

A comparison of model selection criteria¹

Un confronto tra criteri di scelta di un modello

Cinzia Carota²

Dipartimento di Statistica e Matematica Applicata "Diego De Castro"

University of Turin, Italy.

carota@econ.unito.it

Riassunto: In questo lavoro metteremo a confronto alcuni criteri bayesiani di selezione di un modello parametrico contro una alternativa non parametrica sul terreno della capacita' di risolvere l'inconsistenza esibita dal fattore di Bayes sotto certe condizioni. Di tale inconsistenza si fornisce una discussione critica preliminare.

Keywords: Bayes Factor, Fractional Bayes Factor, Inconsistency of the Bayes Factor, Pearson goodness of fit statistic.

1. Introduction and description of the problem

Consider an unknown distribution function Φ and a Dirichlet process prior, say $\mathcal{D}(c\Phi_0)$, with total mass parameter c and mean Φ_0 . Φ_0 is assumed to be a known distribution function and its corresponding probability density function will be denoted by $\phi_0(y)$. Given a random sample $y_1, \dots, y_T$, the problem is to test a parametric model, say M_0 , against the non parametric alternative represented by the random distribution function Φ . This alternative model will be called M_1 . It is well known that the Bayes factor in this setting can show itself severely inadequate. Actually, the chief difficulty arises when the parametric model M_0 is *nested* in the nonparametric alternative (no matter if M_0 corresponds to an absolutely continuous or to a discrete probability density function) if there are *no ties* in the data. In this case: a) the Bayes factor is independent of the data (Carota and Parmigiani, 1996) and b) as the number of distinct observations gets larger, the Bayes factor increasingly favors M_0 even when M_0 is false (Florens, Richard and Rolin, 1996). Put otherwise, its asymptotic "power" is zero. The next section shows that such an inconsistent behaviour cannot be explained by saying that the Dirichlet process prior expects to see ties in the data and introduces a penalty term in the Bayes factor (Florens, Richard and Rolin, 1996; Berger and Guglielmi, 2001).

All the results in this paper can be extended to the case where there are unknown parameters in M_0 and/or M_1 .

2. Results

Let us test $M_0 : \text{Poisson}(\theta_0)$ against $M_1 : \mathcal{D}(c\Phi_0)$ where Φ_0 corresponds to a ge-

¹Il lavoro è stato svolto con finanziamenti M.I.U.R. (60%)

²Indirizzo per corrispondenza: Dipartimento di Statistica e Matematica Applicata, Piazza Arbarello 8, 10122 Torino

ometric probability density function with parameter $1/(\theta_0 + 1)$, and suppose the first T positive natural numbers, $1, 2, \dots, T$, have been observed. Then there are no ties in the data and the Bayes factor is

$$BF = \frac{\Gamma(c+T)e^{-T\theta_0}\theta_0^{T(T+1)/2}}{c^T\Gamma(c)(\theta_0+1)^{-T}(\frac{\theta_0}{\theta_0+1})^{T(T+1)/2}\prod_{i=1}^T\Gamma(i+1)},$$

whose limit, as T increases, is zero.

On the other hand, if we test $M_0 : \text{Geometric}(\theta_0)$ against $M_1 : \mathcal{D}(c\Phi_0)$ where Φ_0 corresponds to a *Geometric*(θ_1),

$$BF = \frac{\Gamma(c+T)}{c^T\Gamma(c)}\left(\frac{\theta_0}{\theta_1}\right)^T\left(\frac{1-\theta_0}{1-\theta_1}\right)^{T(T+1)/2},$$

whose limit, as T increases, is zero when $\theta_0 > \theta_1$ and infinity when $\theta_0 < \theta_1$. Only when M_0 is *nested* into M_1 ($\theta_0 = \theta_1$), the Bayes factor reduces to $\Gamma(c+T)/(c^T\Gamma(c))$. Consequently the interpretation of $\Gamma(c+T)/(c^T\Gamma(c))$ as a penalty term due to the Dirichlet process prior can be misleading, since it does not imply an automatic preference for the parametric model when more and more distinct observations arrive. A *sufficient* condition for such an automatic preference is the concomitant nesting of M_0 into M_1 . To see, however, that this condition is not necessary for the Bayes factor to be inconsistent consider the problem of testing $M_0 : \text{prob}(j) = \frac{2(K-j+1)}{K(K+1)}$, $j = 1, \dots, K$ against $M_1 : \mathcal{D}(c\Phi_0)$ where Φ_0 corresponds to a discrete uniform density function: $\phi_0(j) = 1/K$, $j = 1, \dots, K$. Then, given the data set: $1, 2, \dots, T$,

$$BF = \frac{\Gamma(c+T)}{c^T\Gamma(c)}\prod_{i=1}^T\left(\frac{2(K-j+1)}{K+1}\right).$$

Provided that $c = T = K$ the BF values are huge for all $K > 5$.

3. A possible solution

The difficulties described before gave rise to various solutions such as different choices of the prior for the unknown distribution function Φ (see, e.g. Verdinelli and Wasserman, 1998, Berger and Guglielmi, 2001) or choices of an alternative test statistic, abandoning the Bayes factor (see e.g. Carota, Parmigiani and Polson 1996). The solution proposed here belongs to this second group. In particular we will focus on the way the Bayes factor inconsistency described in sections 1 and 2 is neutralized by this solution.

The condition of no ties, omnipresent when M_0 corresponds to an absolutely continuous density function, can be avoided by suitably grouping the data in $n < T$ arbitrary cells. Under the assumptions:

- (i) M_0 is nested in M_1 ($M_0 = \Phi_0$) and
- (ii) Φ is concentrated around Φ_0 (c large),

let us consider the cumulative distribution function, say Φ_0^* , obtained from the n cell probabilities $\phi_{01}, \dots, \phi_{0n}$ determined according to Φ_0 . Similarly let us denote by $\phi_1, \dots, \phi_n$ the n -dimensional vector of unknown cell probabilities corresponding to Φ .

Then, replacing the Dirichlet process mean Φ_0 by the cumulative distribution function Φ_0^* , it can be shown that an alternative testing tool with respect to the Bayes factor (actually based on a symmetric version of the Kullback -Leibler divergence between prior and pos-

terior distribution of Φ), admits the following representation (see Carota and Parmigiani, 1998):

$$\bar{\Delta}_L = \frac{1}{2} \sum_i^n [E(\phi_i|y) - E(\phi_i)] \left(\frac{\partial BF}{\partial \phi_i} \right) \Big|_{\phi=\phi_0} = \frac{1}{2} \frac{T}{c+T} \chi^2,$$

where the difference between posterior and prior expectation of the cell probabilities ϕ_i s multiplies the relative rate of change of the Bayes factor, while, in the last term, χ^2 represents the Pearson goodness of fit statistic. (This fact suggests that the preliminary grouping of the data could be driven by the criteria usually adopted to maximize the sensitivity of χ^2).

In the second term we can observe that the solution given by $\bar{\Delta}_L$ to the Bayes factor inconsistency has essentially the same nature of the solution given by a partial Bayes factor to the indeterminacy problems usually raised by improper priors in the so called "full" Bayes factor, BF . In this case, if prior information on one or both of the models is weak and is expressed as an improper prior distribution, the undetermined term occurring in the numerator of the partial Bayes factor occurs also in the denominator so that they cancel (see O'Hagan 1994, p. 200). Similarly, in our case, if the diverging term $\Gamma(c+T)/(\Gamma(c)c^T)$ occurs in the BF it will also occur in its first derivative. They will therefore cancel neutralizing the Bayes factor inconsistency. This consideration introduces section 4, while the last expression for $\bar{\Delta}_L$ will be discussed in section 5.

4. Remark

In this section we will note that a partial Bayes factor, such as the fractional one, in general does not solve the inconsistency of the "full" Bayes factor described in this paper. Once again, let us consider the null model $M_0 : prob(j) = \frac{2(K-j+1)}{K(K+1)}$, $j = 1, \dots, K$ but now assume it is nested into the alternative $M_1 : \mathcal{D}(c\Phi_0)$ (or that Φ_0 corresponds to $\phi_0(j) = prob(j)$). Then, given the data set $1, 2, \dots, T$, the fractional Bayes factor in favor of M_0 , say FBF , turns out to be:

$$FBF = \left(\Gamma(c+T) \prod_{j=1}^T \frac{\Gamma(b + c \frac{2(K-j+1)}{K(K+1)})}{\Gamma(c \frac{2(K-j+1)}{K(K+1)})} \right) / \left(c^T \Gamma(c+bT) \prod_{j=1}^T \left(\frac{2(K-j+1)}{K(K+1)} \right)^b \right).$$

For fixed values of b and c and large values of $T = K$ it assumes very quickly huge values.

On the contrary, the FBF performs well in the case where M_0 is the *Geometric*(θ_0) nested in the non parametric alternative described in section 2: it avoids both problems a) and b) of the corresponding BF.

5. Further lines of research and conclusion

Let us consider the expression $\bar{\Delta}_L = \frac{1}{2} \frac{T}{c+T} \chi^2$, where this Bayesian diagnostic tool of the adequacy of M_0 turns out to be increasing with χ^2 and decreasing with c . In particular, on the one hand, note that for a fixed value of χ^2 it becomes more and more favorable to M_0 as c gets larger. On the other hand, consider that the specification of c is traditionally related to the sample size in the sense that c is the prior weight on M_0 (often conceived as a sort of subjective summary of the credit given to M_0 by past observations) which will be compared with T , the weight of the current sample. These facts suggest a new line of research about the specification of c based on the following third consideration. Including c , the linear symmetric diagnostic $\bar{\Delta}_L$ could be used to face another, completely different, "inconsistency" problem affecting the χ^2 test as traditionally practised (see O'Hagan 1994). In fact, when $\Phi = \Phi_0$ (M_0 is the true model), a χ^2 test with size α , no matter how large T is, wrongly rejects M_0 with probability α . In other words, it cannot consistently accept the null hypothesis. From a strictly non Bayesian point of view one can argue that α should decrease as T increases, or, alternatively, that the χ^2 value should be corrected somehow, for instance as the BIC does. But here the Bayesian justification of the BIC needs itself a preliminary analysis of the BF corresponding to the grouping required to maximize the sensitivity of the χ^2 test! Viceversa, if one accept the challenge of suitably modelling the prior weight c , the Bayesian diagnostic $\bar{\Delta}_L$ can justify and do this correction in a direct and very flexible way.

References

- Berger J.O., Guglielmi A. (2001) Bayesian and Conditional Frequentist Testing of a Parametric Model Versus Nonparametric Alternatives, *Journal of the American Statistical Association*, 96 , 174-184.
- Carota, C., Parmigiani, G., Polson, N. G. (1996) Diagnostic Measures for Model Criticism, *Journal of the American Statistical Association*, 91 , 753-762.
- Carota C., Parmigiani G. (1996) On Bayes Factor for Nonparametric Alternatives, in *Bayesian Statistics* (vol.5), eds. J.M. Bernardo, J.O. Berger, A.P.Dawid and A.F.M. Smith, London: Oxford University Press, 507-511.
- Carota C., Parmigiani G. (1998) A Dirichlet process diagnostic for binomial goodness of fit *Test*, 7, N. 1, 133-145.
- Ferguson, T.S. (1973) A Bayesian Analysis of Some Nonparametric Problems *The Annals of Statistics*, 1, 209-230.
- Florens J.P., Richard J.F. and Rolin J.M. (1996) Bayesian Encompassing Specifications Tests of a Parametric model Against a Nonparametric Alternative, Technical Report 96.08, Universite' Catholique de Louvain, Institut de Statistique.
- O'Hagan, A. (1994). *Kendall's Advanced Theory of statistics, vol. 2B- Bayesian Inference*, London: Arnold.
- Verdinelli I, Wasserman L. (1998) Bayesian Goodness of Fit Testing Using Infinite Dimensional Exponential Families *The Annals of Statistics*, 20, 1203-1222.