The University of Auckland

Doctoral Thesis

Large-scale Inference Using Non-parametric Mixtures

Author:
Xiangjie Xue

Supervisor:
Dr. Yong Wang

A thesis submitted in fulfillment of the requirements for the degree of Doctor of Philosophy

in the

Department of Statistics

The University of Auckland

March 4, 2021

Abstract

A new approach to solving large-scale problems using non-parametric mixtures is proposed. The new approach not only takes the advantage of the great flexibility in non-parametric mixture methods and the fast computation of the non-parametric maximum likelihood estimate of a mixing distribution using the constrained Newton method proposed by Wang (2007), but it is also able to make use of the common features among observations in order for a more efficient and accurate estimation.

This new approach is then applied to several practical fields. In the context of multiple hypothesis testing, a new method is proposed to compute the NPMLE given a fixed probability at the location of the null effect, and then is extended to estimate the proportion of null effects based on various threshold functions. Distance-based counterparts under the Cramér-von Mises and the Anderson-Darling distance are also introduced. Furthermore, modifications are also made to the likelihood- and distance-based methods, and hence they can be used in large-scale computation. A new procedure that controls the false discovery rate, based on the estimated null proportion and its estimated mixing distribution, is also introduced. Numerical studies show that the estimators of the null proportion using the non-parametric maximum likelihood estimators and their minimum distance counterparts (the non-parametric minimum distance estimators) perform well and the proposed controlling procedure makes more rejections than existing methods given a pre-specified level.

A new method based on this new approach is also proposed for covariance matrix estimation. A new covariance matrix estimator is constructed using the empirical Bayes on the sample correlation coefficients or the transformed sample correlation coefficients. The estimated density required by the proposed estimator can be computed using the non-parametric maximum likelihood estimators or the non-parametric minimum distance estimators. Estimation under the sparsity assumption is also discussed. The numerical studies using the simulated and the real world datasets all suggest that the proposed covariance matrix estimator performs well and can be applied to a wide range of covariance structures.

Acknowledgements

I would like to thank my supervisor, Dr. Yong Wang, for the continuous help, guidance and support throughout the duration of my PhD. His willingness to give his time so generously has been very much appreciated.

I would like to thank New Zealand eScience Infrastructure (NeSI) for providing high performance computing services. I could not have obtained such a large amount of simulated results for analysis without it.

I would like to thank the university for providing the University of Auckland Doctoral Scholarship.

Last but not least, I would like to thank my parents and all my family for their kind support and continuous motivation.

Contents

List of Figures

List of Tables

List of Abbreviations

AIC	Akaike Information Criterion
AS	The Adaptive Step-down procedure
BH	The Benjamini-Hochberg procedure
BIC	Bayesian Information Criterion
CDF	Cumulative Distribution Function
ECDF	Empirical Cumulative Distribution Function
FDP	False Discovery Proportion
fdr	local false discovery rate
FDR	False Discovery Rate
FWER	Family-Wise Error Rate
GLRT	Generalised Likelihood Ratio Test
HQC	Hannan-Quinn Information Criterion
HT	Hard thresholding
KL	Kullback–Leibler (a measure of divergence)
MCP	Multiple Comparison Procedures
MSE	Mean Squared Error
NASDAQ	National Association of Securities Dealers Automated Quotations (a stock market)
NPMDE	Non-Parametric Minimum Distance Estimator
NPMLE	Non-Parametric Maximum Likelihood Estimator
PCER	Per-Comparison Error Rate
pFDR	positive False Discovery Rate
RBLW	the Rao-Blackwell Ledoit-Wolf estimator
SE	Standard Error
ST	Soft thresholding

List of Symbols

$\boldsymbol{X}$	matrix of the appropriate size or random vector; see Section 1.5
$\boldsymbol{x}_{i,\cdot}$	the i-th row of $\boldsymbol{X}$
$\boldsymbol{x}_j, \boldsymbol{x}_{\cdot,j}$	the j-th column of $\boldsymbol{X}$
$x_{ij}, x_{i,j}$	the (i,j)-th element of $\boldsymbol{X}$
$\boldsymbol{X}_{-i,\cdot}$	$\boldsymbol{X}$ with i-th row removed
$\boldsymbol{X}_{\cdot,-j}$	$\boldsymbol{X}$ with j-th column removed
$\boldsymbol{X}_{-i,j}$	$\boldsymbol{x}_{\cdot,j}$ with i-th row removed
$\boldsymbol{X}_{i,-j}$	$\boldsymbol{x}_{i,\cdot}$ with j-th column removed
$\boldsymbol{x}$	column vector of the appropriate size
x_i	the i-th element in the vector $\boldsymbol{x}$
$x_{(i)}$	the i-th smallest element in the vector $\boldsymbol{x}$
$\boldsymbol{x}_{-i}$	$\boldsymbol{x}$ with i-th element removed
$\boldsymbol{I}$	identity matrix of the appropriate size
$\mathbf{1}$	a vector with 1s of the appropriate size
$\operatorname{diag}(\boldsymbol{X})$	take the diagonal elements of $\boldsymbol{X}$ to a vector
$\operatorname{diag}(\boldsymbol{x})$	take the vector $\boldsymbol{x}$ to a diagonal matrix
$\operatorname{sgn}(\cdot)$	the sign function
$\mathbb{1}(x \geq \mu)$, $\mathbb{1}_\mu$	the step function with jump at μ
$[x]^+$	$x\mathbb{1}(x \geq 0)$
$\phi(\cdot;\cdots)$	component density
$\Phi(\cdot;\cdots)$	component CDF
$f(\cdot;\cdots)$	density of a location mixture
$F(\cdot;\cdots)$	CDF of a location mixture
$\phi_d(\cdot;\cdots)$	discrete component density
$\Phi_d(\cdot;\cdots)$	discrete component CDF
$f_d(\cdot;\cdots)$	density of a discrete location mixture
$F_d(\cdot;\cdots)$	CDF of a discrete location mixture

$o(1)$	$\limsup f(x)/g(x) = 0$
$O(1)$	$\limsup f(x)/g(x) < \infty$
$o_p(1)$	convergence to 0 in probability
$O_p(1)$	stochastically boundedness
$a_n \asymp b_n$	$a_n/b_n + b_n/a_n = O(1)$
$\|\| \cdot \|\|_\infty$	the appropriate sup-norm
$\|\| \cdot \|\|_2$	the appropriate L_2-norm
$\|\| \cdot \|\|_F$	the matrix Frobenius norm
$\lfloor \cdot \rfloor$	round down to the nearest integer
$\lceil \cdot \rceil$	round up to the nearest integer
$\tanh(\cdot)$	hyperbolic tangent function
$\text{arctanh}(\cdot)$	inverse hyperbolic tangent function

Chapter 1

Introduction

1.1 Overview

With the development of new technology, much more information than ever before can be explored from data. Modern devices produce thousands and sometimes millions of observations, each with its own estimation or prediction problem, and this set of estimation problems is known as the *large-scale inferential problem* (Efron, 2010b). The term "large-scale" mentioned in this thesis is in the statistical sense, which are different from the large-scale computation problems in which the large datasets are dealt with, although large-scale statistical problems can also deal with large datasets. When solving large-scale problems, it is certainly not wrong to apply classical methods on each observation since these classical methods are theoretically justified. However, if these observations have common features, estimation using all observations should be more accurate and efficient. In this thesis, we aim to develop new methods to solve large-scale problems by simultaneous estimation using non-parametric mixtures.

This thesis pays a special attention to large-scale inferential problems, in particular multiple hypothesis testing and covariance matrix estimation. The use of hypothesis testing has a long history, from examining whether male and female births are equally likely in Arbuthnot (1710) to investigating demographics and baseline characteristics of patients infected with 2019-nCoV in Huang et al. (2020). It has been the favoured statistical tool in some experimental social sciences. Over 90% of the articles published in the Journal of Applied Psychology during early 1990s included statistical significance testing (Hubbard et al., 1997). Multiple hypothesis testing enables estimation of more than one hypothesis at a time and shifts the focus to finding as many significant hypotheses as possible out of a large number of hypotheses being tested. Frahm et al. (2012) performed multiple hypothesis testing to study the performance of different

investment strategies using US stock returns, and Goeman and Solari (2014) presented an overview of multiple hypothesis testing in genomics data from a user's perspective.

Covariance matrix estimation focuses on exploring the relationships among random variables on the basis of a sample from a multivariate distribution. It is the simplest summary measure of inter-dependence among variables, and plays a prominent role in almost every aspect of multivariate data analysis (Pourahmadi, 2013). It has been widely used in many practical fields. Lindquist (2008) estimated the spatial and temporal covariance in the statistical analysis of fMRI data. Wang and Zou (2010) studied covariance matrix estimation especially when dealing with high-frequency financial data. A review by Fan and Liu (2013) discussed estimating covariance matrices for understanding correlation structure, inverse covariance matrix for network modelling, large-scale simultaneous tests for selecting significantly differently expressed genes and proteins and genetic markers for complex diseases, and high-dimensional variable selection for identifying important molecules for understanding molecule mechanisms in pharmacogenomics.

The motivation of this thesis is shown in Section 1.2 and a list of contributions made is in Section 1.3. The structure and outline of the remainder of the thesis are described in Section 1.4.

1.2 Motivation

We focus on the large-scale problems that the distributions from which the observations are drawn in general belong to the same family with maybe different incidental parameters, including but not limited to the cases where observations are realisations of repeated applications of the same classical method. Taking an empirical Bayes approach, information accrues across the observations in a way that combines Bayesian and frequentist ideas, and hence estimation and prediction may potentially have an increased power in this framework (Efron, 2010b). Since the structure of all incidental parameters is not generally available prior to the estimation process, estimation using non-parametric mixtures is a crucial step in making inferences using the empirical Bayes method. The non-parametric mixture model assumes that the distribution of all underlying incidental parameters is treated as completely unspecified as to whether it is discrete, continuous or in any particular family of distributions (Lindsay, 1995). Methods based on this new approach should be more accurate and have a meaningful interpretation.

Many methods in multiple hypothesis testing make inference using the p-values which are calculated using only the null distribution, and hence do not take into account the locations of non-null effects. If the location of a non-null effect is close to that of the null effect, an inference made could be too pessimistic, and could be too optimistic if the locations of non-null effects are far away from the null. As a result, we aim to propose a new method for estimation, testing and prediction that also considers non-null effects. When t-tests are performed on high-throughput microarrays, each test statistic comes from a t-distribution of the same degrees of freedom with its own non-centrality parameter, and hence the distribution of all test statistics is a t-mixture. The t-mixture automatically takes into account the locations of null and non-null effects. In practice, t-statistics are often transformed to z-statistics and the distribution of all z-statistics can still be thought as a mixture but using normal components. The formulation of the problem is precisely in the form of non-parametric mixtures, and hence non-parametric mixture estimation can be used for multiple hypothesis testing.

The usual estimator of a covariance matrix, the sample covariance matrix, is consistent and efficient by Ráo (1945) and Cramér (1946) in classical theory. However, the disadvantages of the sample covariance matrix are also evident. It is known to be ill-conditioned in high-dimensional settings, and the unbiased estimators typically have a higher variance which may be substantially inefficient from the viewpoint of decision theory. In modern statistics and machine learning, it becomes more popular to choose a model that both accurately captures the features in the observations but also generalises well to unseen observations, and hence we aim to propose a covariance matrix estimator that minimises the mean squared error risk with the help of empirical Bayes. The mean squared error is associated with the bias-variance tradeoff, which combines errors from erroneous assumptions in model specification and sensitivity to small fluctuations in observations. Furthermore, the result by Hotelling (1953) on the moments of a sample correlation coefficient suggests that the distribution of all sample correlation coefficients can be approximated as a one-parameter mixture, and the work by Fisher (1915) suggests that the distribution of all Fisher-transformed sample correlation coefficients can be approximated as a normal mixture. These suggest that non-parametric mixtures can also be used for covariance matrix estimation.

Detailed overviews on the existing methods for multiple hypothesis testing and covariance matrix estimation will be given in Chapters 3 and 8, respectively. To our best knowledge, there is no generic framework in the literature that uses non-parametric mixtures for solving large-scale problems. We aim to show in this thesis the advantages

of applying non-parametric mixtures to large-scale problems, in the two special cases of multiple hypothesis testing and covariance matrix estimation.

1.3 Contributions

For the objectives targeted in this thesis, new methods are proposed to solve large-scale problems and applied under various settings. The main contributions of this thesis are listed as follows.

- Provide a new perspective to solving large-scale problems using non-parametric mixtures.
- Propose a new likelihood-based method for estimation in multiple hypothesis testing.
- Propose a new likelihood-based method for covariance matrix estimation.
- Propose distance-based alternatives for estimating the mixing distribution and modify likelihood-based and distance-based methods to accommodate very large-scale computation.
- Derive some theoretical properties of the new methods proposed in multiple hypothesis testing and covariance matrix estimation.
- Propose a new FDR-controlling procedure based on the estimated null proportion and its estimated density.
- Implement the new methods in R and conduct numerical studies.

1.4 Outline of the Thesis

This thesis is divided into 10 chapters. The topics addressed in each chapter are as follows.

In Chapter 2, we provide an introduction and literature reviews on non-parametric mixture methods of estimating a mixing distribution. Construction of the estimators and theoretical properties are introduced under various distance measures. The non-parametric maximum likelihood estimator (NPMLE) is mentioned on its own in

Section 2.4 as a special type of the distance-based methods due to great statistical properties and wide applications.

In Chapter 3, we first point out a problem that the usual NPMLE can not produce a positive probability at the location representing the null effect in multiple hypothesis testing. The algorithm mentioned in Wang (2007) is then modified such that a consistent NPMLE can be found when the locations and weights of some support points are fixed, under mild conditions. With the modification to the NPMLE, the weight of the location representing the null effects can then be estimated and some theoretical derivations are also shown. Computing the modified NPMLE and the null proportion is shown in Section 3.5.

Chapter 4 contains the numerical studies of the method proposed in Chapter 3. Several threshold functions are used to estimate the proportion of null effects for the proposed method. Synthetic datasets are generated from the estimates of two real world datasets in order to investigate the performance of the proposed method against existing methods. A more systematic simulation study on a two-component model is also shown to see the behaviours of the proposed method when the true proportion of the null effects and the location of the non-null effect vary.

In Chapter 5, we turn our attention to distance-based methods as alternatives to the maximum likelihood. We look into the non-parametric minimum distance estimation using the Cramér-von Mises and the Anderson-Darling distance. We modify the algorithm in Wang (2007) so that the non-parametric minimum distance estimator (NPMDE) under these two distances can be found, and then a further modification is made to find the NPMDE when the locations and weights of some support points are fixed. Since the asymptotic distributions of the Cramér-von Mises and the Anderson-Darling statistic under the true mixing distribution are known, the null proportions can also be consistently estimated under these two distances. Numerical studies are also performed in order to see the computation time and accuracy of the NPMDEs under the Cramér-von Mises and the Anderson-Darling distance against the NPMLE.

In Chapter 6, binning on the observations is used to reduce the computational burden by reducing the effective sample size when the number of observations gets large. The unrestricted and the restricted estimators mentioned in Chapters 3 and 5 are modified for the large-scale computation under the assumption of normal components. Numerical studies are also performed to see how binning affects the estimations of the mixing distribution and the proportion of null effects. The numerical study in Section 4.2 is extended for a even larger number of observations with the help of this

large-scale modification.

In Chapter 7, a new FDR-controlling procedure is proposed based on the restricted NPMLE. The proposed procedure classifies the hypotheses by constructing a critical region, which can be computed by solving a non-linear optimisation problem. Numerical studies are conducted on in order to see the performance of the new FDR-controlling procedure under correct specification of the model and how the estimated null proportion affects the proposed procedure.

Chapter 8 concerns covariance matrix estimation. We propose new ways of estimating the covariance matrix using empirical Bayes on the sample correlation coefficients and the transformed sample correlation coefficients. The new estimators require the estimated density of the (transformed) sample correlation coefficients, which can be estimated using the method mentioned in Chapters 3 and 5. Estimating a covariance matrix under the sparsity constraint is also discussed as well as the theoretical properties of the covariance estimator.

Chapter 9 contains the numerical studies of the covariance matrix estimator proposed in Chapter 8. Two different covariance matrices in the time series are used to conduct the simulation studies. We also consider estimating the covariance matrix of the log-returns of stocks listed in the National Association of Securities Dealers Automated Quotations (NASDAQ). The estimated covariance matrix is also used to generate the synthetic datasets in order to investigate the performance of the proposed estimator under a general covariance structure. In the last part of this chapter, we estimate the covariance matrix of the data used in Chapter 4 in order to assess the feasibility of the independence assumption as well as the ability for the large-scale computation together with the leukemia dataset.

Chapter 10 provides concluding remarks as well as possibilities for future work.

This thesis also has three appendices. Appendix A describes an alternative way to compute the density of t-distributions numerically when degrees of freedom are low. Appendix B contains the numerical algorithms used in finding new support points at each iteration as well as the algorithm for estimating the null proportion. The full list of the 100 publicly listed companies used in Section 9.4 is in Appendix C.

1.5 Notations

For coherency, the notations used in this thesis are described in this section. A quick reference to the notations can be found in the List of Symbols before Chapter 1, and

a quick reference to the abbreviations can be found in the List of Abbreviations.

Let a bold lower-case letter (e.g. $\boldsymbol{v}$) be a column vector, v_i be the i-th element in the column vector, $v_{(i)}$ be the i-th smallest element in the column vector and $\boldsymbol{v}_{-i}$ be the vector with the i-th element removed. Let a bold upper-case letter (e.g. $\boldsymbol{M}$ or $\boldsymbol{\Sigma}$) be a matrix, then $\boldsymbol{m}_{i,\cdot}$ is the i-th row of the matrix $\boldsymbol{M}$, $\boldsymbol{m}_j = \boldsymbol{m}_{\cdot,j}$ are the j-th column of the matrix $\boldsymbol{M}$ and its non-bold lower-case letter with a subscript (e.g $m_{i,j}$ or σ_{ij}) is the (i,j)-th element where i is the row index and j is the column index. We also denote $\boldsymbol{M}_{-i,\cdot}$ to represent the matrix $\boldsymbol{M}$ with the i-th row removed, $\boldsymbol{M}_{\cdot,-j}$ to represent the matrix of $\boldsymbol{M}$ with the j-th column removed, $\boldsymbol{m}_{-i,j}$ be the vector of the j-th column of $\boldsymbol{M}$ with the i-th term removed and $\boldsymbol{m}_{i,-j}$ be the vector of the i-th row of $\boldsymbol{M}$ with the j-th term removed. The symbol $\boldsymbol{I}$ stands for the identity matrix with the appropriate dimension and bold numbers (e.g. $\mathbf{0}$, $\mathbf{1}$) are vectors of the given constants of the appropriate size. Let $\mathcal{N}_p(\boldsymbol{\mu}, \boldsymbol{\Sigma})$ be a p-variate normal distribution with mean $\boldsymbol{\mu}$ and covariance matrix $\boldsymbol{\Sigma}$. For the simplicity, we drop the subscript p and write $\mathcal{N}(\mu, \sigma)$ for the univariate normal distribution ($p = 1$). To avoid inconsistency of notation, we do not distinguish random matrices and the realised matrices by the appearance, and hence attentions should be paid to the definition of the symbol when dealing with the matrices. If no ambiguity arises, bold upper-case letters can sometimes be used to represent random vectors, which typically happens in sections with no reference to matrices. The function $\text{diag}(\boldsymbol{M})$ maps the matrix $\boldsymbol{M}$ to the vector of its diagonal elements while $\text{diag}(\boldsymbol{m})$ maps the vector $\boldsymbol{m}$ to the diagonal matrix with the diagonal entries $\boldsymbol{m}$. The function $\text{sgn}(\cdot)$ produces the element-wise sign of the input. The function $[\cdot]^+$ also operates on each element of the input and makes no change to the element if positive, and change the value to 0 otherwise. The operator $||\cdot||_p$ is the appropriate L_p-norm, with special case $||\cdot||_F$ is the Frobenius norm. The functions $\lceil\cdot\rceil$ and $\lfloor\cdot\rfloor$ round the input up and down to the nearest integer respectively.

Let $\phi(x; \mu, \sigma)$ be the density function where x is the quantile, μ is the location parameter and σ is the scale parameter, e.g., μ and σ are mean and standard deviation respectively when ϕ is a normal density while μ and σ are the non-centrality parameter and the degrees of freedom respectively when ϕ is a t-density. This notation is naturally extended to multivariate distributions by writing $\phi(\boldsymbol{x}; \boldsymbol{\mu}, \boldsymbol{\Sigma})$. The function

$$f(x; G, \sigma) = \int \phi(x; \mu, \sigma) dG(\mu)$$

is denoted as the density of a location mixture, where G is a mixing distribution on

the location parameter. Let $\Phi(x;\mu,\sigma)$ and $F(x;G,\sigma)$ be the corresponding cumulative distribution function (CDF) of $\phi(x;\mu,\sigma)$ and $f(x;G,\sigma)$ respectively. If weights and locations are fixed for some support points in a mixing distribution, we typically append the information to the scale parameter, i.e., in the case of the mixture density,

$$f(x;G,\sigma,\boldsymbol{\pi},\boldsymbol{\beta}) = \sum_{i=1}^{k} \pi_i \phi(x;\beta_i,\sigma) + \left(1 - \sum_{i=1}^{k} \pi_i\right) \int \phi(x;\mu,\sigma) dG(\mu),$$

where $\boldsymbol{\pi}$ and $\boldsymbol{\beta}$ are vectors of fixed weights and locations respectively. The functions with the subscript "d" (e.g. ϕ_d, Φ_d, f_d, F_d) are the discrete versions of the densities or the cumulative distribution functions introduced in Chapter 6. The empirical cumulative distribution function (ECDF) computed using observations of size n is typically denoted as $\hat{F}_n$. The step function with the step up at μ is denoted as $\mathbb{1}(x \geq \mu)$ or $\mathbb{1}_\mu$. We have the non-probabilistic big O notation, if there exists positive numbers ϵ and M such that for all x with $0 < |x - a| < \epsilon$,

$$|f(x)| \leq M g(x),$$

then we can write $f(x) = O(g(x))$. If for all positive constant ϵ, there exists a constant N such that

$$|f(x)| \leq \epsilon g(x) \qquad \forall x \geq N,$$

then we can write $f(x) = o(g(x))$, which is known as the little-o notation. We also have the stochastic little-o notation, i.e.,

$$X_n = o_p(a_n) \qquad \text{or} \qquad X_n/a_n = o_p(1)$$

to represent the convergence in probability, where $X_n = o_p(1)$ is defined as

$$\lim_{n\to\infty} \mathbb{P}(|X_n| \geq \epsilon) = 0,$$

for every positive ϵ and

$$X_n = O_p(a_n) \qquad \text{or} \qquad X_n/a_n = O_p(1)$$

is the stochastic boundedness, where $X_n = O_p(1)$ is defined as

$$\exists M \in \mathbb{R}, N \in \mathbb{N}, \forall n > N, \qquad \mathbb{P}(|X_n/a_n| > M) < \epsilon,$$

for all $\epsilon > 0$. In addition, we write $a_n \asymp b_n$ if a_n and b_n are stochastically of the same order.

Throughout this thesis, we typically use π_0, G, H, etc. as variables to represent their corresponding parameters. We use the superscript "*" for their true values (e.g. π_0^*, G^* and H^*) where the ambiguity may arise.

Chapter 2

Literature Review

2.1 Introduction

The overall goal of this thesis is to solve large-scale inferential problems using non-parametric mixtures. Large-scale inferences encountered in multiple hypothesis testing were discussed in Efron (2010b) while large-scale inferential problems in covariance matrix estimation can be found in Pourahmadi (2013). In this chapter, we introduce non-parametric mixtures first in Section 2.2 and the remainder of the chapter focuses on providing literature reviews on non-parametric mixture methods. The formulation and theoretical aspects of non-parametric distance-based estimation are given in Section 2.3, while the computation and theory of non-parametric maximum likelihood estimation, a special case of non-parametric distance-based estimation, are given in Section 2.4. Rates of convergence for non-parametric estimators of a mixing distribution have also drawn significant attention in recent years, and we will discuss them in Section 2.5.

The literature review on the specific topic in which the proposed methods are applied is in the corresponding chapter. That is, the literature review on multiple hypothesis testing will be given in Section 3.2, and Sections 7.2 and 8.2 give literature reviews on controlling procedures and covariance matrix estimation, respectively.

2.2 Non-parametric Mixtures

The non-parametric mixtures assume that the distribution of all underlying parameters is treated as completely unspecified. It can be discrete, continuous or in any particular family of distributions, i.e.,

$$f(x; G, \sigma) = \int \phi(x; \mu, \sigma) dG(\mu),$$

where G is a mixing distribution and σ is a structural parameter. In particular, if G is discrete, we can write the mixture as

$$f(x; G, \sigma) = \sum_{i=1}^{k} \pi_i \phi(x; \mu_i, \sigma),$$

where μ_i and π_i for $i = 1, \ldots, k$ are the component means and probabilities respectively. In this thesis, we focus on G being the mixing distribution of the location parameter and the structural parameter σ is a known constant or a known mapping of the location parameter.

2.3 Non-parametric Distance-based Estimation

2.3.1 Formulation

In general, the non-parametric minimum distance estimator (NPMDE) of a mixing distribution is the solution to the optimisation problem:

$$\hat{G} = \arg\min_{G \in \Gamma} \ell(G),$$

where Γ is a given space of cumulative distribution functions (CDFs) and ℓ is a distance metric typically depending on the empirical cumulative distribution functions (ECDF) $\hat{F}_n$. Some common distance functions in the non-parametric minimum distance estimation are given below.

Example 2.1 (Kolmogorov-Smirnov)**.** *The Kolmogorov-Smirnov distance (Kolmogorov, 1933; Smirnov, 1948) is*

$$\ell(G) = \sup |F(\cdot; G, \sigma) - \hat{F}_n|.$$

Then the NPMDE is the solution to the problem

$$\hat{G} = \arg\min_{G \in \Gamma} ||F(\boldsymbol{x}; G, \sigma) - \hat{F}_n(\boldsymbol{x})||_\infty.$$

Example 2.2 (Cramér-von Mises)**.** *The Cramér-von Mises distance (Cramér, 1928; von Mises, 1931) is*

$$\ell(G) = \int [F(\cdot; G, \sigma) - \hat{F}_n]^2 dF(\cdot; G, \sigma).$$

Using the transformation in Anderson and Darling (1954), the NPMDE is the solution to the problem

$$\hat{G} = \arg\min_{G \in \Gamma} \frac{1}{12n^2} + \frac{1}{n} \sum_{i=1}^{n} \left[F(x_{(i)}; G, \sigma) - \frac{2i-1}{2n} \right]^2.$$

Example 2.3 (Anderson-Darling)**.** *The Anderson-Darling distance (Anderson and Darling, 1952) is*

$$\ell(G) = \int \frac{[F(\cdot; G, \sigma) - \hat{F}_n]^2}{F(\cdot; G, \sigma)[1 - F(\cdot; G, \sigma)]} dF(\cdot; G, \sigma).$$

Using the transformation in Anderson and Darling (1954), the NPMDE is the solution to the problem

$$\hat{G} = \arg\min_{G \in \Gamma} -n - \frac{1}{n} \sum_{i=1}^{n} (2i-1) \{ \log[F(x_{(i)}; G, \sigma)] + (2n - 2i + 1) \log[1 - F(x_{(i)}; G, \sigma)] \}.$$

Example 2.4 (Squared Error)**.** *The squared error loss (Hall, 1981) is*

$$\ell(G) = \int [F(\cdot; G, \sigma) - \hat{F}_n]^2 d\hat{F}_n.$$

Then the NPMDE is the solution to the problem

$$\hat{G} = \arg\min_{G \in \Gamma} \frac{1}{n} \sum_{i=1}^{n} \left[F(x_{(i)}; G, \sigma) - \frac{i}{n} \right]^2.$$

Example 2.5 (Kullback–Leibler)**.** *The Kullback–Leibler (KL) divergence (Kullback and Leibler, 1951) is*

$$\ell(G) = \int \log \left(\frac{f(\cdot; G^*, \sigma)}{f(\cdot; G, \sigma)} \right) dF(\cdot; G^*, \sigma).$$

Then the NPMDE is the solution to the problem

$$\hat{G} = \underset{G \in \Gamma}{\arg\min} \frac{1}{n} \sum_{i=1}^{n} \log \left(\frac{f(x_i; G^*, \sigma)}{f(x_i; G, \sigma)} \right).$$

It is noted that the KL divergence is not a proper metric, since it does not satisfy the triangle inequality; see Section 5.5 of van der Vaart (1998). However, the KL divergence still plays an important role in information theory and estimation because of its close relationship to the maximum likelihood method. Murphy (2012) mentioned that the maximum likelihood estimate minimises the KL divergence to the empirical distribution. The KL divergence (maximum likelihood) has its own right to be mentioned due to its special properties and wide applications, which will be described in detail in Section 2.4.

2.3.2 Theory and computation

As shown in the previous section, various loss functions depending on the context of the problem can be used to compute the NPMDE. This section gives more details on consistency and computational methods for these loss functions. For the squared error distance, Choi and Bulgren (1968) showed the strong convergence of $\hat{G}$. The consistency of the NPMDE under this distance is simple and easy to verify. It only requires that the underlying distribution function is a continuous mixture, that $\Phi(x; \mu, \sigma)$ is continuous in x and μ and that the limits

$$\Phi(x; \infty, \sigma) = \lim_{\mu \to \infty} \Phi(x; \mu, \sigma), \qquad \Phi(x; -\infty, \sigma) = \lim_{\mu \to -\infty} \Phi(x; \mu, \sigma)$$

exist for every x with neither $\Phi(\cdot; \infty, \sigma)$ nor $\Phi(\cdot; -\infty, \sigma)$ being a distribution function. The consistency of the NPMDE merely has a requirement on Γ. Quadratic programming can be used to find the weights given that the support points are known. In the case of estimating a continuous mixture, the estimator of the mixing distribution is has at most n support points, where n is the number of observations. Bartlett and Macdonald (1968) suggested that using a weighted squared distance would perform better. The weighted least squares are the most sufficient when the weights are close to the true ones. Other choices of the weights can be the geometric mean, since the geometric mean has some convenience over the arithmetic mean in theoretical investigation of the efficiency. A distance-based estimator under the squared error was proposed in Hall

(1981). The weights of the mixing distribution can be obtained by solving a system of linear equations and the support points are the weighted mean in each of the partition of the samples. It was shown that the efficiency approaches 1 as the distance between component distributions increases. Titterington (1983) adapted a similar approach to Hall (1981) and proposed an estimator based on density estimators. This method is applicable to general discrete, multivariate, univariate continuous and ordered categorical data. The asymptotic theory was also developed under squared error loss. Henna (1983) mentioned that a lemma crucial to the consistency proof in Choi and Bulgren (1968) caused much confusion. A new proof was given in the paper and it was shown that the consistency of the NPMDE still holds under the original setting and can be extended under a weaker assumption.

Another family of distances used in minimum distance estimation are the Cramér-von Mises and Anderson-Darling distances. Parr and Schucany (1980) gave the definition of the sequence of asymptotic minimum distance estimators based on the empirical distribution functions with respect to a distance measure and stated the conditions under which the strong consistency holds. The behaviour of some selected distance measures was also mentioned. In cases such as the Anderson-Darling distance, where the weights given to deviations in probability units from the model are high at the tails of the distribution, drastic sensitivity to incorrect tail-width specification can be expected. The Cramér-von Mises distance drastically down-weights discrepancies in the tails. The "trimmed" versions of the weighted Cramér-von Mises distance were designed to further minimise the effect of extreme observations. Parr and DeWet (1981) considered the minimum distance estimation using the weighted Cramér-von Mises distance. Under rather general conditions, the derived estimators are asymptotically normally distributed. Considerations were also given to appropriate weights to produce Fisher-efficient estimators. Boos (1981, 1982) paid attention to the minimum distance estimation under the Anderson-Darling distance in particular, after Parr and Schucany (1980) and Parr and DeWet (1981) mentioned that the weighted Cramér-von Mises distance produces estimators with good robustness and efficiency properties. The existence and consistency of the NPMDE under the Anderson-Darling distance were shown. It was also mentioned that the residual of the Anderson-Darling distance can be used to assess model validity. Woodward et al. (1984) considered the minimum distance estimator based on the Cramér-von Mises distance because Parr and Schucany (1980) found that this estimator performed well in various settings. However, the algorithms for computing, the choice of starting values and the simulation studies

were only done in the case of a two-component normal mixture model with unknown locations, scales and the proportion.

The Hellinger distance has also drawn a significant attention recently. Contrary to squared error and (weighted) Cramér-von Mises distances which compute the distance between two cumulative distribution functions, the Hellinger distance is a measure between two density functions. Karlis and Xekalaki (1998) discussed the advantages of the Hellinger distance in relation to the trade-off between robustness and efficiency according to Lindsay (1994) and proposed a minimum distance estimator for a Poisson mixture under the Hellinger distance. The identifiability, existence and asymptotic normality of the estimator were shown. The algorithm introduced is very similar to the one in Behboodian (1970) when deriving the maximum likelihood estimator (MLE) in the case of finite normal mixtures. This algorithm shares some weaknesses with the EM algorithm. The rate of convergence is slow and the estimate of the mixing distribution depends on the initial values. Karunamuni et al. (2009) and Wu et al. (2009) investigated a two-component mixture model using the Hellinger distance. Under the identifiability condition, it was shown that the estimator of the proportions in this two-component model is asymptotically consistent and asymptotically normally distributed with high efficiency properties. The estimation of each component was done by kernel density estimation with the appropriate choice of the bandwidth.

2.4 Non-parametric Maximum Likelihood Estimation

2.4.1 Formulation

Given a set of observations $x_1, \cdots, x_n$, the log-likelihood function as a function of G can be written as

$$\ell_n(G) = \sum_{i=1}^{n} \log f(x_i; G, \sigma),$$

and the NPMLE is the solution to the following optimisation problem,

$$\hat{G} = \arg\max_{G \in \Gamma} \ell_n(G). \tag{2.1}$$

As mentioned in the previous section, the likelihood-based method can be regarded as a very special case of the distance-based methods due to the close relationship between

the maximum likelihood and the KL divergence. The negative of (2.1) is equivalent to the objective function in Example 2.5. If the number of support points is known and fixed, then the estimation is parametric and the asymptotic theory is complete, thanks to the law of large numbers and the central limit theorem. We are not going to discuss the parametric MLE in detail; our main focus in this thesis is on the non-parametric maximum likelihood estimation, which does not make any assumption on the underlying mixing distribution.

2.4.2 Theoretical Results

The difficulty in a non-parametric mixture estimation is that the parameter G is completely unspecified. Researches have been done by many authors on the theoretical aspect of the non-parametric maximum likelihood methods. The pioneering work in establishing the strong consistency of the non-parametric maximum likelihood estimator (NPMLE) of a mixing distribution is Kiefer and Wolfowitz (1956) based on Wald (1949). The metric chosen by them is

$$d(G', G'') = \int |G'(\mu) - G''(\mu)| e^{-|\mu|} d\mu,$$

for $G', G'' \in \Gamma$, and Γ is assumed to be the set containing all G such that

$$\int_{-\infty}^{\infty} [\log |\mu|]^{+} dG(\mu) < \infty. \tag{2.2}$$

Several assumptions were also made corresponding to assumptions made in Wald (1949) and are listed here since they will be constantly referred to in the remainder of the thesis.

Assumption 1: $\phi(x; \mu, \sigma)$ is a density with respect to a σ-finite measure μ on a Euclidean space which x is the generic point.

Assumption 2: (Continuity) For any sequence of distributions $\{G_n\}$ and G in the completed space of Γ, the weak convergence of $\{G_n\}$ to G implies

$$\lim_{n \to \infty} \int \phi(x; \mu, \sigma) dG_n(\mu) = \int \phi(x; \mu, \sigma) dG(\mu).$$

Assumption 3: (Measurability) For any G' in the completed space of Γ and $\rho > 0$, $w(x; G', \sigma, \rho)$ is a measurable function of x, where

$$w(x; G', \sigma, \rho) = \sup \int \phi(x; \mu, \sigma) dG''(\mu), \tag{2.3}$$

and the supremum is taken over all G'' in the completed space of Γ for which $d(G', G'') < \rho$.

Assumption 4: (Identifiability) If G', G'' are two different points in the complete space of Γ, then for at least one y,

$$\int_{-\infty}^{y} \phi(x; \mu, \sigma) dG'(\mu) \neq \int_{-\infty}^{y} \phi(x; \mu, \sigma) dG''(\mu).$$

Assumption 5: (Integrability) For any G' in the completed space of Γ, as $\rho \to 0$,

$$\lim \mathbb{E}\left[\log \frac{w(X; G', \sigma, \rho)}{f(X; G, \sigma)}\right] < \infty,$$

where $w(\cdot; G', \sigma, \rho)$ is the function defined in (2.3).

The component density ϕ under the assumption (2.2) can, e.g. be normal, Cauchy, uniform and exponential distributions. If σ is fixed, the consistency of $\hat{G}$ can be established without further restrictions. If σ has to be estimated simultaneously, a bound of σ is needed in order for the measurability to hold.

The most demanding condition in showing the consistency of the NPMLE in Kiefer and Wolfowitz (1956) is the integrability assumption which relies on the use of the generalised Jensen's inequality to show

$$\mathbb{E}\left[\frac{f(X; G', \sigma)}{f(X; G, \sigma)}\right] < 0, \tag{2.4}$$

as $\rho \to 0$. Pfanzagl (1988) gave another consistency proof, with emphasis on mixtures over exponential families. Another inequality was proposed to take over the role of the Jensen's inequality to verify the integrability assumption. That is, for any $\epsilon \in (0, 1)$, we have

$$\mathbb{E}\left[1 + \epsilon\left(\frac{f(X; G, \sigma)}{f(X; G', \sigma)} - 1\right)\right] \geq 0, \tag{2.5}$$

and equality holds if and only if $f(X; G') = f(X; G)$ almost surely with respect to

$f(X;G)$. Wilcox (2012) formalised the argument in Leroux (1992) and showed that the NPMLE inconsistently estimates the number of support points, almost surely. Chen (2017) used a Poisson mixture example to show that inequality (2.5) is much easier to verify than (2.4) in practice. A more detailed overview of the consistency of the MLE for mixture models is given in Chen (2017).

For the existence and discreteness of the NPMLE, Laird (1978) showed that the NPMLE exists and that the probabilities are positive only at discrete points. The NPMLE is a discrete mixing distribution if the component density function ϕ is analytic, and either that the component densities for distinct observations are linearly dependent or that the k-th derivatives with respect to μ in $\phi(x_i;\mu,\sigma)$ are of the same sign and non-zero at some point. Since then, various authors obtained the uniqueness and the support size of the NPMLE for specified component densities: Simar (1976) for Poisson, Jewell (1982) for exponential distributions and Geman and Hwang (1982) for the empirical evidence using normal components. Lindsay (1983) showed the existence and the discreteness of the NPMLE in general from a geometric point of view under mild conditions. In addition, it was also established that the number of support points in the NPMLE has no more than the number of observations in the sample and that the NPMLE is unique under certain conditions. Leroux (1992) pointed out that the estimated number of components is shown to be more than the true number of components for large samples. Lindsay and Roeder (1993) provided another proof of the uniqueness and the support size of the NPMLE which is simpler and more widely applicable. It was shown in the paper that either the NPMLE is unique or the gradient functions is identically 0 at the maximum likelihood solution, and that the NPMLE is unique for the strictly totally positive models, which include the one-parameter exponential family, non-central t-distributions and non-central χ^2-distributions.

2.4.3 Computation

It was mentioned that the NPMLE is a discrete distribution with the support points no more than the number of observations, and the exact number of support points in the NPMLE is typically random. Such randomness in the support points creates great difficulties in the estimation process.

Laird (1978) proposed a method using the fact the NPMLE is a step function with finite steps. The approach to computing the NPMLE is to start with a large value close to the number of observations and maximise the (log-)likelihood function. It

turned out that when the value chosen is sufficiently large, then the (log-)likelihood function will be maximised on a ridge and the EM algorithm should be used. The implementation of the EM algorithm for finite mixtures and proof of convergence with the special reference to convergence to a point on a ridge were shown in Dempster et al. (1977). Lindsay (1983) defined the directional derivative from a mixing distribution G' to another mixing distribution G'' as,

$$d(G''; G') \equiv \left. \frac{\partial \ell\{(1-\epsilon)G' + \epsilon G''\}}{\partial \epsilon} \right|_{\epsilon=0} = \sum_{i=1}^{n} \frac{f(x_i; G'', \sigma)}{f(x_i; G', \sigma)} - n.$$

Here, we write $d(G''; G')$ instead of $d(G''; G', \sigma)$ simply due to the reason that we treat σ fixed in the estimation process. If G'' is a step function with the jump at μ, then we write $d(G''; G')$ as $d(\mu; G')$, which is known as the gradient function. This function can characterise the NPMLE $\hat{G}$ thanks to the general equivalence theorem, i.e.,

$$\hat{G} \text{ maximises } \ell(G) \Leftrightarrow \hat{G} \text{ minimises } \sup_{\mu}\{d(\mu; G)\} \Leftrightarrow \sup_{\mu}\{d(\mu; \hat{G})\} = 0.$$

An algorithm based on the vertex direction method was proposed in Fedorov (1972) along with the theory. Set $s = 0$ and let G_s be the mixing distribution at iteration s. For each iteration, the value θ_s at which $d(\theta_s, G_s)$ is maximised can be found and G_{s+1} is set to the mixing distribution such that the log-likelihood is maximised in the path between $\mathbb{1}_\mu$ and G_s. Unfortunately this method has a sub-linear convergence rate and it becomes slower as approaching the convergence. Hall and Titterington (1984) proposed an efficient estimator to find the mixing proportions when the support points are fixed. By constructing a sequence of multinomial approximations and related maximum likelihood estimators, a Cramér-Rao lower bound for the non-parametric estimators of mixture proportions was derived and it was shown that this efficient estimator is asymptotically optimal. Böhning (1985) suggested a simple alternative method called the vertex exchange method, which is considerably faster in terms of the number of iterations. The algorithm can move the mass from one point to another in one step while other masses are kept constant, which retains a low numerical complexity but still shows a good convergence rate. Leroux (1992) proposed an estimation procedure which starts with finding the MLE given a fixed number of components, and uses a penalty function to estimate the number of support points. The MLE with the estimated number of support points is considered as the estimate of the underlying mixing

distribution. Lesperance and Kalbfleisch (1992) suggested two algorithms for computing the NPMLE. The first one is the intra-simplex direction method which focuses on using convexity to compute the optimal weights, subject to non-negativity constraints. This method does not store the current mixing distribution, and hence the number of components might grow explosively. The second method is the semi-infinite programming. It was found that the dual problem to the maximisation of the (log-)likelihood function found in Lindsay (1983) is precisely in the form that can be solved by semi-infinite programming. Newton (2002) proposed a Bayesian-style recursive estimator called the predictive recursion. The algorithm provides an estimator having a smooth density with respect to any specified dominating measure. However, it is not able to be characterised as an optimiser of any objective function, making the predictive recursion estimator difficult to interpret. Wang (2007) developed a method for computing the NPMLE using the constrained Newton method. Groeneboom et al. (2008) used a support reduction algorithm to compute the NPMLE. This algorithm is considered as a vertex direction method and is designed to compute the M-estimators in mixture models, using unconstrained optimisations iteratively. At each iteration, one point is added to the existing iterate and as many support points of the measure as possible are deleted after that. This algorithm is expected to perform well in problems where the solution is a sparse mixture, due to the speed at which low-dimensional optimisations can be performed. Chae et al. (2018) developed an iterative algorithm that shares certain features with the MLE, a Bayesian approach and the predictive recursion. The algorithm updates by taking an average of the current guess and the single-observation update based on the current guess as the prior. The algorithm produces a smooth mixing density as the near-MLE. The convergence was shown by proving the monotone increase in the likelihood as well as the characterisation of the iterative steps.

The constrained Newton method

The computational methods for computing the NPMLEs or NPMDEs in this thesis are based on or modified from the constrained Newton method in Wang (2007). This section gives more details about this method.

Brief steps of the algorithm are described below:

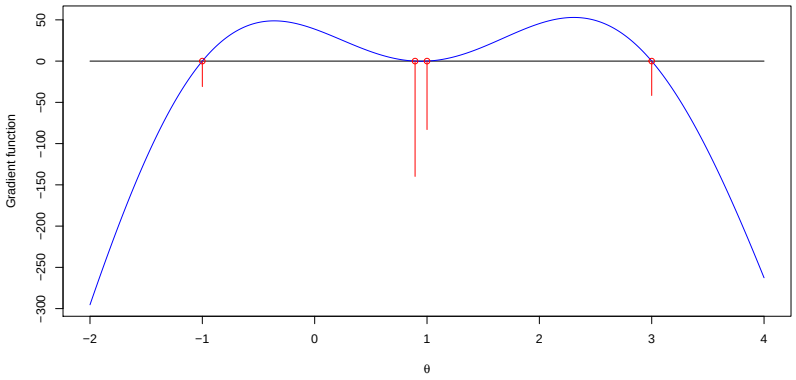

FIGURE 2.1: Gradient function after the first iteration.

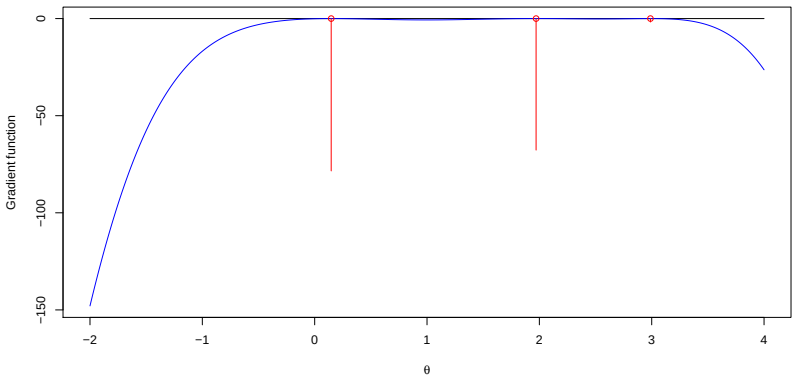

FIGURE 2.2: Gradient function at the NPMLE.

After Iteration 1: log-likelihood is −1742.966						
Support points	−1.000	−0.361	0.894	1.000	2.307	3.000
Probability	0.105	0.000	0.473	0.281	0.000	0.141
After Iteration 2: log-likelihood is −1725.286						
Support points	−1.000	−0.361	0.894	1.000	2.307	3.000
Probability	0.000	0.228	0.485	0.000	0.287	0.000
⋮						
After Iteration 14: log-likelihood is −1722.520						
Support points	0.147	1.971	2.989			
Probability	0.530	0.458	0.012			

TABLE 2.1: The log-likelihood and mixing distribution after Iterations 1, 2 and convergence.

Step 1: Compute all the local maxima $\theta^*_{s1}, \cdots, \theta^*_{sp_s}$ of $d(\theta; G_s)$. If $\max_j\{d(\theta^*_{sj}; G_s)\} = 0$, stop. This stopping criteria is due to the general equivalence theorem. In Figure 2.1, the gradients of two local maxima are greater than 0 and the program heads into Step 2, while in Figure 2.2, two local maxima are 0 suggesting convergence and the program is terminated.

Step 2: Add all the local maxima into the support set and compute the weights of all the points in the support set using the non-negative least squares (NNLS) algorithm in Lawson and Hanson (1974) after the transformation by Dax (1990) under the constraint that the sum of all weights equals 1. As shown in Table 2.1, the points at -0.361 and 0.237 are added into the support set with zero weights, and the NNLS algorithm relocates the weights of the support points.

Step 3: Discard all the points in the support sets with weight 0. The points with non-zero weight form a new mixing distribution G_{s+1}. Set $s = s + 1$ and go to Step 1. As shown in Table 2.1, the points -1.000, 1 and 3.000 are dropped from the support set.

This method uses a quadratic function to approximate the change of the log-likelihood, It will converge and the convergence rate is typically much faster speed than that of other methods (Wang, 2007). This method is implemented in the R package `nspmix` (Wang, 2017).

The constrained Newton method has been widely applied to many areas in the field of statistics. Wang and Stephen (2013) proposed a hierarchical constrained Newton method based on Wang (2007). The points in the current support set can be broken into blocks of a smaller size and the solutions to the smaller blocks can then be back-transformed to the solution to the overall problem, and hence it is suitable for large-sized support sets. Numerical studies were conducted and the method was applied to estimate survival function for interval-censored data. Chee and Wang (2013) investigated the similarity between the maximum likelihood and the minimum distance estimation and proposed a non-parametric density estimator that was modified from the constrained Newton method in Wang (2007). The no-smoothing, the singly-smoothing and the doubly-smoothing strategy were used for defining the quadratic distance in order to measure the discrepancy between the estimator and the data. It was mentioned that the doubly-smoothing strategy helps to reduce the bias in the mixture-based density estimator. The equivalence of the general equivalence theorem and the consistency of the NPMDE were described. Chee and Wang (2014) used the distance function in Balabdaoui and Wellner (2004) and Balabdaoui and Wellner (2010)

to estimate the k-monotone density using least squares. The estimation process is very similar to the non-parametric density estimator proposed in Chee and Wang (2013) with slight modifications. Lesperance et al. (2014) extended the constrained Newton method in Wang (2007) to the multivariate mixing distribution. The local maxima of the gradient function in the multi-dimensional case was done by the direct search method. Two methods for jointly maximising the log-likelihood in semi-parametric models were given and applied to the National Basketball Association data. Wang and Wang (2015) studied the NPMLE when the component density is a multivariate normal distribution. The covariance matrix, which is assumed to be the same across all components, are factored into a positive-definite matrix with unit determinant and a volume parameter. The mixing distribution is estimated under a semi-parametric model with the positive-definite matrix unknown, and the volume parameter is estimated using an information criteria. Titman (2017) considered the problem of estimating the distribution of the failure time for interval-censored data using the NPMLE, subject to misclassification. The expression for interval-censored survival data under misclassification, the support of the NPMLE and the optimality conditions were given. The iterative convex minorant and the constrained Newton algorithm for estimating the NPMLEs were compared. Wang and Fani (2018) discussed the difficulties when computing the U-shaped hazard function. The hazard function is only properly defined when its anti-mode is known. The NPMLE of this problem is proposed and it was shown that the proposed method performs much faster than the method iterating between the support reduction method and the bisection method. Liu and Wang (2018) and Wang (2019) adapted the NPMLE in the estimation of a univariate log-concave density. The logarithm of the estimated density is made up with a finite number of affine pieces. The locations of the mixing distribution are the break points of the piecewise function while the non-negative masses are the changes of the gradient from the previous piece. The method was then applied to the reliability data from Dümbgen and Rufibach (2011).

2.5 Rates of Convergence

The rate of convergence is also an important property of an estimator. It can be used to assess the quality and efficiency of the estimator. The rate of convergence of a parametric estimator can be typically obtained by the central limit theorem. However, obtaining the rate of convergence of non-parametric estimators is often a challenging

task. In this section, we will give an overview on the rate of convergence of non-parametric estimators of a mixing distribution and non-parametric estimators of a mixture density.

Carroll and Hall (1988) obtained the rate of convergence for the class of k-times differentiable densities with bounded supremum when the errors are normal and when deconvolving for general errors. It was also found that the optimal achievable rate of convergence is slower as the distribution of the error becomes smoother. Chen (1995) investigated the rate of convergence of distribution functions in finite mixture models. The metric defined by him is similar to the one defined in Kiefer and Wolfowitz (1956). A two-component model was used to illustrate that the best possible rate of convergence of a mixing distribution is only $n^{-1/4}$ when the true mixing distribution is a step function, with the help of the local asymptotic normality. van de Geer (1996) considered the maximum likelihood estimator of an unknown density in a given convex class of densities. In mixture models, it was shown that the dimension of a collection of kernels and the behaviour of the unknown density near zero determine the rate of convergence of the estimated density in the Hellinger distance. The comparability between the Hellinger distance and the KL divergence from Ibragimov and Has'minskii (1981) was also mentioned. Ghosal and van der Vaart (2001) studied the rate of convergence of the NPMLEs when the densities are location or location-scale normal mixtures with the scale parameter lying between two positive numbers. Assumptions are also made that the true mixing distribution is either compactly supported or having sub-Gaussian tails. Bounds for Hellinger bracketing entropies were derived. From these bounds, the convergence rate of the (sieve) MLE in the Hellinger distance was obtained. Zhang (2009) studied the NPMLE of location and location-scale normal mixtures. A large deviation inequality was used to provide the rate of convergence in the Hellinger distance when the mixing distribution has a finite p-th weak moment and when the mixing distribution has an exponential tail uniformly. Dattner et al. (2011) studied the minimax complexity when the unknown distribution has a density belonging to the Sobolev class and the error density is ordinary-smooth. A discrete deconvolution model was also mentioned. If the distributions of the unknown distribution and the error are compactly supported and the underlying distribution is discrete, a normal density with standard deviation tending to 0 can be used to approximate the discrete distribution with arbitrary accuracy. Nguyen (2013) investigated the convergence behaviour of a latent mixing distribution that arises in finite and infinite mixture models using the transportation distance. The relationship to Hellinger and KL distances was also

discussed using various identifiability conditions. The convergence rates of posterior distributions for latent mixing measures were established, for both finite mixtures of multivariate distributions and infinite mixtures based on the Dirichlet process. Heinrich and Kahn (2018) corrected the rate of convergence obtained in Chen (1995) around a given mixing distribution under some regularity and strong identifiability conditions. It was also mentioned that there exist estimators with a non-uniform point-wise rate of estimation of $n^{-1/2}$ for all mixing distributions with a finite number of components. Saha and Guntuboyina (2020) studied finite sample results on the Hellinger accuracy of the densities estimated using the NPMLE. In particular, densities estimated using the NPMLE achieve a near parametric risk when the true density is a discrete Gaussian mixture without any prior information on the number of mixture components.

Chapter 3

Multiple Hypothesis Testing

3.1 Introduction

In this chapter, we present a new method that is intended to well estimate the proportion of null effects as well as the true distribution of the test statistics. The proposed method can produce a class of estimates of the true distribution of the test statistics for all the hypotheses by using non-parametric maximum likelihood estimation. This class of estimates naturally incorporates the proportion of null effects in the mixing distribution which is easy to interpret. The most appropriate model can be chosen based on various threshold functions which can be subjective depending on the particular problem. All the estimators of the mixing distribution given the proportion of null effects less than or equal to the true proportion are always consistent, and the consistency of the null proportion estimator can be shown in some circumstances.

The rest of this chapter is organised as follows. A literature review on multiple hypothesis testing is given in Section 3.2. Section 3.3 describes the development of the method while Section 3.4 concerns the consistency of the restricted NPMLE and the behaviour of the (log)-likelihood with respect to the null proportion. The algorithms needed for this proposed method are described in Section 3.5.

3.2 Literature Review

Multiple hypothesis testing has been an interest to many researchers for decades because of its importance in practical applications. The goal of traditional hypothesis testing is to test an assumption regarding a population using an appropriate statistic. The earliest traditional hypothesis testing theory was established by Fisher (1925) and Neyman and Pearson (1933) where one case is dealt with at a time. Tukey (1953) and

Duncan (1955) proposed multiple comparison procedures (MCPs) to accommodate the need for testing many hypotheses simultaneously.

The classical approach for the MCPs is to control the probability of overall Type I error in families of comparisons (FWER, the family-wise error rate). The "family" here is the collection of hypotheses being considered for joint testing. Saville (1990) recommended a per comparison error rate (PCER) approach instead of the FWER, since the classical procedures, which control the FWER in the strong sense at levels conventional in single-comparison problems, tend to have much less power than the PCER at the same level. In addition, the control of the FWER is important when a conclusion from various individual inferences is likely to be erroneous when at least one of them is. However, in some practical questions, the overall conclusion needs not be erroneous even if some of the null hypotheses are falsely rejected. Benjamini and Hochberg (1995) pointed out the same problem with the FWER and the difficulties with the classical MCPs which caused under-utilisation in applied research. As a result, a different approach was proposed to solve the problems of multiple significant testing, which called for controlling the expected false discovery proportion (FDP) known as the false discovery rate (FDR). Efron et al. (2001) presented an empirical Bayes interpretation of this (global) FDR with an emphasis on the local false discovery rate (fdr) which is the posterior proportion of a z-value. A hypothesis is rejected if the posterior probability of the non-null effect is higher than a threshold. A non-parametric density estimation based on logistic regression with a natural spline was used to estimate the null density as well as the non-null density. It was also mentioned that the empirical Bayes inferences are closely related to the frequentist theory of the FDR in Benjamini and Hochberg (1995). Dudoit et al. (2002) gave an overview of multiple hypothesis testing to date and described a method of estimating adjusted p-values by permutation. Resampling methods can be used to estimate the unadjusted and the adjusted p-values while avoiding parametric assumptions about the joint distribution of the test statistics. The knock-out and transgenic mouse datasets were used to illustrate the use of permutation in multiple hypothesis testing. In Storey (2002b), the positive false discovery rate (pFDR) was introduced. This modified version of the FDR, was claimed to be arguably more appropriate error measure to use. The term "positive" was used to reflect the fact that the FDR is conditioned on the event that positive findings have occurred. When controlling the FDR and positive findings have occurred, the FDR is only controlled at level $\alpha/\mathbb{P}(R > 0)$, where $\mathbb{P}(R > 0)$ is the probability that the number of rejections is greater than 0. This is quite dangerous for the FDR but it

is not that case for the pFDR. In a series of papers Storey (2002a) and Storey and Tibshirani (2003), q-values were introduced as the pFDR analogue of p-values for the FDR, a measure of each feature's significance, automatically taking into account the fact that thousands are simultaneously being tested. The q-value is most technically defined as the minimum pFDR at which the feature can be called significant. A general algorithm for estimating q-values was also given. Lehmann and Romano (2005) and Romano and Wolf (2007) pointed out that the FWER is too strict. If the number of simultaneous hypothesis tests is large, one might be willing to tolerate more than one false rejection to increase the ability of the procedure in correctly rejecting false null hypotheses. The control of the FWER should be replaced by the k-FWER, the probability of k or more false rejections. Both single step and step-down procedures that control the k-FWER in finite samples or asymptotically were derived.

A basic question faced at the outset of analysing a large set of testing results is whether there is evidence that any of the alternative hypotheses are true. The decision that a single hypothesis is deemed as significant can be made by a controlling procedure. As marked by Langaas et al. (2005), a reliable estimate of the null proportion is important when we want to assess or control multiple error rates. In order to increase the efficiency of the controlling procedures, the proportion of null effects should be estimated before making an inference about a single hypothesis.

Langaas et al. (2005) gave an overview of the estimation of the null proportion to date and proposed an estimator of the null proportion based on decreasing density estimation. The density of p-values was estimated by the Grenander estimator and the estimate of the null proportion was obtained by the minimum of the Grenander estimator. Meinshausen and Rice (2006) proposed a family of methods focusing on the circumstances when there are only few false null hypotheses under the independence assumption, and the confidence bounds for the proportion of zeros were also given. Following the work done by Meinshausen and Rice (2006), Cai et al. (2007) started from a two-point mixture model and constructed a procedure that attains the optimal rate of convergence under similar assumptions. In two papers by Jin and Cai (2007) and Jin (2008), the empirical characteristic function and Fourier analysis were used to estimate the empirical null distribution and the proportion of non-null effects. It was also proved that the estimator is uniformly consistent over a wide class of parameters. Cao and Kosorok (2011) proposed a consistent estimator of the proportion of alternative hypotheses under the condition of a finite fourth moment. The estimator only depends on the test statistics, and hence a better ranking of the hypotheses can be obtained. By

applying the proposed method, a more precise cut-off can be obtained, as long as the asymptotic distribution of the test statistics is known under the null hypothesis. Liang and Nettleton (2012) mentioned that many methods estimating the proportion of null effects by the number of p-values that exceeds a threshold. A finite sample proof of conservative point estimation of the FDR was given for a fixed threshold. A dynamic adaptive procedure, in which the thresholding parameter is determined by data, will lead to a conservative estimation of null proportions and the FDR. Asymptotic results were also given. Oyeniran and Chen (2016) proposed a finite mixture model using uniform components and a multinomial distribution with a mixing proportion to fit the p-values. It was also mentioned that the multinomial distribution can be viewed as a parametric approximation to the non-parametric unknown density, and hence the proposed method was considered as a non-parametric method. Li and Barber (2017) used the accumulation function to construct a cut off point based on the observations to determine the number of hypotheses that can be interpreted as "discoveries", and hence the proportion of non-null effects can be calculated.

3.3 Method

3.3.1 The problem

The method to be proposed is fairly general and can be applied to many types of tests. To make the presentation simpler, we will focus on the z-test. Some other types of tests can be converted to z-test. For example, it is a common practice to transform the t-statistic to the z-statistic; see page 16 of Efron (2010b). It is also possible to apply the method to original statistics directly, as shown in Section 4.3.

Based on $z_1, \cdots, z_n$ where z_i is an observation/statistic of the group i with mean μ_i, the goal of MCPs is to test a series of hypotheses

$$H_{0i} : \mu_i = 0. \tag{3.1}$$

Under the usual assumption of normal sampling, each z_i is an observation from a normal distribution $\mathcal{N}(\mu_i, 1)$ with density $\phi(\cdot; \mu_i, 1)$. According to the notation for the two-group model in page 18 of Efron (2010b), the n cases are either null or non-null with prior probability

$$\pi_0 = \mathbb{P}(\text{null}), \qquad 1 - \pi_0 = \mathbb{P}(\text{non-null}) \tag{3.2}$$

with each statistic z_i having either the null density $f_0(z) = \phi(z; 0, 1)$ or the non-null density $f_A(z)$. Taking the perspective of the empirical Bayes, let G be the true distribution of non-null effects, which can be understood in the asymptotic sense as

$$G(\mu) = \lim_{n\to\infty} \frac{\sum_{i=1}^{n} \mathbb{1}(\mu \geq \mu_i)\mathbb{1}(\mu_i \neq 0)}{\sum_{i=1}^{n} \mathbb{1}(\mu_i \neq 0)}.$$

The non-null density can thus be written as

$$f_A(z) = \int \phi(z; \mu, 1) dG(\mu).$$

As a result, we have the density for all z-values as

$$\begin{aligned} f(z) &= \pi_0 f_0(z) + (1 - \pi_0) f_A(z) \\ &= \pi_0 \phi(z; 0, 1) + (1 - \pi_0) \int \phi(z; \mu, 1) dG(\mu) \\ &= \int \phi(z; \mu, 1) dH(\mu), \end{aligned} \tag{3.3}$$

where

$$H(\mu) = \pi_0 \mathbb{1}(\mu \geq 0) + (1 - \pi_0) G(\mu)$$

is the distribution of μ_i for all $i = 1, \cdots, n$. The goal here is to estimate the proportion of null effects π_0 and the non-null density f_A. Since μ_i might be different from μ_j for all $i \neq j$ and the true structure of the underlying means is usually not available in the estimation process, a non-parametric method should be used to better estimate f_A, and thus the mixing distribution G (as well as the mixing distribution H) is an infinite-dimensional parameter. In addition, π_0 and f_A should be better estimated simultaneously. Producing their estimates separately might result in an unreasonable model. The remainder of this section describes the difficulties in estimating π_0 and f_A jointly and how we solve them.

3.3.2 Estimation of a Non-parametric Mixture

Estimation of f turns out to be a special non-parametric mixture estimation problem. It is equivalent to finding the non-parametric estimator of the mixing distribution H.

The difficulty in such non-parametric mixture estimation is that the parameter H is completely unspecified.

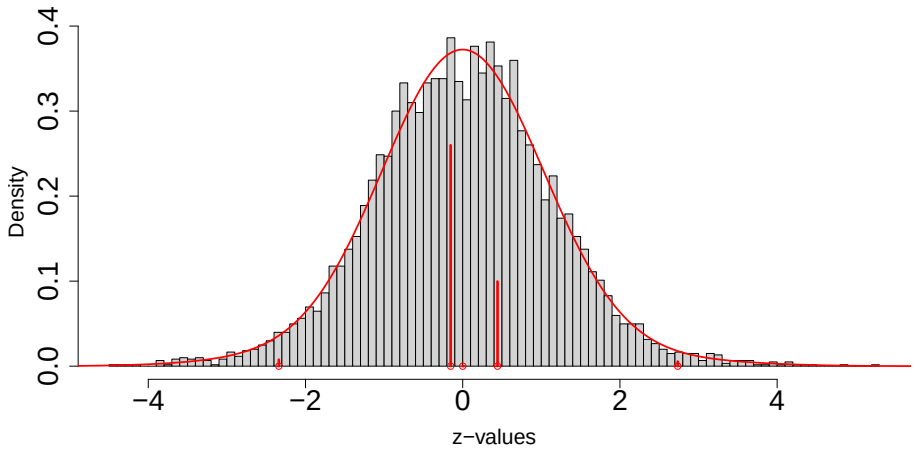

FIGURE 3.1: Histogram of z-values for the prostate data and the estimated density using the NPMLE.

Efron (2010b) discussed making inference using these z-values. These z-values are partitioned into different bins and Poisson regression is used to estimate the count for each bin. The regressors are polynomials of the centre point of each bin. The order is chosen to be 7, which is thought to be roughly a midway between traditional parametric and non-parametric estimation. With the advance of non-parametric mixture estimation, the density of all z-values can be estimated under non-parametric setting easily. In this chapter, only the NPMLE of H is discussed. Distanced-based methods, which are more general, are discussed in Chapter 5. Figure 3.1 shows the histogram of the z-values of the prostate data and the estimated density using the NPMLE. The red vertical lines show the locations and the relative weights of the support points. The NPMLE estimates that the majority of μ_i are close to 0 but there is no positive probability for the support point at 0.

With a finite sample, the NPMLE can not produce an estimate of the mixing distribution that has a positive probability at the support point 0 almost surely. This NPMLE is of little use here since the goal is to find a mixture density naturally containing the proportion of null effects which corresponding to the positive probability for the component centred at 0. It is clear that using the z-values for the prostate data as an example, we can not use the above NPMLE directly to find a good estimate of π_0. We therefore call it the *unrestricted* NPMLE.

It seems to be very difficult to estimate the null proportion. Our proposed approach to estimating the null proportion is as follows:

- For every π_0, compute the NPMLE and obtain the profile likelihood at π_0.
- Obtain an estimate of the null proportion $\hat{\pi}_0$ using the profile-likelihood for all $\pi_0 \in [0, 1]$.

The detailed development of the proposed approach is given in the next two subsections.

3.3.3 Estimation with a Fixed Null Proportion

In this subsection, we discuss the NPMLE with a fixed π_0 and the role of profile likelihood in our approach. Since the unrestricted NPMLE of H cannot produce an estimate with a support point at 0, it is natural to ask if there exists a consistent NPMLE that has a support point at 0 associated with a positive probability? We can achieve this by forcing a fixed positive π_0. Write the density with positive π_0 as

$$f(z; H_{\pi_0}, 1) = \int \phi(z; \mu, 1) dH_{\pi_0}(\mu),$$

where $H_{\pi_0}(\mu) = \pi_0 \mathbb{1}(\mu \geq 0) + (1 - \pi_0) G_{\pi_0}(\mu)$ with $G_{\pi_0} \in \Gamma$ and Γ is a given space of cumulative distribution functions. Indexing G and H by π_0 is to emphasise the fact that H and G varies with π_0. Let $\ell_n(H_{\pi_0})$ be the log-likelihood with n observations, which may be considered having two parameters: $\pi_0 \in [0, 1]$ and $G \in \Gamma$, an infinite-dimensional parameter. We also define the parameter space for H_{π_0}, i.e.,

$$\Gamma_{\pi_0} = \{H : H = \pi_0 \mathbb{1}_0 + (1 - \pi_0) G, G \in \Gamma\}.$$

It can be immediately observed that Γ_{π_0} is nested in Γ, and hence performing maximisation on both parameters (π_0, G) always gives the unrestricted NPMLE which is unable to help detect the proportion of null effects.

As a result, we have to fall back using the NPMLE with a fixed π_0. Let $\tilde{\ell}_n(\pi_0)$ be the profile log-likelihood of π_0, i.e.,

$$\tilde{\ell}_n(\pi_0) = \sup_{G_{\pi_0} \in \Gamma} \sum_{i=1}^{n} \log f(z_i; H_{\pi_0}, 1) = \sum_{i=1}^{n} \log f(z_i; \hat{H}_{\pi_0}, 1) = \ell_n(\hat{H}_{\pi_0}). \tag{3.4}$$

As we will show in Section 3.4 that the NPMLE $\hat{H}_{\pi_0}$ is consistent for all $\pi_0 \in [0, \pi_0^*]$, and that the profile log-likelihood $\tilde{\ell}_{\pi_0}$ is a strictly decreasing function on $[0, 1]$, almost surely. It is expected that these consistent estimators of H^* indexed by π_0 lie in a neighbourhood measured by the (log-)likelihood function. The null proportion can be estimated by defining a boundary of the neighbourhood and checking whether $\tilde{\ell}(\pi_0)$ is in this neighbourhood, which is described in the next subsection.

3.3.4 Estimation of the Null Proportion

In this subsection, a null proportion estimator is proposed and theoretical details of this null proportion estimator are given. We are essentially using generalised likelihood ratio test (GLRT) to obtain an estimate of the null proportion. Given the independence and the identifiability assumptions, we consider testing the hypothesis whether H^* is an element of Γ_{π_0} for a fixed $\pi_0 \in (0, 1]$. The idea is that if H^* is an element in Γ_{π_0}, then the NPMLE $\hat{H}_{\pi_0}$ is an consistent estimator of H^*. Given the observations, the GLRT statistic is defined as

$$t_n(\pi_0) = \tilde{\ell}_n(0) - \tilde{\ell}_n(\pi_0),$$

and the estimator of the null proportion is thus the solution to the following expression,

$$t_n(\hat{\pi}_0) = q(n, \alpha). \tag{3.5}$$

The intuition behind is how similar the model given π_0 is to the unrestricted model which is always strongly consistent. Any model with a log-likelihood value below a certain threshold $\tilde{\ell}_n(0) - q(n, \alpha)$ is regarded as an unreasonable one while any with a likelihood value above as a sensible one. The estimator $\hat{\pi}_0$ is the supremum of π_0 among all sensible models.

The choice of $q(n, \alpha)$ is subjective and depends on the context of a problem. One can choose a constant function $\exp(-q(n, \alpha)) = c$ for $c \in (0, 1)$ which corresponds to the modified maximum likelihood estimator defined in Kiefer and Wolfowitz (1956). In the context of the penalty function for the variable selection, the constant $q(n, \alpha) = 1$ corresponds to the Akaike Information Criterion (AIC; Akaike, 1974) with 1 degree of freedom. Other options are the Bayesian Information Criterion (BIC; Schwarz, 1978) with $q(n, \alpha) = \frac{1}{2} \log n$ and the Hannan-Quinn Criterion (HQC; Hannan and Quinn, 1979) with $q(n, \alpha) = \log \log n$. Theoretically, the constant function can take value

$c = 1$. This corresponds to the unrestricted NPMLE which we do not consider for the reason given earlier. Kiefer and Wolfowitz (1956) mentioned that the NPMLE $\hat{H}_{\pi_0}$ for all π_0 such that $\tilde{\ell}_n(\pi_0) \geq \tilde{\ell}_n(0) - q(n, \alpha)$ are strongly consistent in the infinite-dimensional parameter. However, the consistency of $\hat{H}_{\pi_0}$ does not imply the consistency of its π_0, since the NPMLE $\hat{H}_{\pi_0}$ is consistent for all $0 \leq \pi_0 \leq \pi_0^*$, as we will show in Section 3.4. We will give the theoretical development of consistency of the estimator next.

In order to construct an α-level GLRT test, a critical value $q(n, \alpha)$ should be found such that

$$\mathbb{P}(T_n(\pi_0) > q(n, \alpha)) = \alpha,$$

where

$$T_n(\pi_0) = \sup_{G \in \Gamma} \sum_{i=1}^{n} \log f(Z_i; G, 1) - \sup_{G_{\pi_0} \in \Gamma} \sum_{i=1}^{n} \log f(Z_i; H_{\pi_0}, 1).$$

For parametric models, this likelihood ratio test is based on the fact that the logarithm of the likelihood ratio follows a chi-squared distribution asymptotically and an estimate of the parameter of interest can be found with the quantile of the chi-squared distribution at a pre-specified level. Since H_{π_0} is non-parametric, the usual Wilk theorem can not be used here. Murphy and van der Vaart (2000) investigated the profile log-likelihood in the presence of a non-parametric nuisance parameter, and derived a GLRT for the problem there. Their method is not applicable here since it requires that

$$\hat{\pi}_0 = \arg\max_{\pi' \in [0,1]} \tilde{\ell}_n(\pi')$$

is a consistent estimator of π_0^*. It is shown in Section 3.3.3, it does not converge to π_0^* almost surely as $n \to \infty$. Liu and Shao (2003) showed that the asymptotic distribution of the GLRT statistic is the supremum of a Gaussian process with undiscovered properties while Fan and Jiang (2007) suggested that the Wilk theorem exists in some of the non-parametric cases such as non-parametric regression, varying-coefficient models and varying-coefficient partially linear models but it is not yet clear for the non-parametric mixture estimation. Jiang and Zhang (2016) proposed a GLRT for normal mixtures using a large deviation inequality for the log-likelihood ratio. Jiang and Zhang (2019) showed that the convergence rate is bounded by $\log n$ provided that the support range

of the mixing distribution increases no faster than $(\log n/\log 9)^{1/2}$, under the null hypothesis that all the observations are generated from the standard normal distribution which corresponds to testing $\pi_0^* = 1$ in some sense.

The GLRT proposed in Jiang and Zhang (2016) is used to show the consistency of the estimated null proportions in some circumstances. Define the p-th weak moment as

$$\mu_p(H) = \left\{ \sup_{x>0} x^p \int_{|u|>x} dH(u) \right\}^{1/p},$$

and

$$\mu_p(\Gamma) = \sup_{H\in\Gamma} \mu_p(H).$$

According to Theorem 1 in Jiang and Zhang (2016), if $q(n,\alpha)$ is of equal or smaller order than $t_{n,p}$, where

$$t_{n,p} \asymp \begin{cases} n^{1/(1+p)}(\log n)^{(2+3p)/(2+2p)}, & \mu_p(\Gamma_{\pi_0}) = O(1) \text{ for a fixed } p \\ \log^2 n, & H(M) - H(-M) = 1 \text{ for every } H \in \Gamma_{\pi_0} \end{cases}$$

then under the null hypothesis, there exists a universal constant $k_* > 0$ such that for large n and all $k \geq k_*$,

$$\mathbb{P}(T_n(\pi_0) \geq 3kt_{n,p}) \leq \exp\left(-\frac{kt_{n,p}}{2\log n}\right) \leq n^{-k},$$

and hence the estimator defined in (3.5) can be served as a conservative estimator at α-level due to the use of a large deviation inequality. Further, if $q(n,\alpha)$ is of higher order than $\log n$, the power of the test approaches to 1 as n tends to infinity. If the alternative hypothesis is that H^* has a support point at 0 with probability less than $\pi_0 - \epsilon$ for some small $\epsilon > 0$ independent of n, then Theorem 2 of Jiang and Zhang (2016) shows that the type II error tends to 0 asymptotically hence a consistent estimator of π_0^* can be obtained. This estimator of the null proportion that is consistent but conservative under finite samples should be preferred by practitioners. The estimator acts as an upper bound of the true null proportion under finite sample and converges to the true null proportion asymptotically. Due to the nature of the test statistic $T_n(\pi_0)$, the lower bound of the true null proportion is always 0. The interval $[0, \hat{\pi}_0]$ is analogue to the

confidence interval in traditional hypothesis testing.

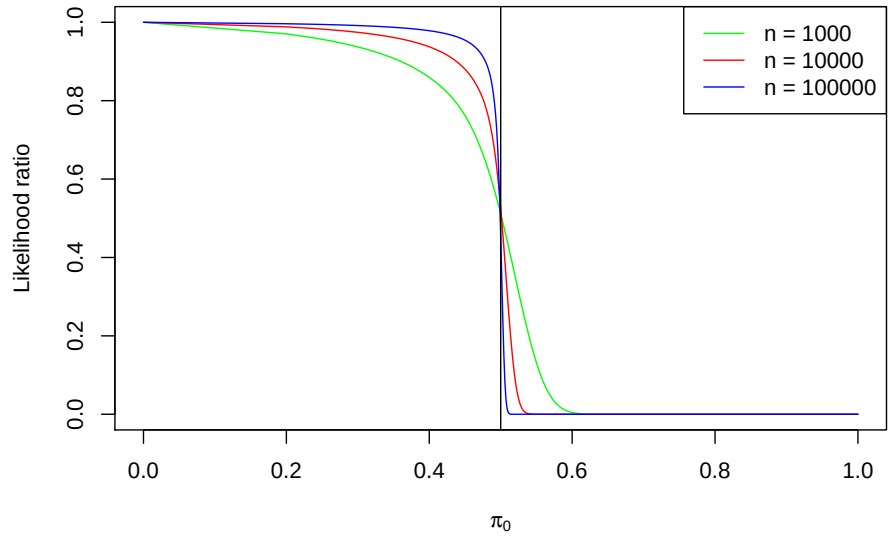

FIGURE 3.2: Plots of the geometric mean of 100 replications of likelihood ratio against π_0 for $n = 1000, 10000$ and 100000.

It remains unclear whether a consistent estimator of $\hat{\pi}_0$ can be obtained if $q(n, \alpha)$ is of equal or smaller order than $\log n$. Figure 3.2 shows the geometric mean of the likelihood ratio as a function of π_0 for various n when the underlying mixing distribution has support point 0 with probability 0.5, -2 with probability 0.25 and 2 with probability 0.25. The geometric mean is used to produce a stable likelihood ratio and the vertical lines shows the true proportion of 0. The convergence to 0 for the geometric mean of the likelihood ratio when $\pi_0 > \pi_0^*$ was shown in Kiefer and Wolfowitz (1956). By a virtual inspection of Figure 3.2, the geometric mean seems to converge to 1 when $\pi_0 < \pi_0^*$, it is equivalent to a flat log-likelihood function in the interval $[0, \pi_0^*)$. The profile likelihood $\tilde{\ell}_n(\pi_0)$ is differentiable with respect to π_0 and the derivative is

$$\frac{\partial \tilde{\ell}_n(\pi_0)}{\partial \pi_0} = \frac{\partial \tilde{\ell}_n(H)}{\partial \pi_0}\bigg|_{H=\hat{H}_{\pi_0}} = \sum_{i=1}^{n} \frac{\phi(z_i; 0, \sigma)}{f(z_i; \hat{H}_{\pi_0}, \sigma)} - n. \tag{3.6}$$

For more detail about the derivative of a profile log-likelihood function with respect to a finite-dimensional structural parameter in the semi-parametric model, see Wang (2010). Our conjecture is hence that the likelihood ratio will converge to an indicator

function with a step down from 1 to 0 at π_0^* asymptotically under a certain condition. In this case, $\hat{\pi}_n$ is consistent for any $\exp(-q(n,\alpha)) \in (0,1)$, and using threshold functions with $o(\log n)$ may produce a consistent estimator of $\hat{\pi}_0$ and it is more efficient under finite samples. However, the exact condition is difficult to obtain since the theory about non-parametric estimation is incomplete, such as the rate of convergence of the NPMLE as well as the rate of convergence of the support points.

3.4 Theory

In this section, we simply focus on the theory when the mixture components are normal; the theory for other families of mixture components can be derived in a similar manner. Proposition 3.1 shows the almost equivalence of the assumptions made on G^* and H^*. Proposition 3.3 is about the consistency of $\hat{H}_{\pi_0}$ for a fixed π_0. The nestedness of our model will be shown in Proposition 3.2, followed by the strict monotonicity of the profile log-likelihood function.

Assumption 3.1 (Independence)**.** *z_i are independent observations from $\mathcal{N}(\mu_i, 1)$.*

Assumption 3.1 is the same as the independence assumption specified in Efron (2010b). It is one of the assumptions made in proving the strong consistency of the NPMLE by Kiefer and Wolfowitz (1956).

Assumption 3.2 (Integrability)**.** *$G^* \in \Gamma$, where*

$$\Gamma = \left\{ G : \int_{-\infty}^{\infty} [\log|\mu|]^+ dG(\mu) < \infty \right\}.$$

Assumption 3.2 gives the restriction to the mixing distribution of the non-null effects when the component density is normal. This assumption certainly can be weakened or tailored to accommodate other component densities or Γ, as long as it meets the requirement in Kiefer and Wolfowitz (1956). Assumption 3.2 is fairly general to accommodate a wide range of common component densities, such as the exponential and Cauchy distributions. Given that $\phi(\cdot;\theta,\sigma)$ satisfies (3.3) and (3.13) of Kiefer and Wolfowitz (1956), Assumption 3.2 can also be applied to H^*. Note that this is the general assumption on the space of mixing distributions. If the space is further restricted, additional conclusions can be made in the context of multiple hypothesis testing, which are discussed in Section 3.3.4. The following proposition shows that applying the restriction to G^* and H^* are almost equivalent.

Proposition 3.1.

$$\int_{-\infty}^{\infty} [\log |\mu|]^+ dG(\mu) < \infty \tag{3.7}$$

if and only if

$$\int_{-\infty}^{\infty} [\log |\mu|]^+ dH_{\pi_0}(\mu) < \infty. \tag{3.8}$$

for all $\pi_0 \in [0, 1)$

Proof.

$$\begin{aligned}\int_{-\infty}^{\infty} [\log |\mu|]^+ dH_{\pi_0}(\mu) &= (1-\pi_0) \int_{-\infty}^{\infty} [\log |\mu|]^+ dG(\mu) + \pi_0 \int_{-\infty}^{\infty} [\log |\mu|]^+ d\mathbb{1}(\mu \geq 0) \\ &= (1-\pi_0) \int_{-\infty}^{\infty} [\log |\mu|]^+ dG(\mu).\end{aligned}$$

For $\pi_0 \in [0, 1)$, the left-hand side of (3.7) is just a finite scalar multiple of the left-hand side of (3.8), and hence one implies the other. □

A side note to Proposition 3.1 is that when $\pi_0 = 1$, (3.7) trivially implies (3.8) but the converse does not necessarily hold since H_1 does not depend on G. That Assumption 3.2 is defined using (3.7) is to accommodate the situation when $\pi_0 = 1$.

Proposition 3.2. $\Gamma_{\pi''} \subset \Gamma_{\pi'}$ *for all* $0 \leq \pi' \leq \pi'' \leq 1$.

Proof. For $G \in \Gamma$, we have

$$\begin{aligned}G''(\mu) &= \pi'' \mathbb{1}(\mu \geq 0) + (1-\pi'')G(\mu) \\ &= \pi' \mathbb{1}(\mu \geq 0) + (\pi'' - \pi')\mathbb{1}(\mu \geq 0) + (1-\pi'')G(\mu) \\ &= \pi' \mathbb{1}(\mu \geq 0) + (1-\pi') \left[\frac{\pi'' - \pi'}{1-\pi'} \mathbb{1}(\mu \geq 0) + \frac{1-\pi''}{1-\pi'} G(\mu) \right] \\ &\in \Gamma_{\pi'},\end{aligned}$$

since the quantity in the square bracket is an element of Γ by Proposition 3.1. □

With Proposition 3.2, the consistency of the NPMLE with a fixed null proportion can be derived in a much simpler way, as shown below.

Proposition 3.3. *Under Assumptions 3.1 and 3.2, the NPMLE $\hat{H}_{\pi_0}$ converges weakly to H^* almost surely for all $\pi_0 \leq \pi_0^*$ and does not converge to H^* for all $\pi_0 > \pi_0^*$.*

Proof. In the case of $\pi_0 > \pi_0^*$, using the metric defined in Kiefer and Wolfowitz (1956), for all $G \in \Gamma$,

$$\begin{aligned} d(H^*, H_{\pi_0}) &= d[\pi_0^* \mathbb{1}(\mu \geq 0) + (1-\pi_0^*)G^*(\mu), \pi_0 \mathbb{1}(\mu \geq 0) + (1-\pi_0)G(\mu)] \\ &= \int |(\pi_0 - \pi_0^*)\mathbb{1}(\mu \geq 0) - (1-\pi_0^*)G^*(\mu) + (1-\pi_0)G(\mu)| e^{-|\mu|} d\tau(\mu) \\ &> 0, \end{aligned}$$

since it is implicit that $\mu_i \neq 0$ for non-null effects by (3.1). The NPMLE $\hat{H}_{\pi_0}$ with $\pi_0 > \pi_0^*$ does not converge to H^*.

In the case of $\pi_0 \leq \pi_0^*$, H^* can be written in the form of H_{π_0},

$$\begin{aligned} H^* &= \pi_0^* \mathbb{1}(\mu \geq 0) + (1-\pi_0^*)G^*(\mu) \\ &= \pi_0^* \mathbb{1}(\mu \geq 0) + (\pi_0^* - \pi_0)\mathbb{1}(\mu \geq 0) + (1-\pi_0^*)G^*(\mu) \\ &= \pi_0^* \mathbb{1}(\mu \geq 0) + (1-\pi_0)\left[\frac{\pi_0^* - \pi_0}{1-\pi_0}\mathbb{1}(\mu \geq 0) + \frac{1-\pi_0^*}{1-\pi_0}G^*(\mu)\right] \\ &= H_{\pi_0} \in \Gamma_{\pi_0}. \end{aligned}$$

The completed space $\bar{\Gamma}_{\pi_0}$ is a subset of $\bar{\Gamma}$ since $\Gamma_{\pi_0} \subset \Gamma$. Under Assumptions 3.1 and 3.2, Proposition 3.1 and the normal component density, $\hat{H}_{\pi_0}$ converges weakly to H^*, almost surely for all $\pi_0 \leq \pi_0^*$ by Kiefer and Wolfowitz (1956). □

Note that the consistency established in Proposition 3.3 is for a fixed π_0. This is important since the identifiability only holds when π_0 is fixed. If π_0 is allowed to vary, then any mixing distribution has infinite number of representations in the function space. This is aligned with the fact that the null proportion must be estimated using the profile log-likelihood.

Proposition 3.3 shows that for a fixed $\pi_0 \leq \pi_0^*$, the NPMLE given π_0 is consistent while it is not consistent if $\pi_0 > \pi_0^*$. The consistency of $\hat{H}_{\pi_0}$ does not imply the consistency of its π_0. Section 3.3.4 explains the method of estimating the null proportion using maximum likelihood.

A direct consequence of Proposition 3.2 is that $\tilde{\ell}_n(\pi') \geq \tilde{\ell}_n(\pi'')$ for all $0 \leq \pi' \leq \pi'' \leq 1$. The following corollary gives a stronger result.

Proposition 3.4. *If $\hat{H}_{\pi_0}$ has a support point at 0 with probability exactly π_0, then for all $\pi' > \pi_0$, $\hat{H}_{\pi'}$ has a support point at 0 with probability exactly π'.*

Proof. Assume

$$\hat{H}_{\pi'}(\mu) = (\pi' + \epsilon)\mathbb{1}(\mu \geq 0) + (1 - \pi' - \epsilon)G(\mu)$$

for some $\epsilon > 0$ and $G \in \Gamma$.

Consider the linear combination of $\hat{H}_{\pi_0}$ and $\hat{H}_{\pi'}$. By the convexity of the log-likelihood and that $\hat{H}_{\pi_0}$ and $\hat{H}_{\pi'}$ are two distinct mixing distributions, we obtain for all $0 < a < 1$,

$$\ell_n(\hat{H}_{\pi'}) < \ell_n(a\hat{H}_{\pi'} + (1 - a)\hat{H}_{\pi_0}).$$

and

$$\frac{\pi' - \pi_0}{\pi' + \epsilon - \pi_0}\hat{H}_{\pi'} + \frac{\epsilon}{\pi' + \epsilon - \pi_0}\hat{H}_{\pi_0} \in \Gamma_{\pi'}.$$

Hence $\hat{H}_{\pi'}$ is not the NPMLE to $\Gamma_{\pi'}$. □

Corollary 3.1. *If $\hat{H}$ has no support point at 0, then $\tilde{\ell}_n$ is a strictly decreasing function on $[0, 1]$.*

The proof of Corollary 3.1 is straightforward. Assume $\tilde{\ell}_n(\pi') = \tilde{\ell}_n(\pi'')$ with $0 \leq \pi' < \pi'' \leq 1$. By Proposition 3.4 and the uniqueness of the NPMLE in $\Gamma_{\pi'}$, $\hat{H}_{\pi''}$ is the same mixing distribution as $\hat{H}_{\pi'}$ hence $\hat{H}_{\pi''}$ should have the support point 0 with probability exactly π' which contradicts with $\hat{H}_{\pi''} \in \Gamma_{\pi''}$ by Proposition 3.2.

3.5 Computation

This section contains some algorithms used in the proposed method. Section 3.5.1 illustrates the algorithm used to compute the NPMLE $\hat{H}_{\pi_0}$ given a fixed null proportion π_0 mentioned in Section 3.3.3. The estimation of $\hat{\pi}_0$ given the family of estimates with respect to π_0 using the restricted NPMLE discussed in Section 3.3.4 is explained in Section 3.5.2.

3.5.1 Computing the Restricted NPMLE

Algorithms for computing the unrestricted NPMLE cannot be used directly to find the restricted NPMLE $\hat{H}_{\pi_0}$ with a fixed π_0. However, we observe that $G_{\pi_0} \in \Gamma$ and thus $\hat{G}_{\pi_0}$ is an unrestricted NPMLE. Modifications should be made so that f is written in terms of G_{π_0} and $\hat{G}_{\pi_0}$ can be computed using the algorithms for the unrestricted NPMLE, followed by transforming $\hat{G}_{\pi_0}$ to $\hat{H}_{\pi_0}$.

Implementation on the `nspmix` package

In order to compute the restricted NPMLE using the algorithm for the unrestricted NPMLE, the log-likelihood (objective) function needs to be modified. We start with the density function for all the observations,

$$\begin{aligned} f(x; H_{\pi_0}, \sigma) &= \pi_0 \phi(x; 0, \sigma) + (1 - \pi_0) \int \phi(x; \mu, \sigma) dG(\mu) \\ &= \int \left[\pi_0 \phi(x; 0, \sigma) + (1 - \pi_0)\phi(x; \mu, \sigma)\right] dG(\mu), \end{aligned}$$

and hence the new component density can be written as

$$f(x; \mu, \sigma, \pi_0) = \pi_0 \phi(x; 0, \sigma) + (1 - \pi_0)\phi(x; \mu, \sigma).$$

Note that the component density is itself a mixture with support points 0 and μ and corresponding probabilities π_0 and $1 - \pi_0$ respectively. Thus the log-likelihood function is

$$\ell_n(G) = \sum_{i=1}^{n} \log \int f(x_i; \mu, \sigma, \pi_0) dG(\mu)$$

and the gradient function that characterises the restricted NPMLE is

$$d(\mu; G') = \sum_{i=1}^{n} \frac{f(x_i; \mu, \sigma, \pi_0)}{f(x_i; G', \sigma, \pi_0)} - n.$$

In order to use the R package `nspmix` to compute the restricted NPMLE, the logarithm of the component density function and its derivative with respect to the location parameter μ are required. For more detail about the implementation of the restricted NPMLE using the R package `nspmix`, see Wang (2007) and the manual in Wang (2017).

Direct Implementation

Implementation of the restricted NPMLE using the existing package is straightforward. However, there may be some issue when using a mixture as the component density. For simplicity, we demonstrate these possible issues using normal components. The logarithm of the component density under the unrestricted NPMLE is

$$\log f(x;\mu,\sigma) = -\frac{(x-\mu)^2}{2\sigma^2} - \frac{1}{2}\log 2\pi\sigma^2,$$

and its derivative with respect to μ is

$$\frac{\partial \log f(x;\mu,\sigma)}{\partial \mu} = \frac{x-\mu}{\sigma^2},$$

while the simplification can not be made in general if the component density is a two-component normal mixture, i.e.,

$$\log f(x;\mu,\sigma,\pi_0) = \log\left[\pi_0\phi(x;0,\sigma) + (1-\pi_0)\phi(x;\mu,\sigma)\right] \tag{3.9}$$

and

$$\frac{\partial \log f(x;\mu,\sigma,\pi_0)}{\partial \mu} = \frac{\frac{x-\mu}{\sigma^2}(1-\pi_0)\phi(x;\mu,\sigma)}{\pi_0\phi(x;0,\sigma) + (1-\pi_0)\phi(x;\mu,\sigma)}. \tag{3.10}$$

Repeatedly evaluating (3.9) and (3.10) also seems to be time-consuming, and the `nspmix` package only deals with computing the NPMLEs, which cannot be extended to other loss functions. This suggests that the computer program for estimating the restricted NPMLE should be constructed directly from the algorithm based on Wang (2007).

The brief steps of the constrained Newton method is already introduced in Chapter 2. In the first step, given the mixing distribution at the current iteration G_s, the gradient function $d(\mu;G_s)$ is needed to compute all local maxima $\mu^*_{s1},\cdots,\mu^*_{sp_s}$. Using the definition of the directional derivative in Lindsay (1983, 1995), we have

$$\begin{aligned} d(\mu;G_s) &\equiv \left.\frac{\partial \ell\{(1-\epsilon)G_s + \epsilon \mathbb{1}_\mu\}}{\partial \epsilon}\right|_{\epsilon=0} \\ &= (1-\pi_0)\sum_{i=1}^{n}\frac{\phi(x_i;\mu,\sigma) - f(x_i;G_s,\sigma)}{f(x_i;G_s,\sigma,\pi_0)}, \end{aligned}$$

and $\mu^*_{s1}, \cdots, \mu^*_{sp_s}$ can be found by partitioning the support space. Different algorithms can be used for finding the local maxima, depending on the differentiability of the gradient function; see Appendix B.

Example 3.1 (Normal distribution)**.** *Since the normal density is infinitely differentiable with respect to μ, in the case of normal mixtures we have*

$$\frac{\partial d(\mu, G_s)}{\partial \mu} = (1-\pi_0)\sum_{i=1}^{n} \frac{(x_i-\mu)\phi(x_i;\mu,\sigma)}{\sigma^2 f(x_i; G_s, \sigma, \pi_0)},$$

and

$$\begin{aligned}\frac{\partial^2 d(\mu, G_s)}{\partial \mu^2} &= \frac{\partial}{\partial \mu}\frac{\partial d(\mu, G_s)}{\partial \mu} \\ &= (1-\pi_0)\sum_{i=1}^{n} \frac{\phi(x_i;\mu,\sigma)[(x_i-\mu)^2-\sigma^2]}{\sigma^4 f(x_i; G_s, \sigma, \pi_0)}.\end{aligned}$$

It is possible to use the Newton-Raphson method to compute the local maxima by finding the roots of $\partial/\partial\mu(d(\mu, G_s)) = 0$ under the constraint that $\partial^2/\partial\mu^2(d(\mu, G_s)) < 0$. Furthermore, according to Jiang and Zhang (2016), the support of $\hat{G}$ under this circumstance is always within the range of the data due to the monotonicity of the normal density in $|x-\mu|$. The support space given the observations can be reduced to the range of the observations.

Example 3.2 (t-distribution)**.** *If ϕ is the non-central t-distribution, i.e.,*

$$\phi(x;\mu,\sigma) = \frac{\sigma^{\frac{\sigma}{2}}\exp\left(-\frac{\sigma\mu^2}{2(x^2+\sigma)}\right)}{\sqrt{\pi}\Gamma\left(\frac{\sigma}{2}\right)2^{\frac{\sigma-1}{2}}(x^2+\sigma)^{\frac{\sigma+1}{2}}}\int_0^\infty y^\sigma \exp\left[-\frac{1}{2}\left(y-\frac{\mu x}{\sqrt{x^2+\sigma}}\right)^2\right]dy.$$

The non-central t-distribution is again infinitely differentiable with respect to μ, we have

$$\frac{\partial d(\mu, G_s)}{\partial \mu} = (1-\pi_0)\sum_{i=1}^{n} \frac{\frac{\partial}{\partial\mu}\phi(x_i;\mu,\sigma)}{f(x_i; G_s, \sigma, \pi_0)},$$

where

$$\frac{\partial}{\partial\mu}\phi(x;\mu,\sigma) = \frac{\sigma^{\frac{\sigma}{2}}\exp\left(-\frac{\sigma\mu^2}{2(x^2+\sigma)}\right)\cdot -\mu}{\sqrt{\pi}\Gamma\left(\frac{\sigma}{2}\right)2^{\frac{\sigma-1}{2}}(x^2+\sigma)^{\frac{\sigma+1}{2}}}\int_0^\infty y^\sigma \exp\left[-\frac{1}{2}\left(y-\frac{\mu x}{\sqrt{x^2+\sigma}}\right)^2\right]dy$$

$$+\frac{\sigma^{\frac{\sigma}{2}}\exp\left(-\frac{\sigma\mu^2}{2(x^2+\sigma)}\right)x}{\sqrt{\pi}\Gamma\left(\frac{\sigma}{2}\right)2^{\frac{\sigma-1}{2}}(x^2+\sigma)^{\frac{\sigma+2}{2}}}\int_0^\infty y^{\sigma+1}\exp\left[-\frac{1}{2}\left(y-\frac{\mu x}{\sqrt{x^2+\sigma}}\right)^2\right]dy.$$

The form for the second derivative will be even more complicated and carries an even heavier computational burden. As a result, for the non-central t-distribution, a derivative-free maximisation algorithm is recommended. Appendix A describes a quick way to evaluate t-density and its first derivative when the degree of freedom is small using integration by parts. According to Aubel and Gawronski (2003), we have the following relationship between the mode of the density and the non-centrality parameter,

$$\sqrt{\frac{2\sigma}{2\sigma+5}}\mu < \textit{Mode of } X < \sqrt{\frac{\sigma}{\sigma+1}}\mu,$$

hence the support space of the non-centrality parameter given the set of observations can be obtained.

These local maxima are then added to the current support set as shown in Step 2 of Wang (2007). The second step is essentially to compute the weights that maximise the log-likelihood function given that the support points are fixed. We also use a quadratic approximation to the log-likelihood function described in Wang (2007). Let $\boldsymbol{\mu}' = (\mu'_1, \cdots, \mu'_{k'})^T$ be the vector of all the points in the support set at the current iteration and $\boldsymbol{\eta}' = (\eta'_1, \cdots, \eta'_{k'})^T$ be the vector of weights corresponding to $\boldsymbol{\mu}'$ and $\boldsymbol{\eta} = (\eta_1, \cdots, \eta_{k'})^T$ corresponds to the vector of weights at the previous iteration with zero probability for the support points added at the current iteration. The second order Taylor expansion of the log-likelihood function around $\boldsymbol{\eta}$ is

$$\ell_n(G_s) = \mathbf{1}^T\boldsymbol{S}(\boldsymbol{\eta}'-\boldsymbol{\eta}) - \frac{1}{2}(\boldsymbol{\eta}'-\boldsymbol{\eta})^T\boldsymbol{S}^T\boldsymbol{S}(\boldsymbol{\eta}'-\boldsymbol{\eta}), \tag{3.11}$$

where

$$\begin{aligned}\boldsymbol{S}^T &= (\boldsymbol{s}_1, \cdots, \boldsymbol{s}_n),\\ \boldsymbol{s}_i &= \left(\frac{\phi(x_i;\mu'_1,\sigma)}{f(x_i;G_{s-1},\sigma,\pi_0)}, \cdots, \frac{\phi(x_i;\mu'_{k'},\sigma)}{f(x_i;G_{s-1},\sigma,\pi_0)}\right)^T.\end{aligned}$$

Let

$$\boldsymbol{s}_{\pi_0} = \left(\frac{\pi_0\phi(x_1,0,\sigma)}{f(x_1;G_{s-1},\sigma,\pi_0)}, \cdots, \frac{\pi_0\phi(x_n,0,\sigma)}{f(x_n;G_{s-1},\sigma,\pi_0)}\right)^T,$$

and the following equations can be obtained,

$$\boldsymbol{S}\boldsymbol{\eta} + \boldsymbol{s}_{\pi_0} = \mathbf{1},$$
$$\mathbf{1}^T\boldsymbol{\eta}' = \mathbf{1}^T\boldsymbol{\eta} = 1 - \mathbf{1}^T\boldsymbol{\pi}.$$

Maximising (3.11) can be turned in to the following optimisation problem

$$\underset{\boldsymbol{\eta}'}{\text{Minimise}} \qquad ||\boldsymbol{S}\boldsymbol{\eta}' - \mathbf{2} + \boldsymbol{s}_{\pi_0}||^2, \tag{3.12}$$

subject to

$$\mathbf{1}^T\boldsymbol{\eta}' = 1 - \mathbf{1}^T\boldsymbol{\pi},$$
$$\eta_i' \geq 0, \ \forall i.$$

This minimisation problem is identical to the semi-parametric problem in Wang and Stephen (2013) and can be solved by the NNLS algorithm in Lawson and Hanson (1974) after the translation suggested by Dax (1990). The program for solving this minimisation problem is implemented as `pnnls` in the R package `lsei` (Wang et al., 2020).

Finally, to ensure the monotone increase and global convergence, a line search is needed. A line search is a search scheme in unconstrained minimisation problems based on the Armijo rule in Armijo (1966). It helps to determine the maximum movement along a given direction such that the objective function is decreasing. Let $G_s^{c,l}$ be the mixing distribution with support points $\boldsymbol{\mu}'$ and the corresponding probabilities $[\boldsymbol{\eta} + c^l(\boldsymbol{\eta}' - \boldsymbol{\eta})]/(1 - \mathbf{1}^T\boldsymbol{\pi})$. Based on the inequality for the unrestricted NPMLE in Wang (2007), the new inequality

$$\ell_n(G_s^{c,l}) \geq \ell_n(G_s^{0,l}) + \alpha c^l \mathbf{1}^T\boldsymbol{S}(\boldsymbol{\eta}' - \boldsymbol{\eta})$$

is tested in the order $l = 0, 1, 2, \cdots$ until it is first satisfied and the resulting $G_s^{c,l}$ is taken as G_s. The values for α and c are taken as the same in Wang (2007). For more details on the line search, see Wang (2007) and Wang and Stephen (2013) and the references therein.

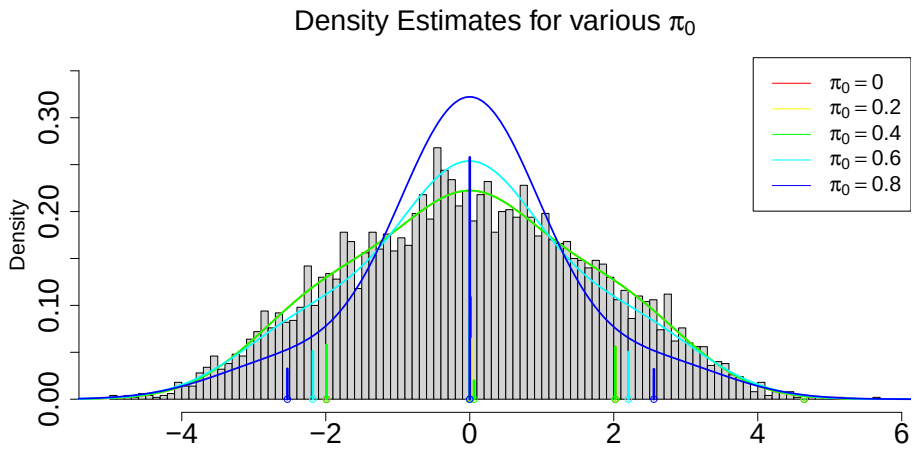

FIGURE 3.3: Plots of estimated densities using the restricted NPMLE for various π_0 when the component distribution is normal.

Figure 3.3 shows the estimated densities using the restricted NPMLE for $\pi_0 = 0, 0.2, 0.4, 0.6$ and 0.8 when the underlying mixing distribution has support point 0 with probability 0.5, -2 with probability 0.25 and 2 with probability 0.25. The locations of the vertical lines are the support points and the height of the lines indicates the corresponding weight. There is no virtual difference among the estimated densities when π_0 is fixed at 0, 0.2 and 0.4 since they are all consistent estimators. The density estimator are not consistent when π_0 is 0.6 or 0.8.

3.5.2 Computing the Null Proportion Given $q(n, \alpha)$

This section describes how to find $\hat{\pi}_0$ given $q(n, \alpha)$. The unrestricted NPMLE is the same as the restricted NPMLE when $\pi_0 = 0$, by the strict monotonicity of the (log-)likelihood, for a fixed $q(n, \alpha)$, the estimate of the null proportion $\hat{\pi}_0$ can be obtained by finding the root of

$$\tilde{\ell}_n(\pi_0) - \tilde{\ell}_n(0) = -q(n, \alpha),$$

and the solution is unique under finite samples. If $\hat{\pi}_0 \notin (0, 1)$, then set $\hat{\pi}_0 = 1$. The profile log-likelihood is differentiable and the derivative is given in (3.6). An efficient root-finding algorithm which utilises the derivative functions can be used, e.g. the secant, the Newton-Raphson or even the Halley method; see Appendix B for the details on implementation. A good initial value to start the root-finding algorithm, if

it requires one, can be

$$\pi_{initial} = \frac{\phi(0;0,\sigma)}{\int \phi(0;\mu,\sigma)d\hat{H}(\mu)},$$

where $\hat{H}$ is the unrestricted NPMLE. If allowed, the restricted NPMLE computed from the previous iteration can be used as a starting mixing distribution at the current iteration with no or minor modification. The performance is expected to be improved since the restricted NPMLEs are similar in a small neighbourhood of the current value π_0 by the continuity assumption in Kiefer and Wolfowitz (1956).

3.6 Conclusion

In this chapter, we introduce a new method of estimating the proportion of null effects as well as the distribution of the test statistics using maximum likelihood. The proposed method in principle can be applied to many types of tests besides the z-statistics, as long as the true distribution G and the component density jointly satisfy the assumptions mentioned in Kiefer and Wolfowitz (1956). This new method produces a class of estimates that naturally incorporate the proportion of null effects in the process of estimating the mixing distribution. It is shown that the restricted NPMLE $\hat{H}_{\pi_0}$ for a fixed π_0 is strongly consistent when $\pi_0 \leq \pi_0^*$, and inconsistent if otherwise. A threshold function depending on the context of the problem can be then used to determine the estimated null proportion. Attempts are made to study the asymptotic properties of the proposed method specifically for estimating the null proportion. Due to the lack of regularity conditions for the mixture models, it is still hard to obtain an efficient estimator of π_0 for inferences, the asymptotic form of the likelihood ratio as a function of π_0 and the exact conditions that the asymptotic form of the likelihood ratio is a step function.

Various threshold functions used in variable selection are introduced in this chapter and the question one may ask is that which threshold function will give the "best" performance for inference when one has no prior knowledge about these information criteria? It is hard to find an answer due to the incompleteness of the asymptotic theory for (log-)likelihood ratio under non-parametric setting. One of the possibilities is that we can take the analogue of the median estimator in distance-based methods shown later in Section 5.6.

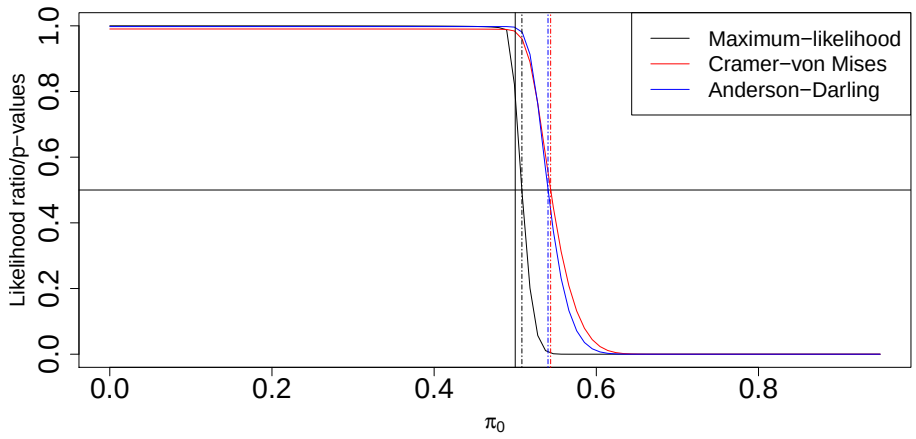

FIGURE 3.4: Plot of the likelihood ratio for the likelihood-based method and the p-value for the Cramér-von Mises and the Anderson-Darling distance against the null proportion.

Figure 3.4 shows the plot of likelihood ratio against π_0 for the likelihood-based method and p-values against π_0 for the Cramér-von Mises and the Anderson-Darling distance. The underlying mixing distribution has support point 0 with probability 0.5, -2 with probability 0.25 and 2 with probability 0.25. An interesting observation from this figure is that the likelihood ratio as a function of π_0 is similar to the p-values with respect to π_0. One may want to consider using $q(n, \alpha) = -\log 0.5$ as the likelihood-based analogue of the median estimator in the distance-based methods. The dashed vertical lines with various colours indicate the estimated proportions of null effects for various methods when the likelihood ratio and the p-values are 0.5. It can be seen that in this particular dataset, using $q(n, \alpha) = -\log 0.5$ gives a reasonably good estimate. However, this statement lacks theoretical justifications completely. If one can find the asymptotic distribution of the (log-)likelihood ratio especially under the presence of the non-parametric nuisance parameter, the median estimator can easily be obtained and it is expected that these estimators should be more efficient and justified in the estimation.

Chapter 4

Numerical Studies For Estimation of the Null Proportion

4.1 Introduction

In this chapter, we use synthetic datasets generated from real world datasets to investigate the performance of our proposed method. Three existing methods are used to compare the performance with the proposed method. The first one is the computation of the local discovery rate (locfdr) described in Efron (2004, 2007a,b, 2010b) which is implemented in the R package `locfdr` (Efron et al., 2015). The second one is the estimation of the proportion of non-null effects using the empirical characteristic function and the Fourier analysis (ECF) described in Jin and Cai (2007) and Jin (2008), where the code is available on the author's homepage. The third one is the empirical Bayes mixture model (mixfdr) in Muralidharan (2010) which is implemented in the R package `mixfdr` (Muralidharan and Efron, 2012). The mixfdr method essentially uses parametric models to estimate the density and requires the number of normal components as an input. As mentioned in Muralidharan (2010), the performance is insensitive to the choice of the parameter and taking the number of normal components as low as 3 gives the same excellent performance. The number of normal components is therefore kept at 3 which is the default setting in the implementation. All of the ECF, the locfdr and the mixfdr method have the capability of estimating the null distribution and computing the proportion of null effects given the null distribution. However, given the nature of the problem in the context of the MCPs, the null distribution is restricted to the standard normal distribution. For the proposed method described in Chapter 3, $q(n, \alpha) = 1$ (the AIC threshold function) and $q(n, \alpha) = \frac{1}{2} \log n$ (the BIC threshold function) are used as the threshold functions when estimating the null proportion. All

the numerical studies in this chapter are performed in R (R Core Team, 2020), version 4.0.0.

The numerical study using the prostate dataset in Section 4.2 focuses on the circumstances when the component distribution is normal or close to normal. The numerical study in Section 4.3, which uses the breast cancer dataset, concerns the estimation of the mixing distribution and the null proportion on the z-values transformed from t-values when the degrees of freedom are small. Synthetic datasets are generated according to the estimates of all the methods on the original dataset. The outline of the numerical studies in Sections 4.2 and 4.3 is as follows:

Step 1: Given a real world dataset, the proportion of null effects is estimated using various methods. A suitable π_0 in the range of the estimated proportions of null effects from various methods is chosen and the restricted NPMLE given π_0 is computed.

Step 2: Synthetic datasets are generated from the mixture based on the restricted NPMLE. In total, 100 synthetic datasets are generated independently for each n.

Step 3: For each synthetic dataset, the proportion of null effects is estimated for each method and all the estimated null proportions are reported.

In Section 4.4, a more systematic numerical study on a two-component model is carried out in order to show the performance of each method when the null proportion and the location of the non-null effect vary. Generation of the synthetic datasets in this section is the same as Steps 2 and 3 in the aforementioned outline.

For the remainder of this chapter, the direct implementation in Section 3.5.1 is used to compute the restricted NPMLE. Two types of component distribution are implemented: normal and t-distribution. The maximisation of (log-)likelihood is converted to a minimisation problem. When finding new support points in every iteration, which is equivalent to finding the local minima with negative gradient, the support space is partitioned into smaller intervals and an algorithm depending on the component distribution is used. For the NPMLE using normal components, local minima are found by computing roots of the first derivative of the gradient function, while the derivative-free minimisation is used to find the local minima of the gradient function when t-components are used. A detailed description of the algorithms implemented can be found in Appendix B. To speed up the process of estimating the null proportion, the NPMLE from the previous iteration is taken as a starting mixing distribution for the new iteration.

4.2 The Prostate Data

In this section, the performance of our proposed method is investigated using z-values converted from t-values when the degrees of freedom are large. The dataset used in this section is the z-values of the prostate data in Singh et al. (2002) and can be obtained in the supplementary material of Efron (2010b). The prostate data concerns the genes of which expression levels are different between the prostate and the normal patients. The data consists of 6033 gene expressions from 102 patients. The first 50 patients are the normal controls while the last 52 are the cancer patients. The goal is to investigate the proportion of null effects (or equivalently, the proportion of the non-null effects) and identify the significant z-values which may lead to the significant gene expressions. Due to the negligible difference between the t-distribution with large degrees of freedom and the standard normal distribution, generation of the synthetic datasets and estimation of the NPMLEs are performed using normal components.

Our proposed method estimates the proportion of null effects to be 0.939 and 0.963 respectively for the AIC and the BIC threshold function while the locfdr, the ECF and the mixfdr method produce 0.932, 0.890 and 0.971 as the estimated null proportion respectively. We fix $\pi_0 = 0.9$ and obtain the estimated mixing distribution for the non-null effects $\hat{G}_{0.9}$. It is particularly interesting to see the performance of our proposed method when the proportion of null effects varies, and hence the model to generate the synthetic datasets is

$$H(\mu) = \pi_0 \mathbb{1}(\mu \geq 0) + (1 - \pi_0)\hat{G}_{0.9},$$

where $\pi_0 = 0.5, 0.7$ and 0.9. In particular, for each synthetic dataset, 6000 observations are generated from the normal mixture with means -2.27, -0.87, 0, 1.14 and 2.75 with their corresponding probabilities. Synthetic datasets with different number of observations (1500 and 3000) are also generated in order to investigate the behaviour of estimating π_0 with changes in n.

π_0	n	npnorm-AIC	npnorm-BIC	locfdr	ECF	mixfdr
0.5	1500	0.675(0.005)	0.732(0.003)	0.739(0.001)	**0.671**(0.003)	0.836(0.001)
	3000	**0.653**(0.006)	0.712(0.003)	0.739(0.001)	0.660(0.003)	0.835(0.001)
	6000	**0.636**(0.006)	0.695(0.002)	0.739(0.001)	0.651(0.002)	0.835(0.001)
0.7	1500	0.813(0.004)	0.860(0.002)	0.839(0.002)	**0.800**(0.003)	0.913(0.001)
	3000	0.806(0.004)	0.847(0.002)	0.839(0.001)	**0.796**(0.003)	0.912(0.001)
	6000	0.795(0.004)	0.836(0.002)	0.839(0.001)	**0.792**(0.002)	0.911(0.001)
0.9	1500	0.946(0.003)	0.973(0.001)	0.947(0.002)	**0.932**(0.003)	0.976(0.000)
	3000	0.942(0.002)	0.966(0.001)	0.947(0.002)	**0.931**(0.002)	0.973(0.000)
	6000	0.940(0.002)	0.961(0.001)	0.947(0.001)	**0.929**(0.002)	0.972(0.000)

TABLE 4.1: The mean (standard error) of estimated null proportions using the z-values from 100 synthetic datasets based on the prostate data for various n and π_0.

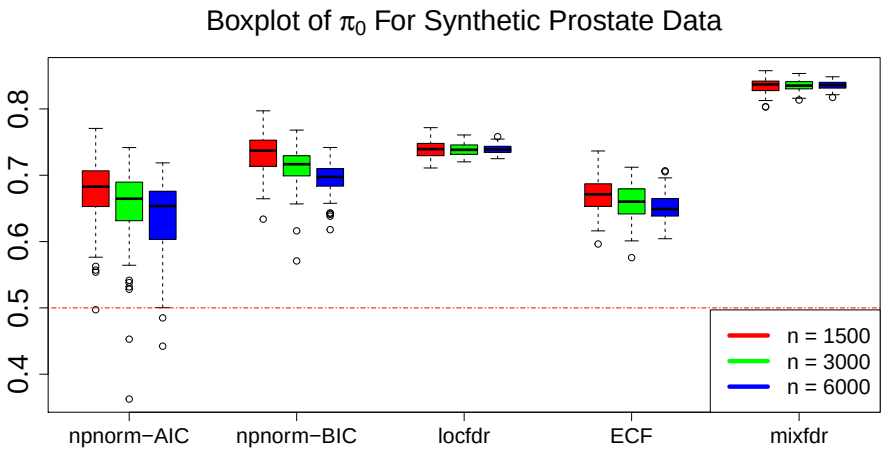

FIGURE 4.1: Boxplot of estimated null proportions for the 100 synthetic datasets generated from the prostate data for each method and n when $\pi_0^* = 0.5$.

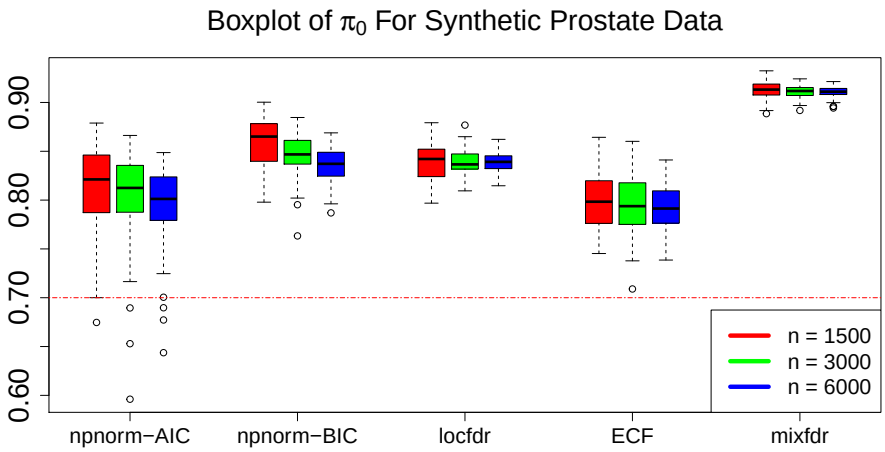

FIGURE 4.2: Boxplot of estimated null proportions for the 100 synthetic datasets generated from the prostate data for each method and n when $\pi_0^* = 0.7$.

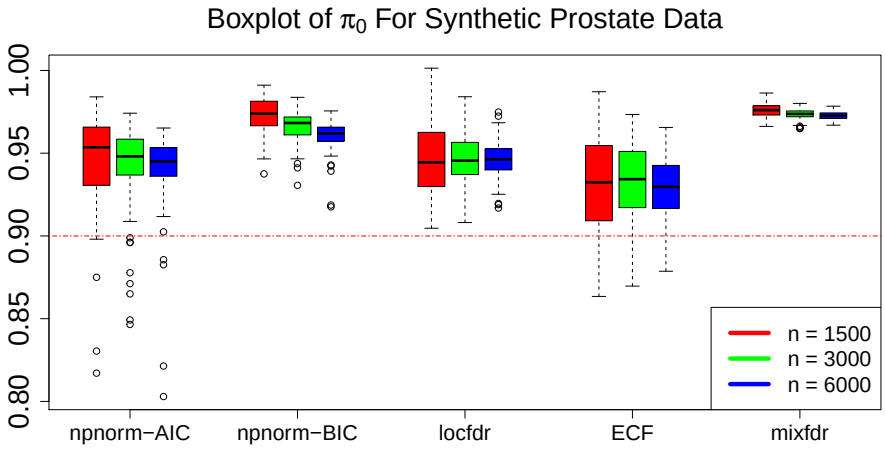

FIGURE 4.3: Boxplot of estimated null proportions for the 100 synthetic datasets generated from the prostate data for each method and n when $\pi_0^* = 0.9$.

Table 4.1 shows the means and standard errors of the estimated null proportions for various methods and numbers of observations while Figures 4.1, 4.2 and 4.3 show the corresponding boxplots. The horizontal line in the boxplots is the true null proportion. The prefix "npnorm" represents that the proportion of null effects is estimated using normal components. The suffixes "AIC" and "BIC" correspond to the threshold functions used in the proposed method when estimating the null proportions. Similar

to the effects of different penalty functions in variable selection, it is expected that the AIC threshold function produces an estimate closer to the true null proportion with a bigger variance while the BIC threshold function produces a conservative estimate with a smaller variance. The estimates of the null proportion using the locfdr and the mixfdr method are relatively constant when the number of observations increases, for all π_0. The estimated means using the proposed method with both the AIC and the BIC threshold function seem to have significant converging trends as n increases. The ECF method gives the best estimated means in terms of the distance to the true proportion of null effects overall for $n = 1500, 3000, 6000$ but our proposed method seems to have a better convergence rate than the ECF method. It is expected that our proposed method may out-perform the ECF method when n increases further; see Section 6.7 for a large-scale comparison.

4.3 The Breast Cancer Data

In this section, our main focus is on a more practical aspect: whether transforming the statistics will affect the estimation of the null proportion. This problem arises from the fact that t-values are often transformed into z-values when conducting two-sample t-tests even when the degrees of freedom are small.

The dataset used to generate the synthetic datasets is the microarray experiment about the difference between genetic mutations causing breast cancers in Efron (2004), the dataset can be obtained from the R package `nspmix` (Wang, 2017). There are 15 breast cancer patients in total with 7 BRCA1s and 8 BRCA2s. For the definition of BRCA1 and BRCA2; see Efron (2003). 3226 genes were reported in the microarray for each woman and z-values can be obtained from the two-sample t-tests comparing 7 BRCA1 responses and 8 BRCA2 responses for each gene with degrees of freedom 13 and apply the transformation

$$z_i = \Phi^{-1}(F_{13}(t_i); 0, 1), \tag{4.1}$$

as shown in Efron (2004). Estimation of the null proportion on t-values using t-components is compared to the corresponding estimation methods applied on the z-values transformed from the t-values using normal components.

First of all, we need to check whether the consistency holds for the NPMLE using t-components. The t-distribution with 1 degree of freedom is the Cauchy distribution

which has heavy tails. As the degrees of freedom increase to infinity, it becomes the standard normal distribution which has exponential tails. According to Kiefer and Wolfowitz (1956) and Proposition 3.3, the consistency of the NPMLE holds for all $\pi_0 \leq \pi_0^*$ when the component density is Cauchy under Assumption 3.2, and hence the consistency of the NPMLE holds for any t-distribution with degrees of freedom greater than 1.

	npnorm-AIC	npnorm-BIC	npt-AIC	npt-BIC	locfdr	ECF	mixfdr
$\hat{\pi}_0$	0.484	0.548	0.551	0.633	0.718	0.565	0.809

TABLE 4.2: Estimated null proportions using the z- or the t-values where appropriate of the breast cancer data.

Table 4.2 shows the estimated null proportions for various methods on the original dataset. Since the two-sample t-test is used to generate the original dataset, the distribution used to generate the synthetic datasets is the t-mixture based on the restricted NPMLE $\hat{H}_{0.6}$. In particular, for each synthetic dataset, 3000 observations in total are generated from the t-mixture with non-centrality parameters -1.81, 0, 1.76 and 4.13 with their corresponding probabilities, and the degrees of freedom are 13. Synthetic datasets with different numbers of observations (750 and 1500) are also generated in order to investigate the behaviour of estimating the null proportion with changes in n.

n	npnorm-AIC	npnorm-BIC	npt-AIC	npt-BIC
750	0.555(0.008)	0.647(0.006)	**0.641**(0.006)	0.713(0.004)
1500	0.537(0.007)	**0.617**(0.004)	0.632(0.006)	0.695(0.004)
3000	0.517(0.005)	0.586(0.003)	0.621(0.004)	0.672(0.003)
n	locfdr	ECF	mixfdr	
750	0.731(0.002)	0.644(0.003)	0.832(0.002)	
1500	0.732(0.001)	0.623(0.003)	0.831(0.002)	
3000	0.730(0.001)	**0.608**(0.003)	0.832(0.001)	

TABLE 4.3: The mean (standard error) of estimated null proportions using the z- and t-values where appropriate from 100 synthetic datasets based on the breast cancer data for various n.

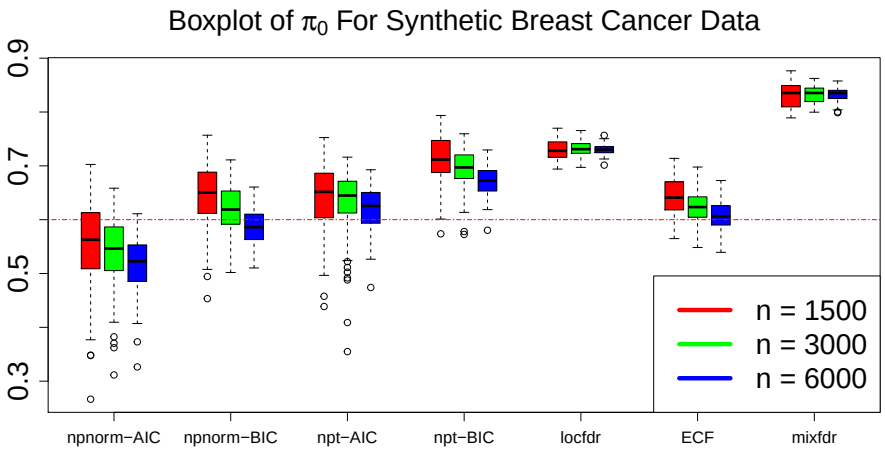

FIGURE 4.4: Boxplot of estimated null proportions for the 100 synthetic datasets generated from the breast cancer data for each method and n.

Table 4.3 shows the means and standard errors of the estimated null proportions for various methods while Figure 4.4 shows the corresponding boxplot. The red horizontal line in the boxplot is the true null proportion. The prefix "npt" represents that the proportion of null effects is estimated using t-components. Overall, for the same threshold function, estimation using t-components performs better than that using normal components. Estimating the mixing distribution using normal components always seems to under-estimate the null proportion, which suggests that converting t-values to z-values may not be a good choice when the degrees of freedom are small.

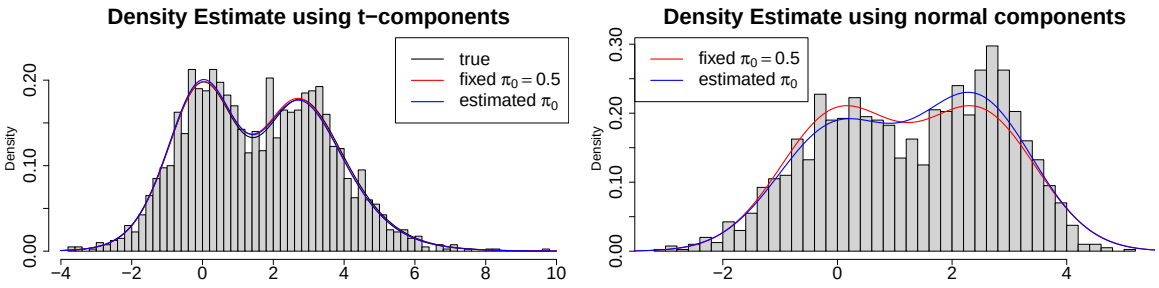

FIGURE 4.5: Plots of the estimated density of 2000 observations using t- and normal components.

Figure 4.5 gives an example where the transformation of t-statistics causes problems in density estimation. The left pane shows the histogram of 1000 observations generated from the central t-distribution and 1000 observations generated from the non-central

t-distribution with the non-centrality parameter 3. The black line is the true density, the red line is the estimated density using the restricted NPMLE based on the true null proportion $\pi_0 = 0.5$ and the blue line is the estimated density when the null proportion is estimated using the AIC threshold function. All three lines are very close to each other, which suggests that there is no issue in estimation. The right pane shows the histogram of the transformed data using (4.1). The red line is the estimated density using the restricted NPMLE based on the true null proportion $\pi_0 = 0.5$ and the blue line is the estimated density when the null proportion is estimated using the AIC threshold function. Both of the estimated densities do not match with the histogram, which suggests that estimation using normal components may be erroneous. The non-null density in the right pane seems to have a higher value at the mode and a smaller standard deviation compared with the standard normal density. The estimating procedure tries to accommodate the observations between 2 and 4 by giving more weight but still fails to produce a reasonable estimate even when the true proportion of null effects is used. This is consistent with results in Figures 4.2 and 4.3 that estimating π_0 using a conservative threshold function still produces a null proportion smaller than the true one. Efron (2010a) also gave some discussions about the non-null distribution of the z-values converted from non-central t-variables.

When conducting the numerical study in this section, we find it computationally much easier to transform t-values to z-values and apply the proposed method with normal components when the degrees of freedom are large, because it is difficult to find an efficient way to evaluate the (log)-density and the derivative of the density for non-central t-distributions. When the degrees of freedom are small, it is still recommended to directly apply the proposed method on t-values. In addition, the estimated mixing distribution when working directly with t-values has a meaningful interpretation. Consider the formulation for the two-sample t-test with the pooled variance,

$$t = \frac{\bar{X}_1 - \bar{X}_2}{s_p\sqrt{\frac{1}{n_1} + \frac{1}{n_2}}},$$

where

$$s_p = \sqrt{\frac{(n_1 - 1)s_{X_1}^2 + (n_2 - 1)s_{X_2}^2}{n_1 + n_2 - 2}},$$

$\bar{X}_1$ and $\bar{X}_2$ are the sample means, $s_{X_1}^2$ and $s_{X_2}^2$ are the sample variances, n_1 and n_2 are

the numbers of observations in two groups respectively. Comparing the formulation of the t-statistic with that of the t-mixture, we can obtain that the support points in the restricted NPMLE $\hat{H}_{\pi_0}$ are the estimate of the scalar multiple $(1/\sqrt{1/n_1 + 1/n_2})$ of the underlying non-centrality parameters. Hence the estimated mixing distribution using t-components has a meaningful interpretation. However, in conventional procedures where t-values are transformed using central t-distribution prior to the estimation process, it is hard to interpret the estimated mixing distribution especially when the degrees of freedom are small, as shown in the right pane of Figure 4.5.

4.4 A Two-Component Model

In this section, a more systematic simulation study is carried out to investigate the performance of the proposed method using a two-component model, assuming the component distribution is normal. The goal in this section is to see the performances among all the methods when the number of observations and the underlying location of the non-null parameter vary.

The two-component model considered is

$$H(\mu) = \pi_0 \mathbb{1}(\mu \geq 0) + (1 - \pi_0)\mathbb{1}(\mu \geq \beta),$$

which the probabilities π_0 and $1 - \pi_0$ are given to the locations at 0 and β respectively. For each combination of (π_0, β), 100 replications are performed with 3000 observations generated in each replication. Choices of the parameters are $\pi_0 = 0.95, 0.9, 0.8, 0.6$ and $\beta = 1, 2, 3, 4$. The methods to be compared here are the proposed method with the AIC, the HQC and the BIC threshold function, the locfdr, the ECF and the mixfdr method.

π_0	AIC	HQC	BIC	locfdr	ECF	mixfdr
			$\beta = 1$			
0.95	**0.964**(0.002)	0.979(0.001)	0.990(0.001)	0.984(0.001)	0.972(0.002)	0.988(0.000)
0.9	**0.904**(0.005)	0.928(0.003)	0.947(0.002)	0.970(0.001)	0.947(0.002)	0.980(0.000)
0.8	**0.810**(0.005)	0.834(0.005)	0.860(0.003)	0.954(0.001)	0.894(0.002)	0.962(0.001)
0.6	**0.606**(0.004)	0.633(0.003)	0.660(0.002)	0.970(0.001)	0.795(0.002)	0.877(0.001)
			$\beta = 2$			
0.95	0.953(0.001)	0.962(0.001)	0.968(0.001)	0.962(0.001)	**0.949**(0.002)	0.974(0.000)
0.9	0.906(0.001)	0.914(0.001)	0.922(0.001)	0.923(0.001)	**0.898**(0.002)	0.945(0.000)
0.8	0.807(0.002)	0.817(0.001)	0.827(0.001)	0.870(0.001)	**0.805**(0.002)	0.878(0.000)
0.6	0.606(0.002)	0.618(0.001)	0.629(0.001)	0.895(0.002)	**0.605**(0.002)	0.733(0.001)
			$\beta = 3$			
0.95	**0.949**(0.002)	0.955(0.001)	0.960(0.000)	0.953(0.001)	0.942(0.002)	0.961(0.000)
0.9	**0.903**(0.001)	0.908(0.001)	0.914(0.000)	0.907(0.001)	0.894(0.001)	0.923(0.000)
0.8	0.803(0.001)	0.811(0.001)	0.818(0.000)	0.819(0.001)	**0.802**(0.001)	0.844(0.000)
0.6	**0.604**(0.001)	0.613(0.001)	0.621(0.001)	0.856(0.001)	**0.604**(0.001)	0.685(0.000)
			$\beta = 4$			
0.95	**0.949**(0.001)	0.954(0.000)	0.958(0.000)	0.958(0.001)	0.942(0.001)	0.956(0.000)
0.9	**0.899**(0.001)	0.906(0.001)	0.911(0.000)	0.908(0.001)	0.895(0.001)	0.916(0.000)
0.8	**0.802**(0.001)	0.809(0.001)	0.815(0.000)	**0.798**(0.001)	0.795(0.001)	0.835(0.000)
0.6	0.605(0.001)	0.612(0.001)	0.619(0.000)	0.842(0.001)	**0.598**(0.001)	0.671(0.000)

TABLE 4.4: The mean (standard error) of estimated null proportions using various methods for each combination of (π_0, β).

Table 4.4 shows the means and standard errors of the estimated proportions of null effects for each combination of (π_0, β). Similar to previous sections, the locfdr and the mixfdr method seem to constantly over-estimate the null proportion while the proposed method produces a better estimate in term of the distance to the true value overall. The proposed method with the BIC threshold function gives a slightly higher estimate with a smaller variance because of the heavy penalty. The ECF method has a performance similar to the proposed method with $q(n, \alpha) = 1$ when β is relatively far away from 0 but when $\beta = 1$, the ECF method also over-estimates the null proportion significantly.

4.5 Conclusion

In this chapter, numerical studies are conducted in order to assess the performance of the proposed method. Overall, when the component distribution is correctly specified,

the proposed method with the AIC threshold function $q(n, \alpha) = 1$ has a competitive performance against the ECF method in terms of the distance to the true null proportion. The numerical study using the breast cancer dataset indicates a potential issue for the z-values converted from t-values when the degrees of freedom are small, which affects the estimation of the proportion of null effects. As we will show in Chapter 6 that our proposed method can be modified in order to accommodate the computation of an even larger dataset. The comparison between the proposed method and the existing methods shows that our proposed method still enjoys a great converging trend while the converging trend for the ECF method does not seem to be obvious, and empirical rates of convergence are estimated. The simulation study using the two-component model suggests that the proposed method performs the best when the location of the non-null effect is close to the location of the null effect, and has a competitive performance otherwise.

Chapter 5

Distance-Based Methods For Estimating a Mixing Distribution

5.1 Introduction

In Chapter 3, a non-parametric method was introduced to estimate a mixing distribution using maximum likelihood with prior knowledge. In this chapter, several distance-based methods are introduced as alternative ways to estimate a mixing distribution as well as the proportion of null effects. The Cramér-von Mises and the Anderson-Darling distance are presented in this chapter, since their estimation processes are easy to implement and the asymptotic properties of the uniformity tests in the standard setting are known under both distances. The Kolmogorov-Smirnov distance is also a common distance measure in testing goodness of fit of a distribution. Finding new support points at each iteration is slightly more complicated and estimating the proportions given fixed support points under the Kolmogorov-Smirnov distance is a linear programming problem, which is rather different from the one in estimating the NPMLE, and hence we are not going to study it in this thesis. For each of the distance-based methods, an algorithm for computing the restricted NPMDE is presented based on the development of the corresponding unrestricted NPMDE. The method of estimating the null proportions is presented in Section 5.4.

In this chapter, we assume here $x_i \neq x_j$ for all i, j. That is, no observations are the same with probability one, which corresponds to the assumption made in Anderson and Darling (1952). The case when $x_i = x_j$ for some i, j is a special case of the large-scale computation and will be discussed in Chapter 6. We use $\ell(\cdot)$ to represent the distance to the ECDF while using $l(\cdot)$ to represent the objective function in computing the NPMDEs.

5.2 The Cramér-von Mises Distance

The Cramér-von Mises test is one of the most common tools for judging the goodness of fit of a cumulative distribution function compared with a given empirical distribution function. The test statistic is also a distance measure which can be used in estimation. In this section, computing the unrestricted and the restricted NPMDE under the Cramér-von Mises distance is discussed as well as some of its theoretical properties.

5.2.1 Computation of the Estimator

The loss function for the Cramér-von Mises distance is

$$\ell(G) = \int [F(\cdot; G, \sigma) - \hat{F}_n]^2 dF(\cdot; G, \sigma).$$

Using the transformation in Anderson and Darling (1954), the unrestricted NPMDE under this metric is the solution to the problem

$$\hat{G} = \arg\min_{G \in \Gamma} \frac{1}{12n^2} + \frac{1}{n} \sum_{i=1}^{n} \left[F(x_{(i)}; G, \sigma) - \frac{2i-1}{2n} \right]^2.$$

In this case we need to only consider minimising the objective function

$$l_n(G) = \sum_{i=1}^{n} \left[F(x_{(i)}; G, \sigma) - \frac{2i-1}{2n} \right]^2.$$

The process of finding the unrestricted NPMDE under this metric is similar to the ones described in Wang (2007) and Chee and Wang (2013). In the first step, given the mixing distribution at the current iteration G_s, the gradient function $d(\mu, G_s)$ is needed for computing all the local minima. For the directional derivative from G' to G'', we have

$$\begin{aligned} d(G''; G') &\equiv \left. \frac{\partial l\{(1-\epsilon)G' + \epsilon G''\}}{\partial \epsilon} \right|_{\epsilon=0} \\ &= \left. \frac{\partial}{\partial \epsilon} \sum_{i=1}^{n} \left[(1-\epsilon) F(x_{(i)}; G', \sigma) + \epsilon F(x_{(i)}; G'', \sigma) - \frac{2i-1}{2n} \right]^2 \right|_{\epsilon=0} \\ &= 2 \sum_{i=1}^{n} (F(x_{(i)}; G'', \sigma) - F(x_{(i)}; G', \sigma)) \left[F(x_{(i)}; G', \sigma) - \frac{2i-1}{2n} \right], \end{aligned}$$

and hence the gradient function is

$$d(\mu, G_s) = 2\sum_{i=1}^{n}(\Phi(x_{(i)};\mu,\sigma) - F(x_{(i)};G_s,\sigma))\left[F(x_{(i)};G_s,\sigma) - \frac{2i-1}{2n}\right]. \tag{5.1}$$

The second step is to compute the weights that minimise the objective function. Let $\boldsymbol{\mu}'$ be the vector of all the support points currently in the support set at iteration s and $\boldsymbol{\eta}'$ be the corresponding weight vector. Then,

$$l_n(G_s) = (\boldsymbol{S}\boldsymbol{\eta}' - \boldsymbol{w})^T(\boldsymbol{S}\boldsymbol{\eta}' - \boldsymbol{w}),$$

where

$$\begin{aligned}
\boldsymbol{S}^T &= (\boldsymbol{s}_1, \cdots, \boldsymbol{s}_n),\\
\boldsymbol{s}_i &= \left(\Phi(x_{(i)};\mu'_1,\sigma), \cdots, \Phi(x_{(i)};\mu'_{k'},\sigma)\right)^T,\\
\boldsymbol{w} &= \left(\frac{1}{2n}, \frac{3}{2n}, \cdots, \frac{2n-1}{2n}\right)^T,
\end{aligned}$$

and hence $\boldsymbol{\eta}'$ can be solved by the NNLS algorithm. Note that the expansion of $l_n(G_s)$ is exact, minimising the objective function at each iteration already ensures the monotone decrease in the objective function, and hence the convergence of the algorithm is guaranteed.

Computing the restricted NPMDE is similar to the process described in Section 3.5.1. The optimisation becomes

$$\hat{G} = \arg\min_{G\in\Gamma}\sum_{i=1}^{n}\left[F(x_{(i)};G,\sigma,\pi_0) - \frac{2i-1}{2n}\right]^2,$$

and hence given the mixing distribution at the current iteration G_s, the gradient function is

$$d(\mu, G_s) = 2(1-\pi_0)\sum_{i=1}^{n}\left(\Phi(x_{(i)};\mu,\sigma) - F(x_{(i)};G_s,\sigma)\right)\left[F(x_{(i)};G_s,\sigma,\pi_0) - \frac{2i-1}{2n}\right].$$

In order to compute the weight vector that minimises the objective function, we use the expansion of $l_n(G_s)$ with respect to $\boldsymbol{\eta}'$, i.e.,

$$l_n(G_s) = (\boldsymbol{S}\boldsymbol{\eta}' + \boldsymbol{s}_{\pi_0} - \boldsymbol{w})^T(\boldsymbol{S}\boldsymbol{\eta}' + \boldsymbol{s}_{\pi_0} - \boldsymbol{w}),$$

where

$$\boldsymbol{s}_{\pi_0} = \big(\pi_0\Phi(x_{(1)},0,\sigma),\cdots,\pi_0\Phi(x_{(n)},0,\sigma)\big)^T,$$

and hence the NNLS algorithm can be used to solve $\boldsymbol{\eta}'$ subject to the constraints

$$\mathbf{1}^T\boldsymbol{\eta}' = 1 - \mathbf{1}^T\boldsymbol{\pi},$$
$$\eta_i' \geq 0,\ \forall i.$$

Example 5.1 (Normal distribution)**.** *If ϕ is the normal density and Φ is the corresponding cumulative distribution function, using the property*

$$\frac{\partial}{\partial\mu}\Phi(x;\mu,\sigma) = -\phi(x;\mu,\sigma),$$

we have for the unrestricted NPMDE,

$$\frac{\partial d(\mu,G_s)}{\partial\mu} = -2\sum_{i=1}^{n}\phi(x_{(i)};\mu,\sigma)\left[F(x_{(i)};G_s,\sigma) - \frac{2i-1}{2n}\right],$$
$$\frac{\partial^2 d(\mu,G_s)}{\partial\mu^2} = -2\sum_{i=1}^{n}(x_{(i)}-\mu)\phi(x_{(i)};\mu,\sigma)\left[F(x_{(i)};G_s,\sigma) - \frac{2i-1}{2n}\right],$$

and

$$\frac{\partial d(\mu,G_s)}{\partial\mu} = -2(1-\pi_0)\sum_{i=1}^{n}\phi(x_{(i)};\mu,\sigma)\left[F(x_{(i)};G_s,\sigma,\pi_0) - \frac{2i-1}{2n}\right],$$
$$\frac{\partial^2 d(\mu,G_s)}{\partial\mu^2} = -2(1-\pi_0)\sum_{i=1}^{n}(x_{(i)}-\mu)\phi(x_{(i)};\mu,\sigma)\left[F(x_{(i)};G_s,\sigma,\pi_0) - \frac{2i-1}{2n}\right],$$

for the restricted NPMDE under the Cramér-von Mises distance.

The procedure above can also be used to find the unrestricted and the restricted NPMDE under the squared distance. The optimisation problem is almost identical except that the terms $(2i-1)/(2n)$ are replaced with i/n.

5.2.2 Theoretical Properties

It is noted that the Cramér-von Mises distance are very similar to the squared distance, we can take the advantage of consistency of the NPMDE under the squared distance

to show the consistency of the NPMDE under the Cramér-von Mises distance.

Proposition 5.1. *Under the same assumptions required by the consistency of the NPMDE for the squared distance in Choi and Bulgren (1968) and Henna (1983), the solution to the following optimisation problem*

$$\hat{G} = \arg\min_{G\in\Gamma} \sum_{i=1}^{n} \left[F(x_{(i)}; G, \sigma) - \frac{2i-1}{2n} \right]^2$$

is a consistent estimator of G.

Proof. Let $\tilde{G}$ be the unrestricted NPMDE under the squared distance, i.e.,

$$\tilde{G} = \arg\min_{G\in\Gamma} \sum_{i=1}^{n} \left[F(x_{(i)}; G, \sigma) - \frac{i}{n} \right]^2.$$

An upper bound can be shown using $\tilde{G}$,

$$\begin{aligned}
&\frac{1}{n}\sum_{i=1}^{n}\left[\frac{2i-1}{2n} - F(x_{(i)}, \hat{G}, \sigma)\right]^2 \\
&\leq \frac{1}{n}\sum_{i=1}^{n}\left[\frac{2i-1}{2n} - F(x_{(i)}, \tilde{G}, \sigma)\right]^2 \\
&= \frac{1}{n}\sum_{i=1}^{n}\left[\frac{2i-1}{2n} + \frac{1}{2n} - \frac{1}{2n} - F(x_{(i)}, \tilde{G}, \sigma)\right]^2 \\
&= \frac{1}{n}\sum_{i=1}^{n}\left[\frac{i}{n} - F(x_{(i)}, \tilde{G}, \sigma)\right]^2 - \frac{1}{n^2}\sum_{i=1}^{n}\left[\frac{i}{n} - F(x, \tilde{G}, \sigma)\right] + \frac{1}{4n^2} \\
&\leq \frac{1}{n}\sum_{i=1}^{n}\left[\frac{i}{n} - F(x_{(i)}, \tilde{G}, \sigma)\right]^2 + \frac{1}{n}\sup_{x\in\mathbb{R}}\left|\frac{i}{n} - F(x_{(i)}, \tilde{G}, \sigma)\right| + \frac{1}{4n^2} \\
&\to 0,
\end{aligned}$$

almost surely, since $\tilde{G}$ is a consistent estimator under the squared distance. We also have

$$\begin{aligned}
&\frac{1}{n}\sum_{i=1}^{n}\left[\frac{2i-1}{2n} - F(x_{(i)}, \hat{G}, \sigma)\right]^2 \\
&= \frac{1}{n}\sum_{i=1}^{n}\left[\frac{i}{n} - \frac{1}{2n} - F(x_{(i)}, \hat{G}, \sigma)\right]^2
\end{aligned}$$

$$
\begin{aligned}
&= \frac{1}{n}\sum_{i=1}^{n}\left[\frac{i}{n} - F(x_{(i)}, \hat{G}, \sigma)\right]^2 - \frac{1}{n^2}\sum_{i=1}^{n}\left[\frac{i}{n} - F(x_{(i)}, \hat{G}, \sigma)\right] + \frac{1}{4n^2} \\
&\geq \frac{1}{n}\sum_{i=1}^{n}\left[\frac{i}{n} - F(x_{(i)}, \hat{G}, \sigma)\right]^2 - \frac{1}{n}\sqrt{\frac{1}{n}\sum_{i=1}^{n}\left[\frac{i}{n} - F(x_{(i)}, \hat{G}, \sigma)\right]^2} + \frac{1}{4n^2} \\
&= \left\{\sqrt{\frac{1}{n}\sum_{i=1}^{n}\left[\frac{i}{n} - F(x_{(i)}, \hat{G}, \sigma)\right]^2} - \frac{1}{2n}\right\}^2 \\
&\to 0,
\end{aligned}
$$

almost surely. The strong consistency of the unrestricted NPMDE under the Cramér-von Mises distance thus holds by Lemma 2 in Henna (1983). □

In the case of the restricted NPMDE, one can always use the nestedness of the restricted space to show the consistency provided that the true mixing distribution is an element in the restricted space. The following proposition gives an alternative proof.

Proposition 5.2. *Under the conditions specified in Henna (1983), if* $H^* \in \Gamma_{\pi_0}$, *then the estimator*

$$\hat{H}_{\pi_0} = \pi_0 \mathbb{1}_0 + (1-\pi_0)\hat{G}_{\pi_0},$$

where

$$\hat{G}_{\pi_0} = \arg\min_{G\in\Gamma} \frac{1}{n}\sum_{i=1}^{n}\left[F(x_{(i)}; G, \sigma, \pi_0) - \frac{2i-1}{2n}\right]^2,$$

is a consistent estimator of H^*.

Proof. Since H^* is an element in Γ_{π_0}, it can be written as

$$H^* = \pi_0 \mathbb{1}_0 + (1-\pi_0)G^*.$$

We can rearrange $F(x; G, \sigma, \pi_0)$ as

$$
\begin{aligned}
F(x; G, \sigma, \pi_0) &= \pi_0 \Phi(x; 0, \sigma) + (1-\pi_0)\int \Phi(x; \mu, \sigma) dG(\mu) \\
&= \int \left[\pi_0 \Phi(x; 0, \sigma) + (1-\pi_0)\Phi(x; \mu, \sigma)\right] dG(\mu). \qquad (5.2)
\end{aligned}
$$

Since $\Phi(x;\mu,\sigma)$ is continuous in μ for each x and continuous in x for each μ, the term in the squared bracket in (5.2) is also continuous in μ for each x and continuous in x for each μ. The strong consistency of $\hat{G}_{\pi_0}$ holds and

$$\left|\left|\int \Phi(x;\mu,\sigma)d\hat{H}(\mu) - \hat{F}_n\right|\right| = \left|\left|\int F(x;\hat{G},\sigma,\pi_0) - \hat{F}_n\right|\right| \to 0.$$

The strong consistency of $\hat{H}_{\pi_0}$ is thus proved. □

The existence of the restricted and the unrestricted NPMDE under the Cramér-von Mises distance is guaranteed, but the estimators may not be unique. It may not hold even in parametric models in general; see Boos (1981). If the space Γ is compact and the component CDF can be written as

$$\Phi(x;\mu,\sigma) = \Phi(x-\mu;0,\sigma),$$

then the NPMDE is unique by Lemma 2.1 in Boos (1981). The CDFs of normal distributions satisfy such conditions, and the NPMDE using normal components is thus unique.

For the computation aspect of the NPMDE, we have the following proposition characterising the NPMDE under the Cramér-von Mises distance.

Proposition 5.3. *For the restricted and the unrestricted NPMDE under the Cramér-von Mises distance, the following three statements are equivalent:*

1. *$\hat{G}$ minimises $l_n(G)$,*
2. *$\hat{G}$ maximises $\inf_\mu\{d(\mu,G)\}$,*
3. *$\inf_\mu\{d(\mu,\hat{G})\} = 0$.*

Furthermore, the support points of $\hat{G}$ are contained in the set

$$\{\mu : d(\mu,\hat{G}) = 0\}.$$

Proof. The proof is similar to Theorems 4 and 5 in Chee and Wang (2013) and the three lemmas required by these theorems. We only need to check that Lemma 1 of Chee and Wang (2013) holds. The restricted NPMDE is used in the proof and the proof of the unrestricted NPMDE is the same by setting $\pi_0 = 0$.

For any $G', G'' \in \Gamma$, we have $(1-\epsilon)G' + \epsilon G'' \in \Gamma$, $\epsilon \in [0,1]$, Γ is a convex set. Then,

$$\frac{\partial^2 l_n[(1-\epsilon)G' + \epsilon G'']}{\partial \epsilon^2}$$

$$
\begin{aligned}
&= \frac{\partial}{\partial \epsilon} \frac{\partial l_n[(1-\epsilon)G' + \epsilon G'']}{\partial \epsilon} \\
&= \frac{\partial}{\partial \epsilon} \left(\sum_{i=1}^{n} 2(F(x_{(i)}; G'', \sigma, \pi_0) - F(x_{(i)}; G, \sigma, \pi_0)) \right. \\
&\quad \left. \cdot \left[(1-\epsilon)F(x_{(i)}; G', \sigma, \pi_0) + \epsilon F(x_{(i)}; G'', \sigma, \pi_0) - \frac{2i-1}{2n} \right] \right) \\
&= \sum_{i=1}^{n} 2[F(x_{(i)}; G'', \sigma, \pi_0) - F(x_{(i)}; G, \sigma, \pi_0)]^2 \\
&\geq 0.
\end{aligned}
$$

The equivalent of Lemma 1 in Chee and Wang (2013) is thus proved, and all the lemmas and the theorems thus follow. □

5.3 The Anderson-Darling Distance

The Anderson-Darling test is also one of the common tools for testing goodness of fit between a cumulative distribution function and a given empirical distribution function. The corresponding distance measure is the weighted Cramér-von Mises distance which can also be used in estimation. As pointed out by Parr and Schucany (1980) that the Anderson-Darling distance gives more weight to observations in the tails of the distribution, which is different from the Cramér-von Mises distance that gives the same weight to all the observations. In this section, the computation of the unrestricted and the restricted NPMDE under the Anderson-Darling distance is discussed as well as some of its theoretical properties.

5.3.1 Computation of the Estimator

The loss function for the Anderson-Darling distance is

$$\ell(G) = \int \frac{[F(\cdot; G, \sigma) - \hat{F}_n]^2}{F(\cdot; G, \sigma)[1 - F(\cdot; G, \sigma)]} dF(\cdot; G, \sigma).$$

Using the transformation in Anderson and Darling (1954), the unrestricted NPMDE under this metric is the solution to the problem

$$\hat{G} = \arg\min_{G \in \Gamma} -n - \sum_{i=1}^{n} \frac{2i-1}{n} \log[F(x_{(i)}; G, \sigma)] + \frac{2n-2i+1}{n} \log[1 - F(x_{(i)}; G, \sigma)]$$

In this case we only need to consider maximising the objective function

$$l_n(G) = \sum_{i=1}^{n} \frac{2i-1}{n} \log[F(x_{(i)}; G, \sigma)] + \frac{2n-2i+1}{n} \log[1 - F(x_{(i)}; G, \sigma)].$$

The process of finding the unrestricted NPMDE under this metric is similar to the ones mentioned in Wang (2007) and Section 5.2.1. In the first step, given the mixing distribution at the current iteration G_s, the gradient function $d(\mu, G_s)$ is needed to compute all the local maxima. For the directional derivative from G' to G'', we have

$$\begin{aligned}
d(G''; G') \equiv & \left. \frac{\partial l\{(1-\epsilon)G' + \epsilon G''\}}{\partial \epsilon} \right|_{\epsilon=0} \\
= & \frac{\partial}{\partial \epsilon} \sum_{i=1}^{n} \frac{2i-1}{n} \log[(1-\epsilon)F(x_{(i)}; G', \sigma) + \epsilon F(x_{(i)}; G'', \sigma)] \\
& + \frac{2n-2i+1}{n} \log[(1-\epsilon)(1 - F(x_{(i)}; G', \sigma)) + \epsilon(1 - F(x_{(i)}; G'', \sigma))] \Big|_{\epsilon=0} \\
= & \sum_{i=1}^{n} \frac{2i-1}{n} \cdot \frac{F(x_{(i)}; G'', \sigma) - F(x_{(i)}; G', \sigma)}{F(x_{(i)}; G', \sigma)} \\
& + \sum_{i=1}^{n} \frac{2n-2i+1}{n} \cdot \frac{[1 - F(x_{(i)}; G'', \sigma)] - [1 - F(x_{(i)}; G', \sigma)]}{1 - F(x_{(i)}; G', \sigma)} \\
= & \sum_{i=1}^{n} \left[\frac{2i-1}{n} \cdot \frac{F(x_{(i)}; G'', \sigma)}{F(x_{(i)}; G', \sigma)} + \frac{2n-2i+1}{n} \cdot \frac{1 - F(x_{(i)}; G'', \sigma)}{1 - F(x_{(i)}; G', \sigma)} - 2 \right],
\end{aligned}$$

and hence the gradient function is

$$d(\mu; G_s) = \sum_{i=1}^{n} \left[\frac{2i-1}{n} \cdot \frac{\Phi(x_{(i)}; \mu, \sigma)}{F(x_{(i)}; G_s, \sigma)} + \frac{2n-2i+1}{n} \cdot \frac{1 - \Phi(x_{(i)}; \mu, \sigma)}{1 - F(x_{(i)}; G_s, \sigma)} - 2 \right]. \quad (5.3)$$

The second step is to compute the weights that maximise the objective function. Let $\boldsymbol{\eta}$ be the vector of weights at the previous iteration with zero probability on the support points added at the current iteration. Up to an additive constant, the second order Taylor expansion of the objective function around $\boldsymbol{\eta}$ with respect to $\boldsymbol{\eta}'$ is

$$\begin{aligned}
l_n(G_s) = & \boldsymbol{w}_1^T \boldsymbol{S}(\boldsymbol{\eta}' - \boldsymbol{\eta}) - \frac{1}{2}(\boldsymbol{\eta}' - \boldsymbol{\eta}) \boldsymbol{S}^T \boldsymbol{W}_1 \boldsymbol{S}(\boldsymbol{\eta}' - \boldsymbol{\eta}) \\
& + \boldsymbol{w}_2^T \boldsymbol{U}(\boldsymbol{\eta}' - \boldsymbol{\eta}) - \frac{1}{2}(\boldsymbol{\eta}' - \boldsymbol{\eta}) \boldsymbol{U}^T \boldsymbol{W}_2 \boldsymbol{U}(\boldsymbol{\eta}' - \boldsymbol{\eta}) \\
= & (\boldsymbol{w}_1^T \boldsymbol{S} + \boldsymbol{w}_2^T \boldsymbol{U})(\boldsymbol{\eta}' - \boldsymbol{\eta}) - \frac{1}{2}(\boldsymbol{\eta}' - \boldsymbol{\eta})^T (\boldsymbol{S}^T \boldsymbol{W}_1 \boldsymbol{S} + \boldsymbol{U}^T \boldsymbol{W}_2 \boldsymbol{U})(\boldsymbol{\eta}' - \boldsymbol{\eta}), \quad (5.4)
\end{aligned}$$

where

$$\begin{aligned}
\boldsymbol{w}_1 &= \left(\frac{1}{n}, \frac{3}{n}, \cdots, \frac{2n-1}{n}\right)^T, \\
\boldsymbol{W}_1 &= \text{diag}(\boldsymbol{w}_1), \\
\boldsymbol{w}_2 &= \left(\frac{2n-1}{n}, \frac{2n-3}{n}, \cdots, \frac{1}{n}\right)^T, \\
\boldsymbol{W}_2 &= \text{diag}(\boldsymbol{w}_2), \\
\boldsymbol{S}^T &= (\boldsymbol{s}_1, \cdots, \boldsymbol{s}_n), \\
\boldsymbol{s}_i &= \left(\frac{\Phi(x_{(i)}; \mu'_1, \sigma)}{F(x_{(i)}; G_{s-1}, \sigma)}, \cdots, \frac{\Phi(x_{(i)}; \mu'_{k'}, \sigma)}{F(x_{(i)}; G_{s-1}, \sigma))}\right)^T, \\
\boldsymbol{U}^T &= (\boldsymbol{u}_1, \cdots, \boldsymbol{u}_n), \\
\boldsymbol{u}_i &= \left(\frac{1-\Phi(x_{(i)}; \mu'_1, \sigma)}{1-F(x_{(i)}; G_{s-1}, \sigma)}, \cdots, \frac{1-\Phi(x_{(i)}; \mu'_{k'}, \sigma)}{1-F(x_{(i)}; G_{s-1}, \sigma)}\right)^T.
\end{aligned}$$

We also have the following equations,

$$\begin{aligned}
\boldsymbol{S}\boldsymbol{\eta} &= \mathbf{1}, \\
\boldsymbol{U}\boldsymbol{\eta} &= \mathbf{1}, \\
\mathbf{1}^T\boldsymbol{\eta}' = \mathbf{1}^T\boldsymbol{\eta} &= 1.
\end{aligned}$$

Maximising (5.4) can be turned into the optimisation problem

$$\underset{\boldsymbol{\eta}'}{\text{Minimise}} \qquad ||\boldsymbol{P}\boldsymbol{\eta}' - \boldsymbol{p}||, \tag{5.5}$$

subject to

$$\begin{aligned}
\mathbf{1}^T\boldsymbol{\eta}' &= 1, \\
\eta'_i &\geq 0, \ \forall i,
\end{aligned}$$

where $\boldsymbol{P}$ and $\boldsymbol{p}$ are the solutions to the system of equations

$$\begin{aligned}
\boldsymbol{P}^T\boldsymbol{P} &= \boldsymbol{S}^T\boldsymbol{W}_1\boldsymbol{S} + \boldsymbol{U}^T\boldsymbol{W}_2\boldsymbol{U}, \\
\boldsymbol{p}^T\boldsymbol{P} &= \boldsymbol{\eta}^T(\boldsymbol{S}^T\boldsymbol{W}_1\boldsymbol{S} + \boldsymbol{U}^T\boldsymbol{W}_2\boldsymbol{U}) + (\boldsymbol{w}_1^T\boldsymbol{S} + \boldsymbol{w}_2^T\boldsymbol{U}) \\
&= 2(\boldsymbol{w}_1^T\boldsymbol{S} + \boldsymbol{w}_2^T\boldsymbol{U}).
\end{aligned}$$

Since $\boldsymbol{S}^T\boldsymbol{W}_1\boldsymbol{S} + \boldsymbol{U}^T\boldsymbol{W}_2\boldsymbol{U}$ is positive definite under finite samples, the spectral or the Cholesky decomposition can be used to find $\boldsymbol{P}$ and $\boldsymbol{p}$ efficiently. Alternatively, (5.4) can be solved directly using the function `pnnqp` in the R package `lsei` (Wang et al., 2020) with the same constraints in (5.5).

Finally, similar to the restricted and the unrestricted NPMLE, the Anderson-Darling distance is not exactly quadratic in $\boldsymbol{\eta}'$ while a quadratic expansion is used to compute the weights. To ensure the monotone increase and global convergence, a line search is needed, the inequality

$$\ell_n(G_s^{c,l}) \geq \ell_n(G_s^{0,l}) + \alpha c^l(\boldsymbol{w}_1^T\boldsymbol{S} + \boldsymbol{w}_2^T\boldsymbol{U})(\boldsymbol{\eta}' - \boldsymbol{\eta}) \tag{5.6}$$

is tested in the order $l = 0, 1, 2, \cdots$ until it is first satisfied. For computing the restricted NPMDE under the Anderson-Darling distance, the optimisation becomes

$$\hat{G} = \arg\max_{G\in\Gamma} \sum_{i=1}^{n} \left\{ \frac{2i-1}{n} \log[F(x_{(i)}; G, \sigma, \pi_0)] + \frac{2n-2i+1}{n} \log[1 - F(x_{(i)}; G, \sigma, \pi_0)] \right\},$$

and hence given the mixing distribution at the current iteration G_s, the gradient function is

$$\begin{aligned} d(\mu; G_s) &= \sum_{i=1}^{n} \left[\frac{2i-1}{n} \cdot \frac{F(x_{(i)}; \mu, \sigma, \pi_0)}{F(x_{(i)}; G', \sigma, \pi_0)} + \frac{2n-2i+1}{n} \cdot \frac{1 - F(x_{(i)}; \mu, \sigma, \pi_0)}{1 - F(x_{(i)}; G', \sigma, \pi_0)} - 2 \right] \\ &= \sum_{i=1}^{n} \left[\frac{2i-1}{n} \cdot \frac{(1-\pi_0)\Phi(x_{(i)}; \mu, \sigma)}{F(x_{(i)}; G', \sigma, \pi_0)} - \frac{2n-2i+1}{n} \cdot \frac{(1-\pi_0)\Phi(x_{(i)}; \mu, \sigma)}{1 - F(x_{(i)}; G', \sigma, \pi_0)} \right] \\ &\quad + \sum_{i=1}^{n} \left[\frac{2i-1}{n} \cdot \frac{1 - \pi_0\Phi(x_{(i)}; 0, \sigma)}{F(x_{(i)}; G', \sigma, \pi_0)} + \frac{2n-2i+1}{n} \cdot \frac{1 - \pi_0\Phi(x_{(i)}; 0, \sigma)}{1 - F(x_{(i)}; G', \sigma, \pi_0)} \right] - 2n. \end{aligned}$$

Example 5.2 (Normal distribution)**.** *If ϕ is the normal density and Φ is the corresponding cumulative distribution function, using the property*

$$\frac{\partial}{\partial\mu}\Phi(x; \mu, \sigma) = -\phi(x; \mu, \sigma),$$

we have for the unrestricted NPMDE,

$$\frac{\partial d(\mu, G_s)}{\partial\mu} = -\sum_{i=1}^{n} \phi(x_{(i)}; \mu, \sigma) \left\{ \frac{2i-1}{nF(x_{(i)}; G', \sigma)} - \frac{2n-2i+1}{n[1 - F(x_{(i)}; G', \sigma)]} \right\},$$

$$\frac{\partial^2 d(\mu, G_s)}{\partial \mu^2} = -\sum_{i=1}^{n} \frac{x_{(i)} - \mu}{\sigma^2} \phi(x_{(i)}; \mu, \sigma) \left\{ \frac{2i-1}{nF(x_{(i)}; G', \sigma)} - \frac{2n-2i+1}{n[1 - F(x_{(i)}; G', \sigma)]} \right\},$$

and

$$\frac{\partial d(\mu, G_s)}{\partial \mu} = -(1-\pi_0) \sum_{i=1}^{n} \phi(x_{(i)}; \mu, \sigma) \left\{ \frac{2i-1}{nF(x_{(i)}; G', \sigma, \pi_0)} - \frac{2n-2i+1}{n[1 - F(x_{(i)}; G', \sigma, \pi_0)]} \right\},$$

$$\frac{\partial^2 d(\mu, G_s)}{\partial \mu^2} = -(1-\pi_0) \sum_{i=1}^{n} \frac{x_{(i)} - \mu}{\sigma^2} \phi(x_{(i)}; \mu, \sigma) \left\{ \frac{2i-1}{nF(x_{(i)}; G', \sigma, \pi_0)} - \frac{2n-2i+1}{n[1 - F(x_{(i)}; G', \sigma, \pi_0)]} \right\}$$

for the restricted NPMDE under the Anderson-Darling distance.

The next step is to compute the weights that maximise the object function. Up to an additive constant, the second order Taylor expansion of the objective function around $\boldsymbol{\eta}$ with respect to $\boldsymbol{\eta}'$ is

$$\begin{aligned} l_n(G_s) &= \boldsymbol{w}_1^T \boldsymbol{S}(\boldsymbol{\eta}' - \boldsymbol{\eta}) - \frac{1}{2}(\boldsymbol{\eta}' - \boldsymbol{\eta}) \boldsymbol{S}^T \boldsymbol{W}_1 \boldsymbol{S}(\boldsymbol{\eta}' - \boldsymbol{\eta}) \\ &\quad + \boldsymbol{w}_2^T \boldsymbol{U}(\boldsymbol{\eta}' - \boldsymbol{\eta}) - \frac{1}{2}(\boldsymbol{\eta}' - \boldsymbol{\eta}) \boldsymbol{U}^T \boldsymbol{W}_2 \boldsymbol{U}(\boldsymbol{\eta}' - \boldsymbol{\eta}) \\ &= (\boldsymbol{w}_1^T \boldsymbol{S} + \boldsymbol{w}_2^T \boldsymbol{U})(\boldsymbol{\eta}' - \boldsymbol{\eta}) - \frac{1}{2}(\boldsymbol{\eta}' - \boldsymbol{\eta})^T (\boldsymbol{S}^T \boldsymbol{W}_1 \boldsymbol{S} + \boldsymbol{U}^T \boldsymbol{W}_2 \boldsymbol{U})(\boldsymbol{\eta}' - \boldsymbol{\eta}), \quad (5.7) \end{aligned}$$

where

$$\begin{aligned} \boldsymbol{S}^T &= (\boldsymbol{s}_1, \cdots, \boldsymbol{s}_n), \\ \boldsymbol{s}_i &= \left(\frac{\Phi(x_{(i)}; \mu_1', \sigma)}{F(x_{(i)}; G_{s-1}, \sigma, \pi_0)}, \cdots, \frac{\Phi(x_{(i)}; \mu_{k'}', \sigma)}{F(x_{(i)}; G_{s-1}, \sigma, \pi_0)} \right)^T, \\ \boldsymbol{U}^T &= (\boldsymbol{u}_1, \cdots, \boldsymbol{u}_n), \\ \boldsymbol{u}_i &= \left(\frac{-\Phi(x_{(i)}; \mu_1', \sigma)}{1 - F(x_{(i)}; G_{s-1}, \sigma, \pi_0)}, \cdots, \frac{-\Phi(x_{(i)}; \mu_{k'}', \sigma)}{1 - F(x_{(i)}; G_{s-1}, \sigma, \pi_0)} \right)^T. \end{aligned}$$

Let

$$\begin{aligned} \boldsymbol{s}_{\pi_0} &= \left(\frac{\pi_0 \Phi(x_{(1)}, 0, \sigma)}{F(x_{(1)}; G_{s-1}, \sigma, \pi_0)}, \cdots, \frac{\pi_0 \Phi(x_{(n)}, 0, \sigma)}{F(x_{(n)}; G_{s-1}, \sigma, \pi_0)} \right)^T, \\ \boldsymbol{u}_{\pi_0} &= \left(\frac{1 - \pi_0 \Phi(x_{(1)}, 0, \sigma)}{1 - F(x_{(1)}; G_{s-1}, \sigma, \pi_0)}, \cdots, \frac{1 - \pi_0 \Phi(x_{(n)}, 0, \sigma)}{1 - F(x_{(n)}; G_{s-1}, \sigma, \pi_0)} \right)^T, \end{aligned}$$

we also have the following equations,

$$\begin{aligned}\boldsymbol{S}\boldsymbol{\eta} + \boldsymbol{s}_{\pi_0} &= \mathbf{1},\\ \boldsymbol{U}\boldsymbol{\eta} + \boldsymbol{u}_{\pi_0} &= \mathbf{1},\\ \mathbf{1}^T\boldsymbol{\eta}' = \mathbf{1}^T\boldsymbol{\eta} &= 1 - \mathbf{1}^T\boldsymbol{\pi}.\end{aligned}$$

the optimisation problem (5.5) can thus be used to solve $\boldsymbol{\eta}'$ with the updated $\boldsymbol{S}$ and $\boldsymbol{U}$, and $\boldsymbol{P}$ and $\boldsymbol{p}$ are the solutions to the system of equations

$$\begin{aligned}\boldsymbol{P}^T\boldsymbol{P} &= \boldsymbol{S}^T\boldsymbol{W}_1\boldsymbol{S} + \boldsymbol{U}^T\boldsymbol{W}_2\boldsymbol{U},\\ \boldsymbol{p}^T\boldsymbol{P} &= \boldsymbol{\eta}^T(\boldsymbol{S}^T\boldsymbol{W}_1\boldsymbol{S} + \boldsymbol{U}^T\boldsymbol{W}_2\boldsymbol{U}) + (\boldsymbol{w}_1^T\boldsymbol{S} + \boldsymbol{w}_2^T\boldsymbol{U})\\ &= (1 - \boldsymbol{s}_{\pi_0})^T\boldsymbol{W}_1\boldsymbol{S} + (1 - \boldsymbol{u}_{\pi_0})^T\boldsymbol{W}_2\boldsymbol{U} + (\boldsymbol{w}_1^T\boldsymbol{S} + \boldsymbol{w}_2^T\boldsymbol{U})\\ &= 2(\boldsymbol{w}_1^T\boldsymbol{S} + \boldsymbol{w}_2^T\boldsymbol{U}) - \boldsymbol{s}_{\pi_0}^T\boldsymbol{W}_1\boldsymbol{S} - \boldsymbol{u}_{\pi_0}^T\boldsymbol{W}_2\boldsymbol{U},\end{aligned}$$

subject to the constraints

$$\begin{aligned}\mathbf{1}^T\boldsymbol{\eta}' &= 1 - \mathbf{1}^T\boldsymbol{\pi},\\ \eta_i' &\geq 0,\ \forall i.\end{aligned}$$

Similar to the process of computing the unrestricted NPMDE, a line search is needed for the monotonic increase of the objective function and global convergence, the inequality to test is the same as (5.6).

5.3.2 Theoretical Properties

The consistency of the unrestricted NPMDE under the Anderson-Darling distance was shown in Boos (1982). It was mentioned that if Γ is a compact set and $\Phi(\cdot;\mu,\sigma)$ is continuous in μ, then the existence of the unrestricted NPMDE is guaranteed. There may be more than one NPMDE but all of them converge to the true mixing distribution, almost surely. Similar to the Cramér-von Mises distance, the consistency of the restricted NPMDE can be shown by the nestedness of the restricted space or changing the component distribution function to a mixture of distribution functions. The following proposition shows the latter one.

Proposition 5.4. *Under the conditions specified in Boos (1982) and* $H^* \in \Gamma_{\pi_0}$*, the estimator*

$$\hat{H}_{\pi_0} = \pi_0 \mathbb{1}_0 + (1-\pi_0)\hat{G}_{\pi_0},$$

where

$$\hat{G}_{\pi_0} = \arg\max_{G\in\Gamma} \sum_{i=1}^{n} \left\{ \frac{2i-1}{n} \log[F(x_{(i)}; G, \sigma, \pi_0)] + \frac{2n-2i+1}{n} \log[1 - F(x_{(i)}; G, \sigma, \pi_0)] \right\},$$

is strongly consistent.

Proof. Since H^* is an element in Γ_{π_0}, it can be written as

$$H^* = \pi_0 \mathbb{1}_0 + (1-\pi_0)G^*.$$

We can rearrange $F(x; G, \sigma, \pi_0)$ as

$$\begin{aligned} F(x; G, \sigma, \pi_0) &= \pi_0 \Phi(x; 0, \sigma) + (1-\pi_0) \int \Phi(x; \mu, \sigma) dG(\mu) \\ &= \int \left[\pi_0 \Phi(x; 0, \sigma) + (1-\pi_0)\Phi(x; \mu, \sigma)\right] dG(\mu). \end{aligned} \tag{5.8}$$

Since $\Phi(x; \mu, \sigma)$ is continuous in μ for each x the term in the squared bracket in (5.8) is also continuous in μ for each x. The strong consistency of $\hat{G}_{\pi_0}$ holds and

$$\begin{aligned} &\lim_{n\to\infty} -n - \sum_{i=1}^{n} \left\{ \frac{2i-1}{n} \log[F(x_{(i)}; \hat{H}_{\pi_0}, \sigma)] + \frac{2n-2i+1}{n} \log[1 - F(x_{(i)}; \hat{H}_{\pi_0}, \sigma)] \right\}, \\ &= \lim_{n\to\infty} -n - \sum_{i=1}^{n} \left\{ \frac{2i-1}{n} \log[F(x_{(i)}; \hat{G}, \sigma, \pi_0)] + \frac{2n-2i+1}{n} \log[1 - F(x_{(i)}; \hat{G}, \sigma, \pi_0)] \right\} \\ &= 0. \end{aligned}$$

The strong consistency of $\hat{H}_{\pi_0}$ is thus proved. □

The next proposition concerns the characterisation of the restricted and the unrestricted NPMDE under the Anderson-Darling distance.

Proposition 5.5. *For the restricted and the unrestricted NPMDE under the Anderson-Darling distance, the following three statements are equivalent:*

1. *$\hat{G}$ maximises $l_n(G)$,*

2. *$\hat{G}$ minimises* $\sup_\mu\{d(\mu, G)\}$,

3. $\sup_\mu\{d(\mu, \hat{G})\} = 0$.

Furthermore, the support points of $\hat{G}$ are contained in the set

$$\{\mu : d(\mu, \hat{G}) = 0\}.$$

Proof. The proof is similar to Theorems 4 and 5 in Chee and Wang (2013) and the three lemmas required by the theorems. We only need to check that Lemma 1 of Chee and Wang (2013) holds. The restricted NPMDE is used in the proof and the proof of the unrestricted NPMDE is the same by setting $\pi_0 = 0$.

For any $G', G'' \in \Gamma$, we have $(1-\epsilon)G' + \epsilon G'' \in \Gamma$, $\epsilon \in [0, 1]$, Γ is a convex set. Then,

$$\begin{aligned}
&\frac{\partial^2 l_n[(1-\epsilon)G' + \epsilon G'']}{\partial \epsilon^2} \\
=& \frac{\partial}{\partial \epsilon}\frac{\partial l_n[(1-\epsilon)G' + \epsilon G'']}{\partial \epsilon} \\
=& \frac{\partial}{\partial \epsilon}\sum_{i=1}^{n}\frac{2i-1}{n}\cdot\frac{F(x_{(i)}; G'', \sigma, \pi_0) - F(x_{(i)}; G', \sigma, \pi_0)}{(1-\epsilon)F(x_{(i)}; G', \sigma, \pi_0) + \epsilon F(x_{(i)}; G'', \sigma, \pi_0)} \\
&+ \frac{\partial}{\partial \epsilon}\sum_{i=1}^{n}\frac{2n-2i+1}{n}\cdot\frac{[1 - F(x_{(i)}; G'', \sigma, \pi_0)] - [1 - F(x_{(i)}; G', \sigma, \pi_0)]}{(1-\epsilon)[1 - F(x_{(i)}; G', \sigma, \pi_0)] + \epsilon[1 - F(x_{(i)}; G'', \sigma, \pi_0)]} \\
=& -\sum_{i=1}^{n}\frac{2i-1}{n}\left[\frac{F(x_{(i)}; G'', \sigma, \pi_0) - F(x_{(i)}; G', \sigma, \pi_0)}{F(x_{(i)}; G', \sigma, \pi_0)}\right]^2 \\
&- \sum_{i=1}^{n}\frac{2i-1}{n}\left\{\frac{[1 - F(x_{(i)}; G'', \sigma, \pi_0)] - [1 - F(x_{(i)}; G', \sigma, \pi_0)]}{1 - F(x_{(i)}; G', \sigma, \pi_0)}\right\}^2 \\
\leq& 0
\end{aligned}$$

The equivalent of Lemma 1 in Chee and Wang (2013) is proved and hence all the lemmas and the theorems follow. □

5.4 Estimation of the Proportion of Zeros

In previous sections, estimates of a mixing distribution can be found by fixing the null proportion under the Cramér-von Mises and the Anderson-Darling distance. This section concerns estimating the null proportions under these distance measures.

5.4.1 The Cramér-von Mises Distance

If X_i are independently and identically distributed with the CDF $\int \Phi(x;\mu,\sigma)dH^*(\mu)$, then under correct specification of the mixing distribution,

$$Y_i = \int \Phi(X_i;\mu,\sigma)dH^*(\mu)$$

follows the uniform distribution $\mathcal{U}(0,1)$ and

$$T_{CvM} = \frac{1}{12n} + \sum_{i=1}^{n}\left[F(X_{(i)};H^*,\sigma) - \frac{2i-1}{2n}\right]^2$$

follows the Cramér-von Mises distribution. The asymptotic CDF of the Cramér-von Mises distribution is given in Anderson and Darling (1952),

$$\mathbb{P}(T_{CvM} \le t) = \frac{1}{\pi\sqrt{t}}\sum_{j=0}^{\infty}(-1)^j\binom{-\frac{1}{2}}{j}(4j+1)^{\frac{1}{2}}e^{-(4j+1)^2/(16t)}K_{\frac{1}{4}}\left(\frac{(4j+1)^2}{16t}\right),$$

where $K_{1/4}(\cdot)$ is the modified Bessel function of the second kind with parameter $\frac{1}{4}$. Csorgo and Faraway (1996) gave the asymptotically precise one-term correction to the asymptotic distribution to approximate the exact distribution under finite samples. In practice, H^* (or G^*) is unknown and we only have a class of non-parametric density estimates indexed by π_0. Define the profile loss function under the Cramér-von Mises distance with respect to π_0 as

$$\tilde{l}_n(\pi_0) = \inf_{G\in\Gamma}\sum_{i=1}^{n}\left[F(X_{(i)};G,\sigma,\pi_0) - \frac{2i-1}{2n}\right]^2 = l_n(\hat{H}_{\pi_0}),$$

and the test statistic can be written as

$$T(\pi_0) = \frac{1}{12n} + \sum_{i=1}^{n}\left[F(X_{(i)};\hat{G}_{\pi_0},\sigma,\pi_0) - \frac{2i-1}{2n}\right]^2.$$

The strong consistency of the restricted NPMDE can be shown for all $\pi_0 \le \pi_0^*$ but the test statistic $T(\pi_0)$ does not follow the Cramér-von Mises distribution even for $\pi_0 = \pi_0^*$ due to the presence of the nuisance parameter. However, given any set of observations, we have the following inequality,

$$t(\pi_1) \le t(\pi_2) \le t(\pi_0^*) \le t_{CvM},$$

for all $0 \leq \pi_1 \leq \pi_2 \leq \pi_0^*$ where

$$t(\pi_0) = \frac{1}{12n} + \sum_{i=1}^{n} \left[F(x_{(i)}; \hat{H}_{\pi_0}, \sigma) - \frac{2i-1}{2n} \right]^2,$$

and

$$\mathbb{P}(T(\pi_1) < t) \geq \mathbb{P}(T(\pi_2) < t) \geq \mathbb{P}(T(\pi_0^*) < t) \geq \mathbb{P}(T_{CvM} < t),$$

for any fixed t. In addition, Massey (1950) and Darling (1957) established that the power of the Cramér-von Mises test approaches 1 as the number of observations tends to infinity, which means the test will asymptotically reject any NPMDE $\hat{H}_{\pi_0}$ that is not consistent. The test statistic under an inconsistent model goes to infinity at a rate proportional to the number of observations, and hence $\mathbb{P}(T(\pi_0) > t) = 1$ asymptotically for all $\pi_0 > \pi_0^*$.

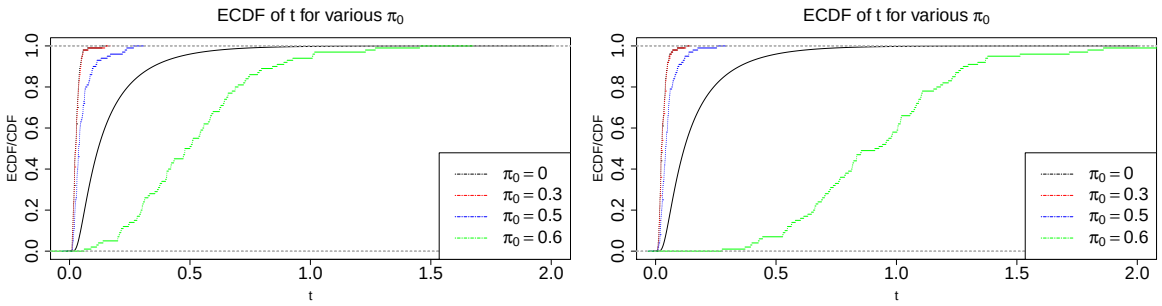

FIGURE 5.1: Plots of ECDFs of 100 Cramér-von Mises statistics for various π_0.

Figure 5.1 shows the ECDFs of 100 Cramér-von Mises statistics for $\pi_0 = 0, 0.3, 0.5$ and 0.6. 500 and 1000 observations are used to estimate the mixing distribution and the test statistics in left and right pane respectively. The step functions are the ECDFs while the smooth black line is the asymptotic CDF of the Cramér-von Mises distribution. The ECDFs to the left of the asymptotic Cramér-von Mises CDF are similar in both figures while the ECDF to the right seems to diverge to infinity as the number of observations increases. The asymptotic CDF can thus be regarded as a boundary between the consistent and the inconsistent NPMDE. As a result, the estimator of the null proportion $\hat{\pi}_0$ is defined to be the root of the expression

$$t(\hat{\pi}_0) = q(n, \alpha). \tag{5.9}$$

By the nestedness of $\Gamma_{\pi'}$ in $\Gamma_{\pi''}$ for all $\pi' \geq \pi''$, $\tilde{l}_n(\pi') \leq \tilde{l}_n(\pi'')$, and thus the test statistic is non-decreasing with respect to π_0. Furthermore, the profile loss function is differentiable, and the derivative is

$$\begin{aligned}\frac{\partial \tilde{l}_n(\pi_0)}{\partial \pi_0} &= \left.\frac{\partial \tilde{l}_n(H)}{\partial \pi_0}\right|_{H=\hat{H}_{\pi_0}} \\ &= \sum_{i=1}^{n} (\Phi(x_{(i)}; 0, \sigma) - F(x_{(i)}; \hat{H}_{\pi_0}, \sigma)) \left[F(x_{(i)}, \hat{H}_{\pi_0}, \sigma) - \frac{2i-1}{2n}\right],\end{aligned}$$

which is the gradient function (5.1) given $\hat{H}_\pi$ evaluated at $\mu = 0$. Calculating $\hat{\pi}_0$ is equivalent to finding the root of (5.9) given α. An efficient root-finding algorithm which utilises the derivative functions can be used, e.g. the secant, the Newton-Raphson or even the Halley method. A good initial value to start a root-finding algorithm, if it requires one, can be

$$\pi_{initial} = \frac{\phi(0; 0, \sigma)}{\int \phi(0; \mu, \sigma) d\hat{H}(\mu)},$$

where $\hat{H}$ is the unrestricted NPMDE under the Cramér-von Mises distance. If allowed, the restricted NPMLE computed from previous iteration can be used as a starting mixing distribution for the current iteration with no or minor modification. The performance is expected to be improved since the restricted NPMLEs are similar in a small neighbourhood of the current value π_0.

5.4.2 The Anderson-Darling Distance

Similar to the Cramér-von Mises distance, under the correct specification of the mixing distribution, the Anderson-Darling test statistic can be written as

$$T_{AD} = -n - \sum_{i=1}^{n} \frac{2i-1}{n} \cdot \log[F(X_{(i)}; H^*, \sigma)] + \frac{2n-2i+1}{n} \cdot \log[1 - F(X_{(i)}; H^*, \sigma)],$$

and the asymptotic distribution function is also given in Anderson and Darling (1952),

$$\mathbb{P}(T \leq t) = \frac{\sqrt{2\pi}}{t} \sum_{j=0}^{\infty} \binom{-\frac{1}{2}}{j} (4j+1)^{-\frac{(4j+1)^2\pi^2}{8t}} \int_{-\infty}^{\infty} e^{\frac{t}{8(w^2+1)} - \frac{(4j+1)^2\pi^2 w^2}{8t}} dw.$$

When computing the above distribution function to unlimited accuracy, the integrals need to be evaluated in a loop within a loop, which might be more complicated and time consuming than one wants to invoke. Lewis (1961) only gave tables for small n based on the simulations. Marsaglia and Marsaglia (2004) extended this idea and provided a correction term to the asymptotic distribution in order to approximate the distribution function under finite samples. Furthermore, a method for a quick approximation of the distribution function with accuracy limited to the computer's machine error and a piecewise function for approximation were also presented. In practice, H^* (or G^*) is unknown and we only have a class of estimated densities indexed by π_0. Define the profile loss function under the Anderson-Darling distance with respect to π_0 as

$$\begin{aligned}\tilde{l}_n(\pi_0) &= \sup_{G\in\Gamma} \sum_{i=1}^{n} \frac{2i-1}{n} \cdot \log[F(x_{(i)}; G, \sigma, \pi_0)] + \frac{2n-2i+1}{n} \cdot \log[1 - F(x_{(i)}; G, \sigma, \pi_0)] \\ &= \sum_{i=1}^{n} \frac{2i-1}{n} \cdot \log[F(x_{(i)}; \hat{G}_{\pi_0}, \sigma, \pi_0)] + \frac{2n-2i+1}{n} \cdot \log[1 - F(x_{(i)}; \hat{G}_{\pi_0}, \sigma, \pi_0)],\end{aligned}$$

and the test statistic can be written as

$$T(\pi_0) = -n - \sum_{i=1}^{n} \frac{2i-1}{n} \cdot \log[F(X_{(i)}; \hat{G}_{\pi_0}, \sigma, \pi_0)] + \frac{2n-2i+1}{n} \cdot \log[1 - F(X_{(i)}; \hat{G}_{\pi_0}, \sigma, \pi_0)].$$

The Anderson-Darling distance is essentially a weighted Cramér-von Mises distance, and hence derivation of the estimator of the null proportion is similar to that in Section 5.4.1. Massey (1950) and Darling (1957) also established that the power of the Anderson-Darling test approaches 1 as the number of observations tends to infinity, which means the test will asymptotically reject any NPMDE $\hat{H}_{\pi_0}$ that is not consistent.

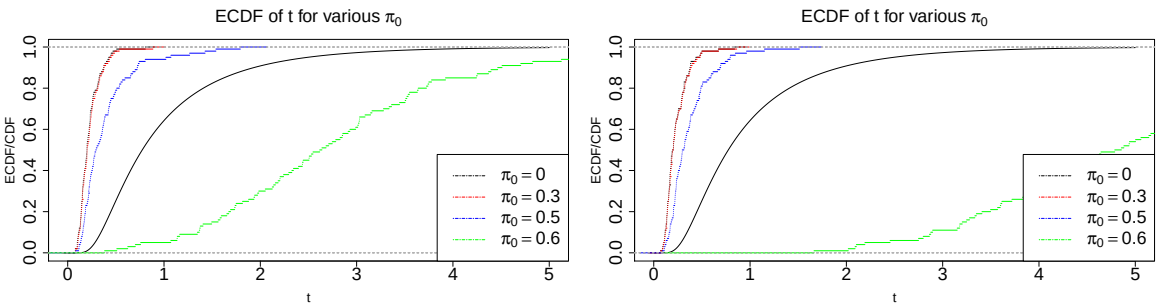

FIGURE 5.2: Plots of ECDFs of 100 Anderson-Darling statistics for various π_0.

Figure 5.2 shows the ECDFs of 100 Anderson-Darling statistics for $\pi_0 = 0, 0.3, 0.5$ and 0.6. The observations used are the same as those in Figure 5.1. The step functions are the ECDFs while the smooth black line is the asymptotic CDF of the Anderson-Darling distribution. The estimator of the null proportion $\hat{\pi}_0$ is thus defined to be the root of the expression

$$t(\hat{\pi}_0) = q(n, \alpha). \tag{5.10}$$

The derivative of the profile loss function, i.e.,

$$\begin{aligned}\frac{\partial \tilde{l}_n(\pi_0)}{\partial \pi_0} &= \left.\frac{\partial \tilde{l}_n(H)}{\partial \pi_0}\right|_{H=\hat{H}_{\pi_0}} \\ &= \sum_{i=1}^{n}\left[\frac{2i-1}{n}\cdot\frac{\Phi(x_{(i)};0,\sigma)}{F(x_{(i)};\hat{H}_{\pi_0},\sigma)} + \frac{2n-2i+1}{n}\cdot\frac{1-\Phi(x_{(i)};0,\sigma)}{1-F(x_{(i)};\hat{H}_{\pi_0},\sigma)} - 2\right],\end{aligned}$$

can be used to find the root.

5.4.3 Choice of the Quantile

In previous subsections, the proportion of null effects can be found with a fixed α-value under various distances. The exact distributions of the Cramér-von Mises and Anderson-Darling statistics for the NPMDEs are unknown, and the asymptotic CDFs can only be regarded as a boundary between the consistent and the inconsistent NPMDEs. The support spaces of the asymptotic Cramér-von Mises and Anderson-Darling distributions are positive real numbers, and hence the null proportion estimator is consistent if $q(n, \alpha)$ in (5.9) approaches infinity at a certain rate. The rate remains unknown since we do not have results under the Cramér-von Mises and Anderson-Darling distances equivalent to that using maximum likelihood by Jiang and Zhang (2016). In practice, we want to use a thresholding function that focuses on performance. The value $\alpha = 0.5$ can be regarded as a distance-based analogue to the AIC thresholding function for the likelihood-based method, and this value will be used in the numerical studies in this Chapter.

5.5 Comparison with the NPMLE

In this section, numerical studies are conducted in order to compare the performance of distance-based methods with the likelihood-based method when the component density is assumed to be normal. The performance is investigated through the computation time, the quality of the estimated mixing distributions measured by the metric in Kiefer and Wolfowitz (1956) and the estimation of the null proportion. The synthetic datasets in Section 4.2 are used. The maximisation problem under the Anderson-Darling distance is converted into a minimisation problem. When finding the negative local minima at each iteration, the support space has been partitioned into smaller intervals and local minima are found by computing the root of the first derivative of the gradient function.

n	method	$\pi_0 = 0$	$\pi_0 = 0.3$	$\pi_0 = 0.6$	$\pi_0 = 0.9$
1500	NPMLE	0.085(0.002)	0.085(0.003)	0.070(0.002)	0.059(0.001)
1500	NPMDE-CvM	0.048(0.001)	0.046(0.001)	0.042(0.001)	0.037(0.001)
1500	NPMDE-AD	0.108(0.003)	0.106(0.002)	0.094(0.004)	0.055(0.002)
3000	NPMLE	0.191(0.005)	0.202(0.009)	0.162(0.006)	0.142(0.003)
3000	NPMDE-CvM	0.101(0.002)	0.096(0.002)	0.094(0.003)	0.082(0.002)
3000	NPMDE-AD	0.244(0.006)	0.228(0.005)	0.179(0.007)	0.114(0.004)
6000	NPMLE	0.400(0.013)	0.389(0.016)	0.308(0.014)	0.254(0.006)
6000	NPMDE-CvM	0.178(0.004)	0.170(0.003)	0.158(0.003)	0.143(0.004)
6000	NPMDE-AD	0.476(0.013)	0.448(0.012)	0.334(0.011)	0.241(0.010)

TABLE 5.1: The mean (standard error) of computation times (in seconds) using various methods for each combination of (n, π_0) from the 100 synthetic datasets generated by the prostate data.

Table 5.1 shows the means and standard errors of the computation time of various methods using synthetic dataset generated from the prostate data when $n = 1500, 3000, 6000$ and $\pi_0 = 0, 0.3, 0.6, 0.9$. The terms "NPMDE-CvM" and "NPMDE-AD" represent the results computed under the Cramér-von Mises and the Anderson-Darling distance respectively. Overall, the computation times for all the methods roughly increase in the same order as the increase in the number of observations. The NPMDE under the Cramér-von Mises distance is the fastest since the distance function is exactly quadratic, the weight vector obtained from the NNLS algorithm is already the global minimum given the current set of the support points, and hence there is no need for a line search at all. The computation times for the NPMLE are slightly

slower than the NPMDE under the Anderson-Darling distance. Across all the methods and numbers of observations, the computing time decreases as π_0 increases, since potentially less support points around 0 are added to the support set at each iteration.

The distance to the true distribution is also investigated. Note that all the methods have their own embedded distance metric. Using any metric in any of the methods will penalise other methods and the results using any metric different from those three are only informative. The metric in Kiefer and Wolfowitz (1956) is used as a measure of performance, since this metric is a distance measure between two mixing distributions, while the Cramér-von Mises, Anderson-Darling and KL distances are distance measures between two mixture distributions.

n	method	$\pi_0 = 0$	$\pi_0 = 0.3$	$\pi_0 = 0.6$	$\pi_0 = 0.9$
1500	NPMLE	**0.109**(0.007)	**0.098**(0.006)	**0.079**(0.003)	**0.029**(0.001)
1500	NPMDE-CvM	0.140(0.009)	0.124(0.007)	0.097(0.004)	0.035(0.001)
1500	NPMDE-AD	0.130(0.008)	0.116(0.006)	0.092(0.004)	0.032(0.001)
3000	NPMLE	**0.094**(0.007)	**0.085**(0.005)	**0.070**(0.003)	**0.025**(0.001)
3000	NPMDE-CvM	0.134(0.009)	0.117(0.007)	0.090(0.004)	0.030(0.001)
3000	NPMDE-AD	0.119(0.008)	0.105(0.007)	0.083(0.004)	0.028(0.001)
6000	NPMLE	**0.096**(0.006)	**0.086**(0.005)	**0.069**(0.003)	**0.024**(0.001)
6000	NPMDE-CvM	0.119(0.008)	0.107(0.006)	0.085(0.004)	0.026(0.001)
6000	NPMDE-AD	0.113(0.008)	0.101(0.006)	0.080(0.004)	0.025(0.001)

TABLE 5.2: The mean (standard error) of the distances to true mixing distribution using various methods for each combination of (n, π_0) from the 100 synthetic datasets generated by the prostate data.

Table 5.2 shows the means and standard errors of the distances to the true mixing distribution for various methods using synthetic dataset generated from the prostate data when $n = 1500, 3000, 6000$ and $\pi_0 = 0, 0.3, 0.6, 0.9$. The distance between the estimated mixing distribution and the true one decreases when π_0 increases toward the true value. The distances are decreasing for all the methods across all π_0, this is guarded by the consistency of the restricted NPMLE and NPMDE when $\pi_0 \leq \pi_0^*$. The NPMLE has the closest distance to the true mixing distribution uniformly followed by the NPMDE under the Anderson-Darling distance, while the NPMDE under the Cramér-von Mises distance seems to have the largest distance among all. This numerical study suggests that the NPMLE seems to be the most efficient. The NPMDE under the Cramér-von Mises distance is the fastest but it has the worst accuracy among all while the Anderson-Darling distance takes slightly shorter time in computation than that of the NPMLE but the result is not as accurate as that produced by the NPMLE.

The performance of estimating the null proportion in the context of the MCPs is also attempted. For the following simulations, we use $\alpha = 0.5$ to compute the proportion of null effects under the Cramér-von Mises and the Anderson-Darling distance. The results in Table 4.1 are also shown here for comparison.

π_0^*	n	CvM	AD	AIC	BIC
0.5	1500	0.746(0.002)	0.743(0.002)	0.675(0.005)	0.732(0.003)
	3000	0.723(0.002)	0.723(0.002)	**0.653**(0.006)	0.712(0.003)
	6000	0.707(0.002)	0.707(0.002)	**0.636**(0.006)	0.695(0.002)
0.7	1500	0.865(0.002)	0.861(0.002)	0.813(0.004)	0.860(0.002)
	3000	0.851(0.002)	0.848(0.002)	0.806(0.004)	0.847(0.002)
	6000	0.840(0.002)	0.837(0.001)	0.795(0.004)	0.836(0.002)
0.9	1500	0.964(0.010)	0.969(0.002)	0.946(0.003)	0.973(0.001)
	3000	0.967(0.001)	0.963(0.001)	0.942(0.002)	0.966(0.001)
	6000	0.951(0.010)	0.957(0.001)	0.940(0.002)	0.961(0.001)
π_0^*	n	locfdr	ECF	mixfdr	
0.5	1500	0.739(0.001)	**0.671**(0.003)	0.836(0.001)	
	3000	0.739(0.001)	0.660(0.003)	0.835(0.001)	
	6000	0.739(0.001)	0.651(0.002)	0.835(0.001)	
0.7	1500	0.839(0.002)	**0.800**(0.003)	0.913(0.001)	
	3000	0.839(0.001)	**0.796**(0.003)	0.912(0.001)	
	6000	0.839(0.001)	**0.792**(0.002)	0.911(0.001)	
0.9	1500	0.947(0.002)	**0.932**(0.003)	0.976(0.000)	
	3000	0.947(0.002)	**0.931**(0.002)	0.973(0.000)	
	6000	0.947(0.001)	**0.929**(0.002)	0.972(0.000)	

TABLE 5.3: The mean (standard error) of estimated null proportions using various methods and n from the 100 synthetic datasets generated by the prostate data.

Table 5.3 shows the means and standard errors of the estimated null proportions for various methods and $n = 1500, 3000, 6000$. Since all the methods use normal components, we drop the prefix "npnorm" for the purpose of presentation. The terms "CvM" and "AD" correspond to the median estimators of the distance-based methods under the Cramér-von Mises and the Anderson-Darling distance respectively. Overall, the proposed method with the AIC threshold function and the ECF method still perform better than other methods. The median estimators for the distance-based methods perform reasonably well. The estimator under the Anderson-Darling distance performs slightly better than that under the Cramér-von Mises distance, which is consistent with

the results shown in Table 5.2. The median estimators for the distance-based methods also show significant decreasing trends in general as the number of observations increases.

5.6 Conclusion

In this chapter, two non-parametric distance-based methods for estimating a mixing distribution are presented in complement to the likelihood-based method. The conditions for the consistency of the NPMDEs are generally less strict than the NPMLE and can be applied to a wider range of problems. However, distance-based methods are generally not as efficient as the likelihood-based method, as shown in Table 5.2. Computing the NPMDE under the Cramér-von Mises distance is much faster but the estimator is not accurate, while under the Anderson-Darling distance the estimator performs slightly better than the Cramér-von Mises distance, but it takes a much longer time in computation. It is still recommended to use the likelihood-based method where applicable.

It is important to notice that the asymptotic distributions of the distance-based test statistics are only known for H^* and these are not the asymptotic distributions of the profile loss function (test statistics) in the presence of the nuisance parameter. Although the asymptotic distribution for the profile loss function remains unknown, we can still consistently estimate the proportion of null effects based on any quantile because of the nestedness of the support space and the nature of the NPMDE.

Chapter 6

Large-Scale Computation

6.1 Introduction

The non-parametric mixture methods of estimating a mixing distribution described in Chapters 3 and 5 already have the speed advantage over most existing methods (Wang, 2007). In this chapter, modifications for large-scale computation are introduced to further reduce the computational burden when the number of observations is large.

The most commonly-used method to reduce the computational burden in density estimation is an approximation technique known as binning (Scott, 1981). Instead of using every single observation in the estimation process, all the observations are placed into m disjoint bins with $m \ll n$, and the estimation process only uses the representation and the frequency of each bin. If original observations are realisations of a continuous distribution, the corresponding discrete version of the distribution should be used. It is important to note that after binning is performed, estimation on the binned data is only an approximation to the original observations. For the remainder of this chapter, we take the normal distribution as an example and show how the discrete version of the normal distribution (discrete normal distribution) can be used to estimate a mixing distribution. First of all, the definition of the discrete normal distribution needs to be introduced.

Definition 6.1 (Discrete Normal Distribution)**.** *Using the definition and Result 2.1 in Roy (2003), the probability mass function of a discrete normal distribution with the mean μ, the variance σ^2 and the bin width h is*

$$\phi_d(x;\mu,\sigma,h) = \begin{cases} \Phi(x+h;\mu,\sigma) - \Phi(x;\mu,\sigma), & x = \cdots, -h, 0, h, \cdots \\ 0, & \textit{otherwise,} \end{cases}$$

with the corresponding cumulative distribution function

$$\Phi_d(x;\mu,\sigma,h) = \Phi(x+h;\mu,\sigma) = \Phi(x;\mu-h,\sigma) \qquad x = \cdots, -h, 0, h, \cdots.$$

According to Definition 6.1, the sufficient representation of each bin is the lower value and the frequency in each bin. In the pre-processing step, all the original observations should be rounded down to the lower values of the bins and the number of original observations should be counted in each bin.

6.2 Maximum Likelihood

The formulation in this section will be based on the restricted NPMLE using discrete normal components. The formulation for the unrestricted NPMLE can be easily obtained by setting $\pi_0 = 0$. Let

$$f_d(\cdot; G, \sigma, \pi_0, h) = \pi_0 \phi_d(\cdot; 0, \sigma, h) + \int \phi_d(\cdot; \mu, \sigma, h) dG(\mu)$$

be the discrete normal mixture given a fixed support point at 0 with probability π_0 and the bin width h. The vector $\boldsymbol{y} = (y_1, \cdots, y_m)^T$ is the lower values of the bins and the vector $\boldsymbol{w} = (w_1, \cdots, w_m)^T$ with $\mathbf{1}^T\boldsymbol{w} = n$ the corresponding frequencies. We can derive the log-likelihood for the discrete normal mixture

$$\begin{aligned}\ell_n(G) &= \sum_{i=1}^{n} \log f_d\left(\left\lfloor \frac{x_{(i)}}{h} \right\rfloor h; G, \sigma, \pi_0, h\right) \\ &= \sum_{i=1}^{m} w_i \log f_d(y_i; G, \sigma, \pi_0, h),\end{aligned}$$

with the gradient function

$$\begin{aligned}d(\mu; G) &\equiv \left.\frac{\partial \ell\{(1-\epsilon)G + \epsilon \mathbb{1}_\mu\}}{\partial \epsilon}\right|_{\epsilon=0} \\ &= (1-\pi_0) \sum_{i=1}^{m} w_i \frac{\phi_d(y_i;\mu,\sigma,h) - f_d(y_i; G, \sigma, h)}{f_d(y_i; G, \sigma, \pi_0, h)}\end{aligned}$$

The first and the second derivative of the gradient function need to be modified if one wants to find all the local maxima, i.e.,

$$\begin{aligned}
\frac{\partial d(\mu, G)}{\partial \mu} &= (1-\pi_0)\sum_{i=1}^{n} w_i \frac{\frac{\partial}{\partial \mu}\phi_d(y_i;\mu,\sigma)}{f_d(y_i; G, \sigma, \pi_0, h)} \\
&= (1-\pi_0)\sum_{i=1}^{n} w_i \frac{\phi(y_i;\mu,\sigma) - \phi(y_i;\mu-h,\sigma)}{f_d(y_i; G, \sigma, \pi_0, h)}, \\
\frac{\partial^2 d(\mu, G)}{\partial \mu^2} &= (1-\pi_0)\sum_{i=1}^{n} w_i \frac{\frac{\partial^2}{\partial \mu^2}\phi_d(y_i;\mu,\sigma)}{f_d(y_i; G_s, \sigma, \pi_0, h)} \\
&= (1-\pi_0)\sum_{i=1}^{n} w_i \frac{(y_i-\mu)\phi(y_i;\mu,\sigma) - (y_i+h-\mu)\phi(y_i;\mu-h,\sigma)}{f_d(y_i; G, \sigma, \pi_0, h)}.
\end{aligned}$$

The derivatives shown above are similar to the forward difference in approximating numerical derivatives. The limits of these derivatives will converge to the ones in Example 3.1 as h tends to 0. The next step is to compute the weights given fixed support points. The second order Taylor expansion of the log-likelihood function now becomes

$$\ell_n(G_s) = \boldsymbol{w}^T \boldsymbol{S}(\boldsymbol{\eta}' - \boldsymbol{\eta}) - \frac{1}{2}(\boldsymbol{\eta}' - \boldsymbol{\eta})^T \boldsymbol{S}^T \boldsymbol{W} \boldsymbol{S}(\boldsymbol{\eta}' - \boldsymbol{\eta}), \tag{6.1}$$

where

$$\begin{aligned}
\boldsymbol{S}^T &= (\boldsymbol{s}_1, \cdots, \boldsymbol{s}_m), \\
\boldsymbol{s}_i &= \left(\frac{\phi_d(y_i; \mu_1', \sigma, h)}{f_d(y_i; G_{s-1}, \sigma, \pi_0, h)}, \cdots, \frac{\phi_d(y_i; \mu_{k'}', \sigma, h)}{f_d(y_i; G_{s-1}, \sigma, \pi_0, h)} \right)^T, \\
\boldsymbol{W} &= \mathrm{diag}(\boldsymbol{w}).
\end{aligned}$$

Let

$$\boldsymbol{s}_{\pi_0} = \left(\frac{\pi_0 \phi_d(y_1, 0, \sigma, h)}{f_d(y_1; G_{s-1}, \sigma, \pi_0, h)}, \cdots, \frac{\pi_0 \phi_d(y_m, 0, \sigma, h)}{f_d(y_m; G_{s-1}, \sigma, \pi_0, h)} \right)^T,$$

the following equations can also be obtained,

$$\begin{aligned}
\boldsymbol{S}\boldsymbol{\eta} + \boldsymbol{s}_{\pi_0} &= \mathbf{1}, \\
\mathbf{1}^T \boldsymbol{\eta}' = \mathbf{1}^T \boldsymbol{\eta} &= 1 - \mathbf{1}^T \boldsymbol{\pi}.
\end{aligned}$$

Maximising (6.1) can be turned into the following optimisation problem,

$$\text{Minimise} \qquad ||\sqrt{\boldsymbol{W}}\boldsymbol{S}\boldsymbol{\eta}' - \sqrt{\boldsymbol{W}}(\mathbf{2} - \boldsymbol{s}_{\pi_0})||^2,$$

subject to

$$\mathbf{1}^T\boldsymbol{\eta}' = 1 - \mathbf{1}^T\boldsymbol{\pi},$$
$$\eta_i' \geq 0, \ \forall i.$$

Note that there is no confusion for $\sqrt{\boldsymbol{W}}$ since the element-wise square root and the matrix square root are identical for a diagonal matrix. For the line search used to ensure the monotone increase and global convergence, the inequality

$$\ell_n(G_s^{c,l}) \geq \ell_n(G_s^{0,l}) + \alpha c^l \mathbf{w}^T \boldsymbol{S}(\boldsymbol{\eta}' - \boldsymbol{\eta})$$

is tested in the order $l = 0, 1, 2, \cdots$ until it is first satisfied.

6.3 The Cramér-von Mises Distance

Since each element in $\boldsymbol{Y}$ follows a discrete normal mixture, discrete test statistics should be used when computing the NPMDEs under the Cramér-von Mises distance. The formulation for the discrete Cramér-von Mises statistics can be found in Choulakian et al. (1994), but we are not going to use the exact formulation since the statistic is cubic in G, which creates unnecessary complications for computation.

Choulakian et al. (1994) also mentioned the usage of the modified version of the Cramér-von Mises distance, which replaces $f_d(y_{(i)}; G, \sigma, \pi_0, h)$ with $1/m$; see Section 4 of Choulakian et al. (1994) for a more detailed discussion. Using the non-hypothesised weights enjoys great computation advantages without losing too much accuracy in the estimation, thanks to the Glivenko-Cantelli theorem. However, we argue that using the empirical weight in each bin is more appropriate and more accurate than $1/m$, although the NPMDE using $1/m$ as weights also consistently estimates the true mixing distribution as h approaches 0. This is because the effective sample size is the same as the original dataset and the loss functions are comparable. This modified version will also makes the formulation and computation much easier since the second order Taylor expansion is exact and there is no need to consider the derivative of weights with respect to G.

We refer to the computational formula for the Cramér-von Mises distance in Anderson and Darling (1952) with the appropriate replacement of the observations, the ECDF and the weights, i.e.,

$$\begin{aligned}\ell_n(G) &= \sum_{i=1}^{m} w_{(i)} \left\{ F^2(y_{(i)}; G, \cdots) - 2 \left(\hat{F}_n(y_{(i)} + h) - \frac{w_{(i)}}{2n} \right) F(y_{(i)}; G, \cdots) \right\} + \frac{n}{3} \\ &= \sum_{i=1}^{m} w_{(i)} \left[F(y_{(i)}; G, \cdots) - \hat{F}_n(y_{(i)} + h) + \frac{w_{(i)}}{2n} \right]^2 + \frac{n}{3} \\ &\quad - \sum_{i=1}^{m} w_{(i)} \left[\hat{F}_n(y_{(i)} + h) - \frac{w_{(i)}}{2n} \right]^2.\end{aligned}$$

The equation above can be used in estimating the null proportion. The optimisation problem for finding the restricted NPMDE becomes

$$\hat{G} = \underset{G \in \Gamma}{\arg\min} \sum_{i=1}^{m} w_{(i)} \left[F_d(y_{(i)}; G, \sigma, \pi_0, h) - \hat{F}_n(y_{(i)} + h) + \frac{w_{(i)}}{2n} \right]^2,$$

and the process of computing the restricted NPMDE under this metric is almost identical to the one described in Section 5.2. The location of the null effect is uniquely defined, and hence the unrestricted NPMLE can be found in a similar way by setting $\boldsymbol{\pi}$ to be a vector of 0 with length 1.

6.4 The Anderson-Darling Distance

The formulation for the discrete Anderson-Darling statistics can also be found in Choulakian et al. (1994), but similarly we will not use the exact formulation. As a result, we use the modified version of the Anderson-Darling statistic by replacing the hypothesised weights with the empirical ones. We also refer to the computational formula for the Anderson-Darling distance in Anderson and Darling (1952) with the appropriate replacement of the observations, the ECDF and the weights, i.e.,

$$\begin{aligned}\ell_n(G) = &- n - \sum_{i=1}^{m} w_{(i)} \left[\left(2\hat{F}_n(y_{(i)} + h) - \frac{w_{(i)}}{n} \right) \log[F(y_{(i)}; G, \cdots)] \right. \\ &\left. + \left(\frac{2n + w_{(i)}}{n} - 2\hat{F}_n(y_{(i)} + h) \right) \log[1 - F(y_{(i)}; G, \cdots)] \right].\end{aligned}$$

The equation above can be used in estimating the null proportion. The optimisation problem for finding the restricted NPMDE becomes

$$\hat{G} = \arg\max_{G\in\Gamma} \sum_{i=1}^{m} w_i \left[\left(2\hat{F}_n(y_{(i)} + h) - \frac{w_{(i)}}{n} \right) \log[F_d(y_{(i)}; G, \sigma, \pi_0, h)] \right.$$
$$\left. + \left(\frac{2n + w_{(i)}}{n} - 2\hat{F}_n(y_{(i)} + h) \right) \log[1 - F_d(y_{(i)}; G, \sigma, \pi_0, h)] \right]$$

and the process of computing the restricted NPMDE under this metric is almost identical to the one described in Section 5.3. The location of the null effect is uniquely defined, and hence the unrestricted NPMLE can be found by setting $\boldsymbol{\pi}$ to be a vector of 0 with length 1.

6.5 Choice of Bin Width

There is no definitive answer for the choice of the bin width h, different values of the bin width can reveal different features of the observations. Bin widths for histograms or the width selection in the kernel density estimation can be used. The conventional relationship between the number of bins and the bin width is

$$k = \left\lceil \frac{\max x - \min x}{h} \right\rceil, \qquad \text{or}$$
$$h = \frac{\max x - \min x}{k}.$$

Once a good choice of the bin width is obtained, the corresponding number of bins can be computed and vice versa. Sturges (1926) derived a formula from a binomial distribution and implicitly assumed an approximated normal distribution. The formula,

$$k = \lceil \log_2 n \rceil + 1,$$

performs poorly when the number of observations is small or the observations are not normally distributed. Doane (1976) proposed a modification of the formula proposed in Sturges (1926) which attempts to improve its performance with non-normal data, i.e.,

$$k = 1 + \log_2 n + \log_2 \left(1 + \frac{|g_1|}{\sigma_{g_1}} \right),$$

where g_1 is the estimated skewness of the distribution and

$$\sigma_{g_1} = \sqrt{\frac{6(n-2)}{(n+1)(n+3)}}.$$

The Scott's normal reference rule in Scott (1979) has the following formula

$$h = 3.49\hat{\sigma}n^{-1/3},$$

where $\hat{\sigma}$ is the sample standard deviation. Scott's normal reference rule is optimal for normally distributed random samples, in the sense that it minimises the integrated mean squared error of the estimated density (Scott, 1992). Freedman and Diaconis (1981) replaced $3.49\hat{\sigma}$ in the Scott's normal reference rule with twice the interquartile range, which is less sensitive than the standard deviation to outliers in data. The idea of minimising the integrated mean squared error from Scott (1979) was further generalised by Stone (1984) beyond normal distributions, by using generalised cross-validation. The choice proposed in Shimazaki and Shinomoto (2007) is based on minimisation of an estimated L_2 risk function, i.e.,

$$h = \arg\min_{h'} \frac{2\hat{\mu} - \hat{\sigma}}{h^2},$$

where $\hat{\mu}$ and $\hat{\sigma}$ are the sample mean and the sample variance of the binned data given bin width h'.

When choosing the bin width for the proposed methods, the distance function used should be taken into consideration. The NPMLE or NPMDE can always be computed for any fixed bin width h, and different choices of h lead to approximations to the NPMLE or NPMDE on the original dataset of different accuracy. For a large h, computing the estimate of the mixing distribution and proportion of null effects is faster but the accuracy might be poor.

In some applications, it is more preferable to use bins with varying width. Estimation with varying-width bins can be done by a further modification to the formulations in this section. The working will not be shown here since it is not of primary interest as well as the working is similar to the one with the fixed bin-width.

6.6 Numerical Studies

In this section, numerical studies are conducted in order to compare the performance of the large-scale modification of the proposed methods. Estimations using binned data under the method of maximum likelihood, the Cramér-von Mises distance and the Anderson-Darling distance are compared with the corresponding one using original data. The performance is investigated through the computation time, the distance between the binned estimates and the estimate using the original dataset measured by the metric in Kiefer and Wolfowitz (1956), and estimation of the null proportion. The synthetic datasets in Section 4.2 with $\pi_0 = 0.9$ are used.

In order to process the data easily, the choice of the bin width is $h = 10^{-a}$ for a non-negative integer a. In this case, all the observations are round down to a decimal places. We choose $h = 10^{-2}, 10^{-4}$ in the numerical studies and these choices of h give reasonable approximations of the estimated mixing distribution computed using the original dataset. It is expected that $h = 10^{-4}$ will give a more accurate result but use more time compared with $h = 10^{-2}$.

n	h	$\pi_0 = 0$	$\pi_0 = 0.3$	$\pi_0 = 0.6$	$\pi_0 = 0.9$
1500	0	0.085(0.002)	0.085(0.003)	0.070(0.002)	0.059(0.001)
1500	10^{-4}	0.409(0.011)	0.398(0.016)	0.318(0.008)	0.263(0.005)
1500	10^{-2}	0.181(0.006)	0.162(0.004)	0.127(0.004)	0.093(0.002)
3000	0	0.194(0.004)	0.204(0.009)	0.160(0.006)	0.138(0.003)
3000	10^{-4}	0.907(0.018)	0.901(0.034)	0.697(0.022)	0.592(0.013)
3000	10^{-2}	0.234(0.007)	0.193(0.005)	0.157(0.006)	0.120(0.003)
6000	0	0.408(0.013)	0.395(0.017)	0.311(0.015)	0.258(0.006)
6000	10^{-4}	1.993(0.045)	1.932(0.064)	1.444(0.040)	1.171(0.025)
6000	10^{-2}	0.287(0.008)	0.266(0.012)	0.199(0.008)	0.156(0.004)

TABLE 6.1: The mean (standard error) of computation times (in seconds) of estimating the mixing distribution using maximum likelihood for each combination of (n, π_0) from the 100 synthetic datasets generated by the prostate data.

n	h	$\pi_0 = 0$	$\pi_0 = 0.3$	$\pi_0 = 0.6$	$\pi_0 = 0.9$
1500	0	0.049(0.001)	0.046(0.001)	0.043(0.001)	0.038(0.001)
1500	10^{-4}	0.156(0.004)	0.151(0.003)	0.150(0.003)	0.133(0.003)
1500	10^{-2}	0.080(0.002)	0.078(0.002)	0.071(0.001)	0.052(0.001)
3000	0	0.098(0.002)	0.096(0.002)	0.095(0.003)	0.087(0.003)
3000	10^{-4}	0.306(0.006)	0.298(0.006)	0.297(0.006)	0.268(0.005)
3000	10^{-2}	0.100(0.002)	0.095(0.002)	0.087(0.002)	0.065(0.002)
6000	0	0.183(0.004)	0.173(0.003)	0.161(0.003)	0.152(0.004)
6000	10^{-4}	0.589(0.011)	0.549(0.011)	0.567(0.014)	0.537(0.010)
6000	10^{-2}	0.119(0.002)	0.112(0.003)	0.100(0.003)	0.077(0.002)

TABLE 6.2: The mean (standard error) of computation times (in seconds) of estimating the mixing distribution using the Cramér-von Mises distance for each combination of (n, π_0) from the 100 synthetic datasets generated by the prostate data.

n	h	$\pi_0 = 0$	$\pi_0 = 0.3$	$\pi_0 = 0.6$	$\pi_0 = 0.9$
1500	0	0.103(0.003)	0.102(0.002)	0.093(0.003)	0.071(0.002)
1500	10^{-4}	0.111(0.002)	0.109(0.002)	0.097(0.003)	0.076(0.002)
1500	10^{-2}	0.042(0.001)	0.040(0.001)	0.035(0.001)	0.024(0.001)
3000	0	0.232(0.007)	0.231(0.006)	0.207(0.009)	0.150(0.005)
3000	10^{-4}	0.206(0.005)	0.204(0.005)	0.191(0.008)	0.145(0.005)
3000	10^{-2}	0.047(0.001)	0.046(0.001)	0.037(0.001)	0.029(0.001)
6000	0	0.473(0.014)	0.465(0.013)	0.391(0.015)	0.317(0.014)
6000	10^{-4}	0.402(0.010)	0.390(0.009)	0.332(0.010)	0.258(0.009)
6000	10^{-2}	0.051(0.001)	0.050(0.001)	0.040(0.001)	0.034(0.001)

TABLE 6.3: The mean (standard error) of computation times (in seconds) of estimating the mixing distribution using the Anderson-Darling distance for each combination of (n, π_0) from the 100 synthetic datasets generated by the prostate data.

n	h	$\pi_0 = 0$	$\pi_0 = 0.3$	$\pi_0 = 0.6$	$\pi_0 = 0.9$
1500	0	0.109(0.007)	0.098(0.006)	0.079(0.003)	0.029(0.001)
1500	10^{-4}	0.109(0.007)	0.098(0.006)	0.079(0.003)	0.029(0.001)
1500	10^{-2}	0.109(0.007)	0.098(0.006)	0.079(0.003)	0.029(0.001)
3000	0	0.094(0.007)	0.085(0.005)	0.070(0.003)	0.025(0.001)
3000	10^{-4}	0.094(0.007)	0.085(0.005)	0.070(0.003)	0.025(0.001)
3000	10^{-2}	0.094(0.007)	0.085(0.005)	0.070(0.003)	0.025(0.001)
6000	0	0.096(0.006)	0.086(0.005)	0.069(0.003)	0.024(0.001)
6000	10^{-4}	0.096(0.006)	0.086(0.005)	0.069(0.003)	0.024(0.001)
6000	10^{-2}	0.096(0.006)	0.086(0.005)	0.069(0.003)	0.024(0.001)

TABLE 6.4: The mean (standard error) of distances to true mixing distribution using maximum likelihood for each combination of (n, π_0) from the 100 synthetic datasets generated by the prostate data.

n	h	$\pi_0 = 0$	$\pi_0 = 0.3$	$\pi_0 = 0.6$	$\pi_0 = 0.9$
1500	0	0.140(0.009)	0.125(0.007)	0.099(0.004)	0.035(0.001)
1500	10^{-4}	0.139(0.008)	0.124(0.007)	0.099(0.004)	0.035(0.001)
1500	10^{-2}	0.153(0.009)	0.130(0.007)	0.101(0.004)	0.034(0.001)
3000	0	0.132(0.009)	0.117(0.007)	0.092(0.004)	0.030(0.001)
3000	10^{-4}	0.134(0.009)	0.118(0.007)	0.091(0.004)	0.030(0.001)
3000	10^{-2}	0.136(0.009)	0.118(0.007)	0.096(0.004)	0.030(0.001)
6000	0	0.116(0.008)	0.105(0.006)	0.085(0.004)	0.025(0.001)
6000	10^{-4}	0.117(0.007)	0.105(0.006)	0.084(0.004)	0.025(0.001)
6000	10^{-2}	0.121(0.008)	0.107(0.006)	0.085(0.004)	0.025(0.001)

TABLE 6.5: The mean (standard error) of distances to true mixing distribution using the Cramér-von Mises distance for each combination of (n, π_0) from the 100 synthetic datasets generated by the prostate data.

n	h	$\pi_0 = 0$	$\pi_0 = 0.3$	$\pi_0 = 0.6$	$\pi_0 = 0.9$
1500	0	0.126(0.008)	0.116(0.006)	0.091(0.004)	0.032(0.001)
1500	10^{-4}	0.127(0.008)	0.116(0.006)	0.092(0.004)	0.032(0.001)
1500	10^{-2}	0.128(0.008)	0.116(0.006)	0.093(0.004)	0.032(0.001)
3000	0	0.114(0.008)	0.102(0.006)	0.085(0.004)	0.028(0.001)
3000	10^{-4}	0.114(0.008)	0.102(0.006)	0.085(0.004)	0.028(0.001)
3000	10^{-2}	0.114(0.008)	0.104(0.006)	0.083(0.004)	0.028(0.001)
6000	0	0.107(0.007)	0.098(0.006)	0.081(0.004)	0.025(0.001)
6000	10^{-4}	0.107(0.007)	0.097(0.006)	0.081(0.004)	0.024(0.001)
6000	10^{-2}	0.109(0.007)	0.099(0.006)	0.081(0.004)	0.024(0.001)

TABLE 6.6: The mean (standard error) of distances to true mixing distribution using the Anderson-Darling distance for each combination of (n, π_0) from the 100 synthetic datasets generated by the prostate data.

Table 6.1 shows the means and standard errors of the computation time of estimating the mixing distribution using maximum likelihood for various n and bin widths. The computation time decreases when π_0 increases, and seems to increase linearly with the increase of sample size. For any fixed π_0 value, by using bin width $h = 10^{-2}$, the computation time is reduced dramatically, as expected. However, the time taken to compute the mixing distribution with bin width $h = 10^{-4}$ is more than the one without binning. Further investigation reviews that the number of observations when $h = 10^{-4}$ is very close to the number of observations in the original dataset. Using bin width $h = 10^{-4}$ gives 1472, 2887 and 5562 observations on average respectively while the mean number of observations is 457, 545 and 630 respectively when bin width $h = 10^{-2}$ is used for $n = 1500, 3000, 6000$. Similar behaviour is also observed when estimating the NPMDE under the Cramér-von Mises and the Anderson-Darling distance, as shown in Tables 6.2 and 6.3. It is expected that when the number of observations are similar, estimation using the binned data will be much slower than that using the original observations, since the original observations will be binned first and extra pre-caution is needed in evaluating the probability mass of the discrete normal distribution especially for values that are far way from the support points. Tables 6.4, 6.5 and 6.6 show the means and the standard errors of the distance between the estimated mixing distribution and the true one. As expected, the distances to the true mixing distribution are decreasing as π_0 and the number of observations increases. The NPMDE under the Cramér-von Mises distance is more sensitive to the choice of the bin width than the NPMLE and the NPMDE under the Anderson-Darling distance. The numerical study

in this section also suggests that the choices of the bin width are sensible and can be applied to the large-scale computation of the real world dataset in the next section.

6.7 Prostate Data Continued

As shown in the previous section the usage of binning significantly reduces the computation time for estimating the mixing distributions without losing too much accuracy. In this section, we are investigating how binning affects the estimation of null proportions. The NPMLEs for all the synthetic datasets in Section 4.2 are computed with $h = 10^{-2}, 10^{-4}$. The results are shown in Table 6.7 and corresponding boxplots are shown in Figures 6.1, 6.2 and 6.3.

		no binning		$h = 10^{-4}$		$h = 10^{-2}$	
π_0^*	n	AIC	BIC	AIC	BIC	AIC	BIC
0.5	1500	0.675(0.005)	0.732(0.003)	0.675(0.005)	0.732(0.003)	0.675(0.005)	0.732(0.003)
	3000	0.653(0.006)	0.712(0.003)	0.653(0.006)	0.712(0.003)	0.653(0.006)	0.712(0.003)
	6000	0.636(0.006)	0.695(0.002)	0.636(0.006)	0.695(0.002)	0.636(0.006)	0.695(0.002)
0.7	1500	0.813(0.004)	0.860(0.002)	0.813(0.004)	0.860(0.002)	0.813(0.004)	0.860(0.002)
	3000	0.806(0.004)	0.847(0.002)	0.806(0.004)	0.847(0.002)	0.806(0.004)	0.847(0.002)
	6000	0.795(0.004)	0.836(0.002)	0.795(0.004)	0.836(0.002)	0.795(0.004)	0.836(0.002)
0.9	1500	0.946(0.003)	0.973(0.001)	0.946(0.003)	0.973(0.001)	0.946(0.003)	0.973(0.001)
	3000	0.942(0.002)	0.966(0.001)	0.942(0.002)	0.966(0.001)	0.942(0.002)	0.966(0.001)
	6000	0.940(0.002)	0.961(0.001)	0.940(0.002)	0.961(0.001)	0.940(0.002)	0.961(0.001)

TABLE 6.7: The mean (standard error) of the estimated null proportions from the 100 synthetic datasets based on the prostate data for various n, π_0 and h.

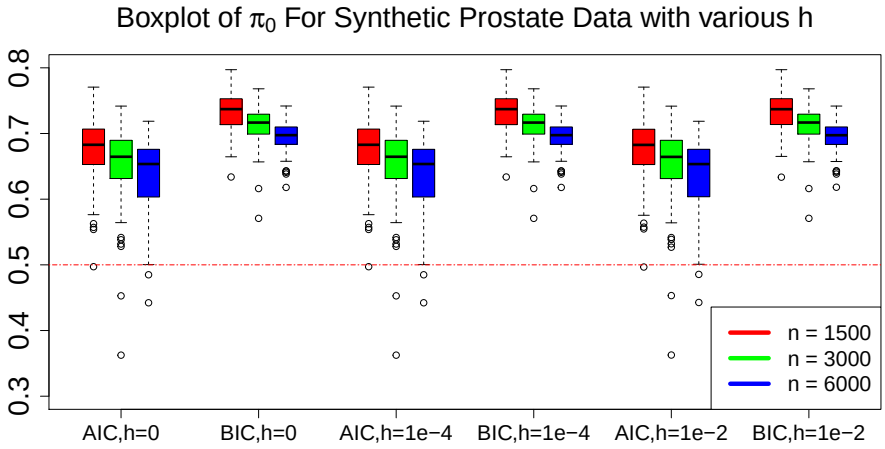

FIGURE 6.1: Boxplot of estimated null proportions from the 100 synthetic datasets generated from the prostate data for various threshold functions, n and bin widths when $\pi_0^* = 0.5$.

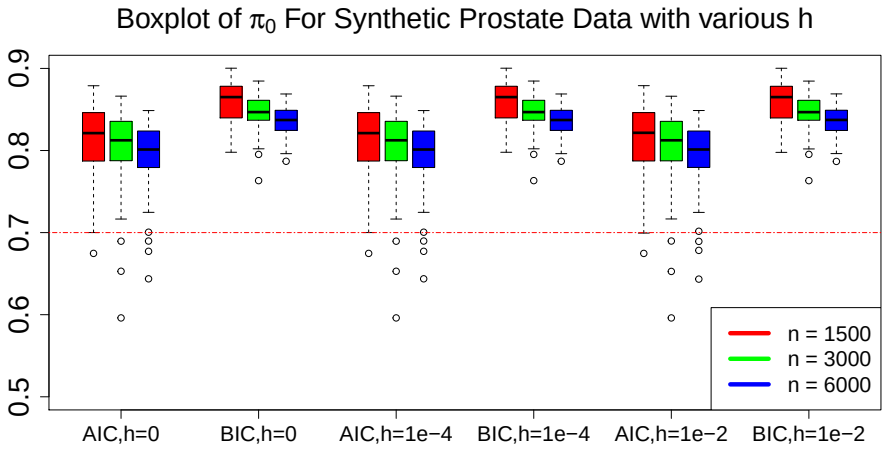

FIGURE 6.2: Boxplot of estimated null proportions from the 100 synthetic datasets generated from the prostate data for various threshold functions, n and bin widths when $\pi_0^* = 0.7$.

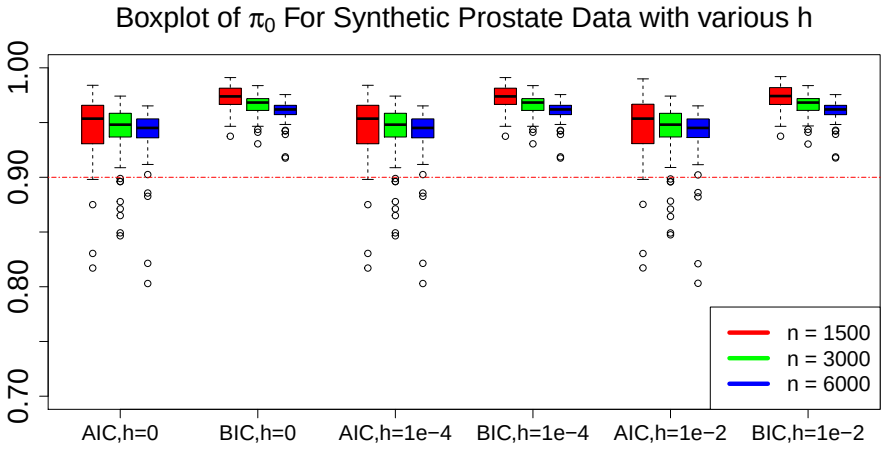

FIGURE 6.3: Boxplot of estimated null proportions from the 100 synthetic datasets generated from the prostate data for various threshold functions, n and bin widths when $\pi_0^* = 0.9$.

Table 6.7 shows that the means and standard errors of the estimated null proportions when different bin widths are used. In general, the mean and standard error of the estimated null proportions when binning is used are very similar to the corresponding ones without binning. The differences are smaller than the standard errors of the estimated null proportions. They are good estimates, considering that the original observations are rounded down to 2 decimal places and that the computation time has dropped significantly. Figures 6.1, 6.2, 6.3 show the corresponding boxplot of the estimated null proportions for various π_0^*. There is virtually no difference between the boxes when different bin widths are used except for some extreme observations.

Since using binning does not lose too much accuracy in estimating the null proportions, we can extend the numerical study in Section 4.2 with a larger number of observations ($n = 12000, 24000, 48000$) and see the performance of the proposed method against other methods. We estimate the null proportions using $h = 10^{-2}$ and compare with other methods. Table 6.8 shows the mean effective number of observations for various n and π_0^* when the bin width $h = 10^{-2}$ is used. The effective sample size increases when the number of observations in the original dataset increases or π_0^* decreases. The theoretical number of effective sample size may be derived using the law of the iterated logarithm under the normality assumption.

π_0^*	$n = 12000$	$n = 24000$	$n = 48000$
0.5	887	957	1016
0.7	830	908	974
0.9	710	794	869

TABLE 6.8: The mean effective sample size when binning is used with $h = 10^{-2}$.

π_0^*	n	npnorm-AIC	npnorm-BIC	locfdr	ECF	mixfdr
	1500	0.675(0.005)	0.732(0.003)	0.739(0.001)	**0.671**(0.003)	0.836(0.001)
	3000	**0.653**(0.006)	0.712(0.003)	0.739(0.001)	0.660(0.003)	0.835(0.001)
0.5	6000	**0.636**(0.006)	0.695(0.002)	0.739(0.001)	0.651(0.002)	0.835(0.001)
	12000	**0.613**(0.005)	0.676(0.002)	0.738(0.001)	0.642(0.002)	0.835(0.001)
	24000	**0.601**(0.005)	0.664(0.002)	0.738(0.000)	0.638(0.001)	0.836(0.000)
	48000	**0.577**(0.005)	0.646(0.002)	0.737(0.000)	0.629(0.001)	0.835(0.000)
	1500	0.813(0.004)	0.860(0.002)	0.839(0.002)	**0.800**(0.003)	0.913(0.001)
	3000	0.806(0.004)	0.847(0.002)	0.839(0.001)	**0.796**(0.003)	0.912(0.001)
0.7	6000	0.795(0.004)	0.836(0.002)	0.839(0.001)	**0.792**(0.002)	0.911(0.001)
	12000	**0.783**(0.003)	0.824(0.001)	0.837(0.001)	0.786(0.002)	0.910(0.000)
	24000	**0.779**(0.003)	0.816(0.001)	0.838(0.000)	0.783(0.001)	0.911(0.000)
	48000	**0.758**(0.003)	0.803(0.001)	0.837(0.000)	0.777(0.001)	0.910(0.000)
	1500	0.946(0.003)	0.973(0.001)	0.947(0.002)	**0.932**(0.003)	0.976(0.000)
	3000	0.942(0.002)	0.966(0.001)	0.947(0.002)	**0.931**(0.002)	0.973(0.000)
0.9	6000	0.940(0.002)	0.961(0.001)	0.947(0.001)	**0.929**(0.002)	0.972(0.000)
	12000	0.937(0.002)	0.956(0.001)	0.947(0.001)	**0.927**(0.002)	0.972(0.000)
	24000	0.936(0.001)	0.952(0.001)	0.947(0.000)	**0.926**(0.001)	0.972(0.000)
	48000	0.930(0.001)	0.947(0.001)	0.948(0.000)	**0.927**(0.001)	0.971(0.000)

TABLE 6.9: The mean (standard error) of estimated null proportions from the 100 synthetic datasets generated from the prostate data for large n and π_0 with bin width $h = 10^{-2}$.

Table 6.9 shows the extended version of Table 4.1 with the estimated proportion of null effects computed using $h = 10^{-2}$ for $n = 12000, 24000, 48000$ and Figures 6.4, 6.5 and 6.6 show the corresponding boxplots. The red dotted lines show the true proportions. The table and all of the figures show a clear decreasing trend for the proposed method. The estimated null proportion seems to be constant for the locfdr and the mixfdr methods. For $\pi_0^* = 0.5$, the mean of the estimated null proportions using the AIC threshold function is similar to the ECF method when $n = 1500$ but is dropped by 0.098 as n increases to 48000 while the mean estimated using the ECF method is only dropped by 0.042. The proposed method using the BIC threshold

function is also dropped by 0.086 when n increases from 1500 to 48000. Similar results can be obtained when $\pi_0^* = 0.7$ and 0.9. Although the ECF method performs the best when $\pi_0^* = 0.9$, it seems that the estimated null proportion converges fairly slowly to π_0^*, whereas the proposed method still has an obvious decreasing trend as n increases.

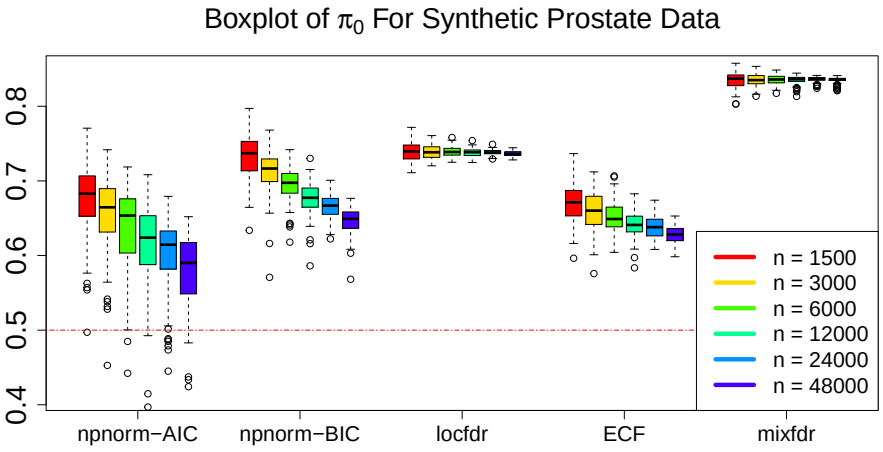

FIGURE 6.4: Boxplot of estimated null proportions for the 100 synthetic datasets generated from the prostate data for each method and each n when $\pi_0^* = 0.5$ using $h = 10^{-2}$.

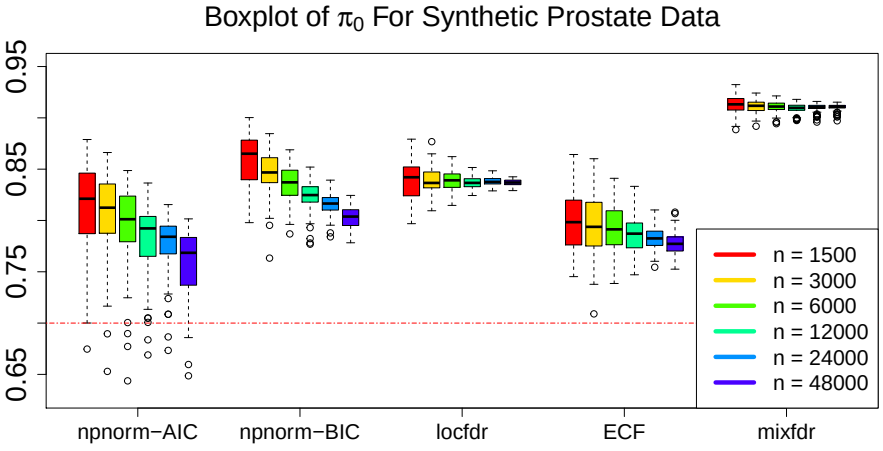

FIGURE 6.5: Boxplot of estimated null proportions for the 100 synthetic datasets generated from the prostate data for each method and each n when $\pi_0^* = 0.7$ using $h = 10^{-2}$.

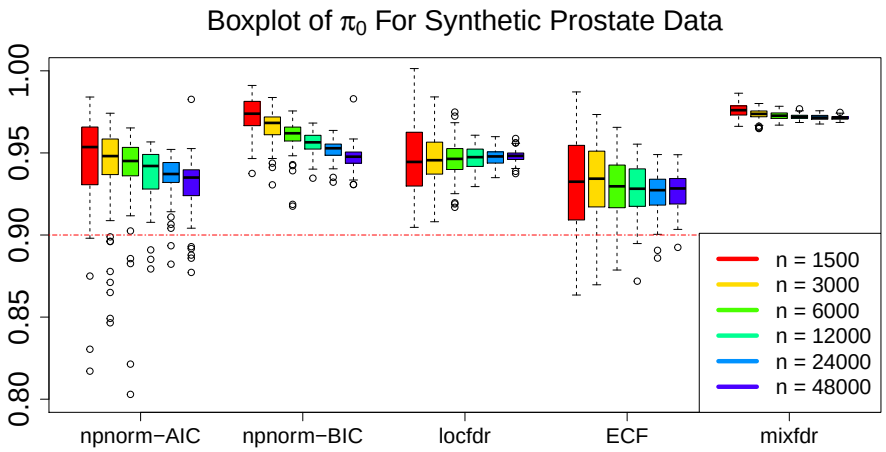

FIGURE 6.6: Boxplot of estimated null proportions for the 100 synthetic datasets generated from the prostate data for each method and each n when $\pi_0^* = 0.9$ using $h = 10^{-2}$.

Due to the incompleteness of the asymptotic theory for (log-)likelihood ratio under non-parametric setting, the exact rate of convergence of the null proportion estimator is unknown. However, a simple linear regression

$$\log|\hat{\pi}_0 - \pi_0^*| = \beta_0 + \beta_1 \log(n),$$

where β_0 is a factor and β_1 is the rate of convergence, is used to assess the empirical rate of convergence, and the results are shown in Table 6.10. The empirical rates for the locfdr and mixfdr methods are very close to 0, which is consistent with a constant trend in Figures 6.4, 6.5 and 6.6. Our proposed methods have the fastest empirical rate of convergence, significantly higher than that of the ECF method.

π_0^*	npnorm-AIC	npnorm-BIC	locfdr	ECF	mixfdr
0.5	$-$**0.272**(0.028)	$-$0.131(0.005)	$-$0.003(0.001)	$-$0.075(0.005)	0.000(0.001)
0.7	$-$**0.202**(0.028)	$-$0.124(0.004)	$-$0.003(0.003)	$-$0.058(0.009)	$-$0.003(0.001)
0.9	$-$0.114(0.020)	$-$**0.123**(0.005)	0.047(0.011)	$-$0.024(0.029)	$-$0.016(0.001)

TABLE 6.10: The empirical rate (standard error) of convergence of the null proportion estimator for various methods and π_0^*.

6.8 Conclusion

In this chapter, binning is used to reduce the computation burden by reducing the effective sample size and it can be shown that all the proposed methods can be modified to accommodate the large-scale computation. The usage of binning does not lose too much accuracy in estimation but significantly reduces the computation time in large-scale setting, as shown in the numerical study in Section 6.6. The numerical study in Section 4.2 is extended to an even larger number of observations and it can be found that the proposed method mentioned in Chapter 3 still has a significant converging trend compared with other existing methods.

Chapter 7

Controlling the False Discovery Rate

7.1 Introduction

In this chapter, a non-conservative method is introduced in order to make inferences about hypotheses as an extension to the non-parametric density estimation and estimation of the null proportion. Similar to the method described in Benjamini et al. (2006), we first estimate the null proportion and use the estimated mixing distribution based on this estimated null proportion to construct a critical region in order to make decisions on whether to reject hypotheses. If an estimator of the proportion of null effects is consistent, then using the Helly-Bray and the continuous mapping theorem, the density estimator for all the statistics is consistent and the target quantity can be controlled at the pre-specified level exactly. This is more powerful and hence can detect more non-null hypotheses than the Bonferroni procedures. Under finite observations, once a model is chosen from the family of estimates, the controlling procedure only depends on the estimated density instead of the observations directly, which should be more robust in estimation and efficient in computation.

The rest of this chapter is organised as follows. A literature review on controlling procedures is given in Section 7.2. A new FDR-controlling procedure based on the estimated null proportion and its estimated mixing distribution is proposed in Section 7.3 and numerical studies are given in Section 7.4.

7.2 Literature Review

In previous chapters, efforts have been made to exploit the information about all the hypotheses by estimating the density of all the statistics as well as the proportion of null effects. With this information in hand, making an inference about a single

hypothesis in a controlling procedure should be more accurate and powerful. There are many controlling quantities such as the FWER, the k-FWER, the FDR, the pFDR and the FDP; see Section 3.2. Various controlling procedures were also proposed, and each controlling procedure can be used in making rejections in one or more controlling quantities.

The most widely-used controlling procedure is the Bonferroni procedure which utilises Boole's inequality and has been adapted to control many quantities. Holm (1979) derived a method that n hypotheses were first tested separately by using the appropriate tests at the level α/n, and then these hypotheses were rejected one at a time until no further rejections could be made. The power of the test and some applications were also discussed. A modification of the Bonferroni procedure was described in Simes (1986) for testing hypotheses. The modification makes the procedure less conservative but still simple to apply. Simulations were shown that probability of type I error does not exceed the nominal significance level in multivariate normal and gamma test statistics. Hommel (1988) mentioned that in contrast to the classical Bonferroni procedures, it was not obvious how statements about individual hypotheses could be made in the method developed by Simes (1986). A new method based on the principle of closed test procedures was proposed by Marcus et al. (1976) in order to allow making statements on individual hypotheses. Hochberg (1988) made modifications to the sequential methods by Holm (1979) and Simes (1986). Instead of rejecting a hypothesis only if its p-value and each of the smaller p-values is less than the corresponding critical value, the new method rejects all hypotheses with smaller or equal p-values to that of any one found less than its critical value. It was shown that this extended Simes's test is more powerful than the corresponding Holm's test. In microarray experiments, the test statistics and hence the p-values are correlated due, for example, to co-regulation of the genes. Westfall et al. (1993) proposed adjusted p-values for less conservative multiple testing procedures which take into account the dependence structure among the test statistics. Tusher et al. (2001) described a method, the Significance Analysis of Microarrays, that assigns a score to each gene on the basis of the change in gene expression relative to the standard deviation of repeated measurements. For genes with scores greater than an adjustable threshold, it uses permutations of the repeated measurements to estimate the FDR. Storey (2002a) used a traditional statistical idea to control the FDR and gave an estimate of the false discovery rate based on the original FDR procedure. Instead of fixing the error rate and estimating its corresponding rejection region, the new approach is to fix the rejection region and then estimate its

corresponding error rate. This new approach potentially offers increased applicability, accuracy and power. This approach was applied on both pFDR and FDR. Genovese and Wasserman (2004) extended the theory of the FDR and developed a framework in which the FDP is treated as a stochastic process. The limiting distribution of the process was obtained and the validity of a class of procedures for controlling the FDR (the expected FDP) was demonstrated. A confidence envelope for the whole FDP process was also constructed. Benjamini et al. (2006) pointed out that although the procedure derived in Benjamini and Hochberg (1995) is quite striking in its ability to control the FDR under the independence assumption regardless of the test statistics, the procedure is conservative by a factor which is the proportion of the true null hypotheses among all the hypotheses when some of the null hypotheses are not true. A two-stage procedure was proposed which estimates the null proportion in the first stage and then uses the estimated null proportion in the first stage to make rejections in the second stage. Neuvial (2008) filled a gap in the theory. A variant of the stochastic process approach proposed by Genovese and Wasserman (2004) was used to study the fluctuations of the FDP achieved with each of the procedures listed in the paper around its expectation, for independent-tested hypotheses. The framework for derivation of the generic central limit theorem for the FDPs of these procedures, characterizing the associated regularity conditions, and comparing the asymptotic power of the various procedures were also introduced. Gavrilov et al. (2009) considered an adaptive step-down procedure instead of the original procedure proposed by Benjamini and Hochberg (1995), motivated by its success as a model selection procedure, as well as by its asymptotic optimality. It was also proved that this step-down procedure controls the FDR at level α for independent test statistics. Cao and Kosorok (2011) proposed an asymptotically valid data-driven procedure to find critical values for rejection regions controlling the k-FWER, the FDR and the tail probability of the FDP. An advantage of this procedure, which only depends on the test statistics, is that it requires minimum conditions rather than other stringent conditions such as the existence of the moment generating function. Sun and Mclain (2012) found out that the conventional framework in multiple testing, which involves rescaling or standardisation, is likely to distort the scientific questions. As a result, a concept of composite null distribution for heteroscedastic models was introduced and the corresponding optimal testing procedure which minimises the false non-discovery rate subject to a constraint on the FDR was proposed. Efforts have also been made when the dependent structures among the hypotheses are known. Ramdas et al. (2019) used p-filter algorithm to improve the power and

precision in multiple testing by partitioning the hypotheses into groups based on the prior knowledge. Under the constraint that the FDR is controlled at a pre-specified level, the decision to reject a single hypothesis depends on other hypotheses in the same layer. In the example of genomics, since genes are not generally independent or even marginally identically distributed, model flexibility is an essential task regardless of dependent structures among genes. Kang (2020) considered a two-stage procedure incorporating a version of the Chen-Stein theorem in Chen (1975). The new procedure seems to have better average power without much elevation of the FDR compared with the traditional Bonferroni procedure.

7.3 False Discovery Rate Control

The method to be proposed in this section is a general one and can be applied to many controlling procedures. In this section, we focus only on the FDR-controlling procedure due to its wide applications. Traditional procedures mentioned in Holm (1979) and Benjamini and Hochberg (1995) sequentially reject one single hypothesis until no further rejections can be made under the constraint that the FWER or FDR is to be controlled at a pre-specified level. With hypotheses tested simultaneously, it is possible that a large number of queries about whether a hypothesis should be deemed significant are made at the same time. Considering one hypothesis at a time seems to be too time-consuming. It would be better to know the set of intervals in which all the statistics are deemed significant and hence rejections can be made simultaneously. The critical regions in some of the traditional procedures are constructed after making rejections, which in this case the FDR can only be considered as a summary to the controlling procedure instead of a crucial piece of information used in decision making.

We aim to identify as many significant hypotheses as possible while the expectation of the proportion of false discovery Q, i.e.,

$$\mathbb{E}[Q] = \mathbb{E}[V/(V+S)] = \mathbb{E}[V/R],$$

is to be controlled at a pre-specified level. Here V is the number of cases declared significant under the true null hypothesis (Type I error), S is the number of cases declared significant under the alternative hypothesis and R is the number of cases declared significant in total. To avoid division by zero, Q is defined to be 0 when

$R = 0$ and hence formally,

$$\text{FDR} = \mathbb{E}[V/R|R > 0]\mathbb{P}(R > 0), \tag{7.1}$$

as introduced in Benjamini and Hochberg (1995). With a slight modification to the post hoc constrained optimisation problem in Benjamini and Hochberg (1995), we define our procedure for finding the critical region as

$$\text{Maximise} \qquad \mathbb{E}[R], \tag{7.2}$$

subject to

$$\mathbb{E}[V/R|R > 0]\mathbb{P}(R > 0) \leq \alpha,$$

where α is the target level for the controlling quantity. Under the independence assumption of hypotheses, the strong convergence of the ECDF of the statistics can be established and hence V and R converge to their own underlying value almost surely as the number of hypotheses increases. Since statistics are univariate and the location of the null effect is uniquely defined at 0, the critical region can be written as the interval of the form $(-\infty, a] \cup [b, \infty)$. The expectation of number of rejections, or equivalently, the expectation of the proportion of rejections can be written as

$$F(a; H^*, \sigma) + 1 - F(b; H^*, \sigma),$$

and the FDR can be written as

$$\frac{\pi_0^*[\Phi(a; 0, \sigma) + 1 - \Phi(b; 0, \sigma)]}{F(a; H^*, \sigma) + 1 - F(b; H^*, \sigma)}.$$

Note that the interval above depends on H^* and may not be symmetrical around 0. This is essentially different from making inferences using p-values based on a distance measure to the location of the null effect. For an estimate of the mixing distribution $\hat{H}_{\hat{\pi}_0}$ and a pre-specified level α, the critical region can be found using a constrained maximisation problem, i.e.,

$$\text{Maximise} \qquad F(a; \hat{H}_{\hat{\pi}_0}, \sigma) + 1 - F(b; \hat{H}_{\hat{\pi}_0}, \sigma), \tag{7.3}$$

subject to

$$\begin{aligned} a - b &\leq 0, \\ \frac{\hat{\pi}_0[\Phi(a;0,\sigma) + 1 - \Phi(b;0,\sigma)]}{F(a;\hat{H}_{\hat{\pi}_0},\sigma) + 1 - F(b;\hat{H}_{\hat{\pi}_0},\sigma)} &\leq \alpha. \end{aligned}$$

As the number of hypotheses increases, the solution to optimisation problem (7.3) converges to the solution in (7.2) almost surely by the Helly-Bray and the continuous mapping theorem provided that $\hat{\pi}_0$ is a consistent estimator of π_0^*. In general, the objective function may contain multiple local maxima over $\mathbb{R}^2$, a global optimisation algorithm should better be used. A reliable one is the DIRECT method (Gablonsky and Kelley, 2001; Jones et al., 1993) which is a deterministic global search algorithm based on a systematic division of the search domain into hyper-rectangles. The optimisation problem only requires $\hat{\pi}_0$ and the restricted NPMLE $\hat{H}_{\hat{\pi}_0}$ and hence computing the solution pair (a, b) does not scale with the number of hypotheses which is computationally feasible for large-scale datasets. The left pane in Figure 7.1 shows the contour of the FDR for $\hat{H}_{0.963}$ using z-values of the prostate data mentioned in Section 3.3.2, the thick contour line is the set of pairs such that the FDR is exactly at the level $\alpha = 0.05$. The right pane in Figure 7.1 shows the contour of the expected proportion of rejections with the same thick contour line in the left pane superimposed. The red star indicates the global solution computed using the DIRECT method.

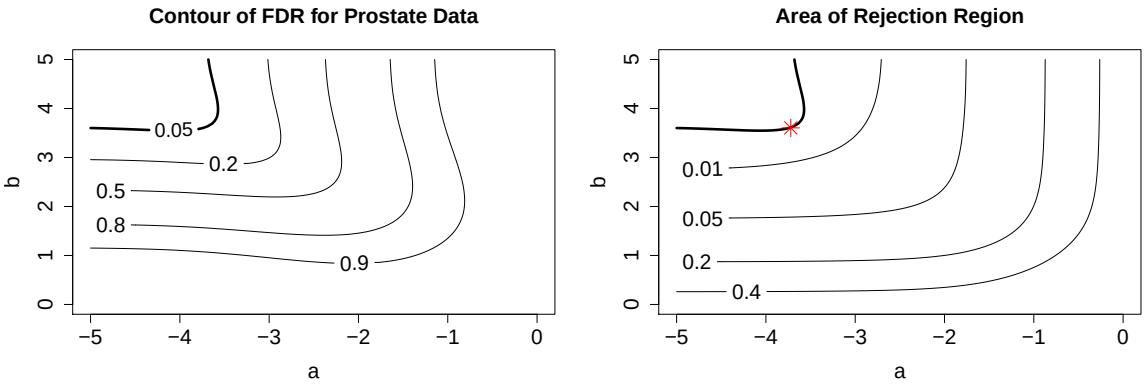

FIGURE 7.1: Contour plots of the FDR of a mixing distribution and the expected proportion of rejections.

Once the estimate of the critical region is obtained, all the hypotheses with the statistic falling inside the critical region is deemed as significant and the ones falling

outside the critical region are regarded as insignificant. Figure 7.2 shows the density estimate of the z-values for the prostate data with the restricted NPMLE $\hat{H}_{0.963}$. The red solid lines show the information about the estimated density and the dashed vertical lines are the estimated boundaries of the critical region under $\alpha = 0.05$. The pair (a, b) is estimated to be $(-3.721, 3.604)$ and 26 hypotheses are deemed as significant.

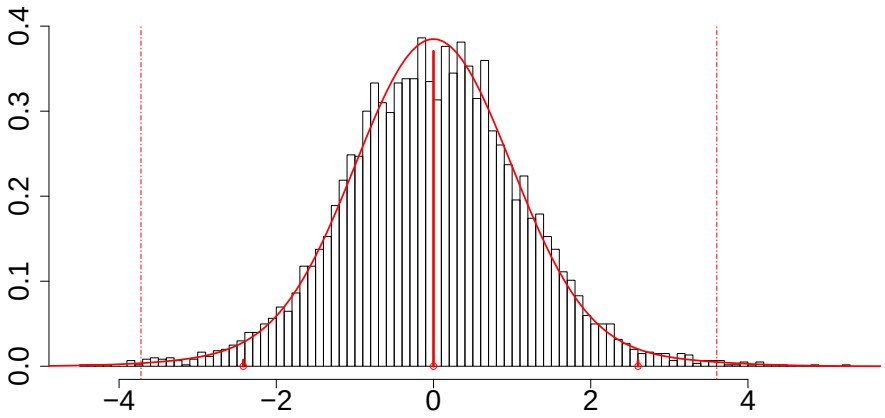

FIGURE 7.2: Plot of the estimated density of the z-values for prostate data using restricted NPMLE $\hat{H}_{0.963}$ with the estimated critical region.

7.4 Numerical Studies

The performance of this FDR-controlling procedure is examined using the breast cancer and the prostate dataset from Chapter 4. The existing methods to compare the performance with are the Benjamini-Hochberg (BH) procedure mentioned in Benjamini and Hochberg (1995) and the adaptive step-down procedure mentioned in Gavrilov et al. (2009). Let $p_{(1)} \leq p_{(2)} \leq \cdots \leq p_{(n)}$ be the ordered p-values, and the FDR is to be controlled at level α. Let k_{BH} be the largest i for which $p_{(i)} \leq i\alpha/n$, and the BH method rejects first k_{BH} ordered hypotheses. For the AS method, let k_{AS} be the largest i for which $p_{(j)} \leq j\alpha/(n+1-j(1-q))$ for all $j \leq i$, and the first k_{AS} ordered hypotheses are rejected. The implementations of these two methods are self-written. There is a function for the BH procedure in the R package `sgof` (Conde and Una Alvarez, 2020), but this implementation is computationally infeasible for large-scale computation. To the best of our knowledge, there is no implementation for the adaptive step-down procedure. The R package `mcp.project` available on R-forge implements various controlling

procedures, but the documentation for these functions is missing, and thus it is hard to determine which procedure was implemented in each function exactly. The DIRECT method which is implemented in the R package `nloptr` (Ypma et al., 2020) is used to perform the global optimisation in order to find the boundary of the critical region.

The steps of the examining process are as follows:

Step 1: For every synthetic dataset, the t- or z-values generated from the non-null parameters are deemed significant, and insignificant if otherwise.

Step 2: For our proposed method, given the mixing distribution $\hat{H}_{\hat{\pi}_0}$ estimated from the synthetic dataset with a particular choice of the component density and the threshold function, the critical region is computed using the procedure described above with $\alpha = 0.05$. The z- or t-values in the critical region are labelled as significant, and insignificant if otherwise. Should any existing method need p-values in the computation, the p-values are computed using z-values and z-values are transformed from t-values if needed. A classification table can be generated and then the FDR can be obtained. If there is no case classified as significant, the FDR for this synthetic dataset will be set to 0 by (7.1). This typically happens when π_0^* is close to 1 and the number of observations is relatively small.

Step 3: The mean, standard error (SE) and the mean squared error (MSE) of FDRs are calculated and reported for each case.

	$n = 750$	$n = 1500$	$n = 3000$
FDRs			
Mean	5.795	4.732	5.393
SE	0.580	0.359	0.272
MSE	0.338	0.129	0.075
Number of Rejections			
Mean	22	40	80
SE	1	1	2

TABLE 7.1: The mean×100%, standard error×100% and MSE×100% of FDRs from the 100 synthetic datasets generated by the breast cancer data using t-mixtures and $\pi_0 = \pi_0^*$.

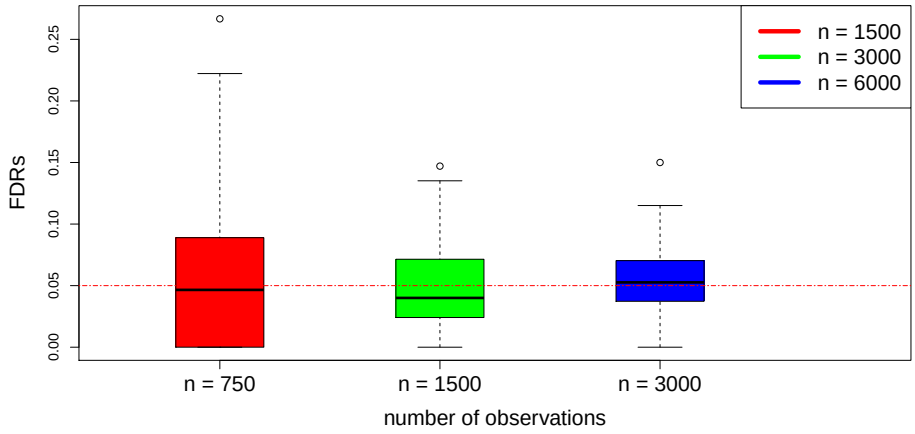

FIGURE 7.3: Boxplot of FDRs from the 100 synthetic datasets generated by the breast cancer data using t-mixtures and $\pi_0 = \pi_0^*$.

Table 7.1 shows the means and standard errors of the FDRs and the numbers of rejections, and Figure 7.3 shows the corresponding boxplot when the component density and π_0 are specified correctly. The means and standard errors of the number of rejections are rounded to the nearest integers. The estimated means of the FDRs are sufficiently close to the pre-specified level and the MSE of the FDRs is decreasing as the number of observations increases. The proposed procedure is able to make at least one rejection in every synthetic dataset for all choices of the component densities and threshold functions. The BH procedure makes no rejection in 6, 8 and 2 synthetic datasets for $n = 750, 1500$ and 3000 respectively while the adaptive step-down procedure makes no rejection in 17, 24 and 17 synthetic datasets for all n.

n	npnorm-AIC	npnorm-BIC	npt-AIC	npt-BIC	BH	AS
750	6.548(0.458)	6.257(0.524)	5.985(0.588)	5.680(0.660)	2.638(0.580)	2.212(0.609)
1500	5.958(0.338)	5.631(0.331)	4.605(0.368)	4.305(0.366)	3.431(0.642)	1.666(0.484)
3000	6.199(0.240)	5.828(0.231)	5.280(0.271)	5.136(0.281)	2.798(0.326)	1.992(0.326)

TABLE 7.2: The mean×100% (standard error×100%) of FDRs from the 100 synthetic datasets generated by the breast cancer data.

n	npnorm-AIC	npnorm-BIC	npt-AIC	npt-BIC	BH	AS
750	30(1)	25(1)	20(1)	17(1)	8(1)	5(1)
1500	59(2)	50(1)	39(2)	33(1)	13(1)	7(1)
3000	123(3)	107(2)	77(2)	69(2)	25(1)	17(2)

TABLE 7.3: The mean (standard error) of numbers of rejections from the 100 synthetic datasets generated by the breast cancer data.

Table 7.2 shows the means and standard errors of the FDRs of the synthetic datasets for the breast cancer data in Section 4.3, while Table 7.3 shows the means and standard errors of the number of rejections. The numbers in Table 7.3 are rounded to the nearest integers. The term "AS" represents the adaptive step-down procedure. The proportion of null effects in the proposed procedure is estimated using various threshold functions and component distributions. The means of the FDRs computed using normal components do not seem to converge to $\alpha = 0.05$, for all threshold functions. Since the null proportions are under-estimated using normal components, more hypotheses than expected are deemed significant and the FDRs are higher than the pre-specified level, for the reason given earlier. The estimated means of the FDRs computed using the t-distribution seem to converge quickly to $\alpha = 0.05$. For all the methods that have the estimated FDR smaller than α-level, the proposed procedure using t-components make the most rejections, around three times as those by the BH and the adaptive step-down procedure.

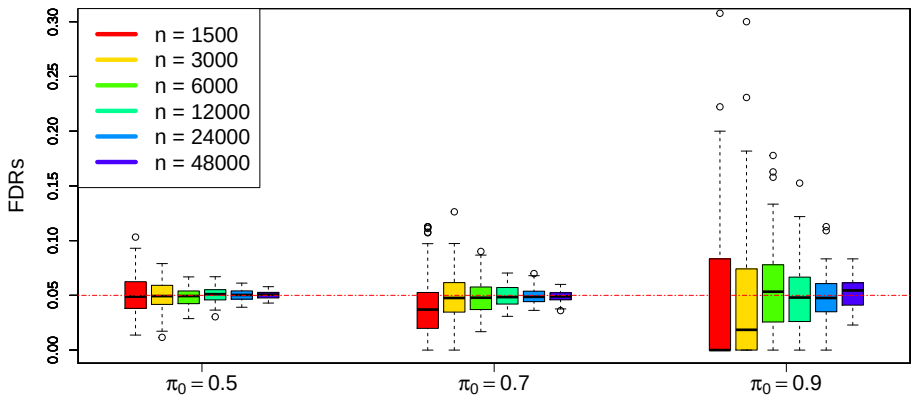

FIGURE 7.4: Boxplot of FDRs from the 100 synthetic datasets generated by the prostate data with normal density and $\pi_0 = \pi_0^*$.

Table 7.4 shows the means and standard errors of the FDRs using the synthetic datasets generated by the prostate data in Sections 4.2 and 6.7, and the means and standard errors of the number of rejections are shown in Table 7.5. The numbers in Table 7.5 are rounded to the nearest integers. The proportions of null effects for $n = 1500, 3000, 6000$ are estimated using the original data while the proportions of null effects for $n = 12000, 24000, 48000$ are carried out using the large-scale computation described in Chapter 6 with $h = 10^{-2}$. The results in the column with the name "npnorm-true" are the means and standard errors of the FDRs when $\pi_0 = \pi_0^*$, and the corresponding boxplot is shown in Figure 7.4. The results in this column are for references only. The estimated means of the FDRs are sufficiently close to the pre-specified level and the most rejections are made for all choices of π_0^*. The boxes for $\pi_0^* = 0.9$ are different from those for $\pi_0^* = 0.5$ or 0.7 but still have a convergent trend to the pre-specified level. Since the null proportion is close to 1, the controlling procedure hardly detects any significant observation. A smaller number of significant hypotheses or even no significant hypothesis is identified, which leads to a bigger standard error of the estimated mean of the FDRs. In this simulation study, the proposed controlling procedure makes no rejection in only one out of all the synthetic datasets under the right specification of the model.

For the proposed procedure with the estimated null proportion as well as the existing methods, the standard errors are decreasing as the number of observations increases. Overall, the proposed procedure performs the best in term of the distance to the pre-specified level, followed by the adaptive step-down procedure. The means of the FDRs estimated from the BH and the adaptive step-down procedure are significantly smaller than $\alpha = 0.05$ for $\pi_0^* = 0.5$ and 0.7. The mean of the FDRs estimated using the proposed procedure shows a clear increasing trend toward the pre-specified level for both of the AIC and BIC threshold functions while the trends for the BH and the adaptive step-down procedure seem to be constant. The proposed procedure also makes more rejections than the BH and the adaptive step-down procedure given the FDR is controlled at the same level. The proposed procedure has a similar performance compared with the BH and the adaptive step-down procedure when $\pi_0^* = 0.9$. However, as shown Table 7.5 that the proposed procedure always make at least the same number of rejections as the BH and adaptive step-down procedures, which indicates that the proposed procedure likely has a higher efficiency. As the number of hypotheses gets large, the estimated means of the FDRs are all within two standard errors of the pre-specified level, despite the difficulty in identifying significant hypotheses.

n	npnorm-AIC	npnorm-BIC	BH	AS	npnorm-true
			$\pi_0^* = 0.5$		
1500	3.629(0.157)	3.313(0.157)	2.285(0.141)	2.503(0.147)	5.051(0.179)
3000	3.724(0.126)	3.482(0.118)	2.417(0.111)	2.645(0.111)	5.018(0.139)
6000	3.853(0.082)	3.513(0.076)	2.448(0.072)	2.687(0.074)	4.845(0.080)
12000	4.176(0.074)	3.773(0.064)	2.535(0.056)	2.729(0.058)	5.077(0.070)
24000	4.210(0.053)	3.802(0.039)	2.550(0.035)	2.725(0.035)	5.029(0.049)
48000	4.367(0.060)	3.858(0.036)	2.507(0.034)	2.672(0.035)	5.012(0.035)
			$\pi_0^* = 0.7$		
1500	3.715(0.290)	3.498(0.279)	2.841(0.280)	2.877(0.281)	4.176(0.289)
3000	4.099(0.216)	4.019(0.222)	3.415(0.224)	3.495(0.226)	4.877(0.215)
6000	4.313(0.145)	4.179(0.153)	3.527(0.155)	3.628(0.154)	4.846(0.141)
12000	4.508(0.087)	4.234(0.087)	3.544(0.090)	3.651(0.092)	4.964(0.096)
24000	4.485(0.076)	4.266(0.074)	3.441(0.075)	3.555(0.074)	4.947(0.074)
48000	4.557(0.056)	4.275(0.048)	3.457(0.049)	3.560(0.050)	4.908(0.050)
			$\pi_0^* = 0.9$		
1500	3.513(0.783)	3.666(0.807)	3.634(0.626)	2.645(0.533)	4.766(1.017)
3000	4.861(0.597)	4.945(0.637)	4.944(0.645)	4.394(0.633)	4.940(0.619)
6000	5.244(0.397)	5.162(0.424)	4.970(0.412)	4.821(0.420)	5.315(0.404)
12000	4.758(0.286)	4.758(0.287)	4.458(0.281)	4.359(0.286)	5.094(0.292)
24000	4.614(0.198)	4.605(0.198)	4.502(0.203)	4.480(0.201)	4.779(0.200)
48000	5.012(0.133)	4.929(0.139)	4.695(0.130)	4.683(0.131)	5.197(0.128)

TABLE 7.4: The mean×100% (standard error×100%) of FDRs from the 100 synthetic datasets generated by the prostate data.

n	npnorm-AIC	npnorm-BIC	BH	AS	npnorm-true
			$\pi_0^* = 0.5$		
1500	133(1)	125(1)	101(1)	106(2)	157(1)
3000	271(2)	258(2)	207(2)	216(2)	316(2)
6000	550(4)	520(3)	407(3)	428(3)	628(3)
12000	1122(7)	1059(4)	819(4)	857(4)	1257(4)
24000	2288(12)	2151(6)	1645(5)	1723(6)	2524(5)
48000	4676(29)	4369(12)	3272(9)	3432(9)	5042(9)
			$\pi_0^* = 0.7$		
1500	52(1)	49(1)	45(1)	45(1)	58(1)
3000	103(2)	98(1)	90(1)	91(1)	114(2)
6000	209(2)	202(2)	180(2)	184(2)	231(2)
12000	428(3)	410(3)	360(3)	368(3)	465(3)
24000	854(5)	820(4)	716(4)	730(4)	930(5)
48000	1751(9)	1666(6)	1441(6)	1472(6)	1863(6)
			$\pi_0^* = 0.9$		
1500	8(0)	7(0)	8(0)	7(0)	8(0)
3000	14(1)	14(1)	14(1)	13(1)	15(1)
6000	30(1)	30(1)	29(1)	29(1)	31(1)
12000	57(1)	56(1)	55(1)	54(1)	58(1)
24000	115(2)	115(2)	111(2)	111(2)	120(2)
48000	232(2)	229(2)	222(2)	221(3)	241(2)

TABLE 7.5: The mean (standard error) of numbers of rejections made from 100 synthetic datasets generated by the prostate data.

7.5 Conclusion

In this chapter, a controlling procedure is proposed as an extension to the non-parametric density estimation and the estimation of null proportions, and is illustrated using procedures that control the FDR. The proposed controlling procedure relies on the estimated mixing distribution of all the statistics to compute the critical region, which does not scale with the number of hypotheses. In the case of controlling the FDR, the proposed procedure is able to control the quantity at the pre-specified level asymptotically, due to the strong convergences of the ECDF and the restricted NPMLE for all $\hat{\pi}_0 \leq \pi_0^*$ under the independence assumption of the hypotheses. The proportion of the null effects in this chapter is estimated through various threshold functions. One can also use the estimators from other methods such as Jin and Cai (2007). If such estimator $\hat{\pi}_0$ is consistent, then the density estimator using the restricted NPMLE based on this

estimator of the null proportion is also consistent and the critical region obtained also has the FDR controlled exactly at the pre-specified level asymptotically. It is possible to relax the independence assumption on the statistics. As shown later in Chapter 8 if the observations are weakly correlated, then strong convergences of the ECDF and the restricted NPMLE also hold under some additional conditions, and hence the proposed controlling procedure can be potentially applied to a wider range of hypothesis testing problems.

Chapter 8

Covariance Matrix Estimation

8.1 Introduction

In this chapter, a new estimator is proposed to estimate the correlation matrix using non-parametric empirical Bayes method. Our approach is slightly different from existing methods. We focus on estimating each element in the correlation matrix instead of estimating the correlation matrix as a whole. In this case, the problem is turned into a problem of estimating the means, and hence a non-parametric empirical Bayes method can be applied to obtain an estimate of every single element.

The rest of this chapter is organised as follows. A literature review on controlling procedures is given in Section 8.2. Formulation of the covariance matrix estimation is shown in Section 8.3 and the framework of our proposed estimation approach is shown in Section 8.4. Sections 8.5 and 8.6 show two ways of estimating correlation coefficients and Sections 8.7, 8.8 and 8.9 give details about sparse estimation, positive-definiteness of the estimator and correlation among the sample correlation coefficients. The theoretical work is presented in Section 8.10.

8.2 Literature Review

Analysis of high-dimensional data often poses challenges so new statistical methods and theories need to be established (Donoho, 2000). In particular, some classical multivariate statistical methods cannot handle high-dimensional data. Pourahmadi (2013) discussed that the number of observations and the number of variables cause opposite effects on statistical results. In general, the aim of multivariate analysis is to make statistical inference about the dependence structures among variables. Increase in the number of observations has the effect of improving the precision and certainty

of the inference while increase in the number of variables has the opposite effect of reducing the precision and certainty. Therefore, the estimation of large covariance and precision (inverse covariance) matrices underlies fundamental problems in modern multivariate analysis.

Two major obstacles in modelling covariance matrices or their inverses are high dimensionality and positive definiteness. The usual estimator of the covariance matrix, the sample covariance matrix, is known to be unstable and ill-conditioned in high-dimensional settings. The aggregation of a massive amount of estimation errors can lead to considerable adverse impacts on the estimation accuracy, and hence there may be a large deviation in modelling and prediction. People started to adapt methods in other fields such as variable selection or hypothesis testing on the sample covariance matrix to obtain a stable and well-conditioned estimator. However, positive definiteness cannot be guaranteed by some methods alone and the estimators often get adjusted to be positive-definite using principal component analysis, Cholesky decomposition or Gaussian graphical models.

A brief review of the estimation of covariance and precision matrices is given in this section. There are various references that give comprehensive overviews of the estimations and theoretical aspects of covariance and precision matrices. Cai and Zhou (2012) considered an optimal estimation of a covariance matrix as well as its inverse over several commonly-used parameter spaces under the matrix L_1-norm. Pourahmadi (2013) brought together and presented some of the most important ideas and methods in high-dimensional covariance estimation. Fan et al. (2016) gave an overview of the estimation of large covariance and precision matrices. Cai et al. (2016) reviewed recent developments on the optimal estimation of structured high-dimensional covariance and precision matrices to date. Buydens et al. (2017) selectively reviewed several modern estimators of covariance and precision matrices using eigenvalue-shrinkage, ridge-type, and structured estimation. The most recent reviews on estimation of covariance matrices can be found in Lam (2020) and the references therein.

8.2.1 Estimation of Covariance Matrices

One of the key and common assumptions in estimating a covariance matrix is that the true covariance matrix or the correlation matrix is sparse, in the sense that the majority of the covariances/correlations are zero. Thresholding has been a very popular non-parametric method in regularising covariance or correlation matrices. Thresholding was

first used in variable selection and then brought into estimation of covariance or correlation matrices. Thresholding methods carry no computational burden besides cross-validation, and are hence suitable in high-dimensional settings (Pourahmadi, 2013). Tibshirani (1996) proposed the soft thresholding rule (LASSO),

$$s_\lambda(z) = \text{sgn}(z)[|z| - \lambda]^+,$$

where $\text{sgn}(\cdot)$ is the sign function. That is, given a thresholding parameter λ, the rule returns an output 0 if the absolute value of the input parameter is less than λ, and thresholds the input value towards 0 by λ otherwise. Antoniadis (1997) and Fan (1997) proposed the hard thresholding rule,

$$s_\lambda(z) = z\mathbb{1}(|z| > \lambda).$$

The values remain unchanged if the absolute value of z is greater than λ. Antoniadis and Fan (2001) showed that the thresholding rules are also the solutions of penalised likelihood problems with different penalty functions. LASSO corresponds to the penalised least-squares estimators with the L_1 penalty $p_\lambda(\theta) = |\theta|$ and the hard-thresholding rule corresponds to $p_\lambda(\theta) = \lambda^2 - (|\theta| - \lambda)^2\mathbb{1}(|\theta| < \lambda)$. Fan and Li (2001) proposed the Smoothly Clipped Absolute Deviation Penalty (SCAD),

$$s_{\lambda,a}(z) = \begin{cases} \text{sgn}(z)[|z| - \lambda]^+, & |z| \leq 2\lambda \\ \{(a-1)z - \text{sgn}(z)a\lambda\}/(a-2), & 2\lambda < |z| \leq a\lambda \\ z, & |z| > a\lambda \end{cases}$$

for some $a > 2$, and the penalty function is

$$p_\lambda(\theta) = \mathbb{1}(\theta \leq \lambda) + \frac{[a\lambda - \theta]^+}{(a-1)\lambda}\mathbb{1}(\theta > \lambda).$$

In practice, the choice of $a = 3.7$ works similarly to that chosen by the generalised cross-validation method. The rates of convergence of the penalised likelihood estimators were established. Furthermore, with a proper choice of regularization parameters, it was shown that the SCAD performs as well as the oracle procedure in variable selection problems. Zou (2006) proposed the adaptive LASSO, which gives each input value a corresponding weight. Large coefficients get penalised less. One of the weight functions proposed is dependent on the solution of the ordinary least squares. The

penalty function is $p_\lambda(\theta) = \lambda w(z)|\theta|$, where $w(z)$ is taken to be $C|z|^{-\eta}$, $\eta \geq 0$ and the thresholding rule is

$$s_\lambda(z) = \text{sgn}(z)[|z| - \lambda^{n+1}|z|^{-\eta}]^+.$$

Based on the banding operator defined in Bickel and Levina (2008b), Bickel and Levina (2008a) defined the uniformity class of covariance matrices invariant under permutations by

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

for $0 \leq q < 1$. When $q = 0$, we have

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

which corresponds to the class of sparse matrices. Rothman et al. (2009) proposed a set of conditions to unify the thresholding rules

$$\begin{cases} |s_\lambda(z)| \leq |z|, & \\ s_\lambda(z) = 0, & \text{for } |z| \leq \lambda \\ |s_\lambda(z) - z| \leq \lambda, & \end{cases}$$

and it was proved that the LASSO, the hard thresholding, the SCAD and the adaptive LASSO are all following this set of conditions. Note that the estimators of a covariance matrix using the thresholding rules are not guaranteed to be positive-definite, but instead it was shown that these estimators converge to positive-definite limits with probability tending to 1. The sparsistency (Lam and Fan, 2009) of the estimator was also shown under the conditions for the class of sparse matrices mentioned in Bickel and Levina (2008a). Zhang (2010) introduced another nearly unbiased thresholding rule under the minimax concave penalty (MCP). The thresholding rule is

$$s_\lambda(z) = \begin{cases} \lambda|z| - \frac{z^2}{2\gamma}, & |z| \leq \gamma\lambda \\ \frac{1}{2}\gamma\lambda^2, & |z| > \gamma\lambda \end{cases}$$

for $\gamma > 1$, and the corresponding penalty function is

$$p_\lambda(\theta) = \lambda \int_0^\theta \left[1 - \frac{x}{\gamma\lambda}\right]^+ dx.$$

Similar to the SCAD, the MCP penalty is made to be differentiable since the derivative of a penalty is the contribution made by the penalty to the penalised estimating equations (Karush-Kuhn-Tucker condition). Cai and Liu (2011) proposed an adaptive thresholding procedure in which the thresholding rule is kept the same but the value of λ is multiplied by the standard deviation of individual entries. Under the appropriate conditions, the rate of convergence and support recovery were shown. The comparison to the original thresholding rule was also mentioned. Cai and Zhou (2012) investigated the minimax rates of convergence for covariance and precision matrices in general. A thresholding estimator which has the same rate of convergence as the lower bound was constructed. Cai and Yuan (2012) considered an adaptive estimation problem and constructed a data-driven block thresholding estimator based on the procedure in Cai and Liu (2011). The estimator was shown to have the optimal rate over a wide range of bandable covariance matrices. It was proved in Fan et al. (2016) that performing an adaptive thresholding on a covariance matrix is equivalent to performing the generalised thresholding to the corresponding correlation matrix. Fang et al. (2016) studied the number of folds in cross-validation and constructed an optimal λ given data in thresholding the sample covariance matrix. Liyanage et al. (2019) conducted a theoretical study for the optimal threshold level, and obtained its analytical expression by minimising the Frobenius risk of the adaptive thresholding estimator. A consistent estimator based on this expression was proposed for the optimal threshold level, which is easy to implement in practice and efficient in computation.

Other structures or methods of estimating a covariance matrix are also considered in the literature. Chiu et al. (1996) proposed an estimator using the matrix logarithm of a covariance matrix to ensure the positive definiteness. Ledoit and Wolf (2004) introduced an estimator that is both well-conditioned and more accurate than the sample covariance matrix asymptotically. The estimator is distribution-free and has a simple explicit formula that is easy to compute and interpret. It is the asymptotically optimal convex linear combination between the sample covariance matrix and the identity matrix. The optimality is meant with respect to the quadratic loss function, asymptotically as the number of observations and the number of variables go to infinity together. Chaudhuri et al. (2007) presented an algorithm called the iterative conditional fitting,

for computing the maximum likelihood estimate of the constrained covariance matrix, under the assumptions that the observations are multivariate-normally distributed and that the zeros in the covariance matrix estimator are pre-specified. This algorithm has guaranteed convergence properties. Dropping the assumption of multivariate normality, a way to estimate a covariance matrix using an empirical likelihood approach was also presented. In a series of two papers Pourahmadi (2007, 2011), estimators using the (modified) Cholesky decomposition were proposed to ensure symmetry. Bickel and Levina (2008b) introduced the class of bandable matrices and proposed a banding estimator. Wu and Pourahmadi (2009) introduced and studied a banding estimator based on sample auto-covariance matrices. Cai et al. (2010) studied further the class of bandable matrices and proposed a tapering estimator. In addition, the optimal rates of convergence for estimating a covariance matrix under both the operator and the Frobenius norm were established using the minimax upper bound and the risk properties based on this special class of tapering estimators. Chen et al. (2010) made an improvement to the estimator proposed in Ledoit and Wolf (2004) by conditioning on a sufficient statistic. This new estimator dominates the Ledoit-Wolf estimator for Gaussian variables under the mean-squared error by the Rao-Blackwell theorem. Bien and Tibshirani (2011) proposed penalising the likelihood with a LASSO penalty on the entries of the covariance matrix, and used a majorisation-minimisation algorithm in Lange et al. (2000) and a generalised gradient descent to solve this concave-convex procedure. Won et al. (2013) proposed a maximum likelihood approach, with the direct goal of obtaining a well-conditioned estimator. No sparsity assumption on either the covariance matrix or its inverse was imposed for this estimator. Cai et al. (2013) extended the problem of the optimal estimation of Toeplitz matrices from a fixed sample size to both the number of observations and the number of variables tending to infinity. Fan et al. (2013) introduced the principal orthogonal complement thresholding method "POET" to explore the approximated factor structure with sparsity. Mathematical insights were provided when the factor analysis is approximately the same as the principal component analysis for high dimensional data. The rates of convergence of the sparse residual covariance matrix and the conditional sparse covariance matrix were studied under various norms. It was shown that the effect of estimating the unknown factors vanishes as the dimensionality increases. Cui et al. (2016) used a convex optimisation formulation for estimating a sparse correlation matrix as opposed to the corresponding covariance matrix. An efficient accelerated proximal gradient algorithm was developed,

and it was shown that this method has a faster rate of convergence. An adaptive version of this approach was also discussed. Motivated by the deep connection between the Stein's unbiased risk estimator (SURE) and the AIC in regression models, Li and Zou (2016) proposed a family of generalized SUREs indexed by c (SURE$_c$) for covariance matrix estimation, where c is a constant. Rajaratnam and Vincenzi (2016) gave a theoretical aspect of the Stein's covariance estimator. Tsukuma (2016) investigated not only the non-singular case but also the singular case of estimated covariance matrices based on Stein's minimax estimator and orthogonally invariant estimator. Ledoit and Wolf (2018) introduced a new method for deriving covariance matrix estimators that are decision-theoretically optimal within a class of non-linear shrinkage estimators as an extension to the estimator in Ledoit and Wolf (2004).

In most circumstances, it is more desirable to have a positive-definite covariance matrix for inference. Under finite samples, some estimators proposed above may not be positive-definite, and hence post-processing is needed. Higham (2002) proposed an alternating projection method to compute the nearest correlation matrix under the Frobenius norm. This projection method converges at best linearly. Qi and Sun (2006) proposed to achieve the positive definiteness under finite samples through solving a constrained optimisation problem. Since this method is of Newton type, it has a quadratic convergence rate. Fan et al. (2013) considered the property that the minimum eigenvalue of a matrix is a function of the thresholding parameter and gave an estimator that is positive-definite. Goulart et al. (2020) quantified the accuracy of projecting a symmetric or a Hermitian matrix to the positive semi-definite cone using the standard eigenvalue perturbation theory.

8.2.2 Estimation of Precision Matrices

Estimating precision matrices is another fundamental problem in multivariate analysis. Precision matrices capture the conditional correlations among these variables and are closely related to the undirected graph under a Gaussian model. Similar to covariance matrix estimation, researchers have been making the assumption that the target precision matrix is sparse. Estimators of a sparse precision matrix can be obtained using penalised likelihood methods with different penalties such as the LASSO, the hard thresholding, the SCAD, the adaptive LASSO and the MCP. Solving the penalised likelihood problem involves convex optimisations which can be solved by the majorisation-minimisation algorithm (Lange et al., 2000). In addition to the penalised

likelihood and thresholding methods used in sparse covariance matrix estimation, several methods for estimating the precision matrices are introduced here. Meinshausen and Buhlmann (2006) used a neighbourhood selection with the LASSO, which is a computationally attractive alternative to the standard covariance selection for high-dimensional sparse graphs. The consistency was shown for the proposed neighbourhood selection scheme. The computation was done by a sequence of Lasso regressions to separately estimate the columns of the precision matrix. Yuan and Lin (2007) proposed a LASSO-type estimator and a non-negative garrotte-type estimator. The computation was done effectively by taking advantage of the efficient maxdet algorithm developed in convex optimisation. A BIC-type criterion was then used for selecting the tuning parameter for the penalised likelihood methods. Friedman et al. (2008) took advantage of the LASSO and the cyclic coordinate descent, and proposed the graphical LASSO to solve for the precision matrix as well as the covariance matrix efficiently. Banerjee et al. (2008) pointed out that the optimisation problem is convex, but the memory requirement and complexity of the existing interior point methods are prohibitive for problems with more than tens of nodes. As a result, two new algorithms were proposed for solving problems with at least a thousand nodes in the Gaussian case. The first algorithm used block coordinate descent, and hence can be interpreted as a recursive L_1-norm penalised regression. The second algorithm, based on Nesterov's first-order method, yields a complexity estimate with a better dependence on the problem size than the existing interior point methods. Yuan (2010) took advantage of the connection between multivariate linear regression and the entries of the inverse covariance matrix, and proposed an estimation procedure that can effectively exploit the sparsity. The proposed method can be computed using linear programming, and therefore has the potential to be used in high-dimensional problems. The regression coefficients were computed using the Dantzig selector in Candes and Tao (2007). Cai et al. (2011) proposed the CLIME estimator which uses a constrained L_1 minimization and the second-stage refitting procedure considered by Candes and Tao (2007) to further improve the numerical performance. However, the CLIME estimator may not be symmetrical and positive-definite, and hence another procedure was used to ensure symmetry and positive-definiteness. Sun and Zhang (2013) proposed applying the scaled LASSO method in Sun and Zhang (2012) in a column-by-column fashion to estimate the precision matrix. This proposed algorithm provides a map from the space of non-negative-definite matrices to the space of symmetric matrices. For each column,

the penalty level is determined by data using convex minimisation instead of cross-validation. Liu et al. (2014) developed a procedure that applies the sparse column-wise inverse operator to a modified sample covariance matrix. The rate of convergence was established under various norms. Under the Frobenius norm, theoretical guarantees were proved using cross-validation to pick tuning parameters. Based on the CLIME estimator in Cai et al. (2011), Pang et al. (2014) proposed an efficient parametric simplex algorithm to implement the CLIME estimator and Cai et al. (2016) proposed the ACLIME estimator which is a data-driven estimator based on the adaptive constrained L_1 minimisation. Fan et al. (2016) pointed out the conditions for column-by-column iterative methods. It requires all columns of the precision matrix to be sparse. Liu and Wang (2017) proposed an asymptotically tuning-free and non-asymptotically tuning insensitive approach for estimating high-dimensional Gaussian graphical models. The precision matrix was estimated in a column-by-column fashion. For each column, a sparse regression was performed. Dalal and Rajaratnam (2017) used an alternating minimization algorithm, which effectively works as a proximal gradient algorithm on the dual problem, to solve the sparse inverse covariance matrix. Zhang and Liu (2019) paid attention to L_0-norm estimation. It was pointed out that the L_0-norm is the most natural function to measure the sparsity. However, when the sample covariance matrix is singular, the problem using the L_0-norm has no optimal solution. As a result, a new method was proposed to estimate the sparse inverse covariance matrix using the L_0-norm and the Tikhonov regularization simultaneously. Deng and Kang (2020) proposed a sparse precision matrix estimator using the modified Cholesky decomposition. The consistent property of the proposed estimator was established under some weak regularity conditions. Jiang and Wang (2020) proposed an algorithm for precision matrix estimation via penalized quadratic loss functions. When the dimension is high and the sample size is low, the computational complexity of our algorithm is linear in both the sample size and the number of parameters.

8.3 The problem

Without loss of generality, let $\boldsymbol{x}_i$ with $i = 1, \ldots, n$ denote independent and identically distributed random vectors sampled from a p-variate Gaussian distribution with $\mathbb{E}[\boldsymbol{x}_i] = \mathbf{0}$ and $\mathbb{E}[\boldsymbol{x}_i \boldsymbol{x}_i^T] = \boldsymbol{\Sigma}$, then $\boldsymbol{M} = \boldsymbol{X}\boldsymbol{X}^T$ where $\boldsymbol{X} = (\boldsymbol{x}_1, \cdots, \boldsymbol{x}_n)$ follows a

Wishart distribution with

$$\mathbb{E}[m_{ij}] = n\sigma_{ij},$$

where σ_{ij} is the corresponding true covariance. The maximum likelihood estimator of the covariance matrix can therefore be written as

$$\boldsymbol{S} = \frac{\boldsymbol{M}}{n}.$$

The goal is to obtain an estimator of the covariance matrix by estimating the correlation matrix based on the sample correlation matrix using the decomposition

$$\boldsymbol{S} = \boldsymbol{DRD},$$

where $\mathbf{D}$ is a diagonal matrix with $d_{ii} = \sqrt{\sigma_{ii}}$, the estimated standard deviation for the i-th variable, and $\mathbf{R}$ is the sample correlation matrix. We do not attempt to make inferences about the variances, although it is also possible to apply the proposed method in this chapter on variances with slight modifications; see Section 8.11 for a brief discussion. Hotelling (1953) gave an approximation of the mean and the variance of r_{ij}, i.e.,

$$\begin{aligned}
\mathbb{E}[r_{ij}] &= \rho_{ij} - \frac{\rho_{ij}(1-\rho_{ij}^2)}{2n} + \mathcal{O}\left(\frac{1}{n^2}\right), \\
\mathrm{Var}(r_{ij}) &= \frac{(1-\rho_{ij}^2)^2}{n} + \frac{23(1-\rho_{ij}^2)^2\rho_{ij}^2}{4n^2} + \mathcal{O}\left(\frac{1}{n^3}\right).
\end{aligned}$$

We can also use the Taylor series expansion to write the covariance between two correlation coefficients r_{ij} and r_{nk},

$$\begin{aligned}
\mathrm{Cov}(r_{ij}, r_{nk}) =& \frac{1}{2n}\left[(\rho_{in} - \rho_{ij}\rho_{jn})(\rho_{jk} - \rho_{jn}\rho_{nk}) + (\rho_{in} - \rho_{ik}\rho_{nk})(\rho_{jk} - \rho_{ij}\rho_{ik}) + \right. \\
& \left. (\rho_{ik} - \rho_{ij}\rho_{jk})(\rho_{jn} - \rho_{jk}\rho_{nk}) + (\rho_i k - \rho_{in}\rho_{nk})(\rho_{jn} - \rho_{ij}\rho_{in})\right] + \mathcal{O}\left(\frac{1}{n^2}\right).
\end{aligned} \tag{8.1}$$

From (8.1) we know that there is a correlation between two sample correlation coefficients r_{ij} and r_{nk} and this correlation is not zero in general. Dealing with correlations

among the sample correlation coefficients has its complications and most of the methods in the literature do not have the capability of dealing with this issue either. Similar to what has been done in the literature, we do not take into account the correlations among sample correlation coefficients. A brief discussion about the correlations among the sample correlation coefficients will be given in Section 8.9.

8.4 Estimation using Empirical Bayes

In this section, a new estimator is proposed to estimate the correlation matrix using empirical Bayes. Herbert Robbins introduced empirical Bayes into statistics (Robbins, 1951, 1956, 1964, 1983) and Stein (1956) investigated the admissibility of the usual maximum likelihood estimator and proposed the earlier version of Stein's estimator. The result was improved by James and Stein (1961), which is known as the James-Stein estimator. Several authors have discussed the relationship between empirical Bayes and the James-Stein estimator; see Efron and Morris (1973) and Petrone et al. (2014). The usage of empirical Bayes in estimating covariance matrices can be found in Dey and Srinivasan (1985), Kubokawa et al. (1992), Ledoit and Wolf (2018), Li and Zou (2016), Rajaratnam and Vincenzi (2016), and Tsukuma (2016). Most of the aforementioned methods are based on the James-Stein estimator, which is essentially a parametric empirical Bayes method. Our approach is slightly different from these methods. We focus on estimating each element in the correlation matrix instead of estimating the correlation matrix as a whole. In this case, the problem is turned into a problem of estimating the means, and hence a non-parametric empirical Bayes method can be applied to obtain an estimate of every single element. The usage of empirical Bayes in estimating the means can be found in Brown (2008), Brown and Greenshtein (2009), Jiang and Zhang (2009), Lu and Stephens (2019), Weinstein et al. (2018), and Zhang (1997).

Let $\boldsymbol{r}$ be the vector containing r_{ij} and $\boldsymbol{\rho}$ be the vector containing ρ_{ij} for all $i, j \in \mathbb{N}$ such that $1 \leq i < j \leq p$. According to the formulation in Efron (2010b), an unknown parameter vector $\boldsymbol{\rho}$ with prior distribution function H gives rise to an observable data vector $\boldsymbol{r}$ according to the univariate density $\phi(r_k; \rho_k)$ under the independence assumption. The Bayes rule is a formula for the conditional density of ρ_k having observed r_k (its posterior distribution), i.e.,

$$g(\rho_k|r_k) = \frac{\phi(r_k; \rho_k)h(\rho_k)}{f(r_k)}, \tag{8.2}$$

where

$$f(r_k) = \int \phi(r_k; \rho_k) dH(\rho_k) \tag{8.3}$$

is the marginal distribution of r_k and h is the corresponding probability density (or mass) function of H. Having observed $\boldsymbol{r}$, we want to estimate $\boldsymbol{\rho}$ with some estimator $\hat{\boldsymbol{\rho}}$. We use the total squared loss to measure the error of estimating $\boldsymbol{\rho}$ by $\hat{\boldsymbol{\rho}}$,

$$L(\boldsymbol{\rho}, \hat{\boldsymbol{\rho}}) = ||\boldsymbol{\rho} - \hat{\boldsymbol{\rho}}||_2^2.$$

The corresponding risk function is the expected value of $L(\boldsymbol{\rho}, \hat{\boldsymbol{\rho}})$ for a given $\boldsymbol{\rho}$, i.e.,

$$R(\boldsymbol{\rho}) = \mathbb{E}_{\hat{\boldsymbol{\rho}}}[L(\boldsymbol{\rho}, \hat{\boldsymbol{\rho}})] = \mathbb{E}_{\hat{\boldsymbol{\rho}}}[||\hat{\boldsymbol{\rho}} - \boldsymbol{\rho}||_2^2] = \mathbb{E}_{\boldsymbol{r}}[||\hat{\boldsymbol{\rho}}(\boldsymbol{r}) - \boldsymbol{\rho}||_2^2].$$

Assuming the simple and the symmetric rule described in Greenshtein and Ritov (2009) together with (8.2) and (8.3), the estimator is the posterior mean

$$\hat{\rho}_k(r_k) = \int \rho g(\rho | r_k) d\rho = \frac{\int \rho \phi(r_k; \rho) dH(\rho)}{\int \phi(r_k; \rho) dH(\rho)}, \tag{8.4}$$

and this estimator is asymptotically optimal for any H in the class of simple, symmetric estimators. The estimator of the correlation matrix can then be constructed by replacing the sample correlation coefficients by the corresponding posterior mean. Note that the class of simple and symmetric estimators contains the maximum likelihood estimator of $\boldsymbol{\rho}$, since the maximum likelihood estimator of $\boldsymbol{\rho}$ is $\hat{\boldsymbol{\rho}}(\boldsymbol{r}) = \boldsymbol{r}$, an element-wise identity operator. In this case, we have $\hat{\rho}_i(\boldsymbol{r}) = r_i$ and if there exists i, j such that $r_i = r_j$, then $\hat{\rho}_i(r_i) = r_i = r_j = \hat{\rho}_j(r_j)$.

In order to find the estimate of a correlation coefficient from the corresponding sample correlation coefficient, say r_k, estimates of $H(\rho)$ and $\phi(r_k; \rho)$ are needed. In practice, the component distribution $\phi(r_k; \rho)$ is assumed to be known while $H(\rho)$ is unknown and must be estimated from the data. The following two sections describe two different ways of estimating $H(\rho)$ by assuming two different known component distributions $\phi(r_k; \rho)$.

8.5 Approximating the Distribution of Sample Correlation Coefficients

In this section, we discuss approximating the distribution of sample correlation coefficients using a normal distribution and illustrate how to compute $\hat{\rho}_k(r_k)$ based on this approximated component distribution.

The distribution of $r_{ij} \in (-1, 1)$ can be well approximated by a one-parameter distribution with mean ρ_{ij} and variance $(1-\rho_{ij}^2)^2/n$ when n is large, since

$$\mathbb{E}[(r_{ij}-\rho_{ij})^3] = -\frac{15\rho_{ij}(1-\rho_{ij}^2)^3}{2n^2} + \mathcal{O}\left(\frac{1}{n^3}\right),$$
$$\mathbb{E}[(r_{ij}-\rho_{ij})^4] = \frac{3(1-\rho_{ij}^2)^4}{n^2} + \frac{(1-\rho_{ij}^2)^4(237\rho_{ij}^2-12)}{2n^3} + \mathcal{O}\left(\frac{1}{n^4}\right).$$

The approximations of the third and the fourth moment shown above are also due to Hotelling (1953). The distribution function becomes degenerated when r_{ij} is -1 or 1, in this case the distribution function is defined to be a step function. Thus, an approximate distribution function of a sample correlation coefficient thus the following piece-wise form,

$$\Phi(r;\rho) = \begin{cases} \mathbb{1}(r \geq \rho), & \rho = -1, 1 \\ \Phi\left(r; \rho, \frac{(1-\rho^2)^2}{n}\right), & \rho \in (-1, 1). \end{cases} \tag{8.5}$$

Since the distribution of each element in the sample correlation matrix can be approximated by this one-parameter distribution, the posterior density is a one-parameter location mixture. The algorithm described in Wang (2007) can then be used to compute the estimate of the prior distribution by defining the component density to be the one-parameter distribution above. The MLEs of the weights at -1 and 1 are the sample proportions respectively.

The consistency of the NPMLE can be obtained. The sample proportion of -1 and 1 in the observations are consistent estimators of the true weights at -1 and 1 respectively. For the NPMLE in the interval $(-1, 1)$, since all the sample correlation coefficients lie in the interval $(-1, 1)$, the support points in the mixing distribution also lie in the same interval by Example 3.1, the space of all the mixing distributions with support points in the interval $(-1, 1)$ is a subset of the space mentioned in

Assumption 3.2. For a fixed n, the density of this one-parameter distribution satisfies the requirement in Kiefer and Wolfowitz (1956), and hence the strong consistency of the NPMLE holds, when the true distribution is a mixture of this approximating one-parameter family.

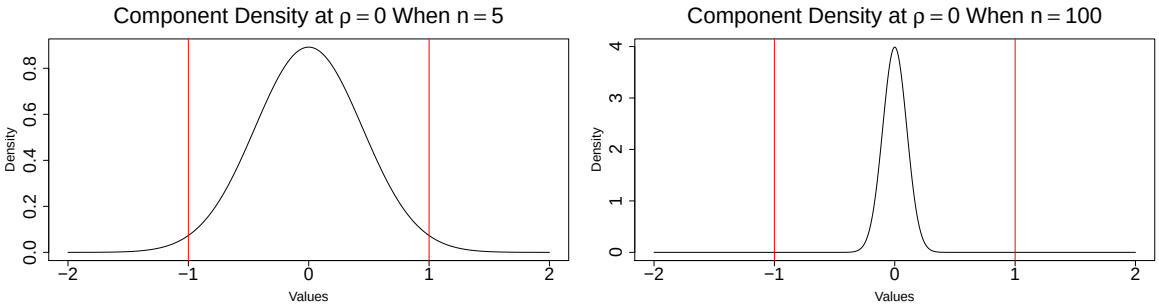

FIGURE 8.1: Plots of the approximated density of sample correlation coefficient at $\rho = 0$ for $n = 5, 100$.

Figure 8.1 shows the approximated density of the sample correlation coefficients at $\rho = 0$ when $n = 5$ (left) and $n = 100$ (right). The probability outside the interval $[-1, 1]$ is small when $n = 100$ while the probability outside the interval $[-1, 1]$ is significant when $n = 5$. This suggests that this normal approximation is not suitable for estimation and inference when n is small. Figure 8.2 shows the log-probability outside the interval $[-1, 1]$ plotted against n at $\rho = 0$. The red horizontal line represents the logarithm of the machine error. It is suggested that n should be no smaller than 70 for this normal approximation at $\rho = 0$ to work well numerically.

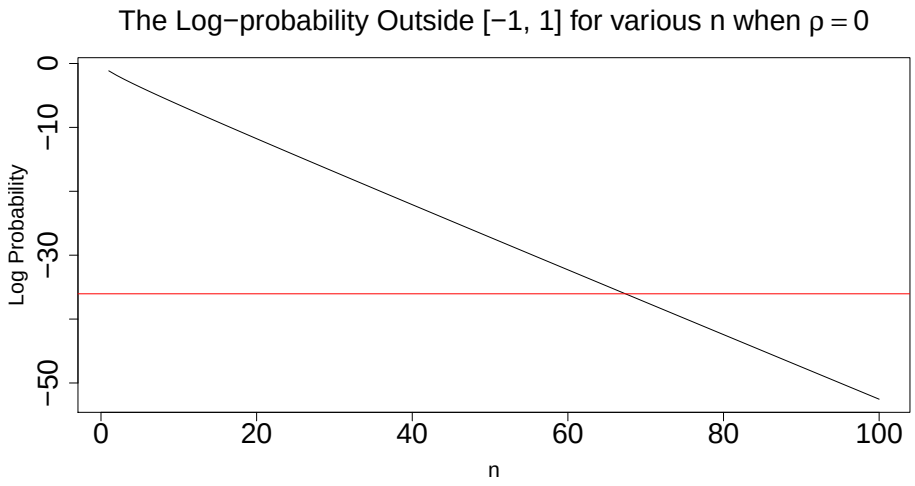

FIGURE 8.2: Plot of the log-probability outside $[-1, 1]$ against n when $\rho = 0$.

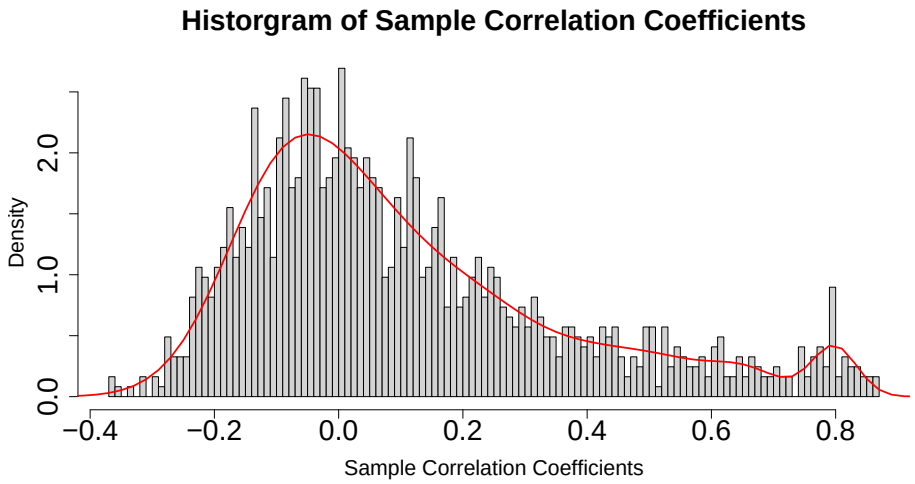

FIGURE 8.3: Histogram of sample correlation coefficients and the estimated density.

Figure 8.3 shows the histogram of sample correlation coefficients and its density estimate using the unrestricted NPMLE when the true correlation matrix is of size 50×50 and has entries $\rho_{ij} = 0.8^{|i-j|}\mathbb{1}(|i-j| \leq 10)$. 100 observations are used to estimate the sample correlation matrix. Since both $\phi(r;\rho)$ and $\rho\phi(r;\rho)$ are bounded functions, by the Helly-Bray theorem, the continuous mapping theorem and the strong

consistency of $\hat{H}$, the strong consistency of $\hat{\rho}$ holds, i.e., for any fixed r,

$$\hat{\rho}(r) = \frac{\int \rho\phi(r;\rho)d\hat{H}(\rho)}{\int \phi(r;\rho)d\hat{H}(\rho)} \to \frac{\int \rho\phi(r;\rho)dH(\rho)}{\int \phi(r;\rho)dH(\rho)},$$

almost surely.

Computational Notes

If one wants to use the constrained Newton method described in Wang (2007) to compute the NPMLE, a new family needs to be defined which requires the log of the component density as a function of ρ as well as the first derivative of the log of the component density with respect to ρ,

$$\log\phi(r;\rho) = -\frac{n(r-\rho)^2}{2(1-\rho^2)^2} - \frac{1}{2}\log\left(\frac{2\pi(1-\rho^2)^2}{n}\right)$$
$$\frac{\partial\log\phi(r;\rho)}{\partial\rho} = \frac{n(r-\rho)(\rho^2-2r\rho+1)}{(1-\rho^2)^3} + \frac{2\rho}{1-\rho^2}.$$

If one wants to implement the algorithm directly, the derivatives

$$\frac{\partial d(\rho;G_s)}{\partial\rho} = \sum_{i=1}^{n} \frac{\frac{\log\phi(r;\rho)}{\partial\rho}\phi(r;\rho)}{f(r;G_s)}$$
$$\frac{\partial^2 d(\rho;G_s)}{\partial\rho^2} = \sum_{i=1}^{n} \frac{\frac{\partial^2\log\phi(r;\rho)}{\partial\rho^2}\phi(r;\rho) + \left(\frac{\partial\log\phi(r;\rho)}{\partial\rho}\right)^2\phi(r;\rho)}{f(r;G_s)}$$

can be used to compute the local maxima. The formulations for computing the unrestricted NPMDEs can be obtained in a similar manner.

8.6 The Distribution of Transformed Sample Correlation Coefficients

An alternative way is to use the Fisher's z-transformation to obtain an estimate of r_{ij}. The Fisher transformation is an approximate variance-stabilizing transformation for the correlation coefficient between two random variables which jointly follow a bivariate normal distribution. Fisher (1915) proved that if $\boldsymbol{x}_i$ with $i = 1,\ldots,n$ are

independent and identically distributed, then

$$z_{ij} = \frac{1}{2}\log\left(\frac{1+r_{ij}}{1-r_{ij}}\right) = \operatorname{arctanh}(r_{ij})$$

is approximately normally distributed with mean $v_{ij} = \operatorname{arctanh}(\rho_{ij})$ and variance $1/(n-3)$. Here, $\operatorname{arctanh}(\cdot)$ is the inverse hyperbolic tangent function. Similarly, the distribution function of a transformed sample correlation coefficient can be written in the piece-wise form

$$\Phi(z;v) = \begin{cases} \mathbb{1}(z \geq v), & v = -\infty, \infty \\ \Phi\left(z; v, \frac{1}{n-3}\right), & v \in (-\infty, \infty). \end{cases}$$

Since distribution of each transformed element in the sample correlation matrix can be approximated by the normal distributions with different means, the posterior distribution follows a normal mixture with a common scale parameter. The algorithm described in Wang (2007) can also be used to compute the estimate of the prior distribution when the component distribution is normal with variance $1/(n-3)$. The goal is still to minimise the risk function

$$R(\boldsymbol{\rho}) = \mathbb{E}_{\hat{\boldsymbol{\rho}}}[||\hat{\boldsymbol{\rho}} - \boldsymbol{\rho}||_2^2] = \mathbb{E}_{\boldsymbol{z}}[||\hat{\boldsymbol{\rho}}(\boldsymbol{z}) - \boldsymbol{\rho}(v)||_2^2],$$

and it can be shown that the estimator now becomes

$$\hat{\rho}_k(z_k) = \int \tanh(v) g(v|z_k) dv = \frac{\int \tanh(v)\phi(z_k;v)dH(v)}{\int \phi(z_k;v)dH(v)}, \tag{8.6}$$

where $\tanh(\cdot)$ is the hyperbolic tangent function. Figure 8.4 shows the histogram of transformed sample correlation coefficients and the estimated density using the unrestricted NPMLE. The observations are transformed from the observations used in Figure 8.3. Figure 8.5 shows the maps from sample correlation coefficients to the corresponding estimated correlation coefficients through the Fisher transformation. Sample correlation coefficients are shown at the top of the figure while the bottom shows estimated correlation coefficients. The density shown in the bottom of the figure is the true density. Points in the middle are sample correlation coefficients under the Fisher transformation and the density is estimated using the unrestricted NPMLE. It is shown that estimating the mixing distribution is truly a deconvolution process.

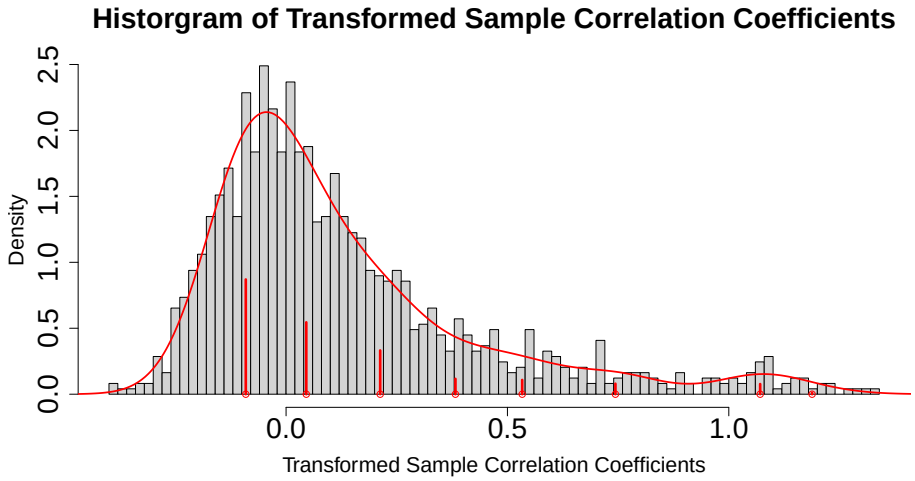

FIGURE 8.4: Histogram of transformed sample correlation coefficients and its estimated density.

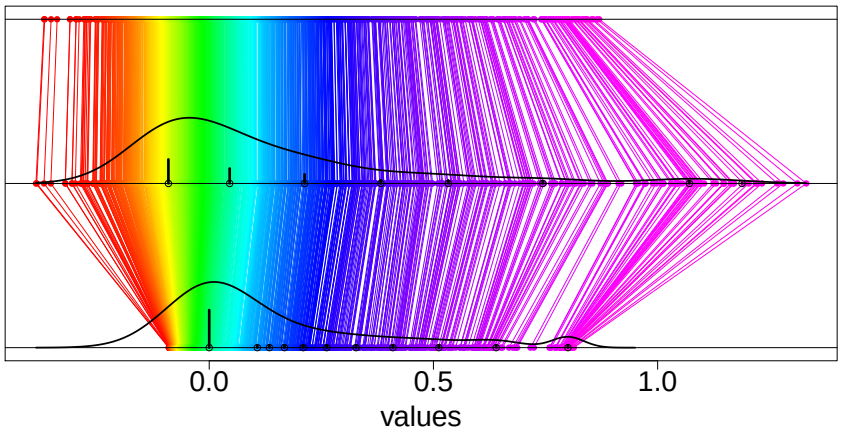

FIGURE 8.5: Plot of the maps from sample correlation coefficients to the corresponding estimated correlation coefficients.

The consistency of the NPMLE can also be shown for a fixed $n > 3$, the MLEs of the weights at $-\infty$ and ∞ are the sample proportions and they are known to be consistent. Since the standard deviation is a function of the fixed n, normal distributions with strictly positive standard deviation satisfies the requirement of the component density in Kiefer and Wolfowitz (1956). The strong consistency holds if $\boldsymbol{v}$ satisfies Assumption 3.2. Since both $\phi(z; v, 1/(n-3))$ and $\tanh(v)\phi(z; v, 1/(n-3))$ are bounded functions, by the Helly-Bray theorem, the continuous mapping theorem and the strong

consistency of $\hat{H}$, the strong consistency of $\hat{\rho}$ holds, i.e., for any fixed r,

$$\hat{\rho}(z) = \frac{\int \tanh(v)\phi(z;v)d\hat{H}(v)}{\int \phi(z;v)d\hat{H}(v)} \to \frac{\int \tanh(v)\phi(z;v)dH(v)}{\int \phi(z;v)dH(v)},$$

almost surely.

Computational Notes

The non-parametric maximum likelihood estimate of the mixing distribution of the Fisher-transformed sample correlation coefficients can be found using the function `cnm` in the R package `nspmix` after the construction of the class of normal mixtures using the function `npnorm` with the structural parameter (standard deviation) $1/\sqrt{n-3}$. Alternatively, the algorithms in Sections 3.5.1, 5.2 and 5.3 can be used to find the unrestricted NPMLE and NPMDEs with $\boldsymbol{\pi} = 0$ and $\sigma = 1/\sqrt{n-3}$.

8.7 Estimating Sparse Correlation Matrices

As mentioned in Section 8.2, numerous methods in the literature were proposed assuming that the true correlation matrix is sparse. The sparsity means that most entries in the covariance matrix are zeros, which is similar to the null effect in multiple hypothesis testing. In this section, the same strategy is used to estimate the correlation matrix under the sparsity assumption. We adapt the definition of the uniformity class of correlation matrices invariant under permutations in Bickel and Levina (2008a),

$$\mathcal{U}(q, c_0(p)) = \left\{ \boldsymbol{P} : \sum_{j=1}^{p} |\rho_{ij}|^q \leq c_0(p), \text{for all } i \right\},$$

with $q = 0$ to represent the class of sparse correlation matrices, i.e.,

$$\mathcal{U}(0, c_0(p)) = \left\{ \boldsymbol{P} : \sum_{j=1}^{p} \mathbb{1}(\rho_{ij} \neq 0) \leq c_0(p), \text{for all } i \right\}.$$

We borrow the term "null" and "non-null" from multiple hypothesis testing for the alternative representations of being zero and non-zero respectively. According to the notation for the two-group model in page 18 of Efron (2010b), the true correlation

coefficients are either null (0) or non-null (not 0) with prior probability

$$\pi_0 = \mathbb{P}(\text{null}), \qquad 1 - \pi_0 = \mathbb{P}(\text{non-null}) \tag{8.7}$$

with any sample correlation coefficient r having either the null density $\phi(r; 0)$ or the non-null density $f_A(r)$. Taking the perspective of empirical Bayes, let G be the true distribution of the non-null, which can be understood in the asymptotic sense as

$$G(\rho) = \lim_{n\to\infty} \frac{\sum_k \mathbb{1}(\rho \geq \rho_k)\mathbb{1}(\rho_k \neq 0)}{\sum_{i=1}^n \mathbb{1}(\rho_k \neq 0)},$$

and the non-null density can thus be written as

$$f_A(r) = \int \phi(r; \rho) dG(\rho).$$

As a result,

$$H(\rho) = \pi_0 \mathbb{1}(\rho \geq 0) + (1 - \pi_0)G(\rho).$$

The formulation in this section is almost identical to the formulation in Chapters 3 and 5, we can compute the restricted NPMLE or NPMDEs as well as estimating π_0 using the likelihood-based or distance-based methods. The estimates of the mixing distribution also has a meaningful interpretation. Computing the restricted NPMLE or NPMDEs is equivalent to estimating the distribution of the (transformed) sample correlation coefficients given a sparsity level in the covariance matrix, while computing $\hat{\pi}_0$ is the same as estimating the level of sparsity in the covariance matrix. The consistency of the NPMLE can also be established without major modifications to the proofs in Sections 3.4, 5.2.2 and 5.3.2.

Finding the proportion of zeroes is also possible. The formulation for the Fisher-transformed sample correlation coefficients is identical to that in the multiple hypothesis testing problem. When finding the proportion of zeros, the threshold function can be of the order smaller than or equal to the logarithm of the number of sample correlation coefficients in addition to the threshold functions mentioned in Jiang and Zhang (2016). However, it may not be possible to find the estimated proportion of zeros on the scale of sample correlation coefficients, the standard deviation is a function of the location parameter instead of being constant, and hence the GLRT in Jiang and Zhang

(2016) may not hold. As a result, it is recommended to transform the sample correlation coefficients before computing the non-parametric estimates or the proportion of zeros in practice. Figure 8.6 shows the maps from sample correlation coefficients to the corresponding estimated correlation coefficients through the Fisher transformation, using the same set of observations as in Figure 8.5. Sample correlation coefficients are shown at the top of the figure while the bottom shows estimated correlation coefficients. The density shown in the bottom of the figure is the true density. Points in the middle are sample correlation coefficients under the Fisher transformation and the density is estimated using the restricted NPMLE given the estimated level of sparsity using the BIC threshold function.

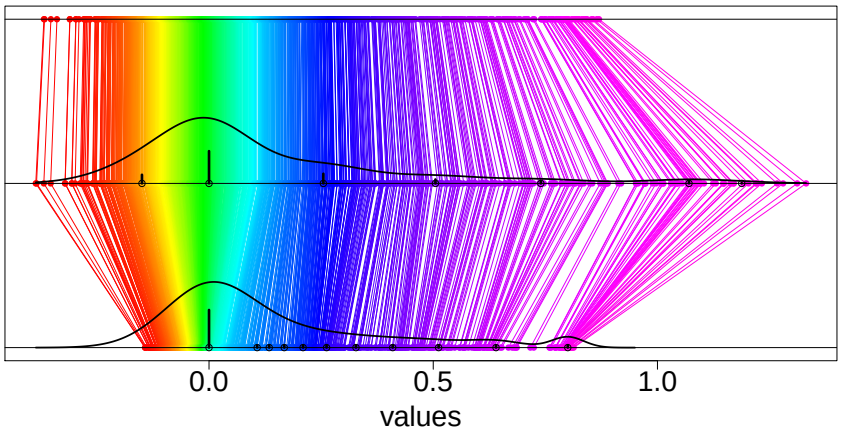

FIGURE 8.6: Plot of the maps from sample correlation coefficients to the corresponding estimated correlation coefficients.

Once the restricted NPMLE is obtained, the estimates of the correlation coefficients can be obtained using (8.4) or (8.6). It is noted that even if we fix a support point at 0 with a positive probability, the estimates computed by the posterior mean can not be 0 exactly since the type II error has been taken into consideration. The proposed estimator can never have zero entries unless the proportion of zeros is fixed at or estimated to be 1.

Computational Notes

When implementing the algorithm directly to estimate the unrestricted NPMLE of the sample correlation coefficients, set $k = 1$, $\boldsymbol{\pi} = \pi_0$, $\boldsymbol{\beta} = 0$. The equations

$$\frac{\partial d(\rho, G_s)}{\partial \rho} = \left(1 - \sum_{j=1}^{k} \pi_j\right) \sum_{i=1}^{n} \frac{\frac{\log \phi(r;\rho)}{\partial \rho} \phi(r;\rho)}{f(r; G_s, \boldsymbol{\pi}, \boldsymbol{\beta})},$$

$$\frac{\partial^2 d(\rho, G_s)}{\partial \rho^2} = \left(1 - \sum_{j=1}^{k} \pi_j\right) \sum_{i=1}^{n} \frac{\frac{\partial^2 \log \phi(r;\rho)}{\partial \rho^2} \phi(r;\rho) + \left(\frac{\partial \log \phi(r;\rho)}{\partial \rho}\right)^2 \phi(r;\rho)}{f(r; G_s, \boldsymbol{\pi}, \boldsymbol{\beta})},$$

where $\phi(r;\rho)$ is the corresponding density function of (8.5) can be used to compute the local maxima. The NPMDEs can be computed in a similar manner. The algorithms in Sections 3.5, 5.2 and 5.3 can be used to calculate the restricted NPMLE or NPMDEs of Fisher-transformed sample correlation coefficients by setting $k = 1$, $\boldsymbol{\pi} = \pi_0$, $\boldsymbol{\beta} = 0$ and $\sigma = 1/(n-3)$.

8.8 Positive Definiteness of the Estimator

The proposed method focuses on estimating each element in a correlation matrix and the element-wise estimation does not in general take into account of the overall structure hence the estimate of the covariance matrix may not be positive-definite and an additional step is needed. Under the assumption of the simple and symmetrical rule when deriving the estimator using the empirical Bayes, the resulting correlation matrix is always symmetric and the Newton method in Qi and Sun (2006) can be used to find the nearest correlation matrix under Frobenius norm.

8.9 Correlations among Sample Correlation Coefficients

In previous sections, the estimator of the correlation matrix is obtained by not taking into account the correlations among the sample correlation coefficients, which is generally not a true statement but used anyway by many in the literature. In this section, we present some of the findings when taking into consideration the correlations

among sample correlation coefficients. The results will be presented using the Fisher-transformed variables since the density function $\phi(z; v)$ is a regular normal distribution. Proofs of the statements are shown in Section 8.10.

We start the discussion with the result shown in Azriel and Schwartzman (2015). The set of standard normal random variables $\boldsymbol{Z}$ of length p is called *weakly correlated* if

$$\lim_{p\to\infty} \frac{1}{p^2} \sum_{k,l}^{p} |\text{cor}(Z_k, Z_l)| = 0.$$

Theorem 1 of Azriel and Schwartzman (2015) states that if $\boldsymbol{Z}$ is weakly correlated, then the ECDF $\hat{F}_p(z)$ computed using $\boldsymbol{z}$ converges to the standard normal CDF in L_2-norm uniformly, and does not converge if $\boldsymbol{Z}$ is not weakly correlated. This result can be easily extended to the set of normal random variables with different means. In other words, if the correlations among the normal random variables are weak, the ECDF is still a consistent estimator. An immediate result is that any non-parametric distance-based estimator with normal components that minimises the distance to the ECDF is also consistent. It is also worth investigating whether the NPMLE is consistent and the conditions when the consistency holds. If we restrict Γ to be the space of distribution functions with a finite second moment, then the consistency of the NPMLE can be obtained if the likelihood function for the independent variables is used as the likelihood function for weakly correlated variables. That is, the NPMLE $\hat{H}$ satisfying

$$\hat{H} = \arg\max_{H\in\bar{\Gamma}} \sum_{k=1}^{p} \log f(z_i; H, \sigma)$$

converges strongly to H^*. Note that the restriction on Γ is not a necessary condition, the consistency may hold for a wider range of distribution functions.

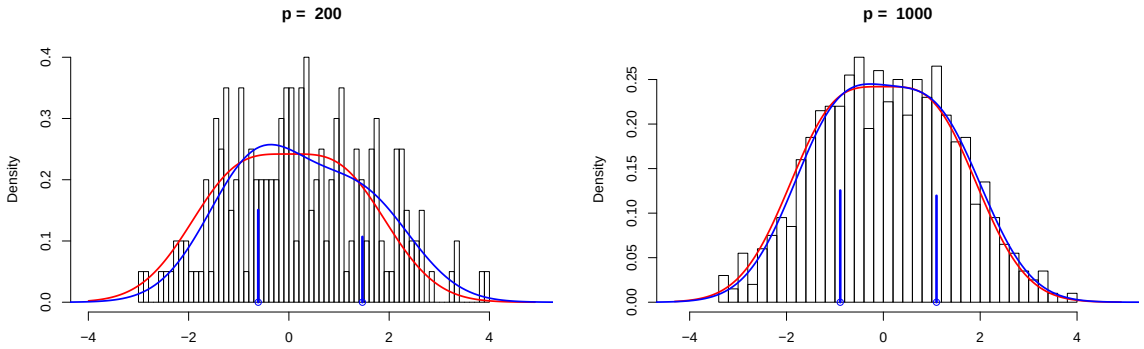

FIGURE 8.7: Histograms of weakly correlated normal random variables for various n and the densities are estimated using the NPMLE.

Figure 8.7 shows histograms of independently generated observations for $p = 200, 1000$ respectively when $\boldsymbol{Z}$ follows a p-variate normal distribution with mean $\boldsymbol{v} = (-1, 1, \cdots)^T$ and covariance matrix Σ with $\sigma_{ij} = \rho^{|i-j|}$ for $\rho = 0.5$. The red line stands for the true density while the blue lines stands for the density where NPMLE is used. The blue vertical lines show the locations and the weights of the component means. Figure 8.8 shows the graph of mean distances between H^* and 100 iid $\hat{H}$ plotted against p for p ranging from 0 to 1400. The distance used is the metric defined in Kiefer and Wolfowitz (1956). The black line represents the mean distance as a function of p and the red dotted lines show the estimated bounds of central 95%-percentile of the distribution of the distance. It can be seen that the mean distance and the width of the bound decreases as p increases.

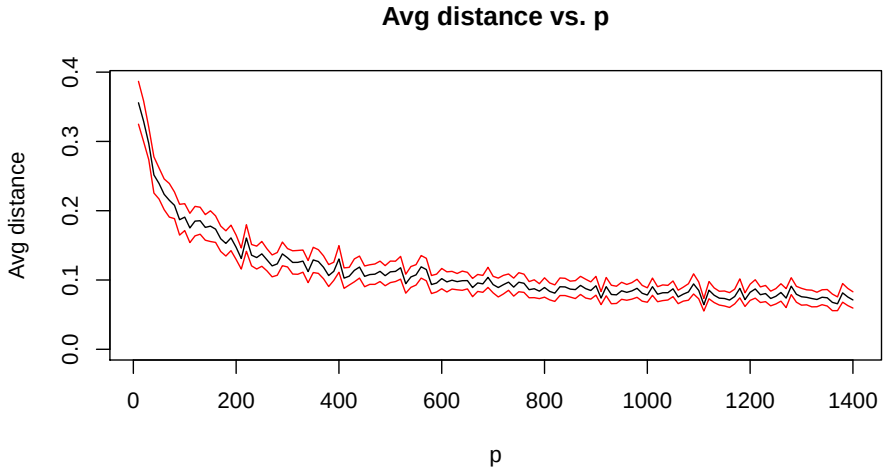

FIGURE 8.8: Plot of the mean distance between H^* and 100 iid $\hat{H}$ plotted against p and the estimated bounds of central 95%-percentile of the distribution of the distance.

In the context of this chapter, if the Fisher-transformed sample correlation coefficients are weakly correlated, then the NPMLE using the likelihood function for the independent variables is strongly consistent. The expression of the correlations among the Fisher-transformed sample correlation coefficients can be obtained from (8.1) and the delta theorem, i.e.,

$$\begin{aligned}\operatorname{cor}(z_{ij}, z_{nk}) = \frac{1}{2(1-\rho_{ij}^2)(1-\rho_{nk}^2)}[&(\rho_{in}-\rho_{ij}\rho_{jn})(\rho_{jk}-\rho_{jn}\rho_{nk}) \\ &+ (\rho_{in}-\rho_{ik}\rho_{nk})(\rho_{jk}-\rho_{ij}\rho_{ik}) + (\rho_{ik}-\rho_{ij}\rho_{jk})(\rho_{jn}-\rho_{jk}\rho_{nk}) \\ &+ (\rho_{ik}-\rho_{in}\rho_{nk})(\rho_{jn}-\rho_{ij}\rho_{in})] + \mathcal{O}\left(\frac{1}{n}\right).\end{aligned}$$

If $\boldsymbol{x}_i \sim \mathcal{N}_p(0, \boldsymbol{\Sigma})$ are weakly correlated, then the sample correlation coefficients under Fisher transformation are asymptotically weakly correlated as the number of observations tends to infinity. This includes all the elements in the set $\mathcal{U}(1, o(p))$ which contains the class of sparse matrices $\mathcal{U}(0, o(p))$. Since the NPMLE using the likelihood functions for the independent variables is consistent when the sample correlation coefficients are weakly correlated, the procedures proposed in Section 8.4 and 8.6 still give consistent estimates of the empirical Bayes estimator. Numerical studies in Chapter 9 show that the performance of the proposed method is competitive with other existing methods under the appropriate conditions.

8.10 Theory

The theoretical aspects of this chapter are given in this section. Proposition 8.1 shows the derivation of the empirical Bayes estimator of the sample correlation coefficients under the symmetrical and simple rule.

Proposition 8.1. *Under the assumption of the simple and symmetric rule, the action* $\hat{\rho}(r)$ *that minimises the risk* $\mathbb{E}_{\hat{\boldsymbol{\rho}}}[L(\boldsymbol{\rho}, \hat{\boldsymbol{\rho}})]$ *for all* $r \in \mathbb{R}$ *is*

$$\hat{\rho}(r) = \frac{\int \rho\phi(r;\rho)dH(\rho)}{\int \phi(r;\rho)dH(\rho)},$$

where $\phi(r;\rho)$ *is the corresponding piece-wise density in (8.5). This estimator is asymptotically optimal for any* $H(\theta)$ *in the class of simple, symmetric estimators.*

Proof. This proof only conerns with $\rho \in (-1, 1)$ and thus the notation $\phi(r;\rho)$ in this proof is the corresponding density for the case $\rho \in (-1, 1)$ in (8.5). The posterior mean for $\rho = -1, 1$ is -1 and 1 respectively. We also extend the notation $\phi(r;\rho)$ by writing $\phi(\boldsymbol{r}, \boldsymbol{\rho})$ to represent the corresponding multivariate distribution. We omit the covariance matrix in this notation since it is not of importance here.

$$\begin{aligned}
R(\boldsymbol{\rho}) &= \mathbb{E}_{\boldsymbol{r}}[||\hat{\boldsymbol{\rho}}(\boldsymbol{r}) - \boldsymbol{\rho}||_2^2] \\
&= \int \sum_k (\hat{\rho}_k(\boldsymbol{r}) - \mu_k)^2 \phi(\boldsymbol{r};\boldsymbol{\rho}) d\boldsymbol{r} \\
&= \sum_k \int (\hat{\rho}_k(\boldsymbol{r}) - \mu_k)^2 \phi(\boldsymbol{r};\boldsymbol{\rho}) d\boldsymbol{r} \qquad (8.8) \\
&= \sum_k \int (\hat{\rho}(r_k) - \rho_k)^2 \phi(\boldsymbol{r};\boldsymbol{\rho}) d\boldsymbol{r} \qquad (8.9) \\
&= \sum_k \int \int (\hat{\rho}(r_k) - \rho_k)^2 \phi\left(\boldsymbol{r}_{-k}; \boldsymbol{\rho}_{-k} + \frac{r_k - \rho_k}{\sigma_{kk}} \boldsymbol{\sigma}_{-k,k}\right) \phi(r_k;\rho_k) d(\boldsymbol{r}_{-k}) dr_k \\
&= \sum_k \int (\hat{\rho}(r_k) - \rho_k)^2 \left[\int \phi\left(\boldsymbol{r}_{-k}; \boldsymbol{\rho}_{-k} + \frac{r_k - \rho_k}{\sigma_{kk}} \boldsymbol{\sigma}_{-k,k}\right) d(\boldsymbol{r}_{-k})\right] \phi(r_i;\rho_k) dr_k \\
&= \sum_k \int (\hat{\rho}(r_k) - \rho_k)^2 \phi(r_k;\rho_k) dr_k \\
&= \sum_k \int (\hat{\rho}(r) - \rho_k)^2 \phi(r;\rho_i) dr \\
&= \int \sum_k [(\hat{\rho}(r_k) - \rho_k)^2 \phi(r_k;\rho_k)] dr_k
\end{aligned}$$

$$= \frac{p(p-1)}{2} \int \int (\hat{\rho}(r) - \rho)^2 f(r;\rho) dG(\rho) dr.$$

Obtaining (8.9) from (8.8) is due to the assumption of the simple and the symmetric rule in Greenshtein and Ritov (2009). The expected loss is minimised if for any observed sample $\boldsymbol{r}$, $\int (\hat{\rho}(r) - \rho)^2 f(r;\rho) dG(\rho)$ is minimal. By taking the derivative with respect to $\hat{\rho}(r)$ and set to 0, the action that minimises the expected loss is

$$\hat{\rho}(r) = \frac{\int \rho \phi(r;\rho) dH(\rho)}{\int \phi(r;\rho) dH(\rho)}.$$

□

It is quite surprising that under the simple and the symmetric rule, the posterior mean does not depend on the covariance/correlation matrix. The posterior mean still minimises the risk if the observations are weakly correlated, as mentioned in Section 8.9. The following proposition shows the strong consistency of the ECDF when the observations are weakly correlated.

Proposition 8.2. *Let $\boldsymbol{Z} \sim \mathcal{N}_p(\boldsymbol{v}, \boldsymbol{\Sigma})$ with $diag(\boldsymbol{\Sigma}) = \sigma^2 \mathbf{1}$ and H be the distribution of $\boldsymbol{v}$. If $\boldsymbol{Z}$ is weakly correlated, then*

$$\lim_{p\to\infty} \sup_{z\in\mathbb{R}} \mathbb{E}\left[\frac{1}{p}\sum_{k=1}^{p} \mathbb{1}(Z_k \geq z) - \int \Phi(z; v, \sigma) dH(v)\right] = 0.$$

Proof. Since $\mathbb{E}[(Z_k - v_k)^2] = \sigma^2 < 1$ and

$$\begin{aligned}
&\lim_{P\to\infty} \sum_{p=1}^{P} \frac{1}{p} \sqrt{\mathbb{E}\left[\frac{1}{p}\sum_{k=1}^{p}(Z_k - v_k)\right]^2} \\
&= \lim_{P\to\infty} \sum_{p=1}^{P} \frac{1}{p} \sqrt{\frac{1}{p^2}\sum_{k,l=1}^{p} \mathbb{E}[(Z_k - v_k)(Z_l - v_l)]} \\
&\leq \lim_{P\to\infty} \sum_{p=1}^{P} \frac{1}{p\sigma} \sqrt{\frac{1}{p^2}\sum_{k,l=1}^{p} |\mathrm{cor}(Z_k, Z_l)|} \\
&= \lim_{P\to\infty} \sum_{p=1}^{p} \frac{1}{p\sigma} \sqrt{\frac{1}{p^2}\sum_{k,l=1}^{p} |\mathrm{cor}(Z_k, Z_l)|} \\
&< \infty
\end{aligned}$$

by Theorem 6 in Lyons (1988), we have

$$\lim_{p\to\infty} \frac{1}{p}\sum_{k=1}^{p}(Z_k - v_k) = 0 \qquad \text{a.s.}$$

Since $|\mathbb{1}(Z_k \le z) - \Phi(z; v_k, \sigma^2)| \le 1$ for all $k \in \mathbb{N}$ and $z \in \mathbb{R}$, with the proof of Theorem 2 in Azriel and Schwartzman (2015), we have

$$\lim_{P\to\infty} \sum_{p=1}^{P} \frac{1}{p}\mathbb{E}\left\{\frac{1}{p}\sum_{k=1}^{p}[\mathbb{1}(Z_k \le z) - \Phi(z; v_i, \sigma)]\right\}^2 < \infty.$$

According to Theorem 1 in Lyons (1988),

$$\begin{aligned}
&\lim_{p\to\infty} \frac{1}{p}\sum_{k=1}^{p}[\mathbb{1}(Z_k \le z) - \Phi(z; v_i, \sigma)] \\
&= \lim_{p\to\infty} \left[\frac{1}{p}\sum_{k=1}^{p}\mathbb{1}(Z_k \le z) - \int \Phi(z; v, \sigma) dH(v)\right] \\
&= 0 \qquad \text{a.s.}
\end{aligned}$$

Using the same proof for Glivenko-Cantelli in van der Vaart (1998), we have that

$$\lim_{n\to\infty} \sup_z \left|\frac{1}{p}\sum_{k=1}^{p}\mathbb{1}(Z_k \le z) - \int \Phi(z; \mu, \sigma) dH(\mu)\right| = 0 \qquad \text{a.s.}$$

□

From the proposition above, the ECDF is a consistent estimator of the cumulative distribution of the true mixture. The following proposition extends this result and shows that any NPMDE is also a consistent estimator.

Proposition 8.3. *Let $d(\cdot,\cdot)$ be a proper metric between two distribution functions. Let $\boldsymbol{Z} \sim \mathcal{N}_p(\boldsymbol{v}, \boldsymbol{\Sigma})$ with $diag(\boldsymbol{\Sigma}) = \sigma^2\mathbf{1}$ and H^* be the distribution of $\boldsymbol{v}$. If $\boldsymbol{Z}$ is weakly correlated, then any NPMDE $\hat{H}$ in the compact space of Γ containing H^* equipped with the metric $d(\cdot,\cdot)$ is consistent.*

Proof. By triangle inequality,

$$d(F(z; \hat{H}, \sigma), F(z; H^*, \sigma)) \le d(F(z; \hat{H}, \sigma), \hat{F}_n(z)) + d(\hat{F}_n(z), F(z; H^*, \sigma)).$$

By Proposition 8.2, $\hat{F}_n(z)$ is a consistent estimator of $F(z; H^*, \sigma)$ and $d(\hat{F}_n(z), F(z; H^*, \sigma))$ goes to 0, almost surely. Since $\hat{H}$ is the solution to the optimisation problem,

$$\hat{H} = \arg\min_{H \in \Gamma} d(F(z; H, \sigma), \hat{F}_n(z)),$$

we have

$$d(F(z; \hat{H}, \sigma), F_n(z)) \leq d(F(z; H^*, \sigma), \hat{F}_n(z)) \to 0$$

and the sequence of estimators $\hat{H}$ converges weakly to H^*, almost surely. □

A similar result can also be shown for the NPMLE if the space Γ is restricted to the set of the distribution functions with a finite second moment. It is relatively easy to prove the strong consistency of the NPMLE in this case, the same statement may hold for a large class of distribution functions.

Proposition 8.4. *Let Γ be the space of distribution functions with a finite second moment. Let $\boldsymbol{Z} \sim \mathcal{N}_p(\boldsymbol{v}, \boldsymbol{\Sigma})$ with $diag(\boldsymbol{\Sigma}) = \sigma^2 \mathbf{1}$ and H^* be the distribution of $\boldsymbol{v}$. If $\boldsymbol{Z}$ is weakly correlated, then the NPMLE $\hat{H}$, which uses the likelihood function for the independent variables,*

$$\hat{H} = \arg\max_{H \in \bar{\Gamma}} \sum_{k=1}^{p} \log f(z_i; H, \sigma)$$

is strongly consistent.

Proof. The proof of consistency follows from the proof of the M-estimators described in van der Vaart (1998) with slight modifications. Let $\bar{\Gamma}$ be the compact space of Γ. It suffices to prove the following two statements

$$\sup_{H \in \bar{\Gamma}} |M_p(H) - M(H)| \to 0 \qquad \text{a.s.}, \tag{8.10}$$

$$\sup_{H: d(H, H^*) > \epsilon} M(H) < M(H^*), \tag{8.11}$$

for every $\epsilon > 0$, where

$$M_p(H) = \int \log f(z_i; H, \sigma) d\hat{F}_n(z) = \mathbb{E}[\log f(Y_n; H, \sigma)],$$
$$M(H) = \int \log f(z; H, \sigma) dF(z; H^*, \sigma) = \mathbb{E}[\log f(Y; H, \sigma)].$$

We will show (8.10) first. Take any event that $\hat{F}_n(z)$ converges weakly to $F(z; H^*, \sigma)$ and a fixed $H \in \bar{\Gamma}$. Since $\log f(y; H, \sigma)$ is not a bounded function, the Helly-Bray theorem can not be used here. It is observed that $\log f(y; H, \sigma)$ is a continuous function in y and by the continuous mapping theorem, we have

$$\log f(Y_n; H, \sigma) \xrightarrow{d} \log f(Y; H, \sigma).$$

It remains to show the integrability of $\log f(Y_n; H, \sigma)$ and $\log f(Y; H, \sigma)$ for all $H \in \bar{\Gamma}$. Up to an additive constant $1/2 \log 2\pi\sigma^2$, both of the random variables are non-positive random variables and

$$\begin{aligned}
&\mathbb{E}\left[\left|\log f(Y_n; H, \sigma) + \frac{1}{2}\log 2\pi\sigma^2\right|\right] \\
&= \mathbb{E}\left[-\log f(Y_n; H, \sigma) - \frac{1}{2}\log 2\pi\sigma^2\right] \\
&\leq \mathbb{E}\left[\int \left(\frac{(Y_n - \mu)^2}{2\sigma^2}\right) dH(\mu)\right] - \frac{1}{2}\log 2\pi\sigma^2 \\
&\leq \frac{1}{2\sigma^2}\mathbb{E}\left[Y_n^2 - 2Y_n \int \mu dH(\mu) + \int \mu^2 dH(\mu)\right] - \frac{1}{2}\log 2\pi\sigma^2. \qquad (8.12)
\end{aligned}$$

Showing the finiteness of (8.12) is equivalent to showing the finiteness of

$$\mathbb{E}[Y_n^2] = \frac{1}{p}\sum_{k=1}^{p} z_k^2. \qquad (8.13)$$

Since $\boldsymbol{Z}$ are weakly correlated,

$$\lim_{P\to\infty} \sum_{p=1}^{P} \frac{1}{p} \sqrt{\mathbb{E}\left[\frac{1}{p}\sum_{k=1}^{p}(Z_k - v_k)^2 - \sigma^2\right]} < \infty$$

and by Theorem 6 in Lyons (1988),

$$\lim_{p\to\infty} \frac{1}{p}\sum_{k=1}^{p}(Z_k - v_k)^2 = \sigma^2$$

almost surely. The integrability of $\log f(Y_n; H, \sigma)$ is thus shown, and integrability of $\log f(Y; H, \sigma)$ can be shown in a similar manner, and hence

$$|M_p(H) - M(H)| \to 0 \qquad \text{a.s.}$$

The next step is to show this happens uniformly in the compact space $\bar{\Gamma}$. Let

$$\mathcal{F} = \left\{ \log f(y; H, \sigma) : H \in \bar{\Gamma} \right\}.$$

According to Example 19.8 in van der Vaart (1998), $\mathcal{F}$ is a collection of measurable functions with an integrable envelope function

$$d(y) = \frac{y^2 - 2yM_1 + M^2}{2\sigma},$$

where

$$M_1 = \sup_{H \in \bar{\Gamma}} \int \mu dH(\mu),$$
$$M_2 = \sup_{H \in \bar{\Gamma}} \int \mu^2 dH(\mu) < \infty.$$

The map $H \to \log f(y; H, \sigma)$ is continuous for every y by the continuity assumption in Kiefer and Wolfowitz (1956), then the L_1-bracketing number of $\mathcal{F}$ is finite and $\mathcal{F}$ is Glivenko-Cantelli, and (8.10) can be obtained. Under the identifiability assumption in Kiefer and Wolfowitz (1956), (8.11) holds by Lemma 5.35 in van der Vaart (1998). By replacing the convergence in probability by almost sure convergence in the proof of Theorem 5.7 in van der Vaart (1998), the NPMLE $\hat{H}$ converges strongly to H^*. □

The following proposition shows that if $\boldsymbol{x}_i$ are weakly correlated, then the sample correlation coefficients are asymptotically weakly correlated, as the number of observations tends to infinity.

Proposition 8.5. *Let $\boldsymbol{X}$ be a random matrix with $\boldsymbol{x}_k \sim \mathcal{N}_p(\mathbf{0}, \Sigma)$ independently for all $k = 1, \cdots, n$. If*

$$\lim_{p \to \infty} \frac{1}{p^2} \sum_{i,j=1}^{p} |\rho_{ij}| = 0,$$

where ρ_{ij} is the underlying correlation between x_{ik} and x_{jk} for all k, then

$$\lim_{p \to \infty} \frac{1}{p^2(1-p)^2} \sum_{\substack{i,j,n,k=1 \\ i \neq j, n \neq k}}^{p} |cor(z_{ij}, z_{nk})| = \left| \mathcal{O}\left(\frac{1}{n}\right) \right|,$$

where $z_{ij} = \operatorname{arctanh}(r_{ij})$ with $\boldsymbol{R}$ the sample correlation matrix.

Before proving the the proposition, the following lemma is needed.

Lemma 8.1. *Let $0 < C < 1$ be a constant. If*

$$\lim_{p\to\infty} \frac{1}{p^2} \sum_{i,j=1}^{p} |\rho_{ij}| = 0,$$

then

$$\lim_{p\to\infty} \frac{1}{p^2} \sum_{i,j=1}^{p} \mathbb{1}(|\rho_{ij}| \geq C) = \mathcal{O}\left(\frac{1}{n}\right).$$

Proof. Since $|\rho_{ij}|$ are non-negative for all i, j and $C > 0$, by Markov's inequality,

$$\frac{1}{p^2} \sum_{i,j=1}^{p} \mathbb{1}(|\rho_{ij}| \geq C) \leq \frac{1}{p^2} \sum_{i,j=1}^{p} \frac{|\rho_{ij}|}{C} = \frac{1}{Cp^2} \sum_{i,j=1}^{p} |\rho_{ij}| \to 0.$$

□

We can now proceed with the proof with the help of Lemma 8.1.

Proof of Proposition 8.5. Using the multivariate delta theorem, the correlation between z_{ij} and z_{nk} with $i \neq j, n \neq k$ is

$$\begin{aligned}
\operatorname{cor}(z_{ij}, z_{nk}) = \frac{1}{2(1-\rho_{ij}^2)(1-\rho_{nk}^2)} &[(\rho_{in} - \rho_{ij}\rho_{jn})(\rho_{jk} - \rho_{jn}\rho_{nk}) \\
&+ (\rho_{in} - \rho_{ik}\rho_{nk})(\rho_{jk} - \rho_{ij}\rho_{ik}) + (\rho_{ik} - \rho_{ij}\rho_{jk})(\rho_{jn} - \rho_{jk}\rho_{nk}) \\
&+ (\rho_{ik} - \rho_{in}\rho_{nk})(\rho_{jn} - \rho_{ij}\rho_{in})] + \mathcal{O}\left(\frac{1}{n}\right).
\end{aligned}$$

Let $0 < C < 1$ be a constant. By Lemma 8.1,

$$\begin{aligned}
&\frac{1}{p^2(p-1)^2} \sum_{\substack{i,j,n,k=1 \\ i\neq j, n\neq k}}^{p} |\operatorname{Cor}(z_{ij}, z_{nk})| \\
= &\frac{1}{p^2(p-1)^2} \sum_{\substack{i,j,n,k=1 \\ i\neq j, n\neq k}}^{p} [|\operatorname{Cor}(z_{ij}, z_{nk})| \mathbb{1}(\rho_{ij} < C, \rho_{nk} < C)] \\
&+ \frac{1}{p^2(p-1)^2} \sum_{\substack{i,j,n,k=1 \\ i\neq j, n\neq k}}^{p} [|\operatorname{Cor}(z_{ij}, z_{nk})| \mathbb{1}(\rho_{ij} \geq C, \rho_{nk} < C)]
\end{aligned}$$

$$+\frac{1}{p^2(p-1)^2}\sum_{\substack{i,j,n,k=1\\ i\neq j,n\neq k}}^{p}[|\mathrm{Cor}(z_{ij},z_{nk})|\mathbb{1}(\rho_{ij}<C,\rho_{nk}\geq C)]$$
$$+\frac{1}{p^2(p-1)^2}\sum_{\substack{i,j,n,k=1\\ i\neq j,n\neq k}}^{p}[|\mathrm{Cor}(z_{ij},z_{nk})|\mathbb{1}(\rho_{ij}\geq C,\rho_{nk}\geq C)]$$

There are four terms in total, we need to look at each term individually. Let $D = 1/(2(1-C^2)^2)$. For the first term, we have

$$\begin{aligned}
&\frac{1}{p^2(p-1)^2}\sum_{\substack{i,j,n,k=1\\ i\neq j,n\neq k}}^{p}[|\mathrm{Cor}z_{ij},z_{nk})|\mathbb{1}(\rho_{ij}<C,\rho_{nk}<C)]\\
\leq&\frac{1}{p^2(p-1)^2}\sum_{\substack{i,j,n,k=1\\ i\neq j,n\neq k}}^{p}[D|\rho_{in}-\rho_{ij}\rho_{jn}||\rho_{jk}-\rho_{jn}\rho_{nk}|\\
&+D|\rho_{in}-\rho_{ik}\rho_{nk}||\rho_{jk}-\rho_{ij}\rho_{ik}|+D|\rho_{ik}-\rho_{ij}\rho_{jk}||\rho_{jn}-\rho_{jk}\rho_{nk}|\\
&+D|\rho_{ik}-\rho_{in}\rho_{nk}||\rho_{jn}-\rho_{ij}\rho_{in}|]+\left|\mathcal{O}\left(\frac{1}{n}\right)\right|\\
\leq&\frac{1}{p^2(p-1)^2}\sum_{\substack{i,j,n,k=1\\ i\neq j,n\neq k}}^{p}[D(|\rho_{in}|+|\rho_{ij}||\rho_{jn}|)(|\rho_{jk}|+|\rho_{jn}||\rho_{nk}|)\\
&+D(|\rho_{in}|+|\rho_{ik}||\rho_{nk}|)(|\rho_{jk}|+|\rho_{ij}||\rho_{ik}|)\\
&+D(|\rho_{ik}|+|\rho_{ij}||\rho_{jk}|)(|\rho_{jn}|+|\rho_{jk}||\rho_{nk}|)\\
&+D(|\rho_{ik}|+|\rho_{in}||\rho_{nk}|)(|\rho_{jn}|+|\rho_{ij}||\rho_{in}|)]+\left|\mathcal{O}\left(\frac{1}{n}\right)\right|\\
\leq&\frac{1}{p^2(p-1)^2}\sum_{\substack{i,j,n,k=1\\ i\neq j,n\neq k}}^{p}[2D(|\rho_{in}|+|\rho_{jn}|)+2D(|\rho_{jk}|+|\rho_{ik}|)\\
&+2D(|\rho_{ik}|+|\rho_{jk}|)+2D(|\rho_{jn}|+|\rho_{in}|)]+\left|\mathcal{O}\left(\frac{1}{n}\right)\right|\\
=&\frac{1}{p^2(p-1)^2}\sum_{\substack{i,j,n,k=1\\ i\neq j,n\neq k}}^{p}[4D(|\rho_{in}|+|\rho_{jn}|+|\rho_{jk}|+|\rho_{ik}|)]+\left|\mathcal{O}\left(\frac{1}{n}\right)\right|
\end{aligned}$$

For the second term, we have

$$\begin{aligned}
&\frac{1}{p^2(p-1)^2}\sum_{\substack{i,j,n,k=1\\ i\neq j,n\neq k}}^{p}[|\mathrm{Cor}(z_{ij},z_{nk})|\mathbb{1}(\rho_{ij}\geq C,\rho_{nk}<C)]\\
\leq\ &\frac{1}{(p-1)p}\sum_{\substack{i,j=1\\ i\neq j}}^{p}\mathbb{1}(\rho_{ij}\geq C)
\end{aligned}$$

For the third term, we have

$$\begin{aligned}
&\frac{1}{p^2(p-1)^2}\sum_{\substack{i,j,n,k=1\\ i\neq j,n\neq k}}^{p}[|\mathrm{Cor}(z_{ij},z_{nk})|\mathbb{1}(\rho_{ij}<C,\rho_{nk}\geq C)]\\
\leq\ &\frac{1}{(p-1)p}\sum_{\substack{n,k=1\\ n\neq k}}^{p}\mathbb{1}(\rho_{nk}\geq C)
\end{aligned}$$

For the last term, we have

$$\begin{aligned}
&\frac{1}{p^2(p-1)^2}\sum_{\substack{i,j,n,k=1\\ i\neq j,n\neq k}}^{p}[|\mathrm{Cor}(z_{ij},z_{nk})|\mathbb{1}(\rho_{ij}\geq C,\rho_{nk}\geq C)]\\
\leq\ &\frac{1}{p^2(p-1)^2}\sum_{\substack{i,j,n,k=1\\ i\neq j,n\neq k}}^{p}\mathbb{1}(\rho_{ij}\geq C)\mathbb{1}(\rho_{nk}\geq C)
\end{aligned}$$

Hence

$$\lim_{p\to\infty}\frac{1}{p^2(p-1)^2}\sum_{\substack{i,j,n,k=1\\ i\neq j,n\neq k}}^{p}|\mathrm{Cor}(z_{ij},z_{nk})|=\left|\mathcal{O}\left(\frac{1}{n}\right)\right|.$$

□

8.11 Conclusion

In this chapter, a new method is proposed to estimate a covariance matrix through estimation of the corresponding correlation matrix using empirical Bayes. Contrary to parametric empirical Bayes estimation of covariance matrices, the proposed method

focuses on the element-wise estimation of the underlying mean. The estimator still has desired theoretical properties even when the correlations among the sample correlation coefficients are slightly relaxed. The proposed method in principle can be modified in order to estimate variances. Recall that each x_{ij} for $j = 1, \cdots, n$ independently and marginally follows a normal distribution $\mathcal{N}(0, \sigma_{ii})$. By the central limit theorem, the MLE of the sample variance, i.e.,

$$\frac{1}{n}\sum_{j=1}^{n} x_{ij}^2 = \frac{\sigma_{ii}}{n}\sum_{j=1}^{n}\frac{x_{ij}^2}{\sigma_{ii}},$$

follows a normal distribution with mean σ_{ii} and variance σ_{ii}/n. Similar to the density estimation on the scale of sample correlation coefficients, the density of all the sample variances follows a one-parameter mixture, and hence the NPMLE can be solved using the constrained Newton method in Wang (2007); see computational notes in Section 8.5 for implementation. As a result, an estimate of the covariance matrix can be obtained using the estimate of the correlation matrix and the estimate of the variances separately, but a more rigorous approach should include the correlations between the sample variances and the sample correlation coefficients. In addition, the estimate using the minimiser of the sample correlation coefficients and the minimiser of the sample variances separately under expected total squared loss does not imply that this estimate is the minimiser of the sample covariance matrix.

Chapter 9

Numerical Studies for Estimating Covariance Matrices

9.1 Introduction

In this chapter, we first use two simulation studies to investigate the performance of the proposed method. In the first simulation study, the underlying covariance matrix used to generate data is weakly correlated but there is no zero anywhere. The underlying covariance matrix in the second simulation study is still weakly correlated and there are zeros in the entries, which allows us to see the performance of the covariance estimator based on the estimated proportion of zeros. These two models are special Auto-Regressive Moving-Average (ARMA) models; see Brockwell and Davis (2017). The NASDAQ stock dataset is then used to estimate the covariance matrix and then synthetic datasets are generated based on an estimate of the underlying covariance matrix to assess the performance of the proposed method in a general context. Lastly, we estimate a large covariance matrix under both sparse and non-sparse settings using the prostate data and the leukemia data respectively.

The existing methods to be compared with are Rao-Blackwell Ledoit-Wolf Estimator in Chen et al. (2010) based on Ledoit and Wolf (2004) which is implemented in the R package `CovTools` (Lee and You, 2019), the hard and soft thresholding of the sample covariance matrix implemented in the R package `CVTuningCov` (Wang, 2014), all of which have been introduced in Section 8.2. It is noted that the package `CovTools` also has implementations of various thresholding procedures of the sample covariance matrix, but there has been an issue with the code causing idling in parallel computing. Should any method require estimating a tuning parameter, the process of conducting

cross-validation can be found in the reference above and therein and the candidate tuning parameters range from 0 to 0.95 with increment of 0.01. For our proposed method, the estimation is conducted on the Fisher-transformed scale so that the proportion of zeros can be estimated and the binning can be used to speed up the estimation when the size of the covariance matrix is large. In particular, binning with $h = 10^{-3}$ is used when the number of observations passed to compute the NPMLE is more than 5000 and the threshold function with AIC type $q(n, \alpha) = 1$ is used for estimating proportion of zeros. The method in Qi and Sun (2006) is applied to the estimate produced by the proposed method to ensure the resulting estimated matrix is positive-definite. Throughout this chapter, all the observations are generated independently using a multivariate normal distribution with mean $\mathbf{0}$. In total, 100 datasets are generated for each p and the number of observations used to evaluate the sample covariance matrix is 100. The Frobenius norm is used to measure the distance between the estimates and the underlying covariance matrix.

9.2 An Autoregressive Model

An autoregressive (AR) model is a time series model used to predict future behaviour based on past behaviour. It is used for forecasting when there is some correlation between values in a time series and the values that precede and succeed them. In this section, the performance of the proposed method is investigated through estimating the covariance matrix of the autoregressive model with lag order 1 (AR1). The true covariance matrix $\mathbf{\Sigma}$ of size $p \times p$ has entries

$$\rho_{ij} = \rho^{|i-j|},$$

for $\rho = 0.1, 0.3, 0.5$ and $p = 50, 100, 200$. This AR1 model is in the class $\mathcal{U}(1, c_0(p))$, with $c_0(p)$ a constant that does not depend on p and contains no zeros anywhere in the covariance matrix unless $\rho = 0$.

ρ	EB	EB0	RBLW	ST	HT	Sample
			$p = 50$			
0.1	1.414(0.008)	1.410(0.007)	**1.012**(0.004)	2.038(0.009)	1.411(0.007)	5.064(0.015)
0.3	2.213(0.013)	**2.209**(0.013)	2.664(0.005)	2.674(0.013)	2.919(0.018)	5.081(0.017)
0.5	2.621(0.015)	**2.618**(0.014)	3.823(0.010)	3.209(0.017)	3.231(0.016)	5.078(0.022)
			$p = 100$			
0.1	2.007(0.008)	2.005(0.008)	**1.437**(0.004)	3.061(0.009)	1.999(0.007)	10.082(0.018)
0.3	3.361(0.012)	**3.356**(0.012)	4.068(0.003)	4.149(0.015)	4.400(0.015)	10.087(0.022)
0.5	3.927(0.015)	**3.921**(0.015)	6.346(0.008)	4.994(0.017)	4.855(0.022)	10.131(0.028)
			$p = 200$			
0.1	2.857(0.007)	2.860(0.009)	**2.051**(0.004)	4.614(0.008)	2.839(0.007)	20.133(0.022)
0.3	5.025(0.012)	**5.021**(0.012)	6.007(0.002)	6.285(0.011)	6.441(0.011)	20.199(0.025)
0.5	5.836(0.015)	**5.835**(0.015)	10.012(0.006)	7.704(0.018)	7.248(0.023)	20.136(0.030)

TABLE 9.1: The mean (standard error) of Frobenius norms between the estimated covariance matrices and the true covariance matrix under the AR1 model.

Table 9.1 shows the means and standard errors of the Frobenius norms between the estimated covariance matrices and the true one. The "EB" and "EB0" represent the our proposed methods without estimating proportion of zeros and with estimating proportion of zeros respectively. The term "RBLW" is the Rao-Blackwell Ledoit-Wolf Estimator, "ST" is the soft-thresholding on the sample covariance matrix, "HT" is the hard-thresholding on the sample covariance matrix and "Sample" is the sample covariance matrix. The Frobenius norms of all the existing methods except the sample covariance matrix are increasing when ρ increases. The Frobenius norm of the sample covariance matrix seems to be increasing linearly with respect to the increase in p while all other existing methods appear to increase slower than the linear rate. The Rao-Blackwell Ledoit-Wolf Estimator has the best performance when ρ is small and deteriorates when ρ increases. The reason is that the Rao-Blackwell Ledoit-Wolf estimator is a shrinkage estimator and it is expected to perform well when the magnitude of the sample correlation coefficients are small. The proposed methods perform the best when ρ is large and has a competitive performance when $\rho = 0.1$.

9.3 A Moving Average Model

The performance of our proposed method is also investigated in the sparse setting using the moving-average model with order 1. The true covariance matrix $\mathbf{\Sigma}$ of size $p \times p$

has entries

$$\rho_{ij} = \rho \mathbb{1}(|i-j| = 1)$$

for $\rho = 0.1, 0.3, 0.5$ and $p = 50, 100, 200$. This MA1 model is also in the class $\mathcal{U}(1, c_0(p))$, with $c_0(p)$ a constant that does not depend on p, and the entries of the covariance matrix are all zeros when the distance to the diagonal $|i-j|$ is greater than 1.

ρ	EB	EB0	RBLW	ST	HT	Sample
			$p = 50$			
0.1	1.428(0.009)	1.423(0.009)	**1.009**(0.004)	2.029(0.009)	1.416(0.008)	5.086(0.014)
0.3	2.060(0.013)	**2.050**(0.013)	2.579(0.003)	2.554(0.012)	2.764(0.020)	5.075(0.014)
0.5	1.441(0.021)	**1.421**(0.021)	3.562(0.006)	2.596(0.013)	1.933(0.024)	5.097(0.018)
			$p = 100$			
0.1	2.006(0.008)	2.000(0.008)	**1.422**(0.003)	3.055(0.009)	1.997(0.007)	10.093(0.017)
0.3	3.120(0.013)	**3.113**(0.013)	3.910(0.002)	3.958(0.011)	4.171(0.017)	10.078(0.021)
0.5	2.096(0.019)	**2.081**(0.018)	5.795(0.004)	4.041(0.014)	2.977(0.027)	10.136(0.023)
			$p = 200$			
0.1	2.853(0.008)	2.862(0.012)	**2.042**(0.004)	4.610(0.009)	2.845(0.008)	20.154(0.023)
0.3	4.701(0.011)	**4.698**(0.011)	5.755(0.002)	6.025(0.011)	6.149(0.014)	20.201(0.023)
0.5	3.111(0.026)	**3.100**(0.026)	8.950(0.003)	6.216(0.017)	4.656(0.036)	20.169(0.027)

TABLE 9.2: The mean (standard error) of Frobenius norms between the estimated covariance matrices and the true covariance matrix under the MA1 model.

Table 9.2 shows the means and standard errors of the Frobenius norms between the estimated covariance matrices and the true one. The Rao-Blackwell Ledoit-Wolf Estimator still has the best performance when ρ is small and deteriorates when ρ increases. The proposed method with estimating proportion of zeros performs the best when ρ is away from 0 and the differences between the proposed methods with estimating zeros and without estimating zeros are small. It is interesting to see that the mean Forbenius norm of the proposed method when $\rho = 0.3$ is even higher than the one when $\rho = 0.5$ and it is worth investigating.

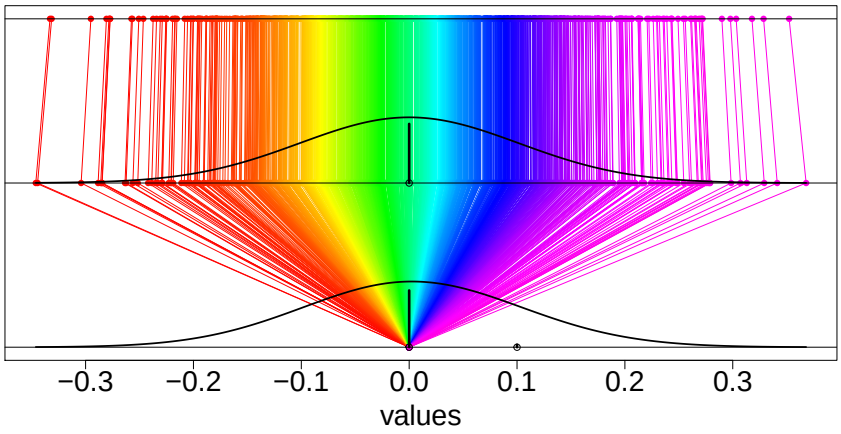

FIGURE 9.1: Plot of the maps from sample correlation coefficients to the corresponding estimated correlation coefficients when $\rho = 0.1$ and $p = 50$.

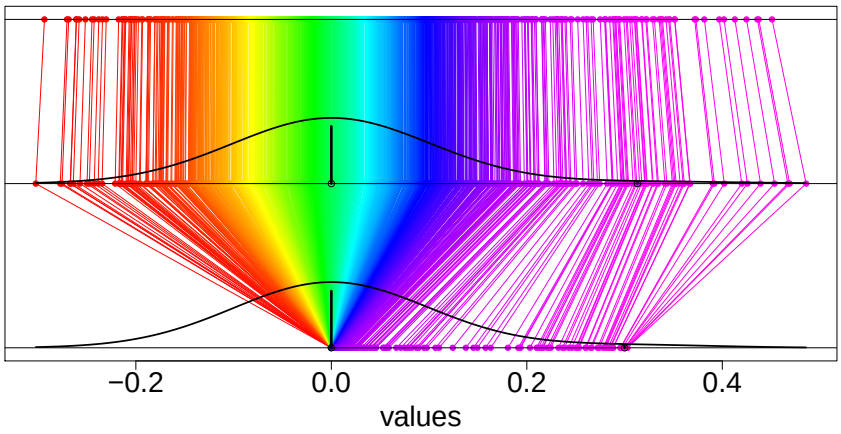

FIGURE 9.2: Plot of the maps from sample correlation coefficients to the corresponding estimated correlation coefficients when $\rho = 0.3$ and $p = 50$.

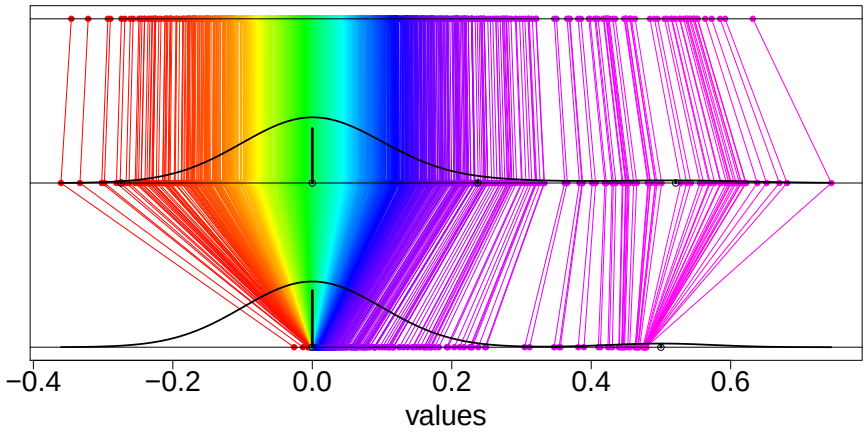

FIGURE 9.3: Plot of the map from sample correlation coefficients to the corresponding estimated correlation coefficients when $\rho = 0.5$ and $p = 50$.

Figures 9.1, 9.2 and 9.3 show the maps from sample correlation coefficients to the corresponding estimated correlation coefficients through the Fisher transformation when $\rho = 0.1, 0.3, 0.5$ respectively. In each figure, sample correlation coefficients are shown at the top of the figure while the bottom shows estimated correlation coefficients. The density shown in the bottom of the figure is the true density. Points in the middle are sample correlation coefficients under the Fisher transformation and the density is estimated using the restricted NPMLE given the estimated level of sparsity. In Figure 9.1, the corresponding non-zero support point on the transformed scale is close to 0 and the posterior mean shrinks all the transformed sample correlation coefficients to 0. In Figure 9.3, the corresponding non-zero support point on the transformed scale is relatively far away from 0 and the transformed sample correlation coefficients in the right shrunk to the estimated non-zero support points. However, as shown in Figure 9.2, the corresponding non-zero support point on the transformed scale can be estimated, but the transformed sample correlation coefficients are still shrunk to 0 even for the observations close to the estimated non-zero support points due to the dominating posterior proportion at 0. Despite this, the proposed methods still perform better than other existing methods, as shown in Table 9.2.

9.4 The NASDAQ Stock Data

Estimation of covariance matrices has broad applications in econometrics, it is commonly used to estimate and forecast the excess returns of financial instruments over time. For instance, Wang and Zou (2010) estimated sparse volatility matrices directly under high-frequency settings and Aït-Sahalia and Xiu (2017) use a factor-based model to estimate the covariance matrix. In this section, we attempt to estimate the covariance matrix of the log-returns for the stock data under the assumption that the covariances among the stocks are constant over time. 100 companies listed in NASDAQ are randomly chosen and the daily closing prices from 31/10/2014 to 27/10/2017 are obtained from the Alpha Vantage website using the R package `alphavantager` (Dancho and Vaughan, 2020). The full list of these 100 publicly listed companies can be found in Appendix C. As introduced in financial courses, the daily log-return approximately follows a normal distribution and hence all the closing prices are converted into the log-returns and a normal distribution is used as the component distribution. The primary focus of the analysis in this section is the covariances among the log-returns of the stocks, the effect of autocorrelation should be removed as much as possible. In detail, for every stock, the daily log-returns are divided into groups of 7 according to the time stamps and the middle of each group is taken as the observation for the group. The new observations are also standardised using the observations in the corresponding group since the means and variances of the new observations may be all different. Hypothesis tests on the new observations are carried out for each stock in order to make sure that the autocorrelations among the observations are removed. The Ljung-Box test (Ljung and Box, 1978) shows that 4 stocks have p-values less than 0.05, the Breusch-Godfrey test (Breusch, 1978; Godfrey, 1978) shows that 4 stocks have p-values less than 0.05 and the Durbin-Watson test (Durbin and Watson, 1950, 1951, 1971) shows that 3 stocks have p-values less than 0.05, which does not need to be of concern. The Ljung-Box test is implemented in the R package `stats` (R Core Team, 2020) and the Breusch-Godfrey and the Durbin-Watson tests are implemented in the R package `lmtest` (Zeileis and Hothorn, 2002).

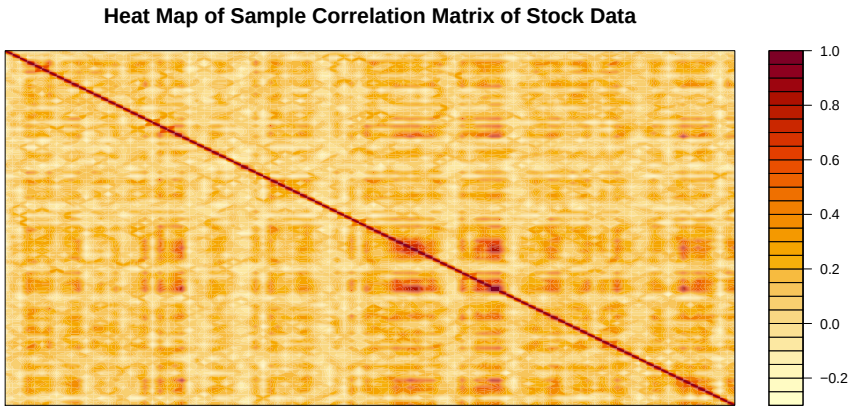

FIGURE 9.4: Heat map of the sample correlation matrix of the log-returns.

Figure 9.4 shows the heat map of the sample correlation matrix. Some blocks have quite high correlation coefficients while some correlation coefficients are negative. It is unclear whether the underlying correlation matrix is weakly correlated and the conditions for the consistency of the NPMLE may not hold. We argue that none of the existing methods used here meet all their assumptions either. Since we do not know the underlying covariance matrix, 10-fold cross-validation is used to assess the performance. The log-returns are randomly assigned to 10 different groups. For every group i, $i = 1, \cdots, 10$, log-returns for all the stocks that are not in group i are used for estimating the covariance matrix (training set) and Frobenius norm between the estimate obtained from the training set and the sample covariance for the i group (testing set) is computed. The Frobenius norm is taken as the measurement for the performance. In order to reduce the standard error of the estimation for each method, this process is repeated 100 times with the observations randomly shuffled in each repetition.

	EB	EB0	RBLW	ST	HT	Sample
mean	**32.954**	32.957	33.327	33.386	33.731	33.706
SE	0.061	0.061	0.062	0.062	0.059	0.059

TABLE 9.3: The Frobenius norm of estimating the covariance matrix of log-returns from the stock data using various methods when 10-fold cross-validation is used.

Table 9.3 shows the Forbenius norms of various methods for estimating the covariance matrix in the 10-fold cross-validation for each repetition as well as the mean and the standard error of the testing errors for each method. The proposed method without estimating the proportion of zeros seems to have the lowest mean testing error, followed by the proposed method with estimating the proportion of zeros, both of which are at least two standard errors smaller than those of the existing methods.

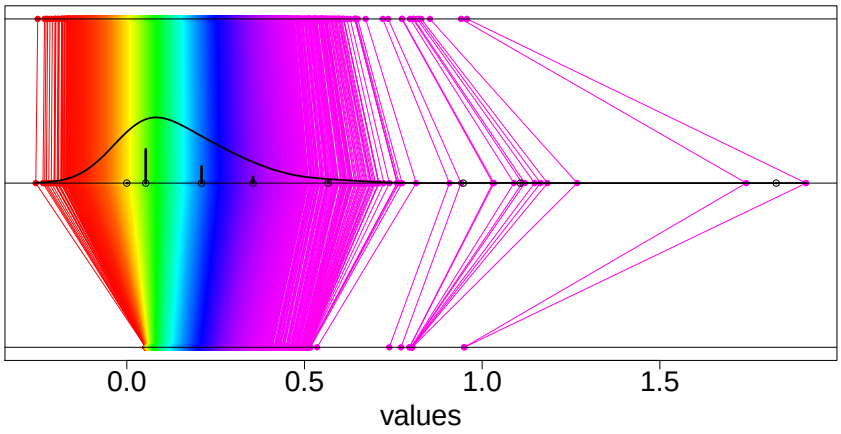

FIGURE 9.5: Plot of the maps from transformed sample correlation coefficients to the corresponding estimated correlation coefficients for the log-returns.

Figure 9.5 shows the maps from sample correlation coefficients to the corresponding estimated correlation coefficients through the Fisher transformation. Sample correlation coefficients are shown at the top of the figure while the bottom shows estimated correlation coefficients. Points in the middle are sample correlation coefficients under the Fisher transformation and the density is estimated using the unrestricted NPMLE. Points in the right are shrunken to towards the support points in the right while points in the middle are shrunken to a small neighbourhood indicating that taking the posterior mean is indeed a deconvolution process. The estimated covariance matrix using "EB" is then used to generate the synthetic datasets in order to further assess the performances of all methods in a general covariance structure that the assumption may not hold.

	EB	EB0	RBLW	ST	HT	Sample
mean	**7.052**	7.055	8.902	9.406	9.951	9.898
SE	0.069	0.069	0.087	0.081	0.074	0.056

TABLE 9.4: The Frobenius norms using various methods from 100 synthetic datasets generated from the estimated covariance matrix using "EB".

Table 9.4 shows the means and standard errors of the Frobenius norms between the estimated covariance matrices and the true one using all the methods. The proposed methods have smaller mean Forbenius norms followed by the Rao-Blackwell Ledoit-Wolf Estimator, and the hard-thresholding estimator has the highest mean Forbenius norm.

9.5 The Prostate Data

The prostate data mentioned in Section 4.2 are the z-values converted from t-values after performing the two-sample t-test assuming equal variance. In this section, we are going back to the samples performing the t-tests and investigate the covariance structure among those genes in order to show that the proposed method can be applied in the high-dimensional setting and to check that the independence assumption, which is needed in multiple hypothesis testing and FDR control, is feasible. The original dataset consists of 6033 expressions for 102 patients and is available in the R package `sda` (Ahdesmaki et al., 2015). Due to the limitation on the memory of the computing device, only the first 3000 gene expression levels out of 6033 are considered.

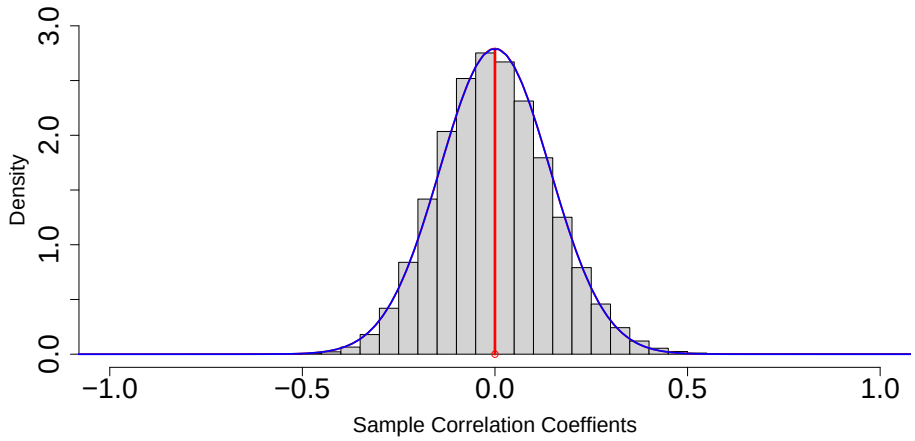

FIGURE 9.6: Histogram of the Fisher-transformed sample correlation coefficients for the first 3000 genes of the cancer patients in prostate data.

Figure 9.6 shows the histogram of the Fisher-transformed sample correlation coefficients for the gene expression levels of the cancer patients. The blue line is the density of the normal distribution centred at 0 and the red lines are the mixture density computed from the unrestricted NPMLE. Two density curves are almost identical by virtual inspection and they are also close fits to the histogram, which suggests that the correlations among the gene expression levels seem to be quite weak. We are interested in the estimation of the covariance matrix of the gene expressions for the cancer patients. Since we do not know the underlying covariance matrix, cross-validation is used to assess the performance. 52 patients are randomly assigned to 4 different groups with 13 patients each. The use of 4-fold is simply because the number of patients are relatively small and 52 is divisible by 4. For every group i, $i = 1, 2, 3, 4$, gene expression levels for all the patients that are not in the group i are used for estimating the covariance matrix (training set) and Frobenius norm between the estimate obtained from the training set and the sample covariance for the i group (testing set) is computed. The Frobenius norm is taken as the measurement of performance. In order to reduce the standard error of the estimation for each method, this process is repeated 100 times with the observations are randomly shuffled in each repetition.

	EB	EB0	RBLW	ST	HT	Sample
mean	**793.999**	**793.999**	794.307	794.413	794.909	905.567
SE	0.122	0.122	0.121	0.121	0.121	0.103

TABLE 9.5: The Frobenius norm of estimating the covariance matrix of expression levels for the cancer patients in the prostate data using various methods when 4-fold cross-validation is used.

Table 9.5 shows the Frobenius norms of various methods for estimating the covariance matrix of gene expression levels. The proposed methods have the lowest mean testing errors, both of which are at least two standard errors smaller than those of the existing methods. The sample covariance matrix has the highest mean testing error. The proposed method with estimating the proportion of zeros is then applied to all of the cancer patients, the estimated sparsity level is 1 and the estimated covariance matrix is a diagonal matrix, which suggests that the first 3000 genes seem to be uncorrelated. Figure 9.7 shows the maps between sample correlation coefficients to the corresponding estimated correlation coefficients through the Fisher transformation, for 1000 randomly selected sample correlation coefficients. Sample correlation coefficients are shown at the top of the figure while the bottom shows estimated correlation coefficients. Points in the middle are sample correlation coefficients under the Fisher transformation and the density is estimated using the restricted NPMLE given the estimated level of sparsity. There is only one support points in the estimated mixing distribution so all the sample correlation coefficients are shrunken to the neighbourhood around 0.

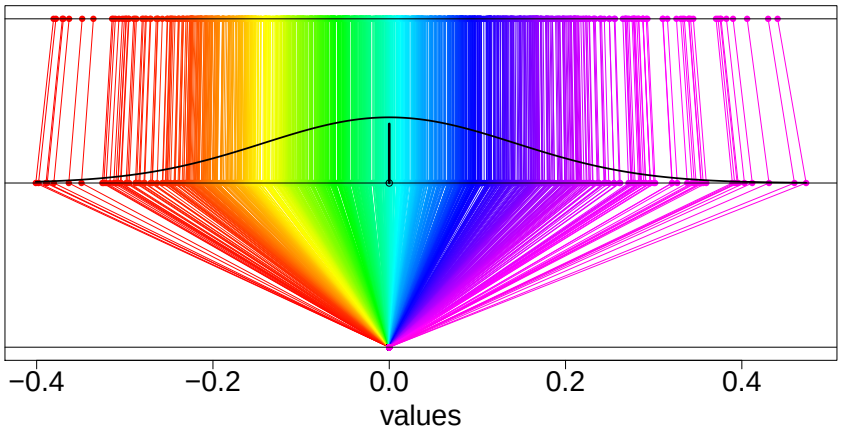

FIGURE 9.7: Plot of the maps from 1000 randomly chosen transformed sample correlation coefficients and the corresponding estimated correlation coefficients for the cancer patients in the prostate data.

Similar analysis can be extended to all of 6033 genes if the computation capacity permits and also applied to the group of normal controls. This allows to check whether the independence assumption is feasible in multiple hypothesis testing and FDR control in Section 4.2.

9.6 The Leukemia Data

The previous section uses the prostate data to assess the performance of the proposed covariance matrix estimator under sparsity. In this section, the performance of the proposed covariance matrix estimator under a non-sparse covariance structure is investigated. The dataset used in this section is the Leukemia data in Golub et al. (1999). High-density oligonucleotide microarrays provided expression levels on 7128 genes for 72 patients, with 45 with acute lymphoblastic leukemia (ALL) and 27 with acute myeloid leukemia (AML). It is noted that the AML has the worse prognosis. The full dataset can be obtained in the supplementary material of Efron (2010b). We only analyse the covariance structure of the first 1000 gene expression levels, which should be sufficient for the demonstration and the result should be a good representative of the entire dataset. In particular, we are interested in the estimation of the covariance matrix of the gene expressions for the patients with ALL.

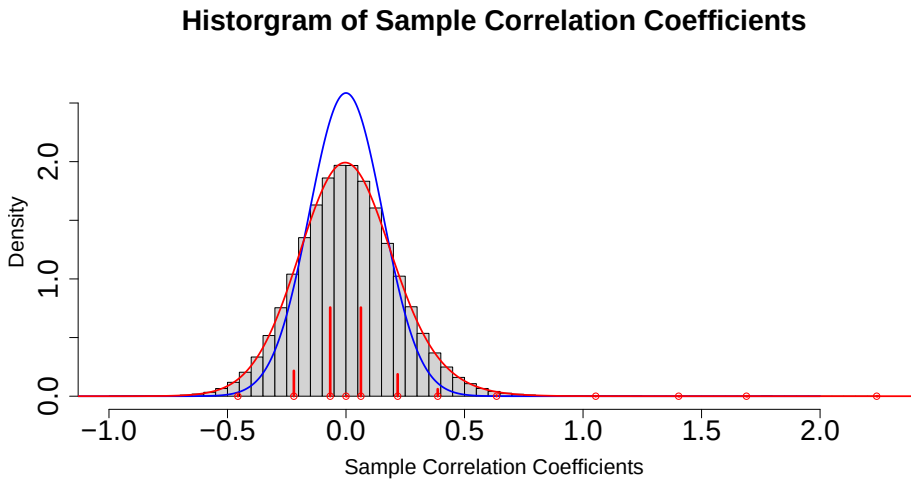

FIGURE 9.8: Histogram of the Fisher-transformed sample correlation coefficients for the first 1000 genes of patients with ALL in Leukemia data.

Figure 9.8 shows the histogram of the Fisher-transformed sample correlation coefficients for the gene expression levels for the patients with ALL in the Leukemia data. The blue line is the density of the normal distribution centred at 0 and the red lines are the normal mixtures computed from the unrestricted NPMLE. There is a large deviation between the normal density and the mixture density, which suggests that the underlying covariance structure embedded in this dataset is not sparse. The red line is much higher than the blue line at around -0.5 and 0.5, suggesting that some of the correlations among the first 1000 gene expressions are moderate. Since we do not know the underlying covariance matrix, 5-fold cross-validation is used to assess the performance. The use of 5-fold is simply because the 45 patients can be partitioned into 5 groups with the same size. For every group i, $i = 1, \cdots, 5$, gene expression levels for all the patients with ALL that are not in the group i are used for estimating the covariance matrix (training set) and Frobenius norm between the estimate obtained from the training set and the sample covariance for the i group (testing set) is computed. The Frobenius norm is taken as the measurement of performance. In order to reduce the standard error of the estimation for each method, this process is repeated 100 times with the observations are randomly shuffled in each repetition.

	EB	EB0	RBLW	ST	HT	Sample
mean	**62.934**	62.942	63.138	63.749	64.173	67.075
SE	0.068	0.068	0.069	0.071	0.071	0.063

TABLE 9.6: The Frobenius norm of estimating the covariance matrix of expression levels for the patients with ALL in Leukemia data using various methods when 5-fold cross-validation is used.

Figure 9.6 shows the Frobenius norms of various methods for estimating the covariance matrix of first 1000 gene expression levels for patients with ALL in Leukemia data. The proposed methods with and without estimating the proportion of zeros have the lowest mean testing errors, followed by the Rao-Blackwell Ledoit-Wolf Estimator. Both of two proposed methods are at least two standard errors smaller than those of the existing methods. The results for the soft and the hard thresholding on the sample covariance matrix is statistically significantly higher than the proposed methods and the Rao-Blackwell Ledoit-Wolf Estimator, but still smaller than the mean testing error for the sample covariance matrix. The proposed method with estimating the proportion of zeros is then applied to all of the normal controls, the estimated sparsity level is 0.588. Figure 9.9 shows the maps between sample correlation coefficients to the corresponding estimated correlation coefficients through the Fisher transformation, for 1000 randomly selected sample correlation coefficients. Sample correlation coefficients are shown at the top of the figure while the bottom shows estimated correlation coefficients. Points in the middle are sample correlation coefficients under the Fisher transformation and the density is estimated using the restricted NPMLE given the estimated level of sparsity. There are two support points that are not in the neighbourhood around 0 but these points are not significantly far way and hence all of these randomly chosen sample correlation coefficients are shrunken towards a smaller value.

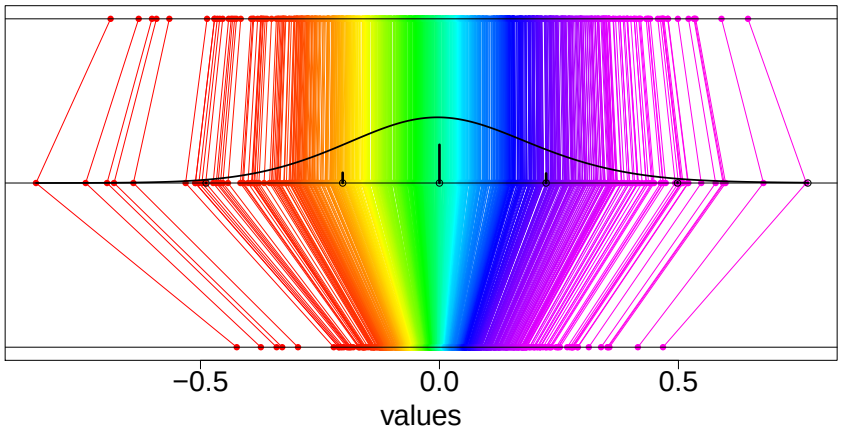

FIGURE 9.9: Plot of the maps from 1000 randomly chosen transformed sample correlation coefficients to the corresponding estimated correlation coefficients for patients with ALL in the Leukemia data.

9.7 Conclusion

In this chapter, numerical studies are conducted in order to assess the performance of the proposed method for estimating the covariance matrix. Using two ARMA models in which the assumptions of all the existing methods are satisfied, the proposed estimator has a competitive performance against the existing methods when the underlying covariance matrix is close to an identity matrix but out-performs the existing methods when ρ is far away from zero. The results based on the NASDAQ stock and the leukemia data show that the proposed estimator can potentially be applied to general covariance structures while the applications on the leukemia and the prostate data show that the proposed estimator is potentially suitable for large-scale covariance estimation.

Chapter 10

Summary and Future Work

10.1 Summary

In this thesis, we give a new perspective and present new methods for solving large-scale inferential problems using non-parametric mixtures. The proposed methods are applied to different statistical inferential problems. In the context of multiple hypothesis testing, we offer a new approach to estimating the proportion of null effects and the underlying distribution of the test statistics based on this proportion of null effects using the NPMLE. If the test statistics are z-values, a consistent estimator of the null proportion can be obtained using GLRT. We also derive distance-based alternatives to estimating the null proportion and its mixing distribution under the Cramér-von Mises and Anderson-Darling distances, and prove the consistency and characteristics of these NPMDEs. The computation details are given for these NPMLE and NPMDEs, and modifications are made in order to accommodate large datasets using the discrete normal distribution. A new FDR-controlling procedure based on the estimated null proportion and its estimated mixing distribution is also proposed. In the problem of covariance matrix estimation, we propose a new covariance matrix estimator by estimating the correlation matrix using empirical Bayes. Numerical studies suggest that the methods based on this new perspective have outstanding performances and meaningful interpretations. While the application of this new approach only covers two main areas in this thesis, it is expected that it can be applied to a wider range of problems.

Chapter 2 gives a review of non-parametric mixture methods to date. More specifically, Section 2.3 introduces distance-based non-parametric methods while Section 2.4 reviews methods for non-parametric maximum likelihood estimation, which is a special and important case of the minimum distance estimation under the KL divergence.

In Chapter 3, we discuss the problem of the unrestricted NPMLE in the context of multiple hypothesis testing and the need of the restricted NPMLE, an estimate of

the mixing distribution assuming that the proportion of the null effects is at a pre-specified level π_0. With the help of the restricted NPMLE, the proportion of null effects can then be estimated using various threshold functions. Section 3.3.4 presents some partial work on the asymptotic behaviour of the likelihood ratio as a function of the null proportion. Theoretical results and the implementation of the proposed method based on Wang (2007) are discussed in Sections 3.4 and 3.5 respectively.

Chapter 4 shows the numerical studies of estimating the null proportions. The prostate dataset in Section 4.2 has large degrees of freedom and can be used to evaluate the performance when the component density is normal or assumed to be normal. The proposed method performs better than the locfdr and the mixfdr method, and is comparable with the ECF method in term of the distance to the true null proportion when the AIC threshold function is used. In section 4.3, the breast cancer data is used to illustrate the problem of converting t-values to z-values when the degrees of freedom are small. The proposed method can estimate the null proportion directly using t-values by setting the component density to be the appropriate t-distribution. Other existing methods require converting t-values into z-values, and hence the null proportions are under-estimated. The two-component model in Section 4.4 shows that the proposed method performs the best when the location of the non-null effect is either close to or far away from the location of the null effect, almost uniformly for all choices of the true null proportion.

Inspired by the restricted and the unrestricted NPMLE, the restricted and the unrestricted NPMDE under the Cramér-von Mises and the Anderson-Darling distance are introduced in Chapter 5 in complement to the likelihood-based method. The proportion of null effects can be consistently estimated under the Cramér-von Mises and the Anderson-Darling distance. The median estimator can be obtained while the mean estimator is not available since the asymptotic distributions of the test statistics are unknown with the presence of the non-parametric nuisance parameter. The computation and theoretical results of the estimators are given. The numerical studies in Section 5.5 show that the NPMLEs seem to be more efficient than the corresponding NPMDEs while the NPMDEs are slightly faster to compute in general.

Chapter 6 shows that the NPMLEs and the NPMDEs under the Cramér-von Mises and the Anderson-Darling distance can be modified to accommodate the large-scale computation, with the help of binning and the discrete normal distributions. The numerical studies in Section 6.6 show that binning significantly reduces the computation time without losing too much accuracy, under an appropriate choice of the bin width.

Using this modification, Section 6.7 continues the investigation in Section 4.2 about the asymptotic behaviour of estimating the proportion of null effects for our proposed method by generating synthetic datasets with an even larger number of observations. It is shown that the proposed method still has a significant converging trend, while other existing methods seem to converge slowly.

A new FDR-controlling procedure based on the estimated null proportion and its estimated mixing distribution is introduced in Chapter 7. The proposed procedure can control the FDR at the pre-specified level exactly as the number of hypotheses tends to infinity, under the independence assumption. The numerical studies are conducted in Section 7.4 to assess the performance of the proposed FDR-controlling procedure. The mean FDRs are sufficiently close to the pre-specified level when the null proportion and the component distribution are correctly specified. The breast cancer dataset in Section 4.3 is used to show the behaviour of the proposed FDR-controlling procedure under misspecification compared with the existing methods. Using the prostate dataset in Section 4.2, the proposed procedure is shown to make more rejections than the original Benjamini-Hochberg and the adaptive step-down procedure, given the FDR is controlled at a pre-specified level.

A covariance matrix estimator is proposed in Chapter 8 using empirical Bayes. The proposed method focuses on estimating each element in the correlation matrix, which is different from other empirical Bayes estimators. The estimating procedure on the sample correlation coefficients is given in Section 8.5, where the density of the sample correlation coefficients are estimated using mixtures of one-parameter distributions and the estimates of the correlation coefficients are computed using the posterior mean. The estimating procedure on the Fisher-transformed sample correlation coefficients is given in Section 8.6, which the density is estimated non-parametrically using normal components. Sparse covariance matrices can also be estimated using the restricted NPMLE, although there is no zero anywhere in the covariance matrix estimator unless the sparsity level is estimated to be 1, as shown in Section 8.7. Section 8.9 presents some theoretical developments of the proposed covariance matrix estimator under a wider range of covariance structures.

The numerical studies of estimating covariance matrices are shown in Chapter 9. Using the ARMA models in Sections 9.2 and 9.3, it can be shown that the proposed estimator is competitive with existing methods when the non-zero entries in the underlying covariance matrix are close to zero and out-performs the existing methods when the non-zero entries are far away from zero. Section 9.4 uses NASDAQ stock data

to show that the proposed estimator performs well in a general covariance structure. Large-scale covariance matrix estimation is performed in Sections 9.5 and 9.6, and it is shown that the proposed method is at least comparable with and may potentially out-perform the existing methods under both the sparse and the non-sparse setting.

10.2 Future Work

The major milestone of this thesis is that a new approach on solving large-scale problems using non-parametric mixtures is proposed and new methods based on this new approach in various settings are derived. This enables the applicability of non-parametric methods on a wider range of problems. The thesis only shows applications to multiple hypothesis testing and covariance matrix estimation. In this section, we briefly discuss the future work in the methods proposed for multiple hypothesis testing and covariance matrix estimation, as well as the possibility of other future work.

10.2.1 Multiple Hypothesis Testing

As shown in Section 3.3.4, due to the incompleteness of asymptotic theoretical results in non-parametric maximum likelihood estimation, it is difficult to obtain asymptotic results for the estimator of the proportion of null effects. We can only rely on the results in Jiang and Zhang (2016), which use the large deviation inequality to provide a rate of divergence of the GLRT under the null hypothesis. The estimator based on this large deviation inequality, however, over-estimates the proportion of null effects significantly. Figure 3.2 shows an interesting behaviour that the likelihood ratio as a function of the null proportion seems to converge to an indicator function stepping down at the underlying null proportion asymptotically, and hence it is possible that any constant threshold function can produce a consistent estimator under certain conditions. Efforts have been made to investigate the conditions under which this interesting behaviour holds and partial work is presented. The difficulty is to find the rate of convergence of the density estimator and the support points near the location of the null effect, when the proportion of null effects is under-specified in the non-parametric setting. This is not an issue in parametric models; when the number of components is correctly specified, the rate of convergence can be obtained by the central limit theorem.

The choice of the threshold function in practice is also a challenging problem when estimating the proportion of null effects. In distance-based methods such as

the Cramér-von Mises and the Anderson-Darling distance, the median can be taken as a generic estimator of the null proportion in estimation. As discussed in Section 3.6, we could make use of the similarity between the likelihood ratio function for the likelihood-based method and the functions of p-values for the distance-based methods in order to obtain a generic estimator of the null proportion for the likelihood-based method. Using $q(n, \alpha) = -\log 0.5$ as a likelihood-based analogue of the median estimator in distance-based methods, the estimated null proportion seems to perform well. However, the theoretical justification needs to be investigated.

10.2.2 Covariance Matrix Estimation

Similar to most of the existing methods in the literature, the proposed covariance matrix estimator is constructed by not taking into account the correlations among the sample correlation coefficients. In Section 8.9, a consistent estimator of the distribution of the Fisher-transformed sample correlation coefficients can be constructed if the sample correlation coefficients are weakly correlated. This covers the class of covariance matrices described in Bickel and Levina (2008a) with $q = 0$ and $c_0(p) = o(p)$. It is of interest to investigate a broader range of covariance matrix structures. For example, sample correlation coefficients in the covariance matrix with off-diagonal entries the same are not weakly correlated. The correlation matrix of the transformed sample correlation coefficients needs to be constructed in order to estimate the stochastic means of the transformed sample correlation coefficients by performing eigen-decomposition or principle component analysis according to Azriel and Schwartzman (2015). However, constructing the correlation matrix of the transformed sample correlation coefficients is not computationally feasible in practice: for a covariance matrix of size 50 by 50, the size of the correlation matrix of the transformed sample correlation coefficients is 1225 by 1225, quadratic in the dimension of the target covariance matrix. Furthermore, as shown in (8.1), correlations among the transformed sample correlation coefficients are non-linear functions, and hence an estimator based on observed sample correlation coefficients is needed. The plug-in estimator may be a slightly simple choice, but the performance of this estimator is unknown without a rigorous analysis.

One can also apply the proposed method to the variances in order to further improve the accuracy of the covariance matrix estimator, but a more rigorous approach should include the correlations among the sample variances and the sample correlation coefficients. In addition, the estimate using the minimiser of the sample correlation

coefficients and the minimiser of the sample variances separately under expected total squared loss does not imply that this estimate is the minimiser of the sample covariance matrix.

10.2.3 Estimating More Than One Weight

In this thesis, we only consider estimating the weight of one particular support point. It is also of interest to estimate the weights of more than one support point. The process of estimating more than one weight should be very similar to the one-dimensional problem. The difficulty is that there are infinite number of solutions to the root-finding problem for more than one parameter while the solution is unique in the one-dimensional case. One may consider the following constrained optimisation problem for estimating the weights using maximum likelihood as an example:

$$\text{Maximise} \qquad ||\boldsymbol{\pi}||_2,$$

subject to

$$\begin{aligned} \tilde{\ell}_n(\mathbf{0}) - \tilde{\ell}_n(\boldsymbol{\pi}) &\leq q(n, \alpha), \\ \mathbf{1}^T \boldsymbol{\pi} &\leq 1, \\ \boldsymbol{l} \leq \boldsymbol{\pi} &\leq \boldsymbol{u}, \end{aligned}$$

where

$$\tilde{\ell}_n(\boldsymbol{\pi}) = \sup_{G \in \Gamma} \sum_{i=1}^{n} \log f(z_i; G, \sigma, \boldsymbol{\pi}, \boldsymbol{\beta})$$

and $\boldsymbol{u}$ and $\boldsymbol{l}$ are the vectors of upper and lower bounds for the parameters. If a parameter needs to be estimated, the bounds can be $[0, 1]$ and the bounds can be set to be the same value if a parameter is fixed. This optimisation problem is equivalent to the one in Section 3.5.2 if $\boldsymbol{\pi}$ is a scalar. The optimisation problems for distance-based methods can be derived in the similar manner. The choice of the norm as the objective function in the aforementioned optimisation problem should be based on the theory. Furthermore, as mentioned in Sections 3.3.4 and 10.2.1 that the asymptotic theory for estimating the weight of one location parameter is already challenging, it would be more difficult to obtain the theoretical properties for estimators for more than one parameter.

Appendix A

Evaluating density of t-distribution and its derivative

In Example 3.2, the density of non-central t-distribution with non-centrality parameter μ and degree of freedom σ is

$$\phi(x;\mu,\sigma) = \frac{\sigma^{\frac{\sigma}{2}} \exp\left(-\frac{\sigma\mu^2}{2(x^2+\sigma)}\right)}{\sqrt{\pi}\Gamma\left(\frac{\sigma}{2}\right) 2^{\frac{\sigma-1}{2}} (x^2+\sigma)^{\frac{\sigma+1}{2}}} \int_0^\infty y^\sigma \exp\left[-\frac{1}{2}\left(y - \frac{\mu x}{\sqrt{x^2+\sigma}}\right)^2\right] dy \quad \text{(A.1)}$$

with the derivative with respect to the non-centrality parameter μ,

$$\begin{aligned}\frac{\partial}{\partial\mu}\phi(x;\mu,\sigma) = &\frac{\sigma^{\frac{\sigma}{2}} \exp\left(-\frac{\sigma\mu^2}{2(x^2+\sigma)}\right)\cdot -\mu}{\sqrt{\pi}\Gamma\left(\frac{\sigma}{2}\right) 2^{\frac{\sigma-1}{2}} (x^2+\sigma)^{\frac{\sigma+1}{2}}} \int_0^\infty y^\sigma \exp\left[-\frac{1}{2}\left(y - \frac{\mu x}{\sqrt{x^2+\sigma}}\right)^2\right] dy \\ &+ \frac{\sigma^{\frac{\sigma}{2}} \exp\left(-\frac{\sigma\mu^2}{2(x^2+\sigma)}\right) x}{\sqrt{\pi}\Gamma\left(\frac{\sigma}{2}\right) 2^{\frac{\sigma-1}{2}} (x^2+\sigma)^{\frac{\sigma+2}{2}}} \int_0^\infty y^{\sigma+1} \exp\left[-\frac{1}{2}\left(y - \frac{\mu x}{\sqrt{x^2+\sigma}}\right)^2\right] dy. \end{aligned} \quad \text{(A.2)}$$

The function for calculating the density of non-central t-distribution in R is based on the relationship between the density and its cumulative distribution function and the derivative is usually done by numerical integration. Here we present another way of calculating the density of t-distribution. Using this new way, the derivative of the density with respect to the non-centrality parameter can also be obtained in an easier way.

The essential problem for calculating (A.1) and (A.2) comes to evaluating the indefinite integral

$$\int_0^\infty y^a \exp\left(-\frac{(y-b)^2}{2}\right) dy,$$

can be rearranged as

$$\sqrt{2\pi}\int_{-b}^\infty (y+b)^a f(y;0,1)dy \tag{A.3}$$

where $f(y;\mu,\sigma^2)$ is the normal density with mean μ and variance σ^2. Expanding (A.3), we have

$$\begin{aligned}&\int_{-b}^\infty (y+b)^a f(y;0,1)dy\\ =&\sum_{i=0}^a \binom{a}{i} b^{a-i}\int_{-b}^\infty y^a f(y;0,1)dy.\end{aligned}$$

We utilise the following three equations in Owen (1980) to evaluate the integral,

$$\int_{-b}^\infty f(x;0,1)dx = 1 - F(-b;0,1)$$

$$\int_{-b}^\infty x^{2k+1} f(x;0,1)dx = f(-b;0,1)\sum_{j=0}^k \frac{(2k)!!}{(2j)!!}(-b)^{2j} \tag{A.4}$$

$$\int_{-b}^\infty x^{2k+2} f(x;0,1)dx = f(-b;0,1)\sum_{j=0}^k \frac{(2k+1)!!}{(2j+1)!!}(-b)^{2j+1} + (2k+1)!![1-F(-b;0,1)]. \tag{A.5}$$

In these integrals, $n!!$ is the double factorial: for even n it is equal to the product of all even numbers from 2 to n, and for odd n it is the product of all odd numbers from 1 to n; additionally it is assumed that $0!! = (-1)!! = 1$. The major difficulty in evaluating the non-central t-distribution numerically is to overcome the cancellation error. The analysis is carried out for the above three equations in order to reduce the cancellation error as much as possible. Terms in the right-hand side of (A.4) are positive as well as the second term in the right-hand side of (A.5). Terms in the summation of (A.5) have the same sign depending on the sign of b. Given b and σ, positive and negative terms are collected separately on the logarithm scale when calculating (A.3).

The method described above should be most efficient when the degree of freedom is small; for evaluating the density or its derivative when the degree of freedom is large, one can use numerical integration to arbitrary accuracy or use the normal distribution as an approximation.

Appendix B

Algorithms

This appendix contains the algorithms implemented in the programs used to carry out numerical studies in this thesis.

B.1 Computing the support points

Estimating a mixing distribution using various methods is implemented as a minimisation problem. As mentioned in Wang (2007) as well as the characterisation of the NPMDE and NPMLE shown in Lindsay (1983) and Chapter 5 respectively, at each iteration the new support points are the local minima of the gradient function based on the mixing distribution from the last iteration. The support range has been partitioned into m smaller grids $(a_1, a_2), (a_2, a_3), \cdots, (a_{m-1}, a_m)$ in order to find all the local minima. There exists a local minimum in the interval $[a, b]$ if the first derivative of the gradient function $d'(\mu, G_s) = \partial d(\mu; G_s)/\partial\mu$ evaluated at a is negative and positive at b. Note that there might be more than one local minimum in a particular grid, or there might be grids which contain a local minimum but can not be spotted by the derivative test. These are the problems existing in all minimisation or root-finding schemes and it is not of the interest in this Appendix. Depending on the differentiability of the component density and the difficulty in evaluating the derivatives of the density with respect to the location parameter, three different algorithms are used to compute the support points: the derivative-free minimisation algorithm that only depends on the values of the gradient function, the derivative-free root-finding algorithm that depends only on the first derivative of the gradient function $d'(\mu; G_s)$ and the first-order root-finding algorithm that depends on the first and the second derivative of the gradient function $d'(\mu; G_s)$ and $d''(\mu; G_s) = \partial^2 d(\mu; G_s)/\partial\mu^2$.

B.1.1 Derivative-free Minimisation Algorithm

A derivative-free minimisation algorithm is used when the gradient function is not differentiable, or difficult or time-consuming to evaluate. In order to find the grids containing local minima, finite differencing is used. In this case, we have to consider the wider interval (a_i, a_{i+2}), and hence the derivative test at a_i is replaced by the forward difference $d(a_{i+1}; G_s) - d(a_i; G_s)$ and the derivative test at a_{i+2} is replaced by the backward difference $d(a_{i+2}; G_s) - d(a_{i+1}; G_s)$. The reason to consider the interval (a_i, a_{i+2}) becomes clear: the sign changes at a_{i+1} depending on the choice of the differencing strategy. The derivative-free minimisation algorithm based on Brent (1973) implemented as the R function `optimise` is used to find the local minimum given an interval.

B.1.2 Derivative-free Root-finding Algorithm

The Brent method mentioned in Brent (1973) is a popular root-finding algorithm which combines the bisection, the secant and the inverse quadratic interpolation method. It has the reliability of the bisection method, but it can be as quick as some of the less-reliable methods. The algorithm tries to use the potentially fast-converging secant method or the inverse quadratic interpolation if possible, but it falls back to the more robust bisection method if necessary. The algorithm used in this thesis is based on the improvement of the Brent method by Zhang (2011) with slight modifications tailored to this specific problem. One iteration in the improvement made by Zhang (2011) is equivalent to two iterations in the original Brent's method, but the interval is reduced by more than half at each iteration, which can not be obtained by the original Brent method. The outline of the algorithm is as follows.

For any interval (a_i, a_{i+1}) with $d'(a_i; G_s) < 0$ and $d'(a_{i+1}; G_s) > 0$,

Step 1: Set $t = 0$, $l_0 = a_i$, $u_0 = a_{i+1}$.

Step 2: Let $c_t = (l_t + u_t)/2$. If $d'(l_t; G_s) \neq d'(c_t; G_s)$ and $d'(u_t; G_s) \neq d'(c_t; G_s)$,

$$\begin{aligned} r_t =& \frac{l_t d'(c_t; G_s) d'(u_t; G_s)}{[d'(l_t; G_s) - d'(c_t; G_s)][d'(l_t; G_s) - d'(u_t; G_s)]} \\ &+ \frac{c_t d'(l_t; G_s) d'(u_t; G_s)}{[d'(c_t; G_s) - d'(l_t; G_s)][d'(c_t; G_s) - d'(u_t; G_s)]} \\ &+ \frac{u_t d'(l_t; G_s) d'(c_t; G_s)}{[d'(u_t; G_s) - d'(l_t; G_s)][d'(u_t; G_s) - d'(c_t; G_s)]} \end{aligned}$$

(the inverse quadratic interpolation), and

$$r_t = u_t - d'(u_t; G_s)\frac{u_t - l_t}{d'(u_t; G_s) - d'(l_t; G_s)}$$

(the secant rule) if otherwise.

Step 3: If $d'(c_t; G_s) < 0$, set $l_{t+1} = c_t$, and set $u_{t+1} = c_t$ if otherwise. If $d'(r_t; G_s) < 0$, set $l_{t+1} = \max(l_{t+1}, r_t)$, and set $u_{t+1} = \min(u_{t+1}, r_t)$ if otherwise. Set $t = t + 1$ and go back to Step 2 until convergence.

B.1.3 First-order Root-finding Algorithm

The Newton-Raphson method is a popular root-finding algorithm that utilises the first derivative of the given function. The order of convergence is at least quadratic in a neighbourhood of the root, and hence the performance may be poor when the value in the current iteration is not in the neighbourhood. The algorithm used in this thesis is a combination of the Newton-Raphson and the bisection method, which takes the advantage of fast order of convergence of the Newton-Raphson method and uses the bisection method in reducing the size of the interval when the Newton-Raphson method performs poorly. The combined method ensures that the convergence is fast and that the interval is reduced by more than half at each iteration. The outline of the algorithm is as follows.

For any interval (a_i, a_{i+1}) with $d'(a_i; G_s) < 0$ and $d'(a_{i+1}; G_s) > 0$,

Step 1: Set $t = 1$, $l_0 = a_i$, $u_0 = a_{i+1}$ and r_0 be an initial guess.

Step 2: If $d'(r_{t-1}; G_s) < 0$, set $l_t = \max(l_{t-1}, r_{t-1})$, and set $u_t = \min(u_{t-1}, r_{t-1})$ if otherwise.

Step 3: Let $c_t = (l_t + u_t)/2$ and $r_t = r_{t-1} - d'(r_{t-1}; G_s)/d''(r_{t-1}; G_s)$.

Step 4: If $d'(c_t; G_s) < 0$, set $l_t = c_t$, and set $u_t = c_t$ if otherwise. If r_t is not in the range of (l_t, u_t), set $r_t = c_t$ and c_t will not be referred in the next iteration. Set $t = t + 1$ and go back to Step 2 until convergence.

B.2 Computing the Proportion of Zeros

In distance-based methods, finding $\hat{\pi}_0$ can always be formulated as computing the root of

$$\tilde{\ell}_n(\hat{\pi}_0) = q,$$

where $q > 0$ is a constant. In this section, computing the NPMLE is taken as computing the NPMDE with the KL divergence. As shown in Wang (2010) that the profile likelihood/loss function is differentiable with respect to $\hat{\pi}_0$ and it is a scalar multiple of the gradient function of the restricted NPMDE evaluated at 0. $\tilde{\ell}_n(\hat{\pi}_0)$ is always an increasing function under finite samples with $\tilde{\ell}_n(0) < q$ and $\tilde{\ell}_n(1) > q$. For the case $\tilde{\ell}_n(1) < q$, we take $\hat{\pi}_0 = 1$. The first-order root-finding algorithm in the previous section can be used. Further analysis shows that for the aforementioned distances in this thesis, the second directional derivative always exists and evaluation of the second directional derivative is always computationally cheaper than one additional step of computing the mixing distribution. The algorithm used for computing $\hat{\pi}_0$ in this thesis is a combination of the Halley and the bisection method, which takes the advantage of cubic order of convergence of the Halley method and uses the bisection method in reducing the size of the interval when the Halley method performs poorly. Efforts have been made in trying to obtain the reference of the original publication by Halley, but failed. For the description and the theoretical aspects of the method, see Alefeld (1981) and the references therein. The combined method ensures that the convergence is fast and that the interval is reduced by more than half at each iteration. The outline of the algorithm is as follows.

Let $\tilde{\ell}'_n(r_t)$ be the first derivative of $\tilde{\ell}_n(\pi_0)$ with respect to π_0 evaluated at $\pi_0 = r_t$ and $\tilde{\ell}''_n(r_t)$ be the second derivative of $\tilde{\ell}_n(\pi_0)$ with respect to π_0 evaluated at $\pi_0 = r_t$. For the interval $(0, 1)$ with $\tilde{\ell}_n(0) < q$ and $\tilde{\ell}_n(1) > q$,

Step 1: Set $t = 1$, $l_0 = 0$, $u_0 = 1$ and r_0 be an initial guess.

Step 2: If $\tilde{\ell}_n(r_t) < q$, set $l_t = \max(l_{t-1}, r_{t-1})$, and set $u_t = \min(u_{t-1}, r_{t-1})$ if otherwise.

Step 3: If r_t can not reduce the interval by at least half, then let $c_t = (l_t + u_t)/2$ and if $\tilde{\ell}_n(c_t) < q$, set $l_t = c_t$, and set $u_t = c_t$ if otherwise.

Step 4: Let

$$r_t = r_{t-1} - \frac{2\tilde{\ell}_n'(r_t)(\tilde{\ell}_n(r_t) - q)}{2[\tilde{\ell}_n'(r_t)]^2 - \tilde{\ell}_n''(r_t)(\tilde{\ell}_n(r_t) - q)},$$

and if r_t is not in the range of (l_t, u_t), then $r_t = (l_t + u_t)/2$. Set $t = t + 1$ and go back to Step 2 until convergence.

It is noted that the bisection method is only deployed when the interval produced by the Hally method can not be reduced by more than half at each iteration. Performing a bisection method is equivalent to estimating the mixing distribution. If the iteration is been reduced by more than half at one iteration, there is no need to perform a bisection method to further reduce the interval at the cost of estimating the mixing distribution one more time, especially at the early stage of the root-finding process.

When finding the NPMLE, it generally takes 5-6 steps for the Newton-Raphon method to converge while it only takes 4-5 steps for the Halley method to converge. In addition, after 1-2 iterations of the Helly's method, the intermediate solution is generally sufficiently close to the root and the mixing distribution from the previous iteration can be used to further accelerate the process of estimating the mixing distribution at current iteration.

Appendix C

The list of 100 publicly listed companies

This appendix contains the full list of the 100 publicly listed companies, of which the daily trading data from 31/10/2014 to 27/10/2017 is used in Section 9.4.

	Symbol	Full Name
1	AABA	Altaba Inc.
2	AAL	American Airlines Group, Inc.
3	AAME	Atlantic American Corporation
4	AAOI	Applied Optoelectronics, Inc.
5	AAON	AAON, Inc.
6	AAPL	Apple Inc.
7	AAWW	Atlas Air Worldwide Holdings
8	AAXJ	iShares MSCI All Country Asia ex Japan Index Fund
9	AAXN	Axon Enterprise, Inc.
10	ABAC	Renmin Tianli Group, Inc.
11	ABAX	ABAXIS, Inc.
12	ABCB	Ameris Bancorp
13	ABCD	Cambium Learning Group, Inc.
14	ABDC	Alcentra Capital Corp.
15	ABEO	Abeona Therapeutics Inc.
16	ABIO	ARCA biopharma, Inc.
17	ABMD	ABIOMED, Inc.
18	ACAD	ACADIA Pharmaceuticals Inc.
19	ACET	Aceto Corporation
20	ACFC	Atlantic Coast Financial Corporation

21	ACGL	Arch Capital Group Ltd.
22	ACHC	Acadia Healthcare Company, Inc.
23	ACHN	Achillion Pharmaceuticals, Inc.
24	ACHV	Achieve Life Sciences, Inc.
25	ACIW	ACI Worldwide, Inc.
26	ACLS	Axcelis Technologies, Inc.
27	ACNB	ACNB Corporation
28	ACOR	Acorda Therapeutics, Inc.
29	ACRX	AcelRx Pharmaceuticals, Inc.
30	ACSF	American Capital Senior Floating, Ltd.
31	ACST	Acasti Pharma, Inc.
32	ACTG	Acacia Research Corporation
33	ACWI	iShares MSCI ACWI Index Fund
34	ACWX	iShares MSCI ACWI ex US Index Fund
35	ACXM	Acxiom Corporation
36	ADBE	Adobe Systems Incorporated
37	ADES	Advanced Emissions Solutions, Inc.
38	ADI	Analog Devices, Inc.
39	ADMA	ADMA Biologics Inc
40	ADMP	Adamis Pharmaceuticals Corporation
41	ADMS	Adamas Pharmaceuticals, Inc.
42	ADP	Automatic Data Processing, Inc.
43	ADRA	BLDRS Asia 50 ADR Index Fund
44	ADRD	BLDRS Developed Markets 100 ADR Index Fund
45	ADRE	BLDRS Emerging Markets 50 ADR Index Fund
46	ADRU	BLDRS Europe 100 ADR Index Fund
47	ADSK	Autodesk, Inc.
48	ADTN	ADTRAN, Inc.
49	ADUS	Addus HomeCare Corporation
50	ADXS	Advaxis, Inc.
51	AEGN	Aegion Corp
52	AEHR	Aehr Test Systems
53	AEIS	Advanced Energy Industries, Inc.
54	AEMD	Aethlon Medical, Inc.
55	AERI	Aerie Pharmaceuticals, Inc.

56	AETI	American Electric Technologies, Inc.
57	AEY	ADDvantage Technologies Group, Inc.
58	AEZS	AEterna Zentaris Inc.
59	AFAM	Almost Family Inc
60	AFH	Atlas Financial Holdings, Inc.
61	AFMD	Affimed N.V.
62	AFSI	AmTrust Financial Services, Inc.
63	AGEN	Agenus Inc.
64	AGII	Argo Group International Holdings, Ltd.
65	AGIIL	Argo Group International Holdings, Ltd.
66	AGIO	Agios Pharmaceuticals, Inc.
67	AGNC	AGNC Investment Corp.
68	AGNCB	AGNC Investment Corp.
69	AGND	WisdomTree Barclays Negative Duration U.S. Aggregate Bond Fund
70	AGRX	Agile Therapeutics, Inc.
71	AGTC	Applied Genetic Technologies Corporation
72	AGYS	Agilysys, Inc.
73	AGZD	WisdomTree Barclays Interest Rate Hedged U.S. Aggregate Bond F
74	AHGP	Alliance Holdings GP, L.P.
75	AHPI	Allied Healthcare Products, Inc.
76	AIA	iShares Asia 50 ETF
77	AIMC	Altra Industrial Motion Corp.
78	AINV	Apollo Investment Corporation
79	AIRR	First Trust RBA American Industrial Renaissance ETF
80	AIRT	Air T, Inc.
81	AKAM	Akamai Technologies, Inc.
82	AKAO	Achaogen, Inc.
83	AKBA	Akebia Therapeutics, Inc.
84	AKER	Akers Biosciences Inc
85	AKRX	Akorn, Inc.
86	ALBO	Albireo Pharma, Inc.
87	ALCO	Alico, Inc.
88	ALDR	Alder BioPharmaceuticals, Inc.

89	ALDX	Aldeyra Therapeutics, Inc.
90	ALGN	Align Technology, Inc.
91	ALGT	Allegiant Travel Company
92	ALIM	Alimera Sciences, Inc.
93	ALJJ	ALJ Regional Holdings, Inc.
94	ALKS	Alkermes plc
95	ALLT	Allot Communications Ltd.
96	ALNY	Alnylam Pharmaceuticals, Inc.
97	ALOG	Analogic Corporation
98	ALOT	AstroNova, Inc.
99	ALQA	Alliqua BioMedical, Inc.
100	ALRN	Aileron Therapeutics, Inc.

Bibliography

Ahdesmaki, M., Zuber, V., Gibb, S., and Strimmer, K. (2015). *sda: Shrinkage Discriminant Analysis and CAT Score Variable Selection.* R package version 1.3.7.

Akaike, H (1974). "A new look at the statistical identification Model". In: *IEEE Trans. Auto. Control* 19, pp. 716–723.

Alefeld, G. (1981). "On the convergence of Halley's method". In: *The American Mathematical Monthly* 88, pp. 530–536.

Anderson, T. W. and Darling, D. A. (1952). "Asymptotic theory of certain "goodness of fit" criteria based on stochastic processes". In: *Ann. Math. Statist.* 23, pp. 193–212.

Anderson, T. W. and Darling, D. A. (1954). "A test of goodness of fit". In: *Journal of the American Statistical Association* 49, pp. 765–769.

Antoniadis, A. (1997). "Wavelets in statistics: A review". In: *Journal of the Italian Statistical Society* 6, pp. 97–130.

Antoniadis, A. and Fan, J. (2001). "Regularization of wavelet approximations". In: *Journal of the American Statistical Association* 96, pp. 939–967.

Arbuthnot, J. (1710). "An argument for divine providence, taken from the constant regularity observed in the births of both sexes". In: *Philosophical Transactions of the Royal Society of London* 27, pp. 186–190.

Armijo, L. (1966). "Minimization of functions having Lipschitz continuous first partial derivatives." In: *Pacific J. Math.* 16, pp. 1–3.

Aubel, A. van and Gawronski, W. (2003). "Analytic properties of noncentral distributions". In: *Applied Mathematics and Computation* 141, pp. 3–12.

Azriel, D. and Schwartzman, A. (2015). "The empirical distribution of a large number of correlated normal variables". In: *Journal of the American Statistical Association* 110, pp. 1217–1228.

Aït-Sahalia, Y. and Xiu, D. (2017). "Using principal component analysis to estimate a high dimensional factor model with high-frequency data". In: *Journal of Econometrics* 201, pp. 384–399.

Balabdaoui, F. and Wellner, J. A. (2004). *Estimation of a k-monotone density, part 2: Algorithms for computation and numerical results*. Tech. rep. 460. Department of Statistics, University of Washington.

Balabdaoui, F. and Wellner, J. A. (2010). "Estimation of a k-monotone density: characterizations, consistency and minimax lower bounds". In: *Statistica Neerlandica* 64, pp. 45–70.

Banerjee, O., El Ghaoui, L., and d'Aspremont, A. (2008). "Model selection through sparse maximum likelihood estimation for multivariate Gaussian or binary data". In: *Journal of Machine Learning Research* 9, pp. 485–516.

Bartlett, M. S. and Macdonald, P. D. M. (1968). "Least-squares estimation of distribution mixtures". In: *Nature* 217, pp. 195–196.

Behboodian, J. (1970). "On a mixture of normal distributions". In: *Biometrika* 57, pp. 215–217.

Benjamini, Y, Krieger, A., and Yekutieli, D (2006). "Adaptive linear step-up procedures that control the false discovery rate". In: *Biometrika* 93, pp. 491–507.

Benjamini, Y. and Hochberg, Y. (1995). "Controlling the false discovery rate: a practical and powerful approach to multiple testing". In: *Journal of the Royal Statistical Society Series B-Statistical Methodology* 57, pp. 289–300.

Bickel, P. J. and Levina, E. (2008a). "Covariance regularization by thresholding". In: *Annals of Statistics* 36, pp. 2577–2604.

Bickel, P. J. and Levina, E. (2008b). "Regularized estimation of large covariance matrices". In: *Annals of Statistics* 36, pp. 199–227.

Bien, J. and Tibshirani, R. J. (2011). "Sparse estimation of a covariance matrix". In: *Biometrika* 98, pp. 807–820.

Boos, D. D. (1981). "Minimum distance estimators for location and goodness of fit". In: *Journal of the American Statistical Association* 76, pp. 663–670.

Boos, D. D. (1982). "Minimum Anderson-Darling estimation". In: *Communications in Statistics* 11, pp. 2747–2774.

Brent, R. (1973). *Algorithms for Minimization without Derivatives*. Englewood Cliffs N.J.: Prentice-Hall.

Breusch, T. S. (1978). "Testing for autocorrelation in dynamic linear models". In: *Australian Economic Papers* 17, pp. 334–355.

Brockwell, P. J. and Davis, R. A. (2017). *Introduction to Time Series and Forecasting*. Cham, Switzerland: Springer.

Brown, L. D. (2008). "In-season prediction of batting averages: A field test of empirical Bayes and Bayes methodologies". In: *Annals of Applied Statistics* 2, pp. 113–152.

Brown, L. D. and Greenshtein, E. (2009). "Nonparametric empirical Bayes and compound decision approaches to estimation of a high-dimensional vector of normal means". In: *Annals of Statistics* 37, pp. 1685–1704.

Buydens, L., Blanchet, L., and Engel, J. (2017). "An overview of large-dimensional covariance and precision matrix estimators with applications in chemometrics". In: *Journal of Chemometrics* 31, pp. 1–19.

Böhning, D. (1985). "Numerical estimation of a probability measure". In: *Journal of Statistical Planning and Inference* 11, pp. 57–69.

Cai, T., Jin, J., and Low, M. G. (2007). "Estimation and confidence sets for sparse normal mixtures". In: *The Annals of Statistics* 35, pp. 2421–2449.

Cai, T. and Liu, W. D. (2011). "Adaptive thresholding for sparse covariance matrix estimation". In: *Journal of the American Statistical Association* 106, pp. 672–684.

Cai, T., Liu, W. D., and Luo, X. (2011). "A constrained ℓ_1 minimization approach to sparse precision matrix estimation". In: *Journal of the American Statistical Association* 106, pp. 594–607.

Cai, T., Ren, Z., and Zhou, H. H. (2013). "Optimal rates of convergence for estimating Toeplitz covariance matrices". In: *Probability Theory and Related Fields* 156, pp. 101–143.

Cai, T., Ren, Z., and Zhou, H. H. (2016). "Estimating structured high-dimensional covariance and precision matrices: Optimal rates and adaptive estimation". In: *Electronic Journal of Statistics* 10, pp. 1–59.

Cai, T. and Yuan, M. (2012). "Adaptive covariance matrix estimation through block thresholding". In: *Annals of Statistics* 40, pp. 2014–2042.

Cai, T., Zhang, C. H., and Zhou, H. H. (2010). "Optimal rates of convergence for covariance matrix estimation". In: *Annals of Statistics* 38, pp. 2118–2144.

Cai, T. and Zhou, H. H. (2012). "Minimax estimation of large covariance matrices under L1-norm". In: *Statistica Sinica*.

Candes, E. and Tao, T. (2007). "The Dantzig selector: Statistical estimation when p is much larger than n". In: *The Annals of Statistics* 35, pp. 2313–2351.

Cao, H. and Kosorok, M. R. (2011). "Simultaneous critical values for t-tests in very high dimensions". In: *Bernoulli* 17, pp. 347–394.

Carroll, R. J. and Hall, P. (1988). "Optimal rates of convergence for deconvolving a density". In: *Journal of the American Statistical Association* 83, pp. 1184–1186.

Chae, M., Martin, R., and Walker, S. G. (2018). "Convergence of an iterative algorithm to the nonparametric MLE of a mixing distribution". In: *Statistics & Probability Letters* 140, pp. 142–146.

Chaudhuri, S., Drton, M., and Richardson, T. S. (2007). "Estimation of a covariance matrix with zeros". In: *Biometrika* 94, pp. 199–216.

Chee, C.-S. and Wang, Y. (2013). "Minimum quadratic distance density estimation using nonparametric mixtures". In: *Computational Statistics & Data Analysis* 57, pp. 1–16.

Chee, C.-S. and Wang, Y. (2014). "Least squares estimation of a k-monotone density function". In: *Computational Statistics & Data Analysis* 74, pp. 209–216.

Chen, J. (1995). "Optimal rate of convergence for finite mixture models". In: *The Annals of Statistics* 23, pp. 221–233.

Chen, J. (2017). "Consistency of the MLE under mixture models". In: *Statist. Sci.* 32, pp. 47–63.

Chen, L. H. Y. (1975). "Poisson approximation for dependent trials". In: *The Annals of Probability* 3, pp. 534–545.

Chen, Y., Wiesel, A., Eldar, Y. C., and Hero, A. O. (2010). "Shrinkage algorithms for MMSE covariance estimation". In: *IEEE Transactions on Signal Processing* 58.10, 5016–5029.

Chiu, T. Y. M., Leonard, T., and Tsui, K. W. (1996). "The matrix logarithmic covariance model". In: *Journal of the American Statistical Association* 91, pp. 198–210.

Choi, K. and Bulgren, W. G. (1968). "An estimation procedure for mixtures of distributions". In: *Journal of the Royal Statistical Society. Series B (Methodological)* 30, pp. 444–460.

Choulakian, V., Lockhart, R. A., and Stephens, M. A. (1994). "Cramér-von Mises Statistics for discrete distributions". In: *The Canadian Journal of Statistics / La Revue Canadienne de Statistique* 22, pp. 125–137.

Conde, I. C. and Una Alvarez, J. de (2020). *sgof: Multiple Hypothesis Testing*. R package version 2.3.2.

Cramér, H. (1928). "On the composition of elementary errors. Second paper: Statistical Applications." In: *Skand. Aktuarietidskrift* 11, pp. 171–180.

Cramér, H (1946). *Mathematical Methods of Statistics*. Princeton: Princeton University Press.

Csorgo, S. and Faraway, J. J. (1996). "The exact and asymptotic distributions of Cramer-von Mises statistics". In: *Journal of the Royal Statistical Society. Series B (Methodological)* 58, pp. 221–234.

Cui, Y., Leng, C. L., and Sun, D. F. (2016). "Sparse estimation of high-dimensional correlation matrices". In: *Computational Statistics & Data Analysis* 93, pp. 390–403.

Dalal, O. and Rajaratnam, B. (2017). "Sparse Gaussian graphical model estimation via alternating minimization". In: *Biometrika* 104, pp. 379–395.

Dancho, M. and Vaughan, D. (2020). *alphavantager: Lightweight R Interface to the Alpha Vantage API*. R package version 0.1.2.

Darling, D. A. (1957). "The Kolmogorov-Smirnov, Cramer-von Mises tests". In: *The Annals of Mathematical Statistics* 28, pp. 823–838.

Dattner, I., Goldenshluger, A., and Juditsky, A. (2011). "On deconvolution of distribution functions". In: *Ann. Statist.* 39, pp. 2477–2501.

Dax, A. (1990). "The smallest point of a polytope". In: *Journal of Optimization Theory and Applications* 64, pp. 429–432.

Dempster, A. P., Laird, N. M., and Rubin, D. B. (1977). "Maximum likelihood from incomplete data via the EM algorithm". In: *Journal of the Royal Statistical Society. Series B (Methodological)* 39, pp. 1–38.

Deng, X. and Kang, X. (2020). "An improved modified cholesky decomposition approach for precision matrix estimation". In: *Journal of Statistical Computation and Simulation* 90, pp. 443–464.

Dey, D. K. and Srinivasan, C. (1985). "Estimation of a covariance matrix under Stein's loss". In: *The Annals of Statistics* 13, pp. 1581–1591.

Doane, D. P. (1976). "Aesthetic frequency classifications". In: *The American Statistician* 30, pp. 181–183.

Donoho, D. (2000). "High-dimensional data analysis: The curses and blessings of dimensionality". In: *AMS Math Challenges Lecture*, pp. 1–32.

Dudoit, S., Yang, Y. H., Callow, M. J., and Speed, T. P. (2002). "Statistical methods for identifying differentially expressed genes in replicated cDNA microarray experiments". In: *Statistica Sinica* 12, pp. 111–139.

Dümbgen, L. and Rufibach, K. (2011). "logcondens: Computations related to univariate log-concave density estimation". In: *Journal of Statistical Software* 39, pp. 1–28.

Duncan, D. B. (1955). "Multiple range and multiple F tests". In: *Biometrics* 11, pp. 1–42.

Durbin, J. and Watson, G. S. (1950). "Testing for serial correlation in least squares regression: I". In: *Biometrika* 37, pp. 409–428.

Durbin, J. and Watson, G. S. (1951). "Testing for serial correlation in least squares regression. II". In: *Biometrika* 38, pp. 159–177.

Durbin, J. and Watson, G. S. (1971). "Testing for serial correlation in least squares regression.III". In: *Biometrika* 58, pp. 1–19.

Efron, B. (2003). "Robbins, empirical Bayes and microarrays". In: *Ann. Statist.* 31, pp. 366–378.

Efron, B. (2004). "Large-scale simultaneous hypothesis testing". In: *Journal of the American Statistical Association* 99, pp. 96–104.

Efron, B. (2007a). "Correlation and large-scale simultaneous significance testing". In: *Journal of the American Statistical Association* 102, pp. 93–103.

Efron, B. (2007b). "Size, power and false discovery rates". In: *Ann. Statist.* 35, pp. 1351–1377.

Efron, B. (2010a). "Correlated z-values and the accuracy of large-scale statistical estimates". In: *J Am Stat Assoc* 105, pp. 1042–1055.

Efron, B. (2010b). *Large Scale Inference: Empirical Bayes Methods for Estimation, Testing, and Prediction*. Cambridge: Cambridge University Press.

Efron, B. and Morris, C. (1973). "Stein's estimation rule and its competitors: An empirical Bayes approach". In: *Journal of the American Statistical Association* 68, pp. 117–130.

Efron, B., Tibshirani, R., Storey, J. D., and Tusher, V. (2001). "Empirical Bayes analysis of a microarray experiment". In: *Journal of the American Statistical Association* 96, pp. 1151–1160.

Efron, B., Turnbull, B., and Narasimhan, B. (2015). *locfdr: Computes Local False Discovery Rates*. R package version 1.1-8.

Fan, J. (1997). "Comments on "wavelets in statistics: A review" by A. Antoniadis". In: *Journal of the Italian Statistical Society* 6, pp. 131–138.

Fan, J. and Jiang, J. (2007). "Nonparametric inference with generalized likelihood ratio tests". In: *Test; Heidelberg* 16, pp. 409–444.

Fan, J. and Li, R. Z. (2001). "Variable selection via nonconcave penalized likelihood and its oracle properties". In: *Journal of the American Statistical Association* 96, pp. 1348–1360.

Fan, J., Liao, Y., and Liu, H. (2016). "An overview of the estimation of large covariance and precision matrices". In: *Econometrics Journal* 19, pp. C1–C32.

Fan, J., Liao, Y., and Mincheva, M. (2013). "Large covariance estimation by thresholding principal orthogonal complements". In: *J R Stat Soc Series B Stat Methodol* 75, pp. 603–680.

Fan, J. and Liu, H. (2013). "Statistical analysis of big data on pharmacogenomics". In: *Adv Drug Deliv Rev* 65, pp. 987–1000.

Fang, Y., Wang, B., and Feng, Y. (2016). "Tuning-parameter selection in regularized estimations of large covariance matrices". In: *Journal of Statistical Computation and Simulation* 86, pp. 494–509.

Fedorov, V. (1972). *Theory of Optimal Experiments Designs(Engl. transl.)* New York: Academic Press.

Fisher, R. A. (1915). "Frequency distribution of the values of the correlation coefficient in samples from an indefinitely large population". In: *Biometrika* 10, pp. 507–521.

Fisher, R. A. (1925). *Statistical Methods for Research Workers.* Oliver & Boyd (Edinburgh).

Frahm, G., Wickern, T., and Wiechers, C. (2012). "Multiple tests for the performance of different investment strategies". In: *Advances in Statistical Analysis* 96, pp. 343–383.

Freedman, D. and Diaconis, P. (1981). "On the histogram as a density estimator: L_2 theory". In: *Zeitschrift für Wahrscheinlichkeitstheorie und Verwandte Gebiete* 57, pp. 453–476.

Friedman, J., Hastie, T., and Tibshirani, R. (2008). "Sparse inverse covariance estimation with the graphical lasso". In: *Biostatistics* 9, pp. 432–441.

Gablonsky, J. M. and Kelley, C. T. (2001). "A locally-biased form of the DIRECT Algorithm". In: *Journal of Global Optimization* 21, pp. 27–37.

Gavrilov, Y., Benjamini, Y., and Sarkar, S. K. (2009). "An adaptive step-down procedure with proven FDR control under independence". In: *The Annals of Statistics* 37, pp. 619–629.

Geman, S. and Hwang, C.-R. (1982). "Nonparametric maximum likelihood estimation by the method of sieves". In: *Ann. Statist.* 10, pp. 401–414.

Genovese, C. R. and Wasserman, L. (2004). "A stochastic process approach to false discovery control". In: *Ann. Statist.* 32, pp. 1035–1061.

Ghosal, S. and van der Vaart, A. W. (2001). "Entropies and rates of convergence for maximum likelihood and Bayes estimation for mixtures of normal densities". In: *The Annals of Statistics* 29, pp. 1233–1263.

Godfrey, L (1978). "Testing against general autoregressive and moving average error models when the regressors include lagged dependent variables". In: *Econometrica* 46, p. 1293.

Goeman, J. J. and Solari, A. (2014). "Multiple hypothesis testing in genomics". In: *Statistics in Medicine* 33, pp. 1946–1978.

Golub, T. R., Slonim, D. K., Tamayo, P., Huard, C., et al. (1999). "Molecular classification of cancer: Class discovery and class prediction by gene expression monitoring". In: *Science* 286, pp. 531–537.

Goulart, P. J., Nakatsukasa, Y., and Rontsis, N. (2020). "Accuracy of approximate projection to the semidefinite cone". In: *Linear Algebra and Its Applications* 594, pp. 177–192.

Greenshtein, E. and Ritov, Y. (2009). "Asymptotic efficiency of simple decisions for the compound decision problem". In: *Optimality*. Vol. 57. Lecture Notes–Monograph Series. Institute of Mathematical Statistics, pp. 266–275.

Groeneboom, P., Jongbloed, G., and Wellner, J. A. (2008). "The support reduction algorithm for computing non-parametric function estimates in mixture models". In: *Scandinavian Journal of Statistics* 35, pp. 385–399.

Hall, P. (1981). "On the non-parametric estimation of mixture proportions". In: *Journal of the Royal Statistical Society. Series B (Methodological)* 43, pp. 147–156.

Hall, P. and Titterington, D. M. (1984). "Efficient nonparametric estimation of mixture proportions". In: *Journal of the Royal Statistical Society. Series B (Methodological)* 46, pp. 465–473.

Hannan, E. J. and Quinn, B. G. (1979). "The determination of the order of an autoregression". In: *Journal of the Royal Statistical Society, Series B* 41, pp. 190–195.

Heinrich, P. and Kahn, J. (2018). "Strong identifiability and optimal minimax rates for finite mixture estimation". In: *Ann. Statist.* 46, pp. 2844–2870.

Henna, J. (1983). "A note on a consistent estimator of a mixing distribution function". In: *Ann. Inst. Statist. Math.* 35, pp. 229–233.

Higham, N. J. (2002). "Computing the nearest correlation matrixa problem from finance". In: *IMA Journal of Numerical Analysis* 22, pp. 329–343.

Hochberg, Y. (1988). "A sharper Bonferroni procedures for multiple tests of significance". In: *Biometrika* 75, pp. 800–802.

Holm, S. (1979). "A simple sequentially rejective multiple test procedure". In: *Scandinavian Journal of Statistics* 6, pp. 65–70.

Hommel, G. (1988). "A stagewise rejective multiple test procedure based on a modified Bonferroni test". In: *Biometrika* 75, pp. 383–386.

Hotelling, H. (1953). "New light on the correlation coefficient and its transforms". In: *Journal of the Royal Statistical Society Series B-Statistical Methodology* 15, pp. 193–232.

Huang, C., Wang, Y., Li, X., Ren, L., et al. (2020). "Clinical features of patients infected with 2019 novel coronavirus in Wuhan, China". In: *The Lancet* 395, pp. 497–506.

Hubbard, R., Parsa, R. A., and Luthy, M. R. (1997). "The spread of statistical significance testing in psychology: The case of the Journal of Applied Psychology, 1917-1994". In: *Theory & Psychology* 7, pp. 545–554.

Ibragimov, I. A. and Has'minskii, R. (1981). *Statistical Estimation–Asymptotic Theory.* Springer-Verlag.

James, W. and Stein, C. (1961). "Estimation with quadratic loss". In: *Proceedings of the Fourth Berkeley Symposium on Mathematical Statistics and Probability, Volume 1: Contributions to the Theory of Statistics.* Berkeley, Calif.: University of California Press, pp. 361–379.

Jewell, N. P. (1982). "Mixtures of exponential distributions". In: *Annals of Statistics* 10, pp. 479–484.

Jiang, B. and Wang, C. (2020). "An efficient ADMM algorithm for high dimensional precision matrix estimation via penalized quadratic loss". In: *Computational Statistics & Data Analysis.* 142, pp. 1–12.

Jiang, W. and Zhang, C. H. (2009). "General maximum likelihood empirical Bayes estimation of normal means". In: *Annals of Statistics* 37, pp. 1647–1684.

Jiang, W. and Zhang, C. H. (2016). "Generalized likelihood ratio test for normal mixtures". In: *Statistica Sinica.* 26, pp. 955–978.

Jiang, W. and Zhang, C. H. (2019). "Rate of divergence of the nonparametric likelihood ratio test for Gaussian mixtures". In: *Bernoulli* 25, pp. 3400–3420.

Jin, J. (2008). "Proportion of non-zero normal means: universal oracle equivalences and uniformly consistent estimators". In: *Journal of the Royal Statistical Society: Series B (Statistical Methodology)* 70, pp. 461–493.

Jin, J. and Cai, T. (2007). "Estimating the null and the proportion of nonnull effects in large-scale multiple comparisons". In: *Journal of the American Statistical Association* 478, pp. 495–506.

Jones, D. R., Perttunen, C. D., and Stuckman, B. E. (1993). "Lipschitzian optimization without the Lipschitz constant". In: *J. Optim. Theory Appl.* 79, pp. 157–181.

Kang, J. (2020). "Two-stage false discovery rate in microarray studies". In: *Communications in Statistics* 49, pp. 894–908.

Karlis, D. and Xekalaki, E. (1998). "Minimum Hellinger distance estimation for Poisson mixtures". In: *Computational Statistics & Data Analysis* 29, pp. 81–103.

Karunamuni, R. J., Wu, J., and Karunamuni, R. J. (2009). "On minimum Hellinger distance estimation". In: *The Canadian Journal of Statistics* 37, pp. 514–533.

Kiefer, J. and Wolfowitz, J. (1956). "Consistency of the maximum likelihood estimator in the presence of infinitely many incidental parameters". In: *The Annals of Mathematical Statistics* 27, pp. 887–906.

Kolmogorov, A. (1933). "Sulla determinazione empirica di una legge di distribuzione". In: *1st. Ital. Attuari, G.* 4, pp. 1–11.

Kubokawa, T., Robert, C., and Saleh, A. K. M. E. (1992). "Empirical Bayes estimation of the covariance matrix of a normal distribution with unknown mean under an entropy loss". In: *The Indian Journal of Statistics, Series A* 54, pp. 402–410.

Kullback, S. and Leibler, R. A. (1951). "On information and sufficiency". In: *The Annals of Mathematical Statistics* 22, pp. 79–86.

Laird, N. (1978). "Nonparametric maximum likelihood estimation of a mixing distribution". In: *Journal of the American Statistical Association* 73, pp. 805–811.

Lam, C. and Fan, J. (2009). "Sparsistency and rates of convergence in large covariance matrix estimation". In: *Ann Stat* 37, pp. 4254–4278.

Lam, C. (2020). "High-dimensional covariance matrix estimation". In: *WIREs Computational Statistics* 12, pp. 1–21.

Langaas, M., Lindqvist, B. H., and Ferkingstad, E. (2005). "Estimating the proportion of true null hypotheses, with application to DNA microarray data". In: *Journal of the Royal Statistical Society: Series B (Statistical Methodology)* 67, pp. 555–572.

Lange, K., Hunter, D. R., and Yang, I. (2000). "Optimization transfer using surrogate objective functions". In: *Journal of Computational and Graphical Statistics* 9, pp. 1–20.

Lawson, C. L. and Hanson, R. J. (1974). *Solving Least Squares Problems*. Englewood Cliffs: Prentice-Hall.

Ledoit, O. and Wolf, M. (2004). "A well-conditioned estimator for large-dimensional covariance matrices". In: *Journal of Multivariate Analysis* 88, pp. 365–411.

Ledoit, O. and Wolf, M. (2018). "Optimal estimation of a large-dimensional covariance matrix under Stein's loss". In: *Bernoulli* 24, pp. 3791–3832.

Lee, K. and You, K. (2019). *CovTools: Statistical Tools for Covariance Analysis*. R package version 0.5.3.

Lehmann, E. L. and Romano, J. P. (2005). "Generalizations of the familywise error rate". In: *The Annals of Statistics* 33, pp. 1138–1154.

Leroux, B. G. (1992). "Consistent estimation of a mixing distribution". In: *Ann. Statist.* 20, pp. 1350–1360.

Lesperance, M. and Kalbfleisch, J. D. (1992). "An algorithm for computing the nonparametric MLE of a mixing distribution". In: *Journal of the American Statistical Association* 87, pp. 120–126.

Lesperance, M., Saab, R., and Neuhaus, J. (2014). "Nonparametric estimation of the mixing distribution in logistic regression mixed models with random intercepts and slopes". In: *Computational Statistics & Data Analysis* 71, pp. 211–219.

Lewis, P. A. W. (1961). "Distribution of the Anderson-Darling statistic". In: *The Annals of Mathematical Statistics* 32, pp. 1118–1124.

Li, A. and Barber, R. F. (2017). "Accumulation tests for FDR control in ordered hypothesis testing". In: *Journal of the American Statistical Association* 112, pp. 837–849.

Li, D. and Zou, H. (2016). "SURE information criteria for large covariance matrix estimation and their asymptotic properties". In: *IEEE Transactions on Information Theory* 62, pp. 2153–2169.

Liang, K. and Nettleton, D. (2012). "Adaptive and dynamic adaptive procedures for false discovery rate control and estimation". In: *Journal of the Royal Statistical Society: Series B (Statistical Methodology)* 74, pp. 163–182.

Lindquist, M. A. (2008). "The statistical analysis of fMRI data". In: *Statist. Sci.* 23, pp. 439–464.

Lindsay, B. G. (1983). "The geometry of mixture likelihoods - a general theory". In: *Annals of Statistics* 11, pp. 86–94.

Lindsay, B. G. (1994). "Efficiency versus robustness: The case for minimum Hellinger distance and related methods". In: *The Annals of Statistics* 22, pp. 1081–1114.

Lindsay, B. G. (1995). *Mixture Models: Theory, Geometry and Applications*. Institute of Mathematical Statistics.

Lindsay, B. G. and Roeder, K. (1993). "Uniqueness of estimation and identifiability in mixture models". In: *The Canadian Journal of Statistics / La Revue Canadienne de Statistique* 21, pp. 139–147.

Liu, H. and Wang, L. (2017). "TIGER: A tuning-insensitive approach for optimally estimating Gaussian graphical models". In: *Electronic Journal of Statistics* 11, pp. 241–294.

Liu, H., Wang, L., and Zhao, T. (2014). "Sparse covariance matrix estimation with eigenvalue constraints". In: *J Comput Graph Stat* 23, pp. 439–459.

Liu, X and Shao, Y. (2003). "Asymptotics for likelihood ratio tests under loss of identifiability". In: *Annals Of Statistics* 31, pp. 807–832.

Liu, Y. and Wang, Y. (2018). "A fast algorithm for univariate log-concave density estimation". In: *Australian & New Zealand Journal of Statistics* 60, pp. 258–275.

Liyanage, J. S. S., Qiu, Y., and Liyanage, J. S. S. (2019). "Threshold selection for covariance estimation". In: *Biometrics* 75, pp. 895–905.

Ljung, G. M. and Box, G. E. P. (1978). "On a measure of lack of fit in time series models". In: *Biometrika* 65, pp. 297–303.

Lu, M. and Stephens, M. (2019). "Empirical Bayes estimation of normal means, accounting for uncertainty in estimated standard errors". In: *arXiv.org:1901.10679*.

Lyons, R. (1988). "Strong laws of large numbers for weakly correlated random variables". In: *Michigan Math. J.* 35.

Marcus, R., Peritz, E., and Gabriel, K. R. (1976). "On closed testing procedures with special reference to ordered analysis of variance". In: *Biometrika* 63, pp. 655–660.

Marsaglia, J. and Marsaglia, G. (2004). "Evaluating the Anderson-Darling distribution". In: *Journal of Statistical Software* 9.

Massey, F. J. (1950). "A note on the power of a non-parametric test". In: *Ann. Math. Statist.* 21, pp. 440–443.

Meinshausen, N. and Buhlmann, P. (2006). "High-dimensional graphs and variable selection with the Lasso". In: *Annals of Statistics* 34, pp. 1436–1462.

Meinshausen, N. and Rice, J. (2006). "Estimating the proportion of false null hypotheses among a large number of independently tested hypotheses". In: *The Annals of Statistics* 34, pp. 373–393.

Muralidharan, O. (2010). "An empirical Bayes mixture method for effect size and false discovery rate estimation". In: *Ann. Appl. Stat.* 4, pp. 422–438.

Muralidharan, O. and Efron, B. (2012). *mixfdr: Computes False Discovery Rates and Effect Dizes using Normal Mixtures*. R package version 1.0.

Murphy, K. P. (2012). *Machine Learning: A Probabilistic Perspective*. Cambridge, MA: MIT press.

Murphy, S. A. and van der Vaart, A. W. (2000). "On profile likelihood". In: *Journal of the American Statistical Association* 95, pp. 449–465.

Neuvial, P. (2008). "Asymptotic properties of false discovery rate controlling procedures under independence". In: *Electronic Journal of Statistics* 2, pp. 1065–1110.

Newton, M. A. (2002). "On a nonparametric recursive estimator of the mixing distribution". In: *Sankhyā: The Indian Journal of Statistics, Series A* 64, pp. 306–322.

Neyman, J. and Pearson, E. S. (1933). "On the problem of the most efficient tests of statistical hypotheses". In: *Philosophical Transactions of the Royal Society A* 231, pp. 289–339.

Nguyen, X. (2013). "Convergence of latent mixing measures in finite and infinite mixture models". In: *Ann. Statist.* 41, pp. 370–400.

Owen, D. B. (1980). "A table of normal integrals". In: *Communications in Statistics - Simulation and Computation* 9, pp. 389–419.

Oyeniran, O. and Chen, H. (2016). "Estimating the Proportion of True Null Hypotheses in Multiple Testing Problems". In: *Journal of Probability and Statistics* 2016, pp. 1–7.

Pang, H., Liu, H., and Vanderbei, R. (2014). "The fastclime package for linear programming and large-scale precision matrix estimation in R". In: *J Mach Learn Res* 15, pp. 489–493.

Parr, W. C. and DeWet, T. (1981). "On minimum cramer-von mises-norm parameter estimation". In: *Communications in Statistics - Theory and Methods* 10, pp. 1149–1166.

Parr, W. C. and Schucany, W. R. (1980). "Minimum distance and robust estimation". In: *Journal of the American Statistical Association* 75, pp. 616–624.

Petrone, S., Rousseau, J., and Scricciolo, C. (2014). "Bayes and empirical Bayes: do they merge?" In: *Biometrika* 101, pp. 285–302.

Pfanzagl, J. (1988). "Consistency of maximum-likelihood estimators for certain nonparametric families, in particular - mixtures". In: *Journal of Statistical Planning and Inference* 19, pp. 137–158.

Pourahmadi, M. (2007). "Cholesky decompositions and estimation of a covariance matrix: orthogonality of variance-correlation parameters". In: *Biometrika* 94, pp. 1006–1013.

Pourahmadi, M. (2011). "Covariance estimation: The GLM and regularization perspectives". In: *Statistical Science* 26, pp. 369–387.

Pourahmadi, M. (2013). *High-dimensional Covariance Estimation*. Wiley Series in Probability and Statistics. Hoboken, New Jersey: Wiley.

Qi, H. D. and Sun, D. F. (2006). "A quadratically convergent Newton method for computing the nearest correlation matrix". In: *Siam Journal on Matrix Analysis and Applications* 28, pp. 360–385.

R Core Team (2020). *R: A Language and Environment for Statistical Computing*. R Foundation for Statistical Computing. Vienna, Austria.

Rajaratnam, B. and Vincenzi, D. (2016). "A theoretical study of Stein's covariance estimator". In: *Biometrika* 103, pp. 653–666.

Ramdas, A. K., Barber, R. F., Wainwright, M. J., and Jordan, M. I. (2019). "A unified treatment of multiple testing with prior knowledge using the p-filter". In: *Ann. Statist.* 47, pp. 2790–2821.

Ráo, C. (1945). "Information and the accuracy attainable in the estimation of statistical parameters". In: *Bulletin of Calcutta Mathematical Society* 37, pp. 81–91.

Robbins, H. (1951). "Asymptotically subminimax solutions of compound statistical decision problems". In: *Proceedings of the Second Berkeley Symposium on Mathematical Statistics and Probability*. Berkeley, Calif.: University of California Press, pp. 131–149.

Robbins, H. (1956). "An empirical Bayes approach to statistics". In: *Proceedings of the Third Berkeley Symposium on Mathematical Statistics and Probability, Volume 1: Contributions to the Theory of Statistics*. Berkeley, Calif.: University of California Press, pp. 157–163.

Robbins, H. (1964). "The empirical Bayes approach to statistical decision problems". In: *The Annals of Mathematical Statistics* 35, pp. 1–20.

Robbins, H. (1983). "Some thoughts on empirical Bayes estimation". In: *The Annals of Statistics* 11, pp. 713–723.

Romano, J. P. and Wolf, M. (2007). "Control of generalized error rates in multiple testing". In: *The Annals of Statistics* 35, pp. 1378–1408.

Rothman, A. J., Levina, E., and Zhu, J. (2009). "Generalized thresholding of large covariance matrices". In: *Journal of the American Statistical Association* 104, pp. 177–186.

Roy, D. (2003). "The discrete normal distribution". In: *Communications in Statistics - Theory and Methods* 32, pp. 1871–1883.

Saha, S. and Guntuboyina, A. (2020). "On the nonparametric maximum likelihood estimator for Gaussian location mixture densities with application to Gaussian denoising". In: *Ann. Statist.* 48, pp. 738–762.

Saville, D. J (1990). "Multiple comparison procedures: The practical solution". In: *The American Statistician* 44, pp. 174–180.

Schwarz, G. E. (1978). "Estimating the dimension of a model". In: *Annals of Statistics* 6, pp. 461–464.

Scott, D. W. (1979). "On optimal and data-based histograms". In: *Biometrika* 66, pp. 605–610.

Scott, D. W. (1981). "Using computer-binned data for density estimation". In: *Computer Science and Statistics: Proceedings of the 13th Symposium on the Interface*. New York, NY: Springer US, pp. 292–294.

Scott, D. W. (1992). *Multivariate Density Estimation : Theory, Practice, and Visualization*. Wiley series in probability and mathematical statistics. New York: Wiley.

Shimazaki, H. and Shinomoto, S. (2007). "A method for selecting the bin size of a time histogram". In: *Neural Computation* 19, pp. 1503–1527.

Simar, L. (1976). "Maximum likelihood estimation of a compound poisson process". In: *Ann. Statist.* 4, pp. 1200–1209.

Simes, R. J. (1986). "An improved Bonferroni procedure for multiple tests of significance". In: *Biometrika* 73, pp. 751–754.

Singh, D., Febbo, P. G., Ross, K., Jackson, D. G., et al. (2002). "Gene expression correlates of clinical prostate cancer behavior". In: *Cancer Cell* 1, pp. 203–209.

Smirnov, N. (1948). "Table for estimating the goodness of fit of empirical distributions". In: *Ann. Math. Statist.* 19, pp. 279 –281.

Stein, C. (1956). "Inadmissibility of the usual estimator for the mean of a multivariate normal distribution". In: *Proceedings of the Third Berkeley Symposium on Mathematical Statistics and Probability, Volume 1: Contributions to the Theory of Statistics*. Berkeley, Calif.: University of California Press, pp. 197–206.

Stone, C. J. (1984). "An asymptotically optimal window selection rule for kernel density estimates". In: *The Annals of Statistics* 12, pp. 1285–1297.

Storey, J. D. (2002a). "A direct approach to false discovery rates". In: *Journal of the Royal Statistical Society. Series B (Statistical Methodology)* 64, pp. 479–498.

Storey, J. D. (2002b). "False discovery rates: Theory and applications to DNA microarrays". PhD thesis. Stanford University.

Storey, J. D. and Tibshirani, R. (2003). "Statistical significance for genomewide studies". In: *Proceedings of the National Academy of Sciences* 100, pp. 9440–9445.

Sturges, H. A. (1926). "The choice of a class interval". In: *Journal of the American Statistical Association* 21, pp. 65–66.

Sun, T. N. and Zhang, C. H. (2012). "Scaled sparse linear regression". In: *Biometrika* 99, pp. 879–898.

Sun, T. N. and Zhang, C. H. (2013). "Sparse matrix inversion with scaled Lasso". In: *Journal of Machine Learning Research* 14, pp. 3385–3418.

Sun, W. and Mclain, A. C. (2012). "Multiple testing of composite null hypotheses in heteroscedastic models". In: *Journal of the American Statistical Association* 107, pp. 673–687.

Tibshirani, R. (1996). "Regression shrinkage and selection via the Lasso". In: *Journal of the Royal Statistical Society Series B-Statistical Methodology* 56 (1), pp. 267–288.

Titman, A. C. (2017). "Non-parametric maximum likelihood estimation of interval-censored failure time data subject to misclassification". In: *Statistics and Computing* 27, pp. 1585–1593.

Titterington, D. M. (1983). "Minimum distance non-parametric estimation of mixture proportions". In: *Journal of the Royal Statistical Society. Series B (Methodological)* 45, pp. 37–46.

Tsukuma, H. (2016). "Estimation of a high-dimensional covariance matrix with the Stein loss". In: *Journal of Multivariate Analysis* 148, pp. 1–17.

Tukey, J. W. (1953). *The Problem of Multiple Comparisons*. Mimeographed monograph.

Tusher, V. G., Tibshirani, R, and Chu, G (2001). "Significance analysis of microarrays applied to the ionizing radiation response." In: *Proceedings of the National Academy of Sciences of the United States of America* 98, pp. 5116–5121.

van de Geer, S. (1996). "Rates of convergence for the maximum likelihood estimator in mixture models". In: *Journal of Nonparametric Statistics* 6, pp. 293–310.

van der Vaart, A. (1998). *Asymptotic Statistics*. Cambridge Series in Statistical and Probabilisitic Mathematics. Cambridge University Press.

von Mises, R. (1931). *Vorlesungen aus dem Gebiete der Angewandten Mathematik*. Vol 1., Wahrscheinlichkeitsrechnung und ihre Anwendung in der Statistik und theoretischen Physik. Springer: Wien.

Wald, A. (1949). "Note on the consistency of the maximum likelihood estimate". In: *The Annals of Mathematical Statistics* 20, pp. 595–601.

Wang, B. (2014). *CVTuningCov: Regularized Estimators of Covariance Matrices with CV Tuning*. R package version 1.0.

Wang, Y. and Zou, J. (2010). "Vast volatility matrix estimation for high-frequency financial data". In: *Ann. Statist.* 38.2, pp. 943–978.

Wang, Y. (2007). "On fast computation of the non-parametric maximum likelihood estimate of a mixing distribution". In: *Journal of the Royal Statistical Society Series B-Statistical Methodology* 69, pp. 185–198.

Wang, Y. (2010). "Maximum likelihood computation for fitting semiparametric mixture models". In: *Statistics and Computing* 20, pp. 75–86.

Wang, Y. (2017). *nspmix: Nonparametric and Semiparametric Mixture Estimation*. R package version 1.4-0.

Wang, Y. (2019). "Computation of the nonparametric maximum likelihood estimate of a univariate log-concave density". In: *WIREs Computational Statistics* 11, pp. 1–9.

Wang, Y. and Fani, S. (2018). "Nonparametric maximum likelihood computation of a U-shaped hazard function". In: *Statistics and Computing* 28, 187–200.

Wang, Y., Lawson, C. L., and Hanson, R. J. (2020). *lsei: Solving Least Squares or Quadratic Programming Problems under Equality/Inequality Constraints*. R package version 1.2-0.1.

Wang, Y. and Stephen, M. T. (2013). "Efficient computation of nonparametric survival functions via a hierarchical mixture formulation". In: *Statistics and Computing* 23, 713–725.

Wang, Y. and Wang, X. (2015). "Nonparametric multivariate density estimation using mixtures". In: *Statistics and computing* 25, pp. 349–364.

Weinstein, A., Ma, Z., Brown, L. D., and Zhang, C. H. (2018). "Group-linear empirical Bayes estimates for a heteroscedastic normal mean". In: *Journal of the American Statistical Association*, pp. 1–13.

Westfall, P. H., Young, S. S., and Wright, S. P. (1993). "On adjusting p-values for multiplicity". In: *Biometrics* 49, pp. 941–944.

Wilcox, J. (2012). "Some aspects of non parametric maximum likelihood for normal mixtures". PhD thesis. University of Sydney.

Won, J., Lim, J., Kim, S., and Rajaratnam, B. (2013). "Condition-number-regularized covariance estimation". In: *Journal of the Royal Statistical Society: Series B (Statistical Methodology)* 75, pp. 427–450.

Woodward, W. A., Parr, W. C., Schucany, W. R., and Lindsey, H. (1984). "A comparison of minimum distance and maximum likelihood estimation of a mixture proportion". In: *Journal of the American Statistical Association* 79, pp. 590–598.

Wu, J., Karunamuni, R., and Wu, J (2009). "Minimum Hellinger distance estimation in a nonparametric mixture model". In: *Journal of Statistical Planning and Inference.* 139, pp. 1118–1133.

Wu, W. B. and Pourahmadi, M. (2009). "Banding sample autocovariance matrices of stationary processes". In: *Statistica Sinica* 19, pp. 1755–1768.

Ypma, J., Johnson, S. G., Borchers, H. W., Eddelbuettel, D., et al. (2020). *nloptr: R Interface to NLopt*. R package version 1.2.2.2.

Yuan, M. (2010). "High dimensional inverse covariance matrix estimation via linear programming". In: *Journal of Machine Learning Research* 11, pp. 2261–2286.

Yuan, M. and Lin, Y. (2007). "Model selection and estimation in the Gaussian graphical model". In: *Biometrika* 94, pp. 19–35.

Zeileis, A. and Hothorn, T. (2002). "Diagnostic checking in regression relationships". In: *R News* 2, pp. 7–10.

Zhang, C. H. (1997). "Empirical Bayes and compound estimation of normal means". In: *Statistica Sinica* 7, pp. 181–193.

Zhang, C. H. (2009). "Generalized maximum likelihood estimation of normal mixture densities". In: *Statistica Sinica* 19, pp. 1297–1318.

Zhang, C. H. (2010). "Nearly unbiased variable selection under minimax concave penalty". In: *Annals of Statistics* 38, pp. 894–942.

Zhang, N. and Liu, X. (2019). "Sparse inverse covariance matrix estimation via the ℓ_0-norm with Tikhonov regularization". In: *Inverse Problems* 35, p. 115010.

Zhang, Z. (2011). "An improvement to the Brent's method". In: *International Journal of Experimental Algorithms* 2, pp. 21–26.

Zou, H. (2006). "The adaptive lasso and its oracle properties". In: *Journal of the American Statistical Association* 101, pp. 1418–1429.