# University of Alberta

# Nonparametric density estimation via regularization

By

**Mu Lin**

A thesis submitted in partial fulfillment of the

requirements for the degree of

Master of Science

In Statistics

Department of Mathematical and Statistical Sciences



Fall, 2009

Edmonton, Alberta

# Examining Committee

Ivan Mizera, Department of Mathematical and Statistical Sciences.
(Supervisor)

Rohana Karunamuni, Department of Mathematical and Statistical Sciences.
(Chair)

Pengfei Li, Department of Mathematical and Statistical Sciences.

Oy Leuangthong, Department of Civil and Environmental Engineering.

# Abstract

The thesis aims at showing some important methods, theories and applications about non-parametric density estimation via regularization in univariate setting.

It gives a brief introduction to non-parametric density estimation, and discuss several well-known methods, for example, histogram and kernel methods. Regularized methods with penalization and shape constraints are the focus of the thesis. Maximum entropy density estimation is introduced and the relationship between taut string and maximum entropy density estimation is explored. Furthermore, the dual and primal theories are discussed and some theoretical proofs corresponding to quasi-concave density estimation are presented. Different the numerical methods of non-parametric density estimation with regularization are classified and compared. Finally, a real data experiment will also be discussed in the last part of the thesis.

# Acknowledgements

I would like to express my sincere appreciation to my supervisor Dr. Ivan Mizera, who guided me to finish this thesis and provided many useful and helpful suggestions.

I am truly grateful to the Department of Mathematical and Statistical Sciences, University of Alberta for its support through my master program.

I also would like to thank my parents. They always support me and give me the courage to overcome the difficulties.

# List of Tables

# List of Figures

# List of Abbreviations

| | |
|---|---|
| $\|f\|$ | $[\int f^2(x)dx]^{1/2},\ f \in L_2$ |
| $f'$ | the first derivative of $f(x)$ |
| $f''$ | the second derivative of $f(x)$ |
| $f^{(k)}(x)$ | the $k$th derivative of $f(x), k > 2$ |
| $< f,\ g >$ | $\int f(x)g(x)dx$ |
| $O_p(1)$ | stochastic order symbols |
| sgn($x$) | sign function of x |
| Leb($x$) | Lebesgue measure of x |
| $\preceq,\ \succeq$ | componentwise inequality |

# Introduction

In Chapter One, the thesis introduces various well investigated and studied non-parametric density estimation methods. We carry out the numerical experiments of all the methods with the geyser data from R source. Specifically, we show the numerical results of histogram, kernel methods and regularization methods with penalization on the same data set.

In Chapter Two, we discuss the maximum entropy density estimation in the univariate setting. Some preliminary functional analysis notions are introduced first. Finally, we adapt all the notations and definitions in univariate case to derive the taut string theory. Detailed discussions and example about the taut string are given; Theorems 2.3.1 and 2.3.2 are proved. Moreover, some special important cases of maximum entropy densities are investigated in the univariate framework.

Chapter Three involves most of the theoretical work in this thesis. We apply primal and dual theories on density estimation and introduce several estimators based on the Theorems 3.2.1 and 3.2.2. In section 3.3 and 3.4, we give the proofs of existence of the solution and Fisher consistency independent of that of Koenker and Mizera (2008a). Finally, we verify the strong dual relationship for a special two data points case by using the primal and dual theory.

Detailed numerical methods are discussed in Chapter Four. We apply the different shape constraint methods to Bright Star velocity data. The comparisons of different methods are shown as well.

# Contents

# Chapter One: Nonparametric Density Estimation

## 1.1 Introduction to Density Estimation

The *probability density function* is a fundamental concept in statistics. Consider any random quantity $X$ that has probability density $f$. Function $f$ gives a natural description of the distribution of $X$, and allows probabilities associated with $X$ to be found from the relation

$$P(a < x < b) = \int_a^b f(x)dx \quad \text{for all}\, a < b \tag{1.1}$$

Typically, we consider a sample $X_1, X_2, \ldots, X_n$ of independent, identically distributed (iid) random variables with common density $f$, and density estimation seeks to estimate $f$ from the data.

To illustrate what we have in mind, let $f_0$ be a univariate probability density function (pdf), with corresponding cumulative distribution function (cdf) $F_0$. The density estimation problem aims at estimating $f_0$, where the data, $X_1, X_2, \ldots, X_n$ are independent, identically distributed univariate random variables, with common density $f_0$, and $n$ is the sample size. It is customary to encode the data $X_1, X_2, \ldots, X_n$ in the empirical distribution function $F_n$, which is defined as

$$F_n(x) = \frac{\#\{X_i : X_i \leq x\}}{n}, \quad -\infty < x < \infty, \tag{1.2}$$

the fraction of the observations not exceeding $x$. What is lost by doing so is the order in which the observations occurred, but by assuming iid observations, this is irrelevant, since $F_n$ is a sufficient statistics for $F_0$.

One approach to density estimation is *parametric*. Assume that the data are drawn

from one of the known parametric families of distributions, for example the normal distribution with mean $\mu$ and variance $\sigma^2$. The density $f_0$ underlying the data could then be estimated by finding estimates of $\mu$ and $\sigma^2$ from the data and substituting these estimates into the formula for the normal density.

In this thesis, we shall not consider parametric estimates of this kind. The approach discussed here is *nonparametric*, which means that less rigid assumptions will be made about the distribution of the observed data. Although we will still assume that the distribution has a probability $f_0$, the data will be allowed to speak for themselves in determining the estimate of $f_0$ more than would be the case if $f_0$ was constrained to belong to a given parametric family.

A very natural use of density estimates is in the informal investigation of the properties of a given set of data. Density estimates can give valuable indication of such features as skewness and multimodality in the data. An important aspect of statistics, often neglected nowdays, is the presentation of data back to the client in order to provide explanation and illustration of the conclusions that may possibly have been obtained by other means. Density estimates are ideal for this purpose, for the simple reason that they are fairly easily comprehensible to non-mathematicians.

## 1.2 Histogram

First of all, we review the oldest and most widely used density estimator, histogram. Given an origin $x_0$ and a *bin width* $h$, we define the *bins* of the histogram to be the intervals $[x_0 + mh, x_0 + (m+1)h)$ for positive and negative integers $m$. The intervals have been chosen closed in the left and open on the right for definiteness.

The histogram is then defined by

$$\hat{f}(x) = \frac{1}{nh}(\#\text{ of } X_i \text{ in the same bin as } x). \tag{1.3}$$

Note that, to construct the histogram, we have to choose both an origin and a bin

width. It is the choice of bin width which, primarily, controls the amount of smoothing in the procedure.

The histogram can be generalized by allowing the bin widths to vary. Formally, suppose that we have any dissection of the real line into bins; then the estimate will be defined by

$$\hat{f}(x) = \frac{1}{n}\frac{(\#\text{ of } X_i \text{ in the same bin as } x)}{\text{width of bin containing } x}. \tag{1.4}$$

The dissection into bins can either be carried out a priori or else in some ways which depend on the observations themselves. We show an example of histogram in Figure 1 from R example by using the function "hist()"; the data set we use in this example is Old Faithful Geyser Data in R, the data we use represents the number of duration period for 299 observations.

## 1.3 Kernel Method

It is worthy to mention that the histogram estimator suffers a extreme drawback: it is discontinuous. When density estimates are needed as intermediate components of other methods, the case for using alternatives to histograms is quite strong. Meanwhile, histogram is quite subjective and depends on the choice of bin width and the origin. Although the histogram remains an excellent tool for data presentation, it is worth at least considering the various alternative density estimates.

From the definition of a probability density, if the random variable X has density *f*, then

$$f(X) = \lim_{h\to 0}\frac{1}{2h}P(x-h < X < x+h). \tag{1.5}$$

For any given $h$, we can, of course, estimate $P(x-h < X < x+h)$ by the proportion of the sample falling in the interval $(x-h, x+h)$. Thus a natural estimator $\hat{f}$ of the

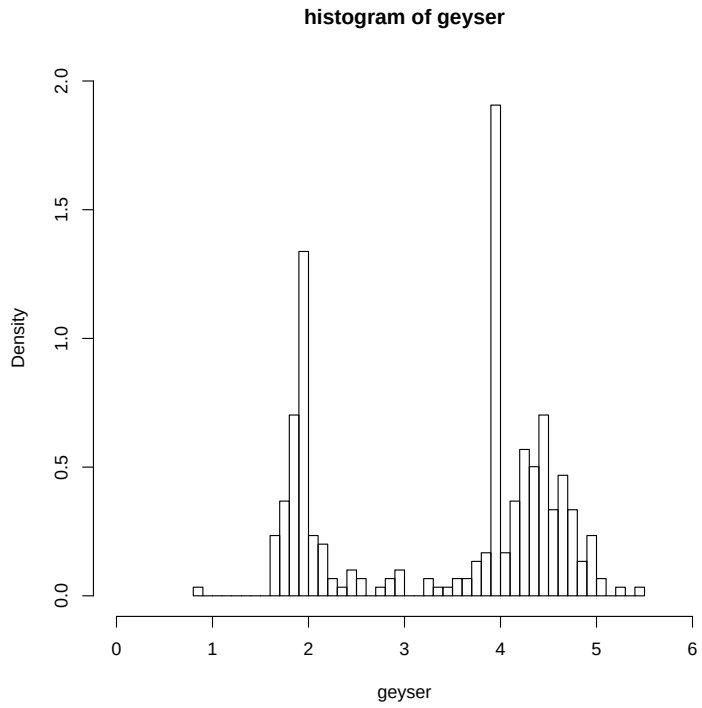


Figure 1: histogram of geyser data

density is given by choosing a small number $h$ and setting,

$$\hat{f}(x) = \frac{1}{2nh}[\#\text{ of } X_1, X_2, \ldots, X_n \text{ falling in } (x-h, x+h)], \tag{1.6}$$

we shall call (1.6) the naive estimator.

To express the estimator more transparently, define the weight function $W$ by,

$$W(x) = \begin{cases} \frac{1}{2} & \text{if } |x| < 1; \\ 0 & \text{otherwise.} \end{cases} \tag{1.7}$$

Then it is easy to see that the naive estimator can be written as,

$$\hat{f}(x) = \frac{1}{n}\sum_{i=1}^{n} \frac{1}{h} W(\frac{x - X_i}{h}). \tag{1.8}$$

It follows from (1.7) that the estimate is constructed by placing a "box" of width $2h$ and height $(2nh)^2$ on each observation and then summing to obtain the estimate.

The naive estimator is not wholly satisfactory from the point of view of using density estimates for presentation. It follows from the definition that $\hat{f}$ is not continuous function, has jumps at the points $X_i$ and zero derivative everywhere else. This gives the estimates a somewhat ragged character which is not only aesthetically undesirable, but, more seriously, could provide the untrained observer with a misleading impression.

It is easy to generalize the naive estimator to overcome some of the difficulties discussed above. Replace the weight $W$ by a kernel function $K$ which satisfies the condition

$$\int_{-\infty}^{+\infty} K(x)dx = 1. \tag{1.9}$$

Usually, but not always, $K$ will be a symmetric probability density function, the normal density, for instance, or the weight function $W$ used in the definition of the naive estimator. By analogy with the definition of the naive estimator, the kernel estimator with kernel $K$ is defined by

$$\hat{f}(x) = \frac{1}{nh}\sum_{i=1}^{n} K(\frac{x - X_i}{h}), \tag{1.10}$$

where $h$ is window width, also called the *smoothing parameter* or *bandwidth*.

Just as the naive estimator can be considered as a sum of 'boxes' centered at the observations, the kernel estimator is a sum of 'bumps' placed at the observations. The kernel function $K$ determines the shape of the bumps while the window width $h$ determines their width. Some elementary properties of kernel estimates follow at once from the definition. Provided the kernel $K$ is everywhere non-negative and satisfies the condition (1.9), in other words is a probability density function, it will follow at once from the definition that $\hat{f}$ will itself be a probability density. Furthermore, $\hat{f}$ will inherit all the continuity and differentiability properties of the kernel $K$, so that if,

for example, $K$ is the normal density function, then $\hat{f}$ will be a smooth curve having derivatives of all orders.

Table 1 lists some useful kernel functions from Silverman (1986), the definitions of efficiency of kernels densities can be found in Silverman. (1986 Chapter 3)

Table 1: some kernels and the efficiency

| *Kernel* | *K(t)* | Efficiency(Exact and to 4 *d.p*) |
|---|---|---|
| Epanechnikov | $\frac{3}{4}(1-\frac{1}{5}t^2)/\sqrt{5}$ for $\lvert t\rvert < \sqrt{5}$<br>0 otherwise | 1 |
| Biweight | $\frac{15}{16}(1-t^2)^2$ for $\lvert t\rvert < 1$<br>0 otherwise | $(\frac{3087}{3125})^{1/2} \approx 0.9939$ |
| Triangular | $1-\lvert t\rvert$ for $\lvert t\rvert < 1$<br>0 otherwise | $(\frac{243}{250})^{1/2} \approx 0.9859$ |
| Gaussian | $\frac{1}{\sqrt{2\pi}}\ \mathrm{e}^{-(1/2)t^2}$ | $(\frac{36\pi}{125})^{1/2} \approx 0.9512$ |
| Rectangular | $\frac{1}{2}$ for $\lvert t\rvert < 1$<br>0 otherwise | $(\frac{108}{125})^{1/2} \approx 0.9295$ |

We also plot the kernel estimates of the geyser data in Figure 2 with Gaussian kernel. We will illustrate the selection of smoothing parameter in penalized methods in next section.

There are many consistency results for kernel estimates. Consistency of the estimate $f$ at a single point $x$ was studied by Parzen (1962). His assumptions on the kernel $K$ were that $K$ was a bounded Borel function, satisfying,

$$\int |K(t)|dt < \infty, \int K(t)dt = 1, \tag{1.11}$$

and,

$$|tK(t)| \to 0 \ \text{ as } \ |t| \to 0. \tag{1.12}$$

Furthermore, the window width was assumed to satisfy

$$h_n \to 0 \quad \text{and } \ nh_n \to \infty \quad \text{as } n \to \infty. \tag{1.13}$$

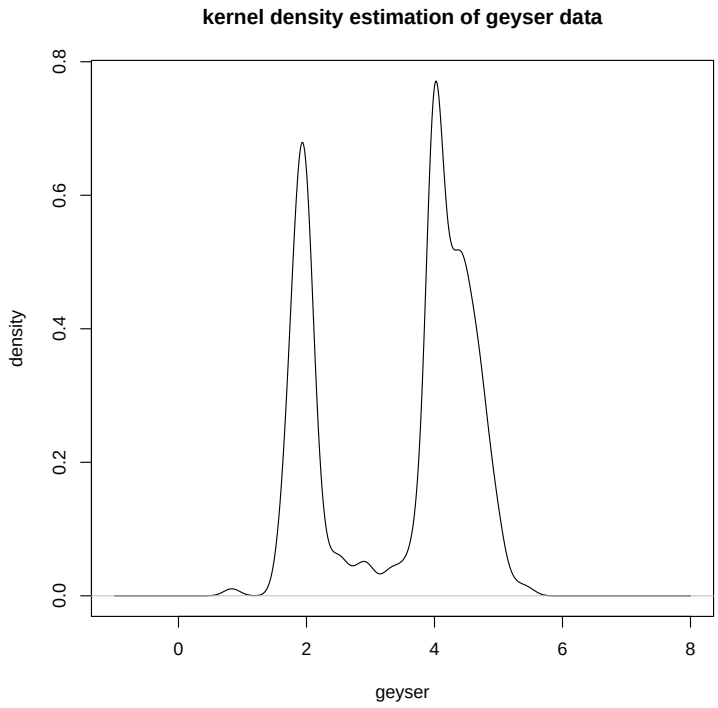


Figure 2: Gaussian kernel estimation of geyser data, with $h = 0.12762$

Under the conditions (1.11)-(1.13) and provided $f$ is continuous at $x$, $\hat{f}(x) \to f(x)$ in probability as $n \to \infty$.

Uniform consistency was considered by several authors, for example, Parzen (1962), Nadaraya (1965), Silverman (1978) and Bertrand-Retail (1978).

Suppose the kernel $K$ is bounded, has the variation and satisfies (1.11), and that the set of discontinuities of $K$ has Lebesgue measure zero. These conditions are satisfied by virtually any conceivable kernel. Suppose that,

$$f \text{ is uniformly continuous on } (-\infty, +\infty) \tag{1.14}$$

and the the window width satisfies,

$$h_n \to 0 \text{ and } nh_n(\log n)^{-1} \to \infty \text{ as } n \to \infty. \tag{1.15}$$

Bertrand-Retail showed, with probability one,

$$\sup_x |\hat{f}(x) - f(x)| \to 0 \text{ as } n \to \infty. \tag{1.16}$$

And conditions (1.14), (1.15) are necessary as well as sufficient for uniform consistency.

Apart from the histogram, the kernel estimator, which depends on the choice of kernel function, is probably the most commonly used estimator and is certainly most studied mathematically. It does, however, suffer from a slight drawback when applied to data from long-tailed distributions. Because the window is fixed across the entire sample, there is a tendency for spurious noise to appear in the tails of the estimates. If the estimates are smoothed sufficiently to deal with this, then essential detail in the main part of the distribution is masked. In order to deal with this difficulty, various adaptive methods have been proposed; their survey can be found in Silverman (1986).

Another problem regarding the kernel method is to choose the smoothing parameter properly. It is crucial for the kernel method. There are many ways to determine how to choose the smoothing parameter: for example, subjective choice, least-square cross-validation, discretization errors in cross-validation and likelihood cross-validation. See Silverman (1986) for details.

## 1.4 Penalized Approach

We already know that the kernel method depends on the choice of kernel and is sensitive at the boundary when the density is on a bounded domain like $(0, +\infty)$. Regularization methods with penalization introduced in this section could overcome these difficulties to some extent.

### 1.4.1 Introduction to Maximum Penalized Likelihood Approach

The likelihood of a density $f$ with a set of independent identically distributed observations $X_1, X_2, \ldots, X_n$ is given by

$$L(f|X_1, X_2, \ldots, X_n) = \prod_{i=1}^{n} f(X_i), \tag{1.17}$$

provided that $f$ is a continuous pdf.

However, the maximum likelihood problem has no solution. Loosely speaking, the solution of maximum likelihood problem would be a sum of point masses at the observations,

$$f^n(x) = \frac{1}{n}\sum_{i=1}^{n} \delta(x - X_i), \quad -\infty < x < +\infty, \tag{1.18}$$

where $\delta(x)$ is the unit mass at 0; but this is not a density.

So we cannot use the maximum likelihood directly for density estimation. There is the chance that we could solve this problem by adding restrictions on the density over which the likelihood is to be maximized.

One method is to add a term into the likelihood which describes the roughness, in some sense, of the curve under consideration. Assuming $R(f)$ is a function which quantifies the roughness of $f$, one choice of $R$ is,

$$R(f) = \int_{-\infty}^{+\infty} [f''(t)]^2 dt. \tag{1.19}$$

The penalized log likelihood is defined by,

$$l_\lambda(f) = \sum_{i=1}^{n} \log f(X_i) - \lambda R(f), \tag{1.20}$$

where $\lambda$ is a positive smoothing parameter.

The probability density function $\hat{f}$ is said to be a maximum penalized likelihood

density estimate if it maximizes (1.20) and is subject to the constraints,

$$\int_{-\infty}^{+\infty} f = 1, \quad f(x) \geq 0 \text{ for all } x, \text{ and } R(f) < \infty. \tag{1.21}$$

The penalized log likelihood can be seen as a way of quantifying the conflict between smoothness and goodness-of-fit to the data, because the log likelihood term measures how well $f$ fits the data, while the penalized term is to avoid curves which exhibit too much roughness or rapid variation. The choice of the smoothing parameter controls the balance between smoothness and goodness-of-fit, while the choice of the roughness penalty determines exactly what kind of behavior in the density estimate is considered to be undesirable in excess. For example, the choice $R(f(x))=\int_{-\infty}^{+\infty}(f(x)'')^2dx$ will have a large value if $f$ exhibits a large amount of local curvature, and equals to zero if $f$ is a straight line.

Unfortunately, the implicit nature of the definition of the estimate $\hat{f}$, as the solution of a maximum problem, is the price to be paid for the explicit statement of the aims of the estimation.

### 1.4.2 $L_2$ Penalized Methods

In this section, we discuss several important penalized methods with respect to $L_2$ penalty. The first authors to apply the penalized likelihood approach to density estimation were Good and Gaskins (1971). They based their roughness penalty on the square root, $\gamma = \sqrt{f}$, of the density. The choice has the added practical advantage that it permits the optimization to be formulated as a convex problem with the (squared) $L_2$ penalty,

$$R(f) = \int [\gamma'(t)]^2 dt. \tag{1.22}$$

Another advantage of working with $\gamma$ rather than $f$ is that the constraint $f(x) \geq 0$ will automatically be satisfied if $\gamma$ is real. Furthermore, the constraint $\int f = 1$ will be replaced by $\int \gamma^2 = 1$, an easier constraint under the numerical method used by Good

and Gaskins (1971), which was described briefly in their paper.

We apply Good and Gaskin's method to the geyser data. The estimated density is plotted in Figure 3. We do automatic $\lambda$ selection based on Koenker and Mizera (2006b). The Kolmogorov distance between the empirical distribution function of the sample, $\hat{F}_n$, and the smoothed empirical, $\tilde{F}_{n,\lambda}$, corresponding to the density estimate

$$\kappa(\lambda) \equiv K(\hat{F}_n, \tilde{F}_{n,\lambda}) = \max |\hat{F}_n(X_i) - \tilde{F}_{n,\lambda}(X_i)| \tag{1.23}$$

is computed. Then we find $\lambda$ which makes $\kappa(\lambda)$ approximately equal to the cutoff $c_\kappa/\sqrt{n}$. In our computation, we used $c_\kappa = 1.4801$, the 0.975 quantile of the Kolmogorov distribution. There was no particular reason for this choice. If we chose 0.95 quantile, some estimates are too rough; on the other hand, if we chose 0.99 quantile, some estimates are too smooth. The most important point here is that all estimates have about the same degree of fit by choosing $c_\kappa = 1.4801$.

From this point of view, we can illustrate the choice of smoothing parameter in kernel density estimation in the previous section. Since we use the geyser data all through this chapter, we want to make the kernel density estimation result to be comparable with all the penalized estimations. Therefore, we choose the kernel smoothing parameter according to (1.23), which leads the smoothing parameter equal to 0.12762. However, we want to mention once again that the $\lambda$ selection strategy we discussed is not necessarily the optimal selection procedure; we just want estimates to be comparable with each other.

The penalty (1.22) penalized for slope rather than curvature in the estimates. In the substantial work of Good and Gaskins, they found the penalty (1.22) somewhat unsatisfactory, producing estimates that sometimes "too rough". This can also be seen from our example in Figure 3. In order to penalize for curvature, Good and Gaskins (1971) proposed using the penalty

$$R(f) = \int [\gamma''(t)]^2 dt \tag{1.24}$$

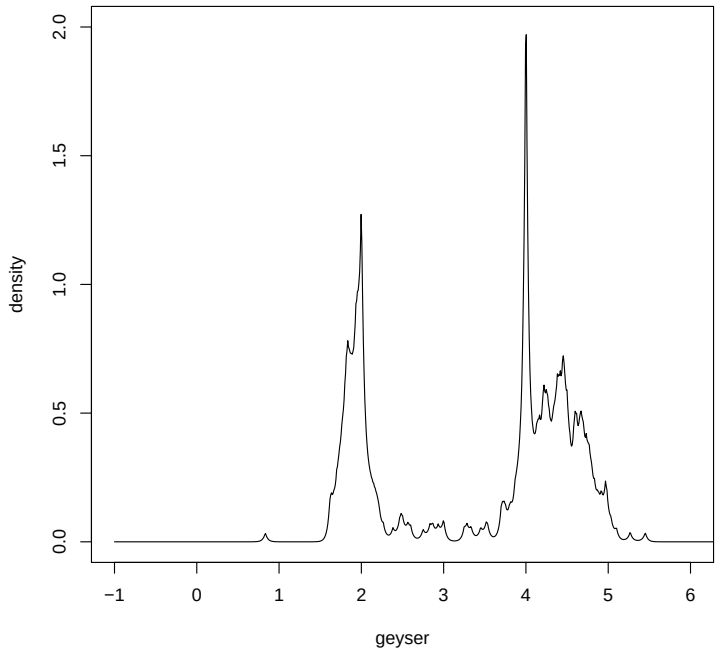


Figure 3: the estimated density by Good's method

or even a linear combination of (1.22) and (1.24). This component has a more direct interpretation as a measure of roughness of the fitted density.

Finally, it should be pointed out that the use of the roughness penalty (1.24) leads to some technical difficulties concerning the uniqueness and definition of the estimates. These are discussed by de Montricher, Tapia and Thompson (1975).

There are certain potential advantages in using a roughness penalty based on the logarithm of the density, as proposed by Silverman (1982). Consider the roughness penalty,

$$R(f) = \int [(\log f(t))^{(3)}]^2 dt, \tag{1.25}$$

when expressed in terms of the logarithm of $f$, the problem of finding the maximum

penalized likelihood density estimate becomes, setting $g(x) = \log f(x)$,

$$\sum g(X_i) - \lambda \int (g^{(3)}(t))^2 dt \to \max \tag{1.26}$$

subject to

$$\int e^{g(t)} dt = 1. \tag{1.27}$$

Working with the logarithm of $f$ has the advantage of eliminating the necessity for a positivity constraint on $f$ and reducing the quantity to be maximized in (1.26) to a quadratic form. The cost to be paid is the awkward nonlinear nature of the constraint (1.27).

The rather strange-looking roughness penalty (1.25) has the important property that it is zero if and only if $f$ is a normal density. Thus normal densities are considered by the method to be 'infinitely smooth' because they are not penalized at all in (1.26) and so cost nothing in terms of roughness. It can be shown, in a sense made clear in Silverman (1982, Theorem 2.1), that as the smoothing parameter $\lambda$ tends to infinity, the limiting estimate will be the normal density with the same mean and variance as the data. As $\lambda$ varies, the method will give a range of estimates from the 'infinitely rough' sum of delta function at the data points to the 'infinitely smooth' maximum likelihood normal fit to the data. Since one of the objects of nonparametric methods is to investigate the effect of relaxing parametric assumptions, it seems sensible that the limiting case of a nonparametric density estimate should be a natural parametric estimate.

We also plot the estimated geyser density by using Silverman's approach in Figure 4.

Gu (2002) adopted a similar point of view of penalizing the logarithm of the density. However, he replaced the third derivative by the second derivative of the logarithm of density. The resulting constrained formulation is

$$\sum_{i=1}^{n} \log g(X_i) - \lambda \int (g''(x))^2 dx \to \max \quad \text{subject to } \int e^{g(x)} dx = 1. \tag{1.28}$$

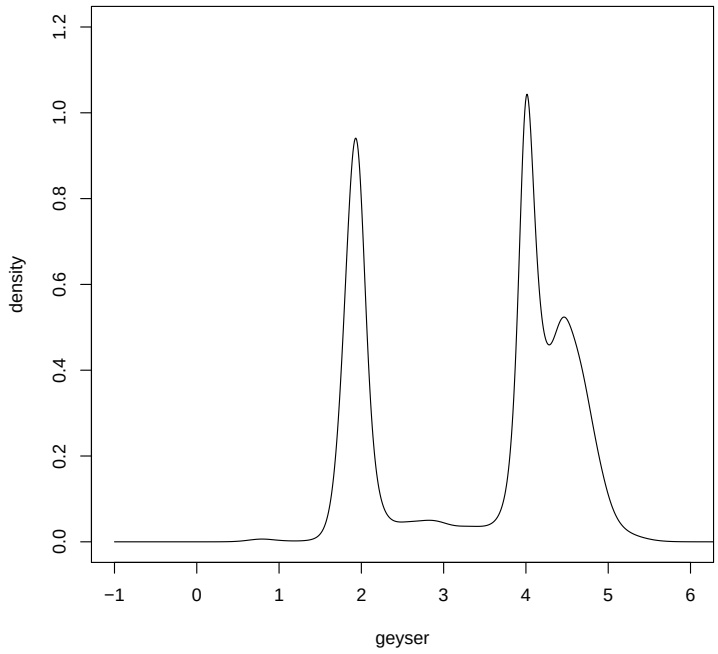


Figure 4: the estimated density by Silverman's method

Comparing to Silverman's method, Gu's penalized approach give the similar result for geyser density as we could see in Figure 5.

It is possible to define other roughness penalties according to other perceptions of 'infinitely smooth' exponential families of densities. The key property is that *R(f)* should be zero if and only if $f$ is in the required family. For example, when working on the half-line $(0, \infty)$ a natural penalty is $\int_0^\infty [(\log f)'']^2$, which gives zero roughness to the exponential densities $\lambda e^{-\lambda x}$.

Detailed numerical work on the density estimation of this section can be found in Ramsay and Silverman (2005). It can be shown (see Theorem 3.1 of Silverman, 1982) that the maximum of (1.26) subject to the constraint (1.27) can be found as the

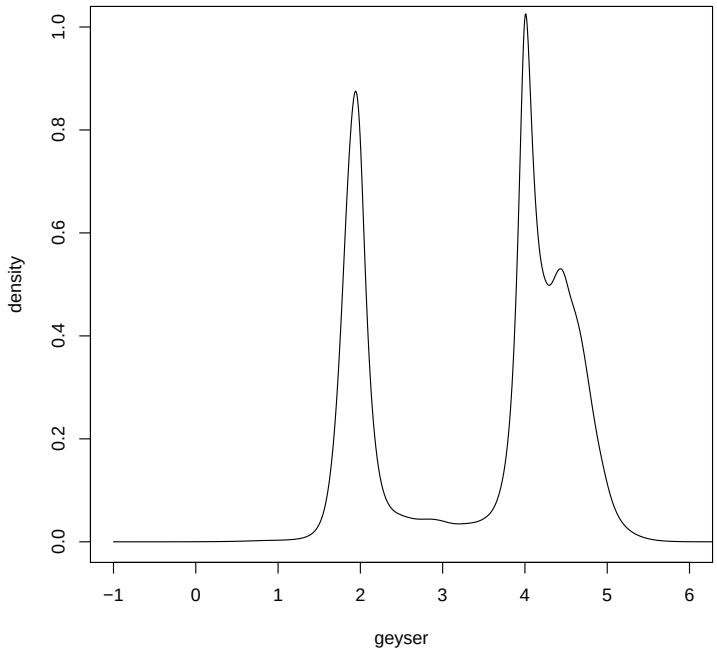


Figure 5: the estimated density by Gu's method

unconstrained maximum of the strictly concave functional

$$\frac{1}{n}\sum g(X_i) - \lambda \int [g^{(3)}(t)]^2 dt - \int e^{g(t)} dt. \tag{1.29}$$

The form (1.29) is remarkable in that it contains no unknown Lagrange multipliers to be determined; its maximum will automatically satisfy the constraint (1.27). The fact that the estimates can be found as the unconstrained maximum of a concave function also makes it possible to derive various theoretical properties of the estimates; see Silverman (1982).

### 1.4.3 $L_1$ Penalized Methods

In this section, we discuss $L_1$ penalized methods. In $L_1$ 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 $L_2$ framework since the penalty evaluates them as 'infinitely rough'. Total variation penalties are happy to tolerate such jumps and bends, and they are therefore better suited to identifying discrete jumps in densities or in their derivatives. This is precisely the property that has made them attractive in imaging applications.

Specifically, given a random sample, $X_1, X_2, \ldots, X_n$ from a density $f_0$, we consider estimators that solve,

$$\max_f \{\sum_{i=1}^{n} \log f(X_i) - \lambda R(f)| \int_\Omega f = 1\}, \tag{1.30}$$

where $R$ denotes a function intended to penalize for the roughness of candidate estimates, and $\lambda$ is the tuning parameter controlling the smoothness of the estimate. Here the domain $\Omega$ may depend on a priori considerations as well as on the observed data.

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

$$\bigvee_\Omega(f) = \sup \sum_{i=1}^{m} |f(u_i) - f(u_{i-1})|, \tag{1.31}$$

where the supremum is taken over all partitions, $u_1, u_2, \ldots, u_m$ of $\Omega$. When $f$ is absolutely continuous, we can write, see e.g. Natanson (1974, P.259),

$$\bigvee_\Omega(f) = \int_\Omega |f'(x)|dx. \tag{1.32}$$

Usually, we will focus on penalizing the total variation of the first derivative of the log density,

$$J(f) = \bigvee_{\Omega}((\log f)') = \int_{\Omega} |(\log f(t))''|dt, \tag{1.33}$$

so letting $g = \log f$ we can write (1.30) as,

$$\max_{g}\{\sum_{i=1}^{n} g(X_i) - \lambda \bigvee_{\Omega}(g')| \int_{\Omega} e^g = 1\}. \tag{1.34}$$

In Figure 6, we can see the estimated geyser data by penalizing the second derivative of $\log f$.

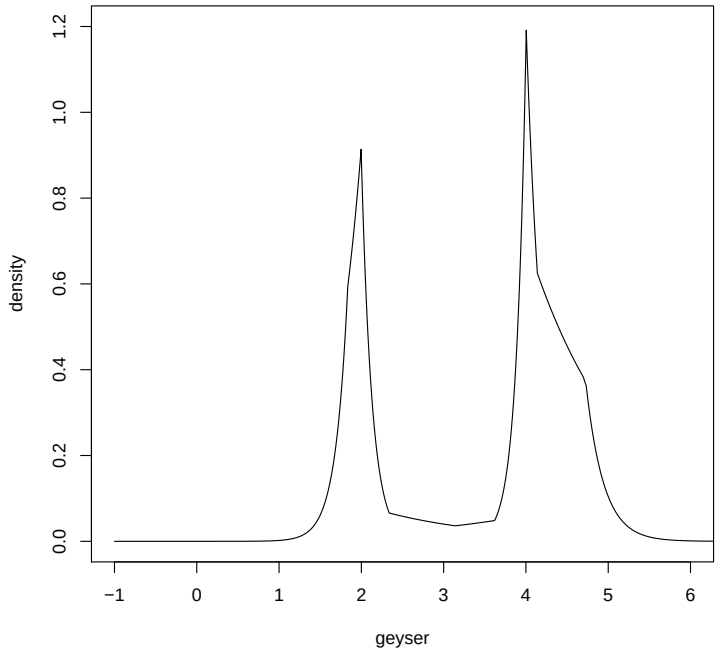


Figure 6: the estimated density by Total variation method penalizing second derivative

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

$$J(f) = \bigvee_{\Omega}(g^{(k)}), \tag{1.35}$$

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

For example, we plot the estimated geyser density by penalizing the second derivative of $\log f$ in Figure 7.

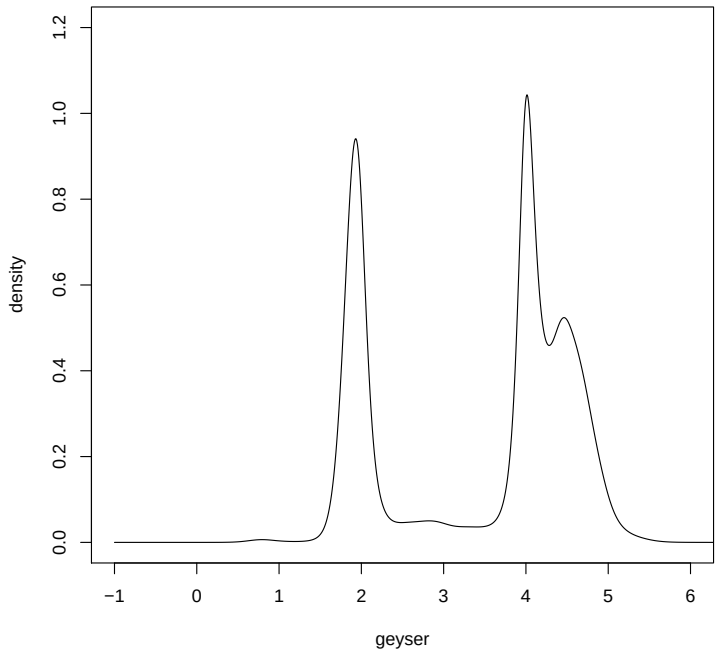


Figure 7: the estimated density by Total variation method penalizing third derivative

We know that even for $L_2$ formulations the presence of the integrability constraint prevents the usual reproducing kernel strategy from finding exact solutions; so iterative algorithm is needed. Koenker and Mizera (2006a) 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 parameterization of the problem in terms of $\log f$ is that it obviates any worries about the non-negativity of $\hat{f}$. However we have still neglected one crucial constraint. We 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 $\alpha's$,

$$\sum_{j=1}^{m} h_j \frac{e^{\alpha_i} - e^{\alpha_{i-1}}}{\alpha_i - \alpha_{i-1}} = 1. \tag{1.36}$$

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 Constraints Approach

Although the penalized approaches successfully overcome some difficulties in density estimation compared to kernel methods, the approaches indeed depend on the choice of the smoothing parameter $\lambda$.

Once we are quite sure about the shape of the data, especially when it is unimodal, we can apply the shape constraint methods which do not have to select $\lambda$. However, the price we need to pay is to assume certain shape constraint.

The maximum penalized likelihood approach to nonparametric density estimation lends itself well to the use of such a side information. Examples of interest include shape constraints on the unknown density $f_0$, such as monotonicity, convexity, unimodality, and log-concavity. It seems logical to require that the estimator satisfies the same constraints as $f_0$. Here, we study monotone and log-concave density estimation.

A univariate density $f$ is monotone on $(a, \infty)$ if it is decreasing there, i.e, $f(x) \leq f(y)$ for all $x \geq y > a$. Likewise, a univariate density is monotone on $(-\infty, a)$ if it is increasing there.

A univariate density is unimodal if there exists an real number (the mode), such that $f$ is monotone on $(-\infty, m)$ and on $(m, \infty)$.

Density estimation under shape constraints was first considered by Grenander (1956), who found that the nonparametric maximum likelihood estimator $\hat{f}_n^{mon}$ of a non-increasing density function $f_0$ on $[0, \infty)$ is given by the left derivative of the least concave majorant of the empirical cumulative distribution function on $[0, \infty)$. This work was continued by Prakasa Rao (1969) and Groeneboom (1985, 1988), who established asymptotic distribution theory for $n^{1/3}(f_0 - \hat{f}_n^{mon})(t)$ at a fixed point $t > 0$ under certain regularity conditions and analyzed the non-gaussian limit distribution.

Let $D$ denote the set of decreasing densities on $(0, \infty)$. Actually, we do not enforce the pdf constraint, so we define $D$ as the set $\{f^1(0, \infty) : f \geq 0, f \text{ decreasing}\}$.

Suppose $X_1, X_2 \dots X_n$ be nonnegative iid random variables with common density $f_0$, which is assumed to be decreasing on $(0, \infty)$. The unpenalized maximum likelihood estimation problem is

$$\text{minimize} \quad L_n(f) = -\frac{1}{n}\sum_{i=1}^{n} \log f(X_i) + \int_0^\infty f(y)dy, \text{ subject to } f \in D. \tag{1.37}$$

Without the restriction of the monotonicity, we know that the maximum likelihood estimator does not exist. We claim that with this monotone shape restriction, the solution exists and it must be a pdf.

Monotone density estimation can be extended to cover unimodal densities. If the true mode is known a priori, unimodal density estimation boils down to monotone estimation in a straightforward way. However, the situation is different if mode is unknown; in that case the likelihood is unbounded, problem being caused by observations too close to a hypothetical mode (Dümbgen Hüsler and Rufibach (2007)). Even if we know the mode, the estimator of the density is not consistent at the mode, the effect is called "spiking". Some methods were proposed to remedy this problem, see Wegman (1970), Woodroofe and Sun (1993), Meyer and Woodroofe (2004) or Kulikov

and Lopuhaa (2006). However, all of them require additional constraints on $f$.

Next, we will discuss a very important type of shape constraint: log-concave. Log-concave density is defined as: A density $f$ is log-concave if $\log f$ is concave. Equivalently, if for all $0 < \theta < 1$ and for all real number $x$ and $y$, $f(\theta x + (1-\theta)y) \geq |f(x)|^{\theta}|f(y)|^{1-\theta}$.

Log-concave density plays a crucial role in a wide variety of economic models, as well as in multivariate settings. Most of the familiar parameter densities employed in economics are log-concave, for instance, the Uniform, Normal, Exponential, Logistic, Weibull, Gamma, all belong to log-concave family. However, the student $t$ family fails to be log-concave. Moreover, according to the definition of unimodal density, we know that log-concave densities are unimodal. It turns out that by imposing the log-concave shape constraint, one could overcome the spiking problem mentioned before; this yields a new approach to estimate a unimodal, possibly skewed density. Another advantage to impose log-concave shape constraint is that the density estimation procedure is fully automatic, which means that there is no need to select any binwidth, kernel function or other tuning parameters. Thus, all these properties make the new estimator appealing for its use in statistical applications.

There are many people working on the density estimation with log-concave shape constraint. For example, Dümbgen and Rufibach (2008), Pal, Woodroofe and Meyer (2006), and Koenker and Mizera (2008a).

Let $X$ be a random variable with distribution function $F$ and Lebesgue density $f(x) = \exp g(x)$ for some concave function $g : R \to [-\infty, \infty)$. The goal is to estimate $f$ based on a random sample with size $n > 1$ from $F$. Let $X_{(1)} < X_{(2)} < \cdots < X_{(n)}$ be the corresponding order statistics. For any log-concave probability density $f$ on R, the normalized log-likelihood function at $f$ is given by,

$$\int \log f dF_n = \int g dF_n, \tag{1.38}$$

where $F_n$ stands for the empirical distribution function of the sample. In order to relax

the constraint of $f$ being a probability density and get a criterion function to maximize over the convex set of all concave functions $g$, we employ the standard trick of adding a Lagrange term to (1.38), leading to the functional

$$\Psi_n(g) = \int g dF_n - \int \exp g(x) dx. \tag{1.39}$$

The non-parametric maximum likelihood estimator of $g = \log f$ is the maximizer of this functional over all concave functions,

$$\hat{g}_n = \arg\max \Psi_n(g), \tag{1.40}$$

and $\hat{f}_n$=exp $\hat{g}_n$.

In Dümbgen and Rufibach (2008), the authors established the existence, uniqueness and shape of $\hat{g}_n$. The following theorem was proved independently by Pal et al (2006) and Rufibach (2006). It follows also from more general considerations in Dümbgen at al.(2007).

**Theorem 1.5.1.** *The non-parametric maximum likelihood estimator $\hat{g}_n$ exists and is unique. It is linear on all intervals $[X_{(j)}, X_{(j+1)}], 1 \leq j < n$. Moreover, $\hat{g}_n = -\infty$ on $R \setminus [X_{(1)}, X_{(n)}]$.*

Dümbgen and Rufibach also provided two characterizations of the estimators $\hat{g}_n, \hat{f}_n$ and the corresponding distribution function $\hat{F}_n$, i.e. $\hat{F}_n(x) = \int_{-\infty}^{x} \hat{f}_n(r) dr$.

When applying this density estimation method, Dümbgen, Hüsler and Rufibach(2007) proposed an active set and EM algorithm for log-concave densities based on complete and censored data. We note that their method can be applied to censored data, e.g. right-censored or interval censored data set; thus, it could be very useful in survival analysis. We will discuss the numerical method and data experiment in Chapter Four.

As mentioned above, we know that unimodality of concave functions implies that

log-concave densities are unimodal. However, unimodal densities need not be log-concave; none of the Student $t_\nu$ densities are log-concave for $\nu < +\infty$. Laplace densities, with their exponential tail behavior, are log-concave; but heavier tails−say, algebraic−are prohibited. This prohibition motivates a relaxation of the log-concavity requirement.

A natural hierarchy of concave functions can be built on the foundation of the weighted means of order $\rho$ studied in Hardy, Littlewood, and Polya (1934),

$$M_\rho(a,p) = M_\rho(a_1,\cdots,a_n;p) = (\sum_{i=1}^{n} p_i a_i^\rho)^{1/\rho}, \tag{1.41}$$

for $p$ in the unit simplex, $\{S = p \in R^n | p \geq 0, \sum p_i = 1)\}$. The familiar arithmetic, geometric and harmonic means correspond to $\rho = 1$, $\rho = 0$ and $\rho = -1$ respectively. Following Avriel (1972), a non-negative, real function $f$, defined on a convex set $C \subset R^d$ is said to be $\rho$-concave if for any $x_0 \in C, x_1 \in C$, and $p \in S$, if

$$f(p_0 x_0 + p_1 x_1) \geq M_\rho(f(x_0), f(x_1); p); \tag{1.42}$$

in this terminology log-concave functions are 0-concave, and concave functions are 1-concave.

Since $M_\rho(a,p)$ is monotone increasing in $\rho$ for $a \geq 0$ and any $p \in S$, it follows that if $f$ is $\rho$-concave, then $f$ is also $\rho'$-concave for any $\rho' < \rho$. Thus, concave functions are log-concave, but not vice-versa. In the limit $(-\infty)$-concave functions satisfy the condition,

$$f(p_0 x_0 + p_1 x_1) \geq \min(f(x_0), f(x_1)), \tag{1.43}$$

and therefore are *quasi-concave*, a class that includes all $\rho$-concave functions.

Koenker and Mizera (2008a) studied the $\rho$-concave shape constrained problem

$$\Psi(g) = \frac{1}{n}\sum_{i=1}^{n} g(X_i) + \int \psi(g)dx = \min_{g \in C(x)} ! \quad \text{subject to } g \in K, \tag{1.44}$$

where $C(X)$ is the collection of all continuous functions on the convex hull of X, and $K$ stands for the set of all convex functions on $R^d$. With some duality and maximum entropy theorems (we will introduce these definitions in the following chapters), they derived,

**Theorem 1.5.2.** *The strong (Fenchel) dual of the primal formulation (1.44) is*

$$-\int \psi^*(-f)dy = \max_f ! \quad \textit{subject to } f = \frac{d(P_n - G)}{dy},\ G \in K^*. \tag{1.45}$$

In the sense that all $\Psi(g)$, the values of the primal object for $g$ satisfying the constraint of (1.44), dominate those of (1.45), and both problems have optimal solutions, $g$ and $f$ respectively, for which the respective values of the primal and dual objective functions coincide. Any function $f$ satisfying the constraint of (1.44) is a probability density with respect to Lebesgue measure: ($f \geq 0$ and $\int f dx = 1$). If $\psi$ is differentiable on the interior of its domain, then the dual and primal optimal solutions satisfy the relationship $f = -\psi'(g)$ and $K^*$ is the cone dual of $K$.

**Corollary 1**. Maximum likelihood estimation of a log-concave density has an equivalent dual formulation,

$$-\int f \log f dy = \max_f ! \quad \text{subject to } f = \frac{d(P_n - G)}{dy},\ G \in K^*, \tag{1.46}$$

whose solution satisfies the relationship $f = e^{-g}$, where $g$ is the solution of (1.44) and $\psi(g) = e^{-g}$. In particular, the solution of (1.44) satisfies $\int e^g = 1$.

We notice the emergence of Shannon entropy in (1.46). It is not surprising in the view of the well-established connections of maximum likelihood estimation to the Kullback-Leibler divergence and maximum entropy. To explore the link to potential

alternatives, we consider the family of entropies originally introduced for $\alpha > 0$ by Rényi (1961, 1965),

$$(1-\alpha)^{-1}\log(\int f^{\alpha}(x)dx), \qquad \alpha \neq 1, \tag{1.47}$$

as an extension of the limiting case for $\alpha = 1$, the Shannon entropy. For $\alpha \neq 1$, maximizing (1.47) is equivalent to the maximization of

$$\frac{\mathrm{sgn}(1-\alpha)}{\alpha}\int f^{\alpha} = -\mathrm{sgn}(\alpha-1)\int \frac{f^{\alpha}}{\alpha}. \tag{1.48}$$

The dependence of convexity/concavity properties of the function $y^{\alpha}$ necessitates a separate treatment of the case with $\alpha > 1$, when the conjugate pair is,

$$\psi(x) = \begin{cases} \frac{x^{\beta}}{\beta} & \text{if } X \leq 0; \\ 0 & \text{if } X > 0. \end{cases} \quad \psi^{*}(y) = \begin{cases} \frac{y^{\alpha}}{\alpha} & \text{if } y \leq 0; \\ 0 & \text{if } y > 0. \end{cases} \tag{1.49}$$

And the cases with $\alpha < 1$, where $1/\alpha + 1/\beta = 1$ is,

$$\psi(x) = \begin{cases} +\infty & \text{if } X \leq 0; \\ -\frac{x^{\beta}}{\beta} & \text{if } X > 0. \end{cases} \quad \psi^{*}(y) = \begin{cases} -\frac{(-y)^{\alpha}}{\alpha} & \text{if } y \leq 0; \\ +\infty & \text{if } y > 0. \end{cases} \tag{1.50}$$

Specifically, some of important cases for different $\alpha$ are considered in Koenker and Mizera (2008a). For example, they introduce $\alpha = 2$, corresponding to a more restrictive form of concavity that log-concave; $\alpha = 1/2$, which requires the estimated density to be only $(-1/2)$-concave, a significant relaxation of the log-concave constraint. In addition to all log-concave densities, all the student $t_{\nu}$ densities with $\nu \geq 1$ satisfy this requirement; the limiting case $\alpha = 0$, the estimate is constrained to be $(-1)$-concave,a yet still weaker requirement that admits all of the student $t_{\nu}$ densities for $\nu > 0$.

We could continue to consider $\alpha < 0$, which corresponds to weaker concavity requirements. The shape constraints corresponding to negative $\alpha$ encompass a wider and wider class of quasi-concave densities, eventually arriving at the $-\infty$-concave constraint.

Finally, the detailed numerical methods of quasi-concave density estimation will be discussed in Chapter Four.

# Chapter Two: Maximum Entropy Density Estimation

In this chapter, we discuss maximum entropy density estimation method, and also the application of the taut string method to maximum entropy density estimation.

## 2.1 Definition of Maximum Entropy Estimation

In the univariate case, we recall from the previous chapter that the Shannon entropy of a density function $p$ is defined as

$$H(p) = -\int p(x) \log p(x) dx. \tag{2.1}$$

More generally, given a probability density $p$, the $\alpha$-Rényi entropy is defined for $\alpha \neq 1$ to be

$$H_\alpha(p) = \frac{1}{1-\alpha} \log \left( \int p(x)^\alpha dx \right). \tag{2.2}$$

Note that by L'Hopital rule,

$$\lim_{\alpha \to 1} H_\alpha(p) = \lim_{\alpha \to 1} \frac{-\int p(x)^\alpha \log p(x) dx}{\int p(x)^\alpha dx} = H(p), \tag{2.3}$$

since $\frac{d}{dt} a^t = a^t \log_e a$. Some properties of the entropy function are: the entropy function is concave, differentiable on their domain containing $(0, +\infty)$, and equals $+\infty$ for any $x < 0$.

Before we discuss the maximum entropy estimation, we first need to review some functional analysis knowledge. In functional analysis, it is said that a *Radon measure* $\mu$ is representable by a function, if there is a function $f$, which is *locally integrable*

(which means $f$ is integrable on any compact set of its domain of definition), such that $\int u(x)d\mu = \int u(x)f(x)dx$ for all $u$ (continuous, or merely differentiable, but with bounded support). Again, such an $f$ is unique up to equality almost everywhere (with respect to the Lebesgue measure). In statistics, we say that $\mu$ is absolutely continuous with respect to Lebesgue measure, and write $f = d\mu/dx$ to denote this.

If $D$ is a continuous linear operator defined on $\mathcal{U}$, then $D^*$ denotes the *adjoint* of $D$, which is an operator defined on the topological dual $\mathcal{U}'$ of $\mathcal{U}$, satisfying

$$\langle D^*T, u\rangle = \langle T, Du\rangle. \tag{2.4}$$

Important instances are the distributional adjoints of standard operators. In such a case, the *adjoint* of the operator of differentiation satisfies, for any $u \in \mathcal{U}$ and distribution $T$ representable by a function $f$,

$$\langle D^*f, u\rangle = \langle f, Du\rangle = \int f(x)\frac{du}{dx}dx = -\int u(x)\frac{df}{dx}dx, \tag{2.5}$$

which makes natural to call the result of the application of $-D^*$ on any distribution $T$ the distributional derivative of $T$. Distributional differentiation is essential for proper expression of some of the definitions coming below; it is returning the underlying probability measures when applied to its cumulative distribution function.

If we consider the primal problem as

$$\frac{1}{n}\sum_{i=1}^{n} g(X_i) + \int \psi(g)dx + \lambda N(Dg) = \min_g! \tag{2.6}$$

we define maximum entropy density estimation with penalization as solution of the following variational problem

$$\int_\Omega \psi^*(P_n - D^*u) + \lambda N^*(\frac{u}{\lambda}) = \min_{u\in\mathcal{U}}! \tag{2.7}$$

where $D^*$ is the adjoint of $D$; $\psi^*$ is the conjugate to $\psi$; $N^*$ is the dual to $N$; $\lambda > 0$ and $P_n$ is the empirical probability measure supported by the datapoints.

We consider $\psi^*$ that satisfies the following: they are convex functions, differentiable on their domain containing $(0, +\infty)$, and $\psi(t) = +\infty$ for any $t < 0$. The reason why the dual is called maximum entropy density estimation with penalization is that if we consider $\psi(g) = e^g$, then $\psi^*(x) = x \log x$; this give the Shannon entropy in (2.7). If we consider $\psi(g) = \frac{1}{\beta} g^\beta$, then $\psi^*(x) = \frac{1}{\alpha} x^\alpha$; it gives general $\alpha$ Rényi entropy, where $1/\alpha + 1/\beta = 1$. Thus, we can generally consider $\psi^*$ as the entropy function.

The operator $D^*$ is a distributional differential operator in typical penalty formulations. We limit our scope to the standard $L^p$ choices for penalized term $N^*$, for $p \geq 1$: we assume that $N^*(u) = \|u\|_{L^p}^p / p$, for $p < \infty$, or $N^*(u)$ is the indicator in $L^\infty$ norm.

To explain things in detail, we say that $P_n - D^* u \in \mathcal{M}(R)$, where $\mathcal{M}(R)$ is finite signed Radon measure on $R$. We require $D^* u \in \mathcal{M}(R)$ as well. For convex function $\psi^*$, we adapt the convention that $\int \psi^*(Q) dx = +\infty$ if $Q$ is not absolutely continuous with respect to Lebesgue measure; and $\int \psi^*(Q) = \int \psi^*(\frac{dQ}{dx}) dx$ if $Q$ is absolutely continuous.

We also need to define what is the set $\mathcal{U}$ here. In this univariate case, we denote $\mathcal{U}$ as set of functions with bounded variation on $R$. And our fitted objects $u$ belongs to $\mathcal{U}$.

Thus, we use the notation $\mathcal{U} = BV(R)$, where $BV$ stands for bounded variation. By Lebesgue decomposition theorem, we conclude that $u$ can be decomposed as three parts: $u = u_{disc} + u_{abs} + u_{sing}$, the discrete, absolutely continuous and singular parts. In the thesis, we have a very important assumption that there is no singular part $u_{sing}$ in the decomposition. Therefore, we can simplify the decomposition as $u = u_{disc} + u_{abs}$, and use $SBV(R)$ to denote sets without the singular part.

Because we employ the entropy function as $\psi^*$ in (2.7), we indeed seek the minimization of the objective function (2.7) instead of maximization. Under this convention, we still name the problem as maximum entropy density estimation, however, we actually do the minimization calculation here.

Given $u \in SBV(R)$, we can express $D^*(u) = P_n - Q$, where $P_n = D^*(u_{disc})$ and $Q = D^*(u_{abs})$. The operation $P_n - D^*(u)$ cancels the discrete part and leaves only absolutely function $Q$. In such a case, it indicates that $\int_\Omega \psi^*(P_n - D^*u) = \int_\Omega \psi^*(\frac{d(P_n - D^*(u))}{dt})dt$. Given the assumptions on $\psi^*$, we express (2.7) as

$$\begin{aligned} \int_\Omega \psi^*(f(t))dt + \lambda N^*(\frac{u}{\lambda}) &= \min_{u\in\mathcal{U}, f\in\mathcal{V}} ! \\ \text{subject to } f = \frac{d(P_n - D^*(u))}{dx} &\text{ and } f \succeq 0, \end{aligned} \tag{2.8}$$

where $f \in \mathcal{V} = L^1(R)$.

In the univariate setting, we can explain the notations and formulas above in simple and familiar statistical language: first, we use the usual statistical definition of distribution function, and we denote $F$ as the distribution function of the density we seek to estimate. We know that $F$ is bounded and nondecreasing on $R$. Secondly, we could define the bounded variation set on R as: $BV(R) = \{x - y : x, y \in F\}$. For all $u \in BV(R)$, $D^*(u)$ belongs to the the set of finite signed Radon measure. Specifically, for every $u \in BV(R)$, we could define a measure $du(a, b] = u(b) - u(a)$. We still use the definition of empirical distribution function defined in (1.1), we denote $F_n$ as the empirical distribution function supported by all $n$ datapoints.

For instance, if $D^* = D'$, which is the first derivative, then $P_n = D^{'}(F_n)$; if $D^* = D^{''}$, the second derivative, then $P_n = D^{''}(\int F_n dt)$. In general, we denote $P_n = D^*(F_n)$, therefore $P_n - D^*(u) = D^*(F_n - u)$. And we assume the Lebesgue decomposition of $u$ has no singular part, in such a case $F_n - u$ cancels the discrete part and leaves only absolutely continuous part with $f(t) = d(D^*(F_n - u))/dt$.

## 2.2 Preservation of Moments

Usually, if the operator $D$ could annihilate functions like $g(x) = x^k$, then we can conclude the estimate $f$ has the $k$-th moment equal to $k$-th empirical moment supported

by data. This is called the preservation of the $k$-th moment.

We say that a *locally integrable* function $g$ is annihilated by the operator $D$, if $Dg = 0$ for the restriction of $g$ to any bounded domain $\Omega \subset R$. (In the spirit of how we understand differential operators, $D$ should be defined in every such case.)

**Theorem 2.2.1.** *If function $g$ is constant and annihilated by the operator $D$, then the solution of (2.8) is indeed a density function.*

**Proof**. As following from (2.7) and (2.8), the solution of (2.8) is sought among $f$ representing positive Radon measure in the form $P_n - D^*(u)$. Suppose that $D$ is defined for $g$ and $g$ is continuous. Then

$$\begin{aligned} &\textstyle\int_\Omega g d(P_n - D^*(u)) \\ &= \langle P_n - D^*(u), g\rangle \\ &= \langle P_n, g\rangle - \langle D^*(u), g\rangle \\ &= \textstyle\int g dP_n - \langle u, Dg\rangle, \end{aligned} \tag{2.9}$$

where $\langle u, Dg\rangle = \int_\Omega u(Dg)$, that is, $u$ is understood as the distribution it represents. If $g$ is annihilated by $D$, then $\langle u, Dg\rangle$ vanishes on any bounded subdomain $\Omega'$ of $\Omega$, then

$$\int_{\Omega'} g f dx = \int_{\Omega'} g dP_n, \tag{2.10}$$

for any $f$ representing a solution of (2.8). Now, let us take $g \equiv 1$ and suppose that $D$ annihilates $g$. If $\Omega$ is an open set containing all datapoints (and hence the support of $P_n$), then there is a bounded subdomain $\Omega'$ of $\Omega$ such that

$$\int_{\Omega'} f dx = \int_{\Omega'} dP_n = 1. \tag{2.11}$$

From the fact that (2.11) holds for any bounded subdomain $\Omega'$ of $\Omega$ containing the support of $P_n$, we obtain that $f$, being nonnegative, integrates to 1; therefore, it is integrable and belongs to $L^1(\Omega)$.

Suppose that $\psi^*$ and $N^*$ satisfy all assumptions made, and $D^*$ is a distributional adjoint of a differential operator $D$ that annihilates constants. For any $\lambda > 0$, we call any solution $f$ of (2.7) or (2.8) a maximum penalized density estimate.

More generally, if $g(x) = x^k$, then it is possible to show in the analogous fashion that the estimate $f$ has the $k$-th moment equal to $k$-th empirical moment supported by data. In particular, we remark that when $D$ is the operator of $k$-th derivative in R, then any solution of (2.8) automatically satisfies

$$\int_\Omega t^i f(t)dt = \int_\Omega t^i dP_n \quad \text{for } i = 0, 1, 2, \ldots, k-1; \tag{2.12}$$

thus, adding the identities (2.12) does not alter their solutions.

So by adding the moment constraints, we can finally express (2.7) as

$$\begin{array}{l} \int_\Omega \psi(f(t))dt + \lambda(\frac{u}{\lambda}) = \min_{u\in\mathcal{U}, f\in\mathcal{V}}!; \\ \text{subject to} \quad f = \frac{d(P_n - D^{'}(u))}{dx}, \quad f \succeq 0; \\ \text{and} \quad \int_\Omega t^i f(t)dt = \int_\Omega t^i dP_n \quad \text{for } i = 0, 1, 2, \ldots, k-1. \end{array} \tag{2.13}$$

## 2.3 Taut String Theory in Maximum Entropy Density Estimation

The stretched, or taut string methods were firstly considered by Hartigan and Hartigan (1985), Davies and Kovac (2001, 2004), and Koenker and Mizera (2006b).

To illustrate what is taut string method is, we consider the special case that the penalty $N^*$ is the indicator of the unit ball in $L^\infty$ norm.

$$N^*(x) = \begin{cases} +\infty & \|x\|_\infty > 1; \\ 0 & \|x\|_\infty \leq 1, \end{cases} \tag{2.14}$$

or equivalently,

$$\lambda N^*(\frac{u}{\lambda}) = \begin{cases} +\infty & \|u\|_\infty > \lambda; \\ 0 & \|u\|_\infty \leq \lambda. \end{cases} \tag{2.15}$$

If we denote $P_n$ as the derivative of the empirical distribution function $F_n$, then $D^*(u) = D'(u) = u'$, which is the first derivative of $u$. By using notations above, (2.8) can be shown as

$$\begin{aligned} &\textstyle\int_\Omega \psi^*(f(t))dt = \min_{u\in\mathcal{U}, f\in\mathcal{V}}! \\ &\text{subject to} \quad f = \tfrac{d(P_n - u')}{dx}, \quad f \succeq 0, \quad \text{and } \|u\|_{L^\infty} \leq \lambda. \end{aligned} \tag{2.16}$$

If we denote $P_n - u' = q$, where $q$ is the derivative of $F = F_n - u$, then following the assumption above that $F$ is absolutely continuous with $\|u\|_\infty \leq \lambda$, the estimate $F$ is within the Kolmogorov distance $\lambda$ of $F_n$, and minimizes the objective function $\int \psi^*(f(t))$.

If the function $\psi^*$ is the Shannon entropy, then minimizing objective function $\int F'(t) \log F'(t)dt$ leads $F$ to be linear between the points where it touches the boundary of the Kolomogorov "tube". This follows the theorem below.

**Theorem 2.3.1.** *Minimizing $\int_\alpha^\beta F'(x) \log F'(x)dx$ under boundary conditions fixing $F'(\alpha)$, $F'(\beta)$ leads to the solution $F$ linear on $[\alpha, \beta]$.*

**Proof**: The theorem could be simply proved by using the techniques of calculus of variations. For example, if we denote $(F' \log F')(\alpha) = A$, $(F' \log F')(\beta) = B$, then for any $\epsilon \to 0$ and function $g$, which satisfies $g(\alpha) = g(\beta) = 0$, we know $((F' + \epsilon g') \log(F' + \epsilon g'))(\alpha) = A$ and $((F' + \epsilon g') \log(F' + \epsilon g'))(\beta) = B$.

Consider the derivative of the objective function with respect to $\epsilon$ at the point $\epsilon = 0$

$$\frac{d}{d\epsilon}|_{\epsilon=0}(\int_\alpha^\beta (F' + \epsilon g') \log(F' + \epsilon g')) = \int_\alpha^\beta (g'(\log F' + 1)) \tag{2.17}$$

By setting (2.17) equal to zero, for any function $g$, $\int_\alpha^\beta (g'(\log F' + 1)) = 0$, with $g(\alpha) = g(\beta) = 0$. It indicates that $\log F' + 1 = const$ over the interval $[\alpha, \beta]$. Thus, we have

$F'$ must be a constant over $[\alpha, \beta]$. In another word, it means $F$ is a linear function between $\alpha$ and $\beta$.

We can generalize the proof to Rényi entropy case. The proof of the following theorem is analogous to that of Therorem 2.3.1.

**Theorem 2.3.2.** *Minimizing* $\frac{1}{q-1} \log \int (F'(x))^q dx$ *under boundary conditions fixing* $F'(\alpha)$, $F'(\beta)$ *leads to the solution* $F$ *linear on* $[\alpha, \beta]$.

Therefore, if there is no "tube" (or $\lambda \to \infty$), the solution is exactly the linear function between $\alpha$ and $\beta$. If there is "tube" constraint, we have to take the following three cases into consideration.

First, $F$ is the stretched string in the neighborhood of $F_n$ given $\lambda$, which also means the solution is the piecewise linear function within the "tube". A simple instance with four data points is showed in Figure 8. It shows the solution within the tube(dashed lines) is the piecewise linear function connecting the points A, B, C, D, E and F, where the A and F are the end points, like $\alpha$ and $\beta$ we have mentioned in the Theorem 2.3.1, and the middle points C, D and E exactly lie on the "tube". It is because between the points C and E, we could not find a straight line connecting C and E due to the "tube" constraint. The solution between C and E will be a piecewise linear solution, for example, CD and DE, with D be exactly on the tube. We will illustrate why we choose point D later. And it is the same for the piecewise linear solution DE and EF. (the solid points are data points and B is also on the left side of the first tube).

Second, the solution is still the piecewise linear function within the tube, however, within some interval, when $\lambda$ is ideally large, the linear solution may be exactly the linear solution pass through the those intervals according to Theorem 2.3.1 and 2.3.2. For example, in Figure 8, the linear function AC illustrates this point because we can directly connect A and C within the tube.

Finally, if $\lambda$ is too small, there is no linear solution within the tube. For example, if we shrink $\lambda$ in Figure 8 into 0.1, then we could not find any piecewise linear or linear solution within the tube. It means the the problem is not feasible, the maximum

entropy will be infinity in such a case, and the estimate does not exist.

For the first and second cases, a natural question is: there are infinitely many piecewise linear functions within the tube, why shall we choose the function like in Figure 8? The answer is the linear function showed in Figure 8 indeed minimizes the objective negative entropy function.

To demonstrate this point, first of all, we can not find a straight line between AD, AE and AF. However, within the tube, we could connect A and C with a straight line. According to Theorem 2.3.1 and 2.3.2, we conclude that the piecewise linear solution between A and C is exactly the linear function between A and C.

Next, it is simple to verify that $f \log f$ is an increasing function defined on positive $f$. If there is no tube constraint ($\lambda$ is sufficient large), then the solution between C and F will be a straight line. Once the tube is considered, according to the monotone property of $f \log f$, we know the piecewise linear function which is most close to the straight line CF is the solution. Thus, the piecewise linear function CDEF in Figure 8 give us the solution, since D and E are on the tube. That is to say, piecewise linear function CDEF gives smallest objective function values than any linear function within the tube. Overall speaking, the solution will be piecewise linear function ABCDEF in Figure 8.

## 2.4 Maximum Entropy Density

If we consider the limiting case: $\lambda \to \infty$, then the penalization term tends to contribute less and less to the minimization problem (2.8), thus we have the limiting situation

$$\begin{aligned} &\textstyle\int_\Omega \psi(f(t))dt = \min_{u\in\mathcal{U}, f\in\mathcal{V}}! \\ &\text{subject to } f = \tfrac{d(P_n - D^*(u))}{dx} \text{ and } f \succeq 0. \end{aligned} \tag{2.18}$$

For example, if we consider $N^*$ as the indicator of unit ball in $L^\infty$ norm, then the constraint $\|u\|_{L^\infty} \leq \lambda$ will relax more and more as $\lambda \to \infty$. Therefore, the goal under

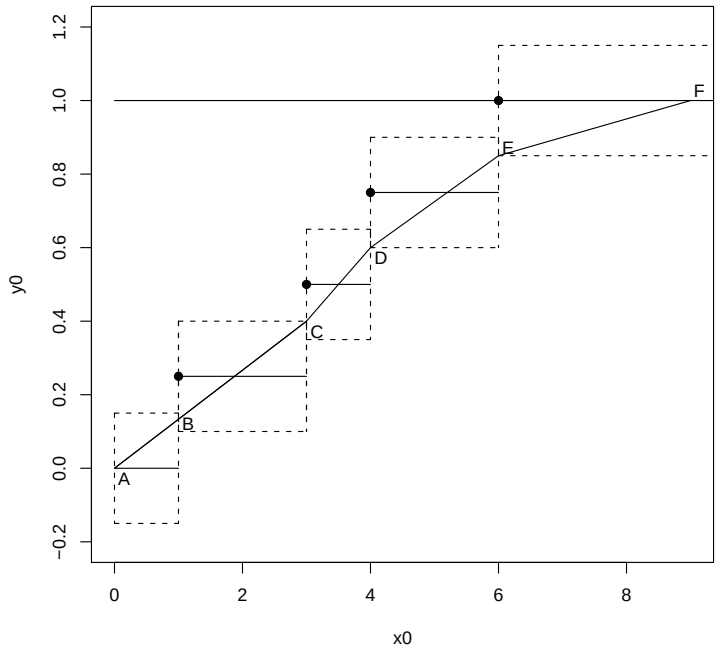


Figure 8: taut string with $\lambda = 0.15$ and the piecewise linear solution within the tube

the limiting case is to find the solution which will minimize the objective negative entropy function.

It is well known that under a variance constraint, Shannon entropy is maximized by Gaussian distribution. So the limiting distribution, the maximizer of Shannon entropy, is normal distribution (under a variance constraint). This is a well known result in information theory. More generally, given certain restrictions, we have already known some results for distributions which can maximize Shannon entropy. For example, in Kagan, Linnik and Rao (1972), they gave the results summarized in Table 2.

Moreover, if we increase our scope to Rényi entropy, then we can obtain some limiting densities, under variance constraint, based on Johnson and Vignat (2006). Johnson and Vignat proved the general results for multivariate case; as the scope of this thesis is only univariate case, we limit out exposition below to the one dimensional

Table 2: Maximum entropy densities

| *Set of values of r.v* | *Restrictions* | *Density function* |
|---|---|---|
| (0,1) | - | $p(x) = 1$ (Uniform) |
| (0,1) | $E\log X = g_1$ <br> $E\log(1-X) = g_2$ | $p(x) = \frac{x^{m-1}(1-x)^{n-1}}{B(m,n)}$ (Beta) |
| (0,$\infty$) | $EX = g_1$ | $p(x) = ae^{-ax}$ (Exponential) |
| (0,$\infty$) | $EX = g_1$ <br> $E\log X = g_2$ | $p(x) = \frac{a^p}{\Gamma(p)}x^{p-1}e^{-ax}$ (Gamma) |
| ($-\infty$, $\infty$) | $EX = g_1$ <br> $EX^2 = g_2$ | $p(x) = \frac{1}{\sigma\sqrt{2\pi}}e^{-(x-\mu)^2/2\sigma^2}$ (Normal) |
| ($-\infty$, $\infty$) | $E\|X\| = g_1$ | $p(x) = \frac{1}{2}ae^{-a\|X\|}$ (Laplace) |

case.

For $1/3 < q$ and $q \neq 1$, define the probability density $g_{q,c}$ as

$$g_{q,c} = A_q(1 - (\frac{q-1}{3q-1})\frac{x^2}{c})_+^{\frac{1}{q-1}}, \tag{2.19}$$

with the normalization constants

$$A_q = \begin{cases} (\Gamma(\frac{1}{1-q})(\frac{1-q}{3q-1})^{1/2})/(\Gamma(\frac{1}{q-1}-\frac{1}{2})(c\pi)^{\frac{1}{2}}) & \text{if } \frac{1}{3} < q < 1, \\ (\Gamma(\frac{q}{q-1}+\frac{1}{2})(\frac{q-1}{3q-1})^{1/2})/(\Gamma(\frac{q}{q-1})(c\pi)^{\frac{1}{2}}) & \text{if } q > 1, \end{cases} \tag{2.20}$$

here $x_+ = \max(x, 0)$ denotes the positive part. And we write $R_{q,c}$ for a random variable with density $g_{q,c}$, which has mean 0 and variance $c$.

Specially, $\lim_{q\to 1}\Gamma(1/(1-q))(1-q)^{1/2}/\Gamma(1/(1-q)-1/2) = 1$, and $\lim_{q\to 1}(1 - \frac{q-1}{3q-1}\frac{x^2}{c}) = \exp(-\frac{x^2}{2c})$, $\lim_{q\to 1} g_{q,c}(x) = g_{1,c}(x) = (2\pi c)^{-1/2}\exp(-\frac{x^2}{2c})$, is the Gaussian density, which corresponds to the maximizer of Shannon entropy.

For $q \neq 1$, given the probability densities $f$ and $g$, define the relative $q$-Rényi

entropy distance from $f$ to $g$ to be

$$D_q(f \parallel g) = \frac{1}{1-q} \log \left( \int g^{q-1}(x) f(x) dx \right) + \frac{1-q}{q} H_q(g) - \frac{1}{q} H_q(f). \tag{2.21}$$

For $q = 1$, we write $D_1(f \parallel g) = \int f(x) \log(f(x)/g(x)) dx$ for the standard relative entropy. We justify this as an extension by continuity; as $q \to 1$, as in (2.20), $D_q(f \parallel g) \to -\int f(x) \log g(x) dx - H_1(f) = D_1(f \parallel g)$.

Johnson and Vignat (2006) also established the following important theorems.

**Theorem 2.4.1.** *For any $q > 0$, and for any probability densities $f$ and $g$ , the relative entropy $D_q(f \parallel g) \geq 0$, with equality if and only if $f = g$ almost everywhere.*

**Theorem 2.4.2.** *Given any $q > 1/3$, and $c > 0$, among all probability densities $f$ with mean 0 and variance $c$, the Renyi-entropy is uniquely maximized by $g_{q,c}$, that is $H_q(f) \leq H_q(g_{q,c})$, with the equality if and only if $f = g_{q,c}$ almost everywhere.*

According to Theorems 2.4.1 and 2.4.2, we have the following conclusions:

First, when $1/3 < q < 1$, if we denote $m = \frac{1+q}{1-q}$, then the density which maximizes the entropy function is:

$$g_{q,c} = \frac{(\Gamma(m + 1/2))}{\Gamma(m/2)((m-2)c\pi)^{1/2}} (1 + \frac{x^2}{(m-2)c})^{-\frac{(m+1)}{2}} \tag{2.22}$$

for $m > 2$, Johnson and Vignat (2006) proposed that $R_{q,c} \sim Z_{(m-2)c}/U$, where $Z_c$ denotes for a $N(0, c)$, $U \sim \chi_m$ and $U$ is independent of $Z$.

Especially, if $c = \frac{m}{m-2}$, then $g_{q,c}$ is the **student t** distribution with $m$ degrees of freedom.

Second, if $q > 1$, we denote $m = \frac{3q-1}{q-1} > 3$, then the density which maximizes the entropy function is:

$$g_{q,c} = \frac{(\Gamma(m/2))}{\Gamma((m-1)/2)(mc\pi)^{1/2}} (1 - \frac{x^2}{mc})^{\frac{(m-3)}{2}}, \tag{2.23}$$

where $x$ satisfies $\{x : x^2 \leq mc\}$. Johnson and Vignat (2006) proposed that $R_{q,c}U \sim Z_{mc}$, where $U \sim \chi_m$ is independent of $R_{q,c}$.

We want to emphasize a very special case when $q = 2(m = 5)$ and $c = 1$, $g_{q,1}(t) = \frac{3}{4\sqrt{5}}(1-\frac{t^2}{5})$ for $|t| < \sqrt{5}$. We notice this distribution is Epanechnikov distribution, which is well known in the theory of kernel density estimation. If we recall the discussion in Section 1.2, among all kernel densities, Epanechnikov density kernel has the largest efficiency; see Silverman (1986).

# Chapter Three: Duality Theorems and Proofs in Shape-Constrained Setting

In Chapter Two, we discussed the maximum entropy density estimation with penalization in continuous setting. In this chapter, we consider the dual and primal formulation applied in density estimation in the discrete case. We will show some additional proofs about the quasi-concave shape constraint density estimation based on Koenker and Mizera (2008a). Finally, we will present an example where the strong duality holds in penalized setting.

## 3.1 Duality Theorem

### 3.1.1 Lagrange Duality Function

In this section, we first introduce an operation that will play an important role in the following sections. All the definitions and notations follow the book of Boyd and Vanderberghe (2004).

Let $f : R^n \to R$. The *conjugate* function of $f$ is defined as,

$$f^*(y) = \sup_{x \in \mathrm{dom} f} (y^T x - f(x)), \tag{3.1}$$

the domain of the conjugate function consists of $y \in R^n$ for which the difference $y^T x - f(x)$ is bounded above on dom $f$.

We see immediately $f^*$ is a convex function, since it is the pointwise supremum of a family of convex (indeed, affine) functions of $y$. This is true whether or not $f$

is convex. From the definition of conjugate function, we could obtain the inequality, $f(x) + f^*(y) \geq x^T y$, for all $x$ and $y$; this is called *Fenchel's inequality* (or *Young inequality* when $f$ is differentiable). The conjugate of the function $\lambda\|\cdot\|_p$, where $\|\cdot\|_p$ stands for the $l^p$ norm, is the indicator of the ball in the dual norm, $\{x : \|x\|_q \leq \lambda\}$, where $q$ satisfies $(1/p) + (1/q) = 1$. The indicator of a convex set $E$ is defined to be 0 for all $x \in E$ and $+\infty$ otherwise. The conjugate of the indicator of the cone $\{x : x \succeq 0\}$ is the indicator of the polar cone $\{x : x \preceq 0\}$.

We consider a optimization problem

$$\begin{aligned} &\text{minimize } f_0(x), \text{ subject to } \ f_i(x) \leq 0,\ i = 1, \dots m \\ &h_i(x) = 0,\ i = 1, \dots p, \end{aligned} \tag{3.2}$$

with variable $x \in R^n$. We assume the domain $D = \cap_{i=0}^m \text{dom}(f_i) \cap_{i=1}^p \text{dom}(h_i)$ is not empty, and further we denote the optimal value of (3.2) as $p^\star$. We do not need to assume problem (3.2) to be convex.

The main idea in Lagrange duality is to take the constraints in (3.2) into consideration by adding the objective function with a weighted sum of constraint functions. We define the *Lagrangian* $L : R^n \times R^m \times R^p \to R$ associated with problem (3.2) as

$$L(x, \lambda, \nu) = f_0(x) + \sum_{i=1}^m \lambda_i f_i(x) + \sum_{i=1}^p \nu_i h_i(x), \tag{3.3}$$

with $\text{dom} L = D \times R^L \times R^p$. The vectors $\lambda$ and $\nu$ are called the *dual variables* or *Lagrange multiplier vectors* associated with the problem (3.2).

We define the Lagrange dual function (or just *dual function*) $g : R^m \times R^p \to R$ as the minimum value of the Lagrangian over $\{x : for\, \lambda \in R^m,\ \nu \in R^p\}$,

$$g(\lambda, \nu) = \inf_{x \in D} L(x, \lambda, \nu) = \inf_{x \in D} (f_0(x) + \sum_{i=1}^m \lambda_i f_i(x) + \sum_{i=1}^p \nu_i h_i(x)). \tag{3.4}$$

When the Lagrangian is unbounded below in $x$, the dual function takes on the value

$-\infty$. Since the dual function is the pointwise infimum of a family of affine functions of $(\lambda, \nu)$, it is concave, even when the problem (3.2) is not convex.

Thus, from the definition of the Lagrange dual function, we have immediately

$$g(\lambda, \nu) \leq p^\star. \tag{3.5}$$

One can easily find some examples for which we can derive an analytical expression for the Lagrange dual function in Boyd and Vanderberghe (2004, chapter 5).

Finally, we indicate the application of conjugate function in Lagrange dual function. The conjugate function and Lagrange dual function are closely related, for instance, we consider an optimization problem with linear inequality and equality constraints,

$$\text{minimize } f_0(x), \text{ subject to } Ax \preceq b, \ Cx = d. \tag{3.6}$$

Recalling the definition of the conjugate function in (3.1), we can express the dual function for the problem (3.6) as

$$\begin{aligned} g(\lambda, \nu) &= \inf_x (f_0(x) + \lambda(Ax - b) + \nu(Cx - d)) \\ &= -b^T\lambda - d^T\nu + \inf_x (f_0(x) + (A^T\lambda + C^T\nu)^T x) \\ &= -b^T\lambda - d^T\nu - f_0^*(-A^T\lambda - C^T\nu). \end{aligned} \tag{3.7}$$

And we will use the relationship between conjugate function and dual function in the next few sections.

### 3.1.2 The Lagrange Dual Problem

For each pair $(\lambda, \nu)$ with $\lambda \succeq 0$, the Lagrange dual function gives us a lower bound on the optimal value $p^\star$ of the optimization problem (3.2). A natural question is: what is the best lower bound that can be obtained from the Lagrange dual function?

This leads us to consider the optimization problem

$$\text{maximize } g(\lambda, \nu) \text{ subject to } \lambda \succeq 0, \tag{3.8}$$

this problem is called the *Lagrange dual problem* associated with the problem (3.2); the original problem (3.2) is sometimes called the *primal problem*. We require the pair of parameters $(\lambda, \nu)$ with $\lambda \succeq 0$ and $g(\lambda, \nu) > -\infty$. We refer to $(\lambda^\star, \nu^\star)$ as *dual optimal* or *optimal Lagrange multipliers* if they are optimal for the problem (3.8).

Therefore, the optimal value of the Lagrange dual problem, which is denoted as $d^\star$, is, by definition, the best lower bound on $p^\star$ that can be obtained from the Lagrange dual function. In particular, we have the simple but important inequality

$$d^\star \leq p^\star, \tag{3.9}$$

which holds even if the original problem (3.2) is not convex. This important property is called *weak duality*.

It is worth to mention that the weak duality (3.9) also holds when $d^\star$ and $p^\star$ are infinite. For instance, if the primal problem is unbounded below, so that $p^\star = -\infty$, we must have $d^\star = -\infty$. Conversely, if the dual problem is unbounded above, so that $d^\star = \infty$, we must have $p^\star = \infty$.

Next, a very natural and interesting problem is under which conditions that we can achieve $d^\star = p^\star$. So if the equality

$$d^\star = p^\star \tag{3.10}$$

holds, then we say that *strong duality* holds. It means that the best bound can be obtained from the Lagrange dual function.

Generally speaking, the strong duality does not hold. But if the primal problem (3.2) is of the form

$$\text{minimize } f_0(x) \text{ subject to } f_i(x) \leq 0 \;\; i = 1, \ldots m, \; Ax = b, \tag{3.11}$$

with $f_0, \ldots, f_m$ convex, we usually(but not always) have strong duality. There are many results that establish conditions on the problem, beyond convexity, under which strong duality holds. These conditions are called *constraint qualifications*. Among all these constraints, one simple constraint qualification is *Slater's condition*: There exists an $x \in D$ such that

$$f_i(x) < 0 \;\; i = 1, \ldots m, \; Ax = b, \tag{3.12}$$

such a point is sometimes called *strictly feasible*, since the inequality constraints hold with strict inequalities. Slater's theorem states that strong duality holds, if Slater's condition holds (and the problem is convex).

Slater's condition can be refined when some of the inequality constraint functions $f_i$ are affine. For example, if the first $k$ constraint functions $f_1, \ldots, f_k$ are affine, then strong duality holds provided the following weaker condition holds: There exists an $x \in D$ with

$$f_i(x) \leq 0 \;\; i = 1, \ldots k, \;\; f_i(x) < 0, \;\; i = k+1, \ldots m, \;\; Ax = b. \tag{3.13}$$

## 3.2 Primal and Dual Formulation in Density Estimation

In this section, we discuss the application of the dual and primal theory in density estimation with penalized and shape constraints within the discrete setting.

In the continuous version, the primal formulation can be traced back to Leonard (1978) and Silverman (1982). As we mentioned in the previous section, the latter established

$$-\int g dP_n + \int e^g dx + \lambda \int (g^{(k)})^2 dx = \min_g ! \tag{3.14}$$

using the third ($k = 3$) derivative to estimate the logarithm, $g$, of a density $f$, with the symbol $P_n$ denoting the empirical probability supported by the datapoints; Gu (2002)

and others championed second ($k = 2$) derivative instead. The total variation penalty

$$\int |g^{(k)}|dx = \bigvee g^{(k-1)} \tag{3.15}$$

was considered by Koenker and Mizera (2006a, 2006b) for $k = 1, 2, 3$. Dümbgen and Rufibach (2008) investigated maximum likelihood estimation of a log-concave density, which in our setting corresponds to $k = 2$ and the penalty in the form of the non-positivity constraint on the second derivative (with no tunning parameter $\lambda$).

In the discrete setting, we replace the $k$th derivative with the difference operator $P$, and the evaluation operator $L$ and vector of weights $w$ by their typical instances described below.

Koenker and Mizera (2008b) discussed the primal and dual formulation relevant for the numerical estimation of a probability via regularization. Under various situations, we study the problem

$$-w^T Lh + s^T \Psi(g) + J(-Ph) = \min_{g,h}!, \quad \text{subject to } h \preceq g, \tag{3.16}$$

where $L$ and $w$ are evaluation operator and averaging function respectively. The estimated density is denoted by the vector $f$ if its values in some collection of points, which is referred as a *grid*. The evaluation operator $L$ indicates the position of $n$ datapoints with respect to the grid via interpolation; for example, if the datapoints are exactly among the gridpoints, then the $i$th row assigns 1 to a gridpoint equal to the $i$th datapoint and zero otherwise. The vector $w$ assigns weights to the datapoints, $1/n$ to each.

$\Psi(g)$ denotes a real convex function $\Psi$ applied to the components of $g$, while $J(h)$ is a convex function applied to the whole vector $-Ph$, the negative of the result of a linear operator $P$ applied on $h$.

As for the penalization term, $P$ is a discretized version of a differential operator. The penalty $J$ involves an $l^p$ norm and a tuning constant, $\lambda$, customary in the context: say,

$J(u) = \lambda\|u\|_1$ or $J(u) = \lambda\|u\|_2^2$. Regularization may be also expressed in a constrained form, in which $J$ is the indicator of a set $\{u : \|u\|_p \leq \Lambda\}$.

All these examples are symmetric: $J(-u) = J(u)$. An asymmetric example is: $J$ equals to the indicator of $\{u : u \preceq 0\}$, the style of penalization used in density estimation under monotonicity or convexity constraints.

We assume that vectors $w$ and $s$ have nonnegative elements. And we use $\preceq, \succeq$ to denote the componentwise inequalities. If $\Psi$ is nondecreasing, i.e, $\Psi(g) = e^g$, the primal formulation (3.16) can be simplified to the unconstrained problem

$$-w^T LH + s^T \Psi(g) + J(-Pg) = \min_g ! \tag{3.17}$$

We assume that all convex functions in (3.16) and (3.17) have the domains with nonempty interiors. Here, the domain is the set that the convex function is finite, and the convex function is allowed to attain $+\infty$ as a value.

According to the definition of the conjugate function and dual formulation in Section 3.1, Koenker and Mizera claimed that the dual of (3.16), or equivalent, (3.17) is the problem

$$-s^T \Psi^\star(f) - J^\star(e) = \max_{f,e} !, \quad \text{subject to } Sf = L^T w + P^T e, \text{ and } f \succeq 0, \tag{3.18}$$

where $S$=diag$(s)$ and $\Psi^\star(f)$ indicates the componentwise application of $\Psi^\star$. The vector $s$ is composed of the integration weights corresponding to the gridpoints: the identity $s^T f = 1$ demonstrates that the estimated density should integrate to 1.

The fact that the estimated $f$ is indeed a probability density can be most conveniently verified through the dual formulation (3.18).

Koenker and Mizera (2008b) proved theorems regarding to the property of the solution of (3.18) and the strong duality between (3.16) (3.17)and (3.18).

**Theorem 3.2.1.** *Suppose that $W^T L1 = 1$ and $P1 = 0$. Then the solution $f$ of (3.18) satisfies $\sum_j s_j f_j = 1$ and $f_j \geq 0$ for every $j$.*

Compared to the dual (3.18) the relationship of variables appearing in the primal formulations (3.16) and (3.17) to the estimated density is not explicit. However, once a strong duality of (3.16) and (3.18) is demonstrated to be true, the relationship of $g$ and $f$ for qualified $\Psi$ is given by

$$f = \Psi^{'}(g), \tag{3.19}$$

where $\Psi^{'}(g)$ indicates the componentwise application of $\Psi^{'}$, the derivative of $\Psi$

**Theorem 3.2.2.** *Problem (3.18) is a strong dual of the problem (3.16). If $\psi$ is differentiable on the interior $I$ of its domain, then the corresponding solutions of (3.18) and (3.16) satisfies (3.19), whenever $g$ and $f$ are componentwise from $I$ and the image of $I$ under $\Psi^{'}$, respectively.*

Since $\psi(x) = e^x$ is nondecreasing, (3.16) is equivalent to the unconstrained formulation (3.17), whose specific form is, for symmetric $J(u) = \lambda\|u\|_p^p$ and $p = 1, 2$,

$$-w^T Lg + s^T e^g + \lambda\|Pg\|_p^p = \min_g ! \tag{3.20}$$

where $e^g$ is understood componentwise. The additional assumptions of Theorem 3.2.2 are satisfied, so $f = e^g$, and

$$\psi^*(y) = \begin{cases} y\log y - y, & \text{for } y > 0, \\ 0, & \text{for } y = 0, \\ +\infty, & \text{otherwise.} \end{cases} \tag{3.21}$$

According to the definition of conjugate function, the feasibility requirement related to the fact dom $\psi^* = [0, +\infty)$ independently enforces the nonnegativity constraint on $f$. In continuous case, Silverman (1982) showed that the result of (3.20) is a probability density. The same conclusion follows, in the discrete setting, from Theorem 3.2.1 and 3.2.2 for all formulations of type (3.20). If the assumptions of Theorem 3.2.1 regarding

$P$, $L$, and $w$ are satisfied, then the dual objective function is

$$-\sum_j s_j f_j \log f_j + \sum_j s_j f_j = \max_{f,e}!, \tag{3.22}$$

which can be further simplified, since the second sum is equal to 1, a constant. The resulting dual of (3.20), writing in the minimization form, for $p = 1$, is

$$\begin{gathered} \textstyle\sum_j s_j f_j \log f_j = \min_{f,e}!, \\ \text{subject to} \quad Sf = L^T w + P^T e, \quad f \succeq 0, \quad \text{and} \quad \|e\|_\infty \leq \lambda, \end{gathered} \tag{3.23}$$

and for $p = 2$,

$$\begin{gathered} \textstyle\sum_j s_j f_j \log f_j + \frac{1}{4\lambda}\|e\|_2^2 = \min_{f,e}!, \\ \text{subject to} \quad Sf = L^T w + P^T e, \quad \text{and} \quad f \succeq 0. \end{gathered} \tag{3.24}$$

The dual of the *penalty-constrained* version of the primal (3.20),

$$-w^T L g + s^T e^g = \min_g!, \qquad \text{subject to} \quad \|Pg\|_p \leq \Lambda, \tag{3.25}$$

is(when $p$ and $q$ being conjugate)

$$\begin{gathered} \textstyle\sum_j s_j f_j \log f_j + \Lambda\|e\|_q = \min_{f,e}!, \\ \text{subject to} \quad Sf = L^T w + P^T e, \quad \text{and} \quad f \succeq 0. \end{gathered} \tag{3.26}$$

Finally, the dual of the shape constrained formulation,

$$-w^T L g + s^T \Psi(g) = \min_g!, \qquad \text{subject to} \quad Pg \leq 0 \tag{3.27}$$

(yielding log-concave $f$ when $P$ is a second order difference operator), is

$$\begin{gathered} \textstyle\sum_j s_j f_j \log f_j = \min_{f,e}!, \\ \text{subject to} \quad Sf = L^T w + P^T e, \quad f \succeq 0 \text{and} \quad e \preceq 0. \end{gathered} \tag{3.28}$$

We notice that the essence of all dual variants is the maximization of the Shannon entropy of $f$. We can generalize the dual function of the penalized likelihood problem by replacing the Shannon entropy term by some of the Rényi entropies we discussed in the previous section.

Let $\psi_\alpha$ be a function equal to $x^\alpha/\alpha$ for $x \geq 0$ and to 0 for $x < 0$. the conjugate of $\psi$, say $\psi^*_\alpha$ is equal to $y^\beta/\beta$ for $y \geq 0$ ($\alpha$ and $\beta$ are conjugate), and to $+\infty$ otherwise. Note that $\psi_p$ is nondecreasing, hence (3.16) is equivalent to (3.17) whenever $\psi = \psi_\alpha$.

**Example**. The special case of the Rényi entropy for $\alpha = 2$ yields $\psi(x) = \psi_2$ and $\psi^*_2$ for $y \geq 0$. The dual can be easily obtained by replacing the Shannon entropy term $\sum_j s_j f_j \log f_j$ in the objective function of (3.23), (3.24), (3.26), and (3.28) by $s^T f^2$, and eliminating the redundant constants in the objective. Specifically, This leads to the dual objective function for $p = 1$,

$$\begin{gathered} \textstyle\sum_j s_j f_j^2 = \min_{f,e}!, \\ \text{subject to} \quad Sf = L^T w + P^T e, \quad f \succeq 0, \quad \text{and} \quad \|e\|_\infty \leq \lambda, \end{gathered} \tag{3.29}$$

and for $p = 2$,

$$\begin{gathered} \textstyle\sum_j s_j f_j^2 \log f_j + \frac{1}{4\lambda}\|e\|_2^2 = \min_{f,e}!, \\ \text{subject to} \quad Sf = L^T w + P^T e, \quad \text{and} \quad f \succeq 0. \end{gathered} \tag{3.30}$$

The dual of penalized-constrained version is

$$\begin{gathered} \textstyle\sum_j s_j f_j^2 + \Lambda\|e\|_q = \min_{f,e}!, \\ \text{subject to} \quad Sf = L^T w + P^T e, \quad \text{and} \quad f \succeq 0. \end{gathered} \tag{3.31}$$

And the dual of shape constrained formulation is

$$\begin{gathered} \textstyle\sum_j s_j f_j^2 \log f_j = \min_{f,e}!, \\ \text{subject to} \quad Sf = L^T w + P^T e, \quad f \succeq 0 \text{and} \quad e \preceq 0. \end{gathered} \tag{3.32}$$

The corresponding primal results from replacing $\sum_j s_j e^{g_j}$ in (3.20), (3.25), and

(3.27) by $\sum_j s_j$. This leads the primal formula

$$-w^T Lg + s^T 1 + \lambda \|Pg\|_p^p = \min_g ! \tag{3.33}$$

The primal of the penalty-constrained version,

$$-w^T Lg + s^T 1 = \min_g !, \qquad \text{subject to} \quad \|Pg\|_p \leq \Lambda. \tag{3.34}$$

And shape constrained primal,

$$-w^T Lg + s^T \Psi(g) = \min_g !, \qquad \text{subject to} \quad Pg \leq 0. \tag{3.35}$$

If instead of $\psi_2$ we consider $\psi(x) = (1/2)x^2$ for all $x$, we can cast both primal and dual in a quadratic programming form. However, we need to pay attention that the correct primal formulation has to be written in the constrained form (3.16) now, because $\psi$ is no longer monotone. In particular, the correct formulation for the setting corresponding to (3.17) is

$$-w^T Lh + \frac{1}{2} s^T g^2 + \lambda \|Ph\|_p^p = \min_{g,h} !, \quad \text{subject to} \quad h \preceq g, \tag{3.36}$$

since $\psi'(x) = x$, both primal and dual yield directly $f = g$.

**Example**. (Hellinger) Another important example from the Rényi system, with $\alpha = 1/2$, set $\psi(x) = -1/x$, for $x < 0$ and $+\infty$ elsewhere. The conjugate is $\psi^*(y) = -2\sqrt{y}$, for $y \geq 0$. The dual (for $p = 1$) can be obtained by replacing $-s^T\sqrt{f}$ in the objective of (3.23), (3.24), (3.26),(3.28), and eliminating the constants ; $\sqrt{f}$ is again applied componentwise.

On the other hand, since $\psi$ is nondecreasing, the primal can be cast in its unconstrained version (3.17), just replacing the $\sum_j s_j e^{g_j}$ term in (3.20), (3.25), and (3.27) by $-s^T g^{-1}$; however, the domain restriction for $\psi$ has to be included as a feasibility

constraint. The resulting primal analog of (3.20) is

$$-w^T Lg - s^T g^{-1} + \lambda \|Pg\|_1 = \min_g !, \qquad \text{subject to} \quad g \preceq 0. \tag{3.37}$$

All the dual and primal formulas mentioned above for $\alpha = 1/2$ can be derived in the similar manner as (3.29)-(3.32) and (3.33)-(3.35).

For symmetric penalties, it is more convenient to recast the primal in terms of $h = -g$:

$$w^T Lh + s^T h^{-1} + \lambda \|Ph\|_1 = \min_h !, \qquad \text{subject to} \quad h \succeq 0. \tag{3.38}$$

The estimated density satisfies $f = 1/g^2 = 1/h^2$.

**Example**. (Maximum empirical likelihood). The limiting variant of the Rényi system for $\alpha = 0$ is $\psi(x) = -1/2 - \log(-x)$ for $x < 0$, and $+\infty$ otherwise. The dual puts $-s^T \log f$ into the objective function of (3.23), (3.24), (3.26), and (3.28), while the primal (unconstrained, but with a feasibility constraint) puts $-s^T \log(-g)$ in (3.20), (3.25), and (3.27). For instance, recasting (3.20) in terms of $h = -g$ gives

$$w^T Lh - s^T \log h + \lambda \|Ph\|_1 = \min_h !, \qquad \text{subject to} \quad h \succ 0. \tag{3.39}$$

## 3.3 The Existence of the Solution

Based on the paper of Koenker and Mizera (2008a), we study the shape constraint problem

$$\Psi_n(g) = \frac{1}{n} \sum_{i=1}^{n} g(X_i) + \int \psi(g) dx = \min !, \tag{3.40}$$

subject to $g \in C(X)$ and $g \in K$, where $C(X)$ is the collection of all continuous functions on the convex hull of X, and $K$ stands for the set of all convex functions on $R^d$

According to (1.49) and (1.50), $\alpha$ and $\beta$ are conjugates in the usual sense that $\frac{1}{\beta} + \frac{1}{\alpha} = 1$. All these $\psi$ functions satisfy the assumptions that they are convex and

decreasing, their domain contains $(0, +\infty)$, and the boundary limit behavior requirements is also satisfied. Moreover, they are all strictly convex and differentiable on their domains. Finally, for $\alpha > 0$ they are all bounded from below by 0. So there is no problem with the Lebesgue existence of the integral in the primal formulation, whose general form for $\alpha \neq 1$ can be written in a unified way as

$$\frac{1}{n}\sum_{i=1}^{n} g(X_i) + \frac{1}{\beta}\int |g|^{\beta} dx = \min !, \qquad g \in C(x), \tag{3.41}$$

as well as the relation between the dual and primal solutions, say, $f = |g|^{\beta-1}$.

Koenker and Mizera (2008a) gave the general proof of the existence of the solution to (3.40) in multi-dimensional case. Within the scope of this thesis, we give a proof for the one-dimensional case by using the continuous functions' properties on bounded domain, techniques distinct from the multi-dimensional ones.

**Theorem 3.3.1.** *The solution of $\Psi_n$ exists and is unique on $[X_{(1)}, X_{(n)}]$.*

**Proof**. In order to prove the existence of the solution of (3.40), we first prove the continuity of $\Psi_n(g)$, and we use two steps to illustrate it.

1) We prove that if a sequence of functions $g_k(x) \to g(x)$ on $x \in [X_{(1)}, X_{(n)}]$, then $\Psi_n(g_k) \to \Psi_n(g)$ for $\alpha > 0$ as $k \to \infty$.

First, we consider the case $\beta > 1$. Since $1/\beta + 1/\alpha = 1$, we have $\alpha > 1$. In such a case, $\psi(g) = g^{\beta}/\beta$ with $g(x) \leq 0$ and $g$ is convex and continuous. We assume that the ordered observations $X_{(i)}, i = 1, 2, \ldots n$ are finite in R. Since $g$ is continuous on $[X_{(1)}, X_{(n)}]$, we could assume that there exists $g_k(x) \to g(x)$ on $[X_{(1)}, X_{(n)}]$. Then,

$$\begin{aligned} &|\Psi_n(g_k) - \Psi_n(g)| \\ &\leq |\tfrac{1}{n}\textstyle\sum_{i=1}^{n}(g_k(x_i) - g(x_i))| + \int |(\psi(g_k - \psi(g)))| dx \end{aligned} \tag{3.42}$$

For the first term on the right side of inequality (3.42), provided $g_k(x) \to g(x)$ on

$[X_{(1)}, X_{(n)}]$, we conclude

$$|\frac{1}{n}\sum_{i=1}^{n}(g_k(x_i) - g(x_i))| \to 0. \tag{3.43}$$

For the second term on the right side of inequality (3.42), $|\psi(g_k) - \psi(g)| = |g_k^\beta - g^\beta|/\beta$. Since the exponential function $f(u) = u^\beta$, $\beta > 1$ is continuous on $[X_{(1)}, X_{(n)}]$, based on the Mean Value Theorem

$$\frac{|g_k^\beta - g^\beta|}{\beta} = \frac{|f(g_k) - f(g))|}{\beta} = \frac{|f'(\epsilon)||g_k - g|}{\beta}, \tag{3.44}$$

where $\epsilon \in [\min(g_k(x), g(x)), \max(g_k(x), g(x))]$.

Due to fact $g$ is continuous and $g_k(x) \to g(x)$, we know both $g$ and $g_k$ are bounded on $[X_{(1)}, X_{(n)}]$ with $|g_k - g| \to 0$. Therefore, $\epsilon$ is also bounded. Especially, $f'(\epsilon) = \beta\epsilon^{\beta-1}$ is bounded. We conclude that $|\psi(g_k) - \psi(g)| \to 0$.

As a consequence, we conclude that $|\Psi_n(g_k) - \Psi_n(g)| \to 0$. It means that $\Psi_n(g_k) \to \Psi_n(g)$ for $\beta > 1$.

In the same manner, we can prove that for $\beta < 0$, $0 < \alpha < 1$, if $g_k(x) \to g(x)$, then $\Psi_n(g_k) \to \Psi_n(g)$.

In summary, we prove the first step that if $g_k(x) \to g(x)$ on $x \in [X_{(1)}, X_{(n)}]$, then $\Psi_n(g_n) \to \Psi_n(g)$.

2) We will show if $\|g_k\| \to +\infty$, then $\Psi_n(g_k) \to +\infty$.

Let $(g_k)_{k=1}^{\infty}$ be such vectors that $\|g_k\| \to +\infty$, and we denote $g_k(x_i) \to r_i \in [-\infty, +\infty]$.

First, we consider the case $\beta > 1$, from (1.49), we know that $g(x) \leq 0$, so $g_k(x_i) \to r_i \in [-\infty, 0]$.

Since $\|g_k\| \to \infty$, we suppose there is at least one $i$, such that $r_i = -\infty$. Then,

$$\begin{aligned} \Psi_n(g_k) &= \tfrac{1}{n}\textstyle\sum_{1=1}^{n} g_k(x_i) + \int \psi(g_k)dx \\ &= \tfrac{1}{n}\textstyle\sum_{1=1}^{n} g_k(x_i) + \int(|g_k(x)|^\beta/\beta)dx. \end{aligned} \tag{3.45}$$

Since $\beta > 1$, if we denote $u(x) = -g(x)$, then $u(x)$ is concave and $u(x) \geq 0$. And we can recast (3.45) as

$$\Psi_n(g_k) = -\frac{1}{n}\sum_{1=1}^{n} u_k(x_i) + \int (u_k(x)^\beta/\beta)dx, \tag{3.46}$$

and $u_k(x_i) \to +\infty$. For the second term on the right side of (3.46), based on the concavity property of $u(x)$

$$\begin{aligned} \int (u_k(x)^\beta/\beta)dx &\geq \tfrac{1}{\beta}\int_{x_{i-1}}^{x_i} u_k^\beta(x)dx \\ &= \tfrac{1}{\beta}\int_{x_{i-1}}^{x_i} u_k^\beta(x)[\tfrac{x-x_{i-1}}{x_i-x_{i-1}}x_i + \tfrac{x_i-x}{x_i-x_{i-1}}x_{i-1}]dx \\ &\geq \tfrac{1}{\beta}\int_{x_{i-1}}^{x_i} [u_k(x_i)\tfrac{x-x_{i-1}}{x_i-x_{i-1}} + u_k(x_{i-1})\tfrac{x_i-x}{x_i-x_{i-1}}]^\beta dx. \end{aligned} \tag{3.47}$$

**Lemma 3.3.2.** *If $a$, $b$, $\alpha$, $\beta \geq 0$ and $\alpha + \beta = 1$, then $a\alpha + b\beta \geq a^\alpha + b^\beta$.*

By using the Lemma 3.3.2, and the nonnegative property of $u$

$$\begin{aligned} \int (u_k(x)^\beta/\beta)dx &\geq \tfrac{1}{\beta}\int_{x_{i-1}}^{x_i} u_k(x_i)^{\beta\frac{x-x_{i-1}}{x_i-x_{i-1}}} u_k(x_{i-1})^{\beta\frac{x_i-x}{x_i-x_{i-1}}} dx \\ &= \tfrac{u_k^\beta(x_i)-u_k^\beta(x_{i-1})}{\beta^2 \ln u_k(x_i)/u_k(x_{i-1})}(x_i - x_{i-1}), \end{aligned} \tag{3.48}$$

therefore,

$$\begin{aligned} \Psi_n(g_k) \geq \tfrac{1}{n}\textstyle\sum_{j=1,j\neq i}^{n}(-)u_k(x_j) - \tfrac{1}{n}u_k(x_i) \\ + \tfrac{u_k^\beta(x_i)-u_k^\beta(x_{i-1})}{\beta^2 \ln u_k(x_i)/u_k(x_{i-1})}(x_i - x_{i-1}). \end{aligned} \tag{3.49}$$

We notice the fact that: $\lim_{x\to+\infty}(\frac{x^\beta}{\beta^2 \ln x} - x) = +\infty$, $u_k(x_i) \to +\infty$, and $u_k(x_{i-1})$ is finite, then

$$\frac{u_k^\beta(x_i) - u_k^\beta(x_{i-1})}{\beta^2 \ln u_k(x_i)/u_k(x_{i-1})} - \frac{1}{n}u_k(x_i) \to +\infty. \tag{3.50}$$

The remaining terms on the right side of (3.49) are finite, so we could conclude that $\Psi_n(g_k) \to +\infty$ for $\beta > 1$.

Secondly, we consider the case $\beta < 0$ and $< \alpha < 1$. In such a case, we have $\psi(g) = -g^\beta/\beta$, $g > 0$, and $g$ is convex. If $g_k(x_i) \to +\infty$ for at least one $i$, then we

know $\frac{1}{n}\sum_{i=1}^{n} g_k(x_i) \to +\infty$ and $-\frac{1}{\beta}\int g^{\beta}dx \to 0$ when $\beta < 0$, it implies the fact that $\Psi_n(g_k) \to +\infty$.

Finally, we conclude that if $\|g_k\| \to +\infty$, then $\Psi_n(g_k) \to +\infty$. In summary, according to 1) and 2), we know that $\Psi_n(g)$ is continuous on $[X_{(1)}, X_{(n)}]$.

Due to the important property that continuous function has minimum value on the closed interval, we know that the solution of (3.40) exists.

For the uniqueness of the solution of (3.40), we know that $\Psi_n$ is a strictly convex functional in the sense that

$$\Psi_n((1-\lambda)g_1 + \lambda g_2) < (1-\lambda)\Psi_n(g_1) + \lambda\Psi_n(g_2), \tag{3.51}$$

for $\lambda \in (0,1)$ and convex function $g_1$ and $g_2$: $R \to [-\infty, +\infty)$ such that $\int \psi(g_i) < +\infty$ and $\text{Leb}(g_1 \neq g_2) > 0$.

So if the solution of (3.40) is not unique, for example, we have two different solutions $g_1$ and $g_2$, then according to (3.51)

$$\Psi_n((1-\lambda)g_1 + \lambda g_2) < (1-\lambda)\Psi_n(g_1) + \lambda\Psi_n(g_2) \leq \min\left(\Psi_n(g_1), \Psi_n(g_2)\right). \tag{3.52}$$

It indicates $(1-\lambda)g_1 + \lambda g_2$ is the minimizer of $\Psi_n$. However, it contradicts the fact that $g_1$ and $g_2$ are two different solutions of (3.40).

Thus we have illustrated the uniqueness of the solution of (3.33).

## 3.4 Fisher's Consistency

Consistency of these estimators in dimension one for selected $\alpha$ has been addressed by several authors: Pal, Woodroofe, and Meyer (2006) prove consistency in the Hellinger metric for the log-concave ($\alpha = 1$) case, Groeneboom, Jongbloed, and Wellner (2001b) establish consistency and rates of convergence for $\alpha \in 1, 2$ in the uniform metric. Dümbgen and Rufibach (2008) improved Groeneboom, Jongbloed, and Wellner (2001b)'s

work and proved the difference between the empirical and log-concave estimated distribution function vanished with the rate $o_p(n^{-1/2})$ under certain regularity assumptions. Koenker and Mizera (2008a) prove that the Fisher consistency, can be verified by Geometric Inequality for the special case of $\alpha = 1/2$ as,

$$\int \frac{1}{\sqrt{f}} dP_n + \int \sqrt{f} dx. \tag{3.53}$$

Replacing $dP_n$ by $f_0 dx$, where $f_0$ is the unknown target density and $P_n$ is the empirical density function supported by the datapoints yields,

$$2\int \sqrt{f_0} \leq \int \frac{f_0}{\sqrt{f}} dx + \int \sqrt{f} dx, \tag{3.54}$$

which follows the inequality,

$$\sqrt{f_0 f} \leq \frac{f_0 + f}{2}, \tag{3.55}$$

with the "=" if and only if $f = f_0$. Thus, for the case $\alpha = 1/2$, we see that the unknown target density $f_0$ really minimize the problem (3.40) and will be the solution we are going to seek for.

We will extend the proof of the Fisher consistency of the problem (3.40) to the general cases for all $\alpha \neq 1$. We consider the proof for $0 < \alpha < 1$ and $\alpha > 1$ separately.

First, if $\alpha > 1$, then $\beta > 1$. In such a case, the Primal formulation is

$$\Psi(g) = -\frac{1}{n}\sum_{i=1}^{n} g(X_i) + \frac{1}{\beta}\int g^{\beta} dx, \quad \text{subject to } g \text{ is concave and } g \geq 0. \tag{3.56}$$

By applying the same techniques,

$$\Psi(g) = \int -g dP_n + \frac{1}{\beta}\int g^{\beta} dx. \tag{3.57}$$

Replacing $dP_n$ by $f_0 dx$, where $f_0$ is the unknown target density yields

$$\Psi(g) = \int -g f_0 dx + \frac{1}{\beta} \int g^\beta dx. \tag{3.58}$$

When strong duality holds, we know $f = g^{\beta-1}$. Therefore, the aim is to prove for all concave and non-negative function $g$

$$\Psi(f_0^{\frac{1}{\beta-1}}) \leq \Psi(g), \tag{3.59}$$

this relationship can be verified by taking the difference

$$\begin{aligned} \Psi(g) - \Psi(f_0^{\frac{1}{\beta-1}}) &= \tfrac{1}{\beta} \textstyle\int g^\beta dx - \int g f_0 dx + \int f_0^{\frac{1}{\beta-1}} f_0 dx - \tfrac{1}{\beta} \int f_0^{\frac{\beta}{\beta-1}} dx \\ &= \tfrac{1}{\beta} \textstyle\int g^\beta dx + \tfrac{\beta-1}{\beta} \int f_0^{\frac{1}{\beta-1}} dx - \int g f_0 dx \\ &= \tfrac{1}{\beta} \textstyle\int g^\beta dx + \tfrac{1}{\beta'} \int f_0^{\beta'} dx - \int g f_0 dx. \end{aligned} \tag{3.60}$$

If we denote $(\beta - 1)/\beta = \beta' > 0$, then $1/\beta + 1/\beta' = 1$, and both $g$ and $f_0$ are nonnegative. We apply Young's inequality

$$\frac{1}{\beta} \int g^\beta dx + \frac{1}{\beta'} \int f_0^{\beta'} dx \geq \int g f_0 dx, \tag{3.61}$$

it indicates $\Psi_n(g) - \Psi(f_0^{\frac{1}{\beta-1}}) \geq 0$, so $g = f_0^{\frac{1}{\beta-1}}$ minimizes the primal formulation (3.56). On the other hand, we know $f = g^{\beta-1}$, thus we conclude that $f_0 = f$, it means $f_0$ indeed minimizes $\Psi(g)$.

Next, we consider the case $0 < \alpha < 1$ and $\beta > 0$. The primal formulation is

$$\Psi(g) = \frac{1}{n} \sum_{i=1}^{n} g(X_i) - \frac{1}{\beta} \int g^\beta dx, \quad \text{subject to } g \text{ is convex and } g > 0. \tag{3.62}$$

By applying the Fenchel's inequality, which one can refer to Boyd and Vanderberghe (2004), for every $g$,

$$\psi(g) + \psi^*(-f_0) \geq -g f_0, \tag{3.63}$$

where $\psi(g) = -\frac{1}{\beta}g^{\beta}$ and $\psi^*(-f_0) = -\frac{1}{\alpha}f_0^{\alpha} = -\frac{1}{\beta'}f_0^{\beta'}$,

$$\psi(g) + \psi^*(-f_0) = -\frac{1}{\beta}g^{\beta} - \frac{1}{\beta'}f_0^{\beta'} \geq -gf_0, \tag{3.64}$$

so we conclude that

$$\Psi(g) - \Psi(f_0^{\frac{1}{\beta-1}}) = -\frac{1}{\beta}g^{\beta} - \frac{1}{\beta'}f_0^{\beta'} + gf_0 \geq 0, \tag{3.65}$$

this again demonstrates $f_0 = f$, so $f_0$ actually minimizes $\Psi(g)$ for $0 < \alpha < 1$.

In summary, we conclude that Fisher consistency holds for all positive $\alpha$. It is equivalent to say the population version of (primal) objective function is uniquely minimized at $f_0$.

## 3.5 The Dual and Primal Formulation for the Maximum Entropy Density Estimation

In Chapter Two, we have discussed the maximum entropy density estimation in the continuous setting without duality theorem. Since we have discussed the dual and primal formulation applied in density estimation in Section 3.2, we will show a special case that the strong duality holds for maximum entropy density estimation with penalization in this section. However, the general duality theorem for maximum entropy density estimation has not been fully proved yet.

We consider the following dual problem

$$-\int \psi(P_n - u') = \max_u, \qquad \text{subject to} \quad \|u\|_{L^\infty} \leq \lambda, \tag{3.66}$$

where the function $\psi$ is the negative Shannon entropy function. The primal of (3.66) is

$$-\frac{1}{n}\sum_{i=1}^{n}\log f(x_i) + \lambda \bigvee(\log f) + \int f dx = \inf_f! \tag{3.67}$$

where $\bigvee$ denotes the total variation of $\log f$ in the domain of $f$.

We consider the very special case that there are only two data points. When only two data points are considered, if $\lambda < 1/4$, from the taut string theory, we know that there is no linear function within the string, which is to say, the dual problem is not feasible. So the dual objective is defined to be infinity.

If we consider $1/4 < \lambda < 1/2$, then we could find a solution within the tube. For different $1/4 < \lambda < 1/2$, the solution could be a straight line between the two end points, or it could be a piecewise linear function as we mentioned in Chapter Two. Here, we first consider the case that the solution is piecewise linear function, we could see this in Figure 3.1.

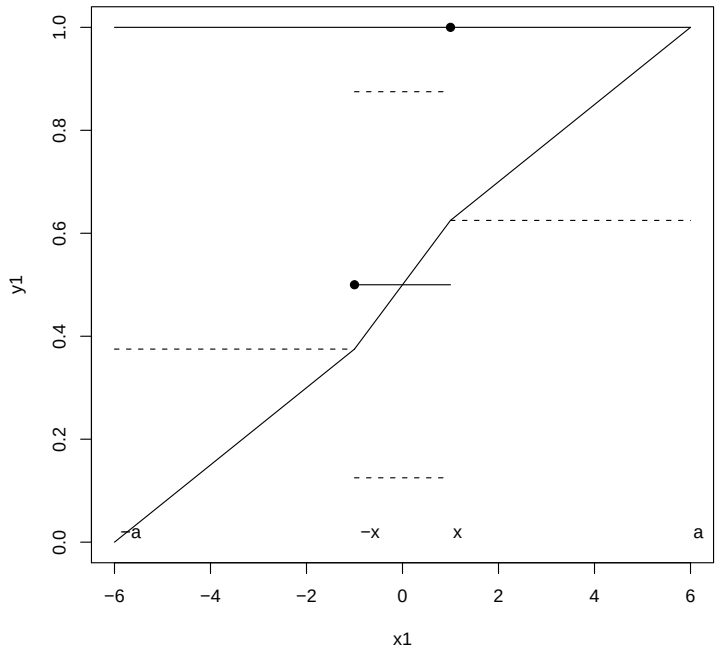


Figure 9: taut string with $1/4 < \lambda < 1/2$ and the piecewise linear solution within the string

In Figure 9, we denote the two data points as $-x$ and $x$ which are the two solid

points, and the two end points as $-a$ and $a$. In such a case, the solution can be treated as three pieces of linear functions. Specifically, the first linear function passes the points $(-a, 0)$ and $(-x, \lambda)$, so it is the function $y = \frac{\lambda}{a-x}t + \frac{a\lambda}{a-x}$. The second linear function is from $(-x, \lambda)$ to $(x, 1-\lambda)$, thus the function is $y = \frac{1-2\lambda}{2x}t + \frac{1}{2}$. At last, the third linear function goes through $(x, 1-\lambda)$ and $(a, 1)$, and the linear function is $y = \frac{\lambda}{a-x}t + \frac{a-x-a\lambda}{a-x}$.

The estimated density $f$ is the first derivative the piecewise linear function above. So $f$ is piecewise constants $f_1 = \frac{\lambda}{a-x}$, $f_2 = \frac{1-2\lambda}{2x}$ and $f_3 = \frac{\lambda}{a-x}$. The width of intervals corresponding to each constant are $d_1 = a - x$, $d_2 = 2x$ and $d_3 = a - x$.

Then we could figure out the dual problem as

$$\begin{aligned} & -\int f \log f dx + f \\ & = -\textstyle\sum_i f_i \log f_i d_i + f_i \\ & = -2\lambda \log(\tfrac{\lambda}{a-x}) - (1-2\lambda)\log(\tfrac{1-2\lambda}{2x}) + 1. \end{aligned} \tag{3.68}$$

On the other hand, when we calculate the primal problem (3.67), we need to define the values for $f_{-x}$ and $f_x$ because the piecewise constants functions $f_i$ are not continuous at the two data points. We use a continuous function in Figure 3.2 to approach the piecewise constants function, with $e \to 0$.

Then we could figure out the value of the primal problem by using the continuous function in Figure 10. The primal problem equals

$$\begin{aligned} & -\tfrac{1}{n}\textstyle\sum_{i=1}^{n} \log f(x_i) + \int f dx + \lambda \bigvee (\log f) \\ & = \lim_{e\to 0} -\tfrac{1}{2}[\log(\tfrac{1-2\lambda}{2x} + \log(\tfrac{1-2\lambda}{2x}] + 1 + \delta(e) + 2\lambda[\log(\tfrac{1-2\lambda}{2x} - \log(\tfrac{\lambda}{a-x})] \\ & = -2\lambda \log(\tfrac{\lambda}{a-x}) - (1-2\lambda)\log(\tfrac{1-2\lambda}{2x}) + 1, \end{aligned} \tag{3.69}$$

where $\delta(e) \to 0$ as $e \to 0$.

We find that the dual and primal problem have the same value in this case; the strong duality holds.

Depending on different values of data points and end points, when $1/4 < \lambda < 1/2$,

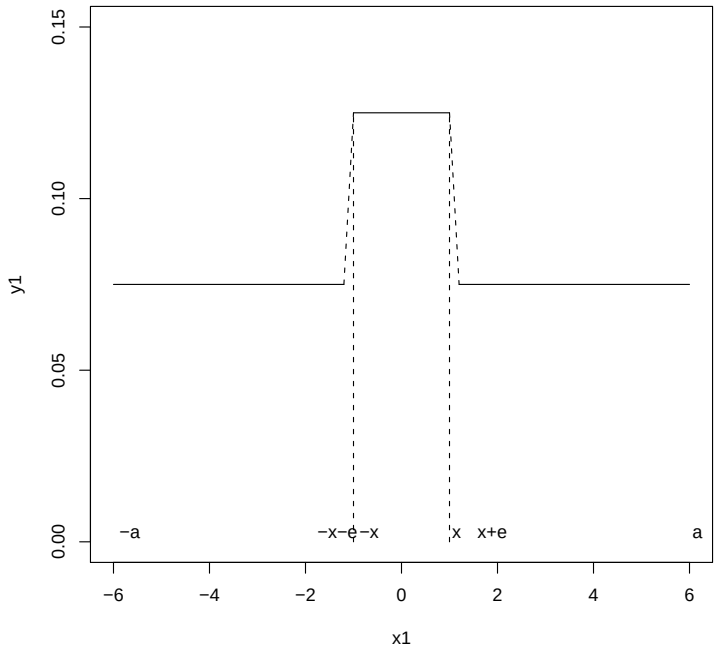


Figure 10: piecewise constants function

the piecewise linear solution could be other situations as described in section 2.3, however, it is in the same manner to verify that the strong duality holds for all data points and end points.

Next, we consider the case $\lambda < 1/4$, we already know that the dual problem is not feasible when $\lambda < 1/4$, thus the dual optimum is negative infinity. We want to verify that the primal optimum is also negative infinity, therefore, the strong duality holds even when the dual problem is not feasible.

In such a case, we could also treat the "solution" as piecewise linear function. However, the "solution" itself is not continuous because the tube constraint. Similarly, the first linear function passes the points $(-a, 0)$ and $(-x, \lambda)$, so it is the function $y = \frac{\lambda}{a-x}t + \frac{a\lambda}{a-x}$. The second linear function is from $(-x, 1/2-\lambda)$ to $(x, 1/2+\lambda)$, which is $y = \frac{\lambda}{x}t + \frac{1}{2}$. The last linear function goes through the points $(x, 1-\lambda)$ and $(a, 1)$,

and the linear function is $y = \frac{\lambda}{a-x}t + \frac{a-x-a\lambda}{a-x}$.

Then we denote $f_1 = f_3 = \frac{\lambda}{a-x}$ and $f_2 = \frac{\lambda}{x}$. Again, we need to use a continuous function to approach the piecewise linear function $f_i$. We use the function in Figure 11 to approach the piecewise linear function, with $cf = gh = d$, and $d, e \to 0$.

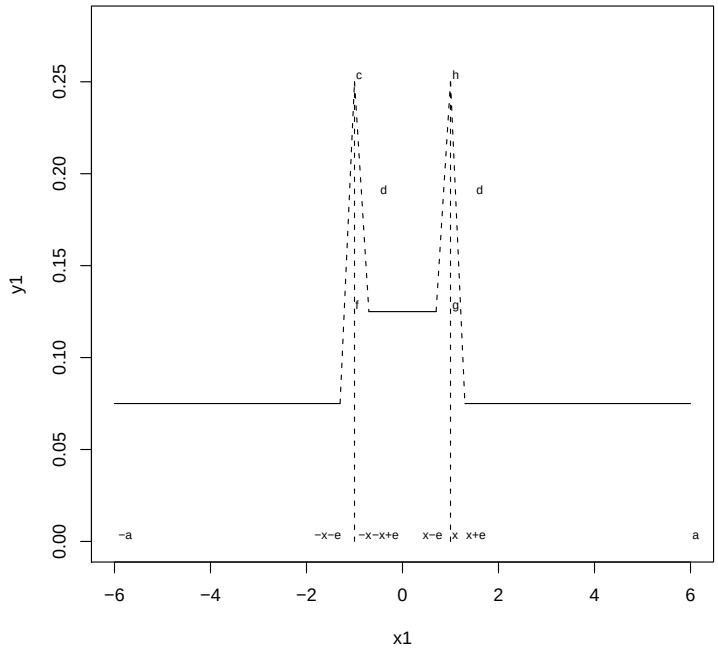


Figure 11: piecewise constants function

Then we could figure out the value of the primal problem by using the continuous function in Figure 11. The primal problem equals

$$\begin{aligned} & -\tfrac{1}{n}\textstyle\sum_{i=1}^{n} \log f(x_i) + \int f dx + \lambda \bigvee(\log f) \\ & = \lim_{d,e\to 0} -\log(d + \tfrac{\lambda}{x}) + 1 + \delta(d,e) + 2\lambda[\log(d) + \log(d + \tfrac{\lambda}{x} - \tfrac{\lambda}{a-x})] \\ & = -\infty, \end{aligned} \tag{3.70}$$

where $\delta(d, e) \to 0$ as $d, e \to 0$ and $\lambda < 1/4$.

Therefore, the dual and primal optima coincide also in this case. We conclude that the strong duality holds.

At last, when $\lambda > 1/2$, the solution is exactly the linear function between the end points, or equivalently, the estimate density is uniform within the interval. And the strong duality could also be verified.

In summary, we conclude the strong duality between (3.66) and (3.67) holds for all $\lambda$ for the two data points' special case. However, for general $n$ data points case, the proof of strong duality requires more complicated techniques which are beyond the scope of this thesis, and there is no such proof so far. This is an important problem for the further research work.

# Chapter Four: Numerical Methods and Data Experiments

As we have already discussed many theoretical techniques in the previous chapters, in this part, we will focus on the numerical methods and data experiments of regularization methods.

## 4.1 Existing Numerical Methods

As we discussed in section 1.4.2, the detailed numerical $L_2$ penalized methods could be found in Ramsay and Silverman (2005).

For $L_1$ penalties, there are also many numerical methods. In Koenker and Mizera (2006a), they established the total variation method based on $L_1$ penalty. Their numerical method will restrict attention to $f'$s for which $\log f$ is piecewise linear on a specified partition of $\Omega$. We can write $J(f)$ as an $L_1$ norm of the second weighted differences of $f$ evaluated at the mesh points of the partition. More explicitly, let $\Omega$ be the closed interval $[x_0, x_m]$ and consider the partition $x_0 < x_1 < \cdots < x_m$ with spacings $h_i = x_i - x_{i-1}, i = 1, \ldots m$. If $\log(f(x))$ is piecewise linear, so that

$$\log(f(x)) = \alpha_i + \beta_i x, \qquad x \in [x_i, x_{i+1}), \tag{4.1}$$

then

$$J(f) = \bigvee_{\Omega}((\log f)') = \sum_{i=1}^{m} |\beta_i - \beta_{i-1}| = \sum_{i=1}^{m} |(\alpha_{i+1} - \alpha_i)/h_{i+1} - (\alpha_i - \alpha_{i-1})/h_i|, \tag{4.2}$$

where we have imposed continuity of $f$ in the last step. We can therefore parameterize functions $f$ by the function values $\alpha_i = \log(f(x_i))$, and this enables us to write our

problem (1.25) as a linear program,

$$\max\{\sum_{i=1}^{n}\alpha_i - \lambda\sum_{j=1}^{m}(u_j) + v_j | D\alpha - u + v, (\alpha, u, v) \in R^n \times R_+^{2m}\}, \tag{4.3}$$

where $D$ denotes a tridiagonal matrix containing the $h_i$ factor for the penalty contribution, and $u$ and $v$ represent the positive and negative parte of the vector $D\alpha$, respectively.

It is worthy to mention that all the penalized numerical data experiments with respect to geyser data in Chapter One are carried out by Matlab.

For the regularization methods with shape constraint, the famous one is established by Dümbgen, Hüsler, and Rufibach (2007). We have discussed the theoretical part of their work in section 1.5, concerning the computation of the log-concave nonparametric maximum likelihood estimation. The authors proposed an active set algorithm based on EM algorithm, which is similar to the vertex reduction algorithms presented by Groeneboom et al. (2007) and are available within the R package "Logcondens", accessible via "CRAN".

Pal, Woodroofe, and Meyer (2005) also considered the nonparametric maximum likelihood estimation in the class of densities with a concave logarithm. The estimation is shown to be the solution of a convex programming problem in the Euclidean space and a numerical algorithm is devised similar to the Iterative Convex Minorant algorithm by Jongbloed (1999).

Koenker and Mizera (2008a) considered the numerical method for log-concave density estimation by primal and dual method. The numerical implementation of Koenker and Mizera's methods are based on two independent algorithms for solving the convex programming problems posed by : mskscopt from the Mosek software package of Andersen (2006), and *PDCO* Matlab procedure of Saunders (2003). Both algorithms are coded in Matlab and employ similar primal-dual, log-barrier methods. The crux of both algorithm is a sequence of Newton-type steps that involving solving large, very

sparse least squares problems, a task that is very efficiently carried out by modern variants of Cholesky decomposition. And this method can also used in multivariate cases, while the log-concave density estimation numerical methods we discussed in previous paragraph are only for the univariate case.

If we consider the primal and dual formulation for density estimation more generally based on the theories in chapter three, for example, from Koenker and Mizera (2008b). The authors concluded that the numerical performance from the dual formulations are always significantly faster and more numerical stable than their primal counterparts.

## 4.2 Data Experiment

Finally, we discuss the application of shape constraint density estimation to the Bright Star data as a numerical experiment. The data source is The Bright Star Catalogue, which contains 9110 objects. The indicators and variables in the data file provides variety of astronomical information out of which radial and rotational velocity are the subjects of concern. There are 9092 objects with radial velocity and 3933 with rotational velocity. The variable rotational velocity is of non-integer type and radial velocity is integer. The densities of these two data sets are the candidates for estimation. The data set and descriptions can be found through the following link *http://cdsarc.u-strasbg.fr/viz-bin/Cat?V/50*

In Pal, Woodroofe and Meyer (2006), the authors also estimated the Polya density of radial velocities of 178 stars by using the data from Walker et al. (2006) in Astrophysical Journal.

Radial velocity is the velocity of an object in the direction of the line of sight (i.e. its speed straight towards or away from an observer). The light of an object with a substantial radial velocity will be subject to Doppler effect, so the frequency of the light decreases for receding objects (redshift) and increases for approaching objects (blueshift).

The radial velocity of a star or other luminous but distant objects can be measured accurately by taking a high-resolution spectrum and comparing the measured wavelengths of known spectral lines to wavelengths from laboratory measurements. By convention, a positive radial velocity indicates the object is receding; if the sign is negative, then the object is approaching.

Rotational speed (sometimes called speed of revolution) indicates how fast a star is running. Rotational speed is equivalent to angular speed, but with different units. Rotational speed tells how many complete rotations (i.e. revolutions or cycles) there are per time unit. It is therefore a cyclic frequency, measured in hertz (revolutions per second) in the SI System. The units revolutions per minute (rpm or 1/min) are more common in everyday life. Angular speed, however, tells the change in angle per time unit, which is measured in radians per second in the SI system.

Before we carry out the numerical experiment, first of all, we draw the histograms of the two velocities in Figure 12. We can see the histogram of radial velocity looks symmetric and centered around zero. On the other hand, the histogram of rotational velocity has a long tail and is obviously skewed.

From the histograms (or one can even plot the kernel estimates) of the two velocities, we are confident of the unimodality of the two densities. This inspires us to use shape constraint methods for the numerical density estimation next.

The first numerical method we use is based on Dümbgen, Hüsler, and Rufibach (2007). The numerical method is available in R, and the density estimation function is "activeSetLogCon". In Figure 13 and 14, we could see the estimated densities of radial and rotational velocities with solid lines. We notice that the estimated radial density is quite similar to the radial velocity histogram. However, the estimated rotational density does not show the spiking shape feature around 0 as the histogram indicated. Thus, Dümbgen, Hüsler, and Rufibach's approach is not wholly satisfied for the rotational velocity data. We consider another method from Koenker and Mizera next.

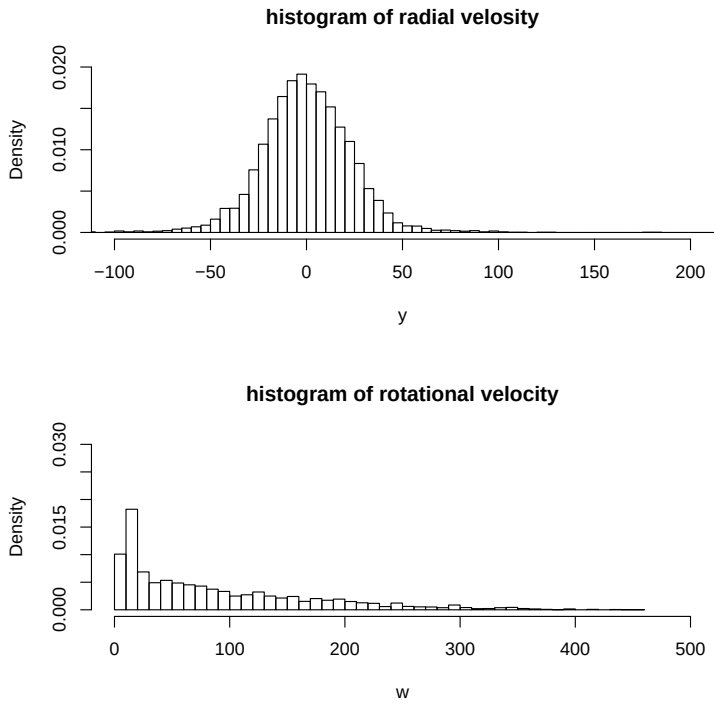


Figure 12: the histogram of the radial and rotational velocity

And the second numerical method is based on Koenker and Mizera (2008a, 2008b), the shape constraint method with the regularization. This numerical method is done by Matlab. Specifically, for each velocity estimation, the first estimated density is calculated by setting the values of the density as zero out of the bound. The second estimated density is calculated by setting the values of the density as $10^{-6}$ out of the bound. Although we find there is almost no difference between the two estimates, the second one always could give the optimal solution of the maximization problem by Matlab, while the first one sometimes can not. The third density estimation is calculated by using the Hellinger estimator mentioned in page 50 ($\alpha = 1/2$).

In Figure 15, we plot the above three estimations of radial density. Actually, we use solid and dashed line to denote the first and second estimate respectively. We observe that these two estimates of radial velocity density almost coincide. While the Hellinger

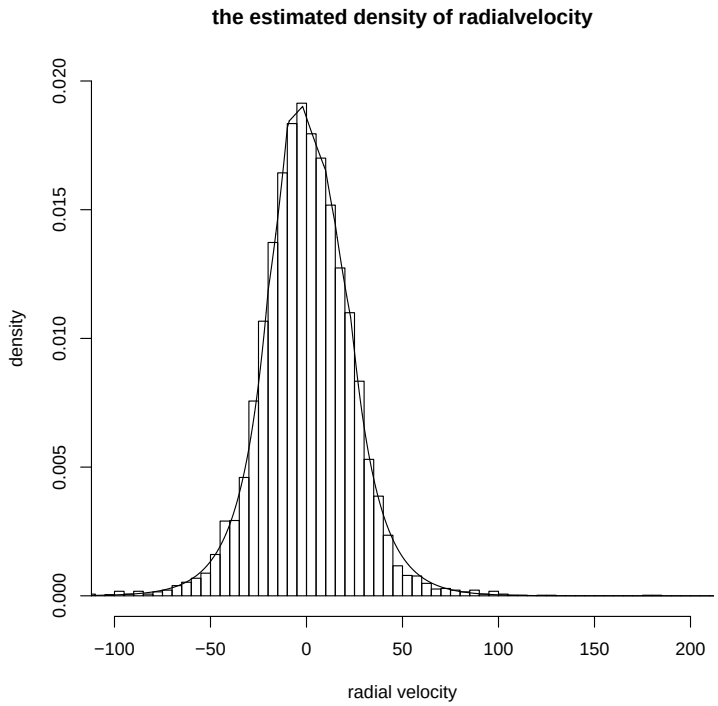


Figure 13: the estimated densities of the radial velocity

estimation is denoted by dotted line, we can see the sight difference around the tails between the Hellinger estimate and the first two estimates.

Overall speaking, the three methods almost give the same estimated densities for radial velocity.

Next, we apply the same methods to rotational velocity. The result is plotted in Figure 16.

Again, we can see there is nearly no difference between the first and second estimator. However, the Hellinger estimator gives more spiking shape around the mode, which is more similar to the histogram.

Finally, we conclude all the shape constraint method we use in this chapter almost gives the same estimated densities for radial velocity data.

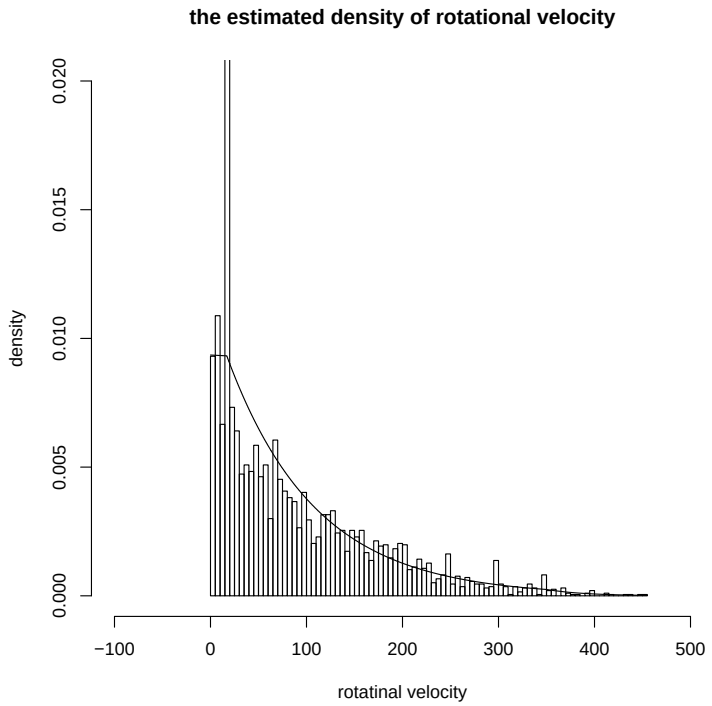


Figure 14: the estimated densities of the rotational velocity

On the other hand, compared with the Dümbgen, Hüsler, and Rufibach's shape constraint method, Koenker and Mizera's methods generate density fits more similar to the histogram for rotational velocity data, especially by applying the Hellinger method. We notice that Koenker and Mizera's numerical methods not only can do log-concave density estimation, but also can estimate the more general quasi-concave density. For example, we recall that the Hellinger estimator corresponds to $-1/2$ concave. In addition, the optimal computation methods used in Dümbgen, Hüsler, and Rufibach's method and Koenker and Mizera's approach are based on different numerical packages in R and Matlab respectively.

It is worth to emphasize that the choice of numerical methods in density estimation depends on the many aspects. In order to make a objective choice of numerical

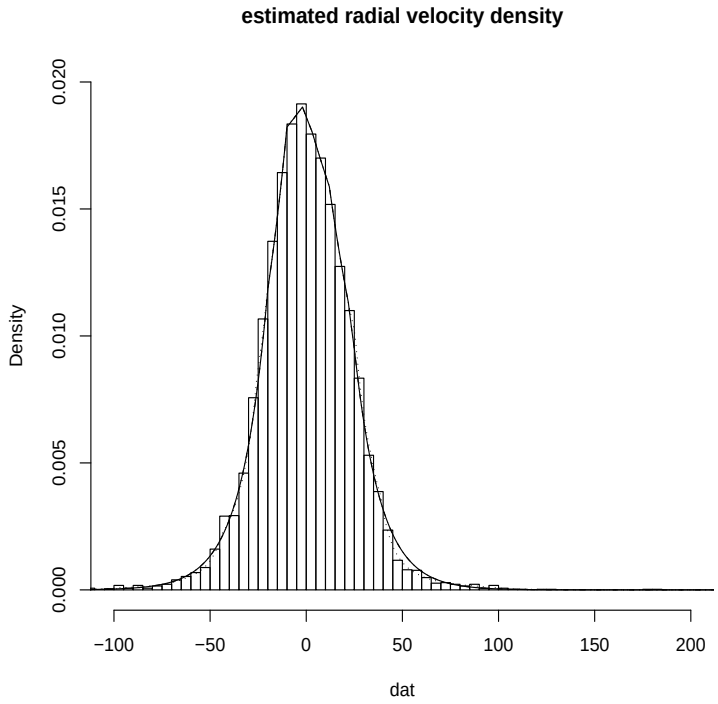


Figure 15: the estimated densities of the radial velocity

approaches, it is beneficial to carry out the data analysis first to obtain the prior information and properties of the data; and then determine the methods according to one's special needs, for instance, the usage of different softwares.

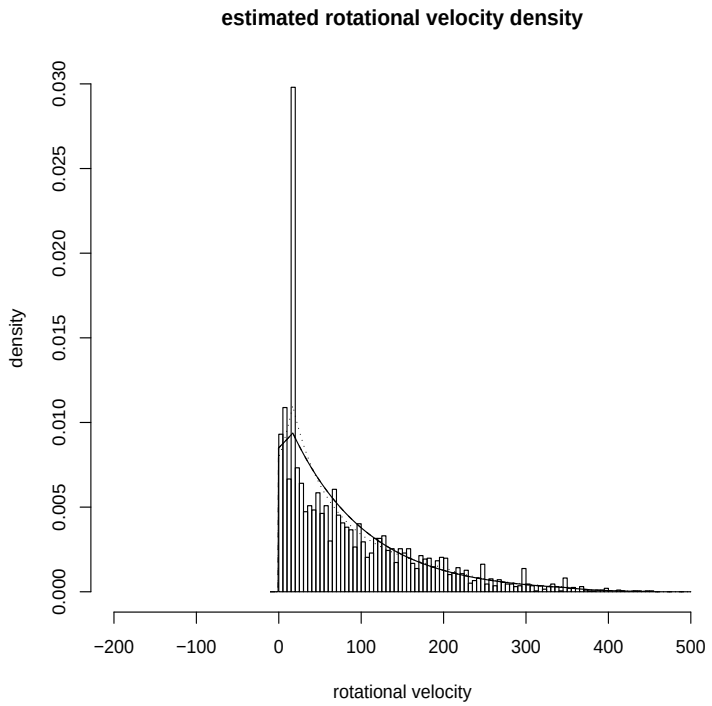


Figure 16: the estimated densities of the radial and rotational velocity

# Conclusion

We have discussed various nonparametric density estimation methods and applications in the thesis. After reviewing all the methods, we can obtain a clear understanding of properties and advantages of each method.

In summary, the histogram method is one of the most simple and basic methods available for nonparametric density estimation. On the other hand, we have to suffer the pain of selecting the origin and binwidth. The discontinuity feature also makes histogram approach unsatisfying for the users. The classical kernel method provide more smooth estimator depending on usage different kernels. However, the performance of kernel method in the boundary of the domain is not very good. For example, when the estimated density has a clear spike around 0 when the domain is $(0, \infty)$. Another drawback is that we have to choose the smoothing parameter.

Regularization methods with penalization are useful since the estimators can show the tail behavior of the density. Especially, for Silverman's method with penalization of third derivative of $\log f$, the domain of the support can be unbounded. The total variation approach by Koenker and Mizera is a good tool to show the spike of the density with second or third derivative of $\log f$. The investigation of the features of penalization methods with a higher derivative of $\log f$ or $f$ is the promising future work.

Regularization methods with shape constraints do not depend on any smoothing parameter. But we need to assume the estimated density has a certain shape, for instance, monotonicity and unimodality. Once we are confident that the density is unimodal, we can add the log-concave or even quasi-concave constraint to estimate the density.

The thesis also discussed some theoretical proofs in chapter two and three, for example, Theorem 2.3.1, 2.3.2, 3.3.1 and the Fisher consistency. One potential problem

is to prove the general consistency result for quasi-concave estimators. And for the general $n$ data points case, the proof of strong duality by using taut string theory is also a complicated but important problem for future research.

# Bibliography


1. Avriel, M.(1972). $\gamma$-Convex Functions, *Math. Programming.*, 2, 309-323.

2. Bertrand-Retali, M.(1978). convergence uniforme d'un estimateur de la densite par la methode de noyau, *Rev. Roumaine Math. Pures. Appl.*, 23, 361-385.

3. Boyd, S and Vandenberghe, L.(2004). Convex Optimization, *Cambridge University Press.*, Cambridge, 2004.

4. Davies, P.L and Kovac, A.(2001). Local extremes, runs, Strings and Multiresolution (with discussion). *Ann. Statist.*, 29:1-65.

5. Davies, P.L and Kovac, A.(2004). Densities, spectral densities and modality. *Ann. statist.*, 32:1093-1136.

6. Dümbgen, L., Hüsler, A. and Rufibach, K.(2007). Active set and EM algorithms for log-concave densities based on complete and censored data. Technical report 61, IMSV, University of Bern (arXiv:0707.4643)

7. Dümbgen, L and Rufibach, K.(2008). Maximum likelihood estimation of a log-concave density: basic properties and uniform consistencies. arXiv:07 09.0334 v3 [math.ST], 2008.

8. Good, I.J and Gaskins, R.A.(1971). Nonparametric roughness penalties for probability densities, *Biometrika*, 58, 255-277.

9. Grenander, P.(1956). On the theory if mortality measurement, part II. *Skandinavisk Aktuarietidskrift* 39, 125-153.

10. Groeneboom, P.(1985). Estimating a monotone density. In: *Proc. Berkeley Conf. in Honor of Jerzy Neyman and Jack Kiefer,* Vol.II (L.M.LeCam and R.A.Ohlsen,eds.),pp. 539-555

11. Groeneboom, P.(1988). Brownian motion with a parabolic drift and Airy functions. *Prob.Theory Rel.Fields* 81, 79-109.

12. Groeneboom, P., Jongbloed, G. and Wellner, J.A.(2001). A canonical process for estimation of convex functions: The "Invelope" of intergrated Brownian motion". *The Annals of Statistics,* 29(6), 1620-1652.

13. Groeneboom, P., Jongbloed, G. and Wellner, J.A.(2007). The support reduction algorithm for computing nonparametric function estimates in the mixture models. Preprint.

14. Gu, C.(2002). Smoothing spline *ANOVA* models. *Springer-Verlag.*

15. Hartigan, J.A and Hartigan, P.M.(1985). The dip test of unimodality. *Ann. Statist.* 13:70-84, 1985.

16. Johnson, O. and Vignat, C.(2005). Some results concerning maximum Renyi Entropy distributions. arXiv:math.PR/0507400 v1, 2005.

17. Jonker, M. and Van Der Vaart, A.(2001). A semi-parametric model for censored and passively registered data. *Bernoulli* 7, 1-31.

18. Kagan, A.M., Linnik, Yu.V. and Rao, C.R.(1972). Characterization Problems in Mathematical Statistics. Nauka, Moskva.

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

20. Koenker, R. and Mizera, I.(2006b). The alter egos of the regularized maximum likelihood density estimators: deregularized maximum-entropy, Shannon, Renyi, Simpson, Gini, and strenched strings. In: proceedings of 7th Prague Symposium on asymptotic Statistics and 15th Prague Conference on information Theory, statistical Decision Functions and Random Processes-Prague Stochastics '06 (M. Huskova, M. Janzura, eds.), Prague, August 21-25, 2006, Matfyzpress, Prague, 2006, pp. 145-157.

21. Koenker, R. and Mizera, I.(2008a). Quasi-Concave Density Estimation. Preprint.

22. Koenker, R. and Mizera, I.(2008b). Primal and Dual Formulations for the Numerical Estimation of a Probability Density via Regularization. In: *Tatra Mountains Mathematical Publications*, ed. by Pazman, A., Volaufova, J. and Witkowsky, V.

23. Kulikov, V.N. and Lopuhaa, H.P.(2006). The behavior of the NPMLE of a decreasing density near the boundaries of the support. *Ann. Statist.*,34, 742-768.

24. Leonard, T.(1978). Density Estimation, Stochastic Processes and prior information. *J. R. Stat. Soc. (B)*, 40, 113-132.

25. Meyer, C.M. and Woodroofe, M.(2004). Consistent maximum likelihood estimation of a unimodal density using shape restrictions. *Canad. J. Statist.*, 32, 85-100.

26. de Montricher, G.M., Tapia, R.A and Thompson, J.R.(1975). Nonparametric maximum likelihood estimation of probability densities by penalty function methods, *Ann. Statist.*, 3, 1329-1348.

27. Nadaraya, E.A.(1965). On nonparametric estimates of density functions and regression curves, *Theor. Probab. Appl.*, 10, 186-190.

28. Natanson, I. (1974). Theory of functions of a real variable. *Ungar.*

29. Pal, J., Woodroofe, M. and Meyer, M.(2006). Estimating a Polya frequency function. In:*Complex datasets and Inverse problems: Tomography, Networks and Beyond (R.Liu, W.Strawderman, C.-H.Zhang,eds.)*,PP. 239-249. IMS Lecture Notes-Monograph Series 54.

30. Parzen, E. (1962). On estimation of a probability density function and mode, *Ann. Math. Statist.*, 33, 1065-1076.

31. Ramsay, J.O. and Silverman, B.W.(2005). Functional data Analysis, second edition. *Springer*, New York, 2005.

32. Rao, P. (1969). Estimation of a unimodal density, *Sankhya Ser.*, A 31, 23-36.

33. Rényi, A.(1961). On measures of entropy and information. In: *Processing of the 4th Berkeley Symposium on Mathematical statistics and probability, Vol I*,pp. 547-561. University of California Press, Berkeley.

34. Rényi, A.(1965). On the foundations of information theory. *Revue d'Institut International de Statistique,* 33, 1-14.

35. Rufibach, K.(2006). Log-Concave Density Estimation and Bump Hunting for I.I.D. observations. PHD Dissertation, University of Bern and Gottingen.

36. Silverman, B.W.(1978). Weak and strong uniform consistency of the kernel estimate of a density function and its derivatives, *Ann. Statist.*, 6, 177-184.

37. Silverman, B.W.(1982). On the estimation of a probability density function by the maximum penalized likelihood method, *Ann.Statist.*, 10, 795-810.

38. Silverman, B.W.(1986). Density Estimation for statistics and data analysis, *Chapman and Hall.*

39. Walker, Matt, Mateo, Mario, Olszewski, Ed Bernstein, Rebecca, Wang, Xiao and Woodroofe, M,B.(2006). Internal Kinematics of the Fornax Dwarf Spheroidal Galaxy. *To appear in Astrophysical Journal,* 2006.

40. Wegman, E.J.(1970). Maximum likelihood estimation of a unimodal density function. *Ann. Math. Statist.*, 41, 457-471.

41. Woodroofe, M. and Sun, J. (1993). A penalized maximum likelihood estimate of $f(0+)$ when $f$ is non-increasing. *Ann. Statist.*, 3, 501-515.