Bayesian Discrete Model Selection Based on Proper Scoring Rules

A. Philip Dawid*
University of Cambridge, Cambridge, U.K. - apd@statslab.cam.ac.uk

Monica Musio
University of Cagliari, Cagliari, Italy - mmusio@unica.it

Silvia Columbu
University of Cagliari, Cagliari, Italy - silvia.columbu@unica.it

Abstract

When improper priors are employed, the Bayes Factor—the standard Bayesian tool for comparing two different models in the light of data—becomes useless, involving an essentially arbitrary multiplicative constant. One way of sidestepping this problem is to replace negative log likelihood by a homogeneous proper scoring rule, which is insensitive to the value of such a multiplier. We have previously studied this for the case of real-valued continuous data, using the Hyvärinen scoring rule, and have shown that, so long as this is applied in a prequential (predictive sequential) manner, it leads, asymptotically, to selection of the simplest true model. In this work we report a parallel investigation for models with discrete data, for which quite different scoring rules are appropriate. For the particular case of distinguishing between Poisson and Negative Binomial models, we conduct simulations that indicate that, applied prequentially, the method will consistently select the true model.

Keywords: Consistent model selection; homogeneous score; prequential

1 Introduction

Bayesian model selection with improper within-model prior distributions is not well-defined, owing to the presence of an arbitrary multiplicative constant in each term of the marginal likelihood function. Recently (Dawid and Musio, 2015) it has been shown how this problem can be overcome if one replaces negative log-likelihood (the *log score*) by another, homogeneous, proper scoring rule (Parry *et al.*, 2012). That paper showed how, for continuous data, this approach can produce consistent selection of the correct model.

Here we study the case of discrete data. Simulations indicate that, for the problem of distinguishing between the Poisson and the Negative Binomial distributions, this method will again deliver consistent selection of the true model.

For an expanded version of this material, see Dawid *et al.* (2017).

2 Local scoring rules

Dawid *et al.* (2012) defined and characterised a *key local scoring rule* $S(x, P)$ on a discrete sample space $\mathcal{X}$, where $x \in \mathcal{X}$, and P is a distribution over $\mathcal{X}$. This will be proper, and *homogeneous* in the sense that its value is unchanged when all probabilities are scaled by the same positive constant.

Suppose the sample space $\mathcal{X}$ is the set of non-negative integers, and we regard x and y as neighbours when they differ by at most 1. A key local scoring rule adapted to this structure has the form

$$S(x,P) = G'_{x-1}\left\{\frac{p(x)}{p(x-1)}\right\} + G_x\left\{\frac{p(x+1)}{p(x)}\right\} - \frac{p(x+1)}{p(x)}\,G'_x\left\{\frac{p(x+1)}{p(x)}\right\} \quad (x=0,1,\ldots) \tag{1}$$

where $p(x) = P(X=x)$, G_x is a concave function on $\mathbb{R}^+$, and the first term in (1) is absent if $x=0$. It is clear from the way in which ratios enter (1) that such a scoring rule is homogeneous.

In the sequel we use the special case of (1) with

$$G_x(v) = -(x+1)^a v^m / m(m-1) \quad (m>0, m\neq 1), \tag{2}$$

giving the scoring rule

$$S(x,P) = \begin{cases} m^{-1}\left\{p(1)/p(0)\right\}^m & (x=0) \\ \{m(m-1)\}^{-1}\left[(m-1)(x+1)^a\left\{p(x+1)/p(x)\right\}^m \right. & \\ \qquad \left. -mx^a\left\{p(x)/p(x-1)\}\right\}^{m-1}\right] & (x>0). \end{cases} \tag{3}$$

3 Bayesian Model Selection

Let $\mathcal{M}$ be a finite or countable class of statistical models for the same observable $X \in \mathcal{X}$. Each $M \in \mathcal{M}$ is a parametric family, with parameter $\theta_M \in \Theta_M$, a d_M-dimensional Euclidean space; when M obtains, with parameter value θ_M, then X has distribution P_{θ_M}, with density function (probability mass function) $p_M(x \mid \theta_M)$. Having observed data $X=x$, we wish to make inference about which model $M \in \mathcal{M}$ generated the data. The Bayesian approach assigns, within each model M, a prior distribution Π_M, with density $\pi_M(\cdot)$ say, for its parameter θ_M. The associated *predictive distribution* P_M of X (given only the validity of model M, but no information on its parameter) has density function

$$p_M(x) = \int_{\Theta_M} p_M(x \mid \theta_M)\,\pi_M(\theta_M)\,d\theta_M. \tag{4}$$

Bayesian model selection, based on data x, involves comparison of $p_M(x)$ across the various models M. "Objective Bayesian" inference attempts to use standardised within-model priors Π_M intended to represent "prior ignorance". These are frequently "improper", with "density" $\pi_M(\cdot)$ that can not be normalised and has an arbitrary associated multiplier. This presents a serious problem for model selection, since that arbitrary multiplier, which can vary with M, persists into $p_M(x)$.

A way round this problem was proposed by Dawid and Musio (2015): if we compare the $\{P_M(x)\}$ using a homogeneous scoring rule, the arbitrary multipliers will not appear. Dawid and Musio (2015) conducted a detailed analysis of this approach for the case of continuous data and the Hyvärinen scoring rule (Hyvärinen, 2005). It was shown that this will typically deliver consistent selection of the true model.

In this work we investigate empirically, for a simple example, the validity of the above results when generalised to the case of discrete data. We shall use the scoring rule (3), and apply this to the choice between a Poisson and a Negative Binomial model.

4 Poisson model

Consider the Poisson model $X \sim \mathcal{P}(k\Lambda)$:

$$p(x \mid \lambda) = e^{-k\lambda}(k\lambda)^x/x! \quad (x=0,1,\ldots), \tag{5}$$

with conjugate prior $\Lambda \sim \Gamma(\alpha, \beta)$:

$$\pi(\lambda) = \frac{\beta^\alpha}{\Gamma(\alpha)} \lambda^{\alpha-1} e^{-\beta\lambda} \qquad (\alpha, \beta > 0). \tag{6}$$

The predictive density function is

$$p(x) = \frac{\Gamma(\alpha + x)}{\Gamma(\alpha) x!} (1-\phi)^\alpha \phi^x \tag{7}$$

with $\phi := k/(\beta + k)$. We find

$$\begin{aligned} S(0, P) &= m^{-1} \alpha^m \phi^m & (8)\\ S(x, P) &= \{m(m-1)\}^{-1} \left\{(m-1)\phi^m (x+1)^{a-m} (x+\alpha)^m \right. \\ &\quad \left. - m\phi^{m-1} x^{a-m+1} (x+\alpha-1)^{m-1}\right\} \quad (x > 0). & (9) \end{aligned}$$

4.1 Multiple observations

Suppose we have N independent and identically distributed observations $\boldsymbol{X}^N = (X_1, \ldots, X_N)$ from the above Poisson distribution. When, for $n \leq N$, we have observed $\boldsymbol{X}^{n-1} = \boldsymbol{x}^{n-1}$, the posterior distribution of Λ is

$$\Lambda \mid \boldsymbol{X}^{n-1} = \boldsymbol{x}^{n-1} \sim \Gamma\left\{\alpha + t_{n-1}, \beta + (n-1)k\right\},$$

with $t_n := \sum_{i=1}^N x_i$. The predictive distribution of X_n, given the previous observations $\boldsymbol{X}^{n-1} = \boldsymbol{x}^{n-1}$, is obtained from (8) and (9) on replacing x with x_n, α with $\alpha + t_{n-1}$, and β with $\beta + (n-1)k$. We henceforth use the standard improper prior, with $\alpha, \beta \downarrow 0$. The incremental contribution to the score is then given by

$$\begin{aligned} S_n^*(0, P) &= t_n^m / m n^m & (10)\\ S_n^*(x_n, P) &= (x_n + 1)^{a-m} t_n^m / m n^m \\ &\quad - x_n^{a-m+1} (t_n - 1)^{a-m+1} / (m-1) n^{m-1} \quad (x_n > 0). & (11) \end{aligned}$$

The *prequential score* is obtained by summing this from $n = 1$ to N.

5 Negative Binomial model

Now consider the alternative Negative Binomial model, $X \sim \mathcal{NB}(s; \Theta)$, having

$$p(x \mid \theta) = \frac{(s+x-1)!}{x!(s-1)!} (1-\theta)^s \theta^x \quad (x = 0, 1, \ldots), \tag{12}$$

with conjugate prior $\Theta \sim \beta(p, q)$:

$$\pi(\theta) = \frac{\Gamma(p+q)}{\Gamma(p)\Gamma(q)} \theta^{p-1} (1-\theta)^{q-1} \qquad (p, q > 0). \tag{13}$$

The predictive density is

$$p(x) = \frac{\Gamma(p+q)}{\Gamma(p)\Gamma(q)} \frac{(s+x-1)!}{x!(s-1)!} \frac{\Gamma(p+x)\Gamma(q+s)}{\Gamma(p+q+s+x)}. \tag{14}$$

Then we find

$$\begin{aligned} S(0, P) &= m^{-1} (sp)^m (p+q+s)^{-m} & (15)\\ S(x, P) &= \{m(m-1)\}^{-1} \left[(m-1)(x+1)^{a-m} \{(x+s)(x+p)\}^m (x+p+q+s)^{-m} \right. \\ &\quad \left. - m x^{a-m+1} \{(x+s-1)(x+p-1)\}^{m-1} (x+p+q+s-1)^{-m+1}\right]. & (16) \end{aligned}$$

5.1 Multiple observations

Now suppose we have already observed $\boldsymbol{X}^{n-1} = \boldsymbol{x}^{n-1}$. The posterior distribution of Θ is

$$\Theta \mid \boldsymbol{X}^{n-1} = \boldsymbol{x}^{n-1} \sim \beta \left\{ p + t_{n-1}, q + (n-1)s \right\}.$$

The predictive density of X_n, given the previous observations $\boldsymbol{X}^{n-1} = \boldsymbol{x}^{n-1}$, is obtained from (15) and (16) on replacing x with x_n, p with $p + t_{n-1}$, and q with $q + (n-1)s$. We henceforth use the standard improper prior, with $p, q \downarrow 0$. The incremental contribution to the prequential score is then given by:

$$S_n^*(0, P) = m^{-1} s^m t_{n-1}^m (t_{n-1} + ns)^{-m} \tag{17}$$

$$\begin{aligned} S_n^*(x_n, P) = \; & \{m(m-1)\}^{-1} \left[(m-1)(x_n+1)^{a-m}(x_n+s)^m t_n^m (t_n+ns)^{-m} \right. \\ & \left. - \, m x_n^{a-m+1}(x_n+s-1)^{m-1}(t_n-1)^{m-1}(t_n+ns-1)^{-m+1} \right]. \end{aligned} \tag{18}$$

6 Simulations

We generated observations from either the Poisson distribution (5) with $k = 1$, $\lambda = 10$, or the Negative Binomial distribution (12) with $s = 90$, $\theta = 0.1$. These both have mean 10, the former having variance 10, and the latter variance 11.1. We used, as the scoring rule, the special case of (3) having $a = m = 2$:

$$S(x, P) = \frac{1}{2}(x+1)^2 \left\{ \frac{p(x+1)}{p(x)} \right\}^2 - x^2 \left\{ \frac{p(x)}{p(x-1)} \right\} \delta(x > 0).$$

For each generating distribution we computed the excess of the cumulative prequential score for the wrong model over that for the correct model. These differences are shown, as a function of increasing data, in Figures 1 and 2 respectively. Each figure displays 10 sample sequences generated from the indicated distribution, as well as the average taken over a sample of 100 sequences.

In each case we see a clear linear upward trend, supporting the expectation of consistent model selection, although even with 1000 observations there is a non-negligible probability of a negative value, giving a misleading preference for the wrong model.

7 Conclusion

We have extended the Bayesian model selection methodology of Dawid and Musio (2015) to apply to problems with discrete data. We have conducted a simulation study to compare Poisson and Negative Binomial distributions. The results suggest that the method will consistently select the correct model as the number of data points increases.

Acknowledgments

Philip Dawid's research was supported through an Emeritus Fellowship from the Leverhulme Trust.

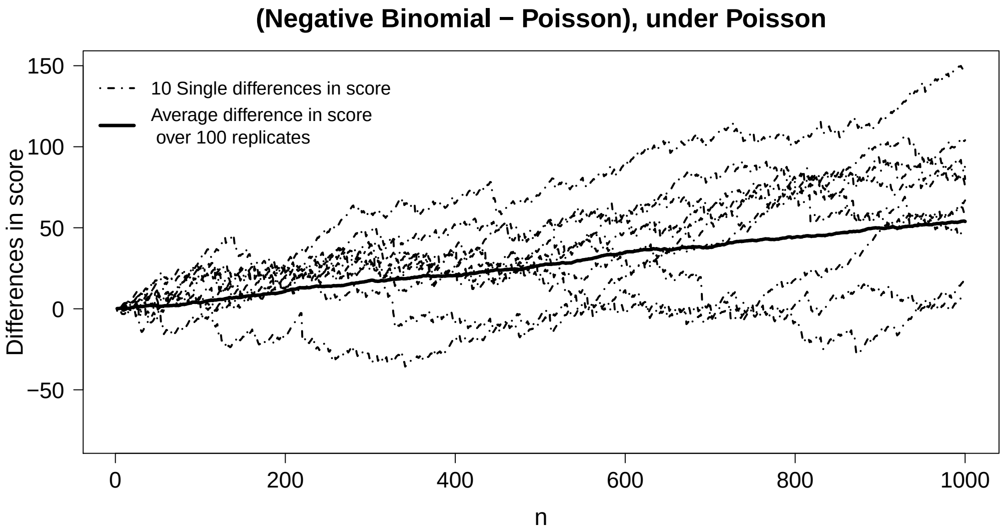

Figure 1: Data from Poisson distribution $\mathcal{P}(10)$

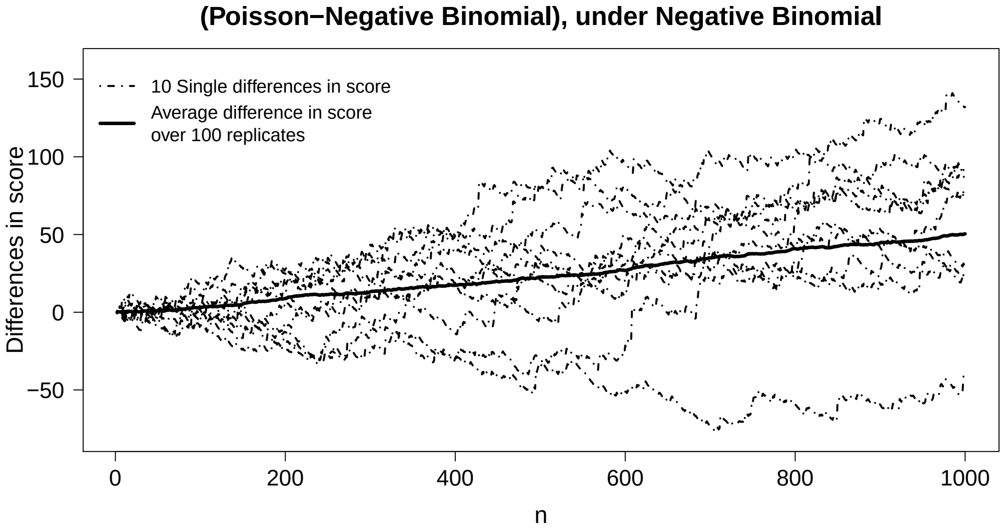

Figure 2: Data from Negative Binomial distribution $\mathcal{NB}(90; 0.1)$

References

Dawid, A. P. (2011). Posterior model probabilities. In *Philosophy of Statistics*, (ed. P. S. Bandyopadhyay and M. Forster), pp. 607–30. Elsevier, New York.

Dawid, A. P., Lauritzen, S. L. and Parry, M. (2012). Proper local scoring rules on discrete sample spaces. *Ann. Statist.* **40**, 593–608.

Dawid, A. P. and Musio, M. (2015). Bayesian model selection based on proper scoring rules (with Discussion). *Bayesian Analysis* **10**, 479–521.

Dawid, A. P., Musio, M. and Columbu, S. (2017). A note on Bayesian model selection for discrete data using proper scoring rules. *Statistics & Probability Letters* **129**, 101–106.

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

Parry, M. F., Dawid, A. P., and Lauritzen, S. L. (2012). Proper local scoring rules. *Ann. Statist.* **561**, 40–92.