University of Alberta

Shape constrained density estimation

By

Mu Lin

A thesis submitted in partial fulfillment of the

requirements for the degree of

Doctor of Philosophy

In

Statistics

Department of Mathematical and Statistical Sciences

©Mu Lin

Spring, 2014

Edmonton, Alberta

Permission is hereby granted to the University of Alberta Libraries to reproduce single copies of this thesis and to lend or sell such copies for private, scholarly or scientific research purposes only. Where the thesis is converted to, or otherwise made available in digital form, the University of Alberta will advise potential users of the thesis of these terms.

The author reserves all other publication and other rights in association with the copyright in the thesis and, except as herein before provided, neither the thesis nor any substantial portion thereof may be printed or otherwise reproduced in any material form whatsoever without the author's prior written permission.

Abstract

We discuss non-parametric shape constrained density estimation methods in univariate setting and their applications to the classic Gaussian compound decision problem. The original contribution of the thesis is establishing various important consistency results of the shape constrained density estimators, which clarify the theoretical properties in the ρ-concave density estimation problem and the mixture density estimation in classical Gaussian compound decision problem.

Our main results begin with the consistency properties of ρ-concave density estimator in quasi-concave density estimation problem, proposed by Koenker and Mizera (2010). We consider a new type of divergence called ρ-divergence and prove the ρ-consistency for the corresponding ρ-concave density estimator when $\rho < 0$. We also generalize this consistency result to the consistencies under the Hellinger and the total variation distance.

Next, we consider the monotone constrained mixture density estimation problem in the classical Gaussian compound decision problem. We first obtain the Hellinger consistency of the mixture density estimator and further adopt the similar formulation of the convex transformed maximum likelihood density estimation method of Seregin and Wellner (2010) to prove the pointwise consistency of the estimated convex function and decision rule in the interior of the domain of the true convex function.

At last, we propose some new mixture density estimation approaches by imposing additional log-concave shape constraint on both the original monotone constrained maximum likelihood estimation and Kiefer-Wolfowitz maximum

likelihood mixing distribution estimation methods respectively. Finally, we perform a simulation study to compare the new methods with various existing ones in the empirical Bayes inference problems.

Acknowledgements

I would like to thank all the people who have supported and helped me during my PhD studies.

I would deeply thank my supervisor Dr. Ivan Mizera for all his invaluable advice, inspirational instructions and tremendous academic guidance during the past four and half years. He introduced me to the shape constrained density estimation problem and spent a lot of time to help me understand the recent research work in nonparametric density estimation field. He also helped me with the numerical experiment and some theoretical proofs in the thesis. I believe the research experience I have gained from him will have a great impact and influence on my future study and work.

My sincere thanks to other examining committee members: Dr. John Bowman, Dr. Rohana Karunamuni, Dr. Narasimha Prasad and Dr. Richard Samworth for their thorough reviews and insightful comments. A special thank to Dr Bowman for all his editorial corrections of this thesis.

I would also like to thank my wife, Weiyu Qiu, for all her support and encouragements when I was working on the thesis. I am so grateful for the endless support from my loved patents in my life and I feel so fortunate to learn from their great personalities and invaluable life experience.

Contents

List of Tables

List of Figures

Abbreviations and Symbols

$\mathbb{R}$	$(-\infty, +\infty)$
$\bar{\mathbb{R}}$	$(-\infty, +\infty]$
$\mathbb{R}_+$	$[0, +\infty)$
$\bar{\mathbb{R}}_+$	$[0, +\infty]$
$\xrightarrow{p}$	convergence in probability
$\xrightarrow{a.s.}$	convergence almost surely
$\text{lev}_\text{y}\text{g}$	$\{x \vert g(x) \leq y\}$
f'	the first derivative of $f(x)$
f''	the second derivative of $f(x)$
$f^{(k)}(x)$	the kth derivative of $f(x), k > 2$
$O_p(1)$	stochastic order symbols
$(\Omega, \mathcal{F}, P)$	probability triple
$\mathcal{B}$	Borel subset of $\mathbb{R}$
ϕ	standard normal density
$\mathbb{P}_n$	empirical measure
H	Hellinger distance with respect to the Lebesgue measure
H_Φ	Hellinger distance with respect to the standard noraml measure

iid	independent and identically distributed
a.s	almost surely
w.p.1	with porbability one
aff	affine hull
conv	convex hull
epi	epigraph
cl	closure
int	interior
ri	relative interior
$\lambda[s]$	Lebesgue measure of S
MLE	Maximum Likelihood estiamte
J-S	James and Stain
K-W	Kiefer and Wolfowitz
dim	dimension
EM	Expectation − maximization
dom	domain

Chapter 1

Estimation of probability densities

1.1 Introduction to density estimation

The probability density function is a fundamental concept in statistics. For a random variable X that has a probability density function f, the latter provides a natural description of the probabilistic law of the random variable X, via the simple relation

$$P(a < X < b) = \int_a^b f(x)dx.$$

Density estimation, as discussed in this thesis, addresses the problem to construct an estimate of the density function from a sequence of observed data points, sampled from an unknown probability density function.

A density estimate can provide many important properties of the given data set. For example, it can provide valuable information regarding the skewness and mutimodality of the probability distribution governing the data, and more importantly points the way to further statistical analyses. For example, density estimation often plays an important role in statistical analyses such as mixture problems, data discrimination, quantile regression, etc.

One approach to density estimation is parametric: it is assumed that the

data comes from a known parametric density family— for instance, the binomial distribution with single parameter p, or the normal distribution with the mean μ and the variance σ^2. The parametric method seeks first the estimates of the parameters p or μ and σ^2 from the data, and then substitutes the estimates into the expression for the corresponding density function. In many situations there is insufficient motivation for using a particular parametric model in statistical analysis. In this thesis, we consider alternative, nonparametric methods, which make less rigid assumptions than the parametric method about the distribution of the observed data. Although it is assumed that the distribution has a probability density f, the data will be allowed to "explain for themselves" in determining the estimate of f.

In the following sections, we will give a brief survey of the current existing methods in nonparametric density estimation. Despite the focus on the method like shape constrained density estimation in later chapters, it is very helpful to have a overall review of the well known methods and their basic properties. Moreover, we will introduce many general notations and definitions in nonparametric density estimation which will be used all through the thesis. For the remaining parts of this chapter, it is assumed that we have a sample of n independent and identically distributed (iid) observations $X_1, \ldots, X_n$ whose underlying density f is to be estimated and that $\hat{f}_n$ denotes the density estimator of the target density f.

1.2 Histogram

The histogram is probably the oldest and most widely used density estimator. Particularly in one dimension, histograms constitute an extremely useful and

simple class of density estimates for the presentation and exploration of the data. The method of histogram requires an origin x_0 and a bin width h; the bins of the histogram then can be defined to be the intervals $[x_0+mh, x_0+(m+1)h)$ for positive and negative integers m. (The intervals can also be chosen closed on the left and open on the right.)

The histogram estimate is defined to be,

$$\hat{f}_n(x) = \frac{1}{nh} \left(\text{number of } X_i \text{ in the same bin as } x\right).$$

Note that in order to construct the histogram, we need to choose both the origin and bin width, and the choice of bin width primarily controls the amount of the smoothing in the procedure.

There are always some users of density estimation who ask why it is ever necessary to use methods more complicated than the simple histogram. From the mathematical point of view, the discontinuity of histogram density estimates causes difficulties if derivatives of the estimates are required—for example, in the quantile regression and the Gaussian compound decision problem. Another mathematical drawback is the inefficient use of data when histograms are used as density estimates in procedures like the cluster analysis and the nonparametric discriminant analysis. In particular, when density estimates are used as intermediate components of other methods, the need for alternative methods is very strong.

The choice of the parameters, the origin and bin width, may also have very significant impact on the density estimates: such examples can be found in many texts on density estimation, for example, Silverman (1986). Histograms for graphical presentation of multivariate data pose also serious difficulties. In

the multidimensional situation, one cannot easily draw the histogram because of the dependence on the choice of the origin and the bin width in different coordinate directions.

1.3 The Kernel and other related density estimation methods

1.3.1 The method of kernel

Let $X_1, \ldots, X_n$ be a random sample of size n from a population with distribution function $F(x)$ and density function $f(x)$. Given that f is the derivative of F almost everywhere, it is natural to expect $[F(x+h) - F(x-h)]/2h$ to be close to $f(x)$ for sufficiently small h. Let

$$F_n(x) = \frac{\text{number of } X_i \leq x}{n}$$

be the empirical distribution function based on the observed sample. Letting $h_n \downarrow 0$, we consider

$$\hat{f}_n(x) = [F_n(x+h_n) - F_n(x-h_n)]/2h_n$$

to be an estimator of $f(x)$. Here $\hat{f}_n(x)$ denotes the proportion of observations falling in the interval $(x-h_n(x), x+h_n(x)]$ divided by the length of the interval.

By using

$$K(x) = \begin{cases} \frac{1}{2} & \text{if} \quad x \in [-1, 1), \\ 0 & \text{otherwise,} \end{cases}$$

the "naive" estimator can be written in the form

$$\hat{f}_n(x) = \frac{1}{n}\sum_{i=1}^{n}\frac{1}{h_n}K\left(\frac{x - X_i}{h_n}\right).$$

In order to produce this estimate, it still remains to choose the bin width h_n, to control the amount by which the data are smoothed. The naive estimator is not fully satisfactory since the density estimator is not a continuous function, but has jumps at the points $X_i \pm h_n$ and has zero derivative everywhere else—that is, it has somewhat ragged character.

It is easy to generalize the naive estimator to a class of estimators of the form

$$\hat{f}_n(x) = \frac{1}{n}\sum_{i=1}^{n}\frac{1}{h_n}K\left(\frac{x - X_i}{h_n}\right),$$

where $h_n \to 0$ as $n \to \infty$ and K is a suitable density function, which satisfies $\int_{-\infty}^{\infty} K(x)dx = 1$. This is called the kernel-type estimator. Usually, but not always, K will be a symmetric probability density function—for instance, the standard normal density. The parameter h_n is the window width, also called smoothing parameter, or bandwidth.

Some basic properties of kernel estimates follow from the definition. For example, since the kernel K is a probability density function, the estimate $\hat{f}_n$ is a probability density; $\hat{f}_n$ will inherit all the continuity and differentiability properties of the kernel function K.

The problem of choosing how much to smooth is of crucial importance in the kernel density estimation method. There are various methods of choosing smoothing parameters for the kernel method— for instance, least-squares cross-validation, likelihood cross-validation, test graph method, etc. It is important

to note that the shape constrained methods we introduce later in this chapter do not require any smoothing parameter selection.

It is also worth to mention some asymptotic properties of the kernel method. Since we will discuss the asymptotic behaviour of various shape constrained methods, we give now some basic consistency definitions of density estimators that will be used throughout the thesis.

Hereafter, we use $\xrightarrow{p}$ to denote convergence in probability. Let $d(x, y)$ be a distance function in $\mathbb{R}^d$—for instance, the Euclidean distance

$$d(x, y) = \|x - y\| = (\sum_{i=1}^{d} \big(x_i - y_i)^2\big)^{1/2}.$$

A sequence of random variables X_n is said to *converge in probability* to X if for every $\varepsilon > 0$

$$P(d(X_n, X) > \varepsilon) \to 0.$$

We denote it by $X_n \xrightarrow{p} X$.

An even stronger mode of convergence is almost sure convergence. The sequence X_n is said to *converge almost surely* to X if $d(X_n, X) \to 0$ with probability one:

$$P(\lim_{n\to\infty} d(X_n, X) = 0) = 1.$$

This is denoted by $X_n \xrightarrow{a.s.} X$. We will also use the notation w.p.1 (with probability one) interchangeably with almost surely (a.s) in the thesis.

Using the notations above, we can define the various types of consistency.

Definition 1.3.1. *A sequence of density estimators $\hat{f}_n$ is said to be weakly consistent if*

$$\hat{f}_n(x) \xrightarrow{p} f(x) \quad as \quad n \to \infty$$

for every x in the domain of f.

Definition 1.3.2. *A sequence of density estimators $\hat{f}_n$ is uniformly weakly consistent if*

$$\sup_x |\hat{f}_n(x) - f(x)| \xrightarrow{p} 0 \quad as \quad n \to \infty.$$

Definition 1.3.3. *A sequence of density estimators $\hat{f}_n$ is strongly consistent if*

$$\hat{f}_n(x) \xrightarrow{a.s.} f(x) \quad as \quad n \to \infty.$$

for every x in the domain of f.

Definition 1.3.4. *A sequence of density estimator $\hat{f}_n$ is uniformly strongly consistent if*

$$\sup_x |\hat{f}_n(x) - f(x)| \xrightarrow{a.s.} 0 \quad as \quad n \to \infty.$$

There is a vast literature on the asymptotic properties of kernel estimators. In this section we just mention a few asymptotic results; for more details, interested readers are referred to Prakasa Rao (1983) and Silverman (1986).

The consistency of the kernel estimator at a single point x was studied by Parzen (1962). He assumes that K should be a bounded Borel function, satisfying

$$\int |K(t)|dt < \infty \quad \text{and} \quad \int K(t)dt = 1 \tag{1.1}$$

and

$$|tK(t)| \to 0 \quad \text{as} \quad |t| \to \infty.$$

These conditions are satisfied by almost all kernels in general use. Assuming

that smoothing parameters h_n satisfy

$$h_n \to 0 \quad \text{and} \quad nh_n \to \infty \quad \text{as } n \to \infty,$$

Parzen (1962) proveed that $\hat{f}_n$ is weakly consistent.

A stronger consistency result is uniform weak consistency of the kernel estimator. Suppose the kernel is bounded and satisfies condition (1.1); assume that f is uniformly continuous on $(-\infty, \infty)$ and h_n satisfies

$$h_n \to 0 \quad \text{and} \quad nh_n(\log n)^{-1} \to \infty \quad \text{as } n \to \infty.$$

Under this assumption, Bertrand-Retali (1978) showed that kernel estimator $\hat{f}_n$ is uniformly weakly consistent.

There are also many results regarding the rates of the convergence of kernel estimators; for example, Prakasa Rao (1983) and Van der Vaart (1998), to name a few.

1.3.2 The nearest neighbour method

The nearest neighbour method focuses on adapting the amount of smoothness to the local density of data. The smoothness is controlled by an integer k which is much smaller than the sample size—typically by $k = n^{4/4+q}$ in q-dimension by minimizing the MSE of the density estimator. Following the usual definition of the distance, let $d(x, y) = |x - y|$, and for each t define the ordered distance from point t to the points of the sample data as

$$d_1(t) \leq d_2(t) \leq \ldots \leq d_n(t).$$

Then the *kth nearest neighbour density estimate* is defined by

$$\hat{f}_n(t) = \frac{k}{2nd_k(t)}. \tag{1.2}$$

A simple illustration of the meaning of (1.2) is: given a sample size n with density function $f(t)$, we would expect to observe $2rnf(t)$ observations to fall within the interval $[t-r, t+r]$ for $r > 0$. Therefore, there are k observations fall within the interval $[t - d_k(t), t + d_k(t)]$, and an estimate of the density can be obtained by the relationship $k = 2d_k n\hat{f}_n(t)$ which yields the definition above.

Like the naive estimator, to which it is related, the nearest neighbour estimate in equation (1.2) is not a smooth curve. While the distance function $d_k(t)$ is continuous, its derivative will have a discontinuity at every point of the form $\frac{1}{2}(X_{(j)} + X_{(j+k)})$, where $X_{(j)}$ are order statistics of the sample. Thus $\hat{f}_n$ is positive and continuous everywhere, but will have discontinuous derivative at all same points as d_k. Unlike kernel estimate, the nearest neighbour estimate will not be itself a probability density function, because it will not integrate to unity. To explain this point, consider t less than the smallest data point, we have $d_k = X_k - t$. For $t > X_{(n)}$, we have $d_k = t - X_{(n-k+1)}$. Substituting into the definition (1.2), it turns out that $\int_{-\infty}^{\infty} \hat{f}_n(t)$ is infinite. Therefore, the nearest neighbour estimate is unlikely to be an appropriate estimate if an estimate of entire density is required.

So it is possible to generalize the nearest neighbour estimate related to the kernel estimate. We can define *generalized kth nearest neighbour estimate* by

$$\hat{f}_n(t) = \frac{1}{nd_k(t)} \sum_{i=1}^{n} K\left(\frac{t - X_i}{d_k(t)}\right). \tag{1.3}$$

From (1.3), we can see that the estimate is the kernel estimate evaluated at t with window width $d_k(t)$. The ordinary nearest neighbour estimate is the special case of (1.3). However, the derivative of the generalized nearest neighbour estimate will be discontinuous at all the points where the function $d_k(t)$ has discontinuous derivative. A general form of the nearest neighbour estimate in multivariate dimension can be found in Silverman (1986).

An important extension of the nearest neighbour method combined with the kernel method is the adaptive kernel method. Adaptive kernel method allows to deal with the long-tailed densities by using a broader kernel in regions of low density. It has been also proposed that when estimating mixture parameters in finite mixture models, adaptive kernel density estimators are preferable over nonadaptive kernel density estimators. Regarding the adaptive kernel method, more detailed formulations and definitions can be found in Silverman (1986) and one can refer to Karunamuni et al. (2006a) and Karunamuni et al. (2006b) for the applications of adaptive kernel method to finite mixture models.

1.3.3 Orthogonal series estimators

The orthogonal series density estimation method is based on the Fourier expansion of the density function. To illustrate this point of view, we will start from an example which estimates a density f on the interval $[0, 1]$. The idea is to represent the density function f on the interval $[0, 1]$ based on its Fourier expansion, then we will only need to estimate its Fourier coefficients in order

to obtain the density estimate. Define the sequence $\phi_v(t)$ for $v = 1, 2, \ldots$ by

$$\begin{aligned} \phi_0(x) &= 1 \\ \phi_{2v-1}(x) &= \sqrt{2}\cos 2\pi vx \\ \phi_{2v}(x) &= \sqrt{2}\sin 2\pi vx. \end{aligned}$$

Then f can be expressed as the Fourier series

$$\sum_{v=0}^{\infty} f_v \phi_v,$$

where for each $v \geq 0$,

$$f_v = \int_0^1 f(x)\phi_v(x)dx. \tag{1.4}$$

Now suppose X is a random variable with density f, then (1.4) can be written as $f_v = E\phi_v(X)$. Thus a natural estimator of f_v based on the sample $X_1, \ldots, X_n$ from f is

$$\hat{f}_v = \frac{1}{n}\sum_{i=1}^{n} \phi_v(X_i).$$

Substituting f_v with $\hat{f}_v$ above yields $\sum_{v=0}^{\infty} \hat{f}_v\phi_v$ as an estimator of f; however Silverman (1986) showed that this estimator will converge to a sum of delta functions at the observations. The easy way to overcome this problem is to truncate the expansion at some point. Select an integer k and define the density estimator by

$$\hat{f}(x) = \sum_{v=0}^{k} \hat{f}_v\phi_v(x); \tag{1.5}$$

the choice of the cutoff point k then determines the amount of smoothness.

A more general approach is to use a sequence of weights to obtain the

estimator

$$\hat{f}(x) = \sum_{v=0}^{\infty} \lambda_v \hat{f}_v \phi_v(x)$$

where $\lambda_v \to 0$ as $v \to \infty$.

Without the restriction of the finite interval as the range of the data, we can use other orthogonal sequences of functions. Suppose $a(x)$ is a weighing function and (ψ_v) is a series satisfying for $u, v \geq 0$

$$\int_{-\infty}^{\infty} \psi_u(x)\psi_v(x)a(x)dx = \begin{cases} 1 & u = v \\ 0 & \text{otherwise.} \end{cases}$$

The sample coefficients are defined by

$$\hat{f}_v = \frac{1}{n} \sum_{i=1}^{n} \psi_v(X_i)a(X_i) \tag{1.6}$$

and the density estimator is

$$\hat{f}(x) = \sum_{v=0}^{k} \hat{f}_v \psi_v(x)$$

or

$$\hat{f}(x) = \sum_{v=0}^{\infty} \lambda_v \hat{f}_v \psi_v(x).$$

The properties of the orthogonal series estimates depend on the details of the Fourier series and weight functions used. For example, the density estimate will integrate to one given $\lambda_0 = 1$. Estimates obtained according to (1.5) will have derivatives of all orders. More details can be found in Silverman (1986).

1.4 Maximum penalized likelihood estimator

The maximum penalized likelihood estimator attempts to apply the idea of maximum likelihood to curve estimation. Let $X_1, X_2, \ldots, X_n$ be a sample of iid observations with common density f. The logarithm of the likelihood function of the sample is

$$L(f) = \sum_{i=1}^{n} \log(f(X_i)).$$

A naive application of the maximum likelihood method would make the estimator the mean of a set of the Dirac delta functions at the n observations and yields a value of $+\infty$ for the likelihood function. Therefore it is not possible to use the maximum likelihood directly without imposing any restrictions on the class of densities over which the likelihood is to be maximized. One approach suggested for using the method of maximum likelihood for density estimation is to add a penalty function to the likelihood, taking into account the degree of roughness or local variability of the density.

Let $\mathcal{F}$ be the class of density functions defined over $\mathbb{R}$. A penalty function $\Phi : \mathcal{F} \to \mathbb{R}$ is a real-valued functional defined over $\mathcal{F}$. The functional $\Psi(\cdot|\alpha) : f \to L(f) - \alpha\Phi(f)$, where $\alpha > 0$ is the smoothing parameter, is called the logarithm of penalized likelihood function. Any measurable function $\hat{f}_n : \mathbb{R} \to \mathcal{F}$ that maximizes $\Psi(\cdot|\alpha)$ over $\mathcal{F}$ is called *penalized maximum likelihood density estimator* of f.

The penalty approach can be explained as aiming at two often conflicting objectives in a single curve estimation problem: one is to maximize fidelity to the data, as measured by $L(f)$, while the other is to avoid curves which exhibit too much roughness or rapid variation, as measured by $\Phi(f)$. The choice of the smoothing parameter α controls the balance between smoothness

and goodness-of-fit.

One of the first penalty functions was suggested by Good and Gaskins (1971). They propose the penalty using $\gamma = \sqrt{f}$, the square root of the density function. The penalty is defined as

$$\Phi(f) = \int (\gamma')^2.$$

The advantage of working with γ instead of f is that the constraint $f(x) \geq 0$ is automatically satisfied if γ is real. Furthermore, the constraint $\int_{-\infty}^{\infty} f = 1$ is replaced by $\int \gamma^2 = 1$, an easier constraint under the numerical method proposed by Good and Gaskins (1971).

It is also convenient to use roughness penalties based on the logarithm of the density. Consider the penalty function that penalizes the third derivative of the log-density,

$$\Phi(f) = \int [(d/dx)^3 (\log f(x))]^2 dx. \tag{1.7}$$

Then the corresponding maximum likelihood problem can be expressed, by denoting $g(x) = \log f(x)$, as maximizing

$$\Psi(g) = \sum_{i=1}^{n} g(X_i) - \alpha \int (g^{(3)})^2, \tag{1.8}$$

subject to the constraint

$$\int e^{g(x)} dx = 1. \tag{1.9}$$

Notice that working with the logarithm of the density means there is no need to add the positivity constraint on f (since $f = e^g$ will be positive automatically).

The penalty (1.7) has the important property that it is zero if and only if f is a normal density. The normal densities are considered to be infinitely smooth since they are not penalized at all in equation (1.8). Silverman (1986) points out that when $\alpha \to \infty$, the limiting estimate is a normal density with the same mean and variance as the data. As α varies, the method provides different estimates from "infinitely rough" sum of the Dirac delta functions to the "infinitely smooth" maximum normal density. It is also possible to define other penalty functions, corresponding to other infinitely smooth exponential families of densities. The key property is that $\Phi(f)$ should be zero if and only if f is in the target family.

The remarkable form of maximum density estimation with penalty function was proposed by Silverman (1982), who showed that the maximum of (1.8) subject to the constraint (1.9) can be found as the *unconstrained* maximum of the strictly concave functional

$$\sum_{i=1}^{n} g(X_i) - \alpha \int (g^{(3)})^2 - n \int e^{g(x)} dx. \tag{1.10}$$

The form (1.10) is remarkable because it contains no unknown Lagrange multipliers to be determined; its maximum will automatically satisfy the constraint (1.9). The fact that the estimates can be found as the unconstrained maximum of a concave functional also makes it possible to derive many theoretical properties of the estimates— for example, see Silverman (1982). The unconstrained form also sheds light on the further density estimation problems that can be expressed in the similar concave (convex) forms—for instance, the shape constrained density estimation problem considered in the next chapters. Computationally, the concave problem (1.10) also makes the maximum penalized

likelihood method possible to calculate by using modern convex optimization methods (adding a negative sign on problem (1.10) makes the problem convex). The details of convex optimization numerical methods will be discussed in the section on shape constrained methods.

So far, we have discussed the L_2 penalty functions $\Phi(f)$. There are also the L_1 alternatives for the penalty functions (the L_1 penalized method is also called total variation method). In the L_1 framework, weighted sums of squared L_2 norms are replaced by weighted L_1 norms as an alternative penalized regularization device. Squaring penalty contributions inherently exaggerates the contribution to the penalty of jumps and sharp bends in the density. Indeed, density jumps and piecewise linear bends are impossible in the L_2 framework since the penalty evaluates them as "infinitely rough". Total variation penalties tolerate such jumps and bends, and they are therefore better suited to identifying discrete jumps in densities or in their derivatives, the property that has made them attractive in imaging applications.

Koenker and Mizera (2006) proposed to use roughness penalties $\Phi(f)$ based on total variation of the transformed density and its derivatives. Recall that the total variation of a real function f on Ω is defined as

$$\bigvee_{\Omega}(f) = \sup \sum_{i=1}^{m} |f(u_i) - f(u_{i-1})|,$$

where the supremum is taken over all partitions, $u_1, u_2, \ldots, u_m$ of Ω. When f is absolutely continuous, we can write

$$\bigvee_{\Omega}(f) = \int_{\Omega} |f'(x)| dx.$$

Usually, we focus on penalizing the total variation of the first derivative of the log-density, which leads to

$$\Phi(f) = \bigvee_{\Omega}((\log f)') = \int_{\Omega} |(\log f(t))''| dt.$$

Equivalently, setting $g = \log f$, the maximum L_1 penalized likelihood problem is maximizing:

$$\Psi(g) = \sum_{i=1}^{n} g(X_i) - \lambda \bigvee_{\Omega}(g'), \quad \text{subject to } \int_{\Omega} e^g = 1.$$

However, this is only one of many choices: one may think about using

$$\Phi(f) = \bigvee_{\Omega}(g^{(k)}),$$

where $g^{(0)} = g, g^{(1)} = g'$, etc, and g may be $\log f$, $\sqrt{f}$, f itself, or more generally $g^k = f$, for $k \in [1, \infty]$, with the convention that $g^\infty = e^g$. Furthermore, the linear combinations of such penalties with positive weights may also be considered.

We know that even for L_2 formulations the presence of the integrability constraint prevents the usual reproducing kernel strategy from finding exact solutions. Therefore iterative algorithms are needed. Koenker and Mizera (2006) suggested to adopt a finite element strategy that enables to exploit the sparse structure of the linear algebra used by modern interior-point algorithms for convex programming.

An advantage of the parametrization of the problem in terms of $\log f$ is that it obviates any worries about the non-negativity of f. However, we still need to ensure that our density estimates integrate to one. In the piecewise

linear model for $\log f$ this involves a awkward nonlinear constraint on the $\alpha's$,

$$\sum_{j=1}^{m} h_j \frac{e^{\alpha_i} - e^{\alpha_{i-1}}}{\alpha_i - \alpha_{i-1}} = 1.$$

This form of the constraint cannot be incorporated directly in its exact form into our optimization framework, nevertheless its approximation by a Riemann sum on a sufficient fine grid provides a numerically satisfactory solution.

1.5 Shape constrained density estimation

In nonparametric density estimation, it is sensible to place a priori restriction on the true density. For example, in the kernel method discussed in the preceding section, we assume the true density is smooth since the kernel method is a smoothing method. However, smoothness is not the only possible restriction. In this section we assume that the true density satisfies some shape constraints—for instance, we may assume the underlying density is monotone, unimodal, log-concave or even more general ρ-concave. The earliest work on the shape constraint density estimation dates back to the Grenander (1956) monotone density estimation which estimates a nonincreasing density in the positive half line. In recent years, the shape constrained density estimation has received a lot of interest—partly due to the development of the efficient computing algorithms which make the complicated mathematical optimization problems easier to solve.

The attractive virtue of the methods involving shape constraints is that, unlike the classical kernel and maximum penalized likelihood density estimation methods, the shape constrained methods are fully automatic: there are no

tuning (smoothing) parameters to choose. This is especially appealing in the multidimensional context, since there the choice of smoothing parameters can be extremely difficult. For instance, when the observations take values in $\mathbb{R}^d$, the general kernel estimator requires the specification of a symmetric, positive definite $d \times d$ bandwidth matrix.

1.5.1 Estimating a monotone density

We start with monotone densities and next view a unimodal density as a combination of two monotone pieces. Monotone density models are often used in survival analysis and reliability analysis in economics—see Huang and Wellner (1995), Huang and Zhang (1994). It is interesting that for the monotone density estimation, we can apply maximum likelihood as the estimating principle. Suppose that $X_1, \ldots, X_n$ is a random sample from a density f on $[0, \infty)$ that is known to be nonincreasing; the maximum likelihood estimator $\hat{f}_n$ then can be defined as the nonincreasing density that maximizes the likelihood

$$f \mapsto \prod_{i=1}^{n} f(X_i).$$

This optimization problem has a unique solution under the monotone assumption and was first proposed by Grenander (1956)—so the estimator is also called the *Grenander estimator*. This estimator is given explicitly by the left derivative of the *least concave majorant* of the empirical distribution function. The least concave majorant of the empirical distribution function F_n is defined as the smallest concave function $\hat{F}_n$ with $\hat{F}_n \geq F_n$ for every x. This can be found by attaching a rope at the origin $(0, 0)$ and winding this from above around the empirical distribution function F_n. Because $\hat{F}_n$ is concave,

its derivative is nonincreasing.

The limiting distribution of the Grenander estimator at a point was first obtained by Prakasa Rao in 1969; see Prakasa Rao (1983). Groeneboom (1988) provided a characterization of the limit distribution and other interesting related results. Van der Vaart (1998) pointed out that the rate of convergence of the monotone density estimator is slower than that of the kernel estimator when the existence at least two derivatives is assumed. The rate of convergence of maximum likelihood estimator, can be found in, for example, Van der Vaart (1998).

It is quite natural to generalize the monotone density to the unimodal density: we can define a density f on the real line to be *unimodal* if there exists a number M such that f is nondecreasing on the interval $(-\infty, M]$ and nonincreasing on $[M, \infty)$. The mode needs not to be unique. If we have a random sample from a unimodal density and we know the location of the true mode M a priori, then a natural extension of the monotone density estimation method is to estimate the distribution function F of the observation by the distribution function $\hat{F}_n$ that is the least concave majorant of F_n in the interval $[M, \infty)$ and the greatest convex minorant on $(-\infty, M]$, and then to estimate the density function f by taking the derivative of $\hat{F}_n$. Provided that none of the observations coincides with the mode M, the estimator maximizes the likelihood. The previous limiting results can also be applied to the unimodal case.

If the mode is not known as a priori, then the maximum likelihood estimator does not exist: the likelihood can be maximized to infinity by placing an arbitrary large mode at some fixed observation. It has been proposed to fix this problem by restricting the likelihood to densities that have a modal interval of a

given length (in which the f must be constant and maximal). Alternatively, we can also estimate the mode by some independent methods and then apply the procedure for a known mode as preceding discussions. However even if the mode is known, the estimator also suffers from the so called spiking problem—the inconsistency happening near the mode.

It seems that due to these problems, unimodality is not a reasonable assumption in shape-constrained density estimation. Thus there is strong need to find an alternative to the class of the unimodal densities. It turns out that log-concave densities are attractive and natural substitution for unimodal densities: the class of log-concave densities is a subset of unimodal densities, but it contains most of the common used parametric distributions and provides a rich and useful nonparametric model.

1.5.2 Log-concave density estimation

In this section, we provide a brief introduction to the log-concave density estimation method. A probability density function f is called *log-concave* if $-\log f$ is a convex function on the support (the smallest closed set whose complement has probability zero) of f.

Log-concave densities play a crucial role in various probability models: in reliability theory, search model, social model and economics, for instance, see An (1995), An (1998) and Bagnoli and Bergstrom (2005). Caplin and Nalebuff (1991) showed that in the theory of elections, under a log-concavity assumption the proposal that is most preferred by the mean voter is unbeatable under a 64% majority rule. Brooks (1998), Mengersen and Tweedie (1996) have developed the properties of the convergence of Markov chain Monte Carlo

sampling procedures on log-concave densities in Bayesian analysis.

Log-concave density functions have a number of properties that are desirable for modeling: the marginal distributions, convolutions and product measures of log-concave distributions are again log-concave. There are also some alternative characterizations for the class of univariate log-concave distributions, for example, Ibragimov (1956) proves that log-concave distributions are precisely the distributions whose convolution with a unimodal distribution is always unimodal, so sometimes log-concave distributions are also called strongly unimodal. Log-concave densities are also precisely the Polya frequency functions of order 2. Log-concave density family contains most of the common used parametric distributions—for instance, the normal density, gamma densities with shape parameter ≥ 1, exponential, Weibull densities with exponent ≥ 1, are all log-concave density functions.

Due to these attractiveness, recently the most intensively studied shape constraint is log-concavity— not only in univariate case, but also in multidimensional situations. In general, let $X_1, \ldots, X_n$ be an iid sample on $\mathbb{R}^d$ with log-concave density f_0 and denote the nonparametric maximum likelihood estimator as $\hat{f}_n$. For $d = 1$, the one dimensional case, the recent research showed that the nonparametric MLE exists and is unique; the log-concave density estimator $\hat{f}_n = e^{\hat{\phi}_n}$, where ϕ_n is continuous and piecewise linear on $[X_{(1)}, X_{(n)}]$ with the set of knots contained in $\{X_1, \ldots, X_n\}$, and $\hat{\phi}_n = -\infty$ on $\mathbb{R}\backslash[X_{(1)}, X_{(n)}]$; more properties of maximum likelihood log-concave density estimators in univariate case can be found in Pal et al. (2007) and Dümbgen and Rufibach (2009).

The consistency results of univariate log-concave density estimators were

developed by many authors: Pal et al. (2007) established the Hellinger consistency, an important type of consistency for log-concave density estimator in both univariate and multivariate settings. The Hellinger consistency can be defined as follows.

Definition 1.5.1. *For two arbitrary densities f and g on $\mathbb{R}^d$, the Hellinger distance between f and g is denoted by*

$$H(f,g) = \left(\frac{1}{2}\int(\sqrt{f(x)} - \sqrt{g(x)})^2 dx\right)^{1/2} = \left(1 - \int_{\mathbb{R}^d}\sqrt{f(x)g(x)}dx\right)^{1/2}.$$

If $H(\hat{f}_n, f_0) \to 0$ (or equivalently, $H^2(\hat{f}_n, f_0) \to 0$) almost surely for $n \to \infty$, then $\hat{f}_n$ is called Hellinger-consistent estimator of f_0.

Besides the Hellinger consistency, Dümbgen and Rufibach (2009) proved the uniform strong consistency on the compact subsets in the interior of the support of $\hat{f}_n$ and also provide the convergence rate of the density estimator.

Regarding the computation of log-concave density estimation in one dimension, Walther (2002), Pal et al. (2007) and Rufibach (2007) employed the ICMA (Iterated Convex Minorant Algorithm) to compute the log-concave density estimator. Rufibach (2007) also gave thorough descriptions of the algorithm and comparisons to the alternative algorithms like interior-point method; Dümbgen and Rufibach (2007) and Dümbgen and Rufibach (2011) computed the log-concave MLE with an active set algorithm. Both ICMA and active set algorithm can be found in R package "`logcondens`" (Rufibach and Dümbgen (2006)), which is accessible from "CRAN".

For $d > 1$, the structure of the multivariate maximum likelihood log-concave density estimators is similar to the univariate case. The support of the MLE is the convex hull of the data, and there is a triangulation of this convex hull

such that $\log \hat{f}_n$ is linear on each simplex of the triangulation. Nonparametric estimation of a log-concave density on $\mathbb{R}^d$ was initiated by Cule et al. (2010). These authors apply the powerful non-differential convex optimization method of the subgradient method of Shor (1985) and Shor's $\gamma-$algorithm implemented by Kappel and Kuntsevich (2000) to produce the estimator. An R version of the algorithm is available in the package "`LogConcDEAD`" (Cule et al. (2007)) and works in arbitrary dimension. Koenker and Mizera (2010) developed a family of penalized criterion functions related to Renyi divergence measures and explored duality in the optimization problems. Schuhmacher and Dümbgen (2010) proved the Hellinger consistency for the nonparametric maximum likelihood estimation of a log-concave probability density estimator on $\mathbb{R}^d$. Cule and Samworth (2010) established the consistency of the maximum likelihood estimator of a log-concave density on $\mathbb{R}^d$ even in a setting of model misclassification: when the true density is not log-concave, they show that the estimator converges to the log-concave density that is closest in the Kullback-Leibler sense to the true density.

1.5.3 ρ-concave density estimation

The log-concave densities discussed in the previous section have property that the tails must decrease exponentially: for example, densities with algebraic tails like t and F families are not log-concave. This restriction motivates a search for weaker forms of the concavity constraint that contains a richer family of densities. Koenker and Mizera (2010) and Seregin and Wellner (2010) both considered the generalization of the log-concave family called the ρ-concave densities defined as follows. For $a, b \in \mathbb{R}$, $\rho \in \mathbb{R}$ and $\lambda \in (0, 1)$, let the

generalized mean of order ρ, $M_\rho(a, b; \lambda)$ be defined for $a, b \geq 0$ as

$$M_\rho(a, b; \lambda) = \begin{cases} ((1-\lambda)a^\rho + \lambda b^\rho)^{1/\rho}, & \rho \neq 0, a, b > 0, \\ 0, & \rho < 0, ab = 0, \\ a^{1-\lambda} b^\lambda, & \rho = 0. \end{cases} \tag{1.11}$$

A density function f is then called $\rho-$concave on $\mathcal{C} \subset \mathbb{R}^d$ if and only if

$$f((1-\lambda)x + \lambda y) \geq M_\rho(f(x), f(y); \lambda) \quad \text{for all} \quad x, y \in \mathcal{C}, \lambda \in (0, 1).$$

In this terminology, log-concave functions are 0-concave, and concave functions are 1-concave. As $M_\rho(a, b; \lambda)$ is monotone increasing in ρ for $a, b \geq 0$ and $\lambda \in (0, 1)$, it follows that if f is ρ-concave, then f is also ρ'-concave for any $\rho' < \rho$. Therefore, concave functions are log-concave, but not vice-versa. The limiting $-\infty$-concave functions satisfy the condition

$$f((1-\lambda)x + \lambda y) \geq \min\{f(x), f(y)\},$$

which means that all ρ-concave functions are *quasi-concave*. The generalization of log-concave densities to ρ-concave was considered by many authors, starting with Avriel (1972); more details can be also found in Prékopa (1973), Borell (1975), and Dharmadhikari and Joag-Dev (1988).

Currently, there are two different approaches to the ρ-concave density estimation problem. Firstly, Koenker and Mizera (2010) considered the maximum likelihood estimation of log-concave density and then generalize the problem to ρ-concave density estimation. They also apply the conjugate dual formulation

of the primal estimation problem to estimate the ρ-concave densities which is not only numerically more efficient than the corresponding primal problem, but also conveys a maximum entropy interpretation.

Koenker and Mizera (2010) defined the general (primal) problem as,

$$\Phi(g) = \frac{1}{n}\sum_{i=1}^{n} g(X_i) + \int \psi(g)dx = \min_{g\in\mathcal{C}(X)} ! \quad \text{subject to } g \in \mathcal{K}(X), \tag{1.12}$$

where $\mathcal{K}(x)$ denote the cone of closed convex function on $\mathcal{H}(X)$, the convex hull of X. This cone is a subset of $\mathcal{C}(X)$, the collection of function continuous on $\mathcal{H}(X)$. Generally speaking, ψ is a nonincreasing convex function on $\mathbb{R}$.

The convex function g in problem (1.12) is not the density function we want to estimate. The relationship between g and the estimated density f can be derived through the dual formulation of (1.12), and is given by the formula $f = -\psi^{'}(g)$, according to Theorem 2 of Koenker and Mizera (2010). To interpret the problem (1.12) as $\rho-$concave density estimation, Koenker and Mizera proposed to use power function $\psi(x)$ with parameter $\alpha < 1$, where

$$\psi(x) = \begin{cases} +\infty & \text{for} \quad x \leq 0, \\ -x^{\beta}/\beta & \text{for} \quad x > 0. \end{cases} \tag{1.13}$$

and $1/\beta + 1/\alpha = 1$.

Given the power function $\psi(x)$ in (1.13), the estimated density function can be derived as $f = g^{\beta-1}$, or equivalently the convex function $g = f^{\alpha-1}$, based on $f = -\psi^{'}(g)$. In particular, for $\alpha < 1$, according to the definition of ρ-concave functions, f is $(\alpha - 1)$-concave density function—and this class is a significant relaxation of the class of the log-concave functions. In the next chapter we

focus on the proof of the consistency of the ρ-concave density estimator for $\rho = \alpha - 1 < 0$, given the power function defined in (1.13).

The crucial advantage of adopting the ρ-concave constraint is that the density estimation problem (1.12) is convex, thus it can be solved by the modern convex optimization tools as shown by Koenker and Mizera (2010). On the other hand, Koenker and Mizera (2010) also provided many theoretical properties of their approach to the ρ-concave density estimation problem: they established both primal and the dual formulations, derived the explicit relationship between them, proved the existence of the solution, the uniqueness of the solution and the Fisher consistency. We provide more detailed formulations of this approach when we discuss the new consistency result in next chapter.

The other way to solve the ρ-concave density estimation problem was proposed by Seregin and Wellner (2010). They denoted the class of all ρ-concave densities on $C \subset \mathbb{R}^d$ by $\hat{\mathcal{P}}(y_+^{1/\rho}; C)$ and write $\hat{\mathcal{P}}(y_+^{1/\rho})$ when $C = \mathbb{R}^d$. By Dharmadhikari and Joag-Dev (1988), for $\rho \leq 0$, it suffices to consider $\hat{\mathcal{P}}(y_+^{1/\rho})$, and $f \in \hat{\mathcal{P}}(y_+^{1/\rho})$ if and only if $f(x) = (g(x))^{1/\rho}$ for some convex function $g : \mathbb{R}^d \to [0, \infty)$. For $\rho > 0$, $f \in \hat{\mathcal{P}}(y_+^{1/\rho}; C)$ if and only if $f(x) = (g(x))^{1/\rho}$, where g mapping C into $(0, \infty)$ is concave.

Motivated by these results, Seregin and Wellner (2010) defined the classes $\mathcal{P}(y_+^{-s}) = \{f(x) = g(x)^{-s} : g \text{ is convex}\}$ for $s \geq 0$ and more generally, for a fixed monotone function h from $\mathbb{R}$ to $\mathbb{R}$,

$$\begin{aligned}\mathcal{P}(h) \equiv &\{h \circ g = h(g) : g \text{ convex and } h \circ g \text{ is a density with respect} \\ &\text{to Lebesgue measure}\}.\end{aligned}$$

Seregin and Wellner (2010) investigated the maximum likelihood estimation in

the class $\mathcal{P}(h)$ corresponding to a fixed monotone (decreasing or increasing) function h. In particular, for decreasing function h, they handle all of $\rho-$concave class $\mathcal{P}(y_+^{1/\rho})$ with $\rho = 1/s$ and $\rho \leq -1/d$ (or $s \geq d$). On the increasing side, they treat the case $h(y) = y1_{[0,\infty)}(y)$ and $h(y) = e^y$ with $C = \mathbb{R}_+^d$. On the decreasing side, this is also the ρ-concave density family considered by Koenker and Mizera (2010) by using the power functions defined in (1.13). Although the two papers use different parameters to denote the ρ-concave density family, they both achieve the solution of the same ρ-concave density estimation problem by setting parameters $s = 1 - \beta$. The equivalence of the ρ-concave density families in two papers can be summarized by the relationship of the parameters as follows

$$\rho = -1/s = 1/(\beta - 1) = \alpha - 1 \tag{1.14}$$

by using the notations from two papers.

For the increasing transformation h, the first situation $h(y) = y1_{[0,\infty)}(y)$ corresponds to an interesting class of models which can be thought as multivariate generalizations of the class of decreasing and convex densities studied by Groeneboom and Wellner (2001), while the second $h(y) = e^y$ corresponds to the multivariate versions of log-convex families studied by An (1998). In particular, the increasing classes $\mathcal{P}(y_+^{1/\rho})$ with $\rho > 0$ are actually ρ-convex densities which are quite different from ρ-concave classes defined in the problem (1.11).

Let $X_1, \ldots, X_n$ be independent random variables distributed according to the density $f_0 = h(g_0(x))$ on $\mathbb{R}^d$, where h is a fixed monotone (either increasing or decreasing) function and g_0 is an unknown convex function; the probability measure on Borel sets $\mathcal{B}_d$ corresponding to f_0 is denoted by P_0. Seregin and Wellner (2010) proposed to solve the problem in a straightforward way, seeking

the convex transformed density estimator by maximizing the likelihood. Let $\mathcal{C}$ denotes the class of all closed proper convex functions $g : \mathbb{R}^d \to (-\infty, \infty]$, the estimator $\hat{g}_n$ of g_0 is the maximizer of the functional

$$\mathbb{L}_n(g) \equiv \int (\log (h \circ g)) d\mathbb{P}_n \tag{1.15}$$

over the class of all convex functions g such that $h \circ g = h(g(x))$ is a density with respect to the Lebesgue measure and $\mathbb{P}_n$ is the empirical measure of the observations. Seregin and Wellner (2010) proved, under regularity conditions that, the maximum likelihood estimator of the convex transformed density $\hat{f}_n = h(\hat{g}_n)$ exists and is unique. Under some other natural assumptions, they also established the consistency in both Hellinger and uniform metrics and provide the asymptotic minimax lower bound and for estimation under curvature hypotheses.

In summary, both methods of Koenker and Mizera (2010) and Seregin and Wellner (2010) can be applied to estimate ρ-concave ($\rho < 0$) densities and are very similar by considering the density as some power function transformations: in Koenker and Mizera (2010), the ρ-concave density estimation problem corresponds to choosing $\psi(g)$ as the power function defined in (1.13), while in Seregin and Wellner (2010) it corresponds to the nonincreasing transformation power function $h(g)$. The natural question is what are the differences between the two ρ-concave density estimation problems. Generally speaking, the most important distinction between problem (1.12) and (1.15) is that the formulation (1.12) in Koenker and Mizera (2010) is convex, adding the integral term in problem (1.12) is a device that ensures that the solution integrates to one without enforcing this condition explicitly. Thus, the convex optimization

problem can be solved efficiently by the numerical algorithms proposed by Koenker and Mizera (2010). But adding the integral term "breaks" the traditional formulation of the maximum likelihood problem, which makes the proof of consistency results more challenging. So far, there was no successful proof of the Hellinger or uniform consistency of Koenker and Mizera's problem. In the next chapter, we provide a proof of a different type of consistency result for the ρ-concave density estimator based on Koenker and Mizera's method and also prove this particular consistency result implies the convergence with respect to Hellinger and total variation distance.

On the other hand, for the latter problem (1.15), Seregin and Wellner (2010) solved the convex transformed density estimation problem directly by maximizing the log-likelihood under the assumptions that the convex transformed function is a density. Seregin and Wellner (2010) successfully proved the Hellinger consistency for the convex transformed density estimators and pointwise convergence of the estimated convex functions, by applying many theorems from convex analysis and borrowing some techniques from the proof of consistency of the maximum likelihood log-concave density estimation problem. However, problem (1.15) is not convex, which makes the computation of the estimators more difficult than (1.12); Seregin and Wellner (2010) do not provide any algorithm to solve (1.15) numerically. For this reason, no numerical comparisons between two methods have been undertaken yet. Given all this, the method of Koenker and Mizera (2013) appears more preferable in practice, especially when density estimators are required in statistical procedures; the value of the method of Seregin and Wellner (2010) is in providing more theoretical insights on the $\rho-$concave density estimators.

1.6 Density estimation in the empirical Bayes inference

The purpose of this section is to introduce empirical Bayes method, James-Stein estimator, Gaussian compound decision problem and more importantly, point out the how the shape constrained density estimation methods can be applied to estimate the mixture density in Gaussian compound decision problem. The theoretical results regarding to the mixture density estimation will be discussed in Chapter 3 and 4. Most of the known results introduced in this section are based on Efron (2010) and Koenker and Mizera (2013).

1.6.1 J-S estimator and the empirical Bayes method

Charles Stein in 1955 proved that the common maximum likelihood methods for Gaussian models are inadmissible beyond simple one or two dimensional situations: the James-Stein estimator everywhere dominates the MLEs in higher dimensions. Efron (2010) discusses the empirical Bayes interpretation of the Stein estimator.

We will start from the introduction to the Bayes rule in normal means estimation. Suppose an observed data vector $z = (z_1, z_2, \ldots, z_n)$ has the density $f_\mu(z)$ given the unknown parameter vector $\mu = (\mu_1, \mu_2, \ldots, \mu_n)$ with prior density $g(\mu)$

$$z|\mu \sim f_\mu(z) \quad \text{and} \quad \mu \sim g(\cdot).$$

The Bayes rule then gives the posterior distribution of μ given the observed data z by

$$g(\mu|z) = g(\mu)f_\mu(z)/f(z), \tag{1.16}$$

where $f(z)$ is the marginal distribution of z,

$$f(z) = \int g(\mu) f_\mu(z) d\mu.$$

Now suppose that $\mu_i \sim N(0, a)$ and $z_i|\mu_i \sim N(\mu_i, 1)$ for $i = 1, 2, \ldots, n$, with pairs (z_i, μ_i) being independent. Given the Bayes rule above we denote in n-dimensions

$$\mu \sim N_n(0, aI) \tag{1.17}$$

and

$$z|\mu \sim N_n(\mu, I) \tag{1.18}$$

where I is $n \times n$ identity matrix. With $b = a/(a+1)$, the posterior distribution is

$$\mu|z \sim N_n(bz, bI). \tag{1.19}$$

From the statistical inference point of view, we denote the estimator $\hat{\mu} = t(z)$, and use the total squared error loss to measure the error between $\hat{\mu}$ and μ by

$$L(\mu, \hat{\mu}) = \| \mu - \hat{\mu} \|^2 \tag{1.20}$$

with the corresponding risk function

$$R(\mu) = E_\mu\{L(\mu, \hat{\mu})\} = E_\mu\{\| \mu - \hat{\mu} \|^2\}, \tag{1.21}$$

where E_μ is the expectation with respect to $z \sim N_n(\mu, I)$ for μ fixed.

The maximum likelihood estimator of μ in model (1.18) is the data itself,

$$\hat{\mu}^{(\mathrm{MLE})} = z,$$

and has the risk

$$R^{(\mathrm{MLE})}(\mu) = n$$

for every choice of μ.

Suppose we have the prior belief (1.17), then the corresponding Bayes estimator based on relationship (1.19) is

$$\hat{\mu}^{(\mathrm{Bayes})} = bz = \left(1 - \frac{1}{a+1}\right) z \tag{1.22}$$

which minimizes the expected squared error given z.

Assume the model is right but if the value a is unknown, so we can not use the Bayes estimator $\hat{\mu}^{(\mathrm{Bayes})}$ in equation (1.22). However, the empirical Bayes estimation method will work in this situation. The assumptions (1.17) and (1.18) imply the marginal distribution of z is

$$z \sim N_n(0, (a+1)I).$$

The sum of squares $S = \|Z\|^2$ follows a chi-square distribution with n degrees of freedom, $S \sim (a+1)\chi^2_n$, so that $E(\frac{n-2}{S}) = \frac{1}{a+1}$. The *James-Stein estimator* is then defined as

$$\hat{\mu}^{(\text{J-S})} = \left(1 - \frac{n-2}{S}\right) Z. \tag{1.23}$$

The James-Stein estimator just replaces the unknown term $1/(a+1)$ in $\hat{\mu}^{(\mathrm{Bayes})}$ in equation (1.22) by the unbiased estimator $(n-2)/S$. Therefore $\hat{\mu}^{(\text{J-S})}$ is

called "empirical Bayes" estimator since the Bayes estimator $\hat{\mu}^{(\text{Bayes})}$ is being empirically estimated from the data. This is feasible because that we have n similar problems, $z_i \sim N(\mu_i, 1)$ for $i = 1, 2, \ldots n$. The James-Stein theorem from James and Stein (1961) below demonstrates the statement at the beginning of the section.

Theorem 1.6.1. *For $n \geq 3$ the James-Stein estimator everywhere dominates the MLE in terms of expected total squared error; that is,*

$$E_\mu\{\|\hat{\mu}^{(J\text{-}S)} - \mu\|^2\} < E_\mu\{\|\hat{\mu}^{(MLE)} - \mu\|^2\} \tag{1.24}$$

for every choice of μ.

Theorem 1.6.1 states that for $n \geq 3$, $\hat{\mu}^{(\text{J-S})}$ dominates $\hat{\mu}^{(\text{MLE})}$ no matter what is the prior of μ. But we should notice that the James-Stein Theorem only concentrates on the total squared loss function, without concerning for the effects on the individual cases. Thus under some circumstances, we still use the MLE rather than empirical Bayes method, for instance, the linear regression problems. More examples can be found in Efron (2010).

1.6.2 Gaussian compound decision problem

The James-Stein estimator can be viewed as a special case of the empirical Bayes procedure for the Gaussian compound decision problem. The compound decision problem concerns the estimation of a vector with iid normal errors under the average squared loss. The problem has been considered as the canonical model or motivating example in the developments of empirical Bayes, admissibility, adaptive nonparametric regression, variable selection and many

other areas in statistics. It also plays a significant role in statistical applications since the observed data are often represented or summarized as the sum of a signal vector and the white noise.

We suppose that the vector $\mu = (\mu_1, \mu_2, \ldots, \mu_n)$ is sampled from

$$\mu_i \sim F. \tag{1.25}$$

Given the parameters μ_i, the observations $(X_1, X_2, \ldots, X_n)$ follow the normal distribution with known variance σ^2 as

$$X_i \mid \mu_i \sim N(\mu_i, \sigma^2). \tag{1.26}$$

The observations X_i then have the mixture (marginal) density,

$$g(x) = \int \phi_\sigma(x - \mu) dF(\mu). \tag{1.27}$$

where $\phi_\sigma(x) = (2\pi\sigma^2)^{-1/2} \exp\{-x^2/2\sigma^2\}$.

The *Tweedie's formula* calculates the posterior expectation of μ given x, which is also known as the Bayes rule (denoted by $\delta(x)$)

$$E(\mu|X = x) = \delta(x) = x + \sigma^2 l'(x). \tag{1.28}$$

where $l(x) = \log(g(x))$.

The attractive advantage of Tweedie's formula is that it works directly with the mixture density g without estimating the mixing distribution F, thus avoiding the difficult deconvolution procedure of F. The formulas above hold for any one parameter exponential family and can be verified by considering

the more general exponential generalization of (1.25) and (1.26),

$$\eta \sim f(\cdot) \quad \text{and} \quad x \mid \eta \sim f_\eta(x) = e^{\eta x - \psi(\eta)} f_0(x) \tag{1.29}$$

where f is the prior density with distribution function F, η is the natural or canonical parameter of the family, $\psi(\eta)$ is the cumulant generating function or cgf (which assures $f_\eta(x)$ integrate to 1), and $f_0(x)$ is the density when $\eta = 0$. The choice $f_0(x) = \phi_\sigma(x)$ in (1.27) is a $N(0, \sigma^2)$ density, yields the normal family $N(\mu, \sigma^2)$ with $\eta = \mu/\sigma^2$ and $\psi(\eta) = \frac{1}{2}\sigma^2\eta^2$.

The Bayes rule then gives the posterior density,

$$f(\eta \mid x) = f_\eta(x) f(\eta)/g(x),$$

where $g(x)$ is the marginal density

$$g(x) = \int f_\eta(x) f(\eta) d\eta.$$

The generalized model (1.29) yields

$$f(\eta \mid x) = e^{x\eta - \lambda(x)}[f(\eta)e^{-\psi(\eta)}] \quad \text{where} \quad \lambda(x) = \log\left(\frac{g(x)}{f_0(x)}\right). \tag{1.30}$$

Equation (1.30) denotes the exponential family with canonical parameter x and cgf $\lambda(x)$. Therefore, differentiating $\lambda(x)$ provides the posterior cumulant of η given x, which yields

$$E\{\eta \mid x\} = \lambda'(x), \quad \text{Var}\{\eta \mid x\} = \lambda''(x). \tag{1.31}$$

It worth to note that (1.31) implies that $E\{\eta \mid x\}$ is an increasing function of x; see Van Houwelingen and Stijnen (1983).

With $l(x) = \log(g(x))$ and $l_0(x) = \log(f_0(x))$, the posterior mean and variance can be expressed as

$$\eta \mid x \sim (l'(x) - l_0'(x), l''(x) - l_0''(x)). \tag{1.32}$$

In the normal translation family case, (1.32) implies

$$\mu \mid x \sim (x + \sigma^2 l'(x), \sigma^2(1 + \sigma^2 l''(x))), \tag{1.33}$$

which illustrates the formula (1.28).

It is also worth to note that if the mixture density $g(x)$ is log-concave, that is $l''(x) \leq 0$, then $\mathrm{Var}(\eta|x)$ is less than σ^2; The log-concavity of mixing density $f(\mu)$ in (1.25) will guarantee the log-concavity of mixture density $g(x)$, see Marshall and Olkin (2007), and this property will be used in Chapter 4 where we investigate the new density estimation methods for the Gaussian compound problem.

We can see that the empirical Bayes formula $\hat{\mu}_i = X_i + \sigma^2 \hat{l}'(X_i)$ requires the estimation of the mixture density $g(x)$ in the Gaussian compound decision problem. Efron (2011) proposes the *Lindsey's method*, a Poisson regression method described in Efron (2008) and Efron (2010) to accomplish this task. On the other hand, we will apply the shape constrained density estimation method to estimate the mixture density g for this problem. In particular, in Chapter 3, we will utilize the increasing property of (1.31) as a constraint in the maximum likelihood estimation of the mixture density and discuss the

consistency property of the density estimator. In Chapter 4, we investigate the new methods to estimate the mixture density by imposing additional shape constraints based on the monotone constrained method discussed in Chapter 3.

1.7 Prerequisites

In this section, we provide some basic definitions from measure theory and convex analysis. The references we use are Rosenthal (2006) and Rockafellar (1970) . These definitions will be used all through Chapter 2 to Chapter 4, since the measure theory and the properties of convex function and convex sets will play important roles in the remaining chapters.

First, we use the notation $(\Omega, \mathcal{F}, P)$ to denote a probability triple, where Ω is the sample space, $\mathcal{F}$ is the σ-algebra, and P is the probability measure. We will also use $\mathcal{B}$ for the Borel σ-algebra of the subsets of $\mathbb{R}$, defined as the smallest $\sigma-$algebra that includes all the intervals.

The distribution or law of a random variable is defined in the usual way:

Definition 1.7.1. *Given a random variable X on a probability triple $(\Omega, \mathcal{F}, P)$, its distribution (or law) is the function μ defined on $\mathcal{B}$, the Borel subsets of $\mathbb{R}$, by*

$$\mu(B) = P(X \in B) = P(x^{-1}(B)), B \in \mathcal{B}.$$

If μ is the law of a random variable, then $(\mathbb{R}, \mathcal{B}, \mu)$ is also a valid probability triple. We will write μ as $\mathcal{L}(X)$ or as PX^{-1}. We will also write $X \sim \mu$ to indicate that μ is the distribution of X. We define the cumulative distribution function of a random variable X by $F_X(x) = P(X \leq x)$, for $x \in \mathbb{R}$.

Given any Borel-measurable function (called a density function) f such

that $f \geq 0$ and $\int_{-\infty}^{\infty} f(t)\lambda(dt) = 1$, we can define a law μ by

$$\mu(B) = \int_{-\infty}^{\infty} f(t)\mathbf{1}_B(t)\lambda(dt), \text{ for every Borel set } B.$$

We will sometimes write this as $\mu(B) = \int_B f(t)\lambda(dt)$, or even as $\mu(dt) = f(t)\lambda(dt)$.

For the example of the distribution of a random variable, consider X with the normal $(0, 1)$ distribution (usually denoted as $N(0, 1)$). We can define the law μ_N by

$$\mu_N(B) = \int_{-\infty}^{\infty} \phi(t)\mathbf{1}_B\lambda(dt), \text{ for every Borel set } B,$$

where λ is the Lebesgue measure on $\mathbb{R}$ and $\phi(t) = \frac{1}{\sqrt{2\pi}}e^{-t^2/2}$.

Since we consider the density on the real line $\mathbb{R}$, by choosing the Borel set B above as interval $(-\infty, x]$, the definition of distribution and cumulative distribution are identical. Using the usual notation, Φ, for the cumulative distribution function of $N(0, 1)$, we have

$$\Phi(x) = \int \phi(t)\mathbf{1}_{(-\infty,x]}\lambda(dt) = \int_{-\infty}^{x} \phi(t)dt.$$

The next Lemma is the Radon-Nikodym Theorem.

Lemma 1.7.2. *A Borel probability measure μ is absolutely continuous (i.e. there is f with $\mu(A) = \int_A f d\lambda$ for all Borel A) if and only if it is dominated by λ (i.e. $\mu \ll \lambda$, i.e. $\mu(A) = 0$ whenever $\lambda(A) = 0$).*

From the Radon-Nikodym Theorem, we know that the standard normal probability measure Φ is dominated by the Lebegue measure λ on $\mathbb{R}$.

In the following, we list some definitions from convex analysis.

Definition 1.7.3. *If x and y are different points in $\mathbb{R}^d$, then the set of points $(1-\lambda)x+\lambda y$, $\lambda \in \mathbb{R}$ is called the line through x and y. A subset $\mathcal{M}$ is called an affine set if $(1-\lambda)x+\lambda y \in \mathcal{M}$ for every $x, y \in \mathcal{M}$ and $\lambda \in \mathbb{R}$.*

The affine set in one dimension is simply the whole real line $\mathbb{R}$.

Definition 1.7.4. *A subset $\mathcal{C}$ is of $\mathbb{R}^d$ is said to be convex if $(1-\lambda)x+\lambda y \in \mathcal{C}$, whenever $x \in \mathcal{C}, y \in \mathcal{C}$ and $0 < \lambda < 1$.*

The intersection of all convex sets containing a given subset $\mathcal{S}$ of $\mathbb{R}^d$ is called the *convex hull* of $\mathcal{S}$ and is denoted by $\operatorname{conv}\mathcal{S}$. Thus $\operatorname{conv}\mathcal{S}$ is the unique smallest convex set that contains $\mathcal{S}$.

Definition 1.7.5. *A subset $\mathcal{K}$ of $\mathbb{R}^d$ is called a cone if it is closed under positive scalar multiplication. A convex cone is a cone which is a convex set.*

Definition 1.7.6. *Let f be a function whose values are real or $\pm\infty$ and whose domain is a subset of $\mathbb{R}^d$. The set $\{(x,\mu)|x \in \mathcal{S}, \mu \in \mathbb{R}, \mu \geq f(x)\}$ is called the epigraph of f and is denoted by* $\operatorname{epi} f$.

We then define f to be a *convex function* on $\mathcal{S}$ if $\operatorname{epi} f$ is convex as a subset of $\mathbb{R}^{d+1}$. A more often used definition of a convex function is the following.

Definition 1.7.7. *Let f be a function from $\mathcal{C}$ to $(-\infty, +\infty]$, where $\mathcal{C}$ is a convex set, then f is convex if and only if*

$$f((1-\lambda)x+\lambda y) \leq (1-\lambda)f(x)+\lambda f(y), \quad 0 < \lambda < 1,$$

for every x and y in $\mathcal{C}$.

A convex function f is said to be *proper* if its epigraph is non-empty and contains no vertical lines, i.e, if $f(x) < +\infty$ for at least one x and $f(x) > -\infty$ for every x.

Definition 1.7.8. *The Euclidean unit ball in* $\mathbb{R}^d$ *is defined to be:* $B = \{x||x| \leq 1\} = \{x|d(x,0) \leq 1\}$. *Then for any set* $\mathcal{C}$ *in* $\mathbb{R}^d$, *the set of points* x *whose distance from* $\mathcal{C}$ *does not exceed* ε *is* $\{x|\exists y \in \mathcal{C}, d(x,y) \leq \varepsilon\} = \cup\{y + \varepsilon B|y \in \mathcal{C}\} = C + \varepsilon B$.

Then the closure $\mathrm{cl}\,\mathcal{C}$ and interior $\mathrm{int}\,\mathcal{C}$ can be defined by the mathematical expressions

$$\mathrm{cl}\,\mathcal{C} = \cap\{\mathcal{C} + \varepsilon B|\varepsilon > 0\}$$
$$\mathrm{int}\,\mathcal{C} = \{\exists \varepsilon > 0, x + \varepsilon B \subset \mathcal{C}\}.$$

or equivalently, the closure $\mathrm{cl}\,\mathcal{C}$ of a set $\mathcal{C} \subset \mathcal{D}$ consists of all points that are the limit of a sequence in $\mathcal{C}$; it is the smallest closed set containing $\mathcal{C}$. The interior $\mathrm{int}\,\mathcal{C}$ is the collection of all points x such that $x \in \mathcal{G} \subset \mathcal{C}$ for some open set $\mathcal{G}$; it is the largest open set contained in $\mathcal{C}$. The *relative interior* of a convex set $\mathcal{C}$ in $\mathbb{R}^d$, which is denoted by $\mathrm{ri}\,\mathcal{C}$ is then defined as the interior which results when $\mathcal{C}$ is regarded as a subset of its affine hull $\mathrm{aff}\,\mathcal{C}$,

$$\mathrm{ri}\,\mathcal{C} = \{x \in \mathrm{aff}\,\mathcal{C}|\exists \varepsilon > 0, (x + \varepsilon B) \cap (\mathrm{aff}\,\mathcal{C}) \subset \mathcal{C}\}.$$

In one dimension, the definition of the relative interior is equivalent to interior of a convex function, which is the largest open set within the domain of the convex function (the domain of a convex function is defined in definition 1.7.9).

Definition 1.7.9. *The effective domain of a convex function f on $\mathcal{S}$, which we denote by* $\operatorname{dom} f$, is the projection on $\mathbb{R}^d$ of the epigraph of f,

$$\operatorname{dom} f = \{x | \exists \mu, (x, \mu) \in \operatorname{epi} f\} = \{x | f(x) < \infty\}.$$

This is a convex set in $\mathbb{R}^d$. Its dimension is called the dimension of f. Trivially, the convexity of f is equivalent to that of the restriction of f to $\operatorname{dom} f$.

Chapter 2

The ρ-consistency of univariate ρ-concave density estimator based on the quasi-concave density estimation method

As mentioned in Chapter 1, there are two methods available regarding the general ρ-concave density estimation problem. One is from Koenker and Mizera (2010) and the other one is proposed by Seregin and Wellner (2010). To make the terminology clearer, we follow the first paper's name "quasi-concave density estimation" to denote the general method of Koenker and Mizera (2010) and use "maximum likelihood convex transformed density estimation" to denote the method of Seregin and Wellner (2010).

The organization of Chapter 2 is: we first propose the ρ-concave density estimation problem in one dimension based on the quasi-concave density estimation approach. Then we introduce a new definition of divergence, that is, the function u-divergence and its corresponding ρ-consistency, by a certain choice of the power function u associated with parameter ρ. The focus of Chapter 2 is to establish the theorem about the ρ-consistency of the ρ-concave density estimator by choosing $\rho = \alpha - 1$ for all $0 < \alpha < 1$. Finally, we

investigate the relationships between the u-divergence and other distances. We prove that the ρ-consistency actually implies the convergence of the density estimator to the true density function under both Hellinger and total variation distances.

2.1 The introduction to the ρ-concave density estimation

The formulation of the primal problem (1.12) of quasi-concave density estimation is a generalization of that for the maximum likelihood estimation for log-concave densities. Koenker and Mizera (2010) proposed the general quasi-concave density estimation problem in $\mathbb{R}^d$. Since we are going to discuss the consistency result in univariate situation, we will adapt the notations and definitions accordingly to $\mathbb{R}$. Suppose that $X_1, \ldots, X_n$ are the data points in $\mathbb{R}$, and the smallest convex set that contains the data collections has a nonempty interior in $\mathbb{R}$; such configuration occurs with probability 1 if $n \geq 1$ and X_i behave like a random sample from f_0, a probability density with respect to the Lebesgue measure on $\mathbb{R}$. We adhere to the conventions introduced in Section 1.7, which allow convex functions to take infinite values— although we will allow only $+\infty$, because all our convex functions will be proper. The domain of a convex (concave) function, $\operatorname{dom} g$, is then the set of x such that $g(x)$ is finite. We also adopt the convention $\log 0 = +\infty$.

Let us first assume that the X_i's are from an unknown, log-concave density

f_0; the usual maximum likelihood estimate of f_0 is defined by solving

$$\prod_{i=1}^{n} f(X_i) = \max_{f}! \quad \text{such that } f \text{ is a log-concave density.} \tag{2.1}$$

It is convenient to rewrite (2.1) in terms of $g = -\log f$ with the density estimate becoming $f = e^{-g}$,

$$\sum_{i=1}^{n} g(X_i) = \min_{g}! \quad \text{such that } g \text{ is convex and } \int e^{-g} dx = 1. \tag{2.2}$$

Following Silverman (1982), we may move the integral constraint over the function g into the objective function,

$$\frac{1}{n} \sum_{i=1}^{n} g(X_i) + \int e^{-g} dx = \min_{g}! \quad \text{such that } g \text{ is convex.} \tag{2.3}$$

Adding the integral term in the objective function is a device that ensures the solution integrates to one without enforcing this condition explicitly. More importantly, it makes the problem (2.3) a convex problem— note that (2.2) is not a convex problem.

Maximum likelihood estimation of log-concave densities constitutes important special case; however, the wider class allows us to include a variety of other shapes as discussed in Chapter 1. Expanding the scope of the investigation, Koenker and Mizera (2010) replaced e^{-g} into a generic function ψ which yields the primal problem of quasi-concave density estimation (1.12) introduced in Chapter 1:

$$\Phi(g) = \frac{1}{n} \sum_{i=1}^{n} g(X_i) + \int \psi(g) dx = \min_{g}!$$

and impose the conditions on the form of ψ as follows:

- (A1) ψ is a nonincreasing, proper convex function on $\mathbb{R}$.
- (A2) The domain of ψ is an open interval containing $(0, +\infty)$.
- (A3) The limit, as $\tau \to +\infty$, of $\psi(y + \tau x)/\tau$ is $+\infty$ for every real y and any $x < 0$.
- (A4) ψ is differentiable on the interior of its domain.
- (A5) ψ is bounded below by 0, with $\psi(x) \to 0$ when $x \to +\infty$.

Given assumptions $(A1) - (A5)$, Koenker and Mizera (2010) developed the dual formulation of (1.12). Since we will not use the dual formulation in the proof of the consistency in the thesis, we are not going to provide this dual formulation, except the most important relationship between the primal and the dual problem: the solution of the dual problem that is directly the estimate of the density function f satisfies $f = -\psi'(g)$, where g is the primal problem solution. Thus unlike log-concave density estimation problem (2.3), $\psi(g)$ is not necessarily the estimated density f. From the computational aspect, solving the dual problem is numerically more efficient than that of its corresponding primal problem. Strong duality, established by Koenker and Mizera (2010), means that both primal and dual problems achieve the same optimal value. Therefore we can calculate the density estimate from either primal or the dual problem.

For the log-concave density estimation problem (2.3), the choice of ψ is $\psi(g) = e^{-g}$. The authors interpret dual formulation of log-concave density estimate as an equivalent maximum (Shannon) entropy problem. A generalization to ρ-concave density estimation is to consider the dual problem as the maximum Renyi entropy problem, which yields the function ψ in the primal to

be: with $\alpha > 1$,

$$\psi(x) = \begin{cases} (-x)^\beta/\beta & \text{for} \quad x \leq 0, \\ 0 & \text{for} \quad x > 0. \end{cases}$$

For the case with $\alpha < 1$,

$$\psi(x) = \begin{cases} +\infty & \text{for} \quad x \leq 0, \\ -x^\beta/\beta & \text{for} \quad x > 0, \end{cases}$$

where $1/\beta + 1/\alpha = 1$.

The case $\alpha > 1$ is not interesting because it imposes a more restrictive form of concavity than log-concave (with $\alpha = 1$). Therefore, from our perspective, it seems more reasonable to focus on the weaker form of concavity corresponding to $\alpha < 1$. There seems no obstacle to consider $\alpha < 0$; the general primal formulation (1.12) is still applicable. The shape constraint corresponding to negative α contains more and more ρ-concave density functions, eventually achieving $(-\infty)$-concavity. However, Koenker and Mizera (2010) pointed that "the formal complications, as well as computational difficulties dictate the more prudent strategy of restricting attention to positive α cases."

Therefore, we only focus on the case $0 < \alpha < 1$. The primal problem (1.12) becomes

$$\Phi(g) = \frac{1}{n}\sum_{i=1}^{n} g(X_i) - \frac{1}{\beta}\int g^\beta dx = \min_{g \text{ is convex}} ! \tag{2.4}$$

together with the relation $f = g^{\beta-1}$ or, equivalently, $g = f^{\alpha-1}$. The density function f is called ρ-concave with $\rho = \alpha - 1$ according to the definition of the ρ-concave density functions.

Apart from the celebrated log-concave case $\alpha = 1$, a specific example of

the general formulation (2.4) that merits special attention is the situation $\alpha = 1/2$. The primal formulation can be written, after the application (2.4), in a particularly simple form,

$$\frac{1}{n}\sum_{i=1}^{n} g(X_i) + \int \frac{1}{g(x)}dx = \min_{g \text{ is convex}} !$$

The estimated density satisfies $f = 1/g^2$, which means that the primal constraint, g is convex, enforces the convexity of $g = 1/\sqrt{f}$. In the terminology of ρ-concave densities, the estimated density is now only $(-1/2)$-concave, a significant relaxation of the log-concavity constraint; in addition to all log-concave densities, all the Student t_ν densities with $\nu \geq 1$ satisfy this requirement.

2.2 The ρ-consistency of the ρ-concave density estimator

For the maximum likelihood log-concave density estimation or ρ-concave density estimation introduced in Chapter 1 and Chapter 2, many researchers considered the Hellinger consistency and uniform consistency of the nonparametric density estimator in both univariate and multivariate cases: Pal et al. (2007), Dümbgen and Rufibach (2009), Cule et al. (2010), Schuhmacher and Dümbgen (2010), and Seregin and Wellner (2010). However, proving the consistency of the ρ-concave density estimators based on problem (2.4) seems to be quite different from the usual maximum likelihood problems since the first term in problem (2.4) is not exactly the log-likelihood of the density. Moreover, adding the second integral term eliminates the need for assuming that the solution integrates to one, but since $\psi(g)$ is not necessarily the estimated density f, this convenience increases

the difficulty in the proof of consistency for the ρ-concave density estimation problem. The consequences of using formulation (2.4) are that some useful techniques in the proof of maximum likelihood log-concave density estimation can not be directly applied for the consistency proof of the ρ-concave density estimation problem (2.4), and some of the results need to be modified according to the ρ-concave problem (2.4).

Due to these obstacles, we are considering a different type of consistency for ρ-concave density functions, which we call ρ-consistency (with $\rho = \alpha - 1$ and $0 < \alpha < 1$). In the following, we first introduce the definition of the function u-divergence and ρ-consistency between the two different probability densities with respect to the Lebesgue measure.

Given the definitions and notations introduced in Section 1.7, we define the function u-divergence between the two probability distributions as follows.

Definition 2.2.1. *Let P and Q be two probability distributions over a space Ω such that P is absolutely continuous with respect to Q. For a convex function u such that $u(1) = 0$, the u-divergence of Q from P is*

$$D_u(P, Q) = \int_\Omega u\left(\frac{dP}{dQ}\right) dQ.$$

If P and Q are both absolutely continuous with respect to a reference distribution μ on Ω, then we have probability densities p and q satisfying $dP = pd\mu$ and $dQ = qd\mu$. In this case the u-divergence can be written as

$$D_u(P, Q) = \int_\Omega u\left(\frac{p(x)}{q(x)}\right) q(x)d\mu(x). \tag{2.5}$$

Lemma 2.2.2. *$D_u(P, Q) \geq 0$; if $u(t)$ is strictly convex at $t = 1$ then $D_u(P, Q) =$*

0 *only when* $P = Q$.

Proof. Following Rockafellar (1970), we have

$$\sum_i b_i f(\frac{a_i}{b_i}) \geq bf(\frac{a}{b}) \quad a = \sum a_i \text{ and } b = \sum b_i.$$

If f is strictly convex at $c = a/b$, then the equality holds if and only if $a_i = cb_i$, for all i.

Replacing the summation with integral and considering p and q to be two density function with respect to the Lebesgue measure (works for other measures as well), we conclude that

$$\int_\Omega u\left(\frac{p(x)}{q(x)}\right) q(x)d(x) \geq u\left(\int \frac{p(x)}{q(x)} q(x)dx\right) = u(1) = 0.$$

□

Some examples of functions $u(t)$ include the following:

$$\begin{aligned}
&(1) \quad u(t) = t\log t \Rightarrow D_u(P,Q) = \int p(x)\log\frac{p(x)}{q(x)}dx,\\
&(2) \quad u(t) = (t-1)^2 \Rightarrow D_u(P,Q) = \int \frac{(p(x)-q(x))^2}{q(x)}dx,\\
&(3) \quad u(t) = 1-\sqrt{t} \Rightarrow D_u(P,Q) = 1 - \int \sqrt{p(x)q(x)}dx,\\
&(4) \quad u(t) = |t-1| \Rightarrow D_u(P,Q) = \int |p(x)-q(x)|dx.
\end{aligned}$$

They correspond to KL-divergence, χ^2-divergence, Hellinger divergence and total variation distance in statistics.

Considering the nature of the power function $\psi(g) = \frac{-1}{\beta}g^\beta$, we adopt the similar definition from Liese and Vajda (2006) and Harremoës and Vajda (2011)

to define the specific function u-divergence by choosing the convex function u in Definition 2.2.1 as

$$u_\rho(t) = \frac{t^\rho - \rho(t-1) - 1}{\rho(\rho-1)} \quad \text{when } \rho \neq 0, 1 \tag{2.6}$$

with the corresponding limits

$$u_0(t) = -\ln(t) + t - 1 \quad \text{and} \quad u_1(t) = t\ln(t) - t + 1.$$

The ρ-divergence is then denoted by

$$D_\rho(P, Q) \stackrel{\text{def}}{=} D_{u_\rho}(P, Q), \quad \rho \in \mathbb{R}. \tag{2.7}$$

A simpler and equivalent formulation of ρ-divergence in equation (2.7) that will be used in the proof of the consistency theorem is

$$\begin{aligned} D_\rho(P, Q) &= \int \frac{1}{\rho(1-\rho)} \left(\frac{p}{q}\right)^\rho q dx - \int \frac{1}{\rho - 1} \left(\frac{p}{q}\right) q dx + \int \frac{q}{\rho} dx \\ &= \frac{1}{\rho(\rho-1)} \int p^\rho q^{1-\rho} dx - \frac{1}{\rho - 1} \int p dx + \frac{1}{\rho} \int q dx \\ &= \frac{1}{\rho(\rho-1)} \left[\int p^\rho q^{1-\rho} dx - 1 \right] \\ &= \frac{1}{\rho(\rho-1)} \int \left[\left(\frac{p}{q}\right)^\rho - 1 \right] q dx. \end{aligned} \tag{2.8}$$

Specifically, we call the divergence by the choice of parameter $\rho = \alpha - 1$ as $(\alpha - 1)$-divergence for all $0 < \alpha < 1$ and denote the divergence of probability distribution P to Q by $D_{\rho=\alpha-1}(P, Q)$. Consequently, the $(\alpha - 1)$-consistency can be defined as follows.

Definition 2.2.3. *If a sequence of density estimators $\{\hat{f}_n\}$ converges to the*

true density f_0 in the sense that

$$D_{\alpha-1}(\hat{f}_n, f_0) \xrightarrow{a.s.} 0,$$

then $\{\hat{f}_n\}$ is called $(\alpha-1)$-consistent estimator of f_0.

Assume that $X_1, X_2, \cdots X_n$ are iid random variables with a probability distribution F_0 on $\mathbb{R}$, with $\rho = (\alpha-1)$-concave density function f_0; let F_n be their empirical distribution function. We assume that the true density f_0 is bounded, with the corresponding convex function g_0 finite. This is a very natural assumption according to the assumption $(A3)$; such bounded assumption is also imposed in the ρ-concave density estimation problem in Seregin and Wellner (2010). Our goal is to estimate f_0 based on the random sample of size $n > 1$ from F_0. The estimator is the solution of the optimization problem

$$\Phi(g) = \frac{1}{n}\sum_{i=1}^{n} g(X_i) - \frac{1}{\beta}\int g^{\beta} dx = \min_{\text{g convex}} !. \tag{2.9}$$

We denote the minimizer of problem (2.9) as $\hat{g}_n = \arg\min_g \Phi(g)$, where the $\arg\min$ is taken over all convex functions g. The corresponding estimated density function is denoted by $\hat{f}_n = \hat{g}_n^{\beta-1}$. We further assume that for any such convex function $g = f^{\alpha-1}$, with $0 < \alpha < 1$, the first moment of g exists, in the sense that $\int g(x)f(x)dx < \infty$. The function $\psi(g) = -g^{\beta}/\beta$ satisfies conditions $(A1) - (A5)$ from Koenker and Mizera (2010). In the following, we assume $\rho = \alpha - 1$, and we will use ρ and $\alpha - 1$ interchangeably in the proof of the consistency.

Theorem 2.2.4. *Under the assumptions mentioned above, we obtain that for $(\alpha-1)$-concave density estimation problem, the $(\alpha-1)$-divergence from $\hat{f}_n$ to*

f_0 satisfies:

$$D_{\alpha-1}(\hat{f}_n, f_0) = \frac{1}{(\alpha-1)(\alpha-2)} \int ((\frac{\hat{f}_n}{f_0})^{\alpha-1} - 1) f_0 dx \xrightarrow{a.s.} 0.$$

That is, the sequence of estimators $\{\hat{f}_n\}$ is $(\alpha-1)$-consistent.

2.3 The proof of the ρ-consistency theorem

We begin with introducing some lemmas which will be used in the proof of Theorem 2.2.4. We use the notation $X_{(1)}, X_{(2)}, \ldots, X_{(n)}$ for the order statistics of $X_1, X_2, \ldots, X_n$. Based on the Theorem 1 of Koenker and Mizera (2010), the primal problem solution $\hat{g}_n$ exists and is unique, it is linear on all intervals $[X_{(j)}, X_{(j+1)}]$ for $1 \leq j < n$. Also according to conditions $(A2)$ and $(A5)$, $\hat{g}_n = +\infty$ on $\mathbb{R} \setminus [X_{(1)}, X_{(n)}]$. Lemma 2.3.1, Lemma 2.3.2 and Lemma 2.3.3 are similar to Theorem 2.2, 2.4 and Corollary 2.5 of Dümbgen and Rufibach (2009) with appropriate adaptations for the ρ-concave density estimation problem (2.9). These lemmas provide some characterizations of the estimator $\hat{g}_n$, $\hat{f}_n$ and the estimated distribution function $\hat{F}_n$.

Lemma 2.3.1. *Let $\tilde{g}_n$ be a convex function such that $\{x : \tilde{g}_n(x) < \infty\} = [X_{(1)}, X_{(n)}]$. Then $\tilde{g}_n = \hat{g}_n$, if, and only if,*

$$\int \Delta(x) dF_n \geq \int \Delta(x) \tilde{g}_n^{\beta-1} dx, \tag{2.10}$$

for any $\Delta : \mathbb{R} \to \mathbb{R}$ such that $\tilde{g}_n + \lambda\Delta$ is convex for some $\lambda > 0$.

Proof. We restrict our attention to convex and real-valued functions g on $[X_{(1)}, X_{(n)}]$ and set $g := \infty$ on $\mathbb{R} \setminus [X_{(1)}, X_{(n)}]$. The set C_n of all such functions

is a convex cone, and for any function $\Delta : \mathbb{R} \to \mathbb{R}$ and $t > 0$, the convexity of $g + t\Delta$ on $\mathbb{R}$ is equivalent to its convexity on $[X_{(1)}, X_{(n)}]$. A function $\tilde{g}_n \in C_n$ is the minimizer of $\Phi(g)$ if and only if

$$\lim_{t \to 0} \frac{\Phi(\tilde{g}_n + t(g - \tilde{g}_n)) - \Phi(\tilde{g}_n)}{t} \geq 0$$

for any $g \in C_n$. But this is equivalent to

$$\lim_{t \to 0} \frac{\Phi(\tilde{g}_n + t\Delta) - \Phi(\tilde{g}_n)}{t} \geq 0,$$

for any function $\Delta : \mathbb{R} \to \mathbb{R}$ such that $\tilde{g}_n + \lambda\Delta$ is convex for some $\lambda > 0$. Now this implies $\int \Delta(x) dF_n \geq \int \Delta(x) \tilde{g}_n^{\beta-1} dx$. □

Our next characterization is in the terms of empirical distribution function F_n and the estimated distribution function $\hat{F}_n$. In Dümbgen and Rufibach (2009), they defined for a continuous and piecewise linear function $h : [X_{(1)}, X_{(n)}] \to \mathbb{R}$, the set of its "knots" to be

$$S_n(h) := \{t \in (X_{(1)}, X_{(n)}) : h'(t-) \neq h'(t+)\} \bigcup \{X_{(1)}, X_{(n)}\}.$$

Recall that $\hat{g}_n$ in the problem (2.9) is an example of such a function h with $S_n(\hat{g}_n) \subset \{X_{(1)}, X_{(2)}, \ldots, X_{(n)}\}$.

Lemma 2.3.2. *Let $\tilde{g}_n$ which is linear on all intervals $[X_{(j)}, X_{(j+1)}]$, $1 \leq j < n$, while $\tilde{g}_n = \infty$ on $\mathbb{R} \backslash [X_{(1)}, X_{(n)}]$. Defining $\tilde{F}_n(x) = \int_{-\infty}^{x} \tilde{g}_n^{\beta-1}(r) dr$, we assume further that $\tilde{F}_n(X_{(n)}) = 1$. Then $\tilde{g}_n = \hat{g}_n$ and $\tilde{F}_n = \hat{F}_n$, if and only if, for arbitrary $t \in [X_{(1)}, X_{(n)}]$,*

$$\int_{X_{(1)}}^{t} \tilde{F}(r)dr \leq \int_{X_{(1)}}^{t} F_n(r)dr. \tag{2.11}$$

Proof. Let G be some distribution function with support $[X_{(1)}, X_{(n)}]$, and let $\Delta : [X_{(1)}, X_{(n)}] \to \mathbb{R}$ be absolutely continuous with L_1 derivative $\Delta^{'}$. Then

$$\int \Delta dG = \Delta(X_{(n)}) - \int_{X_{(1)}}^{X_{(n)}} \Delta^{'}(r)G(r)dr. \tag{2.12}$$

Now assume that $\tilde{g}_n = \hat{g}_n$, and let $t \in (X_{(1)}, X_{(n)}]$. Let Δ be absolutely continuous on $[X_{(1)}, X_{(n)}]$ with L_1 derivative $\Delta^{'}(r) = -1\{r \leq t\}$ and arbitrary value of $\Delta(X_{(n)})$. The absolute continuity of $\Delta(x)$ yields that $\tilde{g}_n + \Delta$ is convex. Thus (2.10) and (2.12) entail

$$\Delta(X_{(n)}) + \int_{X_{(1)}}^{t} \tilde{F}_n(r)dr \leq \Delta(X_{(n)}) + \int_{X_{(1)}}^{t} F_n(r)dr, \tag{2.13}$$

which is equivalent to (2.11).

In the case of $t \in S_n(\tilde{g}_n)\backslash\{X_{(1)}\}$, let $\Delta^{'}(r) = 1\{r \leq t\}$. Then $\tilde{g}_n + \lambda\Delta$ is convex for some $\lambda > 0$, so that

$$\Delta(X_{(n)}) - \int_{X_{(1)}}^{t} \tilde{F}(r)dr \leq \Delta(X_{(n)}) - \int_{X_{(1)}}^{t} F_n(r)dr, \tag{2.14}$$

which yields the equality in (2.11). □

A corollary of Lemma 2.3.2 is that the estimated distribution function $\hat{F}_n$ is very close to the empirical distribution function F_n on $S_n(\hat{g}_n)$.

Lemma 2.3.3. *$F_n - n^{-1} \leq \hat{F}_n \leq F_n$ on $S_n(\hat{g}_n)$.*

Proof. For $t \in S_n(\hat{g}_n)$ and $s < t < u$, it follows from Lemma 2.3.2 that,

$$\frac{1}{u-t}\left(\int_{X_{(1)}}^{t} \hat{F}_n(r)dr + \int_{t}^{u} \hat{F}_n(r)dr\right) \leq \frac{1}{u-t}\left(\int_{X_{(1)}}^{t} F_n(r)dr + \int_{t}^{u} F_n(r)dr\right). \tag{2.15}$$

On the other hand, since $t \in S_n(\hat{g}_n)$, we have

$$\frac{1}{u-t}\int_{X_{(1)}}^{t} \hat{F}_n(r)dr = \frac{1}{u-t}\int_{X_{(1)}}^{t} F_n(r)dr. \tag{2.16}$$

Therefore, (2.15) and (2.16) indicate

$$\frac{1}{u-t}\int_{t}^{u} \hat{F}_n(r)dr \leq \frac{1}{u-t}\int_{t}^{u} F_n(r)dr. \tag{2.17}$$

In the same way, we can conclude

$$\frac{1}{t-s}\int_{s}^{t} \hat{F}_n(r)dr \geq \frac{1}{t-s}\int_{s}^{t} F_n(r)dr. \tag{2.18}$$

Finally, let $u \downarrow t$ and $s \uparrow t$; then we can derive $\hat{F}_n(t) \leq F_n(t)$ based on inequality (2.17), and inequality (2.18) indicates that $\hat{F}_n(t) \geq F_n(t-) = F_n(t) - n^{-1}$. □

After demonstrating some interested characterizations of ρ-concave density estimators, we establish some lemmas for the proof of the consistency theorem. The next lemma provides an important inequality about the ρ-concave densities.

Lemma 2.3.4. *Let $g = f^{\alpha-1}$ be a convex function for $0 < \alpha < 1$, and f is a density function . If $0 < f(a) \leq f(b)$, then*

$$g(a) - g(b) \geq \frac{|b-a|(1-\alpha)}{\alpha}[g(b)^{\alpha/\alpha-1} - g(a)^{\alpha/\alpha-1}]. \tag{2.19}$$

Proof. Without loss of generality, we may assume that $0 < a < b$. Since

$0 < f(a) \leq f(b)$, then $g(a) \geq g(b)$, considering $\alpha - 1 < 0$. Due to the convexity of $g(x)$, for $a < x < b$ in the domain of f,

$$\begin{aligned} g(x) &= g\left(\frac{x-a}{b-a}b + \frac{b-x}{b-a}a\right) \\ &\leq \frac{x-a}{b-a}g(b) + \frac{b-x}{b-a}g(a) \\ &= g(a) + \frac{x-a}{b-a}(g(b) - g(a)). \end{aligned} \tag{2.20}$$

As $f(x)$ is a density function, and $f(x) = g(x)^{1/\alpha-1}$, we can derive the following relationship:

$$\begin{aligned} 1 &\geq \int_a^b f(x)dx \\ &= \int_a^b g(x)^{1/\alpha-1}dx \\ &\geq \int_a^b [g(a) + \frac{x-a}{b-a}(g(b) - g(a))]^{1/\alpha-1}dx \quad \text{according to (2.20)} \\ &= (\frac{1-\alpha}{\alpha})(\frac{b-a}{g(a)-g(b)})(g(b)^{\alpha/\alpha-1} - g(a)^{\alpha/\alpha-1}). \end{aligned}$$

Following the inequality above, we can conclude that

$$g(a) - g(b) \geq (\frac{1-\alpha}{\alpha})(b-a)(g(b)^{\alpha/\alpha-1} - g(a)^{\alpha/\alpha-1}). \tag{2.21}$$

If $a > b$, we can derive the similar inequality as (2.21), so finally, (2.19) holds as asserted. □

In the following, we assume that $\hat{g}_n$ attains its minimum at an order statistics, say $X_{(q)}$, which depends on n, and also reach its maximum on an order statistics, say, $X_{(m)}$, where $X_{(q)}$ and $X_{(m)} \in [X_{(1)}, X_{(n)}]$. This assumption is justified due to the fact that $\hat{g}_n$ is piecewise linear in $[X_{(1)}, X_{(n)}]$ and the

knots of piecewise linear function can only be the data points.

Lemma 2.3.5. *Suppose $g = f^{\alpha-1}$ is a convex function for $0 < \alpha < 1$, where f is a $(\alpha-1)$-concave density function, and the first moment of g exists. If $\hat{g}_n = \arg\min_g \Phi(g)$, then $\sup_{n\geq 1} \hat{g}_n(X_{(m)}) < \infty$.*

Proof. For any convex function g satisfying assumptions above, as $\Phi(g)$ is minimized by $\hat{g}_n$, we have

$$\frac{1}{n}\sum_{i=1}^{n}\hat{g}_n(X_i) - \frac{1}{\beta}\int \hat{g}_n^{\beta}dx \leq \frac{1}{n}\sum_{i=1}^{n} g(X_i) - \frac{1}{\beta}\int g^{\beta}dx.$$

This is equivalent to

$$\frac{n-1}{n}\hat{g}_n(X_{(q)}) + \frac{1}{n}\hat{g}_n(X_{(m)}) - \frac{1}{\beta}\int \hat{g}_n^{\beta}dx \leq \frac{1}{n}\sum_{i=1}^{n} g(X_i) - \frac{1}{\beta}\int g^{\beta}dx. \quad (2.22)$$

For the left-hand side of the equation (2.22), we apply Lemma 2.3.4 with $0 < \alpha < 1, \beta < 0$, $1/\alpha + 1/\beta = 1$ to derive the following inequalities:

$$\begin{aligned}
&\frac{n-1}{n}\hat{g}_n(X_{(q)}) + \frac{1}{n}\hat{g}_n(X_{(m)}) - \frac{1}{\beta(n+1)}\int \hat{g}_n^{\beta}dx \\
&\geq \frac{n-1}{n}\hat{g}_n(X_{(q)}) + \frac{1}{n}\hat{g}_n(X_{(m)}) - \frac{1}{\beta(n+1)}\hat{g}_n^{\beta}(X_{(m)})|X_{(m)} - X_{(q)}| \\
&\geq \frac{n-1}{n}\hat{g}_n(X_{(q)}) + \frac{1}{n}\hat{g}_n(X_{(m)}) + \frac{|X_{(m)} - X_{(q)}|}{n+1}\Big(\frac{\hat{g}_n(X_{(q)}) - \hat{g}_n(X_{(m)})}{|X_{(m)} - X_{(q)}|} \\
&- \frac{1}{\beta}\hat{g}_n^{\beta}(X_{(q)}) \\
&= \frac{n-1}{n}\hat{g}_n(X_{(q)}) - \frac{1}{\beta}\frac{|X_{(m)} - X_{(q)}|}{n+1}\hat{g}_n^{\beta}(X_{(q)}) + \Big(\frac{1}{n} - \frac{1}{n+1}\Big)\hat{g}_n(X_{(m)}) \\
&+ \frac{1}{n+1}\hat{g}_n(X_{(q)}). \qquad (2.23)
\end{aligned}$$

Koenker and Mizera (2010) demonstrated the existence of the solution $\hat{g}_n$ (or $\hat{f}_n$), thus we must have $\hat{g}_n(X_q) < \infty$. Moreover, when we consider the

right-hand side of the equation (2.22), due to the existence of the first moment of g and the strong law of large numbers, we have

$$\frac{1}{n}\sum_{i=1}^{n} g(X_i) \xrightarrow{a.s} \int g(x)f(x)dx = \int g(x)^{\beta}dx < \infty.$$

As a result, $\frac{1}{n}\sum g(X_i) - \frac{1}{\beta}\int g^{\beta}dx < \infty$ almost surely, and we can conclude that $\hat{g}_n(X_{(m)}) < \infty$ almost surely according to (2.23). □

The next lemma is from Schuhmacher and Dümbgen (2010).

Lemma 2.3.6. *Suppose that P and Q are two arbitrary probability measures on R^d. Let ψ: $R^d \to R$ be a bounded unimodal function, where unimodal means that the set $C_{\psi,t} := \{x \in R^d; \psi(x) \geq t\}$ are closed and convex for all $t \in R$. Then,*

$$|\int_{R^d} \psi d(P-Q)| \leq 2 \parallel \psi \parallel_{\infty} \sup_{C \in \omega^*} |P(C) - Q(C)|,$$

where ω^* denotes the set of all closed convex subsets of R^d.

Proof. The quick proof from Schuhmacher and Dümbgen (2010) is by taking $K = \parallel \psi \parallel_{\infty} < \infty$; we have

$$\begin{aligned}
\int_{\mathbb{R}^d} \psi(x)P(dx) &= -K + \int_{\mathbb{R}^d} (\psi(x) + K)P(dx) \\
&= -K + \int_{\mathbb{R}^d} \int_{[0,\infty)} 1\{r \leq \psi(x) + K\} dr P(dx) \\
&= -K + \int_{[0,\infty)} P(\{\psi \geq r - K\}) dr \\
&= -K + \int_{-K}^{K} P(C_{\psi,t}) dt
\end{aligned}$$

and therefore

$$|\int_{R^d} \psi d(P-Q)| = \int_{-K}^{K} (P(C_{\psi,t}) - Q(C_{\psi,t}))dt \leq 2K \sup_{c\in\omega^*} |P(C) - Q(C)|.$$

□

Given the random sample with the true distribution function F_0, the law of large numbers states that the empirical distribution function is consistent in the sense that

$$F_n(t) \xrightarrow{a.s} F_0(t), \quad \text{for every } t.$$

The Glivenko-Cantelli theorem extends the law of large numbers and gives uniform convergence. The uniform distance

$$\|F_n - F_0\|_\infty = \sup_t |F_n(t) - F_0(t)|$$

is also known as the *Kolmogorov-Smirnov* statistic and the Glivenko-Cantelli theorem states that

Lemma 2.3.7. *If $X_1, X_2 \ldots, X_n$ are iid random variables with distribution function F_0, then $\|F_n - F_0\|_\infty \xrightarrow{a.s.} 0$.*

We are now ready to prove Theorem 2.2.4, by using the lemmas in this section.

Proof. As $\Phi(g)$ is minimized by $\hat{g}_n$, we have

$$0 \leq \int (g_0(x) - \hat{g}_n(x))dF_n + \frac{1}{\beta}\int \hat{g}_n(x)^\beta dx - \frac{1}{\beta}\int g_0(x)^\beta dx. \tag{2.24}$$

With $\hat{g}_n^{\beta-1} = \hat{f}_n$ and $g_0^{\beta-1} = f_0$, the equation (2.24) can be written as

$$\begin{aligned} 0 \leq & \frac{\beta-1}{\beta}\int (g_0(x) - \hat{g}_n(x))dF_n + \frac{1}{\beta}\int \hat{g}_n(x)d(\hat{F}_n - F_n) \\ & + \frac{1}{\beta}\int g_0(x)d(F_n - F_0) \\ & = I_n + II_n + III_n, \end{aligned} \tag{2.25}$$

with

$$\begin{aligned} I_n &= \frac{\beta-1}{\beta}\int (g_0(x) - \hat{g}_n(x))dF_n, \\ II_n &= \frac{1}{\beta}\int \hat{g}_n(x)d(\hat{F}_n - F_n), \\ III_n &= \frac{1}{\beta}\int g_0(x)d(F_n - F_0). \end{aligned}$$

We first consider the term I_n in the equation (2.25), for any $b > 0$,

$$\begin{aligned} I_n \leq & \frac{\beta-1}{\beta}\int ((g_0(x)+b) - \hat{g}_n(x))dF_n \\ & \leq \frac{\beta-1}{\beta}\int (g_0(x)+b)d(F_n - F_0) + \frac{\beta-1}{\beta}\int \left((g_0(x)+b) - (\hat{g}_n(x)+b)\right)dF_0 \\ & + \frac{\beta-1}{\beta}\left(\int (\hat{g}_n(x)+b)dF_0 - \int \hat{g}_n(x)dF_n\right) \\ & \leq \epsilon(b) - \frac{\beta-1}{\beta}\int (\hat{g}_n(x) - g_0(x))dF_0, \end{aligned}$$

where

$$\begin{aligned} \epsilon(b) = & \frac{\beta-1}{\beta}\int (g_0(x)+b)d(F_n - F_0) + \frac{\beta-1}{\beta}\left(\int (\hat{g}_n(x)+b)dF_0 - \int \hat{g}_n(x)dF_n\right) \\ & \longrightarrow 0, \end{aligned}$$

as $b \to 0$ and $n \to \infty$. Due to the consistency of the empirical distribution function F_n and the true convex function g_0 is finite, we can apply Lemma 2.3.6 to unimodal function $-g_0$ considering $\beta < 0$ and $\beta - 1 < 0$. Furthermore, we can express $\int(\hat{g}_n(x) - g_0(x))dF_0$ as

$$\begin{aligned}\int(\hat{g}_n(x) - g_0(x))dF_0 &= \int(\hat{g}_n f_0 - g_0 f_0)dx \\ &= \int f_0(\hat{f}_n^{\alpha-1} - f_0^{\alpha-1})dx \\ &= \int \left((\frac{\hat{f}_n}{f_0})^{\alpha-1} - 1\right) f_0 f_0^{\alpha-1} dx.\end{aligned}$$

For the term $II_n = \frac{1}{\beta}\int \hat{g}_n(x)d(\hat{F}_n - F_n)$, we can further split it into two terms $II_n = IV_n + V_n$, where

$$IV_n = \frac{1}{\beta}\int(\hat{g}_n(x))d(\hat{F}_n - F_0),$$

and

$$V_n = \frac{1}{\beta}\int(\hat{g}_n(x))d(F_0 - F_n).$$

The structure of IV_n is similar as term I_n, which can be expressed in the

following way

$$\begin{aligned}
IV_n &= \frac{1}{\beta}\int(\hat{g}_n(x))d(\hat{F}_n - F_0) \\
&= \frac{1}{\beta}\int(\hat{g}_n\hat{f}_n - \hat{g}_n f_0)dx \\
&= -\frac{1}{\beta}\int f_0^{\alpha}(\frac{\hat{f}_n^{\alpha-1}}{f_0^{\alpha-1}} - \frac{\hat{f}_n^{\alpha}}{f_0^{\alpha}})dx \\
&= -\frac{1}{\beta}\int(\frac{\hat{f}_n}{f_0})^{\alpha-1} - 1)f_0 f_0^{\alpha-1}dx - \frac{1}{\beta}\int(f_0^{\alpha} - \hat{f}_n^{\alpha})dx \\
&\leq -\frac{1}{\beta}\int((\frac{\hat{f}_n}{f_0})^{\alpha-1} - 1)f_0 f_0^{\alpha-1}dx - \frac{1}{\beta}\alpha(\alpha-1)C_1 D_{\alpha}(f_0, \hat{f}_n) \\
&\leq -\frac{1}{\beta}\int(\frac{\hat{f}_n}{f_0})^{\alpha-1} - 1)f_0 f_0^{\alpha-1}dx.
\end{aligned}$$

The last two inequalities hold since

$$-\frac{1}{\beta}\int(f_0^{\alpha} - \hat{f}_n^{\alpha})dx \tag{2.26}$$

can be viewed as a negative value of another divergence with term C_1 and $D_{\alpha}(f_0, \hat{f}_n)$ explained by the following: First of all, there exist $C_1 > 0$ which is the lower bound of $g = f^{\alpha-1}$ in the domain of g for all convex function such that $g^{\beta-1} = f$ is a density function and yields the finite value of $\Phi(g)$. Such lower bound exists due to the following: If there is $C_1 > 0$ sufficiently small, then the target function will be infinitely large. For instance, if we assume that $P(g < C_1) > \varepsilon > 0$, then

$$\begin{aligned}
\Phi(g) &> \frac{1}{n}\sum g(X_i) - \frac{1}{\beta}\int_{\{g<C_1\}} C_1^{\beta}dx - \frac{1}{\beta}\int_{\{g>C_1\}} g^{\beta}dx \\
&> \frac{\varepsilon}{-\beta}c_1^{\beta} \longrightarrow \infty.
\end{aligned}$$

Therefore according to the power divergence (2.7) and the derivation of the equivalent formulation (2.8), we have

$$D_\alpha(f_0, \hat{f}_n) = \frac{1}{\alpha(\alpha-1)} \int \left((\frac{f_0}{\hat{f}_n})^\alpha - 1 \right) \hat{f}_n dx$$

for $0 < \alpha < 1$. As a result, the term (2.26) is less or equal to

$$-\frac{1}{\beta} C_1 \alpha(\alpha-1) D_\alpha(f_0, \hat{f}_n) \leq 0$$

given $0 < \alpha < 1$ and $\beta < 0$. Hence we can demonstrate that $IV_n \leq -\frac{1}{\beta} \int \left((\frac{\hat{f}_n}{f_0})^{\alpha-1} - 1 \right) f_0 f_0^{\alpha-1} dx$.

Therefore, by combining I_n and II_n, we obtain

$$I_n + II_n \leq \epsilon(b) - \int \left((\frac{\hat{f}_n}{f_0})^{\alpha-1} - 1 \right) f_0 f_0^{\alpha-1} dx + V_n.$$

According to the definition of $(\alpha-1)$-divergence in the equation (2.7) and (2.8), if $\rho = \alpha - 1$, then the $(\rho = \alpha - 1)$-divergence from $\hat{f}_n$ to f_0, is $D_{\alpha-1}(\hat{f}_n, f_0) = \frac{1}{(\alpha-1)(\alpha-2)} \int \left((\frac{\hat{f}_n}{f_0})^{\alpha-1} - 1 \right) f_0 dx$. The emergence of the power function $t^{\alpha-1}$ is not surprising, and this can be viewed as a connection between $(\alpha-1)$-concave density function and the corresponding function u-divergence. Hence, we first conclude that

$$I_n + II_n \leq \epsilon(b) - C(\alpha-1)(\alpha-2) D_{\alpha-1}(\hat{f}_n, f_0) + V_n.$$

where C again is the lower bound of $g_0 = f_0^{\alpha-1}$ in the domain of the true

density function. Moreover, Lemma 2.3.5 shows that

$$\|\hat{g}_n(x)\|_\infty = \sup |\hat{g}_n(X_{(m)})| < \infty.$$

We next show that $-\hat{g}_n$ is the unimodal function according to Lemma 2.3.6. As $-\hat{g}_n$ is a concave function, by denoting $-\hat{g}_n(x) = G(x)$, we first need to prove $\mathcal{C}_{G,t} = \{x : G(x) \geq t\}$ is a closed convex set. Let $x_1, x_2 \in \mathcal{C}_{G,t}$; the linear combination $(1-\lambda)x_1 + \lambda x_2$ with $\lambda > 0$ yields

$$G((1-\lambda)x_1 + \lambda x_2) \geq (1-\lambda)G(x_1) + \lambda G(x_2) \geq t,$$

due to the convexity of $\hat{g}_n$. Thus $-\hat{g}_n(x) = G(x)$ is a unimodal function as defined in Lemma 2.3.6. Given that $\hat{g}_n$ is bounded, we can apply Lemma 2.3.6 with $\psi = -\hat{g}_n$ $P = \hat{F}_n, Q = F_0$ in one dimension. Therefore, due to the consistency of F_n, we conclude that

$$|V_n| \leq \frac{-1}{\beta} \parallel -\hat{g}_n(x) \parallel_\infty \parallel F_0 - F_n \parallel_\infty \xrightarrow{a.s} 0.$$

Lastly,

$$|III_n| = \frac{-1}{\beta} \left| \int (-g_0) d(F_n - F_0) \right| \xrightarrow{a.s} 0,$$

as $n \to \infty$ due to the consistency of F_n, g_0 is bounded and Lemma 2.3.6.

Finally, following equation (2.25), we can conclude

$$D_{\alpha-1}(\hat{f}_n, f_0) \leq \frac{1}{C(\alpha-1)(\alpha-2)} \left(\epsilon(b) + |V_n| + |III_n|\right) \longrightarrow 0, \tag{2.27}$$

as $b \to 0$ and $n \to \infty$. In another words, we conclude that $\{\hat{f}_n\}$ is the $(\alpha-1)$-consistent estimator of f_0. □

2.4 From the ρ-divergence to the Hellinger and the total variation distance

In previous section, we introduced the definition of ρ-consistency and proved that for $\rho = \alpha - 1$, the $(\alpha-1)$-concave density estimator $\hat{f}_n$ is $(\alpha-1)$-consistent. However, this type of consistency is still dependent on the parameter of the power function α. In this section, we will establish the relationship between $(\alpha-1)$-divergence and other well-known "parameter-independent" distances, for example, the Hellinger distance and the total variation distance, and demonstrate that the $(\alpha-1)$-consistency implies the convergence of $\hat{f}_n$ to f_0 under both the Hellinger and the total variation distance.

According to the definition of the general ρ-divergence, if we choose $\rho = 1/2$, then we have the symmetric Hellinger distance from Definition 1.5.1:

$$D_{1/2}(P, Q) = 4H^2(P, Q).$$

We will prove that for all $0 < \alpha < 1$, $D_{\alpha-1}(\hat{f}_n, f_0) \geq D_{1/2}(\hat{f}_n, f_0) = 4H^2(\hat{f}_n, f_0)$. In another words, the $(\alpha-1)$-consistency implies the Hellinger consistency. An important Lemma is Proposition 2 of Harremoës and Vajda (2011).

Lemma 2.4.1. *Assume that functions f and g are C^2 and that $f''(1) > 0$ and $g''(1) > 0$. Assume that $\liminf_{t\to 0} \frac{g(t)}{f(t)} > 0$, and $\liminf_{t\to\infty} \frac{g(t)}{f(t)} > 0$. Then there exist $\gamma > 0$ such that*

$$D_g(P, Q) \geq \gamma D_f(P, Q)$$

for all distributions P and Q.

Theorem 2.4.2. *If a sequence of density estimators $\{\hat{f}_n\}$ is $(\alpha-1)$-consistent*

estimator of the true density f_0, then $\{\hat{f}_n\}$ is also Hellinger consistent:

$$H(\hat{f}_n, f_0) = \sqrt{\frac{1}{4} D_{1/2}(\hat{f}_n, f_0)} \xrightarrow{a.s.} 0. \tag{2.28}$$

Proof. First, we denote $g = u_{\rho_1}$ with $\rho_1 = 2 - \alpha$ and $f = u_{1/2}$. According to Harremoës and Vajda (2011), for any $\rho \neq 0, 1$ we have

$$D_\rho(P, Q) = D_{1-\rho}(Q, P). \tag{2.29}$$

Therefore, $D_{\alpha-1}(P, Q) = D_{2-\alpha}(Q, P) = D_g(Q, P)$. Next, it is easy to verify that $g''(1) = f''(1) = 1 > 0$. Moreover,

$$\frac{g(t)}{f(t)} = \frac{u_{\rho_1}}{u_{1/2}} = \frac{-1}{4\rho_1(\rho_1 - 1)} \frac{t^{\rho_1} - \rho_1(t-1) - 1}{t^{1/2} - 1/2(t-1) - 1}.$$

Given $0 < \alpha < 1$, then $1 < \rho_1 = 2 - \alpha < 2$, therefore we can conclude that

$$\liminf_{t \to 0} \frac{g(t)}{f(t)} = \frac{g(0)}{f(0)} = \frac{1}{2\rho_1} > \frac{1}{4} > 0$$

and

$$\begin{aligned}
\liminf_{t \to \infty} \frac{g(t)}{f(t)} &= \liminf_{t \to \infty} \frac{-1}{4\rho_1(\rho_1 - 1)} \frac{t^{\rho_1 - 1} - \rho_1 + (\rho_1 - 1)/t}{(\sqrt{t})^{-1} - 1/2 - 1/2t} \\
&= \frac{-1}{4\rho_1(\rho_1 - 1)} \liminf_{t \to \infty} \frac{t^{\rho_1 - 1} - \rho_1}{-1/2} \\
&= \frac{1}{2\rho_1(\rho_1 - 1)} \liminf_{t \to \infty} (t^{\rho_1 - 1} - \rho_1) = \infty > 0.
\end{aligned}$$

Therefore, according to Lemma 2.4.1 and symmetric property of Hellinger

distance, there exists $\gamma > 0$ such that

$$D_{2-\alpha}(f_0, \hat{f}_n) \geq \gamma D_{1/2}(f_0, \hat{f}_n) = \gamma D_{1/2}(\hat{f}_n, f_0).$$

Applying the equation (2.29), we conclude for all $0 < \alpha < 1$,

$$D_{\alpha-1}(\hat{f}_n, f_0) = D_{2-\alpha}(f_0, \hat{f}_n) \geq \gamma D_{1/2}(\hat{f}_n, f_0) = 4\gamma H^2(\hat{f}_n, f_0). \tag{2.30}$$

Therefore, Theorem 2.4.2 is assured by the equation (2.30) and Theorem 2.2.4. □

After discussing the relationship between the $(\alpha - 1)$-divergence and the Hellinger distance, we consider the total variation due to the well-known inequality between the Hellinger distance and the total variation distance. Le Cam (1969) showed that total variation distance $D_{TV}(P, Q)$ is bounded above by the multiple of the Hellinger distance $D_{TV}(P, Q) \leq \sqrt{2}H(P, Q)$. By applying Theorem 2.2.4 and 2.4.2, we conclude that

Theorem 2.4.3. *If a sequence of density estimators $\{\hat{f}_n\}$ is $(\alpha - 1)$-consistent estimator of the true density f_0, then $\{\hat{f}_n\}$ is also consistent estimator of f_0 with respect to the total variation distance:*

$$D_{TV}(\hat{f}_n, f_0) = \int |\hat{f}_n - f_0| dx \xrightarrow{a.s.} 0. \tag{2.31}$$

Therefore, once the $(\alpha - 1)$-concave is established, the convergence under the Hellinger and the total variation metrics are automatically guaranteed. We are safe to conclude that $(\alpha - 1)$-concave density estimator $\hat{f}_n$, obtained through solving the quasi-concave density estimation problem of Koenker and

Mizera (2010), is consistent under both the Hellinger and the total variation metrics.

Chapter 3

The consistency of the estimated mixture density and the decision rule in the classical Gaussian compound decision problem

In this chapter, we introduce more definitions and methods regarding the classical Gaussian compound decision problem and propose the monotone constrained mixture density estimation problem. From the theoretical point of view, we establish the Hellinger consistency of the mixture density estimator in the classical Gaussian compound decision problem. Due to the recent work of Seregin and Wellner (2010), we adopt the similar convex transformed density estimation method to further prove the pointwise convergence of the estimated convex function $\hat{k}_n(x)$ in the interior of the domain of true convex function $k_0(x)$. Last but not least, the pointwise convergence of the estimated decision rule is also demonstrated.

3.1 Mixture density estimation in the Gaussian compound decision problem

In Chapter 1, we have introduced the empirical Bayes method and Gaussian compound decision problem. There are many recent references: for instance, Brown (2008), Brown and Greenshtein (2009), Efron (2010), Efron (2011) and Jiang and Zhang (2009) consider empirical Bayes methods for Gaussian compound decision problems, especially the computational methods for non-parametric mixture models.

Recall that the classical compound decision problem amounts to estimating an n-vector $\mu_1, \ldots, \mu_n$ of parameters, under the squared error loss, where μ_i is iid from a distribution F. Given a conditionally Gaussian random sample, $X_i \sim N(\mu_i, 1), i = 1, 2 \ldots n$ (assume $\sigma^2 = 1$), the mixture density of X_i is

$$g(x) = \int \phi(x - \mu) dF(\mu).$$

If F and g are known, the optimal prediction of the μ's is given by the Bayes rule (the Tweedie's formula in Section 1.6.2)

$$\delta(x) = E(\mu | X = x) = x + l'(x) = x + \frac{g'(x)}{g(x)}, \tag{3.1}$$

where $l(x) = \log(g(x)) = \log(\int \phi(x - \mu) dF(\mu))$.

The estimation of the Bayes prediction rule $\delta(x)$ is directly associated with the estimation of the mixture density function $g(x)$ according to equation (3.1). Therefore, the estimation of the mixture density $g(x)$ plays a key role in the classic Gaussian compound decision problem and has been considered by many

authors. For example, Brown and Greenshtein (2009) showed that the simple kernel density estimates of $g(x)$ can be applied to achieve attractive performance compared with other empirical Bayes procedures. Jiang and Zhang (2009) have recently proposed a variant of the classical Kiefer and Wolfowitz (1956) nonparametric MLE as another promising approach to estimate empirical Bayes rule for the Gaussian compound decision problem. They demonstrated that the good predictive performance for the class of Gaussian compound decision problem can be achieved by an implementation of the Kiefer and Wolfowitz (1956) nonparametric maximum likelihood estimator. The original, infinitely-dimensional, formulation of the Kiefer and Wolfowitz (1956) nonparametric MLE solves,

$$\min\{-\sum_{i=1}^{n}\log(\int \phi(X_i-\mu)dF(\mu))\}, \tag{3.2}$$

where F runs over all mixing distributions. Jiang and Zhang (2009) showed that under mild moment conditions on the means, the MSE of the general maximum likelihood empirical Bayes (GMLEB) is within an infinitesimal fraction of the minimum average MSE among all separable estimators which use a single deterministic estimating function on individual observations, provided that the risk is of greater order than $(\log n)^5/n$. They also proved that the GMLEB is simultaneously uniformly approximately minimax.

The objective function (3.2) is convex and is minimized over a convex set of F, therefore we have a convex problem. This approach estimates the mixing distribution function F instead of mixture density $g(x)$, so the implied empirical Bayes rule is obtained by substituting this estimate for F through

the simple conditional expectation of μ given $X = x$:

$$\delta(x) = \frac{\int \mu\phi(x-\mu)f(\mu)d\mu}{\int \phi(x-\mu)f(\mu)d\mu},$$

where $f = F'$ is the mixing density function. The numerical method proposed by Jiang and Zhang (2009) employs the EM algorithm, a strategy introduced by Laird (1978). Considering the EM algorithm runs slow even for moderately large sample sizes toward their objective (3.2), Koenker and Mizera (2013) proposed to solve this maximum likelihood problem by establishing the dual formulation of (3.2), since it is convex. Their implementation by interior-point algorithm in Mosek is much quicker compared to the EM algorithm, as reported in Koenker and Mizera (2013). We will explore this application with numerical implementations in Chapter 4.

In the following, we will introduce an alternative method for estimating the mixture density $g(x)$. The method is inspired by the monotone property of the prediction rule which comes from the derivation of (1.31) from Chapter 1: the empirical Bayes prediction rule $\delta(x)$ is nondecreasing in x, not only for the conditional Gaussian case, but for more general one-parameter exponential families as well, no matter what the prior distribution F is. The monotonicity of the empirical Bayes prediction rule $\delta(x)$ further suggests the function

$$k(x) = \frac{1}{2}x^2 + \log g(x) \tag{3.3}$$

to be convex, where $k'(x) = \delta(x)$. As the unconstrained kernel estimators do not deliver the monotonicity property of $\delta(x)$, Van Houwelingen and Stijnen (1983) suggested a greatest convex minorant estimator based on a preliminary

histogram-type estimate of the density $g(x)$. Koenker and Mizera (2013) proposed to solve the mixture density $g(x)$ via maximum likelihood, subject to the constraint (3.3), motivated by the close link to the recent work on the maximum likelihood estimation of the log-concave densities and other shape constrained estimation problems.

Although the constraint looks quite different from the constraints used in previous nonparametric shape constrained density estimation methods, the essence of the problem is very similar. Analogous to the development of Koenker and Mizera (2010), Koenker and Mizera (2013) considered maximizing the log-likelihood

$$\sum_{i=1}^{n} \log g(X_i),$$

over the mixture density $g(x)$. Using the same technique as quasi-concave density estimation approach, they add a Lagrange term into the likelihood above to ensure the result is a density. With $\log g(x) = l(x)$, the whole task is to maximize

$$\sum_{i=1}^{n} l(X_i) - \int e^{l(x)} dx,$$

under the constraint $\frac{1}{2}x^2 + l(x)$ to be convex on $\mathbb{R}$. Evidently, this constraint is equivalent to the requirement that

$$k(x) = \log\sqrt{2\pi} + \frac{1}{2}x^2 - cx + \frac{1}{2}c^2 + \log g(x) = \log\sqrt{2\pi} + \frac{1}{2}(x-c)^2 + l(x)$$

is convex for arbitrary c. Rewriting everything in terms of $k(x)$ yields the objective function

$$\sum_{i=1}^{n} k(X_i) - n\log\sqrt{2\pi} - \frac{1}{2}\sum_{i=1}^{n}(X_i - c)^2 - \int e^{k(x)}\phi_c(x)dx,$$

where $\phi_c(x)$ is the normal density function with mean c and unit variance.

Koenker and Mizera (2013) expressed the optimization problem in the minimization form and obtain the formulation without constant terms,

$$\min_k \{-\sum_{i=1}^{n} k(X_i) + \int e^{k(x)} d\Phi_c(x) | k \in \mathcal{K}\},$$

where Φ_c denotes the Gaussian distribution function with unit variance and mean c and $\mathcal{K}$ is the cone of convex functions on $\mathbb{R}$. They further adopted $c = 0$ due to its simplicity which yields the (primal) formulation $\Phi = \Phi_0$ (and also denote $\phi_c = \phi$ as the standard normal density function)

$$\min_k \{-\sum_{i=1}^{n} k(X_i) + \int e^{k(x)} d\Phi(x) | k \in \mathcal{K}\}. \tag{3.4}$$

Problem (3.4) only differs from the general primal form of Koenker and Mizera (2010)—as introduced in Chapter 2, the log-concave density estimation problem (2.3)—in the sign of k (and correspondingly in the requirement that $k(x)$ be convex rather than concave), and in the integration measure $d\Phi(x) = \phi(x)dx$ rather than dx. The fact the objective function in problem (3.4) is convex and minimized over the convex set $\mathcal{K}$ implies that the problem is convex.

Similarly as quasi-concave density estimation, the convexity property of the problem also makes the dual interpretation possible. Koenker and Mizera (2013) provide the dual form of problem (3.4) as well as the numerical algorithms by using two independent convex programming algorithms employing interior-point methods: the PDCO algorithm of Saunders (2003) and the Mosek methods of Andersen (2010). Via the dual formulation, they prove that the solution of the primal (3.4), $\hat{k}_n(x)$, exists and is piecewise linear. Given the relationship

between primal solution $\hat{k}_n(x)$ and dual solution $\hat{f}_n(x)$, we have $e^{\hat{k}_n(x)} = \hat{f}_n(x)$. The estimated empirical Bayes rule is given by $\hat{\delta}_n(x) = \hat{k}_n'(x)$ and $\hat{g}_n(x) = \phi(x)e^{\hat{k}_n(x)} = \phi(x)\hat{f}_n(x)$. The numerical methods and more applications with additional shape constraints will be discussed in Chapter 4.

Once the existence of the maximum likelihood estimator is ensured, our attention shifts to other properties of the estimator: our main concern is to establish the consistency of the mixture density estimator and also that of the Bayes prediction rule. Due to the similarity of the problem to the log-concave density estimation problem of Koenker and Mizera (2010), it is quite plausible to explore the consistency as the similar way as the log-concave density consistency proof. Indeed, the structure of the Hellinger consistency proof for the mixture density is analogous to the work of Pal et al. (2007) along with the characterizations of the mixture density estimator from Koenker and Mizera (2013). Furthermore, the exploration of pointwise consistency of the convex function estimator $\hat{k}_n(x)$ on any compact set in the interior of the domain of the convex function $k_0(x)$ is also possible due to the connection between the Hellinger consistency of the mixture density estimator $\hat{g}_n(x)$ and the uniform consistency (pointwise convergence) of the estimated convex function $\hat{k}_n(x)$ established by Seregin and Wellner (2010). We can demonstrate the pointwise convergence of the estimated convex function $\hat{k}_n(x)$ to the true convex function $k_0(x)$ since we can reformulate problem (3.4) to a similar form as the convex transformed density estimation problem of Seregin and Wellner (2010). We will focus on the proof of the Hellinger consistency of the estimated mixture density, the pointwise convergence of the estimated convex function and the estimated decision rule in the remainder of this chapter.

3.2 The Hellinger consistency of the mixture density estimator $\hat{g}_n$

The purpose of this section is to establish the Hellinger consistency of the density estimator $\hat{g}_n(x)$. If $g_0(x) = e^{(k_0(x))}\phi(x)$ is the true mixture density and $\hat{g}_n(x) = e^{\hat{k}_n(x)}\phi(x)$ is the maximum likelihood estimator according to the problem (3.4), we will show that $\lim_{n\to\infty} H(\hat{g}_n(x), g_0(x)) = 0$ almost surely. We further denote the empirical distribution measure of the observations $X_1, X_2 \ldots, X_n$ as $\mathbb{P}_n$.

The following two lemmas are from Pal et al. (2007).

Lemma 3.2.1. *If p and q are densities, $b > 0$, P is the distribution function of p, then*

$$\int_{\mathbb{R}} \log\left(\frac{b+q}{b+p}\right) dP \leq \epsilon(b) - 2H^2(p, g),$$

where

$$\epsilon(b) = 2\int_{\mathbb{R}} \sqrt{\frac{b}{b+p}} dP.$$

Proof. In this case, according to the inequality

$$\frac{1}{2}\log(x) \leq \sqrt{x} - 1.$$

Then for any $x > 0$, we can derive the following inequalities

$$\begin{aligned}
\int_{\mathbb{R}} \log\left(\frac{b+q}{b+p}\right) dP &\leq 2\left[\int_{\mathbb{R}} \sqrt{\frac{b+q}{b+p}} dP - 1\right] \\
&\leq 2\left[\int_{\mathbb{R}} \sqrt{\frac{b}{b+p}} dP + \int_{\mathbb{R}} \sqrt{\frac{q}{b+p}} dP - 1\right] \\
&\leq \epsilon(b) + 2\left[\int_{\mathbb{R}} \sqrt{\frac{q}{b+p}} p dx - 1\right] \\
&\leq \epsilon(b) + 2\left[\int_{\mathbb{R}} \sqrt{pq} dx - 1\right] \\
&= \epsilon(b) - 2H^2(p,q).
\end{aligned}$$

□

Lemma 3.2.2. *Let $0 < b, c < \infty$. If p is a unimodal density and $\sup_x(g(x)) \leq c$, then*

$$\left| \int_{\mathbb{R}} \log(b+p) d(\mathbb{P}_n - P) \right| \leq 2 \sup_x |\mathbb{P}_n(x) - P(x)| \log\left(1 + \frac{c}{b}\right).$$

Proof. One can refer to the proof of Lemma 2 of Pal et al. (2007). □

The next lemma based on the characterizations of $\hat{k}_n(x)$ gives the bounded property of mixture density estimator $\hat{g}_n(x)$.

Lemma 3.2.3. *For the nonparametric maximum likelihood estiamtor $\hat{g}_n$ of problem* (3.4)*, we obtain that*

$$\sup_{n>1} \| \hat{g}_n \|_\infty \leq \frac{1}{\sqrt{2\pi}}.$$

Proof. We first need to review the characterization of the solution $\hat{k}_n(x)$ of the problem (3.4). Theorem 1 of Koenker and Mizera (2013) states that the

solution $\hat{k}_n(x)$ exists and is piecewise linear. The difference from Theorem 2.1 of Koenker and Mizera (2010) is that although the solution is still piecewise linear, the knots (the breakpoints of linearity in $\hat{k}_n(x)$) do not necessarily occur at the observed X_i. Without loss of generality, we assume that there are m pieces of linear function $\hat{k}_n(x)$. The linearity of $\hat{k}_n(x)$ indicates that the decision rule $\hat{\delta}_n(x)$ is piecewise constant. Denoting the constant as c_i and the density estimator as $\hat{g}_i$ on each piece, on each interval corresponding to $i = 1, 2, \ldots, m$ in $\mathbb{R}$ we have

$$c_i = x + (\log \hat{g}_i(x))'. \tag{3.5}$$

Taking the anti-derivative of the equation (3.5) yields

$$\hat{g}_i(x) = b_i e^{\frac{-1}{2}(c_i - x)^2}.$$

Although the number of pieces m may depend on n, each $b_i e^{\frac{-1}{2}(c_i-x)^2} \leq \hat{g}_n(x)$ where $\hat{g}_n$ is the estimated density on the real line. Hence, we can derive

$$b_i = \int \frac{b_i}{\sqrt{2\pi}} e^{\frac{-1}{2}(c_i-x)^2} dx \leq \frac{1}{\sqrt{2\pi}} \int \hat{g}_n(x) dx = \frac{1}{\sqrt{2\pi}}. \tag{3.6}$$

Since $\hat{g}_n(x) = \max_i \hat{g}_i$ for $i = 1, 2, \ldots, m$, we obtain that

$$\sup \hat{g}_n(x) \leq \sup_i \sup_x \hat{g}_i(x) = \sup b_i \leq \frac{1}{\sqrt{2\pi}}.$$

□

Given all the lemmas in this section, we can establish the Hellinger consistency theorem for the mixture density estimators as follows.

Theorem 3.2.4. *Assume the true density function* $g_0(x) = e^{(k_0(x))}\phi(x)$ *is*

bounded in $\mathbb{R}$ *with distribution function* P_0

$$\int \log[g_0(x)]dP_0 < \infty \tag{3.7}$$

where k_0 *is a unknown convex function. Under the conditions above, we conclude that* $H(\hat{g}_n, g_0) \xrightarrow{a.s} 0$*; that is, the sequence of estimators* $\{\hat{g}_n\}$ *is Hellinger consistent estimator of* g_0*.*

Proof. For $\varepsilon \in (0,1)$, according to condition (3.7) we have

$$\begin{aligned} 0 &\geq \int_{\{g_0(x)\leq 1-\varepsilon\}} \log[\varepsilon + g_0]dP_0 \geq \log[\varepsilon]P_0\{g_0 \leq 1-\varepsilon\} > -\infty \\ 0 &\leq \int_{\{g_0(x)\geq 1\}} \log[\varepsilon + g_0]dP_0 \leq \int_{\{g_0(x)\geq 1\}} \log[2g_0]dP_0 \\ &\leq \int \log[g_0]dP_0(x) + \log 2 < \infty. \end{aligned}$$

Therefore the function $\log[\varepsilon + g_0]$ is integrable with respect to probability measure P_0.

Since $\hat{k}_n(x)$ minimizes the objective function (3.4), we obtain that

$$-\sum_{i=1}^{n} \hat{k}_n(X_i) + \int e^{\hat{k}_n(x)}d\Phi(x) \leq -\sum_{i=1}^{n} k_0(X_i) + \int e^{k_0(x)}d\Phi(x). \tag{3.8}$$

Notice that if $\hat{k}_n(x)$ is the solution of objective function, then the estimated function $\hat{g}_n = e^{\hat{k}_n(x)}\phi(x)$ is automatically a density function. Therefore the

inequality (3.8) is equivalent to

$$\begin{aligned}
0 &\leq \sum_{i=1}^{n} \hat{k}_n(X_i) - \sum_{i=1}^{n} k_0(X_i) = \int \log[\hat{g}_n/\phi] d\mathbb{P}_n - \int \log[g_0/\phi] d\mathbb{P}_n \\
&= \int \log[\hat{g}_n] d\mathbb{P}_n - \int \log[g_0] d\mathbb{P}_n \\
&\leq \int \log[\varepsilon + \hat{g}_n] d\mathbb{P}_n - \int \log[g_0] d\mathbb{P}_n \\
&\leq \int \log[\varepsilon + \hat{g}_n] d(\mathbb{P}_n - P_0) \qquad (3.9) \\
&+ \int \log\left[\frac{\varepsilon + \hat{g}_n}{\varepsilon + g_0}\right] dP_0 \qquad (3.10) \\
&+ \int \log[\varepsilon + g_0] dP_0 - \int \log[g_0] d\mathbb{P}_n \qquad (3.11)
\end{aligned}$$

for any $\varepsilon > 0$.

Denote the terms (3.9), (3.10) and (3.11) as I_n, II_n and III_n respectively.

With Glivenko-Cantelli Theorem 1.6.1, Lemma 3.2.2 and Lemma 3.2.3, we can conclude that

$$|I_n| \leq 2 \sup_x |\mathbb{P}_n - P_0| \log\left(1 + \frac{1}{\sqrt{2\pi}\varepsilon}\right) \longrightarrow 0$$

almost surely as $n \to \infty$.

For the term II_n we apply the Lemma 3.2.1 and derive

$$II_n \equiv \int \log\left[\frac{\varepsilon + \hat{g}_n}{\varepsilon + g_0}\right] \leq 2 \int \sqrt{\frac{\varepsilon}{\varepsilon + g_0}} dP_0 - 2H^2(\hat{g}_n, g_0). \qquad (3.12)$$

At last, the strong law of large numbers implies that

$$\begin{aligned}
III_n &= \int \log[\varepsilon + g_0] dP_0 - \int \log[g_0] d\mathbb{P}_n \\
&\overset{a.s}{\to} \int \log[\varepsilon + g_0] dP_0 - \int \log[g_0] dP_0 = \int \log\left[\frac{\varepsilon + g_0}{g_0}\right] dP_0.
\end{aligned}$$

Therefore, with probability 1, we have

$$\begin{aligned} 0 &\leq \liminf(I + II + III) \\ &\leq -\limsup 2H^2(\hat{g}_n, g_0) \\ &\quad +2\int \sqrt{\frac{\varepsilon}{\varepsilon + g_0}} dP_0 + \int \log\left[\frac{\varepsilon + g_0}{g_0}\right] dP_0, \end{aligned}$$

which yields

$$\limsup H(\hat{g}_n, g_0) \leq \left(\int \sqrt{\frac{\varepsilon}{\varepsilon + g_0}} dP_0 + \frac{1}{2}\int \log\left[\frac{\varepsilon + g_0}{g_0}\right] dP_0\right)^{1/2} \to 0 \quad (3.13)$$

as $\varepsilon \downarrow 0$ by monotone convergence. Therefore, we can conclude $\{\hat{g}_n\}$ is the Hellinger consistent estimator of g_0. □

3.3 The pointwise convergence of $\hat{k}_n$ and $\hat{\delta}_n$

The next consistency property we are interested is the pointwise convergence of $\hat{k}_n(x)$ in the compact set within the interior of the domain of the true convex function $k_0(x)$. We will establish the pointwise consistency of the MLE $\{\hat{k}_n(x)\}$ once the Hellinger consistency of $\{\hat{g}_n(x)\}$ is proved. In order to obtain this result, we will first need to reformulate the problem (3.4) into the equivalent form as follows.

We denote the function $h(k(x)) = e^{k(x)}$, where the convex function $k(x) = \log(\sqrt{2\pi}) + \frac{1}{2}x^2 + \log(g(x))$ satisfies the equivalent relationship

$$e^{k(x)} = \frac{g(x)}{\frac{1}{\sqrt{2\pi}}e^{-1/2x^2}}. \quad (3.14)$$

Due to the fact that g is the mixture density with respect to the Lebesgue

measure, the monotone transformation $h(k(x))$ can be viewed as a density with respect to the standard normal measure $\Phi(x)$. Hence, this motivates us to define a new density function $q(x)$ with respect to the standard normal measure $\Phi(x)$ through

$$q(x) = \frac{g(x)}{\frac{1}{\sqrt{2\pi}} e^{-1/2x^2}} = \frac{g(x)}{\phi(x)}. \tag{3.15}$$

If $g(x)$ is the mixture density with respect to the Lebesgue measure, in the sense that $\int g(x)dx = 1$ over the domain of the density function on the real line, then $q(x)$ defined in equation (3.15) is a density function with respect to the standard normal measure, with $\int q(x)d\Phi(x) = 1$. Following the conventions in the measure theory, for any Borel measurable set $A \in \mathcal{B}$, the Borel subset of $\mathbb{R}$, the corresponding probability measure is $Q(A) = \int_A q(x)d\Phi(x)$, or $q(x) = dQ(x)/d\Phi(x)$. Similarly, the distribution function Fq_X can be defined as $Fq_X(x) = \int_{-\infty}^{x} q(x)d\Phi(x)$.

Now, we can define the new problem in a similar fashion as the problem (3.4): let $X_1, \ldots, X_n$ be n independent random variables distributed according to the Gaussian compound problem with the mixture density function g_0 with respect to the Lebesgue measure on $\mathbb{R}$. Denote the corresponding probability measure as $dP_0 = g_0(x)dx$. Define the probability density function $q_0 = \frac{g_0(x)}{\phi(x)} = h(k_0(x))$ on $\mathbb{R}$ with respect to the standard normal measure $\Phi(x)$ with $h(k_0(x)) = e^{k_0(x)}$ and k_0 an (unknown) convex function. The probability measure on the Borel sets $A \in \mathcal{B}$ corresponding to q_0 is denoted by Q_0 (which means $Q_0(A) = \int_A q_0(x)d\Phi(x)$). Since q_0 is the density function of the random variables with respect to the normal measure and g_0 is the density with respect to the Lebesgue measure, the following relationships $dP_0 = g_0dx = q_0\phi(x)dx = q_0d\Phi(x) = dQ_0$

also hold.

Notice that the problem (3.4),

$$\min_k\{-\sum_{i=1}^{n} k(X_i) + \int e^{k(x)} d\Phi(x) | k \in \mathcal{K}\}$$

is equivalent to

$$\max_g\{\int \log g d\mathbb{P}_n | k(x) = \log\sqrt{2\pi} + \frac{1}{2}x^2 + \log(g(x))\}, \tag{3.16}$$

where $g(x)$ is the mixture density with respect to the Lebesgue measure and $k(x)$ is a convex function. Furthermore, (3.16) is equivalent to $\max_q \int \log q d\mathbb{P}_n$ due to the equation (3.15). Hence we rewrite the maximum likelihood problem (3.16) as maximizing

$$\mathbb{L}_n(k) \equiv \int \log h(k(x)) d\mathbb{P}_n \tag{3.17}$$

subject to $q(x) = h(k(x)) = e^{k(x)}$, where $k(x)$ is a convex function and $q(x)$ is a density with respect to the standard normal measure Φ on $\mathbb{R}$. We denote $\hat{q}_n(x) = h(\hat{k}_n(x))$ as the estimated density function of $q_0(x)$ with respect to the measure Φ, $\hat{g}_n(x)$ as the estimated mixture density with respect to the Lebesgue measure, and $\hat{\delta}_n(x)$ as the corresponding estimated empirical Bayes prediction rule.

Theorem 3.3.1. *Under the assumption of Theorem 3.2.4, we obtain that*

$$H_\Phi(\hat{q}_n, q_0) \longrightarrow 0 \tag{3.18}$$

almost surely, where $H_\Phi(\hat{q}_n, q_0)$ is the Hellinger distance with respect to the standard normal measure Φ. In another words, the sequence of estimators $\{\hat{q}_n\}$

is the Hellinger consistent estimator of q_0.

Proof. Since we have already proved that the Hellinger consistency between $\hat{g}_n$ and g_0 in Theorem 3.2.4, we show that this is equivalent to the Hellinger consistency between $\hat{q}_n$ and q_0 with respect to the standard normal measure.

In Chapter 2, we define the Hellinger distance between the two probability measure with respect to the Lebesgue measure, here we give the similar definition of the Hellinger distance with respect to the standard normal probability measure.

Definition 3.3.2. *Let* $H_\Phi(p,q)$ *denote the Hellinger distance between densities* p *and* q*, which are densities with respect to the standard normal measure* Φ *on* $\mathbb{R}$*. Then*

$$\begin{aligned} H_\Phi(p,q) &= \left(\frac{1}{2}\int_{\mathbb{R}}(\sqrt{p}(x)-\sqrt{q}(x))^2 d\Phi(x)\right)^{1/2} \\ &= \left(1-\int_{\mathbb{R}}\sqrt{p(x)q(x)}d\Phi(x)\right)^{1/2}. \end{aligned}$$

Since the mixture density $g(x)$ is a density with respect to the Lebesgue measure, the Hellinger distance between $\hat{g}_n$ and g_0 follows from Definition 1.5.1. On the other hand from Definition 3.3.2, we obtain

$$\begin{aligned} H^2_\Phi(\hat{q}_n, q_0) &= 1-\int\sqrt{\hat{q}_n q_0}d\Phi(x) \\ &= 1-\int\sqrt{\hat{g}_n g_0}/\phi(x)d\Phi(x) \\ &= 1-\int\sqrt{\hat{g}_n g_0}dx \\ &= H^2(\hat{g}_n, g_0). \end{aligned}$$

Thus, the Hellinger distance between $\hat{g}_n$ and g_0 with respect to the Lebesgue

measure is the same as the Hellinger distance between $\hat{q}_n$ and q_0 with respect to the standard normal measure. Therefore according to Theorem 3.2.4, we conclude that $H_\Phi(\hat{q}_n, q_0) \to 0$ almost surely. □

The next Lemma is Lemma S.A.2 of the supplement to Seregin and Wellner (2010); it is used to prove the Lemma 3.3.4 later.

Lemma 3.3.3. *Let A be a convex set in $\mathbb{R}^d$ such that $\dim(A) = d$ and $\mathrm{ri}(A) \neq \varnothing$. Then:*

- *suppose a sequence of convex sets B_n is such that $A \subseteq B_n$ and $\lim \lambda[B_n \setminus A] = 0$ then $\limsup \mathrm{cl}(B_n) = \mathrm{cl}(A)$;*
- *suppose a sequence of convex sets C_n is such that $C_n \subseteq A$ and $\lim \lambda[A \setminus C_n] = 0$ then $\liminf \mathrm{ri}(C_n) = \mathrm{ri}(A)$.*

Here $\lambda[S]$ is the Lebesgue measure of S.

Analogous to Lemma 3.14 of Seregin and Wellner (2010), the next lemma allows us to obtain the pointwise consistency of the MLEs $\{\hat{k}_n(x)\}$ once the Hellinger consistency of $\{\hat{q}_n(x)\}$ is proved.

Lemma 3.3.4. *Suppose that, for the model (3.17), a sequence of MLEs $\{h(\hat{k}_n(x)) = \hat{q}_n(x)\}$ is Hellinger consistent. The sequence $\hat{k}_n(x)$ is then pointwise consistent. In other words, $\hat{k}_n(x) \xrightarrow{a.s.} k_0(x)$ for x in the interior of* `dom` *$k_0(x)$ and the convergence is uniform on compact sets.*

Proof. Let us denote L_a^0 and L_a^n in the following sublevel sets:

$$\begin{aligned} L_a^0 &= \mathrm{lev}_a k_0 = \{x : k_0(x) \leq a\} \\ L_a^n &= \mathrm{lev}_a \hat{k}_n = \{x : \hat{k}_n(x) \leq a\}. \end{aligned}$$

Consider Ω_0 such that $Pr[\Omega_0] = 1$ and $H^2(h(\hat{k}_n)^w, h(k_0)) \to 0$, where $\hat{k}_n^w$ is the MLE for $w \in \Omega_0$. For all $w \in \Omega_0$ we have:

$$\begin{aligned}\int [\sqrt{h}(k_0) - \sqrt{h}(\hat{k}_n)]^2 d\Phi(x) &\geq \int_{L_a^0 \setminus L_{a+\varepsilon}^n} [\sqrt{h}(k_0) - \sqrt{h}(\hat{k}_n)]^2 d\Phi(x) \\ &\geq (\sqrt{h}(a) - \sqrt{h}(a+\varepsilon))^2 \Phi(L_a^0 \setminus L_{a+\varepsilon}^n) \to 0.\end{aligned}$$

Therefore, $\Phi(L_a^0 \setminus L_{a+\varepsilon}^n) \to 0$. Since the standard normal density ϕ is bounded on $\{L_a^0 \setminus L_{a+\varepsilon}^n\}$, we have that $\lambda(L_a^0 \setminus L_a^n) \to 0$. By Lemma 3.3.3,

$$\liminf \mathrm{ri}(L_a^0 \cap L_{a+\varepsilon}^n) = \mathrm{ri}(L_a^0).$$

Hence $\limsup \hat{k}_n(x) < a + \varepsilon$ for $x \in \mathrm{ri}(L_a^0)$. Since a and ε are arbitrary we have $\limsup \hat{k}_n \leq k_0$ on $\mathrm{ri}(\mathrm{dom}\, k_0)$. Note that in Section 1.7, we explain that in $\mathbb{R}$ the relative interior becomes interior of the domain of the convex function; therefore, we avoid using the relative interior notation from now on, as the problem is defined in one dimension, and we conclude that $\limsup \hat{k}_n \leq k_0$ in the interior of $\mathrm{dom}\, k_0$.

On the other hand, we have

$$\begin{aligned}\int [\sqrt{h}(k_0) - \sqrt{h}(\hat{k}_n)]^2 \Phi(dx) &\geq \int_{L_{a-\varepsilon}^n \setminus L_a^0} [\sqrt{h}(k_0) - \sqrt{h}(\hat{k}_n)]^2 \Phi(dx) \\ &\geq (\sqrt{h}(a-\varepsilon) - \sqrt{h}(a))^2 \Phi(L_{a-\varepsilon}^n \setminus L_a^0) \to 0.\end{aligned}$$

By Lemma 3.3.3

$$\limsup \mathrm{cl}(L_{a-\varepsilon}^n \cup L_a^0) = \mathrm{cl}(L_a^0).$$

Therefore $\liminf \hat{k}_n(x) > a - \varepsilon$ for x such that $k_0(x) \geq a$. Since a and ε are arbitrary we have $\liminf \hat{k}_n \geq k_0$ in the interior of $\mathrm{dom}\, k_0$.

Thus $\hat{k}_n(x) \to k_0(x)$ almost surely in the interior of $\operatorname{dom} k_0$. By Theorem 10.8 of Rockafellar (1970), the convergence is uniform on every compact set $\mathcal{K}$ that belongs to the interior of $\operatorname{dom} k_0$. □

Therefore, by using Theorem 3.3.1 and Lemma 3.3.4, we can establish the following theorem.

Theorem 3.3.5. *Under the assumption of Theorem 3.2.4, we obtain that*

$$\hat{k}_n(x) \to k_0(x) \text{ almost surely in the interior of } \operatorname{dom} k_0, \tag{3.19}$$

and the convergence is uniform on any compact set contained in the interior of $\operatorname{dom} k_0$.

For the empirical Bayes rules estimation problem, estimating the mixture density $g(x)$ is only the intermediate step of the whole procedure. The final goal is to estimate the corresponding empirical Bayes prediction rule $\delta(x)$. Therefore, we consider the consistency of the estimated empirical Bayes rule $\hat{\delta}_n(x) = \hat{k}'_n(x)$ at last.

Theorem 3.3.6. *Following the assumptions above and assume that $k''(x)$ exists, then the derivative of sequence of MLEs $\{\hat{k}'_n(x)\}$ or the estimated empirical Bayes decision rule $\{\hat{\delta}_n(x)\}$ is consistent. That is $\hat{k}'_n(x) \to k'_0(x)$ almost surely and the convergence is uniform on any compact set in the interior of* $\operatorname{dom} k_0$.

Proof. By assuming $k''(x)(> 0)$ exists, then $k'(x)$ is continuous on compact sets. We now need to prove that $\hat{k}'_n(x) = \hat{\delta}_n(x)$ converges to a function $\sigma(x)$ for x in the interior of $\operatorname{dom} k_0$, and the function $\sigma(x)$ is identical to $\delta_0(x)$.

We notice that the function $\hat{\delta}_n(x)$ is the estimated decision rule, so that $\hat{\delta}_n(x)$ is a monotone function of x and bounded on the compact sets according

to Jiang and Zhang (2009). Every subsequence of $\hat{\delta}_n(x)$, say $\hat{\delta}_{n_k}(x)$, is then also bounded on the compact sets. The Bolzano-Weierstrass theorem says there is at least one subsequence $\hat{\delta}_{n_{k_i}}(x)$ of $\hat{\delta}_{n_k}(x)$ such that $\hat{\delta}_{n_{k_i}}(x) \to \sigma_i(x)$. Since we choose the subsequence $\hat{\delta}_{n_k}(x)$ arbitrarily, and each of the subsequences has a convergent subsequence, we will next show that all these sub-subsequences will converge to the same limit.

Actually for any a and x belonging to a compact set, as $n_{k_i} \to \infty$, following Lemma 3.3.4, we have

$$\begin{aligned} \int_a^x \sigma_i(t)dt &= \lim \int_a^x \hat{k}'_{n_{k_i}}(t)dt \\ &= \lim\{\hat{k}_{n_{k_i}}(x) - \hat{k}_{n_{k_i}}(a)\} \\ &= k_0(x) - k_0(a). \end{aligned}$$

Since $k'_0(x)$ exists, then the equation above indicates $\delta_0(x) = k'_0(x) = \sigma_i(x)$ for all i, so $\sigma_i(x) \equiv \sigma(x)$ are all identical. Therefore every subsequence of $\delta_n(x)$ has a convergent subsequence, and all these sub-subsequences converge to the same limit $\delta_0(x)$ (or $\sigma(x)$).

Next we will show that the original sequence $\hat{\delta}_n(x)$ must converge to $\delta_0(x)$. We use two steps to prove it. First of all, we will show that $\hat{\delta}_n(x)$ converges. To see this, if we assume the sequence $\hat{\delta}_n(x)$ diverges, then we can at least find a subsequence which also diverges, and this divergent subsequence satisfies that for all $n_k, n_l > N$, $|\hat{\delta}_{n_k} - \hat{\delta}_{n_l}| > \varepsilon$. Then we can find the subsequence of $\hat{\delta}_{n_k}$ and $\hat{\delta}_{n_l}$, say $\hat{\delta}_{n_{k_i}}$ and $\hat{\delta}_{n_{l_j}}$, the two sub-subsequences all converge to the same limit as in the argument in last paragraph. But mathematically, the subsequence index must have the order that $n_{k_i} > n_k > N$ and $n_{l_j} > n_l > N$, indicating the two sub-subsequences must satisfy the condition $|\hat{\delta}_{n_{k_i}} - \hat{\delta}_{n_{l_j}}| > \varepsilon$ and this

contradicts with the statement that these two sub-subsequences converge to the same limit. Thus, $\hat{\delta}_n(x)$ must be a convergent sequence.

Secondly, we demonstrate the limit of $\hat{\delta}_n(x)$ is $\delta_0(x)$. To see what is the limit of $\hat{\delta}_n(x)$, we assume $\hat{\delta}_n(x) \to \sigma_1(x) \neq \sigma(x)$, then the subsequence $\hat{\delta}_{n_k}(x)$ must converge to $\sigma_1(x)$ as well since $\hat{\delta}_n(x)$ is convergent, and we can find the sub-subsequence $\hat{\delta}_{n_{k_i}}(x) \to \sigma(x)$ according to the sub-subsequence property that already proved. On the other hand, since $\hat{\delta}_{n_{k_i}}(x)$ is a subsequence of $\hat{\delta}_{n_k}(x)$, $\hat{\delta}_{n_{k_i}}(x)$ should have the same limit as parent sequence $\hat{\delta}_{n_k}(x)$ which is $\sigma_1(x)$, which contradicts the assumption $\sigma_1(x) \neq \sigma(x)$.

Therefore, we can finally state that $\hat{\delta}_n(x) \to \delta_0(x)$ for x in the interior of $\operatorname{dom} k_0$, and the convergence is uniform in any compact set in the interior of $\operatorname{dom} k_0$. □

Chapter 4

Shape constraints in empirical Bayes inference

Koenker and Mizera (2013) investigated two different approaches for the estimation of the mixture density $g(x)$ in the classic Gaussian compound decision problem. We have already introduced both methods in Chapter 3: the maximum likelihood estimation of the mixture density subject the monotonicity constraint on the Bayes rule and the maximum likelihood estimation of the mixing distribution F using the mixture representation of the mixture density. In this chapter, we pursue two alternative approaches based on the two methods above, which consider the additional log-concave or the quasi-concave shape constraint(s) on the mixture density. These modifications exhibit superior behavior in the case when the shape assumption reasonably captures the behavior of the data, in particular, when the mixing distribution is unimodal. Similar as in the earlier work, the new proposals are both self-automatic, without choosing any extra tuning parameters and also yield convex optimization problems that can be efficiently solved by modern convex optimization methods. The finite-sample properties of these density estimation procedures are also discussed in this chapter. A small simulation study is presented to compare the new proposals with several existing empirical Bayes methods.

4.1 Empirical Bayes estimation for unimodal distributions

We have seen that the empirical Bayes rule based on the maximum likelihood estimation of the Gaussian mixture densities subject to the monotone constraint provides some improvements over unconstrained kernel based estimation methods. On the other hand, Kiefer-Wolfowitz nonparametric maximum likelihood estimation of mixing distribution offers good performance and the computational burden of the EM implementations of the Kiefer-Wolfowitz estimator can be dramatically reduced by the interior-point methods for solving the convex Kiefer-Wolfowitz maximum likelihood problem.

In this section, we explore some new approaches to the mixture density estimation in the classic Gaussian compound decision problem. Consider the first method we introduced in Chapter 3: the monotone constrained problem (3.4). Rewriting problem (3.4) in terms of $l(x) = \log(g(x))$, where $g(x)$ is the mixture density, yields the equivalent form,

$$\min_l \{-\sum_{i=1}^{n} l(X_i) + \int e^{l(x)} dx \mid \frac{1}{2}x^2 + l(x) \text{ convex}\}. \tag{4.1}$$

The second approach is the maximum likelihood based on the formulation of Kiefer and Wolfowitz (1956),

$$\min_F \{-\sum_{i=1}^{n} \log(\int \phi(X_i - \mu) dF(\mu))\} \tag{4.2}$$

over all mixing distributions F. The Problem (4.2) is convex as illustrated in Chapter 3, which opens the possibility of the dual presentation as well.

Theorem 2 in Koenker and Mizera (2013) provides the dual formulation of (4.2) and states that the solution of (4.2) exists and is an atomic probability measure.

Both methods as a rule yield estimators that are not very smooth—as illustrated in Koenker and Mizera (2013) and also confirmed by our numerical experiment in the next section. The estimated mixture density is piecewise linear and the estimated Bayes rule is piecewise constant by the monotonicity constraint method. The mixing distribution estimated by the maximum likelihood approach is atomic. In certain situations, the density estimate by mixing distribution is believed to be more smooth. Therefore a natural question is whether we can obtain more smooth density estimate and subsequently better prediction rule by using the density estimation with constraints on mixing distribution. Although it is always possible to control the size of the atoms of the mixing distribution by a proper upper bound and other regularization strategies, these methods would introduce additional tuning parameters. In order to avoid this, we can employ further suitable shape constraint, for example, the log-concave on the estimated mixing density.

However, adding the shape constraint on the mixing density usually leads to a non-convex problem with uncertain numerical implementation. As noted by Efron (2011), if the mixing distribution is log-concave, then the mixture density is also log-concave. Since the shape constraint on mixing distribution does not give us a convex problem, the promising way out is via switching the interest from mixing distribution to mixture density. Thus, the remedy of the non-convexity problem is to impose the shape constraint on the mixture distribution $g(x)$, rather than the mixing distribution. This way preserves the convexity of the problem—for both the approaches we introduce in this section.

Moreover, the log-concave shape constraint on the mixture distribution can take place even in the case where the mixing distribution is not log-concave. Therefore, the log-concave shape constraint on mixture density will cover a more general case than when applied to the mixing distribution. The convexity property once again ensures the efficient numerical implementation and also the theoretical insights into the problem.

Following the discussions above, we first apply the shape constraint on mixture density according to the monotonicity constraint maximum likelihood problem (4.1). A slight modification of the problem (4.1) reveals the new problem is

$$\min_{l}\{-\sum_{i=1}^{n} l(X_i) + \int e^{l(x)}dx \mid \frac{1}{2}x^2 + l(x) \text{ convex and } l \text{ concave}\}. \qquad (4.3)$$

The additional constraint is that the function $l(x) = \log g(x)$ is concave, or equivalently, that the mixture density $g(x)$ is log-concave. The constraints can also be expressed through derivatives. First, the monotone constraint gives $l''(x) \geq -1$. The concavity constraint is achieved by adding a upper bound 0 to $l''(x)$, which implies $l(x)$ is concave. The constraint can be roughly expressed in the derivative form $0 \geq l''(x) \geq -1$, which is very similar as the dual constraints in density estimation regularized with total variation. The implementation of the problem (4.3) is a straightforward extension of the problem (4.1), the minor modification brings no increase in computational complexity or running time.

The problem (4.3) is a restrictive form of the original monotone constrained mixture density estimation problem (4.1) or equivalently the problem (3.4) defined in Chapter 3. Although the formulations of the two density estimation

problems are very similar, the theoretical properties of the problem (3.4) can be applied to the problem (4.3) if and only if we impose the same constraint over the true density g_0. Equivalently, the convex function $k_0(x) = \log\sqrt{2\pi} + \frac{1}{2}x^2 + \log g_0(x)$ should have bounded second derivative $k_0''(x) \leq 1$ due to the log-concavity of g_0. Then the existence of the density estimator is guaranteed by Theorem 1 of Koenker and Mizera (2013). With the additional assumption we can also establish the similar consistency results immediately for the problem (4.3) since (4.3) is a special case of the original problem (3.4). The proof of the consistency theorem is very similar to the Hellinger consistency proof of Theorem 3.2.4 for the problem (3.4). The introduction of log-concave constraint will guarantee that the condition $\int \log[g_0(x)]dP_0 < \infty$ is satisfied—thus we do not need to make it an assumption as required in Theorem 3.2.4. The following result is from Schuhmacher and Dümbgen (2010), Lemma 2, in one-dimension.

Lemma 4.1.1. *For any log-concave density g , we have*

$$\int |\log(g(x))|g(x)dx < \infty.$$

With Lemma 4.1.1 and all the other assumptions and lemmas for the proof of the consistency of the problem (3.4), we can follow exactly the same way as the proof of Theorem 3.2.4 to derive the following theorem.

Theorem 4.1.2. *For the mixture density estimation problem* (4.3)*, assuming $k_0''(x) \leq 1$, the sequence of mixture density estimators $\{\hat{g}_n(x)\}$ is Hellinger consistent: $H(\hat{g}_n, g_0) \xrightarrow{a.s} 0$.*

Once the Hellinger consistency of mixture density is established, by using the same trick of convex transformation as we applied to the problem (3.4), we

can similarly obtain the pointwise consistency of estimated convex function $\hat{k}_n(x)$ and the empirical Bayes decision rule $\hat{\delta}_n(x)$ on any compact set in the interior of the domain of convex function $k_0(x)$ as Theorem 3.3.5 and 3.3.6.

Theorem 4.1.3. *For the mixture density estimation problem* (4.3)*, assuming that* $k_0''(x) \leq 1$*, the sequence of convex functions* $\{\hat{k}_n(x)\}$ *is consistent:* $\hat{k}_n(x) \xrightarrow{a.s} k_0(x)$ *for* x *in the interior of* $\operatorname{dom} k_0$ and the convergence is uniform on any compact set.

Theorem 4.1.4. *For the mixture density estimation problem* (4.3)*, assuming that* $k_0''(x) \leq 1$*, the derivative of sequence of estimators* $\{\hat{k}_n'(x)\}$ *or the estimated empirical Bayes decision rule* $\{\hat{\delta}_n(x)\}$ *is consistent. That is* $\hat{k}_n'(x) \xrightarrow{a.s} k_0'(x)$ *for* x *in the interior of* $\operatorname{dom} k_0$ and convergence is uniform on any compact set.

We can also consider to impose the shape constraint on the mixture density in the nonparametric likelihood problem (4.2). But the problem is a bit more tricky than the original problem (4.1) since it can even create a non-convex problem by adding the shape constraint over $g(x)$. The way out is by introducing a slack function and the constraint expressing in an epigraph, inequality form

$$\min_{h,F}\{\sum_{i=1}^{n} h(X_i) \mid h(x) \geq -\log(\int \phi(x-\mu)dF(\mu)) \text{ and } h \text{ convex}\}. \qquad (4.4)$$

The emergence of convex function $h(x)$ dominates the negative logarithm of mixture density and preserves the convexity of the problem (4.4). The inequality in (4.4) should be satisfied over all x from some fine grid (not only over the actual observed X_i). Thus, it brings more computational complexity in this formulation. Nevertheless, due to the dominance of maximum likelihood

approach over the monotonicity constraint of the Bayes rule, it still worth the effort to pursue such application. Moreover, this method also allows for the incorporation of the general ρ-concave constraints considered in previous chapter. Last but not least: from the theoretical point of view, establishing the consistency property the density estimator of the problem (4.4) is an important future work.

4.2 The discrete formulations and numerical comparisons

So far we have discussed the theoretical properties of the monotone constrained mixture density estimation method, the Kiefer-Wolfowitz nonparametric maximum likelihood estimation of mixing distribution method and also the new approaches with additional shape constraints based on the these two methods. In this section, we are going to illustrate the discrete formulations of these four methods and conduct a numerical experiment to compare the performance of the new approaches with other existing methods.

First of all, the numerical implementation of the problem (4.1) requires the discrete formulation of the monotone constrained mixture density estimation method. This is accomplished by choosing a fine grid of points, say $x_1 < x_2 < \cdots < x_m$, setting $\alpha_i = l(x_i) = \log g(x_i)$ to be unknown function values of the mixture density and solving

$$\max_{\alpha}\{w^T\alpha - \sum c_i e^{\alpha_i} \mid D\alpha + 1 \geq 0\}. \tag{4.5}$$

The matrix D represents the finite difference version of the second derivative

operator that in the variational form in the problem (4.1), is involved in the monotonicity constraint, $D\alpha + 1 \geq 0$. The accuracy of the Riemann approximation of the integral is controlled by the fineness the grid, which increases the number of estimated function values as a result. The typical choice of the fine grid with equal spacing is $m \approx 300$, as reported by Koenker and Mizera (2013). The vector w is an evaluation term that simply allows to recover and sum up the contributions to the likelihood term given the expanded vector of function values; the c_i's are Riemann weights of the integral term corresponding to the the fine grid. Koenker and Mizera (2013) also investigate the finite-dimensional formulation for the dual problem of (4.1). The dual form of the estimator is implemented by using two independent convex programming algorithms both utilize the interior-point methods: the PDCO algorithm of Saunders (2003) and Mosek method of Andersen (2010).

The numerical experiments reveal that the fitted $\hat{k}'_n(x)$ is piecewise linear and the estimated Bayes rule $\hat{\delta}_n(x) = \hat{k}'_n(x)$ is piecewise constant which is assured by the Theorem 1 in Koenker and Mizera (2013). But the piecewise linear property of $\hat{k}'_n(x)$ consequently makes $\log \hat{g}_n(x)$ to be piecewise quadratic, which leads the density estimate $\hat{g}_n(x)$ looks rather bizarre compared to the conventional kernel density estimate. The numerical plot of the piecewise constant Bayes rule in Koenker and Mizera (2013) also illustrates that its jumps do not (necessarily) occur at the observed data points. The estimates of the mixture density may look a bit strange, but their implied Bayes rules nevertheless conform to the monotonicity requirement and perform quite well—as we will see in our numerical experiment.

The numerical implementation of problem (4.3) is similar to problem (4.1). A discrete formulation of the shape constrained MLE can be obtain similarly

as

$$\max_{\alpha}\{w^T\alpha - \sum c_i e^{\alpha_i} \mid 0 \geq D\alpha \geq -1\}. \tag{4.6}$$

All notations in the finite discrete form (4.6) are the same as the formulation (4.5). The additional log-concave shape constraint is expressed through the condition $0 \geq D\alpha$.

Regarding the implementation of the Kiefer-Wolfowitz nonparametric maximum likelihood approach (4.2), Jiang and Zhang (2009) recently proposed a fixed EM iteration that requires a grid $\{\mu_1, \ldots, \mu_m\}$ containing the support of the observed sample. This produces a sequence

$$\hat{f}_j^{k+1} = \frac{1}{n}\sum_{i=1}^{n} \frac{\hat{f}_j^k \phi(X_i - \mu_j)}{\sum_{l=1}^{m} \hat{f}_l^k \phi(X_i - \mu_l)},$$

where $\hat{f}_j^k$ denote the value of the estimated "prior" mixing density on the interval (μ_j, μ_{j+1}) at the kth iteration. The decision rule is simply the conditional expectation of μ_i given X_i,

$$\hat{\delta}_n(X_i) = \frac{\sum_{j=1}^{m} \mu_j \phi(X_i - \mu_j)\hat{f}_j^k}{\sum_{j=1}^{m} \phi(X_i - \mu_j)\hat{f}_j^k}.$$

Jiang and Zhang (2009) reported good performance of their simulations applying a design of Johnstone and Silverman (2004), but it is hard for us to reproduce the experiment by the EM algorithm since it yields slow implementation even for the moderate large sample sizes.

An alternative proposed by Koenker and Mizera (2013) is to apply the convex optimization tool to solve the problem (4.2) since the problem itself is convex. Let $\{\mu_1, \ldots, \mu_k\}$ be the fixed grid for μ and $\{x_1, \ldots, x_m\}$ be the fixed grid for the data. Let A be an m by k matrix, with the elements $\phi(X_i - \mu_j)$

on the ith row and jth column. The discrete form of the primal problem (4.2) can be then expressed as

$$\min\{-\log(1_m^T g) \mid Af = g, f \in S\}, \tag{4.7}$$

where S denotes the unit simplex in $\mathbb{R}^k$: $S = \{s \in \mathbb{R}^k \mid 1^T s = 1, s \geq 0\}$. Now f_j is the estimated mixing density estimate $\hat{f}_n$ evaluated at the grid point μ_j, and g_i denotes the estimated mixture density estimate, $\hat{g}_n$, evaluated at x_i. Koenker and Mizera (2013) also provide the dual formulation of problem (4.7) and the interior-point algorithm as implemented in Mosek solves the dual problem quicker and more accurately compared to the EM procedures.

In the simulation experiments, this approach has a slight, but visible edge over the monotone constraint maximum likelihood method; and both dominate all other existing methods. The algorithm employing monotonicity constraint on the Bayes rule easier scales to larger data sets while the nonparametric Kiefer-Wolfowitz maximum likelihood is more comfortable with the inclusion of covariates.

The finite dimensional discrete formulation of the problem (4.4) can be derived in a analogous manner to the discrete work of the problem (4.2):

$$\min\{-1_m^T h \mid Af = g \geq e^{-h}, f \in S, Dh \geq 0\}. \tag{4.8}$$

All the notations in the problem (4.8) are the same as the problem (4.7).

In order to compare the proposed new estimators/prediction schemes with other existing methods, we finally perform a simulation study using the following mixing distributions: the uniform distribution on $[5, 15]$ (featured in

the examples shown in Figure 4.1 and 4.2); the t distribution with 3 degrees of freedom; the χ^2 distribution with 2 degrees of freedom; and four instances from simulation study of Johnstone and Silverman (2004), employed also by Koenker and Mizera (2013): a mixture consisting of two numbers, k of which (we choose $k = 5$ and $k = 50$) are equal to some fixed μ (we use $\mu = 2$ and $\mu = 5$) and the rest are zeros. For each distribution, we performed 1000 repetitions. The results in Table 1 show the averages of the sums, computed for each repetition, of squared errors for all sampled μ_i. The sample size is $n = 100$ for all cases. In addition to the methods considered above, Table 1 shows also the results of the "naive" maximum likelihood predictor $\mu_i = X_i$; the predictor based on the James-Stein estimator

$$X_i - \frac{n-3}{\sum_{i=1}^n (X_i - \bar{X})^2}(X_i - \bar{X}), \quad \bar{X} = \frac{1}{n}\sum_{i=1}^n X_i;$$

and the "oracle" predictor, the optimal Bayes predictor assuming the knowledge of the mixing distribution.

The simulation results indicate that the precision gain of the Kiefer-Wolfowitz maximum likelihood method (4.2) over the monotone constrained method (4.1), as confirmed by similar experiments in Koenker and Mizera (2013), is still somewhat preserved when we compare the maximum likelihood method (4.4) with log-concavity constrained to the monotone and log-concavity constrained approach (4.3). The actual magnitude of the differences becomes rather negligible, and obviously can be offset by the lower computational complexity of the method (4.3) over (4.4). The plots in Figure 4.1 and 4.2 confirm what we have observed about the new approaches: firstly, the monotone constrained method (4.1) produces the piecewise quadratic estimated mixture

	$U[5,15]$	t_3	χ^2_2	$0_{95}\|2_{05}$	$0_{50}\|2_{50}$	$0_{95}\|5_{05}$	$0_{50}\|5_{50}$
br	101.5	112.4	77.8	19.7	57.3	12.6	21.1
kw	92.6	114.4	71.9	17.4	51.3	10.0	17.0
brlc	85.6	98.1	67.6	17.3	51.7	21.6	58.2
kwlc	84.9	98.2	66.8	16.5	50.4	21.2	67.6
mle	100.2	100.1	100.2	100.7	100.4	100.1	99.6
js	89.8	98.5	80.2	18.5	52.1	56.2	86.8
oracle	81.9	97.5	63.9	12.6	44.9	4.9	11.5

Table 1: The empirical risk of several estimators/prediction schemes: the MLE of the mixture density with the monotonicity constraint on the prediction rule (br); the Kiefer-Wolfowitz nonparametric MLE of the mixing distribution (kw); their versions, (brlc) and (kwlc) respectively, with mixture density constrained to be log-concave; the MLE (no shrinkage) predictor (mle); the James-Stein estimator/predictor, assuming the normal mixing distribution (js); and finally the "oracle" predictor, the Bayes rule employing the knowledge of the mixing distribution.

density and piecewise constant estimated Bayes prediction rule; secondly, the new approach (4.3) with log-concave constraint gives more smooth estimator than the original monotone constrained method (4.1); third, the original Kiefer-Wolfowitz maximum likelihood method (4.2) produces the estimate with visible edge, but looks smoother than the original monotone constrained method (4.1); at last, the Kiefer-Wolfowitz maximum likelihood with additional shape constraint method (4.4) generates smoother estimator than its original method (4.2).

It also worth mentioning that a different—reversed—behavior is observed for the heavy-tailed distribution, t_3—not only for the versions with enforced log-concavity, but also for the unrestricted ones—which are, interestingly, in this case dominated by the "naive" predictor $\hat{\mu}_i = X_i$ (which is however, still dominated by the oracle predictor, in accord with the theory). The squared errors of the traditional MLE for all choices of mixing distribution are all

around 100 which is the number of observations as demonstrated in Section 1.6. Finally, we can see that the versions with enforced mixture log-concavity still dominate the James-Stein predictor, based on the assumption of the normality of the mixture distribution—in the case of asymmetric mixing distribution χ^2_2, the difference in efficiency seems to be substantial.

On the other hand, it should be prudent to apply the new approaches described above, since they give better prediction properties and estimates only when the true distribution is unimodal. For example in the last two examples, where we choose μ to be 0 or 5: no matter we select 95 of them to be 0 and remaining of μ are 5, or 50 of them are 0 and the remaining 50 are 5, the convolution of these two mixing distribution functions with the normal density $\phi(x)$ yields the mixture density $g(x)$ to be bimodal. Therefore our new proposals do not perform any better than the original monotone constrained mixture density estimation method and Kiefer-Wolfowitz maximum likelihood, since the additional shape constraint does not capture the true behaviour the data. Hence, one can always first apply the preliminary density estimation approaches introduced in Chapter 1 to check the modality of the density and then decide which method to use. For bimodal cases, the unconstrained methods are preferable.

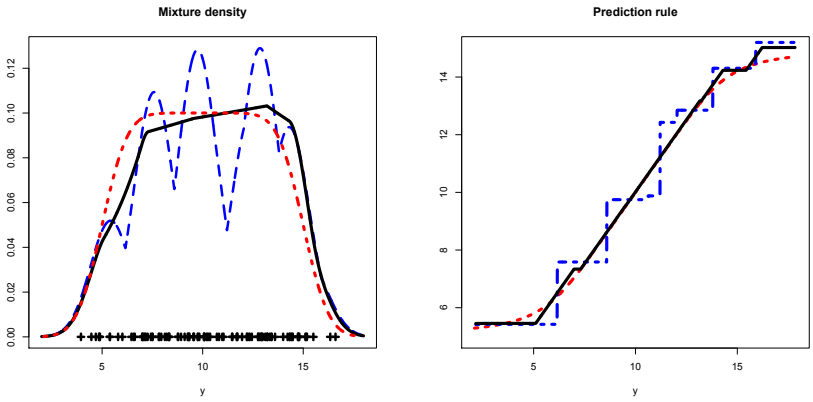

Figure 4.1: Estimated mixture density (left) and corresponding Bayes rule (right) for the monotone constrained maximum likelihood (dashed blue) and the log-concave shape constrained variant (solid). The target, "oracle" mixture density and its Bayes rule are plotted in dotted red line.

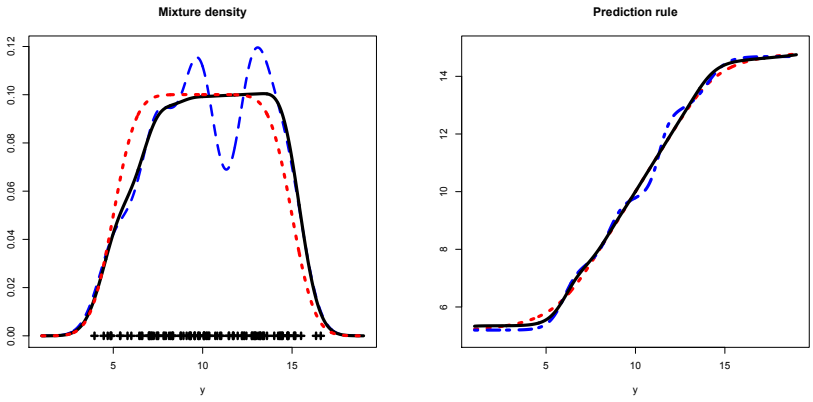

Figure 4.2: Estimated mixture density (left) and corresponding Bayes rule (right) for the Kiefer-Wolfowitz maximum likelihood (dashed blue) and its log-concave shape-constrained variant (solid). The target, "oracle" mixture density and its Bayes rule are again plotted in dotted red line.

Chapter 5

Summary of the results and the future work

5.1 Summary of the results

In this chapter, we present a thorough summary of all the original work/results from Chapter 2 to Chapter 4 and point out some promising and important future work in the $\rho-$concave density estimation and mixture density estimation in the Gaussian compound decision problem .

First of all, in Chapter 2, we apply the definition of function u-divergence from probability distribution P to Q to establish the $(\alpha-1)$-consistency, in one dimension, in $(\alpha-1)$-concave density estimation problem (2.9), based on quasi-concave density estimation method proposed by Koenker and Mizera (2010). The $(\alpha-1)$-concave density family is an important subset of the quasi-concave densities by choosing the function $\psi(g)$ from the power function family and setting $0<\alpha<1$. From our perspective, the $(\alpha-1)$-concave density functions discussed in this chapter are weaker forms of concave functions than well studied log-concave density functions. In Koenker and Mizera (2010), the authors make some conjectures about the convergence of the $(\alpha-1)$-concave density estimators. Thus, the $(\alpha-1)$-consistency result not only can be viewed as a complement to the theoretical results of Koenker and Mizera (2010), but

also confirms the simulation study in the paper.

Given that the $(\alpha-1)$-consistency property depends on the power function parameter α, we strengthen and extend $(\alpha-1)$-consistency into the convergence under other distances that are free of parameters. In view of many recent results regarding the Hellinger consistency of log-concave and ρ-concave density estimators, we also investigate the Hellinger consistency of $(\alpha-1)$-concave density estimators. Specifically, we prove that the $(\alpha-1)$-consistency actually implies the convergence of the sequence of $(\alpha-1)$-concave density estimators $\{\hat{f}_n\}$ to f_0 with respect to the Hellinger distance. Due to the familiar relationship between the Hellinger distance and the total variation distance, we also demonstrate the convergence of density estimators under the total variation distance.

Chapter 3 concerns the mixture density estimation in the classic Gaussian compound decision problem. Following the recently developed monotonely constrained mixture density estimation method in Koenker and Mizera (2013), we give the original proof of the Hellinger consistency property of the mixture density estimators, which substantiates the numerical experiment performance in Koenker and Mizera (2013). As the convexity constraint is imposed on $k(x)$, and the mixture density $g(x)$ is closely associated with the convex function $k(x)$, we further adopt the similar formulation of convex-transformed density estimation approach of Seregin and Wellner (2010) to establish the pointwise consistency of the estimated convex function $\hat{k}_n(x)$ in any compact set in the interior of the domain of the true convex function $k_0(x)$. In addition, since the monotone compound decision rule $\delta(x)$ is the first derivative of the convex function $k(x)$, we also prove the pointwise convergence of the estimated decision rule. Thus, we eventually figure out three important consistency problems in

the classic Gaussian compound decision problem.

Last but not least, we propose two new approaches to estimate the mixture density in the classic Gaussian compound decision problem in Chapter 4. The new proposals are natural extensions of the work of Koenker and Mizera (2013) and enrich the mixture density estimation methods for classic Gaussian compound decision problem. The new methods are developed by imposing additional shape constraints on the monotone constrained mixture density estimation and the Kiefer-Wolfowitz maximum likelihood mixing distribution estimation methods, respectively. In particular, since the shape-constrained version of monotone constrained mixture density estimation problem is a special case of the original monotone constrained maximum likelihood density estimation problem, the theoretical properties of this new method can be also established by assuming that the true density satisfies the additional log-concave shape constraint. Due to the convexity of the new approaches, all methods can be easily implemented by modern convex optimization tools and our numerical experiment shows that the new approaches perform better than their earlier work in terms of lower mean squared error in Table 1 and produce smoother density estimate as shown in Figure 4.1 and 4.2.

5.2 Future work

Although we prove a series consistency results and propose some new density estimation approaches, there are still some open questions left and interesting future work worthy more effort—in both the ρ-concave density estimation and mixture density estimation in the classic Gaussian compound decision problem.

For instance, one of the possible problems of interest regarding the ρ-concave

density estimation is to generalize the current results into multidimensional setting—which is more challenging since many of our univariate proof tricks are not applicable in higher dimensions. These results are plausible, because the numerical experiment in Koenker and Mizera (2010) already confirms the performance of the density estimators in higher dimensions. On the other hand, due to the recent multidimensional work of Schuhmacher and Dümbgen (2010), Cule et al. (2010), Cule and Samworth (2010) and Dümbgen et al. (2013) for the log-concave density estimation, we believe these excellent results can be applied or adapted to ρ-concave density estimation problem as well. Moreover, according to Lemma 2.3.1—2.3.3, we can see that the estimated distribution function is very close to the empirical distribution function on the set $S_n(\hat{g}_n)$. Thus another perspective is to establish the similar convergence result of $\hat{F}_n \to F_n$ as Dümbgen and Rufibach (2009) did, at least in the fixed compact set on the real line.

With respect to the mixture density estimation problem in Chapter 3, an important perspective is establishing the convergence rate of the mixture density estimators. For the new approach (4.4) proposed in Chapter 4, the theoretical results for this method are still open problems for future work. Another possible direction that takes both the ρ-concave density estimation and mixture density estimation into account is considering ρ-concave shape constraints other than log-concavity in the new proposals of mixture/mixing density estimation problem in Chapter 4, since the ρ-concavity with $\rho < 0$ imposes weaker concavity than log-concavity. This is more tricky since the naive application of the ρ-concave constraint on mixture density can produce a non-convex form, which makes the numerical implementation much harder.

Bibliography

An, M. Y. (1995). Log-concave probability distributions: Theory and statistical testing. Technical report, Economic Dept., Duke Univ.

An, M. Y. (1998). Logconcavity versus logconvexity: A complete characterization. *J. Econom. Theory*, 80:350–369.

Andersen, E. D. (2010). The mosek optimization tools manual, version 6.0. Available from http://www.mosek.com.

Avriel, M. (1972). r-concave functions. *Math. Program.*, 2:309–323.

Bagnoli, M. and Bergstrom, T. (2005). Log-concave probability and its applications. *Econometric Theory*, 26:445–469.

Bertrand-Retali, M. (1978). Convergence uniforme d'un estimateur de la densité par la méthode de noyau. *Rev. Roumaine Math. Pures Appl.*, 23:361–368.

Borell, C. (1975). Convex set functions in d-space. *Period. Math. Hungar.*, 6:111–136.

Brooks, S. P. (1998). MCMC convergence diagnosis via multivariate bounds on log-concave densities. *Ann. Statist.*, 26:398–433.

Brown, L. (2008). In-season prediction of batting averages: A field test of empirical Bayes and Bayes methodologies. *Ann. Statist.*, 2:113–152.

Brown, L. and Greenshtein (2009). Non-parametric empirical Bayes and compound decision approaches to estimation of a high dimensional vector of normal means. *Ann. Statist.*, 37:1685–1704.

Caplin, A. and Nalebuff, B. (1991). Aggregation and social choice: A mean voter theorem. *Econometrica*, 59:1–23.

Cule, M. and Samworth, R. (2010). Theoretical properties of the log-concave maximum likelihood estimator of a multidimensional density. *Electron. J. Statist.*, 4:252–270.

Cule, M. L., Gramacy, R. B., and Samworth, R. J. (2007). *LogConcDEAD: maximum likelihood estimation of a log-concave density*. URL http://CRAN.R-project.org/package=LogConcDEAD.

Cule, M. L., Samworth, R. J., and Stewart, M. I. (2010). Maximum likelihood estimation of multidimensional log-concave density. *J. R. Statist. Soc. B*, 72:545–607.

Dharmadhikari, S. and Joag-Dev, K. (1988). *Unimodality, convexity and Applications*. Academic Press, Boston, MA.

Dümbgen, L. Hüsler, A. and Rufibach, K. (2007). Active set and EM algorithms for log-concave densities based on complete and censored data. Preprint.

Dümbgen, L. and Rufibach, K. (2009). Maximum likelihood estimation of a log-concave density: Basic properties and uniform consistency. *Bernoulli*, 15:40–68.

Dümbgen, L. and Rufibach, K. (2011). Logcondens: Computations related to univariate log-concave density estimation. *J. Statist. Software*, 39:1–28.

Dümbgen, L., Samworth, R, J., and Schuhmacher, D. (2013). Stochastic search for semiparamatric linear regression models. *In from probability to statistics and back: High-Dimension models and process—A festschrift in honor of Jon A. Wellner. Eds M. Banerjee, F. Bunea, J. Huang, V. Koltchinskii, M. H Maathuis*, pages 78–90.

Efron, B. (2008). Microarrays, empirical Bayes and the two-groups model. *Statist. Sci.*, 23:1–22.

Efron, B. (2010). *Large-Scale inference: empirical Bayes methods for estimation, testing and prediction*. Cambridge Univ. Press: Cambridge.

Efron, B. (2011). Tweedie's formula and selection bias. *J. Amer. Statist. Assoc.*, 106:1602–1614.

Good, I. and Gaskins, R. (1971). Nonparametric roughness penalties for probability densities. *Biometrika*, 58:255–277.

Grenander, U. (1956). On the theory of mortality measurement, part II. *Skandinavisk Aktuarietidskrift*, 39:125–153.

Groeneboom, P., J. G. and Wellner, J. A. (2001). Estimation of a convex function: Characterizations and asymptotic theory. *Ann. Statist.*, 29:1653–1698.

Groeneboom, P. (1988). Brownian motion with a parabolic drift and Airy functions. *Prob. Theory Related Fields*, 81:79–109.

Harremoës, P. and Vajda, I. (2011). On pairs of f-divergences and their joint range. *IEEE. Trans. Inform. Theory*, 57(6).

Huang, J. and Wellner, J. A. (1995). Estimation of a monotone density or monotone hazard under random censoring. *Scandinavian J. Statist.*, 22(1).

Huang, Y.-P. and Zhang, C.-H. (1994). Estimating a monotone density from censored observations. *Ann. Statist.*, 22(3).

Ibragimov, I. A. (1956). On the composition of unimodal distributions. *Theory Probab. Appl.*, 1:255–260.

James, W. and Stein, C. (1961). Estimation with quadratic loss. In *Proc. 4th Berkeley Sympos. Math. Statist. and Probab.*, volume I, pages 361–379, Berkeley. Univ. California Press.

Jiang, W. and Zhang, C.-H. (2009). General maximum likelihood empirical Bayes estimation of normal means. *Ann. Statist.*, 37:1647–1684.

Johnstone, I. and Silverman, B. (2004). Needles and straw in haystacks: Empirical Bayes estimates of possibly sparse sequences. *Ann. Statist.*, 32:1594–1649.

Kappel, F. and Kuntsevich, A. (2000). An implementaion of Shor's $\gamma-$algorithm. *Computational Optimization and Applications*, 15:193–205.

Karunamuni, R., Sriram, T., and Wu, J. (2006a). Asymptotic normality of an adaptive kernel density estimator for finite mixture models. *Statist. Probab. Lett.*, 76:211–220.

Karunamuni, R., Sriram, T., and Wu, J. (2006b). Rates of convergence of an adaptive kernel density estimator for finite mixture models. *Statist. Probab. Lett.*, 76:221–230.

Kiefer, J. and Wolfowitz, J. (1956). Consistency of the maximum likelihood estimator in the presence of infinitely many incidental parameters. *Ann. Math. Statist.*, 27:887–906.

Koenker, R. and Mizera, I. (2006). Density estimation by total variation regularization. *Advances in statistical modeling and inference, Essays in honor of Kjell A. Doksum (V. Nair, ed)*. World Scientific, Singapore.

Koenker, R. and Mizera, I. (2010). Quasi-concave density estimation. *Ann. Statist.*, 38:2998–3027.

Koenker, R. and Mizera, I. (2013). Convex optimization, shape constraints, compound decisions, and empirical Bayes rules. *J. Amer. Statist. Assoc.*, to appear.

Laird, N. (1978). Nonparametric maximum likelihood estimation of a mixing distribution. *J. Amer. Statist. Assoc.*, 11:86–94.

Le Cam, L. (1969). *Théorie Asymptotique de la Décision Statistique*. Univ. of Montreal Press.

Liese, F. and Vajda, I. (2006). On divergences and informations in statistics and information theory. *IEEE. Trans. Inform. Theory*, 52:4394–4412.

Marshall, A. W. and Olkin, I. (2007). *Life distribution*. Springer Series in Statistics, Springer, New York.

Mengersen, G, L. and Tweedie, R. L. (1996). Rate of convergence of the Hastings and Metropolis algorithms. *Ann. Statist.*, 24:101–121.

Pal, J. K., Woodroofe, M., and Meyer, M. (2007). Estimating a Polya frequency function. *In Complex Datasets and Inverse Problems: Tomography, Networks and Beyond, (R. Liu, W. Strawderman, and C.-H. Zhang, eds.). IMS Lecture Notes-Monograph*, 54:239–249.

Parzen, E. (1962). On estimation of a probability density function and mode. *Ann. Math. Statist.*, 33:105–131.

Prakasa Rao, B. (1983). *Nonparametric functional estimation*. Probability and mathematical statistics. New York: Academic Press.

Prékopa, A. (1973). On logarithmic concave measures and functions. *Acta. Sci. Math. (Szeged)*, 34:335–343.

Rockafellar, R. T. (1970). *Convex analysis*. Number 28 in Princeton Mathematical Series. Princeton Univ. Press, Princeton, N.J.

Rosenthal, J. S. (2006). *A first look at rigorous probability theory*. World Scientific, Singapore, 2nd edition.

Rufibach, K. (2007). Computing maximum likelihood estimators of a log-concave density function. *J. Statist. Comput. Simul.*, 77:561–574.

Rufibach, K. and Dümbgen, L. (2006). *logcondens: Estimate a log-concave probability density from i.i.d observations*. URL http://CRAN.R-project.org/package=logcondens.

Saunders, M, A. (2003). `PDCO`: A primal-dual interior solver for convex optimization. Available at http://www.stanford.edu/group/SOL/software/pdco.html.

Schuhmacher, D. and Dümbgen, L. (2010). Consistency of multivariate log-concave density estimators. *Statist. Probab. Lett.*, 80:376–380.

Seregin, A. and Wellner, J. A. (2010). Nonparametric estimation of multivariate convex-transformed densities. *Ann. Statist.*, 38:3751–3781.

Shor, N. Z. (1985). *Minimization methods for non-differentiable functions.* Berlin: Springer-Verlag.

Silverman, B. W. (1982). On the estimation of a probability density function by the maximum penalized likelihood method. *Ann. Statist.*, 10:798–810.

Silverman, B. W. (1986). *Density Estimation for Statistics and Data Analysis.* Chapman and Hall, London.

Van der Vaart, A. W. (1998). *Asymptotic statistics.* Cambridge series in statistical and probablistic mathematics. Cambridge Univ. Press.

Van Houwelingen, J. and Stijnen, T. (1983). Monotone empirical Bayes estimator for the continuous one-parameter exponential family. *Statist. Neerlandica*, 37:29–43.

Walther, G. (2002). Detecting the presence of mixing with multiscale maximum likelihood. *J. Amer. Statist. Assoc.*, 97:508–513.