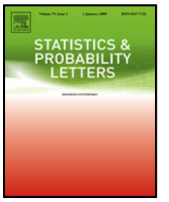
ARTICLE IN PRESS

Statistics and Probability Letters xx (xxxx) xxx–xxx

Contents lists available at ScienceDirect

Statistics and Probability Letters

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



Testing convexity of a discrete distribution

Fadoua Balabdaoui [a,b], Cécile Durot [c,*], Babagnidé François Koladjo [d]

[a] Seminar für Statistik, E.T.H. Zürich, 8092, Zürich, Switzerland
[b] CEREMADE, Université Paris-Dauphine, PSL Research University, 75775, Paris, France
[c] Modal'x, Université Paris Nanterre, F-92001, Nanterre, France
[d] ENSPD, Université de Parakou, BP 55 Tchaourou, Benin

ARTICLE INFO

Article history:
Received 9 November 2016
Received in revised form 28 August 2017
Accepted 16 October 2017
Available online xxxx

ABSTRACT

Based on the convex least-squares estimator, we propose two different procedures for testing convexity of a probability mass function supported on $\mathbb{N}$ with an unknown finite support. The procedures are shown to be asymptotically calibrated.

1. The testing problem

Modeling count data is an important issue in statistical research, see e.g. Gómez-Déniz and Calderín-Ojeda (2011). A popular parametric model for such data is the Poisson model. While non-parametric extensions are conceivable, those that only assume a shape constraint of the underlying probability mass function (pmf) may offer more flexibility. Recent papers on estimating a pmf under a shape constraint are Jankowski and Wellner (2009), Durot et al. (2013, 2015), Balabdaoui et al. (2013), Giguelay (2016), Chee and Wang (2016). In any case, it is sensible to validate the chosen model using a goodness-of-fit test. Goodness-of-fit tests to validate a model connected to the Poisson distribution are given in Karlis and Xekalaki (2000), Meintanis and Nikitin (2008), Ledwina and Wyłupek (2016) and the references therein. However, to the best of our knowledge, no goodness-of-fit test has been proposed to validate an assumed shape constraint for discrete data.

Motivated by the biological application in Durot et al. (2015), where the number of species is estimated assuming a convex abundance distribution, we develop here a goodness-of-fit test for convexity of the underlying pmf on $\mathbb{N}$. To the best of our knowledge, this is the first attempt to build a convexity test for count data. Precisely, based on i.i.d. observations $X_1, \ldots, X_n$ from some pmf p_0 on $\mathbb{N}$, we test the null hypothesis H_0: "p_0 is convex on $\mathbb{N}$" (i.e. $p_0(k+1) - p_0(k) \geq p_0(k) - p_0(k-1)$ for all integers $k \geq 1$) versus H_1: "p_0 is not convex". The test is based on the convex least-squares estimator $\widehat{p}_n := argmin_{p \in \mathcal{C}_1} \|p_n - p\|$, where $\mathcal{C}_1$ is the set of all convex pmfs on $\mathbb{N}$, $\|q\|^2 = \sum_{j \in \mathbb{N}} (q(j))^2$ for any sequence $q = \{q(j), j \in \mathbb{N}\}$, and $p_n(j) = n^{-1} \sum_{i=1}^n \mathbb{1}_{\{X_i = j\}}$, $j \in \mathbb{N}$, is the empirical pmf. It is proved in Durot et al. (2013, Sections 2.1 to 2.3) that $\widehat{p}_n$ exists, is unique, and can be implemented with an appropriate algorithm. We reject H_0 if $\{T_n > t_{\alpha,n}\}$ where $T_n = \sqrt{n}\|p_n - \widehat{p}_n\|$ and $t_{\alpha,n}$ is an appropriate quantile, chosen in such a way that the test has asymptotic level α.

In the sequel, we assume that p_0 has a finite support in $\{0, \ldots, S\}$ with an unknown integer $S > 0$ and we consider two different constructions of $t_{\alpha,n}$. First, we define $t_{\alpha,n}$ as the $(1-\alpha)$-quantile of a random variable whose limiting distribution coincides with the limiting distribution of T_n under H_0. Next, we calibrate the test under a least favorable hypothesis (when the true pmf is triangular). Theoretical justification requires knowledge of the limiting distribution of T_n under H_0. This needs some notation. For all $p = \{p(j), j \in \mathbb{N}\}$ and $k \in \mathbb{N} \setminus \{0\}$ we set $\Delta p(k) = p(k+1) - 2p(k) + p(k-1)$ (hence p is convex on $\mathbb{N}$ iff $\Delta p(k) \geq 0$ for all k) and a given $k \in \mathbb{N} \setminus \{0\}$ is called a knot of p if $\Delta p(k) > 0$. For all $s > 0$ and $u = (u(0), \ldots, u(s+1)) \in \mathbb{R}^{s+2}$,

* Corresponding author.
E-mail address: cecile.durot@gmail.com (C. Durot).

https://doi.org/10.1016/j.spl.2017.10.023
0167-7152/

we set $\|u\|_s^2 = \sum_{k=0}^{s+1}(u(k))^2$. Also, let g_0 be a $(S+2)$ centered Gaussian vector whose dispersion matrix Γ_0 has component $(i+1, j+1)$ equal to $\mathbb{1}_{\{i=j\}}p_0(i) - p_0(i)p_0(j)$ for all $i, j = 0, \ldots, S+1$, and $\widehat{g}_0$ the minimizer of $\|g - g_0\|_S$ over

$$g \in \mathcal{K}_0 := \Big\{g = (g(0), \ldots, g(S+1)) \in \mathbb{R}^{S+2} : \Delta g(k) \geq 0 \text{ for all } k \in \{1, \ldots, S\} \text{ such that } \Delta p_0(k) = 0\Big\}.$$

Existence, uniqueness and characterization of $\widehat{g}_0$ are given in Balabdaoui et al. (2017, Theorem 3.1). The asymptotic distribution of T_n under H_0 is given below.

Theorem 1.1. *(i) The distribution function of* $\widehat{T}_0 := \|\widehat{g}_0 - g_0\|_S$ *is continuous on* $(0, \infty)$. *(ii) Under* H_0, $T_n \xrightarrow{d} \widehat{T}_0$ *and* $\sup_{t\geq 0}|P(T_n \leq t) - P(\widehat{T}_0 \leq t)| \to 0$, *as* $n \to \infty$.

2. Calibrating by estimating the limiting distribution

Here, we build a random variable that weakly converges to $\widehat{T}_0$ and can be approximated via Monte-Carlo simulations. Let $S_n = \max\{X_1, \ldots, X_n\}$, and let g_n be a random vector which, conditionally on $(X_1, \ldots, X_n)$, is a $S_n + 2$ centered Gaussian vector whose dispersion matrix Γ_n has component $(i+1, j+1)$ equal to $\mathbb{1}_{\{i=j\}}p_n(i) - p_n(i)p_n(j)$ for all $i, j = 0, \ldots, S_n + 1$. Now, let $\widehat{g}_n = argmin_{g\in\mathcal{K}_n}\|g - g_n\|_{S_n}$, the least squares projection of g_n on $\mathcal{K}_n$, where $\mathcal{K}_n$ "approaches" $\mathcal{K}_0$ as $n \to \infty$:

$$\mathcal{K}_n = \Big\{g = (g(0), \ldots, g(S_n+1)) \in \mathbb{R}^{S_n+2} : \Delta g(k) \geq 0 \text{ for all } k \in \{1, \ldots, S_n\} \text{ such that } \Delta\widehat{p}_n(k) \leq v_n\Big\}$$

for an appropriate positive sequence $(v_n)_n$. Choosing $v_n = 0$ would make $\mathcal{K}_n$ to be the largest possible and hence $\|\widehat{g}_n - g_n\|_{S_n}$ the smallest possible; this distance would be stochastically smaller than the actual limit of T_n, yielding a large probability of rejection. In fact, choosing $v_n = 0$ amounts to estimate the knots of p_0 by those of $\widehat{p}_n$, which is not desirable since $\widehat{p}_n$ has typically more knots than p_0. The conditions required on v_n are given below.

Theorem 2.1. *Let* g_n, $\mathcal{K}_n$, *and* $\widehat{g}_n$ *be as above, and take* $v_n > 0$ *such that* $v_n = o(1)$ *and* $v_n \gg n^{-1/2}$. *(i) Then,* $\widehat{g}_n$ *uniquely exists, both* $\widehat{g}_n$ *and* $\widehat{T}_n := \|\widehat{g}_n - g_n\|_{S_n}$ *are measurable. (ii) Under* H_0, *conditionally on* $X_1, \ldots, X_n$ *we have* $\widehat{T}_n \xrightarrow{d} \widehat{T}_0$ *in probability as* $n \to \infty$.

By (i) in Theorem 1.1, the conditional convergence $\widehat{T}_n \xrightarrow{d} \widehat{T}_0$ in probability means that

$$\sup_{t\in\mathbb{R}}\left|P(\widehat{T}_n \leq t|X_1, \ldots, X_n) - P(\widehat{T}_0 \leq t)\right| = o_p(1). \tag{2.1}$$

In Balabdaoui et al. (2017, Theorem 3.3) it is shown that $\lim_{n\to\infty}P(\mathcal{K}_n \neq \mathcal{K}_0) = 0$ for any $(v_n)_n$ satisfying the conditions of the theorem. The intuition behind is as follows: when k is a knot of p_0 and $\Delta\widehat{p}_n(k) \leq v_n$, then $\sqrt{n}(\Delta\widehat{p}_n(k) - \Delta p_0(k)) < -\sqrt{n}\epsilon_0$ for some $\epsilon_0 > 0$ and n large enough. Weak convergence of $\widehat{p}_n$ to p_0 implies that this happens with probability converging to zero. In case k is not a knot; i.e., $\Delta p_0(k) = 0$ such that $\Delta\widehat{p}_n(k) > v_n$ then $\sqrt{n}\Delta\widehat{p}_n(k) > \sqrt{n}v_n \to \infty$, which again happens with decreasing probability. We now state the main result of the section, which is proven is the supplement.

Theorem 2.2. *Let* $\widehat{T}_n$ *as in Theorem 2.1. Let* $\alpha \in (0, 1)$ *and* $t_{\alpha,n}$ *the conditional* $(1-\alpha)$*-quantile of* $\widehat{T}_n$ *given* $X_1, \ldots, X_n$. *Under* H_0, $\limsup_{n\to\infty}P\big(T_n > t_{\alpha,n}\big) \leq \alpha$.

Hence, the test is asymptotically calibrated. In fact, it can be shown that the asymptotic Type I error is precisely equal to α for appropriate range of α, i.e. $\lim_{n\to\infty}P(T_n > t_{\alpha,n}) = \alpha$, see the supplementary material. An approximative value of $t_{\alpha,n}$ can be computed using Monte-Carlo simulations as follows. Having observed $X_1, \ldots, X_n$, draw independent sequences $(Z_i^{(b)})_{0\leq i\leq S_n+1}$ for $b \in \{1, \ldots, B\}$, of i.i.d. $\mathcal{N}(0, 1)$ variables $Z_i^{(b)}$, where $B > 0$ is the number of Monte-Carlo runs. For all b, compute $g_n^{(b)} = \Gamma_n^{1/2}(Z_0^{(b)}, \ldots, Z_{S_n+1}^{(b)})^T$ and $\widehat{g}_n^{(b)}$ the minimizer of $\|g_n^{(b)} - g\|_{S_n}$ over $\mathcal{K}_n$ using Dykstra's algorithm, see Balabdaoui et al. (2017). If $(s_j)_j$ is the sequence of successive knots of $\widehat{p}_n$ such that $\Delta\widehat{p}_n(s_j) > v_n$, then the algorithm performs iterative projections on the cones $\{g \in \mathbb{R}^{S_n+2} : \Delta g(k) \geq 0 \text{ for all } k \in \{s_j + 1, \ldots, s_{j+1} - 1\}\}$, whose intersection is precisely $\mathcal{K}_n$. See Dykstra (1983) for more details and a proof of convergence. Then, $t_{\alpha,n}$ can be approximated by the $(1-\alpha)$-quantile of the empirical distribution corresponding to $\|g_n^{(b)} - \widehat{g}_n^{(b)}\|_{S_n}$, with $b \in \{1, \ldots, B\}$.

3. Calibrating under the least favorable hypothesis

We consider below an alternative calibration that is easier to implement than the first one since it does not involve a sequence (v_n). In what follows we denote by $\mathcal{T}_a$ the triangular pmf supported on $\{0, \ldots, a-1\}$ for a given integer $a \geq 1$; i.e., $\mathcal{T}_a(i) = 2(a-i)_+[a(a+1)]^{-1}$. Consider $\widetilde{\mathcal{K}}_0$ the set of all vectors $g = (g(0), \ldots, g(S+1)) \in \mathbb{R}^{S+2}$ such that $\Delta g(k) \geq 0$

Please cite this article in press as: Balabdaoui F., et al., Testing convexity of a discrete distribution. Statistics and Probability Letters (2018), https://doi.org/10.1016/j.spl.2017.10.023.