DOI: 10.1515/auom-2015-0053
An. Şt. Univ. Ovidius Constanţa
VERSITA
Vol. 23(3),2015, 151–166

A Computational Approach to Log-Concave Density Estimation

Fabian Rathke, Christoph Schnörr

Abstract

Non-parametric density estimation with shape restrictions has witnessed a great deal of attention recently. We consider the maximum-likelihood problem of estimating a log-concave density from a given finite set of empirical data and present a computational approach to the resulting optimization problem. Our approach targets the ability to trade-off computational costs against estimation accuracy in order to alleviate the curse of dimensionality of density estimation in higher dimensions.

1 Introduction

Density estimation constitutes a fundamental problem of statistics and machine learning with applications to clustering, classification and various further tasks of data analysis. Given a set of independent and identical distributed (i.i.d.) realizations

$$\mathcal{X}_n = \{x^1, \dots, x^n\} \subset \mathbb{R}^d, \qquad x^i \sim f_0 \tag{1}$$

generated from some unknown distribution with density f_0, the task is to obtain an estimate $\hat{f}$ of f_0 based on $\mathcal{X}_n$. Classical parametric methods are sensitive to model-misspecification. Non-parametric density estimation, on

Key Words: log-convave distributions, probability density estimation, mathematical programming, machine learning.
2010 Mathematics Subject Classification: Primary 90C90, 62G07; Secondary 65C60, 65K05, 90C25.
Received: December, 2014.
Revised: January, 2015.
Accepted: February, 2015.

the other hand, offers a flexible alternative to unbiased density estimation but requires regularization when working with *finite* data sets $\mathcal{X}_n$.

In this connection, the estimation of *log-concave* densities has recently attracted interest [1, 2, 3, 4]. This class is fairly rich as it includes many well-known unimodal parametric densities: all proper normal distributions, Wishart distributions, Gamma distributions with shape-parameter larger than one, Beta(α, β) distributions with $\alpha, \beta \geq 1$, and many more. Thus, this class of distributions constitutes an attractive class of non-parametric models whose flexibility is bounded by constraining the shape of corresponding densities to log-concave distributions, which is plausible for a broad range of applications. For a survey of various statistical aspects, we refer to [5]. Convexity properties related to log-concave distributions are worked out in [6] whereas the sampling problem is addressed in [7]. A major extension of theoretical results to a larger class of convexity-transformed densities has been established by [8].

In this paper, we focus on *computational* aspects of the estimation of log-concave density estimates $\hat{f} \approx f_0$ of the form

$$\hat{f}(x) = e^{-\hat{g}(x)}, \qquad \int_{\mathbb{R}^d} \hat{f}(x)dx = 1, \qquad \hat{g} \text{ is convex.} \tag{2}$$

The objective function is given by the *maximum-likelihood problem* in terms of minimizing the negative log-likelihood

$$-\frac{1}{n} \log \prod_{i=1}^{n} f(x^i) = \frac{1}{n} \sum_{i=1}^{n} g(x^i), \tag{3}$$

that is, $\hat{f}$ will be given by $\hat{g}$ solving

$$\min_g \frac{1}{n} \sum_{i=1}^{n} g(x^i) \quad \text{subject to} \quad \int_{\mathbb{R}^d} e^{-g(x)}dx = 1, \quad g \text{ is convex.} \tag{4}$$

The constraint can be taken into account [9, Thm. 3.1] by considering instead the optimization problem

$$\min_g \frac{1}{n} \sum_{i=1}^{n} g(x^i) + \int_{\mathbb{R}^d} e^{-g(x)}dx \quad \text{subject to} \quad g \text{ is convex.} \tag{5}$$

In order to solve this problem computationally, a finite representation of g and the constraints has to be adopted. Our approach is based on the Legendre-Fenchel transform [10, p. 473] that enables to represent any proper, convex and lower-semicontinuous function as supremum of affine functions

$$g(x) = \sup_{y \in \operatorname{dom} g^*} \big\{ \langle x, y \rangle - g^*(y) \big\}, \tag{6}$$

where the function g^* denotes the convex conjugate of g. A natural finite representation then is obtained by restriction to a finite set of affine functions

$$g(x) \approx g_n(x) := \max_{i=1,\dots,K} \big\{ \langle x, y^i \rangle - g^*(y^i) \big\}. \tag{7}$$

Clearly, because g is unknown g^* is unknown as well. Hence we consider the parametrization

$$g_n(x;\beta) := \max_{i=1,\dots,K} \big\{ \langle x, a^i \rangle - \alpha_i \big\}, \quad \beta := \{(a^1,\alpha_1),\dots,(a^K,\alpha_K)\}. \tag{8}$$

In view of (5), we thus arrive at the optimization problem

$$\hat{\beta} = \arg\min_{\beta} J(\beta), \tag{9a}$$

$$J(\beta) = \frac{1}{n} \sum_{i=1}^{n} g_n(x^i;\beta) + \int_{\mathbb{R}^d} e^{-g_n(x;\beta)} dx. \tag{9b}$$

The log-concave density estimate is then given by

$$\hat{f}_n(x) = \exp(-g_n(x;\hat{\beta})). \tag{10}$$

Our main motivation is the well-known *curse of dimensionality* in connection with density estimation. Ansatz (8) enables to control the number of variables, to exploit sparsity and to trade-off estimation accuracy against computational costs. The latter becomes a serious issues as the dimension d increases. The price to pay is the non-convexity of the optimization problem (9). Our preliminary experimental results demonstrate, however, that a suitable smoothing strategy alleviates this issue and leads to an efficient estimation algorithm that outperforms a recently established state-of-the-art method without significantly compromising estimation accuracy.

Our paper is organised as follows. We summarise relevant results from the literature concerning log-concave density estimation in Section 2. Our computational approach is presented and worked out in Section 3. Preliminary numerical results are discussed in Section 4. We conclude in Section 5.

2 Log-Concave Density Estimation: Related Work

We briefly report available results concerning the maximum-likelihood estimation problem and related work.

Koenker and Mizera [4] adopt the representation

$$g_n(x) := \inf \left\{ \sum_{i=1}^{n} \lambda_i y_i \colon x = \sum_{i=1}^{n} \lambda_i x^i,\ \sum_{i=1}^{n} \lambda_i = 1,\ \lambda_i \geq 0 \right\} \tag{11}$$

of the convex function g, based on the given observations $\mathcal{X}_n$ and the corresponding function values $y_i = g_n(x^i)$, $i = 1, \dots, n$. This choice is motivated by the convexification of an arbitrary function $g\colon \mathbb{R}^d \to [-\infty, +\infty]$ given by

$$(\operatorname{conv} g)(x) = \inf \left\{ \sum_{i=0}^{d} \lambda_i g(x^i) \colon x = \sum_{i=0}^{d} \lambda_i x^i,\ \sum_{i=0}^{d} \lambda_i = 1,\ \lambda_i \geq 0 \right\}, \tag{12}$$

which is the greatest convex function majorized by g (cf. [10, Prop. 2.31]). While in (12) $(d+1)$ points have to be chosen, dependent on x, according to Carathéodory's theorem, *n fixed* observed points $\mathcal{X}_n$ are used in (11) and thus lead to a *finite*-dimensional representation. Authors show, in fact, that solutions to the maximum-likelihood problem based on a finite data set $\mathcal{X}_n$ take the form (11).

Insertion into (5) results in a convex optimization problem with respect to the function values $g_n(x^i)$, $x^i \in \mathcal{X}_n$. In order to approximate the second nonlinear term of (9b) sufficiently accurate by numerical integration, authors work with a regular grid of appropriate cell-size and interpolated functions values at grid vertices. Moreover, the convexity of g_n has to be enforced by local convex constraints in terms of these function values, which amounts to an inequality system if the dimension $d = 1$, to second-order cone constraints if $d = 2$, and to semidefinite constraints if $d > 2$. As a consequence, the problem size quickly becomes computationally intractable in the latter cases.

Cule et al. [2] further exploit the fact that epigraphs

$$\operatorname{epi} g_n = \left\{ (x, \alpha) \in \mathbb{R}^d \times R \colon \alpha \geq g_n(x) \right\} \tag{13}$$

of functions g_n of the form (11) are polyhedral, which due to [10, Prop. 2.31] means that g_n is a *convex piecewise affine* function: The support of the corresponding density $f_n = e^{-g_n}$

$$\operatorname{supp} f_n = \mathcal{C}_n := \operatorname{conv} \mathcal{X}_n = \bigcup_{i=1}^{K} \mathcal{C}_i, \tag{14}$$

which equals the convex hull of the given data $\mathcal{X}_n$, can be represented as union of finitely many polyhedral sets $\mathcal{C}_i$, relative to each of which g_n is affine,

$$g_n(x)\big|_{\mathcal{C}_i} = \langle a^i, x \rangle + \alpha_i, \qquad i = 1, \dots, K. \tag{15}$$

Accordingly, authors of [2] suggest to triangulate $\mathcal{C}_n$ in the case $d = 2$ and the simplicial decomposition in the case $d > 2$, respectively, such that each $\mathcal{C}_i$ is given as convex hull of $d + 1$ points of $\mathcal{X}_n$.

While this suggests the expedient application of highly accurate methods of numerical integration for computing the second term of (5), a major drawback is the non-smoothness of the resulting convex objective function. Accordingly, a numerical optimization strategy based on subgradients is employed in [2] which are known to converge slowly. Our experimental results reported below confirm this fact. Furthermore, the curse of dimensionality obstacle persists: For large n which is desirable for density estimation, and particularly so in higher dimensions $d > 2$, the approach becomes computationally expensive.

Motivation and goal behind our approach to be presented in the subsequent section is

(i) to achieve a problem representation that can be controlled independently of the size n of the data set $\mathcal{X}_n$ and the dimension d, and

(ii) to approximate the objective function for computing an estimate $\hat{f}_n$, so that efficient numerical methods can be applied that scale up to large problem sizes n and to higher dimensions d.

We refer to Figure 1 for a first illustration of our approach in comparison to the state-of-the-art approach of Cule et al. [2].

Clearly, objective (i) is questionable from the viewpoint of *consistency*, that is convergence of the density estimate $\hat{f}_n$ to the true unknown underlying log-concave density f_0 as $n \to \infty$. Yet, we consider this aspect as less important in view of practical scenarios where n will be finite and often relatively small due to application-specific restrictions.

3 Approximate Density Estimation

3.1 Objective Function

The objective function (9b) explicitly reads with (8)

$$J(\beta) = \frac{1}{n} \sum_{i=1}^{n} \max_{1 \le k \le K} \left\{ \langle a^k, x^i \rangle - \alpha_k \right\} + \int_{\mathbb{R}^d} e^{-\max_{1 \le k \le K} \{\langle a^k, x^i \rangle - \alpha_k\}} dx. \tag{16}$$

While the first term is convex, the latter is not. Moreover, the functional is non-smooth. We remove the latter property and alleviate the former one by

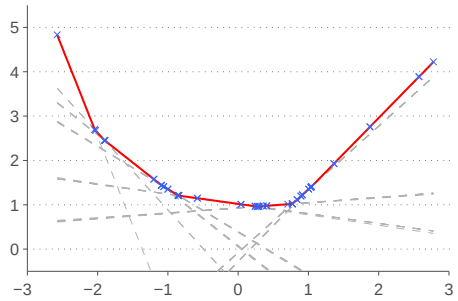


Figure 1: The solution $\hat{g}_n$ returned by the **R** package of Cule et al. [11] for $n = 250$ samples drawn from a standard normal distribution. Crosses mark those samples $x^i \in \mathcal{X}_n$ that span the resulting 35 sets $\mathcal{C}_i$ of (14), and dashed lines indicate the corresponding affine functions (15). A small subset only is relevant for representing $\hat{g}_n$ sufficiently accurate, however. Fig. 4 (a) accordingly shows the density estimate in terms of a sparse approximation of $\hat{g}_n$ obtained by our approach, only using 5 automatically determined affine functions.

defining the one-parameter family of objective functions

$$J_\gamma(\beta) := \frac{1}{n}\sum_{i=1}^{n} g_{\gamma,n}(x;\beta) + \int_{\mathbb{R}^d} e^{-g_{\gamma,n}(x;\beta)}dx, \qquad \gamma > 0 \tag{17a}$$

$$g_{\gamma,n}(x;\beta) := \frac{1}{\gamma}\operatorname{logexp}\big(\gamma h_n(x;\beta)\big), \tag{17b}$$

$$h_n(x;\beta) := \big(\langle a^1, x\rangle + \alpha_1, \ldots, \langle a^K, x\rangle + \alpha_K\big)^\top, \tag{17c}$$

$$\operatorname{logexp}(y) := \log\Big(\sum_{k=1}^{K}\exp(y_k)\Big), \quad y \in \mathbb{R}^K. \tag{17d}$$

The rational behind this is the following uniform approximation property.

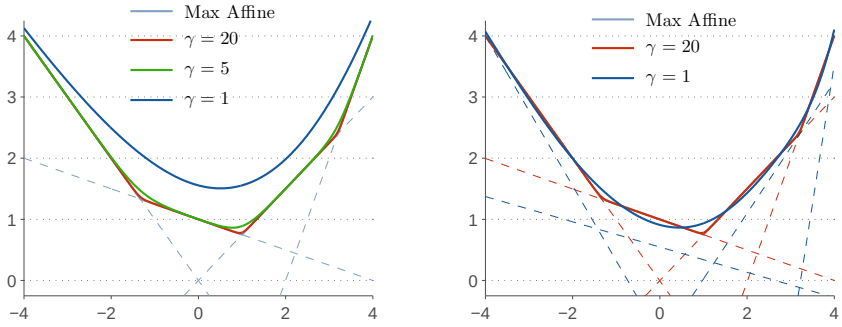
Lemma 3.1 ([10, Example 1.30])**.** *For any $y \in \mathbb{R}^K$ and $\gamma > 0$, we have*

$$\frac{1}{\gamma}\big(\operatorname{logexp}(\gamma\, y) - \log K\big) \le \max_{1\le k\le K} y_k \le \frac{1}{\gamma}\operatorname{logexp}(\gamma\, y). \tag{18}$$

As a consequence,

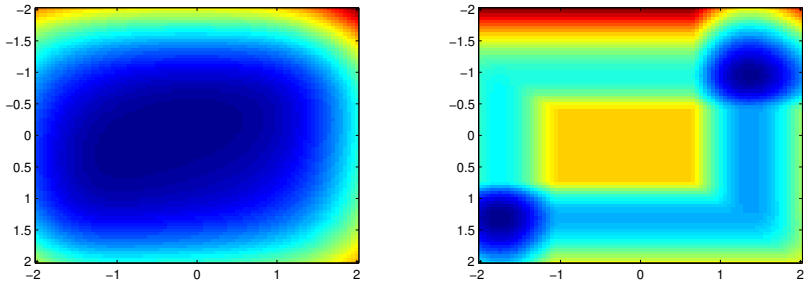
$$g_{\gamma,n}(x;\beta) \;\to\; \max_{1\le k\le K} h_n(x;\beta) \quad \text{as} \quad \gamma \to +\infty \tag{19}$$

and $J_\gamma(\beta) \to J(\beta)$ given by (16). Figures 2 and 3 provide illustrations.



(a) (b)

Figure 2: (a) Approximation of $\max_{1\leq k\leq K} h_n(x,\beta)$ by $g_{\gamma;n}(x;\beta)$ due to (19) for $K = 4$, using the same β. The affine functions comprising $h_n(x;\beta)$ are shown as dashed lines. (b) Approximation by $g_{\gamma;n}(x;\tilde{\beta})$ with $\tilde{\beta}$ obtained as least-squares estimate. This demonstrates that even smooth approximations, corresponding e.g. to $\gamma = 1$, yield accurate approximations in the L^2-sense.



(a) $J(\beta)$ for $\gamma = 1$ (b) $J(\beta)$ for $\gamma = 50$

Figure 3: Objective function $J_\gamma(\beta)$ given by (17a) and $\beta = (a^k, \alpha_k)^\top \in \mathbb{R}^2$ randomly drawn from a uniform distribution with support $(-1,+1)\times(-1,+1)$. Both plots display $J_\gamma(\beta)$ as a function of two slopes a^i and a^j. Increasing γ decreases smoothness and convexity of J_γ, and increases the number of local minima.

3.2 Optimization

We numerically minimize the following approximation of the objective function (17a)

$$J_{\gamma,\delta}(\beta) := \frac{1}{n}\sum_{i=1}^{n} g_{\gamma,n}(x;\beta) + \delta_L \sum_{l=1}^{L} e^{-g_{\gamma,n}(z^l;\beta)}, \tag{20}$$

where the second term simply approximates the integral of (17a) by a step function on a sufficiently fine grid covering the support $\operatorname{conv} \mathcal{X}_n$ of the density to estimated, with vertices z^l, $l = 1, \ldots, L$, and volume δ_L of the corresponding cells centered at the points z^l.

Algorithm. We apply a modified iterative Newton scheme [12, Sec. 9.5] in order to minimize (20):

$$\beta^{k+1} = \beta^k + t_k \Delta\beta^k, \tag{21a}$$

$$\Delta\beta^k = -\big(\nabla^2 J_{\gamma,\delta}(\beta^k) + \eta_k I\big)^{-1} \nabla J_{\gamma,\delta}(\beta^k), \quad k = 0, 1, \ldots \tag{21b}$$

where $\eta_k \geq 0$ is chosen so that the matrix $\nabla^2 J_{\gamma,\delta}(\beta^k) + \eta_k I$ is positive definite, and the step-size t_k is determined by backtracking line-search [12, p. 464]. We set $\eta_k = -\lambda_{\min} + 10^{-3}$, with $\lambda_{\min} < 0$ being the smallest eigenvalue of $\nabla^2 J_{\gamma,\delta}(\beta^k)$. We calculate $\lambda_{\min}$ explicitly, which is inexpensive compared to the evaluations of various terms of (21) at all grid points z^l.

We terminate the iteration at step k if the following two conditions are satisfied.

(a) $\big|1 - \delta_L \sum_{l=1}^{L} e^{-g_{\gamma,n}(z^l;\beta^k)}\big| \leq 10^{-4}$ and

(b) $\frac{1}{2}\big\langle \Delta\beta^k, \nabla J(\beta^k)\big\rangle =: \frac{1}{2}\lambda(\beta^k)^2 \leq 10^{-5}$.

Condition (a) ensures that the estimated density almost integrates to 1. The quantity $\frac{1}{2}\lambda(\beta^k)^2$ of condition (b) upper bounds the gap $J_{\gamma,\delta}(\beta^k) - J_{\gamma,\delta}(\hat{\beta})$, where $\hat{\beta}$ is a local minimum of $J_{\gamma,\delta}(\beta)$ [12, p. 487].

Initialization. We adopt the following strategy for determining an initialisation β^0 of the iteration (21). Choosing $\gamma = 1$ which yields a smooth objective function $J_{\gamma,\delta}$, we fit $10 \cdot d$ affine functions with parameters randomly initialized by sampling the uniform distribution supported on $(-1, +1)$. Then β^0 is found by fitting K affine functions to $g_{1,n}(x;\beta)$ at points that are determined by k-means clustering of $\mathcal{X}_n$.

We demonstrate in Section 4 that this strategy effectively removes the sensitivity of density estimates with respect to the initialization β^0.

Speed-Up Heuristic. In order to accelerate the computations we keep track of which affine functions forming the components of h_n in (17) do not

contribute to the second sum of (20). The corresponding "inactive" parameters are successively removed and ignored during the remaining iterative steps. Our experiments demonstrate that this yields a compact parametrisation without compromising estimation accuracy.

4 Experimental Results

4.1 Set-Up and Evaluation Measures

This section provides an assessment of our approach by numerical experiments. Specifically, we examine

- the influence of the smoothing parameter γ,
- the size K of active affine functions, both when the iteration started and after convergence to a local optimum,
- the effectiveness of the initialization procedure,
- runtime depending on the size n of the given dataset $\mathcal{X}_n$.

We compare our results with the approach of Cule et al. [2] sketched in Section 2, based on the independent implementation provided by a corresponding **R** package `LogConcDead` [11].

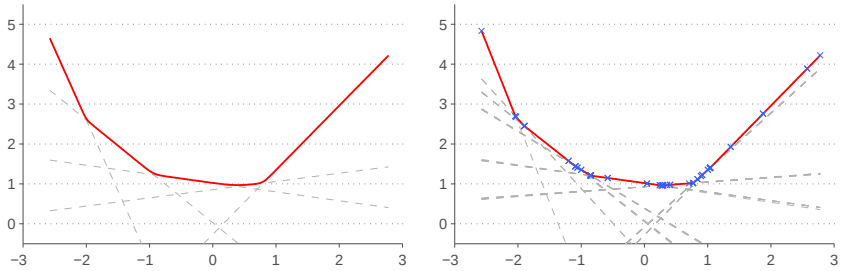
Our own implementation was done using MATLAB (non-tuned research code). All datasets are points samples from the standard normal distribution $\mathcal{N}(0, I_d)$ for $d = 1$ (next section) and $d = 2$ (all remaining sections).

4.2 1-D Toy Example

Revisiting the example from Fig. 1, Fig. 4 demonstrates that our approach returns a density estimate that is very close to the estimate returned by the approach of Cule et al, but with a sparse representation. Specifically, in this example, using $\gamma = 20$ and $K = 20$ initial affine functions, the final representation comprises 5 affine functions whereas the estimate of Cule et al. needs about 7 times more variables. In general, our experiments show that our approach reliably returns a sparse representation whenever a density estimate admits one.

4.3 Influence of γ, K and the Initialization

We generated 20 different 2-D data sets $\{\mathcal{X}_n^i\}_{i=1}^{20}$ with $n = 250$ samples each. We estimated the corresponding densities for each combination of $K = \{10, 20, 50, 100, 200\}$ affine functions and values of the smoothing parameter



(a) Our estimate, $\gamma = 20, K = 20$ (5 after termination).

(b) Cule et al., $K = 35$.

Figure 4: (a) The estimate of our approach for the introductory example of Fig. 1. The resulting density is very close to the optimal estimate of Cule et al. (b) but only comprises five affine functions. In general, if a density estimate admits a sparse polyhedral representation, then our approach determines a representation that is sparse.

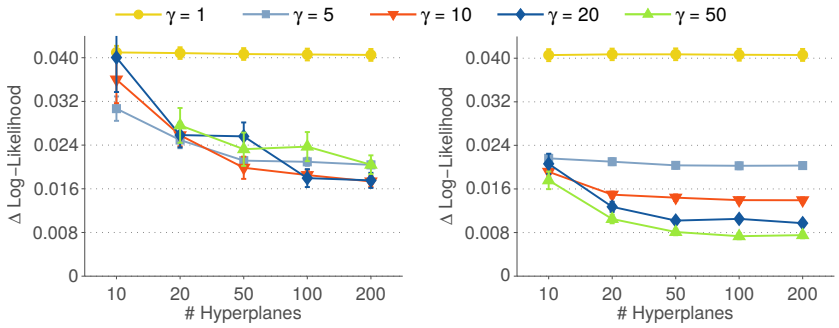
$\gamma = \{1, 5, 10, 20, 50\}$. Furthermore, we compared the results obtained with the initialization procedure described in Section 3.2 with the results based on an entirely random initialization.

Let $\hat{f}^i_{\gamma,n}$ denote the density estimate returned by our approach for each sample set $\mathcal{X}^i_n$, and let $\hat{f}^i_{C,n}$ denote the corresponding estimate obtained using the approach of Cule et al. Figure 5 reports for each pair of values K, γ the empirical mean along with the standard error (standard deviation divided by $\sqrt{20}$) as error bar of the sequence

$$\frac{1}{n} \sum_{x^j \in \mathcal{X}^i_n} \left| \log \hat{f}^i_{\gamma,n}(x^j) - \log \hat{f}^i_{C,n}(x^j) \right|, \qquad i = 1, \ldots, 20. \tag{22}$$

Generally it holds that increasing values of γ and K improves the quality of the approximation, as to be expected. Convergence of the green curve ($\gamma = 50$) towards $0.008 \approx 0$ in the right panel, in particular, shows that both approaches return virtually the same estimate in the sense that the empirical average $(\hat{f}_{\gamma,n}/\hat{f}_{C,n})(x) \to 1$. Comparing both types of initialization, especially estimates with large values of γ benefit from our two-stage initialization procedure, as discussed in the previous section (cf. also Fig. 3).

Additional experiments with bigger data sets of up to 10^4 samples showed that increasing the initial number of affine functions beyond 200 does not improve results any more. Therefore, $K = 200$ seems to be a reasonable default setting if $d \in \{1, 2\}$. We further observed that a maximal value $\gamma = 20$



(a) Random initialization (b) Non-random initialization

Figure 5: The influence of parameters γ and K on the quality of the estimated density, measured as difference to the result of [2] by averaging the sequence (22). Error bars indicate the corresponding standard error. Randomly initializing β leads to suboptimal results for large γ, in comparison to our two-stage initialisation strategy. Convergence of the green curve ($\gamma = 50$) in the right panel towards 0 shows that our approach essentially returns the same density estimates as the approach of Cule et al., in the sense that the empirical average $(\hat{f}_{\gamma,n}/\hat{f}_{C,n})(x) \to 1$.

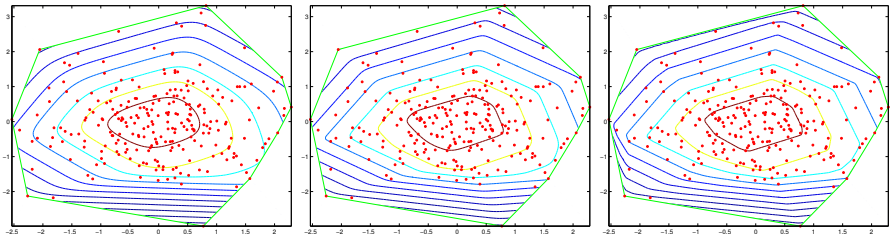
suffices for highly accurate approximation, since larger values of γ do not yield different estimates but may cause numerical problems (floating point arithmetic).

Fig. 6 shows two densities estimated by our approach for $K = 200$ and $\gamma = 5$ (a) and $\gamma = 20$ (b) with the estimate from Cule et al. (c).

We also measured the absolute estimation accuracy in terms of the Hellinger distance to ground truth. These quantitative results, summarized and discussed in the caption of Figure 7, illustrate the following findings: The estimates of our approach are as accurate as those returned by the approach of Cule et al. The dependency on the smoothing parameter γ is insignificant. We also observed superior estimation accuracy when using a strongly smooth objective function $J_{\gamma,n}$ with $\gamma = 1$, which is plausible in view of the smoothness of the ground truth density f_0 (standard Normal distribution). The approach of Cule et al. cannot exploit this prior knowledge if it were available.

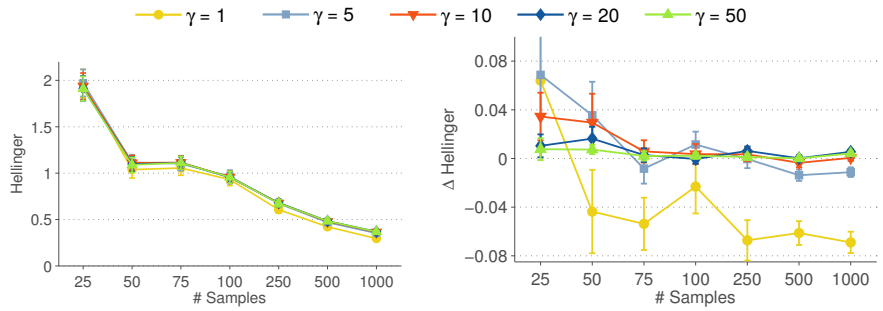
4.4 Sample Size vs. Runtime

As discussed in Section 2, the approach of Cule et al. [2] essentially depends on the size n. The simplicial decomposition defining the partition (14) and the representation (15) has to be performed at *each* iteration [2, Appx. B] because



(a) $\gamma = 5, K = 200$ (121) (b) $\gamma = 20, K = 200$ (26) (c) Cule et al., $K = 383$

Figure 6: Density estimates $\hat{f}_{\gamma,n} = f_{\gamma,n}(x; \hat{\beta})$ in (a) and (b) and the estimate of Cule et al. in (c). The final number of affine functions for our approach is given in parentheses. While the estimate for $\gamma = 20$ only comprises 26 affine functions, the result of Cule et al. requires 383 affine functions. The estimates (b) and (c) are very similar, however.



(a) Hellinger distance $H(\hat{f}_{\gamma,n}, f_0)$ of our density estimate to ground truth.

(b) Difference of Hellinger distances $H(\hat{f}_{\gamma,n}, f_0) - H(\hat{f}_{C,n}, f_0)$ of estimates (our approach, Cule et al.) to ground truth.

Figure 7: Distance of density estimates, for various values of the smoothing parameter γ, to the underlying log-concave density $f_0 = \mathcal{N}(0, I_d)$. Panel (a) shows that our approach consistently approaches ground truth with increasing sample size n. Because the approach determines the required number K of affine functions, estimation accuracy does virtually not depend on the smoothing parameter γ – cf. Fig. 2, right panel, for an "empirical explanation" of this fact. The negative values in the right panel (b) for the case $\gamma = 1$ (strong smoothing) signal that in this case our density estimate is even more accurate than the estimate returned by the approach of Cule et al. Otherwise, the curves approach 0, which demonstrates that our efficient approach (runtime, parametrization) does not compromise estimation accuracy.

Table 1: The runtime (in seconds) (R), number of iterations (I) and number of affine functions (K) used to model g_n, depending on n. Our approach efficiently determines a sparse representation of the density estimate without essentially compromising estimation accuracy. This significantly contrasts with the approach of Cule et al. that has quadratic runtime complexity and a less compact representation, which becomes computationally infeasible for large data sets in higher dimensions d.

	n	100	250	500	1000	2500	5000	10000
Cule et al. [2, 11]	R	0.6	3.2	13.1	59.5	610.6	3073.2	16653.7
	I	255	643	1266	2602	7211	10766	14973
	K	172	366	674	1054	2183	5445	9006
Our approach	R	3.0	9.6	8.9	9.4	8.3	9.6	13.7
	I	26	45	34	53	34	44	41
	K	13	20	30	35	40	44	54

the values $g_n(x^i)$ change during numerical optimization.

As a result, each iteration has a worst-case execution time of $\mathcal{O}(n \log n + n^{\lfloor d/2 \rfloor})$, which for $d = 2$ becomes $\mathcal{O}(n \log n)$. Furthermore, authors of [2] report that the number of iterations grows linearly with n, a fact that our experiments confirmed. Thus, the total dependency on n is $\mathcal{O}(n^2 \log n)$. In contrast, regarding our approach, the number of terms of the first summand of (16) only linearly grows with n. Furthermore, we observed that the number of iterations of the minimization algorithm for determining $\hat{\beta}$ is independent of n, thus resulting in an overall linear complexity $\mathcal{O}(n)$.

To examine experimentally the impact of n on the running time and on the number of iterations, we sampled five data sets of sizes $n = \{100, 250, 500, 1000, 2500, 5000, 10000\}$. For our approach we used parameters $K = \min\{n, 200\}$ and $\gamma = 20$, that is a sufficiently rich (number K of affine functions) and non-smooth accurate representation (large value of γ) of $g_{\gamma,n}(x; \beta)$ (cf. Lemma 3.1 and (19)). The average results are collected as Table 1: runtime in seconds (R), number of iterations (I) until convergence of numerical optimisation and number variables in terms of affine functions (K).

The numbers of Table 1 reveal, for the approach of Cule et al. [2], the expected quadratic runtime dependency on n as well as the linear increase in the number of iterations. For our approach, on the other hand, the number of iterations remained largely constant. Furthermore, the runtime is significantly smaller and does not essentially differ for the smallest and largest data sets. Overall, our approach is more efficient and more compactly parametrised than the approach of Cule et al. without compromising estimation accuracy. The

runtime of our approach for data sizes $n > 100$ is dominated by the numerical integration, that is by the terms of (21) corresponding to the second term of (20). Such evaluations on a regular lattice can be easily parallelized, however.

5 Conclusion

We presented a novel approach to the estimation of a log-concave density. It features a *sparse* approximation of the max-affine function underlying the optimal log-concave density estimate $\hat{f}_n(x)$ (10). We presented an optimization scheme based on the smooth approximation (17b), a numerical evaluation of the integral term in (17a) and a Newton-based line search with modified Hessian for the non-convex objective function (20). We demonstrated that this approach yields densities with almost minimal log-likelihood, while significantly reducing the runtime for medium-sized and large sample sets in comparison to the current state-of-the-art approach by Cule at al. [2].

Future Work. For dimensions $d \geq 4$, the evaluation of the second term of (20) on a fine regular grid is too expensive. *Adaptive multiscale grids* are a natural solution to this issue, using a coarser resolution in regions of low density where a finely spaced grid does not reduce the approximation error. The speed-up heuristic described in Section 3.2 indicates how grid adaption (or local scale selection) may be incorporated into the optimization procedure. Doing this rigorously along with a proof that the iteration terminates, requires more work.

Another point concerns the transition from the estimate obtained with a smooth objective function $\gamma = 1$ to the estimate computed with a larger value of γ, that is a refinement of the two-stage initialization procedure described in Section 3.2. Rather than directly "jumping" from $\gamma = 1$ to, say, $\gamma = 20$, a numerical continuation method instead is conceivable, based on the smooth dependency of our estimates $\hat{f}_{\gamma,n}$ on γ. We also point out that choosing a too large value of γ seems suboptimal *if* the underlying unknown density is smooth, as our experiments summarized by Fig. 7 indicate. Relating optimal values of γ to such smoothness assumptions (viz. prior knowledge) defines another open point of research.

Finally, applications to real problems should be mentioned. An attractive example is provided by the probabilistic shape prior introduced in [13], in terms of a density supported on a convex cone that models ordering constraints. Estimation of such densities from examples is naturally supported by our approach presented here, due to the convexity property (14) of corresponding supports. Yet, making the approach practical for higher dimensions $d = 3, 4, \ldots, 9$ defines an ambitious research task.

Acknowledgements

This work has been supported by the German Research Foundation (DFG), grant GRK 1653, as part of the research training group on "Probabilistic Graphical Models and Applications in Image Analysis".*

References

[1] Y. Chen and R. J. Samworth, "Smoothed log-concave maximum likelihood estimation with applications," *Statist. Sinica*, vol. 23, 2013.

[2] M. Cule, R. Samworth, and M. Stewart, "Maximum likelihood estimation of a multi-dimensional log-concave density," *J. R. Stat. Soc. Series B Stat. Methodol.*, vol. 72, no. 5, pp. 545–607, 2010.

[3] L. Dümbgen and K. Rufibach, "Maximum likelihood estimation of a log-concave density and its distribution function: Basic properties and uniform consistency," *Bernoulli*, vol. 15, no. 1, pp. 40–68, 2009.

[4] R. Koenker and I. Mizera, "Quasi-concave density estimation," *Ann. Stat.*, vol. 38, no. 5, pp. 2998–3027, 2010.

[5] G. Walther, "Inference and modeling with log-concave distributions," *Statist. Sci.*, vol. 24, no. 3, pp. 319–327, 2009.

[6] B. Klartag and V. Milman, "Geometry of log-concave functions and measures," *Geom. Dedicata*, vol. 112, no. 1, pp. 169–182, 2005.

[7] L. Lovász and S. Vempala, "The geometry of log-concave functions and sampling algorithms," *Rand. Structures Alg.*, vol. 30, no. 3, pp. 307–358, 2007.

[8] A. Seregin and J. Wellner, "Nonparametric estimation of multivariate convex-transformed densities," *Ann. Statistics*, vol. 38, no. 6, pp. 3751–3781, 2010.

[9] B. Silverman, "On the estimation of a probability density function by the maximum penalized likelihood method," *Ann. Stat.*, vol. 10, no. 3, pp. 795–810, 1982.

[10] R. Rockafellar and R.-B. Wets, *Variational Analysis*, vol. 317. Springer, 3rd ed., 2009.

*http://graphmod.iwr.uni-heidelberg.de

[11] M. Cule, R. Gramacy, and R. Samworth, "LogConcDEAD: An R package for maximum likelihood estimation of a multivariate log-concave density," *J. Stat. Softw.*, vol. 29, no. 2, pp. 1–20, 2009.

[12] S. Boyd and L. Vandenberghe, *Convex Optimization*. Cambridge University Press, 2004.

[13] F. Rathke, S. Schmidt, and C. Schnörr, "Probabilistic intra-retinal layer segmentation in 3-D OCT images using global shape regularization," *Med. Image Anal.*, vol. 18, no. 5, pp. 781–794, 2014.

Fabian Rathke,
Image and Pattern Analysis Group,
University of Heidelberg,
Speyerer Str. 6, 69126 Heidelberg, Germany.
Email: fabian.rathke@iwr.uni-heidelberg.de

Christoph Schnörr,
Image and Pattern Analysis Group,
University of Heidelberg,
Speyerer Str. 6, 69126 Heidelberg, Germany.
Email: schnoerr@math.uni-heidelberg.de