Journal of the Royal Statistical Society
SERIES B Statistical Methodology

J. R. Statist. Soc. B (2010)
72, *Part* 5, *pp.* 545–607

Maximum likelihood estimation of a multi-dimensional log-concave density

Madeleine Cule and Richard Samworth

University of Cambridge, UK

and Michael Stewart

University of Sydney, Australia

[*Read before* The Royal Statistical Society *at a meeting organized by the* Research Section *on Wednesday, May 12th, 2010*, Professor D. M. Titterington *in the Chair*]

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 1, there is 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 can reformulate the issue of computing $\hat{f}_n$ in terms of a non-differentiable convex optimization 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–maximization algorithm to fit finite mixtures of log-concave densities.

Keywords: Computational geometry; Log-concavity; Maximum likelihood estimation; Non-differentiable convex optimization; Non-parametric density estimation; Shor's r-algorithm

1. Introduction

Modern non-parametric density estimation began with the introduction of a kernel density estimator in the pioneering work of Fix and Hodges (1951), which was 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 (MISE) 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 (see Wand and Jones (1995), chapter 3, and the references therein). Although 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.

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

 1369–7412/10/72545

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 that are 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 (e.g. Duong and Hazelton (2003, 2005), Zhang *et al.* (2006), Chacón *et al.* (2010) and Chacón (2009)) significant practical challenges remain.

Extensions that adapt to local smoothness began with Breiman *et al.* (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, e.g. 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 non-parametric estimator of f, with no tuning parameters to be chosen, under the condition that f is log-concave—i.e. $\log(f)$ 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 1 there is a unique log-concave density $\hat{f}_n$ that maximizes 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 non-parametric 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 *et al.* (2007) have proved the existence and uniqueness of the log-concave maximum likelihood estimator in one dimension and Dümbgen and Rufibach (2009), Pal *et al.* (2007) and Balabdaoui *et al.* (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 *et al.* (2007) also presented an active set algorithm, which has similarities with the vertex direction and vertex reduction algorithms

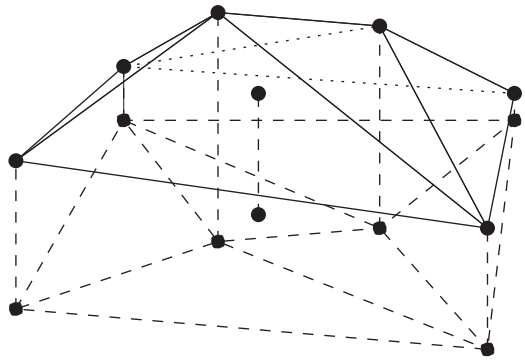


Fig. 1. 'Tent-like' structure of the graph of the logarithm of the maximum likelihood estimator for bivariate data

that were described in Groeneboom *et al.* (2008). Walther (2009) provides a nice recent review of inference and modelling with log-concave densities. Other recent related work includes Seregin and Wellner (2010), Schuhmacher *et al.* (2009), Schuhmacher and Dümbgen (2010) and Koenker and Mizera (2010). For univariate data, it is also well known that there are maximum likelihood estimators of a non-increasing density supported on $[0, \infty)$ (Grenander, 1956) and of a convex, decreasing density (Groeneboom *et al.*, 2001).

Fig. 1 gives a diagram illustrating the structure of the maximum likelihood estimator on the logarithmic scale. This structure is most easily visualized for two-dimensional data, where we 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, which is discussed in Section 3.1, is that it is possible to minimize a modified objective function that is convex (though non-differentiable). This allows us to apply the powerful non-differentiable convex optimization 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, Fig. 2 presents plots of the maximum likelihood estimator, and its logarithm, for 1000 observations from a standard bivariate normal distribution. These plots were created using the LogConcDEAD package (Cule *et al.*, 2007, 2009) 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 that is 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 minimizing the Kullback–Leibler divergence). This latter result amounts to a desirable robustness property.

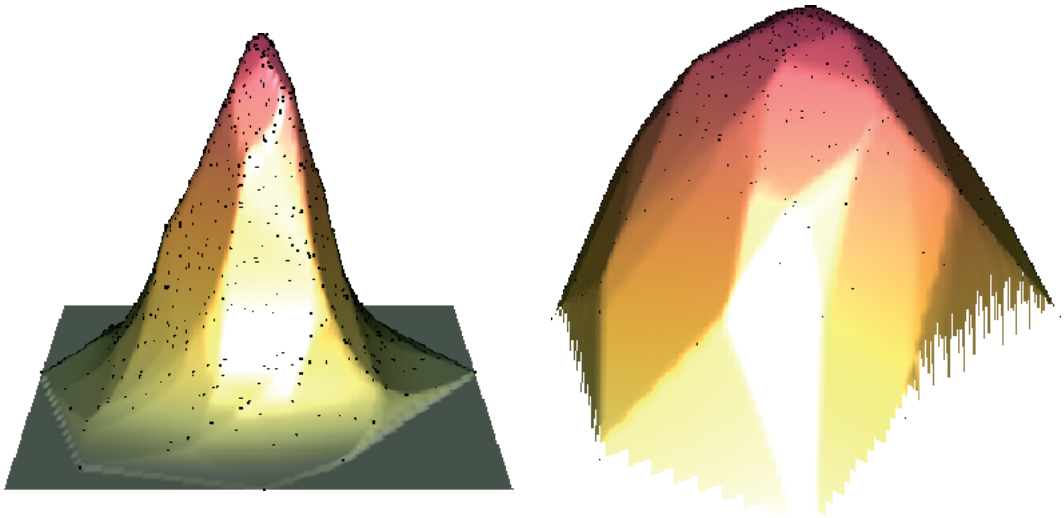


Fig. 2. Log-concave maximum likelihood estimates based on 1000 observations (•) from a standard bivariate normal distribution: (a) density; (b) log-density

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 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 MISE 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.

Non-parametric density estimation is a fundamental tool for the visualization 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 as follows.

(a) In classification problems, we have $p \geqslant 2$ populations of interest, and we 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 and ecology—see Gordon (1981), Hand (1981) or Devroye *et al.* (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, and 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 that were described above. The difference is that, in the above notation, we do not observe $Y_1, \ldots, Y_n$, and must 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 1. 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—see 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 $\int_{\|x\| \geqslant 1} f(x)\,\mathrm{d}x$, moments, e.g. $\int \|x\|^2 f(x)\,\mathrm{d}x$, and the differential entropy, $-\int f(x)\log\{f(x)\}\,\mathrm{d}x$. 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 non-parametric 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 suggests a model that exploits this elliptical symmetry.

(e) In the univariate case, Walther (2002) described methodology based on log-concave density estimation for addressing the problem of detecting the presence of mixing in a distribution. As an application, he cited 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 that is attached to people in the right-hand tail of the blood pressure distribution. As a result of our algorithm for computing the multi-dimensional log-concave maximum likelihood estimator, a similar test may be 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 1, beta(α, β) distributions with $\alpha, \beta \geqslant 1$, Weibull distributions with shape parameter at least 1, 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 log-normal 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 $\|\mu\| \leqslant 2$.

The assumption of log-concavity is popular in economics; Caplin and Nalebuff (1991a) showed that, in the theory of elections and under a log-concavity assumption, the proposal that is most preferred by the mean voter is unbeatable under a 64% majority rule. As another example, in the theory of imperfect competition, Caplin and Nalebuff (1991b) used 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) listed many useful properties of log-concave densities. For instance, if f and g are (possibly multi-dimensional) 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 characterization: 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 generalization of this result to higher dimensions.

As was mentioned in Section 1, this paper concerns multi-dimensional 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 so that f is 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 and
(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 multi-dimensional 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 that are 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 that 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 generalization 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 $n \geqslant d+1$. 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 maximizes $l(f) = \sum_{i=1}^n \log\{f(X_i)\}$ over $f \in \mathcal{F}_0$.

Theorem 1. With probability 1, a log-concave maximum likelihood estimator $\hat{f}_n$ of f_0 exists and is unique.

During the course of the proof of theorem 1, 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 $\bar{h}_y(X_i) \geqslant y_i$ for all $i = 1, \ldots, n$. A typical example of a tent function is depicted in Fig. 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 1 gives no indication of how to find the member of this class (in other words, the $y \in \mathbb{R}^n$) that maximizes the likelihood function. We therefore seek an iterative algorithm to compute the estimator.

3.1. Reformulation of the optimization problem

As a first attempt to find an algorithm which produces a sequence that converges to the maximum likelihood estimator in theorem 1, it is natural to try to minimize numerically the function

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

The first term on the right-hand side of equation (3.1) represents the (normalized) negative log-likelihood of a tent function, whereas the second term can be thought of as a Lagrangian term, which allows us to minimize 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 minimize 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:

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

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

Theorem 2. The function σ is a convex function satisfying $\sigma \geqslant \tau$. It has a unique minimum at $y^* \in \mathbb{R}^n$, say, and $\log(\hat{f}_n) = \bar{h}_{y^*}$.

Thus theorem 2 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 the tent poles touching the tent'.

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 equation (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 Fig. 1. The structure of the estimator and the issue of computing σ are described in greater detail in Appendix B.

3.2. Non-smooth optimization

There is a vast literature on techniques of convex optimization (see 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 Fig. 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 equation (3.2); in contrast, increasing the height of the tent pole does alter the tent function and therefore the integral in equation (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 optimization 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 optimization 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 minimizing non-differentiable, convex functions defined on Euclidean spaces. A slightly specialized 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 that is generated by the formula

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

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

Shor recognized, 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 non-orthogonal) linear transformation, and to 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 non-smooth 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 minimizer. 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 Kappel and Kuntsevich's (2000) 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

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

d	*Running times for the following values of n:*				
	$n=100$	$n=200$	$n=500$	$n=1000$	$n=2000$
2	1.5 s (260)	2.9 s (500)	50 s (1270)	4 min (2540)	24 min (5370)
3	6 s (170)	12 s (370)	100 s (820)	7 min (1530)	44 min (2740)
4	23 s (135)	52 s (245)	670 s (600)	37 min (1100)	224 min (2060)

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

(a) $|\sigma(y^{(l+1)}) - \sigma(y^{(l)})| \leqslant \delta$,
(b) $|y_i^{(l+1)} - y_i^{(l)}| \leqslant \varepsilon$ for $i = 1, \ldots, n$ and
(c) $|1 - \int \exp\{\bar{h}_{y^{(l)}}(x)\}\,\mathrm{d}x| \leqslant \eta$

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

Table 1 gives sample running times and the approximate number of iterations of Shor's r-algorithm that are required for different sample sizes and dimensions on an ordinary desktop computer (1.8 GHz, 2 GBytes random-access memory). Unsurprisingly, the running time increases relatively quickly with the sample size, whereas 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 that are required for the algorithm to terminate decreases as the dimension increases.

When $d = 1$, we recommend the active set algorithm of Dümbgen *et al.* (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

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

can be characterized 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 3 below we present the main result from that paper. See also Schuhmacher and Dümbgen (2010) and Dümbgen *et al.* (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

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

It is a simple consequence of Jensen's inequality that the Kullback–Leibler divergence $d_{\mathrm{KL}}(f_0, f)$ is always non-negative. The first part of theorem 3 asserts under very weak conditions the existence and uniqueness of a log-concave density f^* that minimizes 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 minimized (in fact, made to equal 0) by choosing $f^* = f_0$. The second part of the theorem then gives that, with probability 1, 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 we 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 minimizing the Kullback–Leibler divergence. This establishes a desirable robustness property for $\hat{f}_n$, with the natural practical interpretation that, provided that f_0 is not too far from being log-concave, the estimator is still sensible.

To introduce the notation that is used in the theorem, we write E for the support of f_0, i.e. 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 3. Let f_0 be any density on $\mathbb{R}^d$ with $\int_{\mathbb{R}^d} \|x\| f_0(x)\,\mathrm{d}x < \infty$, $\int_{\mathbb{R}^d} f_0 \log_+(f_0) < \infty$ and $\mathrm{int}(E) \neq \emptyset$. There is a log-concave density f^*, unique almost everywhere, that minimizes 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 $f^*(x) \leqslant \exp(-a_0\|x\| + b_0)$, we have for any $a < a_0$ that

$$\int_{\mathbb{R}^d} \exp(a\|x\|)\,|\hat{f}_n(x) - f^*(x)|\,\mathrm{d}x \to 0 \qquad \text{almost surely}$$

as $n \to \infty$, and, if f^* is continuous, $\sup_{x \in \mathbb{R}^d} \exp(a\|x\|)\,|\hat{f}_n(x) - f^*(x)| \to 0$ almost surely 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 $f^*(x) \leqslant \exp(-a_0\|x\| + b_0)$, 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

and the normal location mixture $0.6\,\phi_d(\cdot) + 0.4\,\phi_d(\cdot - \mu)$ for

(d) $\|\mu\| = 1$,
(e) $\|\mu\| = 2$ and
(f) $\|\mu\| = 3$.

An application of proposition 1 tells us that such a normal location mixture is log-concave if and only if $\|\mu\| \leqslant 2$.

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

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

(i) the theoretically optimal bandwidth, computed by minimizing 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 *et al.*, 1992; Duong, 2004),
(iv) a two-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) and
(vi) Sain's method (Sain, 2002; Scott and Sain, 2004). This divides the sample space 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 that is described in Section 6 to fit a mixture of two log-concave components. Further examples and implementational details can be found in Cule (2009).

Table 2. Summary of features of the example densities†

Density	*Log-concave*	*Dependent*	*Normal*	*Mixture*	*Skewed*	*Bounded*
(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

†Log-concave, log-concave density; dependent, components are dependent; normal, mixture of one or more Gaussian components; mixture, mixture of log-concave distributions; skewed, non-zero skewness; bounded, support of the density is bounded in one or more directions.

Results are given in Fig. 3 and Fig. 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 performed 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 that are 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 the choice of bandwidth, for moderate or large sample sizes. Remarkably, our estimator outperforms the kernel estimator even when the bandwidth is chosen on the basis of knowledge of the true density to minimize 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 cannot 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 be 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 that is described in theorem 3. In this case we may recover good performance at larger sample sizes by using a mixture of two log-concave components.

To investigate the effect of boundary effects further, we performed the same simulations for a bivariate density with independent components having a Unif(0,1) distribution and a beta(2,4) distribution. The results are shown in Fig. 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

Recently, Chang and Walther (2007) introduced an algorithm which combines the univariate log-concave maximum likelihood estimator with the EM algorithm (Dempster *et al.*, 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 1, and the component densities $f_1, \ldots, f_p$ are univariate and log-concave. The method is an extension of the standard Gaussian EM 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 $j = \arg\max_r \{\hat{\pi}_r \hat{f}_r(X_i)\}$. Chang and Walther (2007) showed empirically that, in cases where the true component densities are

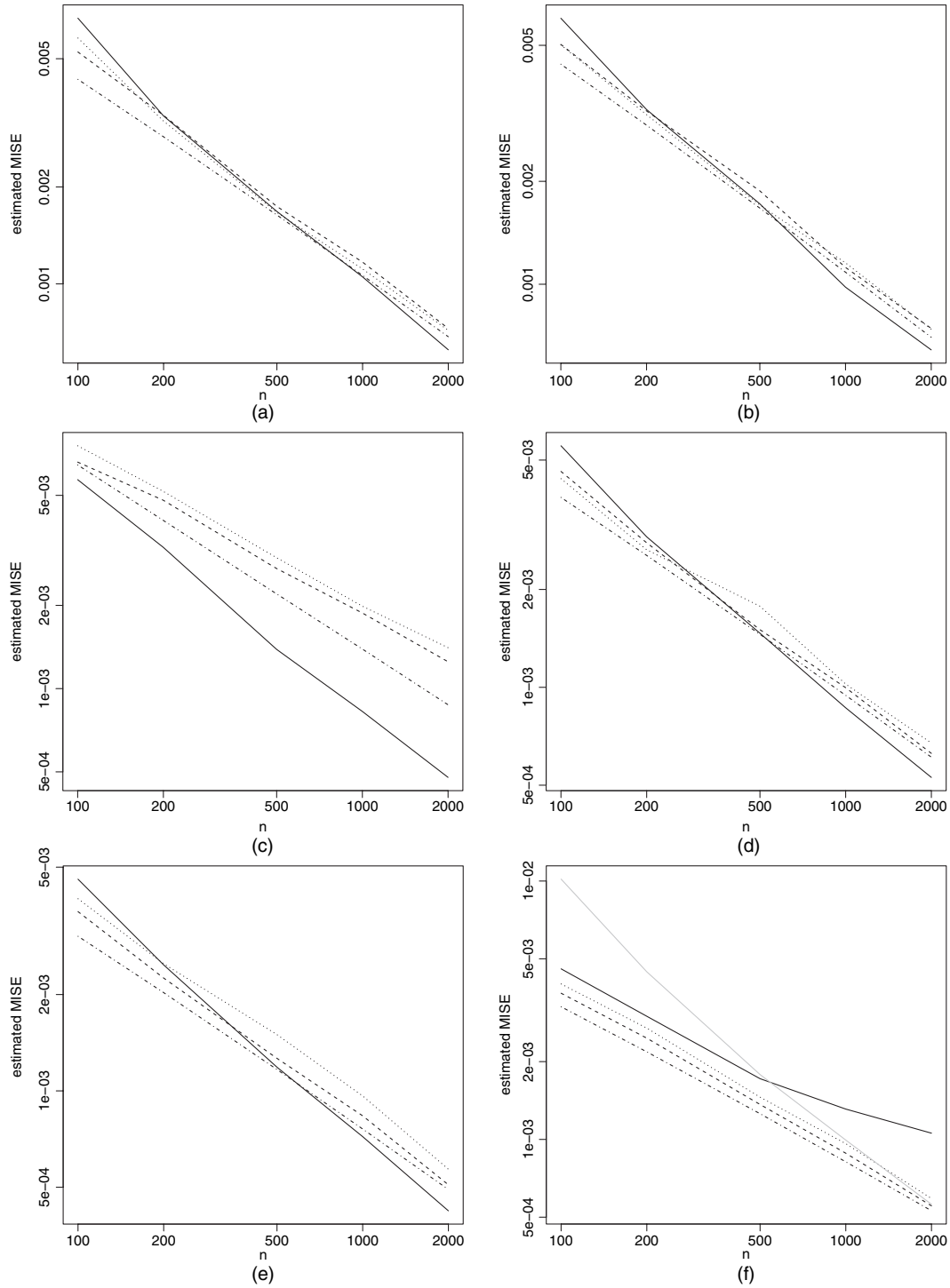


Fig. 3. MISE, $d = 2$: ———, LogConcDEAD estimate; -------, plug-in kernel estimate; · · · · · · ·, Abramson kernel estimate; · – · – ·, MISE optimal bandwidth kernel estimate; ———, ((f) only) two-component log-concave mixture

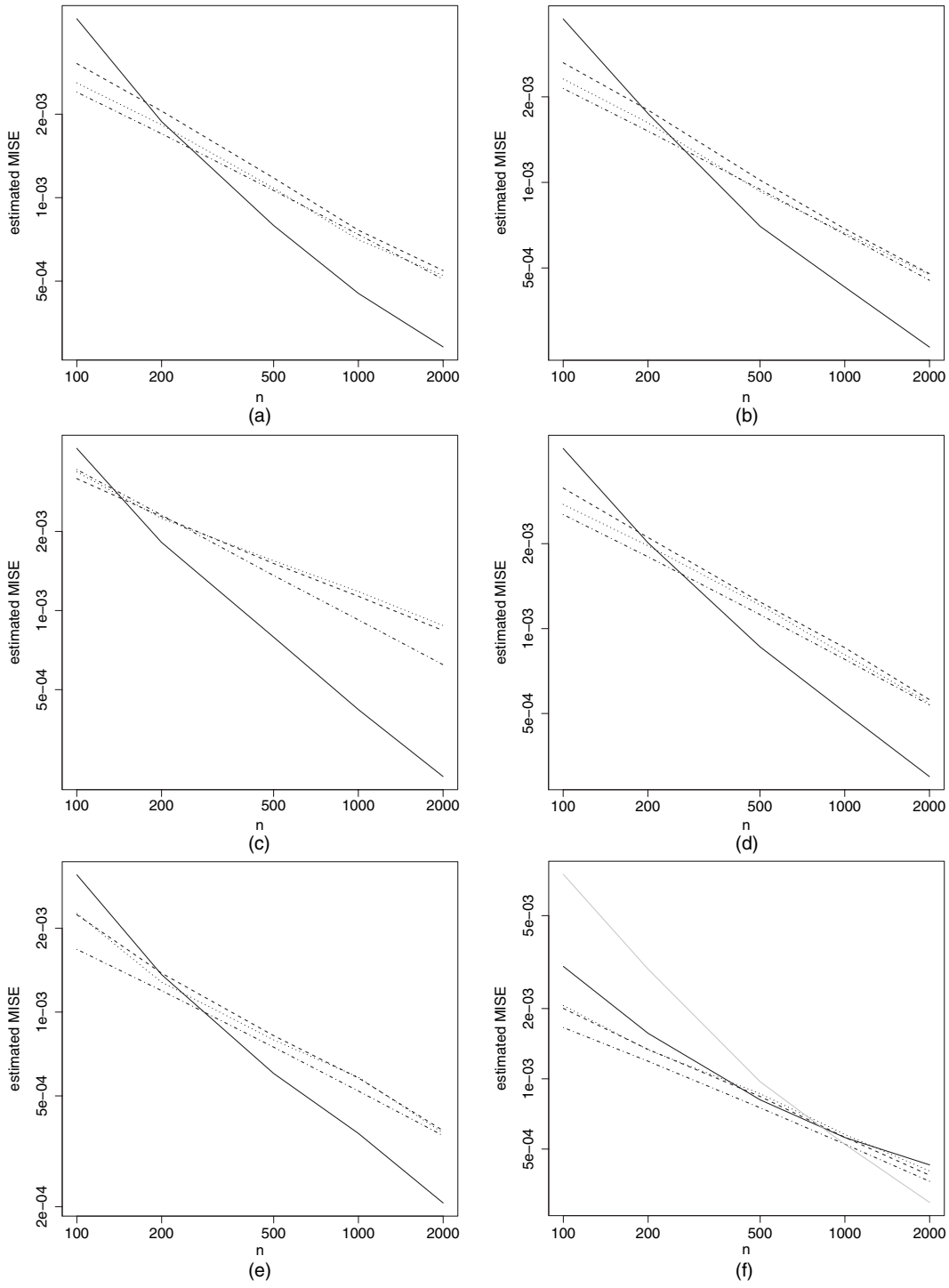


Fig. 4. MISE, $d = 3$: ———, LogConcDEAD estimate; -------, plug-in kernel estimate; ·····, Abramson kernel estimate; · – · – ·, MISE optimal bandwidth kernel estimate; ———, ((f) only) two-component log-concave mixture

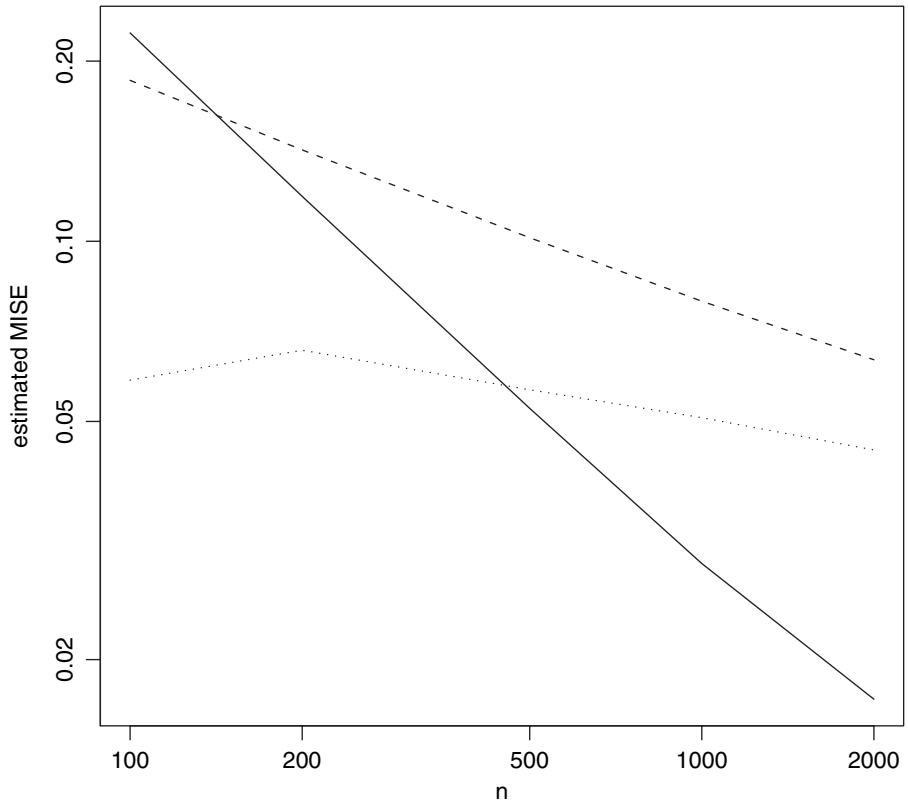


Fig. 5. MISE, $d = 2$, bivariate uniform and beta density: ———, LogConcDEAD estimate; -------, plug-in kernel estimate; · · · · · · ·, Abramson kernel estimate

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 multi-dimensional log-concave density, Chang and Walther (2007) discussed an extension of model (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 can compute the maximum likelihood estimator of a multi-dimensional log-concave density, we can carry this method through to its natural conclusion, i.e., in the finite mixture model (6.1) for a multi-dimensional 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 *et al.* (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 $\hat{\pi}_1^{(l)}, \ldots, \hat{\pi}_p^{(l)}$,

$\hat{f}_1^{(l)}, \ldots, \hat{f}_p^{(l)}$ at the lth iteration of the algorithm, we update the estimates of the mixture proportions by setting $\hat{\pi}_j^{(l+1)} = n^{-1} \Sigma_{i=1}^n \hat{\theta}_{i,j}^{(l)}$ for $j = 1, \ldots, p$, where

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

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 by using the algorithm that was described in Section 3, choosing $\hat{f}_j^{(l+1)}$ to be the log-concave density f_j that maximizes

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

The incorporation of the weights $\hat{\theta}_{1,j}^{(l)}, \ldots, \hat{\theta}_{n,j}^{(l)}$ in the maximization process presents no additional complication, as is easily seen by inspecting the proof of theorem 1. As usual with methods that are 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 that is concentrated on a single observation, so the likelihood becomes arbitrarily high. The same issue can arise when fitting mixtures of Gaussian densities, and in this context Fraley and Raftery (2002) suggested 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 *et al.* (1993), which is available on the machine learning repository Web site at the University of California, Irvine (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 digitized 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 on the basis of the first two principal components of the 30 different measurements, which capture 63% of the variability in the full data set. These data are presented in Fig. 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. 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 Gaussian distributions model may be inadequate, and in Fig. 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 Fig. 6(c), whereas Fig. 6(d) plots the fitted mixture distribution from the log-concave EM algorithm. For

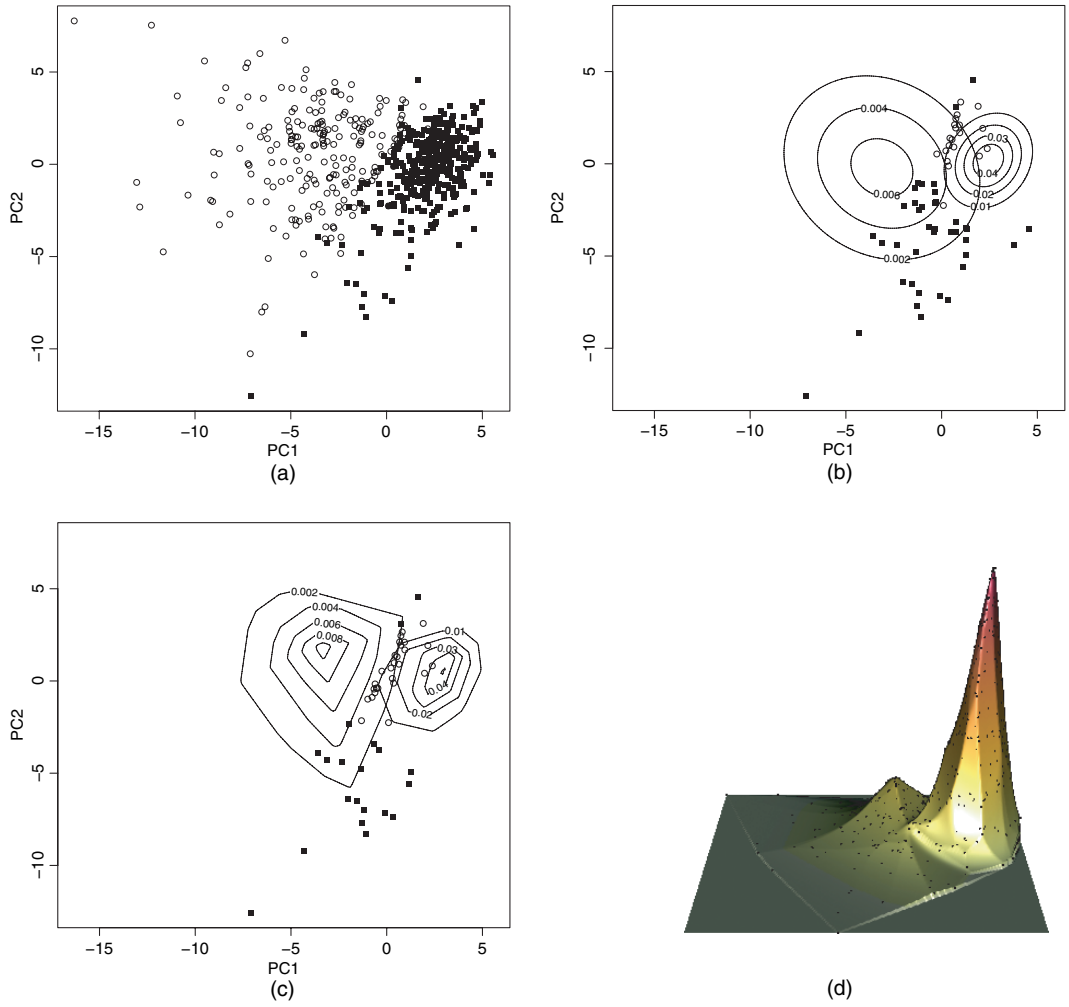


Fig. 6. (a) Wisconsin breast cancer data (■, benign cases; ○, malignant cases), (b) contour plot together with the misclassified instances from the Gaussian EM algorithm, (c) corresponding plot obtained from the log-concave EM algorithm and (d) fitted mixture distribution from the log-concave EM algorithm

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

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. $\mathbb{P}(\|X\| \geqslant 1) = \int \mathbb{1}_{\{\|x\| \geqslant 1\}} \, \mathrm{d}F(x)$. In such cases, it is more natural to use the plug-in estimator that is 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 $\|f\|_{1,a} = \int \exp(a\|x\|)|f(x)| \, \mathrm{d}x$. The conclusion of theorem 3 is that $\|\hat{f}_n - f^*\|_{1,a} \to 0$ almost surely as $n \to \infty$ for a range of values of a, where f^* is the log-concave density that minimizes 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 $\|\hat{\theta} - \theta^*\| \to 0$ almost surely, 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 $\theta(f) = \int f(x) g(x) \, \mathrm{d}x$ for some function g. Such a functional is continuous (so $\hat{\theta}$ is strongly consistent) provided merely that $\sup_{x \in \mathbb{R}^d} \{\exp(-a\|x\|)|g(x)|\} < \infty$ for some a in the allowable range that is provided by theorem 3. 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 $\hat{\theta}_B \to \hat{\theta}$ almost surely as $B \to \infty$. In practice, even when analytic calculation of $\hat{\theta}$ was possible, this method was found to be fast and accurate.

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

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 an $N_2(0, I)$ distribution, and we compare the performance of the LogConcDEAD

estimate with that of a kernel-based plug-in estimate, where the bandwidth was chosen by 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 the densities in Section 5, though we present results only for density (c) and $d = 2$ for brevity. See Cule (2009) for further examples and results. In Fig. 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 $\mu_f(A) = \int_A f(x)\,\mathrm{d}x$ and '$\triangle$' denotes set difference. Highest density regions can be computed once we have approximated the sample versions of f_α by using the density quantile algorithm that was 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 Fig. 8.

In real data examples, we cannot 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 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.

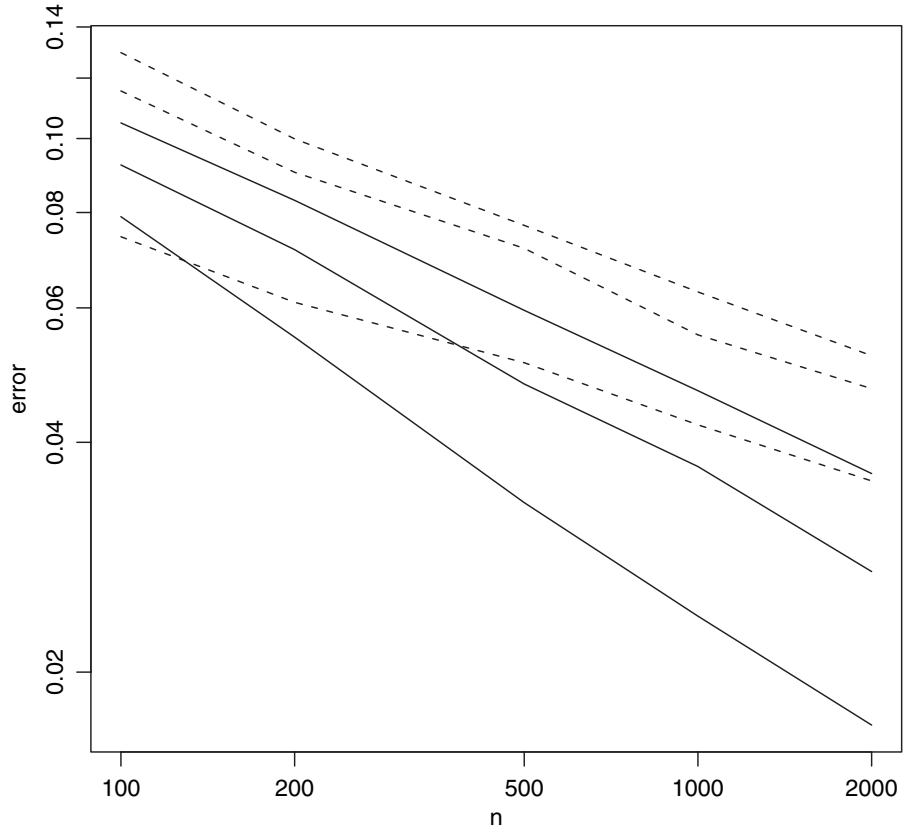


Fig. 7. Error for the highest density regions (the lowest of each set of lines are the 25% highest density region, the middle lines are the 50% highest density region and the highest lines are the 75% highest density region): ———, LogConcDEAD estimates; - - - - - - -, kernel estimates

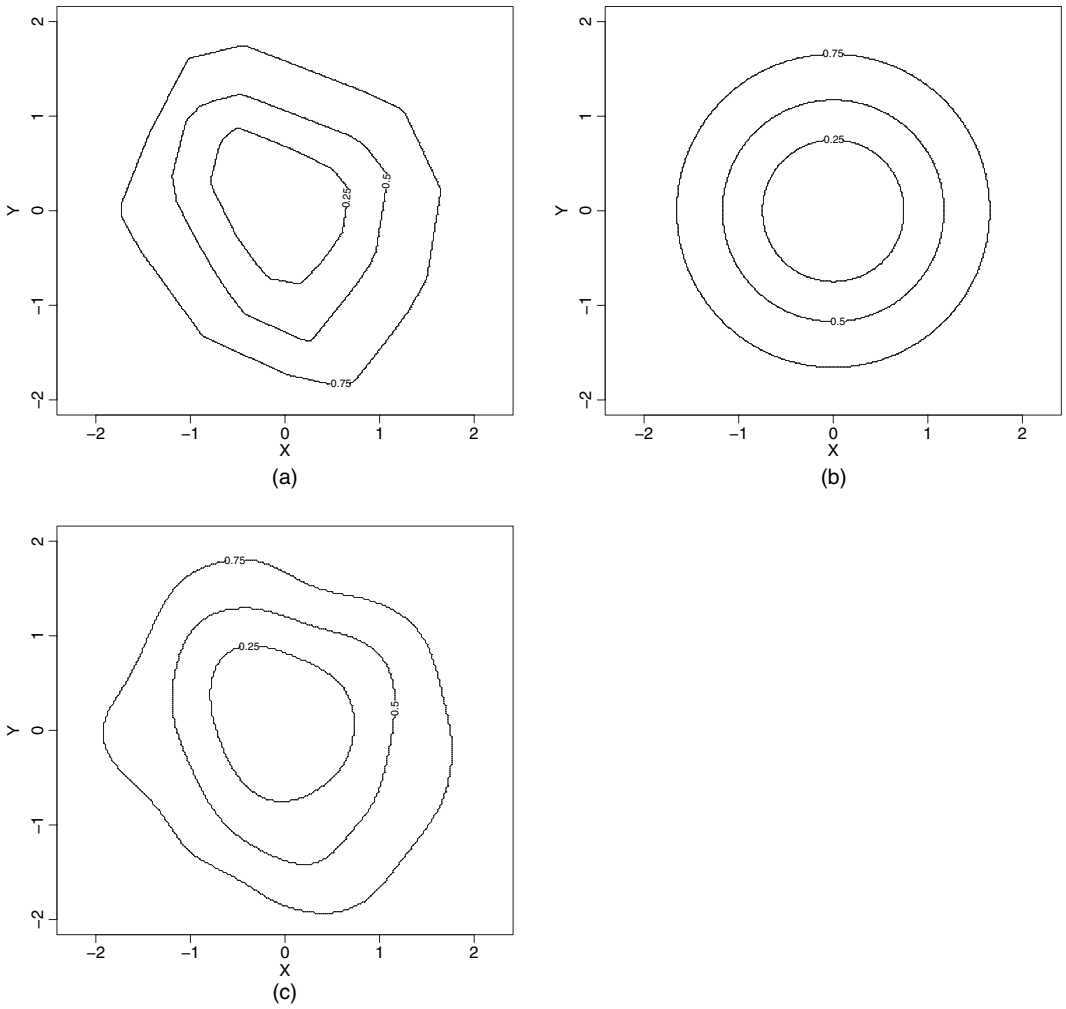


Fig. 8. Estimates of the 25%, 50% and 75% highest density region from 500 observations from the $N_2(0, I)$ distribution: (a) LogConcDEAD estimate; (b) true regions; (c) kernel estimate

8. Assessing log-concavity

In Section 4 we mentioned the fact that we can never be certain that a particular data set comes from a log-concave density. Even though theorem 3 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 that is similar to that described in Walther (2002) to test the null hypothesis that a log-concave model adequately models the data, compared with 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) proved 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. To do this, we compute

$$\hat{f}_n^c = \arg\max_{f \in \mathcal{F}^c} \{L(f)\}$$

for fixed values $c \in \mathcal{C} = \{c_0, \ldots, c_M\}$, where $\mathcal{F}^c = \{f : f(x) = \exp\{\phi(x) + c\|x\|^2\} \text{ with } \phi \text{ concave}\}$. We wish to assess how much $\hat{f}_n^c$ deviates from log-concavity; one possible measure is

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

where $\bar{h}$ is the least concave majorant of $\log(\hat{f}_n^c)$. 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 that was 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

$$\frac{1}{B+1}\#\{b : \max_{c \in \mathcal{C}}\{\tilde{T}(c)\} < \max_{c \in \mathcal{C}}\{\tilde{T}_b^*(c)\}\}.$$

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 an $N(0, \sigma^2)$ 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 non-parametric tests (having a single peak). Fig. 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 that was described above using $B = 99$ and $M = 11$. 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. Fig. 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 formalize this idea with a notion of

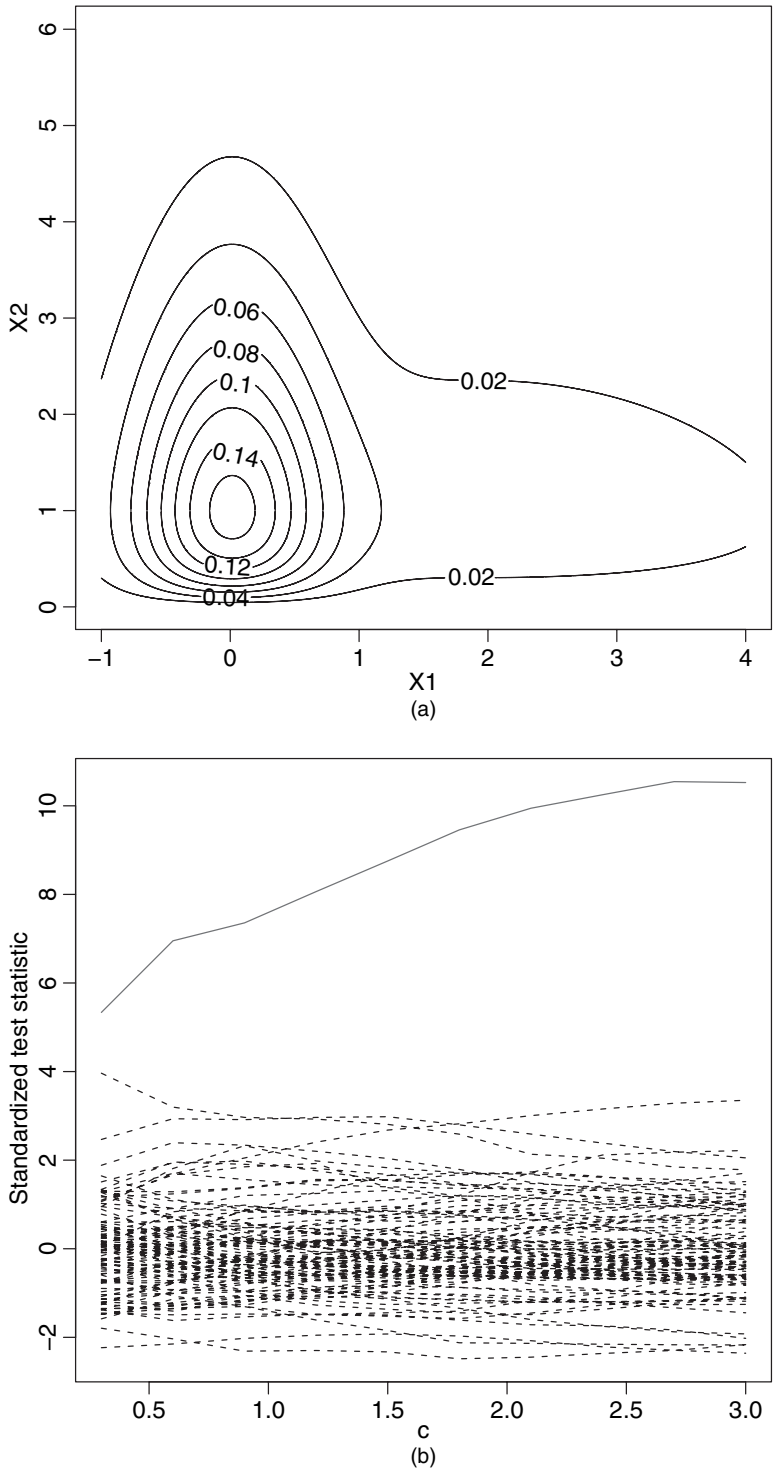


Fig. 9. Assessing the suitability of log-concavity: (a) contour plot of the density; (b) test statistic (———) and bootstrap reference values (--------)

distance, and a fairly natural metric 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

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

where $\mathcal{A}_0$ is the set of all balls centred 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 equation (8.1) is attained and the test statistic is easy to compute.

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 relabel 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 $T_{(1)}^* \leqslant \ldots \leqslant T_{(B)}^*$. 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 we would expect to find clusters of points with the same label (i.e. with or without stars). Thus the supremum in equation (8.1) may be attained at a relatively small ball radius. In contrast, in the permuted samples, the supremum is likely to be attained at a ball radius that includes approximately half of the points in the combined sample, so by restricting the ball radius we shall 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 *et al.* (2008). There it was shown that, under mild regularity conditions, the misclassification rate is minimized 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 normal distributions 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 $\|\mu\| \leqslant 2$. For each simulation set-up, we conducted 200 hypothesis tests with $k = \lfloor n^{4/(d+4)} \rfloor$ and $B = 99$, and we report in Table 3 the proportion of times that 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

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

n	*Proportions for the following values of* $\|\mu\|$*:*				
	$\|\mu\| = 0$	$\|\mu\| = 1$	$\|\mu\| = 2$	$\|\mu\| = 3$	$\|\mu\| = 4$
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

that come from slightly different distributions, tends to be a little smaller than the test statistic that is based on the permuted samples, in which both samples come from the same distribution. The explanation is that the dependence between $\mathcal{X}$ and $\mathcal{X}^*$ means that the realizations of the empirical distributions P_n and P_n^* tend to be particularly close together. Nevertheless, the test can 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 the following:

(a) 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 (2010) and 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);

(b) 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 that was presented in Dümbgen and Rufibach (2009), is as follows. 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. We can therefore define a slightly smoothed version of $\hat{f}_n$ via the convolution

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

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, an analogue of theorem 3 also applies to $\tilde{f}_n$) but offers potential improvements for small sample sizes);

(c) assessing the uncertainty in shape-constrained non-parametric density estimates, through confidence intervals or bands;

(d) developing analogous methodology and theory for discrete data under shape constraints;

(e) examining non-parametric shape constraints in regression problems, such as those studied in Dümbgen *et al.* (2010);

(f) studying methods for choosing the number of clusters in non-parametric, shape-constrained mixture models.

Acknowledgements

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

Appendix A: Proofs

A.1. Proof of proposition 1

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

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

which is 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

$$\begin{aligned}\log[f\{\lambda x_1 + (1-\lambda)x_2\}] &\geqslant \lambda \log\{f_{X|P_V(X)}(x_1|t)\} + (1-\lambda)\log\{f_{X|P_V(X)}(x_2|t)\} + \log\{f_{P_V(X)}(t)\} \\ &= \lambda \log\{f(x_1)\} + (1-\lambda)\log\{f(x_2)\}\end{aligned}$$

so f is log-concave, as required.

A.2. Proof of theorem 1

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 1 when $n \geqslant d+1$). By a standard argument in convex analysis (Rockafellar (1997), page 37), for each $y = (y_1, \ldots, y_n) \in \mathbb{R}^n$ there 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 $\bar{h}_y(X_i) \geqslant y_i$ 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

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

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

(a) $f(x) > 0$ for $x \in C_n$,
(b) $f(x) = 0$ for $x \notin C_n$,
(c) $\log(f) \in \mathcal{H}$,
(d) $f \in \mathcal{F}_0$ and
(e) there exists $M > 0$ such that, if $\max_i |\bar{h}_y(X_i)| \geqslant M$, then $\psi_n\{\exp(\bar{h}_y)\} \leqslant \psi_n(f)$.

First note that, if $x_0 \in C_n$, then by Carathéodory's theorem (theorem 17.1 of Rockafellar (1997)), there are distinct indices $i_1, \ldots, i_r$ with $r \leqslant d+1$, 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,

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

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

Now suppose that $f(x_0) > 0$ 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 part (b).

To prove part (c), 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 $\log\{f(x)\} \geqslant \bar{h}_y(x)$ for all $x \in \mathbb{R}^d$, we may assume that there is $x_0 \in C_n$ 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

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

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 $\bar{h}_y = \mathrm{cl}(\bar{h}_y) = \mathrm{cl}\{\log(f)\} \geqslant \log(f)$, 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 part (c).

The function $\log(f)$ has no *direction of increase* because, if $x \in C_n$, z is a non-zero vector and $t > 0$ is sufficiently large 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 (a) and (b) as well, we may write $\int f(x)\,\mathrm{d}x = c$, say, where $c \in (0, \infty)$. Thus $f(x) = c\,\bar{f}(x)$, for some $\bar{f} \in \mathcal{F}_0$. But then

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

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

To prove part (e), we may assume by property (d) 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, for $\exp(\bar{h}_y)$ to be a density, m must be negative with $|m|$ so large that $\psi_n\{\exp(\bar{h}_y)\} \leqslant \psi_n(f)$. 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

$$\begin{aligned}\bar{h}_y\left\{X_i + \frac{1}{M-m}(x - X_i)\right\} &\geqslant \frac{1}{M-m}\bar{h}_y(x) + \frac{M-m-1}{M-m}\bar{h}_y(X_i)\\ &\geqslant \frac{m}{M-m} + \frac{(M-m-1)M}{M-m} = M - 1.\end{aligned}$$

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

$$\mu[\{x : \bar{h}_y(x) \geqslant M - 1\}] \geqslant \mu\left\{X_i + \frac{1}{M-m}(C_n - X_i)\right\} = \frac{\mu(C_n)}{(M-m)^d}.$$

Thus

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

For $\exp(\bar{h}_y)$ to be a density, then, we require

$$m \leqslant -\tfrac{1}{2}\exp\{(M-1)/d\}\,\mu(C_n)^{1/d}$$

when M is large. But then

$$\psi_n\{\exp(\bar{h}_y)\} \leqslant \frac{(n-1)M}{n} - \frac{1}{2n}\exp\left(\frac{M-1}{d}\right)\mu(C_n)^{1/d} \leqslant \psi_n(f)$$

when M is sufficiently large. This proves part (e).

It is not difficult to see that, for any $M > 0$, the function $y \mapsto \psi_n\{\exp(\bar{h}_y)\}$ 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 maximize $\psi_n(f)$. We may assume that $f_1, f_2 \in \mathcal{F}_0$, $\log(f_1)$, $\log(f_2) \in \mathcal{H}$ and f_1 and f_2 are supported on C_n. Then the normalized geometric mean

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

is a log-concave density, with

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

However, by the Cauchy–Schwarz inequality, $\int_{C_n}\{f_1(y)\,f_2(y)\}^{1/2}\,\mathrm{d}y \leqslant 1$, so $\psi_n(g) \geqslant \psi_n(f_1)$. 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 $f_1 = f_2$.

A.3. Proof of theorem 2

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

$$\bar{h}_{ty^{(1)}+(1-t)y^{(2)}}(X_i) \geqslant ty_i^{(1)} + (1-t)y_i^{(2)}$$

for $i = 1, \ldots, n$, so

$$\bar{h}_{ty^{(1)}+(1-t)y^{(2)}} \leqslant t\bar{h}_{y^{(1)}} + (1-t)\bar{h}_{y^{(2)}}.$$

The convexity of σ follows from this and the convexity of the exponential function. It is clear that $\sigma \geqslant \tau$, since $\bar{h}_y(X_i) \geqslant y_i$ for $i = 1, \ldots, n$.

From theorem 1, 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^* minimizes τ. For any other $y \in \mathbb{R}^n$ which minimizes τ, by the uniqueness part of theorem 1 we must have $\bar{h}_y = \bar{h}_{y^*}$, so $\sigma(y) > \sigma(y^*) = \tau(y^*)$.

Appendix B: Structural and computational issues

As illustrated in Fig. 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 is a finite set J consisting of m such $(d+1)$-tuples, with the following three properties:

(a) $\cup_{j \in J} C_{n,j} = C_n$,
(b) the relative interiors of the sets $\{C_{n,j} : j \in J\}$ are pairwise disjoint and
(c)

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

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 by 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 equation (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 $T_d = \{w = (w_1, \ldots, w_d) : w_l \geqslant 0, \Sigma_{l=1}^d w_l \leqslant 1\}$ 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

$$\begin{aligned} \int_{C_n} \exp\{\bar{h}_y(x)\}\,\mathrm{d}x &= \sum_{j \in J} |\det(A_j)| \int_{T_d} \exp(y_{j_0} w_0 + \ldots + y_{j_d} w_d)\,\mathrm{d}w \\ &= \sum_{j \in J} |\det(A_j)| \exp(y_{j_0}) \sum_{r=1}^{d} \frac{\exp(z_{j,r}) - 1}{z_{j,r}} \prod_{\substack{1 \leqslant s \leqslant d \\ s \neq r}} \frac{1}{z_{j,r} - z_{j,s}}. \end{aligned}$$

The singularities that occur when some of $z_{j,1}, \ldots, z_{j,d}$ may be 0 or equal are removable. However, for stable computation of σ in practice, a Taylor approximation was used—see Cule and Dümbgen (2008) and 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 that is defined in equation (3.2) is differentiable, and we 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

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

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. Note 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 1). Let $w_0 = 1 - w_1 - \ldots - w_d$ and, for $y \in \mathbb{R}^n$ and $i = 1, \ldots, n$, let

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

Proposition 2.

(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. For part (a), by theorem 25.2 of Rockafellar (1997), it suffices to show that, for $y \in \mathcal{Y}$, all the partial derivatives exist and are given by the expression in the statement of proposition 2. 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 co-ordinate 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 $|t|$. However, if $j \in J_i$, then there are two cases to consider.

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

It follows that

$$\begin{aligned}
\frac{\partial \sigma}{\partial y_i}(y) &= -\frac{1}{n} + \lim_{t \to 0} \left\{ \frac{1}{t} \sum_{j \in J_i} \int_{C_{n,j}} \exp(\langle x, b_j^{(t)} \rangle - \beta_j^{(t)}) - \exp(\langle x, b_j \rangle - \beta_j) \, \mathrm{d}x \right\} \\
&= -\frac{1}{n} + \lim_{t \to 0} \left\{ \frac{1}{t} \sum_{j \in J_i} \left(\int_{C_{n,j}} \exp(\langle x, b_j \rangle - \beta_j)[\exp\{t(1 - \langle A_j^{-1}(x - \alpha_j), \mathbf{1}_d \rangle)\} - 1] \, \mathrm{d}x \, \mathbb{1}_{\{j_0 = i\}} \right. \right. \\
&\quad \left. \left. + \sum_{l=1}^{d} \int_{C_{n,j}} \exp(\langle x, b_j \rangle - \beta_j)[\exp\{t\langle A_j^{-1}(x - \alpha_j), e_l^d \rangle\} - 1] \, \mathrm{d}x \, \mathbb{1}_{\{j_l = i\}} \right) \right\} \\
&= -\frac{1}{n} + \sum_{j \in J_i} \left[\int_{C_{n,j}} \exp(\langle x, b_j \rangle - \beta_j)\{1 - \langle A_j^{-1}(x - \alpha_j), \mathbf{1}_d \rangle\} \, \mathrm{d}x \, \mathbb{1}_{\{j_0 = i\}} \right. \\
&\quad \left. + \int_{C_{n,j}} \exp(\langle x, b_j \rangle - \beta_j)\langle A_j^{-1}(x - \alpha_j), e_l^d \rangle \, \mathrm{d}x \, \mathbb{1}_{\{j_l = i\}} \right] \\
&= \partial_i(y),
\end{aligned}$$

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

For part (b), if $y \in \mathcal{Y}^c$, then it can be shown that there is a unit co-ordinate vector $e_i^n in \mathbb{R}^n$ such that the *one-sided directional derivative* at y with respect to e_i^n, which is 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 $\varepsilon > 0$, a point $\tilde{y} \in \mathbb{R}^n$ such that $\|\tilde{y} - y\| < \varepsilon$ and such that σ is differentiable at $\tilde{y}$ with $\|\nabla\sigma(\tilde{y}) - \partial(y)\| < \varepsilon$. 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 2 and equation (B.1) and once we have a formula for

$$\tilde{I}_{d,u}(z) = \int_{T_d} w_u \exp\left(\sum_{r=1}^{d} z_r w_r\right) \mathrm{d}w.$$

An explicit closed formula for $\tilde{I}_{d,u}(z)$ where $z_1, \ldots, z_d$ are non-zero and distinct is derived in Cule *et al.* (2010). Again, for practical purposes, we use a Taylor expansion for cases where $z_1, \ldots, z_d$ are close to 0 or approximately equal. Details are given in Cule and Dümbgen (2008) and Cule (2009).

B.3. Sampling from the fitted density estimate

To use the Monte Carlo procedure that was described in Section 7.1, we must be able to sample from $\hat{f}_n$. Fortunately, this can be done efficiently by 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 $X_{jl} - X_{j0}$ for $l = 1, \ldots, d$, and let $\alpha_j = X_{j0}$, 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 $z_{j,l} = y_{jl}^* - y_{j0}^*$ for $l = 1, \ldots, d$. Write

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

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

(a) Select $j^* \in J$, selecting $j^* = j$ with probability q_j.
(b) 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 this step.

Appendix C: Glossary of terms and results from convex analysis and computational geometry

All 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

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

We say that 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) called a *proper concave* function. If C is a convex subset of $\mathbb{R}^d$ then, provided that $f : C \to [-\infty, \infty)$ is not identically $-\infty$, it is *concave* if and only if

$$f\{tx + (1-t)y\} \geqslant t\, f(x) + (1-t)\, f(y)$$

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 $\int_{\mathbb{R}^d} f(x) \, \mathrm{d}x = 1$. 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 the closure of $\{x \in \mathbb{R}^d : \log\{f(x)\} > -\infty\}$, which is 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, which is 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 $T\{tx + (1-t)y\} = t\, T(x) + (1-t)\, T(y)$ for all $x, y \in M$ and $t \in \mathbb{R}$.

The *closure* of a concave function g on $\mathbb{R}^d$, which is 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 semicontinuous, concave function satisfying $\mathrm{cl}(g) \geqslant g$. 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* (plural, *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 $\langle x, b\rangle \leqslant \beta$ 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, which is denoted $\mathrm{vert}(P)$, is the set of zero-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 that are 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 that is obtained by modifying each $S_l \in S$ as follows.

(a) If $v \notin S_l$, then $S_l \in \tilde{S}$.
(b) If $v \in S_l$ and $\mathrm{conv}[\mathrm{vert}(S_l) \setminus \{v\}]$ is $d-1$ dimensional, then $S_l \in \tilde{S}$.
(c) 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

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

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

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

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

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. Econ. Theor.*, **80**, 350–369.

Asuncion, A. and Newman, D. J. (2007) UCI Machine Learning Repository. University of California, Irvine. (Available from `http://www.ics.uci.edu/~mlearn/MLRepository.html`.)

Bagnoli, M. and Bergstrom, T. (2005) Log-concave probability and its applications. *Econ. Theor.*, **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. Softwr.*, **22**, 469–483.

Barndorff-Nielsen, O. (1978) *Information and Exponential Families in Statistical Theory*. New York: 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), pp. 169–177. New York: Springer.

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 Nalebuff, B. (1991a) Aggregation and social choice: a mean voter theorem. *Econometrica*, **59**, 1–23.
Caplin, A. and Nalebuff, B. (1991b) Aggregation and imperfect competition: on the existence of equilibrium. *Econometrica*, **59**, 25–59.
Chacón, J. E. (2009) Data-driven choice of the smoothing parametrization for kernel density estimators. *Can. J. Statist.*, **34**, 249–265.
Chacón, J. E., Duong, T. and Wand, M. P. (2010) Asymptotics for general multivariate kernel density derivative estimators. *Statist. Sin.*, to be published.
Chang, G. and Walther, G. (2007) Clustering with mixtures of log-concave distributions. *Computnl Statist. Data Anal.*, **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. *PhD Thesis*. University of Cambridge, Cambridge.
Cule, M. L. and Dümbgen, L. (2008) On an auxiliary function for log-density estimation. *Technical Report 71*. Universität Bern, Bern.
Cule, M. L., Gramacy, R. B. and Samworth, R. J. (2007) LogConcDEAD: Maximum Likelihood Estimation of a Log-Concave Density. Statistical Laboratory, Cambridge. (Available from `http://CRAN.R-project.org/package=LogConcDEAD`.)
Cule, M. L., Gramacy, R. B. and Samworth, R. J. (2009) LogConcDEAD: an R package for maximum likelihood estimation of a multivariate log-concave density. *J. Statist. Softwr.*, **29**, issue 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). Statistical Laboratory, Cambridge. (Available from `http://www.statslab.cam.ac.uk/~rjs57/Research.html`.)
Ćwik, J. and Koronacki, J. (1997) Multivariate density estimation: a comparative study. *Neur. Computn Appl.*, **6**, 173–185.
Deheuvels, P. (1977) Estimation non parametrique de la densité par histogrammes generalisés II. *Publ. Inst. Statist. Univ. Paris*, **22**, 1–23.
Dempster, A. P., Laird, N. M. and Rubin, D. B. (1977) Maximum likelihood from incomplete data via the EM algorithm (with discussion). *J. R. Statist. Soc.* B, **39**, 1–38.
Devroye, L., Györfi, L. and Lugosi, G. (1996) *A Probabilistic Theory of Pattern Recognition*. New York: Springer.
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. *Technical Report 61*. Universität Bern, Bern. (Available from `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. *Technical Report 75*. Universität Bern, Bern. (Available from `http://arxiv.org/abs/1002.3448/`.)
Duong, T. (2004) Bandwidth selectors for multivariate kernel density estimation. *PhD Thesis*. University of Western Australia, Perth.
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 bandwidth matrix selectors in multivariate kernel density estimation. *J. Multiv. Anal.*, **93**, 417–433.
Eggermont, P. P. B. and LaRiccia, V. (2001) *Maximum Penalized Likelihood Estimation*, vol. 1, *Density Estimation*. New York: Springer.
Eubank, R. L. (1988) *Spline Smoothing and Nonparametric Regression*. New York: Dekker.
Fix, E. and Hodges, J. L. (1951) Discriminatory analysis—nonparametric discrimination: consistency properties. *Technical Report 4*, project 21-29-004. US Air Force School of Aviation Medicine, Randolph Field.
Fix, E. and Hodges, J. L. (1989) Discriminatory analysis—nonparametric discrimination: consistency properties. *Int. Statist. Rev.*, **57**, 238–247.
Fraley, C. F. and Raftery, A. E. (2002) Model-based clustering, discriminant analysis, and density estimation. *J. Am. Statist. Ass.*, **97**, 611–631.
Gordon, A. D. (1981) *Classification*. London: Chapman and Hall.
Grenander, U. (1956) On the theory of mortality measurement II. *Skand. Akt.*, **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. Theor. Reltd Flds*, **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. *Am. Statistn*, **50**, 120–126.
Ibragimov, A. I. (1956) On the composition of unimodal distributions. *Theor. Probab. Appl.*, **1**, 255–260.
Jongbloed, G. (1998) The iterative convex minorant algorithm for nonparametric estimation. *J. Computnl Graph. Statist.*, **7**, 310–321.
Kappel, F. and Kuntsevich, A. (2000) An implementation of Shor's *r*-algorithm. *Computnl Optimizn Appl.*, **15**, 193–205.
Koenker, R. and Mizera, I. (2010) Quasi-concave density estimation. *Ann. Statist.*, to be published.
Lee, C. W. (2004) Subdivisions and triangulations of polytopes. In *Handbook of Discrete and Computational Geometry* (eds J. E. Goodman and J. O'Rourke), 2nd edn, pp. 383–406. New York: CRC Press.
McLachlan, G. J. and Basford, K. E. (1988) *Mixture Models: Inference and Applications to Clustering*. New York: 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*, 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 Sci. Math.*, **34**, 335–343.
R Development Core Team (2009) *R: a Language and Environment for Statistical Computing*. Vienna: R Foundation for Statistical Computing.
Rockafellar, R. T. (1997) *Convex Analysis*. Princeton: 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. Computn Simuln*, **77**, 561–574.
Rufibach, K. and Dümbgen, L. (2006) logcondens: estimate a log-concave probability density from i.i.d. observations. Universität Bern, Bern. (Available from `http://CRAN.R-project.org/package=logcondens`.)
Sain, S. R. (2002) Multivariate locally adaptive density estimation. *Computnl Statist. Data Anal.*, **39**, 165–186.
Sain, S. R. and Scott, D. W. (1996) On locally adaptive density estimation. *J. Am. Statist. Ass.*, **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. (2009) Multivariate log-concave distributions as a nearly parametric model. *Technical Report 74*. Universität Bern, Bern. (Available from `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. (2010) Nonparametric estimation of convex-transformed densities *Ann. Statist.*, to be published.
Shor, N. Z. (1985) *Minimization Methods for Non-differentiable Functions*. Berlin: Springer.
Street, W. M., Wolberg, W. H. and Mangasarian, O. L. (1993) Nuclear feature extraction for breast tumor diagnosis. In *Proc. Int. Symp. Electronic Imaging: Science and Technology, San Jose*, pp. 861–870.
Swales, J. D. (ed.) (1985) *Platt vs. Pickering: an Episode in Recent Medical History*. Cambridge: Keynes.
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*, pp. 659–665. Cambridge: 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. Am. Statist. Ass.*, **97**, 508–513.
Walther, G. (2009) Inference and modeling with log-concave distributions. *Statist. Sci.*, **24**, 319–327.
Wand, M. P. and Jones, M. C. (1995) *Kernel Smoothing*. Boca Raton: Chapman and Hall–CRC Press.
Zhang, X., King, M. L. and Hyndman, R. J. (2006) Bandwidth selection for multivariate kernel density estimation using MCMC. *Computnl Statist. Data Anal.*, **50**, 3009–3031.

Discussion on the paper by Cule, Samworth and Stewart

Kaspar Rufibach (*University of Zurich*)

The authors are to be congratulated on the extension of log-concave density estimation to more than one dimension. Their work marks the temporary culmination of substantial research activity in shape-constrained density estimation over the last decade and directs attention to (at least) two directions that previously had received little or no regard. First, apart from very recent concurrent papers by Schuhmacher *et al.* (2009), Seregin and Wellner (2009), Koenker and Mizera (2010), Dümbgen *et al.* (2010) and Schuhmacher and Dümbgen (2010), non-parametric estimation of shape-constrained densities in dimension $d \geqslant 2$ has received virtually no attention. Apart from the theoretical obstacles that are related to these problems this neglect may be attributed to the difficulty of implementing algorithms to maximize the underlying likelihood function. The development of an algorithm and its implementation in R (Cule *et al.*, 2009; R Development Core Team, 2009) for the log-concave case is certainly the first highlight of this paper. In dimension $d = 1$, after realizing that the maximizer of the likelihood function must be piecewise linear with kinks only at the observations, finding the log-concave density estimate boils down to maximizing a concave functional on $\mathbb{R}^n$ subject to linear constraints; see Rufibach (2007). In the multivariate case, however, it is not clear how to parameterize the class of *concave* tent functions that hampers the formulation of a (linearly) constrained maximization problem similarly to the univariate scenario. To circumvent this problem the authors modified the initial likelihood function to receive an updated functional whose *unconstrained* maximizer gives rise to the tent function that corresponds to the log-concave density estimate. This updated functional is concave but non-differentiable, disallowing the use of standard optimization algorithms. Instead, the authors successfully implemented (Cule *et al.*, 2009) an algorithm due to Shor (l985) which can handle non-differentiable target functionals.

As a second highlight the authors show that the estimator converges to the log-concave density f^* where this density minimizes the Kullback–Leibler divergence to f_0, the density of the observations. To the best of my knowledge, this general set-up has not previously been considered for shape-constrained density estimation, not even for $d = 1$ for example in Groeneboom *et al.* (2001) or Dümbgen and Rufibach (2009), which dealt only with the well-specified case where f_0 is log-concave. However, to assess robustness properties an analysis of the misspecified model is particularly valuable. A natural link here is: what are the limit distribution results under misspecification? In the well-specified univariate case, the pointwise limiting distribution is known (see Balabdaoui *et al.* (2009)) and it seems worthwile to generalize these results to

(a) higher dimensions and
(b) the misspecified scenario.

Having shown consistency in some strong norms the natural next question is: what rates of convergence can be expected for the log-concave density estimator? For $d = 1$ rates of convergence, either in sup-norm (Dümbgen and Rufibach, 2009) or pointwise (Balabdaoui *et al.*, 2009), have been derived.

If we assume that f_0 belongs to a Hölder class with exponent $\alpha \in [1, 2]$, the minimax optimal rate for estimators within such a class can be derived from the entropy structure of the underlying function space and is $n^{-\alpha/(2\alpha+d)}$; see Birgé and Massart (1993). However, Birgé and Massart (1993) showed that the rate of convergence for *minimum contrast estimators*, which are a class that contains maximum likelihood estimators (MLEs) as a special case, is only $n^{-\alpha/2d}$ once $d > 2\alpha$.

The dependence of the exponents of these rates of convergence on dimension for $\beta = 2$, i.e. for densities with uniformly bounded second derivative, is displayed in Fig. 10, which reveals that up to dimension $d = 4$ we can conjecture the MLE to be rate efficient but beyond that split MLEs do not reach the minimax optimal rate anymore. Future work should aim at

(a) in fact verifying the conjectured rates for the log-concave MLE in arbitrary dimension and
(b) 'fixing' the log-concave MLE for dimensions $d > 4$ to make them also rate efficient for higher dimensions.

How to achieve this goal is another open issue: (additional) penalization comes to mind or consideration of classes of densities that are smaller than that of log-concave densities but yet non-parametric.

In addition to solving some important questions this paper has opened up new directions for research in shape-constrained density estimation and I am convinced that it will stimulate further research in the area. Consequently, I have great pleasure in proposing the vote of thanks.

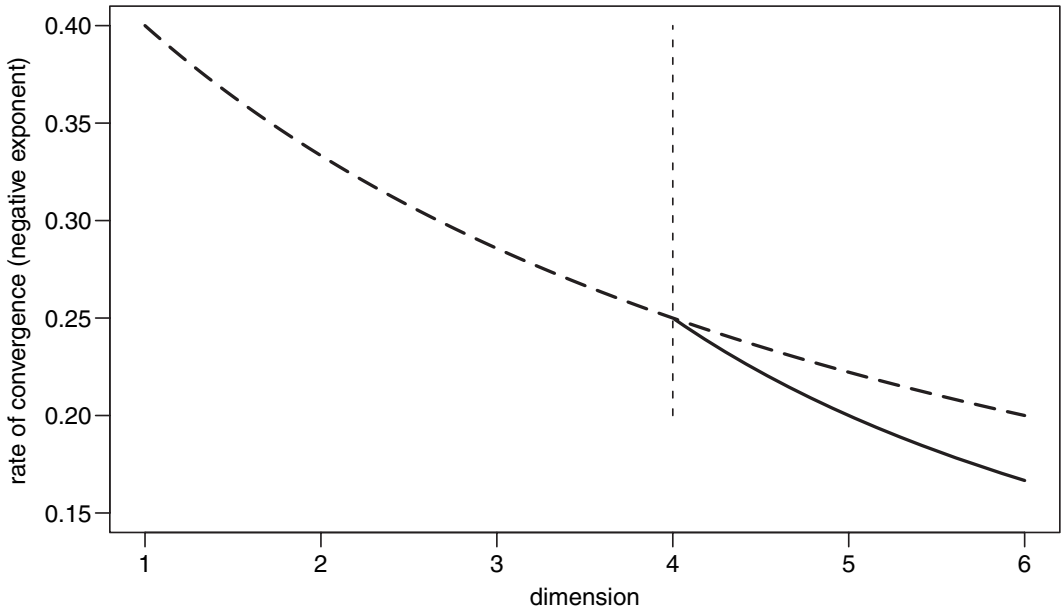


Fig. 10. Rates of convergence for minimax optimal (– – –) and minimum contrast estimators (———): illustration for $\beta = 2$

Aurore Delaigle (*University of Melbourne*)

I congratulate the authors for a very stimulating, innovative and carefully written paper on the topic of non-parametric estimation of a multivariate density f. A popular estimator in this context is the kernel density estimator (KDE). Although this estimator is consistent under mild smoothness conditions on f, its quality degrades quickly as the dimension increases. The authors suggest imposing a structural assumption of log-concavity on a non-parametric (maximum likelihood) estimator, to improve performance in practice. They develop a nice theoretical study and discuss a variety of interesting applications of their method, encompassing plain density estimation, hypothesis testing and clustering problems.

Compared with the KDE, the numerical improvement that is achieved by the new procedure is impressive. However, one may question the suitability of comparison between these two methods. In particular, the KDE that is discussed in the paper is a non-restricted non-parametric estimator, whereas the methodology suggested incorporates a strong log-concavity shape constraint. Are we surprised to do better by incorporating *a priori* knowledge of the density? Perhaps it would be more appropriate to compare the new procedure with a KDE which satisfies the same log-concavity constraint. Such shape-restricted KDEs were developed in the literature more than a decade ago (see, for example, the tilting method of Hall and Presnell (l999) and the discussion in Braun and Hall (2001)). Moreover, these modified KDEs are not restricted to log-concavity constraints; they can be used to impose a variety of shapes. Can the method that is suggested by the authors be extended to more general constraints?

When comparing their procedure with the KDE, the authors highlight the fact that their method does not require a choice of a smoothing parameter. This attracts at least two comments.

(a) If we were to use a shape-constrained KDE, which would make the comparison between methods more fair, then it is not clear that the choice of a bandwidth would be critical.
(b) A consequence of the fact that the authors do not use a smoothing parameter is that their estimator is not smooth, and in fact not differentiable. (See for example their Fig. 2.)

One might say that the estimator is not visually attractive, whereas by introducing a smoothing parameter the authors could make it smooth and differentiable. A simple approach could be to take the convolution between their estimator and a function $K_H(x) = H^{-1}K(\cdot/H)$, where the kernel function K is a smooth and symmetric density, and H denotes a small smoothing parameter. In fact, the approach that is

discussed in Section 9 is of this type, where the kernel function is the standard normal density and H is a matrix of smoothing parameters. Hence, to make their estimator smooth, the authors suggest introducing a kernel function and smoothing parameters. We may wonder how different from the tilted KDE the resulting estimator is. Incidentally, it is not clear that the theoretical mean integrated squared error bandwidth that is used by the authors in their numerical work is systematically better than a data-driven bandwidth (if the authors had employed the integrated squared error bandwidth, this would not have been questionable).

In Section 1, application (d), the authors suggest that their shape-restricted estimator be employed to assess the validity of a parametric model. Since their estimator already contains a rather strong shape constraint, it is not clear that this procedure would be appropriate. In most cases their estimator will contain a systematic bias (which does not vanish as the sample size increases) and it seems a little odd to use such an estimator to infer the validity of a parametric model. For example, an incorrect shape constraint can give the erroneous impression that the systematic bias of a wrong parametric model is smaller than it really is.

In Section 7, it is a little surprising that the authors consider examples (a) and (b) as potential applications of their method. Clearly, in both cases, one could employ empirical estimators which, unlike the authors' procedure, do not rely on any shape restriction. For the other examples that are treated in that section (where an empirical procedure is not available), again, the authors compare their method with the KDE, but the comparison does not seem fully satisfactory. First, as already noted above, the authors did not use the shape-restricted version of the KDE (the choice of the bandwidth is perhaps also questionable). Second, is it clear that, in the examples considered, imposing a shape constraint brings as much improvement as in the context of density estimation? This is particularly questionable for integrated quantities, which are easier to estimate than a full multivariate density (in such problems KDEs can usually achieve very good performance by undersmoothing, i.e. by using a bandwidth that is much smaller than for density estimation). How robust is the new estimator against non-log-concavity in such problems? Is it clear that the gain that can be obtained by imposing the right shape constraint is worth the loss that can occur by imposing a wrong constraint?

The vote of thanks was passed by acclamation.

Wenyang Zhang (*University of Bath*) **and Jialiang Li** (*National University of Singapore*)

We congratulate Dr Cule, Dr Samworth and Dr Stewart for such a brilliant paper. We believe that this paper will have a big influence on the estimation of multivariate density functions and will stimulate many further researches in this direction.

The commonly used approach to estimate density functions is based on kernel smoothing. The authors take a completely different approach; by making use of the log-concavity of the density function, they transform the density estimation to a non-differentiable convex optimization problem, which is quite interesting.

As the authors rightly point out kernel density estimation has a boundary effect problem. There are some boundary correction methods to allay this problem. It would be interesting to see a comparison between the method proposed and the kernel density estimation with boundary correction. We can envisage, even with the boundary correction, that kernel density estimation would still not perform as well as the method proposed does, as kernel density estimation has not made use of the log-concavity information. We guess that it is probably not very easy to make use of log-concavity information in kernel density estimation.

In multivariate kernel density estimation, the dimensionality could be a problem. Would the dimensionality be an issue in the method proposed? How would the dimensionality affect the convergence of the algorithm and the accuracy of the estimator proposed?

In real life, we often want to estimate the conditional density function of a response variable or vector given some covariates. Does the method proposed apply to the case where there are some covariates?

The basic idea of kernel estimation for conditional density functions is as follows: suppose that (X_i^{T}, Y_i), $i = 1, \ldots, n$, are independent identically distributed from (X^{T}, Y). By simple calculation, we have

$$p(y|X = x) \approx E\{K_h(Y - y)|X = x\}, \tag{1}$$

where $p(y|X = x)$ is the conditional density function of Y given $X = x$, $K_h(\cdot) = K(\cdot/h)/h$, $K(\cdot)$ is a kernel

function such that $\int K(u)\,\mathrm{d}u = 1$ and h is a bandwidth. Expression (1) leads to the non-parametric regression model

$$K_h(Y_i - y) = p(y|X = X_i) + \varepsilon_i, \qquad i = 1, \ldots, n. \tag{2}$$

The conditional density function estimation is now transformed to a non-parametric regression problem, and the estimator of conditional density functions can be obtained by non-parametric regression. Like the standard non-parametric modelling, we must impose some conditions on $p(y|X = x)$ when the dimension of X is not very small owing to the 'curse of dimensionality'. Which conditions should be imposed depends on the data set that we analyse and the problem that we are interested in.

Beaumont *et al.* (2002) proposed this regression approach to compute the posterior density function in approximate Bayesian computation with Y being the parameter concerned and X being the vector of selected statistics.

Vikneswaran Gopal and George Casella (*University of Florida, Gainesville*)

In Appendix B.3, the authors suggest an accept–reject (AR) algorithm to sample from the fitted maximum likelihood estimate of the density. We note the following.

(a) As the true density moves away from log-concavity, the acceptance rate falls.
(b) When both algorithms use an equal number of random variables, we show empirically that
  (i) a Metropolis–Hastings (MH) algorithm has a higher acceptance (move) rate and
  (ii) MH sampling yields smaller standard errors.

If MH and AR algorithms use the *same* generating candidate, MH sampling always has a higher acceptance rate (Robert and Casella (2004), lemma 7.9).

Metropolis–Hastings candidate density

As in the main paper, denote the estimated density by $\hat{f}_n$. A natural modification of the authors' AR candidate gives our MH proposal density Q,

$$Q(x) = \sum_{j=1}^{|J|} q_j \frac{I_{C_{n,j}}(x)}{\lambda(C_{n,j})} \tag{3}$$

where $C_{n,j}$ are defined in the paper. The volume of a simplex in $\mathbb{R}^d$ is $\lambda(C_{n,j}) = (1/d!)|\det(A_j)|$ (Stein, 1966), and the resulting MH algorithm at iteration n is given by the following steps.

Step 1: given $X_n = x$ pick $C_{n,j}$ with probability $(q_1, q_2, \ldots, q_{|J|})$ and sample from the uniform distribution on this simplex to obtain the candidate $Y_{n+1} = y$.
Step 2: compute the MH ratio, which is explicitly given by

$$\alpha(x, y) = \min\left\{ \frac{\hat{f}_n(y)\ Q(x)}{\hat{f}_n(x)\ Q(y)}, 1 \right\} = \min\left\{ \frac{\exp(b_j' y - \beta_j)\ q_i/\lambda(C_{n,i})}{\exp(b_i' x - \beta_i)\ q_j/\lambda(C_{n,j})}, 1 \right\} \tag{4}$$

for $x \in C_{n,i}$ and $y \in C_{n,j}$.
Step 3: set

$$X_{n+1} = \begin{cases} y & \text{with probability } a(x, y), \\ x & \text{otherwise.} \end{cases}$$

Simulation

We considered five sampling densities (Table 4), a correlated bivariate normal, two gammas and two t-distributions, and generated 150 observation from each. Both algorithms simulated $n = 5000$ candidates and computed the mean of the random variables returned. To obtain standard error estimates, each mean was computed 100 times. (To avoid burn-in issues, the Metropolis algorithm was started with an AR step.)

As seen in Table 4, the acceptance rate falls as we move from left to right However, MH sampling is consistently better than AR sampling and provides standard errors at least as good as the AR algorithm, with more pronounced differences towards the right-hand side of Table 4. Note also that switching from $\Gamma(1.1, 2)$ to $\Gamma(0.2)$, which crosses the log-concave border, causes the acceptance rate to plummet sharply.

Table 4. True means and estimated means (with standard errors in parentheses) for AR and MH algorithms, from a total of 5000 random variables for each algorithm†

		Results when log-concave			*Results when not log-concave*			
		Bivariate normal		$\Gamma(1.1, 2)$	$\Gamma(0.9, 2)$	$t(1)$	*Bivariate t*	
True sample mean		1.888	−1.980	0.519	0.414	−0.683	−1.144	0.208
Estimate of mean	AR	1.888	−1.980	0.519	0.414	−0.698	−1.117	0.237
		(0.018)	(0.018)	(0.010)	(0.016)	(0.134)	(0.501)	(0.433)
	MH	1.891	−1.979	0.517	0.414	−0.677	−1.135	0.254
		(0.017)	(0.016)	(0.010)	(0.014)	(0.115)	(0.296)	(0.293)
Acceptance rate	AR	0.575		0.402	0.120	0.121	0.020	
	MH	0.856		0.677	0.246	0.266	0.158	

†Standard errors are calculated from 100 replicate runs. Acceptance rates are based on the 5000 random variables. The two leftmost densities are log-concave, and the other three are not.

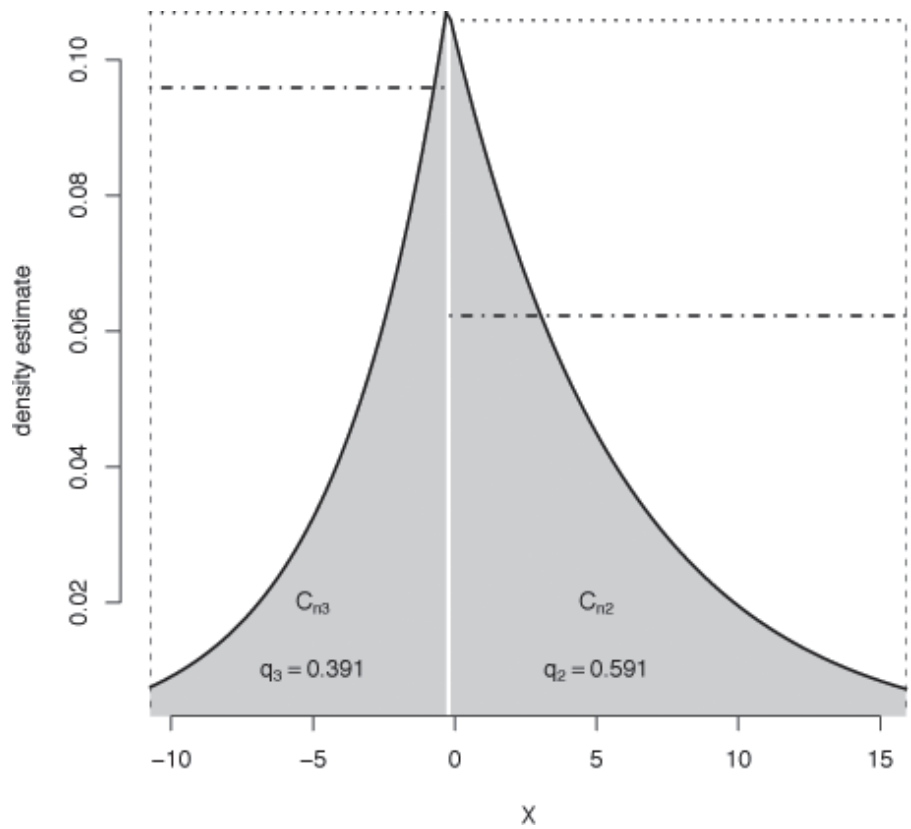


Fig. 11. With $n = 15$ observations from a t_1-distribution, the maximum likelihood estimators (———) of the density as fitted by the R package LogConcDEAD, where the triangulation of the convex hull C_n resulted in three simplices (· - · - · -, Metropolis candidate; · · · · · · ·, accept–reject candidate): the areas above simplices 2 and 3 have been shaded grey, and the sliver of white between the two grey areas corresponds to the region above simplex 1

Conclusion

Fig. 11 provides some insight about the MH–AR performance. If the AR scheme picks simplex 2 it then samples from the conditional density on that simplex using a uniform proposal. But the disparity between the uniform and the conditional density results in a large number of rejected random variables.

When the underlying density is not log-concave, the AR approach has problems because the fitted density will be log-concave, and hence have light tails. This corresponds to steep slopes on the boundary simplices of the convex hull defining the support of $\hat{f}_n$. The fat tails of the true distribution cause the q_is to be large for these simplices, which the AR scheme picks often, but does not generate from efficiently.

Jing-Hao Xue (*University College London*) **and D. M. Titterington** (*University of Glasgow*)

We congratulate the authors on this most impressive paper. In this contribution, we discuss four issues that are related to applying the LogConcDEAD method to clustering or classification problems.

First, as pointed out by the authors, a shortcoming of the LogConcDEAD method is its performance for small samples, which is mainly caused by the restriction of the support of the underlying density estimate to be the convex hull C_n, which is purely decided by the observed data. This C_n is almost inevitably a proper subset of the true support, in which case the integral $\int_{C_n} \exp\{\bar{h}_y(x)\}\,\mathrm{d}x$ in equation (3.2) is less than, not equal to, 1. To mitigate the negative effect of such an underestimated support, it is reasonable to post-process the estimated density $\hat{f}_n$. One post-processing way, as suggested by the authors in Section 9, is to use a Gaussian kernel to smooth $\hat{f}_n$. However, this leads to a virtually infinite support. Alternatively, we may consider post-processing $\hat{f}_n$ by extending the lowest exponential surfaces of $\hat{f}_n$ downwards to zero, such that a larger-than-C_n and finite support can be naturally obtained.

Secondly, classification is challenging when there is class imbalance in data: in the case of two-group discrimination, there are often a majority group and a minority group, with the size of the former being very much larger than that of the latter. We may consider using LogConcDEAD for the majority group while using a kernel-based method for the minority group. Nevertheless, it would be attractive to use LogConcDEAD throughout, if the small sample performance of LogConcDEAD is comparable with that of kernel-based methods.

Thirdly, the authors note from Table 1 an interesting pattern: the number of iterations decreases as the dimension d increases. Is this pattern influenced by the termination criteria, given that we note that in the experiments the criteria are not adaptive to the dimension d? For example, when d increases, is it possible that the integral $\int_{C_n} \exp\{\bar{h}_{y^{(l)}}(x)\}\,\mathrm{d}x$ goes to 1 faster than in the case of a smaller d? If such behaviour implies an undesired convergence, it might be better to make the parameters δ, ε or η adaptive to d; however, this leads to further complexity, which may not be worthwhile.

Finally, it is common for clustering and classification to involve data of moderate or high dimension. Therefore, for the method to be attractive the computational complexity of the LogConcDEAD method must be reduced substantially.

Kevin Lu and Alastair Young (*Imperial College London*)

This is a clever elegant paper and, reassuringly, it gives proper consideration to the question of what happens if the central assumption of log-concavity is violated. But, perhaps the authors undersell the full power of their method in these circumstances: high accuracy can often be obtained if we avoid interpreting model constraints too rigidly.

In the context of likelihood-based parametric inference, conventional aspirations of what might be achieved under model misspecification are typically limited to ensuring asymptotic validity, rather than small sample accuracy.

Let $Y = \{Y_1, \ldots, Y_n\}$ be a random sample from an underlying density $g(y)$, modelled (perhaps incorrectly) by a parametric density $f(y; \theta)$, with $\theta = (\psi, \phi)$, with scalar ψ. Let $\theta_0 = (\psi_0, \phi_0)$ maximize $T(\theta) = \int \log\{f(y; \theta)\}\, g(y)\,\mathrm{d}y$ and suppose that we test $H_0 : \psi = \psi_0$, against a one-sided alternative, say $\psi > \psi_0$.

Let $l(\theta) \equiv l(\theta; Y)$ be the log-likelihood, $\hat{\theta} = (\hat{\psi}, \hat{\phi})$ the overall maximum likelihood estimator of θ, and $\hat{\phi}_\psi$ the constrained maximum likelihood estimator of ϕ, for a fixed value of ψ. The likelihood ratio statistic is $w(\psi_0) = 2\{l(\hat{\theta}) - l(\psi_0, \hat{\phi}_0)\}$, $\hat{\phi}_0 = \hat{\phi}_{\psi_0}$, and its signed square root is $R = r(\psi_0) = \mathrm{sgn}(\hat{\psi} - \psi_0)\, w(\psi_0)^{1/2}$. The asymptotic distribution of R is $N(0, v)$ under H_0, where $v \equiv \nu(g) \neq 1$ in general. If $g(y) = f(y; \theta_0)$, $v = 1$, and an $N(0, 1)$ approximation to the distribution of R is accurate to error $O(n^{-1/2})$. This error rate can be improved to $O(n^{-3/2})$ by

(a) simulating the distribution of R under $f(y; \psi_0, \hat{\phi}_0)$ or
(b) using an adjusted form of R, such as Barndorff-Nielsen's R^*-statistic (Barndorff-Nielsen, 1986).

Under model misspecifIcation, none of the procedures is asymptotically valid. Using an estimate $\hat{v}$ of v we can, however, construct a statistic $R' = R/\sqrt{\hat{v}}$, which is asymptotically $N(0, 1)$, whether the distributional assumption is correct or not. Inference based on an $N(0, 1)$ approximation to the distribution of R' is a sensible safeguard against misspecification and still achieves an $O(n^{-1/2})$ error rate, albeit with some loss of efficiency with small n. But, we can do much better for small n by simulating the distribution of R' under the assumed (wrong) distribution, as this typically does not change much with the underlying (true) distribution g.

For example, suppose that our parametric assumption is of the inverse Gaussian distribution,

$$f(y; \mu, \lambda) = \left(\frac{\lambda}{2\pi y^3}\right)^{1/2} \exp\left\{-\frac{\lambda}{2\mu^2 y}(y-\mu)^2\right\},$$

with interest parameter the shape λ, and the mean μ as nuisance, whereas the true distribution g is gamma, scale parameter 1. Fig. 12 shows for $n = 10$ the densities of the various statistics both under the model assumption and various cases of gamma distribution g. Stability of the distribution of R', and that the distribution is far from its $N(0, 1)$ limit for $n = 10$, is apparent. In Table 5, we compare, from a series of 50 000 replications, the nominal and actual size properties of tests derived by normal approximation to the distributions of R, R' and R^* ($\Phi(R)$, $\Phi(R')$ and $\Phi(R^*)$) with those of the procedure which simulates the

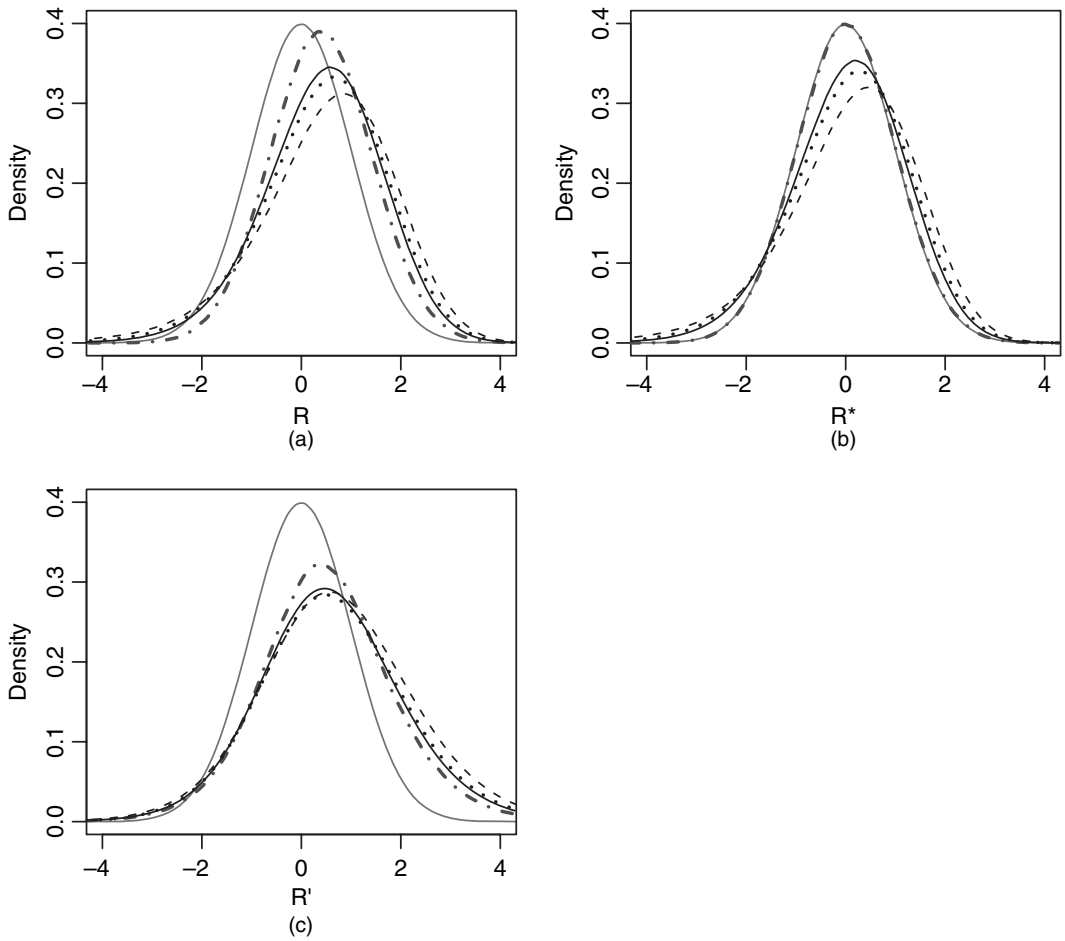


Fig. 12. Densities of statistics (a) R, (b) R^* and (c) R' under inverse Gaussian and gamma underlying distributions, $n = 10$: -------, shape = 3.0; · · · · · · ·, shape = 4.5; ———, shape = 6.0; ———, $N(0, 1)$; · - · - · -, inverse Gaussian *versus* inverse Gaussian

Table 5. Actual sizes of tests of different nominal size, inverse Gaussian shape example, for the two cases g is misspecified and g is correctly specified

	Results for the following nominal sizes:					
	0.010	*0.050*	*0.100*	*0.900*	*0.950*	*0.990*
g is gamma (scale = 1; shape = 5.5)						
$\Phi(R')$	0.022	0.056	0.090	0.696	0.776	0.886
'Bootstrap' R'	0.012	0.055	0.106	0.871	0.931	0.984
$\Phi(R^*)$	0.032	0.082	0.132	0.859	0.923	0.982
g is inverse Gaussian (mean = 1; shape = 2)						
$\Phi(R)$	0.004	0.023	0.051	0.808	0.890	0.971
'Bootstrap' R	0.010	0.050	0.101	0.902	0.950	0.990
$\Phi(R^*)$	0.009	0.050	0.099	0.900	0.950	0.990

distribution of the relevant statistic. The simulation procedure performs well compared with the $N(0, 1)$ approximation under model misspecification, though with noticeable loss of accuracy compared with the case of correct specification. Harnessing the stability of R' allows excellent small sample accuracy.

Mervyn Stone (*University College London*)

The authors of this theoretically impressive paper say that theorem 3 has, in itself, a 'desirable robustness property'. The same should therefore apply to the close analogue of theorem 3, for least squares estimation with an untrue model (Table 6).

The proviso that the data-generating f be 'not too far from' log-concavity rather begs the question of robustness. If 'robustness' means anything, it must accommodate what the real world dictates—and that is not a theoretical question. For empirical least squares, most statisticians would think that there is no such robustness in the research underlying the formulae for funding England's primary care trusts. At the heart of the Department of Health's case for reallocating £10 billion (13%) of England's primary care trust funding (Stone, 2010), there was a supposedly plausible linear model $\{\mathbf{Z}\gamma\}$ with $\mathbf{Z} = (\mathbf{V}\,\mathbf{v})$ and $\gamma^{\mathrm{T}} = (\delta^{\mathrm{T}}\varepsilon)$ in which the least squares estimate of ε (the coefficient of variable v) was, somewhat paradoxically, held to have a 'wrong', implausibly negative sign. The dependent variable y was a local measure of healthcare need. The negative sign was taken to reveal 'unmet need' in areas with high values of the socio-economic variable v, justifying a later reallocation of the £l0 billion to favour those areas. However, there is no intrinsic robustness in such a conclusion. Simply extend $\mathbf{Z}$ to $\mathbf{X} = (\mathbf{Z}\,\mathbf{a})$ by including just one of the variables omitted for one reason or another (the research did plenty of that). The consequences for the estimation of ε in the model $E(\mathbf{y}) = \mathbf{Z}\gamma + \alpha\mathbf{a}$ are not now ascertainable without a historical reanalysis of the data. The outcome would depend on both the magnitude of the omitted component $\alpha\mathbf{a}$ and its orientation to the subspace $\{\mathbf{Z}\gamma\}$—to the vector $\mathbf{v}$ in particular. The 'wrong sign' $\hat{\varepsilon}$ might be restored to 'plausible' positivity and the case for moving billions would have been weakened.

Table 6. Analogous questionable robustnesses

Step	*Density estimation of f*	*Least squares estimation of* β
I	$X_1, \ldots, X_n$ unknown true probability density function f	$E(\mathbf{Y}) = \mathbf{X}\beta$: true $\mathbf{X}$ and β
II	$\{f_0\}$: untrue ('misspecified') log-concave model	$\{\mathbf{Z}_\gamma\}$: untrue model
III	f_0^*: the f_0 that is Kullback–Leibler closest to f	$\mathbf{Z}\gamma^*$: the $\mathbf{Z}\gamma$-vector closest to $\mathbf{X}\beta$
IV	$f_n \to f_0^*$: theorem 3	$\mathbf{Z}\hat{\gamma} \to_{\mathrm{p}} \mathbf{Z}\gamma^* = E(\mathbf{Z}\hat{\gamma})$

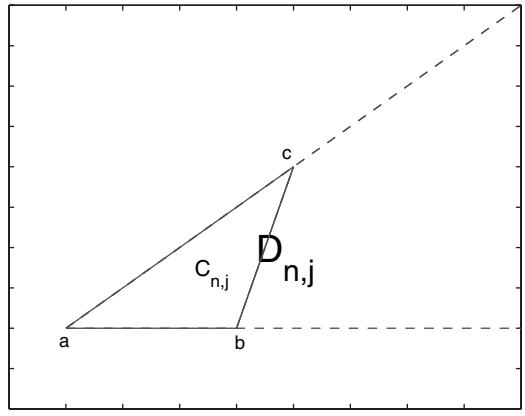


Fig. 13. Extended boundary

Yingcun Xia (*National University of Singapore*) **and Howell Tong** (*London School of Economics and Political Science*)

We congratulate the authors on their breathtaking paper. We have two questions and two comments.

(a) Non-parametric density estimation has a long history in time series analysis. There the likelihood can often be expressed in terms of a product of conditional likelihoods.
   (i) Have the authors considered how their method can be extended to cover this case without assuming any parametric models?
   (ii) If so, will their estimation still be consistent?

(b) In constructing a convex hull C_n, the tuples $C_{n,j}$ are not unique, i.e. there is another set of tuples $C'_{n,i}$ such that

$$C_n = \bigcup_{j \in J} C_{n,j} = \bigcup_{i \in I} C'_{n,i}.$$

What is the difference in the estimates that are based on different decompositions?

(c) Besides good boundary performance, it seems that LogConcDEAD can estimate the density beyond the observed region after suitable modifications. For example, consider $d = 2$. If $C_{n,j} = \mathrm{conv}(a, b, c)$ is a boundary tuple with bc being the boundary of $C_{n,j}$, we define $D_{n,j}$ as the extended tuple with the side bc removed; Fig. 13. If $C_{n,j}$ is not a boundary tuple, let $D_{n,j} = C_{n,j}$. Thus, we have

$$\mathbb{R}^d = \bigcup_{j \in J} D_{n,j}.$$

Then item (c) of Appendix B can be changed to

$$\log\{\hat{f}_n(x)\} = \langle x, b_j \rangle - \beta_j, \qquad \text{if } x \in D_{n,j} \text{ for some } j \in J.$$

Now, the density $\hat{f}_n(x)$ is well defined in the whole space $\mathbb{R}^d$.

(d) Following the late Professor Maurice Bartlett, a probability density function comes to life only if it is related to a stochastic process. Now, let $A > 0$ and ε_t be independent and identically distributed $N(0, 1)$. The self-exciting threshold auto-regressive model $X_t = -A + \varepsilon_t$ if $X_{t-1} > 0$ and $X_t = A + \varepsilon_t$ otherwise is strictly stationary with a marginal probability density function which is a mixture of $N(-A, 1)$ and $N(A, 1)$ distributions. The bimodality is related to the fact that the underlying skeleton (i.e. suppressing ε_t) is a limit cycle (Tong, 2010). Now, when a log-concave density estimate returns a unimodal distribution, a different dynamics (e.g. a limit point) results and the limit cycle is concealed. This suggests that theorem 3 could be a double-edged sword.

Ming-Yen Cheng (*University College London*)

This interesting paper establishes existence and uniqueness of a non-parametric maximum likelihood estimator (MLE) for a multi-dimensional log-concave density function and suggests the use of Shor's r-algorithm to compute the MLE. Some characterization of a multi-dimensional log-concave density

function is also given. Compared with the univariate case, for which the authors provide a comprehensive literature review, such investigations are much more challenging although of no less importance in many areas of inference. By giving illustrative examples, this paper further exploits statistical problems where multi-dimensional log-concave modelling may be useful; this includes classification, clustering, validation of a smaller (parametric) model and detecting mixing. In what follows, a few questions are raised in the hope of stimulating interest in future studies on multi-dimensional log-concave densities and applications.

Basically, log-concavity is a stronger assumption than the unimodal shape constraint. In using log-concave constrained estimators to assess suitability of a smaller model, it is sensible to ask that the model under investigation is log-concave, in which case the present approach is expected to be more powerful than using unimodal smoothing or simply non-parametric smoothing; both have been extensively studied in the literature. There are certain types of mixing that can be differentiated from log-concavity whereas other types may not be. For example, a mixture of two log-concave densities can be either log-concave or not, and the method can detect only mixtures that are no longer log-concave. Further characterization of log-concave densities and their mixtures seems necessary to gain more insight into this problem.

Shifting back to the MLE, although the authors acknowledge that the computational burden remains an issue, it is worthwhile to seek approximations that allow fast implementation or dimension reduction techniques for log-concave densities; in practice the performance deteriorates quickly when the dimension becomes larger and usually one does not go beyond $d = 3$. A question is whether the log-concavity framework allows certain simple dimension reduction transformation. Of course, studying the rate of convergence of the MLE and estimators of functionals of the density is important to understanding or assuring the performance from the theoretical viewpoint. Finally, there is an abundant literature on shape constraint estimation based on alternative approaches such as kernel smoothing and penalized likelihood approach. Interesting questions include what the differences and similarities between these different approaches are and whether ideas for one approach can be used in another.

Peter Hall (*University of Melbourne and University of California, Davis*)

This paper contains fascinating elegant results, and the authors are to be congratulated on a lovely piece of work. Of course, the paper also generates a hunger for still more, e.g. for information about the rate of convergence, but I assume that this will appear in the fullness of time. One cannot help but conjecture that, since log-concavity is essentially a property of the second derivative of the density estimator, the rate of convergence will be the same as for a kernel estimator when the density has two derivatives and the bandwidth is chosen optimally.

The implications of log-concavity for non-parametric inference are perhaps a little unclear, because the severity of the constraint seems difficult to judge. The fact that log-concavity implies unique maximization of the likelihood suggests that it is rather confining, although the authors can perhaps contradict this. Is there an interesting class of constraints that imply unique maximization of the likelihood, and for which analogues of the authors' results can be derived?

Log-concavity can be enforced by using a variety of other approaches, including the biased bootstrap and data sharpening methods of Hall and Presnell (1999) and Braun and Hall (2001) respectively. I have tried the first method in the log-concave case, and, like other applications of the biased bootstrap to impose shape on function estimators (e.g. the constraints of unimodality and monotonicity in the context of density estimation), it makes the estimator much less susceptible than usual to choice of the smoothing parameter, e.g. to the selection of the bandwidth in a kernel estimator. This property resonates with the authors' result that a log-concave density estimator can be constructed by 'maximum likelihood' without the need for a smoothing parameter.

Jon Wellner (*University of Washington, Seattle*)

I congratulate the authors on their very interesting paper. It has already stimulated considerable further research in an area which deserves much further. investigation and which promises to be useful from several perspectives.

I shall focus my comments on some possible avenues for further developments and briefly mention some related work.

An alternative to the log-concave class

The classes of hyperbolically completely monotone and hyperbolically k-monotone densities that were studied by Bondesson (1990, 1992, 1997) offer one way of introducing a very interesting family of shape-constrained densities with a range of smoothness and useful preservation properties on $\mathbb{R}$. As Bondesson (l997) showed,

(a) the hyperbolically monotone densities of order k on $(0, \infty)$ are closed under formation of products of the corresponding (independent) random variables, and hence under sums of the logarithms of these random variables in the transformed classes on $(-\infty, \infty)$,
(b) the logarithm of a random variable with a hyperbolically monotone density of order 1 has a density which is log-concave on $\mathbb{R}$ and
(c) the logarithms of the class of random variables with completely hyperbolically monotone densities yields a class of random variables which contains the Gaussian densities on $\mathbb{R}$.

These facts suggest several further problems and questions.

(i) Can we estimate a hyperbolically monotone density of order k non-parametrically for $k \geqslant 2$, and hence their natural log-transforms on $\mathbb{R}$? (For $k = 1$ such non-parametric estimators follow from the existence of non-parametric estimators of a log-concave density as studied in Dümbgen and Rufibach (2009) and Balabdaoui *et al.* (2009).)
(ii) Do there exist 'natural' generalizations of the hyperbolically k-monotone classes to $\mathbb{R}^d$ which when transformed to $\mathbb{R}^d$ include the Gaussian densities? Such classes, if they exist, would generalize the multi-dimensional log-concave class that was studied by the authors and give the possibility of trading off smoothness and dimension with smaller classes of densities offering many of the advantages of the log-concave class but with more smoothness.

These possibilities might be related to the authors' nice observation in (b) of their discussion concerning the possibility of further smoothing of the maximum likelihood estimator $\hat{f}_n$.

More on regression

Multivariate convex regression, including some work on the algorithmic side, has recently been studied by Seijo and Sen (2010).

Arseni Seregin (*University of Washington, Seattle*)

I thank the authors for their stimulating contribution to shape-constrained estimation and inference. I shall limit my comments to a brief discussion of related classes of shape-constrained families which may be of interest.

The log-concave class may be too small

As mentioned by the authors, log-concave densities have tails which decline at least exponentially fast. Larger classes of densities, the classes of s-concave densities, were introduced in both econometrics and probability in the 1970s and connected with the theory of s-concave measures by Borell (1975). A useful summary of the properties of these classes, including preservation properties under marginalization, formation of products and convolution, has been given by Dharmadhikari and Joag-Dev (l988). An initial study of estimation in such classes via likelihood methods is given in Seregin and Wellner (2010), and via minimum contrast estimation methods in Koenker and Mizera (2010). Much more research concerning properties of the estimators and development of efficient algorithms for various estimators in these classes is needed.

The log-concave class may be too big

The log-concave densities have the feature that they are based on a fixed transform (the exponential function) composed on a class of functions with a fixed degree of smoothness (namely 2) in all dimensions. Thus the entropies of these classes (or, more exactly, slightly smaller classes defined on compact connected subsets of $\mathbb{R}^d$ and satisfying an additional Lipschitz property) grow as $\varepsilon^{-d/2}$; see for example van der Vaart and Wellner (1996), corollary 2.7.10, page 164; this bound is due to Bronšteĭn (1976). This means that these classes are 'trans-Donsker' for $d \geqslant 4$, and hence the results of Birgé and Massart (1993) strongly suggest that maximum likelihood estimators will be rate inefficient for $d \geqslant 4$. Although this has not yet been proved, it raises some interesting questions for estimation in these or related classes.

(a) How can we construct rate efficient estimators of a log-concave density when $d \geqslant 4$?
(b) Can we find smaller (and smoother) classes of densities that include Gaussian densities and that are still closed under marginalization, convolution, etc.?

Theory for smooth functionals

A large number of interesting problems arise from the authors' Section 7 concerning the proposed plug-in estimators of (smooth) functionals of a log-concave density f. To the best of our knowledge the corresponding class of estimators has not been studied thoroughly even in the case of Grenander's maximum likelihood estimators of a monotone decreasing density on $\mathbb{R}^+$.

Qiwei Yao (*London School of Economics and Political Science*)

This paper provides an elegant solution to an important statistical problem. The extension to the estimation for regression functions that is presented in Dümbgen *et al.* (2010) is also attractive. I would like to make two remarks and to pose one open-ended question.

I wonder whether it is necessary for theorem 1 to assume that the observations are independent and identically distributed as the result is largely geometric. Is it enough to assume that all X_i share the same distribution? If so, the method proposed would be applicable to, for example, vector time series data.

Estimation of conditional density $f(y|x)$ is another important and difficult problem. Since $\log\{f(y|x)\} = \log\{f(y,x)\} - \log\{f(x)\}$ the method proposed provides an estimator for $f(y|x)$ by estimating $\log\{f(y,x)\}$ and $\log\{f(x)\}$ separately, and the support of the conditional density $f(\cdot|x)$ is identified as $\{y: f(y,x) > 0\}$. All those involve no smoothing.

Smoothing is a tricky technical issue in multivariate non-parametric estimation. It is associated with many practical difficulties. As illustrated in this paper, we are better off without it if possible. But, if the density function to be estimated is smooth such as having a first derivative, is it possible to incorporate this information in the algorithm?

Roger Koenker (*University of Illinois, Urbana–Champaign*) **and Ivan Mizera** (*University of Alberta, Edmonton*)

We are pleased to have this opportunity to congratulate the authors on this contribution to the growing literature on log-concave density estimation. Having begun to explore regularization of multivariate density estimation via concavity constraints several years ago (Mizera and Koenker, 2006), we can also sympathize with the prolonged gestation period for publication of such work.

We feel that the authors may be too pessimistic about Newton-type methods when rationalizing their gradient descent approach to computation. Interior point algorithms for convex optimization have been remarkably successful in adapting barrier function methods to a variety of non-smooth problems and employing Newton steps. Linear programming has served as a prototype for these developments, but there has been enormous progress throughout the full range of convex optimization.

Our computational experience has focused on finite difference methods that impose both the concavity and the integrability constraints on a grid with increments controlling the accuracy of the approximation. Even on rather fine grids this approach combined with modern interior point optimization is quite quick. For the bivariate example in Koenker and Mizera (2010) with 3000 points, computing the Hellinger estimate subject to the $f^{-1/2}$ concavity constraint takes about 23 s, whereas the maximum likelihood estimate with the log-concavity constraint required 45 min on the same machine with the LogConcDEAD package implementing the authors' algorithm.

The authors express the hope that their results for maximum likelihood estimation of log-concave densities may offer ideas that can be transferred to more general settings. Koenker and Mizera (2010) establish a polyhedral characterization, which is kindred to that exemplified by Fig. 1, for a class of maximum entropy estimators imposing concavity on corresponding transformations of densities. Particular special cases include maximum likelihood estimation of log-concave densities; the instance that we find especially appealing amounts to minimizing a Hellinger entropy criterion for densities f, such that $f^{-1/2}$ is concave. This class of densities covers the Student t_ν-densities with degrees of freedom $\nu \geqslant 1$. Whether any similar polyhedral representation holds for maximum likelihood estimation subject to such concavity requirements, as recently proposed by Seregin and Wellner (2010), is not clear.

In view of this common polyhedral characterization, it would be interesting to know whether the Shor approach can be adapted to this broader class of quasi-concave estimation problems. We are looking forward to the authors' opinion on this.

The following contributions were received in writing after the meeting.

Christoforos Anagnostopoulos (*University of Cambridge*)

The multi-dimensional density estimator that is proposed in this work is a key contribution in the field of non-parametric statistics, owing to its automated operation, computational simplicity and theoretical properties. It represents the culmination of a recent body of work on log-concave probability densities and will certainly stimulate further research into the properties and applications of shape-constrained estimators.

The authors mention classification and clustering as two possible application areas of their method. Indeed, non-parametric class descriptions (or cluster descriptions) have been an increasingly active area of

research in the machine learning community (e.g. Fukunaga and Mantock (1983) and Roberts (1997)). A further challenge that such algorithms face when deployed in realtime environments is the need to process data *on line* without revisiting the data history. Unfortunately, the requirement of a constant time update clearly clashes with the infinite dimensional nature of non-parametric estimators such as that proposed in the paper. It is consequently of great practical interest to investigate the extent to which an on-line approximation could be devised.

A working candidate may be constructed readily, by performing a fixed number of iterations of Shor's r-algorithm per time step, initialized at the previous time step's pole heights estimates. In Section 3.2, the number of iterations required for convergence in the off-line case is reported to increase approximately *linearly* with n. This suggests that, on arrival of each novel data point, a constant number of iterations may indeed suffice for convergence, but early stopping may be employed if necessary. To handle the increasing sample size, we may fix the number of pole heights to a constant number w. In an on-line context, this means dynamically maintaining an active set of w data points, and replacing (at most) one data point per time step. The selection of which data point to replace could be arbitrary (as in a *sliding window* where, at time n, x_{n-w} is replaced by x_n) geometric (for example replace the data point whose removal has the smallest effect on the shape of the estimator) or information theoretic.

Similar work on sequential kernel density estimation has attracted great attention in the machine learning community (e.g. Han *et al.* (2007)). Notably, the lack of bandwidth parameters for the estimator of Cule, Samworth and Stewart represents a crucial comparative advantage, even more so in on-line than in off-line contexts. There is consequently little doubt that a theoretical argument concerning the error in the approximation above as a function of w and the data selection mechanism would be of great interest. Finally, it should be noted that the extension to on-line estimation of mixtures of log-concave densities can be handled by using recent work on on-line expectation–maximization (Cappé and Moulines, 2009).

Dankmar Böhning (*University of Reading*) **and Yong Wang** (*University of Auckland*)

We congratulate the authors on their excellent contribution to multivariate non-parametric density estimation under a log-concavity restriction. This restriction appears to be quite realistic for many practical problems and we expect to see many successful applications of this new methodology.

We acknowledge the authors' detailed use of convex geometry that led to the existence and uniqueness of the log-concave maximum likelihood estimator (LCMLE). Evidently, the algorithmic approach still lacks computational power as their Table 1 indicates and there is likely room for improvements; see Wang (2007) for a fast algorithm on non-parametric mixture estimation, which is a problem that is somewhat related. Also, the authors point out that the kernel density estimator is a natural competitor but has difficulty with bandwidth selection, especially in the multivariate case.

We are also intrigued by the clustering example that the authors discuss in Section 6. For a long time, the mixture community has been looking for a non-parametric replacement for the parametric mixture component distribution. This paper gives a very interesting new solution to this problem. It appears to us that the misclassification rate might be competitive only in comparison with the Gaussian mixture, if cross-validation assessment is used. We also wonder whether the new method can be extended to non-parametric clustering.

Finally, we have explored the predictive performance of the LCMLE in the setting of *supervised learning* and compared it with that of three others: a Gaussian density estimator, logistic regression and a kernel estimator. The same Wisconsin breast cancer data set as shown in Fig. 6(a) was used, but for classification purposes here. Except for logistic regression, which is fitted to all observations, observations in each class are modelled by each specific distribution estimator. A *new* observation is then classified according

Table 7. Numbers of misclassified observations, with standard errors in parentheses

Method	*Parametric results*		*Non-parametric results*	
	Gaussian	*Logistic*	*Kernel*	*Log-concave*
Resubstitution	35	25	32	25
Cross-validation	37.1 (0.35)	27.5 (0.30)	37.5 (0.21)	37.1 (0.32)

to its posterior probability. For bandwidth selection of the kernel estimator, we simply use Silverman's rule of thumb, $h_j = n^{-1/6} s_j$, $j = 1, 2$, where s_j is the sample standard deviation along the jth co-ordinate (Silverman (1986), page 87).

Table 7 gives both the resubstitution and the tenfold cross-validation (averaged over 20 replications) classification errors. Despite its low resubstitution error, the LCMLE performs only comparatively with the less appealing kernel estimator and the apparently biased Gaussian estimator in terms of cross-validation error. With only a small increase from its resubstitution error, the parametric logistic regression gives a remarkably smaller prediction error than the other three. Note that the LCMLE is far more expensive to compute than the others.

It is likely that the fair performance of the LCMLE in this example is due to its non-parametric nature and truncation of the density to zero outside the convex hull of the training data. This increases the estimation variance, which may significantly outweigh its reduced bias.

José E. Chacón (*Universidad de Extremadura, Badajoz*)

First of all, I congratulate the authors on their thorough and interesting paper. Sometimes, the extension of a univariate technique to its multivariate analogue is taken as a trivial or incremental step, but most of the time this is not so, and this paper provides a nice example of the latter type of advance.

Even if the present study is quite exhaustive. I would just like to add further avenues for future research to those which the authors already propose.

Although most of the literature on maximum likelihood (ML) estimation of the density is devoted to the univariate case, there is another recent reference which provides a multivariate method: Carando *et al.* (2009). There, the class of densities is constrained to be Lipschitz continuous, so the problem is of a different nature, but both resulting ML estimators present some similarities in shape. In the univariate case the two estimators are piecewise linear, although the log-concave estimator allows for different slopes; in contrast, the Lipschitz estimate is not necessarily unimodal. In any case, the connections between the two methods surely deserve to be explored.

Probably an undesirable feature of the ML estimate is that it is not smooth (i.e. differentiable). Dümbgen and Rufibach (2009) amended this via convolution, but perhaps it would be more natural in this setting to investigate the ML estimator imposing some smoothness condition on the class of log-concave densities. In the univariate case, for instance, we could think of a smoothness constraint leading to a piecewise quadratic or cubic (instead of linear) ML estimate.

Another possible research direction points to a comparison with kernel methods. I agree with the authors that general bandwidth matrix selection is a difficult task, yet the plug-in method that was recently introduced in Chacón and Duong (2010) looks promising from a practical point of view, being the multivariate analogue of the method by Sheather and Jones (1991). On the theoretical side, it would be interesting to obtain the mean integrated squared error rates (and the asymptotic distribution) for the multivariate log-concave ML estimator, since it seems from the simulations that they might be faster than for the kernel estimator. Nevertheless, in the supersmooth case of, say, the standard d-variate normal density, it looks like this rate should be slower than $n^{-1}\log(n)^{d/2}$, which can be deduced to be the rate for a superkernel estimator, reasoning as in Chacón *et al.* (2007).

Yining Chen (*University of Cambridge*)

I congratulate the authors for developing an innovative and attractive method for non-parametric density estimation. The power of this method was well illustrated in their simulation examples by comparing the mean integrated squared error (MISE) with other kernel-based approaches. To improve its performance at small sample sizes, the authors proposed a smoothed (yet still fully automatic) version of their estimator via convolution in Section 9. Below, we give some justification for this new estimator and argue that it has some favourable properties.

We consider the same simulation examples as in Section 5, for $d = 2$ and $d = 3$, and for small to moderate sample sizes $n = 100, 200, 500$. Results are given in Figs 14 and 15. We see that for cases (a), (b), (d) and (e), where the true density is log-concave and has full support, the smoothed log-concave maximum likelihood estimator has a much smaller MISE than the original estimator. The improvement is most significant (around 60%) for $d = 3$ with small sample sizes, i.e. $n = 100$ and $n = 200$, but is still around 20% even when $d = 2$ and $n = 500$. Interestingly, this new estimator outperforms most kernel-based estimators (including those based on MISE optimal bandwidths, which would be unknown in practice) even at small sample sizes, where the original estimator performs relatively poorly. As shown in case (f), even if the log-concavity assumption is violated, the smoothing process still offers some mild reduction in MISE for small sample sizes.

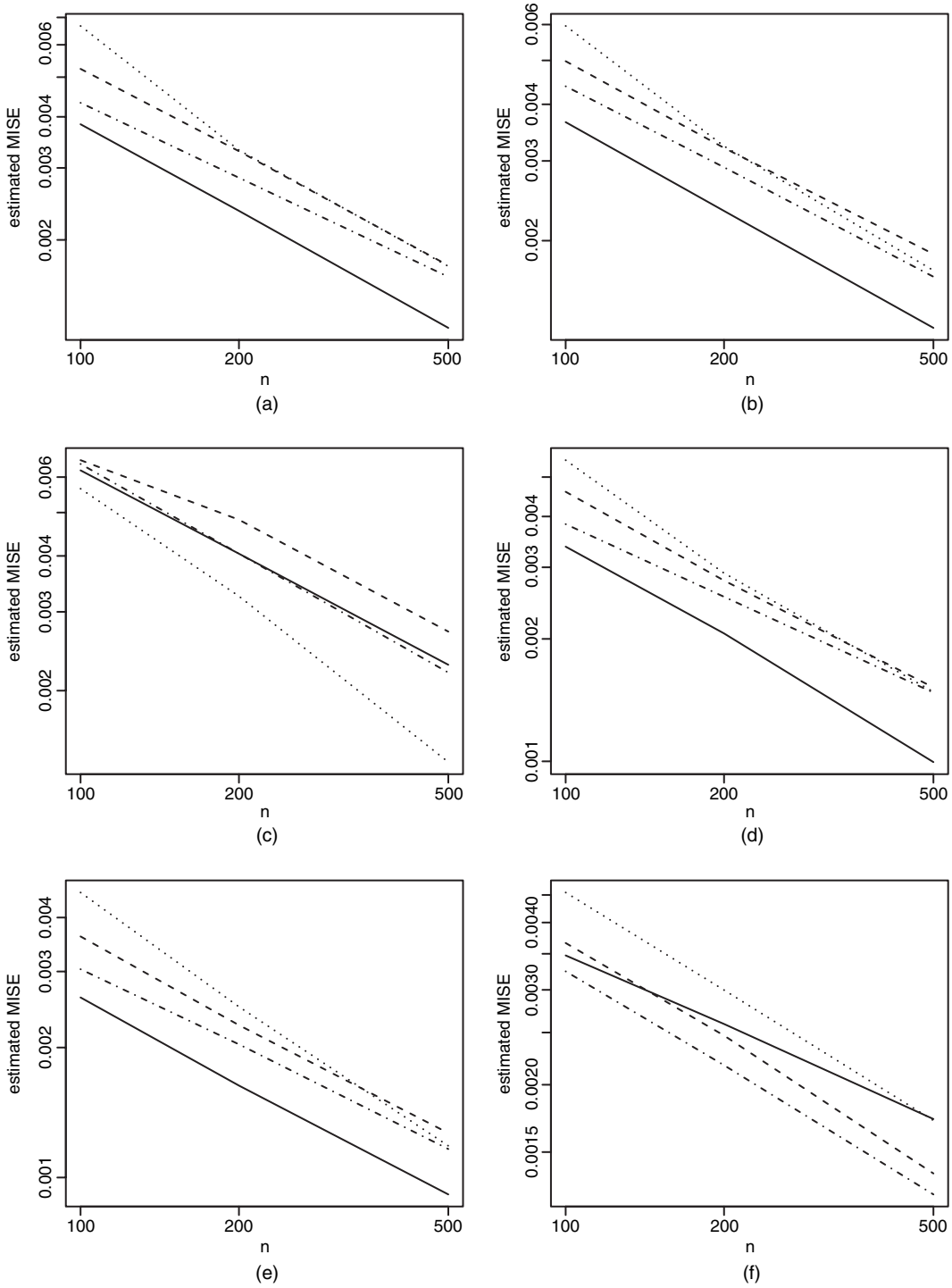


Fig. 14. MISE, $d = 2$: ———, smoothed LogConcDEAD estimate; · · · · · · ·, LogConcDEAD estimate; - - - - - - -, plug-in kernel estimate; - · - · - ·, MISE optimal bandwidth kernel estimate

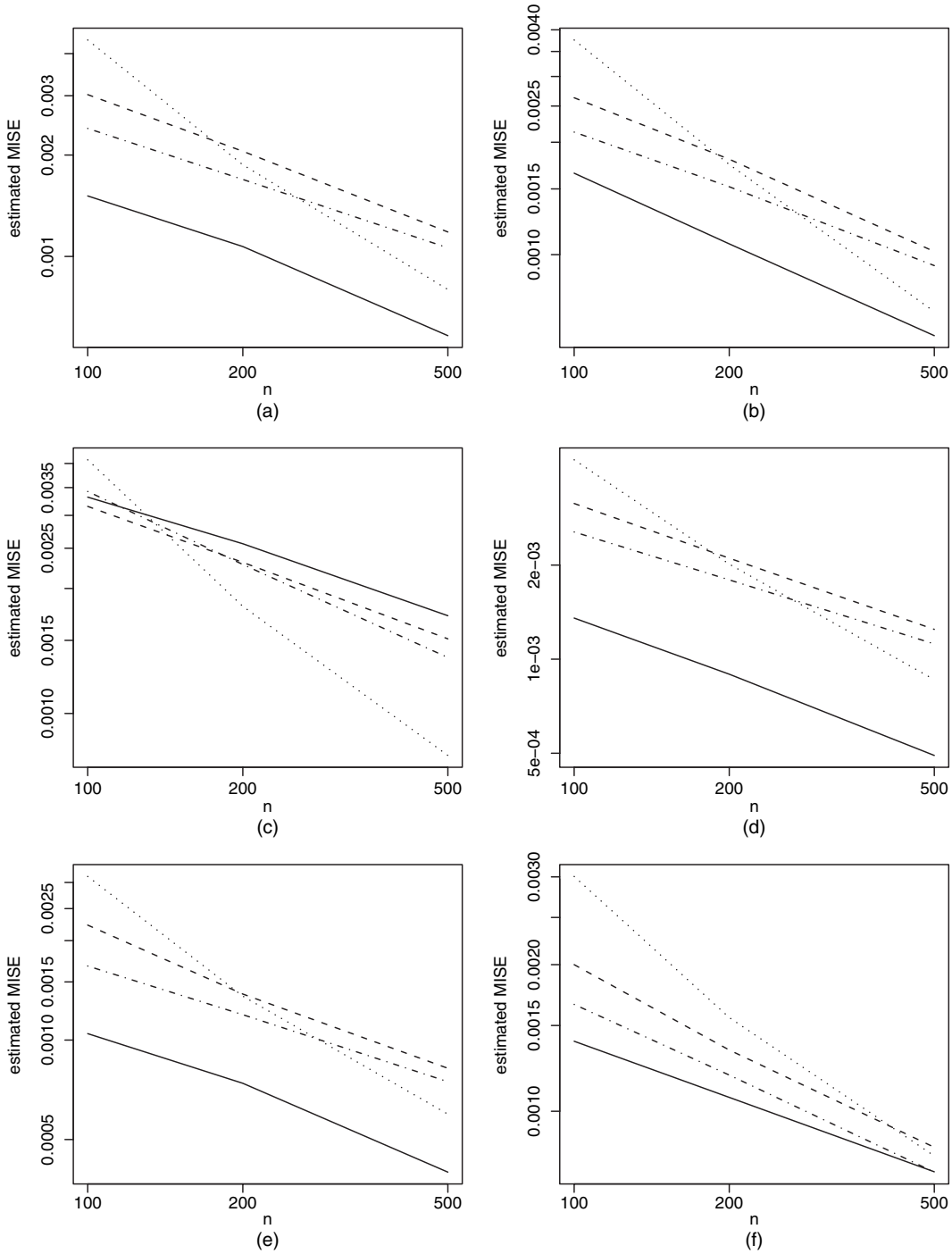


Fig. 15. MISE, $d = 3$: ———, smoothed LogConcDEAD estimate; · · · · · · ·, LogConcDEAD estimate; - - - - - - -, plug-in kernel estimate; - · - · - · -, MISE optimal bandwidth kernel estimate

However, as demonstrated in case (c), this modification can sometimes lead to an increased MISE at large sample sizes. This is mainly due to the boundary effect. Recall that in case (c) the underlying gamma distribution does not have full support. Convolution with the multivariate normal distribution shifts some mass of the estimated density outside the support of the true distribution and thus results in a higher MISE. It is a nice feature of the original estimator that it handles cases of restricted support effectively and automatically.

Finally, we note that the smoothed log-concave maximum likelihood estimator also offers a natural way of estimating the derivative of a density.

Frank Critchley (*The Open University, Milton Keynes*)

It is a great pleasure to congratulate the authors on a splendid paper: I only regret that I could not be there to say this in person!

Like all good papers read to the Society, its depth and originality raise many interesting further questions. The authors themselves allude to a variety of these, implicitly if not explicitly, and I hope that they will forgive any overlap with the following.

(a) What can be said about *which* (mixtures of) shapes admit maximum likelihood estimators?
(b) With log-concavity as target, what scope is there for transformation–retransformation methods?
(c) Notwithstanding the overall thrust of the paper, are there contexts in which there is some advantage to *smoothing* the maximum likelihood estimator that is produced?
(d) Are there potential links with dimension reduction methods in regression?

Jörn Dannemann and Axel Munk (*University of Göttingen*)

We congratulate the authors for their very interesting and stimulating paper which demonstrates that multivariate estimation with a log-concave shape constraint is computationally feasible. Conceptually, this approach seems very appealing, since it is much more flexible than parametric models, but sufficiently restrictive to preserve relevant data structures. Further, we believe that the extension to finite mixtures of log-concave densities for clustering as addressed in Section 6 is of particular practical importance.

As for classical mixture models identifiability is essential for model analysis and interpretation and as almost nothing is known for log-concave models we would like to comment on this issue here. First, note that classical parametric mixtures, namely mixtures of multivariate Gaussian (log-concave) or t-distributions (not log-concave), are identifiable (Yakowitz and Spragins, 1968; Holzmann *et al.*, 2006; Dümbgen *et al.*, 2008).

For non-parametric mixture models identifiability has been investigated by Hall and Zhou (2003) and Allman *et al.* (2009) for multivariate observations and Hunter *et al.* (2007) and Bordes *et al.* (2006) for univariate data under specific assumptions such as independence of components or symmetry.

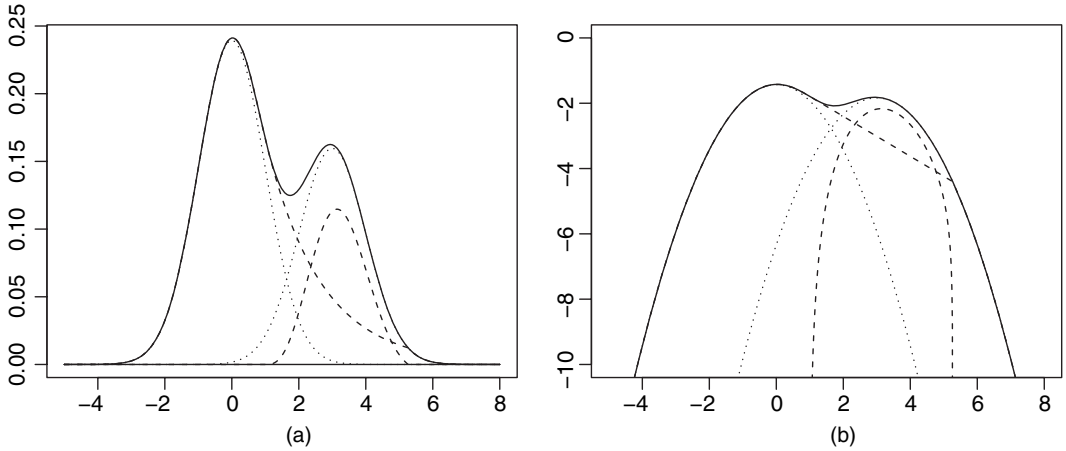
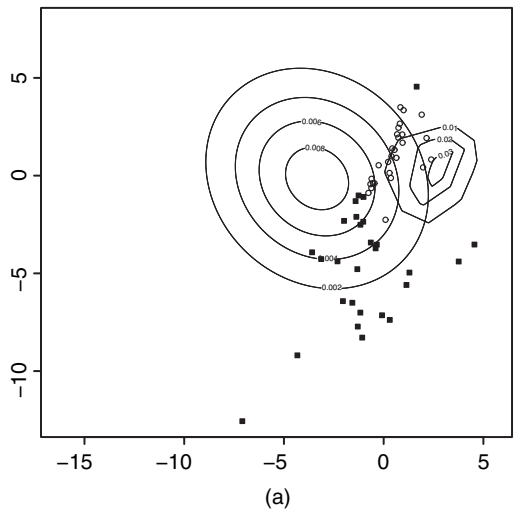


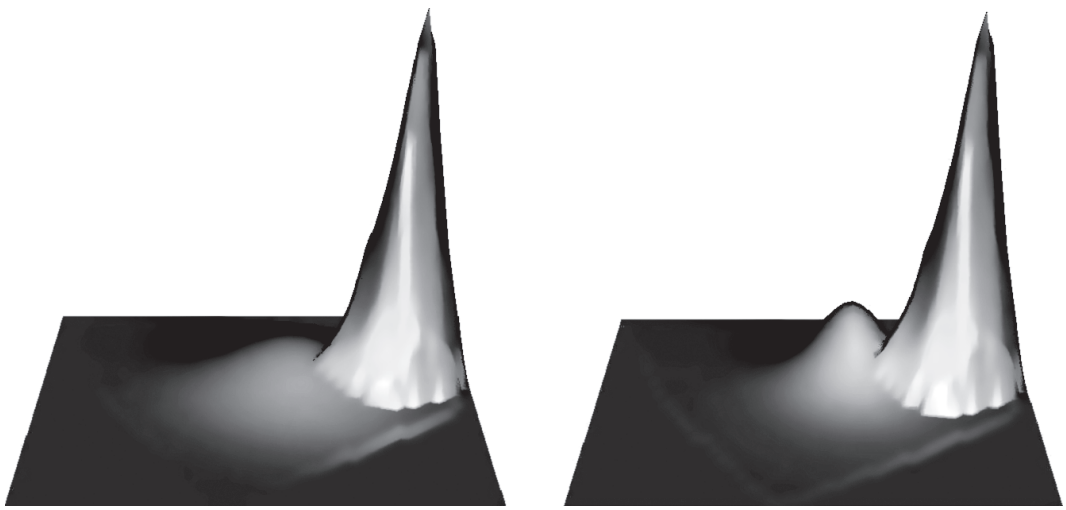
Fig. 16. (a) Density of $f = 0.6\,\mathcal{N}(0,1) + 0.4\,\mathcal{N}(3,1)$ (———) and its components (·····) (modifying $0.6\,\mathcal{N}(0,1)$ and $0.4\,\mathcal{N}(3,1)$ without violating non-negativity and log-concavity yields a remarkable different representation (-------) of the mixture with mixture proportion $\pi = 0.77$) and (b) logarithms of the functions

As pointed out by the authors mixtures of log-concave densities are possibly log-concave themselves. In this case the mixture proportions and components are generically not identifiable. Besides this extreme situation non-identifiability seems more severe for mixtures of log-concave densities than for comparable concepts, where typically only a small subset of the parameter space is not identifiable (for example see Bordes *et al.* (2006)). In contrast, for mixtures $f = \pi f_1 + (1-\pi) f_2$ with f_1 and f_2 strictly log-concave and not completely separated, we can perturb f_1 a little by some function h such that $\tilde{f}_1$ and $\tilde{f}_2$ with $\tilde{f}_1 = \pi f_1 + h$ and $\tilde{f}_2 = (1-\pi) f_2 - h$ are non-negative log-concave functions such that its normalized versions yield a different representation of f. Two different representations of a univariate Gaussian mixture are displayed in Fig. 16.

We would like to draw the authors' attention to mixture models, where one component is modelled as a log-concave density, whereas the others belong to some parametric family, i.e. a Gaussian or t-distribution. For example consider a two-component mixture with a log-concave f_{LC} and a Gaussian component f_{Gauss} This model is identifiable if there is an interval I for which $I \cap \mathrm{supp}(f_{\mathrm{LC}}) = \emptyset$ is a *priori* known.



(a)



(b) (c)

Fig. 17. (a) Contour plot with misclassified instances, (b) estimated mixture with a log-concave and Gaussian component and (c) estimated mixture with a log-concave and t-distributed component (with 3 degrees of freedom) from the EM algorithm

The EM algorithm that was suggested by the authors can easily be adapted to this semiparametric model. Applying it to the Wisconsin breast cancer data in the way that the component that is associated with the malignant cases is modelled as a Gaussian (or multivariate t-distribution) shows that it is intermediate between the purely Gaussian and the purely log-concave EM algorithm with 55 misclassified instances (51 for t-distributions with $\nu = 3$ degrees of freedom). The estimated mixture densities are presented in Fig. 17.

David Draper (*University of California, Santa Cruz*)

The potential usefulness of this interesting paper is indicated by, among other things, the existence of the rather infelicitously named LogConcDEAD package in R that the authors have already made available, for implementing their point estimate of an underlying data-generating density f. I would like to suggest a potentially fruitful area of future work by adding to the paper's reference list a few pointers into the Bayesian non-parametric density estimation literature; this may be seen as a possible small sample competitor, to a bootstrapped version of the authors' point estimate, in creating well-calibrated uncertainty bands for density estimates and functionals based on them. This parallel literature dates back at least to the early 1960s (Freedman, 1963, 1965; Ferguson, 1973, 1974) and has burgeoned since the advent of Markov chain Monte Carlo methods (Escobar and West, 1995): main lines include Dirichlet process mixture modelling (e.g. Hjort *et al.* (2010)) and (mixtures of) Pólya trees (e.g. Hanson and Johnson (2001)). Advantages of the Bayesian non-parametric approach to density estimation include

(a) the automatic creation of a full posterior distribution on the space $\mathcal{F}$ of all cumulative distribution functions, with built-in uncertainty bands arising directly from the Markov chain Monte Carlo sampling, and
(b) a guarantee of asymptotic consistency of the posterior distribution in estimating f (when the prior distribution on $\mathcal{F}$ is chosen sensibly: see, for example Walker (2004)) whether f is log-concave or not.

From the authors' viewpoint, with their emphasis on the lack of smoothing parameters in their point estimate, disadvantages of the Bayesian approach may include the need to specify hyperparameters in the construction of the prior distribution on $\mathcal{F}$, which act like user-specified tuning constants. The small sample performance—both in terms of calibration (e.g. nominal 95% intervals include the data-generating truth x% of the time; $x = ?$) and of useful information obtained per central processor unit second—of these two rather different approaches would seem to be an open problem that is worth exploring.

Martin L. Hazelton (*Massey University, Palmerston North*)

Non-parametric density estimation in high dimensions is a difficult business. It is therefore natural to look at restricted versions of the problem, e.g. by placing shape constraints on the target density f. The authors are to be congratulated on their progress in the case where f is assumed to be log-concave. I offer two (loosely connected) comments on this work: the first with regard to practical performance for bivariate data, and the second to suggest an alternative test for log-concavity.

I would expect the log-concave maximum likelihood estimator to improve markedly on kernel methods when the data are highly multivariate. However, the situation is less clear for bivariate data, where the curse of dimensionality has not really begun to bite. In that important case, kernel estimation using plug-in bandwidth selection is generally very competitive against the log-concave maximum likelihood estimator for $n \leqslant 500$, and only slightly worse when $n = 2000$. Arguably the extra smoothness properties of the kernel estimate are a fair swap for the small loss in performance with respect to mean integrated squared error. The only bivariate setting in which the log-concave maximum likelihood estimator appears much better is for test density (c). However, this is almost certainly a result of boundary bias in the kernel estimator, for which corrections are available (e.g. Hazelton and Marshall (2009)).

Of course, if we are convinced that f is log-concave then kernel estimation with a standard bandwidth selector may be unattractive because it is not guaranteed to produce a density of that form. However, if the kernel is log-concave then so also will be the density estimate for sufficiently large bandwidth h, although this might result in significant oversmoothing from most standpoints. This observation motivates a test for log-concavity of f.

Suppose that we construct a kernel estimate by using an isotropic Gaussian kernel. Then there will be a (scalar) bandwidth $h_0 > 0$ such that the estimate $\hat{f}$ will be log-concave if and only if $h \geqslant h_0$ (because log-concavity is preserved under convolution). This bandwidth is a plausible test statistic for log-concavity, since the larger its value the more we have had to (over)smooth the data to enforce log-concavity. This idea mirrors the bump hunting test that was developed by Silverman (1981). Following Silverman's approach,

bootstrapping could be employed to test significance, although for practical application it would be necessary to refine the basic methodology to mitigate the effects of tail wiggles that are generated by isolated data points.

Woncheol Jang (*University of Georgia, Athens*) **and Johan Lim** (*Seoul National University*)

We congratulate the authors for an interesting and stimulating paper. The methodology in the paper is well supported in theory and is nicely applied to classification and clustering. Here we consider the application of the proposed method to bagging, which is popularly used in the machine learning literature.

The main idea of bagging (Breiman, 1996) is to use a committee network approach. Instead of using a single predictor, bootstrap samples are generated from the original data and the bagged predictions are calculated as averages of the models fitted to the bootstrap samples.

Clyde and Lee (2001) proposed a Bayesian version of bagging based on the Bayesian bootstrap (Rubin, 1981) and proved a variance reduction under Bayesian bagging. A key idea of Bayesian bagging is to use *smoothed* weights for the bootstrap samples whereas the weights in the original bagging can be considered as being generated from a discrete multinomial$(n; 1/n, \ldots, 1/n)$ distribution.

Other related ideas are output smearing of Breiman (2000) and input smearing of Frank and Pfahringer (2006). They suggested adding Gaussian noise to the output and input respectively and applied the bagging to these noise-added data sets. Both smearing methods were shown empirically to work very well in their simulation studies. However, the optimal magnitude (the variance) of the noise to be added is not well understood.

The idea behind smearing methods is indeed equivalent to generating resamples with the *smoothed* bootstrap and the issue of the choice of magnitude of the noise is the same as that of bandwidth selection of the multivariate kernel density estimator that is used in the smoothed bootstrap procedure. In Bayesian bagging, there is a similar issue with the choice of the hyperparameter of the Dirichlet prior that is used in the Bayesian bootstrap.

An advantage of the proposed method against the aforementioned methods is that it needs no tuning. The authors also propose a procedure to sample from the estimated log-concave density in Appendix B.3. Thus, bagging based on resamples from the estimated log-concave density would be a good alternative to the Bayesian bagging or smearing methods.

Hanna K. Jankowski (*York University, Toronto*)

I congratulate the authors on an important and thought-provoking paper. This work will certainly be a catalyst for further research in the area of shape-constrained estimation, and the authors themselves suggest several open problems towards the end of the paper. I shall restrict my discussion to adding another question to this list.

One of the identifying features of non-parametric shape-constrained estimators is their rates of convergence, which are slower than the typical $n^{1/2}$-rate that is achieved by parametric estimators. In one dimension, the Grenander estimator of the decreasing density converges at a local rate of $n^{1/3}$ whereas the estimator of a convex decreasing density converges locally at rate $n^{2/5}$ (Prakasa Rao, 1970; Groeneboom, 1989; Groeneboom *et al.*, 2001). A similar rate is seen for the one-dimensional non-parametric maximum likelihood estimator of a log-concave density, which was recently proved to be $n^{2/5}$, as long as the density is strictly log-concave (Balabdaoui *et al.*, 2009). A heuristic justification of how different local rates arise has been given by Kim and Pollard (1990). The global convergence rates, in contrast, can be quite different. For the Grenander estimator, the convergence rate for functionals $\varphi(\hat{f}_n - f_0)$ is known to be

$$n^{1/6}\{n^{1/3}\varphi(\hat{f}_n - \hat{f}_0) - \mu_\varphi(f_0)\} \Rightarrow \sigma_\varphi(f_0)Z,$$

where Z is a standard normal random variable (Groeneboom, 1985; Groeneboom *et al.*, 1999; Kulikov and Lopuhaä, 2005). Here, f_0 denotes the true underlying monotone density. Thus, smooth functionals with $\mu_\varphi(f_0) = 0$ (such as plug-in estimators of the moments) converge at rate $n^{1/2}$ and recover the faster rate characteristic of parametric estimators.

Global and local convergence rates for the log-concave non-parametric maximum likelihood estimator are sure to be of much interest in the near future. Indeed, it is already conjectured in Seregin and Wellner (2009) that the local convergence rate for the estimator $\hat{f}_n$ that is introduced here is $n^{2/(4+d)}$ when $d = 2$, 3. In Section 7, the authors consider plug-in estimators of the moments or the differential entropy for $\hat{f}_n$. What would the convergence rate be for these functionals? Preliminary simulations for $d = 1$ indicate that the $n^{1/2}$-rate may continue to hold for the log-concave maximum likelihood estimators (Fig. 18). Further

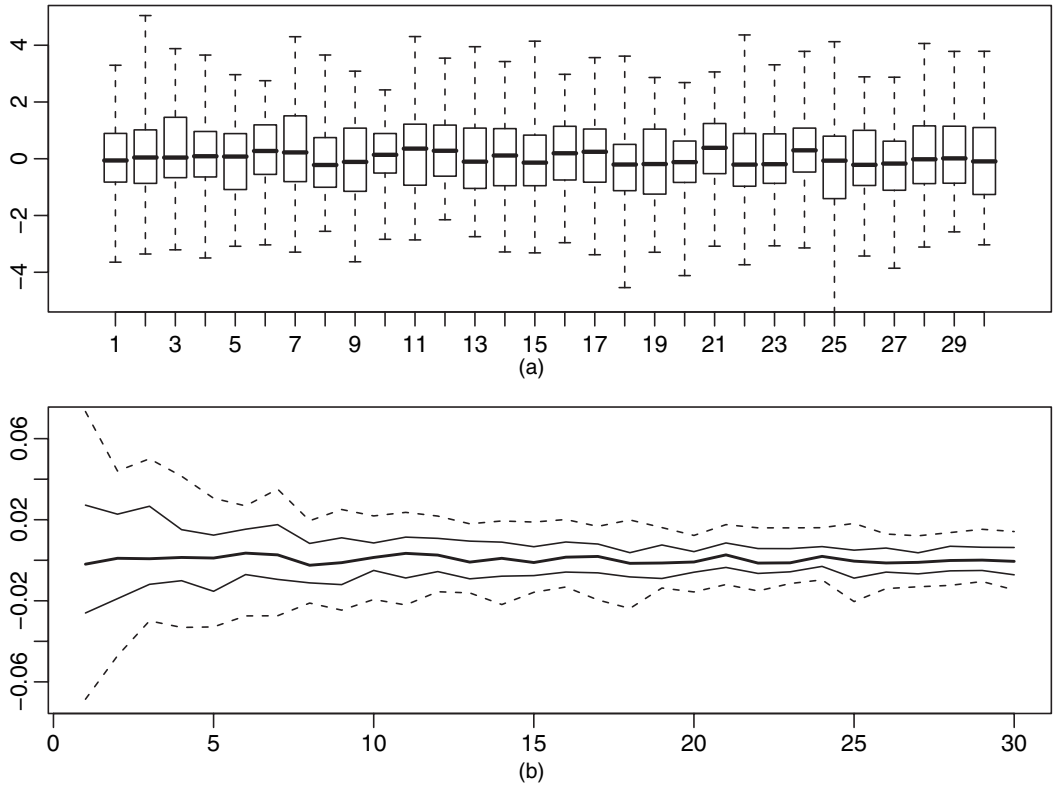


Fig. 18. (a) $n^{1/2}$ rescaled functional *versus* sample size (in thousands) (the non-parametric maximum likelihood estimate of a gamma(2,1) random variable was computed, by using Rufibach and Dümbgen (2006), and the centred mean functional was calculated on the basis of the estimated density; each boxplot consists of $B = 100$ simulations) and (b) quantiles *versus* sample size (in thousands) (quantiles of the unscaled and centred functionals (-------, 0.05 and 0.95; ——, 0.25 and 0.75; ———, median)): a regression of the logarithm on the 0.05 and 0.95 quantiles on the logarithm of the sample size yields a highly signifancant slope estimate of −0.48968

investigation is needed in higher dimensions. A rate of $n^{1/2}$ would, naturally, be very attractive in the application of these methods.

Theodore Kypraios and Simon P. Preston (*University of Nottingham*) **and Simon R. White** (*Medical Research Council Biostatistics Unit, Cambridge, and University of Nottingham*)

We congratulate the authors for this interesting paper. In this discussion, we would like to hear the authors' views on the applicability of their approach in the following context.

Suppose that we interested in a distribution whose probability density function, say $p(x)$, is proportional to a product of other probability density functions $f_i(x)$, $i = 1, \ldots, k$, i.e.

$$p(x) = c \prod_{i=1}^{k} f_i(x) \tag{5}$$

with c being a normalizing constant. Suppose that none of the $f_i(x)$ is known explicitly but that we can draw *independent and identically distributed samples* from each. How should we best calculate functionals of $p(x)$, or draw samples from it?

White *et al.* (2010) consider an exact method of sampling-based Bayesian inference in the context of stochastic population models. This gives rise to a posterior distribution of the parameters of the form (5). Their approach is to use a kernel density estimator for each $f_i(x)$, and then to estimate $p(x)$ as the normalized pointwise product of kernel density estimators. But if the $f_i(x)$ are log-concave then would the methodology that is presented in this paper provide a better alternative? If so, then a clear advantage would

be that we could draw samples from $p(x)$ by making use of the rejection sampling method in Appendix B.3 of this paper. Can the authors comment on the applicability of their method to product densities such as density (5), in particular on issues as k increases?

Chenlei Leng (*National University of Singapore*) **and Yongho Jeon** (*Yonsei University, Seoul*)

Multi-dimensional density estimation without any parametric distributional assumption is known to be difficult. We congratulate Cule, Samworth and Stewart for an impressive piece of work, in which they show that log-concavity is an attractive option compared with non-parametric smoothing. Here we focus on an alternative formulation, which may greatly facilitate numerical implementation. In the following discussion, we use the notation that is used in the paper.

Consider an alternative objective function to function (3.1),

$$\frac{1}{n}\sum_{i=1}^{n}\exp\{-g(X_i)\}+\int_{C_n} g(x)\,\mathrm{d}x, \tag{6}$$

where g is a concave function. Jeon and Lin (2006) showed that its population minimizer is $g=\log(f_0)$. It is easy to see that the sample minimizer of this function is a least log-concave function. An application of theorem 2 in the paper leads to our alternative formulation

$$\omega(y_1, y_2, \ldots, y_n)=\frac{1}{n}\sum_{i=1}^{n}\exp(-y_i)+\int_{C_n}\bar{h}_y(x)\,\mathrm{d}x. \tag{7}$$

Following Appendix B, it is easy to see that the subgradient corresponding to equation (B.l) can be written as

$$\partial_i\omega(y)=-\frac{1}{n}\exp(-y_i)+\sum_{j\in J_i}|\det(A_j)|\int_{T_d}\sum_{l=0}^{d}w_l\mathbb{I}_{\{j_l=i\}}\,\mathrm{d}w.$$

Note that this formulation requires $\int_{T_d} w_1\,\mathrm{d}w=1/(d+1)!$ to be computed *only once, precisely*, whereas the authors require $\tilde{I}_{d,u}$ defined in Appendix B.2 to be computed iteratively, and to use Taylor series expansion in approximating the integral to avoid singularity problems if necessary. The new formulation can be straightforwardly extended to other types of constrained density estimation, by replacing the function $\exp\{-g(x)\}$ in expression (6) with some appropriate function $\psi\{g(x)\}$, which can be formulated to correspond to the quasi-concave function in Koenker and Mizera (2010) or the convex-transformed density in Seregin and Wellner (2010). The computation of our estimator remains effectively the same with respect to the integral.

Another interesting problem is to introduce structures to the density. For example, we may decompose the log-density as an analysis-of-variance model by writing

$$\log(f)=h_0+\sum_{j=1}^{d}h_j+\sum_{j<k}h_{jk},$$

where h_js are the main effects and h_{jk}s are the two-way interactions. Higher order interactions may be considered as well. Some side-conditions are assumed to assure the identifiability of this decomposition. It is known that $h_{jk}=0$ corresponds to conditional independence of the jth variable and the kth variable. The conditional independent structure corresponds to a graphical model, as discussed by Jeon and Lin (2006). Using our formulation, we may decompose y_i as

$$y_i=y_0+\sum_j y_{i,j}+\sum_{j<k}y_{i,jk}$$

and minimize expression (7) with this decomposition. For graphical model building, we may apply the group lasso penalty $\lambda\Sigma_{j<k}\sqrt{(\Sigma_i y_{i,jk}^2)}$ (Yuan and Lin, 2006) which can estimate $y_{i,jk}, i=1,\ldots,n$, as 0.

In on-going work, we are investigating this new density estimator and will report the result elsewhere.

Dominic Schuhmacher (*University of Bern*)

It was a pleasure to read this interesting and elegant paper that covers so much ground on multivariate log-concave density estimation. I would like to comment on two central points.

First, the log-concave maximum likelihood estimator that is studied by the authors may be written as the unique function

$$\hat{f}_n = \hat{f}_n(\cdot|\hat{P}_n) \in \arg\max_{\tilde{f}\in\mathcal{F}_0}\left[\int \log\{\tilde{f}(x)\}\,\hat{P}_n(\mathrm{d}x)\right],$$

where

$$\hat{P}_n = \frac{1}{n}\sum_{i=1}^{n}\delta_{X_i}$$

denotes the empirical distribution of $X_1, \ldots, X_n \in \mathbb{R}^d$. We know now from joint work with one of the authors that this is a special case of a more universal approximation scheme, in which $\hat{P}_n$ is replaced by a general probability measure P on $\mathbb{R}^d$. It is shown in Dümbgen *et al.* (2010) that for a probability measure P that has a first moment and is not concentrated on any hyperplane of $\mathbb{R}^d$ a unique maximizer

$$\hat{f}_n(\cdot|P) \in \arg\max_{\tilde{f}\in\mathcal{F}_0 \text{ upper semi-cont.}}\left[\int \log\{\tilde{f}(x)\}\,P(\mathrm{d}x)\right]$$

exists and depends continuously on P in Mallows distance. If P has a log-concave density f, then $\hat{f}_n(\cdot|P) = f$ almost surely; if f is a general density, $\hat{f}_n(\cdot|P)$ minimizes the Kullback–Leibler divergence $d_{\mathrm{KL}}(\cdot, f)$.

One particular choice is

$$\hat{P}_\theta = \frac{1}{n}\sum_{i=1}^{n}\delta_{Y_i - x_i^{\mathrm{T}}\theta},$$

the empirical measure of the residuals in a univariate linear regression model $Y_i = x_i^{\mathrm{T}}\theta + \varepsilon_i$, $1 \leqslant i \leqslant n$. Assuming the error terms ε_i to be independent and identically distributed with a log-concave density f, existence and consistency of the semiparametric maximum likelihood estimator $(\hat{f}_n^*, \hat{\theta}_n^*)$ in this setting can be shown under very general conditions (Dümbgen *et al.*, 2010).

My second point concerns computation. I congratulate the authors on their algorithm LogConcDEAD, which in view of the adversity of the problem in the multivariate case is surprisingly fast and reliable. However, the computation times in Table 1 mean that the algorithm cannot realistically be applied in higher dimensions, with large samples or many times sequentially. The second limitation is also relevant for an approximate computation of $\hat{f}_n(\cdot|P)$ if P is non-discrete; the third limitation in particular for the multivariate version of the regression setting that was outlined above.

It might also be desirable to have an algorithm which can identify in a natural way (up to numerical precision) the maximal polytopes in $\mathbb{R}^d$ on which $\log(\hat{f}_n)$ is linear, whereas the current algorithm 'only' identifies subsimplices with this property. Consider n points in $\mathbb{R}^2$ that form a regular n-gon. It is easy to see from symmetry considerations that $\hat{f}_n$ is the uniform density on this n-gon and not just log-linear on subsimplices. Although this example is rather contrived, I conjecture that such maximal polytopes that are not simplices appear quite often and can reveal important information about the structure of the underlying distribution.

Guenther Walther (*Stanford University*)

I started looking at log-concave distributions when I was searching for an appropriate model for subpopulations of multivariate flow cytometry data about 10 years ago. The use of log-concave distributions is appealing for this purpose since their unimodal character is commonly associated with a single component population. In addition, log-concave distributions have a certain non-parametric flexibility that is helpful in many problems, but they can still be estimated without having to deal with a tuning parameter. When I worked out how to compute the maximum likelihood estimator (MLE) in the univariate case, I realized that the multivariate case would be much more daunting, requiring a more involved optimization algorithm and a considerable computational overhead for the construction of multivariate tessellations. I considered the task to be too challenging and decided not to pursue it further beyond the univariate work that I had done at that time.

Cule, Samworth and Stewart have shown in their paper how to compute the multivariate MLE by using Shor's r-algorithm, and they provide an accompanying software package that implements their algorithm. I congratulate them on this work and I believe that the paper will inspire much further research into the multivariate case. In particular, they show how, by modifying the objective function for the MLE, the problem becomes amenable to known, albeit slow, convex optimization algorithms. It is desirable to improve on the computation times that are given in Table 1, especially for the higher dimensional cases. I expect that the groundwork that the paper lays in terms of the optimization problem will inspire new research into faster algorithms. Another intriguing result is the outstanding performance of the MLE

vis-à-vis other non-parametric methods as reported in their simulation study. These results provide a strong motivation to establish theoretical results about the finite sample and asymptotic performance of the MLE.

The **authors** replied later, in writing, as follows.

We are very grateful to all the discussants for their many helpful comments, insights and suggestions, which will no doubt inspire plenty of future work. Unfortunately we cannot respond to all of the issues raised in this brief rejoinder, but we offer the following thoughts related to some of these contributions.

Other shape constraints and methods

Several discussants (Delaigle, Hall, Wellner, Seregin, Chacón and Critchley) ask about other possible shape constraints. Indeed, Seregin and Wellner (2010) have recently shown that a maximum likelihood estimator exists within the class of d-variate densities of the form $f = h \circ g$, where h is a known monotone function and g is an unknown convex function. Certain conditions are required on h, but taking $h(y) = \exp(-y)$ recovers log-concavity, whereas taking $h(y) = y_+^{1/r}$ (with $0 > r > -1/d$) yields the larger class of r-concave densities. Questions of uniqueness and computation of the estimate for these larger classes are still open. Of course, such larger classes must still rule out the spiking problem that was mentioned on the second page of the paper. Koenker and Mizera (2010) study maximum entropy estimators within these larger classes, whereas Leng and Jeon propose in their discussion an alternative M-estimation method which again has wide applicability.

As pointed out in Chacón's discussion, Carando *et al.* (2009) have considered maximum likelihood estimation of a multi-dimensional Lipschitz continuous density. The Lipschitz constant κ must be specified in advance and the estimator will be as rough as allowed by the class, but consistency, e.g. in L_1-distance, is achievable provided that κ is chosen sufficiently large (we are not required to let $\kappa \to \infty$). Given the size of the class, slower rates of convergence are to be expected.

Shape-constrained kernel methods, as studied in Braun and Hall (2001) and mentioned by Delaigle, Cheng and Hall, offer a further alternative. The idea here is to choose a distance (or divergence) between an original data point and a perturbed version of it. Starting with a standard kernel estimate, we then minimize the sum of these distances subject to the shape constraint being satisfied by the kernel estimate applied to the perturbed data set. Attractive features are smoothness of the resulting estimates and the generality of the method for incorporating different shape constraints; difficulties include the need to choose a distance as well as a bandwidth matrix and the challenges that are involved in solving the optimization problem, particularly in multi-dimensional cases. Similarly, the related biased bootstrap method of Hall and Presnell (1999) warrants further study in multi-dimensional density estimation contexts.

Wellner mentions the interesting class of hyperbolically k-monotone (completely monotone) densities on $(0, \infty)$. To answer one of his questions, it seems the natural generalization to higher dimensions is to say that a density f on $(0, \infty)^d$ is *hyperbolically k monotone (completely monotone)* if, for all $u \in (0, \infty)^d$, the function $f(uv)\ f(u/v)$ is k monotone (completely monotone) in $w = v + v^{-1} \in [2, \infty)$. We would then be interested, for instance, in the class $\mathcal{C}$ of densities of random vectors $X = (X_1, \ldots, X_d)^{\mathrm{T}}$ such that the density of $\exp(X) = \exp(X_1), \ldots, \exp(X_d))^{\mathrm{T}}$ is hyperbolically completely monotone. It can be shown that $\mathcal{C}$ does indeed contain the Gaussian densities on $\mathbb{R}^d$, and, given the attractive closure and other properties, maximum likelihood estimation within the class $\mathcal{C}$ would seem to be an exciting avenue for future research.

Theoretical properties

We wholeheartedly agree with the many discussants (Rufibach, Zhang and Li, Cheng, Hall, Seregin, Chacón and Jankowski) who identify the problem of establishing the rates of convergence of the log-concave maximum likelihood estimator (and corresponding functional estimates) when $d > 1$ as a key future challenge. The well-known conjectured rates (e.g. Seregin and Wellner (2010)) suggest a suboptimal rate when $d \geqslant 4$. Although this certainly motivates the search for modified rate optimal estimates involving penalization or working with smaller classes of densities, as mentioned by both Rufibach and Seregin, it is also important not to lose sight of the computational demands in these higher dimensional problems. With this is mind, dimension reduction techniques, as mentioned by both Cheng and Critchley, are especially valuable, as are methods which introduce further structure into the density, such as the analysis-of-variance decomposition of the log-density that was mentioned by Leng and Jeon. The fact that log-concavity is preserved under marginalization and conditioning, as described in proposition 1 of the paper, suggests viable methods that certainly deserve further exploration.

Theory for the plug-in functional estimators $\hat{\theta} = \theta(\hat{f}_n)$ that was introduced in Section 7 and discussed by Delaigle, Seregin and Jankowski is also of considerable interest, and the simulations by Jankowski suggesting an $n^{-1/2}$ rate of convergence in one case are noteworthy in this respect. To answer a question that was raised by Delaigle, $\hat{\theta}$ will be robust to misspecification of log-concavity in cases where the true density f_0 is close to the Kullback–Leibler minimizing density f^* and/or where the functional $\theta(f)$ varies only slowly as f moves from f_0 to f^*. In a different context, Lu and Young argue that simulating the distribution of a scaled version of the signed root likelihood ratio statistic under an incorrect fitted distribution is robust to model misspecification. The disturbing story that was recounted by Stone regarding the allocation of primary care trust funding by the Department of Health emphasizes the need for much greater understanding of the properties of statistical procedures under model misspecification.

Dependent data

Zhang and Li, Xia and Tong, and Yao ask about conditional density estimation. In low dimensional contexts, one could use the log-concave maximum likelihood estimate (or its smoothed version) of the joint density and then obtain a conditional density estimate by taking the relevant normalized 'slice' through the joint density estimate. Proposition 1 of course guarantees that this conditional density estimate is log-concave. In the specific time series settings that were mentioned by both Xia and Tong, and Yao, where the likelihood may be expressed as a product of conditional likelihoods, we can in fact extend our ideas to handle these cases. For instance, take the simple example of an auto-regressive model of order 1, where $X_0 = 0$ and

$$X_i = \rho X_{i-1} + \varepsilon_i, \qquad i = 1, \ldots, n.$$

Assuming that the innovations $\varepsilon_1, \ldots, \varepsilon_n$ are independent with common density f, the likelihood function in this semiparametric model is

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

Dümbgen *et al.* (2010) discuss algorithms for maximizing similar functions to obtain the joint maximizer $(\hat{\rho}, \hat{f})$ under the assumption that f is log-concave. These ideas can be extended to certain other types of dependence, which greatly increases the scope of our methodology. Heuristic arguments indicate that consistency results of the sort given for independent data in Dümbgen *et al.* (2010) should continue to hold for these sorts of dependent data, though these require formal verification.

Computational issues

Both Xue and Titterington and Xia and Tong discuss the possibility of modifying the log-concave maximum likelihood estimate so that it is positive beyond the boundary of the convex hull of the data by extending the lowest exponential surfaces (and presumably renormalizing so that the density has unit integral). Unfortunately, in certain cases such an extension is not well defined: for instance, if $d = 1$ and the data are uniformly spaced, the log-concave maximum likelihood estimate is the uniform distribution between the minimum and the maximum data points; extending this density yields a function which cannot be renormalized. The smoothed log-concave estimator that is proposed in Section 9 offers an alternative method for obtaining an estimate with full support.

Gopal and Casella show that the Metropolis–Hastings method for sampling from the fitted log-concave maximum likelihood estimator results in a higher acceptance rate and smaller standard errors than the rejection sampling method that is proposed in Appendix B.3. The (weak) dependence that is introduced into successive sampled observations by this method is probably insignificant for most purposes, so we have incorporated the algorithm into the latest version of the R package LogConcDEAD (Cule *et al.*, 2010).

To answer a question of Xia and Tong, the triangulation of the convex hull of the data into simplices which underpins the maximum likelihood estimator is not unique; however, there is a unique set of maximal polytopes (whose vertices correspond to the set of 'critically supporting tent poles') on which $\log(\hat{f}_n)$ is linear. Schuhmacher comments on identifying these maximal polytopes. Indeed, in one dimension, Dümbgen and Rufibach (2009) showed that, under sufficient smoothness and other conditions, the maximal distance between consecutive knots in the estimator is $O_p(\rho_n^{1/5})$, where $\rho_n = n^{-1} \log(n)$. An analogous result in higher dimensions would certainly be of interest. It would remain a challenge to exploit this information to yield a faster algorithm but, along with Xue and Titterington, Böhning and Wang, Schuhmacher and Walther, we strongly encourage further developments in this area. Such developments may

even facilitate on-line algorithms, which as described by Anagnostpoulos are of great interest particularly in the machine learning community.

Koenker and Mizera report impressive time savings for computing their maximum entropy estimator in a bivariate example. Their algorithm is based on interior point methods for convex programming which enforce convexity on a finite grid through a discrete Hessian and uses a Riemann sum and linear interpolation approximations to estimate the integral in their analogue of equation (3.2) in our paper. It may be desirable, instead of only computing the estimator at grid points, to obtain the triangulation into simplices $C_{n,j}$ and quantities $b_1, \ldots, b_m \in \mathbb{R}^d$ and $\beta_1, \ldots, \beta_m \in \mathbb{R}$ involved in the polyhedral characterization of the estimator (see Appendix B), in which case it seems that it should be possible to adapt Shor's r-algorithm to handle r-concave estimators, though some numerical approximation of the integral term may be necessary. It would be interesting to know whether Koenker and Mizera have had success with their method in more than two dimensions, and whether it is possible to control the error in their approximations in terms of the mesh size of the grid.

Finite sample properties

Several discussants (Delaigle, Chacón, Chen, Hazelton and Walther) discuss the simulation results. Of course the maximum likelihood estimator makes use of additional log-concavity information, but what makes the results interesting is the fact that maximum likelihood estimators are not designed specifically to perform well against integrated squared error (ISE) criteria. Moreover, the log-concave maximum likelihood estimator has other desirable properties, such as affine equivariance, which many other methods do not have.

It is gratifying to see from the additional simulations that are provided by Chen that the smoothed log-concave estimator in Section 9 does appear to offer quite substantial ISE improvements over its unsmoothed analogue for small or even moderate sample sizes. In Fig. 19 we give further detail on these results in the case of density (a), the standard Gaussian density, by providing boxplots of the ISE for various methods based on 50 replications. Apart from giving another demonstration of the performance of the smoothed log-concave estimator, two points are particularly worth noting: firstly, in most cases the variability of the ISE does not appear to be larger for the two log-concave methods compared with the kernel methods (this addresses a question that was raised by Chacón in a personal communication). Secondly, using the optimal ISE bandwidth for the kernel method (which would again be unknown in practice) offers very little improvement over the optimal mean ISE bandwidth. This agrees with the findings for other distributions in a study by Chacón (personal communication) and addresses a point that was raised by Delaigle.

Both Zhang and Li, and Hazelton mention using boundary kernels (Wand and Jones (1995), pages 46–49) to improve the ISE performance of kernel methods in cases where the true density does not have full support. Indeed, as Fig. 20 indicates for the one-dimensional $\Gamma(2, 1)$ true density, some improvements are possible when the bandwidth for the linear boundary kernel is chosen to minimize the ISE (though the method also assumes knowledge of the support of the true density). As envisaged by Zhang and Li, however, even then we can do better with our proposed methods (except in the case of a small sample size, for the unsmoothed log-concave estimator).

Other issues

Both Xue and Titterington, and Böhning and Wang discuss applications of the log-concave maximum likelihood estimator to classification and clustering. Chen (2010) has also observed competitive performance from the log-concave maximum likelihood estimator in classification problems. Using the smoothed log-concave estimator (Section 9) can further improve matters, and finesses the issue of how to classify observations outside the convex hulls of the training data in each class.

Dannemann and Munk make insightful remarks about the identifiability of mixtures of log-concave densities, and their Fig. 16 with two mixture components is particularly instructive. One sensible alternative, as Dannemann and Munk suggest, is to model one of the mixture components parametrically; another possibility in some circumstances might be to model the logarithm of each of the mixture components as a tent function (requiring no change to the algorithm).

Critchley asks a very pertinent question about the possibility of transforming to log-concavity. In this context, Wellner mentions that logarithmic transformations of random variables with hyperbolically monotone densities of order 1 have log-concave densities, but this is an area which deserves much greater exploration.

Draper provides several pointers to the parallel Bayesian non-parametric density estimation literature. As he points out, these methods offer small sample competitors to confidence intervals or bands for densities or functionals of densities that are constructed using the bootstrap or asymptotic theory.

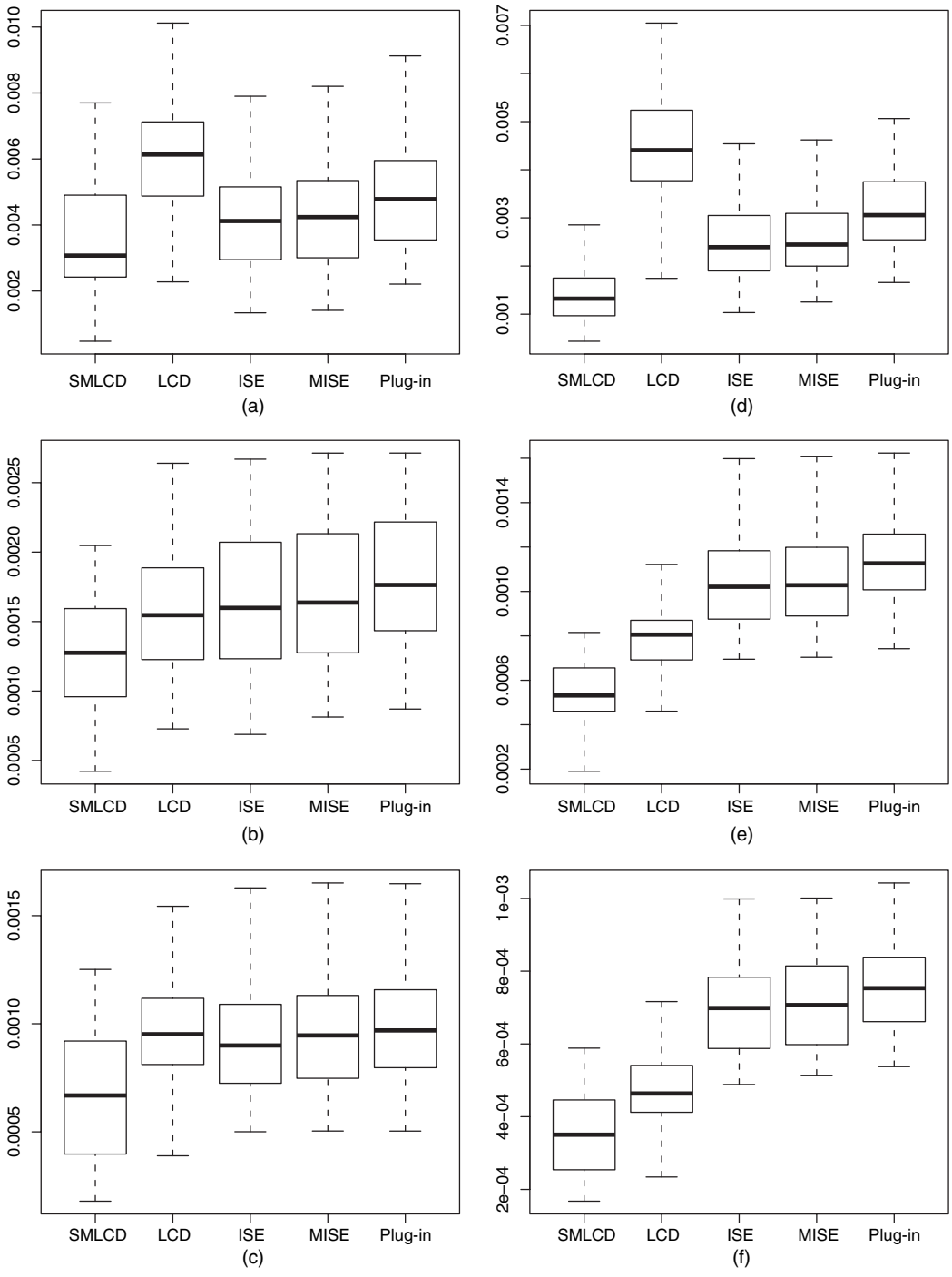


Fig. 19. Boxplots of ISEs with standard Gaussian true density for the smoothed log-concave maximum likelihood estimator SMLCD, log-concave maximum likelihood estimator LCD and three kernel methods—with the optimal ISE bandwidth ISE, the optimal mean ISE bandwidth MISE and a plug-in bandwidth Plug-in: (a) $n = 100$, $d = 2$; (b) $n = 500$, $d = 2$; (c) $n = 1000$, $d = 2$; (d) $n = 100$, $d = 3$; (e) $n = 500$, $d = 3$; (f) $n = 1000$, $d = 3$

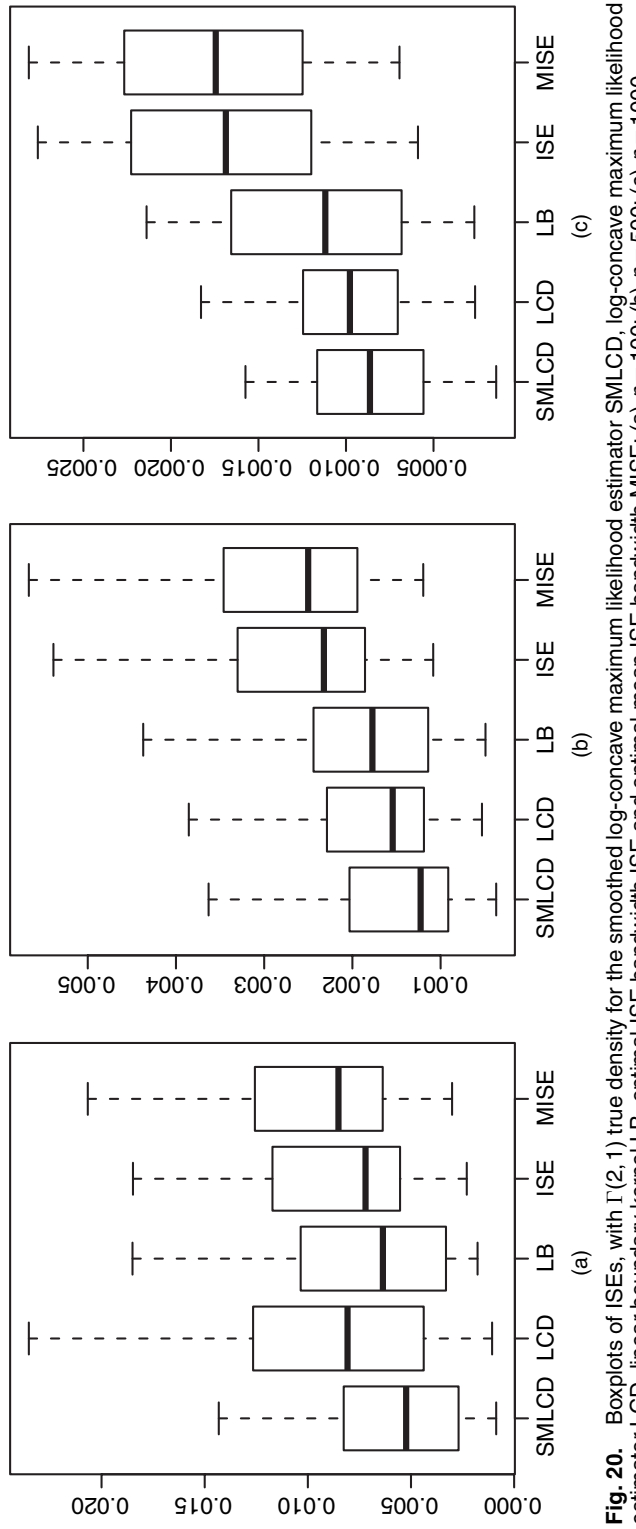


Fig. 20. Boxplots of ISEs, with $\Gamma(2, 1)$ true density for the smoothed log-concave maximum likelihood estimator SMLCD, log-concave maximum likelihood estimator LCD, linear boundary kernel LB, optimal ISE bandwidth ISE and optimal mean ISE bandwidth MISE: (a) $n = 100$; (b) $n = 500$; (c) $n = 1000$

Hazelton presents a nice extension of Silverman's bump hunting idea as an alternative test for log-concavity. It may be that taking bootstrap samples from the fitted smoothed log-concave estimator (which is very straightforward to do) when computing the critical value of the test is a sensible option here. More generally, as mentioned by Jang and Lim, taking bootstrap samples from the fitted smoothed log-concave estimator, or its unsmoothed analogue, can form the basis for many other smoothed bootstrap (bagging with smearing) procedures, which certainly deserve further investigation. Sampling from the smoothed version has a clear advantage in the product density scenario of Kypraios, Preston and White, since, when using the unsmoothed maximum likelihood estimator, the product density would only be positive on the intersection of the convex hulls of the samples. The strategy is viable in principle regardless of the number of terms in the product, though, as with all related methods, estimates in the tails (where the product density is very small) are likely to be highly variable when the number of terms in the product is large.

We thank Yining Chen for his help with the simulations that are reported in this rejoinder. Finally, we record our gratitude to the Research Section for their handling of the paper, and the Royal Statistical Society for organizing the Ordinary Meeting.

References in the discussion

Allman, E. S., Matias, C. and Rhodes, J. A. (2009) Identifiability of parameters in latent structure models with many observed variables. *Ann. Statist.*, **37**, 3099–3132.
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.
Barndoff-Nielsen, O. E. (1986) Inference on full and partial parameters based on the standardized signed log likelihood ratio. *Biometrika*, **73**, 307–322.
Beaumont, M. A., Zhang, W. and Balding, D. J. (2002) Approximate Bayesian computation in population genetics. *Genetics*, **162**, 2025–2035.
Birgé, L. and Massart, P. (1993) Rates of convergence for minimum contrast estimators. *Probab. Theor. Reltd Flds*, **97**, 113–150.
Bondesson, L. (1990) Generalized gamma convolutions and complete monotonicity. *Probab. Theor. Reltd Flds*, **85**, 181–194.
Bondesson, L. (1992) Generalized gamma convolutions and related classes of distributions and densities. *Lect. Notes Statist.*, **76**.
Bondesson, L. (1997) On hyperbolically monotone densities. In *Advances in the Theory and Practice of Statistics*, pp. 299–313. New York: Wiley.
Bordes, L., Delmas, C. and Vandekerkhove, P. (2006) Semiparametric estimation of a two-component mixture model where one component is known. *Scand. J. Statist.*, **33**, 733–752.
Borell, C. (1975) Convex set functions in d-space. *Period. Math. Hung.*, **6**, 111–136.
Braun, W. J. and Hall, P. (2001) Data sharpening for nonparametric inference subject to constraints. *J. Comput. Graph. Statist.*, **10**, 786–806.
Breiman, L. (1996) Bagging predictors. *Mach. Learn.*, **24**, 123–140.
Breiman, L. (2000) Randomizing outputs to increase prediction accuracy. *Mach. Learn.*, **40**, 229–242.
Bronšteĭn, E. M. (1976) ε-entropy of convex sets and functions. *Sibirsk. Mat. Ž.*, **17**, 508–514, 715.
Cappe, O. and Moulines, E. (2008) On-line expectation–maximization algorithm for latent data models. *J. R. Statist. Soc.* B, **71**, 593–613.
Carando, D., Fraiman, R. and Groisman, P. (2009) Nonparametric likelihood based estimation for a multivariate Lipschitz density. *J. Multiv. Anal.*, **100**, 981–992.
Chacón, J. E. and Duong, T. (2010) Multivariate plug-in bandwidth selection with unconstrained pilot bandwidth matrices. *Test*, **19**, 375–398.
Chacón, J. E., Montanero, J. and Nogales, A. G. (2007) A note on kernel density estimation at a parametric rate. *J. Nonparam. Statist.*, **19**, 13–21.
Chen, Y. (2010) A comparison of different nonparametric classification techinques. *MPhil Thesis*. University of Cambridge, Cambridge.
Clyde, M. A. and Lee, H. (2001) Bagging and the Bayesian bootstrap. In *Artificial Intelligence and Statistics* (eds T. Richardson and T. Jaakkola), pp. 169–174. San Francisco: Morgan Kaufmann.
Cule, M., Gramacy, R. and Samworth, R. (2009) LogConcDEAD: an R package for maximum likelihood estimation of a multivariate log-concave density. *J. Statist. Softwr.*, **29**.
Cule, M. L., Gramacy, R. B., Samworth, R. J. and Chen, Y. (2007) LogConcDEAD: maximum likelihood estimation of a log-concave density. *R Package Version 1.4-2*. (Available from `http://CRAN.R-project.org/package=LogConcDEAD`.)
Dharmadhikari, S. and Joag-Dev, K. (1988) *Unimodality, Convexity, and Applications*. Boston: Academic Press.

Dümbgen, L., Igl, B.-W. and Munk, A. (2008) *p*-values for classification. *Electron. J. Statist.*, **2**, 468–493.
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. *Technical Report 75*. Universität Bern, Bern. (Available from `http://arxiv.org/abs/1002.3448/`.)
Escobar, M. D. and West, M. (1995) Bayesian density estimation and inference using mixtures. *J. Am. Statist. Ass.*, **90**, 577–588.
Ferguson, T. S. (1973) A Bayesian analysis of some nonparametric problems. *Ann. Statist.*, **1**, 209–230.
Ferguson, T. (1974) Prior distributions on the space of all probability measures. *Ann. Statist.*, **2**, 615–629.
Frank, E. and Pfahringer, B. (2006) Improving on bagging with input smearing. In *Proc. 10th Pacific-Asia Conf. Knowledge Discovery and Data Mining* (eds W. K. Ng, M. Kit-suregawa and J. Li), pp. 97–106. Berlin: Springer.
Freedman, D. (1963) On the asymptotic behavior of Bayes estimates in the discrete case. *Ann. Math. Statist.*, **34**, 1386–1403.
Freedman, D. (1965) On the asymptotic behavior of Bayes estimates in the discrete case: II. *Ann. Math. Statist.*, **35**, 454–456.
Fukunaga, K. and Mantock, J. M. (1983) Nonparametric discriminant analysis. *IEEE Trans. Pattn Anal. Mach. Intell.*, **5**, 671–677.
Groeneboom, P. (1985) Estimating a monotone density. In *Proc. Berkeley Conf. in Honor of Jerzy Neyman and Jack Kiefer*, vol. II, pp. 539–555. Belmont: Wadsworth.
Groeneboom, P. (1989) Brownian motion with a parabolic drift and Airy functions. *Probab. Theor. Reltd Flds*, **81**, 79–109.
Groeneboom, P., Hooghiemstra, G. and Lopuhaä, H. P. (1999) Asymptotic normality of the L_1 error of the Grenander estimator. *Ann. Statist.*, **27**, 1316–1347.
Groeneboom, P., Jongbloed, G. and Wellner, J. A. (2001) Estimation of a convex function: characterizations and asymptotic theory. *Ann. Statist.*, **29**, 1653–1698.
Hall, P. and Presnell, B. (1999) Biased bootstrap methods for reducing the effects of contamination. *J. R. Statist. Soc.* B, **61**, 661–680.
Hall, P. and Zhou, X.-H. (2003) Nonparametric estimation of component distributions in a multivariate mixture. *Ann. Statist.*, **31**, 201–224.
Han, B., Comaniciu, D., Zhu, Y. and Davis, L. S. (2007) Sequential kernel density approximation and its application to real-time visual tracking. *IEEE Trans. Pattn Anal. Mach. Intell.*, **30**, 1186–1197.
Hanson, T. and Johnson, W. O. (2001) Modeling regression error with a mixture of Pólya trees. *J. Am. Statist. Ass.*, **97**, 1020–1033.
Hazelton, M. L. and Marshall, J. C. (2009) Linear boundary kernels for bivariate density estimation. *Statist. Probab. Lett.*, **79**, 999–1003.
Hjort, N. L., Holmes, C., Müller, P. and Walker, S. G. (2010) *Bayesian Nonparametrics*. Cambridge: Cambridge University Press.
Holzmann, H., Munk, A. and Gneiting, T. (2006) Identifiability of finite mixtures of elliptical distributions. *Scand. J. Statist.*, **33**, 753–763.
Hunter, D. R., Wang, S. and Hettmansperger, T. P. (2007) Inference for mixtures of symmetric distributions. *Ann. Statist.*, **35**, 224–251.
Jeon, Y. and Lin, Y. (2006) An effective method for high dimensional log-density ANOVA estimation, with application to nonparametric graphical model building. *Statist. Sin.*, **16**, 353–374.
Kim, J. K. and Pollard, D. (1990) Cube root asymptotics. *Ann. Statist.*, **18**, 191–219.
Koenker, R. and Mizera, I. (2010) Quasi-concave density estimation. *Ann. Statist.*, **38**, 2998–3027.
Kulikov, V. N. and Lopuhaä, H. P. (2005) Asymptotic normality of the L_k-error of the Grenander estimator. *Ann. Statist.*, **33**, 2228–2255.
Mizera, I. and Koenker, R. (2006) Primal and dual formulations for the estimation of a probability density via regularization: divergences, entropies, and likelihoods. In *Qualitative Assumptions and Regularization in High-dimensional Statistics* (eds L. Dümbgen and J. A. Wellner), pp. 2981–2982. Oberwolfach: Mathematisches Forschungsinstitut. (Available from `www.mfo.de/programme/schedule/2006/45/OWR_2006_49.pdf`.)
Prakasa Rao, B. L. S. (1970) Estimation for distributions with monotone failure rate. *Ann. Math. Statist.*, **41**, 507–519.
R Development Core Team (2009) *R: a Language and Environment for Statistical Computing*. Vienna: R Foundation for Statistical Computing.
Robert, C. P. and Casella, G. (2004) *Monte Carlo Statistical Methods*, 2nd edn. New York: Springer.
Roberts, S. J. (1997) Parametric and non-parametric unsupervised cluster analysis. *Pattn Recogn.*, **30**, 261–272.
Rubin, D. (1981) The Bayesian bootstrap. *Ann. Statist.*, **6**, 461–464.
Rufibach, K. (2007) Computing maximum likelihood estimators of a log-concave density function. *J. Statist. Computn Simuln*, **77**, 561–574.

Rufibach, K. and Dümbgen, L. (2006) logcondens: estimate a log-concave probability density from i.i.d. observations. (Available from `http://CRAN.R-projects.org/package=logcondens`.)
Schuhmacher, D. and Dümbgen, L. (2010) Consistency of multivariate log-concave density estimators. *Statist. Probab. Lett.*, **80**, 376–380.
Schuhmacher, D., Huesler, A. and Duembgen, L. (2009) Multivariate log-concave distributions as a nearly parametric model. *Technical Report*. University of Bern, Bern. (Available from `http://www.citebase.org/abstract?id=oai:arXiv.org:0907.0250`.)
Seijo, E. and Sen, B. (2010) Nonparametric least squares estimation of a multivariate convex regression function. *Technical Report arXiv:1003.4765*. Department of Statistics, Columbia University.
Seregin, A. and Wellner, J. A. (2010) Nonparametric estimation of convex-transformed densities. *Ann. Statist.*, to be published.
Sheather, S. J. and Jones, M. C. (1991) A reliable data-based bandwidth selection method for kernel density estimation. *J. R. Statist. Soc.* B, **53**, 683–690.
Shor, N. Z. (1985) *Minimization Methods for Non-differentiable Functions*. Berlin: Springer.
Silverman, B. (1981) Using kernel density estimates to investigate multimodality. *J. R. Statist. Soc.* B, **43**, 97–99.
Silverman, B. W. (1986) *Density Estimation for Statistics and Data Analysis*. London: Chapman and Hall.
Stein, P. (1966) A note on the volume of a simplex. *Am. Math. Mnthly*, **73**, 299–301.
Stone, M. (2010) Formulas at war over two sorts of inequality in health funding. *Report*. Civitas, London. (Available from `www.civitas.org.uk/pdf/formulasatwarApril2010.pdf`.)
Tong, H. (2010) Threshold models in time series analysis-thirty years on. *Statist. Interface*, to be published.
van der Vaart, A. W. and Wellner, J. A. (1996) *Weak Convergence and Empirical Processes, with Applications to Statistics*. New York: Springer.
Walker, S. (2004) New approaches to Bayesian consistency. *Ann. Statist.*, **32**, 2028–2043.
Wand, M. P. and Jones, M. C. (1995) *Kernel Smoothing*. London: Chapman and Hall.
Wang, Y. (2007) On fast computation of the non-parametric maximum likelihood estimate of a mixing distribution. *J. R. Statist. Soc.* B, **69**, 185–198.
White, S. R., Preston, S. P., Crowe, J. and Kypraios, T. (2010) Simulation-based Bayesian inference for discretely observed Markov-models using an interval-based approach. To be published.
Yakowitz, S. J. and Spragins, J. D. (1968) On the identifiability of finite mixtures. *Ann. Math. Statist.*, **39**, 209–214.
Yuan, M. and Lin, Y. (2006) Model selection and estimation in regression with grouped variables. *J. R. Statist. Soc.* B, **68**, 49–67.