Characterizations of non-normalized discrete probability distributions and their application in statistics

Steffen Betsch and Bruno Ebner and Franz Nestmann

Institute of Stochastics,
Karlsruhe Institute of Technology (KIT),
Karslruhe, BW, Germany
e-mail: steffen.betsch@kit.edu*;* bruno.ebner@kit.edu*;* franz.nestmann2@kit.edu

Abstract: From the distributional characterizations that lie at the heart of Stein's method we derive explicit formulae for the mass functions of discrete probability laws that identify those distributions. These identities are applied to develop tools for the solution of statistical problems. Our characterizations, and hence the applications built on them, do not require any knowledge about normalization constants of the probability laws. To demonstrate that our statistical methods are sound, we provide comparative simulation studies for the testing of fit to the Poisson distribution and for parameter estimation of the negative binomial family when both parameters are unknown. We also consider the problem of parameter estimation for discrete exponential-polynomial models which generally are non-normalized.

Contents

1. Introduction

In recent years, research in probability and statistics has witnessed the rise of Stein's method but also the emergence of methods to tackle the analysis and application of models which are based on non-normalized probability laws. In this work, we seek to apply findings from the research on Stein's method to contribute to the solution of testing and estimation problems involving non-normalized statistical distributions. We focus on the analysis of discrete probability laws, and how the theoretical results can be used to develop statistical methods. As such, we tie on to Betsch and Ebner (2021) who provide similar tools for continuous distributions. A rather well-known approach to the problem of parameter estimation for non-normalized continuous probability distributions is the score matching technique due to Hyvärinen (2005, 2007) (see Yu, Drton and Shojaie, 2019, for recent progress). Another approach is known as noise contrastive estimation (cf. Gutmann and Hyvärinen, 2010), but a number of 2019/20 papers indicate that the proposition and study of new tools remains an important issue, see Matsuda and Hyvärinen (2019), Uehara et al. (2020), and Uehara, Matsuda and Kim (2020).

The tool box which is now known as Stein's method goes back to the work of Stein (1972) (see also Stein, 1986) who sought for an alternative proof of the central limit theorem that provides a bound on the rate of the convergence. This inherent feature made Stein's method popular. The idea is applied in all kinds of settings, as it allows to find bounds on distributional distances between sequences of probability laws and a limit distribution, and as it often applies in the absence of stochastic independence. The application of the method to discrete distributions goes back to Chen (1975), who first derived corresponding results for the Poisson distribution, known as the Stein-Chen method. The method has since been extended to other discrete distributions, like the binomial distribution (by Ehm, 1991), the geometric distribution (by Peköz, 1996), the negative binomial distribution (by Brown and Phillips, 1999), discrete Gibbs measures (by Eichelsbacher and Reinert, 2008), and others. The foundation of the method are characterization results for the underlying probability law. While for the first distributions in consideration specific identities were used or devised, general approaches have emerged that apply to many different distributions at once. In this context, we mention the generator approach of Barbour (1988, 1990) and Götze (1991) who use time-reversible Markov processes, where the stationary distribution is the probability law of interest, to characterize that law. On the other hand, a direct derivation of characterizations is possible, and a well-known class of such identities can be found under the name of 'density approach'. For the continuous case, first ideas on the density approach came from Stein (1986) and Stein et al. (2004), and a more complete version is due to Ley and Swan (2013a). The corresponding characterizations for discrete distributions are given by Ley and Swan (2013b).

The contribution at hand is certainly not the first application of Stein's method in statistics. Indeed, similar problems in the context of non-normalized models are tackled with the use of so-called Stein discrepancies by the machine

learning community, though many of them refer to the continuous setting. Let us mention some papers that explore these tools. Namely, Chwialkowski, Strathmann and Gretton (2016), Liu, Lee and Jordan (2016), and Yang et al. (2018) consider the construction of tests of fit, Gorham and Mackey (2015) build measures of sample quality, and Barp et al. (2019) solve estimation problems for non-normalized models.

Our new work is based, in a strict sense, not on what is generally called Stein's method, but rather on the characterization identities we refer to above. More precisely, we take as a starting point the discrete density approach identity as provided by Ley and Swan (2013b). To sketch the idea, consider a probability mass function p on $\mathbb{N}_0$ as well as an $\mathbb{N}_0$-valued random variable X. Subject to few regularity conditions, X is governed by p if, and only if,

$$\mathbb{E}\left[\Delta^+ f(X) + \frac{\Delta^+ p(X)}{p(X)} f(X+1)\right] = 0$$

holds for a large enough class of test functions f. Hereby Δ^+ denotes the forward difference operator. Our first contribution lies in proving that this characterization can essentially be restated as to X being governed by p if, and only if, the probability mass function ρ_X of X satisfies

$$\rho_X(k) = \mathbb{E}\left[-\frac{\Delta^+ p(X)}{p(X)} \mathbb{1}\{X \geq k\}\right], \quad k \in \mathbb{N}_0.$$

With regard to applications in statistics, this second identity is more accessible. We can, for instance, tackle the goodness-of-fit testing problem as follows. Assume we are to test whether a sample $X_1, \ldots, X_n$ of $\mathbb{N}_0$-valued random variables follows one of the laws of a parametric family of distributions $\{p_\vartheta : \vartheta \in \Theta\}$, where Θ denotes the parameter space. By the above characterization, if $X_1, \ldots, X_n$ are governed by one of the p_ϑ, then the difference between

$$\widehat{\rho}_n(k) = \frac{1}{n} \sum_{j=1}^{n} \mathbb{1}\{X_j = k\}$$

and

$$\frac{1}{n} \sum_{j=1}^{n} -\frac{\Delta^+ p_{\widehat{\vartheta}_n}(X_j)}{p_{\widehat{\vartheta}_n}(X_j)} \mathbb{1}\{X_j \geq k\}$$

ought to be small for each $k \in \mathbb{N}_0$. Here we denote by $\widehat{\vartheta}_n$ an estimator of ϑ based on $X_1, \ldots, X_n$. Thus, in line with the idea of characterization based goodness-of-fit testing, our proposal is to use

$$\sum_{k=0}^{\infty} \left(\frac{1}{n} \sum_{j=1}^{n} \mathbb{1}\{X_j = k\} + \frac{1}{n} \sum_{j=1}^{n} \frac{\Delta^+ p_{\widehat{\vartheta}_n}(X_j)}{p_{\widehat{\vartheta}_n}(X_j)} \mathbb{1}\{X_j \geq k\}\right)^2$$

as a test statistic for the hypothesis

$$\mathbf{H_0} \ : \ \rho_{X_1} \in \{p_\vartheta : \vartheta \in \Theta\},$$

and to reject the hypothesis for large values of the statistic. Supposing that $X_1, \ldots, X_n$ are governed by p_{ϑ_0} for some (unknown) $\vartheta_0 \in \Theta$, the very same heuristic leads us to propose

$$\widehat{\vartheta}_n = \operatorname{argmin}_{\vartheta \in \Theta} \ \sum_{k=0}^{\infty} \left(\frac{1}{n} \sum_{j=1}^{n} \mathbb{1}\{X_j = k\} + \frac{1}{n} \sum_{j=1}^{n} \frac{\Delta^+ p_\vartheta(X_j)}{p_\vartheta(X_j)} \, \mathbb{1}\{X_j \geq k\} \right)^2$$

as an estimator for the unknown ϑ_0. The paper at hand formalizes these ideas and puts them on firm mathematical ground. We also provide examples for the theoretical results as well as for the testing and estimation methods we propose. In Section 2 we introduce basic notation and recall the density approach identity. In Section 3 we prove the new characterization result as indicated above and in Section 4 we discuss examples. We then construct and study empirically, in Section 5, the test of fit for the Poisson distribution. In Section 6 a discrepancy measure as above leads to minimum distance estimators for the negative binomial distribution, which are put to a test in a simulation study. Section 7 deals with similar parameter estimators in the non-normalized class of discrete exponential-polynomial models.

2. The foundation: Stein characterizations for discrete distributions

We denote by $p : \mathbb{Z} \to [0, 1]$ a probability mass function (pmf) defined on the integers. We assume that the support of p, that is, $\operatorname{spt}(p) = \{k \in \mathbb{Z} : p(k) > 0\}$ is connected, in the sense that

(C1) $\operatorname{spt}(p) = \{L, L+1, \ldots, R\}$, where $L, R \in \mathbb{Z} \cup \{\pm\infty\}$, $L < R$.

This prerequisite is quite usual in the context of Stein's method in the discrete setting. We further assume that

(C2) $\displaystyle \sup_{k \in \{L, \ldots, R-1\}} \left| \frac{\Delta^+ p(k) \cdot \min\{P(k), 1 - P(k)\}}{p(k) \, p(k+1)} \right| < \infty,$

where $\Delta^+ f(k) = f(k+1) - f(k)$ is the forward difference operator. Moreover, we denote by $P(k) = \sum_{\ell=L}^{k} p(\ell)$ the distribution function corresponding to p. Assumption (C2) is known from the continuous setting, see Lemma 13.1 of Chen, Goldstein and Shao (2011). The supremum in (C2) runs from L to ∞ whenever $R = \infty$. In what follows, we stick to the convention of setting empty sums to 0.

Definition 2.1 *Let p be a pmf that satisfies (C1) and (C2). We write $\mathcal{F}_p$ for the class of functions $f : \{L, \ldots, R\} \to \mathbb{R}$ such that*

(a) $\displaystyle \sum_{k=L}^{R} \left| \Delta^+ \big(p(k) \, f(k) \big) \right| \ < \ \infty \ \ and \ \ \sum_{k=L}^{R} \Delta^+ \big(p(k) \, f(k) \big) \ = \ 0, \ where \ we \ put$ $f(R+1) = 0 \ if \ R < \infty, \ as \ well \ as$

(b) $\displaystyle\sup_{k\in\{L,\dots,R\}} \big|\Delta^+ f(k)\big| < \infty, \; and \; \sup_{k\in\{L,\dots,R\}} \left|\frac{\Delta^+ p(k)}{p(k)}\, f(k+1)\right| < \infty.$

Conditions (C2) and *(b)* vanish completely whenever the support of p is finite. We now state the characterization theorem that is known as Stein's density approach for discrete distributions. The proof is an easy adaptation of the proof of Theorem 2.1 from Ley and Swan (2013b) taking into account the different class of test functions. We give a full proof in Appendix A to make it possible for the reader to comprehend how the assumptions come into play. Denote by $(\Omega, \mathcal{A}, \mathbb{P})$ the probability space which underlies all random quantities in this work.

Theorem 2.2 (Discrete density approach) *Let p be a pmf which satisfies conditions (C1) and (C2), and let $X : \Omega \to \mathbb{R}$ be a random variable such that $\mathbb{P}\big(X \in \mathrm{spt}(p)\big) > 0$. Then $\mathbb{P}\big(X = k \,|\, X \in \mathrm{spt}(p)\big) = p(k)$, $k \in \mathbb{Z}$, if, and only if,*

$$\mathbb{E}\left[\Delta^+ f(X) + \frac{\Delta^+ p(X)}{p(X)}\, f(X+1) \;\middle|\; X \in \mathrm{spt}(p)\right] = 0,$$

for all $f \in \mathcal{F}_p$, where $\mathbb{E}[\,\cdot\,|\,\cdot\,]$ denotes the conditional expectation.

We use the abbreviation $X|p \sim p$ for $\mathbb{P}\big(X = k \,|\, X \in \mathrm{spt}(p)\big) = p(k)$, $k \in \mathbb{Z}$. There exists a very similar result for continuous probability distributions. This continuous version existed first and was initiated by Stein (1986). For the fully prepared statement we refer to Ley and Swan (2011) and Ley and Swan (2013a), and for further constructions of this type of Stein operators, see Ley, Reinert and Swan (2017).

Remark 2.3 *It follows from the proof of Theorem 2.2 that, if $L > -\infty$, we may assume that $f(L) = 0$ for all $f \in \mathcal{F}_p$.*

3. Distributional characterizations via the probability mass function

In this section we derive explicit distributional characterizations via the probability mass function. The whole theory can be understood as a discrete version of the results from Betsch and Ebner (2021) who established similar characterization identities for continuous probability laws starting with a continuous version of Theorem 2.2 as stated by Ley and Swan (2013a). We make the further assumption that the expectation of p exists, that is,

(C3) $\mathbb{E}|Z| < \infty$, where Z is a discrete random variable with pmf p.

It follows from (C3) that $\mathbb{E}|\Delta^+ p(Z) \cdot Z \,/\, p(Z)| < \infty$, and hence we also have $\mathbb{E}|\Delta^+ p(Z) \,/\, p(Z)| < \infty$. We note our first result, a proof of which is given in Appendix B.

Theorem 3.1 *Let p be a pmf which satisfies (C1) – (C3) with $L > -\infty$. Let $X : \Omega \to \mathbb{R}$ be a random variable with $\mathbb{P}\big(X \in \mathrm{spt}(p)\big) > 0$ as well as*

$$\mathbb{E}\left|\frac{\Delta^+ p(X)}{p(X)}\, X \cdot \mathbb{1}\big\{X \in \mathrm{spt}(p)\big\}\right| < \infty,$$

and denote by $\rho_{X|p}(k) = \mathbb{P}\big(X = k \mid X \in \mathrm{spt}(p)\big)$ *the pmf of* X *given* $X \in \mathrm{spt}(p)$. *Then* $X|p \sim p$ *if, and only if,*

$$\rho_{X|p}(k) = \mathbb{E}\bigg[-\frac{\Delta^+ p(X)}{p(X)}\, \mathbb{1}\{X \geq k\} \,\bigg|\, X \in \mathrm{spt}(p)\bigg], \quad k \in \mathbb{Z}, k \geq L.$$

Notice that the integrability assumption on X implies the existence of the (conditional) expectation that appears in the theorem. Even in stating Theorem 3.1 the ordering of the integers is essential. However, if p is an admissible probability mass function on some arbitrary countable set $\mathbb{S}$ (where $\mathbb{S}$ is endowed with the power set as a σ-field), there exists a bijection $\iota : \mathbb{S} \to \mathbb{N}_0$, which corresponds to imposing an order on the space $\mathbb{S}$, and Theorem 3.1 can be applied. This leads to the following corollary which allows the handling of more general state spaces.

Corollary 3.2 *Let* $\mathbb{S}$ *be a countable set and* $p : \mathbb{S} \to [0,1]$ *with* $\sum_{s \in \mathbb{S}} p(s) = 1$. *Let* $\iota : \mathbb{S} \to \{L, \dots, R\}$, *with* $L > -\infty$, *be a bijection so that* $\widetilde{p} = p \circ \iota^{-1}$ *satisfies (C1) – (C3). Assume that* $X : \Omega \to \mathbb{S}$ *is a random variable such that* $\mathbb{P}\big(X \in \mathrm{spt}(p)\big) > 0$, *and*

$$\mathbb{E}\bigg|\frac{(\Delta^+ \widetilde{p})\big(\iota(X)\big)}{p(X)}\, \iota(X) \cdot \mathbb{1}\big\{X \in \mathrm{spt}(p)\big\}\bigg| < \infty.$$

Then $X|p \sim p$ *if, and only if,*

$$\rho_{X|p}(k) = \mathbb{E}\bigg[-\frac{(\Delta^+ \widetilde{p})\big(\iota(X)\big)}{p(X)}\, \mathbb{1}\big\{\iota(X) \geq \iota(s)\big\} \,\bigg|\, X \in \mathrm{spt}(p)\bigg], \quad s \in \mathbb{S}.$$

Any such ordering on $\mathbb{S}$ gives a characterization result, and if $X \sim p$, the (conditional) expectation is the same for every ordering (as $\rho_{X|p}$ does not depend on ι). However, if one intends to use the converse of the characterization (with general X), the calculation of the expectation depends on the ordering, so in practice the question of choosing an efficient ordering arises. Finding an order such that the conditions (C1) – (C3) are satisfied is a non-trivial endeavor. We give one example of choosing an order such that a pmf with a support that is not bounded from below can be considered. To state the result, we first recall that $\Delta^- f(k) = f(k) - f(k-1)$ defines the backward difference operator.

Corollary 3.3 *Let* p *be a pmf on* $\{L, \dots, R\}$ *which satisfies (C1) – (C3) with* $R < \infty$. *Let* $X : \Omega \to \mathbb{R}$ *be a random variable with* $\mathbb{P}\big(X \in \mathrm{spt}(p)\big) > 0$ *and*

$$\mathbb{E}\bigg|\frac{\Delta^- p(X)}{p(X)}\, X \cdot \mathbb{1}\big\{X \in \mathrm{spt}(p)\big\}\bigg| < \infty.$$

Then $X|p \sim p$ *if, and only if,*

$$\rho_{X|p}(k) = \mathbb{E}\bigg[\frac{\Delta^- p(X)}{p(X)}\, \mathbb{1}\{X \leq k\} \,\bigg|\, X \in \mathrm{spt}(p)\bigg], \quad k \in \mathbb{Z}, k \leq R.$$

Choosing $\iota : \{L, \ldots, R\} \to \{-R, \ldots, -L\}$, $\iota(k) = -k$, the result follows from Corollary 3.2, as

$$(\Delta^+ \widetilde{p})\big(\iota(X)\big) = (\Delta^+ \widetilde{p})(-X) = p(X-1) - p(X) = -\Delta^- p(X).$$

Note that Corollary 3.3 can also be obtained via a different path. With few technical changes in Definition 2.1 and Appendix A, a Δ^--version of Theorem 2.2 can be formulated (see also Ley and Swan, 2013b). Using this result and an adaptation of the proof of Theorem 3.1 yields another proof of Corollary 3.3.

Remark 3.4 *Whenever X is assumed a priori to take values in $\{L, \ldots, R\}$, the conditioning on $X \in \mathrm{spt}(p)$ can be omitted, and when $-\infty < L < R < \infty$, the integrability condition on X is trivially satisfied. As for the regularity assumptions (C1) – (C3), notice that, by Corollary 3.2, (C1) is mostly an issue of notation. Whenever we deal with discrete distributions that have finite support, conditions (C2) and (C3) are trivially satisfied. In case of an infinite support, assumption (C3) is easy to interpret. It is stated to guarantee that the statement of Theorem 3.1 is consistent, as it ensures that a random variable $Z \sim p$ satisfies the integrability condition on X. A drawback in terms of the assumptions is that we cannot give a general treatment of (C2), and that this condition can sometimes be difficult to check for a given distribution. A similar condition (with identical problems) is required by Betsch and Ebner (2021) in the continuous setting. If $L > -\infty$ and $R = \infty$ then (C2) holds if, and only if,*

$$\limsup_{k \to \infty} \left| \frac{\Delta^+ p(k) \cdot (1 - P(k))}{p(k)\, p(k+1)} \right| < \infty. \tag{3.1}$$

Similar thoughts apply to other choices for L and R, but this does not solve the problem in general. However, the Stolz-Cesáro theorem (see Theorem 2.7.1 of Choudary and Niculescu, 2014) provides a useful tool for checking the condition in practice, see Example 4.3.

Remark 3.5 (Non-normalized models) *As explained in the introduction, many statistical models, primarily in machine learning and physics, are too complex for the normalization constant of the distribution to be calculable. As estimation and testing procedures (e.g. the maximum likelihood estimator) usually rely on some knowledge about this constant, they may not be applicable to such models. Thus, we want to emphasize that our explicit characterizations do not need any knowledge about the normalization constants, and neither do any of the statistical applications presented in subsequent sections.*

4. Examples

In this section, we provide examples that fit into our framework. For each distribution we indicate why (C1) – (C3) hold and we explicitly state the characterization via Theorem 3.1. We consider three infinite support examples which are subject of our statistical applications in the subsequent sections. More precisely,

we discuss the Poisson and the negative binomial distribution as well as a discrete version of the exponential-polynomial model. From our previous discussion in Remark 3.4 it is easy to see that many other examples can be found. For such additional examples and further characterization results via transformations of the pmf, we refer to the arXiv version of this paper.

Example 4.1 (Poisson distribution) *The mass function of the Poisson distribution is given as $p(k) = \lambda^k \, e^{-\lambda} \, / \, k!$, $k \in \mathbb{N}_0$, for some rate parameter $\lambda > 0$. In this case, we obtain*

$$\frac{\Delta^+ p(k)}{p(k)} = \frac{\lambda}{k+1} - 1, \quad k \in \mathbb{N}_0.$$

Conditions (C1) and (C3) are obviously true. To see that (C2) holds, note that whenever $\lambda \, / \, (k+2) < 1$, we have

$$\left| \frac{\Delta^+ p(k) \cdot (1 - P(k))}{p(k) \, p(k+1)} \right| = \left| \frac{\lambda}{k+1} - 1 \right| \sum_{\ell=1}^{\infty} \frac{\lambda^{\ell-1}}{(\ell + k)!} \, (k+1)!$$

$$\leq \left| \frac{\lambda}{k+1} - 1 \right| \sum_{\ell=0}^{\infty} \left(\frac{\lambda}{k+2} \right)^{\ell} = \left| \frac{\lambda}{k+1} - 1 \right| \cdot \frac{1}{1 - \frac{\lambda}{k+2}},$$

and therefore (3.1) holds which yields (C2). Theorem 3.1 implies that a random variable $X : \Omega \to \mathbb{N}_0$ with $\mathbb{E}X < \infty$ has the Poisson distribution with parameter λ if, and only if,

$$\rho_X(k) = \mathbb{E}\left[\left(1 - \frac{\lambda}{X+1} \right) \mathbb{1}\{X \geq k\} \right], \quad k \in \mathbb{N}_0.$$

Example 4.2 (Negative binomial distribution) *The probability mass function of the negative binomial distribution with parameters $r > 0$ and $q \in (0, 1)$ is $p(k) = \binom{k+r-1}{k}(1-q)^k \, q^r$, $k \in \mathbb{N}_0$. An important special case arises for $r = 1$, where the negative binomial distribution reduces to the geometric distribution. These laws are frequently used in the analysis of arrival times. We have*

$$\frac{\Delta^+ p(k)}{p(k)} = \frac{r+k}{k+1} \, (1-q) - 1, \quad k \in \mathbb{N}_0.$$

Condition (C1) is trivially satisfied, and (C3) is easily verified. We prove (3.1) to show that (C2) is satisfied. To this end, observe that

$$\left| \frac{\Delta^+ p(k) \cdot (1 - P(k))}{p(k) \, p(k+1)} \right| = \left| \frac{r+k}{k+1} \, (1-q) - 1 \right| \sum_{\ell=0}^{\infty} \frac{(\ell+k+r)! \cdot (k+1)!}{(\ell+k+1)! \cdot (k+r)!} \, (1-q)^{\ell}.$$

If $r \leq 1$, the sum is bounded by $\sum_{\ell=0}^{\infty}(1-q)^{\ell} = 1/q$. If $r > 1$, let k be large enough so that $2 \cdot (r-1) \, / \, (k+2) < q \, / \, (1-q)$, and observe that

$$\sum_{\ell=0}^{\infty} \frac{(r+k+\ell) \cdot (r+k+\ell-1) \cdots (r+k+1)}{(k+\ell+1) \cdot (k+\ell) \cdots (k+2)} \, (1-q)^{\ell}$$

$$= \sum_{\ell=0}^{\infty} \left(1 + \frac{r-1}{k+\ell+1}\right) \cdot \left(1 + \frac{r-1}{k+\ell}\right) \cdots \left(1 + \frac{r-1}{k+2}\right) \cdot (1-q)^{\ell}$$

$$\leq \sum_{\ell=0}^{\infty} \left(1 + \frac{1}{2} \cdot \frac{q}{1-q}\right)^{\ell} (1-q)^{\ell} = \sum_{\ell=0}^{\infty} \left(1 - \frac{q}{2}\right)^{\ell} = \frac{2}{q},$$

where the products in the sum are empty (hence equal to 1) for $\ell = 0$. In any case, (3.1) follows, so (C2) is valid. Theorem 3.1 states that a discrete random variable $X : \Omega \to \mathbb{N}_0$ with $\mathbb{E}X < \infty$ follows the negative binomial law with parameters r and q if, and only if,

$$\rho_X(k) = \mathbb{E}\left[\left(1 - \frac{r+X}{X+1}(1-q)\right)\mathbb{1}\{X \geq k\}\right], \quad k \in \mathbb{N}_0.$$

Note that the statement by Johnson, Kotz and Kemp (1993) (on p. 223) that "only a few characterizations have been obtained for the negative binomial distribution" appears to still hold true. For one recent characterization related to Stein's method, we refer to Arras and Houdré (2019).

Example 4.3 (Exponential-polynomial models) *We consider the following discrete exponential-polynomial parametric model given through*

$$p_{\vartheta}(k) = C(\vartheta)^{-1} \exp\left(\vartheta_1 k + \ldots + \vartheta_d k^d\right), \quad k \in \mathbb{N},$$

where

$$C(\vartheta) = \sum_{k=1}^{\infty} \exp\left(\vartheta_1 k + \ldots + \vartheta_d k^d\right),$$

and $\vartheta = (\vartheta_1, \ldots, \vartheta_d) \in \mathbb{R}^{d-1} \times (-\infty, 0)$. This corresponds to a discrete exponential family in the canonical form with the sufficient statistic containing monomials up to order $d \in \mathbb{N}$, with $d \geq 2$. Clearly condition (C1) is satisfied and the restriction $\vartheta_d < 0$ ensures that (C3) holds for every ϑ as well as that $C(\vartheta) < \infty$. We have

$$\frac{p_{\vartheta}(k+1)}{p_{\vartheta}(k)} = \exp\left(\vartheta_1 + \vartheta_2\left((k+1)^2 - k^2\right) + \ldots + \vartheta_d\left((k+1)^d - k^d\right)\right) \longrightarrow 0,$$

as $k \to \infty$, and the Stolz-Cesáro theorem (Theorem 2.7.1 of Choudary and Niculescu, 2014) yields

$$\lim_{k \to \infty} \frac{p_{\vartheta}(k+1)}{1 - P_{\vartheta}(k)} = -\lim_{k \to \infty} \frac{p_{\vartheta}(k+2)}{p_{\vartheta}(k+1)} + 1 = 1.$$

Consequently, we obtain

$$\limsup_{k \to \infty} \left|\frac{\Delta^+ p_{\vartheta}(k) \cdot (1 - P_{\vartheta}(k))}{p_{\vartheta}(k)\, p_{\vartheta}(k+1)}\right| \leq \limsup_{k \to \infty} \left|\frac{p_{\vartheta}(k+1)}{p_{\vartheta}(k)} - 1\right| \cdot \left|\frac{1 - P_{\vartheta}(k)}{p_{\vartheta}(k+1)}\right| = 1$$

for every $\vartheta \in \mathbb{R}^{d-1} \times (-\infty, 0)$, so (C2) holds. Finally observe that we have $p_\vartheta(k+1)/p_\vartheta(k) < 1$ for all but finitely many $k \in \mathbb{N}$, so an $\mathbb{N}$-valued random variable with $\mathbb{E}X < \infty$ also satisfies

$$\mathbb{E}\left| \frac{\Delta^+ p_\vartheta(X)}{p_\vartheta(X)} X \right| < \infty.$$

Theorem 3.1 yields that a random variable $X : \Omega \to \mathbb{N}$ with $\mathbb{E}X < \infty$ has the pmf p_ϑ if, and only if, $\rho_X(k)$ is given through

$$\mathbb{E}\left[\left(1 - \exp\left(\vartheta_1 + \vartheta_2\big((X+1)^2 - X^2\big) + \ldots + \vartheta_d\big((X+1)^d - X^d\big) \right) \right) \mathbb{1}\{X \geq k\} \right]$$

for each $k \in \mathbb{N}$. In Section 7 we use this characterization to construct an estimation method for this type of parametric model, focusing on a two-parameter case where $d = 3$ and $\vartheta_2 = 0$ fixed.

5. Goodness-of-fit testing for the Poisson distribution

A first application of the characterization results from the previous sections is the construction of a test of fit for the Poisson distribution. Given a sample of $\mathbb{N}_0$-valued independent identically distributed (i.i.d.) random variables, the problem is to test the composite hypothesis that the sample comes from some Poisson distribution $\text{Po}(\lambda)$ with an unknown rate parameter $\lambda > 0$, that is,

$$\mathbf{H_0} \;:\; \mathbb{P}^{X_1} \in \big\{ \text{Po}(\lambda) : \lambda > 0 \big\}.$$

This is a classical statistical problem, well studied in the literature. Apart from Pearson's χ^2 test, see Khmaladze (2013) for recent developments, the hitherto proposed tests are based on the (conditional) empirical distribution function, see Beltrán-Beltrán and O'Reilly (2019); Gürtler and Henze (2000); Henze (1996); Frey (2012), the empirical probability generating function, see Baringhaus and Henze (1992); Puig and Weiß (2020); Rueda and O'Reilly (1999), on the integrated distribution function, see Klar (1999), on a characterization by mean distance, see Székely and Rizzo (2004), on quadratic forms of score vectors, see Inglot (2019), on Charlier polynomials, see Ledwina and Wylupek (2017), on conditional probabilities ratio, see Beltrán-Beltrán and O'Reilly (2019), and on relating first- and second-order moments, see Kyriakoussis, Li and Papadopoulos (1998). For a survey of classical procedures and a comparative simulation study see Gürtler and Henze (2000). However, the construction of new and powerful methods is still of relevance: As Nikitin (2017) stated on p.4 of his contribution that "[...] one should keep in mind that any hypothesis has to be tested with several possible criteria. The point of the matter is that with absolute confidence we can only reject it, while each new test which fails to reject the null-hypothesis gradually brings the statistician closer to the perception that this hypothesis is true".

The idea of our new method is to estimate the two quantities that appear in the characterization via Theorem 3.1 as given in Example 4.1, and to compare these empirical quantities. Based on the sample $X_1, \ldots, X_n$, let

$$\widehat{e}_n(k) = \frac{1}{n} \sum_{j=1}^{n} \left(1 - \frac{\widehat{\lambda}_n}{X_j + 1} \right) \mathbb{1}\{X_j \geq k\}, \quad k \in \mathbb{N}_0,$$

be an estimator of the expectation that arises in the characterization, where $\widehat{\lambda}_n = n^{-1} \sum_{j=1}^{n} X_j$ is a consistent estimator of the rate parameter. Also consider the empirical probability mass function,

$$\widehat{\rho}_n(k) = \frac{1}{n} \sum_{j=1}^{n} \mathbb{1}\{X_j = k\}, \quad k \in \mathbb{N}_0,$$

as an estimator of ρ_{X_1}. By Theorem 3.1 (see Example 4.1), if the sample $X_1, \ldots, X_n$ comes from a Poisson distribution, the absolute difference between $\widehat{e}_n(k)$ and $\widehat{\rho}_n(k)$ ought to be small for every $k \in \mathbb{N}_0$. On the other hand, if the sample does not come from the Poisson law, we expect their absolute difference to be large. Based on this heuristic, we suggest to use as a test statistic the squared difference of $\widehat{e}_n$ and $\widehat{\rho}_n$ summed over $k \in \mathbb{N}_0$, that is,

$$T_n^{Po} = \sum_{k=0}^{\infty} \left(\widehat{e}_n(k) - \widehat{\rho}_n(k) \right)^2,$$

and to reject the Poisson hypothesis $\mathbf{H_0}$ for large values of T_n^{Po}. Note that we do not need to introduce any weight functions to make the infinite sum in the definition of T_n^{Po} converge, and observe that we choose the squared distance to obtain a finite double sum representation for T_n^{Po}, namely

$$T_n^{Po} = \frac{1}{n^2} \sum_{j, \ell=1}^{n} \left[\left(1 - \frac{\widehat{\lambda}_n}{X_j + 1} \right) \left(X_\ell - 1 - \widehat{\lambda}_n \right) \mathbb{1}\{X_j \geq X_\ell\} \right.$$
$$\left. + \left(X_j + 1 - \widehat{\lambda}_n \right) \left(1 - \frac{\widehat{\lambda}_n}{X_\ell + 1} \right) \mathbb{1}\{X_j < X_\ell\} + \mathbb{1}\{X_j = X_\ell\} \right],$$

which is easily implemented in a computer. The calculation of T_n^{Po} involves only straight forward algebra and consists, mainly, of writing the squared difference of $\widehat{e}_n(k)$ and $\widehat{\rho}_n(k)$ as a double sum, multiplying the corresponding terms, and solving the sum over k of the indicator functions.

As a proof of concept, we carry out a simulation study in order to compare our new test of poissonity with established procedures. All simulations are performed in the statistical computing environment R, see R Core Team (2020). We consider the sample size $n = 50$ and the nominal level of significance is set to 0.05. Based on the methodology for asymptotic theory detailed by Henze (1996), we expect the (limit) distribution of the test statistics considered in the following to depend on the unknown parameter λ. Consequently, we use for the implementation of

the tests a parametric bootstrap procedure similar to the one suggested by Gürtler and Henze (2000). For a given sample $X_1, \ldots, X_n$ and a statistic T_n, simulate an approximate critical value $c_{n,B}$ for a level α test procedure as follows:

1) Calculate $\widehat{\lambda}_n(X_1, \ldots, X_n)$ and generate B bootstrap samples of size n with distribution $\mathrm{Po}(\widehat{\lambda}_n)$, i.e., generate i.i.d. $\mathrm{Po}(\widehat{\lambda}_n)$ random variables $X_{j,1}^*, \ldots, X_{j,n}^*$, $j = 1, \ldots, B$.
2) Compute $T_{j,n}^* = T_n(X_{j,1}^*, \ldots, X_{j,n}^*)$ for $j = 1, \ldots, B$.
3) Derive the order statistics $T_{1:B}^* \leq \ldots \leq T_{B:B}^*$ of $T_{1,n}^*, \ldots, T_{B,n}^*$ and put

$$c_{n,B} = T_{k:B}^* + (1 - \alpha) \cdot \left(T_{(k+1):B}^* - T_{k:B}^* \right),$$

where $k = \lfloor (1 - \alpha) \cdot B \rfloor$ and $\lfloor \cdot \rfloor$ denotes the floor function.
4) Reject the hypothesis $\mathbf{H_0}$ if $T_n(X_1, \ldots, X_n) > c_{n,B}$.

This parametric bootstrap procedure was used for all of the following procedures to generate the critical points. We consider the test of Baringhaus and Henze (1992) based on the statistic

$$\mathrm{BH} = \frac{1}{n} \sum_{i,j=1}^{n} \left(\frac{\widehat{\lambda}_n^2}{X_i + X_j + 1} + \frac{X_i X_j}{X_i + X_j - 1} \right) - \widehat{\lambda}_n \left(n - \frac{1}{n} \left(\sum_{j=1}^{n} \mathbb{1}\{X_j = 0\} \right)^2 \right).$$

The mean distance test by Székely and Rizzo (2004) is based on

$$\mathrm{SR} = n \sum_{j=0}^{\infty} \left(\widehat{M}_n(j) - \mathrm{P}(j; \widehat{\lambda}_n) \right)^2 \mathrm{p}(j; \widehat{\lambda}_n),$$

where $\widehat{M}_n(j)$ is an estimator of the CDF based on the mean distance and $\mathrm{P}(j; \lambda)$ (resp. $\mathrm{p}(j; \lambda)$) denotes the distribution function (resp. pmf) of $\mathrm{Po}(\lambda)$. Note that SR is implemented in the R-package energy, see Rizzo and Székely (2019). The test of Rueda, O'Reilly and Pérez-Abreu (1991) is based on

$$\begin{aligned} \mathrm{RU} = {}& \frac{1}{n} \sum_{i,j=1}^{n} \frac{1}{X_i + X_j + 1} + \frac{n \cdot (1 - e^{-2\widehat{\lambda}_n})}{2\widehat{\lambda}_n} \\ & - 2 \sum_{i=1}^{n} \left(\frac{(-1)^{X_i} \cdot X_i! \cdot (1 - e^{-\widehat{\lambda}_n})}{\widehat{\lambda}_n^{X_i+1}} + \sum_{j=1}^{X_i} \frac{(-1)^{j+1} \cdot X_i!}{(X_i - j + 1)! \cdot \widehat{\lambda}_n^j} \right). \end{aligned}$$

Note that in the original paper of Rueda, O'Reilly and Pérez-Abreu (1991) and in a slight handwritten correction thereof available on the internet, as well as in the work of Gürtler and Henze (2000), the explicit formula of the RU-statistic contains errors. We have corrected and numerically checked the formula given above against the integral representation used to introduce the test. The integrated distribution function based tests of Klar (1999) are defined via

$$\mathrm{K}_1 = \sqrt{n} \left(\sum_{j=0}^{M} \left| \widehat{F}_n(j) - \mathrm{P}(j; \widehat{\lambda}_n) \right| + \widehat{\lambda}_n - \sum_{j=0}^{M} \left(1 - \mathrm{P}(j; \widehat{\lambda}_n) \right) \right)$$

and

$$\mathrm{K}_2 = \sqrt{n} \sup_{1 \le k \le M} \left| \sum_{j=0}^{k-1} \left(\widehat{F}_n(j) - \mathrm{P}(j; \widehat{\lambda}_n) \right) \right|,$$

where $M = \max\{X_1, \ldots, X_n\}$ and $\widehat{F}_n$ is the empirical distribution function of $X_1, \ldots, X_n$. For representations of the Kolmogorov-Smirnov statistic and the modified Cramér-von Mises statistic, we follow the representation given by Gürtler and Henze (2000), namely

$$\mathrm{KS} = \sqrt{n} \sup_{0 \le k \le M} \left| \widehat{F}_n(x) - \mathrm{P}(k; \widehat{\lambda}_n) \right|$$

and

$$\mathrm{CM} = n \sum_{j=0}^{M} \left(\widehat{F}_n(j) - \mathrm{P}(j; \widehat{\lambda}_n) \right)^2 \cdot \frac{1}{n} \sum_{k=1}^{n} \mathbb{1}\{X_k = j\}.$$

The simulation study consists of the following 45 representatives of families of distributions. In order to show that all the considered testing procedures maintain the nominal level α of 5%, we consider the $\mathrm{Po}(\lambda)$ distribution with $\lambda \in \{1, 5, 10, 30\}$. As examples for alternative distributions, we consider the discrete uniform distribution $\mathcal{U}\{0, 1, \ldots, m\}$ with $m \in \{1, 2, 3, 5, 6\}$, several different instances of the binomial distribution $\mathrm{Bin}(m, q)$, several Poisson mixtures of the form $\mathrm{PP}(q; \vartheta_1, \vartheta_2) = q \cdot \mathrm{Po}(\vartheta_1) + (1 - q) \cdot \mathrm{Po}(\vartheta_2)$, a $0.9/0.1$ mixture of $\mathrm{Po}(3)$ and point mass in 0 denoted by $\mathrm{Po}(3)\delta_0$, discrete Weibull distributions $\mathrm{W}(\vartheta_1, \vartheta_2)$, zero-modified Poisson distributions $\mathrm{zmPo}(\lambda, q)$, the zero-truncated Poisson distributions $\mathrm{ztPo}(\lambda)$ with $\lambda \in \{2, 3, 5\}$, and the absolute discrete normal distribution $|\mathrm{N}(\mu, 1)|$ with $\mu \in \{0, 2, 3\}$. Note that most distributions were generated by the packages `extraDistr`, see Wolodzko (2019), and `actuar`, see Dutang, Goulet and Pigeon (2008), and that a significant part of these distributions can also be found in the simulation study presented by Gürtler and Henze (2000). Furthermore we indicate that the chosen design of simulation parameters coincides with the study by Gürtler and Henze (2000) which facilitates the comparison to other tests of poissonity not considered here.

Every entry in Table 1 is based on 100000 repetitions and 500 bootstrap samples of size 50. All of the considered procedures maintain the significance level $\alpha = 5\%$ under the hypothesis, which supports the statement that the parametric bootstrap procedure is well calibrated. Overall the best performing tests are K_2, BH and SR. The new test based on T_n^{Po} is competitive to the stated procedures, although it never outperforms them all at once for the considered alternatives.

6. Parameter estimation in the family of negative binomial distributions

The characterizations we employ contain information about the underlying probability law and lead to empirical discrepancy measures being close to zero

TABLE 1

Empirical rejection rates of the tests of poissonity (sample size 50, significance level 5%).

Distr. / Test	T_n^{Po}	BH	SR	RU	K_1	K_2	KS	CM
Po(1)	5	5	5	5	5	5	5	5
Po(5)	5	5	5	5	5	5	5	5
Po(10)	5	5	5	5	5	5	5	5
Po(30)	5	5	5	5	5	5	5	5
$\mathcal{U}\{0,1\}$	99	99	99	99	99	99	99	99
$\mathcal{U}\{0,1,2\}$	39	9	22	15	64	68	50	58
$\mathcal{U}\{0,1,2,3\}$	46	33	20	27	61	16	45	51
$\mathcal{U}\{0,1,2,3,4,5\}$	69	65	58	62	75	39	60	63
$\mathcal{U}\{0,1,2,3,4,5,6\}$	85	85	83	85	86	66	72	76
Bin(2, 0.5)	81	81	89	87	86	90	83	81
Bin(4, 0.25)	18	22	23	24	18	22	21	15
Bin(10, 0.1)	7	7	7	7	6	7	7	6
Bin(10, 0.5)	57	52	49	52	82	88	60	68
Bin(1, 0.5)	73	77	82	81	80	82	76	77
Bin(2, 2/3)	34	38	44	44	43	45	37	39
Bin(3, 0.75)	19	22	26	26	26	27	21	23
Bin(9, 0.9)	6	7	8	8	8	8	7	8
Bin(5, 0.5)	82	85	80	84	88	89	67	71
Bin(10, 2/3)	41	45	41	44	48	50	28	31
Bin(15, 0.75)	23	27	26	26	28	30	16	18
Bin(45, 0.9)	8	9	10	9	9	8	6	7
PP(0.5; 2, 5)	64	64	69	65	72	74	53	57
PP(0.5; 3, 5)	17	19	20	19	20	21	12	14
PP(0.25; 1, 5)	93	95	96	95	87	88	75	73
PP(0.05; 1, 5)	23	33	32	32	13	12	8	7
PP(0.01; 1, 5)	7	9	9	9	6	5	5	5
$Po(3)\delta_0$	54	62	54	59	32	31	32	26
W(0.5, 1)	73	77	82	81	80	82	76	77
W(0.25, 1)	22	24	27	28	26	26	26	25
W(0.5, 2)	49	52	52	51	48	52	51	52
W(0.25, 2)	8	8	7	6	6	8	7	10
W(0.75, 2)	28	32	35	35	26	32	30	21
W(0.1, 1)	10	10	10	8	10	10	10	10
W(0.9, 3)	97	97	99	99	98	99	97	93
zmPo(1, 0.1)	91	93	90	92	81	84	90	64
zmPo(1, 0.5)	17	18	19	19	19	18	18	19
zmPo(1, 0.8)	71	72	74	74	74	74	74	74
zmPo(2, 0.1)	8	9	8	8	6	6	6	6
zmPo(3, 0.1)	24	30	24	27	13	12	11	9
ztPo(2)	93	99	83	95	38	39	56	19
ztPo(3)	12	18	18	18	9	10	7	9
ztPo(5)	4	1	1	1	5	5	5	5
$|N(0,1)|$	45	48	48	50	42	46	47	41
$|N(2,1)|$	44	46	59	53	54	61	42	37
$|N(3,1)|$	88	78	94	86	96	98	85	90

if the distribution generating the data is the one stated in the characterization. These measures can be used for estimation of the parameters of the considered parametric family of distributions. To illustrate this point, we propose a minimum distance estimation procedure for the family of negative binomial

distributions. Our objective is to estimate the unknown parameters $q \in (0,1)$ and $r > 0$ of a negative binomial distribution based on an i.i.d. $\mathbb{N}_0$-valued sample $X_1, \ldots, X_n$. Estimation in this particular family is not trivial, since Aragón, Eberly and Eberly (1992) have shown the conjecture of Anscombe dating back to 1950, that the maximum likelihood equations have a unique solution if, and only if, $\overline{X}_n < S_n^2$ with $\overline{X}_n = n^{-1} \sum_{j=1}^n X_j$, the sample mean, and $S_n^2 = n^{-1} \sum_{j=1}^n (X_j - \overline{X}_n)^2$, the sample variance. However, as Johnson, Kotz and Kemp (1993) state in their Section 8.3, so called "[...] underdispersed samples [...] will occasionally be encountered, even when a negative binomial model is appropriate." The moment estimators defined by $\widetilde{q}_n = \overline{X}_n / S_n^2$ and

$$\widetilde{r}_n = \frac{\left(\overline{X}_n\right)^2}{(1 - \widehat{q}_n) S_n^2} = \frac{\left(\overline{X}_n\right)^2}{S_n^2 - \overline{X}_n},$$

see display (5.49) and (5.50) of Johnson, Kotz and Kemp (1993), perform comparably bad as the maximum likelihood estimators, since, in underdispersed samples, they lead to negative values of $\widetilde{r}_n$ or values of $\widetilde{q}_n$ that are greater than one, see the following simulation study and Figure 2.

The heuristic for our new method is similar to that of the previous section, again based on Theorem 3.1 (see also Example 4.2). Thus, we define

$$\widehat{e}_n(k; r, q) = \frac{1}{n} \sum_{j=1}^n \left(1 - \frac{r + X_j}{X_j + 1}(1 - q)\right) \mathbb{1}\{X_j \geq k\}, \quad k \in \mathbb{N}_0,$$

and let $\widehat{\rho}_n$ be as in the previous section. Similar to the test for the Poisson distribution, we consider the empirical discrepancy measure

$$S_n^{NB}(r, q) = \sum_{k=0}^{\infty} \left(\widehat{e}_n(k; r, q) - \widehat{\rho}_n(k)\right)^2$$

which is given explicitly by the double sum

$$\frac{1}{n^2} \sum_{j,\ell=1}^n \left[\left(1 - \frac{r + X_j}{X_j + 1}(1 - q)\right)\left(q(r + X_\ell) - r - 1\right) \mathbb{1}\{X_j \geq X_\ell\}\right.$$

$$\left. + \left(q(r + X_j) - r + 1\right)\left(1 - \frac{r + X_\ell}{X_\ell + 1}(1 - q)\right) \mathbb{1}\{X_j < X_\ell\} + \mathbb{1}\{X_j = X_\ell\}\right].$$

The proposed estimators for (r, q) are defined by

$$(\widehat{r}_n, \widehat{q}_n) = \mathrm{argmin}_{(r,q)} S_n^{NB}(r, q). \tag{6.1}$$

In this particular example we expect that it is possible to minimize the quadratic equation in r and q explicitly to obtain formulae for the estimators. However, in the following section we cannot hope for an explicit solution to the optimization problem and for reasons of consistency of the presentation, we use a numerical routine to find the values of the estimators in both cases. Note that similar
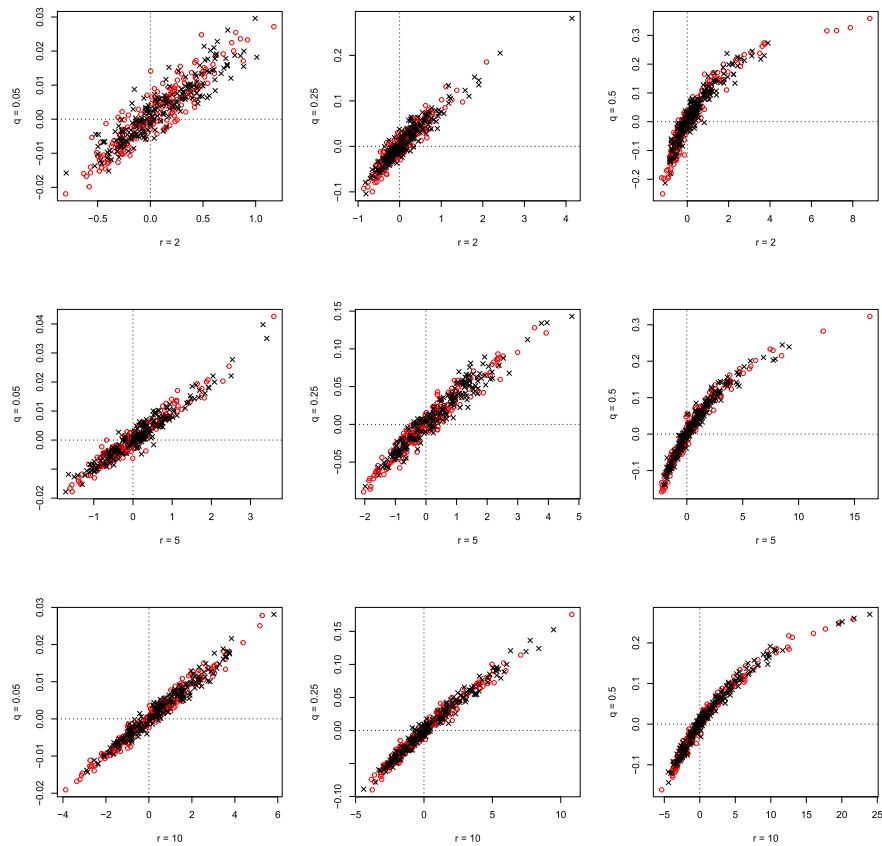
Fig 1. *Biases of the minimum distance (red) and method of moments (black) estimation procedures simulated for different parameters, $(r_0, q_0) \in \{2, 5, 10\} \times \{0.05, 0.25, 0.5\}$, of the negative binomial distribution. The sample size is $n = 100$ and 200 realizations are simulated. A red circle represents a value of $(\widehat{r}_n - r_0, \widehat{q}_n - q_0)$, while a black cross stands for a value of $(\widetilde{r}_n - r_0, \widetilde{q}_n - q_0)$.*

estimators for parametric families of continuous distributions are investigated by Betsch, Ebner and Klar (2021).

For a comparison of the two presented methods we conduct a simulation study in R and use the optim routine to find the minimal values in (6.1). The option method was fixed to L-BFGS-B, thus choosing an implementation of the routine suggested by Byrd et al. (1995), and the maximum number of iterations to maxit=1000. As starting values for the optimization routine we choose independent uniformly distributed random numbers $r_s \sim \mathcal{U}(1, 3)$ and $q_s \sim \mathcal{U}(0.1, 0.9)$. For different $(r_0, q_0) \in (0, \infty) \times (0, 1)$ we simulate 200 i.i.d. samples of size $n = 100$ from a negative binomial distribution with parameters (r_0, q_0) and calculate the minimum distance estimators $(\widehat{r}_n, \widehat{q}_n)$ as well as the moment estimators $(\widetilde{r}_n, \widetilde{q}_n)$. Then the bias of the estimation is derived by subtracting the underlying 'true' parameters (r_0, q_0). In Figures 1 and 2 the results
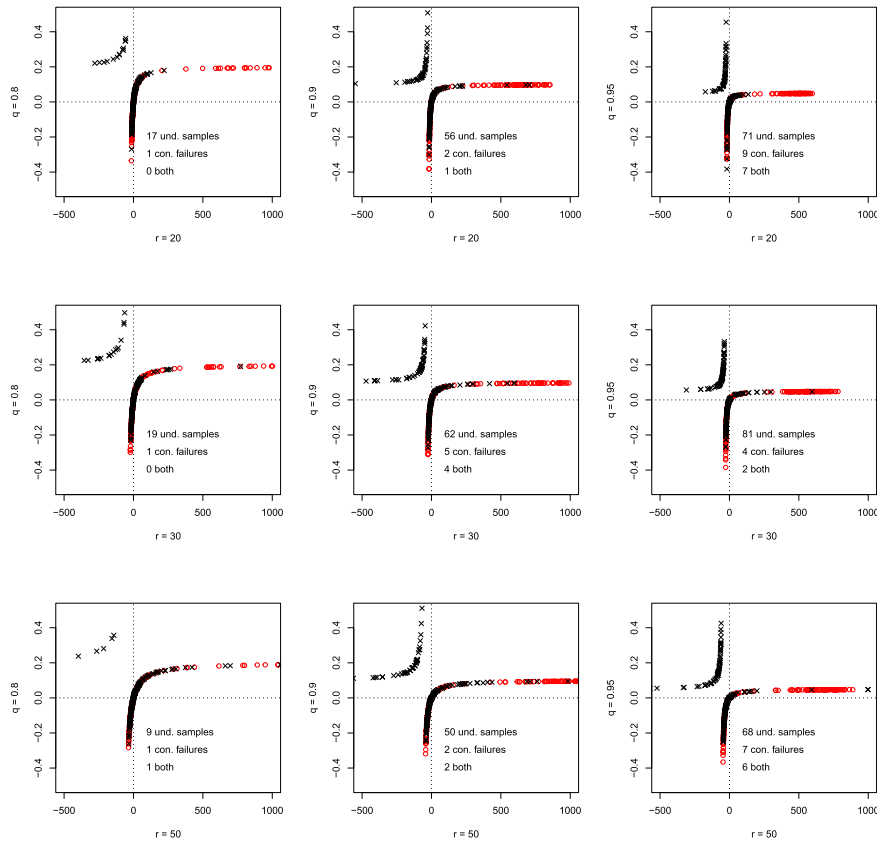
FIG 2. *Biases of the minimum distance (red) and method of moments (black) estimation procedures simulated for different parameters, $(r_0, q_0) \in \{20, 30, 50\} \times \{0.8, 0.9, 0.95\}$, of the negative binomial distribution. The sample size is $n = 100$ and 200 realizations are simulated. A red circle represents a value of $(\widehat{r}_n - r_0, \widehat{q}_n - q_0)$, while a black cross stands for a value of $(\widetilde{r}_n - r_0, \widetilde{q}_n - q_0)$. The number of cases of underdispersed samples, the number of convergence failures of the optimization routine, and the number of cases where both occurred are stated in the plot.*

of the different simulations are plotted, with estimation results of the moment estimators plotted as black crosses and the results of the minimum distance estimators as red circles. It is visible in Figure 1 that for small values of q_0, both procedures perform comparably, although the values of the moments estimators seem to scatter a little more than those of the minimum distance estimators. A completely different picture is seen in Figure 2, where values of q_0 in the neighborhood of 1 and greater values of r_0 are assumed. The moment estimators $(\widetilde{r}_n, \widetilde{q}_n)$ regularly produce values which are clearly outside of the defined parameter space $(0, \infty) \times (0, 1)$ as opposed to the minimum distance estimators $(\widehat{r}_n, \widehat{q}_n)$ which do not show this behavior due to the optimization constraints. Nevertheless, some convergence failures in the optimization routine did occur and they are not exclusively related to the underdispersed samples and only

happen for somewhat extreme parameter configurations. We chose to visually assess the quality of estimation, since empirical versions of the bias and mean squared error are very sensitive to big discrepancies, and hence did not provide valuable information on the quality of the estimation procedures. It would be of interest to find theoretical statements for the estimators $(\widehat{r}_n, \widehat{q}_n)$ such as consistency results or a central limit theorem type asymptotic distribution.

7. Parameter estimation in discrete exponential-polynomial models

In this final section we present an application to a non-normalized model, namely parameter estimation in the discrete exponential-polynomial models introduced in Example 4.3. We follow Betsch, Ebner and Klar (2021) who apply the continuous version of our estimation method to continuous exponential-polynomial models. In their work, they compare the method with two other methods for parameter estimation in non-normalized continuous models. More specifically, they implemented the score matching approach of Hyvärinen (2007) as well as noise-contrastive estimators from Gutmann and Hyvärinen (2012). As another contribution that focuses on the continuous exponential-polynomial distribution and the corresponding parameter estimation problem, let us mention Hayakawa and Takemura (2016). In our search through the literature we have found only few methods for the parameter estimation in the discrete version of the model. As such, contrastive divergence methods based on the initial proposal by Hinton (2002) can be applied in principle though it does not avoid dealing with the normalization constant $C(\vartheta)$ and Lyu (2009) proposes a discrete version of the score matching approach but does not give details on its implementation. More recently Takenouchi and Kanamori (2017) proposed a method (that avoids any calculation or approximation of the normalization constant) based on suitable homogeneous divergences which are empirically localized. We use this latter method as a comparison to our approach.

Assume that $X_1, \ldots, X_n$ is an i.i.d. $\mathbb{N}$-valued sample from the exponential-polynomial model $p_{\vartheta^{(0)}}$ in Example 4.3 with some unknown parameter vector $\vartheta^{(0)} \in \mathbb{R}^{d-1} \times (-\infty, 0)$ (with $d \in \mathbb{N}$, $d \geq 2$, fixed and known). We seek to estimate $\vartheta^{(0)}$ based on $X_1, \ldots, X_n$. Very similar to the previous section, we consider $\widehat{\rho}_n$ as before, and put

$$\widehat{e}_n(k; \vartheta) = \frac{1}{n} \sum_{j=1}^{n} \left(1 - \exp\left(\sum_{m=1}^{d} \vartheta_m \big((X_j + 1)^m - X_j^m \big) \right) \right) \mathbb{1}\{X_j \geq k\}, \quad k \in \mathbb{N}.$$

We define the empirical discrepancy measure

$$S_n^{PE}(\vartheta) = \sum_{k=0}^{\infty} \big(\widehat{e}_n(k; \vartheta) - \widehat{\rho}_n(k) \big)^2.$$

In line with Theorem 3.1, or more precisely, Example 4.3, we propose as an estimator

$$\widehat{\vartheta}_n = \operatorname{argmin}_{\vartheta} S_n^{PE}(\vartheta). \tag{7.1}$$

To see if this approach leads to sensible estimators, we conduct simulations in a two-parameter special case of the model. Following the continuous-case simulation setting of Betsch, Ebner and Klar (2021), we consider $d = 3$ but fix $\vartheta_2 = 0$, thus effectively estimating the parameters of the parametric family given through

$$p_{(\vartheta_1, \vartheta_3)}(k) = C(\vartheta_1, \vartheta_3)^{-1} \exp\left(\vartheta_1 k + \vartheta_3 k^3\right), \quad k \in \mathbb{N}, \quad \vartheta_1 \in \mathbb{R}, \ \vartheta_3 < 0, \quad (7.2)$$

which, though simpler than the general case, is still a non-normalized model and thus inaccessible to explicit maximum likelihood estimation. The discrepancy measure $S_n^{PE}(\vartheta)$ can be calculated as

$$S_n^{PE}(\vartheta) = \frac{1}{n^2} \sum_{j,\ell=1}^{n} \Big[\big(E_j(\vartheta_1, \vartheta_3) - 1\big)\big(E_\ell(\vartheta_1, \vartheta_3) \cdot X_\ell - X_\ell + 2\big) \mathbb{1}\{X_j \geq X_\ell\}$$
$$+ \mathbb{1}\{X_j = X_\ell\} + \big(E_j(\vartheta_1, \vartheta_3) - 1\big)\big(E_\ell(\vartheta_1, \vartheta_3) - 1\big) \cdot X_j \, \mathbb{1}\{X_j < X_\ell\}\Big],$$

where

$$E_i(\vartheta_1, \vartheta_3) = \exp\left(\vartheta_1 + \vartheta_3 + 3\vartheta_3 X_i + 3\vartheta_3 X_i^2\right), \quad i = 1, \ldots, n.$$

For a comparison, we consider the estimator proposed by Takenouchi and Kanamori (2017). For positive constants $\alpha, \alpha', \gamma > 0$, such that $\alpha > \alpha'$, and $\bar{\alpha} = (\alpha + \gamma\alpha') / (1 + \gamma)$, their estimator is given as

$$\widetilde{\vartheta}_n = \operatorname{argmin}_\vartheta \Bigg\{ \frac{1}{1+\gamma} \log \left(\sum_{k \in \mathcal{Z}} \left(\frac{n_k}{n}\right)^\alpha q_\vartheta(k)^{1-\alpha} \right) - \log \left(\sum_{k \in \mathcal{Z}} \left(\frac{n_k}{n}\right)^{\bar{\alpha}} q_\vartheta(k)^{1-\bar{\alpha}} \right)$$
$$+ \frac{\gamma}{1+\gamma} \log \left(\sum_{k \in \mathcal{Z}} \left(\frac{n_k}{n}\right)^{\alpha'} q_\vartheta(k)^{1-\alpha'} \right) \Bigg\}, \quad (7.3)$$

where $\mathcal{Z}$ is the set of all values that appear in the sample $X_1, \ldots, X_n$, the variable n_k denotes how often the value k is found in the sample, and

$$q_\vartheta(k) = \exp\left(\vartheta_1 k + \ldots + \vartheta_d k^d\right), \quad k \in \mathbb{N}.$$

Since Takenouchi and Kanamori (2017) do not propose a specific way of choosing the constants α, α' and γ, we use the values that appear most frequently in their simulation study and therefore set $\alpha = 1.1$, $\alpha' = 0.1$ and $\gamma = 1/9$.

As in the previous section, we use the software R for the simulation and the `optim` routine to find the minimal values in (7.1) and (7.3). Again, the option `method` is fixed to `L-BFGS-B` and the maximum number of iterations to `maxit=1000`. As starting values for the optimization we choose independent uniformly distributed random numbers $\vartheta_1^{(s)} \sim \mathcal{U}(-1,1)$ and $\vartheta_3^{(s)} \sim \mathcal{U}(-1,0)$. For different $\left(\vartheta_1^{(0)}, \vartheta_3^{(0)}\right) \in \mathbb{R} \times (-\infty, 0)$ we simulate 200 i.i.d. samples of size $n = 100$ from the discrete exponential-polynomial model in (7.2) with parameters $(\vartheta_1^{(0)}, \vartheta_3^{(0)})$ and calculate the estimators $\left(\widehat{\vartheta}_{n,1}, \widehat{\vartheta}_{n,3}\right)$ and $\left(\widetilde{\vartheta}_{n,1}, \widetilde{\vartheta}_{n,3}\right)$ presented in (7.1) and (7.3) respectively. The biases of the estimators are given by
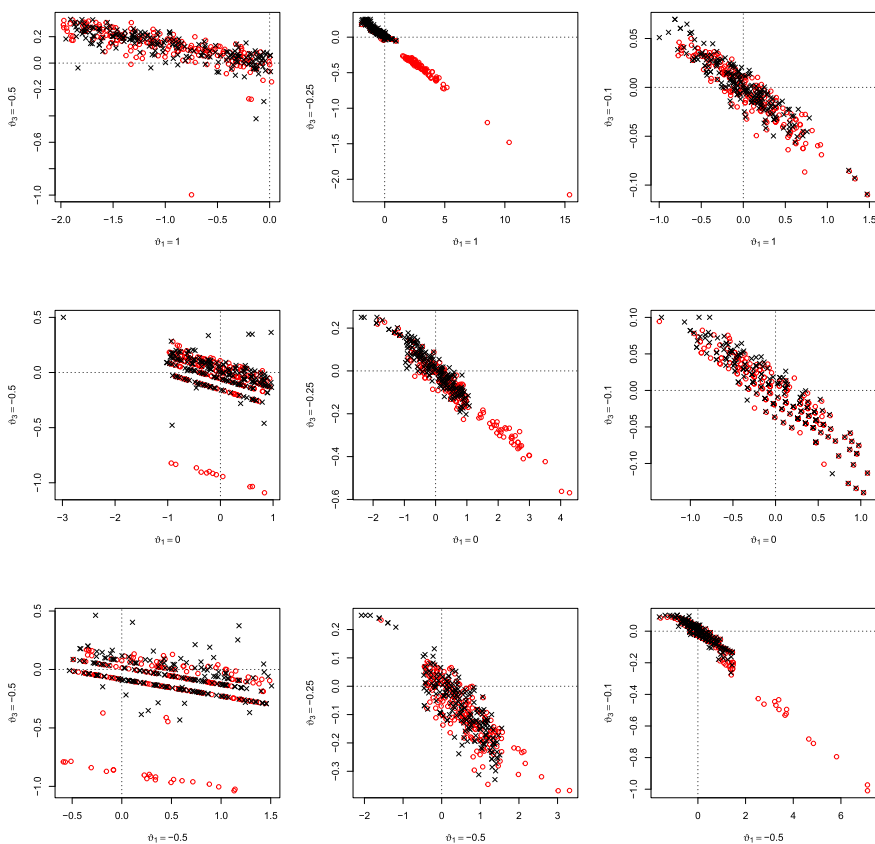
FIG 3. *Simulated biases of the estimators $\widehat{\vartheta}_n$ (red) and $\widetilde{\vartheta}_n$ (black) in the discrete exponential-polynomial model for different parameters $\left(\vartheta_1^{(0)}, \vartheta_3^{(0)}\right)$ (n = 100; 200 repetitions). A red circle represents the value of $\left(\widehat{\vartheta}_{n,1} - \vartheta_1^{(0)}, \widehat{\vartheta}_{n,3} - \vartheta_3^{(0)}\right)$, while a black cross stands for $\left(\widetilde{\vartheta}_{n,1} - \vartheta_1^{(0)}, \widetilde{\vartheta}_{n,3} - \vartheta_3^{(0)}\right)$.*

subtracting the underlying 'true' parameters $\left(\vartheta_1^{(0)}, \vartheta_3^{(0)}\right)$. The simulation of a discrete exponential-polynomial model is rather simple as $\vartheta_3^{(0)} < 0$ ensures that the probability $p_{\left(\vartheta_1^{(0)}, \vartheta_3^{(0)}\right)}(k)$ is rapidly decreasing as k grows. From a practical point of view and minding the usual calculation accuracy, we only need to deal with a discrete distribution with finite support. In Figure 3 the results of the simulations are presented and it is visible that both procedures perform comparable and overall well. The newly proposed estimators tend to scatter more which favors the competing estimators. However, our new estimators require no (data-dependent or quick fix) choice of parameters (like α, α' and γ for the estimators of Takenouchi and Kanamori, 2017). Introducing additional parameters, for instance through suitable weight functions, is also conceivable for our method. It would certainly allow for some choice which improves the overall performance, but it also leads to a less intuitive implementation as these pa-

rameters need to be chosen in practice. Note that, as in the previous simulation, some convergence failures in the optimization routine occurred (less than ten percent per parameter configuration).

Appendix A: Proof of Theorem 2.2

The following proof is, up to technical details involving the class of test functions, due to Ley and Swan (2013b) and given here for the reader's convenience.

Assume that $X|p \sim p$. Then, for $f \in \mathcal{F}_p$,

$$\mathbb{E}\left[\Delta^+ f(X) + \frac{\Delta^+ p(X)}{p(X)} f(X+1) \; \middle| \; X \in \mathrm{spt}(p)\right]$$

$$= \sum_{k=L}^{R} \left(p(k)\,\Delta^+ f(k) + f(k+1)\,\Delta^+ p(k)\right) = \sum_{k=L}^{R} \Delta^+\left(p(k)\,f(k)\right) = 0,$$

using assumption (a). To prove the converse, take a discrete random variable Z with mass function p, independent of X. For $m \in \mathbb{Z}$, define $f_m : \{L, \ldots, R\} \to \mathbb{R}$ via

$$f_m(k) = \frac{1}{p(k)} \sum_{\ell=L}^{k-1} \left(\mathbb{1}\{\ell \leq m\} - \mathbb{P}(Z \leq m)\right) p(\ell).$$

This map satisfies

$$\Delta^+\left(p(k)\,f_m(k)\right) = \left(\mathbb{1}\{k \leq m\} - \mathbb{P}(Z \leq m)\right) p(k), \quad k \in \{L, \ldots, R\}.$$

Therefore,

$$\sum_{k=L}^{R} \left|\Delta^+\left(p(k)\,f_m(k)\right)\right| \leq 2 \sum_{k=L}^{R} p(k) = 2 < \infty,$$

as well as

$$\sum_{k=L}^{R} \Delta^+\left(p(k)\,f_m(k)\right) = \sum_{k=L}^{R} \left(\mathbb{1}\{k \leq m\} - \mathbb{P}(Z \leq m)\right) p(k)$$

$$= \mathbb{P}(Z \leq m) - \mathbb{P}(Z \leq m) = 0.$$

Moreover, we have $f_m(k+1) \leq 2 \cdot P(k)\,/\,p(k+1)$ and, since

$$\sum_{\ell=L}^{R} \left(\mathbb{1}\{\ell \leq m\} - \mathbb{P}(Z \leq m)\right) p(\ell) = 0,$$

we get

$$\left|f_m(k+1)\right| = \frac{1}{p(k+1)} \left| \sum_{\ell=k+1}^{R} \left(\mathbb{1}\{\ell \leq m\} - \mathbb{P}(Z \leq m)\right) p(\ell) \right|$$

$$\leq \frac{2}{p(k+1)} \cdot \big(1 - P(k)\big),$$

which implies

$$\sup_{k \in \{L,\dots,R\}} \left| \frac{\Delta^+ p(k)}{p(k)} f_m(k+1) \right| \leq 2 \sup_{k \in \{L,\dots,R-1\}} \left| \frac{\Delta^+ p(k) \cdot \min\{P(k), 1 - P(k)\}}{p(k)\, p(k+1)} \right|$$

$$< \infty.$$

Now, notice that

$$\begin{aligned}
\Delta^+ f_m(k) &= \left(\frac{1}{p(k+1)} - \frac{1}{p(k)} \right) \sum_{\ell=L}^{k} \big(\mathbb{1}\{\ell \leq m\} - \mathbb{P}(Z \leq m)\big)\, p(\ell) \\
&\quad + \frac{1}{p(k)} \big(\mathbb{1}\{k \leq m\} - \mathbb{P}(Z \leq m)\big)\, p(k) \\
&= \left(1 - \frac{p(k+1)}{p(k)} \right) \frac{1}{p(k+1)} \sum_{\ell=L}^{k} \big(\mathbb{1}\{\ell \leq m\} - \mathbb{P}(Z \leq m)\big)\, p(\ell) \\
&\quad + \mathbb{1}\{k \leq m\} - \mathbb{P}(Z \leq m) \\
&= -\frac{\Delta^+ p(k)}{p(k)} f_m(k+1) + \mathbb{1}\{k \leq m\} - \mathbb{P}(Z \leq m),
\end{aligned}$$

where the calculation is valid for $k \in \{L, \dots, R-1\}$, but the equality obviously also holds for $k = R$ [using our convention $f_m(R+1) = 0$] if $R < \infty$. From this relation, we immediately get that

$$\sup_{k \in \{L,\dots,R\}} \big|\Delta^+ f_m(k)\big| < \infty,$$

so $f_m \in \mathcal{F}_p$, as well as, by the assumption in the converse implication,

$$\begin{aligned}
0 &= \mathbb{E}\left[\Delta^+ f_m(X) + \frac{\Delta^+ p(X)}{p(X)} f_m(X+1) \,\Big|\, X \in \mathrm{spt}(p) \right] \\
&= \mathbb{E}\left[\mathbb{1}\{X \leq m\} - \mathbb{P}(Z \leq m) \,\Big|\, X \in \mathrm{spt}(p) \right],
\end{aligned}$$

which implies the claim.

Appendix B: Proof of Theorem 3.1

First assume that $X|p \sim p$. Then, for all $k \in \mathbb{Z}$, $k \geq L$, we have

$$\begin{aligned}
\rho_{X|p}(k) = p(k) &= -\sum_{\ell=k}^{R} \frac{\Delta^+ p(\ell)}{p(\ell)}\, p(\ell) \\
&= \mathbb{E}\left[-\frac{\Delta^+ p(X)}{p(X)} \mathbb{1}\{X \geq k\} \,\Big|\, X \in \mathrm{spt}(p) \right].
\end{aligned}$$

For the converse implication, assume that

$$\rho_{X|p}(k) = \mathbb{E}\left[-\frac{\Delta^+ p(X)}{p(X)} \, \mathbb{1}\{X \geq k\} \;\middle|\; X \in \mathrm{spt}(p) \right], \quad k \in \mathbb{Z}, k \geq L.$$

We obtain for $f \in \mathcal{F}_p$

$$\begin{aligned}
\mathbb{E}\left[\Delta^+ f(X) \,\middle|\, X \in \mathrm{spt}(p)\right] &= \sum_{\ell=L}^{R} \left(\Delta^+ f(\ell)\right) \rho_{X|p}(\ell) \\
&= \sum_{\ell=L}^{R} \Delta^+ f(\ell) \, \mathbb{E}\left[-\frac{\Delta^+ p(X)}{p(X)} \, \mathbb{1}\{X \geq \ell\} \;\middle|\; X \in \mathrm{spt}(p) \right] \\
&= \mathbb{E}\left[-\frac{\Delta^+ p(X)}{p(X)} \sum_{\ell=L}^{X} \left(f(\ell+1) - f(\ell)\right) \;\middle|\; X \in \mathrm{spt}(p) \right] \\
&= \mathbb{E}\left[-\frac{\Delta^+ p(X)}{p(X)} \, f(X+1) \;\middle|\; X \in \mathrm{spt}(p) \right],
\end{aligned}$$

where we use that $f(L) = 0$ by Remark 2.3, and where Fubini's theorem is applicable as

$$\sum_{\ell=L}^{R} \mathbb{E}\left| \Delta^+ f(\ell) \, \frac{\Delta^+ p(X)}{p(X)} \, \mathbb{1}\{X \geq \ell, \, X \in \mathrm{spt}(p)\} \right|$$

$$\leq \sup_{k \in \{L,\dots,R\}} \left| \Delta^+ f(k) \right| \cdot \mathbb{E}\left| \frac{\Delta^+ p(X)}{p(X)} \, (X - L) \cdot \mathbb{1}\{X \in \mathrm{spt}(p)\} \right| < \infty.$$

Theorem 2.2 implies the claim.

Acknowledgments

References

ARAGÓN, J., EBERLY, D. and EBERLY, S. (1992). Existence and uniqueness of the maximum likelihood estimator for the two-parameter negative binomial distribution. *Statistics & Probability Letters* **15** 375–379. MR1193896

ARRAS, B. and HOUDRÉ, C. (2019). *On Stein's method for infinitely divisible laws with finite first moment. SpringerBriefs in Probability and Mathematical Statistics.* Springer International Publishing, Cham. MR3931309

BARBOUR, A. D. (1988). Stein's method and Poisson process convergence. *Journal of Applied Probability* **25** 175–184. MR0974580

BARBOUR, A. D. (1990). Stein's method for diffusion approximations. *Probability Theory and Related Fields* **84** 297–322. MR1035659

BARINGHAUS, L. and HENZE, N. (1992). A goodness of fit test for the Poisson distribution based on the empirical generating function. *Statistics & Probability Letters* **13** 269–274. MR1160747

BARP, A., BRIOL, F. X., DUNCAN, A. B., GIROLAMI, M. A. and MACKEY, L. W. (2019). Minimum Stein discrepancy estimators. In *Proceedings of the 33rd International Conference on Neural Information Processing Systems* (H. WALLACH, H. LAROCHELLE, A. BEYGELZIMER, F. D'ALCHÉ-BUC, E. FOX and R. GARNETT, eds.). *Advances in Neural Information Processing Systems* **32** 12964–12976. Curran Associates, Inc.

BELTRÁN-BELTRÁN, J. I. and O'REILLY, F. J. (2019). On goodness of fit tests for the Poisson, negative binomial and binomial distributions. *Statistical Papers* **60** 1–18. MR3905436

BETSCH, S. and EBNER, B. (2021). Fixed point characterizations of continuous univariate probability distributions and their applications. *Annals of the Institute of Statistical Mathematics* **73** 31–59. MR4205241

BETSCH, S., EBNER, B. and KLAR, B. (2021). Minimum L^q-distance estimators for non-normalized parametric models. *The Canadian Journal of Statistics* **42** 514–548. MR4267931

BROWN, T. C. and PHILLIPS, M. J. (1999). Negative binomial approximation with Stein's method. *Methodology and Computing in Applied Probability* **1** 407–421. MR1770372

BYRD, R. H., LU, P., NOCEDAL, J. and ZHU, C. (1995). A limited memory algorithm for bound constrained optimization. *SIAM Journal on Scientific Computing* **16** 1190–1208. MR1346301

CHEN, L. H. Y. (1975). Poisson approximation for dependent trials. *Annals of Probability* **3** 534–545. MR0428387

CHEN, L. H. Y., GOLDSTEIN, L. and SHAO, Q. M. (2011). *Normal approximation by Steins method. Probability and its applications.* Springer, Berlin. MR2732624

CHOUDARY, A. D. R. and NICULESCU, C. P. (2014). *Real Analysis on Intervals.* Springer India, New Delhi. MR3243604

CHWIALKOWSKI, K., STRATHMANN, H. and GRETTON, A. (2016). A kernel test of goodness of fit. In *Proceedings of the 33rd International Conference on Machine Learning* (M. F. BALCAN and K. Q. WEINBERGER, eds.). *JMLR: W&CP* **48** 2606–2615. Proceedings of Machine Learning Research.

DUTANG, C., GOULET, V. and PIGEON, M. (2008). actuar: an R package for actuarial science. *Journal of Statistical Software* **25**.

EHM, W. (1991). Binomial approximation to the Poisson binomial distribution. *Statistics & Probability Letters* **11** 7–16. MR1093412

EICHELSBACHER, P. and REINERT, G. (2008). Stein's method for discrete Gibbs measures. *The Annals of Applied Probability* **18** 1588–1618. MR2434182

FREY, J. (2012). An exact Kolmogorov–Smirnov test for the Poisson distribution with unknown mean. *Journal of Statistical Computation and Simulation* **82** 1023–1033. MR2949505

GORHAM, J. and MACKEY, L. (2015). Measuring sample quality with Stein's method. In *Proceedings of the 28th International Conference on Neural In-*

formation Processing Systems (C. CORTES, N. D. LAWRENCE, D. D. LEE, M. SUGIYAMA and R. GARNETT, eds.). *Advances in Neural Information Processing Systems* **28** 226–234. Curran Associates, Inc.

GÖTZE, F. (1991). On the rate of convergence in the multivariate CLT. *The Annals of Probability* **19** 724–739. MR1106283

GÜRTLER, N. and HENZE, N. (2000). Recent and classical goodness-of-fit tests for the Poisson distribution. *Journal of Statistical Planning and Inference* **90** 207–225. MR1795597

GUTMANN, M. U. and HYVÄRINEN, A. (2010). Noise-contrastive estimation: A new estimation principle for unnormalized statistical models. In *Proceedings of the 13th International Conference on Artificial Intelligence and Statistics (AISTATS)* (Y. W. TEH and M. TITTERINGTON, eds.). *JMLR: W&CP* **9** 297–304. Journal of Machine Learning Research – Proceedings Track.

GUTMANN, M. U. and HYVÄRINEN, A. (2012). Noise-contrastive estimation of unnormalized statistical models, with applications to natural image statistics. *Journal of Machine Learning Research* **13** 307–361. MR2913702

HAYAKAWA, J. and TAKEMURA, A. (2016). Estimation of exponential-polynomial distribution by holonomic gradient descent. *Communications in Statistics – Theory and Methods* **45** 6860–6882. MR3544176

HENZE, N. (1996). Empirical-distribution-function goodness-of-fit tests for discrete models. *The Canadian Journal of Statistics / La Revue Canadienne de Statistique* **24** 81–93. MR1394742

HINTON, G. E. (2002). Training products of experts by minimizing contrastive divergence. *Neural Computation* **14** 1771–1800.

HYVÄRINEN, A. (2005). Estimation of non-normalized statistical models by score matching. *Journal of Machine Learning Research* **6** 695–709. MR2249836

HYVÄRINEN, A. (2007). Some extensions of score matching. *Computational Statistics & Data Analysis* **51** 2499–2512. MR2338984

INGLOT, T. (2019). Data driven efficient score tests for Poissonity. *Probability and Mathematical Statistics* **39** 115–126. MR3964387

JOHNSON, N. L., KOTZ, S. and KEMP, A. W. (1993). *Univariate discrete distributions (2nd Edition). Wiley Series in Probability and Mathematical Statistics*. Wiley, New York. MR1224449

KHMALADZE, E. (2013). Note on distribution free testing for discrete distributions. *The Annals of Statistics* **41** 2979–2993. MR3161454

KLAR, B. (1999). Goodness-of-fit tests for discrete models based on the integrated distribution function. *Metrika* **49** 53–69. MR1713644

KYRIAKOUSSIS, A., LI, G. and PAPADOPOULOS, A. (1998). On characterization and goodness-of-fit test of some discrete distribution families. *Journal of Statistical Planning and Inference* **74** 215–228. MR1671701

LEDWINA, T. and WYLUPEK, G. (2017). On Charlier polynomials in testing Poissonity. *Communications in Statistics – Simulation and Computation* **46** 1918–1932. MR3625257

LEY, C., REINERT, G. and SWAN, Y. (2017). Stein's method for comparison of univariate distributions. *Probability Surveys* **14** 1–52. MR3595350

LEY, C. and SWAN, Y. (2011). A unified approach to Stein characterizations. *arXiv e-prints* 1105.4925v3. MR2909575

LEY, C. and SWAN, Y. (2013a). Stein's density approach and information inequalities. *Electronic Communications in Probability* **18** 1–14. MR3019670

LEY, C. and SWAN, Y. (2013b). Local Pinsker inequalities via Stein's discrete density approach. *IEEE Transactions on Information Theory* **59** 5584-5591. MR3096942

LIU, Q., LEE, J. D. and JORDAN, M. (2016). A kernelized Stein discrepancy for goodness-of-fit tests. In *Proceedings of the 33rd International Conference on Machine Learning* (M. F. BALCAN and K. Q. WEINBERGER, eds.). *JMLR: W&CP* **48** 276–284. Proceedings of Machine Learning Research.

LYU, S. (2009). Interpretation and generalization of score matching. In *Proceedings of the 25th Conference on Uncertainty in Artificial Intelligence. UAI'09* 359–366. AUAI Press.

MATSUDA, T. and HYVÄRINEN, A. (2019). Estimation of non-normalized mixture models. In *Proceedings of the 22nd International Conference on Artificial Intelligence and Statistics (AISTATS)* (K. CHAUDHURI and M. SUGIYAMA, eds.). *PMLR* **89** 2555–2563. Proceedings of Machine Learning Research.

NIKITIN, Y. Y. (2017). Tests based on characterizations, and their efficiencies: A survey. *Acta et Commentationes Universitatis Tartuensis de Mathematica* **21** 3–24. MR3672200

PEKÖZ, E. A. (1996). Stein's method for geometric approximation. *Journal of Applied Probability* **33** 707–713. MR1401468

PUIG, P. and WEISS, C. H. (2020). Some goodness-of-fit tests for the Poisson distribution with applications in biodosimetry. *Computational Statistics & Data Analysis* **144** 106878. MR4029592

RIZZO, M. and SZÉKELY, G. (2019). energy: e-statistics: multivariate inference via the energy of data. R package, version 1.7-7.

RUEDA, R., O'REILLY, F. and PÉREZ-ABREU, V. (1991). Goodness of fit for the Poisson distribution based on the probability generating function. *Communications in Statistics – Theory and Methods* **20** 3093–3110. MR1144876

RUEDA, R. and O'REILLY, F. (1999). Tests of fit for discrete distributions based on the probability generating function. *Communications in Statistics – Simulation and Computation* **28** 259–274. MR1702548

STEIN, C. (1972). A bound for the error in the normal approximation to the distribution of a sum of dependent random variables. In *Proceedings of the Sixth Berkeley Symposium on Mathematical Statistics and Probability, Volume 2: Probability Theory* 583–602. University of California Press, Berkeley. MR0402873

STEIN, C. (1986). *Approximate computation of expectations. Lecture Notes – Monograph Series, Vol. 7.* Institute of Mathematical Statistics, Hayward. MR0882007

STEIN, C., DIACONIS, P., HOLMES, S. and REINERT, G. (2004). Use of exchangeable pairs in the analysis of simulations. In *Stein's Method* (P. DIACONIS and S. HOLMES, eds.). *Lecture Notes – Monograph Series* **46** 1–25. Institute of Mathematical Statistics, Beachwood. MR2118600

Székely, G. and Rizzo, M. (2004). Mean distance test of Poisson distribution. *Statistics & Probability Letters* **67** 241–247. MR2053526

Takenouchi, T. and Kanamori, T. (2017). Statistical inference with unnormalized discrete models and localized homogeneous divergences. *Journal of Machine Learning Research* **18** 1–26. MR3687599

R Core Team (2020). R: A language and environment for statistical computing. R Foundation for Statistical Computing, Vienna.

Uehara, M., Matsuda, T. and Kim, J. K. (2020). Imputation estimators for unnormalized models with missing data. In *Proceedings of the 23rd International Conference on Artificial Intelligence and Statistics (AISTATS)* (S. Chiappa and R. Calandra, eds.). *PMLR* **108** 831–841. Proceedings of Machine Learning Research.

Uehara, M., Kanamori, T., Takenouchi, T. and Matsuda, T. (2020). A unified statistically efficient estimation framework for unnormalized models. In *Proceedings of the 23rd International Conference on Artificial Intelligence and Statistics (AISTATS)* (S. Chiappa and R. Calandra, eds.). *PMLR* **108** 809–819. Proceedings of Machine Learning Research.

Wolodzko, T. (2019). extraDistr: additional univariate and multivariate distributions. R package, version 1.8.11.

Yang, J., Liu, Q., Rao, V. and Neville, J. (2018). Goodness-of-fit testing for discrete distributions via Stein discrepancy. In *Proceedings of the 35th International Conference on Machine Learning* (J. Dy and A. Krause, eds.). *PMLR* **80** 5561–5570. Proceedings of Machine Learning Research.

Yu, S., Drton, M. and Shojaie, A. (2019). Generalized score matching for non-negative data. *Journal of Machine Learning Research* **20** 1–70. MR3960930