

Divergence-based tests of statistical models

*Sulla scelta di modelli statistici
basata sulle divergenze*

Cinzia Carota

Dipartimento di Statistica e Matematica Applicata "Diego De Castro"
Piazza Arbarello 8, 10122, Torino, Italy.
carota@econ.unito.it

Riassunto: Si propone una estensione al contesto bayesiano non parametrico delle statistiche di bontà dell'adattamento di un modello ai dati note come "power divergence statistics" (Cressie e Read, 1984). Se si conviene di valutare l'adeguatezza del modello di interesse attraverso la divergenza tra una distribuzione a priori fortemente addensata intorno ad esso e la corrispondente distribuzione a posteriori e se questa legge di probabilità a priori è un processo di Dirichlet con parametro discreto, allora si riesce ad esprimere la famiglia delle power divergences in forma esplicita, si ottengono alcuni casi speciali interessanti e in vario modo legati alle tradizionali statistiche logaritmo del rapporto delle verosimiglianze e chi-quadrato di Pearson e, infine, si possono proporre alcune approssimazioni. L'esempio conclusivo discute la calibrazione degli elementi di tale famiglia.

Keywords: Bayesian diagnostics of goodness of fit, Cressie and Read power divergence family.

1. Introduction.

We are interested in evaluating how well a fixed model describes a given data set. The Bayesian approach to this problem often consists in the computation of an appropriate Bayes factor. This aim is here achieved by embedding the model of interest in a class of suitable alternative models and computing a divergence measure between a prior distribution very peaked around the model of interest and the corresponding posterior obtained from the given data set. More details on this general strategy for model criticism are in Carota et al. (1996). The elements of the random sample, denoted by $\mathbf{Y} = Y_1, \dots, Y_T$, are assumed to be independent and identically distributed conditional on a fixed value of their unknown distribution function F , i.e. exchangeable. The unknown F is taken to be a Dirichlet process, $\mathcal{D}_\alpha$, (Ferguson, 1973) with parameter given by the product of a pre-specified distribution function F_0 , and a given constant c , that is $\alpha(z) = cF_0(z)$. This implies $E(F(z)) = F_0(z)$ and $E(F^2(z)) = cF_0(z)(cF_0(z) + 1)/[c(c + 1)]$. The mean of the process is assumed to be discrete, and the mass parameter c , controlling the variance of the process, to be large, so as to express a high degree of belief in F_0 which is deputed to represent our model of interest. Eventually, F_0 can be specified up to a finite dimensional vector of parameters. Moreover, it is sometimes convenient to express the parameter c as a multiple γ of the sample size, $c = \gamma T$, $\gamma > 0$. Given $\mathbf{Y} = \mathbf{y}$, the corresponding posterior Dirichlet process will be denoted by $\mathcal{D}_{\alpha|\mathbf{y}}$.

Under these assumptions we provide a "power divergence" family of Bayesian diagnostics of the adequacy of the model F_0 , whose characteristics will be discussed comparatively with respect to the power divergence goodness-of-fit statistics by Cressie and Read (1984)

and, particularly, with respect to the most important traditional statistics belonging to this class.

This paper, which is a first contribute in the above delineated direction, is organized as follows: section 2 recalls the definition and the main characteristics of the Cressie and Read power divergence family, section 3 quickly sketches some extensions of this family based on more general measures of divergence and disparity, section 4 contains our power divergence family of Bayesian diagnostics, and, finally, section 5 presents an example where the calibration of the members of this family is discussed.

2. The power divergence family.

Given a model of interest, many classical tests of goodness-of fit can be reduced to testing a hypothesis about the parameter of a multinomial distribution:

$$H_0 : p_i = \pi_i, \quad i = 1, \dots, k,$$

where under the "null" model the unknown cell probabilities p_i are specified to be π_i , and k denotes the number of mutually exclusive cells or classes adopted. Given a random sample $\mathbf{Y}$, the most used tools to test H_0 are by far the Pearson's statistic X^2 and the loglikelihood ratio statistic G^2 , which compare in a different way the expected frequencies in each cell, given by $T\pi_i$ (or $T\hat{\pi}_i$, if the model of interest, and consequently the vector π , contains nuisance parameters which are estimated from the data), against the observed frequencies X_i , $X_i = \{\# \text{ times that values of } Y_1, \dots, Y_T \text{ fall in the cell } i\}$. Under regularity conditions these two statistics are asymptotically equivalent. Other goodness of fit statistics extensively studied in the literature are: the Freedman-Tukey statistic $F^2 = 4 \sum_1^k (\sqrt{X_i} - \sqrt{T\hat{\pi}_i})^2$ (Freedman and Tukey, 1950; Bishop et al., 1975), the minimum discrimination information statistic $M^2 = 2 \sum_1^k T\hat{\pi}_i \log(T\hat{\pi}_i/X_i)$ (Kullback 1959), and the Neyman statistic $N^2 = \sum_1^k (X_i - T\hat{\pi}_i)^2/X_i$, (Neyman 1949). The last two appear as modified versions respectively of $G^2 = 2 \sum_1^k X_i \log(X_i/(T\hat{\pi}_i))$ and $X^2 = \sum_1^k (X_i - T\hat{\pi}_i)^2/(T\hat{\pi}_i)$, while F^2 is connected with the Matusita distance between two probability distributions $\mathbf{p} = p_1, \dots, p_k$ and $\mathbf{q} = q_1, \dots, q_k$ (Matusita, 1954),

$$M(\mathbf{p} : \mathbf{q}) = \left(\sum_i^k (\sqrt{p_i} - \sqrt{q_i})^2 \right)^{1/2},$$

i. e. a discrete version of the Hellinger distance.

Cressie and Read (1984) unified and extended this literature by introducing the power-divergence family of goodness-of-fit statistics,

$$2TI^\lambda(\frac{\mathbf{X}}{T} : \hat{\pi}) = \frac{2}{\lambda(1+\lambda)} \sum_1^k X_i \left[\left(\frac{X_i}{T\hat{\pi}_i} \right)^\lambda - 1 \right], \quad -\infty < \lambda < \infty, \quad (1)$$

where the cases $\lambda = 0$ and $\lambda = -1$ are defined by considering the continuous limits of (1) as $\lambda \rightarrow 0$ and $\lambda \rightarrow -1$, respectively. Then (1) is a measure of divergence between the empirical probability distribution, $\mathbf{X}/T = X_1/T, \dots, X_k/T$, and the hypothesized distribution, $\hat{\pi} = \hat{\pi}_1, \dots, \hat{\pi}_k$, continuous in λ . All the before mentioned traditional goodness-of-fit

statistics are members of this single real-valued parameter family:

$$X^2 = 2TI^1\left(\frac{\mathbf{X}}{T} : \hat{\pi}\right), \quad G^2 = \lim_{\lambda \rightarrow 0} 2TI^\lambda\left(\frac{\mathbf{X}}{T} : \hat{\pi}\right), \quad F^2 = 2TI^{-1/2}\left(\frac{\mathbf{X}}{T} : \hat{\pi}\right),$$

$$M^2 = \lim_{\lambda \rightarrow -1} 2TI^\lambda\left(\frac{\mathbf{X}}{T} : \hat{\pi}\right), \quad \text{and} \quad N^2 = 2TI^{-2}\left(\frac{\mathbf{X}}{T} : \hat{\pi}\right).$$

Important results about the family (1) are summarized in the rest of this section.

- Under the assumption of a fixed number of cells k , when the null model is true, all members of the power divergence family are asymptotically chi-squared with degree of freedom given by k minus the number of efficiently estimated nuisance parameters, s say, minus 1, denoted by χ_{k-s-1}^2 .

When a local alternative $H_{1,T} : p_i = \pi_i + \delta/\sqrt{T}$, $i = 1, \dots, k$, $\sum \delta_i = 1$, is assumed to be true, the limiting distribution becomes a non-central chi-squared, $\chi_{k-s-1}^2(b)$, with parameter independent of λ ($b = \sum \delta_i^2/\pi_i$). Consequently all power divergence family members have the same Pitman asymptotic relative efficiency. On the contrary, when considering non-local alternatives that, as T diverges, are a fixed distance away from the null model, the asymptotic distribution becomes a standard normal, with standardizing mean and variance depending on λ and the true cell probabilities π . In general we cannot establish which λ is dominant, unless the Bahadur efficiency is considered. In this case, when $s = 0$, the loglikelihood ratio statistics G^2 ($\lambda = 0$) is optimal.

- Under the "sparseness" assumptions, that is when the number of cells k increases at a similar rate to the sample size so that T/k remains finite, the asymptotic null distribution of $2TI^\lambda(\frac{\mathbf{X}}{T} : \hat{\pi})$ for $\lambda > -1$ is normal with mean and variance dependent on both λ and π . No general rule emerges as to which λ is the most efficient, except when we are dealing with a local alternative and an equiprobable null model, $H_0 : p_i = \pi_i = 1/k$ for all i , in which case the Pearson X^2 ($\lambda = 1$) is optimal.

- The value of λ is also influential in: a) finite sample approximation to chi-squared distribution with regard to both critical value and moments, b) power in finite samples, c) sensitivity to observed frequencies in each cell. (Extreme values of $|\lambda|$ are useful for detecting a very large or a very small deviation from an expected cell frequency in a single cell, and, in general, large negative values of λ emphasize ratios of $x_i/(T\pi_i)$ near to zero while large positive values of λ emphasize large ratios of $x_i/(T\pi_i)$, where x_i is the realization of the random variable X_i .)

- Provided that the minimum expected cell frequency is no less than 1 and $T \geq 10$, the choice of $\lambda = 2/3$ is the most competitive in the power divergence family when the above criteria are jointly considered.

Moreover the link of the traditional goodness-of fit statistics through the parameter λ provides a useful perspective from which the Authors interestingly describe their relative merits and clarify how a decision can be made when some of them lead to reject the null model, whereas others confirm it. This link is at the same time a formidable unifying factor of the fragmented literature on these issues.

Inspired by this approach, we will consider power-divergence measures for the comparison of stochastic processes, and explore the consequences of such an extension of the family (1) in a Bayesian nonparametric context. More specifically, we provide measures of power divergence between the posterior and the prior Dirichlet processes described in section 1. These results are presented in section 4; the next section briefly illustrates how

the power divergence family is linked to the literature on more general types of divergence.

3. ϕ -divergences and ϕ -disparity statistics.

The power divergence family can be enclosed as a special case in the class of convex directed divergences considered by Ali and Silvey (1966) and Csiszár (1978) for the comparison of two probability measures P and Q on the same probability space,

$$\int_{-\infty}^{+\infty} q(z) \phi\left(\frac{p(z)}{q(z)}\right) d\mu(z), \quad (2)$$

where $p(z)$ and $q(z)$ denote the corresponding densities, ϕ is a convex function, and μ is a measure. Very general P and Q can be considered in this frame by introducing the generalized Radon-Nikodym derivative of P with respect to Q (generalized in the sense that it is allowed to take the value $+\infty$ on a Q -null set if P has a component singular with respect to Q) and defining a corresponding generalized expectation. When considering discrete probability measures on a partition of the data space, in turn the ϕ -divergence coefficients (2) can be viewed as special cases of the ϕ -disparity statistics proposed by Menendez et al. (1998a). This last extension is obtained by restricting the convexity of functions $\phi(t)$, $t > 0$, to local convexity at $t = 1$ and monotonicity on intervals $(0, 1)$ and $(1, \infty)$. For asymptotic results on ϕ -divergence and ϕ -disparity goodness of fit statistics and further generalizations see e.g. Menendez et al. (1995), (1996), (1998b), (2001), Morales et al. (1995), (1997), (2003), Cressie and Pardo (2000), (2002), Cressie et al. (2003), Pardo (2006), Pardo et al. (2006). Various other extensions of the power divergence family have been proposed over the last twenty years; two recent references are Chen et al. (2004) which adapt (1) to analyze grouped survival data and Baranov and Baranov (2005) where a multidimensional analogue of I_λ is suggested for testing homogeneity of m independent polynomial schemes.

With respect to the class of ϕ -divergence coefficients (2) Read and Cressie (1988, p.110) judge their power divergence to provide "sufficient generality together with a more tangible and descriptive structure through the single real-valued parameter λ "; from our nonparametric Bayesian perspective we can add that this family is exactly the one allowing us to provide the next results in explicit form.

4. Results.

Proposition 1 gives the directed power divergence, $I_\lambda(\text{post} : \text{prior})$, between posterior and prior Dirichlet processes, $\mathcal{D}_{\alpha|y}$ and $\mathcal{D}_\alpha$, when the parameter $\alpha(z) = cF_0(z)$ is assumed to be absolutely continuous with respect to the counting measure and c is large. As before the cases $\lambda = 0$ and $\lambda = -1$ are defined by the continuous extensions.

Proposition 1.

$$I^\lambda(post : prior) = \frac{1}{\lambda(1 + \lambda)} \left\{ \frac{\frac{\Gamma(c+T)}{\Gamma(c+(1+\lambda)T)} \prod_1^r \frac{\Gamma(c_i+(1+\lambda)x_i)}{\Gamma(c_i+x_i)}}{\left[\frac{\Gamma(c)}{\Gamma(c+T)} \prod_1^r \frac{\Gamma(c_i+x_i)}{\Gamma(c_i)} \right]^\lambda} - 1 \right\} \quad (3)$$

where $c_i = c[F_0(y_i) - F_0(y_i^-)] = c \pi_i$, r is the number of distinct values in the observed sample $y_1, \dots, y_T$, x_i denotes the observed frequency of the i th distinct value ($\sum_1^r x_i = T$), and λ is such that $\min_{1 \leq i \leq r} (c_i + (1 + \lambda)x_i) > 0$.

A "symmetrized" with respect to *post* and *prior*, or "undirected", form of the power divergence family can be obtained considering, for instance, $\bar{I}^\lambda = \frac{1}{2}(I^\lambda(post : prior) + I^\lambda(prior : post))$ where

$$I^\lambda(prior : post) = \frac{1}{\lambda(1 + \lambda)} \left\{ \frac{\frac{\Gamma(c)}{\Gamma(c-\lambda T)} \prod_1^r \frac{\Gamma(c_i-\lambda x_i)}{\Gamma(c_i)}}{\left[\frac{\Gamma(c)}{\Gamma(c+T)} \prod_1^r \frac{\Gamma(c_i+x_i)}{\Gamma(c_i)} \right]^{-\lambda}} - 1 \right\}, \quad (4)$$

and λ is such that the arguments of all the involved Γ functions are positive.

Note that the divergence (4) can be derived from (3) by observing that $I^\lambda(prior : post) = I^{-\lambda-1}(post : prior)$.

Special cases and approximations.

Interesting special cases of $I^\lambda(post : prior)$ and $\bar{I}^\lambda$ are obtained for fixed values of λ .

$\lambda = 1$.

The first term in the braces of formula (3) becomes the ratio of the marginal distribution of the random vector $\mathbf{Y}$ according to the posterior Dirichlet process, that is the *predictive distribution* $p(\mathbf{y}|\mathbf{y}) = \int p(\mathbf{y}|F)d\mathcal{D}_{\alpha|\mathbf{y}}$, and the marginal distribution according to the prior Dirichlet process, $p(\mathbf{y}) = \int p(\mathbf{y}|F)d\mathcal{D}_\alpha$,

$$I^1(post : prior) = \frac{1}{2} \left[\frac{p(\mathbf{y}|\mathbf{y})}{p(\mathbf{y})} - 1 \right].$$

A different interpretation of $I^1(post : prior)$ is as the variance of the ratio *post/prior* with respect to the prior, that is the most obvious measure of closeness to its prior mean 1.

$\lambda = 0$ and $\lambda = -1$.

The expressions of the symmetric divergence $\bar{I}^\lambda$ corresponding to these two values of λ collapse into a single one,

$$\lim_{\lambda \rightarrow 0} \bar{I}^\lambda = \lim_{\lambda \rightarrow -1} \bar{I}^\lambda = \frac{1}{2} \sum_1^r T_i [\Psi(c_i + x_i) - \Psi(c_i) + \Psi(c) - \Psi(c + T)], \quad (5)$$

where Ψ denotes the digamma function. This is the symmetrized Kullback-Leibler or Jeffreys divergence (Jeffreys, 1948) between prior and posterior Dirichlet processes (Carota 2005). Let us call it $\bar{I}^{0,-1}$. With respect to $4\bar{I}^{0,-1}$ the log-likelihood ratio goodness of fit

statistic G^2 represents an asymptotic lower bound, that is for a large T :

$$4\bar{I}^{0,-1} \leq G^2.$$

Furthermore, when the discrete prior mean of the Dirichlet process is finitely supported, the symmetric divergence $\bar{I}^{0,-1}$ is linearly approximated around $F = F_0$ by

$$\frac{T}{2(c+T)} X^2,$$

where X^2 is the Pearson's goodness of fit statistics (details are in Carota, 2005, Carota and Parmigiani, 1998).

On the other hand, if we consider the directed power divergence $I^\lambda(post : prior)$, then

$$\begin{aligned} \lim_{\lambda \rightarrow 0} I^\lambda(post : prior) &= \\ &= \sum_1^r T_i [\Psi(c_i + x_i) - \Psi(c + T)] - \sum_1^r \log\left(\frac{\Gamma(c_i + x_i)}{\Gamma(c_i)}\right) - \log\left(\frac{\Gamma(c)}{\Gamma(c + T)}\right). \end{aligned} \quad (6)$$

This is the explicit form of the diagnostic suggested by Carota, Parmigiani and Polson (1996). The Kullback-Leibler divergence in the opposite direction, $\lim_{\lambda \rightarrow -1} I^\lambda(post : prior) = \lim_{\lambda \rightarrow 0} I^\lambda(prior : post)$, is easily derivable from (5) and (6).

$$\lambda = -\frac{1}{2}.$$

In this case $I^\lambda(post : prior) = \bar{I}^\lambda$, being $\lambda = -1 - \lambda$, and

$$I^{-\frac{1}{2}}(post : prior) = 4 \left\{ 1 - \left[\frac{Be(c, \frac{1}{2}T) \prod_1^r Be(c_i + \frac{1}{2}x_i, \frac{1}{2}x_i)}{Be(c + \frac{1}{2}T, \frac{1}{2}T) \prod_1^r Be(c_i, \frac{1}{2}x_i)} \right]^{1/2} \right\},$$

where $Be()$ denotes the Beta function.

$$\lambda = -2.$$

$$I^{-2}(post : prior) = \frac{1}{2} \left[\frac{\Gamma(c) \prod_1^r Be(c_i + x_i, c_i - x_i)}{Be(c + T, c - T) \prod_1^r \Gamma(c_i)} - 1 \right].$$

Finally we conclude the special cases noting that, under our assumptions, the expectation on which are based the Chernoff's measure of discriminatory information, whose general expression is

$$-\log \inf_{0 < t < 1} \int q^t(z) p^{1-t}(z) d\mu(z),$$

and the additive and non additive directed divergences (7.18) and (7.19) in Read and Cressie (1988) can be expressed as

$$\left[\frac{\Gamma(tT) Be^t(c, T) \prod_1^r Be(c_i + (1-t)x_i, tx_i) \Gamma^t(x_i)}{\Gamma^t(T) Be(c + (1-t)T, tT) \prod_1^r Be^t(c_i, x_i) \Gamma(tx_i)} \right].$$

Let us now turn to consider possible approximations of $I^\lambda(\text{post} : \text{prior})$. If we write it according to Read and Cressie (1988, section 6.6), the Taylor-series expansion about the value $\text{post}/\text{prior} = 1$ assigns to $I^1(\text{post} : \text{prior})$ the same approximating role played by the Pearson's X^2 statistic in the family (1). If, alternatively, we expand $I^\lambda(\text{post} : \text{prior})$ about $F = F_0$, then the linear approximation of $I^\lambda(\text{post} : \text{prior})$ turns out to be:

$$\frac{1}{\lambda(1+\lambda)}[(BF)^\lambda\{1 + \lambda E(p - \pi|\mathbf{y})S\} - 1],$$

where BF denotes the Bayes factor in favour of the parametric model F_0 against the non-parametric alternative $F \sim \mathcal{D}_\alpha$, $E(p - \pi|\mathbf{y})$ is the posterior expected value of the difference between $\mathbf{p} = p_1, \dots, p_r = [F(y_1) - F(y_1^-)], \dots, [F(y_r) - F(y_r^-)]$ and its null value $\pi = \pi_1, \dots, \pi_r$, and S denotes the score function. When F_0 is finitely supported $E(p - \pi|\mathbf{y})S$ reduces to $X^2T/(2c + 2T)$; it is worth noting that a particular mode of the "sparseness" assumptions, that is when r grows with T and $T/r = 1$ (*no ties*), can cause difficulties with the Bayes factor in our setting, rather than with the Pearson's statistic X^2 . Examples of these difficulties (essentially cases of inconsistency of the Bayes factor) are shown in Carota (2006).

Finally we observe that the Jensen inequality provides a lower bound for $I^\lambda(\text{post} : \text{prior})$,

$$I^\lambda(\text{post} : \text{prior}) \geq \frac{1}{\lambda(1+\lambda)}\left\{\left[\frac{p(\mathbf{y}|\mathbf{y})}{p(\mathbf{y})}\right]^\lambda - 1\right\},$$

where the ratio of *predictive/marginal* is raised to the power λ .

5. Example.

We reconsider the Time Passage and Memory Recall example in Read and Cressie (1988, pp. 12 and 91-2), based on the Haberman's data (Haberman, 1978, pp. 2-23) and aimed at illustrating how their power divergence statistic works for different choices of λ . The possible relationship between the ability to recall specific events and the amount of time that has passed since the event has occurred is analyzed. The left side of Table 1 provides the observed frequencies of subjects who indicated one stressful event between 1 and 18 months prior to interview (column 2), and the expected frequencies under the null model

$$H_0 : \log(\pi_i) = \alpha + \beta i \quad i = 1, \dots, 18, \quad (7)$$

that is a loglinear time-trend model with unknown nuisance parameters, α and β , estimated by using the maximum likelihood method (column 3).

Cressie and Read use this example to show that the traditional goodness-of-fit statistics do not always agree about the rejection of the null model essentially for two reasons: the approximation of the critical value ($\chi_{16,0.05}^2 = 26.3$) is not equally adequate for all values of λ , and different choices of λ measure different types of departures from H_0 . Here the Haberman's null model is assumed to specify only the mean of the true distribution of the data and there is some variability around it controlled by the parameter γ , that is a "first" stage is placed before (7) as follows:

$$Y|F, F \sim D_{cF_0}, F_0 = \sum_j^i \pi_j, c = \gamma T, \text{ and } \log(\pi_i) = \alpha + \beta i, \quad i = 1, \dots, 18.$$

For simplicity the nuisance parameters are treated as if they were known and equal to their maximum likelihood estimated values. There are however no difficulties to consider them as random variables endowed with a prior distribution. Under these assumptions, the right side of Table 1 shows the values of the power divergence $I^\lambda(\text{post} : \text{prior})$ corresponding to a range of λ and γ values.

Table 1: *Time Passage and Memory Recall data. Observed and expected frequencies (left side), and $I^\lambda(\text{post} : \text{prior})$ values corresponding to different values of λ and γ (right side).*

<i>time</i>	<i>x_i</i>	<i>T$\hat{\pi}_i$</i>	γ	λ	$I^\lambda(\text{post} : \text{prior})$	γ	λ	$I^\lambda(\text{post} : \text{prior})$
1	15	15.171	1	1.00	340.27	20	1.00	0.87
2	11	13.952	1	0.66	49.48	20	0.66	0.69
3	14	12.831	1	0.50	22.35	20	0.50	0.63
4	17	11.800	1	0.00	4.67	20	0.00	0.53
5	5	10.852	1	-0.5	3.11	20	-0.5	0.50
6	11	9.980	1	-0.99	8.25	20	-0.99	0.55
7	10	9.178	5	1.00	9.44	30	1.00	0.50
8	4	8.440	5	0.66	4.22	30	0.66	0.43
9	8	7.762	5	0.50	3.10	30	0.50	0.40
10	10	7.138	5	0.00	1.73	30	0.00	0.36
11	7	6.565	5	-0.5	1.50	30	-0.5	0.35
12	9	6.037	5	-0.99	2.04	30	-0.99	0.37
13	11	5.552	10	1.00	2.54	40	1.00	0.35
14	3	5.106	10	0.66	1.62	40	0.66	0.31
15	6	4.696	10	0.50	1.36	40	0.50	0.30
16	1	4.318	10	0.00	0.98	40	0.00	0.27
17	1	3.971	10	-0.5	0.91	40	-0.5	0.27
18	4	3.652	10	-0.99	1.07	40	-0.99	0.28

Note also that the values of the symmetric divergence $\bar{I}^{0,-1}$ corresponding to $c = 1, 5, 10, 20, 30, 40$ respectively are of 6.68, 1.90, 1.03, 0.54, 0.36 and 0.28.

Of course, the $I^\lambda(\text{post} : \text{prior})$ values decrease for all λ as γ increases, while, for a fixed γ , they decrease to a minimum (roughly corresponding to $\lambda = -0.5$) and then start to increase as λ increases. Prudentially λ has been restricted to the interval $(-1, 1]$, but enlarging this range we can observe a behaviour of $I^\lambda(\text{post} : \text{prior})$ qualitatively similar to that of $2TI^\lambda(\mathbf{X}/T : \hat{\pi})$, in the sense that its magnitude increases more rapidly for large negative values than for large positive values of λ .

As regards the calibration of the $I^\lambda(\text{post} : \text{prior})$ values we can conveniently extend the scheme proposed by McCulloch (1989). Let $B(p)$ denote the Bernoulli distribution that assigns probability p to an event. Given $I^\lambda(P : Q) = h$ we find $q(h)$ such that $I^\lambda(B(.5) : B(q(h))) = h$. The number $q(h)$ is our calibration of h . The calibration tell us that, as measured by the power divergence, the difference h , between P and Q , is the same as that between $B(.5)$ and $B(q(h))$. For example, if $h = 2.54$, as in the case ($\gamma = 10, \lambda = 1$), being the corresponding q given by $q(2.54) = 0.957$, then the divergence between *post* and *prior* is large because it is the same as the divergence $I^1(B(.5) : B(0.957))$. If $h =$

0.35, as in the cases $(\gamma = 40, \lambda = 1)$ and $(\gamma = 30, \lambda = -0.5)$, being the corresponding q values respectively of 0.821 and 0.875, the divergence of 0.35 between *post* and *prior* is smaller in the first case (because it is the same as the divergence $I^1(B(.5) : B(0.821))$) than in the second (where it is the same as the divergence $I^{-0.5}(B(.5) : B(0.875))$). It is clearly explained in Cressie and Read (1988) that the choice of λ should be made before looking at the data, being essentially the choice of an alternative model. ("If λ is chosen to highlight some characteristic after observing the sample, then any significance test will not have the specified size.") To this we add that the choice of γ must of course be made before looking at the data. This is the only subjective assumption actually influencing the decision about the null model in a way perfectly parallel to the usual specification of the size of the test (Carota 2005, remark 4).

Table 2 shows a range of large q values with the corresponding $I^\lambda(B(.5) : B(q))$ values when $\lambda \in [-1, 1]$; the last column contains $\bar{I}^{0,-1}(B(.5) : B(q))$.

Table 2: Values of $I^\lambda(B(0.5) : B(q))$ and $\bar{I}^{0,-1}(B(.5) : B(q))$ for different values of λ and q .

q	I^1	$I^{0.66}$	$I^{0.5}$	I^0	$I^{-0.5}$	I^{-1}	$\bar{I}^{0,-1}$
0.90	0.89	0.72	0.65	0.51	0.42	0.37	0.44
0.95	2.13	1.48	1.26	0.83	0.61	0.49	0.66
0.99	12.13	5.49	3.85	1.61	0.90	0.65	1.13

References

- Ali S. M., Silvey S.D. (1966) A general class of coefficients of divergence of one distribution from another. *Journal of the Royal Statistical Society*, B, 28, 131-142.
- Baranov A. P., Baranov Y. A. (2005) A power divergence test in the problem of sample homogeneity for large numbers of outcomes and trials. *Discrete Mathematics and Applications*. 15, 211-240.
- Bishop Y.M.M., Fienberg S.E., Holland P.W. (1975) *Discrete Multivariate Analysis: theory and practice*. Cambridge, MA, The MIT Press.
- Carota C. (2005) Symmetric diagnostics for the analysis of residuals in regression models, *Biometrika* 92, 787-799.
- Carota C. (2006) Some faults of the Bayes factor in nonparametric model selection. *Statistical Methods and Applications* (to appear).
- Carota C., Parmigiani, G. (1998). A Dirichlet process elaboration diagnostic for binomial goodness of fit. *Test*, 7, 133-145.
- Carota C., Parmigiani G., Polson N. G. (1996). Diagnostic measures for model criticism. *Journal of the American Statistical Association*, 91, 753-762.
- Chen H.S., Lai K., Ying Z. (2004) Goodness-of-fit tests and minimum power divergence estimators for survival data. *Statistica Sinica*, 14, 231-248.
- Cressie N., Pardo L. (2000) Minimum ϕ -divergence estimator and hierarchical testing in loglinear models. *Statistica Sinica*, 10, 867-884.
- Cressie N., Pardo L. (2002) Model checking in loglinear models using ϕ -divergences and MLE. *Journal of Statistical Planning Inference*, 103, 437-453.

- Cressie N., Pardo L., Pardo M. (2003) Size and power considerations for testing loglinear models using ϕ -divergence test statistics. *Statistica Sinica*, 13, 555-570.
- Cressie N. and Read T. R. C. (1984) Multinomial goodness of fit tests *Journal of the Royal Statistical Society*, B, 46, 440-464.
- Csizár I. (1978) Information measures: a critical survey. In *Transactions of the seventh Prague Conference on Information Theory, Statistical Decision Functions and Random Processes and of the 1974 European Meeting of Statisticians*, Volume B, 73-86. Dordrecht, Reidel.
- Ferguson T.S. (1973) A Bayesian Analysis of Some Nonparametric Problems. *The Annals of Statistics*, 1, 209-230.
- Freeman M.F., Tukey J.W. (1950) Transformations related to angular and the square root. *Annals of Mathematical Statistics*, 21, 607-611.
- Haberman S.J. (1974) *Analysis of Qualitative Data, Volume 1*. New York. Academic Press.
- Jeffreys H. (1948) *Theory of Probability* (second edition). Oxford University Press, London.
- Kullback S. (1959) *Information Theory and Statistics*. John Wiley, New York.
- McCulloch R. E. (1989). Local model influence. *Journal of American Statist. Ass.*, 84, 473-478.
- Menéndez M., Morales D., Pardo L., Vajda I. (1995) Divergence-based estimation and testing of statistical models of classifications. *Journal of Multivariate Analysis*, 54, 329-354.
- Menéndez M., Morales D., Pardo L., Vajda I. (1996) About divergence-based goodness-of-fit tests in the Dirichlet-multinomial model. *Communications in statistics- Theory Meth.*, 25, 1119-1133.
- Menéndez M., Morales D., Pardo L., Vajda I. (1998a) Two approaches to grouping of data and related disparity statistics. *Communications in statistics- Theory Meth.*, 27 (3), 609-633.
- Menéndez M., Morales D., Pardo L., Vajda I. (1998b) Asymptotic distributions of ϕ -divergences of hypothetical and observed frequencies on refined partitions. *Statistica Neerlandica*, 52, 71-89.
- Menéndez M., Morales D., Pardo L., Vajda I. (2001) Approximations to powers of ϕ -disparity goodness of fit tests. *Communications in statistics- Theory Meth.*, 30, 105-134.
- Morales D., Pardo L., Vajda I. (1995) Asymptotic divergence of estimates of discrete distributions. *Journal of Statistical Planning and Inference*, 48, 347-369.
- Morales D., Pardo L., Vajda I. (1997) Some new statistic for Testing Hypotheses in Parametric Models. *Journal of multivariate analysis*, 62, 137-168.
- Morales D., Pardo L., Vajda I. (2003) Asymptotic laws for disparity statistics in product multinomial models. *Journal of multivariate analysis*, 85, 335-360.
- Neyman J. (1949) Contribution to the theory of χ^2 test. In *Proceedings of the first Berkeley Symposium on Mathematical Statistics and Probability* 239-273.
- Pardo L. (2006) *Statistical Inference based on Divergence measures*, Chapman and Hall CRC, Boca Raton.
- Pardo J.A., Pardo L., Pardo M. (2006) Testing in logistic regression models based on ϕ -divergences measures *Journal of Statistical Planning and Inference* 136, 982-1006.
- Read T. R. C., Cressie N. (1988) *Goodness of fit statistics for discrete multivariate data* Springer. New York.