

Bayesian inference with a pairwise likelihood: an approach based on empirical likelihood

Francesco Pauli and Gianfranco Adimari

Abstract In several applications, the model of interest is such that its likelihood function is difficult, or even impractical, to compute. For these situations, it is common to substitute the likelihood with a surrogate, which resembles the full likelihood but is easier to calculate. Among these surrogates are composite likelihoods and in particular pairwise likelihoods. Their properties in classical inference have been widely discussed in the literature; their use within a Bayesian approach has been seldom considered. The substitution of the likelihood with a surrogate is not straightforward in Bayesian inference: the posterior distribution which is obtained must be validated on a case by case basis, as general results are not available. We propose a Bayesian procedure in which the surrogate is the empirical likelihood derived from the pairwise score equation. This pseudo-likelihood has standard asymptotic properties, so the validation of the posterior distribution is based on its asymptotic behavior.

Key words: Bayesian statistics, empirical likelihood, pairwise likelihood, parametric estimation

1 Introduction

Bayesian inference is based on the posterior distribution: the conditional distribution of the parameters given the data; this is obtained by combining, through Bayes theorem, the prior information on the parameter and the likelihood.

Francesco Pauli

Dip. Scienze Statistiche, Università di Padova, via c. Battisti 241/243, 35121, Padova, e-mail: fpauli@stat.unipd.it

Gianfranco Adimari

Dip. Scienze Statistiche, Università di Padova, via c. Battisti 241/243, 35121, Padova, e-mail: adimari@stat.unipd.it

A difficulty arises when the model structure is such that writing the likelihood function is a difficult, or even impractical, task. One of the possible workarounds is to substitute the likelihood with a surrogate which is easier to treat but has good properties, that is, it resembles to some extent the behavior of the likelihood.

Among the available surrogates are composite likelihoods [9], which are obtained by composition of marginal distributions for subsets of the data (Section 2). This is particularly convenient when it is relatively easy to compute marginal likelihoods as opposed to the full likelihood. In particular, we consider a special case of composite likelihood: the pairwise likelihood [2].

Classical inference based on the composite likelihood has been widely discussed [14], few examples can be found of its use for Bayesian inference [12].

Substituting the likelihood with a surrogate in the Bayes theorem is a non orthodox form of Bayesian inference, on which a wide literature already exists (Section 3). When a surrogate likelihood is employed in the Bayesian paradigm, Bayes theorem can not be invoked as a justification of the procedure – since the surrogate is not the distribution of the data conditional on the parameter – so the (non orthodox) posterior distribution which is obtained has to be validated. The validation of the posterior can rely on the good properties of the surrogate likelihood.

In this work, we propose to use, as a surrogate likelihood, the empirical likelihood derived from the estimating equations defining the pairwise maximum likelihood estimator. It is argued, based on general results for empirical likelihood, that such a surrogate has good asymptotic properties. Moreover, the efficacy of such a posterior distribution is explored by comparing it with the ordinary posterior distribution on simulated datasets.

2 Composite likelihood

Let $\mathbf{y}$ be a q dimensional r.v. and let $f(\mathbf{y}; \theta)$, $\theta \in \Theta \subset \mathbb{R}^d$ be a parametric model for $\mathbf{y}$. Given a set of measurable events $\{A_i; i = 1, \dots, m\}$ and a set of positive weights $\{w_i; i = 1, \dots, m\}$ the composite likelihood [9] is the product of the likelihoods corresponding to each event, $L_c(\theta; \mathbf{y}) = \prod_{j=1}^m f(\mathbf{y} \in A_j; \theta)^{w_j}$. The maximum composite likelihood estimator (MCLE) is defined accordingly. A special case of the composite likelihood, based only on bivariate marginal distributions of the pairs (y_r, y_s) , $f(y_r, y_s; \theta)$, is the pairwise likelihood [2],

$$L_p(\theta; \mathbf{y}) = \prod_{s=1}^{q-1} \prod_{r=s+1}^q f(y_r, y_s; \theta), \quad (1)$$

where we also assume unit weights.

Given a random sample $\mathcal{Y} = (\mathbf{y}^{(1)}, \dots, \mathbf{y}^{(n)})$ from $f(\mathbf{y}; \theta)$, the pairwise likelihood for the entire sample is

$$L_p(\theta; \mathcal{Y}) = \prod_{i=1}^n \prod_{s=1}^{q-1} \prod_{r=s+1}^q f(y_r^{(i)}, y_s^{(i)}; \theta). \quad (2)$$

The associated pairwise loglikelihood is $l_p(\theta; \mathcal{Y}) = \log L_p(\theta; \mathcal{Y})$ and its maximum $\hat{\theta}_{PL}$, which is the solution of the pairwise score equation

$$l_{p*}(\theta; \mathcal{Y}) = \frac{\partial}{\partial \theta} l_p(\theta; \mathcal{Y}) = \sum_{i=1}^n \sum_{s=1}^{q-1} \sum_{r=s+1}^q \frac{\partial}{\partial \theta} \log f(y_r^{(i)}, y_s^{(i)}; \theta) = 0, \quad (3)$$

is, if unique, the maximum pairwise likelihood estimator (MPLE).

Since $l_{p*}(\theta; \mathcal{Y})$ is a linear combination of valid likelihood score functions, the estimating equation (3) is unbiased. Moreover, the MPLE is asymptotically normal with mean θ and variance matrix $I(\theta)^{-1}J(\theta)(I(\theta)^T)^{-1}$, the inverse of Godambe information, where $I(\theta) = E(-\nabla l_{p*}(\theta; \mathbf{y}))$ and $J(\theta) = \text{var}(l_{p*}(\theta; \mathbf{y}))$. The second Bartlett identity does not hold ($I(\theta) \neq J(\theta)$), implying that there is a loss of efficiency with respect to the full likelihood.

Wald and score tests based on the pairwise likelihood are straightforward to derive and have standard asymptotic distributions. On the contrary, it can be shown (for example as a special case of the likelihood ratio test for misspecified models), that the pairwise likelihood ratio statistic has a non standard asymptotic distribution,

$$W_p(\theta) = 2(l_p(\hat{\theta}_{PL}; \mathcal{Y}) - l_p(\theta; \mathcal{Y})) \xrightarrow{d} \sum_{i=1}^p \lambda_i U_i^2, \quad (4)$$

with U_i^2 independent χ_1^2 random variables and λ_i eigenvalues of $I(\theta)^{-1}J(\theta)$, when θ is the true parameter value.

As a workaround, it has been proposed [4] to use the adjusted version $W_a(\theta) = W_p(\theta)/\bar{\lambda}$, where $\bar{\lambda}$ is the mean of the d eigenvalues λ_i . In [6, 7], the authors avoid the calculation of the eigenvalues of $I(\theta)^{-1}J(\theta)$ by substituting their mean $\bar{\lambda}$ with its consistent estimator (see also [13])

$$\tilde{\lambda} = \text{tr}(I(\hat{\theta}_{PL})^{-1}J(\hat{\theta}_{PL}))/d. \quad (5)$$

If $d = 1$, the adjusted version $W_a(\theta) = W_p(\theta)/\tilde{\lambda}$ (where $\tilde{\lambda} = J(\hat{\theta}_{PL})/I(\hat{\theta}_{PL})$) converges in distribution to a χ_1^2 , at the true parameter value. If $d > 1$ the asymptotic null distribution is not a χ_d^2 , nonetheless the χ_d^2 is used as a rough approximation.

3 Bayesian inference and the pairwise likelihood

Let $\pi(\theta)$ represent the prior information on θ and $L(\theta) = L(\theta; \mathcal{Y})$ the likelihood of the model, then the posterior distribution is given by

$$\pi(\theta|\mathcal{Y}) \propto \pi(\theta)L(\theta). \quad (6)$$

When the likelihood $L(\theta)$ is substituted with a pseudo-likelihood $L^*(\theta)$, one can derive a (possibly) non orthodox posterior distribution by

$$\pi^*(\theta|\mathcal{Y}) \propto \pi(\theta)L^*(\theta). \quad (7)$$

There are various useful examples of application of this approach in Bayesian inference (see references in [15]). The problem with them is that the posterior distribution $\pi^*(\theta|\mathcal{Y})$ must be validated, since equation (7) does not represent an application of Bayes theorem because $L^*(\theta)$ is not the conditional distribution of the data.

How to accomplish the validation depends on the nature of $L^*(\theta)$. In some situations, $L^*(\theta)$ has a Bayesian interpretation which makes $\pi^*(\theta|\mathcal{Y})$ a probability (genuine) posterior distribution [15]. In other circumstances, one can validate (7) by simulation methods such as the one proposed in [10] and used in [8]. Another approach is to investigate the asymptotic behaviour of $\pi^*(\theta|\mathcal{Y})$.

The posterior distribution obtained by substituting the likelihood with the pairwise likelihood (2),

$$\pi_p(\theta|\mathcal{Y}) \propto \pi(\theta)L_p(\theta) \quad (8)$$

is a probability posterior distribution since $L_p(\theta)$ is the density function of the fictitious sample $\{(y_r^{(i)}, y_s^{(i)}) \mid i = 1, \dots, n; s = 1, \dots, q-1; r = s+1, \dots, q\}$ so, if the prior $\pi(\theta)$ is proper, $\pi_p(\theta|\mathcal{Y})$ is the result of an orthodox application of Bayes' Theorem. This reasoning mimics that in [3], where a posterior is defined using the square of the likelihood and this is justified by referring to a doubled data set. However, asymptotically, the variance matrix of $\pi_p(\theta|\mathcal{Y})$ is the inverse of the hessian matrix of minus the pairwise likelihood, which, as explained in Section 2, does not reflect the variability of the parameter. For this reason, the inference drawn from $\pi_p(\theta|\mathcal{Y})$ is not reliable.

To correct for this, in [12], the pairwise likelihood $L_p(\theta)$ in (8) is substituted with the adjusted pairwise likelihood

$$L_a(\theta) = L_p(\theta)^{1/\tilde{\lambda}}, \quad (9)$$

where the adjustment factor $1/\tilde{\lambda}$ is the one defined in (5). The motivation is found in [6, 7]: the corresponding pairwise likelihood ratio $W_a(\theta) = 2(l_a(\hat{\theta}_{PL}) - l_a(\theta))$, where $l_a(\theta) = \log L_a(\theta)$, approximates the usual asymptotic χ_d^2 distribution if $d > 1$, and has asymptotically a χ_1^2 distribution if $d = 1$.

Since the pseudo-likelihood $L_a(\theta)$ has approximately all the desired standard first-order properties, it resembles the behavior of true likelihood and thus can be used as a basis for Bayesian inference. More precisely, the posterior distribution

$$\pi_a(\theta|\mathcal{Y}) \propto \pi(\theta)L_a(\theta) \quad (10)$$

roughly behaves as an asymptotic normal distribution with variance matrix the inverse of the information matrix of $L_a(\theta)$ which is equal to the Godambe information matrix of $L_p(\theta)$, i.e. the inverse of the asymptotic variance matrix of $\hat{\theta}_{PL}$. In particular, if interest focuses on a scalar parameter one can employ the same reasoning as

in [8, 5] to show that the posterior distribution $\pi_a(\theta|\mathcal{Y})$ is asymptotically normal. When the parameter of interest has dimension greater than one, since the adjustment (5) does not lead to a χ_d^2 asymptotic null distribution (see Section 2), the normal distribution is only a rough approximation for the asymptotic distribution.

4 Empirical likelihood

The empirical likelihood [11] is a special case of a surrogate likelihood and it can be derived from an estimating equation. Let then

$$\bar{\eta}(\mathcal{Y}; \theta) = \frac{1}{m} \sum_{j=1}^m \eta_j(\mathcal{Y}_j^l; \theta) = 0 \quad (11)$$

be an estimating equation for $\theta \in \mathbb{R}^d$, where $\mathcal{Y}_j^l$ is, for each j , a subset of l elements in $\mathcal{Y} = (\mathbf{y}^{(1)}, \dots, \mathbf{y}^{(n)})$ and η_j are real functions. The estimating equation (11) is associated to the empirical likelihood

$$L_e(\theta) = \prod_{j=1}^m \frac{1}{m(1 + \lambda^T \eta_j(\mathcal{Y}_j^l; \theta))} \quad (12)$$

where λ satisfies

$$\sum_{j=1}^m \frac{\eta_j(\mathcal{Y}_j^l; \theta)}{1 + \lambda^T \eta_j(\mathcal{Y}_j^l; \theta)} = 0. \quad (13)$$

The use of empirical likelihood in the Bayesian paradigm has been discussed in [8]. Results therein suggest that the procedure is reasonable, but no general results are obtained concerning the validity of the procedure, which must then be established on a case by case basis.

We note then that the pairwise score equation (3) has a form as in (11) with $m = n$, $l = 1$ and

$$\eta_j(\mathcal{Y}_j^l; \theta) = \sum_{s=1}^{q-1} \sum_{r=s+1}^{q-1} \frac{\partial}{\partial \theta} \log f(\mathbf{y}_r^{(j)}, \mathbf{y}_s^{(j)}; \theta), \quad (14)$$

If (14) is used, due basically to the fact that the η_j functions are independent, corollary 1 in [1] holds, so the empirical likelihood ratio obtained combining (12) and the score equation (3) follows standard asymptotics. In this case, following the same reasoning which has been put forward in Section 2 and has motivated the correction of the likelihood suggested in [12], it can be shown, along the lines of [8, 5], that the posterior distribution

$$\pi_e(\theta|\mathcal{Y}) \propto \pi(\theta) L_e(\theta) \quad (15)$$

is asymptotically normal where the limit of the variance matrix is the inverse of Godambe information of $L_p(\theta)$. This is analogous to $\pi_a(\theta|\mathcal{Y})$ in the one dimen-

sional case, but, unlike the adjustment proposed in $\pi_a(\theta|\mathcal{Y})$ it also applies without changes to multidimensional parameters.

Fig. 1 Comparison of posterior distributions for the parameter θ of model $\mathcal{M}_1$: black circles represent the quantile quantile plot of the samples obtained via MCMC from $\pi(\theta)$ and $\pi_e(\theta)$; gray circles represent the quantile quantile plot of the samples from $\pi(\theta)$ and $\pi_p(\theta)$

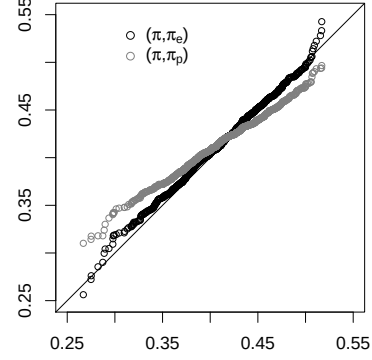
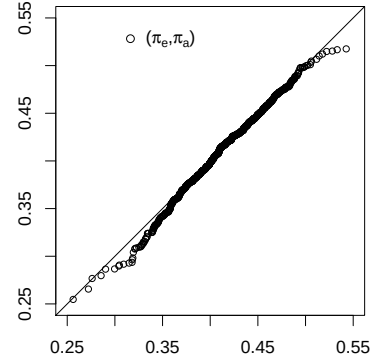


Fig. 2 Comparison of posterior distributions for the parameter θ of model $\mathcal{M}_1$: circles represent the quantile quantile plot of the samples obtained via MCMC from $\pi_e(\theta|\mathcal{Y})$ and $\pi_a(\theta|\mathcal{Y})$



5 Simulation experiment

We consider inference for the correlation coefficients of a q -variate normal distribution with standard margins. In both examples we simulate a sample of length 100 from a q -variate normal with null mean and variance matrix having the diagonal equal to 1 and off diagonal elements equal to 0.5.

The two models differ because in the first one, $\mathcal{M}_1$, we assume, similar to [2], that the correlation coefficient is the same for each pair, so the parameter has dimension

Fig. 3 Comparison of posterior distributions for the parameter θ_1 of model $\mathcal{M}_2$: black circles represent the quantile quantile plot of the samples obtained via MCMC from $\pi(\theta_1|\mathcal{Y})$ and $\pi_e(\theta_1|\mathcal{Y})$; gray circles represent the quantile quantile plot of the samples from $\pi(\theta_1|\mathcal{Y})$ and $\pi_p(\theta_1|\mathcal{Y})$

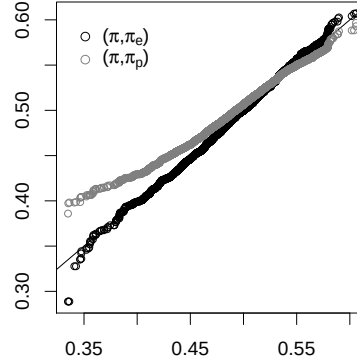
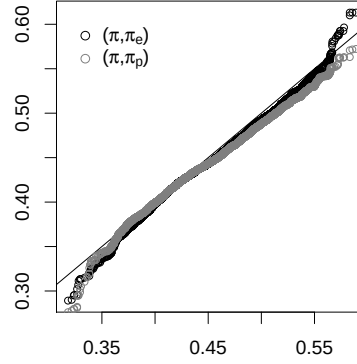


Fig. 4 Comparison of posterior distributions for the parameter θ_2 of model $\mathcal{M}_2$: black circles represent the quantile quantile plot of the samples obtained via MCMC from $\pi(\theta_2|\mathcal{Y})$ and $\pi_e(\theta_2|\mathcal{Y})$; gray circles represent the quantile quantile plot of the samples from $\pi(\theta_2|\mathcal{Y})$ and $\pi_p(\theta_2|\mathcal{Y})$



one; in the second model, $\mathcal{M}_2$, we assume that $\text{cov}(Y_{hi}, Y_{ki}) = \theta_{|h-k|}$ for $h \neq k$, so $q - 1$ parameters are involved.

We estimate the first model on a sample with $q = 5$ and assume for the (unidimensional) parameter θ a uniform prior on $[-1, 1]$. We obtain, by Markov Chain Monte Carlo, simulated samples from four posterior distributions: the posterior distribution (6) based on the full likelihood (π); the posterior (8) based on the pairwise likelihood (π_p); the posterior (10) based on the adjustment of the likelihood proposed in [12], (π_a) and, finally, the posterior (15) based on the empirical likelihood (π_e).

The comparison of these posterior distributions is depicted in Figures 1 and 2. In Figure 1 we compare π_e and π_p with π , it is seen that the π_e resembles the variability of the ordinary posterior, while π_p does not give valid inference. In Figure 2 we compare π_e with π_a , we see that the adjusted and the empirical likelihoods give similar results.

The second model is estimated on a sample with $q = 4$ dimensions, so there are three parameters. Assuming independent uniform priors we obtain the posterior

distributions π , π_p and π_e , defined as above. These are compared in figures 3 for θ_1 and 4 for θ_2 , showing the extent of the correction due to the use of L_e . In this case we do not consider π_a , as the adjustment is based on asymptotic considerations which are only roughly valid.

6 Final remarks

In this work we explore a strategy for using the pairwise likelihood tool in a Bayesian setting. The strategy we put forward relies on using the empirical likelihood function which can be derived from the pairwise score equations. We think that this approach has more well founded theoretical basis as compared to the approach proposed in [12], since the latter relies on rough approximations rather than asymptotic results. On the other hand, it might be more computationally demanding than the latter.

References

1. Adimari, G., Guolo, A.: A note on the asymptotic behaviour of empirical likelihood statistics. *Statist. Methods Appl.*, published on-line, doi:10.1007/s10260-010-0137-9 (2010)
2. Cox, D.R., Reid, N.: A note on pseudolikelihood constructed from marginal densities. *Biometrika* **97**, 729–737 (2004)
3. Efron, B.: Bayes and likelihood calculations from confidence intervals. *Biometrika* **80**, 3–26 (2003)
4. Geys, H., Molenberghs, G., Ryan, L.M.: Pseudolikelihood modelling of multivariate outcomes in developmental toxicology. *J. Amer. Statist. Assoc.* **94**, 734–745 (1999)
5. Greco, L., Racugno, W., Ventura, L.: Robust likelihood functions in Bayesian inference. *J. Statist. Plann. Inf.* **138**, 1258–1270 (2008)
6. Hanfelt, J.J., Liang, K.Y.: Approximate likelihood ratios for general estimating functions. *Biometrika* **82**, 461–477 (1995)
7. Hanfelt, J.J., Liang, K.Y.: Inference for odds ratio regression models with sparse dependent data. *Biometrics* **54**, 136–147 (1998)
8. Lazar, N. A.: Bayesian Empirical Likelihood. *Biometrika* **90**, 319–326 (2003)
9. Lindsay, B.G.: Composite likelihood methods. *Contemp. Math.* **80**, 221–240 (1988)
10. Monahan, J.F., Boos, D.D.: Proper likelihoods for Bayesian analysis. *Biometrika* **79**, 271–278 (1992)
11. Owen, A.B.: *Empirical Likelihood*. Chapman & Hall/CRC, Boston (2001)
12. Pauli, F., Racugno, W., Ventura, L.: Bayesian Composite Marginal Likelihoods. Submitted.
13. Rao, J.N.K., Scott, A.J.: The analysis of categorical data from complex sample surveys: Chi-squared tests for goodness of fit and independence in two-way tables. *J. Amer. Statist. Assoc.* **76**, 221–230 (1981)
14. Varin, C.: On composite marginal likelihoods. *Adv. Statist. Analysis* **92**, 1–28 (2008)
15. Ventura, L., Cabras, S., Racugno, W.: Prior distributions from pseudo-likelihoods in the presence of nuisance parameters. *J. Amer. Statist. Assoc.* **104**, 768–774 (2009)