Statistical Science
2018, Vol. 33, No. 4, 493–509
https://doi.org/10.1214/18-STS666

Recent Progress in Log-Concave Density Estimation

Richard J. Samworth

Abstract. In recent years, log-concave density estimation via maximum likelihood estimation has emerged as a fascinating alternative to traditional nonparametric smoothing techniques, such as kernel density estimation, which require the choice of one or more bandwidths. The purpose of this article is to describe some of the properties of the class of log-concave densities on $\mathbb{R}^d$ which make it so attractive from a statistical perspective, and to outline the latest methodological, theoretical and computational advances in the area.

Key words and phrases: Log-concavity, maximum likelihood estimation.

1. INTRODUCTION

Shape-constrained density estimation has a long history, dating back at least as far as Grenander (1956), who studied the maximum likelihood estimator of a decreasing density on the nonnegative half-line. Unlike traditional nonparametric smoothing approaches, this estimator does not require the choice of any tuning parameter, and indeed it has a beautiful characterisation as the left derivative of the least concave majorant of the empirical distribution function. Over subsequent years, a great deal of work went into understanding its theoretical properties (e.g., Prakasa Rao, 1969, Groeneboom, 1985, Birgé, 1989), revealing in particular its nonstandard cube-root rate of convergence.

On the other hand, the class of decreasing densities on $[0, \infty)$ is quite restrictive, and does not generalise particularly naturally to multivariate settings. In recent years, therefore, alternative families of densities have been sought, and the class of log-concave densities has emerged as one with many attractive properties from a statistical viewpoint. Indeed, the theory of log-concave density estimation has led to applications to a wide variety of problems, including the detection of the presence of mixing (Walther, 2002), filtering (Henningsson and Åström, 2006), tail index estimation (Müller and Rufibach, 2009), clustering (Cule, Samworth and Stewart, 2010), regression (Dümbgen, Samworth and Schuhmacher, 2011), Independent Component Analysis (Samworth and Yuan, 2012), classification (Chen and Samworth, 2013) and censored data problems (Dümbgen, Rufibach and Schuhmacher, 2014).

The main aim of this article is to give an account of the key properties of log-concave densities and their relevance for applications in statistical problems. We focus especially on ideas of log-concave projection, which underpin the maximum likelihood approach to inference within the class. Recent theoretical results and computational aspects will also be discussed. For alternative reviews of related topics, see Saumard and Wellner (2014), which has a greater emphasis on analytic properties, and Walther (2009), with a stronger focus on modelling and applications.

2. BASIC PROPERTIES

We say that $f : \mathbb{R}^d \to [0, \infty)$ is log-concave if $\log f$ is a concave function (with the convention $\log 0 := -\infty$). Let $\mathcal{F}_d$ denote the class of upper semi-continuous log-concave probability density functions on $\mathbb{R}^d$ with respect to d-dimensional Lebesgue measure. The upper semi-continuity is not particularly important in most of what follows, but it fixes a particular version of the density and means we do not need to worry about densities that differ on a set of zero Lebesgue measure.

Many standard families of densities are log-concave. For instance, Gaussian densities with positive-definite

Richard J. Samworth is Professor of Statistical Science, Statistical Laboratory, Wilberforce Road, Cambridge CB3 0WB, United Kingdom (e-mail: r.samworth@statslab.cam.ac.uk).

covariance matrices and uniform densities on convex, compact sets belong to $\mathcal{F}_d$; the logistic density $f(x) = \frac{e^{-x}}{(1-e^{-x})^2}$, Beta$(a, b)$ densities with $a, b \geq 1$, Weibull(α) densities with $\alpha \geq 1$, $\Gamma(\alpha, \lambda)$ densities with $\alpha \geq 1$, Gumbel and Laplace densities (amongst many others) belong to $\mathcal{F}_1$. It is convenient to think of log-concave densities as unimodal densities with exponentially decaying tails. Unimodality here is meant in the sense of the upper level sets being convex, though in one dimension, we have a stronger characterisation.

LEMMA 2.1 (Ibragimov, 1956). *A density f on $\mathbb{R}$ is log-concave if and only if the convolution $f * g$ is unimodal for every unimodal density g.*

A more precise statement about the exponentially decaying tails is as follows:

LEMMA 2.2 (Cule and Samworth, 2010). *If $f \in \mathcal{F}_d$, then there exist $\alpha > 0$, $\beta \in \mathbb{R}$ such that $f(x) \leq e^{-\alpha\|x\|+\beta}$ for all $x \in \mathbb{R}^d$.*

Thus, in particular, random vectors with log-concave densities have moment generating functions that are finite in a neighbourhood of the origin.

One of the features of the class of log-concave densities that makes them so attractive for statistical inference is their stability under various operations. A key result of this type is the following, due to Prékopa (1973), and with a simpler proof given in Prékopa (1980).

THEOREM 2.3. *Let $d = d_1 + d_2$ for some $d_1, d_2 \in \mathbb{N}$, and let $f : \mathbb{R}^d \to [0, \infty)$ be log-concave. Then*

$$x \mapsto \int_{\mathbb{R}^{d_2}} f(x, y)\, dy$$

is log-concave on $\mathbb{R}^{d_1}$.

Hence, marginal densities of log-concave random vectors are log-concave. As a simple consequence, we have the following corollary.

COROLLARY 2.4. *If f, g are log-concave densities on $\mathbb{R}^d$, then their convolution $f * g$ is a log-concave density on $\mathbb{R}^d$.*

PROOF. The function $(x, y) \mapsto f(x - y)g(y)$ is log-concave on $\mathbb{R}^{2d}$, so the result follows from Theorem 2.3. □

Two further straightforward stability properties are as follows.

PROPOSITION 2.5. *Let X have a log-concave density f on $\mathbb{R}^d$.*

(i) *If $A \in \mathbb{R}^{m\times d}$ has $m \leq d$ and $\operatorname{rank}(A) = m$, then AX has a log-concave density on $\mathbb{R}^m$.*

(ii) *If $X = (X_1^\top, X_2^\top)^\top$, then the conditional density of X_1 given $X_2 = x_2$ is log-concave for each x_2.*

Together, Theorem 2.3, Corollary 2.4 and Proposition 2.5 indicate that the class of log-concave densities is a natural infinite-dimensional generalisation of the class of Gaussian densities. Indeed, one can argue that a grand vision in the shape-constrained inference community is to free practitioners from restrictive parametric (often Gaussian) assumptions, while retaining many of the properties of these parametric procedures that make them so convenient for use in applications.

3. LOG-CONCAVE PROJECTIONS

Despite all of the nice properties of $\mathcal{F}_d$ described in the previous section, the class is not convex (again, this is also the case for the class of Gaussian densities). It is therefore by no means clear that there should exist a "closest" element of this set to a general distribution. Nevertheless, it turns out that one can make sense of such a notion, and that the appropriate concept is that of log-concave projection.

Let Φ denote the class of upper semi-continuous, concave functions $\phi : \mathbb{R}^d \to [-\infty, \infty)$ that are coercive in the sense that $\phi(x) \to -\infty$ as $\|x\| \to \infty$. Thus, $\mathcal{F}_d = \{e^\phi : \phi \in \Phi, \int_{\mathbb{R}^d} e^\phi = 1\}$. For $\phi \in \Phi$ and an arbitrary probability measure P on $\mathbb{R}^d$, define a kind of log-likelihood functional by

$$L(\phi, P) := \int_{\mathbb{R}^d} \phi\, dP - \int_{\mathbb{R}^d} e^\phi.$$

Thus, instead of enforcing the (nonconvex) constraint that ϕ should be a log-density explicitly, the functional above has the flavour of a Lagrangian, though the Lagrange multiplier is conspicuous by its absence. Nevertheless it turns out that any maximiser $\phi^* \in \Phi$ of this functional with $L(\phi^*, P) \in \mathbb{R}$ must be a log-density. To see this, note that if $\phi \in \Phi$ has $L(\phi, P) \in \mathbb{R}$ and $c \in \mathbb{R}$, then

$$\frac{\partial}{\partial c} L(\phi + c, P) = 1 - e^c \int_{\mathbb{R}^d} e^\phi.$$

Hence, at a maximum, $c = -\log(\int_{\mathbb{R}^d} e^\phi)$, which is equivalent to $\phi + c$ being a log-density.

Theorem 3.1 below gives a complete characterisation of when there exists a unique maximiser of $L(\phi, P)$ over $\phi \in \Phi$. We first require several further definitions: let $L^*(P) := \sup_{\phi\in\Phi} L(\phi, P)$ and let $\mathcal{P}_d$ denote the class of probability measures P on $\mathbb{R}^d$ satisfying both $\int_{\mathbb{R}^d} \|x\|\, dP(x) < \infty$ and $P(H) < 1$ for

all hyperplanes H. Let $\mathcal{C}_d$ denote the class of closed, convex subsets of $\mathbb{R}^d$, for a probability measure P on $\mathbb{R}^d$, let $\mathcal{C}_d(P) := \{C \in \mathcal{C}_d : P(C) = 1\}$, and let $\operatorname{csupp}(P) := \bigcap_{C \in \mathcal{C}_d(P)} C$ denote the convex support of P. Finally, let $\operatorname{int}(C)$ denote the interior of a convex set C, and for a concave function $\phi : \mathbb{R}^d \to [-\infty, \infty)$, let $\operatorname{dom}(\phi) := \{x : \phi(x) > -\infty\}$ denote its effective domain.

THEOREM 3.1 (Dümbgen, Samworth and Schuhmacher, 2011).

(i) *If* $\int_{\mathbb{R}^d} \|x\| \, dP(x) = \infty$, *then* $L^*(P) = -\infty$.

(ii) *If* $\int_{\mathbb{R}^d} \|x\| \, dP(x) < \infty$ *but* $P(H) = 1$ *for some hyperplane* H, *then* $L^*(P) = \infty$.

(iii) *If* $P \in \mathcal{P}_d$, *then* $L^*(P) \in \mathbb{R}$ *and there exists a unique* $\phi^* \in \Phi$ *that maximises* $L(\phi, P)$ *over* $\phi \in \Phi$. *Moreover*, $\operatorname{int}(\operatorname{csupp}(P)) \subseteq \operatorname{dom}(\phi^*) \subseteq \operatorname{csupp}(P)$.

A consequence of Theorem 3.1 and the preceding discussion is that there exists a well-defined map $\psi^* : \mathcal{P}_d \to \mathcal{F}_d$, given by

$$\psi^*(P) := \operatorname*{argmax}_{f \in \mathcal{F}_d} \int_{\mathbb{R}^d} \log f \, dP.$$

We refer to ψ^* as the *log-concave projection*. In the case where P is the empirical distribution of some data, this tells us that provided the convex hull of the data is d-dimensional, there exists a unique log-concave maximum likelihood estimator (MLE), a result first proved in Walther (2002) in the case $d = 1$, and Cule, Samworth and Stewart (2010) for general d. If P has a log-concave density f_0, then $\psi^*(P) = f_0$; more generally, if P has a density f_0 satisfying $\int_{\mathbb{R}^d} f_0 |\log f_0| < \infty$, then $\psi^*(P)$ minimises the Kullback–Leibler divergence $d^2_{\mathrm{KL}}(f_0, f) := \int_{\mathbb{R}^d} f_0 \log(f_0/f)$ over all $f \in \mathcal{F}_d$. These statements justify the use of the term "projection".

4. COMPUTATION OF LOG-CONCAVE MAXIMUM LIKELIHOOD ESTIMATORS

Let $X_1, \ldots, X_n \overset{\text{i.i.d.}}{\sim} P \in \mathcal{P}_d$, and let $\mathbb{P}_n$ denote their empirical distribution. In this section, we discuss the computation of the log-concave MLE $\hat{f}_n := \psi^*(\mathbb{P}_n)$ when the convex hull C_n of $X_1, \ldots, X_n$ is d-dimensional.

We initially focus on the case $d = 1$, and follow the Active Set approach of Dümbgen, Hüsler and Rufibach (2007), which is implemented in the `R` package `logcondens` (Dümbgen and Rufibach, 2011). Write $X_{(1)} \le \cdots \le X_{(n)}$ for the order statistics of the sample, and let Ψ denote the set of functions $\psi : \mathbb{R} \to [-\infty, \infty)$ that are continuous on $[X_{(1)}, X_{(n)}]$, linear on each $[X_{(k)}, X_{(k+1)}]$ and $-\infty$ on $\mathbb{R} \setminus [X_{(1)}, X_{(n)}]$. Let Ψ_{conc} denote the concave functions in Ψ. Then $\log \hat{f}_n \in \Psi_{\mathrm{conc}}$, because otherwise we could strictly increase $L(\cdot, \mathbb{P}_n)$ by replacing $\log \hat{f}_n$ with the $\psi \in \Psi_{\mathrm{conc}}$ with $\psi(X_i) = \log \hat{f}_n(X_i)$. Since any $\psi \in \Psi$ can be identified with the vector $\underline{\psi} := (\psi(X_{(1)}), \ldots, \psi(X_{(n)}))^\top \in \mathbb{R}^n$, our objective function can be written as

$$\begin{aligned}\tilde{L}(\underline{\psi}) &= \tilde{L}(\psi_1, \ldots, \psi_n) \\ &:= \frac{1}{n} \sum_{i=1}^{n} \psi_i - \sum_{k=1}^{n-1} \delta_k J(\psi_k, \psi_{k+1}),\end{aligned}$$

where $\delta_k := X_{(k+1)} - X_{(k)}$ (assumed positive for simplicity) and

$$J(r, s) := \int_0^1 e^{(1-t)r + ts} \, dt.$$

For $j = 2, \ldots, n-1$, let $v_j = (v_{j,1}, \ldots, v_{j,n})^\top \in \mathbb{R}^n$ have three nonzero components:

$$v_{j,j-1} := \frac{1}{\delta_{j-1}}, \quad v_{j,j} := -\frac{1}{\delta_j} - \frac{1}{\delta_{j-1}}, \quad v_{j,j+1} := \frac{1}{\delta_j}.$$

Then $\mathcal{K} := \{\underline{\psi} \in \mathbb{R}^n : v_j^\top \underline{\psi} \le 0 \text{ for } j = 2, \ldots, n-1\}$ denotes the set of feasible vectors, because $\underline{\psi} = (\psi_1, \ldots, \psi_n)^\top \in \mathcal{K}$ if and only if

$$\frac{\psi_j - \psi_{j-1}}{X_{(j)} - X_{(j-1)}} \ge \frac{\psi_{j+1} - \psi_j}{X_{(j+1)} - X_{(j)}}$$

for $j = 2, \ldots, n-1$. Thus our optimisation problem can be expressed as

$$\text{Maximise } \tilde{L}(\underline{\psi}) \quad \text{over } \underline{\psi} \in \mathcal{K}.$$

For any $\underline{\psi} \in \mathbb{R}^n$, we can define the set of "active" constraints $A(\underline{\psi}) := \{j \in \{2, \ldots, n-1\} : v_j^\top \underline{\psi} \ge 0\}$, so that for $\underline{\psi} \in \mathcal{K}$, the inactive constraints correspond to the "knots" of $\underline{\psi}$, where $\underline{\psi}$ changes slope. Since $\tilde{L}$ is strictly concave and infinitely differentiable, for any $A \subseteq \{2, \ldots, n-1\}$ and corresponding subspace $\mathcal{V}(A) := \{\underline{\psi} \in \mathbb{R}^n : v_j^\top \underline{\psi} = 0 \text{ for } j \in A\}$, it is straightforward to compute

$$\tilde{\psi}(A) \in \mathcal{V}_*(A) := \operatorname*{argmax}_{\underline{\psi} \in \mathcal{V}(A)} \tilde{L}(\underline{\psi})$$

using Newton methods. The basic idea of the Active Set approach is to start at a feasible point with a given

active set of variables A. We then optimise the objective under that set of active constraints, and move there if that new candidate point is feasible. If not, we move as far as we can along the line segment joining our current feasible point to the candidate point while remaining feasible. This new point has a different active set compared with our previous feasible iterate, so we can optimise the objective under this new set of active constraints, and repeat. More precisely, define a basis for $\mathbb{R}^n$ by $b_1 := (1)_{i=1}^n$, $b_j := \min(X_{(i)} - X_{(j)}, 0)_{i=1}^n$ for $j = 2, \ldots, n-1$ and $b_n := (X_{(i)})_{i=1}^n$. By considering the first-order stationarity conditions, it can be shown that any $\underline{\psi} \in \mathcal{V}_*(A)$ maximises $\tilde{L}$ over $\mathcal{K}$ if and only if $b_j^\top \nabla \tilde{L}(\underline{\psi}) \leq 0$ for all $j \in A$. The Active Set algorithm can therefore proceed as in Algorithm 1.

The main points to note in this algorithm are that in each iteration of the inner **while** loop, the active set decreases strictly (which ensures this loop terminates eventually), and that after each iteration of the outer **while** loop, the log-likelihood has strictly increased, and the current iterate $\underline{\psi}$ belongs to $\mathcal{K} \cap \mathcal{V}_*(A)$ for some $A \subseteq \{2, \ldots, n-1\}$. It follows that, up to machine precision, the algorithm terminates with the exact solution in finitely many steps. See Figure 1. Dümbgen, Rufibach and Schuhmacher (2014) study the more involved problem of estimating a log-concave (sub)-probability density in settings where observations may be subject to various different types of censoring, including right and interval censoring. In their `R` package `logconcens`, they propose an EM algorithm for computation (Dümbgen, Rufibach and Schuhmacher, 2013).

Returning to the original problem of computing the log-concave MLE, for $d \geq 2$, the feasible set is much more complicated, and only slower algorithms are available. For $y = (y_1, \ldots, y_n)^\top \in \mathbb{R}^n$, let $\bar{h}_y : \mathbb{R}^d \to \mathbb{R}$ denote the smallest concave function with $\bar{h}_y(X_i) \geq y_i$ for $i = 1, \ldots, n$; these are called *tent functions* in Cule, Samworth and Stewart (2010) (see Figure 2, which is taken from that paper). We can write the objective function in terms of the tent pole heights $y_1, \ldots, y_n$ as

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

This function is hard to optimise over $(y_1, \ldots, y_n)^\top \in \mathbb{R}^n$, partly because τ is not injective. However, Cule, Samworth and Stewart (2010) defined the modified ob-

Algorithm 1: Pseudo-code for an Active Set algorithm to compute $(\log \hat{f}_n(X_{(1)}), \ldots, \log \hat{f}_n(X_{(n)}))^\top$

```
Input: A ← {2, ..., n − 1}
ψ ← ψ̃(A)
while max_{j∈A} b_j^⊤ ∇L̃(ψ) > 0 do
    j* ← min(argmax_{j∈A} b_j^⊤ ∇L̃(ψ))
    ψ_cand ← ψ̃(A \ {j*})
    while ψ_cand ∉ K do
        t* ← max{t ∈ [0, 1] : (1 − t)ψ + tψ_cand ∈ K}
        ψ ← (1 − t*)ψ + t*ψ_cand
        A ← A(ψ)
        ψ_cand ← ψ̃(A)
    end
    ψ ← ψ_cand
    A ← A(ψ)
end
Output: ψ
```

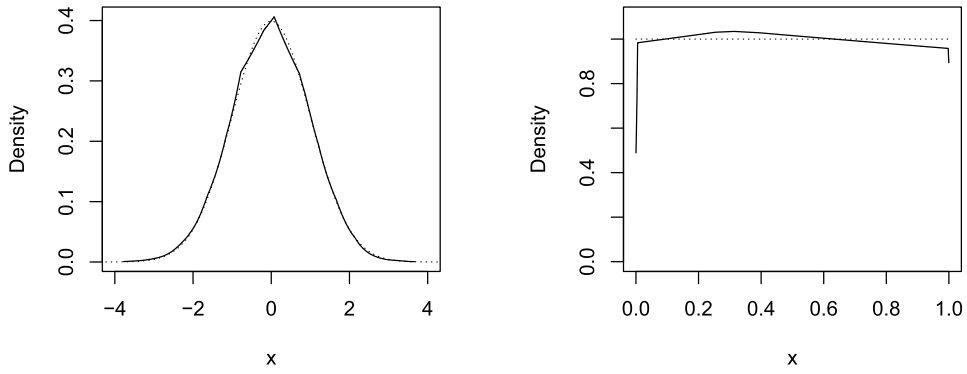

FIG. 1. *Log-concave maximum likelihood estimators (solid) based on* 4000 *observations from a standard normal distribution (left) and the* $U[0, 1]$ *distribution (right). The true densities are shown as dotted lines.*

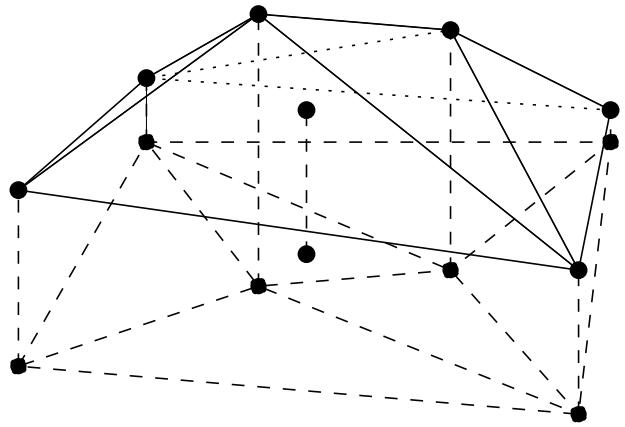

FIG. 2. *A schematic picture of a tent function in the case* $d = 2$.

jective function

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

Thus $\sigma \le \tau$, but the crucial points are that σ is concave and its unique maximum $\hat{y} \in \mathbb{R}^n$ satisfies $\log \hat{f}_n = \bar{h}_{\hat{y}}$. Even though σ is nondifferentiable, a subgradient of $-\sigma$ can be computed at every point, so Shor's r-algorithm (Kappel and Kuntsevich, 2000) can be used, as implemented in the `R` package `LogConcDEAD` (Cule, Gramacy and Samworth, 2009). See Figure 3, which is taken from Cule, Samworth and Stewart (2010). Koenker and Mizera (2010) study an alternative approximate approach based on imposing concavity of the discrete Hessian matrix of the log-density on a grid, and using a Riemann approximation to the integrability constraint.

5. PROPERTIES OF LOG-CONCAVE PROJECTIONS

For general distributions $P \in \mathcal{P}_d$, it is not possible to compute the log-concave projection $\psi^*(P)$ explicitly

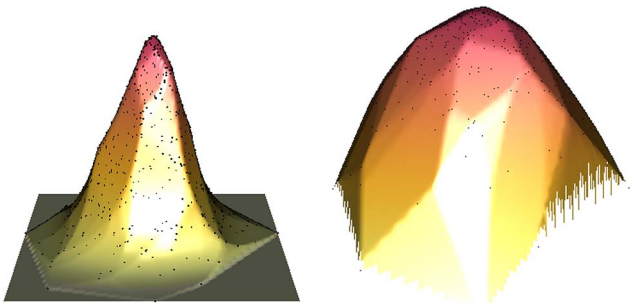

FIG. 3. *The log-concave maximum likelihood estimator (left) and its logarithm (right) based on* 1000 *observations from a standard bivariate normal distribution.*

(though see Section 5.1 below for several exceptions to this). Nevertheless, one can say quite a lot about the properties of log-concave projections, starting with affine equivariance.

LEMMA 5.1 (Dümbgen, Samworth and Schuhmacher, 2011). *Let* $X \sim P \in \mathcal{P}_d$, *let* $A \in \mathbb{R}^{d\times d}$ *be invertible, let* $b \in \mathbb{R}^d$, *and let* $P_{A,b}$ *denote the distribution of* $AX + b$. *Then*

$$\psi^*(P_{A,b})(x) = \frac{1}{|\det A|}\psi^*(P)\big(A^{-1}(x-b)\big).$$

A generic hope for the log-concave projection is that it should preserve as many properties of the original distribution as possible. Indeed, as we will see, such preservation results have motivated several associated methodological developments.

LEMMA 5.2 (Dümbgen, Samworth and Schuhmacher, 2011). *Let* $P \in \mathcal{P}_d$, *let* $\phi^* := \log\psi^*(P)$, *and let* $P^*(B) := \int_B e^{\phi^*}$ *for any Borel set* $B \subseteq \mathbb{R}^d$. *If* $\Delta : \mathbb{R}^d \to [-\infty, \infty)$ *is such that* $\psi^* + t\Delta \in \Phi$ *for sufficiently small* $t > 0$, *then*

$$\int_{\mathbb{R}^d} \Delta\,dP \le \int_{\mathbb{R}^d} \Delta\,dP^*.$$

As a special case of Lemma 5.2, we obtain the following corollary.

COROLLARY 5.3. *Let* $P \in \mathcal{P}_d$. *Then* P *and the log-concave projection measure* P^* *from Lemma 5.2 are convex ordered in the sense that*

$$\int_{\mathbb{R}^d} h\,dP^* \le \int_{\mathbb{R}^d} h\,dP$$

for all convex $h : \mathbb{R}^d \to (-\infty, \infty]$.

Applying Corollary 5.3 to $\Delta(x) = t^\top x$ for arbitrary $t \in \mathbb{R}^d$ allows us to conclude that $\int_{\mathbb{R}^d} x\,dP^*(x) = \int_{\mathbb{R}^d} x\,dP(x)$; in other words, log-concave projection preserves the mean μ of a distribution $P \in \mathcal{P}_d$. On the other hand, we see that the projection shrinks the second moment, in the sense that $A := \int_{\mathbb{R}^d}(x-\mu)(x-\mu)^\top d(P - P^*)(x)$ is nonnegative definite. In fact, we can say more: from the convex ordering in Corollary 5.3 and Strassen's theorem (Strassen, 1965), there exist random vectors $X \sim P$ and $X^* \sim P^*$, defined on the same probability space, such that $\mathbb{E}(X|X^*) = X^*$ almost surely. Thus $\mathbb{E}\{X^*(X - X^*)^\top\} = 0$, and from the decomposition $X = X^* + (X - X^*)$, we deduce that $A = 0$ if *and only if* P has a log-concave density. A different proof of this fact was given in Chen and Samworth (2013), Theorem 5.

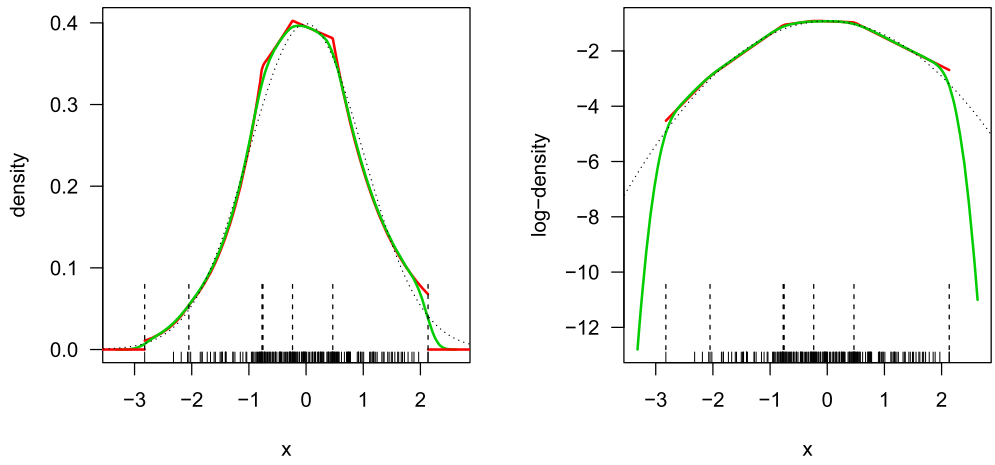

FIG. 4. *Left*: *A comparison of the original log-concave MLE* (*red*) *and smoothed log-concave MLE* (*green*) *based on* 200 *observations from a standard normal density* (*dotted*). *The short vertical lines indicate the observations, and the longer, dashed vertical lines show the locations of the knots of the log-concave MLE. Right*: *The same comparison on the log scale.*

This property validates the definition of the *smoothed log-concave projection*, proposed in the case $d = 1$ by Dümbgen and Rufibach (2009) and studied for general d in Chen and Samworth (2013). Writing $\tilde{\mathcal{P}}_d := \{P \in \mathcal{P}_d : \int_{\mathbb{R}^d} \|x\|^2 \, dP(x) < \infty\}$, this smoothed projection $\tilde{\psi}^* : \tilde{\mathcal{P}}_d \to \mathcal{F}_d$ is given by

$$\tilde{\psi}^*(P) := \psi^*(P) * N_d(0, A) = \int_{\mathbb{R}^d} \psi^*(x - y) \, dN_d(0, A)(y).$$

When P is the empirical distribution of some data, $\tilde{\psi}^*(P)$ is a smooth (real analytic), fully automatic density estimator that is log-concave (cf. Corollary 2.4), matches the first two moments of the data and is supported on the whole of $\mathbb{R}^d$. See Figure 4.

Our next property concerns the preservation of product structure, or, in the language of random vectors, independence of components.

PROPOSITION 5.4 (Chen and Samworth, 2013). *Let $P \in \mathcal{P}_d$ be of the form $P = P_1 \otimes P_2$ for some $P_1 \in \mathcal{P}_{d_1}$, $P_2 \in \mathcal{P}_{d_2}$ with $d_1 + d_2 = d$. Then for every $x = (x_1^\top, x_2^\top)^\top \in \mathbb{R}^{d_1+d_2}$, we have*

$$\psi^*(P)(x) = \psi^*(P_1)(x_1)\psi^*(P_2)(x_2).$$

Similarly, if in addition $P \in \tilde{\mathcal{P}}_d$, then $\tilde{\psi}^*(P)(x) = \tilde{\psi}^*(P_1)(x_1)\tilde{\psi}^*(P_2)(x_2)$.

Proposition 5.4 inspires a new approach to Independent Component Analysis; see Section 8 below. Incidentally, the converse of this result is false: for instance, for $q \in (0, 1]$, consider a distribution P supported on five points in $\mathbb{R}^2$, with

$$P(\{(0, 0)\}) = q,$$
$$P(\{(-1, -1)\}) = P(\{(-1, 1)\}) = P(\{(1, -1)\}) = P(\{(1, 1)\}) = (1 - q)/4.$$

Then it can be shown that $\psi^*(P)$ is the uniform density on the square $[-1, 1] \times [-1, 1]$ for $q \in (0, 1/3]$.

In a similar spirit, it is not necessarily the case that the log-concave projection of a marginal of a joint distribution is equal to the corresponding marginal of the log-concave projection of the joint distribution. For example, if P is the discrete uniform distribution on the three points $\{(-1, -1), (0, 3^{1/2} - 1), (1, -1)\}$ in $\mathbb{R}^2$ (which form an equilateral triangle), then the log-concave projection is the continuous uniform density on the triangle, with corresponding marginal density $f_1(x_1) = (1 - |x_1|)\mathbb{1}_{\{|x_1| \le 1\}}$ on the x-axis. On the other hand, the log-concave projection of the discrete uniform distribution on $\{-1, 0, 1\}$ is the uniform density on $[-1, 1]$.

We conclude this section by mentioning two further properties that are not preserved by log-concave projection, namely stochastic ordering and convolution. More precisely, regarding stochastic ordering, let P and Q be distributions on the real line with[1] $P(\{0\}) = P(\{1\}) = 1/2$ and $Q(\{0\}) = 1/2$, $Q(\{1\}) = 2/5$, $Q(\{2\}) = 1/10$. Then P is stochastically smaller than Q, in the sense that the respective distribution

[1] I thank Min Xu and Yining Chen for helpful conversations leading to this example.

functions F and G satisfy $F(x) \geq G(x)$ with strict inequality for some x_0. Now $\psi^*(P)$ is the uniform density on $[0,1]$, while it can be shown using the ideas in Section 5.1 below that $\psi^*(Q)(x) = e^{bx-\beta}$ for $x \in [0,2]$, where $b \in [-1.337, -1.336]$ is the unique real solution to

$$\frac{1}{b} - \frac{2}{e^{2b}-1} = \frac{7}{5},$$

and where $\beta = \log(\frac{e^{2b}-1}{b}) \in [-0.3619, -0.3612]$. In particular, $\psi^*(Q)(0) = e^{-\beta} \geq 1.4 > 1 = \psi^*(P)(0)$, so $\psi^*(P)$ is not stochastically smaller than $\psi^*(Q)$; see Figure 5.

To see that log-concave projection does not preserve convolution in general,[2] let $P(\{0\}) = P(\{1\}) = 1/2$. Then $Q := P * P$ satisfies $Q(\{0\}) = Q(\{2\}) = 1/4$ and $Q(\{1\}) = 1/2$. We know that $\psi^*(P)$ is the uniform density on $[0,1]$, but $\psi^*(Q)$ maximises

$$\frac{1}{4}\log f(0) + \frac{1}{2}\log f(1) + \frac{1}{4}\log f(2)$$

over $f \in \mathcal{F}_1$, so is log-linear on $[0,1]$ and on $[1,2]$. In particular, $\psi^*(Q)$ is not equal to the triangular density on $[0,2]$, so $\psi^*(Q) \neq \psi^*(P) * \psi^*(P)$ in this example.

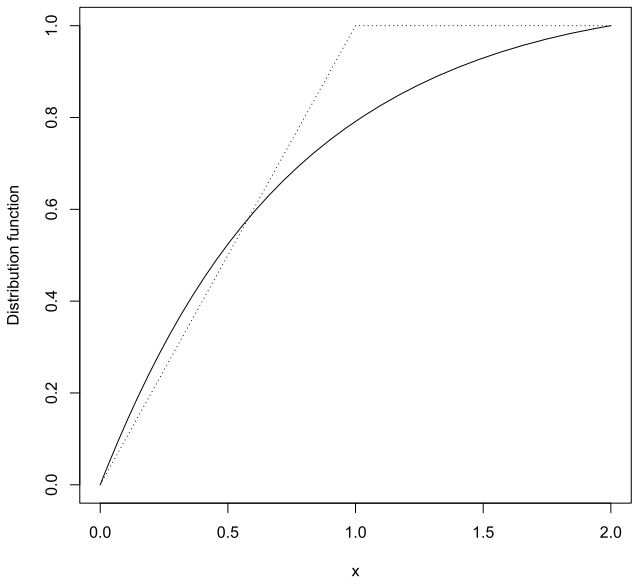

FIG. 5. *The distribution functions corresponding to $\psi^*(P)$ (dotted) and $\psi^*(Q)$ (solid) in the stochastic ordering example at the end of Section 5.*

[2]The question of whether or not log-concave projection preserves convolution was asked to me by Varun Jog.

5.1 The One-Dimensional Case

When $d = 1$, the log-concave projection can be characterised in terms of its integrated distribution function. For $\phi \in \Phi$, let

$$\mathcal{S}(\phi) := \Big\{ x \in \mathrm{dom}(\phi) : \phi(x) > \frac{1}{2}\{\phi(x+\delta) + \phi(x-\delta)\} \text{ for all } \delta > 0 \Big\}$$

denote the closed subset of $\mathbb{R}$ consisting of the points x_0 where ϕ is not affine in a neighbourhood of x_0.

THEOREM 5.5 (Dümbgen, Samworth and Schuhmacher, 2011). *Let $P \in \mathcal{P}_1$ have distribution function F, and let F^* be a distribution function with density $f^* = e^{\phi^*} \in \mathcal{F}_1$. Then $f^* = \psi^*(P)$ if and only if*

$$\int_{-\infty}^{x} \{F^*(t) - F(t)\}\,dt \begin{cases} \leq 0 & \text{for all } x \in \mathbb{R}, \\ = 0 & \text{for all } x \in \mathcal{S}(\phi^*) \cup \{\infty\}. \end{cases}$$

In particular, if P is absolutely continuous with respect to Lebesgue measure with continuous density f, and if $\mathcal{S}(\log \psi^*(P))$ contains an open interval I, then $\psi^*(P) = f$ on I. Theorem 5.5 is especially useful as a way of verifying the form of log-concave projection in cases where one can guess what it might be. For instance, consider the family of symmetrised Pareto densities

$$f(x;\alpha,\sigma) := \frac{\alpha\sigma^\alpha}{2(|x|+\sigma)^{\alpha+1}}, \quad x \in \mathbb{R}, \alpha > 1, \sigma > 0.$$

Theorem 5.5 can be used to verify that the corresponding log-concave projection is

$$f^*(x;\alpha,\sigma) = \frac{\alpha-1}{2\sigma}\exp\Big\{-\frac{(\alpha-1)|x|}{\sigma}\Big\}, \quad x \in \mathbb{R};$$

see Chen and Samworth (2013). Since the preimage under ψ^* of any $f \in \mathcal{F}_d$ is a convex set, this shows that the preimage of the Laplace density $x \mapsto e^{-|x|}/2$ is infinite-dimensional. Theorem 5.5 can also be used to show results such as the following proposition.

PROPOSITION 5.6 (Dümbgen, Samworth and Schuhmacher, 2011). *Suppose that $P \in \mathcal{P}_1$ has log-density ϕ that is differentiable, convex on a bounded interval $[a,b]$ and concave on $(-\infty,a] \cup [b,\infty)$. Then there exist $a' \in (-\infty,a]$ and $b' \in [b,\infty)$ such that $\log \psi^*(P)$ is affine on $[a',b']$ and $\log \psi^*(P) = \phi$ on $(-\infty,a'] \cup [b',\infty)$.*

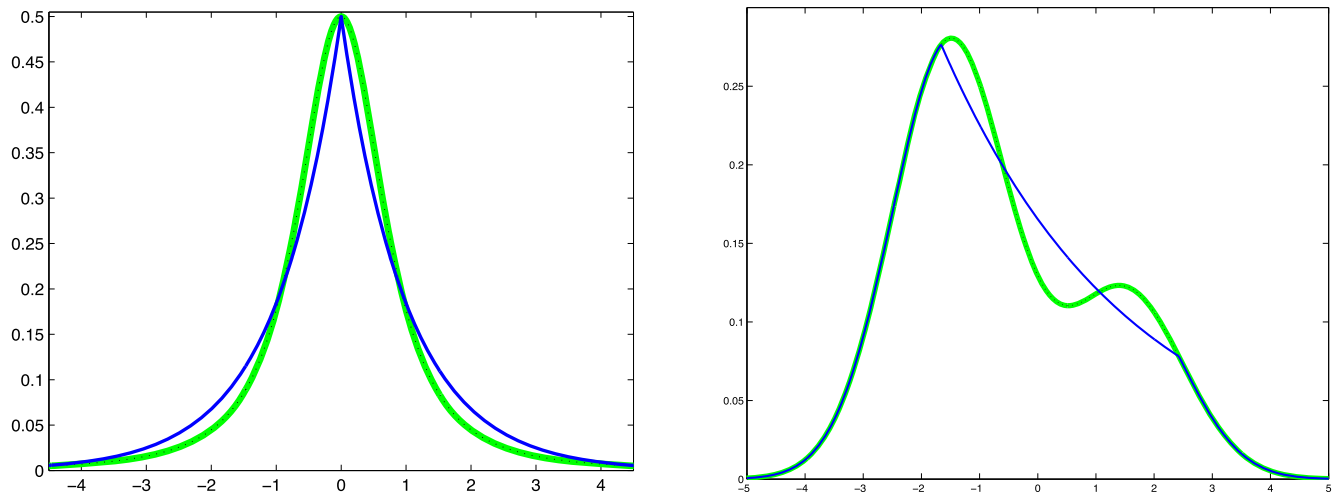

FIG. 6. *Left: the scaled t_2 density $f(x) = (1+x^2)^{-3/2}/2$ (green) and its Laplace log-concave projection $f^*(x) = e^{-|x|}/2$ (blue). Right: the density of the normal mixture $0.7N(-1.5,1) + 0.3N(1.5,1)$ (green) together with its log-concave projection (blue); the normal mixture satisfies the conditions of Proposition 5.6.*

These ideas are illustrated in Figure 6, taken from Dümbgen, Samworth and Schuhmacher (2011).

6. STRONGER FORMS OF CONVERGENCE AND CONSISTENCY

In minor abuse of standard notation, if (f_n), f are densities on $\mathbb{R}^d$, we write $f_n \xrightarrow{d} f$ to mean $\int_{\mathbb{R}^d} g(x) f_n(x)\,dx \to \int_{\mathbb{R}^d} g(x) f(x)\,dx$ for all bounded continuous functions $g : \mathbb{R}^d \to \mathbb{R}$. The constraint of log-concavity rules out certain pathologies and means we can strengthen certain convergence statements:

THEOREM 6.1 (Cule and Samworth, 2010, Schuhmacher, Hüsler and Dümbgen, 2011). *Let (f_n) be a sequence in $\mathcal{F}_d$ with $f_n \xrightarrow{d} f$ for some density f on $\mathbb{R}^d$. Then f is log-concave. Moreover, if $\alpha_0 > 0$ and $\beta_0 \in \mathbb{R}$ are such that $f(x) \le e^{-\alpha_0\|x\|+\beta_0}$ for all $x \in \mathbb{R}^d$, then for all $\alpha < \alpha_0$,*

$$\int_{\mathbb{R}^d} e^{\alpha\|x\|} \big| f_n(x) - f(x) \big|\,dx \to 0$$

as $n \to \infty$.

Thus, in the presence of log-concavity, convergence in distribution statements automatically yield convergence in certain exponentially weighted total variation distances.

A very natural question about log-concave projections, with important implications for the consistency of the log-concave maximum likelihood estimator, is "In what sense does a distribution $Q \in \mathcal{P}_d$ need to be close to $P \in \mathcal{P}_d$ in order for $\psi^*(Q)$ to be close to $\psi^*(P)$"? To answer this, we first recall that the Mallows-1 distance[3] d_1 between probability measures P, Q on $\mathbb{R}^d$ with finite first moment is given by

$$d_1(P,Q) := \inf_{(X,Y)\sim(P,Q)} \mathbb{E}\|X - Y\|,$$

where the infimum is taken over all pairs of random vectors (X, Y) defined on the same probability space with $X \sim P$ and $Y \sim Q$. It is well known that $d_1(P_n, P) \to 0$ if and only if both $P_n \xrightarrow{d} P$ and $\int_{\mathbb{R}^d} \|x\|\,dP_n(x) \to \int_{\mathbb{R}^d} \|x\|\,dP(x)$.

THEOREM 6.2 (Dümbgen, Samworth and Schuhmacher, 2011). *Suppose that $P \in \mathcal{P}_d$ and that $d_1(P_n, P) \to 0$. Then $L^*(P_n) \to L^*(P)$, $P_n \in \mathcal{P}_d$ for sufficiently large n, and, taking $\alpha_0 > 0$ and $\beta_0 \in \mathbb{R}$ such that $\psi^*(P)(x) \le e^{-\alpha_0\|x\|+\beta_0}$ for all $x \in \mathbb{R}^d$, we have for $\alpha < \alpha_0$ that*

$$\int_{\mathbb{R}^d} e^{\alpha\|x\|} \big| \psi^*(P_n)(x) - \psi^*(P)(x) \big|\,dx \to 0$$

as $n \to \infty$.

The Mallows convergence cannot in general be weakened to $P_n \xrightarrow{d} P$. In particular, if $P = U\{-1,1\}$ and $P_n = (1-n^{-1})U\{-1,1\} + n^{-1}U\{-(n+1), n+1\}$, then $P_n \xrightarrow{d} P$ but it can be shown that

$$\int_{-\infty}^{\infty} |\psi^*(P_n) - \psi^*(P)| \to \frac{4}{5^{1/2}+1}.$$

[3] Also known as the Wasserstein distance, Monge–Kantorovich distance and Earth Mover's distance.

Writing $d_{\mathrm{TV}}(f,g) := \frac{1}{2}\int_{\mathbb{R}^d}|f-g|$, Theorem 6.2 implies that the log-concave projection ψ^* is continuous when considered as a map between the metric spaces $(\mathcal{P}_d, d_1)$ and $(\mathcal{F}_d, d_{\mathrm{TV}})$. However, it is not uniformly continuous: for instance, let $P_n = U[-1/n, 1/n]$ and $Q_n = U[-1/n^2, 1/n^2]$. Then $d_1(P_n, Q_n) = \frac{1}{2n} - \frac{1}{2n^2} \to 0$, but since $\psi^*(P_n)(x) = \frac{n}{2}\mathbb{1}_{\{x\in[-1/n,1/n]\}}$ and $\psi^*(Q_n)(x) = \frac{n^2}{2}\mathbb{1}_{\{x\in[-1/n^2,1/n^2]\}}$, we have

$$d_{\mathrm{TV}}\big(\psi^*(P_n), \psi^*(Q_n)\big) = 1 - \frac{1}{n} \to 1.$$

One of the great advantages of working in the general framework of log-concave projections for arbitrary $P \in \mathcal{P}_d$, as opposed to simply focusing on empirical distributions, is that one can study analytical properties of the projection as above, meaning that the only probabilistic arguments required to deduce convergence statements about the log-concave maximum likelihood estimator are simple facts about the convergence of the empirical distribution. This is illustrated in the following corollary.

COROLLARY 6.3 (Dümbgen, Samworth and Schuhmacher, 2011). *Suppose that $X_1, X_2, \ldots$ are independent and identically distributed with distribution $P \in \mathcal{P}_d$, and let $\mathbb{P}_n$ denote the empirical distribution of $X_1, \ldots, X_n$. Then, with probability one, $\hat{f}_n := \psi^*(\mathbb{P}_n)$ is well defined for sufficiently large n, and taking $\alpha_0 > 0$ and $\beta_0 \in \mathbb{R}$ such that $f^*(x) := \psi^*(P)(x) \le e^{-\alpha_0\|x\|+\beta_0}$ for all $x \in \mathbb{R}^d$, we have for $\alpha < \alpha_0$ that*

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

as $n \to \infty$.

PROOF. let $\mathcal{H} := \{h : \mathbb{R}^d \to [-1,1] : |h(x) - h(y)| \le \|x-y\| \text{ for all } x, y \in \mathbb{R}^d\}$, and define the bounded Lipschitz distance between probability measures P and Q on $\mathbb{R}^d$ by

$$d_{\mathrm{BL}}(P,Q) := \sup_{h\in\mathcal{H}}\int_{\mathbb{R}^d} h\,d(P-Q).$$

Then d_{BL} metrises convergence in distribution for probability measures on $\mathbb{R}^d$, and from Varadarajan's theorem (Dudley, 2002, Theorem 11.4.1), we deduce that $d_{\mathrm{BL}}(\mathbb{P}_n, P) \xrightarrow{\text{a.s.}} 0$. In particular, since the set of probability measures P on $\mathbb{R}^d$ with $P(H) < 1$ for all hyperplanes H is an open subset of the set of all probability measures on $\mathbb{R}^d$ in the topology of weak convergence (Dümbgen, Samworth and Schuhmacher, 2011, Lemma 2.13), it follows that with probability one, $\mathbb{P}_n \in \mathcal{P}_d$ for sufficiently large n, and $\hat{f}_n$ is well defined for such n.

Since we also have

$$\int_{\mathbb{R}^d}\|x\|\,d\mathbb{P}_n(x) \xrightarrow{\text{a.s.}} \int_{\mathbb{R}^d}\|x\|\,dP(x)$$

by the strong law of large numbers, it follows that $d_1(\mathbb{P}_n, P) \xrightarrow{\text{a.s.}} 0$. The second part of the result therefore follows by Theorem 6.2. □

Corollary 6.3 yields the (strong) consistency of the log-concave maximum likelihood estimator in exponentially weighted total variation distances, and also provides a robustness to misspecification guarantee in the case where the true distribution P does not have a log-concave density.

7. RATES OF CONVERGENCE AND ADAPTATION

Historically, a great deal of effort has gone into understanding rates of convergence in shape-constrained estimation problems, with both local (pointwise) and global rates being considered. For the log-concave maximum likelihood estimator, the following result, a special case of Balabdaoui, Rufibach and Wellner (2009), Theorem 2.1, establishes the pointwise rates of convergence in the case $d = 1$.

THEOREM 7.1 (Balabdaoui, Rufibach and Wellner, 2009). *Let $X_1, \ldots, X_n \overset{\text{i.i.d.}}{\sim} f_0 \in \mathcal{F}_1$, let $f_0(x_0) > 0$ and suppose that $\phi_0 := \log f_0$ is twice continuously differentiable in a neighbourhood of x_0 with $\phi_0''(x_0) < 0$. Let W be a standard two-sided Brownian motion on $\mathbb{R}$, and let*

$$Y(t) := \begin{cases} \int_0^t W(s)\,ds - t^4 & \text{for } t \ge 0, \\ \int_t^0 W(s)\,ds - t^4 & \text{for } t < 0. \end{cases}$$

Then the log-concave maximum likelihood estimator $\hat{f}_n$ satisfies

$$n^{2/5}\{\hat{f}_n(x_0) - f_0(x_0)\} \xrightarrow{d} \left(\frac{f_0(x_0)^3|\phi_0''(x_0)|}{24}\right)^{1/5} H''(0), \tag{7.1}$$

where $\{H(t) : t \in \mathbb{R}\}$ is the "lower invelope" process of Y, so that $H(t) \le Y(t)$ for all $t \in \mathbb{R}$, H'' is concave and $H(t) = Y(t)$ if the slope of H'' decreases strictly at t.

This lower invelope process was introduced and studied in detail in Groeneboom, Jongbloed and Wellner (2001a). The nonstandard limiting distribution is

characteristic of shape-constrained estimation problems. Balabdaoui, Rufibach and Wellner (2009) study the more general case where more than two derivatives of ϕ_0 may vanish at x_0, in which case a faster rate is obtained; they also study the joint convergence of $\hat{f}_n$ with its derivative $\hat{f}_n'$. The pointwise convergence rate in d dimensions remains an open problem, though Seregin and Wellner (2010) obtained a minimax lower bound for pointwise estimation at x_0 with respect to absolute error loss of order $n^{-2/(d+4)}$, provided ϕ_0 is twice continuously differentiable in a neighbourhood of x_0 and the determinant of the Hessian matrix of ϕ_0 at x_0 does not vanish. This is the familiar rate attained by, e.g. kernel density estimators, under similar smoothness conditions but without the log-concavity assumption.

An interesting feature of (7.1) is that the limiting distribution depends in a complicated way on the unknown true density. This makes it challenging to apply this result directly to construct confidence intervals for $f_0(x_0)$. However, in the special case where x_0 is the mode of f_0, Doss and Wellner (2016a) have recently proposed an approach for confidence interval construction based on comparing the log-concave MLE at x_0 with the constrained MLE $\hat{f}_n^0$, say, where the mode of the density is fixed at $m \in \mathbb{R}$, say. Their key observation is that, under the null hypothesis that the log-concave density f_0 attains its maximum at m, and provided $(\log f_0)''(m) < 0$, the likelihood ratio statistic is asymptotically pivotal, in the sense that

$$2\log\lambda_n := 2\sum_{i=1}^{n} \log \frac{\hat{f}_n(X_i)}{\hat{f}_n^0(X_i)} \xrightarrow{d} \mathbb{D},$$

where $\mathbb{D}$ is a universal limiting distribution (not depending on f_0). Under the alternative hypothesis that the log-concave density f_0 does not have a mode at m, the statistic λ_n tends to be inflated; in fact, $(2/n)\log\lambda_n$ converges in probability to a deterministic, positive limit.

We now turn to global rates of convergence, and write $d_{\mathrm{H}}^2(f,g) := \int_{\mathbb{R}^d}(f^{1/2} - g^{1/2})^2$ for the squared Hellinger distance between densities f and g. The same rate as for pointwise estimation had been expected in the light of the facts that any concave function on $\mathbb{R}^d$ is twice differentiable (Lebesgue) almost everywhere in its domain (Aleksandrov, 1939), and that for twice continuously differentiable functions, concavity is equivalent to a second derivative condition, namely that the Hessian matrix is nonpositive definite. The following minimax lower bound therefore came as a surprise:

THEOREM 7.2 (Kim and Samworth, 2016). *Let $X_1, \ldots, X_n \overset{\text{i.i.d.}}{\sim} f_0 \in \mathcal{F}_d$, and let $\tilde{F}_n$ denote the set of all estimators of f_0 based on $X_1, \ldots, X_n$. Then for each $d \in \mathbb{N}$, there exists $c_d > 0$ such that*

$$\inf_{\tilde{f}_n \in \tilde{\mathcal{F}}_n} \sup_{f_0 \in \mathcal{F}_d} \mathbb{E}_{f_0} d_{\mathrm{H}}^2(\tilde{f}_n, f_0) \geq \begin{cases} c_1 n^{-4/5} & \text{if } d = 1, \\ c_d n^{-2/(d+1)} & \text{if } d \geq 2. \end{cases}$$

Theorem 7.2 yields the expected lower bound when $d = 1, 2$ [note that $2/(d+1) = 4/(d+4) = 2/3$ when $d = 2$]. However, it also reveals that log-concave density estimation in three or more dimensions is fundamentally more challenging in this minimax sense than estimating a density with two bounded derivatives. The reason is that although log-concave densities are twice differentiable almost everywhere, they can be badly behaved (in particular, discontinuous) on the boundary of their support; recall that uniform densities on convex, compact sets in $\mathbb{R}^d$ belong to $\mathcal{F}_d$. It turns out that it is the difficulty of estimating the support of the density that drives the rate in these higher dimensions.

The following complementary result provides the corresponding global rate of convergence for the log-concave MLE in squared Hellinger distance in low-dimensional cases.

THEOREM 7.3 (Kim and Samworth, 2016). *Let $X_1, \ldots, X_n \overset{\text{i.i.d.}}{\sim} f_0 \in \mathcal{F}_d$, and let $\hat{f}_n$ denote the log-concave MLE based on $X_1, \ldots, X_n$. Then*

$$\sup_{f_0 \in \mathcal{F}_d} \mathbb{E}_{f_0} d_{\mathrm{H}}^2(\hat{f}_n, f_0) = \begin{cases} O\big(n^{-4/5}\big) & \text{if } d = 1, \\ O\big(n^{-2/3}\log n\big) & \text{if } d = 2, \\ O\big(n^{-1/2}\log n\big) & \text{if } d = 3. \end{cases}$$

Thus the log-concave MLE attains the minimax optimal rate in terms of squared Hellinger risk when $d = 1$, and attains the minimax optimal rate up to logarithmic factors when $d = 2, 3$. We mention that in the case $d = 1$, Doss and Wellner (2016b) proved that $d_{\mathrm{H}}^2(\hat{f}_n, f_0) = O_p(n^{-4/5})$ for each fixed $f_0 \in \mathcal{F}_1$, and indeed showed that the same rate holds for the MLEs over classes of s-concave densities with $s > -1$; see Section 9.1. The proofs of these results rely on empirical process theory and delicate bracketing entropy bounds for the relevant class of log-concave densities, made more complicated by the fact that the domains of the log-densities can be an arbitrary d-dimensional closed, convex set. The argument proceeds by approximating these domains by convex polytopes, which can be triangulated into simplices, and appropriate bracketing entropy bounds for concave functions on such domains are known (e.g., Gao and Wellner, 2017). Critically, when $d \leq 3$, the region between two nested convex polytopes with p and q vertices respectively can

be triangulated into $O(p+q)$ simplices (e.g., Brass, 2005).

Although Theorem 7.3 provides strong guarantees on the worst case performance of the log-concave MLE in low-dimensional cases, it ignores one of the appealing features of the estimator, namely its potential to adapt to certain characteristics of the unknown true density. Dümbgen and Rufibach (2009) obtained the first such result in the case $d=1$. Recall that given an interval I, $\beta \in [1,2]$ and $L>0$, we say $h:\mathbb{R}\to\mathbb{R}$ belongs to the Hölder class $\mathcal{H}_{\beta,L}(I)$ if for all $x,y\in I$, we have

$$|h(x)-h(y)| \le L|x-y|, \quad \text{if } \beta = 1,$$

$$|h'(x)-h'(y)| \le L|x-y|^{\beta-1}, \quad \text{if } \beta > 1.$$

THEOREM 7.4 (Dümbgen and Rufibach, 2009). *Let $X_1,\dots,X_n \overset{\text{i.i.d.}}{\sim} f_0 \in \mathcal{F}_1$, and assume that $\phi_0 := \log f_0 \in \mathcal{H}_{\beta,L}(I)$ for some $\beta\in[1,2]$, $L>0$ and compact interval $I \subseteq \operatorname{int}(\operatorname{dom}(\phi_0))$. Then*

$$\sup_{x_0\in I} |\hat{f}_n(x_0) - f_0(x_0)| = O_p\left(\left(\frac{\log n}{n}\right)^{\beta/(2\beta+1)}\right).$$

Here the log-concave MLE is adapting to unknown smoothness. When measuring loss in the supremum norm, the need to restrict attention to a compact interval in the interior of support of f_0 is suggested by the right-hand plot in Figure 1.

Other adaptation results are motivated by the thought that since the log-concave MLE is piecewise affine, we might hope for faster rates of convergence in cases where $\log f_0$ is made up of a relatively small number of affine pieces. We now describe two such results. For $k\in\mathbb{N}$ we define $\mathcal{F}^k$ to be the class of log-concave densities f on $\mathbb{R}$ for which $\log f$ is k-affine in the sense that there exist intervals $I_1,\dots,I_k$ such that f is supported on $I_1\cup\cdots\cup I_k$, and $\log f$ is affine on each I_j. In particular, densities in $\mathcal{F}^1$ are uniform or (possibly truncated) exponential, and can be parametrised as

$$f_{\alpha,s_1,s_2}(x) := \begin{cases} \dfrac{1}{s_2-s_1}\mathbb{1}_{\{x\in[s_1,s_2]\}} & \text{if } \alpha = 0, \\ \dfrac{\alpha}{e^{\alpha s_2}-e^{\alpha s_1}} e^{\alpha x}\mathbb{1}_{\{x\in[s_1,s_2]\}} & \text{if } \alpha \neq 0, \end{cases}$$

for $(\alpha,s_1,s_2)\in\mathcal{T} := (\mathbb{R}\times\mathcal{T}_0)\cup((0,\infty)\times\{-\infty\}\times\mathbb{R})\cup((-\infty,0)\times\mathbb{R}\times\{\infty\})$, where $\mathcal{T}_0 := \{(s_1,s_2)\in\mathbb{R}^2 : s_1<s_2\}$. Define a continuous, strictly increasing function $\rho:\mathbb{R}\to(0,\infty)$ by

$$\rho(x) := \begin{cases} \dfrac{2e^x(x-1)-x^2+2}{2e^x-2-2x-x^2} & \text{if } x\neq 0, \\ 2 & \text{if } x = 0; \end{cases} \tag{7.2}$$

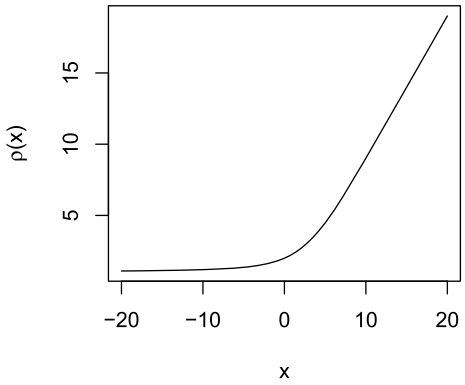

FIG. 7. *The function ρ defined in (7.2).*

cf. Figure 7. It can be shown that $\rho(x)\le\max\{\rho(2),\rho(x)\}\le\max(3,2x)$ for all $x\in\mathbb{R}$.

THEOREM 7.5 (Kim, Guntuboyina and Samworth, 2018). *Let $X_1,\dots,X_n \overset{\text{i.i.d.}}{\sim} f_{\alpha,s_1,s_2}\in\mathcal{F}^1$ with $n\ge 5$, and let $\hat{f}_n$ denote the log-concave MLE. Then, writing $\kappa^* := \alpha(s_2-s_1)$,*

$$\mathbb{E}_{f_0} d_{\mathrm{TV}}(\hat{f}_n, f_0) \le \frac{\min\{2\rho(|\kappa^*|), 6\log n\}}{n^{1/2}}.$$

In fact, Theorem 7.5 is a special case of the result given in Kim, Guntuboyina and Samworth (2018), which allows the true density f_0 to be arbitrary, and includes an additional approximation error term that measures the proximity of f_0 to the class $\mathcal{F}^1$. An important consequence of Theorem 7.5 is the fact that if $|\alpha|$ is small, then the log-concave MLE can attain the parametric rate of convergence in total variation distance. In particular, if f_0 is a uniform density on a compact interval (so that $\kappa^*=0$), then $\mathbb{E}_{f_0} d_{\mathrm{TV}}(\hat{f}_n,f_0)\le 4/n^{1/2}$; cf. the right-hand plot of Figure 1 again. Interestingly, this behaviour is in stark contrast to that of the least squares convex regression estimator with respect to squared error loss in the random design problem where covariates are uniformly distributed on $[0,1]$ and the responses are uniform on $\{-1,1\}$: in that case, the true regression function is zero, but the risk of the estimator is infinite (Balázs, Györgi and Szepesvári, 2015)! The proof of Theorem 7.5 relies on a version of Marshall's inequality for log-concave density estimation. A special case of this result states that if $X_1,\dots,X_n \overset{\text{i.i.d.}}{\sim} f_{\alpha,s_1,s_2}\in\mathcal{F}^1$, then writing $X_{(1)} := \min_i X_i$, $X_{(n)} := \max_i X_i$ and $\kappa := \alpha(X_{(n)}-X_{(1)})$, we have

$$\sup_{x\in\mathbb{R}} |\hat{F}_n(x) - F_0(x)| \le \rho(|\kappa|) \sup_{x\in\mathbb{R}} |\mathbb{F}_n(x) - F_0(x)|, \tag{7.3}$$

where F_0 and $\hat{F}_n$ denote the distribution functions corresponding to the true density and the log-concave MLE respectively, and where $\mathbb{F}_n$ denotes the empirical distribution function.[4]

We now aim to generalise these ideas to situations where f_0 is close to $\mathcal{F}^k$, but assume only that $X_1, \ldots, X_n \overset{\text{i.i.d.}}{\sim} f_0 \in \mathcal{F}_1$. An application of Lemma 5.2 to the function $\Delta(x) = \log \frac{f_0(x)}{\hat{f}_n(x)}$ yields

$$d^2_{\mathrm{KL}}(\hat{f}_n, f_0) \leq \frac{1}{n}\sum_{i=1}^{n} \log \frac{\hat{f}_n(X_i)}{f_0(X_i)} =: d^2_X(\hat{f}_n, f_0).$$

In particular, an upper bound on $d^2_X(\hat{f}_n, f_0)$ immediately provides corresponding bounds on $d^2_{\mathrm{TV}}(\hat{f}_n, f_0)$, $d^2_{\mathrm{H}}(\hat{f}_n, f_0)$ and $d^2_{\mathrm{KL}}(\hat{f}_n, f_0)$.

THEOREM 7.6 (Kim, Guntuboyina and Samworth, 2018). *There exists a universal constant $C > 0$ such that for $n \geq 2$,*

$$\mathbb{E}_{f_0} d^2_X(\hat{f}_n, f_0) \leq \min_{k=1,\ldots,n} \left\{ \frac{Ck}{n} \log^{5/4} \frac{en}{k} + \inf_{f_k \in \mathcal{F}^k} d^2_{\mathrm{KL}}(f_0, f_k) \right\}.$$

To help understand this theorem, first consider the case where $f_0 \in \mathcal{F}^k$. Then $\mathbb{E}_{f_0} d^2_X(\hat{f}_n, f_0) \leq \frac{Ck}{n} \cdot \log^{5/4}(en/k)$, which is nearly the parametric rate when k is small. More generally, this rate holds when $f_0 \in \mathcal{F}_1$ is only close to $\mathcal{F}^k$ in the sense that the approximation error $d^2_{\mathrm{KL}}(f_0, f_k)$ is $O(\frac{k}{n} \log^{5/4} \frac{en}{k})$. The result is known as a "sharp" oracle inequality, because the leading constant for this approximation error term is 1. See also Baraud and Birgé (2016), who also obtain an oracle inequality for their general ρ-estimation procedure. It is worth noting that the techniques of proof, which rely on empirical process theory and local bracketing entropy bounds, are completely different from those used in the proof of Theorem 7.5.

8. HIGHER-DIMENSIONAL PROBLEMS

The minimax lower bound in Theorem 7.2 is relatively discouraging for the prospects of log-concave density estimation in higher dimensions. It is natural, then, to consider additional structures that reduce the complexity of the class $\mathcal{F}_d$, thereby increasing the potential for applications outside low-dimensional settings. The purpose of this section is to explore two ways of imposing such structures, namely through independence and symmetry constraints.

In the simplest, noiseless case of Independent Component Analysis (ICA), one observes independent replicates of a random vector

$$X := AS, \tag{8.1}$$

where $A \in \mathbb{R}^{d \times d}$ is a deterministic, invertible matrix, and S is a d-dimensional random vector with independent components. One can think of the model as being the density estimation analogue of multiple index models in regression. ICA models have found an enormous range of applications across signal processing, machine learning and medical imaging, to name just a few; see Hyvärinen, Karhunen and Oja (2001) for an introduction to the field. The main interest is in estimating the unmixing matrix $W := A^{-1}$, with estimation of the marginal distributions of the components of S as a secondary goal. Let $\mathcal{W}$ denote the set of all invertible $d \times d$ real matrices, let $\mathcal{B}_d$ denote the set of all Borel subsets of $\mathbb{R}^d$, and let $\mathcal{P}_d^{\mathrm{ICA}}$ denote the set of $P \in \mathcal{P}_d$ with

$$P(B) = \prod_{j=1}^{d} P_j(w_j^\top B) \quad \forall B \in \mathcal{B}_d,$$

for some $W = (w_1, \ldots, w_d)^\top \in \mathcal{W}$ and $P_1, \ldots, P_d \in \mathcal{P}_1$. Thus $\mathcal{P}_d^{\mathrm{ICA}}$ is the set of distributions of random vectors X with $\mathbb{E}(\|X\|) < \infty$ satisfying (8.1). As stated, the model (8.1) is not identifiable, as we can write $X = ADPP^\top D^{-1}S$, where D is a diagonal $d \times d$ matrix with nonzero diagonal entries, and $P \in \mathbb{R}^{d \times d}$ is a permutation matrix (note that ADP is invertible and $P^\top D^{-1} S$ has independent components). Fortunately, these can be regarded as "trivial" lack of identifiability problems, because it is typically the directions of the set of rows of $W := A^{-1}$ that are of interest, not their order or magnitude. Eriksson and Koivunen (2004) proved that the pair of conditions that none of $P_1, \ldots, P_d$ are Dirac point masses and at most one of them is Gaussian is necessary and sufficient for the ICA model to be identifiable up to the permutation and scaling transformations described above.

Now let $\mathcal{F}_d^{\mathrm{ICA}}$ denote the set of $f \in \mathcal{F}_d$ with

$$f(x) = |\det W| \prod_{j=1}^{d} f_j(w_j^\top x)$$

for some $W = (w_1, \ldots, w_d)^\top \in \mathcal{W}$ and $f_1, \ldots, f_d \in \mathcal{F}_1$. In this way, $\mathcal{F}_d^{\mathrm{ICA}}$ is the set of densities of random vectors X satisfying (8.1), where each component of S

[4] The original Marshall's inequality (Marshall, 1970) applies to the (integrated) Grenander estimator, in which context $\rho(|\kappa|)$ in (7.3) may be replaced by 1.

has a log-concave density. Define the log-concave ICA projection on $\mathcal{P}_d$ by

$$\psi^{**}(P) := \operatorname*{argmax}_{f \in \mathcal{F}_d^{\mathrm{ICA}}} \int_{\mathbb{R}^d} \log f \, dP.$$

In general, $\psi^{**}(P)$ only defines a nonempty, proper subset of $\mathcal{F}_d^{\mathrm{ICA}}$ rather than a unique element. However, the following theorem gives uniqueness in an important special case, and the form of the log-concave ICA projection here is key to the success of this approach to fitting ICA models.

THEOREM 8.1 (Samworth and Yuan, 2012). *If $P \in \mathcal{P}_d^{\mathrm{ICA}}$, then $\psi^{**}(P)$ defines a unique element of $\mathcal{F}_d^{\mathrm{ICA}}$. In fact, the restrictions of ψ^{**} and ψ^* to $\mathcal{P}_d^{\mathrm{ICA}}$ coincide. Moreover, suppose that $P \in \mathcal{P}_d^{\mathrm{ICA}}$, so*

$$P(B) = \prod_{j=1}^{d} P_j(w_j^\top B) \quad \forall B \in \mathcal{B}_d,$$

for some $W = (w_1, \ldots, w_d)^\top \in \mathcal{W}$ and $P_1, \ldots, P_d \in \mathcal{P}_1$. Then $f^{**} := \psi^{**}(P)$ can be written explicitly as

$$f^{**}(x) = |\det W| \prod_{j=1}^{d} f_j^*(w_j^\top x),$$

where $f_j^* := \psi^*(P_j)$.

The fact that ψ^* preserves the ICA structure is a consequence of Lemma 5.1 and Proposition 5.4. However, the most interesting aspect of this result is the fact that the unmixing matrix W is preserved by the log-concave projection. This suggests that, at least from the point of view of estimating W, there is no loss of generality in assuming that the marginal distributions of the components of S have log-concave densities provided they have finite means. Another crucial result is the fact that the log-concave ICA projection of $P \in \mathcal{P}_d^{\mathrm{ICA}}$ does not sacrifice identifiability: in fact, $\psi^{**}(P)$ is identifiable if and only if P is identifiable.

Given data $X_1, \ldots, X_n \overset{\text{i.i.d.}}{\sim} P \in \mathcal{P}_d$ with empirical distribution $\mathbb{P}_n$, we can therefore fit an ICA model by computing $\hat{f}_n := \psi^{**}(\mathbb{P}_n)$. This estimator has similar consistency properties to the original log-concave projection, and requires the maximisation of

$$\ell(W, f_1, \ldots, f_d; X_1, \ldots, X_n) := \log|\det W| + \frac{1}{n}\sum_{i=1}^{n}\sum_{j=1}^{d} \log f_j(w_j^\top X_i)$$

over $W \in \mathcal{W}$ and $f_1, \ldots, f_d \in \mathcal{F}_1$. For reasons of numerical stability, however, it is convenient to "prewhiten" the estimator by setting $Z_i := \hat{\Sigma}^{-1/2} X_i$ for $i = 1, \ldots, n$, where $\hat{\Sigma}$ denotes the sample covariance matrix. We can then instead obtain a maximiser $(\hat{O}, \hat{g}_1, \ldots, \hat{g}_d)$ of $\ell(O, g_1, \ldots, g_d; Z_1, \ldots, Z_n)$ over $O \in O(d)$, the set of $d \times d$ orthogonal matrices, and $g_1, \ldots, g_d \in \mathcal{F}_1$, before setting $\hat{\hat{W}} := \hat{O}\hat{\Sigma}^{-1/2}$ and $\hat{\hat{f}}_j := \hat{g}_j$. This estimator has the same consistency properties as the original proposal, provided that $\int_{\mathbb{R}^d} \|x\|^2 \, dP(x) < \infty$. In effect, it breaks down the estimation of the d^2 parameters in W into two stages: first, we use $\hat{\Sigma}$ to estimate the $d(d+1)/2$ free parameters of the symmetric, positive definite matrix Σ, leaving only the maximisation over the $d(d-1)/2$ free parameters of $O \in O(d)$ at the second stage. Even after pre-whitening, however, there is an additional computational challenge relative to the original log-concave MLE caused by the fact that the objective function ℓ is only bi-concave[5] in O and $g_1, \ldots, g_d$, but not jointly concave in these arguments. Since we only have to deal with computation of univariate log-concave maximum likelihood estimators, however, marginal updates are straightforward, and taking the solution with highest log-likelihood over several random initial values for the variables can lead to satisfactory solutions (Samworth and Yuan, 2012).

Symmetry constraints provide another alternative approach to extending the scope of shape-constrained methods to higher dimensions. For simplicity of exposition, we focus on the simplest case of spherical symmetry, as studied recently by Xu and Samworth (2017), though more general symmetry constraints may also be considered. We write $\mathcal{F}_d^{\mathrm{SS}}$ for the set of spherically symmetric $f \in \mathcal{F}_d$, and let Φ^{SS} denote the class of upper semi-continuous, decreasing, concave functions $\phi : [0, \infty) \to [-\infty, \infty)$. The starting point for the symmetry-based approach is the observation that a density f on $\mathbb{R}^d$ belongs to $\mathcal{F}_d^{\mathrm{SS}}$ if and only if $f(x) = e^{\phi(\|x\|)}$ for some $\phi \in \Phi^{\mathrm{SS}}$. One can then define the notion of spherically symmetric log-concave projection, which has several similarities with the theory presented in Sections 3 and 5 (though with some notable differences, especially with regard to moment preservation properties). In particular, given data $X_1, \ldots, X_n \in \mathbb{R}^d$

[5] In other words, ℓ is concave in O for fixed $g_1, \ldots, g_d$, and concave in $g_1, \ldots, g_d$ for fixed O.

that are not all zero, there exists a unique spherically symmetric log-concave MLE $\hat{f}_n^{\mathrm{SS}}$. This estimator can be computed using a variant of the Active Set algorithm outlined in Section 4. Importantly, this algorithm only depends on d through the need to compute $Z_i := \|X_i\|$ for $i = 1, \ldots, n$ at the outset, and it therefore scales extremely well to high-dimensional cases, even when d may be in the hundreds of thousands.

The following worst case bound reveals that $\hat{f}_n^{\mathrm{SS}}$ succeeds in evading the curse of dimensionality:

THEOREM 8.2 (Xu and Samworth, 2017). *Let $f_0 \in \mathcal{F}_d^{\mathrm{SS}}$, let $X_1, \ldots, X_n \overset{\text{i.i.d.}}{\sim} f_0$, and let $\hat{f}_n^{\mathrm{SS}}$ denote the corresponding spherically symmetric log-concave MLE. Then there exists a universal constant $C > 0$ such that*

$$\sup_{f_0 \in \mathcal{F}_d^{\mathrm{SS}}} \mathbb{E} d_X^2(\hat{f}_n^{\mathrm{SS}}, f_0) \le C n^{-4/5}.$$

Similar to the ordinary log-concave MLE, we have $d_X^2(\hat{f}_n^{\mathrm{SS}}, f_0) \ge d_{\mathrm{KL}}^2(\hat{f}_n^{\mathrm{SS}}, f_0)$, and the interesting feature of this bound is that it does not depend on d. Nevertheless, a viable alternative, which also satisfies the same worst case risk bound, and which is equally straightforward to compute, is to let $\tilde{h}_n$ denote the (ordinary) log-concave MLE based on $Z_1, \ldots, Z_n$, and then set

$$\tilde{f}_n(x) := \begin{cases} \tilde{h}_n(\|x\|)/(c_d \|x\|^{d-1}) & \text{if } x \neq 0, \\ 0 & \text{if } x = 0, \end{cases} \tag{8.2}$$

where $c_d := 2\pi^{d/2}/\Gamma(d/2)$. This estimator, however, ignores the fact that the density of Z_1 is a "special" log-concave density, belonging to the class

$$\mathcal{H} := \Big\{ r \mapsto r^{d-1} e^{\phi(r)} : \phi \in \Phi^{\mathrm{SS}}, \int_0^\infty r^{d-1} e^{\phi(r)}\, dr = 1 \Big\},$$

and means that $\tilde{f}_n$ does not belong to $\mathcal{F}_d^{\mathrm{SS}}$ in general. Moreover, $\tilde{f}_n$ is inconsistent at $x = 0$ (the estimator is zero for $\|x\| < \min_i Z_i$) and behaves badly for small $\|x\|$; cf. Figure 8, taken from Xu and Samworth (2017).

A further advantage of $\hat{f}_n^{\mathrm{SS}}$ in this context relates to its adaptation behaviour. To describe this, for $k \in \mathbb{N}$, we say $\phi \in \Phi^{\mathrm{SS}}$ is *k-affine*, and write $\phi \in \Phi^{\mathrm{SS},k}$, if there exist $r_0 \in (0, \infty]$ and a partition $I_1, \ldots, I_k$ of $[0, r_0)$ into intervals such that ϕ is affine on each I_j for $j = 1, \ldots, k$, and $\phi(r) = -\infty$ for $r > r_0$. Define $\mathcal{H}^k := \{h \in \mathcal{H} : h(r) = r^{d-1} e^{\phi(r)} \text{ for some } \phi \in \Phi^{\mathrm{SS},k}\}$.

THEOREM 8.3 (Xu and Samworth, 2017). *Let $f_0 \in \mathcal{F}_d^{\mathrm{SS}}$ be given by $f_0(x) = e^{\phi_0(\|x\|)}$, where $\phi_0 \in \Phi^{\mathrm{SS}}$ and let $X_1, \ldots, X_n \overset{\text{i.i.d.}}{\sim} f_0$. Let $\hat{f}_n^{\mathrm{SS}}$ be the spherically symmetric log-concave MLE. Define $h_0 \in \mathcal{H}$ by $h_0(r) := r^{d-1} e^{\phi_0(r)}$ for $r \in [0, \infty)$. Then, writing $v_k^2 := 2 \wedge \inf_{h \in \mathcal{H}^k} d_{\mathrm{KL}}^2(h_0, h)$, there exists a universal constant $C > 0$ such that*

$$\mathbb{E} d_X^2(\hat{f}_n^{\mathrm{SS}}, f_0) \le C \min_{k=1,\ldots,n} \Big(\frac{k^{4/5} v_k^{2/5}}{n^{4/5}} \log \frac{en}{k v_k} + \frac{k}{n} \log^{5/4} \frac{en}{k} \Big).$$

Interestingly, this result implies the following sharp oracle inequality: there exists a universal constant $C >$

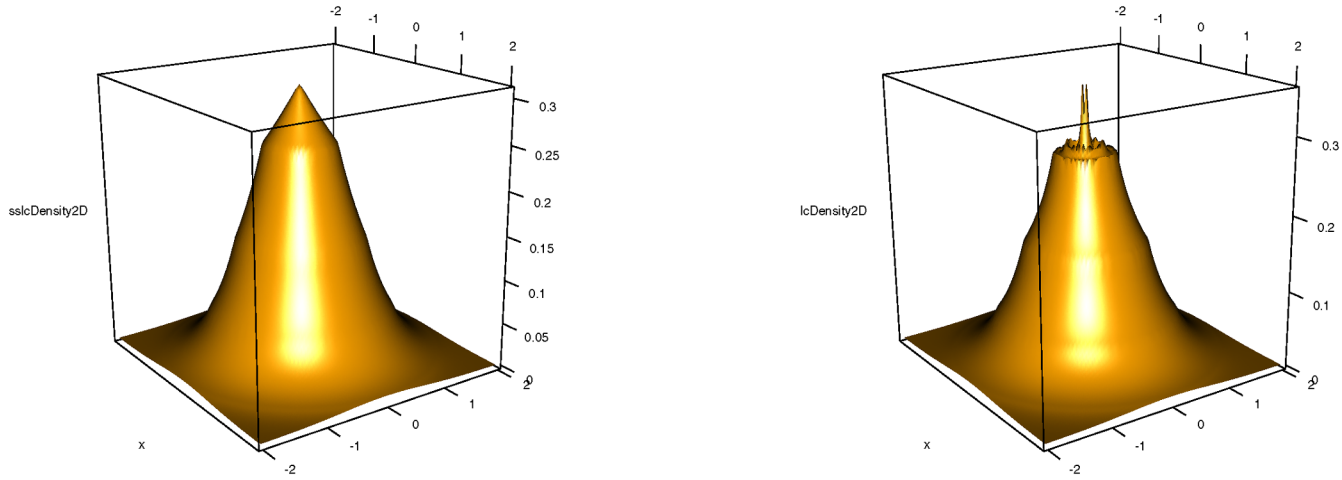

FIG. 8. *A comparison of the spherically-symmetric log-concave MLE $\hat{f}_n^{\mathrm{SS}}$ (left) and the estimator $\tilde{f}_n$ defined in (8.2) (right) based on a sample of size $n = 1000$ from a standard bivariate normal distribution.*

0 such that

$$\mathbb{E}d_X^2(\hat{f}_n^{\mathrm{SS}}, f_0) \le \min_{k=1,\ldots,n}\left(v_k^2 + C\frac{k}{n}\log^{5/4}\frac{en}{k}\right).$$

9. OTHER TOPICS

9.1 s-Concave Densities

As an attempt to allow heavier tails than are permitted by log-concavity, say a density f is *s-concave* with $s < 0$, and write $f \in \mathcal{F}_{d,s}$, if $f = (-\phi)^{1/s}$ for some $\phi \in \Phi$. Such densities have convex upper level sets, but allow polynomial tails, and satisfy $\mathcal{F}_{d,s_2} \supseteq \mathcal{F}_{d,s_1} \supseteq \mathcal{F}_d$ for $s_2 < s_1 < 0$. Some, but not all, of the properties of $\mathcal{F}_d$ translate over to these larger classes (e.g., Dharmadhikari and Joag-Dev, 1988). Results on the maximum likelihood estimator in the case $d = 1$ are recently available (Doss and Wellner, 2016b), but estimation techniques based on Rényi divergences are also attractive here (Koenker and Mizera, 2010, Han and Wellner, 2016a).

9.2 Finite Mixtures of Log-Concave Densities

Finite mixtures offer another attractive way of generalising the scope of log-concave modelling (Chang and Walther, 2007, Eilers and Borgdorff, 2007, Cule, Samworth and Stewart, 2010). The main issue concerns identifiability: for instance, the mixture distribution $pN_d(-\mu, I) + (1-p)N_d(\mu, I)$ with $p \in (0,1)$ has a log-concave density if and only if $\|\mu\| \le 1$ (Cule, Samworth and Stewart, 2010). However, all is not lost: for instance, consider distribution functions on $\mathbb{R}$ of the form

$$G(x) := pF(x-\mu_1) + (1-p)F(x-\mu_2),$$

where $p \in [0,1]$, $\mu_1 \le \mu_2$ and $F(-x) = 1 - F(x)$, so that the distribution corresponding to F is symmetric about zero. Hunter, Wang and Hettmansperger (2007) proved that if $p \notin \{0, 1/2, 1\}$ and $\mu_1 < \mu_2$, then p, μ_1, μ_2 and F are identifiable. Balabdaoui and Doss (2018) have recently exploited this result to fit a two-component location mixture of a symmetric, log-concave density. One can imagine this as a model for a population of adult human heights, where the two components correspond to men and women.

9.3 Log-Concave Probability Mass Functions

Let p denote a probability mass function supported on a subset $\mathcal{S}$ of the integers, so that $\mathcal{S} := \{z \in \mathbb{Z} : p(z) > 0\}$. Balabdaoui et al. (2013) define p to be log-concave if the following two conditions hold:

(a) If $z_1 < z_2 < z_3$ and $\min\{p(z_1), p(z_3)\} > 0$, then $p(z_2) > 0$;

(b) $p(z)^2 \ge p(z-1)p(z+1)$ for all $z \in \mathbb{Z}$.

Writing $\psi(z) := \log p(z)$ and defining the discrete Laplacian $(\Delta\psi)(z) := \psi(z+1) - 2\psi(z) + \psi(z-1)$, it can be seen that a probability mass function p supported on $\mathcal{S} \subseteq \mathbb{Z}$ is log-concave if and only if $\mathcal{S}$ is the intersection of an interval with $\mathbb{Z}$ and $(\Delta\psi)(z) \le 0$ for all $z \in \mathcal{S}$. Log-concave probability mass functions have many of the same properties as log-concave density functions, and Balabdaoui et al. (2013) show how much of the theory and methodology can be adapted to this setting.

9.4 Regression Problems

Consider the basic regression model

$$Y = m(x) + \epsilon,$$

where $x \in \mathbb{R}^d$ is considered fixed for simplicity, m belongs to a class of real-valued functions $\mathcal{M}$ and $\epsilon \sim P$ with $\mathbb{E}(\epsilon) = 0$. There is a large literature on estimating m under different shape constraints (e.g., van Eeden, 1958, Groeneboom, Jongbloed and Wellner, 2001b, Han and Wellner, 2016b, Chen and Samworth, 2016). But log-concavity does not seem to be a natural constraint to impose on a regression function. On the other hand, it may well represent a sensible model for the distribution of the error vector ϵ. Given covariates $x_1, \ldots, x_n \in \mathbb{R}^d$ and corresponding independent responses $Y_1, \ldots, Y_n$, Dümbgen, Samworth and Schuhmacher (2011, 2013) considered estimating $(m, \log\psi^*(P))$ by

$$(\hat{m}, \phi^*) \in \operatorname*{argmax}_{(m,\phi)\in\mathcal{M}\times\Phi} \frac{1}{n}\sum_{i=1}^n \phi(Y_i - m(x_i)) - \int_{\mathbb{R}^d} e^{\phi} + 1.$$

Such a maximiser exists, assuming only that $\mathcal{M}$ is closed under the addition of constant functions, and that $\mathcal{M}(x) := \{(m(x_1), \ldots, m(x_n)) : m \in \mathcal{M}\}$ is a closed subset of $\mathbb{R}^n$. Under a triangular array scheme, it can be shown that in the case of linear regression with a fixed number of covariates, the estimator of the vector of regression coefficients is consistent (Dümbgen, Samworth and Schuhmacher, 2013, Corollary 2.2), while numerical evidence suggests that the estimator can yield significant improvements over the ordinary least squares estimator in settings where ϵ has a log-concave, but not Gaussian, density. Similar to the Independent Component Analysis problem studied in

Section 8, the optimisation problem is again only biconcave, though stochastic search algorithms offer a promising approach (Dümbgen, Samworth and Schuhmacher, 2013).

ACKNOWLEDGEMENTS

Supported by an EPSRC Early Career Fellowship (EP/J017213/1 and EP/P031447/1), an EPSRC Programme grant (EP/N031938/1) and a grant from the Leverhulme Trust (RG81761).

REFERENCES

ALEKSANDROV, A. D. (1939). Almost everywhere existence of the second differential of a convex functions and related properties of convex surfaces. *Uchenye Zapisky Leningrad. Gos. Univ. Math. Ser.* **37** 3–35.

BALABDAOUI, F. and DOSS, C. R. (2018). Inference for a two-component mixture of symmetric distributions under log-concavity. *Bernoulli* **24** 1053–1071. MR3706787

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. MR2509075

BALABDAOUI, F., JANKOWSKI, H., RUFIBACH, K. and PAVLIDES, M. (2013). Asymptotics of the discrete log-concave maximum likelihood estimator and related applications. *J. R. Stat. Soc. Ser. B. Stat. Methodol.* **75** 769–790. MR3091658

BALÁZS, G., GYÖGY, A. and SZEPESVÁRI, C. (2015). Near-optimal max-affine estimators for convex regression. In *Proc. 18th International Conference on Artificial Intelligence and Statistics* (*AISTATS*) 56–64.

BARAUD, Y. and BIRGÉ, L. (2016). Rho-estimators for shape restricted density estimation. *Stochastic Process. Appl.* **126** 3888–3912. MR3565484

BIRGÉ, L. (1989). The Grenander estimator: A nonasymptotic approach. *Ann. Statist.* **17** 1532–1549. MR1026298

BRASS, P. (2005). On the size of higher-dimensional triangulations. In *Combinatorial and Computational Geometry*. *Math. Sci. Res. Inst. Publ.* **52** 147–153. Cambridge Univ. Press, Cambridge. MR2178319

CHANG, G. T. and WALTHER, G. (2007). Clustering with mixtures of log-concave distributions. *Comput. Statist. Data Anal.* **51** 6242–6251. MR2408591

CHEN, Y. and SAMWORTH, R. J. (2013). Smoothed log-concave maximum likelihood estimation with applications. *Statist. Sinica* **23** 1373–1398. MR3114718

CHEN, Y. and SAMWORTH, R. J. (2016). Generalized additive and index models with shape constraints. *J. R. Stat. Soc. Ser. B. Stat. Methodol.* **78** 729–754. MR3534348

CULE, M., GRAMACY, R. B. and SAMWORTH, R. (2009). LogConcDEAD: An R package for maximum likelihood estimation of a multivariate log-concave density. *J. Stat. Softw.* **29**.

CULE, M. and SAMWORTH, R. (2010). Theoretical properties of the log-concave maximum likelihood estimator of a multidimensional density. *Electron. J. Stat.* **4** 254–270. MR2645484

CULE, M., SAMWORTH, R. and STEWART, M. (2010). Maximum likelihood estimation of a multi-dimensional log-concave density. *J. R. Stat. Soc. Ser. B. Stat. Methodol.* **72** 545–607. MR2758237

DHARMADHIKARI, S. and JOAG-DEV, K. (1988). *Unimodality, Convexity, and Applications*. Academic Press, Boston, MA. MR0954608

DOSS, C. R. and WELLNER, J. A. (2016a). Inference for the mode of a log-concave density. https://arxiv.org/abs/1611.10348.

DOSS, C. R. and WELLNER, J. A. (2016b). Global rates of convergence of the MLEs of log-concave and s-concave densities. *Ann. Statist.* **44** 954–981. MR3485950

DUDLEY, R. M. (2002). *Real Analysis and Probability*. *Cambridge Studies in Advanced Mathematics* **74**. Cambridge Univ. Press, Cambridge. MR1932358

DÜMBGEN, L., HÜSLER, A. and RUFIBACH, K. (2007). Active set and EM algorithms for log-concave densities based on complete and censored data. Available at https://arxiv.org/abs/0707.4643v4.

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. MR2546798

DÜMBGEN, L. and RUFIBACH, K. (2011). logcondens: Computations related to univariate log-concave density estimation. *J. Stat. Softw.* **39** 1–28.

DÜMBGEN, L., RUFIBACH, K. and SCHUHMACHER (2013). logconcens: Maximum likelihood estimation of a log-concave density based on censored data. R package available at: https://cran.r-project.org/web/packages/logconcens/index.html.

DÜMBGEN, L., RUFIBACH, K. and SCHUHMACHER, D. (2014). Maximum-likelihood estimation of a log-concave density based on censored data. *Electron. J. Stat.* **8** 1405–1437. MR3263127

DÜMBGEN, L., SAMWORTH, R. and SCHUHMACHER, D. (2011). Approximation by log-concave distributions, with applications to regression. *Ann. Statist.* **39** 702–730. MR2816336

DÜMBGEN, L., SAMWORTH, R. J. and SCHUHMACHER, D. (2013). Stochastic search for semiparametric linear regression models. In *From Probability to Statistics and Back: High-Dimensional Models and Processes*. *Inst. Math. Stat. (IMS) Collect.* **9** 78–90. IMS, Beachwood, OH. MR3186750

EILERS, P. H. C. and BORGDORFF, M. W. (2007). Nonparametric log-concave mixtures. *Comput. Statist. Data Anal.* **51** 5444–5451. MR2370883

ERIKSSON, J. and KOIVUNEN, V. (2004). Identifiability, separability and uniqueness of linear ICA models. *IEEE Signal Process. Lett.* **11** 601–604.

GAO, F. and WELLNER, J. A. (2017). Entropy of convex functions on $\mathbb{R}^d$. *Constr. Approx.* **46** 565–592. MR3735701

GRENANDER, U. (1956). On the theory of mortality measurement. II. *Skand. Aktuarietidskr.* **39** 125–153. MR0093415

GROENEBOOM, P. (1985). Estimating a monotone density. In *Proceedings of the Berkeley Conference in Honor of Jerzy Neyman and Jack Kiefer, Vol. II* (*Berkeley, Calif.*, 1983). 539–555. Wadsworth, Belmont, CA. MR0822052

GROENEBOOM, P., JONGBLOED, G. and WELLNER, J. A. (2001a). A canonical process for estimation of convex functions: The "invelope" of integrated Brownian motion $+t^4$. *Ann. Statist.* **29** 1620–1652. MR1891741

GROENEBOOM, P., JONGBLOED, G. and WELLNER, J. A. (2001b). Estimation of a convex function: Characterizations and asymptotic theory. *Ann. Statist.* **29** 1653–1698. MR1891742

HAN, Q. and WELLNER, J. A. (2016a). Approximation and estimation of s-concave densities via Rényi divergences. *Ann. Statist.* **44** 1332–1359. MR3485962

HAN, Q. and WELLNER, J. A. (2016b). Multivariate convex regression: Global risk bounds and adaptation. Available at https://arxiv.org/abs/1601.06844.

HENNINGSSON, T. and ÅSTRÖM, K. J. (2006). Log-concave observers. In *Proc*. 17*th International Symposium on Mathematical Theory of Networks and Systems*.

HUNTER, D. R., WANG, S. and HETTMANSPERGER, T. P. (2007). Inference for mixtures of symmetric distributions. *Ann*. *Statist*. **35** 224–251. MR2332275

HYVÄRINEN, A., KARHUNEN, J. and OJA, E. (2001). *Independent Component Analysis*. Wiley, Hoboken, New Jersey.

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

KAPPEL, F. and KUNTSEVICH, A. V. (2000). An implementation of Shor's r-algorithm. *Comput*. *Optim*. *Appl*. **15** 193–205. MR1747059

KIM, A. K. H. and SAMWORTH, R. J. (2016). Global rates of convergence in log-concave density estimation. *Ann*. *Statist*. **44** 2756–2779. MR3576560

KIM, A. K. H., GUNTUBOYINA, A. and SAMWORTH, R. J. (2018). Adaptation in log-concave density estimation. *Ann*. *Statist*. To appear.

KOENKER, R. and MIZERA, I. (2010). Quasi-concave density estimation. *Ann*. *Statist*. **38** 2998–3027. MR2722462

MARSHALL, A. W. (1970). Discussion of Barlow and van Zwet's paper. In *Nonparametric Techniques in Statistical Inference*. *Proceedings of the First International Symposium on Nonparametric Techniques Held at Indiana University*, *June* 1969. Cambridge Univ. Press, London.

MÜLLER, S. and RUFIBACH, K. (2009). Smooth tail-index estimation. *J*. *Stat*. *Comput*. *Simul*. **79** 1155–1167. MR2572422

PRAKASA RAO, B. L. S. (1969). Estimation of a unimodal density. *Sankhyā Ser. A* **31** 23–36. MR0267677

PRÉKOPA, A. (1973). Contributions to the theory of stochastic programming. *Math*. *Program*. **4** 202–221. MR0376145

PRÉKOPA, A. (1980). Logarithmic concave measures and related topics. In *Stochastic Programming* (*Proc*. *Internat*. *Conf*., *Univ*. *Oxford*, *Oxford*, 1974) (M. A. H. Dempster ed.) 63–82. Academic Press, London. MR0592596

SAMWORTH, R. J. and YUAN, M. (2012). Independent component analysis via nonparametric maximum likelihood estimation. *Ann*. *Statist*. **40** 2973–3002. MR3097966

SAUMARD, A. and WELLNER, J. A. (2014). Log-concavity and strong log-concavity: A review. *Stat*. *Surv*. **8** 45–114. MR3290441

SCHUHMACHER, D., HÜSLER, A. and DÜMBGEN, L. (2011). Multivariate log-concave distributions as a nearly parametric model. *Stat*. *Risk Model*. **28** 277–295. MR2838319

SEREGIN, A. and WELLNER, J. A. (2010). Nonparametric estimation of multivariate convex-transformed densities. *Ann*. *Statist*. **38** 3751–3781. MR2766867

STRASSEN, V. (1965). The existence of probability measures with given marginals. *Ann*. *Math*. *Stat*. **36** 423–439. MR0177430

VAN EEDEN, C. (1958). *Testing and Estimating Ordered Parameters of Probability Distributions*. Mathematical Centre, Amsterdam. MR0102874

WALTHER, G. (2002). Detecting the presence of mixing with multiscale maximum likelihood. *J*. *Amer*. *Statist*. *Assoc*. **97** 508–513. MR1941467

WALTHER, G. (2009). Inference and modeling with log-concave distributions. *Statist*. *Sci*. **24** 319–327. MR2757433

XU, M. and SAMWORTH, R. J. (2017). High-dimensional nonparametric density estimation via symmetry and shape constraints. Working paper. Available at: http://www.statslab.cam.ac.uk/~rjs57/Research.html.