Bayesian Inference from Composite Likelihoods, with an Application to Spatial Extremes

Mathieu Ribatet[†,*] Daniel Cooley[‡] Anthony C. Davison[§]

June 25, 2011

[†] Department of Mathematics, Université Montpellier II

[‡] Department of Statistics, Colorado State University

[§] Institute of Mathematics, École Polytechnique Fédérale de Lausanne

Abstract

Composite likelihoods are increasingly used in applications where the full likelihood is analytically unknown or computationally prohibitive. Although some frequentist properties of the maximum composite likelihood estimator are akin to those of the maximum likelihood estimator, Bayesian inference based on composite likelihoods is in its early stages. This paper discusses inference when one uses composite likelihood in Bayes' formula. We establish that using a composite likelihood results in a proper posterior density, though it can differ considerably from that stemming from the full likelihood. Building on previous work on composite likelihood ratio tests, we use asymptotic theory for misspecified models to propose two adjustments to the composite likelihood to obtain appropriate inference. We also investigate use of the Metropolis Hastings algorithm and two implementations of the Gibbs sampler for obtaining draws from the composite posterior. We test the methods on simulated data and apply them to a spatial extreme rainfall dataset. For the simulated data, we find that posterior credible intervals yield appropriate empirical coverage rates. For the extreme precipitation data, we are able to both effectively model marginal behavior throughout the study region and obtain appropriate measures of spatial dependence.

Keywords: Bayesian hierarchical model; Composite likelihood; Gibbs sampler; Markov chain Monte Carlo; Max-stable process; Metropolis–Hastings algorithm; Posterior coverage; Rainfall data.

*Corresponding author. Email: mathieu.ribatet@math.univ-montp2.fr. Phone: +33 (0)4 67 14 41 98

1 Introduction

1.1 Motivation

The likelihood function is central to both frequentist and Bayesian inference, but in many modern settings it may be infeasible to calculate it, either because no analytical form is available, or because such a form is known but is computationally prohibitive. The first difficulty arises with max-stable processes, which are used to construct probability models for complex rare events, but for which closed forms are typically available only for the bivariate marginal densities [Smith, 1990; Schlather, 2002; de Haan and Pereira, 2006; Kabluchko et al., 2009], though Genton et al. [2011] show that substantial efficiency gains are possible if trivariate margins can be used. The second difficulty may be experienced when dealing with Gaussian random fields on large lattices. Both of these problems and many other similar ones can be tackled using composite likelihoods. Padoan et al. [2010] and Gholamrezaee [2010] propose the use of composite likelihood based on marginal events to fit max-stable processes, and Rue and Tjelmeland [2002] have used composite likelihoods based on omitting components of the full likelihood in approximating Gaussian random fields. Rydén and Titterington [1998] describe the use of pseudo-likelihood, a form of composite likelihood, in simulation-based inference involving missing data, and show that their approach leads to a valid Markov chain simulation algorithm.

Frequentist methods for composite likelihoods have been used for some time (for an overview, see Varin [2008]), but little work has been done to explore how composite likelihoods could be employed in a Bayesian framework. The motivating application for this work is the spatial modelling of extremes. Recently authors (e.g., Padoan et al. [2010] and Gholamrezaee [2010]) have used composite likelihoods to fit max-stable models, enabling the researchers to successfully model dependence between observations. However, the frequentist methods they employ may not be flexible enough to accurately fit the marginal behavior across the study region. Cooley et al. [2007] and Sang and Gelfand [2009] have used Bayesian hierarchical spatial models to capture the marginal effects for spatial extremes, but have not used the max-stable process models suggested by extreme value theory to describe the dependence in the data. The goal of this work is combine these two approaches, and this entails appropriately deploying a composite likelihood within a Bayesian framework.

1.2 Likelihood asymptotics for composite likelihoods

Although it has numerous antecedents, the notion of a composite likelihood was crystallized by Lindsay [1988], who defined it as a combination of valid likelihood entities. Consider a random vector $Y \in \mathbb{R}^K$ with

probability density function $f(y;\theta)$ where $\theta \in \mathbb{R}^p$ is an unknown parameter vector. Let $\{\mathscr{A}_i : i \in I\}$, $I \subset \mathbb{N}$, be a set of marginal or conditional events for Y and let $\{w_i, i \in I\}$ be a set of non-negative weights. A composite likelihood is defined as

$$L_c(\theta; y) = \prod_{i \in I} f(y \in \mathscr{A}_i; \theta)^{w_i}, \tag{1}$$

with corresponding log-composite likelihood

$$\ell_c(\theta; y) = \sum_{i \in I} w_i \log f(y \in \mathscr{A}_i; \theta). \tag{2}$$

Below we assume that n independent replicates $Y^1, \ldots, Y^n$ of Y are available, yielding a total composite likelihood and log likelihood of the form

$$L_c^{\text{tot}}(\theta; y) = \prod_{j=1}^{n} \prod_{i \in I} f(y^j \in \mathscr{A}_i; \theta)^{w_i}, \quad \ell_c^{\text{tot}}(\theta; y) = \sum_{j=1}^{n} \sum_{i \in I} w_i \log f(y^j \in \mathscr{A}_i; \theta),$$

and consider asymptotics as $n \to \infty$, with a fixed number of observations K in each replicate. The development below is simpler if we work with quantities that remain of order one as $n \to \infty$, and we shall do so wherever possible.

If the true likelihood is unavailable or difficult to work with, θ is often estimated by the maximum composite likelihood estimator $\hat{\theta}_c$. Let θ_0 denote the true value of the parameter. As each term on the right-hand side of equation (2) is a valid loglikelihood, the composite score function $\nabla \ell_c^{\text{tot}}(\theta; y)$ is a linear combination of unbiased estimating functions and so has mean zero. Under appropriate regularity conditions, therefore, the maximum composite likelihood estimator $\hat{\theta}_c$ converges in distribution as follows,

$$\sqrt{n}\{H(\theta_0)J(\theta_0)^{-1}H(\theta_0)\}^{1/2}(\hat{\theta}_c - \theta_0) \xrightarrow{d} N\left(0, \mathrm{Id}_p\right), \qquad n \to \infty, \tag{3}$$

where $M^{1/2}$ denotes a matrix square root, i.e., $\{M^{1/2}\}^T M^{1/2} = M$, Id_p denotes the $p \times p$ identity matrix, $H(\theta_0) = -\mathbb{E}[\nabla^2 \ell_c(\theta_0; Y)]$ and $J(\theta_0) = \mathrm{Var}[\nabla \ell_c(\theta_0; Y)]$, where the expectations are with respect to the full density. Both $H(\theta_0)$ and $J(\theta_0)$ are positive definite in a regular model, and both are of order one as $n \to \infty$.

Essentially the usual regularity conditions for the asymptotic normality of the maximum likelihood estimator as $n \to \infty$ apply [Davison, 2003, Sec. 4.4.2], but the parameter θ must be identifiable from the densities appearing in (2). The limiting distribution in equation (3) also stems from the behavior of the maximum likelihood estimator under mis-specification [Kent, 1982]. The maximum composite likelihood

estimator may thus be viewed as resulting from a mis-specified, or, more accurately, under-specified, statistical model, leading to consistent estimation but with a "sandwich" variance estimator of the type arising in longitudinal data analysis and many other domains.

1.3 Bayesian inference with a composite likelihood

Bayesian inference based on composite likelihoods has been little explored. Motivated by the spatial extremes problem mentioned above, Smith and Stephenson [2009] use a pairwise likelihood and Markov chain Monte Carlo simulation to fit a max-stable model for rainfall at five sites in South-West England. They obtain a posterior by replacing the unavailable full likelihood with the pairwise likelihood, but although they mention that this substitution may lead to overly precise inferences, they do not describe how to correct this. Pauli et al. [2011] independently suggest the adjustment to the composite likelihood called by us the magnitude adjustment in Section 2.1, establish the asymptotic normality of the corresponding composite posterior and apply the method to a five-dimensional data set on air pollution.

Related to Bayesian inference with composite likelihoods is work in Bayesian methods when one lacks or wishes to avoid using the true likelihood. Monahan and Boos [1992] explore the validity of a posterior when the likelihood is not the conditional density of the data given the parameter, and propose both an alternative definition based on the coverage of posterior sets and a test that can be used to invalidate a particular replacement likelihood. Lazar [2003] applies this test when an empirical likelihood is used in place of a parametric one. Other work on Bayesian methods with conditional or pseudo likelihoods (e.g., Efron [1993], Chang and Mukerjee [2006] and Ventura et al. [2009]) is typically motivated by a desire to avoid specifying a full likelihood when there are nuisance parameters, and thus focuses on Bayesian implementation using a pseudo-likelihood, which is often a marginal, conditional or profile likelihood for the parameters of interest. Like Monahan and Boos, our ultimate aim is the practical one of using a composite likelihood to provide valid inferences; for example, the resulting posterior confidence sets should be correctly calibrated.

Provided that $\int L_c^{\mathrm{tot}}(\theta; y)\pi(\theta)\mathrm{d}\theta$ is finite, we use (1) to define a composite posterior density as

$$\pi_c(\theta \mid y) = \frac{L_c^{\mathrm{tot}}(\theta; y)\pi(\theta)}{\int L_c^{\mathrm{tot}}(\theta; y)\pi(\theta)\mathrm{d}\theta}, \tag{4}$$

where $\pi(\cdot)$ is the prior density. The first question arising is under what circumstances $\int L_c^{\mathrm{tot}}(\theta; y)\pi(\theta)\mathrm{d}\theta < \infty$, so that (4) is well-defined. In Bayesian analysis, integrability questions usually arise when discussing improper priors; but here we suppose that $\pi(\cdot)$ is proper. Then a sufficient condition for (4) to be proper is that for

each i there exists a finite b_i such that $\sup_\theta f(y \in \mathscr{A}_i; \theta) \leq b_i$, since in that case

$$\int L_c^{\mathrm{tot}}(\theta; y)\pi(\theta)\mathrm{d}\theta = \int \prod_j \prod_{i \in I} f(y^j \in \mathscr{A}_i; \theta)^{w_i} \pi(\theta)\mathrm{d}\theta \leq \prod_{i \in I} b_i^{n w_i} < \infty. \tag{5}$$

The boundedness of $f(y \in \mathscr{A}_i; \theta)$ holds in many cases, and in cases of doubt it can be imposed by recalling that in practice continuous observations are always rounded to some extent. The correct likelihood is therefore a product of probabilities obtained as differences of cumulative distribution functions, for which $b_i \equiv 1$. The rounding is often ignored so that simpler density function approximations to the correct likelihood may be used, but if these approximations lead to difficulties, then we may choose to work with the correct likelihood; see, e.g., Copas [1972]. As this rounding argument applies to any probability elements in (1), and, with minor changes, also applies to the modified composite likelihoods used below, in practice we may always arrange that $\int L_c^{\mathrm{tot}}(\theta; y)\pi(\theta)\mathrm{d}\theta < \infty$ and thus that (4) is proper.

Since the composite likelihood is not the likelihood believed to have generated the data, the naive implementation of a composite posterior may give misleading inferences, as we now illustrate.

Example 1. Let $\{Y(x)\}$ be a stationary Gaussian process with unknown mean $\mu \in \mathbb{R}$ and with covariance function $\gamma(h) = \tau \exp(-h/\omega)$, where the sill $\tau > 0$ is unknown but the scale $\omega > 0$ is known. Let $\{y(x_1), \ldots, y(x_K)\}$ be one realisation of this process at locations $x_1, \ldots, x_K \in \mathbb{R}$. Now consider a prior density of the form $\pi(\theta) = \pi(\mu)\pi(\tau)$, where $\pi(\mu) \sim N(a, b)$ and $\pi(\tau) \sim \mathrm{IG}(c, d)$, i.e., an inverse Gamma distribution with shape c and scale d.

Here the prior densities are conjugate for $\pi(\theta \mid y)$, so the full conditional distributions needed for Gibbs sampling are easily found to be

$$\pi(\mu \mid \cdots) \sim N\left(\tilde{\mu}, \tilde{\sigma}^2\right), \qquad \pi(\tau \mid \cdots) \sim \mathrm{IG}\left\{c + \frac{K}{2}, d + \frac{1}{2}(y - \mu\mathbf{1})^T \Sigma^{-1} (y - \mu\mathbf{1})\right\},$$

where $\tilde{\sigma}^2 = \left(b^{-1} + \tau^{-1}\mathbf{1}^T\Sigma^{-1}\mathbf{1}\right)^{-1}$, $\tilde{\mu} = \tilde{\sigma}^2\left(ab^{-1} + \tau^{-1}\mathbf{1}^T\Sigma^{-1}y\right)$ and Σ is the correlation matrix derived from $\gamma(\cdot)$.

The full conditional pairwise distributions are also readily available, and are

$$\pi_p(\mu \mid \cdots) \sim N\left(\tilde{\mu}_p, \tilde{\sigma}_p^2\right), \qquad \pi_p(\tau \mid \cdots) \sim \mathrm{IG}\left\{c + \frac{K(K-1)}{2}, d + \frac{1}{2}(y_p - \mu\mathbf{1})^T \Sigma_p^{-1} (y_p - \mu\mathbf{1})\right\},$$

where $\tilde{\sigma}_p^2 = \left(b^{-1} + \tau^{-1}\mathbf{1}^T\Sigma_p^{-1}\mathbf{1}\right)^{-1}$, $\tilde{\mu}_p = \tilde{\sigma}_p^2\left(ab^{-1} + \tau^{-1}\mathbf{1}^T\Sigma_p^{-1}y_p\right)$, Σ_p is a block diagonal matrix with

blocks

$$\begin{bmatrix} 1 & \tau^{-1}\gamma(x_i - x_j) \\ \tau^{-1}\gamma(x_i - x_j) & 1 \end{bmatrix}, \qquad 1 \leq i < j \leq K,$$

and $y_p = (y_1, y_2, y_1, y_3, \ldots, y_1, y_K, y_2, y_3, \ldots, y_2, y_K, \ldots, y_{K-1}, y_K)^T$. □

Example 1 shows that, as might be expected, the full conditional densities derived from the pairwise likelihood differ from those derived from the full likelihood. Since $\mathbf{1}^T A \mathbf{1}$ is the sum of all entries of the matrix A and Σ_p is block diagonal, it is not difficult to show that

$$\mathbf{1}^T \Sigma_p^{-1} \mathbf{1} = 2 \sum_{i=1}^{K-1} \sum_{j=i+1}^{K} \left\{1 + \tau^{-1}\gamma(x_i - x_j)\right\}^{-1} \geq \frac{\tau K(K-1)}{1+\tau}, \qquad \mathbf{1}^T \Sigma^{-1} \mathbf{1} \leq K,$$

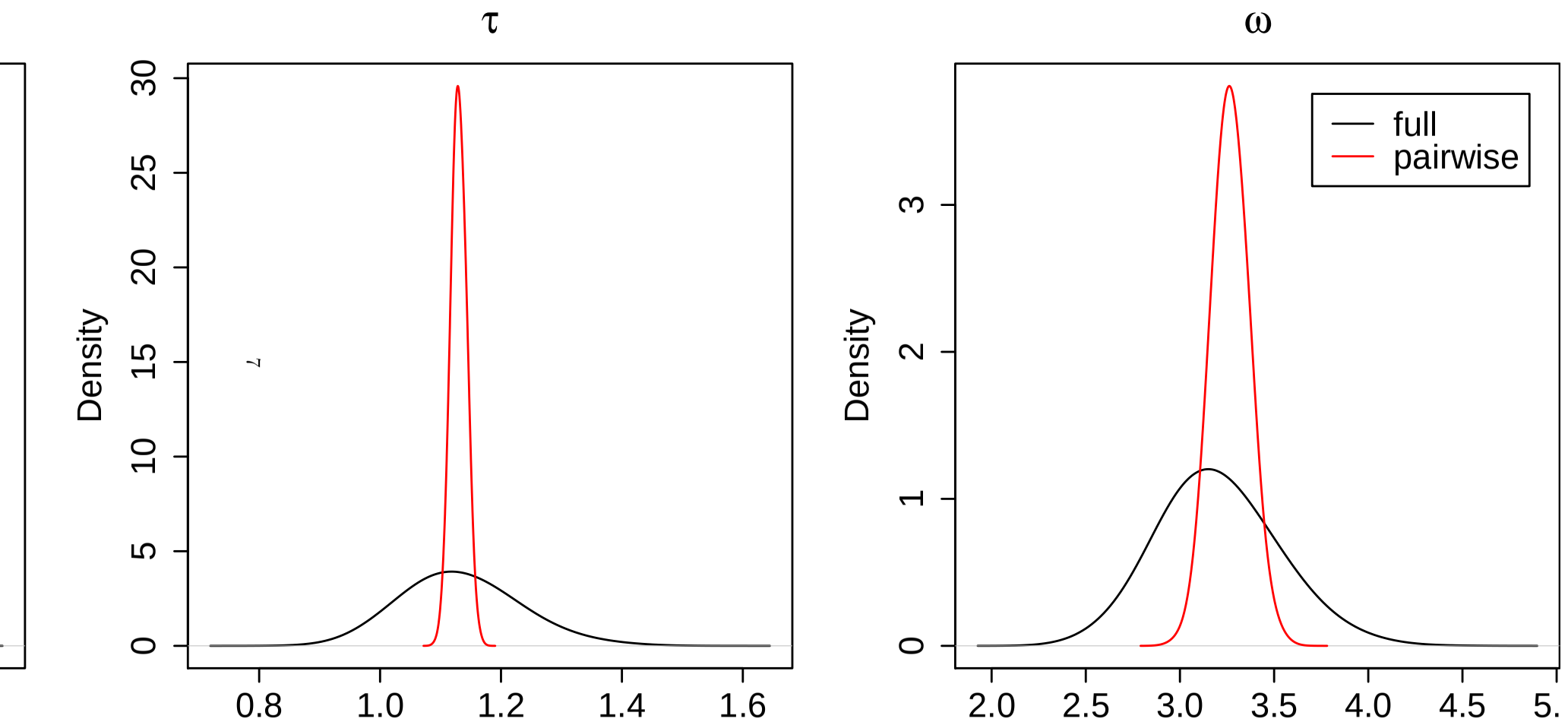
so, in particular,

$$\frac{\tilde{\sigma}_p^2}{\tilde{\sigma}^2} \leq \frac{(1+\tau)(\tau + bK)}{(1+\tau)\tau + b\tau K(K-1)},$$

and when τ is fixed, $\tilde{\sigma}_p^2/\tilde{\sigma}^2 \downarrow 0$ as $K \to \infty$.

To illustrate this discussion, Figure 1 shows posterior marginal density estimates for μ and τ based on the composite and full likelihoods, found using a Gibbs sampler. These densities were obtained by taking the same setting as in Example 1 with $\mu = 0$, $\tau = 1$ and $\omega = 3$, the last taken as constant in the sampling algorithm, with $K = 20$, and with the locations $x_1, \ldots, x_K$ taken uniformly at random in $[0, 20]$. There are $n = 50$ independent replicates of these data. A Gaussian prior with mean 0 and variance 100 was placed on μ, and independently an inverse gamma prior with shape 1/10 and scale 1 was placed on τ. The marginal composite posterior densities are much too concentrated, because the pairwise likelihood treats the pairs of observations as though they were mutually independent and thus uses each observation repeatedly—see the definition of y_p in Example 1.

The aim of this paper is to propose a framework for approximate Bayesian inference from composite likelihoods when the full likelihood is not available. Our aim is to obtain composite posterior distributions that give credible intervals with reasonable coverage. Section 2 introduces two adjustments to the composite likelihood that are intended to retrieve some of the desirable properties given by the usual likelihood. Section 3 shows how these adjustments can be incorporated into Markov chain Monte Carlo samplers, and their performance in simulation studies is discussed in Section 4. Section 5 gives a case study on the modelling of extreme rainfall around Zurich. The paper closes with a brief discussion and two technical appendices.


τ
Density
0
5
10
15
20
25
30
0.8
1.0
1.2
1.4
1.6
ω
Density
0
1
2
3
2.0
2.5
3.0
3.5
4.0
4.5
5.
full
pairwise

2 Adjustment of the composite likelihood

We ultimately wish to perform a Bayesian analysis, in which setting there is no "true" parameter value θ_0. However, we use asymptotic relationships developed under the frequentist paradigm to adjust the likelihood to obtain appropriate inference for the composite posterior, and thus speak of θ_0 throughout this section.

The theory of unbiased estimating functions applied to the score functions of composite likelihood implies that under suitable regularity conditions, the modes of a composite posterior and of the full posterior density will approach one another as the sample size increases; see Figure 1. However, the figure also shows that the composite posterior density can differ significantly in spread from the true one, because the composite likelihood treats the events $\{\mathscr{A}_i, i \in I\}$ as though they were mutually independent. Below we seek to modify the composite likelihood in order to mitigate this.

Suppose that the parameter $\theta = (\phi^T, \psi^T)^T$ has true value $\theta_0 = (\phi_0^T, \psi_0^T)^T$ and that ψ contains q elements. Let $\tilde{\theta}$ be the restricted maximum likelihood estimator, obtained by maximizing the full log likelihood $\ell(\theta; Y)$ over θ with ψ held fixed at ψ_0 and let $\tilde{\theta}_c$ be the restricted maximum composite likelihood estimator, which maximizes (1) with ψ held fixed at ψ_0. Then as $n \to \infty$,

$$\Lambda(\psi_0) = 2\{\ell(\hat{\theta}; Y) - \ell(\tilde{\theta}; Y)\} \xrightarrow{d} \chi^2_q \tag{6}$$

whereas for the composite likelihood,

$$\Lambda_c(\psi_0) = 2\{\ell_c(\hat{\theta}_c; Y) - \ell_c(\tilde{\theta}_c; Y)\} \xrightarrow{d} \sum_{i=1}^{q} \lambda_i X_i \tag{7}$$

where $X_1, \ldots, X_q$ are independent χ^2_1 random variables, $\lambda_1, \ldots, \lambda_q$ are the eigenvalues of the $q \times q$ matrix $\{H(\theta_0)^{-1} J(\theta_0) H(\theta_0)^{-1}\}_\psi [\{H(\theta_0)^{-1}\}_\psi]^{-1}$, and A_ψ denotes the sub-matrix of a matrix A corresponding to the elements of ψ [Kent, 1982]. These relationships have previously been exploited to provide likelihood ratio tests [Rotnitzky and Jewell, 1990; Chandler and Bate, 2007] suitable for misspecified models. Here we aim to recover convergence in distribution to the usual χ^2 distribution through two different modifications of the composite likelihood: a magnitude adjustment and a curvature adjustment. The reasons for such modifications is to make the composite likelihood ratio, which appears in the Metropolis–Hastings algorithm but is hidden in the Gibbs sampler, behave in distribution as it would if a full likelihood were available. In the remainder of this section we consider only the case where ψ has dimension zero, but in Section 3.2.2 we will show how partitioning θ can yield better coverage.

2.1 Magnitude adjustment

The magnitude adjustment to the composite log likelihood is inspired by Rotnitzky and Jewell [1990], who, in the context of hypothesis testing in longitudinal studies, estimate $\lambda_1, \ldots, \lambda_q$ from estimates of $H(\theta_0)$ and $J(\theta_0)$, and use them to calculate the appropriate rejection region for the χ^2 test based on (7).

We define the magnitude adjustment by

$$\ell_{\text{magn}}(\theta; y) = k\ell_c^{\text{tot}}(\theta; y), \qquad \theta \in \Theta, \tag{8}$$

where k is a positive constant; (8) was also suggested by Pauli et al. [2011]. With this modification and as $n \to \infty$ we have

$$\Lambda_{\text{magn}}(\psi_0) = 2\{\ell_{\text{magn}}(\hat{\theta}_c; Y) - \ell_{\text{magn}}(\tilde{\theta}_c; Y)\} \xrightarrow{d} k \sum_{i=1}^{q} \lambda_i X_i \tag{9}$$

and

$$\mathbb{E}[\Lambda_{\text{magn}}(\psi_0)] \longrightarrow k \sum_{i=1}^{q} \lambda_i, \qquad \text{Var}[\Lambda_{\text{magn}}(\psi_0)] \longrightarrow 2k^2 \sum_{i=1}^{q} \lambda_i^2.$$

Setting $k = q/\sum_{i=1}^{q} \lambda_i$ therefore ensures that $\mathbb{E}[\Lambda_{\text{magn}}(\psi_0)]$ converges to $\mathbb{E}[\chi_q^2] = q$, but the higher moments of (9) will not match those of χ_q^2 unless all the λ_i's are equal or $q = 1$. For our purposes, we consider the case where ϕ has dimension zero, i.e., $k = p/\sum_{i=1}^{p} \lambda_i$ where $\lambda_1, \ldots, \lambda_p$ are the eigenvalues of $H(\theta_0)^{-1}J(\theta_0)$. Varin [2008] proposes a Satterthwaite adjustment to match the first two moments of $\Lambda_{\text{magn}}(\psi_0)$ and χ_q^2, though their higher moments would still differ.

2.2 Curvature adjustment

Another strategy is to modify the curvature of the composite likelihood around its global maximum $\hat{\theta}_c$ by considering the adjustment given by

$$\ell_{\text{curv}}(\theta; y) = \ell_c^{\text{tot}}(\theta^*; y), \qquad \theta^* = \hat{\theta}_c + C(\theta - \hat{\theta}_c), \tag{10}$$

for some constant $p \times p$ matrix C. Clearly $\hat{\theta}_c$ is also a global maximum for ℓ_{curv}, and

$$\nabla \ell_{\text{curv}}(\theta; y) = C^T \nabla \ell_c^{\text{tot}}(\theta; y)|_{\theta=\theta^*}, \qquad \nabla^2 \ell_{\text{curv}}(\theta; y) = C^T \nabla^2 \ell_c^{\text{tot}}(\theta; y)|_{\theta=\theta^*} C.$$

Under mild conditions, Taylor expansion of the usual log-likelihood and the asymptotic normality of the

maximum likelihood estimator $\hat{\theta}$ yield convergence of the likelihood ratio statistic in distribution to a χ^2 variable [Davison, 2003, Sec. 4.5]. More precisely, the facts that

$$\Lambda(\theta_0) \xrightarrow{d} n(\hat{\theta}-\theta_0)^T\Sigma(\hat{\theta}-\theta_0), \quad n\to\infty,$$

for some $q \times q$ covariance matrix Σ depending only on $\mathbb{E}[\nabla^2\ell(\theta_0; Y)]$ and

$$\sqrt{n}\Sigma^{1/2}(\hat{\theta}-\theta_0) \xrightarrow{d} N(0, \mathrm{Id}_p), \quad n\to\infty,$$

ensure that $\Lambda(\theta_0)$ converges in distribution to a χ^2_p variable. This occurs because $-n^{-1}\nabla^2\ell(\hat{\theta}; y)$ converges almost surely to the rescaled inverse of the asymptotic covariance matrix of the maximum likelihood estimator, the Fisher information in a single Y.

This suggests that we should try to ensure that $-n^{-1}\nabla^2\ell_{\mathrm{curv}}(\hat{\theta}_c; y)$ converges almost surely to the inverse of the asymptotic covariance matrix of the maximum composite likelihood estimator, i.e., $H(\theta_0)J(\theta_0)^{-1}H(\theta_0)$, by taking any semi-definite negative matrix C such that

$$C^TH(\theta_0)C = H(\theta_0)J(\theta_0)^{-1}H(\theta_0). \tag{11}$$

One possible choice, $C = M^{-1}M_A$, where $M_A^TM_A = H(\theta_0)J(\theta_0)^{-1}H(\theta_0)$ and $M^TM = H(\theta_0)$, corresponds to a suggestion of Chandler and Bate [2007] for hypothesis testing for clustered data using the independence log-likelihood. However, the matrix square roots M and M_A are not unique, and although the choice is immaterial for composite likelihoods that are quadratic in the neighborhood of $\hat{\theta}_c$, it might be necessary to ensure that the mapping (10) preserves any directions of asymmetry. For this reason we use singular value decompositions for M and M_A for the curvature adjustments in this paper.

2.3 Properties of the adjustments

Although both adjustments rely on the idea of recovering the usual convergence to a χ^2 variable, they express different aspects of this. The magnitude adjustment (8) is an "overall" adjustment, intended to scale the composite likelihood down to the appropriate magnitude; in Figure 1 it amounts to raising the narrower curve to a power and thus giving a nonlinear transformation of the vertical axis. Therefore all (local) extrema are left unchanged, because $\nabla\ell_{\mathrm{magn}}(\theta; y) = 0$ implies that $\nabla\ell_c^{\mathrm{tot}}(\theta; y) = 0$, and the composite and full posterior modes should be approximately the same, because the composite score function has mean zero.

The curvature adjustment (10), on the other hand, stretches the horizontal axis linearly so that the curvature of $\ell_{\text{curv}}(\theta; y)$ at $\hat{\theta}_c$ matches that of the large-sample log-density of $\hat{\theta}_c$; thus this changes the locations of any local maxima other than the global maximum at $\hat{\theta}_c$. Therefore the magnitude adjustment might be more appropriate if the full posterior distribution is multi-modal.

However only the curvature adjustment ensures that the convergence to a χ^2 distribution is met; the magnitude adjustment only gets the first moment correct. This may have a strong impact on the shape of the composite likelihood around $\hat{\theta}_c$, and therefore on the composite posterior density.

2.3.1 Asymptotic posterior distributions

We now derive the asymptotic properties of the composite posterior distributions, both adjusted and unadjusted. Provided that the unadjusted composite posterior is a valid distribution, it can be shown under the usual regularity conditions that when n is large enough (Appendix A),

$$\pi_c(\theta \mid y) \dot{\sim} \mathrm{N}\left\{\theta_0, n^{-1} H(\theta_0)^{-1}\right\}. \tag{12}$$

Here and below we abuse notation; (12) means that θ has the stated distribution, conditional on y, not that the posterior density has a distribution. Unlike in the usual case, the unadjusted composite posterior distribution does not converge to the asymptotic distribution of the composite likelihood estimator, given by (3).

In their investigation of the asymptotic distribution of the the magnitude-adjusted posterior, Pauli et al. [2011, page 8] state that the posterior has approximately the correct variance "by the χ^2 approximation for the null distribution." Further Pauli et al. [2011, pages 8 & 9] state that the approximation is asymptotically correct when $p = 1$, and argue that the approximation represents an improvement over the naive composite posterior when $p > 1$. To expand on this, as the scaling constant estimate $\hat{k} = p/\sum_{i=1}^{p} \hat{\lambda}_i$ used for the magnitude adjustment converges almost surely to $p/\mathrm{tr}\{H(\theta_0)^{-1} J(\theta_0)\}$ as $n \to \infty$, we conclude that (Appendix A),

$$\pi_{\text{magn}}(\theta \mid y) \dot{\sim} \mathrm{N}\left\{\theta_0, (np)^{-1} \mathrm{tr}\{H(\theta_0)^{-1} J(\theta_0)\} H(\theta_0)^{-1}\right\}. \tag{13}$$

Thus unless θ_0 is scalar, i.e., unless $p = 1$, π_{magn} will differ from the asymptotic distribution given by (3). Compared to (12), the asymptotic variance is inflated, because $\mathrm{tr}\{H(\theta_0)^{-1} J(\theta_0)\} \geq p$; see Appendix B.

Since the curvature adjustment obtains the correct curvature, it is straightforward to see that

$$\pi_{\text{curv}}(\theta \mid y) \dot{\sim} \text{N}\left\{\theta_0, n^{-1}H(\theta_0)^{-1}J(\theta_0)H(\theta_0)^{-1}\right\}, \tag{14}$$

which is exactly the asymptotic distribution of the maximum composite likelihood estimator.

2.3.2 Comparison of the Adjusted Likelihood to the Full Likelihood

The magnitude or curvature adjustment will ensure only that the *distribution* of the corresponding adjusted composite likelihood ratio, $\Lambda_{\text{adj}}(\theta_0) = 2\{\ell_{\text{adj}}(\hat{\theta}_c; y) - \ell_{\text{adj}}(\theta_0; y)\}$, will approximate the χ^2_p distribution of the true likelihood ratio, $\Lambda(\theta_0)$. However, since the composite likelihood should contain some of the information in the full likelihood, one would hope that $\Lambda_{\text{adj}}(\theta_0) \approx \Lambda(\theta_0)$, i.e., that the values of these ratios should be related. Figure 2 compares values of $\Lambda_{\text{curv}}(\theta_0)$ and $\Lambda(\theta_0)$ for 200 datasets simulated as described in §1.3. Their correlation is $\hat{r} = 0.64$ when the number of replicate Gaussian processes is $n = 50$, and $\hat{r} = 0.79$ when $n = 500$: reasonable correlations, but not overwhelming.

Our aim in adjusting the likelihood is not to approximate the true likelihood—and in turn, approximate the full posterior—but rather to obtain appropriate inference from a composite posterior. If we did wish to approximate $\ell(\theta_1)$ at $\theta_1 \in \Theta$, then it can be shown using the curvature-adjusted likelihood that $2\{\ell_{\text{curv}}(\hat{\theta}_c) - \ell_{\text{curv}}(\theta_1)\} \xrightarrow{d} X^T X$, where $X \sim N(\{H(\theta_0)J^{-1}(\theta_0)H(\theta_0)\}^{1/2}(\theta_1 - \theta_0), \text{Id}_p)$, whereas $2\{\ell(\hat{\theta}) - \ell(\theta_1)\} \xrightarrow{d} Y^T Y$, where $Y \sim N\{I(\theta_0)^{1/2}(\theta_1 - \theta_0), \text{Id}_p\}$ and $I(\theta_0)$ is the Fisher information matrix based on the full likelihood. Obviously, the approximation will degrade as $(\theta_1 - \theta_0)$ grows. Since the true likelihood and information about $I(\theta_0)$ would not be available in a realistic application, it seems unclear how to improve the approximation to the true likelihood away from θ_0. Simply put, by not having the full likelihood available, we lose information.

3 Markov chain Monte Carlo samplers

This section describes implementations of Markov chain Monte Carlo basing Bayesian inference on composite likelihoods. One must take care to show that MCMC algorithms will converge to the correct target distributions, as composite likelihoods, adjusted or not, are not valid likelihoods. We describe the adjusted Metropolis–Hastings algorithm and the Gibbs sampler in turn.

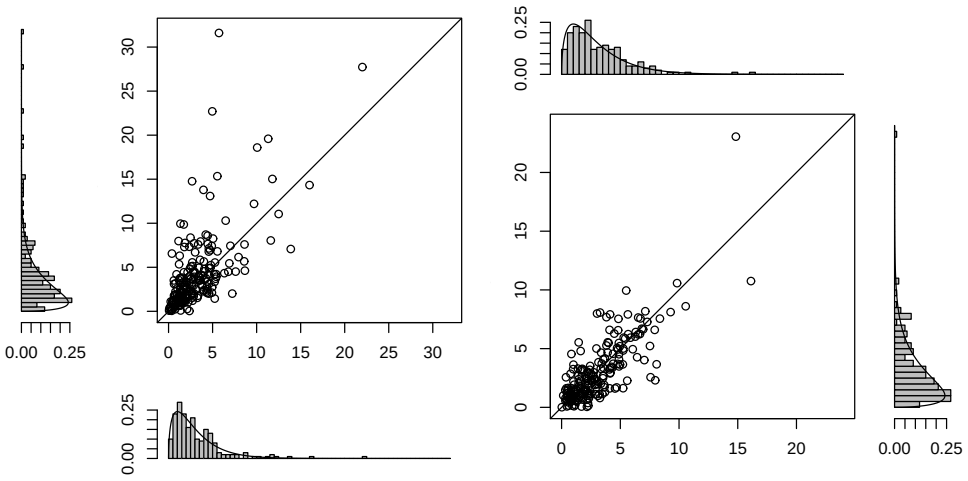


Figure 2: *Comparison of 200 likelihood ratios for the Gaussian process simulation with n replicates: $\Lambda_{\mathrm{curv}}(\theta_0)$ for the curvature-adjusted composite likelihood (y-axis) versus $\Lambda(\theta_0)$ (x-axis). Left: $n = 50$. Right: $n = 200$.*

3.1 Adjusted Metropolis–Hastings algorithm

In Section 2 we suggested two adjustments intended to provide approximations to the full likelihood ratios. We now discuss an adjusted Metropolis–Hastings algorithm, given in Algorithm 1, and verify that it has the desired stationary distribution.

Algorithm 1: Adjusted Metropolis–Hastings algorithm.

Input : $\hat{\theta}_c$, $\hat{H}(\hat{\theta}_c)$, $\hat{J}(\hat{\theta}_c)$, $\theta_1 \in \Theta$, a proposal distribution $q(\cdot \mid \theta)$ and an adjusted composite likelihood $L_{\mathrm{adj}}(\cdot; y)$

Output: A realisation of length $N + 1$ from a Markov chain

for $t \leftarrow 1$ **to** N **do**
  $\theta^{(p)} \sim q(\cdot \mid \theta^{(t)})$;
  $\alpha_{\mathrm{adj}}(\theta^{(t)}, \theta^{(p)}) \leftarrow \min\left\{1, \frac{L_{\mathrm{adj}}(\theta^{(p)};y)\pi(\theta^{(p)})q(\theta^{(t)}|\theta^{(p)})}{L_{\mathrm{adj}}(\theta^{(t)};y)\pi(\theta^{(t)})q(\theta^{(p)}|\theta^{(t)})}\right\}$;
  $U \sim U(0,1)$;
  if $\alpha_{\mathrm{adj}}(\theta^{(t)}, \theta^{(p)}) \leq U$ **then**
    $\theta^{(t+1)} \leftarrow \theta^{(p)}$;
  else
    $\theta^{(t+1)} \leftarrow \theta^{(t)}$;
  end
end
return $\{\theta^{(t)}\}_{t=1,\ldots,N+1}$;

Implementation with one of the adjusted likelihoods, $L_{\mathrm{magn}}(\theta; y)$ or $L_{\mathrm{curv}}(\theta; y)$, requires only a preliminary maximisation of the composite likelihood to estimate the matrices $H(\theta_0)$ and $J(\theta_0)$ for the adjustment. The argument that establishes detailed balance for the original Metropolis–Hastings algorithm [Robert and

Casella, 2005, Theorem 7.2] applies to Algorithm 1, and it can be shown that apart from normalizing constants, the stationary distribution of the Markov chain is given by

$$L_c^{\text{tot}}(\theta; y)^k \pi(\theta), \quad k = p/\sum_{i=1}^{p} \hat{\lambda}_i, \tag{15}$$

for the magnitude adjustment and

$$\exp\{\ell_{\text{curv}}(\theta; y)\}\pi(\theta) \tag{16}$$

for the curvature adjustment. The stationary distributions (15) and (16) should provide better coverage than if an unadjusted composite likelihood was used.

3.2 Gibbs sampling

When the unknown parameter θ has low dimension, Algorithm 1 should provide approximate inference for θ without too much Monte Carlo effort. For models in which θ is of high dimension, however, the probability of acceptance may be too low for Algorithm 1 to be viable, and then the parameter vector is often partitioned and Gibbs sampler employed. Let us write $\theta = (\theta_1^T, \ldots, \theta_G^T)^T$, where $\theta_j \in \mathbb{R}^{p_j}$ and $\sum_{j=1}^{G} p_j = p$, and suppose that we wish to draw from

$$\pi(\theta \mid y) \propto L(\theta; y)\pi(\theta). \tag{17}$$

A typical implementation of a Gibbs sampler will successively draw from

$$\pi(\theta_j \mid \theta_{-j}, y) \propto L(\theta_j \mid \theta_{-j}, y)\pi(\theta_j), \quad j = 1, \ldots, G, \tag{18}$$

where θ_{-j} is the parameter vector θ with the elements of θ_j removed. In this section we propose two Gibbs samplers for use with composite likelihoods.

3.2.1 Overall Gibbs sampler

Since the true likelihood is unobtainable, we use the Gibbs sampler with an adjusted composite likelihood. We could replace $L(\theta; y)$ in (17) with $L_{\text{adj}}(\theta; y)$, where L_{adj} is either the magnitude- or the curvature-adjusted composite likelihood. To perform Gibbs sampling, $\hat{\theta}_c$, $\hat{H}(\hat{\theta}_c)$ and $\hat{J}(\hat{\theta}_c)$ can be estimated once prior to running the algorithm, and $L_{\text{adj}}(\theta; y)$ can be calculated. Gibbs sampling then proceeds as usual.

As the Gibbs sampler is a special case of the Metropolis–Hastings algorithm [Robert and Casella, 2005, sec 10.2.2] and it was shown in Section 3.1 that the latter could accommodate an adjusted composite likelihood,

this overall Gibbs sampler algorithm converges to the stationary distributions given by (15) or (16).

3.2.2 Adaptive Gibbs sampler

In real problems, the dimensions of θ, and hence of $\hat{\theta}_c$, can be quite large. By finding $\hat{\theta}_c$, $\hat{H}(\hat{\theta}_c)$ and $\hat{J}(\hat{\theta}_c)$ only once before implementing the algorithm, the overall Gibbs sampler loses the 'spirit' of Gibbs sampling, which is to sample the lower-dimensional θ_j given the current value of θ_{-j}.

An alternative to adjusting the likelihood in (17) is to replace the likelihood in (18) by an adjusted composite likelihood. That is, the likelihood for θ_j can be adjusted based on the current values of θ_{-j}. Since this adjustment requires knowledge of the maximum composite likelihood estimates, the value of $\hat{\theta}_{j,c} \mid \theta_{-j} = \theta_{-j}^{(t)}$ must be found at each step. This approach has the advantage that the adjusted composite likelihood approximation using the current value of θ_{-j} should be more accurate, as the approximation is made in a lower-dimensional parameter space. In particular if θ_j is scalar, then the magnitude adjustment of the composite likelihood ratio statistic is exact; see (9). This adaptive Gibbs sampler is given in Algorithm 2.

Algorithm 2: Adaptive adjusted Gibbs sampler.

Input : $\theta^{(1)} \in \Theta$
Output: A realisation of length $N+1$ from a Markov chain

for $t \leftarrow 1$ **to** N **do**
  for $j \leftarrow 1$ **to** G **do**
    Get the restricted maximum composite likelihood estimate $\hat{\theta}_{j,c}$ with θ_{-j} held fixed at $\theta_{-j}^{(t)}$;
    Get $\hat{H}_{j,j}(\hat{\theta}_j) = \nabla^2 \ell_c(\hat{\theta}_{j,c} \mid \theta_{-j}^{(t)}, y)$ and $\hat{J}_{j,j}(\hat{\theta}_j)$, the sample covariance matrix of $\nabla \ell_c(\hat{\theta}_{j,c} \mid \theta_{-j}^{(t)}, y_i), i = 1, \ldots n$, and define the adjusted composite log-likelihood $\ell_{\text{adj}}(\theta_j \mid \theta_{-j}^{(t)}, y)$ from either (8) or (10);
    Draw $\theta_j^{(t+1)}$ from $L_{\text{adj}}(\theta_j; y, \theta_{-j}^{(t)})\pi(\theta_j \mid \theta_{-j})$ (using Metropolis–Hastings updates if necessary);
  end
end
return $\{\theta^{(t)}\}_{t=1,\ldots,N+1}$;

It can be shown that Algorithm 2 corresponds to a well-defined posterior by considering the completion [Robert and Casella, 2005, section 10.1.2]:

$$\pi(\hat{\theta}, \theta \mid y) = \prod_{j=1}^{G} \pi(\hat{\theta}_j \mid \theta, y)\pi(\theta \mid y),$$

where $\pi(\theta \mid y)$ represents the target density. Note that

$$\pi(\theta \mid y) = \int \prod_{j=1}^{G} \pi(\hat{\theta}_j \mid \theta, y)\pi(\theta \mid y)d\hat{\theta}$$

as required for a completion, provided that $\pi(\hat{\theta}_j \mid \theta, y)$ is a valid density. Define

$$\pi(\hat{\theta}_j \mid \theta, y) = \delta_{\arg\max L_c(\theta_j|\theta_{-j},y)}(\hat{\theta}_j),$$

that is, a Dirac measure on the value of θ_j that maximizes the composite likelihood given the current values of θ_{-j}. If the maximum composite likelihood estimates $\hat{\theta}_j$ could be found analytically, then Algorithm 2 would simply be a Gibbs sampler on the completion. Since $\hat{\theta}_j$ must be obtained numerically, convergence of the Markov chains must be carefully checked by examining the output.

In the context of the adaptive Gibbs sampler, both the magnitude and curvature adjustments must be understood as adjusting the conditional likelihood $L_c(\theta_j \mid \theta_{-j}, y)$. That is, k in equation (8) now becomes $p_j / \sum_{i=1}^{p_j} \hat{\lambda}_i$ where $\hat{\lambda}_i$ are the eigenvalues of the matrix defined by $\hat{H}(\hat{\theta}_j)$ and $\hat{J}(\hat{\theta}_j)$. Similarly, the matrix C in (11) is defined by $\hat{H}(\hat{\theta}_j)$ and $\hat{J}(\hat{\theta}_j)$.

It is instructive to tie each of the Gibbs samplers to the asymptotic distribution of the posterior. Let $\pi(\theta \mid y)$ denote the composite posterior distribution evaluated at $\theta \in \mathbb{R}^p$ and further assume that the asymptotic posterior distribution corresponds with that of the maximum composite likelihood estimator,

$$\log \pi(\theta \mid y) \dot{\propto} -\frac{1}{2}(\theta - \theta_0)^T H(\theta_0) J^{-1}(\theta_0) H(\theta_0)(\theta - \theta_0), \tag{19}$$

where $\dot{\propto}$ means 'asymptotically proportional to'. Gibbs sampling for a given partition $\theta = (\theta_j, \theta_{-j})^T$, where $\theta \in \mathbb{R}^{p_j}$ and $\theta_{-j} \in \mathbb{R}^{p-p_j}$, would involve drawing from $\pi(\theta_j \mid \theta_{-j}, y)$, the conditional posterior distribution of θ_j given some fixed value for θ_{-j}.

In the overall Gibbs sampler, one begins by approximating (19) with

$$\log \pi_{\text{adj}}(\theta \mid y) \dot{\propto} -\frac{1}{2}(\theta - \hat{\theta}_c)^T H^{\text{adj}}(\hat{\theta}_c)(\theta - \hat{\theta}_c). \tag{20}$$

where $H^{\text{adj}}(\hat{\theta}_c)^{-1}$ is the covariance matrix in (13) or (14) for the magnitude- and curvature-adjusted posteriors respectively.

Let $\hat{\theta}_c = (\hat{\theta}_{c,j}, \hat{\theta}_{c,-j})^T$ and partition

$$H^{\text{adj}}(\hat{\theta}_c) = \begin{bmatrix} H^{\text{adj}}_{j,j}(\hat{\theta}_c) & H^{\text{adj}}_{j,-j}(\hat{\theta}_c) \\ H^{\text{adj}}_{-j,j}(\hat{\theta}_c) & H^{\text{adj}}_{-j,-j}(\hat{\theta}_c) \end{bmatrix}.$$

Since (20) implies that $\pi_{\text{adj}}\{(\theta_j, \theta_{-j})^T\}$ is approximately a Gaussian density with mean $(\hat{\theta}_{c,j}, \hat{\theta}_{c,-j})^T$ and covariance matrix $\Sigma = H^{\text{adj}}(\hat{\theta}_c)^{-1}$, we see that $\pi_{\text{adj}}(\theta_j \mid \theta_{-j}, y)$ is approximately Gaussian, with mean

$$\hat{\theta}_{c,j} + \Sigma_{j,-j}\Sigma^{-1}_{-j,-j}(\theta_{-j} - \hat{\theta}_{c,-j}) = \hat{\theta}_{c,j} - H^{\text{adj}}_{j,j}(\hat{\theta}_c)^{-1}H^{\text{adj}}_{j,-j}(\hat{\theta}_c)(\theta_{-j} - \hat{\theta}_{c,-j}) \tag{21}$$

and covariance matrix

$$\Sigma_{j,j} - \Sigma_{j,-j}\Sigma^{-1}_{-j,-j}\Sigma_{-j,j} = H^{\text{adj}}_{j,j}(\hat{\theta}_c)^{-1}. \tag{22}$$

The adaptive Gibbs sampler makes its approximation later in the algorithm. Starting from (19), let θ_{-j} be given and consider $\log \pi(\theta_j \mid \theta_{-j}, y)$. By partitioning θ_0 and $H(\theta_0)J^{-1}(\theta_0)H(\theta_0)$, it is straightforward to show that the asymptotic conditional posterior is

$$\log \pi(\theta_j \mid \theta_{-j}, y) \dot{\propto} (\theta_j - \mu_{j|-j})^T \Sigma^{-1}_{j|-j}(\theta_j - \mu_{j|-j}), \tag{23}$$

where

$$\mu_{j|-j} = \theta_{0,j} - \left\{H(\theta_0)J^{-1}(\theta_0)H(\theta_0)\right\}^{-1}_{j,j} \left\{H(\theta_0)J^{-1}(\theta_0)H(\theta_0)\right\}_{j,-j} (\theta_{-j} - \theta_{0,-j}), \tag{24}$$

and

$$\Sigma_{j|-j} = \left\{H(\theta_0)J^{-1}(\theta_0)H(\theta_0)\right\}^{-1}_{j,j} \tag{25}$$

analogous to (21) and (22) above. The adaptive Gibbs sampler makes its approximation to the conditional distribution, estimating the conditional mean by finding $\hat{\theta}_{c,j|-j}$, the value which maximizes the, lower-dimensional, conditional composite log-likelihood $\ell_c(\theta_{j,c} \mid \theta^{(t)}_{-j}, y)$, and then adjusting this lower-dimensional likelihood to obtain an estimate for $\left\{H(\theta_0)J^{-1}(\theta_0)H(\theta_0)\right\}_{j,j}$.

The advantage of the overall Gibbs sampler is computational and in its simplicity; the adaptive Gibbs sampler's need to estimate $\hat{\theta}_j$ at every step slows it tremendously. The potential gain from the latter is that the approximation made by employing a composite likelihood is made only for the subvector θ_j and is done with knowledge of the current values of the other parameters. In the next section we explore by simulation whether the adaptive Gibbs sampler improves overall estimation.

4 Simulation study

In this section, we use simulation to assess the performance of the magnitude and the curvature adjustments. Following Monahan and Boos [1992], we assess whether our adjustments yield posteriors that are valid by coverage, i.e., whether $\Pr[\theta \in \mathrm{CI}_\alpha(Y)] = \alpha$, under some probability measure for θ defined on Θ and some credible intervals CI_α with level $0 \leq \alpha \leq 1$.

We first apply the proposed adjustments to the stationary isotropic Gaussian process of Section 1.3 and compare the results obtained using the adjusted composite likelihood to those using both the full likelihood and the naive composite likelihood. We then focus on spatial extremes by considering a Bayesian hierarchical model involving max-stable processes.

4.1 Gaussian processes

We again consider a one-dimensional stationary Gaussian process with mean $\mu \in \mathbb{R}$ and an exponential covariance function $\gamma(h) = \tau \exp(-h/\omega)$, $\tau > 0$, $\omega > 0$. We examine two different forms of dependence, allowing ω to equal 3 and 1.5, which respectively yield effective ranges for the covariance of roughly 9 and 4.5. The priors on μ, τ are those reported in Section 1 while an inverse Gamma density with shape 1/10 and scale 1 is assumed on ω. The stochastic process is replicated $n = 50$ times in each simulation and is observed at $K = 20$ locations uniformly generated in the interval $[0, 20]$. The simulation was repeated 500 times to assess coverage, with $\mu = 0$ and $\tau = 1$ in each case.

Figure 3 compares the posterior densities obtained from the full likelihood, the unadjusted pairwise posterior, and the adjusted composite posterior distributions using the magnitude and the curvature adjustments from a single simulation. There is a large improvement due to the adjustment. Owing to the asymptotic unbiasedness of the maximum composite likelihood estimator, the modes of the marginal composite posterior distributions are close to those obtained from the full likelihood. The use of the adaptive Gibbs sampler for the magnitude adjustment seems to improve the approximation to the full posterior, particularly for the range parameter ω; recall from Section 3.2 that this is not an overall magnitude adjustment. The adaptive sampler used here has three blocks each comprising a single parameter.

Table 1 summarizes the empirical coverages based on 500 replicate data sets. Overall, the adjustments give reasonable credible intervals, whereas the naive composite posterior has poor coverage. The Metropolis–Hastings algorithm and overall Gibbs sampler have the same stationary distribution and give the same coverages for each adjustment. The curvature adjustment performs better overall than the magnitude adjustment,

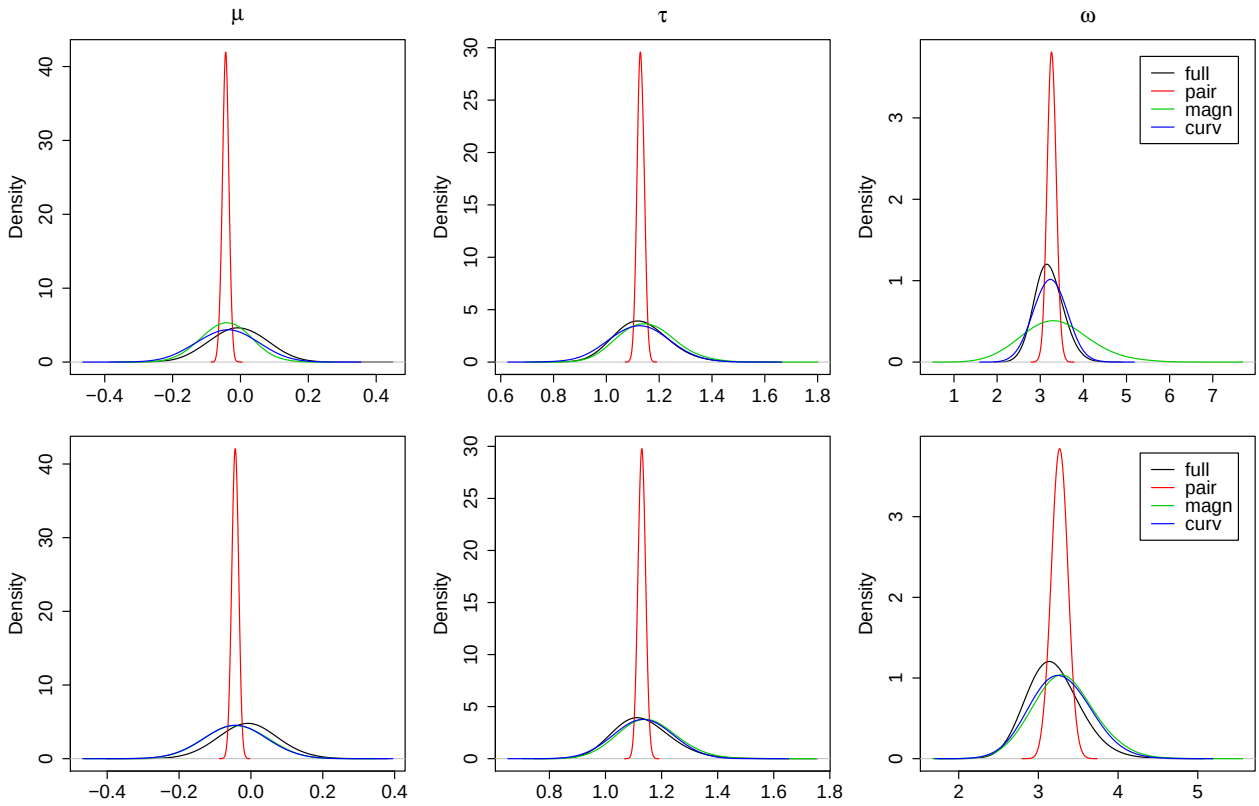


Figure 3: *Comparison between the marginal full posterior (black), the marginal pairwise posterior (red) and the marginal adjusted pairwise posterior densities based on the magnitude (green) and curvature (blue) adjustments. The posterior distributions are derived from $n = 50$ realisations of a Gaussian process having an exponential covariance function with $\mu = 0$, $\tau = 1$ and $\omega = 3$ and observed at $K = 20$ locations. Top row: Metropolis–Hastings algorithm. Bottom row: Adaptive adjusted Gibbs sampler.*

Table 1: *Empirical coverages (%) for nominal 95% credible intervals based on 500 Gaussian process simulations. "Full" denotes coverage with the full posterior, "Magnitude" corresponds to the magnitude adjusted posterior, "Curvature" to the curvature adjusted posterior, and "Unadjusted" to the naive composite posterior.*

	Metropolis–Hastings											
	Full			Magnitude			Curvature			Unadjusted		
	μ	τ	ω	μ	τ	ω	μ	τ	ω	μ	τ	ω
$\omega = 3$	96	94	94	89	92	100	94	93	94	16	21	37
$\omega = 1.5$	94	95	96	85	93	100	94	94	93	19	22	53
	Overall Gibbs sampler											
	Full			Magnitude			Curvature			Unadjusted		
	μ	τ	ω	μ	τ	ω	μ	τ	ω	μ	τ	ω
$\omega = 3$	95	96	95	87	93	100	94	94	90	19	16	41
$\omega = 1.5$	96	96	96	87	94	100	94	94	94	23	21	55
	Adaptive Gibbs sampler											
	Full			Magnitude			Curvature			Unadjusted		
	μ	τ	ω	μ	τ	ω	μ	τ	ω	μ	τ	ω
$\omega = 3$	96	94	95	95	92	93	95	94	93	20	24	39
$\omega = 1.5$	95	96	95	95	95	95	94	97	95	17	24	55

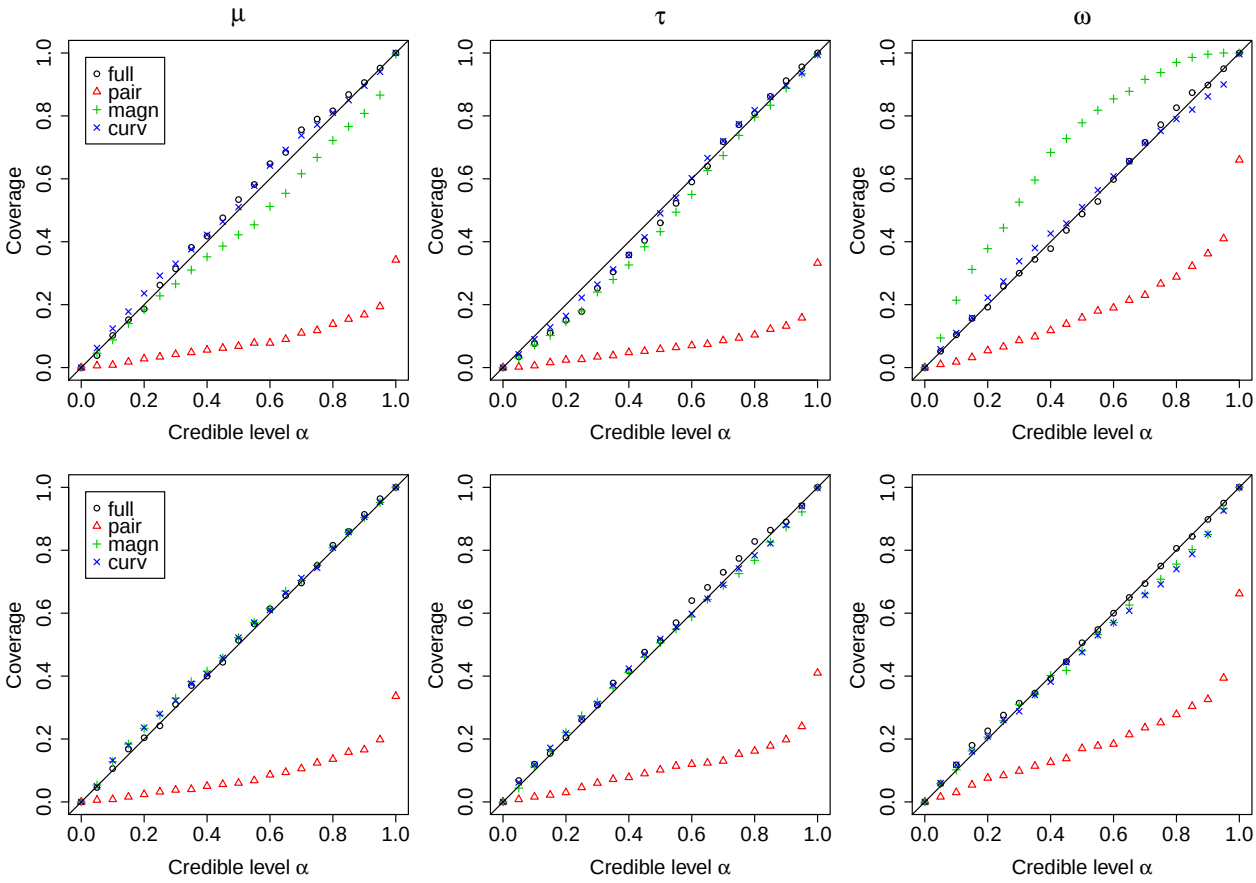


Figure 4: *Variation of the empirical coverages with the credible level α, based on* 500 *replicates of the Gaussian process simulation with $\mu = 0$, $\tau = 1$ and $\omega = 3$, for the full, the non adjusted pairwise and the magnitude/curvature adjusted posteriors. Top row: Overall Gibbs sampler. Bottom row: Adaptive Gibbs sampler.*

particularly for the mean and range parameters μ and ω. The improvement in coverage due to using the adaptive Gibbs sampler appears greater for the magnitude adjustment than for the curvature adjustment, partly because there is more room for improvement, and because the latter was already adjusting each element of θ differently. The curvature and adaptive magnitude adjustments yield the best coverages.

Figure 4, which complements Table 1 by showing how the empirical coverages depend on the credible level for the overall and adaptive Gibbs samplers, corroborates the conclusions drawn from Table 1. Compared to the unadjusted composite posterior, the proposed adjustments clearly improve the coverages and seem to yield essentially the same coverages as the full posterior, though the latter provides shorter intervals, if it is available. The adaptive Gibbs sampler for the magnitude adjustment performs better than its overall counterpart, indicating that the latter might not be flexible enough to provide the correct coverages for each element of the parameter vector. The curvature adjustment again seems to be improved less by the adaptive

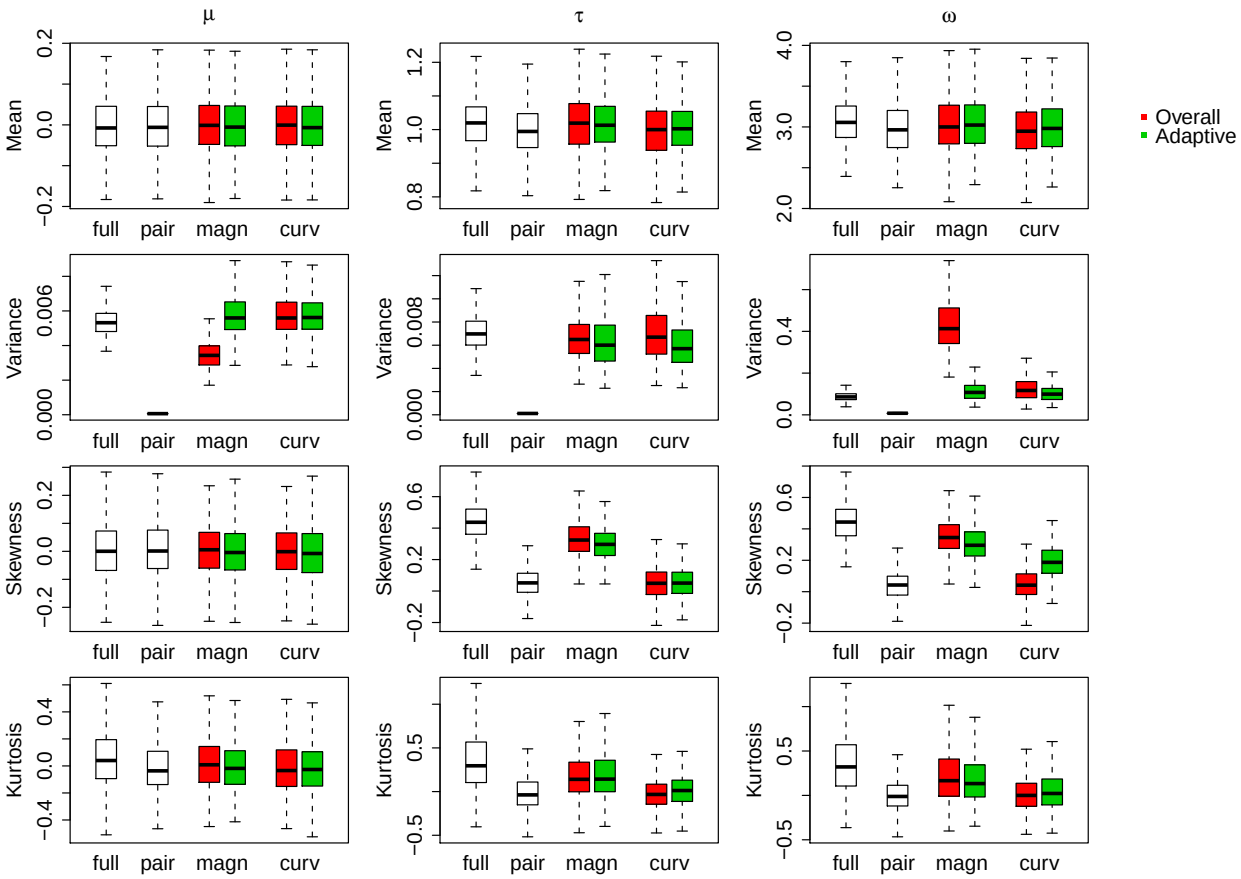


Figure 5: *Boxplots of the sample centered moments of the estimated posterior distribution for each of the* 500 *simulations* ($\mu = 0$, $\tau = 1$, $\omega = 3$) *for the full posterior (full), the unadjusted pairwise posterior (pair), the magnitude adjusted composite posterior (magn) and the curvature adjusted composite posterior (curv). Red boxplots: Overall Gibbs sampler. Green boxplots: Adaptive Gibbs sampler.*

version of the Gibbs sampler.

Figure 4 shows that the proposed adjustments have good coverage properties, but it is also interesting to check to what extent the composite posterior distributions share common features with the full posterior. Figure 5 shows boxplots of the first four centered moments of the estimated posterior distributions. As one would expect from the fact that the composite likelihoods give unbiased estimating equations, the first moments of the composite posterior distributions, including the unadjusted one, match those of the full posterior. The variance of the unadjusted pairwise posterior distribution is much too small, but those of the adjusted posterior distributions are closer to that of the full posterior. The magnitude adjustment combined with the overall Gibbs sampler has a smaller variance for the mean μ and a larger one for the range ω; this clarifies why Table 1 shows that this particular adjustment tends to undercover μ and overcover ω. Except

for μ, none of the adjustments gives the correct skewness and kurtosis, though the magnitude adjustment is slightly better. Nevertheless, both adjustments can capture the first two moments well, and despite the degradation of the approximation with distance from θ_0, yield coverage rates which are very comparable to those obtained using the full likelihood.

Finally, we investigate the difference between the magnitude- and curvature-adjusted posteriors and the effect of the dimension of the blocks used in the adaptive Gibbs sampler. As noted in Section 2.1, the magnitude adjustment will recover the χ^2 null distribution only if the dimension of θ_j is one. In addition to running the adaptive Gibbs sampler with μ, τ, and ω each serving as its own block, we also ran a two-block version of the adaptive Gibbs sampler with $\theta_1 = \mu$ and $\theta_2 = (\tau, \omega)^T$. For the individual block version of the adaptive Gibbs sampler, there is virtually no difference in the estimates of the magnitude- and curvature-adjusted posteriors, reflecting that both adjustments adequately capture the information contained in the composite likelihood. However, for the two-block version of the sampler, the empirical posterior correlations of τ and ω differ: the curvature-adjusted posterior gives $\text{Cov}(\tau, \omega) \approx 0.69$, whereas the magnitude-adjusted posterior gives $\text{Cov}(\tau, \omega) \approx 0.33$. It is difficult to estimate both the sill and range parameters of a Gaussian process [**?**], whose ratio τ/ω is important for applications such as interpolation. The 95% credible intervals for this ratio have an empirical coverage rate of 96% for the curvature-adjusted posterior, but a coverage rate of 100% for the magnitude-adjusted posterior. This suggests that for the two-block Gibbs sampler, the magnitude adjustment fails to fully capture the relationship between these two parameters, thus giving further evidence that the curvature adjustment is to be preferred, since it seems to provide output that can be used more flexibly.[1]

4.2 Bayesian hierarchical model for spatial extremes

Let $Y_m(x)$, $x \in \mathcal{D}$, $m \geq 1$ be independent replications of a stochastic process. Asymptotic theory for extremes implies that, provided the limit exists and is non-degenerate, the process

$$\max_{m=1,\ldots n} a_n(x)^{-1}\{Y_m(x) - b_n(x)\}$$

converges weakly to a max-stable process $Z(x)$ as $n \to +\infty$ [de Haan, 1984]. Given observations that arise as block (e.g., annual) maxima, it is therefore natural to approximate their joint distribution using such a process. The univariate marginal distributions for such a process will be generalised extreme-value (GEV)

[1] Dan: please see tex code which includes the reference for Zhang.

distributions, which depend on three parameters.

Although the general methodology we propose could be applied with any max-stable model [Smith, 1990; Schlather, 2002; Kabluchko et al., 2009], we focus here on the Gaussian extreme value process of Smith [1990],

$$Z(x) = \max_{k \geq 1} \zeta_k \varphi(x - s_k) \tag{26}$$

where $\{(\zeta_k, s_k)\}_{k \geq 1}$ are the points of a Poisson process on $(0, \infty) \times \mathcal{D}$, with $\mathcal{D} \subset \mathbb{R}^d$, having intensity $\mathrm{d}\Lambda(\zeta, s) = \zeta^{-2}\mathrm{d}\zeta\mathrm{d}s$, and φ is the zero mean d-variate normal density with covariance matrix Σ. As formulated, $Z(x)$ has unit Fréchet margins and its bivariate and trivariate marginal distributions can be used to construct a composite likelihood [Padoan et al., 2010; Genton et al., 2011].

A simple approach to fitting max-stable models is to employ a pairwise likelihood [Padoan et al., 2010; Gholamrezaee, 2010]. To account for non-stationarity in the marginal distributions, it is convenient to assume that the GEV parameters follow response surfaces that depend on location and on covariates such as altitude. Often, however, the available covariates do not fully explain the variation of the marginal distribution over the study region. One approach to capturing the regional effects is to construct a hierarchical model in which the marginal parameters of the extreme value distribution follow a stochastic process, such as a Gaussian process, over the study region.

Our approach is to use a max-stable process model within a hierarchical framework; the max-stable model provides a theoretically justified model for the local dependence, i.e., the spatial dependence of the extremes, and the hierarchy allows for flexibility in modeling how the regional effects influence the marginal behavior. The difficulty is that the full likelihood is unavailable, and so fully Bayesian inference cannot be performed. Instead we employ one of the adjusted MCMC samplers suggested in Section 3.

Our chosen model has the data-process-prior framework of most hierarchical models:

$$\begin{aligned}
Z \mid \boldsymbol{\mu}, \boldsymbol{\sigma}, \boldsymbol{\xi}, \Sigma &\sim \text{Smith's max-stable model,} \\
\boldsymbol{\mu} \mid \boldsymbol{\beta}_\mu, \tau_\mu, \omega_\mu &\sim \mathrm{GP}\left(\mathbf{X}_\mu \boldsymbol{\beta}_\mu, \gamma_\mu\right), \\
\log \boldsymbol{\sigma} \mid \boldsymbol{\beta}_\sigma, \tau_\sigma, \omega_\sigma &\sim \mathrm{GP}\left(\mathbf{X}_\sigma \boldsymbol{\beta}_\sigma, \gamma_\sigma\right) \\
\boldsymbol{\xi} \mid \boldsymbol{\beta}_\xi, \tau_\xi, \omega_\xi &\sim \mathrm{GP}\left(\mathbf{X}_\xi \boldsymbol{\beta}_\xi, \gamma_\xi\right),
\end{aligned}$$

where $\boldsymbol{\mu}$, $\boldsymbol{\sigma}$, $\boldsymbol{\xi}$ represent the three GEV parameters, $\mathrm{GP}(m, \gamma)$ denotes a Gaussian process with mean m and covariance function γ, the $\gamma_\cdot$'s represent exponential covariance functions with corresponding sill and range

parameters $\tau_\cdot$ and $\omega_\cdot$, and the $\boldsymbol{\beta}_\cdot$ are regression coefficients associated to the design matrices $\mathbf{X}_\cdot$.

The prior level places independent priors on all parameters introduced at the process level. We take conjugate normal priors for all regression parameters $\beta_\cdot$, conjugate inverse gamma priors for the $\tau_\cdot$, gamma priors for the range parameters $\omega_\cdot$ and a Wishart prior for the covariance matrix Σ appearing in the Smith model. In all cases, the prior variance is set to be large so that the prior densities, though proper, are relatively flat.

We performed a simulation study to evaluate our approach. Gaussian processes were simulated for $\mu(x)$, $\sigma(x)$, and $\xi(x)$, with $\mu(x)$ and $\sigma(x)$ dependent, and with values similar to those found for annual maximum rainfall data. Then, 50 max-stable processes with marginals given by $\mu(x)$, $\sigma(x)$, and $\xi(x)$ were simulated according to the Smith model. Fifty locations were chosen and the 50 observations at each location were used to fit four models:

M1 the hierarchical model with a conditional independence assumption in the data layer, yielding a product of K independent GEV densities, analogous to Cooley et al. [2007] or Sang and Gelfand [2009];

M2 the max-stable process hierarchical model with no adjustment;

M3 the max-stable process hierarchical model with an adaptive curvature-adjusted Gibbs sampler; and

M4 the max-stable process model where the marginals are described by a response surface in the covariates x, as proposed by Padoan et al. [2010].

The left panels of Figure 6 show boxplots of the differences between the true GEV parameters and all the states of the Markov chains for four different stations, with the asymptotic 95% confidence intervals for the max-stable response surface model. The right panels of Figure 6 display the coverage rates, for all 50 stations, of the 95% posterior credible intervals for the three hierarchical models, with the 95% confidence intervals for the max-stable response surface model. As expected, the unadjusted max-stable hierarchical model produces a posterior that is too concentrated and yields very poor coverages, and the max-stable trend surface model is not flexible enough to account for the complicated regional behavior of the GEV parameters, as evidenced by the poor point estimates in the box plots and the corresponding poor coverage rates. The adjusted max-stable hierarchical model and the conditionally independent hierarchical model produce very similar posterior distributions and have similar coverage rates, although the max-stable model does slightly less well.

The advantage of the max-stable hierarchical model over the conditional independence model is that the former can account for local dependence; even with only 50 locations in the region, it seems to be able to

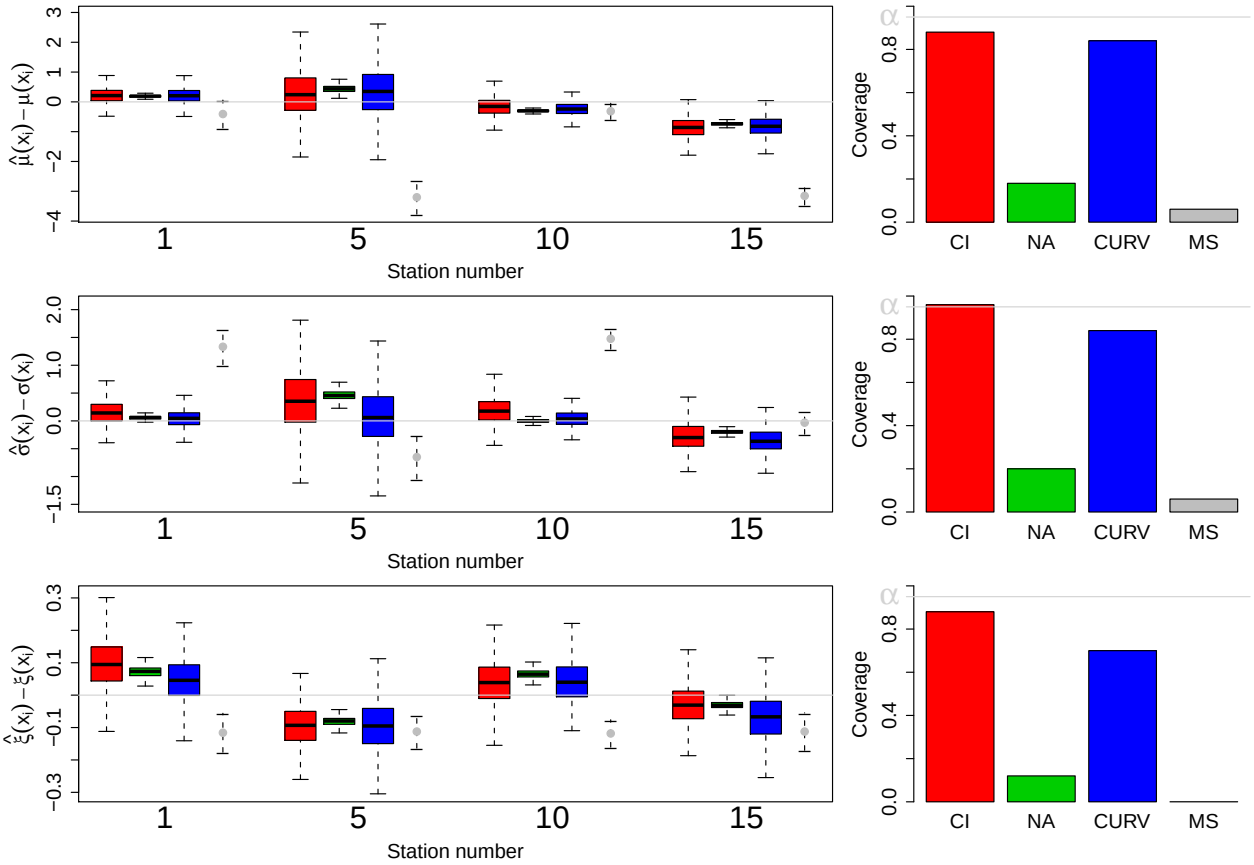


Figure 6: *Boxplots of the difference between the true GEV parameters and all the states of the Markov chains for four stations (left panel). For each of the stations the boxplots are (from left to right) the conditional independence model (M1, red), the non-adjusted hierarchical model (M2, green), the hierarchical model with the curvature adjustment within an adaptive Gibbs sampler (M3, blue), and the asymptotic 95% confidence limit from the max-stable response surface model (M4, grey). The right panel shows the proportion of the credible intervals at level $\alpha = 95\%$ containing the true GEV parameters.*

detect the true pattern of local dependence. The 95% credible intervals for the elements σ_{11}, σ_{12} and σ_{22} of Σ are $(5.39, 8.76)$, $(-1.28, 0.67)$ and $(5.58, 8.37)$, which include the true values 6, 0, and 6. The fitted max-stable model provides a mechanism for producing realistic draws from the spatial process. As Figure 7 shows, a draw from the posterior distribution of the conditional independence model would be inappropriate and unrealistic for spatial phenomena such as rainfall or temperature, annual maxima of which would produce smoother surfaces.

These results are obtained from a (near) perfect model simulation; that is, the max-stable hierarchical model fitted to the data was nearly identical to that from which the data were simulated. Nevertheless, this simulation exercise shows that the adjusted max-stable hierarchical model can flexibly model marginal behavior that captures regional spatial effects and can capture local dependence through the max-stable

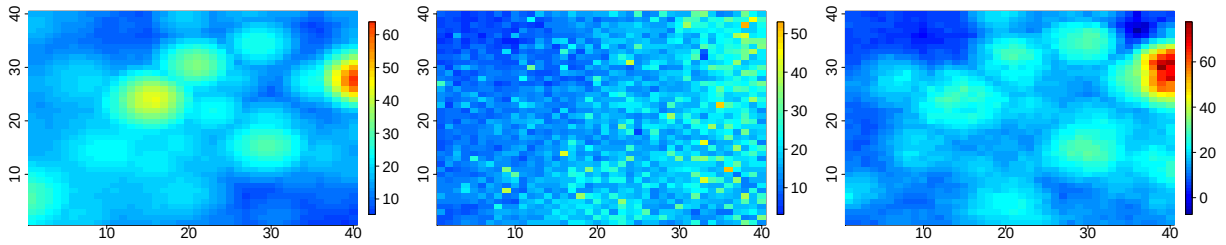


Figure 7: *Comparison between one realization of the observed field and one realization of the different models analyzed. From left to right: observed field; conditional independence model; and max-stable hierarchical model with adjustment. The same seed was used for each simulation.*

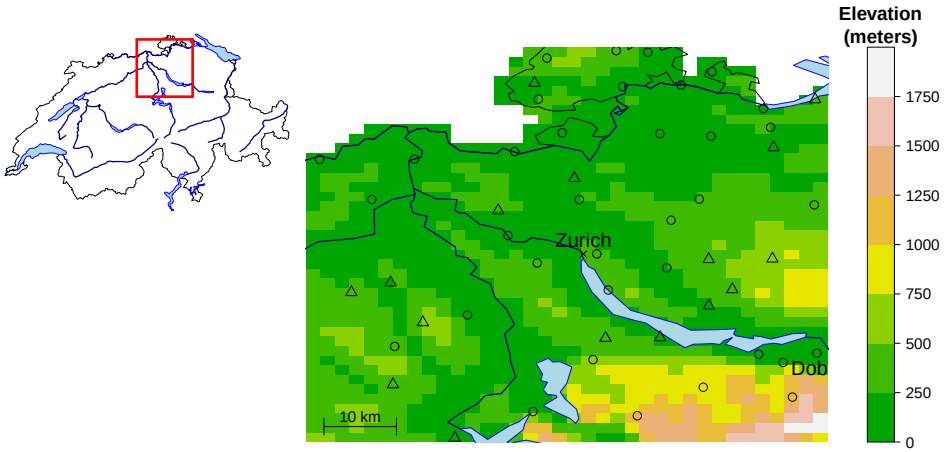


Figure 8: *Map of the study region. The stations used for inference/validation are depicted by circles/triangles.*

process model. Despite the approximation due to employing a composite likelihood, the inference obtained appropriately captures the uncertainty associated with the estimation. In the next section we show that it also seems to perform well on real data.

5 Application

We analyze data on maximum daily rainfall amounts for the years 1962–2008 at 51 sites in the Plateau region of Switzerland; see Figure 8. The area under study is relatively flat, the altitudes of the sites varying from 322 to 910 meters above mean sea level. Data from 16 of the stations were were kept aside for model validation and not used for fitting.

Figure 9 compares the annual maxima over the 16 validation stations, which we term the "groupwise maxima", and the simulated groupwise maxima from the different models. All the max-stable based models seem able to model the distributions of the groupwise maxima, though the simple max-stable model

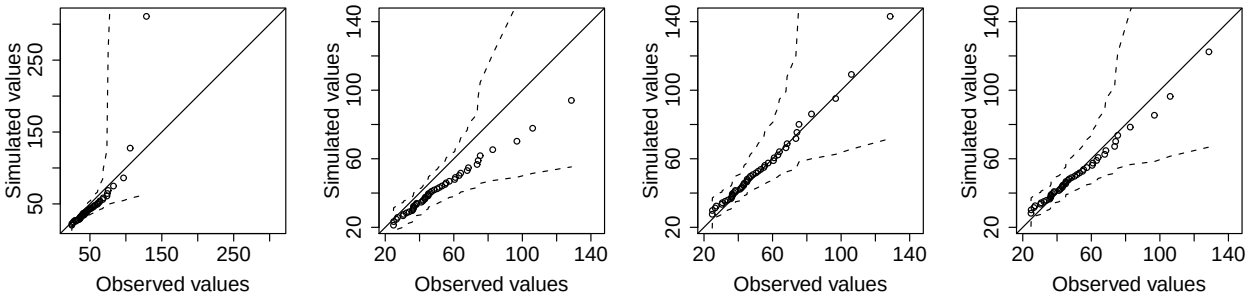


Figure 9: *QQ-plots to compare the observed maxima of the annual maxima from the validation locations and those simulated from various models. From left to right: simple max-stable, conditional independence, unadjusted Bayesian hierarchical, adjusted Bayesian hierarchical models. The* 95% *confidence/credible envelopes are shown as dashed lines.*

badly overestimates the largest value, perhaps due to inaccurate trend surfaces for the GEV parameters, particularly the shape parameter. The conditional independence model shows systematic underestimation, confirming that this model is inappropriate. The unadjusted and adjusted Bayesian hierarchical models yield similar credible envelopes, which seem principally to reflect the variability of simulated conditional Gaussian processes and GEV realizations.

Figure 10 shows three simulated random fields for each model, taken from a large number of such fields. To rank these we took a disk V_{Zurich} of radius 6 km and centered near the Zurich gauging station, and ordered the random fields according to their suprema $S_{\text{Zurich}} = \sup_{x \in V_{\text{Zurich}}} Y(x)$. This allows us to summarize the intensity of a particular realization of a random field. The three rows of Figure 10 correspond to the situation where $\Pr[S_{\text{Zurich}} \leq z_{\text{crit}}] = \alpha$, where $\alpha = 0.05, 0.50, 0.95$ respectively and the level z_{crit} depends on the model considered. Roughly speaking, the three rows show patterns for which S_{Zurich} is expected to be exceeded once every 1.05, 2 and 20 years.

The conditional independence model leads to unrealistic realizations of extreme rainfall fields, but because of the deterministic trend surfaces for the marginal parameters, the simple max-stable model produces fields that are too smooth to be realistic. The unadjusted and adjusted hierarchical models seem to produce the most plausible realizations.

Figure 11 plots return level curves, i.e., graphs of the estimated pth quantile of S_{Zurich} and a similar quantity S_{DOB} for the DOB gauging station, against $1/(1-p)$, and smaller disks of radius 0.3. For the smallest neighborhood, the return level curves are compared to the observations available at the Zurich and DOB gauging stations; see Figure 8.

As the neighbourhoods of radius 0.3km are very small, the return level curves should be close to the empirical curves computed from the data available at the Zurich and DOB gauging stations. This is indeed

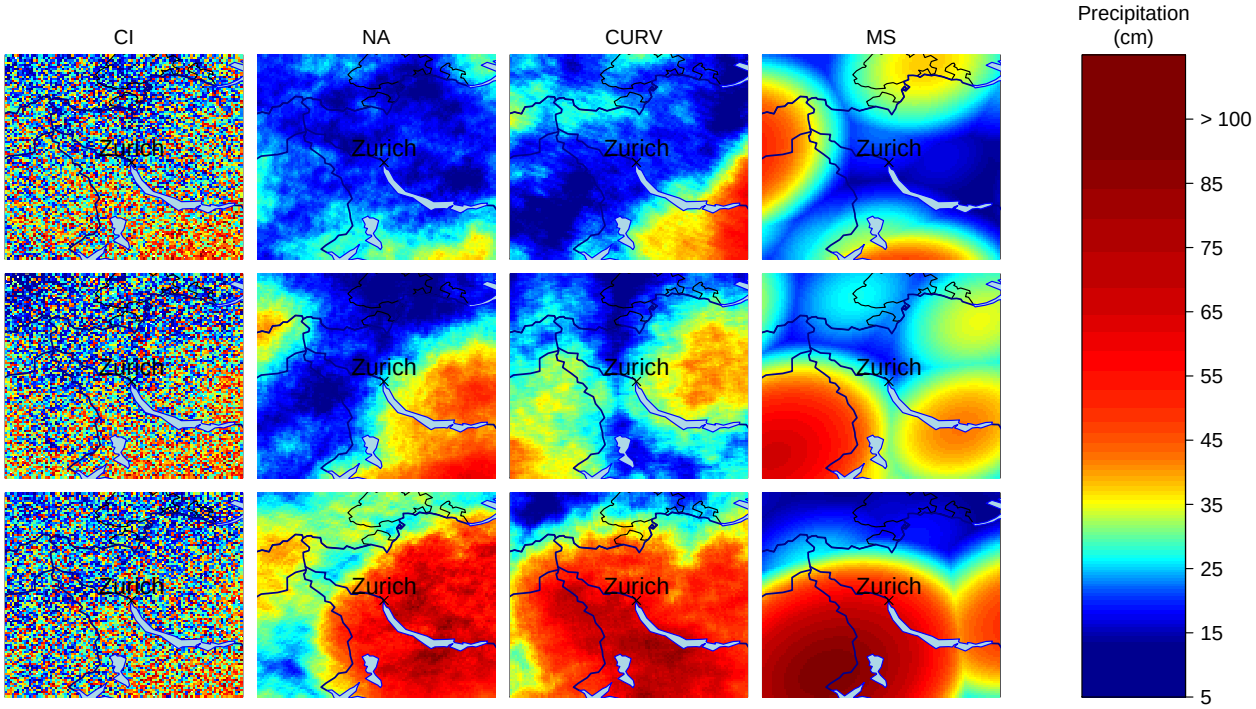


Figure 10: *Three realizations of random fields over the study region for the conditional independent model (M1), hierarchical models without any adjustment (M2) and with the curvature adjustment (M3) and a simple max-stable model with deterministic trend surfaces (M4). The three rows show realizations corresponding to different risk scenarios according to the values of* S_{Zurich} *expected to be exceeded once every* 1.05, 2 *and* 20 *years (from top to bottom).*

the case for Zurich, where all the models apparently reproduce the distribution of extreme rainfall quite well. The results are less convincing for the DOB gauging station where, apart from the adjusted hierarchical model, all the models seem to overestimate the largest extremes. This situation is similar to that seen in Section 4.2: the unadjusted hierarchical model produces a posterior that is too concentrated, while the max-stable trend surface model might not be flexible enough. Both models fail to capture the complicated spatial behavior of the GEV parameters.

For the neighbourhoods of radius 6km, the central panel of the figure shows a very strong discrepancy between the models, because of their different spatial assumptions. The conditional independence model yields unrealistically high return levels, of around 2m for 10-year values, for example. All the max-stable based give approximately the same return levels for return periods shorter than 10 years. For larger return periods, the unadjusted hierarchical model gives the largest estimates. The same plots for 20 other gauging stations depicted the same patterns, suggesting that the unadjusted hierarchical model systematically overestimates the distribution of the supremum in a given neighborhood.

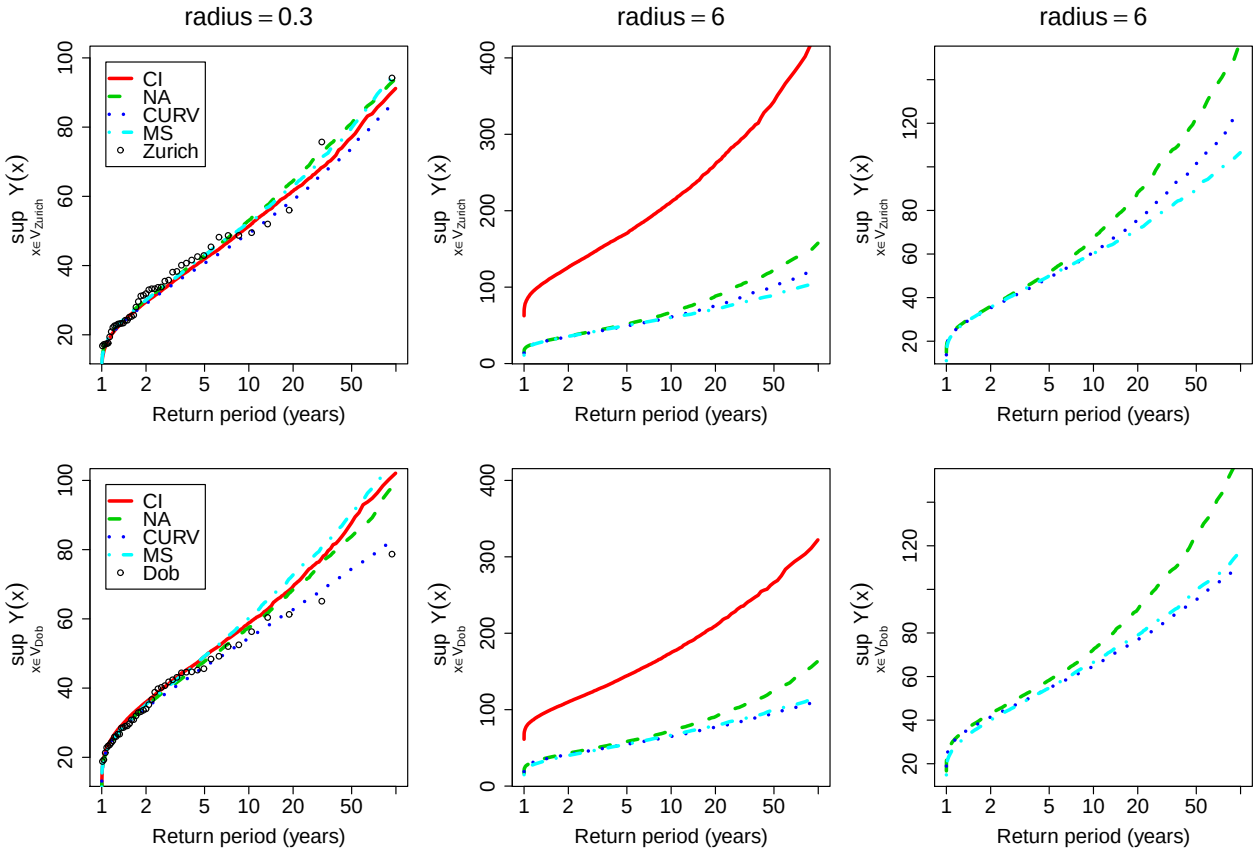


Figure 11: *Comparison between the return level curves (cm) computed on neighborhoods centered at the Zurich (top) and DOB gauging stations (bottom) and having radius* 0.3 *and* 6 *km (left and middle panels) for the conditional independent model (M1), the hierarchical models without any adjustment (M2) and with the curvature adjustment (M3) and a simple max-stable model with deterministic trend surfaces (M4). The left panels compares the return level curves to the observations available at the gauging stations. The right panel is the same as the middle one but shows only the max-stable based models.*

6 Conclusion

In this paper, motivated by a real problem in which Bayesian inference seems natural but a full likelihood is unavailable, we investigate the usefulness of composite likelihood within a Bayesian framework. The posterior distribution obtained from a naive implementation of a composite likelihood can have very poor coverage properties, owing to its inappropriate re-use of the data.

To bypass this hurdle, we propose two modifications of the composite likelihood to recover the usual asymptotic distribution of the likelihood ratio statistic at the true value of the parameters θ_0. We show how these adjustments can be implemented in Markov chain Monte Carlo algorithms and propose two ways of integrating them into the Gibbs sampler. Although the approximation degrades with distance from the parameter underlying the data, simulation studies show that the proposed framework has coverage properties

similar to those obtained using the full posterior.

The work was motivated by a need to flexibly model the marginal distributions when modeling spatial extreme phenomena. We construct a Bayesian hierarchical model whose data layer is driven by a max-stable process while the marginal parameters are modeled as realizations of a stochastic process. A spatial extreme simulation study showed that this framework is able to capture complex marginal behavior as well as the spatial dependence in the data. An application to extreme rainfall around Zurich shows that the approach can capture both local dependence due to individual storms and regional dependence due to similar climatologies, thus broadening the scope of max-stable modelling beyond its current limits.

Acknowledgments

The work of M. Ribatet and A. C. Davison was supported by the CCES Extremes project, `http://www.cces.ethz.ch/projects/hazri/EXTREMES`. D. Cooley's work is partly supported by National Science Foundation grant DMS-0905315.

A Asymptotic distributions of the posterior distributions

The derivation of the asymptotic normality of the posterior distribution heavily relies on Taylor expansions. Let $\hat{\theta}_c$ denote the maximum composite likelihood estimate, let θ_{prior} denote the mode of the prior distribution $\pi(\theta)$, and let

$$h_c^{\text{tot}}(\hat{\theta}_c) = -\nabla_\theta^2 \ell_c^{\text{tot}}(y; \hat{\theta}_c), \qquad h_{\text{prior}}(\theta_{\text{prior}}) = -\nabla_\theta^2 \log \pi(\theta_{\text{prior}}).$$

For n large enough we have

$$\begin{aligned}\pi_c(\theta \mid y) &\dot{\propto} \exp\left\{ \ell_c^{\text{tot}}(y; \hat{\theta}_c) - \frac{1}{2}(\theta - \hat{\theta}_c)^T h_c^{\text{tot}}(\hat{\theta}_c)(\theta - \hat{\theta}_c) + \log \pi(\theta_{\text{prior}}) - \frac{1}{2}(\theta - \theta_{\text{prior}})^T h_{\text{prior}}(\theta_{\text{prior}})(\theta - \theta_{\text{prior}}) \right\} \\ &\dot{\sim} N\left\{ \tilde{\theta}, \tilde{h}(\hat{\theta}_c, \theta_{\text{prior}})^{-1} \right\},\end{aligned}$$

where $\tilde{h}(\hat{\theta}_c, \theta_{\text{prior}}) = h_c^{\text{tot}}(\hat{\theta}_c) + h_{\text{prior}}(\theta_{\text{prior}})$ and $\tilde{\theta} = \tilde{h}(\hat{\theta}_c, \theta_{\text{prior}})^{-1}\{h_c^{\text{tot}}(\hat{\theta}_c)\hat{\theta}_c + h_{\text{prior}}(\theta_{\text{prior}})\theta_{\text{prior}}\}$.

Provided the contribution of the prior distribution $\pi(\theta)$ vanishes as $n \to \infty$, the strong law of large

numbers implies that

$$n^{-1}\tilde{h}(\hat{\theta}_c, \theta_{\text{prior}}) = \left\{ \frac{h_c^{\text{tot}}(\hat{\theta}_c)}{n} + \frac{h_{\text{prior}}(\theta_{\text{prior}})}{n} \right\} \longrightarrow -\mathbb{E}[\nabla^2 \ell_c(\theta_0; Y)] = H(\theta_0),$$

$$\tilde{\theta} = \left\{ \frac{\tilde{h}(\hat{\theta}_c, \theta_{\text{prior}})}{n} \right\}^{-1} \left\{ \frac{h_c^{\text{tot}}(\hat{\theta}_c)}{n} \hat{\theta}_c + \frac{h_{\text{prior}}(\theta_{\text{prior}})}{n} \theta_{\text{prior}} \right\} \longrightarrow \theta_0,$$

almost surely, and thus $\pi_c(\theta \mid y) \dot{\sim} N\left\{\theta_0, n^{-1} H(\theta_0)^{-1}\right\}$.

The derivation of the asymptotic distribution for the magnitude adjustment uses the same argument, with a slight modification. As $n \to \infty$,

$$\hat{k} \longrightarrow p/\text{tr}\left\{H(\theta_0)^{-1} J(\theta_0)\right\}$$

almost surely. Since $\hat{k}$ is estimated prior to running the MCMC algorithm, we can assume that $\hat{k}$ is a (tuning) constant that does not depend on θ. Therefore the analogue of $h_c(\hat{\theta}_c)$ when using ℓ_{magn} in place of ℓ_c^{tot} is

$$h_{\text{magn}}(\hat{\theta}_c) = -\hat{k} \nabla_\theta^2 \ell_{\text{magn}}(y; \hat{\theta}_c) \longrightarrow \text{tr}\{H(\theta_0)^{-1} J(\theta_0)\} H(\theta_0), \qquad n \to \infty,$$

almost surely, from which we conclude that $\pi_{\text{magn}}(\theta \mid y) \dot{\sim} N\left\{\theta_0, (np)^{-1} \text{tr}\{H(\theta_0)^{-1} J(\theta_0)\} H(\theta_0)^{-1}\right\}$.

We conclude with the derivation of the asymptotic distribution of the curvature adjusted composite likelihood. By construction we have

$$n^{-1} h_{\text{curv}}(\hat{\theta}_c) = -n^{-1} \nabla_\theta^2 \ell_{\text{curv}}(y; \hat{\theta}_c) \longrightarrow H(\theta_0) J(\theta_0)^{-1} H(\theta_0), \qquad n \to \infty,$$

almost surely from which we get that $\pi_{\text{curv}}(\theta \mid y) \dot{\sim} N\left\{\theta_0, n^{-1} H(\theta_0)^{-1} J(\theta_0) H(\theta_0)^{-1}\right\}$.

B Asymptotic variance inflation

In this appendix we argue that in many cases in which the densities appearing in the composite likelihood are correct, so that they satisfy the first two Bartlett identities, $\mathbb{E}[\nabla \log f(Y \in \mathcal{A}_i; \theta_0)] = 0$ and $\mathbb{E}[\nabla^2 \log f(Y \in \mathcal{A}_i; \theta_0)] + \text{Var}[\nabla \log f(Y \in \mathcal{A}_i; \theta_0)] = 0$ for all $i \in I$, then $\text{tr}\{H(\theta_0)^{-1} J(\theta_0)\} \geq p = \dim(\theta_0)$. This agrees with our empirical experience, which is that in many cases $\text{tr}\{H(\theta_0)^{-1} J(\theta_0)\} \gg p$.

We first note that

$$\mathrm{tr}\{H(\theta_0)^{-1}J(\theta_0)\} - p = \mathrm{tr}\{H(\theta_0)^{-1}J(\theta_0) - \mathrm{Id}_p\} = \mathrm{tr}[H(\theta_0)^{-1}\{J(\theta_0) - H(\theta_0)\}] \geq 0.$$

Since $H(\theta_0)^{-1}$ is positive semi-definite, the result follows if $J(\theta_0) - H(\theta_0)$ is positive semi-definite, because $\mathrm{tr}\{AB\} \geq 0$ when both A and B are positive semi-definite.

On the one hand we have

$$H(\theta_0) = -\mathbb{E}\left[\nabla^2 \sum_{i\in I} \log f(Y \in \mathcal{A}_i; \theta_0)\right] = -\sum_{i\in I} \mathbb{E}\left[\nabla^2 \log f(Y \in \mathcal{A}_i; \theta_0)\right] = \sum_{i\in I} \mathrm{Var}\left[\nabla \log f(Y \in \mathcal{A}_i; \theta_0)\right],$$

because the variance of the score equals the Fisher information for each individual summand. On the other hand we have

$$\begin{aligned} J(\theta_0) &= \mathrm{Var}\left[\sum_{i\in I} \nabla \log f(Y \in \mathcal{A}_i; \theta_0)\right] \\ &= \sum_{i\in I} \mathrm{Var}\left[\nabla \log f(Y \in \mathcal{A}_i; \theta_0)\right] + \sum_{i,j\in I, i\neq j} \mathbb{E}\left[\nabla \log f(Y \in \mathcal{A}_i; \theta_0) \nabla \log f(Y \in \mathcal{A}_j; \theta_0)^T\right]. \end{aligned}$$

Thus

$$J(\theta_0) - H(\theta_0) = \sum_{i,j\in I, i\neq j} \mathbb{E}\left[\nabla \log f(Y \in \mathcal{A}_i; \theta_0) \nabla \log f(Y \in \mathcal{A}_j; \theta_0)^T\right] = \sum_{i,j\in I, i<j} \mathbb{E}(U_i U_j^T + U_j U_i^T),$$

say; clearly these expectations are symmetric. To see that they will often be positive definite, let A_i and A_j correspond to the events $Y \in \mathcal{A}_i$ and $Y \in \mathcal{A}_j$. If these events are independent, then $\mathbb{E}(U_i U_j^T) = 0$, but if not, suppose that that we may write let $A_i = A_i' \cap A_{ij}$, $A_j = A_j' \cap A_{ij}$, for some event A_{ij} such that A_i' and A_j' are independent conditional on A_{ij}. This arises if, for example, in a Markov chain $Y \in \mathcal{A}_i$ corresponds to $\{Y_1 = y_1, Y_2 = y_2\}$, $Y \in \mathcal{A}_j$ corresponds to $\{Y_2 = y_2, Y_3 = y_3\}$, and we take $A_i' \equiv \{Y_1 = y_1\}$, $A_{ij} \equiv \{Y_2 = y_2\}$ and $A_j' \equiv \{Y_3 = y_3\}$. If we write $\mathrm{pr}(A_i) = \mathrm{pr}(A_i' \mid A_{ij})\mathrm{pr}(A_{ij})$, then the corresponding log likelihood derivative may be written as $U_i = U_i' + U_{ij}$ in a natural notation, and

$$\mathbb{E}(U_i U_j^T) = \mathbb{E}\{(U_i' + U_{ij})(U_j' + U_{ij})^T\} = \mathbb{E}(U_i' U_j'^T) + \mathrm{Var}(U_{ij}) = \mathbb{E}\{\mathrm{Cov}(U_i', U_j' \mid A_{ij})\} + \mathrm{Var}(U_{ij}),$$

because the cross terms $\mathbb{E}(U_i' U_{ij}) = \mathbb{E}(U_j' U_{ij}) = 0$, as may be seen by conditioning on A_{ij}. If U_i' and U_j' are independent conditional on A_{ij}, then $\mathbb{E}(U_i U_j^T) = \mathrm{Var}(U_{ij})$ is positive semi-definite; this would be

the case in the Markov chain example mentioned above. If they are not independent, but are sufficiently weakly correlated conditional on A_{ij} that the term $\mathrm{Var}(U_{ij})$ is dominant, then $\mathbb{E}(U_i U_j^T)$ will also be positive semi-definite, and hence so will be $J(\theta_0) - H(\theta_0)$. This will be the case in typical applications of composite likelihood, as terms that correspond to dependent events A_i, A_j will tend to be positively correlated, because they are proximate in space or time, or both.

References

Chandler, R. E. and Bate, S. (2007). Inference for clustered data using the independence loglikelihood. *Biometrika*, 94(1):167–183.

Chang, I. H. and Mukerjee, R. (2006). Probability matching property of adjusted likelihoods. *Statistics & Probability Letters*, 76(8):838–842.

Cooley, D., Nychka, D., and Naveau, P. (2007). Bayesian spatial modeling of extreme precipitation return levels. *J. Am. Stat. Assoc.*, 102(479):824–840.

Copas, J. B. (1972). The likelihood surface in the linear functional relationship problem. *Journal of the Royal Statistical Society series B*, 34:274–278.

Davison, A. (2003). *Statistical Models*. Cambridge Series in Statistical and Probabilistic Mathematics. Cambridge University Press.

de Haan, L. (1984). A spectral representation for max-stable processes. *The Annals of Probability*, 12(4):1194–1204.

de Haan, L. and Pereira, T. T. (2006). Spatial extremes: Models for the stationary case. *The Annals of Statistics*, 34:146–168.

Efron, B. (1993). Bayes and likelihood calculations from confidence intervals. *Biometrika*, 80(1):3–26.

Genton, M. G., Ma, Y., and Sang, H. (2011). On the likelihood function of Gaussian max-stable processes. *Biometrika*, 98:To appear.

Gholamrezaee, M. M. (2010). *Geostatistics of Extremes: A composite likelihood approach*. PhD thesis, École Polytechnique Fédérale de Lausanne.

Kabluchko, Z., Schlather, M., and de Haan, L. (2009). Stationary max-stable fields associated to negative definite functions. *Ann. Prob.*, 37(5):2042–2065.

Kent, J. T. (1982). Robust properties of likelihood ratio tests. *Biometrika*, 69:19–27.

Lazar, N. A. (2003). Bayesian empirical likelihood. *Biometrika*, 90(2):319–326.

Lindsay, B. (1988). *Composite likelihood methods*. Statistical Inference from Stochastic Processes. American Mathematical Society, Providence.

Monahan, J. and Boos, D. (1992). Proper likelihoods for Bayesian analysis. *Biometrika*, 79(2):271–278.

Padoan, S., Ribatet, M., and Sisson, S. (2010). Likelihood-based inference for max-stable processes. *Journal of the American Statistical Association (Theory & Methods)*, 105(489):263–277.

Pauli, F., Racugno, W., and Ventura, L. (2011). Bayesian composite marginal likelihoods. *Statistica Sinica*, 21:149–164.

Robert, C. P. and Casella, G. (2005). *Monte Carlo Statistical Methods (Springer Texts in Statistics)*. Springer-Verlag New York, Inc., Secausus, NJ, USA.

Rotnitzky, A. and Jewell, N. (1990). Hypothesis testing of regression parameters in semiparametric generalized linear models for cluster correlated data. *Biometrika*, 77:495–497.

Rue, H. and Tjelmeland, H. (2002). Fitting gaussian markov random fields to gaussian fields. *Scandinavian Journal Of Statistics*, 29(1):31–49.

Rydén, T. and Titterington, D. M. (1998). Computational Bayesian analysis of hidden Markov models. *Journal of Computational and Graphical Statistics*, 7(2):194–211.

Sang, H. and Gelfand, A. (2009). Hierarchical modeling for extreme values observed over space and time. *Environmental and Ecological Statistics*, 16(3):407–426.

Schlather, M. (2002). Models for stationary max-stable random fields. *Extremes*, 5(1):33–44.

Smith, E. L. and Stephenson, A. G. (2009). An extended Gaussian max-stable process model for spatial extremes. *Journal of Statistical Planning and Inference*, 139:1266–1275.

Smith, R. L. (1990). Max-stable processes and spatial extreme. *Unpublished manuscript*.

Varin, C. (2008). On composite marginal likelihoods. *AStA Advances in Statistical Analysis*, 92(1):1–28.

Ventura, L., Cabras, S., and Racugno, W. (2009). Prior distributions from pseudo-likelihoods in the presence of nuisance parameters. *Journal of the American Statistical Association*, 104(486):768–774.