Received: 19 November 2018 | Accepted: 20 December 2018
DOI: 10.1002/sta4.218

WILEY

ORIGINAL ARTICLE

Full likelihood inference for max-stable data

Raphaël Huser[1] | Clément Dombry[2] | Mathieu Ribatet[3] | Marc G. Genton[1]

[1]CEMSE Division, King Abdullah University of Science and Technology (KAUST), Thuwal, Saudi Arabia
[2]Department of Mathematics, University of Franche-Comté, Besançon cedex, France
[3]Department of Mathematics, University of Montpellier, Montpellier cedex 5, France

Correspondence
Raphaël Huser, CEMSE Division, King Abdullah University of Science and Technology (KAUST), Thuwal 23955-6900, Saudi Arabia.
Email: raphael.huser@kaust.edu.sa

Funding information
King Abdullah University of Science and Technology

We show how to perform full likelihood inference for max-stable multivariate distributions or processes based on a stochastic expectation–maximization algorithm, which combines statistical and computational efficiency in high dimensions. The good performance of this methodology is demonstrated by simulation based on the popular logistic and Brown–Resnick models, and it is shown to provide computational time improvements with respect to a direct computation of the likelihood. Strategies to further reduce the computational burden are also discussed.

KEYWORDS
full likelihood, max-stable distribution, Stephenson–Tawn likelihood, stochastic expectation–maximization algorithm

1 | INTRODUCTION

Max-stable distributions and processes are useful for studying high-dimensional extreme events recorded in space and/or time (Davis, Klüppelberg, & Steinkohl, 2013; Davison, Huser, & Thibaud, 2013; de Carvalho & Davison, 2014; Huser & Davison, 2014; Huser & Genton, 2016; Padoan, Ribatet, & Sisson, 2010). This broad but constrained class of models may, at least theoretically, be used to extrapolate into the joint tail, hence providing a justified framework for risk assessment of extreme events. The probabilistic justification for using these models is that the max-stable property arises in limiting models for suitably renormalized maxima of independent and identically distributed processes; see, for example, Davison, Padoan, and Ribatet (2012), Davison and Huser (2015), and Davison, Huser, and Thibaud (2019).

Because extremes are rare by definition, it is crucial for reliable estimation and prediction to extract as much information from the data as possible. Thus, efficient estimators play a particularly important role in statistics of extremes. Although nonparametric estimators (Vettori, Huser, & Genton, 2018), *M*-estimators (Einmahl, Kiriliouk, Krajina, & Segers, 2016), and generalized least squares estimators (Buhl & Klüppelberg, 2019) perform quite well and have reasonable efficiency in low dimensions, likelihood-based estimators remain a natural choice in higher dimensions thanks to their appealing large-sample properties. However, the full likelihood function is excessively difficult to compute for high-dimensional data following a max-stable distribution. As detailed in Section 3, likelihood evaluations require the computation of a sum indexed by all elements of a given set $\mathcal{P}_D$, the cardinality of which grows more than exponentially with the dimension, D. In a thorough simulation study, Castruccio, Huser, and Genton (2016) stated that current technologies are limiting full likelihood inference to dimension 12 or 13, and they concluded that without meaningful methodological advances, a direct full likelihood approach will not be feasible.

To circumvent this computational bottleneck, several strategies have been advocated. Padoan et al. (2010) proposed a pairwise likelihood approach, combining the bivariate densities of carefully chosen pairs of observations. Although this method is computationally attractive and inherits many good properties from the maximum likelihood estimator, it also entails a loss in efficiency, which becomes more apparent in high dimensions (Huser, Davison, & Genton, 2016). More efficient triplewise and higher order composite likelihoods were investigated by Genton, Ma, and Sang (2011), Huser and Davison (2013), Sang and Genton (2014), and Castruccio et al. (2016). However, they are still not fully efficient, and it is not clear how to optimally select the composite likelihood terms. Furthermore, because composite likelihoods are generally not valid likelihoods (Varin et al., 2011), the classical likelihood theory cannot be blindly applied for uncertainty assessment, testing, model validation and selection, and so on, and Bayesian inference based on composite likelihoods is tricky, too (Ribatet, Cooley, & Davison, 2012).

Alternatively, Stephenson and Tawn (2005) suggested augmenting the componentwise block maxima data $\boldsymbol{z}^n = (z_1^n, \ldots, z_D^n)^\top$, where n is the block size, with their occurrence times. This extra information may be summarized by a random partition π^n of the set $\{1, \ldots, D\}$, which indicates

© 2019 John Wiley & Sons, Ltd.

whether or not these maxima occurred simultaneously. Essentially, the Stephenson–Tawn likelihood corresponds to the limiting joint "density" of $\boldsymbol{z}^n$ and π^n, as $n \to \infty$, and it yields drastic simplifications and improved efficiency; see also Bienvenüe and Robert (2017). However, Wadsworth (2015) and Huser et al. (2016) noted that this approach may be severely biased for a finite n, especially in low-dependence scenarios. By fixing the limit partition, π, to the observed one, π^n, a strong constraint is imposed, creating model misspecification, to which likelihood methods are very sensitive.

In this paper, to mitigate the subasymptotic bias due to fixing the limit partition to the observed one, we suggest returning to the original likelihood formulation, which integrates out the partition rather than treating it as known. By interpreting the limit partition π as missing data, we show how to design a stochastic expectation–maximization algorithm (Dempster, Laird, & Rubin, 1977; Nielsen, 2000) for efficient inference. The quality of the stochastic approximation to the full likelihood can be controlled and set to any arbitrary precision at a computational cost. We show that higher dimensional max-stable models may be fitted in reasonable time. Importantly, our method is based solely on max-stable data and does not require extra information about the partition or the original processes, unlike the Stephenson–Tawn likelihood or related threshold-based methods (Huser et al., 2016). Our approach exploits the algorithm of Dombry, Éyi-Minko, and Ribatet (2013) for conditional simulation of the partition given the data, and it can be linked to the recent papers of Thibaud, Aalto, Cooley, Davison, and Heikkinen (2016) and Dombry et al. (2017), who in a Bayesian setting developed a Markov chain Monte Carlo algorithm for max-stable processes by treating the partition as a latent variable that is resampled at each iteration.

The paper is organized as follows: In Section 2, we recall preliminaries on max-stable distributions and processes. In Section 3, we detail the full and Stephenson–Tawn likelihoods, and we describe our novel stochastic expectation–maximization algorithm. To illustrate the performance and benefits of our approach, we report simulation results for the logistic and Brown–Resnick max-stable models in Section 4. Finally, Section 5 concludes with a brief discussion.

2 | MAX-STABLE PROCESSES AND DISTRIBUTIONS

2.1 | Definition, construction, and models

Consider a sequence of independent and identically distributed processes $Y_1(\boldsymbol{s}), Y_2(\boldsymbol{s}), \ldots$, indexed by spatial site $\boldsymbol{s} \in \mathcal{S} \subset \mathbb{R}^d$, and assume that there exist sequences of functions $a_n(\boldsymbol{s}) > 0$ and $b_n(\boldsymbol{s})$, such that the renormalized pointwise block maximum process (with block size n),

$$Z^n(\boldsymbol{s}) = a_n(\boldsymbol{s})^{-1}\left[\max\{Y_1(\boldsymbol{s}), \ldots, Y_n(\boldsymbol{s})\} - b_n(\boldsymbol{s})\right], \tag{1}$$

converges in the sense of finite-distributional distributions, as $n \to \infty$, to a process $Z(\boldsymbol{s})$ with nondegenerate marginal distributions; that is, $Y(\boldsymbol{s})$ is in the max-domain of attraction of $Z(\boldsymbol{s})$. Then, the limit $Z(\boldsymbol{s})$ is max-stable (see, e.g., de Haan & Ferreira, 2006, chap. 9). That is, pointwise maxima of independent copies of the limit process $Z(\boldsymbol{s})$ remain in the same location-scale family. Mathematically, this means that for every integer $m \in \mathbb{N}$, there exist functions $\alpha_m(\boldsymbol{s}) > 0$ and $\beta_m(\boldsymbol{s})$ such that

$$\max\{Z_1(\boldsymbol{s}), \ldots, Z_m(\boldsymbol{s})\} \overset{\mathrm{D}}{=} \alpha_m(\boldsymbol{s})Z(\boldsymbol{s}) + \beta_m(\boldsymbol{s}), \tag{2}$$

where $Z_1(\boldsymbol{s}), \ldots, Z_m(\boldsymbol{s})$ are independent copies of $Z(\boldsymbol{s})$ and $\overset{\mathrm{D}}{=}$ denotes equality in distribution. In particular, the process in Equation 2 has the same dependence structure as $Z(\boldsymbol{s})$ itself, whereas its marginal distributions may differ in location and scale and coincide with the generalized extreme value distribution.

Consider now points of a unit rate Poisson point process, $P_1, P_2, \ldots$, on $(0, +\infty)$, and independent copies, $W_1(\boldsymbol{s}), W_2(\boldsymbol{s}), \ldots$, of a stochastic process $W(\boldsymbol{s}) \geq 0$ with unit mean. Then, the process

$$Z(\boldsymbol{s}) = \sup_{j \geq 1} W_j(\boldsymbol{s})/P_j, \quad \boldsymbol{s} \in \mathcal{S} \tag{3}$$

is max-stable with unit Fréchet marginal distributions, that is, $\mathrm{pr}\{Z(\boldsymbol{s}) \leq z\} = \exp(-1/z)$, $z > 0$ (de Haan, 1984; Schlather, 2002). In the remainder of the paper, we shall always consider max-stable processes $Z(\boldsymbol{s})$ with unit Fréchet marginal distributions. Representation (3) is useful to construct a wide variety of max-stable processes, such as the Smith model (Smith, 1990), the Schlather model (Schlather, 2002), the Brown–Resnick model (Kabluchko, Schlather, & de Haan, 2009), the extremal-t model (Opitz, 2013), and the Tukey g-and-h model (Xu & Genton, 2016), and to simulate from them (Dombry, Engelke, & Oesting, 2016; Schlather, 2002). Multivariate max-stable models can be constructed similarly by substituting the processes $W_j(\boldsymbol{s})$ in Equation 3 by analogous random vectors $\boldsymbol{W}_j = (W_{j1}, \ldots, W_{jD})^\top$. From Equation 3, we deduce that the joint distribution of $Z(\boldsymbol{s})$ at a finite collection of sites $\mathcal{S}_D = \{\boldsymbol{s}_1, \ldots, \boldsymbol{s}_D\} \subset \mathcal{S}$ may be expressed as

$$\mathrm{pr}\{Z(\boldsymbol{s}_1) \leq z_1, \ldots, Z(\boldsymbol{s}_D) \leq z_D\} = \exp\{-V(z_1, \ldots, z_D)\}, \tag{4}$$

where the exponent function $V(z_1, \ldots, z_D) = E\left[\max\{W(\boldsymbol{s}_1)/z_1, \ldots, W(\boldsymbol{s}_D)/z_D\}\right]$ satisfies homogeneity and marginal constraints (see, e.g., Davison & Huser, 2015). As an illustration, Figure 1 shows two independent realizations from the same Smith (1990) model on $\mathbb{R}$ defined by taking $W_j(s) = \varphi(s - U_j; \sigma^2)$, $s \in \mathcal{S} = \mathbb{R}$, in Equation 3, where $\varphi(\cdot; \sigma^2)$ is the normal density with zero mean and variance σ^2, and the U_js are points from a unit rate Poisson point process on the real line. Figure 4 below illustrates realizations from the Brown–Resnick process on $\mathbb{R}^2$.

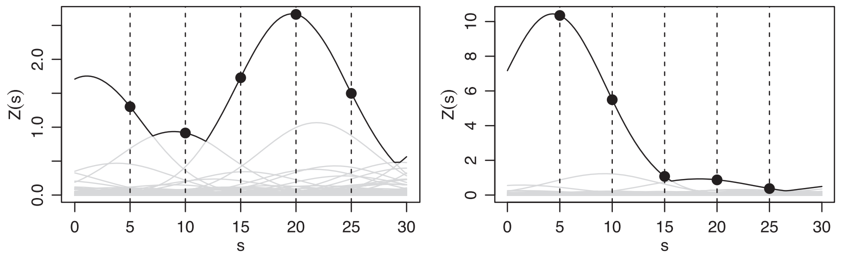

FIGURE 1 Two realizations (black) from the same Smith (1990) max-stable process on the line defined by setting $W_j(s) = \varphi(s - U_j; \sigma^2)$ in Equation 3, $s \in \mathcal{S} = \mathbb{R}$, where the U_js are points from a unit rate Poisson process on $\mathbb{R}$. Grey curves show all latent profiles $W_j(s)/P_j$. Here, $\sigma^2 = 5$. When observed at sites $s = 5, 10, 15, 20, 25$, the corresponding realized partitions are $\pi = \{\{1\}, \{2\}, \{3, 4, 5\}\}$ (left) and $\pi = \{\{1, 2, 3\}, \{4, 5\}\}$ (right)

2.2 | Underlying partition and extremal functions

At each site $\boldsymbol{s} \in \mathcal{S}$, the pointwise supremum, $Z(\boldsymbol{s})$, in Equation 3 is realized by a single profile $W_j(\boldsymbol{s})/P_j$ almost surely. Such profiles are called extremal functions in Dombry et al. (2013). As illustrated in Figure 1, the extremal functions are only partially observed on $\mathcal{S}_D = \{\boldsymbol{s}_1, \ldots, \boldsymbol{s}_D\}$; they define a random partition $\pi = \{\tau_1, \ldots, \tau_{|\pi|}\}$ (of size $|\pi|$) of the set $\{1, \ldots, D\}$, called the hitting scenario in Dombry et al. (2013), that identifies clusters of variables stemming from the same event. For example, the partition $\pi = \{\{1\}, \{2\}, \{3, 4, 5\}\}$ on the left panel of Figure 1 indicates that the max-stable process at these five sites came from three separate independent events; in particular, the maxima at sites $s_3 = 15$, $s_4 = 20$, and $s_5 = 25$ were generated from the same profile.

Similarly, an observed partition π^n of $\{1, \ldots, D\}$ may be defined for the pointwise maximum process $Z^n(\boldsymbol{s})$ in Equation 1, based on the original processes $Y_1(\boldsymbol{s}), \ldots, Y_n(\boldsymbol{s})$. The knowledge of π^n tells us if the block maxima (i.e., extreme events) at different sites occurred simultaneously or not, so π^n carries information about the strength of spatial extremal dependence. If the processes $Y_1(\boldsymbol{s}), \ldots, Y_n(\boldsymbol{s})$ (suitably marginally transformed) are in the max-domain of attraction of $Z(\boldsymbol{s})$ in Equation 3, then the observed partition π^n converges in distribution to π, as $n \to \infty$, on the space of all partitions $\mathcal{P}_D$ of $\{1, \ldots, D\}$ (Stephenson & Tawn, 2005).

We now describe likelihood inference for max-stable vectors: by exploiting the information on the partition π (Stephenson–Tawn likelihood) or by integrating it out (full likelihood).

3 | LIKELIHOOD INFERENCE

3.1 | Full and Stephenson–Tawn likelihoods

By differentiating the distribution (4) with respect to the variables $z_1, \ldots, z_D$, we can deduce that the corresponding density, or the full likelihood for one replicate, may be expressed as

$$g_{\text{Full}}(z_1, \ldots, z_D) = \exp\{-V(z_1, \ldots, z_D)\} \sum_{\pi \in \mathcal{P}_D} \prod_{i=1}^{|\pi|} \{-V_{\tau_i}(z_1, \ldots, z_D)\}, \tag{5}$$

where V_{τ_i} denotes the partial derivative of the function V with respect to the variables indexed by the set $\tau_i \subseteq \{1, \ldots, D\}$ (Huser et al., 2016; Castruccio et al., 2016). The sum in Equation 5 is taken over the set of all possible partitions $\pi = \{\tau_1, \ldots, \tau_{|\pi|}\}$ of $\{1, \ldots, D\}$, denoted by $\mathcal{P}_D$, the size of which equals the Bell number of order D. This leads to an explosion of terms, even for a moderate D. In fact, each partition that appears on the right-hand side of Equation 5 corresponds to a different configuration of the profiles $W_j(\boldsymbol{s})/P_j$ in Equation 3 at the sites $\boldsymbol{s}_1, \ldots, \boldsymbol{s}_D$. Thus, Castruccio et al. (2016) argued that the computation of Equation 5 is limited to dimension $D = 12$ or 13 with modern computational resources.

As demonstrated in Appendix A using a point process argument, and originally shown by Stephenson and Tawn (2005), the joint density of the max-stable data $\boldsymbol{z} = (z_1, \ldots, z_D)^\top$ and the associated partition $\pi = \{\tau_1, \ldots, \tau_{|\pi|}\} \in \mathcal{P}_D$ is simply equal to

$$g_{\text{ST}}(z_1, \ldots, z_D, \pi) = \exp\{-V(z_1, \ldots, z_D)\} \prod_{i=1}^{|\pi|} \{-V_{\tau_i}(z_1, \ldots, z_D)\}, \tag{6}$$

hence reducing the problematic sum to a single term, making likelihood inference possible in higher dimensions and simultaneously improving statistical efficiency. Because the asymptotic partition π is not observed, Stephenson and Tawn (2005) suggested replacing it by the observed partition π^n of occurrence times of maxima, which converges to π under mild conditions provided the asymptotic model is well specified. However, Wadsworth (2015) and Huser et al. (2016) showed that lack of convergence of π^n to π may result in severe estimation bias, which is especially strong in low-dependence cases. To circumvent this problem, Wadsworth (2015) proposed a bias-corrected likelihood; alternatively, we show in the next section how to design a stochastic expectation–maximization algorithm to maximize Equation 5, while taking advantage of the computationally appealing nature of Equation 6.

3.2 | Stochastic expectation–maximization algorithm

It is instructive to rewrite the full likelihood (5) using Equation 6 as $g_{\text{Full}}(z_1, \dots, z_D) = \sum_{\pi \in \mathcal{P}_D} g_{\text{ST}}(z_1, \dots, z_D, \pi)$ because it highlights that the full likelihood simply integrates out the latent random partition π needed for the Stephenson–Tawn likelihood. By interpreting π as a missing observation and the Stephenson–Tawn likelihood as the completed likelihood, an expectation–maximization algorithm (Dempster et al., 1977) may be easily formulated. Assume that the exponent function $V(z_1, \dots, z_D|\boldsymbol{\theta})$ is parametrized by a vector $\boldsymbol{\theta} \in \Theta \subset \mathbb{R}^p$. Starting from an initial guess $\boldsymbol{\theta}_0 \in \Theta$, the expectation–maximization algorithm consists of iterating the following E- and M-steps for R iterations:

- E-step: At the rth iteration, compute the functional

$$Q(\boldsymbol{\theta}, \boldsymbol{\theta}_{r-1}) = E_{\pi|\boldsymbol{z},\boldsymbol{\theta}_{r-1}}\left[\log\{g_{\text{ST}}(\boldsymbol{z}, \pi|\boldsymbol{\theta})\}\right] = \sum_{\pi \in \mathcal{P}_D} g(\pi|\boldsymbol{z}, \boldsymbol{\theta}_{r-1}) \log\{g_{\text{ST}}(\boldsymbol{z}, \pi|\boldsymbol{\theta})\}, \tag{7}$$

where the expectation is computed with respect to the discrete conditional distribution of π given the data $\boldsymbol{z} = (z_1, \dots, z_D)^\top$ and the current value of the parameter $\boldsymbol{\theta}_{r-1}$, that is,

$$g(\pi|\boldsymbol{z}, \boldsymbol{\theta}_{r-1}) = g_{\text{ST}}(\boldsymbol{z}, \pi|\boldsymbol{\theta}_{r-1}) / g_{\text{Full}}(\boldsymbol{z}|\boldsymbol{\theta}_{r-1}). \tag{8}$$

- M-step: At the rth iteration, update the parameter as $\boldsymbol{\theta}_r = \arg\max_{\boldsymbol{\theta}\in\Theta} Q(\boldsymbol{\theta}, \boldsymbol{\theta}_{r-1})$.

Dempster et al. (1977) showed that the expectation–maximization algorithm has appealing properties; in particular, the value of the log-likelihood increases at each iteration, which ensures convergence of $\boldsymbol{\theta}_r$ to a local maximum, as $r \to \infty$. In our case, however, the expectation in Equation 7 is tricky to compute: It contains again the sum over the set $\mathcal{P}_D$, and Equation 8 relies on the full density $g_{\text{Full}}(\boldsymbol{z}|\boldsymbol{\theta}_{r-1})$, which we try to avoid. To circumvent this issue, one solution is to approximate Equation 7 by Monte Carlo as

$$\hat{Q}(\boldsymbol{\theta}, \boldsymbol{\theta}_{r-1}) = \frac{1}{N}\sum_{i=1}^{N} \log\{g_{\text{ST}}(\boldsymbol{z}, \pi_i|\boldsymbol{\theta})\}, \quad \pi_1, \dots, \pi_N \sim g(\pi|\boldsymbol{z}, \boldsymbol{\theta}_{r-1}), \tag{9}$$

where the partitions $\pi_1, \dots, \pi_N$ are conditionally independent at best or form an ergodic sequence at least. As $g(\pi|\boldsymbol{z}, \boldsymbol{\theta}_{r-1}) \propto g_{\text{ST}}(\boldsymbol{z}, \pi|\boldsymbol{\theta}_{r-1})$—see Equation 8—it is possible to devise a Gibbs sampler to generate approximate simulations from $g(\pi|\boldsymbol{z}, \boldsymbol{\theta}_{r-1})$ without explicitly computing the constant factor $g_{\text{Full}}(\boldsymbol{z}|\boldsymbol{\theta}_{r-1})$ in the denominator of Equation 8. Thanks to ergodicity of the resulting Markov chain, the precision of the approximation (9) may be set arbitrarily high by letting $N \to \infty$ (and discarding some burn-in iterations). More details about the practical implementation of the Gibbs sampler are given in Dombry et al. (2013) and in Appendix B. Although the number of iterations of the Gibbs sampler, N, will typically be much smaller than the cardinality of $\mathcal{P}_D$, the approximation (9) to Equation 7 will likely be reasonably good for moderate values of N because only a few partitions $\pi \in \mathcal{P}_D$ may be plausible or compatible with the data $\boldsymbol{z} = (z_1, \dots, z_D)^\top$.

The asymptotic properties of the stochastic expectation–maximization estimator, $\hat{\boldsymbol{\theta}}_{\text{SEM}}$, were studied in details by Nielsen (2000) and compared with the classical maximum likelihood estimator, $\hat{\boldsymbol{\theta}}$ (see §2–3 therein, in particular theorem 2). Dombry, Engelke, and Oesting (2017) showed that the maximum likelihood estimator $\hat{\boldsymbol{\theta}}$ is consistent and asymptotically normal for the most popular max-stable models, including the logistic and Brown–Resnick models used in this paper. This suggests that these appealing asymptotic properties should also be satisfied for the estimator $\hat{\boldsymbol{\theta}}_{\text{SEM}}$, provided some additional rather technical regularity conditions detailed in Nielsen (2000) are satisfied. If so, then the asymptotic performance of $\hat{\boldsymbol{\theta}}_{\text{SEM}}$ is akin to that of $\hat{\boldsymbol{\theta}}$, though with a slightly larger asymptotic variance. Finally, the inherent variability of the stochastic expectation–maximization algorithm may also be a blessing: Unlike the deterministic expectation–maximization algorithm, it is less likely to get stuck at a local maximum of the full likelihood.

4 | SIMULATION STUDY

4.1 | General setting

To assess the performance of the stochastic expectation–maximization estimator $\hat{\boldsymbol{\theta}}_{\text{SEM}}$, we conducted an extensive simulation study using two max-stable models: In Section 4.2, we report results for the multivariate logistic distribution, which is exchangeable and has a single dependence parameter, whereas in Section 4.3, we consider the Brown–Resnick model, which is a popular spatial max-stable process governed by a range parameter and a smoothness parameter. Whereas the logistic model serves as an illustrative “test case,” the Brown–Resnick model is much more challenging and computationally intensive.

All simulations presented below were performed using the KAUST Cray XC40 supercomputer Shaheen II. Nevertheless, our implementation can also be run on a standard laptop, and the computational times reported below correspond to *single core* experiments (at clock frequency 2.3 GHz). Available distributed computing resources may be exploited to significantly accelerate computations.

4.2 | Results for the logistic model

We start by considering the multivariate logistic max-stable distribution with exponent function $V(z_1, \ldots, z_D|\theta) = (\sum_{j=1}^{D} z_j^{-1/\theta})^{\theta}$, $\theta \in \Theta = (0, 1]$. Here, the parameter θ controls the dependence strength, with $\theta \to 0$ and $\theta = 1$ corresponding to perfect dependence and independence, respectively. This model was chosen for two main reasons: First, it is the simplest max-stable distribution, often used as a benchmark, that interpolates between perfect dependence and independence; and second, the full likelihood (5) can be efficiently computed in this case using a recursive algorithm (Shi, 1995), thus allowing us to compare $\hat{\theta}_{\text{SEM}}$ and the classical maximum likelihood estimator $\hat{\theta}$ in high dimensions.

We first investigated the statistical performance of the estimator $\hat{\theta}_{\text{SEM}}$ under different scenarios. We simulated logistic random vectors in dimension $D = 2, 5, 10, 20$, with 20 independent temporal replicates, and $\theta = 0.1, \ldots, 0.9$ (strong to weak dependence). Setting the initial value to $\theta_0 = 0.6$, we chose $R = 30$ iterations for the expectation–maximization algorithm, averaging the last five iterations, and we took $110 \times D$ iterations for the underlying Gibbs sampler. Following our simulations reported in Appendix B, we discarded the first $10 \times D$ iterations as burn-in, and we thinned the Markov chain by a factor D, in order to keep $N = 100$ roughly independent partitions π_i to compute Equation (9). We repeated the experiment 1,024 times to estimate the bias, $B = E(\hat{\theta}_{\text{SEM}}) - \theta$, the standard deviation, $\text{SD} = (E[\{\hat{\theta}_{\text{SEM}} - E(\hat{\theta}_{\text{SEM}})\}^2])^{1/2}$, the root mean squared error, $\text{RMSE} = (B^2 + \text{SD}^2)^{1/2} = [E\{(\hat{\theta}_{\text{SEM}} - \theta)^2\}]^{1/2}$, and the relative error with respect to the maximum likelihood estimator, $\text{RE} = E|(\hat{\theta}_{\text{SEM}} - \hat{\theta})/\hat{\theta}|$. Figure 2 reports the results. As expected, the bias is negligible compared with the standard deviation, and the latter decreases with increasing dimension D but increases as the data approach independence ($\theta \to 1$). The root mean squared error (not shown) is almost only determined by the standard deviation. The relative error is always very small (uniformly less than about 0.6%), and it decreases for the most part with D and also with N as suggested by further unreported simulations.

We turn now our attention to the computational efficiency of the stochastic expectation–maximization algorithm. Considering dimensions up to $D = 100$ under the same setting as before, the leftmost panel of Figure 3 shows that it takes on average 5–6 min to compute $\hat{\theta}_{\text{SEM}}$ when $D = 100$ and $\theta = 0.9$. Recall that, according to Castruccio et al. (2016), a direct evaluation of the likelihood (5) is not possible in dimensions greater than $D = 12$ or 13; thus, this result is a big improvement over the current existing methods. The computational time appears to be roughly linear with D, which is due to the number of iterations of the underlying Gibbs sampler being set proportional to D. To further reduce the

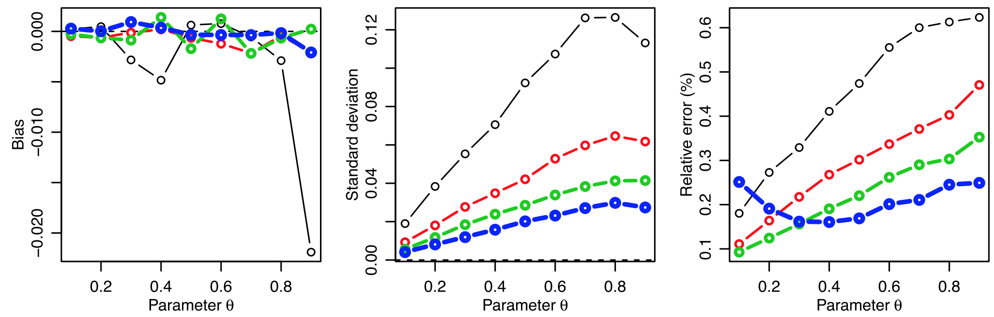

FIGURE 2 Performance of the estimator $\hat{\theta}_{\text{SEM}}$: bias (left), standard deviation (middle), and relative error in % (right), for the logistic model with $\theta = 0.1, \ldots, 0.9$ in dimension $D = 2$ (thinnest black), 5 (thin red), 10 (thick green), and 20 (thickest blue), based on 20 independent temporal replicates. The number of iterations for the expectation–maximization algorithm was set to $R = 30$, averaging the last five iterations, and the number of iterations of the underlying Gibbs sampler was set to $110 \times D$ (thinned by a factor D, after a burn-in of $10 \times D$ iterations). The initial value was set to $\theta_0 = 0.6$

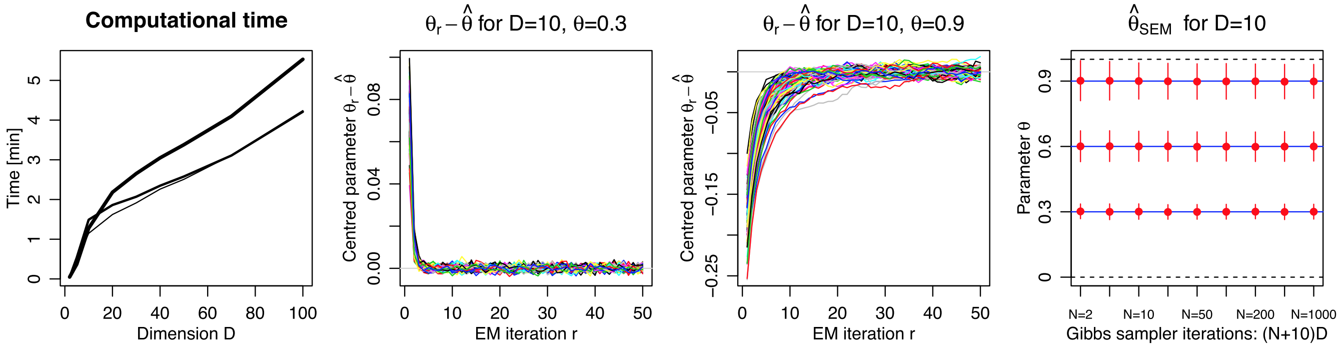

FIGURE 3 Left: Computational time for computing $\hat{\theta}_{\text{SEM}}$ for the logistic model, as function of dimension D, for $\theta = 0.3$ (thin), 0.6 (medium), and 0.9 (thick). We used 20 temporal replicates, 30 expectation–maximization (EM) iterations, and $110 \times D$ iterations for the underlying Gibbs sampler. Middle panels: $\theta_r - \hat{\theta}$ as function of iteration $r = 1, \ldots, 50$, for 100 independent runs in dimension $D = 10$. True values were set to $\theta = 0.3$ (second panel) and 0.9 (third panel), and the initial value was set to $\theta_0 = 0.6$. Right: Mean of estimated parameters $\hat{\theta}_{\text{SEM}}$ (red dots) with 95% confidence intervals (calculated from the 1,024 simulations) for $\theta = 0.3, 0.6, 0.9$, dimension $D = 10$, 30 expectation–maximization iterations, and $(N + 10) \times D$ Gibbs sampler iterations with $N = 2, 5, 10, 20, 50, 100, 200, 500, 1{,}000$ (x-axis). In all simulations, we always thinned the underlying Markov chains by a factor D after discarding a burn-in of $10 \times D$ iterations

computational burden, one possibility is to tune the number of iterations of the stochastic expectation–maximization algorithm. To investigate its speed of convergence, the two middle panels of Figure 3 show the sample path $r \mapsto \theta_r$ for the logistic model as a function of the iteration $r = 1, \ldots, 50$, centred by the maximum likelihood estimator $\hat{\theta}$ for 100 independent runs in dimension $D = 10$. The true values were set to $\theta = 0.3$ (second panel) and 0.9 (third panel), and the initial value was set to $\theta_0 = 0.6$. The convergence is quite fast when $\theta = 0.3$, requiring about five iterations, but when $\theta = 0.9$, it takes between 15 and 25 iterations. Another possibility to reduce the computational time is to play with the number of iterations of the underlying Gibbs sampler. The rightmost panel of Figure 3 displays estimated parameters in dimension $D = 10$ with associated 95% confidence intervals (calculated from the 1,024 simulations) for $\theta = 0.3, 0.6, 0.9$, using 30 expectation–maximization iterations and $(N + 10) \times D$ iterations for the underlying Gibbs sampler with $N = 2, 5, 10, 20, 50, 100, 200, 500,$ 1,000. Again, we discarded the first $10 \times D$ iterations as burn-in, and we thinned the resulting chain by a factor D. Surprisingly, the distribution of $\hat{\theta}_{\text{SEM}}$ is almost stable for all $N \geq 2$, suggesting that the number of Gibbs iterations does not need to be very large for accurate estimation. Overall, major computational savings can be achieved without significant loss of accuracy, by appropriately choosing the number R of iterations for the expectation–maximization algorithm and the number N of iterations of the underlying Gibbs sampler.

4.3 | Results for the Brown–Resnick model

We now provide simulation results for the popular Brown–Resnick model (Kabluchko et al., 2009), defined by taking $W_j(\boldsymbol{s}) = \exp\{\varepsilon_j(\boldsymbol{s}) - \gamma(\boldsymbol{s})\}$ in Equation (3), where the terms $\varepsilon_j(\boldsymbol{s})$ are independent copies of $\varepsilon(\boldsymbol{s})$, taken as an intrinsically stationary Gaussian process with zero mean and variogram $2\gamma(\boldsymbol{h}) = \text{var}\{\varepsilon(\boldsymbol{s}) - \varepsilon(\boldsymbol{s} + \boldsymbol{h})\}$ such that $\varepsilon(\boldsymbol{0}) = 0$ almost surely. The general form of the Brown–Resnick model's exponent function and its partial derivatives may be found in Huser and Davison (2013) and Wadsworth and Tawn (2014), respectively. We chose the isotropic semivariogram $\gamma(\boldsymbol{h}) = (\|\boldsymbol{h}\|/\lambda)^{\nu}$, where $\lambda > 0$ and $\nu \in (0, 2]$ are range and smoothness parameters, respectively, and we considered the scenarios displayed in Table 1. Realizations for each scenario are illustrated in Figure 4.

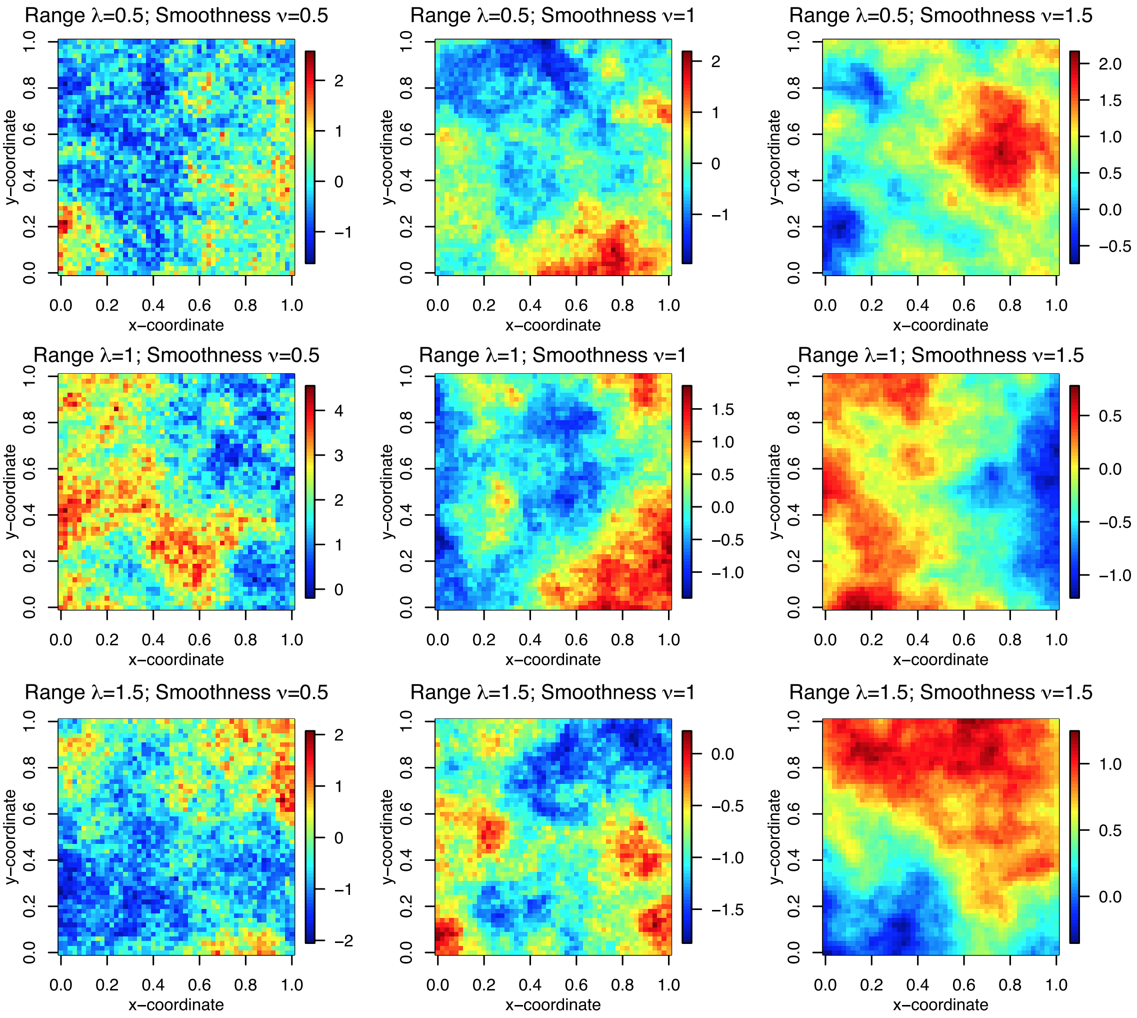

FIGURE 4 Realizations from the Brown–Resnick model on $[0, 1]^2$, with semivariogram $\gamma(\boldsymbol{h}) = (\|\boldsymbol{h}\|/\lambda)^{\nu}$, and parameter values taken according to Table 1, covering short- to long-range-dependent processes (top to bottom) and rough to smooth processes (left to right). Realizations, simulated exactly using the algorithm of Dombry et al. (2016), are displayed on standard Gumbel margins

TABLE 1 Scenarios considered for the simulation study based on the Brown–Resnick model with semivariogram $\gamma(\boldsymbol{h}) = (\|\boldsymbol{h}\|/\lambda)^{\nu}$

Scenario	1	2	3	4	5	6	7	8	9
$\lambda; \nu$	0.5; 0.5	0.5; 1.0	0.5; 1.5	1.0; 0.5	1.0; 1.0	1.0; 1.5	1.5; 0.5	1.5; 1.0	1.5; 1.5

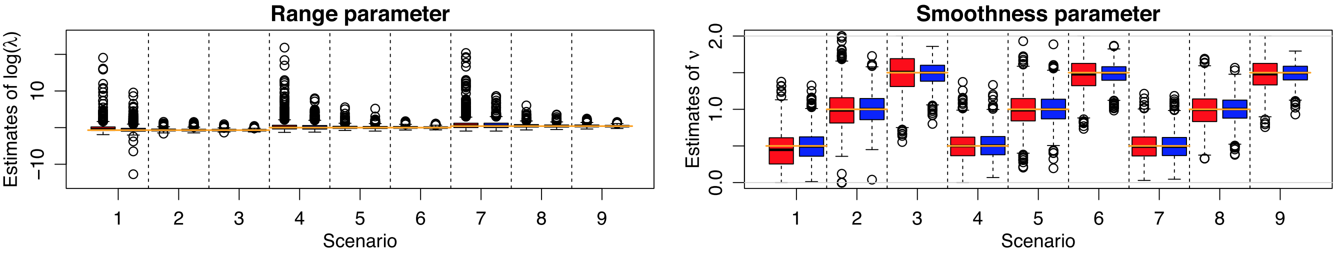

FIGURE 5 Boxplots of estimates of $\log(\lambda)$ (left panel) and ν (right panel) for each scenario in Table 1 based on the Brown–Resnick model with semivariogram $\gamma(\boldsymbol{h}) = (\|\boldsymbol{h}\|/\lambda)^{\nu}$, simulated at $D = 10$ random sites in $[0, 1]^2$, with 10 independent replicates. Red (left) and blue (right) boxplots correspond to $\hat{\boldsymbol{\theta}}_{\mathrm{PAIR}} = (\hat{\lambda}_{\mathrm{PAIR}}, \hat{\nu}_{\mathrm{PAIR}})^{\top}$ and $\hat{\boldsymbol{\theta}}_{\mathrm{SEM}} = (\hat{\lambda}_{\mathrm{SEM}}, \hat{\nu}_{\mathrm{SEM}})^{\top}$, respectively. Each boxplot is based on 1,024 simulations. Five estimates reaching up to $\log(\hat{\lambda}_{\mathrm{PAIR}}) \approx 40$ were omitted in Scenario 1 for visibility purposes. Orange horizontal segments are the true values

TABLE 2 Relative efficiencies for the estimates of $\log(\lambda)$ (number on the left) and ν (number on the right) based on the pairwise likelihood estimator with respect to the stochastic expectation–maximization estimator

$\lambda = 0.5$			$\lambda = 1.0$			$\lambda = 1.5$		
$\nu = 0.5$	$\nu = 1.0$	$\nu = 1.5$	$\nu = 0.5$	$\nu = 1.0$	$\nu = 1.5$	$\nu = 0.5$	$\nu = 1.0$	$\nu = 1.5$
39%; 75%	94%; 80%	77%; 59%	57%; 92%	82%; 87%	73%; 62%	54%; 92%	79%; 82%	71%; 61%

Note. Simulations were based on the Brown–Resnick model with semivariogram $\gamma(\boldsymbol{h}) = (\|\boldsymbol{h}\|/\lambda)^{\nu}$ for each scenario of Table 1, simulated at $D = 10$ random sites in $[0, 1]^2$ with 10 independent replicates.

In order to assess the performance of the stochastic expectation–maximization estimator for each scenario of Table 1, we simulated in each case 10 independent copies of the Brown–Resnick model at $D = 10$ randomly generated sites in $[0, 1]^2$, and then we estimated the range and smoothness parameters. We used (a) the stochastic expectation–maximization estimator $\hat{\boldsymbol{\theta}}_{\mathrm{SEM}} = (\hat{\lambda}_{\mathrm{SEM}}, \hat{\nu}_{\mathrm{SEM}})^{\top}$ based on $60 \times D = 600$ Gibbs sampler iterations in total, and then thinning by a factor $D = 10$, after discarding a burn-in of $10 \times D = 100$ iterations as suggested by the results in Appendix B; and (b) a pairwise likelihood estimator $\hat{\boldsymbol{\theta}}_{\mathrm{PAIR}} = (\hat{\lambda}_{\mathrm{PAIR}}, \hat{\nu}_{\mathrm{PAIR}})^{\top}$ (see, e.g., Huser & Davison, 2013; Padoan et al., 2010; Varin et al., 2011), which maximizes the pairwise likelihood constructed by combining the likelihood contributions from all $\binom{10}{2} = 45$ pairs of sites together with equal weights. We repeated this experiment 1,024 times to compute performance metrics, such as the root mean squared error of parameter estimates.

Figure 5 displays boxplots of estimated parameters for each scenario. Both the stochastic expectation–maximization and pairwise likelihood estimators seem to work well overall and have a negligible bias, although the estimation variability is quite high in some cases due to the tricky estimation exercise with only 10 replicates in dimension $D = 10$. Nevertheless, the interquartile range appears to be quite moderate in all cases. The stochastic expectation–maximization estimator is clearly superior to the pairwise likelihood estimator, as it fully utilizes the information available in the data. To investigate this further, Table 2 reports relative efficiencies of the pairwise likelihood estimator with respect to the stochastic expectation–maximization estimator, defined as the ratio between their root mean squared errors (calculated from the 1,024 replicates). The results suggest that the stochastic expectation–maximization estimator has a much better performance overall, as expected. Moreover, such results are expected to improve in higher dimensional settings, where the loss in efficiency of pairwise likelihood estimators is more significant.

We now check the speed of convergence of the expectation–maximization algorithm, similarly to our simulations for the logistic model reported in Section 4.2. Figure 6 displays the value of the likelihood (left panel) and the parameters (right panel) for each iteration $r = 1, \ldots, 20$ of the expectation–maximization algorithm, for 100 simulations performed in the same setting as above with true values chosen to be $\lambda = \nu = 1.5$. The likelihood values were centred by their average over iterations 16–20 for visibility purposes. The plots show that the expectation–maximization algorithm converges after roughly five iterations in this case, which is similar to the results obtained for the logistic model, despite the fact that the Brown–Resnick model has one more parameter ($p = 2$). Hence, in practice, a small number of iterations could be chosen to speed up the algorithm.

The right panel of Figure 6 also reveals that the estimated range parameter is negatively correlated with the estimated smoothness parameter. This was expected as these two parameters have an opposing effect on the dependence strength, and it suggests that alternative orthogonal parametrizations might be preferable. We leave this problem for future research.

Finally, we investigate the scalability of the stochastic expectation–maximization estimator for the Brown–Resnick model when the dimension D increases. To assess this, we considered the same setting as above with 10 independent replicates, same number of Gibbs sampler iterations, and true values set to $\lambda = \nu = 1.5$, and we measured the computational time needed for the first expectation–maximization iteration in dimensions $D = 5, 10, 15, 20$. Unlike the logistic model, which has an explicit exponent function V and partial derivatives V_{τ_i}, the expressions for

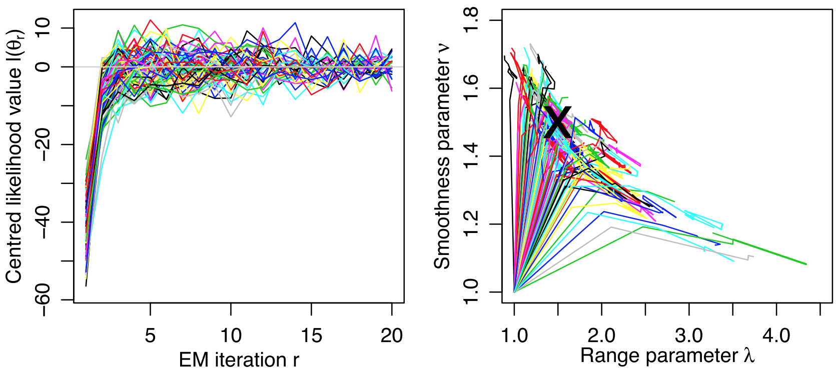

FIGURE 6 Left: Likelihood value $l(\theta_r)$ plotted as a function of the expectation–maximization (EM) iteration $r = 1, \ldots, 20$, for 100 simulations based on the Brown–Resnick model with semivariogram $\gamma(\boldsymbol{h}) = (\|\boldsymbol{h}\|/\lambda)^{\nu}$, simulated at $D = 10$ random sites in $[0,1]^2$, with 10 independent replicates. The likelihood values were centred by their average over iterations 16–20 for visibility purposes. Right: Trace of corresponding parameter values $\boldsymbol{\theta}_r = (\lambda_r, \nu_r)^{\top}$, plotted for each expectation–maximization iteration $r = 0, 1, \ldots, 20$. The true parameter values (black cross ×) are $\lambda = \nu = 1.5$, while the initial values were taken to be $\lambda_0 = \nu_0 = 1$

the Brown–Resnick model involve the multivariate Gaussian distribution in dimension up to $D - 1$ (see Huser & Davison, 2013 and Wadsworth & Tawn, 2014), whose computation with the Genz–Bretz algorithm implemented in the R package mvtnorm is very demanding for large D. This significantly slows down the algorithm, and a single expectation–maximization iteration using a single core takes on average 1.5 min, 12.7 min, 52.2 min, and 19.8 hr in dimensions $D = 5, 10, 15$, and 20, respectively. However, recall that Castruccio et al. (2016) argued that a direct likelihood evaluation was simply impossible beyond dimension $D = 12$ or 13. As the main computational bottleneck relates to the computation of multivariate Gaussian probabilities, strategies to speed up the Genz–Bretz algorithm are crucially needed beyond $D = 20$. de Fondeville and Davison (2018) suggested that major speed-ups can be achieved by appropriately using quasi-Monte Carlo techniques for the calculation of multivariate Gaussian distributions. Alternatively, hierarchical matrix decompositions (Genton, Keyes, & Turkiyyah, 2018) have been proven to be exceedingly accurate and fast in high dimensions. Moreover, significant speed-ups may also be obtained by efficiently exploiting distributed computing resources and running all Gibbs samplers (i.e., one for each temporal replicate) in parallel. We leave these computational improvements for future research.

As far as other max-stable processes are concerned, a similar computational burden is expected for the extremal-t model (Opitz, 2013), which relies on the computation of multivariate Student t distributions, but a better computational efficiency should prevail for the Reich and Shaby (2012) max-stable model, for which the expressions of the exponent function V and its partial derivatives V_{τ_i} are available in explicit form; see the appendix of Castruccio et al. (2016). Overall, our experiments on the logistic and Brown–Resnick models open the road to full likelihood inference for general max-stable models in dimensions higher than what was possible before, with a better overall scalability.

5 | DISCUSSION

To address the problem of inference for max-stable distributions and processes, we have proposed a stochastic expectation–maximization algorithm, which does not fix the underlying partition but, instead, treats it as a missing observation and integrates it out. The beauty of this approach is that it combines statistical and computational efficiency in high dimensions, and it does not suffer from misspecification entailed by lack of convergence of the partition. As a proof of concept, we have validated the methodology by simulation based on the logistic model, and we have shown that in this case it is easy to make inference beyond dimension $D = 100$ in just a few minutes. We have also provided results for the popular Brown–Resnick spatial max-stable model. In this case, our full likelihood inference approach can handle dimensions up to about $D = 20$ in a reasonable amount of time. The difficulty resides in the computation of high-dimensional multivariate Gaussian distributions needed for the exponent function V and its partial derivatives V_{τ_i}. Unbiased Monte Carlo estimates of these quantities can be obtained, and Thibaud et al. (2016) and de Fondeville and Davison (2018) suggest using crude approximations to reduce the computational time while maintaining accuracy; see also Genton et al. (2018), who instead suggest using hierarchical matrix decompositions. Our method is not limited to these two models and could potentially be applied to any max-stable model for which the functions V and V_{τ_i} are known and computable. The main computational bottleneck of our approach is that we need to generate a Gibbs sampler for each independent temporal replicate of the process. Fortunately, as this setting is embarrassingly parallel, we may thus easily take advantage of available distributed computing resources. Finally, there is a large volume of literature on the stochastic expectation–maximization algorithm, and it might be possible to devise automatic stopping criteria and adaptive schemes for the Gibbs sampler to further speed up the algorithm (Booth & Hobert, 1999).

In this paper, we have also explored the variability of the stochastic expectation–maximization estimator by simulation, and we have shown that by fully exploiting the information in the data, it has a better efficiency than composite likelihood estimators, even in the moderate dimension we have considered ($D = 10$ for the Brown–Resnick model). We predict these gains to be even more significant in higher dimensions, although the computational aspect becomes also more challenging. In practice, it may be tricky to assess the uncertainty of the stochastic expectation–maximization estimator and to provide confidence intervals based on a single dataset. Using bootstrap methods might be an option, but this would explode the computational burden and thus require each bootstrap sample to be treated in parallel on a different core. In future research, it would be interesting to study bootstrap confidence intervals and their coverage for the stochastic expectation–maximization estimator.

ACKNOWLEDGEMENTS

This research was supported by King Abdullah University of Science and Technology (KAUST). This research made use of the resources of the KAUST Supercomputing Laboratory.

ORCID

Raphaël Huser https://orcid.org/0000-0002-1228-2071
Marc G. Genton http://orcid.org/0000-0001-6467-2998

REFERENCES

Bienvenüe, A., & Robert, C. Y. (2017). Likelihood inference for multivariate extreme value distributions whose spectral vectors have known conditional distributions. *Scandinavian Journal of Statistics*, *44*(1), 130–149.

Booth, J. G., & Hobert, J. P. (1999). Maximizing generalized linear mixed model likelihoods with an automated Monte Carlo EM algorithm. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, *61*, 265–285.

Buhl, S., & Klüppelberg, C. (2019). Generalised least squares estimation of regularly varying space–time processes based on flexible observation schemes. Extremes. To appear.

Castruccio, S., Huser, R., & Genton, M. G. (2016). High-order composite likelihood inference for max-stable distributions and processes. *Journal of Computational and Graphical Statistics*, *25*, 1212–1229.

Davis, R. A., Klüppelberg, C., & Steinkohl, C. (2013). Statistical inference for max-stable processes in space and time. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, *75*, 791–819.

Davison, A. C., & Huser, R. (2015). Statistics of extremes. *Annual Review of Statistics and its Application*, *2*, 203–235.

Davison, A. C., Huser, R., & Thibaud, E. (2013). Geostatistics of dependent and asymptotically independent extremes. *Mathematical Geosciences*, *45*, 511–529.

Davison, A. C., Huser, R., & Thibaud, E. (2019). Spatial extremes. In Gelfand, A. E., Fuentes, M., & Smith, R. L. (Eds.), *Handbook of Environmental and Ecological Statistics*, CRC Press. To appear. https://www.crcpress.com/Handbook-of-Environmental-and-Ecological-Statistics/Gelfand-Fuentes-Hoeting-Smith/p/book/9781498752022

Davison, A. C., Padoan, S., & Ribatet, M. (2012). Statistical modelling of spatial extremes (with discussion). *Statistical Science*, *27*, 161–86.

de Carvalho, M., & Davison, A. C. (2014). Spectral density ratio models for multivariate extremes. *Journal of the American Statistical Association*, *109*, 764–776.

de Fondeville, R., & Davison, A. C. (2018). High-dimensional peaks-over-threshold inference. *Biometrika*, *105*, 575–592.

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

de Haan, L., & Ferreira, A. (2006). *Extreme value theory: An introduction*. New York: Springer.

Dempster, A. P., Laird, N. M., & Rubin, D. B. (1977). Maximum likelihood from incomplete data via the EM algorithm. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, *39*, 1–38.

Dombry, C., Engelke, S., & Oesting, M. (2016). Exact simulation of max-stable processes. *Biometrika*, *103*, 303–317.

Dombry, C., Engelke, S., & Oesting, M. (2017). Asymptotic properties of the maximum likelihood estimator for multivariate extreme value distributions. arXiv preprint 1612.05178.

Dombry, C., Engelke, S., & Oesting, M. (2017). Bayesian inference for multivariate extreme value distributions. *Electronic Journal of Statistics*, *11*, 4813–4844.

Dombry, C., & Éyi-Minko, F. (2013). Regular conditional distributions of continuous max-infinitely divisible random fields. *Electronic Journal of Probability*, *18*, 1–21.

Dombry, C., Éyi-Minko, F., & Ribatet, M. (2013). Conditional simulation of max-stable processes. *Biometrika*, *100*, 111–124.

Einmahl, J. H. J., Kiriliouk, A., Krajina, A., & Segers, J. (2016). An *M*-estimator of spatial tail dependence. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, *78*, 275–298.

Genton, M. G., Keyes, D. E., & Turkiyyah, G. (2018). Hierarchical decompositions for the computation of high-dimensional multivariate normal probabilities. *Journal of Computational and Graphical Statistics*, *27*, 268–277.

Genton, M. G., Ma, Y., & Sang, H. (2011). On the likelihood function of Gaussian max-stable processes. *Biometrika*, *98*, 481–488.

Huser, R., & Davison, A. C. (2013). Composite likelihood estimation for the Brown–Resnick process. *Biometrika*, *100*, 511–518.

Huser, R., & Davison, A. C. (2014). Space–time modelling of extreme events. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, *76*, 439–461.

Huser, R., Davison, A. C., & Genton, M. G. (2016). Likelihood estimators for multivariate extremes. *Extremes*, *19*, 79–103.

Huser, R., & Genton, M. G. (2016). Non-stationary dependence structures for spatial extremes. *Journal of Agricultural, Biological and Environmental Statistics*, *21*, 470–491.

Kabluchko, Z., Schlather, M., & de Haan, L. (2009). Stationary max-stable fields associated to negative definite functions. *Annals of Probability*, *37*, 2042–2065.

Nielsen, S. F. (2000). The stochastic EM algorithm: Estimation and asymptotic results. *Bernoulli*, *6*, 457–489.

Opitz, T. (2013). Extremal *t* processes: Elliptical domain of attraction and a spectral representation. *Journal of Multivariate Analysis*, *122*, 409–413.

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

Reich, B. J., & Shaby, B. A. (2012). A hierarchical max-stable spatial model for extreme precipitation. *Annals of Applied Statistics*, *6*, 1430–1451.

Ribatet, M. (2013). Spatial extremes: Max-stable processes at work. *Journal de la Société Française de Statistique*, *154*, 156–177.

Ribatet, M., Cooley, D. S., & Davison, A. C. (2012). Bayesian inference from composite likelihoods, with an application to spatial extremes. *Statistica Sinica*, *22*, 813–845.

Sang, H., & Genton, M. G. (2014). Tapered composite likelihood for spatial max-stable models. *Spatial Statistics*, *8*, 86–103.

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

Shi, D. (1995). Fisher information for a multivariate extreme value distribution. *Biometrika*, *82*, 644–649.

Smith, R. L. (1990). Max-stable processes and spatial extremes. Unpublished.

Stephenson, A., & Tawn, J. A. (2005). Exploiting occurrence times in likelihood inference for componentwise maxima. *Biometrika*, *92*, 213–227.

Thibaud, E., Aalto, J., Cooley, D. S., Davison, A. C., & Heikkinen, J. (2016). Bayesian inference for the Brown–Resnick process, with an application to extreme low temperatures. *Annals of Applied Statistics*, *10*, 2303–2324.

Thibaud, E., & Opitz, T. (2015). Efficient inference and simulation for elliptical Pareto processes. *Biometrika*, *102*, 855–870.

Varin, C., Reid, N., & Firth, D. (2011). An overview of composite likelihood methods. *Statistica Sinica*, *21*, 5–42.

Vettori, S., Huser, R., & Genton, M. G. (2018). A comparison of dependence function estimators in multivariate extremes. *Statistics and Computing*, *28*, 525–538.

Wadsworth, J. L. (2015). On the occurrence times of componentwise maxima and bias in likelihood inference for multivariate max-stable distributions. *Biometrika*, *102*, 705–711.

Wadsworth, J. L., & Tawn, J. A. (2014). Efficient inference for spatial extreme value processes associated to log-Gaussian random functions. *Biometrika*, *101*, 1–15.

Xu, G., & Genton, M. G. (2016). Tukey max-stable processes for spatial extremes. *Spatial Statistics*, *18*, 431–443.

How to cite this article: Huser R, Dombry C, Ribatet M, Genton MG. Full likelihood inference for max-stable data. *Stat*. 2019;**8**:1–14. https://doi.org/10.1002/sta4.218

APPENDIX A: LIKELIHOOD DERIVATION VIA POISSON POINT PROCESS INTENSITY

A.1 | General formulation for the full likelihood g_{Full} and Stephenson–Tawn likelihood g_{ST}

In their original paper, Stephenson and Tawn (2005) derived the likelihood functions g_{Full} and g_{ST} by differentiating the cumulative distribution function in Equation 4. Here, we propose a different approach based on the analysis of the Poisson point process representation (3) of the max-stable process. By introducing the functions $\varphi_j = W_j/P_j, j = 1, 2, \dots$, the point process $\Phi = \{\varphi_j, j \geq 1\}$ is a Poisson point process on the space of nonnegative functions defined on $\mathcal{S}$. The max-stable process Z appears as the pointwise maximum of the functions in Φ. Dombry and Éyi-Minko (2013) showed that for all sites $\boldsymbol{s} \in \mathcal{S}$, there almost surely exists a unique function in Φ that reaches the maximum $Z(\boldsymbol{s})$ at $\boldsymbol{s}$. This function is called the extremal function at $\boldsymbol{s}$ and denoted by φ_s^+. Clearly, $Z(\boldsymbol{s}) = \varphi_s^+(\boldsymbol{s})$.

Given D sites $\boldsymbol{s}_1, \dots, \boldsymbol{s}_D \in \mathcal{S}$, there can be repetitions within the extremal functions $\varphi_{s_1}^+, \dots, \varphi_{s_D}^+$, meaning that the maximum at different sites $\boldsymbol{s}_{j_1}, \boldsymbol{s}_{j_2}$, can arise from the same extremal event. The notion of hitting scenario accounts for such possible repetitions. It is defined as the random partition $\pi = \{\tau_1, \dots, \tau_{|\pi|}\}$ (of size $|\pi|$) of $\{1, \dots, D\}$ such that the two indices j_1 and j_2 are in the same block if and only if the extremal functions at $\boldsymbol{s}_{j_1}$ and $\boldsymbol{s}_{j_2}$ are equal. Here, $|\pi|$ denotes the number of blocks of the partition π and is equal to the number of different functions in Φ reaching the maximum $Z(\boldsymbol{s})$ for some point $\boldsymbol{s} \in \{\boldsymbol{s}_1, \dots, \boldsymbol{s}_D\}$. Within the block $\tau_i \in \pi$, all the points $\boldsymbol{s}_j, j \in \tau_i$, share the same extremal function that will hence be denoted by $\varphi_{\tau_i}^+$.

The joint distribution of the hitting scenario $\pi = \{\tau_1, \dots, \tau_{|\pi|}\}$ and extremal functions $\{\varphi_{\tau_1}^+, \dots, \varphi_{\tau_{|\pi|}}^+\}$ was derived by Dombry and Éyi-Minko (2013). The max-stable observations $Z(\boldsymbol{s}_1), \dots, Z(\boldsymbol{s}_D)$ relate to the hitting scenario and extremal functions via the simple equation $Z(\boldsymbol{s}_j) = \varphi_{\tau_i}^+(\boldsymbol{s}_j)$ for $j \in \tau_i$. In this way, we can deduce the joint distribution of the partition $\pi = \{\tau_1, \dots, \tau_{|\pi|}\}$ and max-stable observations $Z(\boldsymbol{s}_1), \dots, Z(\boldsymbol{s}_D)$, that is, the Stephenson–Tawn likelihood g_{ST}. Marginalizing out the random partition, we deduce the full likelihood g_{Full}.

Suppose that the random vectors $\boldsymbol{W}_j = \{W_j(\boldsymbol{s}_1), \dots, W_j(\boldsymbol{s}_D)\}^\top, j \geq 1$, stemming from Equation 3, have a density $f_{\boldsymbol{W}}$ with respect to the Lebesgue measure on $(0, +\infty)^D$. Then, the Poisson point process $\{\boldsymbol{\varphi}_j, j \geq 1\}$ on $(0, +\infty)^D$, where $\boldsymbol{\varphi}_j = \{\varphi_j(\boldsymbol{s}_1), \dots, \varphi_j(\boldsymbol{s}_D)\}^\top$, has intensity

$$\lambda(z_1, \dots, z_D) = \int_0^\infty f_{\boldsymbol{W}}(z_1/r, \dots, z_D/r) r^{-2-D} \mathrm{d}r. \tag{A1}$$

For clarity, we introduce some vectorial notation: Let $\boldsymbol{Z} = \{Z(\boldsymbol{s}_1), \ldots, Z(\boldsymbol{s}_D)\}^\top$, $\boldsymbol{z} = (z_1, \ldots, z_D)^\top$, $\boldsymbol{z}_i = (z_{1i}, \ldots, z_{Di})^\top$, and $\boldsymbol{\varphi}^+_{\tau_i} = \{\varphi^+_{\tau_i}(\boldsymbol{s}_1), \ldots, \varphi^+_{\tau_i}(\boldsymbol{s}_D)\}^\top$, $i = 1, \ldots, |\pi|$. For $\tau_i \subset \{1, \ldots, D\}$, τ_i^c denotes the complementary subset, and $\boldsymbol{z}_{\tau_i}$ and $\boldsymbol{z}_{\tau_i^c}$ are the subvectors of $\boldsymbol{z}$ obtained by keeping only the components from τ_i and τ_i^c, respectively. Proposition 3 in Dombry and Éyi-Minko (2013) yields the following results:

- From the Poisson point process property, one can deduce the joint law of the hitting scenario and extremal functions:
$$\mathrm{pr}\{\pi = \{\tau_1, \ldots, \tau_{|\pi|}\}, \boldsymbol{\varphi}^+_{\tau_1} = \mathrm{d}\boldsymbol{z}_1, \ldots, \boldsymbol{\varphi}^+_{\tau_{|\pi|}} = \mathrm{d}\boldsymbol{z}_{|\pi|}\} = \exp\{-V(\max_{i=1}^{|\pi|} \boldsymbol{z}_i)\} \prod_{i=1}^{|\pi|} \lambda(\boldsymbol{z}_i)\mathrm{d}\boldsymbol{z}_i,$$
provided the partition associated with $\boldsymbol{z}_1, \ldots, \boldsymbol{z}_{|\pi|}$ is π; otherwise, this probability equals zero.
- By definition of the extremal functions, one gets the joint law of the hitting scenario and max-stable observations:
$$\mathrm{pr}\{\pi = \{\tau_1, \ldots, \tau_{|\pi|}\}, \boldsymbol{Z} = \mathrm{d}\boldsymbol{z}\} = \exp\{-V(\boldsymbol{z})\} \left(\prod_{i=1}^{|\pi|} \int_{\boldsymbol{u}_i < \boldsymbol{z}_{\tau_i^c}} \lambda(\boldsymbol{z}_{\tau_i}, \boldsymbol{u}_i)\mathrm{d}\boldsymbol{u}_i \right) \mathrm{d}\boldsymbol{z}. \tag{A2}$$
- By integrating out the hitting scenario, one obtains the law of the max-stable observations:
$$\mathrm{pr}\{\boldsymbol{Z} = \mathrm{d}\boldsymbol{z}\} = \exp\{-V(\boldsymbol{z})\} \sum_{\pi \in \mathcal{P}_D} \left(\prod_{i=1}^{|\pi|} \int_{\boldsymbol{u}_i < \boldsymbol{z}_{\tau_i^c}} \lambda(\boldsymbol{z}_{\tau_i}, \boldsymbol{u}_i)\mathrm{d}\boldsymbol{u}_i \right) \mathrm{d}\boldsymbol{z}. \tag{A3}$$

Equation A2 provides an alternative formula for the Stephenson–Tawn likelihood, g_{ST}, based on the Poisson point process intensity, λ, whereas Equation A3 is the max-stable full likelihood, g_{Full}. Identifying the expressions (A2) and (A3) above with Equations 6 and 5, respectively, we can see that

$$-V_{\tau_i}(z_1, \ldots, z_D) = \int_{\boldsymbol{u}_i < \boldsymbol{z}_{\tau_i^c}} \lambda(\boldsymbol{z}_{\tau_i}, \boldsymbol{u}_i)\mathrm{d}\boldsymbol{u}_i. \tag{A4}$$

This relates a partial derivative of the exponent function V with a partial integral of the point process intensity λ. In particular, Equation (A4) implies that the intensity is the mixed derivative of the exponent function with respect to all arguments, that is,

$$\lambda(z_1, \ldots, z_D) = -\frac{\partial^D}{\prod_{i=1}^D \partial z_i} V(z_1, \ldots, z_D). \tag{A5}$$

Furthermore, the function V corresponds to the integrated intensity of the set $A = [\boldsymbol{0}, \boldsymbol{z}]^c$, that is,

$$V(z_1, \ldots, z_D) = \Lambda([\boldsymbol{0}, \boldsymbol{z}]^c) = \int_A \lambda(\boldsymbol{u})\mathrm{d}\boldsymbol{u}.$$

A.2 | Computing the Poisson point process intensity

The intensity measure λ is an important feature of max-stable models and can be computed for most popular models; see Dombry et al. (2013) for a derivation of λ for the Brown–Resnick model (Kabluchko et al., 2009) and Ribatet (2013) for an expression of λ for the extremal-t model (Opitz, 2013). Partial integrals of λ for these models may be found in Wadsworth and Tawn (2014) and Thibaud and Opitz (2015), respectively. With the use of the relations (A4) and (A5), the intensity λ and its partial integrals can be deduced for the Reich and Shaby (2012) model from the expressions in the appendix of Castruccio et al. (2016).

Here, as a simple pedagogical illustration for many other multivariate or spatial max-stable models, we consider the multivariate logistic model, which we used in our simulation study. In this case, the function V and its partial and full derivatives can be readily obtained by direct differentiation.

Recall that the exponent function for the logistic model is

$$V(z_1, \ldots, z_D|\theta) = \left(z_1^{-1/\theta} + \ldots + z_D^{-1/\theta}\right)^\theta, \quad \theta \in (0, 1].$$

It is known that the multivariate counterpart of the spectral representation (3) for the logistic model is obtained by taking $\boldsymbol{W} = (W_1, \ldots, W_D)^\top$ with independent and identically distributed Fréchet(β, c_β) components, where $\beta = 1/\theta$ and $c_\beta = 1/\Gamma(1 - 1/\beta)$ are shape and scale parameters, respectively; see, for example, Proposition 6 in Dombry et al. (2016). Then,

$$f_W(z_1, \ldots, z_D) = \prod_{i=1}^D \frac{\beta}{c_\beta} \left(\frac{z_i}{c_\beta}\right)^{-1-\beta} e^{-(z_i/c_\beta)^{-\beta}},$$

and we deduce from Equation (A1) that

$$\lambda(z_1, \ldots, z_D) = \int_0^\infty \left[\prod_{i=1}^D \frac{\beta}{c_\beta} \left(\frac{z_i}{rc_\beta}\right)^{-1-\beta} e^{-\{z_i/(rc_\beta)\}^{-\beta}} \right] r^{-2-D}\mathrm{d}r = \frac{\Gamma(D - 1/\beta)}{\beta} \left\{ \sum_{i=1}^D (z_i/c_\beta)^{-\beta} \right\}^{1/\beta - D} \prod_{i=1}^D \frac{\beta}{c_\beta} \left(\frac{z_i}{c_\beta}\right)^{-1-\beta}.$$

Similar computations entail

$$\int_{\mathbf{u}_i<\mathbf{z}_{\tau_i^c}} \lambda(\mathbf{z}_{\tau_i},\mathbf{u}_i)\,\mathrm{d}\mathbf{u}_i = \int_0^\infty \left[\prod_{j\in\tau_i}\frac{\beta}{c_\beta}\left(\frac{z_j}{rc_\beta}\right)^{-1-\beta} e^{-\{z_j/(rc_\beta)\}^{-\beta}}\right]\times\left[\prod_{j\in\tau_i^c} e^{-\{z_j/(rc_\beta)\}^{-\beta}}\right] r^{-2-|\tau_i|}\,\mathrm{d}r$$

$$= \left\{\prod_{j\in\tau_i}\frac{\beta}{c_\beta}\left(\frac{z_j}{c_\beta}\right)^{-1-\beta}\right\}\int_0^\infty e^{-\sum_{j=1}^{D}\{z_j/(rc_\beta)\}^{-\beta}} r^{\beta|\tau_i|-2}\,\mathrm{d}r$$

$$= \frac{\Gamma(|\tau_i|-1/\beta)}{\beta}\left\{\sum_{j=1}^{D}(z_j/c_\beta)^{-\beta}\right\}^{1/\beta-|\tau_i|}\prod_{j\in\tau_i}\frac{\beta}{c_\beta}\left(\frac{z_j}{c_\beta}\right)^{-1-\beta}$$

$$= \beta^{|\tau_i|-1}\frac{\Gamma(|\tau_i|-1/\beta)}{\Gamma(1-1/\beta)}\left(\sum_{j=1}^{D} z_j^{-\beta}\right)^{1/\beta-|\tau_i|}\prod_{i\in\tau_i} z_j^{-1-\beta},$$

where, for the first equality, we used

$$\prod_{j\in\tau_i^c}\int_0^{z_j}\frac{\beta}{c_\beta}\left(\frac{u_j}{rc_\beta}\right)^{-1-\beta} e^{-\{u_j/(rc_\beta)\}^{-\beta}}\,\mathrm{d}u_j = \prod_{i\in\tau_i^c} re^{-\{z_j/(rc_\beta)\}^{-\beta}}.$$

APPENDIX B: DETAILS ON THE UNDERLYING GIBBS SAMPLER

The Gibbs sampler proposed by Dombry et al. (2013) is designed to draw an ergodic sequence of partitions $\pi_1, \dots, \pi_N$, whose limiting stationary distribution is the distribution of the partition π conditional on the observed max-stable data $\mathbf{z} = (z_1, \dots, z_D)^\top$, that is, the discrete distribution $g(\pi|\mathbf{z},\boldsymbol{\theta})$, where $\boldsymbol{\theta}\in\Theta\subset\mathbb{R}^p$ is the parameter vector characterizing the max-stable dependence structure. One has

$$g(\pi|\mathbf{z},\boldsymbol{\theta}) = \frac{g_{\mathrm{ST}}(\pi,\mathbf{z}|\boldsymbol{\theta})}{g_{\mathrm{Full}}(\mathbf{z}|\boldsymbol{\theta})} = \frac{\exp\{-V(\mathbf{z}|\boldsymbol{\theta})\}\prod_{i=1}^{|\pi|}\{-V_{\tau_i}(\mathbf{z}|\boldsymbol{\theta})\}}{\exp\{-V(\mathbf{z}|\boldsymbol{\theta})\}\sum_{\pi\in\mathcal{P}_D}\prod_{i=1}^{|\pi|}\{-V_{\tau_i}(\mathbf{z}|\boldsymbol{\theta})\}}$$

$$= \frac{\prod_{i=1}^{|\pi|}\{-V_{\tau_i}(\mathbf{z}|\boldsymbol{\theta})\}}{\sum_{\pi\in\mathcal{P}_D}\prod_{i=1}^{|\pi|}\{-V_{\tau_i}(\mathbf{z}|\boldsymbol{\theta})\}} \propto \prod_{i=1}^{|\pi|}\{-V_{\tau_i}(\mathbf{z}|\boldsymbol{\theta})\}. \tag{B1}$$

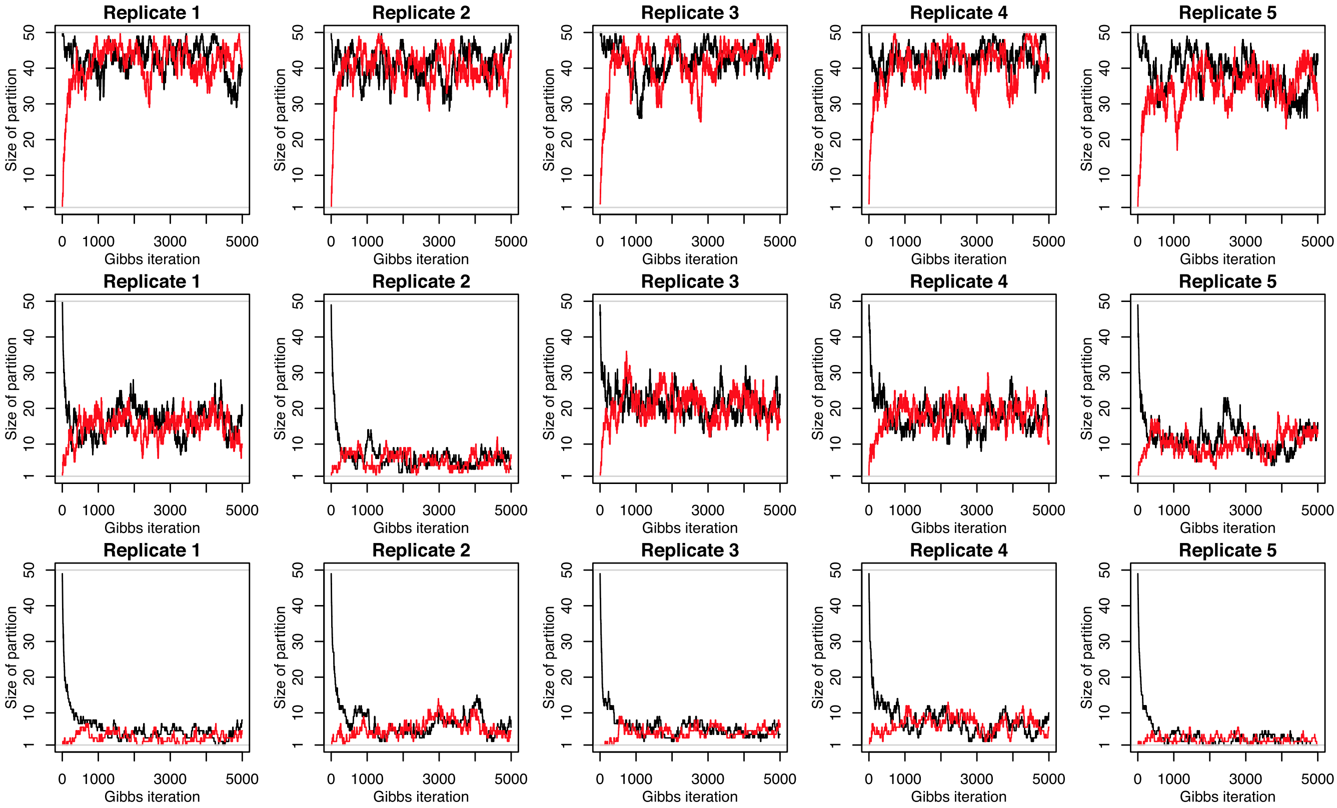

FIGURE B1 Trace plots of the sizes of partitions obtained from the Gibbs samplers for each of the five replicates (columns). We considered the logistic model in dimension $D = 50$ with parameter $\theta = 0.9, 0.6, 0.3$ (top to bottom rows). Initial partitions were taken as $\{\{1\}, \dots, \{D\}\}$ (black) and $\{\{1, \dots, D\}\}$ (red); 5,000 iterations were performed

The normalizing constant in the denominator of Equation B1 is computationally demanding to compute as it involves the sum over all partitions. Nevertheless, the Gibbs sampler of Dombry et al. (2013) provides a way to construct a Markov chain whose stationary distribution is $g(\pi|\boldsymbol{z},\boldsymbol{\theta})$, while avoiding the computation of the normalizing constant. Let $\pi_r = \{\tau_{r;1}, \dots, \tau_{r;|\pi_r|}\} \in \mathcal{P}_D$ be the partition at the rth iteration of the Gibbs sampler. The idea of the Gibbs sampler is to sample the next partition $\pi_{r+1} = \{\tau_{r+1;1}, \dots, \tau_{r+1;|\pi_{r+1}|}\} \in \mathcal{P}_D$ by keeping all but one components fixed. Let $\ell \in \{1, \dots, D\}$ be the component to be updated, and let $\pi_r^{-\ell}$ and $\pi_{r+1}^{-\ell}$ denote the partitions π_r and π_{r+1}, respectively, with the ℓth component removed. We update the partition π_r by modifying the (randomly chosen) ℓth component using the full conditional distribution

$$g(\pi_{r+1}|\pi_{r+1}^{-\ell} = \pi_r^{-\ell}, \boldsymbol{z}, \boldsymbol{\theta}) \propto \frac{\prod_{i=1}^{|\pi_{r+1}|}\{-V_{\tau_{r+1;i}}(\boldsymbol{z}|\boldsymbol{\theta})\}}{\prod_{i=1}^{|\pi_r|}\{-V_{\tau_{r;i}}(\boldsymbol{z}|\boldsymbol{\theta})\}}. \tag{B2}$$

The combinatorial explosion is avoided, because the number of possible updates π_{r+1} such that $\pi_{r+1}^{-\ell} = \pi_r^{-\ell}$ is at most $|\pi_r| + 1$. Moreover, as we update only one component at a time, many terms in the ratio (B2) cancel out, and at most four of them need to be computed, which makes it computationally feasible. However, for the same reason, the resulting partitions will also be heavily dependent, and so, intuitively, we should take the number of Gibbs sampler iterations to be roughly proportional to the dimension D and thin the Markov chain accordingly by a factor D to get approximately independent (or weakly dependent) partitions. A suitable burn-in should also be specified to ensure that the Markov chain has appropriately converged to its stationary distribution.

In order to assess the number of iterations required for the Gibbs sampler to converge (i.e., the burn-in), we considered the logistic model defined by its exponent function $V(z_1, \dots, z_D|\theta) = (z_1^{-1/\theta} + \dots + z_D^{-1/\theta})^{\theta}$, $\theta \in (0, 1]$. We generated five independent copies of a logistic random vector in dimension $D = 50$, and we considered the cases $\theta = 0.9, 0.6, 0.3$ (weak to strong dependence). For each dataset, we ran five Gibbs samplers (one per independent replicate) for 5,000 iterations. To easily visualize the resulting Markov chains and assess convergence, we display in Figure B1 trace plots of the sizes of sampled partitions along the different Markov chains. The initial partitions were taken as $\{\{1\}, \dots, \{D\}\}$ (of size $D = 50$), which reflects weak dependence scenarios, and $\{\{1, \dots, D\}\}$ (of size one), which reflects strong dependence scenarios.

In all cases, we can see that the Gibbs sampler converges rather quickly and that it is enough to discard a burn-in of about $10 \times D = 500$ iterations.

To validate such results for another max-stable model, we did the same experiment for the Brown–Resnick model (Kabluchko et al., 2009) with semivariogram $\gamma(\boldsymbol{h}) = (\|\boldsymbol{h}\|/\lambda)^{\nu}$, where $\lambda > 0$ and $\nu \in (0, 2]$ are the range and smoothness parameters, respectively, at $D = 10$ randomly generated sites $\boldsymbol{s}_1, \dots, \boldsymbol{s}_{10} \in [0, 1]^2$. We considered the cases $\lambda = 0.5, 1, 1.5$ (short- to long-range dependence) with $\nu = 1.5$. For each dataset,

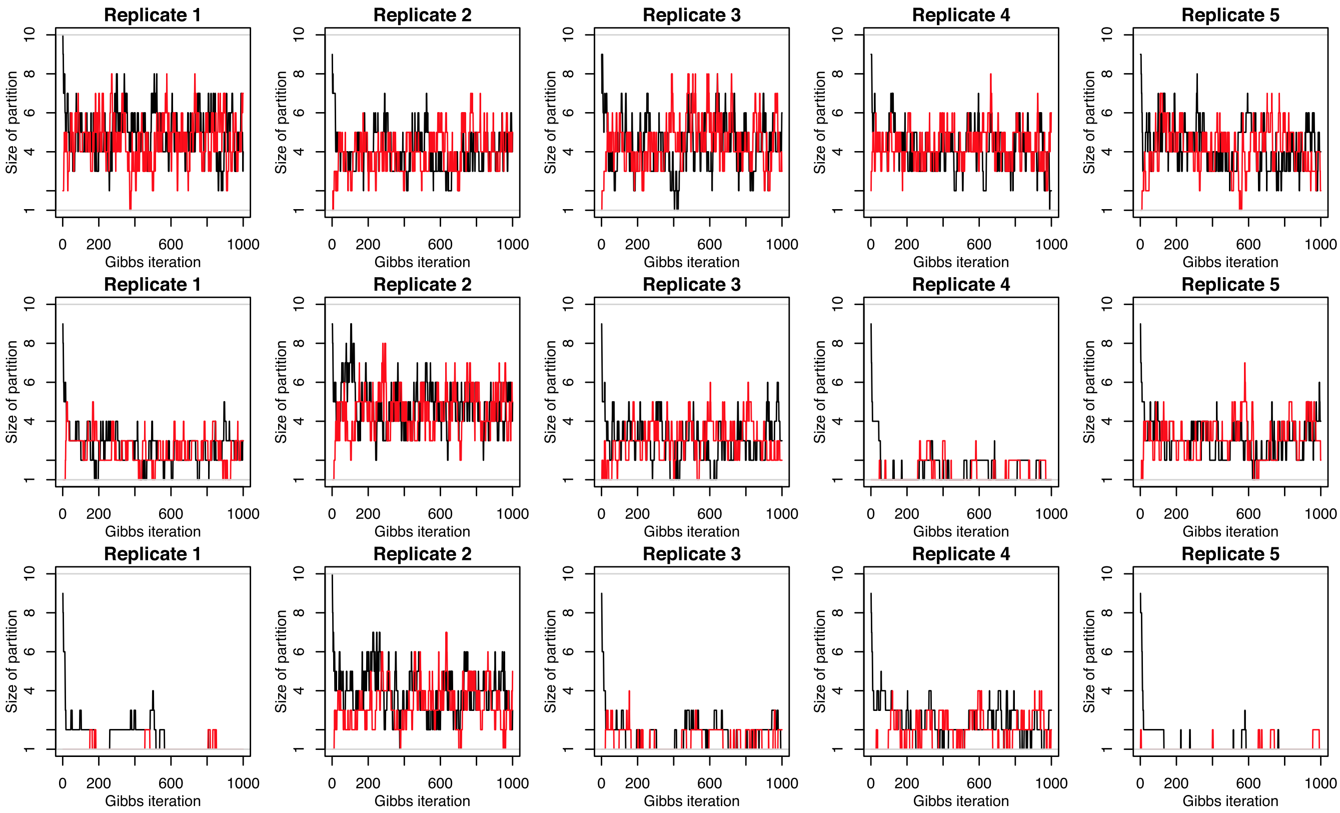

FIGURE B2 Trace plots of the sizes of partitions obtained from the Gibbs samplers for each of the five replicates (columns). We considered the Brown–Resnick model at $D = 10$ sites in $[0, 1]^2$ with semivariogram $\gamma(\boldsymbol{h}) = (\|\boldsymbol{h}\|/\lambda)^{\nu}$ and parameters $\lambda = 0.5, 1, 1.5$ (top to bottom rows) with $\nu = 1.5$. Initial partitions were taken as $\{\{1\}, \dots, \{D\}\}$ (black) and $\{\{1, \dots, D\}\}$ (red); 1,000 iterations were performed

we ran five Gibbs samplers (one per independent replicate) for 1,000 iterations. Figure B2 shows the trace plots of the sizes of sampled partitions along the different Markov chains. As before, the initial partitions were taken as $\{\{1\}, \dots, \{D\}\}$ (of size $D = 10$) and $\{\{1, \dots, D\}\}$ (of size one).

As concluded for the logistic model, we can see that the Gibbs sampler converges quickly and that about $10 \times D = 100$ iterations are enough for the algorithm to converge in all cases.

These results suggest to discard the first $10 \times D$ iterations as burn-in and to thin the resulting Markov chains by a factor D to obtain approximately (conditionally) independent partitions. With this setting, the initial partition has negligible impact on the results. Furthermore, another natural option could be to initialize the partition randomly from its unconditional distribution, which can be easily obtained from an unconditional simulation of the max-stable distribution. This could potentially provide further computational savings by reducing the time it takes for the Gibbs sampler to converge (thus reducing the burn-in).