditional correlations changed a lot and remaining conditional correlations stay roughly the same.

4.4.2 Correlation Paths Conditioned on Linear Statistics of the Data

Next, we apply the correlation path to data that conditioned on a linear statistic of the data. That is, the conditional correlation matrix is a random object. The correlation paths were apply to the match gene sets with p reduced to 10 or 20. Figure 4.17 and Figure plot the correlation paths conditioned on a linear statistics. The data sets used are the labeled correlation paths in previous section. The projection directions are derived from the principal components. Except for the first couple of projection directions, all other directions had small variability. We found that the the correlation path for directions with small variability tend to tangled together. To further investigate, we can plot the range against each projected score or plot all pairwise correlations on each direction.

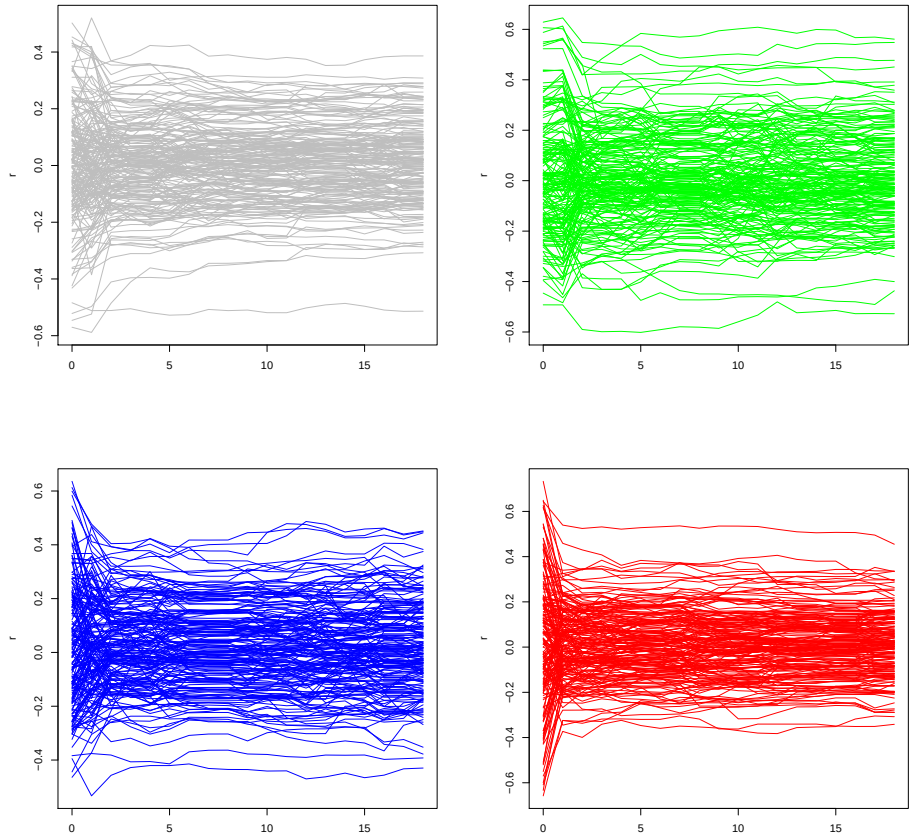


Figure 4.16: All pairwise conditional correlations for 4 selected correlation paths.

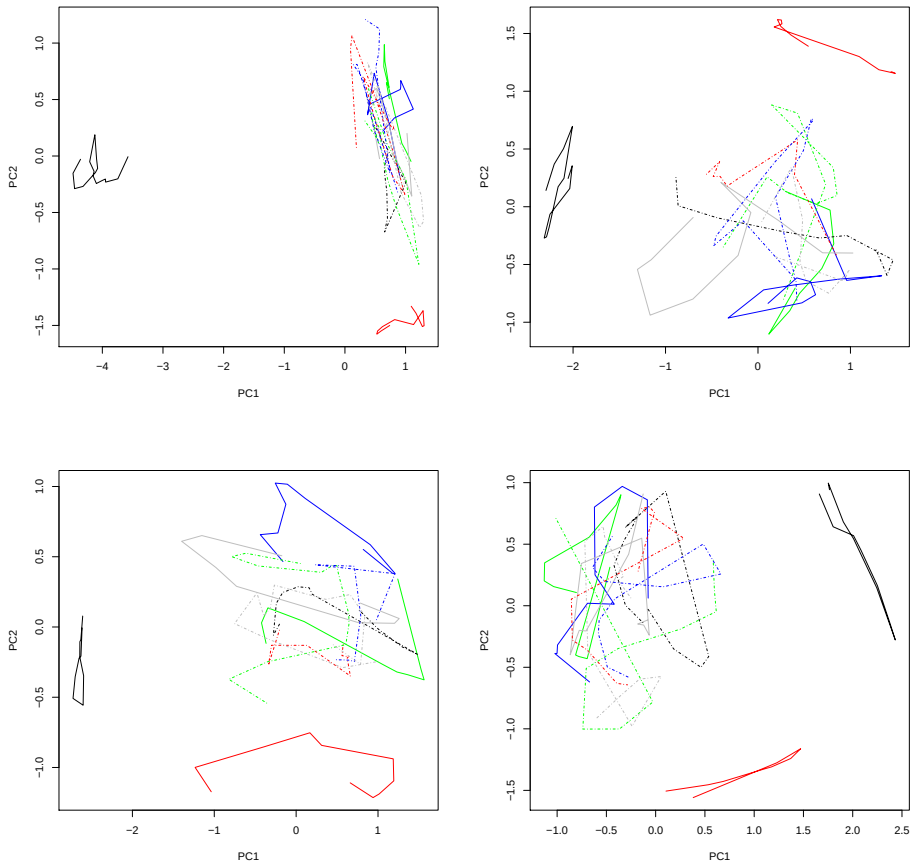


Figure 4.17: Correlation path conditioned on a linear statistics. The data sets are the labeled data sets in Figure 4.14.

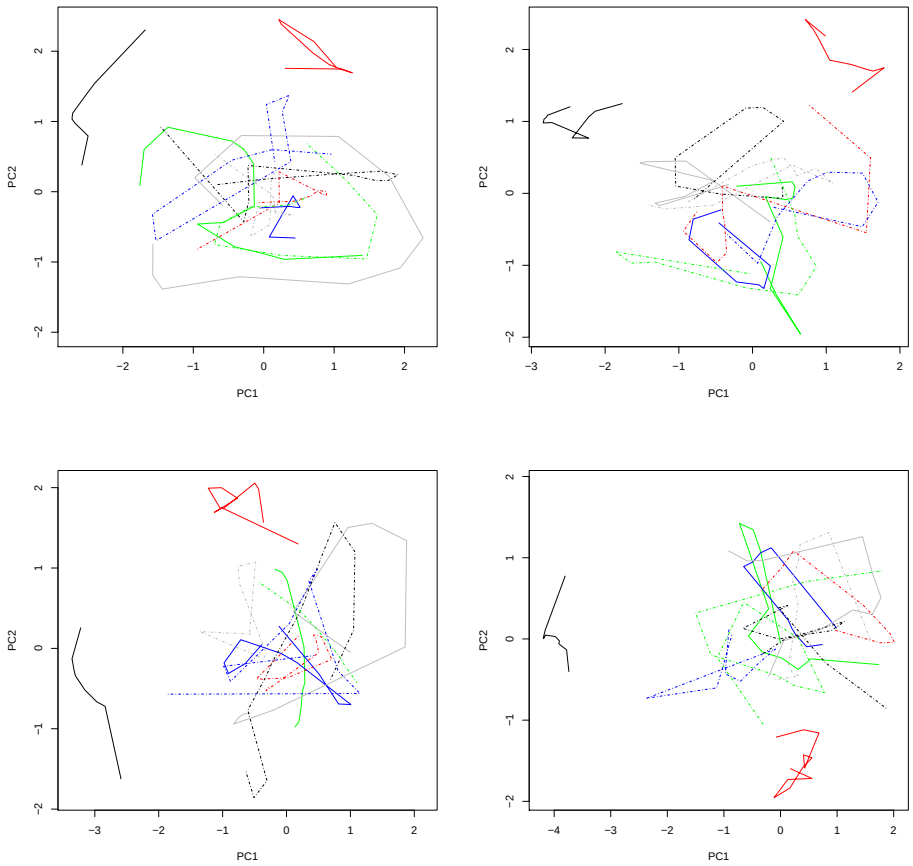


Figure 4.18: Correlation path conditioned on a linear statistics. The data sets are the labeled data sets in Figure 4.16.

Remark We use local estimate to estimate the conditional correlations. In application, some conditional correlations are not estimable. This is because there do not have enough points within the grid. When creating the correlation path conditioned on linear statistics, we connect points according to their order on the domain, and the unestimable correlations are ignored. When creating graphs against the domain, we can use linear extrapolation given there were not many unestimable points.

4.5 Discussion

In this chapter, we propose correlation paths that can be used to visualize the behavior of a sequence of p by p objects. This means we can further extend it to time-varying data. It enables us to figure out the similarity and dissimilarity in the correlation structure. Once a distinct path is found, we can examine the correlations to explore informations. For the correlation paths conditioned on a linear statistic, we conditioned on each principal component. If we have certain directions that we are interested, we use that direction. While conditioned on multivariate statistics, the projection-based approach also works for visualizations.

BIBLIOGRAPHY

BIBLIOGRAPHY

I. Bebu, F. Seillier-Moiseiwitsch, and T. Mathew. Generalized Confidence Intervals for Ratios of Regression Coefficients with Applications to Bioassays. Biometrical Journal, 51(6):1047–1058, 2009.

G. Casella and R. Berger. Statistical Inference. Duxbury Press, 2001.

M. A. Creasy. Limits for the ratio of the means. Journal of the Royal Statistical Society Series B, 16:186–194, 1954.

A. Dempster. Covariance selection. Biometrics, 28(1):157–175, 1972.

P. Diggle and J. Gratton. Monte Carlo methods of inference for implicit statistical models. Journal of the Royal Statistical Society, Series B, 46:193–227, 1984.

S. Dorogovtsev and J. F. F. Mendes. Evolution of networks. Advances in Physics, 51 (4):1079–1187, 2002.

D. Edwards. Introduction to Graphical Modelling. New York: Springer, 2000.

B. Efron. Bootstrap methods: Another look at the jackknife. The Annals of Statistics, 7(1):1–26, 1979.

P. Erdos and A. Renyi. On Random Graphs. I. Publicationes Mathematicae, 6: 290–297, 1959.

J. Faith, R. Mintram, and M. Angelova. Projection pursuit. Bioinformatics, 22(21): 2267–2673, 2006.

E. C. Fieller. Some problems in interval estimation. Journal of the Royal Statistical Society Series B, 16:175–185, 1954.

J. H. Friedman and J. W. Tukey. A Projection Pursuit Algorithm for Exploratory Data Analysis. IEEE Transactions on Computers, C-13:881–890, 1974.

M. Friendly. Corrgrams: Exploratory displays for correlation matrices. The American Statistician, pages 1–18, 2002.

K. Gabriel. The biplot graphic display of matrices with application to principal component analysis. Biometrika, 58(3):453–467, 1932.

A. Gelman, H. Carlin, and D. Rubin. Bayesian data analysis. Chapman and Hall, New York, 2004.

M. Ghosh, M. Yin, and Y. H. Kim. Objective Bayesian inference for ratios of regression coefficients in linear models. Statistica Sinica, 13:409–422, 2003.

M. Ghosh, G. S. Datta, D. Kim, and T. J. Sweeting. Likelihood-based inference for the ratios of regression coefficients in linear models. Annals of the Institute of Statistical Mathematics, 58:457–473, 2006.

E. Gumbel. Statistics of Extremes. Mineloa, NY: Dover, 2004.

A. E. Hoerl. Application of Ridge Analysis to Regression Problems. Chemical Engineering Progress, 58:54–59, 1962.

P. Huber. Projection pursuit. The Annals of Statistics, 13(2):435–475, 1985.

S. Kullback and R. Leibler. On Information and Sufficiency. Annals of Mathematical Statistics, 22(1):79–86, 1951.

G. Oehlert. A Note on the Delta Method. The American Statistician, 46(1):27–29, 1992.

D. Pena and J. Rodriguez. Descriptive measures of multivariate scatter and linear dependence. Journal of Multivariate Analysis, 84(2):361–374, 2003.

D. Rubin. Bayesianly justifiable and relevant frequency calculations for the applies statistician. The Annals of Statistics, 12:11511172, 1984.

G. Seber. Multivariate Observations. New York: Wiley, 1984.

J. Spall. Introduction to Stochastic Search and Optimization: Estimation, Simulation, and Control. Wiley, 2003.

A. Subramanian, P. Tamayo, and .etc. Gene set enrichment analysis: A knowledge-based approach for interpreting genome-wide expression profiles. PNAS, 102(43): 15545–15550, 2005.

M. Sunnaker, A. Busetto, E. Numminen, J. Corander, M. Foll, and C. Dessimoz. Approximate Bayesian Computation. PLoS Comput Biol., 9(1):e1002803, 2013.

S. Tavare, D. Balding, R. Griffiths, and P. Donnelly. Inferring coalescence times from DNA sequence data. Genetics, 145:505–518, 1997.

R. Tibshirani. Regression shrinkage and selection via the LASSO. J. Royal. Statist. Soc. B., 58(1):267–288, 1996.

D. Venzon and S. Moolgavkar. A Method for Computing Profile-Likelihood-Based Confidence Intervals. Journal of the Royal Statistical Society. Series C (Applied Statistics), 37(1):87–94, 1988.

S. Wilks. Certain generalizations in the analysis of variance. Biometrika, 24:471–494, 1932.