Computational Statistics and Data Analysis 67 (2013) 282–298

Contents lists available at SciVerse ScienceDirect

Computational Statistics and Data Analysis

journal homepage: www.elsevier.com/locate/csda

Least-squares estimation of a convex discrete distribution

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

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

ARTICLE INFO

Article history:
Received 28 February 2012
Received in revised form 24 April 2013
Accepted 29 April 2013
Available online 16 May 2013

Keywords:
Convex discrete distribution
Non-parametric estimation
Least squares
Support reduction algorithm
Abundance distribution

ABSTRACT

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

1. Introduction

Recently, the problem of estimating a discrete probability mass function under a shape constraint has attracted attention: Jankowski and Wellner (2009) considered the non-parametric estimation of a monotone distribution and Balabdaoui et al. (in press) considered the case of a log-concave distribution. Although the discrete case is in some ways very different from the continuous case (for example, the convergence rates are typically different in the two cases), the construction of shape-constrained estimators in the discrete case is largely inspired by the construction of shape-constrained estimators of a probability density function. The non-parametric estimation, based on i.i.d. observations, of the distribution of a continuous random variable under a shape constraint has received a great deal of attention in the past decades; see Balabdaoui and Wellner (2007) for a review. The most studied constraint is the monotonicity of the density function. It is well-known that the non-parametric maximum likelihood estimator of a decreasing density function over $[0, \infty)$ is the Grenander estimator defined as the left-continuous slope of the least concave majorant of the empirical distribution function of the observations. This estimator can be easily implemented using the PAVA (pool adjacent violators algorithm) or a similar device; see Barlow et al. (1972). The non-parametric maximum likelihood of a log-concave density function (i.e., a density function f such that $\log(f)$ is a concave function) was introduced by Walther (2002) and algorithmic aspects were treated by Dümbgen et al. (2007); see also the R package in Dümbgen and Rufibach (2011). Another well studied constraint is the convexity (or concavity) of the density function over a given interval. It was shown by Groeneboom et al. (2001) that both the least squares estimator and the non-parametric maximum likelihood estimator under the convexity constraint exist and are unique. However, although a precise characterization of these estimators is given in that paper, their practical implementation

* Corresponding author. Tel.: +33 134652224.
E-mail address: sylvie.huet@jouy.inra.fr (S. Huet).

0167-9473/$ – see front matter © 2013 Elsevier B.V. All rights reserved.
http://dx.doi.org/10.1016/j.csda.2013.04.019

is a non-trivial issue: it requires sophisticated iterative algorithms that use a mixture representation, such as the *support reduction algorithm* described by Groeneboom et al. (2008).

In this paper, we consider the non-parametric estimation of a discrete distribution on $\mathbb{N}$ under the convexity constraint (note that a convex distribution on $\mathbb{N}$ is necessarily non-increasing, so in our setting, convex is equivalent to non-increasing convex). This problem has not yet been considered in the literature, although it has several applications, such as the estimation of species abundance distribution in ecology (Lanumteang and Böhning, 2011). In this field, the term "non-parametric methods" often refers to finite mixtures of parametric distributions where only the mixing distribution is inferred in a non-parametric way, see e.g. Böhning and Kuhnert (2006), Böhning et al. (2005), Chao and Shen (2004).

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

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

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

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

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

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

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

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

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

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

the discrete Laplacian. Let $\mathcal{C}$ be the set of convex discrete functions $f \in \ell_2(\mathbb{N})$, that is, the set of all $f \in \ell_2(\mathbb{N})$ having $\Delta f(j) \geqslant 0$ for all integers $j \geqslant 1$, and let $\mathcal{C}_1$ be the set of all convex probability mass functions on $\mathbb{N}$, that is the set of functions $f \in \mathcal{C}$ satisfying $\sum_{i\geqslant 0} f(i) = 1$. An integer $j \geqslant 1$ is a knot of $f \in \mathcal{C}$ if $\Delta f(j) > 0$.

It should be noticed that any $f \in \mathcal{C}$ has $\lim_{j\to\infty} f(j) = 0$, so by convexity, any $f \in \mathcal{C}$ is non-negative, non-increasing and strictly decreasing on its support. For example, any mixture of triangular distributions is convex, the geometric distribution and the Poisson distribution with parameter smaller than $2 - \sqrt{2}$ are convex (see e.g. Murota, 2009 for more on convex discrete functions).

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

2. The constrained LSE of a convex discrete distribution

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

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

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