Statistica Sinica **28** (2018), 2697-2712
doi:https://doi.org/10.5705/ss.202017.0344

ESTIMATING A DISCRETE LOG-CONCAVE DISTRIBUTION IN HIGHER DIMENSIONS

Hanna Jankowski and Yan Hua Tian

York University

Abstract: We define a new class of log-concave distributions on the discrete lattice $\mathbb{Z}^d$, and study its properties. We show how to compute the maximum likelihood estimator of this class of probability mass functions from an independent and identically distributed sample, and establish consistency of the estimator, even if the class has been incorrectly specified. For finite sample sizes, in our simulations, the proposed estimator outperforms a purely nonparametric approach (the empirical distribution), but is able to remain comparable to the correct parametric approach. Notably, the new class of distributions has a natural relationship with log-concave densities.

Key words and phrases: Log-concave, maximum likelihood estimation, multivariate data, probability mass function estimation, shape-constrained methods.

1. Introduction

1.1. For Peter

This work is part of the field of shape-constrained methods. Although not Peter's primary area of research, he did consider certain variations and related problems, including Braun and Hall (2001); Hall and Heckman (2000); Hall and Presnell (1999); see also his comments in Cule, Samworth and Stewart (2010, p.586).

The first author of this paper (HJ) got to know Peter during her sabbatical at the University of Melbourne in 2013–2014. Some of the fondest and most heart-warming memories from the visit took place over our many group lunches, where discussions varied from how best to teach mathematical statistics to a story about a cat from the UK who regularly takes the bus on his own. Peter was always kind and generous, and is well-known to have been exceptionally supportive of young researchers. When HJ was invited to contribute to a special issue dedicated in Peter's name, she felt this work was appropriate, as it is joint work with HJ's first doctoral student.

1.2. Estimation of log-concave distributions

In recent years, there has been much interest in log-concave probability density function estimation, see, for example, the review article of Walther (2009). As noted by Walther (2002), log-concave densities provide a natural alternative to the class of unimodal densities, while not being too restrictive by specifying a parametric family. One can also view the class of log-concave densities as striking a balance between too large a class (and hence inefficient) and too small a class (and hence not robust). Maximum likelihood estimation for the $d = 1$ setting was studied in Dümbgen and Rufibach (2011); Balabdaoui, Rufibach and Wellner (2009); Doss and Wellner (2016), while the $d > 1$ setting was first considered by Cule, Samworth and Stewart (2010); Cule and Samworth (2010). Furthermore, shape-constrained methods, such as log-concave maximum likelihood density estimation, are known to be naturally adaptive, achieving nearly parametric rates in certain settings (Kim and Samworth (2016)).

For the discrete setting, when $d = 1$, Weyermann (2008) shows the existence and uniqueness of the maximum likelihood estimator (MLE) for the log-concave probability mass function (PMF), and provides an active set algorithm to calculate the MLE that is much in the spirit of Rufibach (2007). Balabdaoui et al. (2013) introduced the log-concave MLE of a discrete distribution in one-dimensional space, and studied consistency and asymptotic properties of the estimator, while Balabdaoui and Jankowski (2016) compare this estimator with the MLE over the class of unimodal probability mass functions on $\mathbb{Z}$.

To our best knowledge, consideration of log-concave probability mass functions in the multidimensional discrete setting is limited to the work of Bapat (1988), see also Dharmadhikari and Joag-Dev (1988). Estimation methods/algorithms are not considered there. We review this class of discrete distributions, called generalized log-concave distributions, in Section 2.1.

In this work, we give a new definition of log-concave probability mass functions defined on the $\mathbb{Z}^d$ lattice, see Definition 1. We call this class extendible-log-concave, as it is closely related to extendible-convex functions (Murota and Shioura (2001)). We also derive some properties of the new class of distributions. Notably, we show that random variables from a continuous log-concave density which are binned (e.g. rounded to some accuracy level) will fall into our new class of mass functions, under certain conditions. In certain instances it might be desirable that the accuracy of the binning depend on the sample size. Tang, Banerjee and Kosorok (2012) consider the problem of analysis of binning in the context of current status data, and propose an adaptive inference procedure. It would be of interest to perform a similar analysis in our setting, although the issue is beyond the context of the current work.

In Section 3, we show that the maximum likelihood estimator exists and is unique, and we discuss its computation. Our algorithm is a modification of the algorithm of Cule, Samworth and Stewart (2010) to compute the MLE of a log-concave density in higher dimensions. We also show that the MLE is consistent, even under misspecification (this is done in Section 4). Furthermore, we study the finite sample size of our estimator via simulations. The proposed MLE exhibits considerable improvement in efficiency over the empirical distribution in the examples we consider. Moreover, in one of the examples we compare our nonparametric MLE to the correct parametric MLE, and the proposed method does not show a great loss of efficiency over the parametric method. Similar behaviour was observed by Balabdaoui et al. (2013). In our opinion, this is one of the key benefits of the balance that the log-concave class is able to strike between robustness and efficiency.

Although some results have recently been established for convergence rates for log-concave densities in $\mathbb{R}^d$ for $d > 1$ (Kim and Samworth (2016)), a more complete analysis of the limiting distribution in this case is still unavailable. On the other hand, the discrete setting is typically easier to handle. Our hope is that this work, aside from the practical utility of our method, will also provide an avenue for such theoretical exploration.

2. Discrete Log-Concavity in Higher Dimensions

A density function f is said to be log-concave if $h(x) = (-\log f)(x)$ is a convex function on $\mathbb{R}^d$. In particular, if h is sufficiently smooth, it is convex if and only if the Hessian matrix of h is positive semi-definite, Rockafellar (1970, Theorem 4.5). Similarly, one can define convex functions in the one-dimensional discrete setting via $(\triangle h)(z) = h(z-1) - 2h(z) + h(z+1) \geq 0$. This naturally leads to a definition of log-concave probability mass functions (Balabdaoui et al. (2013)). Perhaps surprisingly at first, in higher dimensions, the definition of discrete convexity is not so straightforward. For a discrete function defined on $\mathbb{Z}^d$ for $d > 1$ there are multiple definitions of convexity. Murota and Shioura (2001) provide a detailed survey of the various definitions available, and a summary of the relationships between them.

Our goal here is to define and study discrete log-concave distributions in higher dimensions, and we therefore need to select a class of discretely convex (equivalently, concave) functions to work with. Among the various discrete convex definitions introduced by Murota and Shioura (2001), we choose to focus on the class of convex-extendible functions. There are two main reasons for this: They show that the class of convex-extendible functions is closed under addition. Furthermore, using this definition, our class of log-concave probability mass functions is closed under taking limits. We will show this property in Theorem 1.

Define the convex closure of $h(z)$ as

$$\bar{h}(x) = \sup_{\alpha \in \mathbb{R}, \beta \in \mathbb{R}^d} \{\alpha + \beta^T x : \alpha + \beta^T z \leq h(z) \text{ for all } z \in \mathbb{Z}^d\}, x \in \mathbb{R}^d.$$

The function h is **convex-extendible** if $\bar{h}(z) = h(z)$ for all $z \in \mathbb{Z}^d$. Similarly, a set $S \subseteq \mathbb{Z}^d$ is said to be convex-extendible if $\bar{S} \cap \mathbb{Z}^d = S$, where $\bar{S} \subseteq \mathbb{R}^d$ is the convex closure of S, the smallest closed convex set (in $\mathbb{R}^d$) containing S. A related notion is that of a convex extension: A convex function $h^{\mathbb{R}} : \mathbb{R}^d \to \mathbb{R} \cup \{+\infty\}$ is called a convex extension of h if $h^{\mathbb{R}}(z) = h(z)$ for all $z \in \mathbb{Z}^d$. Clearly, a convex closure is a convex extension, but not vice versa. A function $h : \mathbb{Z}^d \to \mathbb{R} \cup \{-\infty\}$ is concave-extendible if $-h$ is convex-extendible.

Definition 1. *A PMF $p(z) : \mathbb{Z}^d \to [0, 1]$ is **e-log-concave** (eLC) if $p(z) = e^{\varphi(z)}$ and $\varphi(z)$ is concave-extendible.*

In what follows, we let $\mathcal{P}_0$ denote the class of all eLC probability mass functions on $\mathbb{Z}^d$.

Remark 1 (Separable-log-concavity). When $d = 1$, the class $\mathcal{P}_0$ agrees with the class of discrete log-concave distributions defined in Balabdaoui et al. (2013). This follows, for example, since $(\Delta\varphi)(z) \leq 0$ by appealing to the properties of a convex extension of φ. The maximum likelihood estimation considered here, when $d = 1$, has been studied in Balabdaoui et al. (2013). Furthermore, if a $\mathbb{Z}^d$-valued random variable $X = \{X_1, \ldots, X_d\}$ has a distribution which is e-log-concave and the elements $X_1, \ldots, X_d$ are known to be mutually independent, then the PMF can be written as $e^{\varphi(z)}$, where $-\varphi(z)$ is separable-convex. In such a situation, the multivariate MLE problem can be solved, again using the work of Balabdaoui et al. (2013); we fit a PMF $e^{\varphi}(z)$, where $-\varphi(z)$ is separable-convex by fitting a univariate log-concave probability mass function to each marginal of the data set and appealing to the independence assumption to obtain the joint distribution.

There is a simple way to verify if a discrete function is convex-extendible.

Lemma 1. Murota and Shioura (2001, Lemma 2.3) *If* $h : \mathbb{Z}^d \to \mathbb{R} \cup \{+\infty\}$ *is some function, then,* $\overline{h}(z) = h(z)$ *for any* $z \in \mathbb{Z}^d$ *if and only if there exists a closed convex extension of* h*.*

For example, consider $h(z) = z^T Az$, where $z \in \mathbb{Z}^d$, and A is a symmetric $d \times d$ positive-definite matrix. The natural convex extension of $h(z)$ is $h^{\scriptscriptstyle R}(x) = x^T Ax$ for $x \in \mathbb{R}^d$. The function is closed because it is continuous. By Lemma 1, $h(z)$ is convex-extendible. For another example, consider the proof of Proposition 2. Alternatively, for certain distributions, the relationship with generalized log-concave distributions is useful; see Section 2.1.

Remark 2 (Alternative lattice structures). In this work we limit ourself to the grid $\mathbb{Z}^d$, although potentially other lattice structures could also be explored. Simple linear transformations and rotations are naturally covered by our work. We conjecture that the convex extendible approach could also be applied to more irregular structures, although we do not explore this here. This is particularly attractive in light of the relationship that our definition has with log-concave densities, see Section 2.2.

Remark 3 (Unimodality). Several notions of unimodality exist for densities in $\mathbb{R}^d$ and mass functions on $\mathbb{Z}^d$ when $d > 1$. The class $\mathcal{P}_0$ is unimodal, in the sense that for all $z \in \mathbb{Z}^d$, the probability mass function is equal to $p(z) = \exp\left(-h^{\scriptscriptstyle R}(z)\right)$, where $h^{\scriptscriptstyle R}$ is a convex function defined not only on $\mathbb{Z}^d$, but also on $\mathbb{R}^d$.

An interesting example of a PMF which is not in the class eLC, is the following. Let

$$p(z) = \begin{cases} \dfrac{a}{(1+a)} & z = (0,0), \\ \dfrac{a(1-a)^k}{(1+a)} & z = (1,k),\ k \geq 0 \\ 0 & \text{otherwise,} \end{cases}$$

for some $a \in (0,1)$. Next, let $\alpha = -\log(a/(1+a))$ and $\beta = -\log(1-a) \geq 0$. Also, let $\mathcal{S}$ denote the support of the PMF p. The plane $\alpha + \beta x_2$ is convex and a minorant of $h(z) = -\log p(z)$ on $\mathbb{Z}^2$. Let conv S denote the convex hull of S. The function

$$\widetilde{h}(x) = \begin{cases} \alpha + \beta x_2 & x \in \operatorname{conv} \mathcal{S}, \\ +\infty & x \notin \operatorname{conv} \mathcal{S}, \end{cases}$$

satisfies $\widetilde{h}(z) = h(z)$ on $z \in \mathbb{Z}^2$, but $\widetilde{h}$ is not a closed function. In fact, $\overline{h}$ is the function

$$\overline{h}(x) = \begin{cases} \alpha + \beta x_2 & x \in \bar{\mathcal{S}}, \\ +\infty & x \notin \operatorname{conv} \mathcal{S}, \end{cases}$$

but it does not match h for $z \in \{(0,k) : k \geq 1\}$. Hence, $p \notin \mathcal{P}_0$. We call this the "shifted geometric" example, when we refer to it below.

2.1. Generalized log-concavity

Bapat (1988) (see also Johnson, Kotz and Balakrishnan (1997)) gave an alternative definition of "generalized log-concavity" on $\mathbb{N}^d$, where $\mathbb{N}$ denotes the natural numbers. A probability mass function p on $\mathbb{N}^d$ with support $\mathcal{S} = \{z \in \mathbb{N}^d : p(z) > 0\}$, is said to be generalized log-concave if

$$p(z) = \prod_{i=1}^{d} p_i(z_i), \quad z \in \mathcal{S}, \tag{2.1}$$

where each p_i satisfies $(\triangle \log p_i)(z_i) \leq 0$. Thus, each p_i is a univariate discrete log-concave function (though not necessarily a PMF). The definition need not be restricted to $\mathbb{N}^d$ and can easily be extended to $\mathbb{Z}^d$. Even with this extension, the definition is still more restrictive than our eLC definition for certain supports.

Proposition 1. *Suppose that p is generalized log-concave with support $\mathcal{S}$. If $\mathcal{S}$ is convex-extendible, then $p \in \mathcal{P}_0$.*

The discrete Gaussian distribution described later is not generalized log-concave. Thanks to Proposition 1 and Bapat (1988); Johnson, Kotz and Balakrishnan (1997), we find that distributions such as the multinomial, negative multinomial, multivariate hypergeometric, multivariate negative hypergeometric, as well as multi-parameter versions of the multinomial and negative multinomial are also extendible log-concave. We do this by checking that their supports are convex-extendible. Hence, Proposition 1 provides another approach to checking if a given probability mass function falls in the class $\mathcal{P}_0$.

2.2. Relationship with continuous log-concave distributions

Proposition 2. *Suppose that f is a log-concave density on $\mathbb{R}^d$, and let $A = [-1/2, 1/2)^d$. A probability mass function $p(z) = \int_{z+A} f(y)dy$ with a convex-extendible support satisfies $p \in \mathcal{P}_0$.*

This result is not tied to the lattice $\mathbb{Z}^d$ nor our particular choice of A. If Y is a random variable with this density f, then the PMF p with $A = [-1/2, 1/2)^d$ corresponds to the probability mass function of the random variable $\lfloor Y + 0.5 \rfloor$ (componentwise). Other choices of lattice and A lead to other discretizations of Y, such as $\delta\lfloor Y/\delta\rfloor$ for some $\delta > 0$. This means that the class $\mathcal{P}_0$ can be used to analyze log-concave random variables which have been discretized or “grouped/binned”.

2.3. Properties

The class $\mathcal{P}_0$ has several attractive properties.

Proposition 3. *Suppose $p \in \mathcal{P}_0$.*

1. *The support of p, $\mathcal{S} = \{z \,|\, p(z) > 0\}$, is a convex-extendible set.*
2. *For $\mathcal{A} \subset \mathcal{S}$, let*
$$\widetilde{p}(z) \propto \begin{cases} p(z) & z \in \mathcal{A}, \\ 0 & \textit{otherwise}. \end{cases}$$
If $\mathcal{A}$ is a convex-extendible set, $\widetilde{p} \in \mathcal{P}_0$.
3. *Let $p_1 \in \mathcal{P}_0$ and $p_2 \in \mathcal{P}_0$ with supports $\mathcal{S}_1 = \{z_1 \in \mathbb{Z}^{d_1} \,|\, p_1(z_1) > 0\}$ and $\mathcal{S}_2 = \{z_2 \in \mathbb{Z}^{d_2} \,|\, p_2(z_2) > 0\}$. Then $p(z) = p_1(z_1)p_2(z_2)$ with support $\mathcal{S} = \mathcal{S}_1 \times \mathcal{S}_2 \subset \mathbb{Z}^{d_1+d_2}$ satisfies $p \in \mathcal{P}_0$.*
4. *Suppose that $p \in \mathcal{P}_0$ with support in $\mathbb{Z}^d$, and let $z = (z_1, z_2)$ where $z_1 \in \mathbb{Z}^{d_1}$ and $z_2 \in \mathbb{Z}^{d_2}$ with $d_1 + d_2 = d$. Then the conditional distribution $p(z_1|z_2) = p((z_1, z_2))/p(z_2) \in \mathcal{P}_0$.*
5. *Let Z be a discrete random variable, with probability mass function $p \in \mathcal{P}_0$ with support $\mathcal{S}$. Let $\widetilde{Z} = AZ + b$, where A is a $d \times d$ matrix and b is a vector of length d, and let $\widetilde{p}$ denote the PMF of $\widetilde{Z}$ with support $\widetilde{\mathcal{S}}$. If $\widetilde{\mathcal{S}}$ is a subset of $\mathbb{Z}^d$, and the matrix A is invertible, then $\widetilde{p} \in \mathcal{P}_0$.*

The shifted geometric example does not have convex-extendible support (the convex hull of its support is not closed). The first property of Proposition 3 thus shows that it cannot be eLC. The class $\mathcal{P}_0$ is closed under limits with some restrictions.

Theorem 1. *Let p_n, p be discrete PMFs on $\mathbb{Z}^d$, and suppose that for each $n \geq 1$, $p_n \in \mathcal{P}_0$. If $p_n \to p$ pointwise and the support of p is convex-extendible, then $p \in \mathcal{P}_0$.*

The Kullback-Leibler (KL) divergence between two PMFs p and p_0 is defined as
$$\rho_{KL}(p \,\|\, p_0) = \sum_{z \in \mathbb{Z}^d} p_0(z) \log\left(\frac{p_0(z)}{p(z)}\right).$$

Although it is not a distance, but a divergence, is the natural notion of "distance" associated with maximum likelihood estimation. Let $\|z\|_\infty$ denote the maximum norm, $\|z\|_\infty = \max\{|z_1|, \ldots, |z_d|\}$.

Theorem 2. *Let p_0 be a probability mass function on $\mathbb{Z}^d$ such that $\sum_{z\in\mathbb{Z}^d} \|z\|_\infty p_0(z) < \infty$ and $\left|\sum_{z\in\mathbb{Z}^d} p_0(z)\log p_0(z)\right| < \infty$. If the convex hull of the support of p_0 is closed, there exists a unique $\widehat{p}_0$, such that*

$$\widehat{p}_0 = \operatorname*{argmin}_{p\in\mathcal{P}_0} \rho_{KL}(p \parallel p_0).$$

Furthermore, if $p_0 \in \mathcal{P}_0$, then $\widehat{p}_0 = p_0$.

We will refer to $\widehat{p}_0$ as the KL projection of p_0 in what follows. Heuristically, the KL projection is the closest element of the class $\mathcal{P}_0$ to the fixed PMF p_0.

Lemma 2. *The support of the KL projection $\widehat{p}_0$ is the intersection of $\mathbb{Z}^d$ with the (closed) convex hull of $\mathcal{S}_0$.*

3. Maximum Likelihood Estimation

The convex hull of a finite number of points is a closed polygon, from which it follows that the support of the empirical distribution is convex-extendible. The following is thus a simple consequence of Theorem 2.

Proposition 4. *If $X_1, \ldots, X_n$ are independent and identically distributed random variables on $\mathbb{Z}^d$ with true PMF p_0, then, with probability one, there exists a unique $\widehat{p}_n$ which maximizes the likelihood $\prod_{i=1}^n p(X_i)$ over the class of probability mass functions $p \in \mathcal{P}_0$.*

In what follows, we denote the MLE by

$$\widehat{p}_n = \text{argmax}_{p\in\mathcal{P}_0} \sum_{i=1}^n \log p(X_i).$$

Computation of this estimator is not an easy problem, see Walther (2009). For $d = 1$, the main options are the iterative convex minorant (ICMA) and active set algorithms, although other methods have also been used (Jongbloed (1998); Rufibach (2007)). These algorithms tend to rely on a special structure of convex functions which holds only for $d = 1$. For $d > 1$, this computational problem was first solved in Cule, Samworth and Stewart (2010), and it is their approach which we adapt to the discrete setting in this work. It is described in the next section.

A useful property of the eLC MLE, that holds in the continuous and discrete $d = 1$ cases, is the following.

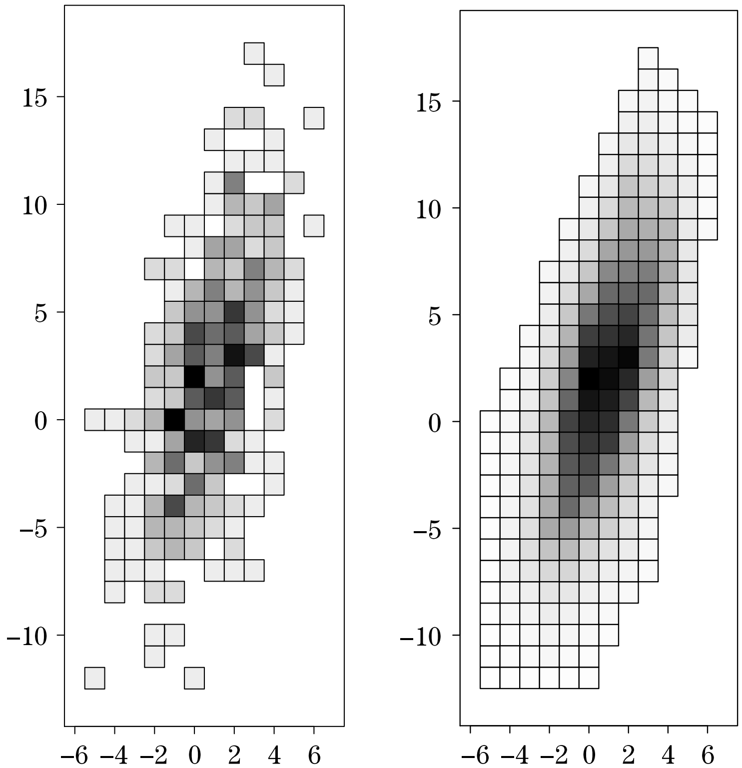

Figure 1. Grayscale heatmaps of the empirical PMF (left) and its eLC projection (right). The true distribution is a discrete Gaussian.

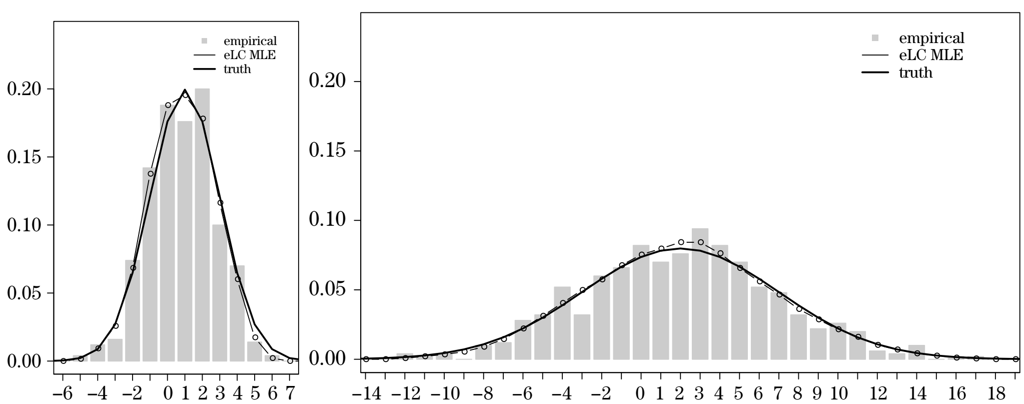

Figure 2. Marginal distributions corresponding to the heatmaps in Figure 1.

Lemma 3. *If $\overline{p}_n$ is the empirical PMF of the data and $h : \mathbb{Z}^d \mapsto \mathbb{R}$ is any convex-extendible function, then*

$$\sum_{z\in\mathbb{Z}^d} h(z)\,\widehat{p}_n(z) \le \sum_{z\in\mathbb{Z}^d} h(z)\,\overline{p}_n(z).$$

In particular, this implies that the mean of the MLE is equal to the observed mean of the data. Furthermore, if $\widehat{\Sigma}_n$ denotes the variance matrix under the MLE, and $\overline{\Sigma}_n$ the empirical variance matrix, then $\widehat{\Sigma}_n \le \overline{\Sigma}_n$, in the sense that $\overline{\Sigma}_n - \widehat{\Sigma}_n$ is positive semi-definite.

An example of the MLE is given in Figures 1 and 2. The data is an IID sample of size $n = 1{,}000$ from the discrete Gaussian distribution, see Section 3.2. Figure 1 shows the empirical distribution (left) and the fitted eLC (right) as a grey-scale heatmap. The marginal distributions are given in Figure 2, where the true marginals are also added.

3.1. Computation of the MLE

It is well-known that maximizing $\sum_{i=1}^n \log p(X_i)$ over $p \in \mathcal{P}_0$ is equivalent to minimizing

$$-\frac{1}{n}\sum_{i=1}^{n} \varphi(X_i) + \sum_{z\in\mathbb{Z}^d} \exp\left(\varphi(z)\right),$$

over all concave-extendible functions φ, see Lemma B.2. However, the values $X_1, \ldots, X_n$ are expected to have duplicates in our setting. Let $z_1, \ldots, z_m$ be the unique observed values of $X_1, \ldots, X_n$, and recall the empirical PMF $\overline{p}_n$. Let $\widehat{\mathcal{S}}_n = \bar{\mathcal{S}}_n \cap \mathbb{Z}^d$, where $S_n = \{z_1, \ldots, z_m\}$. Using also the characterization of the MLE, we can re-write this optimization problem as that of minimizing

$$-\sum_{j=1}^{m} \overline{p}_n(z_j)\,\varphi(z_j) + \sum_{z\in\widehat{\mathcal{S}}_n} \exp\left(\varphi(z)\right),$$

over all concave-extendible functions φ. Following Cule, Samworth and Stewart (2010), for a fixed vector of values $y \in \mathbb{R}^m$, define the "tent" function

$$t_y(x) = \quad \inf\{g(x) : \mathbb{R}^d \to \mathbb{R} \mid g \text{ is concave, and } g(z_j) \ge y_j \text{ for } j = 1, \ldots, m\}.$$

In our minimization problem, we can exchange the functions $\varphi(z)$ with the tent functions $t_y(z)$, and optimize over $y \in \mathbb{R}^m$ instead. A further simplification of the problem follows.

Theorem 3. *If*

$$\sigma(y_1, \ldots, y_m) = -\sum_{j=1}^{m} \overline{p}_n(z_j)\,y_j + \sum_{z\in\widehat{S}_n} \exp\left(t_y(z)\right),$$

then σ is convex and has a unique minimum $\widehat{y}$ such that $\widehat{p}_n(z) = \exp\left(t_{\widehat{y}}(z)\right)$.

Unfortunately, the function σ is not differentiable, and hence subgradient-based methods are used to perform the optimization. Details, including the algorithm, are given in the Appendix, and we refer to Cule (2009); Cule, Samworth and Stewart (2010) for the original development of these methods.

Remark 4. The function $t_{\widehat{y}}$ is a concave extension of $\log \widehat{p}_n$ (see Lemma B.3 in the Appendix). Thus, the algorithm finds not only $\widehat{p}_n$, but the associated concave extension.

A faster algorithm for log-concave density/PMF estimation in dimension greater than one remains an open problem in the field (see Cule, Samworth and Stewart (2010, Sec. 3) and Walther (2009, Sec. 5)). Although our algorithm in no way improves on the one proposed in Cule, Samworth and Stewart (2010), our general approach does reduce the number of data points from n to m, if one considers grouping/binning the data.

3.2. Finite sample preformance

We investigated the finite sample performance of the proposed method via simulations for $d = 2$. We considered two scenarios for the true p_0. For scenario (A), we assumed that $p_0(z) = p_1(z_1)p_2(z_2)$ where p_1 is Poisson ($\lambda = 4$) and p_2 is negative binomial ($p = 0.3, r = 6$). For scenario (B), we assumed that p_0 is discrete Gaussian, in that $p_0(z) \propto \exp(-0.5(z-\mu)^T\Sigma^{-1}(z-\mu))$, where $\mu = (1, 2)$, and

$$\Sigma = \begin{bmatrix} 4 & 6 \\ 6 & 25 \end{bmatrix}.$$

Consider the closed, continuous function $h^R(x) = 0.5(x-\mu)^T\Sigma^{-1}(x-\mu) + c$, $x \in \mathbb{R}^d$. It is easy to see that this is a convex extension of $h(z) = -\log p_0(z) = 0.5(z-\mu)^T\Sigma^{-1}(z-\mu) + c$, for the appropriate constant c. By Rockafellar (1970, Theorem 4.5) and Lemma 1 we conclude that $p_0 \in \mathcal{P}_0$.

For both scenarios, we simulated IID samples with sample sizes $n = 200, 500$, and 1,000. The results of our simulations are shown in Figure 3 with scenario (A) in the top row, and scenario (B) in the bottom row. Each boxplot is the result of $B = 1{,}000$ repetitions and compares the performance of our eLC estimator as well as three others, via the l_2 distance of the estimator to the true PMF p_0. The other estimators are the empirical PMF $\overline{p}_n$, the MLE assuming that $-\log p(z) = h_1(z_1) + h_2(z_2)$, for two convex functions h_1, h_2 on $\mathbb{Z}$ (see Remark 1) and the correct parametric MLE, where the latter is calculated for scenario (A)

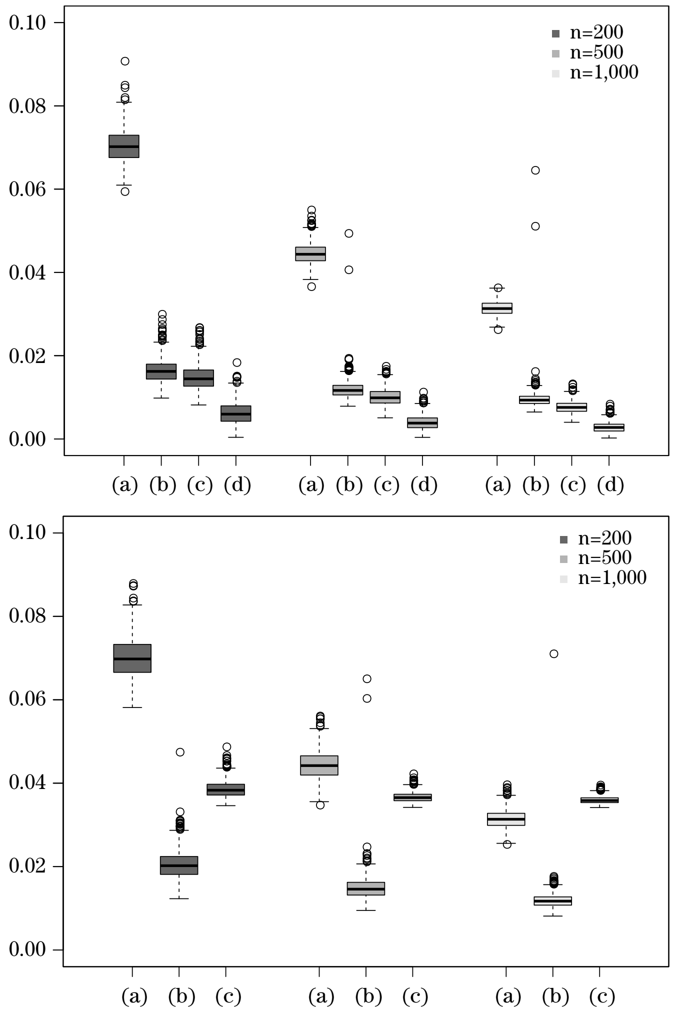

Figure 3. Boxplots of l_2 distance between estimator and true distribution when the true distribution is Poisson and negative binomial product (top) and discrete Gaussian (bottom). The estimators are (a) empirical MLE, (b) eLC MLE, (c) separable log-concave MLE, and (d) parametric MLE (top plot only).

only. Clearly, the more (correct) assumptions we make, the more we increase efficiency without increasing bias – this is seen in the top plot. In the bottom plot, the misspecified MLE (c) has poor performance. In our opinion, the success story of the eLC estimator is seen in the top plot: there is not that much loss of efficiency for the MLE between the nonparametric eLC assumption versus the strong correct parametric assumption.

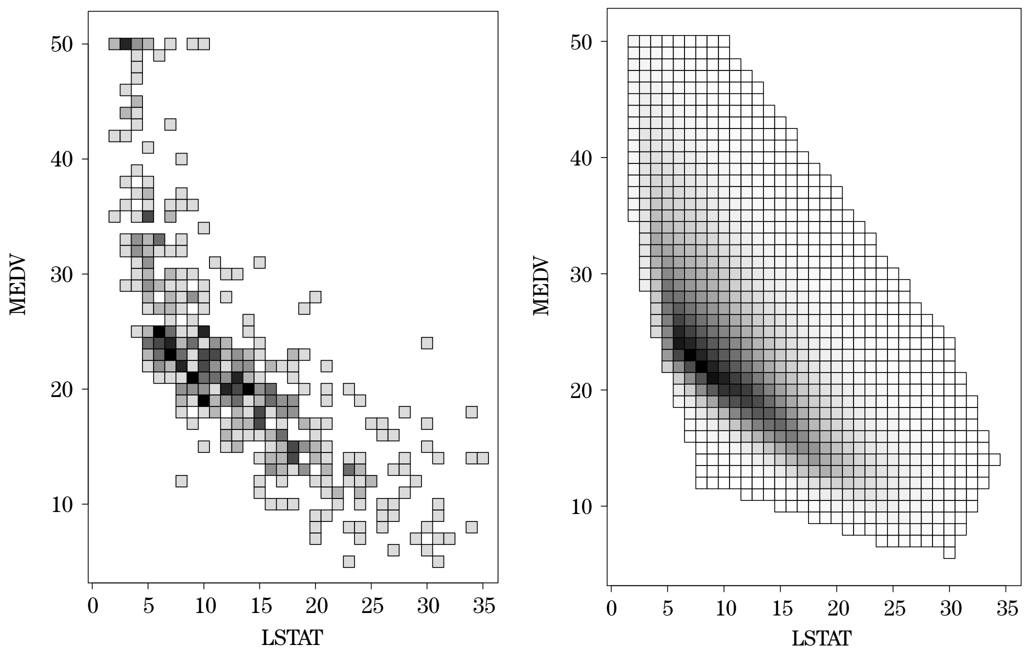

Figure 4. Boston Housing Data: original empirical distribution (left) along with eLC maximum likelihood estimate (right).

3.3. Mixtures and the EM algorithm

As mentioned in Chang and Walther (2007); Walther (2009); Cule, Samworth and Stewart (2010), one of the advantages of the maximum likelihood approach over a fixed family of functions is that it naturally extends to fitting of mixture models via the EM algorithm, for apriori known number of mixtures. Although we do not explore this in detail, this approach could extend our class into possibly multimodal distributions as well. Alternatively, the class of, say, zero-inflated distributions could also be considered.

3.4. Binned data example

We illustrate our estimation technique on the Boston housing data set created by Harrison and Rubinfeld (1978) and available online at Lichman (2013). The data set consists of a sample size of $n = 506$ and 14 variables. We chose to work with the last two variables: LSTAT (percentage lower status of the population) and MEDV (median value of owner occupied homes in \$1,000s). Prior to binning, LSTAT has a range of $(1.73, 37.97)$ with a median/mean value of $11.36/12.65$, while MEDV has a range of $(5.00, 50.00)$ with a median/mean value of 21.20/22.53. We removed observations with missing values (only an issue for MEDV) for a sample size of $n = 452$. We binned the data as described in Section 2.2, using $x_i = \lfloor y_i + 0.5 \rfloor$ for each observation, and for both variables. This creates $m = 270$ unique bins. Figure 4 shows the result of fitting the eLC MLE to the binned data, along with the original histogram.

4. Asymptotic Properties

For two PMFs p and q, we define the l_k and Hellinger distances as

$$l_k(p,q) = \begin{cases} \left\{ \sum_{z\in\mathbb{Z}^d} |p(z)-q(z)|^k \right\}^{1/k} & \text{if } 1 \le k < \infty, \\ \sup_{z\in\mathbb{Z}^d} |p(z)-q(z)| & \text{if } k = \infty, \end{cases}$$

$$h^2(p,q) = \frac{1}{2}\sum_{z\in\mathbb{Z}} \left\{ \sqrt{p(z)} - \sqrt{q(z)} \right\}^2.$$

Our main consistency result follows.

Theorem 4. *Suppose that p_0 is a discrete distribution on $\mathbb{Z}^d$ satisfying*

$$\sum_{z\in\mathbb{Z}^d} \|z\|_\infty \, p_0(z) < \infty \text{ and } \left| \sum_{z\in\mathbb{Z}^d} p_0(z) \log p_0(z) \right| < \infty.$$

If the convex hull of the support of p_0 is closed, then $d(\widehat{p}_n, \widehat{p}_0) \to 0$ almost surely, where d is the distance l_k for any $1 \le k \le \infty$, or the Hellinger distance h.

We see that even if the true distribution p_0 is not in $\mathcal{P}_0$, the MLE still converges, and it converges to $\widehat{p}_0$, the best approximation to p_0 in $\mathcal{P}_0$. Such robustness properties are known to hold for other shape-constrained estimators (based on maximum likelihood), and for other maximum likelihood estimators in general. They are a very appealing aspect of the method and can be interpreted to say that, even if $p_0 \notin \mathcal{P}_0$, then if p_0 is "close" to $\mathcal{P}_0$, our proposed MLE will exhibit desirable behaviour.

Supplementary Materials

All appendices appear in the Supplementary Material. R code (R Core Team, 2017) to recreate the example from Figures 1 and 2 is also provided as Supplementary Material. It is also available as a zipped folder at `http://www.math.yorku.ca/~hkj/Research/code_eLCD.zip`.

References

Balabdaoui, F. and Jankowski, H. (2016). Maximum likelihood estimation of a unimodal probability mass function. *Statist. Sinica* **26**, 1061–1086.

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.

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.

Bapat, R. B. (1988). Discrete multivariate distributions and generalized log-concavity. *Sankhyā Ser. A* **50**, 98–110.

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

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

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

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.

Cule, M., Samworth, R. and Stewart, M. (2010). Maximum likelihood estimation of a multidimensional log-concave density. *J. R. Stat. Soc. Ser. B Stat. Methodol.* **72**, 545–607.

Dharmadhikari, S. and Joag-Dev, K. (1988). *Unimodality, Convexity, and Applications.* Probability and Mathematical Statistics, Academic Press, Inc., Boston, MA.

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

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

Hall, P. and Heckman, N. E. (2000). Testing for monotonicity of a regression mean by calibrating for linear functions. *Ann. Statist.* **28**, 20–39.

Hall, P. and Presnell, B. (1999). Density estimation under constraints. *J. Comput. Graph. Statist.* **8**, 259–277.

Harrison, D. and Rubinfeld, D. L. (1978). Hedonic housing prices and the demand for clean air. *J. Environ. Econ. Manag.* **5**, 81–102.

Johnson, N. L., Kotz, S. and Balakrishnan, N. (1997). *Discrete Multivariate Distributions.* Wiley Series in Probability and Statistics: Applied Probability and Statistics, John Wiley & Sons, Inc., New York. A Wiley-Interscience Publication.

Jongbloed, G. (1998). The iterative convex minorant algorithm for nonparametric estimation. *J. Comput. Graph. Statist.* **7**, 310–321.

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

Lichman, M. (2013). UCI machine learning repository. `http://archive.ics.uci.edu/ml`.

Murota, K. and Shioura, A. (2001). Relationship of M-/L-convex functions with discrete convex functions by Miller and Favati-Tardella. *Discrete Appl. Math.* **115**, 151–176. 1st Japanese-Hungarian Symposium for Discrete Mathematics and its Applications (Kyoto, 1999).

R Core Team (2017). R: A Language and Environment for Statistical Computing. R Foundation for Statistical Computing, Vienna, Austria. `https://www.R-project.org`.

Rockafellar, R. T. (1970). *Convex Analysis.* Princeton Mathematical Series, No. 28, Princeton

University Press, Princeton, N.J.

Rufibach, K. (2007). Computing maximum likelihood estimators of a log-concave density function. *J. Stat. Comput. Simul.* **77**, 561–574.

Tang, R., Banerjee, M. and Kosorok, M. R. (2012). Likelihood based inference for current status data on a grid: A boundary phenomenon and an adaptive inference procedure. *Ann. Statist.* **40**, 45–72.

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

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

Weyermann, K. (2008). An active-set algorithm for the estimations of discrete log-concave densities. Master's thesis, University of Bern, Switzerland.

Department of Mathematics and Statistics, York University, Toronto, ON M3J 1P3 Canada.

E-mail: hkj@yorku.ca

Department of Mathematics and Statistics, York University, Toronto, ON M3J 1P3 Canada.

E-mail: tian.amanda@hotmail.com

(Received February 2017; accepted September 2017)