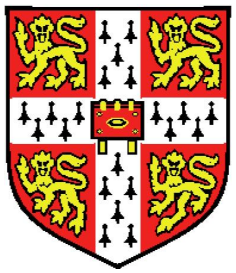
Maximum likelihood estimation of a multivariate log-concave density



Madeleine Cule

Statistical Laboratory, DPMMS

and

Emmanuel College

University of Cambridge

This dissertation is submitted for the degree of

Doctor of Philosophy

14 September 2009

for the Professor

Abstract

Density estimation is a fundamental statistical problem. Many methods are either sensitive to model misspecification (parametric models) or difficult to calibrate, especially for multivariate data (nonparametric smoothing methods). We propose an alternative approach using maximum likelihood under a qualitative assumption on the shape of the density, specifically log-concavity. The class of log-concave densities includes many common parametric families and has desirable properties. For univariate data, these estimators are relatively well understood, and are gaining in popularity in theory and practice. We discuss extensions for multivariate data, which require different techniques.

After establishing existence and uniqueness of the log-concave maximum likelihood estimator for multivariate data, we see that a reformulation allows us to compute it using standard convex optimization techniques. Unlike kernel density estimation, or other nonparametric smoothing methods, this is a fully automatic procedure, and no additional tuning parameters are required.

Since the assumption of log-concavity is non-trivial, we introduce a method for assessing the suitability of this shape constraint and apply it to several simulated datasets and one real dataset. Density estimation is often one stage in a more complicated statistical procedure. With this in mind, we show how the estimator may be used for plug-in estimation of statistical functionals. A second important extension is the use of log-concave components in mixture models. We illustrate how we may use an EM-style algorithm to fit mixture models where the number of components is known. Applications to visualization and classification are presented. In the latter case, improvement over a Gaussian mixture model is demonstrated.

Performance for density estimation is evaluated in two ways. Firstly, we consider Hellinger convergence (the usual metric of theoretical convergence results for nonparametric maximum likelihood estimators). We prove consistency with respect to this metric and heuristically discuss rates of convergence and model misspecification, supported by empirical investigation. Secondly, we use the mean integrated squared error to demonstrate favourable performance compared with kernel density estimates using a variety of bandwidth selectors, including sophisticated adaptive methods.

Throughout, we emphasise the development of stable numerical procedures able to handle the additional complexity of multivariate data.

Declaration

Sections 3.2.1 and 3.2.2 closely follows part of Cule and Dümbgen (2008), which was substantially written by Lutz Dümbgen.

R code to compute bandwidths Abr and Sain in Section 5.5.6 was kindly provided by Tarn Duong.

This dissertation is my own work and contains nothing which is the outcome of work done in collaboration with others, except as specified above, in the Acknowledgements and in the text.

This dissertation is not substantially the same as any that I have submitted for a degree or diploma or other qualification at any other University.

Acknowledgements

This thesis would not have been written without the support of a large number of people. Thanks go to my supervisor Richard Samworth for his guidance. As well as introducing me to this topic and sharing his enthusiasm for statistical theory, he has been very generous with his time and the source of many fresh ideas.

Sections 3.2.1 and 3.2.2 are the outcome of joint work with Lutz Dümbgen conducted during a visit to the Institute of Mathematical Statistics and Actuarial Science at the University of Bern in April 2008. I am indebted to Prof. Dümbgen for his hospitality. This visit was made possible by the support of University of Bern and the Cambridge Statistics Initiative. I also thank Tarn Duong for sharing his code for adaptive bandwidth selection used in Section 5.5.6. Michael Stewart gave valuable suggestions for Section 5.4.

It has been a joy to work in the Statistical Laboratory, and I am grateful to colleagues and friends who made this such an invigorating and supportive environment. In particular, Bobby Gramacy has been an inspiring mentor whose advice on computational statistics, programming and navigating postgraduate study has been invaluable. Pat Altham, Matt Parry and many others gave helpful suggestions. My fellow PhD students, especially my officemate Li Qin, have been a great support throughout.

Finally, I would like to thank my family for their love and support. Many friends in Cambridge and further afield have helped me to maintain a healthy work-life balance. Particular thanks go to Iain Mathieson and Christiana Spyrou who, as well as tireless encouragement, carefully read and provided detailed comments on drafts of this thesis.

This thesis is dedicated to my grandfather.

Contents

List of Figures

List of Figures

List of Algorithms

List of Tables

1 Introduction

1.1 Motivation

This thesis outlines one approach to the fundamental statistical problem of density estimation. In the simplest case we are given an independent sample $\{X_1, \ldots, X_n\}$ drawn from some distribution assumed to have density f_0 in $\mathbb{R}^d$. Our task is to estimate the underlying density. Applications include clustering, exploratory data analysis, classification, and data display (Silverman, 1986; Thompson and Tapia, 1990).

There are many possible approaches to this problem, from the most restrictive parametric models to the most flexible nonparametric methods. In recent years, there has been considerable interest in the area of shape restricted maximum likelihood inference (Balabdaoui, 2004; Groeneboom, Jongbloed, and Wellner, 2001b; Rufibach, 2007; Walther, 2002). We examine one form of shape restriction, namely log-concavity, which has attracted particular attention (Balabdaoui, Rufibach, and Wellner, 2009; Chang and Walther, 2007; Rufibach, 2007; Walther, 2008).

Development of a new technique for an old statistical problem requires both computational considerations (for we must be able to evaluate and compute with our estimator if it is to be of any use) and an evaluation of theoretical and practical performance. After some discussion to motivate our restriction to log-concave densities, we present computational techniques for this estimator. We discuss how this estimator may be used, and investigate the finite-sample performance through simulations.

In this chapter, we review some existing methods of density estimation (Sections 1.2.1 and 1.2.2), consider their advantages and disadvantages, and introduce the basic ideas of shape-constrained maximum likelihood estimation (Section 1.2.3).

1.2 Methods of density estimation

1.2.1 Parametric models and maximum likelihood

One approach to the density estimation problem is to assume that $f_0 = f(\cdot, \theta_0)$ belongs to a family of densities

$$\mathscr{F}_\theta = \{f(\cdot; \theta) \colon \theta \in \Theta\}$$

indexed by a finite-dimensional parameter $\theta \in \Theta \subseteq \mathbb{R}^d$, which, assuming our model is identifiable, reduces our problem to that of constructing an estimate $\widehat{\theta}_n$ of θ_0. Once this has been done, the density may be estimated by $\widehat{f}_n(\cdot) = f(\cdot, \widehat{\theta}_n)$; under conditions on the map $\theta \mapsto f(\cdot; \theta)$, asymptotic results for parameter estimation carry over to the density estimation setting.

There are many general methods for parametric estimation, including the method of moments and maximum likelihood (Rice, 1995). The latter, introduced by Fisher (1922), bases estimation on maximization of the likelihood function

$$L_n(\theta) = \prod_{i=1}^{n} f(X_i; \theta),$$

or equivalently the (averaged) log-likelihood

$$\ell_n(\theta) = \frac{1}{n} \sum_{i=1}^{n} \log f(X_i; \theta) = \int \log f(x; \theta)\, dF_n(x). \tag{1.1}$$

Here F_n denotes the empirical distribution function.

The latter is often simpler to deal with in terms of calculating the maximum likelihood estimator

$$\widehat{\theta}_n = \arg\max_{\theta \in \Theta} \ell_n(\theta).$$

It also simplifies theoretical considerations because the behaviour of (1.1) is easier to study asymptotically, thanks to the laws of large numbers and their generalizations.

The asymptotic properties of these estimators are well understood (at least in the context of exponential families) (Pace and Salvan, 1997; Rice, 1995, and the references therein), and widely used due to the desirable theoretical properties of the estimators (for example, asymptotic efficiency, asymptotic normality and attainment of the Cramer-Rao lower bound) which, under reasonable continuity assumptions, extend to the estimates of the density function as well as those of the parameters. In more complicated models, computational techniques such as the bootstrap (Efron and Tibshirani, 1993) may be used to construct confidence intervals where closed-form theoretical analysis is not possible.

It may be shown that, using $\|\cdot\|_2$ to denote the Euclidean norm on $\mathbb{R}^d$, under regularity conditions

$$\left\| \widehat{\theta}_n - \theta_0 \right\|_2 = O_p(n^{-1/2}). \tag{1.2}$$

As already discussed, under some conditions on the map $\theta \mapsto f(\cdot, \theta)$ and for a suitable choice of norm, the same holds for $\left\| \widehat{f}_n - f_0 \right\|$.

However, the maximum likelihood approach does nothing to address the problem

of finding a suitable family $\mathscr{F}_\theta$, which Fisher (1922) called "entirely a matter for the practical statistician". The consequences of model misspecification can be severe (Huber, 1967; White, 1982), and there have been many attempts to make statistical estimation procedures less sensitive to this (Huber and Ronchetti, 2009).

1.2.2 Nonparametric Smoothing

If no suitable parametric model is available, we may prefer to make fewer assumptions about the underlying density. This comes at a price: our estimate will almost certainly converge more slowly than the $O(n^{-1/2})$ in (1.2), and computation is usually more complicated or expensive than the finite-dimensional maximization problem of (1.1). However, we will be free from making arbitrary assumptions about the shape of the density and can more easily adapt to features of the data in a way that may not be possible using a parametric model.

In a fully nonparametric setting (in which we make no assumptions about f_0 at all), the maximum likelihood approach breaks down. If we attempt to maximize the log-likelihood (1.1) over all densities, we find that $\ell_n(f)$ can be made arbitrarily large, even if we place smoothness restrictions on our density. Informally, this is because we can approximate

$$\frac{1}{n}\delta(x - X_i),$$

a point mass at each observation point, arbitrarily closely.

For example, suppose $X_1, \ldots, X_n$ is a 1-dimensional sample. Set

$$f_h(x) = \frac{1}{n}\sum_{i=1}^{n} \phi_h\left(x - X_i\right)$$

where ϕ is a standard normal density and

$$\phi_h(x) = \frac{1}{h}\phi\left(\frac{x}{h}\right).$$

Then

$$\ell_n(f_h) = \frac{1}{n}\sum_{i=1}^{n} \log f_h(X_i) \to \infty$$

as $h \to 0$.

From this point of view, then, the empirical distribution function can be viewed as the nonparametric maximum likelihood estimate (see Thompson and Tapia (1990) for a more detailed discussion of this viewpoint). However, when our interest is in estimating the density (for example, for classification or visualization purposes), a

different approach is necessary.

Since density estimation is inevitably a tradeoff between fidelity (goodness-of-fit to the data) and parsimony (a simple model), one approach is to simply add a penalty term for "roughness" and maximize the penalized log-likelihood

$$\ell_n(f;\lambda,\rho) = \frac{1}{n}\sum_{i=1}^{n} \log f(X_i) - \lambda\rho(f)$$

where ρ is some roughness measure such as

$$\rho(f) = \int (f''(x))^2\, dx$$

and λ a parameter that must be chosen from the data, for example by cross-validation.

Perhaps the most popular nonparametric density estimator is the kernel density estimator, introduced for univariate data by Fix and Hodges (1951) (see also Fix and Hodges, 1989; Parzen, 1962; Rosenblatt, 1956). This is conceptually simple: we estimate the f_0, a density on $\mathbb{R}$, using a function of the form

$$\widehat{f}_n(x;K,h) = \frac{1}{n}\sum_{i=1}^{n} K_h(x - X_i)$$

where K is (typically) a symmetric density, h a strictly positive bandwidth, and

$$K_h(x) = \frac{1}{h}K\left(\frac{x}{h}\right).$$

Like many common nonparametric density estimates that at first glance appear to have nothing to do with the likelihood approach, the kernel density estimator can also be interpreted as a certain kind of maximum likelihood estimator. Specifically, we may write the kernel estimator as the maximum smoothed likelihood estimator

$$\arg\max_f \ell_{n,h}(f),$$

where

$$\ell_{n,h}(f) = \frac{1}{n}\sum_{i=1}^{n} (K_h * \log f)(X_i).$$

and

$$(f * g)(x) = \int f(x-y)g(y)\, dy.$$

This viewpoint is advocated by Eggermont and LaRiccia (2001).

The extension to multivariate data is, in theory, straightforward. In this case, K is typically a spherically symmetric density, and the smoothing parameter H is a symmetric positive definite matrix. The kernel density estimator is then

$$\widehat{f}_n(x; K, H) = \frac{1}{n}\sum_{i=1}^{n} K_H(x - X_i)$$

where

$$K_H(x) = \frac{1}{|H|^{1/2}} K\left(H^{-1/2}x\right).$$

In order to use kernel density estimation in practice, K and h or H must be specified, with the choice of smoothing parameter being particularly important. The ideal smoothing parameter must strike a balance between a close fit to the data (requiring a small bandwidth, leading to higher variance) and smoothness (requiring a large bandwidth, leading to higher bias). Despite recent progress (surveyed in Section 5.5), choosing a suitable smoothing parameter remains a practical problem.

If H is chosen appropriately, it may be shown that, under regularity conditions on f and K, the asymptotic mean integrated squared error error is $O_p(n^{-4/(d+4)})$ (Scott, 1992; Wand and Jones, 1995). This compares with $O_p(n^{-1})$ for parametric methods as discussed in the previous section.

There are many other smoothing methods, for example, wavelet methods (Donoho, Johnstone, Kerkyacharian, and Picard, 1996), spline methods (Eubank, 1988; Wahba, 1990), penalized likelihood (Eggermont and LaRiccia, 2001), and vector support methods (Vapnik and Mukherjee, 2000). For a review, see Ćwik and Koronacki (1997). However, all suffer from the drawback that some smoothing parameter must be chosen, and the optimal value typically depends on the unknown density. It can be difficult to achieve the correct balance between high variance (caused by undersmoothing) and high bias (caused by oversmoothing), especially when $d > 1$.

For nonparametric density estimation methods, the choice of error criterion can have a significant effect on asymptotic results. Even (in)consistency can vary according to choice of norm (van de Geer, 2000). L_1, L_2 and L_∞ norms, the Kullback-Leibler divergence and Hellinger distance have all been studied. Devroye and Györfi (1985) argue in favour of the L_1 norm, which is the only L_p norm invariant under affine transformation. This is also true of the Hellinger distance and Kullback-Leibler divergence (studied in Chapter 5), which are important for nonparametric maximum likelihood, although they are not so readily interpretable. For kernel density estimation, the L_2 norm has been preferred for its easy decomposition into bias and variance terms (Wand and Jones, 1995). We use this in Section 5.5.

1.2.3 Shape-constrained maximum likelihood estimation

From the arguments at the start of Section 1.2.2, in order for maximum likelihood to be used outside the parametric setting of Section 1.2.1 some restrictions are necessary to ensure that the density does not get too "spiky".

Shape-constrained maximum likelihood inference was first introduced by Grenander (1956) in the context of estimating mortality under the assumption of monotonicity. In this article, an explicit characterization of the nonparametric maximum likelihood estimator as the least concave majorant of the empirical distribution function was given.

Since then a number of other shape constraints have been investigated, including unimodality with known mode (Rao, 1969), convexity (Groeneboom et al., 2001b), k-monotonicity, (Balabdaoui and Wellner, 2007) and log-concavity (Dümbgen and Rufibach, 2008; Walther, 2002). Not all shape constraints are suitable for this approach – for example, the nonparametric maximum likelihood estimator of a unimodal density with unknown mode does not exist (Birgé, 1997).

In some cases a shape constraint may result from the physical problem under consideration (Hampel, 1987; Wang, Woodroofe, Walker, Mateo, and Olzewski, 2005; Watson, 1971). In this case, it is natural to use a shape-constrained estimator. Even if no maximum likelihood estimator exists for the desired class (e.g. unimodal densities with unknown mode, Birgé, 1997), a small additional restriction (e.g. log-concavity) may enable us to use this technique. Even in the absence of a suitable physical model, the lack of tuning parameters makes this approach an appealing alternative to a kernel density estimate. Empirical results (Section 5.5.6) support the claim that these estimators perform well in practice.

Various asymptotic results for the Grenander estimator have been obtained. Pointwise limiting distributional results have been provided by Rao (1969) (see also Groeneboom, 1983), and the L_1 error has been shown to be asymptotically normal, and $O_p(n^{-1/3})$ (Groeneboom, Hooghiemstra, and Lopuhaä, 2001a). This has been extended to the L_p error (Kulikov and Lopuhaä, 2005), where the Grenander estimator was shown to be inconsistent for $p > 2.5$ due to the inconsistency at 0 (identified by Woodroofe and Sun, 1993).

Maximum likelihood estimation for log-concave densities in one dimension was suggested by Walther (2002) and further explored by Dümbgen and Rufibach (2008). Computational issues for $d = 1$ were addressed by Rufibach (2007) and Dümbgen, Hüsler, and Rufibach (2007). Pointwise limit theory was provided by Balabdaoui et al. (2009), and Hellinger consistency has been proved by Pal, Woodroofe, and Meyer (2007). An extension to mixtures of log-concave distributions using the EM algorithm was suggested in Chang and Walther (2007).

1.3 Outline

In Chapter 2, properties of log-concave densities are discussed. Natural applications of log-concave densities are introduced, and we argue in favour of this class as a proxy for other shape-constrained classes for which no maximum likelihood estimator exists (for example, unimodal densities or densities with increasing hazard function). We also argue that this is a suitable class for nonparametric modelling in a general setting. In Section 2.4, we demonstrate the existence and uniqueness of the multivariate log-concave maximum likelihood estimator. We also present some structural features of the maximum likelihood estimator which will be important in Chapter 3.

Chapter 3 discusses the application of subgradient based optimization methods (Kappel and Kuntsevich, 2000; Shor, 1985) to our likelihood maximization problem. These methods were implemented as the package **LogConcDEAD** (Log-Concave Density Estimation in Arbitrary Dimensions) in R (R Development Core Team, 2008), and are discussed in detail in this chapter. We heuristically discuss computational complexity. Further calculations using the maximum likelihood estimator (for example, to compute marginal densities, to evaluate the estimator and to sample from the density) are discussed. This chapter finishes with several examples of the use of **LogConcDEAD** illustrating some important structural features of this estimator.

Log-concavity is a non-trivial assumption, so it is important to have a method for assessing its suitability. The first part of Chapter 4 presents such a method. This is applied to several simulated datasets and to a real dataset, for which we conclude a single component log-concave model is inadequate. We then discuss application to functional estimation using several examples. Finally, we discuss an EM-style algorithm designed to fit finite mixtures of log-concave densities using maximum likelihood. This is a significant extension of the log-concave model and greatly extends its practical relevance. We apply this to the dataset introduced at the beginning of this chapter. We also present an example application to clustering for a second real dataset, and see a significant improvement in misclassification rate over a Gaussian mixture fitted using the EM algorithm.

In Chapter 5, we discuss the asymptotic performance of the estimator. Two points of view are considered. We review the standard approaches to consistency and convergence rates of nonparametric maximum likelihood using empirical process theory. We extend the result of Pal et al. (2007) to prove consistency of our estimator with respect to the Hellinger distance for arbitrary d. Rates of convergence are discussed, and we present simulation results using the Hellinger distance for several examples. Convergence in the case of a misspecified model is also considered. In the second part of this

chapter, we compare the log-concave maximum likelihood estimator to a common competitor, the kernel density estimator. A variety of bandwidth selectors are discussed, and compared empirically with the log-concave maximum likelihood estimator using the mean integrated squared error criterion. The log-concave maximum likelihood estimator is shown to perform well in comparison with kernel methods.

Background material and definitions of terms from convex analysis and computational geometry are given in Appendix A. As an illustration of the package **LogConcDEAD**, the R code used to produce the figures in Section 3.8 is given in Appendix B.

A summary of notation and symbols is given on page 125.

2 Log-concave density estimation

2.1 Introduction

In this chapter, we discuss some basic properties of log-concave densities and their applications, and explain why we find this an attractive and flexible model for multivariate density estimation. In Section 2.4, we prove existence and uniqueness of a maximum likelihood estimator. The insight given by this proof into the structure of the maximum likelihood estimator will be useful in Chapters 3 and 4.

2.1.1 Basic definitions

A function $f : \mathbb{R}^d \to [0,\infty)$ is said to be log-concave if $\log f$ is concave, that is, if

$$\log f(\lambda x + (1-\lambda)y) \geq \lambda \log f(x) + (1-\lambda)\log f(y) \tag{2.1}$$

for all $\lambda \in (0,1)$ and all $x, y \in \mathbb{R}^d$. We adopt the convention $\log 0 = -\infty$.

Note that (2.1) is equivalent to the condition

$$f(\lambda x + (1-\lambda)y) \geq f(x)^{\lambda} f(y)^{1-\lambda}.$$

We say a probability measure $\mathbb{P}$ on $\mathscr{B}_d$, the Borel σ-algebra in $\mathbb{R}^d$, is log-concave if

$$\mathbb{P}(\lambda A + (1-\lambda)B) \geq \mathbb{P}(A)^{\lambda}\mathbb{P}(B)^{1-\lambda} \tag{2.2}$$

for all $\lambda \in (0,1)$ and all $A, B \in \mathscr{B}_d$. A key result is that a probability measure $\mathbb{P}$ is log-concave if and only if $\mathbb{P}$ is absolutely continuous with respect to the affine hull of the support of $\mathbb{P}$, and the corresponding density is log-concave in the sense of (2.1) (Dharmadhikari and Joag-Dev, 1988, Theorem 2.8). We say an $\mathbb{R}^d$-valued random variable X is log-concave if the corresponding probability measure is log-concave. Since we may change the density at a point without altering the probability measure, the density of a log-concave random variable refers to the version of the density satisfying (2.1) throughout. Karlin (1968) showed that the class of log-concave densities are precisely the densities with Pólya frequency of order 2.

2.1.2 Examples

Using the criterion (2.1), it is straightforward to verify that the class of log-concave densities includes many common parametric families (at least for certain parameter values). For $d = 1$, examples include the Gaussian distribution, uniform distribution, Weibull(α) distribution for $\alpha \geq 1$, $\Gamma(n, \lambda)$ distribution for $n \geq 1$ and Beta(a, b) distribution for $a, b \geq 1$. For a more comprehensive list of one-dimensional examples, see Bagnoli and Bergstrom (2005). Multivariate examples include the multivariate normal distribution, Wishart distribution and the Dirichlet distribution.

2.2 Properties of log-concave random variables

The following general fact about log-concave functions gives rise to some additional properties of log-concave densities.

Theorem 2.1 (Prékopa, 1973)**.** *Let $f(x, y)$ be a log-concave function on $\mathbb{R}^m \times \mathbb{R}^n$, with $x \in \mathbb{R}^m$ and $y \in \mathbb{R}^n$. Further, let A be a convex subset of $\mathbb{R}^n$. Then*

$$g(x) = \int_A f(x, y)\, dy$$

is a log-concave function on $\mathbb{R}^m$.

2.2.1 Sums of log-concave random variables

Observing that the product of two log-concave functions is log-concave, we have the following corollary to Theorem 2.1. If $d = 1$, this corollary also arises as a consequence of Theorem 2.9.

Corollary 2.2. *If f and g are two log-concave functions on $\mathbb{R}^m$, then their convolution product $f \star g$ is log-concave. In particular, if X and Y are independent $\mathbb{R}^d$-valued log-concave random variables, then $X + Y$ is log-concave.*

The probabilistic interpretation of Corollary 2.2, namely that the sums of independent log-concave random variables are log-concave, is an attractive feature of this class. As will be seen in Section 2.2.7, this property is not shared by the class of unimodal random variables.

2.2.2 Limits

Using the characterization (2.2), it is proved in Dharmadhikari and Joag-Dev (1988, Theorem 2.10) that the property of log-concavity is preserved under weak limits. As

a direct consequence of Rockafellar (1997, Theorem 10.8), log-concavity is preserved under pointwise limits.

2.2.3 Product measures

It is proved in Dharmadhikari and Joag-Dev (1988, Theorem 2.7) that, if $\mathbb{P}$ and $\mathbb{Q}$ are log-concave measures on $\mathbb{R}^p$ and $\mathbb{R}^q$ respectively, then $\mathbb{P} \times \mathbb{Q}$ is log-concave on $\mathbb{R}^{p\times q}$. This fact is used to prove that log-concave random variables (in the sense of (2.2)) have log-concave densities.

2.2.4 Marginals and conditionals

Let X denote a log-concave $\mathbb{R}^d$-valued random variable, and let $T = PX$, where P is the matrix of some orthogonal projection onto a k-dimensional subspace of $\mathbb{R}^d$. By performing a suitable orthogonal transformation (so that the projection is on to the first k components of X) and observing that $\{x : Px = t\}$ is a convex set (Definition A.2) for $t \in \mathbb{R}^d$, we see immediately using Theorem 2.1 that the marginal density

$$f_T(t) = \int_{\{x\,:\,Px=t\}} f_X(x)\,dx$$

is log-concave. In particular, we have the following corollary.

Corollary 2.3. *If f is a log-concave density on $\mathbb{R}^d$, then all marginal densities are log-concave.*

Considering the same projection, the conditional density $f_{X|T}$ is given by

$$f_{X|T}(x|t) = \begin{cases} \dfrac{f_X(x)\mathbb{1}_{\{Px=t\}}}{f_T(t)} & \text{if } f_T(t) > 0 \\ 0 & \text{else.} \end{cases}$$

Observe that if $f_T(t) = 0$ then $f_{X|T}(x|t) = 0$ for all x, so $f_{X|T}$ is log-concave. If $f_T(t) > 0$, then $f_{X|T}(x|t)$ is log-concave (as a function of x for fixed t) by Theorem 2.1, since $\{x : Px = t\}$ is a convex set.

Combining these two observations, we have the following

Theorem 2.4. *Suppose X is an $\mathbb{R}^d$-valued log-concave random variable with density f, and T is the result of orthogonally projecting X onto some k-dimensional subspace of $\mathbb{R}^d$. Then the marginal density f_T and the conditional density $f_{X|T}(\cdot|t)$ (for fixed t) are log-concave.*

For the converse, we may weaken the conditions slightly.

Theorem 2.5. *Suppose X is an $\mathbb{R}^d$-valued random variable with density f such that, for every projection onto a $(d-1)$-dimensional subspace with corresponding projection matrix P, the conditional density $f_{X|T}(\cdot|t)$ is log-concave, where $T = PX$. Then f is log-concave.*

Proof. Take two points x and y in $\mathbb{R}^d$ and $\lambda \in (0,1)$. Set

$$V = \{\tau(x-y) \colon \tau \in \mathbb{R}\},$$

and let $V^\perp$ denote the orthogonal complement of V. Let P be the matrix corresponding to orthogonal projection onto $V^\perp$, and set $t = Px\ (= Py)$. If $f_T(t) = 0$, we can set $f_X(\lambda x + (1-\lambda)y) = 0$. If $f_T(t) > 0$,

$$f_X(\lambda x + (1-\lambda)y) = f_{X|T}(\lambda x + (1-\lambda)y|t) f_T(t)$$

so that f_X is log-concave. □

2.2.5 Existence of moments

Clearly a log-concave random variable must have light tails (the borderline case being the exponential distribution). This is made precise by the following result. For $d = 1$, a proof may be found in Dharmadhikari and Joag-Dev (1988) or Eggermont and LaRiccia (2001).

Theorem 2.6. *Suppose X is an $\mathbb{R}^d$-valued log-concave random variable. Then there exists $\theta_0 > 0$ such that, for all $\theta \in \mathbb{R}^d$ with $\|\theta\|_2 < \theta_0$, $M_X(\theta) = \mathbb{E}(e^{\theta^T X}) < \infty$.*

Proof. First we consider the case $d = 1$. Suppose X is a real-valued random variable with log-concave density f. Following Eggermont and LaRiccia (2001), observe that for a and $x \in \mathbb{R}$, we have

$$\log f(x) = \log f(a) + \int_a^x \frac{d}{dy} \log f(y)\, dy.$$

Since $\frac{d}{dx} \log f(x)$ is nonincreasing, for $x > a$

$$\log f(x) \le \log f(a) + (x-a) \left.\frac{d}{dy} \log f(y)\right|_{y=a}$$

and for $x < a$

$$\log f(x) \le \log f(a) + (a-x) \left.\frac{d}{dy} \log f(y)\right|_{y=a},$$

so that

$$\log f(x) \leq \log f(a) + |x-a| \left. \frac{d}{dy} \log f(y) \right|_{y=a}.$$

Since f is integrable, we must have

$$\left. \frac{d}{dy} \log f(y) \right|_{y=a} < 0$$

for at least one $a \in \mathbb{R}$, so we conclude that for some strictly positive constants A and c we have

$$f(x) \leq A \exp(-c|x|)$$

and the result follows, taking $\theta_0 = c$.

For the multivariate extension, observe that each of the marginal components $1, \ldots, d$ is log-concave (Theorem 2.4). Set c_i and A_i to be constants such that, for component i, the marginal density f_i satisfies

$$f_i(x) \leq A_i \exp(-c_i|x|),$$

and set $c = \min\{c_1, \ldots, c_d\}$. Then, for each nonzero $\theta \in \mathbb{R}^d$ such that $\sum |\theta_i| < c$,

$$\begin{aligned}
M_X(\theta) &= \mathbb{E}\left[e^{\theta^T X}\right] \\
&= \mathbb{E}\left[\exp\left(\theta_1 X_1 + \ldots + \theta_d X_d\right)\right] \\
&= \mathbb{E}\left[\exp\left(\frac{|\theta_1| \operatorname{sgn}(\theta_1) \sum |\theta_i|}{\sum |\theta_i|} X_1 + \ldots + \frac{|\theta_d| \operatorname{sgn}(\theta_d) \sum |\theta_i|}{\sum |\theta_i|} X_d\right)\right] \\
&\leq \frac{|\theta_1|}{\sum |\theta_i|} \mathbb{E}\left[\exp\left(\operatorname{sgn}(\theta_1) \sum |\theta_i| X_1\right)\right] + \ldots \\
&\qquad + \frac{|\theta_d|}{\sum |\theta_i|} \mathbb{E}\left[\exp\left(\operatorname{sgn}(\theta_d) \sum |\theta_i| X_d\right)\right] \\
&< \infty,
\end{aligned}$$

using Jensen's inequality and the $d = 1$ result. □

2.2.6 Mixtures

A mixture of log-concave densities may be log-concave, but in general it will not be. This contrasts with the situation for log-convex densities, where mixtures are always log-convex (An, 1995, 1998).

2.2 Properties of log-concave random variables

The key to understanding mixtures of log-concave random variables is the following proposition, proved in Walther (2002).

Proposition 2.7 (Walther, 2002)**.** *Let $X_1, \ldots, X_k$ be log-concave random variables on $\mathbb{R}^d$. Let X_i have density f_i and support S_i, and let $S = \cap_{i=1}^k S_i$. Then on any compact subset of S, and for any $\pi_i > 0$ with $\sum_{i=1}^k \pi_i = 1$, we have the following representation:*

$$f(x) = \sum_{i=1}^{k} \pi_i f_i(x) = \exp(\varphi(x) + c\,\|x\|_2^2) \tag{2.3}$$

for some $c \geq 0$ and some concave function φ.

Clearly if the representation above holds for $c = c_0$, then it also holds for $c > c_0$ since the sum of two concave functions is concave (Rockafellar, 1997, Theorem 5.2). Moreover, if this representation holds for all $c > c_0$ for some c_0, then

$$\log f(x) - c_0\,\|x\|_2^2 = \inf_{c > c_0} \left\{ \log f(x) - c\,\|x\|_2^2 \right\}$$

so by Rockafellar (1997, Theorem 10.8), this also holds for $c = c_0$. Thus the representation (2.3) holds if and only if $c \in [c_{\text{true}}, \infty)$ for some $c_{\text{true}} \geq 0$.

The following example illustrates that mixtures of log-concave random variables may be log-concave. In this case, we may also identify the values c for which (2.3) holds.

Example 2.8. Let ϕ denote a d-dimensional standard Gaussian density. Then the mixture

$$f(x) = p\phi(x) + q\phi(x - \mu)$$

(where $q = 1 - p$ and $0 < p < 1$) is log-concave if and only if $\left\|\mu\right\|_2 \leq 2$. Further, the representation (2.3) holds if and only if

$$c \geq \max\left\{0, \frac{1}{4}\left\|\mu\right\|_2^2 - 1\right\}$$

Proof. A smooth function $g : \mathbb{R}^d \to \mathbb{R}$ is concave if and only if $-\nabla\nabla^T g(x)$ is positive semi-definite for all $x \in \mathbb{R}^d$ (Rockafellar, 1997, Theorem 4.5). In this case, recalling that

$$\nabla\phi(x) = -x\phi(x)$$

and

$$\nabla\nabla^T\phi(x) = (xx^T - I)\phi(x),$$

we see that

$$\nabla \log f(x) = \frac{-p\phi(x)x - q\phi(x-\mu)(x-\mu)}{f(x)}$$

and

$$\begin{aligned}\nabla\nabla^T \log f(x) &= \frac{-p^2\phi^2(x)I - q^2\phi^2(x-\mu)I + pq\phi(x)\phi(x-\mu)(\mu\mu^T - 2I)}{f^2(x)} \\ &= -h(x)\left(I - g(x)(\mu\mu^T - 2I) + g^2(x)I\right) \\ &= -h(x)\left((g(x)I - A)(g(x) - I - A)^T - AA^T + I\right)\end{aligned}$$

where

$$h(x) = \frac{p^2\phi^2(x)}{f^2(x)},$$
$$g(x) = \frac{q\phi(x-\mu)}{p\phi(x)}$$

and

$$A = \frac{1}{2}\mu\mu^T - I.$$

Note that h and g are both strictly positive functions, so $-\nabla\nabla^T \log f(x)$ will be positive definite if and only if

$$\begin{aligned}-(AA^T - I) &= I - \frac{1}{4}\left(\mu\mu^T - 2I\right)\left(\mu\mu^T - 2I\right)^T \\ &= (1 - \frac{1}{4}\left\|\mu\right\|_2^2)\mu\mu^T\end{aligned}$$

is positive definite, which occurs if and only if

$$1 - \frac{1}{4}\left\|\mu\right\|_2^2 \geq 0,$$

that is if $\left\|\mu\right\|_2 \leq 2$. The preceding calculation shows that the smallest value of c satisfying (2.3) is

$$\max\left\{0, \frac{1}{4}\left\|\mu\right\|_2^2 - 1\right\}.$$ □

2.2.7 Connection with unimodality

One application of log-concave densities is as a proxy for the class of unimodal densities, since this is a natural shape constraint, but the likelihood within this class is unbounded. We say a function $f : \mathbb{R} \to \mathbb{R}$ is unimodal if there exists a $\mu \in \mathbb{R}$ such that f is nondecreas-

ing on $(-\infty, \mu)$ and nonincreasing on (μ, ∞). We say a random variable is unimodal if its density is unimodal.

There is no universally agreed extension of this definition to $d > 1$ (see Dharmadhikari and Joag-Dev, 1988, Chapter 2 for various possibilities). Following Ibragimov (1956), we say a function $f : \mathbb{R}^d \to \mathbb{R}$ is unimodal if f is quasiconcave, that is, if for each $\alpha \in \mathbb{R}$, $\{x : f(x) \geq \alpha\}$ is a convex set. This corresponds to what Dharmadhikari and Joag-Dev (1988) calls convex unimodality. Note that this coincides with the definition above if $d = 1$.

Any log-concave density is unimodal according to this definition, since for $\alpha > 0$,

$$\{x : f(x) \geq \alpha\} = \{x : \log f(x) \geq \log(\alpha)\},$$

which is convex (Rockafellar, 1997, Theorem 4.6); however, the converse is false. For example, the Cauchy density

$$f(x) = \frac{1}{\pi(1 + x^2)}, x \in \mathbb{R}$$

is unimodal according to this definition. However,

$$\log f(x) = -\log(\pi(1 + x^2))$$

so that

$$\frac{d}{dx} \log f(x) = \frac{-2x}{1 + x^2}$$

and

$$\frac{d^2}{dx^2} \log f(x) = \frac{2x^2 - 2}{(1 + x^2)^2}.$$

This means

$$\frac{d^2}{dx^2} \log f(x) > 0 \text{ for } |x| > 1,$$

so $\log f$ is not concave (Rockafellar, 1997, Theorem 4.5).

If $d = 1$, log-concave densities can be characterized as follows (Ibragimov, 1956, discussed in Eggermont and LaRiccia (2001) and Barndorff-Nielsen (1978)).

Theorem 2.9 (Ibragimov, 1956)**.** *A density f on $\mathbb{R}$ is log-concave if and only if the convolution $f \star g$ is unimodal for any unimodal density g in $\mathbb{R}$.*

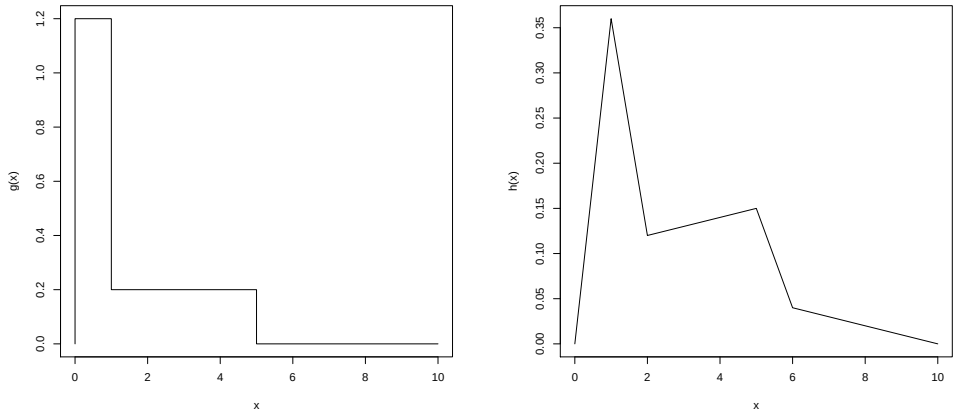
This has led to log-concave densities sometimes being referred to as strong unimodal densities (Ibragimov, 1956). Note that this gives an alternative proof of Corollary 2.2 for $d = 1$.

Closure under convolution does not hold for the class of unimodal distributions

as the following example, suggested by Feller (1971, page 168). A discussion, and further examples, may be found in Dharmadhikari and Joag-Dev (1988). Consider two independent random variables X_1 and X_2, both with density

$$g(x) = \frac{1}{2}\mathbb{1}_{[0,1]}(x) + \frac{1}{2}\mathbb{1}_{[0,5]}(x).$$

This is unimodal (as can be seen from Figure 2.1(a), displaying this density). However, the density h of the random variable $X_1 + X_2$, shown in Figure 2.1(b), is not unimodal.



(a) Density of X_1 and X_2 (independent)

(b) Density of $X_1 + X_2$

Figure 2.1: Example showing the sum of independent unimodal random variables is not necessarily unimodal.

2.3 Applications of log-concave densities

Applications of log-concavity have so far focused on the univariate setting. Useful general references include Bagnoli and Bergstrom (2005), An (1998) and An (1995).

If $d = 1$, a log-concave density implies an increasing hazard function (the "new is better than used" property, An, 1998), and this, combined with the flexibility of this class, has made the class of log-concave densities an important tool for reliability theory (Barlow and Proschan, 1975). An (1995) also introduced robust tests for this property.

There has also been interest in the field of econometrics (An, 1998; Bagnoli and Bergstrom, 2005). Log-concavity of the generalized extreme value distribution and the generalized Pareto distribution for important parameter values is demonstrated in Müller and Rufibach (2007). Müller and Rufibach (2009) suggest replacing the empirical distribution function with an integrated maximum likelihood density estimate to obtain

smooth tail index estimates. Log-concavity has also been used to improve convergence of MCMC algorithms (Brooks, 1998).

Log-concave densities are a subclass of unimodal densities, and may be a useful surrogate since estimation under the restriction of unimodality is difficult if the mode is unknown (Birgé, 1997; Walther, 2008). Using the mode of the log-concave maximum likelihood estimator as an estimator of the mode of a unimodal distribution has been investigated in Balabdaoui et al. (2009).

In the multivariate setting, this is a flexible model, containing many common parametric families. As we have seen, the class has desirable theoretical properties, for example closure under convolution and the taking of weak limits, the preservation of log-concavity for marginal and conditional densities, and the preservation of log-concavity under affine transformation (An, 1998). Unlike, for example, a Gaussian model, a log-concave density can incorporate skewness or other asymmetry.

There are several fast algorithms available to compute the log-concave maximum likelihood estimator for the univariate case (Dümbgen et al., 2007; Rufibach, 2007). The lack of algorithms to compute the estimator for $d > 1$ has so far held back applications for multivariate data. We go some way towards addressing this problem.

This is, of course, still a restrictive family, being unimodal with exponentially light tails. However, the extension to mixtures discussed in Section 4.4 broaden the appeal of this method.

2.4 Existence and uniqueness of maximum likelihood estimator

In this section, we demonstrate the existence and uniqueness of the log-concave maximum likelihood estimator

$$\widehat{f_n} = \arg\max_{f \in \mathscr{F}} \ell_n(f),$$

for data $\{X_1, \ldots, X_n\}$ in $\mathbb{R}^d$, where $\mathscr{F}$ denotes the class of all log-concave densities. The structure of the proof is similar to that used to treat other classes of shape-constrained maximum likelihood estimators (Grenander, 1956; Groeneboom et al., 2001b). Firstly, we show that likelihood maximization over the entire class is equivalent to likelihood maximization over a particular finite-dimensional subclass. Secondly, we show (for example, using the convexity of a modified objective function) the existence of a unique maximizer within this class.

The case $d = 1$ was treated in Dümbgen and Rufibach (2008); Pal et al. (2007); Rufibach (2006); Walther (2002). A proof using Theorem 2.9 was given in Eggermont and LaRiccia (2001, p. 423).

The proof given here gives valuable insight into the structure of the maximum likelihood estimator, which we use in Chapter 3. A summary of the relevant results and definitions from convex analysis can be found in Appendix A.

2.4.1 Main theorem and proof

In this section we prove our main result.

Theorem 2.10. *Suppose that $X_1, \ldots, X_n$ are iid observations taking values in some subset $S \subseteq \mathbb{R}^d$ (with $n > d$) drawn from a distribution with density f_0 with respect to Lebesgue measure. Then, with probability one, there is a unique log-concave maximum likelihood estimator*

$$\widehat{f}_n = \operatorname*{arg\,max}_{f \in \mathscr{F}} \frac{1}{n} \sum_{i=1}^{n} \log f(X_i)$$

where $\mathscr{F}$ is the class of all log-concave densities.

Proof. First, observe that, via an affine transformation taking S to a subset of $\mathbb{R}^d$ with the last $d - \dim(S)$ components being zero, we may restrict attention to the case $\dim S = d$.

With probability one, the observations $X_1, \ldots, X_n$ will be distinct, and

$$C_n = \operatorname{conv}\left(\{X_1, \ldots, X_n\}\right),$$

the convex hull of the observations, will have dimension d.

Now write

$$\mathscr{G} = \left\{ f : \mathbb{R}^d \to \mathbb{R} : f \text{ log-concave} \right\}$$

for the class of all log-concave functions. In order to optimize over $\mathscr{G}$ rather than $\mathscr{F}$, we add a Lagrangian term and minimize

$$\psi_n(f) = -\sum_{i=1}^{n} \frac{1}{n} \log f(X_i) + \int f \tag{2.4}$$

over $f \in \mathscr{G}$.

As a preliminary, for any vector $y \in \mathbb{R}^n$, let

$$\bar{h}_y(x) = \inf\left\{ h(x) : h \text{ concave and } h(X_i) \geq y_i \text{ for } i = 1, \ldots, n \right\} \tag{2.5}$$

and define

$$\mathscr{H} = \left\{ \bar{h}_y : y \in \mathbb{R}^n \right\}. \tag{2.6}$$

We prove this theorem by showing that, if a function $f \in \mathscr{G}$ minimizes ψ_n over $\mathscr{G}$, it also

has the following properties:

(1) $f(x) > 0$ for $x \in C_n$

(2) $f(x) = 0$ for $x \notin C_n$

(3) $\int f = 1$

(4) $\log f \in \mathscr{H}$, where $\mathscr{H}$ is as defined in (2.6)

(5) $\exists\, M > 0$ such that if $\max_i \bar{h}_y(X_i) > M$ then $\psi_n(\exp(\bar{h}_y)) > \psi_n(f)$

This suffices to demonstrate existence, since in this case we may instead minimize

$$\tau(y) = -\frac{1}{n}\sum_{i=1}^{n} \bar{h}_y(X_i) + \int \exp(\bar{h}_y(x))\,dx \tag{2.7}$$

over $y \in [-M, M]^n$. We have dropped the explicit dependence of the objective function τ on n for legibility, but it should be remembered that the objective function depends on the data throughout. The function τ is continuous, and $[-M, M]^n$ is a compact set, so the minimum must be attained by some $y \in [-M, M]^n$.

To show (1), note that if $f(x) = 0$ for some $x \in C_n$, then by Theorem 17.1 of Rockafellar (1997), we must have

$$-\infty = \log f(x) \geq \sum_{j=1}^{r} \log f(X_j)$$

for some $\lambda_j > 0$ with $\sum_j \lambda_j = 1$ and $r < d$. Thus we must have $f(X_i) = 0$ for some i, whence $\psi_n(f) = \infty$. For (2), note that

$$\begin{aligned}\psi_n(f\,\mathbb{1}_{C_n}) &= \psi_n(f) + \int (\mathbb{1}_{C_n}(x) - 1) f(x)\,dx \\ &\leq \psi_n(f),\end{aligned}$$

with equality if and only if the effective domain (Definition A.7) of $\log f$ is C_n.

For (3), suppose that $\int f = c$. By virtue of (1) and (2), we may restrict attention to the case $c \in (0, \infty)$, so let $\widetilde{f} = \frac{1}{c} f$. We then have

$$\begin{aligned}\psi_n(\widetilde{f}) &= \psi_n(f) - c + 1 + \log c \\ &\leq \psi_n(f)\end{aligned}$$

with equality if and only if $c = 1$. Thus f must be a density.

Now, for any $y \in \mathbb{R}^n$, let $\bar{h}_y$ be as defined in (2.5). We can think of the set $\mathscr{H}$ as the set of functions obtained by placing a pole of height y_i at observation X_i and stretching a piece of rubber over them. This is illustrated in Figure 2.2.

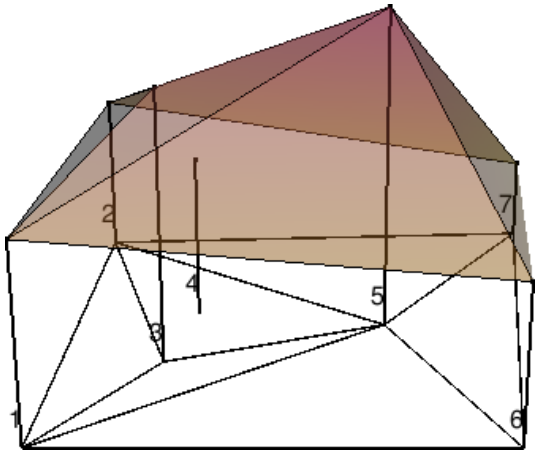


Figure 2.2: The structure of a typical element of $\mathscr{H}$. The surface is divided into simplices on which the function is piecewise affine.

Suppose that $\log f(X_i) = y_i$ for $i = 1, \ldots, n$ but $\log f \neq \bar{h}_y$. Clearly $\log f(x) \geq \bar{h}_y(x)$ for all $x \in \mathbb{R}^d$. If $\log f(x_0) > \bar{h}_y(x_0)$ for some x_0 in the interior of C_n, since f is continuous at x_0 (Rockafellar, 1997, Theorem 10.1) we have $\mu_d\left(\left\{\log f \neq \bar{h}_y\right\}\right) > 0$, where μ_d denotes Lebesgue measure on $\mathbb{R}^d$.This means that

$$\int f(x)\,dx > \int \exp(\bar{h}_y(x))\,dx,$$

so that

$$\psi_n(\exp(\bar{h}_y)) < \psi_n(f).$$

Thus $\log f = \bar{h}_y$ on the interior of C_n.

There remains the possibility that $\log f(x_0) > \bar{h}_y(x_0)$ for some x_0 on the boundary of C_n. But, writing $\mathrm{cl}(g)$ for the closure of a convex function (see Definition A.8),

$$\bar{h}_y = \mathrm{cl}(\bar{h}_y) = \mathrm{cl}(\log f) \geq \log f,$$

by Corollary 17.2.1 and Corollary 7.3.4 of Rockafellar (1997), so that $\log f$ is closed and $\log f \in \mathscr{H}$.

Finally, for (5), consider a log-concave function g with $\max_i \log g(X_i) = M$ and $\min_i \log g(X_i) = m$. Clearly as $m \to -\infty$, we have $\psi_n(g) \to \infty$. We show that, for M sufficiently large, we have $\psi_n(g) \geq \psi_n(f)$, where f satisfies (1)–(4).

Observe that if $x \in C_n$ and $\log g(X_k) = M$,

$$\begin{aligned}\log g\left(X_k + \frac{1}{M-m}(x - X_k)\right) &\geq \frac{1}{M-m}\log g(x) + \frac{M-m-1}{M-m}\log g(X_k)\\ &\geq \frac{m}{M-m} + \frac{(M-m-1)M}{M-m}\\ &= M-1,\end{aligned}$$

so that

$$\begin{aligned}\mu_d\left(\{x\colon \log g(x) \geq M-1\}\right) &\geq \mu_d\left(\left\{X_k + \frac{1}{M-m}\left(C_n - X_k\right)\right\}\right)\\ &= \frac{\mu_d(C_n)}{(M-m)^d.}\end{aligned}$$

Therefore

$$\int g \geq e^{M-1}\frac{\mu_d(C_n)}{(M-m)^d}.$$

This means that, for g to be a density, we must have

$$m \leq -\frac{1}{2}e^{(M-1)/d}\mu_d(C_n)^{1/d}.$$

In this case,

$$\begin{aligned}\psi_n(g) &\geq -\frac{M(n-1)}{n} + \frac{1}{2n}e^{(M-1)/d}\mu_d(C_n)^{1/d} - 1\\ &\to \infty \text{ as } M \to \infty.\end{aligned}$$

For uniqueness, observe that if f_1 and f_2 minimize ψ_n, and

$$g = \frac{f_1^{1/2}f_2^{1/2}}{\int f_1^{1/2}f_2^{1/2}},$$

we have

$$\psi_n(g) = \frac{1}{2}\psi_n(f_1) + \frac{1}{2}\psi_n(f_2) - \log\int f_1^{1/2}f_2^{1/2},$$

so that by Cauchy–Schwarz

$$\psi_n(g) \leq \psi_n(f_1) = \psi_n(f_2)$$

with equality if and only if $f_1 \equiv f_2$. □

2.4.2 Extension to binned observations and weighted log-likelihood

In practice our observations will be made only to finite precision so the observations will not necessarily be distinct, even if the underlying distribution is not degenerate. However, the same method of proof shows that, more generally, if $X_1, \ldots, X_n$ are distinct points in $\mathbb{R}^d$ and $w_1, \ldots, w_n$ are strictly positive weights satisfying $\sum_{i=1}^n w_i = 1$, then there is a unique log-concave density $\widehat{f}_n$ with $\log f \in \mathscr{H}$ which satisfies

$$\widehat{f}_n = \arg\max_{f \in \mathscr{F}} \sum_{i=1}^n w_i \log f(X_i). \tag{2.8}$$

Of course, this includes the case $w_i = \frac{1}{n}$ discussed above. However, this generalization allows us to extend this methodology to binned observations. In more detail, if $Y_1, \ldots, Y_m$ are independent and identically distributed according to a density f_0 and distinct binned values $X_1, \ldots, X_n$ are observed, we may construct a maximum likelihood problem of the form given in (2.8), setting

$$w_i = \frac{\#\text{ of times value } X_i \text{ is observed}}{m}.$$

We also use this formulation in the EM-style algorithm of Section 4.4.

2.5 Conclusion

In this chapter, we saw that log-concave densities have many desirable properties and that this model incorporates a wide range of common parametric distributions. We discussed some univariate applications, and motivated our use of this model for multivariate data. We proved the existence and uniqueness of a log-concave maximum likelihood estimator. Computing, using and understanding this estimator are addressed in the next three chapters.

3 Computation

3.1 Introduction

Having demonstrated the existence and uniqueness of a log-concave maximum likelihood estimator in Section 2.4.1, our next task is to formulate and implement algorithms to compute the estimator in practice.

In this chapter, $\mathscr{X} = \{X_1, \ldots, X_n\}$ denote an iid sample from a distribution on $\mathbb{R}^d$ with log-concave density f_0. The class of log-concave densities is denoted by $\mathscr{F}$, and the log-concave maximum likelihood estimator is denoted by $\widehat{f}_n$.

In light of step (4) in the proof of Theorem 2.10, we know that the likelihood maximization may be reformulated as

$$\text{minimize } \tau(y) \text{ subject to } y \in \mathbb{R}^n \tag{3.1}$$

where

$$\tau(y) = -\frac{1}{n}\sum_{i=1}^{n} \bar{h}_y(X_i) + \int \exp(\bar{h}_y(x))\,dx. \tag{3.2}$$

This appears simple, but there are two practical problems. First, in order to evaluate the objective function we must compute the integral

$$\int \exp(\bar{h}_y(x))\,dx. \tag{3.3}$$

Secondly, the objective function is not convex and therefore optimization is expensive. Indeed, preliminary experiments using general optimization techniques (such as simulated annealing) were prohibitively slow. In addition, the dimension of the optimization problem we must solve is n, rendering a straightforward approach infeasible even for moderate sample sizes. We therefore need to take advantage of the special structure of $\bar{h}_y$ in order to compute the maximum likelihood estimator for reasonably sized examples.

3.2 Evaluation of objective function

In this section we discuss the evaluation of (3.3) for arbitrary $y \in \mathbb{R}^n$. As a first step, we seek to understand the structure of $\bar{h}_y$ better. The following characterizations, which

follow from Corollaries 17.1.3 and 19.1.2 of Rockafellar (1997), are useful:

$$\bar{h}_y(x) = \sup\left\{\sum_{l=1}^{d+1}\lambda_l y_{j_l} : \sum_{l=1}^{d+1}\lambda_l X_{j_l} = x \text{ for some } \lambda_l \geq 0 \text{ with } \sum_{l=1}^{d+1}\lambda_l = 1\right\} \tag{3.4}$$

$$= \min\left\{b_1^T x - \beta_1, \ldots, b_m^T x - \beta_m\right\} - \delta_{C_n}(x) \tag{3.5}$$

for some $b_1, \ldots, b_m \in \mathbb{R}^d$ and $\beta_1, \ldots, \beta_m \in \mathbb{R}$, where

$$\delta_{C_n}(x) = \begin{cases} 0 & \text{if } x \in C_n \\ -\infty & \text{if } x \notin C_n. \end{cases}$$

In fact, from our construction of $\bar{h}_y$, the sets

$$S_j = \left\{i : b_j^T X_i - \beta_j = y_i\right\}, j = 1, \ldots, m$$

form a subdivision of $\mathscr{X}$. Roughly speaking, a subdivision partitions the space into simple shapes called polytopes; see Definitions A.19 and A.15. This is the subdivision induced by projecting the upper hull of the points

$$\{(X_1, y_1), \ldots, (X_n, y_n)\}$$

onto the first d components (see Figure 2.2). We write $\mathscr{S}(y)$ for the subdivision corresponding to a particular y. Note not all $X \in \mathscr{X}$ need be vertices of a polytope in $\mathscr{S}$ (for example, point 4 in Figure 2.2).

We may further refine this subdivision to form a triangulation (a partition into simplices, Definition A.20) $\mathscr{T}$ of $\mathscr{X}$ such that for each $T \in \mathscr{T}$, $\bar{h}_y$ is affine on $\text{conv}(T)$. We identify each such simplex with a $(d+1)$-tuple j such that

$$\text{conv}(T) = C_j = \text{conv}\left(\{X_{j_0}, \ldots, X_{j_d}\}\right),$$

and write $\mathscr{J}$ for the collection of all such j. In a slight abuse of notation, in the following we swap between the two representations of the triangulation.

We may then write

$$\bar{h}_y(x) = \sum_{j \in \mathscr{J}} (b_j^T x - \beta_j)\mathbb{1}_{C_j}(x) - \delta_{C_n}(x) \tag{3.6}$$

and therefore

$$\widehat{f}_n(x) = \sum_{j \in \mathscr{J}} \exp(b_j^T x - \beta_j) \mathbb{1}_{C_j}(x). \tag{3.7}$$

Given such a triangulation $\mathscr{J}$, for each $j \in \mathscr{J}$ corresponding to a simplex with vertices $\{X_{j_0}, \ldots, X_{j_d}\}$, we set

$$A_j = \begin{pmatrix} X_{j_1} - X_{j_0} | & \ldots & | X_{j_d} - X_{j_0} \end{pmatrix} \text{ and } a_j = X_{j_0}. \tag{3.8}$$

Then the map $w \mapsto A_j w + a_j$ sends the unit d-simplex T_d to the simplex C_j, where

$$T_d = \left\{ x \in [0, \infty)^d : \sum_{i=1}^d x_i \leq 1 \right\}.$$

Further, let $z_j \in \mathbb{R}^d$ have components $(y_{j_1} - y_{j_0}, \ldots, y_{j_d} - y_{j_0})$. After some calculation, we have

$$b_j = (A_j^T)^{-1} z_j \text{ and } \beta_j = a_j^T b_j - y_{j_0}. \tag{3.9}$$

In light of this, using a change of variables we may write

$$\begin{aligned}
\int \exp(\bar{h}_y(x))\, dx &= \sum_{j \in \mathscr{J}} \int_{C_j} \exp(b_j^T x - \beta_j)\, dx \\
&= \sum_{j \in \mathscr{J}} \left| A_j \right| e^{y_{j_0}} \int_{T_d} \exp(z_j^T w)\, dw \\
&= \sum_{j \in \mathscr{J}} \left| A_j \right| \int_{T_d} \exp(y_{j_0}(1 - w_1 - \ldots - w_d) + y_{j_1} w_1 + \ldots + y_{j_d} w_d)\, dw \\
&= \sum_{j \in \mathscr{J}} \left| A_j \right| \int_{T_d} \exp(y_{j_0} w_0 + \ldots + y_{j_d} w_d)\, dw,
\end{aligned}$$

where $w_0 = 1 - w_1 - \ldots - w_d$.

3.2.1 Basic properties of G_d

In this section, we discuss the basic properties of the function

$$G_d(y_0, \ldots, y_d) = \int_{T_d} \exp\left(\sum_{k=0}^d y_k w_k \right) dw \tag{3.10}$$

where $w_0 = 1 - w_1 - \ldots - w_d$. This discussion follows closely Cule and Dümbgen (2008). Some discussion of a similar function may also be found in Cule, Samworth, and Stewart

(2008). We define

$$G_0(y_0) = \exp(y_0), \tag{3.11}$$

and observe from direct computation that

$$G_1(y_0, y_1) = \begin{cases} \dfrac{\exp(y_1) - \exp(y_0)}{y_1 - y_0} & \text{if } y_0 \neq y_1 \\ \exp(y_0) & \text{if } y_0 = y_1. \end{cases} \tag{3.12}$$

Note that this is precisely the expression used for the one-dimensional case by Dümbgen et al. (2007).

It may be shown that

$$G_d(y_0, \ldots, y_d) = \frac{1}{d!}\mathbb{E}\left[\exp\left(\sum_{i=0}^{d} B_i y_i\right)\right] \tag{3.13}$$

where

$$B_i = \frac{E_i}{\sum_{i=0}^{d} E_i}$$

and E_i, $i = 0, \ldots, d$ are independent standard exponential random variables. Therefore G_d is symmetric in its arguments.

The key to computing G_d is the following proposition.

Proposition 3.1.

$$G_d(y_0, \ldots, y_d) = \begin{cases} \dfrac{G_{d-1}(y_1, y_2, \ldots, y_d) - G_{d-1}(y_0, y_2, \ldots, y_d)}{y_1 - y_0} & \text{if } y_0 \neq y_1 \\ \dfrac{\partial}{\partial y_1} G_{d-1}(y_1, \ldots, y_d) & \text{if } y_0 = y_1. \end{cases} \tag{3.14}$$

Proof. If $y_0 \neq y_1$, this follows by induction using the the base case (3.11), the recursive identity

$$G_d(y_0, \ldots, y_d) = \int_0^1 t^{d-1} G_{d-1}(t y_1, \ldots, t y_d) \exp((1-t)y_0)\, dt$$

and the symmetry of G_d in its arguments.

If $y_1 = y_0$, we use the first part of (3.14) and a limiting argument. More precisely,

$$\begin{aligned} \frac{\partial}{\partial y_1} G_{d-1}(y_1, \ldots, y_d) &= \lim_{t \to 0} \frac{G_{d-1}(y_1 + t, y_2, \ldots, y_d) - G_{d-1}(y_1, y_2, \ldots, y_d)}{t} \\ &= \lim_{t \to 0} G_d(y_1, y_1 + t, y_2, \ldots, y_d) \end{aligned} \tag{3.15}$$

$$= G_d(y_1, y_1, \ldots, y_d). \qquad \square$$

3.2.2 Taylor expansion of G_d

In this section, we assume that $y_0 \leq y_1 \leq \ldots \leq y_d$. This is no loss of generality because of the symmetry of G_d in its arguments, according to (3.13).

If $y_i - y_{i-1}$ is not too small for any $i = 1, \ldots, d$, we may compute $G_d(y_0, \ldots, y_d)$ recursively using the first part of (3.14) and the base case (3.11). For the other case, we derive a Taylor expansion in terms of $y_d - y_0$.

First of all, observe that for any $\widetilde{y} \in \mathbb{R}$,

$$G_d(y_0, \ldots, y_d) = \exp(\widetilde{y}) G_d(y_0 - \widetilde{y}, \ldots, y_d - \widetilde{y}).$$

Thus, letting $\bar{y} = \frac{1}{d+1} \sum_{i=0}^{d} y_i$ and $v_i = y_i - \bar{y}$,

$$G_d(y_0, \ldots, y_d) = \exp(\bar{y}) G_d(v_0, \ldots, v_d)$$

and $\sum_{i=0}^{d} v_i = 0$.

Then as $\|v\|_2 \to 0$,

$$d! G_d(v_0, \ldots, v_d) = 1 + \sum_{i=0}^{d} \mathbb{E}(B_i) v_i + \frac{1}{2} \sum_{i,j=0}^{d} \mathbb{E}(B_i B_j) v_i v_j + \frac{1}{6} \sum_{i,j,k=0}^{d} \mathbb{E}(B_i B_j B_k) v_i v_j v_k + O(\|v\|_2^4).$$

After some computation using the formulation in (3.13), we find that

$$G_d(y_0, \ldots, y_d) = \exp(\bar{y}) \left(\frac{1}{d!} + \frac{1}{2(d+2)!} \sum_{i=0}^{d} v_i^2 + \frac{1}{3(d+3)!} \sum_{i=0}^{d} v_i^3 + O(\|v\|_2^4) \right)$$

which allows us to compute appropriately if $\left| y_d - y_0 \right|$ is small (in our implementation, this expansion was used when $\left| y_d - y_0 \right| < 10^{-3}$).

Combining this with (3.14), we obtain Algorithm 3.1 for computing $G_d(y_0, \ldots, y_d)$ for $d \geq 1$. Note that the requirement that $y_0 \leq y_1 \leq \ldots \leq y_d$ is for notational convenience only, since G_d is symmetric in its arguments.

Require: d, dimension
Require: $y_0 \leq y_1 \leq \ldots \leq y_d$
Require: $\epsilon > 0$, tolerance
  if $d = 0$ **then**
    return $\exp(y_0)$
  else if $y_d - y_0 < \epsilon$ **then**
    $\bar{y} = \sum_{i=0}^{d} y_i/(d+1)$
    $v^2 = \frac{1}{2}\sum_{i=0}^{d}(y_i - \bar{y})^2$
    $v^3 = \frac{1}{3}\sum_{i=0}^{d}(y_i - \bar{y})^3$
    return $\exp(\bar{y})(\frac{1}{d!} + \frac{v^2}{(d+2)!} + \frac{v^3}{(d+3)!})$
  else
    return $\frac{G_{d-1}(y_0,\ldots,y_{d-1}) - G_{d-1}(y_1,\ldots,y_d)}{y_d - y_0}$

Algorithm 3.1: Computing $G_d(y_0, \ldots, y_d)$ if $y_0 \leq y_1 \leq \ldots \leq y_d$.

3.2.3 Finding an appropriate triangulation

Following Gelfand, Kapranov, and Zelevinsky (1994), for each triangulation $\mathscr{T}$ of $\mathscr{X}$ we define $h_{y,\mathscr{T}}$ to be the function obtained by linearly interpolating the points

$$\{(X_1, y_1), \ldots, (X_n, y_n)\}$$

over each simplex. For a subdivision $\mathscr{S}$, where possible we define $h_{y,\mathscr{S}}$ similarly, although this is not always well-defined.

Finding a triangulation $\mathscr{T}$ such that $\bar{h}_y = h_{y,\mathscr{T}}$ is fundamental to our calculation. In light of the view of $\mathscr{S}$ as the projection onto the first d components of the $d+1$-dimensional hull of the points $(X_1, y_1), \ldots, (X_n, y_n)$, however, it is straightforward. We augment the set

$$\{(X_1, y_1), \ldots, (X_n, y_n)\}$$

with auxiliary points

$$(X_1, y_{\min} - 1), \ldots, (X_n, y_{\min} - 1),$$

where

$$y_{\min} = \min_{i \in \{1,\ldots,n\}} y_i.$$

We then compute the convex hull of this extended set of points. This produces a list of $(d+1)$-tuples triangulating the surface of the convex hull, exactly as required. It is straightforward to remove those faces containing one of the points $\{(X_1, y_{\min} - 1), \ldots, (X_n, y_{\min} - 1)\}$, to be left with precisely the set of $(d+1)$-tuples $\mathscr{J}$. There are many possible convex hull algorithms; in our particular implementation, we use the

Quickhull algorithm (Barber, Dobkin, and Huhdanpaa, 1996; Grasman and Gramacy, 2008), described in detail in Section A.2.2.

It is conjectured by Barber et al. (1996, Conjecture 3.3) that the **Quickhull** algorithm has worst case running time of $O(n \log n + n^{\lfloor d/2 \rfloor})$. Note that, at least close to the optimum $\widehat{y}$, we will be close to the worst case because all points will lie on the surface of the convex hull.

For any subdivision $\mathscr{S}$, we set $C(\mathscr{S})$ to be the set of $y \in \mathbb{R}^n$ such that

1. $h_{y,\mathscr{S}}$ is well-defined and a concave function, and
2. for any i such that X_i is not a point in some $S \in \mathscr{S}$, $h_{y,\mathscr{S}}(X_i) \geq y_i$

We call triangulation *regular* if the interior of $C(\mathscr{T})$ is non-empty.

It is a classical fact of computational geometry (Gelfand et al., 1994, Proposition 1.5) that the cones $C(\mathscr{T})$ of all regular triangulations $\mathscr{T}$ of a point set $\mathscr{X}$ of size n in $\mathbb{R}^d$ form a complete polyhedral fan in $\mathbb{R}^n$ (that is, each $C(\mathscr{T})$ is a polyhedral cone, and the cones partition $\mathbb{R}^n$). Further, the relationship between triangulations $\mathscr{T}$ is intimately connected to the structure of this fan in $\mathbb{R}^n$. It turns out that two cones $C(\mathscr{T})$ and $C(\widetilde{\mathscr{T}})$ share an $(n-1)$-dimensional facet if and only if we may transform from one triangulation to the other by a simple, local geometric operation called a flip (Definition A.22). This is important for our later discussion of subgradients.

3.2.4 Possible improvement

Edelsbrunner and Shah (1996) describe an alternative method for constructing a regular triangulation based on flipping. We expect that, in moving around $\mathbb{R}^n$ in accordance with the subgradient algorithm, the triangulation from one stage to another will not necessarily change very much, and it should be possible to update the triangulation automatically.

Given any regular triangulation and a height vector y, we may update the triangulation using flips as described in Algorithm 3.2. This is a simple modification of the algorithm in Pournin and Liebling (2007) that allows for the insertion of new points, since in our case points may be removed from the triangulation and re-inserted at a later stage. Not surprisingly, this algorithm also has expected running time $O(n \log n + n^{\lfloor d/2 \rfloor})$. According to Pournin and Liebling (2007), a flip may be performed in constant time. Further, each face must be examined at least once. If $d = 2$, according to Euler's formula the number of faces is $O(n)$; this could therefore represent an improvement over the $O(n \log n)$ running time of **Quickhull** if no changes are made. However, for $d > 2$ there are no such bounds on the number of faces in a triangulation. Further, the diameter of the flip graph, which measures the maximum possible minimum distance between two

Require: $\mathscr{T}$, triangulation
Require: y, new height
Require: $\mathscr{E}$, list of edges
Require: $\mathscr{P}$, list of indices p such that $X_p \notin \mathscr{T}$
  for $p \in \mathscr{P}$ **do**
    if $y_p > h_{y,\mathscr{T}}(X_p)$ **then**
      Add p to triangulation
      Update $\mathscr{T}$ and $\mathscr{E}$
  while Some E is irregular and flippable **do**
    Flip E
    Update $\mathscr{T}$ and $\mathscr{E}$

Algorithm 3.2: Updating a regular triangulation.

triangulations, is not understood. Therefore this method has not been implemented as it is unlikely to lead to a significant improvement in the speed of our algorithm, despite its intuitive appeal.

3.3 Reformulation as a convex optimization problem

Now we are able to compute our objective function, we seek to reformulate (3.1) as a convex optimization problem in order to evaluate the estimator in practice. In order to do this, consider the (related) optimization problem

$$\text{minimize } \sigma(y) \text{ subject to } y \in \mathbb{R}^n \tag{3.16}$$

where

$$\sigma(y) = -\frac{1}{n}\sum_{i=1}^{n} y_i + \int \exp(\bar{h}_y(x))\,dx. \tag{3.17}$$

This is useful in light of the following result:

Theorem 3.2. *For $X_1, \ldots, X_n \in \mathbb{R}^d$, the function σ defined in* (3.17) *is convex, and has a unique minimum y^* satisfying*

$$\exp(\bar{h}_{y^*}) = \arg\max_{f \in \mathscr{F}} \frac{1}{n}\sum_{i=1}^{n} \log f(X_i) \tag{3.18}$$

where $\mathscr{F}$ is the class of log-concave densities on $\mathbb{R}^d$.

Proof. Note that, for $y, z \in \mathbb{R}^n$, $\lambda \in (0,1)$ and all $x \in \mathbb{R}^d$ we have

$$\bar{h}_{\lambda y + (1-\lambda)z}(x) \leq \lambda \bar{h}_y(x) + (1-\lambda)\bar{h}_z(x).$$

Therefore

$$\begin{aligned}\sigma(\lambda y+(1-\lambda)z) &= -\lambda\frac{1}{n}\sum_{i=1}^n y_i-(1-\lambda)\frac{1}{n}\sum_{i=1}^n z_i+\int \exp(\bar{h}_{\lambda y+(1-\lambda)z}(x))\,dx\\ &\le -\lambda\frac{1}{n}\sum_{i=1}^n y_i-(1-\lambda)\frac{1}{n}\sum_{i=1}^n z_i+\int \lambda\exp(\bar{h}_y(x))\,dx\\ &\qquad +\int(1-\lambda)\exp(\bar{h}_z(x))\,dx\\ &= \lambda\sigma(y)+(1-\lambda)\sigma(z),\end{aligned}$$

so that σ is a convex function.

To show (3.18), note that since

$$\sigma(y)=\tau(y)+\frac{1}{n}\sum_{i=1}^n\left(\bar{h}_y(X_i)-y_i\right)$$

and the second term on the right hand side is nonnegative, we have $\sigma(y)\ge\tau(y)$, and if

$$\widehat{y}\in\operatorname*{arg\,min}_{y\in\mathbb{R}^n}\tau(y)$$

we may set

$$y_i^*=\bar{h}_{\widehat{y}}(X_i)\text{ for }i=1,\dots,n.$$ □

3.4 Subgradients and subgradient methods

As shown in Section 3.3, σ is a convex function. For smooth convex functions, there are many optimization techniques available (Boyd and Vandenberghe, 2004, Chapter 11, and the references therein). However, as we will see, our objective function σ is not smooth, so these techniques are not appropriate here. However, we may extend the principle of descent methods to those based on subgradients (Shor, 1985), which enable us to calculate the log-concave maximum likelihood estimator.

3.4.1 Subgradients

A subgradient of any function $g:\mathbb{R}^d\to\mathbb{R}$ at a point y is any vector $\partial g(y)$ such that, for all $z\in\mathbb{R}^d$,

$$g(z)\ge g(y)+\partial g(y)^T(z-y)$$

(Rockafellar, 1997). At points where g is differentiable, there is exactly one subgradient (the derivative) but in general there may be many or no subgradients. However,

convex functions are always subdifferentiable in their relative interior (Definition A.14) (Rockafellar, 1997, Theorem 23.3).

3.4.2 Computation of $\partial\sigma$

In this section, we prove that σ is differentiable at y if and only if $\mathscr{S}(y)$ is a triangulation. Further, we provide an explicit formula for $\partial\sigma(y)$.

Case 1: $\mathscr{S}(y)$ is a triangulation

In this case, $C(\mathscr{S}(y))$ is full-dimensional (Gelfand, Kapranov, and Zelevinsky, 1990) and y lies in the interior of this cone, so for any $z \in \mathbb{R}^n$ and for sufficiently small $t > 0$ we have $\mathscr{S}(y + tz) = \mathscr{S}(y)$.

By Theorem 25.2 of Rockafellar (1997), in order to prove that σ is differentiable at y it suffices to show that all the partial derivatives $\partial_i\sigma(y)$ exist. Then the (sub)gradient is given by the vector

$$\big(\partial_1\sigma(y), \ldots, \partial_n\sigma(y)\big)$$

From the expression (3.10) we see that

$$\partial_i\sigma(y) = -\frac{1}{n} + \sum_{j\in\mathscr{J}_i} \left|A_j\right| \frac{\partial}{\partial y_i} G_d(y_{j_0}, \ldots, y_{j_d})$$

where $\mathscr{J}_i = \{j \in \mathscr{J} : j_l = i \text{ for some } l\}$. Therefore, in order to compute subgradients of σ we must compute partial derivatives of G_d.

From (3.15)

$$\frac{\partial G_d(y_0, \ldots, y_d)}{\partial y_k} = G_{d+1}(y_0, \ldots, y_k, y_k, \ldots, y_d),$$

and therefore

$$\partial_i\sigma(y) = -\frac{1}{n} + \sum_{j\in\mathscr{J}_i} \left|A_j\right| G_{d+1}(y_i, y_{j_0}, \ldots, y_{j_d}).$$

This computation may be done in a numerically stable way using Algorithm 3.1.

Case 2: $\mathscr{S}(y)$ not a triangulation

If $\mathscr{S}(y)$ is not a triangulation, the situation is a little more complicated. Since σ is a convex function, directional derivatives

$$\sigma'(y; z) = \lim_{t\downarrow 0} \frac{\sigma(y + tz) - \sigma(y)}{t}$$

exist (Rockafellar, 1997, Theorem 23.1). Moreover, if the function is differentiable, we must have

$$\sigma'(y;z) = -\sigma'(y;-z)$$

for all unit vectors $z \in \mathbb{R}^n$.

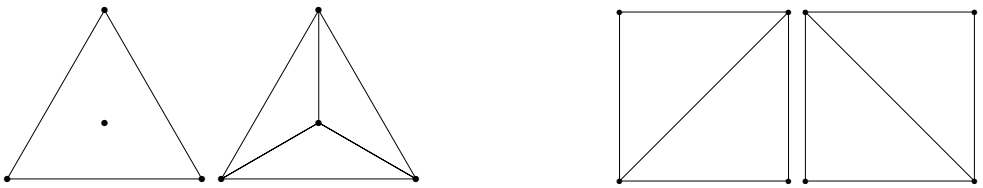
We show that

$$\sigma'(y, e_i) + \sigma'(y, -e_i) > 0 \tag{3.19}$$

for some $i \in \{1,\dots,n\}$, thus showing that σ is not differentiable at y. Intuitively, this is because in this case y lies on the boundary of more than one cone $C(\mathscr{T})$, and so in general $\mathscr{S}(y+tz)$ and $\mathscr{S}(y-tz)$ may be different, even for arbitrarily small values of t.

For notational simplicity, we consider the case where $\mathscr{S}(y)$ contains exactly one polytope V with exactly $d+2$ vertices. The extension to multiple non-simplicial elements, or those with more than $d+2$ vertices, proceeds in exactly the same way.

By reordering if necessary, let $1,\dots,d+2$ be the indices of the non-simplicial d-dimensional face of $\mathscr{S}$. Note that, while we require that the face be d-dimensional, we do not require that the points $X_1,\dots,X_{d+2}$ be in general position. There are exactly 2 ways to triangulate $d+2$ points in $\mathbb{R}^d$ (Lawson, 1986), and therefore there are 2 corresponding refinements of $\mathscr{S}$. An example of possible refinements of 4 points in $\mathbb{R}^2$ is given in Figure 3.1.



(a) Case 1 (b) Case 2

Figure 3.1: Two ways of triangulating four points in $\mathbb{R}^2$.

For each $i = 1,\dots,d+2$ and sufficiently small $t > 0$, $\mathscr{S}(y+te_i)$ and $\mathscr{S}(y-te_i)$ are triangulations, with index sets identical apart from those containing only terms from $\{1,\dots,d+2\}$.

From the definition of $h_{y,\mathscr{T}}$ for a triangulation $\mathscr{T}$ given in Section 3.2.3,

$$\bar{h}_{y+te_i}(x) = \bar{h}_y(x) + t h_{e_i,\mathscr{S}(y+te_i)}$$

and

$$\bar{h}_{y-te_i}(x) = \bar{h}_y(x) + t h_{-e_i,\mathscr{S}(y-te_i)}$$

so that

$$\sigma'(y;e_i)+\sigma'(y,-e_i)=\int\left(h_{e_i,\mathscr{S}(y+te_i)}(x)-h_{-e_i,\mathscr{S}(y-te_i)}(x)\right)\exp(\bar{h}_y(x))\,dx.$$

Now from the way we have chosen $\mathscr{S}(y+te_i)$ and $\mathscr{S}(y-te_i)$, $h_{e_i,\mathscr{S}(y+te_i)}$ and $-h_{-e_i,\mathscr{S}(y-te_i)}$ correspond to the upper and lower hulls of the points

$$\{(X_1,0),\ldots,(X_i,1),\ldots,(X_{d+2},0)\},$$

so provided these points are in general position in $\mathbb{R}^{d+1}$, the difference will be strictly positive, and thus

$$\sigma'(y;e_i)+\sigma'(y,-e_i)>0.$$

Note that, even if the points $X_1,\ldots,X_{d+2}$ are not in general position in $\mathbb{R}^d$, since their convex hull is a d-dimensional face of $\mathscr{S}(y)$, for at least one $i\in\{1,\ldots,d+2\}$ the points

$$\{(X_1,0),\ldots,(X_i,1),\ldots,(X_{d+2},0)\}$$

will be in general position. Thus σ is not differentiable at y.

In order to find a subgradient, by Rockafellar (1997, Theorem 25.6), it suffices to show that, for every $\epsilon>0$, we can find a point $\tilde{y}\in\mathbb{R}^n$ satisfying $\left\|y-\tilde{y}\right\|_2<\epsilon$ such that σ is differentiable at $\tilde{y}$ and $\left\|\nabla\sigma(\tilde{y})-\partial\sigma(y)\right\|_2<\epsilon$. In light of the structure of the sets $C(\mathscr{S})$, this may be done by perturbing y in a direction z by an amount $t>0$ such that $\widetilde{\mathscr{S}}=\mathscr{S}(y+tz)$ is a triangulation that refines $\mathscr{S}(y)$, with index set $\widetilde{\mathscr{J}}$ say. Then, we may set

$$\partial\sigma_i(y)=-\frac{1}{n}+\sum_{j\in\widetilde{\mathscr{J}}:\,i\in j}\left|A_j\right|G_{d+1}(y_i,y_{j_0},\ldots,y_{j_d}).$$

In theory, it is necessary to check that the refinement of $\mathscr{S}(y)$ obtained by the **QuickHull** algorithm corresponds to a regular triangulation (that is, to $\mathscr{S}(y+tz)$ for some $t>0$ and $z\in\mathbb{R}^n$). In principle, the regularity of a triangulation may be checked using the simplex method to determine the feasibility of a solution to to a linear programming problem determined by each edge in the triangulation. In practice, this was not found to be necessary. This is similar to the experience of Pournin and Liebling (2007), who describe an incremental update procedure for triangulations and remark that, even though in principle their algorithm requires a regular triangulation at each step and their algorithm did not explicitly test this, in practice no problems occurred over billions of test cases.

3.4.3 Subgradient-based optimization methods

The following theorem is fundamental to the theory of the optimization of nonsmooth convex functions.

Theorem 3.3 (Shor, 1985)**.** *Let (h_i) be a positive sequence with $h_i \to 0$ as $i \to \infty$ and $\sum_{i=0}^{\infty} h_i = \infty$. Then, for any convex function f, the sequence generated by the formula*

$$y_{i+1} = y_i - h_i \frac{\partial f(y_i)}{\|\partial f(y_i)\|}$$

has the property that either there exists an i_0 and y^* such that $y_{i_0} = y^*$, or $y_i \to y^*$ and $\sigma(y_i) \to \sigma(y^*)$ as $i \to \infty$.

We seem to have considerable freedom in our choice of h. In practice, Shor recognised that, although appropriate choice of step size could improve the rate of convergence, with this method it would never be better than linear (Shor, 1985, Chapter 2). This contrasts with the quadratic convergence near the optimum for Newton's method for smooth convex functions (Boyd and Vandenberghe, 2004, Section 9.5). Slow convergence can be caused by, at each stage, taking a step in a direction nearly orthogonal to the direction towards the optimum, which means that simply adjusting the step size selection scheme will never provide significant improvements in convergence rate.

One solution suggested by Shor (1985, Chapter 3) is to attempt to shrink the angle between the subgradient and the direction towards the minimum through a (necessarily nonorthogonal) linear transformation, and perform the subgradient step in the transformed space. By analogy with Newton's method for smooth convex functions, an appropriate transformation would be some approximation to the inverse of the Hessian matrix at y^*. This is not possible for nonsmooth problems, because the inverse might not even exist (it does not exist at points at which the function is not differentiable, for example, which may include the optimum).

Instead, we perform a sequence of dilations in the direction of the difference between two successive subgradients, in the hope of improving the rate of convergence in the worst-case scenario of steps nearly orthogonal to the direction towards y^*. The theory of Shor (1985, Chapter 3) suggests that the convergence may be quadratic under fairly general restrictions. However, unlike the original subgradient method, no formal proof of convergence is available for this algorithm. It would in principle be possible to follow this with the subgradient algorithm described above in order to guarantee convergence. In practice this has not been found necessary.

3.4.4 Stopping criteria

We terminate our algorithm when all of the following criteria are met:

$$|y_i^{t+1} - y_i^t| \leq \delta |y_i^t| \text{ for } i = 1, \ldots, n \tag{3.20}$$

$$|\sigma(y^{t+1}) - \sigma(y^t)| \leq \epsilon |\sigma(y^t)| \tag{3.21}$$

$$\left| \int_{C_n} \exp\{\bar{h}_{y^t}(x)\}\, dx - 1 \right| \leq \eta \tag{3.22}$$

for some tolerances δ, ϵ and η that must be specified. The criteria (3.20) and (3.21) were suggested by Kappel and Kuntsevich (2000). The criterion (3.22) is based on the observation that the maximum likelihood estimator integrates to 1.

3.5 The special case $d = 1$

In this section we specifically consider the case $d = 1$, which has been considered in some detail by Rufibach (2007) and Dümbgen et al. (2007). Let $X_{(1)}, \ldots, X_{(n)}$ denote the order statistics of the dataset

$$\mathscr{X} = \{X_1, \ldots, X_n\} \subseteq \mathbb{R}.$$

Triangulations of $\mathbb{R}^1$ have a particularly simple structure, being a division into intervals. Moreover, we know that at the optimum

$$\widehat{y} = \operatorname*{arg\,min}_{y \in \mathbb{R}^n} \sigma(y)$$

we have $\bar{h}_{\widehat{y}}(X_i) = \widehat{y}_i$ for all $i = 1, \ldots, n$. We may explicitly characterize the collection of points $y \in \mathbb{R}^n$ such that $\bar{h}_y(X_i) = y_i$ for all i as

$$\mathscr{Z} = \left\{ y \in \mathbb{R}^n : \frac{y_{(i+1)} - y_{(i)}}{X_{(i+1)} - X_{(i)}} \leq \frac{y_{(i)} - y_{(i-1)}}{X_{(i)} - X_{(i-1)}} \text{ for } i = 2, \ldots, n \right\}. \tag{3.23}$$

This means that the likelihood maximization reduces to

$$\text{minimize } \sigma(y) \text{ subject to } y \in \mathscr{Z}.$$

This is optimization of a smooth function over a polyhedral convex cone; moreover, we may write the constraints $y \in \mathscr{Z}$ in a form to which active set techniques, as well as more general convex optimization methods, may be easily applied (Dümbgen et al., 2007;

Rufibach, 2007). For multivariate data we cannot write these restrictions explicitly, and hence these methods cannot be applied directly.

3.6 Other computational aspects

In this section we discuss some further computational methods that will be of use in Chapter 4.

3.6.1 Sampling from density

Recall from Section 3.2 that $\log \widehat{f}_n \in \mathscr{H}$, that is,

$$\log \widehat{f}_n(x) = \sum_{j \in \mathscr{J}} \exp(b_j^T x - \beta_j) \mathbb{1}_{C_j}(x)$$

for simplices C_j in some triangulation $\mathscr{J}$. Further, for each $j \in \mathscr{J}$ we may compute the quantity

$$\begin{aligned} q_j &= \mathbb{P}(X \in C_j) \\ &= \int_{C_j} \widehat{f}_n(x)\, dx \\ &= \left|A_j\right| G_d(y_{j_0}, \ldots, y_{j_d}) \end{aligned}$$

using Algorithm 3.1. The scheme described in Algorithm 3.3 may now be used to draw a sample from the estimated density $\widehat{f}_n$.

Proof that this produces a sample from $\widehat{f}_n$. Observe that, by symmetry, U is uniformly distributed on the unit simplex, so that $A_j U + \alpha_j$ is uniformly distributed on C_j. Then the density $\widehat{f}_n$ (restricted to the simplex j) is dominated by a scalar multiple of the uniform distribution on the simplex C_j. Standard rejection sampling principles give the required result. □

3.6.2 Evaluation of density

The representation (3.5) allows us to evaluate the density at arbitrary points. Of course, we could also use (3.7), but in this case we must locate which (if any) simplex the point belongs to. There are a variety of techniques for doing this in triangulations in $\mathbb{R}^2$, some of which generalize to $d > 2$. The most appropriate depends on the application, as there is a tradeoff between preprocessing, storage and lookup time.

Require: $q_j : j \in \mathscr{J}$
Require: $A_j : j \in \mathscr{J}$
Require: $\alpha_j : j \in \mathscr{J}$
  Draw $U_1, \ldots, U_d \sim \text{Unif}[0,1]$
  Set $U = (U_{(1)}, U_{(2)} - U_{(1)}, \ldots, U_{(d)} - U_{(d-1)})$
  Draw $W \sim \text{Unif}[0,1]$
  Select an $i \in \mathscr{J}$, choosing j with probability q_j
  if

$$W < \frac{\exp(U^T z_i)}{\max_{v \in \mathscr{T}_d} \exp(v^T z_i)}$$

  then
    return $X = A_i U + \alpha_i$
  else
    Repeat

Algorithm 3.3: Sampling from $\widehat{f}_n$.

In our case, however, the number of polytopes m in $\mathscr{S}(\log \widehat{f}_n(X_1), \ldots, \log \widehat{f}_n(X_n))$ is typically much smaller than the number of simplices in $\mathscr{J}$ (which is $O(n^{\lceil d/2 \rceil})$). This phenomenon has also been observed for $d = 1$ (Dümbgen and Rufibach, 2008). Therefore there is little to be gained from using the representation (3.7) together with a point location strategy over the simpler (3.5).

In order to quickly identify whether a point lies in C_n, we create a list of vectors $c_1, \ldots, c_K \in \mathbb{R}^d$ and values $\gamma_1, \ldots, \gamma_K \in \mathbb{R}$ such that $x \in C_n$ if and only if

$$c_k^T x - \gamma_k \leq 0 \text{ for all } k = 1, \ldots, K. \tag{3.24}$$

These equations correspond to the facets (Definition A.16) of C_n, and there are $O(n^{\lfloor d/2 \rfloor})$ of them, although typically far fewer (Seidel, 2004). We may therefore check easily whether a given point $w \in C_n$, and, if so, the representation (3.5) holds, and $\widehat{f}_n(w)$ may be computed in $O(m)$ time. If $w \notin C_n$ then $\widehat{f}_n(w) = 0$ (Section 2.4.1). This is summarised in the pseudocode in Algorithm 3.4.

3.6.3 Evaluation of marginal and conditional densities

Once we have estimated the density, it may be of interest to evaluate the marginal densities or conditional densities, for example for the purpose of visualising certain aspects of the density estimate when $d > 2$. As in Section 2.2.4, let $T = PX$, where P is the matrix of some orthogonal projection onto the k-dimensional subset T.

Dropping the subscript n for legibility, we denote the marginal density on T by $\widehat{f}_T$ and the conditional density by $\widehat{f}_{X|T}$.

Require: $c_1, \ldots, c_K$
Require: $\gamma_1, \ldots, \gamma_K$
Require: $b_1, \ldots, b_m$
Require: $\beta_1, \ldots, \beta_m$
Require: $w \in \mathbb{R}^d$
  if $c_k^T w - \gamma_k \leq 0$ for $k = 1, \ldots, K$ **then**
    return $\exp\left(\min_j \{b_j^T w - \beta_j\}\right)$
  else
    return 0

Algorithm 3.4: Evaluation of density $\widehat{f}_n$ at a point.

From the expression in (3.7), for any $t \in T$

$$\widehat{f}_T(t) = \sum_{j \in \mathscr{J}} \int_{C_j \cap \{x\colon Px = t\}} \exp(b_j^T x - \beta_j)\, dx$$

where integration is with respect to Lebesgue measure on $\{x\colon Px = t\}$.

In order to compute this integral, note that $C_j \cap \{x\colon Px = t\}$ will be either empty or a union of simplices. After triangulating the simplex appropriately (in our implementation we used the Delaunay triangulation (Lee, 2004)), we may use Algorithm 3.1 to compute the appropriate integral.

For the conditional density

$$\widehat{f}_{X|T}(x|t) = \frac{\widehat{f}_n(x) \mathbb{1}_{\{y\colon Py = t\}}(x)}{\widehat{f}_T(t)}$$

normalization using the previous expression is needed.

3.6.4 Extension to binned observations and weighted log-likelihood

We may extend all of the above methodology to minimize the function

$$\sigma_w(y) = -\sum_{i=1}^n w_i y_i + \int \exp(\bar{h}_y(x))\, dx,$$

where w_i are strictly positive and sum to 1. This easily seen to be equivalent to the weighted log-likelihood maximization of Section 2.4.2. This will be particularly useful when we come to use an EM-style algorithm in Section 4.4.

3.7 Running time

In this section, we discuss the running time of the algorithm discussed above for various problem sizes. We concentrate on small values of d because, although in principle this method is valid for arbitrary d, in practice the curse of dimensionality, as well as time constraints, make it less useful for d greater than about 4 (Scott, 1992, p.4-6).

Figure 3.2 illustrates the running times on a 1.86GHz/2GB RAM desktop PC (based on the average of 10 runs using $N_d(0, I)$ data for $d = 1, 2, 3, 4$). Figure 3.3(a) gives the number of function evaluations required for the computation, and Figure 3.3(b) the number of evaluations of σ required (each step may involve several function evaluations, so these numbers are different). The stopping criteria $\delta = 10^{-3}$, $\epsilon = 10^{-6}$, $\eta = 10^{-3}$ were used in all cases.

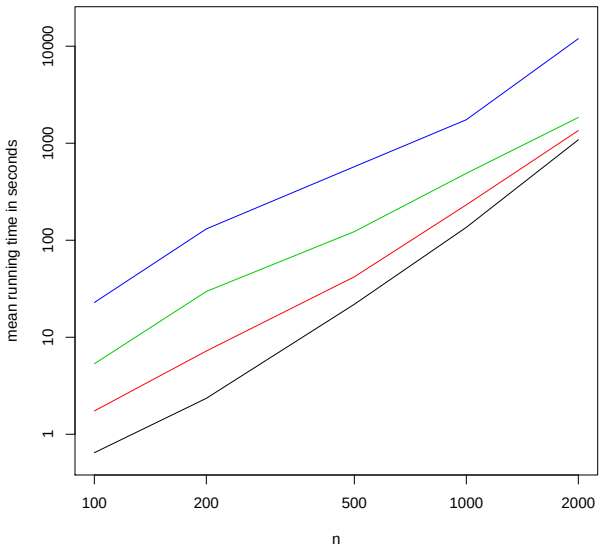
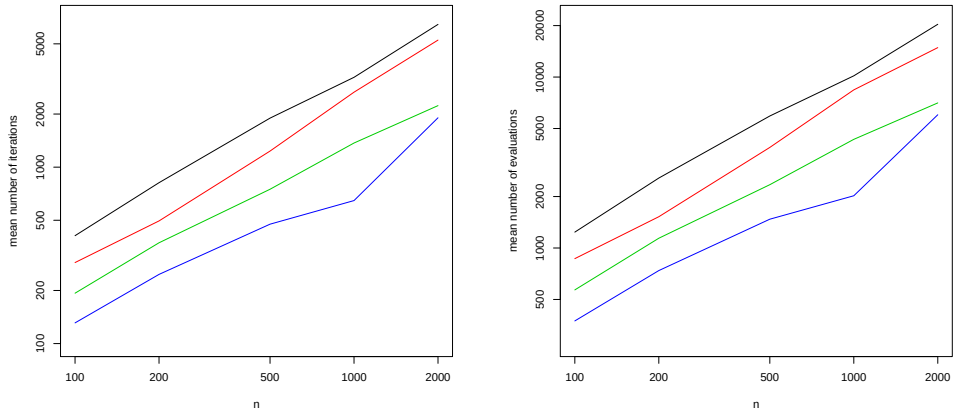


Figure 3.2: Running times for $d = 1$ (black), 2 (red), 3 (green) and 4 (blue) (average from 10 runs with $N_d(0, I)$ data).

Figure 3.2 shows that, as we might expect, the running time increases with both dimension and sample size. The running time appears to be polynomial in the sample size, but this is nevertheless slow for large datasets. Figure 3.3(a) and Figure 3.3(b) illustrate that the number of iterations and the number of function evaluations required decrease with dimension for a fixed sample size. This suggests that each step becomes progressively more expensive. This is not surprising considering the calculation of the objective function requires a convex hull computation at each step. Indeed, a more



(a) Iterations (b) Function evaluations

Figure 3.3: Number of iterations and number of function evaluations for $d = 1$ (black), 2 (red), 3 (blue) and 4(green) (average from 10 runs for $N_d(0, I)$ data).

detailed breakdown of the running time reveals that most computational effort is spent computing the convex hull. This suggests that one way to speed up the computation would be to use a more sophisticated algorithm for finding the appropriate triangulation of C_n, rather than simply computing the convex hull at each step.

For the one dimensional case, as discussed in Section 3.5, it is much faster to use the active set algorithm of Dümbgen et al. (2007). This takes under one second for the examples discussed here. However extension to $d > 1$ is complicated.

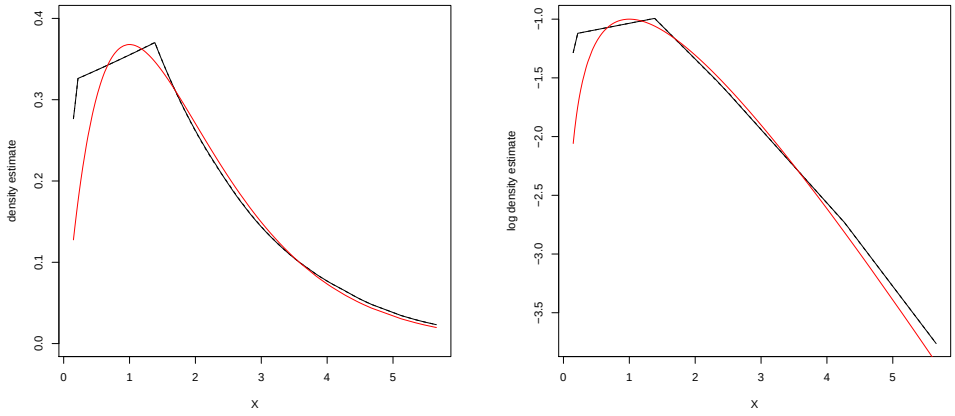
3.8 Examples

In this section we briefly illustrate the computational methods implemented in the package **LogConcDEAD** using simple examples. The code to produce these plots, which has been checked by R (R Development Core Team, 2008) using `Sweave` (Leisch, 2002), is given in Appendix B.

Example: $d = 1$

We begin by demonstrating that, if $d = 1$, our estimator is the same as that produced by the package **logcondens** (Rufibach and Dümbgen, 2006) which uses active set methods to compute the maximum likelihood estimator. We will do this using 200 points drawn from a Gamma$(2, 1)$ distrubtion.

As expected, the two methods produce the same estimate. They are plotted in



(a) Density (b) Log density

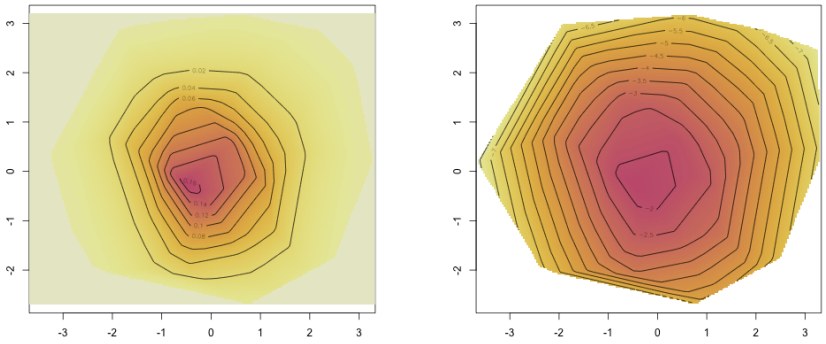
Figure 3.4: Estimated and true density and log density based on 200 points from a Gamma(2,1) distribution. The solid black line is the **LogConcDEAD** estimate, which coincides with the **logcondens** estimate, and the red line is the true value.

Figure 3.4(a). In addition, Figure 3.4(b) illustrates the structure of the log-concave maximum likelihood estimator: its logarithm is piecewise linear with changes of slope only at observation points.

Example: $d = 2$

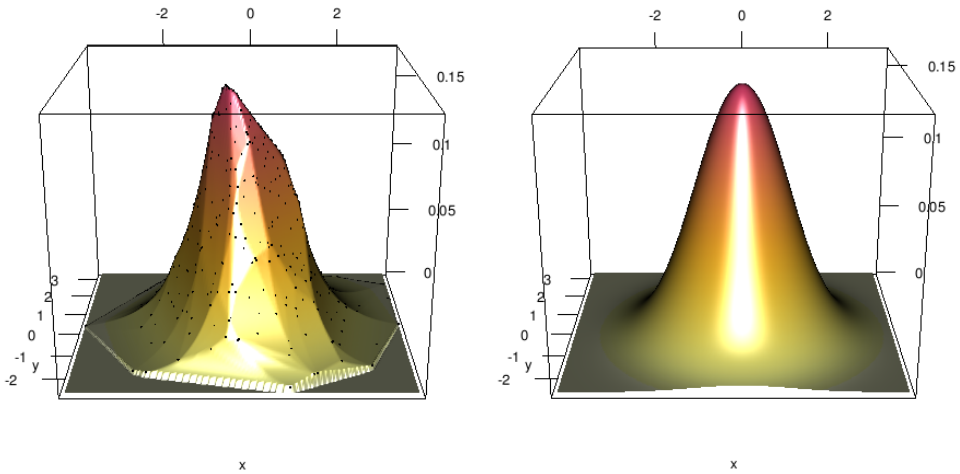
We will now illustrate the use of this method for multivariate data. We begin by generating 500 points from a $N_2(0, I)$ distribution and computing the log-concave maximum likelihood estimator. The resulting estimate is then used to produce the contour plots of the true and estimated log density in in Figure 3.5, the surface plot of the estimated density in Figure 3.6(a) and the surface plot of the log density in Figure 3.7(a). For comparison, the true density and log density are shown in Figure 3.6(b) and Figure 3.7(b) respectively.

These figures illustrate several important points about the structure of the maximum likelihood estimator. The density is supported on the convex hull of the dataset. This can be seen more clearly in Figure 3.8. From Figure 3.7(a), we can see the form of the log estimator: as if a rubber sheet has been stretched over "tent poles" placed at the observation points (see Figure 2.2). The data points used to generate this estimate are visible in Figure 3.6(a) and Figure 3.7(a) to highlight this point.



(a) Estimated density (b) Estimated log density

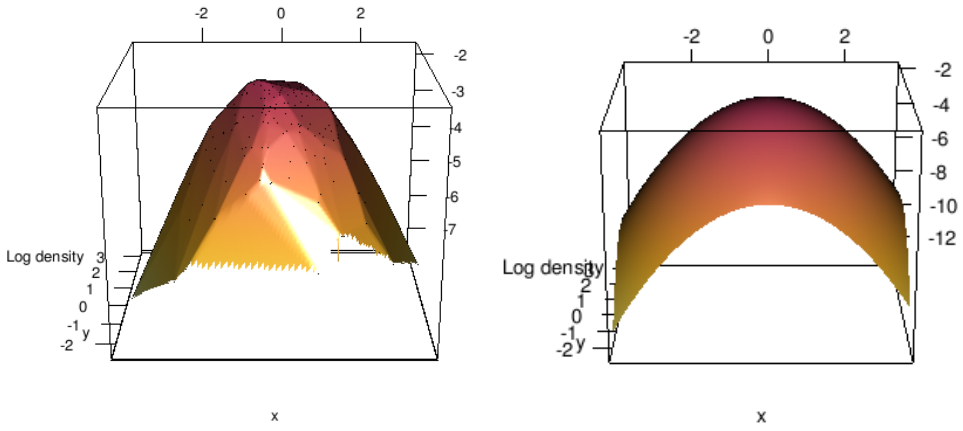
Figure 3.5: Contour plots of estimated density based on 500 points from a standard bivariate normal distribution.



(a) Estimated density (b) True density

Figure 3.6: Surface plots of estimated and true density based on 500 points from a standard bivariate normal distribution.

3.8 Examples



(a) Estimated log density

(b) True log density

Figure 3.7: Surface plots of estimated and true log density based on 500 points from a standard bivariate normal distribution.

Example: Binned data

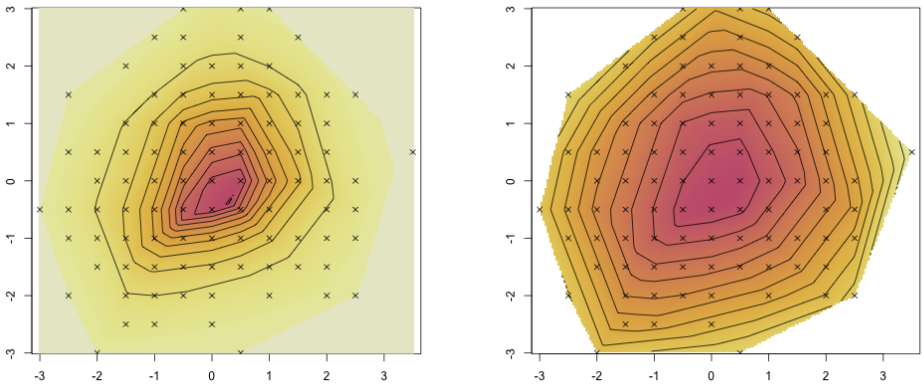
In this section, we illustrate the extension of this algorithm to binned data, as discussed in Section 3.6.4. As an example we use 500 points drawn from a $N_2(0, \Sigma)$ distribution, where

$$\Sigma = \begin{pmatrix} 1 & 0.2 \\ 0.2 & 1 \end{pmatrix}$$

These are then rounded to the nearest 0.5. In total there are then 94 distinct observations. In Figure 3.8 we show the density and log density estimates using a contour plot as before.

Example: $d = 3$

In our final example, we illustrate the use of the log-concave maximum likelihood for higher-dimensional data. In this case we integrate the density estimator to obtain an estimate of the marginal densities, as described in Section 3.6.3. In this example, we use 500 points from a 3-dimensional distribution with independent Gamma(2,1) components. We plot the estimated marginal distributions, computed using the method described in Section 3.6.3. The last panel in this plot shows the true density of each component.



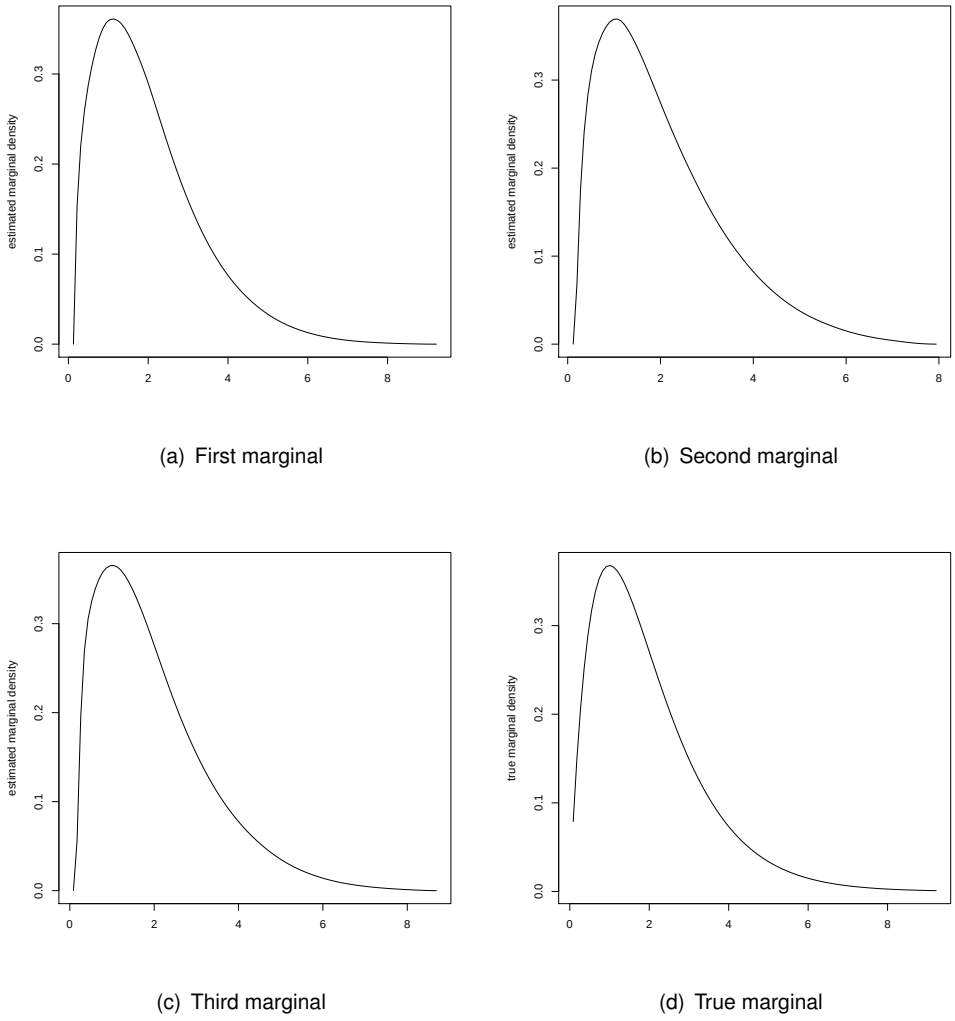
(a) Density (b) Log density

Figure 3.8: Estimated density and log density based on 500 points from a bivariate normal distribution in two dimensions (rounded to nearest 0.5).

3.9 Conclusion

In this chapter, we saw how a reformulation allows us to compute the log-concave maximum likelihood estimator using convex optimization techniques. We discussed some of the computational issues arising and proposed solutions. The proposed algorithm has been implemented, and several examples were used to illustrate important structural features of the density estimate. We touched on several computational issues of practical importance, including extension to weighted likelihood, sampling from the density estimate, evaluation of the density estimate, and computation of marginals and conditionals.

We discussed the running time of the existing algorithm. We acknowledge that this method is slower than existing methods for univariate data. However, as we saw, extension of existing methods to higher dimensions is extremely difficult, whereas our proposed technique works for arbitrary dimensional data.



(a) First marginal

(b) Second marginal

(c) Third marginal

(d) True marginal

Figure 3.9: Estimated and true marginal densities for 3-dimensional-data based on 500 points from a 3-dimensional Gamma distribution.

4 Inference

4.1 Introduction

Having proved the existence and uniqueness of a log-concave maximum likelihood estimator (Section 2.4) and discussed computational aspects (Chapter 3), we discuss its statistical uses.

We present a multivariate extension of the multiscale test suggested in Walther (2002). This enables us to assess the suitability of our assumption of log-concavity for a particular dataset. It is designed to detect mixing more sensitively than nonparametric procedures based on modality. We apply this to several simulated datasets and to one real dataset, for which the assumption of a single log-concave component is found to be inadequate.

Density estimation is often one stage in a more complicated statistical procedure. With this in mind, in Section 4.3 we discuss the use of the log-concave maximum likelihood estimator for plug-in estimation of statistical functionals. Three examples are given, demonstrating the strengths and weaknesses of this compared with alternative estimators. Where a kernel estimator was used, we used a 2-stage plug-in rule to choose the bandwidth. We also tried other selectors but, as discussed further in Section 5.5, their performance tended to be similar or worse, so they are not shown here.

As discussed in Chapter 2, an important extension of the class of log-concave densities is the class of finite mixtures of log-concave densities. In Section 4.4, we discuss fitting mixtures of this form where k is known using an EM-style algorithm. We fit this kind of mixture to the UK university ranking dataset discussed in Section 4.2.2. We discuss a natural clustering rule that arises from this kind of mixture, and apply this to a breast cancer dataset for which k is known. For this problem, we see a significant improvement over a Gaussian mixture.

4.1.1 Example densities

For our evaluation of functional estimation, we consider several example densities chosen to illustrate a range of features. These densities will also be considered in our convergence rate simulations in Chapter 5. For several densities, $d = 1, 2$ and 3 are considered to illustrate the effect of increasing dimension on the various quantities under

consideration. For $d = 2$, we also consider density G which has additional structure. We consider sample sizes $n = 100, 200, 500, 1000$ and 2000.

Here $N_d(\mu, \Sigma)$ denotes a d-dimensional Gaussian distribution with mean μ and covariance Σ. The densities are:

A $N_d(0, I)$.

B ($d > 1$ only) $N_d(0, \Sigma)$, where

$$\Sigma_{ij} = \begin{cases} 1 & \text{if } i = j \\ 0.2 & \text{if } i \neq j. \end{cases}$$

C Independent $\Gamma(2, 1)$ components.

D A mixture with each observation drawn from $N_d(0, I)$ with probability 0.6, and $N_d(\mu, I)$ with probability 0.4, where

$$\mu = \begin{cases} 1 & \text{if } d = 1 \\ (1, 0) & \text{if } d = 2 \\ (1, 0, 0) & \text{if } d = 3 \end{cases}$$

(this is log-concave; see Section 2.2.6).

E As for D, but with

$$\mu = \begin{cases} 2 & \text{if } d = 1 \\ (2, 0) & \text{if } d = 2 \\ (2, 0, 0) & \text{if } d = 3 \end{cases}$$

(this is log-concave; see Section 2.2.6).

F As for D, but with

$$\mu = \begin{cases} 3 & \text{if } d = 1 \\ (3, 0) & \text{if } d = 2 \\ (3, 0, 0) & \text{if } d = 3 \end{cases}$$

(this is not log-concave; see Section 2.2.6).

G ($d = 2$ only) First component is Uniform(0,1); second component is Beta$(2, 4)$, independently.

These features of these densities are summarized in Table 4.1.

Density	Log-c	Depend	Norm	Mix	Skewed	Bded	Edge
A	Yes	No	Yes	No	No	No	No
B	Yes	Yes	Yes	No	No	No	No
C	Yes	No	No	No	Yes	Yes	No
D	Yes	No	Yes	Yes	No	No	No
E	Yes	No	Yes	Yes	No	No	Yes
F	No	No	Yes	Yes	No	No	NA
G	Yes	No	No	No	Yes	Yes	Yes

Table 4.1: Summary of features of example densities:
Log-c: Log-concave density.
Depend: Components are dependent.
Norm: Mixture of one or more Gaussian components.
Mix: Mixture of log-concave distributions.
Skewed: Nonzero skewness.
Bded: Support of the density is bounded in one or more directions.
Edge: Density is on boundary of space of log-concave densities.

4.2 Assessing log-concavity

An important prerequisite for using this density estimate is some means to assess the validity of the assumption of log-concavity for a particular dataset. Log-concavity is an inappropriate assumption for heavy-tailed or multimodal distributions, for example. This has been discussed in detail in Chapter 2.

Moreover, for some applications, it is desirable to test whether data comes from a single component distribution or from a mixture of two or more. Walther (2002) gives an example from flow cytometry. Many parametric tests for mixing are based on a specific exponential family. On the other hand, nonparametric tests are frequently less sensitive to mixing, depending on detecting multimodality. A detailed discussion can be found in Walther (2002).

In Walther (2002), a method is proposed for a multiscale test to detect the presence of mixing using a nonparametric test. In this section, we extend this to multivariate data and to assessing the suitability of a log-concave model, regardless of mixing.

Recall from Proposition 2.7 that a finite mixture of log-concave densities may be represented in the form

$$f(x) = \sum_{j=1}^{k} \pi_j f_j(x) = \exp(c\,\|x\|_2^2 + \varphi(x)), \tag{4.1}$$

where φ is concave and $c \geq 0$. In fact, the same representation holds for any smooth

density f such that

$$\sup_{z\in\mathbb{R}^d}\sup_{x\in\mathrm{dom}(f)} z^T\left(\nabla\nabla^T\log f(x)\right)z$$

is finite. This includes heavy tailed distributions such as the t distribution. An analogous condition may also be developed for the case when f is not smooth.

Recall that the set values of c for which this representation is valid is of the form $[c_{\mathrm{true}},\infty)$ for some $c_{\mathrm{true}}\in\mathbb{R}$ (Section 2.2.6). We aim to test the hypothesis $c_{\mathrm{true}}=0$. If we reject this hypothesis, either a mixture or a different model would be more appropriate. We note that the test cannot detect all mixing, because some mixtures of log-concave distributions are log-concave (and thus satisfy (4.1) with $c_{\mathrm{true}}=0$; see Section 2.2.6).

4.2.1 Description of log-concavity test

Given a sample $X_1,\ldots,X_n$, choose an equally spaced grid $c_0=0<c_1<\ldots<c_M=C$. Suggested values are $M=11$ and $C=3$. For each value of c, we compute the c-maximum likelihood estimator

$$\widehat{f}_n^c=\operatorname*{arg\,max}_{f\in\mathscr{F}_c}\frac{1}{n}\sum_{i=1}^n\log f(X_i),$$

where $\mathscr{F}_c$ is the set of densities of the form (4.1). This may be calculated using the subgradient method described in Chapter 3 after a suitable modification of the objective function and subgradient.

We then assess the deviation of $\log\widehat{f}_n^c$ from concavity for each value of c. For univariate data, since a function is concave if and only if its derivative is nonincreasing, Walther (2002) suggests the metric

$$d(g,\mathscr{M})=\inf_{m\in\mathscr{M}}\left\|(g-m)\widehat{f}_n^0\right\|_\infty,$$

where $\mathscr{M}$ is the class of all monotone decreasing functions, and g the left-hand derivative of $\log\widehat{f}_n^c$. The weight function $\widehat{f}_n^0$ is included to downweight the tails of the distribution of the distribution, which is desirable for applications.

This metric cannot be extended directly to multivariate data. However, an alternative is to measure the L_1 distance between $\log\widehat{f}_n^c$ and its least concave majorant (the smallest concave function exceeding $\log\widehat{f}_n^c$ everywhere), denoted by $\bar{h}$:

$$T_n(c)=\int\left(\bar{h}(x)-\log\widehat{f}_n^c(x)\right)\widehat{f}_n^0(x)\,dx.$$

This captures both long, shallow deviations from concavity and short, deep deviations.

In order to generate a reference distribution, we draw B bootstrap samples of n repli-

cates from $\widehat{f}_n^0$. For each sample $X_1^{*b}, \ldots, X_n^{*b}$ and each value $c = c_0, \ldots, c_M$, we compute the test statistic defined above, $T_n^{*b}(c)$. For each value of c, we compute $m(c)$ and $s(c)$, the sample mean and sample standard deviation respectively of $T_n^{*1}(c), \ldots, T_n^{*B}(c)$. We then standardize the statistics on each scale, computing

$$\widetilde{T}_n(c) = \frac{T_n(c) - m(c)}{s(c)}$$

and

$$\widetilde{T}_n^{*b}(c) = \frac{T_n^{*b}(c) - m(c)}{s(c)}$$

for each $c \in \mathscr{C}$ and $b = 1, \ldots, B$.

To perform the test we compute the (approximate) p-value

$$\frac{1}{B+1} \# \left\{ b : \max_{c \in \mathscr{C}} \widetilde{T}_n(c) > \max_{c \in \mathscr{C}} \widetilde{T}_n^{*b}(c) \right\}.$$

4.2.2 Examples

In this section we apply the log-concavity test to several simulated examples and to a real dataset.

Example: A single-peaked mixture

In order to illustrate this test, we used 500 samples from a mixture distribution. The first component was a mixture with density

$$0.5\,\phi_{0.25}(x) + 0.5\,\phi_5(x-2),$$

where ϕ_{σ^2} is the density of a $N(0, \sigma^2)$ random variable. The second component was an independent $\Gamma(2,1)$ random variable. This is the type of mixture that presents difficulties for both parametric tests (not being easy to capture with a single parametric family) and for many nonparametric tests (having a single peak). Figure 4.1 is a contour plot of this density. Mixing is not immediately apparent because of the combination of components with very different variances.

We performed the test described above using $B = 99$, $m = 11$ and $C = 3$. Before performing this test, both the data and the bootstrap samples were rescaled to have variance 1 in each dimension. This was done because c_{true} is not invariant under rescaling, so we wish to have all dimensions on the same scale before performing the test using our fixed grid $\mathscr{C}$. The resulting p-value was less than 0.01. Figure 4.2 shows the values of the test statistic for various values of c (on the standardized scale). The red

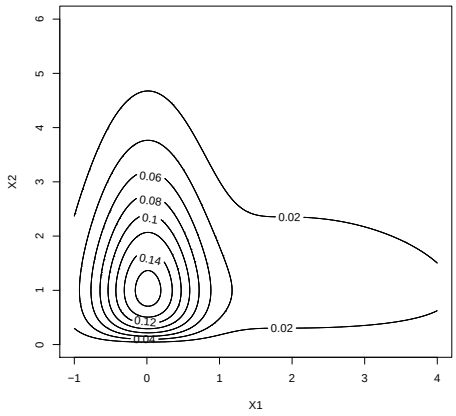


Figure 4.1: Contour plot for the mixture described in Section 4.2.2.

line corresponds to the data and the dashed black lines to the bootstrap samples.

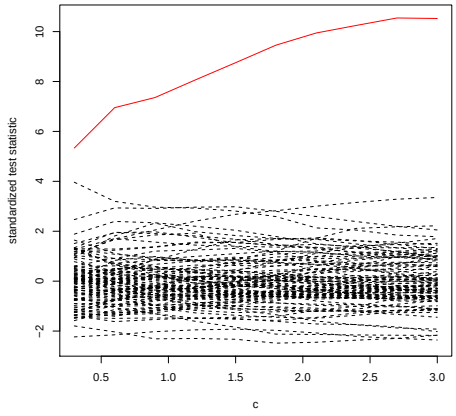


Figure 4.2: Assessing log-concavity for a single-peaked mixture as described in Section 4.2.2, $n = 500$. The red line is the test statistic $\widetilde{T}_n(c)$. The other lines are the bootstrap samples $\widetilde{T}_n^{*b}(c)$ for $b = 1, \ldots, 99$.

Example: A heavy-tailed distribution

For this example, we use a bivariate t_4 distribution with covariance matrix

$$\Sigma = \begin{pmatrix} 1 & 0.4 \\ 0.4 & 1 \end{pmatrix}.$$

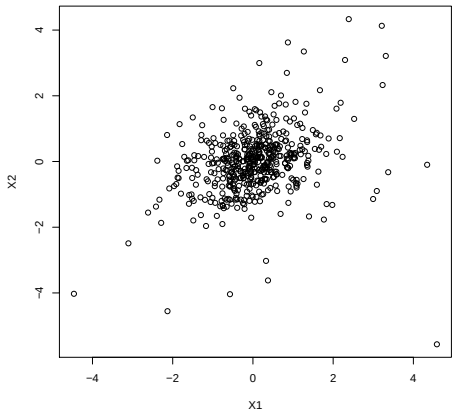


Figure 4.3: 500 points from a bivariate t_4 distribution, after rescaling to have unit variance in each dimension.

This is not log-concave due to its heavy tails. A plot of the data, after rescaling to have unit variance in each dimension, is shown in Figure 4.3. As the data appear unimodal, it is not immediately obvious whether our model is appropriate.

The test was performed, with rescaling, and the test statistics are shown in Figure 4.4. This shows that the log-concave model is not appropriate in this situation (p-value less than 0.01).

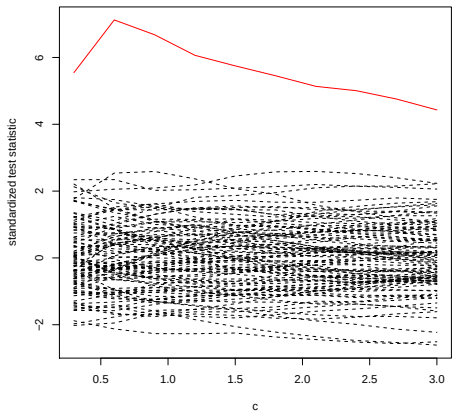


Figure 4.4: Assessing log-concavity: a heavy tailed distribution (t_4 distribution, $d = 2$, $n = 500$). The red line is the test statistic $\widetilde{T}_n(c)$. The other lines are the bootstrap samples $\widetilde{T}_n^{*b}(c)$ for $b = 1, \ldots, 99$.

4.2 Assessing log-concavity

Example: Log-concave distributions

As already mentioned, the test may not be able to detect all forms of mixing. To illustrate this, Figure 4.5 shows the result of performing this test for 500 samples drawn from density D. Once again, the data were rescaled to have variance 1 in each dimension. We see that the test statistic is not significant, as can be expected. We also performed this test for 500 $N_2(0, I)$ observations. The resulting values of $\widetilde{T}_n(c)$ and $\widetilde{T}_n^{*b}(c)$ for $b = 1, \ldots, 99$, are shown in Figure 4.6. From this we see that for this example the test does not produce a significant result. We can therefore be more confident in using a log-concave distribution to model the data.

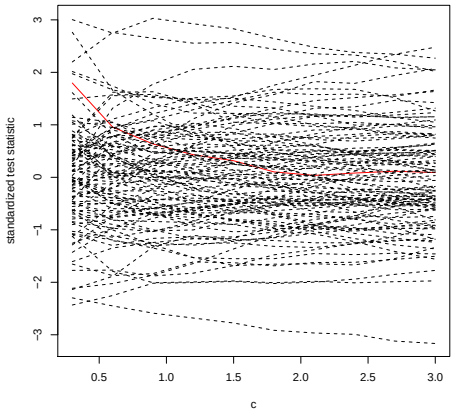


Figure 4.5: Assessing log-concavity: an undetectable mixture (density D, $d = 2$, $n = 500$). The red line is the test statistic $\widetilde{T}_n(c)$. The other lines are the bootstrap samples $\widetilde{T}_n^{*b}(c)$ for $b = 1, \ldots, 99$.

Example: Universities data

In this section we apply this test to a dataset extracted from the Times 2008 Good Universities Guide, April 28 2008 (The Times, 2008). This dataset gives several measures of the quality of UK higher education institutions, which are combined into an annual overall "ranking" of the universities by The Times newspaper. The measures are student satisfaction, research quality, services and facilities spend, entry standards, completion rate, percentage of students getting a good honours degree and graduate prospects. Since student satisfaction had some missing values, this was ignored for this analysis.

The data were rescaled and projected onto the first 2 principal components. These captured most (almost 80%) of the variability of the data. The first two scaled principal components are shown in Figure 4.7.

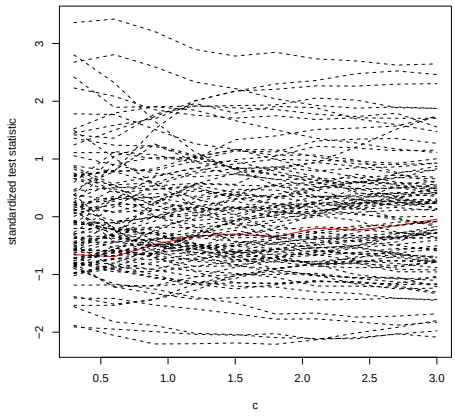


Figure 4.6: Assessing log-concavity: a log-concave distribution (density A, $d = 2$, $n = 500$). The red line is the test statistic $\widetilde{T}_n(c)$. The other lines are the bootstrap samples $\widetilde{T}_n^{*b}(c)$ for $b = 1, \ldots, 99$.

The test described above was performed to see whether it is reasonable to classify the universities into groups. No rescaling was used since the data had already been rescaled before performing the principal components analysis, and further rescaling could distort the relationship among the covariates. The results of this are shown in Figure 4.8. The p-value of less than 0.01 suggests that this dataset should be modelled as a mixture. We return to this in Section 4.4.3.

4.2.3 Further investigation

We have seen how we may in principle use a multivariate extension of the technique described in Walther (2002) to assess the suitability of a log-concave model for a particular dataset. For larger-scale investigation of this procedure, including verifying that the test has the required coverage under the null hypothesis and investigating the power of this test, more efficient algorithms for computing the c-MLE will be required. This is because the procedure described above is currently extremely computationally intensive. However, this nonparametric test for mixing shows great potential for detecting mixing in skewed multivariate samples, and may be of use in application areas such as screening and flow cytometry.

4.3 Functional estimation

As well as density estimation for visualization or clustering purposes, we may be interested in some statistical functional $\theta(F)$ depending on the underlying distribution

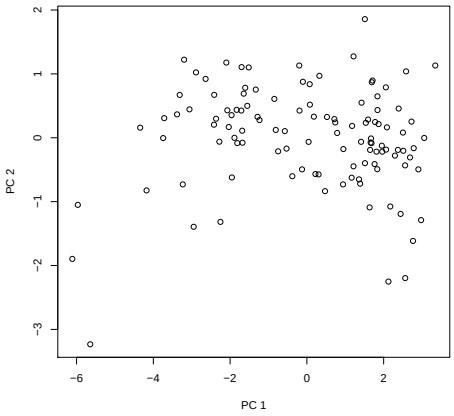


Figure 4.7: First two principal components (rescaled) of the universities data.

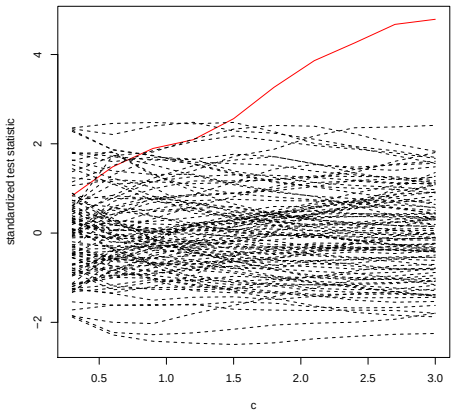


Figure 4.8: Assessing log-concavity: first two principal components of the universities data. The red line is the test statistic $\widetilde{T}_n(c)$. The other lines are the bootstrap samples $\widetilde{T}_n^{*b}(c)$ for $b = 1, \ldots, 99$.

function F, for example the mean and higher moments, level sets or highest density regions, quantiles, differential entropy or tail index. A common method of estimating such functionals is to "plug in" an estimate $\widehat{F}_n$ of the distribution function, that is, to estimate $\theta(F)$ by

$$\widehat{\theta}_n = \theta(\widehat{F}_n).$$

Clearly there is great flexibility in our choice of estimator $\widehat{F_n}$. If a continuous distribution is not required, one method is to plug-in the empirical distribution function. However, the functional may require a smooth density, for example the differential entropy or

highest density region. Even if this is not the case, Müller and Rufibach (2009) argue in the context of tail index estimation that, if additional information on the structure of the density is available, using the log-concave maximum likelihood estimate can lead to significantly reduced variability in some cases. This may lead to better overall estimation.

For a functional that may be written in the form

$$\theta(F) = \mathbb{E}[g(X)]$$

for some function g, the plug-in estimator may be written as

$$\widehat{\theta}_n = \int g(x)\widehat{f}_n(x)\,dx, \tag{4.2}$$

where $\widehat{f}_n$ is some estimate of the density.

Due to the form of the density given in (3.6), we may write the integral in (4.2) as

$$\widehat{\theta}_n = \sum_{j \in \mathscr{J}} \int_{C_j} \exp(b_j^T x - \beta_j) g(x)\,dx \tag{4.3}$$

with appropriate choices of $\{b_j\}$ and $\{\beta_j\}$ as described in Section 3.2.

If g is sufficiently smooth, (4.3) is relatively easy to compute. Sophisticated adaptive algorithms for integration over a simplex which make use of the invariant integration formulae detailed in Stroud (1971) and Grundmann and Möller (1978) are available (Genz, 1991; Genz and Cools, 2003). However, if g is sufficiently well-behaved and d not too large (smaller than 4, say), it is sufficient to use a fixed multivariate Gaussian quadrature method described in Press, Teukolsky, Vetterling, and Flannery (2007, Section 4.8).

If g is not a smooth function, this approach is not suitable. However, as we can sample easily and quickly from the density $\widehat{f}_n$ (see Section 3.6.1), a Monte Carlo approach also works. In more detail, we draw samples $Z_1, \ldots, Z_m$ from $\widehat{f}_n$ and set

$$\widetilde{\theta}_m = \frac{1}{m}\sum_{i=1}^{m} g(Z_i).$$

By the central limit theorem, for large m, conditional on the observed data,

$$\widetilde{\theta}_m \overset{\text{approx}}{\sim} N\left(\widehat{\theta}_n, \frac{\sigma_n^2}{m}\right), \tag{4.4}$$

where

$$\sigma_n^2 = \text{var}(g(Z_1)).$$

In order to construct an approximate confidence interval, we may approximate $\mathrm{var}(g(Z_1))$ by the sample variance of $g(Z_1), \ldots, g(Z_n)$ and use the approximation (4.4). This enables us to choose a large enough m to achieve our desired approximation to $\widehat{\theta}_n$.

In the remainder of this section, we consider three examples of functional estimation using the log-concave maximum likelihood estimator. These are estimation of covariance, estimation of differential entropy, and estimation of highest density regions.

4.3.1 Estimation of covariance

In general, plug-in estimation may be performed using numerical approximations to the integral

$$\int \widehat{f}_n(x) g(x)\, dx.$$

In certain cases, such as the moments, the structure of the maximum likelihood estimator means that this integral may be computed exactly. We use the notation of Section 3.2. Recall that the function G_d was defined in Section 3.2.1. Let x^r denote the rth component of a vector x. Further, let $w_0 = 1 - w_1 - \ldots - w_d$.

For the first moment (the mean), we have

$$\begin{aligned}
\int x^r \widehat{f}_n(x)\, dx &= \sum_{j \in \mathscr{J}} \int_{C_j} x^r \exp(b_j^T x - \beta_j)\, dx \\
&= \sum_{j \in \mathscr{J}} |A_j| \int_{T_d} (A_j w + \alpha_j)^r \exp(y_{j_0} w_0 + \ldots + y_{j_d} w_d)\, dw \\
&= \sum_{j \in \mathscr{J}} |A_j| \int_{T_d} \sum_{i=0}^{d} X_{j_i}^r w_i \exp(y_{j_0} w_0 + \ldots + y_{j_d} w_d)\, dw.
\end{aligned}$$

Observe that

$$\begin{aligned}
\frac{\partial}{\partial y_k} G_d(y_0, \ldots, y_d) &= \lim_{t \to 0} \frac{1}{t} \int_{T_d} \big[\exp(y_0 w_0 + \ldots + (y_k + t) w_k + \ldots + y_d w_d) \\
&\qquad - \exp(y_0 w_0 + \ldots + y_d w_d) \big]\, dw \\
&= \lim_{t \to 0} \frac{1}{t} \int_{T_d} \exp(y_0 w_0 + \ldots + y_d w_d)(\exp(t w_k) - 1)\, dw \\
&= \int_{T_d} w_k \exp(y_0 w_0 + \ldots + y_d w_d)\, dw.
\end{aligned}$$

Thus, using the relationship in (3.14),

$$\int x^r \widehat{f}_n(x)\,dx = \sum_{j\in\mathscr{J}} |A_j| \sum_{i=0}^{d} X_{j_i}^r G_{d+1}(y_{j_i}, y_{j_0}, \ldots, y_{j_d}).$$

This is easy to compute using Algorithm 3.1.

For the second moment, if $r \neq s$, by a similar calculation we have

$$\int x^r x^s \widehat{f}_n(x)\,dx = \sum_{j\in\mathscr{J}} |A_j| \sum_{i=0}^{d}\sum_{k=0}^{d} X_{j_i}^r X_{j_k}^s G_{d+2}(y_{j_i}, y_{j_k}, y_{j_0}, \ldots, y_{j_d}).$$

Further,

$$\int (x^r)^2 \widehat{f}_n(x)\,dx = 2\sum_{j\in\mathscr{J}} |A_j| \sum_{i=0}^{d}\sum_{k=0}^{d} X_{j_i}^r X_{j_k}^r G_{d+2}(y_{j_i}, y_{j_k}, y_{j_0}, \ldots, y_{j_d}).$$

Using the above expressions, we may easily compute the covariance matrix $\widehat{\Sigma}_n$ of the distribution corresponding to $\widehat{f}_n$. We illustrate the performance of this as a plug-in estimate of the distribution covariance. As our error criterion we use the mean squared error

$$\mathrm{MSE}\left(\widehat{\Sigma}_n, \Sigma\right) = \mathbb{E}\left[\frac{1}{d^2}\sum_{i,j=1}^{d}\left(\widehat{\Sigma}_n^{ij} - \Sigma^{ij}\right)^2\right],$$

where Σ^{ij} denotes the (i,j)th component of the covariance and $\widehat{\Sigma}_n^{ij}$ the (i,j)th component of the estimated covariance. This has an appealing decomposition into variance and squared bias terms

$$\mathrm{MSE}\left(\widehat{\Sigma}_n, \Sigma\right) = \frac{1}{d^2}\sum_{i,j=1}^{d}\mathbb{E}\left(\widehat{\Sigma}_n^{ij} - \mathbb{E}\widehat{\Sigma}_n^{ij}\right)^2 + \frac{1}{d^2}\sum_{i,j=1}^{d}\left(\mathbb{E}\widehat{\Sigma}_n^{ij} - \Sigma^{ij}\right)^2.$$

We have estimated this using 100 Monte Carlo samples for each of the densities, dimensions and sample sizes listed in Section 4.1.1. For comparison, in each case we also computed the sample covariance S. This has (i,j)th component

$$S_{ij} = \frac{1}{n}\sum_{k=1}^{n}(X_k^i - \bar{X}^i)(X_k^j - \bar{X}^j),$$

where

$$\bar{X}^i = \frac{1}{n}\sum_{k=1}^{n} X_k^i.$$

A small selection of the results are presented in Figure 4.9. The remainder were qualitatively similar.

In general, the two estimators had similar performance for log-concave densities. This is illustrated by Figures 4.9(a), 4.9(b) and 4.9(c) which show the results for densities A, B and C for $d = 1$, $d = 2$ and $d = 3$ respectively. This shows that the performance is similar for $d = 1$, but the log-concave estimator underperforms in higher dimensions. Other log-concave distributions performed similarly.

Although our estimator does not perform as well with respect to this criterion than an empirical estimate, it is worth more investigation to see why this is the case, since the errors are of the same order of magnitude for some examples. In Figure 4.10, we give a breakdown of the MSE for distribution C for $d = 2$, broken down into bias and variance terms. Here we see (since the black dashed line is almost directly underneath the red dashed line) that, for the empirical estimate, almost all the error is due to variance (not surprising as this estimator is asymptotically unbiased). For the log-concave maximum likelihood estimator, although the bias is much larger there is a more equal split into variance and bias terms. This more stable estimator may be preferred for some applications. We see this phenomenon, which has also been observed by Müller and Rufibach (2009), in Section 4.3.2.

4.3.2 Estimation of differential entropy

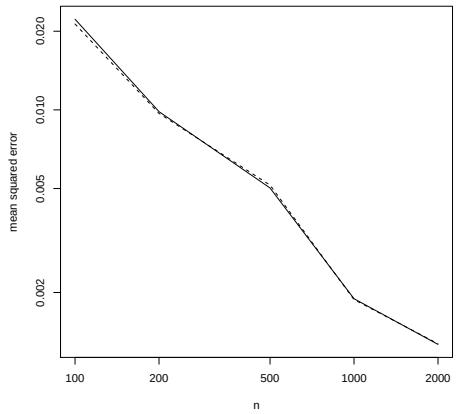
Differential entropy, introduced by Shannon (1948), is a common statistical functional. A review of techniques for estimation and applications is given in Beirlant, Dudewicz, Györfi, and van der Meulen (1997). One technique for estimation proposed in this article is the (exact) plug-in estimate

$$\widehat{\theta}_n = -\int \left(\log \widehat{f}_n(x)\right) \widehat{f}_n(x)\, dx, \tag{4.5}$$

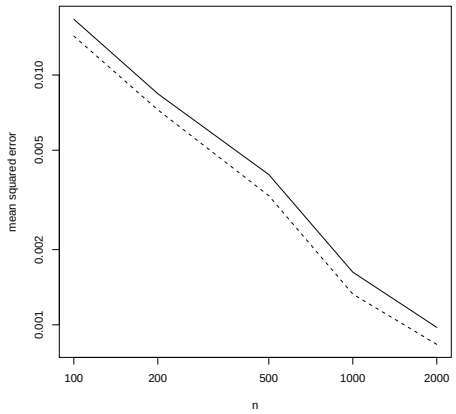
where $\widehat{f}_n$ is some estimate of the density f.

If we take $\widehat{f}_n$ to be the log-concave maximum likelihood estimator by a similar calculation to that in Section 2.2.5, we have

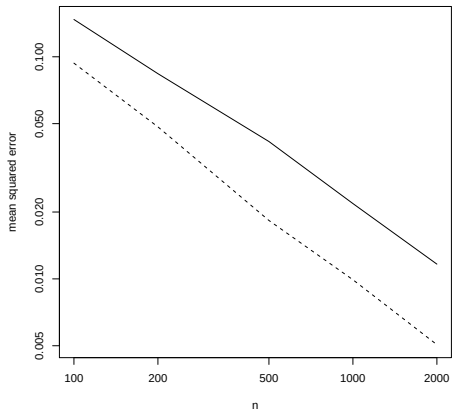
$$\begin{aligned}
\widehat{\theta}_n &= -\int \left(\log \widehat{f}_n(x)\right) \widehat{f}_n(x)\, dx \\
&= -\sum_{j \in \mathscr{J}} \int_{C_j} (b_j^T x - \beta_j) \exp(b_j^T x - \beta_j)\, dx \\
&= -\sum_{j \in \mathscr{J}} \int_{T_d} \left|A_j\right| (y_{j_0} w_0 + \ldots + y_{j_d} w_d) \exp\left(y_{j_0} w_0 + \ldots + y_{j_d} w_d\right)\, dw
\end{aligned}$$



(a) Density A, $d = 1$



(b) Density B, $d = 2$



(c) Density C, $d = 3$

Figure 4.9: MSE of covariance matrix estimates. The solid lines are the log-concave maximum likelihood estimate and the dashed lines are the sample covariance.

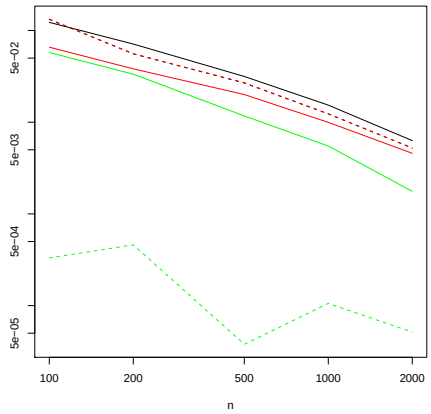


Figure 4.10: Bias-variance decomposition of the MSE of covariance matrix estimates for example distribution C and $d = 2$. The solid lines are the log-concave maximum likelihood estimate and the dashed lines are the kernel estimate. The black lines are the MSE, the green lines the squared bias and the red lines the variance. The red dashed line covers the black dashed line.

$$= -\sum_{j \in \mathscr{J}} \sum_{i=0}^{d} y_{j_i} \left|A_j\right| \int_{T_d} w_i \exp\left(y_{j_0} w_0 + \ldots + y_{j_d} w_d\right) \, dw$$
$$= -\sum_{j \in \mathscr{J}} \sum_{i=0}^{d} \left|A_j\right| y_{j_i} G_{d+1}(y_{j_i}, y_{j_0}, \ldots, y_{j_d}).$$

Here $w_0 = 1 - w_1 - \ldots - w_d$, G_d is as defined in Section 3.2.1 and the rest of the notation is as in Section 3.2.

For other density estimates, computing (4.5) is computationally intensive, especially for $d > 1$ or if the density estimate is multimodal. Therefore several alternatives have been proposed (Beirlant et al., 1997, and the references therein). One alternative is the so-called resubstitution estimate

$$\widetilde{\theta}_n = \frac{1}{n} \sum_{i=1}^{n} \log(\widehat{f}_n(X_i)). \tag{4.6}$$

For the log-concave maximum likelihood estimate, this gives similar results to the integrated plug-in estimate (typically less than 1% difference), so we do not expect this to make much difference.

The plug-in method was implemented for the densities, dimensions and sample sizes listed in Section 4.1. An estimate of the MSE, based on 100 Monte Carlo replications, is shown in Figure 4.11.

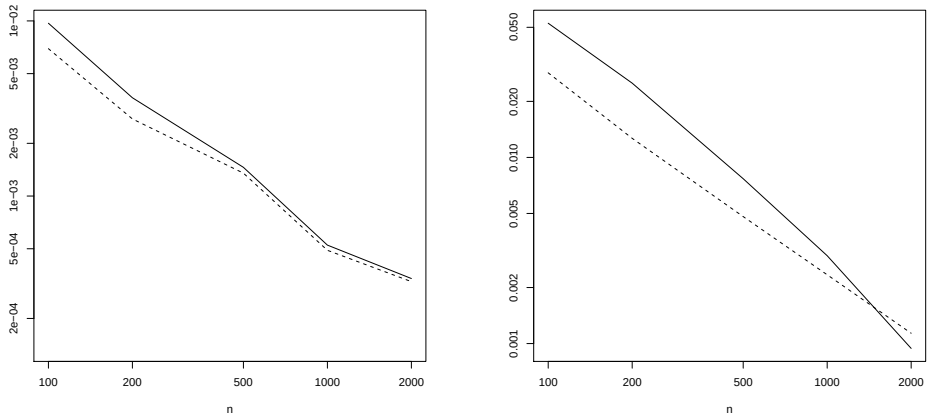
For comparison, we have also computed differential entropy estimates based on a (Gaussian) kernel estimate with bandwidth selected according to a 2-stage plug-in rule. For computational reasons we use the empirical plug-in estimator (4.6) rather than the integrated plug-in estimator (4.5). For the sample size we are considering this effect of using the empirical plug-in estimator rather than the integrated estimator is likely to be negligible. It is true that this kernel and bandwidth are known to be suboptimal for this estimation problem, and more sophisticated methods have been suggested (Beirlant et al., 1997). However, a detailed discussion of the optimal kernel and optimal bandwidth is far outside the scope of this thesis.

Selected results are illustrated in Figure 4.11. As before, a solid line denotes the log-concave maximum likelihood estimate, and a dashed line denotes the kernel estimate. From Figure 4.11(a), we see that for simple one-dimensional data the performance of the two methods is similar, with the kernel enjoying a slight advantage. In higher dimensions, illustrated by Figure 4.11(b) for density C and $d = 2$, the log-concave maximum likelihood estimator performs better than a kernel density estimator for larger sample sizes. The rate of decrease of the MSE is greater for the log-concave maximum likelihood estimator. This effect is particularly strong for our multivariate Beta distribution, shown in Figure 4.11(c) (density G, $d = 2$). This suggests that this is partly due to the well-known boundary bias of the kernel density estimator at the boundary of the support of the density. However, even in situations for which the kernel density estimator is ideally suited (illustrated in Figure 4.11(d), density B, $d = 3$), the log-concave estimator performs better for larger sample sizes.

In order to better understand this phenomenon, we display for two of our examples a decomposition of the estimated MSE into estimated squared bias and estimated variance terms. In Figure 4.12(a), we show this decomposition for density A and $d = 2$. For the log-concave maximum likelihood estimate, we see that the squared bias is much bigger than the variance, and decreases more rapidly with sample size. This helps to explain the better performance of the log-concave maximum likelihood estimate for larger sample sizes. For the kernel estimate, the squared bias and variance are of the same order of magnitude. A bias-variance breakdown for density G and $d = 2$ is shown in Figure 4.12(b). Here it is obvious that the main contributor to the poor performance of the kernel estimator is bias, likely to be due to boundary effects for this density.

4.3.3 Level sets, quantiles and highest density regions

Several authors have argued in favour of level sets and related quantities as quick and informative summaries of multivariate data (Hyndman, 1996, and references therein). A



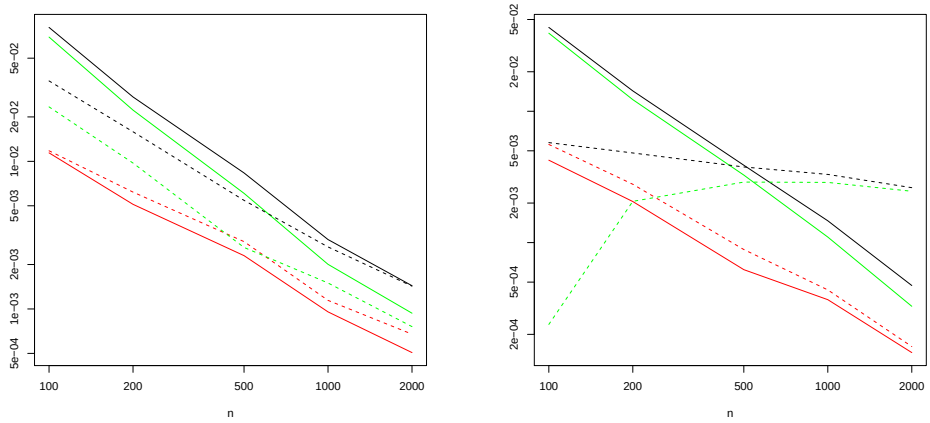
(a) Density A, $d = 1$

(b) Density B, $d = 2$

(c) Density G, $d = 2$

(d) Density F, $d = 3$

Figure 4.11: MSE of the differential entropy estimates. The solid lines are the log-concave maximum likelihood estimate and the dashed lines are the kernel estimate.



(a) Density A, $d = 2$ (b) Density G, $d = 2$

Figure 4.12: Bias-variance decomposition of the MSE of the differential entropy estimates for densities A and G and $d = 2$. The solid lines are the log-concave maximum likelihood estimate and the dashed lines are the kernel estimate. The black lines are the MSE, the green lines the squared bias and the red lines the variance.

level set of a density f at level y is defined as

$$R(f, y) = \{x : f(x) \geq y\}.$$

Setting f_α to be the largest constant such that

$$\mathbb{P}(X \in R(f, f_\alpha)) \geq 1 - \alpha,$$

$R(f, f_\alpha)$ is defined by Hyndman (1996) to be the $(1-\alpha)$ highest density region. This has the appealing properties that it is the smallest volume set with coverage $(1-\alpha)$, and that the density of any point outside this region is smaller than the density of any point inside this region. This summary can reveal interesting features in the density not shown by symmetric regions or quantiles, such as multimodality.

Note that f_α is the $(1-\alpha)$ quantile of the distribution $f(X)$, where X has density f. While in principle this may be computed directly, this is difficult, especially for multivariate data. However, this formulation immediately suggests a Monte Carlo approach to calculating f_α. We draw m independent samples $Z_1, \ldots, Z_m$ from f and estimate f_α with

$$\widetilde{f}_{\alpha,m} = f_{(\lfloor \alpha m \rfloor)},$$

where $f_{(j)}$ is the j/n sample quantile of $\{f(Z_1), \ldots, f(Z_m)\}$. To find a confidence interval

for this Monte Carlo estimate and the resulting region, we may generalize the results of Hyndman (1996, Section 3.3) to multivariate data. Define

$$S(y) = \{x : f(x) = y\}.$$

Provided f is reasonably behaved, this defines a surface in $\mathbb{R}^d$. Define

$$A(y) = \int_{R(f,y)} f(x)\,dx.$$

For small δ

$$A(y+\delta) = A(y) - \delta y \int_{S(y)} \left|\nabla f(x)^T n\right|^{-1} dS + O(\delta^2)$$

where n denotes the outward unit normal vector of $S(y)$ and the integral is over the surface $S(y)$. Then, since $\mathbb{P}(Y \le y) = 1 - A(y)$, the density of Y is given by

$$h(y) = y \int_{S(y)} \left|\nabla f(x)^T n\right|^{-1} dS. \tag{4.7}$$

This coincides precisely with the one-dimensional expression from this paper.

By standard arguments involving quantiles (Cox and Hinkley, 1979), $\widetilde{f}_{\alpha,m}$ has the approximate distribution

$$N\left(f_\alpha, \frac{\alpha(1-\alpha)}{mh^2(f_\alpha)}\right). \tag{4.8}$$

Finding the corresponding regions of $R(f, \widetilde{f}_{\alpha,m})$ is straightforward.

Since (4.7) depends on the unknown density, we are unable to compute a confidence interval directly. However, we may use a bootstrap approximation to the standard error in (4.8) to compute a bootstrap confidence interval (Efron and Tibshirani, 1993). This in turn leads to uniform confidence bands for $S(f_\alpha)$, the boundary of $R(f, f_\alpha)$. This is illustrated in Figure 4.13. Here, we used $m = 500$ observations from density G to estimate f_α for $\alpha = 0.25, 0.5$ and 0.75. The estimated highest density regions, together with uniform confidence bands, are illustrated. Using larger values of m obviously leads to narrower confidence bands. We found $m = 50000$ was fast to run and led to very narrow bands. This value of m was therefore used in our further experiments.

Of course, in general we are given a dataset and wish to estimate the highest density regions, which first requires a density estimate. From the outline above, by use of a sufficient number of Monte Carlo samples we may attain accuracy arbitrarily close to the true value for the estimated density. Moreover, the sampling may be done efficiently as described in Section 3.6.1. We used $m = 50000$ to produce the following examples;

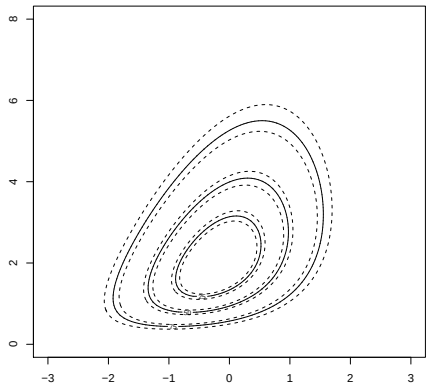


Figure 4.13: Bootstrap confidence bands for density G for the 25%, 50% and 75% highest density region using $m = 500$. The estimated contours are the solid lines.

the resulting confidence bands were so small they are not visible on the plots and so are omitted. The remaining variability is due to our approximation procedure for the density.

In practice, the density f will not be known, so even if we compute the confidence regions exactly, there will be added uncertainty in our confidence region estimate due to the replacement of f with $\widehat{f}_n$. Quantifying this remains an important open problem, and obviously the first steps will be consistency and convergence results for $\widehat{f}_n$. The difficulties of this are discussed in Section 5.3.

As an illustration, in Figure 4.14 we used 2000 observations from density G to compute the log-concave maximum likelihood estimate and a kernel density estimate. We then computed the 25%, 50% and 75% highest density regions for

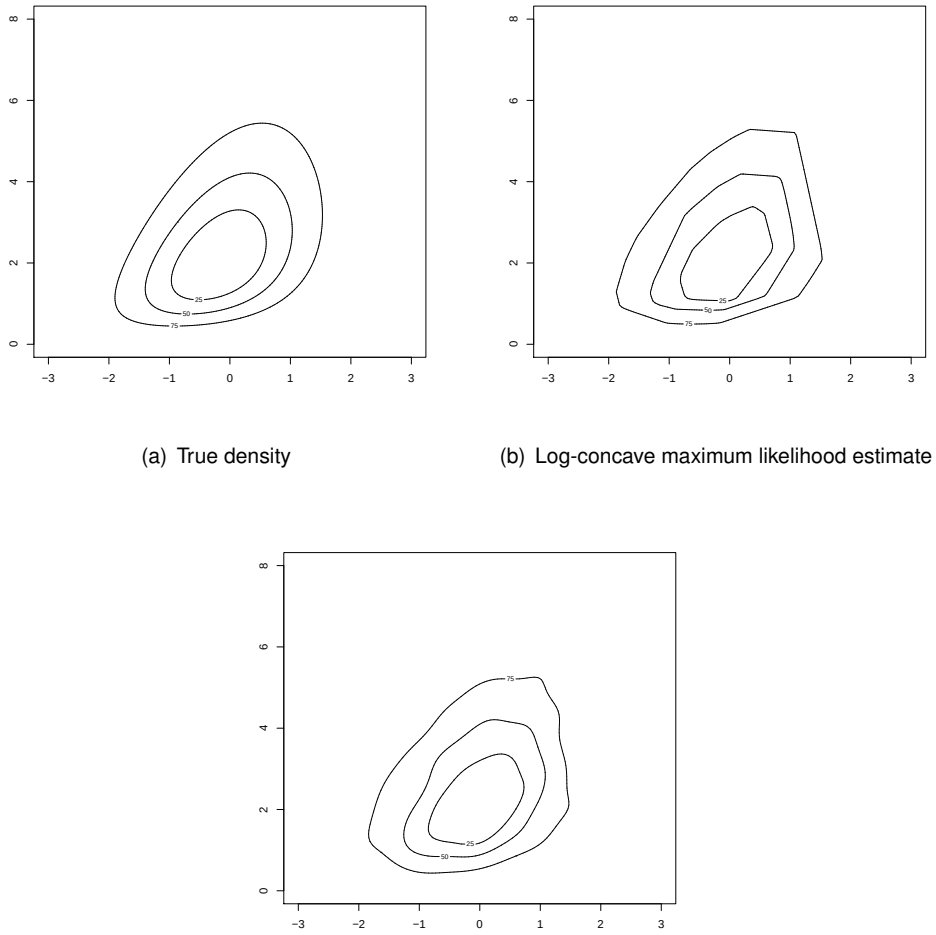
a. the true density,

b. the log-concave maximum likelihood estimate, and

c. a kernel density estimate using a 2-stage plug-in bandwidth.

Here, we see that the log-concave maximum likelihood estimate captures the shape of the density. Note that, since the true density is log-concave, the highest density regions are convex. This is captured by the log-concave maximum likelihood estimate, and not the kernel estimate.

To test this method, for the example densities listed in Section 4.1 we computed the highest density region based on the log-concave maximum likelihood estimate for $\alpha = 0.25, 0.5$ and 0.75. We also computed the highest density region based on a kernel



(a) True density

(b) Log-concave maximum likelihood estimate

(c) Kernel estimate

Figure 4.14: Highest density regions for density G and $n = 2000$. The highest density regions were computed using the Monte Carlo method described above with $m = 50000$.

density estimate with a Gaussian kernel and using a 2-stage plug-in rule to compute the bandwidth. These estimates were compared to the true highest density region using the error criterion

$$\int f(x)\Delta(x)\,dx,$$

where

$$\Delta(x) = \begin{cases} 1 & \text{if } x \text{ in exactly one of } R(f, f_\alpha) \text{ and } R(\widehat{f}_n, \widehat{f}_{n,\alpha}), \\ 0 & \text{else.} \end{cases}$$

This is motivated by a desire to treat both errors symmetrically, and to weight according to the true density.

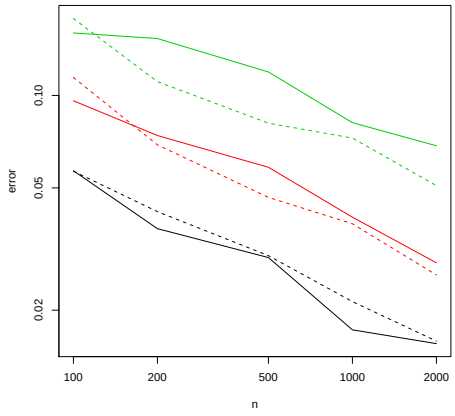
The results for several densities are summarized in Figure 4.15. The results for all example densities were largely similar.

For $d = 1$, there is similar performance for the kernel density estimator and the log-concave maximum likelihood estimator. Figure 4.15(a) illustrates this for the Gaussian data. For the larger (75%) region, the kernel performs better; for the median and upper regions (50% and 25%) the performance is similar, with the log-concave maximum likelihood estimator offering a slight improvement. For densities with bounded support, however, the fact that the log-concave maximum likelihood estimate did not suffer from boundary bias means that it was better able to capture the shape of the density. This is illustrated by the density C (Figure 4.15(b)), which is supported on $[0, \infty)$.

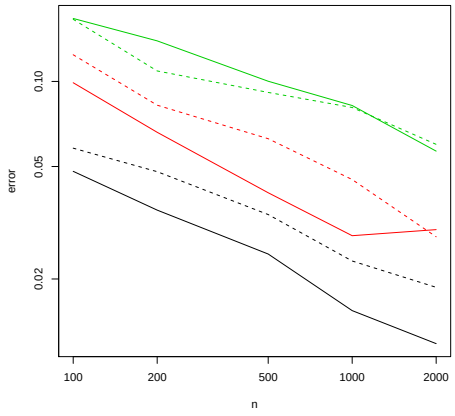
For $d > 1$, we see that the performance of the log-concave maximum likelihood estimator is even more impressive. Even for densities that are easy for the kernel density estimate to capture, such a Gaussian mixture that is only just log-concave (density E, Figure 4.15(c)), the log-concave maximum likelihood estimate performs at least as well as the kernel density estimator. It does a significantly better job of estimating the smaller regions, and improves more rapidly with increasing sample size. As before, we see a strong boundary effect for density G (Figure 4.15(d)).

For $d = 3$, the story is similar but the effects are more pronounced, and the improvement afforded by using the log-concave maximum likelihood estimator rather than a kernel density estimate is more dramatic. Again, the effect is present for simpler Gaussian data (such as density B, Figure 4.15(e)), and more dramatic for skewed data with a finite boundary (density C, Figure 4.15(f)).

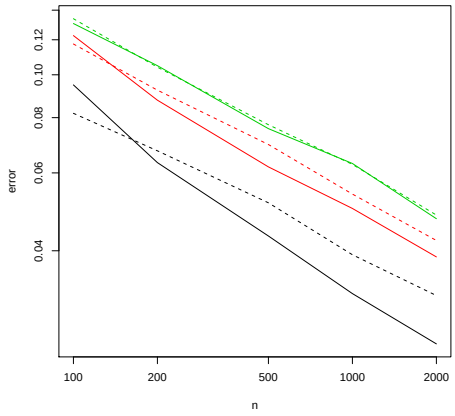
Moreover, once the density estimate has been computed, estimating $\widehat{f}_{n,\alpha}$, and the corresponding highest density regions is significantly less computationally expensive. This is because sampling from the density and evaluating the density at that point may be done in (effectively) constant time. This may be contrasted with the situation for the



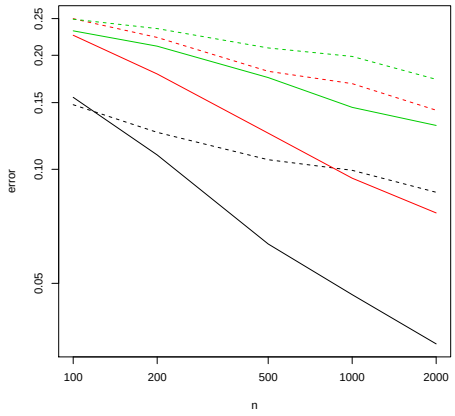
(a) Density A, $d = 1$



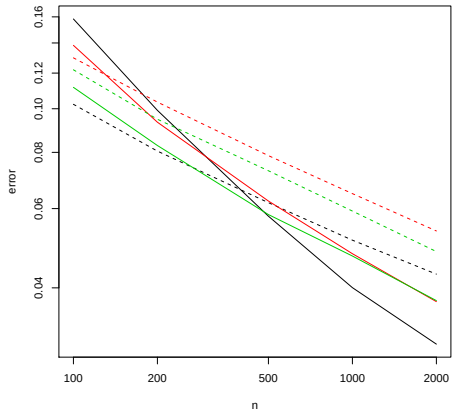
(b) Density C, $d = 1$



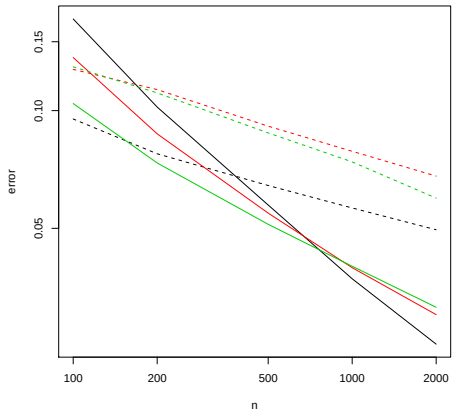
(c) Density E, $d = 2$



(d) Density G, $d = 2$



(e) Density B, $d = 3$



(f) Density C, $d = 3$

Figure 4.15: Mean error of highest density region estimates. Black is 25% HDR, red is 50% HDR, green is 75% HDR. Solid lines are log-concave maximum likelihood estimate and dashed lines are kernel density estimate.

kernel estimate, where $O(n)$ operations are needed to evaluate the density estimate once a sample has been drawn.

4.4 Finite mixture models

It may seem that the assumption that f belongs to the class of log-concave densities is rather restrictive: we cannot model multimodal distributions, for example. However, as we have already discussed in Chapter 1, some restriction is necessary if we are to avoid the need to select smoothing parameters for density estimation. A possible extension of the log-concave model is to extend to finite mixtures of log-concave densities, that is, to densities of the form

$$f(x) = \sum_{j=1}^{k} \pi_j f_j(x), \tag{4.9}$$

where the $\pi_j > 0$, $\sum_j \pi_j = 1$ and each $f_j \in \mathscr{F}$.

In this section, we discuss how to estimate the mixture using an EM-style algorithm for fixed k. There is a large literature on choice of k for both parametric and nonparametric models (see McLachlan and Peel (2000) for an introduction with many examples). Information criteria such as AIC or BIC are commonly used. These are based on an asymptotic approximation to

$$\mathbb{E}\left[\int \log \widehat{f}_n \, d(F_n - F_0)\right]. \tag{4.10}$$

Unfortunately, these approximations are based on the number of parameters, for which there is no easy analogue in the nonparametric setting. An alternative is Efron's information criterion (EIC), which approximates (4.10) directly using the bootstrap (Konishi and Kitagawa, 2008, Chapter 8). This has the advantage that we could also compare a log-concave mixture with, say, a Gaussian mixture. However, this is too computationally intensive to be realistically used with our algorithm. We therefore leave this issue as a topic for future research, alongside development of faster algorithms for $d > 1$.

4.4.1 The EM algorithm

For parametric models, the expectation-maximization (EM) algorithm is commonly used to estimate mixture densities of the form (4.9), where the components f_j are assumed to have a particular parametric form (such as Gaussian) (Dempster, Laird, and Rubin, 1977). However, this approach depends critically on the selection of an appropriate parametric model for the components which can be difficult to assess, particularly in the multivariate context.

Chang and Walther (2007) suggest using log-concave components in the EM framework. Simulation results in the one-dimensional case were promising, showing an improvement over a Gaussian model where this was inadequate and no appreciable loss of performance even for Gaussian mixtures. For the multivariate case a more restricted model (with log-concave components and a copula dependence structure) was used due to the lack of an algorithm for computing the log-concave maximum likelihood estimator. We show in this section how to extend this to model general log-concave mixtures for multivariate data.

In order to derive the EM algorithm, we artificially augment the data so that each observation consists of a pair (X_i, Z_i), with $Z_i \in \{1, \dots, k\}$ being the (unobserved) component from which observation i was drawn. Then the log-likelihood for the whole model is given by

$$\ell_n(\pi, f) = \frac{1}{n} \sum_{i=1}^{n} \log \pi_{Z_i} f_{Z_i}(X_i).$$

Since the values of Z_i are unknown, at each iteration we replace them with their expected values under the current model. In more detail, given current estimates $f_j^{(t)}$ and $\pi_j^{(t)}$ of f_j and π_j respectively, we compute the expected log-likelihood under this distribution, conditional on X_i, namely

$$\ell_n^{(t)}(\pi, f) = \frac{1}{n} \sum_{i=1}^{n} \sum_{j=1}^{k} \theta_{j,i}^{(t)} \left(\log \pi_i + \log f_j(X_i) \right),$$

where

$$\theta_{j,i}^{(t)} = \frac{\pi_j^{(t)} f_j^{(t)}(X_i)}{\sum_{l=1}^{k} \pi_l^{(t)} f_l^{(t)}(X_i)}$$

is the conditional probability that $Z_i = j$, given X_i. This is the so-called "expectation" step.

We then maximize this expected log-likelihood, that is, we set

$$(\pi^{(t+1)}, f^{(t+1)}) = \arg\max \ell_n^{(t)}(\pi, f),$$

where the maximization is over $\{\pi \geq 0 \colon \sum \pi_l = 1\}$ and $\mathscr{F}$ for π and f respectively. A standard calculation involving the addition of a Lagrangian term to enforce $\sum_i \pi_i = 1$ leads to the estimates

$$\pi_j^{(t+1)} = \frac{\sum_{i=1}^{n} \theta_{j,i}^{(t)}}{\sum_{i=1}^{n} \sum_{l=1}^{k} \theta_{l,i}^{(t)}}$$

and

$$f_j^{(t+1)} = \arg\max_{f \in \mathscr{F}} \sum_{i=1}^{n} w_{j,i} f(X_i),$$

where $\mathscr{F}$ is the class of all log-concave densities and

$$w_{j,i} = \frac{\theta_{j,i}^{(t)}}{\sum_{l=1}^{n} \theta_{j,l}^{(t)}}$$

are weights. From Section 3.6.4, we see that the weights $w_{j,i}$ may readily be incorporated into our existing method.

In the parametric case, the expectation and maximization steps are alternated until a local maximum is reached. The algorithm is guaranteed to converge to a local maximum. We mimic these alternating steps in the nonparametric case.

4.4.2 Implementation details

As usual for mixture models, the likelihood may be made arbitrarily large by choosing one extremely narrow component and one wide component. We therefore seek a local maximum in the likelihood surface with both components having a variance strictly greater than some $\epsilon > 0$. In order to start our search at a suitable point in the parameter space, as we initially fitted k clusters using a hierarchical clustering algorithm. We then modelled each of these clusters as a multivariate normal distribution, with the sample mean and variance of the cluster used for the mean and variance of each Gaussian component. This is the starting point suggested for the Gaussian EM algorithm in the R package **mclust** (Fraley and Raftery, 2008). Chang and Walther (2007) suggested using the maximum found by the Gaussian EM algorithm as a starting point. However, for multivariate data we found that this did not offer sufficient flexibility and our EM-style algorithm got stuck (in a local maximum of the likelihood function) at a distribution close to the Gaussian mixture, leading to little improvement.

In contrast to the case for parametric components, there is no formal proof of convergence of the nonparametric EM-style algorithm. Therefore, we run the algorithm until the relative increase in the log likelihood was smaller than some ϵ_ℓ. We used the value $\epsilon_\ell = 10^{-2}$. We also placed an upper bound $N_{\max} = 10$ on the number of iterations. This appeared to be sufficient to capture the shape of the distributions.

4.4.3 Application to visualization

For bivariate data, plotting the estimate can give some insight into the structure of the data. We illustrate that here with the universities data described in Section 4.2.2.

Example: Universities data revisited

The results of the Section 4.2.2 suggest a single log-concave component is insufficient to capture all the features of the universities dataset. We therefore applied our clustering technique to the first two principal components of the universities dataset described in Section 4.2.2. The resulting estimate is shown in Figure 4.16.

If we examine the two groups more closely, we see that, of the 63 universities for which

$$\widehat{\pi}_1 \widehat{f}_1(X_i) > \widehat{\pi}_2 \widehat{f}_2(X_i),$$

59 are in the top 60 of the Times Universities ranking. This suggests that there is a "break point" part way down the list splitting the universities into two broad groups.

4.4.4 Application to clustering

Given an observation X from a distribution with a density of the form (4.9), the Bayes clustering rule assigns this observation to

$$\underset{l \in \{1,\ldots,k\}}{\arg\max}\, \pi_l f_l(X).$$

This observation immediately suggests a plug-in rule for clustering. First, fit a density estimate, and then assign observation i to cluster j, where

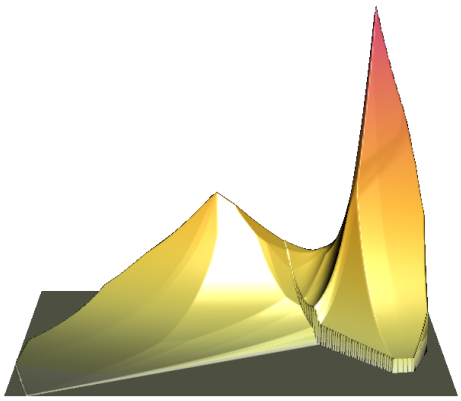
$$j = \underset{l \in \{1,\ldots,k\}}{\arg\max}\, \widehat{\pi}_l \widehat{f}_l(X_i). \tag{4.11}$$

This classification rule is a plug-in version of the optimal Bayes rule for the mixture. This method not only clusters the groups, but also gives a posterior estimate of the probability that an observation comes from each of the two groups, namely

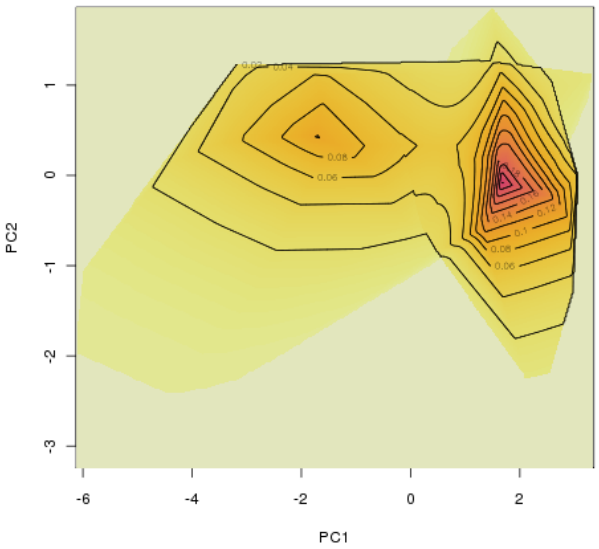
$$\tau_j(X_i) = \frac{\widehat{\pi}_j \widehat{f}_j(X_i)}{\sum_{l=1}^{k} \widehat{\pi}_l \widehat{f}_l(X_i)}.$$

Example: Breast cancer data

To illustrate this, we use the Wisconsin breast cancer diagnostic dataset (Street, Wolberg, and Mangasarian, 1993) available from the UCI machine learning repository (Asuncion and Newman, 2007). The dataset consists of several measurements from a digitized image of a fine needle aspirate of a breast mass, describing characteristics of the cell nucleus. In total, measurements from 569 individuals were taken, 357 being benign and 212 malignant. The full dataset consists 30 real values for each patient. These are



(a) Surface plot



(b) Contour plot

Figure 4.16: Density estimate for universities dataset.

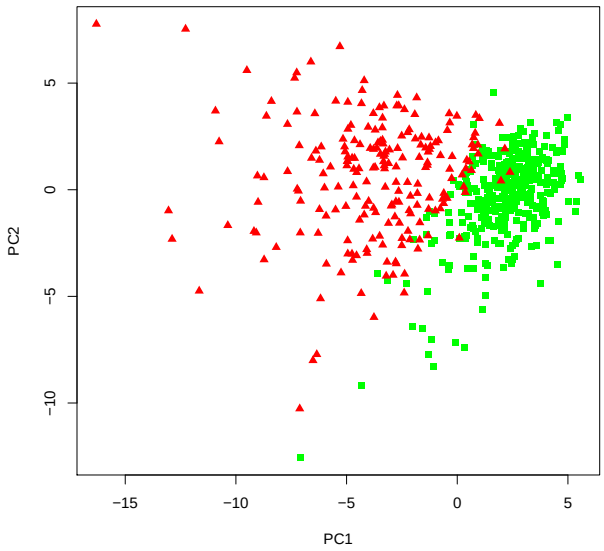
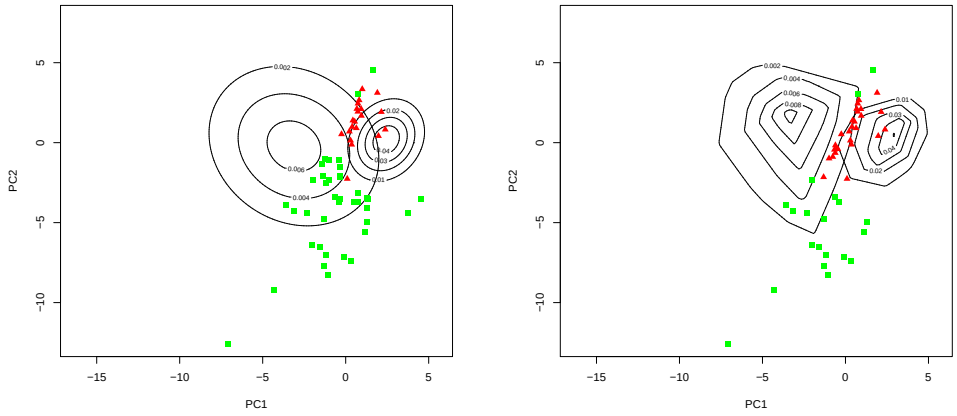


Figure 4.17: First two principal components of the Wisconsin breast cancer dataset. Red triangles are malignant, and green squares are benign.

the mean, standard error and worst (mean of 3 largest values) of ten aspects of the image. In order to use this dataset, we first projected the (rescaled) dataset onto its first two principal components (after rescaling the data to lie on the same scale in each dimension). Because of the inclusion of standard error information, we might expect that a log-concave mixture would better be able to capture the shape of the clusters than a Gaussian mixture. A plot of the first two principal components is given in Figure 4.17 below. These two components capture 63% of the variation in the data. Malignant instances are shown as red triangles, and benign as green squares. Some clustering is clearly present, but as the clusters are skewed a Gaussian mixture may not be adequate. A log-transform of the standard deviation measurements was considered, but was not appropriate because some of the observations had standard deviation zero.

To test the procedure, we fitted a two-component mixture using both the Gaussian and log-concave EM algorithms to the unlabelled dataset. For both the Gaussian and log-concave mixtures, the initial point was chosen by using hierarchical clustering to identify initial groups, and fitting a Gaussian distribution to each cluster. The variances of the Gaussian mixture were unrestricted. After fitting the mixture, we classified each observation according the the plug-in Bayes rule (4.11). To assess the performance of each method, we compared this to the known true classification and counted the number of misclassifications. The two components of the fitted mixture are illustrated



(a) Gaussian mixture using EM algorithm

(b) Log-concave mixture fitted using EM-style algorithm

Figure 4.18: Contours of the components of a two-component mixture fitted to the Wisconsin breast cancer dataset. Misclassified points are shown. Red triangles are malignant, and green squares are benign.

in Figure 4.18(a) for the Gaussian mixture, and Figure 4.18(b) for the log-concave mixture. We see that the log-concave mixture is better able to capture the shape of the two components. The "malignant" component has an artificially large variance in order to capture the skewed part of the data, and this leads to greater classification error. However, the log-concave mixture is able to adapt automatically to the shape of the density. The misclassification rate was reduced from 59/569 to 48/569. In both Figure 4.18(a) and Figure 4.18(b), only misclassified points are shown.

4.5 Conclusion

In this chapter, we saw that, besides being a simple and parsimonious density estimator, the log-concave maximum likelihood may be used for many more complicated applications.

We presented a method, extending the work of Walther (2002), of assessing the suitability of a log-concave model for a particular dataset. We tested the adequacy of a single log-concave component using a multiscale test. We successfully detected heavy tails and mixing on simulated examples, and did not produce a significant result when log-concavity was appropriate. This method was also applied to a real dataset, where we found a single log-concave component was not able to adequately model the data. This informed our decision to use a mixture for this dataset in Section 4.4.

4.5 Conclusion

We considered plug-in estimation of 3 different quantities: the covariance, the differential entropy and the highest density region. For the covariance, an empirical plug-in estimator was found to have superior performance in terms of the mean squared error. However, the log-concave maximum likelihood estimate showed smaller variance, a property that has been found elsewhere. For a more complicated example, the differential entropy, we observed favourable performance compared to a common competitor, the kernel density estimator. A range of bandwidth selectors were tried, and in each case we found that the log-concave maximum likelihood estimator performed better, especially for large sample sizes or higher dimensions. Once again, by decomposing the mean squared error into bias and variance terms we saw that the log-concave estimator was more stable. It also did not suffer from boundary bias where the distribution had bounded support.

The final plug-in estimator considered was for highest density regions. Once again, for multivariate data the log-concave estimator performed well compared to the kernel density estimate. It has the added advantage of producing convex highest density regions, meaning it is well able to capture the essential structure of the density.

Finally, we discussed how an EM-style algorithm may be used to fit mixtures where the number of components is known. This was applied to two examples. Firstly, we fitted a mixture to the universities dataset from the previous section, guided by our previous assessment. Secondly, we investigated the use of this method for clustering using a plug-in Bayes rule using a breast cancer dataset. We saw a significant improvement over a Gaussian mixture because of the improved adaption to the shape of the underlying clusters.

In summary, while acknowledging the computational limitations of our current methods, we have developed a variety of useful applications of this estimator.

5 Performance

5.1 Introduction

In this chapter we examine the asymptotic performance of the log-concave maximum likelihood estimator from two different points of view. Firstly, we discuss convergence in terms of the Hellinger metric. This is motivated by general considerations of the maximum likelihood estimation in Section 5.2. We prove Hellinger consistency in Section 5.3 via a uniform law of large numbers for a suitable class of sets. By bounding the "size" of the class of log-concave functions, in Section 5.4 we derive a rate of convergence of a restricted log-concave maximum likelihood estimator with respect to the Hellinger distance. This is supplemented by simulation results for several example densities.

Secondly, in Section 5.5 we compare the performance of the log-concave maximum likelihood estimator with that of one of its main competitors, the kernel density estimator introduced in Chapter 1. Since using this estimator in practice depends on careful choice of a bandwidth matrix, we compare several possibilities for bandwidth selection, including sophisticated adaptive methods. The results reveal favourable performance of the log-concave maximum likelihood estimator in a range of settings.

5.1.1 Asymptotic behaviour of log-concave maximum likelihood estimators

For the special case $d = 1$, there has already been some interest in the asymptotic performance of log-concave maximum likelihood estimators. Balabdaoui et al. (2009) have provided a pointwise limiting distribution in terms of derivatives at 0 of the "lower invelope" of an integrated Brownian motion minus a drift term. This leads to a pointwise rate of convergence of $O_p(n^{-k/(2k+1)})$ at a point x_0, where k is the smallest integer ≥ 2 such that

$$\frac{d^k}{dx^k} \log f_0(x_0) \neq 0.$$

They also prove consistency of the mode of $\widehat{f}_n$ as an estimator of the mode of f_0.

Uniform convergence at a rate $O_p((\log n/n)^\alpha)$ on compact intervals strictly contained within the support of f_0, with $\alpha \in [1/3, 2/5]$ depending on the smoothness of f_0, has been provided by Dümbgen and Rufibach (2008). In detail, it is shown that if f_0 is Hölder continuous with exponent $\beta \in [1, 2]$, we may set $\alpha = \beta/(2\beta + 1)$. As a corollary of this,

the authors obtain L_1-consistency of the maximum likelihood estimator and uniform consistency of the integrated density as an estimator of the distribution function. Further, $\left|\widehat{F}_n(x) - F_n(x)\right|$ is shown to converge at rate $O_p(n^{3\beta/(4\beta+2)})$ on compact subintervals of the support of f_0, where $\widehat{F}_n$ is the integrated log-concave maximum likelihood estimator. Pal et al. (2007) have proved consistency of the log-concave maximum likelihood estimator with respect to the Hellinger metric, although no rate is provided.

These results rely on the special structure of the space of log-concave densities when $d = 1$. This was mentioned in Section 3.5. In general, results for $d > 1$ require different techniques.

5.2 Nonparametric maximum likelihood estimation

In this section we introduce some tools for understanding convergence of nonparametric likelihood estimators. Key references on this subject are Pollard (1984, Chapter II), van der Vaart and Wellner (1996) and van de Geer (2000). The discussion in Evans (2007) has also been useful.

5.2.1 Consistency

For any class of densities $\mathscr{F}$ and $f_0 \in \mathscr{F}$,

$$f_0 = \arg\max_{f \in \mathscr{F}} \int \log f \, dF_0. \tag{5.1}$$

We may therefore interpret $\widehat{f}_n$ as solving (5.1), where we have replaced F_0 with the empirical distribution function F_n. Note that (5.1) implies that, for all $f \in \mathscr{F}$,

$$\int \log \frac{\widehat{f}_n}{f_0} dF_n \geq 0.$$

Combining this with (5.1), we see that

$$\begin{aligned} 0 &\leq \int \log \frac{f_0}{\widehat{f}_n} \, dF_0 \\ &\leq \int \log \frac{f_0}{\widehat{f}_n} \, dF_0 - \int \log \frac{f_0}{\widehat{f}_n} \, dF_n \\ &= \int \log \frac{f_0}{\widehat{f}_n} \, d(F_0 - F_n). \end{aligned}$$

Note that, for a fixed measurable function g with $\int |g| \, dF_0 < \infty$, by the law of large numbers as $n \to \infty$

$$\left| \int g \, d(F_0 - F_n) \right| \xrightarrow{a.s.} 0.$$

If we can show that this holds uniformly for g in

$$\mathcal{G} = \left\{ \log \frac{f_0}{f} : f \in \mathcal{F} \right\},$$

i.e.

$$\sup_{g \in \mathcal{G}} \left| \int g \, d(F_0 - F_n) \right| \xrightarrow{a.s.} 0$$

as $n \to \infty$, then we certainly have consistency of the maximum likelihood estimator. In this case, we say that the class $\mathcal{G}$ satisfies a uniform law of large numbers.

In certain special cases it may suffice to provide a bound on some different class of functions arising from a different transformation of $\mathcal{F}$. Several examples are given in van de Geer (2000, Chapter 4). Identifying and proving a suitable law of large numbers can be delicate. In the next section, we make a minor diversion into uniform laws of large numbers which is important later.

5.2.2 Uniform laws of large numbers

In the previous section, we said that a class of functions $\mathcal{G}$ satisfied a uniform law of large numbers (ULLN) for f_0 if

$$\sup_{g \in \mathcal{G}} \left| \int g \, d(F_0 - F_n) \right| \xrightarrow{a.s.} 0$$

as $n \to \infty$. In the special case

$$\mathcal{G} = \{ \mathbb{1}_A : A \in \mathcal{A} \}$$

for some collection $\mathcal{A}$ of subsets of $\mathbb{R}^d$, we say $\mathcal{A}$ satisfies a ULLN if

$$\sup_{A \in \mathcal{A}} \left| \mathbb{P}(A) - \mathbb{P}_n(A) \right| \xrightarrow{a.s.} 0$$

where

$$\mathbb{P}(A) = \mathbb{P}_0(A) = \int \mathbb{1}_A \, dF_0$$

and $\mathbb{P}_n$ is defined similarly for the empirical measure. An example of such a result is the classical (multivariate) Glivenko-Cantelli theorem, for which

$$\mathscr{A} = \left\{(-\infty, x_1] \times \ldots \times (-\infty, x_d] : (x_1, \ldots, x_d) \in \mathbb{R}^d\right\}.$$

Proving a law of large numbers depends on the ability of sets in $\mathscr{A}$ to pick out subsets of points in $\mathbb{R}^d$. To state this more precisely, we require an additional definition. Given a set of points $\mathscr{Z} = \{\zeta_1, \ldots, \zeta_n\} \subseteq \mathbb{R}^d$, we say that $\mathscr{A}$ *shatters* $\mathscr{Z}$ if every $Z \subseteq \mathscr{Z}$ can be written in the form $\mathscr{Z} \cap A$ for some $A \in \mathscr{A}$. The exact relationship between shattering and uniform laws of large numbers is provided by the following theorem.

Theorem 5.1 (Pollard, 1984, Chapter II, Theorem 21)**.** *Let $\mathscr{A}$ be a class of measurable subsets of $\mathbb{R}^d$. A necessary and sufficient condition for*

$$\sup_{A \in \mathscr{A}} |\mathbb{P}_n(A) - \mathbb{P}(A)| \xrightarrow{a.s.} 0$$

is that

$$\frac{1}{n} S_n \xrightarrow{p} 0$$

where $S_n = S_n(\zeta_1, \ldots, \zeta_n)$ is the smallest integer such that $\mathscr{A}$ shatters no collection of S_n points from $\{\zeta_1, \ldots, \zeta_n\} \subseteq \mathbb{R}^d$.

We use this in Section 5.3.1.

5.2.3 Hellinger distance

The Kullback-Leibler divergence

$$\mathrm{KL}(g, f) = \int \log \frac{g}{f} \, dF$$

does not define a distance function. However, it is closely related to the Hellinger distance, defined for two nonnegative functions f and g on $\mathbb{R}^d$ to be

$$h(f, g) = \left[\int \left(f^{1/2}(x) - g^{1/2}(x) \right)^2 dx \right]^{1/2}$$

If f and g are densities, $h^2(f, g) \leq 2$ (some authors normalise so that $h^2(f, g) \leq 1$). This is less interpretable than the more common L_p norms (particularly for $p = 1, 2$ or ∞), but is particularly useful for nonparametric convergence rates due to the following inequality.

Lemma 5.2 (van de Geer, 2000, Lemma 1.3)**.** *For densities f and g on $\mathbb{R}^d$,*

$$h^2(g,f) \leq \mathrm{KL}(g,f)$$

Proof. Noting that, for $y > 0$, $\log y \leq y - 1$, and applying this to $y^2 = \frac{f(x)}{g(x)}$, we have

$$\log \frac{f(x)}{g(x)} \leq 2\left(\frac{f^{1/2}(x)}{g^{1/2}(x)} - 1\right)$$

so that

$$\begin{aligned}
\mathrm{KL}(g,f) &= \int \log\left(\frac{g}{f}\right) dF \\
&\geq 2 - 2\int f\left(\frac{g^{1/2}}{f^{1/2}}\right) \\
&= 2 - 2\int f^{1/2} g^{1/2} \\
&= h^2(g,f)
\end{aligned}$$

□

In light of the discussion in Section 5.2.1, a suitable ULLN therefore implies Hellinger consistency. As remarked by van de Geer (2000, p.8), this is far from a necessary condition.

5.2.4 Entropy and bracketing entropy

The rate of convergence of the maximum likelihood estimator over a class $\mathscr{F}$ is intimately connected to the "size" of the $\mathscr{F}$, measured via the notion of the entropy.

The entropy of a metric space $(\mathscr{Y}, d)$ is defined as follows. Fix $\epsilon > 0$ and consider a collection of points $x_1, \ldots, x_N$ (not necessarily in $\mathscr{Y}$) such that, for every $x \in \mathscr{Y}$, there is some i such that $d(x, x_i) \leq \epsilon$. This is called a *covering* of $\mathscr{Y}$. Let $N(\epsilon, \mathscr{Y}, d)$ be the smallest such N; then the *ϵ-entropy* is defined as

$$H(\epsilon, \mathscr{Y}, d) = \log N(\epsilon, \mathscr{Y}, d).$$

We need to extend this notion in the special case in which $\mathscr{Y}$ is a space of real-valued functions $\mathscr{G}$ on some metric space Y, in our case $\mathbb{R}^d$. As before, fix $\epsilon > 0$ and let N_B be the smallest number such that there exist pairs

$$\{(g_i^L, g_i^U), i = 1, \ldots, N_B\}$$

(not necessarily in $\mathscr{G}$) such that $d(g_i^L, g_i^U) \leq \epsilon$ and, for every $g \in \mathscr{G}$, there is some j such that

$$g_j^L(y) \leq g(y) \leq g_j^U(y) \text{ for all } y \in Y.$$

Then $H_B(\epsilon, \mathscr{G}, d) = \log N_B(\epsilon, \mathscr{G}, d)$ is the *ϵ-entropy with bracketing* of $(\mathscr{G}, d)$.

5.2.5 Connection between bracketing entropy and rate of convergence

The relationship between the bracketing entropy of a space of functions as defined in the previous section and the rate of convergence of the maximum likelihood in this space is provided by Wong and Shen (1995), and explored further in van de Geer (2000).

The key result is as follows.

Theorem 5.3 (Wong and Shen, 1995, Theorem 2)**.** *Let $\mathscr{F}$ be a class of densities and let $\widehat{f}_n$ be a sequence of maximum likelihood estimators. Then there exist positive constants $c_1, \ldots, c_4$ such that, for $\delta_n > 0$, if*

$$\int_{\delta_n^2/2^8}^{\sqrt{2}\delta_n} H_B^{1/2}(u/c_1, \mathscr{F}, h)\, du \leq c_2 n^{1/2} \delta_n^2,$$

then for sufficiently large n,

$$\mathbb{P}(h(\widehat{f}_n, f_0) \geq \delta_n) \leq c_3 \exp(-c_4 n \delta_n^2).$$

This shows that the rate of convergence is essentially determined by δ_n, the smallest δ such that

$$\int_{\delta^2}^{\delta} H_B^{1/2}(u, \mathscr{F}, h)\, du \leq n^{1/2} \delta^2. \tag{5.2}$$

5.2.6 Sieves

In some cases it may be possible to improve upon the rate given by Theorem 5.3 by approximating $\mathscr{F}$ by a sequence of spaces $\mathscr{F}_n$ with smaller bracketing entropy. Given such a sequence, it may be possible to define a sequence of sieved maximum likelihood estimators

$$\widetilde{f}_n = \operatorname*{arg\,max}_{f \in \mathscr{F}_n} \int \log f \; dF_n.$$

Given a class of functions $\mathscr{F}_n \subseteq \mathscr{F}$, a suitable index of the degree to which $\mathscr{F}_n$ approximates a given $f_0 \in \mathscr{F}$ is given for $\alpha \in (0, 1]$ by

$$\inf_{f \in \mathscr{F}_n} \rho_\alpha(f_0, f),$$

where, for $\alpha \in [-1, 1]$

$$\rho_\alpha(f_0, f) = \begin{cases} \frac{1}{\alpha}\left[\mathbb{E}\left(\frac{f_0}{f}\right)^\alpha - 1\right] & \text{if } \alpha \neq 0 \\ \mathbb{E}\left[\log\left(\frac{f_0}{f}\right)\right] & \text{else.} \end{cases}$$

Here expectation is under f_0.

The following theorem tells us that the rate of convergence of $\widetilde{f}_n$ to f_0 is determined by the entropy of $\mathscr{F}_n$, and the distance between f_0 and $\mathscr{F}_n$. Along the same lines as Theorem 5.3, the link with the first quantity is via δ_n, the smallest δ such that

$$\int_{\delta^2}^{\delta} H_B^{1/2}(u, \mathscr{F}_n, h)\, du \leq n^{1/2}\delta^2. \tag{5.3}$$

Theorem 5.4 (Wong and Shen, 1995, Theorem 4)**.** *Let $\mathscr{F}_n$ be a sieve, and $\widetilde{f}_n$ be the corresponding sieved maximum likelihood estimators. Let*

$$\tau_n(\alpha) = \inf_{f \in \mathscr{F}_n} \rho_\alpha(f_0, f),$$

and suppose that, for some $\alpha \in (0, 1]$,

$$\tau_n(\alpha) \leq \frac{1}{\alpha}.$$

Then there exist constants $c_1, \ldots, c_5$ such that if

$$\int_{\delta_n^2/2^8}^{\sqrt{2}\delta_n} H_B^{1/2}(u/c_1, \mathscr{F}_n, h)\, du \leq c_2 n^{1/2}\delta_n^2$$

and

$$\epsilon_n = \min\left(\delta_n, \left(\frac{4\tau_n(\alpha)}{c_3}\right)^{1/2}\right)$$

then

$$\mathbb{P}(h(f_0, \widetilde{f}_n) \geq \epsilon_n) \leq c_4 \exp(-c_5 n\epsilon_n^2).$$

When choosing our sieve, if we wish to ensure the optimal rate of convergence we must ensure that

$$\tau_n \asymp \delta_n^2.$$

A useful application of this is to correct a suboptimal rate that sometimes arises in

an attempt to apply Theorem 5.3. This is part of Example 4 in Wong and Shen (1995). If the Hellinger bracketing entropy $H_B(\delta, \mathscr{F}, h)$ is finite for some class $\mathscr{F}$, it is shown in this example that we may, for a suitable choice of τ_n, use a sieve based on the bracketing

$$\mathscr{G}_n = \left\{(f_j^L, f_j^U) \colon j = 1, \ldots, N = \exp(H_B(\tau_n, \mathscr{F}, h))\right\}$$

to improve the rate of convergence. The sieve is

$$\mathscr{F}_n = \left\{ \frac{f_{j,n}^U}{\int f_{j,n}^U} : j = 1, \ldots, N \right\}.$$

It may be shown that the rate of Hellinger convergence is then $O(\delta_n)$, where

$$H_B(\delta_n, \mathscr{F}, h) \asymp n\delta_n^2.$$

5.2.7 Effect of model misspecification

If $f_0 \notin \mathscr{F}$, the arguments above lead to a rate of convergence for $h(\widetilde{f}_n, f_n^*)$, where

$$f_n^* = \operatorname*{arg\,max}_{f \in \mathscr{F}_n} \int \log f \, dF_0.$$

This is the closest approximation to f_0 within the class $\mathscr{F}_n$. Thus, if the model is misspecified, the maximum likelihood estimator converges to the closest approximation to the true density (in the Kullback-Leibler sense). For our purposes this is a desirable robustness property.

5.3 Consistency of the log-concave maximum likelihood estimator

In this section we prove that the maximum likelihood estimator is consistent with respect to the Hellinger distance. The motivation for using this distance function was given in Section 5.2.3. Our result follow Pal et al. (2007), which proved Hellinger consistency if $d = 1$. We first prove a suitable ULLN and several technical results. We combine these to establish consistency in Theorem 5.11.

5.3.1 A uniform law of large numbers

In this section, we prove a uniform Glivenko-Cantelli result needed in the sequel. This proof closely follows Pollard (1984, Example 22, Chapter II), which considered the special case $d = 2$ and $f = \mathbb{1}_C$, where $C = [0,1]^2$. An equivalent result is Theorem 2.1.1 of Bhattacharya and Rao (1976).

Theorem 5.5. *Let f be a density on $\mathbb{R}^d$, and $\mathbb{P}$ the corresponding measure. Then the class $\mathscr{A}$ of all measurable convex subsets of $\mathbb{R}^d$ satisfies a uniform strong law of large numbers, that is,*

$$\sup_{A \in \mathscr{A}} |\mathbb{P}_n(A) - \mathbb{P}(A)| \xrightarrow{a.s.} 0$$

Proof. We use the notation of Theorem 5.1. Suppose for a contradiction that for some $\epsilon > 0$

$$P\left(\frac{1}{n} S_n \geq \epsilon\right) \geq \epsilon \text{ infinitely often.} \tag{5.4}$$

For $m, l \in \mathbb{N}$, consider a grid dividing the hypercube $[-m,m]^d$ into l^d hypercubes of side length $\delta = \frac{2m}{l}$. Let $\mathscr{B}_{m,l}$ be the collection of all subsets of this collection of hypercubes together with $\mathbb{R}^d \backslash [-m,m]^d$. Since this is a finite collection of measurable sets,

$$P\left(\sup_{B \in \mathscr{B}_{m,l}} \left|\mathbb{P}_n(B) - \mathbb{P}(B)\right| \geq \frac{\epsilon}{2}\right) < \frac{\epsilon}{2}$$

for sufficiently large n. Combining this with (5.4), for suitable n,

$$P\left(\frac{1}{n} S_n \geq \epsilon \text{ and } \sup_{B \in \mathscr{B}_{m,\delta}} |\mathbb{P}_n(B) - \mathbb{P}(B)| \leq \frac{1}{2}\epsilon\right) \geq \frac{1}{2}\epsilon.$$

This set cannot be empty, so for some configuration of sample points $\mathscr{A}$ must shatter at least $n\epsilon$ points, and $|\mathbb{P}_n(B) - \mathbb{P}(B)|$ must be less than $\frac{1}{2}\epsilon$ for each $B \in \mathscr{B}_{m,l}$. Write H for the convex hull of a shattered set of size at least $n\epsilon$ and

$$B_H = \bigcap \{B \in \mathscr{B}_{m,l} : H \subseteq B\}.$$

Note that $B_H \in \mathscr{B}_{m,l}$. Further, $\mathbb{P}_n(B_H) \geq \epsilon$ (since H contains at least $n\epsilon$ points) and

$$|\mathbb{P}_n(B_H) - \mathbb{P}(B_H)| \leq \frac{1}{2}\epsilon$$

(since $B_H \in \mathscr{B}_{m,\delta}$). Therefore $\mathbb{P}(B_H) \geq \frac{1}{2}\epsilon$.

This gives us our desired contradiction. Fix m so that all the sample points lie in $[-m,m]^d$. Consider $l = 3^k$ for $k = 1, 2, \ldots$. For $k = 1$, no convex set can have boundary

points in all 3^d of the hypercubes in $\mathscr{B}_{m,l}$ that partition $[-m,m]^d$, so

$$\mathbb{P}(B_H) \leq \frac{c(2m)^d(3^d-1)}{3^d}.$$

Applying the same argument to each hypercube and repeating k times, we find that for $k = 1, 2, \ldots,$

$$\mathbb{P}(B_H) \leq \frac{c(2m)^d(3^d-1)^k}{3^{dk}}.$$

Choosing k sufficiently large leads to the desired contradiction. □

5.3.2 Technical preliminaries

We begin with a lemma.

Lemma 5.6 (Pal et al., 2007, Lemma 1)**.** *If f and g are densities on $\mathbb{R}^d$, F is the distribution function of f and $b > 0$, then*

$$h^2(f,g) \leq \int \log\left(\frac{f(x)+b}{g(x)+b}\right) dF(x) + \epsilon(b)$$

where

$$\epsilon(b) = 2\int \left(\frac{b}{f(x)+b}\right)^{1/2} dF(x).$$

Proof.

$$\begin{aligned}
\int \log\left(\frac{g(x)+b}{f(x)+b}\right) dF(x) &\leq 2\left[\int \left(\frac{g(x)+b}{f(x)+b}\right)^{1/2} dF(x) - 1\right] \\
&\leq 2\left[\int \left(\frac{b}{f(x)+b}\right)^{1/2} dF(x) + \int \left(\frac{g(x)}{f(x)+b}\right)^{1/2} dF(x) - 1\right] \\
&\leq \epsilon(b) + 2\left[\int \big(f(x)g(x)\big)^{1/2}\, dx - 1\right] \\
&\leq \epsilon(b) - h^2(f,g)
\end{aligned}$$

□

Our second lemma is closely related to the second lemma of Pal et al. (2007).

Lemma 5.7. *Let $b > 0$ and $0 < c < \infty$. Let f be a density on $\mathbb{R}^d$. Let $X_1, \ldots, X_n$ an iid sample from f. Denote the empirical distribution function of $X_1, \ldots, X_n$ by F_n, and the distribution function corresponding to f by F. Let g be a log-concave function on $\mathbb{R}^d$ with*

$$\sup_{x \in \mathbb{R}^d} g(x) \leq c.$$

Then

$$\left| \int \log(g(x) + b)\, d(F_n(x) - F(x)) \right| \xrightarrow{a.s.} 0.$$

Proof. Let $\mathscr{A}$ denote the class of all convex (measurable) subsets of $\mathbb{R}^d$. Then by Fubini's theorem, we have

$$\left| \int \log(g(x) + b)\, d(F_n(x) - F(x)) \right| \leq \sup_{A \in \mathscr{A}} \left| \mathbb{P}_n(A) - \mathbb{P}(A) \right| \log\left(1 + \frac{c}{b}\right)$$

The result follows by Theorem 5.5. □

Our next task is to show that, with probability 1, the sequence

$$M_n = \max_{x \in \mathbb{R}^d} \widehat{f}_n(x)$$

is bounded. We accomplish this by generalizing the arguments in Pal et al. (2007). We start with a generalization of their Lemma 4.

Lemma 5.8. *If $a_i, x > 0$ and*

$$x \leq \sum_{i=0}^{k} a_i (\log x)^i \tag{5.5}$$

then

$$x \leq (d+1) \sum_{i=0}^{k} 2^i a_i \log((d+1) a_i i^i)^i.$$

Proof. For $1 \leq m \leq k$, let $\lambda = (d+1)m^m$. Then, for $a > 0$,

$$\begin{aligned}
a(\log x)^m &= a(m \log(x^{1/m}))^m \\
&\leq a \left[m \log\left((\lambda a)^{1/m} \left(1 + \left(\frac{x}{\lambda a} \right)^{1/m} \right) \right) \right]^m \\
&\leq a \left[m \left[\log((\lambda a)^{1/m}) + \left(\frac{x}{\lambda a} \right)^{1/m} \right] \right]^m \\
&\leq 2^m a m^m \left[\log\left((\lambda a)^{1/m} \right)^m + \frac{x}{\lambda a} \right] \\
&= 2^m a \log(\lambda a)^m + \frac{x}{d+1}
\end{aligned}$$

Applying this to each term of (5.5) with $i > 0$ and rearranging, we obtain the result. □

We now need a bound on log-concave densities.

Lemma 5.9. *Let $X_1, \ldots, X_k$ be points in $\mathbb{R}^d$ and let g be a log-concave density on $\mathbb{R}^d$. Let*

$$a = \min_j \log g(X_j) \text{ and } c = \max_j \log g(X_j).$$

Then

$$e^c \leq \frac{1}{V_k} \left(\sum_{i=0}^{d-1} \frac{(c-a)^i}{i!} + (c-a)^d \right).$$

where V_k is the volume of the convex hull of $X_1, \ldots, X_k$.

Proof. Without loss of generality suppose that $\log g(X_1) = a$ and $\log g(X_k) = c$. If $a = c$ or $a = -\infty$ the result is trivial, so assume that neither is the case. Let C_k denote the convex hull of the points $X_1, \ldots, X_k$. Divide this into simplices $T_1, \ldots, T_p$ such that each has X_k as a vertex and the simplices form a triangulation of C_k. Then, for each T_j, using the expression (3.10) and the notation of (3.8), we have

$$\begin{aligned} \int_{T_j} g(x)\,dx &\geq \left|A_j\right| G_d(c, a, \ldots, a) \\ &\geq \left|A_j\right| e^a \left(\frac{e^{c-a} - \left(\sum_{i=0}^{d-1} \frac{(c-a)^i}{i!} \right)}{(c-a)^d} \right) \end{aligned}$$

Adding up the contributions from all T_j and using the fact that $\int g = 1$, we obtain

$$e^c \leq \frac{(c-a)^d}{V_k} + e^a \left(\sum_{i=0}^{d-1} \frac{(c-a)^i}{i!} \right).$$

Using the fact that $e^a \leq \frac{1}{V_k}$ (since $\int g = 1$), we obtain the result. □

We now combine these results to control the rate of growth of M_n.

Lemma 5.10. *Let f_0 be a log-concave density on $\mathbb{R}^d$ and let $X_1, X_2, \ldots$ be an iid sample from f_0. Let $\widehat{f}_n$ the maximum likelihood estimate based on the first n observations. Let*

$$M_n = \sup_x \log \widehat{f}_n(x).$$

Then with probability 1, there exists a constant C such that $M_n \leq C$ for sufficiently large n.

Proof. Since f_0 is log-concave, we automatically have $\int \left|\log f_0\right| dF_0 < \infty$. Let $K_n = \lceil \frac{n}{2} \rceil$.

Without loss of generality, assume that $\widehat{f}_n(X_1) \leq \widehat{f}_n(X_2) \leq \ldots \leq \widehat{f}_n(X_n)$. Then we have

$$\begin{aligned}\ell_n(f_0) &\leq \ell_n(\widehat{f}_n) \\ &\leq \left(1 - \frac{K_n}{n}\right) \log(\widehat{f}_n(X_n)) + \frac{K_n}{n} \log(\widehat{f}_n(X_{K_n})),\end{aligned}$$

so that

$$\begin{aligned}\log\left(\frac{\widehat{f}_n(X_n)}{\widehat{f}_n(X_{K_n})}\right) &\leq \frac{n}{K_n} \log\left(\widehat{f}(X_n)\right) - \frac{n}{K_n}\ell_n(f_0) \\ &\leq 2\log(\widehat{f}_n(X_n)) - \ell_n(f_0). \qquad (5.6)\end{aligned}$$

Moreover, letting V_{K_n} denote the volume of the convex hull of $\{X_1, \ldots, X_{K_n}\}$, we have by Lemma 5.9,

$$\widehat{f}_n(X_n) \leq \frac{1}{V_{K_n}} \left[\sum_{i=0}^{d-1} \frac{1}{i!} \left[\log\left(\frac{\widehat{f}_n(X_n)}{\widehat{f}_n(X_{K_n})}\right)\right]^i + 2\left[\log\left(\frac{\widehat{f}_n(X_n)}{\widehat{f}_n(X_{K_n})}\right)\right]^d\right].$$

Combining this with (5.6),

$$\widehat{f}_n(X_n) \leq \frac{1}{V_{K_n}} \left[1 + \sum_{i=0}^{d-1} \frac{1}{i!} \left[2\log\left[\left(\widehat{f}_n(X_n) - \ell_n(f_0)\right)^i\right) + \left(\log \widehat{f}_n(X_n) - \ell_n(f_0)\right)^d\right]\right].$$

This is precisely of the form to which we may apply Lemma 5.8. By the strong law of large numbers,

$$\ell_n(f_0) \xrightarrow{a.s.} \int \log f_0 \, dF_0$$

which is finite since f_0 is assumed log-concave. Moreover, the convex hull of $X_1, \ldots, X_{K_n}$ is convex and has empirical measure that is certainly greater than $\frac{1}{3}$, say. Also, the volume of any set of probability at least $\frac{1}{3}$ under f_0 is bounded below by some strictly positive value. Therefore V_{K_n} is almost surely bounded away from zero eventually. Combining this with Lemma 5.8 yields the result. □

5.3.3 Consistency

We now join all the pieces in a manner which closely resembles Theorem 3.1 of Pal et al. (2007).

Theorem 5.11. *Let f_0 be a log-concave density and let $\widehat{f}_n$ denote the log-concave maximum likelihood estimator. Then, with probability 1, $h(f_0, \widehat{f}_n) \to 0$ as $n \to \infty$.*

Proof. First observe that f_0 must be bounded since it is log-concave and integrates to 1. Thus we may apply any of the results from the previous section.

If f_0 is log-concave and $b > 0$, then by Lemma 5.6,

$$\begin{aligned} h^2(f_0, \widehat{f}_n) &\leq \epsilon(b) + \int \log\left(\frac{f_0(x)+b}{\widehat{f}_n(x)+b}\right) dF_0(x) \\ &= \epsilon(b) + \int \log(f_0(x)+b)\, dF_0(x) - \int \log(\widehat{f}_n(x)+b)\, dF_0(x) + \\ &\qquad \int \log(\widehat{f}_n(x)+b)\, dF_n(x) - \int \log(\widehat{f}_n(x)+b)\, dF_n(x) + \\ &\qquad \int \log(f_0(x))\, dF_n(x) - \int \log f_0(x)\, dF_n(x) \\ &= \epsilon(b) + A_n + B_n + C_n \end{aligned}$$

where

$$A_n = \int \log(\widehat{f}_n(x)+b)\, d(F_n(x) - F_0(x)),$$

$$B_n = \int [\log f_0(x) - \log(\widehat{f}_n(x)+b)]\, dF_n(x)$$

and

$$C_n = \int \log(f_0(x)+b)\, dF_0(x) - \int \log f_0(x)\, dF_n(x).$$

Now by Lemma 5.10, $\widehat{f}_n$ is (uniformly) bounded above, so we may apply Lemma 5.7 to conclude that with probability 1,

$$\left|A_n\right| \to 0.$$

Now observe that

$$\begin{aligned} B_n &\leq \int [\log f_0(x) - \log \widehat{f}_n(x)]\, dF_n(x) \\ &\leq 0. \end{aligned}$$

Further, by the strong law of large numbers,

$$C_n \xrightarrow{a.s.} \int \left[\log(f_0(x)+b) - \log(f_0(x))\right]\, dF_0(x)$$

Rearranging, with probability 1

$$\limsup h^2(f_0, \widehat{f}_n) \leq \int \left[\log(f_0(x) + b) - \log f_0(x)\right] \, dF_0(x) + \epsilon(b)$$

for each $b > 0$. However, each term on the right hand side approaches 0 as $b \to 0$, so that

$$h^2(f_0, \widehat{f}_n) \xrightarrow{a.s.} 0. \qquad \square$$

5.4 Rate of convergence of the log-concave maximum likelihood estimator

The general results described in Section 5.2 are intuitively reasonable. We expect the rate of convergence to depend on the size of the parameter space. This link is provided by the Hellinger bracketing entropy. However, bounding the bracketing entropy of a particular class of functions is well recognised to be a difficult problem. Therefore, in this section we consider a restricted subclass of log-concave densities for which a bound on the bracketing entropy may be derived relatively easily. Combined with sieving, this leads to a rate of convergence for the maximum likelihood estimator within this class. We finish with a small simulation experiment to illustrate the rates achieved in practice, even when the aforementioned restrictions are not satisfied.

5.4.1 Bracketing entropy of the space of log-concave functions

In this section, $\mathscr{F}$ denotes the class of log-concave densities on $\mathbb{R}^d$. Let $S \subseteq \mathbb{R}^d$ be convex and bounded, and let $a < b$ and $c > 0$ be real numbers. Let $\mathscr{L}(a, b, c, S)$ denote the class of all concave functions $l : \mathbb{R}^d \to [0, \infty)$ such that, for some fixed subset $C \subseteq S$, we have

$$\inf_{x \in C} l(x) \geq a \text{ and } \sup_{x \in C} l(x) \leq b,$$

and the restriction of l to C is Lipschitz with constant c. Further, let

$$\mathscr{F}(a, b, c; S) = \left\{f \in \mathscr{F} : \log f \in \mathscr{L}(a, b, c; S)\right\}.$$

A minor modification of the proof of Corollary 2.7.10 in van der Vaart and Wellner (1996) (which concerns the special case $C = S$) allows us to conclude that for $\epsilon > 0$,

$$H(\epsilon, \mathscr{L}(0, 1, c; S), \|\cdot\|_\infty) \leq K \left(\frac{1 + c}{\epsilon}\right)^{d/2},$$

where K depends on S and d only.

There is a one-to-one correspondence between ϵ-balls in $\mathscr{L}(a,b,c;S)$ and $\epsilon/(b-a)$-balls in $\mathscr{L}(0,1,c/(b-a);S)$, so

$$H(\epsilon,\mathscr{L}(a,b,c;S),\|\cdot\|_\infty) \leq K\left(\frac{b-a+c}{\epsilon}\right)^{d/2}.$$

We can use the centres of the balls forming an $\epsilon/2$-covering to form an ϵ-bracketing of $\mathscr{L}(a,b,c;S)$ with respect to $\|\cdot\|_\infty$. If $\{\ell_i\}$ are the centres of the ϵ-balls covering $\mathscr{Y}$, then

$$\left\{[\ell_i^L,\ell_i^U]\right\} = \left\{[\ell_i-\epsilon/2,\ell_i+\epsilon/2]\right\}$$

form an ϵ-bracketing with respect to the supremum norm.

For each $\ell \in \mathscr{L}(a,b,c;S)$, define a function $f:\mathbb{R}^d \to [0,\infty)$ such that

$$f(x) = \begin{cases} \exp(\ell(x)) & \text{if } x \in \mathrm{cl}(\{y : l(y) > a\}) \\ 0 & \text{else.} \end{cases}$$

Define f^L and f^U corresponding to ℓ^L and ℓ^U similarly.

Given an ϵ-bracket $[f^L,f^U]$ such that, for some $f^L \leq f \leq f^U$ we have $\int f = 1$, observe that

$$\begin{aligned} h^2(f^U,f^L) &= \int \left((f^U)^{1/2}-(f^L)^{1/2}\right)^2 \\ &= \int f^U\left(1-\left(\frac{f^L}{f^U}\right)^{1/2}\right)^2 \\ &\leq \frac{\epsilon^2}{4}\int f^U \\ &\leq \frac{\epsilon^2}{4}\int e^\epsilon f \\ &\leq \frac{\epsilon^2}{4}e^\epsilon \\ &\leq \epsilon^2, \end{aligned}$$

provided we choose ϵ so that $e^\epsilon/4 \leq 1$, i.e. $\epsilon \leq \log 4$.

Therefore

$$H_B(\epsilon,\mathscr{F}(a,b,c;S),h) \leq H(\epsilon/2,\mathscr{L}(a,b,c;S),\|\cdot\|_\infty)$$

$$\leq K\left(\frac{b-a+c}{\epsilon}\right)^{d/2}.$$

5.4.2 Rate of convergence

Let $f_0 \in \mathscr{F}(a,b,c;S)$, and

$$\widehat{f}_n = \operatorname*{arg\,max}_{f\in\mathscr{F}(a,b,c;S)} \int \log f \, dF_n.$$

Note that existence and uniqueness of such an estimator may be established in exactly the same way as that of the unrestricted estimator in Theorem 2.10, since

$$\left\{y \in \mathbb{R}^n : \bar{h}_y \in \mathscr{L}(a,b,c;S)\right\}$$

is convex.

Now we may apply Theorem 5.3 to conclude that

$$h(\widehat{f}_n, f_0) = O_p(\delta_n),$$

where

$$\delta_n = \begin{cases} n^{-2/(d+4)} & d < 4 \\ n^{-1/4}(\log n)^{1/2} & d = 4 \\ n^{-1/d} & d > 4. \end{cases}$$

Using the method of sieves from Section 5.2.6, we may improve on the suboptimal rates for $d > 4$. Recall that the rates are essentially determined by the solution to

$$H_B(\epsilon_n, \mathscr{F}(a,b,c;S), h) \asymp n\epsilon_n^2$$

which gives a rate of $O_p(n^{-2/(d+4)})$ for all d.

We conjecture that, even without the additional restrictions imposed in this section, $h(\widehat{f}_n, f_0)$ will be $O(n^{-d/(d+4)})$. However, bounding bracketing entropy, particularly for the Hellinger metric, is well known to be a difficult problem (van de Geer, 2000). This conjecture will be supported by our simulation results in Section 5.4.3, which include a favourable convergence rate with respect to the Hellinger distance even for densities not satisfying the additional restrictions of this section.

5.4.3 Simulation results

In this section, we study the distance $h(\widehat{f}_n, f_0)$ empirically for the densities, dimensions and sample sizes listed in Section 4.1.1. For each density, n and d we computed the Hellinger distance $h(\widehat{f}_n, f_0)$ for 100 samples. Densities E and G are of particular interest because they lie on the boundary of the class of log-concave densities. It has been observed that pointwise convergence occurs at a slower rate where derivatives vanish when $d = 1$ (Balabdaoui et al., 2009). However, we do not observe the same phenomenon here.

Dümbgen and Rufibach (2008) observe that, while their theoretical results only hold for compact intervals, in fact good behaviour is observed over the whole space. In our case, while our theoretical results hold only in a restricted setting, we see good behaviour even when the restrictions are not satisfied.

Recall from Section 2.4.1 that $\widehat{f}_n(x) = \exp(\bar{h}_y(x))$ for some $y \in \mathbb{R}^n$. Clearly

$$\widehat{f}_n^{1/2}(x) = \exp(\bar{h}_{y/2}(x))$$

so computing the integral

$$\int \widehat{f}_n^{1/2} f_0^{1/2}$$

is straightforward using the numerical techniques of Section 4.3. It follows that computing the Hellinger distance

$$\begin{aligned} h^2(\widehat{f}_n, f_0) &= \int (\widehat{f}_n^{1/2} - f_0^{1/2})^2 \\ &= 2 - 2\int \widehat{f}_n^{1/2} f_0^{1/2} \end{aligned}$$

is also straightforward.

Some of these results are shown in Figure 5.1. In Figure 5.1(a), we show the results for density C. We can see that the rate of convergence is slightly slower for larger d, although not as much as predicted by the above asymptotic analysis. We suggest this is a finite sample effect. The other densities exhibited similar performance. In Figure 5.1(c), we show the results for density G for $d = 2$. This density does have compact support. The rate of convergence observed empirically was not affected much by the features noted in Table 4.1.

In Figure 5.1(b), we show the results for density F. This is not a log-concave density, and we conjectured in Section 5.2.7 that in this case $h(\widehat{f}_n, f_0)$ converge to a finite non-zero value. This can be seen for $d = 1$. It is less obvious for higher dimensions, but this

could be because we did not consider a sufficiently large sample size.

In Table 5.1, the right-hand column shows the approximate rate of convergence based on these simulations for density A. The first column gives the rate of convergence conjectured above for comparison. We see that in practice the rate achieved is at least as fast as the rate we have conjectured. We expect that the rate of convergence would be slower for larger d, but computational considerations currently prohibit experimental verification.

d	$-2/(d+4)$	Slope
1	-0.4	-0.465
2	-0.333	-0.420
3	-0.286	-0.379

Table 5.1: Conjectured and empirical rates of convergence for $N_d(0, I)$ random variables, $d = 1, 2, 3$.

5.5 Comparison with kernel density estimation

In this section, we compare the performance of the log-concave maximum likelihood estimator with one common competitor, namely the kernel density estimator introduced in Section 1.2.2. Recall that the multivariate kernel density estimator is defined by

$$\widehat{f}_n(x; H) = \frac{1}{n} \sum_{i=1}^{n} K_H(x - X_i). \tag{5.7}$$

Here H is a positive definite $d \times d$ matrix and K a kernel, typically a spherically symmetric density, and

$$K_H(x) = \frac{1}{|H|^{1/2}} K(H^{-1/2}x).$$

If $d = 1$, we usually write h for the smoothing parameter, and

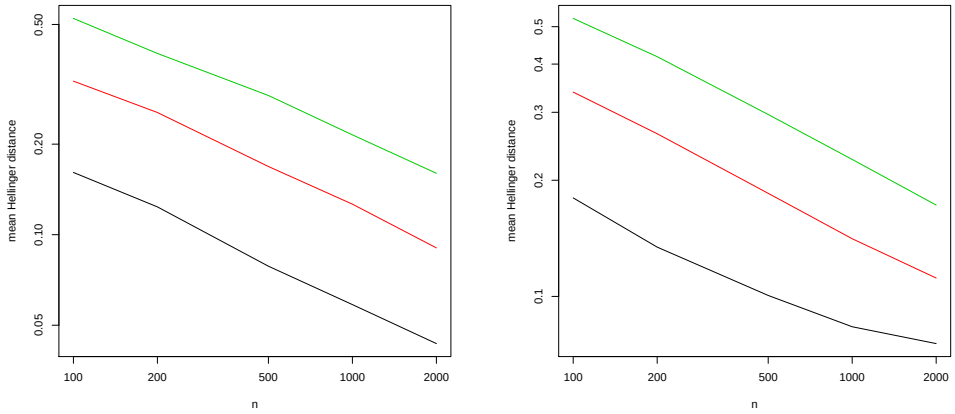
$$\widehat{f}_n(x; h) = \frac{1}{n} \sum_{i=1}^{n} K_h(x - X_i)$$

for the density estimate, where

$$K_h(x) = \frac{1}{h} K\left(\frac{x}{h}\right).$$

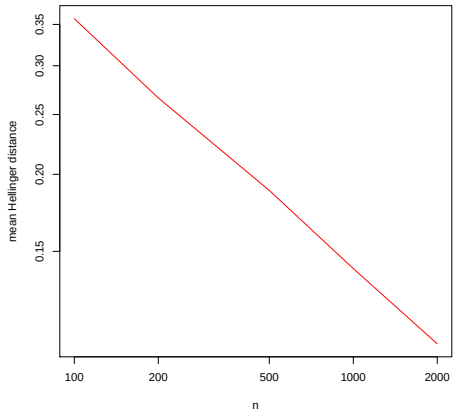
The scalar h or matrix H is known as the bandwidth, and controls the degree of smoothing performed by the density estimator. Choosing a suitable bandwidth is important to ensure

5.5 Comparison with kernel density estimation



(a) Density C

(b) Density F



(c) Density G

Figure 5.1: Hellinger error (estimated). $d = 1$ is black, $d = 2$ red and $d = 3$ green.

good performance of the estimator.

In common with most practitioners, we use a Gaussian kernel

$$K(x) = \frac{1}{(2\pi)^{d/2}} \exp\left(-\frac{1}{2}x^T x\right)$$

though the choice of kernel does not strongly influence the asymptotic performance (Scott, 1992, Section 6.2.3). This kernel is often used for its smoothness (leading to a smooth estimate) and for its computational tractability.

In common with most literature on density estimation, in this section our error

criterion is the mean integrated squared error (MISE)

$$\mathrm{MISE}\left(\widehat{f}_n, f_0\right) = \mathbb{E}\left[\int \left(f_0(x) - \widehat{f}_n(x)\right)^2 dx\right].$$

This is preferred for its ease of computation and intepretability. This is due to the following breakdown into integrated variance and integrated squared bias terms:

$$\mathrm{MISE}\left(\widehat{f}_n, f_0\right) = \int \mathbb{E}\left[\left(\widehat{f}_n(x) - \mathbb{E}\widehat{f}_n(x)\right)^2\right] dx + \int \left(f_0(x) - \mathbb{E}\widehat{f}_n(x)\right)^2 dx \qquad (5.8)$$

As already mentioned, a rate of convergence of $O(n^{-4/(d+4)})$ for the MISE is attainable if H is chosen appropriately. In Section 5.5.1, we discuss bandwidths that in theory allow us to achieve the optimal rate of convergence, at least asymptotically. However, these depend on the unknown density f_0. In Section 5.5.4, we introduce three conceptually simple techniques for data-driven bandwidth selection. In Section 5.5.5, we suggest an extension allowing H to vary over the sample space and discuss how this influences the performance of these estimators. Finally, we present some simulation results in Section 5.5.6 comparing the performance of a kernel density estimator with that of the log-concave maximum likelihood estimator.

For a discussion of the added difficulties for $d > 1$, see Duong (2004). Multivariate bandwidth selection is an area of active research. The main references for Section 5.5.4 are Duong and Hazelton (2003), Duong and Hazelton (2005) and Duong (2007b). Other recent innovations include Zhang, King, and Hyndman (2006), Chacón, Wand, and Duong (2008) and Chacón (2009). Most of there results are generalizations, at least in spirit, of well-studied univariate techniques. A survey of some alternative multivariate techniques can be found in Ćwik and Koronacki (1997).

Extensions that adapt to local smoothness began with Breiman, Meisel, and Purcell (1977) and Abramson (1982). A review of several adaptive methods for univariate data may be found in Sain and Scott (1996). Further multivariate adaptive methods are presented in Sain (2002) and Scott and Sain (2004).

5.5.1 Theoretically optimal bandwidths

The MISE of a univariate kernel density estimator with kernel K and bandwidth h is given by

$$\begin{aligned}\mathrm{MISE}\left(\widehat{f}_n(\cdot,h),f_0\right) = \frac{1}{nh}\int K^2(x)\,dx + \frac{n-1}{n}\int (K_h \star f_0)^2(x)\,dx \\ - 2\int (K_h \star f_0)(x) f_0(x)\,dx + \int f_0^2(x)\,dx.\end{aligned} \tag{5.9}$$

Since this depends on the bandwidth h in a complicated way, it is common to use an asymptotic expansion of the MISE, for which analysis is easier. For $d = 1$, the asymptotic mean integrated squared error (AMISE) of a kernel density estimate with kernel K and bandwidth h is given by

$$\mathrm{AMISE}\left(\widehat{f}_n(\cdot,h),f_0\right) = \frac{1}{nh}R(K) + \frac{1}{4}h^4 m_2(K)^2 R(f_0'') \tag{5.10}$$

where for a density g with finite second moment,

$$m_2(g) = \int x^2 g(x)\,dx$$

and for a square integrable function h

$$R(h) = \int h^2(x)\,dx.$$

In this case, the optimal bandwidth is given by

$$h_n = \left[\frac{R(K)}{m_2(K)^2 R(f_0'')}\right]^{1/5} n^{-4/5}.$$

Plugging this into (5.10), we see that the optimal AMISE is

$$\inf_{h>0} \mathrm{AMISE}\left(\widehat{f}_n(\cdot,h),f_0\right) = \frac{5}{4}\left[m_2(K)^2 R(K)^4 R(f_0'')\right]^{1/5} n^{-4/5}.$$

This is consistent with the rate of $O(n^{-4/(d+4)})$ mentioned in the introduction. Observe that the optimal bandwidth depends on the density f_0.

In the special case that f_0 is a finite Gaussian mixture and K a Gaussian kernel, it is possible to compute the convolution integrals in (5.9) exactly and a MISE-optimal bandwidth may be computed by numerically optimizing (5.9) (Wand and Jones, 1995,

Section 2.6).

In the multivariate case, the calculations are somewhat more intricate. The AMISE is given by

$$\text{AMISE}\left(\widehat{f}(\cdot;H), f_0\right) = \frac{1}{n}R(K)\,|H|^{-1/2} + \frac{1}{4}m_2(K)[\text{vech}\,H]^T\Psi(f_0)[\text{vech}\,H] \qquad (5.11)$$

where vech is the vector half operator. For a general $m \times m$ matrix $A = (a_{ij})$, this stacks the columns of the lower triangle, that is,

$$\text{vech}\,A = (a_{11}, a_{21}, \ldots, a_{m1}, a_{22}, a_{32}, \ldots, a_{mm})^T.$$

The matrix $\Psi(f)$ is defined as

$$\Psi(f) = \int w(x)w(x)^T\,dx,$$

where

$$w(x) = \text{vech}\left(2\nabla\nabla^T f(x) - \text{diag}\,\nabla\nabla^T f(x)\right),$$

and for a matrix A,

$$(\text{diag}\,A)_{ij} = \begin{cases} A_{ii} & \text{if } i = j \\ 0 & \text{else.} \end{cases}$$

For smooth densities f, it may be shown that $\Psi(f)$ depends on f through quantities of the form

$$\psi_r(f) = \int f^{(r)}(x)f(x)\,dx. \qquad (5.12)$$

where $r = (r_1, \ldots, r_d)$, $|r| = \sum_{i=1}^d r_d$ and

$$f^{(r)}(x) = \frac{\partial^{|r|}}{\partial x_1^{r_1}, \ldots, x_d^{r_d}} f(x).$$

Once more, in the special case that f_0 is a finite Gaussian mixture no asymptotic expansion is required (Wand and Jones, 1995, Section 4.4).

If $d > 1$, it is not possible to find a general expression for the AMISE-optimal H, so (5.11) must be minimized numerically.

5.5.2 Restrictions and pre-transformation

In the above discussion, we have made no restriction beyond requiring that H be positive definite. Additional restrictions on the shape of H can simplify the computation considerably. For example, writing $\mathscr{H}_3$ for the class of all positive-definite matrices, further restrictions include

$$H \in \mathscr{H}_2 = \{H : H \text{ diagonal and } H \in \mathscr{H}_3\}$$

or

$$H \in \mathscr{H}_1 = \left\{H : H = h^2 I \text{ for some } h > 0\right\}.$$

We have already remarked that, in general, there is no closed-form expression for the AMISE-optimal bandwidth matrix. However, in the special case $H \in \mathscr{H}_1$, we may show that the optimal bandwidth matrix is

$$h = h_n = \left[\frac{dR(K)}{m_2(K)^2 R(\nabla^2 f) n}\right]^{1/(d+4)}.$$

These restrictions can lead to inferior performance, and are not generally recommended unless the data are known to have a simple structure (Duong, 2007b).

To improve the performance of the kernel density estimate, pre-transforming the data has been suggested Silverman (1986). The idea is to compute a bandwidth matrix (in $\mathscr{H}_1$ or $\mathscr{H}_2$) for the transformed dataset, which may be transformed back to the original scale. This aim is to enjoy the advantages of a simpler bandwidth matrix in terms of computation and stability, while adapting correctly to the underlying structure of the data.

Two pre-transformations are considered. Sphering transforms the data so that the covariance is I, that is, $X^* = S^{-1/2}X$, where S is the sample covariance matrix. Scaling transforms only the marginal variances, that is, $X^* = (\operatorname{diag} S)^{-1/2}X$, and does not alter the correlation structure. The bandwidth matrix H^* for the transformed data may be translated back to one for the original dataset by inverting the transformation ($H = S^{1/2}H^*S^{1/2}$ or $H = (\operatorname{diag} S)^{1/2}H^*(\operatorname{diag} S)^{1/2}$ for sphering and scaling respectively). However, Duong (2007b) advises caution using this approach, especially in conjunction with the assumption $H^* \in \mathscr{H}_1$, since the sample covariance may not adequately capture the structure of the dataset.

5.5.3 Normal scale rule

A crude method for practical bandwidth selection is the normal scale rule. This method uses the AMISE-optimal bandwidth for normal data on the same scale as the sample. For multivariate data, we assume the components are independent and that $H \in \mathscr{H}_2$. Writing $(h^i)^2$ for the ith component of $\operatorname{diag} H$, we set

$$h_n^i = \widehat{\sigma}_i \left(\frac{4}{d+2} \right)^{1/(d+4)} n^{-1/(d+4)}$$

where $\widehat{\sigma}_i$ is some estimate of the marginal standard deviation, for example the sample marginal standard deviation. This has a tendency to oversmooth the data but is a useful rule of thumb for plug-in estimators, where the choice of bandwidth is less critical.

5.5.4 Fixed bandwidth selectors

As discussed in Section 5.5.1, the (asymptotically) optimal bandwidth for a multivariate density can be found by minimizing (5.11). However, this bandwidth depends on f_0, which is unknown in practice. For univariate data, there are two main classes of bandwidth selector used in practice: plug-in selectors and cross-validation selectors. Multivariate counterparts have been developed for both types, with various restrictions on H (Duong and Hazelton, 2003, 2005).

Plug-in methods aim to approximate (5.11) by "plugging in" preliminary estimates of the unknown quantities. Terms of the form (5.12) may be written $\psi_r = \mathbb{E}[f^{(r)}(X)]$, so we have a natural estimator of ψ_r (depending on a so-called "pilot" bandwidth G)

$$\begin{aligned} \widehat{\psi}_r(G) &= \frac{1}{n} \sum_{i=1}^{n} \widehat{f}_n^{(r)}(X_i; G) \\ &= \frac{1}{n^2} \sum_{i=1}^{n} \sum_{j=1}^{n} K_G^{(r)}(X_i - X_j). \end{aligned}$$

We typically assume $G \in \mathscr{H}_1$, leaving a single parameter to select. Although this assumption is problematic when choosing H, it appears not to matter so much for the pilot estimate. Duong and Hazelton (2003) observed choosing a different G for each r can lead to bandwidth matrices that are not positive definite, and proposed choosing a single pilot bandwidth matrix G according to the sum of asymptotic mean squared error (SAMSE) criterion

$$\operatorname{SAMSE}_k(G) = \sum_{r:\, |r|=k} \mathbb{E}[(\widehat{\psi}_r(G) - \psi_r)^2].$$

Duong and Hazelton (2003) provide a closed-form expression for the optimal value

of G, and show that if $|r| = k$, it depends on terms of the type ψ_s where $|s| = k+2$. These terms may be estimated using another plug-in estimator, requiring a further pilot bandwidth G_2. This may be iterated indefinitely. In practice, we choose some fixed number of stages m and use a simple estimate of the corresponding ψ_r, which may be plugged in the previous stages. Duong and Hazelton (2003) suggest a normal scale rule

$$\widehat{\psi}_r^{NR} = (-1)^{|r|}\phi_{2S}^{(r)}(0),$$

where ϕ_Σ is a normal density with mean 0 and covariance Σ, and S the sample covariance matrix, after 1 or 2 stages.

The second class of bandwidth selectors uses cross-validation to choose the optimal bandwidth. This approach may be motivated by the alternative integrated squared error (ISE) criterion

$$\begin{aligned}\mathrm{ISE}\left(\widehat{f}_n, f_0\right) &= \int \left(f_0(x) - \widehat{f}_n(x)\right)^2 dx \\ &= \int f_0^2(x)\,dx - 2\int f_0(x)\widehat{f}_n(x)\,dx + \int \widehat{f}_n^2(x).\end{aligned}$$

The first term does not depend on the estimated density, so may be ignored. In the second, we replace the unknown f_0 with an estimate of the density. For least squares (or unbiased) cross validation (LSCV), we use the leave-one-out density estimate

$$\frac{1}{n}\sum_{j=1}^{n}\widehat{f}_n^{(-j)}(\cdot, H),$$

where

$$\widehat{f}_n^{(-j)}(x; H) = \frac{1}{n-1}\sum_{k\neq j} K_H(x - X_k)$$

is the kernel estimate based on the sample excluding X_j. This leads to the LSCV criterion

$$\mathrm{LSCV}(H) = \int \widehat{f}_n^2(x; H)\,dx - \frac{2}{n}\sum_{i=1}^{n}\widehat{f}_n^{(-i)}(X_i; H).$$

The bandwidth matrix is chosen by minimizing this criterion. Restriction to H in $\mathscr{H}_1$ or $\mathscr{H}_2$ is straightforward. This is a simple criterion, and is unbiased in the sense that

$$\mathbb{E}\left[\mathrm{LSCV}(H) + \int f_0^2\right] = \mathrm{MISE}\left(\widehat{f}_n(\cdot; H), f_0\right).$$

However, it can be quite unstable (especially allowing $H \in \mathscr{H}_3$) and tends to undersmooth

the data (Wand and Jones, 1995).

For a Gaussian kernel, the smoothed cross-validation criterion is given by

$$\mathrm{SCV}(H) = \frac{1}{n}R(K)\,|H|^{-1/2} + \frac{1}{n^2}\sum_{i=1}^{n}\sum_{j=1}^{n}\left[K_{2H+2G}(X_i - X_j) - 2K_{H+2G}(X_i - X_j) + K_{2G}(X_i - X_j)\right],$$

where once more G is a pilot bandwidth. In the degenerate case $G = 0$, this is closely related to LSCV, so the SCV criterion may be interpreted as a smoothed version of LSCV, with G controlling the degree of smoothing. The above discussion on selection of a pilot bandwidth matrix also applies in this case.

A further variant, biased cross-validation, was also explored in Duong and Hazelton (2005), but will not be pursued here because of its poor practical performance and heavy computational burden.

5.5.5 Variable bandwidth selectors

A single bandwidth matrix is necessarily a compromise between the requirements of different regions. Therefore, an obvious extension is to allow the bandwidth matrix to vary over the sample space, that is, to replace the estimator (5.7) with

$$\widehat{f}_n(x; H(\cdot)) = \frac{1}{n}\sum_{i=1}^{n} K_{H(X_i)}(x - X_i).$$

This has the advantage that the degree of smoothing can adapt to the amount of data in each region: where data are sparse, we may use a larger bandwidth (to reduce the variance), whereas in regions of high intensity we may choose a smaller bandwidth (to reduce the bias).

In its most general form, this estimator requires the specification of $nd(d+1)/2$ parameters. For a practical calibration some additional structure is therefore required. Breiman et al. (1977) suggested choosing a bandwidth proportional to $f^{-1/d}(X_i)$; Abramson (1982) suggest using $f^{-1/2}(X_i)$ independent of d. In practice, this means the bandwidth matrix should be chosen to be $hf^{-1/2}(X_i)A$, where A is some shape matrix and h a parameter controlling the global degree of smoothing. Choosing $A = I$ or A diagonal leads to the restrictions already discussed. An alternative is to scale or sphere the data as discussed Section 5.5.2.

In order to use this method in practice, a pilot estimate for f is required. As in Section 5.5.4, choice of pilot bandwidth for the pilot estimate appears to be less critical than that for the density estimate. We use a diagonal bandwidth matrix and a normal

scale rule (Section 5.5.3) for the pilot bandwidth. We also pre-scale the data. This is equivalent to choosing $A = \text{diag}(S)$. The global smoothing parameter h is then chosen by cross-validation.

The Abramson estimator allows the degree of smoothing to adapt to the height of the density, but makes no allowance for curvature. For this reason, Sain (2002) suggested partitioning the data into m bins and using a constant bandwidth matrix on each partition. The bandwidth matrices $H_1, \ldots, H_m$ may then be chosen by least squares cross-validation. Restriction to $H_i \in \mathscr{H}_1$ or $H_i \in \mathscr{H}_2$ is possible to simplify computation and improve stability. Duong (2004) investigated several binning rules and concluded that calibrating this density estimate was very difficult.

5.5.6 Simulation results

In this section, we study the MISE empirically for the densities, dimensions and sample sizes listed in Section 4.1.1 to compare the performance of the log-concave maximum likelihood estimator and various kernel estimators. It is important to remember that the log-concave estimator is fully automatic, requiring no tuning parameters. For the kernel estimator, on the other hand, a large number of decisions must be made by the practitioner: a fixed or variable bandwidth, whether to pre-transform the data, whether to choose bandwidth matrix or matrices in $\mathscr{H}_1, \mathscr{H}_2$ or $\mathscr{H}_3$, and which selection criterion to use. In the case of a variable bandwidth, the situation is even more complicated, with the additional choice of pilot estimate for the Abramson estimator and choice of binning rule for the Sain estimator. It is outside the scope of this thesis to examine every possibility. As we shall see, the choices made can have a big impact on the performance of the estimator, and different methods may be appropriate in different situations. For a detailed discussion of the various options, we refer the reader to Duong and Hazelton (2003, 2005); Scott and Sain (2004).

We illustrate the following bandwidth selectors.

LSCV Least squares cross-validation, $H \in \mathscr{H}_3$

SCV Smoothed cross-validation, $H \in \mathscr{H}_3$

PI 2-stage plug-in with pilot bandwidth chosen by SAMSE and with pre-sphering for the pilot bandwidth

Abr Abramson estimator with normal scale rule for choice of pilot bandwidth, pre-scaling and global smoothing parameter chosen by LSCV

Sain Binned Sain estimator with a fixed number (7) of equally spaced bins in each dimension, $H_i \in \mathscr{H}_1$ for $i = 1, \ldots, 7^d$, and bandwidths chosen by LSCV

Bandwidths LSCV, SCV and PI were computed using the package **ks** (Duong, 2007a). Code to compute Abr and Sain was based on code kindly provided by Tarn Duong. The log-concave estimator was computed using the package **logcondens** for $d = 1$ and **LogConcDEAD** for $d > 1$. For each density, dimension and sample size, 100 Monte Carlo simulations were performed. For each, the ISE was estimated using a similar method to Section 4.3 for the log-concave maximum likelihood estimator. For finite Gaussian mixtures, the ISE may be computed exactly using formulae in Wand and Jones (1995). For others, Monte Carlo integration was used to compute the ISE. The MISE was estimated using the mean of these samples.

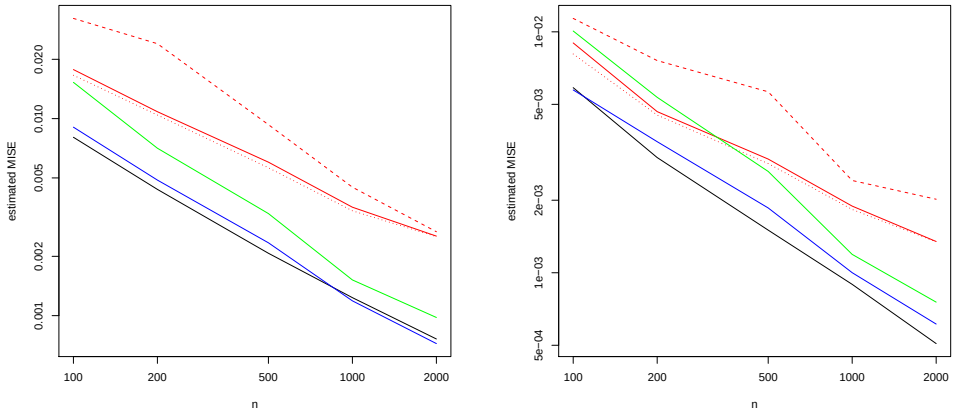
Figure 5.2 shows 2 examples for $d = 1$. In Figure 5.2(a), we see that even for a univariate normal density, the log-concave maximum likelihood estimator performs at least as well as the other estimators. The only competitive kernel estimator is the Abramson estimator. The Sain estimator is worse. Fixed bandwidth selectors all perform worse than the log-concave maximum likelihood estimator, with PI and SCV displaying similar performance, and LSCV being considerably worse for small sample sizes.

Figure 5.2(b) shows a similar performance for density E, which is "only just" log-concave. Again, the log-concave estimator performs best. Among the kernels, the Abramson estimator performs best, with LSCV worst. PI and SCV are similar. The Sain estimator improves with sample size but is never as good as the Abramson estimator and never close to the log-concave maximum likelihood estimator.

Figure 5.3 shows two examples for $d = 2$. Figure 5.3(a) shows density B which has some dependence structure. In this case, the performance of PI, SCV, the Abramson estimator and the log-concave maximum likelihood estimator are similar. For larger sample sizes, the log-concave estimator is slightly better. Again, LSCV and Sain perform worst.

Figure 5.3(b) shows the results for density G. We expect this situation to be more difficult for the kernel density estimator due to the boundary bias. The performance of all the fixed bandwidth estimators is similar. The Abramson estimator performs best among the kernel estimators but its rate of decrease is slow. Once again the Sain estimator is worse, especially for smaller values of n. The log-concave maximum likelihood estimator, on the other hand, does not suffer from these problems. Although initially the error is larger than that of the Abramson estimator, it decreases much faster, and does not appear to have the same difficulties due to the boundary effects.

Figure 5.4 shows two 3-dimensional examples. As in Figure 5.3(a), we see that the Abramson estimator has lost the advantage over PI and SCV that it enjoyed when $d = 1$. PI, SCV and Abr have similar performance, with LSCV and Sain being worse, especially for smaller n. The log-concave estimator is initially slightly worse than PI, SCV and Abr



(a) Density A (b) Density E

Figure 5.2: Estimated MISE, $d = 1$. Black is log-concave maximum likelihood estimator. The rest are kernel density estimators, with bandwidth chosen by PI (red solid), LSCV (red dashed), SCV (red dotted), Abr (blue) or Sain (green).

but soon recovers its advantage. The situation in Figure 5.4(b), which illustrates density D, is almost identical.

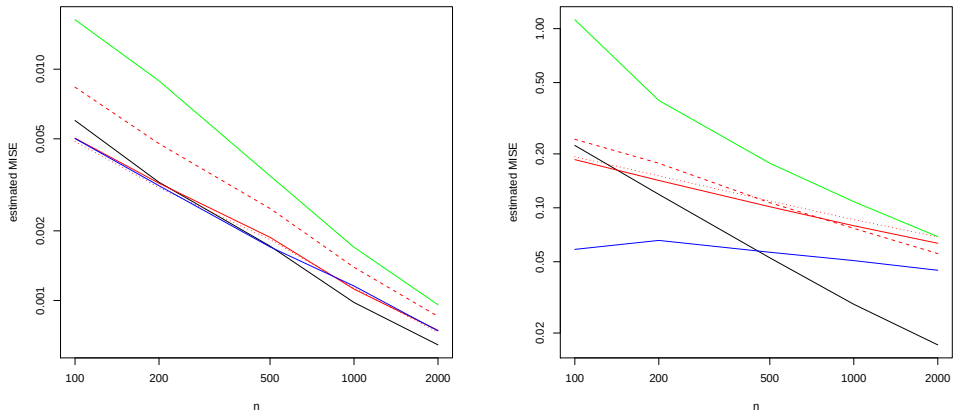
Finally, Figure 5.5 shows what happens when the assumption of log-concavity is violated. By analogy with our conjectures in Section 5.2.7, we might expect the MISE of the log-concave maximum likelihood estimator to converge to a nonzero value. This appears to be the case. Once more, PI, SCV and Abr all have similar performance, with LSCV being slightly worse and Sain more erratic, especially for small n.

Summarising our findings from this small study, we see that, when the true density is log-concave, the log-concave maximum likelihood estimator performs at least as well as a kernel density estimator. The difference can be quite dramatic if there are strong boundary effects or the data are skewed. PI and SCV had similar performance, and, as expected from previous work (summarized in Wand and Jones (1995)), LSCV did not perform as well as these methods. We also compared with adaptive bandwidth methods. When $d = 1$, the Abramson estimator gave an improvement over PI and SCV, but this was mostly lost for $d > 1$. The Sain estimator performed poorly.

5.6 Conclusion

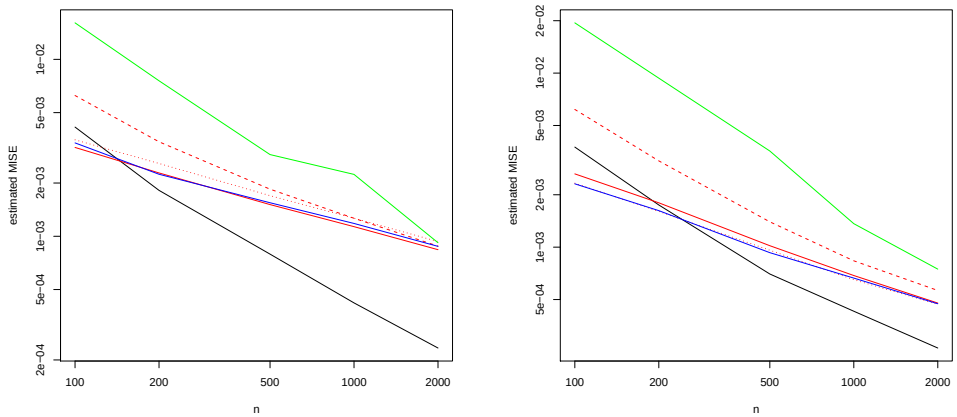
In this chapter, we have studied the asymptotic performance of the log-concave maximum likelihood estimator from two different points of view.

In the first part, we discussed some standard results concerning the rate of conver-



(a) Density B (b) Density G

Figure 5.3: Estimated MISE, $d = 2$. Black is log-concave maximum likelihood estimator. The rest are kernel density estimators, with bandwidth chosen by PI (red solid), LSCV (red dashed), SCV (red dotted), Abr (blue) or Sain (green).



(a) Density C (b) Density D

Figure 5.4: Estimated MISE, $d = 3$. Black is log-concave maximum likelihood estimator. The rest are kernel density estimators, with bandwidth chosen by PI (red solid), LSCV (red dashed), SCV (red dotted), Abr (blue) or Sain (green).

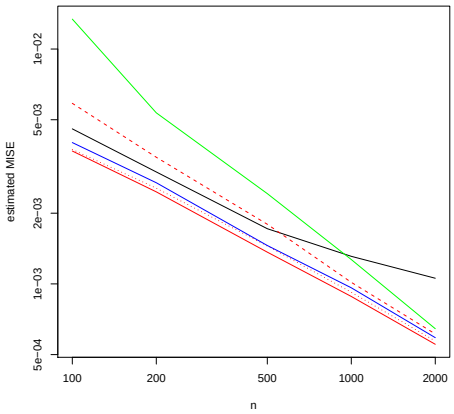


Figure 5.5: Estimated MISE, density F, $d = 2$. Black is log-concave maximum likelihood estimator. The rest are kernel density estimators, with bandwidth chosen by PI (red solid), LSCV (red dashed), SCV (red dotted), Abr (blue) or Sain (green).

gence of maximum likelihood estimators with respect to the Hellinger distance. We then proved a uniform law of large numbers for a suitable class of sets, which enabled us to extend the results of Pal et al. (2007) and prove consistency of the multivariate log-concave maximum likelihood estimator. We proved a rate of convergence of $O_p(n^{-2/(d+4)})$ with respect to the Hellinger metric for a restricted model and conjectured that this will also hold for the full class of log-concave densities. We heuristically discussed model misspecification. Our conjectures about rate of convergence were supported by simulation results for several example densities.

In the second part, we compared our estimator with a kernel density estimator using the MISE error criterion. Bandwidth selection is well known to be crucial for this problem, so we introduced several different bandwidth selectors and compared them across a range of densities, dimensions and sample sizes. These included fixed bandwidth selectors (plug-in, least squares cross-validation and smoothed cross-validation) and two adaptive bandwidths (Abramson and Sain). We found that the log-concave maximum likelihood estimator performed at least as well as the kernel density estimator, even with an adaptive bandwidth. In fact the performance of the Sain estimator was somewhat disappointing, with its ability to adapt to the amount of data in each region cancelled out by the difficulty of calibrating the bandwidths in practice. It may be possible to improve performance using a different binning rule, but this is highly problem-dependent. We agree that with Duong (2004) that, although in principle it would be desirable to adapt to the amount of data in a region, in practice this is very difficult. We also saw support for our conjecture of convergence in the case of misspecification.

In summary, we have opened the door for further work on the theoretical performance of log-concave maximum likelihood estimators, and demonstrated good performance (at least comparable with a widely used competitor) in practice in a range of situations. There was a particular advantage where boundary effects are important. Moreover, we saw that choice of bandwidth has a big influence on the performance of the kernel estimator. There are a range of options, and no one method is best in all situations, leading to a difficult decision for the practitioner. The log-concave maximum likelihood estimator typically performs at least as well as the best kernel density estimator, and sometimes much better, when the density is log-concave.

6 Conclusion

We have explored a novel approach to nonparametric density estimation, using maximum likelihood under the shape restriction of log-concavity. We were motivated by the problems of existing methods of density estimation, and aimed to capture both the flexibility of nonparametric methods and the parsimony of parametric models. We began by surveying some recent developments in shape constrained maximum likelihood estimation, and saw that novel theoretical and computational tools are needed to extend existing methods.

We discussed log-concavity and its consequences. We argued that this is a natural shape restriction for many situations, with desirable properties, and gave some examples of the many parametric families that fall within this model. We proved that a unique log-concave maximum likelihood estimator exists for multivariate data. In the course of this proof we gained insight into the structure of the estimator which was crucial to the rest of the project.

A large part of this work was the development of a stable numerical procedure to compute the log-concave maximum likelihood estimator for multivariate data. We discussed in detail how to exploit the special structure of the maximum likelihood estimator to achieve this. We mentioned methods for one-dimensional data and discussed why they cannot be directly extended to the multivariate case. We considered other computational aspects, including sampling from the maximum likelihood estimator and evaluation of conditional and marginal densities. This was implemented in R in the package **LogConcDEAD** (Log-Concave Density Estimation in Arbitrary Dimensions).

We then turned our attention to inference using the log-concave maximum likelihood estimator. A key development was a method of assessing the suitability of our assumption of log-concavity, in the form of a hypothesis test against the general alternative of local regions of log-convexity. This successfully detected two key departures from log-concavity, heavy tails and mixing, for simulated datasets. We applied this test to a dataset of university rankings and found a single component model to be inadequate.

We presented several examples of functional estimation and demonstrated comparable or improved performance compared to a kernel density estimator for differential entropy and highest density region estimation. We saw that the log-concave maximum likelihood estimator often gives rise to estimates with lower variance, although sometimes at the price of inflated bias.

6 Conclusion

We discussed an extension to finite mixtures of log-concave densities, and proposed an EM-style algorithm which aims to find approximate local maxima in the likelihood function. We used this to fit a 2-component mixture to the universities dataset of the previous section. One major application of mixture models is clustering, and we saw that this method may be used in a similar way to a Gaussian mixture to cluster unlabelled data. We illustrated this using the Wisconsin breast cancer dataset, and saw improved performance over a Gaussian mixture. We attribute this to the fact that the log-concave components can better capture the shape of the distribution, as they are not forced to compensate for skewness by overestimating the variance.

In the last chapter, we investigated the asymptotic performance of the log-concave maximum likelihood estimator. We introduced standard tools for the study of nonparametric maximum likelihood estimators, and used these to generalize an earlier result for one-dimensional data to prove Hellinger consistency of our estimator. We also mentioned the rate of convergence for a related, restricted class of estimators. We conjecture that the same rate holds for the class of all log-concave densities. This was supported by simulation results for various log-concave densities, and is suggested as an area for further investigation. Behaviour in the case of model misspecification was also discussed.

Finally, we compared the performance of our estimator with that of a common competitor, the kernel density estimator introduced in the first chapter. Because the choice of bandwidth is critical for good performance of the kernel density estimator, we included several possible bandwidth selectors in our comparison. We demonstrated that our estimator performs at least as well as the best of our kernel density estimators in most situations, and considerably better where boundary effects are significant or the data are skewed. We included in our comparison two bandwidth selectors that aim to adapt to the amount of data in a region. This did not lead to much improvement in performance because in practice calibrating these bandwidths proved difficult. This highlights one of the main advantages of our estimator, namely its fully automatic nature.

We hope that this work will stimulate future activity in this area, as we have seen that log-concave density estimation is an attractive alternative to existing methods. We suggest further development of computational methods would make this method more widely applicable. This would allow the use of more numerically intensive techniques, such as the bootstrap, to assess uncertainty of functional estimates and obtain empirical confidence bounds. Obtaining stronger or more general theoretical results is the second area of particular interest and current activity.

A Definitions and background material

This section is a review of fundamental concepts from convex analysis and computational geometry used in the text. Key references are Rockafellar (1997) for convex analysis and Bern (2004), Seidel (2004), Snoeyink (2004) and Gelfand et al. (1994) for computational geometry and triangulation of point sets. Many more references may be found in these surveys.

A.1 Convex analysis

Definition A.1 (Affine set)**.** A set $A \subseteq \mathbb{R}^d$ is affine if

$$\lambda x + (1-\lambda)y \in A$$

for every $x, y \in A$ and $\lambda \in \mathbb{R}$.

Definition A.2 (Convex set)**.** A set $C \subseteq \mathbb{R}^d$ is convex if, for all $x, y \in C$ and $\lambda \in [0,1]$, we have

$$\lambda x + (1-\lambda)y \in C.$$

Definition A.3 (Convex or concave function)**.** A function $f : \mathbb{R}^d \to (-\infty, \infty]$ is convex if $f(x) < \infty$ for at least one $x \in \mathbb{R}^d$ and, for all $x, y \in \mathbb{R}^d$ and $\lambda \in [0,1]$, we have

$$f(\lambda x + (1-\lambda)y) \leq \lambda f(x) + (1-\lambda) f(y).$$

We say f is concave if $-f$ is convex.

Note that if f is a convex function from some (convex) set $C \subseteq \mathbb{R}^d$ to $(-\infty, \infty]$, we may extend f to a convex function on all of $\mathbb{R}^d$ by setting $f(x) = \infty$ for $x \notin C$ (for the corresponding extension of a concave function, we set $f(x) = -\infty$ for $x \notin C$). Therefore there is no loss of generality in assuming the domain of a convex or concave function is all of $\mathbb{R}^d$. This corresponds to what Rockafellar (1997) calls a proper convex function.

A.1 Convex analysis

Definition A.4 (Epigraph)**.** The epigraph of a convex function $f : \mathbb{R}^d \to (-\infty, \infty]$ is

$$\operatorname{epi} f = \{(x, \mu) \in \mathbb{R}^{d+1} : f(x) \leq \mu\}.$$

For a concave function $f : \mathbb{R}^d \to [-\infty, \infty)$,

$$\operatorname{epi} f = \{(x, \mu) \in \mathbb{R}^{d+1} : f(x) \geq \mu\}.$$

Definition A.5 (Log-concave function)**.** A function $f : \mathbb{R}^d \to [0, \infty)$ is log-concave if $\log f$ is concave (Definition A.3) (with the convention that $\log 0 = -\infty$).

Definition A.6 (Quasiconcave or quasiconvex)**.** A function $f : \mathbb{R}^d \to (-\infty, \infty]$ is quasiconvex if, for every $a \in \mathbb{R}$,

$$\{x : f(x) \leq a\}$$

is a convex set. f is quasiconcave if $-f$ is quasiconvex.

This is sometimes used as a definition of multivariate unimodality; see Section 2.2.7 for a discussion.

Definition A.7 (Effective domain)**.** The effective domain of a function $f : \mathbb{R}^d \to (-\infty, \infty]$ is

$$\operatorname{dom} f = \left\{x \in \mathbb{R}^d : f(x) < \infty\right\}$$

Definition A.8 (Closed convex function, closure of a convex function)**.** A convex (resp. concave) function is closed if its epigraph is a closed set. The closure of a convex (resp. concave) function is the function $\operatorname{cl} f$ such that $\operatorname{epi} \operatorname{cl} f = \operatorname{cl} \operatorname{epi} f$.

Definition A.9 (Least concave majorant)**.** The least concave majorant of a function $f : \mathbb{R}^d \to \mathbb{R}$ is the smallest concave function g such that $g(x) \geq f(x)$ for all $x \in \mathbb{R}$.

Definition A.10 (Convex hull)**.** The convex hull of a set $S \subseteq \mathbb{R}^d$ is

$$\operatorname{conv}(S) = \bigcap \left\{C \subseteq \mathbb{R}^d : S \subseteq C \text{ and } C \text{ convex}\right\},$$

that is, the smallest convex set containing S.

Definition A.11 (Affine hull)**.** The affine hull of a set $S \subseteq \mathbb{R}^d$ is

$$\operatorname{aff} S = \bigcap \left\{A \subseteq \mathbb{R}^d : S \subseteq S \text{ and } A \text{ affine}\right\},$$

that is, the smallest affine set containing S.

Definition A.12 (Affine independence)**.** A collection of $m+1$ points $X_0, \ldots, X_m$ in $\mathbb{R}^d$ is affinely independent if

$$\dim \operatorname{aff}\{X_0, \ldots, X_m\} = m.$$

Definition A.13 (General position)**.** A collection $\mathscr{X} = \{X_1, \ldots, X_n\}$ of points in $\mathbb{R}^d$ are in general position if, for every $\mathscr{Y} \subseteq \mathscr{X}$ of size $d+1$, $\dim \operatorname{conv}(\mathscr{Y}) = d$.

Definition A.14 (Relative Interior)**.** The relative interior of a convex set C is its interior when regarded as a subset of its affine hull.

Definition A.15 (Polytope)**.** A polytope is the convex hull of a finite number of points in $\mathbb{R}^d$.

Definition A.16 (Face)**.** The intersection of a polytope P and the boundary hyperplane of a halfspace containing P is called a face of P. Facets are faces of dimension $\dim P - 1$. Vertices are faces of dimension 0.

Definition A.17 (Simplex)**.** A simplex is the convex hull of a set $\mathscr{X}$ of $d+1$ affinely independent points in $\mathbb{R}^d$, denoted $\sigma(\mathscr{X})$, say.

The unit simplex in $\mathbb{R}^d$ is

$$\sigma\left(\{0, e_1, \ldots, e_d\}\right)$$

where e_i denotes the ith coordinate vector, and 0 denotes the origin. This is denoted by T_d.

Definition A.18 (Cone)**.** A cone is a set $C \subseteq \mathbb{R}^n$ which is under positive scalar combination, i.e. with the property that, for each $x, y \in C$ and scalars $\lambda, \mu > 0$, $\lambda x + \mu y \in C$.

A.2 Computational geometry

Definition A.19 (Subdivision)**.** A subdivision of a finite point set $\mathscr{X} \subseteq \mathbb{R}^d$ is a collection of subsets $\mathscr{V} = \{V_1, \ldots, V_m\}$ of $\mathscr{X}$ such that

1 $\operatorname{conv}(V_i)$ is a d-dimensional polytope for each $i = 1, \ldots, m$

2 $\bigcup_{i=1}^{m} \operatorname{conv}(V_i) = \operatorname{conv}(\mathscr{X})$

3 If $i \neq j$ then $\operatorname{conv}(V_i \cap V_j)$ is a face of both $\operatorname{conv}(V_i)$ and $\operatorname{conv}(V_j)$.

Note that not every point in $\mathscr{X}$ need appear in some V_i.

Definition A.20 (Triangulation)**.** A triangulation of a point set $\mathscr{X} \subseteq \mathbb{R}^d$ is a subdivision such that each V_i has $d+1$ elements. Note that, as in Definition A.19, not every point in $\mathscr{X}$ need appear in some V_i.

Definition A.21 (Refinement)**.** A subdivision $\mathscr{V}$ refines $\widetilde{\mathscr{V}}$ if, for every $V \in \mathscr{V}$, there is some $\widetilde{V} \in \widetilde{\mathscr{V}}$ such that $V \subseteq \widetilde{V}$.

Definition A.22 (Flip)**.** Given a set $\mathscr{X} = \{X_1, \ldots, X_{d+2}\}$ of size $d+2$, there are precisely two ways to triangulate $\mathscr{X}$, corresponding to the lower and upper hulls of the points

$$\left\{(X_1, \left\|X_1\right\|^2), \ldots, (X_{d+2}, \left\|X_{d+2}\right\|^2)\right\}$$

in $\mathbb{R}^{d+1}$ (Lawson, 1986). An example is given in Figure 3.1. A flip is an operation on (a subset of) the simplices in a triangulation that, if possible without altering the rest of the triangulation, replaces one of these within a triangulation with the other.

A.2.1 Construction of triangulations

We construct a triangulation refining $\mathscr{S}(y)$ by finding the convex hull of points in $\mathbb{R}^{d+1}$. There are many possible algorithms; see Seidel (2004) for some alternatives. We use the **Quickull** algorithm. The following description follows closely Barber et al. (1996).

It is also worth mentioning that algorithms exist for directly constructing a suitable weighted Delaunay triangulation, for example Edelsbrunner and Shah (1996), which uses the notion of a flip (Definition A.22).

A.2.2 Description of Quickhull algorithm

This algorithm is incremental, in the sense that we begin with an initial simplex and a list of points not in the simplex. We then process the outside points one by one and update the convex hull until all the points have been included. We may assume without loss of generality that $0 \in \text{conv}(\mathscr{X})$ (if not, a change of origin will enforce this).

The convex hull is represented as a list of faces F_i and a list of the adjacent faces (that is, faces F_j such that $\dim F_i \cap F_j = d$).

We require two additional notions: the oriented hyperplane through d points (represented by a normal vector n to the hyperplane pointing away from the origin and an offset x), and the signed distance to a hyperplane from a point v given by $n^T(v-x)$. We say a point is above a hyperplane if its signed distance to the hyperplane is positive. Otherwise, we say the point is below the hyperplane.

Let H be a convex hull and let v be a point in $\mathbb{R}^d \backslash H$. Then F is a facet of $\text{conv}(v \cup H)$ if and only if

1 F is a facet of H and p is below F, or

2 F is not a facet of H, and its vertices are v and the vertices of $F_i \cap F_j$, where F_i and F_j are neighbouring facets of H, with v above F_i and below F_j.

This leads to the following pseudocode for computing the convex hull, taken from Barber et al. (1996).

Require: X, a set of points in $\mathbb{R}^d$
  Set $H = \text{conv}\left(\{x_1, \ldots, x_{d+1}\}\right)$
  Set $U = \{x_{d+2}, \ldots, x_n\}$
  Set $\mathscr{F}$ to be the set of faces of H
  for each facet $F \in \mathscr{F}$ **do**
    Set N_F to be the set of neighbours of F
    Set $\text{out}_F = \emptyset$
    for Each point $u \in U$ **do**
      if u is above F **then**
        set $\text{out}_F = \text{out}_F \cup \{u\}$
        set $U = U \backslash \{u\}$
  for each facet F **do**
    if $\text{out}_F \neq \emptyset$ **then**
      set p to be point in out_F furthest from F
      Set visible set $V = F$
      for $N \in N_G$ for faces G of V **do**
        if p is above N **then**
          set $V = V \cup \{N\}$
      Set $H =$ boundary of V
      for each ridge R in H **do**
        create a new facet from $R \cup p$ and add to H
        create links between this facet and its neighbours
      for each new facet F' **do**
        for each unassigned point q in an outside set of a facet in V **do**
          if q is above F' **then**
            add q to F''s outside set
      delete the facets in V

Algorithm A.1: Quickhull algorithm (Barber et al., 1996).

B Example Code

Here we give R code for the examples in Section 3.8.

```
> require("LogConcDEAD")
> require("logcondens")
> require("mvtnorm")

> n <- 200
> x <- sort(rgamma(n, shape = 2))
> out1 <- activeSetLogCon(x)
> out2 <- mlelcd(x)

> ylim <- c(0, 0.4)
> plot(out2, ylim = ylim, lty = 1)
> lines(x, exp(out1$phi), lty = 2)
> lines(x, x * exp(-x), col = "red")

> ylim <- c(-4, -1)
> plot(out2, uselog = TRUE, lty = 1)
> lines(x, out1$phi, lty = 2)
> lines(x, log(x) - x, col = "red")

> n <- 500
> d <- 2
> x <- matrix(rnorm(n * d), ncol = d)
> out <- mlelcd(x)
> g <- interplcd(out, gridlen = 200)
> gnorm <- g
> for (i in 1:200) {
+     for (j in 1:200) {
+         gnorm$z[i, j] <- dmvnorm(c(gnorm$x[i], gnorm$y[j]), log = TRUE)
+     }
+ }

> plot(out, g = g, addp = FALSE, asp = 1, main = "")
> plot(out, g = g, uselog = TRUE, addp = FALSE, asp = 1, main = "")
```

B Example code

```
> plot(out, g = g, type = "r")
> plot(out, g = gnorm, type = "r", addp = FALSE)
> plot(out, g = g, type = "r", uselog = TRUE)
> plot(out, g = gnorm, type = "r", uselog = TRUE, addp = FALSE)

> sigma <- matrix(c(1, 0.2, 0.2, 1), nrow = 2)
> y <- rmvnorm(n, sigma = sigma)
> xall <- 0.5 * round(2 * y)
> tmpw <- getweights(xall)
> outw <- mlelcd(tmpw$x, w = tmpw$w)
> gw <- interplcd(outw, gridlen = 200)

> par(mfrow = c(1, 2), pty = "s", cex = 0.7)
> plot(outw, g = gw, asp = 1, drawlabels = FALSE, pch = 4, main = "")
> plot(outw, g = gw, uselog = TRUE, asp = 1, drawlabels = FALSE,
+     pch = 4, main = "")

> d <- 3
> n <- 500
> x <- matrix(rgamma(n * d, shape = 2), ncol = d)
> out3 <- mlelcd(x)

> par(mfrow = c(2, 2), cex = 0.8)
> plot(out3, marg = 1, main = "", xlab = "")
> plot(out3, marg = 2, main = "", xlab = "")
> plot(out3, marg = 3, main = "", xlab = "")
> tmp <- seq(min(out3$x), max(out3$x), len = 100)
> plot(tmp, dgamma(tmp, shape = 2), type = "l", main = "", xlab = "",
+     ylab = "true marginal density")
```

Notation

Where the domain of integration is not specified, it may be assumed to be all of $\mathbb{R}$ or $\mathbb{R}^d$ as appropriate. We will write

$$\int g\,dF \text{ for } \int g(x)\,dF(x)$$

and

$$\int g \text{ for } \int g(x)\,dx.$$

List of symbols

$\|\cdot\|_p$ L_p norm

A_j, a_j Matrix and vector giving transformation of variables (3.8)

$\nabla\nabla^T f$ Hessian matrix of function f

b_j, β_n Vector and scalar defining log-concave maximum likelihood estimator (3.9)

C_n Convex hull of observed data

δ Stopping criterion (relative change in parameter)

$\delta(\cdot)$ Dirac δ function

$\delta_{C_n}(\cdot)$ $\delta_{C_n}(x) = \begin{cases} 0 & \text{if } x \in C_n \\ -\infty & \text{if } x \notin C_n \end{cases}$

∂_i ith partial derivative

ϵ Stopping criterion (relative change in objective function)

η Stopping criterion (integral)

e_i ith unit vector

$\mathscr{F}$ Class of log-concave densities (unless otherwise specified)

Notation

$\widehat{f}_n$ Log-concave maximum likelihood estimator (unless otherwise specified)

$\widehat{f}_n(\cdot, H)$ Kernel density estimate with bandwidth H

$\widehat{f}_n(\cdot, h)$ Kernel density estimate with bandwidth h

f_α Cutoff point for highest density region

F_n Empirical distribution function

G_d A special function used to evaluate σ, $\int_{T_d} \exp(y_0(1 - \sum_{k=1}^{d} w_k) + \sum_{k=1}^{d} y_k w_k)\, dw$

$\bar{h}_y(x)$ $\inf\{h(x) : h \text{ concave and } h(X_i) \geq y_i \text{ for } i = 1, \ldots, n\}$

$\mathscr{H}$ $\{\bar{h}_y : y \in \mathbb{R}^n\}$

$\mathscr{H}_1$ $\{H : H = h^2 I \text{ for some } h > 0\}$

$\mathscr{H}_2$ $\{H : H \text{ diagonal and } H \in \mathscr{H}_3\}$

$\mathscr{H}_3$ $\{H : |H| > 0\}$

H Bandwidth matrix

h Scalar bandwidth

$H(\epsilon, \mathscr{Y}, d)$ ϵ-entropy of $\mathscr{Y}$, with respect to distance d

$H_B(\epsilon, \mathscr{Y}, d)$ ϵ-bracketing entropy of $\mathscr{Y}$, with respect to distance d

$h_{y,\mathscr{T}}, h_{y,\mathscr{S}}$ Function obtained by interpolating over triangulation $\mathscr{T}$ or subdivision $\mathscr{S}$

$\mathbb{1}_A(\cdot)$ Indicator function of a set A

$\mathscr{J}$ Collection of indices defining a subdivision or triangulation

$\mathscr{J}_i$ $\{j \in \mathscr{J} : j_l = i \text{ for some } l\}$

K Kernel

K_H $K_H(x) = \dfrac{1}{|H|^{1/2}} K(H^{-1/2} x)$

K_h $K_h(x) = \dfrac{1}{h} K\left(\dfrac{x}{h}\right)$

ℓ_n Log-likelihood function

L_n Likelihood function

μ_d Lebesgue measure on $\mathbb{R}^d$

$m_2(f)$ Second moment of density f

$N(\epsilon, \mathscr{Y}, d)$ ϵ-covering number of $\mathscr{Y}$, with respect to distance d

$N_B(\epsilon, \mathscr{Y}, d)$ ϵ-bracketing number of $\mathscr{Y}$, with respect to distance d

$N_d(\mu, \Sigma)$ d-dimensional normal distribution with mean vector μ and covariance matrix Σ

ϕ Standard normal density function

ϕ_Σ Multivariate normal density function with mean 0 and covariance matrix Σ

ϕ_σ Univariate normal density function with mean 0 and variance σ^2

$\mathbb{P}_n$ Empirical measure

ψ_n Objective function for log-concave maximum likelihood (2.4)

ψ_r $\psi_r(f) = \int f^{(r)}(x) f(x)\, dx$

ρ_α Measure of discrepancy between function spaces

$R(f)$ $\int f^2$

$R(f, y)$ Region in which density f exceeds $y \in [0, \infty)$

$\mathscr{S}$ A subdivision (Definition A.19)

σ Function to be minimized to compute $\widehat{f}_n$

σ_w Function to be minimized to compute weighted maximum likelihood estimator

S Sample covariance matrix

S_n Shattering number

τ Objective function for maximum likelihood estimation (3.2)

$\mathscr{T}$ A triangulation (Definition A.20)

T_d d-dimensional unit simplex $\{x \in [0, \infty)^d : \sum_{i=1}^{d} x_i \leq 1\}$

$X_{(i)}$ ith order statistic

Binary relations

$\asymp$	$a_n \asymp b_n$ means that a_n and b_n are of the same order, that is, $a_n = O(b_n)$ and $b_n = O(a_n)$.
$f \star g$	Convolution product: $f \star g(y) = \int f(x)g(y-x)\,dx$

Acronyms and abbreviations

Abr	Abramson estimator
aff	Affine hull
AMISE	Asymptotic Mean Integrated Squared Error
$\xrightarrow{a.s.}$	Converges almost surely
cl	Closure of a convex function
conv	Convex hull
dom	Effective domain of a function
epi	Epigraph of a convex or concave functiong
iid	Independent and identically distributed
LSCV	Least-Squares Cross-Validation
MISE	Mean Integrated Squared Error
MSE	Mean Squared Error
$\xrightarrow{p}$	Converges in probability
PI	Plug-In
Sain	Sain estimator
SAMSE	Sum of Asymptotic Mean Squared Error
SCV	Smoothed Cross-Validation
vech	Vector half operator

References

I. Abramson. On variable bandwidth in kernel estimates – a square root law. *Annals of Statistics*, 10:1217–1223, 1982.

M. Y. An. Log-concave probability distributions: theory and statistical testing. Technical report, Duke University, 1995.

M. Y. An. Logconcavity versus logconvexity: A complete characterization. *Journal of Economic Theory*, 80:350–369, 1998.

A. Asuncion and D. J. Newman. UCI Machine Learning Repository, 2007. URL `http://www.ics.uci.edu/~mlearn/MLRepository.html`.

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

F. Balabdaoui. *Nonparametric estimation of a k-monotone density: a new asymptotic distribution theory*. PhD thesis, University of Washington, 2004.

F. Balabdaoui and J. A. Wellner. Estimation of a k-monotone density: Limiting distribution theory and the spline connection. *The Annals of Statistics*, 35:2536–2564, 2007.

F. Balabdaoui, K. Rufibach, and J. A. Wellner. Limit distribution theory for maximum likelihood estimation of a log-concave density. *Annals of Statistics*, 37(3):1299–1331, 2009.

C. B. Barber, D. P. Dobkin, and H. Huhdanpaa. The **Quickhull** algorithm for convex hulls. *ACM Transactions on Mathematical Software*, 22:469–483, 1996. URL `http://www.qhull.org/`.

R. E. Barlow and F. Proschan. *Statistical theory of reliability and life testing*. Holt, Reinhart and Winston, New York, 1975.

O. Barndorff-Nielsen. *Information and exponential families in statistical theory*. Wiley, New Jersey, 1978.

J. Beirlant, E. J. Dudewicz, L. Györfi, and E. C. van der Meulen. Nonparametric entropy estimation: An overview. *International Journal of Mathematical and Statistical Sciences*, 6(17-40), 1997.

References

M. Bern. Triangulations and mesh generation. In J. E. Goodman and J. O'Rourke, editors, *Handbook of discrete and computational geometry*, pages 563–582. CRC Press, New York, 2004.

R. N. Bhattacharya and R. Ranga Rao. *Normal approximations and asymptotic expansions*. Wiley, New York, 1976.

L. Birgé. Estimation of unimodal densities without smoothness assumptions. *Annals of Statistics*, 25(3):970–981, 1997.

S. Boyd and L. Vandenberghe. *Convex optimization*. Cambridge University Press, Cambridge, 2004.

L. Breiman, W. Meisel, and E. Purcell. Variable kernel estimates of multivariate densitiies. *Technometrics*, 19:135–144, 1977.

S. P. Brooks. MCMC convergence diagnosis via multivariate bounds on log-concave densities. *Annals of Statistics*, 26:398–433, 1998.

J. E. Chacón. Data-driven choice of the smoothing parametrization for kernel density estimators. *Canadian Journal of Statistics*, 34(4):249 – 265, 2009.

J. E. Chacón, M. P. Wand, and T. Duong. Asymptotics for general multivariate kernel density derivative estimators. In preparation, 2008.

G. Chang and G. Walther. Clustering with mixtures of log-concave distributions. *Computational Statistics and Data Analysis*, 51(12):6242–6251, 2007.

D. R. Cox and D. V. Hinkley. *Theoretical statistics*. Chapman and Hall, 1 edition, 1979.

M. L. Cule and L. Dümbgen. On an auxiliary function for log-density estimation. Technical Report 71, Universität Bern, 2008.

M. L. Cule, R. J. Samworth, and M. I. Stewart. Maximum likelihood estimation of a multidimensional log-concave density. Submitted, 2008.

J. Ćwik and J. Koronacki. Multivariate density estimation: A comparative study. *Neural Computation and Applicaitons*, 6(173-185), 1997.

A. Dempster, N. Laird, and D. Rubin. Maximum likelihood from incomplete data via the EM algorithm. *Journal of the Royal Statistical Society Series B*, 39(1):1–38, 1977.

L. Devroye and L. Györfi. *Nonparametric density estimation: The L_1 view*. Wiley, New Jersey, 1985.

S. Dharmadhikari and K. Joag-Dev. *Unimodality, convexity and applications*. Academic Press, Boston, 1988.

D. L. Donoho, I. M. Johnstone, G. Kerkyacharian, and D. Picard. Density estimation by wavelet thresholding. *Annals of Statistics*, 24(2):508–539, 1996.

L. Dümbgen and K. Rufibach. Maximum likelihood estimation of a log-concave density and its distribution function: Basic properties and uniform consistency. *Bernoulli*, 15 (1):40–68, 2008.

L. Dümbgen, A. Hüsler, and K. Rufibach. Active set and EM algorithms for log-concave densities based on complete and censored data. Technical report, Universität Bern, 2007. URL `http://arxiv.org/abs/0709.0334/`.

T. Duong. *Bandwidth selectors for multivariate kernel density estimation*. PhD thesis, University of Western Australia, 2004.

T. Duong. ***ks**: Kernel smoothing*, 2007a. URL `http://cran.r-project.org/package=ks`. R package version 1.56.

T. Duong. **ks**: Kernel density estimation and kernel discriminant analysis for multivariate data in R. *Journal of Statistical Software*, 21(7), 2007b.

T. Duong and M. L. Hazelton. Plug-in bandwidth matrices for bivariate kernel density estimation. *Nonparametric Statistics*, 15(1):1029–0311, 2003.

T. Duong and M. L. Hazelton. Cross-validation bandwidth matrices for multivariate kernel density estimation. *Scandinavian Journal of Statistics*, 32(3):485–506, 2005.

H. Edelsbrunner and N. R. Shah. Incremental topological flipping works for regular triangulations. *Algorithmica*, 15:223–241, 1996.

B. Efron and R. Tibshirani. *An introduction to the bootstrap*. Monographs on statistics and applied probability. Chapman and Hall, New York, 1993.

P. P. B. Eggermont and V. LaRiccia. *Maximum penalized likelihood estimation*, volume 1: Density estimation of *Springer series in statistics*. Springer-Verlag, New York, 2001.

R. L. Eubank. *Spline smoothing and nonparametric regression*. Marcel Dekker, 1988.

R. Evans. Rates of convergence of maximum likelihood estimates via entropy methods. Part III Essay, 2007.

W. Feller. *An introduction to probability theory and its applications*, volume 2. Wiley, New York, 2nd edition, 1971.

References

R. A. Fisher. On the mathematical foundations of theoretical statistics. *Philisophical Transations of the Royal Society of London Series A*, 222(594-604):309–368, 1922.

E. Fix and J. L. Hodges. Discriminatory analysis – nonparametric discrimination: Consistency properties. Technical Report 4, Project no. 21-29-004, USAF School of Aviation Medicine, Randolph Field, Texas, 1951.

E. Fix and J. L. Hodges. Discriminatory analysis – nonparametric discrimination: Consistency properties. *International Statistical Review*, 57:238–247, 1989.

C. Fraley and A. Raftery. ***mclust****: Model-based clustering / Normal mixture modeling*, 2008. URL `http://CRAN.R-project.org/package=mclust`. R package version 3.1-3.

I. M. Gelfand, M. M. Kapranov, and A. V. Zelevinsky. Discriminants of polynomials of several variables and triangulations of Newton polyhedra. *Leningrad mathematics journal*, 2:449–505, 1990.

I. M. Gelfand, M. M. Kapranov, and A. V. Zelevinsky. *Discriminants, Resultants and Multidimensional Determinants*. Birkhäuser, Boston, 1994.

A. Genz. An adaptive numerical integration algorithm for simplices. In N. A. Sherwani, E. de Doncker, and J. A. Kapenga, editors, *Computing in the 90s, proceedings of the first Great Lakes computer science conference*, volume 1 of *Lecture Notes in Computer Science*. Springer-Verlag, New York, 1991.

A. Genz and R. Cools. An adaptive numerical cubature algorithm for simplices. *ACM Transactions on Mathematical Software*, 29(3):297–308, 2003.

R. Grasman and R. B. Gramacy. ***geometry****: Mesh Generation and Surface Tesselation*, 2008. URL `http://CRAN.R-project.org/package=geometry`. R package version 0.1.

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

P. Groeneboom. Estimating a monotone density. In *Proceedings of the Berkeley conference in honor of Jerzy Neyman and Jack Keifer*, volume 2, pages 539–555, London, 1983. Chapman and Hall.

P. Groeneboom, G. Hooghiemstra, and H. P. Lopuhaä. Asymptotic normality of the L_1 error of the Grenander estimator. *Annals of Statistics*, 27(4):1316–1347, 2001a.

P. Groeneboom, G. Jongbloed, and J. A. Wellner. Estimation of a convex function: Characterizations and asymptotic theory. *The Annals of Statistics*, 29:1653–1698, 2001b.

A. Grundmann and H. Möller. Invariant integration formulas for the n-simplex by combinatorial methods. *SIAM Journal of Numerical Analysis*, 15:282–290, 1978.

F. R. Hampel. Design, modelling and analysis of some biological datasets. In C. L. Mallows, editor, *Design, data and analysis: By some friends of Cuthbert Daniel*. Wiley, New Jersey, 1987.

P. J. Huber. The behaviour of maximum likelihood estimates under nonstandard conditions. In *Proceedings of the fifth Berkeley symposium on mathematical statistics and probability*, pages 221–233, Berkeley, CA, 1967. Univerity of California Press.

P. J. Huber and E. M. Ronchetti. *Robust statistics*. Wiley series in probability and mathematical statistics. Wiley, New Jersey, 2nd edition, 2009.

R. J. Hyndman. Computing and graphing highest density regions. *The American Statistician*, 50:120–126, 1996.

A. I. Ibragimov. On the composition of unimodal distributions. *Theory of Probability and its Applications*, 1(2):255–260, 1956.

F. Kappel and A. Kuntsevich. An implementation of Shor's r-algorithm. *Computational Optimization and Applications*, 15:193–205, 2000.

S. Karlin. *Total positivity*, volume 1. Stanford University Press, Palo Alto, CA, 1968.

S. Konishi and G. Kitagawa. *Information criteria and statistical modeling*. Springer series in statistics. Springer, New York, 2008.

V. N. Kulikov and H. P. Lopuhaä. Asymptotic normality of the L_k error of the Grenander estimator. *Annals of Statistics*, 33(5):2228–2255, December 2005.

C. L. Lawson. Properties of n-dimensional triangulations. *Computer Aided Geometric Design*, 3:231–246, 1986.

C. W. Lee. Subdivisions and triangulations of polytopes. In J. E. Goodman and J. O'Rourke, editors, *Handbook of discrete and computational geometry*, pages 383–406. CRC Press, New York, 2004.

F. Leisch. Dynamic generation of statistical reports using literate data analysis. In W. Härdle and B. Rönz, editors, *COMPSTAT 2002 – Proceedings in Computational Statistics*, pages 575–580, Heidelberg, 2002. Physica-Verlag.

G. McLachlan and D. Peel. *Finite mixture models*. Wiley series in probability and mathematical statistics. Wiley, New York, 2000.

S. Müller and K. Rufibach. Smooth tail index estimation. *Journal of Statistical Computation and Simulation*, 79(9):1155 – 1167, 2009.

S. Müller and K. Rufibach. On the max-domain of attraction of distributions with log-concave densities. *Statistics and Probability Letters*, 78(12):1440–1444, 2007.

L. Pace and A. Salvan. *Principles of statistical inference from a neo-Fisherian perspective*. Advanced series on statistical science and applied probability. World Scientific Publishing, Singapore, 1997.

J. K. Pal, M. Woodroofe, and M. Meyer. Estimating a Polya frequency function. In *Complex Datasets and Inverse Problems: Tomography, Networks and Beyond*. Institute of Mathematical Statistics, Ohio, 2007.

E. Parzen. On the estimation of a probability density function and the mode. *Annals of Mathematical Statistics*, 33:1065–76, 1962.

D. Pollard. *Convergence of stochastic processes*. Springer series in statistics. Springer-Verlag, New York, 1984.

L. Pournin and T. M. Liebling. Constrained paths in the flip-graph of regular triangulations. *Computational Geometry: Theory and Applications*, 37:134–140, 2007.

A. Prékopa. On logarithmically concave measures and functions. *Acta Scientarium Mathematicarum*, 34:335–343, 1973.

W. H. Press, S. A. Teukolsky, W. T. Vetterling, and B. P. Flannery. *Numerical recipes: The art of scientific computing*. Cambridge University Press, Cambridge, 2007.

B. L. S. Prakasa Rao. Estimation of a unimodal density. *Sankhya Series A*, 31:23–36, 1969.

R Development Core Team. *R: A Language and Environment for Statistical Computing*. R Foundation for Statistical Computing, Vienna, Austria, 2008. URL `http://www.R-project.org/`. ISBN 3-900051-07-0.

J. A. Rice. *Mathematical statistics and data analysis*. Dunbury Press, Belmont, CA, 1995.

R. T. Rockafellar. *Convex analysis*. Princeton University Press, Princeton, New Jersey, 1997.

M. Rosenblatt. Remarks on some nonparametric estimates of a density function. *Annals of Mathematical Statistics*, 27:832–837, 1956.

K. Rufibach. *Log-concave density estimation and bump hunting for iid observations*. PhD thesis, University of Bern, 2006.

K. Rufibach. Computing maximum likelihood estimators of a log-concave density function. *Journal of Statistical Computation and Simulation*, 77:561–574, 2007.

K. Rufibach and L. Dümbgen. ***logcondens****: Estimate a Log-Concave Probability Density from i.i.d. Observations*, 2006. URL `http://CRAN.R-project.org/package=logcondens`. R package version 1.3.2.

S. R. Sain. Multivariate locally adaptive density estimation. *Computational Statistics and Data Analysis*, 39(2):165–186, 2002.

S. R. Sain and D. W. Scott. On locally adaptive density estimation. *Journal of the American Statistical Association*, 91(436):1525–1534, 1996.

D. W. Scott. *Multivariate density estimation*. Wiley, New York, 1992.

D. W. Scott and S. R. Sain. Multi-dimensional density estimation. In C. R. Rao and E. J. Wegman, editors, *Handbook of statistics*, volume 23: Data mining and computational statistics. Elsevier, Amsterdam, 2004.

R. Seidel. Convex hull computations. In J. E. Goodman and J. O'Rourke, editors, *Handbook of discrete and computational geometry*, pages 495–512. CRC Press, New York, 2004.

C. E. Shannon. A mathematical theory of communication. *Bell System Technical Journal*, 27:379–423, 1948.

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

B. W. Silverman. *Density estimation for statistics and data analysis*. Chapman and Hall, London, 1986.

J. Snoeyink. Point location. In J. E. Goodman and J. O'Rourke, editors, *Handbook of discrete and computational geometry*, pages 767–785. CRC Press, New York, 2nd edition, 2004.

W. M. Street, W. H. Wolberg, and O. L. Mangasarian. Nuclear feature extraction for breast tumor diagnosis. *IS & T/SPIE International Symposium on Electronic Imaging: Science and Technology*, 1905:861–870, 1993.

References

A. H. Stroud. *Approximate calculation of multiple integrals*. Prentice-Hall Series in Automatic Computation. Prentice-Hall, New Jersey, 1971.

The Times. Good university guide, 2008. `http://extras.timesonline.co.uk/gug/gooduniversityguide.php`.

J. R. Thompson and R. A. Tapia. *Nonparametric function estimation, modeling and simulation*. Society for Industrial and Applied Mathematics, Philadelphia, 1990.

S. van de Geer. *Applications of empirical process theory*. Cambridge series in statistical and probabilistic mathematics. Cambridge University Press, Cambridge, 2000.

A. W. van der Vaart and J. A. Wellner. *Weak convergence and empirical processes*. Springer-Verlag, New York, 1996.

V. N. Vapnik and S. Mukherjee. Support vector method for multivariate density estimation. In *Advances in Neural Information Processing Systems*, pages 659–665, Cambridge, MA, 2000. MIT press.

G. Wahba. *Spline models for observational data*. Society for Industrial and Applied Mathematics, Philadelphia, 1990.

G. Walther. Detecting the presence of mixing with multiscale maximum likelihood. *Journal of the American Statistical Association*, 97(458):508–513, 2002.

G. Walther. Inference and modeling with log-concave distributions. Unpublished manuscript, 2008.

M. P. Wand and M. C. Jones. *Kernel smoothing*. Chapman and Hall, CRC Press, Florida, 1995.

X. Wang, M. Woodroofe, M. Walker, M. Mateo, and E. Olzewski. Estimating dark matter distributions. *Astrophysics Journal*, 626:245–158, 2005.

G. S. Watson. Estimating functionals of particle size distributions. *Biometrika*, 58(3): 483–490, 1971. doi: 10.1093/biomet/58.3.483.

H. White. Maximum likelihood estimation of misspecified models. *Econometrica*, 50: 1–25, 1982.

W. H. Wong and X. Shen. Probability inequalities for likelihood ratios and convergence rates of sieve MLEs. *Annals of Statistics*, 23(2):339–362, 1995.

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

X. Zhang, M. L. King, and R. J. Hyndman. Bandwidth selection for multivariate kernel density estimation using MCMC. *Computational Statistics and Data Analysis*, 50: 3009–3031, 2006.