Kernel-based association measures

Ying Liu

Submitted in partial fulfillment of the
requirements for the degree
of Doctor of Philosophy
in the Graduate School of Arts and Sciences

COLUMBIA UNIVERSITY

2013

ABSTRACT

Kernel-based association measures

Ying Liu

Measures of associations have been widely used for describing the statistical relationships between two sets of variables. Traditional association measures tend to focus on specialized settings (specific types of variables or association patterns). Based on an in-depth summary of existing measures, we propose a general framework for association measures unifying existing methods and novel extensions based on kernels, including practical solutions to computational challenges. The proposed framework provides improved feature selection and extensions to a variety of current classifiers. Specifically, we introduce association screening and variable selection via maximizing kernel-based association measures. We also develop a backward dropping procedure for feature selection when there are a large number of candidate variables. We evaluate our framework using a wide variety of both simulated and real data. In particular, we conduct independence tests and feature selection using kernel association measures on diversified association patterns of different dimensions and variable types. The results show the superiority of our methods to existing ones. We also apply our framework to four real-word problems, three from statistical genetics and one of gender prediction from handwriting. We demonstrate through these applications both the *de novo* construction of new kernels and the adaptation of existing kernels tailored to the data at hand, and how kernel-based measures of associations can be naturally applied to different data structures including functional input and output spaces. This shows that our framework can be applied to a wide range of real world problems and work well in practice.

Table of Contents

List of Figures

List of Tables

Acknowledgments

I would like to express the deepest appreciation to my advisor, Professor Tian Zheng, who continually and convincingly supported and encouraged me in regard to both my research and life. This dissertation would not have been possible without her guidance and persistent help.

I would like to thank my committee members, Professor David Madigan, Professor Victor de la Pena, Professor Pei Wang and Professor Rongling Wu, for their insightful comments and suggestions.

I would also like to thank Professor Victor de la Pena for bringing to our attention the important papers by Szekely et al. and Sejdinovic et al. and his helpful suggestions during my study, Professor Andrew Gelman for his help on the Spin project, Professor Shaw-Hwa Lo for his help on statistical genetics, and Professor Rongling Wu for providing the poplar data set.

In addition, a thank you to all my friends in the Department of Statistics at Columbia University, especially members of the statistical genetics group, who gave me so much encouragement and joy. I also have had the opportunity to learn from many professors in the department.

Last but most importantly, I would like to express my deepest gratitude to my parents for their everlasting love, tolerance and encouragement.

To my parents

Chapter 1

Introduction

Measures of associations are used in a broad range of applications including genetics, communication, economics, physics, etc. For example, in biology, we strive to identify associations between variants in a person's genome and the risk of a certain disease. The existence of such associations between two sets of variables may suggest the influence of one set of variables on the other, which is of great importance in practice. In communication, call center managers are interested in whether having an agent or an interactive voice response (IVR) answer incoming calls (among other factors) will have an influence on callers' patience.

Many of today's real-world applications involve big data, where the number of variables involved is commonly very large. For instance, in genetics, the number of variables (variants in genome sequences) can easily reach several millions. This puts more challenges on traditional statistical methods for the detection of associations.

A lot of statistical learning problems are deeply rooted in searching for association or nonrandom patterns among variables. Especially, *supervised learning* is the learning task of inferring a relationship between two sets of variables from an input space and an output space, respectively. For example, the output space of a *classification* problem contains labeled data. The dimensionality of the input space can be very high nowadays as aforementioned, thus a key step in tackling such learning problems is (supervised) *variable selection*, which seeks to identify the relevant variables (or features). This is usually done based on some criteria that measure the strength of the association between the two sets of variables.

In the statistics literature, a number of measures of associations (or independence) have been used to evaluate and test associations between two sets of objects. The most commonly used measure is the Pearson correlation coefficient between a univariate X and a univariate Y, which measures *correlation* (or linear association) between individual continuous-valued variables (see Section 2.1.1). However, association patterns of interest in real-world applications are very likely to be much more complicated. Therefore an ideal association measure should have the flexibility to account for such complex patterns.

In this chapter we provide the formal definition of associations in statistics, and overview very briefly several existing measures commonly used in practice.

1.1 Definition of association

In statistics, the definition of association is derived from the definition of independence. Under a hypothesis test setting, the problem becomes conducting a test of independence, with the hypotheses

$$\begin{aligned} H_0 : f_{X,Y} &= f_X f_Y \\ &\text{vs} \\ H_1 : f_{X,Y} &\neq f_X f_Y \end{aligned} \tag{1.1}$$

where for the continuous case,

$$X \in \mathbb{R}^p \text{ and } Y \in \mathbb{R}^q \tag{1.2}$$

with p and q positive integers. $f_{X,Y}$ denotes the joint probability density function (PDF) of (X, Y), f_X and f_Y denote the PDFs of X and Y, respectively; for the discrete case,

$$X \in \prod_{i=1}^{p} \{x_{i1}, x_{i2}, ..., x_{ic_i}\} \text{ and } Y \in \prod_{j=1}^{q} \{y_{j1}, y_{j2}, ..., y_{jr_j}\} \tag{1.3}$$

$f_{X,Y}$ denotes the joint probability mass function (PMF) of (X, Y), f_X and f_Y denote the PMFs of X and Y, respectively. *Association* is defined as the opposite of independence (H_0 in (1.1)). Measures of associations can be treated as measures of "dependence" that are constructed based on a sample $(\mathbf{X}, \mathbf{Y}) = \{(X_i, Y_i)\}, i = 1, 2, ..., n$, and can therefore be naturally used as test statistics.

Problem (1.1) is difficult in the following senses. Random variables X and Y can be of different types (discrete or continuous) and dimensions. The test problem is totally nonparametric, i.e., it does not rely on what kind of distributions X and Y follow. The alternative hypothesis is composite and includes all potential association patterns (e.g., functional relationships such as linear vs nonlinear, monotonic vs non-monotonic, periodic vs non-periodic, and non-functional relationships such as superposition of two or more functions). Therefore the main difficulty stems from the fact that it is hard to quantify all possible departures from the null hypothesis of independence using a test statistic. Unlike simple testing problems such as a test of a normal mean where one can find a "good" test statistic (which leads to an unbiased uniformly most powerful test in the normal case), a general test statistic with good properties for test of independence is still unknown. The ideal measure of association would be nonparametric and flexible enough to accommodate all potential dependence patterns.

1.2 An overview of this thesis

Different measures of associations have been proposed with different motivations. Each of the measures has its own assumptions, which may give a hint on the senarios where the measure will and will not work well. Commonly-used association measures include (among others) Pearson's correlation coefficient, the chi-squared statistic, rank-based measures, and influential measures such as Partition-Retention (PR)'s I [Zheng *et al.*, 2011; Chernoff *et al.*, 2009]. Especially, two new measures have attracted a lot of interests in recent years, the maximal information coefficient (MIC) [Reshef *et al.*, 2011] and the distance covariance (dCov) [Szekely and Rizzo, 2009; Szekely *et al.*, 2007]. MIC's strength is in capturing local patterns while dCov is supported by an elegant theoretical framework. In this thesis, we propose a general kernel-based extension of traditional distance-based measures that will be flexible enough to capture both global and local association patterns.

In Chapter 2 we will review several existing association measures and connect them (including novel connections established in Propositions 1 and 2). Chapter 3 develops the general framework for kernel-based association measures. Chapter 4 considers practical

issues such as selecting kernel parameters. Chapter 5 illustrates the application of the proposed framework to real-world problems. Chapter 6 concludes.

Chapter 2

Association measures

In this chapter, we start with an overview of some representative measures often adopted in practice in Section 2.1, with a focus on the strengths and weaknesses (assumptions and limitations) of each of the measures. In Section 2.2, we examine connections between some of these existing measures of associations.

2.1 Existing measures of associations

In this section we give a brief review of some popular association measures. We will begin with ones with more assumptions and (thus) limitations, followed by more recent ones which are intended to be more general.

2.1.1 Distribution and variable-type specific measures

As mentioned earlier, one of the major difficulties for test of independence is that there are too many possibilities under the alternative. It would be easier to find a test with a good power when one limits attention to a particular type of variables (either continuous or discrete) from a specific distribution family.

For bivariate normal,

$$\begin{aligned} f(x,y) &= \frac{1}{2\pi\sigma_x\sigma_y\sqrt{1-\rho^2}}\exp(-\frac{1}{2(1-\rho^2)}[\frac{(x-\mu_x)^2}{\sigma_x^2}+\frac{(y-\mu_y)^2}{\sigma_y^2}-\frac{2\rho(x-\mu_x)(y-\mu_y)}{\sigma_x\sigma_y}]) \\ f(x) &= \frac{1}{\sigma_x\sqrt{2\pi}}e^{-\frac{1}{2}(\frac{x-\mu_x}{\sigma_x})^2}, f(y)=\frac{1}{\sigma_y\sqrt{2\pi}}e^{-\frac{1}{2}(\frac{y-\mu_y}{\sigma_y})^2} \end{aligned}$$

where ρ is the *correlation* between X and Y

$$\rho = \frac{\text{cov}(X,Y)}{\sigma_x \sigma_y} = \frac{E[(X-\mu_x)(Y-\mu_y)]}{\sigma_x \sigma_y}$$

In this case, the test in (1.1) is equivalent to

$$H_0 : \rho = 0 \text{ vs } H_1 : \rho \neq 0$$

For a data set with n pairs of observations, $(X_i, Y_i), i = 1, ..., n$, sample Pearson's correlation coefficient

$$r = \frac{\sum_{i=1}^{n}(X_i - \bar{X})(Y_i - \bar{Y})}{\sqrt{\sum_{i=1}^{n}(X_i - \bar{X})^2}\sqrt{\sum_{i=1}^{n}(Y_i - \bar{Y})^2}}$$

is shown to be the maximum likelihood estimate (MLE) of the population correlation ρ. r ranges from -1 to 1 and measures *linear* associations, thus it is invariant with respect to linear transformations. As a test statistic, r has an asymptotic normal distribution. Fisher's variance-stabilizing transformation

$$z = \frac{1}{2}\log[(1+r)/(1-r)]^1$$

results in a z that approaches normality faster [Fisher, 1915].

In general, ρ characterizes "correlation" instead of dependence, i.e.,

$$X \perp Y \Longrightarrow X \text{ and } Y \text{ are uncorrelated } \iff \rho = 0$$

but

$$\rho = 0 \not\Rightarrow X \perp Y$$

where $\perp$ denotes independence. This suggests that tests based on r may have low power for non-normal data or nonlinear relationship.

For multivariate variables, *canonical correlation analysis* is used to measure the association by finding linear combinations of the X's and the Y's which have maximum correlation with each other. Specifically, canonical correlation seeks vectors a and b such that the random variables $a^T X$ and $b^T Y$ maximize the correlation $r = \text{cor}(a^T X, b^T Y)$. This can be

[1] All log used in this thesis is natural log unless otherwise mentioned.

treated as a generalization of the correlation ρ (or r) to measure the dependence between two vectors.

The multivariate counterpart of covariance (correlation) is the *covariance (correlation) matrix*. Given a p dimensional vector X with finite second moments, the covariance matrix Σ is defined to be the p by p matrix whose (i, j)th entry is the covariance $\text{cov}(X_i, X_j)$. If the covariance matrix has a block structure, the variables in different blocks are independent with each other. The inverse of the covariance matrix is called the *precision matrix*, which characterizes *conditional independence* in the special case of the multivariate normal distribution. However its goal is different from measuring the overall dependence between variables.

The *odds ratio* [CORNFIELD, 1951] describes the strength of association between two *binary* data values. It is defined in terms of the joint distribution of the two random variables, which can be written as

	$X = 0$	$X = 1$
$Y = 0$	p_{00}	p_{01}
$Y = 1$	p_{10}	p_{11}

where p_{00}, p_{01}, p_{10} and p_{11} are "cell probabilities" that sum to one. The *odds* for Y within the two subpopulations indicated by $X = 1$ and $X = 0$ are defined by the ratio of the conditional probabilities given X, i.e., $P(Y = 1|X)/P(Y = 0|X)$. For example, the odds of $Y = 1$ when $X = 0$ is

$$\frac{p_{10}/(p_{00} + p_{10})}{p_{00}/(p_{00} + p_{10})}$$

The odds ratio is just the ratio of the odds, i.e.,

$$\frac{\frac{p_{00}/(p_{00}+p_{10})}{p_{10}/(p_{00}+p_{10})}}{\frac{p_{01}/(p_{01}+p_{11})}{p_{11}/(p_{01}+p_{11})}} = \frac{p_{00}p_{11}}{p_{10}p_{01}} \tag{2.1}$$

It can be easily seen that the odds ratio defined in (2.1) is symmetric about X and Y. Given sampled data, one can estimate the probabilities in the joint distribution first and define analogously the sample odds ratio. The distribution of the sample *log* odds ratio is approximately normal [Agresti, 2002]. It can be shown that the odds ratio *characterizes*

independence, i.e., it equals one if and only if X and Y are independent. However, it is only available for univariate binary variables.

One generalization of the odds ratio is the Cochran-Mantel-Haenszel (CMH) statistics, in cases where data can be arranged in a series of associated 2×2 tables. The corresponding CMH test has increased ability to detect associations (see, for example, [Wallenstein and Wittes, 1993] for details on this test).

For two general discrete random variables, chi-squared tests are commonly used for test of independence. In this case, data can be allocated to a two-way contingency table

	x_1	...	x_c	margin
y_1	$O_{1,1}$	...	$O_{1,c}$	$O_{1,.}$
...	...	...	...	...
y_r	$O_{r,1}$	...	$O_{r,c}$	$O_{r,.}$
margin	$O_{.,1}$	...	$O_{.,c}$	$O_{.,.}$

where $O_{i,j}$ is the number of observations with $X = x_j$ and $Y = y_i$ (here the notations become simpler than those in (1.3) since we are considering the bivariate case). The expected frequency for a cell, under the null hypothesis of independence, is

$$E_{i,j} = \frac{O_{i,.}O_{.,j}}{n}$$

The value of the test statistic is

$$\chi^2 = \sum_{i=1}^{r}\sum_{j=1}^{c}\frac{(O_{i,j} - E_{i,j})^2}{E_{i,j}}$$

Under the null hypothesis, χ^2 follows asymptotically a chi-squared distribution with $(r-1)(c-1)$ degrees of freedom.

Chi-squared tests are in fact approximations of the log-likelihood ratio test, so they may have low power in some circumstances. It is well known that the approximation to the chi-squared distribution breaks down if expected frequencies are too low. In such cases it is found to be more appropriate to use the G-test (recommended by [Sokal and Rohlf, 1981]), a likelihood-ratio based test statistic

$$G = 2\sum_{ij} O_{i,j}\log(O_{i,j}/E_{i,j})$$

where the sum is taken over all non-empty cells. If we write $O_{i,j} = E_{i,j} + \delta_{i,j}$, with $\sum \delta_{i,j} = 0$ so that the total number of observations stays the same, the G-test is then

$$G = 2\sum (E_i, j + \delta_{i,j}) \ln(1 + \frac{\delta_{i,j}}{E_{i,j}})$$

If we Taylor expand this around $\frac{\delta_{i,j}}{E_{i,j}} = 0$ (the point at which $O_{i,j}$ and $E_{i,j}$ agree), we get

$$\begin{aligned} G &= 2\sum (E_{i,j} + \delta_{i,j})(\frac{\delta_{i,j}}{E_{i,j}} - \frac{1}{2}\frac{\delta_{i,j}^2}{E_{i,j}^2} + O(\delta_{i,j}^3)) \\ &= 2\sum (\delta_{i,j} + \frac{1}{2}\frac{\delta_{i,j}^2}{E_{i,j}}) + O(\delta_{i,j}^3) \\ &\approx \sum \frac{(O_{i,j} - E_{i,j})^2}{E_{i,j}} \\ &= \chi^2 \end{aligned}$$

For very small samples an appropriate exact test (such as Fisher's exact test, which assumes fixed marginals) is preferable to either the chi-squared test or the G-test. Note that the reason why one can find the likelihood ratio test in the current situation is that a specific type (bivariate discrete) of variables is under consideration.

For categorical $X \in \{x_1, x_2, ..., x_c\}$ and continuous $Y \in \mathbb{R}$, test of independence in (1.1) is equivalent to one-way analysis of variance (ANOVA). Let $\bar{Y}_1, ..., \bar{Y}_c$ be the sample mean within each value of X calculated on $n_1, ..., n_c$ observations, respectively. One can define

$$\text{SSR} = \sum_{j=1}^{c} n_j (\bar{Y}_j - \bar{Y})^2 \tag{2.2}$$

to measure the association between X and Y. An F test can be carried out to test the dependence of Y on X, under the assumption that Y_i's with the same X_i values are independent and identically distributed (i.i.d.) normal random variables.

2.1.2 Rank-based measures

Pearson's correlation coefficient r is known to be not robust [Wilcox, 2005], so its value can be misleading if outliers are present or the distribution has a departure from Gaussian [DEVLIN *et al.*, 1975; Huber, 2004]. Rank-based measures have been proposed to address such sensitivity problems. The most natural generalization is to calculate r based on rank

values, which is named Spearman's rho. Another rank-based measure, Kendall's tau, is defined as

$$\begin{aligned}\tau_n &= \frac{\#\text{agreements} - \#\text{disagreements}}{\text{total number of pairs}} \\ &= \frac{\sum(\mathbf{1}\{(X_i - X_j)(Y_i - Y_j) > 0\} - \mathbf{1}\{(X_i - X_j)(Y_i - Y_j) < 0\})}{\frac{1}{2}n(n-1)}\end{aligned}$$

where #disagreements can be treated as a distance metric (called the Kendall tau distance) between two lists. The population version τ of Kendall's tau is defined similarly. Specifically, let (X_1, Y_1) and (X_2, Y_2) be independent random vectors with the same distribution as (X, Y). Then

$$\begin{aligned}\tau &= \mathrm{P}[(X_1 - X_2)(Y_1 - Y_2) > 0] - \mathrm{P}[(X_1 - X_2)(Y_1 - Y_2) < 0] \\ &= \rho[\mathrm{sign}(X_1 - X_2), \mathrm{sign}(Y_1 - Y_2)]\end{aligned}$$

$-1 \leq \tau \leq 1$ with $\tau = 0$ under independence. The testing problem then becomes

$$H_0 : \tau = 0 \text{ vs } H_1 : \tau \neq 0$$

Note that in general the above test is *not* equivalent to that in (1.1) since one can find dependent X and Y that also satisfy $\tau = 0$. In other words, τ does not characterize independence.

If X and Y are independent, the sampling distribution of τ_n has an expected value of zero. If the agreement between the two rankings is perfect (i.e., the two rankings are the same) the coefficient has value 1, while if the disagreement between the two rankings is perfect (i.e., one ranking is the reverse of the other) the coefficient has value -1. For hypothesis testing, exact probability can be calculated for small samples for significance evaluation. Normal approximation or permutation is used for larger samples. For bivariate normal distribution,

$$E(\tau_n) = \frac{2}{\pi}\mathrm{arcsin}\rho$$

Since both Spearman's rho and Kendall's tau are based only on the ranks of the data, they are invariant under rank-preserving (e.g., strictly increasing) transformations. For the same reason, they may not have good power for non-monotonic relationships.

As shown above, converting to rank values is one of the many procedures used to transform data that do not meet the assumptions of normality. The *Kruskal-Wallis H statistic* is another measure designed for situations when the normality assumption has been violated. It conducts a standard ANOVA on the rank-transformed data [Kruskal and Wallis, 1952; Conover and Iman, 1976].

Another group of rank-based measures is developed from the empirical copula. The corresponding test problem is a little different from that in (1.1). Consider $\tilde{X} \in \mathbb{R}^p$, the hypotheses of interest here are

$$\begin{aligned} H_0: &\quad \text{the elements } \{X_1, ..., X_p\} \text{ in } \tilde{X} \text{ are mutually independent} \\ &\qquad\qquad \text{vs} \\ H_1: &\qquad\qquad \text{otherwise} \end{aligned}$$

We consider a special case that is relavent to problem (1.1) with $\tilde{X} \in \mathbb{R}^2$. We shall denote the two elements in $\tilde{X}$ by X and Y following our notation in (1.1). The idea underlying tests based on the empirical copula relies on the fact that X and Y are independent if and only if $C = uv$, where C is the copula defined implicitly by $C(F_X(u), F_Y(v))$ where F_X and F_Y are the cumulative distribution functions (CDFs) of X and Y, respectively. The test problem now becomes

$$H_0 : C = uv \text{ vs } H_1 : C \neq uv$$

which is equivalent to (1.1). A natural thought would be to define some type of norm for the (scaled) difference

$$\mathbb{C}_n(u, v) = \sqrt{n}\{C_n(u, v) - uv\}$$

where C_n is the empirical copula

$$C_n(u, v) = \frac{1}{n} \sum_{i=1}^{n} \mathbf{1}\{\frac{R_i}{n+1} \leq u, \frac{S_i}{n+1} \leq v\}$$

with R_i and S_i ranks of X_i and Y_i, respectively. This leads Deheuvels [Deheuvels, 1979] to a test of independence based on

$$\int_0^1 \int_0^1 \{\mathbb{C}_n(u, v)\}^2 dudv$$

which in turn leads to the Cramér-von Mises test statistic as a function of the ranks through (see [Genest *et al.*, 2007] for details)

$$\begin{aligned} B_n &= \frac{1}{n}\sum_{i=1}^{n}\sum_{j=1}^{n}(1-\frac{R_i \vee R_j}{n})(1-\frac{S_i \vee S_j}{n}) \\ &- 2\sum_{i=1}^{n}\{\frac{n(n-1)-R_i(R_i-1)}{2n^2}\}\{\frac{n(n-1)-S_i(S_i-1)}{2n^2}\} \\ &+ n\{\frac{(n-1)(2n-1)}{6n^2}\}^d \end{aligned} \tag{2.3}$$

Like Kendall's tau, B_n is based only on ranks, thus it is invariant under rank-preserving transformations. However, it is not rotation invariant. Although it has been shown that $C_n(u,v)$ is a consistent estimator of C, the convergence rate is quite slow. Tests based on B_n are asymptotically distribution-free. There are also many parametric copula families available, which usually have parameters that control the strength of dependence. Tests based on different copula families may only have high power if the underlying data structure coincides with the specified family.

It turns out that Kendall's tau can be expressed in terms of copula when X and Y are continuous,

$$\tau = 4\int_0^1\int_0^1 C(u,v)dC(u,v) - 1$$

2.1.3 Influential measures and variance-component scores

If one treats X as the independent variable, and Y as the response variable, association measures can be considered as quantities that measure the influence of X on Y. An example of such measures is Partition-Retention (PR)'s I [Zheng *et al.*, 2011; Chernoff *et al.*, 2009].

PR's I was motivated by one-way ANOVA described in 2.1.1. For simplicity, consider a subset of k binary valued variables from X denoted by $\{X_1, X_2, ..., X_k\}$, and Y either continuous or discrete. The k X variables define a partition Π of the sample into $m = 2^k$ subsets. The resulting partition elements are denoted by $\{A_1, A_2, ..., A_m\}$ corresponding to the possible values of those k binary variables. Let n_j denote the number of observations in A_j. Each nonempty partition element A_j yields a mean of the Y values $\bar{Y}_j$ and the overall

mean is denoted by $\bar{Y} = \sum_{j=1}^{m} n_j \bar{Y}_j / n$. The central *influence* measure is defined as

$$I_{\Pi} = \sum_{j=1}^{m} n_j^2 (\bar{Y}_j - \bar{Y})^2 \tag{2.4}$$

(2.4) is similar to SSR defined in (2.2), but with different weights for the squared differences. In practice, PR's I is shown to be more robust.

For two-way tables as in the case of chi-squared tests, PR's I can be written as a particular form called the genotype-trait distortion (GTD) score. For simplicity, we consider the single-nucleotide polymorphism (SNP) data type where $X \in \{0, 1, 2\}$ and $Y \in \{0, 1\}$. The backward genotype-trait association (BGTA) method was proposed for this type of data with the key statistic, the GTD score [Zheng *et al.*, 2006b]

$$\text{GTD} = \sum_{i=1}^{3} (\frac{O_{1,i}}{O_{1,.}} - \frac{O_{2,i}}{O_{2,.}})^2 \tag{2.5}$$

following the same notations in Section 2.1.1. It is easy to show that [Zheng *et al.*, 2011; Chernoff *et al.*, 2009]

$$I_{\Pi} = \frac{O_{1,.}^2 O_{2,.}^2}{(O_{1,.} + O_{2,.})^2} \text{GTD} \tag{2.6}$$

For binary data (two-by-two tables), it is also easy to show that

$$I_{\Pi} = 2\chi^2 \frac{O_{1,.} O_{2,.} O_{.,1} O_{.,2}}{n^2} \tag{2.7}$$

where $O_{1,.} = \sum_{i=1}^{2} O_{1,i}, O_{2,.} = \sum_{i=1}^{2} O_{2,i}, O_{.,1} = \sum_{i=1}^{2} O_{i,1}$ and $O_{.,2} = \sum_{i=1}^{2} O_{i,2}$ are the row and column sums.

The Sequencing Kernel Association Test (SKAT) [Wu *et al.*, 2011] is motivated by the same applications. It conducts rare-variant association testing for sequencing data (see Chapter 5 for such applications). There the variance-component score statistic is defined as

$$Q = (\mathbf{Y} - \hat{\mu})' \mathbf{K} (\mathbf{Y} - \hat{\mu})$$

where $\mathbf{K} = \mathbf{X}\mathbf{W}\mathbf{X}'$, $\hat{\mu}$ is the predicted mean of $\mathbf{Y}$ based on some covariates, $\mathbf{X}$ is an $n \times p$ matrix with the (i, j)-th element being the j-th value of the i-th X, and $\mathbf{W}$ =diag$(w_1, ..., w_p)$ contains the weights of the p dimensions.

2.1.4 Brownian distance covariance and maximal information coefficient

In recent years more general association measures have been proposed in order to capture complex relationships for various data types. In this section, we discuss two recent general measures that motivated this thesis, Brownian distance covariance (dCov) [Szekely and Rizzo, 2009; Szekely *et al.*, 2007] and maximal information coefficient (MIC) [Reshef *et al.*, 2011].

The motivation behind Brownian distance covariance is morally similar to that of copula based measures described in 2.1.2. Instead of working with the empirical copula process, consider hypotheses (1.1) in terms of characteristic functions. The joint characteristic function of (X, Y) is defined as

$$\phi_{X,Y}(t,s) = E\exp\{i\langle t, X\rangle + i\langle s, Y\rangle\} \tag{2.8}$$

The marginal characteristic functions of X and Y are

$$\phi_X(t) = E\exp\{i\langle t, X\rangle\}, \phi_Y(s) = E\exp\{i\langle s, Y\rangle\} \tag{2.9}$$

respectively. In terms of characteristic functions, X and Y are independent if and only if $\phi_{X,Y} = \phi_X\phi_Y$. Thus the testing problem in (1.1) is equivalent to

$$H_0 : \phi_{X,Y} = \phi_X\phi_Y \text{ vs } H_1 : \phi_{X,Y} \neq \phi_X\phi_Y$$

As in the case of copula based tests, a natural thought would be to measure the difference between $\phi_{X,Y}$ and $\phi_X\phi_Y$ with a suitable distance $||\phi_{X,Y} - \phi_X\phi_Y||$. This distance turns out to be defined through the $||\cdot||_w$-norm in the weighted L_2 space of functions on $\mathbb{R}^{p+q}$

$$\begin{aligned} \mathcal{V}^2(X,Y;w) &= ||\phi_{X,Y}(t,s) - \phi_X(t)\phi_Y(s)||_w^2 \\ &= \int_{\mathbb{R}^{p+q}} |\phi_{X,Y}(t,s) - \phi_X(t)\phi_Y(s)|^2 w(t,s)dtds \end{aligned} \tag{2.10}$$

where $w(t,s)$ is an arbitrary positive weight function for which the above integral exists, and for complex-valued function $f(\cdot)$, $|f|^2 = f\bar{f}$ with $\bar{f}$ the complex conjugate of f. The next step would be to choose the weight function such that the resulting measure have some desirable properties, namely,

(i) One can also define

$$\mathcal{V}^2(X;w) = \mathcal{V}^2(X,X;w)$$

and similarly $\mathcal{V}^2(Y;w)$, and then

$$\mathcal{R} = \begin{cases} \frac{\mathcal{V}(X,Y;w)}{\sqrt{\mathcal{V}(X;w)\mathcal{V}(Y;w)}}, & \mathcal{V}(X;w)\mathcal{V}(Y;w) > 0 \\ 0, & \mathcal{V}(X;w)\mathcal{V}(Y;w) = 0 \end{cases}$$

(ii) $\mathcal{R}$ characterizes independence in the sense that $\mathcal{R} = 0$ if and only if independence holds.

(iii) $\mathcal{R}$ is scale invariant.

The above considerations lead to the definition of the *distance covariance* (dCov) between X and Y with finite first moments as the nonnegative square root of (see [Szekely and Rizzo, 2009] for details)

$$\begin{aligned} \mathcal{V}^2(X,Y;w) &= ||\phi_{X,Y}(t,s) - \phi_X(t)\phi_Y(s)||^2 \\ &= \frac{1}{c_p c_q}\int_{\mathbb{R}^{p+q}} \frac{|\phi_{X,Y}(t,s) - \phi_X(t)\phi_Y(s)|^2}{|t|_p^{1+p}|s|_q^{1+q}} dtds \end{aligned} \tag{2.11}$$

where

$$c_d = \frac{\pi^{(1+d)/2}}{\Gamma((1+d)/2)} \tag{2.12}$$

and $|x|_p$ denotes the Euclidean norm of x in $\mathbb{R}^p$.

One can also define naturally the sample version of the Brownian distance covariance by

$$\mathcal{V}_n^2(\mathbf{X},\mathbf{Y}) = ||\phi_{X,Y}^n(t,s) - \phi_X^n(t)\phi_Y^n(s)||^2$$

where $\phi_{X,Y}^n(t,s)$, $\phi_X^n(t)$, and $\phi_Y^n(s)$ are the joint and marginal empirical characteristic functions, respectively. More specifically, these functions are defined by replacing the expectations in (2.8) and (2.9) by sample averages.

A useful alternative expression of $\mathcal{V}_n^2(\mathbf{X},\mathbf{Y})$ is

$$\mathcal{V}_n^2(\mathbf{X},\mathbf{Y}) = T_1 + T_2 - 2T_3 \tag{2.13}$$

where

$$\begin{aligned} T_1 &= \frac{1}{n^2}\sum_{k,l=1}^{n}|X_k - X_l|_p|Y_k - Y_l|_q, \\ T_2 &= \frac{1}{n^2}\sum_{k,l=1}^{n}|X_k - X_l|_p\frac{1}{n^2}\sum_{k,l=1}^{n}|Y_k - Y_l|_q, \\ T_3 &= \frac{1}{n^3}\sum_{k=1}^{n}\sum_{l,m=1}^{n}|X_k - X_l|_p|Y_k - Y_m|_q \end{aligned}$$

with $|x|_p$ the Euclidean norm of x in $\mathbb{R}^p$. It is not hard to find some analogy in the expressions of $\mathcal{V}_n^2$ in (2.13) and B_n in (2.3). This is not suprising considering the similar motivations of the two methods. In fact, we will show later (see Proposition 2 in 2.2.4) that B_n can be considered as a special case of $\mathcal{V}_n^2$ under a specific "distance" rather than the Euclidean distance in (2.13). One can define the sample $\mathcal{R}_n$ analogously.

Properties such as almost sure convergence $\mathcal{V}_n \to \mathcal{V}$ and $\mathcal{R}_n^2 \to \mathcal{R}^2$, weak convergence and the limit distribution of $n\mathcal{V}_n^2$, and statistical consistency have been proved (see [Szekely and Rizzo, 2009; Szekely *et al.*, 2007] for details). A deterministic relationship between $\mathcal{R}^2(X,Y)$ and correlation ρ has been established for the bivariate normal case. In particular, $\mathcal{R}(X,Y) \leq |\rho|$. This implies that in the normal case where ρ is optimal, tests based on dCov will lose some power, which is as expected. It is also worth noting that the restriction to the $||\cdot||_w$-norm with the specific choice of w and the consequent Euclidean norm in the expression of $\mathcal{V}_n^2$ imply potential limitations of the corresponding tests. It is not clear under what data structures the tests may have good power (which is important information in real applications), the reason being that the measure is not data-driven but rather defined for population first. Our simulation studies showed that the test did fail under some circumstances (see Chapter 4). This is actually the motivation of our current study with the aim of generalizing current measures to a broader kernel-based family and adapting kernels in a data-driven manner.

Now we discuss the maximal information coefficient (MIC), another method based on the idea of partitions. Recall that the reason why one can find an (approximate) likelihood-ratio test with good properties (chi-squared and G-test) in the discrete case is that everything is multinomial (i.e., a specific distribution family). The task becomes more challenging when it comes to continuous random variables. A natural way to tackle this problem is

discretizing, as was done by MIC [Reshef *et al.*, 2011]. For $X \in \mathbb{R}$ and $Y \in \mathbb{R}$, a grid G is drawn on the scatterplot of the two variables. Let I_G denote the mutual information of the probability distribution induced on the cells of G, where the probability of a cell is proportional to the number of observations falling inside that cell. More specifically, the mutual information is defined as

$$MI = H(\text{row}) + H(\text{col}) - H(\text{row}, \text{col})$$

where the entropy of a discrete random variable X is defined as

$$H(X) = -\sum_x p(x) \ln p(x)$$

Mutual information characterizes independence in the following sense: $MI = 0$ if and only if the two random variables are independent. Define the characteristic matrix $M = (m_{a,b})$, where $m_{a,b}$ =max$\{I_G\}$/ln min$\{a, b\}$, with the maximum taken over all a-by-b grids G for the pair of integers (a, b). MIC is the maximum of $m_{a,b}$ over ordered pairs (a, b) such that $ab < B$, where $B = n^{0.6}$.

In fact, MIC is equivalent to conducting the G-test (discussed in 2.1.1) to the induced discrete distribution, since it has been shown that

$$G = 2 \cdot n \cdot MI$$

The default choice of the maximal grid size B is a balance of sensitivity and specificity. However, the power 0.6 is chosen somewhat arbitrarily. This remains an open problem and is worth studying in the future.

MIC falls between 0 and 1, is symmetric [i.e., MIC(X,Y)=MIC(Y,X)], and because I_G depends only on the rank order of the data, MIC is invariant under order-preserving transformations of the axes. MIC was intended to capture a wide range of association types, although our experiments revealed some senarios where it did not have good power (Chapter 4). One can consider the shape of the grids as a rectangular "kernel" imposed on the data, which may explain the observed limitations of MIC.

2.1.5 Hilbert-Schmidt independence criterion (HSIC)

Researchers in the machine learning field have also been developing association measures, one of which is the Hilbert-Schmidt independence criterion (HSIC) [Gretton *et al.*, 2007]. Its motivation goes back to the original definition of the test problem (1.1). As in the case of copula-based measures and the distance covariance, a natural thought is to define a distance directly between the joint distribution $f_{X,Y}$ and the product of its marginals $f_X f_Y$. This is done in HSIC with the kernel embeddings of probability measures into reproducing kernel Hilbert spaces (RKHS, on which more in Chapter 3), which is a common approach in machine learning. Specifically, the maximum mean discrepancy (MMD) between two probability measures is just defined by the norm-induced metric in the RKHS. The HSIC is then defined as the maximum mean discrepancy between the joint and the product distributions, with the form

$$\begin{aligned} H &= E\{k(X,X')l(Y,Y')\} + E\{k(X,X')\}E\{l(Y,Y')\} \\ &- 2E\{E\{k(X,X')|X\}E\{l(Y,Y')|Y\}\} \end{aligned}$$

where (X,Y) and (X',Y') are $\overset{i.i.d.}{\sim} P_{XY}$, $k(X,X')$ and $l(Y,Y')$ are *kernel functions*. One can readily see that this involves transforming the Euclidean distances of $\mathcal{V}_n^2$ in Section 2.1.4, equation (2.13), by passing them through a kernel distortion [Gretton *et al.*, 2009]. This is interesting considering the fact that dCov and HSIC are independently discovered in the two separate fields. The similar formulations of these two types of measures imply that a richer family of association measures could potentially be defined, which is pursued in the current study.

2.2 Equivalence between different association measures

In this section we first provide a diagram of association measures, then prove the equivalence between some of the association measures discussed in Section 2.1.

2.2.1 A map of association measures

Figure 2.1 shows a diagram of the most commonly used measures of associations, with relationships between different measures indicated by lines with arrows. It can be seen that most of the measures are only suitable for variables of specific types (as indicated in bold on the borders of Figure 2.1). Different measures are designed in order to capture different types of associations as mentioned earlier. For example, Pearson's correlation, Spearman's rho and Kendall's tau can detect monotone relationships, while MIC is powerful in identifying local patterns. It is also worth noting that the distance covariance has a lot of connections with other measures as special cases. This actually motivates our current work to develop a more general framework for association measures utilizing kernels (as an example, the kernel distance covariance will be defined in Section 3.2).

2.2.2 Equivalence of I_Π and $\mathcal{V}_n(\mathbf{X}, \mathbf{Y})$

In this section we show the equivalence of $\mathcal{V}_n(\mathbf{X}, \mathbf{Y})$ in (2.13) and the influence measure I_Π in (2.4). This is stated in the following Proposition.

Proposition 1. *For two-way tables with binary X and Y,*

$$I_\Pi = \frac{n^2}{2}\mathcal{V}_n^2(\mathbf{X}, \mathbf{Y}) \tag{2.14}$$

For m-by-two tables with $m > 2$, use m indicator variables to code X,

$$I_\Pi = \frac{n^2}{2\sqrt{2}}\mathcal{V}_n^2(\mathbf{X}, \mathbf{Y}) \tag{2.15}$$

Proof. We shall use $n_{ij}, n_{i.}$ and $n_{.j}$ instead of $O_{i,j}, O_{i,.}$ and $O_{.,j}$ as in Section 2.1.

(i) Binary case.

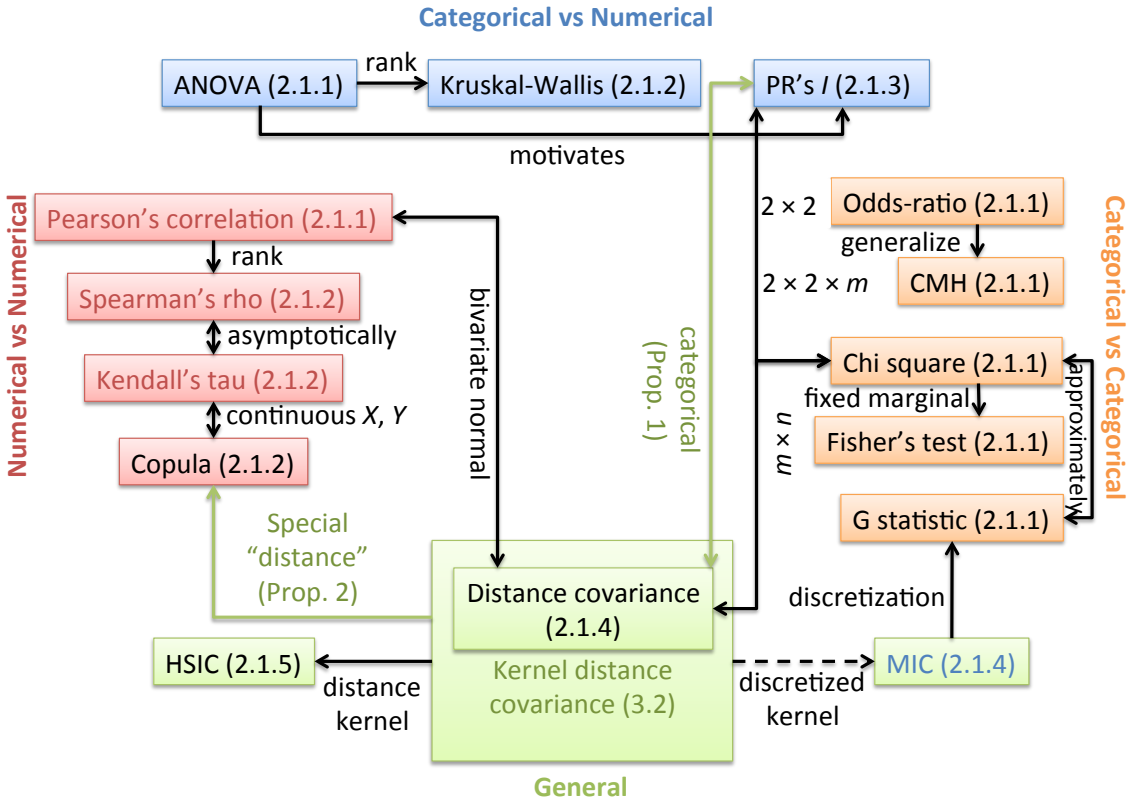


Figure 2.1: A map of association measures. Our contribution is highlighted in green. Measures in red detect monotone associations, while measure in blue (MIC) is powerful in capturing local patterns. The broken line implies heuristic relation. The corresponding parts in the main texts (indicated by numbers in the parentheses) provide more details on each measure and relationships between different measures.

For two-way tables with binary X and Y,

$$\begin{aligned}
T_1 &= \frac{2}{n^2}(n_{00}n_{11} + n_{01}n_{10}) \\
&= \frac{2}{n^4}(n_{00} + n_{01} + n_{10} + n_{11})^2(n_{00}n_{11} + n_{01}n_{10}) \\
&= \frac{2}{n^4}(n_{00}^2 + n_{01}^2 + n_{10}^2 + n_{11}^2 + 2n_{00}n_{01} + 2n_{00}n_{10} + 2n_{00}n_{11} \\
&+ 2n_{01}n_{10} + 2n_{01}n_{11} + 2n_{10}n_{11})(n_{00}n_{11} + n_{01}n_{10}) \\
&= \frac{2}{n^4}(n_{00}^3n_{11} + n_{00}^2n_{01}n_{10} + n_{01}^2n_{00}n_{11} + n_{01}^3n_{10} + n_{00}n_{11}n_{10}^2 \\
&+ n_{01}n_{10}^3 + n_{00}n_{11}^3 + n_{01}n_{10}n_{11}^2 + 2n_{00}^2n_{01}n_{11} + 2n_{00}n_{01}^2n_{10} \\
&+ 2n_{00}n_{10}^2n_{01} + 2n_{00}^2n_{10}n_{11} + 2n_{00}^2n_{11}^2 + 2n_{00}n_{01}n_{10}n_{11} + 2n_{01}n_{10}n_{00}n_{11} \\
&+ 2n_{01}^2n_{10}^2 + 2n_{01}n_{00}n_{11}^2 + 2n_{01}^2n_{11}n_{10} + 2n_{10}n_{11}^2n_{00} + 2n_{10}^2n_{01}n_{11})
\end{aligned}$$

$$\begin{aligned}
T_2 &= \frac{4}{n^4}(n_{.0}n_{.1})(n_{0.}n_{1.}) \\
&= \frac{4}{n^4}(n_{00} + n_{10})(n_{01} + n_{11})(n_{00} + n_{01})(n_{10} + n_{11}) \\
&= \frac{4}{n^4}(n_{00}n_{01} + n_{00}n_{11} + n_{10}n_{01} + n_{10}n_{11})(n_{00}n_{10} + n_{00}n_{11} + n_{01}n_{10} + n_{01}n_{11}) \\
&= \frac{4}{n^4}(n_{00}^2n_{01}n_{10} + n_{00}^2n_{01}n_{11} + n_{00}n_{01}^2n_{10} + n_{00}n_{01}^2n_{11} + n_{00}^2n_{10}n_{11} + n_{00}^2n_{11}^2 \\
&+ n_{00}n_{01}n_{10}n_{11} + n_{00}n_{01}n_{11}^2 + n_{00}n_{10}^2n_{01} + n_{00}n_{10}n_{01}n_{11} + n_{10}^2n_{01}^2 + n_{10}n_{01}^2n_{11} \\
&+ n_{10}^2n_{00}n_{11} + n_{10}n_{00}n_{11}^2 + n_{10}^2n_{01}n_{11} + n_{10}n_{01}n_{11}^2)
\end{aligned}$$

$$\begin{aligned}
T_3 &= \frac{1}{n^3}[n_{00}(n_{01} + n_{11})(n_{10} + n_{11}) + n_{01}(n_{10} + n_{11})(n_{00} + n_{10}) \\
&+ n_{10}(n_{00} + n_{01})(n_{01} + n_{11}) + n_{11}(n_{00} + n_{01})(n_{00} + n_{10})] \\
&= \frac{1}{n^3}(n_{00}n_{10}n_{01} + n_{00}n_{01}n_{11} + n_{00}n_{11}n_{10} + n_{00}n_{11}^2 + n_{01}n_{10}n_{00} \\
&+ n_{01}n_{10}^2 + n_{01}n_{11}n_{10} + n_{01}n_{11}n_{00} + n_{10}n_{00}n_{01} + n_{10}n_{00}n_{11} \\
&+ n_{10}n_{01}^2 + n_{10}n_{01}n_{11} + n_{11}n_{00}^2 + n_{11}n_{00}n_{10} + n_{11}n_{01}n_{00} + n_{11}n_{01}n_{10}) \\
&= \frac{1}{n^4}(n_{00}n_{10}n_{01} + n_{00}n_{01}n_{11} + n_{00}n_{11}n_{10} + n_{00}n_{11}^2 + n_{01}n_{10}n_{00} \\
&+ n_{01}n_{10}^2 + n_{01}n_{11}n_{10} + n_{01}n_{11}n_{00} + n_{10}n_{00}n_{01} + n_{10}n_{00}n_{11} \\
&+ n_{10}n_{01}^2 + n_{10}n_{01}n_{11} + n_{11}n_{00}^2 + n_{11}n_{00}n_{10} + n_{11}n_{01}n_{00} + n_{11}n_{01}n_{10}) \\
&\quad (n_{00} + n_{01} + n_{10} + n_{11})
\end{aligned}$$

$$\begin{aligned}
\mathcal{V}_n^2(\mathbf{X},\mathbf{Y}) &= T_1 + T_2 - 2T_3 \\
&= \frac{4n_{01}^2 n_{10}^2 - 8n_{00}n_{01}n_{10}n_{11} + 4n_{00}^2 n_{11}^2}{n^4} \\
&= \frac{4(n_{00}n_{11} - n_{01}n_{10})^2}{n^4}
\end{aligned}$$

$$\begin{aligned}
I_\Pi &= \frac{n_{.0}^2 n_{.1}^2}{n^2}[(\frac{n_{00}}{n_{.0}} - \frac{n_{01}}{n_{.1}})^2 + (\frac{n_{10}}{n_{.0}} - \frac{n_{11}}{n_{.1}})^2] \\
&= \frac{n_{.0}^2 n_{.1}^2}{n^2}[\frac{(n_{00}n_{.1} - n_{01}n_{.0})^2}{(n_{.0}n_{.1})^2} + \frac{(n_{10}n_{.1} - n_{11}n_{.0})^2}{(n_{.0}n_{.1})^2}] \\
&= 2\frac{n_{.0}^2 n_{.1}^2}{n^2}\frac{(n_{00}n_{11} - n_{01}n_{10})^2}{(n_{.0}n_{.1})^2} \\
&= \frac{n^2}{2}\mathcal{V}_n^2(\mathbf{X},\mathbf{Y})
\end{aligned}$$

Thus $\mathcal{V}_n^2(\mathbf{X},\mathbf{Y})$ and I_Π are equivalent conditional on the sample size n.

(ii) m-by-two tables with $m > 2$.

In this case, use m indicator variables to code X. We first rewrite T_1, T_2 and T_3 in terms of partitions based on X.

$$\begin{aligned}
T_1 &= \frac{\sqrt{2}}{n^2}\sum_{j=1}^{m}(n_{0j}(n_{1.} - n_{1j}) + n_{1j}(n_{0.} - n_{0j})) \\
T_2 &= \frac{2\sqrt{2}n_{0.}n_{1.}}{n^4}\sum_{j=1}^{m} n_{.j}(n - n_{.j}) \\
T_3 &= \frac{\sqrt{2}}{n^3}\sum_{j=1}^{m}(n_{0j}(n - n_{0.})(n - n_{.j}) + n_{1j}(n - n_{1.})(n - n_{.j}))
\end{aligned}$$

Calculations then give

$$\begin{aligned}
n^4\mathcal{V}_n^2(\mathbf{X},\mathbf{Y}) &= n^4(T_1+T_2-2T_3)\\
&= \sqrt{2}\sum_{j=1}^{m}(-n_{0.}^2n_{0j}n_{1.}-2n_{0.}n_{0j}n_{1.}^2-n_{0j}n_{1.}^3-n_{0.}^3n_{1j}-2n_{0.}^2n_{0j}n_{1j}-2n_{0.}^2n_{1.}n_{1j}\\
&- 4n_{0.}n_{0j}n_{1.}n_{1j}-n_{0.}n_{1.}^2n_{1j}-2n_{0j}n_{1.}^2n_{1j}+2n_{0.}^2n_{1.}n_{.j}+2n_{0.}n_{0j}n_{1.}n_{.j}+2n_{0.}n_{1.}^2n_{.j}\\
&+ 2n_{0j}n_{1.}^2n_{.j}+2n_{0.}^2n_{1j}n_{.j}+2n_{0.}n_{1.}n_{1j}n_{.j}-2n_{0.}n_{1.}n_{.j}^2)\\
&= \sqrt{2}(-n_{0.}^3n_{1.}-2n_{0.}^2n_{1.}^2-n_{0.}n_{1.}^3-n_{0.}^3n_{1.}-2n_{0.}^2\sum n_{0j}n_{1j}-2n_{0.}^2n_{1.}^2\\
&- 4n_{0.}n_{1.}\sum n_{0j}n_{1j}-n_{0.}n_{1.}^3-2n_{1.}^2\sum n_{0j}n_{1j}+2n_{0.}^2n_{1.}n+2n_{0.}n_{1.}\sum n_{0j}n_{.j}\\
&+ 2n_{0.}n_{1.}^2n+2n_{1.}^2\sum n_{0j}n_{.j}+2n_{0.}^2\sum n_{1j}n_{.j}+2n_{0.}n_{1.}\sum n_{1j}n_{.j}-2n_{0.}n_{1.}\sum n_{.j}^2)\\
&= \sqrt{2}(-2n_{0.}^3n_{1.}-4n_{0.}^2n_{1.}^2-2n_{0.}n_{1.}^3-\sum n_{0j}n_{1j}(2n_{0.}^2+4n_{0.}n_{1.}+2n_{1.}^2)+2n_{0.}^2n_{1.}n\\
&+ 2n_{0.}n_{1.}^2n+(2n_{0.}n_{1.}+2n_{1.}^2)\sum n_{0j}n_{.j}+(2n_{0.}^2+2n_{0.}n_{1.})\sum n_{1j}n_{.j}-2n_{0.}n_{1.}\sum n_{.j}^2)\\
&= \sqrt{2}(-2n_{0.}n_{1.}(n_{0.}+n_{1.})^2-2\sum n_{0j}n_{1j}(n_{0.}+n_{1.})^2+2n_{0.}n_{1.}n^2\\
&+ (2n_{0.}n_{1.}+2n_{1.}^2)\sum n_{0j}n_{.j}+(2n_{0.}^2+2n_{0.}n_{1.})\sum n_{1j}n_{.j}-2n_{0.}n_{1.}\sum n_{.j}^2)\\
&= \sqrt{2}(-2n^2\sum n_{0j}n_{1j}+2n_{1.}n\sum n_{0j}n_{.j}+2n_{0.}n\sum n_{1j}n_{.j}-2n_{0.}n_{1.}\sum n_{.j}^2)\\
&= \sqrt{2}(-2n^2\sum n_{0j}n_{1j}+2nn_{1.}\sum n_{0j}(n_{0j}+n_{1j})+2n_{0.}n\sum n_{1j}(n_{0j}+n_{1j})\\
&- 2n_{0.}n_{1.}\sum(n_{0j}+n_{1j})^2)\\
&= \sqrt{2}(-2n^2\sum n_{0j}n_{1j}+2nn_{1.}\sum(n_{0j}^2+n_{0j}n_{1j})+2nn_{0.}\sum(n_{0j}n_{1j}+n_{1j}^2)\\
&- 2n_{0.}n_{1.}\sum(n_{0j}^2+n_{1j}^2+2n_{0j}n_{1j}))\\
&= \sqrt{2}(-2n^2\sum n_{0j}n_{1j}+2nn_{1.}\sum n_{0j}^2+2nn_{1.}\sum n_{0j}n_{1j}+2nn_{0.}\sum n_{0j}n_{1j}\\
&+ 2nn_{0.}\sum n_{1j}^2-2n_{0.}n_{1.}\sum n_{0j}^2-2n_{0.}n_{1.}\sum n_{1j}^2-4n_{0.}n_{1.}\sum n_{0j}n_{1j})\\
&= 2\sqrt{2}(n_{1.}^2\sum n_{0j}^2+n_{0.}^2\sum n_{1j}^2-2n_{0.}n_{.j}\sum n_{0j}n_{1j}) \qquad (2.16)
\end{aligned}$$

by noticing that $n = n_{0.} + n_{1.} = \sum n_{.j}, n_{0.} = \sum n_{0j}, n_{1.} = \sum n_{1j}$, and $n_{.j} = n_{0j} + n_{1j}$. On the other hand,

$$\begin{aligned}
n^2I_\Pi &= \sum_{j=1}^{m}(n_{0j}n_{1.}-n_{1j}n_{0.})^2\\
&= \sum_{j=1}^{m}(n_{0j}^2n_{1.}^2-2n_{0.}n_{1.}n_{0j}n_{1j}+n_{0.}^2n_{1j}^2)\\
&= n_{1.}^2\sum n_{0j}^2-2n_{0.}n_{1.}\sum n_{0j}n_{1j}+n_{0.}^2\sum n_{1j}^2 \qquad (2.17)
\end{aligned}$$

Combining (2.16) and (2.17), we have

$$I_{\Pi} = \frac{n^2}{2\sqrt{2}} \mathcal{V}_n^2(\mathbf{X}, \mathbf{Y})$$

Thus conditional on n, $\mathcal{V}_n(\mathbf{X}, \mathbf{Y})$ and I_{Π} are equivalent.

It is worth mentioning that there is no essential difference between cases (i) and (ii). Specifically, the additional factor of $\frac{1}{\sqrt{2}}$ in (2.15) just comes from the Euclidean distance between different X values coded as m dimensional vectors. □

We have validated the derived equivalence using randomly generated data of different sample sizes n and numbers of possible values m (results not shown).

Remark 1. *The equivalence shown by Proposition 1 is quite interesting, considering the very different motivations of the two classes of association measures. The definition of I_{Π} for categorical Y with more than two categories has not been fully studied. The above results may provide some insight into defining I_{Π} in such cases.*

2.2.3 Equivalent form of $\mathcal{V}_n(\mathbf{X}, \mathbf{Y})$ for continuous Y

(i) The form of $\mathcal{V}_n(\mathbf{X}, \mathbf{Y})$.

For continuous Y,

$$\begin{aligned}
T_1 &= \frac{1}{n^2} \sum_{X_k \neq X_l} |Y_k - Y_l| \\
T_2 &= \frac{2}{n^4} \#(X_k \neq X_l) \sum_{k,l=1}^{n} |Y_k - Y_l| \\
T_3 &= \frac{1}{n^3} \sum_{k} \underset{l}{\#}(X_l \neq X_k) \sum_{m} |Y_k - Y_m| \\
\mathcal{V}_n^2(\mathbf{X}, \mathbf{Y}) &= T_1 + T_2 - 2T_3 \\
&= \frac{1}{n^4} \sum_{k} [n^2 \sum_{X_l \neq X_k} |Y_k - Y_l| + 2(\#(X_a \neq X_b) - n \underset{m}{\#}(X_m \neq X_k)) \\
&\quad \sum_{l} |Y_k - Y_l|]
\end{aligned} \tag{2.18}$$

(ii) $\mathcal{V}_n(\mathbf{X}, \mathbf{Y})$ under squared distance.

Using squared distance for Y and with binary X,

$$\begin{aligned}\mathcal{V}_n^2(\mathbf{X},\mathbf{Y}) &= T_1 + T_2 - 2T_3 \\ &= \frac{1}{n^4}\sum_k [n^2 \sum_{X_l \neq X_k} (Y_k - Y_l)^2 + 2(\#(X_a \neq X_b) - n \#_m(X_m \neq X_k)) \sum_l (Y_k - Y_l)^2]\end{aligned}$$

where

$$\begin{aligned}&\sum_k \sum_{X_l \neq X_k} (Y_k - Y_l)^2 \\ =\ & \sum_{l,k} (Y_k - Y_l)^2 1(X_l = 0)1(X_k = 1) + \sum_{l,k} (Y_k - Y_l)^2 1(X_l = 1)1(X_k = 0)\end{aligned}$$

$$\begin{aligned}&\sum_{l,k} (Y_k - Y_l)^2 1(X_l = 0)1(X_k = 1) \\ =\ & \sum_{l,k} (Y_k - \bar{Y_1} + \bar{Y_1} - \bar{Y_0} + \bar{Y_0} - Y_l)^2 1(X_l = 0)1(X_k = 1) \\ =\ & \sum_{l,k} (Y_k - \bar{Y_1})^2 1(X_l = 0)1(X_k = 1) + \sum_{l,k} (\bar{Y_1} - \bar{Y_0})^2 1(X_l = 0)1(X_k = 1) \\ +\ & \sum_{l,k} (\bar{Y_0} - Y_l)^2 1(X_l = 0)1(X_k = 1) + \sum 2(Y_k - \bar{Y_1})(\bar{Y_1} - \bar{Y_0})1(X_k = 1)1(X_l = 0) \\ +\ & \sum 2(Y_k - \bar{Y_1})(Y_l - \bar{Y_0})1(X_k = 1)1(X_l = 0) \\ +\ & \sum 2(Y_l - \bar{Y_0})(\bar{Y_1} - \bar{Y_0})1(X_k = 1)1(X_l = 0) \\ =\ & n_0 \sum_{X_k=1} (Y_k - \bar{Y_1})^2 + n_0 n_1 (\bar{Y_1} - \bar{Y_0})^2 + n_1 \sum_{X_k=0} (Y_k - \bar{Y_0})^2\end{aligned}$$

By symmetry,

$$\begin{aligned}&\sum_{l,k} (Y_k - Y_l)^2 1(X_l = 1)1(X_k = 0) \\ =\ & n_0 \sum_{X_k=1} (Y_k - \bar{Y_1})^2 + n_0 n_1 (\bar{Y_1} - \bar{Y_0})^2 + n_1 \sum_{X_k=0} (Y_k - \bar{Y_0})^2\end{aligned}$$

$$\#(X_a \neq X_b) = n_0 n_1$$

$$\#_m(X_m \neq X_k) = n_0 1(X_k = 1) + n_1 1(X_k = 0)$$

Thus,

$$\#(X_a \neq X_b) - n \#_m(X_m \neq X_k) = n_0 n_1 - n n_0 1(X_k = 1) - n n_1 1(X_k = 0)$$

$$\begin{aligned}
& \sum_{l,k} 2(\#(X_a \neq X_b) - n \underset{m}{\#}(X_m \neq X_k))(Y_k - Y_l)^2 \\
= & \sum_{k} 2[n_0 n_1 - n n_0 1(X_k = 1) - n n_1 1(X_k = 0)] \\
\times & [\sum_{l}(Y_k - Y_l)^2 1(X_l = 0) + \sum_{l}(Y_k - Y_l)^2 1(X_l = 1)]
\end{aligned}$$

$$\begin{aligned}
& \sum_{X_k = X_l = 0} (Y_k - Y_l)^2 \\
= & \sum_{X_k = X_l = 0} (Y_k - \bar{Y_0} + \bar{Y_0} - Y_l)^2 \\
= & n_0 \sum_{X_k = 0} (Y_k - \bar{Y_0})^2 + n_0 \sum_{X_l = 0} (Y_l - \bar{Y_0})^2 \\
= & 2 n_0 \sum_{X_k = 0} (Y_k - \bar{Y_0})^2
\end{aligned}$$

Similarly,

$$\begin{aligned}
& \sum_{X_k = X_l = 1} (Y_k - Y_l)^2 \\
= & 2 n_1 \sum_{X_k = 1} (Y_k - \bar{Y_1})^2
\end{aligned}$$

Thus,

$$\begin{aligned}
& \sum_{k} [n^2 \sum_{X_l \neq X_k} (Y_k - Y_l)^2 + 2(\#(X_a \neq X_b) - n \underset{m}{\#}(X_m \neq X_k)) \sum_{l} (Y_k - Y_l)^2] \\
= & (2n^2 n_0 - 2n n_0^2 - 6n n_0 n_1) \sum_{X_k = 1} (Y_k - \bar{Y_1})^2 \\
+ & (2n^2 n_1 - 2n n_1^2 - 6n n_0 n_1) \sum_{X_k = 0} (Y_k - \bar{Y_0})^2 \\
+ & (2n^2 - 2n n_0 - 2n n_1) n_0 n_1 (\bar{Y_1} - \bar{Y_0})^2 + 2 n_0 n_1 \sum_{l,k} (Y_k - Y_l)^2 \\
= & -4 n n_0 n_1 \sum_{X_k = 1} (Y_k - \bar{Y_1})^2 - 4 n n_0 n_1 \sum_{X_k = 0} (Y_k - \bar{Y_0})^2 + 2 n_0 n_1 \sum_{l,k} (Y_k - Y_l)^2
\end{aligned}$$

where

$$\begin{aligned}
&\sum_{l,k}(Y_k - Y_l)^2 \\
= &\sum_{X_l=0,X_k=0}(Y_k - Y_l)^2 + \sum_{X_l=0,X_k=1}(Y_k - Y_l)^2 \\
+ &\sum_{X_l=1,X_k=0}(Y_k - Y_l)^2 + \sum_{X_l=1,X_k=1}(Y_k - Y_l)^2 \\
= &\ 2n_0\sum_{X_k=1}(Y_k - \bar{Y_1})^2 + 2n_0n_1(\bar{Y_1} - \bar{Y_0})^2 \\
+ &\ 2n_1\sum_{X_k=0}(Y_k - \bar{Y_0})^2 + 2n_0\sum_{X_k=0}(Y_k - \bar{Y_0})^2 + 2n_1\sum_{X_k=1}(Y_k - \bar{Y_1})^2
\end{aligned}$$

Thus,

$$\begin{aligned}
&\sum_k[n^2\sum_{X_l\neq X_k}(Y_k - Y_l)^2 + 2(\#(X_a \neq X_b) - n\,\underset{m}{\#}(X_m \neq X_k))\sum_l(Y_k - Y_l)^2] \\
= &\ 4n_0^2n_1^2(\bar{Y_0} - \bar{Y_1})^2
\end{aligned}$$

$$\mathcal{V}_n^2(\mathbf{X},\mathbf{Y}) = \frac{4}{n^4}n_0^2n_1^2(\bar{Y_0} - \bar{Y_1})^2 \tag{2.19}$$

2.2.4 Equivalence of $\mathcal{V}_n(\mathbf{X},\mathbf{Y})$ and B_n

In this section we establish the equivalence of $\mathcal{V}_n(\mathbf{X},\mathbf{Y})$ and B_n.

Proposition 2. *Define a "distance" between X_k and X_l as* [2]

$$1 - \frac{R_k \vee R_l}{n} \tag{2.20}$$

then,

$$B_n = n\mathcal{V}_n^2(\mathbf{X},\mathbf{Y}) \tag{2.21}$$

Proof. Using the "distance" defined by (2.20),

$$T_1 = \frac{1}{n^2}\sum_{k,l=1}^{n}(1 - \frac{R_k \vee R_l}{n})(1 - \frac{S_k \vee S_l}{n})$$

[2]Strictly speaking, (2.20) is not a distance. For example, it is not necessarily 0 when $X_k = X_l$. Here we use the term "distance" only to highlight the analogy between the two measures.

$$T_2 = \frac{1}{n^2}\sum_{k,l=1}^{n}(1-\frac{R_k \vee R_l}{n})\frac{1}{n^2}\sum_{k,l=1}^{n}(1-\frac{S_k \vee S_l}{n})$$

where (by reordering the X's)

$$\begin{aligned}
&\sum_{k,l=1}^{n}(1-\frac{R_k \vee R_l}{n}) \\
= &\sum_{k=1}^{n}[\sum_{l=k+1}^{n}(1-\frac{R_k \vee R_l}{n})+\sum_{l=1}^{k}(1-\frac{R_k \vee R_l}{n})] \\
= &\sum_{k=1}^{n}[\sum_{l=k+1}^{n}(1-\frac{l}{n})+\sum_{l=1}^{k}(1-\frac{k}{n})] \\
= &\frac{1}{3}n^2-\frac{n}{2}+\frac{1}{6}
\end{aligned}$$

$\sum_{k,l=1}^{n}(1-\frac{S_k \vee S_l}{n})$ has the same value since here only rank matters. Thus,

$$\begin{aligned}
T_2 &= \frac{1}{n^4}(\frac{1}{3}n^2-\frac{n}{2}+\frac{1}{6})^2 \\
&= [\frac{(2n-1)(n-1)}{6n^2}]^2
\end{aligned}$$

$$\begin{aligned}
T_3 &= \frac{1}{n^3}\sum_{k=1}^{n}\sum_{l,m=1}^{n}(1-\frac{R_k \vee R_l}{n})(1-\frac{S_k \vee S_m}{n}) \\
&= \frac{1}{n^3}\sum_{k=1}^{n}[\sum_{l=1}^{n}(1-\frac{R_k \vee R_l}{n})][\sum_{m=1}^{n}(1-\frac{S_k \vee S_m}{n})]
\end{aligned}$$

where (by reordering the X's)

$$\begin{aligned}
&\sum_{l=1}^{n}(1-\frac{R_k \vee R_l}{n}) \\
= &\sum_{l=1}^{k}(1-\frac{R_k \vee R_l}{n})+\sum_{l=k+1}^{n}(1-\frac{R_k \vee R_l}{n}) \\
= &\sum_{l=1}^{k}(1-\frac{R_k}{n})+\sum_{l=k+1}^{n}(1-\frac{l}{n}) \\
= &n-\frac{R_k^2}{n}-\frac{(R_k+1+n)(n-R_k)}{2n} \\
= &\frac{n^2-n-R_k^2+R_k}{2n}
\end{aligned}$$

Similar calculations can be done for $\sum_{m=1}^{n}(1-\frac{S_k \vee S_m}{n})$ by reordering the Y's. Thus,

$$\begin{aligned} T_3 &= \frac{1}{n^3}\sum_{k=1}^{n}[\frac{n(n-1)-R_k(R_k-1)}{2n}][\frac{n(n-1)-S_k(S_k-1)}{2n}] \\ &= \frac{1}{n}\sum_{k=1}^{n}[\frac{n(n-1)-R_k(R_k-1)}{2n^2}][\frac{n(n-1)-S_k(S_k-1)}{2n^2}] \end{aligned}$$

Therefore,

$$B_n = n\mathcal{V}_n^2(\mathbf{X},\mathbf{Y}) \tag{2.22}$$

□

Chapter 3

A general framework for kernel-based association measures

In this chapter we develop a general framework to incorporate kernels into the traditional measures of associations. Kernel machines have been widely used to take into account complex structures contained in data in the machine learning field. Here we provide a unified framework linking kernels from machine learning and association measures in the statistics literature. The hope is that taking advantage of kernels allows us to detect richer association patterns, which is actually demonstrated by the comprehensive empirical studies shown later.

3.1 Kernel-based association measures

Recall that in Propositions 1 and 2 in Section 2.2 we established the equivalence between association measures with very different motivations. Especially, Proposition 2 shows that by choosing a particular form of "distance" instead of the Euclidean distance, the distance covariance (2.13) and the copula-based measure (2.3) can be treated as special cases of a family of more general association measures. Besides distance covariance, most commonly-used association measures can be decomposed to a common set of elements, specifically, inner products and Euclidean distances within the same set of variables, and their cross product terms. This allows general association measures to be defined by replacing the Eu-

clidean distances in traditional measures by kernel distances, where the "distance" function need not satisfy the triangle inequality.

3.1.1 Reproducing kernel Hilbert spaces (RKHS) and kernel distances

Here we introduce definitions and notations required to understand kernel-based association measures. One can refer to [Bertinet and Agnan, 2004; Phillips and Venkatasubramanian, 2011; Sejdinovic *et al.*, 2012] for a more comprehensive treatment on this topic.

Definition 1. *(RKHS) Let $\mathcal{H}$ be a Hilbert space of real-valued functions defined on $\mathcal{Z}$. A function $k : \mathcal{Z} \times \mathcal{Z} \to \mathbb{R}$ is called a reproducing kernel of $\mathcal{H}$ if*

- $\forall z \in \mathcal{Z}, k(, z) \in \mathcal{H}$
- $\forall z \in \mathcal{Z}, \forall f \in \mathcal{H}, \langle f, k(, z)\rangle_{\mathcal{H}} = f(z)$

If $\mathcal{H}$ has a reproducing kernel, it is called a reproducing kernel Hilbert space (RKHS).

The map $z \mapsto k(., z)$ is called the canonical feature map of k. Moore-Aronszajn theorem (e.g., see [Bertinet and Agnan, 2004]) states that, every symmetric, positive definite function is a reproducing kernel associated with some RKHS. This allows us to take advantage of RKHS without considering explicitly the canonical map.

There are several kernel functions successfully used in the literature, such as

The linear kernel:	$z^T z'$
The polynomial kernel:	$(\gamma z^T z' + \gamma_0)^p$
The radial basis function (RBF) kernel :	$\exp(-\sum_i \sigma_i (z_i - z'_i)^2)$ or $\exp(-\sum_i \sigma_i \lvert z_i - z'_i\rvert)$
	obtain Gaussian and Laplace kernels respectively
The sigmoid kernel:	$tanh(\gamma z^T z' + \gamma_0)$

Definition 2. *(Kernel distance) Consider two points z,z' and a reproducing kernel k. The kernel distance between z and z' is defined as*

$$\rho_k(z, z') = k(z, z) + k(z', z') - 2k(z, z') \tag{3.1}$$

We shall sometimes omit the subscript k if it is self-clarified in the context. Definition 2 ensures that given a kernel, one can define a distance by (3.1).

Definition 3. *(Negative type) The distance ρ is said to have negative type if $\forall n \geq 2$, $z_1, ..., z_n \in \mathcal{Z}$, and $\alpha_1, ..., \alpha_n \in \mathbb{R}$ with $\sum_{i=1}^{n} \alpha_i = 0$,*

$$\sum_{i=1}^{n} \sum_{j=1}^{n} \alpha_i \alpha_j \rho(z_i, z_j) \leq 0 \tag{3.2}$$

It has been shown that negative type is essential for the isometric embedding of the original space into RKHS [Sejdinovic *et al.*, 2012]. Also, (3.1) in Definition 2 defines a valid distance ρ of negative type on $\mathcal{Z}$.

Lemma 1. *[Sejdinovic* et al.*, 2012] Let $\mathcal{Z}$ be a nonempty set, and let ρ be a distance on $\mathcal{Z}$. Let $z_0 \in \mathcal{Z}$, and denote $k(z, z') = \rho(z, z_0) + \rho(z', z_0) - \rho(z, z')$. Then k is positive definite if and only if ρ satisfies (3.2).*

Thus given a distance of negative type, there is an *induced* kernel (called the distance kernel). In addition, k is a distance kernel if and only if $k(z_0, z_0) = 0$ for some $z_0 \in \mathcal{Z}$. Hence, it is clear that any strictly positive definite kernel, e.g., the Gaussian kernel, is *not* a distance kernel.

We shall define general association measures based on either kernel distances defined in (3.1) (Chapters 4 and 5), or directly-defined distance functions (Chapter 5). When working in the latter way, checking whether the distance is of negative type makes sure that it is a kernel induced distance.

3.1.2 Kernel-based association measures

Here we provide the general definition of kernel-based association measures.

Definition 4. *(Kernel-based association measures) Given an association measure $\mathcal{A} = \mathcal{A}(d)$, which is a functional of the Euclidean distance denoted by d, a corresponding kernel-based measure can be defined as*

$$\mathcal{A}_k = \mathcal{A}(\rho_k)$$

In other words, a kernel-based association measure is of the same form as a traditional measure, with Euclidean distances replaced by kernel distances. Definition 4 is quite general in the sense that corresponding to any association measure that is a functional of (Euclidean) distance, there is a family of implicit measures based on kernels. Hence it actually provides a general framework to define more flexible association measures. The rationale is that by mapping the current space into an RKHS of a higher dimension, the association patterns in that RKHS would be simpler enough to be captured by the original measure.

Example 1. *(Kernel influential measure) The kernel influential measure is defined as*

$$I_{\Pi}^{\rho} = n^{-1} \sum_{j=1}^{m} n_j^2 \rho(\bar{Y}_j, \bar{Y})$$

following the notations in (2.4), where ρ is a (kernel) distance. In particular, taking the Mahalanobis distance,

$$I_{\Pi}^{M} = n^{-1} \sum_{j=1}^{m} n_j^2 (\bar{Y}_j - \bar{Y})^T S^{-1} (\bar{Y}_j - \bar{Y})$$

where S is the sample covariance matrix.

Example 2. *(Kernel distance covariance) The population kernel distance covariance is defined as the nonnegative square root of*

$$\begin{aligned} \mathcal{V}^2_{\rho_x,\rho_y}(\mathbf{X},\mathbf{Y}) &= E_{XY} E_{X'Y'} \rho_x(X,X') \rho_y(Y,Y') \\ &+ E_X E_{X'} \rho_x(X,X') E_Y E_{Y'} \rho_y(Y,Y') \\ &- 2 E_{X'Y'} [E_X \rho_x(X,X') E_Y \rho_y(Y,Y')] \end{aligned}$$

where (X,Y) and (X',Y') are $\overset{i.i.d.}{\sim} P_{XY}$, and ρ_x and ρ_y are kernel distances defined on the spaces of X and Y, respectively. Analogously, the sample kernel distance covariance is defined by

$$\begin{aligned} \mathcal{V}^2_{n,\rho_x,\rho_y}(\mathbf{X},\mathbf{Y}) &= \frac{1}{n^2} \sum_{k,l=1}^{n} \rho_x(X_k,X_l) \rho_y(Y_k,Y_l), \\ &+ \frac{1}{n^2} \sum_{k,l=1}^{n} \rho_x(X_k,X_l) \frac{1}{n^2} \sum_{k,l=1}^{n} \rho_y(Y_k,Y_l), \\ &- \frac{2}{n^3} \sum_{k=1}^{n} \sum_{l,m=1}^{n} \rho_x(X_k,X_l) \rho_y(Y_k,Y_m) \end{aligned}$$

3.2 Kernel distance covariance

In Chapters 4 and 5 we will study the kernel distance covariance numerically by intensive simulations and real data applications. In this section we discuss connections to some previous work in order to provide insights into the kernel distance covariance.

[Kosorok, 2009] extends the original Brownian distance covariance [Szekely and Rizzo, 2009; Szekely *et al.*, 2007] to arbitrary normed spaces. There the author does not consider RKHS embedding but rather explicitly the norm induced distances. [Lyons, 2013] studies the distance covariance in the context of embeddings to general Hilbert spaces, and the relation with the theory of RKHS is not exploited. [Sejdinovic *et al.*, 2012] establishes the equivalence between the population versions of Hilbert-Schmidt Independence Criterion and the kernel distance covariance when *distance kernels* (see Lemma 1) are used. The current work is more general in the sense that no restrictions are forced on the kernels. In addition, the aforementioned work focuses on developing theoretical properties of the measures under certain conditions, while the current work focuses more on practical issues inevitably encountered when applying such measures to real applications, such as developing criteria for kernel and parameter selection, and procedures for feature selection with kernel-based association measures (Chapter 4).

Chapter 4

Kernel and parameter selection

It can be imagined that a key factor in the framework presented in Chapter 3 is the choice of the kernel. There are several kernel functions successfully used in the literature, such as the linear kernel, the polynomial kernel, and the radial basis function (RBF) kernel (with the Gaussian kernel as a special case). In this chapter we present systematic procedures for the selection of kernels and their parameters for the proposed kernel-based association measures.

4.1 Selection of kernels and their parameters

Let $\mathcal{A}(k_{x,\theta_x}, k_{y,\theta_y})$ denote the association measure at hand, where k_{x,θ_x} and k_{y,θ_y} are kernels for X and Y, respectively, with θ_x and θ_y corresponding kernel parameters. Then for kernels within the same category, multiple parameters could be tuned according to

$$\theta^* = \arg\max_{\theta} \mathcal{A}(k_{x,\theta_x}, k_{y,\theta_y})$$

A nice feature of the above optimization problem is that most kernels are differentiable w.r.t. θ, so one can compute the gradient easily. This may guide kernel and feature selection for kernel machines such as support vector machines (SVMs, see Section 4.3).

For kernels of different categories, one can choose appropriate kernels according to

$$\{k_x, k_y\} = \arg\max_{m_x, m_y} \mathcal{A}(k_{x,m_x}, k_{y,m_y})$$

where m_x and m_y are indices for candidate kernels for X and Y, respectively. Normalization may be needed for a fair comparison between different kernels.

More generally, methods have been proposed to combine multiple kernel functions instead of selecting a specific one for kernel machines recently (see, for example, [Gönen and Alpaydin, 2011]). Similar ideas can be adopted here for multiple kernel learning in the association testing framework. Specifically, if there are P kernels for X under consideration, the combined kernel is then (omitting x in the subscript)

$$k_\eta(X_i, X_j) = f_\eta(\{k_m(X_i, X_j)\}_{m=1}^P)$$

where the combining function $f_\eta : \mathbb{R}^P \to \mathbb{R}$, can be a linear or a nonlinear function, and η is its parameter. η can be optimized with a set of predefined kernels (i.e., we know the kernel functions and the corresponding kernel parameters before training), such that the association measure is maximized. Or it can be integrated into the kernel functions and optimized during training.

Remark 2. *The idea behind maximizing the association measure is that we want to find a suitable transformation of the original data by kernels that reflect the underlying structure. The maximum value can be treated as an indication of the "strength" of the association pattern. A maximum of 0 indicates independence.*

4.2 Parameter selection for kernel distance covariance with RBF kernels

In this section we give an example of parameter selection for the kernel distance covariance with RBF kernels in association testing. Recall that the kernel distance covariance is defined by its square

$$\mathcal{V}^2_{n,\rho_x,\rho_y}(\mathbf{X}, \mathbf{Y}) = T_1 + T_2 - 2T_3$$

where

$$\begin{aligned} T_1 &= \frac{1}{n^2}\sum_{k,l=1}^{n}\rho_x(X_k,X_l)\rho_y(Y_k,Y_l), \\ T_2 &= \frac{1}{n^2}\sum_{k,l=1}^{n}\rho_x(X_k,X_l)\frac{1}{n^2}\sum_{k,l=1}^{n}\rho_y(Y_k,Y_l), \\ T_3 &= \frac{1}{n^3}\sum_{k=1}^{n}\sum_{l,m=1}^{n}\rho_x(X_k,X_l)\rho_y(Y_k,Y_m), \end{aligned}$$

with ρ_x and ρ_y kernel distances in the X and Y spaces, respectively. Specifically (omitting the subscript),

$$\rho(x,z) = k(x,x) + k(z,z) - 2k(x,z)$$

Here we focus on the RBF kernels with different scaling factors, defined as

$$k_\sigma(x,z) = \exp(-\sum_i \sigma_i(x_i - z_i)^2),$$

where $\sigma_i \geq 0$ are the scaling factors, and i is the index for different dimensions. The induced distance is then

$$\rho_\sigma(x,z) = 2(1 - \exp(-\sum_i \sigma_i(x_i - z_i)^2)).$$

Traditional strategies for choosing σ_i's are usually based on exhaustive search. In our framework, one can compute the gradient directly to maximize the association measure.

4.2.1 Optimizing the RBF kernel parameters

We use Newton's method to find the maximum of the RBF kernel distance covariance. Specifically (omitting the subscript x or y to σ_i),

$$\delta\sigma_i = -(\Delta_\sigma \mathcal{V}^2_{n,\rho_x,\rho_y})^{-1}\frac{\partial \mathcal{V}^2_{n,\rho_x,\rho_y}}{\partial \sigma_i}$$

where the Laplacian operator Δ is defined by

$$(\Delta_\sigma \mathcal{V}^2_{n,\rho_x,\rho_y})_{i,j} = \frac{\partial^2 \mathcal{V}^2_{n,\rho_x,\rho_y}}{\partial\sigma_i\partial\sigma_j}$$

In this study we use the parameterization on the logarithmic scale, i.e., $\log\sigma_i$, to avoid positivity constraints in the optimization problem. This also turns out to give more stable

Figure 4.1: Patterns for the two angles used in the simulations (adapted from Wikipedia).

results. The gradient and the second order derivatives of the RBF kernel distance is then

$$\frac{\partial \rho_\sigma}{\partial \log\sigma_k} = 2(1+\exp(\sum_i -\sigma_i(x_i-z_i)^2)(x_k-z_k)^2\sigma_k) \quad (4.1)$$

$$\begin{aligned}\frac{\partial^2 \rho_\sigma}{\partial (\log\sigma_k)^2} &= 2(1+\exp(\sum_i -\sigma_i(x_i-z_i)^2)(x_k-z_k)^2\sigma_k) \\ &- \sigma_k^2(x_k-z_k)^4\exp(\sum_i -\sigma_i(x_i-z_i)^2)) \quad (4.2)\end{aligned}$$

$$\frac{\partial^2 \rho_\sigma}{\partial (\log\sigma_k)(\log\sigma_l)} = 2(1-\exp(\sum_i -\sigma_i(x_i-z_i)^2)(x_k-z_k)^2\sigma_k(x_l-z_l)^2\sigma_l) \quad (4.3)$$

(4.1), (4.2) and (4.3) can be directly used when computing the gradient and the Hessian matrix of $\mathcal{V}^2_{n,\rho_x,\rho_y}$.

4.2.2 Numerical results

In this section we compare the performance of the RBF kernel distance covariance with optimized scaling factors and three state-of-the-art association measures introduced in Chapter 1, namely, the Brownian distance covariance, the maximal information coefficient, and the copula-based Cramér-von Mises test statistic. We generated the two angles (ϕ, θ) in the spherical coordinate system following seven unusual association patterns, mimicking those at the wikipedia.org page on Pearson correlation (Figure 4.1, R code is modified from supplementary files of [Newton, 2009]). Then we converted to Cartesian coordinates by

$$\begin{aligned}x &= \cos\phi\sin\theta \\ y &= \sin\phi\sin\theta \\ z &= \cos\theta\end{aligned}$$

We first conduct association tests with the aforementioned four measures as test statistics on each of the constructed pairwise patterns. In each case we randomly sample $n = 50$

and 300 points. Permutation tests are used to get P-values for tests based on the RBF kernel distance covariance. Specifically, we permute the order of the generated data in one dimension so that the association pattern is destroyed. Two hundred permutations are done, thus the minimum possible P-value is 0.005. For both the original and each of the permuted data, Newton's method as described in Section 4.2.1 is used to find the scaling factor in the RBF kernel for each dimension, such that the type I error is well controlled. We set the initial values $\log\sigma_i = 3$. Each component is normalized by its standard deviation. Two hundred independent replicates are generated for each of the association patterns. The constructed association patterns and the resulted parallel P-value box plots from different tests are shown in Figure 4.2 and Figure 4.3. Given the fact that there is dependence between horizontal and vertical components in all of the cases, smaller P-values indicate higher power. It can be seen that the power of the considered tests tend to be higher when the sample size is bigger. The three competing tests exhibit low power in several cases especially when the sample size is small. In all the cases, P-values from the Monte Carlo test of independence based on the kernel distance covariance with parameters tuned by the proposed strategy are the smallest among the four tests.

One can roughly divide the generated association patterns into two groups: functional (e.g., all but the case on the far right in the middle row of Figures 4.2 and 4.3) and non-functional (most of the other patterns shown in Figures 4.2 and 4.3) relationships. It can be seen that existing methods tend to perform better for functional associations. Non-functional associations contain superposition of two or more functions (such as the lower-left pattern shown in Figures 4.2 and 4.3). More complicated non-functional associations include ones without *predictability*, i.e., Y cannot be predicted from X (e.g., all the patterns shown in the first row of Figures 4.2 and 4.3 except the second one). In such patterns Y has a constant mean across the range of X, and only the variance changes. The three competing tests do not have a good power for such patterns, especially when the sample size is small (Figure 4.2). As mentioned earlier, MIC has the strength in capturing local patterns, while the distance covariance and the Cramér-von Mises test statistic can both be treated as specialized versions of the kernel distance covariance with specific "kernels". As expected, the dependence is revealed by the RBF kernel distance covariance with optimized

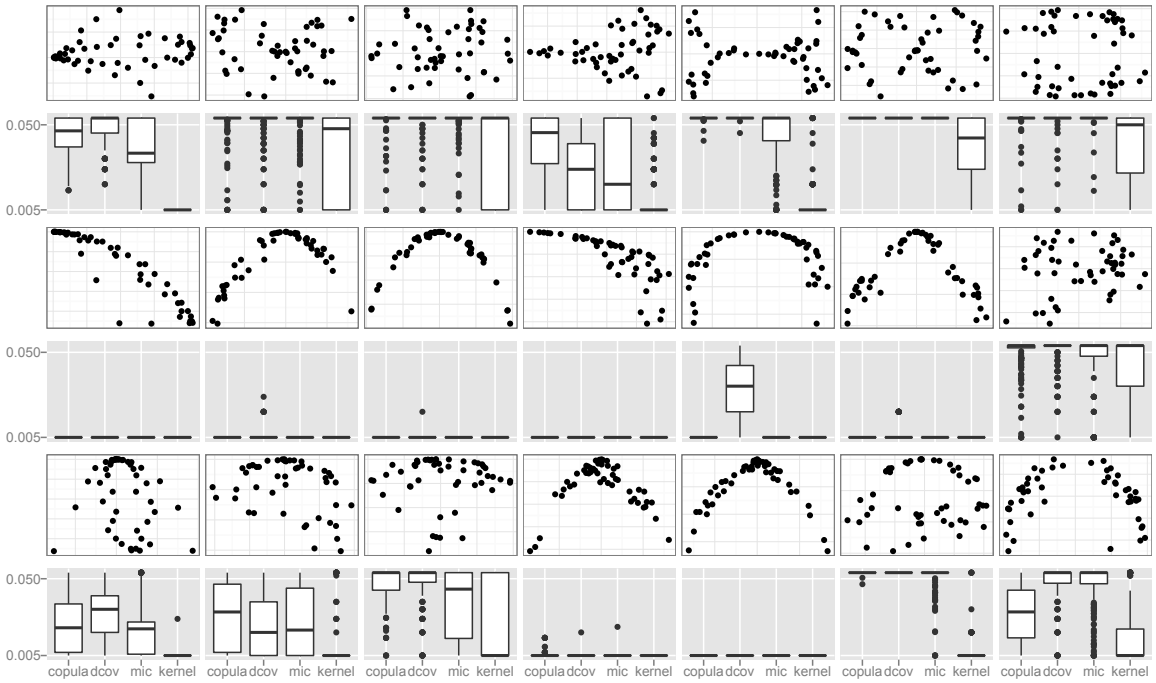


Figure 4.2: Association patterns and corresponding P-value box plots from four different tests based on 50 samples. The scatter plots are based on one random simulation and the box plots of P-values are based on 200 simulations and the P-values are calculated using 200 permutations.

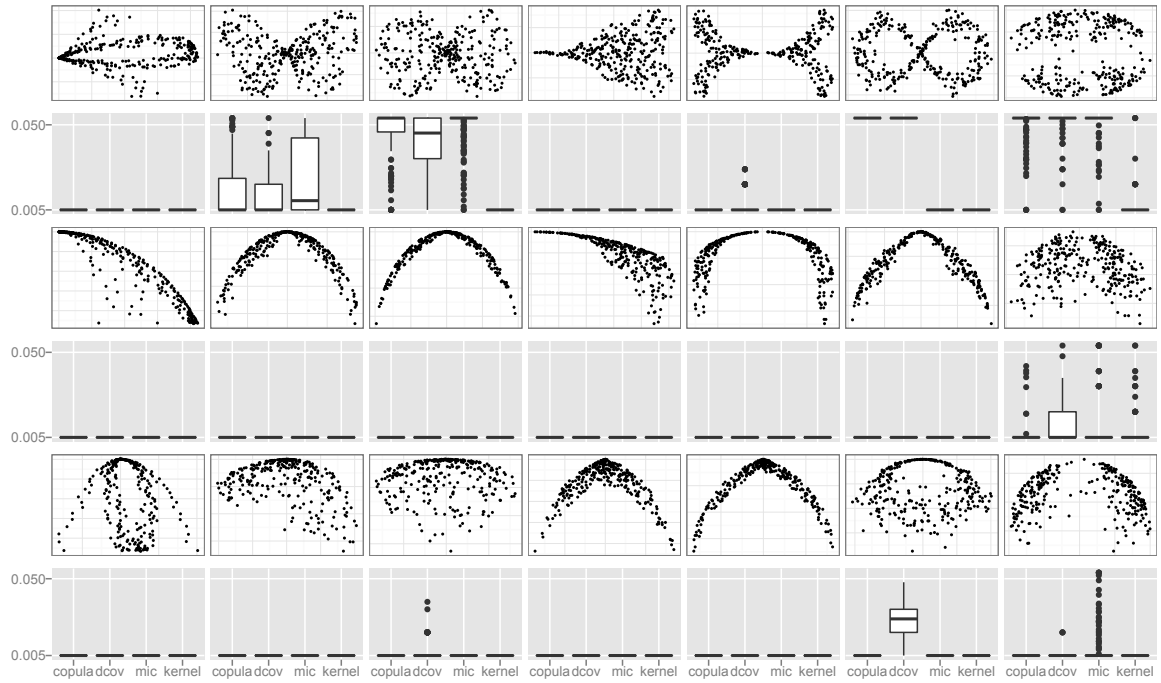


Figure 4.3: Association patterns and corresponding P-value box plots from four different tests based on 300 samples. The scatter plots are based on one random simulation and the box plots of P-values are based on 200 simulations and the P-values are calculated using 200 permutations.

scaling factors in all of the cases.

In practice problems are usually of higher dimensions. However, there has been virtually no systematic empirical evaluation on the performance of association tests for higher (more than two) dimensional data to the author's best knowledge. Using the patterns generated above, we can treat a pair of the variables as a two dimensional vector, and test for the association with the other remaining variable. We then conduct tests based on the RBF kernel distance covariance on the 21 three-dimensional patterns as described above. Of the three methods we evaluated on the pairwise patterns, only Brownian distance covariance can deal with such problems. Thus we compare its results with those from our approach. The P-value box plots in Figure 4.4 and Figure 4.5 again show the impact of the sample size and superiority of the RBF kernel distance covariance relative to the original Brownian distance covariance. This is because the RBF kernel is more flexible, and our approach adapts kernel parameters in a data-driven manner.

It is a little difficult to visualize the corresponding three-dimensional association patterns. Here we show a "slice plot", i.e., scatterplots of two variables conditional on different ranges of the third variable (Figure 4.6), for the sixth association pattern corresponding to the sixth column in Figure 4.3 (all the other six patterns can be found in Appendix B). This is a hard problem as indicated by the results shown in Figures 4.4 and 4.5, especially X and Z versus Y as shown in the middle row in Figure 4.6. It can be seen that X and Z together are associated in a certain way with Y, but the relationships between X and Z within different ranges of Y are quite similar. Even in such cases the kernel distance covariance can detect the association with a high power.

4.3 Parameter and feature selection with kernel-based association measures in classification

A lot of classifiers are built upon a metric of the feature space, such as SVM [Vapnik, 1998; Vapnik, 2000], regression-based methods, and K nearest neighbors (KNN). Especially, feature selection for these methods can be treated as training 0/1 weights on the candidate features, which is widely known to be critical for the performance of the classifiers. There-

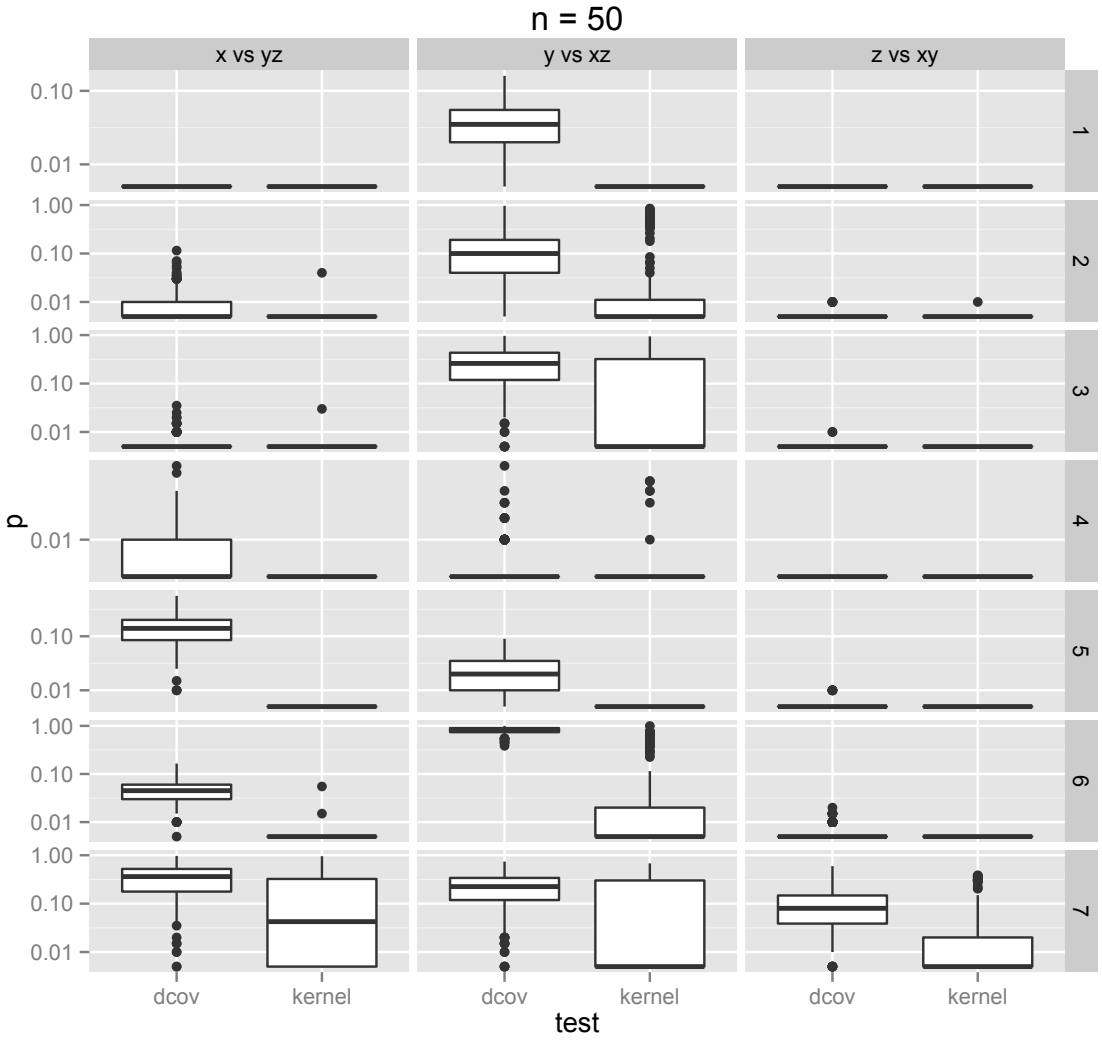


Figure 4.4: P-value box plots from two competing tests for each of the three-dimensional association patterns based on 50 samples. The box plots are based on 200 simulations and the P-values are calculated using 200 permutations.

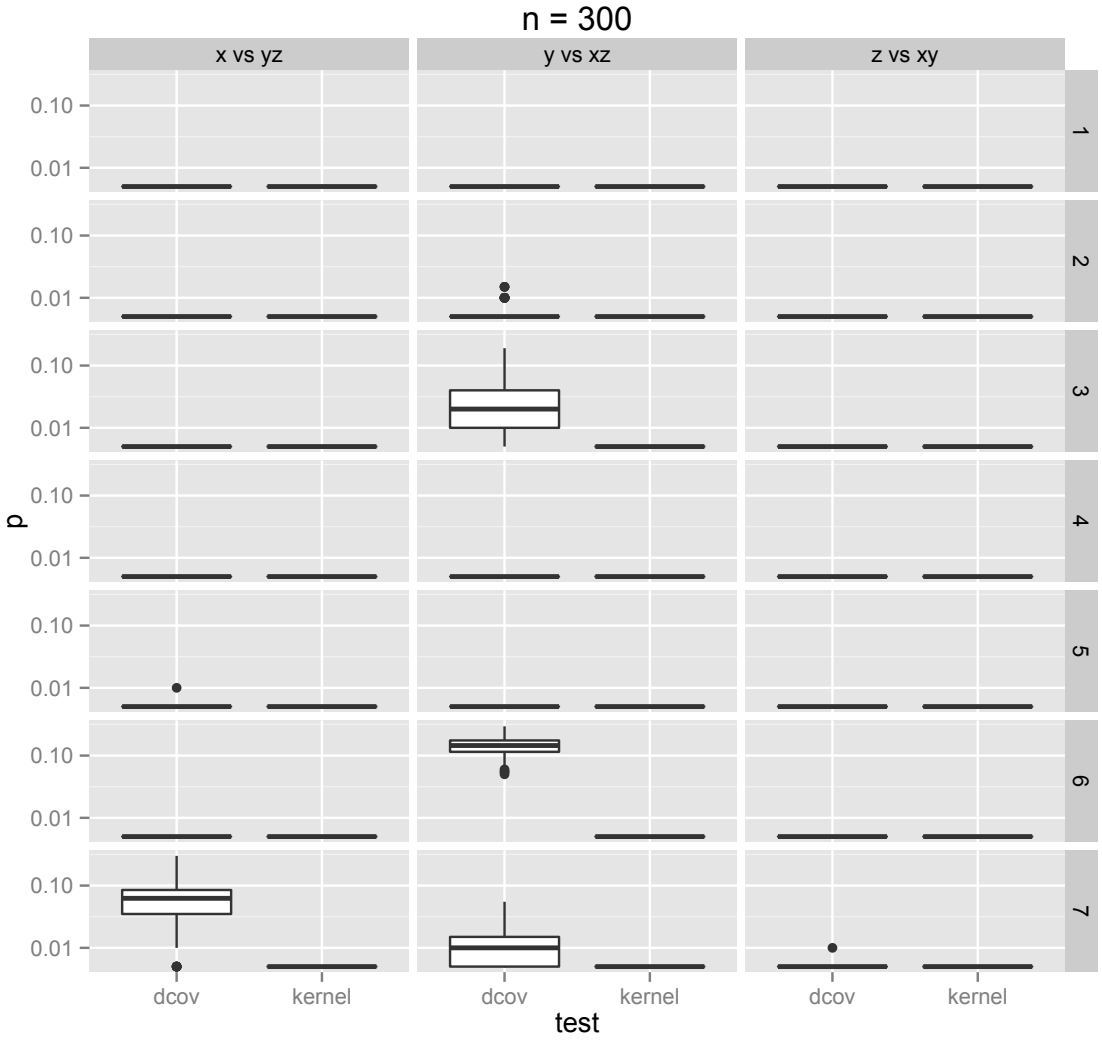


Figure 4.5: P-value box plots from two competing tests for each of the three-dimensional association patterns based on 300 samples. The box plots are based on 200 simulations and the P-values are calculated using 200 permutations.

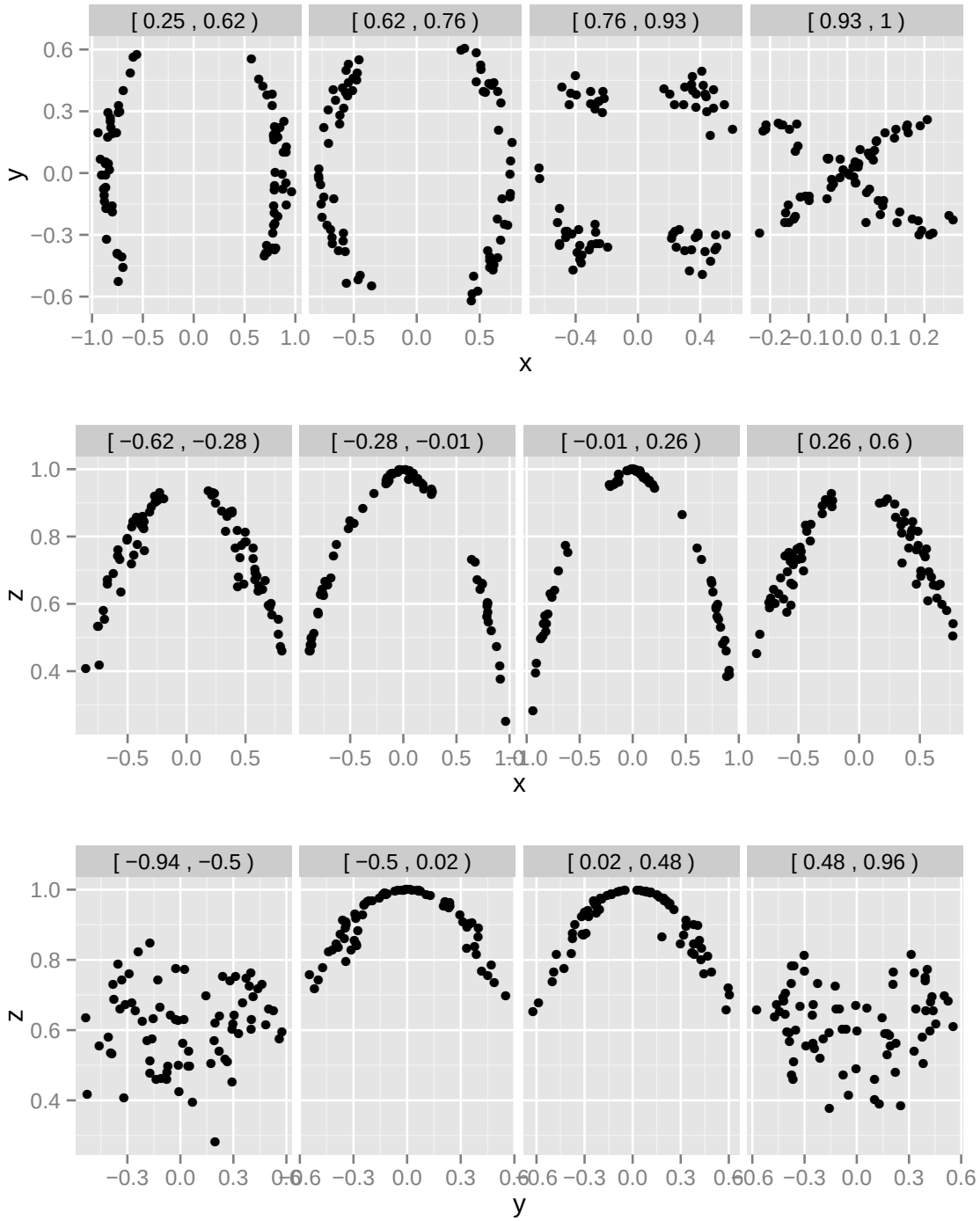


Figure 4.6: Slice plots of the sixth association pattern in Figure 4.1. The scatter plots are based on one random simulation.

fore, our framework of maximizing the association between two sets of variables, Y and X, via algorithmic search of an optimal set of kernel parameters (e.g., weights for the dimensions of Y and X) can be naturally utilized for variable selection.

There are two groups of approaches for feature selection in the literature: filtering methods and wrapper methods. Filtering methods rank the features according to some criterion that intends to measure each feature's individual prediction strength to the classification question. For example, one can compute for each feature the correlation with the outcome and then rank them accordingly. Kernel-based association measures developed in this thesis can also be used to rank the variables, as evidenced in Section 4.2. Multi-feature filtering criteria are also possible, one example being the pairwise influential measure (see Section 2.1.3). Note that typically the filtering criteria are *not* related with the follow-up classification. Wrapper methods select features by classification. Specifically, it starts from all features as candidates, using all candidate features to build a certain classifier, e.g., SVM or a neural network, and uses classification accuracy to assess the performance. The contribution of each feature in the classifier is evaluated by some criterion, for example, for linear classifier, according to the weights (coefficients) of the features in the classifier. This process is carried out recursively, i.e., a new subset of candidate features is selected according to the above criterion and a new classifier is built by restricting the data to the dimensions in this subset. Wrapper methods have consistent criteria for both classification and feature selection with no extra model/criterion for the selection step. Interactions between variables are also considered. However, there might be more of a risk of overfitting in selection. Especially when the sample size is small, the risk of getting a "good" combination of features by chance might be larger comparing to that of single-feature selection. Furthermore, the computational burden for wrapper methods is heavier due to the construction of a new classifier at each step. Hybrid procedures are also possible, e.g., certain initial variable selection before the recursive procedure. Ideally, a feature selection method should combine the advantages of the two types of methods, i.e., less computation and classification-relavent criteria. We shall see (e.g., in Section 5.4) that procedures based on kernel association measures can have this desirable characteristic by considering the same metric space during both feature selection and classification. But before that, let us discuss

a method developed in [Chapelle *et al.*, 2002] with this flavor.

The problem of parameter and feature selection for Support Vector Machines (SVMs) is very crucial to the performance of constructed classifiers and have long been studied. One traditional strategy often used in practice is exhaustive search, i.e., running the algorithm on a grid of parameter values. This is usually very time consuming when there are multiple parameters, which is usually the case for SVM. Another direction is to minimize errors. For example, [Chapelle *et al.*, 2002] proposed such a method, which includes the following steps. First, test errors are estimated, e.g., given a validation set $\{(\boldsymbol{x}_i', y_i')\}_{1 \le i \le m}$,

$$T = \frac{1}{p} \sum_{i=1}^{m} \Psi(-y_i' f(\boldsymbol{x}_i'))$$

where Ψ is a step function: $\Psi(x) = 1$ when $x > 0$ and 0 otherwise. This function is not differentiable so that it is replaced by a smoothed version, e.g.,

$$\Psi(x) = (1 + \exp(-Ax + B))^{-1}$$

where constants A and B have to be chosen (which is difficult). Kernel parameters can then be optimized by computing the gradient of the error estimate. It can be seen that the above procedure contains several steps and is not very straightforward.

The situation becomes simpler in our framework. Because most commonly-used kernels are smooth functions, one can compute the gradient directly to maximize the association measure. We show how this can be further simplified into a greedy search procedure, and evaluate its performance using simulations in this section. This can guide feature selection in practice as shown in the following.

4.3.1 RBF kernel association measures for variable selection

In this section we consider a naive procedure for feature selection through finding optimal scaling parameters as shown in Section 4.2. This is based on the conjecture that if one of the input dimensions is irrelevant for the classification problem, its scaling factor is likely to become small. In other words, if a scaling factor becomes small enough, it means that it is possible to remove the corresponding component without affecting the classification performance [Chapelle *et al.*, 2002]. This naturally leads to the following idea for feature selection: keep the features whose scaling factors are the largest.

We generate a double helix data set as shown in Figure 4.7 to test the above idea. Specifically, three dimensions out of 10 are relevant (as shown in the pairwise scatter plots in Figure 4.8), and the rest are just Gaussian noise. The double helix consists of two "swiss rolls", which makes the first two dimensions symmetric and marginally non-informative (the left panel in Figure 4.8). The third dimension also contains complete information for the classification problem (the y axis in left panel in Figure 4.8). In other words, one can build a perfect classifier based on either the first two features or the third feature. In this sense, the first two dimensions and the third dimension contain redundant information.

We standardize all the features to zero mean and unit variance. We consider the kernel distance covariance with the RBF kernel for the input space and the linear kernel on class labels, and choose the scaling factors in the RBF kernel by maximizing the association measure. It might be interesting to have a look at the trace of the scaling coefficients through the optimization process, which is shown in Figure 4.9 from a simulated data of 80 points. It can be seen that the scaling parameters are initialized to the same value $(1/(2d)$, where d is the dimension of the data. See, e.g., [Song *et al.*, 2012]), and stabilized after around 20 iterations. The result is as expected in such a situation: the coefficients for the irrelevant features are smaller, so that these coefficients may be interpreted as measures of the relevance of the corresponding features to the classification problem. Interestingly, the adopted measure also automatically picks one of the two redundant groups in the input space (the first two features) for this realization. This whole optimization process runs for only a few seconds.

4.3.2 A backward recursion procedure for kernel-based feature selection

The experiments in 4.3.1 show the promise of using kernel-based association measures for feature selection. However the procedure presented therein requires optimizing over a large number of parameters if the dimension is high, which is often the case in practice. One possibility is to consider a few features at a time and use the trained weights for feature selection. This may require shrinking some of the weights to zero, which we shall discuss in Chapter 6. For now we treat the results shown in 4.3.1 more as a demonstration of the meaning of the trained weights (we will come back to this in Section 5.3).

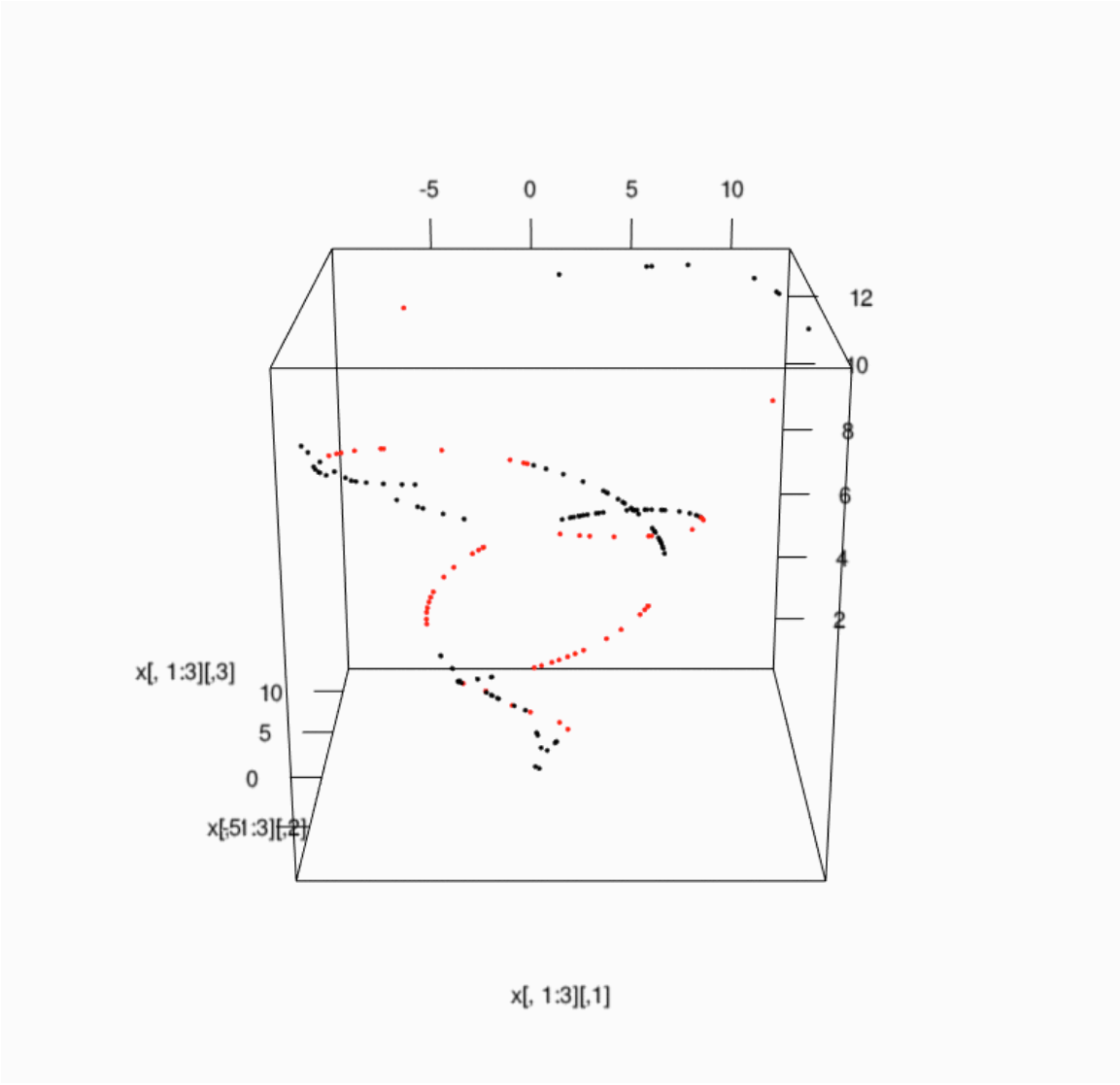


Figure 4.7: Double helix with class labels indicated by colors.

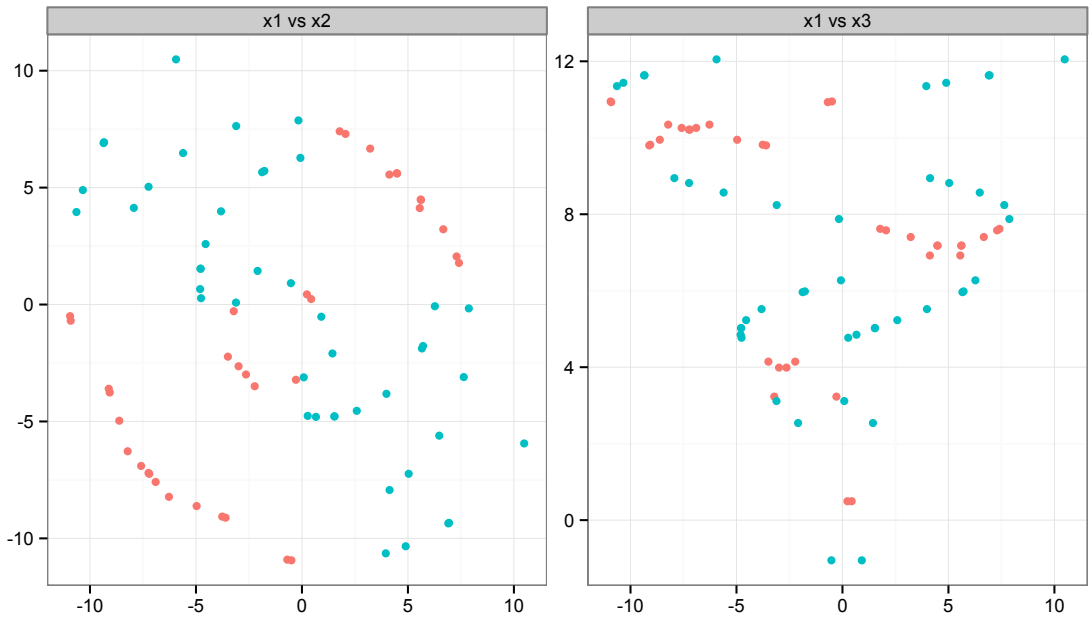


Figure 4.8: Pairwise scatter plots for the three informative features in the simulated double-helix data, with class labels indicated by colors.

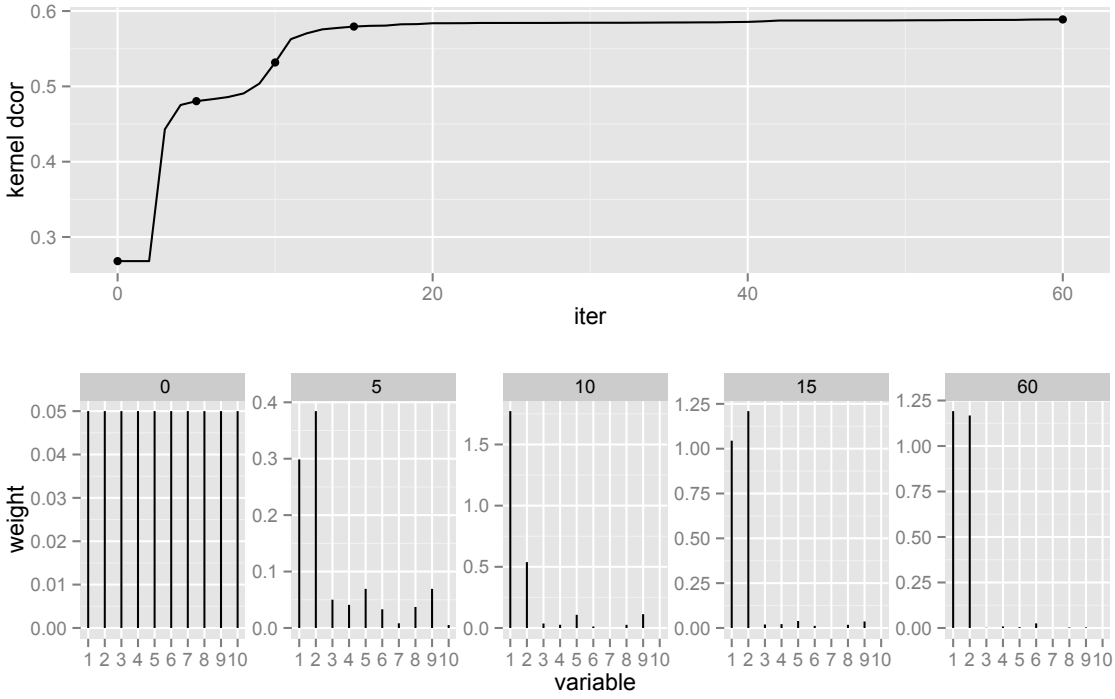


Figure 4.9: The evolution of the scaling factors (weights) through the optimization procedure. For this particular realization there are 60 iterations in the optimization process (the upper panel), of which 5 are plotted (the lower panels, indicated by dots in the upper panel).

Actually there are two important aspects of the feature selection problem: the choice of a criterion and a selection algorithm ([Song *et al.*, 2012]). An ideal criterion quantifies the relevance of a feature subset (to the class label), so that a global optimum for the criterion (by restricting the data to a feature subset) is the solution to the variable selection problem (under this criterion). Unfortunately, finding a global optimum is typically NP-hard [Weston *et al.*, 2003]. Therefore an efficient algorithm is needed to solve the combinatorial optimization problem approximately.

Sections 4.2 and 4.3.1 shows that kernel-based association measures can be a powerful candidate criterion to perform feature selection. In this section we shall develop a backward elimination algorithm to be used together with kernel-based criteria. We first describe the general procedure, and then study in detail two realizations with the criteria being the kernel influential measure and the kernel brownian distance defined in Examples 1 and 2, respectively. Specifically, for the realization with the kernel influential measure, we give the conditions for the corresponding procedure to be consistent (Theorem 1); for the realization with the kernel distance covariance, we illustrate different choices of kernels using simulations.

As discussed earlier, the goal of feature selection is to pick out the informative features. In other words, one can treat the scaling factors in association measures as indicators. Specifically, $\sigma_i = 1$ suggests keeping the ith dimension, while $\sigma_i = 0$ means the ith dimension is irrelevant. Thus maximizing a criterion with scaling factors is done in a discretized space of the σ_i's. This becomes a search problem, the solution of which can be approximated by the procedure described bellow.

1. Calculate the kernel-based criterion using all the X variables.
2. While there are still variables left
2.1 Remove each variable and re-calculate the kernel-based measure.
2.2 Eliminate the variable that results in the maximum criterion value.

We first consider the procedure with the kernel influential measure as the criterion. For classification problems, the kernel influential measure with Mahalanobis distance can be redefined as

$$I_{\Pi}^{M}(X,Y) = n^{-1}\sum_{j=1}^{2} n_j^2(\bar{X}_j - \bar{X})^T S_x^{-1}(\bar{X}_j - \bar{X}) \tag{4.4}$$

Note that here we are treating the class label Y as X in the original definition and using it for partition. Consider a diagonal matrix Σ with its diagonal elements scaling factors for the corresponding dimensions, $\sigma_i \geq 0$. Then one can define the rescaled features as

$$X' = \Sigma X \tag{4.5}$$

This corresponds to a linear kernel on X. Plugging (4.5) into the definition (4.4), one can maximize $I_\Pi^M(X', Y)$ with respect to Σ as discussed earlier. One difficulty is that (4.4) may be monotone increasing in σ_i. This can be solved by imposing some constraint on the sum of σ_i's. We will not pursue the mathematical details here since we will investigate the more straightforward search algorithm described above.

Now let us discuss the consistency of the kernel influential measure used as the criterion for variable selection. Model selection consistency is defined as

$$P(\{i : \hat{\sigma}_i^n \neq 0\} = \{i : \sigma_i^* \neq 0\}) \to 1, \text{as } n \to \infty$$

where σ_i^*'s are the true scaling factors, and $\hat{\sigma}_i^n$'s are the outputs from a feature selection algorithm based on n samples. This requires

$$I_\Pi^M(\Sigma^* X, Y) = \max_\Sigma I_\Pi^M(\Sigma X, Y) \text{ with probability 1} \tag{4.6}$$

where Σ^* is the scaling matrix corresponding to the true features. For simplicity, assume different dimensions of X are independent, and each component of X is normalized, such that the covariance S_x becomes the identity matrix. (4.4) then becomes

$$I_\Pi^M = n^{-1} \sum_{j=1}^{2} n_j^2 \sum_{i=1}^{p} \sigma_i^2 (\bar{X}_{ji} - \bar{X}_i)^2 \tag{4.7}$$

where $\bar{X}_{ji}$ and $\bar{X}_i$ are the corresponding averages on the ith dimension. This suggests the following conditions (which we shall call the *identifiability condition*)

$$\begin{gathered} EX_{ji} \neq EX_i \text{ for the informative dimensions} \\ EX_{ji} = EX_i \text{ for the non-informative dimensions} \\ \text{for } j = 1, 2 \end{gathered} \tag{4.8}$$

Clearly (4.6) holds under the identifiability condition (4.8) due to central limit theorem. Thus we have proved the following theorem for features with independent components:

Theorem 1. *Under the identifiability condition in (4.8), the kernel influential measure is consistent in terms of model selection.*

Now we move on to the kernel distance correlation as the criterion in the backward dropping procedure. Here we use the same double-helix example as in Section 4.3.1, except that we increase the dimensions to 23. We impose the proposed procedure with different kernels in the distance correlation, namely, the linear kernel (which results in the original distance correlation), the RBF kernel, and the angular kernel. The angular kernel distance is defined between two vectors z and z' as (see, e.g., [van der Laan and Pollard, 2003])

$$\theta(z, z') = \arccos(\frac{z^T z'}{||z||||z'||}) \tag{4.9}$$

where $||\cdot||$ is the Euclidean norm. Thus it is simply the angle between the two vectors under consideration. Indeed, the angular distance is extensively used in the literature for hyperspectral data classification due to its invariance to the spectral energy [Keshava, 2004; Honeine and Richard, 2010]. Considering the fact that each of the swiss rolls in the double-helix in our simulated data is actually a one-dimensional manifold indexed just by the angle in the three-dimensional space (Figure 4.7), the angular distance (4.9) can be treated as the "right" kernel for this question. Our numerical studies as described later actually shows that backward elimination with the angular distance covariance discovers all the three informative features.

Consider the following example to best understand the relation between the deletion and information changes: Figure 4.10 shows a typical application to a simulated data containing 80 points. Initially, before any feature is deleted, the class information measured by the angular distance correlation is relatively low. This is because the amount of information has been swamped by the noises and irrelevant dimensions due to these features. As screening out more and more features, the information grows. The algorithm will stop at the peak of the score, which in this case is around 0.38 (remember the kernel distance correlation is between 0 and 1). If the deletion process is forced to continue, the information would start to drop dramatically. For comparison, we also include in Figure 4.10 an information flow plot for the same data but with permuted class labels. It is easily seen that there is no important information contained in those features since the curve stays relatively constant

throughout all deletions and eventually, the algorithm returns no features.

We then apply the backward elimination procedure with linear, RBF and angular distance correlations to randomly generated data sets to illustrate the difference in the performance. We compare backward elimination with the marginal RBF distance correlation, Pearson's correlation, and MIC (Section 2.1.4). We aim to show that when complex dependencies exist in the data, our backward dropping algorithm with appropriate kernels is very competent in finding them.

We instantiate the artificial data sets over a range of sample sizes (from 40 to 160), and plot the median rank, produced by various methods, for the first three dimensions of the data. All numbers in Figure 4.11 are from 300 independent runs, with the sizes of the ovals proportional to the standard errors. It can be seen that backward elimination with kernel distance correlations shows good performance. More specifically, we observe backward elimination with angular dCor correctly selects the first three dimensions of the data almost every time even for small sample sizes. Backward elimination with RBF-kernel dCor selects the first two interactive features (one of the two groups of redundant features), which is consistent with the observations in Section 4.3.1. The third dimension is dropped out because only the first two features contain complete information for the classification problem as measured by the RBF kernel. We manually remove one of the first two dimensions, and the third dimension gets a lower rank (results not shown), which confirms the above conjecture. This implies that in practice when the data structure is unknown, RBF kernels are still a safe choice. Backward elimination with the original dCor also converges to the first two features as the sample size increases, but cannot pick up the correct answer for small samples. Since the first two features have nonlinear interactions between each other and do not contain marginal information, the three marginal methods should only find the third feature that is marginally associated with the class label. This is what is observed for the marginal RBF-kernel dCor. Pearson's correlation fails because it evaluates the goodness of each feature independently. Hence it is unable to capture the nonlinear interaction between the first two features, and the nonlinear association between the third feature and the class label (See Figure 4.8). MIC captures the third feature as expected, but may have overfitting problems for the first two marginally non-informative

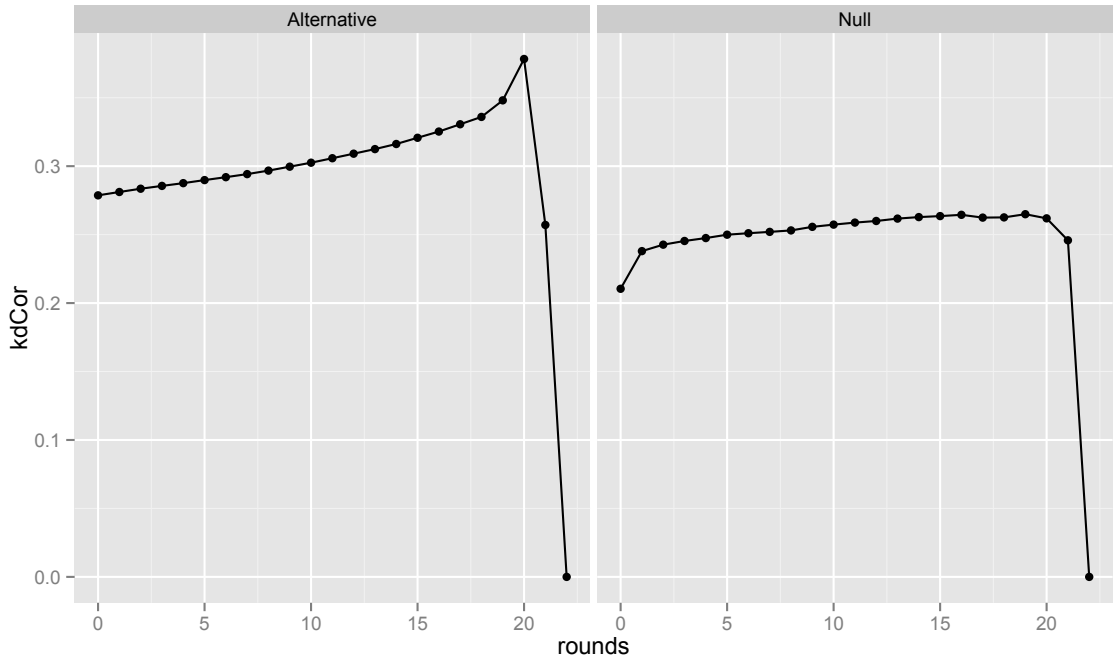


Figure 4.10: Information (the angular distance correlation) flow during screening (due to the properties of the angular distance, the distance correlation for the last round is set to 0). As shown in the left plot: at first the information regarding the class label is contaminated by the noise due to the unassociated features; as screening out more and more irrelevant features, the information indicated by the distance correlation begins to grow and the algorithm will stop at the peak (due to a significant drop after this deletion), thereby returning the 3 important features. The right plot: the class labels are permuted so that no feature has an association with the labels. Thus the information stays relatively low throughout the screening and the algorithm will return no features.

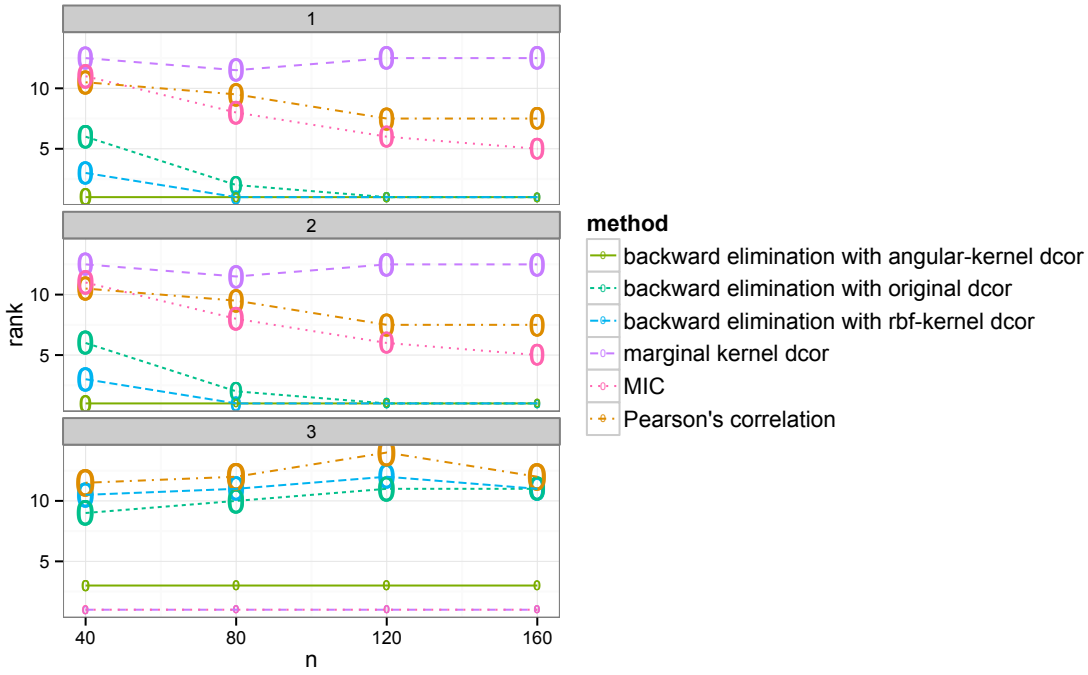


Figure 4.11: The performance of different methods when varying the number of observations. Plotted is the median rank (y-axis) of the three relevant features as a function of sample size (x-axis) for the double helix data sets. The sizes of the ovals are proportional to the standard deviations from the 300 simulations.

features when the sample size is large. This may be because MIC conducts finer partitions for larger samples (see Section 2.1.4).

In practice there may be a lot more variables. The above procedure might become computationally intractable. In this case one can select a subset of variables of a moderate size at random from the original set of all the variables (as in the Partition Retention method [Chernoff *et al.*, 2009] or the random forest [Breiman, 2001]). This process is to be carried out many times, and variables can be selected based on the return frequencies (see Section 5.4 in Chapter 5).

Chapter 5

Real data applications

In this chapter we show how kernel-based association measures can be used in practical problem by four real data applications. Specifically, the first three are problems from statistical genetics, while the last one is gender prediction from handwriting. We focus on four aspects in each of the applications, respectively: (1) *de novo* kernel construction, i.e., defining kernels that are tailored for the problems at hand; (2) adopting existing kernels; (3) functional output spaces; (4) functional input spaces.

5.1 Association screening for genes with multiple potential rare variants using inverse-probability weighted kernel

Here we consider applications of the proposed framework to genome-wide association studies (GWAS). GWAS is an examination of many genetic variants in different individuals to see if any variants or their combinations are *associated* with a trait. It has been believed that both common variants and rare variants are involved in the etiology of most complex diseases in humans. Developments in sequencing technology have led to the identification of a high density of rare variant single-nucleotide polymorphisms (SNPs) on the genome, each of which affects only at most 1% of the population. Genotypes derived from these SNPs allow one to study the involvement of rare variants in common human disorders. In this section, we propose an association screening approach that treats genes as units of analysis. SNPs within a gene are used to create partitions of individuals, and inverse-probability weighting

is used to overweight genotypic differences observed on rare variants. The focus is to show how kernels based on genotypes can be defined for the particular application. The partition of individuals allows association between a phenotype trait and the constructed cluster label readily evaluated by existing association measures. Specifically, we consider three association tests (one-way ANOVA, chi-square test, and the partition retention method) and compare these strategies using the data from the Genetic Analysis Workshop 17 [Almasy *et al.*, 2011]. The proposed method identifies several genes that contain causal SNPs as top genes.

5.1.1 Background

Rare variants are common on the genome and have long been speculated to be involved in the etiology of most human disorders [Pritchard, 2001]. In the 2000s, a large number of genome-wide association studies (GWAS) were conducted using relatively more common single-nucleotide polymorphisms (SNPs) (with minor allele frequency [MAF] $> 5\%$). Most of the common variants identified in these studies have borderline odds ratios and can explain only a small fraction of susceptibility to a disease [Asimit and Zeggini, 2010]. As a result, there has been increasing interest in the study of rare variants for complex diseases. This concern has also been fueled by advancements in sequencing technology. In particular, the availability of such technology has directly led to the implementation of the 1000 Genomes Project (http://www.1000genomes.org/), in which 1,000 genomes from individuals of different ethnic backgrounds were sequenced, consequently leading to the identification of a large number of rare variants (SNPs) with MAF $< 1\%$ and some very rare variants with MAF $< 0.5\%$. Because of such low MAFs, association methods developed for common variants have limited efficiency for mapping rare variants in population studies. For these methods to have adequate power to detect individual rare variants, the sample size needs to increase substantially as the MAF decreases.

It is also more likely for a rare variant to contribute to the susceptibility of a disease as part of a group of rare variants in the same gene or pathway. Therefore grouping or collapsing rare variants is the most feasible option to improve efficiency in studying rare variants. Usually, the grouping is constructed on the basis of functional relevancy, physical

proximity, or both. Once rare variants have been grouped, their genotypic information is combined, or collapsed, into a usually univariate score, and the association between the group of rare variants and the disease is then studied using the association between the univariate score and the disease traits. See [Asimit and Zeggini, 2010] and [Dering *et al.*, 2011] for excellent reviews of different methods for rare variant association analysis, including single-marker, multimarker, and various collapsing strategies. A popular alternative to collapsing genotypic information is to combine single-SNP statistics.

In this application, we consider a gene-based association analysis for rare variants. This is equivalent to grouping based on the gene affiliation of SNPs. We propose using a clustering-based method for collapsing genotypic information of multiple SNPs within each gene. The clustering is based on an inverse-probability weighted sum of genotypic distances that highlights the variation at rare variant loci. Association between the collapsed partition label and the disease traits can then be readily evaluated using single-marker association methods, such as one-way analysis of variance (ANOVA), a chi-square test, and the partition retention method [Zheng *et al.*, 2011; Chernoff *et al.*, 2009]. We apply our approach to the data of the Genetic Analysis Workshop 17 (GAW17) without knowledge of the models that generate the association based on real sequencing data. After the workshop, a comparison of our results with the answers led to interesting observations regarding both the method and the data. We discuss these observations in Section 5.1.5.

5.1.2 Data set

The data set of GAW17 is a combination of real sequence data and simulated phenotypes. An exome of 3,205 autosomal genes, corresponding to 24,487 SNPs, was selected. Sequences of these SNPs were obtained from the 1000 Genomes Project on 697 unrelated subjects. SNPs with missing values were imputed using fastPhase. A majority of the SNPs (74%) were rare variants (MAF $< 1\%$). Two hundred phenotype sets were simulated based on these common genotype data. Each simulated unrelated-individual data set has three quantitative trait values (Q1, Q2, Q4) and the Affected status Y, with 209 case subjects and 488 control subjects. Gene information and SNP information were provided. Especially, whenever available, SNPs were labeled as synonymous or nonsynonymous [Almasy *et al.*, 2011].

Table 5.1: Inverse probability kernel: allelic similarity scores

	Individual 1	
Individual 2	a	A
a	$\frac{1}{p_a^2}$	$-\frac{1}{p_a(1-p_a)}$
A	$-\frac{1}{p_a(1-p_a)}$	$\frac{1}{(1-p_a)^2}$

5.1.3 Inverse probability weighted kernels for gene-based grouping and collapsing of SNP genotypes

We propose to evaluate an individual gene's association with disease traits. SNPs within a gene are grouped for the association analysis. Our main focus is a collapsing strategy for multiple-SNP genotypes within a gene. We propose to create partitions of individuals (or observed genotypes) based on their genotypic similarities evaluated by inverse-probability weighted kernels. It is easier to start with considering alleles at a single SNP locus first. For two individuals, we can count when they have the same alleles or different alleles. When the MAF is small, the chance of having a random match for the major allele is high. On the other hand, if a rare variant is involved in the etiology of a disease, then the case subjects are more likely to have the same rare variants than the control subjects are. Therefore for rare variant association analysis we want to overweight the allelic or genotypic similarity for the minor alleles but not that for the major alleles.

We use the inverse-probability weighted kernel, as defined in Table 5.1 (where p_a is the population frequency of minor allele a). Actually, the kenrel is used as a similarity measure here. This measure has a mean similarity 0, which is also a desirable property. The allelic kernel can be straightforwardly generalized to the genotypic kernel in Table 5.2. For example, an individual 1 with genotype aa and an individual 2 with genotype Aa will have one match (a, a) and one mismatch (a, A). Because a is the minor allele, the (a, a) match will dominate the (a, A) mismatch, and these two individuals will have a high similarity score. Such a weighting scheme implicitly assumes that individuals with the same rare variants will be clustered together for association analysis with the disease outcomes.

We denote the genotypic similarity score between two individuals i and j at SNP k by

Table 5.2: Inverse probability kernel: genotypic similarity scores

	Individual 1		
Individual 2	aa	aA	AA
aa	$\frac{2}{p_a^2}$	$\frac{1}{p_a^2} - \frac{1}{2p_a(1-p_a)}$	$-\frac{1}{p_a(1-p_a)}$
aA	$\frac{1}{p_a^2} - \frac{1}{2p_a(1-p_a)}$	$\frac{1}{2}[\frac{1}{p_a^2} + \frac{1}{(1-p_a)^2} - \frac{1}{p_a(1-p_a)}]$	$\frac{1}{(1-p_a)^2} - \frac{1}{2p_a(1-p_a)}$
AA	$-\frac{1}{p_a(1-p_a)}$	$\frac{1}{(1-p_a)^2} - \frac{1}{2p_a(1-p_a)}$	$\frac{2}{(1-p_a)^2}$

$\text{sim}(i, j; k)$. For a given gene G, the similarity between i and j is defined as the sum of the similarity scores on SNPs within the gene:

$$\text{sim}(i, j) = \sum_{k \in G} \text{sim}(i, j; k) \tag{5.1}$$

For the 697 individuals, pairwise similarity scores, the $\text{sim}(i, j)$'s, are evaluated first and are then converted to a distance measure using the transformation

$$d(i, j) = \exp[-a\text{sim}(i, j)] \tag{5.2}$$

where a is a normalizing constant such that the distance calculated at each gene is bounded by e^{20}. We then apply hierarchical clustering using Wards method [Dasgupta *et al.*, 2011] and partition individuals into groups by cutting the hierarchical clustering tree into a pre-specified number of groups (we consider partition sizes of 5 to 10). See Figure 5.1 for an example using *FLT1*. We also take advantage of the synonymy information about the SNPs by carrying out two separate analyses using nonsynonymous SNPs only or every SNP in a gene.

5.1.4 Partition-based association analysis

After obtaining the partition of individuals, for each gene we tested the association between the partition indexes obtained from the SNPs in that particular gene and the disease phenotypes. For the disease status Y, we considered one-way ANOVA, the chi-square test of independence, and the partition retention method [Zheng *et al.*, 2011]. For continuous-valued disease outcomes Q1, Q2, and Q4, we considered one-way ANOVA and the partition retention method.

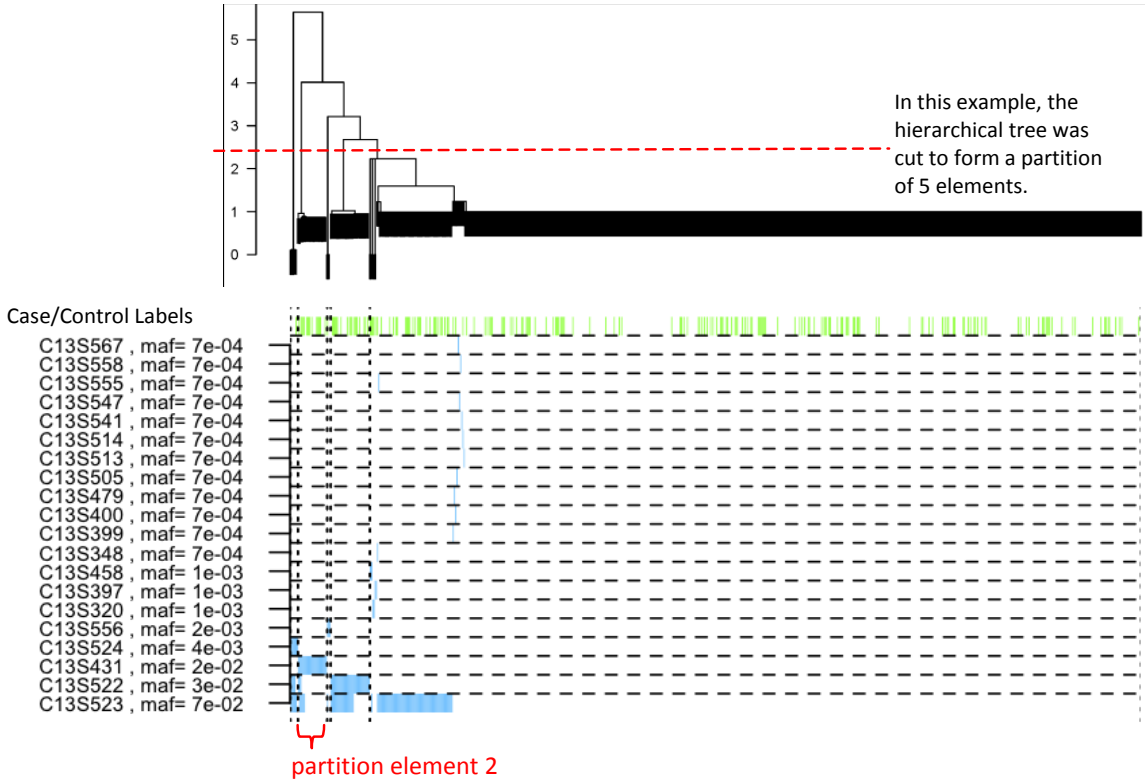


Figure 5.1: Clustering of individuals using nonsynonymous SNPs for *FLT1*. Each row is a SNP, and each column is an individual. Green vertical bars indicate case subjects. Genotype aA is plotted in blue, and genotype AA is plotted in white (a is the minor allele); genotype aa was not observed. The partitions of the 697 individuals are indicated by dotted lines. Partition element 2 is driven by similarity on SNP C13S431 but not on the more common SNPs C13S522 and C13S523.

Recall that the partition retention method is based on association measure I defined between an outcome variable Y and a partition Π (see Section 2.1.3). More specifically, here we use the normalized measure

$$I = \sum_{\Pi_i} \frac{n_i}{n} \frac{(\bar{Y}_i - \bar{Y})^2}{s^2/\sqrt{n}} \tag{5.3}$$

where n_i is the number of individuals in partition element i and $\bar{Y}_i$ is the sample mean of element i. $\bar{Y}$ and s are the sample mean and the standard deviation of all n individuals, respectively. Under the null hypothesis, I asymptotically converges to a weighted sum of chi-square distributions with 1 degree of freedom and therefore has mean 1. The partition retention method is more robust to sparse partition than the chi-square test and can be applied to both dichotomous disease status and continuous-valued traits [Zheng *et al.*, 2011]. Intuitively, the I in the partition retention method evaluates the amount of influence a particular gene has on the disease phenotypes.

P-values for the ANOVA test and the chi-square test are derived from corresponding asymptotic distributions. To address the multiple testing issue, we control the family-wise error rate using the conservative Bonferroni correction. For the evaluation using the partition retention I, we simply chose the top 0.1% of genes for each trait. A further examination of results from chromosome 4 revealed that, by using a cutoff of the top 0.1%, only 15 of the 200 replicates returned any null gene (a family-wise type I error rate), which suggests that the top 0.1% is a reasonable threshold. In practice, we suggest evaluating P-values using permutations (see Section 5.3) and controlling the false discovery rate in order to have better sensitivity to real genetic signals.

5.1.5 Results

Because we have 200 simulated sets of phenotypes, for each gene we counted the number of times it was selected (either in the top 0.1% for I using the partition retention method or significant by Bonferroni correction for ANOVA and the chi-square test) for each trait for each method. We also compared the effects of partition sizes (results not shown). The significance varied between different partition sizes, and the partition size that corresponded to the most significant results also changed from simulation to simulation. Therefore we

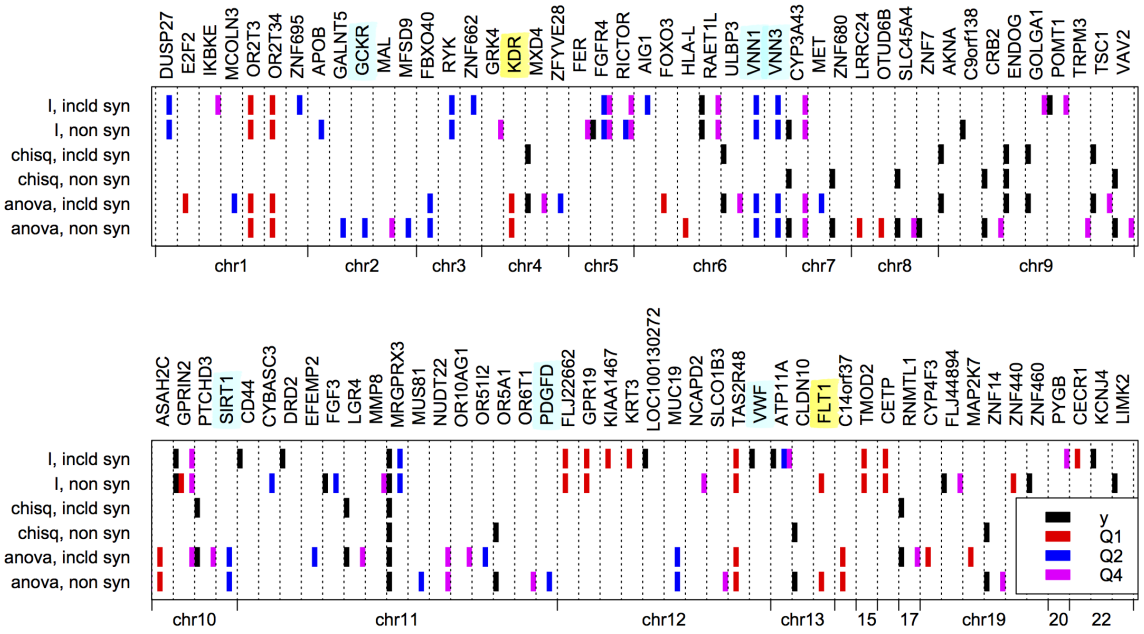


Figure 5.2: Top ten genes identified by each of the methods and for each of Y, Q1, Q2, and Q4. Ninety-one genes are shown, displayed by chromosome. Genes with causal SNPs are highlighted (yellow for Q1 and blue for Q2).

used the average count across six partition sizes (from 5 to 10) to rank genes. By visually examining the average counts (not shown), we observed that Q1 had strong genetic signals and that Q2 and Affected status were harder to map. For Q4, the one-way ANOVA identified many noncausal genes, or false positives, to which the partition retention method was relatively more immune.

Figure 5.2 summarizes the results from the 200 simulations. The top 10 genes for each method and each trait are plotted by chromosome. Note that for Q2 the top 10 genes are identified less than 25% of the time and that the six genes that contain "answers" or causal genes are identified as top genes but with less than 5% probability, with the exception of *VNN1*, which is identified by the partition retention method 22% of the time. Two genes for Q1 (*FLT1* and *KDR*) are identified in more than 50% of the simulated replicates. It is interesting to note that excluding synonymous SNPs led to better identification of *FLT1* and had less effect on identification of *KDR*.

To better understand the "consistent false positives" problem that arose during GAW17,

Table 5.3: Association between a consistent false-positive gene (*OR2T3*) and a causal SNP at C13S523 ($p = 1.8 \times 10^{18}$ by Fishers exact test)

C13S523 genotype	Partition based on SNPs of *OR2T3*				
	1	2	3	4	5
1	41	29	3	9	11
2	525	59	5	8	7

we studied several consistent false-positive genes identified by our methods. All of them were found to be significantly associated with multiple causal SNPs. See Table 5.3 for an example between the gene *OR2T3* on chromosome 1 and a causal SNP at C13S523.

We further investigated the relation between power to detect (probability of true positive) and the effect size of a gene. The effect size for each SNP is provided by [Almasy *et al.*, 2011]. For each gene, we define its total effect size as

$$\text{effect}_g = \sum_{\text{SNP } i \in g} \text{MAF}_i \beta_i \tag{5.4}$$

where β_i is the effect size β used in the simulation model for SNP i, which is 0 for noncausal SNPs.

Figure 5.3 plots the frequencies of each gene with causal genes identified by the best performing method for each trait against the gene-wise effect size, that is, the one-way ANOVA with Bonferroni correction for Q1 and Y and the I from the partition retention method for Q2. The power of our approach suffers greatly for extreme rare variants if the effect size does not scale up as MAF drops.

5.1.6 Discussion

In this section, we propose a novel strategy for gene-based association analysis for genes with multiple potentially rare variants. The inverse-probability weighted clustering approach automatically adjusts weights for rare variants and overweights their genotypic variation when comparing individuals for an association study. Individuals are first partitioned on the basis of their genetic similarity on multiple SNPs in a gene, and this partition is then

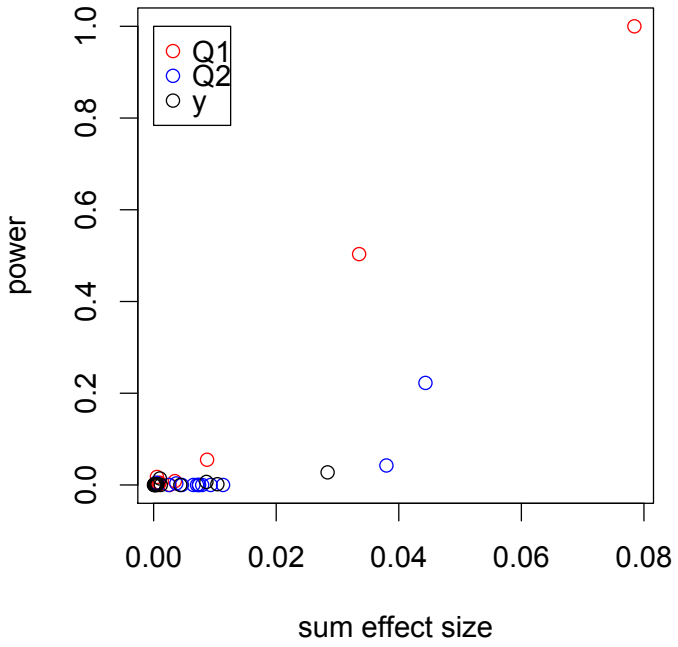


Figure 5.3: Power to identify a causal gene versus effect size. For each trait, we plot the power to detect using the best performing method against the effect size used in the simulation model. That is, we plot the one-way ANOVA with Bonferroni correction for Q1 and Y, and the I from the partition retention method for Q2. The gene-wise effect size is defined as the sum of SNP-wise MAF $\times$ causal SNP effect in the simulation model.

used to calculate association between a gene and a disease trait.

We also considered several association scores and the effect of including synonymous variants. Different methods seem to focus on nonoverlapping signals, which suggests a multimethod approach for future association studies. From our results, we can conclude that our method gains power by considering multiple rare variants in a gene, as illustrated in Figure 5.1 for one of our identified causal genes. It is probably beneficial to consider synonymous and nonsynonymous SNPs in future practice. Filtering out synonymous SNPs corresponds to a weight of 0 being assigned to synonymous SNPs and a weight of 1 being assigned to nonsynonymous SNPs, which can be extended to a smoother weighting scheme as a possible future direction.

For this study, we used asymptotic P-values and the conservative Bonferonni correction because we needed to analyze 200 sets of data. In practice, we suggest evaluating P-values using permutations and controlling the false discovery rate in order to have better sensitivity to real genetic signals. Population information is provided with the data. Some consistent false positives may have resulted from confounding due to population admixture. We recommend using existing methods, such as Eigensoft [Price *et al.*, 2006], to adjust for population stratification in other real applications when applying our method. It should be pointed out that algorithms such as Eigensoft [Price *et al.*, 2006] may convert the original discrete genotype data to continuous values, which requires modification to the kernels defined in Tables 5.1 and 5.2.

5.2 A dual clustering framework for association screening with whole genome sequencing data and longitudinal traits

We have discussed the *de novo* kernel construction for the independent variables (i.e., genetic variants) in the previous section. There we only considered scalar responses (e.g., disease status). In this section we shall present a framework incorporating the inverse probability kernels (defined in Tables 5.1 and 5.2), and able to deal with multivariate response variables.

As mentioned earlier, current sequencing technology enables generation of whole genome sequencing data sets that contain a high density of rare variants, each of which is carried

by at most 5% of the sampled subjects. Such variants are involved in the etiology of most common diseases in humans. These diseases can be studied by relevant *longitudinal* phenotype traits. Tests for association between such genotype information and longitudinal traits allow the study of the function of rare variants in complex human disorders. In this section, we propose an association-screening framework that highlights both the genotypic differences observed on rare variants and the longitudinal nature of phenotypes. In particular, both variants within a gene and longitudinal phenotypes are used to create partitions of subjects. Association between the two sets of constructed partitions is then evaluated. We apply the proposed strategy to the sequencing data from the Genetic Analysis Workshop 18 and compare the obtained results with those from sequence kernel association test (Section 2.1.3) using the receiver operating characteristic curves.

5.2.1 Background

Rare variants have been speculated to be involved in the etiology of complex human diseases [Pritchard, 2001]. Such diseases usually progress over time so that measures of relevant traits at different time points can provide information on the disease development process. For example, the Type 2 Diabetes Genetic Exploration by Next-generation sequencing in Ethnic Samples (T2D-GENES) Project 2 aims to identify rare variants influencing susceptibility to type 2 diabetes using information from whole genome sequencing (WGS) and measurements of related traits (such as blood pressure) at up to four time points. Such WGS genotype and longitudinal phenotype data present new challenges to commonly used statistical methods for association testing in genome-wide studies.

As mentioned in Section 5.1, many genetic variants are rare variants (here we refer to rare variants with minor allele frequencies [MAFs] $< 5\%$). Due to their low MAFs, traditional association methods may suffer from low power. A natural idea for improving power is grouping or collapsing together certain variants. Such collapsing methods are based on the assumption that rare variants in a group (e.g., gene or pathway) may function in combination [Bailey-Wilson *et al.*, 2011]. For example, sequence kernel association test (SKAT) [Wu *et al.*, 2011] assigns different weights to variants in a region and incorporates them into a kernel matrix. We have proposed the inverse-probability weighted clustering

approach in Section 5.1 (see also [Liu *et al.*, 2011]), a gene-based method where inverse-probability weighting is used to overweigh genotypic differences observed on rare variants. The above methods can deal with both continuous and dichotomous traits and have obtained insightful results in different studies. However, leveraging them in an effort to efficiently address longitudinal traits remains a major obstacle.

Longitudinal traits (i.e., time-series phenotypes) provide valuable information regarding the progression of diseases. Traditionally, such longitudinal data can be analyzed using the so-called cross-sectional strategies. In particular, such methods involve repeating the same analysis at various, specific points in time. Since at each time point the trait under consideration reduces to a scalar, methods such as inverse-probability clustering can be conducted for association screening. Then, variants can be selected based on the results from each time point. The assumption underlying this type of strategy is that genetic variants maintain similar influences at different time points. However, it is more likely that those variants influence the pattern of the traits across time; e.g., a group of variants may affect how blood pressure changes in a time-dependent manner. Cross-sectional analysis may fail under such circumstances. A method that takes full consideration of the longitudinal nature of traits is thus desired to capture such genetics-time interactions.

In this section we propose a dual clustering framework, which highlights both rare variants and the longitudinal structure of traits. Here by "dual" clustering we mean individuals are clustered based on both genotypic information through inverse-probability weighting and longitudinal traits through ordinary hierarchical clustering. The focus here is to apply different kernels (corresponding to different metric spaces) to input and output spaces, and use them for partitioning. Association between the two sets of partition labels can then be readily evaluated using existing single-marker and scalar-trait association methods, such as one-way analysis of variance (ANOVA) or the partition retention (PR) method [Zheng *et al.*, 2011; Chernoff *et al.*, 2009]. We apply the proposed approach to the data of the Genetic Analysis Workshop 18 (GAW18) and compare the obtained results with those by SKAT, with some interesting findings.

5.2.2 Data set

The data set of GAW18 is a combination of real WGS data and simulated longitudinal traits. The sequence data is drawn from T2D-GENES Project 2. In this section we use the dosage genotype data on chromosome 3, which include 773,088 SNPs that can be mapped to the genome. Two hundred phenotype sets were simulated based on genotype data. For each data set, we analyze systolic blood pressure (SBP) and diastolic blood pressure (DBP), each with measurements at 3 time points, for 849 related subjects. We map the SNPs to its host gene, resulting in 1,426 genes.

5.2.3 Inverse-probability clustering based on genotypes

Here we use the same inverse-probability kernels as defined in Tables 5.1 and 5.2 to measure the similarity between individuals based on genotypes. In Section 5.1 we adopted an exponential type transformation to induce a distance from the inverse-probability kernel. Other bounded monotone-decreasing transformations can also be applied. Here we try another form of transformation. Specifically, for the 849 individuals, pairwise similarity scores, $\text{sim}(i, j)$s, are converted to a distance measure using the transformation: $d(i, j) = -\text{sim}(i, j)+\max(\text{sim}(i, j))$, such that the pair with the largest similarity has distance 0. We then conduct hierarchical clustering based on the above distances as in Section 5.1 and partition individuals into groups by cutting the hierarchical clustering tree into a pre-specified number of groups (again we consider partition sizes of 5 to 10, Figure 5.4(a)).

5.2.4 Hierarchical clustering based on longitudinal phenotypes

The main difficulty of dealing with longitudinal traits is that most existing association methods only consider scalar phenotypes. Thus it is natural to transform longitudinal traits into some one-dimensional summary statistics. Here we adopt ordinary hierarchical clustering using phenotype vectors and treat the resulting class labels as a summary statistic. Since hierarchical clustering uses the whole longitudinal trait as features, we expect that it can capture the structure contained in the phenotypes. In this study we cluster the 849 individuals into two groups. Results show that these two groups can be treated as with high and low blood pressures (Figure 5.4(b)). Our main focus here is a strategy that turns

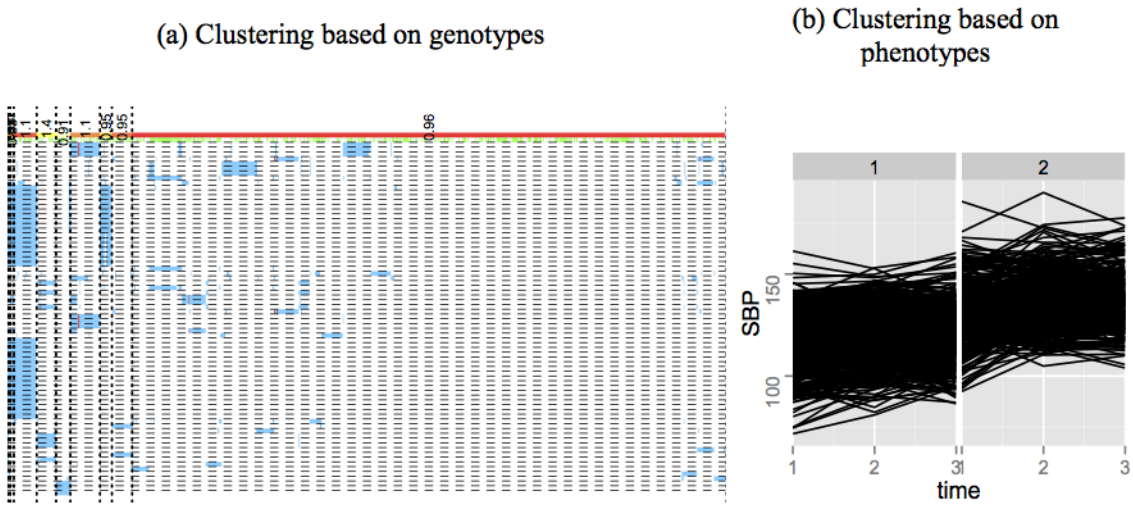


Figure 5.4: Clustering of individuals using SNPs with MAFs between 0.01 and 0.05 for *MAP4*. (a) Shown are 10 clusters, with the numbers at the top odds ratios within each partition block based on blood pressures (see Section 2.1.1 for the definition of odds ratios). Each row is a SNP, and each column is an individual. SNPs are ordered with decreasing MAFs (from top to bottom). Green vertical bars indicate subjects with higher blood pressures (see text). Genotype aa is plotted in red, aA is plotted in blue, and AA is plotted in white (a denotes the minor allele). The partitions of the 849 individuals are indicated by dotted lines. Most partition elements are driven by similarity on rarer SNPs but not on more common SNPs. (b) Clustering of individuals using their SBP curves from the first simulation. It can be seen that individuals are reasonably grouped into one high blood pressure cluster and one low blood pressure cluster.

longitudinal traits into one-dimensional summaries. Other dimension reduction techniques can also be used for this task. We choose to adopt hierarchical clustering for illustration purpose here due to its simplicity, and get reasonable results (see Section 5.2.6).

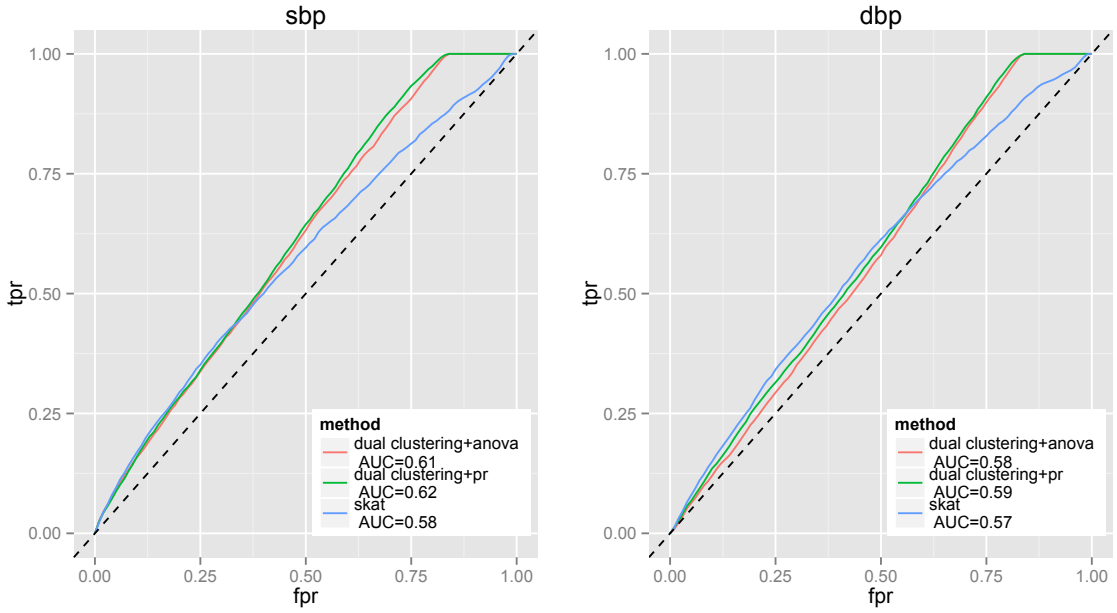
5.2.5 Association analysis based on obtained clusters

After clustering individuals base on both genotypes and phenotypes, for each gene we test the association between the corresponding two sets of partition indices. We consider one-way ANOVA and the partition retention method [Zheng *et al.*, 2011; Chernoff *et al.*, 2009]. Recall that the partition retention method is based on an association measure I defined between an outcome variable Y and a partition Π. Here we take the variable that indicates which cluster an individual is in from longitudinal traits as Y.

5.2.6 Results

We first apply the proposed method to the WGS dosage data including all the 773,088 SNPs and the SBP trait. Three genes are discovered after Bonferroni correction, of which one gene, *Y_RNA*, is significant in 15 out of the 200 replicates. It turns out that this gene resides within *MAP4*, which has the strongest signal in the simulated model, and produces a non-coding RNA.

One reason for the relatively few significant genes obtained above may be that there is a very high density of variants within most genes. We then conduct similar analysis using only SNPs with MAFs between 0.01 and 0.05 to increase power. SBP and DPB are regressed on age, sex, age $\times$ sex, and medication, and the residuals are used in the clustering analysis. For method comparison, we treat genes containing at least one causal SNP in the simulated model as causal genes, resulting in 21 genes for SBP and 26 genes for DBP. We compare the receiver operating characteristic (ROC) curves by the proposed dual clustering framework and SKAT (Figure 5.5). SKAT cannot get results for some of the replicates. It can be seen from Figure 5.5 that all the three methods have relatively low power, among which our dual clustering approach with PR's I has a bigger area under curve (AUC). Results are similar for other partition sizes resulted from inverse-probability clustering.



P-values from paired Wilcoxon signed rank test	SBP	DBP
anova vs pr	0.0001	4×10^{-7}
anova vs skat	4×10^{-11}	0.3664
pr vs skat	2×10^{-14}	2×10^{-5}

Figure 5.5: Average ROC curves across simulation replicates for three methods. Shown are results by 10 clusters using inverse-probability weighting. Areas under curve (AUCs) by different methods are compared using paired Wilcoxon signed rank tests based on the 200 replicates, with resulting P-values shown in the table below.

5.2.7 Discussion

In this section, we propose a dual clustering framework for gene-based association analysis with WGS and longitudinal traits. The first clustering is based on the inverse-probability weighted kernel, which automatically increases weights for rare variants. The kernel values are calculated from empirical MAF estimates. If better estimates are available, the proposed method can incorporate these to achieve improved power. The second clustering treats trait vectors of individuals as features, which accounts for the longitudinal nature of the phenotypes. Individuals are then partitioned based on their genetic similarity on the SNPs in a gene, as well as the similarity of their traits. These two partitions are then used to calculate association between a gene and a longitudinal trait.

Our proposed framework is actually quite general. We define the kernel (used as a similarity measure) based on inverse-probability weighting. Other kernels, such as the one used in SKAT, can also be incorporated into our framework. Other distance-based clustering approaches can be adopted for the first clustering based on similarity measures. The proposed kernel can detect variants with variable directions of the effects. For longitudinal traits, we choose hierarchical clustering due to its simplicity. Hierarchical clustering does not take into account the correlation induced by time. Considering there are only 3 time points in the GAW18 data, we believe that not much information has been lost. If more time points are available, time series clustering methods can be used (see [Liao, 2005] for a survey on commonly-used time-series clustering algorithms). More generally, we use clustering as a means of summarization, so other summarization strategies can also be integrated into the proposed framework. After obtaining the two sets of clustering indices, any association method can be used to measure the association between them. In this study, we choose ANOVA and PR's I. The obtained results are similar but a little better than that from SKAT in terms of ROC curves (Figure 5.5). SKAT has shown superiority to more traditional methods in the simulation studies presented in [Wu *et al.*, 2011]. Many of those traditional methods assume that causal variants have effects with the same direction and magnitude, and do not consider the potential effects of rarer variants to boost power. The purpose of the current study is not to show the absolute superiority of our method, but rather to present a general framework that can incorporate different choices of kernels and

association measures, such as that from SKAT.

Although the simulation model did not take family structures into account, the ANOVA P-values may be inflated due to such structures. However, P-values will be inflated (if any) for both causal and non-causal variants. Therefore the main conclusion based on ROC curves is still valid. In practice, we suggest evaluating P-values using permutations and controlling the false discovery rate in order to have better sensitivity to real genetic signals. This may introduce more computational burden, but it is worth mentioning that the two clustering tasks can be done independently and simultaneously so that the computational time can be reduced. Multilevel models with MCMC techniques may also address the multiple comparisons problem encountered here by partial pooling [Gelman *et al.*, 2012].

5.3 Adaptive kernels for association screening with longitudinal traits

In Section 5.1 we demonstrated the use of customized kernels with pre-defined weights. Such inverse-probability weights are defined so as to highlight the effect of rare variants. In this section we study a more flexible approach, allowing the weights in the kernels to be selected adaptively with data. In Section 5.2 we started dealing with longitudinal traits, where we used the Euclidean distance to cluster the individuals. In this section we use the more flexible kernel distance covariance instead, so that clustering is not required.

5.3.1 Background

There are several methods that use pre-defined weights for testing the association between rare variants and diseases, including the weighted sum statistic [Madsen and Browning, 2009] and our inverse-probability weighted approach as discussed in Section 5.1 and [Liu *et al.*, 2011]. Different methods tend to define different weights that reflect their particular assumptions. One natural question is that, given a particular data set, which weights are optimal. Or rather, can we find a set of optimal weights in a data-driven manner instead of pre-defining them a priori? In this section we explore this question by using the strategies developed in Chapter 4. We also study longitudinal data in this section. The ability of the

kernel distance covariance (Chapter 3) to deal with both X and Y in arbitrary dimensions makes it a natural choice for such data.

We discussed the possibility of permutation tests but did not actually conduct permutations in Sections 5.1 and 5.2. One challenge for such tests is the large number of candidate variants and (thus) the huge number of permutations needed to reach the genome-wide significance level. This makes the computation very expensive and often infeasible. In this section we propose a stepwise multiple test procedure considering randomness in P-values. This procedure can save a large number of permutations and produce credible intervals for P-values as a by-product.

We apply the proposed methods to a GWAS data set for poplars. Poplars are widely distributed in the Northern Hemisphere, from subtropical to boreal forests. They are cultivated for pulp and paper, wood products, lumber and energy. The data set contains measurements of different wood property indices, as well as growth records across 24 years (1987-2010). We obtain interesting results by using our methods, which may lead to novel findings.

5.3.2 Data set

The poplar data contain genotypes of 156,362 SNPs for 66 samples (trees), two of which are parents producing the other 64 progenies. For the genome-wide association screening in this study we only include the 64 progenies. The 156,362 SNPs are on different scaffolds, of which we analyze the first 19 scaffolds corresponding to the 19 chromosomes in the poplar genome. The phenotype data contains various wood property and tree growth traits, including longitudinal data. The eight wood property indices are: fiber length, fiber lumen width, fiber double-wall thickness, fiber diameter (the sum of width and thickness), wood basic density, air-dry density, the rankle ratio defined as double-wall thickness divided by lumen width, and the slenderness ratio defined as fiber length divided by diameter. An increase in wood density is partially due to the increase in the cell wall thickness. The above traits also have the following indications on paper production: higher fiber lengths will result in higher resistance of paper; higher slenderness ratios (>33) are preferred; rankle ratios greater than 1 indicate thick walls and least suitable for paper production, a rankle

ratio of 1 indicates medium thickness and suitable for paper production, while rankle ratios less than 1 indicate thin cell walls and most suitable for paper production. The eight growth records are: the total growth of DBH (diameter at breast height), the average growth of DBH, the annual increment of DBH, the total growth of tree height, the average growth of tree height, the annual increment of tree height, the total growth of volume, the average growth of volume, and the annual increment of volume for 24 years, thus each record is a longitudinal curve of 24 dimensions. Climate information (annually average temperature and participation) is also available.

5.3.3 Association between tree growth and wood properties

We first do some exploratory analysis. Since there are two types of traits, namely, scalar wood properties and longitudinal growth records, a natural question is if there is any association between them. Here we adopt similar strategies as in Section 5.2: first clustering the samples based on the growth time series, then conducting ANOVA test with the cluster index as the group variable, and wood properties as the response. For example, Figure 5.6 shows by different colors the 3 clusters according to the total growth of DBH. It can be seen that the 66 trees are reasonably clustered into three groups featuring fast, medium and slow DBH growth. The following three wood properties are significantly associated with DBH growth according to the ANOVA results: basic density (P-value 0.016), lumen width (P-value 0.026), and slenderness ratio (P-value 0.031).

5.3.4 Association screening with the kernel distance correlation

We use a fixed-bin approach with 10 SNPs per region, thus X is the SNP phenotype (10 dimensional). This relies on the assumption that proximity and similarity go hand in hand. Tree growth is regressed on temperature and precipitation with tree indicator as a random effect, and the residuals are treated as Y and used in the association analysis. We use the kernel distance correlation with RBF kernels for both X and Y as defined in Section 4.2 for association screening. The 10 scaling factors in the RBF kernel for the SNPs are tuned with the approach described in Section 4.2. The idea is to let the data pick the right weighting for each SNP instead of presetting the weights. Such a strategy would adaptively

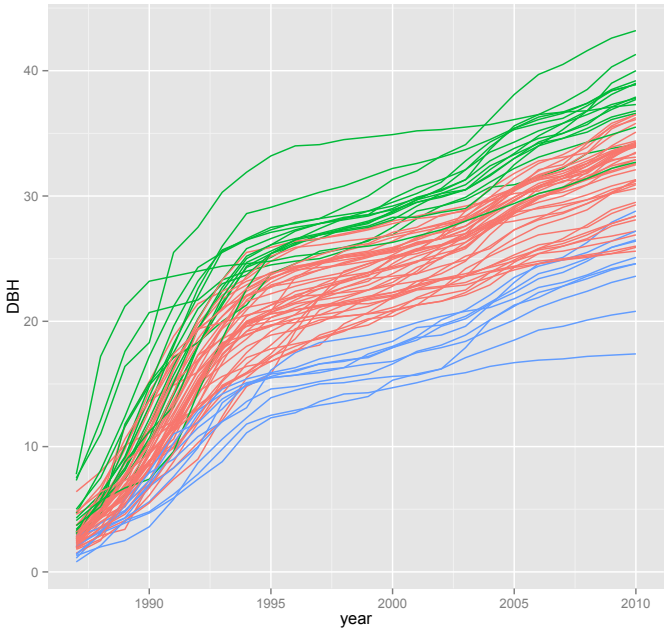


Figure 5.6: Clustering of the trees according to DBH. Green, red and blue curves indicate three groups featuring fast, medium and slow DBH growth.

highlight the SNPs that are more relevant to the traits, as evidenced by the experiments in Section 4.3.1. We run permutations to get the significance level for each of the bins, with the approach described in the next section.

5.3.5 A stepwise multiple test procedure considering randomness in P-values

Permutation tests have become a standard nonparametric tool for the assessment of statistical significance. On one hand, there are some methods aiming fewer permutations (e.g., [Knijnenburg *et al.*, 2009; Gandy, 2012]), often producing P-value estimates with a guaranteed error bound. On the other hand, confidence intervals for P-values have been constructed with several existing methods available in the literature [Li *et al.*, 2009]. In this section we propose a novel stepwise multiple test procedure that combines the above two aspects: estimating P-values with fewer permutations, *by* considering (credible) intervals for P-values. Our approach is based on Bayesian sample size determination for binomial

proportions [M'lan *et al.*, 2007].

Consider a one-sided test with a large value of the test statistic (denoted by T) favoring the alternative hypothesis. The P-value for such a test is defined as

$$P = P_{H_0}(T^* \geq T) \tag{5.5}$$

where T^* has a probability distribution in accordance with the specification of H_0. In permutation tests, the P-value is estimated as the fraction of permutation values at least as extreme as the original statistic derived from non-permuted data, i.e. (with B permutations),

$$\begin{aligned} \hat{P} &= \sum_{b=1}^{B} I(T_b^* \geq T)/B \\ &= M_B/B \end{aligned} \tag{5.6}$$

where T_b^* is the same statistic calculated from the bth permuted sample, thus M_B is the number of more extreme values out of the B permuted statistics. Thus conditional on the original data, the estimation of P-values in permutation tests is just estimating binomial proportions.

A naive way to save computation when conducting multiple tests is to run only a few (say, 10) permutations for all the bins at the first stage, then drop the bins with non-zero estimated P-values (this is because we know those bins will have P-values at least 10%). Since most of the bins come from the null hypothesis, this would leave ∼10% of the bins for the next stage. One can then conduct more permutations (say, 100) for those remaining bins. The process goes on until no bins are left. One problem of the above naive procedure is that even if the estimated P-value is greater than 0 at the first stage, one cannot be 100 percent sure that the true P-value is greater than 10% due to random error. Such randomness should be taken into account when designing the permutation procedure so as to avoid false negatives. We do this by controlling the precision (in terms of the length of the credible interval) of the estimation at each stage using the approach as described briefly below (one is referred to [M'lan *et al.*, 2007] for details).

We assume the following Bayesian model following [Joseph *et al.*, 1995]: $P \sim$Beta(a, b), and $M_B|P \sim$Binomial$(B, P), B \geq 2$. As a result, the marginal predictive distribution of

M_B, $p_{M_B}(m_B|B,a,b)$, is Beta-Binomial, and the posterior distribution of P is Beta($a + M_B, B + b - M_B$). If M_B is known, one can construct the highest posterior density (HPD) interval for P with Monte Carlo methods. Let HPD$(M_B, B, a, b, 1-\alpha)$ be an HPD interval for P of a given coverage $1-\alpha$, and define $l_{1-\alpha}(M_B|B,a,b) = \int_{\text{HPD}(M_B,B,a,b,1-\alpha)} dP$ to be the actual length of that interval. The idea of the proposed test procedure is that at different stages one can specify the maximum length l (minimum precision) of the say, 50% HPD interval, so as to find the smallest number of permutations B required. This can be easily done with M_B known. However, the value of M_B is unavailable before conducting any permutations. Thus one need to define alternative criteria to determine B.

Here we adopt the average length criterion as defined in [M'lan *et al.*, 2007], which seeks the minimum B such that

$$\sum_{m_B=0}^{B} l_{1-\alpha}(m_B|B,a,b)p_{M_B}(m_B|B,a,b) \leq l$$

For this criterion, an approximate sample size formula has been developed as

$$B = 4\frac{z^2_{\alpha/2}}{l^2}\Big(\frac{\mathbf{B}(a+1/2, b+1/2)}{\mathbf{B}(a,b)}\Big)^2 - a - b, \text{ for } a \geq 1, b \geq 1 \tag{5.7}$$

where $\mathbf{B}(a,b)$ indicates the beta function with parameters a and b.

The procedure begins with specifying a ladder of desired precisions, for example, 0.1, 0.01, ..., for each of the stages. The number of permutations can then be calculated using (5.7) for every stage (e.g., for the first stage, $B = 27$), and bins with non-zero estimated P-values are dropped. One issue is the choice of the prior parameters a and b. For the first stage, $a = b = 1$, for which values the prior distribution is the uniform distribution. For the following stages, one needs to incorporate the information from previous stages in the prior. Since a and b correspond to the number of pseudo counts, and the mean of the beta distribution is $a/(a+b)$, it is reasonable to keep $a = 1$ and assign b to be the cumulative number of permutations done so far. In our study we use a more conservative setting with $b = 1/$(precision in the previous stage) which is less than B.

Figure 5.7 shows the required numbers of permutations for the conventional (non-stepwise) and the proposed stepwise procedures. One can see that the number of permutations is much reduced by our strategy. Credible intervals can be produced for the estimated

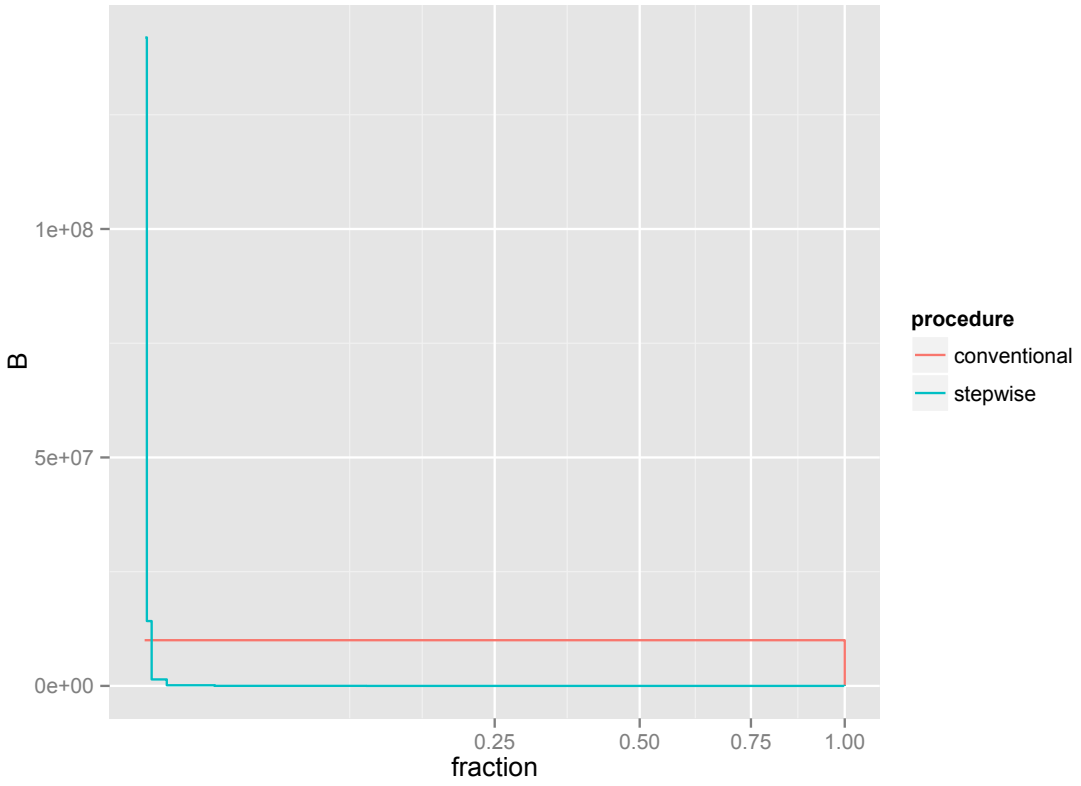


Figure 5.7: The numbers of permutations for the conventional procedure (red) and the proposed stepwise procedure (blue). The x axis (on the square-root scale) is the fraction of the total number of tests. Thus the number of permutations required corresponds to the area under the curve. It shows that the total number of permutations by our procedure is negligible compared to the traditional non-stepwise procedure.

P-values using Monte Carlo methods (we develop a method named Spin for error-reduced credible intervals, see Appendix A).

5.3.6 Results

Figures 5.8 and 5.9 are the Manhattan plots (negative logarithm P-values) for chromosome 1 with the traits being the eight wood properties and the nine longitudinal tree growth records, respectively (Appendix C contains Manhattan plots for the other chromosomes and tree growth). Focusing on tree growth, the bin with the smallest P-value (HPD $[3\times10^{-6}, 3\times10^{-4}]$

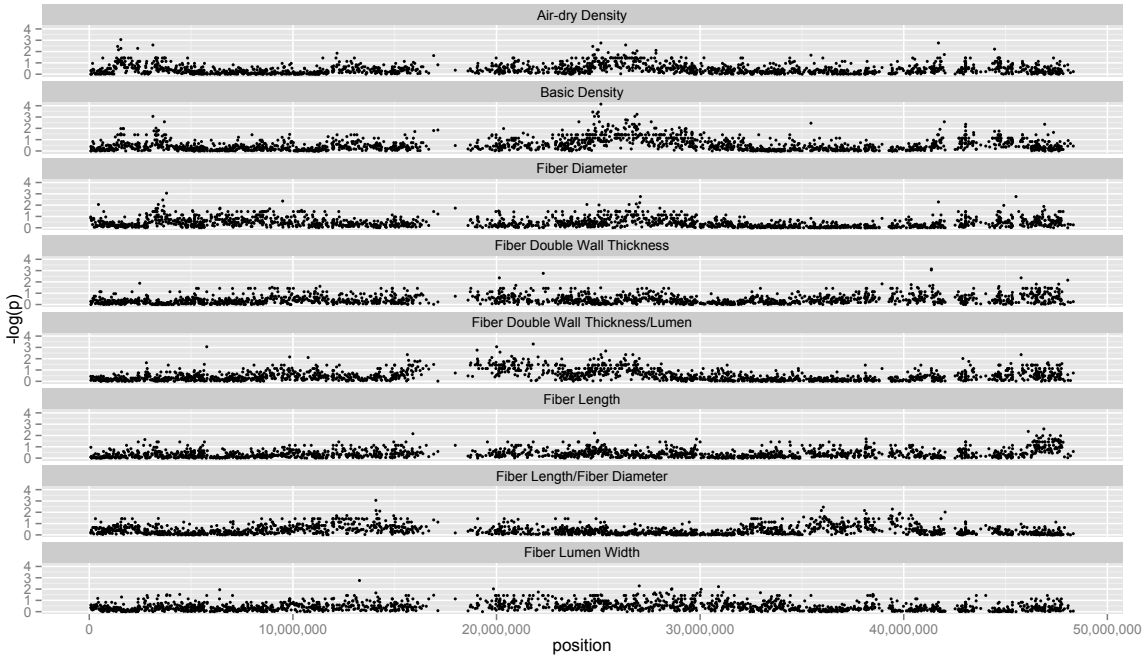


Figure 5.8: Manhattan plot for chromosome 1 and the wood properties.

from the Spin method in Appendix A) is a region (positions 29505890-29531308) associated with total growth of DBH (diameter at breast height). Table 5.4 lists the four genes within this region, together with functional annotations if available.

5.3.7 Discussion

In this section we have proposed a data-driven method for association screening with adaptive kernel distance correlations. The weights for SNPs within a region are selected so

Table 5.4: The four genes in chromosome 1 most associated with total growth of DBH

Gene	Function
POPTR_0001s31170.1	Protein binding / zinc ion binding
POPTR_0001s31180.1	Plant protein of unknown function (DUF869)
POPTR_0001s31190.1	Unknown
POPTR_0001s31200.1	Unknown

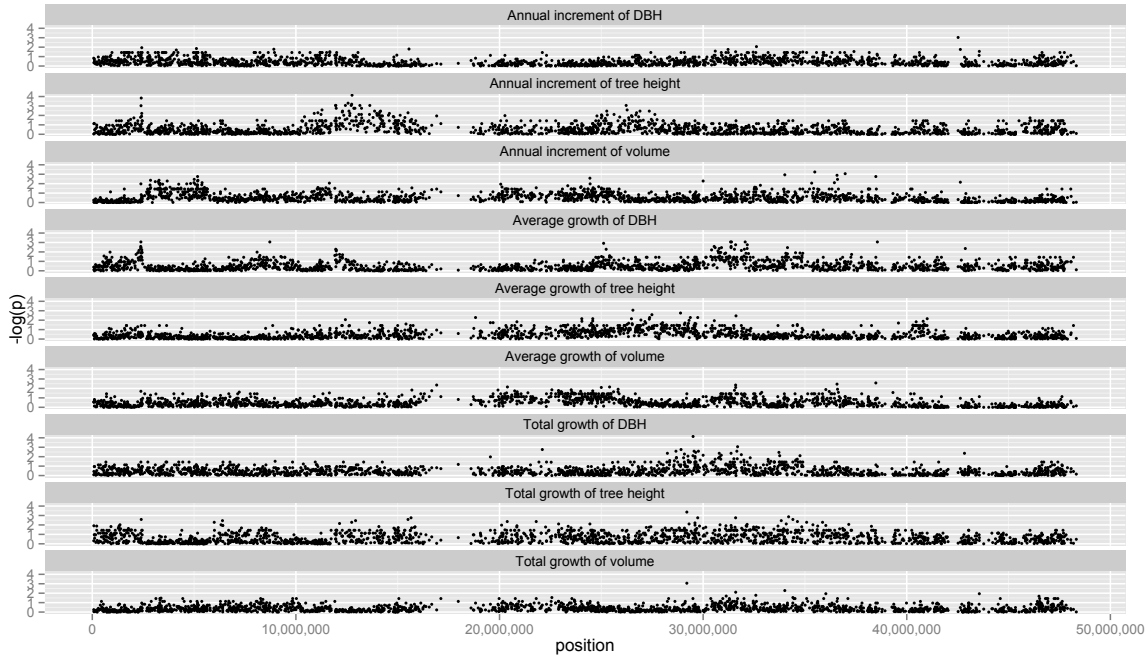


Figure 5.9: Manhattan plot for chromosome 1 and the longitudinal tree growth.

as to maximize the association measure. This is supposed to highlight (in an adaptive way) the SNPs that are more associated with the outcome. We also develop a novel stepwise procedure for multiple permutation tests. The bayesian framework introduced in this procedure naturally incorporates available information from previous steps, which reduces computation to a great extent.

5.4 Feature selection with functional input spaces

In the previous section we demonstrated the use of adaptive kernel distance covariance with functional output spaces. In this section we consider functional input spaces and show how carefully-chosen kernel association measures can be applied in such cases.

5.4.1 Background

We consider feature selection in classification problems in this section. As discussed in Chapter 4, traditional feature selection methods either consider one feature at a time (filtering

methods), or integrate variable selection with classification (wrapper methods). However, most of those methods do not take into account the specific structures of the feature space, especially when the space is functional (the features are samples from one or more functional forms). If such information is available, one can choose an appropriate kernel in the association measures to accomodate the corresponding structure.

In this section we analyse a handwriting data set for gender prediction (i.e., the goal is to predict if a handwritten document has been produced by a male or a female writer) from Kaggle (http://www.kaggle.com/c/icdar2013-gender-prediction-from-handwriting). The prediction of gender from handwriting is a very interesting research field [Bandi and Srihari, 2005; Liwicki *et al.*, 2007; Marcus Liwicki, 2010]. It has many applications, e.g., the forensic application where it can help investigators focus more on a certain category of suspects.

5.4.2 Data set

The handwriting data set consists of the same Arabic handwritten texts produced by 282 writers (Arabic native speakers), for which the genders are provided. We have extracted 4584 non-constant features for each writer. The features have a group structure, specifically, features within the same group form a histogram of a certain geometrical characteristic of the handwriting (curvatures, directions, tortuosities) with different numbers of bins (see [Hassane *et al.*, 2012] for descriptions of those features). In other words, each group of the features is an approximation to the corresponding density function. Readers are referred to [Al-Madeed *et al.*, 2012] for a more detailed description of this dataset. In this study we separate 1/5 of the data (57 individuals) to be a test set.

5.4.3 Classification using RBF kernel distance covariances

We first try RBF kernel distance correlations with both marginal screening and the backward elimination procedure discussed in Chapter 4. Since the dimension is relatively high, we first compute all the pairwise RBF distance correlations with scaling factors optimized, then use the top 400 features to conduct the following backward elimination. Specifically, we randomly pick 10 out of these 400 features and conduct the backward dropping procedure. This process is repeated for 5,000 times. The final set of selected features are based on

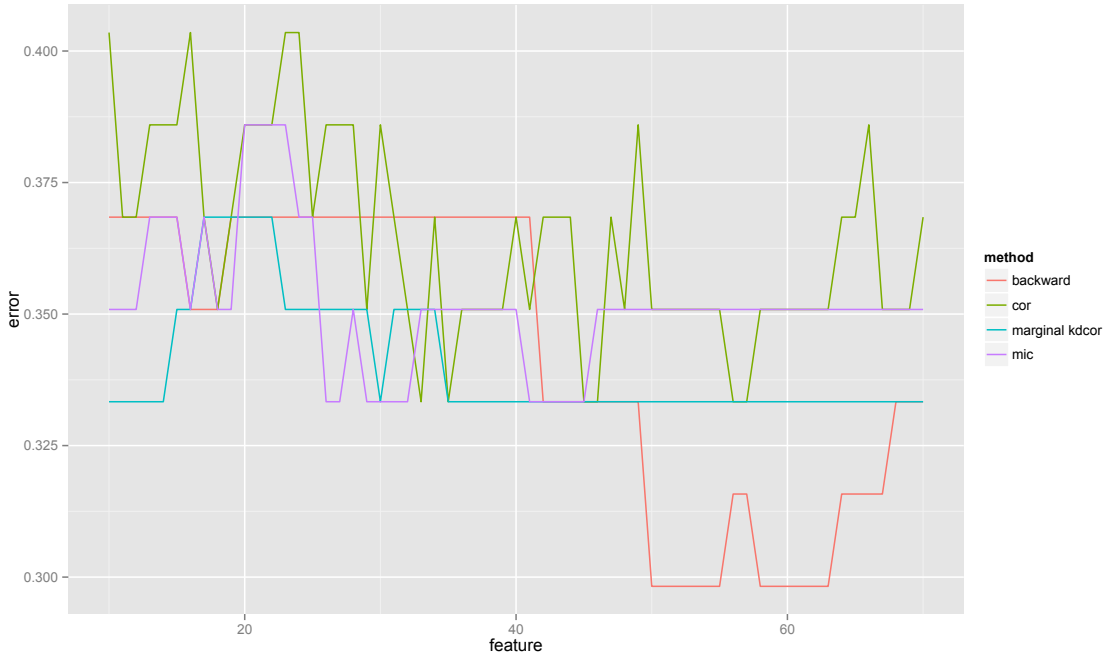


Figure 5.10: Classification errors for the handwriting data set with different numbers of features from different feature selection methods (shown in different colors).

the return frequencies. We also use Pearson's correlation (Section 2.1.1) and MIC (Section 2.1.4) to select the features for comparison purposes.

Figure 5.10 shows the prediction errors on the testing set for different methods with different numbers (1-70) of top features using the RBF-kernel SVM. It can be seen that the lowest error is from our backward elimination procedure with the RBF distance correlation. However, all the errors are relatively high ($\sim$30%). This may be due to the fact that some of the informative features did not show up in the prescreening stage by only pairwise considerations. This problem may be overcome by a strategy called *resuscitation* (See Chapter 6 for more discussion). Another explanation may be that the above tried methods do not take the specific functional structure of the feature space into account. This is confirmed by the experiments described in the next section.

5.4.4 Adapting the kernel distance covariance

Here we treat features within the same group as a "functional" feature. As described above, each of these features can actually be treated as a discrete distribution. This makes the Kullback-Leibler (KL) distance and the maximum distance between components of two vectors (supremum norm) naturally choices for the distance on the feature space. Specifically, for discrete probability distributions P and Q, the Kullback-Leibler divergence of Q from P is defined to be (see [Kullback and Leibler, 1951] for details)

$$D_{KL}(P||Q) = \sum_i \ln\Big(\frac{P(i)}{Q(i)}\Big)P(i) \tag{5.8}$$

The KL divergence defined in (5.8) is not symmetric so it is not a metric. We therefore use the following *KL distances* for our experiments.

$$D_{KLD}(P||Q) = D_{KL}(P||Q) + D_{KL}(Q||P) \tag{5.9}$$

We use the Euclidean distance on the labels as usual.

We compute the KL distance correlation for each group of the features and rank them accordingly. Figure 5.11 shows the two feature groups ranked top and bottom, respectively. It can be seen that the informative feature group has a bigger variance, and a more discrepancy between the two classes. We then use this feature group (shown in the left plot in Figure 5.11), and K nearest neighbors (K = 1, 3, ..., 15) with the KL distance (5.9) to classify the writers. An error rate of 24.6% is obtained with K=5 (compared to the error rates shown in Figure 5.10). This implies that by considering the structure of the feature space by appropriate kernels (distances) in the association measure, and using the same metric for the classifier, one can improve the classification accuracy.

5.4.5 Discussion

In this section we briefly describe an application of the kernel distance covariance to a problem with a functional feature space. The results shows that one can improve the classification performance by taking into account the specific structures of the feature space, when such information is available. In practice when such information is not available, our

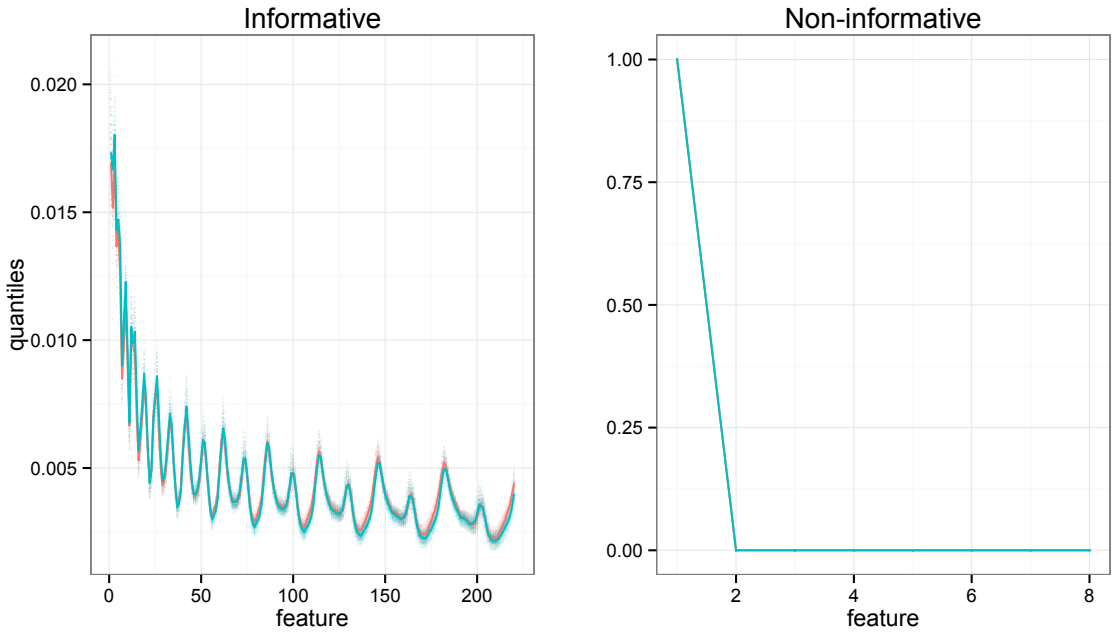


Figure 5.11: The most informative (left) and the least informative (right) feature groups in the handwriting data set according to KL distance correlation. Male is plotted with blue and female with red. For each feature and each class, 21 quantiles are plotted using dots. The connected lines show the medians of the features for male and female, respectively.

backward elimination procedure with kernel distance correlation is still a safe choice as evidence by the experiments in Section 5.4.3 (Figure 5.10).

Chapter 6

Conclusions

In this thesis, we propose a general framework for kernel-based measures of associations. Our work makes three main contributions. First, we show the connections (including some novel relationships) between several existing association measures (Chapter 2). Second, we propose a general framework for kernel-based association measures unifying existing methods and novel extensions based on kernels (Chapter 3). Kernel introduces flexibility. Third, more importantly, we show how to implement algorithms incorporating such measures, specifically, optimization and backward elimination (Chapter 4). Our general framework has been applied to a diversified set of simulations and applications with different variable types, where we have observed improved performance (Chapters 4 and 5). We demonstrate *de novo* construction of kernels tailored to the data at hand. We also provide a way for selecting informative dimensions for classification problems. In our results, we have shown that kernel association measures can adapt to the quantity of information in different dimensions of the data and determine an optimal weighting strategy. We have also shown that the distance metric used in the association measures can be coupled with that used in the classification scheme, which bridges the filtering methods and the wrapper methods in the traditional feature selection literature (Section 5.4.4). More accurate classifiers can be built this way via considering the structure of the feature space. Kernel association measures can handle different data types, including functional input and output spaces in a natural way. As two by-products, we propose a novel stepwise multiple test procedure that can save computation dramatically and (at the same time) produce interval estimates for

P-values (Section 5.3.5); we also propose a method to construct simulation-efficient shortest probability intervals (Spin, Appendix A).

Although optimization issues are not the focus of this work, the performance of the proposed algorithms may depend on the quality of the optimization method adopted. We mainly use the Newton-Raphson method for optimization in our experiments. We observe that it may be sensitive to initial values for the parameters, which should be carefully chosen in practice especially for high dimensions. One needs the gradient and the Hessian matrix of the objective function (the association measure) for such optimization. These are easy to compute analytically for simple functions, which can save computational time. Finite-difference gradients and Hessians are computed for more complex functions, as in most of our examples. Other optimization methods are available, e.g., the BFGS method [Broyden, 1970; Fletcher, 1970; Goldfarb, 1970; Shanno, 1970] and simulated annealing [Bélisle, 1992]. Those methods may be less sensitive to the starting point.

We have shown how to search a large space of variables mainly in Chapter 5. There we use two types of strategies to deal with computational issues: considering one group of variables at a time (e.g., Section 5.3), or prescreening based on say, pairwise association measures (Section 5.4.3). The latter may be sensitive to the pairwise effects of two variables, and an influential variable may not show up well in this process if it does not have a high value of the association measure when combined with another variable. This may be overcome by resuscitating variables in a later stage [Chernoff *et al.*, 2009]. Specifically, variables previously neglected are considered again using the association measure combined with variables selected by prescreening.

We use smoothed weights in kernel-based association measures to highlight informative dimensions (Sections 4.2, 4.3 and 5.3). One can also add an L_1 penalty when maximizing the association measure with free scaling factors (denoted by θ), the same penalty in the lasso method [Tibshirani, 1996]

$$\theta^* = \arg\max_{\theta}\{\mathcal{A}(k_{x,\theta_x}, k_{y,\theta_y}) + \lambda|\theta|\} \tag{6.1}$$

This will shrink some of the weights to zero, so that achieve the goal of feature selection. Efficient algorithms are available for computing the entire path of solutions for lasso [Efron *et al.*, 2004]. Similar methods may need to be developed for the problem shown in (6.1).

We show how to use the kernel-based association measures for feature selection in binary classification problems in Section 4.3. The same backward elimination approach is directly applicable to multi-class and regression problems as well. Different kernels can be defined on the corresponding output spaces. For multi-class classification, the simplest kernel would be [Song *et al.*, 2012]

$$k(y, y') = c\delta_{y,y'} \text{ where } c > 0$$

For regression, kernels such as linear and RBF can be readily used. Thus our framework is quite unified and can deal with different supervised learning problems.

Bibliography

[Agresti, 2002] A. Agresti. *Categorical Data Analysis*. Wiley Series in Probability and Statistics - Applied Probability and Statistics Section Series. Wiley, 2002.

[Al-Madeed *et al.*, 2012] Somaya Al-Madeed, Wael Ayouby, Abdelaali Hassane, and Jihad Mohamad Aljaam. Quwi: An arabic and english handwriting dataset for offline writer identification. In *ICFHR*, pages 746–751, 2012.

[Almasy *et al.*, 2011] Laura Almasy, Thomas D Dyer, Juan Manuel Peralta, Jack W Kent, Jr, Jac C Charlesworth, Joanne E Curran, and John Blangero. Genetic analysis workshop 17 mini-exome simulation. *BMC Proc*, 5 Suppl 9:S2, 2011.

[Asimit and Zeggini, 2010] Jennifer Asimit and Eleftheria Zeggini. Rare variant association analysis methods for complex traits. *Annu Rev Genet*, 44:293–308, 2010.

[Bailey-Wilson *et al.*, 2011] Joan E Bailey-Wilson, Jennifer S Brennan, Shelley B Bull, Robert Culverhouse, Yoonhee Kim, Yuan Jiang, Jeesun Jung, Qing Li, Claudia Lamina, Ying Liu, Reedik Mägi, Yue S Niu, Claire L Simpson, Libo Wang, Yildiz E Yilmaz, Heping Zhang, and Zhaogong Zhang. Regression and data mining methods for analyses of multiple rare variants in the genetic analysis workshop 17 mini-exome data. *Genet Epidemiol*, 35 Suppl 1:S92–100, 2011.

[Bandi and Srihari, 2005] Karthik R Bandi and Sargur N Srihari. Writer demographic classification using bagging and boosting. In *In Proc. 12th Int. Graphonomics Society Conference*, pages 133–137, 2005.

[Bélisle, 1992] C. J. P. Bélisle. Convergence theorems for a class of simulated annealing algorithms on $
mathbb R\hat{d}$. *Journal of Applied Probability*, pages 885–895, 1992.

[Bertinet and Agnan, 2004] A. Bertinet and Thomas C. Agnan. *Reproducing Kernel Hilbert Spaces in Probability and Statistics*. Kluwer Academic Publishers, 2004.

[Box and Tiao, 1973] George E. P Box and George C. Tiao. *Bayesian inference in statistical analysis*. Addison-Wesley series in behavioral science. Addison-Wesley Pub. Co, Reading, Mass., 1973.

[Breiman, 2001] Leo Breiman. Random forests. *Machine Learning*, 45:5–32, 2001.

[Broyden, 1970] C. G. Broyden. The Convergence of a Class of Double-rank Minimization Algorithms 1. General Considerations. *IMA Journal of Applied Mathematics*, 6(1):76–90, March 1970.

[Chapelle *et al.*, 2002] Olivier Chapelle, Vladimir Vapnik, Olivier Bousquet, and Sayan Mukherjee. Choosing multiple parameters for support vector machines. *Machine Learning*, 46(1-3):131–159, 2002.

[Chen and Shao, 1998] Ming-Hui Chen and Qi-Man Shao. Monte carlo estimation of bayesian credible and hpd intervals. *Journal of Computational and Graphical Statistics*, 8:69–92, 1998.

[Chernoff *et al.*, 2009] Herman Chernoff, Shaw-Hwa Lo, and Tian Zheng. Discovering influential variables: A method of partitions. *Annals of Applied Statistics*, 3(4):1335–1369, December 2009.

[Conover and Iman, 1976] W.J. Conover and R.L. Iman. On some alternative procedures using ranks for the analysis of experimental designs. *Communications in Statistics - Theory and Methods*, 5(14):1349–1368, 1976.

[CORNFIELD, 1951] J CORNFIELD. A method of estimating comparative rates from clinical data; applications to cancer of the lung, breast, and cervix. *J Natl Cancer Inst*, 11(6):1269–75, Jun 1951.

[Dasgupta *et al.*, 2011] Abhijit Dasgupta, Yan V Sun, Inke R König, Joan E Bailey-Wilson, and James D Malley. Brief review of regression-based and machine learning methods in genetic epidemiology: the genetic analysis workshop 17 experience. *Genet Epidemiol*, 35 Suppl 1:S5–11, 2011.

[David and Nagaraja, 2003] H. A. David and H. N. Nagaraja. *Order Statistics, Third Edition.* Wiley, New York, 2003.

[Deheuvels, 1979] P Deheuvels. La fonction de dependance empirique et ses proprietes. un test non parametrique dindependance. *Bulletin de la Classe des Sciences, V. Serie, Academie Royale de Belgique*, 65:274–292, 1979.

[Dering *et al.*, 2011] Carmen Dering, Claudia Hemmelmann, Elizabeth Pugh, and Andreas Ziegler. Statistical analysis of rare sequence variants: an overview of collapsing methods. *Genet Epidemiol*, 35 Suppl 1:S12–7, 2011.

[DEVLIN *et al.*, 1975] SJ DEVLIN, R GNANADESIKAN, and JR KETTENRING. Robust estimation and outlier detection with correlation-coefficients. *Biometrika*, 62(3):531–545, 1975.

[Efron *et al.*, 2004] Bradley Efron, Trevor Hastie, Iain Johnstone, and Robert Tibshirani. Least angle regression. *Annals of Statistics*, 32:407–499, 2004.

[Efron, 1979] B. Efron. Bootstrap methods: another look at the jackknife. *Ann. Statistics*, 7:1–26, 1979.

[Fisher, 1915] Ronald Aylmer Fisher. Frequency distribution of the values of the correlation coefficient in samples from an indefinitely large population. *Biometrika*, 10(4):507–521, 1915.

[Fletcher, 1970] R. Fletcher. A new approach to variable metric algorithms. *The Computer Journal*, 13(3):317–322, March 1970.

[Gandy, 2012] Alex Gandy. Sequential Computation of p-values based on (Re-)Sampling with a Guaranteed Error Bound. *http://arxiv.org/abs/math/0612488v1*, December 2012.

[Gelman *et al.*, 2012] A Gelman, J Hill, and M Yajima. Why we (usually) don't have to worry about multiple comparisons. *Journal of Research on Educational Effectiveness*, 5:189–211, 2012.

[Gelman, 2004] Andrew Gelman. *Bayesian data analysis*. Chapman & Hall/CRC, Boca Raton, Fla., 2nd ed edition, 2004.

[Genest *et al.*, 2007] Christian Genest, Jean-Francois Quessy, and Bruno Remillard. Asymptotic local efficiency of cramer-von mises tests for multivariate independence. *Annals of Statistics*, 35(1):166–191, February 2007.

[Goldfarb, 1970] Donald Goldfarb. A Family of Variable-Metric Methods Derived by Variational Means. *Mathematics of Computation*, 24(109):23–26, January 1970.

[Gönen and Alpaydin, 2011] Mehmet Gönen and Ethem Alpaydin. Multiple kernel learning algorithms. *Journal of Machine Learning Research*, 12:2211–2268, 2011.

[Gretton *et al.*, 2007] Arthur Gretton, Kenji Fukumizu, Choon Hui Teo, Le Song, Bernhard Schölkopf, and Alex J. Smola. A kernel statistical test of independence. In John C. Platt, Daphne Koller, Yoram Singer, and Sam T. Roweis, editors, *NIPS*. Curran Associates, Inc., 2007.

[Gretton *et al.*, 2009] Arthur Gretton, Kenji Fukumizu, and Bharath K. Discussion of: Brownian distance covariance. *The Annals of Applied Statistics*, 3:1285–1294, 2009.

[Hassane *et al.*, 2012] Abdelaali Hassane, Somaya Al-Madeed, and Ahmed Bouridane. A set of geometrical features for writer identification. In Tingwen Huang, Zhigang Zeng, Chuandong Li, and Chi-Sing Leung, editors, *ICONIP (5)*, volume 7667 of *Lecture Notes in Computer Science*, pages 584–591. Springer, 2012.

[Honeine and Richard, 2010] Paul Honeine and Cdric Richard. The angular kernel in machine learning for hyperspectral data classification. In *Proc. IEEE Workshop on Hyperspectral Image and Signal Processing : Evolution in Remote Sensing*, Reykjavik, Iceland, 14 - 16 June 2010.

[Huber, 2004] Peter J Huber. *Robust statistics*. Wiley-Interscience, Hoboken, N.J., 2004.

[Joseph *et al.*, 1995] L. Joseph, D.B. Wolfson, and R.D. Berger. Sample size calculations for binomial proportions via highest posterior density intervals. *The Statistician: Journal of the Institute of Statisticians*, 44:143–154, 1995.

[Keshava, 2004] N. Keshava. Distance metrics and band selection in hyperspectral processing with applications to material identification and spectral libraries. *Geoscience and Remote Sensing, IEEE Transactions on*, 42(7):1552–1565, July 2004.

[Knijnenburg *et al.*, 2009] Theo A. Knijnenburg, Lodewyk F. A. Wessels, Marcel J. T. Reinders, and Ilya Shmulevich. Fewer permutations, more accurate P-values. *Bioinformatics*, 25(12):i161–i168, June 2009.

[Kosorok, 2009] Michael R. Kosorok. Discussion of: Brownian distance covariance. *The Annals of Applied Statistics*, 3(4):1270–1278, December 2009.

[Kruskal and Wallis, 1952] William H. Kruskal and W. Allen Wallis. Use of ranks in One-Criterion variance analysis. *Journal of the American Statistical Association*, 47(260):583–621, 1952.

[Kullback and Leibler, 1951] S. Kullback and R. A. Leibler. On information and sufficiency. *Ann. Math. Statist.*, 22(1):79–86, 1951.

[Li *et al.*, 2009] Jialiang Li, Bee Choo Tai, and David J. Nott. Confidence interval for the bootstrap P-value and sample size calculation of the bootstrap test. *J. Nonparametric Stat.*, 21(5):649–661, 2009.

[Liao, 2005] T. Warren Liao. Clustering of time series data - a survey. *Pattern Recognition*, 38(11):1857–1874, 2005.

[Liu *et al.*, 2011] Ying Liu, Chien Hsun Huang, Inchi Hu, Shaw-Hwa Lo, and Tian Zheng. Association screening for genes with multiple potentially rare variants: an inverse-probability weighted clustering approach. *BMC Proc*, 5 Suppl 9:S106, 2011.

[Liwicki *et al.*, 2007] M. Liwicki, A. Schlapbach, P. Loretan, and H. Bunke. Automatic detection of gender and handedness from on-line handwriting. In *Proc. 13th Conf. of the Graphonomics Society*, pages 179–183, 2007.

[Lyons, 2013] Russell Lyons. Distance covariance in metric spaces. *To appear in Annals of Probability*, 2013.

[Madsen and Browning, 2009] Bo E. Madsen and Sharon R. Browning. A Groupwise Association Test for Rare Mutations Using a Weighted Sum Statistic. *PLoS Genet*, 5(2):e1000384+, February 2009.

[Marcus Liwicki, 2010] Horst Bunke Marcus Liwicki, Andreas Schlapbach. Automatic gender detection using on-line and off-line information. *Pattern Analysis and Applications (PAA)*, 1:1–6, 2010.

[M'lan *et al.*, 2007] C.É. M'lan, L. Joseph, and D.B. Wolfson. *Bayesian Sample Size Determination for Binomial Proportions*. Technical report (University of Connecticut. Dept. of Statistics). University of Connecticut, Department of Statistics, 2007.

[Newton, 2009] Michael A. Newton. Introducing the discussion paper by székely and rizzo. *The Annals of Applied Statistics*, 3(4):1233–1235, 2009.

[Phillips and Venkatasubramanian, 2011] Jeff M. Phillips and Suresh Venkatasubramanian. A gentle introduction to the kernel distance. *CoRR*, abs/1103.1625, 2011.

[Price *et al.*, 2006] Alkes L Price, Nick J Patterson, Robert M Plenge, Michael E Weinblatt, Nancy A Shadick, and David Reich. Principal components analysis corrects for stratification in genome-wide association studies. *Nat Genet*, 38(8):904–9, Aug 2006.

[Pritchard, 2001] J K Pritchard. Are rare variants responsible for susceptibility to complex diseases? *Am J Hum Genet*, 69(1):124–37, Jul 2001.

[R Development Core Team, 2009] R Development Core Team. *R: A language and environment for statistical computing*. R Foundation for Statistical Computing, Vienna, Austria, 2009. ISBN 3-900051-07-0.

[Reshef *et al.*, 2011] David N. Reshef, Yakir A. Reshef, Hilary K. Finucane, Sharon R. Grossman, Gilean McVean, Peter J. Turnbaugh, Eric S. Lander, Michael Mitzenmacher, and Pardis C. Sabeti. Detecting novel associations in large data sets. *Science*, 334(6062):1518–1524, December 2011.

[Rubin, 1981] D. B. Rubin. Estimation in parallel randomized experiments. *Journal of Educational Statistics*, 6:377–401, 1981.

[Sejdinovic *et al.*, 2012] Dino Sejdinovic, Arthur Gretton, Bharath K. Sriperumbudur, and Kenji Fukumizu. Hypothesis testing using pairwise distances and associated kernels (with appendix). *CoRR*, abs/1205.0411, 2012.

[Shanno, 1970] D. F. Shanno. Conditioning of Quasi-Newton Methods for Function Minimization. *Mathematics of Computation*, 24(111):647–656, July 1970.

[Sokal and Rohlf, 1981] Robert R Sokal and F. James Rohlf. *Biometry: the principles and practice of statistics in biological research*. W. H. Freeman, San Francisco, 2d ed edition, 1981.

[Song *et al.*, 2012] Le Song, Alex Smola, Arthur Gretton, Justin Bedo, and Karsten Borgwardt. Feature selection via dependence maximization. *J. Mach. Learn. Res.*, 98888:1393–1434, June 2012.

[Spiegelhalter *et al.*,] D. Spiegelhalter, A. Thomas, N. Best, and W. Gilks. Bugs: Bayesian inference using gibbs sampling.

[Szekely and Rizzo, 2009] Gabor J. Szekely and Maria L. Rizzo. Brownian distance covariance. *Annals of Applied Statistics*, 3(4):1236–1265, December 2009.

[Szekely *et al.*, 2007] Gabor J. Szekely, Maria L. Rizzo, and Nail K. Bakirov. Measuring and testing dependence by correlation of distances. *Annals of Statistics*, 35(6):2769–2794, December 2007.

[Tibshirani, 1996] R. Tibshirani. Regression shrinkage and selection via the lasso. *Journal of the Royal Statistical Society (Series B)*, 58:267–288, 1996.

[van der Laan and Pollard, 2003] M. J. van der Laan and K. S. Pollard. A new algorithm for hybrid hierarchical clustering with visualization and the bootstrap. *Journal of Statistical Planning and Inference*, 117(2):275–303, December 2003.

[van der Vaart, 1998] A. W. van der Vaart. *Asymptotic statistics*. Cambridge Series in Statistical and Probabilistic Mathematics. Cambridge University Press, 1998.

[Vapnik, 1998] Vladimir Naumovich Vapnik. *Statistical learning theory.* Wiley, New York, 1998.

[Vapnik, 2000] Vladimir Naumovich Vapnik. *The nature of statistical learning theory.* Springer, New York, 2nd ed edition, 2000.

[Wallenstein and Wittes, 1993] Sylvan Wallenstein and Janet Wittes. The power of the mantel-haenszel test for grouped failure time data. *Biometrics*, 49(4):pp. 1077–1087, 1993.

[Weston *et al.*, 2003] Jason Weston, André Elisseeff, Bernhard Schölkopf, and Mike Tipping. Use of the zero norm with linear models and kernel methods. *J. Mach. Learn. Res.*, 3:1439–1461, March 2003.

[Wilcox, 2005] Rand R Wilcox. *Introduction to robust estimation and hypothesis testing.* Elsevier/Academic Press, Amsterdam, 2nd ed edition, 2005.

[Wu *et al.*, 2011] Michael C Wu, Seunggeun Lee, Tianxi Cai, Yun Li, Michael Boehnke, and Xihong Lin. Rare-variant association testing for sequencing data with the sequence kernel association test. *Am J Hum Genet*, 89(1):82–93, Jul 2011.

[Zheng *et al.*, 2006a] Tian Zheng, Matthew J. Salganik, and Andrew Gelman. How many people do you know in prison?: using overdispersion in count data to estimate social structure in networks. *J. Amer. Statist. Assoc.*, 101(474):409–423, 2006.

[Zheng *et al.*, 2006b] Tian Zheng, Hui Wang, and Shaw-Hwa Lo. Backward genotype-trait association (bgta)-based dissection of complex traits in case-control designs. *Human Heredity*, 62(4):196–212, 2006.

[Zheng *et al.*, 2011] Tian Zheng, Herman Chernoff, Inchi Hu, Iuliana Ionita-Laza, and Shaw-Hwa Lo. Discovering influential variables: A general computer intensive method for common genetic disorders. In Henry Horng-Shing Lu, Bernhard Schlkopf, and Hongyu Zhao, editors, *Handbook of Statistical Bioinformatics*, Springer Handbooks of Computational Statistics, pages 87–107. Springer, 2011.

Appendix A

Efficient shortest probability intervals

Here we present the details of the Spin method discussed in Chapter 5.

Bayesian highest posterior density (HPD) intervals can be estimated directly from simulations via empirical shortest intervals. Unfortunately, these can be noisy (that is, have a high Monte Carlo error). We derive an optimal weighting strategy using bootstrap and quadratic programming to obtain a more computationally stable HPD, or in general, shortest probability interval (Spin). We prove the consistency of our method. Simulation studies on a range of theoretical and real-data examples, some with symmetric and some with asymmetric posterior densities, show that intervals constructed using Spin have better coverage (relative to the posterior distribution) and lower Monte Carlo error than empirical shortest intervals. We implement the new method in an R package (`SPIn`) so it can be routinely used in post-processing of Bayesian simulations.

A.1 Introduction

It is standard practice to summarize Bayesian inferences via posterior intervals of specified coverage (for example, 50% and 95%) for parameters and other quantities of interest. In the modern era in which posterior distributions are computed via simulation, we most commonly see central intervals: the $100(1-\alpha)\%$ central interval is defined by the $\frac{\alpha}{2}$ and $1-\frac{\alpha}{2}$ quantiles.

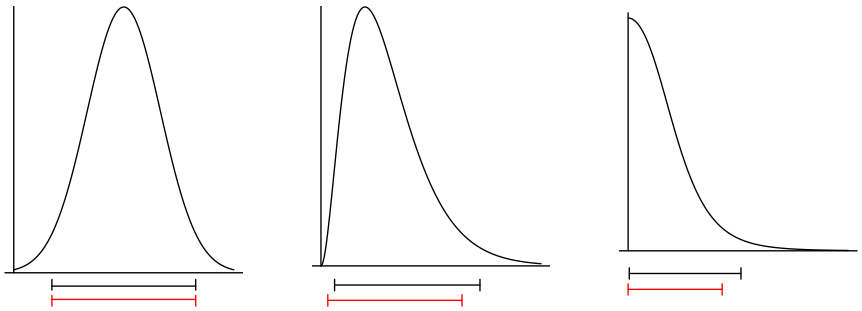


Figure A.1: Simple examples of central (black) and highest probability density (red) intervals. The intervals coincide for a symmetric distribution; otherwise the HPD interval is shorter. The three examples are a normal distribution, a gamma with shape parameter 3, and the marginal posterior density for a variance parameter in a hierarchical model.

Highest-posterior density (HPD) intervals (recommended, for example, in the classic book of [Box and Tiao, 1973]) are easily determined for models with closed-form distributions such as the normal and gamma but are more difficult to compute from simulations.

We would like to go back to the HPD, solving whatever computational problems necessary to get it to work. Why? Because for an asymmetric distribution, the HPD interval can be a more reasonable summary than the central probability interval. Figure A.1 shows these two types of intervals for three distributions: for symmetric densities (as shown in the left panel in Figure A.1), central and HPD intervals coincide; whereas for the two examples of asymmetric densities (the middle and right panels in Figure A.1), HPDs are shorter than central intervals (in fact, the HPD is the shortest interval containing a specified probability).

In particular, when the highest density occurs at the boundary (the right panel in Figure A.1), we strongly prefer the shortest probability interval to the central interval; the HPD covers the highest density part of the distribution and also the mode. In such cases, central intervals can be much longer and have the awkward property at cutting off a narrow high-posterior slice that happens to be near the boundary, thus ruling out a part of the distribution that is actually strongly supported by the inference.

One concern with highest posterior density intervals is that they depend on parameterization. For example, the left endpoint of the HPD in the right panel of Figure A.1

is 0, but the interval on the logarithmic scale does not start at $-\infty$. Interval estimation is always conditional on the purposes to which the estimate will be used. Beyond this, univariate summaries cannot completely capture multivariate relationships. Thus all this work is within the context of routine data analysis (e.g., [Spiegelhalter *et al.*,]) in which interval estimates are a useful way to summarize inferences about parameters and quantities of interest in a model in understandable parameterizations. We do not attempt a conclusive justification of HPD intervals here; we merely note that in the pre-simulation era such intervals were considered the standard, which suggests to us that the current preference for central intervals arises from computational reasons as much as anything else.

For the goal of computing an HPD interval from posterior simulations, the most direct approach is the *empirical shortest probability interval*, the shortest interval of specified probability coverage based on the simulations [Chen and Shao, 1998]. For example, to obtain a 95% interval from a posterior sample of size n, you can order the simulation draws and then take the shortest interval that contains $0.95n$ of the draws. This procedure is easy, fast, and simulation-consistent (that is, as $n \to \infty$ it converges to the actual HPD interval assuming that the HPD interval exists and is unique). The only trouble with the empirical shortest probability interval is that it can be too noisy, with a high Monte Carlo error (compared to the central probability interval) when computed from the equivalent of a small number of simulation draws. This is a concern with current Bayesian methods that rely on Markov chain Monte Carlo (MCMC) techniques, where for some problems the effective sample size of the posterior draws can be low (for example, hundreds of thousands of steps might be needed to obtain an effective sample size of 500).

Figure A.2 shows the lengths of the empirical shortest 95% intervals based on several simulations for the three distributions shown in Figure A.1, starting from the kth order statistic. For each distribution and each specified number of independent simulation draws, we carried out 200 replications to get a sense of the typical size of the Monte Carlo error. The lengths of the 95% intervals are highly variable when the number of simulation draws is small.

In this section, we develop a quadratic programming strategy coupled with bootstrapping to estimate the endpoints of the shortest probability interval. Simulation studies show

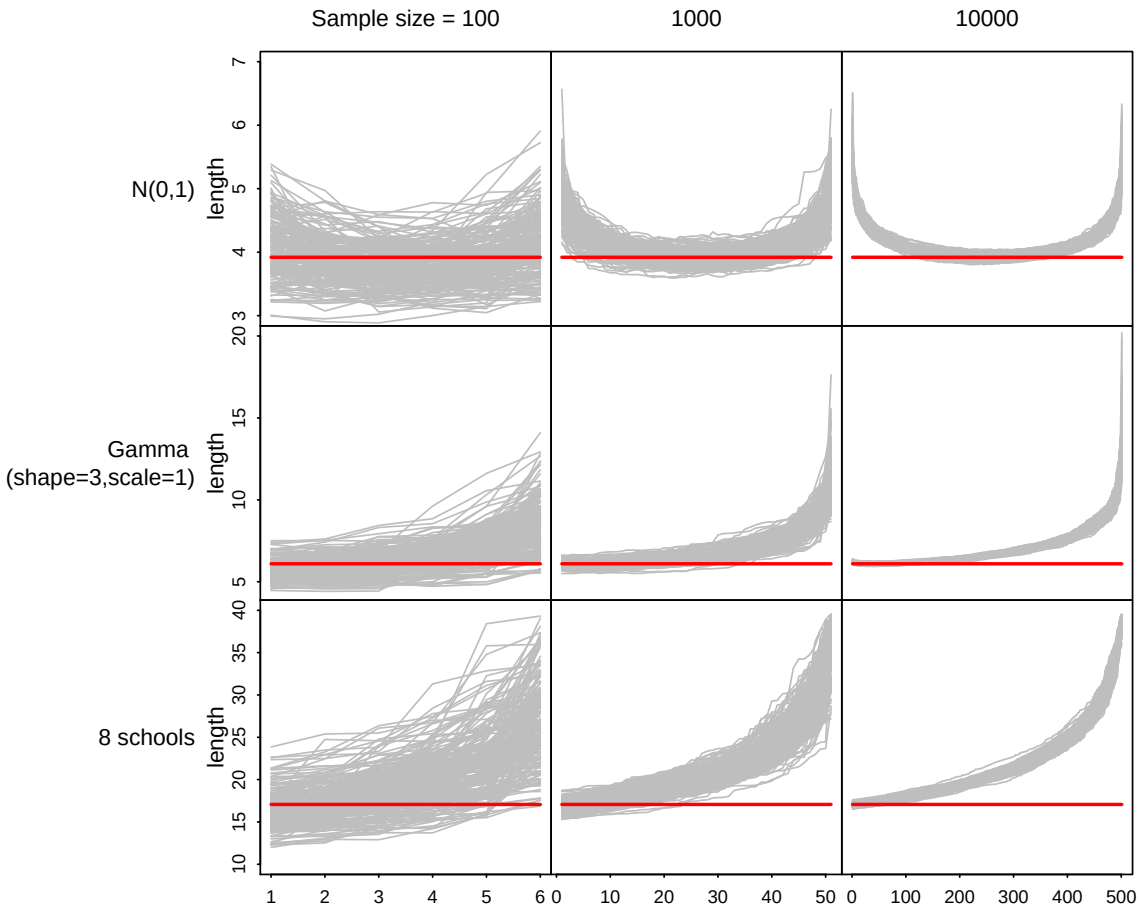


Figure A.2: Lengths of 95% empirical probability intervals from several simulations for each of three models. Each gray curve shows interval length as a function of the order statistic of the interval's lower endpoint; thus, the minimum value of the curve corresponds to the empirical shortest 95% interval. For the (symmetric) normal example, the empirical shortest interval is typically close to the central interval (for example, with a sample of size 1000, it is typically near $(x_{(26)}, x_{(975)})$). The gamma and eight-schools examples are skewed with a peak near the left of the distribution, hence the empirical shortest intervals are typically at the left end of the scale. The red lines show the lengths of the true shortest 95% probability interval for each distribution. The empirical shortest interval approaches the true value as the number of simulation draws increases but is noisy when the number of simulation draws is small, hence motivating a more elaborate estimator.

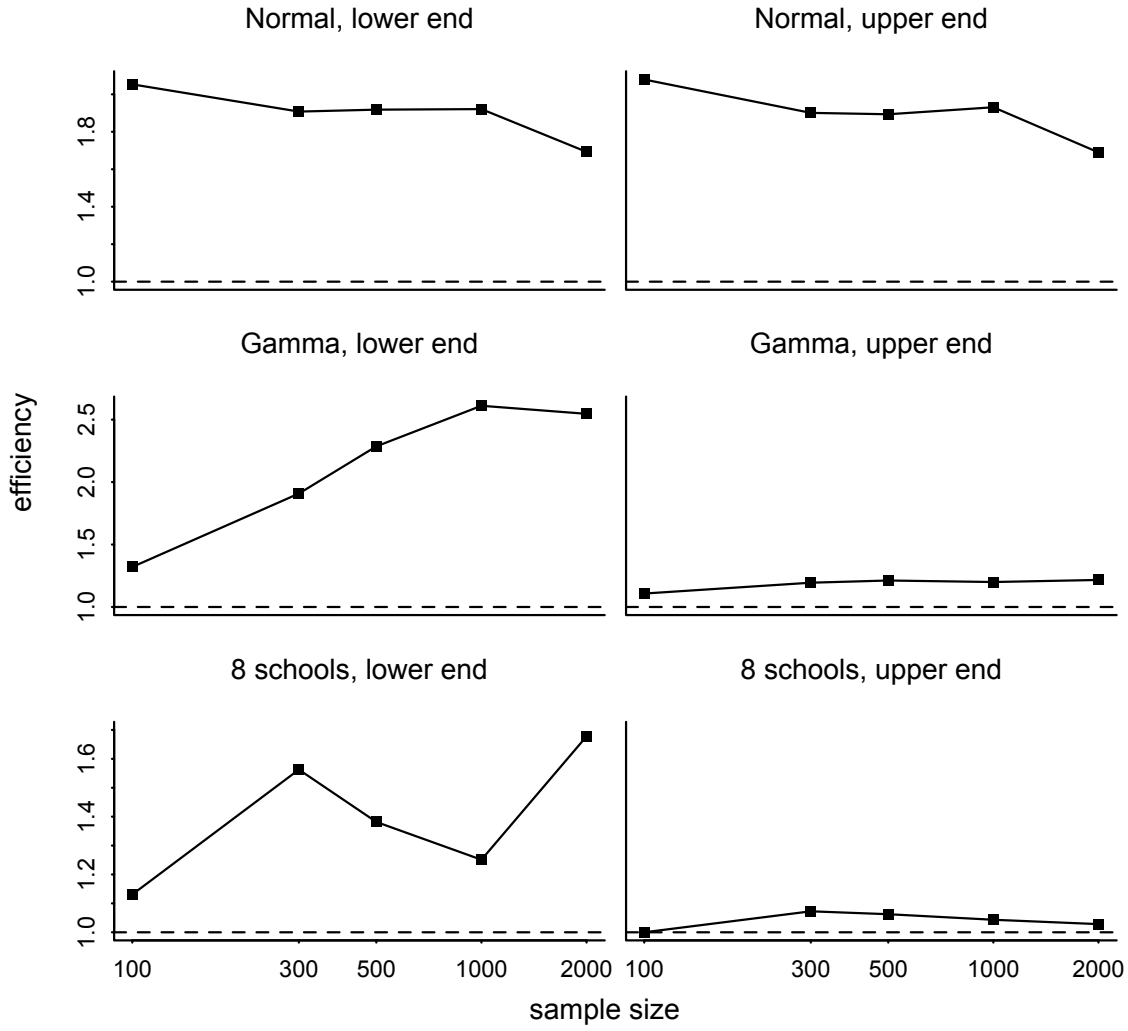


Figure A.3: Efficiency of Spin for 95% shortest intervals for the three distributions shown in Figure A.1. For the eight-schools example, Spin is compared to a modified empirical HPD that includes the zero point in the simulations. The efficiency is always greater than 1, indicating that Spin always outperforms the empirical HPD. The jagged appearance of some of the lines may arise from discreteness in the order statistics for the 95% interval.

that our procedure, which we call Spin, results in more stable endpoint estimates compared to the empirical shortest interval (Figure A.3). Specifically, define the efficiency as

$$\text{efficiency} = \frac{\text{MSE(empirical shortest interval)}}{\text{MSE(Spin)}},$$

so that an efficiency greater than 1 means that Spin is more efficient. We show in Figure A.3 that, in all cases that we experimented on, Spin is more efficient than the competition. We derive our method in Section A.2, apply it to some theoretical examples in Section A.3 and in two real-data Bayesian analysis problems in Section A.4. We have implemented our algorithm as `SPIn`, a publicly available package in R [R Development Core Team, 2009].

A.2 Methods

A.2.1 Problem setup

Let $X_1,\ldots, X_n \overset{iid}{\sim} F$, where F is a continuous unimodal distribution. The goal is to estimate the $100(1-\alpha)\%$ shortest probability interval for F. Denote the true shortest probability interval by $(l(\alpha), u(\alpha))$. Define $G = 1 - F$, so that $F(l(\alpha)) + G(u(\alpha)) = \alpha$.

To estimate the interval, for $0 \leq \Delta \leq \alpha$, find Δ such that $G^{-1}(\alpha - \Delta) - F^{-1}(\Delta)$ is a minimum, i.e.,

$$\Delta^* = \text{argmin}_{\Delta\in[0,\alpha]}\{G^{-1}(\alpha - \Delta) - F^{-1}(\Delta)\}.$$

Taking the derivative,

$$\frac{\partial}{\partial\Delta}[(1-F)^{-1}(\alpha-\Delta) - F^{-1}(\Delta)] = 0,$$

we get

$$\frac{1}{f(G^{-1}(\alpha-\Delta))} - \frac{1}{f(F^{-1}(\Delta))} = 0, \tag{A.1}$$

where f is the probability density function of X. The minimum can only be attained at solutions to (A.1), or $\Delta = 0$ or α (Figure A.4). It can easily be shown that if $f'(x) \neq 0$ a.e., the solution to (A.1) exists and is unique. Then

$$\begin{aligned} l(\alpha) &= F^{-1}(\Delta^*), \\ u(\alpha) &= G^{-1}(\alpha - \Delta^*). \end{aligned}$$

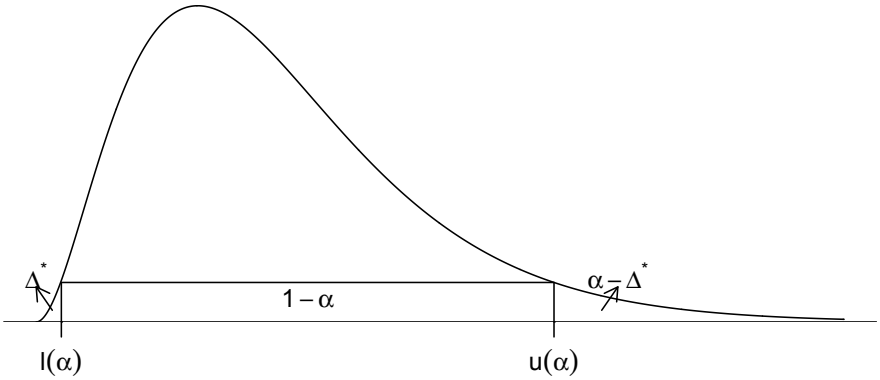


Figure A.4: Notation for shortest probability intervals.

Taking the lower end for example, we are interested in a weighting strategy such that $\hat{l} = \sum_{i=1}^{n} w_i X_{(i)}$ (where $\sum w_i = 1$) has the minimum mean squared error (MSE), $\mathrm{E}\left(\left|\left|\sum_{i=1}^{n} w_i X_{(i)} - l(\alpha)\right|\right|^2\right)$. It can also be helpful to calculate $\mathrm{MSE}(X_{([n\Delta^*])}) = \mathrm{E}\left(||X_{([n\Delta^*])} - l(\alpha)||^2\right)$. In practice we estimate Δ^* by $\hat{\Delta}$ such that

$$\hat{\Delta} = \mathrm{argmin}_{\Delta \in [0,\alpha]} \{\hat{G}^{-1}(\alpha - \Delta) - \hat{F}^{-1}(\Delta)\}, \tag{A.2}$$

where $\hat{F}$ represents the empirical distribution and $\hat{G} = 1 - \hat{F}$. This yields the widely used empirical shortest interval, which can have a high Monte Carlo error (as illustrated in Figure A.2). We will denote its endpoints by l^* and u^*. The corresponding MSE for the lower endpoint is $\mathrm{E}(||X_{([n\hat{\Delta}])} - l(\alpha)||^2)$.

A.2.2 Quadratic programming

Let $\hat{l} = \sum_{i=1}^{n} w_i X_{(i)}$. Then

$$\begin{aligned}
\mathrm{MSE}(\hat{l}) &= \mathrm{E}(\hat{l} - F^{-1}(\Delta^*))^2 \\
&= \mathrm{E}\,(\hat{l} - \mathrm{E}\,\hat{l} + \mathrm{E}\,\hat{l} - F^{-1}(\Delta^*))^2 \\
&= \mathrm{E}\,(\hat{l} - \mathrm{E}\,\hat{l})^2 + (\mathrm{E}\,\hat{l} - F^{-1}(\Delta^*))^2 \\
&= \mathrm{Var} + \mathrm{Bias}^2,
\end{aligned}$$

where $\mathrm{E}(\hat{l}) = \sum_{i=1}^{n} w_i \mathrm{E}X_{(i)}$ and $\mathrm{Var} = \sum_{i=1}^{n} w_i^2 \mathrm{Var}X_{(i)} + 2\sum_{i<j} w_i w_j \mathrm{cov}(X_{(i)}, X_{(j)})$. It has been shown (e.g., [David and Nagaraja, 2003]) that

$$\mathrm{E}(X_{(i)}) = Q_i + \frac{p_i q_i}{2(n+2)} Q_i'' + o(n^{-1}),$$

where $q_i = 1 - p_i$, $Q = F^{-1}$ is the quantile function, $Q_i = Q(p_i) = Q(\mathrm{E}U_{(i)}) = Q(\frac{i}{n+1})$, and $Q_i'' = \frac{Q_i}{f^2(Q_i)}$. Thus

$$\mathrm{E}(\hat{l}) \doteq \sum_{i=1}^{n} w_i \left(Q_i + \frac{p_i q_i}{2(n+2)} Q_i'' \right). \tag{A.3}$$

It has also been shown (e.g., [David and Nagaraja, 2003]) that

$$\begin{aligned} \mathrm{Var}\, X_{(i)} &= \frac{p_i q_i}{n+2} Q_i'^2 + o(n^{-1}) \\ \mathrm{cov}(X_{(i)}, X_{(j)}) &= \frac{p_i q_j}{n+2} Q_i' Q_j' + o(n^{-1}), \text{ for } i < j, \end{aligned}$$

where $Q_i' = \frac{1}{dp_i/dQ_i} = \frac{1}{f(Q_i)}$ ($f(Q_i)$ is called the density-quantile function). Thus,

$$\mathrm{Var}(\hat{l}) = \sum_{i=1}^{n} w_i^2 \frac{p_i q_i}{n+2} Q_i'^2 + 2\sum_{i<j} w_i w_j \frac{p_i q_j}{n+2} Q_i' Q_j' + o(n^{-1}). \tag{A.4}$$

Putting (A.3) and (A.4) together yields,

$$\begin{aligned} \mathrm{MSE}(\hat{l}) &= \sum_{i=1}^{n} w_i^2 \frac{p_i q_i}{n+2} Q_i'^2 + 2\sum_{i<j} w_i w_j \frac{p_i q_j}{n+2} Q_i' Q_j' + \\ &\quad + \left[\sum_{i=1}^{n} w_i (Q_i + \frac{p_i q_i}{2(n+2)} Q_i'') - Q(\Delta^*) \right]^2 + o(n^{-1}). \end{aligned} \tag{A.5}$$

Finding the optimal weights that minimize MSE as defined in (A.5) is then approximately a quadratic programming problem.

In this study we impose triangle kernels centered around the endpoints of the empirical shortest interval on the weights for computational stability. Specifically, the estimate of the lower endpoint has the form,

$$\hat{l} = \sum\nolimits_{i=i^*-b/2}^{i^*+b/2} w_i X_{(i)},$$

where i^* is the index of the endpoint of the empirical shortest interval, b is the bandwidth in terms of data points, and w_i decreases linearly when i moves away from i^*. We choose

b to be of order $\sqrt{n}$ in this study. This optimization problem is equivalent to minimizing MSE with the following constraints:

$$\begin{aligned}
\sum_{i=i^*-b/2}^{i^*+b/2} w_i &= 1 \\
\frac{w_i - w_{i-1}}{X_{(i)} - X_{(i-1)}} &= \frac{w_{i-1} - w_{i-2}}{X_{(i-1)} - X_{(i-2)}} \text{ for } i = i^* - b/2+2, \ldots, i^*, i^*+2, \ldots, i^*+b/2 \\
\frac{w_{i^*} - w_{i^*-1}}{X_{(i^*)} - X_{(i^*-1)}} &= \frac{w_{i^*} - w_{i^*+1}}{X_{(i^*+1)} - X_{(i^*)}} \\
w_{i^*-b/2} &\geq 0 \\
w_{i^*+b/2} &\geq 0 \\
w_{i^*} - w_{i^*+1} &\geq 0. \qquad \text{(A.6)}
\end{aligned}$$

The above constraints reflect the piecewise linear and symmetric pattern of the kernel. In practice, Q, f, and Δ^* can be substituted by the corresponding sample estimates $\hat{Q}$, $\hat{f}$, and $\hat{\Delta}$.

The above quadratic programming problem can be rewritten in the conventional matrix form,

$$\text{MSE}(\hat{l}) \doteq \frac{1}{2} w^T \mathbf{D} w - d^T w,$$

where

$$w = (w_{i^*-b/2}, \ldots, w_{i^*+b/2})^T,$$

and $\mathbf{D} = (d_{ij})$ is a symmetric matrix with

$$d_{ij} = \begin{cases} 2(Q_i^2 + \frac{p_i q_i}{n+2} Q_i'^2), & i = j \\ 2(\frac{Q_i' Q_j'}{n+2} p_i q_j + Q_i Q_j), & i < j, \end{cases}$$

$$d^T = 2Q(\Delta^*) Q_i,$$

subject to

$$\mathbf{A}^T w \geq w_0,$$

with appropriate $\mathbf{A}$ and w_0 derived from the linear constraints in (A.6).

A.2.3 Proof of simulation-consistency of the estimated HPD

The following result ensures the simulation-consistency of our endpoint estimators when we use the empirical distribution and kernel density estimate.

Under regularity conditions, with probability 1,

$$\lim_{n\to\infty} \min_{w} \left(\frac{1}{2} w^T \hat{\mathbf{D}}_n w - \hat{d}_n^T w \right) = \min_{w} \left(\frac{1}{2} w^T \mathbf{D} w - d^T w \right),$$

where $\hat{\mathbf{D}}_n$ and $\hat{d}_n$ are empirical estimates of $\mathbf{D}$ and d based on empirical distribution function and kernel density estimates.

To see this, we first show that $\hat{\mathbf{D}}_n \to \mathbf{D}$ and $\hat{d}_n \to d$ uniformly as $n \to \infty$ almost surely. By the Glivenko-Cantelli theorem, $||\hat{F} - F||_\infty \overset{a.s.}{\to} 0$, which implies $\hat{Q} \rightsquigarrow Q$ almost surely, where $\rightsquigarrow$ denotes weak convergence, i.e., $\hat{Q}(t) \to Q(t)$ at every t where Q is continuous (e.g., [van der Vaart, 1998]). It has also been shown that $\int \mathrm{E}_f(\hat{f}(x) - f(x))^2 dx = O(n^{-4/5})$ under regularity conditions (see, e.g., [van der Vaart, 1998]), which implies that $\hat{f}(x) \to f(x)$ almost surely for all x. The endpoints of the empirical shortest interval are simulation-consistent [Chen and Shao, 1998].

The elements in matrix $\hat{\mathbf{D}}_n$ result from simple arithmetic manipulations of $\hat{Q}$ and $\hat{f}$, so $\hat{d}_{ij} \to d_{ij}$ with probability 1, which implies,

$$\hat{\mathbf{D}}_n \to \mathbf{D} \text{ uniformly and almost surely,}$$

given $\mathbf{D}$ is of finite dimension. We can prove the almost sure uniform convergence of $\hat{d}_n$ to d in a similar manner.

The optimization problem $\min_w(\frac{1}{2} w^T \hat{\mathbf{D}}_n w - \hat{d}_n^T w)$ corresponds to calculating the smallest eigenvalue of an augmented matrix constructed from $\hat{\mathbf{D}}_n$ and $\hat{d}_n$. The above uniform convergence then implies,

$$\lim_{n\to\infty} \min_{w}(w^T \hat{\mathbf{D}}_n w - \hat{d}_n^T w) = \min_{w}(w^T \mathbf{D} w - d^T w).$$

The same proof works for the upper endpoint.

A.2.4 Bootstrapping the procedure to get a smoother estimate

Results from quadratic programming as described above show that, as expected, Spin has a much reduced bias than the empirical shortest intervals. This is because the above procedure

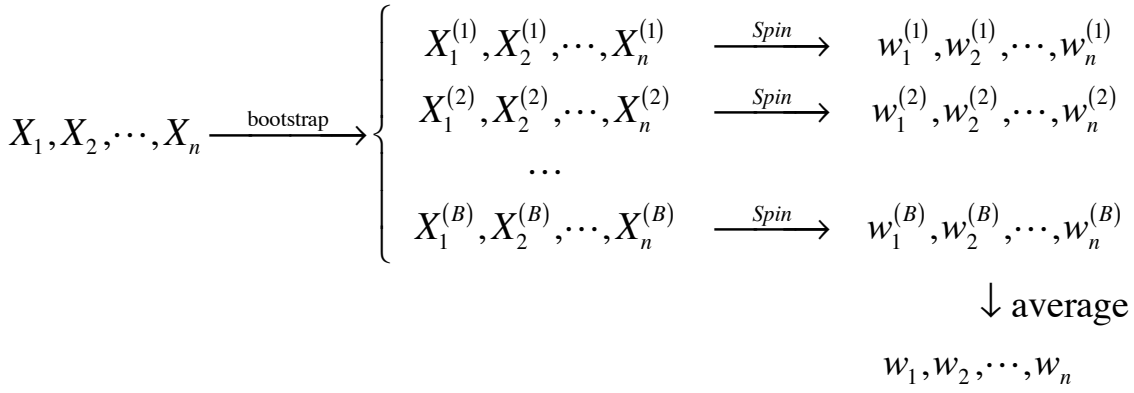


Figure A.5: Bootstrapping procedure to get more stable weights.

takes the shape of the empirical distribution into account. However, the variance remains at the same magnitude as that of the empirical shortest interval (as we shall see in the left panel in Figure A.10), because the optimal weights derived from the empirical distribution are also subject to the same level of variability as the empirical shortest intervals. We can use the bootstrap [Efron, 1979] to smooth away some of this noise and thus further reduce the variance in the interval. Specifically, we bootstrap the original posterior draws B times (in this study we set $B = 50$) and calculate the Spin optimal weights for each of the bootstrapped samples. Here, we treat the weights as general functions of the posterior distribution under study rather than the endpoints of HPD interval of the posterior samples. We then compute the final weights as the average of the B sets of weights obtained from the above procedure (Figure A.5).

A.2.5 Bounded distributions

As defined so far, our procedure necessarily yields an interval within the range of the simulations. This is undesirable if the distribution is bounded with the boundary included in the HPD interval (as in the right graph in Figure 1). To allow boundary estimates, we augment our simulations with a pseudo-datapoint (or two, if the distribution is known to be bounded on both sides). For example, if a distribution is defined on $(0, \infty)$ then we insert another datapoint at 0; if the probability space is $(0, 1)$, we insert additional points at 0

and 1.

A.2.6 Discrete and multimodal distributions

If a distribution is continuous and unimodal, the highest posterior density region and shortest probability interval coincide. More generally, the highest posterior density region can be formed from disjoint intervals. For distributions with known boundary of disjoint parts, Spin can be applied to different regions separately and a HPD region can be assembled using the derived disjoint intervals. When the nature of the underlying true distribution is unknown and the sample size is small, the inference of unimodality can be difficult. Therefore, in this study, we have focused on estimating the shortest probability interval, recognizing that, as with interval estimates in general, our procedure is less relevant for multimodal distributions.

A.3 Results for simple theoretical examples

We conduct simulation studies to evaluate the performance of our methods. We generate independent samples from the normal, t(5), and gamma(3) distributions and construct 95% intervals using these samples. We consider sample sizes of 100, 300, 500, 1000 and 2000. For each setup, we generate 20,000 independent replicates and use these to compute root mean squared errors (RMSEs) for upper and lower endpoints. We also construct empirical shortest intervals as defined in (A.2), parametric intervals and central intervals for comparison. For parametric intervals, we calculate the sample mean and standard deviation. For the normal distribution, the interval takes the form of mean $\pm$ 1.96 sd (for the t distribution we also implement the same form as "Gaussian approximation" for comparison); for the gamma, we use the mean and standard deviation to estimate its parameters first, and then numerically obtain the HPD interval using the resulted density with the two estimates plugged in. The empirical 95% central interval is defined as the 2.5%th and 97.5%th percentiles of the sampled data points. We also use our methods to construct optimal central intervals (see Section A.5) for the two symmetric distributions.

Figure A.6 shows the intervals constructed for the standard normal distribution and the

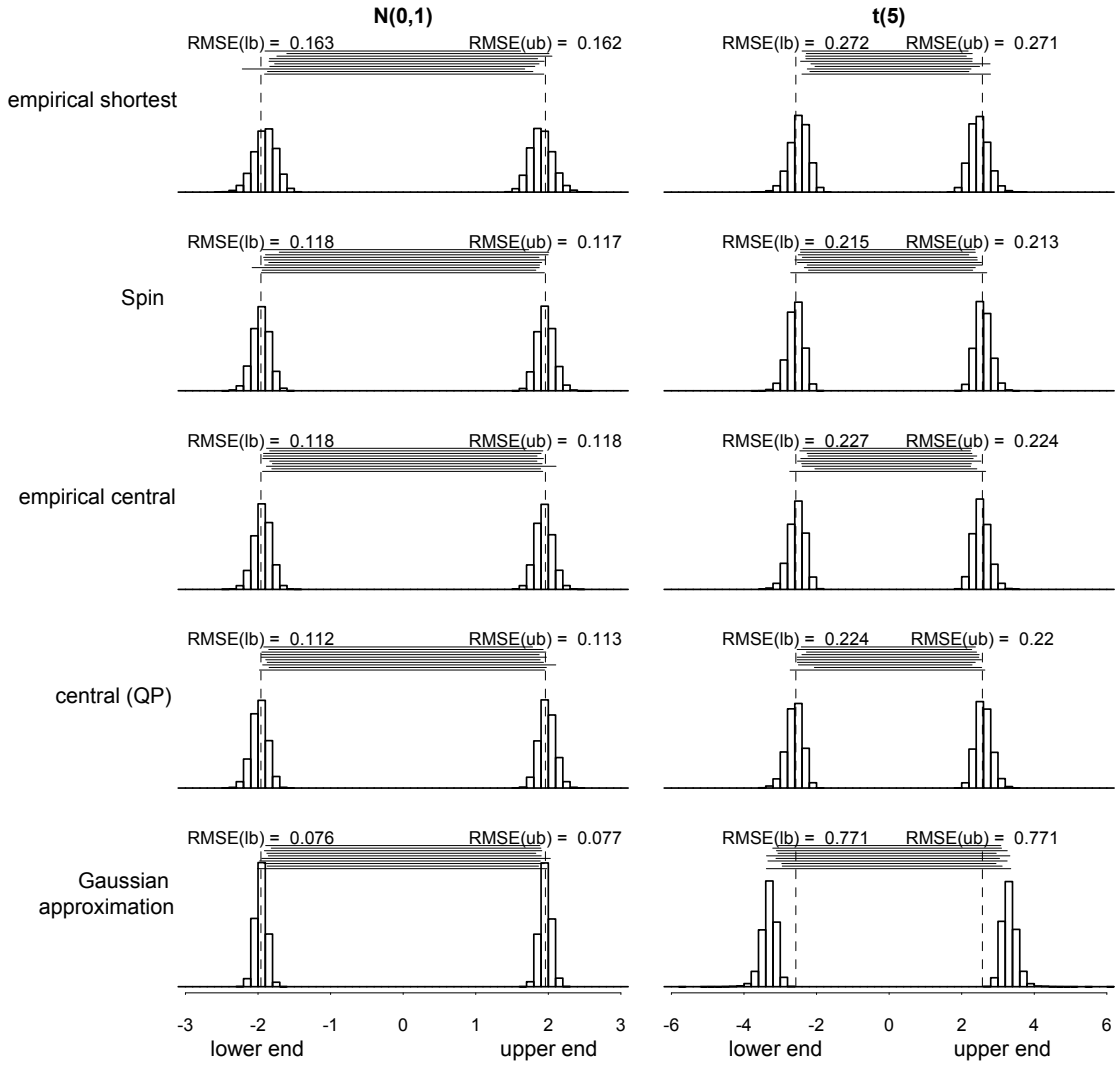


Figure A.6: Spin for symmetric distributions: 95% intervals for the normal and t(5) distributions, in each case based on 500 independent draws. Each horizontal bar represents an interval from one simulation. The histograms of the lower ends and the upper ends are based on results from 20,000 simulations. The dotted vertical lines represent the true endpoints of the HPD intervals. Spin greatly outperforms the raw empirical shortest interval. The central interval (and its quadratic programming improvement) does even better for the Gaussian but is worse for the t(5) and in any case does not generalize to asymmetric distributions. The intervals estimated by fitting a Gaussian distribution do the best for the normal model but are disastrous when the model is wrong.

t(5) distribution based on 500 simulation draws. The empirical shortest intervals tend to be too short in both cases, while Spins have better endpoint estimates. Empirical central intervals are more stable than empirical shortest intervals, and Spins have comparable RMSE for $N(0, 1)$ and smaller RMSE for t(5). Our methods can further improve RMSE based on the empirical central intervals as shown in the "central (QP)" row in Figure A.6. The RMSE is the smallest if one specifies the correct parametric distribution and uses that information to construct interval estimates, while in practice the underlying distribution is usually not totally known, and mis-specifying it can result in far-off intervals (the right bottom panel in Figure A.6).

Figure A.7 shows the empirical shortest, Spin, and parametric intervals constructed from 500 samples of the gamma distribution with shape parameter 3. Spin gets more accurate endpoint estimates than empirical shortest intervals. Specifically, for the lower end where the density is relatively high, Spin estimates are less variable, and for the upper end at the tail of the density, Spin shows a smaller bias. Again, the lowest RMSE comes from taking advantage of the parametric form of the posterior distribution, which is rarely practical in real MCMC applications. Hence the RMSE using the parametric form represents a rough lower bound on the Monte Carlo error in any HPD computed from simulations.

Figure A.8 shows the intervals constructed for MCMC normal samples. Specifically, the Gibbs sampler is used to draw samples from a standard bivariate normal distribution with correlation 0.9. We use this example to explore how Spin works on simulations with high autocorrelation. Two chains each with 1000 samples are drawn with Gibbs sampling. For one variable, every ten draws are recorded for Spin construction, resulting in 200 samples, which is roughly the level of the effective sample size in this case. This is a typical senario in practice when MCMC techniques are adopted for multivariate distributions. Again Spin greatly outperforms the empirical shortest interval in case of highly correlated draws.

We further investigate coverage probabilities of the different intervals constructed (Figure A.9). Empirical shortest intervals have the lowest coverage probability, which is as expected since they are biased towards shorter intervals (see Figure A.6 and Figure A.7). Coverage probabilities of Spin are closer to the nominal coverage (95%) for both normal and gamma distributions. Comparable coverage is observed for central intervals. As expected,

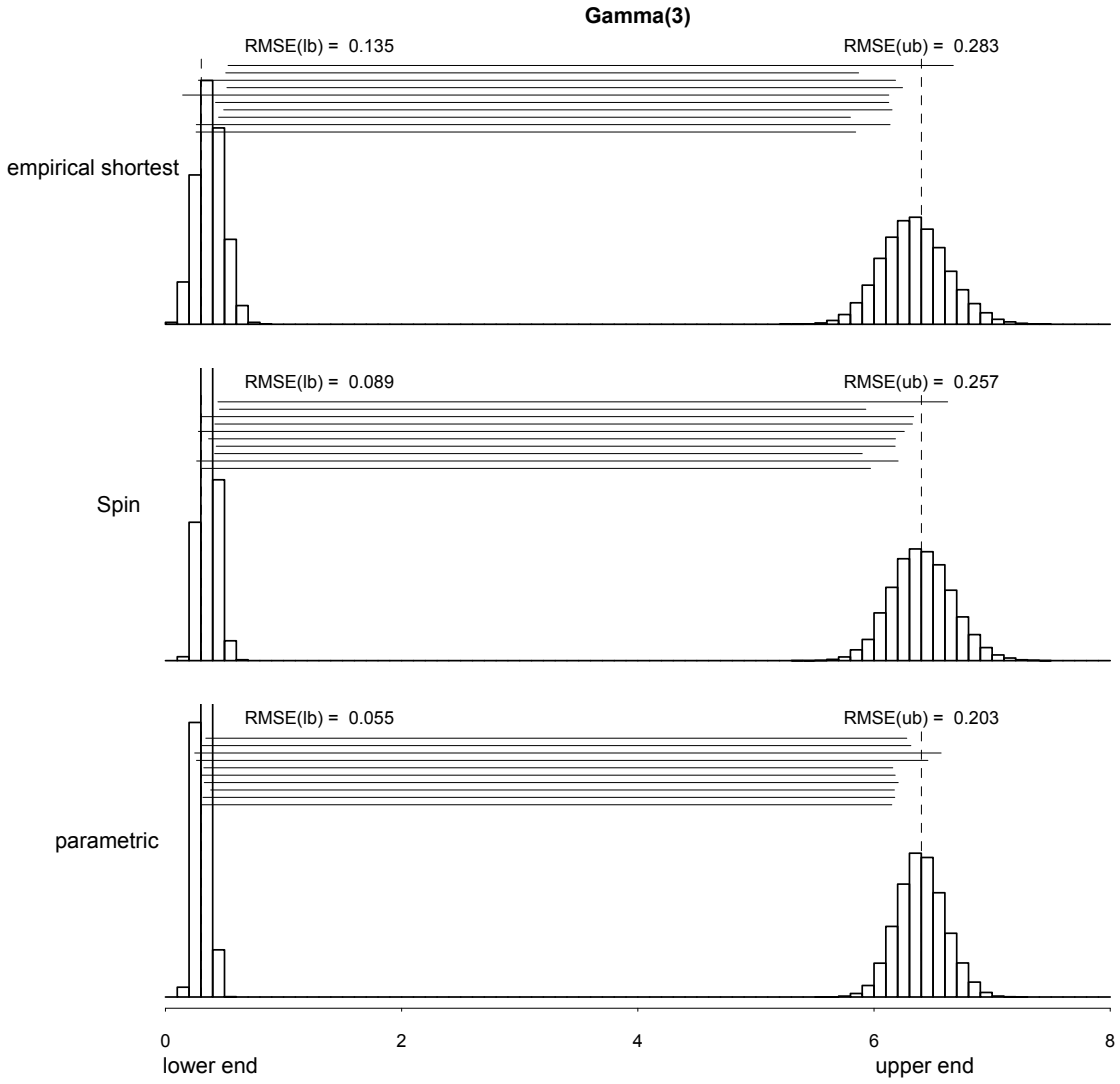


Figure A.7: Spin for an asymmetric distribution. 95% intervals for the gamma distributions with shape parameter 3, as estimated from 500 independent draws. Each horizontal bar represents an interval from one simulation. The histograms are based on results from 20,000 simulations. The dotted vertical lines represent the true endpoints of the HPD interval. Spin outperforms the empirical shortest interval. The interval obtained from a parametric fit is even better but this approach cannot be applied in general; rather, it represents an optimality bound for any method.

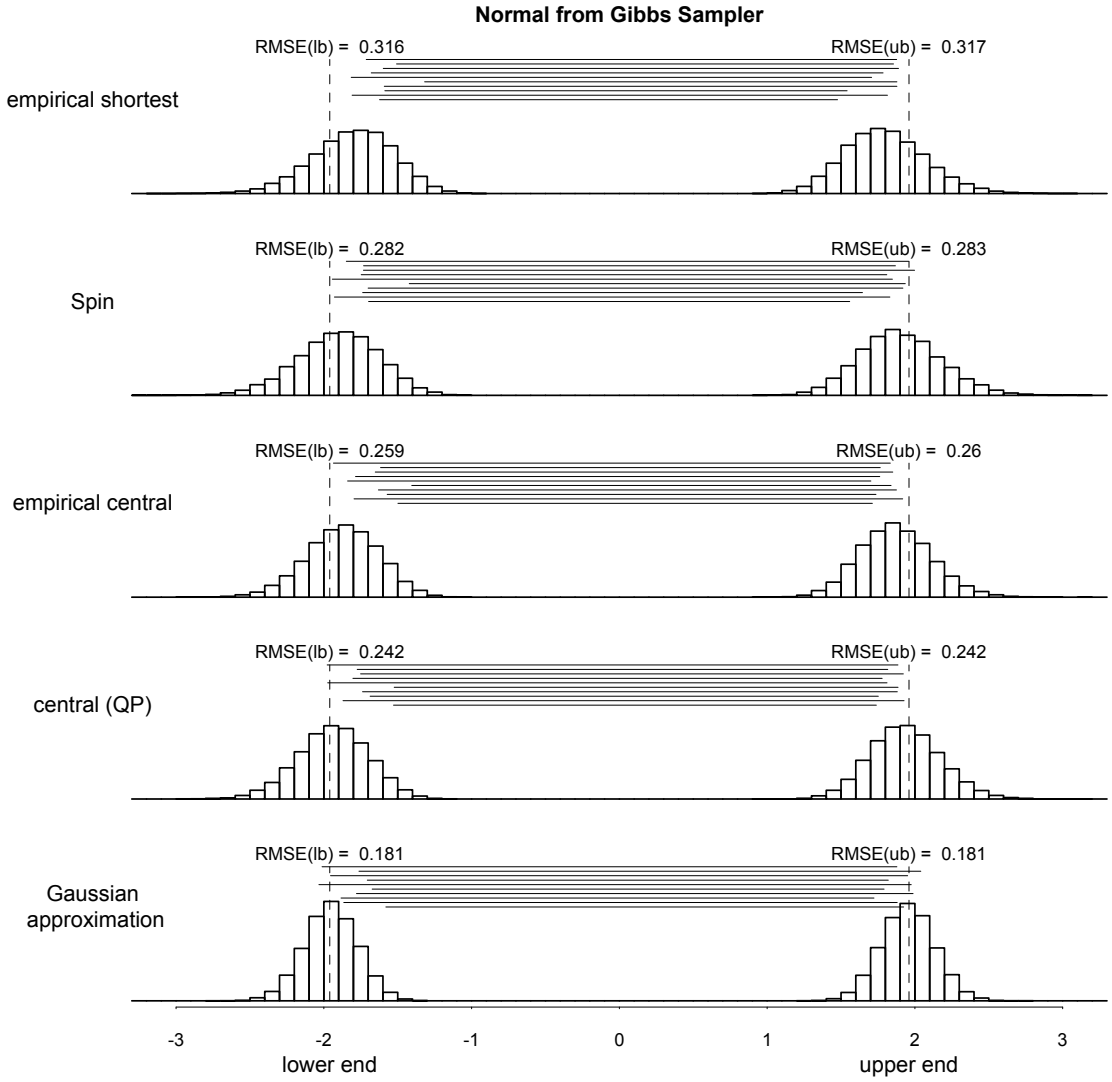


Figure A.8: Spin for MCMC samples. 95% intervals for normal samples from Gibbs sampler, in each case based on 200 draws. Each horizontal bar represents an interval from one simulation. The histograms are based on results from 20,000 simulations. The dotted vertical lines represent the true endpoints of the HPD intervals. Spin greatly outperforms the raw empirical shortest interval. The central interval (and its quadratic programming improvement) does even better. Again the intervals estimated by fitting a Gaussian distribution do the best.

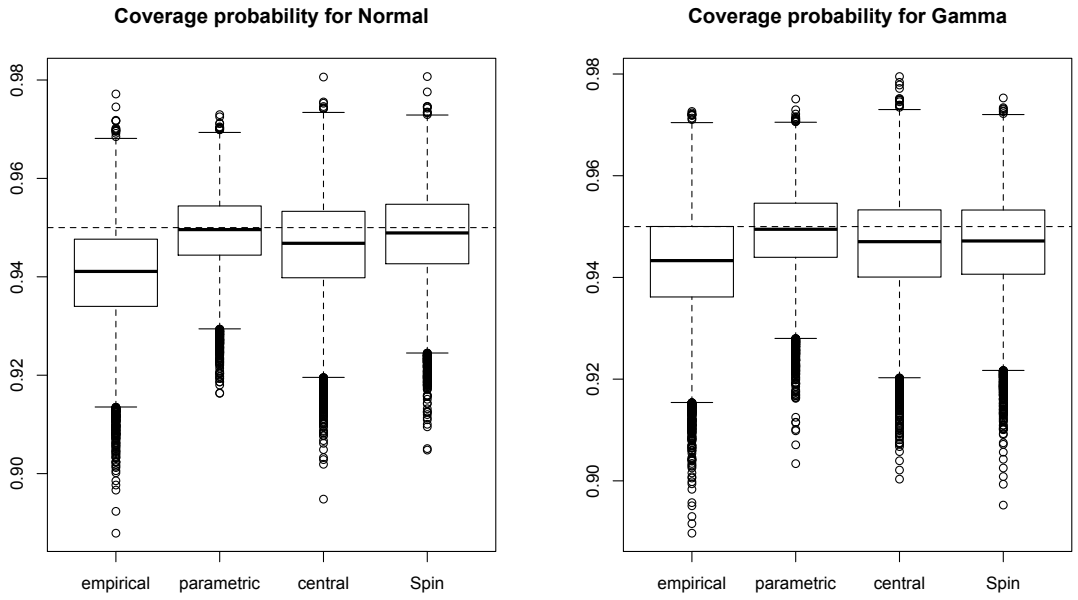


Figure A.9: Distribution of coverage probabilities for Spin and other 95% intervals calculated based on 500 simulations for the normal and gamma(3) distributions.

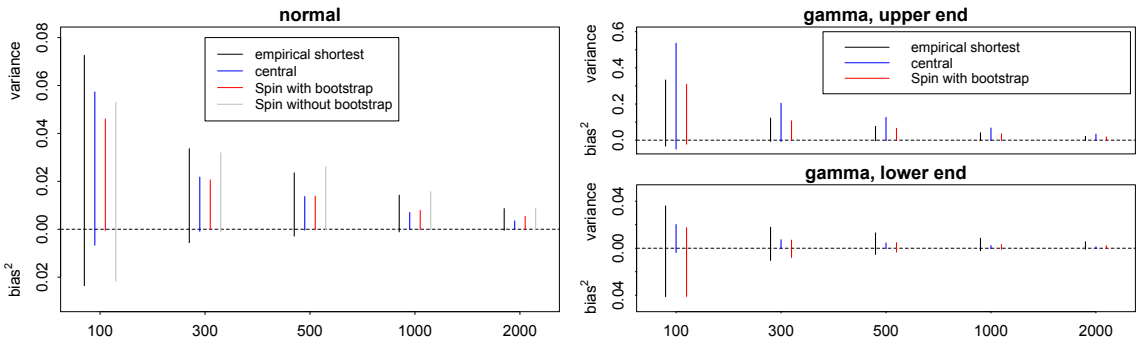


Figure A.10: Bias-variance decomposition for 95% intervals for normal and gamma(3) examples, as a function of the number of simulation draws. Because of the symmetry of the normal distribution, we averaged its errors for upper and lower endpoints. Results from Spin without bootstrap are shown for normal for description purpose.

parametric intervals represent a gold standard and have the most accurate coverage.

Figure A.10 shows the bias-variance decomposition of different interval estimates for normal and gamma distributions under sample sizes 100, 300, 500, 1,000 and 2,000. We average lower and upper ends for the normal case due to symmetry. For both distributions, Spin has both well-reduced variance and bias compared to the empirical shortest intervals. The upper end estimates of empirical central intervals for the gamma have a large variance since the corresponding density is low so the observed simulations in this region are more variable. It is worth pointing out that the computational time for Spin is negligible compared to sampling, thus it is a more efficient way to obtain improved interval estimates. In the normal example shown in the left panel in Figure A.10, rather than increasing the sample size from 300 to 500 to reduce error, one can spend less time to compute Spin with the 300 samples and get an even better interval.

We also carried out experiments with even bigger samples and intervals of other coverages (90% and 50%), and got similar results. Spin beats the empirical shortest interval in RMSE (which makes sense, given that Spin is optimizing over a class of estimators that includes the empirical shortest as a special case).

A.4 Results for two real-data examples

In this section, we apply our methods to two applied examples of hierarchical Bayesian models, one from education and one from sociology. In the first example, we show the advantages of Spin over central and empirical shortest intervals; in the second example, we demonstrate the routine use of Spin to summarize posterior inferences.

Our first example is a Bayesian analysis from [Rubin, 1981] of a hierarchical model of data from a set of experiments performed on eight schools. The group-level scale parameter (which corresponds to the between-school standard deviation of the underlying treatment effects) has a posterior distribution that is asymmetric with a mode at zero (as shown in the right panel of Figure A.1). Central probability intervals for this scale parameter (as presented, for example, in the analysis of these data by [Gelman, 2004]) are unsatisfying in that they exclude a small segment near zero where the posterior distribution is in fact largest. Figure A.11 shows the 95% empirical shortest intervals and Spin constructed from 500 draws. The results of empirical shortest intervals for 8 schools are from including the zero point in the simulations. Spin has smaller RMSE than both empirical shortest and central intervals (Figure A.11 and Figure A.12).

For our final example, we fit the social network model of [Zheng *et al.*, 2006a] using MCMC and construct 95% Spins for the overdispersion parameters based on 200 posterior draws. The posterior is asymmetric and bounded below at 1. Figure A.13 is a partial replot of Figure 4 from [Zheng *et al.*, 2006a] with Spins added. For this type of asymmetric posterior we prefer the estimated HPDs to the corresponding central intervals as HPDs more precisely capture the values of the parameter that are supported by the posterior distribution.

A.5 Discussion

We have presented a novel optimal approach for constructing reduced-error shortest probability intervals (Spin). Simulation studies and real data examples show the advantage of Spin over the empirical shortest interval. Another commonly used interval estimate in Bayesian inference is the central interval. For symmetric distributions, central intervals and

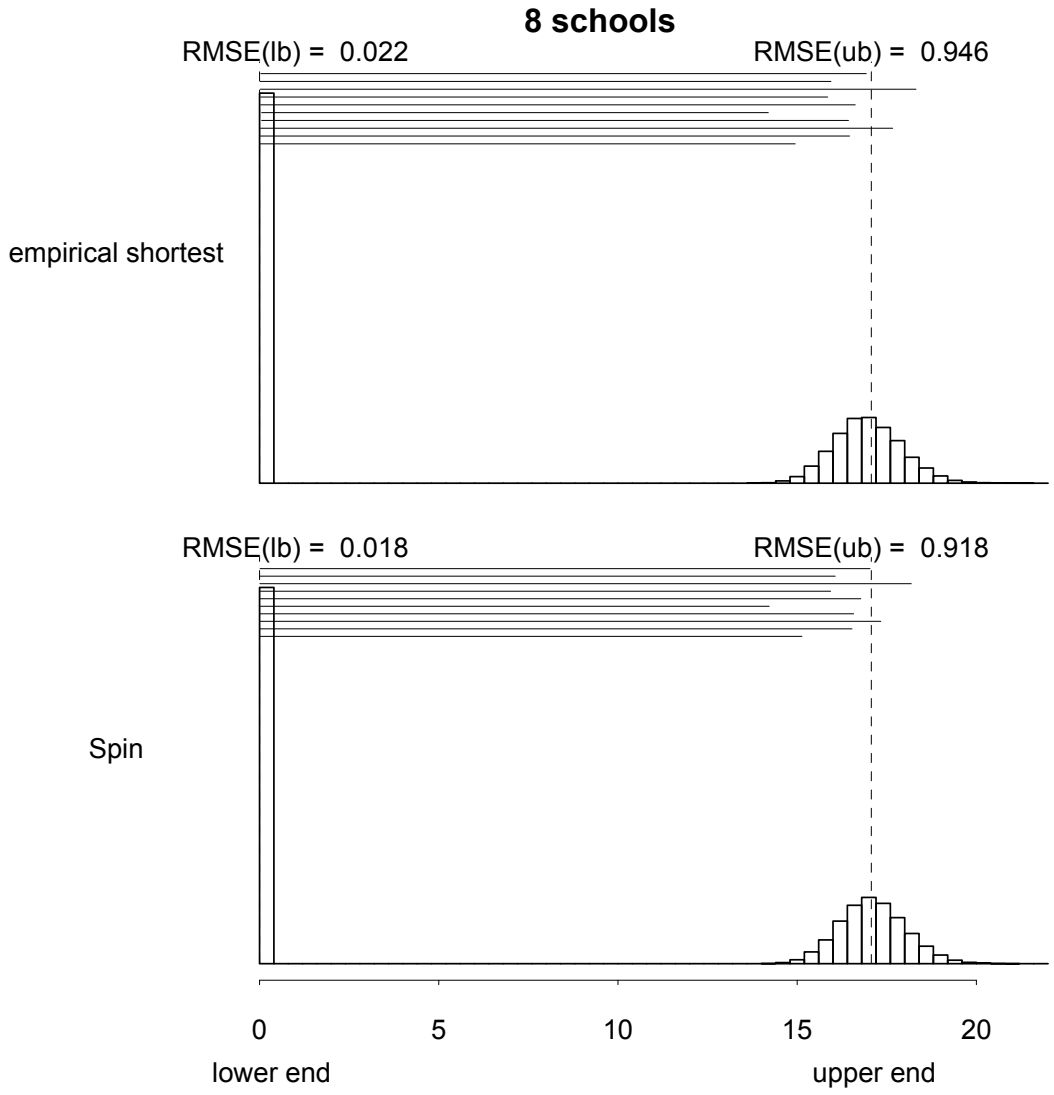


Figure A.11: Spin for the group-level standard deviation parameter in the eight schools example, as estimated from 500 independent draws from the posterior distribution (which is the right density curve in Figure A.1, a distribution that is constrained to be nonnegative and has a minimum at zero). The histograms in this figure are based on results from 20,000 simulations. The dotted vertical lines represent the true endpoints of the HPD interval as calculated numerically from the posterior density. Spin does better than the empirical shortest interval, especially at the left end, where its smoothing tends to (correctly) pull the lower bound of the interval all the way to the boundary at 0.

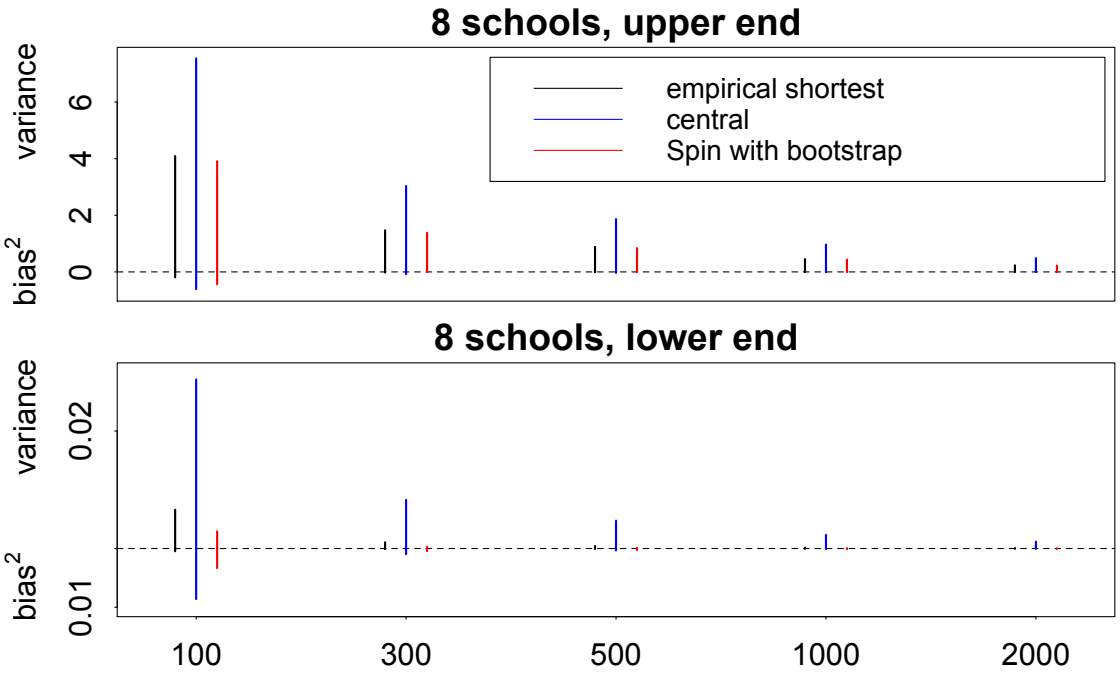


Figure A.12: Bias-variance decomposition for 95% intervals for the eight-school example, as a function of the number of simulation draws.

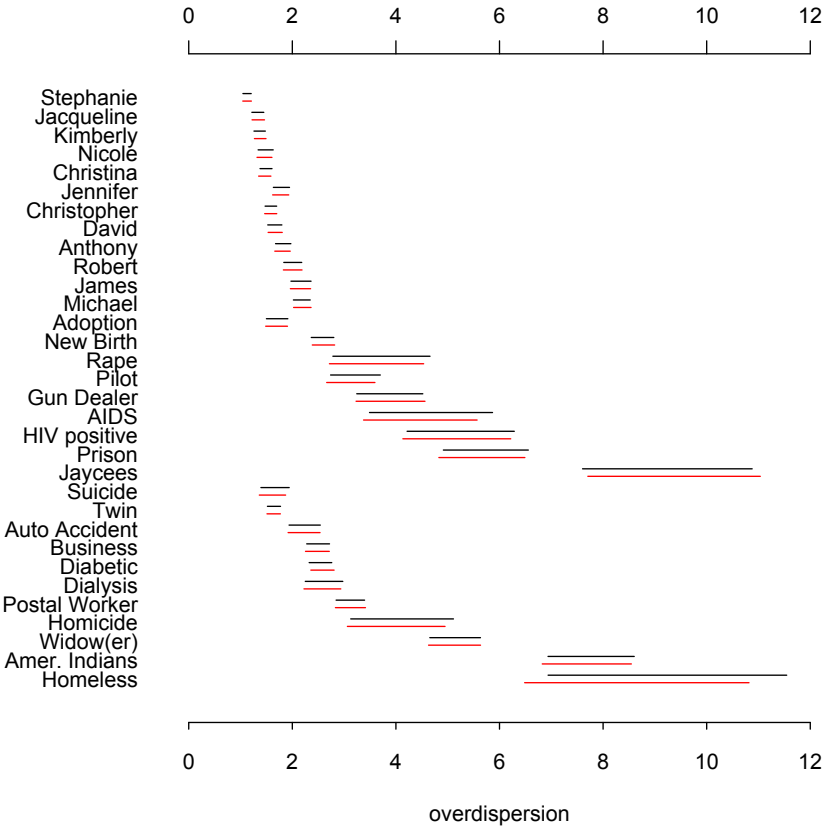


Figure A.13: 95% central intervals (black lines) and Spins (red lines) for the overdispersion parameters in the "How many X's do you know?" study. The parameter in each row is a measure of the social clustering of a certain group in the general population: groups of people identified by first names have low overdispersion and are close to randomly distributed in the social network, whereas categories such as airline pilots or American Indians are more overdispersed (that is, non-randomly distributed). We prefer the Spins as providing better summaries of these highly skewed posterior distributions. However, the differences between central intervals and Spins are not large; our real point here is not that the Spins are much better but that they will work just fine in routine applied Bayesian practice, satisfying the same needs as were served by central intervals but without that annoying behavior when distributions are highly asymmetric.

HPDs are the same; otherwise we agree with [Box and Tiao, 1973] that the HPD is generally preferable to the central interval as an inferential summary (Figure A.1). In our examples we have found that for symmetric distributions Spin and empirical central intervals have comparable RMSEs and coverage probabilities (Figures A.6, A.9, and A.10). Therefore we recommend Spin as a default procedure for computing HPD intervals from simulations, as it is as computationally stable as the central intervals which are currently standard in practice.

We set the bandwidth parameter b in (A.6) to $\sqrt{n}$, which seems to work well for a variety of distributions. We also carried out sensitivity analysis by varying b and found that large b tends to result in more stable endpoint estimates where the density is relatively high but can lead to noisy estimates where the density is low. This makes sense: in low-density regions, adding more points to the weighted average may introduce noise instead of true signals. Based on our experiments, we believe the default value $b = \sqrt{n}$ is a safe general choice.

Our approach can be considered more generally as a method of using weighted averages of order statistics to construct optimal interval estimates. One can replace $Q(\Delta^*)$ in (A.5) by the endpoints of any reasonable empirical interval estimates, and obtain improved intervals by using our quadratic programming strategy (such as the improved central intervals shown in Figure A.6).

One concern that arises is the computational cost of performing Spin itself. Our simulations show Spin intervals to have better simulation coverage and appreciably lower mean squared error compared to the empirical HPD, but for simple problems in which one can quickly draw direct posterior simulations, it could be simpler to forget Spin and instead just double the size of the posterior sample. More and more, though, we find ourselves computing Bayesian models using elaborate Markov chain simulation algorithms for which it can take hundreds of thousands of steps, and hours or even days of computing time, to obtain an effective sample size of a few hundred posterior simulation draws. In this case, the Spin calculations are a small price to pay for obtaining more accurate and stable HPD intervals.

We have demonstrated that our Spin procedure works well in a range of theoretical and

applied problems, that it is simulation-consistent, computationally feasible, addresses the boundary problem, and is optimal within a certain class of procedures that include the empirical shortest interval as a special case. We do not claim, however, that the procedure is optimal in any universal sense. We see the key contribution of the present work as developing a practical procedure to compute shortest probability intervals from simulation in a way that is superior to the naive approach and is competitive (in terms of simulation variability) with central probability intervals. Now that Spin can be computed routinely, we anticipate further research improvements on posterior summaries.

Appendix B

Two-vs-one dimensional association patterns

In this appendix we show the other six two-vs-one dimensional patterns tested in Chapter 4.

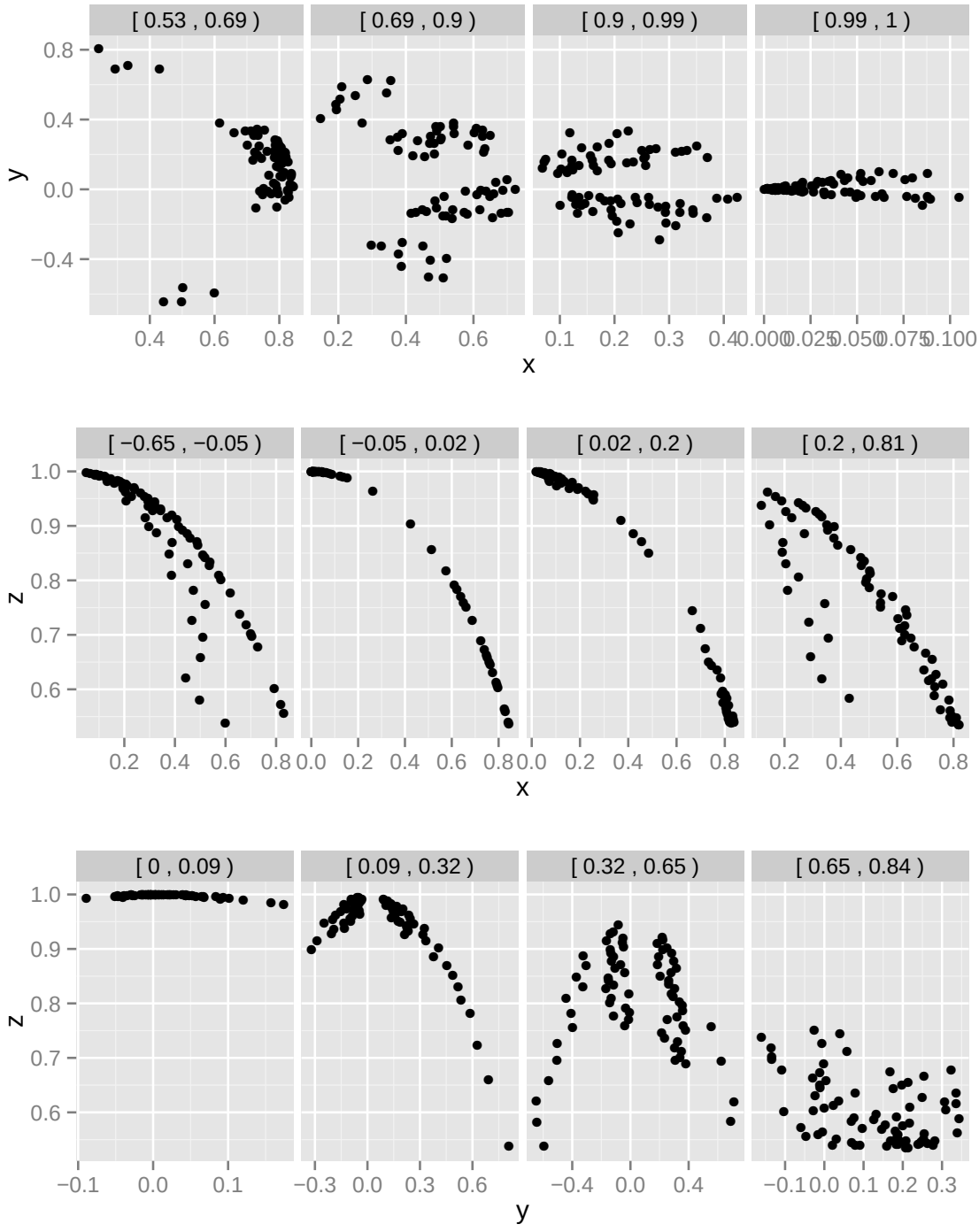


Figure B.1: Slice plots of the first association pattern in Figure 4.1.

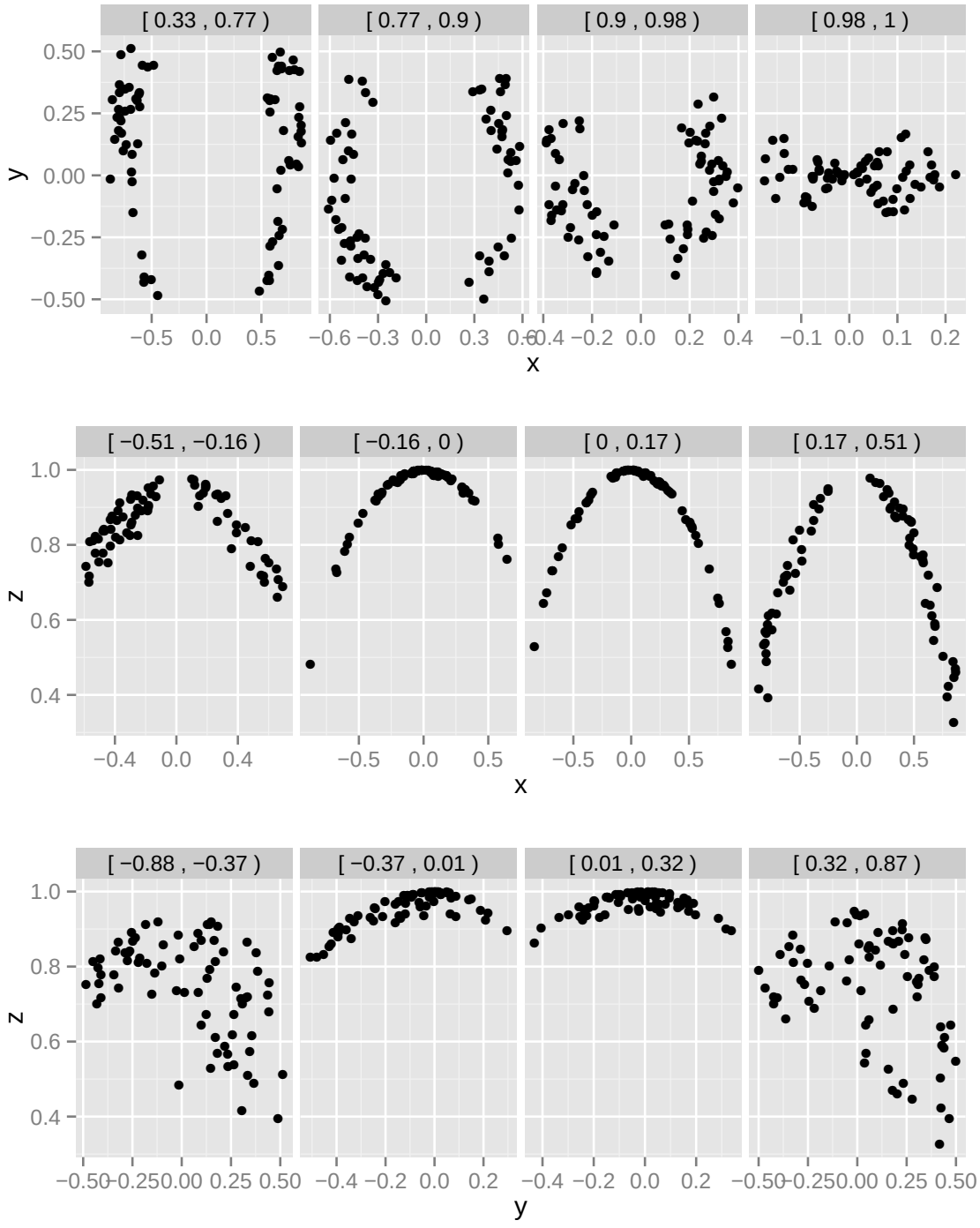


Figure B.2: Slice plots of the second association pattern in Figure 4.1.

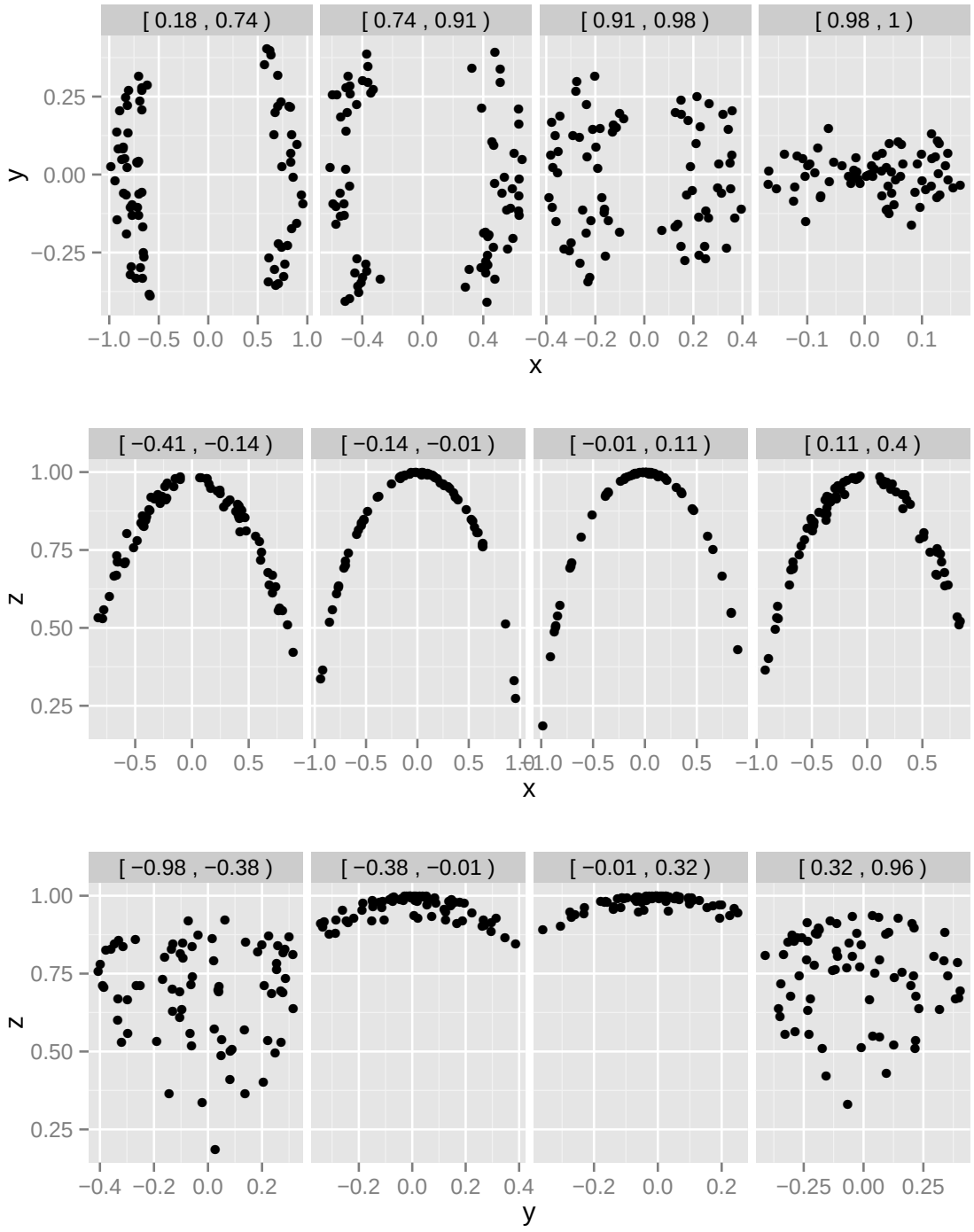


Figure B.3: Slice plots of the third association pattern in Figure 4.1.

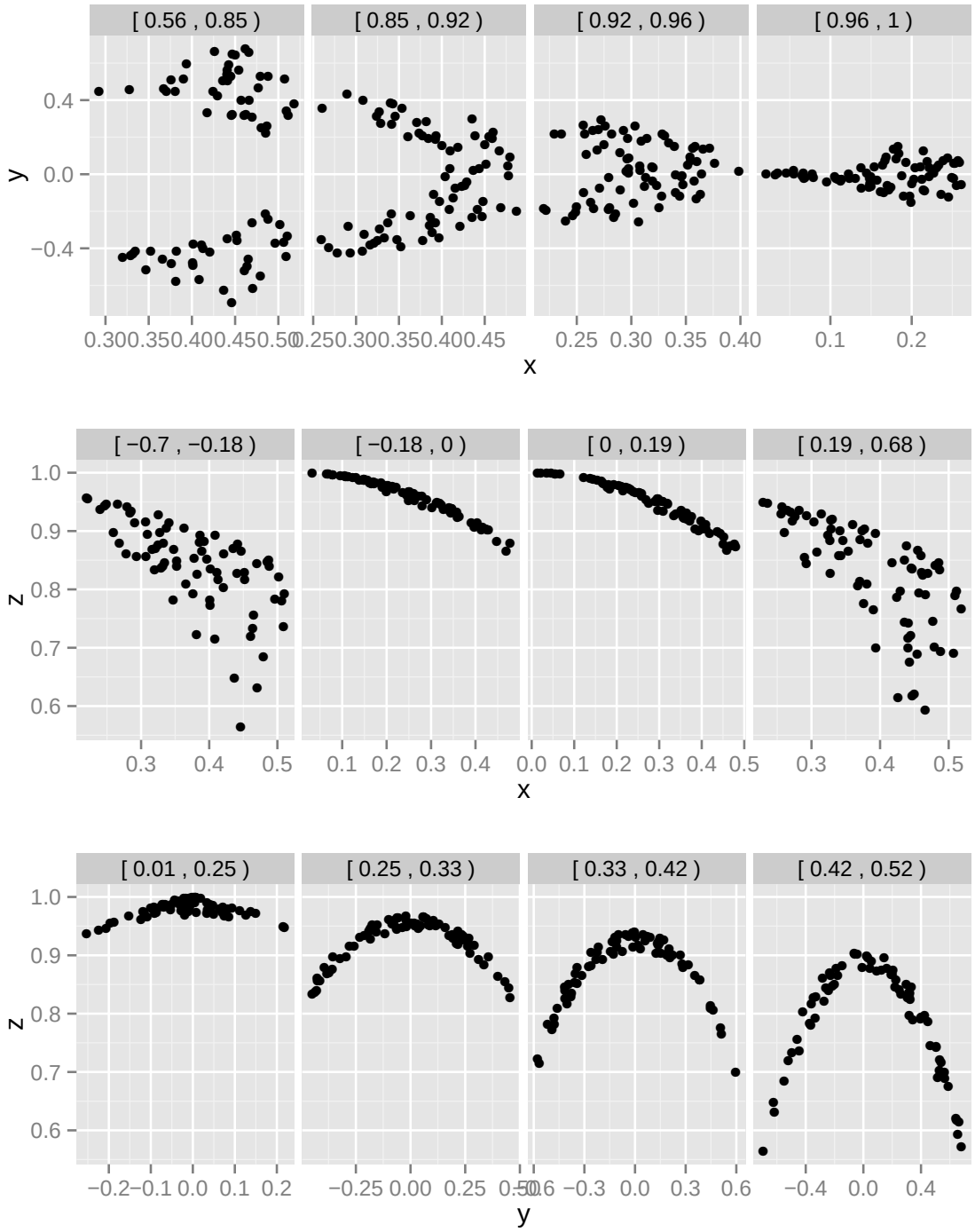


Figure B.4: Slice plots of the fourth association pattern in Figure 4.1.

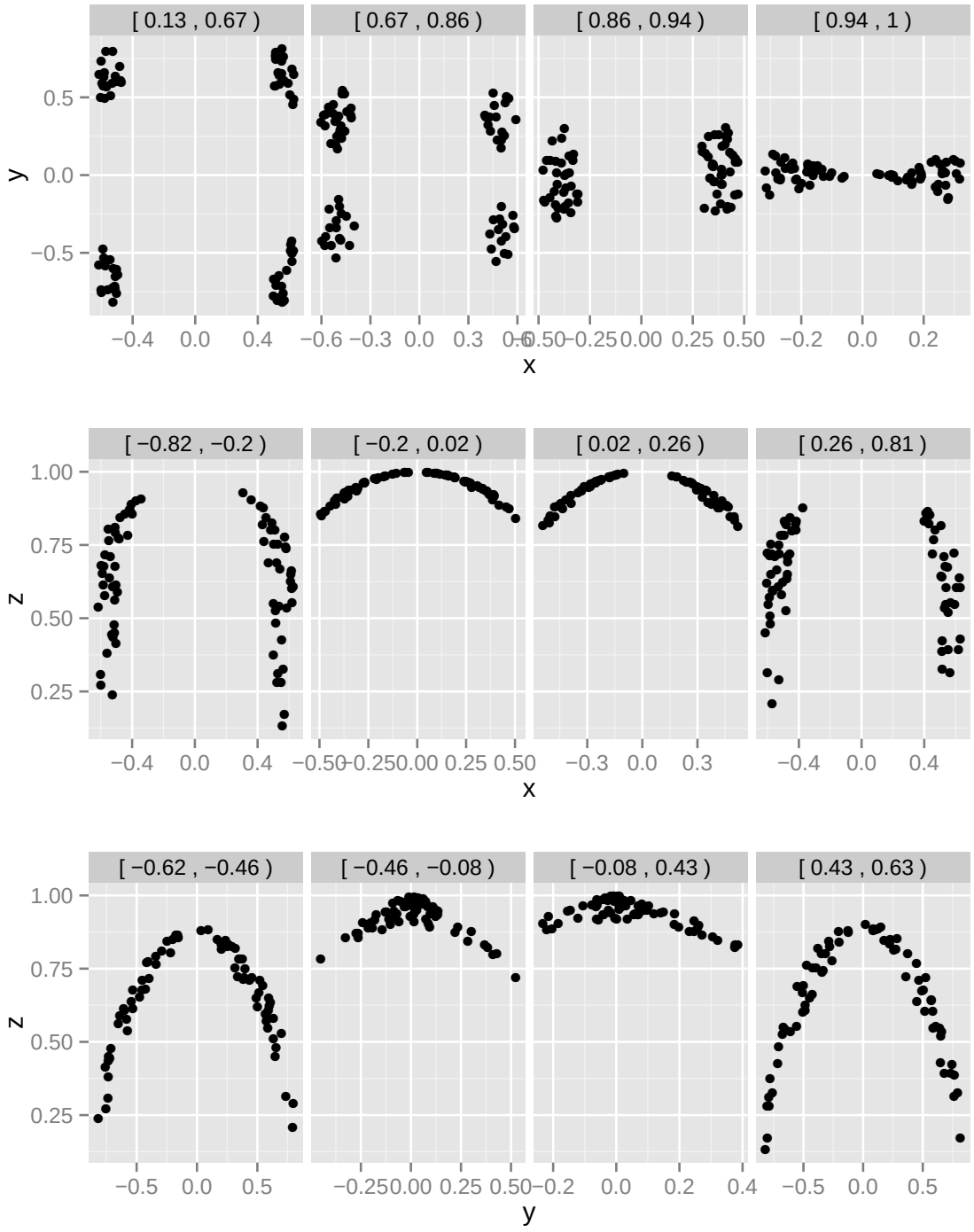


Figure B.5: Slice plots of the fifth association pattern in Figure 4.1.

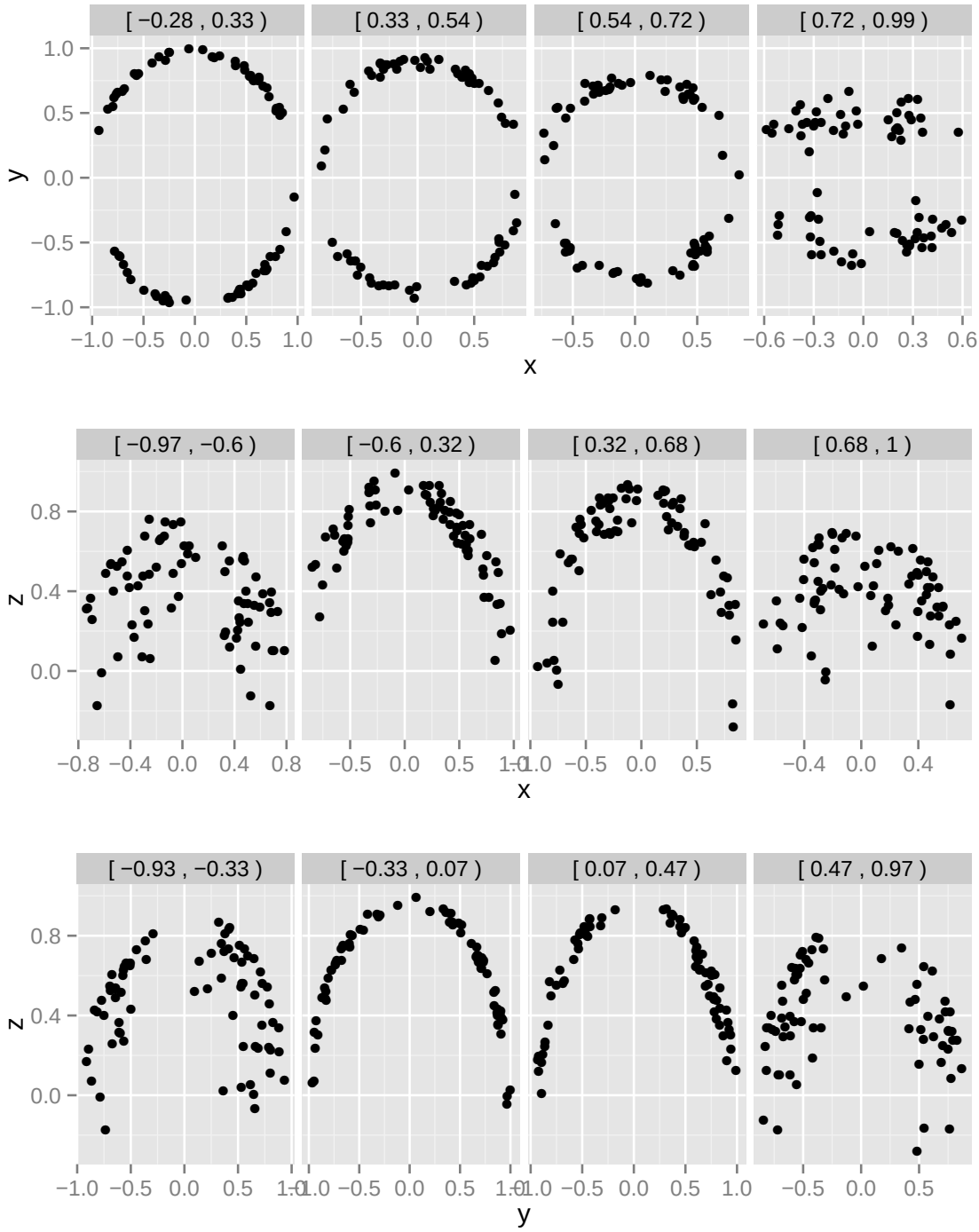


Figure B.6: Slice plots of the seventh association pattern in Figure 4.1.

Appendix C

Manhattan plots for tree growth

In this appendix we show manhattan plots for the other chromosomes (supplementary to Chapter 5).

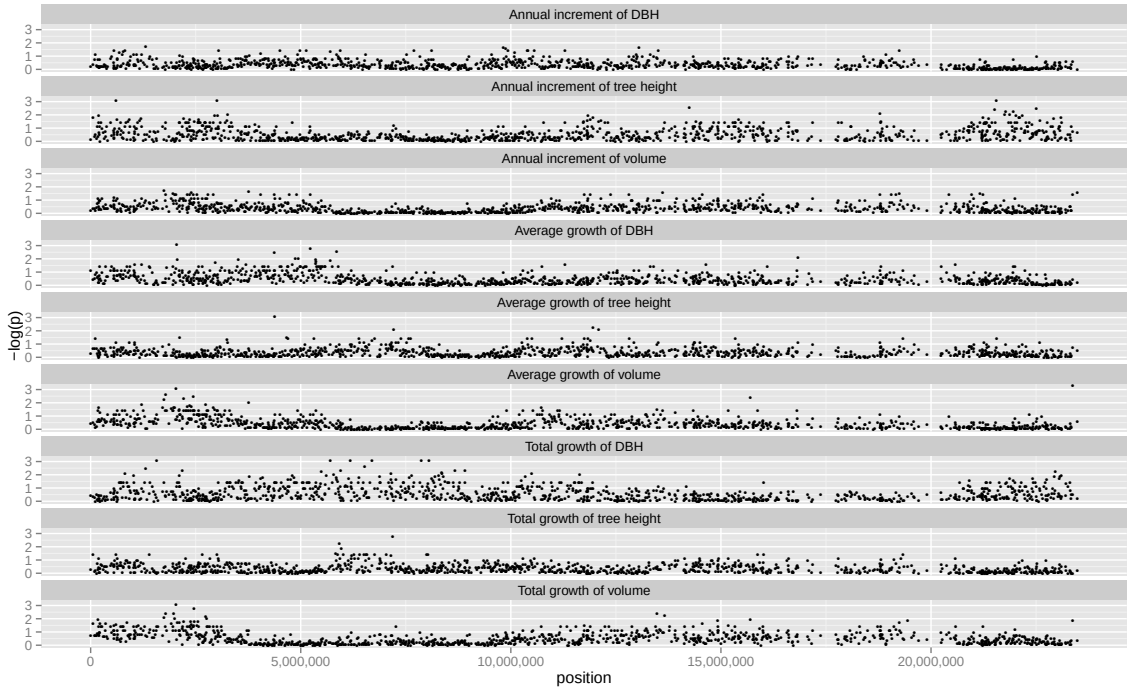


Figure C.1: Manhattan plots for chromosome 2.

Figure C.2: Manhattan plots for chromosome 3.

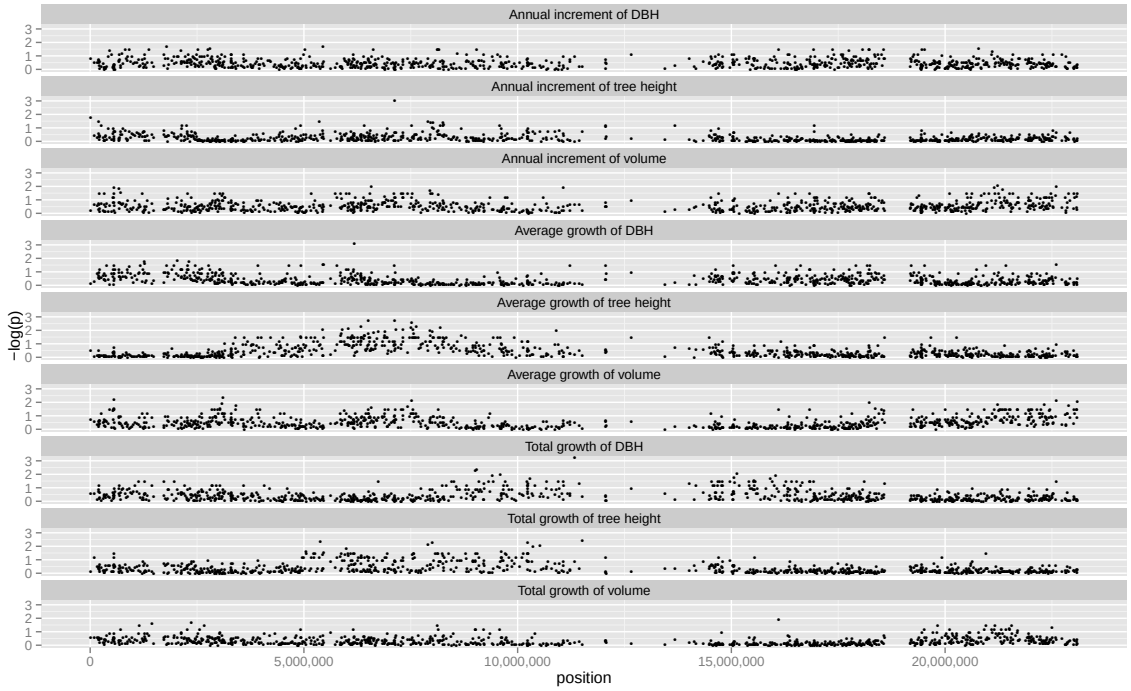


Figure C.3: Manhattan plots for chromosome 4.

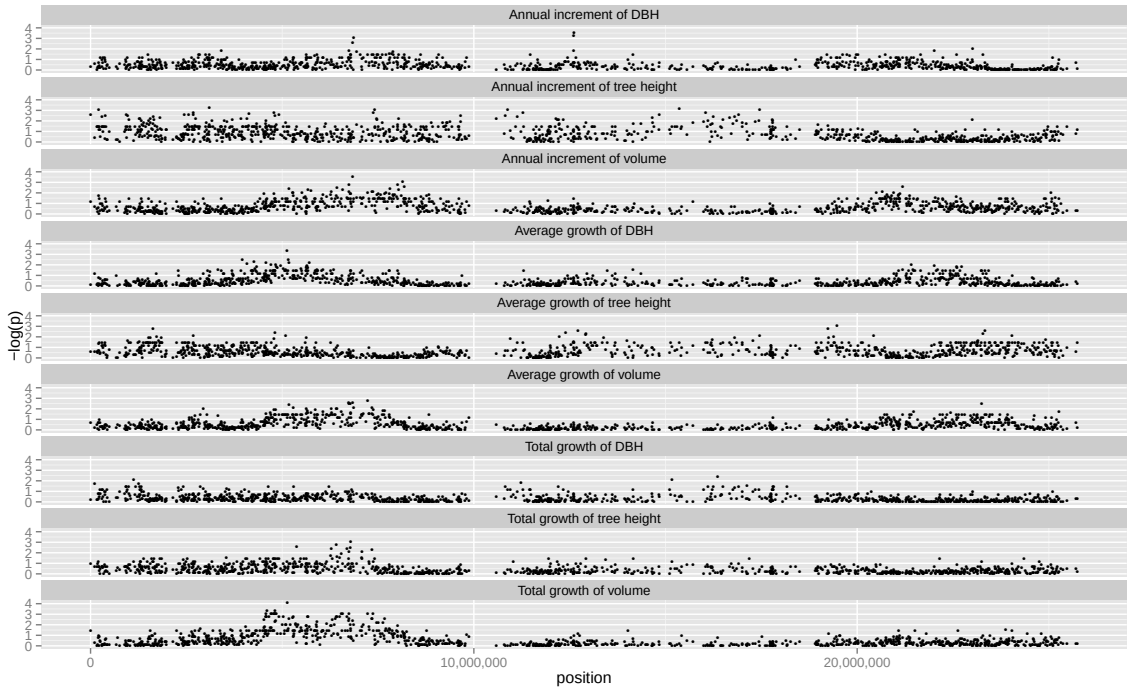


Figure C.4: Manhattan plots for chromosome 5.

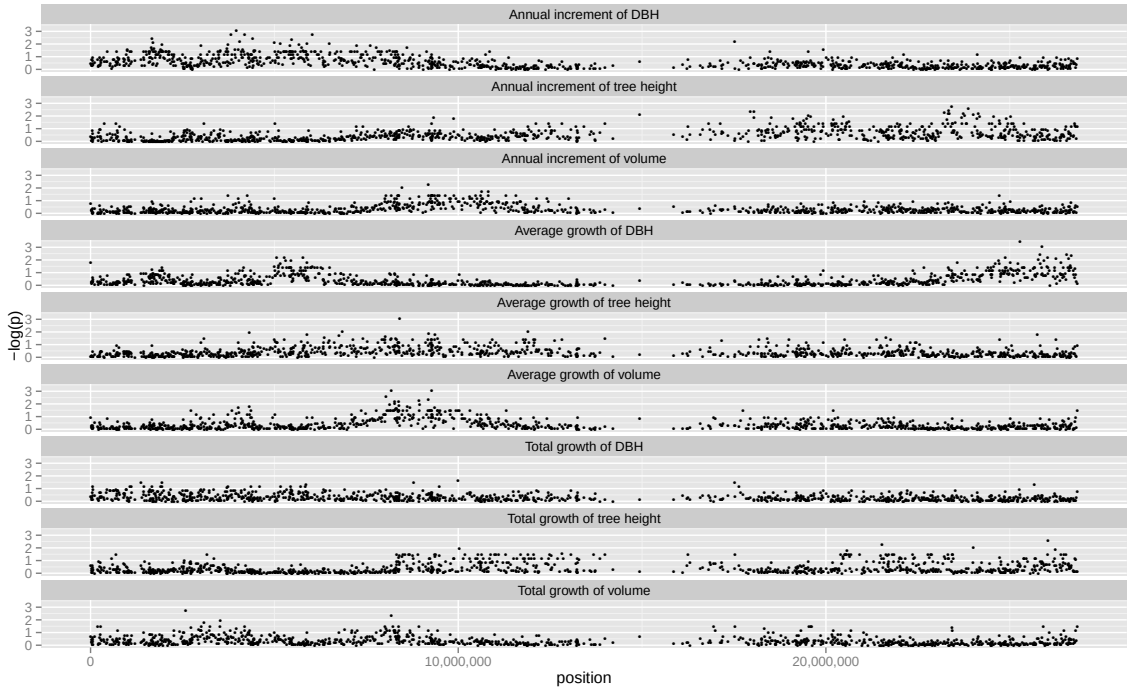


Figure C.5: Manhattan plots for chromosome 6.

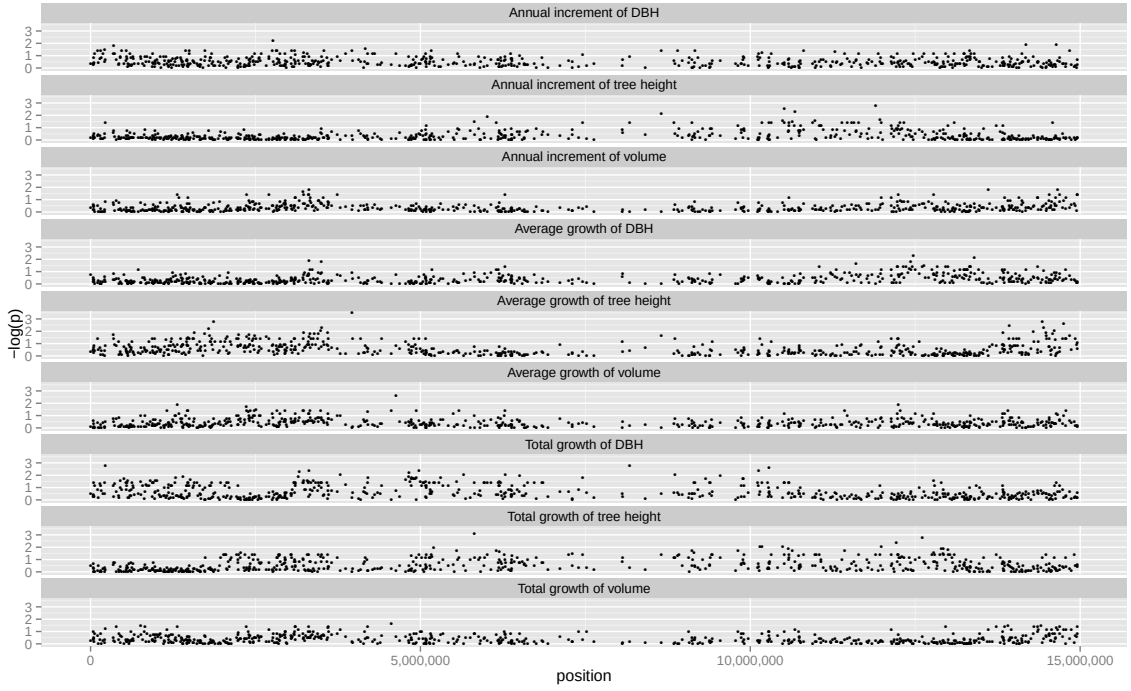


Figure C.6: Manhattan plots for chromosome 7.

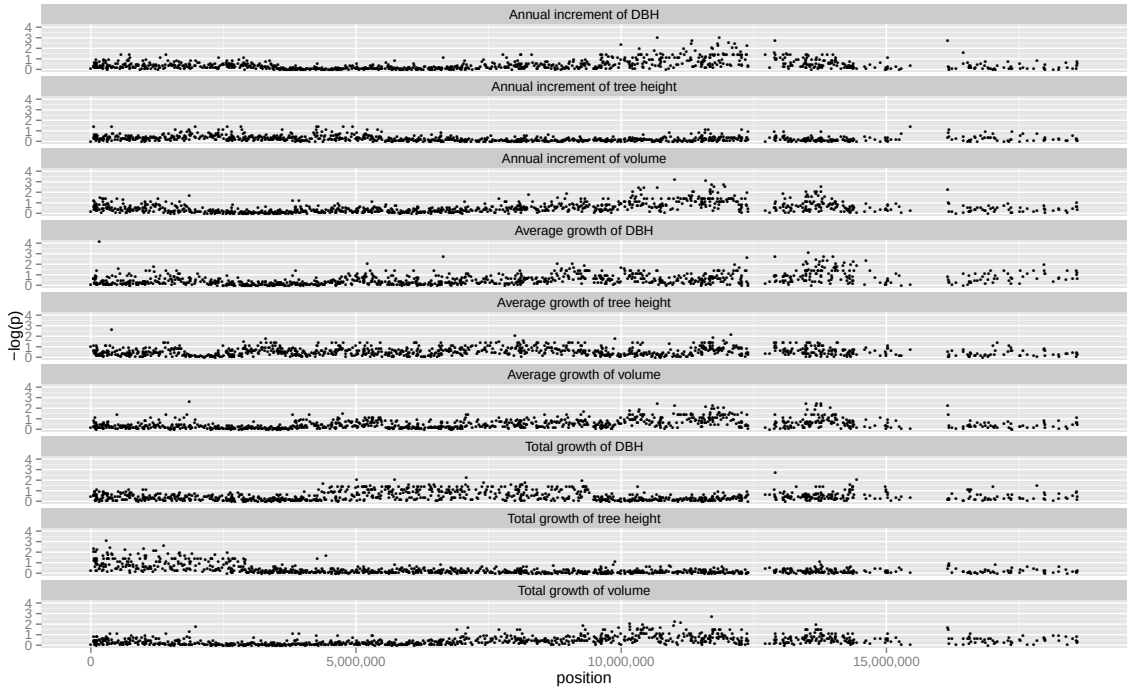


Figure C.7: Manhattan plots for chromosome 8.

Figure C.8: Manhattan plots for chromosome 9.

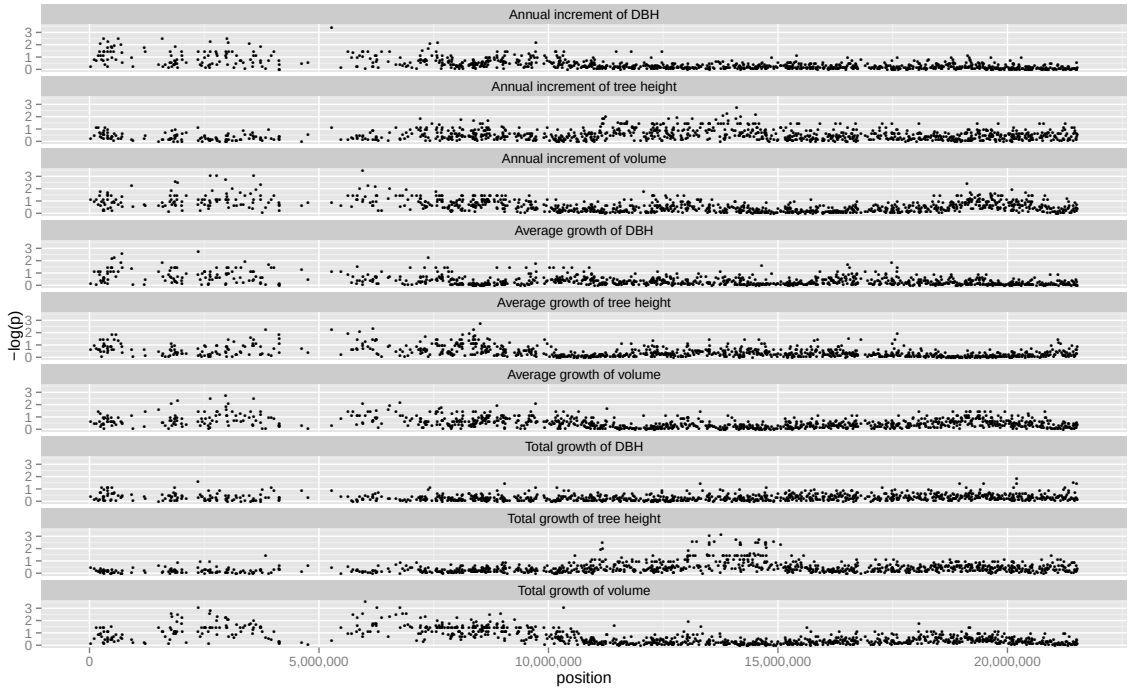


Figure C.9: Manhattan plots for chromosome 10.

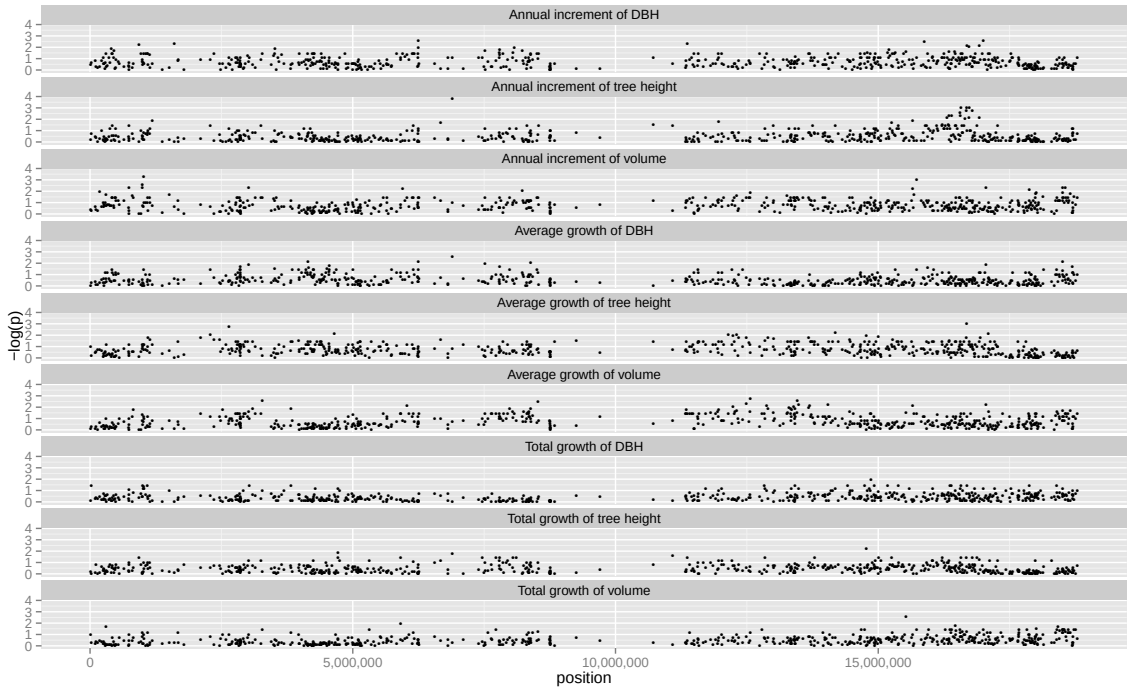


Figure C.10: Manhattan plots for chromosome 11.

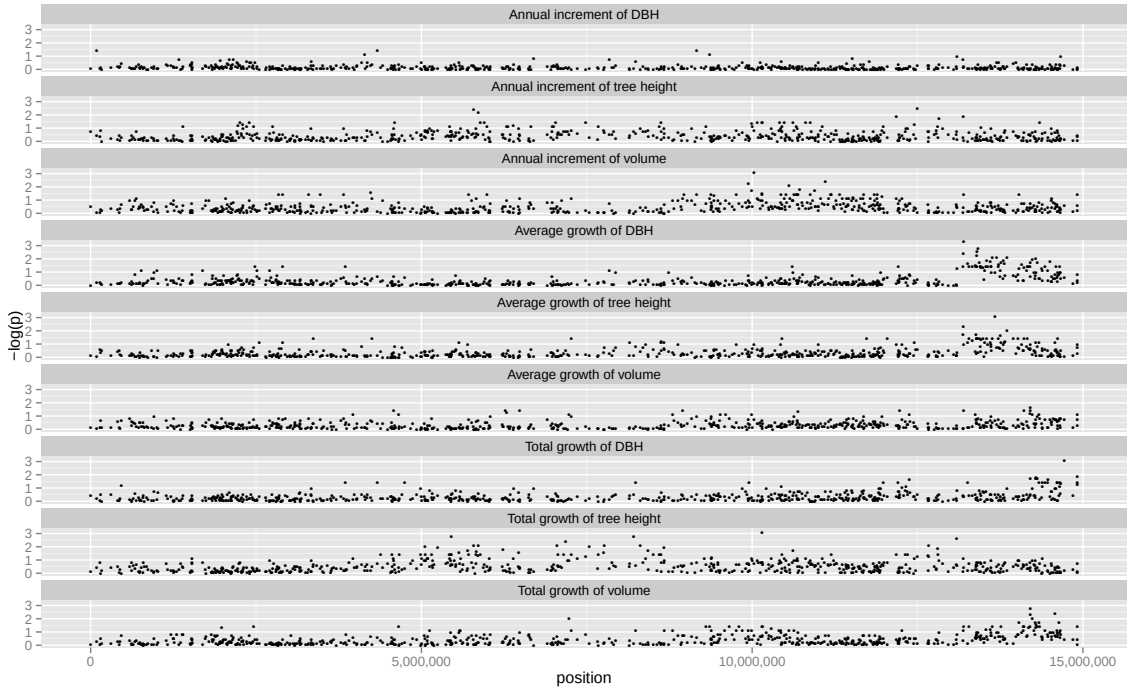


Figure C.11: Manhattan plots for chromosome 12.

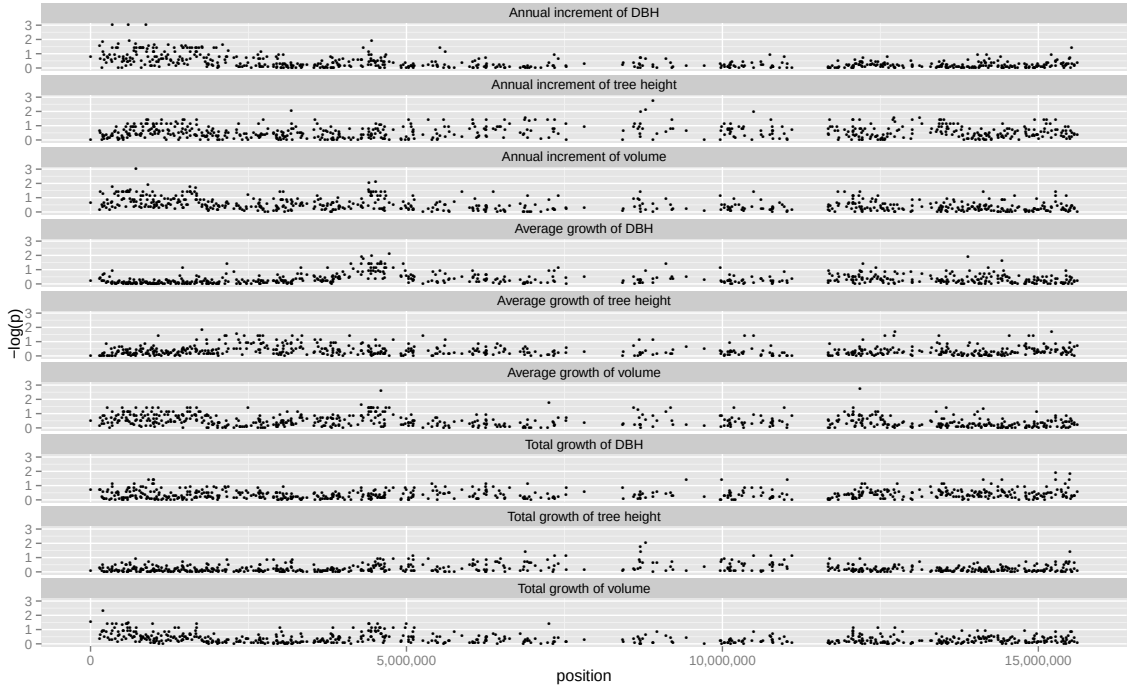


Figure C.12: Manhattan plots for chromosome 13.

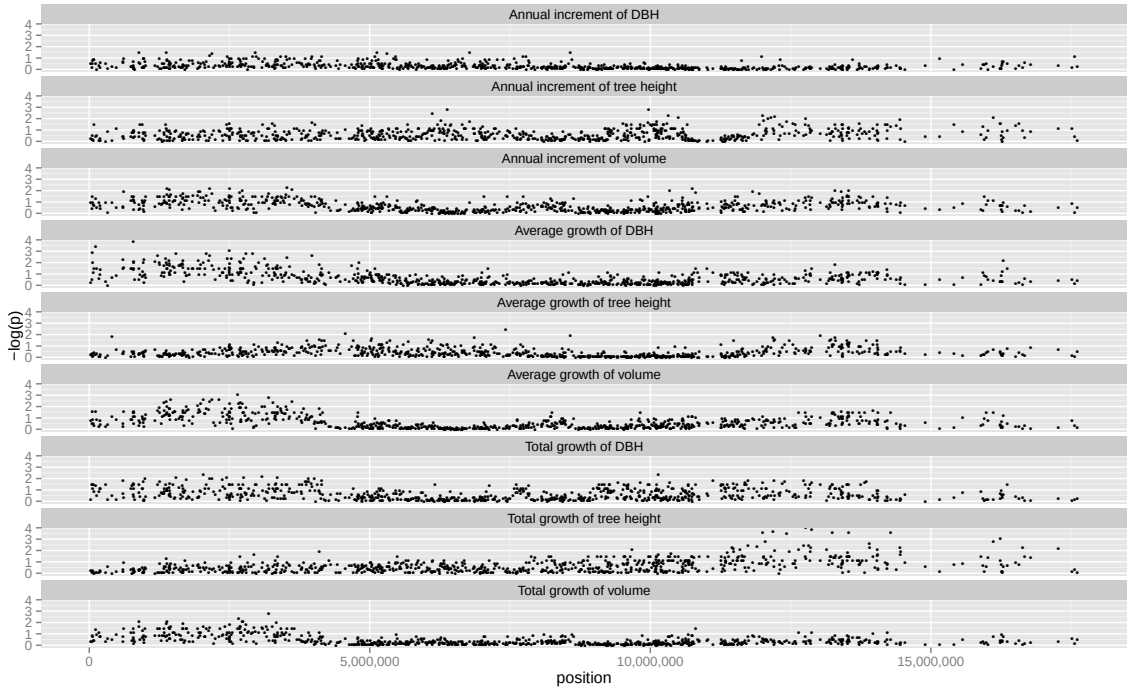


Figure C.13: Manhattan plots for chromosome 14.

Figure C.14: Manhattan plots for chromosome 15.

Figure C.15: Manhattan plots for chromosome 16.

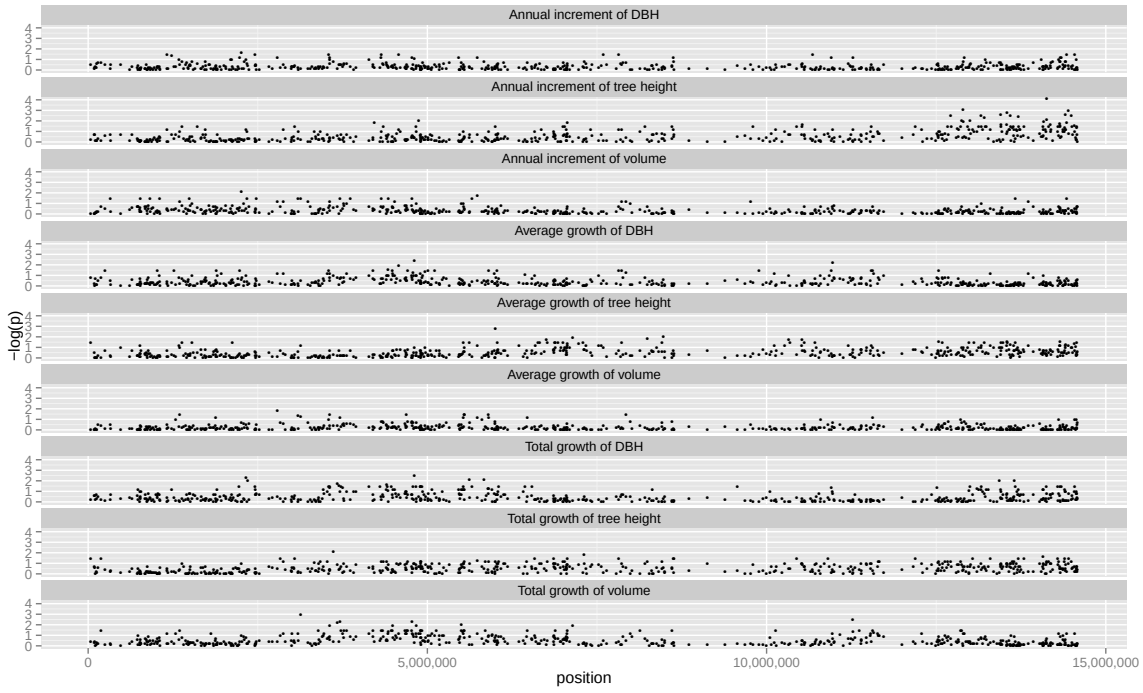


Figure C.16: Manhattan plots for chromosome 17.

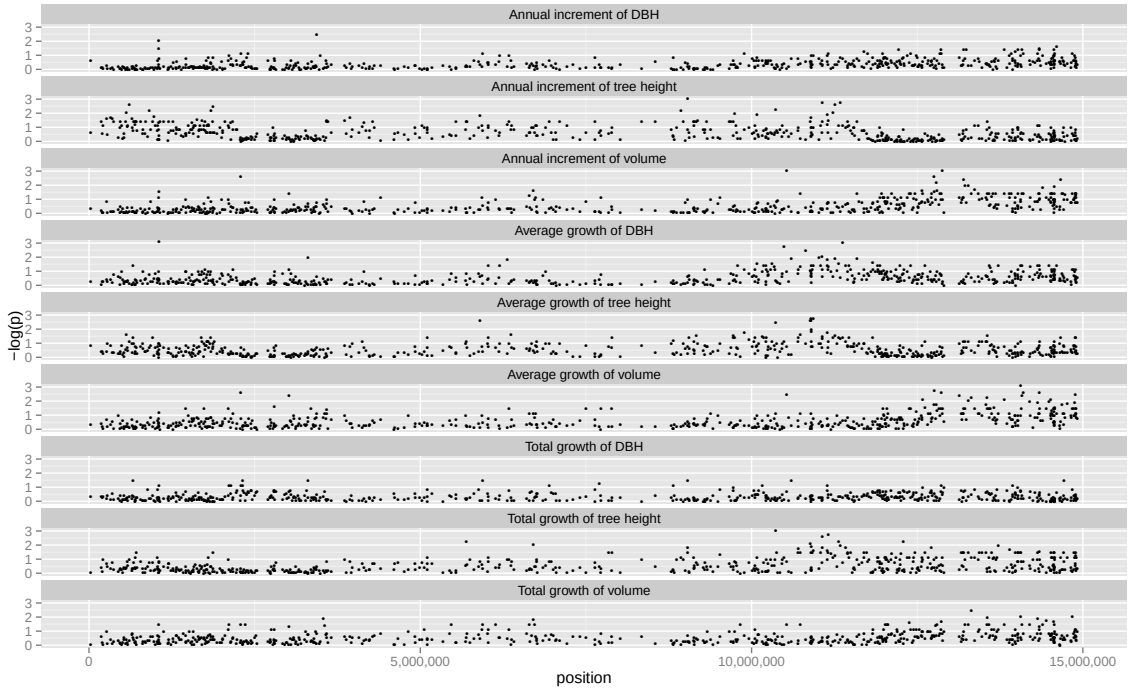


Figure C.17: Manhattan plots for chromosome 18.

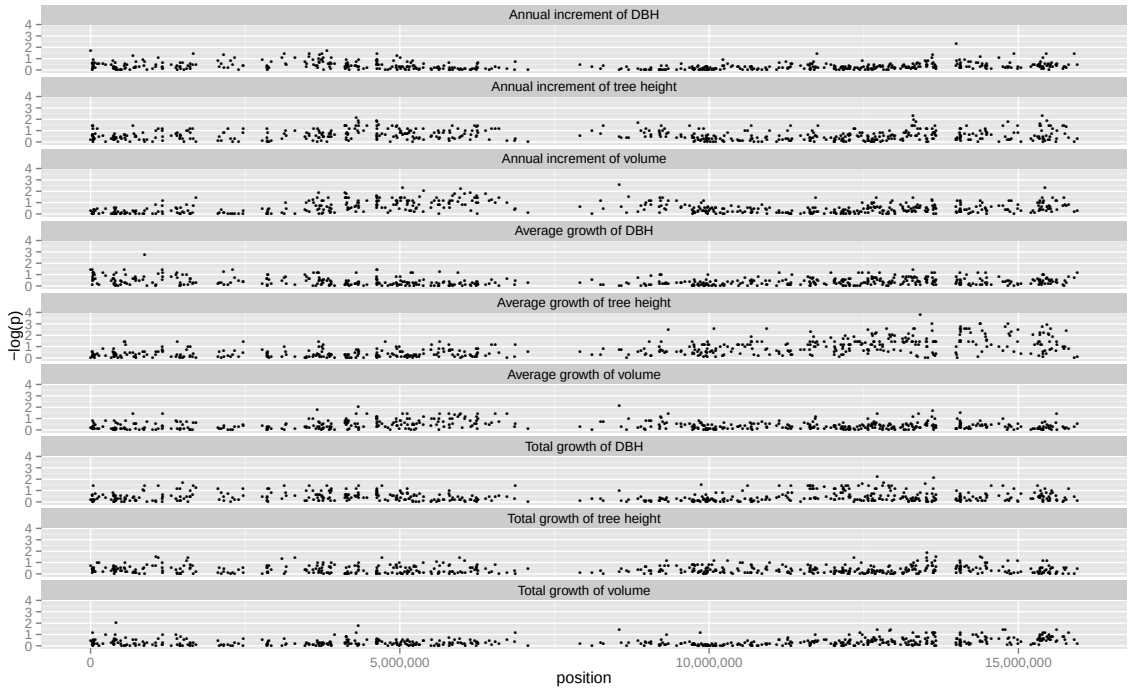


Figure C.18: Manhattan plots for chromosome 19.