Maximum likelihood estimation of a multidimensional log-concave density

Madeleine Cule and Richard Samworth†

University of Cambridge, UK

and Michael Stewart

University of Sydney, Australia

Summary. Let $X_1, \ldots, X_n$ be independent and identically distributed random vectors with a (Lebesgue) density f. We first prove that, with probability one, there exists a unique log-concave maximum likelihood estimator $\hat{f}_n$ of f. The use of this estimator is attractive because, unlike kernel density estimation, the method is fully automatic, with no smoothing parameters to choose. Although the existence proof is non-constructive, we are able to reformulate the issue of computing $\hat{f}_n$ in terms of a non-differentiable convex optimisation problem, and thus combine techniques of computational geometry with Shor's r-algorithm to produce a sequence that converges to $\hat{f}_n$. An R version of the algorithm is available in the package **LogConcDEAD** – Log-Concave Density Estimation in Arbitrary Dimensions. We demonstrate that the estimator has attractive theoretical properties both when the true density is log-concave and when this model is misspecified. For the moderate or large sample sizes in our simulations, $\hat{f}_n$ is shown to have smaller mean integrated squared error compared with kernel-based methods, even when we allow the use of a theoretical, optimal fixed bandwidth for the kernel estimator that would not be available in practice. We also present a real data clustering example, which shows that our methodology can be used in conjunction with the Expectation–Maximisation (EM) algorithm to fit finite mixtures of log-concave densities.

Keywords: Computational geometry, log-concavity, maximum likelihood estimation, non-differentiable convex optimisation, nonparametric density estimation, Shor's r-algorithm

1. Introduction

Modern nonparametric density estimation began with the introduction of a kernel density estimator in the pioneering work of Fix and Hodges (1951), later republished as Fix and Hodges (1989). For independent and identically distributed real-valued observations, the appealing asymptotic theory of the mean integrated squared error was provided by Rosenblatt (1956) and Parzen (1962). This theory leads to an asymptotically optimal choice of the smoothing parameter, or bandwidth. Unfortunately, however, it depends on the unknown density f through the integral of the square of the second derivative of f. Considerable effort has therefore been focused on finding methods of automatic bandwidth selection (cf. Wand and Jones, 1995, Chapter 3, and the references therein). Although

†*Address for correspondence*: Richard Samworth, Statistical Laboratory, Centre for Mathematical Sciences, Wilberforce Road, Cambridge, UK. CB3 0WB.
E-mail: r.j.samworth@statslab.cam.ac.uk.

this has resulted in algorithms, e.g. Chiu (1992), that achieve the optimal rate of convergence of the relative error, namely $O_p(n^{-1/2})$, where n is the sample size, good finite sample performance is by no means guaranteed.

This problem is compounded when the observations take values in $\mathbb{R}^d$, where the general kernel estimator (Deheuvels, 1977) requires the specification of a symmetric, positive definite $d \times d$ bandwidth matrix. The difficulties involved in making the $d(d+1)/2$ choices for its entries mean that attention is often restricted either to bandwidth matrices that are diagonal, or even to those that are scalar multiples of the identity matrix. Despite recent progress (for example, Duong and Hazelton (2003), Duong and Hazelton (2005), Zhang, King and Hyndman (2006), Chacón, Duong and Wand (2008), Chacón (2009)), significant practical challenges remain.

Extensions that adapt to local smoothness began with Breiman, Meisel and Purcell (1977) and Abramson (1982). A review of several adaptive kernel methods for univariate data may be found in Sain and Scott (1996). Multivariate adaptive techniques are presented in Sain (2002), Scott and Sain (2004) and Duong (2004). There are many other smoothing methods for density estimation, for example methods based on wavelets (Donoho *et al.*, 1996), splines (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, the optimal value of which depends on the unknown density, so achieving an appropriate level of smoothing is difficult.

In this paper, we propose a fully automatic nonparametric estimator of f, with no tuning parameters to be chosen, under the condition that f is log-concave – that is, $\log f$ is a concave function. The class of log-concave densities has many attractive properties and has been well-studied, particularly in the economics, sampling and reliability theory literature. See Section 2 for further discussion of examples, applications and properties of log-concave densities.

In Section 3, we show that if $X_1, \ldots, X_n$ are independent and identically distributed random vectors, then with probability one there exists a unique log-concave density $\hat{f}_n$ that maximises the likelihood function,

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

Before continuing, it is worth noting that without any shape constraints on the densities under consideration, the likelihood function is unbounded. To see this, we could define a sequence (f_n) of densities that represent successively close approximations to a mixture of n 'spikes' (one on each X_i), such as $f_n(x) = n^{-1} \sum_{i=1}^{n} \phi_{d,n^{-1}I}(x - X_i)$, where $\phi_{d,\Sigma}$ denotes the $N_d(0, \Sigma)$ density. This sequence satisfies $L(f_n) \to \infty$ as $n \to \infty$. In fact, a modification of this argument may be used to show that the likelihood function remains unbounded even if we restrict attention to unimodal densities.

There has been considerable recent interest in shape-restricted nonparametric density estimation, but most of it has been confined to the case of univariate densities, where the computational algorithms are more straightforward. Nevertheless, as was discussed above, it is in multivariate situations that the automatic nature of the maximum likelihood estimator is particularly valuable. Walther (2002), Dümbgen and Rufibach (2009) and Pal, Woodroofe and Meyer (2007) have proved the existence and uniqueness of the log-concave maximum likelihood estimator in one dimension and Dümbgen and Rufibach (2009), Pal, Woodroofe and Meyer (2007) and Balabdaoui, Rufibach and Wellner

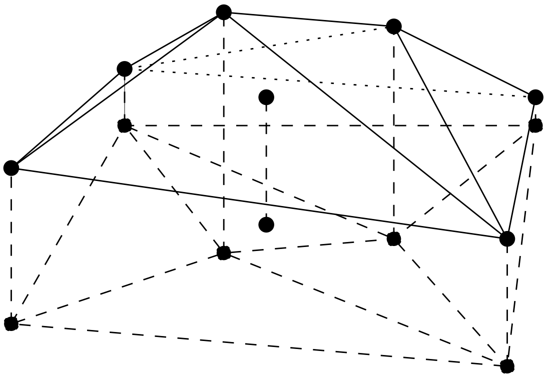


Fig. 1. The ‘tent-like’ structure of the graph of the logarithm of the maximum likelihood estimator for bivariate data.

(2009) have studied its theoretical properties. Rufibach (2007) compared different algorithms for computing the univariate estimator, including the iterative convex minorant algorithm (Groeneboom and Wellner, 1992; Jongbloed, 1998), and three others. Dümbgen, Hüsler and Rufibach (2007) also present an Active Set algorithm, which has similarities with the vertex direction and vertex reduction algorithms described in Groeneboom, Jongbloed and Wellner (2008). Walther (2010) provides a nice recent review article on inference and modelling with log-concave densities. Other recent related work includes Seregin and Wellner (2009), Schuhmacher, Hüsler and Dümbgen (2010), Schuhmacher and Dümbgen (2010) and Koenker and Mizera (2010). For univariate data, it is also well-known that there exist maximum likelihood estimators of a non-increasing density supported on $[0, \infty)$ (Grenander, 1956) and of a convex, decreasing density (Groeneboom, Jongbloed and Wellner, 2001).

Figure 1 gives a diagram illustrating the structure of the maximum likelihood estimator on the logarithmic scale. This structure is most easily visualised for two-dimensional data, where one can imagine associating a ‘tent pole’ with each observation, extending vertically out of the plane. For certain tent pole heights, the graph of the logarithm of the maximum likelihood estimator can be thought of as the roof of a taut tent stretched over the tent poles. The fact that the logarithm of the maximum likelihood estimator is of this ‘tent function’ form constitutes part of the proof of its existence and uniqueness.

In Sections 3.1 and 3.2, we discuss the computational problem of how to adjust the n tent pole heights so that the corresponding tent functions converge to the logarithm of the maximum likelihood estimator. One reason that this computational problem is so challenging in more than one dimension is the fact that it is difficult to describe the set of tent pole heights that correspond to concave functions. The key observation, discussed in Section 3.1, is that it is possible to minimise a modified objective function that is convex (though non-differentiable). This allows us to apply the powerful non-differentiable convex optimisation methodology of the subgradient method (Shor, 1985) and a variant called Shor's r-algorithm, which has been implemented by Kappel and Kuntsevich (2000).

As an illustration of the estimates obtained, Figure 2 presents plots of the maximum likelihood estimator, and its logarithm, for 1000 observations from a standard bivariate normal distribution.

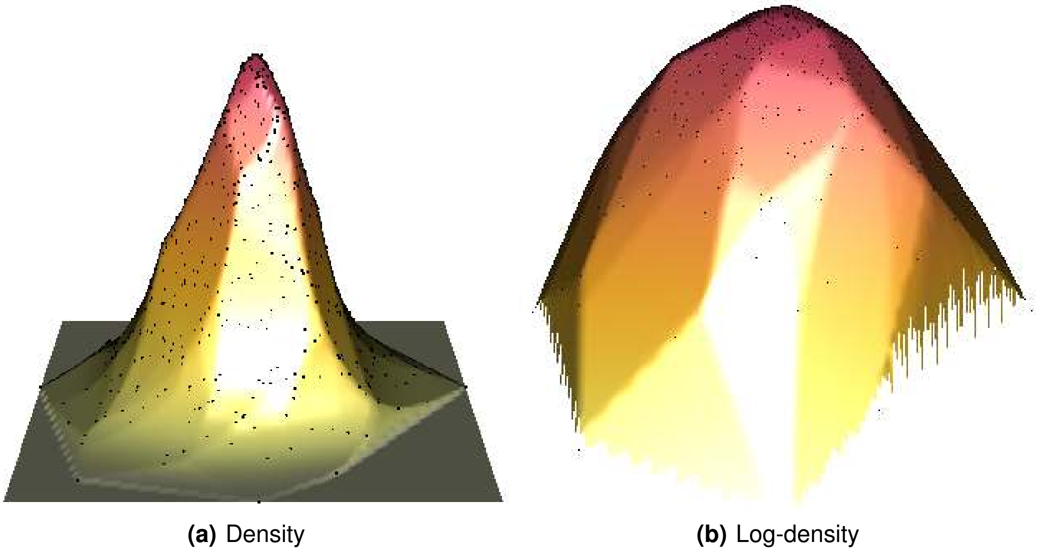

(a) Density **(b)** Log-density

Fig. 2. Log-concave maximum likelihood estimates based on 1000 observations (plotted as dots) from a standard bivariate normal distribution.

These plots were created using the **LogConcDEAD** package (Cule, Gramacy and Samworth, 2007) in R (R Development Core Team, 2009).

Theoretical properties of the estimator $\hat{f}_n$ are presented in Section 4. We describe the asymptotic behaviour of the estimator both in the case where the true density is log-concave, and where this model is misspecified. In the former case, we show that $\hat{f}_n$ converges in certain strong norms to the true density. The nature of the norm chosen gives reassurance about the behaviour of the estimator in the tails of the density. In the misspecified case, $\hat{f}_n$ converges to the log-concave density that is closest to the true underlying density (in the sense of minimising the Kullback–Leibler divergence). This latter result amounts to a desirable robustness property.

In Section 5 we present simulations to compare the finite-sample performance of the maximum likelihood estimator with kernel-based methods with respect to the mean integrated squared error (MISE) criterion. The results are striking: even when we use the theoretical, optimal bandwidth for the kernel estimator (or an asymptotic approximation to this when it is not available), we find that the maximum likelihood estimator has a rather smaller mean integrated squared error for moderate or large sample sizes, despite the fact that this optimal bandwidth depends on properties of the density that would be unknown in practice.

Nonparametric density estimation is a fundamental tool for the visualisation of structure in exploratory data analysis. Our proposed method may certainly be used for this purpose; however, it may also be used as an intermediary stage in more involved statistical procedures. For instance:

(a) In classification problems, we have $p \geq 2$ populations of interest, and assume in this discussion that these have densities $f_1, \ldots, f_p$ on $\mathbb{R}^d$. We observe training data of the form $\{(X_i, Y_i) :$

$i = 1, \ldots, n\}$, where if $Y_i = j$, then X_i has density f_j. The aim is to classify a new observation $z \in \mathbb{R}^d$ as coming from one of the populations. Problems of this type occur in a huge variety of applications, including medical diagnosis, archaeology, ecology etc. – see Gordon (1981), Hand (1981) or Devroye, Györfi and Lugosi (1996) for further details and examples. A natural approach to classification problems is to construct density estimates $\hat{f}_1, \ldots, \hat{f}_p$, where $\hat{f}_j$ is based on the n_j observations, say, from the jth population, namely $\{X_i : Y_i = j\}$. We may then assign z to the jth population if $n_j \hat{f}_j(z) = \max\{n_1 \hat{f}_1(z), \ldots, n_p \hat{f}_p(z)\}$. In this context, the use of kernel-based estimators in general requires the choice of p separate $d \times d$ bandwidth matrices, while the corresponding procedure based on the log-concave maximum likelihood estimates is again fully automatic.

(b) Clustering problems are closely related to the classification problems described above. The difference is that, in the above notation, we do not observe $Y_1, \ldots, Y_n$, and have to assign each of $X_1, \ldots, X_n$ to one of the p populations. A common technique is based on fitting a mixture density of the form $f(x) = \sum_{j=1}^{p} \pi_j f_j(x)$, where the mixture proportions $\pi_1, \ldots, \pi_p$ are positive and sum to one. We show in Section 6 that our methodology can be extended to fit a finite mixture of log-concave densities, which need not itself be log-concave – cf. Section 2. A simple plug-in Bayes rule may then be used to classify the points. We also illustrate this clustering algorithm on a Wisconsin breast cancer data set in Section 6, where the aim is to separate observations into benign and malignant component populations.

(c) A functional of the true underlying density may be estimated by the corresponding functional of a density estimator, such as the log-concave maximum likelihood estimator. Examples of functionals of interest include probabilities, such as $\int_{\|x\| \geq 1} f(x)\, dx$, moments, e.g. $\int \|x\|^2 f(x)\, dx$, and the differential entropy, $-\int f(x) \log f(x)\, dx$. It may be possible to compute the plug-in estimator based on the log-concave maximum likelihood estimator analytically, but in Section 7, we show that even if this is not possible, we can sample from the log-concave maximum likelihood estimator $\hat{f}_n$, and hence in many cases of interest obtain a Monte Carlo estimate of the functional. This nice feature also means that the log-concave maximum likelihood estimator can be used in a Monte Carlo bootstrap procedure for assessing uncertainty in functional estimates.

(d) The fitting of a nonparametric density estimate may give an indication of the validity of a particular smaller model (often parametric). Thus, a contour plot of the log-concave maximum likelihood estimator may provide evidence that the underlying density has elliptical contours, and thus suggest a model that exploits this elliptical symmetry.

(e) In the univariate case, Walther (2002) describes methodology based on log-concave density estimation for addressing the problem of detecting the presence of mixing in a distribution. As an application, he cites the Pickering/Platt debate (Swales, 1985) on the issue of whether high blood pressure is a disease (in which case observed blood pressure measurements should follow a mixture distribution), or simply a label attached to people in the right tail of the blood pressure distribution. As a result of our algorithm for computing the multidimensional log-concave maximum likelihood estimator, a similar test may devised for multivariate data – see Section 8.

In Section 9, we give a brief concluding discussion, and suggest some directions for future research. We defer the proofs to Appendix A and discuss structural and computational issues in Appendix B. Finally, we present in Appendix C a glossary of terms and results from convex analysis and computational geometry that appear in italics at their first occurrence in the main body of the paper.

2. Log-concave densities: examples, applications and properties

Many of the most commonly-encountered parametric families of univariate distributions have *log-concave densities*, including the family of normal distributions, gamma distributions with shape parameter at least one, Beta(α, β) distributions with $\alpha, \beta \geq 1$, Weibull distributions with shape parameter at least one, Gumbel, logistic and Laplace densities; see Bagnoli and Bergstrom (2005) for other examples. Univariate log-concave densities are unimodal and have fairly light tails – it may help to think of the exponential distribution (where the logarithm of the density is a linear function on the positive half-axis) as a borderline case. Thus Cauchy, Pareto and lognormal densities, for instance, are not log-concave. Mixtures of log-concave densities may be log-concave, but in general they are not; for instance, for $p \in (0,1)$, the location mixture of standard univariate normal densities $f(x) = p\phi(x) + (1-p)\phi(x-\mu)$ is log-concave if and only if $\|\mu\| \leq 2$.

The assumption of log-concavity is a popular one in economics; Caplin and Naelbuff (1991b) show that in the theory of elections and under a log-concavity assumption, the proposal most preferred by the mean voter is unbeatable under a 64% majority rule. As another example, in the theory of imperfect competition, Caplin and Naelbuff (1991a) use log-concavity of the density of consumers' utility parameters as a sufficient condition in their proof of the existence of a pure-strategy price equilibrium for any number of firms producing any set of products. See Bagnoli and Bergstrom (2005) for many other applications of log-concavity to economics. Brooks (1998) and Mengersen and Tweedie (1996) have exploited the properties of log-concave densities in studying the convergence of Markov chain Monte Carlo sampling procedures.

An (1998) lists many useful properties of log-concave densities. For instance, if f and g are (possibly multidimensional) log-concave densities, then their convolution $f * g$ is log-concave. In other words, if X and Y are independent and have log-concave densities, then their sum $X + Y$ has a log-concave density. The class of log-concave densities is also closed under the taking of pointwise limits. One-dimensional log-concave densities have increasing hazard functions, which is why they are of interest in reliability theory. Moreover, Ibragimov (1956) proved the following characterisation: a univariate density f is log-concave if and only if the convolution $f * g$ is unimodal for every unimodal density g. There is no natural generalisation of this result to higher dimensions.

As was mentioned in Section 1, this paper concerns multidimensional log-concave densities, for which fewer properties are known. It is therefore of interest to understand how the property of log-concavity in more than one dimension relates to the univariate notion. Our first proposition below is intended to give some insight into this issue. It is not formally required for the subsequent development of our methodology in Section 3, although we did apply the result when designing our simulation study in Section 5.

PROPOSITION 1. *Let X be a d-variate random vector having density f with respect to Lebesgue measure on $\mathbb{R}^d$. For a subspace V of $\mathbb{R}^d$, let $P_V(x)$ denote the orthogonal projection of x onto V. Then in order that f be log-concave, it is:*

(a) necessary that for any subspace V, the marginal density of $P_V(X)$ is log-concave and the conditional density $f_{X|P_V(X)}(\cdot|t)$ of X given $P_V(X) = t$ is log-concave for each t
(b) sufficient that for every $(d-1)$-dimensional subspace V, the conditional density $f_{X|P_V(X)}(\cdot|t)$

of X given $P_V(X) = t$ is log-concave for each t.

The part of Proposition 1(a) concerning marginal densities is an immediate consequence of Theorem 6 of Prékopa (1973). One can regard Proposition 1(b) as saying that a multidimensional density is log-concave if the restriction of the density to any line is a (univariate) log-concave function.

It is interesting to compare the properties of log-concave densities presented in Proposition 1 with the corresponding properties of Gaussian densities. In fact, Proposition 1 remains true if we replace 'log-concave' with 'Gaussian' throughout (at least, provided that in part (b) we also assume there is a point at which f is twice differentiable). These shared properties suggest that the class of log-concave densities is a natural, infinite-dimensional generalisation of the class of Gaussian densities.

3. Existence, uniqueness and computation of the maximum likelihood estimator

Let $\mathcal{F}_0$ denote the class of log-concave densities on $\mathbb{R}^d$. The degenerate case where the support is of dimension smaller than d can also be handled, but for simplicity of exposition we concentrate on the non-degenerate case. Let f_0 be a density on $\mathbb{R}^d$, and suppose that $X_1, \ldots, X_n$ are a random sample from f_0, with $n \geq d+1$. We say that $\hat{f}_n = \hat{f}_n(X_1, \ldots, X_n) \in \mathcal{F}_0$ is a log-concave *maximum likelihood estimator* of f_0 if it maximises $\ell(f) = \sum_{i=1}^n \log f(X_i)$ over $f \in \mathcal{F}_0$.

Theorem 2. *With probability one, a log-concave maximum likelihood estimator $\hat{f}_n$ of f_0 exists and is unique.*

During the course of the proof of Theorem 2, it is shown that $\hat{f}_n$ is supported on the convex hull of the data, which we denote by $C_n = \mathrm{conv}(X_1, \ldots, X_n)$. Moreover, as was mentioned in Section 1, $\log \hat{f}_n$ is a 'tent function'. For a fixed vector $y = (y_1, \ldots, y_n) \in \mathbb{R}^n$, a tent function is a function $\bar{h}_y : \mathbb{R}^d \to \mathbb{R}$ with the property that $\bar{h}_y$ is the least concave function satisfying $\bar{h}_y(X_i) \geq y_i$ for all $i = 1, \ldots, n$. A typical example of a tent function is depicted in Figure 1.

Although it is useful to know that $\log \hat{f}_n$ belongs to this finite-dimensional class of tent functions, the proof of Theorem 2 gives no indication of how to find the member of this class (in other words, the $y \in \mathbb{R}^n$) that maximises the likelihood function. We therefore seek an iterative algorithm to compute the estimator.

3.1. Reformulation of the optimisation problem

As a first attempt to find an algorithm which produces a sequence that converges to the maximum likelihood estimator in Theorem 2, it is natural to try to minimise numerically the function

$$\tau(y_1, \ldots, y_n) = -\frac{1}{n} \sum_{i=1}^n \bar{h}_y(X_i) + \int_{C_n} \exp\{\bar{h}_y(x)\}\, dx. \tag{3.1}$$

The first term on the right-hand side of (3.1) represents the (normalised) negative log-likelihood of a tent function, while the second term can be thought of as a Lagrangian term, which allows us to minimise over the entire class of tent functions, rather than only those $\bar{h}_y$ such that $\exp(\bar{h}_y)$ is a density. Although trying to minimise τ might work in principle, one difficulty is that τ is not convex, so this approach is extremely computationally intensive, even with relatively few observations. Another reason for the numerical difficulties stems from the fact that the set of y-values on which τ attains its minimum is rather large: in general it may be possible to alter particular components y_i without changing $\bar{h}_y$. Of course, we could have defined τ as a function of $\bar{h}_y$ rather than as a function of the vector of tent pole heights $y = (y_1, \ldots, y_n)$. Our choice, however, motivates the following definition of a modified objective function:

$$\sigma(y_1, \ldots, y_n) = -\frac{1}{n}\sum_{i=1}^{n} y_i + \int_{C_n} \exp\{\bar{h}_y(x)\}\, dx. \tag{3.2}$$

The great advantages of minimising σ rather than τ are seen by the following theorem.

THEOREM 3. *The function σ is a convex function satisfying $\sigma \geq \tau$. It has a unique minimum at $y^* \in \mathbb{R}^n$, say, and $\log \hat{f}_n = \bar{h}_{y^*}$.*

Thus Theorem 3 shows that the unique minimum $y^* = (y_1^*, \ldots, y_n^*)$ of σ belongs to the minimum set of τ. In fact, it corresponds to the element of the minimum set for which $\bar{h}_{y^*}(X_i) = y_i^*$ for $i = 1, \ldots, n$. Informally, then, $\bar{h}_{y^*}$ is 'a tent function with all of the tent poles touching the tent'.

In order to compute the function σ at a generic point $y = (y_1, \ldots, y_n) \in \mathbb{R}^n$, we need to be able to evaluate the integral in (3.2). It turns out that we can establish an explicit closed formula for this integral by *triangulating* the convex hull C_n in such a way that $\log \hat{f}_n$ coincides with an *affine function* on each *simplex* in the triangulation. Such a triangulation is illustrated in Figure 1. The structure of the estimator and the issue of computing σ are described in greater detail in Appendix B.

3.2. Nonsmooth optimisation

There is a vast literature on techniques of convex optimisation (cf. Boyd and Vandenberghe (2004), for example), including the method of steepest descent and Newton's method. Unfortunately, these methods rely on the differentiability of the objective function, and the function σ is not differentiable. This can be seen informally by studying the schematic diagram in Figure 1 again. If the ith tent pole, say, is touching but not critically supporting the tent, then decreasing the height of this tent pole does not change the tent function, and thus does not alter the integral in (3.2); on the other hand, increasing the height of the tent pole does alter the tent function and therefore the integral in (3.2). This argument may be used to show that at such a point, the ith partial derivative of σ does not exist.

The set of points at which σ is not differentiable constitute a set of Lebesgue measure zero, but the non-differentiability cannot be ignored in our optimisation procedure. Instead, it is necessary to derive a *subgradient* of σ at each point $y \in \mathbb{R}^n$. This derivation, along with a more formal discussion of the non-differentiability of σ, can be found in Appendix B.2.

The theory of non-differentiable, convex optimisation is perhaps less well-known than its differentiable counterpart, but a fundamental contribution was made by Shor (1985) with his introduction of the subgradient method for minimising non-differentiable, convex functions defined on Euclidean spaces. A slightly specialised version of his Theorem 2.2 gives that if $\partial\sigma(y)$ is a subgradient of σ at y, then for any $y^{(0)} \in \mathbb{R}^n$, the sequence generated by the formula

$$y^{(\ell+1)} = y^{(\ell)} - h_{\ell+1} \frac{\partial\sigma(y^{(\ell)})}{\|\partial\sigma(y^{(\ell)})\|}$$

has the property that either there exists an index ℓ^* such that $y^{(\ell^*)} = y^*$, or $y^{(\ell)} \to y^*$ and $\sigma(y^{(\ell)}) \to \sigma(y^*)$ as $\ell \to \infty$, provided we choose the step lengths h_ℓ so that $h_\ell \to 0$ as $\ell \to \infty$, but $\sum_{\ell=1}^{\infty} h_\ell = \infty$.

Shor recognised, however, that the convergence of this algorithm could be slow in practice, and that although appropriate step size selection could improve matters somewhat, the convergence would never be better than linear (compared with quadratic convergence for Newton's method near the optimum – see Boyd and Vandenberghe (2004, Section 9.5)). Slow convergence can be caused by taking at each stage 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 produce the desired improvements in convergence rate.

One solution (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 functions, an appropriate transformation would be an approximation to the inverse of the Hessian matrix at the optimum. This is not possible for nonsmooth problems, because the inverse might not even exist (and will not exist at points at which the function is not differentiable, 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 convergence in the worst-case scenario of steps nearly perpendicular to the direction towards the minimiser. This variant, which has become known as Shor's r-algorithm, has been implemented in Kappel and Kuntsevich (2000). Accompanying software `SolvOpt` is available from `http://www.uni-graz.at/imawww/kuntsevich/solvopt/`.

Although the formal convergence of the r-algorithm has not been proved, we agree with the authors' claims that it is robust, efficient and accurate. Of course, it is clear that if we terminate the r-algorithm after any finite number of steps and apply the original Shor algorithm using our terminating value of y as the new starting value, then formal convergence is guaranteed. We have not found it necessary to run the original Shor algorithm after termination of the r-algorithm in practice.

If $(y^{(\ell)})$ denotes the sequence of vectors in $\mathbb{R}^n$ produced by the r-algorithm, we terminate when

- $|\sigma(y^{(\ell+1)}) - \sigma(y^{(\ell)})| \leq \delta$
- $|y_i^{(\ell+1)} - y_i^{(\ell)}| \leq \epsilon$ for $i = 1, \ldots, n$
- $|1 - \int \exp\{\bar{h}_{y^{(\ell)}}(x)\}\, dx| \leq \eta$

Table 1. Approximate running times (with number of iterations in brackets) for computing the log-concave maximum likelihood estimator

	$n = 100$	$n = 200$	$n = 500$	$n = 1000$	$n = 2000$
$d = 2$	1.5 secs (260)	2.9 secs (500)	50 secs (1270)	4 mins (2540)	24 mins (5370)
$d = 3$	6 secs (170)	12 secs (370)	100 secs (820)	7 mins (1530)	44 mins (2740)
$d = 4$	23 secs (135)	52 secs (245)	670 secs (600)	37 mins (1100)	224 mins (2060)

for some small δ, ϵ and $\eta > 0$. The first two termination criteria follow Kappel and Kuntsevich (2000), while the third is based on our knowledge that the true optimum corresponds to a density. Throughout this paper, we took $\delta = 10^{-8}$ and $\epsilon = \eta = 10^{-4}$.

Table 1 gives sample running times and the approximate number of iterations of Shor's r-algorithm required for different sample sizes and dimensions on an ordinary desktop computer (1.8GHz, 2GB RAM). Unsurprisingly, the running time increases relatively quickly with the sample size, while the number of iterations increases approximately linearly with n. Each iteration takes longer as the dimension increases, though it is interesting to note that the number of iterations required for the algorithm to terminate decreases as the dimension increases.

When $d = 1$, we recommend the Active Set algorithm of Dümbgen, Hüsler and Rufibach (2007), which is implemented in the R package `logcondens` (Rufibach and Dümbgen, 2006). However, this method relies on the particularly simple structure of triangulations of $\mathbb{R}$, which means that the cone

$$\mathcal{Y}_c = \bigl\{y \colon \bar{h}_y(X_i) = y_i \text{ for } i = 1, \ldots, n\bigr\}$$

can be characterised in a simple way. For $d > 1$, the number of possible triangulations corresponding to a function $\bar{h}_y$ for some $y \in \mathbb{R}^n$ (the so-called regular triangulations) is very large – $O(n^{(d+1)(n-d)})$ – and the cone $\mathcal{Y}_c$ has no such simple structure, so unfortunately the same methods cannot be used.

4. Theoretical properties

The theoretical properties of the log-concave maximum likelihood estimator $\hat{f}_n$ are studied in Cule and Samworth (2010), and in Theorem 4 below we present the main result from that paper. See also Schuhmacher and Dümbgen (2010) and Dümbgen, Samworth and Schuhmacher (2010) for related results. First recall that the Kullback–Leibler divergence of a density f from the true underlying density f_0 is given by

$$d_{KL}(f_0, f) = \int_{\mathbb{R}^d} f_0 \log \frac{f_0}{f}.$$

It is a simple consequence of Jensen's inequality that the Kullback–Leibler divergence $d_{KL}(f_0, f)$ is always non-negative. The first part of Theorem 4 asserts under very weak conditions the existence and uniqueness of a log-concave density f^* that minimises the Kullback–Leibler divergence from f_0 over the class of all log-concave densities.

In the special case where the true density is log-concave, the Kullback–Leibler divergence can be minimised (in fact, made to equal zero) by choosing $f^* = f_0$. The second part of the theorem then gives that with probability one, the log-concave maximum likelihood estimator $\hat{f}_n$ converges to f_0 in certain exponentially weighted total variation distances. The range of possible exponential weights is explicitly linked to the rate of tail decay of f_0. Moreover, if f_0 is continuous, then the convergence also occurs in exponentially weighted supremum distances. We note that when f_0 is log-concave, it can only have discontinuities on the boundary of the (convex) set on which it is positive, a set of zero Lebesgue measure. We therefore conclude that $\hat{f}_n$ is strongly consistent in these norms. It is important to note that the exponential weighting in these distances makes for a very strong notion of convergence (stronger than, say, convergence in Hellinger distance, or unweighted total variation distance), and therefore in particular gives reassurance about the performance of the estimator in the tails of the density.

However, the theorem applies much more generally to situations where f_0 is not log-concave; in other words, where the model has been misspecified. It is important to understand the behaviour of $\hat{f}_n$ in this instance, because one can never be certain from a particular sample of data that the underlying density is log-concave. In the case of model misspecification, the conclusion of the second part of the theorem is that $\hat{f}_n$ converges in the same strong norms as above to the log-concave density f^* that is closest to f_0 in the sense of minimising the Kullback–Leibler divergence. This establishes a desirable robustness property for $\hat{f}_n$, with the natural practical interpretation that provided f_0 is not too far from being log-concave, the estimator is still sensible.

To introduce the notation used in the theorem, we write E for the support of f_0; that is, the smallest closed set with $\int_E f_0 = 1$. We write $\mathrm{int}(E)$ for the interior of E – the largest open set contained in E. Finally, let $\log_+(x) = \max(\log x, 0)$.

Theorem 4. *Let f_0 be any density on $\mathbb{R}^d$ with $\int_{\mathbb{R}^d} \|x\| f_0(x)\,dx < \infty$, $\int_{\mathbb{R}^d} f_0 \log_+ f_0 < \infty$ and $\mathrm{int}(E) \neq \emptyset$. There exists a log-concave density f^*, unique almost everywhere, that minimises the Kullback–Leibler divergence of f from f_0 over all log-concave densities f. Taking $a_0 > 0$ and $b_0 \in \mathbb{R}$ such that $f^*(x) \leq e^{-a_0\|x\| + b_0}$, we have for any $a < a_0$ that*

$$\int_{\mathbb{R}^d} e^{a\|x\|} |\hat{f}_n(x) - f^*(x)|\,dx \overset{a.s.}{\to} 0$$

as $n \to \infty$, and, if f^* is continuous, $\sup_{x\in\mathbb{R}^d} e^{a\|x\|} |\hat{f}_n(x) - f^*(x)| \overset{a.s.}{\to} 0$ as $n \to \infty$.

We remark that the conditions of the theorem are very weak indeed, and in particular are satisfied by any log-concave density on $\mathbb{R}^d$. It is also proved in Cule and Samworth (2010, Lemma 1) that given any log-concave density f^*, we can always find $a_0 > 0$ and $b_0 \in \mathbb{R}$ such that $f^*(x) \leq e^{-a_0\|x\| + b_0}$, so there is no danger of the conclusion being vacuous.

5. Finite sample performance

Our simulation study considered the following densities:

(a) standard normal, $\phi_d \equiv \phi_{d,I}$
(b) dependent normal, $\phi_{d,\Sigma}$, with $\Sigma_{ij} = \mathbb{1}_{\{i=j\}} + 0.2\mathbb{1}_{\{i\neq j\}}$
(c) the joint density of independent $\Gamma(2, 1)$ components

Table 2. 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.

	Log-c	Depend	Norm	Mix	Skewed	Bded
(a)	Yes	No	Yes	No	No	No
(b)	Yes	Yes	Yes	No	No	No
(c)	Yes	No	No	No	Yes	Yes
(d)	Yes	No	Yes	Yes	No	No
(e)	Yes	No	Yes	Yes	No	No
(f)	No	No	Yes	Yes	No	No

(d-f) the normal location mixture $0.6\phi_d(\cdot) + 0.4\phi_d(\cdot - \mu)$ for (d) $\|\mu\| = 1$, (e) $\|\mu\| = 2$, (f) $\|\mu\| = 3$. An application of Proposition 1 tells us that such a normal location mixture is log-concave if and only if $\|\mu\| \leq 2$.

These densities were chosen to exhibit a variety of features, summarised in Table 2. For each density, for $d = 2$ and 3, and for sample sizes $n = 100, 200, 500, 1000$ and 2000, we computed an estimate of the MISE of the log-concave maximum likelihood estimator by averaging the integrated squared error (ISE) over 100 iterations.

We also estimated the MISE for a kernel density estimator using a Gaussian kernel and a variety of bandwidth selection methods, both fixed and variable. These were:

(i) The theoretically optimal bandwidth, computed by minimising the MISE (or asymptotic MISE where closed-form expressions for the MISE were not available)

(ii) Least-squares cross-validation (Wand and Jones, 1995, Section 4.7)

(iii) Smoothed cross-validation (Hall, Marron and Park, 1992; Duong, 2004)

(iv) A 2-stage plug-in rule (Duong and Hazelton, 2003)

(v) Abramson's method. This method, proposed in Abramson (1982), chooses a bandwidth matrix of the form $h\hat{f}^{-1/2}(x)A$, where h is a global smoothing parameter (chosen by cross-validation), $\hat{f}$ a pilot estimate of the density (a kernel estimate with bandwidth chosen by a normal scale rule) and A a shape matrix (chosen to be the diagonal of the sample covariance matrix to ensure appropriate scaling). This is viewed as the benchmark for adaptive bandwidth selection methods.

(vi) Sain's method (Sain, 2002; Scott and Sain, 2004). This divides the sample space up into m^d equally spaced bins and chooses a bandwidth matrix of the form hI for each bin, with h selected by cross-validation. We used $m = 7$.

For density (f), we also used the log-concave EM algorithm described in Section 6 to fit a mixture of two log-concave components. Further examples and implementation details can be found in Cule (2009).

Results are given in Figure 3 and Figure 4. These show only the log-concave maximum likelihood estimator, the MISE-optimal bandwidth, the plug-in bandwidth and Abramson's bandwidth. The other fixed bandwidth selectors (least-squares cross-validation and smoothed cross-validation) performed similarly to or worse than the plug-in estimator (Cule, 2009). This is consistent with the experience of Duong and Hazelton (2003, 2005) who perform a thorough investigation of these methods.

The Sain estimator is particularly difficult to calibrate in practice. Various other binning rules have been tried (Duong, 2004), with little success. Our version of Sain's method performed consistently worse than the Abramson estimator. We suggest that the relatively simple structure of the densities considered here means that this approach is not suitable.

We see that, for cases (a)-(e), the log-concave maximum likelihood estimator has a smaller MISE than the kernel estimator, regardless of choice of bandwidth, for moderate or large sample sizes. Remarkably, our estimator outperforms the kernel estimator even when the bandwidth is chosen based on knowledge of the true density to minimise the MISE. The improvements over kernel estimators are even more marked for $d = 3$ than for $d = 2$. Despite the early promise of adaptive bandwidth methods, they are unable to improve significantly on the performance of fixed bandwidth selectors for our examples. The relatively poor performance of the log-concave maximum likelihood estimator for small sample sizes appears to caused by the poor approximation of the convex hull of the data to the support of the underlying density. This effect becomes negligible in larger sample sizes; see also Section 9. Note that the dependence in case (b) and restricted support in case (c) do not hinder the performance of the log-concave estimator.

In case (f), where the assumption of log-concavity is violated, it is not surprising to see that the performance of our estimator is not as good as that of the optimal fixed bandwidth kernel estimator, but it is still comparable for moderate sample sizes with data-driven kernel estimators (particularly when $d = 3$). This illustrates the robustness property described in Theorem 4. In this case we may recover good performance at larger sample sizes by using a mixture of two log-concave components.

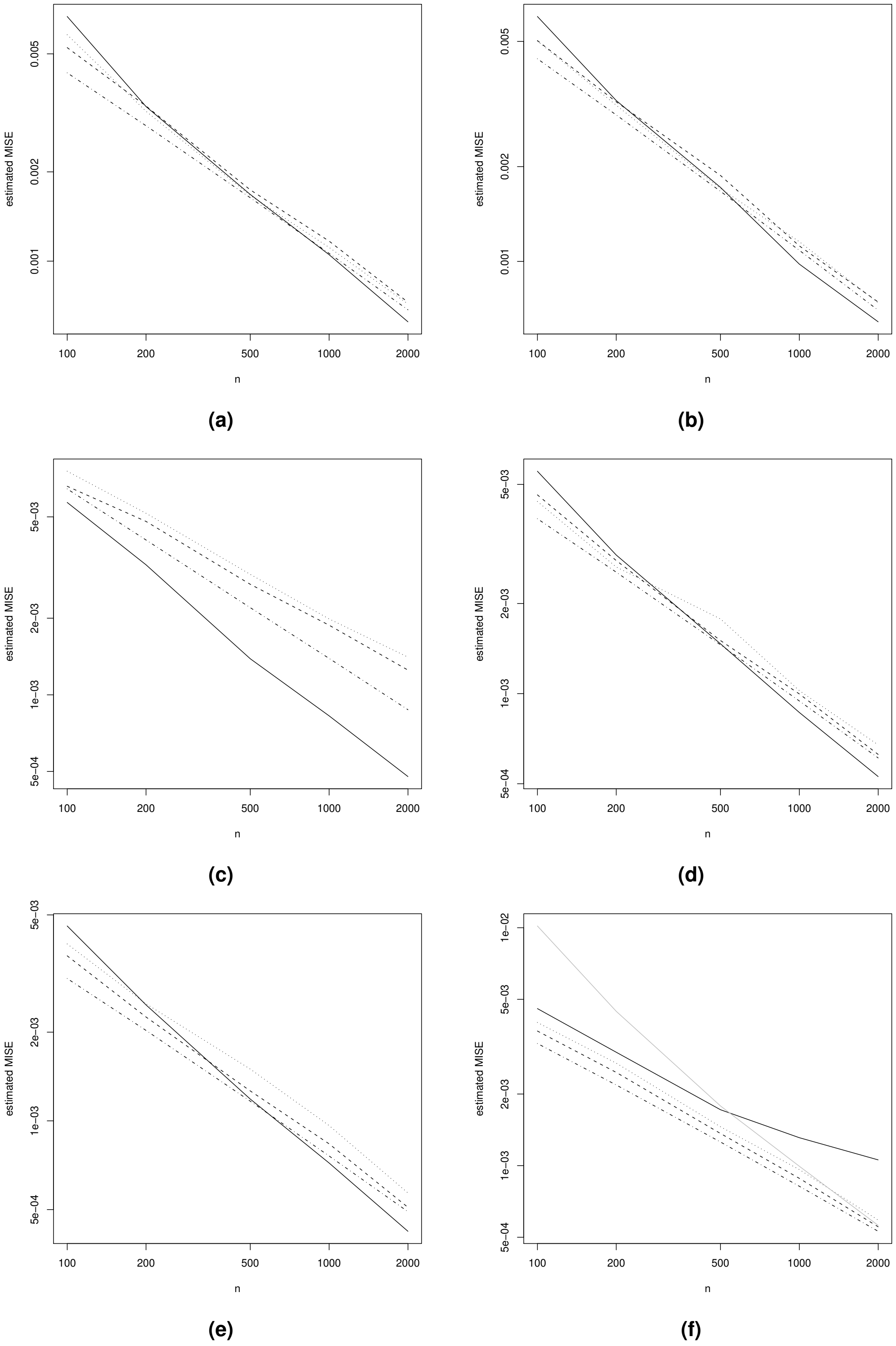
To further investigate the impact of boundary effects, we performed the same simulations for a bivariate density with independent components having a Unif(0,1) distribution and a Beta(2,4) distribution respectively. The results are shown in Figure 5. In this case, boundary bias is particularly problematic for the kernel density estimator, but does not inhibit the performance of the log-concave estimator.

6. Clustering example

In a recent paper, Chang and Walther (2007) introduced an algorithm which combines the univariate log-concave maximum likelihood estimator with the EM algorithm (Dempster, Laird and Rubin, 1977), to fit a finite mixture density of the form

$$f(x) = \sum_{j=1}^{p} \pi_j f_j(x), \tag{6.1}$$

where the mixture proportions $\pi_1, \ldots, \pi_p$ are positive and sum to one, and the component densities $f_1, \ldots, f_p$ are univariate and log-concave. The method is an extension of the standard Gaussian EM



(a)

(b)

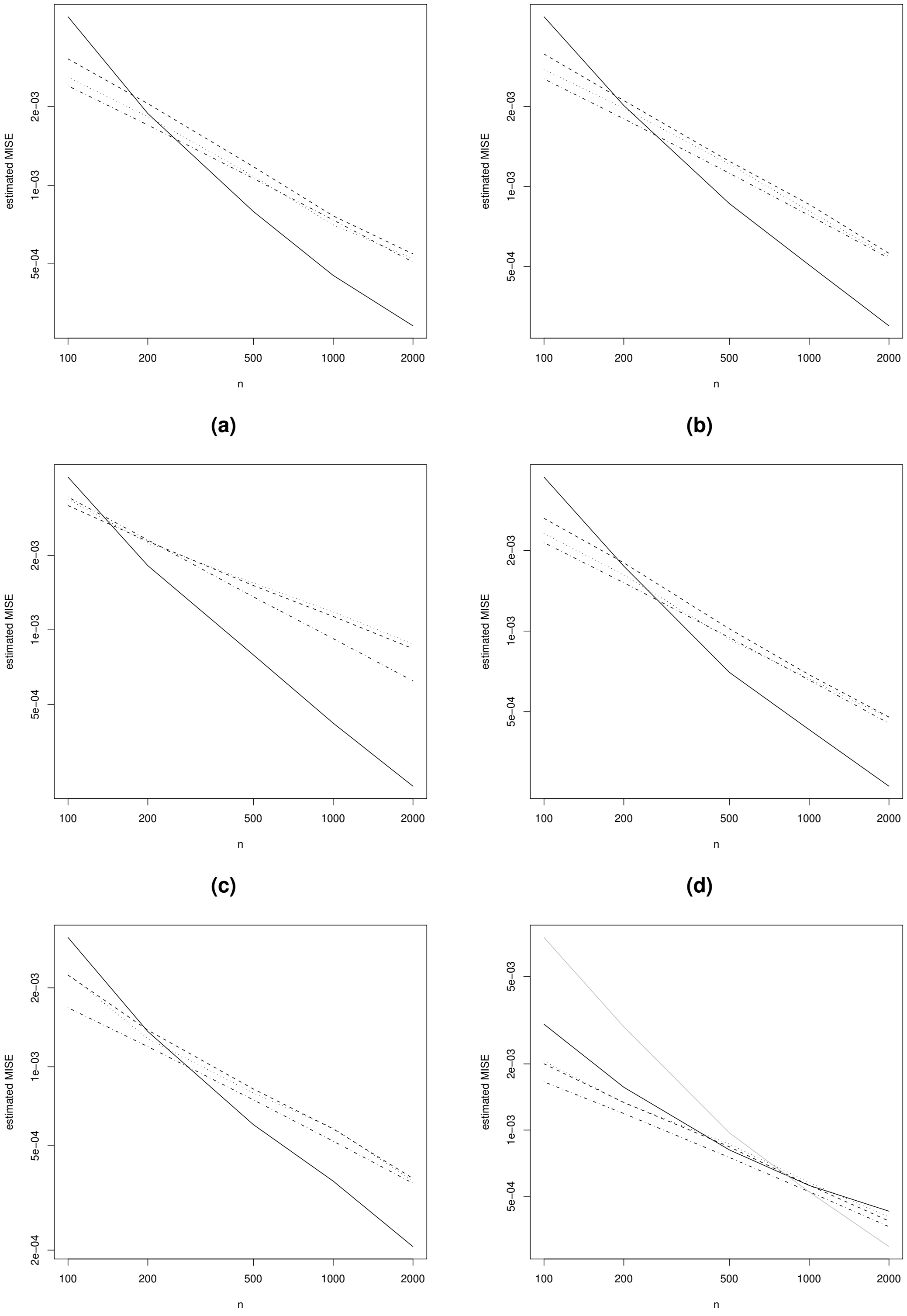
(c)

(d)

(e)

(f)

Fig. 3. MISE, $d = 2$. The solid line is the **LogConcDEAD** estimate, the dashed line the plug-in kernel estimate, the dotted line the Abramson kernel estimate and the dot-dashed line the MISE-optimal bandwidth kernel estimate. The grey line (density f only) is for a 2-component log-concave mixture.



(a) **(b)**

(c) **(d)**

(e) **(f)**

Fig. 4. MISE, $d = 3$. The solid line is the **LogConcDEAD** estimate, the dashed line the plug-in kernel estimate, the dotted line the Abramson kernel estimate and the dot-dashed line the MISE-optimal bandwidth kernel estimate. The grey line (density f only) is for a 2-component log-concave mixture.

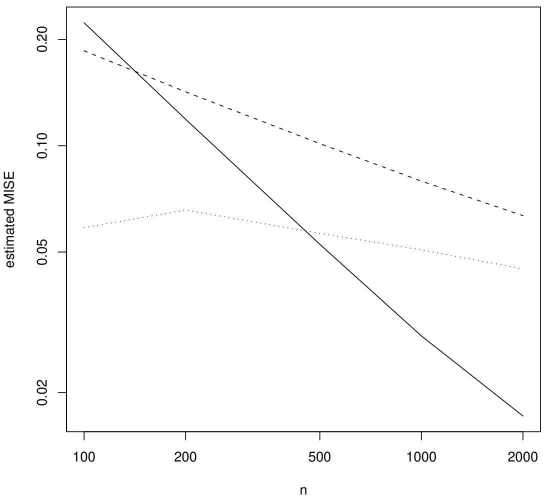


Fig. 5. MISE, $d = 2$, Bivariate uniform/Beta density. The solid line is the **LogConcDEAD** estimate, the dashed line the plug-in kernel estimate, and the dotted line the Abramson kernel estimate.

algorithm, e.g. Fraley and Raftery (2002), which assumes that each component density is normal. Once estimates $\hat{\pi}_1, \ldots, \hat{\pi}_p, \hat{f}_1, \ldots, \hat{f}_p$ have been obtained, clustering can be carried out by assigning to the jth cluster those observations X_i for which $j = \mathrm{argmax}_r\, \hat{\pi}_r \hat{f}_r(X_i)$. Chang and Walther (2007) show empirically that in cases where the true component densities are log-concave but not normal, their algorithm tends to make considerably fewer misclassifications and have smaller mean absolute error in the mixture proportion estimates than the Gaussian EM algorithm, with very similar performance in cases where the true component densities are normal.

Owing to the previous lack of an algorithm for computing the maximum likelihood estimator of a multidimensional log-concave density, Chang and Walther (2007) discuss an extension of the model in (6.1) to a multivariate context where the univariate marginal densities of each component in the mixture are assumed to be log-concave, and the dependence structure within each component density is modelled with a normal copula. Now that we are able to compute the maximum likelihood estimator of a multidimensional log-concave density, we can carry this method through to its natural conclusion. That is, in the finite mixture model (6.1) for a multidimensional log-concave density f, we simply assume that each of the component densities $f_1, \ldots, f_p$ is log-concave. An interesting problem that we do not address here is that of finding appropriate conditions under which this model is identifiable – see Titterington, Smith and Makov (1985, Section 3.1) for a nice discussion.

6.1. EM algorithm

An introduction to the EM algorithm can be found in McLachlan and Krishnan (1997). Briefly, given current estimates of the mixture proportions and component densities $\hat{\pi}_1^{(\ell)}, \ldots, \hat{\pi}_p^{(\ell)}, \hat{f}_1^{(\ell)}, \ldots, \hat{f}_p^{(\ell)}$ at the ℓth iteration of the algorithm, we update the estimates of the mixture proportions by setting $\hat{\pi}_j^{(\ell+1)} = n^{-1}\sum_{i=1}^n \hat{\theta}_{i,j}^{(\ell)}$ for $j = 1, \ldots, p$, where

$$\hat{\theta}_{i,j}^{(\ell)} = \frac{\hat{\pi}_j^{(\ell)} \hat{f}_j^{(\ell)}(X_i)}{\sum_{r=1}^p \hat{\pi}_r^{(\ell)} \hat{f}_r^{(\ell)}(X_i)}$$

is the current estimate of the posterior probability that the ith observation belongs to the jth component. We then update the estimates of the component densities in turn using the algorithm described in Section 3, choosing $\hat{f}_j^{(\ell+1)}$ to be the log-concave density f_j that maximises

$$\sum_{i=1}^n \hat{\theta}_{i,j}^{(\ell)} \log f_j(X_i).$$

The incorporation of the weights $\hat{\theta}_{1,j}^{(\ell)}, \ldots, \hat{\theta}_{n,j}^{(\ell)}$ in the maximisation process presents no additional complication, as is easily seen by inspecting the proof of Theorem 2. As usual with methods based on the EM algorithm, although the likelihood increases at each iteration, there is no guarantee that the sequence converges to a global maximum. In fact, it can happen that the algorithm produces a sequence that approaches a degenerate solution, corresponding to a component concentrated on a single observation, so that the likelihood becomes arbitrarily high. The same issue can arise when fitting mixtures of Gaussian densities, and in this context Fraley and Raftery (2002) suggest that a Bayesian approach can alleviate the problem in these instances by effectively smoothing the likelihood. In general, it is standard practice to restart the algorithm from different initial values, taking the solution with the highest likelihood.

In our case, because of the computational intensity of our method, we first cluster the points according to a hierarchical Gaussian clustering model and then iterate the EM algorithm until the increase in the likelihood is less than 10^{-3} at each step. This differs from Chang and Walther (2007), who used a Gaussian mixture as a starting point. We found that this approach did not allow sufficient flexibility in a multivariate context.

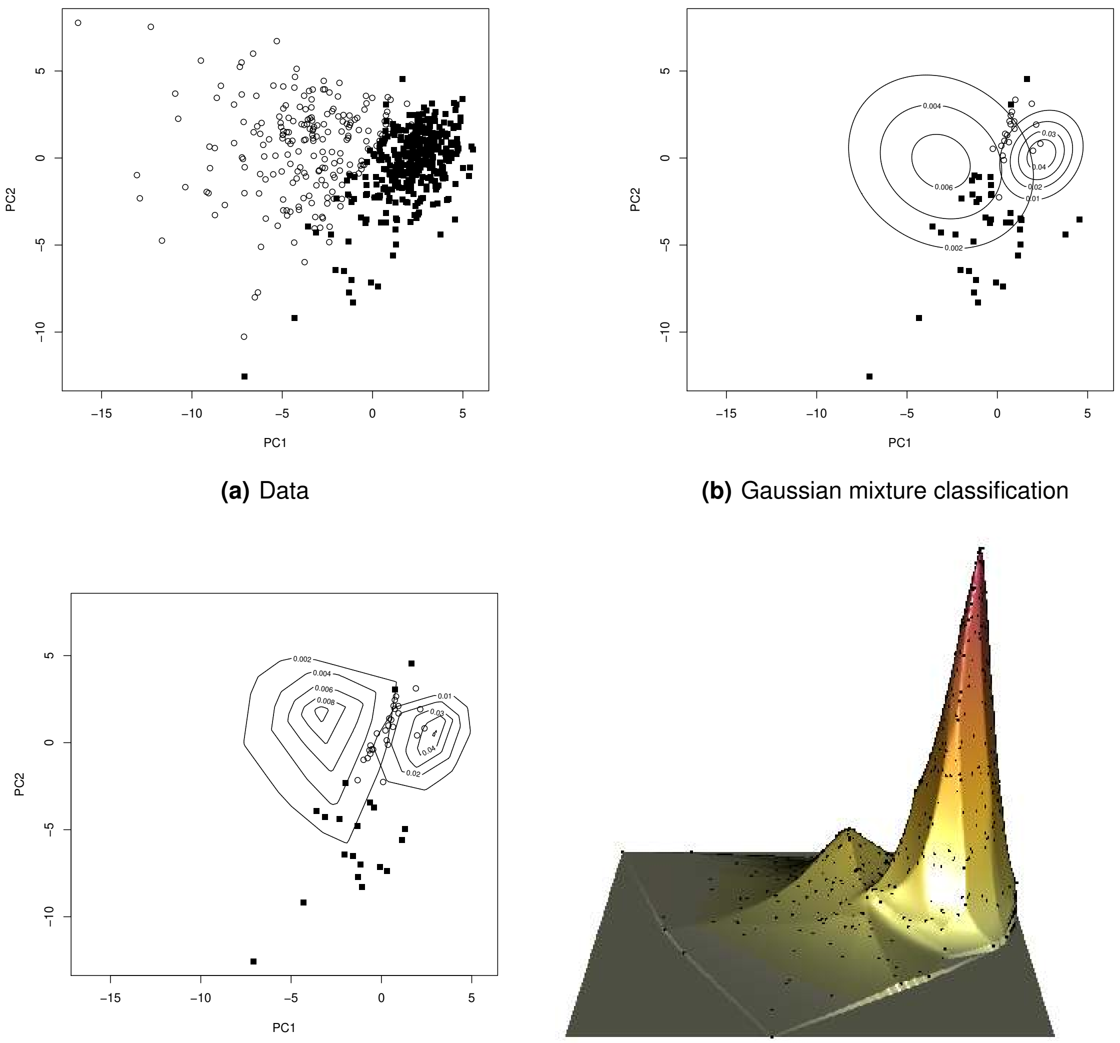
6.2. Breast cancer example

We illustrate the log-concave EM algorithm on the Wisconsin breast cancer data set of Street, Wolberg and Mangasarian (1993), available on the UCI Machine Learning Repository website (Asuncion and Newman, 2007):

`http://archive.ics.uci.edu/ml/datasets/Breast+Cancer+Wisconsin+%28Diagnostic%29.`

The data set was created by taking measurements from a digitised image of a fine needle aspirate of a breast mass, for each of 569 individuals, with 357 benign and 212 malignant instances. We study the problem of trying to diagnose (cluster) the individuals based on the first two principal components of the 30 different measurements, which capture 63% of the variability in the full dataset. These data are presented in Figure 6(a).

It is important also to note that although for this particular data set we do know whether a particular instance is benign or malignant, we did not use this information in fitting our mixture model.



(a) Data

(b) Gaussian mixture classification

(c) Log-concave mixture classification

(d) Estimated log-concave mixture

Fig. 6. Panel (a) plots the Wisconsin breast cancer data, with benign cases as solid squares and malignant ones as open circles. Panel (b) gives a contour plot together with the misclassified instances from the Gaussian EM algorithm, while the corresponding plot obtained from the log-concave EM algorithm is given in Panel (c). Panel (d) plots the fitted mixture distribution from the log-concave EM algorithm.

Instead this information was only used afterwards to assess the performance of the method, as reported below. Thus we are studying a clustering (or unsupervised learning) problem, by taking a classification (or supervised learning) data set and 'covering up the labels' until it comes to performance assessment.

The skewness in the data suggests that the mixture of Gaussians model may be inadequate, and in Figure 6(b) we show the contour plot and misclassified instances from this model. The corresponding plot obtained from the log-concave EM algorithm is given in Figure 6(c), while Figure 6(d) plots the fitted mixture distribution from the log-concave EM algorithm. For this example, the number of misclassified instances is reduced from 59 with the Gaussian EM algorithm to 48 with the log-concave EM algorithm.

In some examples, it will be necessary to estimate p, the number of mixture components. In the general context of model-based clustering, Fraley and Raftery (2002) cite several possible approaches for this purpose, including methods based on resampling (McLachlan and Basford, 1988) and an information criterion (Bozdogan, 1994). Further research will be needed to ascertain which of these methods is most appropriate in the context of log-concave component densities.

7. Plug-in estimation of functionals, sampling and the bootstrap

Suppose X has density f. Often, we are less interested in estimating a density directly than in estimating some functional $\theta = \theta(f)$. Examples of functionals of interest (some of which were given in Section 1), include:

(a) $\mathbb{P}(\|X\| \geq 1) = \int f(x)\mathbb{1}_{\{\|x\|\geq 1\}}\,dx$
(b) Moments, such as $\mathbb{E}(X) = \int xf(x)\,dx$, or $\mathbb{E}(\|X\|^2) = \int \|x\|^2 f(x)\,dx$
(c) The differential entropy of X (or f), defined by $H(f) = -\int f(x)\,\log f(x)\,dx$
(d) The $100(1-\alpha)\%$ highest density region, defined by $R_\alpha = \{x \in \mathbb{R}^d : f(x) \geq f_\alpha\}$, where f_α is the largest constant such that $\mathbb{P}(X \in R_\alpha) \geq 1-\alpha$. Hyndman (1996) argues that this is an informative summary of a density; note that subject to a minor restriction on f, we have $\int f(x)\mathbb{1}_{\{f(x)\geq f_\alpha\}}\,dx = 1-\alpha$.

Each of these may be estimated by the corresponding functional $\hat{\theta} = \theta(\hat{f}_n)$ of the log-concave maximum likelihood estimator. In examples (a) and (b) above, $\theta(f)$ may also be written as a functional of the corresponding distribution function F, e.g. $\mathbb{P}(\|X\| \geq 1) = \int \mathbb{1}_{\{\|x\|\geq 1\}}dF(x)$. In such cases, it is more natural to use the plug-in estimator based on the empirical distribution function, $\hat{F}_n$, of the sample $X_1, \ldots, X_n$, and indeed in our simulations we found that the log-concave plug-in estimator did not offer an improvement on this method. In the other examples, however, an empirical distribution function plug-in estimator is not available, and the log-concave plug-in estimator is a potentially attractive procedure.

To provide some theoretical justification for this, observe from Section 4 that we can think of the sequence $(\hat{f}_n)$ as taking values in the space $\mathcal{B}$ of (measurable) functions with finite $\|\cdot\|_{1,a}$ norm for some $a > 0$, where $\|f\|_{1,a} = \int e^{a\|x\|}|f(x)|\,dx$. The conclusion of Theorem 4 is that $\|\hat{f}_n - f^*\|_{1,a} \overset{a.s.}{\to} 0$

as $n \to \infty$ for a range of values of a, where f^* is the log-concave density that minimises the Kullback–Leibler divergence from the true density. If the functional $\theta(f)$ takes values in another normed space (e.g. $\mathbb{R}$) with norm $\|\cdot\|$ and is a continuous function on $\mathcal{B}$, then $\|\hat{\theta} - \theta^*\| \overset{a.s.}{\to} 0$, where $\theta^* = \theta(f^*)$. In particular, when the true density is log-concave, $\hat{\theta}$ is strongly consistent.

7.1. Monte Carlo estimation of functionals

For some functionals we can compute $\hat{\theta} = \theta(\hat{f}_n)$ analytically. Suppose now that this is not possible, but that we can write $\theta(f) = \int f(x)g(x)\,dx$ for some function g. Such a functional is continuous (so $\hat{\theta}$ is strongly consistent) provided merely that $\sup_{x\in\mathbb{R}^d} e^{-a\|x\|}|g(x)| < \infty$ for some a in the allowable range provided by Theorem 4. In that case, we may approximate $\hat{\theta}$ by

$$\hat{\theta}_B = \frac{1}{B}\sum_{b=1}^{B} g(X_b^*),$$

for some (large) B, where $X_1^*, \ldots, X_B^*$ are independent samples from $\hat{f}_n$. Conditional on $X_1, \ldots, X_n$, the strong law of large numbers gives that $\hat{\theta}_B \overset{a.s.}{\to} \hat{\theta}$ as $B \to \infty$. In practice, even when analytic calculation of $\hat{\theta}$ was possible, this method was found to be fast and accurate.

In order to use this Monte Carlo procedure, we must be able to sample from $\hat{f}_n$. Fortunately, this can be done efficiently using the rejection sampling procedure described in Section B.3 in the Appendix.

7.2. Simulation study

In this section we illustrate some simple applications of this idea to functionals (c) and (d) above. An expression for computing (c) may be found in Cule (2009). For (d), closed-form integration is not possible, so we use the method of Section 7.1. Estimates are based on random samples of size $n = 500$ from a $N_2(0, I)$ distribution, and we compare the performance of the **LogConcDEAD** estimate with that of a kernel-based plug-in estimate, where the bandwidth was chosen using a plug-in rule (the choice of bandwidth did not have a big influence on the outcome; see Cule (2009)).

This was done for all of the densities in Section 5, though we present results only for density (c) and $d = 2$ for reasons of space. See Cule (2009) for further examples and results. In Figure 7 we study the plug-in estimators $\hat{R}_\alpha$ of the highest density region R_α, and measure the quality of the estimation procedures through $\mathbb{E}\{\mu_f(\hat{R}_\alpha \triangle R_\alpha)\}$, where $\mu_f(A) = \int_A f(x)\,dx$ and $\triangle$ denotes set difference. Highest density regions can be computed once we have approximated the sample versions of f_α using the density quantile algorithm described in Hyndman (1996, Section 3.2). The log-concave estimator provides a substantial improvement on the kernel estimator for each of the three levels considered. See also Figure 8.

In real data examples, we are unable to assess uncertainty in our functional estimates by taking repeated samples from the true underlying model. Nevertheless, the fact that we can sample from the log-concave maximum likelihood estimator does mean that we can apply standard bootstrap methodology to compute standard errors or confidence intervals, for example. Finally, we remark

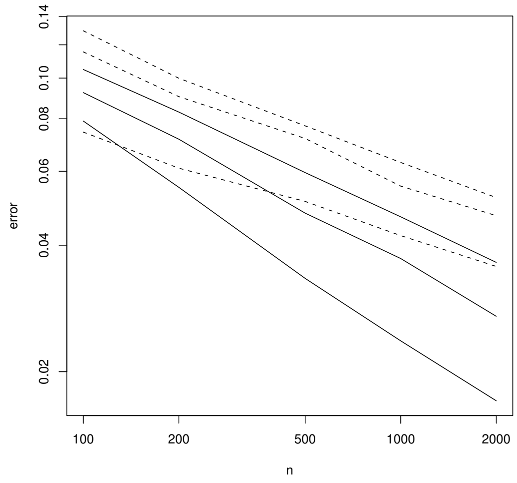


Fig. 7. Error for the highest density regions. The solid lines are the **LogConcDEAD** estimates; the dashed lines are the kernel estimates. The lowest lines are the 25% HDR, the middle lines are the 50% HDR and the highest lines are the 75% HDR.

that the plug-in estimation procedure, sampling algorithm and bootstrap methodology extend in an obvious way to the case of a finite mixture of log-concave densities.

8. Assessing log-concavity

In Section 4 we mentioned the fact that one can never be certain that a particular data set comes from a log-concave density. Even though Theorem 4 shows that the log-concave maximum likelihood estimator has a desirable robustness property, it is still desirable to have diagnostic tests for assessing log-concavity. In this section we present two possible hypothesis tests of the null hypothesis that the underlying density is log-concave.

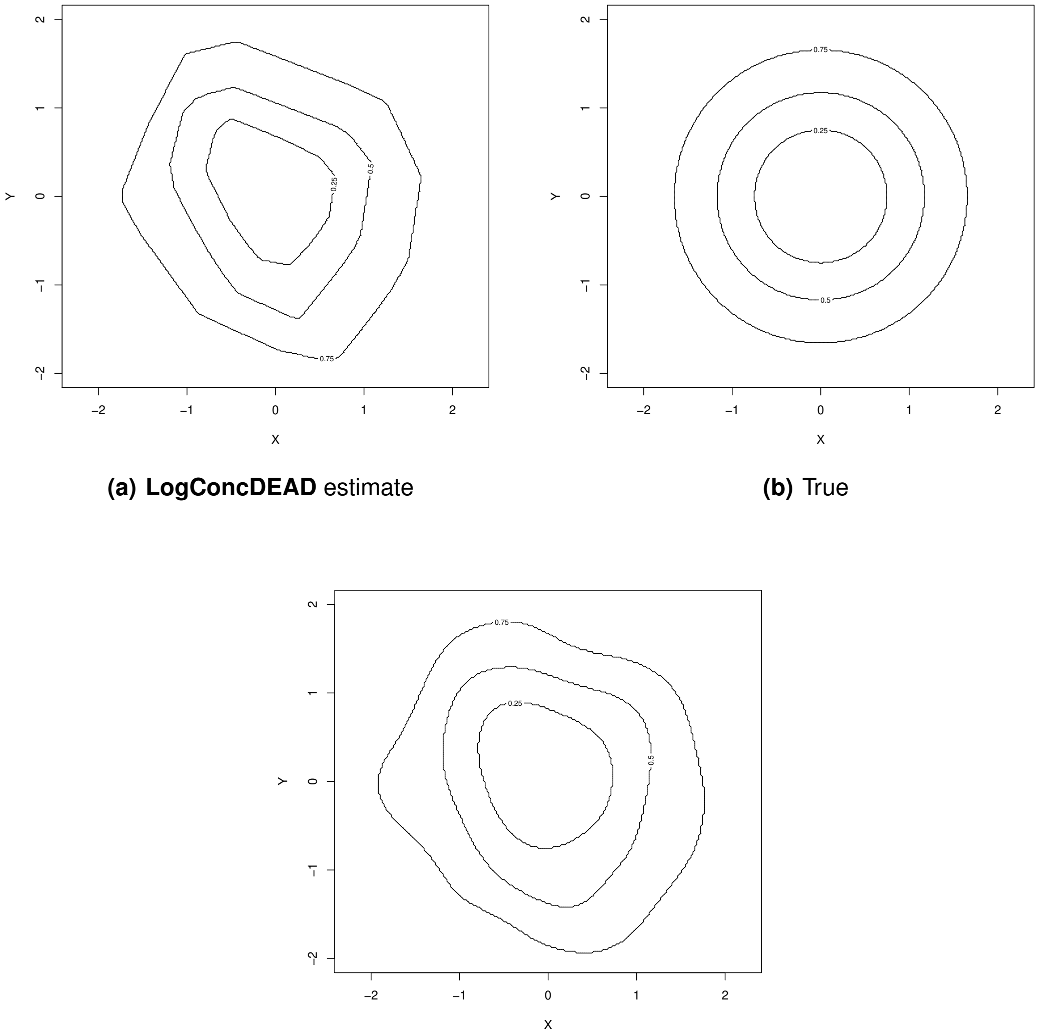
The first uses a method similar to that described in Walther (2002) to test the null hypothesis that a log-concave model adequately models the data, compared to the alternative that

$$f(x) = \exp(\phi(x) + c\|x\|^2)$$

for some concave function ϕ and $c > 0$. This was originally suggested to detect mixing, as Walther (2002) proves that a finite mixture of log-concave densities has a representation of this form, but in fact captures more general alternatives to log-concavity such as heavy tails. In order to do this, we compute

$$\hat{f}_n^c = \operatorname*{arg\,max}_{f \in \mathcal{F}^c} L(f)$$

for fixed values $c \in \mathcal{C} = \{c_0, \ldots, c_M\}$, where $\mathcal{F}^c = \bigl\{f \colon f(x) = \exp(\phi(x) + c\|x\|^2) \text{ with } \phi \text{ concave}\bigr\}$.



(a) LogConcDEAD estimate

(b) True

(c) Kernel estimate

Fig. 8. Estimates of the 25%, 50% and 75% highest density region from 500 observations from the $N_2(0, I)$ distribution.

We wish to assess how much $\hat{f}_n^c$ deviates from log-concavity; one possible measure is

$$T(c) = \int \{\bar{h}(x) - \log \hat{f}_n^c(x)\}\hat{f}_n^0(x)\,dx$$

where $\bar{h}$ is the least concave majorant of $\log \hat{f}_n^c$. In order to generate a reference distribution, we draw B bootstrap samples from $\hat{f}_n^0$. For each bootstrap sample and each value $c = c_0, \ldots, c_M$, we compute the test statistic defined above, to obtain $T_b^*(c)$ for $b = 1, \ldots, B$. Let $m(c)$ and $s(c)$ denote the sample mean and sample standard deviation respectively of $T_1^*(c), \ldots, T_B^*(c)$. We then standardize the statistics on each scale, computing

$$\tilde{T}(c) = \frac{T(c) - m(c)}{s(c)}$$

and

$$\tilde{T}_b^*(c) = \frac{T_b^*(c) - m(c)}{s(c)}$$

for each $c = c_0, \ldots, c_M$ and $b = 1, \ldots, B$. To perform the test we compute the (approximate) p-value

$$\frac{1}{B+1}\#\left\{b\colon \max_{c\in\mathcal{C}} \tilde{T}(c) > \max_{c\in\mathcal{C}} \tilde{T}_b^*(c)\right\}.$$

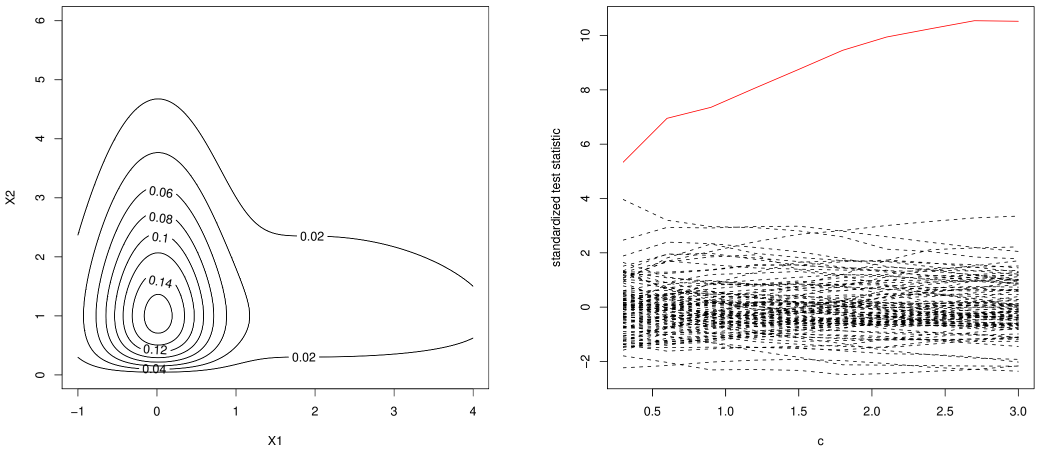
As an illustration, we applied this procedure to a sample of size $n = 500$ from a mixture distribution. The first component was a mixture with density

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

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 density is not log-concave, and 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 9(a) 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 the smallest c such that $f(x) = \exp(\phi(x) + c\|x\|^2)$ for concave ϕ is not invariant under rescaling, so we wish to have all dimensions on the same scale before performing the test. The resulting p-value was less than 0.01. Figure 9(b) shows the values of the test statistic for various values of c (on the standardized scale). See Cule (2009) for further examples. Unfortunately, this test is currently not practical except for small sample sizes because of the computational burden of computing the test statistics for the many bootstrap samples.

We therefore introduce a permutation test that involves fitting only a single log-concave maximum likelihood estimator, and which tests against the general alternative that the underlying density f_0 is not log-concave. The idea is to fit the log-concave maximum likelihood estimator $\hat{f}_n$ to the data $X_1, \ldots, X_n$, and then to draw a sample $X_1^*, \ldots, X_n^*$ from this fitted density. The intuition is that if f_0 is not log-concave, then the two samples $\mathcal{X} = \{X_1, \ldots, X_n\}$ and $\mathcal{X}^* = \{X_1^*, \ldots, X_n^*\}$ should look different. We would like to formalise this idea with a notion of distance, and a fairly natural metric



(a) Contour plot

(b) Test statistic

Fig. 9. Assessing the suitability of log-concavity. The left-hand panel gives a contour plot of the density. In the right-hand panel, the grey line illustrates the value of the test statistic and the dotted lines the bootstrap reference values.

between distributions P and Q in this context is $d(P, Q) = \sup_{A \in \mathcal{A}} |P(A) - Q(A)|$, where $\mathcal{A}$ denotes the class of all (Euclidean) balls in $\mathbb{R}^d$. A sample version of this quantity is

$$T = \sup_{A \in \mathcal{A}_0} |P_n(A) - P_n^*(A)|, \qquad (8.1)$$

where $\mathcal{A}_0$ is the set of all balls centered at a point in $\mathcal{X} \cup \mathcal{X}^*$, and P_n and P_n^* denote the empirical distributions of $\mathcal{X}$ and $\mathcal{X}^*$ respectively. For a fixed ball centre and expanding radius, the quantity $|P_n(A) - P_n^*(A)|$ only changes when a new point enters the ball, so the supremum in (8.1) is attained and the test statistic is easy to compute.

In order to compute the critical value for the test, we 'shuffle the stars' in the combined sample $\mathcal{X} \cup \mathcal{X}^*$; in other words, we re-label the points by choosing a random (uniformly distributed) permutation of the combined sample and putting stars on the last n elements in the permuted combined sample. Writing $P_{n,1}$ and $P_{n,1}^*$ for the empirical distributions of the first n and last n elements in the permuted combined sample respectively, we compute $T_1^* = \sup_{A \in \mathcal{A}_0} |P_{n,1}(A) - P_{n,1}^*(A)|$. Repeating this procedure a further $(B-1)$ times, we obtain $T_1^*, \ldots, T_B^*$, with corresponding order statistics $T_{(1)}^* \leq \ldots \leq T_{(B)}^*$. For a nominal size α test, we reject the null hypothesis of log-concavity if $T > T_{((B+1)(1-\alpha))}^*$.

In practice, we found that some increase in power could be obtained by computing the maximum over all balls containing at most k points in the combined sample instead of computing the maximum over all balls. The reason for this is that if f_0 is not log-concave, then one would expect to find clusters of points with the same label (i.e. with or without stars). Thus the supremum in (8.1) may well be attained at a relatively small ball radius. On the other hand, in the permuted samples, the supremum is likely to be attained at a ball radius that includes approximately half of the points in

Table 3. : Proportion of times out of 200 repetitions that the null hypothesis was rejected

n	$\|\mu\| = 0$	$\|\mu\| = 1$	$\|\mu\| = 2$	$\|\mu\| = 3$	$\|\mu\| = 4$
200	0.01	0	0.015	0.06	0.475
500	0.01	0	0.015	0.065	0.88
1000	0	0.005	0.005	0.12	0.995

the combined sample, so by restricting the ball radius, we will tend to reduce the critical value for the test (potentially without altering the test statistic). Of course, this introduces a parameter k to be chosen. This choice is similar to the problem of choosing k in k-nearest neighbour classification, as studied in Hall, Park and Samworth (2008). There it was shown that, under mild regularity conditions, the misclassification rate is minimised by choosing k to be of order $n^{4/(d+4)}$, but that in practice the performance of the classifier was relatively insensitive to a fairly wide range of choices of k.

To illustrate the performance of the hypothesis test, we ran a small simulation study. We chose the bivariate mixture of normals density $f_0(x) = \frac{1}{2}\phi_2(x) + \frac{1}{2}\phi_2(x - \mu)$, with $\|\mu\| \in \{0, 1, 2, 3, 4\}$, which is log-concave if and only if $\|\mu\| \leq 2$. For each simulation set-up, we conducted 200 hypothesis tests with $k = \lfloor n^{4/(d+4)} \rfloor$ and $B = 99$, and report in Table 3 the proportion of times the null hypothesis was rejected in a size $\alpha = 0.05$ test.

One feature of the test that is apparent from Table 3 is that the test is conservative. This is initially surprising because it indicates that the original test statistic, which is based on two samples that come from slightly different distributions, tends to be a little smaller than the test statistic based on the permuted samples, in which both samples that come from the same distribution. The explanation is that the dependence between $\mathcal{X}$ and $\mathcal{X}^*$ means that the realisations of the empirical distributions P_n and P_n^* tend to be particularly close together. Nevertheless, the test is able to detect the significant departure from log-concavity (when $\|\mu\| = 4$), particularly at larger sample sizes.

9. Concluding discussion

We hope that this paper will stimulate further interest and research in the field of shape-constrained estimation. Indeed, there remain many challenges and interesting directions for future research. As well as the continued development and refinement of the computational algorithms and graphical displays of estimates, and further studies of theoretical properties, these include:

(i) Studying other shape constraints. These have received some attention for univariate data, dating back to Grenander (1956), but in the multivariate setting these are an active area of current development; see, for example, Seregin and Wellner (2009); Koenker and Mizera (2010). Computational, methodological and theoretical questions arise for each different shape constraint, and we hope that this paper might provide some ideas that can be transferred to these different settings.

(ii) Addressing the issue of how to improve performance of shape-constrained estimators at small sample sizes. One idea here, based on an extension of the univariate idea presented in Dümbgen and Rufibach (2009), is the following: we first note that an extension of Theorem 2.2 of Dümbgen and Rufibach (2009) to the multivariate case gives that the covariance matrix $\tilde{\Sigma}$ corresponding to the fitted log-concave maximum likelihood estimator $\hat{f}_n$ is smaller than the sample covariance matrix $\hat{\Sigma}$, in the sense that $A = \hat{\Sigma} - \tilde{\Sigma}$ is non-negative definite. One can therefore define a slightly smoothed version of $\hat{f}_n$ via the convolution

$$\tilde{f}_n(x) = \int_{\mathbb{R}^d} \phi_{d,A}(x-y)\hat{f}_n(y)\,dy.$$

The estimator $\tilde{f}_n$ is still a fully automatic, log-concave density estimator. Moreover, it is supported on the whole of $\mathbb{R}^d$, infinitely differentiable, and the covariance matrix corresponding to $\tilde{f}_n$ is equal to the sample covariance matrix. The estimator $\tilde{f}_n$ will exhibit similar large-sample performance to $\hat{f}_n$ (indeed, Theorem 4 also applies to $\tilde{f}_n$), but offers potential improvements for small sample sizes.

(iii) Assessing the uncertainty in shape-constrained nonparametric density estimates, through confidence intervals/bands.

(iv) Developing analogous methodology and theory for discrete data under shape constraints.

(v) Examining nonparametric shape constraints in regression problems, such as those studied in Dümbgen, Samworth and Schuhmacher (2010), for example.

(vi) Studying methods for choosing the number of clusters in nonparametric, shape-constrained mixture models.

A. Proofs

Proof of Proposition 1

(a) If f is log-concave, then for $x \in \mathbb{R}^d$, we can write

$$f_{X|P_V(X)}(x|t) \propto f(x)\mathbb{1}_{\{P_V(x)=t\}},$$

a product of log-concave functions. Thus $f_{X|P_V(X)}(\cdot|t)$ is log-concave for each t.

(b) Let $x_1, x_2 \in \mathbb{R}^d$ be distinct and let $\lambda \in (0,1)$. Let V be the $(d-1)$-dimensional subspace of $\mathbb{R}^d$ whose orthogonal complement is *parallel* to the *affine hull* of $\{x_1, x_2\}$ (i.e. the line through x_1 and x_2). Writing $f_{P_V(X)}$ for the marginal density of $P_V(X)$ and t for the common value of $P_V(x_1)$ and $P_V(x_2)$, the density of X at $x \in \mathbb{R}^d$ is

$$f(x) = f_{X|P_V(X)}(x|t)f_{P_V(X)}(t).$$

Thus

$$\begin{aligned}\log f\bigl(\lambda x_1 + (1-\lambda)x_2\bigr) &\geq \lambda \log f_{X|P_V(X)}(x_1|t) + (1-\lambda)\log f_{X|P_V(X)}(x_2|t) + \log f_{P_V(X)}(t) \\ &= \lambda \log f(x_1) + (1-\lambda)\log f(x_2),\end{aligned}$$

so f is log-concave, as required. □

Proof of Theorem 2

We may assume that $X_1, \ldots, X_n$ are distinct and their convex hull, $C_n = \mathrm{conv}(X_1, \ldots, X_n)$, is a d-dimensional *polytope* (an event of probability one when $n \geq d+1$). By a standard argument in convex analysis (Rockafellar, 1997, p. 37), for each $y = (y_1, \ldots, y_n) \in \mathbb{R}^n$ there exists a function $\bar{h}_y : \mathbb{R}^d \to \mathbb{R}$ with the property that $\bar{h}_y$ is the least concave function satisfying $\bar{h}_y(X_i) \geq y_i$ for all $i = 1, \ldots, n$. Let $\mathcal{H} = \{\bar{h}_y : y \in \mathbb{R}^n\}$, let $\mathcal{F}$ denote the set of all log-concave functions on $\mathbb{R}^d$, and for $f \in \mathcal{F}$, define

$$\psi_n(f) = \frac{1}{n}\sum_{i=1}^n \log f(X_i) - \int_{\mathbb{R}^d} f(x)\,dx.$$

Suppose that f maximises $\psi_n(\cdot)$ over $\mathcal{F}$. We show in turn that

(i) $f(x) > 0$ for $x \in C_n$
(ii) $f(x) = 0$ for $x \notin C_n$
(iii) $\log f \in \mathcal{H}$
(iv) $f \in \mathcal{F}_0$
(v) there exists $M > 0$ such that if $\max_i |\bar{h}_y(X_i)| \geq M$, then $\psi_n\bigl(\exp(\bar{h}_y)\bigr) \leq \psi_n(f)$.

First note that if $x_0 \in C_n$, then by Carathéodory's theorem (Theorem 17.1 of Rockafellar (1997)), there exist distinct indices $i_1, \ldots, i_r$ with $r \leq d+1$, such that $x_0 = \sum_{l=1}^r \lambda_l X_{i_l}$ with each $\lambda_l > 0$ and $\sum_{l=1}^r \lambda_l = 1$. Thus, if $f(x_0) = 0$, then by Jensen's inequality,

$$-\infty = \log f(x_0) \geq \sum_{l=1}^r \lambda_l \log f(X_{i_l}),$$

so $f(X_i) = 0$ for some i. But then $\psi_n(f) = -\infty$. This proves (i).

Now suppose $f(x_0) > 0$ for some $x_0 \notin C_n$. Then $\{x : f(x) > 0\}$ is a convex set containing $C_n \cup \{x_0\}$, a set which has strictly larger d-dimensional Lebesgue measure than that of C_n. We therefore have $\psi_n(f) < \psi_n(f\mathbb{1}_{C_n})$, which proves (ii).

To prove (iii), we first show that $\log f$ is *closed*. Suppose that $\log f(X_i) = y_i$ for $i = 1, \ldots, n$ but that $\log f \neq \bar{h}_y$. Then since $\log f(x) \geq \bar{h}_y(x)$ for all $x \in \mathbb{R}^d$, we may assume that there exists $x_0 \in C_n$ such that $\log f(x_0) > \bar{h}_y(x_0)$. If x_0 is in the *relative interior* of C_n, then since $\log f$ and $\bar{h}_y$ are continuous at x_0 (by Theorem 10.1 of Rockafellar (1997)), we must have

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

The only remaining possibility is that x_0 is on the *relative boundary* of C_n. But $\bar{h}_y$ is closed by Corollary 17.2.1 of Rockafellar (1997), so writing $\mathrm{cl}(g)$ for the *closure* of a concave function g, we have $\bar{h}_y = \mathrm{cl}(\bar{h}_y) = \mathrm{cl}(\log f) \geq \log f$, where we have used Corollary 7.3.4 of Rockafellar (1997) to obtain the middle equality. It follows that $\log f$ is closed and $\log f = \bar{h}_y$, which proves (iii).

Note that $\log f$ has no *direction of increase*, because if $x \in C_n$, z is a non-zero vector and $t > 0$ is large enough that $x + tz \notin C_n$, then $-\infty = \log f(x + tz) < \log f(x)$. It follows by Theorem 27.2 of Rockafellar (1997) that the supremum of f is finite (and is attained). Using properties (i) and (ii) as well, we may write $\int f(x)\,dx = c$, say, where $c \in (0, \infty)$. Thus $f(x) = c\bar{f}(x)$, for some $\bar{f} \in \mathcal{F}_0$. But then

$$\psi_n(\bar{f}) - \psi_n(f) = -1 - \log c + c \geq 0,$$

with equality only if $c = 1$. This proves (iv).

To prove (v), we may assume by (iv) that $\exp(\bar{h}_y)$ is a density. Let $\max_i \bar{h}_y(X_i) = M$ and let $\min_i \bar{h}_y(X_i) = m$. We show that when M is large, in order for $\exp(\bar{h}_y)$ to be a density, m must be negative with $|m|$ so large that $\psi_n\bigl(\exp(\bar{h}_y)\bigr) \leq \psi_n(f)$. First observe that if $x \in C_n$ and $\bar{h}_y(X_i) = M$, then for M sufficiently large we must have $M - m > 1$, and then

$$\begin{aligned}\bar{h}_y\Bigl(X_i + \frac{1}{M-m}(x - X_i)\Bigr) &\geq \frac{1}{M-m}\bar{h}_y(x) + \frac{M-m-1}{M-m}\bar{h}_y(X_i) \\ &\geq \frac{m}{M-m} + \frac{(M-m-1)M}{M-m} = M - 1.\end{aligned}$$

(The fact that $\bar{h}_y(x) \geq m$ follows by Jensen's inequality.) Hence, denoting Lebesgue measure on $\mathbb{R}^d$ by μ, we have

$$\mu(\{x : \bar{h}_y(x) \geq M - 1\}) \geq \mu\Bigl(\Bigl\{X_i + \frac{1}{M-m}(C_n - X_i)\Bigr\}\Bigr) = \frac{\mu(C_n)}{(M-m)^d}.$$

Thus

$$\int_{\mathbb{R}^d} \exp\{\bar{h}_y(x)\}\,dx \geq e^{M-1}\frac{\mu(C_n)}{(M-m)^d}.$$

For $\exp(\bar{h}_y)$ to be a density, then, we require $m \leq -\frac{1}{2}e^{(M-1)/d}\mu(C_n)^{1/d}$ when M is large. But then

$$\psi_n\bigl(\exp(\bar{h}_y)\bigr) \leq \frac{(n-1)M}{n} - \frac{1}{2n}e^{(M-1)/d}\mu(C_n)^{1/d} \leq \psi_n(f)$$

when M is sufficiently large. This proves (v).

It is not hard to see that for any $M > 0$, the function $y \mapsto \psi_n\bigl(\exp(\bar{h}_y)\bigr)$ is continuous on the compact set $[-M, M]^n$, and thus the proof of the existence of a maximum likelihood estimator is complete. To prove uniqueness, suppose that $f_1, f_2 \in \mathcal{F}$ and both f_1 and f_2 maximise $\psi_n(f)$. We may assume $f_1, f_2 \in \mathcal{F}_0$, $\log f_1, \log f_2 \in \mathcal{H}$ and f_1 and f_2 are supported on C_n. Then the normalised geometric mean

$$g(x) = \frac{\{f_1(x)f_2(x)\}^{1/2}}{\int_{C_n}\{f_1(y)f_2(y)\}^{1/2}\,dy},$$

is a log-concave density, with

$$\begin{aligned}\psi_n(g) &= \frac{1}{2n}\sum_{i=1}^n \log f_1(X_i) + \frac{1}{2n}\sum_{i=1}^n \log f_2(X_i) - \log\int_{C_n}\{f_1(y)f_2(y)\}^{1/2}\,dy - 1 \\ &= \psi_n(f_1) - \log\int_{C_n}\{f_1(y)f_2(y)\}^{1/2}\,dy.\end{aligned}$$

However, by Cauchy–Schwarz, $\int_{C_n}\{f_1(y)f_2(y)\}^{1/2}\,dy \leq 1$, so $\psi_n(g) \geq \psi_n(f_1)$. Equality is obtained if and only if $f_1 = f_2$ almost everywhere, but since f_1 and f_2 are *continuous relative* to C_n (Theorem 10.2 of Rockafellar (1997)), this implies that $f_1 = f_2$. □

Proof of Theorem 3

For $t \in (0, 1)$ and $y^{(1)}, y^{(2)} \in \mathbb{R}^n$, the function $\bar{h}_{ty^{(1)}+(1-t)y^{(2)}}$ is the least concave function satisfying

$\bar{h}_{ty^{(1)}+(1-t)y^{(2)}}(X_i) \geq ty_i^{(1)} + (1-t)y_i^{(2)}$ for $i = 1, \ldots, n$, so $\bar{h}_{ty^{(1)}+(1-t)y^{(2)}} \leq t\bar{h}_{y^{(1)}} + (1-t)\bar{h}_{y^{(2)}}$. The convexity of σ follows from this and the convexity of the exponential function. It is clear that $\sigma \geq \tau$, since $\bar{h}_y(X_i) \geq y_i$ for $i = 1, \ldots, n$.

From Theorem 2, we can find $y^* \in \mathbb{R}^n$ such that $\log \hat{f}_n = \bar{h}_{y^*}$ with $\bar{h}_{y^*}(X_i) = y_i^*$ for $i = 1, \ldots, n$, and this y^* minimises τ. For any other $y \in \mathbb{R}^n$ which minimises τ, by the uniqueness part of Theorem 2 we must have $\bar{h}_y = \bar{h}_{y^*}$, so $\sigma(y) > \sigma(y^*) = \tau(y^*)$. □

B. Structural and computational issues

As illustrated in Figure 1, and justified formally by Corollary 17.1.3 and Corollary 19.1.2 of Rockafellar (1997), the convex hull of the data, C_n, may be *triangulated* in such a way that $\log \hat{f}_n$ coincides with an *affine function* on each *simplex* in the triangulation. In other words, if $j = (j_0, \ldots, j_d)$ is a $(d+1)$-tuple of distinct indices in $\{1, \ldots, n\}$, and $C_{n,j} = \mathrm{conv}(X_{j_0}, \ldots, X_{j_d})$, then there exists a finite set J consisting of m such $(d+1)$-tuples, with the following three properties:

(i) $\cup_{j \in J} C_{n,j} = C_n$
(ii) the relative interiors of the sets $\{C_{n,j} : j \in J\}$ are pairwise disjoint
(iii)

$$\log \hat{f}_n(x) = \begin{cases} \langle x, b_j \rangle - \beta_j & \text{if } x \in C_{n,j} \text{ for some } j \in J \\ -\infty & \text{if } x \notin C_n \end{cases}$$

for some $b_1, \ldots, b_m \in \mathbb{R}^d$ and $\beta_1, \ldots, \beta_m \in \mathbb{R}$. Here and below, $\langle \cdot, \cdot \rangle$ denotes the usual Euclidean inner product in $\mathbb{R}^d$.

In the iterative algorithm that we propose for computing the maximum likelihood estimator, we need to find convex hulls and triangulations at each iteration. Fortunately, these can be computed efficiently using the `Quickhull` algorithm of Barber *et al.* (1996).

B.1. Computing the function σ

We now address the issue of computing the function σ in (3.2) at a generic point $y = (y_1, \ldots, y_n) \in \mathbb{R}^n$. For each $j = (j_0, \ldots, j_d) \in J$, let A_j be the $d \times d$ matrix whose lth column is $X_{j_l} - X_{j_0}$ for $l = 1, \ldots, d$, and let $\alpha_j = X_{j_0}$. Then the *affine transformation* $w \mapsto A_j w + \alpha_j$ takes the unit simplex $T_d = \{w = (w_1, \ldots, w_d) : w_l \geq 0, \sum_{l=1}^d w_l \leq 1\}$ to $C_{n,j}$.

Letting $z_{j,l} = y_{j_l} - y_{j_0}$ and $w_0 = 1 - w_1 - \ldots - w_d$, we can then establish by a simple change of variables and induction on d that if $z_{j,1},\ldots,z_{j,d}$ are non-zero and distinct, then

$$\int_{C_n} \exp\{\bar{h}_y(x)\}\,dx = \sum_{j\in J} |\det A_j| \int_{T_d} \exp(y_{j_0}w_0 + \ldots + y_{j_d}w_d)\,dw$$
$$= \sum_{j\in J} |\det A_j| e^{y_{j_0}} \sum_{r=1}^{d} \frac{e^{z_{j,r}}-1}{z_{j,r}} \prod_{\substack{1\le s\le d\\ s\neq r}} \frac{1}{z_{j,r}-z_{j,s}} \qquad \text{(B.1)}$$

The singularities that occur when some of $z_{j,1},\ldots,z_{j,d}$ may be zero or equal are removable. However, for stable computation of σ in practice, a Taylor approximation was used – see Cule and Dümbgen (2008); Cule (2009) for further details.

B.2. Non-differentiability of σ and computation of subgradients

In this section, we find explicitly the set of points at which the function σ defined in (3.2) is differentiable, and compute a subgradient of σ at each point. For $i = 1,\ldots,n$, define

$$J_i = \{j = (j_0,\ldots,j_d) \in J : i = j_l \text{ for some } l = 0,1,\ldots,d\}.$$

The set J_i is the index set of those simplices $C_{n,j}$ that have X_i as a vertex. Let $\mathcal{Y}$ denote the set of vectors $y = (y_1,\ldots,y_n) \in \mathbb{R}^n$ with the property that for each $j = (j_0,\ldots,j_d) \in J$, if $i \neq j_l$ for any l then

$$\big\{(X_i,y_i),(X_{j_0},y_{j_0}),\ldots,(X_{j_d},y_{j_d})\big\}$$

is affinely independent in $\mathbb{R}^{d+1}$. This is the set of points for which no tent pole is touching but not critically supporting the tent. Notice that the complement of $\mathcal{Y}$ has zero Lebesgue measure in $\mathbb{R}^n$, provided that every subset of $\{X_1,\ldots,X_n\}$ of size $d+1$ is *affinely independent* (an event of probability one). Let $w_0 = 1 - w_1 - \ldots - w_d$, and for $y \in \mathbb{R}^n$ and $i = 1,\ldots,n$, let

$$\partial_i(y) = -\frac{1}{n} + \sum_{j\in J_i} |\det A_j| \int_{T_d} e^{\langle w,z_j\rangle + y_{j_0}} \sum_{l=0}^{d} w_l \mathbb{1}_{\{j_l=i\}}\,dw. \qquad \text{(B.2)}$$

Proposition 5. *(a) For $y \in \mathcal{Y}$, the function σ is differentiable at y and for $i = 1,\ldots,n$ satisfies*

$$\frac{\partial\sigma}{\partial y_i}(y) = \partial_i(y).$$

(b) For $y \in \mathcal{Y}^c$, the function σ is not differentiable at y, but the vector $(\partial_1(y),\ldots,\partial_n(y))$ is a subgradient of σ at y.

Proof. By Theorem 25.2 of Rockafellar (1997), it suffices to show that for $y \in \mathcal{Y}$, all of the partial derivatives exist and are given by the expression in the statement of the proposition. For $i = 1,\ldots,n$ and $t \in \mathbb{R}$, let $y^{(t)} = y + te_i^n$, where e_i^n denotes the ith unit coordinate vector in $\mathbb{R}^n$. For sufficiently small values of $|t|$, we may write

$$\bar{h}_{y^{(t)}}(x) = \begin{cases} \langle x, b_j^{(t)}\rangle - \beta_j^{(t)} & \text{if } x \in C_{n,j} \text{ for some } j \in J \\ -\infty & \text{if } x \notin C_n, \end{cases}$$

for certain values of $b_1^{(t)}, \ldots, b_m^{(t)} \in \mathbb{R}^d$ and $\beta_1^{(t)}, \ldots, \beta_m^{(t)} \in \mathbb{R}$. If $j \notin J_i$, then $b_j^{(t)} = b_j$ and $\beta_j^{(t)} = \beta_j$ for sufficiently small $|t|$. On the other hand, if $j \in J_i$, then there are two cases to consider:

(i) If $j_0 = i$, then for sufficiently small t, we have $z_j^{(t)} = z_j - t1_d$, where 1_d denotes a d-vector of ones, so that $b_j^{(t)} = b_j - t(A_j^T)^{-1}1_d$ and $\beta_j^{(t)} = \beta_j - t(1 + \langle A_j^{-1}\alpha_j, 1_d\rangle)$

(ii) If $j_l = i$ for some $l \in \{1, \ldots, d\}$, then for sufficiently small t, we have $z_j^{(t)} = z_j + te_l^d$, so that $b_j^{(t)} = b_j + t(A_j^T)^{-1}e_l^d$ and $\beta_j^{(t)} = \beta_j + t\langle A_j^{-1}\alpha_j, e_l^d\rangle$.

It follows that

$$\begin{aligned}
\frac{\partial\sigma}{\partial y_i}(y) &= -\frac{1}{n} + \lim_{t\to 0}\frac{1}{t}\sum_{j\in J_i}\int_{C_{n,j}} \exp\bigl\{\langle x, b_j^{(t)}\rangle - \beta_j^{(t)}\bigr\} - \exp\{\langle x, b_j\rangle - \beta_j\}\,dx \\
&= -\frac{1}{n} + \lim_{t\to 0}\frac{1}{t}\sum_{j\in J_i}\Biggl[\int_{C_{n,j}} e^{\langle x,b_j\rangle-\beta_j}\Bigl\{e^{t(1-\langle A_j^{-1}(x-\alpha_j),1_d\rangle)} - 1\Bigr\}\,dx\mathbb{1}_{\{j_0=i\}} \\
&\qquad\qquad + \sum_{l=1}^{d}\int_{C_{n,j}} e^{\langle x,b_j\rangle-\beta_j}\Bigl\{e^{t\langle A_j^{-1}(x-\alpha_j),e_l^d\rangle} - 1\Bigr\}\,dx\mathbb{1}_{\{j_l=i\}}\Biggr] \\
&= -\frac{1}{n} + \sum_{j\in J_i}\Biggl[\int_{C_{n,j}} e^{\langle x,b_j\rangle-\beta_j}\bigl(1 - \langle A_j^{-1}(x-\alpha_j), 1_d\rangle\bigr)\,dx\mathbb{1}_{\{j_0=i\}} \\
&\qquad\qquad + \int_{C_{n,j}} e^{\langle x,b_j\rangle-\beta_j}\langle A_j^{-1}(x-\alpha_j), e_l^d\rangle\,dx\mathbb{1}_{\{j_l=i\}}\Biggr] \\
&= \partial_i(y),
\end{aligned}$$

where to obtain the final line we have made the substitution $x = A_j w + \alpha_j$, after taking the limit as $t \to 0$.

(b) If $y \in \mathcal{Y}^c$, then it can be shown that there exists a unit coordinate vector e_i^n in $\mathbb{R}^n$ such that the *one-sided directional derivative* at y with respect to e_i^n, denoted $\sigma'(y; e_i^n)$, satisfies $\sigma'(y; e_i^n) > -\sigma'(y; -e_i^n)$. Thus σ is not differentiable at y. To show that $\partial(y) = (\partial_1(y), \ldots, \partial_n(y))$ is a subgradient of σ at y, it is enough by Theorem 25.6 of Rockafellar (1997) to find, for each $\epsilon > 0$, a point $\tilde{y} \in \mathbb{R}^n$ such that $\|\tilde{y} - y\| < \epsilon$ and such that σ is differentiable at $\tilde{y}$ with $\|\nabla\sigma(\tilde{y}) - \partial(y)\| < \epsilon$. This can be done by sequentially making small adjustments to the components of y in the same order as that in which the vertices were *pushed* in constructing the triangulation. □

A subgradient of σ at any $y \in \mathbb{R}^n$ may be computed using Proposition 5 and (B.2) and once we have a formula for

$$\tilde{I}_{d,u}(z) = \int_{T_d} w_u \exp\biggl(\sum_{r=1}^{d} z_r w_r\biggr)\,dw.$$

An explicit closed formula for $\tilde{I}_{d,u}(z)$ where $z_1, \ldots, z_d$ are non-zero are distinct is derived in Cule, Samworth and Stewart (2010). Again, for practical purposes, we use a Taylor expansion for cases where $z_1, \ldots, z_d$ are close to zero or approximately equal. Details are given in Cule and Dümbgen (2008); Cule (2009).

B.3. Sampling from the fitted density estimate

In order to use the Monte Carlo procedure described in Section 7.1, we must be able to sample from $\hat{f}_n$. Fortunately, this can be done efficiently using the following rejection sampling procedure. As above, for $j \in J$ let A_j be the $d \times d$ matrix whose lth column is $X_{j_l} - X_{j_0}$ for $l = 1, \ldots, d$, and let $\alpha_j = X_{j_0}$, so that $w \mapsto A_j w + \alpha_j$ maps the unit simplex T_d to $C_{n,j}$. Recall that $\log \hat{f}_n(X_i) = y_i^*$, and let $z_j = (z_{j,1}, \ldots, z_{j,d})$, where $z_{j,l} = y_{j_l}^* - y_{j_0}^*$ for $l = 1, \ldots, d$. Write

$$q_j = \int_{C_{n,j}} \hat{f}_n(x)\, dx.$$

We may then draw an observation X^* from $\hat{f}_n$ as follows:

(i) Select $j^* \in J$, selecting $j^* = j$ with probability q_j

(ii) Select $w \sim \mathrm{Unif}(T_d)$ and $u \sim \mathrm{Unif}([0,1])$ independently. If

$$u < \frac{\exp(\langle w, z_{j^*}\rangle)}{\max_{v \in T_d} \exp(\langle v, z_{j^*}\rangle)},$$

accept the point and set $X^* = A_j w + \alpha_j$. Otherwise, repeat (ii).

C. Glossary of terms and results from convex analysis and computational geometry

All of the definitions and results below can be found in Rockafellar (1997) and Lee (2004). The *epigraph* of a function $f : \mathbb{R}^d \to [-\infty, \infty)$ is the set

$$\mathrm{epi}(f) = \{(x, \mu) : x \in \mathbb{R}^d, \mu \in \mathbb{R}, \mu \leq f(x)\}.$$

We say f is *concave* if its epigraph is non-empty and convex as a subset of $\mathbb{R}^{d+1}$; note that this agrees with the terminology of Barndorff-Nielsen (1978), but is what Rockafellar (1997) calls a *proper concave* function. If C is a convex subset of $\mathbb{R}^d$ then provided $f : C \to [-\infty, \infty)$ is not identically $-\infty$, it is *concave* if and only if

$$f\bigl(tx + (1-t)y\bigr) \geq tf(x) + (1-t)f(y)$$

for $x, y \in C$ and $t \in (0,1)$. A non-negative function f is *log-concave* if $\log f$ is concave, with the convention that $\log 0 = -\infty$. It is a *log-concave density* if it agrees almost everywhere with a log-concave function and $\int_{\mathbb{R}^d} f(x)\, dx = 1$. Note that all densities on $\mathbb{R}^d$ will be assumed to be with respect to Lebesgue measure on $\mathbb{R}^d$. The *support* of a log-concave function f is $\{x \in \mathbb{R}^d : \log f(x) > -\infty\}$, a convex subset of $\mathbb{R}^d$.

A subset M of $\mathbb{R}^d$ is *affine* if $tx + (1-t)y \in M$ for all $x, y \in M$ and $t \in \mathbb{R}$. The *affine hull* of M, denoted $\mathrm{aff}(M)$, is the smallest affine set containing M. Every non-empty affine set M in $\mathbb{R}^d$ is *parallel* to a unique subspace of $\mathbb{R}^d$, meaning that there is a unique subspace L of $\mathbb{R}^d$ such that $M = L + a$, for some $a \in \mathbb{R}^d$. The *dimension* of M is the dimension of this subspace, and more generally the dimension of a non-empty convex set is the dimension of its affine hull. A finite set of points $M = \{x_0, x_1, \ldots, x_d\}$ is *affinely independent* if $\mathrm{aff}(M)$ is d-dimensional. The *relative interior* of a convex set C is the interior which results when we regard C as a subset of its affine hull. The *relative boundary* of C is the set difference between its closure and its relative interior. If M is an affine set in $\mathbb{R}^d$, then an *affine transformation* (or *affine function*) is a function $T : M \to \mathbb{R}^d$ such that $T\bigl(tx + (1-t)y\bigr) = tT(x) + (1-t)T(y)$ for all $x, y \in M$ and $t \in \mathbb{R}$.

The *closure* of a concave function g on $\mathbb{R}^d$, denoted $\mathrm{cl}(g)$, is the function whose epigraph is the closure in $\mathbb{R}^{d+1}$ of $\mathrm{epi}(g)$. It is the least upper semi-continuous, concave function satisfying $\mathrm{cl}(g) \geq g$. The function g is *closed* if $\mathrm{cl}(g) = g$. An arbitrary function h on $\mathbb{R}^d$ is *continuous relative* to a subset S of $\mathbb{R}^d$ if its restriction to S is a continuous function. A non-zero vector $z \in \mathbb{R}^d$ is a *direction of increase* of h on $\mathbb{R}^d$ if $t \mapsto h(x + tz)$ is non-decreasing for every $x \in \mathbb{R}^d$.

The convex hull of finitely many points is called a *polytope*. The convex hull of $d + 1$ affinely independent points is called a *d-dimensional simplex* (pl. *simplices*). If C is a convex set in $\mathbb{R}^d$, then a *supporting half-space* to C is a closed half-space which contains C and has a point of C in its boundary. A *supporting hyperplane* H to C is a hyperplane which is the boundary of a supporting half-space to C. Thus $H = \{x \in \mathbb{R}^d : \langle x, b\rangle = \beta\}$, for some $b \in \mathbb{R}^d$ and $\beta \in \mathbb{R}$ such that $\langle x, b\rangle \leq \beta$ for all $x \in C$ with equality for at least one $x \in C$.

If V is a finite set of points in $\mathbb{R}^d$ such that $P = \mathrm{conv}(V)$ is a d-dimensional polytope in $\mathbb{R}^d$, then a *face* of P is a set of the form $P \cap H$, where H is a supporting hyperplane to P. The *vertex set* of P, denoted $\mathrm{vert}(P)$, is the set of 0-dimensional faces (*vertices*) of P. A *subdivision* of P is a finite set of d-dimensional polytopes $\{S_1, \ldots, S_t\}$ such that P is the union of $S_1, \ldots, S_t$ and the intersection of any two distinct polytopes in the subdivision is a face of both of them. If $S = \{S_1, \ldots, S_t\}$ and $\tilde{S} = \{\tilde{S}_1, \ldots, \tilde{S}_{t'}\}$ are two subdivisions of P, then $\tilde{S}$ is a *refinement* of S if each S_l is contained in some $\tilde{S}_{l'}$. The *trivial subdivision* of P is $\{P\}$. A *triangulation* of P is a subdivision of P in which each polytope is a simplex.

If P is a d-dimensional polytope in $\mathbb{R}^d$, F is a $(d-1)$-dimensional face of P and $v \in \mathbb{R}^d$, then there is a unique supporting hyperplane H to P containing F. The polytope P is contained in exactly one of the closed half-spaces determined by H, and if v is in the opposite open half-space, then F is *visible* from v. If V is a finite set in $\mathbb{R}^d$ such that $P = \mathrm{conv}(V)$, if $v \in V$ and $S = \{S_1, \ldots, S_t\}$ is a subdivision of P, then the result of *pushing* v is the subdivision $\tilde{S}$ of P obtained by modifying each $S_l \in S$ as follows:

(i) If $v \notin S_l$, then $S_l \in \tilde{S}$
(ii) If $v \in S_l$ and $\mathrm{conv}(\mathrm{vert}(S_l) \setminus \{v\})$ is $(d-1)$-dimensional, then $S_l \in \tilde{S}$
(iii) If $v \in S_l$ and $S_l' = \mathrm{conv}(\mathrm{vert}(S_l) \setminus \{v\})$ is d-dimensional, then $S_l' \in \tilde{S}$. Also, if F is any $(d-1)$-dimensional face of S_l' that is visible from v, then $\mathrm{conv}(F \cup \{v\}) \in \tilde{S}$.

If σ is a convex function on $\mathbb{R}^n$, then $y' \in \mathbb{R}^n$ is a *subgradient* of σ at y if

$$\sigma(z) \geq \sigma(y) + \langle y', z - y\rangle$$

for all $z \in \mathbb{R}^n$. If σ is differentiable at y, then $\nabla\sigma(y)$ is the unique subgradient to σ at y; otherwise the set of subgradients at y has more than one element. The *one-sided directional derivative* of σ at

y with respect to $z \in \mathbb{R}^n$ is

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

which always exists (allowing $-\infty$ and ∞ as limits) provided $\sigma(y)$ is finite.

Acknowledgments

The authors would like to thank the anonymous referees for their many helpful comments, which have greatly helped to improve the manuscript.

References

Abramson, I. (1982) On variable bandwidth in kernel estimates – a square root law. *Ann. Statist.*, **10**, 1217–1223.

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

Asuncion, A. and Newman, D. J. (2007) UCI Machine Learning Repository. URL `http://www.ics.uci.edu/~mlearn/MLRepository.html`.

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

Balabdaoui, F., Rufibach, K. and Wellner, J. A. (2009) Limit distribution theory for maximum likelihood estimation of a log-concave density. *Ann. Statist.*, **37**, 1299–1331.

Barber, C. B., Dobkin, D. P. and Huhdanpaa, H. (1996) The **Quickhull** algorithm for convex hulls. *ACM Trans. Math. Software*, **22**, 469–483. URL `http://www.qhull.org/`.

Barndorff-Nielsen, O. (1978) *Information and exponential families in statistical theory.* New Jersey: Wiley.

Boyd, S. and Vandenberghe, L. (2004) *Convex optimization.* Cambridge: Cambridge University Press.

Bozdogan, H. (1994) Choosing the number of clusters, subset selection of variables, and outlier detection on the standard mixture-model cluster analysis. In *New Approaches in Classification and Data Analysis* (eds. E. Diday, Y. Lechevallier, M. Schader, P. Bertrand and B. Burtschy), 169–177. New York: Springer-Verlag.

Breiman, L., Meisel, W. and Purcell, E. (1977) Variable kernel estimates of multivariate densities. *Technometrics*, **19**, 135–144.

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

Caplin, A. and Naelbuff, B. (1991a) Aggregation and imperfect competition: On the existence of equilibrium. *Econometrica*, 25–59.

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

Chacón, J. E. (2009) Data-driven choice of the smoothing parametrization for kernel density estimators. *Canad. J. Statist.*, **34**, 249–265.

Chacón, J. E., Duong, T. and Wand, M. P. (2008) Asymptotics for general multivariate kernel density derivative estimators. Preprint.

Chang, G. and Walther, G. (2007) Clustering with mixtures of log-concave distributions. *Computational Statistics and Data Analysis*, **51**, 6242–6251.

Chiu, S.-T. (1992) An automatic bandwidth selector for kernel density estimation. *Biometrika*, **79**, 771–782.

Cule, M. L. (2009) *Maximum likelihood estimation of a multivariate log-concave density.* Ph.D. thesis, University of Cambridge.

Cule, M. L. and Dümbgen, L. (2008) On an auxiliary function for log-density estimation. Tech. Rep. 71, Universität Bern.

Cule, M. L., Gramacy, R. B. and Samworth, R. J. (2007) **LogConcDEAD***: Maximum Likelihood Estimation of a Log-Concave Density.* URL http://CRAN.R-project.org/package=LogConcDEAD. R package version 1.3-2.

Cule, M. L. and Samworth, R. J. (2010), Theoretical properties of the log-concave maximum likelihood estimator of a multidimensional density. *Electron. J. Statist.*, **4**, 254–270.

Cule, M. L., Samworth, R. J. and Stewart, M. I. (2010) Maximum likelihood estimation of a multidimensional log-concave density (long version). URL http://www.statslab.cam.ac.uk/~rjs57/Research.html.

Ćwik, J. and Koronacki, J. (1997) Multivariate density estimation: A comparative study. *Neural Computation and Applications*, **6**, 173–185.

Deheuvels, P. (1977) Estimation non parametrique de la densité par histogrammes generalisés II. *Publ. l'Inst. Statist. l'Univ Paris*, **22**, 1–23.

Dempster, A., Laird, N. and Rubin, D. (1977) Maximum likelihood from incomplete data via the EM algorithm. *J. Roy. Statist. Soc, Ser. B*, **39**, 1–38.

Devroye, L., Györfi, L. and Lugosi, G. (1996) *A Probabilistic Theory of Pattern Recognition.* New York: Springer-Verlag.

Donoho, D. L., Johnstone, I. M., Kerkyacharian, G. and Picard, D. (1996) Density estimation by wavelet thresholding. *Ann. Statist.*, **24**, 508–539.

Dümbgen, L., Hüsler, A. and Rufibach, K. (2007) Active set and EM algorithms for log-concave densities based on complete and censored data. Tech. rep. 61, Universität Bern. URL http://arxiv.org/abs/0707.4643/.

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

Dümbgen, L., Samworth, R. J. and Schuhmacher, D. (2010) Approximation by log-concave distributions with applications to regression. Tech. rep. 75, Universität Bern. URL `http://arxiv.org/abs/1002.3448/`

Duong, T. (2004) *Bandwidth selectors for multivariate kernel density estimation.* Ph.D. thesis, University of Western Australia.

Duong, T. (2007) **ks***: Kernel smoothing.* URL `http://cran.r-project.org/package=ks`. R package version 1.56.

Duong, T. and Hazelton, M. L. (2003) Plug-in bandwidth matrices for bivariate kernel density estimation. *J. Nonparam. Statist.*, **15**, 17–30.

Duong, T. and Hazelton, M. L. (2005) Convergence rates for unconstrained bandwith matrix selectors in multivariate kernel density estimation. *J. Mult. Anal.*, **93**, 417–433.

Eggermont, P. P. B. and LaRiccia, V. (2001) *Maximum penalized likelihood estimation.* Vol. 1: Density estimation. New York: Springer-Verlag.

Eubank, R. L. (1988) *Spline smoothing and nonparametric regression.* New York: Marcel Dekker.

Fix, E. and Hodges, J. L. (1951) Discriminatory analysis – nonparametric discrimination: Consistency properties. Tech. Rep. 4, Project no. 21-29-004, USAF School of Aviation Medicine, Randolph Field, Texas.

Fix, E. and Hodges, J. L. (1989) Discriminatory analysis – nonparametric discrimination: Consistency properties. *Internat. Statist. Rev.*, **57**, 238–247.

Fraley, C. F. and Raftery, A. E. (2002) Model-based clustering, discriminant analysis, and density estimation. *J. Amer. Statist. Assoc.*, **97**, 611–631.

Gordon, A. D. (1981) *Classification.* London: Chapman and Hall.

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

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

Groeneboom, P., Jongbloed, G. and Wellner, J. A. (2008) The support reduction algorithm for computing nonparametric function estimates in mixture models. *Scand. J. Statist.*, **35**, 385–399.

Groeneboom, P. and Wellner, J. A. (1992) *Information Bounds and Nonparametric Maximum Likelihood Estimation.* Basel: Birkhäuser.

Hall, P., Marron, J. S. and Park, B. U. (1992) Smoothed cross-validation. *Probab. Theory Related Fields*, **92**, 1–20.

Hall, P., Park, B. U. and Samworth, R. J. (2008) Choice of neighbour order in nearest-neighbour classification. *Ann. Statist.*, **36**, 2135–2152.

Hand, D. J. (1981) *Discrimination and Classification.* New York: Wiley.

Hyndman, R. J. (1996) Computing and graphing highest density regions. *The American Statistician*, **50**, 120–126.

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

Jongbloed, G. (1998) The iterative convex minorant algorithm for nonparametric estimaton. *J. Computational and Graphical Statist.*, **7**, 310–321.

Kappel, F. and Kuntsevich, A. (2000) An implementation of Shor's r-algorithm. *Computational Optimization and Applications*, **15**, 193–205.

Koenker, R. and Mizera, I. (2010) Quasi-concave density estimation *Ann. Statist.*, to appear.

Lee, C. W. (2004) Subdivisions and triangulations of polytopes. In *Handbook of discrete and computational geometry* (eds. J. E. Goodman and J. O'Rourke), second edition, 383–406. New York: CRC Press.

McLachlan, G. J. and Basford, K. E. (1988) *Mixture Models: Inference and Applications to Clustering.* New York: Marcel Dekker.

McLachlan, G. J. and Krishnan, T. (1997) *The EM Algorithm and Extensions.* New York: Wiley.

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

Pal, J. K., Woodroofe, M. and Meyer, M. (2007) Estimating a Polya frequency function. In *Complex Datasets and Inverse Problems: Tomography, Networks and Beyond.* Vol. 54 of *Lecture Notes - Monograph Series*, pp. 239–249. Ohio: Institute of Mathematical Statistics.

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

Prékopa, A. (1973) On logarithmically concave measures and functions. *Acta Scientarium Mathematicarum*, **34**, 335–343.

R Development Core Team (2009) *R: A Language and Environment for Statistical Computing.* R Foundation for Statistical Computing, Vienna, Austria. URL `http://www.R-project.org/`. ISBN 3-900051-07-0.

Rockafellar, R. T. (1997) *Convex analysis.* Princeton, New Jersey: Princeton University Press.

Rosenblatt, M. (1956) Remarks on some nonparametric estimates of a density function. *Ann. Math. Statist.*, **27**, 832–837.

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

Rufibach, K. and Dümbgen, L. (2006) **logcondens***: Estimate a Log-Concave Probability Density from i.i.d. Observations.* URL `http://CRAN.R-project.org/package=logcondens`. R package version 1.3.2.

Sain, S. R. (2002) Multivariate locally adaptive density estimation. *Computational Statistics and Data Analysis*, **39**, 165–186.

Sain, S. R. and Scott, D. W. (1996) On locally adaptive density estimation. *J. Amer. Statist. Assoc.*, **91**, 1525–1534.

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

Schuhmacher, D., Hüsler, A. and Dümbgen, L. (2010) Multivariate log-concave distributions as a nearly parametric model. Tech. rep. 74, Universität Bern. URL `http://arxiv.org/pdf/0907.0250v2`

Scott, D. W. and Sain, S. R. (2004) Multi-dimensional density estimation. In *Handbook of statistics* (eds. C. R. Rao and E. J. Wegman), vol. 23: Data mining and computational statistics. Amsterdam: Elsevier.

Seregin, A. and Wellner, J. A. (2009) Nonparametric estimation of convex-transformed densities URL `http://arxiv.org/abs/0911.4151`

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

Street, W. M., Wolberg, W. H. and Mangasarian, O. L. (1993) Nuclear feature extraction for breast tumor diagnosis. *IS & T/SPIE International Symposium on Electronic Imaging: Science and Technology*, **1905**, 861–870.

Swales, J. D., ed. (1985) *Platt Vs. Pickering: An Episode in Recent Medical History.* Cambridge: The Keynes Press.

Titterington, D. M., Smith, A. F. M. and Makov, U. E. (1985) *Statistical Analysis of Finite Mixture Distributions.* Chichester: Wiley.

Vapnik, V. N. and Mukherjee, S. (2000) Support vector method for multivariate density estimation. In *Advances in Neural Information Processing Systems*, 659–665. Cambridge, MA: MIT press.

Wahba, G. (1990) *Spline models for observational data.* Philadelphia: Society for Industrial and Applied Mathematics.

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

Walther, G. (2010) Inference and modeling with log-concave distributions. *Statist. Science*, to appear.

Wand, M. P. and Jones, M. C. (1995) *Kernel smoothing.* CRC Press, Florida: Chapman and Hall.

Zhang, X., King, M. L. and Hyndman, R. J. (2006) Bandwidth selection for multivariate kernel density estimation using MCMC. *Computational Statistics and Data Analysis*, **50**, 3009–3031.

Further technical arguments

We first establish (B.1). We have

$$\begin{aligned}\int_{C_n} \exp\{\bar{h}_y(x)\}\,dx &= \sum_{j\in J} |\det A_j| e^{-\beta_j} \int_{T_d} \exp\{\langle A_j w + \alpha_j, b_j\rangle\}\,dw \\ &= \sum_{j\in J} |\det A_j| e^{y_{j_1}} \int_{T_d} \exp\biggl(\sum_{r=1}^{d} z_{j,r} w_r\biggr)\,dw.\end{aligned}$$

For ease of notation, we drop the j subscript, and consider

$$I_d(z) = \int_{T_d} \exp\biggl(\sum_{r=1}^{d} z_r w_r\biggr)\,dw.$$

Observe that $I_1(z_1) = z_1^{-1}(e^{z_1} - 1)$, in agreement with (B.1). Assume the result for $d-1$ as an inductive hypothesis. Then

$$\begin{aligned}I_d(z) &= \frac{1}{z_d} e^{z_d} \int_{T_{d-1}} \exp\biggl(\sum_{r=1}^{d-1} (z_r - z_d) w_r\biggr)\,dw - \frac{1}{z_d} \int_{T_{d-1}} \exp\biggl(\sum_{r=1}^{d-1} z_r w_r\biggr)\,dw \\ &= \frac{1}{z_d} e^{z_d} I_{d-1}(z_1 - z_d, \ldots, z_{d-1} - z_d) - \frac{1}{z_d} I_{d-1}(z_1, \ldots, z_d) \\ &= \frac{1}{z_d} \sum_{r=1}^{d-1} \frac{e^{z_r} - e^{z_d}}{(z_r - z_d)} \prod_{\substack{1\le s\le d-1\\ s\ne r}} \frac{1}{(z_r - z_s)} - \frac{1}{z_d} \sum_{r=1}^{d-1} \frac{e^{z_r} - 1}{z_r} \prod_{\substack{1\le s\le d-1\\ s\ne r}} \frac{1}{(z_r - z_s)} \\ &= \sum_{r=1}^{d-1} \frac{e^{z_r}}{z_r} \prod_{\substack{1\le s\le d\\ s\ne r}} \frac{1}{(z_r - z_s)} - \frac{e^{z_d}}{z_d} \sum_{r=1}^{d-1} \prod_{\substack{1\le s\le d\\ s\ne r}} \frac{1}{(z_r - z_s)} + \frac{1}{z_d} \sum_{r=1}^{d-1} \frac{1}{z_r} \prod_{\substack{1\le s\le d-1\\ s\ne r}} \frac{1}{(z_r - z_s)}. \qquad \text{(C.1)}\end{aligned}$$

To deal with the middle term in the last line of (C.1) above, define a polynomial

$$P_d(t) = \sum_{r=1}^{d} \prod_{\substack{1\le s\le d\\ s\ne r}} \Bigl(\frac{t - z_s}{z_r - z_s}\Bigr) - 1.$$

This polynomial $P_d(t)$ is of degree at most $d-1$, but has roots $z_1, \ldots, z_d$, so is identically zero. Examining the coefficient of t^{d-1} in this polynomial, we find

$$\sum_{r=1}^{d} \prod_{\substack{1\le s\le d\\ s\ne r}} \frac{1}{(z_r - z_s)} = 0.$$

To deal with the final term in (C.1), observe that

$$\frac{1}{z_d}\sum_{r=1}^{d-1}\frac{1}{z_r}\prod_{\substack{1\leq s\leq d-1\\ s\neq r}}\frac{1}{(z_r-z_s)} = \frac{1}{z_1z_2\ldots z_d}\sum_{r=1}^{d-1}\prod_{\substack{1\leq s\leq d-1\\ s\neq r}}\frac{z_s}{(z_r-z_s)} = \frac{(-1)^d}{z_1z_2\ldots z_d}\{P_{d-1}(0)+1\}$$

$$= \frac{(-1)^d}{z_1z_2\ldots z_d} = -\sum_{r=1}^{d}\frac{1}{z_r}\prod_{\substack{1\leq s\leq d\\ s\neq r}}\frac{1}{(z_r-z_s)}.$$

Substituting these expressions into (C.1) yields (B.1). □

Our final task is to establish that

$$\tilde{I}_{d,u}(z) = \sum_{\substack{1\leq r\leq d\\ r\neq u}}\frac{e^{z_r}}{z_r(z_r-z_u)}\prod_{\substack{1\leq s\leq d\\ s\neq r}}\frac{1}{(z_r-z_s)} - \sum_{\substack{1\leq r\leq d\\ r\neq u}}\frac{e^{z_u}}{z_r(z_r-z_u)}\prod_{\substack{1\leq s\leq d\\ s\neq r}}\frac{1}{(z_r-z_s)} + \frac{(-1)^d(e^{z_u}-1)}{z_u\prod_{s=1}^d z_s} + \frac{e^{z_u}}{z_u}\prod_{\substack{1\leq s\leq d\\ s\neq u}}\frac{1}{(z_u-z_s)}. \quad (C.2)$$

To this end, observe that for $u=1$, we have the recurrence relation

$$\tilde{I}_{d,1}(z) = \sum_{r=0}^{R}\frac{e^{z_{d-k}}\tilde{I}_{d-R-1,1}(z_1-z_{d-r},\ldots,z_{d-R-1}-z_{d-r})}{z_{d-r}\prod_{\substack{1\leq s\leq R\\ s\neq r}}(z_{d-r}-z_{d-s})} + \frac{(-1)^{R+1}}{\prod_{r=0}^R z_{d-r}}\tilde{I}_{d-R-1,1}(z_1,\ldots,z_{d-R-1}),$$

which holds for $R=0,\ldots,d-2$, and may be proved by induction on R. The formulae for other values of u may be deduced by symmetry. The formula for $\tilde{I}_{d,u}(z)$ in (C.2) is found by using this expression with $R=d-2$ together with the fact that for $z\neq 0$,

$$\int_0^1 w\exp(zw)\,dw = \frac{1}{z}\Big(1-\frac{1}{z}\Big)e^z + \frac{1}{z^2}.$$

□