Computational Statistics and Data Analysis 124 (2018) 87–103

Contents lists available at ScienceDirect

Computational Statistics and Data Analysis

journal homepage: www.elsevier.com/locate/csda

A family of the information criteria using the phi-divergence for categorical data☆

Haruhiko Ogasawara

Otaru University of Commerce, 3-5-21, Midori, Otaru 047-8501, Japan

ARTICLE INFO

Article history:
Received 2 August 2017
Received in revised form 27 February 2018
Accepted 3 March 2018
Available online 10 March 2018

Keywords:
Power divergence
Risk
Model selection
Asymptotic bias
Akaike information criterion

ABSTRACT

The risk of the phi-divergence of a statistical model for categorical data is defined using two independent sets of data. The asymptotic bias of the phi-divergence based on current data as an estimator of the risk is shown to be equal to the negative penalty term of the Akaike information criterion (AIC). Though the higher-order asymptotic bias is derived, the higher-order bias depends on the form of the phi-divergence and the estimation method of parameters using a possible different form of the phi-divergence. An approximation to the higher-order bias is obtained based on the simple result of the saturated model. The information criteria using this approximation yield improved results in simulations for model selection. Some cases of the phi-divergences show advantages over the AIC in simulations.

1. Introduction

Various information criteria have been proposed. Many of them are based on the likelihood of parameters in a statistical model, whose typical cases are the Akaike information criterion (AIC, Akaike, 1973), the Takeuchi information criterion (TIC, Takeuchi, 1976; for the TIC see e.g., Burnham and Anderson, 2010, pp. 65–66) and the Bayes information criterion (BIC, Schwarz, 1978). The Mallows (1973) C_p for model selection in linear regression takes a least squares (LS) form, which is also seen as the Gauss discrepancy (Linhart and Zucchini, 1986, p. 18) based on a likelihood. It is known that the C_p is asymptotically equivalent to the AIC.

In covariance structure analysis, the normal-theory (NT) or asymptotically-distribution-free (ADF) generalized LS (GLS) criteria for model selection have also been proposed (Browne and Cudeck, 1989; Yanagihara et al., 2010; Ogasawara, 2017), which are asymptotically equal to the AIC and TIC under some conditions. The criteria based on cross validation (cross validation criteria, CVCs; see Allen, 1971; Stone, 1974; Yanagihara et al., 2013) are similarly used for model selection, and can be shown to be asymptotically equal to the AIC or TIC (Stone, 1977).

In this paper, models for categorical or multinomial data are dealt with. For these models, among the criteria shown above, the AIC, TIC, BIC and CVC can be used, where the likelihood based on multinomial or categorical distributions are used for the criteria except the CVC. The ϕ-divergence statistic (see e.g., Cressie and Pardo, 2002b; Pardo, 2006) is a generalization of the log-likelihood ratio statistic for evaluating the goodness-of-fit of a model. While the original definition of the AIC is based on the log-likelihood rather than the log-likelihood ratio, the latter can also be used for model selection in essentially the same way, since an added term in the log-likelihood ratio common to candidate models is irrelevant to model selection.

☆ This work was partially supported by a Grant-in-Aid for Scientific Research from the Japanese Ministry of Education, Culture, Sports, Science and Technology (JSPS KAKENHI, Grant No. 17K00042).

E-mail address: emt-hogasa@emt.otaru-uc.ac.jp.

https://doi.org/10.1016/j.csda.2018.03.001
0167-9473/

The ϕ-divergences are also used for estimation of parameters as well as criteria for the badness of a model, whose estimators are the minimizing ϕ-divergence estimators (MϕEs; Morales et al., 1995, p. 350; Pardo, 2006, Chapter 5). The ϕ-divergences and MϕEs are used in log-linear models (Cressie and Pardo, 2000, 2002a; Cressie et al., 2003), logistic regression for grouped data (Pardo et al., 2005) and latent class models (Felipe et al., 2015).

Since the main term of the AIC can be replaced by the -2 times the log-likelihood ratio, it is natural to consider the information criterion using the ϕ-divergence, whose special case is the AIC using the ratio. The so-called penalty term in the AIC i.e., 2 times the number of independent parameters in a model is given by the negative asymptotic bias of -2 times the log-likelihood ratio using the maximum likelihood estimator (MLE) as an estimator of the corresponding risk. Note that the risk is defined by the two-fold expectation of -2 times the log-likelihood ratio i.e., one for the expectation of data independent of current data, and the other for the expectation of the current data yielding the MLE. Note that the CVC is a numerical evaluation of the risk.

It will be shown that the asymptotic bias of order $O(1)$ for the ϕ-divergence as an estimator of the risk is equal to -2 times the number of independent parameters, which is the negative penalty term in the AIC. Note that the asymptotic bias is common to different ϕ-divergences using different MϕEs. The corresponding higher-order asymptotic bias of a ϕ-divergence based on MϕEs using different ϕ-divergences will be shown, where the results depend on types of ϕ-divergences and MϕEs.

The ϕ information criterion denoted by ϕIC or PIC will be defined similarly to the AIC. It will be shown that the AIC does not necessarily give best results in model selection among typical ϕICs. The higher-order bias term can be used for correction of the remaining bias yielding the modified ϕIC (MϕIC or MPIC). Since the higher order term is complicated, a simple approximation (M$^*\phi$IC or M*PIC) will be developed. It will be shown that the M$^*\phi$IC performs better than the ϕIC in simulations for model selection.

2. The bias of the ϕ-divergence

The ϕ-divergence statistic for K-category multinomial data is defined by

$$C_\phi = \frac{2n}{\phi''(1)} D_\phi = \frac{2n}{\phi''(1)} D_\phi(\mathbf{p}, \boldsymbol{\pi}) \quad \text{with} \quad D_\phi = \sum_{k=1}^{K} \pi_k \phi(p_k/\pi_k), \tag{2.1}$$

where $\mathbf{p} = (p_1, \ldots, p_K)'$ is a $K \times 1$ vector of sample proportions for K categories based on n observations; $\boldsymbol{\pi} = \boldsymbol{\pi}(\boldsymbol{\theta}) = (\pi_1, \ldots, \pi_K)'$ with $\pi_k = \pi_k(\boldsymbol{\theta})(k = 1, \ldots, K)$ is a $K \times 1$ vector of model-based probabilities, which are functions of a $q \times 1$ vector $\boldsymbol{\theta} = (\theta_1, \ldots, \theta_q)'$ of parameters ($q \leq K - 1; \boldsymbol{\theta} \in \Theta,\ \Theta \subset R^q$); the convex function $\phi(x)$ is assumed to have the following properties:

$$\begin{aligned} &x > 0, \phi(1) = 0, \phi'(1)(\text{the first derivative at } x = 1) = 0, \\ &\phi''(1)(\text{the second derivative at } x = 1) > 0, \\ &0\phi(0/0) = 0, 0\phi(v/0) = \lim_{u \to +\infty} \{\phi(u)/u\} \end{aligned} \tag{2.2}$$

(see e.g., Cressie and Pardo, 2000, 2002b; Pardo, 2006, Section 1.2), where D_ϕ was introduced by Csiszár (1963) and Ali and Silvey (1966).

When $\hat{\boldsymbol{\pi}} = \boldsymbol{\pi}(\hat{\boldsymbol{\theta}})$, where $\hat{\boldsymbol{\theta}}$ is the $q \times 1$ vector of MϕEs with $\phi(\cdot)$ being possibly different from $\phi(\cdot)$ in $D_\phi(\cdot)$ of (2.1), we have the ϕ-divergence statistic

$$\hat{C}_\phi = \frac{2n}{\phi''(1)} \hat{D}_\phi = \frac{2n}{\phi''(1)} D_\phi(\mathbf{p}, \hat{\boldsymbol{\pi}}). \tag{2.3}$$

Probably, the most important sub-family of the ϕ-divergence is that of the power divergences (Cressie and Read, 1984; Read and Cressie, 1988), where

$$\phi(x) = \frac{x^{\lambda+1} - x}{\lambda(\lambda + 1)} - \frac{x - 1}{\lambda + 1} \quad (-\infty < \lambda < +\infty,\ \lambda \neq 0, -1). \tag{2.4}$$

The cases of $\lambda = 0, -1$ are defined as the limiting values of $\phi(x)$ when $\lambda \to 0$ and $\lambda \to -1$, respectively. Eqs. (2.1) and (2.4) with $\phi''(1) = 1$ give an alternative expression of the power divergence

$$\begin{aligned} C_\phi &= 2n \sum_{k=1}^{K} \pi_k \left\{ \frac{(p_k/\pi_k)^{\lambda+1} - (p_k/\pi_k)}{\lambda(\lambda + 1)} - \frac{(p_k/\pi_k) - 1}{\lambda + 1} \right\} \\ &= \frac{2n}{\lambda(\lambda + 1)} \left(\sum_{k=1}^{K} \frac{p_k^{\lambda+1}}{\pi_k^{\lambda}} - 1 \right). \end{aligned} \tag{2.5}$$

Typical cases of the power divergences are as follows.