The Information Geometry of Sparse Goodness-of-Fit Testing

Paul Marriott[1], Radka Sabolová[2], Germain Van Bever[3], Frank Critchley[2]

[1] University of Waterloo, 200 University Avenue West, Waterloo, Ontario, Canada
[2] The Open University, Walton Hall, Milton Keynes, Buckinghamshire, UK MK7 6AA
[3] Université libre de Bruxelles & ECARES, Avenue F.D. Roosevelt 42, 1050 Brussels, Belgium

Abstract

This paper takes an information-geometric approach to the challenging issue of goodness-of-fit testing in the high dimensional, low sample size context where, potentially, boundary effects dominate. The main contributions of this paper are threefold: first, we present, and prove, two new theorems on the behaviour of commonly used test statistics in this context; second, we investigate, in the novel environment of the extended multinomial model, the links between information geometry based divergences and standard goodness-of-fit statistics, allowing us to formalise relationships which have been missing in the literature; finally, we use simulation studies to validate and illustrate our theoretical results and to explore, currently open, research questions about the way that discretisation effects can dominate sampling distributions near the boundary. Novelly accommodating these discretisation effects contrasts sharply with the essentially continuous approach of skewness and other corrections flowing from standard higher-order asymptotic analysis.

Keywords: Extended Multinomial Models; Goodness-of-fit testing; Information geometry.

1 Introduction

We start by emphasising the threefold achievements of this paper, spelt out, in detail, in terms of the paper's section structure below. First, we present and prove two new theorems on the behaviour of some standard goodness-of-fit statistics in the high dimensional, low sample size context, focusing on behaviour 'near the boundary' of the extended multinomial family. We also comment on the methods of proof which allow explicit calculations of higher order moments in this context. Second, working again explicitly in the extended multinomial context, we fill a hole in the literature in linking information-geometric based divergences and standard goodness-of-fit statistics. Finally, we use simulation studies to explore discretisation effects, that can dominate sampling distributions 'near the boundary'. Indeed we illustrate and explore how, in the high dimensional, low sample size context, all distributions are effected by boundary effects. We also used these simulation results to explore, currently open, research questions. As can be seen, the overarching theme is the importance of working in

the geometry of the extended exponential family, [16], rather than the traditional, manifold based, structure of information geometry.

In more detail, the paper extends, and builds on, the results of [35] and we use notation and definitions consistently across these two papers. Both papers investigate the issue of goodness-of-fit testing in the high dimensional, sparse extended multinomial context, using the tools of Computational Information Geometry (CIG), [16].

Section 2 gives formal proofs of two results, Theorems 1 and 2, which were announced in [35]. These results explore the sampling performance of standard goodness-of-fit statistics – Wald, Pearson's χ^2, score and deviance – in the sparse setting. In particular, they look at the case where the data generation process is 'close to the boundary' of the parameter space where one or more cell probabilities vanish. This complements results in much of the literature where the centre of the parameter space – i.e. the uniform distribution – is often the focus of attention.

Section 3 starts with a review of the links between Information Geometry (IG), [4], and goodness-of-fit testing. In particular, it looks at the power family of Cressie and Read, [15], [39], in terms of the geometric theory of divergences. In the case of regular exponential families these links have been well-explored in the literature, [29], as has the corresponding sampling behaviour, [1]. What is novel here is the exploration of the geometry with respect to the closure of the exponential family, i.e. the extended multinomial model, a key tool in CIG. We illustrate how the boundary can dominate the statistical properties in ways that are surprising compared to standard – and even high-order – analyses, which are asymptotic in sample size.

Section 4 explores, through simulation experiments, the consequences of working in the sparse multinomial setting, with the design of the numerical experiments being inspired by the information geometry.

2 Sampling distributions in the sparse case

One of the first major impacts that Information Geometry had on statistical practice was through the geometric analysis of higher order asymptotic theory, see [2], [9]. Geometric interpretations and invariant expressions of terms in the higher order corrections to approximations of sampling distributions are a good example, [2, Chapter 4]. Geometric terms are used to correct for skewness and other higher order moment (cumulant) issues in the sampling distributions. However, these correction terms grow very large near the boundary, [16], [5]. Since this region plays a key role in modelling in the sparse setting – the MLE often being on the boundary – extensions to the classical theory are needed. This paper and [35] together start such a development. This work is related to similar ideas in categorical, (hierarchical) log-linear and graphical models, [33], [27], [24], and [16]. As stated in [24] '[their] statistical properties under sparse settings are still very poorly understood. As a result, [analysis of such data] remains exceptionally difficult'.

In this section we show why the Wald – equivalently, the Pearson χ^2 and score statistics – are unworkable when near the boundary of the extended multinomial model, but that the deviance has a simple, accurate and tractable sampling distribution even for moderate sample sizes. We also show how the higher moments of the deviance are easily computable, allowing in principle for higher order adjustments. However, we also make some observations about

the appropriateness of these classical adjustments in Section 4.

First, we define some notation, consistent with that of [35]. With i ranging over $\{0, 1, ..., k\}$, let $n = (n_i) \sim$ Multinomial $(N, (\pi_i))$, where here each $\pi_i > 0$. In this context the Wald, Pearson's χ^2, and score statistics all coincide, their common value, W, being

$$W := \sum_{i=0}^{k} \frac{(\pi_i - n_i/N)^2}{\pi_i} \equiv \frac{1}{N^2} \sum_{i=0}^{k} \frac{n_i^2}{\pi_i} - 1.$$

Defining $\pi^{(\alpha)} := \sum_i \pi_i^\alpha$ we note the inequality, for each $m \geq 1$,

$$\pi^{(-m)} - (k+1)^{m+1} \geq 0,$$

in which equality holds if and only if $\pi_i \equiv 1/(k+1)$ – i.e. iff (π_i) is uniform. We then have the following theorem, which establishes that the statistic W is unworkable as $\pi_{\min} := \min(\pi_i) \to 0$ for fixed k and N.

Theorem 1. *For $k > 1$ and $N \geq 6$, the first three moments of W are:*

$$E(W) = \frac{k}{N}, Var(W) = \frac{\left\{\pi^{(-1)} - (k+1)^2\right\} + 2k(N-1)}{N^3}$$

and $E[\{W - E(W)\}^3]$ given by

$$\frac{\left\{\pi^{(-2)} - (k+1)^3\right\} - (3k + 25 - 22N)\left\{\pi^{(-1)} - (k+1)^2\right\} + g(k, N)}{N^5},$$

where $g(k, N) = 4(N-1)k(k+2N-5) > 0$.

In particular, for fixed k and N, as $\pi_{\min} \to 0$

$$Var(W) \to \infty \text{ and } \gamma(W) \to +\infty,$$

where $\gamma(W) := E[\{W - E(W)\}^3]/\{Var(W)\}^{3/2}$.

A detailed proof is found in Appendix A.1 and we give here an outline of its important features. The machinery developed is capable of delivering much more than a proof of Theorem 1. As indicated there, it provides a generic way to explicitly compute arbitrary moments or mixed moments of multinomial counts and could in principle be implemented by computer algebra. Overall, there are four stages. A key recurrence relation is first established and then, second, exploited to deliver moments of a single cell count. Third, mixed moments of any order are derived from those of lower order, exploiting a certain functional dependence. Finally, results are combined to find the first three moments of W, higher moments being similarly obtainable.

The practical implication of Theorem 1 is that standard first, and higher-order, asymptotic approximations to the sampling distribution of the Wald, χ^2 and score statistics break down when the data generation process is 'close to' the boundary where at least one cell probability is zero. This result is qualitatively similar to results in [5] which shows how asymptotic approximations to the distribution of the maximum likelihood estimate fail, for example in the case of logistic regression, when the boundary is close in terms of distances as defined by the Fisher information.

Unlike statistics considered in Theorem 1, the deviance has a workable distribution in the same limit: that is, for fixed N and k as we approach the boundary of the probability simplex. In sharp contrast to that theorem we see the very stable and workable behaviour of the k-asymptotic approximation to the distribution of the deviance, in which the number of cells increases without limit.

Define the deviance D via

$$\begin{aligned} D/2 &= \sum\nolimits_{\{0\le i\le k: n_i>0\}} n_i \log(n_i/N) - \sum_{i=0}^{k} n_i \log(\pi_i) \\ &= \sum\nolimits_{\{0\le i\le k: n_i>0\}} n_i \log(n_i/\mu_i), \end{aligned}$$

where $\mu_i := E(n_i) = N\pi_i$. We will exploit the characterisation that the multinomial random vector (n_i) has the same distribution as a vector of independent Poisson random variables conditioned on their sum. Specifically, let the elements of (n_i^*) be *independently* distributed as Poisson $Po(\mu_i)$. Then, $N^* := \sum_{i=0}^{k} n_i^* \sim Po(N)$, while $(n_i) := (n_i^* | N^* = N) \sim$ Multinomial$(N, (\pi_i))$. Define the vector

$$S^* := \begin{pmatrix} N^* \\ D^*/2 \end{pmatrix} = \sum_{i=0}^{k} \begin{pmatrix} n_i^* \\ n_i^* \log(n_i^*/\mu_i) \end{pmatrix},$$

where D^* is defined implicitly and $0 \log 0 := 0$. The terms ν, τ and ρ are defined by the first two moments of S^* via the vectors

$$\begin{pmatrix} N \\ \nu \end{pmatrix} := E(S^*) = \begin{pmatrix} N \\ \sum_{i=0}^{k} E(n_i^* \log(n_i^*/\mu_i)) \end{pmatrix}, \tag{1}$$

$$\begin{pmatrix} N & \rho\tau\sqrt{N} \\ \cdot & \tau^2 \end{pmatrix} := Cov(S^*) = \begin{pmatrix} N & \sum_{i=0}^{k} C_i \\ \cdot & \sum_{i=0}^{k} V_i \end{pmatrix}, \tag{2}$$

where $C_i := Cov(n_i^*, n_i^* \log(n_i^*/\mu_i))$ and $V_i := Var(n_i^* \log(n_i^*/\mu_i))$.

Theorem 2. *Each of the terms ν, τ and ρ remains bounded as $\pi_{\min} \to 0$.*

We start with some preliminary remarks. We use the following notation. $\mathcal{N} := \{1, 2, ...\}$ denotes the natural numbers, while $\mathcal{N}_0 := \{0\} \cup \mathcal{N}$. Throughout, $X \sim Po(\mu)$ denotes a Poisson random variable having positive mean μ – that is, X is discrete with support $\mathcal{N}_0$ and probability mass function $p : \mathcal{N}_0 \to (0, 1)$ given by:

$$p(x) := e^{-\mu}\mu^x/x! \ (\mu > 0). \tag{3}$$

Putting:

$$\forall m \in \mathcal{N}_0, \ F^{[m]}(\mu) := \Pr(X \le m) = \textstyle\sum_{x=0}^{m} p(x) \in (0, 1), \tag{4}$$

for given μ, $\{1 - F^{[m]}(\mu)\}$ is strictly decreasing with m, vanishing as $m \to \infty$. For all $(x, m) \in \mathcal{N}_0^2$, we define $x_{(m)}$ by:

$$x_{(0)} := 1; \ x_{(m)} := x(x-1)...(x-(m-1)) \ (m \in \mathcal{N}) \tag{5}$$

so that, if $x \ge m$, $x_{(m)} = x!/(x-m)!$.

The set $\mathcal{A}_0$ comprises all functions $a_0 : (0, \infty) \to R$ such that, as $\xi \to 0_+$:

(i) $a_0(\xi)$ tends to an infinite limit $a_0(0_+) \in \{-\infty, +\infty\}$, while: (ii) $\xi a_0(\xi) \to 0$.

Of particular interest here, by l'Hôpital's rule,

$$\forall m \in \mathcal{N},\ (\log)^m \in \mathcal{A}_0, \tag{6}$$

where $(\log)^m : \xi \to (\log \xi)^m$ $(\xi > 0)$. For each $a_0 \in \mathcal{A}_0$, $\overline{a_0}$ denotes its continuous extension from $(0, \infty)$ to $[0, \infty)$ – that is: $\overline{a_0}(0) := a_0(0_+)$; $\overline{a_0}(\xi) := a_0(\xi)$ $(\xi > 0)$ – while, appealing to continuity, we also define $0\overline{a_0}(0) := 0$. Overall, denoting the extended reals by $\overline{R} := R \cup \{-\infty\} \cup \{+\infty\}$, and putting

$$\mathcal{A} := \{a : \mathcal{N}_0 \to \overline{R} \text{ such that } 0a(0) = 0\}$$

we have that $\mathcal{A}$ contains the disjoint union:

$$\{\text{all functions } a : \mathcal{N}_0 \to R\} \cup \{\overline{a_0}|_{\mathcal{N}_0} : a_0 \in \mathcal{A}_0\}.$$

We refer to $\overline{a_0}|_{\mathcal{N}_0}$ as *the member of* $\mathcal{A}$ *based on* $a_0 \in \mathcal{A}_0$.

We make repeated use of two simple facts. First:

$$\forall x \in \mathcal{N}_0,\ 0 \leq \log(x+1) \leq x, \tag{7}$$

equality holding in both places if, and only if, $x = 0$. And, second, (3) and (5) give:

$$\forall (x, m) \in \mathcal{N}_0^2 \text{ with } x \geq m,\ x_{(m)} p(x) = \mu^m p(x-m) \tag{8}$$

so that, by definition of $\mathcal{A}$:

$$\forall m \in \mathcal{N}_0, \forall a \in \mathcal{A},\ E(X_{(m)} a(X)) = \mu^m E(a(X+m)), \tag{9}$$

equality holding trivially when $m = 0$. In particular, taking $a = 1 \in \mathcal{A}$ – that is, $a(x) = 1$ $(x \in \mathcal{N}_0)$ – (9) recovers, at once, the Poisson factorial moments:

$$\forall m \in \mathcal{N}_0,\ E(X_{(m)}) = \mu^m$$

whence, in further particular, we also recover:

$$E(X) = \mu,\ E(X^2) = \mu^2 + \mu \text{ and } E(X^3) = \mu^3 + 3\mu^2 + \mu. \tag{10}$$

We are ready now to prove Theorem 2.

Proof of Theorem 2. In view of (1) and (2), it suffices to show that the first two moments of S^* remain bounded as $\pi_{\min} \to 0$. By the Cauchy-Schwarz inequality this, in turn, is a direct consequence of the following result.

□

Lemma 3. *Let* $X \sim Po(\mu)$ $(\mu > 0)$ *and put* $X_\mu := X \log(X/\mu)$, *with* $0 \log 0 := 0$. *Then, there exist* $b^{(1)}, b^{(2)} : (0, \infty) \to (0, \infty)$ *such that:*
(a) $0 \leq E(X_\mu) \leq b^{(1)}(\mu)$ *and* $0 \leq E(X_\mu^2) \leq b^{(2)}(\mu)$, *while:*
(b) for $i = 1, 2:$ $b^{(i)}(\mu) \to 0$ *as* $\mu \to 0_+$.

Proof. By (6), $a_0^{(1)}(\xi) := \log(\xi/\mu) \in \mathcal{A}_0$. Taking $m = 1$ and $a \in \mathcal{A}$ based on $a_0^{(1)}$ in (9), and using (7), gives at once the stated bounds on $E(X_\mu)$ with $b^{(1)}(\mu) = \mu(\mu - \log\mu)$ which does, indeed, tend to 0 as $\mu \to 0_+$.

Further, let $a_0^{(2)}(\xi) := \xi(\log(\xi/\mu))^2$. Taking $m = 1$ and a as the restriction of $a_0^{(2)}$ to $\mathcal{N}_0$ in (9) gives $E(X_\mu^2) = \mu E(a^{(2)}(X+1))$. Noting that

$$\{x \in \mathcal{N}_0 : \log((x+1)/\mu) < 0\} = \begin{cases} \emptyset & (\mu \le 1) \\ \{0, ..., \overline{\mu} - 2\} & (\mu > 1) \end{cases},$$

in which $\overline{\mu}$ denotes the smallest integer greater than or equal to μ, and putting

$$B(\mu) := \begin{cases} 0 & (\mu \le 1) \\ \mu \sum_{x=0}^{\overline{\mu}-2} a^{(2)}(x+1)p(x) & (\mu > 1) \end{cases},$$

(7), (10) and l'Hôpital's rule give the stated bounds on $E(X_\mu^2)$, with

$$\begin{aligned} b^{(2)}(\mu) &= B(\mu) + \mu \textstyle\sum_{x=0}^{\infty}(x+1)(x - \log\mu)^2 p(x) \\ &= B(\mu) + \mu E\{X^3 + X^2(1 - 2\log\mu) + X((\log\mu)^2 - 2\log\mu) + (\log\mu)^2\} \\ &= B(\mu) + \mu^4 + 4\mu^3 + 2\mu^2 + \mu(\log\mu)^2 + (\mu\log\mu)^2 - 2\mu(\mu+2)(\mu\log\mu) \end{aligned}$$

which does, indeed, tend to 0 as $\mu \to 0_+$. □

As a result of Theorem 2 the distribution of the deviance is stable in this limit. Further, as noted in [35], each of ν, τ and ρ can be easily and accurately approximated by standard truncate and bound methods in the limit as $\pi_{\min} \to 0$. These are detailed in Appendix A.2.

3 Divergences and Goodness-of-fit

The emphasis of this section is the importance of the boundary of the extended multinomial when understanding the links between information geometric divergences and families of goodness-of-fit statistics. A set of well-known results linking the Power-Divergence family and information geometry in the manifold sense are surveyed, for completeness, in Sections 3.1, 3.2 and 3.3. The extension to the extended multinomial family is discussed in Section 3.4, where we make clear how the global behaviour of divergences is dominated by boundary effects. This complements the usual local analysis which links divegences with the Fisher information, [2]. Perhaps the key point is, since counts in the data can be zero, information geometric structures should also allow probabilities to be zero. Hence closures of exponential families seem the correct geometric object to work on.

3.1 The Power-Divergence family

The results of Section 2 concern the boundary behaviour of two important members of a rich class of goodness-of-fit statistics. An important unifying framework which encompasses these, and other important, statistics can be found in [39, page 16] with the, so-called, Power-Divergence statistics. These are defined, for $-\infty < \lambda < \infty$ by

$$2NI^\lambda\left(\frac{n}{N} : \pi\right) := \frac{2}{\lambda(\lambda+1)} \sum_{i=0}^{k} n_i \left[\left(\frac{n_i}{N\pi_i}\right)^\lambda - 1\right], \tag{11}$$

with the cases $\lambda = -1, 0$ being defined by taking the appropriate limit to give

$$\lim_{\lambda \to -1} 2NI^{\lambda}\left(\frac{n}{N} : \pi\right) = 2\sum_{i=0}^{k} N\pi_i \log\left(N\pi_i/n_i\right), \ \lim_{\lambda \to 0} 2NI^{\lambda}\left(\frac{n}{N} : \pi\right) = 2\sum_{i=0}^{k} n_i \log\left(n_i/N\pi_i\right).$$

Important special cases are shown in Table 1, (whose first column is described below in Section 3.3) and we also note the case $\lambda = 2/3$, which Read and Cressie recommend, [39, page 79], as a reasonably robust statistic with an easily calculable critical value for small N. It, in a sense, lies 'between' the Pearson χ^2 and deviance statistics which we compared in §2.

$\alpha := 1 + 2\lambda$	λ	Formula	Name
3	1	$\sum_{i=0}^{k} \frac{(n_i - N\pi_i)^2}{N\pi_i}$	Pearson χ^2
7/3	2/3	$\frac{9}{5}\sum_{i=0}^{k} n_i \left[\left(\frac{n_i}{N\pi_i}\right)^{\frac{2}{3}} - 1\right]$	Read-Cressie
1	0	$2\sum_{i=0}^{k} n_i \log\left(n_i/N\pi_i\right)$	twice log-likelihood (deviance)
0	$-\frac{1}{2}$	$4\sum_{i=0}^{k} \left(\sqrt{n_i} - \sqrt{N\pi_i}\right)^2$	Freeman-Tukey or Hellinger
-1	-1	$2\sum_{i=0}^{k} N\pi_i \log\left(N\pi_i/n_i\right)$	twice modified log-likelihood
-3	-2	$\sum_{i=0}^{k} \frac{(n_i - N\pi_i)^2}{n_i}$	Neyman χ^2

Table 1: Special cases of the Power-divergence statistics.

This paper is primarily concerned with the sparse case where many of the n_i counts are zero and we also are interested in letting probabilities, π_i, becoming arbitrarily small, or even zero.

3.2 Literature review

Before we look at this, we briefly review the literature on the geometry of goodness-of-fit statistics. A good source for the historical developments, in the discrete context, can be found in [39, pages 131-153] and [1]. Important examples include the analysis of contingency tables, log-linear and discrete graphical models. Testing is often used to check the consistency of a parametric model with given data, and to check dependency assumptions such as independence between categorical variables. We note an important *caveat* though: as pointed out by [21], [14], the fact that a parametric model 'passes' a goodness-of-fit test constrains the resulting inference only weakly. The essential point here is that goodness-of-fit is a necessary, but not sufficient, condition for model choice since, in general, *many* models will be empirically supported. This issue has recently been explored geometrically in [6] using CIG.

There have been many possible test statistics proposed for goodness-of-fit testing and one of the attractions of the Power-Divergence family, defined in (11), is that the most important ones are included in the family and indexed by a single scalar λ. When there is a choice of test statistic, of course, different inferences can result from different choices. One of the main themes of [39] is to give the analyst insight about selecting a particular λ. Key considerations for making the selection of λ include: the tractability of the sampling distribution, its power against important alternatives, and interpretation when hypotheses are rejected.

The first order, asymptotic in N, χ^2-sampling distribution for all members of the Power-Divergence family, which is appropriate when all observed counts are 'large enough', is the most commonly used tool, and a very attractive feature of the family. However, this can fail badly in the 'sparse' case and when the model is close to the boundary. Elementary, moment based corrections, to improve small sample performance, are discussed in [39, Chapter 5]. More formal asymptotic approaches to these issues include the doubly asymptotic, in N and k, approach of [37], discussed in §2 and similar normal approximation ideas in [38]. See also [28]. Extensive simulation experiments have been undertaken to learn in practice what 'large enough' means, see [31], [39, Section 5.3], and [32].

When, as is common, there are nuisance parameters to be estimated, [36] points out that it is the sampling distribution *conditional* upon these estimates which needs to be approximated, and proposes higher order methods based on the Edgeworth expansion. Simulation approaches are often used in the conditional context due to the common intractability of the conditional distribution, [25], [30], and importance sampling methods play an important role, see [10], [13] and [34]. Other approaches to investigate the sampling distribution include jackknifing, [42], using the Chen-Stein method, [26] and detailed asymptotic analysis in [23], [43], and [7].

In very high dimensional model spaces, considerations of the power of tests rarely generates uniformly best procedures but, we feel, geometry can be an important tool in understanding the choices that need to be made. Further, [39, Section 5.4], states the situation is 'complicated', showing this through simulation experiments. One of the reasons for Read and Cressie's preferred choice of $\lambda = 2/3$ is its good power against some important types of alternative – the so-called bump or dip cases – as well as the relative tractability of its sampling distribution under the null. Other considerations about power can be found in [44] which looks specifically at mixture model based alternatives.

3.3 Links with Information Geometry

At the time that the Power-Divergence family was being examined there was a parallel development in Information Geometry but, oddly, it seemed to have taken some time before the links between the two areas were fully recognised. A good treatment of these links can be found in [29, Chapter 9]. Since it is important to understand the extreme values of divergence functions, considerations of convexity clearly can play an important role. The general class of Bregman divergences, [12], [29, Page 240] and [3, Page 13], is very useful here. For each Bregman divergence there will exist affine parameters of the exponential family in which the divergence function is convex. In the class of product Poisson models – which are the key building blocks of log-linear models – all members of the Power-Divergence family have the Bregman property. These are then α-divergences, capable of generating the complete Information Geometry of the model, [3], with the link between α and λ given in Table 1. The α-representation highlights the duality properties, which are a cornerstone of Information Geometry, but which is rather hidden in the λ representation. The Bregman divergence representation for the Poisson is given in Table 2. The divergence parameter, in which we have convexity, is shown for each λ, as is the so-called potential function which generates the complete information geometry for these models.

λ	α	Divergence $D_\lambda(\mu_1,\mu_2)$	Divergence parameter ξ	Potential
-1	-1	$\mu_1-\mu_2-\mu_2\left(\log(\mu_1)-\log(\mu_2)\right)$	$\xi=\log(\mu)$	$\exp(\xi)$
0	1	$\mu_2-\mu_1-\mu_1\left(\log(\mu_2)-\log(\mu_1)\right)$	$\xi=\mu$	$\xi\log(\xi)-\xi$
$\lambda\neq 0,-1$	$\alpha\neq\pm 1$	$\frac{\left(\lambda^*\mu_1-\lambda^*\mu_2-\mu_2\left(\left(\frac{\mu_1}{\mu_2}\right)^{\lambda^*}-1\right)\right)}{\lambda^*(1-\lambda^*)}$	$\xi=\frac{1}{\lambda^*}\mu^{\lambda^*}$	$\frac{(\lambda^*\xi)^{1/\lambda^*}}{1-\lambda^*}$

Table 2: Power-divergence in the Poisson model with mean μ, where $\lambda^*=1-\lambda$.

3.4 Extended multinomial case

In this paper we are focusing on the class of log-linear models where the multinomial is the underlying class of distributions – that is we condition on the sample size, N, being fixed in the product Poisson space. In particular we focus on extended multinomials which includes the closure of the multinomials, so we have a boundary. Due to the conditioning, which induces curvature, only the cases where $\lambda=0,-1$ remain Bregman divergences, but all are still divergences in the sense of being Csiszár f-divergences, [18] and [19].

The closure of an exponential family, see [8], [11], [33] and [20], and its application in the theory of log-linear models have been explored by [22], [27] , [40] and [24]. The key here is understanding the limiting behaviour in the natural, $\alpha=1$ in the sense of [2], parameter space. This can be done by considering the polar dual [17] or, alternatively, the directions of recession, [27] or [40]. The boundary polytope determines key statistical properties of the model including the behaviour of the sampling distribution of (functions of) the MLE and the shape of level sets of divergence functions.

Figures 1 and 2 show level sets of the $\alpha=\pm 1$ Power-Divergences in the $(+1)$-affine and (-1)-affine parameters, panels (a) and (b) respectively, for the $k=2$ extended multinomial model. The boundary polytope in this case is a simple triangle 'at infinity' and the shape of this is strongly reflected in the behaviour of the level sets. In Fig. 1 we show, in the simplex $\left\{(\pi_0,\pi_1,\pi_2)|\sum_{i=0}^2\pi_i=1,\pi_i\geq 0\right\}$, the level sets of the $\alpha=-1$ divergence, which is, in the Csiszár f-divergence form,

$$K(\pi^0,\pi):=\sum_{i=0}^{2}\log\left(\frac{\pi_i^0}{\pi_i}\right)\pi_i^0.$$

The figures show how in Panel (a) the directions of recession dominate the shape of level sets and in Panel (b) the duals of these directions, (i.e. the vertices of the simplex) each have different maximal behaviour. The lack of convexity of the level sets in Panel (a) corresponds to the fact that the natural parameters are not the affine divergence parameters for this divergence, so we do not expect convex behaviour. In Panel (b) we do get non-convex level sets, as expected.

Figure 2 shows the same story but this time for the dual divergence,

$$K^*(\pi,\pi^0):=K(\pi^0,\pi).$$

Now the affine divergence parameters are shown in Panel (a), the natural parameters. We see that in the limit the shape of the divergence is converging to that of the polar of the boundary polytope. In general, local behaviour is quadratic but boundary behaviour is polygonal.

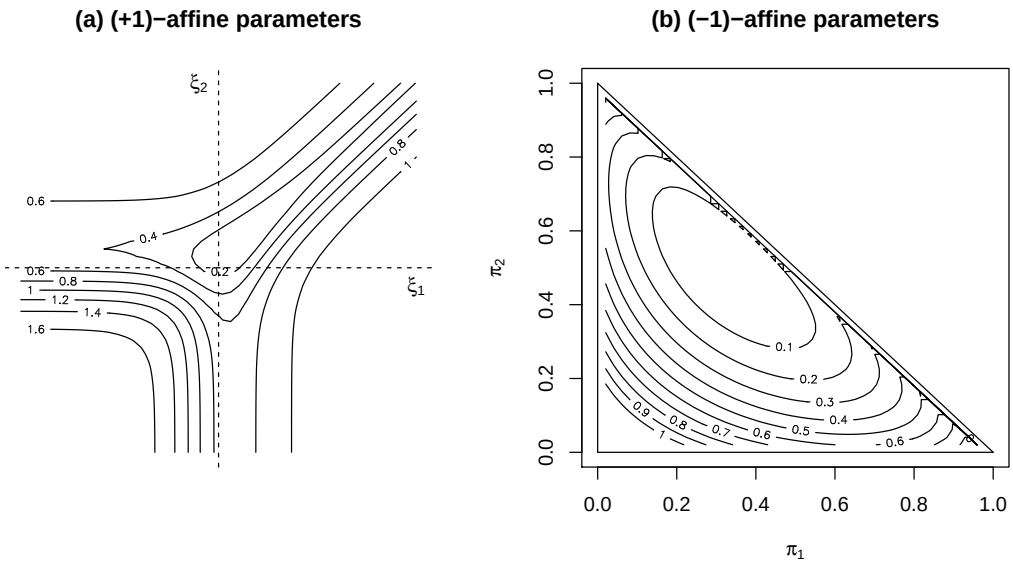

Figure 1: Level sets of $K(\pi^0, \pi)$, for fixed $\pi^0 = (\frac{1}{6}, \frac{2}{6}, \frac{3}{6})$ in: (a) the natural parameters, and (b) the mean parameters.

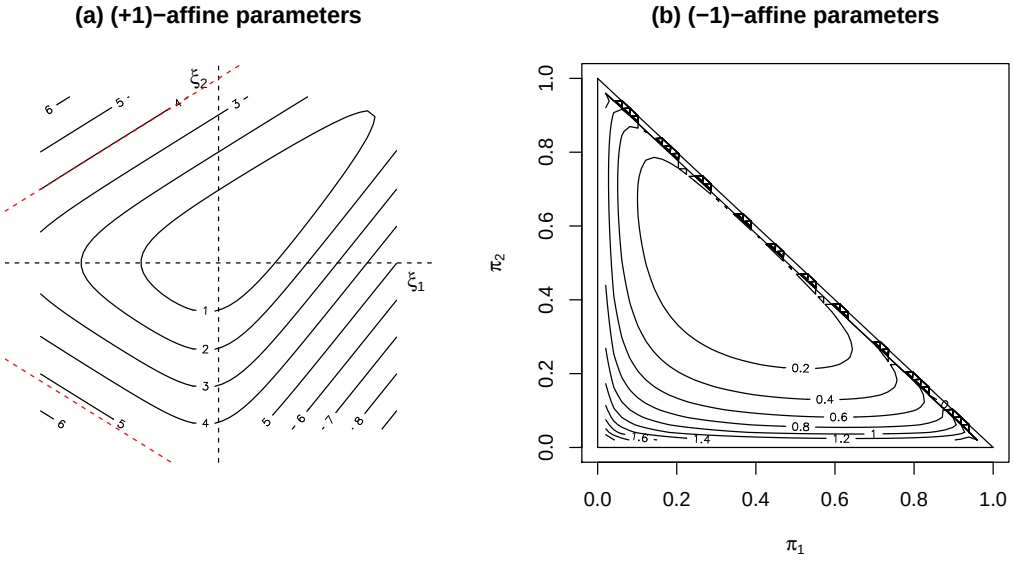

Figure 2: Level sets of $K^*(\pi^0, \pi)$, for fixed $\pi^0 = (\frac{1}{6}, \frac{2}{6}, \frac{3}{6})$ in: (a) the natural parameters, and (b) the mean parameters.

4 Simulation studies

In this section, we undertake simulation studies to numerically explore what has been discussed above. Separate sub-sections address three general topics – focusing on one particular

instance of each, as follows:

1. the transition as (N, k) varies between discrete and continuous features of the sampling distributions of goodness-of-fit statistics – focusing on the behaviour of the deviance at the uniform discrete distribution;
2. the comparative behaviour of a range of Power-Divergence statistics – focusing on the relative stability of their sampling distributions near the boundary; and:
3. the lack of uniformity, across the parameter space, of the finite sample adequacy of standard asymptotic sampling distributions – focusing on testing independence in 2×2 contingency tables.

For each topic, the results presented invite further investigation.

4.1 Transition between discrete and continuous features of sampling distributions

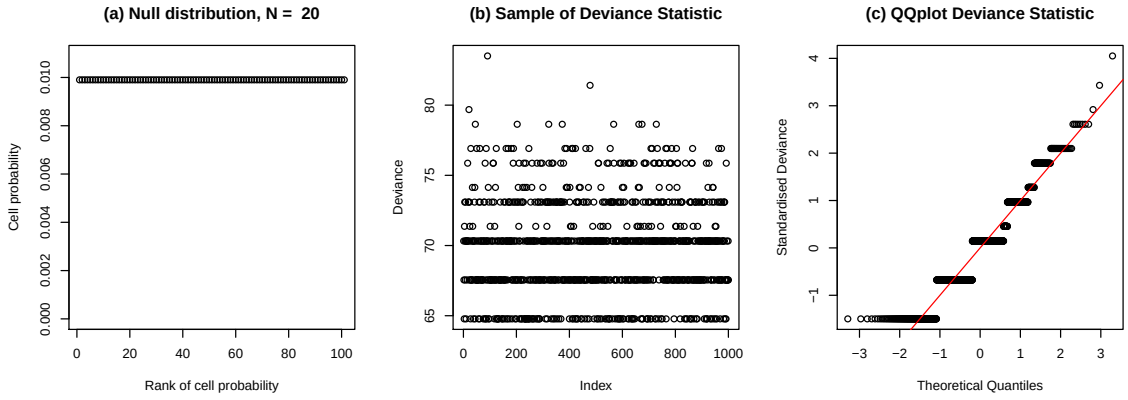

Figure 3: $k = 100$, $N = 20$

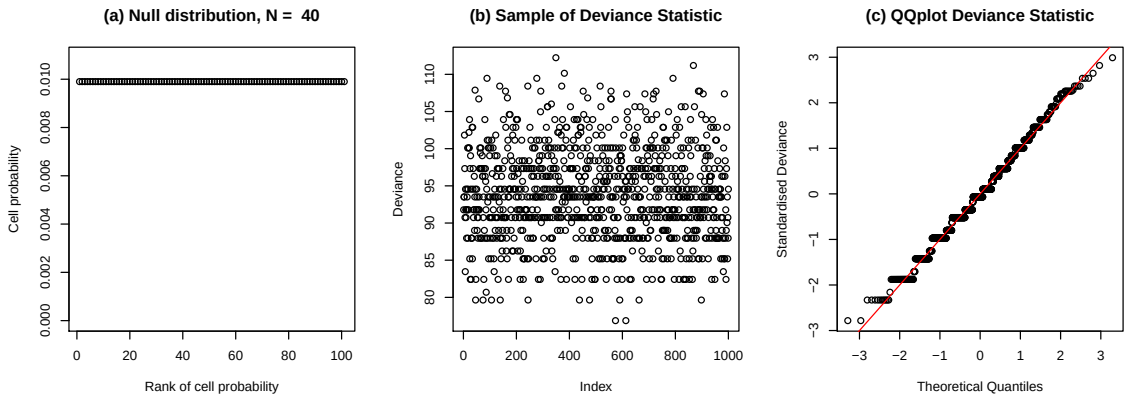

Figure 4: $k = 100$, $N = 40$

Earlier work [35] used the decomposition:

$$D^*/2 = \sum_{\{0\le i\le k: n_i^*>0\}} n_i^* \log(n_i^*/\mu_i) = \Gamma^* + \Delta^*,$$

$$\Gamma^* := \sum_{i=0}^{k} \alpha_i n_i^* \text{ and } \Delta^* := \sum_{\{0\le i\le k: n_i^*>1\}} n_i^* \log n_i^* \ge 0, \text{ where } \alpha_i := -\log \mu_i,$$

to show that a particularly bad case for the adequacy of any continuous approximation to the sampling distribution of the deviance $D := D^*|(N^* = N)$ is the uniform discrete distribution: $\pi_i = 1/(k+1)$. For, in this case, the Γ^* term contributes a constant to the deviance, while the Δ^* term has no contributions from cells with 0 or 1 observations – these being in the vast majority in the $N << k$ situation considered here. In other words, *all* of the variability in D comes from that between the $n_i \log n_i$ values for the, relatively rare, cell counts above 1. This gives rise to a discreteness phenomenon termed 'granularity' in [35], whose meaning was conveyed graphically there in the case $N = 30$ and $k = 200$. Work by Holst [28] predicts that continuous – indeed, normal – approximations will improve with larger values of N/k, as is intuitive. Remarkably, simply doubling the sample size to $N = 60$ was shown in [35] to be sufficient to give a good enough approximation for most goodness-of-fit testing purposes. In other words, N being 30% of $k = 200$ was found to be good enough for practical purposes.

Here, we illustrate the role of k-asymptotics (Section 2) in this transition between discrete and continuous features by repeating the above analyses for different values of k. Figures 3 and 4, where $k = 100$ while $N = 20$ and 40 respectively, are qualitatively the same as those presented in [35]. The difference here is that the smaller value of k means that a higher value of N/k (40%) is needed in Fig. 4 to adequately remove the granularity evident in Fig. 3. For $k = 400$, the figures with $N = 50$ and $N = 100$ (omitted here for brevity) are, again, qualitatively the same as in [35], the larger value of k needing only a smaller value of N/k (25%) for practical purposes. Note the QQ-plots used in these two figures are relative to normal quantiles.

The results of this section show the universality of boundary effects. The simulations of Figs. 3 and 4 are undertaken under the uniform model, which might be felt to be far from the boundary. The results show, in fact, that in the high dimensional, low sample size case, all distributions are 'close to' the boundary and that discretisation effects can dominate.

4.2 Comparative behaviour of Power-Divergence statistics near the boundary

We study here the relative stability, near the boundary of the simplex, of the sampling distributions of a range of Power-Divergence statistics indexed by Amari's parameter α. Fig. 5 shows histograms for six different values of α, $N = 50$, $k = 200$, and exponentially decreasing values of $\{\pi_i\}$, as plotted in Fig. 6. In it, red lines depict kernel density estimates using the bandwidth suggested in [41].

These sampling distributions differ markedly. The instability for $\alpha = 3$ expected from Theorem 1 is clearly visible: very large values contribute to high variance and skewness. Analogous instability features, albeit at a lower level, remain with the Cressie-Read recommended value $\alpha = 7/3$. In contrast, as expected from the discussion around Theorem 2, the

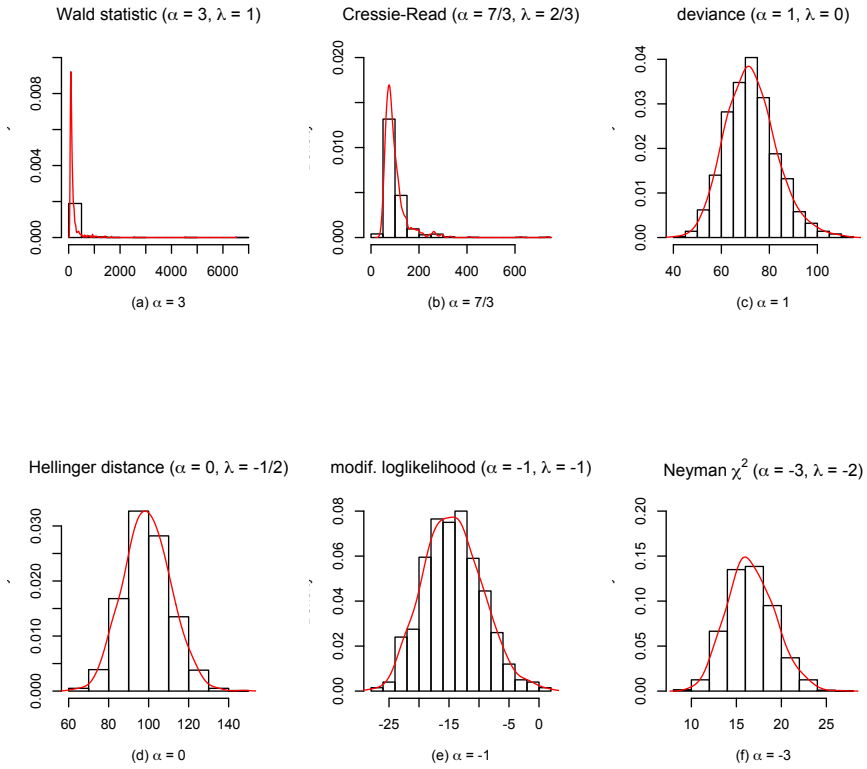

Figure 5: Sampling distributions for six members of the Power-Divergence family

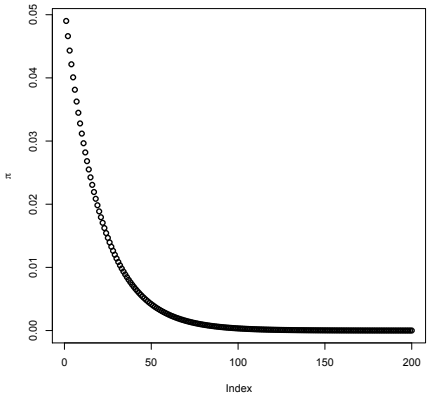

Figure 6: Exponentially decreasing values of π_i

distribution of the deviance ($\alpha = 1$) is stable and roughly normal. Lower values of α retain these same features.

4.3 Variation in finite sample adequacy of asymptotic distributions across the parameter space

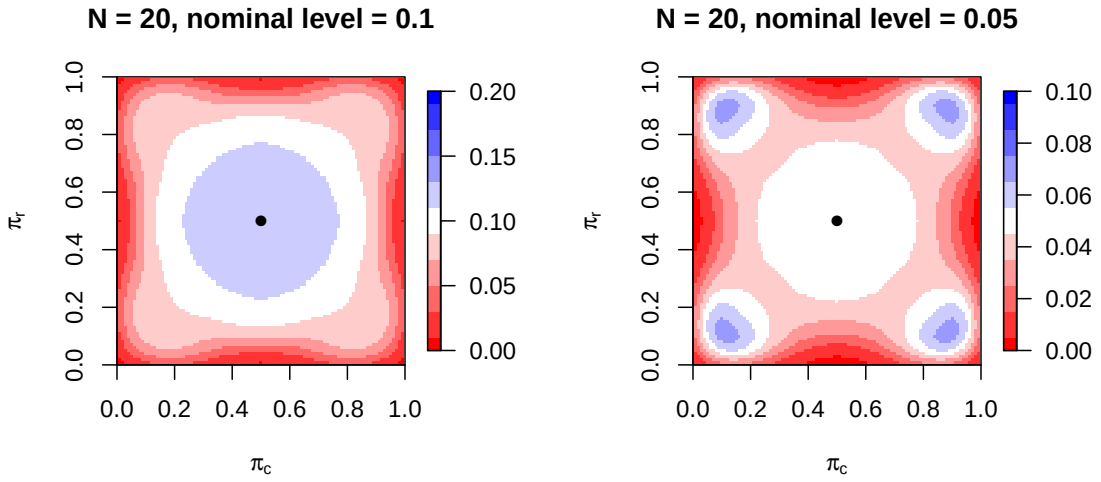

Figure 7: Heatmap of the actual level of the test for $N = 20$ at nominal levels 0.1 and 0.05; the standard rule-of-thumb, where expected counts are greater than 5, applies only at the black dot

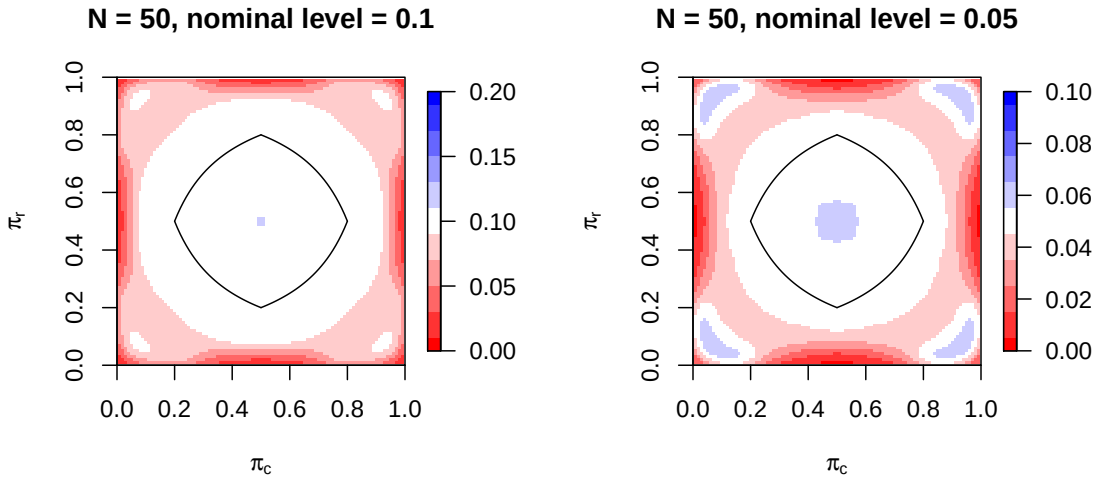

Figure 8: Heatmap of the actual level of the test for $N = 50$ at nominal levels 0.1 and 0.05; the standard rule-of-thumb, where expected counts are greater than 5, applies inside the closed black curved region

Pearson's χ^2 statistic ($\alpha = 3$) is widely used to test independence in contingency tables, a standard rule-of-thumb for its validity being that each expected cell frequency should be at least 5. For illustrative purposes, we consider 2×2 contingency tables, the relevant N-asymptotic null distribution being χ^2_1. We assess the adequacy of this asymptotic approximation by comparing nominal and actual significance levels of this test, based on 10,000 replications. Particular interest here lies in how these actual levels vary across different data generation processes within the same null hypothesis of independence.

Figures 7 and 8 show the actual level of the Pearson χ^2 test for nominal levels 0.1 and 0.05 for sample sizes $N = 20$ and $N = 50$, π_r and π_c denoting row and column probabilities respectively. The above rule-of-thumb applies only at the central black dot in Figure 7, and inside the closed black curved region in Figure 8. The actual level was computed for all pairs of values of π_r and π_c and then averaged using the symmetry of the parameter space and smoothed using the kernel smoother for irregular 2D data (implemented in the package *fields* in `R`). In each case, the white tone contains the nominal level, while red tones correspond to liberal and blue tones to conservative actual levels.

The finite sample adequacy of this standard asymptotic test clearly varies across the parameter space. In particular, its nominal and actual levels agree well at some parameter values outside the standard rule-of-thumb region; and, conversely, disagree somewhat at other parameter values inside it. Intriguingly, the agreement between nominal and actual levels does not improve everywhere with sample size. Overall, the clear patterns evident in this lack of uniformity invite further theoretical investigation.

5 Discussion

This paper has illustrated the key importance of working with the boundary of the closure of exponential families when studying goodness-of-fit testing in the high dimensional, low sample size context. Some of this work is new, Section 2, while some uses the structure of extended exponential families to add insight to standard results in the literature, Section 3. The last section, 4, uses simulation studies to start to explore open questions in this area.

One open question, related to the results of Theorems 1 and 2, is to see if a unified theory, for all values of α, and over large classes of extended exponential families, can be developed.

A APPENDICES

A.1 Appendix: Proof of Theorem 1

We start by noting an important recurrence relation which will be exploited in the computations below. By definition, for any $t := (t_i) \in R^{k+1}$, $n = (n_i)$ has moment generating function

$$M(t; N) := E\{\exp(t^T n)\} = [m(t)]^N$$

with $m(t) = \sum_{i=0}^{k} a_i$ and $a_i = a_i(t_i) = \pi_i e^{t_i}$. Putting

$$f_{N,i}(t; r) := N_{(r)} \, [m(t)]^{N-r} \, a_i^r \quad (0 \le r \le N),$$

where

$$N_{(r)} := {}^{N}P_r = \begin{cases} 1 & \text{if } r = 0 \\ N(N-1)...(N-(r-1)) & \text{if } r \in \{1,...,N\} \end{cases},$$

we have

$$M(t;N) = f_{N,i}(t;0) \quad (0 \le i \le k) \tag{12}$$

and the recurrence relation:

$$\frac{\partial f_{N,i}(t;r)}{\partial t_i} = f_{N,i}(t;r+1) + r f_{N,i}(t;r) \quad (0 \le i \le k;\ 0 \le r < N)\,. \tag{13}$$

When there is no risk of confusion, we may abbreviate $M(t;N)$ to M and $f_{N,i}(t;r)$ to $f_N(r)$, or even to $f(r)$ – so that (12) becomes $M = f(0)$. Again, we may write $\partial^r M(t;N)/\partial t_i^r$ as M_r, $\partial^{r+s} M(t;N)/\partial t_i^r \partial t_j^s$ as $M_{r,s}$ and $\partial^{r+s+u} M(t;N)/\partial t_i^r \partial t_j^s \partial t_l^u$ as $M_{r,s,u}$, with similar conventions for higher order mixed derivatives.

We can now use this to explicitly calculate low order moments of the count vectors. Using $E(n_i^r) = \partial^r M(t;N)/\partial t_i^r|_{t=0}$, the first N moments of n_i now follow from (12) and repeated use of (13), noting that $m(0) = 1$ and $a_i(0) = \pi_i$.

In particular, the first 6 moments of each n_i can be obtained as follows, where $N \ge 6$ is assumed. Using (12) and (13), we have

$$\begin{aligned}
M_1 &= f(1) \\
M_2 &= f(2) + f(1) \\
M_3 &= f(3) + 2f(2) + f(2) + f(1) = f(3) + 3f(2) + f(1) \\
M_4 &= f(4) + \ 6f(3) + \ 7f(2) + \ f(1) \\
M_5 &= f(5) + 10f(4) + 25f(3) + 15f(2) + \ f(1) \\
M_6 &= f(6) + 15f(5) + 65f(4) + 90f(3) + 31f(2) + f(1).
\end{aligned}$$

Substituting in, we have

$$\begin{aligned}
E(n_i) &= \ N\pi_i \\
E(n_i^2) &= N_{(2)}\pi_i^2 + \ N\pi_i \\
E(n_i^3) &= N_{(3)}\pi_i^3 + \ 3N_{(2)}\pi_i^2 + \ N\pi_i \\
E(n_i^4) &= N_{(4)}\pi_i^4 + \ 6N_{(3)}\pi_i^3 + \ 7N_{(2)}\pi_i^2 + \ N\pi_i \\
E(n_i^5) &= N_{(5)}\pi_i^5 + 10N_{(4)}\pi_i^4 + 25N_{(3)}\pi_i^3 + 15N_{(2)}\pi_i^2 + \ N\pi_i \\
E(n_i^6) &= N_{(6)}\pi_i^6 + 15N_{(5)}\pi_i^5 + 65N_{(4)}\pi_i^4 + 90N_{(3)}\pi_i^3 + 31N_{(2)}\pi_i^2 + N\pi_i.
\end{aligned}$$

This can be formalised in the following Lemma

Lemma 4. *The integer coefficients in any expansion*

$$M_r = \sum_{s=1}^{r} c_r(s) f(s) \quad (1 \le r \le N)$$

can be computed using $c_r(1) = c_r(r) = 1$ together, for $r \ge 3$, with the update:

$$c_r(s) = c_{r-1}(s-1) + s c_{r-1}(s) \quad (1 < s < r).$$

We note that if M_r is required for $r > N$, we may repeatedly differentiate

$$M_N = \sum_{s=1}^{N} c_N(s) f(s)$$

w.r.t. t_i, noting that $f(N) = N! a_i^N$ no longer depends on $m(t)$ so that, for all $h > 0$, $\partial^h f(N)/\partial t_i^h = N^h f(N)$.

Mixed moments of any order can be derived from those of lower order, exploiting the fact that a_i depends on t only via t_i. We illustrate this by deriving those required for the second and third moments of W.

First consider the mixed moments required for the second moment of W. Of course, $Var(W) = 0$ if $k = 0$. Otherwise, $k > 0$ and computing $Var(W)$ requires $E(n_i^2 n_j^2)$ for $i \neq j$. We find this as follows, assuming $N \geq 4$.

The relation $M_2 = f(2) + f(1)$ established above gives

$$\partial^2 M/\partial t_j^2 = N_{(2)} a_j^2 f_{N-2}(0) + N a_j f_{N-1}(0). \tag{14}$$

Repeated use of (14) now gives

$$M_{2,2} = N_{(4)} a_i^2 a_j^2 f_{N-4}(0) + N_{(3)} a_i a_j (a_i + a_j) f_{N-3}(0) + N_{(2)} a_i a_j f_{N-2}(0) \tag{15}$$

so that

$$E(n_i^2 n_j^2) = N_{(4)} \pi_i^2 \pi_j^2 + N_{(3)} \pi_i \pi_j (\pi_i + \pi_j) + N_{(2)} \pi_i \pi_j.$$

We further look at the mixed moments needed for the third moment of W. For the skewness of W, we need $E(n_i^2 n_j^4)$ for $i \neq j$ and, when $k > 1$, $E(n_i^2 n_j^2 n_l^2)$ for i, j, l distinct. We find these similarly, as follows, assuming $k > 1$ and $N \geq 6$.

Equation (15) above gives

$$\partial^2 M/\partial t_j^2 \partial t_l^2 = N_{(4)} a_j^2 a_l^2 f_{N-4}(0) + N_{(3)} a_j a_l (a_j + a_l) f_{N-3}(0) + N_{(2)} a_j a_l f_{N-2}(0)$$

from which, using (14) repeatedly, we have

$$\begin{aligned} M_{2,2,2} = {} & a_j^2 a_l^2 \{N_{(6)} a_i^2 f_{N-6}(0) + N_{(5)} a_i f_{N-5}(0)\} + a_j a_l (a_j + a_l)\{N_{(5)} a_i^2 f_{N-5}(0) + N_{(4)} a_i f_{N-4}(0)\} + \\ & a_j a_l \{N_{(4)} a_i^2 f_{N-4}(0) + N_{(3)} a_i f_{N-3}(0)\} \\ = {} & N_{(6)} a_i^2 a_j^2 a_l^2 f_{N-6}(0) + N_{(5)} a_i a_j a_l \{a_i a_j + a_j a_l + a_l a_i\} f_{N-5}(0) + N_{(4)} a_i a_j a_l \{a_i + a_j + a_l\} f_{N-4}(0) + \\ & N_{(3)} a_i a_j a_l f_{N-3}(0) \end{aligned}$$

so that $E(n_i^2 n_j^2 n_l^2)$ equals

$$N_{(6)} \pi_i^2 \pi_j^2 \pi_l^2 + N_{(5)} \pi_i \pi_j \pi_l \{\pi_i \pi_j + \pi_j \pi_l + \pi_l \pi_i\} + N_{(4)} \pi_i \pi_j \pi_l \{\pi_i + \pi_j + \pi_l\} + N_{(3)} \pi_i \pi_j \pi_l.$$

Finally, the relation $M_4 = f(4) + 6f(3) + 7f(2) + f(1)$ established above gives

$$\partial^4 M/\partial t_j^4 = \; N_{(4)} a_j^4 f_{N-4}(0) + 6 N_{(3)} a_j^3 f_{N-3}(0) + 7 N_{(2)} a_j^2 f_{N-2}(0) + \; N a_j f_{N-1}(0)$$

so that, again using (14) repeatedly, yields

$$E(n_i^2 n_j^4) = N_{(6)} \pi_i^2 \pi_j^4 + N_{(5)} \pi_i \pi_j^3 (6\pi_i + \pi_j) + N_{(4)} \pi_i \pi_j^2 (7\pi_i + 6\pi_j) + N_{(3)} \pi_i \pi_j (\pi_i + 7\pi_j) + N_{(2)} \pi_i \pi_j.$$

Combining above results, we obtain here the first three moments of W. Higher moments may be found similarly.

We first look at $E(W)$. We have $W = \frac{1}{N^2}\sum_{i=0}^{k}\frac{n_i^2}{\pi_i} - 1$ and $E(n_i^2) = N_{(2)}\pi_i^2 + N\pi_i$, so that

$$E(W) = \frac{N_{(2)}}{N^2} + \frac{(k+1)}{N} - 1 = \frac{k}{N}.$$

The variance is computed by recalling that $N^2(W+1) = \sum_i \frac{n_i^2}{\pi_i}$, while $E(W) = \frac{k}{N}$,

$$Var(W) = Var(W+1) = \frac{A^{(2)}}{N^4} - \left(\frac{k}{N} + 1\right)^2,$$

where

$$A^{(2)} := N^4 E\{(W+1)^2\} = \sum\nolimits_i \frac{E(n_i^4)}{\pi_i^2} + \sum\sum\nolimits_{i\neq j} \frac{E(n_i^2 n_j^2)}{\pi_i \pi_j}.$$

Using expressions for $E(n_i^4)$ and $E(n_i^2 n_j^2)$ established above, and putting

$$\pi^{(\alpha)} := \sum\nolimits_i \pi_i^\alpha,$$

we have

$$\begin{aligned}\sum\nolimits_i \frac{E(n_i^4)}{\pi_i^2} &= \sum\nolimits_i \{N_{(4)}\pi_i^2 + 6N_{(3)}\pi_i + 7N_{(2)} + N\pi_i^{-1}\} \\ &= N_{(4)}\pi^{(2)} + 6N_{(3)} + 7N_{(2)}(k+1) + N\pi^{(-1)}\end{aligned}$$

and

$$\begin{aligned}\sum\sum\nolimits_{i\neq j} \frac{E(n_i^2 n_j^2)}{\pi_i \pi_j} &= \sum\nolimits_{i\neq j} \{N_{(4)}\pi_i\pi_j + N_{(3)}(\pi_i + \pi_j) + N_{(2)}\} \\ &= N_{(4)}(1 - \pi^{(2)}) + 2N_{(3)}k + N_{(2)}k(k+1),\end{aligned}$$

so that

$$A^{(2)} = N_{(4)} + 2N_{(3)}(k+3) + N_{(2)}(k+1)(k+7) + N\pi^{(-1)},$$

whence

$$\begin{aligned}Var(W) &= \frac{N_{(4)} + 2N_{(3)}(k+3) + N_{(2)}(k+1)(k+7) + N\pi^{(-1)}}{N^4} - \left(1 + \frac{k}{N}\right)^2 \\ &= \frac{\left\{\pi^{(-1)} - (k+1)^2\right\} + 2k(N-1)}{N^3}, \quad \text{after some simplification.}\end{aligned}$$

Note that $Var(W)$ depends on (π_i) *only* via $\pi^{(-1)}$ while, by strict convexity of $x \to 1/x$ $(x > 0)$,

$$\pi^{(-1)} \geq (k+1)^2, \text{ equality holding iff } \pi_i \overset{i}{\equiv} 1/(k+1).$$

Thus, for given k and N, $Var(W)$ is strictly increasing as (π_i) departs from uniformity, tending to ∞ as one or more $\pi_i \to 0_+$.

Finally, for these calculations, we look at $E[\{W - E(W)\}^3]$. Recalling again that $N^2(W+1) = \sum_i \frac{n_i^2}{\pi_i}$,

$$\begin{aligned} E[\{W - E(W)\}^3] &= E[\{(W+1) - E(W+1)\}^3] \\ &= N^{-6}A^{(3)} - 3\,Var(W)(E(W)+1) - (E(W)+1)^3, \end{aligned}$$

where $A^{(3)} := N^6 E\{(W+1)^3\}$ is given by

$$A^{(3)} = \sum\nolimits_i \frac{E(n_i^6)}{\pi_i^3} + 3\sum\sum\nolimits_{i\neq j} \frac{E(n_i^2 n_j^4)}{\pi_i \pi_j^2} + \sum\sum\sum\nolimits_{i,j,l \text{ distinct}} \frac{E(n_i^2 n_j^2 n_l^2)}{\pi_i \pi_j \pi_l}.$$

Given that

$$E(W) = k/N \text{ and } Var(W) = \frac{\left\{\pi^{(-1)} - (k+1)^2\right\} + 2k(N-1)}{N^3},$$

it suffices to find $A^{(3)}$.

Using expressions for $E(n_i^6)$, $E(n_i^2 n_j^2 n_l^2)$ and $E(n_i^2 n_j^4)$ established above, we have

$$\sum\nolimits_i \frac{E(n_i^6)}{\pi_i^3} = N_{(6)}\pi^{(3)} + 15N_{(5)}\pi^{(2)} + 65N_{(4)} + 90N_{(3)}(k+1) + 31N_{(2)}\pi^{(-1)} + N\pi^{(-2)}$$

$$\begin{aligned} \sum\sum\nolimits_{i\neq j} \frac{E(n_i^2 n_j^4)}{\pi_i \pi_j^2} &= N_{(6)}\pi_i\pi_j^2 + N_{(5)}\pi_j(6\pi_i + \pi_j) + N_{(4)}(7\pi_i + 6\pi_j) + N_{(3)}(\pi_i/\pi_j + 7) + N_{(2)}\pi_j^{-1} \\ &= N_{(6)}\{\pi^{(2)} - \pi^{(3)}\} + N_{(5)}\{6 + (k-6)\pi^{(2)}\} + \\ &\qquad 13N_{(4)}k + N_{(3)}\{\pi^{(-1)} + (7k-1)(k+1)\} + N_{(2)}k\pi^{(-1)} \end{aligned}$$

and

$$\begin{aligned} \sum\sum\sum\nolimits_{i,j,l \text{ distinct}} \frac{E(n_i^2 n_j^2 n_l^2)}{\pi_i \pi_j \pi_l} &= N_{(6)}\{1 + 2\pi^{(3)} - 3\pi^{(2)}\} + 3N_{(5)}(k-1)\{1 - \pi^{(2)}\} + \\ &\qquad 3N_{(4)}k(k-1) + N_{(3)}k(k^2-1) \end{aligned}$$

so that, after some simplification,

$$\begin{aligned} A^{(3)} &= N_{(6)} + 3N_{(5)}(k+5) + N_{(4)}\{3k(k+12) + 65\} + \\ &\quad N_{(3)}\{k^3 + 21k^2 + 107k + 87\} + 3N_{(3)}\pi^{(-1)} + N_{(2)}(31 + 3k)\pi^{(-1)} + N\pi^{(-2)}. \end{aligned}$$

Substituting in and simplifying, we find $E[\{W - E(W)\}^3]$ to be:

$$\frac{\left\{\pi^{(-2)} - (k+1)^3\right\} - (3k + 25 - 22N)\left\{\pi^{(-1)} - (k+1)^2\right\} + g(k,N)}{N^5},$$

where

$$g(k,N) = 4(N-1)k(k+2N-5) > 0.$$

Note that $E[\{W - E(W)\}^3]$ depends on (π_i) *only* via $\pi^{(-1)}$ and the *larger* quantity $\pi^{(-2)}$. In particular, for given k and N, the skewness of W tends to $+\infty$ as one or more $\pi_i \to 0_+$.

A.2 Appendix: Truncate and bound approximations

In the notation of Lemma 3, it suffices to find truncate and bound approximations for each of $E(X_\mu)$, $E(X.X_\mu)$ and $E(X_\mu^2)$.

For all r, s in $\mathcal{N}$, define $h_{r,s}(\mu) := E\{(\log(X+r))^s\}$. Appropriate choices of $m \in \mathcal{N}_0$ and $a \in \mathcal{A}$ in (9), together with (10), give:

$$\begin{aligned} E(X_\mu) &= \mu h_{1,1}(\mu) - \mu \log \mu, \\ E(X.X_\mu) &= \{\mu^2 h_{2,1}(\mu) + \mu h_{1,1}(\mu)\} - (\mu^2 + \mu)\log \mu, \text{ and:} \\ E(X_\mu^2) &= \mu^2 h_{2,2}(\mu) + \mu h_{1,2}(\mu) + (\mu^2+\mu)(\log \mu)^2 - 2\log\mu\{\mu^2 h_{2,1}(\mu) + \mu h_{1,1}(\mu)\}, \end{aligned}$$

so that it suffices to truncate and bound $h_{r,s}(\mu)$ for $r, s \in \{1, 2\}$.

For all r, s in $\mathcal{N}$, and for all $m \in \mathcal{N}_0$, we write:

$$h_{r,s}(\mu) = h_{r,s}^{[m]}(\mu) + \varepsilon_{r,s}^{[m]}(\mu)$$

in which:

$$h_{r,s}^{[m]}(\mu) := \textstyle\sum_{x=0}^{m}\{(\log(x+r))^s\}p(x) \text{ and } \varepsilon_{r,s}^{[m]}(\mu) := \sum_{x=m+1}^{\infty}\{(\log(x+r))^s\}p(x).$$

Using again (7), the 'error term' $\varepsilon_{r,s}^{[m]}(\mu)$ has lower and upper bounds:

$$0 < \varepsilon_{r,s}^{[m]}(\mu) < \overline{\varepsilon}_{r,s}^{[m]}(\mu) := \textstyle\sum_{x=m+1}^{\infty}(x+(r-1))^s p(x).$$

Restricting attention now to $r, s \in \{1, 2\}$, as we may, and requiring $m \geq s$ so that $F^{[m-s]}(\mu)$ given by (4) is defined, (8) gives:

$$\begin{aligned} \overline{\varepsilon}_{1,1}^{[m]}(\mu) &= \textstyle\sum_{x=m+1}^{\infty} x p(x) = \mu \sum_{x=m}^{\infty} p(x) = \mu\{1 - F^{[m-1]}(\mu)\}, \\ \overline{\varepsilon}_{2,1}^{[m]}(\mu) &= \textstyle\sum_{x=m+1}^{\infty}(x+1)p(x) = \overline{\varepsilon}_{1,1}^{[m]}(\mu) + \{1 - F^{[m]}(\mu)\}, \\ \overline{\varepsilon}_{1,2}^{[m]}(\mu) &= \textstyle\sum_{x=m+1}^{\infty} x^2 p(x) = \sum_{x=m+1}^{\infty}\{x(x-1)+x\}p(x) \\ &= \mu^2\{1 - F^{[m-2]}(\mu)\} + \overline{\varepsilon}_{1,1}^{[m]}(\mu) \end{aligned}$$

and:

$$\begin{aligned} \overline{\varepsilon}_{2,2}^{[m]}(\mu) &= \textstyle\sum_{x=m+1}^{\infty}(x+1)^2 p(x) = \sum_{x=m+1}^{\infty}\{x^2 + (x+1) + x\}p(x) \\ &= \overline{\varepsilon}_{1,2}^{[m]}(\mu) + \overline{\varepsilon}_{2,1}^{[m]}(\mu) + \overline{\varepsilon}_{1,1}^{[m]}(\mu). \end{aligned}$$

Accordingly, for given μ, each $\overline{\varepsilon}_{r,s}^{[m]}(\mu)$ decreases strictly to zero with m providing, to any desired accuracy, truncate and bound approximations for each of ν, τ and ρ. In this connection, we note that the upper tail probabilities involved here can be bounded by standard Chernoff arguments.

Acknowledgements
The authors would like to thank the EPSRC for the support of grant number EP/L010429/1. Germain Van Bever would also like to thank FRS-FNRS for its support through the grant FC84444.

References

1. Agresti, A. *Categorical data analysis*, 3rd ed.; Wiley: Hoboken, NJ, USA, 2013.
2. Amari, S.-I. *Differential-geometrical methods in statistics. Lecture Notes in Statistics, Vol. 28*; Springer-Verlag Inc: New York, NY, USA, 1985.
3. Amari, S.-I. *Information geometry and its applications*; Springer: Japan, 2015.
4. Amari, S.-I. and Nagaoka, H. *Methods of information geometry*; Translations of mathematical monographs; v. 191, American Mathematical Society, 2000.
5. Anaya-Izquierdo, K.; Critchley, F.; Marriott, P. When are first-order asymptotics adequate? A diagnostic. *STAT* **2014**, *3*, 17–22.
6. Anaya-Izquierdo, K.; Critchley, F.; Marriott, P.; Vos, P. On the geometric interplay between goodness-of-fit and estimation: illustrative examples. In *Computational Information Geometry: For Image and Signal Processing*, Lecture Notes in Computer Science (LNCS); Nielsen, F., Dodson, K., Critchley, F., Eds.; Springer, 2016.
7. Asylbekov, Z.A.; Zubov, V.N.; Ulyanov, V.V. On approximating some statistics of goodness-of-fit tests in the case of three-dimensional discrete data. *Sib. Math. J.* **2011**, *52*, 571–584.
8. Barndorff-Nielsen, O. *Information and exponential families in statistical theory*; John Wiley & Sons, Ltd: Chichester, UK , 1978.
9. Barndorff-Nielsen, O.E.; Cox, D.R. *Asymptotic techniques for use in statistics*; Chapman & Hall: London, UK, 1989.
10. Booth, J.G.; Butler, R.W. An importance sampling algorithm for exact conditional tests in log-linear models. *Biometrika* **1999**, *86*, 321–332.
11. Brown, L.D. *Fundamentals of statistical exponential families with applications in statistical decision theory, Lecture Notes - monograph series, Vol. 9*; IMS: Hayward, CA, USA, 1986.
12. Bregman, L.M. The relaxation method of finding the common point of convex sets and its application to the solution of problems in convex programming. *USSR Comp. Math. Math+.* **1967**, *7*, 200–217.
13. Caffo, B.S.; Booth, J.G. Monte Carlo conditional inference for log-linear and logistic models: a survey of current methodology. *Stat. Methods in Med. Res.* **2003**, *12*, 109–123,
14. Copas, J.; Eguchi, S. Likelihood for statistically equivalent models, *J. R. Stat. Soc. B* **2010**, *72*, 193–217.
15. Cressie, N.; Read, T.R.C. Multinomial goodness-of-fit tests. *J. R. Stat. Soc. B* **1984**, *46*, 440–464.
16. Critchley, F.; Marriott P. Computational Information Geometry in Statistics: theory and practice. *Entropy* **2014**, *16*, 2454 –2471.
17. Critchley, F.; Marriott P. Computing with Fisher geodesics and extended exponential families. *Statistics and Computing* **2016**, *26*, 325–332.
18. Csiszár, I. On topological properties of f-divergences. *Studia Sci. Math. Hungar.* **1967**, *2*, 329–339.

19. Csiszár, I. Information measures: A critical survey. In *Transactions of the Seventh Prague Conference on Information Theory, Statistical Decision Functions, Random Processes and of the 1974 European Meeting of Statisticians, Volume B*, Kozesnik, J,, Ed.; Springer: Netherlands, 1977, 73–86.

20. Csiszár, I.; Matúš, F. Closures of exponential families. *Ann. Prob.* **2005**, *33*, 582–600.

21. Eguchi, S.; Copas, J. Local model uncertainty and incomplete-data bias. *J. R. Stat. Soc. B* **2005**, *67*, 1–37.

22. Eriksson, N.; Fienberg, S.E.; Rinaldo, A.; Sullivant, S. Polyhedral conditions for the nonexistence of the MLE for hierarchical log-linear models. *J. Symb. Comput.* **2006**, *41*, 222–233.

23. Fan, J.; Hung, H.-N.; Wong, W.-H. Geometric understanding of likelihood ratio statistics. *JASA* **2000**, *95*, 836–841.

24. Fienberg, S.E.; Rinaldo, A. Maximum likelihood estimation in log-linear models. *Ann. Stat.* **2012**, *40*, 996–1023.

25. Forster, J.J.; McDonald, J.W.; Smith, P.W.F. Monte Carlo exact conditional tests for log-linear and logistic models, *J. R. Stat. Soc. B*, **1996**, *58*, 445–453.

26. Gaunt, R.E.; Pickett, A.; Reinert, G. Chi-square approximation by Stein's method with application to Pearson's statistic. *arXiv preprint arXiv:1507.01707*, **2015**

27. Geyer, C.J. Likelihood inference in exponential families and directions of recession. *Electron. J. Stat.* **2009**, *3*, 259–289.

28. Holst, L. Asymptotic normality and efficiency for certain goodness-of-fit tests. *Biometrika* **1972**, *59*, 137–145.

29. Kass, R.E.; Vos, P.W. *Geometrical foundations of asymptotic inference*, John Wiley & Sons, Inc.: Hoboken, NJ, USA, 1997.

30. Kim, D.; Agresti, A. Nearly exact tests of conditional independence and marginal homogeneity for sparse contingency tables. *Comput. Stat. Data Anal.* **1997**, *24*, 89–104.

31. Koehler, K.J.; Larntz, K. An empirical investigation of goodness-of-fit statistics for sparse multinomials. *JASA* **1980**, *75*, 336–344.

32. Koehler, K.J. Goodness-of-fit tests for log-linear models in sparse contingency tables. *JASA* **1986**, 81, 483–493.

33. Lauritzen, S.L. *Graphical models*; Clarendon Press: Oxford, UK, 1996.

34. Lloyd, C.J. Computing highly accurate or exact P-values using importance sampling. *Comput. Stat. Data Anal.* **2012**, *56*, 1784–1794.

35. Marriott, P.; Sabolova, R.; Van Bever, G.; Critchley, F. Geometry of goodness-of-fit testing in high dimensional low sample size modelling. In *Geometric Science of Information: Second International Conference, GSI 2015, Palaiseau, France, October 28-30, 2015, Proceedings*, Nielsen, F., Barbaresco, F., Eds.; Springer, 2015, 569–576.

36. McCullagh, P. The conditional distribution of goodness-of-fit statistics for discrete data. *JASA* **1986**, *81*, 104–107.

37. Morris, C. Central limit theorems for multinomial sums. *Ann. Stat.* **1975**, *3*, 165–188.

38. Osius, G.; Rojek, D. Normal goodness-of-fit tests for multinomial models with large degrees of freedom. *JASA* **1992**, *87*, 1145–1152.

39. Read, T.R.C.; Cressie, N.A.C. *Goodness-of-fit statistics for discrete multivariate data*; Springer-Verlag: New York, NY, USA, 1988.

40. Rinaldo, A.; Feinberg, S.; Zhou, Y. On the geometry of discrete exponential families with applications to exponential random graph models. *Electron. J. Statist.* **2009**, *3*, 446–484.

41. Sheather, S.J.; Jones, M.C. A reliable data-based bandwidth selection method for kernel density estimation. *J. R. Stat. Soc. B* **1991**, *53*, 683–690.

42. Simonoff, J.S. Jackknifing and bootstrapping goodness-of-fit statistics in sparse multinomials. *JASA* **1986**, *81*, 1005–1011.

43. Ulyanov, V.V.; Zubov, V.N. Refinement on the convergence of one family of goodness-of-fit statistics to chi-squared distribution. *Hiroshima Math. J.* **2009**, *39*, 133–161.

44. Zelterman, D. Goodness-of-fit tests for large sparse multinomial distributions. *JASA* **1987**, *82*, 624–629.