The Pennsylvania State University

The Graduate School

Department of Statistics

MATRIX DISTANCES WITH THEIR APPLICATION TO FINDING DIRECTIONAL DEVIATIONS FROM NORMALITY IN HIGH-DIMENSIONAL DATA

A Dissertation in

Statistics

by

Guodong Hui

Submitted in Partial Fulfillment
of the Requirements
for the Degree of

Doctor of Philosophy

May 2008

The dissertation of Guodong Hui was reviewed and approved* by the following:

Bruce G. Lindsay
William Professor of Statistics
Head of the Department of Statistics
Dissertattion Adviser, Chair of Committee

Thomas P. Hettmansperger,
Professor of Statistics

Runze Li
Associate Professor of Statistics

Jesse Barlow
Professor of Computer Science and Engineering

*Signatures are on file in the Graduate School.

Abstract

Projection pursuit is a technique locating projections from high- to low-dimensional space that reveal interesting non-linear features of a data set, such as clustering and outliers. The two key components of projection pursuit are the measure of interesting features (projection index) and its algorithm. In this thesis, two projection matrix indices based on Fisher information matrix are presented. Both matrix indices are easily estimated by the kernel method. The eigenanalysis of the estimated matrix index provides all solution projections. The asymptotic distribution of the estimated index is studied using the Von-Mises expansion and kernel-based quadratic distance theory. The application to simulated data and real data sets shows that our algorithm successfully reveals interesting features in fairly high dimensions with a practical sample size.

Table of Contents

List of Tables

List of Figures

Acknowledgments

I would like to thank my thesis advisor Dr. Bruce Lindsay for his great help. His guidance, insight and creative thinking made it all possible. I learned a lot from our weekly meetings. I would like to thank the other members of my committee, Dr.Thomas Hettmansperger, Dr.Runze Li and Dr.Jesse Barlow for their time and the assistance they provided at all levels of the research project. I would also like to thank Dr. Donald Richards for his suggestions and provision of an important paper.

I owe a tremendous amount of gratitude to my lovely wife and my two daughters who have given me great pleasure and their precious love. Finally, I must thank my parents and my parents-in-law as being so supportive. They have spent more than two years to take care of my babies. I can not get this far without their continuous support.

Chapter 1

Introduction

Projection pursuit is a technique to explore interesting structures(such as clustering, skewness or outliers) of multivariate data set by projecting the data onto some low-dimensional spaces. The two basic components of projection pursuit are its index and its algorithm.

A ***projection index*** is designed to measure "interesting" features. Usually it is a distance between the marginal distribution of the data projection in a direction and some "uninteresting distributions" for that marginal distribution. Based on both theoretical and empirical evidence, researchers have reached the consensus that normality best represents the notion of "uninterestingness" (Diaconis and Freedman 1984, Huber 1985). Theoretically, any statistic minimized by the normal distribution, or a test statistic for non-normality can be used as projection index. These indexes are thus optimized to find projections showing departures form normality. Obviously different indexes pick up different departures from normality. Another requirement is that projection index for non-normality should be affinely invariant since linear transformations preserve normality.

A good projection index should be rapidly computable in practice. In current researches, most projection indexes are scalar measures (e.g., determinant or trace of Fisher information, standardized negative Shannon entropy, Hellinger metric). And only a few

projection indexes can be maximized algebraically. Most projection pursuit algorithms have the drawback of a high computational cost. In order to find the optimal projection, the projection index needs to be calculated or estimated for every possible projection. When the dimension increases, the computation cost increases exponentially. After the optimal one-dimensional projection is found, another search has to be done to obtain the optimal two-dimensional projection. Friedman (1987) partially solved the computation problem by expanding his projection index using orthogonal polynomials, and calculating its derivative. After an interesting projection has been found, a transformation is performed to remove the most interesting projection, but still keep all other features unchanged. Then the procedure can be restarted from the beginning to reveal more of the structure of the data set.

This thesis presents two new projection pursuit algorithms based on the standardized Fisher information matrix J_f for a density. One projection index J_{f_2} is the standardized Fisher information matrix for the density square transformed distribution. The other one Q_f is the second term of the Von-Mises expansion of the standardized Fisher information. Both the two new projection indices are matrix measures of non-normality. Compared to the classical standardized Fisher information matrix, the two new indices have a big computation advantage. The least normal projection from the new projection indices can be estimated algebraically just as in principal component analysis, provided the data is standardized (i.e., linear effects are removed). One only needs to estimate the matrix measures by the kernel method, and then do eigenanalysis for the estimated matrices. From the eigenanalysis, we could find the most interesting linear projections for future study, or from a converse point of view, we could find and

discard the least interesting linear projections. We will call the first principal component with the largest eigenvalue the least normal projection. If an eigenvalue reaches the lower bound, the corresponding linear projection is **white noise coordinate**, where a white noise coordinate is marginally normally distributed and is independent of all the other solution projections. So white noise coordinates can be discarded in further study.

For our estimated projection indices, statistical performance highly depends on sample size, dimensionality and a smoothing parameter. When the true distribution is normal but sample size is small, the eigenvalues may be not close to the theoretical lower bound $\frac{1}{4}$. We construct tests based on eigenvalues to detect the white noise coordinates within the solution projections.

In order to find the asymptotic distributions of the two new projection indices, we treat them as functions of distributions. For example, the Fisher information for the density square transformed distribution is a measure of non-normality and reaches the minimum at normal distribution. The asymptotic distribution of the estimation is determined by the second order term of the Von-Mises expansion, which is itself a kernel-based quadratic distance between the estimated distribution and the normal distribution. The asymptotic distribution of a kernel-based quadratic distance can be found using spectral decomposition of the distance kernel(Lindsay et al. 2006).

Chapter 2 introduces the framework of projection pursuit(Huber 1985) and some popular projection indices. In Chapter 3, we introduce standardized Fisher information matrix J_f and the eigenanalysis of J_f. The two new projection indices J_{f_2} and Q_f are developed in order to solve the computation problem of the classical standardized Fisher information matrix. Chapter 4 studies the standardized Fisher information matrix J_{f_2}

after the density square transformation. We present the explicit form of J_{f_2} for mixture normal models. The consistent estimator of J_{f_2} is constructed. The eigenanalysis of J_{f_2} is introduced to find the least normal projections and white noise coordinates. In Chapter 5, we investigate the projection index Q_f from the Von-Mises expansion of J_f. Similar to the study of J_{f_2}, we introduce the consistent estimator and the eigenanalysis. In Chapter 6, we introduce the kernel-based quadratic distance theory(Lindsay et al. 2006), and extend the theory to multivariate version. In Chapter 7, the two new projection indices are applied to simulated data sets and real data sets. The tests for detecting white noise coordinates are presented. In Chapter 8, we will summarize the performance of the two projection indices and present the future work. All proofs are listed in the appendix.

Chapter 2

Projection Pursuit

2.1 Overview

When one maps a multidimensional space into a space of fewer dimensions, one is performing dimension reduction. Dimension reduction allows us to visualize, categorize, or simplify large data sets. Some information of the data is lost unless the data fall exactly on the object subspace. Dimensionality reduction is effective if the loss of information due to mapping to a lower-dimensional space is less than the gain due simplifying the problem. Linear projection is the most widely used dimension reduction transformation in theory and practice, because it is easy to construct and interpret. In this context, our goal is to find the most interesting projections to study, and discard the least interesting projections.

The first successful projection pursuit methodology was the contribution of Friedman and Tukey(1974). Their idea was to assign a certain objective function to every projection, and then search the interesting lower dimensional projections by maximizing the objective function. This method was first termed *projection pursuit*. The objective function was called the *projection index*. Projection pursuit consists of two basic elements: projection index and the algorithm. Projection pursuit techniques were originally proposed by Kruskal(1969), but no successful implementation was given.

A unified framework for projection pursuit was introduced by Huber(1985). It provided a basis for further research of this subject.

In the following sections, we will introduce Huber's framework on projection index,and reasons for our interest in non-normal projections. Some important projection indices are also introduced.

2.2 The Framework of Projection Pursuit

Let X be a d-dimensional random variable and α be a d-dimensional vector. The projection index $Q(\alpha^T X)$ is a objective function that measures how interesting the projection αX is. Usually the larger the projection index, the more interesting the projection is. Huber(1985) distinguishes three classes of projection indices:

Class I Location-scale equivariant:

$$Q_I(sX + t) = sQ_I(X) + t;$$

Class II Location invariant, scale equivariant:

$$Q_{II}(sX + t) = |s|Q_{II}(X);$$

Class III Affine invariant:

$$Q_{III}(sX + t) = Q_{III}(X), s \neq 0,$$

where s, t are real numbers. Generally, the class I indices are "kind of" location estimators, and the class II indices estimate the dispersion of the data. See Huber(1985) and Jee(1985) for details. When projection index is standard deviation, it is a class II index. Then, projection pursuit becomes principal component analysis, which is the most popular dimension reduction technique in practice.

Since many interesting structures, such as clustering and special shapes, can not be detected by mean and deviation, a reasonable projection index should be affine invariant. Another reason for an affine invariant index is that we usually locate and scale pictures at will(Huber 1985). The projection index in Friedman and Tukey (1974) is a class III index, which is a product of two functions $Q(X) = S(X)D(X)$, where $S(X)$ measures the spread of the data, and $D(X)$ describes the local density. Huber(1985) shows that the index in their framework is

$$Q(X) = \sigma_\alpha(X) \int f^2(x)dx,$$

where σ_α is the α trimmed standard deviation, and f is the density.

There is no universal agreement on what it means for a projection to be interesting. But both theoretical and practical evidences show that a lack of marginal normality for a chosen set of linear projections would make it interesting. (Diaconis and Freedman 1984, Huber 1985, Jee 1985). First, theoretically, a linear projection being sum of random variables tends to be normal as dimensionality increases under conditions given in Diaconis and Freedman 1984. Second, the multivariate normal distribution is elliptically symmetric and has the least information (Fisher information, negative entropy) for a

fixed variance. Third, if the optimal projection from a algorithm is not significantly different form normal, then the whole data is believed to be normal. Finally, assuming that the normal distribution defining uninteresting structures has a computation advantage. Based on these reasons, researchers have given the heuristic agreement that a projection is less interesting, the more nearly normal it is(Huber 1985, Friedman 1987).

2.3 Previous work

According to the above discussion, we believe that the least non-normal projections should contain interesting features of the multivariate data. A suitable projection index essentially amounts to a test of non-normality in the projected data.

Huber (1985) recommended several indices:

Example 1 Standardized absolute cumulants:

$$Q_c(X) = |c_m(X)|/c_2(X)^{m/2}, m > 2,$$

where c_m is mth cumulant of X.

Example 2 Standardized Fisher information:

$$Q_F(X) = \sigma^2(X) \int (\frac{f'}{f})^2 f dx - 1.$$

Example 3 Standardized negative Shannon entropy

$$Q_S(X) = \int log(f) f dx + log((2\pi e)^{\frac{1}{2}} \sigma(X)).$$

All of three indices are non-negative, with equality if X is normal. The thesis of Jee'(1985) studied the projection pursuit indices based on classic Fisher information matrix and the standardized negative Shannon entropy. These are estimated using over-smoothed histogram method. His algorithm is applied to particle physics data and Minnesota forests data.

Friedman (1987) presented a new projection pursuit algorithm to find the least normal projections. First a uniform transformation is performed

$$R = 2\Phi(X) - 1,$$

where Φ is the standard normal cdf. If X is standard normal, R will be uniformly distributed in $[-1, 1]$. And then Friedman defined an integrated squared error of densities to measure the non-uniformity of R

$$Q(R) = \int_{-1}^{1} (f_R(r) - \frac{1}{2})^2 dr,$$

where f_R is the density of R, $f_R = \frac{1}{2}$ when R is uniform(X is normal). $Q(R)$ is expanded using Legendre polynomials, and the derivatives are calculated via chain rule and recursion relation of Legendre polynomials. The optimal projection can be found relatively quickly. The algorithm for two-dimensional projection pursuit is applied to three real data sets: the states data, the cars data and the Boston neighborhood data, which we will be examined in chapter 6.

Theoretically, any normality test statistic can be used to construct a projection index. However, in order to find optimal projections in a high dimensional space, computational properties are crucial(Friedman 1987). The above three projection pursuit methods have proved their usefulness in finding interesting projections in real data analysis. Their biggest drawback remaining is the computation. The algorithms must search the whole space to find an optimal projection(Friedman 1984, Jee 1985). The solution projections usually are only local maxima of projection index, not the real largest maxima. Friedman (1987) partially solved the problem. His algorithm is rapidly computable, after the derivatives are found. A transformation is provided to remove the structure of the optimal projection and keep the structures not captured, but it is still not easy to find the high dimensional projections. The derivative computation is very complex for high dimensional projections. A sequential approach is usually needed in current projection pursuit algorithms.

In order to find a faster and more reliable methodology, we borrow the eigenanalysis idea from principal component analysis. In a principal component analysis(PCA), the projection index is the covariance matrix. The optimization is solved by performing an eigenanalysis of the sample covariance matrix. The eigenvectors with the largest eigenvalues correspond to the dimensions that have the strongest correlation in the data. The ith principal component has largest variance in all vectors orthogonal to the previous $i-1$ principal components.

In the following chapters, we will introduce how to use standardized Fisher information matrix to construct a new projection index. It is easy to estimate using kernel method. The solution projections arise from the eigenanalysis of the estimated matrix

measure provided the data is standardized. The projections with large eigenvalues reveal the most interesting features of the data. Here an *interesting* set of linear combinations might mean those that display nonlinear structural relationships, clustering, or other forms of dependence that can occur despite zero correlation. If an eigenvalue reaches the lower bound of eigenvalue, the corresponding projection is white noise. It is marginally normal and independent of other projections. So it can be discarded. Thus in contrast to classical projection pursuit, our methodology is projection pursuit plus *white noise detection*.

Chapter 3

Standardized Fisher Information Matrix

3.1 Overview

In this chapter we directly consider the standardized Fisher information matrix as a matrix non-normality measure for multivariate data. We will show that, similar to PCA, an eigenanalysis of standardized Fisher information matrix provides the optimal solution projections.

This leaves the problem of how to estimate standardized Fisher information matrix, which ordinarily require numerical integration for a kernel estimator of $f(x)$. We will show that the density square transformation $f_2(x) = f^2(x)/\int f^2(y)dy$ preserves the most important structure of the data. The standardized Fisher information matrix for f_2 has an explicit form for normal mixture models, and so we can compute it without numerical integration when using a normal kernel density estimator for $f(x)$.

3.2 Standardized Fisher Information Matrix

Let $X = (X_1, X_2, ..., X_d)$ be a d-dimensional random vector with the density function $f(x)$, mean μ and covariance matrix V_f.

DEFINITION 1. $V_f^{1/2}\Big(\int \frac{\nabla_x f \cdot \nabla_x f^T}{f} dx\Big) V_f^{1/2} := J_f$ *is called Standardized Fisher Information Matrix where* $\nabla_x f(x) = (\frac{\partial}{\partial x_1} f, ..., \frac{\partial}{\partial x_d} f)$.

Since the covariance matrix V_f is a nonnegative-definite symmetric matrix, by the finite-dimensional case of the spectral theorem, the matrix square root $V_f^{1/2}$ exists and is nonnegative symmetric. The Fisher information matrix J_f is called standardized because the mean μ and covariance matrix V_f do not affect the value. For fixed variance V_f, the non-standardized Fisher Information $\int \frac{\nabla_x f \cdot \nabla_x f^T}{f} dx$ has been studied in literature. In the case d=1, it is called Fisher information number(Terrell 1995, Papaioannou 2005).

In the standardized Fisher information matrix, we consider the derivative with respect to x rather than the parameters as done in ordinal Fisher information matrix. So it measures the information in the density, not the parameters. Kagan(2001) demonstrated the connection between the Fisher information for a density and Fisher information for parameters as follows given. Create a location family of distributions by $f(x+\mu_f - A\theta)$, for $\mu_f = E_f(X)$ and arbitrary matrix A, so that $E(X) = A\theta$, $V(X) = V_f$. Kagan (2001) showed a matrix inequality for the Fisher information in parameter θ:

$$\begin{aligned} & E(\frac{\partial log f(x+\mu_f - A\theta)}{\partial \theta})(\frac{\partial log f(x+\mu_f - A\theta)}{\partial \theta})^T \\ = & \int \frac{\nabla_\theta f(x+\mu_f - A\theta) \nabla_\theta f^T(x+\mu_f - A\theta)}{f} dx \\ \geq & A^T V_f^{-1} A. \end{aligned}$$

The normal distribution $N(A\theta, V_f)$ has the least Fisher information for θ because the above inequality becomes equality. Applying the inequality for $A = V_f^{\frac{1}{2}}$ giving the result

$$J_f = V_f^{1/2} \Big(\int \frac{\nabla_x f \cdot \nabla_x f^T}{f} dx \Big) V_f^{1/2} \geq I_d.$$

When f is normal, the equality holds. So the standardized Fisher information matrix J_f measures the non-normality of X. And normal distribution has the smallest Standardized Fisher information matrix in the positive definite sense among all continuous distributions.

PROPOSITION 1. *If $f(x)$ is a differentiable density function, the following matrix inequality holds:*

$$J_f = V_f^{1/2}\Big(\int \frac{\nabla_x f \cdot \nabla_x f^T}{f} dx\Big) V_f^{1/2} \geq I_d. \tag{3.1}$$

where, as always, $A_1 \geq A_2$ means that $A_1 - A_2$ is positive semi-definite matrix. If and only if f is a normal density function, 3.1 becomes equality.

We next make some remarks on the interpretation of J_f. The inequality may not hold if the density $f(x)$ is not differentiable for every point x. For example, consider the bivariate distribution with the density

$$f(x_1, x_2) = 2\phi\Big((x_1, x_2)^T, (0,0)^T, I_2\Big) I(x_1 x_2 > 0),$$

where ϕ is the normal density function. We can show that the matrix inequality does not hold for this density $f(x_1, x_2)$, because it is not differentiable at $(0, 0)$. However, the standardized Fisher information matrix for $f(x_1, x_2)$ still exists. In section 6.3, we will use it as an example to show that the projection pursuit using the standardized Fisher information matrix index provides reasonable solution projections even though the matrix inequality does not hold.

Suppose X has been standardized: $V_f = I_d$. So there are no linear relationships left in the distribution, because the correlations are all zero. The ith diagonal term can be expressed as

$$\begin{aligned} & \int \Big(\frac{\partial}{\partial x_i} f(x_1, ...x_d)/f(x) \Big)^2 f(x) dx \\ = & \int \Big(\frac{\partial}{\partial x_i} log f(x_1, ...x_d) \Big)^2 f(x) dx \\ = & \int \Big(\int \Big(\frac{\partial}{\partial x_i} log f(x_i|x_{-i}) \Big)^2 f(x_i|x_{-i}) dx_i \Big) f(x_{-i}) dx_{-i} \\ = & \int J_{X_i|x_{-i}} f(x_{-i}) dx_{-i}, \end{aligned} \tag{3.2}$$

where $x_{-i} = (x_1, ..., x_{i-1}, x_{i+1}, ..., x_d)$, and $J_{X_i|X_{-i}}$ is the Fisher information for the conditional distribution $f(x_i|x_{-i})$. That is, the ith diagonal term of J_f is not the Fisher information of the marginal distribution of X, but the weighted average of Fisher information of x_i conditioned on the rest of the uncorrelated variables. The weight is the density function $f(x_{-i})$. Obviously the matrix inequality for J_f holds for the conditional distribution $f(x_i|x_{-i})$:

$$J_{X_i|x_{-i}} \geq 1, \forall x_{-i}.$$

We can conclude that, if the ith diagonal term of J_f reaches the lower bound 1, then

$$J_{X_i|x_{-i}} = 1, \forall x_{-i},$$

That is $X_i|x_{-i}$ is standard normal for any x_{-i}. We can conclude that X_i is marginally normal and independent of all other variables. We will then call X_i a **white noise**

coordinate. If the ith diagonal term of J_f is bigger than one, some combination of non-normality and dependence exists. Since X has been standardized, all X_i are uncorrelated. So any dependence must arise from a nonlinear structural relationship, clustering, or other forms of dependence that can occur despite zero correlation.

3.3 Eigenanalysis of J_f

In the section, we present how to use an eigenanalysis of J_f to find solution projections.

PROPOSITION 2. *Let A be a $d \times d$ nonsingular matrix and $Y = AX$. Then*

$$J_g = V_g^{1/2} A^{-T} V_f^{-1/2} \cdot J_f \cdot V_f^{-1/2} A^{-1} V_g^{1/2}, \tag{3.3}$$

where $g(y)$ is the density of Y, V_g is the covariance matrix of g, $f(x)$ is the density of X, V_f is the covariance matrix of f.

We next apply this result. Let $A = V_f^{-\frac{1}{2}}$, and $Y = V_f^{-\frac{1}{2}} X$. Suppose $g(y)$ is the density of Y. Then

$$V_g = Cov(Y) = V_f^{-\frac{1}{2}} V_f V_f^{-\frac{1}{2}} = I_d.$$

Thus, a principal components analysis yields no structure. From the above proposition, we have

$$J_g = V_g^{1/2} A^{-T} V_f^{1/2} J_f V_f^{1/2} A^{-1} V_g^{1/2} = J_f.$$

So, standardizing a vector leaves the Fisher information unchanged. Secondly, consider an orthogonal transformation $Z = \Gamma^T Y$, where $\Gamma = [\gamma_1^T, \gamma_2^T, ..., \gamma_d^T]^T$ is an orthogonal

matrix. These transformations preserve the standardized structure,

$$V_h = Cov(Z) = \Gamma^T V_g \Gamma = I_d,$$

and so do not create any new linear relationships. We can think of the new vector $Z = (\gamma_1^T Y, ..., \gamma_d^T Y)$ as a vector of projections, as each coordinate Z_j is the projection of Y onto the linear space spanned by γ_j.

The density of the projection Z is $h(z) = g(\Gamma^T z)$. Then and

$$\begin{aligned} J_h &= V_h^{1/2}\Gamma^{-T}V_g^{1/2}J_g V_g^{1/2}\Gamma^{-1}V_h^{1/2} \\ &= \Gamma J_f \Gamma^T \\ &= \Big(\gamma_i^T J_f \gamma_j\Big). \end{aligned}$$

Consider the ith diagonal term of J_h:

$$\begin{aligned} & \int \Big(\frac{\partial}{\partial z_i} h(z_1, ...z_d)/h(z)\Big)^2 h(z)dz \\ = & \int \Big(\frac{\partial}{\partial z_i} log h(z_1, ..., z_d)\Big)^2 h(x)dx \\ = & \int \Big(\int \Big(\frac{\partial}{\partial z_i} log h(z_i|z_{-i})\Big)^2 h(z_i|z_{-i})dz_i\Big) h(z_{-i})dz_{-i} \\ = & \int J_{Z_i|z_{-i}} h(z_{-i})dz_{-i}, \end{aligned} \tag{3.4}$$

where $J_{Z_i|z_{-i}}$ is the standardized Fisher information matrix for $Z_i|z_{-i}$. This leads to a way to interpret the least normal directions. Let $\lambda_1 \geq \lambda_2, ... \geq \lambda_d$ be the eigenvalues of $J_g (= J_f)$ and $\gamma_1, ..., \gamma_d$ be the corresponding eigenvectors. Then $Z_1 = \gamma_1^T Y$ is the optimal

projection, in the sense of maximizing $\beta^T J_f \beta$, where β is a d-component column vector such that $\beta^T \beta = 1$. The maximal value of $\beta^T J_f \beta$ is equal to $\gamma_1^T I_Y \gamma_1 = \lambda_1 \gamma_1^T \gamma_1 = \lambda_1$. The Fisher information matrix of the orthogonal projection $Z = \Gamma Y$ can be expressed diagonally by the eigenvalues of the Fisher information matrix of Y:

$$\begin{aligned} J_h = & \int \frac{\nabla_z h(z) (\nabla_z h(z))^T}{h(z)} dz \\ = & \left(\gamma_i^T J_Y \gamma_j \right) \\ = & \left(\gamma_i^T \lambda_j \gamma_j \right) \\ = & \; Diag(\lambda_1, \lambda_2, ..., \lambda_d). \end{aligned}$$

Note that Z_1 is not the optimal projection for maximizing the marginal Fisher information of $\beta^T Y$, because $J_{\beta^T Y}$ may not be equal to $\beta^T J_Y \beta$. The projection $Z_1 = \gamma_1^T Y$ from the eigenanalysis of J_f has the **least conditional normality** conditioned on all the other uncorrelated variables. The λ_1 is the measure of its non-normality. By a similar analysis, $Z_i = \gamma_i^T Y$ has the least conditional normality in all projections which are uncorrelated to $Z_1, ...Z_{i-1}$.

Because the normal distribution has the least Fisher Information: $J_h \geq I_d$, the eigenvalues have the lower bound 1. The $J_{Z_i|z_{-i}}$ also satisfies the Fisher information inequality:$J_{Z_i|z_{-i}} \geq 1$. When $\lambda_i = 1$,

$$J_{Z_i|z_{-i}} = 1, \forall z_{-i}.$$

This equation implies that $Z_i|z_{-i}$ is normal for any z_{-i}. So Z_i is not only marginally normal, it is also independent of the other variables $(Z_1, ..., Z_{i-1})$. If $\lambda_i = 1$, then $\lambda_j = 1, j \geq i$, and the Z_j are **white noise coordinates**, which can be discarded, as they are not only marginally normal, but also independent of the remaining variables. From the point of view of our interest in non-linear relationships, we consider the white noise coordinates to be discardable. In other words, the projections $(Z_1, ..., Z_{i-1})$ are sufficient for further analysis. The logic is that the white noise coordinates are not only marginally uninteresting, but their independence implies that they have no interesting relationships with $(Z_1, ..., Z_{i-1})$.

When the smallest eigenvalue $\lambda_d > 1$, the projection $Z_d = \gamma_d^T Y$ is **most similar** to white noise among all linear projections orthogonal to the white noise coordinates. Generally, the eigenvector γ_{k-1} should generate the linear projection most similar to white noise in the subspace orthogonal to $\gamma_k, ..., \gamma_d$, and so forth, so that the projection $Z_1 = \gamma_1^T Y$ corresponding to the largest eigenvalue λ_1 can be thought of as the most interesting projection, in the sense of being least similar to white noise. Or, if we were to use Z_1 alone, then we can say that we have discarded a subspace of projections that is most similar to white noise.

According to the above analysis, our projection pursuit procedure include two steps:

1. Do standardization $Y = V_f^{-\frac{1}{2}} X$.

2. Do eigenanalysis of $J_g = J_f$ to find the solution projections $Z = \Gamma^T Y$.

We use this standardization $Y = V_f^{-\frac{1}{2}} X$ because it can remove the linear effect and keep the Fisher information unchanged.

We next verify that we can use any standardization $Y = AX$ so that $V_g = AV_fA^T = I_d$. Finally we will get the same solution projections from the eigenanalysis of J_g. According to the above proposition, we have

$$J_g = A^{-T}V_f^{-1/2}J_fV_f^{-1/2}A^{-1} := B^TJ_fB,$$

where $B = V_f^{-1/2}A^{-1}$ is a orthogonal matrix because

$$B^TB = A^{-T}V_f^{-1/2}V_f^{-1/2}A^{-1} = I_d.$$

The new Fisher information matrix J_g has the same eigenvalues as J_f, because

$$\begin{aligned} |J_g - \lambda I_d| &= |B^TJ_fB - \lambda I_d| \\ &= |B^T||J_f - \lambda I_d||B| \\ &= |J_f - \lambda I_d|. \end{aligned}$$

Suppose λ_i is an eigenvalue, γ_{if} is the corresponding eigenvector of J_f. Then, we have

$$\begin{aligned} J_gB^T\gamma_{if} &= B^TJ_fBB^T\gamma_{if} \\ &= B^TJ_f\gamma_{if} \\ &= \lambda_iB^T\gamma_{if}, \end{aligned}$$

so $\gamma_{ig} = B^T \gamma_{if}$ is an eigenvector of J_g. The corresponding projection is $Z_i = \gamma_{ig}^T Y = \gamma_{if}^T BAX = \gamma_{if}^T V_f^{-1/2} X$, which does not depend on the transformation matrix A. So the method of standardization will not affect the final solution projections.

3.4 Computation Problem of J_f

In last section, we showed that an eigenanalysis provides all solution projections, if the standardized Fisher information matrix J_f is explicit for some distribution, or it is estimated by a numerical method. Unfortunately, in theory, usually J_f does not have explicit form for most population models, even for a mixture of normals; in practice, if we estimate the measures by replacing the density with kernel density estimate, the integration will not have explicit form because of the denominator f. Monte Carlo integration is required to calculate these measures. Jee(1985) employed the averaged shifted histogram theory of Scott(1985) to estimate these measures. According to the simulation study in Jee(1985), a large sample is required in order to find good estimate even in a low dimensional space. Jee's algorithm has another computational drawback: it needs to search the whole subspace to find the optimal projection. For example, in order to find the optimal projection in dimension 1, we need to project the data onto one direction in R and then to perform a univariate density estimate. When the subspace dimension increases, it takes long time to find the needle in a haystack.

We will present two ways to solve the computation problem: density square transformation and Von-Mises expansion.

3.4.1 Density Square Transformation

The first way is to transform X to $T(X)$ so that the density function of $T(X)$ is equal to $\frac{f^2(x)}{\int f^2(y)dy} := f_2(x)$. The density square transformation has some good properties:

1. The ordering of the density values is unchanged: $f(x_1) < f(x_2) <=> f_2(x_1) < f_2(x_2)$, and $f(x_1) = f(x_2) <=> f_2(x_1) = f_2(x_2)$.

2. The number and locations of density modes is unchanged.

3. It accentuates the peaks of density, and decreases the variance by flatting the tails.

4. It preserves the normality: X is normal if and only if $T(X)$ is normal,

$$X \sim N(\mu, \Sigma) \Leftrightarrow T(X) \sim N(\mu, \Sigma/2).$$

Plugging f_2 into (3.1) provides the inequality

$$J_{f_2} := \frac{V_{f_2}^{1/2} \int \nabla_x f \cdot \nabla_x f^T dx V_{f_2}^{1/2}}{\int f^2(x)dx} \geq \frac{1}{4} I_d, \tag{3.5}$$

where V_{f_2} is the covariance matrix of $T(X)$:

$$V_{f_2} = \frac{\int x x^T f^2(x)dx}{\int f^2(x)dx} - \Big(\frac{\int x f^2(x)dx}{\int f^2(x)dx}\Big)\Big(\frac{\int x f^2(x)dx}{\int f^2(x)dx}\Big)^T.$$

J_{f_2} is a non-normality matrix measure for $T(X)$, and also a non-normality matrix measure of X since the density square transformation T preserves the normality. Compared to J_f, J_{f_2} has at least two big advantages:

1. it has explicit form for some distributions, for example, mixture of normals;

2. it is easy to estimate using the kernel method.

We will further investigate J_{f_2} in Chapter 4.

3.4.2 Von-Mises Expansion

In this subsection, We treat the Fisher information J_f as a functional $T(F)$ on the space of distributions. We will introduce the idea of Von-Mises expansion and apply it to J_f to solve the computation problem.

Suppose F_0 is the distribution of interest. In order to expand the function $T(F)$ at F_0, we first generate a path in the space of distributions from F_0 to the distribution F, by letting

$$F_\varepsilon = (1-\varepsilon)F_0 + \varepsilon F.$$

For any ε, it generates an intermediate distribution. Assume that the ordinary derivative with respect to ε exist at $\varepsilon = 0$. Then we have the following expansion at $\varepsilon = 0$:

$$T(F_\varepsilon) - T(F_0) = \varepsilon T^{'}_{F_0}(F) + \frac{1}{2}\varepsilon^2 T^{''}_{F_0}(F) + o(\varepsilon^2),$$

where

$$T^{'}_{F_0}(F) = \int T^{'}_{F_0}(s) d(F(s) - F_0(s)),$$

$$T^{'}_{F_0}(s) = \frac{d}{d\varepsilon} T((1-\varepsilon)F_0 + \varepsilon\delta_s)|_{\varepsilon=0},$$

$$T^{''}_{F_0}(F) = \int T^{''}_{F_0}(s,t) d(F(s) - F_0(s)) d(F(t) - F_0(t)),$$

$$T^{''}_{F_0}(s,t) = \frac{\partial^2}{\partial\varepsilon_1\partial\varepsilon_2} T((1-\varepsilon_1-\varepsilon_2)F_0 + \varepsilon_1\delta_s + \varepsilon_2\delta_t)|_{\varepsilon_1=0,\varepsilon_2=0}.$$

We only expand it to the second order term because it is enough for our problem. $T^{'}_{F_0}(s)$ and $T^{''}_{F_0}(s,t)$ are known as influence functions. After evaluating the expansion at $\varepsilon = 1$, we have

$$T(F) - T(F_0) := T^{'}_{F_0}(F) + \frac{1}{2}T^{''}_{F_0}(F) + error.$$

Handling the error term can be technically difficult, but the above expansion generally leads to an asymptotically correct approximation (Serfling 1980).

In our problem, the distribution F_0 is a normal distribution with the same mean and variance as the distribution F. Let

$$f_\varepsilon = \phi(x,\mu_f,V_f) + \varepsilon(f(x) - \phi(x,\mu_f,V_f)) := \phi + \varepsilon\delta,$$

where ϕ is the normal density function. Thus

$$J_{f_\varepsilon} = V_{f_\varepsilon}^{1/2}\Big(\int \frac{\nabla_x f_\varepsilon \cdot \nabla_x f_\varepsilon^T}{f_\varepsilon} dx\Big) V_{f_\varepsilon}^{1/2}.$$

Because $\int x\delta dx = 0$ and $\int xx^T\delta dx = 0_d$, we have

$$\begin{aligned} V_{f_\varepsilon} &= \int xx^T(\phi+\varepsilon\delta)dx - (\int x(\phi+\varepsilon\delta)dx)(\int x(\phi+\varepsilon\delta)dx)^T \\ &= \int xx^T\phi dx - (\int x\phi dx)(\int x\phi dx)^T \\ &= V_f, \forall\varepsilon\in[0,1]. \end{aligned}$$

So we only need to consider $\int \frac{\nabla_x f_\varepsilon \cdot \nabla_x f_\varepsilon^T}{f_\varepsilon} dx$:

$$
\begin{aligned}
& \int \frac{\nabla_x f_\varepsilon \cdot \nabla_x f_\varepsilon^T}{f_\varepsilon} dx \\
= & \int \frac{(\nabla_x \phi + \varepsilon \nabla_x \delta)(\nabla_x \phi + \varepsilon \nabla_x \delta)^T}{\phi + \varepsilon \delta} dx \\
= & \int \Big(\nabla_x \phi \nabla_x \phi^T + \varepsilon (\nabla_x \delta \nabla_x \phi^T + \nabla_x \delta \nabla_x \phi^T) + \varepsilon^2 \nabla_x \delta \nabla_x \delta^T \Big) \\
& \cdot \phi^{-1} (1 + \varepsilon \frac{\delta}{\phi})^{-1} dx \\
= & \int \Big(\nabla_x \phi \nabla_x \phi^T + \varepsilon (\nabla_x \delta \nabla_x \phi^T + \nabla_x \delta \nabla_x \phi^T) + \varepsilon^2 \nabla_x \delta \nabla_x \delta^T \Big) \\
& \cdot \phi^{-1} (1 - \varepsilon \frac{\delta}{\phi} + \varepsilon^2 \Big(\frac{\delta}{\phi} \Big)^2 + o(\varepsilon^2)) dx \\
= & \int \phi^{-1} \nabla_x \phi \nabla_x \phi^T dx \\
& + \varepsilon \int \phi^{-1} \Big(\nabla_x \delta \nabla_x \phi^T + \nabla_x \delta \nabla_x \phi^T - \nabla_x \phi \nabla_x \phi^T \frac{\nabla}{\phi} \Big) dx \\
& + \varepsilon^2 \int \phi^{-1} [\nabla_x \phi \nabla_x \phi^T (\frac{\delta}{\phi})^2 + \nabla_x \phi \nabla_x \phi^T - (\nabla_x \delta \nabla_x \phi^T + \nabla_x \delta \nabla_x \phi^T) \frac{\delta}{\phi}] dx \\
& + o(\varepsilon^2) \\
= & V_f^{-1} \int (x - \mu_f)(x - \mu_f)^T \phi dx V_f^{-1} \\
& - \varepsilon [\int \nabla_x \delta (x - \mu_f)^T dx V_f^{-1} + \int V_f^{-1} (x - \mu_f) \nabla_x \delta^T dx + 0] \\
& + \varepsilon^2 \int \phi^{-1} (\nabla_x \phi \frac{\delta}{\phi} - \nabla_x \delta)(\nabla_x \phi \frac{\nabla}{\phi} - \nabla_x \delta)^T dx \\
& + o(\varepsilon^2) \\
= & V_f^{-1} + \varepsilon^2 \int \phi^{-1} (V_f^{-1} (x - \mu_f) + \nabla_x f)(V_f^{-1} (x - \mu_f) + \nabla_x f)^T dx + o(\varepsilon^2). \quad (3.6)
\end{aligned}
$$

Evaluating the expansion at $\varepsilon = 1$ derives

$$\begin{aligned} J_f &= I_d + V_f^{\frac{1}{2}} \int \phi^{-1}(V_f^{-1}(x-\mu_f) + \nabla_x f)(V_f^{-1}(x-\mu_f) + \nabla_x f)^T dx V_f^{\frac{1}{2}} + o(1) \\ &:= I_d + Q_f + o(1). \end{aligned} \tag{3.7}$$

In order to use this as an approximation, the magnitude of the error term would need to be checked. In our problem, it is not necessary because the statistic Q_f is also a non-normality measure of f that we can use it directly. The Q_f is non-negative and only when f is normal with the mean μ_f and variance V_f, $V^{-1}(x-\mu_f) + \nabla_x f = 0$. Compared to the Fisher information matrix J_f, Q_f can be easily estimated by the kernel method, because the denominator becomes the normal density ϕ. We will investigate the projection index Q_f in Chapter 5. In Chapter 7, the new projection index will be applied to the simulated and real data sets.

Chapter 4

Standardized Fisher Information for f_2

4.1 Overview

In this chapter, we will present the explicit form of J_{f_2} for the normal mixture model in Section 4.1. We study the explicit form of J_{f_2} for this model for at least three reasons: first, finite mixture models have been widely used in a great variety of fields, where it is common to assume there is a finite mixture structure. Also the normal mixture model is the most important and popular of the mixture models because of the flexibility of its density shape and usefulness in practice. Second, the normal mixture model provides a simple example for the study of the power of non-normal projection pursuit. For two-component normal mixture models, one can determine the least normal direction without any computation. So normal mixture model can be used to check if a projection pursuit method is effective. Finally, when we use the kernel method to estimate $f(x)$, and the normal kernel is employed, $\hat{f}$ and $\hat{f}^2$ are, in effect, mixtures of normal densities.

In section 4.2, kernel method is used to estimate the component piece of J_{f_2}. Putting them together provides a consistent estimator $\hat{J}_{f_2} = J_{\hat{f}_2}$. The consistency requires two conditions: $n \to \infty$ and that the smoothing parameter H decreases to zero at

a suitable rate. The selection of the smoothing parameter H is not a easy job in practice. Our simulation study shows that the asymptotic distributions based on vanishing H assumption does not provide a good result.

In fact, the kernel estimator $\hat{J}_{f_2} = J_{\hat{f}_2}$ directly better measures the non-normality of the kernel-smoothed distribution $f_2^* = (f^*(x))^2(x)/(f^*(y))^2 dy$, where

$$f^*(x) = \int f(y) \frac{1}{|H|} K_d(H^{-1}(x-y)) dy.$$

Note that $\hat{J}_{f_2}$ is also a consistent estimator of $J_{f_2^*}$ without H going to zero. This double-smoothing idea (smooth the model and the data with the same kernel) has big theoretical advantage. By a Von-Mises expansion, $\hat{J}_{f_2}$ can be approximated by quadratic distance between the empirical distribution and the corresponding normal distribution. Asymptotic properties will be studied in the Chapter 6.

4.2 J_{f_2} for Normal Mixture Models

Suppose $f(x)$ is the density function of normal mixture model:

$$f(x) = f(x_1, ...x_d) = \Sigma_{i=1}^m \pi_i \phi(x, \mu_i, \Sigma_i),$$

where ϕ is the probability density function of d-variate normal distribution with $x, \mu_i \in R^d$ and Σ_i is $(d \times d)$ covariance matrix. First we give the mean and variance for mixture of normals:

$$E_f(X) = \Sigma_{i=1}^m \pi_i \mu_i = (\mu_1, ..., \mu_n)^T \pi, \tag{4.1}$$

$$\begin{aligned} V_f &= (Cov(X_i, X_j)) \\ &= \Sigma_{i=1}^m \pi_i \Sigma_i + \Sigma_{i=1}^m \pi_i \mu_i \mu_i^T \\ &\quad -(\Sigma_{k=1}^m \pi_k \mu_k)(\Sigma_{k=1}^m \pi_k \mu_k)^T. \end{aligned} \tag{4.2}$$

We will calculate all terms of (3.5). The following result will be used many times.

PROPOSITION 3.

$$\phi(x, \mu_k, \Sigma_k)\phi(x, \mu_l, \Sigma_l) = \phi(\mu_k, \mu_l, \Sigma_k + \Sigma_l)\phi(x, \mu_{kl}, \Sigma_{kl}), \tag{4.3}$$

where

$$\Sigma_{kl} = \Sigma_k(\Sigma_k + \Sigma_l)^{-1}\Sigma_l = \Sigma_l(\Sigma_k + \Sigma_l)^{-1}\Sigma_k,$$

$$\mu_{kl} = \Sigma_l(\Sigma_k + \Sigma_l)^{-1}\mu_k + \Sigma_k(\Sigma_k + \Sigma_l)^{-1}\mu_l.$$

PROPOSITION 4.

$$\int f^2 dx = \sum_k \sum_l \pi_k \pi_l \phi(\mu_k, \mu_l, \Sigma_k + \Sigma_l). \tag{4.4}$$

PROPOSITION 5.

$$\begin{aligned} &\int \nabla_x f \cdot \nabla_x f^T dx \\ = &\sum_k \sum_l \pi_k \pi_l \phi(\mu_k, \mu_l, \Sigma_k + \Sigma_l)[(\Sigma_k + \Sigma_l)^{-1} \\ &+(\Sigma_k + \Sigma_l)^{-1}(\mu_k - \mu_l)(\mu_k - \mu_l)^T(\Sigma_k + \Sigma_l)^{-1}] \end{aligned} \tag{4.5}$$

PROPOSITION 6. *If f is the density of mixture model of m normals, then $f_2(x) = f^2(x)/\int f^2(y)dy$ is the density of a mixture of normals with m^2 uncombined components.The component proportions $\{\pi_{kl}\}$ are proportional to*

$$\pi_k \pi_l \phi(\mu_k, \mu_l, \Sigma_k + \Sigma_l).$$

the component means and variances are μ_{kl}, and Σ_{kl}.

Using the forms (4.1) and (4.2) for f_2, we can get the explicit form of the covariance matrix V_{f_2} :

$$\begin{aligned} V_{f_2} &= \sum_k \sum_l \pi_{kl}\Sigma_{kl} + \sum_k \sum_l \pi_{kl}\mu_{kl}\mu_{kl}^T \\ &\quad -(\sum_k \sum_l \pi_{kl}\mu_{kl})(\sum_k \sum_l \pi_{kl}\mu_{kl})^T. \end{aligned} \tag{4.6}$$

Plugging (4.4), (4.5) and (4.6) into the left part of (3.5), we get an explicit form for J_{f_2}.

4.3 Kernel Estimator of J_{f_2}

The kernel density estimator is a very popular non-parametric estimator of $f(x)$. Let

$$\hat{f}_H(x) = \sum \frac{1}{n|H|} K_d(H^{-1}(x - X_i)),$$

where K_d is the kernel function, which is usually a symmetric probability density function and H is the bandwidth, also called the smoothing parameter. The kernel estimator is a sum of "bumps" placed at the observations(Silverman 1986). The shape of bumps is determined by kernel function and the bandwidth. The quality of a kernel estimator

generally depends less on the shape of the K than on the value of its bandwidth H. In this section, we use the normal density as the kernel because of its computational advantage. Let

$$K_d(u) = (\frac{1}{2\pi})^{\frac{d}{2}} exp(-\frac{u^T u}{2}).$$

Substituting the kernel estimator $\hat{f}_H(x)$ for $f(x)$ in J_{f_2}, and applying equation (4.3), we can get the explicit form of the estimator of J_{f_2} :

$$\hat{J}_{f_2} = J_{\hat{f}_2} = \frac{V_{\hat{f}_2}^{1/2} \int \nabla_x \hat{f}_H(x) \cdot \nabla_x \hat{f}_H(x)^T dx V_{\hat{f}_2}^{1/2}}{\int \hat{f}_H^2(x) dx},$$

where

$$\hat{f}_2(x) = \hat{f}_H^2(x) / \int \hat{f}_H^2(y) dy,$$

$$\int \hat{f}_H^2(x) dx = \frac{1}{n^2} \sum_{i,j} \phi_d(X_i, X_j, 2H^2),$$

$$\begin{aligned}
& \int \nabla_x \hat{f}_H(x) \cdot \nabla_x \hat{f}_H^T(x) dx \\
= \; & \frac{1}{n^2} \sum_{i,j} \int (\frac{1}{2\pi})^d \frac{1}{|H|^2} exp\Big(-\frac{1}{2}[(X_i - x)^T H^{-2}(X_i - x) + (X_j - x)^T H^{-2}(X_j - x)]\Big) \\
& H^{-2}(X_i - x)(X_j - x)^T H^{-2} \\
= \; & \frac{1}{n^2} \sum_{i,j} \phi(X_i - X_j, 0, 2H^2) \phi(x, \frac{X_i + X_j}{2}, \frac{H^2}{2}) H^{-2}(X_i - x)(X_j - x)^T H^{-2} \\
= \; & \frac{1}{n^2} \sum_{i,j} \phi(X_i - X_j, 0, 2H^2) H^{-2}(\frac{H^2}{2} - \frac{1}{4}(X_i - X_j)(X_i - X_j)^T) H^{-2} \\
= \; & \frac{1}{n^2} \sum_{i,j} \phi(X_i - X_j, 0, 2H^2) \Big(\frac{H^{-2}}{2} - \frac{H^{-2}}{4}(X_i - X_j)(X_i - X_j)^T H^{-2}\Big),
\end{aligned}$$

$$V_{\hat{f}_2} = \frac{\int xx^T \hat{f}_H^2(x)(x)dx}{\int \hat{f}_H^2(x)(x)dx} - \Big(\frac{\int x\hat{f}_H(x)^2(x)dx}{\int \hat{f}_H^2(x)(x)dx}\Big)\Big(\frac{\int x\hat{f}_H^2(x)(x)dx}{\int \hat{f}_H^2(x)(x)dx}\Big)^T,$$

$$\int xx^T \hat{f}_H^2(x)(x)dx = \frac{1}{n^2}\sum_{i,j}\phi_d(X_i, X_j, 2H^2)\Big(\frac{H^2}{2} + \frac{(X_i + X_j)(X_i + X_j)^T}{4}\Big),$$

and

$$\int x\hat{f}_H^2(x)dx = \frac{1}{n^2}\sum_{i,j}\phi_d(X_i, X_j, 2H^2)\Big(\frac{X_i + X_j}{2}\Big).$$

All above pieces estimators are V-statistics, so consistent under some assumptions involving a large sample size n, and a vanishing bandwidth H. In practice, it's important and very difficult to choose the most appropriate bandwidth as a value that is too small or too large will not be informative. Small bandwidth leads to very spiky estimate, which makes the estimated pieces unreliable, even when the sample is really from the true distribution. And large bandwidth leads to over-smoothing, which will make non-normal data look more normal.

Bowman (1995) proposed a criteria to select an optimal bandwidth: mean integrated squared error (MISE)

$$\int E(\hat{f}_H(x) - f(x))^2 dx.$$

Usually it needs to be calculated by numerical integration. However, for the normal density kernel, the following bandwidth is optimal for a normal density f according to the MISE criteria:

$$H_{op} = \Big(\frac{4}{d+2}\Big)^{\frac{1}{d+4}}\Sigma^{1/2}n^{-\frac{1}{d+4}},$$

where Σ can be estimated by the sample covariance matrix.

The kernel density estimator $\hat{f}_H$ is biased for any fixed bandwidth. The true mean of $\hat{f}_H(x)$ is $f^*(x)$:

$$f^*(x) = \int f(y) \cdot \frac{1}{|H|} K_d(H^{-1}(x-y))dy.$$

The density transformation $f \rightarrow f^*$also preserves the normality of $f(x)$, when the Gaussian kernel K is used in the kernel density estimator.

PROPOSITION 7. *If $f \sim N(\mu_f, V_f)$, then*

$$f^*(x) = \int f(y) \cdot \frac{1}{|H|} K_d(H^{-1}(x-y))dy \sim N(\mu_f, V_f + H^2).$$

If we estimate all terms of J_{f_2} using the kernel method, the expectations of these terms can be expressed by f^* :

$$\begin{aligned}
E_f \int \hat{f}_H^2(x)dx &= \int E_f\Big(\frac{1}{n|H|}\sum_i K_d(H^{-1}(x-X_i))\Big)^2 dx \\
&= \frac{1}{n^2}\sum_{i,j}\Big(\phi(x, X_i, H^2)\phi(x, X_j, H^2)\Big)dx \\
&= \frac{1}{n^2}\sum_{i\neq j}\int E_f\phi(x, X_i, H^2)\phi(x, X_j, H^2)dx + \frac{1}{n^2}\sum_i \int E_f\phi^2(x, X_i, H^2)dx \\
&= \frac{1}{n^2}\sum_{i\neq j} E_f\phi(x, X_i, H^2)E\phi_d(x, X_j, H^2) + \frac{1}{n^2}\sum_i \int E_f\phi^2(x, X_i, H^2)dx \\
&= (1-\frac{1}{n})\int (f^*)^2(x)dx + \frac{1}{n}E_f\int \phi^2(x, Y, H^2)dx \\
&= (1-\frac{1}{n})\int (f^*)^2(x)dx + \frac{1}{n|H|}\phi(0, 0, 2I_d);
\end{aligned}$$

$$
\begin{aligned}
& E_f \int \nabla_x \hat{f}_H(x) \cdot \nabla_x \hat{f}_H^T(x) dx \\
= & \frac{1}{n^2} \sum_{i \neq j} \int \Big(E_f \phi(x, X_i, H^2) H^{-2} (X_i - x) E_f \phi(x, X_j, H^2) (X_j - x)^T H^{-2} \Big) dx \\
+ & \frac{1}{n^2} \sum_{i} \int \Big(E_f \phi^2(x, X_i, H^2) H^{-2} (X_i - x)(X_i - x)^T H^{-2} \Big) dx \\
= & (1 - \frac{1}{n}) \int \nabla_x f^*(x) (\nabla_x f^*(x))^T dx + \frac{1}{n} E_f \int \Big(\phi^2(x, X_i, H^2) H^{-2} (X_i - x)(X_i - x)^T H^{-2} \Big) dx \\
= & (1 - \frac{1}{n}) \int \nabla_x f^*(x) (\nabla_x f^*(x))^T dx + \frac{1}{2n|H|} \phi(0, 0, 2I_d) H^{-2};
\end{aligned}
$$

$$
\begin{aligned}
E_f \int x x^T \hat{f}_H^2(x) dx &= \frac{1}{n^2} \sum_{i,j} \int x x^T E_f \Big(\phi(x, X_i, H^2) \phi(x, X_j, H^2) \Big) dx \\
&= \frac{1}{n^2} \sum_{i \neq j} \int x x^T E_f \phi(x, X_i, H^2) E_f \phi(x, X_j, H^2) dx \\
&+ \frac{1}{n^2} \sum_{i} E_f \int x x^T \phi^2(x, X_i, H^2) dx \\
&= (1 - \frac{1}{n}) \int x x^T (f^*)^2(x) dx + \frac{1}{n} E_f \int x x^T \phi(x, Y, H^2/2) \phi(0, 0, 2H^2) dx \\
&= (1 - \frac{1}{n}) \int x x^T (f^*)^2(x) dx + \frac{1}{n} \phi(0, 0, 2H^2) E_f [H^2/2 + Y Y^T]
\end{aligned}
\tag{4.7}
$$

$$
\begin{aligned}
E_f \int x\hat{f}_H^2(x)dx &= \frac{1}{n^2}\sum_{i,j}\int xE_f\Big(\phi(x,X_i,H^2)\phi(x,X_j,H^2)\Big)dx \\
&= \frac{1}{n^2}\sum_{i\neq j}\int xE_f\phi(x,X_i,H^2)E_f\phi_d(x,X_j,H^2)dx + \frac{1}{n^2}\sum_{i}\int E_f x\phi^2(x,X_i,H^2)dx \\
&= (1-\frac{1}{n})\int x(f^*)^2(x)dx + \frac{1}{n}E_f\int x\phi^2(x,Y,H^2)dx \\
&= (1-\frac{1}{n})\int x(f^*)^2(x)dx + \frac{1}{n|H|}\phi(0,0,2I_d)E_fY.
\end{aligned}
$$

For fixed H, putting all these kernel estimators together and letting $n \to \infty$ derives a new standardized Fisher information matrix:

$$
\frac{V_{f_2^*}^{1/2}\int \nabla_x f^* \cdot (\nabla_x f^*)^T dx V_{f_2^*}^{1/2}}{\int (f^*)^2(x)dx}, \tag{4.8}
$$

where

$$
f_2^* = (f^*)^2 / \int (f^*)^2(x)dx,
$$

$$
V_{f_2^*} = \frac{\int yy^T(f^*)^2(y)dy}{\int (f^*)^2dx} - \Big(\frac{\int y(f^*)^2(y)dy}{\int (f^*)^2dx}\Big)\Big(\frac{\int y(f^*)^2(y)dy}{\int (f^*)^2dx}\Big)^T.
$$

The above matrix is just the standardized Fisher information matrix of f_2^*. The matrix inequality still holds since f^* is also a density function:

$$
\frac{V_{f_2^*}^{1/2}\int \nabla_x f^* \cdot (\nabla_x f^*)^T dx V_{f_2^*}^{1/2}}{\int (f^*)^2(x)dx} := J_{f_2^*} \geq \frac{1}{4}I_d. \tag{4.9}
$$

So, in order to measure the non-normality of the original density $f(x)$, we first smooth it using a normal kernel, and then transform the smoothed density $f^*(x)$: $f_2^*(x) =$

$(f^*)^2(x)/\int (f^*)^2(y)dy$, finally construct the standardized Fisher information matrix for f_2^*.

In the following part, we use $J_{f_2^*}$ to study the asymptotic property of the estimator $\hat{J}_{f_2}$. Compared to J_{f_2}, this new Fisher information matrix $J_{f_2^*}$ has some good properties: first, for any fixed H, $f^*(x)$ is the mean of kernel density estimator $\hat{f}_H(x)$. The estimator $\hat{J}_{f_2}$ is closer to $J_{f_2^*}$ than J_{f_2}. Second, it is easy to construct the unbiased estimators for all pieces of $J_{f_2^*}$ by U-statistics. Putting all corresponding V-statistics together derives $\hat{J}_{f_2}$. The consistency of the estimator $\hat{J}_{f_2}$ holds for every fixed H. Finally, the asymptotic distribution of $\hat{J}_{f_2}$ can be found using the quadratic distance method in Lindsay et al. (2006).

4.4 U-statistic Estimator of $J_{f_2^*}$

In practice, if the distribution is known and J_{f_2} is known, we use J_{f_2} as our measure of non-normality. For the real data set, we use $\hat{J}_{f_2}$ to estimate J_{f_2}. In order to study asymptotic properties, we fix the smoothing parameter H and treat $\hat{J}_{f_2}$ as the estimator of $J_{f_2^*}$. In this section, we will introduce how to estimate $J_{f_2^*}$ using U-statistics.

First we rename all pieces of the matrix $J_{f_2^*}$ to have a simple expression.

Definition 2.

$$J_{f_2^*} := \frac{V_{f_2^*}^{1/2} \int \nabla_x f^* \cdot (\nabla_x f^*)^T dx V_{f_2^*}^{1/2}}{\int (f^*)^2(x)dx},$$

$$\theta_0 := \int (f^*)^2(x)dx,$$

$$\theta_1 := \int \nabla_x f^* \cdot (\nabla_x f^*)^T dx,$$

$$\theta_2 := \int yy^T (f^*)^2(y)dy,$$

$$\theta_3 := \int y(f^*)^2(y)dy.$$

So

$$J_{f_2^*} = \frac{V_{f_2^*}^{1/2} \int \nabla_x f^* \cdot (\nabla_x f^*)^T dx V_{f_2^*}^{1/2}}{\int (f^*)^2(x)dx} = \Big(\frac{\theta_2}{\theta_0} - \Big(\frac{\theta_3}{\theta_0}\Big)\Big(\frac{\theta_3}{\theta_0}\Big)^T\Big)^{\frac{1}{2}} \theta_1 \Big(\frac{\theta_2}{\theta_0} - \Big(\frac{\theta_3}{\theta_0}\Big)\Big(\frac{\theta_3}{\theta_0}\Big)^T\Big)^{\frac{1}{2}}$$

We will estimate all $\theta_i, i = 0, 1, 2, 3$.

PROPOSITION 8. *Suppose $\{X_1, ..., X_n\}$is a random sample from f, then the U-statistic*

$$\begin{aligned} U_0 &= \frac{1}{n(n-1)|H|^2} \sum_{i \neq j} \int K(H^{-1}(x - X_i))K(H^{-1}(x - X_j))dx \\ &= \frac{1}{n(n-1)} \sum_{i \neq j} \phi(X_i, X_j, 2H^2) \\ &:= \frac{1}{n(n-1)} \sum_{i \neq j} h_0(X_i, X_j) \end{aligned}$$

is an unbiased estimator of θ_0, where $h_0 = \phi(X_i, X_j, 2H^2)$ is the kernel function of the U-statistic.

The projection of U_0 is

$$P_0 = \frac{2}{n} \sum_{i=1}^{n} (h_0^*(X_i) - \theta_0),$$

where $h_0^*(X_i) = Eh_0(X_i, X_j|X_i) = \phi(X_i, \mu_f, A)$, $A = V_f + 2H^2$.

When f is a normal density,

$$\theta_0 = \int \phi(x, \mu_f, A)\phi(x, \mu_f, V_f)dx = \phi(0, 0, 2V_f + 2H^2).$$

Similarly, we can find the U-statistics for other pieces.

PROPOSITION 9. *Suppose* $\{X_1, ..., X_n\}$*is a random sample from* f*, then the U-statistic*

$$\begin{aligned} U_1 &= \frac{1}{n(n-1)|H|^2}\sum_{i\neq j}\int \frac{\partial K(H^{-1}(x-X_i))}{\partial x}\frac{\partial K(H^{-1}(x-X_j))}{\partial x}dx \\ &= \frac{1}{n(n-1)|H|^2}\sum_{i\neq j}\int K(H^{-1}(x-X_i))K(H^{-1}(x-X_j))H^{-2}(x-X_i)(x-X_j)^T H^{-2}dx \\ &= \frac{1}{n(n-1)}\sum_{i\neq j}\phi(X_i, X_j, 2H^2)\Big(\frac{H^{-2}}{2} - \frac{H^{-2}}{4}(X_i-X_j)(X_i-X_j)^T H^{-2}\Big) \qquad (4.10) \\ &:= \frac{1}{n(n-1)}\sum_{i\neq j} h_1(X_i, X_j) \end{aligned}$$

is an unbiased estimator of θ_1*. The projection of* U_1 *is*

$$P_1 = \frac{2}{n}\sum_{i=1}^{n}(h_1^*(X_i) - \theta_1),$$

where $h_1^*(X_i) = Eh_1(X_i, X_j|X_i) = \phi(X_i, \mu_f, A)(A^{-1} - A^{-1}(X_i-\mu_f)(X_i-\mu_f)^T A^{-1})$.

When f *is a normal density,*

$$\theta_1 = \theta_0(2V_f + 2H^2)^{-1}.$$

PROPOSITION 10. *Suppose* $\{X_1, ..., X_n\}$*is a random sample from* f*, then the U-statistic*

$$\begin{aligned} U_2 &= \frac{1}{n(n-1)|H|^2}\sum_{i\neq j}\int xx^T K(H^{-1}(x-X_i))K(H^{-1}(x-X_j))dx \\ &= \frac{1}{n(n-1)}\sum_{i\neq j}\phi(X_i, X_j, 2H^2)\Big(\frac{H^2}{2} + \frac{(X_i+X_j)(X_i+X_j)^T}{4}\Big) \\ &:= \frac{1}{n(n-1)}\sum_{i\neq j} h_2(X_i, X_j) \end{aligned}$$

is an unbiased estimator of $\int xx^T(f^*(x))^2dx$.

The projection of U_2 *is*

$$P_2 = \frac{2}{n}\sum_{i=1}^{n}(h_2^*(X_i) - \theta_2),$$

where

$$\begin{aligned} h_2^*(X_i) &= Eh_2(X_i, X_j|X_i) \\ &= \phi(X_i, 0, A)(\frac{(2H^2 + (\Sigma_f^{-1} + (2H^2)^{-1})^{-1})}{4} \\ &+ \frac{(I + (V_f + 2H^2)^{-1}V_f)X_iX_i^T(I + (V_f + 2H^2)^{-1}V_f)}{4}). \end{aligned}$$

When f *is a normal density,*

$$\theta_2 = \phi(0, 0, 2V_f + 2H^2)[\frac{2V_f + 2H^2}{4} + \mu_f\mu_f^T] = \theta_0\frac{2V_f + 2H^2}{4}.$$

PROPOSITION 11. *Suppose* $\{X_1, ..., X_n\}$*is a random sample from* f*, then the U-statistic*

$$\begin{aligned} U_3 &= \frac{1}{n(n-1)|H|^2}\sum_{i\neq j}\int xK(H^{-1}(x - X_i))K(H^{-1}(x - X_j))dx \\ &= \frac{1}{n(n-1)}\sum_{i\neq j}\phi(X_i, X_j, 2H^2)\frac{(X_i + X_j)}{2} \\ &:= \frac{1}{n(n-1)}\sum_{i\neq j}h_3(X_i, X_j) \end{aligned}$$

is an unbiased estimator of $\int x(f^*(x))^2dx$.

The projection of U_3 *is*

$$P_3 = \frac{2}{n}\sum_{i=1}^{n}(h_3^*(X_i) - \theta_3),$$

where

$$h_3^*(X_i) = Eh_3(X_i, X_j|X_i) = \phi(X_i, \mu_f, A)\frac{1}{2}(X_i + V_f A^{-1} X_i + 2H^2 A^{-1}\mu_f).$$

When f is a normal density,

$$\theta_3 = \theta_0 \mu_f.$$

After combining the above four U-statistic estimators, we get the U-estimator of the standardized Fisher information matrix $J_{f_2^*}$:

$$\begin{aligned} \hat{U}^* &= \frac{(\frac{U_2}{U_0} - (\frac{U_3}{U_0})(\frac{U_3}{U_0})^T)^{\frac{1}{2}} U_1 (\frac{U_2}{U_0} - (\frac{U_3}{U_0})(\frac{U_3}{U_0})^T)^{\frac{1}{2}}}{U_0} \\ &= (U_2 U_0 - U_3 U_3^T)^{\frac{1}{2}} U_1 (U_2 U_0 - U_3 U_3^T)^{\frac{1}{2}} / (U_0^3). \end{aligned}$$

When X is standardized: $\mu_f = 0$, $V_f = I_d$, θ_3 is zero. Thus, we have

$$\begin{aligned} \hat{U}^* &= \frac{(\frac{U_2}{U_0})^{\frac{1}{2}} U_1 (\frac{U_2}{U_0})^{\frac{1}{2}}}{U_0} \\ &= (U_2)^{\frac{1}{2}} U_1 (U_2)^{\frac{1}{2}} / (U_0^2). \end{aligned}$$

When f is normal and n goes to infinity, the U-estimator converges in probability to $\frac{1}{4} I_d$. Note that the matrix inequality $\hat{U}^* \geq \frac{1}{4} I_d$ does not hold for this estimated matrix $\hat{U}^*$, because the four U-statistic estimators are unbiased and $\hat{U}^*$ is not a Fisher information matrix of some distribution. If we use the corresponding V-statistics to estimate the θ_i, we get $\hat{J}_{f_2} = J_{\hat{f}_2}$. The inequality then holds because the estimated matrix is just the

standardized Fisher information matrix of the empirical distribution of f_2^*:

$$\begin{aligned}\hat{V}^* = \hat{J}_{f_2} &= \frac{(\frac{V_2}{V_0} - (\frac{V_3}{V_0})(\frac{V_3}{V_0})^T)^{\frac{1}{2}} V_1 (\frac{V_2}{V_0} - (\frac{V_3}{V_0})(\frac{V_3}{V_0})^T)^{\frac{1}{2}}}{V_0} \\ &= (V_2V_0 - V_3V_3^T)^{\frac{1}{2}} V_1 (V_2V_0 - V_3V_3^T)^{\frac{1}{2}}/(V_0^3) \geq \frac{1}{4} I_d,\end{aligned}$$

where V_i is the corresponding V-statistic of $U_i, i = 0, 1, 2, 3$. $\hat{V}^*$ is still a consistent estimate of $J_{f_2^*}$. In Chapter 6, we will study the asymptotic distribution of $\hat{V}^*$ by expanding it at a normal distribution with the same mean and covariance as f. The second order of the expansion is a kernel-based quadratic distance(Lindsay et al. 2006).

4.5 Eigenanalysis of J_{f_2}

In this section, by extending the results for J_f, we will study the transformation property of J_{f_2}, and show how to use an eigenanalysis of J_{f_2} to find the least normal projections.

PROPOSITION 12. *Let A be a $d \times d$ nonsingular matrix and $Y = AX$. Then*

$$J_{g_2} = V_{g_2}^{1/2} A^{-T} V_{f_2}^{-1/2} J_{f_2} V_{f_2}^{-1/2} A^{-1} V_{g_2}^{1/2}, \tag{4.11}$$

where $g_2(y) = g^2(y)/\int g^2(s)ds$, $g(y)$ is the density of Y, V_{g_2} is the covariance matrix of the distribution with the density g_2; $f_2(x) = f^2(x)/\int f^2(s)ds$, $f(x)$ is the density of X, V_{f_2} is the covariance matrix of the distribution with the density f_2.

First we standardize the vector X. Let $A = V_{f_2}^{-\frac{1}{2}}$, and $Y = V_{f_2}^{-\frac{1}{2}} X$. This standardization does not change the Fisher information. Suppose $g(y)$ is the density of Y. Then

$$V_{g_2} = V_{f_2}^{-\frac{1}{2}} V_{f_2} V_{f_2}^{-\frac{1}{2}} = I_d;$$

and

$$J_{g_2} = V_{g_2}^{1/2} A^{-T} V_{f_2}^{-1/2} J_{f_2} V_{f_2}^{-1/2} A^{-1} V_{g_2}^{1/2} = J_{f_2}.$$

Secondly, we consider the orthogonal transformation of Y. Let

$$Z = (Z_1, ..., Z_d)^T = (\gamma_1^T Y, .., \gamma_d^T Y)^T := \Gamma Y,$$

where Γ is an orthogonal matrix. So the density of Z is $h(z) = g(\Gamma^T z)$, and

$$V_{h_2} = \Gamma V_{g_2} \Gamma^T = I_d.$$

Then, we have

$$\begin{aligned} J_{h_2} &= V_{h_2}^{1/2} \Gamma^{-T} V_{g_2}^{1/2} J_{g_2} V_{g_2}^{1/2} \Gamma^{-1} V_{h_2}^{1/2} \\ &= \Gamma J_{f_2} \Gamma^T \\ &= \left(\gamma_i^T J_{f_2} \gamma_j \right). \end{aligned}$$

Let $\lambda_1 \geq \lambda_2, ... \geq \lambda_d$ be the eigenvalues of the standardized Fisher information matrix of Y and $\gamma_1, ..., \gamma_d$ be the corresponding eigenvectors. By the similar analysis of J_f in section

3.3, we know that the projection $Z_1 = \gamma_1^T Y$ has the least conditional normality and λ_1 is the measure of its non-normality. Moreover $Z_i = \gamma_i^T Y$ has the least conditional normality in all projections which are "uncorrelated" to $Z_1, ...Z_{i-1}$. Note that here "uncorrelated" means that after the density square transformation, the transformed Z_i are uncorrelated because $V_{h_2} = I_d$.

The Fisher information matrix of the optimal orthogonal projection $Z = \Gamma Y$ also can be expressed diagonally by the eigenvalues of the Fisher information matrix of Y:

$$\begin{aligned} J_{h_2} = & \quad \frac{\int \nabla_z h(z)(\nabla_z h(z))^T dz}{\int h^2(z)dz} \\ = & \quad \left(\gamma_i^T J_{f_2} \gamma_j\right) \\ = & \quad \left(\gamma_i^T \lambda_j \gamma_j\right) \\ = & \quad Diag(\lambda_1, \lambda_2, ..., \lambda_d) \geq \frac{1}{4} I_d. \end{aligned}$$

When Z is normal, it becomes equality. So the eigenvalues of J_{h_2} have the lower bound $\frac{1}{4}$. Consider the ith diagonal term of J_Z:

$$\begin{aligned} & \frac{\int \left(\frac{\partial}{\partial z_i} h(z_1, ...z_d)\right)^2 h dz}{\int h^2(z)dz} \\ = & \int \Big(\frac{\int (\frac{\partial}{\partial z_i} h(z_i|z_{-i}))^2 dz_i}{\int h^2(z_i|z_{-i})dz_i} \Big) h^2(z_{-i}) \frac{\int h^2(z_i|z_{-i})dz_i}{\int h^2(z)dz} dz_{-i} \\ = & \int J^2_{Z_i|Z_{-i}} h^2(z_{-i}) \frac{\int h^2(z_i|z_{-i})dz_i}{\int h^2(z)dz} dz_{-i} \\ \geq & \int \frac{1}{4} h^2(z_{-i}) \frac{\int h^2(z_i|z_{-i})dz_i}{\int h^2(z)dz} dz_{-i} \\ = & \frac{1}{4}, \end{aligned}$$

where $J^2_{Z_i|Z_{-i}}$ is the Fisher information matrix for $Z_i|Z_{-i}$. When $\lambda_i = \frac{1}{4}$, $J_{Z_i|z_{-i}} = \frac{1}{4}, \forall z_{-i}$, which implies that $Z_i|z_{-i}$ is normal for any z_{-i}. So Z_i is a white noise variable, which is not only marginally normal, but also independent of the remaining variables. If $\lambda_i = \frac{1}{4}$, then $\lambda_j = \frac{1}{4}, j \geq i$, and Z_j are white noise variables. So, by applying an eigenanalysis of J_{f_2}, we thus not only can find the interesting projection using the largest eigenvalues and the corresponding eigenvectors, but also can discard directions as white noise when the eigenvalues are close to $\frac{1}{4}$.

The diagonal form of the Fisher information matrix J_{h_2} also provides a new test statistic to test the normality of the whole X:

$$\sum_{i=1}^{d} \lambda_i = trace(J_{h_2}) = trace(J_{f_2}).$$

The asymptotic distribution will be discussed in Chapter 6.

Chapter 5

Von-Mises Approximation Q_f

In Chapter 3, we showed that the eigenanalysis of the Fisher information matrix J_f can provide all solution projections. We also found a new projection index Q_f by expanding the Fisher information matrix J_f:

$$Q_f = V_f^{\frac{1}{2}} \cdot \int \phi^{-1}(V_f^{-1}(x-\mu_f) + \nabla_x f)(V_f^{-1}(x-\mu_f) + \nabla_x f)^T dx \cdot V_f^{\frac{1}{2}}.$$

Compared to the Fisher information, the projection index Q_f is easy to calculate and estimate, because the denominator is a normal density function, not f. We still can get the explicit form of Q_f for the normal mixture model by the similar computation in section 4.2. However, we do not list the results because the form is very complicated. We can use the kernel method to estimate Q_f by simulation.

5.1 Kernel Estimator for Q_f

In Chapter 6, we will show that the standardization leaves the index Q_f unchanged. So, without losing generality, we assume the data has been standardized:

$\mu_f = 0, V_f = I_d$. The standardization simplifies the index Q_f:

$$\begin{aligned} Q_f &= \int \phi^{-1}(xf(x) + \nabla_x f(x))(xf(x) + \nabla_x f(x))^T dx \\ &= \int xx^T f^2(x)\phi^{-1}(x,0,1)dx + \int xf(x)\phi^{-1}(x,0,1)\nabla_x f^T(x)dx \\ &\quad + \int \nabla_x f(x) \cdot x^T f(x)\phi^{-1}(x,0,1)dx + \int \nabla_x f(x)\nabla_x f(x)^T \phi^{-1}(x,0,1)dx. \end{aligned}$$

We replace the density function $f(x)$ with the kernel estimator $\hat{f}_H(x)$ to get the estimator $Q_{\hat{f}}$.

Proposition 13.

$$\begin{aligned} &\int xx^T \hat{f}_H^2(x)\phi^{-1}(x,0,1)dx = \frac{1}{N^2}\sum_i\sum_j A_{ij}(\mu_{ij}\mu_{ij}^T + \Sigma) \\ &\int x\hat{f}_H(x)\phi^{-1}(x,0,1)\nabla_x \hat{f}_H^T(x)dx = \frac{1}{N^2}\sum_i\sum_j A_{ij}(\mu_{ij}\mu_{ij}^T + \Sigma - \mu_{ij}X_j^T)H^{-2} \\ &\int \nabla_x \hat{f}_H(x) \cdot x^T \hat{f}_H(x)\phi^{-1}(x,0,1)dx = \frac{1}{N^2}\sum_i\sum_j A_{ij}H^{-2}(\mu_{ij}\mu_{ij}^T + \Sigma - X_i\mu_{ij}^T) \\ &\int \nabla_x \hat{f}_H(x)\nabla_x \hat{f}_H(x)^T \phi^{-1}(x,0,1)dx = \frac{1}{N^2}\sum_i\sum_j A_{ij}H^{-2}((\mu_{ij} - X_i)(\mu_{ij} - X_j)^T + \Sigma)H^{-2}, \end{aligned}$$

where $A_{ij} = \frac{\phi(X_i, X_j, 2H^2)}{|I_d - H^2/2| \cdot \phi((X_i + X_j)/2, 0, I_d - H^2/2)}$, $\Sigma = (2H^2 - I_d)^{-1}$, $\mu_{ij} = (2I_d - H^2)^{-1}(X_i + X_j)$

The kernel estimator $Q_{\hat{f}}$ actually measures the non-normality of the smoothed density $f^*(x) = \int f(y)\frac{1}{H}K_d(H^{-1}(x-y))dy$. We will show that the kernel estimator $Q_{\hat{f}}$ is the V-statistic estimator of Q_{f^*}. It will help us to study the asymptotic properties of the estimator $Q_{\hat{f}}$.

First We will express the terms of $Q_{(f^*)}$ as the expectation form.

PROPOSITION 14. *Suppose Y, Z are independent variables with the same density function f, then the projection index $Q_{(f^*)}$ can be expressed as an expectation*

$$\begin{aligned} Q_{(f^*)} &= |I_d - \frac{H^2}{2}|^{-1} E_{Y,Z} \frac{\phi(Y, Z, 2H^2)}{\phi((Y+Z)/2, 0, I_d - H^2/2)} \\ &\quad \cdot\Big((2I_d - H^2)^{-1}(Y+Z)(Y+Z)^T(2I_d - H^2)^{-1} + \Sigma \\ &\quad +((2I_d - H^2)^{-1}(Y+Z)(Y+Z)^T(2I_d - H^2)^{-1} + \Sigma - (2I_d - H^2)^{-1}(Y+Z)Z^T)H^{-2} \\ &\quad +H^{-2}((2I_d - H^2)^{-1}(Y+Z)(Y+Z)^T(2I_d - H^2)^{-1} + \Sigma - Y(Y+Z)^T(2I_d - H^2)^{-1}) \\ &\quad +H^{-2}(\Sigma + ((2I_d - H^2)^{-1}(Y+Z) - Y)((2I_d - H^2)^{-1}(Y+Z) - Z)^T)H^{-2}\Big) \\ &:= E_{Y,Z} h(Y, Z) \end{aligned} \tag{5.1}$$

where $\Sigma = (2H^{-2} - I_d)^{-1}$.

So the V-statistic estimator of $Q_{(f^*)}$ is

$$\frac{1}{N^2} \sum_{i=1}^{n} \sum_{j=1}^{n} h(X_i, X_j) = Q_{\hat{f}}.$$

5.2 Eigenanalysis of Q_f

In the section, we present how to use an eigenanalysis of Q_f to find solution projections. The transformation equation for Q_f still holds, which is the base of eigenanalysis of Q_f.

PROPOSITION 15. *Let A be a $d \times d$ nonsingular matrix and $Y = AX$. Then*

$$Q_g = V_g^{1/2} A^{-T} V_f^{-1/2} Q_f V_f^{-1/2} A^{-1} V_g^{1/2}, \tag{5.2}$$

where $g(y)$ is the density of Y, V_g is the covariance matrix of g, $f(x)$ is the density of X, V_f is the covariance matrix of f.

We first standardize the variable $X : Y = V_f^{-\frac{1}{2}} X$. It leaves the index unchanged

$$Q_g = V_g^{1/2} A^{-T} V_f^{1/2} Q_f V_f^{1/2} A^{-1} V_g^{1/2} = Q_f.$$

Secondly, consider an orthogonal transformation $Z = \Gamma^T Y$.

$$\begin{aligned} Q_h &= V_h^{1/2} \Gamma^{-T} V_g^{1/2} Q_g V_g^{1/2} \Gamma^{-1} V_h^{1/2} \\ &= \Gamma Q_f \Gamma^T \\ &= \left(\gamma_i^T Q_f \gamma_j\right). \end{aligned}$$

The projection index $Z = \Gamma Y$ still can be expressed diagonally by the eigenvalues of Q_f:

$$\begin{aligned} Q_h = & \int \frac{\nabla_z h(z) (\nabla_z h(z))^T}{h(z)} dz \\ &= \left(\gamma_i^T Q_f \gamma_j\right) \\ &= \left(\gamma_i^T \lambda_j \gamma_j\right) \\ &= Diag(\lambda_1, \lambda_2, ..., \lambda_d). \end{aligned}$$

Similar to the analysis of J_{f_2}, the eigenanalysis of Q_f can provide the least normal projections. The projection $Z_1 = \gamma_1^T Y$ has the least normality in all projections, and the largest eigenvalue λ_1 is the non-normality measure. Because the projection Q_f is always

non-negative, the eigenvalues have the lower bound 0. When $\lambda_i = 0$, the corresponding projection Z_i is a white noise coordinate, which can be discarded in future study.

Chapter 6

Asymptotic Distribution of $\hat{V}^*$

6.1 Overview

In Chapter 4, we constructed estimators for the four terms of $J_{f_2^*}$ using U-statistics or V-statistics, and put them together to get a consistent estimator of $J_{f_2^*}$. Applying the U-statistic asymptotic theorem, we can get the limiting distribution for every term. The convergence rate is $\sqrt{n}$. Unfortunately, the convergence rate of the whole estimator $\hat{V}^*$ (or $\hat{U}^*$)is not $\sqrt{n}$ but n. In Section 6.2, we will use the projection formulas for U-statistics to show that the $\sqrt{n}$ term of $\hat{V}^*$ is exactly zero. In order to find the asymptotic distribution of $\hat{V}^*$, we need to expand all V-statistic estimators to the order n, and then plug them into $\hat{V}^*$ to find the n term. But this computation is really messy.

To find a more direct analysis, we look at the problem in another way. We may think $J_{f_2^*}$ as of a function of a distribution. Then $\hat{V}^*$ is the value of the function at the empirical distribution F_n. Because the empirical distribution converges to the true distribution, we may find the limiting distribution by expanding the function around the true distribution F_0, which is normal under the null hypothesis. Because the function reaches the minimum $\frac{1}{4}I_d$ at the normal distribution, the first Von-Mises derivative at the normal distribution is expected to be zero. So the second "derivative" should determine the limiting distribution. In section, we will express all terms of $J_{f_2^*}$ as the functions of true distribution, so the V-statistic estimated matrix $\hat{V}^*$ is a function of true

distribution. And then we will employ Von Mises expansion(Section 6.3) to show that the second order term of $\hat{V}^*$ is equivalent to a kernel-based matrix quadratic distance between the empirical distribution and the true distribution. The Von Mises expansion computation is much easier than U-statistic method, because it is very easy to find the functions of all terms, and we only need to use the derivative chain rule to find the expansion of $\hat{V}^*$.

In sections 6.4, we introduce the kernel-based quadratic distance theory of Lindsay et al.(2006) for $d = 1$. The limiting distribution of the scalar quadratic distance can be determined from the spectral decomposition of the distance kernel(Lindsay et al. 2006). Lindsay et al.(2006) showed that many important statistical distances have the kernel-based quadratic distance form. In the case $d = 1$, the limiting distribution is a weighted sum of chi-squared variables. The important approximations will be checked by simulation in Section 6.5. In Section 6.6, we apply the quadratic distance theory to the sum of eigenvalues of $\hat{V}^*$, which is equal to the trace of $\hat{V}^*$. Section 6.7 presents how to find the asymptotic distribution of the matrix $\hat{V}^*$ by using the multivariate spectral decomposition theorem and permutation invariant property of $\hat{V}^*$.

6.2 The Zero $\sqrt{n}$-term

In this section, we will introduce the U-statistics projection method, and apply it to show that the $\sqrt{n}$ term of $\hat{V}^*$ is exactly zero.

Let $X_1, ..., X_N$ be a random sample from the distribution f. Suppose $h(X_1, ..., X_r)$ is an unbiased estimator for the parameter of interest θ and permutation symmetric in

its r arguments. A U-statistic with kernel h is defined as

$$U = \frac{1}{C_N^r} \sum_{\beta} h(X_{\beta_1}, ..., X_{\beta_r}),$$

where the sum is taken over the set of all unordered subsets β of r different integers chosen from $\{1, ..., N\}$. The statistic U is also an unbiased estimator and has smaller variance than h, because U is the conditional expectation of h on the order statistic $\{X_{(1)}, ..., X_{(N)}\}$. The U-statistic is asymptotically normal given $Eh^2 < \infty$. But U-statistic is not the sum of independent random vectors. To get the asymptotic normality of a sequence of U-statistics, we can consider the projection of the U-statistic onto the set of all statistics of the form $\sum_{i=1}^N g_i(X_i)$. The projection of $U - \theta$ is given by

$$P = \sum_{i=1}^{N} E(U - \theta | X_i) = \frac{r}{N} \sum_{i=1}^{N} (h^*(X_i) - \theta),$$

where $h^*(x) = Eh(x, X_2, ...X_N)$.

The sequence of the projection P is asymptotically normal by the central limit theorem provided $E(h^*)^2(X) < \infty$. And the difference between $U - \theta$ and P is asymptotically negligible.

PROPOSITION 16. *If $Eh^2(X_1, ..., X_r) < \infty$, then $\sqrt{N}(U - \theta - P) = o_p(1)$. Consequently, the sequence $\sqrt{N}(U - \theta) \to N(0, r^2\zeta_1)$, where, with $X_1, ...X_r, X_{1'}, ..., X_{r'}$ denoting i.i.d. variables,*

$$\zeta_1 = cov(h(X_1, ..., X_r), h(X_{1'}, ..., X_{r'})).$$

By applying the result, we get the expansion of the four piece U-statistics of $\hat{U^*}$.

PROPOSITION 17. *Let U_i be the U-statistic estimator of θ_i, and P_i be the projection of U_i for a sample size n. We have*

$$U_i = \theta_i + P_i + o_p\Big(\frac{1}{\sqrt{n}}\Big);$$

$$P_i = O_p\Big(\frac{1}{\sqrt{n}}\Big),$$

and

$$P_i P_j = o_p\Big(\frac{1}{\sqrt{n}}\Big), i, j = 0, 1, 2, 3.$$

Plugging the above expansions into the U-statistic estimator $\hat{U^*}$ derives the expansion of $\hat{U^*}$ at $\sqrt{n}$. Because the difference between a U-statistic and corresponding V-statistic is $O_p(1/n)$, it is easy to find the expansion of $\hat{V^*}$.

PROPOSITION 18. *Let $H = hI_d, h > 0$. Then the $\sqrt{n}$ term of $\hat{U^*}$ is zero.*

$$\hat{U^*} - \frac{1}{4}I_d = o_p\Big(1/\sqrt{n}\Big);$$

$$\hat{V^*} - \frac{1}{4}I_d = o_p\Big(1/\sqrt{n}\Big).$$

In the proof, we assume X is standardized: $\mu_f = 0$, $V_f = I_d$. The result is true for any μ_f and V_f.

6.3 Von-Mises Expansion of $\hat{V^*}$

In Section 3.4.2, we introduced the Von Mises expansion of J_f to get a easily computable projection index Q_f. The idea is to treat a statistic as a distribution function

$T(F)$, and find the expansion at F_0 by calculating the influence functions:

$$T(F) - T(F_0) := T^{'}_{F_0}(F) + \frac{1}{2}T^{''}_{F_0}(F) + error.$$

In our problem, $F = \hat{F}_n$ and F_0 is normal distribution. The function $T(F) = \hat{V}^*$ reaches the minimum $\frac{1}{4}I_d$ at F_0. So we believe the first derivative $T^{'}_{F_0}(F)$ at F_0 is zeros. Thus, we need to expand $T(F)$ to the second order. We will find the distribution function and influence function for every piece of $\hat{V}^*$ and use the chain law to get the whole derivative.

Proposition 19.

$$\begin{aligned}
\theta_0 &= \int f^{*^2} dx = \int\int \Big(\int \frac{1}{|H|} K(H^{-1}(x-y)) \frac{1}{|H|} K(H^{-1}(x-z)) dx \Big) dF_0(y) dF_0(z) \\
&:= \int\int K_0^*(y,z) dF_0(y) dF_0(z) \\
&:= K_0^*(F_0, F_0) := T_0(F_0).
\end{aligned}$$

And then we can get the influence functions:

$$\begin{aligned}
T_{0F_0}(s) &:= \frac{\partial}{\partial \varepsilon} T_0((1-\varepsilon)F_0 + \varepsilon\delta_s)|_{\varepsilon=0} \\
&= -2\int\int K_0^*(y,z) dF_0(y) dF_0(z) + 2\int K_0^*(y,s) dF_0(y) \\
&:= 2K_0^*(F_0, s) - 2K_0^*(F_0, F_0), \\
T_{0F_0}(s,t) &:= \frac{\partial^2}{\partial \varepsilon_1 \partial \varepsilon_2} T_0((1-\varepsilon_1-\varepsilon_2)F_0 + \varepsilon_1\delta_s + \varepsilon_2\delta_t)|_{\varepsilon_1=0, \varepsilon_2=0} \\
&= -2\int\int K_0^*(y,z) dF_0(y) dF_0(z) + 2\int K_0^*(y,s) dF_0(y) \\
&:= 2K_0^*(F_0, F_0) + 2K_0^*(s,t) - 2(K_0^*(s, F_0) + K_0^*(F_0, t)).
\end{aligned}$$

Finally we get the first and second order terms:

$$\begin{aligned}
T'_{0F_0}(F) &= \int T_{0F_0}(s)d(F(s)-F_0(s)) \\
&= 2\int K_0^*(y,s)dF_0(y)dF(s) - 2\int\int K_0^*(y,z)dF_0(y)dF_0(z) \\
&:= 2K_0^*(F_0,F) - 2K_0^*(F_0,F_0), \\
T''_{0F_0}(F) &= \int\int T_{0F_0}(s,t)d(F(s)-F_0(s))d(F(t)-F_0(t)) \\
&= 2\int\int K_0^*(s,t)d(F(s)-F_0(s))d(F(t)-F_0(t)) \\
&:= 2K_0^*(F_0,F_0) + 2K_0^*(F,F) - 4K_0^*(F_0,F).
\end{aligned}$$

Similarly, we can get the functions and derivatives for all other pieces:

PROPOSITION 20.

$$
\begin{aligned}
\theta_1 &= \int \nabla_x f^* (\nabla_x f^*)^T dx \\
&= \int\int \Big(\int \frac{1}{|H|} K(H^{-1}(x-y)) \frac{1}{|H|} H^{-2}(x-y)(x-z)^T H^{-2} K(H^{-1}(x-z)) dx \Big) dF_0(y) dF_0(z) \\
&:= \int\int K_1^*(y,z) dF_0(y) dF_0(z) \\
&= K_1^*(F_0, F_0) := T_1(F_0).
\end{aligned}
$$

$$
\begin{aligned}
\theta_2 &= \int x x^T f^{*^2} dx \\
&= \int\int \Big(\int \frac{1}{|H|} x x^T K(H^{-1}(x-y)) \frac{1}{|H|} H^{-2} K(H^{-1}(x-z)) dx \Big) dF_0(y) dF_0(z) \\
&:= \int\int K_2^*(y,z) dF_0(y) dF_0(z) \\
&= K_2^*(F_0, F_0) := T_2(F_0).
\end{aligned}
$$

$$
\begin{aligned}
\theta_3 &= \int x f^{*^2} dx \\
&= \int\int \Big(\int \frac{1}{|H|} x K(H^{-1}(x-y)) \frac{1}{|H|} H^{-2} K(H^{-1}(x-z)) dx \Big) dF_0(y) dF_0(z) \\
&:= \int\int K_3^*(y,z) dF_0(y) dF_0(z) \\
&= K_3^*(F_0, F_0) := T_3(F_0).
\end{aligned}
$$

PROPOSITION 21.

$$\begin{aligned}
T^{'}_{iF_0}(F) &= \int T_{iF_0}(s)d(F(s)-F_0(s)) \\
&= 2\int K_i^*(y,s)dF_0(y)dF(s) - 2\int\int K_i^*(y,z)dF_0(y)dF_0(z) \\
&= 2K_i^*(F_0,F) - 2K_i^*(F_0,F_0), i=0,1,2,3 \\
T^{''}_{iF_0}(F) &= \int\int T_{iF_0}(s,t)d(F(s)-F_0(s))d(F(t)-F_0(t)) \\
&= 2\int\int K_i^*(s,t)d(F(s)-F_0(s))d(F(t)-F_0(t)) \\
&= 2K_i^*(F_0,F_0) + 2K_i^*(F,F) - 4K_i^*(F_0,F), i=0,1,2,3.
\end{aligned}$$

So the Fisher information matrix has the following form:

$$T(F) = (T_2T_0 - T_3T_3^T)^{\frac{1}{2}}T_1(T_2T_0 - T_3T_3^T)^{\frac{1}{2}}/T_0^3.$$

Using the above results and chain rule, we can find the first two orders of $T(F)$:

$$T^{'}_{F_0}(F) = 0, \tag{6.1}$$

$$\begin{aligned}
T^{''}_{F_0}(s,t) &= -\frac{2}{\theta_0^2}\Big(\frac{1}{2}\theta_0K_0^*(s,t)I_d - \theta_2K_1^*(s,t) - \theta_1K_2^*(s,t) + K_0^*(s,F_0)K_0^*(t,F_0)I_d \\
&-2(K_1^*(s,F_0)K_2^*(t,F_0) + K_1^*(t,F_0)K_2^*(s,F_0)) + 4K_3^*(s,F_0)K_3^*(t,F_0)\frac{\theta_1}{\theta_0}\Big) \qquad (6.2) \\
&\triangleq 2K^*(s,t).
\end{aligned}$$

Finally we have the expansion:

$$\begin{aligned} \hat{V}^* - \frac{1}{4} I_d &= T(F) - T(F_0) \\ &= \int\int K^*(s,t) d(F(s) - F_0(s)) d(F(t) - F_0(t)) + o_p(1/n). \end{aligned} \tag{6.3}$$

Here we have assumed, not proved, that the remainder term is negligible. We will later investigate the accuracy of the approximation by simulation because a proof of this equivalence is beyond our main emphasis. The first term of the right side is called a kernel-based quadratic distance with kernel K^*. In the univariate case $d = 1$, Lindsay et al.(2006) discussed how to use the spectral decomposition of the kernel K^* to find the limiting distribution of the quadratic distance estimate. We introduce the quadratic distance theory in Section 6.4, and apply it to find the limiting distribution of $trace(\hat{V}^*)$ in Section 6.6. In Section 6.7, we extend the quadratic distance theory to the multivariate case $d > 1$.

6.4 Kernel-based Quadratic Distance and Spectral Decomposition

We use the setting of Lindsay et al.(2006). Let S be a sample space with measurable sets B. Let $K(s,t)$ be a bounded symmetric kernel function on $S \times S$. The kernel $K(s,t)$ is called nonnegative definite(NND), if the quadratic form $\int\int K(s,t) d\sigma(s) d\sigma(t)$ is nonnegative for all bounded signed measures σ, and it is conditionally NND if nonnegativity holds for all σ satisfying the condition $\int d\sigma(s) = 0$.

Definition 3. *Given a conditionally nonnegative definite $K_G(s,t)$, possibly depending on G, the Kernel-based matrix quadratic distance between two probability measures F and*

G is defined as

$$\begin{aligned} d_K(F,G) &= \int\int K_G(s,t)d(F-G)(s)d(F-G)(t) \\ &:= K_G(F,F) - K_G(F,G) - K_G(G,F) + K_G(G,G), \end{aligned}$$

where, for example, $K_G(F,G) = \int\int K_G(x,y)dF(x)dG(y)$. The conditional NND property implies the non-negative definiteness of the kernel-based matrix quadratic distance. Note that the kernel function may depend on the distribution G.

When F and G are continuous with densities f and g, one can write

$$d_K(F,G) = \int\int K_G(s,t)(f(s)-g(s))(f(t)-g(t))dsdt.$$

Many important scalar statistical distances can be written in this form, for example,Pearson-Chisquared distance, Cramer-Von-Mises distance and L_2 distances(see Lindsay et al. 2006). In several physics-related statistical articles, the kernel-based quadratic distance is given another name: energy because of its relation to the energy of electric charge distributions(Aslan and Zech 2006).

In a goodness-of-fit test problem, we can select a suitable kernel function R to measure the distance between the estimated distribution $\hat{F}$ and the null hypothesis distribution F_0:

$$\int\int R(x,y)d(\hat{F}-F_0)(x)d(\hat{F}-F_0)(y).$$

Usually the empirical distribution F_n is used as the estimated distribution $\hat{F}$. For example, Bowman and Foster (1993) use the special kernel function

$$R(x,y) = \int \frac{1}{h}K(\frac{x-z}{h})\frac{1}{h}K(\frac{y-z}{h})dz$$

for a test of multivariate normality. This kernel-based quadratic distance can be written as the quadratic distance between the kernel smoothed densities of the estimated distribution F_n and the null hypothesis distribution F_0. The kernel K is not unique for any given distance $d_K(F,G)$. The kernel $K^*(s,t) = K(s,t) + a(s) + a(t) + b$ generates the exactly same quadratic distance as $K(s,t)$, for any function $a(x)$ and constant b, since F and G are probability measures. The problem can be fixed by centering the kernel K.

Definition 4. *The G-Centered Kernel for kernel K is*

$$K_{cen(G)}(s,t) = K(s,t) - K(s,G) - K(G,t) + K(G,G).$$

In our problem, the $K^*(s,t)$ in (3.3) has been centered automatically: $K^*(s,t) = K_{cen(F_0)}(s,t)$. When G is the hypothesized true distribution, the G-centered kernel $K_{cen(G)}$ simplifies the distance representation:

$$d(F,G) = \int\int K_{cen(G)}(x,y)dF(x)dF(y).$$

The empirical distance between the empirical distribution $\hat{F}$ and the true distribution in null hypothesis is of our interest. For the empirical distance, $d(\hat{F},G) = K_{cen(G)}(\hat{F},\hat{F}) =$

$1^T\hat{K}1/n^2$, where $\hat{K}(i,j) = K_{cen(G)}(x_i, x_j)$. This is a V-statistic, which is a biased estimate of $d(F,G)$. The corresponding U-statistic is

$$U_n = \frac{1}{n(n-1)} \sum_i \sum_{j \neq i} K_{cen}(x_i, x_j).$$

The only difference between $d(\hat{F}, G)$ and U_n is the diagonal terms $K_{cen(G)}(x_i, x_j)$ missing from U_n. When $G = F$, $E_G(U_n) = 0 = d(F,G)$, but $E_G(d(\hat{F}, G)) = E[K_{cen}(X,X)]/n$. A possible numerical problem arises when $K_{cen(G)}$ does not have a explicit form. In this case, numerical calculation may be needed, like Monte Carlo integration. Because the centered kernel only depends on the kernel function K and the distribution G, we can sometimes select suitable kernel K to get an explicit expression of distance. For example, when F_0 is normal distribution, Bowman and Foster (1993) use the normal kernel to estimate density f, and then the quadratic distance is explicit:

$$\begin{aligned} & \int\int R(x,y) d(\hat{F} - F_0)(x) d(\hat{F} - F_0)(y) \\ = \;& \phi(0,0,2(1+h^2)) - \frac{2}{n} \sum_i \phi(x_i, 0, (1+2h^2)) + \frac{1}{n} \sum_{i,j} \phi(x_i, x_j, 2h^2), \end{aligned}$$

where ϕ is normal density and h is the bandwidth.

The asymptotic distribution of $d(\hat{F}, G)$ is determined by the spectral decomposition of kernel.

DEFINITION 5. *Eigenfunction,Eigenvalue*

A function $\phi(y)$ is an eigenfunction of $K(x,y)$ under measure M if the following equation

holds for eigenvalue λ*:*

$$\int K(x,y)\phi(y)dM(y) = \lambda\phi(x)$$

We assume that every eigenfunction ϕ is normalized:

$$\int \phi^2(x)dM(x) = 1$$

and any two different eigenfunctions ϕ_1 and ϕ_2 are orthogonal:

$$\int \phi_1(x)\phi_2(x)dM(x) = 0,$$

provided they are from different eigenvalues. The following spectral decomposition theorem is summarized in Lindsay et al.(2006). The original form can be found in Yosida(1980).

THEOREM 1. *A nonnegative definite kernel* K *can be written as*

$$K(x,y) = \sum_{j=1}^{\infty} \lambda_j \phi_j(x)\phi_j(y), \tag{6.4}$$

if K *satisfies*

$$\int\int K^2(x,y)dM(x)dM(y) < \infty,$$

where λ_j *and* ϕ_j *are eigenvalues and corresponding normalized eigenfunctions of* K *under the measure* M*. The series in (3.4) converges strongly to* K*:*

$$lim_{n\to\infty} \int \Big(\int K(x,y)g(y)dM(y) - \sum_{j=1}^{n} \int \lambda_j \phi_j(x)\phi_j(y)g(y)dM(y) \Big)^2 dM(x) = 0, \forall g \in L_2.$$

The spectral decomposition theorem only ensures the existence of the decomposition of any nonnegative kernel, but there is not a general easy way to find the explicit form of eigenfunctions and eigenvalues. Even when the decomposition of K is available, it is still not easy to find the decomposition of the centered kernel K_{cen}. Some examples of spectral decomposition are listed in Lindsay et al.(2006). Fortunately, the most important attributes of the asymptotic distribution of the estimated kernel-based quadratic distance are summarized in the following two quantities, which can be estimated without the explicit spectral decomposition.

PROPOSITION 22. *If $K_{cen}(x,y)$ is continuous at (x,x) for almost all x with respect to the measure M, and $\sum_{j=1}^{\infty}\lambda_j < \infty$, then*

$$\sum_{j=1}^{\infty}\lambda_j = \int K_{cen}(x,x)dM(x) := trace_M(K_{cen}), \tag{6.5}$$

$$\sum_{j=1}^{\infty}\lambda_j^2 = \int\int K_{cen}^2(x,y)dM(x)dM(y) := trace_M(K_{cen}^2). \tag{6.6}$$

The $trace_M(K_{cen})$ can be unbiasedly estimated by $trace_{\hat{F}}(K_{cen})$. Similarly, $trace_{\hat{F}}(K_{cen}^2)$ is a consistent estimator of $trace_M(K_{cen}^2)$, but it is a V statistic, so biased. The unbiased estimate is a U-statistic: $\frac{1}{n(n-1)}\sum_i\sum_{j>i}K_{cen}^2(x_i,x_j)$. The two estimates do not need the explicit decomposition of kernel and they will be used to approximate the asymptotic distribution of the estimated kernel-based distance.

Suppose we have the spectral decomposition of the centered kernel $K_{cen(G)}$: $K_{cen(G)}(x,y) = \sum_{j=1}^{\infty} \lambda_j \phi_j(x)\phi_j(y)$. The empirical distance can then be written as

$$\begin{aligned} d(\hat{F}_n, G) &= \int\int \sum_{j=1}^{\infty} \lambda_j \phi_j(x)\phi_j(y) d\hat{F}_n(x) d\hat{F}_n(y) \\ &= \sum_{j=1}^{\infty} \lambda_j \Big(\int \phi_j(x) d\hat{F}_n(x) \Big)^2 \\ &= \sum_{j=1}^{\infty} \lambda_j \Big(\frac{1}{n} \sum_{i=1}^{n} \phi_j(X_i) \Big)^2 \\ &\triangleq \sum_{j=1}^{\infty} \lambda_j (\overline{\phi}_j)^2. \end{aligned}$$

The $\overline{\phi}_j$ are uncorrelated over j because of the orthogonality of eigenfunctions and they have zero mean and variance one because of the centered kernel and normalized eigenfunctions.When $\sum_{j=1}^{\infty} \lambda_j < \infty$, Lindsay et al.(2006) presents the asymptotic distribution of $nd(\hat{F}_n, G)$:

$$nd(\hat{F}_n, G) \longmapsto \sum \lambda_j Z_j^2 := \chi^*(\lambda), \lambda = (\lambda_1, \lambda_2, ...) \tag{6.7}$$

$$E(\chi^*(\lambda)) = \sum_{j=1}^{\infty} \lambda_j = \int K_{cen}(x,x) dG(x) \tag{6.8}$$

$$Var(\chi^*(\lambda)) = 2 \sum_{j=1}^{\infty} \lambda_j^2 = 2 \int\int K_{cen}^2(x,y) dG(x) dG(y). \tag{6.9}$$

Under the condition $\sum_{j=1}^{\infty}\lambda_j^2 < \infty$, Liu and Rao(1995) discussed the asymptotic distribution of U_n for quadratic entropy, which has the similar form.

$$\sqrt{n(n-1)}U_n \longmapsto \sum \lambda_j(Z_j^2 - 1) := \chi^*_{cen}(\lambda), \tag{6.10}$$

$$E(\chi^*_{cen}(\lambda)) = 0, \tag{6.11}$$

$$Var(\chi^*_{cen}(\lambda)) = 2\sum_{j=1}^{\infty}\lambda_j^2 = 2\int\int K^2(x,y)dG(x)dG(y). \tag{6.12}$$

Liu and Rao(1995) also showed that under the fixed alternative $G \neq F_0$, the asymptotic distribution of U_n is normal.

$$\sqrt{n}[U_n - d(F_0, G)] \longmapsto n(0, \sigma^2_{\triangle}), \tag{6.13}$$

where $\sigma^2_{\triangle} = 4Var(\triangle\ (X))$, $\triangle\ (x) = K(x, F_0) - K(x, G) - K(F_0, F_0) + K(F_0, G)$. The convergence rate is not continuous between null and alternative.

In the above discussion of asymptotic distributions, we only require the sample size n going to infinity. This is not the end of the story because it is usually difficult to get the eigenvalues and eigenfunctions. To get a simple distribution, we can use Satterthwaite approximation or a normal approximation. Lindsay et al.(2006) showed that Satterthwaite approximation is always better than normal approximation in sense of having closer cumulants to the χ^* distribution. The Satterthwaite approximation of $\chi^*(\lambda)$ is

$$\frac{\chi^*(\lambda) - \sum_{j=1}^{\infty}\lambda_j}{\sqrt{2\sum_{j=1}^{\infty}\lambda_j^2}} \approx \frac{\chi^2_{DOF} - DOF}{\sqrt{2DOF}}, \tag{6.14}$$

where $DOF = (\sum_{j=1}^{\infty} \lambda_j)^2 / \sum_{j=1}^{\infty} \lambda_j^2$. Lindsay et al.(2006) gave conditions that make the approximation valid. In particular, a large value of DOF is required for the Satterthwaite approximation to work well.

Now we go back to our problem: the limiting distribution of $\hat{V}^*$. When $d = 1$ and G is normal, the above results can be applied to K^* directly. Note that we have used a total of three approximations when we use the Satterthwaite distribution approximation:

PROPOSITION 23.

$$\hat{V}^* - \frac{1}{4} = \int\int K^*(s,t) dF_n(s) dF_n(t) + o_p(1/n), \tag{6.15}$$

$$n \int\int K^*(s,t) F_n(s) dF_n(t) \to \chi^*(\lambda), \tag{6.16}$$

$$\frac{\chi^*(\lambda) - \sum_{j=1}^{\infty} \lambda_j}{\sqrt{2 \sum_{j=1}^{\infty} \lambda_j^2}} \approx \frac{\chi^2_{DOF} - DOF}{\sqrt{2DOF}}, \tag{6.17}$$

where DOF is the spectral degrees of freedom under G of centered kernel K^*. The first two approximations only require a large sample size n and the third approximation requires a large degrees of freedom. The degrees of freedom depends on the value of the smoothing parameter h, when G and K^* are fixed. We will use our simulation to show that the degrees of freedom goes to infinity as the smoothing parameter h goes to zero. So the third approximation needs a small h.

Substituting $n(\hat{V}^* - \frac{1}{4}) \approx \chi^*(\lambda)$ into the third approximation, we get:

$$n(\hat{V}^* - \frac{1}{4}) \sum_{j=1}^{\infty} \lambda_j / \sum_{j=1}^{\infty} \lambda_j^2 \approx \chi^2_{DOF}, \tag{6.18}$$

where $\sum_{j=1}^{\infty} \lambda_j$, $\sum_{j=1}^{\infty} \lambda_j^2$ can be estimated using (6.5).

6.5 Simulation

In this section, we assume $d = 1$. We will verify the asymptotic results by simulation. First we consider the approximation:

$$n(\hat{V}^* - \frac{1}{4}) \approx n \int \int K^*(s,t) dF_0(s) dF_0(t).$$

We use $K^*(\hat{F}, \hat{F})$ to estimate $\int \int K^*(s,t) dF_0(s) dF_0(t)$. For every fixed h, the difference decreases to zero as $n \to \infty$. It verifies our guess that the error term of the Von-Mises expansion (6.3) is $o_p(1/n)$. The convergence depends on the sample size n and the smoothing parameter h (Table 6.1). When h is large, the data is heavily smoothed, so it converges to zero faster for large h than small h. For fixed sample size n, the larger h is, the better the approximation is. Suppose the critical value of the difference is 0.1, any $h > 0.533$ is good for the sample size $n = 500$.

For the approximation

$$n \int \int K^*(s,t) F_n(s) dF_n(t) \to \chi^*(\lambda),$$

we check the mean and variance of $n(\hat{V}^* - \frac{1}{4})$ (Table 6.2). When n is large(n=500), the sample mean and sample variance of $n(\hat{V}^* - \frac{1}{4})$ are very close to the mean and variance of $\chi^*(\lambda)$ for every fixed h. The degrees of freedom clearly depends monotonically on the smoothing parameter h. A large h value means more smoothing. The DOF converges to ∞ as $h \to 0$, and 1 as $h \to \infty$.

In order to examine the approximation

$$n(\hat{V}^* - \frac{1}{4}) \sum_{j=1}^{\infty} \lambda_j / \sum_{j=1}^{\infty} \lambda_j^2 \approx \chi^2_{DOF},$$

for every h, we first estimate DOF, and then generate 1000 samples of the left term(every sample contains $n = 500$ observations) to check if they are from χ^2_{DOF} using Kolmogorov-Smirnov test. The p-values and the corresponding h and DOF are shown in Table 6.3. When $DOF > 3 (h < 0.63)$, K-S test$(p - value > 0.05)$ does not reject the hypothesis that the samples are from the distribution of χ^2_{DOF}.

Another useful tool to examine the approximation is a quantile-quantile plot. For $h = 0.63$, we generate 1000 samples from χ^2_{DOF}, and draw a quantile-quantile plot for the simulated left term and the simulated data from χ^2_{DOF}. The points fall approximately along the 45-degree reference line (See Figure 6.1), so we believe that the two data sets come from the same distribution.

Combining all these simulation results, we can conclude that when n is large enough, all approximations are very good; for a fixed sample size n, we can select a suitable smoothing parameter h, which is good for all approximations. For example, when $n = 500$, any h in the interval $(0.533, 0.63)$ is fine for all these approximations.

Table 6.1 The mean of $|n(\hat{V}^* - \frac{1}{4}) - nK^*(\hat{F}, \hat{F})|$ over $B = 100$ simulated data sets.

	h=0.35	0.46	0.533	0.63	2	h_{op}
n=500	0.9616	0.2196	0.0714	0.0093	0.0003	1.8634(h=0.31)
1000	0.6268	0.1897	0.0481	0.0075	0.0000	1.8539(h=0.27)
2000	0.0103	0.0093	0.0093	0.0014	0.0000	0.5593(h=0.23)

Table 6.2 Moments comparisons for the approximation $n(\hat{V}^* - \frac{1}{4}) \rightarrow \chi^*(\lambda)$,n=500, based on $B = 1000$ simulated data sets

h	0.05	0.35	0.63	1	2	5
$E(n(\hat{V}^* - \frac{1}{4}))$	1997.880	5.4601	0.7016	0.0988	0.0039	0.000002
$\sum \lambda_j$	2005.582	5.4768	0.6925	0.1036	0.0037	0.000002
$V(n(\hat{V}^* - \frac{1}{4}))$	293738.6	10.23789	0.3373	0.0115	0.0000022	9.997116e-10
$2\sum \lambda_j^2$	214565.9	11.4889	0.3148	0.0106	0.0000019	9.477387e-10
DOF	37.49298	5.221545	3.0455	1.986791	1.4727232	1.020690

Table 6.3 Evaluating the approximation $n(\hat{V}^* - \frac{1}{4}) \sum_{j=1}^{\infty} \lambda_j / \sum_{j=1}^{\infty} \lambda_j^2 \approx \chi^2_{DOF}, n = 500$ using Kolmogorov-Smirnov test based on $B = 1000$ samples

h	0.46	0.533	0.63	0.755	1
DOF	4	3.5	3	2.6	2
p-value	0.5361	0.3136	0.2193	0.0295	0.0053

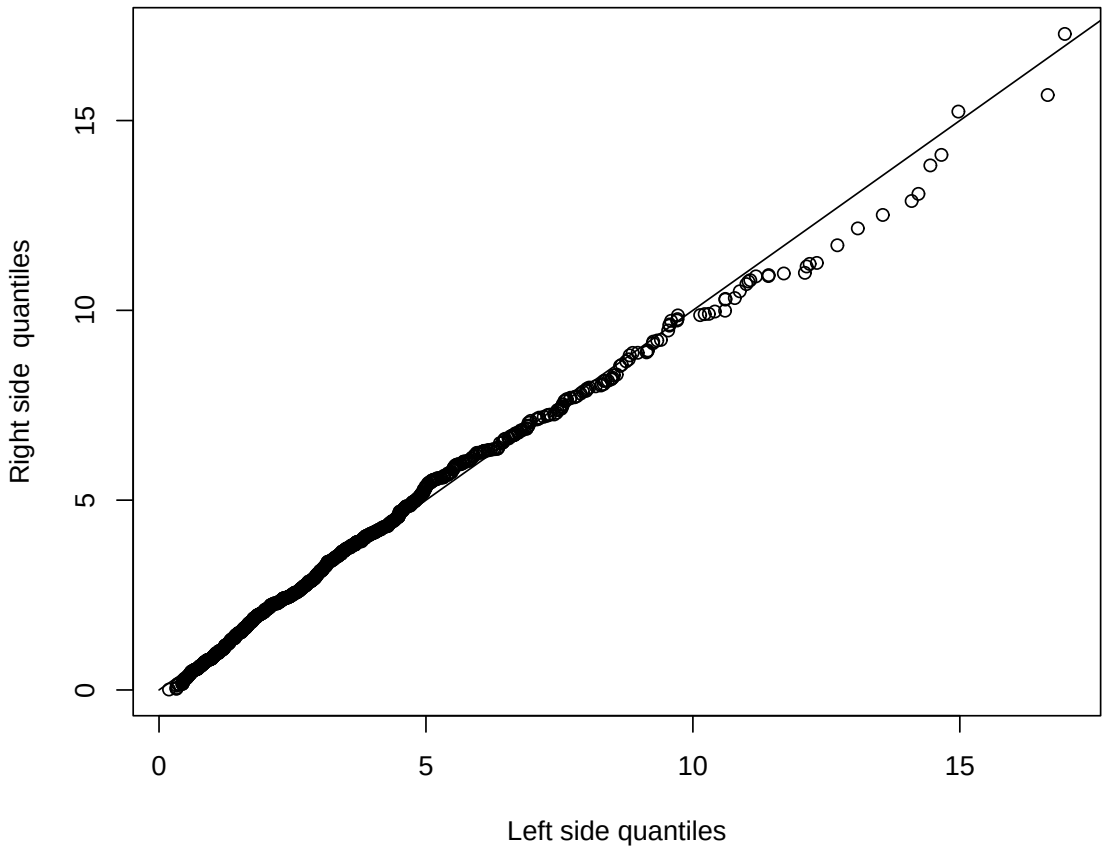

Fig. 6.1 Q-Q plot of randomly generated data from the two sides of (6.18)
Vertical axis: Estimated quantiles from $n(\hat{V}^* - \frac{1}{4}) \sum_{j=1}^{\infty} \lambda_j / \sum_{j=1}^{\infty} \lambda_j^2$
Horizontal axis: Estimated quantiles from $\chi^*(\lambda)$

6.6 Trace of J_{f_2}

In Section 4.5, we showed that the eigenanalysis of J_{f_2} provides all solution projections, and the sum of eigenvalues $\sum_{i=1}^{\infty}\lambda_i$ is a test statistic for non-normality of X. In this section, we assume $\mu_f = 0, V_f = I_d$, and $H = hI_d, h > 0$. We apply the quadratic distance theory to find the asymptotic distribution of the trace of J_{f_2}, which is equal to the sum of eigenvalues $\sum_{i=1}^{\infty}\lambda_i$.

In Section 6.3, we found the Von-Mises expansion of $\hat{V}^*$:

$$\begin{aligned}\hat{J}_{f_2} - \frac{1}{4}I_d &= \hat{V}^* - \frac{1}{4}I_d \\ &= \int\int K^*(s,t)dF_n(s)dF_n(t) + o_p(1/n).\end{aligned}$$

So we have

$$Trace(\hat{J}_{f_2}) - \frac{d}{4} = \int\int trace(K^*(s,t))d\hat{F}_n(s)d\hat{F}_n(t) + o_p(1/n), \qquad (6.19)$$

where

$$
\begin{aligned}
& trace(K^*(s,t)) \\
= \; & -\frac{1}{\theta_0^2} trace\Big(\frac{1}{2}\theta_0 K_0^*(s,t) I_d - \theta_2 K_1^*(s,t) - \theta_1 K_2^*(s,t) + K_0^*(s,F_0) K_0^*(t,F_0) I_d \\
& -2(K_1^*(s,F_0) K_2^*(t,F_0) + K_1^*(t,F_0) K_2^*(s,F_0)) + 4K_3^*(s,F_0) K_3^*(t,F_0) \frac{\theta_1}{\theta_0}\Big) \\
= \; & -\frac{1}{\theta_0^2}\Big(\frac{d}{2}\theta_0 \phi(s,t,2H^2) - \theta_0 \phi(s,t,2H^2) \frac{1+h^2}{2}(\frac{d}{2h^2} - \frac{(s-t)^T(s-t)}{4h^4}) \\
& -\theta_0 \phi(s,t,2H^2) \frac{1}{2+2h^2}(\frac{dh^2}{2} + \frac{(s+t)^T(s+t)}{4}) + d\phi(s,0,A)\phi(t,0,A) \\
& -2\phi(s,0,A)\phi(t,0,A)[\frac{2h^2(1+h^2)}{(1+2h^2)^2} + \frac{1+h^2}{(1+2h^2)^3}(t^T t + s^T s) - \frac{2(1+h^2)^2}{(1+2h^2)^4}(t^T t s^T s)] \\
& +\phi(s,0,A)\phi(t,0,A)\frac{2+2h^2}{(1+2h^2)^2} s^T t\Big),
\end{aligned}
$$

$\theta_0 = \phi(0,0,2I_d + 2H^2)$, and $A = I_d + 2H^2$. The left term $\int\int trace(K^*(s,t)) d\hat{F}_n(s) d\hat{F}_n(t)$ is still a kernel-based quadratic distance. Suppose the trace function has the spectral decomposition

$$trace(K^*(s,t)) = \sum_{j=1}^{\infty} t_j \phi_j(s) \phi_j(t).$$

Under the null hypothesis $X \sim N(0, I_d)$, by the quadratic distance theory, we have

$$n(Trace(\hat{J}_{f_2}) - \frac{d}{4}) \to \sum_{j=1}^{\infty} t_j Z_j^2, \tag{6.20}$$

where $Z_j's$ are iid standard normal variables.

6.7 Asymptotic Distribution of $\hat{V}^*$

When $d = 1$, the limiting distribution of $\hat{V}^*$ can be found by the spectral decomposition of the kernel. The eigenfunctions are normalized and orthogonal:

$$\int \phi_i^2(x) dF_0(x) = 1,$$

$$\int \phi_i(x)\phi_j(x) dF_0(x) = 0,$$

which makes the terms $\int \phi_i(x) d\hat{F}_n(x)$ are uncorrelated variables with mean zero and variance one. So we have $\int \phi_i(x) d\hat{F}_n(x) \to Z_i$, and all $Z_i's$ are iid standard normal. When $d > 1$, the spectral decomposition still holds. We will use the permutation invariant property of $\hat{V}^*$ to study the properties of the terms $\int \phi_i(x) d\hat{F}_n(x)$, and then get the form of the limiting distribution of $\hat{V}^*$.

Withers(1974) presented the multivariate version of the spectral decomposition. Let $K(x, y)$ be a matrix measurable $d \times d$ symmetric kernel function on $S \times S$. A $d \times 1$ function $\phi(y)$ is an eigenfunction of $K(x, y)$ under measure M if the following equation holds for eigenvalue λ:

$$\int K(x, y)\phi(y) dM(y) = \lambda\phi(x).$$

The eigenfunctions ϕ_i are normalized and orthogonal:

$$\int \phi_i^T(x)\phi_j(x) dM(x) = \begin{cases} 1, & i = j; \\ 0, & i \neq j. \end{cases}$$

THEOREM 2. *Under the condition*

$$0 \leq \int \int \sum_{i,j} |K_{ij}(x,y)|^2 dM(x) dM(y) \leq \infty,$$

a symmetric kernel $K(x,y)$ *can be written as*

$$K(x,y) = \sum_{j=1}^{\infty} \lambda_j \phi_j(x) \phi_j^T(y), \tag{6.21}$$

where λ_j *and* ϕ_j *are eigenvalues and corresponding eigenfunctions of* $K(x,y)$ *under the measure* M*. The series in (6.21) converges (elementwise) absolutely and uniformly to* K*.*

Suppose G is a hypothetical true model, we center the matrix kernel $K(x,y)$ to get an uniform decomposition:

$$K_{cen(G)}(x,y) = K(x,y) - K(x,G) - K(G,y) + K(G,G).$$

So any constant vector $a = (a_1, a_2, ..., a_d)^T$ in R^d is a eigenfunction of $K_{cen(G)}(x,y)$ with eigenvalue zero:

$$\int K_{cen(G)}(x,y) \cdot a dG(y) = 0.$$

For any eigenfunction $\phi_i = (\phi_{i1}, \phi_{i2}, ..., \phi_{id})$ with non-zero eigenvalue, by the orthogonality of eigenfunctions, we have

$$\int \phi_i^T(x) \cdot a dG(x) = \sum_{j=1}^{d} a_j \int \phi_{ij}(x) dG(x) = 0, \forall a \in R^d,$$

and

$$\int \phi_i^T(x)\phi_i(x)dG(x) = \sum_{j=1}^{d} \int \phi_{ij}^2(x)dG(x) = 1.$$

Thus, we have $\int \phi_{ij}(x)dG(x) = 0, \forall i, j$, and $\sum_{j=1}^{d} \int \phi_{ij}^2(x)dG(x) = 1$. Now we consider the decomposition of our kernel function $K^*(s,t)$

$$K^*(s,t) = \sum_{j=1}^{\infty} \lambda_j \phi_j(s)\phi_j'(t), \lambda_j > 0.$$

Substituting it into the Von-Mises approximation of $\hat{V}^*$:

$$\begin{aligned} n(\hat{V}^* - \frac{1}{4})I_d &= n\int\int K^*(s,t)\hat{F}(s)d\hat{F}(t) + o_p(1) \\ &= n\int\int \sum_{j=1}^{\infty} \lambda_j \phi_j(s)\phi_j'(t)d\hat{F}_n(s)d\hat{F}_n(t) + o_p(1) \\ &= n\sum_{j=1}^{\infty} \lambda_j \Big(\int \phi_j(s)d\hat{F}_n(s)\Big)\Big(\int \phi_j(t)d\hat{F}_n(t)\Big)' + o_p(1). \end{aligned} \tag{6.22}$$

When the null hypothesis $X = (X_1, X_2, ..., X_d) \sim N(0, I_d)$ is true, any permutation of X: $X_\sigma = (X_{\sigma 1}, X_{\sigma 2}, ..., X_{\sigma d})$ has the same distribution $N(0, I_d)$. We get the same Fisher information matrix J_{f_2} for X_σ. Consider any two realizations $X_i = (X_{i1}, X_{i2}, ..., X_{id})$, $X_j = (X_{j1}, X_{j2}, ..., X_{jd})$ from X, and any permutations $X_{\sigma_i} = (X_{\sigma_{i1}}, X_{\sigma_{i2}}, ..., X_{\sigma_{id}})$, and $X_{\sigma_j} = (X_{\sigma_{j1}}, X_{\sigma_{j2}}, ..., X_{\sigma_{jd}})$. The two matrices have the same distribution:

$$K^*(X_i, X_j) \doteq K^*(X_{\sigma_i}, X_{\sigma_j}).$$

So, for any eigenfunction $\phi_i(x)$, all elements have the same distribution

$$\phi_{ik}(X_i) \doteq \phi_{il}(X_i), \forall k, l.$$

Thus, we have

$$\begin{aligned} & \int \phi_{ik}^2(x) dF_0(x) \\ = & \frac{1}{d} \sum_{k=1}^{d} \int \phi_{ik}^2(x) dF_0(x) \\ = & \frac{1}{d}, \forall 1 \leq i \leq \infty, 1 \leq k \leq d, \end{aligned} \tag{6.23}$$

$$\begin{aligned} & \int \phi_{ik}(x)\phi_{il}(x) dF_0(x) \\ = & \int \phi_{ik'}(x)\phi_{il'}(x) dF_0(x) \\ := & \frac{1}{d}\rho_i, \forall 1 \leq i \leq \infty, 1 \leq k \neq l, k' \neq l' \leq d, \end{aligned} \tag{6.24}$$

and

$$\begin{aligned} & \int \phi_{ik}(x)\phi_{jl}(x) dF_0(x) \\ = & \int \phi_{ik}(x)\phi_{jk}(x) dF_0(x) \\ = & \frac{1}{d} \sum_{k=1}^{d} \int \phi_{ik}(x)\phi_{jk}(x) dF_0(x) \\ = & \frac{1}{d} \int \phi_i^T(x)\phi_j(x) dF_0(x) \\ = & 0, \forall 1 \leq i \neq j \leq \infty, 1 \leq k, l \leq d. \end{aligned} \tag{6.25}$$

Let Σ_i be the covariance matrix of $\phi_i(X)$:

$$\Sigma_i = \frac{1}{d}\begin{pmatrix} 1 & \rho_i & ... & \rho_i \\ \rho_i & 1 & \rho_i & \rho_i \\ ... & ... & ... & ... \\ \rho_i & \rho_i & ... & 1 \end{pmatrix}.$$

The covariance matrix Σ_i has the d eigenvalues $a_{i1} = \frac{1}{d}(1-(d-1)\rho_i)$ and $a_{ik} = \frac{1}{d}(1-\rho_i), k = 2, ..., d$. Then we have

$$\sqrt{n}\int \phi_i(x)d\hat{F}_n(x) \to \Sigma_i^{\frac{1}{2}} Z_i, i = 1, 2, ...d,$$

where $Z_i's$ are iid d-variate standard normal variables. Applying the above results to (6.22) derives

$$\begin{aligned} n(\hat{V}^* - \frac{1}{4}I_d) &= n\sum_{j=1}^{\infty}\lambda_j\Big(\int \phi_j(s)d\hat{F}_n(s)\Big)\Big(\int \phi_j(t)d\hat{F}_n(t)\Big)^T + o_p(1) \\ &\to \sum_{i=1}^{\infty}\lambda_i\Sigma_i^{\frac{1}{2}} Z_i Z_i^T \Sigma_i^{\frac{1}{2}} \\ &:= \sum_{i=1}^{\infty}\lambda_i W_i, \end{aligned} \tag{6.26}$$

where $W_i \sim W_d(\Sigma_i, 1)$ are independent Wishart variables with one degree of freedom. So the limiting distribution of $n(\hat{V}^* - \frac{1}{4}I_d)$ is a infinite sum of weighted of independent Wishart variables. Similar to Satterthwaite approximation of the sum of independent chi-squared variables, we can find a Wishart variable $W(\gamma, \Sigma)$ to approximate the sum of independent Wishart variables by matching the first two moments(See details in Nel

and Merwe 1986). However, we still do not know the correlation parameters ρ_i. More investigation is needed to find the decomposition of the kernel matrix or an easy way to estimate the correlation parameters. By using the limiting distribution of $n(\hat{V}^* - \frac{1}{4}I_d)$, we can find the limiting distribution of $trace(\hat{V}^*)$. Let $D = diag(a_{i1}, a_{i2}, ..., a_{id})$. Then, we have

$$\begin{aligned} n(trace(\hat{V}^*) - \frac{d}{4}) &\rightarrow \sum_{i=1}^{\infty} \lambda_i trace(\Sigma_i^{\frac{1}{2}} Z_i Z_i^T \Sigma_i^{\frac{1}{2}}) \\ &= \sum_{i=1}^{\infty} \lambda_i Z_i^T \Sigma_i Z_i \\ &= \sum_{i=1}^{\infty} \lambda_i Z_i^T D Z_i \\ &= \sum_{i=1}^{\infty} \frac{\lambda_i}{d} \Big((1-(d-1)\rho_i) Z_{i1}^2 + (1-\rho_i) \sum_{k=2}^{d} Z_{ik}^2 \Big), \end{aligned} \tag{6.27}$$

which is different from another form (6.20) we found by directly considering the trace of the kernel function in Section 6.5, but it is still a weighted sum of independent chi-squared variables. The two expressions for the same limiting distribution may help us to find the values of parameters.

Chapter 7

Non-normality Direction

7.1 Overview

Let $X = (X_1, X_2, ..., X_d)$ be a d-dimensional random vector with the density function $f(x)$. Let a be a d-dimensional vector and $Y = a^T X$ be a linear projection with the density $g(y)$. Our aim is to find a subspace of projections that has the least similarity to white noise coordinates. In Section 3.3, we showed that the eigenanalysis of J_f provides all the solution projections. The projection's similarity to white noise depends on the corresponding eigenvalue. When the eigenvalue reaches the lower bound, the corresponding projection is a white noise, which can be discarded in further study because it is marginally normal and independent of the other projections. However, J_f is not easy to estimate because of the denominator f. In order to solve the computation problem of J_f, we applied the density square transformation to get the new Fisher information matrix J_{f_2}; we also constructed a new projection index Q_f by Von-Mises expansion. Both of the two projection indices are easy to estimate by the kernel method. And we can find all the solution projections only by the eigenanalysis of J_{f_2} or Q_f.

In Section 7.2, we will apply the two new projection indices to the simulated data sets to investigate their power in detecting known non-normal structures. In order to assess whether a coordinate is white noise, we develop a simulation-based test procedure. In Section 7.3, the projection pursuit methods will be applied to the real data sets to

compare the performance with classical methods. We will summarize the advantages and possible problems of the two projection pursuit methods in the next chapter.

7.2 Simulation Study

In this section, we apply our method to the artificially generated data sets, in which the structures are known. Without other specification, the smoothing parameter used will be

$$H_{op} = \Big(\frac{4}{d+2}\Big)^{\frac{1}{d+4}} \Sigma^{1/2} n^{-\frac{1}{d+4}},$$

where Σ can be estimated by the sample covariance matrix. See Chapter 4 Section 3 for a discussion of its optimality properties.

7.2.1 Matrix Index vs Scalar Index

A scalar projection index measures the marginal non-normality of a projection. Many projection pursuit algorithms based on scalar projection indices (e.g. Friedman 1984, Jee 1985) require searching the whole subspace to find interesting features within a data set. As showed in Section 3.3, 4.5 and 5.2, the Fisher information matrix J_{f_2} and the Von-Mises approximation Q_f measure the conditional non-normality of X. For example, the ith diagonal term of J_{f_2} is the weighted average of the Fisher information $J_{X_i|X_{-i}}$ of X_i conditioned on all other uncorrelated variables. In order to find the least normal directions, we only need to calculate or estimate the Fisher information matrix J_{f_2}, and do an eigenanalysis of the estimated matrix. Furthermore, we might hope that

the least conditional normal projection from the eigenanalysis of J_{f_2} and Q_f also has the least marginal normality.

Consider the distribution

$$f(x_1, x_2) = 2\phi\Big((x_1, x_2)^T, (0,0)^T, I_2\Big) I(x_1 x_2 > 0).$$

The center is the origin. The covariance matrix is

$$V_f = \begin{pmatrix} 1 & 2/\pi \\ 2/\pi & 1 \end{pmatrix}.$$

This distribution arises from a standard bivariate normal distribution by the transformation:

$$(x_1, x_2) = \begin{cases} (z_1, z_2), & \text{if } z_1 * z_2 \geq 0 \\ (z_1, -z_2), & \text{if } z_1 * z_2 < 0. \end{cases}$$

Both X_1 and X_2 are marginally normal, but the conditional distribution of $X_1|X_2$ or $X_2|X_1$ are not. First, let $Y = V_f^{-\frac{1}{2}} X$. So the density function of Y is

$$\begin{aligned} g(y) &= f(V_x^{\frac{1}{2}} y) I_{V_x^{\frac{1}{2}} y > 0} |V_x^{\frac{1}{2}}| \\ &\propto exp\Big(-\frac{1}{2}(y_1^2 + y_2^2 + \frac{4}{\pi} y_1 y_2)\Big) I_{y_1^2 + y_2^2 + \pi y_1 y_2 > 0}. \end{aligned}$$

From the contour plot(Figure 7.1), we might guess that the projection $\frac{1}{\sqrt{2}} y_1 + \frac{1}{\sqrt{2}} y_2$ has the least marginal normality. By the symmetry of y_1 and y_2, it is easy to show that all

the three indices J_f, J_{f_2} and Q_f have the form

$$\begin{pmatrix} a & b \\ b & a \end{pmatrix}$$

Thus, the eigenvector with the largest eigenvalue is $(\frac{1}{\sqrt{2}}, \frac{1}{\sqrt{2}})$ and the eigenvector with the smallest eigenvalue is $(\frac{1}{\sqrt{2}}, -\frac{1}{\sqrt{2}})$. So, on average, the projection $\frac{1}{\sqrt{2}}y_1 + \frac{1}{\sqrt{2}}y_2$ has least normality conditioned on the uncorrelated variable $\frac{1}{\sqrt{2}}y_2 - \frac{1}{\sqrt{2}}y_1$. The simulation results are consistent with the theoretical analysis (See Table 7.1). A much bigger sample size is needed to get a good estimator of Q_f. When the sample size is only 500, the estimators of J_{f_2} are more robust than those of Q_f.

The result also verifies our conjecture that the least conditional normal projection also has the least marginal normality in our problem. Note that the smaller eigenvalue of J_{f_2} is less than the lower bound 1, because the density function $f(x_1, x_2)$ is not differentiable at the origin $(0, 0)$, and so the matrix inequality does not hold for this density function.

Table 7.1 eigenanalysis of J_{f_2} and Q_f for $f(y_1, y_2) \propto exp\Big(-\frac{1}{2}(y_1^2 + y_2^2 + \frac{4}{\pi}y_1y_2)\Big)I_{y_1^2+y_2^2+\pi y_1y_2>0}$ based on simulated samples

Projection Index	Sample size	Eigenvalue	Eigenvector
J_{f_2}	n=500	0.3455	$(-0.6985, 0.7156)$
		0.6058	$(0.7156, 0.6985)$
J_{f_2}	n=500	0.3488	$(-0.6981, 0.7160)$
		0.5060	$(0.7160, 0.6981)$
Q_f	n=500	3.9131	$(-0.9246, -0.3810)$
		6.2443	$(0.3810, -0.9246)$
Q_f	n=500	6.1705	$(0.1922, 0.9813)$
		13.0338	$(0.9813, -0.1922)$
Q_f	n=2000	1.2001	$(-0.7170, 0.6971)$
		2.0628	$(0.6971, 0.7170)$

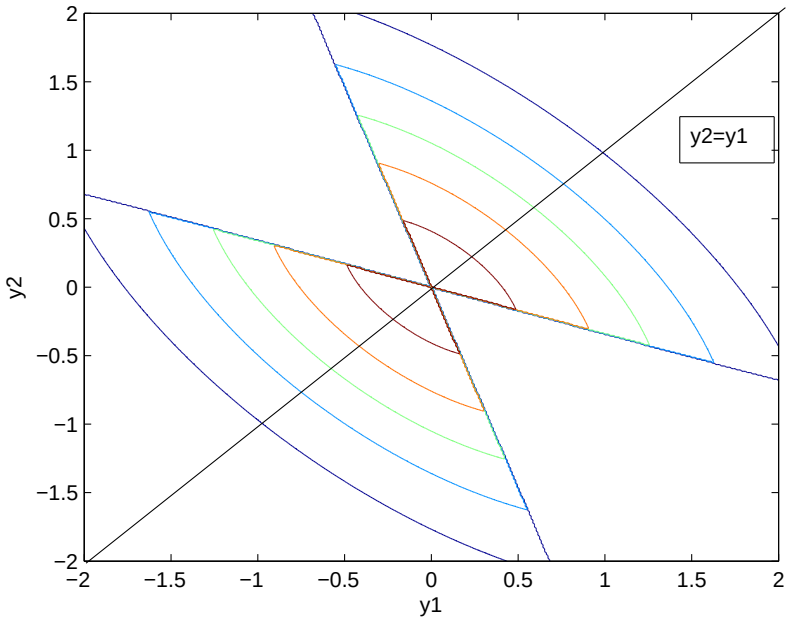

Fig. 7.1 Contour plot for the density of Y: $g(y) = f(V_x^{\frac{1}{2}}x)|V_x^{\frac{1}{2}}|$

7.2.2 Normal Mixture Model

We used the density square transformation to get a rapidly computable Fisher information matrix J_{f_2}. The density square transformation preserves the most important structures of the original data, so we believe the optimal directions from the eigenanalysis of J_{f_2} should be close to those from Q_f. In order to verify our guess, we will compare the results from Q_f and J_{f_2} for the simplest case: projection from dimension 2 to dimension 1. The results are presented graphically.

The first model is two-component mixture of normals:

$$f(x_1, x_2) = 0.5\phi(x_1, 0, 1)\phi(x_2, 0, 1) + 0.5\phi(x_1, 3, 1)\phi(x_2, \mu, 1), \mu \geq 0$$

Both components have unit covariance matrix I_2. The center of the first component is fixed at the origin. The second mean vector is $(3, \mu)$. So the two components are well separated. In order to simplify the computation, we use the angle $\theta \in [0, \pi/2]$ clockwise the positive axis as the parameter of the line. The projection Y on the line $X_2 = tan(\theta)X_1$ is $Y = cos(\theta)X_1 + sin(\theta)X_2$. Let $\alpha = arctan(\mu_2/3)$. Then the line between two component centers is: $X_2 = tan(\alpha)X_1$. From the contour plot of the density (Figure 7.2), we think that the ideal solution should be the projection on the line between the two centers. This projection $Y = cos(\alpha)X_1 + sin(\alpha)X_2$ is still a two-component normal mixture, so it has least conditional normality.

We do the standardization $X_S = V_f^{-\frac{1}{2}}X$. It is easy to show that the transformed model is still a two-component normal mixture model and the standardization will not change the optimal direction. The least normal direction is the line between the two

centers and the optimal direction is still $(cos(\alpha), sin(\alpha))$. In the eigenanalysis of J_f, J_{f_2} and Q_f, we showed that the standardization preserves the non-normality measure. So we do not need to compute the new density and the new projection index for the standardized variable.

For every fixed $\theta \in [0, \pi]$, the exact form of J_{f_2} can be computed numerically (see section 4.2 for formulas). An easy way is to use a random sample from $f(x_1, x_2)$ to estimate J_{f_2}. The asymptotically optimal H is used. The sample size is 1000. In order to ensure good estimator of Q_f, we use a bigger sample size $n = 2000$. The asymptotically optimal H is also used to estimate Q_f.

The simulation results (Table 7.3 and Table 7.2) agree with what we expected: for two-component mixture of normals, the eigenanalysis of Q_f and J_{f_2} provides the same least normal projection $Y = cos(\alpha)X_1 + sin(\alpha)X_2$, which also has the least marginal normality.

Table 7.2 The least normal direction from the eigenanalysis of estimated J_{f_2} based on 1000 samples. The ideal solution direction is $(cos(\alpha), sin(\alpha))$.

μ	α	$(cos(\alpha), sin(\alpha))$	least normal direction
0	0	(1,0)	(0.9995,-0.0301)
$\sqrt{3}$	$\pi/6$	(0.8660, 0.5000)	(0.8659,0.5003)
3	$\pi/4$	(0.7071, 0.7071)	(0.7089,0.7053)
$3\sqrt{3}$	$\pi/3$	(0.5000,0.8660)	(0.5027,0.8643)
1000	$\pi/2$	(0,1)	(0.0002,1.0000)

Table 7.3 The least normal direction from the eigenanalysis of estimated Q_f based on 2000 samples. The ideal solution direction is $(cos(\alpha), sin(\alpha))$.

μ	α	$(cos(\alpha), sin(\alpha))$	least normal direction
0	0	(1,0)	(1,0)
$\sqrt{3}$	$\pi/6$	(0.8660, 0.5000)	(0.8659, 0.5002)
3	$\pi/4$	(0.7071, 0.7071)	(0.7071, 0.7071)
$3\sqrt{3}$	$\pi/3$	(0.5000,0.8660)	(0.5000,0.8660)
1000	$\pi/2$	(0,1)	(0,1)

The second model is a three-component normal mixture model of equal proportions. Each component has unit covariance matrix I_2. The mean vectors are $(5, 5)$, $(-5, -5)$ and$(5, -5)$. According to the contour plot of the density (Figure 7.3), there are three natural solution projections. The first one is the projection $Y_1 = cos(\frac{\pi}{4})X_1 + sin(\frac{\pi}{4})X_2$, which has three separated components. The second one is the projection $Y_2 = X_1$, or $Y_2 = X_2$, which has two components. The density of this projection is $f(x) = \frac{1}{3}\phi(x, -5, 1)+\frac{2}{3}\phi(x, 5, 1)$. A third possible solution projection is $Y_3 = cos(\frac{3\pi}{4})X_1+ sin(\frac{3\pi}{4})X_2$, which has two components.

It is easy to show that, after the standardization$X_S = V_f^{-\frac{1}{2}}X$, the transformed model is still a three-component normal mixture model and the standardization will not

change the possible optimal directions.

In this example, we do not need to find the exact value of J_{f_2} or Q_f It is easy to show that both of them have the form

$$\begin{pmatrix} a & b \\ b & a \end{pmatrix}.$$

For the above matrix, the eigenvectors must be $(\frac{1}{\sqrt{2}}, \frac{1}{\sqrt{2}})$ for the largest eigenvalue and $(\frac{1}{\sqrt{2}}, -\frac{1}{\sqrt{2}})$ for the smallest eigenvalue. So both the projections indices J_{f_2} and Q_f favor the first projection $Y_1 = cos(\frac{\pi}{4})X_1 + sin(\frac{\pi}{4})X_2$. Alternatively, the projection $Y_3 = cos(\frac{3\pi}{4})X_1 + sin(\frac{3\pi}{4})X_2$ is closest to white noise. We may get a different solution projection if a different projection index is used, for example, L_1 metric index or Hellinger metric index (See the similar model in Jee 1985).

This example reveals a possible drawback of our projection pursuit method based on eigenanalysis. Some interesting non-normal projections(e.g. Y_2) are not found, because they are not orthogonal to the least normal projection Y_1. However, the solution found does capture all the mixture components.

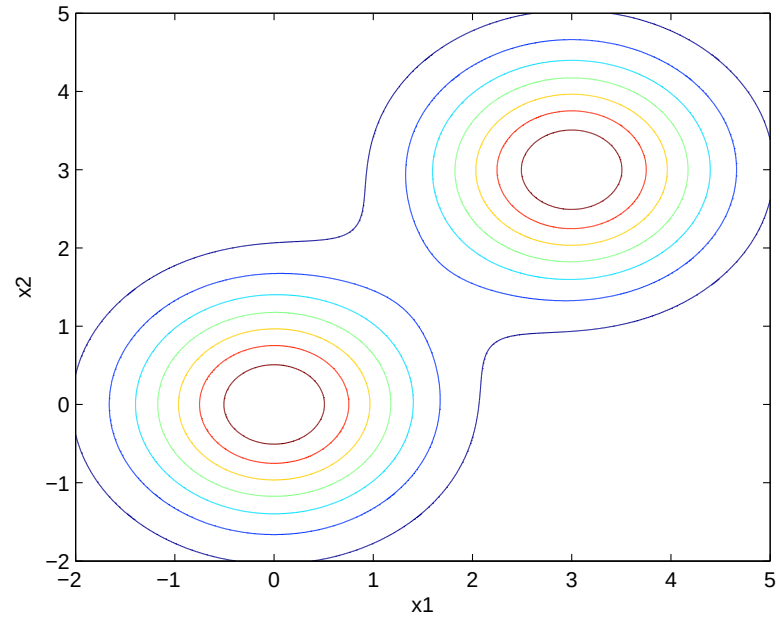

Fig. 7.2 Contour plot for $0.5\phi(x_1, 0, 1)\phi(x_2, 0, 1) + 0.5\phi(x_1, 3, 1)\phi(x_2, 3, 1)$

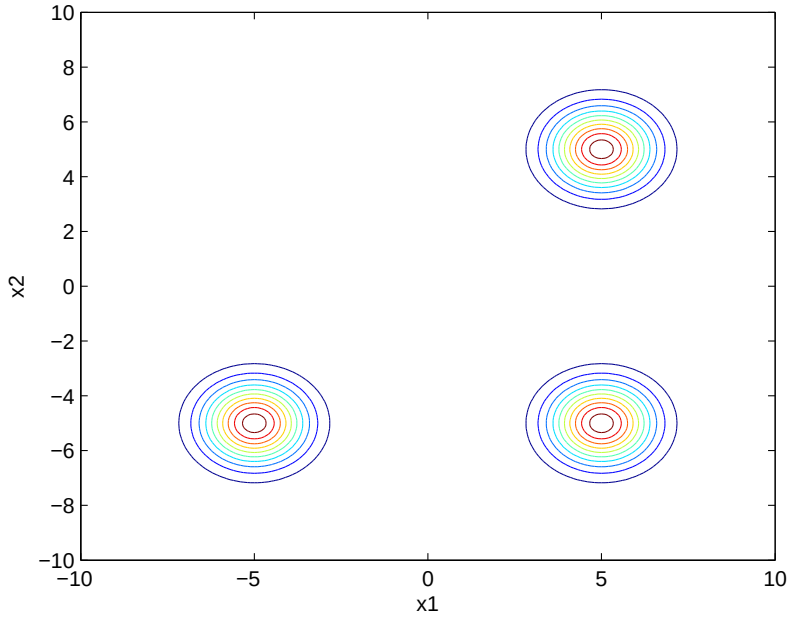

Fig. 7.3 Contour plot for $f(x_1, x_2) = \frac{1}{3}\phi(x_1, 5, 1)\phi(x_2, 5, 1) + \frac{1}{3}\phi(x_1, 5, 1)\phi(x_2, -5, 1) + \frac{1}{3}\phi(x_1, -5, 1)\phi(x_2, -5, 1)$

7.2.3 Needle in a Haystack

In this subsection, we will investigate the power of our projection pursuit methods in finding an interesting structure in high dimensional space. We also study the effect of the selection of smoothing parameter h. The simulation results show that when dimensionality d increases, the required sample size increases rapidly. For the Fisher information matrix J_{f_2}, the rate of increase is less the exponential. The Von-Mises approximation matrix Q_f requires a very large sample size, compared to J_{f_2}.

In the example, we construct a spiral in the first two dimensional space(Figure 7.4), and fill the remaining coordinates with white noise. The standardization $Y = V_f^{-\frac{1}{2}}$ preserves the spiral structure(Figure 7.5).

Our purpose is to find the two-dimensional needle in a high dimensional haystack. This example was used in Posse(1995) for comparing the efficiency of projection pursuit methods. It is a big challenge for an algorithm to find the spiral structure in high dimensional space, because "the density of the spiral is nearly normal i.e., nearly radial and decreasing when going away from the center"(Posse 1995).

First we use the same setting as Posse(1995): consider the spiral structure in R^8 and use the sample size of $n = 400$. The first two principal components from the estimated J_{f_2} reveal the spiral structure very well (Figure 7.6). This structure from the eigenanalysis of J_{f_2} is similar to the result of Posse's method, which has been shown better than Friedman's algorithm in this example. However, the eigenanalysis of Q_f does not reveal the spiral structure (Figure 7.7). The reason is that the sample size $n = 400$ is too small for the dimensionality $d = 8$. When the dimensionality d increases to ten,

the structure found by the eigenanalysis of J_{f_2} becomes vague (Figure 7.8). In order to reveal the structure, we need a bigger sample size(n=800)(Figure 7.9). We conclude that J_{f_2} was remarkably successful in finding a spiral buried in a haystack of white noise.

Now we use different smoothing parameters $H = hI_d$ to estimate Q_f, and then reveal the spiral structure in R^3 using the eigenanalysis of the estimated Q_f. The sample size is $n = 700$. When $h = h_{op} = 0.3799$, the eigenanalysis of Q_f successfully reveals the spiral structure(Figure 7.10). As h increases, the revealed structure becomes vague(Figure 7.11, Figure 7.12). When $h = 0.8$, the eigenanalysis of Q_f fails(Figure 7.13). When h is smaller than $h_{op}(h = 0.2)$, the revealed spiral structure becomes much clearer(Figure 7.14). But if h is too small($h = 0.04$), the revealed structure becomes vague again(Figure 7.15).

This example reveals an important aspect of our problem, that the choice of bandwidth can have a critical role in the success of our methods. We will not pursue bandwidth selection further in this thesis, but leave it to future work. One immediate practical solution is to try a range of bandwidths.

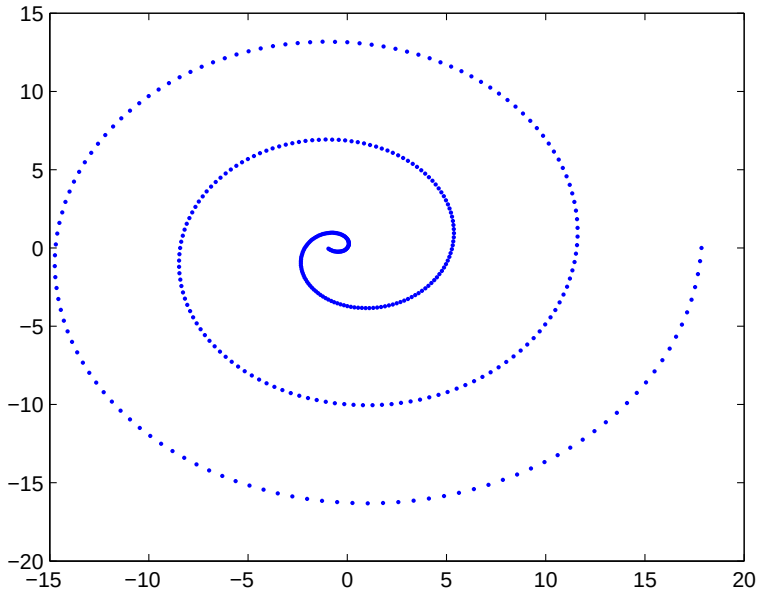

Fig. 7.4 The spiral structure in the first two-dimensional space.

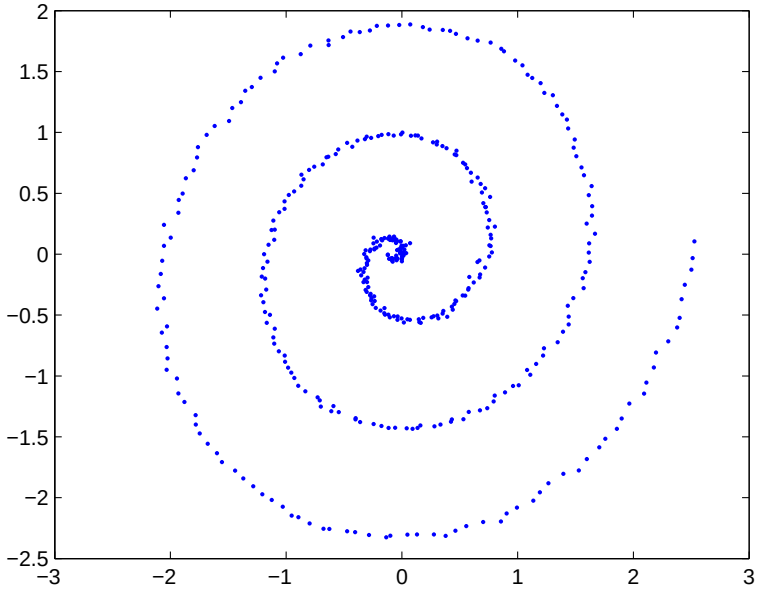

Fig. 7.5 The spiral structure in the first two-dimensional space after the standardization $Y = V_f^{-\frac{1}{2}}$.

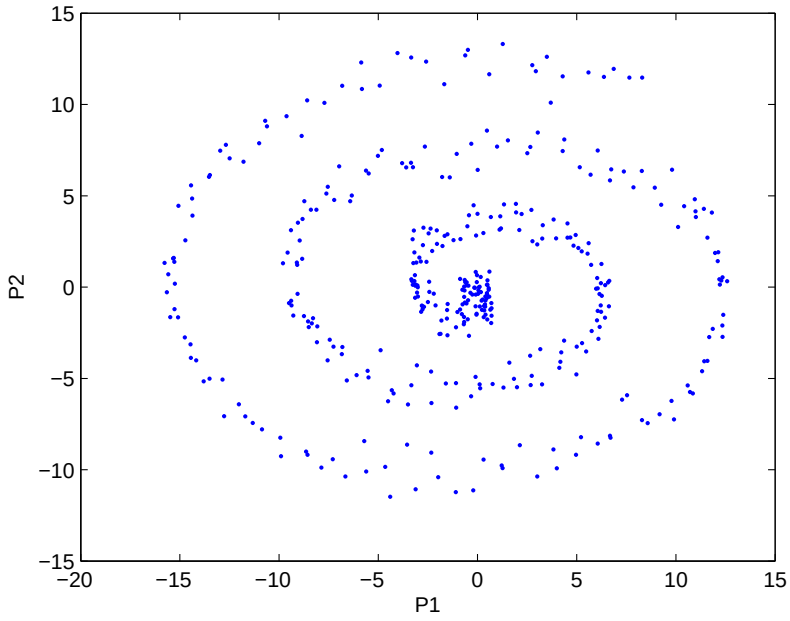

Fig. 7.6 The structure found by the eigenanalysis of $J_{\hat{f}_2}$, $d = 8, n = 400$

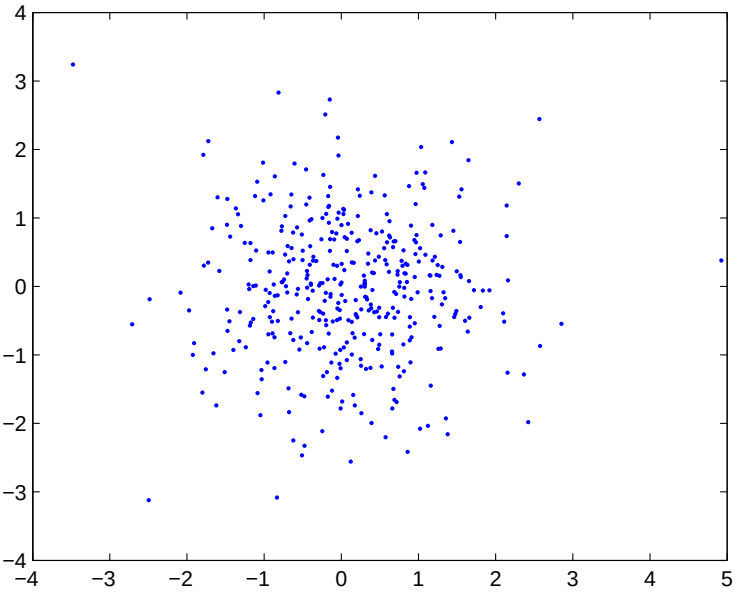

Fig. 7.7 The structure found by the eigenanalysis of Q_f, $d = 8, n = 400$

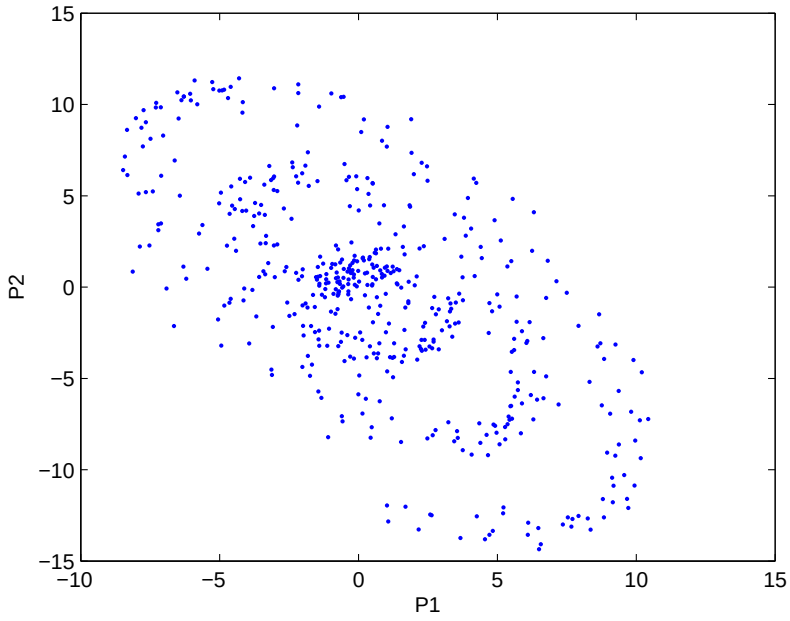

Fig. 7.8 The structure found by the eigenanalysis of $J_{\hat{f}_2}$, $d = 10, n = 400$

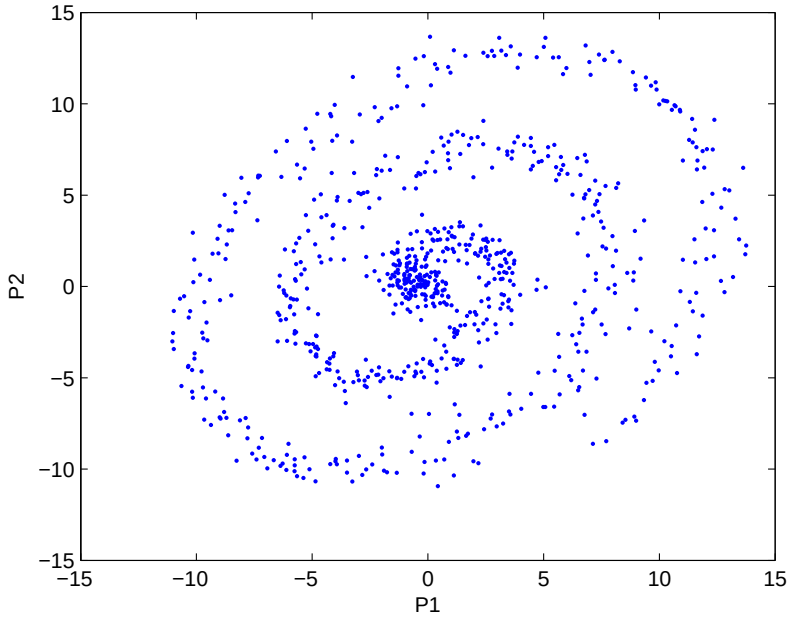

Fig. 7.9 The structure found by the eigenanalysis of $J_{\hat{f}_2}$, $d = 10, n = 800$

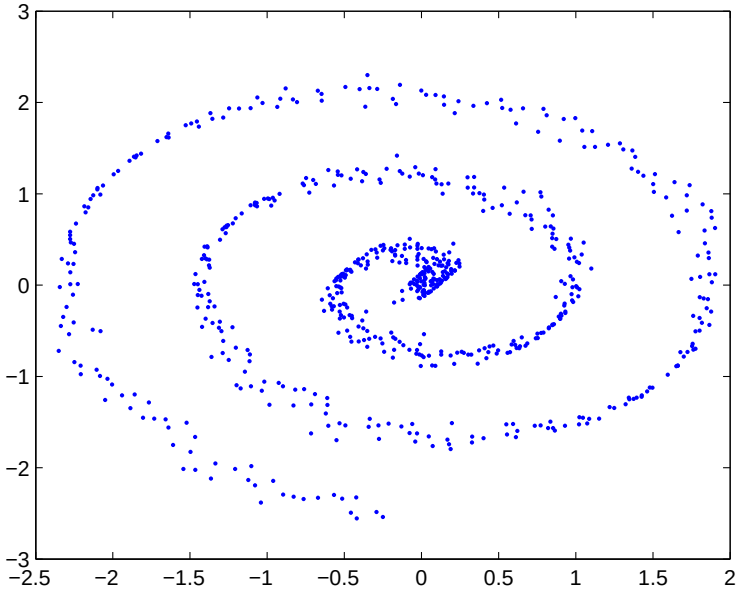

Fig. 7.10 The structure found by the eigenanalysis of Q_f, $d = 3, n = 700, h = 0.3799$

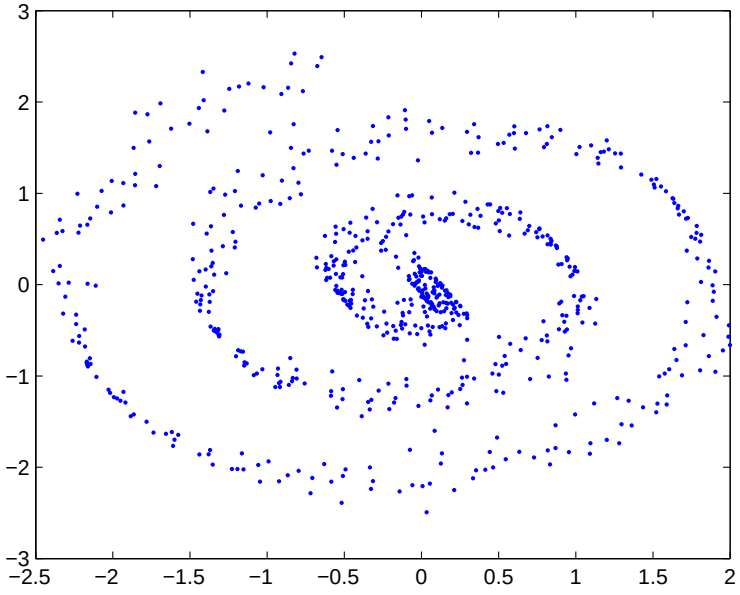

Fig. 7.11 The structure found by the eigenanalysis of Q_f, $d = 3, n = 700, h = 0.5$

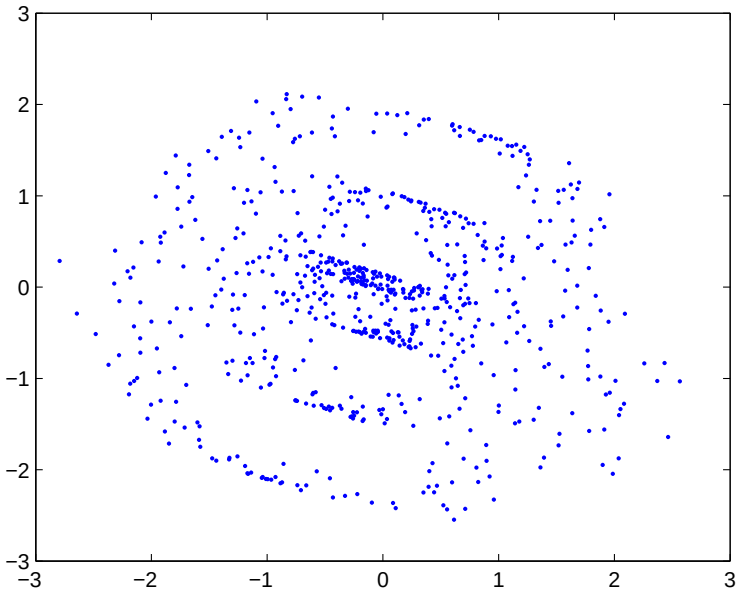

Fig. 7.12 The structure found by the eigenanalysis of Q_f, $d = 3, n = 700, h = 0.6$

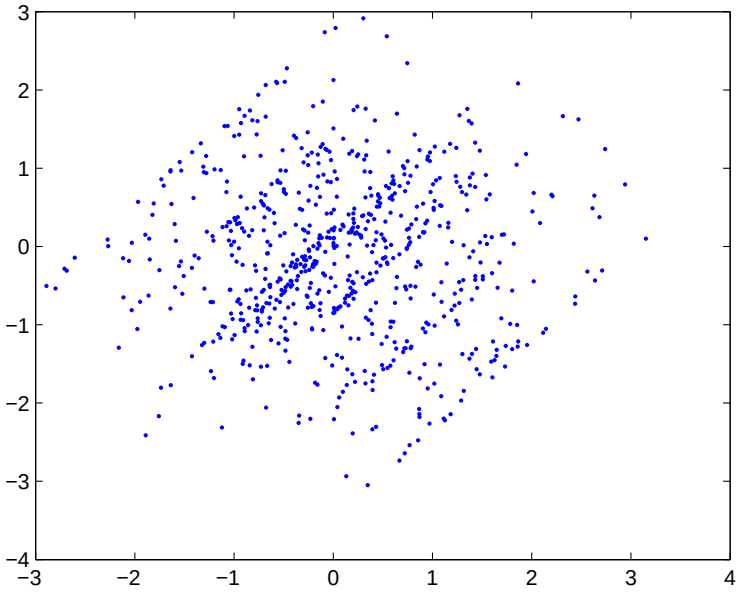

Fig. 7.13 The structure found by the eigenanalysis of Q_f, $d = 3, n = 700, h = 0.8$

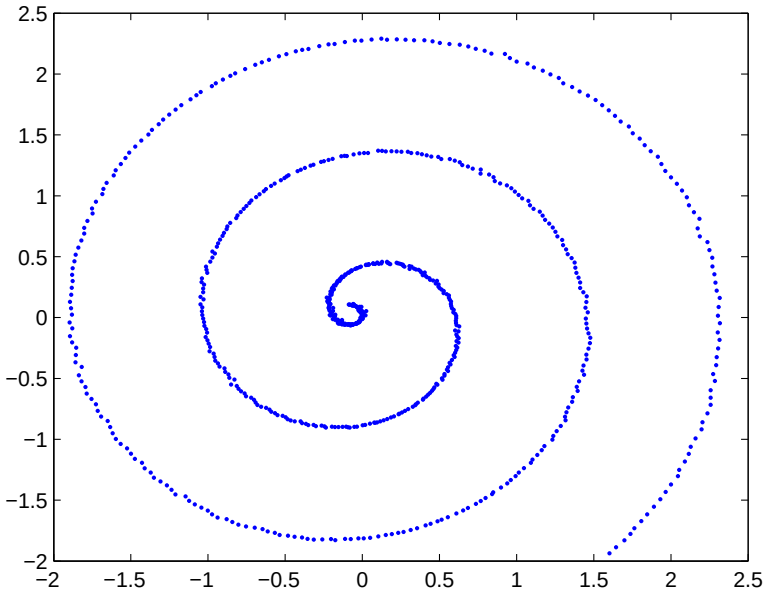

Fig. 7.14 The structure found by the eigenanalysis of Q_f, $d = 3, n = 700, h = 0.2$

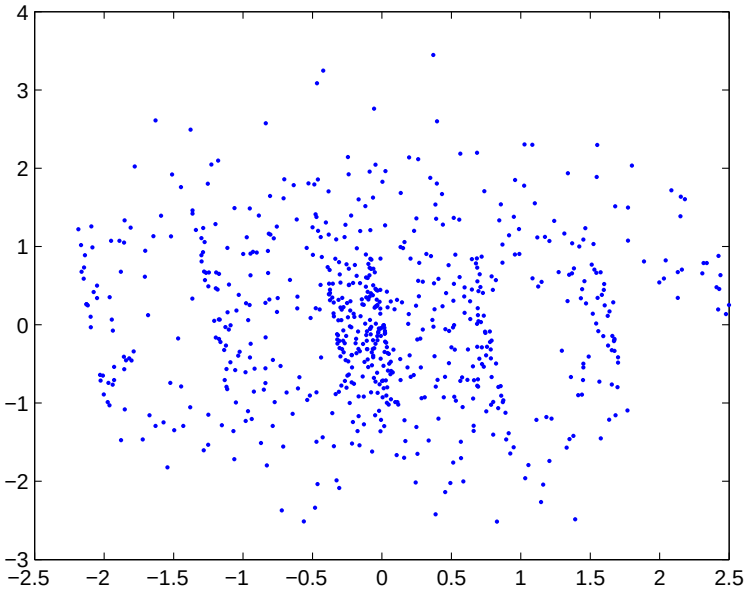

Fig. 7.15 The structure found by the eigenanalysis of Q_f, $d = 3, n = 700, h = 0.04$

7.2.4 White Noise Detection

Suppose the eigenvalues of J_{f_2} are ordered: $\lambda_1 \geq \lambda_2 \geq ... \geq \lambda_d$ with corresponding eigenvectors are $\gamma_1, \gamma_2, ..., \gamma_d$. Suppose the solution projection P has the density h. The Fisher information J_{h_2} is a diagonal matrix $J_{h_2} = diag(\lambda_1, \lambda_2, ...\lambda_d)$. The eigenvalue λ_i is the measure of the non-normality of the corresponding projection $P_i = \gamma_i X$. In theory, when the eigenvalue λ_k reach a lower bound 0.25, then the eigenvalues $\lambda_i = 0.25, i \geq k$, and the corresponding projections $P_i = \gamma_i X$ are white noise coordinates, which are standard normal and independent of the other projections. However, in practice, the eigenvalues from white noise are much bigger than 0.25, unless the sample size is huge.

In this subsection, we will propose a sequential test to detect the white noise coordinates within the solution projections from the eigenanalysis of J_{f_2}. First we test the null hypothesis that all eigenvalues are equal to 0.25. The alternative hypothesis is that the largest eigenvalue $\lambda_1 \geq 0.25$. If we reject the null hypothesis, we will consider the next hypothesis:$H_0 : \lambda_2 = \lambda_3 = ... = \lambda_d = 0.25$ vs $H_a : \lambda_2 \geq 0.25$. We propose to continue in this fashion until we fail to reject.

For the general null hypothesis $H_0 : \lambda_k = \lambda_{k+1} = ... = \lambda_d = 0.25$ vs the alternative hypothesis $H_a : \lambda_k \neq 0.25$, we propose two different test statistics: $\hat{\lambda}_k$ and $\hat{S}_k = \sum_{j=k}^{d} \hat{\lambda}_j$. Under the null hypothesis, the projections $(P_k, P_{k+1}, ..., P_d)$ are white noise coordinates. Suppose $Z_k, Z_{k+1}, ..., Z_d$ are independent standard normal variables. Then the variable vector $P^* = (P_1, P_2, ..., P_{k-1}, Z_k, Z_{k+1}, ..., Z_d)$ should have the same distribution as the solution projection $P = (P_1, P_2, ..., P_k, P_{k+1}, ..., P_d)$. For a fixed sample

size n and dimensionality d, we draw 1000 random samples of size n from a $d-k+1$-dimensional standard normal distribution. For every sample $(Z_k, Z_{k+1}, ..., Z_d)$, we use the data $(P_1, P_2, ..., P_{k-1}, Z_k, Z_{k+1}, ..., Z_d)$ to estimate the Fisher information matrix. The jth eigenvalue is a sample of $\lambda_j, k \leq j \leq d$. We construct the empirical distributions of $\hat{\lambda}_k$ and $\hat{S}_k = \sum_{j=k}^{d} \hat{\lambda}_j$ using the 1000 samples, and then get the critical values $\hat{F}_{\lambda_k,0.05}$ and $\hat{F}_{S_k,0.05}$. If the estimated λ_k (S_k) is less than the critical value $\hat{F}_{\lambda_k,0.05}$ ($\hat{F}_{S_k,0.05}$), we will fail to reject the null hypothesis: $H_0 : \lambda_k = \lambda_{k+1} = ... = \lambda_d = 0.25$, i.e., we think that there are $d-k+1$ white noise coordinates.

When we test the null hypothesis that all eigenvalues are equal to 0.25, another way to find the critical value of the trace $\hat{S}_1 = \sum_{j=1}^{d} \hat{\lambda}_j$ is the asymptotic distribution we found in Chapter 6. The critical values from random samples and the asymptotic distributions are listed in Table 7.4 for some n and d. The results show that the Satterthwaite approximation provides good approximation for fairly big sample size.

After a dimension reduction that removes all white noise coordinates, all remaining projections are significantly non-normal, then similar to principal component analysis, one can use the cumulative proportion of eigenvalues $\sum_{i=1}^{k} \lambda_i / \sum_{i=1}^{d} \lambda_i$ as an index of how much of the non-normality of the data explained by the selected projections: $P_1, ..., P_k$.

We will use these tests and diagnostics on our examples in the next section.

Table 7.4 Critical values of S_1: $\hat{F}_{S_1,0.05}$ from 1000 random normal samples; $\hat{F}^*_{S_1,0.05}$ from asymptotic distributions

(n,d)	$(50,7)$	$(100,4)$	$(150,4)$	$(392,6)$	$(500,7)$
$\hat{F}_{S_1,0.05}$	6.4963	1.7884	1.9006	3.3830	4.6306
$\hat{F}^*_{S_1,0.05}$	9.1469	1.9072	1.7825	3.4430	4.9274

7.3 Real Data Analysis

The following real data sets have been used to illustrate projection pursuit methods by Friedman and Tukey (1974), Friedman (1987), and Jee (1985). We will apply the eigenanalysis of the estimated Fisher information matrix $J_{\hat{f}_2}$ and the Von-Mises approximation Q_f to these data sets, and compare the results with those from the above three algorithms.

7.3.1 Particle Physics Data

This data set, having 500 observations, was derived from a high-energy particle physics scattering experiment(Ballam 1974, Friedman 1974, and Jee 1985). In the nuclear reaction, a positively charged pi-meson becomes a proton, two positively charged pi-mesons and a negatively charged pi-meson. Every observation consists of seven independent measurements.

The results of the eigenanalysis of $J_{\hat{f}_2}$ are listed in Table 7.5. The results from our two tests are contradictory. According to the test procedure based on $\hat{\lambda}_k$, all solution projections are sequentially found to be significantly non-normal because all eigenvalues are bigger than the corresponding critical values. However, the sum of all eigenvalues is less than the critical value 4.6306, so on this basis we would not reject the hypothesis that all solution projections are normal.

The first two largest classical principal components from the covariance matrix are shown in Figure 7.16, which indicates the data may consist of three clusters, but the structure is not very clear. The first two solution projections from $J_{\hat{f}_2}$ (Figure 7.18)

shows a triangular shape with points concentrated around two of three corners, which is obviously non-normal. So the test based on $\hat{\lambda}_k$ seems to better detect non-normal structure. The first two projections explains about 36.68% non-normality of the whole data.

The scatter plot of the first two solution projections from the index Q_f fails to reveal interesting structure because of its high requirement of sample size (Figure 7.17).

Jee (1985) also found the similar triangular structures using the trace of J_f as projection index. It verifies our conjecture that the density square transformation preserves the main structure of the original data.

Table 7.5 Eigenanalysis of $J_{\hat{f}_2}$ and critical values from 1000 samples for Particle Physics Data

k	1	2	3	4	5	6	7
λ_k	0.7697	0.7053	0.6492	0.5196	0.5055	0.4453	0.3930
$\hat{F}_{\lambda_k,0.05}$	0.7170	0.6978	0.6685	0.4069	0.5069	0.4095	0.3420
$\sum_{i=k}^{d} \lambda_i$	3.9876	3.2179	2.5126	1.8634	1.3438	0.8383	0.3930
$\hat{F}_{S_k,0.05}$	4.6306	3.8902	3.1485	2.3925	1.4689	0.8001	0.3420
$\sum_{i=1}^{k} \lambda_i / \sum_{i=1}^{d} \lambda_i$	0.1930	0.3699	0.5327	0.6630	0.7898	0.9014	1.00

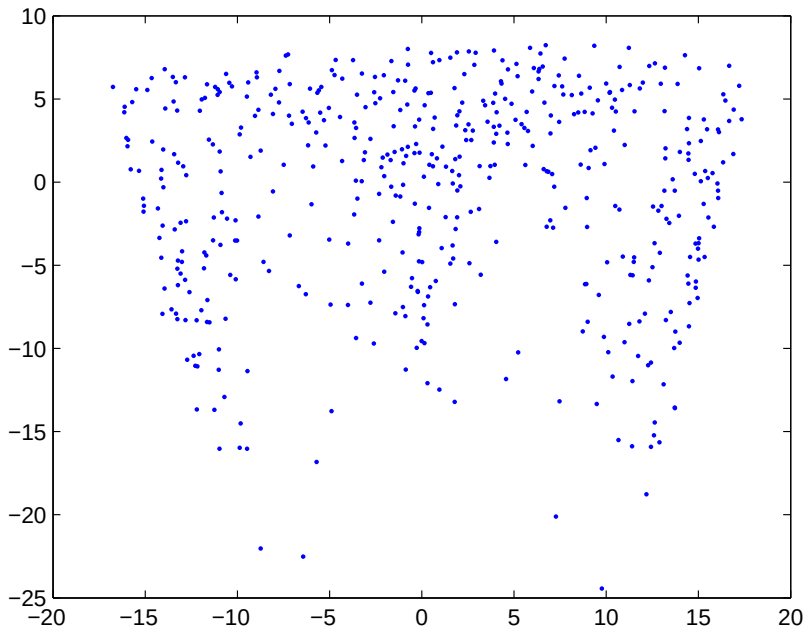

Fig. 7.16 Particle Physics Data
The scatter plot of the first two largest principal components

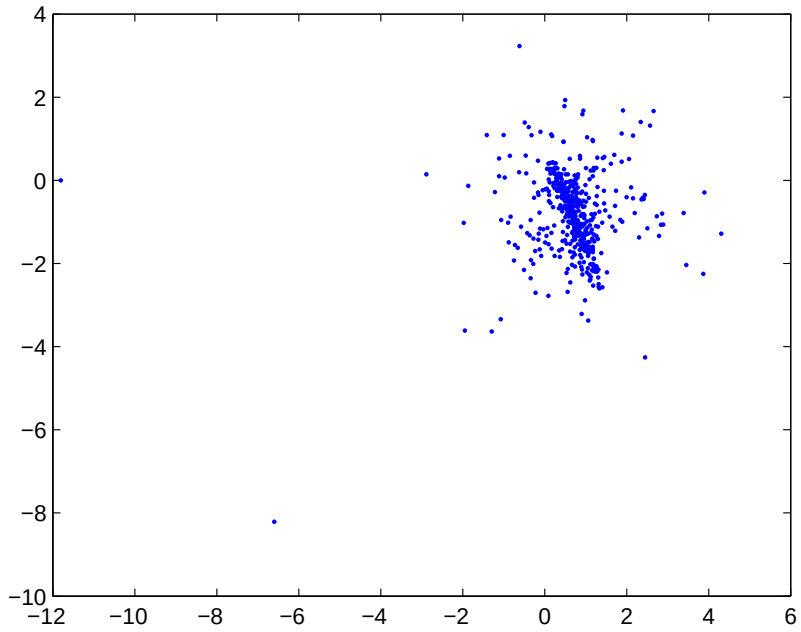

Fig. 7.17 Particle Physics Data
The scatter plot of the first two largest principal components from Q_f

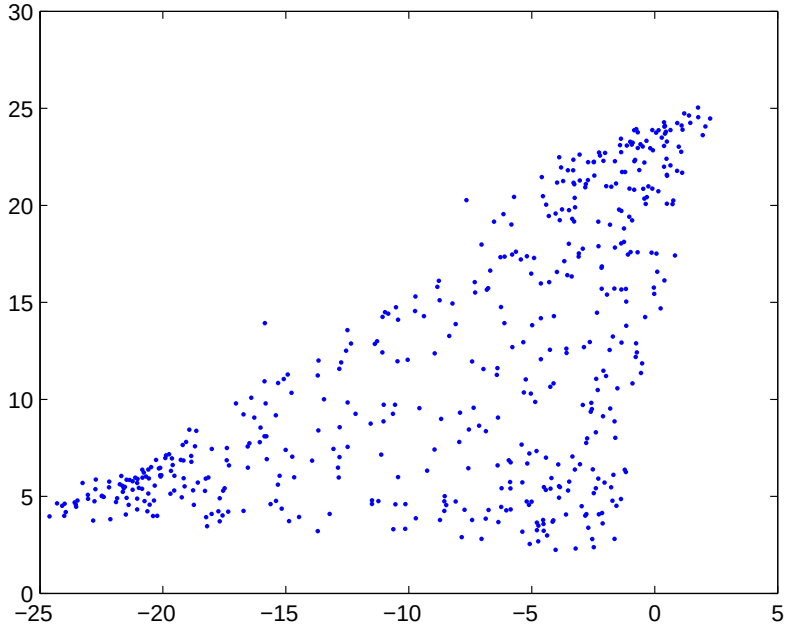

Fig. 7.18 Particle Physics Data
The scatter plot of the two-dimensional solution projections from $\hat{J}_{f_2}$

7.3.2 Iris Data

This well known data set was used by Fisher and many researchers. It contains three classes of fifty observations each, where each class refers to a species of iris plant. One class is quite different from the other two. Every observation consists of four independent measurements: sepal length, sepal width, petal length, petal width.

First we consider the eigenanalysis of Q_f. The plots of the solution projections do not reveal the structure of the iris data (Figure 7.19, Figure 7.20).

For the Fisher information matrix J_{f_2}, the results of the eigenanalysis for the whole iris data are listed in Table 7.6. According to the critical values from 1000 samples, both tests agree that only the first solution projection is significantly non-normal. And the first solution projection explains 48.69% of non-normality of the whole data. The histogram of the one-dimensional solution projection and the scatter plot of the two-dimensional solution projections are shown in Figure 7.21 and Figure 7.22. The projected data are well separated into two clusters: the first 50 observations(one species) and the remaining 100 observations(two more species). The 100 observations are also separated into clusters according to the true tags, but the boundary is not so apparent.

After deleting the first class data, we apply our eigenanalysis to the remaining 100 observations. The results of the eigenanalysis for the reduced iris data are listed in Table 7.7. Because the largest eigenvalue $\lambda_1 = 0.6581$ is bigger than the critical value $\hat{F}_{\lambda_1,0.05} = 0.5122$, but the second largest eigenvalue $\lambda_2 = 0.5007$ is less than the critical value $\hat{F}_{\lambda_2,0.05} = 0.5064$, we think that only the first solution projection is

significantly non-normal. The test based on $\hat{S}_k$ comes to the same conclusion. The non-normal projection explains 32.31% of non-normality of the data. The scatter plot of the two-dimensional projections from $\hat{J}_{\hat{f}_2^*}$ is shown in Figure 7.23. The result is similar to that of Friedman and Tukey(1974) and the two-dimensional Fisher linear discriminant projections (see Figure 2e of Friedman 1974). Because these two species are overlapped in the full four dimensional space(Sammon 1969, Zahn 1971, Friedman and Tukey 1974), it is virtually impossible to separate the two classes perfectly.

Table 7.6 Eigenanalysis of $J_{\hat{f}_2}$ and critical values from 1000 samples for the whole Iris Data

k	1	2	3	4
λ_k	1.2366	0.4766	0.4260	0.4006
$\hat{F}_{\lambda_k,0.05}$	0.5549	0.4871	0.4556	0.4160
$\sum_{i=k}^{d} \lambda_i$	2.5398	1.3032	0.8266	0.4006
$\hat{F}_{S_k,0.05}$	1.9006	1.3220	0.8572	0.4160
$\sum_{i=1}^{k} \lambda_i / \sum_{i=1}^{d} \lambda_i$	0.4869	0.6745	0.8423	1.00

Table 7.7 Eigenanalysis of $J_{\hat{f}_2}$ and critical values from 1000 samples for the remaining 100 observations

k	1	2	3	4
λ_k	0.6581	0.5007	0.4587	0.4192
$\hat{F}_{\lambda_k,0.05}$	0.5122	0.5064	0.4716	0.4423
$\sum_{i=k}^{d} \lambda_i$	2.0367	1.3786	0.8779	0.4192
$\hat{F}_{S_k,0.05}$	1.7884	1.3982	0.8942	0.4423
$\sum_{i=1}^{k} \lambda_i / \sum_{i=1}^{d} \lambda_i$	0.3231	0.5690	0.7942	1.00

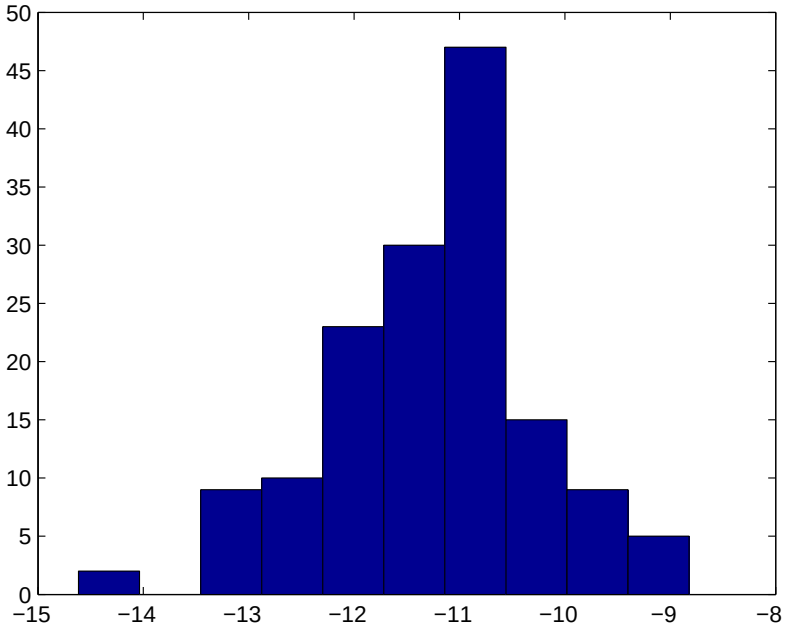

Fig. 7.19 Iris Data(150 points)
The histogram of the one-dimensional solution projection from Q_f

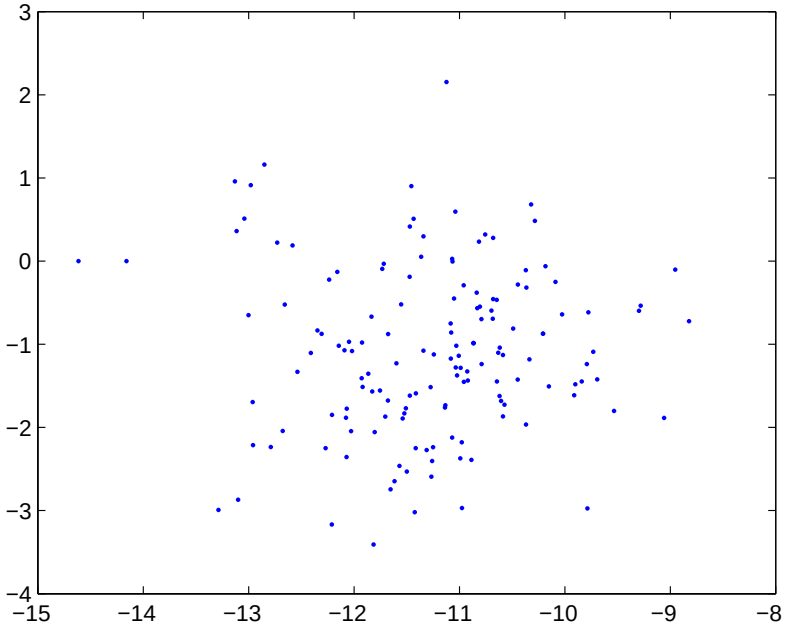

Fig. 7.20 Iris Data(150 points)
The scatter plot of the two-dimensional solution projections from Q_f

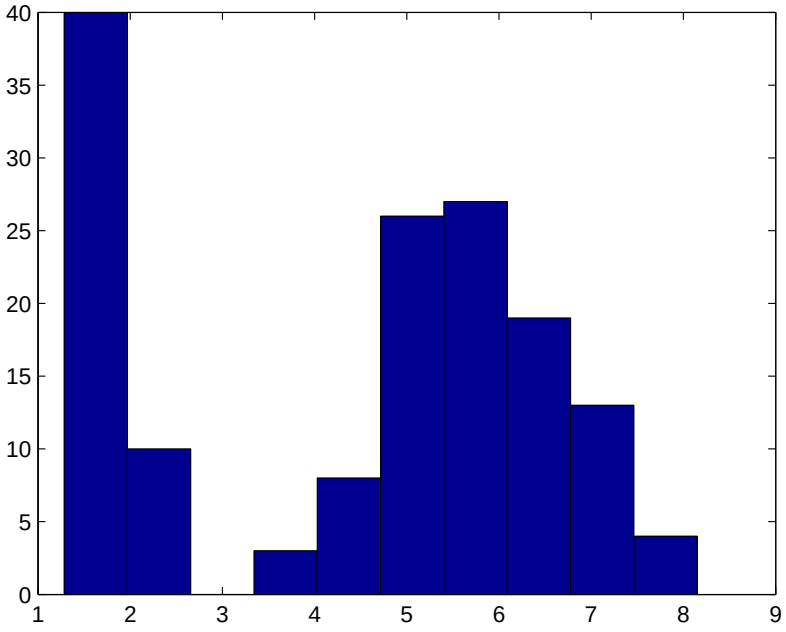

Fig. 7.21 Iris Data(150 points)
The histogram of the one-dimensional solution projection from $J_{\hat{f}_2}$

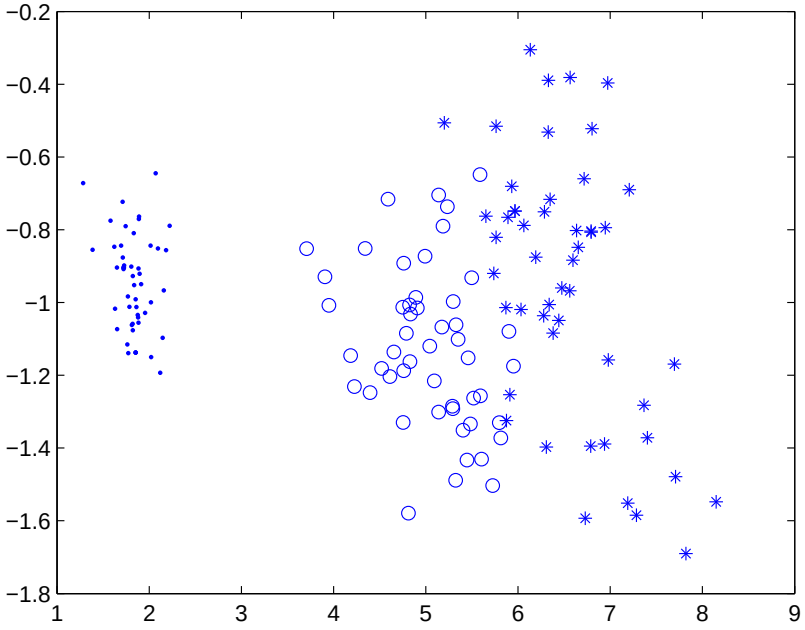

Fig. 7.22 Iris Data(150 points)
The scatter plot of the two-dimensional solution projections from $J_{\hat{f}_2}$

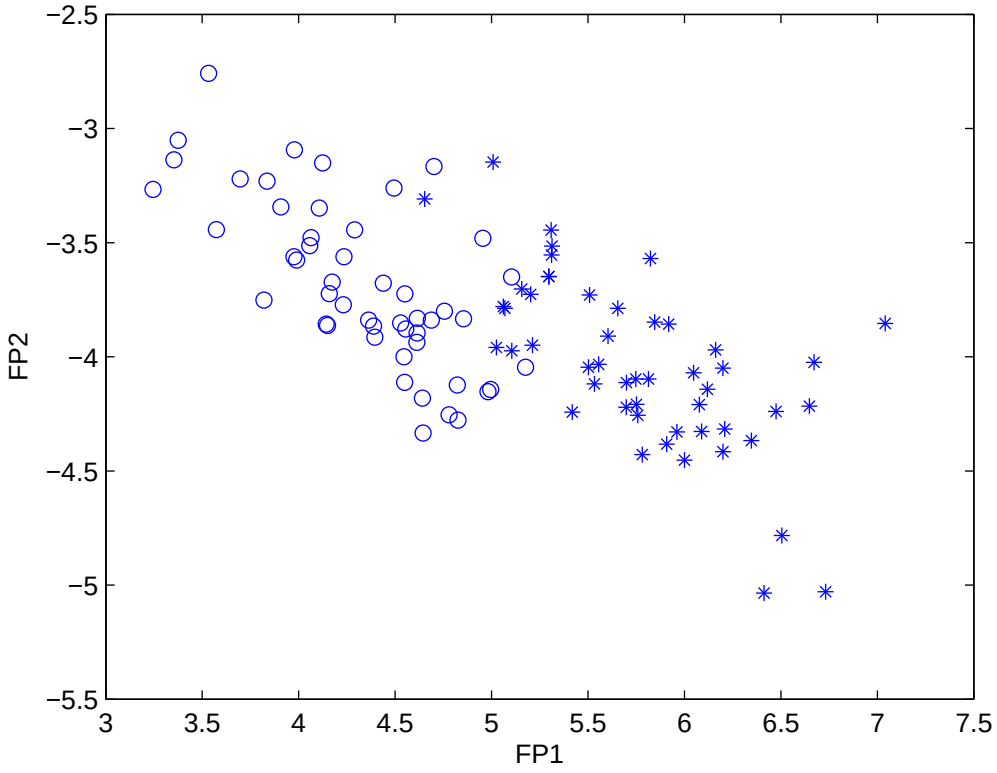

Fig. 7.23 Iris Data(100 points)
The scatter plot of the two-dimensional solution projections from $J_{\hat{f}_2}$

7.3.3 States Data

This data set was used in Becker and Chambers (1984) and Friedman (1987). It includes seven summary variables for 50 United States (from Table 1 of Friedman 1987):

1. Y_1 population estimate as of July 1, 1975;
2. Y_2 average income (1974);
3. Y_3 illiteracy rate (1970);
4. Y_4 life expectancy (1969-1971);
5. Y_5 homicide rate (1976);
6. Y_6 high-school graduation rate (1970);
7. Y_7 average number of days below freezing temperature (1931-1960) in capital or large city.

The sample size 50 is small for seven dimensions. We use the States data set to check the performance of our algorithm for a small sample size. Before estimating the Fisher information matrix J_{f_2} and Q_f, we standardize the variables to remove the scale influence.

The ordered eigenvalues of $J_{\hat{f}_2}$ and the cumulative proportions are listed in Table 7.9. We can not reject the null hypothesis: the whole data is normal, because the largest eigenvalue 0.9302 is less than the critical value 1.0600, and the sum of all eigenvalues 5.0858 is also less than the corresponding critical value 6.4963. Of course, the small sample size gives small power.

The scatter plot of the two-dimensional solution projections from J_{f_2} is shown in Figure 7.24. The 50 states are separated into two clusters, and two states are outliers. The 10 states in the smaller cluster are listed in Table 7.8. The first projection mainly depends on three variables: homicide rate (negative), high school graduation rate(positive), and illiteracy rate(negative):

$$P_1 = 0.0753Y_1 - 0.1080Y_2 - 0.2070Y_3 + 0.0003Y_4 - 0.4410Y_5 + 0.4383Y_6 + 0.0995Y_7.$$

Life expectancy is the least important index. The states in the larger cluster tend to have higher high-school graduation rate, lower homicide rate, and lower illiteracy rate. The means of these rates in the two clusters verify this conclusion (Table 7.10). The two outlier states are Alaska and New Mexico. Alaska has high rate of high-school graduation(0.67) and a high average income(6315). New Mexico does not have one extreme variable. It is distant from most other states because of its whole performance.

The scatter plot of the two-dimensional solution projections from Q_f is shown in Figure 7.25. The four states are outliers: New York, California, Alaska and Nevada. The solution projections do not separate the remaining states well.

Table 7.8 The ten states in the small cluster

Alabama	Arkansas	Georgia	Kentucky	Louisiana
Mississippi	North Carolina	South Carolina	Tennessee	Texas

Table 7.9 Eigenanalysis of $\hat{J}_{f_2}$ and critical values from 1000 samples for State Data

k	1	2	3	4	5	6	7
λ_k	0.9302	0.7600	0.7300	0.7081	0.6863	0.6541	0.6171
$\hat{F}_{\lambda_k,0.05}$	1.0600	1.0017	0.9332	0.8694	0.8394	0.7605	0.6784
$\sum_{i=k}^{d} \lambda_i$	5.0858	4.1556	3.3956	2.6656	1.9575	1.2712	0.6171
$\hat{F}_{S_k,0.05}$	6.4963	5.4761	4.3595	3.3208	2.4081	1.4835	0.6784
$\sum_{i=1}^{k} \lambda_i / \sum_{i=1}^{d} \lambda_i$	0.1829	0.3323	0.4759	0.6151	0.7500	0.8787	1.00

Table 7.10 Means of important variables in the two clusters

Variables	illiteracy rate	homicide rate	high-school graduation rate
Smaller Cluster	2.08	12.13	40.9
Bigger Cluster	0.89	5.96	55.91

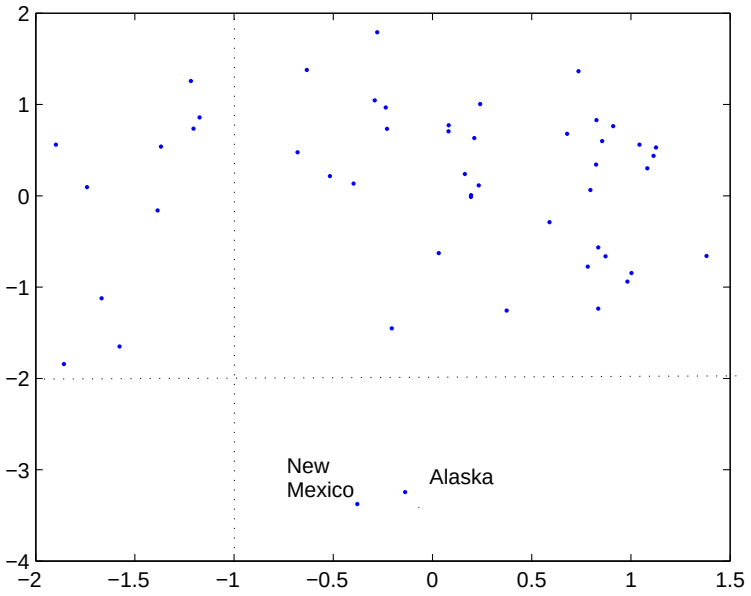

Fig. 7.24 State Data (d=7, n=50)
The scatter plot of the two-dimensional solution projections from $J_{\hat{f}_2}$

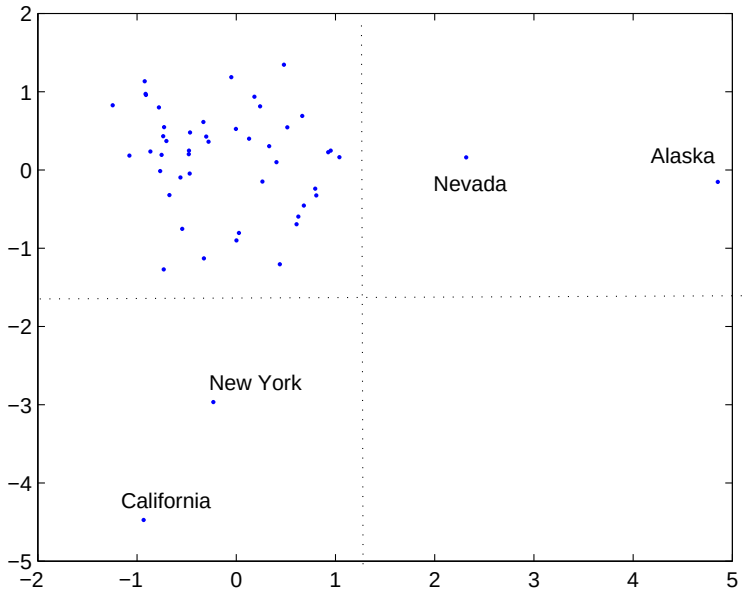

Fig. 7.25 State Data (d=7, n=50)
The scatter plot of the two-dimensional solution projections from Q_f

Friedman and Tukey(1987) analyzed the State data using their projection pursuit method. The results are similar, but the solution projections are different:

1. The 50 states were also separated into two clusters and two outliers. However, the smaller cluster contained the 10 states of our smaller cluster plus New Mexico and West Virginia. The two outliers were Alaska and Nevada.

2. The critical variables are different. In Friedman and Tukey(1987), the solution projection mainly depends on population, life expectancy and homicide rate.

7.3.4 Cars Data

The cars data consists of 392 complete observations on the following 8 variables: Y_1 MPG (miles per gallon), Y_2 number of cylinders, Y_3 engine size, Y_4 horsepower, Y_5 vehicle weight (lbs.), Y_6 time to accelerate from 0 to 60 mph (sec.), Y_7 model year (modulo 100), and Y_8 origin of car (American, European, or Japanese). We only use the first six variables because the last two variables are dummy variables. So we consider 392 points in a 6 dimensional space.

First we consider the eigenanalysis of Q_f. The histogram of the one-dimensional solution projection (Figure 7.26) and the scatter plot of the two-dimensional solution projections (Figure 7.27) show that the data are roughly separated into two clusters. One cluster mainly includes the European and Japanese cars. Most American cars are included in the second cluster.

The eigenanalysis results of J_{f_2} are listed in Table 7.11. Only the first solution projection is significantly non-normal according to the critical values. The first solution

projection explains about 35.56% non-normality of the whole cars data. The histogram and the scatter plot of the two-dimensional solution projections (Figure 7.28 ,Figure 7.29) show that the data are separated into three clusters. Most European and Japanese cars are in one cluster with a few American cars. The second cluster mainly consists of American cars, also includes several European cars and Japanese cars. The remaining American cars form the third cluster. The largest solution projection is $P_1 = -0.1186Y_1 - 1.0449Y_2 + 0.2076Y_3 + 0.2135Y_4 - 0.2922Y_5 - 0.0107Y_6$, which mainly depends on the number of cylinders in engine, and likely arises from the discreteness of this factor. Four Japanese cars associated with larger projected values are outliers. The separation is more clear in the plot of the three-dimensional solution projections (Figure 7.30). The plot of the solution projection of Friedman and Tukey also shows the trimodal pattern, but not as clear as ours (See Figure 7 in Friedman and Tukey 1987). The reason may be that the dummy variables, model year and origin of car, are included in their method and then data with the same values are randomly ordered.

Table 7.11 Eigenanalysis of $J_{\hat{f}_2}$ and critical values from 1000 samples for Cars Data

k	1	2	3	4	5	6
λ_k	1.3433	0.5825	0.5222	0.5008	0.4352	0.3936
$\hat{F}_{\lambda_k,0.05}$	0.6180	0.6046	0.5658	0.5280	0.4756	0.4232
$\sum_{i=k}^{d} \lambda_i$	3.7776	2.4343	1.8518	1.3296	0.8288	0.3936
$\hat{F}_{S_k,0.05}$	3.3830	2.7934	2.1496	1.5161	0.9201	0.4232
$\sum_{i=1}^{k} \lambda_i / \sum_{i=1}^{d} \lambda_i$	0.3556	0.5098	0.6480	0.7806	0.8958	1.00

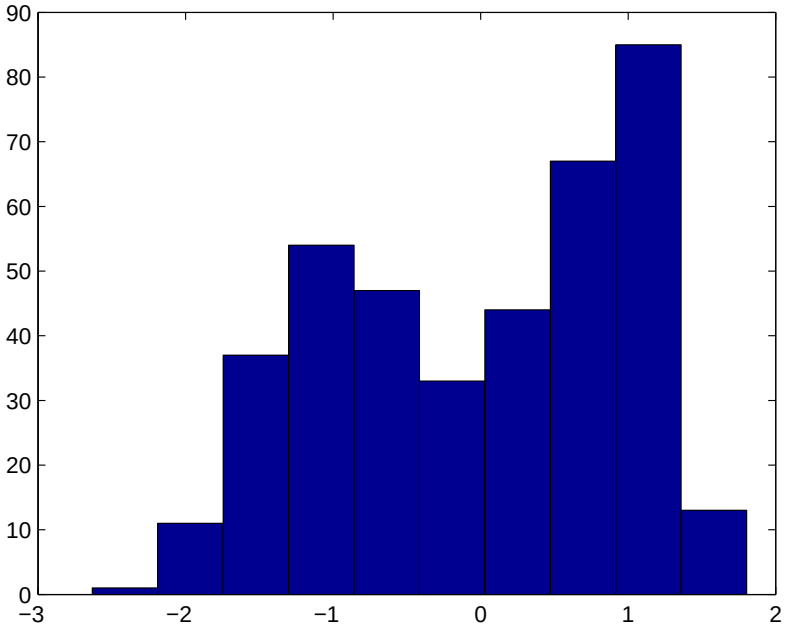

Fig. 7.26 Cars Data (d=7, n=50)
The histogram of the one-dimensional solution projection from Q_f

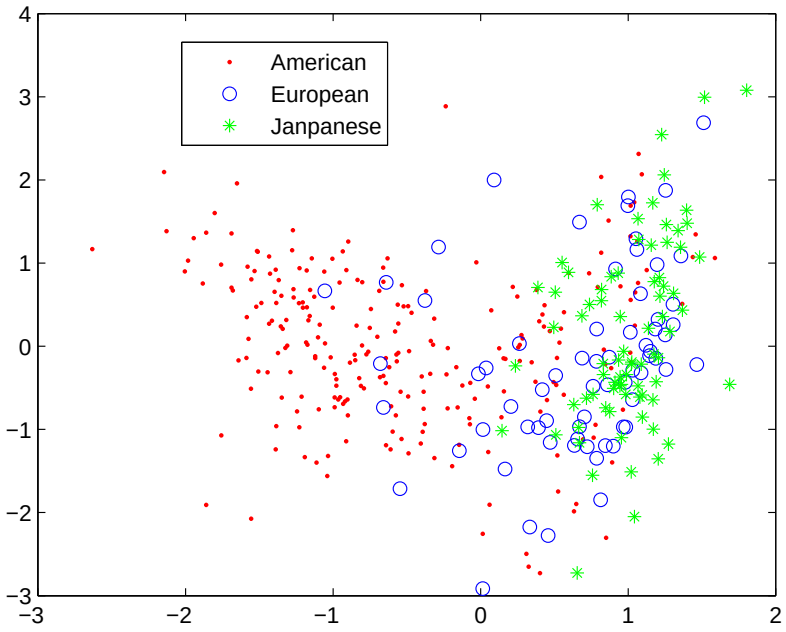

Fig. 7.27 Cars Data (d=7, n=50)
The scatter plot of the two-dimensional solution projections from Q_f

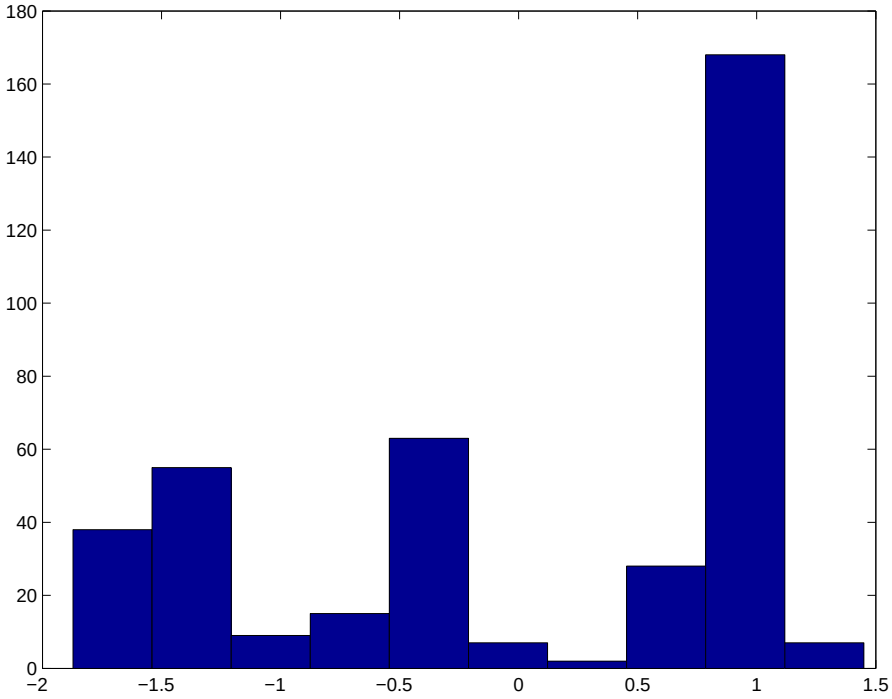

Fig. 7.28 Cars Data (d=7, n=50)
The histogram of the one-dimensional solution projection from $J_{\hat{f}_2}$

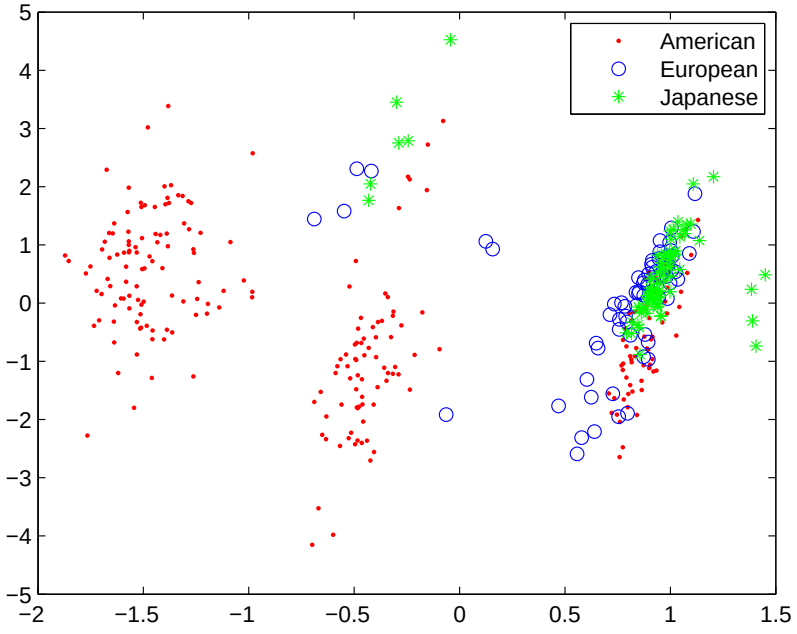

Fig. 7.29 Cars Data (d=7, n=50)
The scatter plot of the two-dimensional solution projections from $J_{\hat{f}_2}$

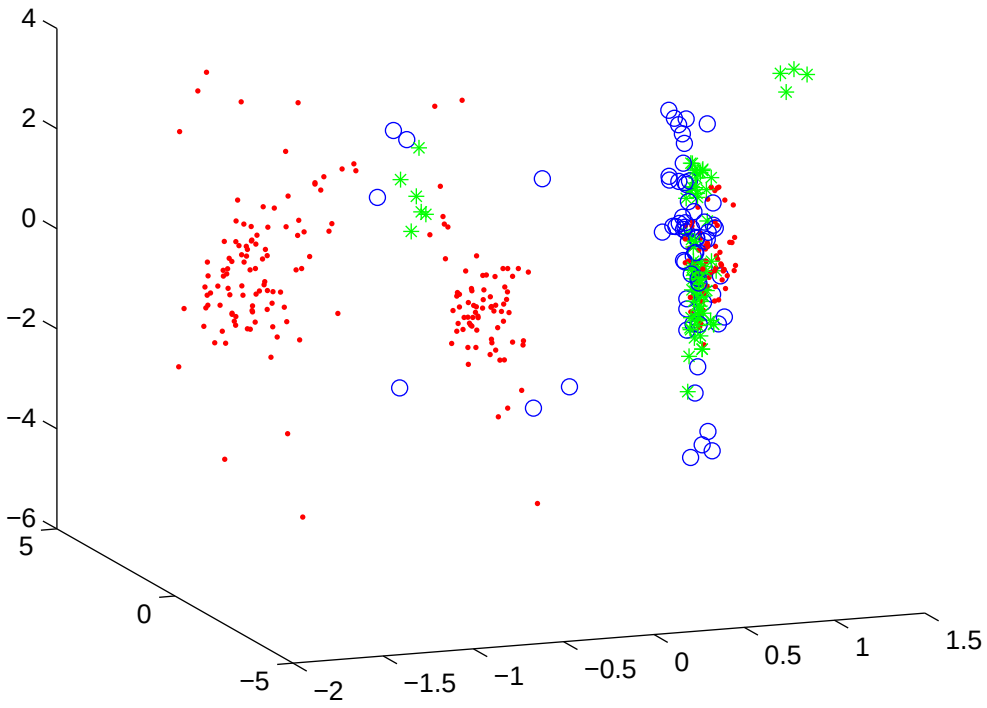

Fig. 7.30 Cars Data (d=7, n=50)
The scatter plot of the three-dimensional solution projections from $J_{\hat{f}_2}$

Chapter 8

Conclusion and Future Work

8.1 Projection Index J_{f_2}

In the simulated samples and real data analysis in Section 7.3 and 7.4, the projection pursuit method based on the eigenanalysis of J_{f_2} successfully revealed the interesting non-linear structures in fairly high dimensions with a practical sample size. Compared to current projection indices, the matrix J_{f_2} has been shown to be a rapidly computable and effective projection matrix index. The aspects contributing to its good performance include

1. An eigenanalysis of Fisher information matrix provides all solution projections; No sequential procedure is needed.

2. The density square transformation makes Fisher information matrix easy to estimate.

3. The density square transformation preserves important geometric structures including normality, and decreases the variance. A standard bandwidth selection method seems to provide good results.

4. After density square transformation, the new Fisher information matrix J_{f_2} measures non-normality in the main body of the distribution rather than in the tails.

In the simulation analysis(Section 7.3), we checked two conjectures:

1. The eigenvectors of J_{f_2} tend to be the same as those from the eigenanalysis of J_f, because density square transformation preserves most important structures.

2. The optimal projections from the matrix index J_{f_2} which have least conditional normality, also tend to have poor marginal normality.

In these examples, these conjectures are true. However, it is just preliminary conclusion due to the small number of population examples.

The efficiency of the algorithm in locating interesting structures in high dimensional space depends on the sample size and the smoothing parameter H. For the projection index J_{f_2}, we recommend the "optimal" bandwidth

$$H_{op} = \Big(\frac{4}{d+2}\Big)^{\frac{1}{d+4}} \Sigma^{1/2} n^{-\frac{1}{d+4}},$$

because of its stable performance in simulation study and real data analysis.

In the application to the real data sets, we have presented the low-(one-,two-,three-) dimensional projections, and showed the improvement over competitors in revealing the most interesting views of the whole data, like clustering, outliers. However, we should be cautious about interpretation on the revealed structures, because projection pursuit is only a part of exploratory data analysis, providing the most informative projections for further study.

8.2 Projection Index Q_f

In theory, the eigenanalysis of projection index Q_f provides all solution projections, but it is not a robust statistic because of the weight function ϕ^{-1}. As a non-normality measure, Q_f puts more weight on tail departures, which may be good for a normality test, but not so good for features detection. The requirement for a large sample size makes this projection pursuit not practically useful without further bandwidth tuning, as shown in real data analysis.

We do think the index Q_f needs further study, because it is not only the Von-Mises approximation of Fisher information J_f, but also a direct non-normality measure of f. The only problem is the denominator. To construct a goodness-of fit test, we could use a different weight function, for example ϕ or any constant, to make the statistic more robust. But the transformation equation (5.2) does not hold any more, so we may not get useful projections from the eigenanalysis of the new matrix. More investigation is required to find new robust and rapidly computable projection indices.

8.3 Tests based on Eigenvalues

We proposed two sequential tests to detect the white noise coordinates using eigenvalues of the estimated matrix $\hat{J}_{f_2}$. According to their performance in above examples, we think that the test based on ordered eigenvalues λ_k is more powerful than the tests based on the sum of eigenvalues S_k. The possible reason is that the test based on ordered eigenvalue λ_k only considers the largest eigenvalue for the tested projections, but the latter measures a sum of eigenvalues.

In Chapter 6, by Von-Mises expansion and spectral decomposition, we found the asymptotic distribution of the trace $\sum_{i=1}^{d} \lambda_i$, and the form of asymptotic distribution of $\hat{J}_{f_2}$, which is a weighted sum of independent Wishart variables. However, there is still no easy way to get the asymptotic distribution of eigenvalues from the asymptotic distribution of $\hat{J}_{f_2}$, for example, partial sum of eigenvalues $\sum_{i=2}^{d} \lambda_i$. We discussed how to use simulation to get the critical values for our proposed test statistics. A rigorous analysis is proposed for our future work.

8.4 Mixture Direction

We have successfully applied the eigenanalysis of matrix distances to projection pursuit. The same eigenanalysis idea can also be used to find number of components in a mixture model. We will present an example to illustrate it.

Consider a two-component normal mixture model:

$$f(x) = \pi_1 \phi(x, \mu_1, \Sigma_1) + \pi_2 \phi(x, \mu_2, \Sigma_2).$$

Suppose both components have the same covariance matrix:$\Sigma_1 = \Sigma_2$. Without losing generality, we assume $\Sigma_1 = \Sigma_2 = I_d$. Let $\mu = E(X)$. We define a matrix to find the number of components:

$$A_f = \int \Big(\frac{\nabla_x f(x) + x f(x)}{f(x)} - \mu \Big) \Big(\frac{\nabla_x f(x) + x f(x)}{f(x)} - \mu \Big)^T f(x) dx.$$

When X is normal, $A_f \equiv 0$. Under the hypothesis: the normal mixture model has two components, we have

$$\begin{aligned} A_f &= \int \Big(\frac{\nabla_x f(x) + x f(x)}{f(x)} - \mu\Big)\Big(\frac{\nabla_x f(x) + x f(x)}{f(x)} - \mu\Big)^T f(x)dx \\ &= \int \Big(\frac{\pi_1 \phi(x, \mu_1, I_d)}{f(x)}\mu_1 + \frac{\pi_2 \phi(x, \mu_2, I_d)}{f(x)}\mu_2 - \mu\Big) \\ &\quad \Big(\frac{\pi_1 \phi(x, \mu_1, I_d)}{f(x)}\mu_1 + \frac{(1-\pi_1) \phi(x, \mu_2, I_d)}{f(x)}\mu_2 - \mu\Big)^T f(x)dx \\ &:= \int \Big(\alpha(x)\mu_1 + (1-\alpha(x))\mu_2 - \mu\Big)\Big(\alpha(x)\mu_1 + (1-\alpha(x))\mu_2 - \mu\Big)^T f(x)dx. \end{aligned}$$

Because $E(\alpha(X)\mu_1 + (1-\alpha(X))\mu_2) = \mu$, the matrix A_X is just the covariance matrix of the variable $\alpha(X)\mu_1 + (1-\alpha(X))\mu_2$. And any realization of $\alpha(X)\mu_1 + (1-\alpha(X))\mu_2$ is a convex combination of μ_1 and μ_2. So, the matrix A_f has only one non-zero eigenvalue and the corresponding eigenvector is $(\mu_2 - \mu_1)/||\mu_2 - \mu_1||$.

More generally, when the true model is a m-component normal mixture model and all components have the identity covariance matrix, the matrix A_f is the covariance matrix of $\sum_{k=1}^m \alpha_k(x)\mu_k$, which is a linear combination of $\mu_i, i = 1, ..., m$. So the number of non-zero eigenvalues of A_f is equal to the dimensionality of the plane determined by the mean points $\mu_i, i = 1, ..., m$. Note that the dimensionality may be less than $m-1$, because some mean points could be in the plane determined by other points. For example, μ_3 is on the line between μ_1 and μ_2. From a converse point of view, suppose we get $k \leq d$ non-zero eigenvalues from the eigenanalysis of A_f, the normal mixture model must have at least $k+1$ components.

The matrix A_f presents the same computation problem as the standard Fisher information matrix J_f. We can not use the density square transformation here, because

the number of components after transformation is changed. The Von-Mises method may help us to find an approximation, which is easy to estimate. But the Von-Mises approximation is not a covariance matrix of a convex combination of mean points. More investigation is needed to solve the computation problem.

Appendix

Proofs

PROOF 1 (**Proof of Proposition 1**). *First,*

$$\begin{aligned} 0_d &\leq E(V_f^{\frac{1}{2}}\nabla_x log(f) + V_f^{-\frac{1}{2}}x)(V_f^{\frac{1}{2}}\nabla_x log(f) + V_f^{-\frac{1}{2}}x)^T \\ &= V_f^{1/2}\Big(\int \frac{\nabla_x f \cdot \nabla_x f^T}{f} dx\Big)V_f^{1/2} + 2V_f^{\frac{1}{2}}\int \nabla_x log(f) x^T f(x) dx V_f^{-\frac{1}{2}} + V_f^{-\frac{1}{2}} E(XX^T) V_f^{-\frac{1}{2}}, \end{aligned}$$

where 0_d *is a* $d \times d$ *zero matrix. Consider*$\int \nabla_x log(f) x^T f(x) dx = \Big(\int \frac{\partial f(x)}{\partial x_i} x_j dx\Big)$.

$$\begin{aligned} & \int \frac{\partial f(x)}{\partial x_i} x_i dx \\ = & \int \Big(f(x)x_i|_{-\infty}^{\infty} - \int f dx_i\Big) dx_1 ... dx_{i-1} dx_i ... dx_d \\ = & 0 - \int f(x) dx \\ = & -1 \\ & \int \frac{\partial f(x)}{\partial x_i} x_j dx \\ = & \int \Big(\frac{\partial f(x)}{\partial x_i} dx_i\Big) x_j dx_1 ... dx_{i-1} dx_i ... dx_d \\ = & 0, i \neq j. \end{aligned}$$

Thus the second term of the right side is

$$2V_f^{\frac{1}{2}} \int \nabla_x log(f) x^T f(x) dx V_f^{-\frac{1}{2}} = -2I_d.$$

By the definition of covariance matrix, the third term is equal to I_d*. Moving the second and third term to another side yields*

$$V_f^{1/2}\Big(\int \frac{\nabla_x f \cdot \nabla_x f^T}{f} dx\Big)V_f^{1/2} \geq I_d.$$

□

PROOF 2 (**Proof of Proposition 2**). *Because* $Y = AX$*, we have*

$$g(y) = f(A^{-1}y)|A^{-1}|,$$

$$\nabla_y g(y) = A^{-T}\nabla_x f(A^{-1}y)|A^{-1}|.$$

Thus, we can conclude that

$$\begin{aligned}
J_g &= V_g^{1/2}\int \frac{\nabla_y g(y)(\nabla_y g(y))^T}{g(y)} dy V_g^{1/2} \\
&= V_g^{1/2}\int \frac{A^{-T}\nabla_x f(A^{-1}y)|A^{-1}||A^{-1}|(\nabla_x f(A^{-1}y))^T A^{-1}|}{f(A^{-1}y)|A^{-1}|} dy V_g^{1/2} \\
&= V_g^{1/2}A^{-T}\int \nabla_x f(x)(\nabla_x f(x))^T dx A^{-1} V_g^{1/2} \\
&= V_g^{1/2}A^{-T}\int \frac{\nabla_x f(x)(\nabla_x f(x))^T}{f(x)} dy A^{-1} V_g^{1/2} \\
&= V_g^{1/2}A^{-T}V_f^{1/2}J_f V_f^{1/2}A^{-1}V_g^{1/2}.
\end{aligned}$$

□

Proof 3 (**Proof of Proposition 3**). *We expand the two densities and merge the x terms:*

$$\begin{aligned}
& \phi(x,\mu_k,\Sigma_k)\phi(x,\mu_l,\Sigma_l) \\
= & (\frac{1}{2\pi})^{d/2}|\Sigma_k|^{-1/2} exp(-1/2(x-\mu_k)^T\Sigma_k^{-1}(x-\mu_k)) \\
& (\frac{1}{2\pi})^{d/2}|\Sigma_l|^{-1/2} exp(-1/2(x-\mu_l)^T\Sigma_l^{-1}(x-\mu_l)) \\
= & (\frac{1}{2\pi})^{d}|\Sigma_k|^{-1/2}|\Sigma_l|^{-1/2} \\
& exp(-1/2(x^T(\Sigma_k^{-1}+\Sigma_l^{-1})x - 2x^T(\Sigma_k^{-1}\mu_k+\Sigma_l^{-1}\mu_l) + \mu_k^T\Sigma_k^{-1}\mu_k + \mu_l^T\Sigma_l^{-1}\mu_l) \\
= & (\frac{1}{2\pi})^{d/2}\frac{|\Sigma_{kl}|^{1/2}}{|\Sigma_k|^{1/2}|\Sigma_l|^{1/2}} exp(-\frac{1}{2}(\mu_k^T\Sigma_k^{-1}\mu_k + \mu_l^T\Sigma_l^{-1}\mu_l - \mu_{kl}^T\Sigma_{kl}\mu_{kl})) \\
& (\frac{1}{2\pi})^{d/2}|\Sigma_{kl}|^{-1} exp(-1/2(x-\mu_{kl})^T\Sigma_{kl}^{-1}(x-\mu_{kl})) \\
= & (\frac{1}{2\pi})^{d/2}\frac{1}{|\Sigma_k+\Sigma_l|^{1/2}} exp(-\frac{1}{2}(\mu_k-\mu_l)^T(\Sigma_k+\Sigma_l)^{-1}(\mu_k-\mu_l))\phi(x,\mu_{kl},\Sigma_{kl}) \\
= & \phi(\mu_k,\mu_l,\Sigma_k+\Sigma_l)\phi(x,\mu_{kl},\Sigma_{kl}).
\end{aligned}$$

□

Proof 4 (**Proof of Proposition 4**). *Applying the above equation derives the result:*

$$\begin{aligned}
\int f^2 dx &= \Big(\int \sum_k \pi_k \phi(x,\mu_k,\Sigma_k)\Big)\Big(\int \sum_l \pi_l \phi(x,\mu_l,\Sigma_l)\Big) \\
&= \sum_k\sum_l \pi_k\pi_l\phi(\mu_k,\mu_l,\Sigma_k+\Sigma_l)\int \phi(x,\mu_{kl},\Sigma_{kl}) \\
&= \sum_k\sum_l \pi_k\pi_l\phi(\mu_k,\mu_l,\Sigma_k+\Sigma_l).
\end{aligned}$$

□

PROOF 5 (**Proof of Proposition 5**). *First,*

$$\nabla_x f(x) = \sum_i \pi_i \phi(x, \mu_i, \Sigma_i)(-\Sigma_i^{-1}(x - \mu_i)).$$

Thus,

$$\begin{aligned}
& \int \nabla_x f \cdot \nabla_x f^T dx \\
= & \sum_k \sum_l \pi_k \pi_l \Sigma_k^{-1} \int \phi(x, \mu_k, \Sigma_k)\phi(x, \mu_l, \Sigma_l)(x - \mu_k)(x - \mu_l)^T dx \Sigma_l^{-1} \\
= & \sum_k \sum_l \pi_k \pi_l \Sigma_k^{-1} \phi(\mu_k, \mu_l, \Sigma_k + \Sigma_l)[\Sigma_{kl} + (\mu_{kl} - \mu_k)(\mu_{kl} - \mu_l)^T]\Sigma_l^{-1} \\
= & \sum_k \sum_l \pi_k \pi_l \phi(\mu_k, \mu_l, \Sigma_k + \Sigma_l)[(\Sigma_k + \Sigma_l)^{-1} + (\Sigma_k + \Sigma_l)^{-1}(\mu_k - \mu_l)(\mu_k - \mu_l)^T(\Sigma_k + \Sigma_l)^{-1}].
\end{aligned}$$

□

PROOF 6 (**Proof of Proposition 8**).

$$\begin{aligned}
& \int (f^*)^2(x)dx \\
= & \int \Big(\int f(y) \frac{1}{|H|} K(H^{-1}(x - y))dy \Big) \Big(\int f(z) \frac{1}{|H|} K(H^{-1}(x - z))dz \Big) dx \\
= & \int \int \Big(\int \frac{1}{|H|} K(H^{-1}(x - y)) \frac{1}{|H|} K(H^{-1}(x - z))dx \Big) f(z)f(y)dydz \\
= & E_{Y,Z} \Big(\int \frac{1}{|H|} K(H^{-1}(x - Y)) \frac{1}{|H|} K(H^{-1}(x - Z))dx \Big) \\
:= & E_{Y,Z} h_0(Y, Z),
\end{aligned}$$

where $h_0(X_1, X_2)$ is an unbiased estimator for θ_0, and permutation symmetric in the 2 arguments.
So $U_0 \stackrel{\Delta}{=} \frac{1}{n(n-1)} \sum_{i \neq j} h_0(X_i, X_j)$ is the U-statistic with the kernel h_0 for θ_0. □

PROOF 7 (**Proof of Proposition 13**). *First, we consider*

$$
\begin{aligned}
& (U_2U_0 - U_3U_3^T)^{\frac{1}{2}} \\
= & \Big[\Big(P_0 + \theta_0 + o_p\Big(\frac{1}{\sqrt{n}}\Big)\Big)\Big(P_2 + \theta_2 + o_p\Big(\frac{1}{\sqrt{n}}\Big)\Big) - \Big(P_3 + \theta_3 + o_p\Big(\frac{1}{\sqrt{n}}\Big)\Big)\Big(P_3 + \theta_3 + o_p\Big(\frac{1}{\sqrt{n}}\Big)^T\Big]^{\frac{1}{2}} \\
= & \Big(\theta_0\theta_2 + \theta_0P_2 + \theta_2P_0 + o_p\Big(\frac{1}{\sqrt{n}}\Big)\Big)^{\frac{1}{2}} \\
= & \Big(\theta_0\theta_2\Big)^{\frac{1}{2}} + \frac{1}{2}\Big(\theta_0\theta_2\Big)^{-\frac{1}{2}}\Big(\theta_0P_2 + \theta_2P_0 + o_p\Big(\frac{1}{\sqrt{n}}\Big)\Big) + o_p\Big(\frac{1}{\sqrt{n}}\Big) \\
= & \Big(\theta_0\theta_2\Big)^{\frac{1}{2}} + \frac{1}{2}\Big(\theta_0\theta_2\Big)^{-\frac{1}{2}}\Big(\theta_0P_2 + \theta_2P_0\Big) + o_p\Big(\frac{1}{\sqrt{n}}\Big),
\end{aligned}
$$

thus, plugging it into $(U_2U_0 - U_3U_3^T)^{\frac{1}{2}}U_1(U_2U_0 - U_3U_3^T)^{\frac{1}{2}}$ *provides*

$$
\begin{aligned}
& (U_2U_0 - U_3U_3^T)^{\frac{1}{2}}U_1(U_2U_0 - U_3U_3^T)^{\frac{1}{2}} \\
= & \Big(\Big(\theta_0\theta_2\Big)^{\frac{1}{2}} + \frac{1}{2}\Big(\theta_0\theta_2\Big)^{-\frac{1}{2}}\Big(\theta_0P_2 + \theta_2P_0\Big) + o_p\Big(\frac{1}{\sqrt{n}}\Big)\Big) \\
& \cdot\Big(P_1 + \theta_1 + o_p\Big(\frac{1}{\sqrt{n}}\Big)\Big) \\
& \cdot\Big(\Big(\theta_0\theta_2\Big)^{\frac{1}{2}} + \frac{1}{2}\Big(\theta_0\theta_2\Big)^{-\frac{1}{2}}\Big(\theta_0P_2 + \theta_2P_0\Big) + o_p\Big(\frac{1}{\sqrt{n}}\Big)\Big) \\
= & \theta_0\theta_2P_1 + \theta_0\theta_2\theta_1 + \Big(\theta_0\theta_2\Big)^{\frac{1}{2}}\theta_1\frac{1}{2}\Big(\theta_0\theta_2\Big)^{-\frac{1}{2}}\Big(\theta_0P_2 + \theta_2P_0\Big) \\
& + \frac{1}{2}\Big(\theta_0\theta_2\Big)^{-\frac{1}{2}}\Big(\theta_0P_2 + \theta_2P_0\Big)\theta_1\Big(\theta_0\theta_2\Big)^{\frac{1}{2}} + o_p\Big(\frac{1}{\sqrt{n}}\Big) \\
= & \theta_0\theta_1\theta_2 + \theta_1\theta_2P_0 + \theta_0\theta_2P_1 + \theta_0\theta_1P_2 + o_p\Big(\frac{1}{\sqrt{n}}\Big),
\end{aligned}
$$

then

$$\begin{aligned}
& (U_2U_0 - U_3U_3^T)^{\frac{1}{2}}U_1(U_2U_0 - U_3U_3^T)^{\frac{1}{2}}/(U_0^3) - \frac{1}{4}I \\
= & \frac{\theta_0\theta_1\theta_2 + \theta_1\theta_2P_0 + \theta_0\theta_2P_1 + \theta_0\theta_1P_2 + o_p\Big(1/\sqrt{n}\Big)}{U_0^3} - \frac{\theta_0\theta_1\theta_2}{\theta_0^3} \\
= & \frac{\theta_0^4\theta_1\theta_2 + \theta_0^3\theta_1\theta_2P_0 + \theta_0^4\theta_2P_1 + \theta_0^4\theta_1P_2 - U_0^3\theta_0\theta_1\theta_2}{U_0^3\theta_0^3} + o_p\Big(1/\sqrt{n}\Big) \\
= & \frac{\theta_0^3\theta_1\theta_2P_0 + \theta_0^4\theta_2P_1 + \theta_0^4\theta_1P_2 - (U_0^3 - \theta_0^3)\theta_0\theta_1\theta_2}{U_0^3\theta_0^3} + o_p\Big(1/\sqrt{n}\Big) \\
= & \frac{\theta_0^3\theta_1\theta_2P_0 + \theta_0^4\theta_2P_1 + \theta_0^4\theta_1P_2 - (U_0 - \theta_0)(U_0^2 + \theta_0U_0 + \theta_0^2)\theta_0\theta_1\theta_2}{U_0^3\theta_0^3} + o_p\Big(1/\sqrt{n}\Big) \\
= & \frac{\theta_0^3\theta_1\theta_2P_0 + \theta_0^4\theta_2P_1 + \theta_0^4\theta_1P_2 - P_0(U_0^2 + \theta_0U_0 + \theta_0^2)\theta_0\theta_1\theta_2}{U_0^3\theta_0^3} + o_p\Big(1/\sqrt{n}\Big) \\
= & \frac{\theta_0^3\theta_1\theta_2P_0 + \theta_0^4\theta_2P_1 + \theta_0^4\theta_1P_2 - P_0(\theta_0^2 + \theta_0\theta_0 + \theta_0^2)\theta_0\theta_1\theta_2}{\theta_0^3\theta_0^3} + o_p\Big(1/\sqrt{n}\Big) \\
= & \frac{\theta_1\theta_2P_0 + \theta_0\theta_2P_1 + \theta_0\theta_1P_2 - 3\theta_1\theta_2P_0}{\theta_0^3} + o_p\Big(1/\sqrt{n}\Big) \\
= & \frac{\theta_0\theta_2P_1 + \theta_0\theta_1P_2 - 2\theta_1\theta_2P_0}{\theta_0^3} + o_p\Big(1/\sqrt{n}\Big)
\end{aligned}$$

However,

$$\begin{aligned}
& \theta_0\theta_2P_1 + \theta_0\theta_1P_2 - 2\theta_1\theta_2P_0 \\
= & \theta_0\theta_2\frac{2}{n}\sum_i\Big(\phi(X_i,0,A)(A^{-1} - A^{-1}X_iX_i^TA^{-1}) - \theta_1\Big) \\
& +\theta_0\theta_1\frac{2}{n}\sum_i\Big(\phi(X_i,0,A)(\frac{H^2B}{2} + \frac{BX_iX_i^TB}{4}) - \theta_2\Big) \\
& -2\theta_1\theta_2\frac{2}{n}\sum_i\Big(\phi(X_i,0,A) - \theta_0\Big) \\
= & 0
\end{aligned}$$

as needed to finish the proof. □

Proof 8 (**Proof of Proposition 19**). *First, because* $Y = AX$,

$$V_g = AV_fA^T,$$

$$g(y) = f(A^{-1}y)|A^{-1}|,$$

$$\nabla_y g(y) = A^{-T}\nabla_x f(A^{-1}y)|A^{-1}|.$$

Thus, we can conclude that

$$\begin{aligned}\int g^2(y)dy &= \int f^2(A^{-1}y)|A^{-2}|dy \\ &= \int f^2(x)|A^{-1}|dx \\ &= |A^{-1}|\int f^2(x)dx,\end{aligned}$$

$$\begin{aligned}&\int \nabla_y g(y)(\nabla_y g(y))^T dy \\ =\ &\int A^{-T}\nabla_x f(A^{-1}y)|A^{-1}|(\nabla_x f(A^{-1}y))^T A^{-1}|A^{-1}|dy \\ =\ &|A^{-1}|A^{-T}\int \nabla_x f(x)(\nabla_x f(x))^T dx A^{-1},\end{aligned}$$

$$\begin{aligned}&\int yy^T g^2(y)dy \\ =\ &\int Axx^T A^T f^2(x)|A^{-2}||A|dx \\ =\ &|A^{-1}|A\int xx^T f^2(x)dx A^T,\end{aligned}$$

and

$$\begin{aligned}&\int yg^2(y)dy \\ =\ &\int Axf^2(x)|A^{-2}||A|dx \\ =\ &|A^{-1}|A\int xf^2(x)dx.\end{aligned}$$

Combining above equations, we have

$$\begin{aligned}V_{g_2} &= \frac{\int yy^T g^2(y)dy}{\int g^2(y)dy} - \frac{\int yg^2(y)dy}{\int g^2(y)dy}\Big(\frac{\int yg^2(y)dy}{\int g^2(y)dy}\Big)^T \\ &= \frac{|A^{-1}|A\int xx^T f^2(x)dxA^T}{|A^{-1}|\int f^2(x)dx} - \frac{|A^{-1}|A\int xf^2(x)dx}{|A^{-1}|\int f^2(x)dx}\Big(\frac{|A^{-1}|A\int xf^2(x)dx}{|A^{-1}|\int f^2(x)dx}\Big)^T \\ &= A\frac{\int xx^T f^2(x)dx}{\int f^2(x)dx}A^T - A\frac{\int xf^2(x)dx}{\int f^2(x)dx}\Big(\frac{\int xf^2(x)dx}{\int f^2(x)dx}\Big)A^T \\ &= AV_{f_2}A^T,\end{aligned}$$

and

$$\begin{aligned}
& \frac{\int \nabla_y g(y)(\nabla_y g(y))^T dy}{\int g^2(y)dy} \\
= & \frac{|A^{-1}|A^{-T}\int \nabla_x f(x)(\nabla_x f(x))^T dx A^{-1}}{|A^{-1}|\int f^2(x)dx} \\
= & A^{-T}\frac{\int \nabla_x f(x)(\nabla_x f(x))^T dx}{\int f^2(x)dx}A^{-1}.
\end{aligned}$$

Finally, we have

$$\begin{aligned}
J_{g_2} &= V_{g_2}^{1/2}\frac{\int \nabla_y g(y)(\nabla_y g(y))^T dy}{\int g^2(y)dy}V_{g_2}^{1/2} \\
&= V_{g_2}^{1/2}A^{-T}\frac{\int \nabla_x f(x)(\nabla_x f(x))^T dx}{\int f^2(x)dx}A^{-1}V_{g_2}^{1/2} \\
&= V_{g_2}^{1/2}A^{-T}V_{f_2}^{1/2}J_{f_2}V_{f_2}^{1/2}A^{-1}V_{g_2}^{1/2}.
\end{aligned}$$

□

Bibliography

[1] Bowman, A.W. and Foster, P.J. (1993), "Adaptive Smoothing and Density-Based Teste of Multivarilate Normality," Journal of American Statistical Association, Vol. 88, No. 422, pp. 529-539.

[2] Day, N.E. (1969), "Estimating the Components of a Mixture of Normal Distributions," Biometrika, Vol. 56, No. 3. (Dec., 1969), pp. 463-474.

[3] Friedman, J.H. and Tukey, J.W. (1974), "A Projection Pursuit Algorithm for Exploatory Data Analysis," IEEE Transactions on Computers, Vol. C-23, pp.881-889.

[4] Friedman, J.H. (1987), "Exploratory Projection Pursuit," Journal of the American Statistical Association, Vol. 82,. No.397. pp.249-266.

[5] Jee, R.J (1985), "A Study of Projection Pursuit Methods," PhD theisis, Rice University.

[6] Kagan, A. (2001), "Aother look at Cramer-Rao Inequality," The American Statistician Vol. 55, No. 3, pp. 211-212(2).

[7] Lindsay, B.G. Markatou, M., Ray, S, Yang, K. and Chen, S. (2006), "Quadratic Distance on Probability: A unified Foundation," submitted to Annals of Statistics.

[8] Nel, D.G. and Van Der Merwe, C.A. (1986), "A Solution to the Multivariate Behrens-Fisher Problem," Communications in Statistics-Theory and Methods, Vol.15(12), pp.3719-3735.

[9] Persic, I. and Posse, C. (2005) "Projection Pursuit Indices Baed on the Empirical Distribution Function," Journal of Computational and Graphical Statistics, Vol. 14, Number 3, pp.700-715.

[10] Posse, C. (1995), "Projection Pursuit Exploratory Data Analysis," Compuatational Statistics and Data Analysis, Vol.20, pp.669-687.

[11] Richard, J. B. and Wojtek, J.K. (1999) "A Charaterization of Princila Component for Projection Pursuit," The American Statistican, Vol. 53, No. 2, pp.108-109.

[12] Sammon, Jr., J.W., (1969), "A Nonlineat Mapping for Data Structure Analysis," IEEE Transactions on Computers, Vol. C-18, pp.401-409.

[13] Silverman, B.W. (1986), "Density Estimation for Statistics and Data Analysis," Chapman and Hall.

[14] Van, A.W. (1998), "Asymptotic Statistics," Cambridge University Press.

[15] Wand, M.P. and Jones, M.C. (1995), "Kernel Smoothing," Chapman and Hall.

[16] Withers, C.S. (1974), "Mercer's Theorem and Fredholm Revolvents," Bull.Austral.Math.Soc.11, pp.373–380.

[17] Zahn, C.T. (1971), "Graph–theoretical methods for detecting and describing Gestalt clusters," IEEE Transactions on Computers, Vol.C-20,pp.68-86.

Vita

Guodong Hui

Guodong Hui was born in Wuxi, China 1977. He received his B.S. degree in Probability and Statistics from Peking University in 2000, and received his M.S. degree in Statistics from Fudan University in 2003. He enrolled in the Ph.D. program in Statistics at The Pennsylvania State University in 2003. His research interest involves quadratic distance, mixture model and goodness-of-fit test.