MAXIMUM LIKELIHOOD ESTIMATION OF DISCRETE LOG-CONCAVE DISTRIBUTION WITH APPLICATIONS

YAN HUA TIAN

A DISSERTATION SUBMITTED TO THE FACULTY OF GRADUATE STUDIES IN PARTIAL FULFILLMENT OF THE REQUIREMENTS FOR THE DEGREE OF DOCTOR OF PHILOSOPHY

GRADUATE PROGRAM IN MATHEMATICS AND STATISTICS
YORK UNIVERSITY
TORONTO, ONTARIO

March 2018

Abstract

Shape-constrained methods specify a class of distributions instead of a single parametric family. The approach increases the robustness of the estimation without much loss of efficiency. Among these, log-concavity is an appealing shape constraint in distribution modeling, because it falls into the popular unimodal shape-constraint and many parametric models are log-concave. This is therefore the focus of our work.

First, we propose a maximum likelihood estimator of discrete log-concave distributions in higher dimensions. We define a new class of log-concave distributions on $\mathbb{Z}^d$, and study its properties. We show how to compute the maximum likelihood estimator 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, the proposed estimator outperforms a purely nonparametric approach (the empirical distribution), but is able to remain comparable to the correct parametric approach. Furthermore, the new class has a natural relationship with log-concave densities when data has been grouped or discretized. We show how this property can be used in a real data example.

Secondly, we apply the discrete log-concave maximum likelihood estimator in one-dimensional space to a clustering problem. Our work mainly focuses on the categorical nominal data. We develop a log-concave mixture model using the discrete log-concave maximum likelihood estimator. We then apply the log-concave mixture

model to our clustering algorithm. We compare our proposed clustering algorithm with the other two clustering methods. Comparing results show that our proposed algorithm has a good performance.

Contents

List of Tables

List of Figures

Part I

eLC maximum likelihood estimator

Chapter 1

Introduction

1.1 Motivation

Estimating probability density/mass function is very important in statistics. Parametric and nonparametric methods are two popular methodologies in this area. The parametric method is efficient and has good performance, because specific assumptions are added to the estimation. The disadvantage is that the estimation has large bias when the assumption is wrong. Some example parametric models include: normal distribution, uniform distribution, poisson distributions. The nonparametric method such as the empirical distribution does not need specific parameter assumption, is therefore more robust and data-adaptive, but it is not efficient and has larger variance.

Shape-constrained estimation method is one kind of nonparametric method, which means that the probability density/mass function is estimated with some shape-constrained assumptions. The shape constraint includes but is not limited to unimodality, monotonicity, symmetric, log-concavity. We refer to Wolters and Braun (2017) for the recent works, as well as the advantages and the challenges of shape-constrained estimation models. Grenander (1956) introduced a nonparametric method to estimate the MLE of non-increasing densities, which is known as the Grenander estimator. Other works focused on monotony densities include Groeneboom (1985); Huang and Wellner (1995). Based on the Grenander estimator, later research attempted to estimate the unimodal densities. Birgé (1997) introduced an estimator for estimating non-smooth unimodal densities, when the mode is known. Other works related to unimodal density estimation include Rao (1969); Bickel and Fan (1996).

Note that shape-constrained estimation method stands in the middle of parametric and nonparametric methods. Adding shape-constrained assumptions to distribution improves the performance and reduces the variance comparing to nonparametric method. On the other hand, shape-constrained models specify a larger class than any specific parametric model. Hence comparing to parametric method, it is more robust. For example, Bickel and Fan (1996) introduced density estimation

methods with unimodal shape-constraint and compared their methods with non-parametric kernel estimation. Their methods do a better job for the tail of density and when the distribution is asymmetric. In Section 3.2, we will show that our proposed shape-constrained estimator has an outstanding better performance than the non-parametric estimator, and remains comparable performance comparing to the strong parametric estimator. Of course, there is no guarantee of this in all circumstances. Among those shape constraints, log-concavity is an attractive shape constraint. It includes a wide range of parametric models. For example: normal, uniform, gamma(r, λ) with $r \geq 1$, beta(a, b) with $a \geq 1, b \geq 1$ in continuous setting, and multinomial, negative multinomial, multivariate hypergeometric in discrete setting, they are all log-concave. Furthermore, log-concave densities provide a natural alternative to the class of unimodal densities, while not being too restrictive by specifying a parametric family. Notably, estimating the maximum likelihood estimator of a unimodal density is not an easy problem. Let d denotes the dimension of space, the maximum likelihood estimator does not even exist when $d = 1$, because the likelihood function can go to infinity (Birgé, 1997). Log-concave estimation falls in the larger paradigm of shape-constrained estimation, inherits the advantage of striking a balance of efficiency and robustness, also benefits from such properties as (local) adaptivity and not requiring bandwidth.

Walther (2009) introduced the attractive properties of log-concavity and gave a brief review of recent works. For example, Balabdaoui et al. (2009) found the limiting distributions of the nonparametric MLE of a log-concave density; Lutz and Kaspar (2011) and Dümbgen et al. (2007) discussed algorithms and R packages (**logcondens**) to compute the log-concave MLE in continuous setting; Cule et al. (2010) presented theoretical properties of MLE of log-concave density in multiple dimensional space, developed algorithm to compute the MLE and proved the consistency of the MLE; Works of Doss and Wellner (2016) and Kim and Samworth (2016) focused on global convergence rates for the MLE of log-concave densities. For discrete setting, when $d = 1$, Weyermann (2008) showed the existence and uniqueness of the MLE for the log-concave probability mass function (PMF), and provided an active set algorithm to calculate the MLE, which is much in the spirit of Dümbgen et al. (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) compared this estimator with the MLE over the class of unimodal probability mass functions on $\mathbb{Z}$.

Among those work, we highlight the works of Cule et al. (2010); Balabdaoui et al. (2013) (will be introduced in the following sections). Our work is much in the spirit of their works. Some of our theoretical results are inspired by the work of Balabdaoui

et al. (2013), for example, the uniqueness, existence and consistency of the MLE. Our algorithm to compute the MLE is much inspired by the work of Cule et al. (2010). We adapt their algorithm and code for discrete setting. In our algorithm, we design and solve the optimization problem following their work. The main difference is that their MLE is in continuous setting, but ours is in discrete setting. More details about the difference will be given in later chapter.

1.2 Overview of maximum likelihood estimation

Maximum likelihood estimation is a well known statistic methodology. With observed data, it attempts to estimate distribution and parameters by maximizing the likelihood function. For example, MLE can be used to estimate parameters in a parametric model or the density function with nonparametric shape-constrained methods.

Let $x_1, ..., x_n$ be n independent and identically distributed observations from some unknown distribution. We denote the distribution with $f(x|\theta)$, where θ is a parameter vector. The likelihood function is defined as:

$$\mathcal{L}(\theta; x_1, ..., x_n) = \prod_{i=1}^{n} f(x_i|\theta).$$

For computation convenience, we apply logarithm to both sides of the function, and obtain the log-likelihood function:

$$\ln\mathcal{L}(\theta; x_1, ..., x_n) = \sum_{i=1}^{n} \ln f(x_i|\theta).$$

The maximum likelihood estimator $\widehat{\theta}_{mle}$ is defined as:

$$\widehat{\theta}_{mle} = \text{argmax}_{\theta \epsilon \Theta} \prod_{i=1}^{n} f(x_i|\theta),$$

where Θ is the family of parameter θ. Equivalently, maximum likelihood estimator (MLE) can also be expressed by maximizing the average of log-likelihood function:

$$\widehat{\theta}_{mle} = \text{argmax}_{\theta \epsilon \Theta} \frac{1}{n} \sum_{i=1}^{n} \ln f(x_i|\theta).$$

The idea behind maximum likelihood estimator is that the most likely parameter maximizes the likelihood function.

MLE is widely used in statistics. Indeed, under certain regulations the MLE has the following properties.

- consistency: MLE converges in probability to the true value θ_0, when sample

size n large enough. That is,

$$\widehat{\theta}_{mle} \xrightarrow{p} \theta_0, \text{ as } n \longrightarrow \infty.$$

- asymptotic normality:

$$\sqrt{n}(\widehat{\theta}_{mle} - \theta_0) \xrightarrow{d} \mathcal{N}(0, \Sigma_{mle}).$$

- efficiency: MLE is asymptotically efficient since its variance approaches Cramer-Rao lower bound.

Geman and Hwang (1982) mentioned that MLE methodology usually fails when the parameters is in an infinite dimensional space. Hence MLE can not be applied to completely nonparametric statistic models. The MLE of such model does not exist because the likelihood function is unbounded. But situation changes when some constraints are added to the model. MLE works efficiently with shape-constrained nonparametric model. for example Cule et al. (2010) introduced an algorithm to compute the MLE of log-concave densities in multiple dimensional space.

1.3 Background

1.3.1 MLE of log-concave density on $\mathbb{R}^d$

A density function f is said to be log-concave if $(-\log f)(x)$ is a convex function on $\mathbb{R}^d$. A function h is convex if for all $x, x' \in \mathbb{R}^d$ and for all $\alpha \in [0,1]$ it satisfies $h(\alpha x + (1-\alpha)x') \leq \alpha h(x) + (1-\alpha)h(x')$. Let $\mathcal{F}$ denote the set of log-concave density functions on $\mathbb{R}^d$, given $x_1, ..., x_n$ are a random sample drawn from a log-concave density. The log-concave MLE on $\mathbb{R}^d$ is defined as:

$$\widehat{f_n} = \text{argmax}_{f \in \mathcal{F}} \sum_{i=1}^{n} \log f(x_i).$$

Theorem 1.3.1. *(Cule et al., 2010) With probability one, a log-concave maximum likelihood estimator $\widehat{f_n}$ of f_0 exists and is unique, where f_0 is the true density on $\mathbb{R}^d$.*

The above Theorem proved the existence and uniqueness of the log-concave MLE. The algorithm computing the MLE is derived based on the tent function t_y. Following Cule (2009), given $y \in \mathbb{R}^n$, a tent function can be explicitly defined as

$$t_y(x) = \inf\{g(x) : \mathbb{R}^d \longrightarrow \mathbb{R} \mid \text{g is concave, and } g(x_j) \geq y_j, j = 1, ..., n\}.$$

To understand the tent function, we can imagine a two-dimensional problem. That

is, $x_1, ..., x_n \in \mathbb{R}^2$. Hence $x_1, ..., x_n$ correspond points on a plane. We then put poles with height of $y_1, ..., y_n$ on those points, respectively. Finally, we stretch a piece of rubber over those poles. The surface of rubber is exactly the tent function for the given $y_1, ..., y_n$. The tent function actually reflects concave shape constraint, and it is a piece-wise affine function.

Cule et al. (2010) compute their log-concave MLE by minimizing the objective function over $y = (y_1, ..., y_n) \in \mathbb{R}^n$. The objective function is defined as follows:

$$\sigma_R(y_1, ..., y_n) = -\frac{1}{n}\sum_{i=1}^{n} y_i + \int_{C_n} \exp\{t_y(x)\}dx,$$

where C_n is the convex hull of $\{x_1, ..., x_n\}$, that is $C_n = \text{conv}\{x_1, ..., x_n\}$. Note that the objective function is convex (Cule et al., 2010), and it has a unique minimum. Unfortunately, it is proved not to be smooth everywhere, hence many efficient optimization techniques can not be applied to this. Subgradient methodology is finally chosen to solve the problem.

The general idea of subgradient algorithms is to proceed iteratively as follows:

Theorem 1.3.2. *(Shor, 1985) Let (h_i) be a positive sequence with $h_i \to 0$ as $i \to \infty$ and $\sum_{i=0}^{\infty} h_i = \infty$. Then, for any convex function σ, the sequence generated by the*

formula

$$y_{i+1} = y_i - h_i \frac{\partial \sigma(y_i)}{\| \partial \sigma(y_i) \|}$$

has the property that either there exists an i_0 *and* y^* *such that* $y_{i_0} = y^*$, *or* $y_i \to y^*$ *and* $\sigma(y_i) \to \sigma(y^*)$ *as* $i \to \infty$.

1.3.2 MLE of log-concave mass function on $\mathbb{Z}$

In this section we introduce the recent work of log-concave MLE in 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, and provides an active set algorithm to calculate the MLE, which is much in the spirit of Dümbgen et al. (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). They defined a class of generalized log-concave distribution. But there is no further study about the property or algorithm

to approximate the distribution. Their definition is a more restrictive class than our proposed definition. More detailed review will be given in later section.

Other than the formal definition of the convex function we mentioned in previous section. Alternatively, if h is twice differentiable, then h is convex if and only if $h''(x) \geq 0$ for all $x \in \mathbb{R}$ and if and only if the Hessian matrix of h is positive semi-definite for all $x \in C$, where C is an open convex set on $\mathbb{R}^d$ for $d > 1$ (Rockafellar, 1970, Theorem 4.5, page 27).

Similarly, one can define convex functions in the one-dimensional discrete setting, which naturally leads to a definition of log-concave probability mass functions. That is, let $p(z) : \mathbb{Z} \to [0,1]$ denote a probability mass function, where $\mathbb{Z}$ denotes the integers

$\{\ldots, -2, -1, 0, 1, 2, \ldots\}$. The PMF p is said to be log-concave if for any $z \in \mathbb{Z}$

$$(\triangle h)(z) \quad = \quad h(z-1) - 2h(z) + h(z+1) \quad \geq \quad 0, \tag{1.1}$$

where $h(z) = (-\log p)(z)$ (Balabdaoui et al., 2013, Proposition 1). In the notation above $(\triangle h)$ denotes the discrete Laplacian operator, which can also be expressed as $(\triangle h)(z) = \{h(z+1) - h(z)\} - \{h(z) - h(z-1)\}$. This is the second difference of the function h, and hence this definition matches well that of the continuous setting.

For two probability distributions p_0 and p, KL divergence (Kullback and Leibler,

1951) is used to measure the information gain/loss when we use one probability distribution p to approximate another probability distribution p_0, which is usually the true distribution. The KL divergence is defined as

$$\rho_{KL}(p\|p_0) = -\sum_{z\in\mathbb{Z}^d} p_0(z)\log\frac{p(z)}{p_0(z)}.$$

Although it is not a distance but a divergence, it is the natural notion of "distance" associated with maximum likelihood estimation.

Let $\mathcal{P}_1$ denote the class of log-concave PMFs on $\mathbb{Z}$. Balabdaoui et al. (2013) proved the following Theorem.

Theorem 1.3.3. *(Balabdaoui et al., 2013) Suppose that p_0 is a discrete PMF on $\mathbb{Z}$ with finite mean such that*

$$|\sum_{z\in\mathbb{Z}} p_0(z)\log\{p_0(z)\}| < \infty.$$

Then there is a unique log-concave PMF on $\mathbb{Z}$, $\widehat{p}_0$, such that

$$\widehat{p}_0 = \text{argmin}_{p\in\mathcal{P}_1}\,\rho_{KL}(p\|p_0).$$

They call $\widehat{p}_0$ the KL projection of p_0. Because of the natural relation between KL

divergence and likelihood function, it is not hard to show that the log-concave MLE exists and is unique on $\mathbb{Z}$. We denote the log-concave MLE ($d = 1$) as $\widehat{p}_n$.

Let $g(z)$ denote a finite concave function on $\mathbb{Z}$. A point $\mathcal{K} \in \mathbb{Z}$ is a knot of g if $g(\mathcal{K}) > -\infty$, and g changes slope at $\mathcal{K}$. Further, a knot $\mathcal{K}$ is called an internal knot if $g(\mathcal{K}-1) > -\infty$ and $g(\mathcal{K}+1) > -\infty$. Let $(x_1, ..., x_m)$ be a random sample from p_0, we assume there are m distinct ordered values in the sample: $z_1 < ... < z_m$.

Discrete concave function $g(z)$ when $d = 1$ can be decomposed to the following form (Balabdaoui et al., 2013):

$$g(z) = a + bz + \sum_{i=1}^{p} c_i(\mathcal{K}_i - z)_+, z \in \mathbb{Z} \cap [z_1, z_m]$$

where $a, b \in \mathbb{R}$ and $c_i < 0$, $\mathcal{K}_1, ..., \mathcal{K}_p$ denote the internal knots of $g(z)$. Here they used the standard notation $z_+ = zI_{\{z \geq 0\}}$. This decomposition is the key to active set algorithm (Weyermann, 2008; Dümbgen et al., 2007). When $d > 1$ the discrete concavity is defined in a totally different way. The definition of discrete concavity is even not unique. There are no decomposition methods for higher dimensional discrete concave function, hence active set algorithm can not be applied to higher dimensions.

For two PMFs p and p_0, we define the l_k and Hellinger distances as

$$\begin{aligned} l_k(p, p_0) &= \begin{cases} \left(\sum_{z \in \mathbb{Z}^d} |p(z) - p_0(z)|^k \right)^{1/k} & \text{if } 1 \le k \le \infty, \\ \sup_{z \in \mathbb{Z}^d} |p(z) - p_0(z)| & \text{if } k = \infty, \end{cases} \\ h^2(p, p_0) &= \frac{1}{2} \sum_{z \in \mathbb{Z}} \left(\sqrt{p(z)} - \sqrt{p_0(z)} \right)^2. \end{aligned}$$

Balabdaoui et al. (2013) proved that the log-concave MLE on $\mathbb{Z}$ is consistent in term of the distance $l_k, 0 < k \le \infty$ or the Hellinger distance.

Theorem 1.3.4. *(Balabdaoui et al., 2013) Suppose that p_0 is a discrete distribution on $\mathbb{Z}$ with finite mean such that*

$$\left| \sum_{z \in \mathbb{Z}} p_0(z) \log\{p_0(z)\} \right| < \infty.$$

Then $d(\widehat{p}_n, \widehat{p}_0) \to 0$ almost surely, where d is the distance $l_k, 0 < k \le \infty$ or the Hellinger distance.

1.3.3 Discrete convex in higher dimensions

In higher dimensions, the definition of a discrete convex (equivalently, concave) function is not 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 sur-

vey of convex functions and sets in the higher-dimensional discrete setting, including a summary of the relationships between the various definitions. Among these definitions there are three which are relevant to our initial considerations: discretely-convex, separable-convex, and convex-extendible. To this end, consider a function $h : \mathbb{Z}^d \to \mathbb{R} \cup \{+\infty\}$ and define the domain $\mathrm{dom}(h) = \{z \in \mathbb{Z}^d \,|\, h(z) < \infty\}$.

– The function h is said to be **separable-convex** if $h(z) = \sum_{i=1}^{d} h_i(z_i)$ $(z \in \mathbb{Z}^d)$ for a finite family of discrete convex functions $h_i : \mathbb{Z} \to \mathbb{R} \cup \{+\infty\}$, $i \in \{1, ..., d\}$. That is, $(\triangle h_i)(z) \geq 0$ for all $z \in \mathbb{Z}$ and all $i \in \{1, \ldots, d\}$.

– For $x \in \mathbb{R}^d$, let $\lfloor x \rfloor$ (respectively, $\lceil x \rceil$) denote the floor (respectively, the ceiling) of the vector x, obtained by rounding down (respectively, up) each component of x to its nearest integer. Next, define the set $N_0(x) = \{z \in \mathbb{Z}^d \,|\, \lfloor x \rfloor \leq z \leq \lceil x \rceil\}$. The function h is said to be **discretely-convex** if, for any $z', z'' \in \mathrm{dom}(h)$ and any $\alpha \in [0, 1]$, it holds that

$$\min\{h(z) \,|\, z \in N_0(\alpha z' + (1-\alpha)z'')\} \quad \leq \quad \alpha h(z') + (1-\alpha)h(z'').$$

Similarly, a set $S \subseteq \mathbb{Z}^d$ is said to be discretely-convex if, for any $z', z'' \in S$ and any $\alpha \in [0, 1]$, it holds that $N_0(\alpha z' + (1-\alpha)z'') \cap S$ is non-empty.

– Define the convex closure of $h(z)$

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

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, that is, it is the smallest closed convex set (in $\mathbb{R}^d$) containing S. Another useful definition is convex extension: a closed convex function $h^R : \mathbb{R}^d \to \mathbb{R} \cup \{+\infty\}$ is called a convex extension of h if $h^R(z) = h(z)$ for all $z \in \mathbb{Z}^d$. For a discrete convex-extendible function, its convex extension is a closed continuous convex function which goes through all its points. We may also take a discrete convex-extendible function as a "sub" function of a closed convex function in continuous setting. Note that affine functions are convex. By (Rockafellar, 1970, Theorem 5.5, page 35), the pointwise supremum of an arbitrary collection of convex functions is convex. Hence we conclude that the convex closure $\bar{h}$ is convex. It is also a convex extension of h. Any convex function is a composition of collections of pointwise affine functions, so convex closure is the greatest convex extension of h. But clearly not all convex extension is convex closure.

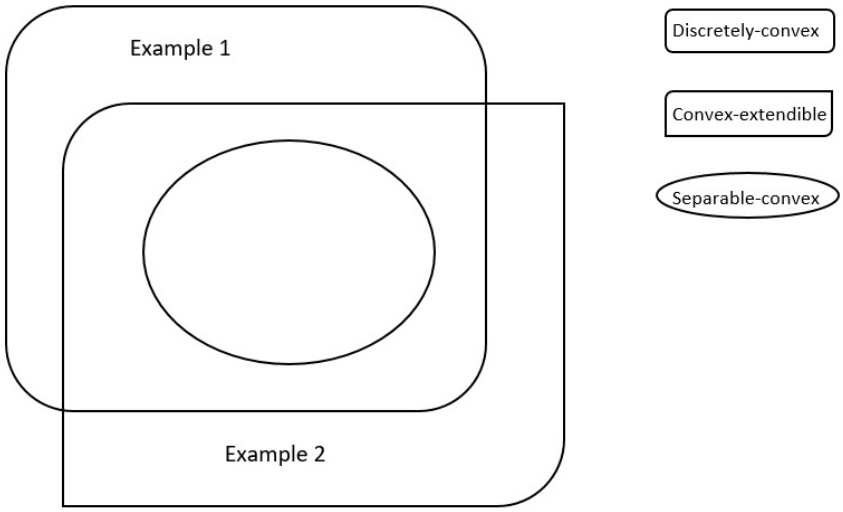


Figure 1.1: Relations between three definitions: discrete-convex, convex-extendible, and separable-convex.

Murota and Shioura (2001) summarize the relationships between the various definitions of convexity. In particular, some but not all discretely-convex functions are convex-extendible functions and vice versa, while separable-convex functions are both discrete-convex and convex-extendible. Figure 1.1 shows the relations between above three discrete convex definitions. We can see that

$$\text{separable-convex} \subset (\text{discretely-convex} \cap \text{separable-convex}).$$

Example 1: Consider the set

$$\begin{aligned} S &= \{z \in \mathbb{Z}^3 | z_1 + z_2 + z_3 = 2, z_i \geq 0, i = 1,2,3\} \cup \{(1,2,0),(0,1,2),(2,0,1)\} \\ &= \{(0,1,1),(1,0,1),(1,1,0),(0,0,2),(0,2,0),(2,0,0),(1,2,0),(0,1,2),(2,0,1)\}. \end{aligned}$$

This set, as well as the function h equal to zero on S and $+\infty$ on $\mathbb{Z} \setminus S$, are discrete-convex. However,

$$\tfrac{1}{3}(1,2,0) + \tfrac{1}{3}(0,1,2) + \tfrac{1}{3}(2,0,1) = (1,1,1)$$

is an element of $\bar{S} \cap \mathbb{Z}^d$, but $(1,1,1) \notin S$, and hence h is not convex-extendible.

Example 2: Let $S = \{(0,0),(2,1)\}$ and again define the function h equal to zero on S and $+\infty$ on $\mathbb{Z} \setminus S$. The convex closure of S is the segment between points (0,0) and (2,1), hence $\bar{S} \cap \mathbb{Z}^d = \{(0,0),(2,1)\} = \{z_1, z_2\} = S$, we conclude that h is convex-extendible. On the other hand, $N_0(0.5z_1 + 0.5z_2) = N_0((1,0.5)) = \{(1,0),(1,1)\}$, whence $N_0(\alpha x' + (1-\alpha)x'') \cap S = \varnothing$ and h is not discrete-convex. Both examples appear in Murota and Shioura (2001).

1.3.4 Generalized log-concave probability mass function

Bapat (1988) (see also Johnson et al. (1997, page 28)) 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

$$p(z) = \prod_{i=1}^{d} p_i(z_i), \quad z \in \mathcal{S}, \tag{1.2}$$

where each p_i satisfies $(\triangle \log p_i)(z_i) \leq 0$. That is, each p_i is a univariate discrete log-concave function (though not necessarily a PMF - therefore, this is a much different definition than separable-log-concavity from Remark 2.1.1). We will compare our new defined PMF class with generalized log-concave PMF in later chapter.

1.4 Outline

Note that this thesis is divided into two parts, in this Part I, we focus on the maximum likelihood estimator of discrete log-concave distribution in higher dimensions.

In Chapter 2, we give a new definition of log-concave probability mass functions defined on $\mathbb{Z}^d$ (see Definition 2.1.1). We call this class extendible-log-concave, as it is closely related to extendible-convex functions (Murota and Shioura, 2001). We show that the new definition is equivalent to discrete log-concave distribution when $d = 1$. We introduce a Lemma which can be used to check if a function falls into our new defined class. We also show its unimodality and derive some properties of the new class of distribution.

Notably, We show that random variables from a continuous log-concave density can be grouped/binned (e.g. rounded to some accuracy level), the resulted discrete mass function will fall into our new defined distribution class under certain conditions (Proposition 2.3.1). We also show that under which condition the class of generalized log-concave is extendible-log-concave (Proposition 2.2.1). Moreover we prove there exist a unique extendible-log-concave PMF which minimize the distance to a given true PMF in term of KL divergence, and this minimizer is the true PMF itself if the true PMF is extendible-log-concave.

In Chapter 3, we show that the maximum likelihood estimator of our new defined class PMF exists and is unique. We show some attactive properties of the MLE of extendible-log-concave PMF. We discuss how to compute the MLE, and how to derive the objective functions. We compare the performance of our MLE with other parametric and nonparametric method through simulations. We developed two simulation scenarios with finite sample size. 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 behavior 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. Furthermore we give example to show that our estimator can be applied to "binned/grouped" continuous data set.

In Chapter 4, we introduce the detailed algorithm computing the MLE, the explicit

form of the objective function and its gradient. We proved that the objective function is convex but not differentiable everywhere. Hence Subgradient methodology is applied to compute the MLE. An R package is developed to make methodology widely available.

In Chapter 5, we prove the consistency of the MLE. If the true PMF is extendible-log-concave, then our MLE converges to the true PMF in term of KL divergence; if the true PMF is not extendible-log-concave, but is close to extendible-log-concave, our MLE still reveals desirable behavior.

Chapter 2

Introduction of discrete log-concave PMF

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 previous chapter, we choose to focus on the class of convex-extendible functions. There are two main reasons for this: It was shown in Murota and Shioura (2001, Theorem 4.1) 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 limits. We will show this property later in Theorem 2.4.1.

2.1 Definition of discrete log-concave PMF

Following the definition of convex-extendible function, naturally a function $h : \mathbb{Z}^d \to \mathbb{R} \cup \{-\infty\}$ is concave-extendible if $-h$ is convex-extendible.

Definition 2.1.1. *A PMF $p(z) : \mathbb{Z}^d \to [0,1]$ is **e-log-concave** (eLC) if $\log(p(z))$ 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 2.1.1 (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) by Murota (2009, Theorem 2.1). The maximum likelihood estimation considered here, when $d = 1$, have already 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 $p(z) = e^{\varphi(z)}$, where $-\varphi(z)$ is separable-convex. In such a situation, the multivariate MLE problem can be solved using the work of Balabdaoui et al. (2013). Recall that the active set algorithm is for one dimensional discrete log-concave MLE. We apply active set algorithm to compute each marginal distribution, then get the joint distribution by multiplication of marginal distributions. Due to the independence, $\widehat{p}(z) = \prod_{i=1}^{d} \widehat{p}_i(z_i)$, where $z = \{z_1, ..., z_d\} \in \mathbb{Z}^d$ and $\widehat{p}_i(z_i)$ is one dimensional log-concave MLE imputed by active set algorithm. Note that*

$$\widehat{\varphi}(z) = \log\{p(z)\} = \log \prod_{i=1}^{d} \widehat{p}_i(z_i) = \sum_{i=1}^{d} \widehat{\varphi}_i(z_i),$$

where $\widehat{\varphi}_i = \log \widehat{p}_i$ is discrete concave (or eLC). Hence $-\widehat{\varphi}$ is separable-convex. We call $\widehat{p}$ as separable log-concave estimation.

Remark 2.1.2 (Checking the class eLC)**.** *The following result gives one simple way to verify if a discrete function is convex-extendible.*

Lemma 2.1.1. *Murota and Shioura (2001, Lemma 2.3) Let $h : \mathbb{Z}^d \to \mathbb{R} \cup \{+\infty\}$ be 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 "obvious" convex extension of $h(z)$ is $h^R(x) = x^T Ax$ for $x \in \mathbb{R}^d$. Note that by Rockafellar (1970, Theorem 4.5, page 27) $h^R(x)$ is convex. The function is closed because it is continuous. By Murota and Shioura (2001, Lemma 2.3) $h(z)$ is therefore, convex-extendible.

Remark 2.1.3 (Alternative lattice structures)**.** *In this work we limit ourselves 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 it here. This is particularly attractive in light of the relationship that our definition has with log-concave densities, see Proposition 2.3.1.*

Remark 2.1.4 (Unimodality)**.** *Several notions of unimodality exist for both 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

$$p(z) = \exp\{-h(z)\} = \exp\{-h^R(z)\},$$

where h^R is a convex extension of $h(z)$ defined not only on $\mathbb{Z}^d$ but also on $\mathbb{R}^d$.

2.2 Relationship with generalized log-concave PMF

Clearly, the definition of generalized log-concave needs 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. In fact, the following relationship holds.

Proposition 2.2.1. *Suppose that p is generalized log-concave with support $\mathcal{S}$. If $\mathcal{S}$ is convex-extendible, then $p \in \mathcal{P}_0$.*

Proof. By definition, for any $z \in \mathcal{S}$, we have that

$$h(z) \;=\; -\log p(z) \;=\; \sum_{i=1}^{d} \{-\log p_i(z_i)\} \;=\; \sum_{i=1}^{d} h_i(z_i),$$

where each $p_i(z_i)$ is discrete log-concave on $\mathbb{Z}$, and hence $h_i(z_i)$ is discrete convex. Note that the definition of generalized log-concavity does not specify the support $\mathcal{S}$, which means that $\mathcal{S}$ can be any form. Hence we make the assumption that $\mathcal{S}$ is convex-extendible to make it work for convex-extendible. For each i, let $\mathcal{S}_i = \{k \in \mathbb{Z} : (k)_i \in \mathcal{S}\}$, where $(k)_i$

denotes any point of $\mathbb{Z}^d$ with its ith element equal to k. Each function h_i is defined on $\mathcal{S}_i$. Hence $h(z) = \sum_{i=1}^{d} h_i(z_i)$ is separable-convex on $\mathcal{S}$. Therefore $h(z)$ is convex-extendible by Murota and Shioura (2001), which implies $p \in \mathcal{P}_0$. □

Based on the Proposition 2.2.1 and the work from Bapat (1988); Johnson et al. (1997), we easily 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 set. Hence Proposition 2.2.1 provides another approach to checking if a given probability mass function falls in the class $\mathcal{P}_0$.

2.3 Relationship with continuous log-concave distributions

In the following proposition, we show the relation between our discrete log-concave distribution and continuous log-concave distribution.

Proposition 2.3.1. *Suppose that f is a log-concave density on $\mathbb{R}^d$, and let $A = [-1/2, 1/2)^d$. Define the probability mass function $p(z) = \int_{z+A} f(y)dy$. Suppose that the support of p is convex-extendible. Then $p \in \mathcal{P}_0$.*

Proof of Proposition 2.3.1. Let f denote a log-concave density on $\mathbb{R}^d$. For $A = [-1/2, 1/2)^d$, consider the function $q(x) = \int_{x+A} f(y)dy = P(Y \in A + x)$, letting Y denote the random

variable with density f. Then, by the property of log-concave distributions (see e.g. Dharmadhikari and Joag-Dev (1988, (2.6) on page 47)), for any $\alpha \in (0,1)$ and any $x, y \in \mathbb{R}^d$ we have that

$$q(\alpha x + (1-\alpha)y) \quad \geq \quad q(x)^{\alpha} q(y)^{1-\alpha}$$

which implies that the function $h^R(x) = -\log q(x)$ is convex. The function $q(x)$ is continuous by properties of integrals (applying, for example, the dominated convergence theorem and the fact that f must be bounded). In fact, letting B denote an upper bound on f, we have that

$$|q(x) - q(y)| \quad \leq \quad B\lambda\{(A+x)\Delta(A+y)\} \quad \leq \quad 4Bd\lambda\{A\}||x-y||_{\infty},$$

where $\lambda\{A\}$ denotes the Lebesgue measure of the set A, and Δ denotes set difference symbol. It follows that $h^R(x) = -\log q(x)$ is continuous on its effective domain, and therefore it is lower semi-continuous. Therefore, it is closed (Rockafellar, 1970, Theorem 7.1, page 51) on its effective domain. Lastly, $h^R(z) = -\log q(x) = -\log p(z)$ by definition on $\mathbb{Z}^d$. It follows that the restriction of h^R to $\bar{S}$ is a closed convex extension of $-\log p(z)$, and hence $p \in \mathcal{P}_0$ by Murota and Shioura (2001, Lemma 2.3) (Lemma 2.1.1). □

A quick look at the proof reveals that our result is not tied to the lattice $\mathbb{Z}^d$ nor our particular choice of A. Letting Y denote a random variable with density f as above.

Then the PMF p with $A = [-1/2, 1/2)^d$ corresponds to the probability mass function of the random variable $X = \lfloor Y + 0.5 \rfloor$ (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 random variable lives on the lattice $\delta\mathbb{Z}^d$). This means that the class $\mathcal{P}_0$ can be used to analyze log-concave random variables which have been discretized or "grouped/binned". An example is given in Section 3.3.

2.4 Properties

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

Proposition 2.4.1. *Suppose that $p \in \mathcal{P}_0$.*

1. *The support of p, $\mathcal{S} = \{z \mid p(z) > 0\}$, is a convex-extendible set.*

   Proof. Let $h(z) = -\log p(z)$, then $h(z)$ is convex-extendible by assumption, and $\mathcal{S} = \{z \mid h(z) < \infty\}$. Hence, by Lemma 2.1.1 (Murota and Shioura, 2001), there exists a convex extension $h^R(x)$ of $h(z)$, which is a closed convex function on $\mathbb{R}^d$. Therefore, the effective domain of h^R, $\{x \mid h^R(x) < +\infty\}$, is a closed convex set in $\mathbb{R}^d$ (Rockafellar, 1970, page 23 and Theorem 7.1 on page 51). The latter follows since for a closed function, its epigraph must be closed (Rockafellar, 1970, Theorem 7.1 on page 51) and the effective domain is the projection of the epigraph onto $\mathbb{R}^d$, (Rockafellar, 1970, page 23). Since such a projection of a closed set must be closed (appealing to the characterization of closed sets via Cauchy sequences), it follows that the effective

domain is closed. Therefore, $\mathcal{S} \subset \bar{\mathcal{S}} \subseteq \{x \,|\, h^R(x) < +\infty\}$. Furthermore, we have that $\mathcal{S} = \mathbb{Z}^d \cap \{x \,|\, h^R(x) < +\infty\}$. Therefore, it follows that $\bar{\mathcal{S}} \cap \mathbb{Z}^d = \mathcal{S}$, and hence $\mathcal{S}$ is convex extendible. □

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, $\widetilde{p} \in \mathcal{P}_0$.

Proof. Let $h(z) = -\log p(z)$, then $h(z)$ is convex-extendible by assumption. By Lemma 2.1.1 (Murota and Shioura, 2001), there exist a convex extension $h^R(x)$ of $h(z)$, which is a closed convex function on $\mathbb{R}^d$. We define a function

$$\widetilde{h}^R(x) = \begin{cases} h^R(x) - \log c, & x \in \bar{\mathcal{A}} \\ +\infty, & x \notin \bar{\mathcal{A}}, \end{cases}$$

for $c = \frac{1}{\sum_{z \in \mathcal{A}} p(z)}$, and where $\bar{\mathcal{A}}$ denotes the convex closure of $\mathcal{A}$. It is obvious that $\widetilde{h}^R$ is also a closed convex function. Also,

$$-\log \widetilde{p}(z) = -\log(cp(z)) = h(z) - \log c = h^R(z) - \log c = \widetilde{h}^R(z),$$

for $z \in \mathcal{A} \subset \operatorname{conv} \mathcal{A} \subset \bar{\mathcal{A}}$, and hence $\widetilde{h}^R$ is a convex extension of $-\log \widetilde{p}$. Therefore

$\widetilde{p} \in \mathcal{P}_0$. □

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* $\mathcal{S} = \mathcal{S}_1 \times \mathcal{S}_2 \subset \mathbb{Z}^{d_1+d_2}$ *also satisfies* $p \in \mathcal{P}_0$.

Proof. Letting $h_1(z_1) = -\log p_1(z_1), h_2(z_2) = -\log p_2(z_2)$ then $h_1(z), h_2(z)$ are both convex-extendible by assumption. Let $x_1 \in \mathbb{R}^{d_1}, x_2 \in \mathbb{R}^{d_2}$. By Lemma 2.1.1 (Murota and Shioura, 2001), there exist convex extensions $h_1^R(x_1), h_2^R(x_2)$ respectively, of $h_1(z_1), h_2(z_2)$. These are closed convex functions on $\mathbb{R}^{d_1}, \mathbb{R}^{d_2}$. Next, $h^R(x_1, x_2) = h_1^R(x_1) + h_2^R(x_2)$ is also a convex function on $\mathbb{R}^{d_1+d_2}$ (see below proof *). Furthermore, it is closed, since it is the sum of lower semi-continuous functions, and hence lower semi-continuous (Rockafellar, 1970, Theorem 7.1, page 51). Finally,

$$h^R(z) = h_1^R(z) + h_2^R(z) = h_1(z) + h_2(z) = -\log(p_1(z_1)p_2(z_2)) = -\log p(z),$$

where $z = (z_1, z_2)$. Therefore $p \in \mathcal{P}_0$.

* We now show that $h^R(x_1, x_2) = h_1^R(x_1) + h_2^R(x_2)$ is also convex function on $\mathbb{R}^{d_1+d_2}$.

Let $x^{'}, x^{''} =\in \mathbb{R}^{d_1+d_2}, x_1^{'}, x_1^{''} \in \mathbb{R}^{d_1}, x_2^{'}, x_2^{''} \in \mathbb{R}^{d_2}$ and $x^{'} = (x_1^{'}, x_2^{'}), x^{''} = (x_1^{''}, x_2^{''})$.

$$
\begin{aligned}
h^R(\alpha x^{'} + (1-\alpha)x^{''}) &= h^R(\alpha x_1^{'} + (1-\alpha)x_1^{''}, \alpha x_2^{'} + (1-\alpha)x_2^{''}) \\
&= h_1^R(\alpha x_1^{'} + (1-\alpha)x_1^{''}) + h_2^R(\alpha x_2^{'} + (1-\alpha)x_2^{''}) \\
&< \alpha h_1^R(x_1^{'}) + (1-\alpha)h_1^R(x_1^{''}) + \alpha h_2^R(x_2^{'}) + (1-\alpha)h_2^R(x_2^{''}) \\
&= \alpha(h_1^R(x_1^{'}) + h_2^R(x_2^{'}) + (1-\alpha)(h_1^R(x_1^{''}) + h_2^R(x_2^{''})) \\
&= \alpha h^R(x_1^{'}, x_2^{'}) + (1-\alpha)(h^R(x_1^{''}, x_2^{''})) \\
&= \alpha h^R(x^{'}) + (1-\alpha)h^R(x^{''})
\end{aligned}
$$

□

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 $p(z_1|z_2) = p((z_1, z_2))/p(z_2) \in \mathcal{P}_0$.*

Proof. Let $h(z) = -\log p(z)$, then $h(z)$ is convex-extendible by assumption and fix $z_2 \in \mathbb{Z}^{d_2}$. By Lemma 2.1.1, there exists a convex extension $h^R(x)$ of $h(z)$, which is a closed convex function on $\mathbb{R}^d$. Let p_2 denote the marginal of p : $p_2(z_2) = \sum_{z_1 \in \mathbb{Z}^{d_1}} p(z_1, z_2)$. We then define $\widetilde{h}^R(x_1) = h^R(x_1, x_2 = z_2) + \log p_2(z_2)$, where $x_1 \in \mathbb{R}^{d_1}$, and $z_2 \in \mathbb{Z}^{d_2} \subset \mathbb{R}^{d_2}$ is fixed. We will show that $\widetilde{h}^R$ is the convex extension of $-\log p(z_1|z_2)$, and therefore $p(z_1|z_2)$ is eLC.

Firstly, we have that for any $z_1 \in \mathbb{Z}^{d_1}$

$$\begin{aligned}\widetilde{h}^R(z_1) &= h^R(z_1, z_2) + \log p_2(z_2) = -\log p(z_1, z_2) + \log p_2(z_2) \\ &= -\log p(z_1|z_2).\end{aligned}$$

Secondly, $\widetilde{h}^R(x_1)$ is convex since $h^R(x)$ is convex in x_1 and $\log p_2(z_2)$ is a constant. Finally, we need that $\widetilde{h}^R$ is closed. This follows from Rockafellar (1970, Theorem 7.1, page 51) by appealing to the definition of closed sets via Cauchy sequences.

Note that $h^R(x) = h^R(x_1, x_2)$ is closed implies the level set of $h^R(x_1, x_2)$ is closed, let denote the level set as $C = \{(x \in \mathbb{R}^d \,|\, h^R(x) \leq \alpha, \alpha < \infty\}$. By Krantz (1991, Proposition 5.5), for any Cauchy sequence $\{x\}_n$ inside C, where $x_i = \{x_{i_1}, x_{i_2}, ..., x_{i_d}\}$, for $i = 1, 2, ..., n, ...$, it's limit $x_0 = \{x_{0_1}, x_{0_2}, ..., x_{0_d}\}$ is also an element of C.

For a fix $z_2 = \{z_{2_1}, ..., z_{2_{d_2}}\} \in \mathbb{Z}^{d_2}$, we firstly show that $h^R(x_1, z_2)$ is closed. For the same α, the level set of $h^R(x_1, z_2)$ can be expressed by $\tilde{C} = \{x_1 \in \mathbb{R}^{d_1} \,|\, h^R(x_1, z_2) \leq \alpha\}$. Note that chosen of z_2 may lead to a empty level set of $h^R(x_1, z_2)$, we then consider two cases:

(a) if $(x_1, z_2) \notin C$, then $\tilde{C} = \varnothing$, which is closed.

(b) if $(x_1, z_2) \in C$, then for each Cauchy sequence $\{\tilde{x}\}_n$ inside $\tilde{C}$, where $\tilde{x_i} = \{\tilde{x}_{i_1}, ..., \tilde{x}_{i_{d_1}}\}$, we can extend it to a d dimensional Cauchy sequence $\{x\}_n$ with $x_i = \{\tilde{x}_{i_1}, ..., \tilde{x}_{i_{d_1}}, z_{2_1}, ..., z_{2_{d_2}}\}, i = 1, 2, ..., n, ...$ Note that $\{x\}_n \in C$, hence its

limit, denoted by $x_0 = \{x_{0_1}, ..., x_{0_a}, z_{2_1} ..., z_{2_{d_2}}\} = (\tilde{x}_0, z_2)$, is inside C. We then have $h^R(\tilde{x}_0, z_2) \leq \alpha$. We conclude that $\tilde{x}_0 \in \tilde{C}$. By the process we construct $\tilde{x}_0$, it is obviously the limit of Cauchy sequence $\{\tilde{x}\}_n$. Hence by Krantz (1991, Proposition 5.5), $\tilde{C}$ is closed.

Therefore $h^R(x_1, z_2)$ is closed. It is obvious that $\widetilde{h}^R(x_1) = h^R(x_1, z_2) + \log p_2(z_2)$ is also closed since $\log p_2(z_2)$ is a constant.

Hence our constructed function $\widetilde{h}^R$ is a closed convex function, and is the convex extension of $-\log p(z_1|z_2)$. Hence $p(z_1|z_2)$ is also eLC.

□

5. *Let Z be a discrete random variable, with probability mass function $p \in \mathcal{P}_0$ with support $\mathcal{S}$. Consider the linear transformation $\widetilde{Z} = AZ + b$, where A is a $d \times d$ matrix and b is a vector of length d. Let $\widetilde{p}$ denote the PMF of $\widetilde{Z}$ with support $\widetilde{\mathcal{S}}$. If*

   (a) $\widetilde{\mathcal{S}}$ is a subset of $\mathbb{Z}^d$,

   (b) the matrix A is invertible,

   then $\widetilde{p} \in \mathcal{P}_0$.

   Proof. Firstly, we add the first condition because our work focus on $\mathbb{Z}^d$. We need the linear transformed random variable be also defined on $\mathbb{Z}^d$. But if we relax our

work to other the lattice structure (which is possible), this condition can be removed easily.

Let $h(z) = -\log p(z)$, then $h(z)$ is convex-extendible by assumption. Hence, by Lemma 2.1.1 (Murota and Shioura, 2001), there exists a convex extension $h^R(x)$ of $h(z)$, which is a closed convex function. Note that $\widetilde{p}(z) = p(A^{-1}(z-b))$ for any $z \in \widetilde{\mathcal{S}}$. We then construct $\widetilde{h}^R(x) = h^R(A^{-1}(x-b))$, for any $x \in \mathrm{conv}(\widetilde{\mathcal{S}})$, where $\widetilde{\mathcal{S}}$ denote the convex hull of $\mathcal{S}$ Clearly, $\widetilde{h}^R$ is also convex and closed. Moreover,

$$\widetilde{h}^R(z) = h^R(A^{-1}(z-b)) = h(A^{-1}(z-b)) = -\log p(A^{-1}(z-b)) = -\log \widetilde{p}(z),$$

for any $z \in \mathbb{Z}$. Hence $\tilde{h}^R$ is the convex extension of $\widetilde{p}$, and therefore $\widetilde{p} \in \mathcal{P}_0$.

□

The following Theorem shows that the class $\mathcal{P}_0$ is closed under limits under some assumptions.

Theorem 2.4.1. *Let $p_n (n = 1, 2, ...), p$ 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 $p \in \mathcal{P}_0$.*

Proof. Define $\mathcal{S}_0 = \{z \in \mathbb{Z}^d \mid p(z) > 0\}$ and assume (for the moment) that $\mathcal{S}_0 = \mathbb{Z}^d$. Define also $h_n(z) = -\log p_n(z)$, for each $n \geq 1$ and $h(z) = -\log p(z)$. By assumption, h_n is convex-extendible, and converges to h pointwise on S_0. To prove that p is eLC, we need to show

that h is convex-extendible. To do this, we will use Lemma 2.1.1 (Murota and Shioura, 2001), and find a closed convex extension of h.

By Lemma 2.1.1 (Murota and Shioura, 2001), there exists a closed convex extension of h_n, for each n. We denote this by $h_n^R : \mathbb{R}^d \to \mathbb{R} \cup \{+\infty\}$. By definition, h_n^R is a closed convex function, and $h_n^R(z) = h_n(z)$ for any $z \in \mathbb{Z}^d$.

Fix $K \in \mathbb{Z}^+$ to be a large, positive integer, and let $\mathcal{B}_K = \{x \in \mathbb{R}^d : \|x\|_\infty \leq K\}$, a closed (in $\mathbb{R}^d$) and bounded set. Since $p_n \to p$ for all $z \in \mathbb{Z}^d$, there exists an n_0 such that for all $n \geq n_0$, $p_n(z) > 0$, and hence $h_n(z) < \infty$ for all $z \in \mathcal{B}_K$.

Note that $\mathcal{B}_K$ is a subset of $\mathbb{R}^d$, and also the convex hull of $\mathcal{B}_K \cap \mathbb{Z}^d$ (in $\mathbb{R}^d$). Since each h_n^R is closed and convex, we can apply Theorem A.1.8 (Rockafellar, 1970) in the Appendix, and conclude that for each n,

$$\sup_{x \in \mathcal{B}_K} h_n^R(x) \leq \sup_{z \in \mathcal{B}_K \cap \mathbb{Z}^d} h_n^R(z) = \sup_{z \in \mathcal{B}_K \cap \mathbb{Z}^d} h_n(z).$$

Therefore,

$$\sup_{n \geq n_0} \sup_{x \in \mathcal{B}_K} h_n^R(x) \leq \sup_{n \geq n_0} \sup_{z \in \mathcal{B}_K \cap \mathbb{Z}^d} h_n(z) = M_{K,n_0}, \tag{2.1}$$

where M_{K,n_0} is finite because $S_0 = \mathbb{Z}^d$. Therefore the sequence $\{h_n^R(x)\}_{n \geq n_0}$ is finite and pointwise bounded (uniformly) for all $x \in \mathcal{B}_K$. The statement continues to hold on the relative interior of $\mathcal{B}_K$ (again, in $\mathbb{R}^d$), which we denote rl $\mathcal{B}_K$. By Rockafellar (1970, Theo-

rem 10.6, page 88), Theorem A.1.4 in the Appendix, we conclude that $h_n^R(x)$ is uniformly bounded and equi-Lipschitzian relative to, say, $\mathcal{B}_{K-1}$. By the Arzelà-Ascoli theorem, we conclude that h_n^R is compact and hence there is a subsequence of h_n^R that converges uniformly on $\mathcal{B}_{(K-1)}$. We denote this subsequence as $h_{n_K}^R$, and its limit as h^R.

We now argue that $h^R(x)$ is a convex extension of $h(z)$ on $\mathcal{B}_{(K-1)}$:

— Since h^R is the limit of a sequence of convex functions defined on $\mathcal{B}_{(K-1)}$, it follows that h^R is convex on $\mathcal{B}_{(K-1)}$.

— By definition, $h^R(z) = \lim_{n_K \to \infty} h_{n_K}^R(z) = \lim_{n_K \to \infty} h_{n_K}(z) = h(z)$, for any $z \in \mathcal{B}_{(K-1)}$.

— For any K, $h_{n_K}^R(x)$ is finite by inequality (2.1). We also know that it is continuous and uniformly converges to $h^R(x)$ on $\mathcal{B}_{(K-1)}$. Hence $h^R(x)$ is also finite, and continuous on $\mathcal{B}_{(K-1)}$ by Krantz (1991, Theorem 9.1, page 201), and therefore $h^R(x)$ is closed on $\mathcal{B}_{(K-1)}$ by the definition of continuous functions (Krantz, 1991, Theorem 6.9, page 142).

Hence we can conclude that $h^R(x)$ is a closed convex extension of $h(z)$ on $\mathcal{B}_{(K-1)}$. Therefore $h(z)$ is convex-extendible by Murota and Shioura (2001, Lemma 2.3). Recall that $h(z) = -\log p(z)$, we conclude that $p(z)$ is also eLC for $z \in \mathcal{B}_{(K-1)}$. Since the above conclusion is true for any $K \in \mathbb{Z}^+$, therefore $p(z)$ is eLC for $z \in \mathbb{Z}^d$.

Now we consider the situation that $\mathcal{S}_0 \subset \mathbb{Z}^d$. Note that p_n is eLC, hence the support $\mathcal{S}_n$ is convex-extendible set by Proposition 2.4.1 1. Therefore, there exists a convex closure $\bar{\mathcal{S}}_n$ of $\mathcal{S}_n$, which is closed and convex on $\mathbb{R}^d$. For large enough n_0, We may repeat the above proof, but considering $\mathcal{B}_K \cap \bar{\mathcal{S}}_{n_0}$ instead of $\mathcal{B}_K$ throughout. Note that $\mathcal{B}_K$ is closed and

convex by definition, and intersection of two closed convex sets is also closed and convex. The proof may now be repeated as above, and h will be convex-extendible on $\mathcal{S}_{n_0}$, which equals to $\mathcal{S}_0$ eventually.

□

Let $\|z\|_\infty$ denote maximum norm, $\|z\|_\infty = \max\{|z_1|, ..., |z_d|\}$.

Theorem 2.4.2. *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 $|\sum_{z\in\mathbb{Z}^d} p_0(z)\log p_0(z)| < \infty$. Suppose also that the convex hull of the support of p_0 is closed. Then, there exists a unique $\widehat{p}_0$, such that*

$$\widehat{p}_0 \;=\; \operatorname*{argmin}_{p\in\mathcal{P}_0} \rho_{KL}(p \parallel p_0). \tag{2.2}$$

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.

Before we proof this Theorem, we will give the proof of a Lemma.

Lemma 2.4.1. *Suppose $p_1, p_2 \in \mathcal{P}_0$. Then a PMF $p \propto (p_1 p_2)^\alpha$ for any $\alpha \in (0,1)$ also satisfies $p \in \mathcal{P}_0$.*

Proof. Let $h_1 = -\log p_1, h_2 = -\log p_2$ and $h = \alpha(h_1 + h_2) + c$ (defined on $\mathbb{Z}^d$) for some appropriate constant $c \in \mathbb{R}$. Let ${h_1}^R$ and ${h_2}^R$ denote the closed convex extensions of h_1, h_2

(respectively), which exist by assumption. Then $h^R = \alpha({h_1}^R + {h_2}^R) + c$ is closed, convex (see previous proof of *), and by definition satisfies

$$h^R(z) = \alpha({h_1}^R(z) + {h_2}^R(z)) + c = \alpha(h_1(z) + h_2(z)) + c = h(z) = -\log p(z)$$

on $\mathbb{Z}^d$. Therefore, $p \in \mathcal{P}_0$. □

We now proof the Theorem 2.4.2.

Proof of Theorem 2.4.2. We firstly prove the existence part of the theorem. Let $\mathcal{S}_0 = \{z \in \mathbb{Z}^d | p_0(z) > 0\}$ denote the support of p_0. Without loss of generality, we assume $\mathcal{S}_0 = \mathbb{Z}^d$. Let $\widetilde{q} \propto e^{-\|z\|_\infty}$, where $z \in \mathbb{Z}^d$, such that $\widetilde{q} \neq \widehat{p}_0$ (if $\widehat{p}_0 \propto e^{-\|z\|_\infty}$, then we can put $\widetilde{q} \propto e^{-0.5\|z\|_\infty}$ instead, say). Note that $-\log \widetilde{q}(z) = \|z\|_\infty$, and since all norms on $\mathbb{R}^d$ are closed convex functions, $\|x\|_\infty$, $x \in \mathbb{R}^d$ is a convex extension of $\|z\|_\infty$. Hence $\widetilde{q}$ is eLC by Murota and Shioura (2001, Lemma 2.3).

We can also show that $\rho_{KL}(\widetilde{q} \parallel p_0) < \infty$ under our assumption.

$$\begin{aligned} \rho_{KL}(\widetilde{q} \parallel p_0) &= \sum_{z \in \mathbb{Z}^d} p_0 \log p_0(z) - \sum_{z \in \mathbb{Z}^d} p_0 \log \widetilde{q}(z) \\ &= \sum_{z \in \mathbb{Z}^d} p_0 \log p_0(z) - \sum_{z \in \mathbb{Z}^d} \|z\|_\infty p_0(z) \quad < \quad \infty. \end{aligned}$$

Hence, $\inf_{q \in \mathcal{P}_0} \rho_{KL}(q \parallel p_0) < \infty$.

Therefore, there exists a sequence of eLC PMFs $\{q_n\}$, such that

$$\rho_{KL}(q_n \parallel p_0) \quad \to \quad \inf_{q \in \mathcal{P}_0} \rho_{KL}(q \parallel p_0).$$

Because $\inf_{q \in \mathcal{P}_0} \rho_{KL}(q \parallel p_0) < \rho_{KL}(\widetilde{q} \parallel p_0)$, there exists an $N > 0$, such that for all $n > N$, we have

$$\rho_{KL}(q_n \parallel p_0) \quad \leq \quad \rho_{KL}(\widetilde{q} \parallel p_0).$$

Because both $-\log q_n(z), -\log \widetilde{q}(z)$ are positive, hence,

$$\sup_{n>N} \sum_{z \in \mathbb{Z}^d} |-\log q_n(z)| \, p_0(z) \leq \sum_{z \in \mathbb{Z}^d} |-\log \widetilde{q}(z)| \, p_0(z) = \sum_{z \in \mathbb{Z}^d} \|z\|_\infty \, p_0(z) < \infty.$$

Let $M > 0$ and consider $\mathcal{S}_M = \{z : \|z\|_\infty \leq M\}$. Let $\alpha_M = \min_{z \in \mathcal{S}_M} p_0(z)$, and note that as $M \to \infty$, we have that $\alpha_M \to 0$, since p_0 is summable. It follows that

$$\begin{aligned}
\sup_{n>N} \sum_{z \in \mathcal{S}_M} |-\log q_n(z)| \quad &\leq \quad \left\{ \max_{z \in \mathcal{S}_M} \frac{1}{p_0(z)} \right\} \left\{ \sup_{n>N} \sum_{z \in \mathcal{S}_M} |-\log q_n(z)| \, p_0(z) \right\} \\
&\leq \quad \left\{ \max_{z \in \mathcal{S}_M} \frac{1}{p_0(z)} \right\} \left\{ \sup_{n>N} \sum_{z \in \mathbb{Z}^d} |-\log q_n(z)| \, p_0(z) \right\} \\
&\leq \quad \left\{ \max_{z \in \mathcal{S}_M} \frac{1}{p_0(z)} \right\} \left\{ \sum_{z \in \mathbb{Z}^d} \|z\|_\infty \, p_0(z) \right\} \quad = \quad \frac{B}{\alpha_M},
\end{aligned}$$

where $B = E_{p_0}[\|Z\|_\infty] < \infty$. Hence,

$$\sup_{n>N} \sup_{z \in \mathcal{S}_M} |-\log q_n(z)| \leq B/\alpha_M,$$

$$\exp\left\{-\sup_{n>N} \sup_{z \in \mathcal{S}_M} |-\log q_n(z)|\right\} \geq \exp\left\{-B/\alpha_M\right\}.$$

Hence

$$\inf_{n>N} \min_{x \in \mathcal{S}_M} q_n(z) \geq e^{-B/\alpha_M} = \delta_M.$$

Furthermore, we can find an integer $M_1 > M$ large enough so that

$$\sup_{n>N} \sup_{z \in \mathcal{S}_{M_1}^c} q_n(z) < \delta_M/2.$$

Therefore, we can find an envelope function $e^{l(z)}$, where $l(z) = -\alpha\|z\|_\infty + \beta$ with $\alpha, \beta \in \mathbb{R}^+$, such that $\sup_{n>N} q_n(z) \leq e^{l(z)}$.

Let X_n be a sequence of random vectors with PMF q_n. Since $e^{l(z)}$ is summable it follows that X_n is tight. Hence, there exists a convergent subsequence q_{n_l}, and a limit point q_0 (Rosenthal, 2006). As q_n is eLC, by Theorem 2.4.1 q_0 is also eLC.

By Fatou's lemma, we have

$$\rho_{KL}(q_0 \parallel p_0) = \sum_{z \in \mathbb{Z}^d} p_0 \log \frac{p_0}{q_0} \leq \liminf_{n_l} \sum_{z \in \mathbb{Z}^d} p_0 \log \frac{p_0}{q_{n_l}} = \liminf_{n_l} \rho_{KL}(q_{n_l} \parallel p_0).$$

Since $\rho_{KL}(q_{n_l} \parallel p_0) \to \inf_{q \in \mathcal{P}_0} \rho_{KL}(q \parallel p_0)$, we have $\rho_{KL}(q_0 \parallel p_0) \leq \inf_{q \in \mathcal{P}_0} \rho_{KL}(q \parallel p_0)$. That

is, a minimizer $\widehat{p}_0$ exists, and the proof of existence is done.

We now prove uniqueness. Let's assume that $\widehat{p}_1, \widehat{p}_2$ are both eLC and minimize $\rho_{KL}(\cdot \,\|\, p_0)$. Let $\widetilde{p} \propto (\widehat{p}_1 \widehat{p}_2)^{1/2}$ is a proper PMF. Note that by Lemma 2.4.1, $\widetilde{p}$ is also eLC. Now,

$$\begin{aligned}
\rho_{KL}(\widetilde{p} \,\|\, p_0) &= \sum p_0 \log \frac{p_0}{\widetilde{p}} \\
&= (1/2) \sum p_0 \log \frac{p_0}{\widehat{p}_1} + (1/2) \sum p_0 \log \frac{p_0}{\widehat{p}_2} + \log \sum (\widehat{p}_1 \widehat{p}_2)^{1/2} \\
&= \rho_{KL}(\widehat{p}_1 \,\|\, p_0) + \log \sum (\widehat{p}_1 \widehat{p}_2)^{1/2} \quad \leq \quad \rho_{KL}(\widehat{p}_1 \,\|\, p_0).
\end{aligned}$$

The last inequality follows that $\sum (\widehat{p}_1 \widehat{p}_2)^{1/2} \leq \sum \widehat{p}_1 \sum \widehat{p}_2 = 1$ by Cauchy-Schwarz. However, since $\rho_{KL}(\widetilde{p} \,\|\, p_0) \geq \rho_{KL}(\widehat{p}_1 \,\|\, p_0)$, we find that $\sum (\widehat{p}_1 \widehat{p}_2)^{1/2} = \sum \widehat{p}_1 \sum \widehat{p}_2 = 1$. Therefore $\widehat{p}_1 = \widehat{p}_2$, again by Cauchy-Schwarz. This completes the proof. □

The following lemma characterizes the support of $\widehat{p}_0$.

Lemma 2.4.2. *The support of the KL projection $\widehat{p}_0$ is the intersection of $\mathbb{Z}^d$ with the (closed) convex hull of $\mathcal{S}_0$, which is the support of p_0.*

Proof of Lemma 2.4.2. Let $\widehat{\mathcal{S}}_0$ denote the support of $\widehat{p}_0$. Let $\widetilde{\mathcal{S}}_0 = \mathrm{conv}(\mathcal{S}_0) \cap \mathbb{Z}^d$. Our goal is to show that $\widehat{\mathcal{S}}_0 = \widetilde{\mathcal{S}}_0$. Here, we denote the convex hull of $\mathcal{S}_0$ as $\mathrm{conv}(\mathcal{S}_0)$, and note that by assumption, this is closed.

Firstly, note that if $p_0(z_0) > 0$, then $\widehat{p}_0(z_0) > 0$ (we call this **fact one**). This follows directly from the form of the KL divergence, as PMFs with support strictly smaller than that of p_0 have an infinite KL divergence, and can therefore not act as minimizers. We

thus have that $\mathcal{S}_0 \subseteq \widehat{\mathcal{S}}_0$.

Next, consider there exists $z_0 \in \widetilde{\mathcal{S}}_0 = \text{conv}(\mathcal{S}_0) \cap \mathbb{Z}^d$, but $z_0 \notin \widehat{\mathcal{S}}_0$, that is $\widehat{p}_0(z_0) = 0$. Then by Carathéodory's Theorem (Rockafellar, 1970, Theorem 17.1 page 155) we can write $z_0 = \sum_{i=1}^{d+1} \lambda_i z_i$, where $\lambda_i > 0, \sum_{i=1}^{d+1} \lambda_i = 1$ and $z_i \in \mathcal{S}_0$ for each $i = 1, \ldots, d+1$. Since $\widehat{p}_0$ is eLC and therefore $\log \widehat{p}_0$ has a concave extension equal to $\log \widehat{p}_0$ on $\mathbb{Z}^d$. By the concave property of the concave extension, we find that

$$\log \widehat{p}_0(z_0) = \log \widehat{p}_0(\sum_{i=1}^{d+1} \lambda_i z_i) \geq \sum_{i=1}^{d+1} \lambda_i \log \widehat{p}_0(z_i).$$

But then $\widehat{p}_0(z_0) = 0$ implies that $\log \widehat{p}_0(z_0) = -\infty$ and hence $\widehat{p}_0(z_i) = 0$ for at least one $1 \leq i \leq d+1$. But $p_0(z_i) > 0$ because $z_i \in \mathcal{S}_0$. This is a direct contradiction with **fact one**. It follows that $\widetilde{\mathcal{S}}_0 \setminus \mathcal{S}_0 \subseteq \widehat{\mathcal{S}}_0$. Together with **fact one** ($\mathcal{S}_0 \subseteq \widehat{\mathcal{S}}_0$), this yields $\widetilde{\mathcal{S}}_0 \subseteq \widehat{\mathcal{S}}_0$.

Finally, consider a $z_0 \in \mathbb{Z}^d$ such that $z_0 \in \widehat{\mathcal{S}}_0$, that is $\widehat{p}_0(z_0) > 0$. But $z_0 \notin \widetilde{\mathcal{S}}_0$. Construct a PMF

$$\widetilde{p}(z) = \begin{cases} c\widehat{p}_0(z) & z \in \widetilde{\mathcal{S}}_0 \\ 0 & z \notin \widetilde{\mathcal{S}}_0 \end{cases}$$

where c denotes an appropriate normalizing constant. By Proposition 2.4.1, $\widetilde{p}$ is also eLC.

Also, note that $c > 1$ by assumption, hence $\widetilde{p}(z) > \widehat{p}_0(z)$. Then,

$$\begin{aligned}
\rho_{KL}(\widehat{p}_0 \parallel p_0) &= \sum_{z \in \mathbb{Z}^d} p_0(z) \log p_0(z) - \sum_{z \in \mathbb{Z}^d} p_0(z) \log \widehat{p}_0(z) \\
&= \rho_{KL}(\widetilde{p} \parallel p_0) + \sum_{z \in \mathbb{Z}^d} p_0(z) \log \widetilde{p}(z) - \sum_{z \in \mathbb{Z}^d} p_0(z) \log \widehat{p}_0(z) \\
&= \rho_{KL}(\widetilde{p} \parallel p_0) + \sum_{z \in \mathbb{Z}^d} p_0(z) \Big\{ \log \widetilde{p}(z) - \log \widehat{p}_0(z) \Big\} \\
&= \rho_{KL}(\widetilde{p} \parallel p_0) + \sum_{z \in S_0} p_0(z) \left\{ \log \widetilde{p}(z) - \log \widehat{p}_0(z) \right\} \\
&> \rho_{KL}(\widetilde{p} \parallel p_0).
\end{aligned}$$

Therefore, $\widehat{p}_0$ cannot minimize the KL divergence. Therefore, $\widehat{\mathcal{S}}_0 \subseteq \widetilde{\mathcal{S}}_0$.

Together with the previous conclusion $\widetilde{\mathcal{S}}_0 \subseteq \widehat{\mathcal{S}}_0$, we have $\widetilde{\mathcal{S}}_0 = \widehat{\mathcal{S}}_0$.

□

Chapter 3

Maximum likelihood estimation of eLC

Note that 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 result is thus a simple consequence of Theorem 2.4.2.

Proposition 3.0.1. *Suppose that $X_1, \ldots, X_n$ are independent and identically distributed random variables on $\mathbb{Z}^d$ with true PMF p_0. Then, with probability one, there exists a unique eLC maximum likelihood estimator. That is, 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 will use the notation $\widehat{p}_n$ to denote the MLE

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

Proof. Let $\bar{p}_n$ denotes empirical PMF. We replace $\widehat{p}_0$ with $\bar{p}_n$ in Equation (2.2), then we

can show that minimizing KL divergence between eLC PMF and the empirical distribution is equivalent to maximizing the likelihood function.

$$\widehat{p}_n = \operatorname{argmin}_{p \in \mathcal{P}_0} \rho_{KL}(p \parallel \bar{p}_n) = \operatorname{argmin}_{p \in \mathcal{P}_0} (- \sum_{z \in \mathbb{Z}^d} \bar{p}_n \log p + \sum_{z \in \mathbb{Z}^d} \bar{p}_n \log \bar{p}_n),$$

is equivalent to

$$\widehat{p}_n = \operatorname{argmin}_{p \in \mathcal{P}_0} (- \sum_{z \in \mathbb{Z}^d} \bar{p}_n \log p),$$

which is equivalent to

$$\widehat{p}_n = \operatorname{argmax}_{p \in \mathcal{P}_0} \frac{1}{n} \sum_{z \in \mathbb{Z}^d} \log p.$$

Therefore the existence and uniqueness of eLC MLE is a quick consequence of Theorem 2.4.2. □

Computation of this estimator is, unfortunately, not an easy problem. We again refer to Walther (2009) for a review. In $d = 1$, for example the active set algorithm tends 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 et al. (2010), and it is their approach which we adapt to the discrete setting in this work. This is described in more detail in the next chapter.

The following is another useful property of the eLC MLE, also known to hold in the continuous and discrete $d = 1$ cases.

Lemma 3.0.1. *Let $\overline{p}_n$ denote the empirical PMF of independent and identically distributed*

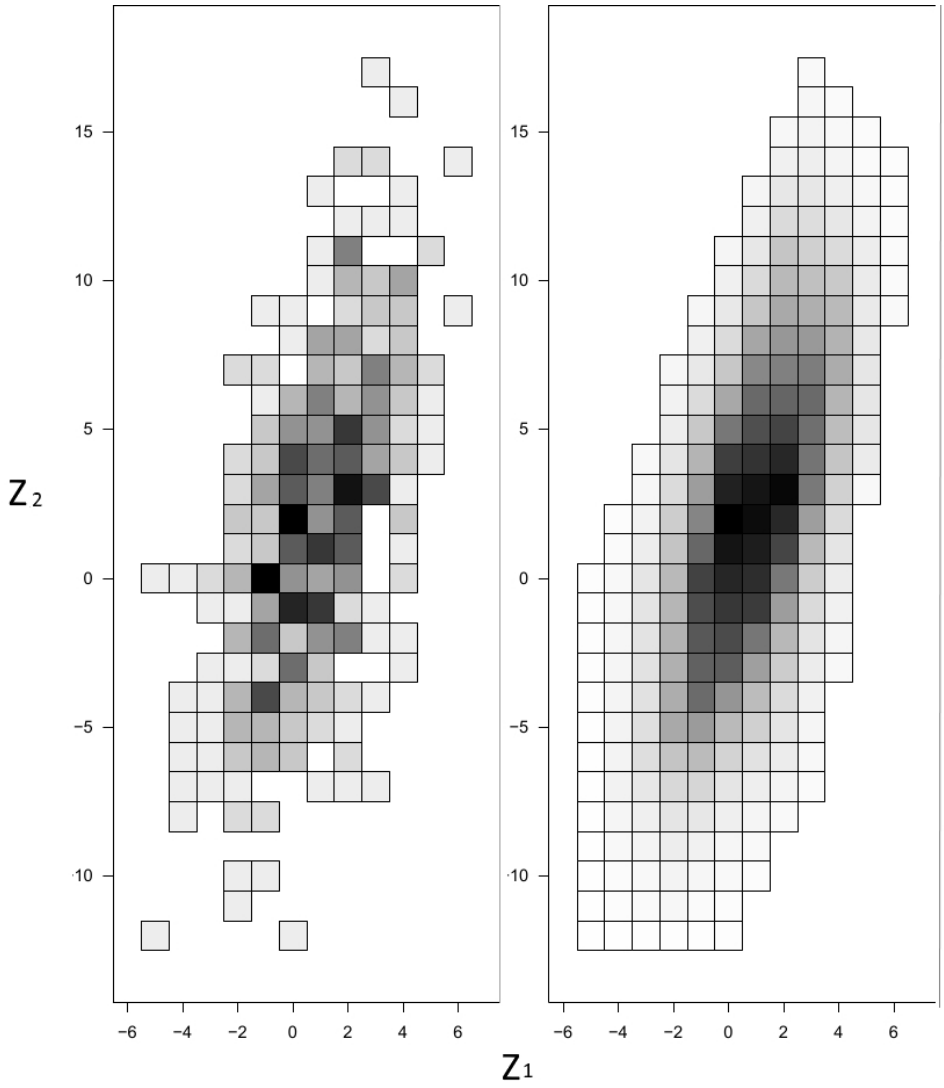


Figure 3.1: Grayscale heatmaps of the empirical PMF (left) and its eLC projection (right). The true distribution is a discrete Gaussian.

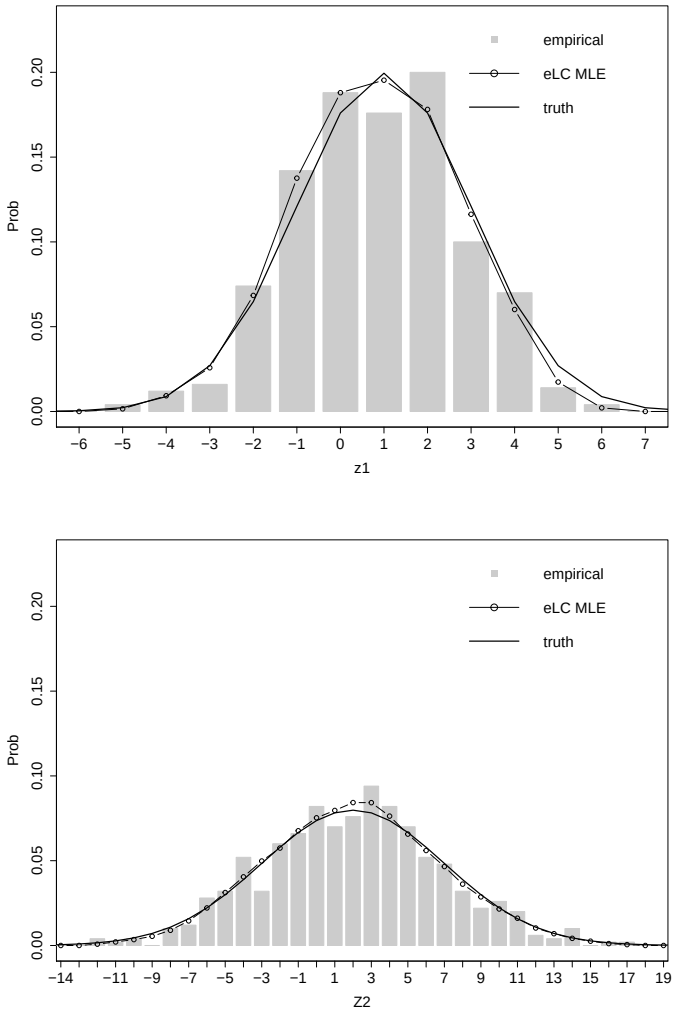


Figure 3.2: From the example of Figure 3.1, we compute the marginal distributions of our eLC MLE, we compare the marginals of elC MLE with empirical marginals and the true marginals in above Figures.

random variables $X_1, ..., X_n$ *on* $\mathbb{Z}^d$*, let* $h : \mathbb{Z}^d \mapsto \mathbb{R}$ *be any convex-extendible function, then*

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

Proof. First, note that $\widehat{p}_n$ is obtained by maximizing the following functional

$$\Phi(\varphi) = \sum_{i=1}^{n} \varphi(z_i)\bar{p}_n - \sum_{z \in \mathbb{Z}^d} \exp\{\varphi(z)\}$$

over all concave-extendible functions, where $\varphi(z) = \log p(z)$ (see Lemma 3.1.1). Letting $\widehat{\varphi}_n = \operatorname{argmax} \Phi(\varphi)$, we then have $\widehat{p}_n(z) = \exp\{\widehat{\varphi}_n(z)\}$.

Let $g(z) : \mathbb{Z}^d \mapsto \mathbb{R}$ be any concave-extendible function, and hence for any $\varepsilon > 0$, $\varphi + \varepsilon g$ is also concave-extendible (Murota and Shioura, 2001, Theorem 4). Therefore, $\Phi(\widehat{\varphi}_n + \varepsilon g) \leq$

$\Phi(\widehat{\varphi}_n)$ since $\widehat{\varphi}_n$ maximize Φ. This implies that

$$
\begin{aligned}
&\lim_{\varepsilon\to 0}\frac{\Phi(\widehat{\varphi}_n+\varepsilon g)-\Phi(\widehat{\varphi}_n)}{\varepsilon}\\
=&\lim_{\varepsilon\to 0}\frac{\sum_{i=1}^{n}\{\widehat{\varphi}_n(z_i)+\varepsilon g(z_i)\}\bar{p}_n-\sum_{z\in\mathbb{Z}^d}\exp\{\widehat{\varphi}_n(z)+\varepsilon g(z)\}-\sum_{i=1}^{n}\widehat{\varphi}_n(z_i)\bar{p}_n+\sum_{z\in\mathbb{Z}^d}\exp\{\widehat{\varphi}_n(z)\}}{\varepsilon}\\
=&\sum_{i=1}^{n}g(z_i)\bar{p}_n-\lim_{\varepsilon\to 0}\sum_{z\in\mathbb{Z}^d}\frac{\exp\{\widehat{\varphi}_n(z)+\varepsilon g(z)\}-\exp\{\widehat{\varphi}_n(z)\}}{\varepsilon}\\
=&\sum_{i=1}^{n}g(z_i)\bar{p}_n-\sum_{z\in\mathbb{Z}^d}g(z)\exp\{\widehat{\varphi}_n\}\\
=&\sum_{i=1}^{n}g(z_i)\bar{p}_n-\sum_{z\in\mathbb{Z}^d}g(z)\widehat{p}_n\\
=&\sum_{z\in\mathbb{Z}^d}g(z)\bar{p}_n-\sum_{z\in\mathbb{Z}^d}g(z)\widehat{p}_n \quad\leq\quad 0.
\end{aligned}
$$

Note that the third last equality comes from:

$$
\begin{aligned}
&\lim_{\varepsilon\to 0}\frac{\exp\{\widehat{\varphi}_n(z)+\varepsilon g(z)\}-\exp\{\widehat{\varphi}_n(z)\}}{\varepsilon}\\
=&\exp\{\widehat{\varphi}_n(z)\}\lim_{\varepsilon\to 0}\frac{\exp\{\varepsilon g(z)\}-1}{\varepsilon}\\
=&\exp\{\widehat{\varphi}_n(z)\}\lim_{\varepsilon\to 0}\frac{g(z)\exp\{\varepsilon g(z)\}}{1}\\
=&g(z)\exp\{\widehat{\varphi}_n(z)\}
\end{aligned}
$$

Similarly, for any convex-extendible function h, we have

$$
\sum_{z\in\mathbb{Z}^d}h(z)\widehat{p}_n \quad\leq\quad \sum_{z\in\mathbb{Z}^d}h(z)\bar{p}_n.
$$

□

In particular, this implies that the mean of the MLE is equal to the observed mean of the data, because affine functions $h(z) = z, z \in \mathbb{Z}^d$ are both convex-extendible and concave-extendible functions. Hence we have

$$E_{\widehat{p}_n}(z) = \sum_{z \in \mathbb{Z}^d} z\widehat{p}_n(z) \leq E_{\bar{p}_n}(z) = \sum_{z \in \mathbb{Z}^d} z\bar{p}_n(z),$$

$$E_{\widehat{p}_n}(z) = \sum_{z \in \mathbb{Z}^d} z\widehat{p}_n(z) \geq E_{\bar{p}_n}(z) = \sum_{z \in \mathbb{Z}^d} z\bar{p}_n(z).$$

Hence $E_{\widehat{p}_n}(z) = E_{\bar{p}_n}(z)$.

Furthermore, the following Lemma shows that the variance matrix under the eLC MLE is smaller than the variance matrix under empirical distribution, in the sense that $\overline{\Sigma}_n - \widehat{\Sigma}_n$ is positive semi-definite.

Lemma 3.0.2. *let $\widehat{\Sigma}_n$ denotes the variance matrix of random variable which follows the distribution $\widehat{p}_n$. Let $\overline{\Sigma}_n$ denotes the empirical variance matrix, which is the variance matrix of random variable following the empirical distribution $\overline{p}_n$. Then $\overline{\Sigma}_n - \widehat{\Sigma}_n$ is positive semi-definite.*

Proof. Let V be any non zero vector on $\mathbb{R}^d$, we define function $h(z) = V^T z z^T V, z \in \mathbb{Z}^d$. Note that $h(z)$ is convex-extendible, because $h(z) = (z^T V)^T (z^T V) = (z^T V)^2$, it has convex extension $h(x) = (x^T V)^2, x \in \mathbb{R}^d$, which is closed convex function on $\mathbb{R}^d$.

Hence

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

$$\sum_{z\in\mathbb{Z}^d} V^T z z^T V\widehat{p}_n - \sum_{z\in\mathbb{Z}^d} V^T z z^T V\bar{p}_n \le 0,$$

$$V^T\Big(\sum_{z\in\mathbb{Z}^d} z z^T\widehat{p}_n - \sum_{z\in\mathbb{Z}^d} z z^T\bar{p}_n\Big)V \le 0.$$

Therefore $\overline{\Sigma}_n - \widehat{\Sigma}_n$ is positive semi-definite. □

An example of the MLE is given in Figures 3.1 and 3.2. The data is an IID sample of size $n = 1000$ from the discrete Gaussian distribution given on later Section. Figure 3.1 shows the empirical distribution (left) and the fitted eLC (right) as a grey-scale heatmap. The marginal distributions are given in Figure 3.2, where the true marginals are also added.

3.1 Computation of the MLE

For shape constraint method, it is a standard trick to add a summation term to the log-likelihood function, such that the optimization problem can be relaxed to general functions set instead of density functions set. That is, maximizing $\sum_{i=1}^{n} \log p(X_i)$ over $p \in \mathcal{P}_0$ is equiv-

alent to minimizing

$$-\frac{1}{n}\sum_{i=1}^{n}\varphi(X_i) + \sum_{z\in\mathbb{Z}^d}\exp\{\varphi(z)\}, \tag{3.1}$$

over all concave-extendible functions φ. Complete proof please see below Lemma 3.1.1.

Note, however, that the values $X_1, \ldots, X_n$ are expected to have duplicates in our setting. Therefore, let $z_1, \ldots, z_m$ denote the unique observed values of $X_1, \ldots, X_n$.

Lemma 3.1.1. *When the criterion function*

$$\Phi(\varphi) \quad = \quad -\sum_{j=1}^{m} w_j\varphi(z_j) + \sum_{z\in\mathbb{Z}^d} e^{\varphi(z)}$$

is minimized over all concave extendible functions φ, the minimizer satisfies $\sum_{z\in\mathbb{Z}^d} e^{\varphi(z)} = 1$.

Proof. Consider any concave extendible φ_0 minimize $\Phi(\varphi)$, and $p_0 = \exp\{\varphi_0\}$ such that $\sum_{z\in\mathbb{Z}^d}\exp\{\varphi_0(z)\} = c \neq 1$. Let $\widetilde{\varphi}_0 = \varphi_0 - \log c$. Then $\sum_{z\in\mathbb{Z}^d}\exp\{\widetilde{\varphi}_0(z)\} = 1$, because

$$\begin{aligned}\sum_{z\in\mathbb{Z}^d}\exp\{\widetilde{\varphi}_0(z)\} &= \sum_{z\in\mathbb{Z}^d}\exp\Big\{\exp\{\varphi_0(z)\} - \log c\Big\}\\ &= \sum_{z\in\mathbb{Z}^d}\frac{\exp\{\varphi_0(z)\}}{c}\\ &= \frac{1}{c}\sum_{z\in\mathbb{Z}^d}\exp\{\varphi_0(z)\} = 1\end{aligned}$$

Now,

$$\begin{aligned}
\Phi(\widetilde{\varphi}_0) &= -\sum_{j=1}^{m} w_j \widetilde{\varphi}_0(z_j) + \sum_{z \in \mathbb{Z}^d} e^{\widetilde{\varphi}_0(z)} \\
&= -\sum_{j=1}^{m} w_j \big\{ \varphi_0(z_j) - \log c \big\} + 1 \\
&= \sum_{j=1}^{m} w_j \varphi_0(z_j) + \sum_{j=1}^{m} w_j \log c + 1 \\
&= \sum_{j=1}^{m} w_j \varphi_0(z_j) + \log c + 1 \\
&= \sum_{j=1}^{m} w_j \varphi_0(z_j) + \sum_{z \in \mathbb{Z}^d} e^{\varphi_0(z)} - \sum_{z \in \mathbb{Z}^d} e^{\varphi_0(z)} + \log c + 1 \\
&= \Phi(\varphi_0) - c + \log c + 1.
\end{aligned}$$

Since $\log c \leq c - 1$ for any $c > 0$, we get $\Phi(\widetilde{\varphi}_0) \leq \Phi(\varphi_0)$, which is a contradiction. □

Let $\widehat{\mathcal{S}}_n = \bar{\mathcal{S}}_n \cap \mathbb{Z}^d$, where $S_n = \{z_1, \ldots, z_m\}$. Also recall the empirical PMF $\overline{p}_n$. Using also the characterization of the MLE, we can further re-write the optimization problem above to be equivalent to minimizing

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

again, over all concave-extendible functions φ. We denote the minimizer of above function as $\widehat{\varphi}(z) = \log \widehat{p}_n$.

For a fixed vector of values $y \in \mathbb{R}^m$, following Cule et al. (2010), the tent function is

defined as

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

It turns out, that in the above we can exchange the function $\varphi(z)$ with the tent functions $t_y(z)$, and optimize over the vector $y \in \mathbb{R}^m$ instead. Lemma 3.1.2 shows a further simplification of the optimization problem .

Lemma 3.1.2. *Consider the function*

$$\tau(y_1, \ldots, y_m) \quad = \quad -\sum_{i=1}^{m} w_j \, t_y(z_j) + \sum_{z \in \widetilde{S}_n} \exp\{t_y(z)\}. \tag{3.2}$$

Then τ has a minimum over $y \in \mathbb{R}^m$. We denote the minimizer as $\widehat{y}$, and $\widehat{p}_n(z) = \exp\{t_{\widehat{y}}(z)\}$. Furthermore, $t_{\widehat{y}}$ is a concave extension of $\log \widehat{p}_n$.

Proof. Let $\widehat{\varphi}_n = \log \widehat{p}_n$, which minimizes $\Phi(\varphi)$. Let $\widehat{y}_i = \widehat{\varphi}_n(z_i)$, for $i = 1, \ldots, m$, and consider

$$t_{\widehat{y}}(x) \quad = \quad \inf\{g(x) : \mathbb{R}^d \mapsto \mathbb{R} \mid g \text{ is concave, and } g(z_i) \geq \widehat{y}_i \text{ for } i = 1, \ldots, m\}.$$

Let $\widehat{\varphi}_n^R : \mathbb{R}^d \mapsto \mathbb{R}$ denote the concave extension of $\widehat{\varphi}_n$, and note that $\widehat{\varphi}_n^R(z_i) = \widehat{\varphi}_n(z_i) = \widehat{y}_i, i = 1 \ldots, m$. Therefore $\widehat{\varphi}_n^R$ belongs to the set

$$\{g(x) : \mathbb{R}^d \to \mathbb{R} \mid g \text{ is concave, and } g(z_i) \geq \widehat{y}_i \text{ for } i = 1, \ldots, m\}.$$

As $t_{\widehat{y}}$ is the infimum of the above class of functions, we have $t_{\widehat{y}}(z) \le \widehat{\varphi}_n^R(z), z \in \mathbb{Z}^d$.

Assume that for some $z_0 \in \mathbb{Z}^d$, $\widehat{\varphi}_n(z_0) > t_{\widehat{y}}(z_0)$. Then $\sum_{z\in\mathbb{Z}^d} \exp \widehat{\varphi}_n(z) > \sum_{z\in\mathbb{Z}^d} \exp\{t_{\widehat{y}}(z)\}$. Also note that $t_{\widehat{y}}(z_i) \ge \widehat{y}_i, i = 1, ..., m$. Hence

$$\begin{aligned}
\Phi(\widehat{\varphi}_n) &= -\sum_{i=1}^{m} w_i\, \widehat{\varphi}_n(z_i) + \sum_{z\in\mathbb{Z}^d} \exp \widehat{\varphi}_n(z) \\
&= -\sum_{i=1}^{m} w_i \widehat{y}_i + \sum_{z\in\mathbb{Z}^d} \exp \widehat{\varphi}_n(z) \\
&> -\sum_{i=1}^{m} w_i\, t_{\widehat{y}}(z_i) + \sum_{z\in\mathbb{Z}^d} \exp\{t_{\widehat{y}}(z)\} \\
&= \Phi(t_{\widehat{y}}).
\end{aligned}$$

However, this creates a contradiction since $\widehat{\varphi}_n$ minimizes Φ. Therefore, $\widehat{\varphi}_n(z) = t_{\widehat{y}}(z)$, for any $z \in \mathbb{Z}^d$. This also implies that $t_{\widehat{y}}$ is a concave extension of $\log \widehat{p}_n$. □

While the above τ function has extended our optimization problem over $y \in \mathbb{R}^m$. It is still not simple enough for computation purpose. We show the objective function in the following Theorem 3.1.1 is convex. Because of convexity, the objective function has a unique minimizer. This objective function is the one we will work on in the sequel.

Theorem 3.1.1. *Consider the function*

$$\sigma(y_1, \ldots, y_m) = -\sum_{j=1}^{m} \overline{p}_n(z_j)\, y_j + \sum_{z\in\widehat{S}_n} \exp\{t_y(z)\}. \tag{3.3}$$

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

Proof of Theorem 3.1.1. We first prove that σ is convex. For $u, v \in \mathbb{R}^m$, $\lambda \in (0,1)$, we have

$$
\begin{aligned}
&\lambda t_u(x) + (1-\lambda) t_v(x) \\
&= \lambda \inf\{g_1(x) |\ g_1 \text{ concave, and } g_1(z_i) \geq u_i, i = 1, \ldots, m\} \\
&\quad + (1-\lambda) \inf\{g_2(x) |\ g_2 \text{ concave, and } g_2(z_i) \geq v_i, i = 1, \ldots, m\} \\
&= \inf\{\lambda g_1(x) : \mathbb{R}^d \to \mathbb{R} \mid g_1 \text{ is concave, and } g_1(z_i) \geq u_i, i = 1, 2, ..., n\} \\
&\quad + \inf\{(1-\lambda) g_2(x) : \mathbb{R}^d \to \mathbb{R} \mid g_2 \text{ is concave, and } g_2(z_i) \geq v_i, i = 1, 2, ..., n\} \\
&= \inf\{g_1(x) |\ g_1 \text{ concave, and } g_1(z_i) \geq \lambda u_i, i = 1, \ldots, m\} \\
&\quad + \inf\{g_2(x) |\ g_2 \text{ concave, and } g_2(z_i) \geq (1-\lambda) v_i, i = 1, \ldots, m\} \\
&\geq \inf\{g_1(x) + g_2(x) |\ g_1, g_2 \text{ are concave, } g_1(z_i) \geq \lambda u_i,\ g_2(z_i) \geq (1-\lambda) v_i, i = 1, \ldots, m\}.
\end{aligned}
$$

We also have

$$
\begin{aligned}
&\bar{h}_{\lambda u + (1-\lambda) v}(x) \\
&= \inf\{h(x) : \mathbb{R}^d \to \mathbb{R} \mid h \text{ is concave, } h(z_i) \geq \lambda u_i + (1-\lambda) v_i, i = 1, 2, ..., n\}.
\end{aligned}
$$

Since $\{g_1(x) + g_2(x) |\ g_1, g_2 \text{ concave, } g_1(z_i) \geq \lambda u_i,\ g_2(z_i) \geq (1-\lambda) v_i, i = 1, \ldots, m\}$ is a subset

of $\{g(x)|\ h \text{ concave}, g(z_i) \geq \lambda u_i + (1-\lambda)v_i, i = 1, \ldots, m\}$, we have

$$\lambda t_u(x) + (1-\lambda)t_v(x) \geq t_{\lambda u+(1-\lambda)v}(x), x \in \mathbb{R}^d.$$

Finally, by convexity of e^x (apply to the 2nd inequality below),

$$\begin{aligned}
& \sigma(\lambda u + (1-\lambda)v) \\
= & -\sum_{i=1}^{m} w_i\{\lambda u_i + (1-\lambda)v_i\} + \sum_{z\in\mathbb{Z}^d} \exp\left\{t_{\lambda u+(1-\lambda)v}(z)\right\} \\
\leq & -\sum_{i=1}^{m} w_i\{\lambda u_i + (1-\lambda)v_i\} + \sum_{z\in\mathbb{Z}^d} \exp\left\{\lambda t_u(z) + (1-\lambda)t_v(z)\right\} \\
\leq & -\sum_{i=1}^{m} w_i\{\lambda u_i + (1-\lambda)v_i\} + \sum_{z\in\mathbb{Z}^d} \left\{\lambda e^{t_u(z)} + (1-\lambda)e^{t_v(z)}\right\} \\
= & -\sum_{i=1}^{m} w_i\{\lambda u_i + (1-\lambda)v_i\} + \lambda \sum_{z\in\mathbb{Z}^d} e^{t_u(z)} + (1-\lambda) \sum_{z\in\mathbb{Z}^d} e^{t_v(z)} \\
= & \lambda\sigma(u) + (1-\sigma)\sigma(v).
\end{aligned}$$

Hence, $\sigma(y)$ is convex.

Next, for any $y \in \mathbb{R}^m$,

$$\sigma(y) \quad = \quad \tau(y) + \sum_{i=1}^{m} w_i \left(t_y(z_i) - y_i\right) \quad \geq \quad \tau(y),$$

by definition of the tent function t_y, the 2nd term $\sum_{i=1}^{m} w_i \left(t_y(z_i) - y_i\right)$ is always positive. Hence $\sigma(y)$ get its minimum value when $\sum_{i=1}^{m} w_i \left(t_y(z_i) - y_i\right) = 0$. It implies $\widehat{y}_i = t_{\widehat{y}}(z_i), i = 1, \ldots, m$, which minimize both $\sigma(y)$ and $\tau(y)$. Note that multiple $y \in \mathbb{R}^m$ may lead to same

tent function t_y, but $\widehat{y}_i = t_{\widehat{y}}(z_i)$ is unique since it makes poles of the tent. Furthermore, recall from Lemma 3.1.2 Therefore, $t_{\widehat{y}}(z) = \widehat{p}_n(z)$. □

We have introduced three objective functions (3.1), (3.2), and (3.3). We showed that minimizing all these three objective functions are equivalent to maximizing the likelihood function. When we maximize likelihood function, the optimization problem is over elC PMFs set. In objective function (3.1), we added a sum term such that the optimization problem is relaxed to concave-extendible functions set. In objective function (3.2), we replace φ with the tent function, such that the optimization problem is relaxed to m dimensional real number set. In objective function(3.3), we replace the tent function with y_j (the "poles" of the tent function) in the first term, to make the optimization problem is easier to compute.

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 Chapter 4, and we refer to Cule (2009); Cule et al. (2010) for the original development of these methods.

Remark 3.1.1. *It also turns out (see Lemma 3.1.2), that the function $t_{\widehat{y}}$ is a concave extension of* $\log \widehat{p}_n$. *Thus, the algorithm finds not only $\widehat{p}_n$, but the associated concave extension.*

Lastly, we note that a faster algorithm for log-concave density/PMF estimation in dimension greater than one remains an open problem in the field (see Cule et al. (2010,

Section 3) and Walther (2009, Section 5)). Although our algorithm in no way improves on the one proposed in Cule et al. (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 performance

We investigate the finite sample performance of the proposed method via simulations for $d = 2$. We consider two different scenarios for the true p_0.

- For scenario (A), we assume 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 assume that p_0 is discretely Gaussian, in that $p_0(z) \propto \exp\{-0.5(z-\mu)^T\Sigma^{-1}(z-\mu)\}$, where $\mu = (1, 2)$, and

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

  Considering the closed and 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, page 27, Lemma 2.1.1) we conclude that $p_0 \in \mathcal{P}_0$.

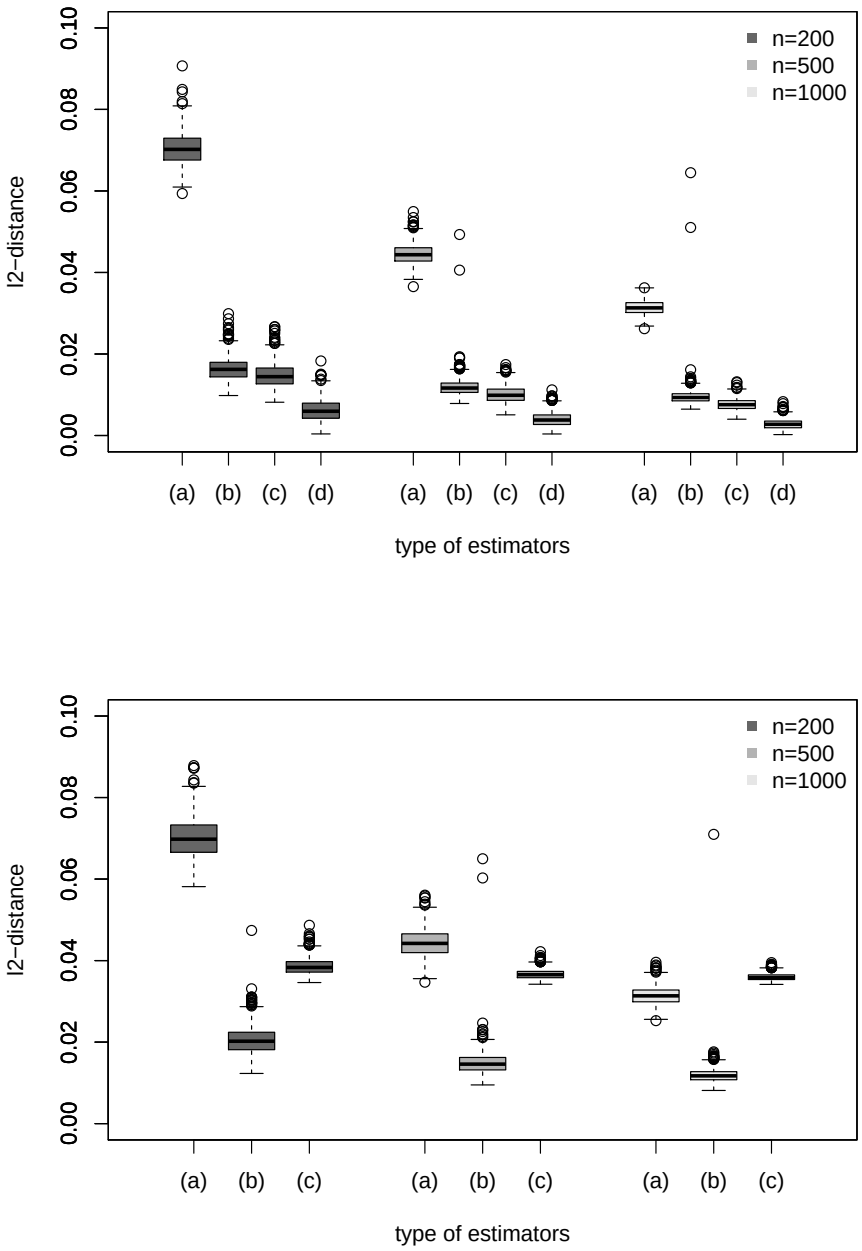


Figure 3.3: Boxplots of l_2 distance between estimator and true distribution.

For both scenarios, we simulated independent and identically distributed samples with samples sizes $n = 200, 500$, and 1000. The results of our simulations are shown in Figure 3.3. Top figure is when the true distribution is Poisson and negative binomial product, the bottom figure is when the true distribution is discrete Gaussian. The estimators are (a) empirical MLE, (b) eLC MLE, (c) separable log-concave MLE, and (d) parametric MLE (top plot only). The y axis is the l_2 distance. Each boxplot is the result of $B = 1000$ 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 separable log-concave 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 2.1.1) and the correct parametric MLE, where the latter is calculated for scenario (A) only. Clearly, the more (correct) assumptions we make, the more we increase efficiency without increasing bias - this is seen in the top plot. However, in the bottom plot, the incorrect MLE has poor performance, our proposed MLE works well when the marginal distributions are dependent to each other. 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.

3.3 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 choose to work with the last two variables: LSTAT (percentage lower status of the population) and MEDV (median value of owner-occupied homes in \$1000s). 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 remove observations with missing values (only an issue for MEDV) for a sample size of $n = 452$. We bin the data as described in Section 2.3, using the formula $x_i = \lfloor y_i + 0.5 \rfloor$ for each observation and for both variables. This creates $m = 270$ unique bins. Figure 3.4 shows the result of fitting the e-LC MLE to the binned data, along with the original histogram. This example has relative large sample size ($m = 270$ unique bins). Note that the larger the sample size, the better the quality of the estimation. But exact rule for large enough sample size is difficult to determine. Roughly, Figure 3.4 shows there is negative relation between MEDV and LSTAT. When the LSTAT increases, the MEDV decreases accordingly. That is when there are more lower status people in the community, people more likely to buy a house with affordable price, hence the average price of the owner-occupied house decreases.

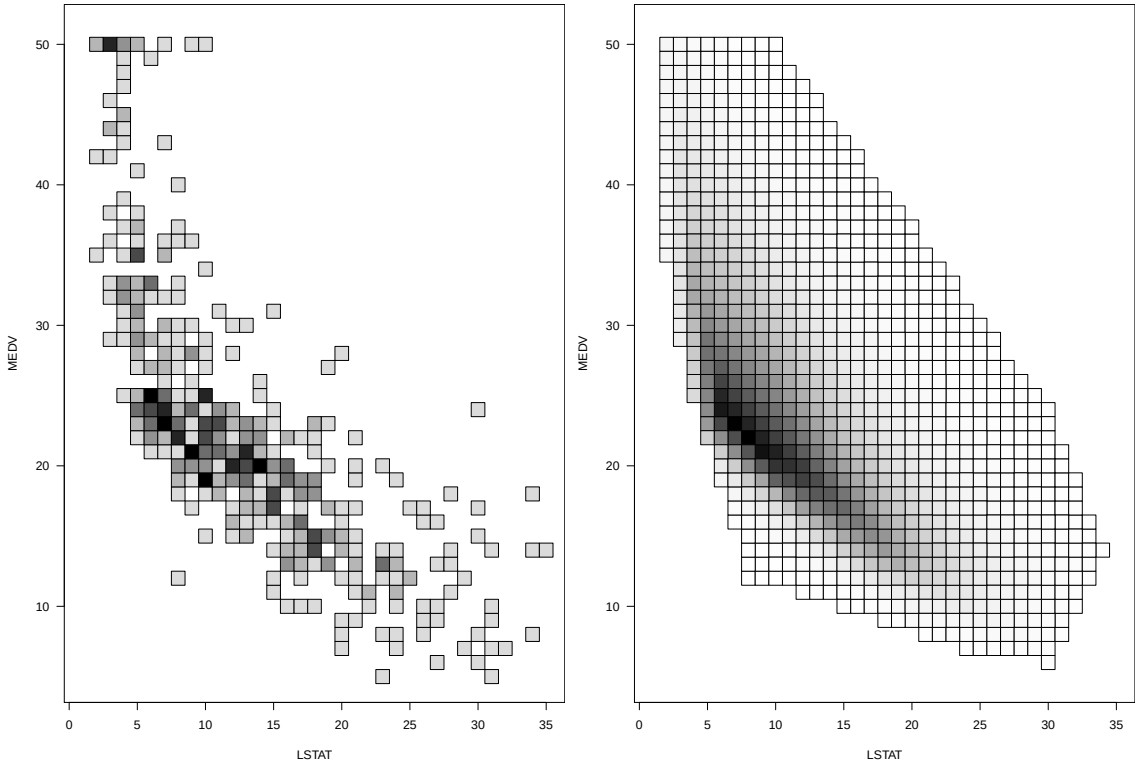


Figure 3.4: Boston Housing Data: original empirical distribution (left) along with eLC maximum likelihood estimate (right).

3.4 Mixtures and the EM algorithm

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

Chapter 4

Computing algorithm

4.1 Derive the explicit form of $\sigma(y)$

We begin by deriving an explicit formula for $\sigma(y)$. To this end, a few definitions are necessary. For our observations $z_1, \ldots, z_m$ and $y \in \mathbb{R}^m$, consider the set $\mathscr{Z} = \{(z_1, y_1), \ldots, (z_m, y_m)\}$. The convex hull of the set $\mathscr{Z} \subset \mathbb{R}^{d+1}$ is made up of the upper hull and the lower hull. Projecting the upper hull on the first dimensional subspace $\mathbb{Z}^d$, the facets of the upper hull create a subdivision of the points $z_1, \ldots, z_m$. We denote the subdivision as $\mathscr{S}(y)$, to emphasize its dependence on the vector y. This notion is best illustrated with examples. Consider the observations $\{z_i\}_{i=1}^4 = \{(0,0), (2,0), (3,1), (1,1)\}$ with $m = 4$. We compute, for three different y vectors, the associated $\mathscr{S}(y)$:

a. Let $y = (1, 1.9, 2, 1)$, then $\mathscr{Z} = \{(0,0,1), (2,0,1.9), (3,1,2), (1,1,1)\}$. The upper hall of $\mathscr{Z}$ consists of two facets with vertices $\{(0,0,1), (2,0,1.9), (3,1,2)\}$ and $\{(0,0,1), (1,1,1), (3,1,2)\}$.

Then we project of the upper hull to $\mathbb{Z}^2$, the resulted $\mathscr{S}(y)$ has two subdivisions: $\{(0,0),(2,0),(3,1)\}$, $\{(0,0),(1,1),(3,1)\}$. See Figure 4.1 left picture.

b. Let $y = (1,2,2,1)$, then $\mathscr{Z} = \{(0,0,1),(2,0,2),(3,1,2),(1,1,1)\}$. The upper hall of $\mathscr{Z}$ consists of one facets with vertices $\{(0,0,1),(2,0,1.9),(3,1,2),(1,1,1)\}$. Then we project of the upper hull to $\mathbb{Z}^2$, the resulted $\mathscr{S}(y)$ has one subdivision: $\{(0,0),(2,0),(3,1),(1,1)\}$. See Figure 4.1 middle picture.

c. Let $y = (1,2.1,2,1)$, then $\mathscr{Z} = \{(0,0,1),(2,0,2.1),(3,1,2),(1,1,1)\}$. The upper hall of $\mathscr{Z}$ consists of two facets with vertices $\{(0,0,1),(2,0,2.1),(1,1,1)\}$ and $\{(2,0,2.1),(1,1,1),(3,1,2)\}$. Then we project of the upper hull to $\mathbb{Z}^2$, the resulted $\mathscr{S}(y)$ has one subdivision: $\{(0,0),(2,0),(1,1)\}$, $\{(0,0),(2,0),(3,1)\}$.

These three examples are illustrated in Figure 4.1. We can refine each subdivision into a triangulation (a partition into simplices). Note that $\mathscr{S}(y_1)$ and $\mathscr{S}(y_3)$ are both triangulations, while $\mathscr{S}(y_2)$ needs further partitioning. Let $\mathscr{T}(y) = \{S_j, j \in \mathscr{J}\}$ denote the triangulation, where $\mathscr{J}$ is the index set of simplices for the triangulation, S_j is the jth simplex of the triangulation. Let $J_j = \{j_0, ..., j_d\}$ denote the vertex indicies of jth simplex, then each S_j is determined by $d+1$ vertices: $\{z_{j_0}, ..., z_{j_d}\}$. Finally, let C_j denote the convex hull of $S_j, j = 1, \ldots, |\mathscr{J}|$. In case of $y_1 = (1, 1.9, 2, 1)$, there are two simplices, $\mathscr{J} = \{1,2\}$. S_1 is determined by vertices: $\{(0,0),(2,0),(3,1)\}$, and S_1 is determined by vertices: $\{(0,0),(1,1),(3,1)\}$. Corresponding convex hulls are $C_1 = \{(0,0),(1,0),(2,0),(3,1)\}$, $C_2 = \{(0,0),(1,1),(2,1),(3,1)\}$.

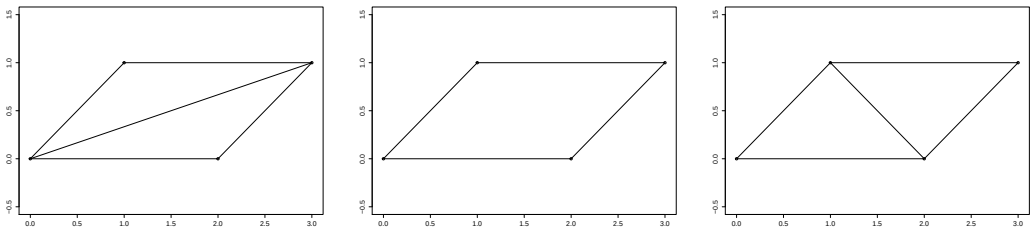


Figure 4.1: Subdivisions $\mathscr{S}(y)$ for cases a.(left), b.(centre), and c.(right).

From this example we can see that for same observations, the subdivision can be totally different by different y.

For finitely many points, the tent functions can be written explicitly via the triangulations (Cule, 2009, Equation 3.6, page 26)

$$t_y(z) = \sum_{j \in \mathscr{J}} (b_j^T z - \beta_j) \mathbb{I}_{C_j}(z) + \delta_{\widehat{\mathcal{S}}_n}(z),$$

for some b_j, β_j. Here, $\mathbb{I}_{C_j}(z)$ is an indicator function and $\mathscr{J}$ indicates the simplices indicies set of the triangulation by y. Finally,

$$\delta_{\widehat{\mathcal{S}}_n}(z) = \begin{cases} 0 & \text{if } z \in \widehat{\mathcal{S}}_n, \\ -\infty & \text{if } z \notin \widehat{\mathcal{S}}_n. \end{cases}$$

Hence we can take the tent function as a composition of affine functions over each simlex. The following works will show how to decomposite the tent function by a chosen y.

Let θ denote an element in a unit d–simplex: $\theta \in [0, \infty)^d$, and $\sum_{i=1}^d \theta_i \leq 1$. Following Cule (2009, page 27), we perform a translation to re-write the above formulas over the unit

simplex. Define $A_j = (z_{j_1} - z_{j_0}| \ ... \ |z_{j_d} - z_{j_0})$ to be a $d \times d$ matrix and let $a_j = z_{j_0}$. Then for $z \in C_j$, $\theta = (A_j)^{-1}(z - a_j)$ is in the unit simplex. Next, let $\widetilde{y}_j \in \mathbb{R}^d$ have components $(y_{j_1} - y_{j_0}, \ldots, y_{j_d} - y_{j_0})$. Then we can write, $b_j = (A_j^T)^{-1}\widetilde{y}_j$ and $\beta_j = a_j^T b_j - y_{j_0}$. Thus,

$$\begin{aligned}
b_j^T z - \beta_j &= [(A_j^T)^{-1}\widetilde{y}_j]^T z - a_j^T (A_j^T)^{-1}\widetilde{y}_j + y_{j_0} \\
&= \widetilde{y}_j^T A_j^{-1} z - a_j^T (A_j^T)^{-1}\widetilde{y}_j + y_{j_0} \\
&= \widetilde{y}_j^T A_j^{-1}(A_j\theta + a_j) - a_j^T (A_j^T)^{-1}\widetilde{y}_j + y_{j_0} \\
&= \widetilde{y}_j^T \theta + + \widetilde{y}_j^T A_j^{-1} a_j - a_j^T (A_j^T)^{-1}\widetilde{y}_j + y_{j_0} \\
&= \widetilde{y}_j^T \theta + + \widetilde{y}_j^T A_j^{-1} a_j - (\widetilde{y}_j^T A_j^{-1} a_j)^T + y_{j_0} \\
&= \widetilde{y}_j^T \theta + y_{j_0} \\
&= (y_{j_1} - y_{j_0})\theta_1 + ... + (y_{j_d} - y_{j_0})\theta_d + y_{j_0} \\
&= y_{j_0}(1 - \theta +_1 - ... - \theta +_d) + y_{j_1}\theta +_1 + ... + y_{j_d}\theta +_d \\
&= y_{j_0}\theta_0 + y_{j_1}\theta_1 + ... + y_{j_d}\theta_d \quad \equiv \quad \theta^T y_{J_j},
\end{aligned}$$

where $\theta_0 = 1 - \theta_1 - \ldots - \theta_d$. Therefore,

$$\begin{aligned}
\sigma(y) &= -\sum_{i=1}^{m} w_i y_i + \sum_{z \in \mathbb{Z}^d} \exp\Big\{t_y(z)\Big\} \\
&= -\sum_{i=1}^{m} w_i y_i + \sum_{z \in \widetilde{\mathcal{S}}_n} \exp\Big\{\sum_{j \in \mathscr{J}} (b_j^T z - \beta_j)\mathbb{I}_{C_j}(z)\Big\} \\
&= -\sum_{i=1}^{m} w_i y_i + \sum_{j \in \mathscr{J}} \sum_{z \in C_j} \exp\Big\{(b_j^T z - \beta_j)\Big\},
\end{aligned}$$

We then obtain

$$\sigma(y) \;=\; -\sum_{i=1}^{m} w_i\, y_i + \sum_{j\in\mathscr{J}} \sum_{z\in C_j, z\notin \cup_{k=1}^{j-1} C_k} \exp\left\{y_{j_0}\theta_0 + y_{j_1}\theta_1 + ... + y_{j_d}\theta_d\right\},$$

where $\theta = A_j^{-1}(z - a_j)$ for $z \in C_j$. Note that some z may belong to more than one simplex, and hence the need to exclude these cases in the second summand above. Recall that the projection of upper hull of $\mathscr{Z}$ onto the first d component can be refined to simplices. To compute $\sigma(y)$, we will then work on each simplex. For each discrete point z inside or on the boundary of the simplex, compute w by $w = A_j^{-1}(z - a_j)$, and add the exponential term to the σ function. Quick Hull algorithm (c package qhull) is applied to compute convex hull and triangulation.

4.2 Derive the explicit form of gradient of $\sigma(y)$

We also need to compute the derivatives, or when not differentiable, the directional derivative of $\sigma(y)$. As in Cule (2009, Section 3.4.2, page 34), $\sigma(y)$ is differentiable if $\mathscr{S}(y)$ is a triangulation, while if $\mathscr{S}(y)$ is not a triangulation, it is not differentiable. This is relatively straightforward to see from Figure 4.1, as small changes to the second element of y yield very different subdivisions.

The following Theorem 4.2.1 conclude that the objective function $\sigma(y)$ is not differentiable everywhere.

Proposition 4.2.1. *The function* $\sigma(y) = -\sum_{i=1}^{m} w_i y_i + \sum_{z \in \widehat{\mathcal{S}}_n} \exp\{t_y(z)\}$ *is not differentiable everywhere.*

Proof. Denote the directional derivatives as

$$\partial\sigma(y;u) = \lim_{t \to 0} \frac{\sigma(y+tu) - \sigma(y)}{t}.$$

Since σ is convex, the directional derivative exist (Rockafellar, 1970, Theorem 23.1 page 213). Furthermore, the function is differentiable if $\partial\sigma(y;u) = -\partial\sigma(y;-u)$. We will show that $\partial\sigma(y;e_i) + \partial\sigma(y;-e_i) > 0$ occurs when $\mathscr{S}(y)$ is not a triangulation, where $e_i \in \{0,1\}^m$ is the ith row of the m dimensional identity matrix. For simplicity, consider the case when there are $m = d+2$ elements in general position, as the more complex case is similar.

For each i and $\varepsilon_0 > 0$ sufficiently small, we have that $\mathscr{S}(y+\varepsilon_0 e_i), \mathscr{S}(y-\varepsilon_0 e_i)$ both form triangulations. Following Cule (2009, Section 3.4.2, page 35), we may write

$$t_{y+\varepsilon e_i}(x) = t_y(x) + \varepsilon g_{e_i, \mathscr{S}(y+\varepsilon_0 e_i)}(x) \qquad t_{y-\varepsilon e_i}(x) = t_y(x) + \varepsilon g_{-e_i, \mathscr{S}(y-\varepsilon_0 e_i)}(x),$$

where $g_{e_i, \mathscr{S}(y+\varepsilon e_i)}$ is the function obtained by:

1 project $\{(z_1,y_1),...,(z_{i-1},y_{i-1}),(z_i,y_i+\varepsilon),(z_{i+1},y_{i+1}),...,(z_n,y_n)\}$ on $\mathbb{Z}^d$, and get the triangulation $\mathscr{S}(y+\varepsilon e_i)$.

2 linearly interpolating the points

$$\{(z_1,0),...,(z_{i-1},0),(z_i,1),(z_{i+1},0),...,(z_n,0)\}$$

over each simplex of $\mathscr{S}(y+\varepsilon e_i)$.

And $g_{-e_i,\mathscr{S}(y-\varepsilon e_i)}$ is obtained by the similar way. Note that $g_{e_i,\mathscr{S}(y+\varepsilon_0 e_i)}, -g_{-e_i,\mathscr{S}(y-\varepsilon_0 e_i)}(x)$ are the upper and lower hulls of the points

$$\{(z_1,0),...,(z_{i-1},0),(z_i,1),(z_{i+1},0),...,(z_m,0)\},$$

respectively.

Letting e_{ij} denote the (i,j)-element of the $n \times n$ identity matrix, we can write

$$\begin{aligned}
\partial\sigma(y;e_i) &= \lim_{\varepsilon\to 0}\frac{\sigma(y+\varepsilon e_i)-\sigma(y)}{\varepsilon} \\
&= \lim_{\varepsilon\to 0}\frac{\sum_{j=1}^{m} -w_j(y_j+\varepsilon e_{ij}) + \sum_{z\epsilon\widehat{\mathcal{S}}_n}\exp\left\{t_{y+\varepsilon e_i}(z)\right\} + \sum_{j=1}^{m} w_j y_j - \sum_{z\epsilon\widehat{\mathcal{S}}_n}\exp\left\{t_y(z)\right\}}{\varepsilon} \\
&= \lim_{\varepsilon\to 0}\frac{\sum_{j=1}^{m} -w_j(y_j+\varepsilon e_{ij}) + \sum_{z\epsilon\widehat{\mathcal{S}}_n}\exp\left\{t_y(z)+\varepsilon g_{e_i,\mathscr{S}(y+\varepsilon_0 e_i)}(z)\right\} + \sum_{j=1}^{m} w_j y_j - \sum_{z\epsilon\widehat{\mathcal{S}}_n}\exp\left\{t_y(z)\right\}}{\varepsilon} \\
&= -w_i + \lim_{\varepsilon\to 0}\frac{\left(\sum_{z\epsilon\widehat{\mathcal{S}}_n}\exp\left\{t_y(z)+\varepsilon g_{e_i,\mathscr{S}(y+\varepsilon_0 e_i)}(z)\right\} - \sum_{z\epsilon\widehat{\mathcal{S}}_n}\exp\left\{t_y(z)\right\}\right)}{\varepsilon} \\
&= -w_i + \sum_{z\epsilon\widehat{\mathcal{S}}_n}\exp\left\{t_y(z)\right\}\lim_{\varepsilon\to 0}\frac{\exp\left\{\varepsilon g_{e_i,\mathscr{S}(y+\varepsilon_0 e_i)}(z)\right\}-1}{\varepsilon} \\
&= -w_i + \sum_{z\epsilon\widehat{\mathcal{S}}_n}\exp\left\{t_y(z)\right\} g_{e_i,\mathscr{S}(y+\varepsilon_0 e_i)}(z).
\end{aligned}$$

Similarly, we find

$$\partial\sigma(y;-e_i) \;=\; w_i + \sum_{z\in\widehat{\mathcal{S}}_n} \exp\{t_y(z)\}\, g_{-e_i,\mathscr{S}(y-\varepsilon_0 e_i)}(z).$$

Hence,

$$\partial\sigma(y;e_i)+\partial\sigma(y;-e_i) \;=\; \sum_{z\in\widehat{\mathcal{S}}_n} \exp\{t_y(z)\}\left\{g_{e_i,\mathscr{S}(y+\varepsilon_0 e_i)}(z)+g_{-e_i,\mathscr{S}(y-\varepsilon_0 e_i)}(z)\right\}.$$

Recall that $g_{e_i,\mathscr{S}(y+\varepsilon_0 e_i)}, -g_{-e_i,\mathscr{S}(y-\varepsilon_0 e_i)}(x)$ are the upper and lower hulls of the points

$$\{(z_1,0),...,(z_{i-1},0),(z_i,1),(z_{i+1},0),...,(z_m,0)\},$$

respectively. It follows that $g_{e_i,\mathscr{S}(y+te_i)}+g_{-e_i,\mathscr{S}(y-te_i)}>0$ and hence $\partial\sigma(y;e_i)+\partial\sigma(y;-e_i)>0$. □

The following Proposition gives the explicit form of $\partial\sigma(y)$.

Proposition 4.2.2. *1 The function σ is differentiable at y, and for $i=1,...,m$ we have*

$$\partial_i\sigma(y) \;=\; -w_i + \sum_{j\in\mathscr{J}} \mathbb{I}_{C_j}(y_i) \sum_{z\in C_j, z\notin\cup_{k=1}^{j-1}C_k} \partial_i\left\{\theta^T y_{J_j}\right\}\exp\left\{\theta^T y_{J_j}\right\},$$

2 The function σ is not differentiable at y, but $\left(\partial_1\sigma(y),...,\partial_m\sigma(y)\right)$ is a subgradient of

σ *at* y.

Proof. (1) When $\sigma(y)$ is differentiable, we easily obtain that

$$\partial_i \sigma(y) \quad = \quad -w_i + \sum_{j \in \mathscr{J}} \mathbb{I}_{C_j}(y_i) \sum_{z \in C_j, z \notin \cup_{k=1}^{j-1} C_k} \partial_i \left\{ \theta^T y_{J_j} \right\} \exp \left\{ \theta^T y_{J_j} \right\},$$

Note that when we compute the ith partial derivative, we only need to consider those simplices which involve y_i, so the indicator function above ensures that only the simplex involving y_i will be counted.

(2) Proposition 4.2.1 showed that σ is not differentiable everywhere. If $\sigma(y)$ is not differentiable at a given $y \in \mathbb{R}^m$, Cule et al. (2010) Proposition 5(b) can also be applied to our case. Rockafellar (1970) shows that for any $\varepsilon > 0$, we can find a point $\tilde{y} \in \mathbb{R}^n$ with $\| y - \tilde{y} \|_2 < \varepsilon$, such that σ is differentiable at $\tilde{y}$ and $\| \nabla\sigma(\tilde{y}) - \partial\sigma(y) \|_2 < \varepsilon$. Hence we conclude that $\big(\partial_1 \sigma(y), ..., \partial_m \sigma(y)\big)$ is a subgradient of σ at y. Technically, this can be done by sequentially making small adjustments to the components of y in the same order as that in which the vertices were pushed in constructing the triangulation. Cule (2009) also mentioned that theoretically, it is necessary to check if the refinement of $\mathscr{S}(y)$ by **QuickHull** algorithm is a triangulation. But in practice, this was not found to be necessary. □

4.3 Subgradient algorithm

Since the function $\sigma(y)$ is not differentiable everywhere, following Cule et al. (2010), we apply a subgradient method to solve the problem. While the convergence rate of subgradient method might be slow due to differential direction. Cule et al. (2010) perform a sequence of dilations in the direction between two successive subgradient, which is known as Shor's r-algorithm. Cule et al. (2010) also state that the formal convergence of the r-algorithm has not been proved theoretically, but it is found to be robust, efficient and accurate in practice. The idea here is to "make steps in the direction opposite to a sub-gradient" (Kappel and Kuntsevich, 2000, page 193). These steps are made in a transformed, "dilated", space. Kappel and Kuntsevich (2000) describe further improvements to the method via modified stopping criteria.

As in Cule et al. (2010), we use this latter modification with stopping criteria

$$\begin{aligned} |y_i^{k+1} - y_i^k| &\leq \delta |y_i^k| \text{ for } i = 1, ..., n \\ |\sigma(y^{k+1}) - \sigma(y^k)| &\leq \varepsilon |\sigma(y^k)| \\ \left| 1 - \sum_{z \in \mathbb{Z}^d} \exp\left\{ t_{y^k}(z) \right\} \right| &\leq \eta \end{aligned}$$

for fixed tolerances δ, ε and η. The last criterion above is not one suggested by Kappel and Kuntsevich (2000), but is there to ensure that the algorithm returns close to a proper probability mass function. In our current implementation, the tolerances are set to $\delta = \varepsilon =$

$\eta = 10^{-4}$.

In our algorithm, we did not change the subgradient iteration process, we follow the work and code of Cule et al. (2010). Although our algorithm is in discrete setting, the objective function $\sigma(y)$ is still a continuous function. The only difference between our algorithm and that of Cule et al. (2010) is that the objective function is different. Hence our main work is to derive the explicit form of the objective function and its gradient, as well as to compute the values of the objective function and its gradient with a given $y \in \mathbb{R}^m$. For detailed subgradient algorithm description, we refer to Cule et al. (2010).

4.4 Computing algorithm and comparison

To compute $\sigma(y)$, we refine the projection of $\mathscr{L}$ into simplices. We then work on each simplex, and find all lattice points inside or on the boundary of the simplex. If the discrete point has not been counted, we compute the corresponding θ, and add in the exponential term. The quickhull algorithm is applied to compute convex hulls and triangulations. Details of these calculations, as well as gradient and subgradient calculations are given in Algorithm 1.

Recall that our optimization problem is

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

Algorithm 1 Calculate $\sigma(y)$ and gradient of $\sigma(y)$, input z_{obs}, y

Compute convex hull of observations: $C = \text{conv}(z_{obs})$
Compute extreme/out points of C : z_{out}, corresponding subset of y: y_{out}
Compute inner points of C : $z_{in} = z_{obj} \setminus z_{out}$, corresponding subset of y: y_{in}
$y_{max} = \max\{y_1, ..., y_m\}$
$y_{min} = \min\{y_1, ..., y_m\}$
Combine $\frac{y_{out}}{y_{max} - y_{min}}$ and z_{out} to get $d+1$ dimensional data set: zz_{out}
Combine $\frac{y_{in}}{y_{max} - y_{min}}$ and z_{in} to get $d+1$ dimensional data set: zz_{in}
Combine $\frac{y_{min} - 1}{y_{max} - y_{min}}$ and z_{out} to get $d+1$ dimensional data set: zz_{xtr}
All points set: $zz_{all} = zz_{out} \cup zz_{in} \cup zz_{xtr}$
Compute convex hull of All points set: $C_{all} = \text{conv}(zz_{all})$
Compute facet set of C_{all} : $fct = \{fct_1, ..., fct_k\}$
Initial $\sigma(y) = -(\bar{p}_1 * y_1 + ... + \bar{p}_m * y_m)$
Initial $\partial_i \sigma(y) = -\bar{p}_i, i = 1, ..., m$
Initial E_{all} as an empty list ▷ Used to check duplication
for each facet fct_j, $1 \le j \le k$ **do**
    if fct_j is a true facet **then**
        The extreme (out) points set of fct_j : $p_j = \{z_{j_0}, ..., z_{j_d}\}$
        Matrix $A = [z_{j_1} - z_{j_0} | ... | z_{j_d} - z_{j_0}]$
        Vector $a_j = z_{j_0}$
        Inverse matrix of A : A^{-1}
        Vector $y_{tmp} = \{y_{j_0}, ..., y_{j_d}\}$
        Generate a rectangle of p_j : $rec = \{r \in \mathbb{Z}^d\}$, such that $r_i = \{z \in \mathbb{Z} | \min\{z^i_{j_0}, ..., z^i_{j_d}\} \le z \le \max\{z^i_{j_0}, ..., z^i_{j_d}\}\}$, for $1 \le i \le d$
        for each point r in rec **do**
            if r is inside convex hull of p_j **then**
                Add r to enumerate list: E_j
            end if
        end for
        for each point of E_j : e **do**
            if e is not duplicated with any points of E_{all} **then**
                Vector $w = A^{-1}(e - a_j), w_0 = 1 - w_1 - ... - w_d$
                Sigma function: $\sigma(y) += \exp\{y_{j_0} w_0 + ... + y_{j_d} w_d\}$
                for $i \in \{j_0, ..., j_d\}$ **do**
                    Gradient: $\partial_i \sigma(y) += w_i \exp\{y_{j_0} w_0 + ... + y_{j_d} w_d\}$
                end for
                Add e to enumerate list E_{all}
            end if
        end for
    end if
end for
Return $\sigma(y)$, $\partial_i \sigma(y)$ for $i = 1, ..., m$

Firstly the log-likelihood function is a convex function of p. Secondly the following Lemma shows that $\mathcal{P}_0$ is a convex set. Hence our optimization problem is a convex optimization problem and we are proposing an algorithm to solve our particular problem.

Lemma 4.4.1. *By our definition,* $\mathcal{P}_0 = \{p = \exp\{\varphi\} \,|\, \varphi$ *is concave-extendible and* p *is PMF on* $\mathbb{Z}^d\}$ *is a convex set.*

Proof. Let $\mathcal{L}_1 = \{p = \exp\{\varphi\} \,|\, \varphi \text{ is concave-extendible}\}, \mathcal{L}_2 = \{p \text{ is PMF on } \mathbb{Z}^d\}$. We can re-write $\mathcal{P}_0$ as: $\mathcal{P}_0 = \mathcal{L}_1 \cap \mathcal{L}_2$. Firstly, We show that $\mathcal{L}_1$ is a convex set. Exponential function is a convex function, it is sufficient if we can show that $\{\varphi \,|\, \varphi$ is concave-extendible function$\}$ is a convex set. Note that convex-extendible is closed under addition by Murota and Shioura (2001). We now show that convex-extendible is also closed under multiplication by a constant. Let $h(z)$ be a convex-extendible function on $\mathbb{Z}^d$, hence there exists a closed convex function $h^R(x)$ on $\mathbb{R}^d$, which is the convex extension of $h(z)$. Let $k \in \mathbb{R}$, it is not hard to see that $k * h^R(x)$ is the convex extension of $k * h(z)$. Hence $k * h(z)$ is also convex-extendible. Therefore convex-extendible functions set is convex. It follows that concave-extendible functions set is also convex.

Secondly, $\mathcal{L}_2$ is also a convex set. Hence $\mathcal{P}_0$, as the intersection of two convex sets, is also a convex set. □

Algorithms to solve the convex feasibility problem are usually called projections

onto convex sets (POCS) method. More details of convex feasibility problem and POCS, we refer to Bauschke and Borwein (1996). The subgradient method is one of the approach to solve the convex feasibility problem per Bauschke and Borwein (1996). Since minimizing the objective function $\sum_{i=1}^{n} \log p(X_i)$ over $\mathcal{P}_0$ is equivalent to minimizing the KL divergence between the empirical distribution $\bar{p}_n$ and any eLC PMF ($p \in \mathcal{P}_0$), the minimizer $\widehat{p}_n$ can be viewed as the projection of the empirical distribution $\bar{p}_n$ onto $\mathcal{P}_0$ in the KL divergence sense.

Chapter 5

Asymptotic Properties

For two PMFs p and p_0, we define the l_k and Hellinger distances as

$$\begin{aligned} l_k(p, p_0) &= \begin{cases} \left(\sum_{z \in \mathbb{Z}^d} |p(z) - p_0(z)|^k \right)^{1/k} & \text{if } 1 \le k \le \infty, \\ \sup_{z \in \mathbb{Z}^d} |p(z) - p_0(z)| & \text{if } k = \infty, \end{cases} \\ h^2(p, p_0) &= \frac{1}{2} \sum_{z \in \mathbb{Z}} \left(\sqrt{p(z)} - \sqrt{p_0(z)} \right)^2. \end{aligned}$$

Our main consistency result follows.

Theorem 5.0.1. *Suppose that p_0 is a discrete distribution on $\mathbb{Z}^d$ with finite expected value of $\|z\|_\infty$ and finite entropy, that is,*

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

Assume also that 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 $1 \leq k \leq \infty$ or the Hellinger distance h.

We have thus shown that even if the true distribution p_0 is not in $\mathcal{P}_0$, then 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 behavior.

We will prove the following lemma firstly.

Lemma 5.0.1. *Let p_n, p be discrete probability mass functions on $\mathbb{Z}^d$, and $p_n \to p$ for all $z \in \mathbb{Z}^d$, then $l_k(p_n, p) \to 0$ for $1 \leq k \leq \infty$, and $h^2(p_n, p) \to 0$, as $n \to 0$.*

Proof. Clearly, it is sufficient to show that pointwise convergence implies the other types of convergence. To this end, fix $\varepsilon > 0$. Then, there exists a K such that $\sum_{\|z\|_\infty \leq K} p(z) \geq 1 - \varepsilon/4$. Furthermore, since $p_n \to p$, for all $z \in \mathbb{Z}^d$, there exists large enough N, such that

$$\sup_{\|z\|_\infty \leq K} |p_n(z) - p(z)| \quad \leq \quad \frac{\varepsilon}{4(2K+1)^d},$$

for all $n \geq N$. Hence, for any z such that $\|z\|_\infty \leq K$, and any $n \geq N$,

$$|p_n(z) - p(z)| \leq \sup_{\|z\|_\infty \leq K} |p_n(z) - p(z)| \leq \frac{\varepsilon}{4(2K+1)^d}.$$

Note that the size of set $\{z \mid \|z\|_\infty \leq K\}$ is $(2K+1)^d$. Following above inequality, we can deduce the below two inequalities:

$$\sum_{\|z\|_\infty \leq K} |p_n(z) - p(z)| \leq \frac{\varepsilon}{4(2K+1)^d} * (2K+1)^d = \frac{\varepsilon}{4},$$

and

$$p_n(z) \geq p(z) - \frac{\varepsilon}{4(2K+1)^d}.$$

From the above, it also follows that for all $n \geq N$,

$$\sum_{\|z\|_\infty \leq K} p_n(z) \quad \geq \quad \sum_{\|z\|_\infty \leq K} p(z) - \frac{\varepsilon}{4(2K+1)^d} * (2K+1)^d \quad \geq \quad 1 - \varepsilon/4 - \varepsilon/4 = 1 - \varepsilon/2.$$

Putting these facts together, we find that

$$\begin{aligned}
\sum_{z\in\mathbb{Z}^d} |p_n(z)-p(z)| &= \sum_{\|z\|_\infty\le K} |p_n(z)-p(z)| + \sum_{\|z\|_\infty> K} |p_n(z)-p(z)| \\
&\le \sum_{\|z\|_\infty\le K} |p_n(z)-p(z)| + \sum_{\|z\|_\infty> K} p_n(z) + \sum_{\|z\|_\infty> K} p(z) \\
&= \sum_{\|z\|_\infty\le K} |p_n(z)-p(z)| + \Big(1-\sum_{\|z\|_\infty\le K} p_n(z)\Big) + \Big(1-\sum_{\|z\|_\infty\le K} p(z)\Big) \\
&\le \varepsilon/4+\varepsilon/2+\varepsilon/4 \;=\; \varepsilon.
\end{aligned}$$

We have thus shown that pointwise convergence implies $l_1(p_n,p)\to 0$.

Note that for any fixed z_0, we have $|p_n(z_0)-p(z_0)| \le \sum_{z\in\mathbb{Z}^d}|p_n(z)-p(z)|$, and hence

$$l_\infty(p_n(z),p) = \sup_{z\in\mathbb{Z}^d}|p_n(z)-p(z)| \le \sum_{z\in\mathbb{Z}^d}|p_n(z)-p(z)| = l_1(p_n,p)\to 0, \text{ as } n\to\infty.$$

Moreover, $0\le p_n(z),p(z)\le 1$ implies that $|p_n(z)-p(z)|\le 1$. Hence $|p_n(z)-p(z)|^k \le |p_n(z)-p(z)|$, for any $1<k<\infty$. Therefore, for any $1<k\le\infty$,

$$l_k^k(p_n,p) \;=\; \sum_{z\in\mathbb{Z}^d}|p_n(z)-q(z)|^k \le \sum_{z\in\mathbb{Z}^d}|p_n(z)-q(z)| = l_1(p_n,p)$$

Lastly, recall that $2h^2(p,q)\le l_1(p,q)$. We conclude that pointwise convergence implies all other types of convergence as well. □

We now prove Theorem 5.0.1

Proof. Let X_n be a random vector with PMF $\widehat{p}_n$. Then by Markov's inequality and Lemma 3.0.1, we have that

$$P(\|X_n\|_\infty \geq m) \leq \frac{\sum_{\|z\|_\infty \geq m} \|z\|_\infty \widehat{p}_n(z)}{m} \leq \frac{\sum_{z\in\mathbb{Z}^d} \|z\|_\infty \widehat{p}_n(z)}{m} \leq \frac{\sum_{z\in\mathbb{Z}^d} \|z\|_\infty \overline{p}_n(z)}{m},$$

since the norm $||\cdot||_\infty$ is convex-extendible. By strong law of large number and the finite mean assumption of the Theorem, $\sum_{z\in\mathbb{Z}^d} \|z\|_\infty \overline{p}_n(z) \leq 2\sum_{z\in\mathbb{Z}^d} \|z\|_\infty p_0(z)$, say, almost surely for all n sufficiently large. Note that $|\sum_{z\in\mathbb{Z}^d} \|z\|_\infty p_0(z)| < \infty$ by assumption. Hence, $\sum_{z\in\mathbb{Z}^d} \|z\|_\infty \bar{p}_n(z) < \infty$ almost surely for sufficient large n. Therefore for any $\varepsilon > 0$, there exists large enough $m > 0$, such that

$$\frac{\sum_{z\in\mathbb{Z}^d} \|z\|_\infty \bar{p}_n(z)}{m} < \varepsilon.$$

It follows that the sequence X_n is tight. Therefore, there exists a subsequence of $\widehat{p}_n$ and a $\widetilde{p}$, which we denote again by n, such that $\widehat{p}_n \to \widetilde{p}$. By Theorem 2.4.1 we conclude that $\widetilde{p}$ is also eLC. It remains to show that $\widetilde{p} = \widehat{p}_0$ to finish the proof.

Because log function is strictly increasing function and $\bar{p}_n(z) > 0$ for any $z \in \mathbb{Z}^d$,

we have $\bar{p}_n(z)\log(\widehat{p}_n(z)+b) \geq \bar{p}_n(z)\log(\widehat{p}_n(z))$ for any $z \in \mathbb{Z}^d$, hence

$$\sum_{z\in\mathbb{Z}^d} \bar{p}_n(z)\log(\widehat{p}_n(z)+b) \geq \sum_{z\in\mathbb{Z}^d} \bar{p}_n(z)\log(\widehat{p}_n(z)).$$

Also by definition of MLE, $\widehat{p}_n$ maximize log-likelihood $\sum_{z\in\mathbb{Z}^d} \bar{p}_n \log p(z)$ over $\mathcal{P}_0$, for any $b > 0$, we have

$$\sum_{z\in\mathbb{Z}^d} \bar{p}_n(z)\log(\widehat{p}_n(z)+b) \geq \sum_{z\in\mathbb{Z}^d} \bar{p}_n(z)\log(\widehat{p}_n(z)) \geq \sum_{z\in\mathbb{Z}^d} \bar{p}_n(z)\log(\widehat{p}_0(z)).$$

The 2nd inequality comes from $\widehat{p}_n$ is MLE, which maximize log likelihood. Hence

$$\sum_{z\in\mathbb{Z}^d} \overline{p}_n(z)\log(\widehat{p}_n(z)+b) \quad \geq \quad \sum_{z\in\mathbb{Z}^d} \overline{p}_n(z)\log(\widehat{p}_0(z)).$$

Therefore,

$$\begin{aligned}
&\sum_{z\in\mathbb{Z}^d} \overline{p}_n(z)\log(\widehat{p}_n(z)+b) - \sum_{z\in\mathbb{Z}^d} \overline{p}_n(z)\log(\widehat{p}_0(z)) \\
&= \sum_{z\in\mathbb{Z}^d} (\overline{p}_n(z)-p_0(z))\log(\widehat{p}_n(z)+b) + \sum_{z\in\mathbb{Z}^d} (p_0(z)-\overline{p}_n(z))\log(\widehat{p}_0(z)) \\
&\qquad + \sum_{z\in\mathbb{Z}^d} p_0(z)\log\frac{\widehat{p}_n(z)+b}{\widehat{p}_0(z)+b} + \sum_{z\in\mathbb{Z}^d} p_0(z)\log\frac{\widehat{p}_0(z)+b}{\widehat{p}_0(z)} \quad \geq \quad 0. \qquad (5.1)
\end{aligned}$$

We next get rid of the first two terms on the right-hand side. For the first term on

the right-hand side,

$$\begin{aligned} |\overline{p}_n(z) - p_0(z)| &\leq |\mathbb{F}_n(z) - F_0(z)| + |\mathbb{F}_n(z-1) - F_0(z-1)| \\ &\leq 2\sup_{z\in\mathbb{Z}^d} |\mathbb{F}_n(z) - F_0(z)| \xrightarrow{a.s.} 0, \text{ as } n \to \infty, \end{aligned}$$

where $\mathbb{F}_n(z), F_0(z)$ denote the empirical and true cumulative distribution functions (CDFs), corresponding to $\overline{p}_n$ and p_0, respectively. By Lemma 5.0.1, we conclude that $\sum_{z\in\mathbb{Z}^d} |\overline{p}_n(z) - p_0(z)| \to 0$. Since

$$\log(b) \leq \log(\widehat{p}_n(z) + b) \leq \log(b+1),$$

we have

$$|\log(\widehat{p}_n(z) + b)| \leq M < \infty,$$

where $M = \max\{|\log(b)|, |\log(b+1)|\} < \infty$ is afinite constant. Therefore the first term converges to 0 (showing as following).

$$\begin{aligned} \left| \sum_{z\in\mathbb{Z}^d} (\overline{p}_n(z) - p_0(z)) \log(\widehat{p}_n(z) + b) \right| &\leq \sum_{z\in\mathbb{Z}^d} |\overline{p}_n(z) - p_0(z)| |\log(\widehat{p}_n(z) + b)| \\ &\leq M \sum_{z\in\mathbb{Z}^d} |\overline{p}_n(z) - p_0(z)| \xrightarrow{a.s.} 0, \quad \text{as } n \to 0. \end{aligned}$$

We relax this fixed sample space to any sample space, above conclusion is true almost

surely.

We now show that the 2nd term converges to zero. Since $\widehat{p}_0$ minimizes KL divergence, that is,

$$-\sum_{z\in\mathbb{Z}^d} p_0(z)\log\frac{\widehat{p}_0(z)}{p_0(z)} - \Big(\sum_{z\in\mathbb{Z}^d} p_0(z)\log\frac{p_0(z)}{p_0(z)}\Big) = -\sum_{z\in\mathbb{Z}^d} p_0(z)\log\widehat{p}_0(z) + \sum_{z\in\mathbb{Z}^d} p_0(z)\log p_0(z) \le 0.$$

Note that $\log p_0(z), \log\widehat{p}_0(z) < 0$, hence by assumption

$$E_{p_0}[|\log\widehat{p}_0(z)|] = \sum_{z\in\mathbb{Z}^d} p_0(z)\Big(-\log\widehat{p}_0(z)\Big) \le \sum_{z\in\mathbb{Z}^d} p_0(z)\Big(-\log p_0(z)\Big) < \infty.$$

$$E_{\widehat{p}_0}[|\log\widehat{p}_0(Z)|] = -\sum_{z\in\mathbb{Z}^d} p_0(z)\log\widehat{p}_0(z) \;\le\; -\sum_{z\in\mathbb{Z}^d} p_0(z)\log\widehat{p}_0(z) = E_{p_0}[|\log\widehat{p}_0(Z)|] < \infty.$$

Therefore, by the strong law of large numbers,

$$\begin{aligned}\sum_{z\in\mathbb{Z}^d}(p_0(z)-\overline{p}_n(z))\log\widehat{p}_0(z) &= \sum_{z\in\mathbb{Z}^d}\overline{p}_n(z)\Big(-\log\widehat{p}_0(z)\Big) - \sum_{z\in\mathbb{Z}^d} p_0(z)\Big(-\log\widehat{p}_0(z)\Big)\\ &= E_{\overline{p}_n}[|\log\widehat{p}_0(Z)|] - E_{p_0}[|\log\widehat{p}_0(Z)|]\\ &\xrightarrow{\text{a.s.}} 0, \text{ as } n\to\infty,\end{aligned}$$

where Z is random vector on $\mathbb{Z}^d$. Since the first two terms on the right-hand side of (5.1) both converge to zero, we have

$$\limsup_n \sum_{z\in\mathbb{Z}^d} p_0(z)\log\frac{\widehat{p}_0(z)+b}{\widehat{p}_n(z)+b} \quad\leq\quad \sum_{z\in\mathbb{Z}^d} p_0(z)\log\frac{\widehat{p}_0(z)}{\widehat{p}_0(z)+b}.$$

By Fatou's Lemma, we have

$$\liminf_{b\to 0}\sum_{z\in\mathbb{Z}^d}\left\{-p_0(z)\log\frac{\widehat{p}_0(z)}{\widehat{p}_0(z)+b}\right\} \quad\geq\quad \sum_{z\in\mathbb{Z}^d}\liminf_{b\to 0}\left\{-p_0(z)\log\frac{\widehat{p}_0(z)}{\widehat{p}_0(z)+b}\right\} \quad=\quad 0,$$

Note that $-\liminf -f = \limsup f$, therefore

$$\begin{aligned}\limsup_{b\to 0}\limsup_n \sum_{z\in\mathbb{Z}^d} p_0(z)\log\frac{\widehat{p}_0(z)+b}{\widehat{p}_n(z)+b} &\leq \limsup_{b\to 0}\sum_{z\in\mathbb{Z}^d} p_0(z)\log\frac{\widehat{p}_0(z)}{\widehat{p}_0(z)+b}\\ &\leq \sum_{z\in\mathbb{Z}^d}\limsup_{b\to 0} p_0(z)\log\frac{\widehat{p}_0(z)}{\widehat{p}_0(z)+b}\\ &= 0.\end{aligned}$$

For each n, fixed b and any $z\in\mathbb{Z}^d$, since $\log(x), x\in\mathbb{R}$ is monotonically increasing and $0\leq\widehat{p}_n(z), p_0(z), \widehat{p}_0(z)\leq 1$, we have

$$p_0(z)\log\frac{b}{b+1}\leq p_0(z)\log\frac{\widehat{p}_0(z)+b}{\widehat{p}_n(z)+b}\leq p_0(z)\log\frac{b+1}{b}.$$

Hence, by dominated convergence theorem, we have

$$\begin{aligned}\limsup_{b\to 0}\limsup_{n}\sum_{z\in\mathbb{Z}^d} p_0(z)\log\frac{\widehat{p}_0(z)+b}{\widehat{p}_n(z)+b} &= \limsup_{b\to 0}\sum_{z\in\mathbb{Z}^d}\limsup_{n} p_0(z)\log\frac{\widehat{p}_0(z)+b}{\widehat{p}_n(z)+b}\\ &= \limsup_{b\to 0}\sum_{z\in\mathbb{Z}^d} p_0(z)\log\frac{\widehat{p}_0(z)+b}{\widetilde{p}(z)+b}.\end{aligned}$$

Without loss of generality, we can restrict $0 < b \le 1$, and hence $-\log(\widehat{p}_0+b) \ge -\log 2$, which implies that $-\log(\widehat{p}_0+b)$ is bounded below and increases as $b \to 0$. Therefore, by monotone convergence theorem, we have

$$\begin{aligned}\limsup_{b\to 0}\sum_{z\in\mathbb{Z}^d} p_0(z)\left\{-\log(\widehat{p}_0(z)+b)\right\} &= -\sum_{z\in\mathbb{Z}^d}\limsup_{b\to 0} p_0(z)\log(\widehat{p}_0(z)+b)\\ &= -\sum_{z\in\mathbb{Z}^d} p_0(z)\log\widehat{p}_0(z),\end{aligned}$$

and similarly when $\widehat{p}_0$ is replaced by $\widetilde{p}$. Putting together the above arguments, we

thus arrive at

$$\begin{aligned}
&\limsup_{b\to 0} \sum_{z\in\mathbb{Z}^d} p_0(z) \log \frac{\widehat{p}_0(z)+b}{\widetilde{p}(z)+b} \\
&= -\left(\limsup_{b\to 0} \sum_{z\in\mathbb{Z}^d} -p_0(z) \log(\widehat{p}_0(z)+b) - \limsup_{b\to 0} \sum_{z\in\mathbb{Z}^d} -p_0(z) \log(\widetilde{p}_0(z)+b) \right) \\
&= \sum_{z\in\mathbb{Z}^d} p_0(z) \log \widehat{p}_0(z) - \sum_{z\in\mathbb{Z}^d} p_0(z) \log \widetilde{p}_0(z) \\
&= \sum_{z\in\mathbb{Z}^d} p_0(z) \log \frac{\widehat{p}_0(z)}{\widetilde{p}(z)} \quad \leq \quad 0.
\end{aligned}$$

Furthermore,

$$\begin{aligned}
\sum_{z\in\mathbb{Z}^d} p_0(z) \log \frac{\widehat{p}_0(z)}{\widetilde{p}(z)} &= \sum_{z\in\mathbb{Z}^d} p_0(z) \log \frac{p_0(z)}{\widetilde{p}(z)} \frac{\widehat{p}_0(z)}{p_0(z)} \\
&= \sum_{z\in\mathbb{Z}^d} p_0(z) \log \frac{p_0(z)}{\widetilde{p}(z)} - \sum_{z\in\mathbb{Z}^d} p_0(z) \log \frac{p_0(z)}{\widehat{p}_0(z)} \quad \leq \quad 0.
\end{aligned}$$

Rearranging, we find that

$$\sum_{z\in\mathbb{Z}^d} p_0(z) \log \frac{p_0(z)}{\widetilde{p}(z)} \quad \leq \quad \sum_{z\in\mathbb{Z}^d} p_0(z) \log \frac{p_0(z)}{\widehat{p}_0(z)}.$$

However, as $\widehat{p}_0$ is the unique minimizer of the quantity on the right hand side, we obtain that $\widetilde{p} = \widehat{p}_0$. Therefore $\widehat{p}_n \to \widehat{p}_0$, and by Lemma 5.0.1 we have $d(\widehat{p}_0, \widehat{p}_0) \to 0$, as $n \to \infty$. □

Part II

Application of discrete log-concave in clustering

Chapter 6

Introduction

6.1 Motivation and overview

Clustering has a broad application in practice, such as biology, medicine, business and marketing. Comparing to the continuous setting, less work has been done for categorical data clustering, although categorical data is very popular for a wide range of applications.

Categorical data is usually divided into two types: nominal and ordinal. With the former, levels of data attributes are not ordinal, and have no order. For example, eye color can take the values: black, green, blue, and brown. Clearly, these colors are not ordered. For ordinal categorical data, levels of attributes have an order. For example, income may have three levels: low, medium, and high. Our work focuses

on nominal categorical data. It is very different from ordinal data, where distance can be measured using a metric. For nominal categorical data, the main challenge is that the points' attributes levels have no order, such that metric distance calculated vary significantly when the attributes' levels are coded in different way. Hamming distance (Gordon, 1999) measures the similarity or dissimilarity between two given data points. It counts how many attributes with different values there are for two data points. This gives us a way to measure how "far" two nominal data points are. Hamming distance actually measures the similarity or dissimilarity of two vectors (data points).

Zhang et al. (2006) developed an algorithm to cluster nominal categorical data based on Hamming distance. They refer to it as the HD vector algorithm. This algorithm can automatically detect significant clusters in a given data. Each cluster is defined by a cluster center and cluster radius. To identify if a cluster exists, the algorithm uses something called HD vector, which is akin to empirical distribution. From this HD vector, the cluster center and radius is identified. This procedure is sensitive to the lack of smoothness in the HD vector. We propose to apply log-concave method to smooth out this HD vector. Applying log-concave method to approximate the HD vector can significantly improve the smoothness, increase the stability and efficiency of the algorithm. We call our proposed algorithm as HD-LCD

algorithm.

6.2 Techniques review and background

6.2.1 Nominal categorical data set

Consider a data set $x = \{x_1, ..., x_n\}$, where n denotes the sample size. Each observation $x_i = \{x_{i1}, x_{i2}, \ldots, x_{iq}\}$ is a sequence of length q, where the jth element of the sequence is a categorical variable, taking on one of m_j possible levels. The data set is thus a collection of observations from the space Ω_q, with $|\Omega_q| = M = \prod_{i=1}^{q} m_i$. Formally we define $\Omega_q = \{\omega = (\omega_1, ..., \omega_q) \,|\, \omega_1 \in A_1, ..., \omega_q \in A_q\}$, where A_j has m_j elements and is the set of states of the jth attribute.

We now take zoo data as an example, which is posted on the UCI Machine Learning Repository (Lichman, 2013), created and donated by Richard Forsyth in 1990. The data were collected from 101 animals. Animals' 17 attributes were recorded, such as: hair, feather, eggs, milk. All attributes' value are nominal type. 101 animals were classified into 7 groups: Mammals, birds, reptiles, fish, amphibians, insects, and mollusks. In this case, $n = 101, q = 17$. The number of clusters is $K = 7$. The number of levels for all 17 attributes are: $\{101, 2, 2, 2, 2, 2, 2, 2, 2, 2, 2, 2, 2, 2, 6, 2, 2, 2\}$. The first attribute is animals' name, so each row (data point) has different value. This attribute

(column) will not be used for clustering, because it does not contribute any helpful information. For the remaining attributes, most of them are boolean type: "Yes" or "No". The 14th attribute "leg" has $m_{14} = 6$ levels, and the corresponding state set : $A_{14} = \{0, 2, 4, 5, 6, 8\}$, which indicates the number of legs. We can still treat it as nominal data in the sense of if animals have same number of legs. For zoo data, the sample space has size of $|\Omega_{17}| = \prod_{i=1}^{17} m_i = 19,857,408$.

6.2.2 Hamming distance vector and cluster center

Unlike how we measure distance for ordinal data, it is hard to define "distance" beween nominal data points. Following Zhang et al. (2006) we apply Hamming distance to measure distance between two nominal data points. Let $I()$ denote the indicator function, let $d(x_i, x_j) = \sum_{k=1}^{q} I(x_{ik} \neq x_{jk})$ denote the Hamming distance between two points x_i and x_j, which is equal to the number of attributes with different values. Hence it actually measure how different two nominal data points are. In the following example (Table 6.1), we take part of the zoo data. Animals are valued by four attributes: hair, legs, domestic, and breathes. They are all categorical nominal data. We can see that the similarity between antelope and calf is higher than that of antelope and carp. The Hamming distance between antelope and calf is one because there is only one attribute (Domestic) with different values; the Hamming distance

Table 6.1: Example of Hamming distance for zoo data.

Animal	**Hair**	**Legs**	**Domestic**	**Breathes**
antelope	Yes	4	No	Yes
calf	Yes	4	Yes	Yes
carp	No	0	Yes	No

between antelope and carp is four because all four attributes have different values.

Note that in a data set, for a given data point x_i, we can calculate Hamming distance between x_i and any points. The Hamming distances between them are from 0 to q. We can then derive a frequency vector, named as Hamming distance (HD) vector, denoted by $H(x_i) = \Big(H_0(x_i), H_1(x_i), ..., H_q(x_i)\Big)$, and $H_k(x_i) = \sum_{j=1}^{n} I\big(d(x_i, x_j) = k\big), k = 0, 1, 2, ..., q$. Hence HD vector is a vector of $q + 1$ elements, the kth element indicates the number of data points which have HD of $k - 1$ to the given point x_i. Also note that $\sum_{k=0}^{q} H_k(x_i) = n$. Here is an example: let $x = \{x_1 = (a, a, a), x_2 = (a, b, c), x_3 = (b, a, c), x_4 = (c, c, c), x_5 = (a, b, a)\}$. Then Hamming distance between points x_1 and x_2 is: $d(x_1, x_2) = 2$. The HD vector for $x_i : H(x_1) = (1, 1, 2, 1)$, where the first element "1" indicates there is one data point $\{x_1\}$ which has Hamming distance of 0 with x_1, and the third elements two indicates there are two data points $\{x_2, x_3\}$ which have Hamming distance of two to x_1.

Follow Zhang et al. (2006)[Definition 2], the position $c = (c_1, ..., c_q) \in \Omega_q$ is said to be the center of a data set $x = \{x_1, ..., x_n\}$ if it minimizes the sum of the Hamming

distance over all data points, namely

$$D(c,x) = \sum_{i=1}^{n} d(c,x_i),$$

where $c = \text{argmin}_{\omega \in \Omega_q} \sum_{i=1}^{n} d(\omega, x_i)$.

6.2.3 Uniform Hamming distance vector

Zhang et al. (2006) also provide a reference HD vector when no clustering pattern appear in the data set, so we can compare a potential center's HD vector with the reference HD vector case and determine the most likely center. When all points of the state space Ω_q are equally likely the HD vector is called the reference HD vector. We, here, will call it as the uniform Hamming distance (UHD) vector. It is denoted by $U(\Omega_q) = \{U_0(\Omega_q), U_1(\Omega_q), ..., U_q(\Omega_q)\}$. Intuitively, the UHD vector is a special case of HD vector. Given an uniformly distributed data $x = \{x_1, ..., x_n\}$, for an arbitrary position $\omega \in \Omega_q$, the corresponding HD vector $U(\omega)$ has $q+1$ elements. the kth element is (Zhang et al., 2006):

$$U_k(\omega) = \sum_{i=1}^{n} d(x_i, \omega), k = 0, 1, ...q,$$

which is the number of possible outcomes that have an exact distance k to the

position ω. Also note that the value of U_k does not depend on the position ω. By Zhang et al. (2006)[Theorem 3] the UHD vector has the form:

$$
\begin{aligned}
U_0(\Omega_q) &= \frac{n}{M}, \\
U_1(\Omega_q) &= \frac{n}{M}\{(m_1 - 1) + ... + (m_q - 1)\}, \\
U_2(\Omega_q) &= \frac{n}{M}\sum_{i<j}^{q}(m_i - 1)(m_j - 1), \\
&... \\
U_q(\Omega_q) &= \frac{n}{M}(m_1 - 1)(m_2 - 1)...(m_q - 1).
\end{aligned}
$$

Figure 6.1 is an example of UHD vector for sample space where $q = 25, m_1 = , ..., m_{25} = 2$. Note that UHD vector is symmetric when $m_1 =, ..., = m_q = 2$, but it is not true for other cases.

For a given point, we can calculate the Hamming distance between this point with all other points. Hence we have a sequence of Hamming distance. Then the HD vector is the frequency vector of those distances. The kth element of the vector counts the number of points which are exactly distance k away from the given point. If the given point is a center c, we expect a local bump in the HD vector representing all the points in the local cluster, followed by a dip. And a second bump for all the other points not in the local cluster. Figure 6.2 shows an example center's HD

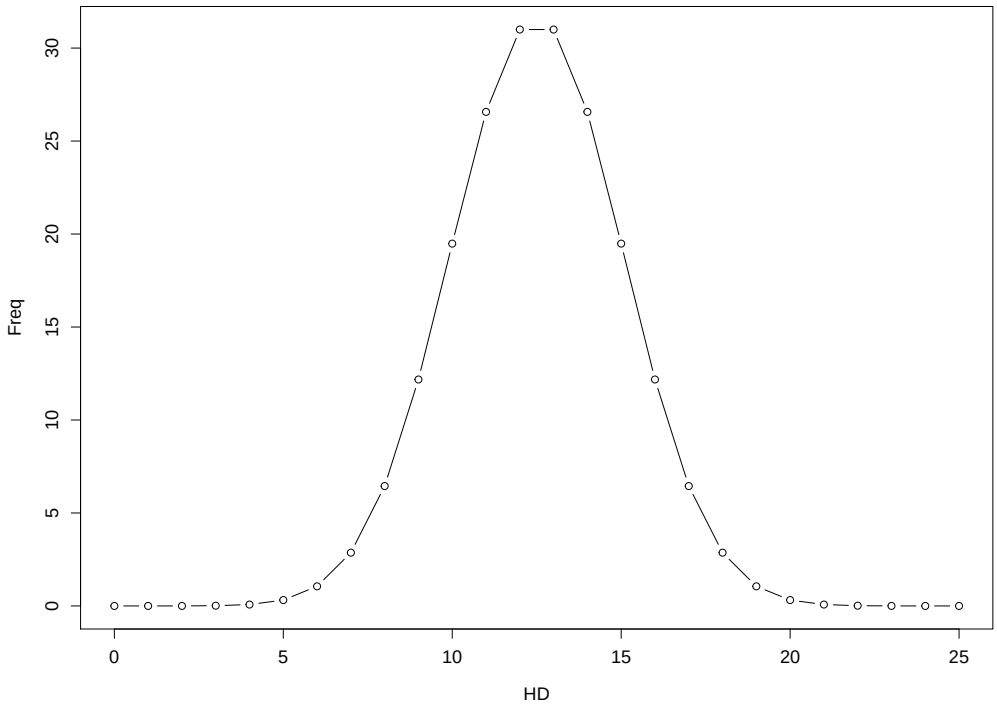


Figure 6.1: UHD with 25 binary attributes

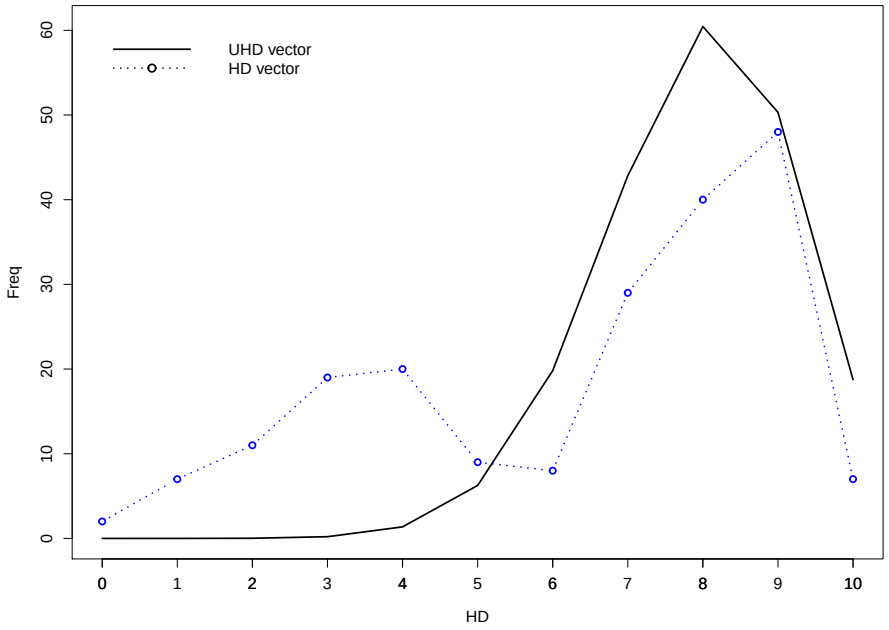


Figure 6.2: We take one sample (n = 200) which is simulated by original simulation method, we choose one of the simulated center, compare the center's HD vector against the corresponding UHD vector.

vector against UHD vector from one of our simulated data set, which is simulated with the original simulation method (will be illustrated in later chapter). Clear bimodal pattern is observed. Hence it becomes possible to identify potential cluster center by comparing the bimodal HD vector with unimodal UHD vector. Note that in this case the UHD vector is not symmetric, unlike the previous example.

6.2.4 HD vector algorithm

The goal of the algorithm is to find a collection of clusters $C_1, \ldots, C_K$ such that

$$x = C_1 \uplus C_2 \uplus \ldots \uplus C_K,$$

where $\uplus$ indicates the disjoint union. Notably, the number of clusters K is not assumed to be known apriori. Each cluster C_k is defined by a cluster center c_k, and a radius R_k^*, where $C_k = \{x_j \in x : d(c_k, x_j) \leq R_k^*\}$. The goal of the algorithm is thus to find the cluster centers $c_1, \ldots, c_K$ and radius $R_1^*, \ldots, R_K^*$.

Zhang et al. (2006) state if each observation in the data set has an equal probability of locating at any position in the space Ω_q (no clear cluster pattern appear in data set), then the resulted UHD vector has a typical unimodal pattern distribution. While for a data set with significant cluster pattern, the cluster center has a bimodal pattern distribution. This provides a way to detect the cluster's center: a point which HD vector is the "farthest" to the UHD vector might be the cluster center. Note that we are mostly interested in the first bump of the HD vector, which reflects the local cluster and involve most relevant information. The effect of the second bump should be minimized to avoid the "noise". Therefore there are two things we need to consider:

1. where to separate two bumps,

2. How to calculate the distance between selected point's HD vector and the UHD vector.

For a given position $\omega \in \Omega_q$, to find the location where to separate the two peaks of $H(\omega)$, the HD vector algorithm version cutoff is defined as (Zhang et al., 2006)

$$r^*_\omega(t) \quad = \quad \min_{i>0}\{i|H_i(\omega) < U_i(\Omega_q)\} - t,$$

where t is a tuning constant to tune the upper cutting edge of first bump, t = 1,...,T, for a suitable T. When $t = 1$, the cutoff is the maximum value of the range on which the HD vector is larger than the UHD vector. For example, in Figure 6.2 $r^*_\omega(1) = 5$. The most optimal cutoff point is chosen by minimizing the p-value of the modified chi-squared statistic described later.

$$r^*_\omega \quad = \quad \text{argmin}_{r^*_\omega(t)}\{p\text{-value of } \chi^2_M(r^*_\omega(t), \omega)\}, \tag{6.1}$$

For a given position $\omega \in \Omega_q$, to compute the distance between its HD vector to the UHD vector. Zhang et al. (2006) applied the Pearson's chi-squared statistics, which

is one of the common way to evaluate the discrepancy of two frequency distribution.

$$\chi^2(\omega) = \sum_{i=0}^{q} \frac{(H_i(\omega) - U_i(\Omega_q))^2}{U_i(\Omega_q)}.$$

Intuitively, chi-squared statistics evaluate the discrepancy of two frequency vectors by adding up the normalized difference of each element. Motivated by minimizing the effect from the second bump, they separate it into two parts using HD vector algorithm version cutoff r^*_ω. For the second bump, the value of $H_{r^*_\omega+1}(\omega), ..., H_q(\omega)$ are reallocated to minimize the difference, the resulted statistic is called as the modified chi-squared statistic:

$$\chi^2(r^*_\omega, \omega) = \sum_{i=0}^{r^*_\omega} \frac{(H_i(\omega) - U_i(\Omega_q))^2}{U_i(\Omega_q)} + \min_{H_i(\omega)} \sum_{i=r^*_\omega+1}^{q} \frac{(H_i(\omega) - U_i(\Omega_q))^2}{U_i(\Omega_q)}.$$

For a given position $\omega \in \Omega_q$, Zhang et al. (2006)[Theorem 5] state that the modified chi-squared statistic takes the form

$$\chi^2_M(r^*_\omega, \omega) = \sum_{i=0}^{r^*_\omega} \frac{(H_i(\omega) - U_i(\Omega_q))^2}{U_i(\Omega_q)} + \frac{(\sum_{i=0}^{r^*_\omega} H_i(\omega) - \sum_{i=0}^{r^*_\omega} U_i(\Omega_q))^2}{\sum_{i=r^*_\omega+1}^{q} U_i(\Omega_q)}. \tag{6.2}$$

Without theoretical support, Zhang et al. (2006) stated that the modified chi-squared statistic follows the chi-squared distribution with degrees of freedom $r^*_\omega + 1$. Hence

they use the upper tail critical value of chi-squared distribution $\chi^2_\alpha(v)$ to determine the significance of the cluster, where $P(X > \chi^2_\alpha(v)) = \alpha$, and X is a chi-squared random variable with v degrees of freedom. Significant cluster exists if the modified chi-squared statistic is larger than $\chi^2_\alpha(r^*_\omega + 1)$. α is set to 0.05 in their simulation and data analysis. Note that Zhang et al. (2006) choose cluster center from set $\mathcal{D}_q = \cup_{x_i \epsilon x}\{\omega \in \Omega_q : d(x_i, \omega) \leq 1\}$. Which is an augmentation of the original data set to include also all nearest neighbours. Point with the largest modified chi-squared statistic value is chosen as cluster center, if significant cluster exists.

After the cluster center is chosen, the next step is to determine the member of cluster. The cluster radius for a given cluster center c_k is defined as:

$$R^*_k = \min_{0<j<q}\{j \,|\, H_j(c_k) < \min(H_{j-1}(c_k), H_{j+1}(c_k))\} - 1, \tag{6.3}$$

which is the first local minimum of the HD vector distribution. We also call it the first local minimum (FLM).

The algorithm proceeds the following iterating procedure to detect all significant clusters.

1. Is there a cluster in the remaining data set?

2. If yes, the algorithm finds the cluster, and removes it from the current data

set. The algorithm then returns to step one.

3. If no, all of the remaining/isolated points in the data set are re-categorized to one of the other existing clusters, and the algorithm terminates.

The detailed HD vector algorithm is shown in Algorithm 2. Notably, the HD vector algorithm much depends on HD vector. It is involved in two essential parts of the algorithm: it is used to calculate the modified chi-squared statistic ("distance" between HD vector and UHD vector), therefore to determine cluster center; it is also used to determine cluster radius (FLM). For the above two essential parts, HD vector algorithm applies the empirical distribution of HD vector in both cases, it may decrease the algorithm accuracy due to the un-smoothness of empirical distribution. We will show this in later chapter.

Algorithm 2 HD vector algorithm, input $x = \{x_1, ..., x_n\}$

Compute $\mathcal{D}_q$.
Compute the UHD vector $U_0(\Omega_q), ..., U_q(\Omega_q)$.
Set the list $l_{center} = NULL$, cluster counter $k = 0$, initial $\chi^2 = +\infty, r^*_{c_0} = 0$.
while $\chi^2 \geq \chi^2_{0.05}(r^*_{c_k} + 1)$ **do**
    for each position $\omega : \omega \in \mathcal{D}_q$ **do**
        Compute the HD vector $H(\omega)$.
        Compute the cutoff r^*_ω using (6.1).
        if $r^*_\omega = 0$ **then**
            $\chi^2_M(r^*_\omega, \omega) = 0$, and label ω as an isolated position.
        else
            Compute $\chi^2_M(r^*_\omega, \omega)$, using (6.2)
        end if
    end for
    Set $\chi^2 = \max_{\omega \in \mathcal{D}_q} \chi^2_M(r^*_\omega, \omega)$, $c_k = \text{argmax}_{\omega \in \mathcal{D}_q} \chi^2_M(r^*_\omega, \omega)$
    if $\chi^2 \geq \chi^2_{0.05}(r^*_{c_k} + 1)$ **then**
        $k = k + 1$.
        $l_{center} = l_{center} \cup \{c_k\}$.
        Compute radius R^*_k using (6.3).
        $C_k = \{x_i \in x \,|\, d(x_i, c_k) \leq R^*_k\}$.
        Set $x : x = x \setminus C_k$.
        Set $\mathcal{D}_q : \mathcal{D}_q = \mathcal{D}_q \setminus \{\omega \in \mathcal{D}_q \,|\, d(\omega, c_k) \leq R^*_k\}$.
    end if
end while

6.2.5 K-modes algorithm

Another clustering algorithm we used to compare is the K-modes algorithm. It is a well-known clustering algorithm for categorical data, proposed by Huang (1998). The K-modes algorithm is an extension of K-means algorithm. The goal of K-modes algorithm is to partition a data set into K clusters. It is also an algorithm based on the Hamming distance (dissimilarity measure). Let's define a cluster centers (cluster modes) set $C = \{c_1, ..., c_K\}$, where $c_i = \{c_{i1}, ..., c_{iq}\}$ is a center, and $c_i \in \Omega_q$. A $n \times K$ partition matrix is defined as:

$$\Lambda = \begin{bmatrix} \lambda_{11} & \lambda_{12} & \dots & \lambda_{1K} \\ \lambda_{21} & \lambda_{22} & \dots & \lambda_{2K} \\ \vdots & \vdots & \ddots & \vdots \\ \lambda_{n1} & \lambda_{n2} & \dots & \lambda_{nK} \end{bmatrix}.$$

For a given data set $x = \{x_1, ..., x_n\}$, the K-modes algorithm is to solve the opti-

mization problem:

$$\text{minimize: } P(\Lambda, C) = \sum_{l=1}^{K}\sum_{i=1}^{n}\sum_{j=1}^{q}\lambda_{il}I(x_{ij}, c_{lj})$$

$$\text{subject to: } \lambda_{ik} \in \{0,1\}, 1 \le i \le n, 1 \le k \le K,$$

$$\sum_{k=1}^{K}\lambda_{ik} = 1, i = 1, ...K,$$

where $P(\Lambda, C)$ is called the cost function. Briefly the K-modes algorithm aims to find the partition of the data, which minimizes the within cluster Hamming distance between cluster members and cluster center. For a given cluster $C_i, i = 1, ..., K$, Let $c_i = \{c_{i1}, ..., c_{iq}\}$ be its center, then

$$c_{ij} = \text{argmax}_{a \in A_j} \sum_{x_l \in C_i} I(x_{lj} = a), j = 1, ...q. \tag{6.4}$$

To determine the value of the jth attribute of cluster center, we count the different state's frequency for all cluster members' jth attribute. The one with largest relative frequency is chosen for c_{ij}.

Briefly, the K-modes algorithm follows the steps:

1. Randomly initial K cluster centers (modes).

2. Each data point is assigned to a cluster which center is closed to it in the

Hamming distance sense.

3. After all data points are allocated, compute and determine the center for all clusters following (6.4).

4. Repeat steps 2-3 until no data point has changed clusters.

The K-modes algorithm can not determine the number of clusters automatically. Manually setting up the number of cluster is required before user runs the algorithm. This is a big problem when user have no idea how many clusters there are in data set. Also K-modes algorithm heavily depends on the start seeds chosen, different orders may lead to significantly different clustering result, hence the clustering result is not unique for one data set.

6.3 Outline

In Part II, we apply log-concave MLE ($d = 1$) to clustering methodology and discuss the performance.

In Chapter 7, we define the log-concave mixture model. An example of log-concave mixture model is given in this chapter. We also describe how to compute the model and how to determine cutoff of two modes in a bimodal pattern distribution. In Chapter 8 we discuss the details of our proposed algorithm, and illustrate why our

model improve the smoothness and increase the accuracy. In Chapter 9, we show the comparison of our proposed algorithm with HD vector algorithm and K-modes algorithm through simulated data and real data.

Chapter 7

Log-concave mixture model

Log-concavity is an appealing shape constraint in distribution modeling. Comparing to parametric and nonparametric methods, log-concavity provides a good trade off between robustness and efficiency. This is because log-concavity is a natural shape constraint. A broad range of parametric models are log-concave, such as, normal, uniform, gamma(r, λ) with $r \geq 1$, beta(a, b) with $a \geq 1, b \geq 1$ in continuous setting, and multinomial, negative multinomial, multivariate hypergeometric in discrete setting (in term of eLC definition). Hence the log-concave class is a much broader class than a specific parametric model, which therefore reduce the bias. On the other hand, comparing with the nonparametric method, adding log-concave shape constraint to estimation improves the efficiency and reduce variance. Another advantage of the log-concavity is that no arbitrary choices such as bandwidth, kernel or prior are

involved in the estimation. Hence, the log-concave model can provide more local information without the effect of the bandwidth.

Recall the definition of a discrete log-concave PMF. Let $p(z) : \mathbb{Z} \to [0,1]$ denote a PMF, where $\mathbb{Z}$ denotes the integers set $\{\ldots, -2, -1, 0, 1, 2, \ldots\}$. Then p is said to be log-concave if for any $z \in \mathbb{Z}$

$$(\triangle h)(z) \quad = \quad h(z-1) - 2h(z) + h(z+1) \quad \geq \quad 0,$$

where $h(z) = -\log p(z)$. The notation above $(\triangle h)$ denotes the discrete Laplacian operator, which can also be expressed as $(\triangle h)(z) = \{h(z+1) - h(z)\} - \{h(z) - h(z-1)\}$. This is the second difference of the function h, and hence this definition matches well with the one in the continuous setting.

Given independent and identically distributed observations $z_1, ..., z_n \in \mathbb{Z}$, Balabdaoui et al. (2013) provides an algorithm to compute the MLE under the assumption that the distribution is log-concave. The log-concave MLE is defined as $\widehat{p}_n = \text{argmax}_{p \in \mathcal{P}_1} \frac{1}{n} \sum_{i=1}^{n} \log p(z_i)$, where $\mathcal{P}_1$ indicates the set of all log-concave probability mass functions on $\mathbb{Z}$. It is theoretically proved to exist and be unique.

In a data set with n points $x = \{x_1, ..., x_n\}$, for a given point $x_i \in x$, recall that the HD vector is defined as $H(x_i) = \Big(H_0(x_i), H_1(x_i), ..., H_q(x_i)\Big)$, and $H_k(x_i) = \sum_{j=1}^{n} I\big(d(x_i, x_j) = k\big), k = 0, 1, 2, ..., q$. We further define the empirical Hamming dis-

tance distribution via the formula:

$$\overline{p}_{x_i}(k) = H_k(x_i)/n,\ k = 0, ..., q.$$

It is the empirical distribution of Hamming distances to all points of a data set from a given point x_i. The support of this empirical Hamming distance distribution is $\{0, ..., q\}$. As for the UHD vector, we also define the uniform Hamming distance distribution $\overline{p}_u(k) = \frac{U_k(\Omega_q)}{n}, k = 0, ..., q$. It can also be expressed as:

$$\begin{aligned}
\overline{p}_u(0) &= 1/M, \\
\overline{p}_u(1) &= \{(m_1 - 1) + ... + (m_q - 1)\}/M, \\
\overline{p}_u(2) &= \{\sum_{i<j}^{q}(m_i - 1)(m_j - 1)\}/M, \\
&... \\
\overline{p}_u(q) &= \{(m_1 - 1)(m_2 - 1)...(m_q - 1)\}/M.
\end{aligned}$$

This is also a distribution on $\{0, \ldots, q\}$. Each value $\overline{p}_u(k)$ represents the expected value of $\overline{p}_{x_i}(k)$ assuming that each observation in the data set has an equal probability to appear at any position in the space Ω_q. Note that the UHD distribution does not depend on either the sample size n or the location x_i.

We develop the log-concave mixture model to approximate a discrete bimodal distribution. Let $\mathcal{P}_q$ denote the class of log-concave PMFs on $\{0,\ldots,q\}$. Let $\mathcal{P}_m = \{p : p = \alpha p_1 + (1-\alpha)p_2, p_1, p_2 \in \mathcal{P}_q, \alpha \in (0,1)\}$ denote the class of mixtures of log-concave distributions. The mixture log-concave projection is the distribution defined by

$$\begin{aligned} \widehat{p}_{x_i} &= \operatorname{argmin}_{g\in\mathcal{P}_m} \sum_{k=0}^{q} \overline{p}_{x_i}(k) \log\left(\frac{\overline{p}_{x_i}(k)}{g(k)}\right) \\ &= \operatorname{argmax}_{g\in\mathcal{P}_m} \prod_{k=0}^{q} \{g(k)\}^{\overline{p}_{x_i}} \end{aligned} \tag{7.1}$$

The mixture log-concave projection $\widehat{p}_{x_i}$ is computed by minimizing the KL divergence between any mixture of log-concave distributions and the empirical Hamming distance distribution $\overline{p}_{x_i}$, hence it is the projection of the empirical Hamming distance distribution $\overline{p}_{x_i}$ onto $\mathcal{P}_m$ in the KL divergence sense. It is the "closest" log-concave mixture distributions to the empirical Hamming distance distribution $\overline{p}_{x_i}$, and intends to provide a smoother approximation than the empirical distribution. Note that the support of the mixture log-concave projection is $\{0,...,q\}$. Consider the following example: the HD vector of position $\omega : H(\omega) = \{1,4,3,2,5,2\}$ represents the frequency vector of set $A = \{0,1,1,1,1,2,2,2,3,3,4,4,4,4,4,5,5\}$. In this case $q = 5$ and the support is $\{0,1,2,3,4,5\}$. Given the observations in set A, we apply

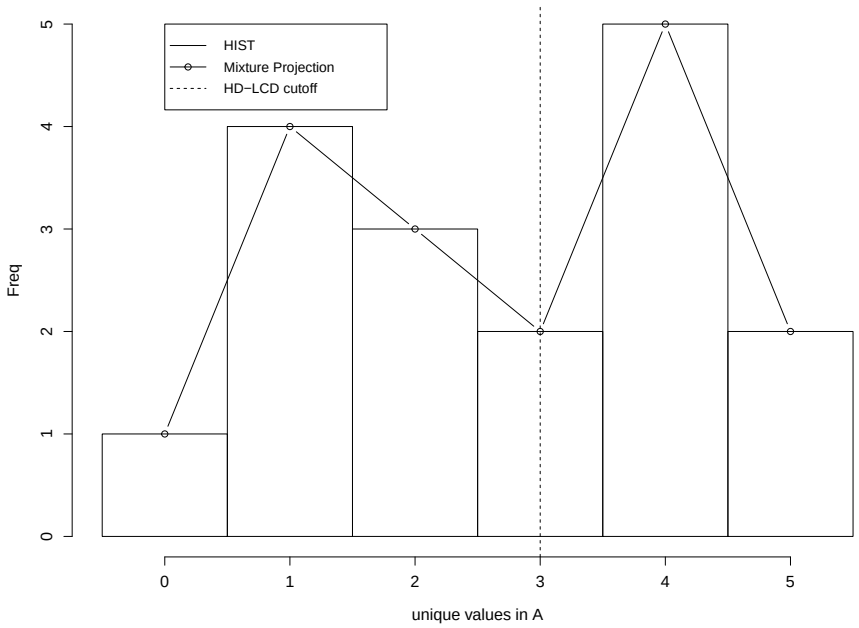


Figure 7.1: Histogram plot against the mixture log-concave projection $\widehat{p}_\omega$. Simulated from set $A = \{0, 1, 1, 1, 1, 2, 2, 2, 3, 3, 4, 4, 4, 4, 4, 5, 5\}$.

the log-concave mixture model to approximate the distribution. Figure 7.1 compares the histogram and the mixture log-concave projection for above example.

The projection $\widehat{p}_{x_i}$ is computed using the EM algorithm (Dempster et al., 1977) and the R package `logcondiscr`. We apply naive Bayes classifier (Rish, 2001) to find the cutoff of two modes. Consider the log-concave mixture projection $\widehat{p}_{x_i}$ and write it in terms of its components $\widehat{\alpha}\widehat{p}_1 + (1 - \widehat{\alpha})\widehat{p}_2$, where we now drop the dependence on x_i in the notation. Let $\widehat{\alpha}_1 = \widehat{\alpha}$ and $\widehat{\alpha}_2 = 1 - \widehat{\alpha}$. The mixture model represents the overall population as a mixture of two sub-populations. Let S_1 and S_2 be the

sub-populations corresponding to $\widehat{p}_1$ and $\widehat{p}_2$, respectively. For a point $k \in \{0, \ldots, q\}$, the probability that k belongs to population $S_i (i = 1, 2)$ is defined as $pr(S_i|k) = \frac{pr(S_i) * pr(k|S_i)}{pr(k)}$, where $pr(A)$ is the probability of event A. Since $pr(k)$ is identical in both populations, we can get rid of it and define the Bayes discriminant functions $D_i(k) = pr(S_i) * pr(k|S_i), i = 1, 2$. Hence the Bayes classifier is defined as

$$B(k) = \text{argmax}_{i=1,2} \{D_i(k)\}.$$

In our case, $D_i(k) = pr(S_i) * pr(k|S_i) = \widehat{\alpha}_i \widehat{p}_i(k)$. Hence the Bayes classifier in our mixture model is $B(k) = \text{argmax}_{i=1,2} \{\widehat{\alpha}_i \widehat{p}_i(k)\}$. Then, the "cut-off" value of r between the two regions can be defined by

$$r = \min \left\{ \max_{B(k)=1} \{k\}, \max_{B(k)=2} \{k\} \right\}.$$

Let us assume that the notation is chosen in such a way that k satisfying $B(k) = 1$ is smaller than those k satisfying $B(k) = 2$. Then, we formally define the HD-LCD version cut-off as

$$r_{x_i} = r = \min_{0 \le k \le q} \{k : \widehat{\alpha}_1 \widehat{p}_1(k) \le \widehat{\alpha}_2 \widehat{p}_2(k)\} - 1. \tag{7.2}$$

Note that the above definition chooses a convention to resolve any possible ties where $\widehat{\alpha}_1 \widehat{p}_1(k) = \widehat{\alpha}_2 \widehat{p}_2(k)$. One example is illustrated in Figure 7.1, the HD-LCD version cutoff $r_\omega = 3$, which perfectly separates two modes.

Chapter 8

Algorithm

8.1 Modified reversed KL divergence

For two probability distributions p_0 and p, KL divergence (Kullback and Leibler, 1951) is used to measure the information gain/loss when we use one probability distribution p to approximate another probability distribution p_0, which is usually the true distribution. The KL divergence defined on $\mathbb{N}_0 = \{0, 1, 2, 3...\}$ can be expressed as.

$$\rho(p\|p_0) = -\sum_{z\epsilon\mathbb{N}_0} p_0(z) \log \frac{p(z)}{p_0(z)}.$$

In our case, we are trying to compare potential cluster centers' HD vector to

UHD vector, such that we can detect the most likely cluster center. Recall that KL divergence is not symmetric, we need to consider which version to apply, the common version: KL divergence from selected point's HD vector to UHD vector? Or the reversed version: KL divergence from UHD vector to selected point's HD vector. In practice, particularly when working on simulated data, we found that common version KL divergence could not identify cluster center well, but the reversed version is doing a much better job. The reason is, as we described in previous chapter, we want to focus on the information of first bump but minimize the "noise" from 2nd bump. We now list the two versions of KL divergence: the common version, $\rho_{KL}(p\|p_0)$, and the reversed version $\rho_{RKL}(p_0\|p)$.

$$\rho_{KL}(p\|p_0) = -\sum_{z\epsilon\mathbb{N}_0} p_0(z)\log\frac{p(z)}{p_0(z)},$$
$$\rho_{RKL}(p_0\|p) = -\sum_{z\epsilon\mathbb{N}_0} p(z)\log\frac{p_0(z)}{p(z)}.$$

Compare these two formulas, the log term: $\log\frac{p(z)}{p_0(z)}$ and $\log\frac{p_0(z)}{p(z)}$ are actually the difference between two probability distributions in log scale. The multiplication factor $p_0(z)$ or $p(z)$ is essential, it can be taken as a weight term. Recall that we want to boost the information contained in the first bump, but eliminate the noise from the second bump. It is more appropriate that the weight term has relative large

value for first bump, but small value for the remaining bumps. As showed in Figure 6.2, obviously PMF of unimodal pattern UHD does not satisfy our situation well. As a weight term, the peak of UHD will boost the difference ("noise") between estimator and true PMF at the second bump. But bimodal pattern HD vector works better as the weight term in this case. So we determined to apply reversed KL divergence to our algorithm, which can be explicitly expressed as:

$$\rho_{RKL}(\overline{p}_u \| \widehat{p}) = -\sum_{k=0}^{q} \widehat{p}(k) \log \frac{\overline{p}_u(k)}{\widehat{p}(k)},$$

where $\widehat{p}$ is the mixture log-concave projection of a potential center's HD vector, and $\overline{p}_u$ indicates the distribution of UHD. We drop the dependence on x_i in the notation in the following content.

To minimize the "noise" from the second bump, we separate the sum term of reversed KL divergence into two parts:

$$\rho_{RKL}(\overline{p}_u \| \widehat{p}) = -\sum_{k=0}^{r} \widehat{p}(k) \log \frac{\overline{p}_u(k)}{\widehat{p}(k)} - \sum_{k=r+1}^{q} \widehat{p}(k) \log \frac{\overline{p}_u(k)}{\widehat{p}(k)},$$

where r is the cutoff to separate two bumps. The following theorem provides us the explicit form of modified reversed KL divergence (MRKL), which we used to evaluate the distance between two distributions.

Theorem 8.1.1. *Let $p_{r+1}, ..., p_q \in [0,1]$. Minimize the effect of $r+1$ to q summation terms of the reversed KL divergence is equivalent to*

$$\begin{aligned} &\textit{Maximize: } \sum_{k=r+1}^{q} p_k \log \frac{\overline{p}_u(k)}{p_k} \\ &\textit{Subject to: } \sum_{k=r+1}^{q} p_k = \sum_{k=r+1}^{q} \widehat{p}(k). \end{aligned}$$

The modified reversed KL divergence takes the form:

$$\rho^*_{RKL}(\overline{p}_u \| \widehat{p}, r) = -\sum_{k=0}^{r} \widehat{p}(k) \log \frac{\overline{p}_u(k)}{\widehat{p}(k)} - \sum_{k=r+1}^{q} \widehat{p}(k) \log \frac{\sum_{k=r+1}^{q} \overline{p}_u(k)}{\sum_{k=r+1}^{q} \widehat{p}(k)}.$$

Proof. The reversed KL divergence from the baseline PMF $\overline{p}_u$ to a given PMF $\widehat{p}$ is

$$\begin{aligned} \rho_{RKL}(\overline{p}_u \| \widehat{p}, r) &= -\sum_{k=0}^{q} \widehat{p}(k) \log \frac{\overline{p}_u(k)}{\widehat{p}(k)} \\ &= -\sum_{k=0}^{r} \widehat{p}(k) \log \frac{\overline{p}_u(k)}{\widehat{p}(k)} - \sum_{k=r+1}^{q} \widehat{p}(k) \log \frac{\overline{p}_u(k)}{\widehat{p}(k)} \end{aligned}$$

We re-write the optimization problem as the following:

$$\begin{aligned} &\text{Maximize: } f(p_{r+1}, ..., p_q) = \sum_{k=r+1}^{q} p_k \log \frac{\overline{p}_u(k)}{p_k} \\ &\text{Subject to: } g(p_{r+1}, ..., p_q) = \sum_{k=r+1}^{q} p_k - l = 0, \text{ where } l = 1 - \sum_{k=0}^{r} \widehat{p}(k). \end{aligned}$$

Let λ be the Lagrange multiplier, and Lagrange function $L = f - \lambda g$. Follow the Lagrange optimization method, we have

$$\frac{\partial L}{\partial p_k} = \log \overline{p}_u(k) - \log p_k - 1 - \lambda = 0, \ k = r+1, ..., q.$$

Hence,

$$\log p_k = \log \overline{p}_u(k) - \lambda - 1,$$

$$p_k = \overline{p}_u(k) \exp\{-(\lambda + 1)\}, \ k = r+1, ..., q.$$

Plug in the above equations to $g(p_{r+1}, ..., p_q)$, we have

$$\exp\{-(\lambda + 1)\} \sum_{k=r+1}^{q} \overline{p}_u(k) - l = 0.$$

So $\exp\{-(\lambda + 1)\} = \frac{l}{\sum_{k=r+1}^{q} \overline{p}_u(k)}$, and

$$p_k = \frac{\overline{p}_u(k) * l}{\sum_{k=r+1}^{q} \overline{p}_u(k)} = \frac{\overline{p}_u(k)\left(1 - \sum_{k=0}^{r} \widehat{p}(k)\right)}{\sum_{k=r+1}^{q} \overline{p}_u(k)} = \frac{\overline{p}_u(k) \sum_{k=r+1}^{q} \widehat{p}(k)}{\sum_{k=r+1}^{q} \overline{p}_u(k)}.$$

Modified reversed KL divergence:

$$\begin{aligned}\rho^{*}_{RKL}(\overline{p}_u \| \widehat{p}, r) &= -\sum_{k=0}^{r} \widehat{p}(k) \log \frac{\overline{p}_u(k)}{\widehat{p}(k)} - \sum_{k=r+1}^{q} \frac{\overline{p}_u(k) \sum_{k=r+1}^{q} \widehat{p}(k)}{\sum_{k=r+1}^{q} \overline{p}_u(k)} \log \frac{\overline{p}_u(k) \sum_{k=r+1}^{q} \overline{p}_u(k)}{\overline{p}_u(k) \sum_{k=r+1}^{q} \widehat{p}(k)} \\ &= -\sum_{k=0}^{r} \widehat{p}(k) \log \frac{\overline{p}_u(k)}{\widehat{p}(k)} - \sum_{k=r+1}^{q} \widehat{p}(k) \log \frac{\sum_{k=r+1}^{q} \overline{p}_u(k)}{\sum_{k=r+1}^{q} \widehat{p}(k)}.\end{aligned}$$

□

8.2 Test cluster pattern using bootstrap

Zhang et al. (2006) determined the significance of cluster pattern with a hypothesis test, their test statistic is the modified chi-squared statistic. In our proposed algorithm, the significance of the cluster pattern is determined using a hypothesis test with test statistic equal to the largest (over all elements of the data set) modified reversed KL divergence (MRKL). The distribution of the test statistic is difficult to compute, so a nonparametric bootstrap procedure is used. The unusual idea behind our hypothesis test is that we reverse the role of the typical null and alternative hypothesis test. Our Null hypothesis states that significant cluster pattern exists, while alternative hypothesis states that no cluster pattern exists. This hypothesis setting corresponds to the reversed version of KL divergence, which we are measuring the distance from UHD vector to chosen point's HD vector. We accept the null

hypothesis if the observed p-value is larger than 0.05.

H_0 : Cluster pattern exists in the data.

H_a : Cluster pattern does not exist in the data.

To implement the test, the idea is that we assume the given data has cluster pattern, hence we create testing samples under "cluster" assumption by randomly drawing points from the given data set with replacement. The testing sample has the same sample size of the given data set. We then calculate the largest MRKL of these testing samples. The testing samples' largest MRKL approximate the behavior of the test statistic. We describe the testing process as following.

1. For a given data set x, we create testing samples with the same sample size of x, the testing samples are created by randomly drawing points from x with replacement.

2. Calculate the MRKL for each point of the testing sample created in step 1. We denote the largest MRKL as ρ_i^*, where i is the index of the testing sample.

3. Repeat step 1-2 B=100 times, then we get a sequence of the largest MRKL under "cluster" assumption: $\rho_1^*, ..., \rho_{100}^*$. This is used to determine if the cluster pattern appear in the given data set x.

4. Let ρ^* be the largest MRKL of the given data set x. The cluster pattern

does exist if the p-value is greater than $\alpha = 0.05$, where p-value is defined as $\frac{1}{B+1}\left(\sum_{i=1}^{B} I(\rho_i^* < \rho^*) + 1\right)$.

8.3 HD-LCD algorithm with the bootstrap

The goal of the HD-LCD algorithm is similar to the HD vector algorithm: automatically detect and find a collection of clusters $C_1, \ldots, C_K$ from a given data set x. We still use $c_1, ..., c_K$ to denote cluster centers, $R_1, ..., R_K$ to denote cluster radius. A cluster is defined as $C_k = \{x_j \in x : d(c_k, x_j) \leq R_k\}$, $k = 1, ..., K$. The process of our algorithm is similar to that of the HD vector algorithm:

1. Is there a cluster in the remaining data set?

2. If yes, the algorithm finds the cluster, and removes it from the current data set. The algorithm then returns to step one.

3. If no, all of the remaining points in the data set are re-categorized to one of the other existing clusters, and the algorithm terminates.

Comparing with HD vector algorithm, there are three main differences between them.

1. For a given data set, the way to determine the significance of cluster pattern. Zhang et al. (2006) tested the significance using test statistic of chi-squared

statistic. In our proposed algorithm, it is done using the largest reversed KL divergence.

2. The range where to choose the cluster center. Zhang et al. (2006) chose center from an augmented set $\mathcal{D}_q = \cup_{x_i \in x}\{\omega \in \Omega_q : d(x_i, \omega) \leq 1\}$. We tried to include the neighbours in our algorithm, but it does not improve the efficiency of our algorithm. On the other hand, it slows down algorithm significantly. We hence restrict our algorithm only on all points of x.

3. The way to determine cluster radius. Zhang et al. (2006) determined the cluster radius by the first local minimum of the center's HD vector. For the kth cluster, we define the cluster radius as $R_k = r_{c_k} - 2$. Note that c_k is the cluster center, and r_{c_k} is the HD-LCD version cutoff of the cluster center which separate two bumps of HD vector distribution. We define the radius as two less than the cutoff, because we believe the edge of the first bump is not relevant to the local cluster.

We also note that it is possible for the algorithm to find no clusters in the data set. The detailed algorithm proceeds as shown on the following page (Algorithm 3). Note that if the data set x has an attribute $j \in \{1, \ldots, q\}$ such that the values of $\{x_{ij}, i = 1, \ldots, n\}$ are either all the same or all different (the latter is, of course, less common), then we remove these attributes from the data set when applying the

algorithm.

We have not theoretically studied the consistency of the bootstrap test process. In later chapter, we will provide two simulated data examples to illustrate the validity of the test process. We apply the HD-LCD algorithm with the bootstrap to two simulated data sets. We will show the cluster test process works well. Unfortunately, the cluster pattern testing process is very time consuming. Roughly, for $B = 100$, calculating the largest MRKL of testing samples cost $100 * K$ times as long, where K is the number of clusters. To improve the time cost, we remove the cluster pattern test process. The algorithm stops when a small number of points remaining in the data set. Usually, we set the threshold as a specific number, or the ratio of number of remaining points over original sample size. For example, in our simulations, the algorithm stops when the ratio is less than 0.1. Our simulation study shows that the HD-LCD algorithm without the bootstrap also works well. Following is the detailed algorithm (Algorithm 4).

Algorithm 3 HD-LCD algorithm with the bootstrap, input $x_0 = \{x_1, ..., x_n\}$

Set $x = x_0$.
Compute the UHD distribution $\overline{p}_u$.
Set the list $l_{center} = NULL$, cluster counter $k = 0$, remaining sample size $l = n$, p-value $= 1$.
while p-value ≥ 0.05 **do**
  for each point of the data set x: $x_i, i = 1, 2, ..., l$. **do**
    Compute the empirical Hamming distance distribution $\overline{p}_{x_i}$.
    Compute its mixture log-concave projection $\widehat{p}_{x_i}$, see (7.1).
    Compute the cutoff value r_{x_i} using (7.2).
    Compute $\rho^*_{RKL}(\overline{p}_u \| \widehat{p}_{x_i}, r_{x_i})$.
  end for
  Set $\rho^* = \max_{i=1,...,l} \rho^*_{RKL}(\overline{p}_u \| \widehat{p}_{x_i}, r_{x_i})$, $c_k = x_{i^*}$ where $i^* = \text{argmax}_{i=1,...,l} \rho^*_{RKL}(\overline{p}_u \| \widehat{p}_{x_i}, r_{x_i})$, and set $R_k = r_{x_{i^*}} - 2$.
  for b = 1,..., B **do**
    Assuming all points in x is equally likely, create testing sample x^* by randomly select l points from x with replacement.
    for each point of $x^* : x^*_j, j = 1, ..., l$ **do**.
      Compute the empirical Hamming distance distribution $\overline{p}_{x^*_j}$.
      Compute its mixture log-concave projection $\widehat{p}_{x^*_j}$, and the cutoff value $r_{x^*_j}$, see (7.1) and (7.2) .
      Compute $\rho^*_{RKL}(\overline{p}_u \| \widehat{p}_{x^*_j}, r_{x^*_j})$.
    end for
    Set $\rho^*_b = \max_{j=1,...,l} \rho^*_{RKL}(\overline{p}_u \| \widehat{p}_{x^*_j}, r_{x^*_j})$,
  end for
  p-value $= (\{\sum_{b=1}^{B} I(\rho^*_b < \rho^*)\} + 1)/(B + 1)$.
  if p-value ≥ 0.05 **then**
    $k = k + 1$.
    $l_{center} = l_{center} \cup \{c_k\}$.
    $C_k = \{x_i \in x \,|\, d(x_i, c_k) \leq R_k\}$.
    Set $x : x = x \setminus C_k$, and update the sample size l for x.
  end if
end while
if k ≥ 1 **then**
  for each point of x_0: $x_i, i = 1, 2, ..., n$. **do**
    Assign x_i to cluster C_{j^*} if $j^* = \text{argmin}_{j \in \{1,...,k\}} d(c_j, x_i)$.
  end for
end if

Algorithm 4 HD-LCD algorithm without the bootstrap, input $x_0 = \{x_1, ..., x_n\}$

Set $x_0 = x$.
Compute the UHD distribution $\overline{p}_u$.
Set the list $l_{center} = NULL$, cluster counter $k = 0$, remaining sample size $l = n$, p-value $= 1$.
while $l \geq n * 0.1$ **do**
    for each point of the data set x: $x_i, i = 1, 2, ..., l$. **do**
        Compute the empirical Hamming distance distribution $\overline{p}_{x_i}$.
        Compute its mixture log-concave projection $\widehat{p}_{x_i}$, see (7.1).
        Compute the cutoff value r_{x_i} using (7.2).
        Compute $\rho^*_{RKL}(\overline{p}_u \| \widehat{p}_{x_i}, r_{x_i})$.
    end for
    Set $\rho^* = \max_{i=1,...,l} \rho^*_{RKL}(\overline{p}_u \| \widehat{p}_{x_i}, r_{x_i})$, $c_k = x_{i^*}$ where $i^* = \text{argmax}_{i=1,...,l} \rho^*_{RKL}(\overline{p}_u \| \widehat{p}_{x_i}, r_{x_i})$, and set $R_k = r_{x_{i^*}} - 2$.
    $k = k + 1$.
    $l_{center} = l_{center} \cup \{c_k\}$.
    $C_k = \{x_i \in x \,|\, d(x_i, c_k) \leq R_k, i = 1, ..., l\}$.
    Set $x : x = x \setminus C_k$, and update the sample size l for x.
end while
if k ≥ 1 **then**
    for each point of x_0: $x_i, i = 1, 2, ..., n$. **do**
        Assign x_i to cluster C_{j^*} if $j^* = \text{argmin}_{j \in \{1,...,k\}} d(c_j, x_i)$.
    end for
end if

8.4 Smoothness and limitations of HD-LCD algorithm

Comparing two algorithms, the main advantage of HD-LCD algorithm is that mixture log-concave projection is much smoother than the empirical distribution. We choose two data points from Soybean disease data, Figure 8.1 and 8.2 compare how radius is determined by two algorithms. By Figure 8.1, we can see that cluster radius (FLM) chosen by HD vector algorithm is $R = 1$, which is obviously not relevant. For HD-LCD algorithm, the HD-LCD version cutoff $r = 7$, we determine the radius is $R = r - 2 = 5$. It is more relevant. Figure 8.2 shows another example where the FLM does not exist in HD vector algorithm, because there is a tie between Hamming distance 1 and 2. In this case HD vector algorithm is not able to get clustering result. While HD-LCD version radius is $R = r - 2 = 8 - 1 = 6$. Mixture log-concave projection does a much better job in both cases.

Our proposed HD-LCD algorithm is based on the assumption that the distribution of the cluster center's Hamming distance vector has a bimodal pattern, hence a data set with less number of attributes can not be used with our algorithm or the HD vector algorithm. For example, if there are less than 6 attributes in a data set ($q < 6$), it is hard to define a bimodal shape distribution for the HD vector when

there is less than six unique observations. Also through our simulation, we found that the performance of our algorithm decreases when some of the clusters have relative small cluster size. In this case small clusters member might be wrongly classified to big neighbor clusters. To improve the clustering accuracy, we applied mixture log-concave projection instead of empirical distribution. At the same time computing mixture log-concave projection slows down the algorithm, especially for data set with large sample size.

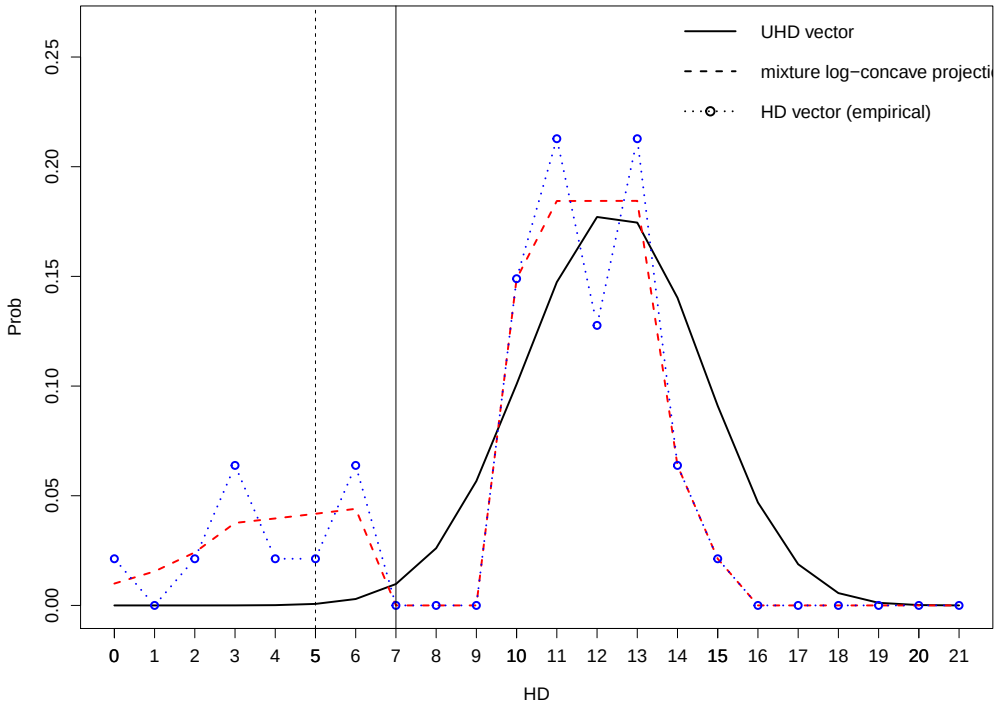


Figure 8.1: Mixture log-concave projection vs empirical distribution of a selected point from Soybean disease data set (1). HD-LCD version cutoff $r = 7$, cluster radius $R = r - 2 = 5$.

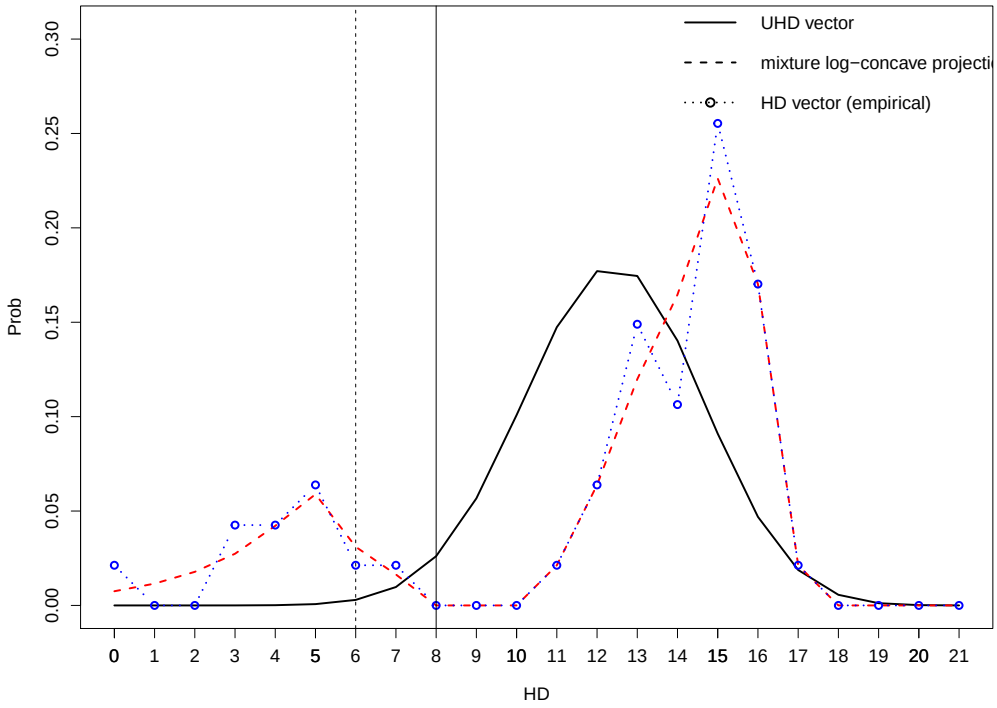


Figure 8.2: Mixture log-concave projection vs empirical distribution of a selected point from Soybean disease data set (2). HD-LCD version cutoff $r = 8$, cluster radius $R = r - 2 = 6$.

Chapter 9

Clustering result comparison

9.1 Algorithm evaluation

Following Zhang et al. (2006), we use two criteria to evaluate accuracy of clustering algorithms. They are classification rate (CR) and information gain (IG)(Bradley et al., 1998). Given K clusters, the CR is defined as $CR(K) = \sum_{k=1}^{K} \frac{\widetilde{n}_k}{n}$, where n is sample size, and $\widetilde{n}_k$ is the number of data points that have been correctly assigned to cluster k. The IG is defined as IG = total entropy - weighted entropy, where

total entropy $= -\sum_{k=1}^{K} \frac{n_k}{n} \log_2 \frac{n_k}{n}$,

weighted entropy $= -\sum_{k=1}^{K} \frac{n_k}{n} (\sum_{l=1}^{L_k} \frac{\widetilde{n}_l^k}{n_k} \log_2 \frac{\widetilde{n}_l^k}{n_k})$, where K is the number of true clusters, L_k is the number of algorithm resulted clusters in a given cluster k, $\widetilde{n}_l^k$ is the number of data points with algorithem resulted cluster label l but belongs to

true cluster label k, n_k is the number of data points in a true cluster k. Here we take a ratio of IG/total entropy, which is similar to CR, the value is between 0 and 1. But Zhang et al. (2006) also mentioned that IG may lead to a wrong conclusion.

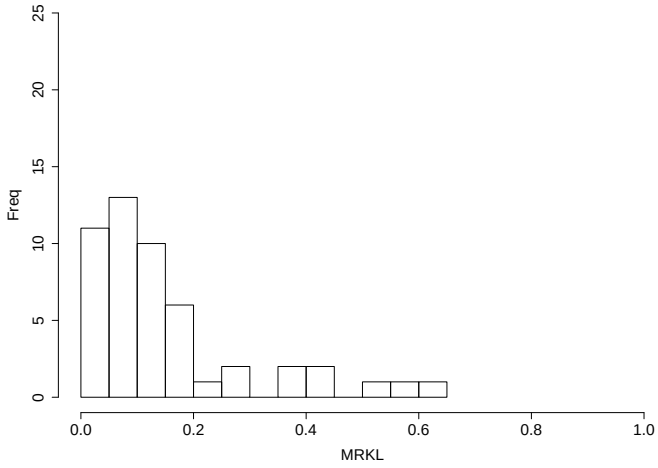
9.2 Examples for HD-LCD algorithm with the bootstrap

In this section, we illustrate two examples which we apply HD-LCD algorithm with the bootstrap. Two example data sets are both simulated by the modified simulation method, which will be introduced in later section. they have q = 7 attributes, each attribute has three levels, that is, m_1 = ... = m_7 = 3, A_1 = ... = A_7 = $\{1, 2, 3\}$. For example 1, sample size n = 40. Points are grouped into two clusters with cluster size n_1 = n_2 = 20. For example 2, sample size n = 60. Points are grouped into three clusters with cluster size n_1 = n_2 = n_3 = 20. For both cases, centers are chosen satisfying $d(c_i, c_j) \geq 5, i \neq j$.

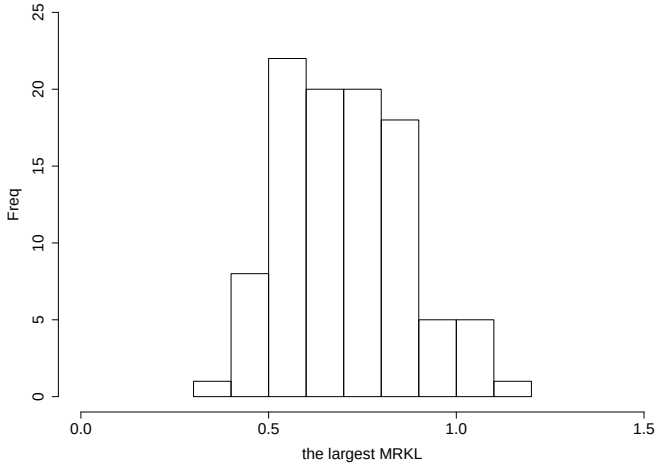
We apply HD-LCD algorithm with the bootstrap to the example data sets, Table 9.1 shows the clustering result of both examples. Hypothesis testing p-value are listed for each round. We can see that the testing process works well.

Figure 9.1 compares the histogram of testing samples' largest MRKL and the

histogram of example 2's MRKL in the first round. The distribution of the MRKL of example 2 (Figure 9.1 (a)) is right skewed, its value is between 0-0.62. The distribution of largest MRKL of testing samples (Figure 9.1 (b)) has an approximate bell shape, the value is between 0.3-1.1. We can see the range 0.3 - 1.1 in Figure 9.1 (a) is exactly the skewed tail. That is, most of the point in the example 2, as members of clusters, do not reveal the cluster pattern. Only a small number of points (right of value 0.3) is detected significant cluster pattern, they are very likely to be cluster centers, or they are close to the centers.



(a) Given data x_0.



(b) Testing samples.

Figure 9.1: (a) is the histogram of modified reversed KL divergence of example data, computed in the first round, sample size n = 60. (b) is the histogram of largest modified reversed KL divergence of "cluster" samples with sample size of 60.

Table 9.1: Examples for HD-LCD algorithm with the bootstrap.

	Example 1 ($K = 2$)	**Example 2** ($K = 3$)
CR	0.925	0.87
IG	0.68	0.62
p**-value (1st round)**	0.44	0.35
p**-value (2nd round)**	0.21	0.07
p**-value (3rd round)**	0.01	0.43
p**-value (4th round)**	not available	0.01
Description	Algorithm runs three round, testing process stops the algorithm at the third round because p-value $= 0.01 < \alpha = 0.05$	Algorithm runs four round, testing process stops the algorithm at the fourth round because p-value $= 0.01 < \alpha = 0.05$.

9.3 Simulation study

Simulation study can help us to evaluate the performance of algorithms. We implement two simulation studies comparing our proposed algorithm with HD vector algorithm and K-modes algorithm. We call the first simulation method as the original simulation method, which we follow the simulation introduced by Zhang et al. (2006). With the original simulation method, the simulated sample has $q = 10$ attributes, the sample size is $n = 200$. Five clusters are simulated ($n_1 = 70, n_2 = 50, n_3 = 40, n_4 = 25, n_5 = 15$). Clusters' centers are set to be well-separated, that is, pairwise Hamming distance between centers is greater or equal to five. The cluster members' attribute value are selected following multinomial distribution.

The second simulation method is based on the original simulation, but we did some modifications in order to increase the non-smoothness of the empirical distribution of Hamming distance. We call it as the modified simulation method. The simulated data still has $q = 10$ attributes, but we decrease the sample size to $n = 100$. Four clusters are simulated with cluster size: $n_1 = 40, n_2 = 30, n_3 = 20, n_4 = 10$. Clusters' centers are chosen following the same rule (pairwise Hamming distance is greater or equal to five). The main difference between two simulation methods is at the last step, where to determine the cluster members' attribute value. For the ith attribute, to determine the ith attribute's value for cluster member, we set the center's ith

attribute's value (state) with a relative higher probability than other states, to make sure cluster members are "close" to their center. For the original simulation method, the same state with the center's ith attribute's value is set to a fixed probability 0.7, all other states' probability is evenly set to $(1-0.7)/(m_i-1)$. Intuitively, within each cluster, distance between each data point and the cluster center follow a normal distribution with mean equals to $q * 0.7$, which very likely results in a smooth curve for first peak. For the modified simulation, we decrease sample size to increase the sparsity of clusters. Cluster member's ith attribute's value still follow multinomial distribution. But the same state with center's ith attribute's value does not have a fixed probability. Its probability is randomly chosen from $\{0.5, 0.6, 0.7, 0.8, 0.9\}$. We make this modification in order to add some "noise". For each simulation, we create 200 samples. Table 9.2 describes details of two simulation methods.

We apply HD-LCD algorithm without the bootstrap to the simulated data, Tables 9.3 and 9.4 compare the results of two simulations. For HD vector algorithm, due to the limitation of not existing FLM, it runs successfully for part of simulated samples. That is, 152 out of 200 for original simulation method; 132 out of 200 for modified simulation method. To provide an equal comparison, the result in the tables for all three algorithms is the average of 152/132 simulations, respectively. Recall that the K-modes algorithm chooses initial cluster center by seed, therefore same data set

may get different clustering results. We reorder each simulation data 100 times by random, to reduce the dependence of the order/seed of input data. Hence the results in table is the average of 15200/13200 clustering results. Also because the number of clusters need to be specified for K-modes algorithm, we tried different k values to have a comprehensive comparison.

Occasionally, the IG criterion can give misleading result as we described in previous chapter, we rely more on the CR criteria when we interpret the comparison results. For the original simulation study, our proposed algorithm's CR and IG are all better than the other two algorithms. Our HD-LCD algorithm's standard deviation is the smallest among three algorithms. It is slightly more accurate than HD vector algorithm and K-modes algorithm (when the correct number of clusters is specified) in terms of classification rate. For the modified simulation study, the performance of HD-LCD algorithm is obviously better than HD vector algorithm, and is about the same level comparing to the K-modes algorithm (when the correct number of clusters is specified). Note that K-modes algorithm has two significant drawbacks: need to specify cluster's number and clustering result is not unique. We can conclude that HD-LCD algorithm's performance of simulation study is the best among three algorithms.

9.4 Soybean disease data and zoo data

We introduced zoo data in previous chapter. About soybean disease data posted on UCI Machine Learning Repository (Lichman, 2013), there are two data sets available. We choose to use the smaller data set without missing values. In this data set, diseased soybean plant's data were collected to classify the type of disease. The data has 35 attributes, for example, seeds' color, leaf's shape, and rotted roots. In our algorithm, 21 attributes are used for clustering after unique value columns are removed. Per the data information, 47 data points are grouped into four clusters. The clusters size are 17, 10, 10, and 10.

Tables 9.5 and 9.6 shows the clustering results comparison between HD-LCD algorithm without the bootstrap, HD vector algorithm and the K-modes algorithm. For zoo data, our algorithm is much better than K-modes algorithm, and has a comparable accuracy with HD vector algorithm. For soybean data, our algorithm provides similar accuracy level with K-modes algorithm (when the correct number of clusters is specified), but it is less accurate than the HD-vector algorithm. Our algorithm's CR rate is 85.1%, while the HD-vector algorithm has a CR rate of 100%. About 7 data points are incorrectly clustered in our algorithm. Our algorithm divides the data into five clusters with clusters size of 21, 10, 7, 6, and 3. We can see that the first cluster has a relative big cluster size. There are four data points are wrongly

grouped into the fist cluster, it is the major reason of the lower accuracy rate. The HD-LCD algorithm's accuracy is sensitive to the cluster centers chosen, because the cluster radius is determined by the width of the first bump of the cluster center. It is hard to tell the exact reason of the mis-clustering of soybean disease data. Sparsity might be one possible reason although we have not theoretically proved it.

Table 9.2: Comparison of two simulations

#	**Original simulation**	**Modified simulation**
1	set number of attributes $q = 10$, Randomly select 10 numbers from the set $\{4, 5, 6\}$ as the levels of attributes $m_1, ..., m_{10}$.	set number of attributes $q = 10$, Randomly select 10 numbers from the set $\{2, 3, 4, 5\}$ as the levels of attributes $m_1, ..., m_{10}$.
2	set 5 clusters, choose five cluster centers, $c_k, k = 1, ..., 5$, such that $d(c_i, c_j) \geq 5$ for all $i \neq j$.	set 4 clusters, choose four cluster centers, $c_k, k = 1, ..., 4$, such that $d(c_i, c_j) \geq 5$ for all $i \neq j$.
3	set the sample size $n = 200$ with cluster sizes $n_1 = 70, n_2 = 50, n_3 = 40, n_4 = 25, n_5 = 15$.	set the sample size $n = 100$ with cluster sizes $n_1 = 40, n_2 = 30, n_3 = 20, n_4 = 10$.
4	generate data points for each cluster. Let take example of first cluster with center c_1, we will generate $n_1 - 1 = 69$ data points. Those 69 data points are generated with the following rule: for ith $(i = 1, ..., 10)$ attribute, it has m_i states $\{a_{i1}, ..., a_{im_i}\}$. Let ith attribute of center c_1 be c_{1i}, where $c_{1i} \in \{a_{i1}, ..., a_{im_i}\}$. Then for each 69 data points, the ith attribute can be chosen from $\{a_{i1}, ..., a_{im_i}\}$, which follow multinomial distribution. The occurrence probability of c_{1i} is 0.7, the occurrence probability of other values is $\frac{(1-0.7)}{m_i - 1}$.	generate data points for each cluster. Let take example of first cluster with center c_1, we will generate $n_1 - 1 = 39$ data points. Those 39 data points are generated with the following rule: for ith $(i = 1, ..., 10)$ attribute, it has m_i states $\{a_{i1}, ..., a_{im_i}\}$. Let ith attribute of center c_1 be c_{1i}, where $c_{1i} \in \{a_{i1}, ..., a_{im_i}\}$. Then for each 39 data points, the ith attribute can be chosen from $\{a_{i1}, ..., a_{im_i}\}$. And the occurrence probability of c_{1i}, indicated by $p(c_{1i})$ is randomly chosen from $\{0.5, 0.6, 0.7, 0.8, 0.9\}$, the occurrence probability of other states is randomly assigned, which sum up to $1 - p(c_{1i})$. Hence $\{a_{i1}, ..., a_{im_i}\}$ still follows multinomial distribution.

Table 9.3: Original simulation study comparison

			K-modes		
Original	**HD-LCD**	**HD vector**	**k=4**	**k=5 (True)**	**k=6**
mean of CR	94.3%	93.1%	87.8%	93.2%	93.2%
SD of CR	3.6%	9.6%	5.2%	5.9%	4.8%
mean of IG	92.7%	91.7%	75.8%	86.9%	90.2%
SD of IG	5.1%	4.0%	7.0%	8.4%	6.4%

Table 9.4: Modified simulation study comparison

			K-modes		
Modified	**HD-LCD**	**HD vector**	**k=3**	**k=4 (True)**	**k=5**
mean of CR	84.0%	80.7%	80.6%	84.7%	81.9%
SD of CR	7.6%	14.9%	8.0%	9.4%	8.5%
mean of IG	78.8%	79.4%	58.0%	70.2%	74.3%
SD of IG	11.1%	9.8%	11.3%	13.7%	11.6%

Table 9.5: Comparison on soybean disease data

			K-modes		
Soybean	**HD-LCD**	**HD vector**	**k=3**	**k=5 (True)**	**k=5**
mean of CR	85.1%	100%	75.7%	86.1%	83.1%
SD of CR	NA	NA	5.8%	15.3%	9.3%
mean of IG	84.0%	100%	66.5%	85.8%	94%
SD of IG	NA	NA	9.6%	14.2%	10.2%

Table 9.6: Comparison on zoo data

			K-modes		
Zoo	**HD-LCD**	**HD vector**	**k=6**	**k=7 (True)**	**k=8**
mean of CR	91.1%	95.1%	78.5%	77.5%	76.3%
SD of CR	NA	NA	8.9%	8.7%	8.2%
mean of IG	84.0%	91.6%	79.0%	81.8%	83.8%
SD of IG	NA	NA	6.0%	5.0%	4.2%

Chapter 10

Conclusion and future work

This thesis includes two Parts. In Part I, we defined a new class of shape-constraint discrete distribution in higher dimensional space: extendible-log-concave (eLC) PMF. We studied and discovered the properties of the eLC PMF. We proved the existence and uniqueness of the eLC MLE, and introduced the algorithm to compute the MLE. We also showed the performance of our eLC MLE through simulations, and compared our estimator with parametric method and empirical nonparametric method. In Part II, we applied the univariate log-concave MLE (see Balabdaoui et al. (2013)) in clustering. We developed a log-concave mixture model to approximate bimodal pattern discrete distributions. We then applied the log-concave mixture model to clustering algorithm (HD vector algorithm) and studied its performance through simulated and real data, and compared our HD-LCD algorithm with other clustering

algorithms. The two Parts of this thesis are not directly related. In Part I, our work focus on higher dimensions, while in Part II the application is based on one dimensional space. What the two parts have in common is that they both focus on discrete log-concavity. Part I proposed a discrete log-concave MLE in dimensions larger than one, and Part II apply the discrete log-concave MLE to clustering application.

Developing eLC mixture model in higher dimensions is a possible future work. The main challenge is how to obtain acceptable running speed. In our current setting, we applied the EM algorithm to compute the mixture log-concave projection. It computes the value of likelihood function for each candidate distribution and update the candidate distribution in each iteration. The mixture log-concave projection is determined when the value of likelihood function converges. It means that in higher dimensions we need to compute eLC MLE in each iteration until it converges. Computing eLC MLE is more complicated than in one dimension. Hence the key is how to improve the running speed. Additional suggested future work includes:

- to derive the asymptotic distribution and convergence rate of the eLC MLE. Balabdaoui et al. (2013) showed the asymptotic behavior of discrete log-concave MLE in one-dimensional space, which might be the starting point.

- to verity the eLC assumption for a given data set. Checking the eLC assumption in higher dimensions is not easy. If one can derive the relations between

eLC joint PMF and the eLC marginals, then the problem might be easier. Cule (2009) showed that the marginal distribution of log-concave density is also log-concave in continuous setting.

In Section 8.4, we listed some limitations of our current HD-LCD algorithm. Those limitations should be the food for the future work.

Appendix A

Appendix: background material

A.1 Important Definitions, Theorems and Lemma

Definition A.1.1. *(?, Definition 5.1, page 85) A function* $f : \mathbb{R}^d \to \mathbb{R}$ *is said to be convex if*

$$pf(x) + qf(y) \geq f(px + qy)$$

for all $x, y \in \mathbb{R}^d$, *and* $p, q \in (0, 1)$ *with* $p + q = 1$.

Definition A.1.2. *(?, Definition 5.3, page 85) A subset* C *of* $\mathbb{R}^d$ *is said to be convex if*

$$px + qy \in C$$

for all $x, y \in C$, *and* $p, q \in (0, 1)$ *with* $p + q = 1$.

Definition A.1.3. *(Rockafellar, 1970, pages 24) A convex function f is said to be proper if its epigraph is non-empty and contains no vertical lines, i.e. $f(x) < +\infty$ for at least one x, and $f(x) > -\infty$ for every x.*

Definition A.1.4. *(Rockafellar, 1970, Theorem 7.1, pages 51–52) A proper convex function*

$f : \mathbb{R}^d \to (-\infty, +\infty]$ is said to be closed if the function is lower semi-continuous. Or, equivalently, if

$$\{x \mid f(x) \leq \alpha\} \text{ is closed for every } \alpha \in \mathbb{R}.$$

The only closed improper convex functions are the constant functions $+\infty$. and $-\infty$.

Definition A.1.5. *For a function $f : \mathbb{Z}^d \to \mathbb{R} \cup \{\infty\}$, We define the convex closure of f as*

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

Definition A.1.6. *A sequence of probability measures p_n on metric space S is defined to be tight if for every $\varepsilon > 0$ there exists n_0 and a compact set $K \subset S$, such that $p_n(K) > 1 - \varepsilon$ for all $n > n_0$.*

Theorem A.1.1. *(Rockafellar, 1970, Theorem 4.5, page 27) Let f be a twice continuously differentiable real-valued function on an open convex set C in $\mathbb{R}^n$. Then f is convex on C if and only if its Hessian matrix*

$$Q_x = (q_{ij}(x)),\ q_{ij}(x) = \frac{\partial^2 f}{\partial \xi_1 \partial \xi_2},$$

is positive semi-definite for every $x \in C$.

Theorem A.1.2. *(Rockafellar, 1970, Theorem 5.5, page 35) The pointwise supremum of an arbitrary collection of convex functions is convex.*

Theorem A.1.3. *(Rockafellar, 1970, Theorem 7.1, page 51) Let f be an arbitrary function from $\mathbb{R}^n$ to $[-\infty, \infty]$. Then the following conditions are equivalent:*

1. *f is lower semi-continuous through out $\mathbb{R}^n$;*
2. *$\{x | f(x) \leq \alpha\}$ is closed for every $\alpha \in \mathbb{R}$;*
3. *The epigraph of f is a closed set in $\mathbb{R}^{n+1}$.*

Theorem A.1.4. *(Rockafellar, 1970, Theorem 10.6, page 88) Let $\mathcal{C}$ be a relatively open convex set, and let $\{f_i | i \in \mathcal{I}\}$ be an arbitrary collection of convex functions finite and pointwise bounded on $\mathcal{C}$. let $\mathcal{S}$ be any closed bounded subset of $\mathcal{C}$. Then $\{f_i | i \in \mathcal{I}\}$ is uniformly bounded on $\mathcal{S}$ and equi-Lipschitzian relative to $\mathcal{S}$.*

The conclusion remains valid if the pointwise boundedness assumption is weakened to the following pair of assumptions:

(a). There exists a subset $\mathcal{C}'$ *of* $\mathcal{C}$ *such that* $\text{conv}(\text{cl } \mathcal{C}') \supset \mathcal{C}$ *and* $\sup\{f_i(x) | i \in \mathbb{I}\}$ *is finite for every* $x \in \mathcal{C}'$*;*

(b). There exists at least one $x \in \mathcal{C}$ *such that* $\inf\{f_i(x) \,|\, i \in \mathbb{I}\}$ *is finite.*

$\text{conv}(\mathcal{C})$ denotes convex hull of $\mathcal{C}$, and cl $\mathcal{C}$ denotes closure of $\mathcal{C}$.

Theorem A.1.5 (Carathéodory's Theorem)**.** *(Rockafellar, 1970, Theorem 17.1, page 155) Let* S *be any set of points and directions in* $\mathbb{R}^n$*, and let* $C = \text{conv}(S)$*. Then* $x \in C$ *if and only if* x *can be expressed as a convex combination of* $n+1$ *of the points and directions in* S *(not necessarily distinct). In fact* C *is the union of all the generalized* $d-$*dimensional simplices whose vertices belong to* S*, where* $d = dim C$*.*

Theorem A.1.6. *(Rockafellar, 1970, Theorem 23.1, page 213) Let* f *be a convex function, and let* x *be a point where* f *is finite. For each* y*, the difference quotient in the definition of* $f'(x;y)$ *is a non-decreasing function of* $\lambda > 0$*, so that* $f'(x;y)$ *exists and*

$$f'(x;y) = \inf_{\lambda > 0} \frac{f(x+\lambda y) - f(x)}{\lambda}.$$

Moreover, $f'(x;y)$ *is a positively homogeneous convex function of* y*, with* $f'(x;y) = 0$

and

$$-f'(x;-y) \le f'(x;y), \quad \forall y.$$

Theorem A.1.7. *(Rockafellar, 1970, Theorem 25.2, page 244) Let f be a convex function on $\mathbb{R}^d$, and let x be a point at which f is finite. A necessary and sufficient condition for f to be differentiable at z is that the directional derivative function $f'(x;\cdot)$ be linear. Moreover, this condition is satisfied if merely the n two-sided partial derivatives $\frac{\partial f(z)}{\partial \xi_j}$ exist at x and are finite.*

Theorem A.1.8. *(Rockafellar, 1970, Theorem 32.2, page 343) Let f be a convex function, and let $\mathcal{C} = \mathrm{conv}(\mathcal{S})$, where $\mathcal{S}$ is an arbitrary set of points. Then*

$$\sup\{f(x) \,|\, x \in \mathcal{C}\} = \sup\{f(x) \,|\, x \in \mathcal{S}\},$$

where the first supremum is attained only when the second (more restrictive) supremum is attained.

Theorem A.1.9. *(?, The Arzelà-Ascoli Theorem, page 221) Let $C = C[0,1]$ be the space of continuous, real-valued functions on the unit interval [0,1] with the uniform metric. A subset A of C has compact closure if and only if*

$$\sup_{x \in A} |x(0)| < \infty$$

and

$$\lim_{\delta \to 0} \sup_{x \in A} w_x(\delta) = 0,$$

where $w_x(\delta) = \sup_{|s-t|<\delta} |x(s) - x(t)|$.

Theorem A.1.10. *(?, The Monotone Convergence Theorem, page 83) Let* $\{f_n\}$ *be an increasing sequence of nonnegative measurable functions on* E*. If* $\{f_n\} \to f$ *pointwise a.e. on* E*, then* $\lim_{n\to\infty} \int_E f_n = \int_E f$.

Theorem A.1.11. *(?, The Lebesgue Dominated Convergence Theorem, page 88) Let* $\{f_n\}$ *be a sequence of measurable functions on* E*. Suppose there is a function* g *that is integrable over* E *and dominates* $\{f_n\}$ *on* E *in the sense that* $|f_n| \leq g$ *on* E *for all* n*.*

If $\{f_n\} \to f$ *pointwise a.e. on* E*, then* f *is integrable over* E *and* $\lim_{n\to\infty} \int_E f_n = \int_E f$.

Theorem A.1.12. ? *Let* $f(x, y)$ *be a log-concave functions on* $\mathbb{R}^m \times \mathbb{R}^n$*, with* $x \in \mathbb{R}^m$ *and* $y \in \mathbb{R}^n$*. Further, let* A *be a convex subset of* $\mathbb{R}^n$*. Then*

$$g(x) = \int_A f(x, y)dy$$

is a log-concave function on $\mathbb{R}^m$.

Theorem A.1.13. *(Murota and Shioura, 2001, Theorem 4.17) For a convex-extendible*

function $f : \mathbb{Z}^d \to \mathbb{R} \cup \{+\infty\}$, *and a value* $\lambda \in \mathbb{R} \cup \{+\infty\}$, *the level set* $L(f, \lambda)$ *is a convex-extendible set.*

Where the level set is defined as $L(f, \lambda) = \{z \in \mathbb{Z}^d \mid f(z) \leq \lambda\}$.

Theorem A.1.14. *(Murota, 2009, Theorem 2.1) A function* $f : \mathbb{Z} \to \bar{\mathbb{R}}$ *is convex-extendible is and only if it satisfies*

$$f(x-1) + f(x+1) \geq 2f(x),$$

where $\bar{\mathbb{R}} = \mathbb{R} \cup \{+\infty\}$.

Theorem A.1.15. *(Murota, 2009, Theorem 2.1) A function* $f : \mathbb{Z} \to \bar{\mathbb{R}}$ *is convex-extendible is and only if it satisfies*

$$f(x-1) + f(x+1) \geq 2f(x),$$

where $\bar{\mathbb{R}} = \mathbb{R} \cup \{+\infty\}$.

Theorem A.1.16. *(Rosenthal, 2006, Theorem 11.1.10) If* $\{u_n\}$ *is a tight sequence of probability measures, then there is a subsequence* $\{u_{n_k}\}$ *and a probability measure* u, *such that* $u_{n_k} \Rightarrow u$, *i.e.* $\{u_{n_k}\}$ *converges weakly to* u.

A.2 List of important symbols and notations

$\|\cdot\|_\infty$: Maximum norm.

f : Probability density function on $\mathbb{R}^d$.

$\widehat{f_n}$: MLE of log-concave density on $\mathbb{R}^d$.

f_0 : True probability density function on $\mathbb{R}^d$.

$\mathcal{F}$: set of log-concave density functions on $\mathbb{R}^d$.

g : Concave function on $\mathbb{R}^d$.

$h(x)$: Convex function on $\mathbb{R}^d$.

$h(z)$: Convex function on $\mathbb{Z}^d$.

$\bar{h}$: Convex closure of h on $\mathbb{R}^d$.

h^R : Convex extension of h on $\mathbb{R}^d$.

$h^2(p,q)$: Hellinger distance between p and q.

$l_k(p,q)$: L_k distance between p and q.

p : Probability mass function on $\mathbb{Z}^d$.

p_0 : True PMF on $\mathbb{Z}^d$, $\sum_{z\in\mathbb{Z}^d}\|z\|_\infty p_0(z) < \infty$ and $|\sum_{z\in\mathbb{Z}^d} p_0(z)\log p_0(z)| < \infty$.

p_n : a sequence of eLC PMF on $\mathbb{Z}^d$.

$\widehat{p}$: Mixture log-concave projection/MLE on $\{0, ..., q\}$.

$\widehat{p_0}$: eLC KL projection of p_0, which minimize KL divergence to p_0 over eLC PMF set.

$\widehat{p_n}$: eLC MLE on $\mathbb{Z}^d, d \in \{1, 2, 3, ...\}$.

$\mathcal{P}_0$: Set of eLC PMF on $\mathbb{Z}^d$.

$\mathcal{P}_1$: Set of log-concave PMF on $\mathbb{Z}$.

$\mathcal{P}_q$: Set of log-concave PMF on $\{0, ..., q\}$.

$\mathcal{P}_m$: Set of log-concave mixture PMF on $\{0, ..., q\}$.

t_y : Tent function on $\mathbb{R}^d$.

φ : $\varphi = \log p$ is convex-extendible function on $\mathbb{Z}^d$.

C : Open convex set on $\mathbb{R}^d$.

$(k)_i$: point of $\mathbb{Z}^d$ with ith element equals to $k, k \in \mathbb{Z}$.

$\mathbb{N}$: Natural number set.

S : Set on $\mathbb{Z}^d$.

$\bar{S}$: Convex closure of S, on $\mathbb{R}^d$.

$\mathcal{S}$: Support of PMF p, on $\mathbb{Z}^d$.

$\mathcal{S}_0$: Support of true PMF p_0, on $\mathbb{Z}^d$.

$\bar{\mathcal{S}}_n$: Convex closure of $S_n = \{z_1, ..., z_m\}$.

$\widehat{\mathcal{S}}_n$: $\bar{\mathcal{S}}_n \cap \mathbb{Z}^d$.

$\mathbb{Z}$: Integers set.

$\widehat{\Sigma}_n$: Variance matrix under the MLE.

$\overline{\Sigma}_n$: Empirical variance matrix.

Δ : Set difference symbol.

σ : objective function to estimate eLC MLE on $\mathbb{Z}^d$.

σ_R : objective function to estimate eLC MLE on $\mathbb{R}^d$.

eLC : log-concave-extendible functions on $\mathbb{Z}^d$.

MLE : Maximum likelihood estimator.

PMF : Probability mass function.

Bibliography

F. Balabdaoui and H. Jankowski. Maximum likelihood estimation of a unimodal probability mass function. *Statist. Sinica*, 26(3):1061–1086, 2016.

F. Balabdaoui, K. Rufibach, and J. A. Wellner. Limit distribution theory for maximum likelihood estimation of a log-concave density. *Ann. Statist.*, 37(3):1299–1331, 2009.

F. Balabdaoui, H. Jankowski, K. Rufibach, and M. Pavlides. Asymptotics of the discrete log-concave maximum likelihood estimator and related applications. *J. R. Stat. Soc. Ser. B. Stat. Methodol.*, 75(4):769–790, 2013.

R. B. Bapat. Discrete multivariate distributions and generalized log-concavity. *Sankhyā Ser. A*, 50(1):98–110, 1988.

H. H. Bauschke and J. M. Borwein. On projection algorithms for solving convex feasibility problems. *SIAM Review*, 38(3):367–426, 1996.

P. J. Bickel and J. Fan. Some problems on the estimation of unimodal densities. *Statistica Sinica*, 1996.

L. Birgé. Estimation of unimodal densities without smoothness assumptions. *Ann. Statist.*, 25:970–981, 1997.

P. S. Bradley, Usama Fayyad, and Cory Reina. Scaling clustering algorithms to large databases. In *Proceedings of the Fourth International Conference on Knowledge Discovery and Data Mining*, KDD'98, pages 9–15. AAAI Press, 1998.

G. T. Chang and G. Walther. Clustering with mixtures of log-concave distributions. *Comput. Statist. Data Anal.*, 51(12):6242–6251, 2007.

M. Cule. *Maximum likelihood estimation of a multivariate log-concave density.* PhD thesis, University of Cambridge, UK, 2009.

M. Cule, R. Samworth, and M. Stewart. Maximum likelihood estimation of a multi-dimensional log-concave density. *J. R. Stat. Soc. Ser. B Stat. Methodol.*, 72(5): 545–607, 2010.

A. P. Dempster, N. M. Laird, and D. B. Rubin. Maximum likelihood from incomplete data via the EM algorithm. *J. Roy. Statist. Soc. Ser. B*, 39(1):1–38, 1977.

S. Dharmadhikari and K. Joag-Dev. *Unimodality, convexity, and applications*. Probability and Mathematical Statistics. Academic Press, Inc., Boston, MA, 1988.

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

L. Dümbgen, A. Hüsler, and K. Rufibach. Active set and EM algorithms for log-concave densities based on complete and censored data. Technical Report 61, IMSV, University of Bern, 2007.

Lutz Dümbgen, Andre Hüsler, and Kaspar Rufibach. Active set and EM algorithms for log-concave densities based on complete and censored data. *Technical report, Universität Bern*, 2007. URL http://arxiv.org/abs/0707.4643.

S. Geman and C. Hwang. Nonparametric maximum likelihood estimation by the method of sieves. *Ann. Statist.*, 10(2):401–414, 1982.

A. D. Gordon. *Classification*. Chapman & Hall/CRC, 1999.

U. Grenander. On the theory of mortality measurement. *Scandinavian Actuarial Journal*, pages 70–96, 1956.

P. Groeneboom. Estimating a monotone density. *Proceedings of the Berkeley Conference in Honor of Jerzy Neyman and Jack Kiefer*, 1985.

D. Harrison and D. L. Rubinfeld. Hedonic housing prices and the demand for clean air. *Journal of Environmental Economics and Management*, 5(1), 1978.

J. Huang and J. A. Wellner. Estimation of a monotone density or monotone hazard under random censoring. *Scandinavian Journal of Statistics*, pages 3–33, 1995.

Z. Huang. Extensions to the k-means algorithm for clustering large data sets with categorical values. *Data Mining and Knowledge Discovery*, 2:(283), 1998.

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

F. Kappel and A. V. Kuntsevich. An implementation of Shor's r-algorithm. *Comput. Optim. Appl.*, 15(2):193–205, 2000.

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

S. G. Krantz. *Real analysis and foundations*. Textbooks in Mathematics. CRC Press, Boca Raton, FL, third edition, 1991.

S. Kullback and R. A. Leibler. On information and sufficiency. *Ann. Math. Statistics*, 22:79–86, 1951.

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

D. Lutz and R. Kaspar. logcondens: Computations related to univariate log-concave density estimation. *Journal of Statistical Software*, 39(1):1–28, 2011.

K. Murota. Recent developments in discrete convex analysis. In *Research trends in combinatorial optimization*, pages 219–260. Springer, Berlin, 2009.

K. Murota and A. Shioura. Relationship of M-/L-convex functions with discrete convex functions by Miller and Favati-Tardella. *Discrete Appl. Math.*, 115(1-3): 151–176, 2001. 1st Japanese-Hungarian Symposium for Discrete Mathematics and its Applications (Kyoto, 1999).

B. L. S. Prakasa Rao. Estimation of a unimodal density. *The Indian Journal of Statistics*, 1969.

I. Rish. An empirical study of the naive bayes classifier. In *IJCAI 2001 workshop on empirical methods in artificial intelligence*, volume 3, pages 41–46. IBM New York, 2001.

R. T. Rockafellar. *Convex analysis*. Princeton Mathematical Series, No. 28. Princeton University Press, Princeton, N.J., 1970.

J. S. Rosenthal. *A first look at rigorous probability theory*. World Scientific Publishing Co. Pte. Ltd., Hackensack, NJ, second edition, 2006.

N. Z. Shor. *Minimization methods for nondifferentiable functions*, volume 3 of *Springer Series in Computational Mathematics*. Springer-Verlag, Berlin, 1985. Translated from the Russian by K. C. Kiwiel and A. Ruszczyński.

G. Walther. Inference and modeling with log-concave distributions. *Statist. Sci.*, 24 (3):319–327, 2009.

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

M. A. Wolters and W. John Braun. Enforcing shape constraints on a probability density estimate using an additive adjustment curve. *Communications in Statistics - Simulation and Computation*, 0(0):1–20, 2017.

P Zhang, X Wang, and P. X.-K. Song. Clustering categorical data based on distance vectors. *Journal of the American Statistical Association*, 101(473), 2006.