Least-squares estimation of a convex discrete distribution

Cécile Durot, Sylvie Huet, Francois Koladjo, Stephane S. Robin

Least-squares estimation of a convex discrete distribution

Cécile Durot[a], Sylvie Huet[1,*], François Koladjo[1,c], Stéphane Robin[d,e]

[a] *UFR SEGMI, Université Paris Ouest Nanterre La Défense, F-92001, Nanterre, France*
[b] *UR341 MIA, INRA, F-78350 Jouy-en-Josas, France*
[c] *CIPMA-Chaire UNESCO, FAST, UAC, 072BP50 Cotonou, Bénin*
[d] *UMR518 MIA, INRA F-75005 Paris, France*
[e] *UMR518 MIA, AgroParisTech F-75005 Paris, France*

Abstract

The least squares estimator of a discrete distribution under the constraint of convexity is introduced. Its existence and uniqueness are shown and consistency and rate of convergence are established. Moreover it is shown that it always outperforms the classical empirical estimator in terms of the Euclidean distance. Results are given both in the well- and the mis-specified cases. The performance of the estimator is checked throughout a simulation study. An algorithm, based on the support reduction algorithm, is provided. Application to the estimation of species abundance distribution is discussed.

Keywords:
convex discrete distribution, nonparametric estimation, least squares, support reduction algorithm, abundance distribution

1. Introduction

Recently, the problem of estimating a discrete probability mass function under a shape constraint has attracted attention: Jankowski and Wellner (2009) considered the non-parametric estimation of a monotone distribution and Balabdaoui et al. (2012) considered the case of a log-concave distribution. Although the discrete case is in some ways very different from the continuous case (for example, the convergence rates are typically different in the

*Corresponding author.
E-mail address: sylvie.huet@jouy.inra.fr

two cases), the construction of shape constrained estimators in the discrete case is largely inspired by the construction of shape constrained estimators of a probability density function. The nonparametric estimation, based on i.i.d. observations, of the distribution of a continuous random variable under a shape constraint, has received a great deal of attention in the past decades, see Balabdaoui and Wellner (2007) for a review. The most studied constraint is the monotonicity of the density function. It is well-known that the nonparametric maximum likelihood estimator of a decreasing density function over $[0, \infty)$ is the Grenander estimator defined as the left-continuous slope of the least concave majorant of the empirical distribution function of the observations. This estimator can be easily implemented using the PAVA (pool adjacent violators algorithm) or a similar device, see Barlow et al. (1972). The nonparametric maximum likelihood of a log-concave density function (i.e., a density function f such that $\log(f)$ is a concave function) was introduced in Walther (2002) and algorithmic aspects were treated in Dümbgen et al. (2007), see also the R package in Dümbgen and Rufibach (2011). Another well studied constraint is the convexity (or concavity) of the density function over a given interval. It was shown by Groeneboom et al. (2001) that both the least squares estimator and the nonparametric maximum likelihood estimator under the convexity constraint exist and are unique. However, although a precise characterization of these estimators is given in that paper, their practical implementation is a non-trivial issue: it requires sophisticated iterative algorithms that use a mixture representation, such as the *support reduction algorithm* described in Groeneboom et al. (2008).

In this paper, we consider the nonparametric estimation of a discrete distribution on $\mathbb{N}$ under the convexity constraint (note that a convex distribution on $\mathbb{N}$ is necessarily non-increasing, so in our setting, convex is equivalent to non-increasing convex). This problem has not yet been considered in the literature, although it has several applications, such as the estimation of species abundance distribution in ecology. In this field, the terms "nonparametric methods" often refer to finite mixtures of parametric distributions where only the mixing distribution is inferred in a nonparametric way, see e.g. (Böhning and Kuhnert (2006), Böhning et al. (2005), Chao and Shen (2004)).

We study the least squares estimator of a discrete distribution on $\mathbb{N}$ under the convexity constraint. First, we prove that the constrained least squares estimator exists and is unique, and we consider computational issues. Similar to the continuous case, we prove that a representation of convex discrete distributions can be given in terms of a – possibly infinite – mixture of triangular

functions on $\mathbb{N}$, and, based on this characterization, we derive an algorithm that provides the least squares estimate, although both the number of components in the mixture and the support of the estimator are unknown. This algorithm is an adaptation to our problem of the support reduction algorithm in Groeneboom et al. (2008). Then, we adress theoretical performance of the estimator: we prove that it always outperforms the classical empirical estimator in terms of the ℓ_2-error and that it is consistent with $\sqrt{n}$-rate of convergence (where as usual, n denotes the sample size), and we consider also the case of a misspecified model. All these results are new. Finally, we assess the performance of the least squares estimator under the convexity constraint through a simulation study. Starting from the mixture representation, we finally give a definition of a convex abundance distribution and illustrate how it applies to data sets analyzed in the literature.

The paper is organized as follows. The characterization of the constrained least squares estimator is given in Section 2 and Section 2.3 is devoted to computational issues. In Section 3 the theoretical properties of the estimator are established and a simulation study allowing to assess its performances is reported in Section 4. The application to abundance distribution is introduced in Section 5. Finally the proofs are postponed to Section 6.

Notation. Below is a list of notation and definitions that will be used throughout the paper.

The same notation is used to denote a discrete function $f : \mathbb{N} \to \mathbb{R}$ and the corresponding sequence of real numbers $(f(j))_{j\in\mathbb{N}}$. The ℓ_r-norm of a real sequence f is

$$\|f\|_r = \left(\sum_{j\geqslant 0} |f(j)|^r\right)^{1/r}$$

for all $r \in \mathbb{N}\backslash\{0\}$ and

$$\|f\|_r = \sup_{j\geqslant 0} |f(j)|$$

for $r = \infty$. For all r, $\ell_r(\mathbb{N})$ is the set of real sequences with a finite ℓ_r- norm.

For all functions $f : \mathbb{N} \to \mathbb{R}$ and all positive integers j, denote by

$$\Delta f(j) = f(j+1) - 2f(j) + f(j-1)$$

the discrete Laplacian. Let $\mathcal{C}$ be the set of convex discrete functions $f \in \ell_2(\mathbb{N})$, that is, the set of all $f \in \ell_2(\mathbb{N})$ having $\Delta f(j) \geqslant 0$ for all integers

$j \geqslant 1$, and let $\mathcal{C}_1$ be the set of all convex probability mass functions on $\mathbb{N}$, that is the set of functions $f \in \mathcal{C}$ satisfying $\sum_{i \geqslant 0} f(i) = 1$ (see e.g. Murota (2009) for more on convex discrete functions). An integer $j \geqslant 1$ is a knot of $f \in \mathcal{C}$ if $\Delta f(j) > 0$.

It should be noticed that any $f \in \mathcal{C}$ has $\lim_{j \to \infty} f(j) = 0$ so by convexity, any $f \in \mathcal{C}$ is non-negative, non-increasing and strictly decreasing on its support.

We say that a function $f : \mathbb{N} \to \mathbb{R}$ is linear over a set of consecutive integers $\{k, \ldots, l\}$, where $l > k+1$, if $\Delta f(j) = 0$ for all $j \in \{k+1, \ldots, l-1\}$.

2. The constrained LSE of a convex discrete distribution

Suppose that we observe n i.i.d. random variables $X_1, \ldots, X_n$ that take values in $\mathbb{N}$, and that the common probability mass function p_0 of these variables is convex on $\mathbb{N}$ with an unknown support. We aim to build an estimator of p_0 that satisfies the convexity constraint. For this task, we consider the constrained least-squares estimator (LSE) $\widehat{p}_n$ of p_0, defined as the minimizer of $\|f - \widetilde{p}_n\|_2$ over $f \in \mathcal{C}$, where $\widetilde{p}_n$ is the empirical estimator:

$$\widetilde{p}_n(j) = \frac{1}{n} \sum_{i=1}^{n} I_{(X_i = j)} \tag{1}$$

for all $j \in \mathbb{N}$. Recall that from the Hilbert projection theorem, it follows that the minimizer is uniquely defined, see Section 2.1 below. Moreover, we will prove that $\widehat{p}_n$ is a probability mass function on $\mathbb{N}$.

It is a common challenge in statistics to postulate an appropriate modeling. In many applications, one may face the difficulties of misspecification, which means for instance that, although we think that the underlying probability mass function is convex, this probability mass function is in fact non-convex. Therefore, an interesting issue is on how behaves a given estimator even in situations where some of the modeling assumptions are violated. For this reason, we will study our estimator even in situations where the true probability mass function p_0 is non-convex.

The rest of the section is organized as follows. From now on, unless otherwise stated, we do not assume convexity of p_0 anymore. We collect in Subsection 2.1 a few theoretical results about $\ell_2(\mathbb{N})$ and convex analysis. A more precise description of the estimator $\widehat{p}_n$ is given in Subsection 2.2, and the practical implementation of the estimator is discussed in Subsection 2.3.

2.1. Preliminaries

In this subsection, we collect a few theoretical results on $\ell_2(\mathbb{N})$ and convex analysis that will be used in the paper. Let us recall that the space $\ell_2(\mathbb{N})$, equipped with the ℓ_2-norm and the corresponding scalar product

$$\langle f, g \rangle = \sum_{j \in \mathbb{N}} f(j) g(j), \tag{2}$$

is a Hilbert space. Thus, it follows from the Hilbert projection theorem that for any closed convex $\mathcal{S} \subset \ell_2(\mathbb{N})$, and any $f \in \ell_2(\mathbb{N})$, there exists a unique sequence in $\mathcal{S}$, that we denote here by $f_{\mathcal{S}}$, such that

$$\|f - f_{\mathcal{S}}\|_2 \leqslant \|f - g\|_2 \text{ for all } g \in \mathcal{S}.$$

For any closed convex $\mathcal{S} \subset \ell_2(\mathbb{N})$, the projection operator $f \mapsto f_{\mathcal{S}}$, defined from $\ell_2(\mathbb{N})$ to $\mathcal{S}$, is known to have

$$\|f_{\mathcal{S}} - g_{\mathcal{S}}\|_2 \leqslant \|f - g\|_2 \tag{3}$$

for all $f, g \in \ell_2(\mathbb{N})$ and

$$\langle f - f_{\mathcal{S}}, g - f_{\mathcal{S}} \rangle \leqslant 0 \tag{4}$$

for all $f \in \ell_2(\mathbb{N})$ and $g \in \mathcal{S}$.

Note that both $\mathcal{C}$ and $\mathcal{C}_1$ are convex closed subsets of $\ell_2(\mathbb{N})$, so the projections $f_{\mathcal{C}}$ and $f_{\mathcal{C}_1}$ are well defined for every $f \in \ell_2(\mathbb{N})$. But any probability mass function p on $\mathbb{N}$ belongs to $\ell_2(\mathbb{N})$ since, from usual inequalities between ℓ_r-norms, $\|p\|_2 \leqslant \|p\|_1 = 1$. So the projections $p_{\mathcal{C}}$ and $p_{\mathcal{C}_1}$ are well defined for every probability mass function p on $\mathbb{N}$.

2.2. Characterizing the constrained LSE

The aim of the subsection is to provide more insight of the estimator $\widehat{p}_n$.

Recall that $\widehat{p}_n$ is defined as the unique minimizer of $\|f - \widetilde{p}_n\|_2$ over $f \in \mathcal{C}$, with $\widetilde{p}_n$ given in (1). In the following theorem, it is proved that $\widehat{p}_n$ is a probability mass function. It minimizes $\|f - \widetilde{p}_n\|_2$ over $f \in \mathcal{C}_1$, and has a finite support that contains the support of $\widetilde{p}_n$. We will denote by $\widehat{s}_n$, respectively $\widetilde{s}_n$, the maximum of the support of $\widehat{p}_n$, respectively $\widetilde{p}_n$.

Theorem 1. *We have $\widehat{p}_n \in \mathcal{C}_1$ and*

$$\|\widehat{p}_n - \widetilde{p}_n\|_2 = \inf_{f \in \mathcal{C}_1} \|f - \widetilde{p}_n\|_2 = \inf_{f \in \mathcal{C}} \|f - \widetilde{p}_n\|_2.$$

Moreover, the support of $\widehat{p}_n$ is non-empty and finite, and $\widehat{s}_n \geqslant \widetilde{s}_n$.

The following proposition gathers together a number of properties of $\widehat{p}_n$ that compare to those of the constrained least squares estimator of a convex density function over $[0, \infty)$, see Groeneboom et al. (2001): in the continuous case the constrained LSE has a bounded support, is piecewise linear, has no changes of slopes at the observation points, and has at most one change of slope between two consecutive observation points. In the discrete case, the constrained LSE is also piecewise linear with a bounded support. However, due to the fact that $\mathbb{N}$ is a discrete set, $\widehat{p}_n$ can have two knots (in the discrete case, changes of slopes correspond to knots) between consecutive observations. In the following proposition, N_n denotes the number of distinct values of the X_i's and $X_{(1)} < \cdots < X_{(N_n)}$ denote these values rearranged in increasing order.

Proposition 1. *The constrained LSE $\widehat{p}_n$ is linear over $\{0, \ldots, X_{(1)}+1\}$ and also over $\{\widetilde{s}_n - 1, \ldots, \widehat{s}_n\}$; in the case where $N_n \geqslant 2$, it has at most two knots on $\{X_{(j)}+1, \ldots, X_{(j+1)}-1\}$ for any given $j = 1, \ldots, N_n - 1$ such that $X_{(j+1)} - X_{(j)} > 1$. In the case where it has two knots on this set, these knots are consecutive points in $\mathbb{N}$.*

Such a description of $\widehat{p}_n$ does not suffice to implement the estimator on concrete examples. The difficulty in implementing the estimator comes from the fact that it is defined as a projection on a non-linear space (rather, it is a projection on a convex set). Thus, the estimator is not a linear function of the observations $X_1, \ldots, X_n$, and no closed form is available for $\widehat{p}_n$. The practical implementation of $\widehat{p}_n$ thus requires the use of a specific algorithm that will be investigated in the following subsection.

2.3. Implementing the constrained LSE

The algorithm we use to compute $\widehat{\pi}_n$ is based on the support reduction algorithm that was proposed by Groeneboom et al. (2008). It relies on the decomposition of convex discrete functions into a combination of triangular functions, as defined below. Other choices are conceivable (one could consider for instance algorithms designed for convex optimization, or algorithms designed for nonparametric mixtures) but the merit of our algorithm is that we are able to prove that it provides the target $\widehat{p}_n$ in a finite number of steps. Moreover, the support reduction algorithm shows good performance as compared to competitors in the context of convex regression, see Groeneboom et al. (2008).

Let us describe the combination mentioned above. For every integer $j \geqslant 1$, let T_j be the j-th triangular function on $\mathbb{N}$:

$$T_j(i) = \begin{cases} \dfrac{2(j-i)}{j(j+1)} \text{ for all } i \in \{0, \ldots, j-1\} \\ 0 \text{ for all integers } i \geqslant j. \end{cases} \tag{5}$$

Note that T_j is a probability mass function, i.e., $T_j(i) \geqslant 0$ for all i and $\sum_{i \geqslant 0} T_j(i) = 1$. Moreover, T_j is non-increasing and convex on $\mathbb{N}$. It can be shown (see Appendix A for a more precise statement) that any $f \in \mathcal{C}$ is a combination of the T_j's, and that the combination is unique: for any $f \in \mathcal{C}$, there exists a unique non-negative measure π on $\mathbb{N}\backslash\{0\}$ such that

$$f(i) = \sum_{j \geqslant 1} \pi_j T_j(i) = \sum_{j \geqslant i+1} \pi_j T_j(i) \text{ for all } i \geqslant 0, \tag{6}$$

where the π_j's denote the components of π. Moreover, π has a finite support if, and only if, the support of f is finite. The decomposition compares with Propositions 2.1 and 2.2 in Balabdaoui and Wellner (2007), which deals with the case of convex (and more generally, k-monotone) functions on $[0, \infty)$.

Hereafter, we denote by $\mathcal{M}$ the convex cone of non-negative measures on $\mathbb{N}\backslash\{0\}$. Note that for $\pi \in \mathcal{M}$ with a finite support, the function f defined by (6) belongs to $\mathcal{C}$. We consider the criterion function

$$\Psi_n(\pi) = \frac{1}{2} \sum_{i \geqslant 0} \left(\sum_{j \geqslant i+1} \pi_j T_j(i) \right)^2 - \sum_{i \geqslant 0} \widetilde{p}_n(i) \sum_{j \geqslant i+1} \pi_j T_j(i)$$

for all $\pi \in \mathcal{M}$, where the π_j's denote the components of π. Since

$$2\Psi_n(\pi) + \sum_{i \geqslant 0} \widetilde{p}_n(i)^2 = \|f - \widetilde{p}_n\|_2$$

for all $f \in \mathcal{C}$ and $\pi \in \mathcal{M}$ satisfying (6), and $\widehat{p}_n$ has a finite support, we have

$$\widehat{p}_n(i) = \sum_{j \geqslant i+1} \widehat{\pi}_{nj} T_j(i) \text{ for all } i \geqslant 0, \tag{7}$$

where $\widehat{\pi}_n$ is the unique minimizer of $\Psi_n(\pi)$ over the set of measures $\pi \in \mathcal{M}$ with a finite support. Therefore, computing $\widehat{p}_n$ amounts to computing $\widehat{\pi}_n$.

We describe now the algorithm for computing $\widehat{\pi}_n$. We distinguish two parts: the first part consists in computing, for a given integer $L \geqslant \widetilde{s}_n + 1$, the minimizer of $\Psi_n(\pi)$ over $\pi \in \mathcal{M}^L$, where $\mathcal{M}^L$ is the set of measures $\pi \in \mathcal{M}$ with support included in $\{1, \ldots, L\}$. It can easily be shown that the minimizer, that we denote by $\widehat{\pi}^L$, exists and is unique. Since its support is included in the known finite set $\{1, \ldots, L\}$, it can be computed using the support reduction algorithm of Groeneboom et al. (2008), as follows.

Algorithm for computing $\widehat{\pi}^L$ for a fixed L. In the sequel, for all $j \geqslant 1$ and $\mu \in \mathcal{M}$ we set

$$[d_j(\Psi_n)](\mu) = \sum_{l=0}^{j-1} T_j(l) \left(\sum_{j' \geqslant l+1} \mu_{j'} T_{j'}(l) - \widetilde{p}_n(l) \right). \tag{8}$$

1. Initialisation
   Let $S = \{L\}$ and choose $\pi^L \in \mathcal{M}^L$ such that $\pi^L_j = 0$ for all $j = 1, \ldots, L-1$ and π^L_L minimizes $\sum_{i=0}^{L-1} \left(\widetilde{p}_n(i) - \pi T_L(i)\right)^2$ over $\pi \in \mathbb{R}$.
2. Optimisation over $\mathcal{M}^L$

   Step 1: For $1 \leqslant j \leqslant L$ compute the quantities $[d_j(\Psi_n)](\pi^L)$. If all are non negative, then set $\widehat{\pi}^L = \pi^L$, and the optimisation over $\mathcal{M}^L$ is achieved. If not, choose j such that $[d_j(\Psi_n)](\pi^L) < 0$, and set $S' = S + \{j\}$. For example, one can take j as the minimizer of $[d_j(\Psi_n)](\pi^L)$. Go to step 2.

   Step 2: Let $\pi^\star_{S'}$ be the minimizer of $\Psi_n(\pi)$ over all functions π with support included in S'. The components of $\pi^\star_{S'}$ are denoted $\pi^\star_{S',l}$ for $l \in S'$. Two cases must be considered:
   (a) If for all $l \in S'$, $\pi^\star_{S',l} \geqslant 0$, then set $\pi^L = \pi^\star_{S'}$, $S = S'$ and return to Step 1.
   (b) If not, let l be defined as follows:

$$l = \arg\min_{j'} \left\{ \varepsilon_{j'} = \frac{\pi^L_{j'}}{\pi^L_{j'} - \pi^\star_{S,j'}} \text{ for } j' \text{ such that } \pi^\star_{S,j'} < \pi^L_{j'} \right\}.$$

   Set $S' = S' - \{l\}$ and return to Step 2.

Theorem 2. *The above algorithm gives $\widehat{\pi}^L$.*

The second part in computing $\widehat{\pi}_n$ consists in choosing L such that $\widehat{\pi}^L = \widehat{\pi}_n$. The following theorem provides a characterization of such a L.

Theorem 3. *Let $L \geqslant \widetilde{s}_n + 1$. If $\widehat{\pi}^L$ is a probability measure, then $\widehat{\pi}^L = \widehat{\pi}_n$.*

Thus, to compute $\widehat{\pi}_n$, we carry out the optimisation over $\mathcal{M}^L$ for increasing values of $L \geqslant \widetilde{s}_n + 1$ until the condition $\sum_{j \geqslant 1} \widehat{\pi}_j^L = 1$ is satisfied. The support of $\widehat{\pi}_n$ is finite, so the condition is fulfilled in a finite number of steps.

3. Theoretical properties of the constrained LSE

The aim of this section is to compare the constrained LSE with the unconstrained estimator $\widetilde{p}_n$ and to investigate its consistency and rate of convergence. In Subsection 3.1, it is proved that $\widehat{p}_n$ is a probability mass function that outperforms the empirical estimator $\widetilde{p}_n$ in the ℓ_2-sense in the case p_0 is convex. Moreover, we provide a comparison between the absolute moments of the constrained and unconstrained estimated distributions. Consistency and rate of convergence of $\widehat{p}_n$ are investigated in Subsection 3.2, in the misspecified setting.

3.1. Comparing the constrained and the unconstrained estimators

In this subsection we investigate the benefits of using the constrained LSE rather than the (unconstrained) empirical estimator $\widetilde{p}_n$. It is proved in the following theorem that the constrained LSE is closer, in the ℓ_2-sense, to any convex $f \in \ell^2(\mathbb{N})$ than is $\widetilde{p}_n$.

Theorem 4. *Let p_0, $\widetilde{p}_n$ and $\widehat{p}_n$ be defined as in Section 2.2. Then,*

$$\|\widehat{p}_n - f\|_2 \leqslant \|\widetilde{p}_n - f\|_2 \tag{9}$$

for all $f \in \mathcal{C}$, with a strict inequality if $\widetilde{p}_n$ is non-convex. Moreover, if $\Delta p_0(i) = 0$ for at least an integer $i \geqslant 1$, then, for all $f \in \mathcal{C}$, we have

$$\liminf_{n \to \infty} \mathbb{P}\Big(\|f - \widehat{p}_n\|_2 < \|f - \widetilde{p}_n\|_2 \Big) \geqslant 1/2. \tag{10}$$

It immediatly follows that if p_0 is convex on $\mathbb{N}$, then both (9) and (10) hold true with $f = p_0$ since $p_0 \in \ell_2(\mathbb{N})$, see Section 2.1. This shows that the probability for $\widehat{p}_n$ to be strictly closer to p_0 than is $\widetilde{p}_n$, is strictly positive

(and even, it is at least 1/2) whenever p_0 is convex with a linear part on its support.

In what follows, we consider the estimation of some characteristics of the distribution p_0, namely the expectation, the centered absolute moments and the probability at 0. As estimators for these characteristics, we naturally consider similar characteristics of the constrained and the unconstrained estimators. Theorem 5 states that the distributions $\widetilde{p}_n$ and $\widehat{p}_n$ have the same expectation, but the centered absolute moments of the distribution $\widetilde{p}_n$ are lower than those of the distribution $\widehat{p}_n$. In particular, the variance of the distribution $\widehat{p}_n$ is greater than the variance of $\widetilde{p}_n$. Moreover, the constrained estimator $\widehat{p}_n(0)$ is greater than or equal to the unconstrained estimator $\widetilde{p}_n(0)$. The performance of $\widehat{p}_n$ is compared with that of $\widetilde{p}_n$ through simulation studies in Section 4.

Theorem 5. *Let $\widetilde{p}_n$ and $\widehat{p}_n$ be defined as in Section 2.2. We have for all $u \geqslant 1$, and $0 \leqslant a \leqslant \widehat{s}_n$*

$$\sum_{i \geqslant 0} |i-a|^u \widetilde{p}_n(i) \leqslant \sum_{i \geqslant 0} |i-a|^u \widehat{p}_n(i). \tag{11}$$

Moreover, $\sum_{i \geqslant 0} i\widetilde{p}_n(i) = \sum_{i \geqslant 0} i\widehat{p}_n(i)$ and $\widehat{p}_n(0) \geqslant \widetilde{p}_n(0)$.

In the case of discrete log-concave distribution, Equations (3.6) and (3.5) in Balabdaoui et al. (2012) show that in contrast to our case, the absolute moments of the constrained maximum likelihood estimator distribution are smaller than those of the empirical distribution whereas similar to our case, the empirical distribution and the constraint estimated distribution have the same expectation.

3.2. Consistency and rate of convergence

An important issue is on how the estimator behaves asymptotically, as the sample size n goes to infinity. It is expected that, at least in the case of a well specified model, the estimator converges to the true p_0 as fast as possible. The two following theorems give the asymptotic behaviour of $\widehat{p}_n$, both in the case of a well specified model and in the case of a misspecified model.

We begin with the case of a well specified model, that is the case where p_0 is convex. We prove that $\widehat{p}_n$ is a consistent estimator of p_0 in the ℓ_r-sense with the rate of convergence $\sqrt{n}$, for all $r \in [2, \infty]$. Note that the $\sqrt{n}$-rate

is expected in the discrete setting. See for instance Jankowski and Wellner (2009) for the constrained least-squares estimator of a monotone probability mass function, and Balabdaoui et al. (2012) for the constrained maximum likelihood estimator of a log-concave probability mass function.

Theorem 6. *Let p_0, $\widetilde{p}_n$ and $\widehat{p}_n$ be defined as in Section 2.2. If p_0 is convex, then $\sqrt{n}\|p_0 - \widehat{p}_n\|_r = O_P(1)$ for all $r \in [2, \infty]$.*

Let us proceed with the misspecified setting, where p_0 is possibly non convex. Since $\widehat{p}_n$ is convex, the limit of the sequence $(\widehat{p}_n)_n$ (if it exists in some sense) is typically convex. This means that $\widehat{p}_n$ cannot converge to p_0 if p_0 is non convex. Instead, it is expected that $\widehat{p}_n$ converges to a convex approximation of p_0. Such convergence results for shape-constrained estimators in a misspecified model were already obtained, for instance in Balabdaoui et al. (2012) and Cule et al. (2010) in the case of the log-concave constraint. See also Patilea (2001) for the case of the constrained MLE in the general case of convex dominated models.

The following theorem proves $\sqrt{n}$-convergence of $\widehat{p}_n$ to $p_{0\mathcal{C}}$, the convex discrete function which is closest, in the ℓ_2-sense, to the true p_0. Specifically, $p_{0\mathcal{C}}$ is the unique minimizer (see Subsection 2.1) of $\|f - p_0\|_2$ over $f \in \mathcal{C}$. It can be proved that $p_{0\mathcal{C}}$ is a probability mass function, but as this property is not used in the sequel, the proof is omitted for brevity. Note that $p_{0\mathcal{C}}$ is well defined whether p_0 is convex or not: it reduces to $p_{0\mathcal{C}} = p_0$ in the case where p_0 is convex whereas $p_{0\mathcal{C}} \neq p_0$ in the case where p_0 is non convex.

Theorem 7. *Let p_0, $\widetilde{p}_n$ and $\widehat{p}_n$ be defined as in Section 2.2. We have $\sqrt{n}\|p_{0\mathcal{C}} - \widehat{p}_n\|_r = O_P(1)$ for all $r \in [2, \infty]$.*

Thus, even if the convex hypothesis is violated, $\widehat{p}_n$ converges in the ℓ_r-sense at the $\sqrt{n}$-rate to the probability mass function that is closest in the ℓ_2-sense to p_0. In particular, if p_0 is not too far from being convex, then $\widehat{p}_n$ is still sensible. This is illustrated on simulations in Section 4 below.

4. Simulation study

4.1. Simulation design

We designed a simulation study to assess the quality of the constrained estimator $\widehat{p}_n$ as compared to the unconstrained estimator $\widetilde{p}_n$.

We considered two shapes for the distribution p_0: the geometric $\mathcal{G}(\gamma)$ with $\gamma = .9, .5, .1$, the support of which is infinite, and the pure triangular distribution T_j with $j = 20, 5, 2$. For each distribution, we considered 9 sample sizes: $n = 10^\alpha$ with $\alpha \in \{1, 1.25, 1.5, 1.75, 2, 2.25, 2.5, 2.75, 3\}$. We also considered the Poisson distribution with mean λ, which is convex as long as λ is smaller than $\lambda^* = 2 - \sqrt{2} \simeq .59$. We considered $\lambda = .59$, .8 and 1. For each simulation configuration, 1000 random samples were generated. The simulations were carried out with `R (www.r-project.org)`, using functions available at the following web-site `http://www.jouy.inra.fr/mia/sylviehuet_en`.

4.2. Global fit

We first compared the fit of the estimated distributions $\widehat{p}_n$ and $\widetilde{p}_n$ to the entire distribution p_0. To this aim, for each simulated sample, we computed the ℓ_2-error for $\widehat{p}_n$

$$\ell_2(\widehat{p}_n, p_0) = \sum_i [\widehat{p}_n(i) - p_0(i)]^2,$$

and likewise for $\widetilde{p}_n$. The ℓ_2-loss is estimated by the mean of 1000 independent replications of the ℓ_2-error calculated on the basis of 1000 simulations and the results are displayed in Figure 1.

As expected from Theorem 4, the constrained estimator $\widehat{p}_n$ outperforms the empirical estimator in all configurations in the ℓ_2-sense. The difference is more sensitive in the triangular case because of the existence of a region where p_0 is linear. The empirical estimator $\widetilde{p}_n$ gets better and closer to $\widehat{p}_n$ as the true distribution p_0 becomes more convex, i.e., for $\gamma = .9$ or $j = 2$. Note that the fit of the unconstrained estimator improves when the true distribution gets more convex.

Although it is not investigated in Theorem 4, we also considered the Kolmogorov loss: $K(\widehat{p}, p_0) = \sup_i |\widehat{P}_n(i) - P_0(i)|$, where P_0 is the true cumulative distribution function (cdf) and $\widehat{P}_n$ is the constrained cdf, the Hellinger loss: $\sum_i \left(\sqrt{\widehat{p}_n(i)} - \sqrt{p_0(i)}\right)^2 /2$ and the total variation loss: $\sum_i |\widehat{p}_n(i) - p_0(i)|/2$. The same behavior was observed for these loss functions than for the ℓ_2-loss (results not shown). We thus observed that the constrained estimator $\widehat{p}_n$ outperforms the empirical estimator for all considered losses.

4.3. Some characteristics of interest

In this section, we consider the estimation of some characteristics of the distribution, namely the variance, the entropy and the probability at 0. For

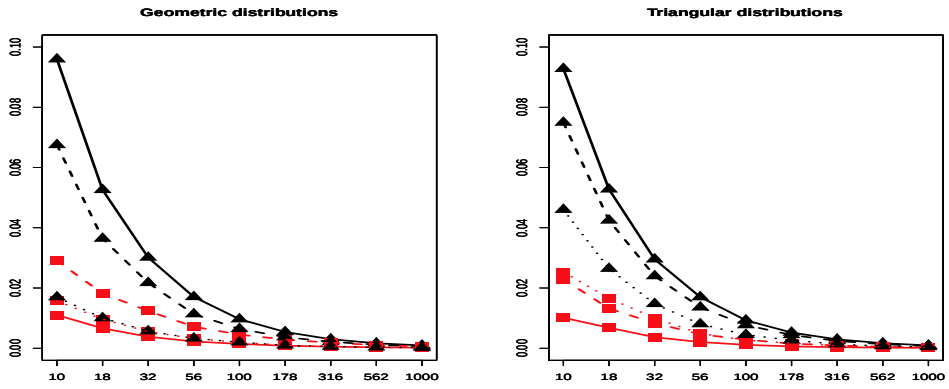

Figure 1: ℓ_2**-loss.** Empirical risk as a function of the sample size n. Black: $\widetilde{p}_n$, red: $\widehat{p}_n$. Solid (–): $\gamma = 1$ or $j = 20$, dashed (- -): $\gamma = .5$ or $j = 5$, dotted ($\cdots$): $\gamma = .9$ or $j = 2$.

each of these characteristics, denoted by $\theta(p)$, we measured the performance in terms of relative standard error:

$$\sqrt{\mathbb{E}\left(\theta(\widehat{p}_n) - \theta(p_0)\right)^2} / \theta(p_0) \; .$$

The expectation was estimated by the mean over 1000 simulations.

As shown in Section 2, the means of the empirical and constrained distributions are equal, whereas the variance of the constrained distribution is larger than the variance of the empirical one. Denoting by μ_k the centered moment of order k of p_0, the mean and variance of the empirical variance are respectively

$$\frac{n-1}{n}\mu_2 \qquad \text{and} \qquad \frac{n-1}{n^3}\left((n-1)\mu_4 - (n-3)\mu_2^2\right) .$$

Figure 2 shows that the relative standard error of the empirical estimator is smaller than that of the constrained one.

We also investigated the estimation of the entropy

$$H(p) = -\sum_{i \geqslant 0} p(i) \log p(i),$$

which is often used in ecology as a diversity index as it is maximal when all species have the same abundancies (p is the uniform distribution) and

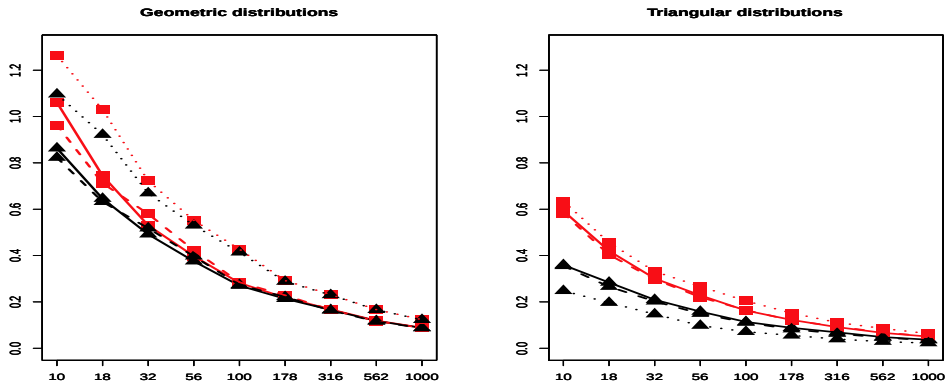

Figure 2: **Variance.** Relative standard error of the variance as a function of the sample size n. Same legend as Figure 1.

is zero when only one species is present. As shown in Figure 3, $H(\widehat{p}_n)$ is a better estimate of the true entropy than $H(\widetilde{p}_n)$, in most situations; the difference between the two estimators vanishes when the true distribution becomes more convex. The worst performance of $H(\widehat{p}_n)$ are obtained when the true distribution is T_2. Note that this distribution is a special case since more than half of the estimation errors consist in adding a component T_j ($j > 2$) in the mixture (6), which result in an increase of the entropy.

We then considered the estimation of the probability mass $p(0)$. Theorem 5 showed that the constrained estimator $\widehat{p}_n(0)$ is greater than or equal to the empirical estimator $\widetilde{p}_n(0)$, which is known to be unbiased. However, Figure 4 shows that the constrained estimator $\widehat{p}_n$ still provides a more accurate estimate of $p_0(0)$ than $\widetilde{p}_n$.

4.4. Robustness to non-convexity

To investigate the robustness of the constrained estimator to non-convexity, we consider the Poisson distribution with mean λ, which is convex as long as λ is smaller that $\lambda^* = 2 - \sqrt{2} \simeq .59$. We studied how $\widetilde{p}_n$ and $\widehat{p}_n$ behave, in terms of ℓ_2-loss, when λ exceeds λ^*.

The left panel of Figure 5 displays the Poisson distributions with respective means λ^* equal to .8 and 1. Figure 5 (right) shows that the ℓ_2-loss of the constrained estimator increases with λ. However for small sample sizes, $\widehat{p}_n$ still provides a better fit than $\widetilde{p}_n$, at least for $\lambda \leqslant 1$. The performance of $\widehat{p}_n$ is

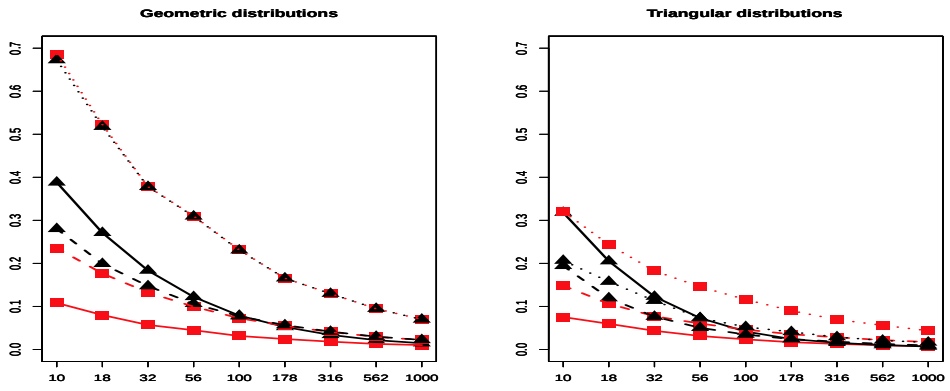

Figure 3: **Entropy.** Relative standard error of the estimated entropy estimators as a function of the sample size n. Same legend as Figure 1.

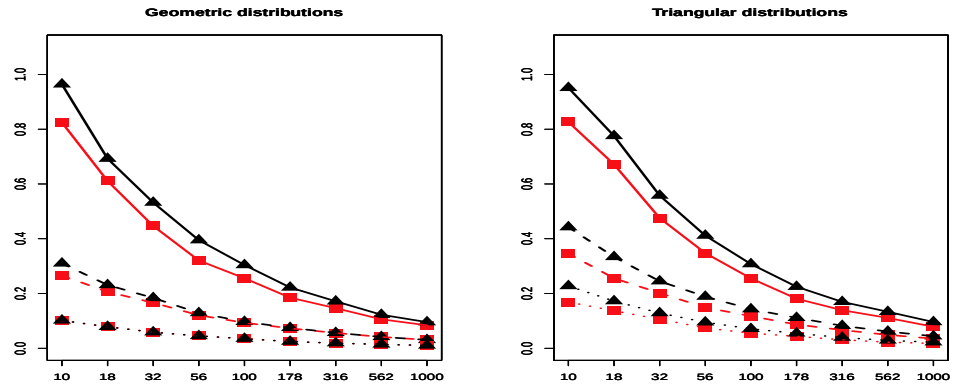

Figure 4: **Probability mass in 0.** Relative standard error of the estimated probability mass in zero as a function of the sample size n. Same legend as Figure 1.

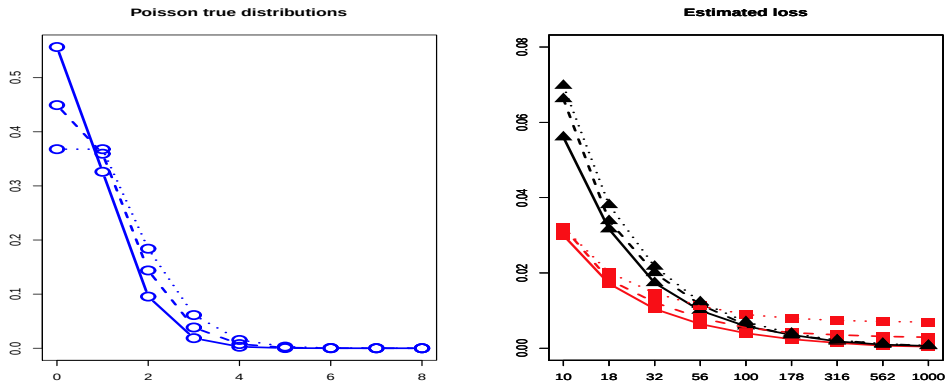

Figure 5: Left: Three different Poisson distributions. Solid (–):$\lambda = \lambda^*$, dashed (- -):$\lambda = .8$, dotted $(\cdots)$:$\lambda = 1$. Right: empirical ℓ_2-loss as a function of n. Black: $\widetilde{p}_n$, red: $\widehat{p}_n$.

dramatically altered when the sample size becomes large and the convexity assumption is strongly violated.

5. Estimation of abundance distribution

As recalled in Section 1, estimating the abundance distribution of species in a community is an old and classical problem in ecology. Similar problems also arise in other fields such as insurance, as shown in the examples below. Yet, we outline here the problem in ecological terms.

Suppose that n different species have been observed and for $s = 1, \dots n$ denote by X_s the number of sampled individuals belonging to species s. The X_s's are supposed to be i.i.d.. Some species may be present in the area whereas they have not been observed, so the X_s's are distributed according to the truncated version p^+ of the true abundance distribution p:

$$p^+(i) = p(i)/(1 - p(0)), \qquad i \geqslant 1.$$

The goal is to estimate the proportion of unobserved species, which is equivalent to estimate $p(0)$ based on an i.i.d. sample from p^+.

5.1. A convexity-based estimate of abundance

We propose here an estimator of $p(0)$ based on the hypothesis that p is a discrete convex distribution. More precisely, we assume that p is a discrete

convex abundance distribution as defined below. Our hypothesis relies on the representation of a discrete convex distribution as a mixture of triangular distributions, see Appendix A. Our interpretation of the mixture is that the set of species is separated into groups, each species having probability π_j to belong to group j of species whose abundance distribution is the triangular distribution T_j. As the first component T_1 is a Dirac mass in 0, it refers to species for which the only abundance that could be observed is 0. This group simply defines *absent* species, and therefore π_1 has to be zero.

Definition 1. *A distribution p over $\mathbb{N}$ is a convex abundance distribution if there exist positive $\pi_j, j \geqslant 2$ such that $p(i) = \sum_{j \geqslant 2} \pi_j T_j(i)$.*

Note that the Poisson distribution with parameter λ is a convex abundance distribution if, and only if, $\lambda = 2 - \sqrt{2}$.

Our estimator of $p(0)$ is based on the fact that $\pi_1 = 0$ implies that $\Delta p(1) = 0$. This means that $p(0) = 2p(1) - p(2)$. Denoting by $\widehat{p}_n^+$ the constraint convex LSE of p^+, we define

$$\widehat{p}_n(0) = (2\widehat{p}_n^+(1) - \widehat{p}_n^+(2))/(1 + 2\widehat{p}_n^+(1) - \widehat{p}_n^+(2)).$$

The denominator is here to ensure that the $\widehat{p}_n(i)$'s sum to one.

5.2. Examples

Datasets. To illustrate the convex estimator we propose for estimating $p(0)$, we considered three datasets available in the SPECIES R package developed by Wang (2011): Butterfly, Traffic and Microbial. We also considered the Bird dataset analysed in Norris and Pollock (1998), as it displays a more complex distribution.

Fit of the convex estimate. Figure 6 shows that the fit of the convex estimate $\widehat{p}^+$ is quite good for all the datasets. This suggests that the convexity assumption is quite reasonable. We also considered other examples from the literature that all display similar fits (not shown). Note that the prolongation at $i = 0$ corresponds to an estimate of $p(0)/(1 - p(0))$, and not to $p(0)$.

Comparison with truncated Poisson mixtures. A penalized non-parametric maximum likelihood method was proposed by Wang and Lindsay (2005) for estimating a mixture of truncated Poisson distributions where

$$p(i) = \sum_{k \geqslant 1} \omega_k \mathcal{P}(i, \lambda_k) \text{ with } \mathcal{P}(i, \lambda_k) = \frac{\lambda_k^i \exp(-\lambda_k)}{i!}.$$

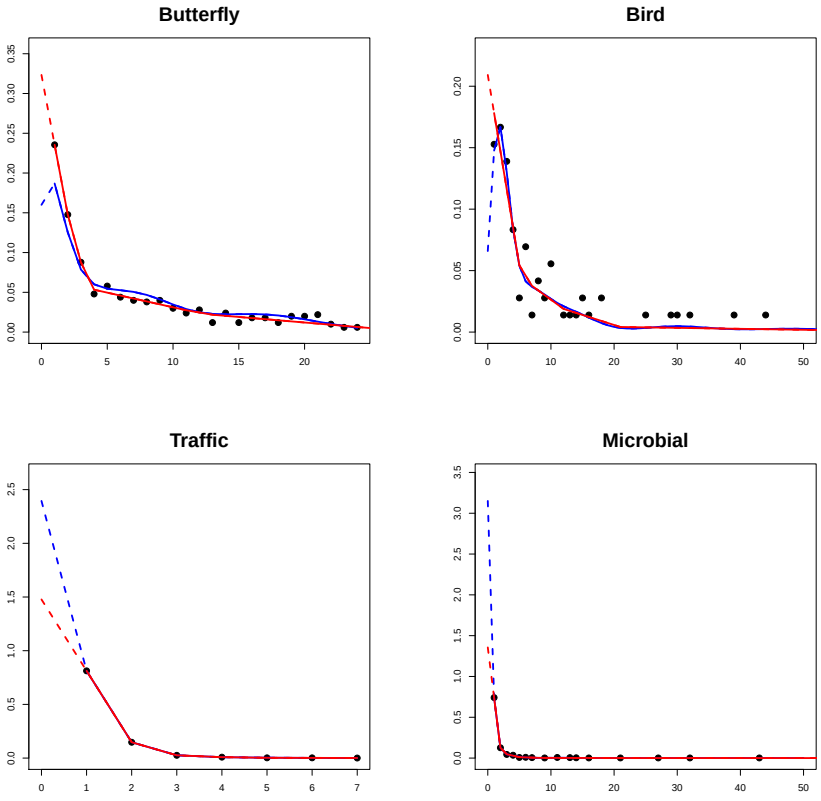

Figure 6: Fit of the convex abundance distribution. Dots (•) = empirical distribution $\widetilde{p}^{+}$, solid red (–) = convex estimates $\widehat{p}^{+}$, solid blue (–) = Poisson mixture estimates $\overline{p}^{+}$.

Dataset	d	K	$\overline{\mathcal{L}}^+$	$\overline{p}(0)$	$\widehat{\mathcal{L}}^+$	$\widehat{p}(0)$
Butterfly	501	4	-1371.9	0.138	-1366.3	0.244
Bird	72	5	-215.3	0.062	-216.4	0.173
Traffic	1621	2	-1007.6	0.705	-1006.7	0.596
Microbial	514	5	-540.2	0.759	-540.2	0.576

Table 1: Comparison of the truncated Poisson mixture (with K components) and of the convex estimates. Log-likelihood of the truncated distribution ($\overline{\mathcal{L}}^+$ and $\widehat{\mathcal{L}}^+$, resp.) and estimates of $p(0)$ ($\overline{p}(0)$ and $\widehat{p}(0)$, resp.).

Its estimator will be denoted by $\overline{p}^+(i) = \sum_{k\geqslant 1} \overline{\omega}_k \mathcal{P}(i, \overline{\lambda}_k)$.

To compare the quality of the two approaches, we computed the log-likelihood of the truncated convex estimate:

$$\widehat{\mathcal{L}}^+ = n \sum_{i\geqslant 1} \widetilde{p}_n(i) \log \widehat{p}_n^+(i), \quad \overline{\mathcal{L}}^+ = n \sum_{i\geqslant 1} \widetilde{p}_n(i) \log \overline{p}_n(i).$$

Figure 6 and Table 1 show that the fit of $\widehat{p}_n^+$ and $\overline{p}_n^+$ are similar, according to both the graphical representation and the likelihood. Note that the value at $i = 0$ can not be compared between the two methods, as the normalizing constants are different. However, due to very different underlying assumptions on the abundance distribution p (Poisson mixture versus convex), the estimates of $p(0)$ strongly differ, the ratio between them reaching almost 3 for the smallest dataset (Bird).

The results display different behaviors and more can be said when considering the components of the fitted Poisson mixture. For the Butterfly and Bird datasets, none of the Poisson components is convex ($\overline{\lambda}_k > 2 - \sqrt{2}$) so the mixture is not convex, which results in $\overline{p}(0) < \overline{p}(1)$. In each of the two other datasets (Traffic and Microbial), the first Poisson component is convex ($\overline{\lambda}_1 = .336$ and $.225$, resp.), but does not satisfy Definition 1 so the corresponding π_1 is not 0. This distribution is not a convex abundance distribution which explains the differences in the estimations of $p(0)$.

As the true distribution p is unknown, it is not possible to decide which approach provides the most accurate estimate. Still, we emphasize that the estimate we propose relies on very mild assumption (convexity) about the true distribution.

6. Proofs

In the case $\widetilde{s}_n = 0$ i.e., $\widetilde{p}_n$ is the dirac distribution with mass 1 at point 0, we have $\widehat{p}_n = \widetilde{p}_n \in \mathcal{C}_1$ and we face a trivial case. Thus, in the sequel, we restrict ourselves to the case $\widetilde{s}_n \geqslant 1$.

Notation. For notational convenience, unless otherwise stated we denote by $\| \, . \, \|$ the ℓ_2-norm on $\ell_2(\mathbb{N})$.

We denote by N_n the number of distinct values of the X_i's and by $X_{(1)}, \ldots, X_{(N_n)}$ these distinct values rearranged in increasing order, i.e., such that $X_{(1)} < \cdots < X_{(N_n)}$. We set $\widetilde{r}_n = X_{(1)}$ and $\widetilde{s}_n = X_{(N_n)}$. We define

$$Q_n(f) = \frac{1}{2}\|f\|^2 - \langle f, \widetilde{p}_n \rangle$$

and

$$\bar{f}(i) = \begin{cases} f(i) \text{ for all } i \in \{0, \ldots, \widetilde{s}_n\} \\ \max\{f(\widetilde{s}_n) + (f(\widetilde{s}_n) - f(\widetilde{s}_n - 1))(i - \widetilde{s}_n)\,,\, 0\} \text{ for all } i \geqslant \widetilde{s}_n \end{cases} \tag{12}$$

for all $f \in \ell_2(\mathbb{N})$. It should be noticed that minimizing $\|f - \widetilde{p}_n\|^2$ is equivalent to minimizing $Q_n(f)$ since $\|f - \widetilde{p}_n\|^2 = 2Q_n(f) + \|\widetilde{p}_n\|^2$. Thus by definition, $\widehat{p}_n$ is the unique minimizer of $Q_n(f)$ over $f \in \mathcal{C}$.

6.1. Proof of Theorem 1

In order to prove Theorem 1, we first prove in Lemma 1 that $\widehat{p}_n$ has a non-empty finite support. Then, after some intermediate results, we prove in Lemma 3 below that $\widehat{s}_n \geq \widetilde{s}_n$ and $\widehat{p}_n \in \mathcal{C}_1$. Since $\mathcal{C}_1 \subset \mathcal{C}$, Theorem 1 follows.

Lemma 1. *The constrained LSE $\widehat{p}_n$ is the unique minimizer of $Q_n(f)$ over the set of all $f \in \mathcal{C}$ satisfying $f = \bar{f}$. Moreover, $\widehat{p}_n$ has a non-empty finite support.*

Proof. Note that for all $f \in \mathcal{C}$, setting $f_+(i) = \max\{f(i), 0\}$ for all $i \geqslant 0$ yields $f_+ \in \mathcal{C}$ and

$$\|f_+ - \widetilde{p}_n\| \leqslant \|f - \widetilde{p}_n\|$$

with a strict inequality if $f \neq f_+$ since $\widetilde{p}_n(i) \geqslant 0$ for all i. Therefore, any candidate $f \in \mathcal{C}$ to be a minimizer of Q_n takes non-negative values. By convexity, any canditate f satisfies $\bar{f}(i) \leqslant f(i)$ for all $i \geqslant \widetilde{s}_n$ and therefore,

$$Q_n(f) - Q_n(\bar{f}) = \sum_{i > \widetilde{s}_n} (f^2(i) - \bar{f}^2(i))/2 \geqslant 0$$

with a strict inequality in the case $f \neq \bar{f}$. Since $\bar{f} \in \mathcal{C}$, any candidate f to be a minimizer of Q_n over $\mathcal{C}$ has $f = \bar{f}$. This proves the first assertion.

If $\widehat{p}_n(\widetilde{s}_n) = 0$, then $\widehat{p}_n$ clearly has a finite support included in $\{0, \ldots, \widetilde{s}_n - 1\}$. On the other hand, if $\widehat{p}_n(\widetilde{s}_n) > 0$, then we must have $\widehat{p}_n(\widetilde{s}_n - 1) > \widehat{p}_n(\widetilde{s}_n)$, since otherwise, we would have $\widehat{p}_n(i) = \widehat{p}_n(\widetilde{s}_n)$ for all $i \geqslant \widetilde{s}_n$ so that $Q_n(\widehat{p}_n) = \infty$. Therefore, $\widehat{p}_n = \widehat{\widetilde{p}_n}$, which has a finite support. To conclude the proof of the lemma, let us prove by contradiction that this support is non-empty. Let $k = 1 + \min_j \{\widetilde{p}_n(j) \neq 0\}$. It is easy to check that there exists a strictly positive a such that $Q_n(aT_k) < 0$. As $Q_n(0) = 0$, $\widehat{p}_n$ cannot be identically zero, so the support of $\widehat{p}_n$ is not empty. □

The following lemma provides a precise characterization of $\widehat{p}_n$. It is the counterpart, in the discrete case, of Lemma 2.2 in Groeneboom et al. (2001) for the continuous case. For every $f \in \mathcal{C}$, we define

$$F_f(j) = \sum_{i=0}^{j} f(i) \text{ and } H_f(j) = \sum_{i=0}^{j} F_f(i) \tag{13}$$

for all integers $j \geqslant 0$, and $F_f(j) = H_f(j) = 0$ for all integers $j < 0$. Thus, F_f is a distribution function in the case $f \in \mathcal{C}_1$.

Lemma 2. *For all $l \geqslant 1$ we have*

$$H_{\widehat{p}_n}(l-1) \geqslant H_{\widetilde{p}_n}(l-1) \tag{14}$$

with an equality if l is a knot of $\widehat{p}_n$.

Conversely, if $p \in \mathcal{C}$ satisfies $H_p(l-1) \geqslant H_{\widetilde{p}_n}(l-1)$ for all $l \geqslant 1$ with an equality if l is a knot of $\widehat{p}_n$, then $p = \widehat{p}_n$.

Proof. For every $\varepsilon > 0$ and $l \geqslant 1$, define $q_{\varepsilon l}$ by $q_{\varepsilon l}(i) = \widehat{p}_n(i)$ for all $i \geqslant l$ and

$$q_{\varepsilon l}(i) = \widehat{p}_n(i) + \varepsilon(l - i)$$

for all $i \in \{0, \ldots, l\}$. Thus, $q_{\varepsilon l}$ is the sum of convex functions, which implies that $q_{\varepsilon l} \in \mathcal{C}$ for all ε, l. Since $\widehat{p}_n$ minimizes Q_n over $\mathcal{C}$, we have $Q_n(q_{\varepsilon l}) \geqslant Q_n(\widehat{p}_n)$ for all ε, l and therefore,

$$\liminf_{\varepsilon \downarrow 0} \frac{1}{\varepsilon}\big(Q_n(q_{\varepsilon l}) - Q_n(\widehat{p}_n)\big) \geqslant 0$$

for all $l \geqslant 1$. This simplifies to

$$\sum_{i=0}^{l-1} \widehat{p}_n(i)(l-i) \geqslant \sum_{i=0}^{l-1} \widetilde{p}_n(i)(l-i)$$

for all $l \geqslant 1$ and can be rewritten as

$$\sum_{j=0}^{l-1} \sum_{i=0}^{j} \widehat{p}_n(i) \geqslant \sum_{j=0}^{l-1} \sum_{i=1}^{j} \widetilde{p}_n(i)$$

for all $l \geqslant 1$, which is precisely (14). To prove the equality case, note that $(1+\varepsilon)\widehat{p}_n \in \mathcal{C}$ for all $\varepsilon > -1$. Therefore, for all $\varepsilon > -1$ we have

$$Q_n\big((1+\varepsilon)\widehat{p}_n\big) \geqslant Q_n(\widehat{p}_n).$$

Distinguishing the cases $\varepsilon > 0$ and $\varepsilon < 0$ we obtain

$$\liminf_{\varepsilon \downarrow 0} \frac{1}{\varepsilon}\big(Q_n((1+\varepsilon)\widehat{p}_n) - Q_n(\widehat{p}_n)\big) \geqslant 0$$

and

$$\limsup_{\varepsilon \uparrow 0} \frac{1}{\varepsilon}\big(Q_n((1+\varepsilon)\widehat{p}_n) - Q_n(\widehat{p}_n)\big) \leqslant 0.$$

Both limits are equal, so their common value is equal to zero, which can be written as

$$\sum_{i \geqslant 0} \widehat{p}_n(i)\big(\widehat{p}_n(i) - \widetilde{p}_n(i)\big) = 0.$$

Since $f(i) = F_f(i) - F_f(i-1)$ for all $f \in \mathcal{C}$ and $i \in \mathbb{N}$, we arrive at

$$\begin{aligned}
0 &= \sum_{i \geqslant 0} \widehat{p}_n(i)\Big(F_{\widehat{p}_n}(i) - F_{\widehat{p}_n}(i-1) - F_{\widetilde{p}_n}(i) + F_{\widetilde{p}_n}(i-1)\Big) \\
&= \sum_{i \geqslant 0} \widehat{p}_n(i)\Big(F_{\widehat{p}_n}(i) - F_{\widetilde{p}_n}(i)\Big) - \sum_{i \geqslant 1} \widehat{p}_n(i)\Big(F_{\widehat{p}_n}(i-1) - F_{\widetilde{p}_n}(i-1)\Big). \\
&= \sum_{i \geqslant 0} \Big(\widehat{p}_n(i) - \widehat{p}_n(i+1)\Big)\Big(F_{\widehat{p}_n}(i) - F_{\widetilde{p}_n}(i)\Big).
\end{aligned}$$

Since $F_f(i) = H_f(i) - H_f(i-1)$ for all $f \in \mathcal{C}$ and $i \in \mathbb{N}$, a similar change of indices as above then yields

$$\begin{aligned}
0 &= \sum_{i \geqslant 0} \Big((\widehat{p}_n(i) - \widehat{p}_n(i+1)) - (\widehat{p}_n(i+1) - \widehat{p}_n(i+2))\Big)\Big(H_{\widehat{p}_n}(i) - H_{\widetilde{p}_n}(i)\Big) \\
&= \sum_{i \geqslant 0} \Delta\widehat{p}_n(i+1)\Big(H_{\widehat{p}_n}(i) - H_{\widetilde{p}_n}(i)\Big).
\end{aligned}$$

It follows from (14) that $H_{\widehat{p}_n}(i) \geqslant H_{\widetilde{p}_n}(i)$ for all $i \geqslant 0$, and we have $\Delta\widehat{p}_n(i+1) \geqslant 0$ by convexity of $\widehat{p}_n$. A sum of non-negative numbers is equal to zero if and only if all of these numbers are equal to zero, so we conclude that

$$\Delta\widehat{p}_n(i+1)\Big(H_{\widehat{p}_n}(i) - H_{\widetilde{p}_n}(i)\Big) = 0$$

for all $i \geqslant 0$. Hence, $H_{\widehat{p}_n}(i) = H_{\widetilde{p}_n}(i)$ for all $i \geqslant 0$ with $\Delta\widehat{p}_n(i+1) > 0$. Setting $l = i+1$, this means that we have an equality in (14) if l is a knot of $\widehat{p}_n$.

Conversely, consider $p \in \mathcal{C}$ such that $H_p(i) \geqslant H_{\widetilde{p}_n}(i)$ for all $i \geqslant 0$ with an equality if $\Delta p(i+1) > 0$. Then we have

$$0 = \sum_{i \geqslant 0} \Delta p(i+1)\big(H_p(i) - H_{\widetilde{p}_n}(i)\big). \tag{15}$$

On the other hand, for all $f \in \mathcal{C}$ we have

$$\begin{aligned} Q_n(f) - Q_n(p) &= \frac{1}{2}\|f-p\|^2 + \langle p - \widetilde{p}_n, f - p\rangle \\ &\geqslant \langle p - \widetilde{p}_n, f - p\rangle. \end{aligned} \tag{16}$$

Rearranging the indices as above yields

$$\langle p - \widetilde{p}_n, f - p\rangle = \sum_{i \geqslant 0} \Big(\Delta\big(f-p\big)(i+1)\Big)\big(H_p(i) - H_{\widetilde{p}_n}(i)\big).$$

Combining this with (15) and (16) yields

$$Q_n(f) - Q_n(p) \geqslant \sum_{i \geqslant 0} \Delta f(i+1)\big(H_p(i) - H_{\widetilde{p}_n}(i)\big).$$

The right-hand side is non-negative since $H_p(i) \geqslant H_{\widetilde{p}_n}(i)$ for all $i \geqslant 0$ and f is convex on $\mathbb{N}$, so we conclude that $Q_n(f) \geqslant Q_n(p)$ for all $f \in \mathcal{C}$, whence $p = \widehat{p}_n$. □

Lemma 3. *We have $\widehat{s}_n \geqslant \widetilde{s}_n$, and $\widehat{p}_n \in \mathcal{C}_1$.*

Proof. We first prove that

$$F_{\widetilde{p}_n}(\widehat{s}_n + 1) = F_{\widehat{p}_n}(\widehat{s}_n + 1). \tag{17}$$

By definition of $\widehat{s}_n$, $\widehat{s}_n + 1$ is a knot of $\widehat{p}_n$, so it follows from Lemma 2 that

$$\sum_{j=0}^{\widehat{s}_n} F_{\widehat{p}_n}(j) = \sum_{j=0}^{\widehat{s}_n} F_{\widetilde{p}_n}(j). \tag{18}$$

Using Lemma 2 again we obtain

$$\sum_{j=0}^{\widehat{s}_n+1} F_{\widehat{p}_n}(j) \geqslant \sum_{j=0}^{\widehat{s}_n+1} F_{\widetilde{p}_n}(j),$$

which, combined with (18) shows that $F_{\widehat{p}_n}(\widehat{s}_n + 1) \geqslant F_{\widetilde{p}_n}(\widehat{s}_n + 1)$.

Let us consider first the case where $\widehat{s}_n \geqslant 1$. We have

$$\sum_{j=0}^{\widehat{s}_n-1} F_{\widehat{p}_n}(j) \geqslant \sum_{j=0}^{\widehat{s}_n-1} F_{\widetilde{p}_n}(j)$$

which, combined with (18) shows that $F_{\widehat{p}_n}(\widehat{s}_n) \leqslant F_{\widetilde{p}_n}(\widehat{s}_n)$. But $\widehat{p}_n(\widehat{s}_n+1) = 0$ by definition of $\widehat{s}_n$, so we also have $F_{\widehat{p}_n}(\widehat{s}_n + 1) = F_{\widehat{p}_n}(\widehat{s}_n)$ and therefore,

$$F_{\widetilde{p}_n}(\widehat{s}_n) \geqslant F_{\widehat{p}_n}(\widehat{s}_n + 1) \geqslant F_{\widetilde{p}_n}(\widehat{s}_n + 1).$$

By definition, $F_{\widetilde{p}_n}$ is non-decreasing, so we conclude that (17) holds.

Consider now the case $\widehat{s}_n = 0$. Then, $\widetilde{p}_n(1) = 0$, since otherwise we could find $q \in \mathcal{C}$ such that $Q_n(q) < Q_n(\widehat{p}_n)$: take $q(0) = \widehat{p}_n(0)$, $0 < q(1) \leqslant \widetilde{p}_n(1)$ and $q(i) = 0$ for all $i > 1$. Similarly, we have $\widehat{p}_n(0) = \widetilde{p}_n(0)$, since otherwise, we could find $q \in \mathcal{C}$ such that $Q_n(q) < Q_n(\widehat{p}_n)$: take $q(0) = \widetilde{p}_n(0)$ and $q(i) = 0$ for all $i > 0$. Hence,

$$F_{\widehat{p}_n}(1) = \widehat{p}_n(0) = \widetilde{p}_n(0) = F_{\widetilde{p}_n}(1),$$

which completes the proof of (17).

To prove that $\widehat{s}_n \geqslant \widetilde{s}_n$, we argue by contradiction. Assume for a while that $\widehat{s}_n = \widetilde{s}_n - 1$. This means that $\widehat{p}_n(i) = 0$ for all $i \geqslant \widetilde{s}_n$ and $\widehat{p}_n(\widetilde{s}_n - 1) > 0$. In this case, we can modify $\widehat{p}_n$ to a $q \in \mathcal{C}$ such that $q(i) = \widehat{p}_n(i)$ for all $i < \widetilde{s}_n$, $0 < q(\widetilde{s}_n) \leqslant \widetilde{p}_n(\widetilde{s}_n)$, and $q(i) = 0$ for all $i > \widetilde{s}_n$. Then we have

$$2\big(Q_n(q) - Q_n(\widehat{p}_n)\big) = \big(q(\widetilde{s}_n) - \widetilde{p}_n(\widetilde{s}_n)\big)^2 - \big(\widetilde{p}_n(\widetilde{s}_n)\big)^2 < 0.$$

This is a contradiction since $\widehat{p}_n$ minimizes Q_n and therefore, $\widehat{s}_n \neq \widetilde{s}_n - 1$. Assume now that $\widehat{s}_n < \widetilde{s}_n - 1$. Then, $F_{\widetilde{p}_n}(\widehat{s}_n + 1) < 1$, so (17) yields

$$F_{\widehat{p}_n}(j) = F_{\widehat{p}_n}(\widehat{s}_n + 1) < 1$$

for all $j \geqslant \widehat{s}_n + 1$. Therefore, for all $l > \widetilde{s}_n$ we have

$$\sum_{j=0}^{l-1} \big(F_{\widehat{p}_n}(j) - F_{\widetilde{p}_n}(j)\big) = \sum_{j=0}^{\widetilde{s}_n - 1} \big(F_{\widehat{p}_n}(j) - F_{\widetilde{p}_n}(j)\big) + (l - \widetilde{s}_n)\big(F_{\widehat{p}_n}(\widehat{s}_n + 1) - 1\big),$$

which tends to $-\infty$ as $l \to \infty$. This is a contradiction since from Lemma 2, this has to remain non-negative for all l. We conclude that $\widehat{s}_n \geqslant \widetilde{s}_n$. Combining this with (17) yields $F_{\widehat{p}_n}(\widehat{s}_n + 1) = 1$, which means that $\widehat{p}_n$ is a genuine probability mass function. This completes the proof of the lemma. □

6.2. Proof of Proposition 1

From Lemma 1, $\widehat{p}_n$ belongs to the set of functions $f \in \mathcal{C}$ with $f = \bar{f}$, so $\widehat{p}_n$ is linear on $\{\widetilde{s}_n - 1, \ldots, \widehat{s}_n\}$. Consider $p \in \mathcal{C}$, fix $j \in \{1, \ldots, N_n - 1\}$, and define $p_l, p_r \in \mathcal{C}$ by $p_l(i) = p(i)$ for all $i \leqslant X_{(j)} + 1$ and all $i \geqslant X_{(j+1)}$ and p_l is linear on $\{X_{(j)}, ..., X_{(j+1)} - 1\}$, whereas $p_r(i) = p(i)$ for all $i \leqslant X_{(j)}$ and all $i \geqslant X_{(j+1)} - 1$ and p_r is linear on $\{X_{(j)} + 1, ..., X_{(j+1)}\}$. Setting $q(i) = \max\{p_l(i), p_r(i)\}$ for all $i \in \mathbb{N}$, we obtain that $q \in \mathcal{C}$ has at most two knots on $\{X_{(j)} + 1, ..., X_{(j+1)} - 1\}$ and in case it has two knots, they occur at consecutive points. We have $q(X_{(j)}) = p(X_{(j)})$ for all j, and $q \leqslant p$ by convexity of p. Since $\widetilde{p}_n(i) > 0$ if and only if $i = X_{(j)}$ for some j, this implies that $Q_n(q) \leqslant Q_n(p)$ with a strict inequality if $p \neq q$. Therefore, p could be a minimizer of Q_n only if $p = q$. This implies that $\widehat{p}_n$ has at most two knots on $\{X_{(j)} + 1, \ldots, X_{(j+1)} - 1\}$. A similar argument shows that $\widehat{p}_n$ is linear on $\{0, \ldots, X_{(1)} + 1\}$. □

6.3. Proof of Theorem 2

The theorem follows from Lemmas 4 and 5 given below.

Let us define the following notation. Let ν, μ be two measures in $\mathcal{M}$. The derivative of Ψ_n in the direction ν calculated in μ is defined as follows:

$$\left[D_\nu(\Psi_n)\right](\mu) = \lim_{\varepsilon \downarrow 0} \frac{1}{\varepsilon} \left(\Psi_n(\mu + \varepsilon\nu) - \Psi_n(\mu)\right),$$

for all μ and ν such that $\Psi_n(\mu)$ and $\Psi_n(\nu)$ are finite. It can be written as

$$[D_\nu(\Psi_n)](\mu) = \sum_{j \geqslant 1} \nu_j [d_j(\Psi_n)](\mu) \tag{19}$$

where $[d_j(\Psi_n)](\mu)$, which can be computed thanks to (8), is defined by

$$[d_j(\Psi_n)](\mu) = \lim_{\varepsilon \downarrow 0} \frac{1}{\varepsilon} \left(\Psi_n(\mu + \varepsilon\delta_j) - \Psi_n(\mu)\right),$$

where δ_j is the dirac measure at point j.

Lemma 4. *Let $\widetilde{s}_n$ be the maximum of the support of $\widetilde{p}_n$ and $L \geqslant \widetilde{s}_n + 1$. Then we have the following result: $\widehat{\pi}^L = \arg\min_{\mu \in \mathcal{M}^L} \Psi_n(\mu)$ is equivalent to*

$$[d_j(\Psi_n)](\widehat{\pi}^L) \geqslant 0 \; \forall 1 \leqslant j \leqslant L, \text{ and } [d_j(\Psi_n)](\widehat{\pi}^L) = 0 \; \forall j \in Supp(\widehat{\pi}^L) \tag{20}$$

This lemma is the same as Lemma 1 in Groeneboom et al. (2008) and its proof is omitted.

Lemma 5. *Let us define the following quantities: let $\pi = \sum_{i=1}^{L-1} a_i \delta_{j_i}$ be the minimizer of Ψ_n over the set of positive measures spanned by $\{\delta_{j_i},\ 1 \leqslant i \leqslant L-1\}$, let j_L be an integer such that $j_L \neq j_i$ for all $1 \leqslant i \leqslant L-1$, and $[d_{j_L}(\Psi_n)](\pi) < 0$ and let $\pi^\star = \sum_{i=1}^{L} b_i \delta_{j_i}$ be the minimizer of Ψ_n over the set spanned by $\{\delta_{j_i},\ 1 \leqslant i \leqslant L\}$. We have the two following results.*

1. *The coefficient b_L is strictly positive, and there exists $\varepsilon > 0$ such that $\pi + \varepsilon(\pi^\star - \pi)$ is non negative, and such that $\Psi_n(\pi + \varepsilon(\pi^\star - \pi)) < \Psi_n(\pi)$.*
2. *Assume that some coefficients $b_i, 1 \leqslant i \leqslant L-1$ are negative and let*

$$\ell = \arg \min_{1 \leqslant i \leqslant L-1} \left\{ \frac{a_i}{a_i - b_i} \text{ for } i \text{ such that } b_i < a_i \right\}.$$

Let $\pi^{\star\star} = \sum_{i=1, i \neq \ell}^{L} c_i \delta_{j_i}$ be the minimizer of Ψ_n over the set spanned by $\{\delta_{j_i},\ 1 \leqslant i \leqslant L, i \neq \ell\}$. Then c_L is positive.

Proof. The first part of the Lemma corresponds to Lemma 2 in Groeneboom et al. (2008) and its proof is omitted.

The second part of the lemma follows by noting that removing ℓ from the support remains to take $\varepsilon = a_\ell/(a_\ell - b_\ell)$. Therefore

$$\Psi_n(\pi^{\star\star}) \leqslant \Psi_n\left(\pi + \frac{a_\ell}{a_\ell - b_\ell}(\pi^\star - \pi)\right) < \Psi_n(\pi).$$

Just as above, we have

$$\lim_{\varepsilon \downarrow 0} \frac{1}{\varepsilon}\left(\Psi_n((1-\varepsilon)\pi + \varepsilon\pi^{\star\star}) - \Psi_n(\pi)\right) = c^L\left[d_{j_L}(\Psi_n)\right](\pi).$$

It follows that $c^L > 0$. □

6.4. Proof of Theorem 3

Let us begin with the following lemma.

Lemma 6. *If $\widehat{\pi}^L = \arg\min_{\mu \in \mathcal{M}^L} \Psi_n(\mu)$, then for all $j \geqslant 1$,*

$$\left[d_{L+j}(\Psi_n)\right](\widehat{\pi}^L) = b\left(\sum_{i=1}^{L} \widehat{\pi}_i^L - 1\right),$$

for some positive constant b depending on j and on the maximum of the support of $\widehat{\pi}^L$.

Proof. Let us consider two cases according to wether $\widehat{\pi}_L^L$ equals 0 or not.

Suppose that $\widehat{\pi}_L^L > 0$, and write

$$\left[d_{L+j}(\Psi_n)\right](\widehat{\pi}^L) = \sum_{l=1}^{L-1} T_{L+j}(l)\left(\sum_{j'=l+1}^{L} \widehat{\pi}_{j'}^L T_{j'}(l) - \widetilde{p}_n(l)\right).$$

Because for $0 \leqslant l \leqslant L-1$, $T_{L+j}(l) = aT_L(l) + b$, for constants a and b depending on L and j, we get

$$\left[d_{L+j}(\Psi_n)\right](\widehat{\pi}^L) = a\left[d_L(\Psi_n)\right](\widehat{\pi}^L) + b\left[\sum_{l=1}^{L-1}\sum_{j'=l+1}^{L} \widehat{\pi}_{j'}^L T_{j'}(l) - 1\right].$$

Following Lemma 4, $\left[d_L(\Psi_n)\right](\widehat{\pi}^L) = 0$, and we get

$$\left[d_{L+j}(\Psi_n)\right](\widehat{\pi}^L) = b\left(\sum_{i=1}^{L} \widehat{\pi}_i^L \sum_{l=0}^{j-1} T_j(l) - 1\right) = b\left(\sum_{i=1}^{L} \widehat{\pi}_i^L - 1\right).$$

If $\widehat{\pi}_L^L = 0$, then $\widehat{\pi}^L \in \mathcal{M}^{L_1}$ for some $L_1 < L$. Thanks to Lemma 4, we know that $\widehat{\pi}^L$ is the minimizer of Ψ_n over $\mathcal{M}^{L_1}$. Then we can show that $[d_{L_1+j}(\Psi_n)](\widehat{\pi}^L) = 0$ for all $j \geqslant 1$ exactly as we have done in the case $\widehat{\pi}_L^L > 0$.

To conclude the proof of Theorem 3, note first that for all $L' \leqslant L$, we have $\mathcal{M}^{L'} \subset \mathcal{M}^L$, which implies

$$\Psi_n(\widehat{\pi}^L) \leqslant \Psi_n(\widehat{\pi}^{L'}). \tag{21}$$

Second, it follows from Lemmas 4 and 6 that if $\sum_{i=1}^L \widehat{\pi}_i^L = 1$, then for all $L' \geqslant L$, $\widehat{\pi}^{L'} = \widehat{\pi}^L$. Therefore Equation (21) holds for all positive integers L', which implies that $\Psi_n(\widehat{\pi}^L) \leqslant \Psi_n(\pi)$ for all measures $\pi \in \mathcal{M}$ with a finite support. Therefore $\widehat{\pi}^L = \widehat{\pi}_n$. □

6.5. Proof of Theorem 4

Inequality (9) follows from (3) when applied with $\mathcal{S} = \mathcal{C}$ and $g = \widetilde{p}_n$. To investigate the strict inequality case, note that for all $f \in \mathcal{C}$, we have

$$\begin{aligned}\|f - \widetilde{p}_n\|^2 &= \|f - \widehat{p}_n\|^2 + \|\widehat{p}_n - \widetilde{p}_n\|^2 + 2\langle f - \widehat{p}_n, \widehat{p}_n - \widetilde{p}_n\rangle \\ &\geqslant \|f - \widehat{p}_n\|^2 + 2\langle \widehat{p}_n - \widetilde{p}_n, f - \widehat{p}_n\rangle\end{aligned}$$

with a strict inequality in the case where $\widetilde{p}_n$ is non-convex since in that case, $\widetilde{p}_n \neq \widehat{p}_n$. But from (4), where we replace f by $\widetilde{p}_n$ and g by f, it follows that $\langle \widehat{p}_n - \widetilde{p}_n, f - \widehat{p}_n\rangle \geqslant 0$. Hence the strict inequality case in (9).

In order to prove (10), it now suffices to prove that

$$\liminf_{n\to\infty} \mathbb{P}(\widetilde{p}_n \text{ is non-convex}) \geqslant 1/2. \tag{22}$$

For this task, note that

$$\begin{aligned}\mathbb{P}(\widetilde{p}_n \text{ is non-convex}) &\geqslant \mathbb{P}\left(\Delta\widetilde{p}_n(i) < 0\right) \\ &\geqslant \mathbb{P}\left(\sqrt{n}\Delta(\widetilde{p}_n - p_0)(i) < 0\right),\end{aligned}$$

since $\Delta p_0(i) = 0$. From the central limit theorem, the random variable $\sqrt{n}\Delta(\widetilde{p}_n - p_0)(i)$ converges, as $n \to \infty$, to a centered Gaussian variable X with a non-degenerate variance and therefore,

$$\liminf_{n\to\infty} \mathbb{P}\left(\widetilde{p}_n \text{ is non-convex}\right) \geqslant \mathbb{P}(X \leqslant 0).$$

Inequality (22) follows since $\mathbb{P}(X \leqslant 0) = 1/2$. □

6.6. Proof of Theorem 5

Since $\widehat{p}_n$ minimizes $Q_n(f)$ over $f \in \mathcal{C}$, for all $q : \mathbb{N} \to \infty$ such that $\widehat{p}_n - \varepsilon q \in \mathcal{C}$ for all $\varepsilon > 0$ sufficiently close to zero, we have

$$0 \;\leqslant\; \lim_{\varepsilon \searrow 0} \frac{Q_n(\widehat{p}_n - \varepsilon q) - Q_n(\widehat{p}_n)}{\varepsilon},$$

that is

$$0 \leqslant \langle \widetilde{p}_n - \widehat{p}_n, q \rangle. \tag{23}$$

For $u \geqslant 1$ and $0 \leqslant a \leqslant \widehat{s}_n$, let us consider

$$q(i) = 1 - \left(\frac{|i-a|}{\widehat{s}_n + 1 - a} \right)^u, \text{ for } 1 \leqslant i \leqslant \widehat{s}_n,$$

and $q(i) = 0$ for $i > \widehat{s}_n$. Then $\widehat{p}_n - \varepsilon q \in \mathcal{C}$ for all $\varepsilon \in (0, \widehat{p}_n(\widehat{s}_n)/q(\widehat{s}_n)]$, so we have (23). Recall that $\widehat{s}_n \geqslant \widetilde{s}_n$, see Theorem 1, so that $\widehat{p}_n(i) = \widetilde{p}_n(i) = 0$ for all $i > \widehat{s}_n$. Since $\sum_{i \geqslant 0} \widehat{p}_n(i) = \sum_{i \geqslant 0} \widetilde{p}_n(i) = 1$, (23) implies (11).

Similarly, for all $q : \mathbb{N} \to \infty$ such that $\widehat{p}_n + \varepsilon q \in \mathcal{C}$ for all ε sufficiently close to zero, we have

$$0 \leqslant \lim_{\varepsilon \searrow 0} \frac{Q_n(\widehat{p}_n + \varepsilon q) - Q_n(\widehat{p}_n)}{\varepsilon} \quad \text{and} \quad 0 \geqslant \lim_{\varepsilon \nearrow 0} \frac{Q_n(\widehat{p}_n + \varepsilon q) - Q_n(\widehat{p}_n)}{\varepsilon},$$

which implies $0 = \langle \widetilde{p}_n - \widehat{p}_n, q \rangle$. Setting

$$q(i) = \left(1 - \frac{i}{\widehat{s}_n + 1} \right) \text{ for } 1 \leqslant i \leqslant \widehat{s}_n, \text{ and } q(i) = 0, \text{ for } i > \widehat{s}_n,$$

we have $\widehat{p}_n + \varepsilon q \in \mathcal{C}$ for all $\varepsilon \geqslant -\widehat{p}_n(\widehat{s}_n)/q(\widehat{s}_n)$, hence $\sum_{i \geqslant 0} i \widetilde{p}_n(i) = \sum_{i \geqslant 0} i \widehat{p}_n(i)$.

It remains to prove that $\widehat{p}_n(0) \geqslant \widetilde{p}_n(0)$. Argue by contradiction and assume that $\widehat{p}_n(0) < \widetilde{p}_n(0)$. Define $q(0) = \widetilde{p}_n(0)$ and $q(i) = \widehat{p}_n(i)$ for all $i \geqslant 1$. Then, $q \in \mathcal{C}$ and we have $Q_n(q) < Q_n(\widehat{p}_n)$. This is a contradiction since $\widehat{p}_n$ minimizes Q_n over $\mathcal{C}$. This completes the proof of the theorem. □

6.7. Proof of Theorems 6 and 7

Invoking (3) with $\mathcal{S} = \mathcal{C}$, $f = p_0$ and $g = \widetilde{p}_n$ yields

$$\sqrt{n} \| p_{0\mathcal{C}} - \widehat{p}_n \|_2 \leqslant \sqrt{n} \| p_0 - \widetilde{p}_n \|_2 = O_P(1),$$

see for instance Theorem 3.1 in Jankowski and Wellner (2009) for the right equality. Moreover, we have $\|f\|_r \leqslant \|f\|_2$ for all $r \in [2, \infty]$ and all $f \in \ell_2(\mathbb{N})$ so with $f = p_0 - \widehat{p}_n$ we obtain:

$$\sqrt{n}\|p_{0\mathcal{C}} - \widehat{p}_n\|_r \leqslant \sqrt{n}\|p_0 - \widehat{p}_n\|_2 = O_P(1),$$

which completes the proof of Theorem 7. In the particular case where p_0 is convex, we have $p_{0\mathcal{C}} = p_0$, which yields Theorem 6. □

Appendix A. More about discrete convex functions

The aim of this appendix is to prove that any discrete convex function in $\ell_2(\mathbb{N})$ is a combination of the triangular probabilities defined in (5).

Theorem 8. *Let $f \in \ell_2(\mathbb{N})$.*

1. *We have $f \in \mathcal{C}$ if and only if there exists $\pi \in \mathcal{M}$ satisfying (6).*
2. *Assume $f \in \mathcal{C}$. Then, π in (6) is uniquely defined by*

$$\pi_j = \frac{j(j+1)}{2}\Delta f(j) \text{ for all } j \geqslant 1. \tag{A.1}$$

Moreover, π is a probability measure on $\mathbb{N}\backslash\{0\}$ if and only if $f \in \mathcal{C}_1$.

Note that according to (A.1), the support of π is included in the set of knots of f. In the case where f has a non-empty finite support with maximal point denoted by s, the greatest knot of f is $s+1$, since $\Delta f(s+1) = f(s) > 0$, so the support of π is included in the finite set $\{1, \ldots, s+1\}$.

Proof of Theorem 8. Assume $f \in \mathcal{C}$ and define π by (A.1). As f is convex, π takes non-negative values, so $\pi \in \mathcal{M}$. Moreover, for all $i \in \mathbb{N}$ we have

$$\sum_{j \geqslant i+1} \pi_j T_j(i) = \sum_{j \geqslant i+1} \Delta f(j)(j-i) = \sum_{j \geqslant i+1} \sum_{k=1}^{j-i} \Delta f(j)$$

and therefore,

$$\sum_{j \geqslant i+1} \pi_j T_j(i) = \sum_{k \geqslant 1} \sum_{j \geqslant i+k} \Delta f(j) = \sum_{k \geqslant 1} \big(f(i+k-1) - f(i+k)\big).$$

This reduces to (6), which proves that π exists. Conversely, every $f \in \ell_2(\mathbb{N})$ satisfying (6) for some $\pi \in \mathcal{M}$ is convex, hence the first assertion of the theorem. To prove the second assertion, we consider $f \in \mathcal{C}$. In view of what precedes, f satisfies (6) for some $\pi \in \mathcal{M}$, so we have

$$f(i-1) - f(i) = \sum_{j \geqslant i} \pi_j \big(T_j(i-1) - T_j(i)\big) = \sum_{j \geqslant i} \frac{2\pi_j}{j(j+1)}$$

for all $i \geqslant 1$. By convexity of f we conclude that

$$0 \leqslant \big(f(i-1) - f(i)\big) - \big(f(i) - f(i+1)\big) = \frac{2\pi_i}{i(i+1)}$$

for all $i \geqslant 1$, which implies that π is uniquely defined by (A.1). Moreover,

$$\sum_{i \geqslant 0} f(i) = \sum_{i \geqslant 0} \sum_{j \geqslant i+1} \pi_j T_j(i) = \sum_{i \geqslant 0} \sum_{j \geqslant 1} \pi_j T_j(i)$$

since $T_j(i) = 0$ for all $j \leqslant i$. This implies that

$$\sum_{i \geqslant 0} f(i) = \sum_{j \geqslant 1} \pi_j \big(\sum_{i \geqslant 0} T_j(i) \big)$$

where $\sum_{i \geqslant 0} T_j(i) = 1$. This completes the proof of the theorem. □

References

Balabdaoui, F., Jankowski, H., Rufibach, K., Pavlides, M., 2012. Asymptotic distribution of the log-concave mle and some application. J. Roy. Statist. Soc., Ser. B To appear.

Balabdaoui, F., Wellner, J. A., 2007. Estimation of a k-monotone density: limit distribution theory and the spline connection. Annals of Statistics 35, 2536–2564.

Barlow, R. E., Bartholomew, D. J., Bremmer, J. M., Brunk, H. D., 1972. Statistical Inference under Order Restrictions. Wiley, New-York.

Böhning, D., Dietz, E., Kuhnert, R., Schön, D., 2005. Mixture models for capture-recapture count data. Stat. Methods Appl. 14, 29–43.

Böhning, D., Kuhnert, R., 2006. Equivalence of truncated count mixture distributions and mixtures of truncated count distributions. Biometrics 62 (4), 1207–1215.

Chao, A., Shen, T., 2004. Nonparametric prediction in species sampling. Journal of Agricultural Bioogical an Environmental Statistics 9 (3), 253–269.

Cule, M., Samworth, R., Stewart, M., 2010. Maximum likelihood estimation of a multi-dimensional log-concave density. J. Roy. Stat. Soc., Ser. B 72 (5), 545–607.
URL http://dx.doi.org/10.1111/j.1467-9868.2010.00753.x

Dümbgen, L., Hüsler, A., Rufibach, K., 2007. Active Set and EM Algorithms for Log-Concave Densities Based on Complete and Censored Data. ArXiv e-prints.
URL http://adsabs.harvard.edu/abs/2007arXiv0707.4643D

Dümbgen, L., Rufibach, K., 2011. logcondens: Computations related to univariate log-concave density estimation. Journal of Statistical Software 39 (6), 1–28.
URL http://www.jstatsoft.org/v39/i06/

Groeneboom, P., Jongbloed, G., Wellner, J. A., 2001. Estimation of a convex function: characterizations and asymptotic theory. Ann. Statist. 29, 1653–1698.

Groeneboom, P., Jongbloed, G., Wellner, J. A., 2008. The support reduction algorithm for computing non-parametric function estimates in mixture models. Scand. J. Statist. 35 (3), 385–399.
URL http://dx.doi.org/10.1111/j.1467-9469.2007.00588.x

Jankowski, H. K., Wellner, J., 2009. Estimation of a discrete monotone distribution. Electron J Stat 3, 1567–1605.

Murota, K., 2009. Recent Developments in Discrete Convex Analysis. In: Cook, WJ and Lovasz, L and Vygen, J (Ed.), Research Trends in Combinatorial Optimization. Springer-Verlag, Berlin, pp. 219–260.

Norris, J., Pollock, K., 1998. Non-parametric MLE for Poisson species abundance models allowing for heterogeneity between species. Environmental and Ecological Statistics 5 (4), 391–402.

Patilea, V., 2001. Convex models, MLE and misspecification. Ann. Statist. 29, 94–123.

Walther, G., 2002. Detecting the presence of mixing with multiscale maximum likelihood. J. Amer. Statist. Assoc. 97 (458), 508–513.
URL http://dx.doi.org/10.1198/016214502760047032

Wang, J.-P., 2011. SPECIES: An R package for species richness estimation. Journal of Statistical Software 40 (9), 1–15.
URL http://www.jstatsoft.org/v40/i09/

Wang, J.-P., Lindsay, B., 2005. A penalized nonparametric maximum likelihood approach to species richness estimation. Journal of the American Statistical Association 100 (471), 942–959.