TWO COMPONENT SEMIPARAMETRIC DENSITY MIXTURE MODELS WITH A KNOWN COMPONENT

A Dissertation

Submitted to the Faculty

of

Purdue University

by

Zhou Shen

In Partial Fulfillment of the

Requirements for the Degree

of

Doctor of Philosophy

December 2018

Purdue University

West Lafayette, Indiana

THE PURDUE UNIVERSITY GRADUATE SCHOOL
STATEMENT OF DISSERTATION APPROVAL

Dr. Michael Levine, Chair

Associate Professor of the Department of Statistics, Purdue University

Dr. Anindya Bhadra

Associate Professor of the Department of Statistics, Purdue University

Dr. Chuanhai Liu

Professor of the Department of Statistics, Purdue University

Dr. Xiao Wang

Professor of the Department of Statistics, Purdue University

Approved by:

Dr. Jun Xie

Graduate Chair of the Department of Statistics, Purdue University

For my family.

ACKNOWLEDGMENTS

I would like to thank my advisor Dr. Michael Levine for his invaluable guidance and support to my research, and for his passion to thinking and pursuit of details which inspire me to achieve more in my Ph.D. career and in my future. I would also like to thank my thesis committee members, Dr. Anindya Bhadra and Dr. Chuanhai Liu for their precious advice to my research work, and Dr. Xiao Wang for his kindly advice and care throughout my Ph.D. life. And I really appreciate the faculty and staff in the Department of Statistics of Purdue University for their great work and help to me.

I want to thank my fellow colleagues in Statistics Department, Jincheng Bai, Chen Chen, Donglai Chen, Jinyuan Chen, Yao Chen, Eric Gerber, Botao Hao, Cheng Li, Jiapeng Liu, Yaowu Liu, Ryan Murphy, Yixuan Qiu, Simeng Qu, Min Ren, Yuying Song, Hui Sun, Qi Wang, Xiaoguang Wang, Wutao Wei, Zizhuang Wu, Yixi Xu, Jiasen Yang, Bing Yu, Boqian Zhang, Hao Zhang, Rongrong Zhang, Yumin Zhang, so many that I can't list all here, and all of my friends. I appreciate all kinds of help from you and the happy time we spent together, as my important memory of my years at Purdue.

Last but not least, I would like to thank my parents for their love and support forever.

TABLE OF CONTENTS

Page

LIST OF TABLES vii

LIST OF FIGURES viii

SYMBOLS ix

ABBREVIATIONS x

ABSTRACT xi

1 INTRODUCTION 1
- 1.1 Finite Parametric Mixture Models 1
- 1.2 Recent Work in Semiparametric Mixture Models 4
- 1.3 Our Contribution 7

2 AN MM ALGORITHM FOR ESTIMATION OF A TWO COMPONENT SEMIPARAMETRIC MIXTURE MODEL 8
- 2.1 Introduction 8
- 2.2 Identifiability 10
- 2.3 Algorithm 14
- 2.4 Convergence of Algorithm 18
- 2.5 Empirical Version 29

3 MINIMIZATION OF THE PENALIZED SMOOTHED LIKELIHOOD FUNCTIONAL 35
- 3.1 Introduction 35
- 3.2 Consistency 36
- 3.3 Convergence 40

4 NUMERICAL STUDY 52
- 4.1 Introduction 52
- 4.2 Simulation Examples 52
- 4.3 Bandwidth Selection 59
- 4.4 Comparison 61
- 4.5 A real data example 63

5 DISCUSSION AND FUTURE WORK 67
- 5.1 Convergence to Stationary Points 67
- 5.2 Convergence Rates 68
- 5.3 Efficiency of Algorithm 74

Page

REFERENCES 76
VITA 79

LIST OF TABLES

Table Page

4.1 Mean(SD) of estimated p/μ obtained by the symmetrization method . . . 63

4.2 Mean(SD) of estimated p/μ obtained by our algorithm 63

LIST OF FIGURES

Figure Page

4.1 Fitted mixture density for a mixture of Gaussian(6,1) and Gamma(2,1) . . 54

4.2 Fitted unknown component density for a mixture of Gaussian(6,1) and Gamma(2,1) 54

4.3 Fitted mixture density for a mixture of Beta(0.5,0.5) and Beta(2,2) 55

4.4 Fitted unknown component density for a mixture of Beta(0.5,0.5) and Beta(2,2) 56

4.5 Fitted mixture density for a mixture of unimodal and bimodal distributions 57

4.6 Fitted unknown component density for a mixture of unimodal and bimodal distributions 57

4.7 MISE of $\hat{f}$ and MSE of $\hat{p}$ in Normal-Exponential mixture model 58

4.8 MISE of $\hat{f}$ and MSE of $\hat{p}$ in Normal-Exponential mixture model 59

4.9 A plot of $CV(h)$ used for bandwidth selection 61

4.10 Fitted mixture densities 65

4.11 Fitted component densities 66

SYMBOLS

$g(x)$	p.d.f. of mixture distribution
$f_0(x)$	p.d.f. of the known component in a mixture distribution
$f(x)$	p.d.f. of the unknown component in a mixture distribution
$\ell(p, f)$	log-likelihood type objective functional
$\ell_n(p, f)$	empirical log-likelihood type objective functional
$KL(\cdot, \cdot)$	Kullback-Leibler distance
$K(x)$	kernel function
$K_h(x)$	rescaled kernel function with bandwidth h
$\mathcal{S}_h$	linear smoothing operator with kernel $K_h(x)$
$\mathcal{N}_h$	nonlinear smoothing operator with kernel $K_h(x)$
$V(\mu_f)$	variance of a p.d.f. $f(x)$ as a function of its mean μ_f
$\|\| \cdot \|\|_{C^1}$	C^1 norm
$\mathbb{E}$	expectation on the Lebesgue measure
$\mathbb{E}_n$	expectation on the counting measure
$r(x)$	excess loss function
$S(\cdot, \cdot)$	fitting functional
$\Omega_h(x)$	stabilizing functional
X	a Banach space $\mathbb{R} \times C^1$

ABBREVIATIONS

EM	expectation maximization
PCA	principal component analysis
KL	Kullback-Leibler (distance)
MM	majorization minimization
NEF	natural exponential family
PVF	power variance functions
GCT	global convergence theorem
SD	standard deviation
IQR	inter-quartile range
MSE	mean squared error
MISE	mean integrated squared error
CV	cross validation
ANC	acid neutralizing capacity

ABSTRACT

Shen, Zhou PhD, Purdue University, December 2018. Two Component Semiparametric Density Mixture Models with A Known Component. Major Professor: Michael Levine.

Finite mixture models have been successfully used in many applications, such as classification, clustering, and many others. As opposed to classical parametric mixture models, nonparametric and semiparametric mixture models often provide more flexible approaches to the description of inhomogeneous populations. As an example, in the last decade a particular two-component semiparametric density mixture model with a known component has attracted substantial research interest. Our thesis provides an innovative way of estimation for this model based on minimization of a smoothed objective functional, conceptually similar to the log-likelihood. The minimization is performed with the help of an EM-like algorithm. We show that the algorithm is convergent and the minimizers of the objective functional, viewed as estimators of the model parameters, are consistent.

More specifically, in our thesis, a semiparametric mixture of two density functions is considered where one of them is known while the weight and the other function are unknown. For the first part, a new sufficient identifiability condition for this model is derived, and a specific class of distributions describing the unknown component is given for which this condition is mostly satisfied. A novel approach to estimation of this model is derived. That approach is based on an idea of using a smoothed likelihood-like functional as an objective functional in order to avoid ill-posedness of the original problem. Minimization of this functional is performed using an iterative Majorization-Minimization (MM) algorithm that estimates all of the unknown parts of the model. The algorithm possesses a descent property with respect to the objective functional. Moreover, we show that the algorithm converges even when the

unknown density is not defined on a compact interval. Later, we also study properties of the minimizers of this functional viewed as estimators of the mixture model parameters. Their convergence to the true solution with respect to a bandwidth parameter is justified by reconsidering in the framework of Tikhonov-type functional. They also turn out to be large-sample consistent; this is justified using empirical minimization approach. The third part of the thesis contains a series of simulation studies, comparison with another method and a real data example. All of them show the good performance of the proposed algorithm in recovering unknown components from data.

1. INTRODUCTION

1.1 Finite Parametric Mixture Models

In statistics, mixtures of distributions have successfully provided mathematical-based approaches to represent the presence of component populations from the overall population. Given an observed data set, finite mixture models can derive the probability distributions of finite sub-populations from the pooled population without knowing the identity information of an individual observation. Under valid assumptions and criteria, mixture models can reveal the underlying structure of the overall population, which is meaningful from both a practical and theoretical point of view.

We let $X_1, ..., X_n$ denote an observed *i.i.d.* random sample of size n from the overall population with *mixture density* function $g(x)$, where $X_i \in \mathbb{R}$. In a typical finite mixture model, the overall population is assumed to consist of K components or sub-populations, each of which has a probability density function $f_i(x)$ called *component density*. A typical finite mixture model assumes that target density $g(x)$ can be represented as

$$g(x) = \sum_{i=1}^{K} \theta_i f_i(x), \tag{1.1}$$

where the θ_i's are *mixture proportions* (also called weights), that is,

$$0 \leq \theta_i \leq 1, \quad i = 1, ..., K \tag{1.2}$$

and

$$\sum_{i=1}^{K} \theta_i = 1. \tag{1.3}$$

Because of its practical usefulness and extreme flexibility in modeling, mixture models (1.1) have been widely applied in many fields such as biology, genetics, machine learning, economics, engineering, social sciences, etc. A variety of techniques

in areas of cluster analysis, discriminant analysis, image analysis and so on are based on applications of mixture models. For example, *Gaussian mixture models* are successfully used to model the heterogeneity in cluster analysis as an industry standard. And *Hidden Markov models* relax the independence assumption of observations in mixture models to model the time-dependent.

The simplest case of a density mixture model is the parametric one where each density component is viewed as belonging to a parametric family of distributions with an unknown parameter. Theory of *parametric* mixture models is fairly well developed by now. For example, Poisson distributed components with different means can be used to model the counts from a mixture of sources. And mixtures of Gaussian distributions with different means and covariance matrices are probably the best studied type of parametric density mixtures. These mixtures can approximate any continuous distribution arbitrarily well. A parametric density mixture distribution can be written in the form

$$g(x) = \sum_{i=1}^{K} \theta_i f(x|\gamma_i), \tag{1.4}$$

where γ_i is the parameter vector corresponding to the i-th component, and $f(x|\gamma_i)$ is the corresponding density function in the parametric family of distributions. K is also a parameter which determines the number of components in the mixture model.

To estimate the parameters $\boldsymbol{\theta} = (\theta_1, ..., \theta_K)$ and $\boldsymbol{\gamma} = (\gamma_1, ..., \gamma_K)$ in (1.4), given a sample, we can write down and maximize the log-likelihood function, that is

$$\begin{aligned} \ell(\boldsymbol{\theta}, \boldsymbol{\gamma}) &= \sum_{i=1}^{n} \log g(x_i|\boldsymbol{\theta}, \boldsymbol{\gamma}) \\ &= \sum_{i=1}^{n} \log \sum_{j=1}^{K} \theta_j f_j(x_i|\gamma_j). \end{aligned}$$

Due to the complexity of the functional form, direct maximum likelihood estimation is unavailable, which can be successfully solved by the *EM* algorithms. We let $\mathbf{Y}_1, ..., \mathbf{Y}_n$ denote the corresponding K-dimensional component label vectors with $Y_{ij} = 0$ or 1

which are unobserved. The complete data is therefore $X_1, \mathbf{Y}_1, ..., X_n, \mathbf{Y}_n$, and the complete data log-likelihood function is given by

$$\ell_c(\boldsymbol{\theta}, \boldsymbol{\gamma}) = \sum_{i=1}^{n} \sum_{j=1}^{K} y_{ij}(\log \theta_j + \log f_j(x_i|\gamma_j)).$$

Let $\boldsymbol{\gamma}^{(k)}$ and $\boldsymbol{\theta}^{(k)}$ be the current fit of the parameter $\boldsymbol{\gamma}$ and $\boldsymbol{\theta}$. In the expectation step (*E-step*), we take the conditional expectation of the complete data log-likelihood function given the observed data by

$$\begin{aligned} Q(\boldsymbol{\theta}, \boldsymbol{\gamma}; \boldsymbol{\theta}^{(k)}, \boldsymbol{\gamma}^{(k)}) &= \mathbb{E}_{\boldsymbol{\theta}^{(k)}, \boldsymbol{\gamma}^{(k)}}[\ell_c(\boldsymbol{\theta}, \boldsymbol{\gamma})|X_1, ..., X_n] \\ &= \sum_{i=1}^{n} \sum_{j=1}^{K} \frac{\theta_j^{(k)} f_j(x_i|\gamma_j^{(k)})}{\sum_{h=1}^{K} \theta_h^{(k)} f_h(x_i|\gamma_h^{(k)})} [\log \theta_j + \log f_j(x_i|\gamma_j)]. \end{aligned}$$

In the maximization step (*M-step*), the estimate of $\boldsymbol{\theta}$ and $\boldsymbol{\gamma}$ is updated by maximizing the conditional expectation above. We can derive

$$\theta_j^{(k+1)} = \frac{1}{n} \sum_{i=1}^{n} \frac{\theta_j^{(k)} f_j(x_i|\gamma_j^{(k)})}{\sum_{h=1}^{K} \theta_h^{(k)} f_h(x_i|\gamma_h^{(k)})},$$

and $\boldsymbol{\gamma}^{(k+1)}$ is obtained by solving

$$\sum_{i=1}^{n} \sum_{j=1}^{K} \frac{\theta_j^{(k)} f_j(x_i|\gamma_j^{(k)})}{\sum_{h=1}^{K} \theta_h^{(k)} f_h(x_i|\gamma_h^{(k)})} \partial \log f_j(x_i|\gamma_j)/\partial \boldsymbol{\gamma} = 0.$$

$\boldsymbol{\theta}$ and $\boldsymbol{\gamma}$ are thus estimated by alternating between an E-step and an M-step in an iterative way. More details can be found in [1].

However, strong assumptions in parametric mixture models still bring difficulties in realistic application. First, one needs to recognize the proper distribution families to unveil the reasonable hidden structure under the mixture distribution. Sometimes, there exist components which can not be perfectly modeled by any parametric distribution families. Second, the number of components K has to be predefined. But usually, it is unknown and inferred from sample. As an extreme example, a kernel density estimator using Gaussian kernel will approximate the mixture distribution very well, which can also be regarded as a mixture model with n Gaussian components. This provides no useful generalization of the mixture structure. Some work like cross validation has been attempted to these problems.

1.2 Recent Work in Semiparametric Mixture Models

A parametric density mixture (1.4) is not always the best choice for data analysis. A *nonparametric* extension can be easily introduced if we do not specify a particular parametric family(ies) for individual density components. This implies that a component density is not defined all the way down to finite-dimensional (Euclidean) parameter. Of course, such a setting is much more flexible than a parametric setting; moreover, even if there is a reason to believe that a particular distribution family should be used, nonparametric modeling provides us with a tool for checking such a hypothesis. The first nonparametric mixtures of density functions appearing in the literature were probably those of [2] and [3]. However, this new setting can be rather problematic on some occasions because of accompanying identifiability issues. It is easy to see that different nonparametric component distributions may generate the same mixture density distribution, so it is usually non-identifiable without strong restrictions. In practice, when the dimensionality is large, fitting general density curves becomes difficult and a simplifying assumption must be made. One such assumption that is commonly used is the assumption of conditional independence, i.e., the overall conditional density can be written in the product of marginal density functions from each dimension. It is conceptually similar to assumptions in dimension reduction like principal component analysis (PCA) where principle components are orthogonal directions. In this situation, it has been established, in order for such a model to be identifiable, first, the marginal density functions must be linearly independent across components and, second, the dimensionality is greater than 3. This powerful result has been established in [4] who focused proved it using algebraic arguments.

Sometimes, instead of just assuming that all of the components are general density functions, we may assume that some belong to a specific algebraic, e.g. location-scale family. In this case, location or scale parameters are additional scalar (Euclidean) parameters that have to be estimated. These models are commonly called *semiparametric* mixture models. Sometimes, a component or a subset of components in

these models are assumed to be known. For example, two-component semiparametric density mixtures with a known component attracted certain attention in the last 10-12 years, partly due to their usefulness in modelling differential gene expression in microarray data. From the practical viewpoint, it can be related to the multiple testing problem where p-values are uniformly distributed on $[0, 1]$ under the null hypothesis but their distribution under the alternative is unknown. In the setting of two-component semiparametric density mixtures, this means that the known distribution is uniform while the goal is to estimate the proportion of the false null hypothesis p and the distribution of the p-values under the alternative. More detailed descriptions in statistical literature can be found in e.g. [5] and [6].

Historically, whenever a two-component mixture model with a known component was considered, some assumptions were imposed on the unknown density function $f(x)$. Most commonly, it would be assumed that an unknown distribution belongs to a particular parametric family. In such a situation, [7] and [8] used the maximum likelihood-based method to fit it; [9] used the minimum χ^2 method, while [10] used the method of moments. [11] and [12] used empirical characteristic functions to estimate the unknown cumulative density function under a semiparametric normal mixture model. A less stringent assumption would be to assume that the unknown density function belongs to a location family with an unknown location parameter μ. An example of this approach is the paper [13] that served as a motivation for our work. It considered a particular two-component semiparametric density mixture model defined as

$$g(x) = (1 - p)f_0(x|\boldsymbol{\gamma}) + pf(x - \mu), \quad x \in \mathbb{R}, \tag{1.5}$$

where $f_0(x|\boldsymbol{\gamma})$ is fully known, while the unknown parameters are the mixture proportion $p \in (0, 1)$, and the non-null location parameter μ. The nonparametric component $f(x)$ is an even density function and has to be estimated as well. The model (1.5) was motivated by the problem of detection of differentially expressed genes under two or more conditions in microarray data. [13] established some sufficient conditions to achieve identifiability for model (1.5). They also suggested two methods,

i.e. symmetrization method and the method of moments, to estimate the unknown parameters. Moreover, they also showed that the resulting estimators are consistent if sufficient identifiability conditions are satisfied. A sequel paper [14] further established a joint central limit theorem for these estimators, showing weak convergence to a multivariate Gaussian process, and proposed methods to test various hypotheses about parameters. [15] considered a setting similar to (1.5) but with f_0 only known down to some Euclidean parameter. They discussed the corresponding identifiability issues, proposed a family of explicit estimators and explained how to achieve the optimal semiparametric efficiency in estimation. However, methods proposed by both [13] and [15] depend heavily on the fact that the density function of the unknown component is symmetric. Actually in many applications, there is no particular practical reason to make the unknown component symmetric. In particular, [13] noted that "In our opinion, a challenging problem would be to consider model (1.5) without the symmetry assumption on the unknown component". That is why we decided to consider a more general version of (1.5) with the function $f(x)$ now being a completely arbitrary density function. While working on this problem, we found out that [16] were considering an almost identical problem at the same time. More specifically, they assumed that $f_0(x|\boldsymbol{\gamma})$ is known and $f(x)$ is an arbitrary density function. A tuning-parameter-free heuristic estimator of p was given along with a finite sample lower confidence bound. Then, they provided a methodology to estimate the nonparametric $f(x)$ without assuming any constraint on its form. They also derived the rate of convergence and asymptotic limit for one of their estimators, and proposed some general identifiability criteria. However, their approach is based on the ideas of shape constrained estimation, which is very hard to generalize to multivariate density components.

1.3 Our Contribution

We consider a general case of a two-component univariate mixture model where one component distribution is known while the mixing proportion and the other component distribution are unknown, i.e., $g(x) = (1-p)f_0(x) + pf(x)$ where f_0 is the only known component. The mixture proportion $p \in (0, 1)$ and the other component density function $f(x)$ are both unknown. We would like to provide a nonparametric estimation to $f(x)$ by imposing some kernel functions. Part of this work is published in our paper [17].

In Chapter 2, we provide a sufficient condition of identifiability when modeling the mixture density, and propose an iterative algorithm based on minimizing a log-likelihood type objective functional. Descent property and algorithmic convergence are also discussed. In Chapter 3, we generalize a class of estimators which are the minimizers in our proposed minimization problem and can not be written in closed forms. The consistency of these estimators are derived by using empirical minimization. And the convergence with respect to a parameter of bandwidth in the problem is also discussed in the framework of Tikhonov-type regularization. In Chapter 4, various numerical studies are implemented including simulation in different settings, bandwidth selection, comparison with a symmetrization method in [13], and application in a real data example. In Chapter 5, some further issues are discussed and proposed as our future work.

2. AN MM ALGORITHM FOR ESTIMATION OF A TWO COMPONENT SEMIPARAMETRIC MIXTURE MODEL

2.1 Introduction

We consider a general case of a two-component univariate mixture model where one component distribution is known while the mixing proportion and the other component distribution are unknown. Such a model can be defined at its most general as

$$g(x) = (1-p)f_0(x) + pf(x), \tag{2.1}$$

where f_0 is a known density component, while $p \in (0,1)$ and $f(x)$ are the unknown weight and the unknown density component, respectively.

[18] proposed a nonlinear smoothing operator which are important to our modeling. Let h be a positive bandwidth and K a symmetric positive-valued kernel function that is also a true density; as a technical assumption, we will assume that K is continuously differentiable. The rescaled version of this kernel function is denoted $K_h(x) = K(x/h)/h$ for any $x \in \mathbb{R}$. We will also need a linear smoothing operator

$$\mathcal{S}f(x) = \int K_h(x-u)f(u)du \tag{2.2}$$

and a nonlinear smoothing operator

$$\mathcal{N}_h f(x) = \exp(\mathcal{S}\log f(x)) \tag{2.3}$$

for any generic density function f. More properties of $\mathcal{N}_h f(x)$ can be reviewed in [18]. For simplicity, let us assume that our densities are defined on a closed interval, e.g. $[0,1]$. This assumption is here for technical convenience only when proving algorithmic convergence related results. In the future, we will omit these integration limits whenever doing so doesn't cause confusion.

Our estimation approach is based on selecting p and f that minimize the following log-likelihood type objective functional

$$\ell(p, f) = \int g(x) \log \frac{g(x)}{(1-p)f_0(x) + p\mathcal{N}_h f(x)} dx. \tag{2.4}$$

The reason the functional (2.4) is of interest as an objective functional is as follows. First, recall that

$$KL(a(x), b(x)) = \int \left[a(x) \log \frac{a(x)}{b(x)} + b(x) - a(x) \right] dx \tag{2.5}$$

is a *Kullback-Leibler distance* between the two arbitrary positive integrable functions $a(x)$ and $b(x)$ which are not necessarily distribution densities; as usual, $KL(a, b) \geq 0$. This version of the Kullback-Leibler distance is a special case of the so-called *Bregman divergence*; one can find its definition in e.g. [19] p. 16. Note that the functional (2.4) can be represented as a penalized Kullback-Leibler distance between the target density $g(x)$ and the smoothed version of the mixture $(1-p)f_0(x) + p\mathcal{N}_h f(x)$; indeed, we can represent $\ell(p, f)$ as

$$\ell(p, f) = KL(g(x), (1-p)f_0(x) + p\mathcal{N}_h f(x)) + p \left\{ 1 - \int \mathcal{N}_h f(x)\, dx \right\}. \tag{2.6}$$

The quantity $1 - \int \mathcal{N}_h f(x)\, dx = \int [f(x) - \mathcal{N}_h f(x)]\, dx$ is effectively the penalty on the smoothness of the unknown density. Thus, the functional (2.4) can be interpreted as a penalized smoothed likelihood functional.

Our method to solve the minimization problem of (2.4) belongs to a family of algorithms called *MM* algorithms. MM algorithms represent a generalization of the classical EM framework. In minimization problems, MM stands for majorization minimization, while in maximization problems, MM stands for minorization maximization. They are commonly used whenever optimization of a difficult objective function is best avoided and a series of simpler objective functions is optimized instead. The concept underlying MM algorithms was stated originally in [20] in the context of line search methods. Suppose $\theta^{(m)}$ is the current estimate of some param-

eter θ, and let $g(\theta|\theta^{(m)})$ denote a real-valued function of θ depending on $\theta^{(m)}$. Given a function $f(\theta)$, if for all θ the following is true,

$$g(\theta|\theta^{(m)}) \geq f(\theta),$$
$$g(\theta^{(m)}|\theta^{(m)}) = f(\theta^{(m)}),$$

we say that the function $g(\theta|\theta^{(m)})$ majorize $f(\theta)$. Let $\theta^{(m+1)}$ denote the minimizer of $g(\theta|\theta^{(m)})$. The decent property

$$f(\theta^{(m+1)}) \leq f(\theta^{(m)})$$

will follow directly from the fact

$$g(\theta^{(m+1)}|\theta^{(m)}) \leq g(\theta^{(m)}|\theta^{(m)}),$$

which brings numerical stability to this MM algorithm. A general introduction to MM algorithms from the statistical viewpoint is available in, for example, [21].

In Section 2.2, we present sufficient conditions and discuss identifiability issues. In Section 2.3, we derive an an algorithm to solve the minimization problem of (2.4) in an iterative way, an show that it is an MM algorithm. In Section 2.4, we prove the descent property of the estimator sequence with respect to the objective functional (2.4) and algorithmic convergence of estimators of unknown parameters. We will see that (2.4) converges to its stationary point under some mild conditions. In Section 2.5, we propose an empirical version of the algorithm for real applications and generalize it to multivariate cases.

2.2 Identifiability

In general, the model (2.1) is not identifiable. In what follows, we investigate some special cases. For an unknown density function f, let us denote its mean by μ_f and its variance by σ_f^2. To state a sufficient identifiability result, we consider a general equation

$$(1-p)f_0(x) + pf(x) = (1-p_1)f_0(x) + p_1 f_1(x). \tag{2.7}$$

We also denote variance of the distribution $f(x)$ as a function of its mean μ_f, i.e., $V(\mu_f)$.

Theorem 2.2.1 *Consider the model* (2.1) *with the unknown density function* f. *Without loss of generality, assume that the first moment of* f_0 *is zero while its second moment is finite. We assume that the function* f *belongs to a set of density functions whose first two moments are finite, whose means are not equal to zero and that are all of the same sign; that is,* $f \in \mathcal{F} = \{f : \int x^2 f(x)\,dx < +\infty; \mu_f > 0 \text{ or } \mu_f < 0\}$. *Moreover, we assume that for any* $f \in \mathcal{F}$ *the function* $G(\mu_f) = \frac{V(\mu_f)}{\mu_f}$ *is strictly increasing. Then, the equation* (2.7) *has the unique solution* $p_1 = p$ *and* $f_1 = f$.

Proof First, let us assume that the mean $\mu_f > 0$. Then, the assumption of Theorem (2.2.1) implies that the function $V : (0, \infty) \mapsto (0, \infty)$ is strictly increasing. Let us use the notation θ_0 for the second moment of f_0. If we assume that there are distinct $p_1 \neq p$ and $f_1 \neq f$ such that

$$(1-p)f_0(x) + pf(x) = (1-p_1)f_0(x) + p_1 f_1(x),$$

the following two moment equations are easily obtained:

$$\zeta = p_1\mu_{f_1} = p\mu_f \tag{2.8}$$

and

$$(p_1 - p)\theta_0 = \zeta(\mu_{f_1} - \mu_f) + p_1 V(\mu_{f_1}) - pV(\mu_f), \tag{2.9}$$

where $\zeta > 0$. Our task is now to show that if (2.8) and (2.9) are true, then $p = p_1$ and $f = f_1$. To see this, let us assume $p_1 > p$ (the case $p_1 < p$ can be treated in exactly the same way). Then from the first equation we have immediately that $\mu_{f_1} < \mu_f$; moreover, since the function $G(\mu_f)$ is a strictly increasing one, then so is the function $\mu_f + G(\mu_f)$. With this in mind, we have

$$\mu_{f_1} + \frac{V(\mu_{f_1})}{\mu_{f_1}} < \mu_f + \frac{V(\mu_f)}{\mu_f}. \tag{2.10}$$

On the other hand, $(p_1 - p)\theta_0 \geq 0$ which implies

$$\begin{aligned} 0 &\leq \zeta(\mu_{f_1} - \mu_f) + p_1 V(\mu_{f_1}) - pV(\mu_f) \\ &= \zeta(\mu_{f_1} - \mu_f) + \zeta\left(\frac{V(\mu_{f_1})}{\mu_{f_1}} - \frac{V(\mu_f)}{\mu_f}\right). \end{aligned}$$

Therefore, this implies that

$$\mu_{f_1} + \frac{V(\mu_{f_1})}{\mu_{f_1}} \geq \mu_f + \frac{V(\mu_f)}{\mu_f}, \tag{2.11}$$

and we end up with a contradiction. Therefore, we must have $p = p_1$. This, in turn, implies immediately that $f = f_1$.

The case where $\mu_f < 0$ proceeds similarly. Let us now consider the case where the variance function $V : (-\infty, 0) \to (0, \infty)$ and is strictly monotonically increasing. As a first step, again take $p_1 > p$. Clearly, the first moment equation is yet again (2.8) where now $\zeta < 0$. If $p_1 > p$, we now have $\mu_{f_1} > \mu_f$ and, due to the strict monotonicity of $G(\mu)$, we have $\mu_{f_1} + \frac{V(\mu_{f_1})}{\mu_{f_1}} > \mu_f + \frac{V(\mu_f)}{\mu_f}$. On the other hand, since $(p_1 - p)\theta_0 \geq 0$, we have

$$\begin{aligned} 0 &\leq \zeta(\mu_{f_1} - \mu_f) + p_1 V(\mu_{f_1}) - pV(\mu_f) \\ &= \zeta\left(\left\{\mu_{f_1} + \frac{V(\mu_{f_1})}{\mu_{f_1}}\right\} - \left\{\mu_f + \frac{V(\mu_f)}{\mu_f}\right\}\right). \end{aligned} \tag{2.12}$$

Because $\zeta < 0$, the above implies that $\left\{\mu_{f_1} + \frac{V(\mu_{f_1})}{\mu_{f_1}}\right\} - \left\{\mu_f + \frac{V(\mu_f)}{\mu_f}\right\} < 0$ which contradicts the assumption that the function $G(\mu)$ is strictly increasing. ■

To understand better what is going on here, it is helpful if we can suggest a more specific density class which satisfies the sufficient condition in Theorem 2.2.1. The form of Theorem 2.2.1 suggests one such possibility - a family of natural exponential families with power variance functions (NEF-PVF). For convenience, we give the definition due to [22].

Definition 2.2.1 *A natural exponential family (NEF for short) is said to have a power variance function if its variance function is of the form $V(\mu) = \alpha\mu^\gamma$, $\mu \in \Omega$, for some constants $\alpha \neq 0$ and γ, called the scale and power parameters, respectively.*

This family of distributions is discussed in detail in [23] and [22]. In particular, they establish that the parameter space Ω can only be $\mathbb{R}$, $\mathbb{R}^+$ and $\mathbb{R}^-$; moreover, we can only have $\gamma = 0$ if and only if $\Omega = \mathbb{R}$. The most interesting property is that (see Theorem 2.1 from [22] for details) for any NEF-PVF, it is necessary that $\gamma \notin (-\infty, 0) \cup (0, 1)$; in other words, possible values of γ are 0, corresponding to the normal distribution, 1, corresponding to Poisson, and any positive real numbers that are greater than 1. In particular, the case $\gamma = 2$ corresponds to gamma distribution. Out of these choices, the only one that does not result in a monotonically increasing function $G(\mu)$ is $\gamma = 0$ that corresponds to the normal distribution; thus, we have to exclude it from consideration. With this exception gone, the NEF-PVF framework includes only density families with either strictly positive or strictly negative means; due to this, it seems a rather good fit for the description of the family of density functions f in the Theorem 2.2.1.

Note that the exclusion of the normal distribution is also rather sensible from the practical viewpoint because it belongs to a location family; therefore, it can be treated in the framework of [13]. More specifically, Proposition 1 of [13] suggests that, when $f(x)$ is normal, the equation (2.7) has at most two solutions if f_0 is an even pdf and at most three solutions if f_0 is not an even pdf.

It is also of interest to compare our Theorem 2.2.1 with the Lemma 4 of [16] that also establishes an identifiability result for the model (2.1). The notions of identifiability that are considered in the two results differ: whereas we discuss the identifiability based on the first two moments, Lemma 4 of [16] looks at a somewhat different definition of identifiability. At the same time, the interpretation given in the previous Remark, suggests an interesting connection. For example, the case where the unknown density function f is gamma corresponds to the power parameter of the NEF-PVF family being equal to 2. According to our identifiability result Theorem 2.2.1, the mixture model (2.1) is, then, identifiable with respect to the first two moments. On the other hand, let us assume that the known density function f_0 is the standard normal. Since its support fully contains the support of any density from

the gamma family, identifiability in the sense of [16] now follows from their Lemma 4.

We only assumed that the first moment of f_0 is equal to zero for simplicity. It is not hard to reformulate the Theorem 2.2.1 if this is not the case. The proof is analogous.

Corollary 2.2.2 *Consider the model* (2.1) *with the unknown density function* f. *We assume that the known density* f_0 *has finite first two moments and denote its first moment* μ_{f_0}. *We also assume that the function* f *belongs to a set of density functions whose first two moments are finite, and whose means are all either greater than* μ_{f_0} *or less than* μ_{f_0}*:*

$$f \in \mathcal{F} = \{f : \int x^2 f(x)\,dx < +\infty; \mu_f > \mu_{f_0} \text{ or } \mu_f < \mu_{f_0}\}. \tag{2.13}$$

Let us assume that $G(\mu_f) = \frac{V(\mu_f)}{\mu_f - \mu_{f_0}}$ *is a strictly increasing function in* μ_f *for a fixed, known* f_0. *Then, the equation* (2.7) *has the unique solution* $p_1 = p$ *and* $f_1 = f$.

2.3 Algorithm

Now we are going to introduce our algorithm that would search for unknown p and $f(x)$ in (2.1). The first result that we need is the following technical Lemma.

Lemma 2.3.1 *For any pdf* $\widetilde{f}$ *and any real number* $\widetilde{p} \in (0,1)$,

$$\begin{aligned} &\ell(\widetilde{p}, \widetilde{f}) - \ell(p, f) \\ &\leq -\int g(x) \left[(1 - w(x)) \log\left(\frac{1-\widetilde{p}}{1-p}\right) + w(x) \log\left(\frac{\widetilde{p}\mathcal{N}_h \widetilde{f}(x)}{p\mathcal{N}_h f(x)}\right) \right] dx, \end{aligned} \tag{2.14}$$

where

$$w(x) = \frac{p\mathcal{N}_h f(x)}{(1-p) f_0(x) + p\mathcal{N}_h f(x)}. \tag{2.15}$$

Proof The result follows by the following straightforward calculations:

$$\begin{aligned}\ell(\widetilde{p},\widetilde{f})-\ell(p,f) &= -\int g(x)\log\left(\frac{(1-\widetilde{p})f_0(x)+\widetilde{p}\mathcal{N}_h\widetilde{f}(x)}{(1-p)f_0(x)+p\mathcal{N}_h f(x)}\right)dx \qquad (2.16)\\ &= -\int g(x)\log\left((1-w(x))\frac{1-\widetilde{p}}{1-p}+w(x)\frac{\widetilde{p}\mathcal{N}_h\widetilde{f}(x)}{p\mathcal{N}_h f(x)}\right)dx\\ &\le -\int g(x)\left[(1-w(x))\log\left(\frac{1-\widetilde{p}}{1-p}\right)+w(x)\log\left(\frac{\widetilde{p}\mathcal{N}_h\widetilde{f}(x)}{p\mathcal{N}_h f(x)}\right)\right],\end{aligned}$$

where the last inequality follows by convexity of the negative logarithm function. ■

Suppose at iteration t, we get the updated pdf f^t and the updated mixing proportion p^t. Let $w^t(x)=\frac{p^t\mathcal{N}_h f^t(x)}{(1-p^t)f_0(x)+p^t\mathcal{N}_h f^t(x)}$, and define

$$\begin{aligned}p^{t+1} &= \int g(x)w^t(x)dx,\\ f^{t+1}(x) &= \alpha^{t+1}\int K_h(x-u)g(u)w^t(u)du,\end{aligned}$$

where α^{t+1} is a normalizing constant needed to ensure that f^{t+1} integrates to one. Then the following result holds.

Theorem 2.3.1 *For any $t\ge 0$, $\ell(p^{t+1},f^{t+1})\le\ell(p^t,f^t)$.*

Proof By Lemma 2.3.1, for an arbitrary density function $\widetilde{f}$ and an arbitrary number $0<\widetilde{p}<1$,

$$\begin{aligned}&\ell(\widetilde{p},\widetilde{f})-\ell(p^t,f^t) \qquad (2.17)\\ &\le -\int g(x)\left[(1-w^t(x))\log\left(\frac{1-\widetilde{p}}{1-p^t}\right)+w^t(x)\log\left(\frac{\widetilde{p}\mathcal{N}_h\widetilde{f}(x)}{p^t\mathcal{N}_h f^t(x)}\right)\right]dx.\end{aligned}$$

Let $(\widehat{p},\widehat{f})$ be the minimizer of the right hand side of (2.17) with respect to $\widetilde{p}$ and $\widetilde{f}$. Note that the right-hand side becomes zero when $\widetilde{p}=p^t$ and $\widetilde{f}=f^t$; therefore, the minimum value of the functional on the right hand side must be less then or equal to 0. Therefore, it is clear that $\ell(\widehat{p},\widehat{f})\le\ell(p^t,f^t)$. To verify that the statement of the Theorem 2.3.1 is true, it remains only to show that $(\widehat{p},\widehat{f})=(p^{t+1},f^{t+1})$.

Note that the right hand side of (2.17) can be rewritten as

$$-\int g(x)[(1-w^t(x))\log(1-\widetilde{p})+w^t(x)\log\widetilde{p}]dx$$
$$-\int g(x)w^t(x)\log\mathcal{N}_h\widetilde{f}(x)dx+T,$$

where the term T only depends on (p^t, f^t). The first integral in the above only depends on $\widetilde{p}$ but not on $\widetilde{f}$. It is easy to see that the minimizer of this first integral with respect to $\widetilde{p}$ is $\widehat{p}=\int g(x)w^t(x)dx$. The second integral, on the contrary, depends only on $\widetilde{f}$ but not on $\widetilde{p}$. It can be rewritten as

$$\begin{aligned}
&-\int g(x)w^t(x)\log\mathcal{N}_h\widetilde{f}(x)dx\\
&=-\int\int g(x)w_t(x)K_h(x-u)\log\widetilde{f}(u)dudx\\
&=-\int\left(\int K_h(u-x)g(x)w^t(x)dx\right)\log\widetilde{f}(u)du\\
&=-\frac{1}{\alpha^{t+1}}\int f^{t+1}(u)\log\widetilde{f}(u)du\\
&=\frac{1}{\alpha^{t+1}}\int f^{t+1}(u)\log\frac{f^{t+1}(u)}{\widetilde{f}(u)}du-\frac{1}{\alpha^{t+1}}\int f^{t+1}(u)\log f^{t+1}(u)du.
\end{aligned}$$

The first term in the above is the Kullback-Leibler divergence between f^{t+1} and $\widetilde{f}$ scaled by α^{t+1}, which is minimized at f^{t+1}, i.e., for $\widehat{f}=f^{t+1}$. Since the second term does not depend on $\widetilde{f}$ at all, we arrive at the needed conclusion. ■

The above suggests that the following algorithm can be used to estimate the parameters of the model (2.1). First, we start with initial values p_0, f^0 at the step $t=0$. Then, for any $t=1,2,\ldots$

- Define the weight

$$w^t(x)=\frac{p^t\mathcal{N}_hf^t(x)}{(1-p^t)f_0(x)+p^t\mathcal{N}_hf^t(x)}. \tag{2.18}$$

- Define the updated probability

$$p^{t+1}=\int g(x)w^t(x)dx. \tag{2.19}$$

- Define

$$f^{t+1}(u) = \alpha^{t+1} \int K_h(u-x) g(x) w^t(x) dx. \tag{2.20}$$

Note that the proposed algorithm is an MM (majorization minimization) algorithm. As a first step, let (p^t, f^t) denote the current parameter values in our iterative algorithm. The main goal is to obtain a new functional $b^t(p, f)$ such that, when shifted by a constant, it majorizes $\ell(p, f)$. In other words, there must exist a constant C^t such that, for any (p, f),

$$b^t(p, f) + C^t \geq \ell(p, f), \tag{2.21}$$

where the equality holds if and only if $(p, f) = (p^t, f^t)$. The use of t as a superscript in this context indicates that the definition of the new functional $b^t(p, f)$ depends on the parameter values (p^t, f^t); these change from one iteration to the other. The benefit of using a functional b^t instead of the original one is that it separates $\tilde{p}$ and $\tilde{f}$.

In our case, we define a functional

$$\begin{aligned} b^t(\tilde{p}, \tilde{f}) = &- \int g(x)[(1-\omega^t(x)) \log(1-\tilde{p}) + \omega^t(x) \log \tilde{p}]\, dx \\ &- \int g(x) \omega^t(x) \log \mathcal{N}_h \tilde{f}(x)\, dx. \end{aligned} \tag{2.22}$$

Note that the dependence on f^t is through weights ω^t. From the proof of the Theorem 2.3.1, it follows that, for any argument $(\tilde{p}, \tilde{f})$ we have

$$\ell(\tilde{p}, \tilde{f}) - \ell(p^t, f^t) \leq b^t(\tilde{p}, \tilde{f}) - b^t(p^t, f^t). \tag{2.23}$$

This means, that $b^t(\tilde{p}, \tilde{f})$ is a majorizing functional; indeed, it is enough to select the constant C^t such that $C^t = \ell(p^t, f^t) - b^t(p^t, f^t)$. In the proof of the Theorem 2.3.1 it is the series of functionals $b^t(\tilde{p}, \tilde{f})$ (note that they are different at each step of iteration) that is being minimized with respect to $(\tilde{p}, \tilde{f})$, and not the original functional $\ell(\tilde{p}, \tilde{f})$. This, indeed, establishes that our algorithm is an MM algorithm.

2.4 Convergence of Algorithm

The following lemma shows that the sequence $\xi_t = \ell(p^t, f^t)$, defined by our algorithm, also has a non-negative limit (which is not necessarily a global minimum of $\ell(p, f)$).

Lemma 2.4.1 *There exists a finite limit of the sequence $\xi_t = \ell(p^t, f^t)$ as $t \to \infty$:*

$$L := \lim_{t\to\infty} \xi_t \tag{2.24}$$

for some $L \geq 0$.

Proof First, note that ξ_t is a non-increasing sequence for any integer t due to the Theorem 2.3.1. Thus, if we can show that it is bounded from below by zero, the proof will be finished. Then, the functional $\ell(p^t, f^t)$ can be represented as

$$\begin{aligned}
\ell(p^t, f^t) =& KL(g(x), (1-p^t)f_0(x) + p^t\mathcal{N}_h f^t(x)) + \int g(x)\,dx \\
&- \int [(1-p^t)f_0(x) + p^t\mathcal{N}_h f^t(x)]\,dx \\
=& KL(g(x), (1-p^t)f_0(x) + p^t\mathcal{N}_h f^t(x)) + 1 \\
&- (1-p^t) - p^t \int \mathcal{N}_h f^t(x)\,dx \\
=& KL(g(x), (1-p^t)f_0(x) + p^t\mathcal{N}_h f^t(x)) + p^t\left[1 - \int \mathcal{N}_h f^t(x)\,dx\right].
\end{aligned} \tag{2.25}$$

Now, since K is a proper density function, by Jensen's inequality,

$$\begin{aligned}
\mathcal{N}_h f^t(x) &= \exp\left\{\int K_h(x-u)\log f^t(u)\,du\right\} \\
&\leq \int K_h(x-u) f_t(u)\,du \equiv \mathcal{S}f^t(x).
\end{aligned} \tag{2.26}$$

Moreover, using Fubini's theorem, one can easily show that $\int \mathcal{S}f^t(x)\,dx = 1$ since f^t is a proper density function. Therefore, one concludes easily that

$$\int \mathcal{N}_h f^t(x)\,dx \leq \int \mathcal{S}f^t(x)\,dx = 1. \tag{2.27}$$

Thus, $\ell(p^t, f^t) \geq 0$ is non-negative due to non-negativity of the Kullback-Leibler distance. ■

It is, of course, not clear directly from the Lemma 2.4.1 if the sequence (p^t, f^t), generated by this algorithm, also converges. Being able to answer this question requires establishing a lower semicontinuity property of the functional $\ell(p, f)$. Some additional requirements have to be imposed on the kernel function K in order to obtain the needed result that is given below. First, we denote a compact set Δ the domain of the kernel function K, and consider all density functions with a compact support.

Theorem 2.4.1 *Let the kernel $K : \Delta \to \mathbb{R}$ be bounded from below and Lipschitz continuous with the Lipschitz constant C_K. Then, the minimizing sequence (p^t, f^t) converges to (p_h^*, f_h^*) that depends on the bandwidth h such that $L = \ell(p_h^*, f_h^*)$.*

Proof We prove this result in two parts. First, let us introduce a subset of functions $B = \{\mathcal{S}\phi : 0 \leq \phi \in L_1^+(\Delta), \int \phi = 1\}$ where $L_1^+(\Delta)$ denotes a subset of all non-negative functions from $L_1(\Delta)$. Such a subset represents all densities on a closed compact interval that can be represented as linearly smoothed integrable functions. Every function f_t generated in our algorithm except, perhaps, the initial one, can clearly be represented in this form. This is because, at every step of iteration,

$$f^{t+1}(x) = \alpha^{t+1} \int K_h(x-u) g(u) w^t(u)\, du = \int K_h(x-u)\phi(u)\, du, \tag{2.28}$$

where $\phi(u) = \alpha^{t+1} g(u) w^t(u)$. Moreover, we observe that

$$\int \phi(u)\, du = \alpha^{t+1} \int g(u) w^t(u)\, du = \alpha^{t+1} p^{t+1}. \tag{2.29}$$

Next, one concludes, by using Fubini theorem that, for any $t = 1, 2, \ldots$

$$\int f^{t+1}(x)\, dx = \alpha^{t+1} \int g(u) w^t(u) \left[\int K_h(x-u)\, dx \right] du = 1. \tag{2.30}$$

Since the iteration step t in the above is arbitrary, we established that $\alpha^t p^t = 1$ and, therefore, $\int \phi(u)\, du = 1$.

By definition of set B, it is clear that, as long as the kernel function is a proper density function (and so is non-negative), any $f \in B$ is non-negative and so every

function in the set B is bounded from below. If the kernel function is Lipschitz continuous on Δ it is clearly bounded from above by some positive constant M : $\sup_{x\in\Delta} K(x) < M$. Thus, every function $f \in B$ satisfies $f(x) \leq M < \infty$. This implies that the set B is uniformly bounded. Also, by definition of set B, for any two points $x, y \in \Delta$ we have

$$\begin{aligned} |f(x) - f(y)| &\leq \int |K_h(x-u) - K_h(y-u)|\phi(u)\, du \\ &\leq C_K|x-y|, \end{aligned} \tag{2.31}$$

where the constant C_K depends on the choice of kernel K but not on the function f. This establishes the equicontinuity of the set B. Therefore, by Arzela-Ascoli theorem the set of functions B is a compact subset of $C(\Delta)$ with a sup metric.

Since for every $t = 2, 3, \ldots$ $f^t \in B$, by Arzela-Ascoli theorem we have a subsequence $f^{t_k} \to f_h^*$ as $k \to \infty$ uniformly over Ω. Since for every $t = 1, 2, \ldots$ p^t is bounded between 0 and 1, there exists, by Bolzano-Weierstrass theorem, a subsequence $p^{t_k} \to p_h^*$ as $k \to \infty$ in the usual Euclidean metric. Consider a Cartesian product space $\{(p, f)\}$ where every $p \in [0, 1]$ and $f \in C(\Delta)$. To define a metric on such a space we introduce an m-product of individual metrics for some non-negative m. This means that, if the first component space has a metric d_1 and the second d_2, the metric on the Cartesian product is $(|d_1|^m + |d_2|^m)^{1/m}$ for some non-negative m. For example, the specific case $m = 0$ corresponds to $|d_1| + |d_2|$ and $m = \infty$ corresponds to $\max(d_1, d_2)$. For such an m-product metric, clearly, we have a subsequence $(p^{t_k}, f^{t_k}) \to (p_h^*, f_h^*)$ that converges to (p_h^*, f_h^*) in the m-product metric. Without loss of generality, assume that the subsequence coincides with the whole sequence (p^t, f^t). Of course, such a sequence $(p^t, f^t) \in [0, 1] \times C(\Delta)$ for any t.

Now, that we know that there is always a converging sequence (p^t, f^t), we can proceed further. Since each f^t is bounded away from zero and from above, then so is the limit function $f_h^*(x)$ in the limit (p_h^*, f_h^*). This implies that $(p^t, \log f^t) \to (p_h^*, \log f_h^*)$ uniformly in the m-product topology as well and the same is true also

for $(p^t, \mathcal{S}\log f^t)$. Analogously, the uniform convergence follows also in $(p^t, \mathcal{N}_h f^t) \to (p_h^*, \mathcal{N}_h f_h^*)$; moreover,

$$(1-p^t)f_0 + p^t\mathcal{N}_h f^t \to (1-p_h^*)f_0 + p_h^*\mathcal{N}_h f_h^* \tag{2.32}$$

uniformly in the m-product topology. Since we have the function

$$\psi(t) = -\log t + t - 1 \geq 0,$$

Fatou Lemma implies that

$$\begin{aligned} &\int g(x)\psi((1-p_h^*)f_0(x) + p_h^*\mathcal{N}_h f_h^*(x))\,dx \\ &\leq \liminf \int g(x)\psi((1-p^t)f_0(x) + p^t\mathcal{N}_h f^t(x))\,dx. \end{aligned} \tag{2.33}$$

The lower semicontinuity of the functional $\ell(p, f)$ follows immediately and with it the conclusion of the Theorem 2.4.1. ■

The above result can also be proved in the case where the densities involved have their support on the entire real line. To do so, it is necessary to impose constraints on the tails of these densities. The following result from the functional analysis forms the cornerstone of this analysis.

Lemma 2.4.2 *(Fréchet-Kolmogorov theorem) Let B be a bounded subset in $L_p(\mathbb{R})$ with $p \in [1, \infty)$. The subset B is relatively compact if and only if the following properties hold for any function $f \in B$:*

1. $\lim_{r\to\infty} \int_{|x|>r} |f|^p = 0$ *uniformly on B,*

2. $\lim_{a\to 0} ||\tau_a f - f||_p = 0$ *uniformly on B.*

where $\tau_a f$ denotes the translation of f by a, that is, $\tau_a f(x) = f(x-a)$.

A very nice proof of this result can be found in e.g. an expository paper [24]. Now we can formulate the following result.

Corollary 2.4.2 *Let all of the conditions of Theorem 2.4.1 be true but assume that the unknown density $f(x)$ and the known density $f_0(x)$ are now defined on the entire real line $\mathbb{R}$. Also, assume that $f_0(x)$ is bounded everywhere from above. Then, the convergence result of Theorem 2.4.1 remains correct.*

Proof The only part of the proof of Theorem 2.4.1 that needs updating is that of establishing compactness of the subset B. Now, we need to establish its compactness as a subset of $L^1(\mathbb{R})$. To get this done, we will use Lemma 2.4.2. Fist, recall that our algorithm updates the density estimate at each step as

$$\begin{aligned} f^{t+1}(x) &= \alpha^{t+1} \int K_h(x-u) g(u) w^t(u)\, du \\ &= \int K_h(x-u)\phi(u)\, du, \end{aligned}$$

where $\phi(u) = \alpha^{t+1} g(u) w^t(u)$ is a density function belonging to $L_1^+(\mathbb{R})$, and as a result, $\int \phi(u)\, du = 1$. Earlier, we showed that there exists a subsequence $p^{t_k} \to p_h^*$ by Bolzano-Weierstrass theorem. In order to use Lemma 2.4.2, we first show that at any step of iteration p_{t+1} is bounded away from zero. Indeed, from our algorithm we can see that $p^{t+1} = \int g(x) w^t(x)\, dx$; thus, if we show that the weight $w^t(x)$ is always bounded away from zero for any t, the probability p^{t+1} is bounded away from zero as well. Since the kernel function K is bounded from below, we can easily claim that for any $f \in B$,

$$\begin{aligned} f &= \int K_h(x-u)\phi(u)\, du \qquad (2.34) \\ &\geq \inf_{x\in\Omega} K_h(x-u) \int \phi(u)\, du \\ &= K^* > 0. \end{aligned}$$

Next, by definition of the smoothing operator $\mathcal{N}_h f^t(x)$, and since $f^t \in B$ for any step of iteration t, we have

$$\begin{aligned} \mathcal{N}_h f^t(x) &= \exp\{\int K_h(x-u)\log f_t(u)\,du\} \\ &\geq \exp\{\log K^* \int K_h(x-u)\,du\} \\ &= K^* > 0, \end{aligned} \tag{2.35}$$

since the kernel function is a proper density function. Now, recall that at each step t the weight is

$$w^t(x) = \frac{p^t \mathcal{N}_h f^t(x)}{(1-p^t) f_0(x) + p^t \mathcal{N}_h f^t(x)}.$$

If we assume that $f_0(x)$ is bounded from above, and since $\mathcal{N}_h f^t(x)$ is always bounded from above by M, we can conclude that the denominator of the integrand in the definition of $w^t(x)$ is bounded from above and, therefore, $w^t(x)$ is always bounded from below as long as p^t is bounded from below. Using the above argument, it is easy to see that as long as we start with $p^0 > 0$, p^t will stay bounded away from zero at every step of iteration. Thus, we can claim that the limit $p_h^* > 0$. Since $a_k p_k = 1$ for all k, there must be $a^{t_k} \to a_h^*$ and a^{t_k} is bounded from above by some $M_a > 0$.

Now, we can check the first condition in Lemma 2.4.2 for functions that belong to the set B. For any fixed mixture density $g(x)$, $\lim_{r\to\infty} \int_{|x|>r} g(x)\,dx = 0$; therefore, for any $\epsilon_g > 0$, there exists $r' > 0$ such that $\int_{|x|>r'} g(x)\,dx < \epsilon_g$. Since the kernel function K is a proper density function defined on a finite interval support Δ, for any

$\epsilon_K > 0$ there exists $r > r'$ such that $\int_{|x|>r} K_h(x-u)\,dx < \epsilon_K$ for any $|u| \le r'$. This implies that

$$\begin{aligned}
&\int_{|x|>r} |f^{t_k}|\,dx \qquad (2.36)\\
=&\int_{|x|>r} \alpha^{t_k}\,dx \int_{-\infty}^{\infty} K_h(x-u)g(u)w^{t_k-1}(u)\,du\\
\le&\alpha^{t_k}\int_{|x|>r} dx \int_{-\infty}^{\infty} K_h(x-u)g(u)\,du\\
=&\alpha^{t_k}\int_{|x|>r} dx \left(\int_{|u|\le r'} K_h(x-u)g(u)\,du + \int_{|u|>r'} K_h(x-u)g(u)\,du\right)\\
=&\alpha^{t_k}\int_{|u|\le r'} g(u)\,du \int_{|x|>r} K_h(x-u)\,dx\\
&+\alpha^{t_k}\int_{|u|>r'} g(u)\,du \int_{|x|>r} K_h(x-u)\,dx\\
\le&\alpha^{t_k}\left(\int_{|u|\le r'} \epsilon_K\, g(u)\,du + \int_{|u|>r'} g(u)\,du\right) \le M_a\left(\epsilon_K+\epsilon_g\right),
\end{aligned}$$

and so the first condition of the Lemma 2.4.2 has been verified. To verify the second condition we note first that, due to Lipschitz continuity of the kernel function and the fact that it is defined on a finite interval, we have

$$\begin{aligned}
&\int_{-\infty}^{\infty} |f^{t_k}(x-a) - f^{t_k}(x)|\,dx \qquad (2.37)\\
=&\int_{-\infty}^{\infty} \left|\int_{-\infty}^{\infty} (K_h(x-a-u) - K_h(x-u)\phi(u))\,du\right| dx\\
\le&\int_{-\infty}^{\infty} dx \int_{-\infty}^{\infty} |(K_h(x-a-u) - K_h(x-u)|\,\phi(u))\,du\\
=&\int_{-\infty}^{\infty} \phi(u)\,du \int_{-\infty}^{\infty} |(K_h(x-a-u) - K_h(x-u)|\,dx\\
\le&\int_{-\infty}^{\infty} |a|C_K|\Delta|\phi(u)\,du = |a|C_K|\Delta|,
\end{aligned}$$

and so for any $|a| < \rho$ we have the integral bounded from above by $C_K\rho|\Delta|$ that does not depend on the function $f \in B$. ■

Theorem 2.4.3 *If (p^t, f^t) is a minimizer of $\ell(p,f)$, then it is a fixed point of our algorithm $(p^{t+1}, f^{t+1}) = G(p^t, f^t)$.*

Proof By Lemma (2.3.1),

$$\begin{aligned}
&\ell(p^t, f^t) - \ell(\widetilde{p}, \widetilde{f}) \qquad (2.38)\\
&\geq \int g(x) \left[(1 - w^t(x)) \log \left(\frac{1-\widetilde{p}}{1-p^t} \right) + w^t(x) \log \left(\frac{\widetilde{p}}{p^t} \right) + w^t(x) \log \left(\frac{\mathcal{N}_h \widetilde{f}(x)}{\mathcal{N}_h f^t(x)} \right) \right] dx \\
&:= I(p^t, f^t, \widetilde{p}, \widetilde{f})
\end{aligned}$$

First, we would like to prove that $I(p^t, f^t, \widetilde{p}, \widetilde{f}) = 0$ if and only if $(\widetilde{p}, \widetilde{f}) = (p^t, f^t)$. Let's apply the iteration step $(p^{t+1}, f^{t+1}) = G(p^t, f^t) = (\widetilde{p}, \widetilde{f}$ according to our algorithm, and consider the first two terms in $I(p^t, f^t, p^{t+1}, f^{t+1})$,

$$\begin{aligned}
&\int g(x) \left[(1 - w^t(x)) \log \left(\frac{1-\widetilde{p}}{1-p^t} \right) + w^t(x) \log \left(\frac{\widetilde{p}}{p^t} \right) \right] dx \qquad (2.39)\\
=&(1 - p^{t+1}) \log \left(\frac{1-p^{t+1}}{1-p^t} \right) + p^{t+1} \log \left(\frac{p^{t+1}}{p^t} \right) \\
=&KL(p^{t+1} || \, p^t)
\end{aligned}$$

This is the Kullback-Leibler divergence from the discrete probability distributions p^{t+1} to p^t. Therefore, it is non-negative but disappears when $p^{t+1} = p^t$. For the third term in $I(p^t, f^t, p^{t+1}, f^{t+1})$,

$$\begin{aligned}
&\int g(x) w^t(x) \log \left(\frac{\mathcal{N}_h \widetilde{f}(x)}{\mathcal{N}_h f^t(x)} \right) dx \qquad (2.40)\\
=&\int g(x) w^t(x) \left[\int K_h(x-u) \log \left(\frac{f^{t+1}(u)}{f^t(u)} \right) du \right] dx \\
=&\int \log \left(\frac{f^{t+1}(u)}{f^t(u)} \right) \left[\int g(x) K_h(x-u) w^t(x) dx \right] du \\
=&\int f^{t+1}(u) \log \left(\frac{f^{t+1}(u)}{f^t(u)} \right) du \\
=&KL(f^{t+1} || \, f^t)
\end{aligned}$$

This is the Kullback-Leibler divergence from f^{t+1} to f^t, which is also non-negative but disappears when $f^{t+1} = f^t$.

We have seen that

$$\ell(p^t, f^t) - \ell(p^{t+1}, f^{t+1}) \geq I(p^t, f^t, p^{t+1}, f^{t+1}) \geq 0, \qquad (2.41)$$

where $I(p^t, f^t, p^{t+1}, f^{t+1}) = 0$ if and only if $(p^t, f^t) = (p^{t+1}, f^{t+1})$. Since (p^t, f^t) is a minimizer of $\ell(p, f)$, there must be

$$\ell(p^t, f^t) - \ell(p^{t+1}, f^{t+1}) \leq 0. \tag{2.42}$$

Therefore, $I(p^t, f^t, p^{t+1}, f^{t+1}) = 0$ and it follows $(p^{t+1}, f^{t+1}) = G(p^t, f^t) = (p^t, f^t)$. ■

Before stating the convergence property of $\ell(p^t, f^t)$, we need to introduce the following definition and lemma, as referenced in the proof of convergence properties of EM algorithms by [25].

Definition 2.4.1 *A map G from points of X to a subset of X is called a point-to-set map on X. The map G is said to be closed at x if $x_k \to x$, $x_k \in X$ and $y_k \to y$, $y_k \in G(x_k)$, imply $y \in G(x)$. Moreover, if a point-to-point map is continuous, then it is closed.*

Lemma 2.4.3 ***Zangwill's Global Convergence Theorem (GCT)***
Let the sequence $\{x_k\}_{k=0}^{\infty}$ be generated by $x_{k+1} \in G(x_k)$, where G is a point-to-set map on X. Let a solution set $\Gamma \subset X$ be given, and suppose that:

1. *The sequence $\{x_k\}_{k=0}^{\infty}$ are contained in a compact subset $S \subset X$.*
2. *G is closed on $X \backslash \Gamma$.*
3. *There is a continuous function ℓ on X such that (a) if $x \notin \Gamma$, then $\ell(y) < \ell(x)$ for all $y \in G(x)$, and (b) if $x \in \Gamma$, $\ell(y) \leq \ell(x)$ for all $y \in G(x)$.*

Then all the limit points of $\{x_k\}_{k=0}^{\infty}$ are in the solution set Γ and $\ell(x_k)$ converges monotonically to $\ell(x)$ for some $x \in \Gamma$.

By following the similar logic, we can have the following conjecture for the convergence of our algorithm.

Conjecture 2.4.4 *Consider algorithm defined in equations* (2.18), (2.19) *and* (2.20). *We assume that $K_h(u)$ used in the definition of the smoothing operators $\mathcal{S}$ and $\mathcal{N}_h$ is*

a kernel function bounded away from zero on a compact interval Δ*:* $\inf_{u\in\Delta} K_h(u) > 0$*, and* $K_h(u)$ *is Lipschitz continuous on* Δ*. Moreover, the unknown density function* f *is assumed to belong to* $C^1(\Omega)$*, and both of* $f_0(x)$ *and* $f(x)$ *are bounded away from zero and from above on* Ω*. Then, this algorithm always converges to a set of stationary points of* (2.4)*.*

Proof Let us denote Γ a set of all stationary points of $\ell(p, f)$. As a first step, we show that the sequence (p^t, f^t), generated by our algorithm, belongs to a compact set of the parameter space. Indeed, let us introduce a subset of functions $B = \{\mathcal{S}\phi : 0 \le \phi \in L_1(\Omega), \int \phi = 1\}$. Such a subset represents all densities on a closed compact interval that can be represented as linearly smoothed integrable functions. Every function f_t generated in our algorithm except, perhaps, the initial one, can clearly be represented in this form. This is because, at every step of iteration, $f^{t+1}(x) = \alpha^{t+1} \int K_h(x-u) g(u) w^t(u)\, du = \int K_h(x-u)\phi(u)\, du$ where $\phi(u) = \alpha^{t+1} g(u) w^t(u)$. Moreover, we observe that $\int \phi(u)\, du = \alpha^{t+1} \int g(u) w^t(u)\, du = \alpha^{t+1} p^{t+1}$. Next, one concludes, by using Fubini theorem that, for any $t = 1, 2, \dots$,

$$\int f^{t+1}(x)\, dx = \alpha^{t+1} \int g(u) w^t(u) \left[\int K_h(x-u)\, dx \right] du = 1. \tag{2.43}$$

Since the iteration step t in the above is arbitrary, we established that $\alpha^t p^t = 1$ and, therefore, $\int \phi(u)\, du = 1$. Next, since the kernel function K is bounded from below, we can easily claim that for every $f \in B$, $f = \int K_h(x-u)\phi(u)\, du \ge \inf_{x\in\Omega} K_h(x-u) \int \phi(u)\, du = \inf_{x\in\Omega} K_h(x-u) > 0$ and, therefore, every function in the set B is bounded from below. If the kernel function is Lipschitz continuous on Ω it is clearly bounded from above by some positive constant $M : \sup_{x\in\Omega} K(x) < M$. Thus, every function $f \in B$ satisfies $f(x) \le M < \infty$. This implies that the set B is uniformly bounded. Also, by definition of set B, for any two points $x, y \in \Omega$, we have

$$\begin{aligned} |f(x) - f(y)| &\le \int |K_h(x-u) - K_h(y-u)| \phi(u)\, du \\ &\le C_K |x - y|, \end{aligned} \tag{2.44}$$

where the constant C_K depends on the choice of kernel K but not on the function f. This establishes the equicontinuity of the set B. Therefore, by Arzela-Ascoli theorem the set of functions B is a compact subset of $C(\Omega)$ with a sup metric. Clearly, at each step t of our algorithm, $0 \leq p^t \leq 1$ and so all of p^t's belong to the compact subset of $[0,1]$ with the standard Euclidean metric. The last step is to define the m-product of the sup metric for functions and the usual Euclidean metric in $\mathbb{R}^1$, for example, the maximum of the two metrics. Let us denote d_1 the sup metric in $C(\Omega)$ and d_2 the Euclidean metric in $\mathbb{R}^1$. Then, the maximum metric is $d = \max(d_1, d_2)$. Thus, all points (p^t, f^t) form a compact subset of $\mathbb{R}^1 \times C[\Omega]$ with respect to metric d.

As a next step, it is necessary to verify that the map $G : (p^t, f^t) \mapsto (p^{t+1}, f^{t+1})$ is a continuous one, thus closed as well. Looking at the definitions in (2.19) and (2.20), it is clear that it will follow from the continuity of the functional $\ell(p, f)$ with respect to both of its arguments. The continuity with respect to p in the standard Euclidean metric is clear; as for the continuity with respect to the functional argument f, it is clear if we note that

$$
\begin{aligned}
& |\ell(p, \tilde{f}) - \ell(p, f)| \qquad (2.45)\\
=& \int g(x) \log \frac{(1-p)f_0(x) + p\mathcal{N}_h f(x)}{(1-p)f_0(x) + p\mathcal{N}_h \tilde{f}}\, dx \\
=& \int g(x) \log \frac{(1-p)f_0(x) + p\mathcal{N}_h \tilde{f} + p\mathcal{N}_h f(x) - p\mathcal{N}_h \tilde{f}}{(1-p)f_0(x) + p\mathcal{N}_h \tilde{f}}\, dx \\
=& \int g(x) \log \left\{ 1 + \frac{p\mathcal{N}_h f(x) - p\mathcal{N}_h \tilde{f}}{(1-p)f_0(x) + p\mathcal{N}_h \tilde{f}} \right\} dx \\
\leq& \int g(x) \frac{2pM \sup |f - \tilde{f}|}{(1-p)f_0(x) + p\mathcal{N}_h \tilde{f}}\, dx \\
\leq& C||f - \tilde{f}||_{C^1},
\end{aligned}
$$

where M is the upper bound of the kernel K_h utilized above and C is some positive constant according to the integral.

Finally, it is necessary to show that the functional $\ell(p, f)$ exhibits only strict descent for any points $(p, f) \notin \Gamma$. In other words, we would like to show that, if $(p^t, f^t) \notin \Gamma$, then $\ell(p^{t+1}, f^{t+1}) < \ell(p^t, f^t)$. Note that $\ell(p, f)$ consists of a Euclidean

parameter p and a functional parameter $f(x)$. If (p^t, f^t) is not a stationary point of $\ell(p, f)$, we may assume that it follows $\frac{\partial \ell}{\partial p} \neq 0$. The dependence on $f(x)$ is not clear for now, which makes this result a conjecture, rather than a theorem (A little more discussion will be made in Chapter 5). Suppose

$$\frac{\partial \ell(p^t, f^t)}{\partial p^t} = \int g(x) \frac{f_0(x) - \mathcal{N}_h f^t(x)}{(1-p^t) f_0(x) + p^t \mathcal{N}_h f^t(x)} \, dx \neq 0. \tag{2.46}$$

It is trivial to see that

$$\int g(x) \frac{(1-p^t) f_0(x) + p^t \mathcal{N}_h f^t(x)}{(1-p^t) f_0(x) + p^t \mathcal{N}_h f^t(x)} \, dx = 1. \tag{2.47}$$

If $p^t = p^{t+1}$, it follows by our algorithm that

$$p^t = p^{t+1} = \int g(x) \frac{p^t \mathcal{N}_h f^t(x)}{(1-p^t) f_0(x) + p^t \mathcal{N}_h f^t(x)} \, dx. \tag{2.48}$$

(2.47) and (2.48) imply that

$$\begin{aligned} & \int g(x) \frac{f_0(x) - \mathcal{N}_h f^t(x)}{(1-p^t) f_0(x) + p^t \mathcal{N}_h f^t(x)} \, dx \\ = & \frac{1}{1-p^t} \int g(x) \frac{(1-p^t) f_0(x) + p^t \mathcal{N}_h f^t(x)}{(1-p^t) f_0(x) + p^t \mathcal{N}_h f^t(x)} \, dx \\ & - \frac{1}{p^t(1-p^t)} \int g(x) \frac{p^t \mathcal{N}_h f^t(x)}{(1-p^t) f_0(x) + p^t \mathcal{N}_h f^t(x)} \, dx \\ = & \frac{1}{1-p^t} - \frac{p^t}{p^t(1-p^t)} = 0 \end{aligned} \tag{2.49}$$

which contradicts (2.46). So there must be $p^{t+1} \neq p^t$. We have seen in the proof of Theorem 2.4.3 that $I(p^t, f^t, p^{t+1}, f^{t+1}) = 0$ if and only if $(p^t, f^t) = (p^{t+1}, f^{t+1})$. Now we can confirm that $\ell(p^{t+1}, f^{t+1}) - \ell(p^t, f^t) < 0$ for any $(p^t, f^t) \notin \Gamma$. Then the theorem is clear by GCT as in Lemma (2.4.3). ■

2.5 Empirical Version

In practice, the number of observations n sampled from the target density function g is finite. This necessitates the development of the empirical version of our algorithm that can be implemented in practice. Many proof details here are similar to proofs

of properties of the algorithm we introduced in the previous chapter. Therefore, we will be brief in our explanations. Denote the empirical cdf of the observations X_i, $i = 1, \ldots, n$ $G_n(x)$ where $G_n(x) = \frac{1}{n}\sum_{i=1}^{n} I_{X_i \leq x}$. Then, we define a functional

$$\begin{aligned} l_n(f,p) &= -\int \log((1-p)f_0(x) + p\mathcal{N}_h f(x))\, dG_n(x) \\ &\equiv -\sum_{i=1}^{n} \log((1-p)f_0(X_i) + p\mathcal{N}_h f(X_i)). \end{aligned} \tag{2.50}$$

The following analogue of the Lemma 2.3.1 can be easily established.

Lemma 2.5.1 *For any pdf* $\widetilde{f}$ *and* $\widetilde{p} \in (0,1)$,

$$\begin{aligned} & l_n(\widetilde{f},\widetilde{p}) - l_n(f,p) \\ & \leq -\int \left[(1-w(x)) \log\left(\frac{1-\widetilde{p}}{1-p}\right) + w(x) \log\left(\frac{\widetilde{p}\mathcal{N}_h\widetilde{f}(x)}{p\mathcal{N}_h f(x)}\right)\right] dG_n(x), \end{aligned} \tag{2.51}$$

where the weight $w(x) = \frac{p\mathcal{N}_h f(x)}{(1-p)f_0(x) + p\mathcal{N}_h f(x)}$.

Proof By convexity of negative logarithm function we get that

$$\begin{aligned} & l_n(\widetilde{p},\widetilde{f}) - l_n(p,f) \\ & = -\int \log\left(\frac{(1-\widetilde{p})f_0(x) + \widetilde{p}\mathcal{N}_h\widetilde{f}(x)}{(1-p)f_0(x) + p\mathcal{N}_h f(x)}\right) dG_n(x) \\ & = -\int \log\left((1-w(x))\frac{1-\widetilde{p}}{1-p} + w(x)\frac{\widetilde{p}\mathcal{N}_h\widetilde{f}(x)}{p\mathcal{N}_h f(x)}\right) dG_n(x) \\ & \leq -\int \left[(1-w(x)) \log\left(\frac{1-\widetilde{p}}{1-p}\right) + w(x) \log\left(\frac{\widetilde{p}\mathcal{N}_h\widetilde{f}(x)}{p\mathcal{N}_h f(x)}\right)\right] dG_n(x). \end{aligned} \tag{2.52}$$

■

Now we can define the empirical version of our algorithm. Denote (p_n^t, f_n^t) values of the density f and probability p at the iteration step t. Define the weights as $w_n^t(x) = \frac{p_n^t \mathcal{N}_h f_n^t(x)}{(1-p_n^t)f_0(x) + p_n^t \mathcal{N}_h f_n^t(x)}$. We use the subscript n everywhere intentionally to

stress that these quantities depend on the sample size n. For the next step, define (p_n^{t+1}, f_n^{t+1}) as

$$p_n^{t+1} = \int w_n^t(x) dG_n(x) = \frac{1}{n}\sum_{i=1}^{n} w_n^t(X_i) \tag{2.53}$$

$$\begin{aligned} f_n^{t+1}(x) &= \alpha_n^{t+1} \int K_h(x-u) w_n^t(u) dG_n(u) \\ &= \frac{\alpha_n^{t+1}}{n} \sum_{i=1}^{n} K_h(x - X_i) w_n^t(X_i), \end{aligned} \tag{2.54}$$

where α_n^{t+1} is a normalizing constant such that f_n^{t+1} is a valid pdf. Since $\int K_h(X_i - u)du = 1$ for $i = 1, \ldots, n$, we get

$$1 = \int f_n^{t+1}(u)du = \frac{\alpha_n^{t+1}}{n}\sum_{i=1}^{n} w_n^t(X_i), \tag{2.55}$$

and hence,

$$\alpha_n^{t+1} = \frac{n}{\sum_{i=1}^{n} w_n^t(X_i)}. \tag{2.56}$$

The following result establishes the descent property of the empirical version of our algorithm.

Theorem 2.5.1 *For any $t \geq 0$, $\ell_n(p_n^{t+1}, f_n^{t+1}) \leq \ell_n(p_n^t, f_n^t)$.*

Proof It follows by Lemma 2.5.1 that

$$\begin{aligned} & l_n(\widetilde{p}, \widetilde{f}) - l_n(p_n^t, f_n^t) \\ & \leq -\int \left[(1 - w_n^t(x)) \log\left(\frac{1-\widetilde{p}}{1-p_n^t}\right) + w_n^t(x) \log\left(\frac{\widetilde{p}\mathcal{N}_h\widetilde{f}(x)}{p_n^t \mathcal{N}_h f_n^t(x)}\right) \right] dG_n(x). \end{aligned} \tag{2.57}$$

Let $(\widehat{p}, \widehat{f})$ be the minimizer of the right hand side; note that the right hand side is equal to zero when $\widetilde{p} = p_n^t$ and $\widetilde{f} = f_n^t$, so the smallest possible value of the right hand side will be less than or equal to zero. Next, we show that the minimizer is $(\widehat{p}, \widehat{f}) = (p_n^{t+1}, f_n^{t+1})$.

The right hand side of (2.57) equals

$$\begin{aligned} & -\log\left(\frac{1-\widetilde{p}}{1-p_n^t}\right) \int (1 - w_n^t(x)) dG_n(x) - \log\left(\frac{\widetilde{p}}{p_n^t}\right) \int w_n^t(x) dG_n(x) \\ & - \int w_n^t(x) \log\left(\frac{\mathcal{N}_h\widetilde{f}(x)}{\mathcal{N}_h f_n^t(x)}\right) dG_n(x). \end{aligned} \tag{2.58}$$

Note that the last term does not depend on $\widetilde{p}$. Minimizing the sum of the first two terms with respect to $\widetilde{p}$, we get that the minimizer is $\widehat{p} = \int w_n^t(x) dG_n(x)$ which is equal to p_n^{t+1}. To minimize the last term with respect to $\widetilde{f}$, note that

$$
\begin{aligned}
& -\int w_n^t(x) \log \mathcal{N}_h \widetilde{f}(x) dG_n(x) \qquad (2.59) \\
& = -\int w_n^t(x) \left(\int K_h(x-u) \log \widetilde{f}(u) du \right) dG_n(x) \\
& = -\int \left(\int w_n^t(x) K_h(x-u) dG_n(x) \right) \log \widetilde{f}(u) du \\
& = -\frac{1}{\alpha_n^{t+1}} \int f_n^{t+1}(u) \log \widetilde{f}(u) du \\
& = \frac{1}{\alpha_n^{t+1}} \int f_n^{t+1}(u) \log \left(\frac{f_n^{t+1}(u)}{\widetilde{f}(u)} \right) du - \frac{1}{\alpha_n^{t+1}} \int f_n^{t+1}(u) \log f_n^{t+1}(u) du,
\end{aligned}
$$

The second term above does not depend on $\widetilde{f}$; by definition of Kullback-Leibler distance, we find that $\widehat{f}(\cdot) = f_n^{t+1}(\cdot)$ is the minimizer of $-\int w_n^t(x) \log \mathcal{N}_h \widetilde{f}(x) dG_n(x)$. ■

As before, the empirical version of the proposed algorithm is an MM (majorization - minimization) algorithm that represents a generalization of the classical EM setting. More specifically, we can show that there exists another functional $b_n^t(p, f)$ such that, when shifted by a constant, it majorizes $l_n(p, f)$. It is easy to check that such a functional is

$$
\begin{aligned}
b_n^t(\tilde{p}, \tilde{f}) = & -\int \left[(1 - \omega_n^t(x)) \log(1 - \tilde{p}) + \omega_n^t(x) \log \tilde{p} \right] dG_n(x) \qquad (2.60) \\
& -\int \omega_n^t(x) \log \mathcal{N}_h \tilde{f}(x) \, dG_n(x).
\end{aligned}
$$

Note that in the proof of the Theorem 2.5.1 it is the series of functionals $b_n^t(\tilde{p}, \tilde{f})$ that is being minimized with respect to $(\tilde{p}, \tilde{f})$, and not the original functional $l_n(\tilde{p}, \tilde{f})$.

Note also that this algorithm can be easily generalized to the multivariate case. Let $f : \mathbb{R}^d \to \mathbb{R}$ be the unknown density function and $f_0 : \mathbb{R}^d \to \mathbb{R}$ be the unknown one. We assume that the target density $g : \mathbb{R}^d \to \mathbb{R}$ is a two-component mixture of the unknown component f and the known component f_0 with the weight $0 < p < 1$. Our

data consist of sample $\vec{X}_1, \ldots, \vec{X}_n$ generated by g. Since Lemma 2.5.1 and Theorem 2.5.1 of our manuscript only depend on some fairly basic tools, such as Jensen's inequality and convexity of the negative logarithm function, both of them remain true in the multivariate case and the following algorithm can be defined.

Denote (p_n^t, f_n^t) values of the density f and probability p at the iteration step t. We use the subscript n everywhere intentionally to stress that these quantities depend on the sample size n. For the next step, define (p_n^{t+1}, f_n^{t+1}) as

$$p_n^{t+1} = \frac{1}{n} \sum_{i=1}^{n} w_n^t(\mathbf{X}_i) \tag{2.61}$$

$$f_n^{t+1}(\mathbf{x}) = \frac{\alpha_n^{t+1}}{n} \sum_{i=1}^{n} K_h(\mathbf{x} - \mathbf{X}_i) w_n^t(\mathbf{X}_i), \tag{2.62}$$

where $w_n^t(\mathbf{x}) = \frac{p_n^t \mathcal{N}_h f_n^t(\mathbf{x})}{(1-p_n^t) f_0(\mathbf{x}) + p_n^t \mathcal{N}_h f_n^t(\mathbf{x})}$ is the weight (probability) that an observation $\mathbf{x}$ has been generated by an unknown component density, and α_n^{t+1} is a normalizing constant such that f_n^{t+1} is a valid density function. Since $\int K_h(\mathbf{X}_i - \mathbf{u}) d\mathbf{u} = 1$ for $i = 1, \ldots, n$, and we assume that K is a symmetric density function, we find that

$$1 = \int f_n^{t+1}(\mathbf{u}) d\mathbf{u} = \frac{\alpha_n^{t+1}}{n} \sum_{i=1}^{n} w_n^t(X_i), \tag{2.63}$$

and hence,

$$\alpha_n^{t+1} = \frac{n}{\sum_{i=1}^{n} w_n^t(\mathbf{X}_i)}. \tag{2.64}$$

It can be verified immediately that the resulting algorithm possesses the descent property and is an MM algorithm, as before.

As before, we can also show that the sequence $\ell_n(p_n^t, f_n^t)$ generated by our algorithm does not only possess the descent property but is also bounded from below.

Lemma 2.5.2 *There exists a finite limit of the sequence $\xi_n^t = \ell_n(p_n^t, f_n^t)$ as $t \to \infty$:*

$$L_n = \lim_{t \to \infty} \xi_n^t \tag{2.65}$$

for some $L_n \geq 0$.

The proof is almost exactly the same as the proof of the Lemma 2.4.1 and is omitted in the interest of brevity. Finally, one can also show that the sequence (p_n^t, f_n^t) generated by our algorithm converges to (p_n^*, f_n^*) such that $L_n = l_n(p_n^*, f_n^*)$. The proof is almost the same as that of the Theorem 2.4.1 and is omitted for conciseness.

3. MINIMIZATION OF THE PENALIZED SMOOTHED LIKELIHOOD FUNCTIONAL

3.1 Introduction

We have seen an estimate sequence (p^t, f^t) to minimize the log-likelihood type objective functional $\ell(p, f)$ in Chapter 2. It may not be the unique iterative algorithm which solves the minimization problem. In fact, one can definitely propose different algorithms which converge to the estimator of the minimizer iteratively, based on a set of observations. We consider a class of estimators which are minimizers of $\ell(p, f)$. It is impossible for them to be written in a closed form. We would like to see if this class of estimators are good solutions of modeling the target mixture density (2.1).

These estimators are based on kernel functions $K(x)$ and the bandwidth h, because we introduce a nonlinear smoothing operator $\mathcal{N}_h$ in the objective functional. One question is the consistency of these estimators, i.e., their behavior when the number of observations increases infinitely. The second question is how the dependence of h influences the estimation of the unknown parameters. Obviously, these estimators can only be safe to use if they converge to the underlying structure of the target mixture density (2.1) when the bandwidth h approaches some values like 0. These are not trivial discussion, because we can not know the closed forms of this class of iterative estimators.

The first question regarding to the consistency of estimators can be answered through the *empirical risk minimization*. [26] investigated the behavior of the empirical minimization algorithm. They compared the empirical, random, structure and the original one on the class via the uniform law of large numbers and isomorphic coordinate projections. They also provided a bound for the estimates by a direct analysis of the empirical minimization algorithm which is essentially sharp.

Our work for the second question regarding convergence of estimators with respect to the bandwidth h is based on *Tikhonov-type regularization*. Consider to solve for x in an equation $F(x) = y$ where F is an operator and y is given. If no x can exactly satisfy the equation or the solution x is not unique or not stable, the inverse problem is said to be ill-posed, e.g., [27]. Ill-posed inverse problems requires regularization techniques for obtaining a stable approximate solution. Classical Tikhonov regularization can be extended to very general settings to avoid overfitting issues. [28] and [29] described and analyzed a general framework for solving ill-posed operator equations by minimizing Tikhonov-like functionals.

In Section 3.2, we show the consistency of estimators without closed forms which minimize $\ell(p, f)$ by using empirical minimization. In Section 3.3, we modify the framework of Tikhonov-type regularization and establish the stability of these estimators and its convergence depending on bandwidth h.

3.2 Consistency

Consider the log-likelihood type objective functional we would like to minimize, and rewrite it as

$$\begin{aligned} \ell(p, f) &= \int g(x) \log \frac{g(x)}{(1-p)f_0(x) + p\mathcal{N}_h f(x)} dx \\ &= \int g(x) \log g(x) dx - \int g(x) \log[(1-p)f_0(x) + p\mathcal{N}_h f(x)] dx \\ &= \mathbb{E}L(x|p, f) + C(g), \end{aligned} \tag{3.1}$$

where $L(x|p, f) = -\log[(1-p)f_0(x) + p\mathcal{N}_h f(x)]$ and $C(g)$ is constant to a given mixture density $g[0, 1] \to \mathbb{R}^+$. $L(x|p, f)$ is a loss function which are to be optimized in the sense of expectation. Similarly, the empirical version of the log-likelihood type objective functional can be rewritten as

$$\begin{aligned} \ell_n(p, f) &= -\frac{1}{n} \sum_{i=1}^{n} \log[(1-p)f_0(x_i) + p\mathcal{N}_h f(x_i)] \\ &= \mathbb{E}_n L(x|p, f), \end{aligned} \tag{3.2}$$

which is the average of the loss given a sample from the mixture distribution g. Suppose the minimizer of $\mathbb{E}L(x|p,f)$ is $(p^*,f^*) \in X$ defined in previous sections, satisfying

$$\mathbb{E}L(x|p^*,f^*) = \inf_{(p,f)\in X} \mathbb{E}L(x|p,f). \tag{3.3}$$

Now we define an excess loss function $r : [0,1] \to \mathbb{R}$ by

$$r(x) = L(x|p,f) - L(x|p^*,f^*), \tag{3.4}$$

and the class F of such excess loss functions is defined by

$$F = \{x \mapsto L(x|p,f) - L(x|p^*,f^*) : (p,f) \in X\}. \tag{3.5}$$

Since $(p^*,f^*) \in X$ are fixed, choosing $(p,f) \in X$ to minimize $\mathbb{E}L(x|p,f)$ or $\mathbb{E}_n L(x|p,f)$ corresponds to choosing $r \in F$ to minimize $\mathbb{E}r(x)$ or $\mathbb{E}_n r(x)$. Note that both of $\mathbb{E}r(x)$ and $\mathbb{E}_n r(x)$ are non-negative by definition, while $r(x)$ can take negative values.

Suppose we have n observations $\{x_1,...,x_n\}$ sampled from the target mixture density g. An empirical minimizer $\hat{r}$ is defined by

$$\hat{r} = \operatorname*{argmin}_{r\in F} \mathbb{E}_n r(x), \tag{3.6}$$

corresponding to an empirical minimizer $(\hat{p},\hat{f}) \in X$. We would like to study the consistency of $\hat{r}$ from the conditional expectation of the empirical minimizer

$$\mathbb{E}\left[\hat{r}(x)|x_1,...,x_n\right], \tag{3.7}$$

and for brevity, we write this conditional expectation as $\mathbb{E}\hat{r}$.

To get the upper bound on $\mathbb{E}\hat{r}$, some concentration inequalities are required and presented below. The first is Bernstein's inequality.

Lemma 3.2.1 *Let P be a probability measure and $g : (0,1) \to \mathbb{R}^+$ be the corresponding probability density and $X_1,...,X_n$ be independent random variables generated by g. Given a function $r : [0,1] \to \mathbb{R}$, set $Z = \sum_{i=1}^n r(X_i)$. Then for any $t > 0$,*

$$Pr\{|Z - \mathbb{E}Z| \geq t\} \leq 2\exp\left(-\frac{t^2}{2(\sigma^2 + bt/3)}\right),$$

where $b = ||r||_\infty$ and $\sigma^2 = n\mathbb{E}r^2$.

The second concentration result is a functional version of Bernstein's inequality.

Lemma 3.2.2 *Let $g : (0,1) \to \mathbb{R}^+$ be a probability density and $X_1, ..., X_n$ be independent random variables generated by g. Suppose F be a class of functions defined on $[0,1]$. For every $r \in F$ and $\mathbb{E}r = 0$, there exists a constant b such that $||r||_\infty \leq b$. Define*

$$Z = \sup_{r \in F} \sum_{i=1}^{n} r(X_i),$$
$$\bar{Z} = \sup_{r \in F} |\sum_{i=1}^{n} r(X_i)|.$$

Then for any $t > 0$,

$$Pr\{|Z - \mathbb{E}Z| \geq t\} \leq C \exp\left(-\frac{t}{Kb} \log\left(1 + \frac{bt}{\sigma^2 + b\mathbb{E}\bar{Z})}\right)\right),$$

where C and K are absolute constants, and $\sigma^2 = n \sup_{r \in F} var(r)$.

The consistency of $\hat{r}$ to be derived is based on the uniform law of large numbers. Recall that a class of functions F satisfies the uniform law of large numbers with respect to a probability measure P, if for every $\epsilon > 0$,

$$\lim_{n \to \infty} Pr\{||P - P_n||_F \geq \epsilon\} = 0,$$

where

$$||P - P_n||_F = \sup_{r \in F} |\mathbb{E}r - \mathbb{E}_n r|.$$

This leads to the following notion of similarity between the empirical and actual structures.

Definition 3.2.1 *Given a probability measure P, the empirical and actual structures on F are (λ, ϵ)-close if*

$$Pr\{||P - P_n||_F \geq \lambda\} \leq \epsilon.$$

If the empirical and actual structures are (λ, ϵ)-close and $\hat{r}$ is an empirical minimizer, it will follow $\mathbb{E}\hat{r} \leq \lambda$ with probability at least $1 - \epsilon$. Actually, $\mathbb{E}_n\hat{r} = \inf_{r \in F} \mathbb{E}_n r \leq 0$ since $r^*(x) = L(x|p^*, f^*) - L(x|p^*, f^*) = 0$, and the result follows.

Lemma 3.2.3 *There exists an absolute constant C for which the following holds. For any class of functions F and every $0 < \epsilon < 1$, the empirical and actual structures are (λ_n, ϵ)-close, provided that*

$$\lambda_n \geq C \max \left\{ \mathbb{E}||P - P_n||_F, \sigma_F \sqrt{\frac{\log(1/\epsilon)}{n}}, \frac{b \log(1/\epsilon)}{n} \right\},$$

where $\sigma_F^2 = \sup_{r \in F} var(r)$ and $b = \sup_{r \in F} ||r||_\infty$.

The following lemma provides bounds to the first component of the right-hand side. It shows that the estimate above cannot be improved by more than a constant factor when n is sufficiently large.

Lemma 3.2.4 *There are absolute constants c, c' and C for which the following holds. Let F be a class of functions satisfying $\sup_{r \in F} ||r||_\infty \leq 1$ and set $\sigma_F^2 = \sup_{r \in F} var(r)$. Then*

$$\mathbb{E}||P - P_n||_F \geq c \frac{\sigma_F}{\sqrt{n}}.$$

Furthermore, for every integer $n \geq 1/\sigma_F^2$, with probability at least c',

$$||P - P_n||_F \geq C\mathbb{E}||P - P_n||_F$$

Proof of these two lemmas can be found in the appendix of [26], thus omitted here. They show that it would be impossible to use this notion of similarity to obtain an asymptotic result stronger than $\mathbb{E}\hat{r} \leq 1/\sqrt{n}$ with high probability. However, it is enough to derive the consistency property of the empirical minimizer in the next theorem.

Theorem 3.2.1 *Suppose $(p^*, f^*) \in X$ is a unique minimizer of the log-likelihood type functional $\ell(p, f)$, and $(\hat{p}_n, \hat{f}_n)$ is the corresponding empirical minimizer given a sample of size n. Then $(\hat{p}_n, \hat{f}_n) \to (p^*, f^*)$ in probability as $n \to \infty$.*

Proof By Lemma 3.2.3 and Lemma 3.2.4, for any $\lambda, \sigma > 0$, there is a constant integer $N > 0$ such that for any $n > N$, the empirical and actual structures are (λ_n, ϵ)-close and $\lambda_n < \lambda$. Thus the empirical and actual structures are (λ, ϵ)-close as well. Let $\hat{r}_n \in F$ be the empirical minimizer given a sample of size n and let $(\hat{p}_n, \hat{f}_n) \in X$ correspond to it, then

$$Pr\{|\mathbb{E}\hat{r}_n| \leq \lambda\} = Pr\{\mathbb{E}\hat{r}_n \leq \lambda\} \geq 1 - \epsilon \tag{3.8}$$

according to the discussion above. Therefore, $\mathbb{E}\hat{r}_n \to 0$ in probability which is equivalent to $\mathbb{E}L(x|\hat{p}_n, \hat{f}_n) - \mathbb{E}L(x|p^*, f^*) \to 0$ in probability. By the assumption of the uniqueness of the minimizer $(p^*, f^*) = \text{argmin}_{(p,f)\in X} \mathbb{E}L(x|p, f)$ and the continuity of $\mathbb{E}L(x|p, f)$ with respect to p and f, it follows that $(\hat{p}_n, \hat{f}_n) \to (p^*, f^*)$ in probability. ■

3.3 Convergence

We begin with the general setting of the type of inverse problem we consider. The setting is similar, but distinct from that recently introduced in a number of publications on the use of Tikhonov-type regularization to solve inverse problems with Poisson data; see e.g. [28], [30], and [31]. In their work, they begin with three Banach spaces with properly chosen topologies: $(X, \tau_X), (Y, \tau_Y)$, and (Z, τ_Z). The first two topologies τ_X and τ_Y are weak topologies in X and Y, respectively while τ_Z is the strong (norm-based) topology on Z. They also let $F : X \to Y$ be a mapping defined on X and taking values in Y. The three Banach spaces can be interpreted in this way, i.e. X is a solution space, Y is the space of the right-hand sides as the underlying structure, and Z is the data space we can observe. For simplicity and the current task, we will not distinguish the data space from the pace of the right-hand sides, i.e. $Z = Y$. Our goal is approximation of a solution of the ill-posed equation

$$F(x) = y, \tag{3.9}$$

with $x \in X$ which is the space of the pair (p, f), and the given right-hand side $y \in Y$ which is the space of the mixture density g. For us, ill-posedness means that the solutions do not depend continuously on the right-hand side. This definition is considerably narrower than the classical definition of ill-posedness due to Hadamard that is also tied up with existence and uniqueness of solutions. In using a narrow definition of ill-posedness, we follow some of the modern research on inverse problems as in e.g. [29].

Typically, to produce numerical solutions one needs to overcome ill-posedness first. Direct numerical solution of ill-posed equations is impossible, because the slightest difference like rounding errors in y can lead to arbitrarily large deviations of the calculated solution from the exact solution in X. The first step in overcoming ill-posedness will be the switch from the original mapping F to the slightly "perturbed" mapping F_h that depends on the parameter (bandwidth) $h \in (0, \infty)$. With the new operator F_h in mind, we define an updated problem

$$F_h(x) = y. \tag{3.10}$$

instead of the original problem (3.9). Typically, when handling inverse problems, the distance between $F_h(x)$ and y is quantified using some *fitting functional* $S : Y \times Y \to [0, \infty)$. The choice of the functional must be such that, for two elements y_1, y_2 of Banach space Y, $S(y_1, y_2) = 0$ if and only if $y_1 = y_2$. However, this is typically not enough to avoid ill-posedness. We also introduce an appropriate regularizing or stabilizing functional $\Omega : X \to (-\infty, \infty]$. In out case, this functional also depends on the bandwidth parameter h, which is thus denoted $\Omega_h(x)$. It is assumed that $\Omega_h(x) \to \Omega(x)$ in the appropriate operator norm as $h \to 0$. With this in mind, a minimization problem

$$T_h(x) := S(y, F_h(x)) + \Omega_h(x) \to \min_{x \in X} \tag{3.11}$$

is considered. The functional $T_h(x)$ is referred to as *Tikhonov-type functional*.

The reason this framework is of interest is as follows. Consider a two-component univariate nonparametric density mixture with one known component. Such a problem can be formulated as

$$(1-p)f_0(v) + pf(v) = g(v), \tag{3.12}$$

where $f_0(v)$ is a known density component, $g(v)$ is a known target density while $0 < p < 1$ is an unknown weight and $f(v)$ is an unknown density function. For the most part, we will be viewing functions f and g as elements of a functional space. Because of this, their arguments will, in general, be omitted to make the notation less cluttered. Several assumptions have to be imposed on the unknown function f to ensure solution convergence in the sense soon to be defined. First, for ease of handling, we assume that the function f is defined on a compact set that can be assumed without loss of generality to be $[0, 1]$. Some smoothness assumptions on f will also be necessary; in particular, we will assume that $f \in C^1[0, 1]$ where $C^1[0, 1]$ is a Banach space of continuously differentiable functions on $[0, 1]$. In the future, we will omit $[0, 1]$ in this notation and write simply C^1 unless specified to the contrary. It is also convenient to define the Banach space $X = \mathbb{R} \times C^1$ to which pairs of unknown probability p and unknown density function f will belong. Finally, we also assume that a kernel function K used in the definition of the nonlinear smoother $\mathcal{N}_h$ is once continuously differentiable as well. With this in mind, we can now define a linear forward operator F as

$$F(p, f) := (1-p)f_0 + pf. \tag{3.13}$$

This implies that the problem (3.12) can be thought of as a special case of (3.9).

Now we can consider our minimization problem with respect to the log-likelihood like functional $\ell(p, f)$ in a similar framework. Define another nonlinear smoothing forward operator $F_h : \mathcal{D}(F) \subseteq \{[0, 1] \times C^1\} \to C^1$ as

$$F_h(p, f) := (1-p)f_0 + p\mathcal{N}_h f \tag{3.14}$$

where $\mathcal{N}_h$ is the nonlinear smoothing operator and where $\mathcal{D}(F) = [0,1] \times \{f(v) : f \in C^1, \exists A \in R : |f''(v)| \leq A \, \forall v \in [0,1], f(v) \geq \eta > 0 \text{ a.e. }, \int_0^1 f(v)\, dv = 1\}$. Note that the functions included in this domain are not just continuously differentiable but their second derivatives are also uniformly bounded (but not necessarily continuous). Instead of the original problem (3.12), we consider the approximation of another ill-posed equation

$$F_h(p, f) = g \tag{3.15}$$

Now, we define a stabilizing functional $\Omega_h(x) := p\left\{1 - \int \mathcal{N}_h(f)\right\}$ and the Kullback-Leibler divergence (KL) as the fitting functional S. Then, the minimization problem of functional $\ell(p, f)$ becomes exactly a minimization problem of Tikhonov-type functional

$$\ell(p, f) = T_h(g, F_h(x)) = S(g, F_h(p, f)) + p\left\{1 - \int \mathcal{N}_h(f)\right\} \to \min_{x \in X}. \tag{3.16}$$

Depending on the context, we will use either notation $T_h(g, x)$ or, if we need to stress the role of the forward operator F_h, $T_h(g, F_h(x))$.

The resulting framework is similar to that of [28] and [29] if $Z = Y$ is assumed, i.e.

$$T_\alpha(x) := S(F(x), y) + \alpha\Omega(x) \to \min_{x \in X}, \tag{3.17}$$

except for two aspects. First, the fitting functional $\Omega_h(x)$ depends on the regularization parameter h and converges to zero as $h \to 0$ for any $x \in X$. In the earlier framework [28] and [29], the fitting functional takes the form $\Omega(x)$ which does not depend on the regularization parameter; instead, the second term of the functional T that they minimize is defined as $\alpha\Omega(x)$ for some stabilizing functional $\Omega(x)$ that does not depend on the regularization parameter α. Second, in our case,the forward operator F_h depends on our regularization parameter h which is not the case in the original framework.

To establish practically important results of existence and stability, we need to impose a set of assumptions on the fitting functional, forward operator, and sta-

bilizing functional. In our problem, the following properties meet those necessary assumptions.

Assumption 1 *Assumptions imposed on F_h: $\mathcal{D}(F) \subseteq X \to C^1$*

1. *F_h is sequentially continuous with respect to the weak topology of the space X, i.e. if $x_k \rightharpoonup x$ for $x, x_k \in \mathcal{D}(F)$, then we have $F_h(x_k) \rightharpoonup F_h(x)$*

2. *$\mathcal{D}(F)$ is sequentially closed with respect to the weak topology on X, that is $x_k \rightharpoonup x$ for $\{x_k\} \in \mathcal{D}(F)$ implies that $x \in \mathcal{D}(F)$.*

Assumption 2 *Assumptions imposed on the fitting functional $S : C^1 \times C^1 \to [0, \infty)$:*

3. *$S(g, v)$ is sequentially lower semi-continuous with respect to the weak topology on $C^1 \times C^1$, that is if $g_k \rightharpoonup g$ and $v_k \rightharpoonup v$, then $S(g, v) \leq \liminf_{k\to\infty} S(g_k, v_k)$.*

4. *If $S(g, v_k) \to 0$ then there exists some $v \in C^1$ such that $v_k \rightharpoonup v$.*

5. *If $v_k \rightharpoonup v$ and $S(g, v) < \infty$, then $S(g, v_k) \to S(g, v)$.*

Assumption 3 *Assumptions imposed on $\Omega_h : \mathcal{D}(F) \times (0, \infty) \to [0, 1]$:*

6. *$\Omega_h(x)$ is sequentially lower semicontinuous with respect to the weak topology in X, that is, if $f_k \rightharpoonup f$ for $f, f_k \in C^1$, $p_k \to p$, we have $\Omega_h(x) \leq \liminf_{k\to\infty} \Omega_h(x_k)$ for any positive h.*

7. *The sets*

$$M_{\Omega_h}(c) := \{x \in \mathcal{D}(F) : \Omega_h(x) \leq c\}$$

are sequentially compact with respect to the weak topology on X for all $c \in \mathbb{R}$, that is each sequence in $M_{\Omega_h}(c)$ has a subsequence that is convergent in the weak topology on X.

Note that in the case of our optimization problem (3.16) the domain of the forward operator is not the same as the Banach space $X = \mathbb{R} \times C^1(D)$. The reason larger

space X has to be considered is that it is a Banach space, unlike $\mathcal{D}(F)$ itself. This, however, makes it necessary that assumptions concerning the operator F and the functional Ω_h be satisfied on $\mathcal{D}(F)$. To verify this, we set first $\tilde{X} = \mathcal{D}(F)$ and $\tau_{\tilde{X}}$ as the topology induced on $\tilde{X}$ by τ_X. Then, the restriction $\tilde{F} := F|_{\tilde{X}}$ is clearly sequentially continuous. Next, let us define the restriction $\tilde{\Omega}_h := \Omega_{h,\tilde{X}}$ and note that its sublevel set $M_{\tilde{\Omega}_h}(c) = M_{\Omega_h}(c) \cap \mathcal{D}(F)$ and so is closed as an intersection of closed sets. Since $M_{\tilde{\Omega}_h}(c) \subseteq M_{\Omega_h}(c)$, it is a closed subset of a compact set, and is thus a compact set. With this discussion in mind, in the future we will conduct the exposition of ideas as if the domain $\mathcal{D}(F)$ coincided with the Banach space X.

Lemma 3.3.1 *Assume that the kernel function K is once continuously differentiable. Then, the optimization problem* (3.16) *satisfies all of the three assumptions listed above.*

Proof We start with the Assumption 1(i). Note that since the operator F_h depends linearly on p and the weak convergence for a sequence $\{p_k\} \in \mathbb{R}$ is just an ordinary convergence of a sequence, it is enough to prove that $f_k \rightharpoonup f$ implies $F_h(p, f_k) \rightharpoonup F_h(p, f)$ for a fixed p, or $\mathcal{N}_h f_k \rightharpoonup \mathcal{N}_h f$ equivalently. Recall that the weak convergence for a sequence of functions $\{f_k\} \in C^1$ implies that $f_k'(v) \to f'(v)$ for any $v \in [0,1]$, $f_k(0) \to f(0)$ and $\sup_k \sup_{v\in[0,1]} |f_k'(v)| < \infty$. These properties also imply pointwise convergence of $\{f_k\}$. With these in mind, it is easy to show for any $v \in [0,1]$, $\mathcal{N}_h f_k(v) = \exp\{\int K_h(v-u) \log f_k(u) du\}$ converges to $(\mathcal{N}_h f)(v) = \exp\{\int K_h(v-u) \log f(u) du\}$, since $\{f_k\}$ and f are bounded away from zero. It follows that $(\mathcal{N}_h f_k)'(v) = (\mathcal{N}_h f_k)(v) \int K_h'(v-u) \log f_k(u) du$ converges to $(\mathcal{N}_h f)'(v) = (\mathcal{N}_h f)(v) \int K_h'(v-u) \log f(u) du$ for any $v \in [0,1]$. Next, note that $\sup_k \sup_{v\in[0,1]} |(\mathcal{N}_h f_k)'(v)|$ is clearly bounded. Finally, $(\mathcal{N}_h f_k)(0) \to (\mathcal{N}_h f)(0)$ by the dominated convergence theorem for any $f \in \mathcal{D}(F_h)$. All of the above imply weak convergence of $\{\mathcal{N}_h f_k\}$ to $\mathcal{N}_h f$.

To prove the Assumption 1(ii), we first note that if $f_k(v) \geq \eta > 0$ for any k and $v \in [0,1]$, we have immediately that $f(v) = \lim_{k\to\infty} f_k(v) \geq \eta > 0$ for any

f that is a pointwise limit of $\{f_k\}$. Moreover, if $|f_k^{'}(v)| \leq L$ for some $L > 0$ and any k, it implies that $|f^{'}(v)| \leq L$ as well due to continuity of the absolute value function. Finally, if a sequence f_k converges to f weakly, the integral $\int_0^1 f(v)\,dv = \int_0^1 \lim_{k\to\infty} f_k(v)\,dv = \lim_{k\to\infty} \int_0^1 f_k(v)\,dv = 1$ by the dominated convergence theorem (because density functions belonging to $\mathcal{D}(F_h)$ are bounded on $[0,1]$).

The fitting functional S is a Kullback-Leibler functional; the fact that it satisfies Assumption 2(iii)(iv)(v) has been demonstrated several times in optimization literature concerned with variational regularization with non-metric fitting functionals. The details can be found in e.g. [28] and [29].

The sequential lower semi-continuity of the stabilizing functional Ω_h in Assumption 3(vi) is guaranteed by Fatou's Lemma. Indeed, let us define

$$\phi_k(v) = p_k \left[\mathcal{S} f_k(v) - \mathcal{N}_h f_k(v)\right]. \tag{3.18}$$

Then, due to Jensen's inequality, $\{\phi_k\}$ is a sequence of non-negative measurable functions. Let f_k converge weakly to f in $C^1[0,1]$ and recall that this implies pointwise convergence. Define the function $\phi(v) = \liminf_{k\to\infty} \phi_k(v)$, and observe that

$$\begin{aligned} \Omega_h(x) &= p \int_0^1 (\mathcal{S}_h f - \mathcal{N}_h f)(v)\,dv \\ &= \int_0^1 \liminf_{k\to\infty} p_k[\mathcal{S}_h f_k - \mathcal{N}_h f_k](v)\,dv \\ &\leq \liminf_{k\to\infty} \int_0^1 \phi_k(v)\,dv = \liminf_{k\to\infty} \Omega_h(x_k) \end{aligned} \tag{3.19}$$

Therefore, $\Omega_h : \mathcal{D}(F) \times (0,\infty) \to [0,1]$ is lower semi-continuous with respect to the weak topology on X.

Last, we will justify the Assumption 3(vii). Consider a sequence $\{p_k, f_k\} \in M_{\Omega_h}(c)$ for any $c > 0$. By definition of $M_{\Omega_h}(c)$, $\{p_k, f_k\} \in \mathcal{D}(F)$ implies that $\{f_k\}$ are uniformly bounded and equicontinuous respectively. Since all $f_k \in \mathcal{D}(F_h)$, all of them have a uniformly bounded second derivative and so the sequence $\{f_k'\}$ is also equicontinuous as well. By Arzelà-Ascoli theorem, there is a subsequence of $\{f_k\}$ convergent in the norm topology of C, and there is also a subsequence of the subsequence which

has a convergent first derivative in the norm topology of C. This implies that the last subsequence, if indexed by k_l and written as $\{f_{k_l}\}$, is convergent in the norm topology of C^1. And by Bolzano-Weierstrass theorem, $\{p_{k_l}\}$ has a convergent subsequence $\{p_{k_{l_m}}\}$. This, of course, means that $\{p_{k_{l_m}}, f_{k_{l_m}}\}$ converges in the weak topology on X as well.

■

F_h and Ω_h are also continuous with respect to the bandwidth h. Note that both F_h and Ω_h have the nonlinearly smoothing operator $\mathcal{N}_h f$ as a component. If the kernel function is chosen with good properties like Lipschitz continuity, for any $f \in C^1$, it is easy to see $\mathcal{N}_h f(v) = \exp(\int K_h(v-u)\log(f(u))\,du)$ is continuous with respect to h by Taylor's expansion, so is F_h. And Ω_h has the integration of $\mathcal{N}_h f$ on a compact set, thus being continuous with respect to h, too.

The first result we want to prove is that of existence.

Theorem 3.3.1 ***(Existence)*** *For any choice of h, the minimization problem of* (3.16) *has a solution. A minimizer $x^* = (p^*, f^*) \in X$ satisfies $T_h(g, x^*) < \infty$ if and only if there exists an element $\bar{x} = (\bar{p}, \bar{f}) \in X$ such that $S(g, F_h(\bar{x})) < \infty$.*

Proof First, define $c := \inf_{x \in X} T_h(x)$. The trivial case $c = \infty$ can only occur if there is no x such that $S(g, F(\bar{x})) < \infty$. Excluding that case, we can choose a sequence $\{x_k\} \in X$ such that $T_h(g, x_k) \to c$. By definition of the functional $T_h(g, x)$, we have

$$\Omega_h(x_k) \leq T_h(g, x_k) \leq c + 1$$

for sufficiently large k. By compactness of sublevel sets of Ω_h there is a subsequence x_{k_l} that converges to some $\tilde{x} \in X$. The continuity of F_h implies that $F_h(x_{k_l}) \to F_h(\tilde{x})$. Since the fitting functional S and the stabilizing functional Ω_h are lower semicontinuous, we have

$$T_h(g, \tilde{x}) \leq \liminf_{l \to \infty} T_h(g, x_{k_l}) = c.$$

Thus, $\tilde{x}$ is a minimizer of $T_h(g, x)$. ■

As a next step, we need to check if the problem can be solved numerically in a meaningful way. In other words, it is necessary to establish that small changes in the "input" g and the amount of regularization used (that is characterized by the bandwidth h) cannot result in arbitrarily large changes in minimizers of the problem (3.16). The following result suggests that it is true.

Theorem 3.3.2 ***(Stability)*** *Let $g \in Y$ and $h \in (0, \infty)$ be fixed. Assume that $\{g_k\}$ is a sequence in Y such that $g_k \to g$ and $\{h_k\}$ is a sequence of bandwidths in $(0, \infty)$ converging to h. Also, let $\{\varepsilon_k\}$ be a sequence in $[0, \infty)$ converging to zero. Finally, assume that there exists an element $\bar{x} \in X$ with $S(g, F_h(\bar{x})) < \infty$.*

Then, each sequence $\{x_k\}$ with $T_{h_k}(g_k, x_k) \leq \inf_{x \in X} T_{h_k}(g_k, x) + \varepsilon_k$ has a τ_X-convergent subsequence, and for sufficiently large k the elements x_k are such that $T_{h_k}(g_k, x_k) < \infty$. Each limit $\tilde{x} \in X$ of a τ_X-convergent subsequence $\{x_{k_l}\}$ is a minimizer of $T_h(g, x)$ and we have $T_{h_{k_l}}(g_{k_l}, x_{k_l}) \to T_h(g, \tilde{x})$, $\Omega_{h_{k_l}}(x_{k_l}) \to \Omega_h(\tilde{x})$ and thus also $S(g_{k_l}, F_{h_{k_l}}(x_{k_l})) \to S(g, F_h(\tilde{x}))$.

Proof First, since $g_k \to g$ and $S(g, F_h(\bar{x})) < \infty$, we have $S(g_k, F_h(\bar{x})) \to S(g, F_h(\bar{x}))$. Thus, $S(g_k, F_h(\bar{x})) < \infty$ for all sufficiently large k. Therefore, without loss of generality, we can assume that $S(g_k, F_h(\bar{x})) < \infty$ for all k. By Theorem 3.3.1, there exist minimizers $x_k^* \in argmin_{x \in X} T_{h_k}(g_k, x)$ and that $T_{h_k}(g_k, x_k^*) < \infty$.

Note that we imposed rather strong assumptions on the domain of the forward operator $\mathcal{D}(F)$. Because of these assumptions, a sequence $\{x_k\}$ defined in the statement of the Theorem has a τ_X-convergent subsequence. Now, let $\{x_{k_l}\}$ be such a convergent subsequence with the limit $\tilde{x} \in X$. We know that for all $x_{h,g} \in argmin_{x \in X} T_h(g, x)$ we have, by Theorem 3.3.1, $S(g, F_h(x_{h,g})) < \infty$; therefore, using lower semicontinuity of the functional T, we have

$$\begin{aligned} T_h(g, \tilde{x}) &\leq \liminf_{l \to \infty} T_h(g_{k_l}, x_{k_l}) \\ &\leq \limsup_{l \to \infty} T_h(g_{k_l}, x_{k_l}) \\ &= \limsup_{l \to \infty} T_{h_{k_l}}(g_{k_l}, x_{k_l}) + \left(\Omega_h(x_{k_l}) - \Omega_{h_{k_l}}(x_{k_l}) \right). \end{aligned} \tag{3.20}$$

As $\Omega_h(x)$ is continuous with respect to h for any fixed value of x, the last difference term in the above inequality can be bounded by an arbitrarily small δ_{k_l} as h_{k_l} converges to h. Using the already developed argument, we continue to obtain

$$\begin{aligned} T_h(g,\tilde{x}) &\leq \limsup_{l\to\infty}[T_{h_{k_l}}(g_{k_l},x_{k_l})+\delta_{k_l}] \\ &\leq \limsup_{l\to\infty}[T_{h_{k_l}}(g_{k_l},x^*_{k_l})+\varepsilon_{k_l}+\delta_{k_l}] \\ &\leq \limsup_{l\to\infty}[T_{h_{k_l}}(g_{k_l},x_{h,g})+\varepsilon_{k_l}+\delta_{k_l}] \\ &= \lim_{l\to\infty} S(g_{k_l},F_{h_{k_l}}(x_{h,g}))+\Omega_{h_{k_l}}(x_{h,g})+\varepsilon_{k_l}+\delta_{k_l}] = T_h(g,x_{h,g}) \end{aligned} \tag{3.21}$$

and so $\tilde{x}$ minimizes $T_h(g,x)$.

Assume $\Omega_{h_{k_l}}(x_{k_l}) \nrightarrow \Omega_h(\tilde{x})$. Then the sequentially lower semicontinuity of $\Omega_h(x)$ implies

$$c := \limsup_{l\to\infty}\Omega_{h_{k_l}}(x_{k_l}) > \liminf_{l\to\infty}\Omega_{h_{k_l}}(x_{k_l}) \geq \Omega_h(\tilde{x}). \tag{3.22}$$

If $\{x_{k_{l_m}}\}$ is a subsequence of $\{x_{k_l}\}$ with the limit of c, there must be

$$\begin{aligned} &\lim_{m\to\infty} S(g_{k_{l_m}},F_{h_{k_{l_m}}}(x_{k_{l_m}})) \\ &= \lim_{m\to\infty}(T_{h_{k_{l_m}}}(g_{k_{l_m}},x_{k_{l_m}})-\Omega_{h_{k_{l_m}}}(x_{k_{l_m}})) \\ &= T_h(g,\tilde{x})-c \\ &= S(g,F_h(\tilde{x}))+\Omega_h(\tilde{x})-c \\ &< S(g,F_h(\tilde{x})), \end{aligned} \tag{3.23}$$

which contradicts the lower semicontinuity of S. ■

As the last step, we expect the minimizers of the problem (3.16) converges to the true solution as the bandwidth $h \to 0$. This is the rationale of considering to minimize the penalized smoothed likelihood functional and one of the main purpose in this chapter.

Theorem 3.3.3 ***(Convergence)*** *Let $\{h_k\}$ be a sequence of bandwidths converging to zero. Further, let $\{x_k\}$ be a sequence in X with $x_k \in argmin_{x\in X} T_{h_k}(g,x)$. If*

$S(g, F_{h_k}(x_k)) \to 0$, then $\{x_k\}$ has a τ_X-convergent subsequence and each limit of a τ_X-convergent subsequence is a solution to (3.9).

Proof It is necessary to check if $\{h_k\}$ and $\{x_k\}$ can guarantee $S(g, F_{h_k}(x_k)) \to 0$ as $k \to \infty$. Suppose there exists a solution $\bar{x} \in X$ of (3.12), i.e. $S(g, F(\bar{x})) = 0$ (so $\bar{x}$ is an S-generalized solution) and $\bar{x} = \operatorname{argmin}_{x \in X} T(g, x)$. Then, since $x_k \in \operatorname{argmin}_{x \in X} T_{h_k}(g, x)$, we have

$$\begin{aligned} S(g, F_{h_k}(x_k)) &= T_{h_k}(g, F_{h_k}(x_k)) - \Omega_{h_k}(x_k) \\ &\leq T_{h_k}(g, F_{h_k}(\bar{x})) - \Omega_{h_k}(x_k) \\ &= S(g, F_{h_k}(\bar{x})) + \Omega_{h_k}(\bar{x}) - \Omega_{h_k}(x_k) \to 0. \end{aligned} \tag{3.24}$$

Next, we can show directly that, since $\lim_{h \to 0} \mathcal{N}_h f(v) = f(v)$ for any $f \in C$, we automatically obtain $\Omega_{h_k}(\bar{x}) \to 0$. The convergence of $\Omega_{h_k}(x_k)$ to zero will be shown separately.

The existence of τ_X-convergent subsequence of $\{x_k\}$ is again guaranteed by definition of the domain of $\mathcal{D}(F)$. Let $\{x_{k_l}\}$ be an arbitrary subsequence of $\{x_k\}$ converging to some element $\tilde{x} \in X$. Since $S(g, v)$ is continuous with respect to the second component v, and $F_h(x)$ is continuous with respect to x and h, it implies

$$S(g, F(\tilde{x})) = \lim_{l \to \infty} S(g, F_{h_{k_l}}(x_{k_l})) = 0, \tag{3.25}$$

that is $S(g, F(\tilde{x})) = 0$. Thus $\tilde{x} \in X$ is a solution of (3.9).

■

In the proof of Theorem 3.3.3, we need to show that $\Omega_{h_k}(x_k)$ goes to zero as $h_k \to 0$. First of all, by definition,

$$\begin{aligned} \Omega_{h_k}(x_k) &= p_k \left\{ 1 - \int_0^1 \mathcal{N}_{h_k} f_k(v) dv \right\} \\ &= p_k \left\{ \int_0^1 \left[f_k(v) - \mathcal{N}_{h_k} f_k(v) \right] dv \right\}. \end{aligned} \tag{3.26}$$

Now, represent the difference $f_k(v) - \mathcal{N}_{h_k} f_k(v)$ as

$$\exp\left\{\int K_{h_k}(v-u)\log f_k(v)\,du\right\} - \exp\left\{\int K_{h_k}(v-u)\log f_k(u)\,du\right\} \tag{3.27}$$
$$:= \exp(A) - \exp(B)$$

and apply Taylor's formula for the exponent function at A or B. If the first derivative is bounded everywhere for the densities we consider, the first term can be bounded away from zero. The difference will be $\int K_{h_k}(v-u)\log f_k(v)\,du - \int K_{h_k}(v-u)\log f_k(u)\,du$. To make it easier, recall that $K_{h_k}(v-u) = \frac{1}{h_k}K\left(\frac{v-u}{h_k}\right)$ and so

$$\begin{aligned}\int K_{h_k}(v-u)\log f_k(u)\,du &= \frac{1}{h_k}\int K(u')\log f_k(x-h_k u')(-h_k)\,du \\ &= \int K(u')\log f_k(x-h_k u')\,du'\end{aligned} \tag{3.28}$$

using the substitution $\frac{x-u}{h_k} = u'$. For simplicity, from now on u is used instead of u'. Using the same substitution in the first integral of the difference, we get

$$\begin{aligned}&\int K(u)\log f_k(x)\,du - \int K(u)\log f_k(x-h_k u)\,du \\ &= \int K(u)\left[\log f_k(x) - \log f_k(x) + \frac{h_k u}{f_k(\theta)}f_k'(\theta)\right]du \\ &= \int K(u)\frac{h_k u}{f_k(\theta)}f_k'(\theta)\,du\end{aligned} \tag{3.29}$$

for some $0 < \theta < h_k$. For densities in $\mathcal{D}(F)$ and $\int uK(u)\,du$ chosen to be finite, the last term will go to zero as $h_k \to 0$ for any reasonable sequence.

4. NUMERICAL STUDY

4.1 Introduction

In this chapter, we focus on the performance of our algorithm. In Section 4.2, we apply our algorithm on simulated data from different settings. In Section 4.3, we talk about method of selecting bandwidth h to improve the estimation. In Section 4.4, we show the advantages of our algorithm by the comparison of our algorithm with the symmetrization method of [13]. In Section 4.5, we present an application on a real dataset.

4.2 Simulation Examples

For our first experiment, we generate n independent and identically distributed observations from a two component normal-gamma mixture with the density $g(x)$ supported on the positive half real line. We will use the notation $I_{[x>0]}$ for the indicator function of the positive half of the real line and $\phi(x)$ for the standard Gaussian distribution. Thus, the known component is

$$f_0(x) = \frac{2}{\sigma}\phi\left(\frac{x-\mu}{\sigma}\right) I_{[x>0]}, \tag{4.1}$$

while the unknown component is $Gamma(\alpha, \beta)$, i.e.,

$$f(x) = \frac{\beta^\alpha}{\Gamma(\alpha)} x^{\alpha-1} e^{-\beta x} I_{[x>0]}. \tag{4.2}$$

Note that we truncate the normal distribution so that it stays on the positive half of the real line. We choose the sample size $n = 500$, the probability $p = 0.6$, $\mu = 6$, $\sigma = 1$, $\alpha = 2$ and $\beta = 1$. The initial weight is $p_0 = 0.2$ and the initial assumption

for the unknown component distribution is $Gamma(4, 2)$. The rescaled triangular function

$$K_h(x) = \frac{1}{h}\left(1 - \frac{|x|}{h}\right) I(|x| \leq h) \tag{4.3}$$

is used as the kernel function. We use a fixed bandwidth throughout the sequence of iterations and this fixed bandwidth is selected according to the classical Silverman's rule of thumb that we describe here briefly for completeness; for more details, see [32]. Let SD and IQR be the standard deviation and interquartile range of the data, respectively. Then, the bandwidth is determined as $h = 0.9 \min\left\{SD, \frac{IQR}{1.34}\right\} n^{-1/5}$. We use the absolute difference $|p_n^{t+1} - p_n^t|$ as a stopping criterion; at every iteration step, we check if this difference is below a small threshold value d that depends on required precision. If it is, the algorithm is stopped. The analogous rule has been described for classical parametric mixtures in [1]. In our setting, we use the value $d = 10^{-5}$. The computation ends after 259 iterations, with an estimate $\hat{p} = 0.6661$; the Figure 4.2 shows the true and estimated mixture density function $g(x)$ while the Figure 4.2 shows both true and estimated second component density f. Both figures show a histogram of the observed target distribution $g(x)$ in the background. Both the fitted mixture density $\hat{g}(x)$ and the fitted unknown component density function $\hat{f}(x)$ are quite close to their corresponding true density functions everywhere.

In the second experiment, we generate n i.i.d. observations from a two component beta-beta mixture with the density $g(x)$ supported on the compact interval $(0, 1)$. The known component is $f_0(x) = Beta(0.5, 0.5)$, while the unknown component is $f(x) = Beta(2, 2)$. The sample size is set as $n = 1000$ and the probability weight $p = 0.6$. We assume that the starting value of the probability weight is $p_0 = 0.3$ and the initial assumption for the unknown component distribution is $Beta(4, 4)$. The rescaled triangular kernel $K_h(x) = \frac{1}{h}\left(1 - \frac{|x|}{h}\right) I(|x| \leq h)$ is used with a fixed bandwidth $h = 0.9 \min\left\{SD, \frac{IQR}{1.34}\right\} n^{-1/5}$. The algorithm is stopped when the absolute difference $|p_n^{t+1} - p_n^t| < 10^{-5}$. The computation ends after around 80 iterations, with an estimate $\hat{p} = 0.601$; the Figure 4.2 shows the true and estimated mixture density function $g(x)$ while the Figure 4.2 shows both true and estimated second component density f. Both

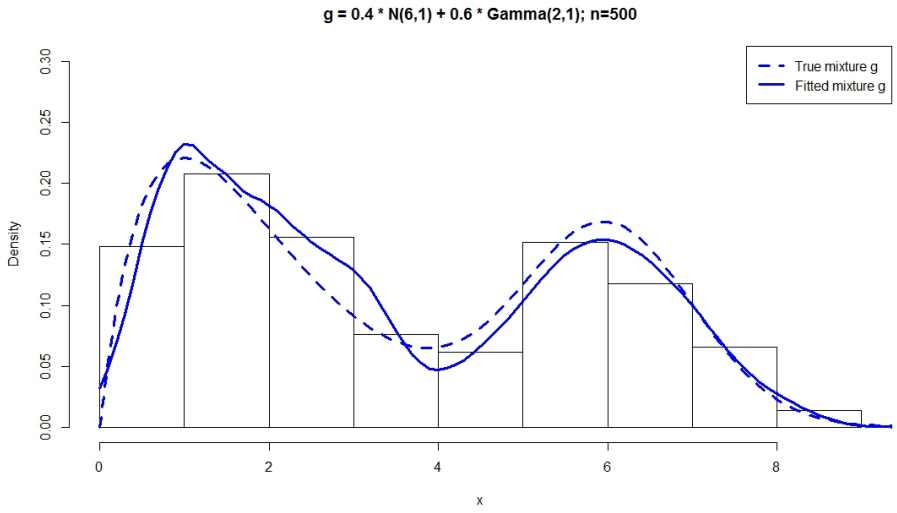

Figure 4.1. Fitted mixture density for a mixture of Gaussian(6,1) and Gamma(2,1)

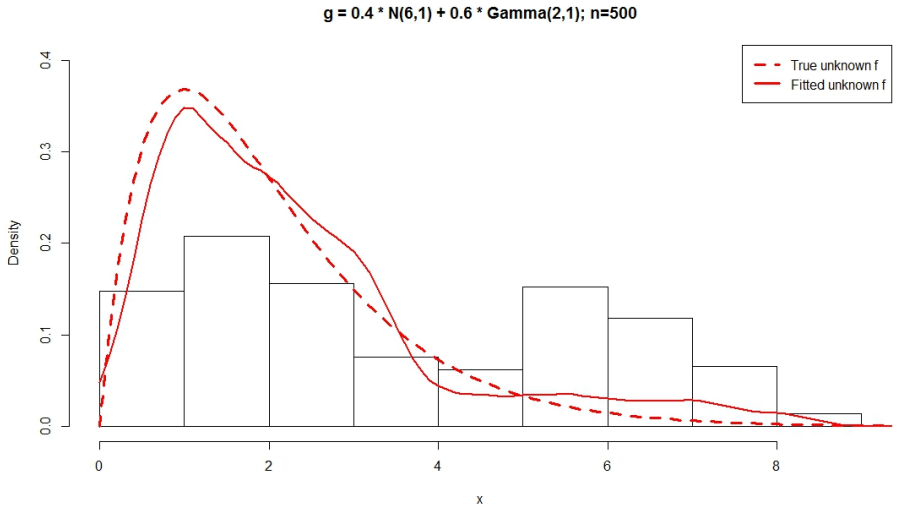

Figure 4.2. Fitted unknown component density for a mixture of Gaussian(6,1) and Gamma(2,1)

figures show a histogram of the observed target distribution $g(x)$ in the background. Note that both the fitted mixture density $\hat{g}(x)$ and the fitted unknown component density function $\hat{f}(x)$ are quite close to corresponding true density functions.

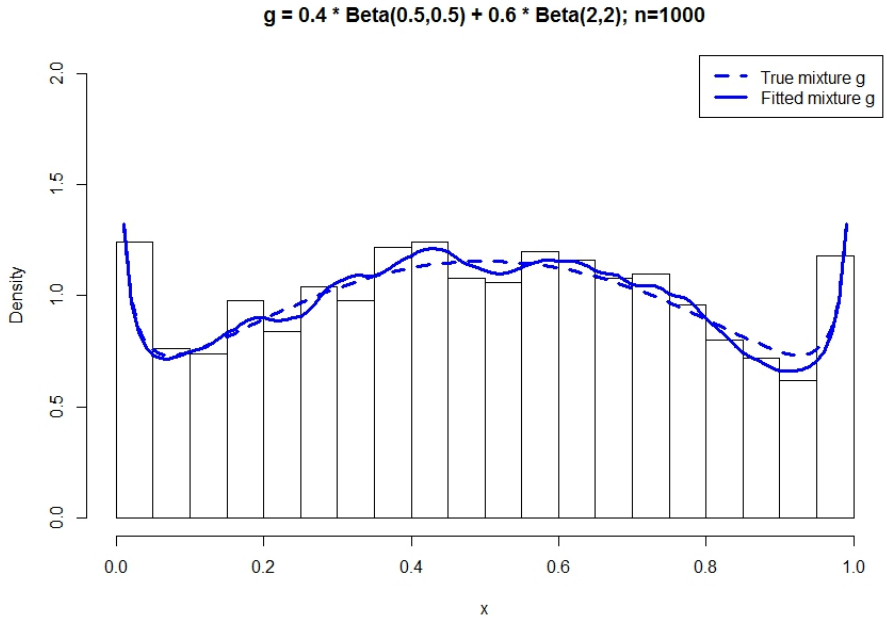

Figure 4.3. Fitted mixture density for a mixture of Beta(0.5,0.5) and Beta(2,2)

In the third experiment, we would like to see how well our algorithm can recover the unknown component which is not unimodal. We generate n i.i.d. observations from a two component Gaussian-bimodal mixture with the density $g(x)$ defined as $g(x) = (1-p)f_0(x) + pf(x)$. The known component is $f_0(x) = Gaussian(0,1)$, while the unknown component is bimodal distribution, i.e. a mixture of $Gaussian(5,1)$ and $Gaussian(8,0.5)$ with equal proportion. Thus both of the known and unknown components having a support on the real line. The sample size is set as $n = 1000$ and the mixture proportion is $p = 0.5$. We assume that the starting value of the mixture proportion is $p_0 = 0.4$ and the initial assumption for the unknown component distribution is a unimodal distribution $Gaussian(4,2)$. The rescaled triangular kernel $K_h(x) = \frac{1}{h}\left(1 - \frac{|x|}{h}\right) I(|x| \leq h)$ is used with a fixed bandwidth $h = 0.9 \min\left\{SD, \frac{IQR}{1.34}\right\} n^{-1/5}$. The algorithm is stopped when the absolute difference $|p_n^{t+1} - p_n^t| < 10^{-5}$. The computation ends after around 120 iterations, with an estimate $\hat{p} = 0.5059$; the Figure 4.2 shows the true and estimated mixture density

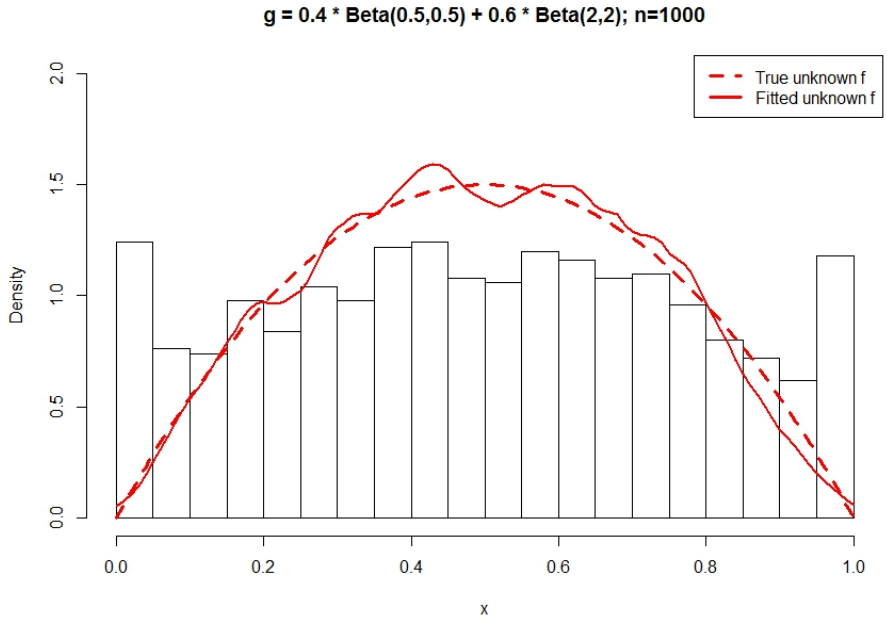

Figure 4.4. Fitted unknown component density for a mixture of Beta(0.5,0.5) and Beta(2,2)

function $g(x)$ while the Figure 4.2 shows both true and estimated second component density f. Both figures show a histogram of the observed target distribution $g(x)$ in the background. Note that both the fitted mixture density $\hat{g}(x)$ and the fitted unknown component density function $\hat{f}(x)$ are quite close to corresponding true density functions.

In the forth experiment, we apply our algorithm on the multivariate cases. We first consider the multivariate kernel function of $K_h(\mathbf{x}) = K_h(|\mathbf{x}|)$, where $|\mathbf{x}| = (\sum_{j=1}^{d} x_j^2)^{\frac{1}{2}}$ is the L^2-norm. We generate n independent and identically distributed observations from a two component two-dimensional Gaussian-Gaussian mixture with the density $g(\mathbf{x})$ defined as $g(\mathbf{x}) = (1-p)f_0(\mathbf{x}) + pf(\mathbf{x})$. Thus, the known component is $f_0(x) = \phi_2(\mathbf{x}; \mu_\mathbf{0}, \mathbf{\Sigma_0})$ while the unknown component is $f(x) = \phi_2(\mathbf{x}; \mu, \mathbf{\Sigma})$, both of which are two-dimensional Gaussian distributions. We choose the sample size $n = 500$, the mixture proportion $p = 0.5$, $\mu_0 = \left(\begin{smallmatrix} 0 \\ 0 \end{smallmatrix}\right)$, $\Sigma_0 = \left(\begin{smallmatrix} 1 & 0 \\ 0 & 1 \end{smallmatrix}\right)$, $\mu = \left(\begin{smallmatrix} 2 \\ 2 \end{smallmatrix}\right)$ and $\Sigma = \left(\begin{smallmatrix} 1 & 0.5 \\ 0.5 & 1 \end{smallmatrix}\right)$. The initial weight is $p_0 = 0.4$ and the initial assumption for the unknown component

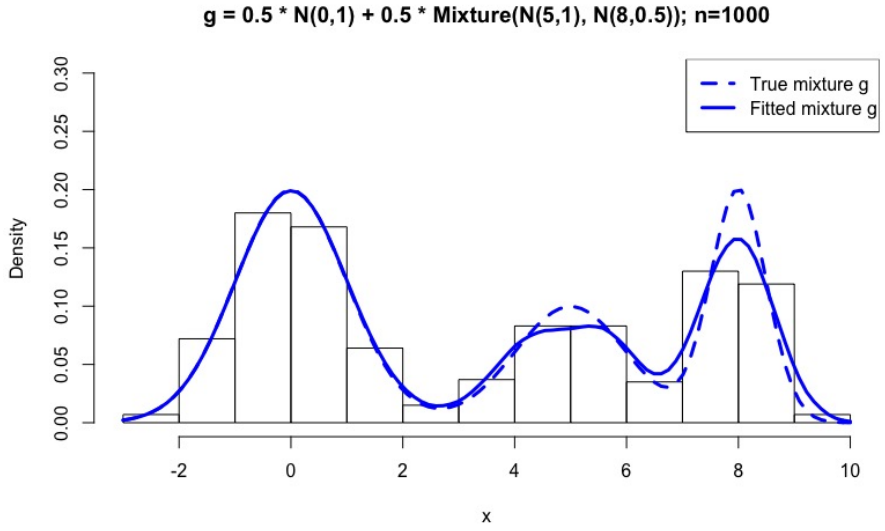

Figure 4.5. Fitted mixture density for a mixture of unimodal and bimodal distributions

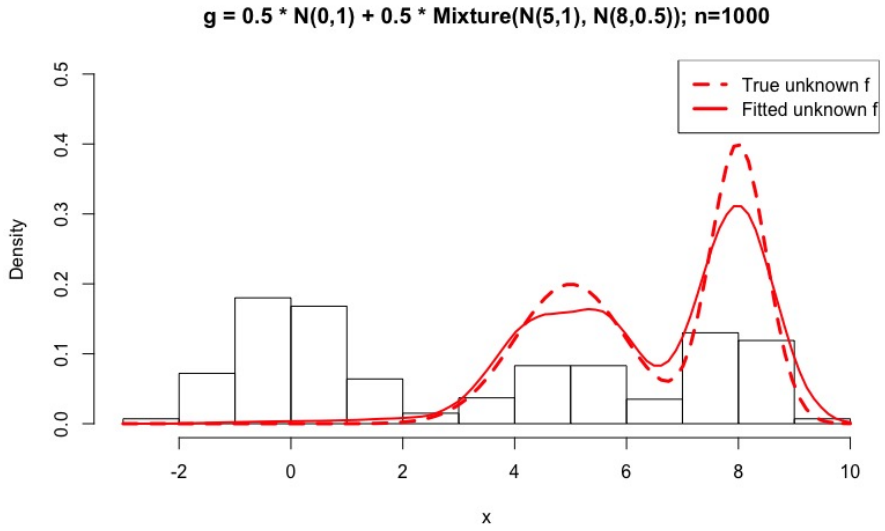

Figure 4.6. Fitted unknown component density for a mixture of unimodal and bimodal distributions

distribution is Gaussian with mean $\mu = \binom{3}{1}$ and $\Sigma = \left(\begin{smallmatrix} 0.5 & 0 \\ 0 & 2 \end{smallmatrix}\right)$. The rescaled Gaussian kernel $K_h(\mathbf{x}) = \frac{1}{h}\phi\left(|\mathbf{x}|/h; 0, 1\right)$ is used as the kernel function. The bandwidth we fixed

is $h = 0.85$. In the two-dimensional case, the computation converges after around 40 iterations, much faster than the one-dimensional cases. The estimate is $\hat{p} = 0.5164$, very close to the setting.

We also analyze performance of our algorithm in terms of the mean squared error (MSE) of estimated weight $\hat{p}$ and the mean integrated squared error (MISE) of $\hat{f}$. We will use two models for this purpose. The first model is the normal exponential model where the (known) normal component is the same as before while the second (unknown) component is an exponential density function $f(x) = \lambda e^{-\lambda x} I_{[x>0]}$ with $\lambda = 0.5$; the value of p used is $p = 0.6$. The second model is the same normal-gamma model as before. For each of the two models, we plot MSE of $\hat{p}$ and MISE of $\hat{f}$ against the true p for sample sizes $n = 500$ and $n = 1000$. Here, we use 30 replications. The algorithm appears to show rather good performance even for the sample size $n = 500$. Note that MISE of the unknown component f seems to decrease with the increase in p. Possible reason for this is the fact that, the larger p is, the more likely it is that we are sampling from the unknown component and so the number of observations that are actually generated by f grows; this seems to explain better precision in estimation of f when p is large.

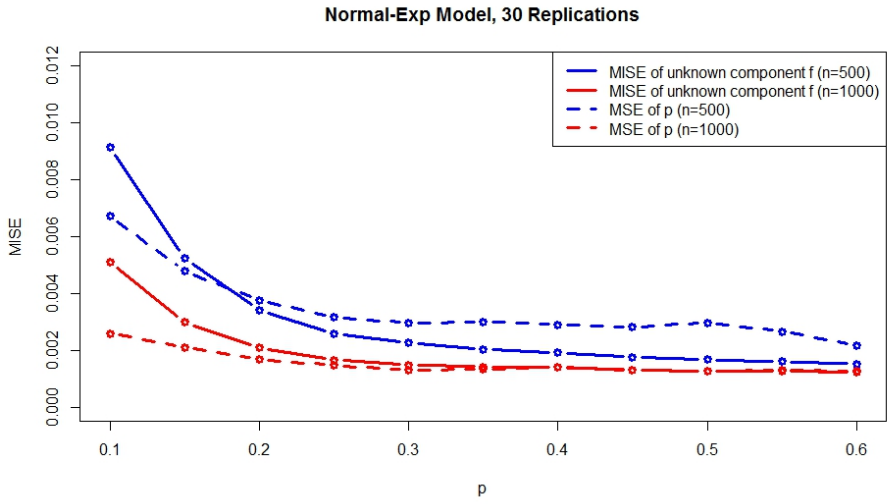

Figure 4.7. MISE of $\hat{f}$ and MSE of $\hat{p}$ in Normal-Exponential mixture model

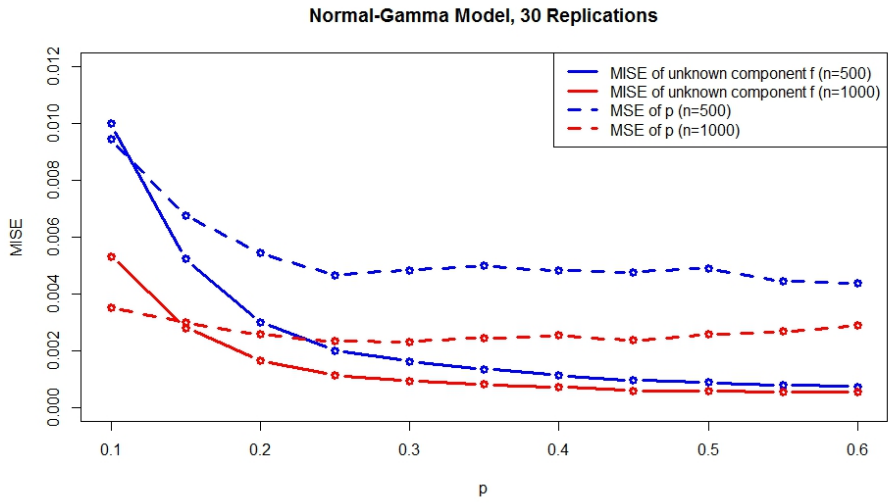

Figure 4.8. MISE of $\hat{f}$ and MSE of $\hat{p}$ in Normal-Exponential mixture model

4.3 Bandwidth Selection

Another important issue in practice is, of course, the bandwidth selection. Earlier, we simply used a fixed bandwidth selected using the classical Silverman's rule of thumb [32]. In general, when the unknown density is not likely to be normal, the use of Silverman's rule may be a somewhat rough approach. Moreover, in an iterative algorithm, every successive step of iteration brings a refined estimate of the unknown density component; therefore, it seems a good idea to put this knowledge to use. Such an idea was suggested earlier in [33].

Here we suggest using a version of the K-fold cross validation method specifically adopted for use in an iterative algorithm. First, let us suppose we have a sample $X_1, \ldots, X_n$ of size n; we begin with randomly partitioning it into K approximately equal subsamples. For ease of notation, we denote each of these subsamples X^k, $k = 1, \ldots, K$. Randomly selecting one of the K subsamples, it is possible to treat the remaining $K - 1$ subsamples as a training dataset and the selected subsample as the validation dataset. We also need to select a grid of possible bandwidths. To

do so, we compute the preliminary bandwidth h_s first according to the Silverman's rule of thumb; the bandwidth grid is defined as lying in an interval $[h_s - l, h_s + l]$ where $2 * l$ is the range of bandwidths we plan to consider. Within this interval, each element of the bandwidth grid is computed as $h_i = h_s \pm \frac{i}{M} l$, $i = 0, 1, \ldots, M$ for some positive integer M. At this point, we have to decide whether a fully iterative bandwidth selection procedure is necessary. It is worth noting that a fully iterative bandwidth selection algorithm leads to the situation where the bandwidth changes at each step of iteration. This, in turn, implies that the monotonicity property of our algorithm derived in Theorem 2.5.1 is no longer true. To reconcile these two paradigms, we implement the following scheme. As in earlier simulations, we use the triangular smoothing kernel. First, we iterate a certain number of times T to obtain a reasonably stable estimate of the unknown f; if we do it using the full range of the data, we denote the resulting estimate

$$\hat{f}_{nh}^{T}(x) = \frac{\alpha_n^T}{n} \sum_{i=1}^{n} K_h(x - X_i) w_n^{T-1}(X_i). \tag{4.4}$$

Integrating the resulting expression, we can obtain the squared L_2-norm of $\hat{f}_{nh}^{T}(x)$ as

$$||\hat{f}_{nh}^{T}||_2^2 = \int \left[\frac{\alpha_n^T}{n} \sum_{i=1}^{n} w_n^{T-1}(X_i) K_h(x - X_i) \right]^2 dx. \tag{4.5}$$

For each of K subsamples of the original sample, we can also define a "leave-kth subsample out" estimator of the unknown component f as $\hat{f}_{nh,-X_k}^{T}(x)$, $k = 1, \ldots, K$ obtained after T steps of iteration. At this point, we can define the CV optimization criterion as (see, for example, [19]) as

$$CV(h) = ||\hat{f}_{nh}^{T}||_2^2 - \frac{2}{n} \sum_{k=1}^{K} \sum_{x_i \in X_k} \hat{f}_{nh,-X_k}^{T}(x_i). \tag{4.6}$$

Finally, we select

$$h^* = argmin\, CV(h) \tag{4.7}$$

as a proper bandwidth. Now, we fix the bandwidth h^* and keep it constant beginning with the iteration step $T+1$ until the convergence criterion is achieved and the process

is stopped. An example of a cross validation curve of $CV(h)$ is given in Figure 4.9. Here, we took a sample of size 500 from a mixture model with a known component of $N(6,1)$, an unknown component of $Gamma(2,1)$ and a mixing proportion $p = 0.5$; we also chose $K = 50$, $l = 0.4$, $M = 10$, and $T = 5$. We tested the possibility of using larger number of iterations before selecting the optimal bandwidth h; however, already $T = 10$ results in the selection of h^* close to zero. We believe that the likeliest reason for that is the overfitting of the estimate of the unknown component f. The minimum of $CV(h)$ is achieved at around $h = 0.68$. Using this bandwidth and running the algorithm until the stopping criterion is satisfied, gives us the estimated mixing proportion $\hat{p} = 0.497$. As a side remark, in this particular case the Silverman's rule of thumb gives a very similar estimated bandwidth $\hat{h} = 0.72$.

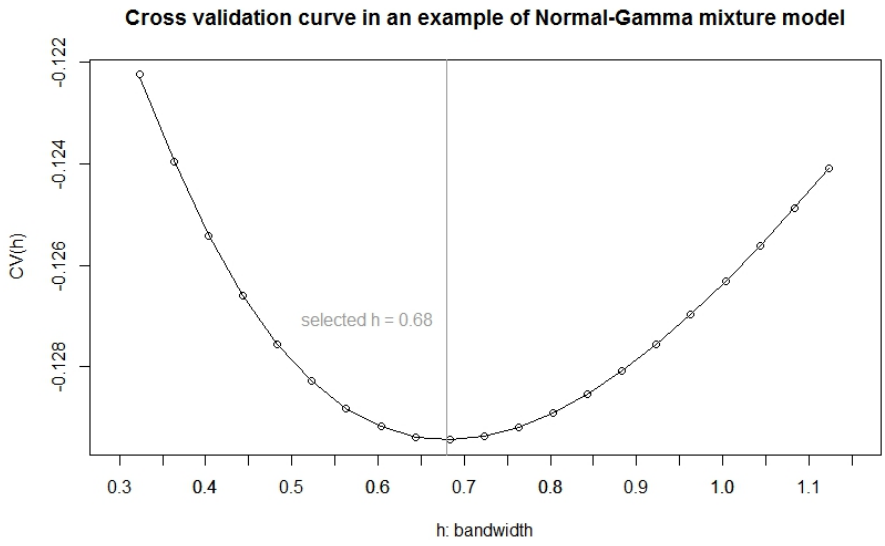

Figure 4.9. A plot of $CV(h)$ used for bandwidth selection

4.4 Comparison

As a last step, we want to compare our method with the symmetrization method of [13]. To do this, we will use a normal-normal model since the method of [13] is

only applicable when an unknown component belongs to a location family. Although such a model does not satisfy the sufficient criterion of the Lemma 2.2.1, it satisfies the necessary and sufficient identifiability criterion given in Lemma 4 of [16] (see also Remark 3 from the Supplement to [16] for even clearer statement about identifiability for normal-normal models in our context); therefore, we can use it for testing purposes. The known component has Gaussian distribution with mean 0 and standard deviation 1, the unknown has mean 6 and standard deviation 1, and we also consider two possible choices of mixture weight, $p = 0.3$ and $p = 0.5$. The results for two different sample sizes, $n = 500$, and $n = 1000$, and 200 replications, are given below in Tables 4.1 and 4.2. Each estimate is accompanied by its standard deviation in parentheses. Note that the proper expectation here is that our method should perform similarly to the method of [13] but not beat it, for several reasons. First, the mean of the unknown Gaussian distribution is directly estimated as a parameter in the symmetrization method, while it is the nonparametric probability density function that is directly estimated by our method. Thus, in order to calculate the mean of the second component, we have to take an extra step when using our method and employ numerical integration. This is effectively equivalent to estimating a functional of an unknown (and so estimated beforehand) density function; therefore, somewhat lower precision of our method when estimating the mean, compared to symmetrization method, where the mean is just a Euclidean parameter, should be expected. Second, when using symmetrization method, we followed an acceptance/rejection procedure exactly as in [13]. That procedure amounts to dropping certain "bad" samples whereas our method keeps all the samples. Third, the method of [13], when estimating an unknown component, uses the fact that this component belongs to a location family - something that our method, more general in its assumptions, does not do. Keeping all of the above in mind, we can see from Tables 4.1 and 4.2 that both methods produce comparable results, especially when the sample size is $n = 1000$. Also, as explained above, it does turn out that our method is practically as good as the method of [13] when it comes to estimating probability p and slightly worse when

Table 4.1.
Mean(SD) of estimated p/μ obtained by the symmetrization method

$K = 200$	$n = 500$	$n = 1000$
$p = 0.3/\mu = 6$	0.302(0.022)/5.989(0.095)	0.302(0.016)/5.998(0.064)
$p = 0.5/\mu = 6$	0.502(0.024)/5.999(0.067)	0.502(0.017)/6.003(0.050)

Table 4.2.
Mean(SD) of estimated p/μ obtained by our algorithm

$K = 200$	$n = 500$	$n = 1000$
$p = 0.3/\mu = 6$	0.315(0.024)/5.772(0.238)	0.312(0.018)/5.818(0.178)
$p = 0.5/\mu = 6$	0.516(0.026)/5.855(0.155)	0.512(0.018)/5.883(0.117)

estimating the mean of the unknown component. However, even when estimating the mean of the unknown component, increase in sample size from 500 to 1000 reduces the difference in performance substantially.

4.5 A real data example

The acidification of lakes in parts of North America and Europe is a serious concern. In 1983, the US Environmental Protection Agency (EPA) began the EPA National Surface Water Survey (NSWS) to study acidification as well as other characteristics of US lakes. The first stage of NSWS was the Eastern Lake Survey, focusing on particular regions in Midwestern and Eastern US. Variables measured include acid neutralizing capacity (ANC), pH, dissolved organic carbon, and concentrations of various chemicals such as iron and calcium. The sampled lakes were selected systematically from an ordered list of all lakes appearing on $1 : 250,000$ scale US Geological Survey topographic maps. Only surface lakes with the surface area of at least 4 hectares were chosen.

Out of all these variables, ANC is often the one of greatest interest. It describes the capability of the lake to neutralize acid; more specifically, low (negative) values of ANC can lead to a loss of biological resources. We use a dataset containing, among others, ANC data for a group of 155 lakes in north-central Wisconsin. This dataset has been first published in [34] in Table 1 and analyzed in the same manuscript. [34] argue that this dataset is rather heterogeneous due to the presence of lakes that are very different in their ANC within the same sample. In particular, seepage lakes, that have neither inlets nor outlets tend to be very low in ANC whereas drainage lakes that include flow paths into and out of the lake tend to be higher in ANC. Based on this heterogeneity, [34] suggested using an empirical mixture of two lognormal densities to fit this dataset. [35] also considered that same dataset; they suggested using a modification of Laplace method to estimate posterior component density functions in the Bayesian analysis of a finite lognormal mixture. Note that [35] viewed the number of components in the mixture model as a parameter to be estimated; their analysis suggests a mixture of either two or three components.

The sample histogram for the ANC dataset is given on Figure 1 of [35]. The histogram is given for a log transformation of the original data $log(ANC + 50)$. [35] selected this transformation to avoid numerical problems arising from maximization involving a truncation; the choice of 50 as an additive constant is explained in more detail in [35]. The empirical distribution is clearly bimodal; moreover, it exhibits a heavy upper tail. This is suggestive of a two-component mixture where the first component may be Gaussian while the other is defined on the positive half of the real line and has a heavy upper tail. We estimate a two-component density mixture model for this empirical distribution using two approaches. First, we follow the Bayesian approach of [35] using the prior settings of Table 4 in that manuscript. Switching to our framework next, we assume that the normal component is a known one while the other one is unknown. For the known normal component, we assume the mean $\mu_1 = 4.375$ and $\sigma_1 = 0.416$; these are the estimated values obtained in [35] under the assumption of two component Gaussian mixture for the original (not log transformed)

data and given in their Table 4. Next, we apply our algorithm in order to obtain an estimate of the mixture proportion and a non-parametric estimate of the unknown component to compare with respective estimates in [35]. We set the initial value of the mixture proportion as $p^0 = 0.3$ and the initial value of the unknown component as a normal distribution with mean $\mu_2^0 = 8$ and standard deviation $\sigma_2^0 = 1$. The iterations stop when $|p^{t+1} - p^t| < 10^{-4}$. After 171 iterations, the algorithm terminates with an estimate of mixture proportion $\hat{p} = 0.4875$; for comparison purposes, [35] produces an estimate $\hat{p}_{Bayesian} = 1 - 0.533 = 0.4667$. The Figure 4.10 shows the resulting density mixtures fitted using the method of [35] and our method against the background histogram of the log-transformed data. The Figure 4.11 illustrates the fitted first component of the mixture according to the method of [35] as well as the second component fitted according to both methods. Once again, the histogram of the log-transformed data is used in the background.

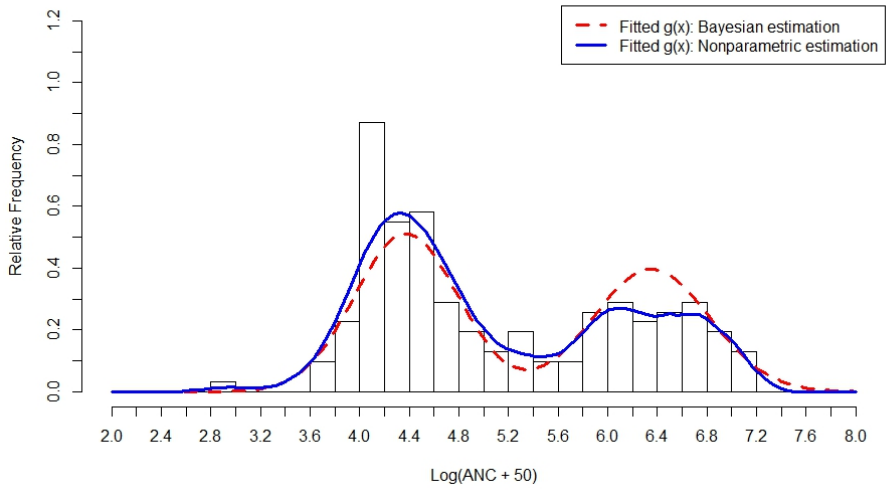

Figure 4.10. Fitted mixture densities

Note that the mixture density curves based on both methods are rather similar in Figure 4.10. One notable difference is that the method of [35] suggests mixture with a peak at the value of transformed ANC of about 6.4 whereas our method produces

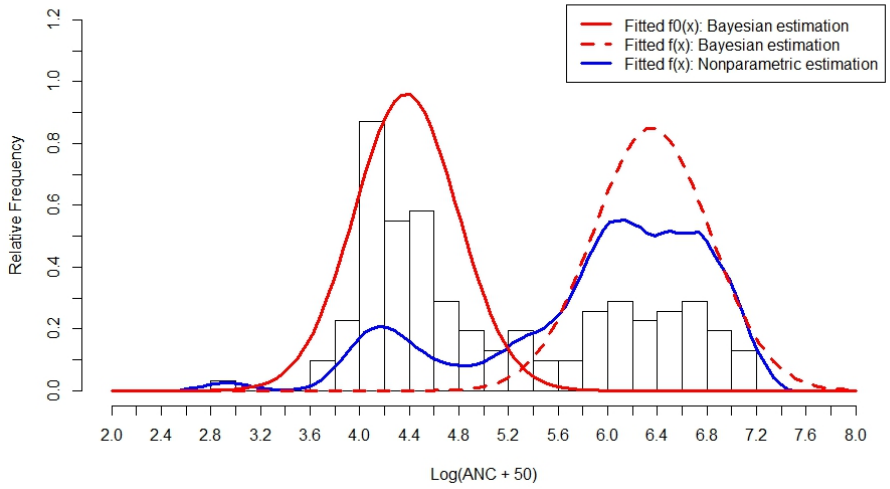

Figure 4.11. Fitted component densities

a curve that seems to be following the histogram more closely in that location. The Figure 4.11 also seems to show that our method describes the data more faithfully than that of [35]. Indeed, the second parametric component fitted by the method of [35] is unable to reproduce the first peak around 4.2 at all. By doing so, the method of [35] suggests that the first peak is there only due to the first component. Our method, on the contrary, suggests that the first peak is at least partly due to the second component as well. Note that [35] discusses the possibility of a three component mixture for this dataset; results of our analysis suggest a possible presence of the third component as well based on a bimodal pattern of our fitted second component density curve. Finally, note that the method of [35] produces an estimated second component that implies a much higher second peak than the data really suggests whereas our method gives a more realistic estimate.

5. DISCUSSION AND FUTURE WORK

In this chapter, we will discuss some issues to be solved in each of previous chapters.

5.1 Convergence to Stationary Points

In Conjecture 2.4.4, we states that our iterative MM algorithm converges to a set of stationary points. The proof stands on the GCT where the satisfaction of condition 3 is the key to the point. However, it can only be proved partially for now. Usually for stationary points of an arbitrary function $g(p, f)$ which consists of an integral of a Euclidean parameter p and a functional parameter f, there are two requirements. Firstly, the partial derivative of $g(p, f)$ with respect to p should be equal to 0; this has been done in proof of the conjecture. Secondly, the functional derivative of $g(p, f)$ with respect to f should be 0 as well, which may not be obvious to everyone.

In the calculus of variations, functionals are usually expressed in terms of an integral of functions and corresponding derivatives. Consider such an example,

$$J(f) = \int L(x, f(x), f'(x))\, dx. \tag{5.1}$$

If f is varied by adding to itself a tiny function δf, and the resulting integrand $L(x, f + \delta f, f' + \delta f')$ is expanded in powers of δf, then the change in the value of $J(f)$ to first order in δf can be expressed as

$$\delta J = \int \frac{\delta J}{\delta f(x)} \delta f(x)\, dx. \tag{5.2}$$

$\frac{\delta J}{\delta f(x)}$ is the functional derivative of $J(f)$ with respect to f at point x. According to Euler-Lagrange equation, if f is a stationary point of $J(f)$, then there must be

$$\frac{\delta J}{\delta f} = \frac{\partial L}{\partial f} - \frac{d}{dx}\frac{\partial L}{\partial f'} = 0. \tag{5.3}$$

However, in our functional $\ell(p, f)$, f appears as the convolutional and transformed version $\mathcal{N}_h f$. Thus it contains the global information of $f(x)$ on its domain. This property prohibits us to partition the interval of integration in $\ell(p, f)$, and therefore the properties of functional derivative in this case is unknown.

One way to conquer it is to discuss the functional derivative of $\ell(p, f)$ with respect to $\mathcal{N}_h f$ instead. Note that the functional derivative of $\mathcal{N}_h f$ with respect to f can also be derived. However, the chain rule for the functional derivatives is to question. Another way is to assume that the bandwidth h is close to 0 enough and approximate $\mathcal{N}_h f(x)$ by the following

$$\begin{aligned}
\mathcal{N}_h f(x) &= \exp\left\{\int K_h(x-u)\log f(u)\,du\right\} \\
&= \exp\left\{\frac{1}{h}\int K(\frac{x-u}{h})\log f(u)\,du\right\} \\
&= \exp\left\{\int K(u)\log f(x-hu)\,du\right\} \\
&\approx \exp\left\{\int K(u)\log\left\{f(x)-huf'(x)\right\}\,du\right\} \\
&\approx \exp\left\{\int K(u)\log\left\{f(x)\right\}\left\{-hu\frac{f'(x)}{f(x)}\right\}\,du\right\} \\
&= f(x)\exp\left\{-hR_K\frac{f'(x)}{f(x)}\right\},
\end{aligned} \tag{5.4}$$

where $R_K = \int_{\Delta/h} K(u)u\,du$ is a non-zero constant according to the selected kernel function $K(u)$. However, this approximation changes the nature of $\ell(p, f)$.

5.2 Convergence Rates

This is discussion follows the Section 3.3. The framework is similar to [29]. Convergence rate describes the relationship between the accuracy of solution and the deviation from the true mixture density to the noisy data. Note that the solution x_h of (3.10) with non-zero bandwidth h is an approximation to the true solution $\bar{x}$ of (3.9). Also, when we only have access to a finite sample $X_1, \ldots, X_n \sim g$, we can assume that a smooth estimate of the density g is available. Such an estimate can

be, for example, a kernel density estimate of g based on the kernel function of an appropriate smoothness. Let us denote this estimate of the function g by g_z. The difference between the estimated g_z and the true g will be described by the *data error* functional $D_g(g_z) : Z \to [0, \infty)$. The solution that corresponds to g_z (which is the minimizer of the Tikhonov-type functional $T_h(g_z, x)$) is denoted by $x_h^{g_z}$ while the solution corresponding to the original g is denoted x_h. It is important to keep in mind that this latter one is a solution of (3.10). Now, we can also introduce the *solution error* functional $E_{x_h} : X \to [0, \infty)$. We would like to obtain the bounds for the solution error $E_{x_h}(x_h^{g_z})$ with respect to the data error $D_g(g_z)$.

The exact data error $D_g(g_z)$ is often replaced by some upper bound in the convergence rate results, which is the so called *noise level* $\delta \in [0, \infty)$. All data elements bounded by a noise level δ constitute the set $Z_g^\delta := \{g_z \in Z : D_g(g_z) \leq \delta\}$. If we assume that $D_g(g) = 0$, then, since $g_z \in Z$ as well as g, we can assume that Z_g^δ is non-empty. In other words, one can always choose $g_z = g$ and obtain $D_g(g) = 0$ equivalently.

It is also necessary to assume a connection between the data error functional D_g and the fitting functional S to discuss convergence rate.

Assumption 4 *There exists a monotonically increasing function $\psi : [0, \infty) \to [0, \infty)$ satisfying $\lim_{\delta \to 0} \psi(\delta) \to 0$, $\psi(\delta) = 0$ if and only if $\delta = 0$, and*

$$S(g_z, g) \leq \psi(D_g(g_z)) \tag{5.5}$$

for all $g_z \in Z$ with $D_g(g_z) < \infty$. Therefore, for all solutions x of (3.10) and $g_{z^\delta} \in Z_g^\delta$, this assumption implies

$$S(g_{z^\delta}, F_h(x)) \leq \psi(D_g(g_{z^\delta})) \leq \psi(\delta) \tag{5.6}$$

The rationality of Assumption 4 can be justified by the following specific example.

Example 1 *Suppose the data error is quantified by the Kolmogorov Distance, i.e.* $D_g(g_z) = \sup_x |g_z(x) - g(x)|$. *In our case, we only consider densities on a compact set and bounded away from zero by some positive constant* η. *Then it is easy to see*

$$S(g_z, g) = \int_0^1 g_z(v)(\log g_z(v) - \log g(v))\, dv \leq \int_0^1 g_z(v)\frac{D_g(g_z)}{\eta}\, dv = \frac{D_g(g_z)}{\eta}. \quad (5.7)$$

Let $\psi(x) := \frac{x}{\eta}$ *and Assumption 4 is satisfied.*

Corollary 5.2.1 *Let* $\{\delta_k\}$ *be a sequence in* $[0, \infty)$ *converging to zero and* $\{g_{z_k}\}$ *be a sequence with* $g_{z_k} \in Z_g^{\delta_k}$. *Choose a sequence of bandwidths* $\{h_k\}$ *converging to zero. Further, let* $\{x_k\}$ *be a sequence in* X *with* $x_k \in argmin_{x\in X} T_{h_k}(g_{z_k}, x)$. *Then* $\{x_k\}$ *has a* τ_X*-convergent subsequence and each limit of a* τ_X*-convergent subsequence is a solution to* (3.9).

Proof By Assumption 4, $S(g_{z_k}, g) \leq \psi(\delta_k) \to 0$. Further, $F_h(x) \to F(x)$ as $h \to 0$ implies $S(g_{z_k}, F_{h_k}(\bar{x})) \to 0$ as $k \to \infty$. Then similar to proof of Theorem (3.3.3),

$$\begin{aligned} S(g_{z_k}, F_{h_k}(x_k)) &= T_{h_k}(g_{z_k}, F_{h_k}(x_k)) - \Omega(x_k, h_k) \quad (5.8)\\ &\leq T_{h_k}(g_{z_k}, F_{h_k}(\bar{x})) - \Omega(x_k, h_k) \\ &= S(g_{z_k}, F_{h_k}(\bar{x})) + \Omega(\bar{x}, h_k) - \Omega(x_k, h_k) \to 0. \end{aligned}$$

Again, let $\{x_{k_l}\}$ be an arbitrary subsequence of $\{x_k\}$ converging to some element $\tilde{x} \in X$. The lower semi-continuity of S and the continuity of F_h imply

$$S(g, F(\tilde{x})) \leq \liminf_{l\to\infty} S(g_{z_{k_l}}, F_{h_{k_l}}(x_{k_l})) = 0, \quad (5.9)$$

that is $S(g, F(\tilde{x})) = 0$. Thus $\tilde{x} \in X$ is a solution of (3.9).

■

We need the following definition and lemma to connect solution error and data error.

Definition 5.2.1 *We define the distance* $S_Y : Y \times Y \to [0, \infty]$ *as*

$$S_Y(y_1, y_2) = \inf_{y \in Y} \left(S(y, y_1) + S(y, y_2) \right). \tag{5.10}$$

Then we have the following triangle-type inequality

$$S_Y(y_1, y_2) \leq S(y, y_1) + S(y, y_2) \tag{5.11}$$

for all $y_1, y_2, y \in Y$.

Lemma 5.2.1 *Let* $\delta \in [0, \infty)$, $g_{z^\delta} \in Z_g^\delta$ *and* $x_h^\delta \in argmin_{x \in X} T_h(g_{z^\delta}, x)$. *For a monotonically increasing function* $\phi : [0, \infty) \to [0, \infty)$,

$$\begin{aligned} &\Omega(x_h^\delta, h) - \Omega(x_h, h) + \phi\left(S_Y(F_h(x_h^\delta), F_h(x_h))\right) \\ &\leq \psi(\delta) - S(g_{z^\delta}, F_h(x_h^\delta) + \phi\left(\psi(\delta) + S(g_{z^\delta}, F_h(x_h^\delta)\right) \end{aligned} \tag{5.12}$$

Proof The definition of x_h^δ, Assumption 4 and Definition 5.2.1 immediately imply

$$\begin{aligned} &\Omega(x_h^\delta, h) - \Omega(x_h, h) + \phi\left(S_Y(F_h(x_h^\delta), F_h(x_h))\right) \\ &= T_h(g_{z^\delta}, x_h^\delta) - \Omega(x_h, h) - S(g_{z^\delta}, F_h(x_h^\delta)) + \phi\left(S_Y(F_h(x_h^\delta), F_h(x_h))\right) \\ &\leq T_h(g_{z^\delta}, x_h) - \Omega(x_h, h) - S(g_{z^\delta}, F_h(x_h^\delta)) + \phi\left(S(g_{z^\delta}, F_h(x_h^\delta)) + S(g_{z^\delta}, F_h(x_h))\right) \\ &= S(g_{z^\delta}, F_h(x_h)) - S(g_{z^\delta}, F_h(x_h^\delta)) + \phi\left(S(g_{z^\delta}, F_h(x_h^\delta)) + S(g_{z^\delta}, F_h(x_h))\right) \\ &\leq \psi(\delta) - S(g_{z^\delta}, F_h(x_h^\delta)) + \phi\left(S(g_{z^\delta}, F_h(x_h^\delta)) + \psi(\delta)\right). \end{aligned} \tag{5.13}$$

■

The most important step is to build the connection between the solution error E_{x_h} and the Tikhonov-type functional. The following assumption is such a key.

Assumption 5 *Assume that there exist a constant* $\beta \in (0, \infty)$ *and a monotonically increasing function* ϕ *such that*

$$\beta E_{x_h}(x) \leq \Omega(x, h) - \Omega(x_h, h) + \phi\left(S_Y(F_h(x), F_h(x_h))\right) \tag{5.14}$$

5.14 is a type of variational inequalities. [29] gives a comprehensive review on this topic. In our work, finding the format of a variational inequality corresponding to our stabilizing functional is still open to question. In standard situations, Ω is convex and the usual choice of E_{x_h} is the Bregman distance with respect to some subgradient of Ω. However, our Ω_h is not convex, and so the Bregman distance with respect to it is out of the question (such a Bregman distance is not positive).

Assumption 6 *The function $\phi : [0, \infty)$ satisfies:*

1. *ϕ is monotoniically increasing, $\phi(0) = 0$, and $\phi(t) \to 0$ if $t \to 0$;*
2. *there exists a constant $\gamma > 0$ such that ϕ is concave, strictly monotonically increasing and $\phi(t) \leq t$ on $[0, \gamma]$;*
3. *the inequality*

$$\phi(t) \leq \phi(\gamma) + \left(\inf_{\tau \in [0,\gamma)} \frac{\phi(\gamma) - \phi(\tau)}{\gamma - \tau} \right) (t - \gamma) \tag{5.15}$$

is satisfied for all $t > \gamma$.

Definition 5.2.2 *Let $f : X \to (-\infty, \infty)$ be a functional on X which is finite at least at one point. For $\xi \in X^*$, the functional $f^* : X^* \to (-\infty, \infty)$ defined by*

$$f^*(\xi) := \sup_{x \in X} (< \xi, x > -f(x)) \tag{5.16}$$

is the conjugate function of f.

Lemma 5.2.2 *Let x_h satisfy Assumption 5. Then*

$$\beta E_{x_h}(x_h^\delta) \leq 2\psi(\delta) + (-\phi)^*(-1) \tag{5.17}$$

where $\alpha > 0$, $\delta \geq 0$

Proof By Lemma 5.2.1 and inequality (5.14), we have

$$\begin{aligned} \beta E_{x_h}(x_h^\delta) &\leq \Omega(x_h^\delta, h) - \Omega(x_h, h) + \phi\left(S_Y(F_h(x_h^\delta), F_h(x_h))\right) \\ &\leq \psi(\delta) - S(g_{z^\delta}, F_h(x_h^\delta) + \phi\left(\psi(\delta) + S(g_{z^\delta}, F_h(x_h^\delta)\right) \\ &= 2\psi(\delta) + \phi\left(\psi(\delta) + S(g_{z^\delta}, F_h(x_h^\delta))\right) - (\psi(\delta) + S(g_{z^\delta}, F_h(x_h^\delta))) \\ &= 2\psi(\delta) + \sup_{t \geq 0} (\phi(t) - t)\,. \end{aligned} \tag{5.18}$$

Then the results follows from the Definition 5.2.2

$$\sup_{t\geq 0}\left(\phi(t)-t\right)=\sup_{t\geq 0}\left(-t-(-\phi)(t)\right)=(-\phi)^{*}(-1). \tag{5.19}$$

■

Conjecture 5.2.2 *Let x_h satisfy Assumption 5, and find a $\delta\to k(\delta)$ satisfying*

$$\inf_{\tau\in[0,\psi(\delta))}\frac{\phi(\psi(\delta))-\phi(\tau)}{\psi(\delta)-\tau}\geq k(\delta)\geq\sup_{\tau\in(\psi(\delta),\gamma]}\frac{\phi(\tau)-\phi(\psi(\delta))}{\tau-\psi(\delta)} \tag{5.20}$$

for all $\delta>0$ with $\psi(\delta)<\gamma$. Then there is some C_δ such that

$$E_{x_h}(x_h^{\delta})\leq\frac{2}{\beta k(\delta)}\phi(\psi(\delta)) \tag{5.21}$$

for all $\delta\in(0,C_\delta]$.

Proof First, it is necessary to assert that for proposed δ and $k(\delta)$ in the theorem, $k(\delta)\leq 1$ and

$$\phi(\tau)-k(\delta)\tau\leq\phi(\psi(\delta))-k(\delta)\psi(\delta) \tag{5.22}$$

for all $\tau\geq 0$.

To see this, concave property in item 2 of Assumption 6 implies

$$1\geq\inf_{\tau\in[0,t)}\frac{\phi(t)-\phi(\tau)}{t-\tau}\geq\sup_{\tau\in(t,\gamma]}\frac{\phi(\tau)-\phi(t)}{\tau-t}>0. \tag{5.23}$$

for all $t\in(0,\gamma)$, which guarantees the existence of $k(\delta)\leq 1$. And for fixed $t\in(0,\gamma)$ and all $\tau>\gamma$, item 3 of Assumption 6 implies

$$\begin{aligned}\frac{\phi(\tau)-\phi(t)}{\tau-t}&\leq\frac{1}{\tau-t}\left(\phi(\gamma)+\left(\inf_{\epsilon\in[0,\gamma)}\frac{\phi(\gamma)-\phi(\epsilon)}{\gamma-\epsilon}\right)(\tau-\gamma)-\phi(t)\right)\\&\leq\frac{1}{\tau-t}\left(\phi(\gamma)+\frac{\phi(\gamma)-\phi(t)}{\gamma-t}(\tau-\gamma)-\phi(t)\right)=\frac{\phi(\gamma)-\phi(t)}{\gamma-t}.\end{aligned} \tag{5.24}$$

Set $t=\psi(\delta)$ and extend the supremum in the lower bound in (5.20) from $\tau\in(\psi(\delta),\gamma]$ to $\tau\in(\psi(\delta),\infty)$, namely

$$\inf_{\tau\in[0,\psi(\delta))}\frac{\phi(\psi(\delta))-\phi(\tau)}{\psi(\delta)-\tau}\geq k(\delta)\geq\sup_{\tau\in(\psi(\delta),\infty)}\frac{\phi(\tau)-\phi(\psi(\delta))}{\tau-\psi(\delta)}. \tag{5.25}$$

which is equivalent to $\frac{\phi(\psi(\delta))-\phi(\tau)}{\psi(\delta)-\tau} \geq k(\delta)$ for all $\tau \in [0, \psi(\delta))$ and $\frac{\phi(\tau)-\phi(\psi(\delta))}{\tau-\psi(\delta)} \leq k(\delta)$ for all $\tau \in (\psi(\delta), \infty)$. Then the assertion follows.

Then from Lemma 5.2.2, we obtain

$$\begin{aligned}
\beta E_{x_h}(x_h^\delta) &\leq 2\psi(\delta) + \sup_{t\geq 0}(\phi(t) - t) \leq 2\psi(\delta) + \sup_{t\geq 0}(\phi(t) - k(\delta)t) \qquad (5.26)\\
&\leq (2 - k(\delta))\psi(\delta) + \phi(\psi(\delta))\\
&\leq \frac{2 - k(\delta)}{k(\delta)}\psi(\delta) \inf_{\tau\in[0,\psi(\delta))} \frac{\phi(\psi(\delta)) - \phi(\tau)}{\psi(\delta) - \tau} + \phi(\psi(\delta))\\
&\leq \frac{2 - k(\delta)}{k(\delta)}\psi(\delta)\frac{\phi(\psi(\delta)) - \phi(0)}{\psi(\delta) - 0} + \phi(\psi(\delta))\\
&= \frac{2}{k(\delta)}\phi(\psi(\delta))
\end{aligned}$$

■

5.3 Efficiency of Algorithm

This discussion are inspired from the application of empirical version of our iterative MM algorithm. In the simulation, when the sample size it kept in thousands, each single iteration takes up to seconds. Since the algorithm usually converges around 100 or 200 iterations, this is acceptable. However, it will become very slow for larger sample.

The reason is the computation of convolution in our algorithm coming from the non-smoothing operator $\mathcal{N}_h$. The convolution is calculated over and over again in each iteration for all sample points and slows down the speed. One way to speed up the computation is to choose kernel functions on counting measure. Then the convolution by integration will be replaced by summation as an rough estimation.

The second way is to apply gradient descent algorithm. From the MM algorithm we derived, we know the iterative updating rule of $(p, f(x))$ as following:

$$\begin{aligned} p_n^{t+1} &= \int w_n^t(x) dG_n(x) = \frac{1}{n}\sum_{i=1}^{n} w_n^t(X_i) \\ f_n^{t+1}(x) &= \alpha_n^{t+1} \int K_h(x-u) w_n^t(u) dG_n(u) \\ &= \frac{\alpha_n^{t+1}}{n} \sum_{i=1}^{n} K_h(x-X_i) w_n^t(X_i) \end{aligned}$$

Since we get the format of the iterative estimator which is the weighted sum of kernel functions, it is reasonable to look for a family of estimations of (p, f), where

$$f(x) = \sum_{i=1}^{n} w_i K_h(x - X_i) \tag{5.27}$$

It may minimize the empirical functional of $(p, \mathbf{w})$

$$l_n(f, p) = l_n(\mathbf{w}, p) = -\sum_{i=1}^{n} \log\left((1-p) f_0(X_i) + p\mathcal{N}f(X_i)\right).$$

Now, the new task is to estimate $n+1$ parameters, i.e. p and w_i's for i from 1 to n, given a target function which would like to be minimized. Gradient descent algorithms can be applied here. The gradients w.r.t. parameters are

$$\frac{\partial l_n}{\partial p} = -\sum_{i=1}^{n} \frac{\mathcal{N}f(X_i) - f_0(X_i)}{(1-p) f_0(X_i) + p\mathcal{N}f(X_i)} \tag{5.28}$$

$$\frac{\partial l_n}{\partial w_j} = -\sum_{i=1}^{n} \frac{p\mathcal{N}f(X_i)}{(1-p) f_0(X_i) + p\mathcal{N}f(X_i)} \int K_h(X_i - u) \frac{K_h(u - X_j)}{f(u)} du \tag{5.29}$$

Suppose a learning rate of α. Therefore, the update rules will be

$$p_n^{t+1} = p_n^t - \alpha \frac{\partial l_n}{\partial p}(p^t, \mathbf{w_n}^t) \tag{5.30}$$

$$w_{n,j}^{t+1} = w_{n,j}^t - \alpha \frac{\partial l_n}{\partial w_j}(p^t, \mathbf{w_n}^t) \tag{5.31}$$

REFERENCES

REFERENCES

[1] Geoffrey McLachlan and David Peel. *Finite mixture models*. John Wiley & Sons, 2004.

[2] P. Hall and X. Zhou. Nonparametric estimation of component distributions in multivariate mixture. *The Annals of Statistics*, 31(1):201–224, 2003.

[3] P. Hall, A. Neeman, R. Pakyari, and R. Elmore. Nonparametirc inference in multivariate mixtures. *Biometrika Trust*, 92(3):667–678, 2005.

[4] Elizabeth S. Allman, Catherine Matias, and John A. Rhodes. Identifiability of parameters in latent structure models with many observed variables. *The Annals of Statistics*, 37(6(A)):3099–3132, 2009.

[5] Bradley Efron. *Large-scale inference: empirical Bayes methods for estimation, testing, and prediction*, volume 1. Cambridge University Press, Cambridge, United Kingdom, 2012.

[6] Stéphane Robin, Avner Bar-Hen, Jean-Jacques Daudin, and Laurent Pierre. A semi-parametric approach for mixture models: Application to local false discovery rate estimation. *Computational Statistics and Data Analysis*, 51(12):5483–5493, 2007.

[7] A Clifford Cohen. Estimation in mixtures of two normal distributions. *Technometrics*, 9(1):15–28, 1967.

[8] Bruce G Lindsay. The geometry of mixture likelihoods: a general theory. *The Annals of Statistics*, 11(1):86–94, 1983.

[9] Neil E Day. Estimating the components of a mixture of normal distributions. *Biometrika*, 56(3):463–474, 1969.

[10] Bruce G Lindsay and Prasanta Basak. Multivariate normal mixtures: a fast consistent method of moments. *Journal of the American Statistical Association*, 88(422):468–476, 1993.

[11] Jiashun Jin. Proportion of non-zero normal means: universal oracle equivalences and uniformly consistent estimators. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 70(3):461–493, 2008.

[12] T. T. Cai and J. Jin. Optimal rates of convergence for estimating the null density and proportion of nonnull effects in large-scale multiple testing. *The Annals of Statistics*, 38(1):100–145, 2010.

[13] L. Bordes, C. Delmas, and P. Vandekerkhove. Semiparametric estimation of a two-component mixture model where one component is known. *Scandinavian Journal of Statistics*, 33(4):733–752, 2006.

[14] Laurent Bordes and Pierre Vandekerkhove. Semiparametric two-component mixture model with a known component: an asymptotically normal estimator. *Mathematical Methods of Statistics*, 19(1):22–41, 2010.

[15] Yanyuan Ma and Weixin Yao. Flexible estimation of a semiparametric two-component mixture model with one parametric component. *Electronic Journal of Statistics*, 9(1):444–474, 2015.

[16] Rohit Kumar Patra and Bodhisattva Sen. Estimation of a two-component mixture model with applications to multiple testing. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 78(4):869–893, 2015.

[17] Zhou Shen, Michael Levine, and Zuofeng Shang. An mm algorithm for estimation of a two component semiparametric density mixture with a known component. *Electronic Journal of Statistics*, 12:1181–1209, 2018.

[18] P. P. B. Eggermont and V. N. LaRiccia. Maximum smoothed likelihood density estimation for inverse problems. *The Annals of Statistics*, 23(1):199–220, 1995.

[19] Paulus Petrus Bernardus Eggermont, Vincent N LaRiccia, and VN LaRiccia. *Maximum penalized likelihood estimation*, volume 1. Springer, New York, 2001.

[20] JM Ortega and WC Reinboldt. Iterative solution of nonlinear equations with multiple variables, 1970.

[21] David R Hunter and Kenneth Lange. A tutorial on mm algorithms. *The American Statistician*, 58(1):30–37, 2004.

[22] S. K. Bar-Lev and O. Stramer. Characterizations of natural exponential families with power variance functions by zero regression properties. *Probability Theory and Related Fields*, 76(4):509–522, 1987.

[23] Shaul K Bar-Lev and Peter Enis. Reproducibility and natural exponential families with power variance functions. *The Annals of Statistics*, 14(4):1507–1522, 1986.

[24] Harald Hanche-Olsen and Helge Holden. The kolmogorov-riesz compactness theorem. *Expositiones Mathematicae*, 28(4):385–394, 2010.

[25] C. F. Jeff Wu. On the convergence properties of the em algorithm. *The Annals of Statistics*, 11(1):95–103, 1983.

[26] Peter L. Bartlett and Shahar Mendelson. Empirical minimization. *Probability Theory Related Fields*, 135(3):311–334, 2006.

[27] S. I. Kabanikhin. Definitions and examples of inverse and ill-posed problems. *Journal of Inverse and Ill-posed Problems*, 16:317–357, 2008.

[28] Jens Flemming. Theory and examples of variational regularization with non-metric fitting functionals. *Journal of Inverse and Ill-Posed Problems*, 18(6):677–699, 2010.

[29] Jens Flemming. *Generalized Tikhonov regularization: basic theory and comprehensive results on convergence rates*. PhD thesis, 2011.

[30] Frank Werner and Thorsten Hohage. Convergence rates in expectation for tikhonov-type regularization of inverse problems with poisson data. *Inverse Problems*, 28(10):104004, 2012.

[31] Thorsten Hohage and Frank Werner. Inverse problems with poisson data: statistical regularization theory, applications and algorithms. *Inverse Problems*, 32(9):093001, 2016.

[32] Bernard W Silverman. *Density estimation for statistics and data analysis*, volume 26. CRC press, 1986.

[33] Didier Chauveau, David R Hunter, Michael Levine, et al. Semi-parametric estimation for conditional independence multivariate finite mixture models. *Statistics Surveys*, 9:1–31, 2015.

[34] Sybil L Crawford, Morris H DeGroot, Joseph B Kadane, and Mitchell J Small. Modeling lake-chemistry distributions: Approximate bayesian methods for estimating a finite-mixture model. *Technometrics*, 34(4):441–453, 1992.

[35] Sybil L Crawford. An application of the laplace method to finite mixture distributions. *Journal of the American Statistical Association*, 89(425):259–267, 1994.

VITA

Zhou Shen was born in 1990 in Hefei, Anhui Province, mid-east of China. He received a bachelors degree in Statistics at University of Science and Technology of China in 2013. Then he joined the Statistics Department at Purdue University and started his graduate study. He received a master degree in Statistics with specialization in Computational Finance in 2015 and earned a doctoral degree in Statistics in 2018. His research interests include nonparametric statistics, machine learning and mathematical finance, and his Ph.D. research focuses on nonparametric estimation of mixture models. During his graduate career, he served as a teaching assistant for various graduate and undergraduate level courses in Statistics Department. He also had internship experience at investment bankings including Credit Suisse and J.P.Morgan. He would like to pursue a professional career in quantitative analysis and data science in industry.