Electronic Journal of Statistics
Vol. 11 (2017) 1–49
ISSN: 1935-7524
DOI: 10.1214/16-EJS1220

Estimation of a discrete probability under constraint of k-monotonicity

Jade Giguelay

Laboratoire de Mathématiques d'Orsay, Université Paris-Sud, CNRS, Université Paris-Saclay, 91405 Orsay, France
MaIAGE INRA, Université Paris-Saclay, 78350 Jouy-en-Josas, France
e-mail: jade.giguelay@ens-paris-saclay.fr
url: http://maiage.jouy.inra.fr/jgiguelay

Abstract: We propose two least-squares estimators of a discrete probability under the constraint of k-monotonicity and study their statistical properties. We give a characterization of these estimators based on the decomposition on a spline basis of k-monotone sequences. We develop an algorithm derived from the Support Reduction Algorithm and we finally present a simulation study to illustrate their properties.

MSC 2010 subject classifications: Least squares, non-parametric estimation, k-monotone discrete probability, shape constraint, Support Reduction Algorithm.

Received July 2016.

1. Introduction

The estimation of a density under shape constraint is a statistical problem that was first raised by Grenander [18] in 1956 in the case of a density under monotonicity constraint. Over the past 30 years, there has been several studies of estimators under shape constraint, most of them being maximum likehood estimators or least squares estimators. In these cases, the authors characterize the estimators, study the asymptotic law and the rate of convergence and discuss the implementation. For such studies, the constraints are, for example, the monotonicity, the convexity or the log-concavity (if $\log(f)$ is concave, f is log-concave) and the k-monotonicity.

The k-monotonicity notion was introduced by Knopp [23] in 1929 for discrete functions: it generalizes to k^{th} order the notion of convex series (or 2-monotone series) and corresponds to the positivity of a k-th derivative function. In 1941, Feller [16] extended that definition to k-monotone continuous functions and Williamson [27] enabled characterizing these k-monotone functions with their decomposition in spline basis:

Property 1 (Williamson, 1955)**.** *Let g be a continuous function. Let $k \geqslant 2$. The function g is k-monotone if and only if there exists a nonnegative mesure μ on $\mathbb{R}^*$ such that:*

$$g(x) = \int_0^\infty (t-x)_+^{k-1} d\mu(t).$$

Consequently k-monotone functions can be described with an integral form. The estimation of a k-monotone distribution has been studied by Balabdaoui et al. ([1],[6],[5]): they proposed the maximum likehood and the least-squares estimators under k-monotonicity constraint for the continuous space and studied their theoretical properties (consistency and rate of convergence) as well as their limit distribution. They also discussed the adaptation of an algorithm proposed by Groeneboom et al. [20].

The k-monotonicity may seem as a restrictive assumption but the study of this shape constraint is motivated by the existence of widely-used k-monotone parametric families like Gamma's, Weibull's or Beta's laws (for particular choices of parameters) for the continuous case and their discretized versions or the Geometric law for the discrete case. The Poisson distribution is also k-monotone for some choice of its parameter λ. Moreover, for each λ, it is possible to determine the exact degree l of k-monotonicity (i.e. l-monotone and not $(l+1)$-monotone), see Property 8 page 14 for more details.

Most of the work on estimation under shape constraint was focused on densities with a support on $\mathbb{R}$ or on an interval, but recently, discrete probabilities have gained interest because of their numerous applications in ecology or financial mathematics (see [14] or [25]). Jankowski and Wellner [21] recently studied the estimation under monotonicity constraint and Balabdaoui et al. [3] investigated the log-concave discrete densities. More recently, the estimation of a convex discrete distribution was treated by Durot et al. ([13], [14]) and Balabdaoui et al. [3].

In this article we propose two least-squares estimators of a k-monotone discrete probability with $k \geqslant 2$. The first one is the projection of the empirical estimator on the set of k-monotone sequences, the second one is the projection of the empirical estimator on the set of k-monotone probabilities. We show the existence of these estimators and give a characterization for each one of them which is based on the decomposition on a spline basis of k-monotone sequences showed by Lefevre and Loisel [25]. Thanks to this characterization we generalize some results for the convex case ($k = 2$, see [13]) to $k > 2$, as for example the comparison with the empirical estimator (Theorem 3).

However differences between the convex case and the case $k > 2$ arised. First the projection of a discrete probability on the set of k-monotone sequences is not a probability in general when $k > 2$. This structural property of the set of k-monotone functions, $k \geqslant 3$, justifies the definition of two different estimators while they are equivalent in the convex case. Secondly the proofs of some other properties require new tools. In fact the results about the support of our estimator require control of the decreasing of the tail of k-monotone probabilities while truncation is sufficient in the convex case.

Because k-monotone sequences are l-monotone for $l \leqslant k$ (See Property 2 page 4) one can ask what is the advantage of using in the estimation procedure the *correct* k (i.e. the integer k such that the true distribution p is k-monotone and not $(k+1)$-monotone) instead of a smaller k, in particular $k = 2$. In fact one could expect that projection on a smaller set decreases the l_2-loss of the estimator. This phenomenon is illustrated in a simulation study in Section 6.

Although the construction of our estimators is inspired by the work of Balabdaoui [1] our results are not deduced from the continuous case. In fact, for $k \geqslant 3$, unlike for the convex case, we could neither construct a k-monotone density that goes through the points of a k-monotone sequence nor approach a k-monotone sequence with a k-monotone density. It is however interesting to note that connecting the points of a convex sequence can provide a convex continuous function because no differentiability assumption is required in the definition of convexity. Moreover the practical implementation of the estimator is structurally different from the continuous case. For the discrete case we implement the estimators using exact iterative algorithms inspired by the Support Reduction Algorithm described in Groeneboom et al. [20] and we discuss a practical stopping criterion (see Section 4).

Differences with the continuous case also emerged when we consider the rate of convergence in terms of l_2-error since our estimators are consistent with typical parametric $\sqrt{n}$-rate of convergence (see Theorem 3).

The paper is organized as follows: the definition of the k-monotonicity and some properties about k-monotone discrete sequences are reminded in Section 2, and a characterization of the estimator is given in Section 3.1. Statistical properties about this estimator are presented in Section 3.3. In Section 4, a method to implement the estimator in practice using the Support Reduction Algorithm of Groeneboom et al. [20] is presented. The stopping criterion for this algorithm, which differs from the convex case ($k = 2$) is also discussed. In Section 5 we discuss the possibility to choose an estimator on the set of k-monotone sequences instead of the set of k-monotone probabilities. Finally a simulation study is given in Section 6. All functions mentionned in this article are implemented as an R package named **pkmon** and available on the Comprehensive R Archive Network (`https://CRAN.R-project.org/package=pkmon`).

2. Characterizing k-monotone sequences

Let us begin with 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 $(f(j), j \in \mathbb{N})$. For all $r \in \mathbb{N} \setminus \{0\}$, the classical L^r-norm of f is defined as follows:

$$\|f\|_r = \left(\sum_{j \geqslant 0} |f(i)|^r \right)^{1/r}, \quad \|f\|_\infty = \sup_{i \geqslant 0} |f(i)|,$$

and we denote by $L^r(\mathbb{N})$ the set of functions f such that $\|f\|_r$ is finite. In particular L^2 is an Hilbert space and the associated scalar product is denoted $\langle , \rangle$.

For any integer $k \geqslant 1$, let $\Delta^k f$ be the k^{th} differential operator of f defined for all $i \geqslant 0$ by the following recurrence equation:

$$\Delta^1 f(i) \quad = \quad f(i+1) - f(i)$$

$$\Delta^k f(i) \quad = \quad \Delta^{k-1} f(i+1) - \Delta^{k-1} f(i).$$

It is easy to see that the operator Δ^k satisfies the following equation:

$$\forall i \in \mathbb{N},\ \Delta^k f(i) = \sum_{h=0}^{k} \binom{k}{h} (-1)^{k-h} f(h+i).$$

Definitions

- A sequence f on $\mathbb{N}$ is k-monotone if
$$(-1)^k \Delta^k f(i) \geqslant 0 \text{ for all } i \in \mathbb{N}.$$
- Let f be a k-monotone sequence. The integers i such that $(-1)^k \Delta^k f(i) > 0$ are called the k-knots of f. If for all integer i in the support of f, the quantities $(-1)^k \Delta^k f(i)$ are strictly positive, f is said to be strictly k-monotone.
- The maximum s_f of the support of f is defined as
$$s_f = \min_{j \geqslant 0} \{\forall i > j, p(i) = 0\}$$
and may be infinite.

Let us remark that if the support of a k-monotone sequence f is finite, then s_f is a k-knot.

A k-monotone function on $L^1(\mathbb{N})$ is for example a non-negative and non-increasing polynomial function of degree $k-1$, such as $f(i) = \max(0, m-i)^{k-1}$ for some positive constant m.

It should be noticed that there exists a link between the k-monotonicity and the $(k-1)$-monotonicity stated in the following property:

Property 2. *For all $k \geqslant 2$, if $p \in L^1(\mathbb{N})$ is a k-monotone discrete sequence then p is j-monotone and strictly j-monotone on its support for all $j < k$.*

This property, shown in Section 7.3.1, is not true in general in the continuous case (see Balabdaoui [1] for example).

Finally we will denote by $\mathcal{S}^k$ the set of k-monotone sequences that are in $L^1(\mathbb{N})$, and by $\mathcal{P}^k$ the set of k-monotone probabilities on $\mathbb{N}$. We denote by $\mathcal{P}$ the set of probabilities on $\mathbb{N}$.

Decomposition on a spline basis

The characterization of k-monotone functions defined on $\mathbb{R}$ as a mixture of polynomial functions has been established by Lévy [26], and the inversion formula that specifies the mixture function follows from the results of Williamson [27] (see Lemma 1. in [6] for example). In the case of k-monotone sequences a similar decomposition has been simultaneously established for convex sequences by Durot et al. [13], and in the more general case of k-monotonicity by Lefevre and Loisel [25]. Many of our proofs will rely on this decomposition.

For any integer k, let us define a basis of spline sequences $(\bar{Q}_j^k)_{j\in\mathbb{N}\setminus\{0\}}$ as follows:

$$\forall i \in \mathbb{N},\ \bar{Q}_j^k(i) = \binom{j-i+k-1}{k-1}\mathbb{1}_{\{j\geqslant i\}} = \frac{(j-i+k-1)\dots(j-i+1)}{(k-1)!}\mathbb{1}_{\{j\geqslant i\}}. \tag{1}$$

Let m_j^k be the mass of $\bar{Q}_j^k$: $m_j^k = \sum_{i=0}^{j}\bar{Q}_j^k(i)$ and $Q_j^k = \bar{Q}_j^k/m_j^k$ the normalized spline. We can now formulate the mixture representation of k-monotone sequences.

Property 3. *Let $f \in L^1(\mathbb{N})$.*

- *The sequence f is k-monotone if and only if there exists a positive measure π on $\mathbb{N}$, such that for all $i \in \mathbb{N}$, $f(i)$ satisfies:*

$$f(i) = \sum_{j\geqslant 0}\pi(j)Q_j^k(i) = \sum_{j\geqslant i}\pi(j)Q_j^k(i). \tag{2}$$

- *If f is k-monotone, the measure π is unique and defined as follows:*

$$\forall j \geqslant 0,\ \pi(j) = (-1)^k\Delta^k f(j)m_j^k. \tag{3}$$

- *If f is k-monotone, $\sum_{i=0}^{\infty} f(i) = \sum_{j=0}^{\infty}\pi_j$.*

In particular Q_j^k is k-monotone, and the set of k-knots of f is the set of integers j such that $\pi(j)$ is strictly positive.

These properties are shown in Lefevre and Loisel [25].

From this property, it appears that monotone discrete probabilities are mixture of uniform distributions, convex probabilities are mixture of triangular distribution, ..., k-monotone probabilities are mixture of splines with degree $k-1$.

3. Constrained least-squares estimation on the set of k-monotone discrete probabilities

Suppose that we observe n i.i.d random variables, $X_1, \dots, X_n$ with distribution p defined on $\mathbb{N}$, such that for all $i = 1, \dots, n$, and $j \in \mathbb{N}$, $p(j) = P(X_i = j)$. We propose to build an estimator of p that satisfies the k-monotonicity constraint. Since the projection on the set of k-monotone sequences $\mathcal{S}_k$ is not a probability in general for $k \geqslant 3$ (see Section 5) we consider the least-squares estimator $\widehat{p}$ defined as follows:

$$\widehat{p} = \operatorname{argmin}\left\{\|f-\widetilde{p}\|_2^2, f \in \mathcal{P}^k\right\} \tag{4}$$

where $\widetilde{p}$ is the empirical estimator of p:

$$\forall j \in \mathbb{N},\ \widetilde{p}(j) = \frac{1}{n}\sum_{i=1}^{n}\mathbb{1}_{\{X_i=j\}}.$$

Since the set $\mathcal{P}^k$ of k-monotone discrete probabilities is convex and closed in the Hilbert space $L^2(\mathbb{N})$, it follows, from the projection theorem for Hilbert spaces, that $\widehat{p}$ exists and is unique.

3.1. Characterizing the constrained least-squares estimator

A connerstone for deriving some statistical properties of our estimator is the following characterization of $\widehat{p}$. Let us begin with a few notation. For any positive sequence f in $L^1(\mathbb{N})$ we define the j-th primitive of f as follows: for all $l \in \mathbb{N}$

$$\begin{aligned} F_f^1(l) &= \sum_{i=0}^{l} f(i), \\ F_f^j(l) &= \sum_{i=0}^{l} F_f^{j-1}(i) \text{ for all } j \geqslant 2. \end{aligned}$$

Moreover, we define the quantity $\beta(f)$

$$\beta(f) = \sum_{i=0}^{\infty} f(i)(f(i) - \widetilde{p}(i)). \tag{5}$$

Theorem 1. *Let $f \in \mathcal{P}$. The projection $\widehat{p}$ defined at Equation (4) is the unique k-monotone probability f satisfying:*

1. *For all $l \in \mathbb{N}$,*

$$F_f^k(l) - F_{\widetilde{p}}^k(l) \geqslant \beta(f) m_l^k. \tag{6}$$

2. *If l is a k-knot of f, then the previous inequality is an equality.*

The proof of this theorem is given in Section 7.1.1. It uses the connections between successive primitives of the spline sequences $(Q_j^k, j \in \mathbb{N})$.

In the particular case of convexity the same result can be established with 0 in place of $\beta(f)$, see Lemma 2 in [13]. Let us recall that in that case, we have the nice property that the least-squares estimator over convex sequences is a convex probability distribution. This property is no longer satisfied when $k \geqslant 3$. We will come back to this point in Section 5.

3.2. Support of $\widehat{p}$

A key feature of the estimator $\widehat{p}$ is that its support is finite. Let us denote by $\widehat{s} = s_{\widehat{p}}$, respectively $\widetilde{s} = s_{\widetilde{p}}$, the maximum of the support of $\widehat{p}$, respectively $\widetilde{p}$.

Theorem 2. *Let $\widehat{p}$ be the least-squares estimator defined by Equation (4).*

1. *The support of $\widehat{p}$ is finite.*
2. *If $\widehat{s} \geqslant \widetilde{s} + 1$, then $\Delta^k \widehat{p}(i) = 0$*
   - *for all $i \in [\widetilde{s} - k + 2, \widehat{s} - 1]$ if k is even,*
   - *for all $i \in [\widetilde{s} - k + 2, \widehat{s} - 2]$ if k is odd.*

The proof of this theorem is given in Section 7.1.2. In the particular case of convexity, when $k = 2$, it is shown that $\widehat{s} \geqslant \widetilde{s}$. The question whether such a property still holds for $k \geqslant 3$ remains open.

3.3. Statistical Properties of $\widehat{p}$ when p is k-monotone

Let us now evaluate the behaviour of $\widehat{p}$ for estimating a probability, in particular, how does it compare with the empirical estimator $\widetilde{p}$. It is proved in the following theorem that the constrained least-squares estimator is closer (with respect to the L^2-norm) to any k-monotone probability than is $\widetilde{p}$.

Theorem 3. *For any k-monotone probability f, the following inequality is satisfied:*

$$\|f - \widehat{p}\|_2 \leqslant \|f - \widetilde{p}\|_2. \tag{7}$$

If $\widetilde{p}$ is not k-monotone, then the inequality is strict.

Moreover if p is k-monotone and if there exists $i \in \mathbb{N}$ such as $\Delta^k p(i) = 0$, then for all k-monotone probability f, we have:

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

In particular, if p is k-monotone and not strictly k-monotone, the estimator $\widehat{p}$ is strictly closer to p than is $\widetilde{p}$ with probability at least $1/2$. This theorem is a straightforward generalization of Theorem 4 in [13] for the convex case and its proof is omitted. Other statistical results have been shown in the convex case. In particular Balabdaoui and Durot [2] considered the difference between the cumulative distribution of the estimator under convex constraint, $F^1_{\widehat{p}}$, and F^1_f for any convex discrete probability f. They showed a Marshall lemma which states the following inequality: $\|F^1_{\widehat{p}} - F^1_f\|_\infty \leqslant 2\|F^1_{\widetilde{p}} - F^1_f\|_\infty$. Such a result could possibly be generalized for $k \geqslant 3$.

In the following theorem the moments of the distributions $\widehat{p}$ and $\widetilde{p}$ are compared. The proof of this theorem is given in Section 7.1.3.

Theorem 4. *For all $u \geqslant \max(1, k-3)$ and $0 \leqslant a \leqslant \widehat{s}$ the following inequality is satisfied:*

$$\sum_{i\geqslant 0} |i - a|^u \left(\widehat{p}(i) - \widetilde{p}(i)\right) \geqslant \beta\left(\widehat{p}\right) m(a, u), \tag{8}$$

where β is defined at Equation (5) and $m(a,u) = \sum_{i=0}^{a}(a-i)^u$.

Moreover $\widehat{p}(0) - \widetilde{p}(0) \geqslant \beta(\widehat{p})$.

If $\widehat{p}$ satisfies $\beta\left(\widehat{p}\right) = 0$, the result is the same as the one obtained in the convex case. In fact, it will be stated in Section 5, that $\beta(\widehat{p}) \leqslant 0$, and that $\beta(\widehat{p}) = 0$ if the mimimizer of $\|f - \widetilde{p}\|^2$ over $f \in \mathcal{S}^k$ equals the mimimizer of $\|f - \widetilde{p}\|^2$ over $f \in \mathcal{P}^k$. This is the case if $\widetilde{p}$ is k-monotone, or if $k = 2$.

3.4. Asymptotic properties of $\widehat{p}$

In this section we consider the asymptotic properties of $\widehat{p}$ when the sample size n tends to infinity. We first establish the consistency of $\widehat{p}$ both in the case of a well-specified model or a misspecified model.

Theorem 5. *Let $p_{\mathcal{S}^k}$ be the orthogonal projection of p on the set $\mathcal{P}^k$. Then, for all $r \in [2, +\infty]$, the random variable $\sqrt{n}\,\|p_{\mathcal{S}^k} - \widehat{p}\|_r$ is bounded in probability.*

In particular, this theorem states that if the distribution p is k-monotone, then the convergence of $\widehat{p}$ to p is of the order $\sqrt{n}$ with respect to the L^r-norm.

The case of a finite support In the particular case where the distribution p is k-monotone and has a finite support, we characterize the asymptotic behaviour of the k-knots of $\widehat{p}$, and give an upper bound for $\widehat{s}$, the maximum of the support of $\widehat{p}$.

Theorem 6. *Let p be a k-monotone probability with finite support.*

1. *Let $j \in \mathbb{N}$ be a k-knot of p. Then with probability one there exists $n_0 \in \mathbb{N}$ such that for all $n \geqslant n_0$, j is a k-knot of $\widehat{p}$.*
2. *Let s, respectively $\widehat{s}$, be the maximum of the support of p, respectively $\widehat{p}$. Then, with probability one, there exists $n_0 \in \mathbb{N}$ such that for all $n \geqslant n_0$ we have*
   - *$\widehat{s} \leqslant s$ if k is even.*
   - *$\widehat{s} \leqslant s + 1$ if k is odd.*

The proof of the first part of the theorem (see Section 7.1.5) is based on the fact that for all $j \in \mathbb{N}$,

$$\mathbb{P}\left(\lim_{n\to\infty} (-1)^k \Delta^k \widehat{p}(j) = (-1)^k \Delta^k p(j)\right) = 1.$$

It follows that if j is a k-knot of p, then $(-1)^k\Delta^k\widehat{p}(j)$ will be strictly positive for n large enough. Conversely, if j is not a k-knot of p, which means that $\Delta^k p(j) = 0$, then $(-1)^k\Delta^k\widehat{p}(j)$ may be strictly positive for all n. Therefore, the set of k-knots of $\widehat{p}$ does not estimate consistently the set of k-knots of p.

Concerning the second part of the theorem, we can notice that the result we get concerning $\widehat{s}$ is weaker than what was obtained in the convex case. Indeed, when $k = 2$, we know that $\widehat{s} \geqslant \widetilde{s}$, and consequently that, $\widehat{s} = s$ for n large enough if p has a finite support.

4. Implementing the estimator $\widehat{p}$

The practical implementation of $\widehat{p}$ requires the use of a specific algorithm that is composed of two parts. The first part consists in solving the problem defined at Equation (4) for sequences f whose support is finite. More precisely, for a chosen positive integer L, we compute $\widehat{p}_L$, the minimizer of $\|f - \widetilde{p}\|^2$ over probabilities $f \in \mathcal{P}^k$ whose support is included in $\{0, \ldots, L\}$. This part is similar to the first part of the algorithm described by Durot et al [13] except that an adaptation is needed to compute the minimizer over the set of probabilities (and not over the set of sequences). The second part consists in checking if $\widehat{p}_L = \widehat{p}$. For that purpose, starting from Theorem 1, we propose a stopping criterion that can be calculated in practice.

All functions mentionned in this article are implemented as an R package named **pkmon** and available on the Comprehensive R Archive Network. (https://CRAN.R-project.org/package=pkmon).

4.1. Constrained least-squares estimation on a given finite support

We know from Property 2 that if $f \in \mathcal{P}^k$, there exists a unique probability π on $\mathbb{N}$, such that f and π satisfy Equation (2). Therefore, solving (4) is equivalent to minimizing on the set of probabilities π on $\mathbb{N}$, the criterion $\Psi(\pi)$ defined as follows:

$$\Psi(\pi) = \sum_{i \geqslant 0} \left(\sum_{j \geqslant i} \pi(j) Q_j^k(i) - \widetilde{p}(i) \right)^2 .$$

The first part of our algorithm computes

$$\widehat{p}_L = \operatorname{argmin} \left\{ \|f - \widetilde{p}\|_2^2,\ f \in \mathcal{P}^k, s_f \leqslant L \right\}.$$

The solution is given by $\widehat{p}_L = \sum_{j \geqslant 0} \widehat{\pi}_L(j) Q_j^k$ where $\widehat{\pi}_L$ is the minimizer of $\Psi(\pi)$ over probabilities π whose support is included in $\{0, \ldots, L\}$:

$$\widehat{\pi}_L = \operatorname{argmin} \left\{ \Psi(\pi),\ \pi \in \mathcal{P}, s_\pi \leqslant L \right\}. \tag{9}$$

The algorithm we use to compute $\widehat{\pi}_L$ is based on the support reduction algorithm introduced by Groeneboom et al. [20]. An adaptation is needeed to guarantee that $\widehat{\pi}_L$ is a probability. Let us underline that this algorithm gives the exact solution in a finite number of steps.

For all $\nu \in \mathcal{P}$, let $D_\nu \Psi$ be the directionnal derivative function of Ψ in the direction ν defined as follows:

$$\forall \mu \in \mathcal{P}, D_\nu \Psi(\mu) = \lim_{\varepsilon \searrow 0^+} \frac{1}{\varepsilon} \big(\Psi((1-\varepsilon)\mu + \varepsilon\nu) - \Psi(\mu) \big).$$

The Support Reduction Algorithm is based on the property that $\widehat{\pi}_L$ is solution of (9) if and only if the directionnal derivative functions calculated in $\widehat{\pi}_L$ in the directions $\nu = \delta_j$, where δ_j denotes the Dirac probability in $\{j\}$, are non negative for all $0 \leqslant j \leqslant L$. Moreover, these derivatives are exactly 0 for all j in the support of $\widehat{\pi}_L$.

Starting from this property, the Support Reduction Algorithm is composed of two steps. In the first step, the support of the current probability μ is augmented by a point j where $D_{\delta_j} \Psi(\mu)$ is strictly negative (if any). In the second step the minimisation of $\Psi(\mu)$ over sequences μ such that $\sum_{j \geqslant 0} |\mu(j)| = 1$ and whose support is the current support, is performed. The current support is reduced to obtain a positive sequence. This second step performs minimization over the set of probabilities and differs from [13] and [20] which minimizes over all sequences. This step requires the introduction of KKT's conditions.

Let us introduce notation used in the second step of the algorithm. For a set $S = \{j_1, \ldots, j_s\} \subset \{0, \ldots, L\}$ we note

TABLE 1
Algorithm for computing $\widehat{\pi}_L$ for a fixed L.

- **Initialisation:**
  S $\leftarrow \{L\}$
  $\pi \leftarrow \delta_L$
- **Step 1:**
  For all $j \in \{0, \dots, L\}$ compute $D_{\delta_j}\Psi(\pi)$.
  - If for all $j \in \{0, \dots, L\}$ $D_{\delta_j}\Psi(\pi) \geqslant 0$, stop.
  - Else choose $j \in \{0, \dots, L\}$ such as $D_{\delta_j}\Psi(\pi) < 0$.
    $S' \leftarrow S + \{j\}$.
    Go to step 2.
- **Step 2:**
  $\lambda \leftarrow \lambda_{S'}$
  $\pi_{S'} \leftarrow \text{argmin}\{\Psi(\mu) + \lambda(\sum_{j\in S'} \mu(j) - 1), \text{supp}(\mu) \subset S'\}$.
  - If for all $l \in S', \pi_{S'}(l) \geqslant 0$,
    $\pi \leftarrow \pi_{S'}$
    $S \leftarrow S'$
    Return to step 1.
  - Else
    $l \leftarrow \text{argmin}\{\varepsilon_{j'} = \dfrac{\pi_{j'}}{\pi_{j'} - \pi_{S'}(j')},\ j' \in S', \pi_{S'}(j') < \pi_{j'}\}$
    $S' \leftarrow S' - \{l\}$
    Return to step 2.

- Q_S the matrix whose component $(Q_S)_{i+1,\ell} = Q^k_{j_\ell}(i)$ for $0 \leqslant i \leqslant L$ and $j_\ell \in S$, $l = 1, \dots, s$,
- H_S the projection matrix $H_S = Q_S(Q_S^T Q_S)^{-1} Q_S^T$,
- λ_S the Lagrange multiplier

$$\lambda_S = \frac{\langle H_S \widetilde{p}, \mathbb{I}\rangle - 1}{\langle H_S \mathbb{I}, \mathbb{I}\rangle},$$

where $\mathbb{I}$ is the vector with $L+1$ components all equal to 1.

The parameter λ_S comes from the KKT's condition linked to the following problem of minimization:

$$\pi_S = \underset{\substack{\sum_{j\in S}\pi(j)=1 \\ \pi \in \mathcal{M}_S}}{\text{argmin}} \big(\Psi(\pi)\big).$$

Its value is calulated in Section 7.2.2.

The algorithm for computing $\widehat{\pi}_L$ for a fixed L is given at Table 1. It is shown in Section 7.2 that this algorithm returns $\widehat{\pi}_L$ in a finite number of steps.

4.2. Stopping criterion

The second step of the algorithm is to find a stopping criterion, that is to say a characterization of $\widehat{p}$ allowing to decide for which L we have $\widehat{p}_L = \widehat{p}$. We will

developp two stopping criteria in this section. The first one is available for all $k \geqslant 2$ but offers no guarantee for the complexity. The second one is available only for $k \in \{3, 4\}$. This is the analoguous of the stopping criterion for $k = 2$ given in [13].

The characterization of $\widehat{p}$ given by Theorem 1 cannot help for practical implementation because the necessary condition in that theorem requires an infinite number of calculations. Nevertheless, if f is a k-monotone probability, with maximum support s_f, it is possible to find an integer M such that if f satisfies Inequality (6) for all $l \leqslant M$, then f satisfies Inequality (6) for all $l > M$. Such a property results from the writting of

$$P_f(l) = F_f^k(l) - F_{\widetilde{p}}^k(l) - \beta(f) m_l^k$$

as a polynomial function in the variable l. On the one hand, Property 4, shown in Section 7.3.2, states that $F_f^k(l) - F_{\widetilde{p}}^k(l)$ is a polynomial function in l of degree $k-1$ as soon as l is greater than the maxima of the support of f and $\widetilde{p}$.

Property 4. *Let f be a discrete sequence with finite support and s_f be the maximum of its support. Let $\tau = \max(s_f, \widetilde{s})$, then for all $l \geqslant \tau + 1$, we have the following equalities:*

$$\begin{aligned} F_f^k(l) - F_{\widetilde{p}}^k(l) &= \sum_{j=1}^{k} \bar{Q}_{l-1}^{k-j+1}(\tau)\big(F_f^j(\tau) - F_{\widetilde{p}}^j(\tau)\big) \\ &= \sum_{j=1}^{k} \frac{F_f^j(\tau) - F_{\widetilde{p}}^j(\tau)}{(k-j)!}\left((l-\tau+k-j-1)\dots(l-\tau)\right) \end{aligned} \tag{10}$$

On the other hand starting from Pascal's rule, it is easy to see that for all $k \geqslant 2$ and $l \geqslant 0$,

$$m_l^k = \bar{Q}_l^{k+1}(0) = \frac{(l+k)(l+k-1)\dots(l+1)}{k!}. \tag{11}$$

Putting Equations (10) and (11) together, it appears that for all $l \geqslant \tau = \max(s_f; \widetilde{s})$, there exist coefficients $(a_0, a_1, \dots, a_{k-1})$ such that

$$P_f(l) = \sum_{j=0}^{k-1} a_j l^j.$$

Let d be the degree of this polynomial (the smallest j such that $a_j = 0$ for all $j \geqslant d+1$) and let M be defined by

$$M = \max\left(1 + \frac{a_{d-1}}{a_d}, \dots, 1 + \frac{a_0}{a_d}\right),$$

By Cauchy's Theorem for localization of polynomial's roots, the largest root of $P_f(l)$ is bounded by M. Therefore if a_d is positive, $P_f(l)$ is positive beyond M. This leads to the following characterization of $\widehat{p}$ which is a corollary of Theorem 1. Its proof is omitted.

Theorem 7. *Let f be a sequence in $\mathcal{P}$ with a finite support. Let M and a_d be defined as above. The two following assertions are equivalent:*

1. *The sequence f satisfies*
   (a) *a_d is positive.*
   (b) *$\forall l \leqslant M, F_f^k(l) - F_{\widehat{p}}^k(l) \geqslant \beta(f) m_l^k$.*
   (c) *If l is a k-knot of f, the previous inequality is an equality.*
   (d) *$\beta(f) \leqslant 0$.*
2. *The sequence f is exactly $\widehat{p}$.*

This Theorem answers our initial problem, namely checking if $\widehat{p}_L$ equals $\widehat{p}$: the coefficients a_j depending on $\widehat{p}_L$ and $\widetilde{p}$, can be calculated in practice, as well as M. Nevertheless M can be very large, leading to tedious computation. For small values of k more efficient criteria can be proposed. In particular, for $k = 3$ and $k = 4$, it is possible to obtain a characterization of $\widehat{p}$ that only depends on $\widetilde{s}$ and $\widehat{s}$. This is the object of the following theorem shown in Section 7.1.6.

Theorem 8. *Let f be a sequence in $\mathcal{P}$ with a finite support and $s' = \max\{s_f, \widetilde{s}\}$. Let $k \in \{3, 4\}$. The two following assertions are equivalent:*

1. *The sequence f satisfies*
   (a) *$\forall l \leqslant s' + 1, F_f^k(l) - F_{\widetilde{p}}^k(l) \geqslant \beta(f) m_l^k$.*
   (b) *If l is a k-knot of f, the previous inequality is an equality.*
   (c) *for all $2 \leqslant j \leqslant k-1$, $F_f^j(s'+1) - F_{\widetilde{p}}^j(s'+1) \geqslant \beta(f) m_{s'+1}^j$.*
   (d) *$\beta(f) \leqslant 0$.*
2. *The sequence f is exactly $\widehat{p}$.*

The algorithm we propose here can also be applied when $k = 2$, as an alternative to the algorithm proposed by [13]. For each L, $\widehat{p}_L$ is calculated under the constraint that $\widehat{p}_L$ sums to one. Necessary and sufficient condition to have $\widehat{p} = \widehat{p}_L$ are reduced to conditions *(a)* and *(b)* with $\beta(f) = 0$.

When $k > 4$ we are not able to propose a similar stopping criterion. Indeed the proof is based on the properties of the spline function Q_j^k for $k > 4$ and in particular, requires the calculation of the number of k'-knots of Q_j^k for $k' \geqslant k$, which is intractable.

5. Constrained least-squares estimation on the set of k-monotone sequences

By definition, our estimator $\widehat{p}$ is a probability. We could have proposed to estimate p by minimizing the least-squares criterion under the constraint of k-monotonicity only. Let $\widehat{p}^*$ be that estimator:

$$\widehat{p}^* = \operatorname{argmin}\left\{\|f - \widetilde{p}\|_2^2, f \in \mathcal{S}^k\right\} \tag{12}$$

The following property, shown in Section 7.3.3, establishes the link between both estimators.

Property 5. *Let $\widehat{p}$ and $\widehat{p}^*$ be defined at Equations (4) and (12), and let β be defined at Equation (5). The coefficient $\beta(\widehat{p})$ is null if and only if $\widehat{p} = \widehat{p}^*$.*

In the particular case where $k = 2$, $\widehat{p}^*$ is exactly $\widehat{p}$ (see Theorem 1 in [13]). As soon as $k \geqslant 3$ this property is no longer satisfied in general and the following property is proven in Section 7.3.4:

Property 6. *The mass of $\widehat{p}^*$ is greater than or equal to 1.*

This result was expected because a similar property was shown by Balabdaoui [1] in the continuous framework.

To illustrate this point in the discrete framework, let us consider the projection of δ_1 (the Dirac probability in 1) on the set of 3-monotone sequences. Some calculation (see Section 7.3.5 for a proof) leads to the following result:

$$\mathrm{Proj}_{\mathcal{S}^3}(\delta_1) = \frac{3}{238}\bar{Q}_5^3 + \frac{1}{238}\bar{Q}_6^3,$$

and its mass is close to 1.06.

Nevertheless we can show (see Section 7.3.6) the following asymptotic result.

Property 7. *Let p be a k-monotone probability, and $\widehat{p}^*$ be defined at Equation (12). Then, with probability one the mass of $\widehat{p}^*$ converges to one.*

The properties shown in Section 3 for the estimator $\widehat{p}$ hold true for $\widehat{p}^*$. More precisely, the estimator $\widehat{p}^*$ satisfies Theorems 2, 5, 6. Theorem 3 is also true for $\widehat{p}^*$ apart from the first assertion, where Equation (7) is satisfied for any k-monotone sequence f. Finally Theorem 4 is true with 0 in place of $\beta(\widehat{p})(\widehat{s}+1-a)$ in Equation (8).

The implementation of $\widehat{p}^*$ is similar to that of $\widehat{p}$ except that in the first part of the algorithm (see Section 4.1) the Support Reduction Algorithm can be used without any modification at Step 2 (where the estimator of π is constraint to have a sum of one). The stopping criterion used in the second part is obtained in the same way as for $\widehat{p}$. The proofs of these last results are omitted. They are based on Property 5.

6. Simulation

We designed a simulation study to assess the quality of the least-squares estimator $\widehat{p}$ on the set of k-monotone probabilities, as compared to the empirical estimator $\widetilde{p}$, and to the least-squares estimator $\widehat{p}^*$ on the set of k-monotone sequences for $k \in \{2, 3, 4\}$. We considered both the case where the true distribution is k-monotone and the case where it is not.

6.1. Simulation design

We considered mainly two shapes for the distribution p: the spline distribution Q_j^ℓ with $j = 10$ and $\ell \in \{2, 3, 4, 10\}$, and the Poisson distribution $\mathcal{P}(\lambda)$ for

$\lambda \in \{0.3, 0.35, 0.45, 2-\sqrt{2}, 0.7, 1\}$. Those two families of distribution differ by the finiteness of their support, and by the number of knots in their decomposition on the spline basis. Precisely, the distribution Q_j^ℓ has one ℓ-knot in j while a ℓ-monotone Poisson distribution has an infinite number of ℓ-knots. The following proposition, shown in Section 7.3.7, gives the property of k-monotonicity for Poisson distributions.

Property 8. *Let $\mathcal{P}(\lambda)$ be the Poisson distribution with parameter λ. For each $\ell \geqslant 1$, let λ_ℓ be defined as the smallest root of the following polynomial function:*

$$P_\ell(\lambda) = \sum_{h=0}^{\ell} (-1)^h \frac{(\ell!)^2}{h!((\ell-h)!)^2} \lambda^h.$$

Then $\mathcal{P}(\lambda)$ is ℓ-monotone if and only if $\lambda \leqslant \lambda_\ell$

Some simple calculation gives the following values: $\lambda_1 = 1$, $\lambda_2 = 2-\sqrt{2} \simeq 0.585$, $\lambda_3 \simeq 0.415$, $\lambda_4 \simeq 0.322$, $\lambda_5 \simeq 0.264$. Therefore the considered Poisson distributions $\mathcal{P}(\lambda)$ are $\{4, 3, 2, 2, 1\}$-monotone when λ belongs to $\{0.3, 0.35, 0.45, 2-\sqrt{2}, 0.7\}$. When $\lambda = 1$, the Poisson distribution is not strictly decreasing.

Moreover, for $k = 3$, we consider an other shape for the distribution p when p is not k-monotone. More precisely we study the behaviour of the l_2-loss of $\widehat{p}^k$ when $(-1)^k \Delta^k p(i) < 0$ for only one $i \in \mathbb{N}$, as $R_\alpha = -\alpha Q_2^3 + (1+\alpha) Q_{10}^3$ with $\alpha \in \{0.2, 0.1, 0.05, 0.01\}$.

For each distribution p, we considered several values for the sample size n: $n \in \{20, 50, 100, 250, 500, 1000\}$. In some cases we also considered very large values of n in order to illustrate the asymptotic framework. We denote by $\widetilde{p}_n$ the empirical estimator and by $\widehat{p}_n^k$, respectively $\widehat{p}_n^{*k}$, the least-squares estimator of p on the set of k-monotone probabilities, respectively sequences. For each simulation configuration, 1000 random samples were generated.

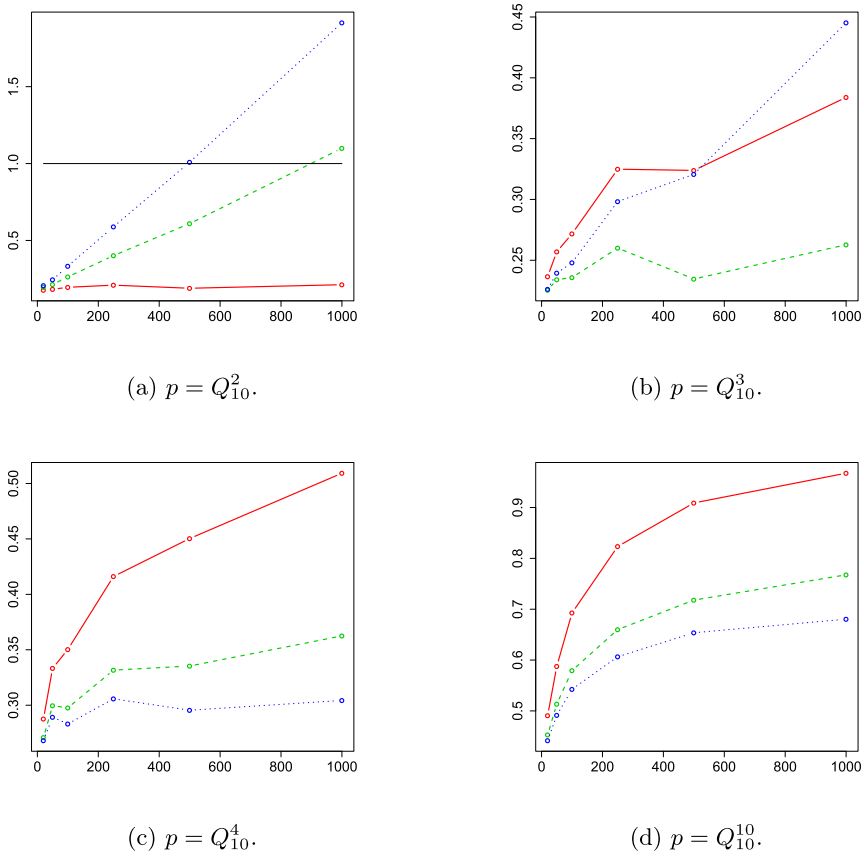
6.2. Global fit

To assess the quality of the estimators for estimating the distribution p we consider the l_2-loss and the Hellinger loss. We have also considered the total variation loss, but the results are not shown because they are very similar to those obtained for the l_2-loss.

6.2.1. Estimators comparison based on the l_2-loss

The l_2-loss between p and any estimator of p, say $\widehat{q}$, is defined as the expectation of the l_2-error, $l_2(p, \widehat{q}) = E(\|p - \widehat{q}\|_2^2)$.

Spline distributions We first compared the quality of the fit of the estimators $\widehat{p}_n^k$ and $\widetilde{p}_n$ by computing for each simulated sample $\|p - \widehat{p}_n^k\|_2^2$ and $\|p - \widetilde{p}_n\|_2^2$. The l_2-losses were estimated by the mean of 1000 independant replications of the l_2-errors. In all simulation configurations, the l_2-losses are decreasing towards 0 when n increases. In what follows we will consider the ratios $l_2(p, \widehat{p}_n^k)/l_2(p, \widetilde{p}_n)$ to compare the estimators.



(a) $p = Q_{10}^2$. (b) $p = Q_{10}^3$.

(c) $p = Q_{10}^4$. (d) $p = Q_{10}^{10}$.

FIG 1. *Spline distributions: ratio between the l_2-loss of $\widehat{p}_n^k$ and the l_2-loss of $\widetilde{p}_n$ versus the sample size n: for $k = 2$ in "—", $k = 3$ in "- -", $k = 4$ in "...". Each subfigure corresponds to the results obtained with $p = Q_j^\ell$, for $\ell \in \{2, 3, 4, 10\}$.*

The results for the spline distributions Q_j^ℓ are presented on Figure 1. When n is small, $\widehat{p}_n^k$ has smaller l_2-loss than $\widetilde{p}_n$ whatever the value of k. When n tends to infinity, we have to consider two cases according to the discrepancy between k which defines the degree of monotonocity of the estimator, and ℓ which is the degree of monotonicity of p. As it was expected considering Theorem 3, when $k \leqslant \ell$, the ratio is smaller than 1.

Moreover we note that the smaller the deviation $\ell - k$ is, the smaller the ratio. In particular when $k = \ell$, the ratio tends to a constant strictly smaller than 1, while when $k < \ell$, the ratio tends to 1. For example, when $\ell = 4$, $k = 3$, the ratio of the l_2-losses equals 0.45 for $n = 10000$ and 0.80 for $n = 100000$. This illustrates the benefit to choosing the *correct* k, i.e. $k = \ell$, instead of $k < \ell$, and matches our intuition. Indeed, the sets $\mathcal{S}^k$ being nested, we are ledding to think that one could gain in l_2-loss when we project the empirical estimator on the set $\mathcal{S}^\ell$ instead on the set $\mathcal{S}^k$ with $k < \ell$.

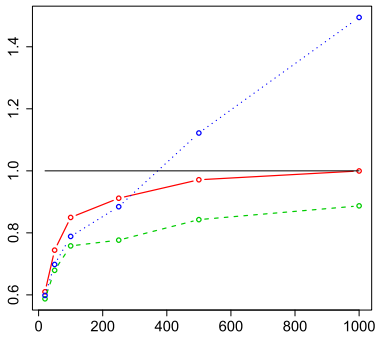


FIG 2. *Poisson distribution with parameter $\lambda = 0.35$: ratio between the l_2-loss of $\widehat{p}_n^k$ and the l_2-loss of $\widetilde{p}_n$ versus the sample size n: for $k = 2$ in "—", $k = 3$ in "- -", and $k = 4$ in "...".*

When $k > \ell$, the ratio tends to infinity. For example, when $\ell = 2$, $k = 3$, the ratio of the l_2-losses equals 9.93 for $n = 10000$ and 259 for $n = 100000$. This result was expected because the empirical estimator $\widetilde{p}_n$ is consistent while our estimator is not. Indeed, following Theorem 5, $\widehat{p}_n^k$ converges in probability to $p_{\mathcal{S}^k}$, the projection of p on $\mathcal{S}^k$. Since p is ℓ-monotone and $k > \ell$, the l_2-loss $l_2(p, \widehat{p}_n^k)$ is greater than a strictly positive constant whereas $l_2(p, \widetilde{p}_n)$ converges to 0. Therefore the ratio of the l_2-losses converges to infinity.

Poisson distribution The results for the Poisson distribution are similar to those obtained for the spline distributions except that the asymptotic is achieved for smaller values of the sample size n. Only the case $\lambda = 0.35$, where the corresponding Poisson distribution is 3-monotone, is presented in Figure 2. It appears that when $k = 2$ the ratio of l_2-losses tends to one, when $k = 3$ it tends to a value close to 0.9, and when $k = 4$ it tends to infinity.

Comparaison between $\widehat{p}$ and $\widehat{p}^*$ Now we compare the l_2-losses for the estimators $\widehat{p}_n^k$, $\widehat{p}_n^{*k}$ and $\widetilde{p}_n$ for $k = 3$ and $k = 4$ (recall that for $k = 2$, $\widehat{p}_n^{*k} = \widehat{p}_n^k$). The ratios $l_2(p, \widehat{p}_n^{*k})/l_2(p, \widetilde{p}_n)$ behave similarly to the ratios $l_2(p, \widehat{p}_n^k)/l_2(p, \widetilde{p}_n)$ (not shown).

Next we compare the values of the l_2 losses for $\widehat{p}_n^{*k}$ and $\widehat{p}_n^k$. When we consider the spline distributions Q_j^ℓ with $l = 2$ and $l = 3$, the difference between the l_2 losses are not significant (they are smaller than 2-times their empirical standard-error calculated on the basis of 1000 simulations). When l increases, the distribution p is more hollow and it appears that $l_2(p, \widehat{p}_n^{*k})$ is greater than $l_2(p, \widehat{p}_n^k)$, see Table 2.

Distributions R_α In this paragraph we take $k = 3$. When looking into model misspecification we face the problem of the deviation from k-monotonicity. A natural issue is to consider the case where the true distribution p is k-monotone except for only one $i \in \mathbb{N}$ where $(-1)^k \Delta^k p(i) < 0$. We assess the l_2-loss of the estimator $\widehat{p}$ in case of such a misspecification, when $k = 3$ and $p = R_\alpha =$

TABLE 2
Spline distributions: difference (×1000) between the l_2-loss of $\widehat{p}_n^{*k}$ and the l_2-loss of $\widehat{p}_n^k$, for different values of n, for $k = 3$ in green and $k = 4$ in blue. The symbol "-" is for non-significant result.

	$n = 20$	$n = 100$	$n = 1000$
$p = Q_{10}^4$	-	-	-
	-	-	0.06
$p = Q_{10}^{10}$	0.89	0.13	0.02
	0.92	0.24	0.01

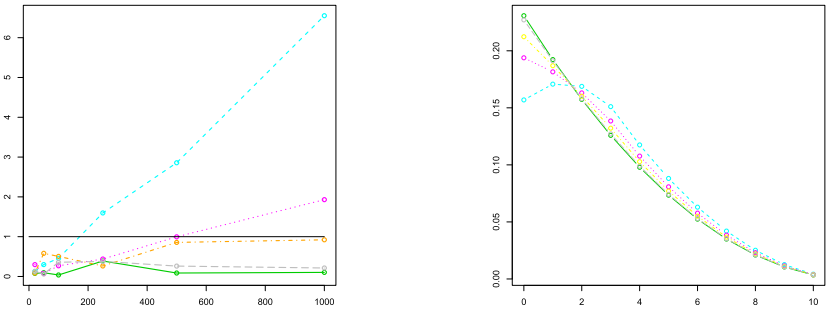


FIG 3. *Mixture of splines with shape $R_\alpha = \alpha\bar{Q}_2^3 + (1+\alpha)\bar{Q}_{10}^3$: In the right: distributions of the R_α and in the left: ratio between the l_2-loss of $\widehat{p}_n^k$ and the l_2 loss of $\widetilde{p}_n$ versus the sample size n: $\alpha = 0$ in "—", $\alpha = 0.01$ in "- -", $\alpha = 0.05$ in "- -", $\alpha = 0.1$ in "..." and $\alpha = 0.2$ in "- -".*

$\alpha\bar{Q}_2^3 + (1+\alpha)\bar{Q}_{10}^3$ is 3-monotone except for $i = 2$. Figure 3 illustrate the results for the different values of α.

When n is small, $\widehat{p}_n^k$ has smaller loss than $\widetilde{p}_n$, although the distributions R_{alpha} are not k-monotone when $\alpha > 0$. When $\alpha \geqslant 0.05$ and n tends to infinity the ratio between the l_2-loss of $\widehat{p}_n^k$ and the l_2-loss of $\widetilde{p}_n$ increases (except for $\alpha = 0$, when the model is not misspecified) whereas when α is small enough ($\alpha \leqslant 0.01$) the ratio stay small. The bigger α is, the more the distribution deviates from de 3-monotonicity and the more the l_2-loss is.

6.2.2. Estimators comparison based on the Hellinger loss

Let us now consider the Hellinger loss defined, for any estimator $\widehat{q}$, as $H(p, \widehat{q}) = E\left(\|\sqrt{p} - \sqrt{\widehat{q}}\|_2^2\right)$.

Spline distributions The results for the spline distributions Q_j^ℓ are similar to those obtained for the l_2-loss, except that the ratios $H(p, \widehat{p}_n^k)/H(p, \widetilde{p}_n)$ are not necessary smaller than 1 when $k \leqslant \ell$, see Figure 4 for the Triangular distribution Q_j^2.

Poisson distribution In the case of the Poisson distributions the differences between the l_2-loss and the Hellinger loss are more obvious. As it is illustrated by Figure 5, if ℓ the degree of monotonicity of p is strictly greater than k, then

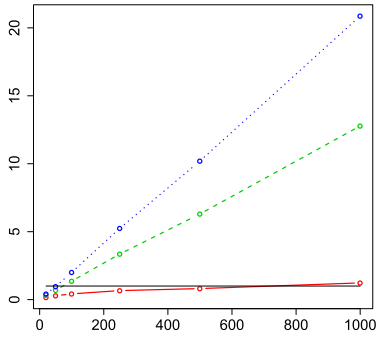
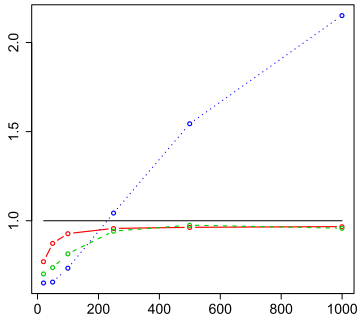
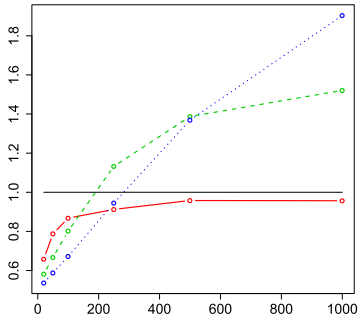


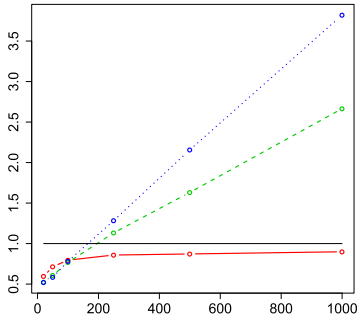
FIG 4. *Triangular distribution Q_j^2: ratio between the Hellinger loss of $\widehat{p}_n^k$ and the Hellinger loss of $\widetilde{p}_n$ versus the sample size n: for $k = 2$ in "—", $k = 3$ in "- -", and $k = 4$ in "...".*



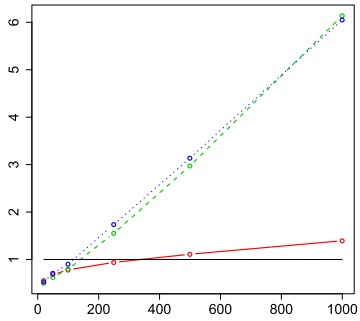
(a) $\lambda = 0.3; \ell = 4$.



(b) $\lambda = 0.35; \ell = 3$.



(c) $\lambda = 0.45; \ell = 2$.



(d) $\lambda = 2 - \sqrt{2}; \ell = 2$.

FIG 5. *Poisson distributions: ratio between the Hellinger loss of $\widehat{p}_n^k$ and the Hellinger loss of $\widetilde{p}_n$ versus the sample size n: for $k = 2$ in "—", $k = 3$ in "- -", and $k = 4$ in "...". Each subfigure corresponds to the results obtained with $p = \mathcal{P}(\lambda)$ for $\lambda \in \left\{0.3, 0.35, 0.45, 2 - \sqrt{2}\right\}$. The degree of monotonicity of these distributions is given by ℓ.*

the ratio is smaller than 1 (see case (a) with $k = 2, 3$ and case (b) with $k = 2$). If $k = \ell$, then $H(p, \widetilde{p}_n^k)$ is smaller than $H(p, \widehat{p}_n)$ if the distribution p is "ℓ-monotone enough", that is to say if the parameter λ of the Poisson distribution is such that $\lambda_\ell - \lambda$ is large enough, where λ_ℓ has been defined in Property 8, see for example cases (c) and (d) with $k = 2$, where $\lambda_2 = 2 - \sqrt{2}$.

6.3. Some characteristics of interest

We consider the estimation of some charactéeristics that may be of interest as the entropy, the variance and the probability at 0. For each of these characteristics denoted $L(p)$, we measure the performance in terms of the root mean squared error of prediction calculated as follows:

$$\text{RMSEP} = \sqrt{\text{BIAS}^2 + \text{SE}^2},$$

where BIAS and SE are the estimated bias and standard-error of the estimator based on the simulations. Let $\widehat{L}$ be an estimator of $L(p)$, then $\text{BIAS} = \widehat{L}_\cdot - L$, where $\widehat{L}_\cdot = \sum_s \widehat{L}_s/1000$ with $\widehat{L}_s$ being the estimate of $L(p)$ at simulation s, and $\text{SE}^2 = \sum_s (\widehat{L}_s - \widehat{L}_\cdot)^2/1000$.

6.3.1. Entropy

The entropy is defined as

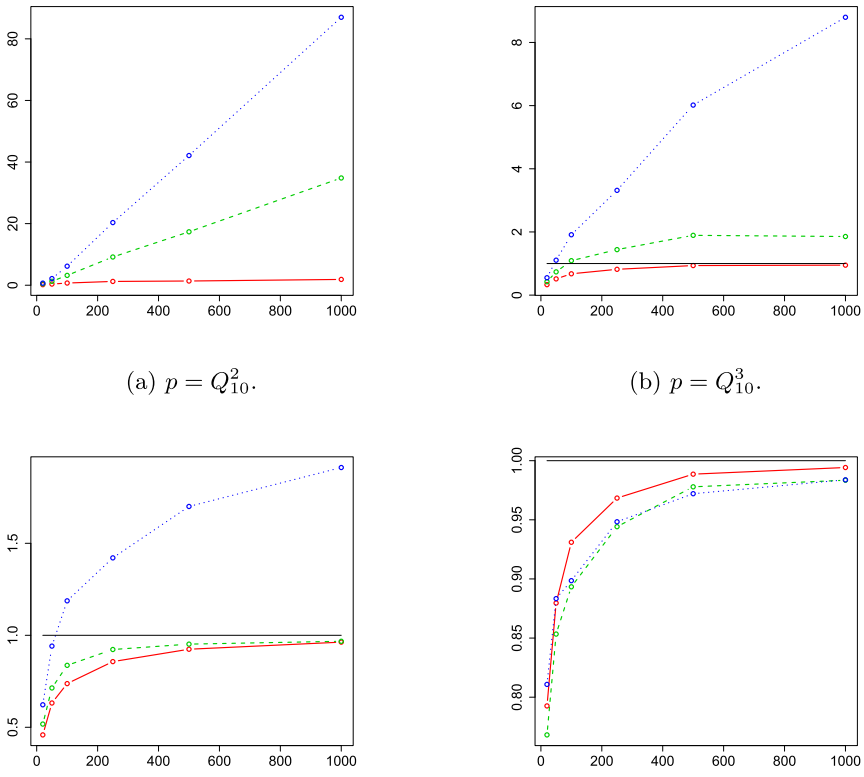
$$\text{Ent}(f) = \sum_{i=0}^{\infty} f(i) \log(f(i)).$$

We compare the estimators $\text{Ent}(\widehat{p}_n^k)$ and $\text{Ent}(\widetilde{p}_n)$ by the ratio of their RMSEP. The results differ according to the family of distributions. For the spline distributions Q_j^ℓ, see Figure 6, it appears that if $k < \ell$, then $\text{Ent}(\widehat{p}_n^k)$ has smaller RMSEP than $\text{Ent}(\widetilde{p}_n)$. However, when $k = l$, the ratio of the RMSEP's increases and reaches an asymptote greater than 1. For example, in Figure 6, case (b) with $k = 3$, the ratio tends to 0.96, in case (c) with $k = 4$, the ratio tends to 1.93. In fact, if we consider the space of ℓ-monotone distributions with maximum support j, the distribution Q_j^ℓ may appear as a "limiting case" in this space, in that it admits only one ℓ-knot in j. It seems that for these Q_j^ℓ distributions, the projection on the space of $\ell - 1$-monotone discrete probabilities give better results than on the space of ℓ-monotone discrete probabilities.

For the Poisson distributions, see Figure 7, when n is small, the estimator based on the emprirical distribution, $\text{Ent}(\widetilde{p}_n)$, has a smaller RMSEP than $\text{Ent}(\widehat{p}_n^k)$. When n is large the RMSEP ratio tend to one if $k \leqslant \ell$, and tend to infinity if $k > \ell$.

6.3.2. Probability mass in 0

We compare the performances of $\widehat{p}_n^k(0)$ and $\widetilde{p}_n(0)$ by comparing the corresponding renormalized SE and BIAS.



(a) $p = Q_{10}^2$. (b) $p = Q_{10}^3$.

(c) $p = Q_{10}^4$. (d) $p = Q_{10}^{10}$.

FIG 6. *Spline distributions: ratio between the* RMSEP *of Ent*$(\widehat{p}_n^k)$ *and the* RMSEP *of Ent*$(\widetilde{p}_n)$ *versus the sample size n: for* $k = 2$ *in "—",* $k = 3$ *in "- -",* $k = 4$ *in "...". Each subfigure corresponds to the results obtained with* $p = Q_j^\ell$*, for* $\ell \in \{2, 3, 4, 10\}$.

The results for the spline distributions are presented in Table 3.

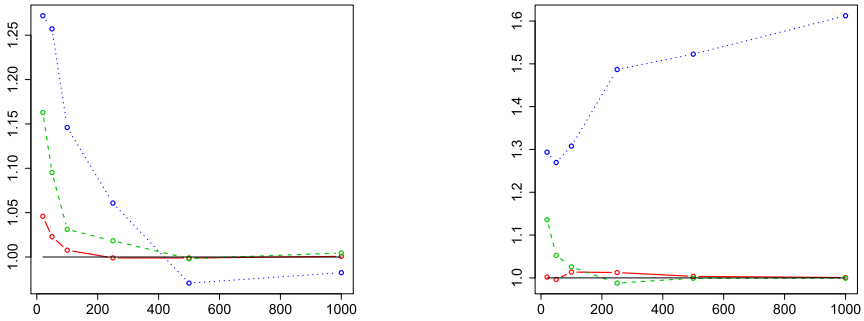
When $k \leqslant l$, $\widehat{p}_n^k(0)$ has smaller SE than $\widetilde{p}_n(0)$. Its bias is greater in absolute value and always negative, but the RMSEP stays smaller. For each k, the variations of $\sqrt{n}SE/p(0)$ versus n are very small and tend to stabilize around a value that increases with $l - k$.

When $k > l$, $\widehat{p}_n^k(0)$ keeps a smaller RMSEP than $\widetilde{p}_n(0)$ for small n. But, when n increases the absolute bias as well as the standard error increase.

The results for the Poisson distributions are similar and omitted.

6.3.3. Variance

We compare the estimators of the variance of p, denoted $\text{var}(\widehat{p}_n^k)$ and $\text{var}(\widetilde{p}_n)$ comparing the ratio of their RMSEP. The results are similar for the spline



(a) $\lambda = 0.3; \ell = 4$. (b) $\lambda = 0.35; \ell = 3$.

FIG 7. *Poisson distributions: ratio between the* RMSEP *of Ent*$(\widehat{p}_n^k)$ *and the* RMSEP *of Ent*$(\widetilde{p}_n)$ *versus the sample size n: for* $k = 2$ *in "—", k* $= 3$ *in "- -", k* $= 4$ *in "...". Each subfigure corresponds to the results obtained with* $p = \mathcal{P}(\lambda)$ *with* $\lambda \in \{0.3, 0.35\}$.

TABLE 3
Spline distributions: $\sqrt{n}$SE$/p(0)$, $\sqrt{n}|$BIAS$|/p(0)$ *and* $\sqrt{n}$RMSEP$/p(0)$ *for* $\widetilde{p}_n(0)$ *in black,* $\widehat{p}_n^k(0)$ *for* $k = 2$ *in red,* $k = 3$ *in green and* $k = 4$ *in blue.*

	$n = 20$			$n = 100$			$n = 1000$		
	SE	BIAS	RMSEP	SE	BIAS	RMSEP	SE	BIAS	RMSEP
$p = Q_{10}^2$	2.25	7e-4	2.25	2.234	0.002	2.234	2.284	0.017	2.284
	1.800	0.181	1.809	1.819	0.170	1.82	1.745	0.162	1.752
	1.757	0.157	1.764	1.783	0.188	1.792	2.231	0.334	2.255
	1.742	0.155	1.748	1.780	0.196	1.790	2.622	0.408	2.653
$p = Q_{10}^4$	1.634	0.008	1.634	1.601	0.013	1.601	1.626	0.006	1.626
	1.362	0.143	1.369	1.389	0.120	1.394	1.488	0.052	1.489
	1.354	0.137	1.361	1.372	0.132	1.378	1.439	0.088	1.442
	1.340	0.135	1.347	1.353	0.136	1.359	1.362	0.109	1.366
$p = Q_{10}^{10}$	1.010	2e-4	1.010	0.98	6e-4	0.98	0.984	0.006	0.984
	0.884	0.058	0.886	0.934	0.022	0.934	0.982	0.006	0.982
	0.886	0.057	0.888	0.919	0.039	0.920	0.957	0.009	0.957
	0.887	0.053	0.889	0.921	0.042	0.922	0.940	0.018	0.940

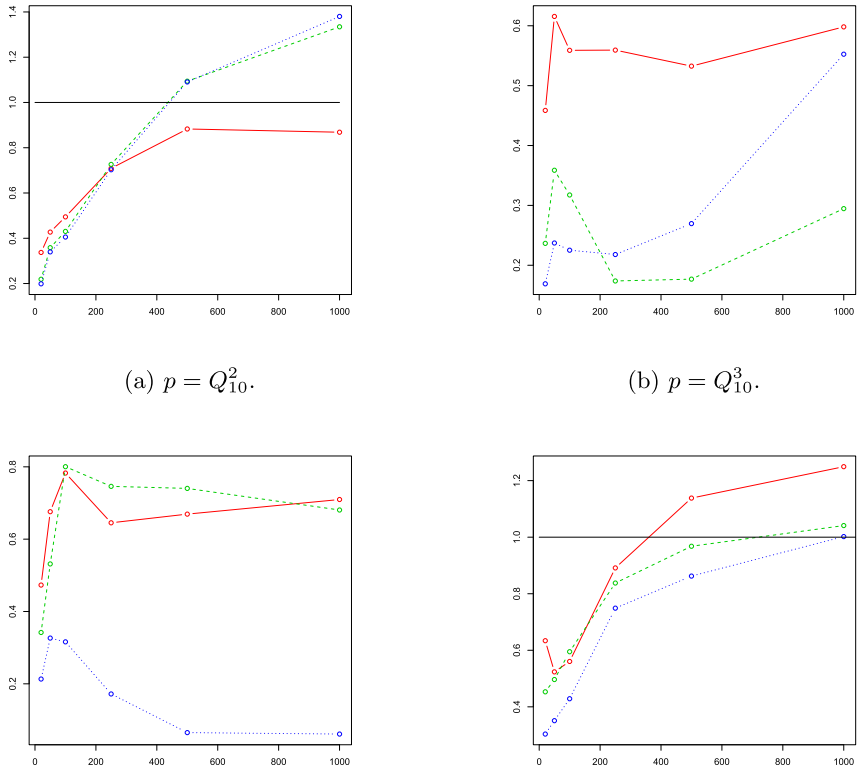
distributions and the Poisson's distributions and we present only the RMSEP for the spline distributions Q_j^l in Figure 8.

When $k = l$, the ratio of the RMSEP tends to a constant smaller than 1 when n tends to infinity. Conversely if we are not in a good model ($k > l$) the ratio of the RMSEPs tends to infinity when n tends to infinity.

When $k < l$ and n large the ratio of the RMSEPs increases with $l - k$ and goes beyond 1. For example for $k = 3$ and $l = 4$ the ratio of the RMSEPs is equal to 0.68 when $n = 10000$, while if $l = 10$ the ratio is greater than 1 as soon as $k \leqslant 3$ and $n \geqslant 1000$.

When $k > l$ the ratio of the RMSEPs tends to infinity when n tends to infinity.

When n is small var$(\widehat{p}_n^k)$ has smaller RMSEP than var$(\widetilde{p}_n)$ whatever the value of k and l.



(a) $p = Q_{10}^2$. (b) $p = Q_{10}^3$.

(c) $p = Q_{10}^4$. (d) $p = Q_{10}^{10}$.

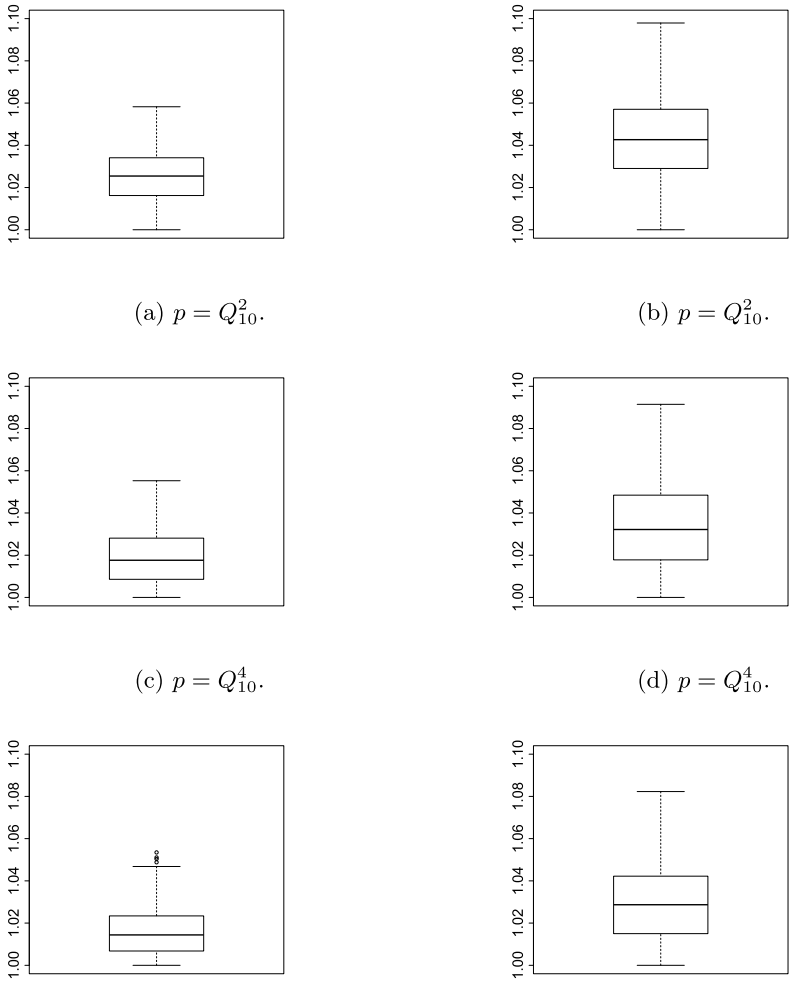
FIG 8. *Spline distributions: ratio between the* RMSEP *of var*$(\widehat{p}_n^k)$ *and the* RMSEP *of var*$(\widetilde{p}_n)$ *versus the sample size n: for* $k = 2$ *in "—", k* $= 3$ *in "- -", k* $= 4$ *in "...". Each subfigure corresponds to the results obtained with* $p = Q_j^\ell$*, for* $\ell \in \{2, 3, 4, 10\}$.

6.4. About the mass of the non-constrained estimator $\widehat{p}^{*k}$

We were also interested in the estimation of the mass of the non-constrained estimator $\widehat{p}^{*k}$. Figures 9 and 10 illustrate the results for the spline distributions with $n = 20$ and $n = 100$. As expected the mass is always larger than 1 and whatever k, the distribution of the mass comes closer to one when n increases (compare figures 9 and 10). The larger l is, the smaller the median and the dispersion around the median are. On the other hand when k increases the distributions are more scattered and their medians move away from 1 (compare the lines of each figure).

6.5. Conclusion

Let us consider the case where p is l-monotone and $\widehat{p}_n^k$ is the least-squares estimator of p on $\mathcal{S}_k$ for $k \leqslant l$. In this case the model is well-specified.



(a) $p = Q_{10}^2$. (b) $p = Q_{10}^2$.

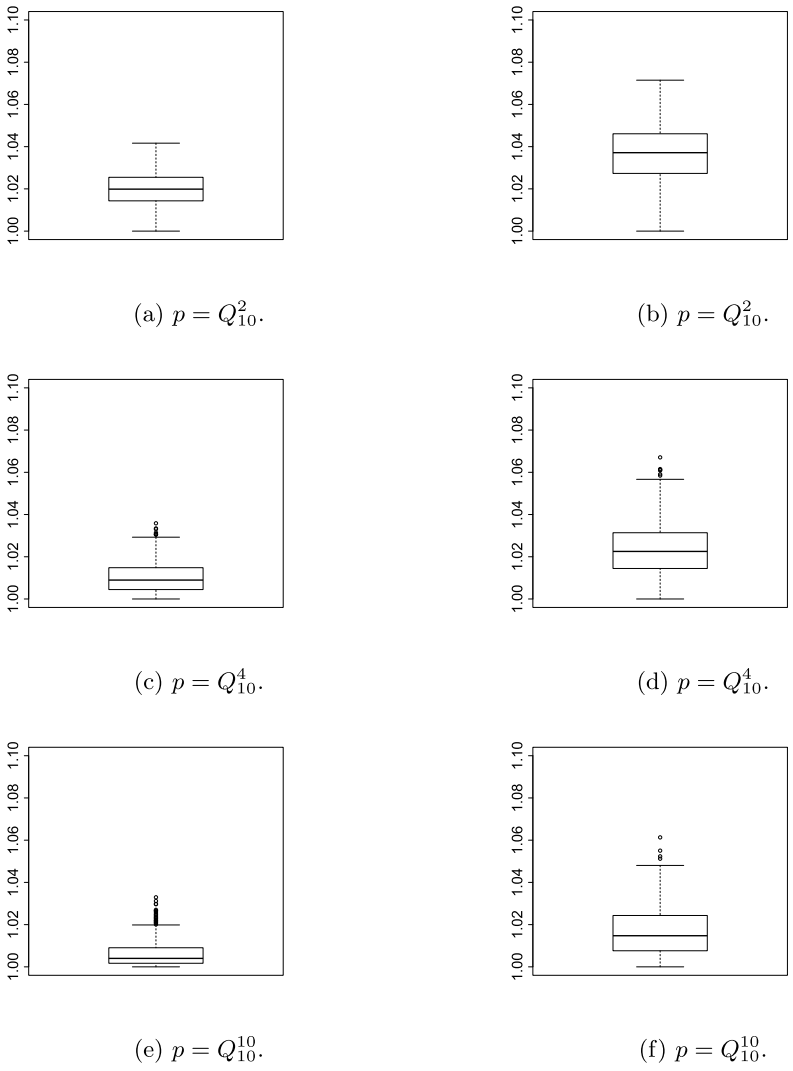
(c) $p = Q_{10}^4$. (d) $p = Q_{10}^4$.

(e) $p = Q_{10}^{10}$. (f) $p = Q_{10}^{10}$.

FIG 9. *Splines distributions; $n = 20$: Boxplot of the mass of $\widehat{p}^{*k}$ for $k = 3$ at the top and $k = 4$ at bottom. Each column corresponds to the results obtained with $p = Q_j^l$ for $l = 2, 4, 10$.*

Concerning the l_2-loss, the total variation loss and the estimation of $p^*(0)$, $\widehat{p}_n^k$ performs better than the empirical estimator $\widetilde{p}_n$. Moreover the superiority of the performance of $\widehat{p}_n^k$ is larger when n is small.

Concerning the Hellinger loss, or the estimation of the variance and the entropy, we get the following results. For small n, as before, the least-squares estimator is always better than the empirical estimator $\widetilde{p}_n$. When n is large, $\widehat{p}_n^k$ and $\widetilde{p}_n$ behave similarly. If p is a *frontier* distribution in $\mathcal{S}_l$, as for example the



(a) $p = Q_{10}^2$. (b) $p = Q_{10}^2$.

(c) $p = Q_{10}^4$. (d) $p = Q_{10}^4$.

(e) $p = Q_{10}^{10}$. (f) $p = Q_{10}^{10}$.

FIG 10. *Repartition of the mass of $\widehat{p}^{*k}$ for $n = 100$. Each column represents the estimation of a different probability p explained in subtitle. The first line is for the mass of $\widehat{p}^{*3}$ and the second line for the mass of $\widehat{p}^{*4}$.*

Poisson distribution with $\lambda = \lambda_l$ or a spline distribution Q_j^l, then $\widehat{p}_n^{l-1}$ performs better than $\widehat{p}_n^l$. If not, then $\widehat{p}_n^l$ performs better than $\widehat{p}_n^k$ for all $k \leqslant l$.

Finally, for all considered criteria, the estimator $\widehat{p}_n^k$ performs better than $\widehat{p}_n^{*k}$ when n is small and both estimators perform similarly when n is large.

Let us now consider the case of p is l-monotone and $k > \ell$. When n is small the estimator under constraint of k-monotonocity performs better than the empirical estimator for all criterion except the estimation of the entropy.

When n is large the estimator under shape constraint is biaised and the the empirical estimator performs better. In the particular case where $p = R_\alpha$ we can measure how misspecified the model is. In this case we showed that the more the model is misspecified the worst is the l_2-loss.

7. Proofs

For all discrete function f, let $\mathcal{Q}(f) = \frac{1}{2}\|f - \widetilde{p}\|^2$. When no confusion is possible we write $\widehat{p}$ instead of $\widehat{p}^k$.

7.1. Proofs of the theorems

7.1.1. Proof of Theorem 1 page 6: Characterization of $\widehat{p}$

Let us first prove that $\widehat{p}$ satisfies *1.* or equivalently, that for all integer l the following inequality is satisfied:

$$F^k_{\widehat{p}}(l) - F^k_{\widetilde{p}}(l) \geqslant \beta(f) m^k_l. \tag{13}$$

By definition $\beta(\widehat{p}) = \langle \widehat{p}, \widehat{p} - \widetilde{p} \rangle$, then (13) is equivalent to:

$$\frac{1}{m^k_l}(F^k_{\widehat{p}}(l) - F^k_{\widetilde{p}}(l)) - \sum_{i=0}^{\infty} \widehat{p}(i)(\widehat{p}(i) - \widetilde{p}(i)) \geqslant 0. \tag{14}$$

Let us rewrite this equation by considering limits of the directionnal derivatives.

For all $\varepsilon \in]0, 1]$, $l \geqslant 0$ we define a function $q_{\varepsilon l}$ as follows:

$$q_{\varepsilon l}(i) = (1-\varepsilon)\widehat{p}(i) + \varepsilon \frac{\bar{Q}^k_l(i)}{m^k_l} = \begin{cases} (1-\varepsilon)\widehat{p}(i) + \frac{\varepsilon}{m^k_l}\bar{Q}^k_l(i) \text{ if } i \in \{0, \dots, l\} \\ (1-\varepsilon)\widehat{p}(i) \text{ if } i \geqslant l+1. \end{cases} \tag{15}$$

The function $q_{\varepsilon l}$ is a k-monotone probability then, using the first point of Lemma 5 (see Section 7.4) with $q_\varepsilon = q_{\varepsilon l}$ we obtain:

$$\sum_{i=0}^{\infty} (\widehat{p}(i) - \widetilde{p}(i)) \frac{\bar{Q}^k_l(i)}{m^k_l} - \sum_{i=0}^{\infty} \widehat{p}(i)(\widehat{p}(i) - \widetilde{p}(i)) \geqslant 0.$$

Now, using Lemma 6 (see Section 7.4) we have that for all $k \geqslant 2$ and for all positive discrete measure f:

$$\forall l \in \mathbb{N}^*, \sum_{i=0}^{\infty} f(i)\bar{Q}^k_l(i) = \sum_{i=0}^{l} f(i)\bar{Q}^k_l(i) = F^k_f(l).$$

We choose $f = \widehat{p}$ and we obtain exactly (14).

Let us now show that $\widehat{p}$ satisfies *2.*. Let l be a k-knot of $\widehat{p}$, we need to show that Inequality (14) is an equality. As before we consider $q_{\varepsilon l}$ defined at Equation (15) and show that $q_{\varepsilon l}$ is a k-monotone probability for ε nonpositive small enough. Thanks to the following equality:

$$(-1)^k\Delta^k\bar{Q}_l^k(i) = \begin{cases} 0 \text{ if } i \neq l \\ 1 \text{ if } i = l \end{cases}$$

we get:

$$(-1)^k\Delta^k q_{\varepsilon l}(i) = \begin{cases} (1-\varepsilon)(-1)^k\Delta^k\widehat{p}(l) + \varepsilon/m_l \text{ if } i = l \\ (1-\varepsilon)(-1)^k\Delta^k\widehat{p}(i) \text{ if } i \neq l. \end{cases}$$

Because $\widehat{p}$ is k-monotone, $(-1)^k\Delta^k q_{\varepsilon l}(i) \geqslant 0$ for ε nonpositive small enough and $i \neq l$. As l is a k-knot of $\widehat{p}$, $(-1)^k\Delta^k\widehat{p}(l) > 0$ then $(-1)^k\Delta^k q_{\varepsilon l}(i) \geqslant 0$. Then $q_{\varepsilon l}$ is a k-monotone probability for ε nonpositive small enough and therefore using Lemma 5 (see Section 7.4) with $q_\varepsilon = q_{\varepsilon l}$ we obtain:

$$\sum_{i=0}^{\infty}(\widehat{p}(i) - \widetilde{p}(i))\frac{\bar{Q}_l^k(i)}{m_l^k} - \sum_{i=0}^{\infty}\widehat{p}(i)(\widehat{p}(i) - \widetilde{p}(i)) \leqslant 0,$$

which together with (14) give that the left hand side of the equation is exactly 0.

Conversely assuming that f is a k-monotone probability that satisfies:

$$\frac{F_f^k(l) - F_{\widetilde{p}}^k(l)}{m_l^k} \geqslant \beta(f), \tag{16}$$

with equality if l is a k-knot of f, we have to show that $f = \widehat{p}$. By definition of $\widehat{p}$ we need to show that for all k-monotone probability g we have $\mathcal{Q}(g) \geqslant \mathcal{Q}(f)$.

Let g be a k-monotone probability. Using Lemma 7 (see Section 7.4):

$$\begin{aligned}
\mathcal{Q}(g) - \mathcal{Q}(f) &= \frac{1}{2}\|g - f\|_2^2 + \langle f - \widetilde{p}, g - f\rangle \\
&\geqslant \langle f - \widetilde{p}, g - f\rangle \\
&= \sum_{i=0}^{\infty}(g(i) - f(i))(f(i) - \widetilde{p}(i)) \\
&= \sum_{i=0}^{\infty}(-1)^k\Delta^k(g-f)_i(F_f^k(i) - F_{\widetilde{p}}^k(i)).
\end{aligned}$$

The function g is k-monotone then for all i, $(-1)^k\Delta^k g(i) \geqslant 0$ and using (16) and lemma 7 (see Section 7.4):

$$\begin{aligned}
\mathcal{Q}(g) - \mathcal{Q}(f) &\geqslant \sum_{i=0}^{\infty}(-1)^k\Delta^k g(i)(F_f^k(i) - F_{\widetilde{p}}^k(i)) \\
&\quad - \sum_{i=0}^{\infty}(-1)^k\Delta^k f(i)(F_f^k(i) - F_{\widetilde{p}}^k(i)) \\
&\geqslant \sum_{i=0}^{\infty}(-1)^k\Delta^k g(i)\beta(f)m_i^k - \sum_{i=0}^{\infty}(-1)^k\Delta^k f(i)(F_f^k(i) - F_{\widetilde{p}}^k(i))
\end{aligned}$$

$$\geqslant \beta(f) \sum_{i=0}^{\infty} (-1)^k \Delta^k g(i) m_i^k - \beta(f).$$

Moreover g being a k-monotone probability, according to Property 2, we have the decomposition on the spline basis:

$$\sum_{i=0}^{\infty} (-1)^k \Delta^k g(i) m_i^k = 1.$$

Finally for all k-monotone probability g we find:

$$\mathcal{Q}(g) - \mathcal{Q}(f) \geqslant \beta(f) - \beta(f) = 0.$$

By unicity of the projection we have $f = \widehat{p}$.

7.1.2. Proof of Theorem 2 page 6: Support of $\widehat{p}$

Proof of 1.: The support of $\widehat{p}$ is finite Let us first consider the case where $\beta(\widehat{p}) = 0$. According to Property 5 this is equivalent to $\widehat{p} = \widehat{p}^*$.

The result is proved by contradiction. Let us assume that $\widehat{p}$ has an infinite support then we can build a discrete function $\bar{p}$ satisfying the following properties:

i) $\bar{p} \leqslant \widehat{p}$.
ii) for all $i \leqslant \widetilde{s}$, $\bar{p}(i) = \widehat{p}(i)$.
iii) there exists i such as $\bar{p}(i) < \widehat{p}(i)$.
iv) $\bar{p}$ is k-monotone and non-negative,

with $\widetilde{s}$ the maximum of the support of $\widetilde{p}$.

For this $\bar{p}$ we have the inequality $\|\bar{p} - \widetilde{p}\|_2 < \|\widehat{p} - \widetilde{p}\|_2$ which contradicts the definition of $\widehat{p} = \widehat{p}^*$.

The probability $\bar{p}$ is constructed as follows.

We define for all $j \in \{1, \ldots, k-1\}$ and for all $i \in \mathbb{N}$, the j^{th} derivative function q_j of $\widehat{p}$:

$$\begin{aligned} q_1(i) &= -\widehat{p}(i+1) + \widehat{p}(i) = -\Delta^1 \widehat{p}(i), \\ q_2(i) &= -q_1(i+1) + q_1(i) = \Delta^2 \widehat{p}(i), \\ &\vdots \\ q_{k-1}(i) &= -q_{k-2}(i+1) + q_{k-2}(i) = (-1)^{k-1} \Delta^{k-1} \widehat{p}(i) \end{aligned} \tag{17}$$

We have $q_{j+1}(i) = \Delta^1 q_j(i)$ so for all $i \in \mathbb{N}$:

$$(-1)^k \Delta^k \widehat{p}(i) = (-1)^{k-1} \Delta^{k-1}(q_1(i)) = \ldots = \Delta^1 q_{k-1}(i).$$

Then $\widehat{p}$ is k-monotone (and non-negative) if and only if q_{k-1} is non-increasing (and non-negative).

Because $\widehat{p}$ has an infinite support, all the functions q_j have infinite suppport too. Moreover for all $i \in \mathbb{N}$ we have the following inequalities:

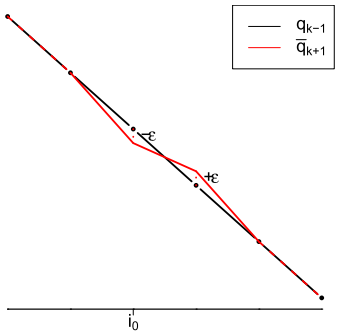


FIG 11. *Functions $q_{k-1}(i)$ and $\bar{q}_{k-1}(i)$ versus i (k odd).*

$$\widehat{p}(i) = -\sum_{h=0}^{i-1} q_1(h) + \widehat{p}(0),$$
$$\forall j \in \{1, \dots, k-2\}, q_j(i) = -\sum_{h=0}^{i-1} q_{j+1}(h) + q_j(0).$$

The next step is to modify q_{k-1} to $\bar{q}_{k-1}$ such that if $\bar{q}_j$ is defined as:

$$\bar{q}_j(i) = -\sum_{h=0}^{i-1} \bar{q}_{j+1}(h) + q_j(0), \forall j \in \{1, \dots, k-2\},$$

and if $\bar{p}$ is defined as:

$$\bar{p}(i) = -\sum_{h=0}^{i-1} \bar{q}_1(h) + \widehat{p}(0), \forall j \in \{1, \dots, k-2\}, \tag{18}$$

then $\bar{p}$ satisfies i)ii)iii)iv).

The function q_{k-1} has an infinite support and is non-increasing, therefore it has an infinity of 1-knots (points where q_{k-1} is strictly non-increasing).

Assume first that k is odd ($k \geqslant 3$). Let i_0 be a 1-knot of q_{k-1} such that $i_0 > \widetilde{s}$. We define $\bar{q}_{k-1}$ as follows:

$$\begin{cases} \bar{q}_{k-1}(i) = q_{k-1}(i) \text{ if } i \neq i_0, i_0 + 1 \\ \bar{q}_{k-1}(i_0) = q_{k-1}(i_0) - \varepsilon \\ \bar{q}_{k-1}(i_0 + 1) = q_{k-1}(i_0 + 1) + \varepsilon. \end{cases} \tag{19}$$

where ε is some positive real number chosen such that $\bar{q}_{k-1}$ is still non-increasing. For example take $\varepsilon = (\bar{q}_{k-1}(i_0) - \bar{q}_{k-1}(i_0 + 1))/2$. The function $\bar{q}_{k-1}$ is shown at Figure 11.

Then the distribution $\bar{p}$ defined at Equation (18) satisfies iv).

To show the properties *i*) to *iii*), we will use the following equality whose proof is straightforward and omitted:

$$\forall i \in \mathbb{N}, \widehat{p}(i) - \bar{p}(i) = (-1)^{k-1} \sum_{h_1=0}^{i-1} \sum_{h_2=0}^{h_1-1} \dots \sum_{h_{k-1}=0}^{h_{k-2}-1} \big(q_{k-1}(h_{k-1}) - \bar{q}_{k-1}(h_{k-1})\big), \tag{20}$$

where the indice h_{k-1} is in the set $\{0, \dots, i-k+1\}$ which is empty if $i \leqslant k-1$.

Let $i \leqslant \widetilde{s}$. According to Equation (20) we get $\widehat{p}(i) - \bar{p}(i) \geqslant 0$ because $\bar{q}_{k-1}(h_{k-1}) = q_{k-1}(h_{k-1})$ for all $h_{k-1} \in \{0, \dots, \widetilde{s} - k + 1\}$. Then the point *ii*) is true.

Let $i = i_0 + k - 1$. Noting that $q_{k-1}(h_{k-1}) = \bar{q}_{k-1}(h_{k-1})$ except in $h_{k-1} = i_0$ we get

$$\widehat{p}(i) - \bar{p}(i) = (-1)^{k-1}(q_{k-1}(i_0) - \bar{q}_{k-1}(i_0)) = +\varepsilon \text{ (because } k \text{ is odd).}$$

and point *iii*) is shown.

It remains to show that $\bar{p} \leqslant \widehat{p}$. By construction of $\bar{q}_{k-1}$, the primitive of q_{k-1} is greater than the primitive of $\bar{q}_{k-1}$, and because $\widehat{p}(i) - \bar{p}(i)$ is nonnegative and the following equality:

$$\widehat{p}(i) - \bar{p}(i) = \sum_{h_1=0}^{i-1} \dots \sum_{h_{k-2}=0}^{h_{k-3}-1} \left(\sum_{h_{k-1}=0}^{h_{k-2}-1} q_{k-1}(h_{k-1}) - \sum_{h_{k-1}=0}^{h_{k-2}-1} \bar{q}_{k-1}(h_{k-1}) \right).$$

we get point *i*).

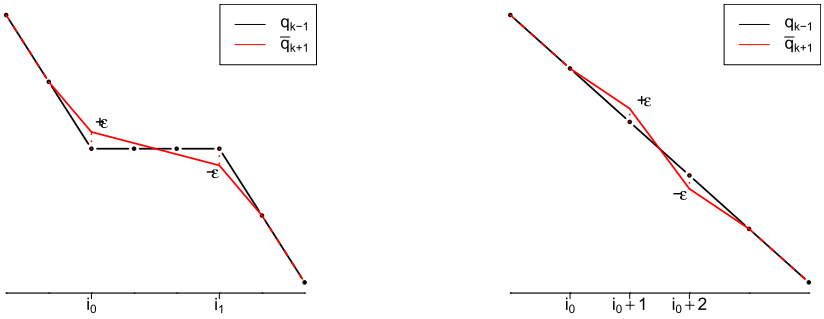
If k is even the proof is based on another construction of $\bar{q}_{k-1}$. Let us first recall that i is a 1-knot of q_{k-1} if $\Delta^1 q_{k-1}(i) = q_{k-1}(i+1) - q_{k-1}(i)$ is strictly negative (because k is even). We have two cases:

Case 1: There exists (i_0, i_1) such that $\widetilde{s} \leqslant i_0 < i_1$, $i_1 - i_0 \geqslant 2$, $\Delta^1 q_{k-1}(i_0)$ and $\Delta^1 q_{k-1}(i_1)$ are strictly negative and $\Delta^1 q_{k-1}(i) = 0$. The probability $\bar{q}_{k-1}$ is defined as follows:

$$\begin{cases} \bar{q}_{k-1}(i) = q_{k-1}(i) \text{ if } i < i_0 + 1, i > i_1 \\ \bar{q}_{k-1}(i_0 + 1) = q_{k-1}(i_0) + \varepsilon \\ \bar{q}_{k-1}(i_1) = q_{k-1}(i_1) - \varepsilon \\ \bar{q} \text{ is an affine function on } [i_0 + 1, i_1]. \end{cases}$$

Case 2: For all $i \geqslant \widetilde{s} + 1$, $\Delta^1 q_{k-1}(i) < 0$. let $i_0 = \widetilde{s} + 1$, then the probability $\bar{q}_{k-1}$ is defined as follows:

$$\begin{cases} \bar{q}_{k-1}(i) = q_{k-1}(i) \text{ if } i < i_0 + 1, i > i_0 + 2 \\ \bar{q}_{k-1}(i_0 + 1) = q_{k-1}(i_0) + \varepsilon \\ \bar{q}_{k-1}(i_0 + 2) = q_{k-1}(i_0 + 2) - \varepsilon. \end{cases}$$



(a) Case 1. (b) Case 2.

FIG 12. *Functions $q_{k-1}(i)$ and $\bar{q}_{k-1}(i)$ versus i (k even).*

The functions $\bar{q}_{k-1}$ are presented in Figure 12. The rest of the proof is similar to the one when k is odd.

Therefore, Theorem 2 is proved in the case $\beta(\widehat{p}) = 0$. Assume now that $\beta(\widehat{p}) \neq 0$. By Theorem 1 we know that if l is a k-knot of $\widehat{p}$, $\beta(\widehat{p})$ is written as follows:

$$\frac{F^k_{\widehat{p}}(l) - F^k_{\widetilde{p}}(l)}{m^k_l} = \beta(\widehat{p}). \tag{21}$$

Let us proved that if the support of $\bar{p}$ is infinite then $\beta(\widehat{p}) = 0$. Indeed if the support of $\bar{p}$ is infinite, $\bar{p}$ has an infinite numbers of k-knots and Equation (21) is true for an infinite numbers of integers l.

Moreover by Equation (11) the term m^k_l is a polynomial function in the variable l with degree k and by Lemma 9 (see Section 7.4) the term $F^k_{\widehat{p}}(l) - F^k_{\widetilde{p}}(l)$ is a polynomial function with degree less than $k-1$. Therefore the left side in Equation (21) tends to zero when l tends to infinity, showing that $\beta(\widehat{p}) = 0$.

Proof of 2.: k-knots' repartition beyond $\widetilde{s}-k+2$ Let us assume that k is odd and prove that if $\widehat{s} \geqslant \widetilde{s}+1$ then $\Delta^k\widehat{p}(r) = 0$ for all $r \in [\widetilde{s}-k+2, \widehat{s}-2]$. We consider $q_1, \ldots, q_{k-1}$ the derivative functions of $\widehat{p}$ defined as before in Equation (17).

As $\widehat{s}$ is a k-knot of $\widehat{p}$ there exist two consecutive 1-knots between r and $\widetilde{s}$. This allows to define the function $\bar{q}_{k-1}$ and $\bar{p}$ as before in Equation (18) and Equation (19).

By construction $\bar{q}_{k-1}$ is non-increasing (and nonnegative) and therefore $\bar{p}$ is k-monotone (and nonnegative). Moreover $\bar{p}$ is lower than $\widehat{p}$, equal to $\widehat{p}$ on $\{0, \ldots, \widetilde{s}\}$ and for $i = r+k-1$ we have $\bar{p}(i) < \widehat{p}(i)$. Moreover $r+k-1 > \widetilde{s}$ because $r \in \{\widetilde{s}-k+2, \ldots, \widehat{s}-1\}$. It follows that $\|\bar{p}-\widetilde{p}\|_2 < \|\widehat{p}-\widetilde{p}\|_2$ which contradicts the definition of $\widehat{p}$. Therefore $\widehat{p}$ does not have any k-knot on $\{\widetilde{s}-k+2, \ldots, \widehat{s}-1\}$.

The proof is similar when k is even.

Remark 1. *The second case requires $q_{k-1}(r+2) > 0$ for $\bar{q}$ to be nonnegative. That is to say we need that $r+2 \leqslant \widehat{s}$. This is the reason why the two sets $\{\widetilde{s}-k+2,\ldots,\widehat{s}-2\}$ and $\{\widetilde{s}-k+2,\ldots,\widehat{s}-1\}$ are different if k is odd or k is even.*

7.1.3. Proof of Theorem 4 page 7: Comparison between the moments of $\widetilde{p}$ and the moments of $\widehat{p}$

Let q a sequence and let ε a real number such that $(1-\varepsilon)\widehat{p}+\varepsilon q$ is a k-monotone probability. Using Lemma 5 (see Section 7.4) we obtain:

$$\sum_{i=0}^{\infty}(\widehat{p}(i)-\widetilde{p}(i))q(i) - \sum_{i=0}^{\infty}\widehat{p}(i)(\widehat{p}(i)-\widetilde{p}(i)) \geqslant 0.$$

By definition $\beta(\widehat{p}) = \langle \widehat{p}, \widehat{p}-\widetilde{p}\rangle$ therefore we have for all $k \geqslant 2$:

$$\sum_{i=0}^{\infty}(\widehat{p}(i)-\widetilde{p}(i))q(i) \geqslant \beta(\widehat{p}).$$

For $q(i) = |i-a|_+^u/m(a,u)$ we get the result. Moreover for $q=\delta_0$ we find $\widehat{p}(0)-\widetilde{p}(0) \geqslant \beta(\widehat{p})$.

7.1.4. Proof of Theorem 5 page 8: Rate of convergence of $\widehat{p}$

The proof is based on Lemma 6.2 of Jankowski and Wellner (2009) [21]. First we assume that $r=2$. Banach's Theorem for projection on a closed convex set says that the projection on the set of k-monotone probabilities is 1-lipschitzienne. Then if $p_{\mathcal{S}_k}$ is the projection of p on the set $\mathcal{S}_k$ we have:

$$\sqrt{n}\|p_{\mathcal{S}_k}-\widehat{p}\|_2 \leqslant \sqrt{n}\|p-\widetilde{p}\|_2$$

We need to show that $\sqrt{n}\|p-\widetilde{p}\|_2 = O_{\mathbb{P}}(1)$, or equivalently that the series $W_n = \sqrt{n}(p-\widetilde{p})$ is tight in $L_2(\mathbb{N})$. Using Lemma 6.2 of Jankowski and Wellner (2009), we have to show that:

$$\begin{cases} \sup\limits_{n\in\mathbb{N}} \mathbb{E}\big[\|W_n\|_2^2\big] < \infty, \\ \lim\limits_{m\to\infty}\sup\limits_{n\in\mathbb{N}} \sum_{j\geqslant m}\mathbb{E}\big[\|W_n\|_2\big] = 0. \end{cases}$$

This is easily verified by noting that $\mathrm{var}(\widetilde{p}(j)) = p(j)(1-p(j))/n$. Then for all $r\in[2,\infty]$, $\sqrt{n}\|p_{S_k}-\widehat{p}\|_r \leqslant \sqrt{n}\|p_{S_k}-\widehat{p}\|_2 = O_{\mathbb{P}}(1)$.

7.1.5. Proof of Theorem 6 page 8: The case of a finite support for p

First part For all integer i, by the strong law of large numbers $\widetilde{p}(i)$ tends a.s. to $p(i)$. Because the maximum of the support s of p is finite we have the following result:

$$a.s. \lim_{n\to\infty} \|\widetilde{p} - p\|_2^2 = \lim_{n\to\infty} \sum_{i=0}^{s} (\widetilde{p}(i) - p(j))^2 = 0.$$

Then by Theorem 3 we get that for all integer i:

$$a.s. \lim_{n\to\infty} (\widehat{p}(i) - p(i))^2 \leqslant \lim_{n\to\infty} \|\widehat{p} - p\|_2^2 \leqslant \lim_{n\to\infty} \|\widetilde{p} - p\|_2^2 = 0.$$

It follows that:

$$a.s. \lim_{n\to\infty} \left[(-1)^k \Delta^k \widehat{p}(j) - (-1)^k \Delta^k p(j)\right] = 0.$$

Because $(-1)^k \Delta^k p(j) > 0$, almost surely for n large enough $(-1)^k \Delta^k \widehat{p}(j) > 0$, which proves that j is a k-knot of $\widehat{p}$.

Second part If $\widehat{s} \leqslant s$ the theorem is true. We assume now that $\widehat{s} > s$.

Let us first consider the case where k is odd. Thanks to the second point of Theorem 2, if we note $\widetilde{s}$ the maximum of the support of $\widetilde{p}$ then $\widehat{p}$ has no k-knot on $\{\widetilde{s} - k + 2, \dots, \widehat{s} - 2\}$.

Moreover as $\widetilde{s} \leqslant s$, $\widehat{p}$ has no k-knot in $\{s - k + 2, \dots, \widehat{s} - 2\}$ (this set may be empty).

The function p is k-monotone and s is a k-knot of p, then by Theorem 6 almost surely there exists n_0 such as for all $n \geqslant n_0$, s is a k-knot of $\widehat{p}$.

It follows that (almost surely) s is not in the set $\{s - k + 2, \dots, \widehat{s} - 2\}$ and therefore $s \geqslant \widehat{s} - 1$ or $\widehat{s} \leqslant s + 1$.

The proof of the result in the case where k is even is similar.

7.1.6. Proof of Theorem 8 page 12: Stopping criterion when $k \in \{3, 4\}$

We first show that $\widehat{p}$ satisfies the four properties stated in *1.* We know by Theorem 1 that it satisfies *1.(a)* and *1.(b)*. Moreover by Lemma 8 (see Section 7.4) it satisfies *1.(d)*. It remains to show 1.(c).

For ε a real number, and for any $j \in \{2, k-1\}$, the function q_ε is defined as follows:

$$q_\varepsilon(i) = (1 - \varepsilon)\widehat{p}(i) + \varepsilon \frac{\bar{Q}^j_{\widehat{s}+1}(i)}{m^j_{\widehat{s}+1}}$$

where $\bar{Q}^j_{s+1}$ is defined at Equation (1).

The function q_ε is a k-monotone probability for ε small enough. Indeed $(-1)^k \Delta^k \bar{Q}^j_{\widehat{s}+1}(i)$ is strictly nonpositive only for $i = \widehat{s}$. Moreover $(-1)^k \Delta^k \widehat{p}(\widehat{s}) = \widehat{p}(\widehat{s}) > 0$ then for ε smaller enough, $(-1)^k \Delta^k q_\varepsilon(\widehat{s}) = (1-\varepsilon)\widehat{p}(\widehat{s}) - \varepsilon Q^j_{\widehat{s}+1}(i)/m^j_{\widehat{s}+1}$ is nonnegative.

Using Lemma 5 (see Section 7.4) we obtain:

$$\sum_{i=0}^{\infty} (\widehat{p}(i) - \widetilde{p}(i)) \frac{\bar{Q}^k_l(i)}{m^k_l} - \sum_{i=0}^{\infty} \widehat{p}(i)(\widehat{p}(i) - \widetilde{p}(i)) \geqslant 0.$$

By Lemma 6 (see Section 7.4) and the fact that $\beta(\widehat{p}) = \langle \widehat{p}, \widehat{p} - \widetilde{p} \rangle$ we deduce that:

$$\frac{F^j_{\widehat{p}}(\widehat{s}+1) - F^j_{\widetilde{p}}(\widehat{s}+1)}{m^j_{\widehat{s}+1}} \geqslant \beta(\widehat{p}),$$

which is exactly *2.(c)*.

Reciprocally we assume now that f satisfies *1.* and we show that $f = \widehat{p}$, which by Theorem 1, is equivalent to show that

$$F^k_f(l) - F^k_{\widetilde{p}}(l) \geqslant \beta(f) m^k_l,$$

for all $l \in \mathbb{N}$ with equality if l is a k-knot of f.

This is true for $l \leqslant s+1$ because f satisfies *2.(a)* and *2.(b)*. Because f has no k-knot after s it remains to show that the inequality is true for $l \geqslant s+2$.

We begin with the case $k = 3$. Because $s \geqslant \widetilde{s}$ and f and $\widetilde{p}$ are probabilities, applying Theorem 4, we obtain that for all $l \geqslant s+1$,

$$\begin{aligned} F^3_f(l) - F^3_{\widetilde{p}}(l) &= \sum_{j=2}^{3} Q^{3-j+1}_{l-1}(s+1)\big(F^j_f(s+1) - F^j_{\widetilde{p}}(s+1)\big) \\ &= \big(F^3_f(s+1) - F^3_{\widetilde{p}}(s+1)\big) + (l-s-1)\big(F^2_f(s+1) - F^2_{\widetilde{p}}(s+1)\big). \end{aligned}$$

As f satisfies *1.(a)* we have:

$$F^3_f(l) - F^3_{\widetilde{p}}(l) \geqslant \beta(f) m^3_{s+1} + (l-s-1)\big(F^2_f(s+1) - F^2_{\widetilde{p}}(s+1)\big).$$

Moreover as f satisfies *1.(c)* we have:

$$\begin{aligned} &F^3_f(l) - F^3_{\widetilde{p}}(l) \\ &\quad \geqslant \beta(f) m^3_{s+1} + (l-s-1)\beta(f) m^2_{s+1} = \beta(f)\big(m^3_{s+1} + (l-s-1)m^2_{s+1}\big). \end{aligned}$$

Finally, because $\beta(f) \leqslant 0$ by *1.(d)*, it remains to show that:

$$m^3_{s+1} + (l-s-1)m^2_{s+1} \leqslant m^3_l. \tag{22}$$

By Equation (11), Equation (22) may be written as follows:

$$\frac{(s+4)(s+3)(s+2)}{6} + (l-s-1)\frac{(s+3)(s+2)}{2} \leqslant \frac{(l+3)(l+2)(l+1)}{6}.$$

After some calculations, we can show that (22) is satisfied if and only if $P_3(l) \geqslant 0$ where P_3 is the polynomial function $P_3(l) = (l-s)(l-(s+1))(l+2s+7)$. This is true because $l \geqslant s+1$.

Let us now prove the case $k = 4$. By Theorem 4 we obtain for all $l \geqslant s+1$:

$$\begin{aligned} F^4_f(l) - F^4_{\widetilde{p}}(l) = \big(F^4_f(s+1) - F^4_{\widetilde{p}}(s+1)\big) + (l-s-1)\big(F^3_f(s+1) - F^3_{\widetilde{p}}(s+1)\big) \\ + Q^3_{l-1}(s+1)\big(F^2_f(s+1) - F^2_{\widetilde{p}}(s+1)\big). \end{aligned}$$

Let $A_I(s) = \Pi_{i\in I}(s+i)$, then using Equation (11) we need to show that:

$$A_{[3,5]}(s) + 4A_{[3,4]}(s)(l-s-s) + 6(l-s-1)(l-s-2)A_{\{3\}}(s) \leqslant \frac{(l+1)A_{[1,4]}(l)}{s+2}. \tag{23}$$

After some calculations, we can show that (23) is satisfied if and only if $P_4(l) \geqslant 0$ where P_4 is the polynomial function

$$P_4(l) = A_{[1,4]}(l) - 6(l-(s+1))(l-(s+2))A_{[2,3]}(s) + 4(l-(s+1))A_{[2,4]}(s) - A_{[2,5]}(s).$$

We have $P_4(l+1) - P_4(l) = 4(P_3(l+1) + 3(s+2)(s+3))$ and $P_4(s+2) = 12(s+3)(s+4) > 0$ then $P_4(l) \geqslant 0$ because $l \geqslant s+2$.

7.2. Proof of the algorithm: Estimating π on a finite support

We show the following theorem for the algorithm described in Section 4:

Theorem 9. *The algorithm described at Table 1 page 10 returns $\widehat{p}_L$ in a finite number of steps.*

7.2.1. Proof of Theorem 9

During step 1 the set S is a subset of $\{0,\dots,L\}$ and π is the minimizer of Ψ on the set S.

The criterion allowing us to determine if $\pi = \widehat{\pi}_L$ (and to stop the algorithm) is given by Lemma 2 (see Section 7.2.2).

In order to show that the algorithm returns $\widehat{\pi}_L$ in a finite number of steps we need to show the both assertions:

- **Assertion 1**: Going from Step 2 to Step 1 is done in a finite number of runs.
- **Assertion 2**: If π_m denotes the value of π at iteration m of the algorithm, then $(\Psi(\pi_m))$ converges to the minimum of Ψ on the set of probabilities with support on $\{0,\dots,L\}$ that is to say to $\widehat{\pi}_L$.

At Step 2 the set S' may be reduced up to one element, but it can not be empty because the minimizer of Ψ on a singleton is non-negative. That proves Assertion 1.

Let us show Assertion 2 by proving that for all $m \in \mathbb{N}^*$, $\Psi(\pi_{m+1}) < \Psi(\pi_m)$. Let S be the support of π_m at iteration m, and let $j \in \{0,\dots,L\}$ be an integer such as $D_{\delta_j}\Psi(\pi_m) < 0$. We have $S' = S + j$ and $\Psi(\pi_{S'}) < \Psi(\pi_S)$ by Lemma 4 (see Section 7.2.2).

We consider two cases:

1: If $\pi_{S'}$ is a nonnegative measure we go to Step 1 with $\pi = \pi_{S'}$. In other terms $\pi_{m+1} = \pi_{S'}$ and therefore $\Psi(\pi_{m+1}) < \Psi(\pi_m) = \Psi(\pi_S)$.

2: If $\pi_{S'}$ is not a nonnegative measure the algorithm iterates inside Step 2 and $\pi_{S'}$ is updated at each loop. We need to verify that at the end of this iterative procedure:

$$\Psi(\pi_{S''}) < \Psi(\pi_S).$$

Let r be the number of times when we go in Step 2 during the m-th loop and let S''_h be the value of the set S'' the h-th time we go in Step 2. We have $\pi_{m+1} = \pi_{S''_h}$.

We show by induction the following property:

$$HR_h : \Psi(\pi_{S''_h}) < \Psi(\pi_S).$$

Thanks to the property *2.* in Lemma 4 (see Section 7.2.2) the property HR_1 is true. Assume now that HR_h is true for some $h \leqslant r-1$, $\Psi(\pi_{S''_h}) < \Psi(\pi_S)$. Let l and ε be defined as follows:

$$l = \underset{i \in S'}{\operatorname{argmin}} \{ \frac{a_i}{a_i - \pi_{S''_h}(i)}, \text{pour } i, \pi_{S''_h}(i) < a_i \},$$
$$\varepsilon = \frac{a_l}{a_l - \pi_{S''_h}(l)}.$$

Then $(1-\varepsilon)\pi_S + \varepsilon\pi_{S''_h}$ is a 1-mass function with support $S''_{h+1} = S''_h - \{l\}$. It follows, by convexity of Ψ that:

$$\begin{aligned}\Psi(\pi_{S''_{h+1}}) &\leqslant \Psi((1-\varepsilon)\pi_S + \varepsilon\pi_{S''_h}) \\ &\leqslant (1-\varepsilon)\Psi(\pi_S) + \varepsilon\Psi(\pi_{S''_h}).\end{aligned}$$

Thanks to HR_h, it follows that:

$$\Psi(\pi_{S''_{h+1}}) < \Psi(\pi_S),$$

and HR_{h+1} is true.

Then HR_r is true, that is to say $\Psi(\pi_{m+1}) < \Psi(\pi_m)$ for all integer m, and $(\Psi(\pi_m))_{m \in \mathbb{N}})$ converges when m tends to infinity (because it is a nonincreasing and bounded sequence). The limit is the minimum of Ψ because the nondecreasing is strict.

7.2.2. Proof of the lemmas used in the proof of Theorem 9

The proof of Theorem 9 is based on the following lemmas whose proofs are given afterwards. All the notations used in this section were defined in Section 4.

Lemma 1. *Let π and μ be two probabilities with support on the set $\{0, \ldots, L\}$. Then we have the following equality:*

$$D_\mu \Psi(\pi) = \sum_{j=0}^{L} \mu(j) D_{\delta_j} \Psi(\pi).$$

Lemma 2. *There is equivalence between:*

1. $\pi = \widehat{\pi}_L$.
2. $\forall j \in \{0, \ldots, L\}$, $D_{\delta_j}\Psi(\pi) \geqslant 0$.

Moreover if $\pi = \widehat{\pi}_L$ *then for all* $j \in supp(\pi)$ *we have* $D_{\delta_j}\Psi(\pi) = 0$.

Lemma 3. *Let* $\mathcal{M}_S$ *be the set of positive measure* π *whose support is included in the set* S. *Let* π_S *and* $\pi_{S'}$ *be defined as follows:*

$$\pi_S = \underset{\substack{\sum_{j\in S}\pi(j)=1 \\ \pi\in\mathcal{M}_S}}{argmin} \big(\Psi(\pi)\big),$$

$$\pi'_S = \underset{\pi\in\mathcal{M}_S}{argmin} \left(\Psi(\pi) + \lambda_S(\sum_{j\in S}\pi(j) - 1)\right).$$

Then we have $\pi_S = \pi'_S$.

The proof of the following lemma is in Durot and al. [13]:

Lemma 4. *Let* $\pi_S = \sum_{i\in S} a_i\delta_i$ *be the minimizer of* Ψ *over the set of nonnegative sequences with support* $S \subset \{0,\dots,L\}$.

Let j *an integer such that* $j \notin S$ *and* $D_{\delta_j}\Psi(\pi_L) < 0$.

Let $\pi^*_{S'} = \sum_{i\in S'} b_i\delta_i$ *be the minimizer of* Ψ *over the set of sequences with support* $S' = S + \{j\}$.

Then, the two following results hold:

1. $\Psi(\pi_{S'}) < \Psi(\pi_S)$.
2. *Assume that* b_i *for some* $i \in S$ *is strictly nonpositive and let:*

$$l = \underset{1\in S'}{argmin}\{\frac{a_i}{a_i - b_i}, pour\ i, b_i < a_i\}.$$

If $\pi_{S''}$ *is the minimizer of* Ψ *over the set of sequences with support* $S'' = S' - \{l\}$, *then* $\Psi(\pi_{S''}) < \Psi(\pi_S)$.

Proof of Lemma 1 Let μ be a probability with support included in $\{0,\dots,L\}$. We write $\mu = \sum_{j=0}^{L}\mu(j)\delta_j$ then, for $L \leqslant \widetilde{s}$:

$$\begin{aligned}
D_\mu\Psi(\pi) &= \lim_{\varepsilon\searrow 0^+}\frac{1}{\varepsilon}\big(\Psi\big((1-\varepsilon)\pi + \varepsilon\mu\big) - \Psi(\pi)\big) \\
&= \lim_{\varepsilon\searrow 0^+}\frac{1}{\varepsilon}\left(\sum_{i=0}^{L}\left(\sum_{l=0}^{L}[(1-\varepsilon)\pi(l) + \varepsilon\mu(l)]Q^k_l(i)) - \widetilde{p}(i)\right)^2\right. \\
&\quad \left. - \sum_{i=0}^{L}\left(\sum_{l=0}^{L}\pi(l)Q^k_l(i)) - \widetilde{p}(i)\right)^2\right) \\
&= \lim_{\varepsilon\searrow 0^+}\frac{1}{\varepsilon}\sum_{i=0}^{L}\left[2\varepsilon\left(\sum_{l=0}^{L}(\mu(l) - \pi(l))Q^k_l(i)\right)\left(\sum_{l=0}^{L}\pi(l)Q^k_l(i) - \widetilde{p}(i)\right)\right] \\
&\quad \left. + \varepsilon^2\left(\sum_{l=0}^{L}(\mu(l) - \pi(l))Q^k_l(i)\right)^2\right]
\end{aligned}$$

$$=2\sum_{i=0}^{L}\left(\sum_{l=0}^{L}\mu(l)Q_l^k(i)-\sum_{l=0}^{L}\pi(l)Q_l^k(i)\right)\left(\sum_{l=0}^{L}\pi(l)Q_l^k(i)-\widetilde{p}(i)\right).$$

In particular for $\mu=\delta_j$ we find:

$$D_{\delta_j}\Psi(\pi)=2\sum_{i=0}^{L}\left(Q_j^k(i)-\sum_{l=0}^{L}\pi(l)Q_l^k(i)\right)\left(\sum_{l=i}^{L}\pi(l)Q_l^k(i)-\widetilde{p}(i)\right).$$

Then, by noting that $\sum_j\mu(j)=1$ we have the following equalities:

$$\begin{aligned}&\sum_{j=0}^{L}\mu(j)D_{\delta_j}\Psi(\pi)\\&\quad=2\sum_{i=0}^{L}\left(\sum_{j=0}^{L}\mu(j)\left(Q_j^k(i)-\sum_{l=0}^{L}\pi(l)Q_l^k(i)\right)\right)\left(\sum_{l=i}^{L}\pi(l)Q_l^k(i)-\widetilde{p}(i)\right)\\&\quad=2\sum_{i=0}^{L}\left(\sum_{j=0}^{L}\mu(j)Q_j^k(i)-\sum_{l=0}^{L}\pi(l)Q_l^k(i)\right)\left(\sum_{l=i}^{L}\pi(l)Q_l^k(i)-\widetilde{p}(i)\right)\end{aligned}$$

and the lemma is proved.

Proof of Lemma 2 We first show that $\widehat{\pi}_L$ satisfies *2.*

For all $0<\varepsilon<1$ and $j\in\{0,\dots,L\}$ the function $(1-\varepsilon)\widehat{\pi}_L+\varepsilon\delta_j$ is a probability and then by definition of $\widehat{\pi}_L$ we have the following inequality:

$$\lim_{\varepsilon\searrow 0^+}\frac{1}{\varepsilon}\big(\Psi((1-\varepsilon)\widehat{\pi}_L+\varepsilon\delta_j)-\Psi(\widehat{\pi}_L)\big)\geqslant 0,$$

which leads to $D_{\delta_j}\Psi(\widehat{\pi}_L)\geqslant 0$, showing the point *2.*

Reciprocally, for π a probability that satisfies *2.*, let us show that $\pi=\widehat{\pi}_L$. Precisely we show that for all probability μ with support on $\{0,\dots,L\}$ we have $\Psi(\mu)-\Psi(\pi)\geqslant 0$. Because Ψ is convex we have:

$$\begin{aligned}D_\mu\Psi(\pi)&=\lim_{\varepsilon\searrow 0^+}\frac{1}{\varepsilon}\big(\Psi\big((1-\varepsilon)\pi+\varepsilon\mu\big)-\Psi(\pi)\big)\\&\leqslant\lim_{\varepsilon\searrow 0^+}\frac{1}{\varepsilon}\big((1-\varepsilon)\Psi(\pi)+\varepsilon\Psi(\mu))-\Psi(\pi)\big)\\&\leqslant\Psi(\mu)-\Psi(\pi),\end{aligned}$$

and by Lemma 1 we have:

$$D_\mu\Psi(\pi)=\sum_{j=0}^{L}\mu(j)D_{\delta_j}\Psi(\pi).$$

Because π satisfies *2.*, $D_\mu\Psi(\pi)\geqslant 0$, and finally $\Psi(\mu)-\Psi(\pi)\geqslant 0$ and $\pi=\widehat{\pi}_L$.

To conclude assume now that $j\in\mathrm{supp}(\widehat{\pi}_L)$. Then the function $(1+\varepsilon)\widehat{\pi}_L-\varepsilon\delta_j$ is a probability for ε positive small enough, and we have the following inequality:

$$-D_{\delta_j}\Psi(\pi) = \lim_{\varepsilon \searrow 0^+} \frac{1}{\varepsilon}\big(\Psi((1+\varepsilon)\widehat{\pi}_L - \varepsilon\delta_j) - \Psi(\widehat{\pi}_L)\big) \geqslant 0,$$

which concludes the proof of the lemma.

Proof of Lemma 3 Let π_S be the solution of the first problem of minimization. Let Q_S and H_S be defined as in Section 4. The KKT's conditions give us that π_S is the unique sequence that satisfies:

$$\exists \lambda_S \in \mathbb{R}, \begin{cases} \sum_{j\in S} \pi_S(j) = 1 \\ \frac{\partial}{\partial \pi}\mathcal{L}(\pi_S, \lambda_S) = 0 \end{cases} \tag{24}$$

where $\mathcal{L}$ is the Lagrange's function:

$$\mathcal{L}(\pi, \lambda) = \Psi(\pi) + \lambda(\sum_{j\in S} \pi(j) - 1).$$

The partial derivative function of $\mathcal{L}$ is:

$$\frac{\partial}{\partial \pi}\mathcal{L}(\pi, \lambda) = -Q_S^T(\widetilde{p} - Q_S\pi) + \lambda Q_S^T\mathbb{I},$$

where $\mathbb{I}$ is the vector with $L+1$ components equal to 1. We have

$$\pi_S = (Q_S^T Q_S)^{-1} Q_S^T(\widetilde{p} - \lambda_S\mathbb{I}) \text{ and } Q_S\pi_S = H_S(\widetilde{p} - \lambda_S\mathbb{I}),$$

leading to:

$$1 = < Q_S\pi_S, \mathbb{I} > = < H_S\widetilde{p}, \mathbb{I} > - \lambda_S < H\mathbb{I}, \mathbb{I} > .$$

Finally we obtain:

$$\lambda_S = \frac{< H\widetilde{p}, \mathbb{I} > -1}{< H\mathbb{I}, \mathbb{I} >}.$$

Then for all π with support included on S we have $\mathcal{L}(\pi_S, \lambda_S) \leqslant \mathcal{L}(\pi, \lambda_S)$ and π_S is solution for the second problem:

$$\pi_S = \underset{\text{supp}(\pi)\subset S}{\text{argmin}} \big(\mathcal{L}(\pi, \lambda_S)\big).$$

Because we are considering strictly convex minimization problems, we get $\pi_S = \pi_S^{'}$.

7.3. Proof of properties

7.3.1. Proof of Property 2 page 4: The link between k-monotonicity and $(k-1)$-monotonicity

We show this result by iteration. First a convex (or 2-monotone) discrete function on $L^1(\mathbb{N})$ is nonincreasing (see [24]).

Let now $k \geqslant 3$. Let $p \in L^1(\mathbb{N})$ be a k-monotone function. We denote q the following discrete function:

$$\forall i \in \mathbb{N}, q(i) = (-1)^{k-2}\Delta^{k-2}p(i).$$

The function q is in $L^1(\mathbb{N})$ and $\Delta^2 q(i) = (-1)^k \Delta^k p(i) \geqslant 0$ for all $i \in \mathbb{N}$. Therefore q is convex and nonincreasing.

It follows that for all $i \in \mathbb{N}$, $(-\Delta^1)((-1)^{k-2}\Delta^{k-2}p(i)) = q(i) - q(i+1) \geqslant 0$ i.e. $(-1)^{k-1}\Delta^{k-1}p(i) \geqslant 0$ and p is $(k-1)$-monotone.

7.3.2. Proof of Property 4 page 11

We prove this property by induction. First for $k = 2$, we have the following equalities:

$$\begin{aligned}
F_f^2(l) - F_{\widetilde{p}}^2(l) &= \sum_{h_1=0}^{l}\sum_{h_2=0}^{h_1}(f(h_2) - \widetilde{p}(h_2)) \\
&= \sum_{h_1=0}^{s}\sum_{h_2=0}^{h_1}(f(h_2) - \widetilde{p}(h_2)) + \sum_{h_1=s+1}^{l}\sum_{h_2=0}^{h_1}(f(h_2) - \widetilde{p}(h_2)) \\
&= F_f^2(s) - F_{\widetilde{p}}^2(s) + \sum_{h_1=s+1}^{l}\sum_{h_2=0}^{s}(f(h_2) - \widetilde{p}(h_2)) \\
&= F_f^2(s) - F_{\widetilde{p}}^2(s) + (F_f^1(s) - F_{\widetilde{p}}^1(s))(l-s)_+.
\end{aligned}$$

Because $\bar{Q}_{l-1}^2(s) = (l-s)_+$ the property is true for $k = 2$.

Assume now that the property is true until the rank $k-1$. We have the following properties:

$$\begin{aligned}
F_f^k(l) - F_{\widetilde{p}}^k(l) &= \sum_{h=0}^{l}(F_f^{k-1}(h) - F_{\widetilde{p}}^{k-1}(h)) \\
&= \sum_{h=0}^{s}(F_f^{k-1}(h) - F_{\widetilde{p}}^{k-1}(h)) + \sum_{h=s+1}^{l}(F_f^{k-1}(l) - F_{\widetilde{p}}^{k-1}(l)) \\
&= F_f^k(s) - F_{\widetilde{p}}^k(s) + \sum_{h=s+1}^{l}\Big(\sum_{j=1}^{k-1}\bar{Q}_{h-1}^{k-1-j+1}(s)\big(F_f^j(s) - F_{\widetilde{p}}^j(s)\big)\Big).
\end{aligned}$$

The last equality is obtained by iteration. Using the definition of the Q_j^k we get:

$$\begin{aligned}
\bar{Q}_{h-1}^{k-j}(l) &= \bar{Q}_{l-1}^{k-j}(l-h+s) \\
&= \bar{Q}_{l-1}^{k-j+1}(l-h+s) - \bar{Q}_{l-1}^{k-j+1}(l-h+s+1),
\end{aligned}$$

and the following equalities:

$$F_f^k(l) - F_{\widetilde{p}}^k(l) = F_f^k(s) - F_{\widetilde{p}}^k(s) + \sum_{j=1}^{k-1}\big(F_f^j(s) - F_{\widetilde{p}}^j(s)\big)$$

$$\times \sum_{h=s+1}^{1} \big(\bar{Q}_{l-1}^{k-j+1}(l-h+s) - \bar{Q}_{l-1}^{k-j+1}(l-h+s+1)\big)$$
$$= F_f^k(s) - F_{\widetilde{p}}^k(s) + \sum_{j=1}^{k-1} \big(F_f^j(s) - F_{\widetilde{p}}^j(s)\big)\big(\bar{Q}_{l-1}^{k-j+1}(s) - \bar{Q}_{l-1}^{k-j+1}(l)\big).$$

Because $\bar{Q}_{l-1}^{k-j+1}(l) = 0$ and $\bar{Q}_{l-1}^{k-k+1}(s) = 1$, we finally obtain:

$$F_f^k(l) - F_{\widetilde{p}}^k(l) = \sum_{j=1}^{k} \bar{Q}_{l-1}^{k-j+1}(s)\big(F_f^j(s) - F_{\widetilde{p}}^j(s)\big).$$

7.3.3. Proof of Property 5 page 13: Link between $\widehat{p}$ and $\widehat{p}^*$

The following property gives a characterization of the estimator $\widehat{p}^*$:

Property 9. *Let $f \in L^1(\mathbb{N})$. There is an equivalence between:*

1. - *For all $l \in \mathbb{N}$ we have $F_f^k(l) \geqslant F_{\widetilde{p}}^k(l)$.*
   - *If l is a k-knot of f, then the previous inequality is an equality.*
2. *$f = \widehat{p}^*$.*

The proof is similar to the proof of Theorem 1 and is omitted. Property 5 is deduced from Property 9.

7.3.4. Proof of Property 6 page 13: The mass of $\widehat{p}^*$ is greater than or equal 1

Let $s_{\max}$ the maximum of $\widehat{s}^*$ and $\widetilde{s}$ (the maxima of the supports of $\widehat{p}^*$ and $\widetilde{p}$ respectively), then using Property 4, for all $l \geqslant s_{\max}$ we have:

$$F_{\widehat{p}^*}^k(l) - F_{\widetilde{p}}^k(l) = \sum_{j=1}^{k} \bar{Q}_{l-1}^{k-j+1}(s_{\max})\big(F_{\widehat{p}^*}^j(s_{\max}) - F_{\widetilde{p}}^j(s_{\max})\big).$$

Because the quantities $\bar{Q}_{l-1}^{j}(s_{\max})$ are polynomial functions of $l - s_{\max}$ with degree $j-1$, we get:

$$F_{\widehat{p}^*}^k(l) - F_{\widetilde{p}}^k(l) = F_{\widehat{p}^*}^1(s_{\max}) - F_{\widetilde{p}}^1(s_{\max})\frac{(l-s_{\max})^{k-1}}{(k-1)!} + o(l^{k-1})$$
$$= (\sum_{j=0}^{s_{\max}} \widehat{p}^*(j) - 1)\frac{(l-s_{\max})^{k-1}}{(k-1)!} + o(l^{k-1}).$$

If $\sum_{j=0}^{s_{\max}} \widehat{p}^*(j) < 1$ then, when l tends to infinity, the right-hand term tends to $-\infty$ and the left-hand term is non-negative by Property 9. Therefore $\sum_{j=0}^{s_{\max}} \widehat{p}^*(j) \geqslant 1$.

7.3.5. Proof of the projection of δ_1 onto $\mathcal{S}^3$ in Section 5

Our purpose is to show that the projection of δ_1 on the cone $\mathcal{S}_3$ has a mass strictly greater than one. After some computational results, we guess that this projection is written as $f = \alpha\bar{Q}_5^3 + \beta\bar{Q}_6^3$. We will now establish a necessary and sufficient condition which makes sure that f is $\widehat{p}^{*3}$. This condition is given in Property 9 (see Section 7.3.3).

We search α and β such as $f = \alpha\bar{Q}_5^3 + \beta\bar{Q}_6^3$ satisfies the stopping criterion. For this p we have $f = (21\alpha+28\beta, 15\alpha+21\beta, 10\alpha+15\beta, 6\alpha+10\beta, 3\alpha+6\beta, \alpha+3\beta, \beta, 0\ldots)$. With elementary calculations we obtain the following necessary conditions for α and β:

$$S1 = \begin{cases} F_p^3(0) = 21\alpha + 28\beta \geqslant 0 = F_{\delta_1}^3(0) \\ F_p^3(1) = 78\alpha + 105\beta \geqslant 1 = F_{\delta_1}^3(1) \\ F_p^3(2) = 181\alpha + 246\beta \geqslant 3 = F_{\delta_1}^3(2) \\ F_p^3(3) = 336\alpha + 461\beta \geqslant 6 = F_{\delta_1}^3(3) \\ F_p^3(4) = 546\alpha + 756\beta \geqslant 10 = F_{\delta_1}^3(4) \\ F_p^3(5) = 812\alpha + 1134\beta \geqslant 15 = F_{\delta_1}^3(5) \\ F_p^3(6) = 1134\alpha + 1596\beta \geqslant 21 = F_{\delta_1}^3(6) \\ F_p^3(7) = 1512\alpha + 2142\beta \geqslant 28 = F_{\delta_1}^3(7) \end{cases}$$

and

$$S2 = \begin{cases} F_p^1(7) = 61\alpha + 89\beta \geqslant 1 = F_{\delta_1}^1(7) \\ F_p^2(2) = 378\alpha + 546\beta \geqslant 7 = F_{\delta_1}^2(7) \end{cases}$$

and

$$S3 = \begin{cases} (-1)^3\Delta^3 p(i) = 0 \text{ if } i > 8 \text{ or if } i < 5 \\ (-1)^3\Delta^3 p(6) = 2\beta \\ (-1)^3\Delta^3 p(5) = 2\alpha. \end{cases}$$

The condition $S1$ assure that f satisfies 1., $S2$ that p satisfies 2.(c) and $S3$ that p satisfies 2.(b).

If we assume that α and β are strictly nonnegative we find the more restrictive necessary condition:

$$S4 = \begin{cases} 812\alpha + 1134\beta = 15 \\ 1134\alpha + 1596\beta = 21 \end{cases} = \begin{cases} 812\alpha + 1134\beta = 15 \\ 54\alpha + 76\beta = 1, \end{cases}$$

whose unique solution is

$$\begin{cases} \alpha = \frac{3}{238} \\ \beta = \frac{1}{238}. \end{cases}$$

Reciprocally if we take α and β like before then f satisfies the conditions $S1$, $S2$ and $S3$. Using Property 9 it follows that f is the projection of δ_1 on the set of 3-monotone sequences.

Let us recall that m_j^k is the mass of $\bar{Q}_j^k$. Then the mass of f is equal to $m(f) = \alpha m_5^3 + \beta m_6^3 = \frac{3}{238} \times 56 + \frac{1}{238} \times 84 = 1.058824$.

7.3.6. Proof of Property 7 page 13: The mass of $\widehat{p}^*$ converges to 1

Let $\varepsilon > 0$ be a real number. Since the set of k-monotone probabilities is included in the set of decreasing probabilities there exists an integer s_ε such as for all k-monotone probability q, the following inequality is true:

$$\sum_{i=s_\varepsilon+1}^{\infty} q(i) \leqslant \varepsilon/4.$$

Let p be a discrete k-monotone probability. The following inequalities are true:

$$\begin{aligned}\Big|\sum_{i=0}^{\infty} \widehat{p}^*(i) - 1\Big| &\leqslant \Big|\sum_{i=0}^{s_\varepsilon} (\widehat{p}^*(i) - p(i))\Big| + \Big|\sum_{i=s_\varepsilon+1}^{\infty} (\widehat{p}^*(i) - p(i))\Big| \\ &\leqslant \sum_{i=0}^{s_\varepsilon} |\widehat{p}^*(i) - p(i)| + \varepsilon/2.\end{aligned}$$

Moreover, by Theorem 3 $\|p - \widehat{p}^*\|_2 \leqslant \|p - \widetilde{p}_n\|_2$ almost surely and therefore we have:

$$\lim_{n\to\infty} \sum_{i=0}^{\infty} (\widehat{p}^*(i) - p(i))^2 = 0.$$

Then almost surely for all $i \in \mathbb{N}$ we have $\lim_{n\to\infty} (\widehat{p}^*(i) - p(i)) = 0$ and $\lim_{n\to\infty} \sum_{i=0}^{s_\varepsilon} |\widehat{p}^*(i) - p(i)| = 0$. Finally almost surely we have:

$$\lim_{n\to\infty} \Big|\sum_{i=0}^{\infty} \widehat{p}^*(i) - 1\Big| = 0.$$

7.3.7. Proof of Property 8 page 14: k-monotonicity for the Poisson law

We prove that the Poisson distribution $\mathcal{P}(\lambda)$ is l-monotone if and only if $\lambda \leqslant \lambda_l$. The distribution $q = \mathcal{P}(\lambda)$ is l-monotone if and only if for all $i \in \mathbb{N}$ we have $(-1)^k \Delta^k q(i) \geqslant 0$. We have for all $l \in \mathbb{N}$ the following equalities:

$$(-1)^k \Delta^k q(i) = \sum_{h=0}^{l} (-1)^h \binom{l}{h} \frac{\lambda^{h+i} e^{-\lambda}}{(h+i)!} = \frac{\lambda^i e^{-\lambda}}{(h+l)!} R_l(\lambda, i)$$

where R_l is the polynomial function defined as follows:

$$R_l(\lambda, i) = \sum_{h=0}^{l} (-1)^h \binom{l}{h} \lambda^h (h+l)\dots(h+i+1).$$

Therefore a necessary condition for $\mathcal{P}(\lambda)$ to be l-monotone is $R_l(\lambda, 0)$ nonnegative which is equivalent to $P_l(\lambda)$ nonnegative where $P_l(\lambda)$ is defined as follows:

$$P_\ell(\lambda) = \sum_{h=0}^{\ell} (-1)^h \frac{(\ell!)^2}{h!((\ell-h)!)^2} \lambda^h.$$

Conversely, because $i \hookrightarrow R_l(\lambda, i)$ is an increasing function for $\lambda \in [0,1]$, the condition is sufficient.

When λ tends to infinity, $P_l(\lambda, 0)$ tends to $+\infty$ then $P_l(\lambda)$ is nonnegative until the smallest root of P_l which is nonnegative. In other terms the previous condition is true in particular for $\lambda \leqslant \lambda_l$.

7.4. Proofs of the technical lemmas

Let us first state technical lemmas used in the proofs given before. Their proofs are given afterwards.

Lemma 5. *Let q be a sequence. For all ε real number we note $q_\varepsilon = (1-\varepsilon)\widehat{p}^k + \varepsilon q$. We note also $D(\widehat{p}^k, \widetilde{p}, q) = \sum_{i=0}^{\infty} \big(\widehat{p}^k(i) - \widetilde{p}(i)\big) q(i) - \sum_{i=0}^{\infty} \big(\widehat{p}^k(i) - \widetilde{p}(i)\big) \widehat{p}^k(i)$.*

1. *We assume that for all $\varepsilon > 0$ the sequence q_ε is a k-monotone probability. Then $D(\widehat{p}^k, \widetilde{p}, q) \geqslant 0$.*
2. *We assume that for all $\varepsilon < 0$ the sequence q_ε is a k-monotone probability. Then $D(\widehat{p}^k, \widetilde{p}, q) \leqslant 0$.*

Lemma 6. *For all integer $k \geqslant 2$, for all $l \in \mathbb{N}$ and for all $f \in \mathcal{P}$, the following assumption is true:*

$$\sum_{i=0}^{l} f(i) \bar{Q}_l^k(i) = F_f^k(l). \tag{25}$$

Lemma 7. *For all $k \geqslant 0$, for all $f \in \mathcal{S}_k$, for all $g \in L^1(\mathbb{N})$:*

$$\sum_{i=0}^{\infty} f(i) g(i) = \sum_{i=0}^{\infty} (-1)^k \Delta^k f(i) F_g^k(i).$$

In particular for all $f \in \mathcal{S}_k$ the coefficient $\beta(f)$ defined at Equation (5) satisfies:

$$\beta(f) = \sum_{i=0}^{\infty} f(i)(f(i) - \widetilde{p}(i)) = \sum_{i=0}^{\infty} (-1)^k \Delta^k f(i) \big(F_f^k(i) - F_{\widetilde{p}}^k(i)\big).$$

Lemma 8. *The coefficient $\beta(\widehat{p})$ defined at Equation (5) is always non-positive*

Lemma 9. *Let $k \geqslant 2$. Let $f \in L^1(\mathbb{N})$, $s \in \mathbb{N}$ and $l \geqslant s$. The following equality is true:*

$$F_f^k(l) - F_{\widetilde{p}}^k(l) = \frac{(l-s)^{k-1}}{(k-1)!} (F_f^1(s) - F_{\widetilde{p}}^1(s)) + o(l^{k-1}).$$

Proof of Lemma 5 We proove the first point. The function q_ε is a k-monotone probability and $\widehat{p}^k$ minimizes $\mathcal{Q}$ on the set of k-monotone probabilities then for all $\varepsilon > 0$ we have $\mathcal{Q}(q_\varepsilon) \geqslant \mathcal{Q}(\widehat{p}^k)$ and:

$$\liminf_{\varepsilon \searrow 0} \frac{1}{\varepsilon}(\mathcal{Q}(q_\varepsilon) - \mathcal{Q}(\widehat{p}^k)) \geqslant 0,$$

that is equivalent to:

$$\liminf_{\varepsilon \searrow 0} \frac{1}{\varepsilon}\left(\sum_{i=0}^{\infty}\left((1-\varepsilon)\widehat{p}^k(i) + \varepsilon q(i) - \widetilde{p}(i)\right)^2 - \sum_{i=0}^{\infty}(\widehat{p}^k(i) - \widetilde{p}(i))^2\right) \geqslant 0.$$

Therefore we have the following inequality:

$$\liminf_{\varepsilon \searrow 0} \frac{1}{\varepsilon}\left(\sum_{i=0}^{\infty}\left[(\widehat{p}^k(i) - \widetilde{p}(i))^2 + 2\varepsilon\big(\widehat{p}^k(i) - \widetilde{p}(i)\big)\big(q(i) - \widehat{p}^k(i)\big)\right.\right.$$
$$\left.\left. + \varepsilon^2\big(q(i) - \widehat{p}^k(i)\big)^2\right] - \sum_{i=0}^{\infty}(\widehat{p}^k(i) - \widetilde{p}(i))^2\right) \geqslant 0,$$

leading to:

$$\liminf_{\varepsilon \searrow 0} \frac{1}{\varepsilon}\left(\varepsilon^2 \sum_{i=0}^{\infty}(q(i) - \widehat{p}^k(i))^2 + 2\varepsilon \sum_{i=0}^{\infty}(\widehat{p}^k(i) - \widetilde{p}(i))(q(i) - \widehat{p}^k(i))\right) \geqslant 0,$$

and finally to:

$$\sum_{i=0}^{\infty}(\widehat{p}^k(i) - \widetilde{p}(i))q(i) - \sum_{i=0}^{\infty}(\widehat{p}^k(i) - \widetilde{p}(i))\widehat{p}^k(i) \geqslant 0.$$

The proof of the second point is similar. The function q_ε is a k-monotone probability and $\widehat{p}^k$ minimizes $\mathcal{Q}$ on the set of k-monotone probabilities then for all $\varepsilon < 0$ we have $\mathcal{Q}(q_\varepsilon) \leqslant \mathcal{Q}(\widehat{p}^k)$ and:

$$\liminf_{\varepsilon \searrow 0} \frac{1}{\varepsilon}(\mathcal{Q}(q_\varepsilon) - \mathcal{Q}(\widehat{p}^k)) \leqslant 0,$$

The following calculations are the same that for the first point.

Proof of Lemma 6 The lemma is proved by induction. Let us first consider $k = 2$. Let f be a positive sequence and $l \in \mathbb{N}$. We have:

$$F_f^2(l) = \sum_{h=0}^{l}\sum_{i=0}^{h} f(i) = \sum_{i=0}^{l}\sum_{h=i}^{l} f(i) = \sum_{i=0}^{l} f(i)(l+1-i) = \sum_{i=0}^{l} f(i)\bar{Q}_l^2(i),$$

and Equation (25) is shown. Assume that Equation (25) is true for $k-1 \geqslant 2$. We have the following equalities:

$$F_f^k(l) = \sum_{h=0}^{l} F_f^{k-1}(h) = \sum_{h=0}^{l}\sum_{i=0}^{h} f(i)\bar{Q}_h^{k-1}(i) = \sum_{i=0}^{l} f(i)\sum_{h=i}^{l}\bar{Q}_h^{k-1}(i).$$

Using Pascal's Triangle and the definition of $\bar{Q}_j^k$, we get:

$$F_f^k(l) = \sum_{i=0}^{l} f(i) \sum_{h=i}^{l} \big(\bar{Q}_h^k(i) - \bar{Q}_h^k(i+1)\big)$$
$$= \sum_{i=0}^{l} f(i) \big(\sum_{h=i}^{l} \bar{Q}_h^k(i) - \sum_{h=i}^{l} \bar{Q}_{h-1}^k(i) \big)$$

where the last equality comes from $\bar{Q}_h^k(i+1) = \bar{Q}_{h-1}^k(i)$ with the convention $\bar{Q}_0^k = 0$. Finally we obtain:

$$F_f^k(l) = \sum_{i=0}^{l} f(i) \big(\bar{Q}_l^k(i)\big),$$

and the lemma is shown.

Proof of Lemma 7 The lemma is proved by induction. First it is true for $k = 0$ with the convention $\Delta^0 f(i) = f(i) = F_f^0(i)$. Assume now that the result if true for some $k - 1 \geqslant 0$. We have the following inequalities:

$$\sum_{i=0}^{\infty} \Delta^k f(i) F_g^k(i) = \sum_{i=0}^{\infty} (\Delta^{k-1} f(i+1) - \Delta^{k-1} f(i)) F_g^k(i)$$
$$= \sum_{i=1}^{\infty} \Delta^{k-1} f(i) F_g^k(i-1) - \sum_{i=0}^{\infty} \Delta^{k-1} f(i) F_g^k(i)$$
$$= \sum_{i=1}^{\infty} \Delta^{k-1} f(i) [F_g^k(i-1) - F_g^k(i)] - \Delta^{k-1} f(0) F_g^k(0)$$
$$= - \sum_{i=1}^{\infty} \Delta^{k-1} f(i) F_g^{k-1}(i)) - \Delta^{k-1} f(0) F_g^{k-1}(0)$$

because $F_g^k(0) = \sum_{h_1=0}^{0} \cdots \sum_{h_k=0}^{0} g(h_k) = f(0) = F_g^{k-1}(0)$.

Remark 2. *This sums are well-defined because thanks to Lemma 6 ((see Section 7.4) we have:*

$$\sum_{l=0}^{\infty} |F_g^k(l) \Delta^k f(l)| = \sum_{l=0}^{\infty} \sum_{i=0}^{l} g(i) \bar{Q}_l^k(i) (-1)^k \Delta^k f(l)$$
$$= \sum_{i=0}^{\infty} \left(\sum_{l=i}^{\infty} (-1)^k \Delta^k f(l) \bar{Q}_l^k(i) \right) g(i).$$

By Property 3 (see Section 2) we have the equality:

$$f(i) = \sum_{l=0}^{\infty} (-1)^k \Delta^k f(l) \bar{Q}_l^k(i).$$

Then $\sum_{l=i}^{\infty} (-1)^k \Delta^k f(l) \bar{Q}_l^k(i) \leqslant 1$ *and finally:*

$$\sum_{l=0}^{\infty} |F_f^k(l)\Delta^k f(l)| \leqslant \sum_{i=0}^{\infty} f(i) < \infty.$$

It follows that $\sum_{i=0}^{\infty} \Delta^k f(i) F_g^k(i) = -\sum_{i=0}^{\infty} \Delta^{k-1} f(i) F_g^{k-1}(i)$ and the lemma is proved.

Proof of Lemma 8 We note $\widehat{s}$ and $\widetilde{s}$ the maxima of the supports of $\widehat{p}$ and $\widetilde{p}$ respectively. We note $s_{\max} = \max(\widehat{s}, \widetilde{s})$. We use Property 4 with $f = \widehat{p}$ and we obtain that for all $l \geqslant s_{\max} + 1$:

$$\begin{aligned} F_{\widehat{p}}^k(l) - F_{\widetilde{p}}^k(l) &= \sum_{j=1}^{k} \bar{Q}_{l-1}^{k-j+1}(s_{\max})\big(F_{\widehat{p}}^j(s_{\max}) - F_{\widetilde{p}}^j(s_{\max})\big) \\ &= \sum_{j=2}^{k} \bar{Q}_{l-1}^{k-j+1}(s_{\max})\big(F_{\widehat{p}}^j(s_{\max}) - F_{\widetilde{p}}^j(s_{\max})\big). \end{aligned}$$

The last equality comes from $F_{\widehat{p}}^1(s_{\max}) = F_{\widetilde{p}}^1(s_{\max}) = 1$ because $\widehat{p}$ and $\widetilde{p}$ are probabilities and $s_{\max}$ is greater than $\widehat{p}$ and $\widetilde{p}$.

Because the quantities $\bar{Q}_{l-1}^j(s_{\max})$ are polynomial functions with degree $j-1$ in the variable $l - s_{\max}$ we write $F_{\widehat{p}}^k(l) - F_{\widetilde{p}}^k(l)$ in the following form:

$$F_{\widehat{p}}^k(l) - F_{\widetilde{p}}^k(l) = \frac{(F_{\widehat{p}}^2(s_{\max}) - F_{\widetilde{p}}^2(s_{\max}))}{(k-2)!}(l-s)^{k-2} + o(l^{k-2}).$$

Thanks to Equation (11), m_l^k is a polynomial function with degree k and we have the following limit:

$$\lim_{l\to\infty} \frac{F_{\widehat{p}}^k(l) - F_{\widetilde{p}}^k(l)}{m_l^k} = 0.$$

Moreover for all $l \in \mathbb{N}$ the characterization of $\widehat{p}$ gives us:

$$\frac{F_{\widehat{p}}^k(l) - F_{\widetilde{p}}^k(l)}{m_l^k} \geqslant \beta(\widehat{p}).$$

Necessarily $\beta(\widehat{p}) \leqslant 0$.

Proof of Lemma 9 We show this result by induction. For $k = 2$ the result is shown in [13]. Assume that the result is true for some $k - 1 \geqslant 2$. We have the following equalities:

$$\begin{aligned} &F_f^k(l) - F_{\widetilde{p}}^k(l) \\ &\quad = \sum_{h=0}^{l} \Big(F_f^{k-1}(h) - F_{\widetilde{p}}^{k-1}(h)\Big) \end{aligned}$$

$$
\begin{aligned}
&= \sum_{h=0}^{s} \left(F_f^{k-1}(h) - F_{\tilde{p}}^{k-1}(h) \right) + \sum_{h=s+1}^{l} \left(F_f^{k-1}(h) - F_{\tilde{p}}^{k-1}(h) \right) \\
&= \left(F_f^{k}(s) - F_{\tilde{p}}^{k}(s) \right) + \sum_{h=s+1}^{l} \left(\frac{(h-s)^{k-2}}{(k-2)!} \left(F_f^{1}(s) - F_{\tilde{p}}^{1}(s) \right) + o(h^{k-2}) \right) \\
&= \frac{(l-s)^{k-1}}{(k-1)!} (F_f^1(s) - F_{\tilde{p}}^1(s)) + o(l^{k-1}).
\end{aligned}
$$

The last equality is due to a result of Bernoulli for Faulhaber's sum: the k-th sum of Faulhaber is denoted by S_k and defined as follows:

$$S_k(m) = \sum_{i=1}^{m} i^k.$$

It is shown that:

$$S_k(m) = \frac{1}{k+1} \sum_{j=0}^{k} C_{k+1}^{j} B_j m^{k+1-j} = \frac{1}{k+1} \left(m^{k+1} + \frac{k+1}{2} m^k + \dots \right)$$

where the B_j are Bernoulli's numbers (with the convention $B_1 = \frac{1}{2}$). A proof of this result can be found in [10].

Acknowledgment

The author thanks Sylvie Huet and Christophe Giraud for helpful comments and the careful reading.

References

[1] Balabdaoui, F. (2004). Nonparametric estimation of a k-monotone density: A new asymptotic distribution theory. PhD thesis, University of Washington. MR2705939

[2] Balabdaoui, F. and Durot, C. (2015). Marshall lemma in discrete convex estimation. *Statistics & Probability Letters* **99** 143–148. MR3321508

[3] Balabdaoui, F., Durot, C. and Koladjo, F. (2014). On asymptotics of the discrete convex LSE of a pmf. *arXiv preprint arXiv:1404.3094*.

[4] Balabdaoui, F., Jankowski, H., Rufibach, K. and Pavlides, M. (2013). Asymptotics of the discrete log-concave maximum likelihood estimator and related applications. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)* **75** 769–790. MR3091658

[5] Balabdaoui, F., Rufibach, K. and Wellner, J. A. (2009). Limit distribution theory for maximum likelihood estimation of a log-concave density. *Annals of Statistics* **37** 1299. MR2509075

[6] Balabdaoui, F. and Wellner, J. A. (2007). Estimation of a k-monotone density: Limit distribution theory and the spline connection. *The Annals of Statistics* **35** 2536–2564. MR2382657

[7] Balabdaoui, F. and Wellner, J. A. (2010). Estimation of a k-monotone density: Characterizations, consistency and minimax lower bounds. *Statistica Neerlandica* **64** 45–70. MR2830965

[8] Boyd, S. and Vandenberghe, L. (2004). *Convex Optimization*. Cambridge university press. MR2061575

[9] Bunge, J., Willis, A. and Walsh, F. (2014). Estimating the number of species in microbial diversity studies. *Annual Review of Statistics and Its Application* **1** 427–445.

[10] Conway, J. H. and Guy, R. K. (1996). *The Book of Numbers*. Springer-Verlag. MR1411676

[11] Dümbgen, L. and Rufibach, K. (2011). logcondens: Computations related to univariate log-concave density estimation. *Journal of Statistical Software* **39** 1–28.

[12] Dümbgen, L., Rufibach, K. et al. (2009). Maximum likelihood estimation of a log-concave density and its distribution function: Basic properties and uniform consistency. *Bernoulli* **15** 40–68. MR2546798

[13] Durot, C., Huet, S., Koladjo, F. and Robin, S. (2013). Least-squares estimation of a convex discrete distribution. *Computational Statistics & Data Analysis* **67** 282–298. MR3079603

[14] Durot, C., Huet, S., Koladjo, F. and Robin, S. (2015). Nonparametric species richness estimation under convexity constraint. *Environmetrics* **26** 502–513. MR3415569

[15] Fejér, L. (1936). Trigonometrische Reihen und Potenzreihen mit mehrfach monotoner Koeffizientenfolge. *Transactions of the American Mathematical Society* **39** 18–59. MR1501832

[16] Feller, W. et al. (1939). Completely monotone functions and sequences. *Duke Math. J* **5** 661–674. MR0000315

[17] Fisher, R. A., Corbet, A. S. and Williams, C. B. (1943). The relation between the number of species and the number of individuals in a random sample of an animal population. *The Journal of Animal Ecology* **12** 42–58.

[18] Grenander, U. (1956). On the theory of mortality measurement: Part ii. *Scandinavian Actuarial Journal* **1956** 125–153. MR0093415

[19] Groeneboom, P., Jongbloed, G. and Wellner, J. A. (2001). Estimation of a convex function: Characterizations and asymptotic theory. *Annals of Statistics* **29** 1653–1698. MR1891742

[20] Groeneboom, P., Jongbloed, G. and Wellner, J. A. (2008). The Support Reduction Algorithm for computing non-parametric function estimates in mixture models. *Scandinavian Journal of Statistics* **35** 385–399. MR2446726

[21] Jankowski, H. K. and Wellner, J. A. (2009). Estimation of a discrete monotone distribution. *Electronic Journal of Statistics* **3** 1567. MR2578839

[22] Jewell, N. P. (1982). Mixtures of exponential distributions. *The Annals of Statistics* **10** 479–484. MR0653523

[23] Knopp, K. (1925). Mehrfach monotone Zahlenfolgen. *Mathematische Zeitschrift* **22** 75–85. MR1544715

[24] Koladjo, F. (2013). Estimation d'une distribution discrete sous contrainte de convexité: Application a l'estimation du nombre d'especes de la faune ichtyologique du bassin du fleuve Ouémé. PhD thesis, Université Paris-Sud XI et d'Abomey-Calavi.

[25] Lefevre, C., Loisel, S. et al. (2013). On multiply monotone distributions, continuous or discrete, with applications. *Journal of Applied Probability* **50** 827–847. MR3102517

[26] Lévy, P. (1962). Extensions d'un théorème de D. Dugué et M. Girault. *Probability Theory and Related Fields* **1** 159–173. MR0145565

[27] Williamson, R. E. et al. (1955). On Multiply Monotone Functions and Their Laplace Transforms. PhD thesis, Graduate School of Arts and Sciences, University of Pennsylvania. MR2612312