Rejoinder to *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

We are very grateful to all the discussants for their many helpful comments, insights and suggestions, which will no doubt inspire a great deal of future work. Unfortunately we are unable to 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, 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 $h(y) = e^{-y}$ recovers log-concavity, while 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 mentioned on p.2 of the paper. Koenker and Mizera (2010) study maximum entropy estimators within these larger classes, while 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) consider 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 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 minimise the sum of these distances subject to

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

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 involved in solving the optimisation 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 generalisation to higher dimensions is to say 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 $X = (X_1, \ldots, X_d)^T$ such that the density of $e^X = (e^{X_1}, \ldots, e^{X_d})^T$ 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, 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 $d \geq 4$. While this certainly motivates the search for modified rate-optimal estimates involving penalisation 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 ANOVA decomposition of the log-density mentioned by Leng and Jeon. The fact that log-concavity is preserved under marginalisation 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 $\hat{\theta} = \theta(\hat{f}_n)$ introduced in section 7 and discussed by Delaigle, Seregin and Jankowski are also of considerable interest, and the simulations by Jankowski suggesting an $n^{-1/2}$ convergence rate in one case are noteworthy in this respect. To answer a question 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 minimising 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 recounted by Stone regarding the allocation of primary care trust funding by the Department of Health emphasises 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 normalised '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 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 autoregressive model of order 1, where $X_0 = 0$ and

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

Assuming the innovations $\epsilon_1, \ldots, \epsilon_n$ are independent with common density f, the likelihood function in this semi-parametric model is

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

Dümbgen *et al.* (2010) discuss algorithms for maximising similar functions to obtain the joint maximiser $(\hat{\rho}, \hat{f})$ 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 it is positive beyond the boundary of the convex hull of the data by extending the lowest exponential surfaces (and presumably renormalising so 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 renormalised. The smoothed log-concave estimator 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 proposed in section B.3. The (weak) dependence 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 exists 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 online 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 (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 characterisation of the estimator (see Appendix B), in which case it seems it should be possible to adapt Shor's r-algorithm to handle r-concave estimators, though some numerical approximation of the integral term may well be necessary. It would be interesting to know if the authors 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, 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 possess.

It is gratifying to see from the additional simulations provided by Chen that the smoothed log-concave estimator in section 9 does indeed appear to offer quite substantial ISE improvements over its unsmoothed analogue for small or even moderate sample sizes. In Figure 1 we give further detail on these results in the case of density (a), the standard Gaussian density, by providing box plots of the ISE for different 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 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-MISE bandwidth. This agrees with the findings for other distributions in a study by Chacón (personal communication), and addresses a point raised by Delaigle.

Both Zhang and Li and Hazelton mention using boundary kernels (Wand and Jones, 1995, pp.46–49) to improve the ISE performance of kernel methods in cases where the true density does not have full support. Indeed, as Figure 2 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 minimise 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 Figure 1 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 nonparametric density estimation literature. As he points out, these methods offer small-sample competitors to confidence intervals/bands for densities or functionals of densities constructed using the bootstrap or asymptotic theory.

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.

Acknowledgements

We would like to thank Yining Chen for his help with the simulations 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 organising the Ordinary Meeting.

References

Braun, W. J. and Hall, P. (2001) Data sharpening for nonparametric inference subject to constraints. *J. Comput. Graph. Statist.*, **10**, 786–806.

Carando, D., Fraiman, R. and Groisman, P. (2009) Nonparametric likelihood based estimation for a multivariate Lipschitz density. *J. Mult. Anal.*, **100**, 981–992.

Chen, Y. (2010) A comparison of different nonparametric classification techniques. MPhil thesis, University of Cambridge.

Cule, M. L., Gramacy, R. B., Samworth, R. J. and Chen, Y. (2010) *LogConcDEAD: Maximum Likelihood Estimation of a Log-Concave Density.* http://CRAN.R-project.org/package=LogConcDEAD. R package version 1.4-2.

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

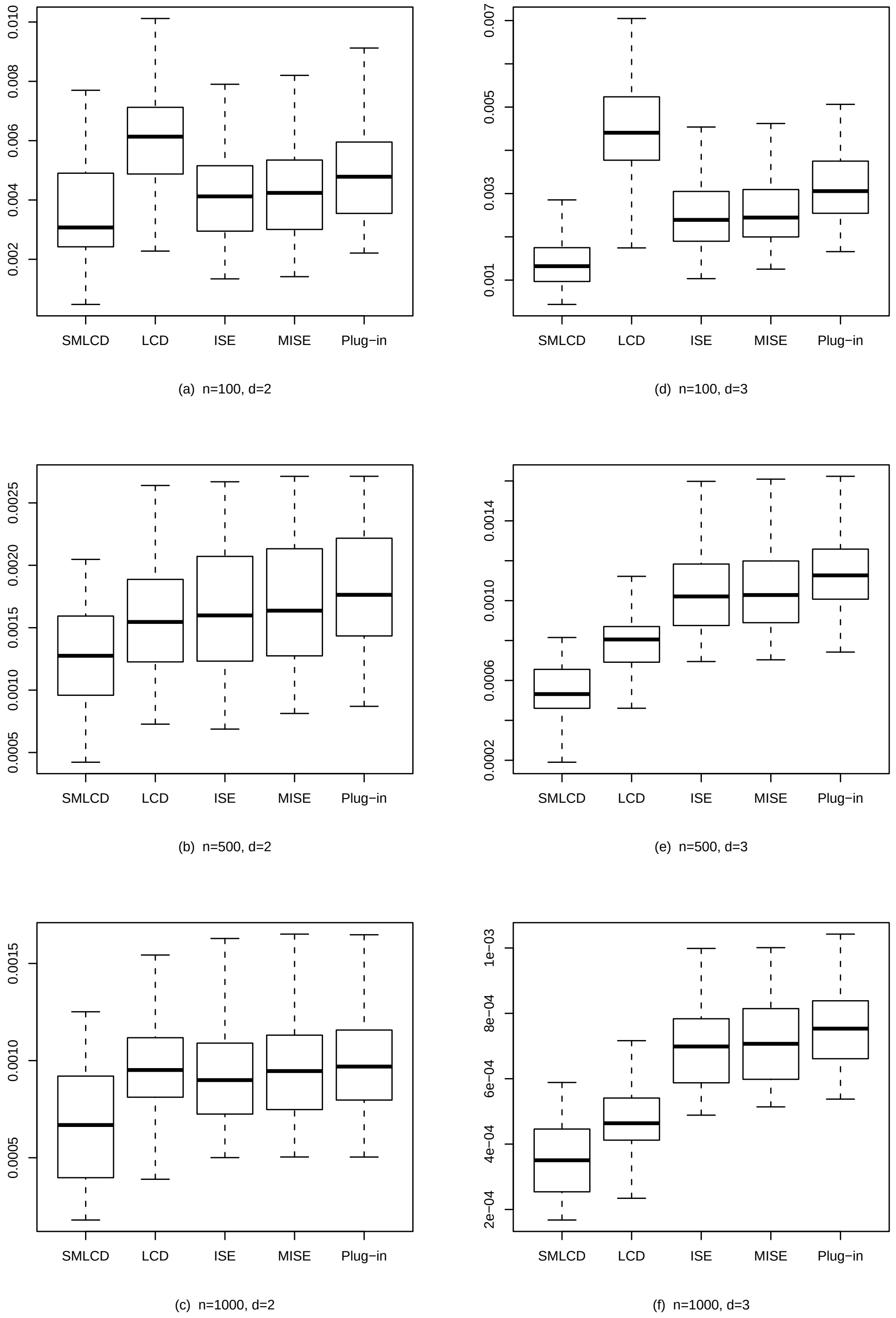
Dümbgen, L., Samworth, R. J. and Schuhmacher, D. (2010) Approximation by log-concave distributions with applications to regression. Tech. rep. 75, Universität Bern. http://arxiv.org/abs/1002.3448/

Hall, P. and Presnell, B. (1999) Biased bootstrap methods for reducing the effects of contamination. *J. Roy. Statist. Soc., Ser B*, **61**, 661–680

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

Seregin, A. and Wellner, J. A. (2010) Nonparametric estimation of convex-transformed densities. *Ann. Statist.*, to appear.

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



(a) n=100, d=2

(d) n=100, d=3

(b) n=500, d=2

(e) n=500, d=3

(c) n=1000, d=2

(f) n=1000, d=3

Fig. 1. Box plots of integrated squared errors 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 MISE bandwidth (MISE) and a plug-in bandwidth (Plug-in).

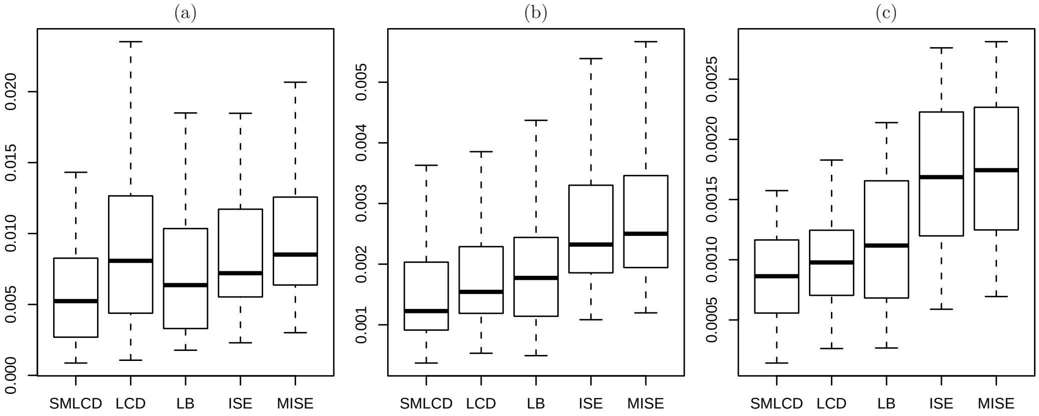


Fig. 2. Box plots of integrated squared errors with $\Gamma(2, 1)$ true density for the smoothed log-concave maximum likelihood estimator (SMLCD), log-concave maximum likelihood estimator (LCD), linear boundary kernel (LB), optimal ISE bandwidth (ISE), and optimal MISE bandwidth (MISE). Panel (a): $n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