Robustness and multivariate analysis

Fatimah Salem Alashwali

Submitted in accordance with the requirements for the degree of Doctor of Philosophy

The University of Leeds

Department of Statistics

August 2013

The candidate confirms that the work submitted is her own and that appropriate credit has been given where reference has been made to the work of others.

Abstract

Invariant coordinate selection (ICS) is a method for finding structures in multivariate data using the eigenvalue-eigenvector decomposition of two different scatter matrices. The performance of the ICS depends on the structure of the data and the choice of the scatter matrices.

The main goal of this thesis is to understand how ICS works in some situations, and does not in other. In particular, we look at ICS under three different structures: two-group mixtures, long-tailed distributions, and parallel line structure.

Under two-group mixtures, we explore ICS based on the fourth-order moment matrix, $\hat{K}$, and the covariance matrix S. We find the explicit form of $\hat{K}$, and the ICS criterion under this model. We also explore the projection pursuit (PP) method, a variant of ICS, based on the univariate kurtosis. A comparison is made between PP, based on kurtosis, and ICS, based on $\hat{K}$ and S, through a simulation study. The results show that PP is more accurate than ICS. The asymptotic distributions of the ICS and PP estimates of the groups separation direction are derived.

We explore ICS and PP based on two robust measures of spread, under two-group mixtures. The use of common location measures, and pairwise differencing of the data in robust ICS and PP are investigated using simulations. The simulation results suggest that using a common location measure can be sometimes useful.

The second structure considered in this thesis, the long-tailed distribution, is modelled by two dimensional errors-in-variables model, where the signal can have a non-normal distribution. ICS based on $\hat{K}$ and S is explored. We gain insight into how ICS finds the signal direction in the errors in variables problem. We also compare the accuracy of the ICS estimate of the signal direction and Geary's fourth-order cumulant-based estimates through simulations. The results suggest

that some of the cumulant-based estimates are more accurate than ICS, but ICS has the advantage of affine equivariance.

The third structure considered is the parallel lines structure. We explore ICS based on the W-estimate based on the pairwise differencing of the data, $\hat{V}$, and S. We give a detailed analysis of the effect of the separation between points, overall and conditional on the horizontal separation, on the power of ICS based on $\hat{V}$ and S.

Acknowledgments

All thanks and praise to Allah for helping me through this journey.

I would like to express my thanks and gratitude to my supervisor Professor John T. Kent for his help, support, and patience during my PhD years. It has been a great pleasure to work with him.

I am profoundly grateful to my husband Abdulaziz for his continuous encouragement and support. My thanks are also to my children Omar and Albaraa who bring joy to my life.

Lastly, I would like to express my heartfelt thanks and appreciations to my parents, sisters, brothers, nieces and nephews for their support and prayers.

Contents

List of Figures

List of Tables

Chapter 1

Introduction

The subject of this thesis is the invariant coordinate selection (ICS) method, developed to discover structures in multivariate data using the eigenvalue-eigenvector decomposition of two different scatter matrices. The performance of the ICS method depends on the structure of the data and the pair of scatter matrices used.

The main goal of this thesis is to understand why ICS works in some situations, and not others. Another goal is to compare ICS to the projection pursuit method (PP).

The general tools used in this thesis are multivariate analysis and robust statistics. We start by giving a brief introduction of these two topics.

1.1 Multivariate analysis

Multivariate analysis is concerned with the analysis of data of dimension p higher than one. References in multivariate analysis include Mardia et al. (1980), and Everitt (2005).

A p-dimensional dataset, with n observations can be represented by an $n \times p$ data matrix, X, say. Each row is denoted by x_i^T, where $i = 1, \ldots, n$. The data

matrix X can be written in terms of its rows as

$$X = \begin{pmatrix} x_1^T \\ \vdots \\ x_n^T \end{pmatrix},$$

where

$$x_i^T = (x_{i1}, \ldots, x_{ip}).$$

1.1.1 Transformations

As a preliminary step of the analysis, multivariate data can be transformed using one of the following transformations:

(1) Non-singular transformation: for a non-singular $p \times p$ matrix, Q, say, and a vector $b \in R^p$, suppose that X is transformed as follows

$$X \to XQ^T + 1_n b^T, \tag{1.1}$$

where 1_n is a vector of length n with all its components equal to one.

Standardization is an example of affine transformation. In standardization, each row of X is shifted to have zero mean, and scaled to have unit variance,

$$X \to (X - 1_n \bar{x}) S^{-1/2},$$

where $\bar{x}$ is the mean vector, defined in (1.5), and $S^{-1/2}$ is the inverse square root of the sample covariance matrix S, defined in (1.8), and (1.6), respectively. Standardization removes the correlation effect between variables, and scale the variance of each variable to 1. In this case the covariance matrix of the standardized data is equal to the identity matrix.

(2) Diagonal scaling: for a $p \times p$ diagonal matrix A =diag$(a_1, \ldots, a_p)$, where

$a_j \neq 0$, $j = 1, \ldots, p$. Suppose that X is transformed as follows:

$$X \to XA. \tag{1.2}$$

For example, if the components of X are measured in different scales, we can unify the measurement scales by choosing A as $\text{diag}(1/s_{jj})$, where s_{jj} is the standard deviation of the jth component of X, $j = 1, \ldots, p$.

(3) Orthogonal rotation: Let R be a $p \times p$ rotation matrix, such that $R^T R = I_p$, and $|R| = 1$. A rotation of X is defined as

$$X \to XR^T. \tag{1.3}$$

The two dimensional rotation matrix is equal to

$$R = \begin{pmatrix} \cos(\omega) & -\sin(\omega) \\ \sin(\omega) & \cos(\omega) \end{pmatrix}. \tag{1.4}$$

It is often important that multivariate techniques are affine equivariant. A multivariate technique is said to be affine equivariant if under any of transformations (1), (2), and (3), the results of the analysis are not affected. As we have mentioned earlier, if the scales of the measurements are unified, or the data is standardized, it is desirable that the performance of method is not affected by the transformation.

Methods that are equivariant under non-singular transformations in (1) are preferable. Most methods are equivariant under at least one of the transformations (1), (2) or (3).

For example, Mahalanobis distances, and linear discriminant analysis are equivariant methods under non-singular transformations. Factor analysis is equivariant under scale change transformation in (2). PCA is equivariant under orthogonal transformations.

1.1.2 Summary statistics

The sample mean vector is defined as

$$\bar{x} = \frac{1}{n} X^T 1_n. \tag{1.5}$$

The sample covariance matrix is defined as

$$S = \frac{1}{n}(X - 1_n \bar{x}^T)^T (X - 1_n \bar{x}^T). \tag{1.6}$$

The spectral decomposition of S is given by

$$S = ULU^T, \tag{1.7}$$

where L=diag$(l_1, \ldots, l_p)$ is a diagonal matrix containing the ordered eigenvalues, and $U = (u_1, \ldots, u_p)$ is a $p \times p$ matrix whose its columns are the corresponding eigenvectors.

The inverse square root of S can be defied as follows

$$S^{-1/2} = UL^{-1/2}U^T, \tag{1.8}$$

where $L^{-1/2}$ =diag$(l_1^{-1/2}, \ldots, l_p^{-1/2})$.

The sample mean and the sample covariance matrix are affine equivariant under linear transformations. Consider the non-singular transformation in (1.1). The sample mean and covariance matrix are given by

$$\bar{x} \to Q\bar{x} + b,$$
$$S \to QSQ^T.$$

1.1.3 Exploratory data analysis

Usually, the first step of the analysis is to explore the multivariate data. Exploratory analysis helps to

- choose appropriate models by detecting departure from normality, including groups, and outliers;
- reduce the dimension of the data by assessing the linear relationships between variables.

In the following we define some of the classical methods that can be used to explore multivariate data.

Principal component analysis

Principal component analysis finds the linear combination for which the data have maximal variance. The aim is to reduce the dimension, and understand the covariance structure of the data.

Consider the linear transformation Xa, where $a \in R^p$ is a unit vector. The mean of Xa is equal to $a^T\bar{x}$, and the variance is equal to a^TSa.

PCA reduces the dimension by projecting the data onto the subspace spanned by the k eigenvectors corresponding to the $k < p$ largest eigenvalues,

$$Y = (X - 1_n\bar{x}^T)V,$$

where V is a $p \times k$ matrix, its columns are the k largest eigenvectors. The dimension of the reduced dataset Y is $n \times k$.

Factor analysis

In Factor analysis it is assumed that each measurement depends on unobservable common factors. The goal is to find the linear relationship between the com-

mon factors and the measurements. This relationship can be used to reduce the dimension of the data.

In a factor model, a p-variate random vector is written as the sum of a linear combination of $k < p$ dimension vector of common factors plus unique factors. Let x be a p-variate random vector, the factor model is given by

$$x = \Lambda f + u,$$

where Λ is a $p \times k$ $(k < p)$, matrix of constants, f is a $k \times 1$ vector of common factors, and u is a $p \times 1$ vector of unique factors.

The model assumptions are

$$\begin{aligned} &\mathrm{E}(x) = \mathrm{E}(f) = 0, \mathrm{var}(f) = I_k, \\ &\mathrm{E}(u) = 0, \mathrm{var}(u) = \Psi = \mathrm{diag}(\psi_1, \ldots, \psi_p), \mathrm{cov}(f, u) = 0. \end{aligned}$$

Thus, the covariance matrix can written as

$$\Sigma = \Lambda\Lambda^T + \Psi,$$

where

$$\sigma_{ii} = \sum_{j=1}^{k} \lambda_{ij}^2 + \psi_i. \tag{1.9}$$

This means that the variance of each variable can be divided into two parts: the communalities,

$$h_i^2 = \sum_{j=1}^{k} \lambda_{ij}^2,$$

and the unique variance ψ_i.

Given multivariate data, there are two methods to estimate Λ and Ψ: the first is the principal factor analysis and the second is the maximum likelihood method, where the common factors and unique factors are assumed to be normally

distributed. We explain briefly the first method in the following paragraph.

Since the factor model is equivariant under the change of scale, as in (1.2), the variables can be scaled to have unit variances. Hence, the covariance matrix of the scaled data is equal to the correlation matrix $\hat{R}$.

The method is based on finding the reduced correlation matrix as follows

$$\hat{R} - \hat{\Psi} = \hat{\Lambda}\hat{\Lambda}^T.$$

We need first to estimate the cummunalites h_i^2, which can be estimated iteratively using the previous equation and the following equation, from (1.9),

$$\hat{\psi}_i = 1 - h_i^2.$$

The spectral decomposition of $\hat{R} - \hat{\Psi}$ is given by

$$\hat{R} - \hat{\Psi} = \sum_{i=1}^{p} a_i \gamma_i \gamma_i^T,$$

where $a_1 \geq \ldots \geq a_p$ are the eigenvalues of $\hat{R} - \hat{\psi}$, and $\hat{\gamma}_1, \ldots, \hat{\gamma}_p$ are the corresponding eigenvectors.

The k eigenvectors corresponding to the largest k eigenvalues can be used as an estimate of Λ.

1.2 Location vectors and Scatter matrices

A vector-valued function $T(X) \in R^p$ is a location vector if it is equivariant under affine transformations, which is defined in (1.1), as follows

$$T(X) \to QT(X) + b.$$

A matrix-valued function $S(X)$ is a scatter matrix if it is positive definite $p \times p$ symmetric matrix, and affine equivariant

$$S(X) \to QS(X)Q^T.$$

1.3 Robust statistics

Many statistical methods rely on the normality assumption of the data. In practice, however, the normality assumption holds at best approximately. This means that the model describes the majority of the data, but a small proportion of the data do not fit into this model. This kind of atypical data are called outliers. Outliers are not always bad data, but may contain significant information.

Using classical statistical methods in the presence of outliers would give unreliable results. Many robust statistical methods have been developed to tackle this problem. Maronna et al. (2006) and Jureckova and Picek (2005) are general references in robust statistics.

Maronna et al. (2006), page xvi, gives the following definition of robust methods:

> The robust approach to statistical modeling and data analysis aims at deriving methods that produce reliable parameter estimates and associated tests and confidence intervals, not only when the data follow a given distribution exactly, but also when this happens approximately....

1.3.1 Measuring robustness

Let $x_1, \ldots, x_n$ be an independent identically distributed replicate of a univariate random variable x. Let $F(x; \varphi)$ be the distribution function of x, where $\varphi = T(x)$

is an unknown parameter. The parameter φ can be written as a functional

$$\varphi = T(F).$$

An estimator of $\hat{\varphi}_n = T(x_1, \ldots, x_n)$, can be written as

$$\hat{\varphi}_n = T(F_n),$$

where F_n is the empirical distribution function defined as follows

$$F_n(t) = \frac{1}{n}\sum_{i=1}^{n} I(x_i \leq t),$$

where $I(A)$ is the indicator random variable which equals 1 when A holds.

The ϵ-contamination distribution is defined as follows,

$$F_\epsilon = (1-\epsilon)F + \epsilon\delta_{x_\circ}, \tag{1.10}$$

where $0 \leq \epsilon \leq 1$, $\delta_{x_\circ}$ is the point mass distribution where $P_r(x = x_\circ) = 1$.

Sensitivity curve

The effect of on the estimator $\hat{\varphi}_n = T(x_1, \ldots, x_n)$ adding a new observation $x_\circ$ can be measured by the difference

$$\begin{aligned} SC(x_\circ; T, F_n) &= T(x_1, \ldots, x_n, x_\circ) - T(x_1, \ldots, x_n) \\ &= T(F_{n+1}) - T(F_n). \end{aligned}$$

The empirical distribution function of the contaminated sample F_{n+1} is given by

$$F_{n+1}(t) = \frac{1}{n+1}(\sum_{i=1}^{n} I(x_i \leq t) + I(x_\circ \leq t)) = \frac{n}{n+1}F_n + \frac{1}{n+1}\delta_{x_\circ}.$$

Hence the sensitivity curve is defined by

$$SC(x_\circ; T, F_n) = T[\frac{n}{n+1}F_n + \frac{1}{n+1}\delta_{x_\circ}] - T(F_n). \tag{1.11}$$

For example, the sample mean $\hat{\varphi}_n = T(F_n)$ is defined as

$$\hat{\varphi}_n = T(F_n) = \int x dF_n = \frac{\sum_{i=1}^{n} x_i}{n} = \bar{x}.$$

Define $\hat{\varphi}_{n+1} = T(F_{n+1})$ as follows

$$\hat{\varphi}_{n+1} = T(F_{n+1}) = \frac{n}{n+1}\bar{x} + \frac{1}{n+1}x_\circ.$$

The sensitivity curve is given by

$$\begin{aligned} SC(x_\circ; T, F_n) &= \left\{(\frac{n}{n+1})\bar{x} + (\frac{1}{n+1})x_\circ - \bar{x}\right\} \\ &= \frac{1}{n+1}(x_\circ - \bar{x}). \end{aligned}$$

From the sensitivity curve of $\bar{x}$, as the value of the outlier $x_\circ$ increases, the sensitivity curve becomes unbounded.

Influence function

The influence function for an estimator $\hat{\varphi}_n = T(F_n)$ is the population version of its sensitivity curve. To derive the influence function, consider the contaminated distribution in (1.10). The influence function is defined as

$$\mathrm{IF}(x_0; T, F) = \lim_{\epsilon \to 0} \frac{T\left((1-\epsilon)F + \epsilon\delta_{x_0}\right) - T(F)}{\epsilon} = \frac{\partial}{\partial \epsilon}\, T\left((1-\epsilon)F + \epsilon\delta_{x_0}\right)\,|_{\epsilon=0}\,. \tag{1.12}$$

For example, the expected value as the parameter of interest: $\varphi = T(F) =$

$E(X)$. Then,

$$\begin{aligned} T(F_\epsilon) &= T\left((1-\epsilon)F + \epsilon\delta_{x_0}\right) = (1-\epsilon)T(F) + \epsilon\ T(\delta_{x_0}) \\ &= (1-\epsilon)E(X) + \epsilon\ x_0. \end{aligned}$$

The influence function of the expected value is

$$\begin{aligned} \mathrm{IF}(x_0; T, F) &= \lim_{\epsilon \to 0} \frac{(1-\epsilon)E(X) + \epsilon\ x_0 - E(X)}{\epsilon} \\ &= x_0 - E(X). \end{aligned}$$

Breakdown point

The breakdown point of an estimator $\hat{\varphi} = T(x_1, \ldots, x_n)$ is the largest possible fraction of contamination such that the estimator remains bounded.

An estimator with a high breakdown point means it is a robust estimator. For example the breakdown point of the mean is 0, since changing a single observation may increase the mean without bound, while the for the median is equal to 1/2.

1.3.2 Robust location and scale estimates

If the data are normally distributed, then the sample mean and the sample variance are the optimal location and scale estimates. Since we only know F approximately, we want the estimates of location and scale to be reliable in the presence of outliers.

In the following we discuss briefly some of robust location and scale estimates. We consider first univariate estimates. After that, we discuss the multivariate analogues of some of the univariate estimates discussed in this section.

Univariate estimates

The simplest alternative to the sample mean is the median (Med); alternatives to the sample variance include the interquartile range (IQR) and mean absolute

deviation (MAD).

In the following, we define the univariate robust estimates that are used in this thesis.

- M-estimate: Suppose that $\varphi = (\mu, \sigma^2)$. The M-estimates of location and scale are the solutions of the following estimation equations

$$\hat{\mu} = \frac{\sum_{i=1}^{n} w_1(r_i)x_i}{\sum_{i=1}^{n} w_1(r_i)},\ \hat{\sigma}^2 = \frac{\sum_{i=1}^{n} w_2(r_i)}{n}, \tag{1.13}$$

  where w_1 and w_2 are non-negative weight functions, and

$$r_i = \frac{x_i - \hat{\mu}}{\hat{\sigma}}.$$

  For example, the maximum likelihood estimate of t_ν-distribution have the following weight functions

$$w_1(r_i) = w_2(r_1) = (\nu + 1)/(\nu + r^2).$$

  Equations (1.13) are solved iteratively, i.e we start by assigning initial values to $\hat{\mu}$ and $\hat{\sigma}^2$ to compute the weights. Then we can solve equations (1.13) iteratively, update the weights in each iteration, until convergence.

- The lshorth: The lshorth is defined as the length of the shortest interval that contains half of observations. Its associated location measure is the midpoint of the shortest interval (shorth).

  The shorth was first introduced by Andrews et al. (1972) as the mean of data points in the shortest interval which contains half of observations. Andrews et al. (1972) showed that the mean of the shorth has bad asymptotic performance, its limiting distribution is not normal, and its rate of convergence is of order n^{-3}. Grubel (1988) suggested the use of the shorth as a

scale measure, lshorth. The lshorth is defined as follows

$$l_n(x_1, \ldots, x_n) = \min\{x_{(i+j)} - x_{(i)} : 1 \leq i \leq i+j \leq n, (i+j)/n \geq 1/2\}. \tag{1.14}$$

Multivariate estimates

- M-estimate: The multivariate M-estimate of location and scale are the solution of the following two equations:

$$\begin{aligned} \hat{\mu} &= \frac{\sum_{i=1}^n w_1(r_i) x_i}{\sum_{i=1}^n w_1(r_i)}, \\ \hat{\Sigma} &= \frac{\sum_{i=1}^n w_2(r_i)(x_i - \hat{\mu})(x_i - \hat{\mu})^T}{n}, \end{aligned} \tag{1.15}$$

  where

$$r_i = (x_i - \hat{\mu})^T \hat{\Sigma} (x_i - \hat{\mu}).$$

- The minimum volume ellipsoid (MVE): The minimum volume ellipsoid, Van Aelst and Rousseeuw (2009), is defined as the smallest ellipsoid containing at least half of observations. The MVE is the multivariate version of the lshorth. Like the lshorth, MVE has a high breakdown point, but bad asymptotic behavior with $n^{-1/3}$ rate of convergence. In practice, MVE is computationally expensive.

1.4 Thesis outline

When investigating ICS and PP, we need to specify:

- pair of spread measures;
- structure of the data;
- The computation of the measures of spread: based on the associated location measures, based on a common location measure, or based on the

pairwise differencing of the data to force the symmetry of the data around the origin.

In Chapter 2, we explore the rationale of the ICS method. We also link PP to ICS.

In Chapter 3, we explore the ICS criterion based on the fourth-order moment matrix and the covariance matrix and the PP based on kurtosis, under two-group mixtures of bivariate normal distributions. We also compare the accuracy of the two methods through a simulation study.

Under equal mixtures of two bivariate normal distributions, we derive the asymptotic distributions of the ICS and PP estimates of the group separation direction in Chapter 4.

ICS and PP kurtosis criteria are not robust, in the sense that they are highly affected by outliers. Both ICS and PP criteria can be defined based on two robust measures of scale. In Chapter 5, we investigate the feasibility of using robust measures of spread in ICS and PP. We also explore the effect of using a common location measure in the measures of spread used in ICS and PP criteria, and the role of pairwise differencing of data.

The errors in variables model, EIV, is a regression model with both measurements are subject to errors. In Chapter 6 we explore using ICS based on the fourth-order moment matrix and the variance in fitting EIV line. We also compare the ICS method to Geary's fourth-order cumulant-based estimators.

The performance of ICS depends on the choice of the pair of scatter matrices, and the structure of the data at hand. For example, ICS based on the fourth-order moment matrix is not able to find the structure direction in the RANDU data set. The points in the RANDU data set are arranged on 15 parallel planes, lying in three dimensional space. The structure direction in the RANDU data set is the direction that views the parallel line structure. In Chapter 7, we explore the choice of the two scatter matrices in ICS that can find such the parallel line structure in such data. Namely, the W-estimate and the covariance matrix. We

also explore the effect of using the pairwise differencing of the data in W-estimate.

Two potential applications of ICS are discussed in Chapter 8. We discuss the role ICS can play in finding principal curves and in the analysis of fingerprint images.

Chapter 2

Invariant coordinate selection and related methods

2.1 Introduction

Suppose we have a multivariate data set that has a lower dimensional structure. One way to detect structure is by projecting the data onto a line for which the data is maximally non-normal. Hence, methods that are sensitive to non-normality can be used to detect structure. Two such methods from the literature: invariant coordinate selection (ICS), introduced by Tyler et al. (2009), and projection pursuit(PP), introduced by Friedman and Tukey (1974).

ICS and PP find structure direction by optimizing criteria sensitive to non-normality. Any summary statistic that is sensitive to non-normality can be used as a criterion. For example, the univariate kurtosis is zero when a random variable has normal distribution. For non-normal distributions the kurtosis is mostly non-zero, positive or negative.

Motivated by kurtosis, the ICS and PP optimality criteria can be defined as ratios of any two measures of spread.

The structure of this chapter is as follows. In Section 2.2, we explore the rationale of the ICS method. After that we link the ideas of ICS to PP in Section

2.3. In Section 2.4, the use of kurtosis as a criterion to identify non-normal directions of the data is reviewed. In Section 2.5, the idea of using a common location measure in the calculations of the pair of spread measures in ICS and PP is introduced. In Section 2.6, the role of differencing on ICS and PP is discussed. In Section 2.7, we go through non-normal structures, considered in this thesis, and their corresponding models. In Section 2.8, we discuss how to choose the pair of spread measures for ICS and PP. Notations are introduced in Section 2.9.

2.2 Rationale of the invariant coordinate selection method

Let X be an $n \times p$ data matrix, where its rows are $x_i^T = (x_{i1}, \ldots, x_{ip})$, $i = 1, \ldots, n$. ICS finds a direction $a \in R^p$ for which Xa is maximally non-normal, using the relative eigenvalue-eigenvector decomposition of two affine equivariant scatter matrices with different level of robustness.

Let $S_1 = S_1(X)$ and $S_2 = S_2(X)$, be two affine equivariant scatter matrices. Each scatter matrix is associated with a location measure, $\hat{\mu}_1 = \hat{\mu}_1(X)$ and $\hat{\mu}_2 = \hat{\mu}_2(X)$, say.

The ICS criterion, based on S_1 and S_2, is to find a direction a that minimizes/maximizes the following criteria

$$\hat{\kappa}_{\text{ICS}}(a) = \frac{a^T S_1 a}{a^T S_2 a}. \tag{2.1}$$

The minimum/maximum value of (2.1) is the smallest/largest eigenvalue of $S_2^{-1}S_1$, obtained when a is the corresponding eigenvector.

An eigenvalue λ and eigenvector a of $S_2^{-1}S_1$ are the solution of the following equation

$$S_2^{-1}S_1 a = \lambda a, \tag{2.2}$$

Suppose that X is transformed by the affine transformation in (1.1). Denote the transformed data matrix by X^*. Then, the scatter matrices $S_1^* = S_1(X^*)$, and $S_2^* = S_2(X^*)$ become

$$S_1^* = QS_1Q^T$$
$$S_2^* = QS_2Q^T,$$

since $S_1(\cdot)$ and $S_2(\cdot)$ are affine equivariant.

Let

$$b = Q^{-T}a. \tag{2.3}$$

The eigenvalues and eigenvectors of ${S_2^*}^{-1}S_1^*$ are the solution of the following equation

$$\begin{aligned} {S_2^*}^{-1}S_1^*b &= Q^{-T}S_2^{-1}Q^{-1}QS_1Q^TQ^{-T}a \\ &= Q^{-T}S_2^{-1}S_1a. \end{aligned}$$

From (2.2),

$$\begin{aligned} {S_2^*}^{-1}S_1^*b &= Q^{-T}\lambda a \\ &= \lambda b. \end{aligned} \tag{2.4}$$

From (2.2) and (2.4), the eigenvalues of ${S_2^*}^{-1}S_1^*$ and $S_2^{-1}S_1$ are the same, whereas the eigenvectors of ${S_2^*}^{-1}S_1^*$ are the eigenvectors of $S_2^{-1}S_1$ scaled by Q^{-T}.

The PCA can be related to (2.2), with S_2 taken as the identity matrix. Since ICS is affine equivariant, as we have shown earlier in (2.2) to (2.4), we may standardize X with respect to $Q = S_2^{-1/2}$, such that $S_2^* = I_p$. In this case, ICS can be seen as applying the PCA of the standardized data X^* with respect to S_1^*.

2.3 Projection pursuit as a variant of invariant coordinate selections

As ICS, PP finds a direction a, such that the projection Xa is maximally non-normal. The PP criterion can be defined as a ratio of two univariate affine equivariant measures of scale, s_1 and s_2, say:

$$\kappa_{\mathrm{PP}}(a) = \frac{s_1(Xa)}{s_2(Xa)}. \tag{2.5}$$

In contrast to ICS, minimizing/maximizing (2.5) is computationally expensive because it is carried out numerically. PP method searches in all projection directions to find the direction that minimizing/maximizing (2.5).

Suppose that X is standardized as in (1.1). Criterion (2.5) based on the standardized data X^* becomes

$$\kappa_{\mathrm{PP}}(b) = \frac{s_1(X^*b)}{s_2(X^*b)}. \tag{2.6}$$

Where b is as in (2.3), and can be noted by comparing the following two linear transformations,

$$\begin{aligned} Xa &\propto X^*b \\ &= XQ^TQ^{-T}a. \end{aligned} \tag{2.7}$$

We explore the effect of standardization on PP in Section 3.6.

2.4 Univariate kurtosis

The kurtosis of a univariate random variable u, say, is defined as follows.

$$\mathrm{kurt}(u) = \frac{\mathrm{E}\{(u-\mu_u)^4\}}{\left[\mathrm{E}\{(u-\mu_u)^2\}\right]^2} - 3. \tag{2.8}$$

where μ_u, is the mean value of u.

The kurtosis takes the following possible values:

(1) $\text{kurt}(u) = 0$: satisfied under normality.

(2) $\text{kurt}(u) < 0$: this case is called sub-Gaussian.

(3) $\text{kurt}(u) > 0$: this case is called super-Gaussian.

The Sub-Gaussian case appears in distributions flatter than the normal and have thinner tails; examples include the uniform distribution. On the other hand, the super-Gaussian case appears in distributions that are more peaked than the normal distribution and have longer tails; examples include t, and Laplace distributions.

The extreme case of sub-Gaussianity occurs for two-point distribution. That is, let s be a random variable that has a two-point distribution, defined as follows.

$$s = \begin{cases} 1 & \text{with probability } q \\ -1 & \text{with probability } (1-q) \end{cases}. \tag{2.9}$$

The kurtosis of s is equal to

$$\begin{aligned} \text{kurt}(s) &= \frac{-3[4q(1-q)]^2 + 16q(1-q)}{(4q(1-q))^2} - 3 \\ &= \frac{1}{q(1-q)} - 6. \end{aligned} \tag{2.10}$$

The minimum value of $\text{kurt}(s)$ is -2, attained when $q = 1/2$. The value of $\text{kurt}(s)$ increases as q increases away from $1/2$, as shown in the following lemma.

Lemma 2.4.1. *The kurtosis of any random variable takes the values between* -2 *and* ∞.

Proof. This lemma can be proved using the Cauchy-Schwarz inequality. The Cauchy-Schwarz inequality for a random variable u, where $\text{E}(u) = 0$, is given as

follows

$$(\mathrm{E}\{u^2\})^2 \leq \mathrm{E}\{u^4\}.$$

Dividing both sides by $(\mathrm{E}\{u^2\})^2$ and subtracting 3

$$\frac{\mathrm{E}\{u^4\}}{(\mathrm{E}\{u^2\})^2} - 3 \geq 1 - 3$$
$$\mathrm{kurt}(u) \geq -2.$$

□

The kurtosis has been used as a criterion for identifying non-normal projections of multivariate data in PP and independent component analysis (ICA); see for example, Huber (1985), Jones and Sibson (1987), Peña and Prieto (2001), and Bugrien and Kent (2005).

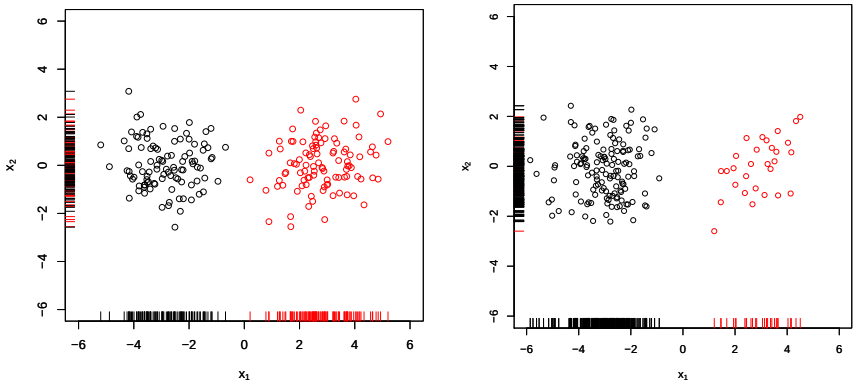
Peña and Prieto (2001) suggested that minimizing kurtosis is appropriate if the purpose is identifying clusters, whereas maximizing kurtosis can be used to detect outliers. In particular, if q is near half, minimizing kurtosis is more useful than maximizing it; if q is far from half maximizing kurtosis is more useful than minimizing it. Peña and Prieto (2001) also gave a threshold of q that distinguishes q from being far from half or near half. The threshold can be noted from (2.9). Namely, if $q(1-q) = 1/6$ the kurtosis equals to zero, otherwise the kurtosis will be negative or positive.

To illustrate the idea of minimizing and maximizing kurtosis, consider n bivariate data points generated by adding a bivariate isotropic noise to points generated from s in (2.9). That is, the data points x_i are generated from the following model

$$x_i = s_i \begin{pmatrix} \alpha \\ 0 \end{pmatrix} + \epsilon_i, \tag{2.11}$$

where $i = 1, \ldots, n$, $\alpha > 0$ is a separation parameter between groups, and ϵ_i is

a bivariate normal isotropic noise. The structure direction in this model is the direction that best separate the two groups, at the horizontal direction, and the noise direction is in the direction the data are normally distributed, the vertical direction, as shown in Figures 2.1(a) and (b), with $q = 1/2$ and $1/7$, and $\alpha = 3$.



(a) $q = 1/2$ (b) $q = 1/7$

Figure 2.1: Bivariate data points generated from (2.11) with, $q = 1/2$, $q = 1/7$, $n = 200$, $\alpha = 3$, and $a = (1, 0)^T$.

In (a), the value of the sample kurtosis in the horizontal direction is -1.7, while in the vertical is 0.2. In (b), the sample kurtosis in the horizontal direction is 1.3, while in the vertical direction is -0.8.

The structure direction is in the direction that minimizes or maximizes the kurtosis relative to the kurtosis of the data in the noise direction.

2.5 Common location measure

The computation of each measure of spread in criterion (2.1) and (2.6) is based on an associated location measure. Sometimes different location measures are used in the denominator and numerator, which makes the methods unreliable.

One possible way to solve this problem is by using a common location measure in the denominator and numerator. Another way is by computing the scale measures based on pairwise differencing of the data to force the symmetry of the data around the origin. The pairwise differencing of the data is defined in the following section. This problem is explored in Chapter 5.

2.6 Role of differencing

Let X be an $n \times p$ data matrix, with its rows $x_i^T = (x_{i1}, \ldots, x_{ip})$, $i = 1, \ldots, n$. The pairwise differencing of X, denoted by X^d, is defined as follows,

$$x_k^d = x_i - x_j, \quad i \neq j = 1, \ldots, n, \tag{2.12}$$

where $k = 1, \ldots, n(n-1)$.

As we noted earlier in Section 2.5, pairwise differencing can be used to force the symmetry of the data around the origin.

Another problem that involves differencing of the data is when the structure is similar to the RANDU example from Tyler et al. (2009). The data points in RANDU data set, explained in Section 7.2, are arranged in 15 parallel planes, evenly spaced. The structure direction in this data set is the direction that views the parallel lines structure.

The pairwise differences of RANDU-like data produce inliers. Inlers are points with small lengths, arise as a result of the difference of two points with small distance.

A sensible choice of the pair of scatter matrices for such data should accentuates inliers to emphasize the parallel line structures. Chapter 7 discusses the choice of the pair of scatter matrices in this kind of structure.

2.7 Types of non-normal structures

Examples of departure from non-normality include skewed, long-tailed, and mixture distributions. The focus in this thesis will be on the following non-normal structures:

(I) Two-group mixtures: this structure is defined in Section 3.2. Figure 2.2 (a) shows plot the two-group structure.

(II) Long-tailed: this structure can be defined by the errors-in-variables model in Section 6.2, when the signal has non-normal distribution. The long-tailed structure is shown in Figure 2.2 (b).

(III) Parallel lines structure: this structure is defined in the RANDU-type model defined in Section 7.3. The parallel line structure is shown in Figure 2.2 (c).

2.8 How to choose the pair of scatter matrices

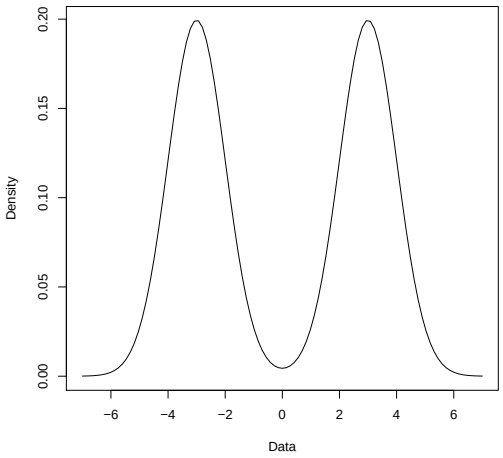
In this Section, we largely follow the classification of scatter matrices from Tyler et al. (2009), who divided the scatter matrices into three classes. We have added a new class, Class I. The classes of scatter matrices are given as follows:

- Class I is the class of highly non-robust scatter matrices, with zero breakdown point and unbounded influence function. The scatter matrices included in this class are highly affected by inliers or outliers. Examples include weighted scatter matrices that up-weight outliers, or inliers, such as the fourth-order scatter matrix $\hat{K}$, defined in (3.23), and the one-step W-estimate $\hat{V}$, defined in (7.1).

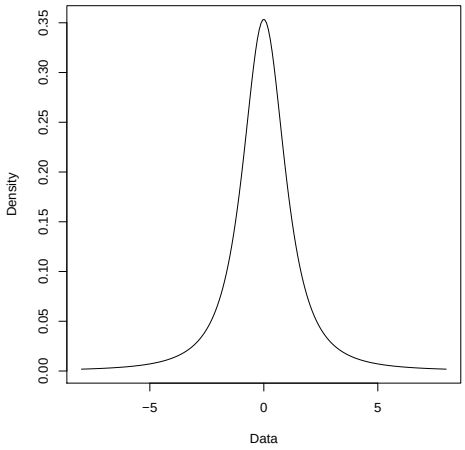
- Class II is the class of non-robust scatter matrices with zero breakdown points and unbounded influence function. Examples include the covariance matrix.

- Class III is the class of scatter matrices that are locally robust, in the sense that they have bounded influence function and positive breakdown points not greater than $\frac{1}{p+1}$. An example from this class is the class of multivariate M-estimators, e.g M-estimate for t-distribution, Arslan et al. (1995).

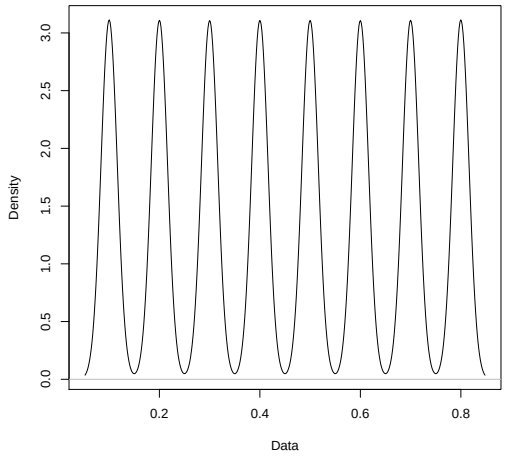
- Class IV is the class of scatter matrices with high breakdown points such as the Stahel-Donoho estimate, the minimum volume ellipsoid, Van Aelst



(a) Two-group mixture



(b) Long-tailed



(c) Parallel lines

Figure 2.2: Illustration of the three non-normal structures considered in the thesis.

and Rousseeuw (2009) and the constrained M-estimates, Kent and Tyler (1996).

Motivated by the univariate kurtosis, the ICS pair of scatter matrices can be chosen from different classes. By convention, the scatter matrix in the denominator is more robust that the one in the numerator.

In the following, we list the scatter matrices that are used in ICS:

1. The fourth-order moment matrix, defined in (3.23).
2. The covariance matrix, defined in (1.6).
3. The M-estimator for t_2 distribution, defined in Section 1.3.2.
4. The minimum volume ellipsoid, defined in Section 1.3.2.
5. The one-step wighted scatter matrix , defined in (7.1), usually computed based on pairwise differencing.

Similarly, in PP the univariate analogues of some of the scatter matrices, listed above, are used,

1. The univariate kurtosis, defined in (2.8).
2. The variance.
3. The univariate M-estimate for t_2 distribution, defined in (1.13).
4. The lshorth, defined in (1.14).

2.9 Notation

Throughout the thesis the following notation will be used:

- Univariate random variables and multivariate random vectors, and their realizations, are denoted by small letters, x, say.

- Data matrices will be denoted by capital letters, X, say.
- The notations of ICS and PP methods are as follows:
  - ICS and PP, based on two different measures of scales, computed with respect to the associated location measures, are denoted by

    $$\text{ICS} : \text{Spread 1} : \text{Spread 2},$$
    $$\text{PP} : \text{Spread 1} : \text{Spread 2}.$$

    For example, ICS and PP based on kurtosis and variance are denoted by

    $$\text{ICS} : \text{kurtosis} : \text{variance},$$
    $$\text{PP} : \text{kurtosis} : \text{variance}.$$

  - ICS and PP with respect to a common location measure will be denoted by

    $$\text{ICS(Location)} : \text{Spread 1} : \text{Spread 2},$$
    $$\text{PP(Location)} : \text{Spread 1} : \text{Spread 2}.$$

  - ICS and PP with respect to differenced data are given the superscript d, i.e

    $$\text{ICS}^d : \text{Spread 1} : \text{Spread 2},$$
    $$\text{PP}^d : \text{Spread 1} : \text{Spread 2}.$$

- The ICS criterion based on any pair of scatter matrices will be denoted by: in p-dimension $\kappa_{\text{ICS}}(a)$, as function of the unit vector $a \in R^p$; two-dimension $\kappa_{\text{ICS}}(\theta)$, as a function of the group separation direction $\theta \in [-\pi/2, \pi/2)$.

- Similarly, PP criterion is denoted by $\kappa_{\mathrm{PP}}(a)$ in p-dimension, and $\kappa_{\mathrm{PP}}(\theta)$ in two-dimension.

Chapter 3

Using ICS and PP based on fourth-order moments in two-group mixtures

3.1 Introduction

In this chapter we explore the theory and practice of ICS:kurtosis:variance and PP:kurtosis:variance, under mixtures of two bivariate normal distributions. The ICS and PP methods are defined in Sections 2.2 and 2.3. The main goal is to compare the accuracy of ICS and PP in identifying the group separation direction.

The structure of this chapter is given as follows. In section 3.2, we define the model, mixtures of two bivariate normal distributions. In Section 3.3 we discuss the theory of ICS:kurtosis:variance. In Section 3.4, the relationship between Mardia's multivariate measure of kurtosis and ICS:kurtosis:variance. In Section 3.5, the theory of PP:kurtosis:variance under the mixture model is discussed. In Section 3.6, we define ICS:kurtosis:variance, and PP:kurtosis:variance criteria in population and sample cases. We define a measure of spread between axes in Section 3.7, that is used to compare the accuracy of the two methods. In Section 3.8, a simulation study is conducted to compare ICS and PP.

3.2 The model: Mixtures of two bivariate normal distributions

3.2.1 The model assumptions

Let $z = (z_1, z_2)^T$ be a bivariate random vector, with mean μ and covariance matrix Σ_z, distributed as a mixture of two bivariate normal distributions, with mixing proportion q.

Assume for simplicity that the two groups have equal within-group covariance matrices, W_z. The density function of z is given by

$$f(z) = qg(z; \mu_1, W_z) + (1-q)g(z, \mu_2, W_z). \tag{3.1}$$

where μ_1 and μ_2 are the group mean vectors, and g is the marginal normal distribution given by

$$g(z; \mu_i, W_z) = |2\pi W_z|^{-1/2} \exp\{-\frac{1}{2}(z-\mu_i)^T W_z^{-1}(z-\mu_i)\}.$$

where $|W_z| > 0$.

Since ICS and PP are equivariant methods, under translation, rotation, and affine transformations, we may, without loss of generality, assume the following

(1) The random vector z is translated and rotated such that the group means lie on the horizontal direction, and become symmetric around the origin,

$$z^{'} = Rz,$$

where R is a rotation matrix defined in (1.4). The within-group covariance

matrix and the total covariance matrix of $z^{'}$ are

$$W_{z'} = RW_zR^T,$$
$$\Sigma_{z'} = R\Sigma_zR^T.$$

(2) The random vector $z^{'}$ is standardized in either of two ways: with respect to $W_{z'}$; or with respect to $\Sigma_{z'}$. Each standardization method gives a different coordinate system, as shown in the following.

(i) Suppose that $z^{'}$ is standardized with respect to $W_{z'}$, as follows:

$$x = W_{z'}^{-1/2}z^{'}. \tag{3.2}$$

The group means are given by

$$\mu_1 = (\alpha, 0)^T, \quad \mu_2 = (-\alpha, 0)^T, \tag{3.3}$$

where the parameter $\alpha \geq 0$ is used here as a separation parameter between the group means. If $\alpha = 0$, the distribution of x will be normal, and as α increases the groups become more separated.

The total mean vector is given by

$$\mu_x = q\mu_1 + (1-q)\mu_2 = ((2q-1)\alpha, 0)^T. \tag{3.4}$$

The total covariance matrix is given by

$$\Sigma_x = I_2 + B_x, \tag{3.5}$$

where B_x is the between-group scatter matrix,

$$B_x = q(\mu_1 - \mu_x)(\mu_1 - \mu_x)^T + (1-q)(\mu_2 - \mu_x)(\mu_2 - \mu_x)^T$$
$$= \begin{pmatrix} 4q(1-q)\alpha^2 & 0 \\ 0 & 0 \end{pmatrix}. \tag{3.6}$$

Substituting (3.6) in (3.5) gives

$$\Sigma_x = \begin{pmatrix} 1 + 4q(1-q)\alpha^2 & 0 \\ 0 & 1 \end{pmatrix}. \tag{3.7}$$

The standardized random vector $x = (x_1, x_2)^T$ can be written as follows

$$\begin{pmatrix} x_1 \\ x_2 \end{pmatrix} = s \begin{pmatrix} \alpha \\ 0 \end{pmatrix} + \begin{pmatrix} \epsilon_1 \\ \epsilon_2 \end{pmatrix}. \tag{3.8}$$

where $\epsilon = (\epsilon_1, \epsilon_2)^T$ is a normally distributed random vector, with zero mean vector and covariance matrix $W_x = I_2$, and the random variable s has a two-point distribution, defined in (2.9).

(ii) Suppose that $z^{'}$ is standardized with respect to $\Sigma_{z'}$, as follows:

$$y = \Sigma_{z'}^{-1/2} z^{'}. \tag{3.9}$$

The group means are given by

$$\mu_1 = (\delta, 0)^T, \quad \mu_2 = (-\delta, 0)^T,$$

where $0 \leq \delta \leq 1$. The total mean vector is given by

$$\mu_y = q\mu_1 + (1-q)\mu_2 = ((2q-1)\delta, 0)^T. \tag{3.10}$$

The between-group variance is given by

$$B_y = \begin{pmatrix} 4q(1-q)\delta^2 & 0 \\ 0 & 0 \end{pmatrix}. \tag{3.11}$$

The within-group variance W_y is equal to

$$\begin{aligned} W_y &= I_2 - B_y \\ &= \begin{pmatrix} 1 - 4q(1-q)\delta^2 & 0 \\ 0 & 1 \end{pmatrix}. \end{aligned} \tag{3.12}$$

The standardized random vector $y = (y_1, y_2)^T$ can be written as

$$\begin{pmatrix} y_1 \\ y_2 \end{pmatrix} = s \begin{pmatrix} \delta \\ 0 \end{pmatrix} + \begin{pmatrix} \epsilon_1^* \\ \epsilon_2^* \end{pmatrix}. \tag{3.13}$$

where $\epsilon^* = (\epsilon_1^*, \epsilon_2^*)^T$ is a normally distributed random vector, with zero mean vector and covariance matrix W_y, and the random variable s has a two-point distribution, defined in (2.9).

3.2.2 Univariate moments

In this section, we derive the univariate moments of the components of x, x_1, x_2, and the components of y, y_1, y_2, up to fourth order, to use them in the derivation of ICS:kurtosis:variance and PP:kurtosis:variance criteria in Sections 3.3 and 3.4.

Let the rth-order non-central moment of a univariate random variable u, say, be denoted by $\mu_u^{'}(r)$, and the rth order central moment be denoted by $\mu_u(r)$, defined as follows, respectively

$$\begin{aligned} \mu_u^{'}(r) &= \mathrm{E}\{u^r\}, \\ \mu_u(r) &= \mathrm{E}\{(u - \mu_u^{'}(1))^r\}. \end{aligned} \tag{3.14}$$

Let the $(h+j)$-order non-central cross moment of a bivariate random variable $u = (u_1, u_2)^T$, say, be denoted by $\mu'_{u_1u_2}(h, j)$, and the $(h + j)$ central order cross moment about the mean vector be denoted by $\mu_{u_1u_2}(h, j)$, defined as follows, respectively,

$$\begin{aligned}\mu_{u_1u_2}(h, j) &= \mathrm{E}\{u_1^h u_2^j\},\\ \mu_{u_1u_2}(h, j) &= \mathrm{E}\{(u_1 - \mu'_{u_1}(1))^h (u_2 - \mu'_{u_2}(1))^j\}.\end{aligned} \tag{3.15}$$

If u_1 and u_2 are independent, $\mu_{u_1u_2}(h, j) = \mu_{u_1}(h)\mu_{u_2}(j)$.

From (3.8), the components of x are independent; the first component is distributed as a mixture of two normal distributions,

$$x_1 \sim qN(\alpha, 1) + (1 - q)N(-\alpha, 1),$$

and the second component is distributed as $N(0, 1)$. Similarly, from (3.13), the first component of y is distributed as

$$y_1 \sim qN(\delta, 1 - 4q(1 - q)\delta^2) + (1 - q)N(-\delta, 1 - 4q(1 - q)\delta^2),$$

and the second component of y is distributed as $N(0, 1)$. We begin by finding the moments of s.

Moments of s

- The mean of s is equal to

$$\mu'_s(1) = 2q - 1.$$

- All odd non-central moments are equal to $\mu'_s(1)$.
- All even non-central moments are equal to 1.

- The central moments are given by

$$\begin{aligned}\mu_s(2) &= 4q(1-q),\\ \mu_s(3) &= -8q(1-q)(2q-1),\\ \mu_s(4) &= -48q^2(1-q)^2 + 16q(1-q).\end{aligned} \tag{3.16}$$

- The kurtosis of s is defined in (2.10).

Moments of x_1 and x_2

Since $\epsilon = (\epsilon_1, \epsilon_2)^T$ is normally distributed with mean zero, and covariance matrix $W_x = I_p$,

- all odd moments of ϵ_1 and ϵ_2 are equal to zero,
- the fourth-order moments are $\mu_{\epsilon_1}(4) = \mu_{\epsilon_2} = 3$.
- Since $x_2 = \epsilon_2$, the moments of x_2 are equal to the moments of ϵ_2.
- The kurtosis of x_2 is equal to zero.

From (3.8), since the variable x_1 can be written as

$$x_1 = \alpha s + \epsilon_1,$$

the moments of x_1 can be computed with the help of the moments of s and ϵ_1 as follows.

- The non-central moments of x_1 are given by

$$\begin{aligned}\mu_{x_1}^{'}(1) &= \mathrm{E}\{\alpha s + \epsilon_1\} = \alpha(2q-1),\\ \mu_{x_1}^{'}(2) &= \mathrm{E}\{\alpha s + \epsilon_1\}^2 = 1 + \alpha^2,\\ \mu_{x_1}^{'}(3) &= \mathrm{E}\{(\alpha s + \epsilon_1)^3\} = (2q-1)(3\alpha + \alpha^3),\\ \mu_{x_1}^{'}(4) &= \mathrm{E}\{(\alpha s + \epsilon_1)^4\} = 3 + 6\alpha^2 + \alpha^4.\end{aligned} \tag{3.17}$$

- Using (3.17), the central moments of x_1, up to fourth order are given by:

$$\mu_{x_1}(2) = \mathrm{E}\{(x_1 - \mu_{x_1}^{'}(1))^2\} = 1 + 4\alpha^2 q(1-q),$$
$$\mu_{x_1}(4) = \mathrm{E}\{(x_1 - \mu_{x_1}^{'}(1))^4\} = 3 + 24\alpha^2 q(1-q) + \alpha^4[-48q^2(1-q)^2 + 16q(1-q)]. \tag{3.18}$$

- The kurtosis of x_1 is

$$\mathrm{kurt}(x_1) = \frac{16q(1-q)\alpha^4\{1-6q(1-q)\}}{\{1+4\alpha^2 q(1-q)\}^2}. \tag{3.19}$$

Moments of y_1 and y_2

From (3.13), the variable y_1 can be written as

$$y_1 = \delta s + \epsilon_1^*,$$

the moments of y_1 can be computed using the moments of s and ϵ_1^*.

- The non-central moments of y_1 are given by

$$\mu_{y_1}^{'}(1) = \mathrm{E}\{\delta s + \epsilon_1^*\} = \delta(2q-1),$$
$$\mu_{y_1}^{'}(2) = \mathrm{E}\{(\delta s + \epsilon_1^*)^2\} = 1 + \delta^2\{1 - 4q(1-q)\},$$
$$\mu_{y_1}^{'}(3) = \mathrm{E}\{(\delta s + \epsilon_1^*)^3\} = (2q-1)\{3\delta + \delta^3(1 - 12q(1-q))\},$$
$$\mu_{y_1}^{'}(4) = \mathrm{E}\{(\delta s + \epsilon_1^*)^4\} = \delta^4 + 6\delta^2\{1 - 4q(1-q)\delta^2\} + 3\{1 - 4q(1-q)\delta^2\}^2. \tag{3.20}$$

- The central moments of y_1, up to fourth order, using (3.20), are given by:

$$\mu_{y_1}(2) = \mathrm{E}\{(y_1 - \mu_{y_1}^{'}(1))^2\} = 1,$$
$$\mu_{y_1}(4) = \mathrm{E}\{(y_1 - \mu_{y_1}^{'}(1))^4\} = 3 + 16\delta^4 q - 112\delta^4 q^2 + 192\delta^4 q^3 - 96\delta^4 q^4. \tag{3.21}$$

- The kurtosis of y_1 is

$$\text{kurt}(y_1) = 16\delta^4 q(1-q)\{1-6q(1-q)\}. \tag{3.22}$$

3.3 Invariant coordinate selection based on fourth-order moments matrix in population

Let $x = (x_1, x_2)^T$ be a bivariate random vector distributed as model (3.1). Without loss of generality, assume that x is standardized beforehand with respect to the within-group scatter matrix, as in (3.2). Th total mean vector μ_x and the total covariance matrix Σ_x are given in (3.4) and (3.7), respectively.

The fourth-order moment matrix, denoted by $K_x = K(x)$, Tyler et al. (2009), is defined as follows

$$K_x = K(X) = E\{(x-\mu_x)^T\Sigma_x^{-1}(x-\mu_x)(x-\mu_x)(x-\mu_x)^T\}. \tag{3.23}$$

The ICS:kurtosis:variance optimality criterion is to minimize/maximize the following criterion

$$\kappa_{ICS}(\theta) = \frac{a^T K_x a}{a^T \Sigma_x a}, \tag{3.24}$$

where a, a unit vector, can be written as

$$a = (\cos(\theta), \sin(\theta))^T. \tag{3.25}$$

The minimum/maximum value of (3.24) is the smallest/largest eigenvalue of $\Sigma_x^{-1}K_x$, obtained when a is the corresponding eigenvector.

To gain an insight into criterion (3.24), we find its explicit formula under the mixture model (3.1). To proceed, we first find the form of K_x, as follows.

The first factor in K_x, from (3.23) is given by

$$(x-\mu_x)^T\Sigma_x^{-1}(x-\mu_x)=\frac{(x_1-\alpha(2q-1))^2}{1+4q(1-q)\alpha^2}+x_2^2. \tag{3.26}$$

The second factor in K_x is given by

$$(x-\mu)(x-\mu)^T=\begin{pmatrix}\{x_1-\alpha(2q-1)\}^2 & \{x_1-\alpha(2q-1)\}x_2\\ \{x_1-\alpha(2q-1)\}x_2 & x_2^2\end{pmatrix}. \tag{3.27}$$

Multiplying (3.26) and (3.27), and taking the expectation, gives

$$K_x=\mathrm{E}\begin{pmatrix}\frac{(x_1-\alpha(2q-1))^4}{1+4q(1-q)\alpha^2}+(x_1-\alpha(2q-1))^2x_2^2 & \frac{(x_1-\alpha(2q-1))^3(x_2)}{1+4q(1-q)\alpha^2}+(x_1-\alpha(2q-1))x_2^3\\ \frac{(x_1-\alpha(2q-1))^3x_2}{1+4q(1-q)\alpha^2}+(x_1-\alpha(2q-1))x_2^3 & \frac{(x_1-\alpha(2q-1))^2(x_2^2)}{1+4q(1-q)\alpha^2}+x_2^4\end{pmatrix}.$$

The matrix K_x can be expressed in terms of moments, as follows

$$K_x=\begin{pmatrix}\frac{\mu_{x_1}(4)}{1+4q(1-q)\alpha^2}+\mu_{x_1}(2)\mu_{x_2}(2) & 0\\ 0 & \frac{\mu_{x_1}(2)\mu_{x_2}(2)}{1+4q(1-q)\alpha^2}+\mu_{x_2}(4)\end{pmatrix}. \tag{3.28}$$

Substituting by the moments of x_1 and x_2, from Section 3.2, in (3.28) gives

$$\begin{aligned}K_x&=\begin{pmatrix}\frac{3+24q(1-q)\alpha^2+[-48q^2(1-q)^2+16q(1-q)]\alpha^4}{1+4q(1-q)\alpha^2}+(1+4q(1-q)\alpha^2) & 0\\ 0 & \frac{(1+4\alpha^2q(1-q))}{1+4\alpha^2q(1-q)}+3\end{pmatrix}.\\ &=\begin{pmatrix}\frac{3+24q(1-q)\alpha^2+[-48q^2(1-q)^2+16q(1-q)]\alpha^4}{1+4q(1-q)\alpha^2}+1+4q(1-q)\alpha^2 & 0\\ 0 & 4\end{pmatrix}.\end{aligned} \tag{3.29}$$

Substituting (3.7) and (3.29) in (3.24) gives an explicit form of $\kappa_{\mathrm{ICS}}(\theta)$, as follows

$$\kappa_{\mathrm{ICS}}(\theta)=4+\frac{16q(1-q)(1-6q(1-q))\alpha^4\cos^2(\theta)}{\{1+4q(1-q)\alpha^2\}\{1+4q(1-q)\alpha^2\cos^2(\theta)\}}. \tag{3.30}$$

Minimizing or maximizing $\kappa_{\mathrm{ICS}}(\theta)$ in (3.30) depends on q, as discussed in Section 2.4. Namely, if q is near half, θ is in the direction that minimizes $\kappa_{\mathrm{ICS}}(\theta)$. If q is

far from half, θ is in the direction that maximizes $\kappa_{\text{ICS}}(\theta)$.

The form of $\Sigma^{-1}K$ is given as follows

$$\begin{aligned}\Sigma^{-1}K &= \begin{pmatrix} \frac{3+24q(1-q)\alpha^2+[-48q^2(1-q)^2+16q(1-q)]\alpha^4}{(1+4q(1-q)\alpha^2)^2}+1 & 0 \\ 0 & 4 \end{pmatrix} \\ &= \begin{pmatrix} 4+\frac{16q(1-q)\alpha^4(1-6q(1-q))}{[1+4\alpha^2q(1-q)]^2} & 0 \\ 0 & 4 \end{pmatrix}. \end{aligned} \tag{3.31}$$

Note that (3.31) can be written as follows

$$\Sigma^{-1}K = \begin{pmatrix} 4+\text{kurt}(x_1) & 0 \\ 0 & 4 \end{pmatrix}. \tag{3.32}$$

The eigenvalues and eigenvectors of $\Sigma^{-1}K$ are

$$\begin{aligned} &\lambda_1 = 4+\text{kurt}(x_1),\ \lambda_2 = 4, \\ &\gamma_1 = (1,0)^T,\ \gamma_2 = (0,1)^T. \end{aligned} \tag{3.33}$$

From (3.32) and (3.33), the group separation direction θ is in the direction of the eigenvector, γ_1, corresponding to the eigenvalue $4+\text{kurt}(x_1)$. The value of $\text{kurt}(x_1)$ depends on the group separation parameter α, and the mixing proportion q. If $\alpha = 0$, $\text{kurt}(x_1) = 0$. And as $\alpha \to \infty$, $\text{kurt}(x_1)$ reduces to $\text{kurt}(s)$, defined in (2.10), which in turn depends on q as explained earlier.

Now, we follow similar calculations to derive K_y, where $y = (y_1, y_2)$ is standardized with respect to $\Sigma^{-1/2}$ as in (3.9), such that $\Sigma_y = I_2$.

The definition of the fourth-order moment matrix in (3.23), $K_y = K(y)$,reduces to

$$K_y = K(y) = \text{E}\{yy^Ty^Ty\}. \tag{3.34}$$

By following similar calculations to (3.26)-(3.29), K_y takes the following form

$$K_y = \begin{pmatrix} 4 + 16\delta^4 q(1-q)\{1 - 6q(1-q)\} & 0 \\ 0 & 4 \end{pmatrix}.$$

From (3.22), K_y can be written as

$$K_y = \begin{pmatrix} 4 + \text{kurt}(y_1) & 0 \\ 0 & 4 \end{pmatrix}. \tag{3.35}$$

The ICS:kurtosis:variance criterion (3.24) reduces to

$$\begin{aligned} \kappa_{\text{ICS}}(\phi) &= b^T K_y b \\ &= 4 + 16\delta^4 q(1-q)\{1 - 6q(1-q)\}\cos^2(\phi) \\ &= 4 + \text{kurt}(y_1)\cos^2(\phi). \end{aligned} \tag{3.36}$$

where

$$b = (\cos(\phi)\sin(\phi))^T. \tag{3.37}$$

The minimum/maximum value of (3.36) is the smallest/largest eigenvalue of K_y when ϕ is in the direction of the corresponding eigenvector.

The eigenvalues and eigenvectors of K_y are

$$\begin{aligned} &\lambda_1 = 4 + \text{kurt}(y_1), \lambda_2 = 4 \\ &\gamma_1 = (1,0)^T, \gamma_2 = (0,1)^T. \end{aligned} \tag{3.38}$$

As we have shown in Section 2.2, the eigenvalues of $\Sigma_x^{-1}K_x$ and K_y are the same. In our model, the eigenvectors of $\Sigma_x^{-1}K_x$ and K_y are also the same, since Σ_x and K_x are diagonal. The effect of standardization is explored in Section 3.6.

3.4 Relationship between ICS:kurtosis:variance and Mardia's multivariate kurtosis measure

ICS:kurtosis:variance can be related to Mardia's measure of kurtosis, Mardia et al. (1980). Consider the bivariate random vector y, defined in (3.13), Mardia's kurtosis is defined as follows

$$\beta_{2,2} = \mathrm{E}\{(y-\mu_y)^T(y-\mu_y)\}^2. \tag{3.39}$$

Then $\beta_{2,2}$ can be expressed in terms of moments as follows

$$\begin{aligned}\beta_{2,2} &= \mu_{y_1}(4) + \mu_{y_2}(4) + 2\mu_{y_1y_2}(2,2)\\ &= \mu_{y_1}(4) + 5 = \mathrm{kurt}(y_1) + 8.\end{aligned} \tag{3.40}$$

From (3.40) and (3.35), and as Peña et al. (2010) pointed out,

$$\beta_{2,2} = \mathrm{trace}(K_y). \tag{3.41}$$

This means that $\beta_{2,2}$ cannot be used as a criterion to identify the groups separation direction, since it gives the aggregate fourth-order moments and cross moments of the random vector components. On the other hand, ICS:kurtosis:variance partitions those moments into combination of fourth-order moments that form the eigenvalue corresponding to the eigenvector which is in the direction of the groups separation, and another combination of fourth-order moments that form the eigenvalue corresponding to the eigenvector which is in the direction of the normal noise.

3.5 Projection pursuit based on kurtosis in population

Let $x = (x_1, x_2)^T$ be a bivariate random vector, standardized with respect to within-group covariance matrix as in (3.2), such that $W_x = I_2$. Consider the following linear transformation, where the unit vector $a \in R^2$ is defined as in (3.25),

$$a^T x = x_1 \cos(\theta) + x_2 \sin(\theta). \tag{3.42}$$

Substituting for x_1, as defined in (3.8), $a^T x$ becomes

$$\begin{aligned} a^T x &= \alpha s \cos(\theta) + \epsilon_1 \cos(\theta) + x_2 \sin(\theta) \\ &= \nu + \beta s, \end{aligned} \tag{3.43}$$

where $\nu = \epsilon_1 \cos(\theta) + x_2 \sin(\theta) \sim N(0, 1)$, and $\beta = \alpha \cos(\theta)$.

The PP:kurtosis:variance criterion is given as follows

$$\kappa_{\text{PP}}(\theta) = \text{kurt}(\theta) = \frac{\text{E}\{\nu + \beta s - \text{E}(\beta s)\}^4}{[\text{E}\{\nu + \beta s - \text{E}(\beta s)\}^2]^2} - 3. \tag{3.44}$$

The numerator can be computed with the help of the central moments of s from (3.16) and the moments of the normally distributed variable ν, as follows

$$\begin{aligned} \text{E}\{(\nu + \beta s - \text{E}(\beta s))^4\} &= \text{E}\{\Big(\nu + \beta(s - \mu_s^{'}(1))\Big)^4\} \\ &= \text{E}\{\nu^4\} + 6\text{E}\{\nu^2\}\beta^2\text{E}\{(s - \mu_s^{'}(1))^2\} + \beta^4\text{E}\{(s - \mu_s^{'}(1))^4\} \\ &= 3 + 24\beta^2 q(1-q) + \beta^4(-48q^2(1-q)^2 + 16q(1-q)). \end{aligned}$$

Substituting by $\beta = \alpha \cos(\theta)$ in the previous equation, gives

$$\text{E}\{(\nu + \beta s - \text{E}(\beta s))^4\} = 3 + 24q(1-q)\alpha^2 \cos^2(\theta) + (16q(1-q) - 48q^2(1-q)^2)\alpha^4 \cos^4(\theta). \tag{3.45}$$

The denominator is given by

$$\begin{aligned}[\mathrm{E}\{(\nu+\beta(s-\mathrm{E}(s)))^2\}]^2 &= [\mathrm{E}\{\nu^2\}+\beta^2\mathrm{E}\{\Big(s-\mu_s^{'}(1)\Big)^2\}]^2 \\ &= [1+4q(1-q)\beta^2]^2 = [1+4q(1-q)\alpha^2\cos^2(\theta)]^2.\end{aligned} \tag{3.46}$$

Substituting (3.46) and (3.45) in (3.44) gives a formula for the criterion PP:kurtosis:variance, as follows.

$$\mathrm{kurt}(\theta) = \frac{\{16q(1-q)[1-6q(1-q)]\}\alpha^4\cos^4\theta}{\{1+4q(1-q)\alpha^2\cos^2\theta\}^2}. \tag{3.47}$$

Minimizing or maximizing (3.47) over θ depends on a number of parameters, the mixing proportion q, the group separation parameter α. As $\alpha \to \infty$, $\mathrm{kurt}(\theta)$ reduces to $\mathrm{kurt}(s)$, defined (2.10).

Let y be a random variable, standardized beforehand with respect to the total covariance, as in (3.9), such that $\Sigma_y = I_2$. Consider the following linear transformation

$$\begin{aligned}b^T y &= y_1\cos(\phi) + y_2\sin(\phi) \\ &= \delta s\cos(\phi) + \epsilon_1^*\cos(\phi) + y_2\sin(\phi) \\ &= \beta^* s + \nu^*,\end{aligned}$$

where $\nu^* = \epsilon_1^*\cos(\phi)+y_2\sin(\phi) \sim N(0, 1-4\delta^2 q(1-q)\cos^2(\phi))$, and $\beta^* = \delta\cos(\phi)$, $0 \le \delta \le 1$ is a separation parameter.

The PP:variance:kurtosis criterion is given by

$$\kappa_{\mathrm{PP}}(\phi) = \mathrm{kurt}(\phi) = \mathrm{E}\{\nu^* + \beta^* s - \mathrm{E}(\beta^* s)\}^4 - 3. \tag{3.48}$$

First, we calculate the fourth-order moment in (3.48) as follows

$$\mathrm{E}\{\nu^* + \beta^* s - \mathrm{E}(\beta^* s)\}^4 = 3 + 16q(1-q)\{1-6q(1-q)\}\delta^4\cos^4(\phi). \tag{3.49}$$

Substituting (3.49) in (3.48) gives

$$\kappa_{\mathrm{PP}}(\phi) = 16q(1-q)\{1-6q(1-q)\}\delta^4\cos^4(\phi). \tag{3.50}$$

Like (3.47), optimizing (3.50) requires numerical optimization. The effect of standardization is explored in Section 3.6.

3.6 A comparison between ICS and PP

3.6.1 In population

Optimization

In this Section we compare $\kappa_{\mathrm{ICS}}(\theta)$ and $\kappa_{\mathrm{PP}}(\theta)$, from (3.30) and (3.47). We consider the case of equal groups, i.e $q = 1/2$. In this case, the ICS:kurtosis:variance criterion from (3.30) becomes

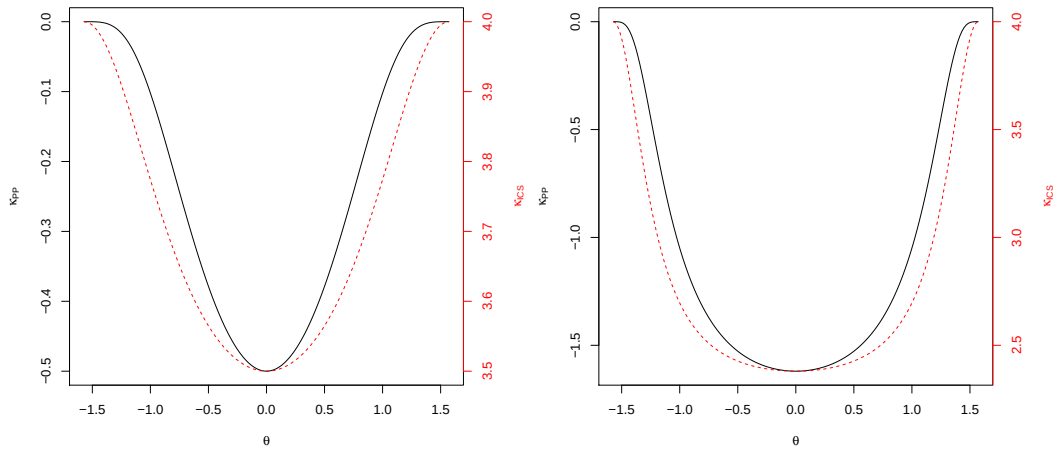
$$\kappa_{\mathrm{ICS}}(\theta) = 4 - \frac{2\alpha^4\cos^2(\theta)}{\{1+\alpha^2\}\{1+\alpha^2\cos^2(\theta)\}}, \tag{3.51}$$

and the PP:kurtosis:variance criterion from (3.47) becomes

$$\kappa_{\mathrm{PP}}(\theta) = -\frac{2\alpha^4\cos^4(\theta)}{\{1+\alpha^2\cos^2(\theta)\}^2}. \tag{3.52}$$

To compare ICS:kurtosis:variance and PP:kurtosis:variance criteria, we plot (3.51) and superimpose it onto the plot of (3.52).

From Figure 3.1, the plot of $\kappa_{\mathrm{PP}}(\theta)$ have a similar behavior to the plot of $\kappa_{\mathrm{ICS}}(\theta)$, minimized at the group separation direction $\theta = 0°$. Also, as α increases, the ratio of the maximum to the minimum values of $\kappa_{\mathrm{ICS}}(\theta)$ and $\kappa_{\mathrm{PP}}(\theta)$ increases. And as $\alpha \to \infty$, $\kappa_{\mathrm{ICS}}(\theta) = 2$, and $\kappa_{\mathrm{PP}}(\theta) = -2$, for $\theta \neq \pi/2$.



(a) $\alpha = 1$ (b) $\alpha = 3$

Figure 3.1: Plot of the population criteria $\kappa_{\text{ICS}}(\theta)$ (red dotted line), and $\kappa_{\text{PP}}(\theta)$ (solid black line) versus θ, for $q = 1/2$, $\alpha = 1$, and 3.

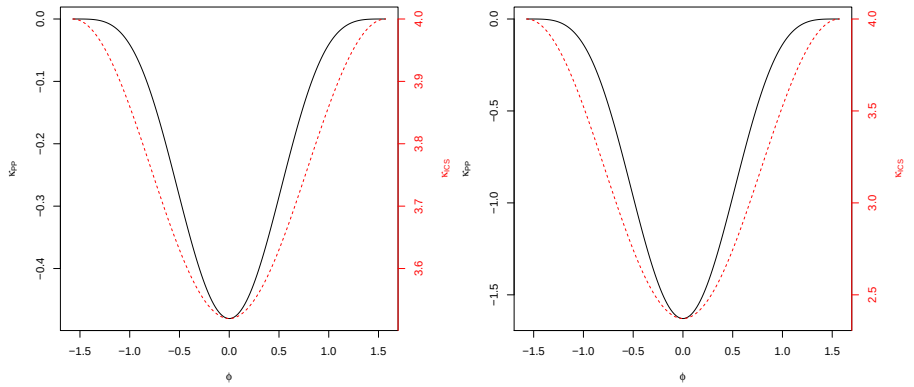
The effect of standardization

To illustrate the effect of standardization on the ICS:kurtosis:variance and PP:kurtosis:variance, we plot $\kappa_{\text{ICS}}(\phi)$ and $\kappa_{\text{PP}}(\phi)$ from (3.36) and (3.50), respectively, for $q = 1/2$.

For $q = 1/2$,

$$\begin{aligned} \kappa_{\text{ICS}}(\phi) &= 4 - 2\delta^4 \cos^2(\phi), \\ \kappa_{\text{PP}}(\phi) &= -2\delta^4 \cos^4(\phi). \end{aligned} \tag{3.53}$$

Figure 3.2 shows plots of (3.53). From Figures (3.1) and (3.2), standardization does not change the minimum and maximum values of κ_{ICS} and κ_{PP}. The difference is only in the scale of the angles θ and ϕ. In particular, ϕ is a scaled version of θ.



(a) $\delta = 0.7$ (b) $\delta = 0.95$

Figure 3.2: Plot of the population criteria $\kappa_{\text{ICS}}(\phi)$ (red dotted line), and $\kappa_{\text{PP}}(\phi)$ (solid black line) versus ϕ, for $q = 1/2$, $\delta = 0.7$, and 0.95.

3.6.2 In sample

Optimization

Let X be an $n \times 2$ data matrix, its rows are given by $x_i^T = (x_{i1}, x_{i2})$, for $i = 1, \ldots, n$. Suppose that X is shifted such that it has a zero mean vector.

The sample ICS:kurtosis:variance and PP:kurtosis:variance criteria are defined as follows

$$\begin{aligned} \hat{\kappa}_{\text{ICS}}(\theta) &= \frac{a^T \hat{K}_x a}{a^T S_x a}, \\ \hat{\kappa}_{\text{PP}}(\theta) &= \text{kurt}(Xa), \end{aligned} \tag{3.54}$$

where a is defined in (3.25), $\hat{K}_x$ is the sample version of K_x, defined in (3.23),

$$\hat{K}_x = \frac{1}{n} \sum_{i=1}^{n} x_i x_i^T (x_i^T S_x^{-1} x_i), \tag{3.55}$$

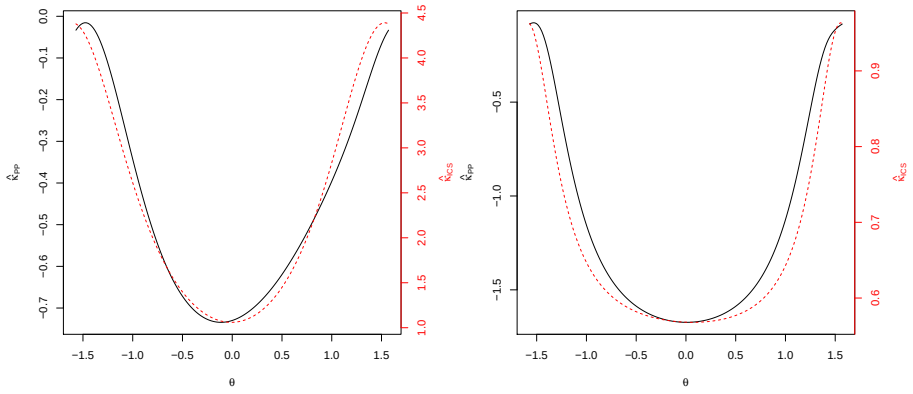
and the kurtosis is defined in (3.10).

The ICS and PP estimates of the groups separation direction, denoted by $\hat{\theta}_{\text{ICS}}$ and $\hat{\theta}_{\text{PP}}$, are the directions that minimizes/maximizes $\hat{\kappa}_{\text{ICS}}(\theta)$ and $\hat{\kappa}_{\text{PP}}(\theta)$.

Minimizing/maximizing $\hat{\kappa}_{\text{ICS}}(\theta)$ is computationally simple and carried out an-

alytically. The estimate $\hat{\theta}_{\text{ICS}}(\theta)$ is in the direction of the smallest/largest eigenvector of $S_x^{-1}\hat{K}_x$. On the other hand, minimizing/maximizing $\hat{\kappa}_{\text{PP}}(\theta)$ is computationally expensive, and requires numerical optimization.

Before discussing the numerical optimization procedure used for $\hat{\kappa}_{\text{PP}}$, we must first plot the sample criteria $\hat{\kappa}_{\text{PP}}(\theta)$ and $\hat{\kappa}_{\text{ICS}}(\theta)$, as a function of θ. The population plots are perfectly symmetric around zero, while the sample plots are not due to sampling noise. Figure 3.3 shows plots of $\hat{\kappa}_{\text{ICS}}(\theta)$ and $\hat{\kappa}_{\text{PP}}(\theta)$.



(a) $\alpha = 1$ (b) $\alpha = 3$

Figure 3.3: Plot of $\hat{\kappa}_{\text{ICS}}(\theta)$ (red dotted line), and $\hat{\kappa}_{\text{PP}}(\theta)$ (solid black line) versus θ, for $q = 1/2$, $\alpha = 1$, and 3.

The plots of Figure 3.3 are based on two samples of size 200, generated from (3.8), with $q = 1/2$, and $\alpha = 1$ and 3, to investigate the behaviour of $\hat{\kappa}_{\text{PP}}$ in the case of slightly separated groups and well separated groups. Figure 3.3 shows that the curve of $\hat{\kappa}_{\text{PP}}(\theta)$ is smooth as a function of θ. More importantly, it has a similar behavior to the curve $\hat{\kappa}_{\text{ICS}}(\theta)$. Thus, we can use a local search method with $\hat{\theta}_{\text{ICS}}$ being used as a starting point.

Another optimization approach of $\hat{\kappa}_{\text{PP}}(\theta)$ is to find the global minimum of $\hat{\kappa}_{\text{PP}}$. This can be done by partitioning the domain $-\pi/2 < \theta \leq \pi/2$ into N, say, equal points, then evaluate $\hat{\kappa}_{\text{PP}}(\theta)$ at each point. The value of θ which has the minimum kurtosis will be taken as the estimate $\hat{\theta}_{\text{PP}}$.

This approach is feasible in $p = 2$, i.e one-dimension optimization over θ. But

as the dimension p increases, the complexity of the search will increase.

The effect of standardization

Without loss of generality, assume that X is standardized with respect to $S_x^{-1/2}$ as follows

$$Y = XS^{-1/2},$$

such that $S_y = I_2$.

Consider the linear transformation Yb, where $b = (\cos(\phi), \sin(\phi))^T$. The sample criterion are

$$\hat{\kappa}_{\text{ICS}}(\phi) = b^T \hat{K}_y b = b^T S_x^{-1/2} \hat{K}_x S_x^{-1/2} b,$$
$$\hat{\kappa}_{\text{PP}}(\phi) = \text{kurt}(Yb) = \sum_{i=1}^{n} \{[b^T y_i]^4\} - 3,$$

From (2.4),

$$b \propto S_x^{1/2} a. \tag{3.56}$$

We have shown in Section 2.2 that $S_x^{-1}\hat{K}_x$ and $\hat{K}_y$ have the same eigenvalues, but the eigenvectors of $\hat{K}_y$ are the eigenvectors of $S_x^{-1}\hat{K}_x$ scaled by $S^{1/2}$.

For PP, we can find a mapping between $\hat{\theta}_{\text{PP}}$ and $\hat{\phi}_{\text{PP}}$ to explore the effect of standardization on PP:kurtosis:variance. Let $S_x^{1/2}$ be given by

$$S_x^{1/2} = \begin{pmatrix} c_1 & c_{12} \\ c_{12} & c_2 \end{pmatrix}.$$

From (3.56),

$$\begin{aligned}\begin{pmatrix}\cos(\phi)\\ \sin(\phi)\end{pmatrix} &\propto \begin{pmatrix}c_1 & c_{12}\\ c_{12} & c_2\end{pmatrix}\begin{pmatrix}\cos(\theta)\\ \sin(\theta)\end{pmatrix}\\ &= \begin{pmatrix}c_1\cos(\theta)+c_{12}\sin(\theta)\\ c_{12}\cos(\theta)+c_2\sin(\theta)\end{pmatrix}\\ &= \begin{pmatrix}R_1\cos(\theta+\beta_1)\\ R_2\sin(\theta+\beta_2)\end{pmatrix}, \end{aligned} \tag{3.57}$$

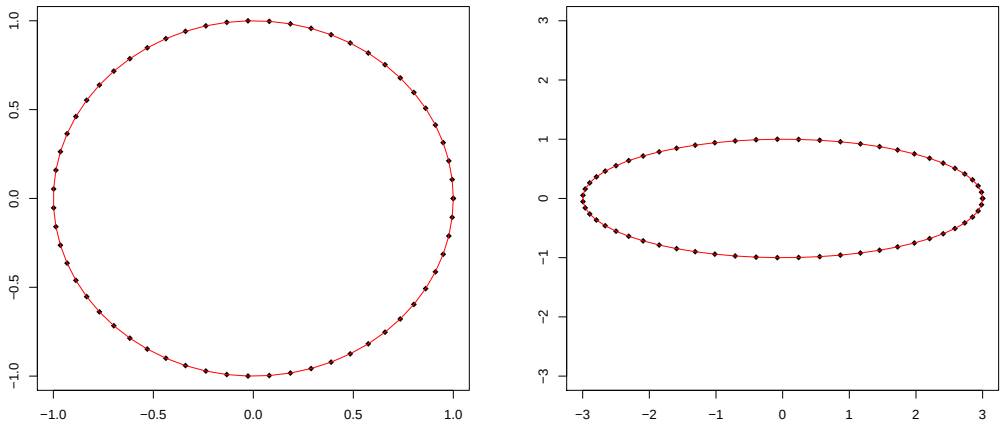
where

$$\begin{aligned} R_1 &= \sqrt{c_1^2+c_{12}^2}, \quad \beta_1 = \operatorname{atan}(c_{12}/c_1),\\ R_2 &= \sqrt{c_2^2+c_{12}^2}, \quad \beta_2 = \operatorname{atan}(c_{12}/c_2). \end{aligned} \tag{3.58}$$

For example if $S_x^{-1/2}$ is diagonal, i.e $c_{12}=0$, (3.57) reduces to

$$\begin{pmatrix}\cos(\phi)\\ \sin(\phi)\end{pmatrix} = \begin{pmatrix}c_1\cos(\theta)\\ c_2\sin(\theta)\end{pmatrix}.$$

This means that standardization maps θ from evenly spaced points in the unit circle to unevenly spaced points in an ellipse with points become dense at the corner of the ellipse and sparse elsewhere, as shown in the Figure 3.4. The optimum value ϕ might be missed if it is not located at the corner of the ellipse, and the ratio c_1/c_2 is large.



(a) θ (b) ϕ

Figure 3.4: Plots of θ and ϕ, when $c_1 = 3$, $c_2 = 1$.

3.7 Axis measure of dispersion

There are two ways to define an axis in two dimensional space. One way is by an angle ψ, noting that a direction and the opposite direction are equivalent, i.e

$$\psi \equiv \psi + \pi.$$

Another way to define an axis is by doubling the angle ψ, noting that each direction 2ψ and $2\psi + 2\pi$ refer to the same angle.

Let ψ and ω be two angles defining two different axes. To measure the distance between the axis defined by ψ and the one defined by ω we can use the following distance measure, Mardia and Jupp (2009),

$$\begin{aligned} d^2(\psi, \omega) &= \frac{1}{2}(1 - \cos 2(\psi - \omega)) \\ &= \sin^2(\psi - \omega). \end{aligned} \tag{3.59}$$

The squared distance measure $d^2(\psi, \omega)$ takes the following possible values

$$d^2(\psi,\omega) = \begin{cases} 0 & \text{if } \psi = \omega \\ 1 & \text{if } \psi = \omega + \pi/2 \text{ or } \psi = \omega + 3\pi/2 \\ (0,1) & \text{if } 0 < |\psi - \omega| < \pi/2. \end{cases} \tag{3.60}$$

In the following lemma, we show that the distance measure d satisfies the triangle inequality.

Lemma 3.7.1. *For any three angles, ω_1, ω_2, and ω_3, defining axes, d satisfies the triangle inequality as follows,*

$$d(\omega_1,\omega_2) + d(\omega_2,\omega_3) \geq d(\omega_1,\omega_3).$$

Proof. We first relate d^2, defined in (3.59), to the Euclidean squared distance between the points

$$\begin{pmatrix} \cos(2\omega_1) \\ \sin(2\omega_1) \end{pmatrix} \text{ and } \begin{pmatrix} \cos(2\omega_2) \\ \sin(2\omega_2) \end{pmatrix}. \tag{3.61}$$

First we compute the Euclidean distances, d_E^2 between the points $(\cos(2\omega_1), \sin(2\omega_1))$, and $(\cos(2\omega_2), \sin(2\omega_2))$,

$$\begin{aligned} d_E^2(\omega_1,\omega_2) &= (\cos(2\omega_1) - \cos(2\omega_2))^2 + (\sin(2\omega_1) - \sin(2\omega_2))^2 \\ &= (\cos^2(2\omega_1) + \sin^2(2\omega_1)) + (\cos^2(2\omega_2) + \sin^2(2\omega_2)) \\ &- 2(\cos(2\omega_1)\cos(2\omega_2) + \sin(2\omega_1)\sin(2\omega_2)) \\ &= 2 - 2\cos 2(\omega_1 - \omega_2). \end{aligned} \tag{3.62}$$

From (3.59), (3.62) can be written as

$$\begin{aligned} d_E^2(\omega_1,\omega_2) &= 4d^2(\omega_1,\omega_2) \\ d_E(\omega_1,\omega_2) &= 2d(\omega_1,\omega_2). \end{aligned} \tag{3.63}$$

Equations (3.63) means that the Euclidean distance between the two points defined in (3.61) is proportional to the axis distance in (3.59). Therefore, d satisfies the triangle inequality. □

Let ψ be a random angle defining an axis. A measure of spread of ψ around an axis ω can be defined as follows

$$v(\psi) = \mathrm{E}(\sin^2(\psi - \omega)). \tag{3.64}$$

If the distribution of ψ is concentrated around $\omega + \pi/2$ or $\omega + 3\pi/2$, $v(\psi) = 1$. If the distribution of ψ is concentrated around ω, $v(\psi) = 0$. If ψ is uniformly distributed, $v(\psi) = 1/2$.

Let $\hat{\theta} = (\hat{\theta}_1, \ldots, \hat{\theta}_m)$ be estimates of the group separation direction θ estimated by ICS and PP, let θ_0 be the true group separation direction. From (3.64), the sample axes measure of spread of $\hat{\theta}$ about the true value θ_0 is

$$\hat{v}(\hat{\theta}) = \frac{1}{m}\sum_{j=1}^{m} \sin^2(\hat{\theta}_j - \theta_0). \tag{3.65}$$

From (3.65), if $\hat{\theta}$ is highly concentrated around θ_0, $v(\hat{\theta})$ will be approximately equal to zero. If $\hat{\theta}$ is concentrated around $\theta_0 + \pi/2$ or $\theta_0 + 3\pi/2$, $v(\hat{\theta})$ will be approximately equal to 1. If $\hat{\theta}$ is uniformly distributed around the circle, $v(\hat{\theta})$ will be approximately equal to 1/2.

3.8 Simulation study

The data sets used in this simulation study are generated from the mixture model (3.1), with means $\mu_1 = (\alpha, 0)^T$ and $\mu_2 = (-\alpha, 0)^T$, for $\alpha > 0$, and $W = I_2$. The true separation direction is $\theta = 0°$.

We vary a number of parameters that have an effect on the performance of ICS:kurtosis:variance and PP:kurtosis:variance. These parameters are given as

follows:

- The mixing proportion q: We consider first the case of equal groups with $q = 1/2$, then we consider the case of unequal groups with $q = 1/4$, not far from half.

- The separation parameter α: $\alpha = 1$ gives two slightly separated groups, and $\alpha = 3$, gives two well separated groups.

- The sample size: The sample sizes used in this study are $n = 20, 50, 200, 500$.

For each choice of the parameters, we simulate $m = 1000$ samples. Applying ICS:kurtosis:variance and PP:kurtosis:variance give two sets of estimates: $\hat{\theta}^{(j)}_{\text{ICS}}$ and $\hat{\theta}^{(j)}_{\text{PP}}$, $j = 1, \ldots, 1000$.

From (3.65), the measure of spread of $\hat{\theta}_{\text{ICS}}$ and $\hat{\theta}_{\text{PP}}$ are

$$\hat{v}(\hat{\theta}_{\text{ICS}}) = \frac{1}{m}\sum_{j=1}^{m} \sin^2(\hat{\theta}^{(j)}_{\text{ICS}})$$
$$\hat{v}(\hat{\theta}_{\text{PP}}) = \frac{1}{m}\sum_{j=1}^{m} \sin^2(\hat{\theta}^{(j)}_{\text{PP}}).$$

Figure 3.3 shows plots of $\hat{v}(\hat{\theta}_{\text{ICS}})$, the red dashed line, and $\hat{v}(\hat{\theta}_{\text{PP}})$, the solid black line, for different values of parameters.

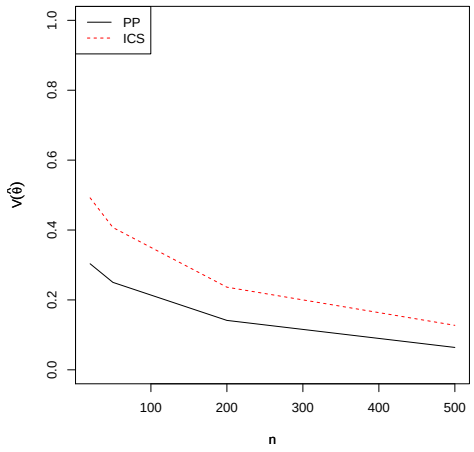
The plots in Figure 3.5 show the following

- PP:kurtosis:variance is more accurate than ICS:kurtosis:variance when $q = 1/2$, and $q = 1/4$.

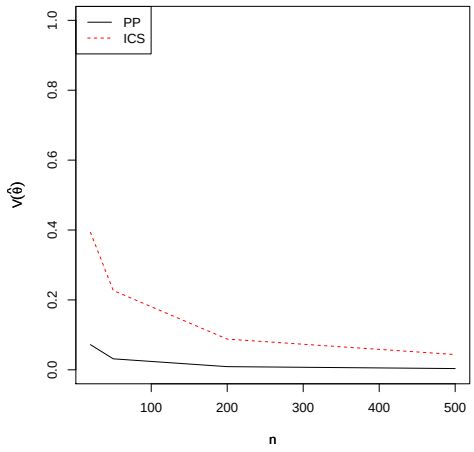
- The accuracy of PP:kurtosis:variance is not affected by changing q from $1/2$ to $1/4$, whereas ICS:kurtosis:variance seems to be highly affected by changing q.

- PP:kurtosis:variance is more accurate than ICS:kurtosis:variance for small n. This result agrees with Peña et al. (2010) who pointed out that if the ratio n/p is small, PP:kurtosis:variance is more accurate.

$q = 1/2$

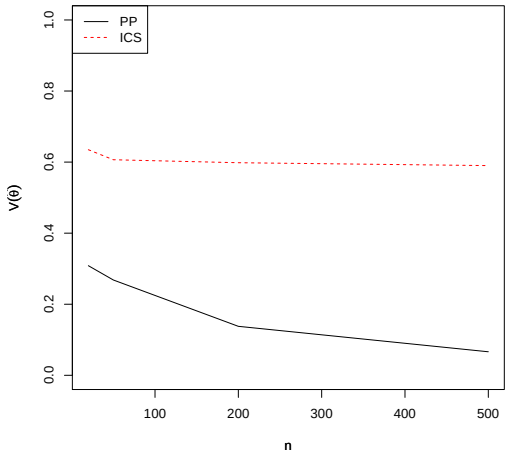


(a) $\alpha = 1$

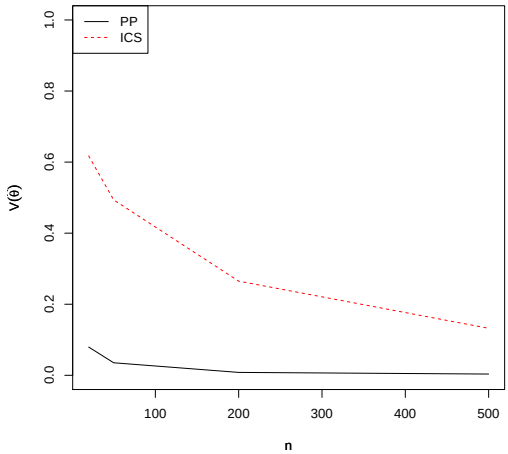


(b) $\alpha = 3$

$q = 1/4$



(c) $\alpha = 1$



(d) $\alpha = 3$

Figure 3.5: For $\alpha = 1, 3$ and $q = 1/2,\ 1/4$, the plots of $\hat{v}(\hat{\theta}_{\mathrm{PP}})$ (black solid curves) and $\hat{v}(\hat{\theta}_{\mathrm{ICS}})$ (red dashed curves).

3.9 Discussion

We have studied ICS:kurtosis:variance and PP:kurtosis:variance criteria under mixtures of two bivariate normal distributions. We also compared the accuracy of the two methods through a simulation study. The simulation results show that PP:kurtosis:variance is more accurate than ICS:kurtosis:variance.

In the following, we discuss extending the dimension p, and the number of groups k.

Most of the work done can be extended to higher dimensions. Suppose that the model in Section 3.2 extended to dimension $p > 2$. Let $x = (x_1, \ldots, x_p)$ be a p-variate random vector distributed a s a mixture of two p-variate normal distributions. The random vector x can be written as,

$$\begin{pmatrix} x_1 \\ x_2 \\ \vdots \\ x_p \end{pmatrix} = s \begin{pmatrix} \alpha \\ 0 \\ \vdots \\ 0 \end{pmatrix} + \begin{pmatrix} \epsilon_1 \\ \epsilon_2 \\ \vdots \\ \epsilon_p \end{pmatrix}, \tag{3.66}$$

where s is defined as in (2.11), $\epsilon = (\epsilon_1, \ldots, \epsilon_p)^T \sim N(0, I_p)$. Hence, the following concepts can be extended as follows.

ICS criterion

The ICS:kurtosis:variance criterion is,

$$\kappa_{\text{ICS}}(a) = \frac{a^T K_x a}{a^T \Sigma_x a}, \tag{3.67}$$

where $a = (a_1, \ldots, a_p)$ is a unit vector. The minimum value of $\kappa_{\text{ICS}}(a)$ is obtained

when a is the smallest/largest eigenvector of $\Sigma_x^{-1}K_x$

$$\Sigma_x^{-1}K_x = \begin{pmatrix} p+2+\text{kurt}(x_1) & 0 & \dots & 0 \\ 0 & p+2 & \dots & 0 \\ \vdots & \vdots & \ddots & \vdots \\ 0 & 0 & \dots & p+2 \end{pmatrix}.$$

Criterion (3.67), takes the following form,

$$\kappa_{\text{ICS}}(a) = (p+2) + \frac{16q(1-q)(1-6q(1-q))\alpha^4 a_1^2}{\{1+4q(1-q)\alpha^2\}\{1+4q(1-q)\alpha^2 a_1^2\}}.$$

PP criterion

Consider the following linear transformation,

$$\begin{aligned} a^T x &= a_1 x_1 + \ldots + a_p x_p \\ &= a_1\alpha s + a_1\epsilon_1 + a_2 x_2 + \ldots + a_p x_p \\ &= \beta s + \nu, \end{aligned}$$

where $\beta = a_1\alpha_1$, $\nu \sim N(0,1)$. Thus, the PP:kurtosis:variance criterion, defined in (3.43), takes a similar form of (3.47), as follows

$$\kappa_{\text{PP}}(a) = \frac{\{16q(1-q)[1-6q(1-q)]\}\alpha^4 a_1^4}{\{1+4q(1-q)\alpha^2 a_1^2\}^2}. \tag{3.68}$$

Optimization

As in the two-dimensional case, optimizing (3.67) is carried out analytically, whereas (3.68) requires numerical optimization. In Section 3.6, we have mentioned that we can use a local or global search methods with κ_{PP}. Local search, using the ICS estimate as a starting point, becomes less accurate as p increase.

The axis measure of spread

Let ν, $h \in R^p$ be two unit vectors defining axes, i.e $\nu \equiv -\nu$, $h \equiv -h$. A useful

distance measure between ν and h is

$$\begin{aligned} d^2(\nu, h) &= 1 - (\nu^T h)^2 \\ &= \text{tr}((\nu\nu^T - hh^T)^2). \end{aligned} \tag{3.69}$$

If $\nu = (1, 0, \ldots, 0)^T$, (3.69) reduces to (3.59).

Suppose now that ν is a random axis and h is the mean axis, the axial variance is defined as

$$V = 1 - \text{E}\{(\nu^T h)^2\}.$$

Number of groups

A second issue is extending the number of groups in model (3.1) to $k > 2$. The performance of the methods depend on the arrangement of the groups in the p-dimensional space.

For example, suppose we have three groups on the three dimensional space. If the groups are arranged in one line, the optimization used will be the same as discussed in the previous sections. If the groups are arranged on a triangle vertices, we will need to find a two-dimensional subspace that view the three groups.

The two directions can be found by sequentially optimizing (3.67) and (3.68). As Tyler et al. (2009) pointed out, the sequential optimization of (3.67) is straightforward, pick the two eigenvectors corresponding to the two eigenvalues that have extreme values.

On the other hand, the sequential optimization of (3.68) is more complicated. An additional constraint must be added, the two directions must be orthogonal, $a_1^T a_2 = 0$, say. Sequential optimization of PP criteria is discussed in Croux et al. (2007).

Chapter 4

An analytical comparison between ICS and PP

4.1 Introduction

In Chapter 3, we have compared the accuracy of ICS:kurtosis:variance and PP:kurtosis:variance, under two-group mixtures of bivariate normal distribution. The results show that PP:kurtosis:variance is more accurate than ICS:kurtosis:variance.

In this chapter, we find the asymptotic behaviour of ICS:kurtosis:variance and PP:kurtosis:variance estimates of the group separation direction.

The chapter is organized as follows. The model is defined in Section 4.2. The model used in this chapter is a special case of the mixture model explained in Section 3.2. We review the asymptotic theory of sample moments Section 4.3. The asymptotic behaviour of the estimates of group separation parameter of ICS:kurtosis:variance and PP:kurtosis:variance are discussed in Sections 4.4 and 4.5. In Section 4.6, we compare the asymptotic variances and the sample variances of ICS:kurtosis:variance and PP:kurtosis:variance.

4.2 The model

4.2.1 Assumptions

Let $y^T = (y_1, y_2)$ be a bivariate random vector distributed as an equal mixture of two normal distributions, as in (3.1), with $q = 1/2$.

Without loss of generality, assume that y is standardized with respect to the total covariance matrix as in (3.9), such that $\Sigma_y = I_2$.

There is only one parameter that describes this model, which is $0 \leq \delta \leq 1$. The case of $\delta = 0$ produces one group distributed as a bivariate standard normal distribution. The case of $\delta = 1$ produces two widely separated parallel lines.

The total mean vector is $(0, 0)^T$. The between-group scatter matrix B_y is given by

$$\begin{aligned} B_y &= \frac{1}{2}\mu_1\mu_1^T + \frac{1}{2}\mu_2\mu_2^T \\ &= \begin{pmatrix} \delta^2 & 0 \\ 0 & 0 \end{pmatrix}. \end{aligned}$$

The within-group scatter matrix

$$\begin{aligned} W_y &= I_2 - B_y \\ &= \begin{pmatrix} 1-\delta^2 & 0 \\ 0 & 1 \end{pmatrix}. \end{aligned}$$

The random vector y can be written as in (3.14).

4.2.2 Moments

We derive the univariate and cross moments of the components of y, y_1 and y_2, up to eighth order, to use them in Sections 4.4 and 4.5.

Since the mean vector of y equals to $(0, 0)^T$, the the central and non-central

moments and cross moments are equivalent. The univariate moments and cross moments are defined in (3.15) and (3.16), respectively.

Also since, y_1 and y_2 are independent,

$$\mu_{y_1 y_2}(h, j) = \mu_{y_1}(h)\mu_{y_2}(j).$$

In the following, a list of all moments of y_1 and y_2, up to eighth order, is given.

- All odd moments y_1 and y_2 are equal to zero.
- The even univariate moments of y_1 are

$$\begin{aligned}
&\mu_{y_1}(2) = 1, \quad \mu_{y_1}(4) = \mathrm{E}\{(s\delta + \epsilon_1^*)^4\} = 3 - 2\delta^4, \\
&\mu_{y_1}(6) = \mathrm{E}\{(s\delta + \epsilon_1^*)^6\} \\
&\qquad = 15 - 30\delta^4 + 16\delta^6, \\
&\mu_{y_1}(8) = \mathrm{E}\{(s\delta + \epsilon_1^*)^8\} \\
&\qquad = 105 - 420\delta^4 + 448\delta^6 - 132\delta^8.
\end{aligned} \tag{4.1}$$

- The even univariate moments of y_2 are

$$\begin{aligned}
&\mu_{y_2}(2) = 1, \quad \mu_{y_2}(4) = 3, \\
&\mu_{y_2}(6) = 15, \quad \mu_{y_2}(8) = 105.
\end{aligned} \tag{4.2}$$

- The even cross moments of y_1 and y_2 are

$$\begin{aligned}
&\mu_{y_1 y_2}(2, 4) = \mu_{y_2}(4), \quad \mu_{y_1 y_2}(4, 2) = \mu_{y_1}(4), \\
&\mu_{y_1 y_2}(2, 6) = \mu_{y_2}(6), \quad \mu_{y_1 y_2}(6, 2) = \mu_{y_1}(6) \\
&\mu_{y_1 y_2}(4, 4) = \mu_{y_1}(4)\mu_{y_2}(4) = 9 - 6\delta^4.
\end{aligned} \tag{4.3}$$

4.3 The asymptotic theory of sample moments

For a bivariate $n \times 2$ data matrix U, where its rows can be written as $u_i^T = (u_{i1}, u_{i2})$, $i = 1, \ldots, n$. To simplify the algebra, assume that the population mean vector is $(0, 0)^T$, thus the $(h + j)$-order sample moment $m(h, j) = m_{u_1 u_2}(h, j)$ is defined as follows

$$m(h, j) = \frac{1}{n} \sum_{i=1}^{n} u_{i1}^h u_{i2}^j. \tag{4.4}$$

By the central limit theorem, $\sqrt{n}(m(h, j) - \mu(h, j))$ is asymptotically normal with mean 0, variance and covariance given as follows, (Kendall and Stuart (1977), page 244)

$$\text{var}\{m(h, j)\} = \{\mu(2h, 2j) - \mu^2(h, j)\}, \tag{4.5}$$

$$\text{cov}\{m(h, j), m(u, v)\} = \{\mu(h + u, j + v) - \mu(h, j)\mu(u, v)\}. \tag{4.6}$$

In general, a vector of l sample moments is distributed asymptotically as an l-variate normal distribution with mean vector and covariance matrix given in the following theorem (which follows Serfling (1980), Theorem B, page 72, and (4.5) and (4.6)).

Theorem 4.3.1. *Let $m = (m(h_1, j_1), m(h_2, j_2), \ldots, m(h_l, j_l))$ be a vector of sample moments, such that $m(2h_k, 2j_k) < \infty$. The vector $\sqrt{n}(m - \mu)$, where $\mu = (\mu(h_1, j_1), \mu(h_2, j_2), \ldots, \mu(h_l, j_l))$, converges in distribution to an l-variate normal distribution with mean vector 0 , and $l \times l$ covariance matrix Σ, given by*

$$\Sigma = (\sigma_{ki}), \tag{4.7}$$

where $\sigma_{ki} = \{\mu(h_k + h_i, j_k + j_i) - \mu(h_k, j_k)\mu(h_i, j_i)\}$, $k = i = 1 \ldots, l$.

Suppose now that we have a vector-valued function g of m. The asymptotic distribution of g is shown in the following theorem by Serfling (1980),

Theorem 4.3.2. *Let $m = (m(h_1, j_1), m(h_2, j_2), \ldots, m(h_l, j_l))$ be a vector of sample moments, where $\sqrt{n}(m - \mu)$ distributed asymptotically as $N(0, \Sigma)$, where Σ is shown in (4.7). Let $g(m) = (g_1(m), \ldots, g_r(m))$ be a vector-valued function with each g_k is a real-valued function and has non-zero differential at $m = \mu$ i.e*

$$d_{ki} = \frac{\partial g_k}{\partial m_i}\Big|_{m=\mu} \neq 0, \tag{4.8}$$

The distribution of $\sqrt{n}\{g(m) - g(\mu)\}$ is asymptotically $N(0, D\Sigma D^T)$, where $D = (d_{ki})$, $k = 1, \ldots, r$, $i = 1, \ldots, l$.

The following corollary is a special case of Theorem 4.3.2 when $r = 1$, i.e g is a real-valued function.

Corollary 4.3.1. *Let $m = (m(h_1, j_1), m(h_2, j_2), \ldots, m(h_l, j_l))$ be a vector of sample moments, where $\sqrt{n}(m - \mu)$ is distributed asymptotically as $N(0, \Sigma)$, where Σ is shown in (4.7). Let $g(m)$ be a real-valued function that has non-zero differential at $m = \mu$. Then, $\sqrt{n}\{g(m) - g(\mu)\}$ is asymptotically $N(0, d^T\Sigma d)$ where*

$$d_i = \frac{\partial g}{\partial m_i}\Big|_{m=\mu} \neq 0.$$

4.4 The asymptotic distribution of the ICS estimates

Let y be an $n \times 2$ data matrix, with rows $y_i^T = (y_{i1}, y_{i2})$, $i = 1, \ldots, n$ Consider the ICS:kurtosis:variance criterion in (3.24). The sample covariance matrix is given by

$$S_y = \begin{pmatrix} m(2,0) & m(1,1) \\ m(1,1) & m(0,2) \end{pmatrix}. \tag{4.9}$$

The sample fourth-order moment matrix $\hat{K}_y$ is defined as

$$\hat{K}_y = \frac{1}{n}\sum_{i=1}^{n} y_i y_i^T (y_i^T S_y^{-1} y_i), \tag{4.10}$$

where

$$S_y^{-1} = \frac{1}{m(2,0)m(0,2) - m^2(1,1)} \begin{pmatrix} m(0,2) & -m(1,1) \\ -m(1,1) & m(2,0) \end{pmatrix}. \tag{4.11}$$

The components of $\hat{K}_y$ can be written in terms of the sample moments as follows

$$\hat{K}_y = \frac{1}{w} \begin{pmatrix} k_{11} & k_{12} \\ k_{12} & k_{22} \end{pmatrix}, \tag{4.12}$$

where

$$\begin{aligned} w &= (m(2,0)m(0,2) - m^2(1,1)), \\ k_{11} &= m(4,0)m(0,2) + m(2,2)m(2,0) - 2m(1,1)m(3,1), \\ k_{12} &= m(0,2)m(3,1) + m(2,0)m(1,3) - 2m(1,1)m(2,2), \\ k_{22} &= m(0,4)m(2,0) + m(2,2)m(0,2) - 2m(1,1)m(1,3). \end{aligned} \tag{4.13}$$

The population K_y is given by substituting the sample moments by the population moments from (4.1), (4.2) and (4.3), as follows

$$\begin{aligned} K_y &= \begin{pmatrix} \mu(4,0)\mu(0,2) + \mu(2,2)\mu(2,0) & 0 \\ 0 & \mu(0,4)\mu(2,0) + \mu(2,2)\mu(0,2) \end{pmatrix} \\ &= \begin{pmatrix} 4 - 2\delta^4 & 0 \\ 0 & 4 \end{pmatrix}. \end{aligned} \tag{4.14}$$

The population covariance matrix is the identity matrix, $\Sigma = I_2$.

In our model, the true group separation direction is at 0°. Therefore, we are

only interested in ϕ near zero, i.e $b = (\cos(\phi), \sin(\phi))$, can be written as, to second order

$$b = \left(1 - \phi^2/2, \phi\right)^T. \tag{4.15}$$

Substituting (4.11), (4.12) and (4.15) in (3.24) gives the ICS:kurtosis:variance criterion as follows

$$\hat{\kappa}_{ICS}(\phi) = \frac{1}{w(m(2,0) + 2\phi m(1,1) + \phi^2(m(0,2) - m(2,0)))}\{k_{11}+2\phi k_{12}+\phi^2(k_{22}-k_{11})\}, \tag{4.16}$$

where w, k_{11}, k_{12}, and k_{22} are in (4.13).

The criterion $\kappa_{ICS}(\phi)$ and its sample version $\hat{\kappa}_{ICS}(\phi)$ can be written as quadratic functions in ϕ as follows

$$\begin{aligned} \kappa_{\text{ICS}}(\phi) &= A_1 + B_1\phi + C_1\phi^2/2 + O(\phi^3), \\ \hat{\kappa}_{\text{ICS}}(\phi) &= \hat{A}_1 + \hat{B}_1\phi + \hat{C}_1\phi^2/2 + O(\phi^3). \end{aligned} \tag{4.17}$$

Let $\phi_\circ$ be the value of ϕ that minimizes $\kappa_{\text{ICS}}(\phi)$, and $\hat{\phi}_{\text{ICS}}$ be the value of ϕ that minimizes $\hat{\kappa}_{\text{ICS}}(\phi)$.

We know from our model that $\phi_\circ = 0°$, whereas $\hat{\phi}_{\text{ICS}}$ is given by

$$\begin{aligned} \hat{\phi}_{\text{ICS}} &= \frac{\partial \hat{\kappa}_{\text{ICS}}(\phi)}{\partial \phi} = 0 \\ &= -\frac{\hat{B}_1}{\hat{C}_1}. \end{aligned} \tag{4.18}$$

The quantities $\hat{B}_1$, and $\hat{C}_1$ are given by:

$$\hat{B}_1 = \frac{\partial \hat{\kappa}_{ICS}(\phi)}{\partial \phi}\Big|_{\phi=0} = \frac{2(b_{11} + b_{12} - 6b_{13} - 2b_{14} + 4b_{15})}{m^2(0,2)w}, \tag{4.19}$$

where w is given in (4.13), and

$$b_{11} = m(2,0)m(0,2)m(3,1), \quad b_{12} = m^2(2,0)m(1,3),$$
$$b_{13} = m(2,0)m(1,1)m(2,2), \quad b_{14} = m(1,1)m(0,2)m(4,0),$$
$$b_{15} = m^2(1,1)m(2,0).$$

$$\begin{aligned}\hat{C}_1 &= \frac{\partial^2 \hat{\kappa}_{ICS}(\phi)}{2\partial\phi^2}\Big|_{\phi=0} \\ &= \frac{2(c_{11} + 6c_{12} + 2c_{13} - 12c_{14} - 4c_{15} + 8c_{16} + c_{17})}{m^3(0,2)w},\end{aligned} \tag{4.20}$$

where w is given in (4.13), and

$$c_{11} = m^3(2,0)m(0,4), \quad c_{12} = m^2(2,0)m(1,1)m(1,3),$$
$$c_{13} = m(1,1)m(2,0)m(0,2)m(3,1), \quad c_{14} = m^2(1,1)m(2,0)m(2,2),$$
$$c_{15} = m^2(1,1)m(0,2)m(4,0), \quad c_{16} = m^3(1,1)m(3,1),$$
$$c_{17} = m(2,0)m^2(0,2)m(4,0).$$

The population values of $\hat{B}_1$ and $\hat{C}_1$ are given by

$$\begin{aligned}B_1 &= 0, \\ C_1 &= 4\delta^4.\end{aligned} \tag{4.21}$$

The following Lemma gives an approximation of $\hat{\phi}_{\text{ICS}}$, in (4.18).

Lemma 4.4.1. *If* $|\hat{B}_1 - B_1| = O_p(1/\sqrt{n})$ *and* $|\hat{C}_1 - C_|| = O_p(1/\sqrt{n})$*, then*

$$\hat{\phi}_{ICS} = -\frac{\hat{B}_1}{\hat{C}_1} \approx \frac{-\hat{B}_1}{C_1},$$

Proof. We first find the Taylor expansion of $\hat{\phi}_{ICS}$,

$$\begin{aligned}\hat{\phi}_{\text{ICS}} &= \phi_\circ + (\hat{B}_1 - B_1)\frac{\partial\hat{\phi}_{ICS}}{\partial\hat{B}_1}\Big|_{\hat{B}_1=B_1,\hat{C}_1=C_1} + (\hat{C}_1 - C_1)\frac{\partial\hat{\phi}_{ICS}}{\partial\hat{C}_1}\Big|_{\hat{B}_1=B_1,\hat{C}_1=C_1} + O_p(1/\sqrt{n}) \\ &= (\hat{B}_1 - B_1)\Big\{-\frac{1}{C_1}\Big\} + (\hat{C}_1 - C_1)\Big\{\frac{B_1}{C_1^2}\Big\} + O_p(1/\sqrt{n}) \\ &= -\frac{\hat{B}_1}{C_1} + O_p(1/\sqrt{n}).\end{aligned}$$

□

From Lemma 4.4.1,

$$\hat{\phi}_{\text{ICS}} \approx -\frac{\hat{B}_1}{C_1}. \tag{4.22}$$

The estimate $\sqrt{n}\hat{\phi}_{\text{ICS}}$ is asymptotically normal, with mean zero and variance equal to

$$V(\hat{\phi}_{\text{ICS}}) = \frac{\text{var}(\hat{B}_1)}{C_1^2}, \tag{4.23}$$

since $\hat{B}_1$ is normally distributed as shown in the following. The quantity $\hat{B}_1$ is a function of $(m(1,1), m(2,0), m(0,2), m(2,2), m(3,1), m(1,3), m(4,0))$. From Corollary 4.3.1, $\sqrt{n}\hat{B}_1$ is distributed asymptotically as $N(0, \text{var}(\hat{B}_1))$, where

$$\begin{aligned}\text{var}(\hat{B}_1) &= d^T\Sigma d, \\ &= \{48 - 72\delta^4 + 64\delta^6 - 16\delta^8\},\end{aligned} \tag{4.24}$$

where

$$d^T = (-12 + 4\delta^4, 0, 0, 0, 2, 2, 0),$$

and Σ is a 7×7 matrix given by

$$\Sigma = \begin{pmatrix} \sigma_{11} & 0 & 0 & 0 & \sigma_{15} & \sigma_{16} & 0 \\ 0 & \sigma_{22} & 0 & \sigma_{24} & 0 & 0 & \sigma_{27} \\ 0 & 0 & \sigma_{33} & \sigma_{34} & 0 & 0 & 0 \\ 0 & \sigma_{42} & \sigma_{43} & \sigma_{44} & 0 & 0 & \sigma_{47} \\ \sigma_{51} & 0 & 0 & 0 & \sigma_{55}\sigma_{56} & 0 & \\ \sigma_{61} & 0 & 0 & 0 & \sigma_{65} & \sigma_{66} & 0 \\ 0 & \sigma_{72} & 0 & \sigma_{74} & 0 & 0 & \sigma_{77}, \end{pmatrix}$$

where

$\sigma_{11} = 1,\ \sigma_{15} = 3 - 2\delta^4,\ \sigma_{16} = 3,$

$\sigma_{22} = 2 - 2\delta^4,\ \sigma_{24} = 2 - 2\delta^4,\ \sigma_{27} = 12 - 28\delta^4 + 16\delta^6,$

$\sigma_{33} = 2,\ \sigma_{34} = 2,$

$\sigma_{42} = 2 - 2\delta^4,\ \sigma_{43} = 2,\ \sigma_{44} = 8 - 6\delta^4,\ \sigma_{47} = 12 - 28\delta^4 + 16\delta^6,$

$\sigma_{51} = 3 - 2\delta^4,\ \sigma_{55} = 15 - 30\delta^4 + 16\delta^6,\ \sigma_{56} = 9 - 6\delta^4,$

$\sigma_{61} = 3,\ \sigma_{65} = 9 - 6\delta^4,\ \sigma_{66} = 15,$

$\sigma_{72} = 2 - 28\delta^4 + 16\delta^6,\ \sigma_{74} = 12 - 28\delta^4 + 16\delta^6,\ \sigma_{77} = 96 - 408\delta^4 + 448\delta^6 + 136\delta^8.$

Substituting (4.24) in (4.23) gives

$$V(\hat{\phi}_{ICS}) = \{\frac{6 - 9\delta^4 + 8\delta^6 - 2\delta^8}{2\delta^8}\}. \tag{4.25}$$

If $\delta = 0$, the data will be one normally distributed isotropic group. In this case, the estimates $\hat{\phi}_{ICS}$ will be uniformly distributed, and $V(\hat{\phi}_{ICS}) = \infty$. As $\delta \to 1$, $V(\hat{\phi}_{\text{ICS}})$ becomes smaller. If $\delta = 1$, $V(\hat{\theta}_{ICS}) = 1.5$, which is the smallest value that $V(\hat{\theta}_{ICS})$ takes.

4.5 The asymptotic distribution of PP:kurtosis:variance estimate

Consider the linear transformation Yb, where $b = (\cos(\phi), \sin(\phi))^T$,

Using b defined in (4.15), the linear transformation Yb can be written as

$$Yb = y_{i1}(1 - \phi^2/2) + y_{i2}\phi + O(\phi^3). \tag{4.26}$$

The univariate kurtosis of Ya as a function of ϕ is defined as

$$\hat{\kappa}_{\text{PP}}(\phi) = \frac{\frac{1}{n}\sum_{i=1}^{n}(y_{i1}(1-\phi^2) + y_{i2}\phi)^4}{\{\frac{1}{n}\sum_{i=1}^{n}(y_{i1}(1-\phi^2) + y_{i2}\phi)^2\}^2} - 3. \tag{4.27}$$

Equation (4.27) can be written in terms of moments as follows,

$$\hat{\kappa}_{\text{PP}}(\phi) = \frac{m(4,0) + 4\phi m(3,1) + \phi^2(6m(2,2) - 2m(4,0))}{m^2(2,0) + 4\phi m(2,0)m(1,1) + \phi^2(2m(2,0)m(0,2) - 2m^2(2,0) + 4m^2(1,1))} - 3. \tag{4.28}$$

As we did in (4.17), the true kurtosis $\kappa_{\text{PP}}(\phi)$ and sample kurtosis $\hat{\kappa}_{\text{PP}}(\phi)$ can be written as quadratic functions in ϕ

$$\begin{aligned} \kappa_{PP}(\phi) &= A_2 + B_2\phi + C_2\phi^2/2 + O(\phi^3), \\ \hat{\kappa}_{PP}(\phi) &= \hat{A}_2 + \hat{B}_2\phi + \hat{C}_2\phi^2/2 + O(\phi^3). \end{aligned} \tag{4.29}$$

Let $\phi_\circ$ be the value of ϕ that minimizes $\kappa_{PP}(\phi)$, and $\hat{\phi}_{PP}$ be the value of ϕ that minimizes $\hat{\kappa}_{PP}(\phi)$. We know that, $\phi_\circ = 0$, whereas $\hat{\phi}_{\text{PP}}$ is given by

$$\begin{aligned} \hat{\phi}_{\text{PP}} &= \frac{\partial \hat{\kappa}_{PP}(\phi)}{\partial \phi} = 0 \\ &= -\frac{\hat{B}_2}{\hat{C}_2}. \end{aligned} \tag{4.30}$$

The sample quantities $\hat{B}_2$ and $\hat{C}_2$ are the first and the second derivatives of $\hat{\kappa}_{PP}(\phi)$

at $\phi_\circ = 0$,

$$\begin{aligned}\hat{B}_2 &= \frac{\partial \hat{\kappa}_{PP}(\phi)}{\partial \phi}\Big|_{\phi_\circ=0} \\ &= \frac{4m(3,1)}{m^2(2,0)} - \frac{4m(4,0)m(1,1)}{m^3(2,0)}. \end{aligned} \tag{4.31}$$

$$\hat{C}_2 = \frac{\partial^2 \hat{\kappa}_{PP}(\phi)}{2\partial \phi^2}\Big|_{\phi_\circ=0} = c_{21} - c_{22} + c_{23} - c_{24}, \tag{4.32}$$

where

$$c_{21} = \frac{-4m(4,0) + 12m(2,2)}{m^2(2,0)},\ c_{22} = \frac{32m(3,1)m(1,1)}{m^3(2,0)},$$
$$c_{23} = \frac{32m(4,0)m^2(1,1)}{m^4(2,0)},\ c_{24} = \frac{m(4,0)(-4m^2(2,0) + 8m^2(1,1) + 4m(2,0)m(0,2))}{m^4(2,0)}.$$

The population quantities B_2 and C_2 are given by

$$\begin{aligned} B_2 &= 0 \\ C_2 &= \frac{-4\mu(4,0) + 12\mu(2,2)}{\mu^2(2,0)} - \frac{\mu(4,0)(-4\mu^2(2,0) + 4\mu(2,0)\mu(0,2))}{\mu^4(2,0)} \\ &= 8\delta^4. \end{aligned} \tag{4.33}$$

From Lemma 4.4.1, $\hat{\phi}_{\text{PP}}$ in (4.30) can be approximated by

$$\hat{\phi}_{\text{PP}} \approx -\frac{\hat{B}_2}{C_2}. \tag{4.34}$$

The estimate $\sqrt{n}\hat{\phi}_{\text{PP}}$ is asymptotically normal with mean 0 and variance

$$V(\hat{\phi}_{\text{PP}}) = \frac{\text{var}(\hat{B}_2)}{C_2^2}. \tag{4.35}$$

The quantity $\hat{B}_2$ is a function of moments $m = (m_{11}, m_{20}, m_{31}, m_{40})$. By Corollary

4.3.1, $\sqrt{n}\hat{B}_2$ is distributed asymptotically as $N(0, \text{var}(\hat{B}_2))$, where

$$\begin{aligned}\text{var}(\hat{B}_2) &= d^T \Sigma d \\ &= \{96 - 288\delta^4 + 256\delta^6 - 64\delta^8\}, \end{aligned} \tag{4.36}$$

and

$$\Sigma = \begin{pmatrix} 1 & 0 & 3-2\delta^4 & 0 \\ 0 & 2-2\delta^4 & 0 & 12-28\delta^4+16\delta^6 \\ 3-2\delta^4 & 0 & 15-30\delta^4+16\delta^6 & 0 \\ 0 & 12-28\delta^4+16\delta^6 & 0 & 96-480\delta^4+448\delta^6-136\delta^8 \end{pmatrix},$$

and d is a vector of length 4, with elements

$$d_i = \frac{\partial B_2}{\partial m_i}\Big|_{m=\mu},$$

$$\begin{aligned} d_1 &= \frac{-4\mu_{40}}{\mu_{20}^3} = -12 + 8\delta^4, \\ d_2 &= \frac{-8\mu_{31}\mu_{20} + 12\mu_{40}\mu_{11}}{\mu_{20}^4} = 0, \\ d_3 &= \frac{4}{\mu_{20}^2} = 4, \\ d_4 &= \frac{4\mu_{11}}{\mu_{20}^3} = 0. \end{aligned}$$

Substituting (4.36) in (4.35), gives the variance of $\hat{\phi}_{\text{PP}}$,

$$\begin{aligned} V(\hat{\phi}_{\text{PP}}) &= \frac{\text{var}\{\hat{B}_2\}}{4C_2^2} \\ &= \{\frac{3 - 9\delta^4 + 8\delta^6 - 2\delta^8}{2\delta^8}\}. \end{aligned} \tag{4.37}$$

Comparing (4.25) and (4.37), the variance $V(\hat{\phi}_{PP})$ is smaller than the variance of $V(\hat{\phi}_{ICS})$. If $\delta = 0$, $V(\hat{\phi}_{PP}) = \infty$, and as $\delta \to 1$, $V(\hat{\phi}_{\text{PP}})$ becomes smaller. If

$\delta = 1$, $V(\hat{\phi}_{PP}) = 0$. The reason for that is when $\delta = 1$, the data will be two parallel lines at ± 1. Hence, the projection in 0° will produce two points, while in all other directions $-\pi/2 < \phi < \pi/2$, the projection will be bimodal. Thus, PP:kurtosis:variance will always pick the 0° direction. Therefore, the variance of $\hat{\phi}_{\text{PP}}$ will be zero.

4.6 A comparison between $V(\hat{\phi}_{\mathbf{ICS}})$ and $V(\hat{\phi}_{\mathbf{PP}})$

We first compare the variance formulas $V(\hat{\phi}_{\text{ICS}})$ and $V(\hat{\phi}_{\text{PP}})$, defined in (4.24) and (4.37), with different values of δ, as shown in Table 4.1.

Let Y be an $n \times 2$ data matrix generated from the model explained in Section 4.2, with $\delta = 0.1, 0.5, 0.7, 0.9, 1$. The simulation is repeated $m = 1000$ times.

Applying ICS and PP gives the estimates: $\hat{\phi}^{(j)}_{\text{ICS}}$ and $\hat{\phi}^{(j)}_{\text{PP}}$, $j = 1, \ldots, 1000$. then find the sample variances of the ICS and PP estimates, used in Section 3.8.

Table 4.2 shows that the sample variances and the asymptotic variances of $\hat{\phi}^{(j)}_{\text{ICS}}$ and $\hat{\phi}^{(j)}_{\text{PP}}$, defined in (4.24) and (4.37), match very well, except for $\delta = 0.5$. Also for $\delta = 0.7$, sample and asymptotic variances of $\hat{\phi}_{\text{PP}}$ differ slightly. The reason for these discrepancies is that for small values of δ, ICS and PP methods do not estimate the group separation precisely. Hence, assumption (4.15) is not satisfied adequately.

Table 4.1: The values of asymptotic variances of $\hat{\phi}_{\text{ICS}}$ and $\hat{\phi}_{\text{PP}}$, for $\delta = 0.1, 0.5, 0.7, 0.9, 1$.

δ	$V(\hat{\phi}_{PP})$	$V(\hat{\phi}_{ICS})$
0.1	$1.49e+08/n$	$2.994e+08/n$
0.5	$327/n$	$711/n$
0.7	$14.44/n$	$40.46/n$
0.9	$0.564/n$	$4.049/n$
1	0	$1.5/n$

Table 4.2: The sample variances and asymptotic variances of $\hat{\phi}_{\text{ICS}}$ and $\hat{\phi}_{\text{PP}}$, for different δ and n

δ	n	$V(\hat{\phi}_{\text{ICS}})$	$\hat{V}(\hat{\phi}_{\text{ICS}})$	$V(\hat{\phi}_{\text{PP}})$	$V(\hat{\phi}_{\text{PP}})$
0.5	200	8.47	0.4018	4.01	0.2852
	500	3.39	0.3577	1.6	0.2553
	1000	1.69	0.2962	0.8	0.209
0.7	200	0.202	0.1712	0.0722	0.112
	500	0.0809	0.0796	0.0289	0.0478
	1000	0.0404	0.0383	0.0144	0.0161
0.9	200	0.0202	0.0183	0.0028	0.0034
	500	0.0081	0.0082	0.0011	0.0013
	1000	0.004	0.0042	0.0006	0.0006
1	200	0.0075	0.0076	0	2.52e-04
	500	0.003	0.0033	0	1.05e-04
	1000	0.0015	0.0018	0	4.94e-05

Chapter 5

Robust ICS and PP

5.1 Introduction

In Chapter 3, we have explored how ICS:kurtosis:variance and PP:kurtosis:variance work to identify group separation direction under two-group mixtures of bivariate normal distributions. ICS and PP find the direction that has an extreme kurtosis value.

However, ICS:variance:kurtosis and PP:variance:kurtosis are not robust, in the sense that they are highly affected by outliers. That is, a single outlier can make the ICS and PP kurtosis criteria extremely large.

The ICS and PP criteria defined in (2.1) and (2.8) can be defined with respect to robust measures of spread. By convention, the measure of spread in the denominator is more robust than the one in the numerator.

In this chapter, we investigate the feasibility of robust ICS and robust PP in identifying group separation direction under equal mixtures of two bivariate normal distributions.

In practice some of the robust ICS and PP do not work well in identifying the group separation direction. Sometimes the separation between the minimum and maximum values of the robust ICS and PP criteria is small, and due to sampling variation. Also, sometimes the robust PP criterion have a local maximum when

it is expected to be a global minimum.

One problem with robust ICS and PP is that different location measures, associated with the pair of spread measures, are used in the numerator and denominator.

Using a common location measure in the numerator and denominator sometimes solves this problem. Alternatively, the measures of spread can be computed based on the pairwise differenced data, as explained briefly in Sections 2.5 and 2.6.

The structure of this chapter is given as follows. In Section 5.2, we explore the behaviour of the robust ICS and robust PP methods through simulations. In Section 5.3, we analyse the problems arising as a result of using robust ICS and PP criteria. In Section 5.4, we discuss the use of a common location measure to improve the performance of robust ICS and robust PP. In Section 5.5, we discuss computing robust measures of spread based on pairwise differencing of the data.

5.2 The behavior of robust ICS and PP in sample case

Robust ICS and PP criteria are defined based on combinations of some of the estimates listed in Section 2.8. In the following, we list the pairs of scatter matrices used in robust ICS methods, and their univariate analogues used in robust PP methods,

- Robust ICS:

  (1) ICS:variance:t_2M-estimate, ICS based on the covariance matrix, $(S, \bar{x})$, and M-estimate for t_2, $(S_t, \bar{x}_t)$.

  (2) ICS:variance:mve, ICS based on the covariance matrix, $(S, \bar{x})$, and the minimum volume ellipsoid, $(S_m, \bar{x}_m)$.

(3) ICS:t_2M-estimate:mve, ICS based on M-estimate for t_2,$(S_t, \bar{x}_t)$, and the minimum volume ellipsoid, $(S_m, \bar{x}_m)$.

- Robust PP:

(1) PP:variance:t_2M-estimate, PP based on the variance, $(s^2, \bar{x})$, and univariate M-estimate for t_2, $(s_t, \bar{x}_t)$.

(2) PP:variance:lshorth, PP based on the variance, $(s^2, \bar{x})$, and the lshorth, (l, m_l).

(3) PP:t_2M-estimate:lshorth, PP based on the univariate M-estimate $(s_t, \bar{x}_t)$, and the lshorth, (l, m_l).

Let Y, be an $n \times 2$ data matrix with $n = 500$, where its rows are $y_i^T = (y_{i1}, y_{i2})$, $i = 1, \ldots, n$, generated from the mixture model (3.1), with mixing proportion $q = 1/2$. Without loss of generality, assume that Y is standardized with respect to the total covariance matrix, as in (3.9), such that $S_y = I_2$.

To investigate the behaviour of robust ICS and PP criteria, we first plot the criteria of robust ICS and robust PP, $\kappa_{\text{ICS}}(\theta)$ and $\kappa_{\text{PP}}(\theta)$, $-\pi/2 \leq \theta \leq \pi/2$. The plots are shown in Figure 5.1.

There are three different criteria for each robust ICS and robust PP. Each criteria will be plotted for data sets with $\delta = 0.7$, and $\delta = 0.9$.

The following `R` packages and functions are used to compute the robust estimates:

- The function `tM`, from the package `ICS`, Nordhausen et al. (2008), is used to compute multivariate and univariate M-estimate for t_2 .

- The function `rob.cov`, from the package `MASS`, Venables and Ripley (2010), is used to compute the MVE.

- The function `lshorth`, from the package `lshorth`, Einmahl et al. (2010), is used to compute the lshorth.

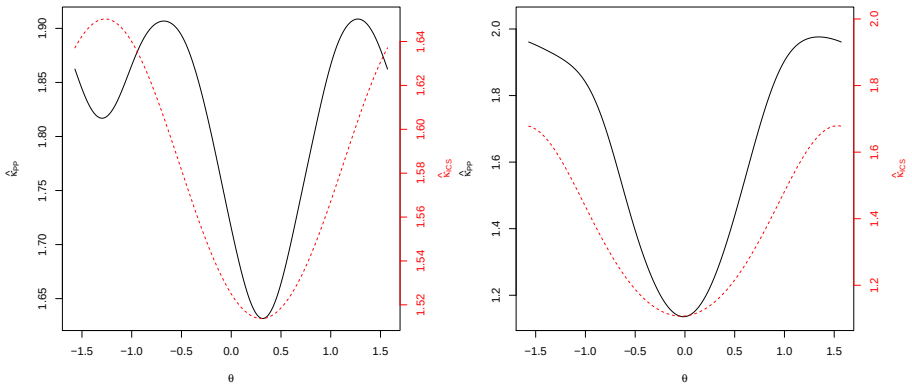
From Figure 5.1, we find the following:

(1) From Figures (a) and (b), ICS:variance:t_2M-estimate and PP:variance:t_2M-estimate seem to work well, and their behaviours are similar.

(2) From Figures (c) and (d), ICS:variance:mve also seems to work well, whereas PP:variance:lshorth does not. PP:variance:lshorth has two problems. The first problem is the plot of PP:t_2M-estimate:lshorth criterion vs θ is not smooth as a function of θ. The second problem is the curve of PP:t_2M-estimate:lshorth sometimes produces a local maximum when it is supposed to be a local minimum, in the direction of the true separation direction, as shown in Figure 5.1 (d).

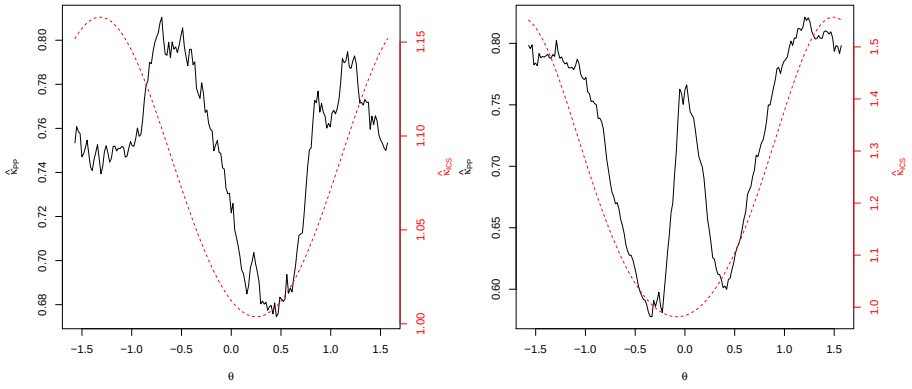
(3) From Figures (e) and (f), the criterion of ICS:t_2M-estimate:mve is minimised near 0°. However, since all the eigenvalues of $S_t^{-1}S_{mve}$ are approximately equal, ICS:t_2M-estimate:mve is not working reliably. That is, the separation between the smallest and largest eigenvalues is due to sampling variation. From Figure 5.1 (f), the smallest eigenvalue is approximately 0.88, and the largest eigenvalue is approximately 0.98. PP:t_2M-estimate:lshorth criterion is maximized in the 0° direction when it is supposed to be minimised.

ICS:variance:t_2M-estimate and PP:variance:t_2M-estimate



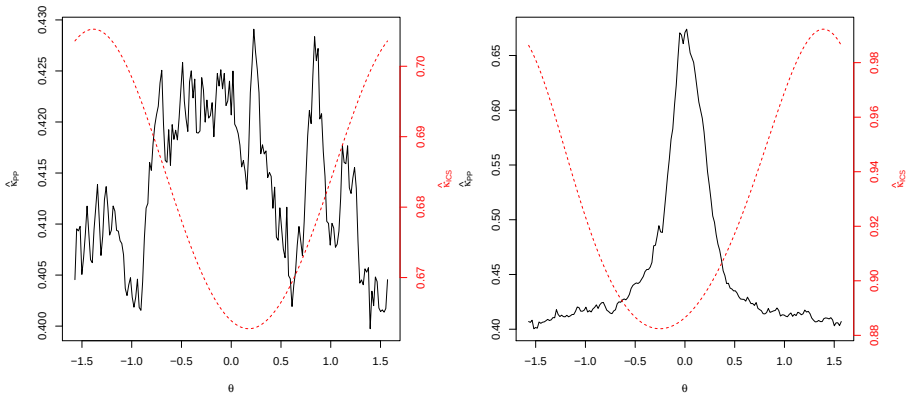
(a) $\delta = 0.7$ (b) $\delta = 0.9$

ICS:variance:MVE and PP:variance:lshorth



(c) $\delta = 0.7$ (d) $\delta = 0.9$

ICS:t_2M-estimate:MVE and PP:t_2M-estimate:lshorth



(e) $\delta = 0.7$ (f) $\delta = 0.9$

Figure 5.1: Plot of ICS criteria $\hat{\kappa}_{\text{ICS}}(\theta)$ (red dotted line), and PP criteria $\hat{\kappa}_{\text{PP}}(\theta)$ (solid black line) versus θ, for $q = 1/2$, $\delta = 0.7$, and 0.9.

One problem with robust ICS and PP methods is that different location measures are used in the denominator and numerator. Moreover, in robust PP criteria, these location measures change unevenly as θ changes.

Two possible ways to solve this problem; one is by forcing a common location measure in the denominator and numerator, another way is to compute the robust measures based on the pairwise differences of the data to force the symmetry of the data around the origin.

Another problem, is that two robust measures of spread can be approximately equal, under our model. Hence, methods based on two robust measures of spread sometimes do not work reliably.

In the following sections, we discuss the two aforementioned problems, and the possible solutions.

5.3 Analysis of the problems arising in robust ICS and PP

5.3.1 PP:variance:lshorth

We analyze the behavior of PP:variance:lshorth by plotting histograms of the following projections: 0°, 15°, 30°, and 90°, for the same data set plotted in Figure 5.1 (d), i.e for $\delta = 0.9$.

The shape of the histograms depend on of the projection directions as shown in the following:

(a) The 0° projection produces two widely separated groups with one group being slightly bigger than the other.

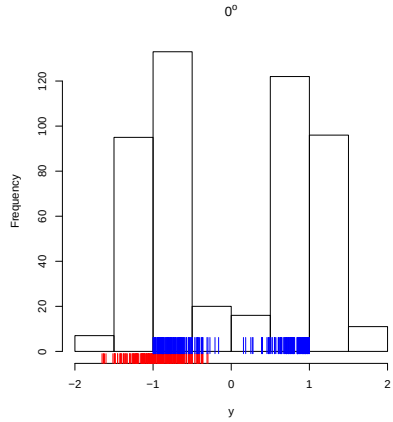
(b) The 15° projection produces two slightly separated groups with within-group variance is larger than in the 0° projection.

(c) The 30° projection produces one group, with a pseudo-uniform distribution.

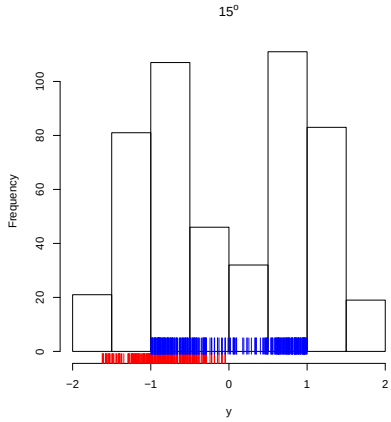
(d) The 90° projection produces one normally distributed group.

For each histogram, we plot the shorth interval and the interval $\bar{x} \pm s$, which is approximately equals to $(-1, 1)$ in all projections, since all projections have zero mean and unit variance.

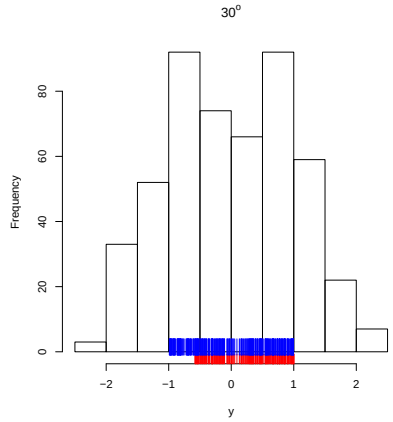
Figure 5.2 shows the changes in the shorth interval, compared to the interval $(-1, 1)$. In Figure 5.2 (a), the shorth interval is located at the bigger group, in which the lshorth will take its smallest value. In Figures (b), and (c), the shorth interval becomes larger as the within-group variance increases. In Figure (d), the shorth becomes small again. In other words, the lshorth takes small value in bimodal projections and unimodal projections, which produces the local maximum at 0° direction.



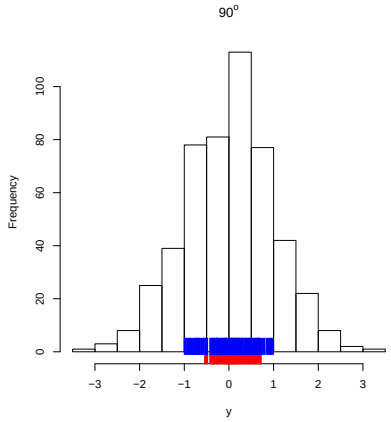
(a) shorth= $(-1.653, -0.298)$, $\bar{x} \pm s = (-1, 1)$



(b) shorth= $(-1.62, -0.04)$, $\bar{x} \pm s = (-1, 1)$



(c) shorth= $(-0.58, 1)$, $\bar{x} \pm s = (-1, 1)$



(d) shorth= $(-0.543, 0.719)$, $\bar{x} \pm s = (-1, 1)$

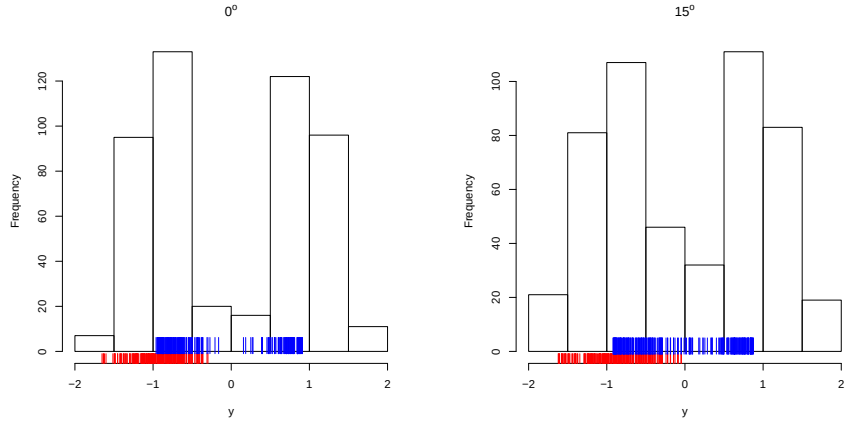
Figure 5.2: Histograms of $0°, 15°, 30°$ and $90°$ projections, with the vectors of data contained in the shorh interval (the lower red lines), and in $\bar{x} \pm s$ interval the (upper blue lines).

5.3.2 PP:t_2M-estimate:lshorth

We follow the same analysis done in analyzing PP:variance:lshorth, for the same data set. For each histogram we plot the shorth interval and $\bar{x}_t \pm s_t$, as shown in Figure 5.3.

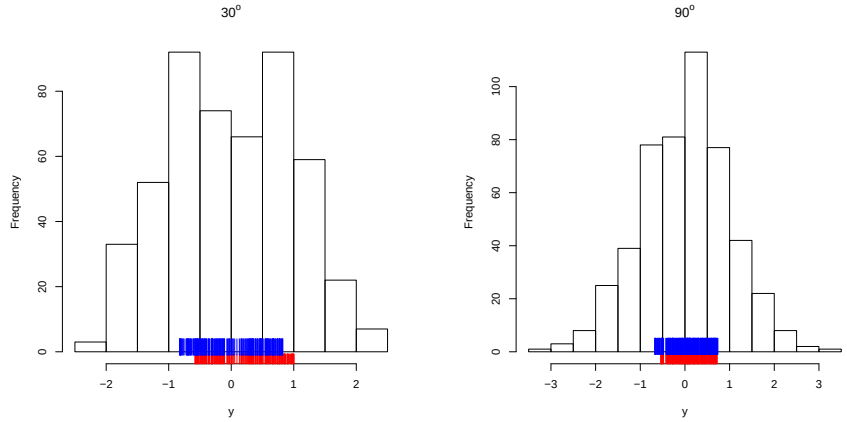
In Figure 5.3 (a) and (b), the $\bar{x}_t \pm s_t$ interval is located at the origin, while the shorth interval is located at one group.

In Figures (c) and (d), the shorth and $\bar{x}_t \pm s_t$ intervals are located in the same place. We can also see from (b), (c) and (d) that the lengths of the two intervals



(a) shorth= $(-1.653, -0.298)$, $\bar{x}_t \pm s_t = (-0.96, 0.91)$

(b) shorth= $(-1.62, -0.04)$, $\bar{x}_t \pm s_t = (-0.92, 0.88)$



(c) shorth= $(-0.58, 1)$, $\bar{x}_t \pm s_t = (-0.828, 0.826)$

(d) shorth= $(-0.543, 0.719)$, $\bar{x}_t \pm s_t = (-0.678, 0.75)$

Figure 5.3: Histograms of $0°, 15°, 30°$ and $90°$ projections, with shoth interval (the lower red lines), and $\bar{x}_t \pm s_t$ interval the (upper blue lines).

are

- In Figure (b), the lshorth= 1.66, while the length of $\bar{x}_t \pm s_t$ is equal to 1.8. The ratio $s_t/l \approx 0.49$.
- In Figure (c), the lshorth= 1.58, while the length of $\bar{x}_t \pm s_t$ is equal to 1.654. The ratio $s_t/l \approx 0.43$.
- In Figure (d), the shorth= 1.26, while the length of $\bar{x}_t \pm s_t$ is equal to 1.428. The ratio $s_t/l \approx 0.405$.

5.3.3 ICS:t_2M-estimate:MVE

The plot of ICS:t_2M-estimate:mve, in Figure 5.1 (f), shows that the smallest and the largest eigenvalues of $S_m^{-1}S_t$ are not widely separated. The M-estimate for t_2, S_t, and the MVE, S_m, scatter matrices, and their associated location vectors, μ_t and μ_m, respectively, for the data in Figure 5.1 (f) are given by

$$S_t = \begin{pmatrix} 0.903 & -0.012 \\ -0.012 & 0.6 \end{pmatrix}, \quad \bar{x}_t = (-0.001, 0.03)^T$$
$$S_m = \begin{pmatrix} 1.016 & -0.025 \\ -0.025 & 0.615 \end{pmatrix}, \quad \bar{x}_m = (0.005, 0.107)^T.$$

The eigenvalues and eigenvectors of $S_m^{-1}S_t$ are equal to

$$\begin{aligned} l_1 = 0.973,\ l_2 = 0.89, \\ u_1 = (-0.155, -0.99)^T,\ u_2 = (-0.98, 0.214)^T. \end{aligned} \tag{5.1}$$

5.4 Using common location measures

5.4.1 PP(Mean):variance:lshorth

The lshorth can be computed based on the sample mean, $\bar{x} \approx 0$. To find out the effect of forcing a common location measure in PP(Mean):var:lshorth, we reproduce Figure 5.1 (d), with the lshorth computed around the origin.

Figure 5.4 suggests that forcing a common location measure in the denominator and numerator fixes the problem in PP:var:lshorth.

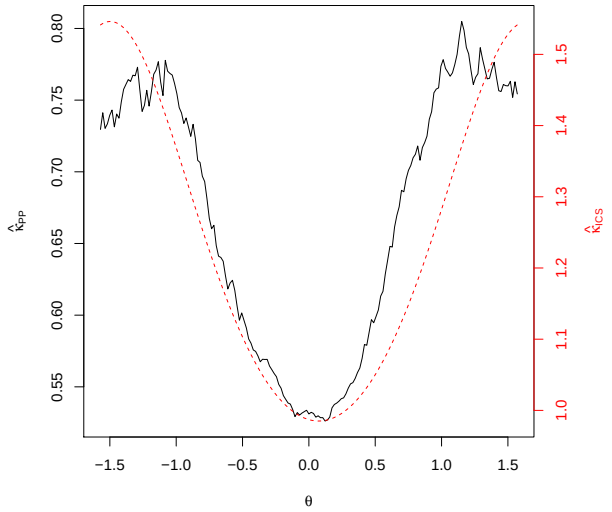


Figure 5.4: The plot of ICS:variance:mve (red dashed curve) and PP(Mean):variance:lshorth (black solid curve) using a common mean $\bar{x}$.

5.4.2 PP(lshorth):t_2M-estimate:lshorth and ICS(MVE):t_2M-estimate:MVE

In PP(lshorth):t_2M-estimate:lshorth, both measures are computed based on the lshorth location measure. Similarly, in ICS(MVE):t_2M-estimate:MVE, both scatter matrices computed based on the MVE location measure.

Figure 5.5 shows plots of the two methods criteria. Comparing the plots in Figure 5.5, using a common location measure does not have an apparent effect.

Computing the M-estimate scatter matrix S_t based on $\bar{x}_m$ does not change the eigenvalues and and eigenvectors in (5.1), since $\bar{x}_t$ and $\bar{x}_m$ are close to each other.

5.5 Using pairwise differencing

5.5.1 PP^d:variance:lshorth and ICS^d:variance:MVE

Figure 5.6 shows that pairwise differencing of the data can be useful in PP:variance:lshorth, but not useful for ICS:variance:MVE, since it can reduce the spread between the

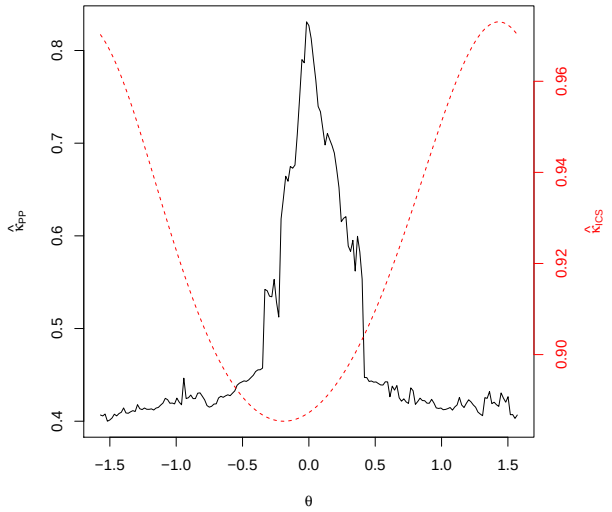


Figure 5.5: The plot of ICS(MVE):t_2M-estimate:MVE and PP(lshorth):t_2M-estimate:lshorth.

eigenvalues of S_m^{-1}.

5.5.2 PP^d:t_2M-estimate:lshorth and ICS^d:t_2M-estimate:MVE

Pairwise differencing has not worked in PP:t_2M-estimate:lshorth, or in ICS:variance:MVE. Figure 5.7 shows that the plot of PP:t_2M-estimate:lshorth criterion (the black solid curve) has local minimums wich complicates the optimization procedure; the spread between the eigenvalues is small in the plot of ICS:variance:MVE criterion (the red dashed curve).

5.6 Conclusion

When applying ICS and PP based on robust measures of spread, different location measures are associated with the pair of spread measures.

Without a common location measure, several methods behave strangely, especially ICS:t_2M-estimate:MVE, PP:t_2M-estimate:lshorth, and PP:variance:lshorth. The problem can be seen clearly for PP.

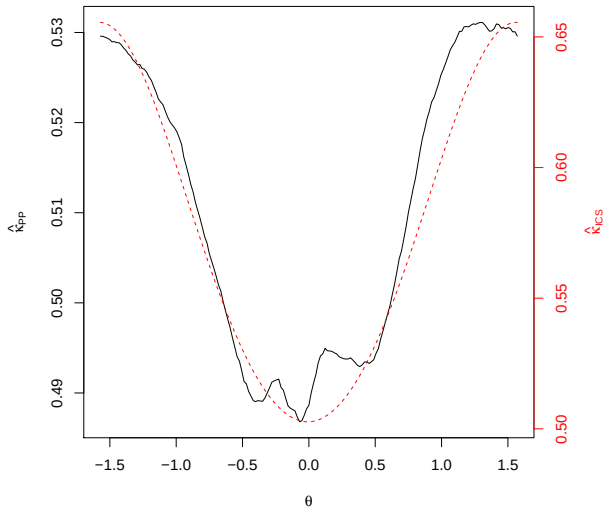


Figure 5.6: The plots of PP^d:variance:lshorth (the black solid curve), and ICS^d:variance:MVE (the red dashed curve).

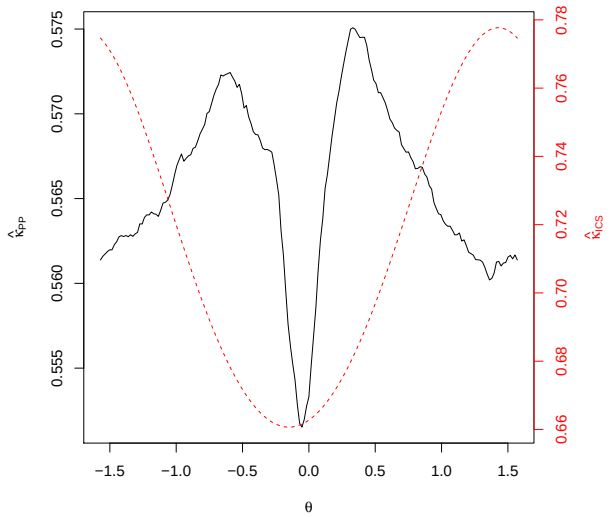


Figure 5.7: The plots of PPd:t_2M-estimate:lshorth (the black solid curve) and ICS^d:t_2M-estimate:MVE (the red dashed curve).

Using a common location measure can solve the problem in PP:variance:lshorth. Preliminary investigation using alternative robust estimators suggests that when using a common location measure (especially the mean) ICS and PP behave better.

Chapter 6

ICS in the errors in variables model

6.1 Introduction

Suppose we have data made on two measurements, and we want to fit a line that represents the linear relationship between the two measurements.

One way to fit a line is by using the ordinary least squares regression (OLS). In the classical regression settings, one of the measurements is assumed to be associated with errors, whereas the other is made with no error. The OLS criterion is to fit a line that minimizes the vertical or horizontal squared distances from the data points.

In practice, both measurements are subject to error. This problem has long been studied as the errors in variables model (EIV). Madansky (1959) and Gillard (2010) provide detailed reviews of the EIV model.

If the data are normally distributed, second moments are sufficient statistics. The case of normally distributed data has been discussed in Kendall and Stuart (1979), Sprent (1969).

If the data are non-normally distributed, higher order moments, up to fourth-order, can be used. The first paper discussed using high order moments was by

Geary (1941).

The goal of this chapter is to explore the use of ICS:kurtosis:variance, studied in Chapter 3, in the EIV model. We also compare the accuracy of Geary's fourth-order moment-based EIV estimators and the ICS:kurtosis:variance.

The structure of this chapter is as follows. In Section 6.2, the EIV model is defined. In Section 6.3, the classical theory of EIV is reviewed when the data are normally distributed. In Section 6.4, we review the use of high-order moments, up to fourth order, when the data are non-normally distributed. In Section 6.5, we explore the use of ICS:kurtosis:variance in EIV. In Section 6.6, we compare the accuracy of the ICS:kurtosis:variance and a selection of Geary's fourth-order moment-based estimates.

6.2 The errors in variables model

Let ζ_1 and ζ_2 be two random variables, with mean 0, that have an exact linear relationship given as

$$\zeta_2 = \beta\zeta_1. \tag{6.1}$$

Suppose that ζ_1 and ζ_2 are not observed, instead we observe the random vector $z = (z_1, z_2)^T$, where z_1 and z_2 are given by

$$\begin{pmatrix} z_1 \\ z_2 \end{pmatrix} = \begin{pmatrix} \mu_1 \\ \mu_2 \end{pmatrix} + \begin{pmatrix} \zeta_1 \\ \zeta_2 \end{pmatrix} + \begin{pmatrix} \epsilon_1 \\ \epsilon_2 \end{pmatrix}. \tag{6.2}$$

where $\mu = (\mu_1, \mu_2)^T$ is the mean vector of $z = (z_1, z_2)^T$, and $\epsilon = (\epsilon_1, \epsilon_2)^T$ is a bivariate random vector, distributed as a normal distribution with mean zero and covariance matrix Σ_ϵ, independent from ζ_1 and ζ_2.

From (6.1), (6.2) can be written as

$$\begin{pmatrix} z_1 \\ z_2 \end{pmatrix} = \begin{pmatrix} \mu_1 \\ \mu_2 \end{pmatrix} + \begin{pmatrix} 1 \\ \beta \end{pmatrix} \zeta_1 + \begin{pmatrix} \epsilon_1 \\ \epsilon_2 \end{pmatrix}. \tag{6.3}$$

where the slope $\beta = \tan(\psi)$, $-\pi/2 < \psi < \pi/2$.

Alternatively, the linear transformation $(1, \beta)^T$ can be expressed in terms of angles as follows

$$\begin{pmatrix} \zeta_1 \\ \zeta_2 \end{pmatrix} = \begin{pmatrix} \cos(\psi) \\ \sin(\psi) \end{pmatrix} \zeta_\circ. \tag{6.4}$$

Equation (6.2) becomes

$$z = \mu + a\zeta_\circ + \epsilon, \tag{6.5}$$

where $a = (\cos(\psi), \sin(\psi))$. The variable $\zeta_\circ$ is called the signal.

The EIV model assumptions are summarized as follows

$$\begin{aligned} \mathrm{E}(\epsilon_1) &= \mathrm{E}(\epsilon_2) = 0 \\ \Sigma_\epsilon &= \begin{pmatrix} \sigma^2_{\epsilon_1} & \sigma_{\epsilon_1\epsilon_2} \\ \sigma_{\epsilon_1\epsilon_2} & \sigma^2_{\epsilon_2} \end{pmatrix}, \\ \mathrm{E}(\zeta_\circ) &= 0 \\ \mathrm{var}(\zeta_\circ) &= \sigma^2_{\zeta_\circ} \\ \Sigma_z &= aa^T\sigma^2_{\zeta_\circ} + \Sigma_\epsilon \\ &= \begin{pmatrix} \cos^2(\psi)\sigma^2_{\zeta_\circ} + \sigma^2_{\epsilon_1} & \cos(\psi)\sin(\psi)\sigma^2_{\zeta_\circ} + \sigma_{\epsilon_1\epsilon_2} \\ \cos(\psi)\sin(\psi)\sigma^2_{\zeta_\circ} + \sigma_{\epsilon_1\epsilon_2} & \sin^2(\psi)\sigma^2_{\zeta_\circ} + \sigma^2_{\epsilon_2} \end{pmatrix}. \end{aligned} \tag{6.6}$$

The EIV model is equivariant under shifting, and affine transformations. Without loss of generality, assume that z is standardized as follows

$$A(z - \mu), \tag{6.7}$$

where A is a 2×2 non-singular matrix. The standardization in (6.7) can be done in two different ways: with respect to the error covariance matrix Σ_ϵ, or with respect to the total covariance matrix Σ_z, as shown in the following.

(1) Standardization with respect to $A \propto \Sigma_\epsilon^{-1/2}$ is given by

$$\begin{aligned} x &= A \begin{pmatrix} \cos(\psi) \\ \sin(\psi) \end{pmatrix} \zeta_\circ + A \begin{pmatrix} \epsilon_1 \\ \epsilon_2 \end{pmatrix}. \\ &= \gamma \zeta_\circ + \epsilon^*, \end{aligned} \tag{6.8}$$

where

$$\gamma = A \begin{pmatrix} \cos(\psi) \\ \sin(\psi) \end{pmatrix} = \begin{pmatrix} \cos(\theta) \\ \sin(\theta) \end{pmatrix}, \text{ say, and } \epsilon^* = A\epsilon,$$

and $\epsilon^* \sim N(0, cI_2)$, for some constant $c > 0$. The covariance matrix of x is given by

$$\begin{aligned} \Sigma_x &= \gamma\gamma^T \sigma^2_{\zeta_\circ} + cI_2 \\ &= \begin{pmatrix} \cos^2(\theta) & \cos(\theta)\sin(\theta) \\ \cos(\theta)\sin(\theta) & \sin^2(\theta) \end{pmatrix} \sigma^2_{\zeta_\circ} + \begin{pmatrix} c & 0 \\ 0 & c \end{pmatrix}. \end{aligned} \tag{6.9}$$

(2) Standardization with respect to $A = \Sigma_z^{-1/2}$ is given by

$$\begin{aligned} y &= A \begin{pmatrix} \cos(\psi) \\ \sin(\psi) \end{pmatrix} \zeta_\circ + A \begin{pmatrix} \epsilon_1 \\ \epsilon_2 \end{pmatrix}. \\ &= \nu \zeta_\circ + \epsilon^{'}, \end{aligned} \tag{6.10}$$

where

$$y = \Sigma_z^{-1/2} z, \quad \nu = \Sigma_z^{-1/2} \begin{pmatrix} \cos(\psi) \\ \sin(\psi) \end{pmatrix} = \begin{pmatrix} \cos(\phi) \\ \sin(\phi) \end{pmatrix}, \text{ say, and } \epsilon^{'} = \Sigma_z^{-1/2}\epsilon.$$

and $\epsilon^{'} \sim N(0, \Sigma_{\epsilon^{'}})$. The covariance matrix of y is $\Sigma_y = I_2$.

The goal is to estimate the signal direction from replications of the observed vector. The estimation procedure depends on the distribution of the signal $\zeta_\circ$. We distinguish two different cases for the signal $\zeta_\circ$: (1) $\zeta_\circ$ has a normal distribution; (2) $\zeta_\circ$ has a non-normal distribution.

6.3 Normal signals

If $\zeta_\circ$ is normally distributed, as a result x will be normally distributed. In this case, second moments are sufficient statistics for the unknown parameters. We discuss the theory of EIV in two different cases: the error variances are unknown, Section 6.3.1, and the error variances are known in Section 6.3.2. The two cases discussed under both assumptions correlated and uncorrelated error variances.

6.3.1 Unknown error variances

Let $z_1, \ldots, z_n$ be a sample from the bivariate normal distribution, where $z_i^T = (z_{i1}, z_{i2})^T$, $i = 1, \ldots, n$. Without loss of generality, assume that z is shifted such that it has 0 mean vector. The covariance matrix Σ_z is defined as in (6.6).

In order to estimate the signal direction ψ, second sample moments can be used as follows

$$
\begin{aligned}
s^2_{z_1} &= \cos^2(\hat{\psi})\hat{\sigma}^2_{\zeta_\circ} + \hat{\sigma}^2_{\epsilon_1}, && \text{(a)} \\
s^2_{z_2} &= \sin^2(\hat{\psi})\hat{\sigma}^2_{\zeta_\circ} + \hat{\sigma}^2_{\epsilon_2}, && \text{(b)} \\
s_{z_1 z_2} &= \cos(\hat{\psi})\sin(\hat{\psi})\hat{\sigma}^2_{\zeta_\circ} + \hat{\sigma}_{\epsilon_1\epsilon_2}. && \text{(c)}
\end{aligned} \tag{6.11}
$$

Then we can solve the equations for $\hat{\psi}$. But, there are three equations and five unknowns: $\hat{\psi}$, $\hat{\sigma}^2_{\zeta_\circ}$, $\hat{\sigma}^2_{\epsilon_1}$, $\hat{\sigma}^2_{\epsilon_2}$, $\hat{\sigma}_{\epsilon_1\epsilon_2}$. Thus, the model is unidentifiable.

Although in the classical EIV theory, e.g. Sprent (1969), Kendall and Stuart (1979), it is assumed that the errors are uncorrelated, i.e $\sigma_{\epsilon_1,\epsilon_2} = 0$, the model

will remain unidentifiable under this assumption with three equations and four unknowns. However, assuming that $\sigma_{\epsilon_1,\epsilon_2} = 0$ helps to set boundaries of $\hat{\psi}$, Kendall and Stuart (1979), as follows:

(1) From (c), since $\hat{\sigma}^2_{\zeta_\circ} \geq 0$, the sign of $\sin(\hat{\psi})\cos(\hat{\psi})$ depends on the sign of $s_{z_1 z_2}$.

(2) From (a), since $\hat{\sigma}^2_{\epsilon_1} \geq 0$, then

$$s^2_{z_1} \geq \cos^2(\hat{\psi})\hat{\sigma}^2_{\zeta_\circ}. \tag{6.12}$$

Multiplying both sides of (6.12) by $|\sin(\hat{\psi})/\cos(\hat{\psi})|$, and compare it to (c), assuming $\sigma^2_{\epsilon_1\epsilon_2} = 0$, gives

$$\begin{aligned} \Big|\frac{\sin(\hat{\psi})}{\cos(\hat{\psi})}\Big| s^2_{z_1} &\geq |\sin(\hat{\psi})|\cos(\hat{\psi})\sigma^2_{\zeta_\circ} \\ \Big|\frac{\sin(\hat{\psi})}{\cos(\hat{\psi})}\Big| s^2_{z_1} &\geq |s_{z_1 z_2}|. \end{aligned} \tag{6.13}$$

(3) Similarly, from (b), since $\hat{\sigma}^2_{\epsilon_2} \geq 0$, then

$$s^2_{z_2} \geq \sin^2(\hat{\psi})\hat{\sigma}^2_{\zeta_\circ}. \tag{6.14}$$

Multiplying both sides of (6.14) by $|\cos(\hat{\psi})/\sin(\hat{\psi})|$ and compare it to (c), gives

$$\begin{aligned} \Big|\frac{\cos(\hat{\psi})}{\sin(\hat{\psi})}\Big| s^2_{z_2} &\geq |\sin(\hat{\psi})|\cos(\hat{\psi})\sigma^2_{\zeta_\circ} \\ \Big|\frac{\cos(\hat{\psi})}{\sin(\hat{\psi})}\Big| s^2_{z_2} &\geq |s_{z_1 z_2}|. \end{aligned} \tag{6.15}$$

From (6.13) and (6.15)

$$\frac{|s_{z_1 z_2}|}{s^2_{z_1}} \leq \Big|\frac{\sin(\hat{\psi})}{\cos(\hat{\psi})}\Big| \leq \frac{s^2_{z_2}}{|s_{z_1 z_2}|}$$

$$\text{atan}\Big\{\frac{|s_{z_1z_2}|}{s^2_{z_1}}\Big\} \leq |\hat{\psi}| \leq \text{atan}\Big\{\frac{s^2_{z_2}}{|s_{z_1z_2}|}\Big\}. \tag{6.16}$$

where $|\psi| \in (0, \pi/2)$.

Equation (6.16) means that, assuming $\sigma_{\epsilon_1,\epsilon_2} = 0$, the EIV fitted line lies between the OLS line of regressing z_2 on z_1, and the one of regressing z_1 on z_2.

Consider the case of points which lie perfectly on a horizontal line. In this case, $\hat{\psi} = 0°$. Suppose that we added some noise in the perpendicular direction. As long as the variance of the added noise in the vertical direction is less than the variance of the points in the horizontal direction, $\hat{\psi}$ will be near 0°. As the variance of the noise is increased vertically, $\hat{\psi}$ will be more fluctuated between 0 and $\pi/2$. In the case of the vertical variance of the noise is equal to the horizontal noise, the data will be isotropic, $s_{z_1z_2} = 0$.

6.3.2 Known error variances

We have shown in the previous section that when the error variances are unknown, ψ cannot be estimated. Hence, we need an additional knowledge about the error variances.

Kendall and Stuart (1979) discussed the following four cases, assuming that $\sigma_{\epsilon_1,\epsilon_2} = 0$,

(i) If $\sigma^2_{\epsilon_1}$ is known, the estimate of ψ is found by multiplying equation (c) in (6.11) by $\sin(\hat{\psi})/\cos(\hat{\psi})$, then substituting by (e). The estimate of ψ is

$$\hat{\psi} = \text{atan}\Big\{\frac{s_{z_1z_2}}{s^2_{z_1} - \sigma^2_{\epsilon_1}}\Big\}. \tag{6.17}$$

(ii) If $\sigma^2_{\epsilon_2}$ is known, the estimate of ψ is found, by multiplying equations (d) by $\cos(\hat{\psi})/\sin(\hat{\psi})$, then substituting by (e),

$$\hat{\psi} = \text{atan}\Big\{\frac{s^2_{z_2} - \sigma^2_{\epsilon_2}}{s_{z_1z_2}}\Big\}. \tag{6.18}$$

(iii) If the ratio $\lambda = \sigma^2_{\epsilon_2}/\sigma^2_{\epsilon_1}$ is known, the estimate of ψ is given by

$$\hat{\psi} = \text{atan}\Big\{\frac{s^2_{z_2} - \lambda s^2_{z_1} + \sqrt{(s^2_{z_2} - \lambda s^2_{z_1})^2 + 4\lambda s^2_{z_1 z_2}}}{2s_{z_1 z_2}}\Big\}. \tag{6.19}$$

(iv) If both $\sigma^2_{\epsilon_1}$ and $\sigma^2_{\epsilon_2}$ are known, there will be three equations and two unknowns. By solving equations (c) and (e) from (6.11),

$$\hat{\psi} = \text{atan}\Big\{\frac{s_{z_1 z_2}}{s^2_{z_1} - \sigma^2_{\epsilon_1}}\Big\}. \tag{6.20}$$

Madansky (1959) pointed out that all estimates in (6.17)-(6.20) are not maximum likelihood estimates. He also argued that the case of both error variances $\sigma^2_{\epsilon_1}$ and $\sigma^2_{\epsilon_2}$ are known is an over-identified case, if $\sigma_{\epsilon_1\epsilon_2} = 0$, because there would be three equations and two unknowns. Thus, it can be assumed that $\sigma_{\epsilon_1\epsilon_2} \neq 0$, this will give three equations and three unknowns, and the resulting estimate will be the maximum likelihood estimate.

Estimators in (6.17)-(6.20) are obtained under the assumption $\sigma_{\epsilon_1\epsilon_2} = 0$. But, we can show that it is no harder to estimate ψ, whether the errors are correlated or not.

Suppose that the error variance $\Sigma_\epsilon \propto \Sigma^\circ_\epsilon$, where Σ°_ϵ is known, and the errors are correlated. Without loss of generality, z can be standardized using $A = (\Sigma^\circ_\epsilon)^{-1/2}$ as in (6.8), such that the error covariance matrix is proportional to the identity matrix.

Consider the covariance matrix of the standardized random vector x from (6.9). The eigenvalues of Σ_x are $\sigma^2_{\zeta_\circ} + c$ and c, the corresponding eigenvectors are γ, and $\gamma^\perp$, where $\gamma^\perp$ denotes the second eigenvector, orthogonal to γ.

Let $x_1, \ldots, x_n$, be a sample of size n from x in (6.8). The log likelihood function is given by

$$\frac{-2}{n}\log L = \log|\Sigma_x| + \text{tr}\Sigma_x^{-1} S_x, \tag{6.21}$$

where S_x is the sample covariance matrix. The determinant of Σ_x is the product of the eigenvalues, given by $c(\sigma^2_{\zeta_\circ} + c)$. The matrix Σ_x^{-1} can be written as

$$\begin{aligned}\Sigma_x^{-1} &= a\gamma\gamma^T + b\gamma^\perp\gamma^{\perp^T}, \quad 0 < a < b. \\ &= b(\gamma\gamma^T + \gamma^\perp\gamma^{\perp^T}) - (b-a)\gamma\gamma^T \\ &= bI_2 - (b-a)\gamma\gamma^T. \end{aligned} \tag{6.22}$$

Substituting (6.22) in (6.20) gives

$$\frac{-2}{n}\log L = \mathrm{tr}(\gamma\gamma^T S_x) = \gamma^T S_x \gamma. \tag{6.23}$$

From (6.23), maximizing the log likelihood depends on $\gamma^T S_x \gamma$. Equation (6.23) is maximized in the direction of the first principal component of S_x. The first principal component of S_x, is proportional to

$$\hat{\gamma}^T = (2s_{x_1x_2}, s^2_{x_2} - s^2_{x_1} + \sqrt{(s^2_{x_1} - s^2_{x_2})^2 + 4s^2_{x_1x_2}}),$$

e.g. Mardia et al. (1980). Thus,

$$\hat{\theta} = \mathrm{atan}\Big\{\frac{s^2_{x_2} - s^2_{x_1} + \sqrt{(s^2_{x_1} - s^2_{x_2})^2 + 4s^2_{x_1x_2}}}{2s_{x_1x_2}}\Big\}. \tag{6.24}$$

By back-transforming $\hat{\theta}$ into the z coordinate system, we can find $\hat{\psi}$. In this case $\hat{\psi}$ will be given as in (6.19).

In conclusion, when the signal $\zeta_\circ$ is normally distributed, it is not possible to estimate θ without additional knowledge about the error variances.

6.4 Non-normal signals

If $\zeta_\circ$ has a non-normal distribution, and its moments exist up to order four, the moments and cumulants can be used to find the signal direction of the EIV

model, defined in Section 6.2. In this section, we explore the use of cumulants in EIV, when the errors variances are unknown. We consider the cases of correlated and uncorrelated errors. We first start by reviewing univariate and bivariate cumulants in Section 6.4.1, and 6.4.2, respectively.

6.4.1 Univariate moments and cumulants

Moments and cumulants are quantities that describe the distribution of a random variable.

For a univariate random variable u, say, the jth order central and non-central moments, $\mu_u^{'}(j)$, and $\mu_u(j)$, are defined as in (3.15). For ease of notation, sometimes the subscript u is omitted from $\mu_u(j)$ or $\mu_u^{'}(j)$.

Moments and cumulants can be related explicitly using the moment generating function (m.g.f) of a random variable u, say , $\phi_u(s)$, defined as follows

$$\phi_u(s) = \mathrm{E}\{\exp(su)\} = 1 + \sum_{j=1}^{\infty} \frac{\mu_u(j)}{j!} s^j, \tag{6.25}$$

Taking the log of $\phi_u(s)$ gives

$$\log\, \phi_u(s) = \sum_{j=1}^{\infty} \frac{\kappa_u(j)}{j!} s^j. \tag{6.26}$$

By finding the coefficients of s^j, $j \geq 1$, in (6.25) and (6.26), explicit formulas that relate moments and cumulants can be found.

A list of cumulants, up to fourth-order, in terms of moments, is given as follows, Kendall and Stuart (1977),

$$\kappa(1) = \mu^{'}(1),\ \kappa(2) = \mu(2),\ \kappa(3) = \mu(3),\ \kappa(4) = \mu(4) - 3\mu^2(2). \tag{6.27}$$

Similarly, a list of moments up to fourth order, in terms of cumulants is given as

follows

$$\mu(2) = \kappa(2), \quad \mu(3) = \kappa(3), \quad \mu(4) = \kappa(4) + 3\kappa^2(2). \tag{6.28}$$

Indeed, the relations in (6.27) and (6.28) depend on the existence of cumulants/-moments up to an appropriate order.

Cumulants have properties that make them theoretically simpler than moments for certain purposes. Some of the properties are:

(i) For $j > 1$, and constants c_1 and $c_2 \neq 0$,

$$\kappa_{c_1 u + c_2}(j) = c_1^j \kappa_u(j). \tag{6.29}$$

This property also applies to jth order moments.

(ii) The jth order cumulant of the sum of two independent random variables is the sum of their jth order cumulants. This property is not true for moments of order $j > 2$.

(iii) For a normally distributed random variable, with mean zero, and variance σ^2, the log of m.g.f is given by

$$\log\{\phi(s)\} = \frac{1}{2} s^2 \sigma^2.$$

This means that all cumulants of order higher than 2 are equal to zero. In contrast, higher order moments do not vanish.

6.4.2 Joint moments and cumulants

For a bivariate random vector $u = (u_1, u_2)^T$, say, let the $(j_1 + j_2)$-order population and sample joint moments about the origin be denoted by $\mu_u'(j_1, j_2)$ and $m_u'(j_1, j_2)$, the central and non-central joint moments, $\mu_u(j_1, j_2)$, and $m_u(j_1, j_2)$, defined as in (3.16).

As in the univariate case, joint cumulants can be explicitly related to each other using the m.g.f $\phi_{u_1,u_2}(s,t)$, defined as follows

$$\phi_{u_1,u_2}(s,t) = \mathrm{E}[\exp\{su_1 + tu_2\}] = 1 + \sum_{j_1+j_2\geq 1} \frac{\mu_{u_1,u_2}(j_1,j_2)}{j_1!j_1!} s^{j_1}t^{j_2}. \tag{6.30}$$

Taking the log, gives

$$\log\, \phi_{u_1,u_2}(s,t) = \sum_{j_1+j_2\geq 1} \frac{\kappa_{u_1,u_2}(j_1,j_2)}{j_1!j_1!} s^{j_1}t^{j_2}. \tag{6.31}$$

From (6.30) and (6.31), the following list of cumulants in terms of moments, up to fourth order, can be found,

$$\begin{aligned}
&\kappa(1,1) = \mu(1,1), \quad \kappa(2,0) = \mu(2,0), \quad \kappa(0,2) = \mu(0,2),\\
&\kappa(2,1) = \mu(2,1), \quad \kappa(1,2) = \mu(1,2), \quad \kappa(3,0) = \mu(3,0),\\
&\kappa(0,3) = \mu(0,3), \quad \kappa(2,2) = \mu(2,2) - \mu(2,0)\mu(0,2) - 2\mu^2(1,1),\\
&\kappa(3,1) = \mu(3,1) - 3\mu(2,0)\mu(1,1), \quad \kappa(1,3) = \mu(1,3) - 3\mu(0,2)\mu(1,1),\\
&\kappa(4,0) = \mu(4,0) - 3\mu^2(2,0), \quad \kappa(0,4) = \mu(0,4) - 3\mu^2(0,2).
\end{aligned} \tag{6.32}$$

Similarly, a list of formulas of moments in terms of cumulants is given as follows.

$$\begin{aligned}
&\mu(1,1) = \kappa(1,1), \quad \mu(2,0) = \kappa(2,0), \quad \mu(0,2) = \kappa(0,2),\\
&\mu(2,1) = \kappa(2,1), \quad \mu(1,2) = \kappa(1,2), \quad \mu(3,0) = \kappa(3,0),\\
&\mu(0,3) = \kappa(0,3), \quad \mu(3,1) = \kappa(3,1) + 3\kappa(2,0)\kappa(1,1),\\
&\mu(1,3) = \kappa(1,3) + 3\kappa(0,2)\kappa(1,1), \mu(2,2) = \kappa(2,2) + \kappa(2,0)\kappa(0,2) + 2\kappa^2(1,1),\\
&\mu(4,0) = \kappa(4,0) + 3\kappa(2,0)^2, \quad \mu(0,4) = \kappa(0,4) + 3\kappa(0,2)^2.
\end{aligned} \tag{6.33}$$

The properties of univariate cumulants can be generalized to the bivariate case. Some of the properties of joint cumulants are given by

(i) The $(j_1 + j_2)$-order cumulant $\kappa_{u_1,u_2}(j_1,j_2)$ is equivariant under diagonal

scaling, i.e for any constants a_1, and $a_2 \neq 0$,

$$\kappa_{a_1 z_1, a_2 z_2}(j_1, j_2) = a_1^{j_1} a_2^{j_2} \kappa_{z_1, z_2}(j_1, j_2). \tag{6.34}$$

Note that, cumulants are not equivariant under affine transformations, or orthogonal rotations.

(ii) The $(j_1 + j_2)$-order cumulant of the sum of two independent bivariate random vectors is the sum of their $(j_1 + j_2)$-order cumulants.

(iii) For normally distributed random vector, with covariance matrix

$$\begin{pmatrix} \sigma_{u_1}^2 & \sigma_{u_1 u_2} \\ \sigma_{u_1 u_2} & \sigma_{u_2}^2 \end{pmatrix},$$

its m.g.f is given by

$$\phi_{u_1, u_2}(s, t) = \exp[\frac{1}{2}(s^2 \sigma_{u_1}^2 + t^2 \sigma_{u_2}^2 + st\sigma_{u_1 u_2})].$$

This means that cumulants of order higher than two are equal to zero.

6.4.3 Cumulants in EIV

There is an extensive literature that has discussed the use of moments and cumulants in EIV, including Geary (1941), Pal (1980), Cragg (1997) and Gillard (2010).

Since the signal $\zeta_\circ$ is non-normal, second moments are not sufficient statistics. In this case high order moments/cumulants can provide more information about the signal direction.

Geary (1941) was the first to use the joint cumulants to estimate ψ. Consider the random vector z in (6.5). Suppose that z is shifted in advance to have zero

mean vector. Geary (1941) noted the following formula for ψ

$$\operatorname{atan}\Big[\Big\{\frac{\kappa(j,l+k)}{\kappa(j+k,l)}\Big\}^{1/k}\Big], \tag{6.35}$$

for $j+k+l>2$. Formula (6.35) is well defined if $\kappa(j+k,l)\neq 0$. Geary's formula follows from the properties of joint cumulants, as shown from (6.40) to (6.43).

Since $\zeta_\circ$ and ϵ are independent,

$$\kappa_{z_1,z_2}(j,l+k)=\cos^j(\psi)\sin^{(l+k)}(\psi)\kappa_{\zeta_\circ}(j+l+k)+\kappa_{\epsilon_1\epsilon_2}(j_1,j_2). \tag{6.36}$$

For $j+k+l>2$, (6.36) reduces to the following identity,

$$\kappa_{z_1,z_2}(j,l+k)=\cos^j(\psi)\sin^{(l+k)}(\psi)\kappa_{\zeta_\circ}(j+l+k). \tag{6.37}$$

Using (6.37), we can evaluate $\kappa_{\zeta_\circ}(4)$ as follows,

$$\kappa_{\zeta_\circ}(4)=\kappa(4,0)+\kappa(0,4)+2\kappa(2,2). \tag{6.38}$$

The following cumulant-based equations are deduced from (6.36), for $j+l+k=$

$3, 4$, subject to $\kappa_{\zeta_\circ}(3)$ and $\kappa(4)_{\zeta_\circ} \neq 0$, as follows.

$$\begin{aligned}
\kappa(2,1) &= \cos^2(\psi)\sin(\psi)\kappa_{\zeta_\circ}(3),\\
\kappa(1,2) &= \cos(\psi)\sin^2(\psi)\kappa_{\zeta_\circ}(3),\\
\kappa(3,0) &= \cos^3(\psi)\kappa_{\zeta_\circ}(3),\\
\kappa(0,3) &= \sin^3(\psi)\kappa_{\zeta_\circ}(3),\\
\kappa(2,2) &= \sin^2(\psi)\cos^2(\psi)\kappa_{\zeta_\circ}(4),\\
\kappa(3,1) &= \cos^3(\psi)\sin(\psi)\kappa_{\zeta_\circ}(4),\\
\kappa(1,3) &= \cos(\psi)\sin^3(\psi)\kappa_{\zeta_\circ}(4),\\
\kappa(4,0) &= \cos^4(\psi)\kappa_{\zeta_\circ}(4),\\
\kappa(0,4) &= \sin^4(\psi)\kappa_{\zeta_\circ}(4). \qquad (6.39)
\end{aligned}$$

By dividing any pair of cumulants, of the same order, we can find an infinite number of formulas of ψ. There are three basic third-order, and four basic fourth-order cumulant-based formulas of ψ, and all others are functions of these formulas, given as follows.

(i) Third-order cumulant based formulas,

- The three third-order basic formulas are:

$$\text{atan}\Big[\frac{\kappa(1,2)}{\kappa(2,1)}\Big], \quad \text{atan}\Big[\frac{\kappa(0,3)}{\kappa(1,2)}\Big], \quad \text{atan}\Big[\frac{\kappa(2,1)}{\kappa(3,0)}\Big]. \qquad (6.40)$$

- Some of other third-order formulas are given by:

$$\text{atan}\Big[\pm\{\frac{\kappa(1,2)}{\kappa(3,0)}\}^{1/2}\Big], \quad \text{atan}\Big[\pm\{\frac{\kappa(0,3)}{\kappa(2,1)}\}^{1/2}\Big] \quad \text{atan}\Big[\{\frac{\kappa(0,3)}{\kappa(3,0)}\}^{1/3}\Big]. \qquad (6.41)$$

(ii) Fourth-order cumulant based formulas,

- The four fourth-order basic formulas are:

$$\operatorname{atan}\Big[\frac{\kappa(3,1)}{\kappa(4,0)}\Big], \quad \operatorname{atan}\Big[\frac{\kappa(0,4)}{\kappa(1,3)}\Big], \quad \operatorname{atan}\Big[\frac{\kappa(2,2)}{\kappa(3,1)}\Big], \operatorname{atan}\Big[\frac{\kappa(1,3)}{\kappa(2,2)}\Big]. \quad (6.42)$$

- Some of other fourth-order formulas are given by:

$$\operatorname{atan}\Big[\pm\{\frac{\kappa(1,3)}{\kappa(3,1)}\}^{1/2}\Big], \quad \operatorname{atan}\Big[\pm\{\frac{\kappa(0,4)}{\kappa(4,0)}\}^{1/4}\Big], \quad \operatorname{atan}\Big[\pm\{\frac{\kappa(2,2)}{\kappa(4,0)}\}^{1/2}\Big]. \quad (6.43)$$

Geary's formula in (6.35) is implied by (6.40)-(6.43). For $k = 1$, and $j + l = 2$, Geary's formula, in (6.35), gives the three basic third-order formulas in (6.40). Similarly, for $k = 1$, and $j + l = 3$, Geary's formula gives the four fourth-order formulas in (6.42).

The sign of ψ is not determined in any of the formulas in (6.41), or (6.43), hence it must be determined separately, e.g by $\operatorname{sign}(k(1,1))$.

Let $z_1, \ldots, z_n$ be a random sample from (6.5), where $z_i^T = (z_{i1}, z_{i2})^T$, $i = 1, \ldots, n$. The sample version of Geary's formula in (6.35) is

$$\hat{\psi} = \operatorname{atan}\Big[\Big\{\frac{\hat{\kappa}(j, l+k)}{\hat{\kappa}(j+k, l)}\Big\}^{1/k}\Big], \quad (6.44)$$

for $j + k + l > 2$. When the power k is positive, $\hat{\psi}$ will be not defined if the ratio of the sample cumulants is negative. There are two possible solutions to this problem: (1) both denominator and numerator can be given the same sign. (2) $\hat{\psi}$ can be given a zero value.

6.4.4 The effect of rotation on cumulant based formulas

We know from joint cumulants properties, explained in Section 6.4.2, that cumulants are not equivariant under orthogonal rotations. But we can show that under 90° and 45° rotation, the four basic fourth-order estimates, in (6.42), can be related to each other.

Recall the four estimates from (6.42),

$$\psi_1 = \operatorname{atan}\Big[\frac{\kappa(3,1)}{\kappa(4,0)}\Big],$$
$$\psi_2 = \operatorname{atan}\Big[\frac{\kappa(0,4)}{\kappa(1,3)}\Big],$$
$$\psi_3 = \operatorname{atan}\Big[\frac{\kappa(2,2)}{\kappa(3,1)}\Big],$$
$$\psi_4 = \operatorname{atan}\Big[\frac{\kappa(1,3)}{\kappa(2,2)}\Big].$$

Note that when $\psi = 0°$, ψ_2 will be inefficient, and when $\psi = 90$, ψ_1 will be inefficient.

To proceed suppose that z is rotated by an angle ω, using the rotation matrix in (1.4). The rotated random vector $z^{'}$ is given by

$$\begin{pmatrix} z_1^{'} \\ z_2^{'} \end{pmatrix} = \begin{pmatrix} \cos(\psi+\omega) \\ \sin(\psi+\omega) \end{pmatrix} \zeta_\circ + \begin{pmatrix} \epsilon_1^{'} \\ \epsilon_1^{'} \end{pmatrix}.$$

where

$$Rz = \begin{pmatrix} z_1^{'} \\ z_2^{'} \end{pmatrix} = \begin{pmatrix} z_1\cos(\omega) - z_2\sin(\omega) \\ z_1\sin\omega) + z_2\cos(\omega) \end{pmatrix}, R\begin{pmatrix} \cos(\psi) \\ \sin(\psi) \end{pmatrix} = \begin{pmatrix} \cos(\psi+\omega) \\ \sin(\psi+\omega) \end{pmatrix},$$
$$\epsilon^{'} = R\epsilon. \tag{6.45}$$

The m.g.f of $z^{'}$ is given by

$$\begin{aligned}
\phi_{z_1^{'},z_2^{'}}(s,t) &= \mathrm{E}\{\exp(z_1^{'}s + z_2^{'}t)\} \\
&= \mathrm{E}\{\exp[(z_1\cos(\omega) - z_2\sin(\omega))s + (z_1\sin(\omega) + z_2\cos(\omega))t]\} \\
&= \mathrm{E}\{\exp[z_1(s\cos(\omega) + t\sin(\omega)) + z_2(-s\sin(\omega) + t\cos(\omega))]\} \\
&= \phi_{z_1,z_2}\{(s\cos(\omega) + t\sin(\omega)), (-s\sin(\omega) + t\cos(\omega))\}.
\end{aligned}$$

Thus,

$$\sum_{j_1+j_2\geq 1} \kappa_{z_1',z_2'}(j_1,j_2)\frac{s^{j_1}t^{j_2}}{j_1!j_2!} = \\ = \sum_{j_1+j_2\geq 1} \kappa_{z_1,z_2}(j_1,j_2)\frac{(s\cos(\omega)+t\sin(\omega))^{j_1}}{j_1!}\frac{(-s\sin(\omega)+t\cos(\omega))^{j_2}}{j_2!}. \quad (6.46)$$

Using (6.46), we can find the forms of the fourth-order cumulants, under $\omega = 90°$, and $\omega = 45°$, as follows.

- Under $\omega = 90°$ rotation:

$$\begin{aligned} \kappa_{z_1'z_2'}(4,0) &= \kappa_{z_1z_2}(0,4), \\ \kappa_{z_1'z_2'}(0,4) &= \kappa_{z_1z_2}(4,0), \\ \kappa_{z_1'z_2'}(2,2) &= \kappa_{z_1z_2}(2,2), \\ \kappa_{z_1'z_2'}(3,1) &= -\kappa_{z_1z_2}(1,3), \\ \kappa_{z_1'z_2'}(1,3) &= -\kappa_{z_1z_2}(3,1). \end{aligned} \quad (6.47)$$

From (6.47),

$$\begin{aligned} \hat{\psi}_1' &= \operatorname{atan}\Big[\frac{\kappa_{z_1'z_2'}(3,1)}{\kappa_{z_1'z_2'}(4,0)}\Big] = \operatorname{atan}\Big[-\frac{\kappa_{z_1z_2}(1,3)}{\kappa_{z_1z_2}(0,4)}\Big] = -\frac{1}{\hat{\psi}_2}, \\ \hat{\psi}_2' &= \operatorname{atan}\Big[\frac{\kappa_{z_1'z_2'}(0,4)}{\kappa_{z_1'z_2'}(1,3)}\Big] = \operatorname{atan}\Big[-\frac{\kappa_{z_1z_2}(4,0)}{\kappa_{z_1z_2}(3,1)}\Big] = -\frac{1}{\hat{\psi}_1}, \\ \hat{\psi}_3' &= \operatorname{atan}\Big[\frac{\kappa_{z_1'z_2'}(2,2)}{\kappa_{z_1'z_2'}(3,1)}\Big] = \operatorname{atan}\Big[-\frac{\kappa_{z_1z_2}(2,2)}{\kappa_{z_1z_2}(1,3)}\Big] = -\frac{1}{\hat{\psi}_4}, \\ \hat{\psi}_4' &= \operatorname{atan}\Big[\frac{\kappa_{z_1'z_2'}(1,3)}{\kappa_{z_1'z_2'}(2,2)}\Big] = \operatorname{atan}\Big[-\frac{\kappa_{z_1z_2}(3,1)}{\kappa_{z_1z_2}(2,2)}\Big] = -\frac{1}{\hat{\psi}_3}. \end{aligned} \quad (6.48)$$

- Under $\omega = 45°$ rotation:

$$\begin{aligned}
\kappa_{z_1' z_2'}(4,0) &= 0.25\{\kappa_{z_1 z_2}(4,0) + \kappa_{z_1 z_2}(0,4) + 6\kappa_{z_1 z_2}(2,2) - 4\kappa_{z_1 z_2}(3,1) - 4\kappa_{z_1 z_2}(1,3)\},\\
\kappa_{z_1' z_2'}(0,4) &= 0.25\{\kappa_{z_1 z_2}(4,0) + \kappa_{z_1 z_2}(0,4) + 6\kappa_{z_1 z_2}(2,2) + 4\kappa_{z_1 z_2}(3,1) + 4\kappa_{z_1 z_2}(1,3)\},\\
\kappa_{z_1' z_2'}(2,2) &= 0.25\{\kappa_{z_1 z_2}(4,0) + \kappa_{z_1 z_2}(0,4) - 2\kappa_{z_1 z_2}(2,2)\},\\
\kappa_{z_1' z_2'}(3,1) &= 0.25\{\kappa_{z_1 z_2}(4,0) - \kappa_{z_1 z_2}(0,4) - 2\kappa_{z_1 z_2}(3,1) + 2\kappa_{z_1 z_2}(1,3)\},\\
\kappa_{z_1' z_2'}(1,3) &= 0.25\{\kappa_{z_1 z_2}(4,0) - \kappa_{z_1 z_2}(0,4) + 2\kappa_{z_1 z_2}(3,1) - 2\kappa_{z_1 z_2}(1,3)\}.
\end{aligned} \tag{6.49}$$

$$\begin{aligned}
\hat{\psi}_1' &= \hat{\psi}_2',\\
\hat{\psi}_3' &= \hat{\psi}_4'.
\end{aligned} \tag{6.50}$$

6.5 ICS:kurtosis:variance in EIV

In Chapter 3, we have explored the use of ICS:kurtosis:variance, in finding groups separation direction. In this section, we explore using ICS:kurtosis:variance in estimating the signal direction.

Consider the random vector z defined in (6.5). Without loss of generality, assume that z is standardized with respect to $\Sigma_z^{-1/2}$, as in (6.10), such that $\Sigma_y = I_2$.

Recall that the ICS:kurtosis:variance criterion, from (3.36), is to maximize/minimize

$$\begin{aligned}
\kappa_{\text{ICS}}(b) &= b^T \Sigma_z^{-1/2} K_z \Sigma_z^{-1/2} b\\
&= b^T K_y b,
\end{aligned} \tag{6.51}$$

where b is in the direction of the smallest/largest eigenvector,K_z is the fourth-order moment matrix, defined in (3.23), and $K_y = \Sigma_z^{-1/2} K_z \Sigma_z^{-1/2}$.

The component of the matrix K_y can be written in terms of moments as follows

$$K_y = \begin{pmatrix} \mu_y(4,0) + \mu_y(2,2) & \mu_y(3,1) + \mu_y(1,3) \\ \mu_y(3,1) + \mu_y(1,3) & \mu_y(2,2) + \mu_y(0,4) \end{pmatrix}.$$

The matrix K_y can also be expressed in terms of cumulants,

$$K_y = \begin{pmatrix} \kappa_y(4,0) + \kappa_y(2,2) + 4 & \kappa_y(3,1) + \kappa_y(1,3) \\ \kappa_y(3,1) + \kappa_y(1,3) & \kappa_y(2,2) + \kappa_y(0,4) + 4 \end{pmatrix}. \tag{6.52}$$

Substituting by the cumulant equations from (6.39), K_y can be written as

$$K_y = \nu\nu^T \kappa_{\zeta_\circ}(4) + 4I_2. \tag{6.53}$$

The signal direction $\zeta_\circ$ will be in the direction that maximizes or minimizes $b^T K_y b$.

From (6.53), one of the eigenvalues of K_y will be in the signal direction, and the other will be in the noise direction, as follows

$$\begin{aligned} \lambda_{\text{sig}} &= \kappa_{\zeta_\circ}(4) + 4, \\ \lambda_{\text{noise}} &= 4, \end{aligned} \tag{6.54}$$

with the corresponding eigenvectors ν and $\nu^\perp$, respectively.

Choosing whether to maximize or minimize depends on the sign of $\kappa_{\zeta_\circ}(4)$, which can be evaluated using (6.38).

If $\kappa_{\zeta_\circ}(4)$ is negative, ν will be the smallest eigenvector. If $\kappa_{\zeta_\circ}(4)$ is positive, ν will be in the largest eigenvector.

The case of negative $\kappa_{\zeta_\circ}(4)$ is satisfied when $\zeta_\circ$ has a bimodal distribution, and the case of positive $\kappa_{\zeta_\circ}(4)$ is satisfied when $\zeta_\circ$ has a long-tailed distribution.

The eigenvalues and eigenvectors of K_y, defined in (6.54), are given as follows,

in terms of cumulants

$$\lambda_1 = \frac{1}{2}(\kappa_y(4,0) + \kappa_y(0,4)) + \kappa_y(2,2) + 4 + \frac{1}{2}\Delta,$$
$$\lambda_2 = \frac{1}{2}(\kappa_y(4,0) + \kappa_y(0,4)) + \kappa_y(2,2) + 4 - \frac{1}{2}\Delta. \tag{6.55}$$

where

$$\Delta = \sqrt{(\kappa_y(4,0) - \kappa_y(0,4))^2 + 4(\kappa_y(3,1) + \kappa_y(1,3))^2}.$$

The largest eigenvalue is λ_1, and the smallest is λ_2. Their corresponding eigenvectors are

$$\nu_1 = ((\kappa_y(4,0) - \kappa_y(0,4) + \Delta), 2(\kappa_y(3,1) + \kappa_y(1,3)))^T,$$
$$\nu_2 = ((\kappa_y(4,0) - \kappa_y(0,4) - \Delta), 2(\kappa_y(3,1) + \kappa_y(1,3))^T, \tag{6.56}$$

Therefore, from ν_1 in (6.56), ϕ takes the following form

$$\phi = \operatorname{atan}\Big[\frac{2(\kappa_y(3,1) + \kappa_y(1,3))}{(\kappa_y(4,0) - \kappa_y(0,4) + \Delta)}\Big], \tag{6.57}$$

Substituting by the formulas in (6.39), which gives

$$\operatorname{atan}\Big[\frac{\kappa_{\zeta_\circ}(4)\cos(\phi)\sin(\phi)}{\kappa_{\zeta_\circ}(4)(\cos(\phi)^2 - \sin(\phi)^2) + 1}\Big]$$
$$= \operatorname{atan}\Big[\frac{\sin(2\phi)}{\cos(2\phi) + 1}\Big] = \operatorname{atan}\Big[\tan(\phi)\Big] = \phi.$$

6.6 Simulation study

In this simulation study, we compare the accuracy of Geary's four basic fourth-order cumulant-based estimators from (6.42), and examine the effect of rotating the data by 45°, and 90°. We also compare the accuracy of these cumulant-based estimators with the ICS:kurtosis:variance.

The eigenvalue of ICS is chosen by evaluating $\kappa_{\zeta_\circ}(4)$, defined in (6.38). If

$\kappa_{\zeta_\circ}(4) > 0$, the largest eigenvector is chosen. If $\kappa_{\zeta_\circ}(4) < 0$, the smallest eigenvector is chosen. The percentage of picking the right eigenvector will be examined in Table 6.4.

Let X,be an $n \times 2$ data matrix, its rows are given by $x_i^T = (x_{i1,x_{i2}})$, $i = 1, \ldots, n$, $n = 50, 100, 200, 500$. The data matrix X is generated from x in (6.5), where the signal $\zeta_\circ$ is distributed as a t_2 distribution, and the noise ϵ is normally distributed, with covariance matrix equals to $\sigma^2 I_2$, with $\sigma^2 = 0.4$. Without loss of generality assume that the signal is standardized to have zero mean and unit variance. For n, the simulation is repeated 1000 times.

From (6.42), the fourth-order cumulant-based estimators considered here are:

$$\hat{\theta}_1 = \operatorname{atan}\Big[\frac{\hat{\kappa}(3,1)}{\hat{\kappa}(4,0)}\Big],$$
$$\hat{\theta}_2 = \operatorname{atan}\Big[\frac{\hat{\kappa}(0,4)}{\hat{\kappa}(1,3)}\Big],$$
$$\hat{\theta}_3 = \operatorname{atan}\Big[\frac{\hat{\kappa}(2,2)}{\hat{\kappa}(3,1)}\Big],$$
$$\hat{\theta}_4 = \operatorname{atan}\Big[\frac{\hat{\kappa}(1,3)}{\hat{\kappa}(2,2)}\Big].$$

The ICS:kurtosis:variance estimator will be denoted by: $\hat{\theta}_{\text{ICS}}$. The accuracy will be measured using (3.65).

The results are shown in Tables 6.1, 6.2, and 6.3.

Table 6.1: Variances of different estimates of θ, where the true signal direction is $\theta = 0°$.

n	$\text{var}(\hat{\theta}_1)$	$\text{var}(\hat{\theta}_2)$	$\text{var}(\hat{\theta}_3)$	$\text{var}(\hat{\theta}_4)$	$\text{var}(\hat{\theta}_{ICS})$
50	0.074	0.548	0.223	0.277	0.225
100	0.021	0.538	0.162	0.222	0.1
200	0.006	0.534	0.123	0.207	0.042
500	0.002	0.541	0.085	0.169	0.017

Table 6.2: Variances of different estimates of θ, where the true signal direction is $\theta = 45°$.

n	$\text{var}(\hat{\theta}_1)$	$\text{var}(\hat{\theta}_2)$	$\text{var}(\hat{\theta}_3)$	$\text{var}(\hat{\theta}_4)$	$\text{var}(\hat{\theta}_{ICS})$
50	0.132	0.133	0.081	0.093	0.224
100	0.048	0.048	0.024	0.027	0.103
200	0.013	0.011	0.007	0.007	0.042
500	0.002	0.002	0.002	0.002	0.017

Table 6.3: Variances of different estimates of θ, where the true signal direction is $\theta = 90°$.

n	$\text{var}(\hat{\theta}_1)$	$\text{var}(\hat{\theta}_2)$	$\text{var}(\hat{\theta}_3)$	$\text{var}(\hat{\theta}_4)$	$\text{var}(\hat{\theta}_{ICS})$
50	0.543	0.078	0.256	0.222	0.214
100	0.532	0.02	0.231	0.169	0.098
200	0.546	0.006	0.204	0.135	0.043
500	0.547	0.002	0.162	0.093	0.016

Table 6.4: Percentage of picking the right eigenvector.

n	percentage
50	0.927
100	0.993
200	1
500	1

- From Table 6.1, for $\theta = 0°$, $\hat{\theta}_1$ is most accurate among all other estimates, then $\hat{\theta}_{\text{ICS}}$ is the second most accurate. The remaining estimates do not work well.

- From Table 6.2, for $\theta = 45°$, the the accuracy of the estimates $\hat{\theta}_1$ and $\hat{\theta}_2$ are similar. Also, $\hat{\theta}_3$ and $\hat{\theta}_4$ have similar accuracy. Moreover, $\hat{\theta}_1$, $\hat{\theta}_2$,$\hat{\theta}_3$ and $\hat{\theta}_4$ outperform $\hat{\theta}_{\text{ICS}}$.

- From Table 6.3, $\hat{\theta}_2$ is the most accurate. Note that $\hat{\theta}_2$ when $\theta = 90°$ is equal to $\hat{\theta}_1$ when $\theta = 0°$.

- Note that $\hat{\theta}_{\text{ICS}}$ has the advantage of being affine equivariant.

6.7 Conclusion

we have discussed the theory of the EIV model when the signal has normal and non-normal distribution.

When the signal is normally distributed, additional knowledge is required to estimate the signal direction. We have shown that when the errors variances are known, whether the errors are correlated or not, it is possible to find the maximum likelihood estimator of the signal direction.

When the signal has a non-normal distribution, ICS:kurtosis:variance can be used to estimate the signal direction. We also compared the efficiencies of ICS estimators with Geary's fourth-order cumulant-based estimators. The results show that some of the cumulant-based estimators are more efficient than the ICS estimator, but ICS method has the advantage of being affine equivaraint.

Chapter 7

ICS for RANDU data set

7.1 Introduction

The RANDU data set contains points arranged on 15 parallel planes, lying in the dimensional space. Any pairwise scatter plot of the RANDU data does not reveal the parallel plane structure.

Tyler et al. (2009) applied ICS^d:W-estimate:variance to the RANDU data set to reveal the parallel plane structure. The W-estimate, based on pairwise differencing of the data, accentuates inliers. Inliers appear more often for points on the same line than on different lines.

The main goal of this chapter is to understand how ICS^d:W-estimate:variance works to discover the structure in the RANDU data set.

This chapter is organized as follows. In Section 7.2, we show the RANDU example from Tyler et al. (2009). In Section 7.3, we explain a RANDU type model. In Section 7.4, we study the behavior of the W-estimate based on pairwise differencing of the data, under the model explained in Section 7.3. A detailed analysis of the behavior of the W-estimate in two-dimension is given in Section 7.5. In Section 7.6, we explain a noisy version of the RAND-type model by allowing small variation in each plane. We also explore the behaviour of ICS^d:W-estimate:variance under the noisy RANDU-type model.

7.2 ICS for RANDU data set

7.2.1 RAND data set

The RANDU data set contains 400 three dimensional data points arranged in 15 equally spaced parallel planes, generated by the 1960s linear congruential generator RANDU.

Any linear congruential generator (LCG) takes the following form, starting from seed x_0 (see for example **?**),

$$x_i = (ax_{i-1} + b) \mod M,$$

where a, b and M are constants. The sequence of generated numbers are periodic, with period length equals to $M - 1$.

For the RANDU data set, the choices of the constants were not good (see for example **?**):

$$a = 655539, \quad b = 0, \text{ and } \quad M = 2^{31}.$$

The RANDU generator is given by

$$\begin{aligned} x_i &= 65539x_{i-1} \bmod 2^{31} \\ &= (2^{16} + 3)x_{i-1} \bmod 2^{31} \\ &= (2^{16} + 3)^2 x_{i-2} \bmod 2^{31} \\ &= 6x_i - 9x_{i-1} \bmod 2^{31}. \end{aligned}$$

Hence, any three successive points have the following linear relationship

$$x_i - 6x_i + 9x_{i-1} = c2^{31}.$$

Every three successive numbers lie on 15 parallel planes, since $0 \leq x_i \leq 2^{31}$. The structure direction in the RANDU data set is the direction normal to the planes,

whereas the noise directions are the directions parallel to the planes.

The parallel plane structure is not revealed in the pairwise scatter plots of the RANDU data set, shown in Figure 7.1 (a).

7.2.2 RANDU example

Let the RANDU data set be denoted by $X = (x_i)$, where $i = 1, \ldots, 400$, $x_i \in R^3$. The W-estimate, $\hat{V}$, Tyler et al. (2009) is defined as follows

$$\hat{V} = \frac{2}{n(n-1)} \sum_{i=1}^{n-1} \sum_{j=i+1}^{n} \frac{(x_i - x_j)(x_i - x_j)^T}{\{(x_i - x_j)^T S^{-1} (x_i - x_j)\}^2}. \tag{7.1}$$

The matrix $\hat{V}$ is a weighted scatter matrix computed with respect to the pairwise differences of the rows of RANDU, with weight function

$$\frac{1}{\{(x_i - x_j)^T S^{-1} (x_i - x_j)\}^2}. \tag{7.2}$$

The weight (7.2) is large if the difference $\|x_i - x_j\|$ is small, and small otherwise.

The covariance matrix S is equal to

$$S = \begin{pmatrix} 0.081 & -0.004 & 0.005 \\ -0.004 & 0.086 & -0.005 \\ 0.005 & -0.005 & 0.078 \end{pmatrix},$$

and $\hat{V}$ matrix is equal to

$$\hat{V} = \begin{pmatrix} 0.135 & 0.138 & 0.028 \\ 0.138 & 0.264 & 0.041 \\ 0.028 & 0.041 & 0.075 \end{pmatrix}.$$

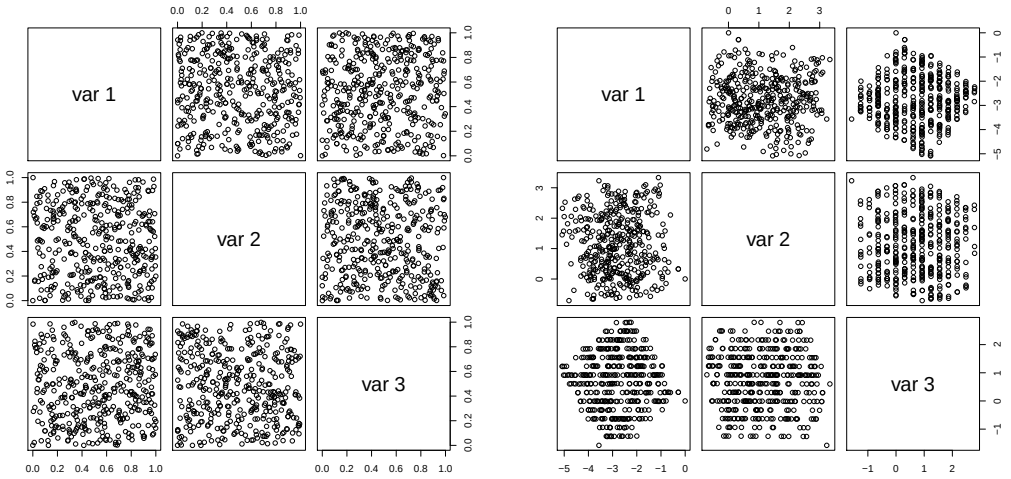
The eigenvalues and eigenvectors of $S^{-1}\hat{V}$ are

$$
\begin{aligned}
l_1 = 2.247, l_2 = 0.429, l_3 = 0.269,\\
u_1^T = (-0.555, -0.806, -0.205),\\
u_2^T = (-0.231, -0.061, 0.971),\\
u_3^T = (-0.829, 0.553, -0.086).
\end{aligned}
$$

The pairwise scatter plot of XU, where $U = (u_1, u_2, u_3)$ is shown in Figure 7.1 (b).

The plot shows that the parallel line structure direction is in the direction of the smallest eigenvector, and the remaining eigenvectors are in the noise direction.

Since the structure direction and the noise directions in the RANDU data set are uniformly distributed, applying ICS:kurtosis:variance, explored in Chapter 3, will fail to detect the parallel line structure. The eigenvalues of $S^{-1}K$ are approximately equal to each other.



(a) (b)

Figure 7.1: The scatter plot of RANDU data set (a) and the transformed data set (b).

Since the covariance matrix is approximately proportional to the identity, we explore the eigenvalues and eigenvectors of $\hat{V}$ in more detail.

To gain more insight into $\hat{V}$, we look at the RANDU data set in a simpler setting. We use two-dimensional subset from the projected RANDU data set, which is shown in Figure 7.1 (b); the second and the third component. The reduced RANDU data set is denoted by X^s, where its rows are given by $x_i^{sT} = (x_{i1}^s, x_{i2}^s)$, $i = 1, \ldots, 400$.

Figure 7.2 shows plots of X^s. The structure direction in X^s is the direction orthogonal to the lines, and the noise direction is the direction parallel to the lines.

Consider two points from X^s, x_i^s and x_j^s, say. There are two possibilities:

(i) x_i^s and x_j^s are from the same group, or

(ii) x_i^s and x_j^s are from different groups.

Under (i), the difference $|x_i^s - x_j^s|$ can range from small to large, whereas under (ii) the length of $x_i^s - x_j^s$ is at least 1/15.

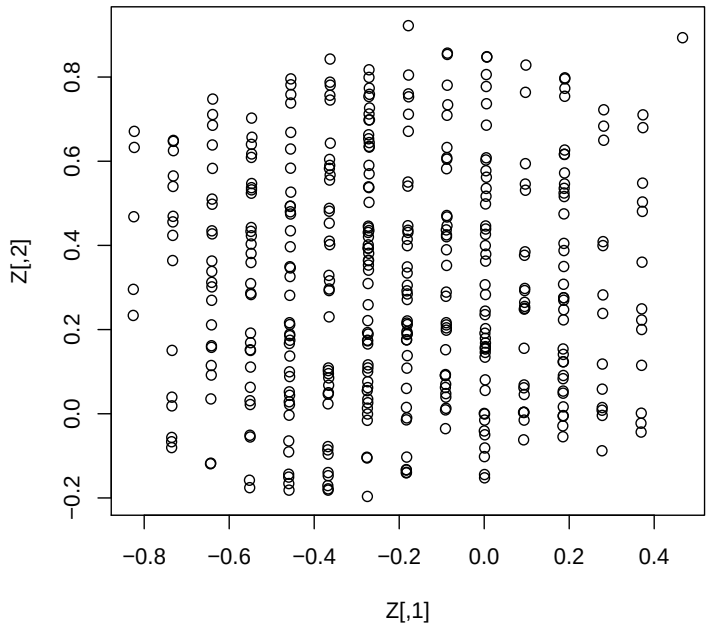


Figure 7.2: The scatter plot of a subset of RANDU data set.

The weights in (7.2) take larger values as $|x_i^s - x_j^s|$ have smaller values. Differ-

ences that have small lengths are called inliers. As we mentioned earlier, inliers appear only when x_i^s and x_j^s are in the same group. Hence, the scatter matrix $\hat{V}$ accentuates inliers by giving them large weights. Therefore, the dominant eigenvectors of $\hat{V}$ will be in the noise directions, while the smallest eigenvector will be in the structure direction.

The direction of the dominant eigenvector of $\hat{V}$ depends on the angles between x_i^s and x_j^s. The difference $x_i^s - x_j^s$ can be expressed by polar coordinate (r_{ij}, θ_{ij}) as follows

$$r_{ij}^2 = (x_i^s - x_j^s)^T(x_i^s - x_j^s), \quad i \neq j = 1, \ldots, n,$$
$$\theta_{ij} = \text{atan2}(x_{i2}^s - x_{j2}^s, x_{i1}^s - x_{j1}^s),$$

where atan2 is defined as follows

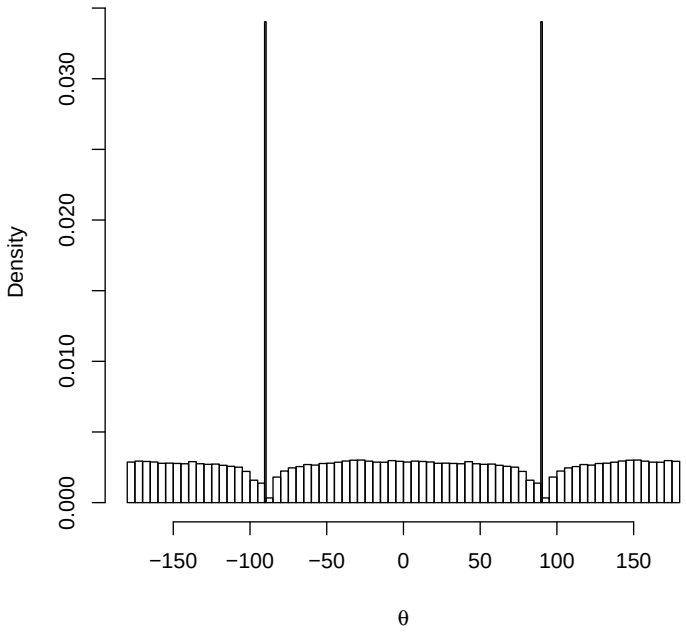
$$\text{atan2}(v_2, v_1) = \begin{cases} \text{atan}(v_2/v_1) & \text{if } v_1 > 0 \\ \pi/2 & \text{if } v_1 = 0, v_2 > 0 \\ -\pi/2 & \text{if } v_1 = 0, v_2 < 0 \\ \pi + \text{atan}(v_2/v_1) & \text{if } v_1 < 0. \end{cases}.$$

Each angle θ_{ij}, describes the direction of the line from x_i^s to x_j^s. The histogram of θ_{ij} under (i) and (ii) are

- Under (i), θ_{ij} will always be in the direction $\pm 90°$, because their horizontal components x_{i1}^s and x_{j1}^s are equal, hence, $x_{i1}^s - x_{j1}^s = 0$.
- Under (ii), θ_{ij} will be scattered within the range from $-180°$ to $180°$, but they will never take the values $\pm 90°$.

Figure 7.3 shows the histogram of θ_{ij}.

The analysis, in the previous paragraph, shows how $\hat{V}$ depends on the angles between points. A further investigation on the effect of the angles between points conditional on the separation between two lines will be carried out in Section 7.5,



(a)

Figure 7.3: The histograms of θ_{ij}.

for $p = 2$.

7.3 Randu-type model

The parallel plane structure in the RANDU data set can be modeled by mixtures of singular p-variate normal distributions.

Let $z \in R^p$ be a p-variate random vector, with mean μ_z and covariance matrix Σ_z, distributed as a mixture of k singular p-variate normal distributions.

For simplicity, assume that the mixture components have equal mixing proportions, $1/k$, and equal within-group covariance matrices, W_z, where W_z is of rank $p - 1$.

The density function of z is given by

$$f(z) = \frac{1}{k} \sum_{j=1}^{k} g(z; \mu_j, W_z), \tag{7.3}$$

where μ_j, $j = 1, \ldots, k$, are the component mean vectors, W_z, the within-group scatter matrix, of rank $p-1$, and g is given by

$$g(z, \mu_j, W_z) = \frac{(2\pi)^{-(p-1)/2}}{(\lambda_1, \ldots, \lambda_{(p-1)})^{1/2}} \exp\{-\frac{1}{2}(z-\mu_j)^T W_z^-(z-\mu_j)\},$$

where W_z^- is the Moore-Penrose generalized inverse, and $\lambda_1, \ldots, \lambda_{p-1}$ are the nonzero eigenvalues of W_z.

The group mean vectors μ_j are assumed to be collinear and equally spaced,

$$\mu_j = j\alpha\nu, \; j = 1, \ldots, k. \tag{7.4}$$

where, $\alpha > 0$, $\nu \notin \text{span}(W)$.

The random vector z can be written as follows

$$z = \epsilon + \alpha r \nu,$$

where $\epsilon \sim N_p(0, W_z)$, r is a random variable distributed as a discrete uniform distribution, takes the values $1, \ldots, k$, and $\nu \in R^p$ is a unit vector.

The total mean vectors is given by

$$\mu = \sum_{j=1}^{k} \mu_j = \frac{1}{2}(k+1)\alpha\nu.$$

The space of z constitutes of k parallel and equally spaced hyperplanes. This model provides a good model for the RANDU data set, and will be called a RANDU-type model.

The total covariance matrix Σ_z is equal to the sum of W_z and the between-group scatter matrix B_z, given as

$$B_z = \frac{(k+1)(k-1)}{12}\alpha^2\nu\nu^T.$$

Since ICS method is an affine invariant transformation, we may, without loss of generality, make the following assumptions:

1. The random vector is rotated, such that

$$\nu^T = e_1^T = (1, 0, \ldots, 0).$$

2. The random vector is standardized with respect to the total covariance matrix Σ_z.

The standardized random vector, denoted by y, can be written as

$$\begin{aligned} y &= \Sigma_z^{-1/2}\epsilon + \Sigma_z^{-1/2}\alpha r e_1, \\ &= u + te_1. \end{aligned} \tag{7.5}$$

where

$$W_y = \Sigma_z^{-1/2} W^{1/2} \Sigma^{-1/2} = \text{diag}(0, 1, \ldots, 1), \tag{7.6}$$

and $t = r/\sqrt{c_k}$,

$$c_k = \frac{(k+1)(k-1)}{12}. \tag{7.7}$$

Since, y is rotated, such that $\nu = e_1$, and standardized, y can be written as follows

$$\begin{pmatrix} y_1 \\ y_2 \\ \vdots \\ y_p \end{pmatrix} = \begin{pmatrix} t \\ u_2 \\ \vdots \\ u_p \end{pmatrix}, \tag{7.8}$$

where $(u_2, \ldots, u_p) \sim N_{p-1}(0, I_{(p-1)})$.

7.4 The behaviour of V in p dimensions

Let y_1 and y_2 be two random p-variate random vectors, distributed as the mixture model (7.3). Without loss of generality, assume that y_1 and y_2 are rotated, such that the line structure direction in the direction of e_1, and standardized as in (7.5).

Consider the difference between y_1 and y_2, $y = y_1 - y_2$. The differenced random vector y has the following density function

$$f^d(y) = \sum_{i,j=1}^{k} \frac{k - |i-j|}{k^2} g(y; \mu_i - \mu_j, 2W_y), \tag{7.9}$$

where W_y =diag$(0, 1, \ldots, 1)$.

The random vector y can be written as in (7.8).

The W-estimate scatter matrix, V, is a weighted covariance matrix, defined as follows

$$V = \mathrm{E}\frac{yy^T}{\{y^T y\}^2}. \tag{7.10}$$

Substituting (7.8) in (7.10) gives

$$V = \mathrm{E}\frac{1}{(t^2 + u_2^2 \ldots + u_p^2)^2} \begin{pmatrix} t^2 & tu_2 & \ldots & tu_p \\ tu_2 & u_2^2 & \ldots & u_p^2 \\ \vdots & \vdots & \vdots & \vdots \\ tu_p & u_p u_2 & \ldots & u_p^2 \end{pmatrix}. \tag{7.11}$$

In order to check the finiteness of the elements of V, i.e to check whether V is well defined or not, we find the trace of V. The trace of V is

$$\mathrm{tr}(V) = \mathrm{E}(wy^T y), \tag{7.12}$$

where $w = w(y) = 1/\{y^T y\}^2$. If V is well defined, then the trace of V will be finite. We show in the following that the converse is true using the Cauchy-

Schwartz inequality; if the trace of V is finite, then the elements of V are finite.

For $a \in R^p$, if $\mathrm{tr}(V) < \infty$, then

$$\mathrm{E}\{wa^T yy^T a\} \leq a^T a \mathrm{E}\{wy^T y\} < \infty,$$

then $\mathrm{E}\{wyy^T\}$ has finite elements.

The trace of V is

$$\begin{aligned} \mathrm{tr}\{V\} &= \mathrm{E}\Big[\frac{1}{y^T y}\Big] \\ &= \mathrm{E}\Big[\frac{1}{y_1^2 + \ldots + y_p^2}\Big]. \end{aligned} \tag{7.13}$$

We show in the following that the expected value of (7.13) depends on p; infinite if $p = 2$, finite if $p > 2$, as shown on the following paragraph.

Let

$$r^2 = y^T y = y_1^2 + y_2^2 + \ldots + y_p^2.$$

The expected value of $1/r^2$ is given by

$$\mathrm{E}\frac{1}{r^2} = \int \frac{1}{r^2} h(r) d(r).$$

We proceed by transforming $y = (y_1, \ldots, y_p)$ to the hyper-spherical coordinates as follows. Let $y = r\omega$, where $\omega = y/|y|$. The Jacobian J is equal to $r^{p-1} q(\omega)$. Hence,

$$\begin{aligned} g(r) &= \int J h(r\omega) q(\omega) d\omega. \\ &= r^{p-1} h(r). \end{aligned}$$

If $h(r) \geq c$, for $0 < r < \varepsilon$, for $c > 0$, and $\epsilon > 0$,

$$\begin{aligned} \mathrm{E}\frac{1}{r^2} &\geq c \int_0^{\varepsilon} \frac{1}{r^2} r^{p-1} dr \\ &= c \int_0^{\varepsilon} r^{p-3} dr \\ &= \infty \quad \text{for } p = 1, 2 \\ &< \infty \quad \text{for } p > 3. \end{aligned} \tag{7.14}$$

So far, we have shown that V is well defined when the dimension $p > 3$. Now, we derive the components of V. From (7.11), the diagonal elements of V are:

$$\begin{aligned} v_{11} &= \mathrm{E}\Big[\frac{t^2}{(t^2 + u_2^2 + \ldots + u_p^2)^2}\Big] \\ &= \mathrm{E} \sum_{i \in t \neq 0} \frac{k - |i|}{k^2} \Big[\frac{i^2}{(i^2 + u_2^2 + \ldots + u_p^2)^2}\Big] \\ &< \frac{k - |i|}{k} \Big[\frac{1}{i^2}\Big] < \infty, \end{aligned}$$

For $l = 2, \ldots, p$,

$$\begin{aligned} v_{ll} &= \frac{1}{k} \mathrm{E}\{\frac{{u_l}^2}{({u_2}^2 + \ldots + u_p)^2}\} + \\ &\quad \frac{k-1}{k} \mathrm{E}\{\frac{{u_l}^2}{(t^2/c_k + {u_2}^2 + \ldots + {u_p}^2)^2}\} \\ &> \frac{1}{k} \mathrm{E}\{\frac{{u_l}^2}{({u_2}^2 + \ldots + u_p)^2}\} \\ &= \infty \qquad \text{if} \quad p \leq 2, \end{aligned} \tag{7.15}$$

The off-diagonal elements are, for $l = 2, \ldots, p$

$$v_{1l} = v_{l1} = \frac{k-1}{k} \mathrm{E}\{\frac{u_l t}{(t^2 + {u_2}^2 + \ldots + {u_p}^2)^2}\}.$$

For $l \neq m = 2, \ldots, p$

$$v_{lm} = \frac{1}{k}\mathrm{E}\{\frac{u_l u_m}{(u_2^2 + \ldots + u_p^2)^2}\} + \frac{k-1}{k}\mathrm{E}\{\frac{u_l u_m}{(t^2 + {u_2}^2 + \ldots + {u_p}^2)^2}\}.$$

When $p = 2$, the eigenvalue associated with the eigenvector in the noise direction has an infinite value. In practice, this can help to distinguish between the noise and the parallel line structure vividly. Thus, V is most powerful when $p = 2$.

Simulation study

In order to compare the effect of the dimension p on the power of V, we conducted a simulation study. The data sets generated from (7.5), with $k = 10$, in dimensions $p = 2, 3, 4, 6$, and sample sizes $n = 40p, 160p, 400p$.

For each choice of n and p the simulation is repeated $m = 1000$ times. The accuracy is measured by the mean of axis squared distance, defined in (3.68),

$$v(\hat{\gamma}_p) = \frac{1}{m}\sum_{i=1}^{m}\{1 - (\hat{\gamma}_p^T a)^2\},$$

where $\hat{\gamma}_p$ is the smallest eigenvector of $\hat{V}$, and $a^T = (1, 0, \ldots, 0) \in R^p$ is a unit vector defining the true line structure direction. We also record the averages of the ratios between the largest to smallest eigenvalues.

The results of the simulations are shown in Table 7.1. From Table 7.1, ICS^d:W-estimate:variance is more accurate when $p = 2$ and $p = 3$. For $p = 4$ and 6, the method does not work.

7.5 A detailed analysis of V in two dimensions

In Section 7.2, we have shown how the dominant eigenvector of $\hat{V}$ of a two dimensional subset from the RANDU data set depends on the distribution of the angles between point. In this section, we will investigate how the angles between points conditional on the horizontal separation affect V, under model (7.3), when $p = 2$. In particular, we want to study the modes of the angular distributions

Table 7.1: The means of axis squared distance of the smallest eigenvector of $\hat{V}$ for simulated bivariate data consists $k = 10$ parallel lines, with dimensions $p = 2, 3, 4, 6$, and sample sizes $40p, 160p, 400p$.

p	n	λ_p/λ_1	$v(\hat{\gamma}_p)$
2	80	5.11e+03	0.0013
	320	6.07e+07	0.0003
	800	2.46e+09	0.0001
3	120	2.304	0.556
	480	4.136	0.004
	1200	3.51	0.001
4	160	1.419	0.85
	640	1.0982	0.789
	1600	1.166	0.571
6	240	1.4538	0.915
	960	1.177	0.956
	2400	1.118	0.975

conditional on the horizontal separations. The angular distribution for model (7.3) is the projected normal distribution.

To gain insight, we use the wrapped Cauchy distribution to model the angles between points conditional on the horizontal separation. The wrapped Cauchy model is an artificial model, but used as a substitute to the projected normal because of its analytical tractability.

7.5.1 Projected normal model analysis

Let $z_1 = (j_1, u_1)^T$ and $z_2 = (j_2, u_2)^T$, for some $j_1, j_2 \in R^+$, u_1 and u_2 are two independent random variables distributed as normal distribution, $N(0, 1)$.

The difference $z = z_1 - z_2$ can be written as $z = (j, u)^T$ where $j = (j_1 - j_2) \in R$, $u = (u_1 - u_2)$ is distributed as $N(0, 2)$.

To find the distribution of the angle between z_1 and z_2, let

$$\theta = 2\tan^{-1}(u/|j|). \tag{7.16}$$

If j is small, the density of θ will be bimodal at $\pm 180°$. If j is large, the density of θ will be unimodal at $0°$.

The density function of θ (for $j \not= 0$) is given by

$$g_j(\theta) = \frac{|j|}{2\sqrt{\pi})(1+\cos(\theta))} \exp\{-\frac{j^2}{4}\frac{1-\cos(\theta)}{1+\cos(\theta)}\}. \tag{7.17}$$

Proof. Using the change of variable formula, from (7.16),

$$u = |j| \tan(\theta/2).$$

Differentiating u with respect to θ gives

$$\frac{\partial u}{\partial \theta} = \frac{|j|}{2} \sec^2(\theta/2). \tag{7.18}$$

The density function of θ is given by

$$g_j(\theta) = \frac{\partial u}{\partial \theta} f(|j| \tan\frac{\theta}{2}), \tag{7.19}$$

where f is the normal density function with mean 0 and variance 2. Substituting by f, and (4.18) in (4.19) gives

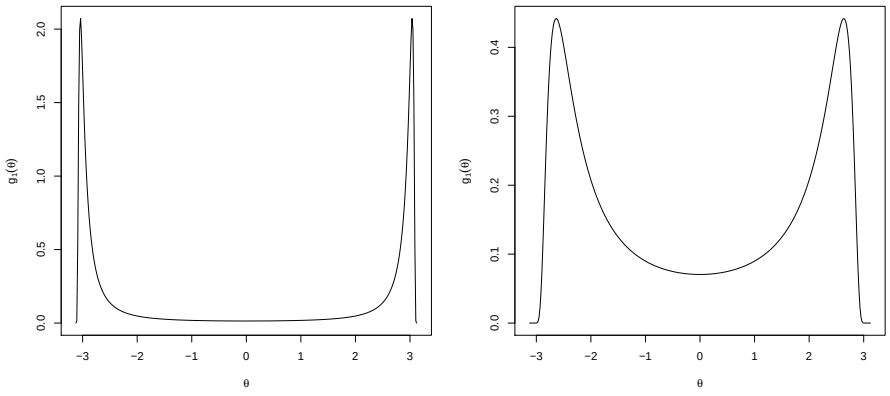
$$\begin{aligned} g_j(\theta) &= \frac{|j|}{2}\sec^2(\theta/2)\frac{1}{2\sqrt{\pi}}\exp\{-\frac{j^2}{4}\tan^2(\theta/2)\} \\ &= \frac{|j|}{2\sqrt{\pi}(1+\cos\theta)}\exp\{-\frac{j^2}{4}\frac{1-\cos\theta}{1+\cos\theta}\}. \end{aligned}$$

□

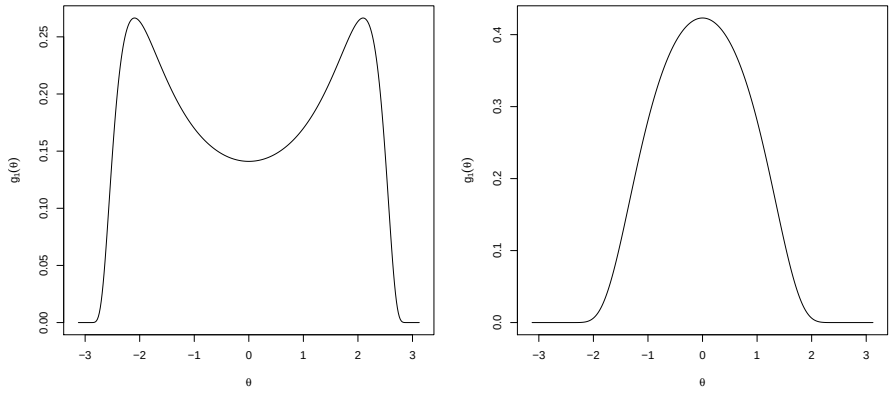
The density plots are shown in Figure 7.4, $j = 0.1, 0.5, 1, 3$.

Let Z be an $n \times p$, $n = 200$, be a data matrix. Each row of Z can be written as $z_i = (r_i, u_i)$, where r_i is an integer that takes two possible values j_1 and j_2 with equal probabilities, and the difference $(j_1 - j_2) = j$, and $u_i \sim N(0, 1)$.

We computed the pairwise differencing of the data points between lines, $z_i -$



(a) $|j| = 0.1$ (b) $|j| = 0.5$



(c) $|j| = 1$ (d) $|j| = 3$

Figure 7.4: The density plots of the projected normal density function $g_j(\theta)$, for $|j| = 0.1, 0.5, 1, 3$.

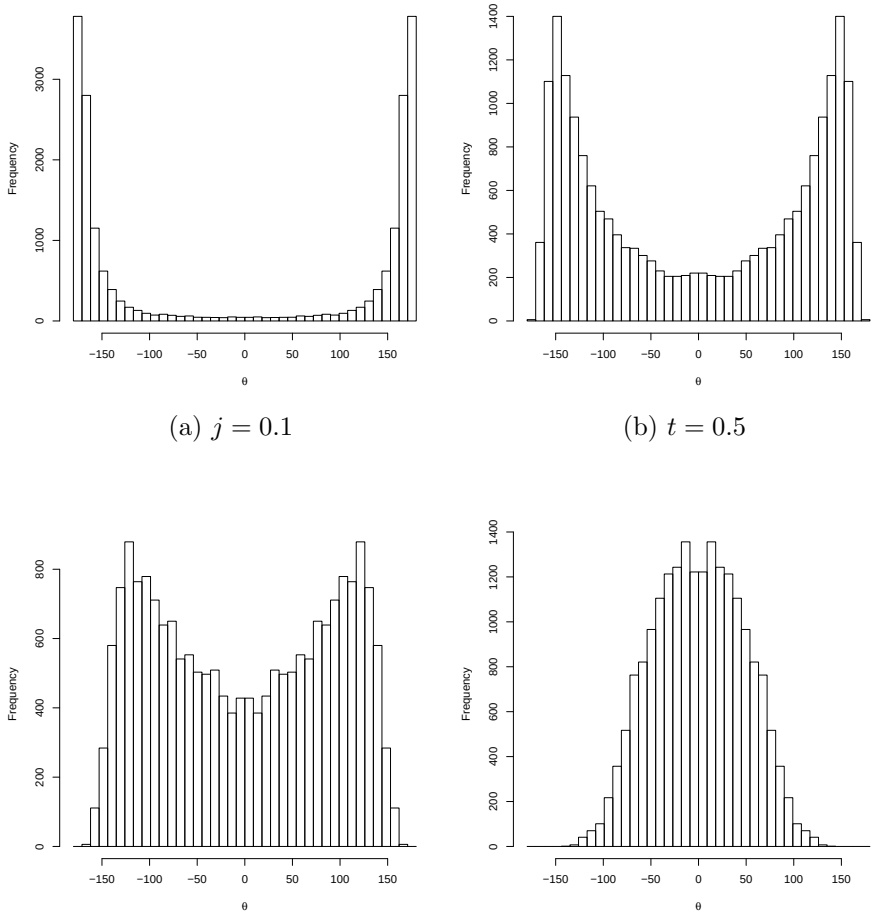
$z_k = (\pm j, u_i - u_k)$. We compute θ_{ik}, as defined in (7.16),

$$\theta_{ik} = 2\text{atan}\{(u_i - u_k)/j\}.$$

The histograms of θ_{ik} are shown in Figure 7.5 for $|j| = 0.1, 0.5, 1$, and 3.

The plots show that when the lines are close to each other, as in Figure 7.5 (a), the distribution of the angles will have two modes at ± 180. We have discussed earlier in Section 7.2 that the the overall angles between points will have bimodal distribution at $\pm 180^\circ$ (if we double the angle). This means that if we have only two lines close to each other the differencing within each line and between lines

will be indistinguishable. As the separation between lines increases, the histogram tends to a normal distribution.



(a) $j = 0.1$ (b) $t = 0.5$

(c) $j = 1$ (d) $j = 3$

Figure 7.5: Histograms of θ of data points simulated from $N(0, 1)$ lying on two parallel lines separated by $j = 0.1, 0.5, 1, 3$.

7.5.2 Cauchy model

To gain more insight into the change of mode of the distribution of θ, conditionally to the spacing between lines, we use an artificial Cauchy model.

It is hard to study the relation between the line spacing and the modal axis in the normal model analytically. So, we model the points along the lines using a Cauchy distribution and the angles between points using a wrapped Cauchy.

Although second moments do not exist for the Cauchy distribution, the Cauchy distribution is used because the wrapped Cauchy has an explicit form.

Cauchy and wrapped Cauchy

Let s be a random variable distributed as a Cauchy distribution, $C(\mu, c)$, with density function

$$f(s) = \frac{1}{\pi} \frac{c}{c^2 + (s-\mu)^2}. \tag{7.20}$$

It can be shown that the wrapped Cauchy distribution can be obtained using a one-to-one transformation as follows, Mardia and Jupp (2009),

$$\theta = 2\tan^{-1}(s-\mu), \qquad \theta \in (-\pi, \pi). \tag{7.21}$$

The wrapped Cauchy distribution with parameters $\mu_c = 0$, and $-1 < \rho_c < 1$, is given by

$$h(\theta) = \frac{\sqrt{1-\rho_c^2}}{2\pi(1-\rho_c\cos(\theta))}, \qquad 0 \leq \theta < 2\pi. \tag{7.22}$$

where

$$\rho_c = \frac{1-c^2}{1+c^2}. \tag{7.23}$$

There are three different cases:

- If $c = 1$, $\rho_c = 0$, and $h(\theta)$ will be the uniform distribution.
- If $c < 1$, $\rho_c > 0$, θ will be distributed as $C(0, \rho_c)$.
- If $c > 1$, $\rho_c < 0$, θ will be distributed as $C(\pi, |\rho_c|)$.

As $\rho_c \to 0$, $h(\theta)$ tends to the uniform distribution, as $|\rho_c| \to 1$, $h(\theta)$ becomes concentrated at 0 or π.

Proof. Using the change of variable technique to find the density function of θ, from (7.21)

$$(s-\mu) = \tan(\theta/2). \tag{7.24}$$

Differentiating with respect to θ

$$\begin{aligned}\frac{\partial s}{\partial \theta} &= \frac{1}{2}\sec^2\frac{\theta}{2} \\ &= \frac{1}{2\cos^2\frac{\theta}{2}} \\ &= \frac{1}{(1+\cos\theta)}.\end{aligned} \tag{7.25}$$

The density function of θ is given by,

$$\begin{aligned}h(\theta) &= \frac{\partial s}{\partial \theta} f(\tan\frac{\theta}{2}) \\ &= \frac{c}{\pi(1+\cos\theta)\{c^2+\tan^2(\theta/2)\}} \\ &= \frac{2c}{2\pi(1+\cos\theta)\{c^2+\frac{1-\cos\theta}{1+\cos\theta}\}} \\ &= \frac{2c}{2\pi\{c^2(1+\cos\theta)+(1-\cos\theta)\}} \\ &= \frac{2c}{2\pi\{(c^2+1)+(c^2-1)\cos\theta\}} \\ &= \frac{\frac{2c}{c^2+1}}{2\pi\{1+\frac{(c^2-1)}{(c^2+1)}\cos\theta\}} \\ &= \frac{\frac{2c}{1+c^2}}{2\pi\{1-\frac{(1-c^2)}{(1+c^2)}\cos\theta\}}.\end{aligned} \tag{7.26}$$

Let ρ takes the form

$$\rho_c = \frac{1-c^2}{1+c^2}. \tag{7.27}$$

Substituting by ρ_c in (7.26) gives the density of the wrapped Cauchy in (7.22). □

Cauchy analysis of RANDU

Let $z_1 = (j_1, u_1)^T$ and $z_2 = (j_2, u_2)^T$, for some $j_1, j_2 \in \{1, \ldots, k\}$, u_1 and u_2 are two independent random variables distributed as Cauchy distribution, $C(\mu, c)$, with density function (7.20).

Consider the difference $z = z_1 - z_2$, where z can be written as $z = (j, u)^T$ where $j = (j_1 - j_2) \in \{0, \pm 1, \ldots, \pm(k-1)\}$, $u = (u_1 - u_2)$ is distributed as

$C(0, 2c)$.

As in (7.23), wrapping z is obtained using the following transformation

$$\theta = 2\tan^{-1}\frac{u}{j}. \tag{7.28}$$

The density function of θ is given by

$$h_j(\theta) = \frac{\sqrt{1-\rho_{jc}^2}}{2\pi(1-\rho_{jc}\cos(\theta))}, \tag{7.29}$$

where ρ_j is given by

$$\rho_{jc} = \frac{j^2-4c^2}{j^2+4c^2}. \tag{7.30}$$

As shown in (7.23), $-1 \leq \rho_{jc} < 1$. The density $h_j(\theta)$ depends on j, as follows:

(1) If $j = 0$, then $\rho_{jc} = -1$ and $h_j(\theta)$ will be concentrated at π.

(2) If $j > 2c$, $\rho_{jc} > 0$, and $h_j(\theta)$

$$\begin{aligned} h_j(\theta) &\propto \frac{1}{1-\rho_{jc}\cos(\theta+\pi)} \\ &= WC(0, \rho_{jc}). \end{aligned}$$

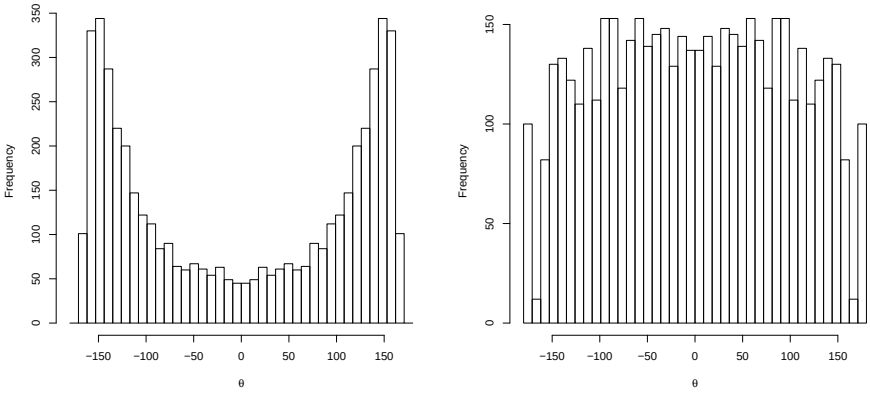
(3) If $j < 2c$, $\rho_{jc} < 0$, and the density of θ

$$\begin{aligned} h_j(\theta) &\propto \frac{1}{1-|\rho_{jc}|\cos(\theta)} \\ &= WC(\pi, |\rho_{jc}|). \end{aligned}$$

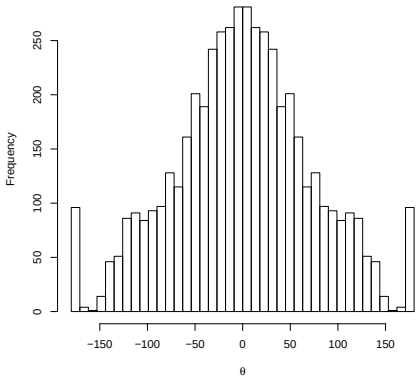
(4) At $j = 2c$, $\rho_{jc} = 0$, then θ will be uniformly distributed.

Let Z,be an 200×2 data matrix. Each row of Z is $z_i^T = (r_i, y_i)$, where r_i takes two values j_1 and j_2 with equal probabilities, such that

$$j_1 - j_2 = j,$$



(a) $j = 0.5$, $c = 1$ (b) $j = 2$, $c = 1$



(c) $j = 4$, $c = 1$

Figure 7.6: Histograms of θ of data points simulated from $C(0, 1)$ lying on two parallel lines separated by $j = 0.5, 2, 4$.

and y_i is generated from $C(0, 1)$.

We computed the pairwise differencing of the data points between lines, $z_i - z_k = (\pm j, y_i - y_k)$. We compute θ_{ik}, as defined in (7.28).

The histograms of θ_{ik} are shown in Figure 7.6 for $j = 0.5, 2$, and 4. Figure 7.6 shows that when the line separation equal to a threshold $j = 2c$, then the points will be uniformly distributed.

7.6 The behaviour of V under mixtures of normal distributions

Suppose that the random vector y, defined in (7.8), has some noise added in the horizontal direction with variance w. In this case y can be written as

$$\begin{pmatrix} y_1 \\ y_2 \\ \vdots \\ y_p \end{pmatrix} = \begin{pmatrix} \delta r + w^{1/2} u_1 \\ u_2 \\ \vdots \\ u_p \end{pmatrix}, \tag{7.31}$$

where $r = 1, \ldots, k$ with equal probabilities, $u^T = (u_1, u_2, \ldots, u_p) \sim N(0, I_p)$, $0 \leq \delta \leq 1$ is a separation parameter between groups and its value depends on k, and $w = 1 - \delta^2$. The total covariance matrix is the identity matrix $\Sigma_y = I_p$.

For $k = 2$, the model will be the two-group mixture model explained in Section 3.2.

Let y_1 and y_2 be two random variables defined as in (7.31). Consider the difference $y = y_1 - y_2$. The W-estimate is defined as in (7.10).

Following the same calculations of Section 7.4. The elements of V are given as follows

- The diagonal elements:

$$v_{11} = \frac{1}{k}\mathrm{E}\{\frac{w{u_1}^2}{(wu_1^2 + {u_2}^2 + \ldots + u_p)^2}\} + \frac{k-1}{k}\mathrm{E}\{\frac{(\sqrt{w}u_1 + \delta r)^2}{((\sqrt{w}u_l + \delta r)^2 + {u_2}^2 + \ldots + {u_p}^2)^2}\}.$$

For $l = 2, \ldots, p$,

$$v_{ll} = \frac{1}{k}\mathrm{E}\{\frac{{u_l}^2}{(wu_1^2 + {u_2}^2 + \ldots + u_p)^2}\} + \frac{k-1}{k}\mathrm{E}\{\frac{{u_l}^2}{((\sqrt{w}u_l + \delta r)^2 + {u_2}^2 + \ldots + {u_p}^2)^2}\}.$$

- The off diagonal elements: for $l = 2, \ldots, p$

$$v_{1l} = v_{l1} = \frac{1}{k}\mathrm{E}\frac{\sqrt{w}u_1u_l}{(wu_1^2 + u_2^2 + \ldots + u_p^2)^2} + \frac{k-1}{k}\mathrm{E}\{\frac{u_l(\sqrt{w}u_1 + \delta r)^2}{((\sqrt{w}u_l + \delta r)^2 + {u_2}^2 + \ldots + {u_p}^2)^2}\}.$$

For $2 \leq l \neq m \leq 2$

$$v_{lm} = \frac{1}{k}\mathrm{E}\{\frac{u_lu_m}{(wu_1^2 + u_2^2 + \ldots + u_p^2)^2}\} + \frac{k-1}{k}\mathrm{E}\{\frac{u_lu_m}{((\sqrt{w}u_l + \delta r)^2 + {u_2}^2 + \ldots + {u_p}^2)^2}\}.$$

The value of v_{11} will be the smallest value, but it will become closer to the values of other elements as w takes values far from zero.

The case of $\delta = 0$ is the case of one isotropic group. In this case, $w = 1$. In this case V will be proportional to the identity. As δ goes near zero the separation between v_{11} and v_{ll} increases.

Simulation study

Consider data sets distributed as mixtures of two equal bivariate normal distributions, generated from (7.31), with $r = \pm 1$, and $\delta = 1, 0.99, 0.9, 0.7$, with $\omega = 0, 0.0199, 0.19, 0.51$, respectively. For each value of ω, the simulation is repeated 2000 times with sample sizes $n = 20, 50, 200, 500$. We use (3.65) to measure the accuracy.

The simulation results, Table 7.2, shows that $\hat{V}$ breaks down even for small w.

Table 7.2: The accuracy of the estimate of ICS^d:W-estimate:variance, for two-group data, for $\delta = 1, 0.99, 0.9, 0.7$, ($\omega = 0, 0.0199, 0.19, 0.51$).

δ	n	$v(\hat{\theta})$
1	20	2.83e-06
	50	2.33e-08
	200	2.25e-11
	500	1.59e-13
0.99	20	0.385
	50	0.365
	200	0.372
	500	0.341
0.9	20	0.495
	50	0.489
	200	0.464
	500	0.467
0.7	20	0.497
	50	0.507
	200	0.489
	500	0.507

Chapter 8

Conclusions, and potential applications

In this chapter, we discuss the thesis results, and suggest some potential applications.

8.1 Conclusions

The goal of this thesis was mainly to understand why ICS works in some situations and does not in others. Our main results are summarized as follows:

8.1.1 ICS vs. PP

We have compared ICS and PP under two-group mixtures of normal distributions with equal covariance matrices.

First, we considered ICS based on the fourth-order moments matrix, $\hat{K}$, and the covariance matrix, S, and PP based on the univariate kurtosis. Under two-group mixtures, we found explicit formulas for ICS based on $\hat{K}$ and S, and PP based on based on kurtosis. We also derived the asymptotic distributions of the ICS and PP estimates of the group separation direction.

The simulation results show that PP based on kurtosis is more accurate than

ICS based on $\hat{K}$ and S. The asymptotic results largely agree with the simulations.

We have compared three different ICS and PP criteria based on robust measures of spread. The results show that robust ICS seems to be more feasible.

8.1.2 Common location measures

When applying ICS and PP based on robust measures of spread, two problems arise. These two problems have a substantial effect the performances of robust ICS and PP. The first problem is the criteria has a local maximum when it supposed to be a local minimum, because the pairs of spread measures were computed based on two different location measures. The second problem is the separation between the smallest and largest value of the criteria is small, and due to sampling variation.

We have explored two possible solutions to improve the performances of ICS and PP: computing the pair of spread measures based on a common location measure, the second is by computing the pair of spread measures based on the pairwise differencing of the data to force the symmetry of the data around the origin. Our simulation results suggest that using a common location measure and pairwise differencing of the data are not always useful.

8.1.3 The role of differencing

Another situation where we looked at the use of pairwise differencing of the data is when the underlying structure has a parallel line structure.

In this case, ICS based on $\hat{K}$ and S fails to find the structure direction, whether the fourth order moment matrix were computed based on the original data or the pairwise differencing of the data.

Applying ICS based on the W-estimate, $\hat{V}$, and S, computed with the pairwise differencing of the data works well. However, if some noise is added to the lines horizontally, ICS based on $\hat{V}$ and S with the pairwise differencing of the data

breaks down.

8.1.4 A new insight into the parallel line structure

We have gained an insight into the effect of the separation between points, overall and conditional on the horizontal separation, on the power of the ICS based $\hat{V}$ and S.

We explored the distribution of the angles between points. We have shown that, when $S \propto I$, the dominant eigenvector of $\hat{V}$ is in the direction of the mode of the distribution of the angles.

We also have found a threshold for the separation between two lines in which the distribution of the angles between points, conditional on the horizontal separation, will be uniform.

8.1.5 A new insight into the errors in variables

We have explored the use of ICS, based on $\hat{K}$ and S, in the errors-in-variables model, when the signal has a non-normal distribution. We found the form of the signal eigenvalue and its corresponding eigenvector, which can give an insight of the way ICS finds the signal direction.

We have also compared the accuracy of the ICS estimates and Geary's fourth-order cumulant-based estimates. Although some of fourth Ceary's estimate are more accurate than ICS estimates, they are lacking the affine equivariance property.

8.2 Applications of ICS

In this section, we give two tentative new applications of ICS. ICS can be used in the construction of principal curves, and the analysis of fingerprint images.

8.2.1 Principal curves

The majority of algorithms used to construct principal curves, introduced by Hastie and Stuetzle (1989), work by applying principal component analysis iteratively. At each iteration, only points that are close to a current point are considered. The radius of the circle, h, say, that covers the points around the current point needs to be specified in advance. The choice of the tuning parameter h has a substantial effect on the performance of principal curve algorithms.

If h is chosen small enough such that the local covariance matrix is proportional to the identity matrix, the algorithm will not work well. In this case ICS can play a role if applied locally.

In this section we consider a recent principal curve algorithm by Einbeck et al. (2005). The algorithm can be explained as follows. Given bivariate data, choose a random starting point, and draw a circle around this point with radius h. After that compute the local mean, and update it in the direction of the local first principal component.

Let X be an $n \times 2$ data matrix, where each row can be written as $x_i^T = (x_{i1}, x_{i2})$, $i = 1, \ldots, n$. Let $K_h(\cdot)$ be the flat kernel function defined as follows

$$K_h(x_i - x) = \begin{cases} 1 & \text{if } \parallel x_i - x \parallel < h \\ 0 & \text{otherwise,} \end{cases}$$

Define $w_i^x = K_h(x_i - x) / \sum_{i=1}^n K_h(x_i - x)$. Einbeck et al. (2005) algorithm is given by

(1) Choose a point $x_\circ$ as a starting point. Set $x = x_\circ$

(2) Calculate the local center of mass

$$\mu^x = \sum_{i=1}^n w_i^x x_i \tag{8.1}$$

at x.

(3) Estimate the local covariance matrix $S^x = (s_{jk})$, where

$$s_{jk} = \sum_{i=1}^{n} w_i^x (x_{ij} - \mu_j^x)(x_{ik} - \mu_k^x).$$

(4) Find the first principal component of S^x, denoted by γ^x.

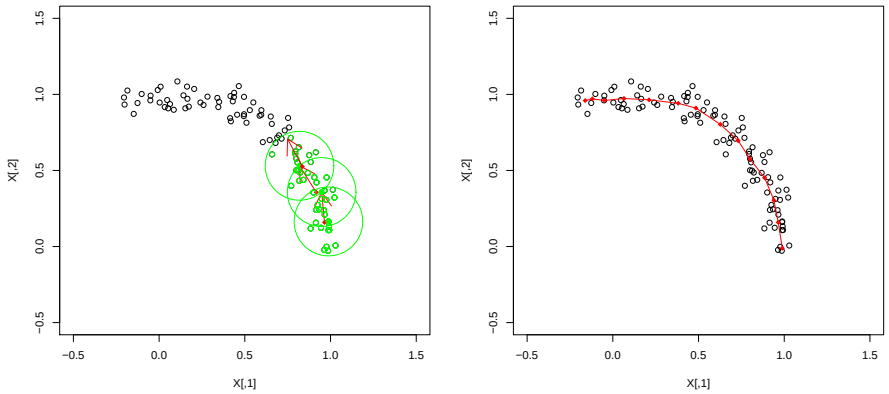
(5) Update x by $x = \mu^x + l\gamma^x$, where l is the step length.

(6) Repeat steps (2)-(5) until the sequence of μ^x remains constant. Set $x = x_\circ$, $\gamma^x = -\gamma^x$ and continue.

For example, let $x_1, \ldots, x_{100}$ be bivariate points generated using the following model

$$\begin{pmatrix} x_1 \\ x_2 \end{pmatrix} = \begin{pmatrix} \cos(\theta) \\ \sin(\theta) \end{pmatrix} + \begin{pmatrix} \epsilon_1 \\ \epsilon_2 \end{pmatrix}, \tag{8.2}$$

where $0 < \theta < \pi/2$, $\epsilon^T = (\epsilon_1, \epsilon_2) \sim N(0, \sigma^2 I_2)$, with $\sigma^2 = 0.06$. Figure 8.1 (a) shows an illustration of Einbeck's algorithm algorithm for this example, (b) shows the final principal curve.



(a) (b)

Figure 8.1: (a) Illustration of the algorithm with $h = 0.2$ and $l = 0.2$, and (b) the local means that form the principal curve.

Now we give an example that illustrates when ICS can help to construct the principal curve. For example, consider the data shown in Figure 8.2. The data

consist of two parallel and closely spaced lines, each line generated from (8.2), with separation equal to 0.07, and $\sigma^2 = 0.01$.

At a selected iteration, shown in Figure 8.2 (a), the eigenvalues and eigenvectors of the local covariance matrix are approximately equal to,

$$l_1 = 0.0027, l_2 = 0.0024,$$
$$\gamma_1^T = (-0.99, 0.07), \gamma_2^T = (-0.07, -0.99).$$

The eigenvalues l_1 and l_2 are close to each other. This means that applying local PCA is meaningless at this iteration.

Let $S^{x^{-1}}$ and K^x be the local covariance and fourth-order moment matrix, respectively, where the fourth-order moment matrix is defined in (3.23). The eigenvalues and eigenvectors of $S^{x^{-1}}K^x$ are given by,

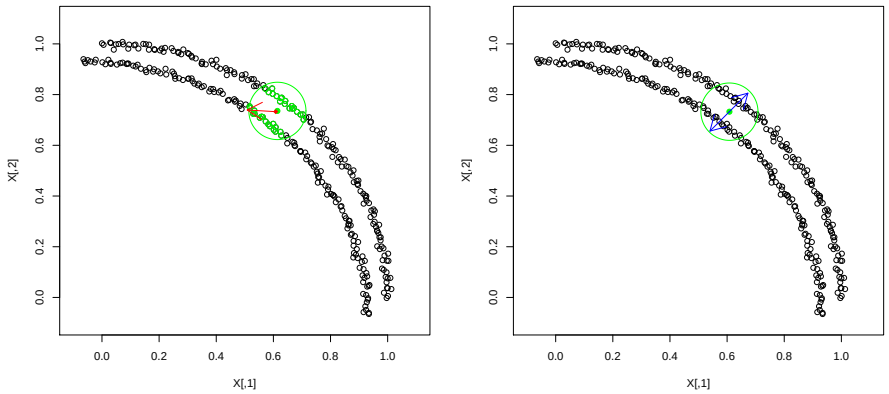
$$a_1 = 0.689, a_2 = 0.494,$$
$$u_1^T = (0.774, -0.632), u_2^T = (0.664, 0.748).$$

As we have shown in Chapter 3 that if we have two equal groups, the smallest eigenvector u_2 is in the group separation direction, whereas the largest eigenvector is in the direction of the normal noise. Figure 8.2 (b) shows u_2.

8.2.2 Fingerprint images

Fingerprint identification is based on information extracted from fingerprint images. The most significant extracted information is the ridge type.

For example, consider the fingerprint image shown in Figure 8.3, which is taken from Dario et al. (2002) database. From Figure 8.3, when the ridges are parallel, the structure resembles the parallel line structure in the RANDU data set, defined in Chapter 7. A subset from Figure 8.3 that shows the parallel ridges as in Figure 8.4.



(a) PCA (b) ICS

Figure 8.2: Illustration of the solutions of the local PCA and ICS at a selected iteration, for two parallel curves with step length $l = 0.1$, and radius $h = 0.1$.

Applying ICS^d:W-estimate:variance can find the direction of the line structure, as shown in Figure 8.5. The eigenvectors and eigenvalues of $S^{-1}\hat{V}$ are given as follows

$$l_1 = 3.513, \quad l_2 = 2.222,$$

$$u_1^T = c(-0.653, -0.757), \quad u_2^T = (-0.689, 0.724).$$

Figure 8.3: A full fingerprint image of dimension 379×388.

Figure 8.4: A subset, of dimension 40×40, from the full fingerprint image that shows the parallel ridges.

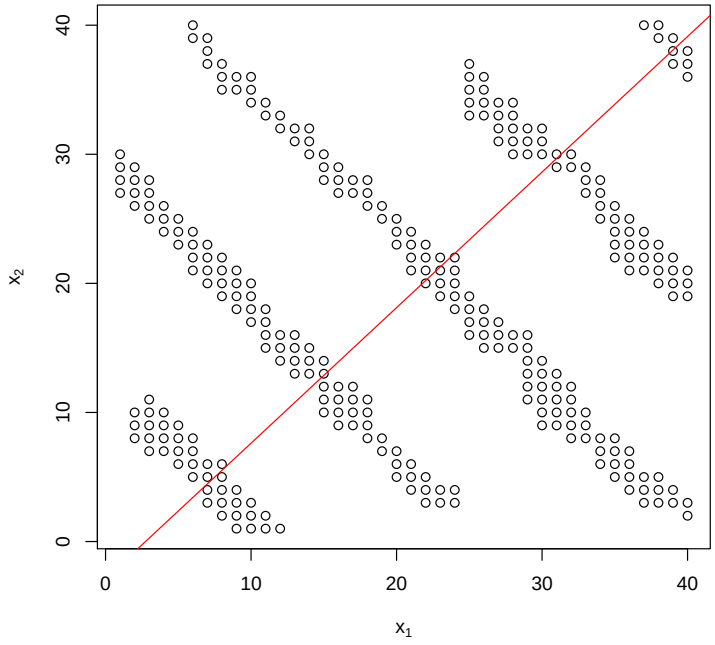


Figure 8.5: The direction of the smallest eigenvector of $S^{-1}\hat{V}$, for the parallel line structure.

Bibliography

Maio Dario and Davide Maltoni and Raffaele Cappelli and J. L. Wayman and Anil K Jain. The Second International Competition for Fingerprint Verification Algorithms website. `http://bias.csr.unibo.it/fvc2002/`. 2002.

David F Andrews, PJ Bickel, FR Hampel, PJ Huber, WH Rogers, and JW Tukey. Robust estimates of location: survey and advances. 1972.

Olcay Arslan, Patrick DL Constable, and John T Kent. Convergence behavior of the em algorithm for the multivariate t-distribution. *Communications in statistics-theory and methods*, 24(12):2981–3000, 1995.

Jamal B Bugrien and John T Kent. Independent component analysis: An approach to clustering. *signs*, 98:8a, 2005.

John G Cragg. Using higher moments to estimate the simple errors-in-variables model. *Rand Journal of Economics*, pages S71–S91, 1997.

Christophe Croux, Peter Filzmoser, and M Rosario Oliveira. Algorithms for projection–pursuit robust principal component analysis. *Chemometrics and Intelligent Laboratory Systems*, 87(2):218–225, 2007.

Jochen Einbeck, Gerhard Tutz, and Ludger Evers. Local principal curves. *Statistics and Computing*, 15(4):301–313, 2005.

John HJ Einmahl, Maria Gantner, and Günther Sawitzki. The shorth plot. *Journal of Computational and Graphical Statistics*, 19(1), 2010.

Brian Everitt. *An R and S-PLUS companion to multivariate analysis*. Springer, 2005.

Jerome H Friedman and John W Tukey. A projection pursuit algorithm for exploratory data analysis. *Computers, IEEE Transactions on*, 100(9):881–890, 1974.

Robert C Geary. Inherent relations between random variables. In *Proceedings of the Royal Irish Academy. Section A: Mathematical and Physical Sciences*, volume 47, pages 63–76. JSTOR, 1941.

Jonathan Gillard. An overview of linear structural models in errors in variables regression. *REVSTAT–Statistical Journal*, 8(1):57–80, 2010.

R Grubel. The length of the shorth. *The Annals of Statistics*, pages 619–628, 1988.

Trevor Hastie and Werner Stuetzle. Principal curves. *Journal of the American Statistical Association*, 84(406):502–516, 1989.

Peter J Huber. Projection pursuit. *The annals of Statistics*, pages 435–475, 1985.

M Chris Jones and Robin Sibson. What is projection pursuit? *Journal of the Royal Statistical Society. Series A (General)*, pages 1–37, 1987.

Jana Jureckova and Jan Picek. *Robust statistical methods with R*. Chapman and Hall/CRC, 2005.

M.G. Kendall and A. Stuart. *The Advanced Theory of Statistics (volume 1)*. Charles Griffin (4th Edition), 1977.

M.G. Kendall and A. Stuart. *The Advanced Theory of Statistics (volume 2)*. Charles Griffin (4th Edition), 1979.

John T Kent and David E Tyler. Constrained m-estimation for multivariate location and scatter. *The Annals of Statistics*, 24(3):1346–1370, 1996.

Albert Madansky. The fitting of straight lines when both variables are subject to error. *Journal of the American Statistical Association*, 54(285):173–205, 1959.

Kanti V Mardia and Peter E Jupp. *Directional statistics*, volume 494. Wiley, 2009.

Kantilal Varichand Mardia, John T Kent, and John M Bibby. *Multivariate analysis*. Academic press, 1980.

Ricardo A Maronna, R Douglas Martin, and Victor J Yohai. *Robust statistics*. J. Wiley, 2006.

Klaus Nordhausen, Hannu Oja, and David E Tyler. Tools for exploring multivariate data: the package ics. *Journal of Statistical Software*, 28(6):1–31, 2008.

Manoranjan Pal. Consistent moment estimators of regression coefficients in the presence of errors in variables. *Journal of Econometrics*, 14(3):349–364, 1980.

Daniel Peña and Francisco J Prieto. Cluster identification using projections. *Journal of the American Statistical Association*, 96(456):1433–1445, 2001.

Daniel Peña, Francisco J Prieto, and Júlia Viladomat. Eigenvectors of a kurtosis matrix as interesting directions to reveal cluster structure. *Journal of Multivariate Analysis*, 101(9):1995–2007, 2010.

Robert J Serfling. *Approximation theorems of mathematical statistics*. Wiley-Interscience, 1980.

Peter Sprent. *Models in regression and related topics*. Methuen London, 1969.

D. E. Tyler, F. Critvhly, L. Dumbgen, and H. Oja. Invariant co-ordinate selection. *Royal Statistical society*, 71:549–592, 2009.

Stefan Van Aelst and Peter Rousseeuw. Minimum volume ellipsoid. *Wiley Interdisciplinary Reviews: Computational Statistics*, 1(1):71–82, 2009.

WN Venables and BD Ripley. Package mass. *Online at: http://cran. r-project. org/web/packages/MASS/MASS. pdf*, 2010.