Geometry of Goodness-of-Fit Testing in High Dimensional Low Sample Size Modelling

Paul Marriott[1], Radka Sabolova[2], Germain Van Bever[2], and Frank Critchley[2]

[1] University of Waterloo, Waterloo, Ontario, Canada
[2] The Open University, Milton Keynes, UK

Abstract. We introduce a new approach to goodness-of-fit testing in the high dimensional, sparse extended multinomial context. The paper takes a computational information geometric approach, extending classical higher order asymptotic theory. We show why the Wald – equivalently, the Pearson χ^2 and score statistics – are unworkable in this context, but that the deviance has a simple, accurate and tractable sampling distribution even for moderate sample sizes. Issues of uniformity of asymptotic approximations across model space are discussed. A variety of important applications and extensions are noted.

1 Introduction

A major contribution of classical information geometry to statistics is the geometric analysis of higher order asymptotic theory, see the seminal work [2] and for example [5]. It has excellent tools for constructing higher order corrections to approximations of sampling distributions, an example being the work on the geometry of Edgeworth expansions in [2, Chapter 4]. These expressions use curvature terms to correct for skewness and other higher order moment (cumulant) issues and provide good, operational corrections to sampling distributions, such as those in Fig. 3 (b) and (c) below. However, as discussed in [6] and [3], these curvature terms grow unboundedly as the boundary of the probability simplex is approached. 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 starts such a development.

Independently, there has been increased interest in categorical, (hierarchical) log-linear and graphical models. See, in particular, [10], [8], [7], and [6]. As stated by [7] '[their] statistical properties under sparse settings are still very poorly understood. As a result, [analysis of such data] remains exceptionally difficult'.

This paper is an introduction to a novel approach which combines and extends these two areas. The extension comes from using approximations based on the asymptotics of high dimensionality (k-asymptotics) rather than the more familiar sample size approach (N-asymptotics). This is connected to, but distinct from, the landmark paper by [12] and related work. In particular, for a practical example of so-called sparse-data asymptotics, see [1, §6.3]. Computational information geometry – in all its forms: see, for example, [11], [13] [4], and [6] – has

been a significant recent development, and this paper is a further contribution to it.

We address the challenging problems which arise in the high dimensional sparse extended multinomial context where the dimension k of the underlying probability simplex, one less than the number of categories or cells, is much more than the number of observations N, so that boundary effects necessarily occur, see [3]. In particular, arbitrarily small (possibly, zero) expected cell frequencies must be accommodated. Hence we work with *extended* multinomial models thus taking us out of the manifold structure of classical information geometry, [4].

For practical relevance, our primary focus is on (a) accurate, finite sample and dimension approximation, rather than asymptotic limiting results *per se*; and (b) performance at or near the boundary, rather than (as in earlier studies) the centre of the simplex.

Section 2.1 shows why the Wald statistic – identical, here, to the Pearson χ^2 or score statistic – is unworkable in this context. In contrast analysis and simulation exercises (§2.2) indicate that the same is not true of the deviance D. We demonstrate that a simple normal (or shifted χ^2) approximation to the distribution of D is accurate and tractable even as the boundary is approached. In contrast to other approaches, this appears to hold effectively *uniformly* across the simplex. The worst place is at its centre (where all cells are equiprobable), due to discretisation effects. However, further theory shows that, even here, the accuracy of approximation improves without limit when $N, k \to \infty$ with $N/k \to c > 0$.

Section 3 considers the uniformity of asymptotic approximations. Its three subsections address issues associated with the boundary, higher moments and discreteness, respectively.

2 Analysis

2.1 Why the Wald statistic is unworkable

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 $m \geq 1$,

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

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

2.2 The deviance statistic

Unlike the triumvirate of statistics above, 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. The paper [3] demonstrated the lack of uniformity across this simplex of standard first order N-asymptotic approximations. In sharp contrast to this we see the very stable and workable behaviour of the k-asymptotic approximation to the distribution of the deviance.

Define the deviance D via

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

where $\mu_i := E(n_i) = N\pi_i$. We will exploit the characterisation that the multinomial random vector n 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

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

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

$$\binom{N}{\nu} := E(S^*) = \binom{N}{\sum_{i=0}^{k} E(n_i^* \log\{n_i^*/\mu_i\})},$$

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

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

Theorem 2. *Each of these terms ν, τ and ρ are bounded as $\pi_{\min} \to 0$ and hence the distribution of the deviance is stable in this limit.*

Moreover, these terms can be easily and accurately approximated using standard truncate and bound computational methods, exploited below.

Under standard conditions, discussed in detail in §3, the multivariate central limit theorem (CLT) gives for large k, but for fixed N, that S^* is approximately distributed as a bivariate normal $N_2\left(E(S^*), cov(S^*)\right)$. Standard normal theory then gives, in the same limit,

$$D/2 = D^*/2|\{N^* = N\} \sim N_1(\nu, \tau^2(1-\rho^2)). \tag{1}$$

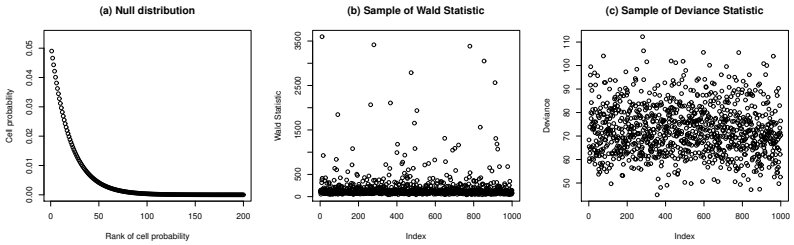

Fig. 1. Stability of the sampling distributions

3 Uniformity of asymptotic approximations

3.1 Uniformity near the boundary

In general asymptotic approximations are not uniformly accurate as is shown in [3]. Consider the consequences of Theorem 1 when $\pi_{\min}$ is close to zero as illustrated in Fig. 1. This shows, in panel (a), the distribution, π, where we see that $\pi_{\min}$ is indeed very small. Here, and throughout, we plot the distributions in rank order without loss since all sampling distributions considered are invariant to permutation of the labels of the multinomial cells. Panel (b) shows a sample of 1000 values of W drawn from its distribution when there are $N = 50$ observations in dimension $k = 200$. The extreme non-normality, and hence the failure of the standard N-asymptotic approximation, is evident. In contrast, consider panel

(c), which shows 1000 replicates of D for the same (N, k) values. The much greater stability, which is implied by Approximation (1), is extremely clear in this case.

The performance of Approximation (1) can, in fact, be improved by simple adjustments. Here we show a couple of examples in Fig. 2. Panel (a) shows a QQ-plot of the deviance, against the normal, in the case where the underlying distribution is shown in Fig. 1 (a) – one that is very close to the boundary. We see the normal approximation is good but shows some skewness. Panel (b) shows a scaled χ^2-approximation, designed to correct for skewness in the sampling distribution, while panel (c) shows a symmetrised version of the deviance statistic which, if it is used for testing against a two tailed alternative, is a valid procedure. Both these simple corrections show excellent performance.

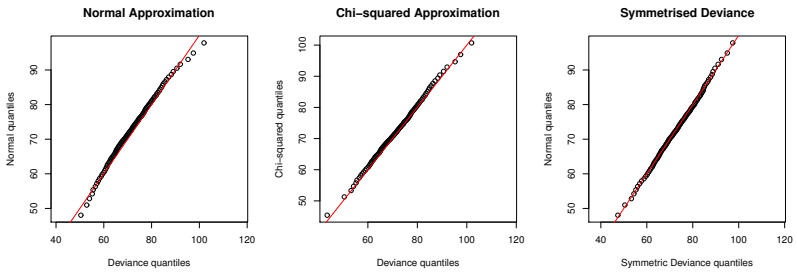

Fig. 2. Evaluation of the quality of k-asymptotic approximations

Having seen that the N-asymptotic approximation does not hold uniformly across the simplex, it is natural to investigate the uniformity of the k-asymptotic approximation given by (1). This approximation exploited a bivariate normal approximation to the distribution of $S^* = (N^*, D^*/2)^T$ and it is sufficient to check the normal approximation to any linear function of N^* and $D^*/2$. In particular, initially, we focus on the component D^*. We note that we can express $D^*/2$ via

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

where

$$\mathit{\Gamma}^* := \sum_{i=0}^{k} \alpha_i n_i^* \text{ and } \mathit{\Delta}^* := \sum\nolimits_{\{0 \le i \le k : n_i > 1\}} n_i^* \log n_i^* \ge 0$$

and $\alpha_i := -\log \mu_i$. It is insightful to consider the terms $\mathit{\Gamma}^*$ and $\mathit{\Delta}^*$ separately.

3.2 Uniformity and higher moments

One way of assessing the quality of the k-asymptotic approximation for the distribution of $\mathit{\Gamma}^*$ would be based on how well the moment generating function

of the (standardised) Γ^* is approximated by that of a (standard) normal. Writing the moment generating function as

$$M_\gamma(t) = \exp\left(-\frac{E(\Gamma^*)}{\sqrt{Var(\Gamma^*)}}\right) \exp\left[\sum_{i=0}^{k}\left\{\sum_{h=1}^{\infty}(-1)^h \mu_i (\log \mu_i)^h \left(\frac{t}{\sqrt{Var(\Gamma^*)}}\right)^h\right\}\right]$$

then, when analysing where the approximation would break down, it is natural to make the third order term (i.e. the skewness)

$$\sum_{i=0}^{k} \mu_i (\log \mu_i)^3$$

as large as possible for fixed mean $E(\Gamma^*) = -\sum_{i=0}^{k} \mu_i \log(\mu_i)$ and $Var(\Gamma^*) = \sum_{i=0}^{k} \mu_i (\log \mu_i)^2$.

Solving this optimisation problem gives a distribution with three distinct values for μ_i. An example of this is shown in Fig. 3, where $k = 200$. Panels (b) and (c) are histograms for a sample of 1000 values of W and D, respectively, drawn from their distribution when $N = 30$. In this example, we see both the Wald and deviance statistics are close to normal but with significant skewness which disappears with a larger sample size. This is to be expected from the analysis

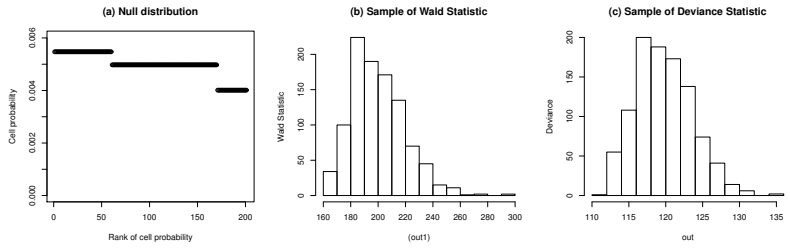

Fig. 3. Worst case solution for normality of Γ^*

of [9] and [12] who look at the behaviour of deviance, when bounded away from the boundary of the simplex, when both N and k tend to infinity together. In particular [9] shows the accuracy of this normal approximation improves without limit when $N, k \to \infty$ with $N/k \to c > 0$.

3.3 Uniformity and discreteness

In fact the hardest cases for the normal approximation (1) to the distribution of the deviance are in complementary parts of the simplex to the hardest cases from the Wald statistic. For W, it is the boundary where there are problems, while for (1) the worst place is the centre of the simplex, i.e. the uniform distribution.

The difficulties there are not due to large higher order moments, but rather to discreteness.

In this analysis consider again decomposition (2). Note that Γ^* is completely degenerate here, while there are never any contributions to the Δ^* term from cells for which n_i is 0 or 1. However, for $k >> N$, we would expect that for all i, $n_i^* \in \{0, 1\}$, with high probability, hence, after conditioning on $N^* = N$ there is no variability in D – it has a completely degenerate (singular) distribution. In the general case all the variability comes from the cases where $n_i^* > 1$ and these events can have a very discrete distribution – so the approximation given by the continuous normal must be poor.

We illustrate this 'granular' behaviour in Fig. 4. Panel (a) shows the uniform distribution when $k = 200$, panel (b) displays 1000 realisations of D when $N = 30$. The discreteness of the distribution is very clear here, and is also illustrated in panel (c) which shows a QQ-plot of the sample against a normal distribution. Note that any given quantile is not far from the normal, but the discreteness of the underlying distribution means that not all quantiles can be attained. This may, or may not, be a problem in a goodness-of-fit testing situation.

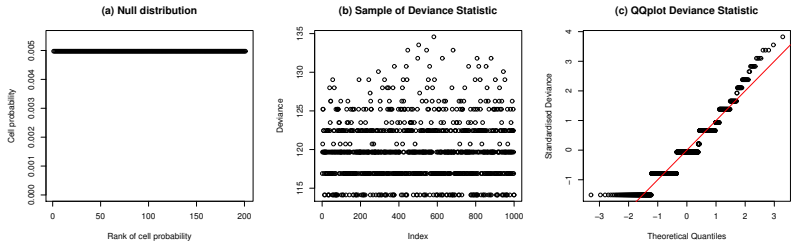

Fig. 4. Behaviour at the centre of the simplex, N=30

Again following the analysis of [9] this behaviour will disappear as N gets larger relative to k. This is shown in Fig. 5 where the N is now 60 – twice what it was in Fig. 4. The marked drop in granularity of panel (b) between Figures 4 and 5 is due to the much greater variability in the maximum observed value of n_i^* as N increases. Clearly, for the distribution of any discrete random variable to be well approximated by a continuous one, it is necessary that it have a large number of support points, close together. The good news here is that, for the deviance, this condition appears also to be sufficient.

4 Discussion

Overall, we have seen that the deviance remains stable and eminently useable in high-dimensional, sparse contexts – of accelerating practical importance. Discreteness issues are rare, predictable and well-understood, while simple modifications are available to deal with any higher moment concerns, such as skewness.

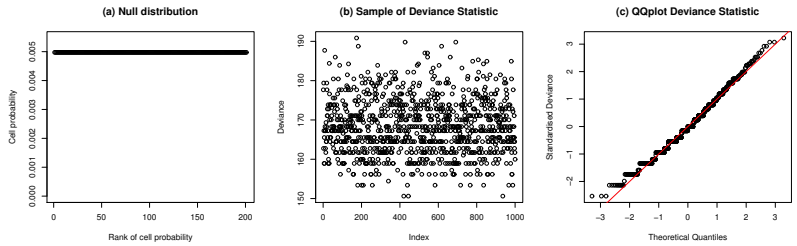

Fig. 5. Behaviour at the centre of the simplex, N=60

When using the deviance, computational information geometry can be used to gain insight into the power of the implicit likelihood ratio test, exploiting the fact that D is constant on high-dimensional affine subspaces in the mean parameterisation, [6], while both its null and alternative approximating distributions depend only on a few low-order moments, inducing a pivotal foliation.

Acknowledgements: The authors acknowledge with gratitude the support of EPSRC grant EP/L010429/1.

References

1. A. Agresti. *Categorical Data Analysis*. Wiley: Hoboken NJ, 2002.
2. S.-I. Amari. *Differential-geometrical methods in statistics*. Springer-Verlag, 1985.
3. K. Anaya-Izquierdo, F. Critchley, and P. Marriott. When are first order asymptotics adequate? a diagnostic. *STAT*, 3:17–22, 2014.
4. K. Anaya-Izquierdo, F. Critchley, P. Marriott, and P. Vos. Computational information geometry: foundations. *Proceedings of GSI 2013, LNCS*, 2013.
5. O.E. Barndorff-Nielsen and D.R. Cox. *Inference and asymptotics*. Chapman & Hall, 1994.
6. F. Critchley and Marriott P. Computational information geometry in statistics: theory and practice. *Entropy*, 16:2454 –2471, 2014.
7. S.E. Fienberg and A. Rinaldo. Maximum likelihood estimation in log-linear models. *Annals of Statistics*, 40:996–1023, 2012.
8. C. J. Geyer. Likelihood inference in exponential families and directions of recession. *Electronic Journal of Statistics*, 3:259–289, 2009.
9. L. Holst. Asymptotic normality and efficiency for certain goodness–of–fit tests. *Biometrika*, 59:137–145, 1972.
10. S.L. Lauritzen. *Graphical Models*. Clarendon Press, Oxford, 1996.
11. M. Liu, B.C. Vemuri, S.-I. Amari, and F. Nielsen. Shape retrieval using heirarchical total Bregman soft clustering. *IEEE Transactions on pattern analysis and machine intelligence*, 34:2407–2419, 2012.
12. C. Morris. Central limit theorems for multinomial sums. *Annals of Statistics*, 3:165–188, 1975.
13. F. Nielsen and N. Nock. Optimal interval clustering: Application to Bregman clustering and statistical mixture learning. *IEEE Transactions on pattern analysis and machine intelligence*, 21(10):1289–1292, 2014.