Statistica Sinica **21** (2011), 149-164

BAYESIAN COMPOSITE MARGINAL LIKELIHOODS

Francesco Pauli[1], Walter Racugno[2] and Laura Ventura[1]

[1] *University of Padua and* [2] *University of Cagliari*

Abstract: This paper proposes and discusses the use of composite marginal likelihoods for Bayesian inference. This approach allows one to deal with complex statistical models in the Bayesian framework, when the full likelihood - and thus the full posterior distribution - is impractical to compute or even analytically unknown. The procedure is based on a suitable calibration of the composite likelihood that yields the right asymptotic properties for the posterior probability distribution. In this respect, an attractive technique is offered for important settings that at present are not easily tractable from a Bayesian perspective, such as, for instance, multivariate extreme value theory. Simulation studies and an application to multivariate extremes are analysed in detail.

Key words and phrases: Asymptotic theory, Bayesian inference, estimating equation, extreme value theory, pairwise likelihood, pseudo-likelihood.

1. Introduction

In such areas as geostatistics, spatial extremes, and genetics, there is scope for using complex models for which writing the full likelihood function poses not only theoretical but computational challenges, for instance, clustered and longitudinal studies or space-time models. In these situations, for frequentist or Bayesian inference, surrogates of the ordinary likelihood are desirable. An important contribution here is given by approximate likelihoods based on the composition of marginal distributions (Varin (2008)), particularly for bivariate marginal distributions (Cox and Reid (2004)), termed composite marginal likelihoods. Their use has been widely advocated by several authors in different complex applications of frequentist inference (see Varin (2008), and Varin, Reid and Firth (2009) for recent reviews). Although there exist models for which composite likelihoods can be helpful to Bayesian inference, inference based on composite marginal likelihoods has not been explored in a Bayesian setting; an exception here is Smith and Stephenson (2009), where a specific application to max-stable processes for modeling spatial extremes is discussed.

Bayesian techniques are essential to incorporate information other than the data into the model in the form of prior distributions. This is a particularly attractive feature since the inclusion of substantive prior information can help in

mitigate the problem of scarce data. Other reasons for the Bayesian approach include the simplicity of predictive inference, and the fact that one can easily handle large numbers of parameters (for example, employing hierarchical models). The aim of this paper is to discuss the use of the composite marginal likelihood as a basis for Bayesian inference. To this end, we propose posterior probability distributions as a useful tool when we are not able to write down the full joint distribution.

Let Y be a $(q \times 1)$ random variable with joint density $f(y;\theta)$, $\theta \in \Theta \subseteq \mathbb{R}^d$, $d \geq 1$, and let $y = (y^{(1)}, \ldots, y^{(n)})$ be a random sample of size n from Y. When $f(y;\theta)$ is complex, composite marginal likelihoods may be useful for approximating the likelihood. Composite likelihood is defined through a set of measurable events $\{A_i; i = 1, \ldots, m\}$ in the sample space: given the likelihood contributions corresponding to each A_i, the composite likelihood is defined as the weighted product (see, e.g., Lindsay (1988))

$$\mathrm{CL}(\theta) = \mathrm{CL}(\theta; y) = \prod_{i=1}^{m} f(y \in A_i; \theta)^{w_i} \ , \tag{1.1}$$

where w_i are positive weights, $i = 1, \ldots, m$. The class of composite likelihoods contains the usual full likelihood as well as many other alternatives: the Besag pseudo-likelihood (Besag (1974, 1977)), the m-order likelihood (Azzalini (1983)), the partial likelihood (Cox (1975)), and the composite marginal likelihood (Cox and Reid (2004)). Complex models pose a serious problem in the Bayesian setting as well as in the frequestist framework. Indeed, when the full likelihood is computationally cumbersome or when a fully specified model is out of reach, a posterior distribution for the parameter of interest may be unavailable. In these situations, alternative posterior distributions based on composite marginal likelihoods may be helpful. In particular, a posterior distribution can be formally obtained with the composite likelihood (1.1) substituting for the full likelihood, i.e.,

$$\pi_{CL}(\theta|y) \propto \pi(\theta)\ \mathrm{CL}(\theta) \ , \tag{1.2}$$

where $\pi(\theta)$ is a suitable prior on θ. Although $\mathrm{CL}(\theta)$ is not a genuine likelihood function, there is an extensive literature on the use of alternative likelihoods in the Bayesian setting: the partial likelihood in the context of survival analysis (Raftery, Madigan and Volinsky (1996)); marginal, conditional, profile, and modified profile likelihoods for the elimination of nuisance parameters (see, among others, Sweeting (1987), Monahan and Boos (1992), Reid (1995), Severini (1999), Fraser et al. (2003), Chang and Mukerjee (2006), Ventura Cabras, and Racugno (2009), Racugno, Salvan, and Ventura (2010)); quasi- and empirical likelihoods

derived from estimating functions (Lazar (2003), Chernozhukov and Hong (2003), Lin (2006), Greco, Racugno, and Ventura (2008), Ventura, Cabras, and Racugno (2010)). In this paper, we state the properties of (1.2), focusing on a particular choice of the weights w_i, $i = 1, \ldots, m$, in (1.1). A calibrated composite likelihood is needed to reach the right asymptotic variance in the normal approximation, as well as a correct shape of the posterior distribution in its Laplace approximation. Indeed, the variability of the non-calibrated composite likelihood, $w_i = 1$ for $i = 1, \ldots, m$, in a posterior distribution is substantially lower than that one should use, as it leads to a falsely precise inference (see also Smith and Stephenson (2009)).

The outline of the paper is as follows. After a brief review of definitions and properties of composite marginal likelihoods, Section 2 discusses the procedure to obtain and use the proposed posterior probability distribution (1.2). To assess our proposal, numerical investigations and an application to multivariate extremes are reported in Sections 3 and 4, respectively. Some final remarks (Section 5) conclude the paper.

2. Marginal Composite Posterior Distributions

In this section we show that the composite likelihood can be central in the Bayesian approach when the full likelihood function $L(\theta)$, needed to obtain the full posterior distribution

$$\pi(\theta|y) \propto \pi(\theta)\ L(\theta)\ , \tag{2.1}$$

is difficult or even impossible to compute. Indeed, the fact that composite likelihoods share many asymptotic properties of the full likelihood (see Subsection 2.1) suggests that they can be used as the basis for Bayesian inference (see Subsection 2.2). The validation of the proposed posterior distribution relies on its asymptotic behaviour.

2.1. Background on composite marginal likelihood inference

Composite marginal likelihoods are based on the composition of low-dimensional margins. For instance, when the events A_i in (1.1) are defined in terms of pairs of observations, the pairwise likelihood can be obtained from the bivariate marginal densities $f(y_i, y_j; \theta)$, $i, j = 1, \ldots, q$, as (Le Cessie and Van Houwelingen (1994), Cox and Reid (2004), Varin (2008))

$$\mathrm{PL}(\theta) = \prod_{i=1}^{q} \prod_{j>i} f(y_i, y_j; \theta)^{w_{ij}}\ , \tag{2.2}$$

where w_{ij} are positive weights, for $i, j = 1, \ldots, q$. In some applications, it may be preferable to consider higher dimensional margins as triplets of observations

(Varin and Vidoni (2005), Engler et al. (2006)), or to combine different composite marginal likelihoods in some optimal way (Cox and Reid (2004)).

Under broad assumptions (see, e.g., Molenberghs and Verbeke (2005)), the maximum composite likelihood estimator $\hat{\theta}_c$ is the solution of the composite score function $s(\theta) = \nabla \log \mathrm{CL}(\theta) = \sum_{i=1}^{m} w_i s_i(\theta)$, with $s_i(\theta) = \nabla \log f(y \in A_i; \theta)$, $i = 1, \ldots, m$. The composite score $s(\theta)$ is unbiased, since it is a linear combination of valid score functions associated with each log-likelihood term. Moreover, $\hat{\theta}_c$ is consistent and asymptotically normal, with mean θ and variance $V(\theta) = H(\theta)^{-1} J(\theta) (H(\theta)^{\intercal})^{-1}$, where $H(\theta) = E(-\nabla s(\theta))$ and $J(\theta) = \mathrm{var}(s(\theta))$. The matrix $G(\theta) = V(\theta)^{-1}$ is the well-known Godambe information (Godambe (1960)) or sandwich information matrix. The form is due to the failure of the second Bartlett identity since in general $H(\theta) \neq J(\theta)$, indicating loss of efficiency with respect to the maximum likelihood estimator.

Wald tests and confidence intervals for θ based on composite likelihoods can be obtained in a standard way using consistent estimates of the matrices $H(\theta)$ and $J(\theta)$; we refer the reader to Varin (2008) for a detailed discussion. However, as is well known, Wald-type statistics lack invariance under reparameterization and force confidence regions to have an elliptical shape. In this respect, a likelihood ratio type statistic may be preferable. However, the composite likelihood ratio statistic $W_c(\theta) = 2(c\ell(\hat{\theta}_c) - c\ell(\theta))$, with $c\ell(\theta) = \log \mathrm{CL}(\theta)$, has a non-standard asymptotic null distribution. Indeed, the asymptotic distribution of $W_c(\theta)$ is a linear combination of independent chi-squared variates, $W_c(\theta) \xrightarrow{d} \sum_{i=1}^{d} \lambda_i(\theta) Z_i^2$, where the Z_i are independent standard normal variates, and the $\lambda_i(\theta)$ are eigenvalues of $H(\theta)^{-1} J(\theta)$, $i = 1, \ldots, d$. This calls for adjustments to $W_c(\theta)$ and several proposals have been considered (see, e.g., Varin, Reid and Firth (2009), Section 2.3). There are adjustments based on moment matching conditions (Geys, Molenberghs and Ryan (1999)), there is a Satterthwaite-type adjustment suggested in Varin (2008), and there are other proposals based on suitable vertical scalings of $W_c(\theta)$ (Chandler and Bate (2007)). Here we focus on the adjustment based on first-order moment matching, Geys, Molenberghs and Ryan (1999), who propose to use $W_c^{\dagger}(\theta) = W_c(\theta)/\tilde{\lambda}$, with

$$\tilde{\lambda} = \frac{1}{d} \sum_{i=1}^{d} \lambda_i(\hat{\theta}_c) = \frac{\mathrm{tr}(H(\hat{\theta}_c)^{-1} J(\hat{\theta}_c))}{d} \ ,$$

which has, in general, an approximate χ^2_d distribution (see Varin (2008), and Hanfelt and Liang (1995)). When $d = 1$, $W_c^{\dagger}(\theta) \xrightarrow{d} \chi^2_1$ and a similar result holds when interest is on a scalar component, say θ_j, of θ, and a profile version of the adjusted composite likelihood ratio test is considered with $\tilde{\lambda} = H(\hat{\theta}_c)^{jj} J(\hat{\theta}_c)_{jj}$,

where $H(\theta)^{jj}$ and $J(\theta)_{jj}$ denote the (θ_j, θ_j)-components of $H(\theta)^{-1}$ and $J(\theta)$, respectively.

Our application of marginal composite likelihoods in the Bayesian framework is based on a particular choice of the weights in (1.1) or in (2.2), that alleviates inefficiency of composite likelihood methods and recovers, approximately, the asymptotic properties of (1.2). In particular, we consider the calibrated composite likelihood

$$\mathrm{CL}_c(\theta) = \prod_{i=1}^{m} f(y \in A_i; \theta)^{1/\tilde{\lambda}} \,, \tag{2.3}$$

with associated loglikelihood $c\ell_c(\theta) = \log \mathrm{CL}_c(\theta)$. Note that (2.3) is simply the composite likelihood with a particular choice of weights, namely those that yield the composite likelihood ratio statistic $W_c^\dagger(\theta)$.

2.2. Asymptotics for marginal composite posterior distributions

The marginal composite likelihood $\mathrm{CL}_c(\theta)$ can be used for Bayesian inference on θ, incorporating prior information in the form of a prior distribution $\pi(\theta)$ to obtain the posterior distribution

$$\pi_{CL_c}(\theta|y) = \frac{\pi(\theta)\ \mathrm{CL}_c(\theta)}{\int \pi(\theta)\ \mathrm{CL}_c(\theta)\, d\theta} \,. \tag{2.4}$$

The validation of (2.4) relies on its asymptotic behaviour (see also the papers by Lazar (2003), and Greco, Racugno, and Ventura (2008)). In particular, paralleling the results for the full posterior distribution, we focus on the Laplace expansion and the asymptotic normality of (2.4); see, for instance, Reid (2003).

The approximation to (2.4) based on the Laplace expansion (see, e.g., Tierney and Kadane (1986)) gives

$$\begin{aligned} \pi_{CL_c}(\theta|y) &\doteq (2\pi)^{-d/2} |j_c(\hat{\theta}_c)|^{1/2} \exp\{c\ell_c(\theta) - c\ell_c(\hat{\theta}_c)\} \frac{\pi(\theta)}{\pi(\hat{\theta}_c)} \\ &\doteq (2\pi)^{-d/2} |j_c(\hat{\theta}_c)|^{1/2} \exp\left\{-\frac{1}{2} W_c^\dagger(\theta)\right\} \frac{\pi(\theta)}{\pi(\hat{\theta}_c)} \,, \end{aligned} \tag{2.5}$$

with relative error of order $O(n^{-1})$. In (2.5), $j_c(\theta) = -\partial^2 c\ell_c(\theta)/(\partial\theta\partial\theta^\top)$ is the composite observed information. The result that the posterior distribution (2.4) has approximately the asymptotic behavior of (2.1) follows by the χ^2_d approximation used for the null distribution of $W_c^\dagger(\theta)$, and by the assumption that the prior has order $O(1)$. When using $W_c^\dagger(\theta)$, a correct shape of the posterior distribution in the Laplace approximation is reached.

Under standard regularity conditions, as $n \to \infty$, straightforward calculations show that $\pi_{CL_c}(\theta|y)$ is approximately a random normal density with mean $\hat{\theta}_c$ and variance $j_c(\hat{\theta}_c)^{-1}$. Indeed, expanding the logarithm of $\pi(\theta)$ and $W_c^\dagger(\theta)$ at (2.5) around their maxima, θ_0 and $\hat{\theta}_c$ respectively, one gets

$$\begin{aligned}\pi_{CL_c}(\theta|y) &\dot{\propto} \exp\left\{-\frac{1}{2}W_c^\dagger(\theta) + \log\pi(\theta)\right\} \\ &\dot{\propto} \exp\left\{-\frac{1}{2}(\theta-\hat{\theta}_c)^\intercal j_c(\hat{\theta}_c)(\theta-\hat{\theta}_c) - \frac{1}{2}(\theta-\theta_0)^\intercal J_0(\theta-\theta_0)\right\} \\ &\dot{\propto} \exp\left\{-\frac{1}{2}(\theta-\hat{\theta}_{c_0})^\intercal J_{c_0}(\theta-\hat{\theta}_{c_0})\right\}\ , \end{aligned} \tag{2.6}$$

where $J_0 = -\partial\log\pi(\theta)/(\partial\theta\partial\theta^\intercal)|_{\theta=\theta_0}$, $J_{c_0} = J_0 + j_c(\hat{\theta}_c)$, and $\hat{\theta}_{c_0} = J_{c_0}^{-1}(J_0\theta_0 + j_c(\hat{\theta}_c)\hat{\theta}_c)$ (see also Lazar (2003), and Greco, Racugno, and Ventura (2008)). From (2.6) it follows that $\pi_{CL_c}(\theta|y)$ is asymptotically normal with mean $\hat{\theta}_{c_0}$ and variance $J_{c_0}^{-1}$. However, for large n, $\hat{\theta}_{c_0}$ and $J_{c_0}^{-1}$ are essentially indistinguishable from $\hat{\theta}_c$ and $j_c(\hat{\theta}_c)^{-1}$, i.e., the influence of the prior vanishes in the limit, as is expected by requiring the prior to supply an amount of information equivalent to that contained in one observation. Note also that $j_c(\theta) = J(\theta)/\tilde{\lambda} + o_p(1)$. Thus, for $d = 1$ or if interest focuses on a scalar component of θ, the asymptotic variance of $\pi_{CL_c}(\theta|y)$ reduces to $V(\hat{\theta}_c)$, i.e., to the asymptotic variance of $\hat{\theta}_c$. When $d > 1$, $J(\theta)/\tilde{\lambda}$ approximates $V(\theta)$ better than $J(\theta)$. In this respect, the calibration of the composite likelihood is needed to reach the right asymptotic variance in the normal approximation. Indeed, when using the non-calibrated composite likelihood in (2.4), the asymptotic variance of the posterior probability distribution is $J(\theta)^{-1}$.

3. A Simulation Study

In this section we discuss Bayesian inference based on the pairwise likelihood (2.2) for the correlation coefficient ρ of a multivariate normal distribution. This illustrative example is considered in Cox and Reid (2004) and, as in that case, attention is restricted to positive ρ.

Let Y be a q-variate normal random variable with standard margins, and let $\text{corr}(Y_r, Y_s) = \rho$, for $r, s = 1, \ldots, q$, $r \neq s$. Following Cox and Reid (2004), the non-calibrated pairwise loglikelihood $p\ell(\rho) = \log \text{PL}(\rho) = \sum_{i=1}^{q}\sum_{j>i}\log f(y_i, y_j; \rho)$ is

$$p\ell(\rho) = -\frac{nq(q-1)}{4}\log(1-\rho^2) - \frac{q-1+\rho}{2(1-\rho^2)}SS_W - \frac{(q-1)(1-\rho)}{2(1-\rho^2)}\frac{SS_B}{q}, \tag{3.1}$$

where $SS_W = \sum_{i=1}^{n}\sum_{r=1}^{q}(y_r^{(i)} - \bar{y}^{(i)})^2$, $SS_B = \sum_{i=1}^{n}\bar{y}^{(i)2}$, and $\bar{y}^{(i)} = \sum_{r=1}^{q} y_r^{(i)}/q$.

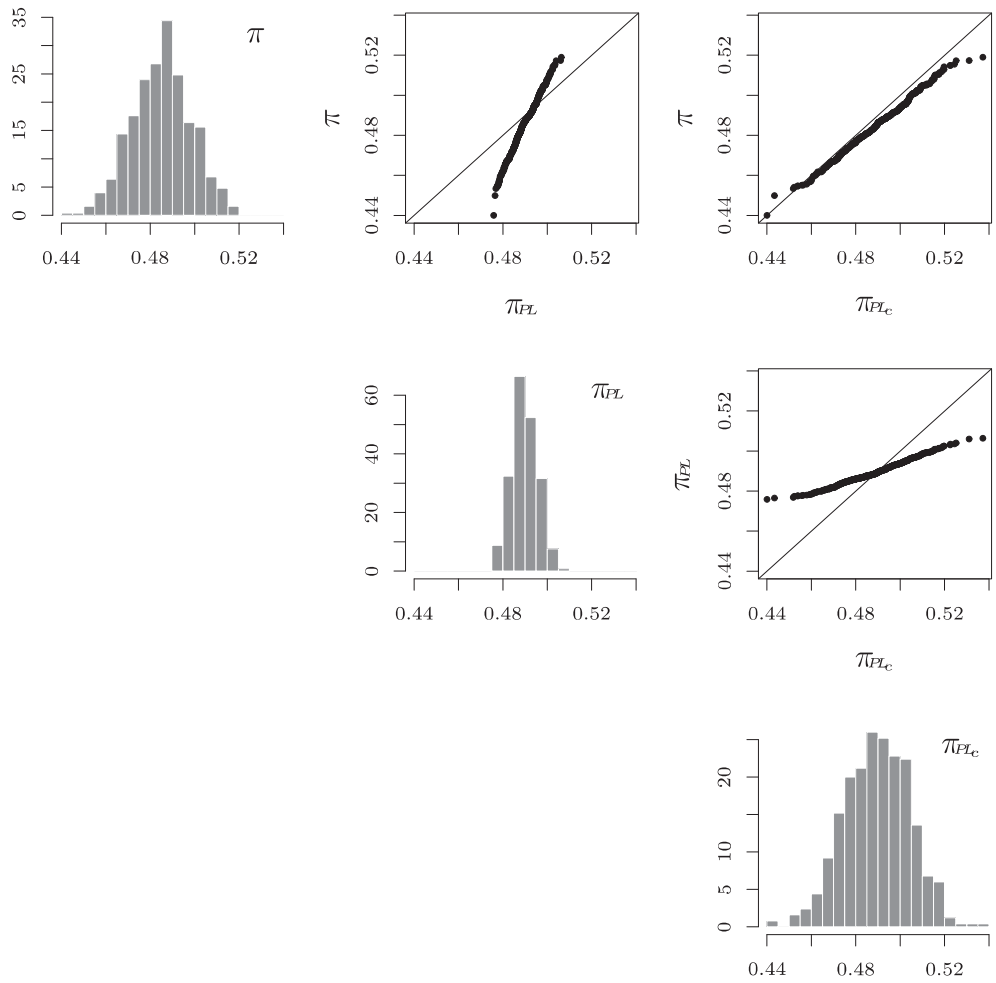

Figure 1. Comparison of the posterior distributions based on the full likelihood (π), the non-calibrated pairwise likelihood (π_{PL}), and the calibrated pairwise likelihood (π_{PL_c}), for a simulated sample with $n = 30$, $q = 10$, and $\rho = 0.5$.

The associated score function is

$$s(\rho) = \frac{nq(q-1)\rho}{2(1-\rho^2)} - \frac{1+\rho^2+2(q-1)\rho}{2(1-\rho^2)^2} SS_W + \frac{(q-1)(1-\rho)^2}{2(1-\rho^2)^2} \frac{SS_B}{q} \ ,$$

and the asymptotic variance of the maximum pairwise likelihood estimator $\hat{\rho}_p$ is $V(\hat{\rho}_p) = 2(1-\rho)^2 c(q,\rho)/(nq(q-1)(1+\rho^2)^2)$, where $c(q,\rho) = (1-\rho)^2(3\rho^2+1) + q\rho(-3\rho^3+8\rho^2-3\rho+2) + q^2\rho^2(1-\rho)^2$.

In this example it is of interest to compare three different posterior distributions: the posterior $\pi(\rho|y) \propto \pi(\rho)\, L(\rho)$ based on the full likelihood; the posterior

$\pi_{PL}(\rho|y) \propto \pi(\rho)\,\mathrm{PL}(\rho)$ based on the non-calibrated pairwise likelihood as proposed in Smith and Stephenson (2009); the posterior $\pi_{PL_c}(\rho|y) \propto \pi(\rho)\,\mathrm{PL}_c(\rho)$ based on the calibrated pairwise likelihood $\mathrm{PL}_c(\rho) = \prod_{i=1}^{q}\prod_{j>i} f(y_i, y_j; \rho)^{1/\tilde{\lambda}}$, with $\tilde{\lambda} = J(\hat{\rho}_p)/H(\hat{\rho}_p)$. Samples from these posterior distributions can be obtained by straightforward MCMC simulations.

The comparison between the three posterior distributions is illustrated in Figure 1, for a sample of size $n = 30$, with $\rho = 0.5$, $q = 10$, and assuming a uniform prior in $(0, 1)$ for ρ. It is clearly seen that the variability of $\pi_{PL}(\rho|y)$ is small with respect to the variability of $\pi(\rho|y)$, as noted by Smith and Stephenson (2009), while $\pi_{PL_c}(\rho|y)$ implies a greater variability than both. The difference in variability between $\pi_{PL_c}(\rho|y)$ and $\pi(\rho|y)$ is not surprising in view of the fact that a misspecified likelihood is used. This can be interpreted analogously to the loss of efficiency of the pairwise maximum likelihood estimator with respect to the maximum likelihood estimator. In this respect consider, as in Cox and Reid (2004), the ratio between the asymptotic variance of the maximum likelihood estimator $\hat{\rho}$ to that of $\hat{\rho}_p$, as a function of $\rho \in (0, 1)$, for a range of values of q. The inspection of such ratios, reported in Figure 2 (the continuous lines in each panel) for $q = 3, 5, 8, 10$, reveals that the loss of efficiency is, for fixed q, a u-shaped function of ρ with a minimum at $\rho = 0.5$ (corresponding to the maximum loss of efficiency) while, for a fixed value of ρ, a greater loss of efficiency occurs for a higher q. A similar feature is seen when comparing the variance of $\pi(\rho|y)$ to that of $\pi_{PL_c}(\rho|y)$. This was done by performing a series of experiments: in each experiment we simulated 800 samples of size 30 from a q-variate equicorrelated normal distribution with standard margins. For each sample we obtained the posterior distributions for ρ using the full likelihood and the calibrated pairwise likelihood, and calculated the ratio between their variances. The experiment was repeated for $q = 3, 5, 8, 10$, and for $\rho = 0.1, 0.2, \ldots, 0.9$. The results are depicted in Figure 2 (where each panel contains results for a value of q as ρ varies), where each boxplot represents the distribution of the ratio of the variance of $\pi(\rho|y)$ to the variance of $\pi_{PL_c}(\rho|y)$ across the 800 simulated samples for fixed ρ and q. The behaviour of the ratio of the posterior variances agrees with the findings in Cox and Reid (2004): in the Bayesian setting also, the loss of information, as measured by the variance ratio, is greater when q is higher and has a u-shaped behaviour with respect to ρ.

The results on the simulated example illustrate the need for using the calibrated pairwise likelihood rather than the non-calibrated pairwise likelihood to perform Bayesian inference. The variability of $\pi_{PL}(\rho|y)$, in fact, is substantially lower than that of $\pi(\rho|y)$, thus it would lead to a falsely precise inference. By contrast, the variability of $\pi_{PL_c}(\rho|y)$ is greater than that of $\pi(\rho|y)$ to a degree that

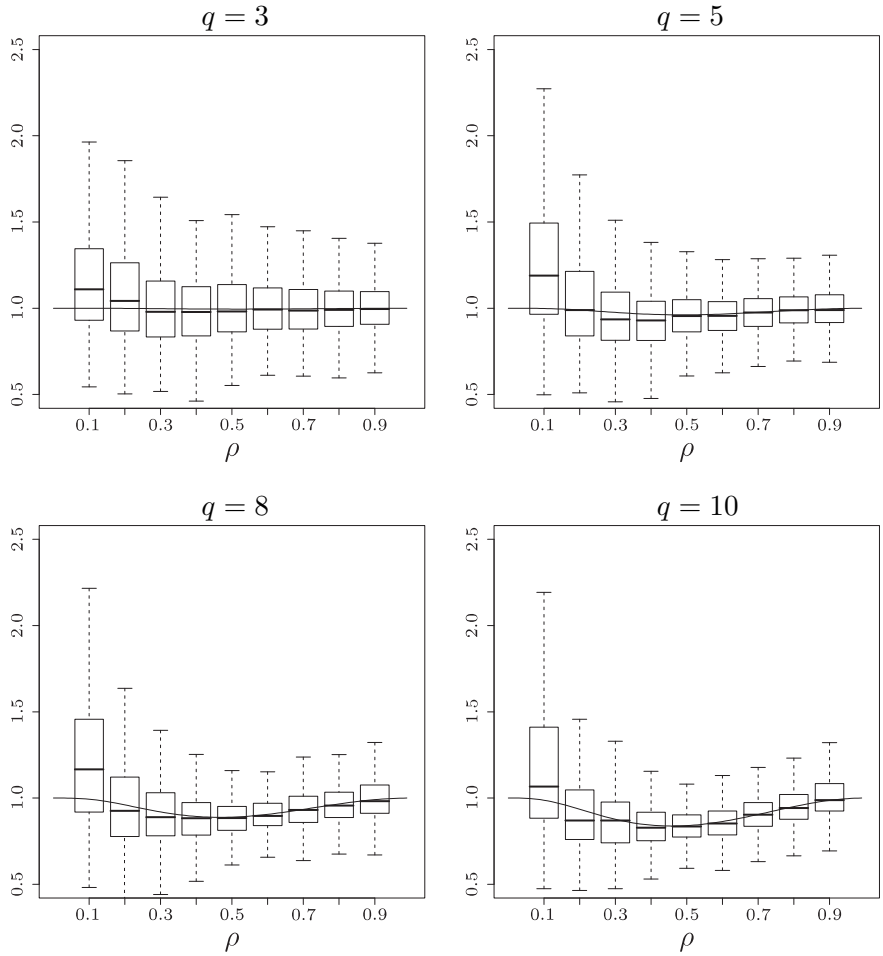

Figure 2. Ratio of the variance of $\pi(\rho|y)$ to the variance of $\pi_{PL_c}(\rho|y)$. Lines are the ratio of asymptotic variance of $\hat{\rho}$ to $\hat{\rho}_p$.

can be justified by the model misspecification (which leads to a loss of efficiency in frequentist inference).

4. Multivariate Extremes: A Simulation Study and an Application

We discuss two examples, both in the field of multivariate extreme values and, in particular, in the context of the multivariate extreme value distributions (MEVD). The first example employs simulated data from the logistic distribution, a sub-family of the MEVD for which the full likelihood is available, thus making possible the comparison between the posterior distributions $\pi(\theta|y)$ and $\pi_{PL_c}(\theta|y)$. The second example illustrates an application to a data set for which $\pi_{PL_c}(\theta|y)$ is used, since a sub-family of the MEVD, for which the full likelihood is not

available, is employed.

There are two basic approaches to dealing with multivariate extremes of a random vector: one can either model the exceedances of a suitable threshold or one can model the componentwise block maxima (see Beirlant et al. (2004)) for a recent review). The classical approach to model maxima is founded on a limit result analogous to the generalized extreme value theorem (Pickands (1981)). It states that if X_n is an independent and identically distributed random sequence from the random vector X in $\mathbb{R}^q_+$, then the limit distribution function of suitably normalized componentwise maxima $Y^{(n)} = \max_{j=1,\ldots,n} X_{i,j}$ $(i = 1, \ldots, q)$, belongs to the MEVD family:

$$G(y) = \exp\left\{ \int_{S_q} \max_{=1,\ldots,q} \left(\frac{u_j}{y_j} \right) dH(u) \right\} , \tag{4.1}$$

where H is a positive finite measure on the unit simplex in $(q-1)$ dimensions, denoted by S_q, satisfying the constraints $\int_{S_q} u_j dH(u) = 1$, $j = 1, \ldots, q$. The density h associated to the measure H is called the spectral density and charachterizes MEVD. Although this is a limit result, it is customary in applications to employ (4.1) as a model for the maxima of finite sequences of random variables like, for example, monthly maxima of concentrations of pollutants.

Contrary to what happens in the univariate case, it is not possible to describe parametrically the whole family of MEVD. Non-parametric estimation has been proposed in the literature (see, e.g., Beirlant et al. (2004)), but it is difficult to implement when, as is common in the field of extremes, few observations are available. To avoid these drawbacks, a number of parametric subfamilies of the MEVD have been proposed. Nonetheless, even after restricting to one of these families of distributions, likelihood inference is not an obvious task since the calculation of the multivariate density, which is needed for the full likelihood, can be tedious or practically infeasible. Finally, we note that to deal with multivariate extremes, other strategies have been proposed based on models outside of the MEVD family. For instance, Heffernan and Tawn (2004) suggest modelling extremes in one variable conditional on other variables, while Boldi and Davison (2007) build a family of multivariate models by defining the spectral density as a mixture of Dirichlet distributions (with appropriate restrictions) and show that the class of densities which is obtained is dense in the class of spectral densities associated to MEVD. We do not further discuss these proposals; we consider the issue of estimating the parameter of a distribution within the MEVD family, for which the use of an alternative likelihood is relevant.

The simplest sub-family of the MEVD is the logistic, with distribution function $G(y; \alpha) = \exp\left\{ \left(\sum_{j=1}^q y_j^{-1/\alpha} \right)^\alpha \right\}$, where $y_j \geq 0$, $j = 1, \ldots, q$, and $\alpha \in (0, 1]$.

The parameter α represents the strength of the dependence between the variables. We employ the routines included in the R package `evd` (Stephenson (2003)) to simulate from the logistic distribution and to compute the full likelihood $L(\alpha)$ and the bivariate densities needed for the calculation of $\mathrm{PL}(\alpha)$ and $\mathrm{PL}_c(\alpha)$. We simulated samples for a range of values of q and α and, for each of them, we compared the posterior distributions $\pi(\alpha|y)$, $\pi_{PL}(\alpha|y)$, and $\pi_{PL_c}(\alpha|y)$, assuming a uniform prior in (0,1) for α. In Figure 3 we show the comparison for a selection of values for α and q, i.e., with weak and strong dependence structure and with a different dimension of the random vector to be modeled. Note that, as in the example of Section 3. $\pi_{PL}(\alpha|y)$ implies less variability than $\pi(\alpha|y)$ in all the scenarios considered, but particularly when $q = 10$, while $\pi_{PL_c}(\alpha|y)$ is closer to $\pi(\alpha|y)$.

The logistic distribution entails a symmetric dependence structure modeled by a single parameter and does not allow for different degrees of dependence between the variables. This restriction is of course unrealistic and, in view of this, the scope of application of the logistic distribution is rather limited. Other sub-families of the MEVD have been proposed which do not share this limitation. An interesting example is the Coles and Tawn (CT) distribution (Coles and Tawn (1991)). The CT distribution is characterized by the spectral density

$$h(u) = \frac{\Gamma(1+\sum_i \alpha_i)}{(\sum_i \alpha_i u_i)^{d+1}} \prod_{j=1}^{d} \frac{\alpha_j}{\Gamma(\alpha_j)} \left(\frac{\alpha_j u_j}{\sum_i \alpha_i u_i} \right)^{\alpha_j - 1},$$

with q positive parameters α_j, $j = 1, \ldots, q$. If $q = 2$, the bivariate distribution function is

$$G(y_1, y_2; \alpha_1, \alpha_2) = \exp\left\{ \frac{1}{y_1} \left(1 - \frac{\Gamma(\alpha_1+\alpha_2+1)}{\Gamma(\alpha_1+1)\Gamma(\alpha_2)} \int_0^v w^{\alpha_1}(1-w)^{\alpha_2-1} dw \right) + \frac{1}{y_2} \frac{\Gamma(\alpha_1+\alpha_2+1)}{\Gamma(\alpha_1)\Gamma(\alpha_2+1)} \int_0^v w^{\alpha_1-1}(1-w)^{\alpha_2} dw \right\}, \tag{4.2}$$

with $v = \alpha_1 y_1/(\alpha_1 y_1 + \alpha_2 y_2)$. When $d > 2$, it is impractical to write the closed form expression for the multivariate density and applications are rather limited; one exception is Coles and Tawn (1994).

Using the composite likelihood approach, we can exploit an interesting property of the CT distribution. If $(Y_1, \ldots, Y_q)$ is distributed as a CT with parameters $(\alpha_1, \ldots, \alpha_q)$, then (Y_i, Y_j) is distributed as a bivariate CT with parameters (α_i, α_j), for each $i \neq j$. The parameter (α_i, α_j) is related to extremal dependence: it is greater the higher the values of the pair, while independence is approached as at least one of the two parameters approaches 0.

To illustrate the use of the posterior distribution (2.4), we employ the CT distribution in a typical application of extreme value theory: the concentration

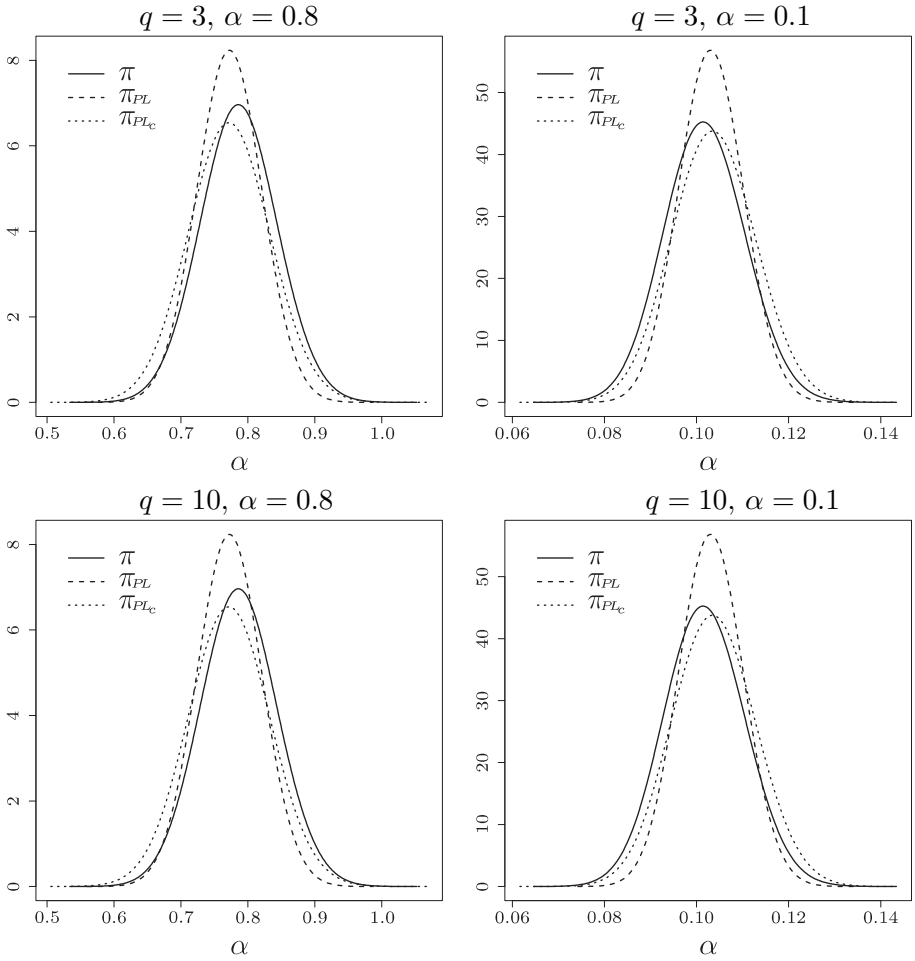

Figure 3. Comparisons of the posterior distributions based on the full likelihood (π), the non-calibrated pairwise likelihood (π_{PL}), and the calibrated pairwise likelihood (π_{PL_c}).

of atmospheric pollutants. We consider daily concentrations of five pollutants: NO, NO_2, O_3, SO_2, and $PM10$, measured in Leeds city centre from January 1993 to February 2009 (Source: UK National Air Quality Data Archive). Since the behaviour of the concentrations of the pollutants under consideration differs between seasons, winter months only (November, December, January, and February) are considered. For these months, the monthly maxima of the concentrations are computed for each pollutant, yielding a sample of 66 months, which reduce to 57 due to missing values (we keep observations only for those months for which data for all pollutants are available). Margins are standardized to unit-Fréchet distributions based on the parametric fitting of a generalized ex-

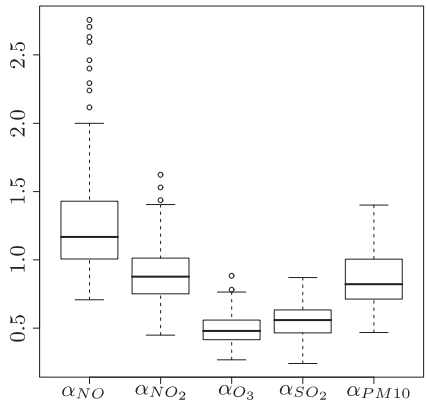

Figure 4. Margins of the posterior distribution $\pi_{PL_c}(\alpha|y)$ for the parameter of the CT distribution given the monthly maxima of pollutant concentrations in Leeds city centre.

treme value model to each pollutant time series, employing, when necessary, a time varying location parameter to eliminate any time trend. The standardized observations are then modeled according to a five-dimensional CT distribution with parameter $\alpha = (\alpha_N O, \alpha_{NO_2}, \alpha_{O_3}, \alpha_{SO_2}, \alpha_{PM10})$ (named after each variable). A flat proper prior implying independence among the α_i is assumed and a posterior distribution is obtained using a pairwise likelihood derived from the bivariate model (4.2), since the full likelihood $L(\alpha)$ is not available in a closed form. The margins of the corresponding posterior distribution $\pi_{PL_c}(\alpha|y)$ are depicted in Figure 4. It is worth noting that, although we justify our approach based on asymptotic properties of $\pi_{PL_c}(\alpha|y)$, in applications one can obtain asymmetric densities as for α_{NO} in Figure 4 (as also underlined in Smith and Stephenson (2009)).

Extremes of pollutant concentrations in Leeds city centre were also analyzed by Boldi and Davison (2007) and Heffernan and Tawn (2004), although with a different model and referring to a different time period. Nonetheless, we remark that the results are broadly consistent in revealing that NO, NO_2, and $PM10$ show a stronger extremal dependence than the other two pollutants.

5. Final Remarks

In this paper, the suitability of Bayesian inference with composite marginal likelihoods is investigated. It is argued that the use of composite likelihoods with unit weights may lead, due to model misspecification, to an unreasonable inference. In particular, the posterior variability may not reflect the uncertainty on the parameter just, like the variability of the composite maximum likelihood estimator is not reflected by the shape of the composite likelihood. For this reason

the proposed posterior distribution involves a calibrated composite likelihood, i.e., a composite likelihood with a particular choice of the weights. The proposed adjustment alleviates inefficiency of composite likelihood methods and approximates the usual asymptotic behavior of the resultant posterior distribution. In the literature, other adjustments to the composite likelihood ratio statistic have been proposed; see for a review Varin, Reid and Firth (2009, Sec. 2.3). Here we focused on the simplest one based on first order moment matching, but an investigation of different adjustments could be of interest.

The example and simulation results presented in this paper show that $\pi_{PL_a}(\theta|y)$ can be exploited to perform Bayesian inference in complex models. In all the examples standard choices for the prior distribution have been made, in order to allow, when possible, the comparison between $\pi_{PL_a}(\theta|y)$ and the full posterior distribution (2.1). One might resort to different elicitations of the prior distribution $\pi(\theta)$; this is an intriguing prospect when referring to default priors, such as Jeffreys' type or matching priors; see, for instance, Ventura, Cabras, and Racugno (2010).

Acknowledgements

The authors acknowledge an associate editor and anonymous referees for many useful comments that greatly improved the paper. This work was supported in part by grants from Ministero dell'Università e della Ricerca Scientifica e Tecnologica, Italy.

References

Azzalini, A. (1983). Maximum likelihood of order m for stationary stochastic processes. *Biometrika* **70**, 381-367.

Beirlant, J., Goegebeur, Y., Segers, J., Teugels, J., De Waal, D. and Ferro, C. (2004). *Statistics of Extremes: Theory and Applications*. Wiley, UK.

Besag, J. E. (1974). Spatial interaction and the statistical analysis of lattice systems (with discussion). *J. Roy. Statist. Soc. Ser. B* **34**, 192-236.

Besag, J. E. (1977). Efficiency of pseudolikelihood estimation for simple Gaussian fields. *Biometrika* **64**, 616-618.

Boldi, M. O. and Davison, A. C. (2007). A mixture model for multivariate extremes. *J. Roy. Statist. Soc. Ser. B* **69**, 217-229.

Chandler, R. E. and Bate, S. (2007). Inference for clustered data using the independence log-likelihood. *Biometrika* **94**, 167-183.

Chang, H. and Mukerjee, R. (2006). Probability matching property of adjusted likelihoods. *Statist. Prob. Lett.* **76**, 838-842.

Chernozhukov, V. and Hong, H. (2003). An MCMC approach to classical estimation. *J. Econom.* **115**, 293-346.

Coles, S. G. and Tawn, J. A. (1991). Modelling extreme multivariate events. *J. Roy. Statist. Soc. Ser. B* **53**, 377-392.

Coles, S. G. and Tawn, J. A. (1994). Statistical methods for multivariate extremes: an application to structural design. *J. Roy. Statist. Soc. Ser. C* **43**, 1-48.

Cox, D. R. (1975). Partial likelihood. *Biometrika* **62**, 269-276.

Cox, D. R. and Reid, N. (2004). A note on pseudolikelihood constructed from marginal densities. *Biometrika* **91**, 729-737.

Engler, D. A., Mohapatra, M., Louis, D. N. and Betensky, R. A. (2006). A pseudolikelihood approach for simultaneous analysis of array comparative genomic hybridizations. *Biostatistics* **7**, 399-421.

Fraser, D. A. S., Reid, N., Wong, A. and Yun Yi, G. (2003). Direct Bayes for interest parameters. *Valencia* **7**, 529-533.

Geys, H., Molenberghs, G. and Ryan, L. M. (1999). Pseudolikelihood modelling of multivariate outcomes in developmental toxicology. *J. Amer. Statist. Assoc.* **94**, 734-745.

Godambe, V. P. (1960). An optimum property of regular maximum likelihood equation. *Ann. Math. Statist.* **31**, 1208-1211.

Greco, L., Racugno,W. and Ventura, L. (2008). Robust likelihood functions in Bayesian inference. *J. Statist. Plann. Inference* **138**, 1258-1270.

Hanfelt, J. J. and Liang, K. Y. (1995). Approximate likelihood ratios for general estimating functions. *Biometrika* **82**, 461-477.

Heffernan, J. E. and Tawn, J. A. (2004). A conditional approach for multivariate extreme values, *J. Roy. Statist. Soc. Ser. B* **66**, 497-546.

Lazar, N. A. (2003). Bayesian empirical likelihood. *Biometrika* **90**, 319-326.

Le Cessie, S. and Van Houwelingen, J. C. (1994). Logistic regression for correlated binary data. *J. Roy. Statist. Soc. Ser. B* **43**, 95-108.

Lin, L. (2006). Quasi Bayesian likelihood. *Statist. Method.* **3**, 444-455.

Lindsay, B. G. (1988). Composite likelihood methods. *Contemporary Mathematics* **80**, 221-240.

Molenberghs, G. and Verbeke, G. (2005). *Models for Discrete Longitudinal Data*. Springer-Verlag, New York.

Monahan, J. F. and Boos, D. D. (1992). Proper likelihoods for Bayesian analysis. *Biometrika* **79**, 271-278.

Pickands, J. (1981). Multivariate extreme value distributions. *Proc. 43rd Sess. Int. Statist. Inst.*, 859-878.

Racugno, W., Salvan, A. and Ventura, L. (2010). Bayesian analysis in regression models using pseudo-likelihoods. *Comm. Statist. Theory Methods*, to appear.

Raftery, A. E., Madigan, D. and Volinsky, C. T. (1996). Accounting for model uncertainty in survival analysis improves predictive performance. *Bayesian Statistics* **6**, 323-349.

Reid, N. (1995). Likelihood and Bayesian approximation methods. *Bayesian Statistics* **5**, 351-368.

Reid, N. (2003). The 2000 Wald memorial lectures: Asymptotics and the theory of inference. *Ann. Statist.* **31**, 1695-1731.

Severini, T. A. (1999), On the relationship between Bayesian and non-Bayesian elimination of nuisance parameters. *Statist. Sinica* **9**, 713-724.

Stephenson, A. G. (2003). Simulating multivariate extreme value distributions of logistic type. *Extremes* **6**, 49-60.

Smith, E. L. and Stephenson, A. G. (2009). An extended Gaussian max-stable process model for spatial extremes. *J. Statist. Plann. Inference* **139**, 1266-1275.

Sweeting, T. J. (1987). Approximate Bayesian analysis of censored survival data. *Biometrika* **74**, 809-816.

Tierney, L. and Kadane, J. B. (1986). Accurate approximations for posterior moments and marginal densities. *J. Amer. Statist. Assoc.* **81**, 82-86.

Varin, C. (2008). On composite marginal likelihoods. *Adv. Statist. Anal.* **92**, 1-28.

Varin, C., Reid, N. and Firth, D. (2009). An overview of composite likelihood methods. *CRiSM Working Paper*, University of Warwick.

Varin, C. and Vidoni, P. (2005). A note on composite likelihood inference and model selection. *Biometrika* **92**, 519-528.

Ventura, L., Cabras, S. and Racugno, W. (2009). Prior distributions from pseudo-likelihoods in the presence of nuisance parameters. *J. Amer. Statist. Assoc.* **104**, 768-774.

Ventura, L., Cabras, S. and Racugno, W. (2010). Default prior distributions from quasi- and quasi-profile likelihoods. *J. Statist. Plann. Inference* **140**, 2937-2942.

Department of Statistics, University of Padua, Via C. Battisti 241, 35121 Padova, Italy.

E-mail: fpauli@stat.unipd.it

Department of Mathematics and Informatics, University of Cagliari, Via Ospedale 72, 09123 Cagliari, Italy.

E-mail: racugno@unica.it

Department of Statistics, University of Padua, Via C. Battisti 241, 35121 Padova, Italy.

E-mail: ventura@stat.unipd.it

(Received September 2009; accepted July 2010)