



Truncated pair-wise likelihood for the Brown-Resnick process with applications to maximum temperature data

Zhendong Huang¹ · Olga Shulyarenko¹ · Davide Ferrari²

Received: 24 May 2018 / Revised: 7 October 2019 / Accepted: 13 October 2019 /

Published online: 1 June 2021

© The Author(s) 2021

Abstract

Max-stable processes are a natural extension of multivariate extreme value theory important for modeling the spatial dependence of environmental extremes. Inference for max-stable processes observed at several spatial locations is challenging due to the intractability of the full likelihood function. Composite likelihood methods avoid these difficulties by combining a number of low-dimensional likelihood objects, typically defined on pairs or triplets of spatial locations. This work develops a new truncation procedure based on ℓ_1 -penalty to reduce the complexity and computational cost associated with the composite likelihood function. The new method is shown to achieve a favorable trade-off between computational burden and statistical efficiency by dropping a number of noisy sub-likelihoods. The properties of the new method are illustrated through numerical simulations and an application to real extreme temperature data.

Keywords Max-stable processes · Extreme temperature · Composite likelihood inference · Likelihood truncation

AMS 2000 Subject Classifications 62P12 · 62M30 · 62F10

1 Introduction

Weather and climate extremes are well-known for their environmental, social and economic impact, with heat waves, droughts, floods, and hurricanes being common examples. The widespread use of geo-referenced data together with the need to

✉ Davide Ferrari
davferrari@unibz.it

¹ School of Mathematics and Statistics, University of Melbourne, Victoria, Australia

² School of Economics and Management, Free University of Bozen-Bolzano, Bolzano, Italy

monitor extreme events have motivated a growing interest in statistical methods for spatial extremes. On the other hand, the availability of accurate inference methods able to estimate accurately the severity of heat extremes is important to understand, prepare for and adapt to future environment changes. This work is motivated by the analysis of extreme temperature data recorded in the state of Victoria, Australia, by Bureau of Meteorology (BoM) (<http://www.bom.gov.au/climate/data>), the national meteorological service of Australia monitoring local climate, including extreme weather events.

An important class of statistical models for spatial extremes are the so-called max-stable processes, which provide a theoretically-justified description of extreme events measured at several spatial locations. Smith (1990) proposed an easy-to-interpret max-stable model based on storm profiles. Despite its widespread use, the Smith model is often criticized for its lack of realism due to excessive smoothness. A more useful max-stable process is the Brown-Resnick process, a generalization of the Smith model able to describe a wider range of extreme dependence regimes (Brown and Resnick 1977; Kabluchko et al. 2009). Reviews of max-stable models and inference are given by Davison et al. (2012) and Davison and Huser (2015).

Inference for max-stable models is generally difficult due to the computational intractability of the full likelihood function. These challenges have motivated the use of composite likelihood (CL) methods, which avoid dealing with intractable full likelihoods by taking a linear combination of low-dimensional likelihood score functions (Lindsay 1988; Varin et al. 2011). Various composite likelihood designs have been studied for max-stable models. Davison et al. (2012) and Huser and Davison (2013) consider pair-wise likelihood estimation based on marginal likelihoods defined on pairs of sites. In the context of the Smith model, Genton et al. (2011) show that estimation based on triple-wise likelihoods (i.e. combining sublikelihoods defined on three sites) is more efficient compared to pair-wise likelihood. For the more realistic Brown-Resnick model, however, Huser and Davison (2013) show that the efficiency gains from using triple-wise likelihood are modest.

The choice of the linear coefficients combining the partial log-likelihood score objects has important repercussions on both efficiency and computation for the final estimator. Cox and Reid (2004) discuss the substantial loss of efficiency for pair-wise likelihood estimators when a large number of correlated scores is included. In the context of max-stable processes, various works aimed at improving efficiency and computing based on the idea that sub-likelihoods defined on nearby locations are generally more informative about dependence parameters than those for distant locations. Sang and Genton (2014) consider a weighting strategy for sub-likelihoods based on tapering to exclude distant pairs or triples and improve statistical efficiency. Their method improves efficiency compared to uniform weights, but tuning of the tapering function is computationally intensive. Castruccio et al. (2016) consider combining a number of sub-likelihoods by taking more than three locations at the time, and show the benefits from likelihood truncation obtained by retaining partial likelihood pairs for nearby locations. In a different direction, other studies have focused on direct approximation of the full likelihood. Huser et al. (2019) consider full-likelihood based inference through a stochastic Expectation-Maximisation algorithm. Thibaud et al. (2016) considers Bayesian approach where the full-likelihood is constructed

by considering a partition of the data based on occurrence times of maxima within blocks. Although the current full likelihood approaches do not directly require the computation of the full likelihood as a sum over all partitions, their application is still hindered by issues related to computational efficiency when the number of measuring sites is large. On the other hand, composite likelihood methods offer considerable computational advantages compared to full likelihood approaches although they may lack of statistical efficiency when too many correlated sub-likelihood objects are considered.

The main contribution of this work is the application of the general composite likelihood truncation methodology of Huang and Ferrari (2017) in the context of max-stable models and pair-wise likelihood for the analysis of extreme temperature data. The new method, referred to as truncated pair-wise likelihood (TPL) hereafter. In the proposed TPL estimation, a data-driven combination of pair-wise log-likelihood objects is obtained by optimizing statistical efficiency, subject to a ℓ_1 -penalty discouraging the inclusion of too many terms in the final estimating equations. Whilst the basic method of Huang and Ferrari (2017) had a single linear coefficient for each sub-likelihood object, here we extend that approach by allowing parameter-specific coefficients within each pair-wise likelihood score. This generalization is shown to improve stability of the truncated estimating equations and the statistical accuracy of the final estimator. The proposed ℓ_1 -penalty enables us to retain only informative sub-likelihood objects corresponding to nearby pairs. This reduces the final computational cost and yields estimators with considerable efficiency compared to pair-wise likelihood estimator with equal coefficients commonly adopted in the spatial extremes literature.

The rest of the paper is organized as follows. In Section 2, we review max-stable processes and the Brown-Resnick model. In Section 3, we describe the main methodology for likelihood truncation and parameter estimation within the pair-wise likelihood estimation framework. In Section 4, we carry out Monte Carlo simulations to illustrate the properties of the method and compare it with other pair-wise likelihood strategies in terms of computational burden and statistical efficiency. In Section 5, we apply the method to analyze extreme temperature data recorded in the state of Victoria, Australia. In Section 6, we conclude and give final remarks.

2 Brown-Resnick process for spatial extremes

2.1 The Brown-Resnick process

Following Huser and Davison (2013), the Brown-Resnick process (Brown and Resnick 1977; Kabluchko et al. 2009) can be defined as the stationary max-stable process with spectral representation given by $Z(x) = \sup_{i \in \mathbb{N}} W_i(x)/T_i$, $x \in \mathcal{X} \subseteq \mathbb{R}^2$, where $0 < T_1 < T_2 < \dots$ are points of a Poisson process on \mathbb{R}^+ , $W_1(x), W_2(x), \dots$ are independent replicates of the random process $W(x) = \exp\{\varepsilon(x) - \gamma(x)\}$, $x \in \mathcal{X}$, $\varepsilon(x)$ represents a Gaussian process with stationary increments such that $\varepsilon(0) = 0$ almost surely and $\gamma(h)$ is the semi-variogram of

$\varepsilon(x)$ defined by $\gamma(h) = \text{var}\{Z(x) - Z(x+h)\}/2, x, x+h \in \mathcal{X}$. The process $Z(x)$ may be interpreted as the maximum of random storms $W_i(x)$ of size $1/T_i$.

Let s be the total number of locations being considered. The s -dimensional distribution function for the process $\{Z(x), x \in \mathcal{X}\}$ measured at the set of locations $\mathcal{S} \in \mathcal{X}$ can be written as

$$P(Z(x) \leq z(x), x \in \mathcal{S}) = \exp(-V\{z(x)\}), \tag{1}$$

where $V\{z(x)\} = E[\sup_{x \in \mathcal{S}} W(x)/z(x)]$ is the so-called exponent measure function. Different max-stable models are obtained by specifying the exponent measure $V(\cdot)$ through the choice of semi-variogram $\gamma(\cdot)$. For example, the Brown-Resnick model can be specified by the parametric variogram model with $\gamma(h; \theta) = (\|h\|/\rho)^\alpha$ and $\theta = (\alpha, \rho)'$, where $\rho > 0$ and $0 < \alpha \leq 2$ are the range and the smoothness parameters, respectively. When $\alpha = 2$ the Brown-Resnick process has maximum smoothness with semi-variogram $\gamma(h) = h'\Sigma h$ for some covariance matrix Σ . In this case, the Brown-Resnick process is equal to the Smith process (Kablichko et al. 2009; Padoan et al. 2010). Figure 1 shows semi-variograms for different parameter values (top row) with realizations of the corresponding Brown-Resnick process at single site (bottom row). The variogram increases as ρ decreases, whilst its shape can be convex ($\alpha > 1$), linear ($\alpha = 1$), or concave ($\alpha < 1$).

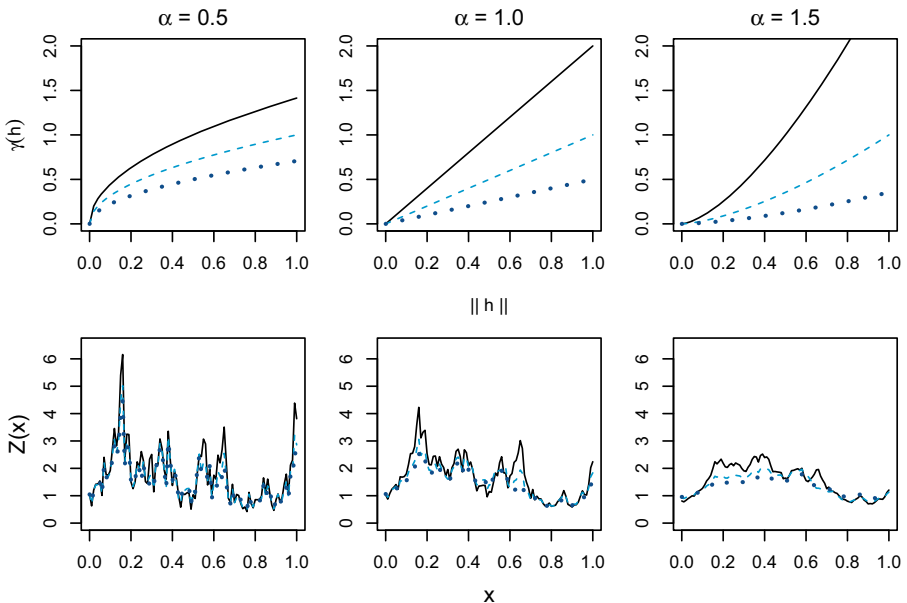


Fig. 1 Top row: Variogram $\gamma(h)$ for the Brown-Resnick process $Z(x)$ for different specifications of smoothness (α) and range (ρ) parameters. Solid, dashed and dotted curves in each plot correspond to range parameter $\rho = 0.5, 1.0$ and 1.5 , respectively. Bottom row: each plot shows individual realizations of the Brown-Resnick processes at a single site with parameters corresponding to the variograms in the top row

2.2 Marginal pair-wise density functions

Let $z = (z_1, \dots, z_s)$ be a sample with $z_j = Z(x_j)$ denoting the realization at site j and $\mathcal{S} = \{x_1, \dots, x_s\} \in \mathcal{X}$. The density of $\{Z(x_1), \dots, Z(x_s)\}$ can be written as

$$f(z_1, \dots, z_s) = \exp \left\{ -V(z_1, \dots, z_s) \right\} \sum_{\xi \in \mathcal{P}_s} v(\xi), \tag{2}$$

where \mathcal{P}_s denotes the set of all possible partitions of the set $\{x_1, \dots, x_s\}$, $\xi = (\xi_1, \dots, \xi_l)$, $|\xi| = l$ is the size of the partition ξ , and

$$v(\xi) = (-1)^{|\xi|} \prod_{j=1}^{|\xi|} \frac{d^{|\xi_j|}}{dz_{\xi_j}} V(z_1, \dots, z_s), \tag{3}$$

where $d^{|\xi_j|}/dz_{\xi_j}$ denotes the mixed partial derivatives with respect to the ξ_j element of z (Padoan et al. 2010). Since the the cardinality of \mathcal{P}_s increases quickly with the number of sites s , the density and the full likelihood functions are unavailable for arbitrary number of sites s due to the storage and computation of an exponentially increasing number of derivatives.

Although the full density cannot be computed unless s is trivially small, low-dimensional densities are readily available. The bivariate exponent measure for the Brown-Resnick process for $\mathcal{S} = \{x_j, x_k\}$ is

$$V(z_j, z_k; \theta) = \frac{1}{z_j} \Phi \left\{ \frac{a(\theta)}{2} - \frac{1}{a(\theta)} \log \left(\frac{z_j}{z_k} \right) \right\} + \frac{1}{z_k} \Phi \left\{ \frac{a(\theta)}{2} - \frac{1}{a(\theta)} \log \left(\frac{z_k}{z_j} \right) \right\}, \tag{4}$$

where $a(\theta) = \sqrt{2\gamma(x_j - x_k; \theta)}$ and $\Phi(\cdot)$ is the standard normal distribution function. Let $m = s(s - 1)/2$ be the total number of pairs (z_j, z_k) , $1 \leq j < k \leq s$, obtained from elements of z . Let $r = 1, \dots, m$ be the subscript corresponding to a site pair $\{(j, k) : 1 \leq j < k \leq s\}$. The bivariate density function $f_r(z_j, z_k; \theta)$ is obtained by direct differentiation as

$$f_r(z_j, z_k; \theta) = \frac{d^2}{dz_j dz_k} \exp\{-V_r\} = (\dot{V}_j \dot{V}_k - \ddot{V}_{jk}) \exp\{-V_r\}, \tag{5}$$

where $V_r = V(z_j, z_k; \theta)$ and $\dot{V}_j, \dot{V}_k, \ddot{V}_{jk}$ are the corresponding partial derivatives $dV_r/dz_j, dV_r/dz_k, d^2V_r/(dz_j dz_k)$ given in the [Appendix](#).

2.3 Extremal coefficient

The dependence structure for the elements in the random vector $\{Z(x_1), \dots, Z(x_s)\}$ is completely determined by the exponent function V defined in (1). Particularly, the exponent measure $V(\cdot)$ is a positive homogeneous function of order -1 , i.e. $V(cz_1, \dots, cz_s) = c^{-1}V(z_1, \dots, z_s)$. Moreover, for all $c > 0$, we have $V(\infty, \dots, z_j, \dots, \infty) = z_j^{-1}$, $j = 1, \dots, s$, meaning that the marginal distributions at each site are unit Fréchet (Davison et al. 2012). The dependence regime falls between two limit cases: $V(z_1, \dots, z_s) = 1/z_1 + \dots + 1/z_s$ (perfect independence)

and $V(z_1, \dots, z_s) = 1 / \min(z_1, \dots, z_s)$ (complete dependence). Note that for $z > 0$, we have

$$P(Z(x_1) \leq z, \dots, Z(x_s) \leq z) = \exp\{-V(z, \dots, z)\} = \exp\{-z^{-1}V(1, \dots, 1)\},$$

and the quantity $\eta_s = V(1, \dots, 1)$ is the so-called s -variate extremal coefficient (Smith 1990; Schlather and Tawn 2003). Although η_s does not completely define the dependence structure of $\{Z(x_1), \dots, Z(x_s)\}$, it provides a useful summary for extreme dependence. Specifically, the s -variate extremal coefficient satisfies $1 \leq \eta_s \leq s$, with the two extreme cases $\eta_s = 1$ and $\eta_s = s$ representing perfect dependence and independence, respectively.

For the case of two sites ($s = 2$), the bivariate extremal coefficient is $\eta_2(h) = V(1, 1)$, with value depending on the Euclidean distance $\|h\|$ between two locations. For the Brown-Resnick model, we have $\eta_2(h; \theta) = V(1, 1; \theta) = 2\Phi\left\{\sqrt{2\gamma(h; \theta)}/2\right\}$ where $\gamma(h; \theta)$ is the Brown-Resnick semi-variogram and Φ is the standard normal distribution function. With this variogram specification, $\eta_2(0; \theta) = 1$ if $\|h\| = 0$, while $\eta_2(h; \theta) \rightarrow 2$ as $\|h\| \rightarrow \infty$. Figure 2 (top row) shows estimated pairwise extremal coefficients based on 40 samples generated from a Brown-Resnick process observed at 30 locations on $[0, 1]^2$. The pairwise extremal coefficients is estimated by $\hat{\eta}_2(h) = (1+2\hat{v}(h))/(1-\hat{v}(h))$, where $\hat{v}(h)$ is the empirical F-madogram proposed by Cooley et al. (2006). Figure 2 (bottom row) shows corresponding realizations of the Brown-Resnick process on $[0, 1]^2$.

3 Truncated pair-wise likelihood by ℓ_1 -penalization

In this section we describe our likelihood truncation approach and related pair-wise inference. For concreteness, we focus on pair-wise inference and the Brown-Resnick model with variogram $\gamma(h) = (\|h\|/\rho)^\alpha$. In principle, the proposed approach may be applied also in the context of composite likelihood designs besides pair-wise likelihood (e.g. triple-wise likelihood) and other max-stable models.

3.1 Pair-wise likelihood estimation

Let $\{z^{(i)}, 1 \leq i \leq n\}$ be independent observations of a Brown-Resnick process at s sites with $z^{(i)} = (z_1^{(i)}, \dots, z_s^{(i)})'$ and let $m = s(s - 1)/2$ be the number of pairs $(z_j^{(i)}, z_k^{(i)})$, $1 \leq j < k \leq s$, obtained from the elements of $z^{(i)}$. Let $u_\alpha^{(i)}(\theta)$ and $u_\rho^{(i)}(\theta)$ be the i th realization of $m \times 1$ pair-wise scores defined by $u_\alpha(\theta) = (u_{\alpha,1}(\theta), \dots, u_{\alpha,m}(\theta))'$ and $u_\rho(\theta) = (u_{\rho,1}(\theta), \dots, u_{\rho,m}(\theta))'$, with elements

$$u_{\alpha,r}(\theta) = \frac{\partial}{\partial \alpha} \log f_r(z_j, z_k; \theta), \quad u_{\rho,r}(\theta) = \frac{\partial}{\partial \rho} \log f_r(z_j, z_k; \theta), \quad (6)$$

$r = 1, \dots, m$, where r is the subscript corresponding to a site pair $\{(j, k) : 1 \leq j < k \leq s\}$. Write $\tilde{u}(\theta) = (u_\alpha(\theta)', u_\rho(\theta)')'$ for the $2m \times 1$ vector collecting all partial scores elements and let $\tilde{u}^{(i)}(\theta)$ be the corresponding i th realization of $\tilde{u}(\theta)$. A

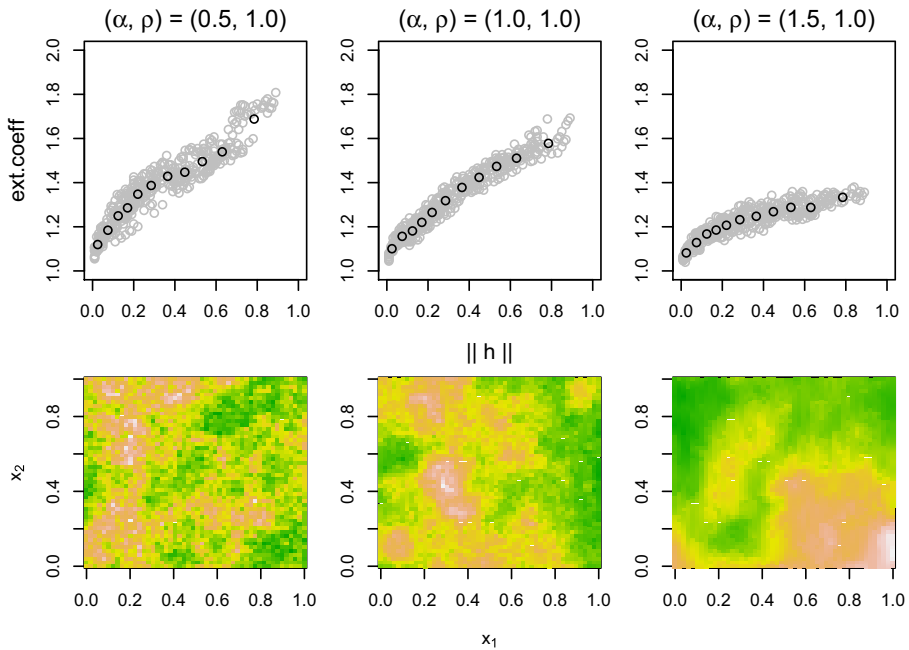


Fig. 2 Top Row: The lighter points are pair-wise extremal coefficient estimates $\hat{\eta}_2(h; \theta)$ for the Brown-Resnick process as a function of the distance $\|h\|$ for smoothness $\alpha = 0.5, 1.0, 1.5$ and range $\rho = 1.0$. Estimates are based on 40 realizations of the Brown-Resnick process at 30 randomly selected spatial locations on $[0, 1]^2$. The darker points are binned averages. Bottom Row: Individual realizations for the Brown-Resnick process on $[0, 1]^2$ with smoothness $\alpha = 0.5, 1$ and 1.5

closed-form expression for the elements of the pair-wise scores $\tilde{u}(\theta)$ for the Brown-Resnick process is reported in the [Appendix](#).

For inference, we consider weighted pair-wise likelihood estimators (PLEs), $\hat{\theta}_w$, found by solving the estimating equations

$$0 = \sum_{i=1}^n u^{(i)}(\theta, w) := \sum_{i=1}^n \mathcal{W}' \tilde{u}^{(i)}(\theta), \tag{7}$$

where \mathcal{W} is the $2m \times 2$ matrix

$$\mathcal{W} = \begin{pmatrix} w_\alpha & \underline{0} \\ \underline{0} & w_\rho \end{pmatrix}, \tag{8}$$

with $w_\alpha = (w_{\alpha,1}, \dots, w_{\alpha,m})'$ and $w_\rho = (w_{\rho,1}, \dots, w_{\rho,m})'$ being the vectors containing specific coefficients for the score components, $\underline{0}$ is a $m \times 1$ vector of zeros, and $u^{(i)}(\theta, w)$ is the i th realization of $u(\theta, w)$ defined as $u(\theta, w) = \mathcal{W}' \tilde{u}(\theta)$. Here $w = (w'_\alpha, w'_\rho)'$ is the $2m \times 1$ vector containing all the coefficients, which we refer to as composition rule in the rest of the paper. A popular choice for w in applications is the vector with uniform elements $w_\alpha = w_\rho = (1, \dots, 1)'$, corresponding to the uniform pair-wise likelihood estimator (UPLE).

The PLE is a popular estimator for max-stable models due to its flexibility and well-known asymptotic properties. Particularly, $\sqrt{n}(\hat{\theta}_w - \theta)$ converges to a bivariate normal distribution with zero mean vector and asymptotic covariance matrix $G_w(\theta)^{-1}$ as n increases, where

$$G_w(\theta) = H_w(\theta)J_w(\theta)^{-1}H_w(\theta) \tag{9}$$

is the so-called Godambe information matrix, and $H_w(\theta) = E[\partial u(\theta, w)/\partial \theta]$ and $J_w(\theta) = \text{Var}[u(\theta, w)]$ are the 2×2 sensitivity and variability matrices.

Although the PLE is consistent, its variance can be much larger than that of the maximum likelihood estimator depending on the choice of the composition rule w . If the composition rule w has all nonzero elements, the matrices H_w and J_w involve $O(s^2)$ and $O(s^4)$ terms, respectively. Thus, when the number of sites s is moderate or large, the presence of many correlated pair-wise scores can inflate J_w and the implied asymptotic variance $G_w(\theta)^{-1}$. From a computational viewpoint, another drawback is that finding the standard errors of the PLE is computationally expensive for large s due to the presence of many terms in J_w . In the following section, we describe a likelihood truncation methodology able to reduce the computational burden while avoiding issues related to variance inflation.

3.2 Truncation by ℓ_1 -norm penalization

To increase statistical performance of pair-wise likelihood estimation for max-stable models while reducing the computing costs, we adopt a new truncation strategy of the estimating equations. The resulting composition rule contains a number of zero elements, which implies simplified pair-wise likelihood equations with less terms. We propose find such a composition rule by minimizing the distance between the unknown full likelihood score, subject to an ℓ_1 -norm penalty representing the likelihood complexity. The procedure may be regarded as to maximize the statistical accuracy for a certain level of afforded computing. Specifically, we aim to solve the PL estimating equations $0 = \sum_{i=1}^n u^{(i)}(\theta, w)$ in (7) with respect to θ with $w = w(\theta)$ found by minimizing with respect to w the ideal criterion

$$Q_\lambda(\theta, w) = \sum_{l \in \{\alpha, \rho\}} \left\{ E \left[u_l^{ML}(\theta) - w'_l u_l(\theta) \right]^2 + \sum_{r=1}^m \lambda_l |w_{l,r}| \right\}, \tag{10}$$

where $\lambda = (\lambda_\alpha, \lambda_\rho)'$ is a 2×1 vector of tuning parameters with non-negative elements and $u_l^{ML}(\theta) = \partial \log f(z_1, \dots, z_s; \theta) / \partial l$, $l \in \{\alpha, \rho\}$ denotes the elements of the unknown maximum likelihood score function. It is worth to keep in mind that the minimizer $w(\theta)$ also depends on the tuning value λ .

The term $E \left[u_l^{ML}(\theta) - w'_l u_l(\theta) \right]^2$ in (10) represents the distance between the pair-wise score and the maximum likelihood score. Thus, the particular case when $\lambda_l = 0$ gives fixed-sample optimality (O_F -optimality) defined as the projection of the the ML score onto the linear space spanned by the partial scores (Heyde 2008). Without additional constraints, however, we have no way to reduce the likelihood complexity since all the pair-wise score terms are in principle included in the final estimating equation. On the other hand, for sufficiently large $\lambda_l > 0$, the penalty $\sum_{r=1}^m \lambda_l |w_{l,r}|$,

$l \in \{\alpha, \rho\}$ implies truncated estimating equations by avoiding the inclusion of noisy terms in the pairwise likelihood score $u(\theta, w)$. This is analogous to ℓ_1 -penalized least-squares approaches for regression (e.g. see Efron et al. (2004)). However, while in regression the penalty involves directly regression coefficients, our penalty does not involve the statistical parameter θ but only the composition rule w .

Due to the geometry of the ℓ_1 -norm penalty, the composition rule $w(\theta)$ minimizing (10) contains an increasing number of zero elements as λ_l grows. Therefore, such a penalty is effectively a constraint on the computing cost (or, equivalently, on the likelihood complexity). This means that the truncated solution $w(\theta)$ can be interpreted as one that maximizes statistical efficiency for a given level of computing. Alternatively, it may be interpreted as one maximizing computational efficiency for a given level of efficiency.

Direct minimization of $Q_\lambda(\theta, w)$ is not useful in practice due to the presence of the intractable likelihood score u_l^{ML} and expectations in (10). To eliminate the explicit dependence on the ML score, note that the expectation in (10) can be written as

$$E \left[u_l^{ML}(\theta) - w'_l u_l(\theta) \right]^2 = w'_l E \left[u_l(\theta) u_l(\theta)' \right] w_l - 2w'_l E \left[u_l^{ML}(\theta) u_l(\theta) \right] + c, \tag{11}$$

where c is a term not depending on w_l . Dependence on the ML score is avoided by replacing the term $E \left[u_l^{ML}(\theta) u_l(\theta) \right]$ in (11) by $\text{diag}\{E[u_l(\theta) u_l(\theta)']\}$, i.e. by the diagonal vector of the score covariance matrix. To see this, note that each partial score $u_{l,r}(\theta)$ defines an unbiased estimating equation, i.e. satisfying $E u_{l,r}(\theta) = 0$. This implies the important relationship

$$E[u_l^{ML}(\theta) u_{l,r}(\theta)] = E \left[-\frac{\partial}{\partial l} u_{l,r}(\theta) \right] = E \left[u_{l,r}(\theta)^2 \right], \tag{12}$$

where the first equality in (12) is obtained by differentiating $E u_{l,r}(\theta) = 0$ under the integral, whilst the second equality is the Bartlett's identity. Unbiasedness implies the important relationship

$$\begin{aligned} E[u_l^{ML}(\theta) u_{l,r}(\theta)] &= E \left[u_l^{ML}(\theta) (u_{l,1}(\theta), \dots, u_{l,m}(\theta))' \right] \\ &= E \left[(u_{l,1}(\theta)^2, \dots, u_{l,m}(\theta)^2)' \right]. \end{aligned} \tag{13}$$

The last expression in (13) can be written $\text{diag}\{E[u_l(\theta) u_l(\theta)']\}$ with $\text{diag}(A)$ denoting the vector collecting the diagonal elements of the square matrix A .

Finally, replacing the expression of the covariance matrix $E \left[u_l(\theta) u_l(\theta)' \right]$ by its empirical counterpart $\hat{S}_l(\theta) = n^{-1} \sum_{i=1}^n u_l^{(i)}(\theta) u_l^{(i)}(\theta)'$ leads to the formulation of the empirical criterion

$$\hat{Q}_\lambda(\theta, w) = \sum_{l \in \{\alpha, \rho\}} \left\{ w'_l \hat{S}_l(\theta) w_l - 2w'_l \text{diag}\{\hat{S}_l(\theta)\} + \sum_{r=1}^m \lambda_l |w_{l,r}| \right\}, \tag{14}$$

For a given θ , we minimize $\hat{Q}_\lambda(\theta, w)$ to obtain the empirical composition rule $\hat{w}(\theta)$.

Further insight on the solution from the above minimization program may be helpful. The truncated composition rule solving the empirical objective (14)

contains elements that are exactly zero when the corresponding sub-likelihood scores are weakly correlated to the maximum likelihood score. To see this, let $\hat{w} = \hat{w}(\theta)$ be the minimizer of (14) with θ fixed and equal to the true parameter value for simplicity. Then the truncated composition rule $\hat{w} = (\hat{w}'_{\alpha}, \hat{w}'_{\rho})'$ minimizing the empirical objective (14) has the form

$$\hat{w}_{l,\mathcal{A}} = \hat{S}_{l,\mathcal{A}}^{-1} \left\{ \text{diag}(\hat{S}_{l,\mathcal{A}}) - \lambda_l \text{sign}(\hat{w}_{l,\mathcal{A}}) \right\} \quad \text{and} \quad \hat{w}_{l,/\mathcal{A}} = 0, \quad l \in \{\alpha, \rho\}, \quad (15)$$

where $\mathcal{A} \subseteq \{1, \dots, m\}$ is the index set of selected scores such that

$$\min_{r \in \mathcal{A}} \left| \frac{1}{n} \sum_{i=1}^n u_{l,r}^{(i)T} (u_{l,r}^{(i)} - \sum_j \hat{w}_{l,j} u_{l,j}^{(i)}) \right| \geq \lambda_l, \quad (16)$$

$/\mathcal{A} = \{1, \dots, m\} \setminus \mathcal{A}$, and function $\text{sign}(w)$ denotes the vector sign function with r th element taking values $-1, 0$ and 1 if $w_r < 0, w_r = 0$ and $w_r > 0$, respectively. The details of the derivation of the solution (3.10) to the optimization problem (3.9) are found in Theorem 3.2 of Huang and Ferrari (2017).

Here $u_{l,r}^{(i)} - \sum_j \hat{w}_{l,j} u_{l,j}^{(i)}$ is the residual difference between the r th score component and the composite likelihood score, and $\hat{S}_{l,\mathcal{A}}$ is the covariance sub-matrix for the selected scores. One can show that the empirical average in the left hand side of (16) approximates $\text{Cov}(u_{l,r}, u_l^{ML} - \sum_j \hat{w}_{l,j} u_{l,j})$, i.e. the covariance between the score for the r th pair and the residual difference between maximum likelihood and pairwise likelihood scores. This means that our truncation approach retains only pairwise score terms $u_{l,r}$, able to explain the gap between the full likelihood score and the pair-wise score, while dropping the remaining scores.

One should note that our methodology relies on certain desirable asymptotic properties including unbiasedness of the truncated composite likelihood estimating equations. These are guaranteed only under certain regularity conditions and, unfortunately, are not straightforward to verify for the Brown-Resnick model. For unbiasedness of the selected equations with fixed weights, one important condition is differentiability in quadratic mean of each pairwise log-likelihood function. Following Corollary 4.6 in Dombry et al. (2016), the Brown-Resnick model on a pair of sites automatically satisfies their conditions A1-A3. This also implies that the overall pairwise log-likelihood with fixed coefficients is differentiable in quadratic mean. One complication is that, differently from the usual composite likelihood setting (e.g. see Padoan et al. (2010)), in our method the weights for pairwise likelihoods depend on the parameter θ , but in practice such weights are estimated from the data by plugging-in a root- n consistent estimator. As a consequence, additional regularity conditions concerning convergence of such weights in probability are needed. Using arguments analogous to Huang and Ferrari (2017), one main condition is that the matrix of pairwise scores is dominated by an integrable function not depending on the parameters. Finally, following Dombry et al. (2016), identifiability for pair-wise estimation holds if euclidean distances for any three sites are not equal.

3.3 Implementation and computational aspects

The analysis in Huang and Ferrari (2017) show that $\hat{Q}_\lambda(\theta, w)$ is a consistent estimate of the population criterion $Q_\lambda(\theta, w)$ (up to some irrelevant additive term not depending on w) as long as θ is in a root- n neighborhood of the true parameter value. Thus, we start by taking a computationally cheap and consistent preliminary estimate and then use the truncation method described in Section 3.2 to improve upon such initial estimate. In practice, our truncation procedure is applied through the following steps:

Step 0) Initialization: Find a root- n consistent estimate $\hat{\theta}$. This can be achieved by solving the estimating equation (7) with $w_k \sim \text{Bernoulli}(\pi)$, $1 \leq k \leq 2m$, where π is a desired fraction of initial nonzero coefficients.

Step 1) Truncation: Compute the truncated composition rule \hat{w} given in (15), by minimizing the empirical criterion $\hat{Q}_\lambda(\hat{\theta}, w)$. For sufficiently large λ_l , this step will result in a likelihood function with a number of terms set exactly equal to zero.

Step 2) Estimation: Obtain the updated estimate $\hat{\theta}_\lambda$ by solving the estimating equations (7) with $w = \hat{w}$.

The criterion $\hat{Q}_\lambda(\hat{\theta}, w)$ in Step 1 is a quadratic function of w , with a ℓ_1 constraint term. To solve the minimization problem in Step 1, we implement a step-up algorithm which essentially coincides with the least angle regression (LARS) algorithm of Efron et al. (2004). LARS starts with a large initial value of λ_l ($l \in \{\alpha, \rho\}$) which yields an initial solution of \hat{w}_l with all elements equal to zero. Then in each subsequent step, the algorithm includes exactly one score component at the time, say $u_{l,r}(\theta)$, in the current composite score $u(\theta, w)$, by decreasing λ_l in such a way that the correspondent coefficient in \hat{w}_l becomes different from zero. The included score components $u_{l,r}$ have their covariance with residuals $1/n \sum_{i=1}^n u_{l,r}^{(i)}(\hat{\theta})[u_{l,r}^{(i)}(\hat{\theta}) - \hat{w}_l u_l^{(i)}(\theta)]$ higher than those not included as discussed in (16). In the last step, the algorithm yields m coefficients $\hat{w}_l^{(1)}, \dots, \hat{w}_l^{(m)}$ corresponding to the nodes $\lambda_l^{(1)}, \dots, \lambda_l^{(m)}$, where $\hat{w}_l^{(t)}$ contains exactly t non-zero elements and $\lambda_l^{(t)}$ represents the node at which the value of λ_l is just small enough to allow for the inclusion of exactly t non-zero terms. Thus, the values $\lambda_l^{(1)}, \dots, \lambda_l^{(m)}$ represent different level of sparsity in \hat{w}_l and the selection of λ through such values is discussed in Section 3.5.

3.4 Standard errors

For a given composition rule w , the matrices $H_w(\theta)$ and $J_w(\theta)$ forming the Godambe information matrix given in (9) are estimated by their empirical counterparts

$$\hat{H}_w(\theta) = \mathcal{W}' \hat{D}(\theta), \quad \hat{J}_w(\theta) = \mathcal{W}' \hat{S}(\theta) \mathcal{W}, \quad (17)$$

where $\hat{S}(\theta) = n^{-1} \sum_{i=1}^n \tilde{u}^{(i)}(\theta) \tilde{u}^{(i)}(\theta)'$ is the empirical score covariance matrix, and $\hat{D}(\theta)$ is a $2m \times 2$ matrix with the first m rows and last m rows stacks of elements

$(\hat{S}(\theta)_{j,j}, \hat{S}(\theta)_{j,j+m})$ and $(\hat{S}(\theta)_{j+m,j}, \hat{S}(\theta)_{j+m,j+m})$, $j = 1, \dots, m$, respectively. A plug-in estimate $\widehat{\text{var}}(\hat{\theta}_\lambda)$ of the variance of the final estimator $\hat{\theta}_\lambda$ is found by replacing $\hat{\theta}_\lambda$ and its composition rule \hat{w} in (17) to obtain:

$$\widehat{\text{var}}(\hat{\theta}_\lambda) = \frac{1}{n} \hat{H}_{\hat{w}}(\hat{\theta}_\lambda)^{-1} \hat{J}_{\hat{w}}(\hat{\theta}_\lambda) \hat{H}_{\hat{w}}(\hat{\theta}_\lambda)^{-1}. \tag{18}$$

Estimating the asymptotic variance of composite likelihood estimators is notoriously difficult. When the composition rules w contains all non-zero coefficients, $\hat{J}_w(\theta)$ may involve a very large number of noisy score covariance terms. When the number of sites s (and the corresponding number of sublikelihoods $2m$) is moderate or large, this increases the computational cost and implies inaccurate estimates of PLE’s variance. The proposed plug-in estimate (18), on the other hand, represents a computationally efficient and stable alternative. For an appropriate choice of $\lambda = (\lambda_\alpha, \lambda_\rho)'$, the truncated composition rule \hat{w} does not include elements corresponding to the noisiest pairwise scores. As a results, the plug-in variance estimator $\widehat{\text{var}}(\hat{\theta}_\lambda)$ is expected to be more accurate and compationally stable compared to the variance estimator that uses all nonzero elements in w .

3.5 Selection of λ

Let \hat{k}_l , $l \in \{\alpha, \rho\}$, be the number of non-zero elements in the selected composition rule \hat{w}_l found by minimizing the empirical objective (14). Recall that for the LARS-type algorithm described in Section 3.3 selecting the number of non-zero components in \hat{k}_l is equivalent to setting corresponding tuning constant λ_l . We choose \hat{k}_l such that at least a fraction of the total information available on parameter l is reached.

Let $\hat{S}_l^{(t)}$ be the $t \times t$ empirical covariance between sub-scores for parameter l after t steps of the LARS algorithm (i.e. after including t terms in the pair-wise likelihood equation), and $\epsilon_l^{(t)}$ be the smallest eigenvalue of $\hat{S}_l^{(t)}$. If we regard $\epsilon_l^{(t)}/\text{tr}(\hat{S}_l^{(t)})$ as the reduction on variability (information gain) on l in step t . The Min-Max Theorem of linear algebra implies that including the remaining non-selected sub-likelihood components will increase the information on l by at most $1 + (m - t)\epsilon_l^{(t)}/\text{tr}(\hat{S}_l^{(t)})$.

We propose to find \hat{k}_l using the empirical rule

$$\hat{k}_l = \min\{t \in \{1, \dots, m\} : \phi_l(t) > \tau\},$$

for some user-specified constant $\tau \in (0, 1)$, where $\phi_l(t)$ is the criterion

$$\phi_l(t) = \left\{ 1 + \frac{(m - t)\epsilon_l^{(t)}}{\text{tr}(\hat{S}_l^{(t)})} \right\}^{-1}. \tag{19}$$

The proportion of information obtained up to step t has to be greater than $\phi_l(t)$. In practice, we choose values τ close to 1. Particularly, the value $\tau = 0.9$ is found to select a number of pair-wise likelihood components that balance well computing and statistical efficiency in most of our numerical examples.

The advantage of our application of the LARS algorithm is that it does not require re-estimation of θ and of the Godambe information for each value of λ . As a consequence estimates of the asymptotic variance are not necessarily computed for each λ . On the other hand, the pair-wise scores are only estimated once at the beginning of the algorithm and can be used to guide selection of λ as described in the above criterion.

While in principle one may select λ by optimizing a criterion based on the estimated Godambe information, this would require additional computations. Namely, at each step of the algorithm w is updated entirely, meaning that re-estimation of θ and re-computation of the matrices in the Godambe information would be also necessary for each value of λ . While this is feasible in small problems, it might be challenging for certain data sets containing a large number of observations.

3.6 Missing data

In our numerical applications there are no missing data. In practice, however, often not all sites have data for all years. Some insight on how to proceed in such a setting may be helpful. Suppose that at time i , we have only k sites. Without loss of generality, let $Z_{obs}^{(i)} = (Z_1^{(i)}, \dots, Z_k^{(i)})$, $k < s$, be the vector observed data at time i , where s is the total number of available sites. The missing data are denoted by $Z_{mis}^{(i)} = (Z_{k+1}^{(i)}, \dots, Z_s^{(i)})$. Let $T^{(i)} = (T_1^{(i)}, \dots, T_s^{(i)})$ be a random vector with binary entries indicating the missing data ($T_j^{(i)} = 0$ if the observation at time i and location j is missing and $T_j^{(i)} = 1$ otherwise). Assume that $T^{(i)}$ is an independent draw from the distribution depending on an unknown parameter ζ . Here θ denotes the max-stable parameter of interest ($\theta = (\alpha, \rho)'$ for the Brown-Resnik model).

The type of treatment for missing data depends on the specific model for the missing data mechanism. For simplicity, here we limit our discussion to the the case of missing completely at random (MCAR) data. The observed data-likelihood function for r th pair $\{(j, k), 1 \leq j < k \leq s\}$ evaluated at the i th observation can be written as

$$L_r^{(i)}(z_j^{(i)}, z_k^{(i)}, t_j^{(i)}, t_k^{(i)}; \theta, \zeta) = g_r(t_j^{(i)}, t_k^{(i)}; \zeta) \times f_{obs}(z_j^{(i)}, z_k^{(i)}; t_j^{(i)}, t_k^{(i)}, \theta),$$

where $g(\cdot; \zeta)$ is the bivariate pmf of (T_j, T_k) , and

$$f_r(z_j^{(i)}, z_k^{(i)}; t_j^{(i)}, t_k^{(i)}, \theta) = \begin{cases} f_r(z_j, z_k; \theta) & \text{if } t_j^{(i)} = 1, t_k^{(i)} = 1, \\ \int f_r(z_j^{(i)}, z_k; \theta) dz_k & \text{if } t_j^{(i)} = 1, t_k^{(i)} = 0, \\ \int f_r(z_j, z_k^{(i)}; \theta) dz_j & \text{if } t_j^{(i)} = 0, t_k^{(i)} = 1, \\ 1 & \text{if } t_j^{(i)} = 0, t_k^{(i)} = 0, \end{cases}$$

where $f_r(\cdot; \theta)$ is the bivariate max-stable model defined in (5). Note that when observation is missing in either site j or k (i.e., $t_j^{(i)} = 0$ or $t_k^{(i)}$), $f_{obs}(\cdot; \theta) = 0$ is actually independent of the parameter θ . This is because marginalization of bivariate max-stable model leads to unit Frèchét univariate distributions. This means that the truncated pair-wise likelihood estimator (TPLE) in Section 3.1 can be computed

as usual, but the pair-wise likelihood scores terms in the estimatin equation will be $u_r^{(i)} = \partial \log f_r(z_j^{(i)}, z_k^{(i)}; \theta) / \partial \theta$ if $t_j^{(i)} = t_k^{(i)} = 1$, and $u_r^{(i)} = 0$ otherwise.

4 Monte Carlo simulations

We simulate from the Brown-Resnick model described in Section 2 for various settings of the parameter $\theta = (\rho, \alpha)'$ using the R package `SpatialExtremes` (Ribatet 2015). We implement the two-step approach described in Section 3.3 to find the truncated PLE (TPLE) $\hat{\theta}_\lambda$. The preliminary estimate $\hat{\theta}$ is found by setting $\pi = 0.3$. We investigate the statistical efficiency and computational cost of TPLE. For comparison, we consider the PLE with uniform coefficients $w_{\text{unif1}} = (1, \dots, 1)$ (UPLE) due to its widespread use, and the PLE with coefficients w_{unif2} set to 1 if the corresponding pairs of locations have distance less than one third of the radius of study region, or 0 otherwise (UPLE₂). We also consider the random PLE (RPLE) with coefficients w_{rand} containing $0.3 \times 2m$ elements equal 1 at random positions, where m is the total number of pair-wise likelihoods.

The performance of our method is measured by Monte Carlo estimates of the relative mean squared error of $\hat{\theta}_\lambda = (\hat{\rho}_\lambda, \hat{\alpha}_\lambda)'$ and required CPU time compared the other composition rules. Particularly, we estimate the relative mean squared errors $\text{RE}^{(1)} = \text{RE}(w_{\text{unif1}})$, $\text{RE}^{(2)} = \text{RE}(w_{\text{rand}})$ $\text{RE}^{(3)} = \text{RE}(w_{\text{unif2}})$, and the relative computing times $\text{RC}^{(1)} = \text{RC}(w_{\text{unif1}})$, $\text{RC}^{(2)} = \text{RC}(w_{\text{rand}})$, $\text{RC}^{(3)} = \text{RC}(w_{\text{unif2}})$, where

$$\text{RE}(w) = \frac{\text{MSE}\{\hat{\theta}_\lambda\}}{\text{MSE}\{\hat{\theta}(w)\}}, \quad \text{RC}(w) = \frac{\text{CPU time for } \hat{\theta}_\lambda}{\text{CPU time for } \hat{\theta}(w)},$$

where $\hat{\theta}(w)$ is the pairwise likelihood estimator obtained using the composition rule w .

Simulation 1 In our first simulation, we illustrate the impact of the tuning constants $(\lambda_\alpha, \lambda_\rho)$ – or, equivalently, the number of selected pair-wise likelihood terms – on statistical accuracy and computational efficiency. Figure 3 (top row) shows the number of pairs of sites selected, i.e. the numbers of nonzero elements in the estimated coefficients $\hat{w}_\alpha = (\hat{w}_{\alpha,1}, \dots, \hat{w}_{\alpha,m})'$ and $\hat{w}_\rho = (\hat{w}_{\rho,1}, \dots, \hat{w}_{\rho,m})'$ against the criterion $\phi_l(t)$, $l \in \{\alpha, \rho\}$ defined in (3.5). Recall that $\phi_l(t)$ represents a lower bound on the explained variability in the selected pair-wise scores after t terms are included in the pairwise likelihood equations. The curves are obtained from a single simulation at 30 randomly selected locations on $[0, 100]^2$. Figure 3 (bottom rows) shows Monte Carlo estimates of the relative efficiency of the TPLE compared to the UPLE, separately for parameters α and ρ against $\phi_\alpha(t)$ and $\phi_\rho(t)$ ($\text{RE}_\alpha^{(1)}$ and $\text{RE}_\rho^{(1)}$, respectively). Estimates are based on 1000 Monte Carlo samples of size 50 from a Brown-Resnick process at 30 randomly selected locations on $[0, 100]^2$, which are not varied throughout simulations. Remarkably, selecting just 20 to 30 pair-wise score terms (i.e. 5 to 7% of the entire set of feasible terms), already gives dramatic improvements in terms of relative efficiency compared to UPLE.

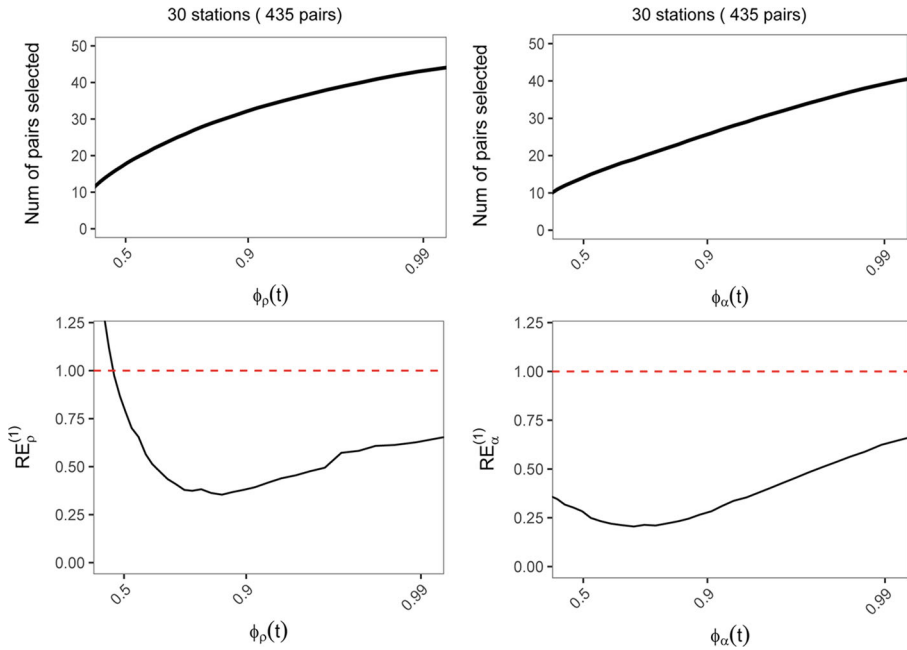


Fig. 3 Top row: Numbers of nonzero elements in composition rule \hat{w}_ρ and \hat{w}_α obtained by minimizing the empirical objective $\hat{Q}_\lambda(\hat{\theta}, w)$ in (14), against the lower bounds on scores variability, $\phi_\rho(t)$ and $\phi_\alpha(t)$, after including t terms as defined (19). The top part of each plot shows the number of pair-wise terms included. Plots are obtained from a single realization of the Brownick-Resnick process with $(\rho, \alpha) = (2.8, 1.5)$ at 30 random sites on $[0, 100]^2$. Bottom rows: Monte Carlo estimates of relative efficiencies $RE_\rho^{(1)}$ and $RE_\alpha^{(1)}$ given by the mean squared error for the TPLEs $\hat{\rho}_\lambda$ and $\hat{\alpha}_\lambda$ divided by the mean squared error of the PLE with uniform composition rule $w = (1, \dots, 1)'$. Estimates are based on 1000 samples of size 50 from a Brown-Resnick process with $(\rho, \alpha) = (28, 1.5)$ at 30 random sites on $[0, 100]^2$

The computational complexity (Fig. 3, top rows) increases when the number of pair-wise scores with coefficients different from zero increases (equivalently, when λ_l decreases). Thus, the computing cost is maximum when $\lambda_\alpha = \lambda_\rho = 0$, since all the pair-wise scores are included. The relative error (Fig. 3, bottom rows) follows a U-shape behavior which is explained as follows. The optimal theoretical weights are given by $\lambda_\alpha = \lambda_\rho = 0$, corresponding to the optimal estimating equations described in Heyde (2008). However, such optimal weights are not achievable in practice due to the substantial correlation between pair-wise scores and the presence of estimation error. This means that for λ_l close to zero the estimated composite likelihood coefficients \hat{w} becomes increasingly unstable yielding parameter estimates $\hat{\theta}(\hat{w})$ with large variance. Specifically, the optimal weights depend on the inverse of the estimated pair-wise score covariance matrix, which is nearly singular in presence of pronounced spatial correlation. This behavior is exacerbated when the number of sites increases. On the other hand, by including too few pair-wise scores in the likelihood equation (i.e. setting too large $\lambda_\alpha, \lambda_\rho$), some important information on the parameter may be missed thus resulting in poor accuracy of the parameter estimator.

Simulation 2 In our second Monte Carlo experiment, we carry out a systematic assessment of the performance of the TPLE compared to UPLE and RPLE. For the TPLE, we consider various choices for the minimum proportion of explained score variability ($\tau = 0.9, 0.95$ and 0.99). Tables 1 and 2 show results based on 1000 Monte Carlo samples of size 50 from a Brown-Resnick process with different smoothness and range parameters observed, respectively, at 20 and 30 random locations on $[0, 100]^2$, which are not varied throughout simulations. We report Monte Carlo estimates for the following quantities: mean number of pair-wise score terms included ($\#Terms$), $E(\hat{\alpha}_\lambda)$ and $E(\hat{\rho}_\lambda)$, $sd(\hat{\alpha}_\lambda)$ and $sd(\hat{\rho}_\lambda)$, relative mean squared error and relative computing cost of the TPLE compared to UPLE and RPLE. Whilst the TPLE generally outperforms the UPLE in terms relative efficiency, it also performs comparably to the random PLE in terms of computational cost. Both the accuracy and computing efficiency of TPLE become more pronounced as the number of sites increases. Finally, note that when α and ρ decrease, the TPLE tends to perform similarly to the UPLE in terms of efficiency. This is not surprising since in this situation observations between sites become increasingly independent and all sub-likelihoods contain roughly the same information on the parameters.

Simulation 3 In our third Monte Carlo experiment, we examine the estimator of the extremal coefficient, a useful quantity in spatial analysis of extremes. The accuracy of our method is compared with UPLE and RPLE. We also assess the accuracy of the estimated extremal coefficients $\hat{\eta}_2(h)$, obtained by plugging-in parameter estimates $\hat{\theta}$ in the formula $\eta_2(h; \theta)$ given in Section 2.3. Figure 4 (top row) shows the fitted extremal coefficients curves $\hat{\eta}_2(h)$ based on the estimated and the true parameters. Figure 4 (bottom row) shows the corresponding mean square errors of the estimates obtained by plugging-in TPL, UPL and RPL estimates. The lighter circles in the plots correspond to empirical estimates of the pairwise coefficients. Results are based on 1000 Monte Carlo samples of size 50, generated from 20 and 30 randomly selected sites on $[0, 100]^2$ with true parameters $(\alpha, \rho) = (1.5, 28)$. Whilst all the estimators tend to underestimate the extremal coefficient for relatively large $\|h\|$, our truncation approach clearly outperforms the other two methods.

5 Analysis of Victoria extreme temperature data

In this section, we apply the new estimation method to maximum temperature data recorded in the state of Victoria, Australia. Daily temperature maxima from 1971 to 2017 are provided by the national Australian meteorological service, the Bureau of Meteorology (data are available at <http://www.bom.gov.au/climate/data>). The final dataset contains the highest annual temperature recordings measured at 26 stations over 47 years from 1971 to 2017. The distances between these stations range between 13 and 1100 kilometers. The locations for the measuring stations are shown in Figure 6 (left). Sites colored in blue (red) correspond to average maximum temperatures below (above) the average maximum temperature across all sites.

The main objective of our analysis is to estimate the correlation structure pertaining extreme temperatures. As a pre-processing step, we transform the data at each

Table 1 Monte Carlo results for TPL estimation for the Brownick-Resnick process measured at 20 locations

(ρ, α)	(28,0.5)	(42,0.5)	(28, 1.5)	(42, 1.5)	(28,1.9)	(42,1.9)
τ	0.90/0.95/0.99	0.90/0.95/0.99	0.90/0.95/0.99	0.90/0.95/0.99	0.90/0.95/0.99	0.90/0.95/0.99
#Terms	48.3/55.9/68.3	44.5/51.8/64.8	36.6/43.6/58.4	31.6/37.0/50.5	33.6/39.9/54.4	27.2/31.6/43.8
$\hat{\rho}_\lambda$	25.7/26.3/26.7	42.3/42.4/42.6	28.4/28.9/29.4	41.4/41.9/42.5	29.9/30.3/31.0	41.7/42.2/43.0
$\hat{\alpha}_\lambda$	0.50/0.49/0.49	0.47/0.47/0.47	1.40/1.38/1.36	1.48/1.47/1.45	1.64/1.62/1.57	1.84/1.83/1.80
$sd(\hat{\rho}_\lambda)$	7.63/7.50/6.87	13.09/12.40/11.85	2.44/2.37/2.44	4.70/4.68/4.46	2.37/2.37/2.42	3.80/3.57/3.46
$sd(\hat{\alpha}_\lambda)$	0.09/0.09/0.09	0.09/0.09/0.09	0.10/0.10/0.11	0.10/0.09/0.09	0.12/0.13/0.13	0.09/0.09/0.09
$RE_\rho^{(1)}$	1.06/0.99/0.81	0.73/0.65/0.60	0.52/0.54/0.68	1.08/1.05/0.97	0.34/0.40/0.53	0.64/0.56/0.58
$RE_\alpha^{(1)}$	0.96/0.97/1.01	0.43/0.45/0.50	0.47/0.55/0.70	0.60/0.58/0.62	0.46/0.55/0.70	0.35/0.38/0.51
$RE_\rho^{(2)}$	0.82/0.76/0.63	0.71/0.64/0.58	0.41/0.43/0.55	0.64/0.62/0.57	0.28/0.33/0.44	0.50/0.44/0.45
$RE_\alpha^{(2)}$	0.67/0.68/0.71	0.37/0.38/0.42	0.38/0.44/0.57	0.28/0.27/0.28	0.41/0.49/0.64	0.22/0.24/0.32
$RE_\rho^{(3)}$	1.04/0.93/0.79	0.55/0.50/0.45	0.63/0.66/0.83	0.52/0.51/0.47	0.68/0.80/1.06	0.54/0.48/0.49
$RE_\alpha^{(3)}$	0.43/0.46/0.50	0.29/0.30/0.34	0.59/0.68/0.88	0.55/0.54/0.57	1.32/1.58/1.03	0.25/0.27/0.35
$RC^{(1)}$	0.16/0.17/0.20	0.09/0.10/0.13	0.26/0.28/0.32	0.26/0.28/0.32	0.29/0.31/0.36	0.34/0.36/0.40
$RC^{(2)}$	0.55/0.62/0.72	0.34/0.38/0.48	0.82/0.89/1.03	0.84/0.91/1.05	0.95/1.00/1.15	1.06/1.12/1.25
$RC^{(3)}$	0.17/0.19/0.20	0.14/0.15/0.16	0.68/0.72/0.85	0.63/0.68/0.78	0.67/0.72/0.83	0.69/0.73/0.82

Rows 1–2: True parameter values (ρ, α) and minimum proportion of explained score variability (τ) . Rows 3–7: Monte Carlo estimates for: mean number of selected pair-wise score terms (#Terms), $E(\hat{\rho})$, $E(\hat{\alpha})$, $sd(\hat{\rho})$, and $sd(\hat{\alpha})$. Rows 8–16: Relative efficiency of TPLE compared to UPLE ($RE_\rho^{(1)}$, $RE_\alpha^{(1)}$), RPLE ($RE_\rho^{(2)}$, $RE_\alpha^{(2)}$) and UPLE₂ ($RE_\rho^{(3)}$, $RE_\alpha^{(3)}$), relative computing cost of TPLE compared to UPLE ($RC^{(1)}$), RPLE ($RC^{(2)}$) and UPLE₂ ($RC^{(3)}$). Results are based on 1000 samples of size 50 from a Brown-Resnick model on $[0, 100]^2$

Table 2 Monte Carlo results for TPL estimation for the Brownick-Resnick process measured at 30 locations

(ρ, α)	(28,0.5)	(42,0.5)	(28, 1.5)	(42, 1.5)	(28,1.9)	(42,1.9)
τ	0.90/0.95/0.99	0.90/0.95/0.99	0.90/0.95/0.99	0.90/0.95/0.99	0.90/0.95/0.99	0.90/0.95/0.99
#Terms	59.5/66.5/77.2	55.7/62.6/73.4	46.1/52.8/65.9	42.4/48.2/60.6	41.6/48.5/62.3	35.8/41.1/53.9
$\hat{\rho}_\lambda$	25.1/25.5/26.2	40.3/40.5/40.8	28.1/28.5/29.1	40.7/41.2/42.2	29.5/29.8/30.5	41.0/41.5/42.3
$\hat{\alpha}_\lambda$	0.50/0.50/0.49	0.48/0.47/0.47	1.43/1.41/1.39	1.49/1.48/1.46	1.72/1.69/1.64	1.87/1.85/1.83
$sd(\hat{\rho}_\lambda)$	5.63/5.45/5.22	11.46/10.91/10.36	2.42/2.26/2.41	4.34/4.13/4.09	2.23/2.24/2.26	3.24/3.21/3.14
$sd(\hat{\alpha}_\lambda)$	0.06/0.06/0.06	0.07/0.07/0.08	0.08/0.08/0.09	0.08/0.07/0.08	0.10/0.10/0.11	0.07/0.07/0.07
$RE_\rho^{(1)}$	1.43/1.27/1.08	0.76/0.68/0.61	0.44/0.40/0.54	1.01/0.86/0.83	0.27/0.31/0.42	0.54/0.50/0.47
$RE_\alpha^{(1)}$	0.90/0.95/1.05	0.35/0.37/0.41	0.32/0.38/0.57	0.53/0.53/0.66	0.29/0.38/0.54	0.23/0.28/0.43
$RE_\rho^{(2)}$	1.18/1.05/0.89	0.71/0.64/0.57	0.42/0.38/0.52	0.71/0.61/0.59	0.22/0.26/0.35	0.52/0.48/0.46
$RE_\alpha^{(2)}$	0.64/0.68/0.75	0.30/0.32/0.35	0.27/0.33/0.49	0.26/0.26/0.32	0.26/0.33/0.48	0.18/0.22/0.34
$RE_\rho^{(3)}$	1.04/0.93/0.79	0.55/0.50/0.45	0.81/0.74/1.00	0.30/0.26/0.25	0.58/0.69/0.92	0.08/0.08/0.07
$RE_\alpha^{(3)}$	0.43/0.46/0.50	0.29/0.30/0.34	0.73/0.89/1.31	0.06/0.06/0.08	1.04/1.34/1.91	0.14/0.17/0.26
$RC^{(1)}$	0.09/0.10/0.10	0.06/0.06/0.07	0.16/0.17/0.18	0.16/0.17/0.18	0.17/0.18/0.20	0.19/0.19/0.21
$RC^{(2)}$	0.30/0.33/0.35	0.21/0.23/0.25	0.53/0.55/0.59	0.54/0.57/0.58	0.55/0.59/0.65	0.62/0.64/0.71
$RC^{(3)}$	0.17/0.19/0.20	0.14/0.15/0.16	0.27/0.28/0.30	0.29/0.31/0.32	0.26/0.28/0.31	0.32/0.33/0.36

Rows 1–2: True parameter values (ρ, α) and minimum proportion of explained score variability (τ) . Rows 3–7: Monte Carlo estimates for: mean number of selected pairwise score terms (#Terms), $E(\hat{\rho})$, $E(\hat{\alpha})$, $sd(\hat{\rho})$, and $sd(\hat{\alpha})$. Rows 8–16: Relative efficiency of TPLE compared to UPLE ($RE_\rho^{(1)}$, $RE_\alpha^{(1)}$), RPLE ($RE_\rho^{(2)}$, $RE_\alpha^{(2)}$) and UPLE₂ ($RE_\rho^{(3)}$, $RE_\alpha^{(3)}$), relative computing cost of TPLE compared to UPLE ($RC^{(1)}$), RPLE ($RC^{(2)}$) and UPLE₂ ($RC^{(3)}$). Results are based on 1000 samples of size 50 from a Brown-Resnick model on $[0, 100]^2$

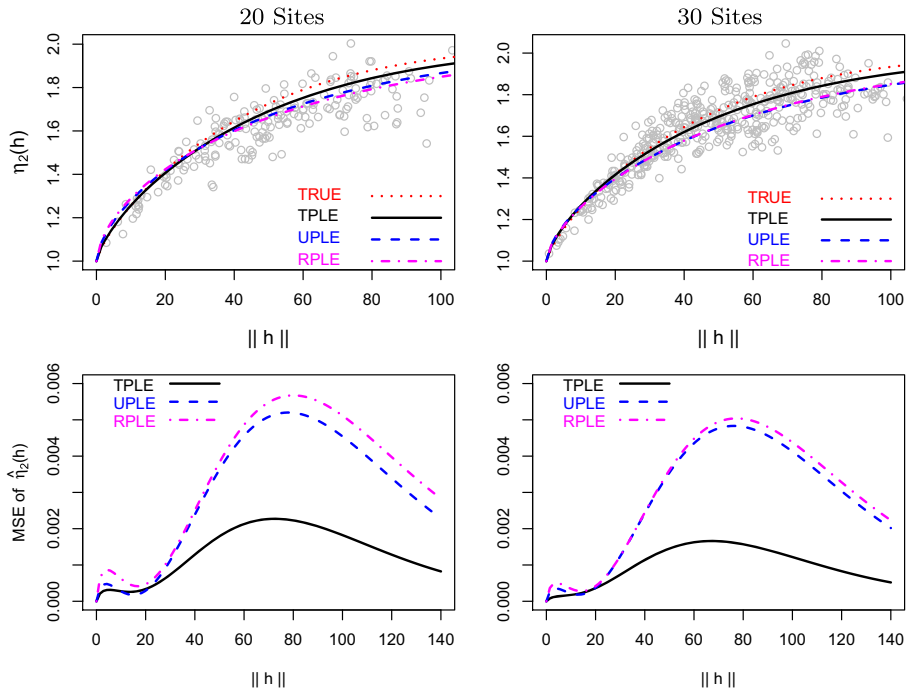


Fig. 4 Plug-in estimates of the pair-wise extremal coefficient based on truncated (TPLE), uniform (UPL) and random composition rules (RPL). Top row: Empirical pair-wise extremal coefficients based on simulated data (light gray circles) and plug-in estimates of the extremal coefficient curve, $\eta_2(h; \theta_w)$, based on TPLE, UPL and RPL. Bottom row: Monte Carlo estimates of the mean square error for the estimated extremal coefficient against distance $\|h\|$. Plots are based on 1000 Monte Carlo samples of size 50, generated from 20 (left column) and 30 (right column) random selected locations on $[0, 100]^2$ with parameters $\alpha = 1.5$ and $\rho = 28$

location into a unit Fréchet distribution with marginal parameters obtained by fitting Generalized Extreme Value models at each location. Extreme dependence parameters under the Brown-Resnick model are obtained using truncated, random and uniform PLEs. Standard deviations and covariances of the pairwise likelihood estimators are calculated by the sandwich approximation of the inverse Godambe information matrix described in Section 3.4.

Figure 5 (top) depicts the entire trajectory for range and smoothness parameters fitted using the TPLE for increasing explained score variability $\phi_\rho(t)$ and $\phi_\alpha(t)$, respectively, along with 95% confidence bands. For comparison, the horizontal dot-dashed line represent the UPLE estimate. Figure 5 (bottom) gives the number of nonzero elements for the truncated composition rules \hat{w}_ρ and \hat{w}_α with number of selected pair-wise score terms reported on the top axes. Note that just by including a small fraction pair-wise likelihoods, the TPLE is very close to the UPLE involving all 325 pair-wise likelihood terms. However, we find that the TPLE has much smaller standard errors compared to the UPLE. For example, when $\tau = 0.95$, the 95%

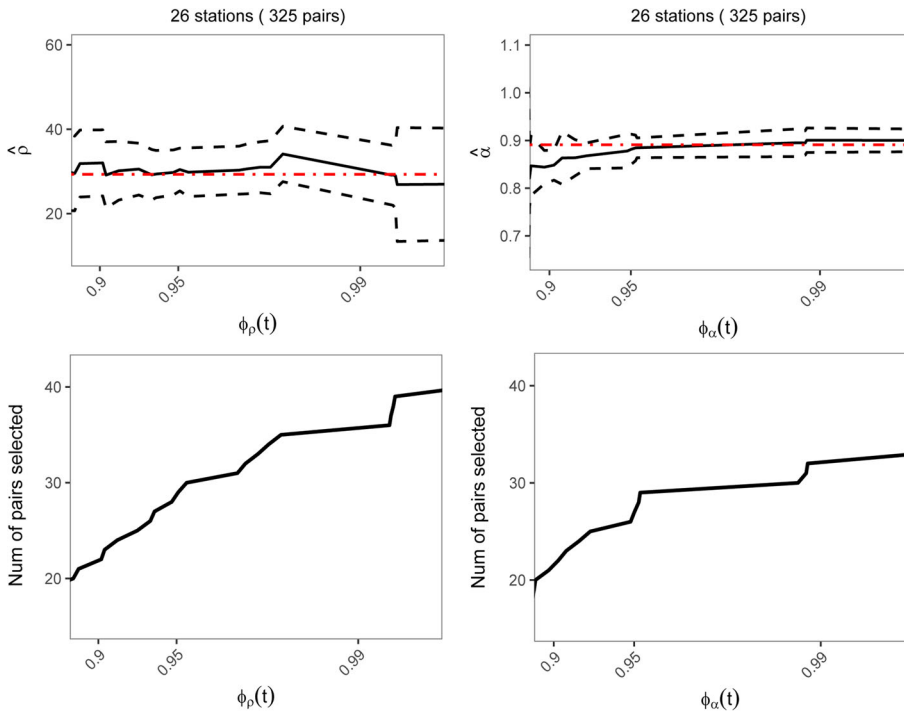


Fig. 5 Estimation of the Brown-Resnick model for the BoM extreme temperature data. Top: TPL estimates of range and smoothness parameters ($\hat{\rho}_\lambda$, and $\hat{\alpha}_\lambda$) against the lower bound of information explained $\phi_\rho(t)$ and $\phi_\alpha(t)$. The horizontal dash-dot line correspond to the uniform PL estimate. The dashed lines are 95% confidence bands obtained by the estimated inverse Godambe information in Section 3.4. Bottom: Number of nonzero elements in the truncated composition rules \hat{w}_ρ and \hat{w}_α with the number of selected pair-wise likelihoods reported on the top axis

confidence intervals for α and ρ based on TPLE are (25.38, 35.52) and (0.84, 0.91), respectively. Those based on the UPLE are (18.23, 40.39) and (0.75, 1.02).

In Figure 6 (left), edges joining site locations represent the selected pair-wise scores by our truncation method when $\tau = 0.9$. Note that the selected scores represent a small fraction of the available 325 pair-wise terms, and generally corresponding to pairs of sites close to each other. This is not surprising, since pairs of neighboring sites are expected provide more information on extreme correlation compared to those far apart from each other. Figure 6 (right) shows extremal coefficient estimates obtained by plugging-in the truncated, uniform and random pair-wise likelihood estimators of the smoothness and range parameters. Whilst fo TPL and UPL estimates are generally very close, the TPL extremal coefficient estimate becomes smaller than the UPL estimate as the distance $\|h\|$ increases. Finally, in Fig. 7 we show fitted extremal dependence coefficients and realizations of temperature maxima in degrees Celsius simulated from fitted max-stable models by TPLE (with $\tau = 0.9$) and UPLE on the map of Victoria (bottom row).

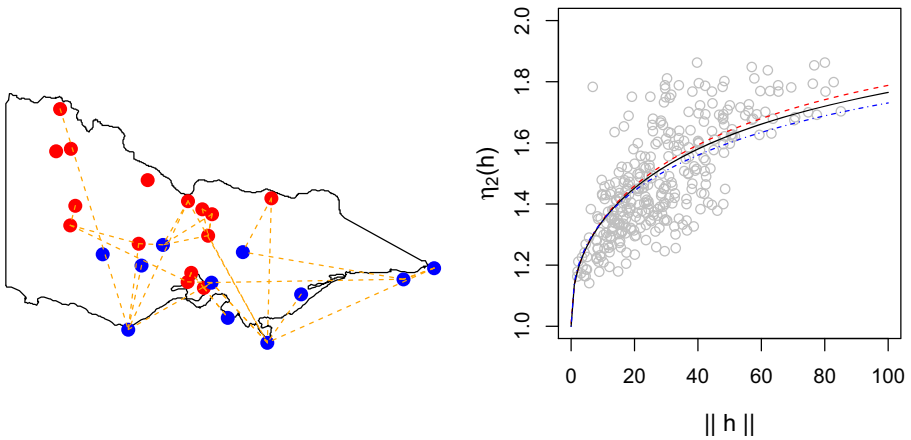


Fig. 6 Left: Points on the map represents sites of weather stations in the state of Victoria, Australia. Dashed lines connecting the sites represent pair-wise likelihood scores selected by our truncation methodology described in Section 3.2. Right: Light gray points represent empirical extremal coefficient estimates against distances between corresponding locations. The smooth curves represent fitted extremal coefficients obtained by plugging-in the truncated (solid), uniform (dashed) and random (dot-dashed) pair-wise likelihood estimators of the smoothness and range parameters (α and ρ) for the Brown-Resnick model

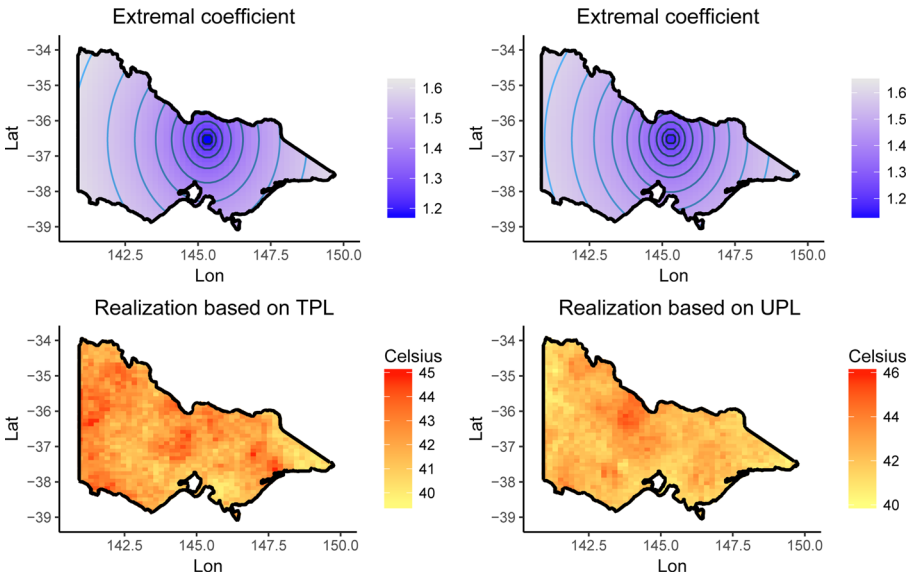


Fig. 7 Maps of the state of Victoria, Australia, with fitted extremal dependence coefficients (top row) and temperature maxima in degrees Celsius simulated from fitted max-stable models (bottom row) using the proposed truncated pair-wise likelihood (TPLE, left column) and the classic uniform pair-wise likelihood (UPLE, right column) estimators

6 Conclusion and final remarks

Building on the general methodology introduced in Huang and Ferrari (2017), we have developed and applied a new truncation strategy for pair-wise likelihood estimation in the Brown-Resnick model, a popular dependence model for spatial extremes. Our method represents a statistically efficient and computationally parsimonious approach for estimating parameters of spatial max-stable models for extreme values. The pair-wise likelihoods constructed by our new method is obtained by minimizing an estimate of the ℓ_2 -distance between the composite likelihood score and full likelihood score subject to a ℓ_1 -norm penalty representing the computational cost (or composite likelihood complexity). When the number of pair-wise likelihood terms considered for estimation is relatively large compared to sample size, traditional CL estimators with uniform weights may be inaccurate due to potentially many large correlations between the sub-likelihood scores (Cox and Reid 2004; Lindsay et al. 2011). This issue is crucial for pairwise likelihood estimation in the context of spatial max-stable models (Sang and Genton 2014).

The proposed ℓ_1 -penalization strategy carries out principled selection of pair-wise likelihood terms. Particularly, only the pair-wise scores with the largest correlation with the maximum likelihood score are kept in the estimating equations (see Section 3.2). This mechanism is shown to improve statistical efficiency whilst reducing the computational burden associated with extensive combinations of equally weighted CL objects. These features of our truncation method make it particularly effective for analyzing large datasets with measurements taken at many sites, where the number of noisy or redundant likelihood terms often increases very quickly with the number of sites. Huang and Ferrari (2017) show that ℓ_1 -truncation yields CL estimators with the asymptotic variance approaching that of optimal CL estimators (with no penalty). The results in this paper supports their theoretical results and suggest that CL truncation by ℓ_1 -penalization is a valid option when dealing with complex likelihood estimation problems.

Various research directions based on modification the current approach may be pursued in the future. One research direction concerns the choice of the penalty function. Inspired by the literature in variable selection in regression (see, e.g., ?Efron04, Fan10 ()), we achieve sparsity of estimating equations via the ℓ_1 -penalty described in (10). Note, however, that in our context the penalty involves the composition rule w but does not the model parameter θ , which is low-dimensional and is treated as fixed. This means that, differently from penalized regression, our estimating equations (and resulting parameter estimates) remain approximately unbiased even when λ is not zero. In the future, exploring other suitable sparsity-promoting penalties strategies may be valuable to deal with cases where both the CL complexity and size of the parameter space are large. Another research direction concerns the choice of the CL design. In this paper we have focused on pair-wise likelihood inference, but our approach may be used for higher-order CL designs with sub-likelihood constructed on site pairs, triplets, quadruples, etc, similarly to Cas-truccio et al. (2016). Finally, since settings beyond the Brown-Resnick model were not directly pursued in the present paper, numerical studies and applications of the TCLE in the context of other max-stable processes would be also valuable.

Appendix

Calculation of pair-wise score functions

Let

$$\begin{aligned}\dot{V}_j &= \frac{dV_r}{dz_j} = \frac{\varphi(m_j)}{a(\theta)z_j^2} + \frac{\varphi(m_k)}{a(\theta)z_j z_k} - \frac{\Phi(m_j)}{z_j^2}, \\ \dot{V}_k &= \frac{dV_r}{dz_k} = \frac{\varphi(m_k)}{a(\theta)z_k^2} + \frac{\varphi(m_j)}{a(\theta)z_j z_k} - \frac{\Phi(m_k)}{z_k^2}, \\ \ddot{V}_{jk} &= \frac{d^2V_r}{dz_j dz_k} = \frac{m_j \varphi(m_j)}{a(\theta)^2 z_j^2 z_k} + \frac{m_k \varphi(m_k)}{a(\theta)^2 z_j z_k^2} - \frac{\varphi(m_j)}{a(\theta) z_j^2 z_k} - \frac{\varphi(m_k)}{a(\theta) z_j z_k^2},\end{aligned}$$

where φ and Φ are the standard normal PDF and CDF, $a(\theta) = \sqrt{2\gamma(x_1 - x_2; \theta)}$, $m_j = 2^{-1}a(\theta) - a(\theta)^{-1} \log(z_j/z_k)$ and $m_k = 2^{-1}a(\theta) - a(\theta)^{-1} \log(z_k/z_j)$.

Moreover, let $V'_r = \partial V_r / \partial \theta$, $\dot{V}'_j = \partial \dot{V}_j / \partial \theta$ and $\dot{V}'_k = \partial \dot{V}_k / \partial \theta$. Then the r th and $(r + m)$ th element of $\tilde{u}(\theta)$ is given by the vector

$$\frac{\dot{V}'_j \dot{V}'_k + \dot{V}_j \dot{V}'_k}{\dot{V}_j \dot{V}_k - \ddot{V}_{jk}} - V'_r.$$

Funding Open access funding provided by Libera Università di Bolzano within the CRUI-CARE Agreement.

Open Access This article is licensed under a Creative Commons Attribution 4.0 International License, which permits use, sharing, adaptation, distribution and reproduction in any medium or format, as long as you give appropriate credit to the original author(s) and the source, provide a link to the Creative Commons licence, and indicate if changes were made. The images or other third party material in this article are included in the article's Creative Commons licence, unless indicated otherwise in a credit line to the material. If material is not included in the article's Creative Commons licence and your intended use is not permitted by statutory regulation or exceeds the permitted use, you will need to obtain permission directly from the copyright holder. To view a copy of this licence, visit <http://creativecommons.org/licenses/by/4.0/>.

References

- Brown, B.M., Resnick, S.I.: Extreme values of independent stochastic processes. *J. Appl. Probab.* **14**(04), 732–739 (1977)
- Castruccio, S., Huser, R., Genton, M.G.: High-order composite likelihood inference for max-stable distributions and processes. *J. Comput. Graph. Stat.* **25**(4), 1212–1229 (2016)
- Cooley, D., Naveau, P., Poncet, P.: Variograms for spatial max-stable random fields. In: *Dependence in probability and statistics*, pp. 373–390. Springer (2006)
- Cox, D.R., Reid, N.: A note on pseudolikelihood constructed from marginal densities. *Biometrika* **91**(3), 729–737 (2004)
- Davison, A.C., Huser, R.: Statistics of extremes. *Annual Review of Statistics and its Application* **2**, 203–235 (2015)
- Davison, A.C., Padoan, S.A., Ribatet, M., et al.: Statistical modeling of spatial extremes. *Stat. Sci.* **27**(2), 161–186 (2012)
- Dombry, C., Engelke, S., Oesting, M.: Asymptotic properties of the maximum likelihood estimator for multivariate extreme value distributions. arXiv:1612.05178 (2016)

- Efron, B., Hastie, T., Johnstone, I., Tibshirani, R., et al.: Least angle regression. *The Annals of statistics* **32**(2), 407–499 (2004)
- Fan, J., Lv, J.: A selective overview of variable selection in high dimensional feature space. *Stat. Sin.* **20**(1), 101 (2010)
- Genton, M.G., Ma, Y., Sang, H.: On the likelihood function of gaussian max-stable processes. *Biometrika* **98**(2), 481–488 (2011)
- Heyde, C.C.: Quasi-likelihood and its application: a general approach to optimal parameter estimation. Springer Science & Business Media, Berlin (2008)
- Huang, Z., Ferrari, D.: Fast and efficient construction of composite estimating equations. ArXiv:1709.03234 Available online at <https://arxiv.org/pdf/1709.03234.pdf> (2017)
- Huser, R., Davison, A.C.: Composite likelihood estimation for the brown–resnick process. *Biometrika* **100**(2), 511–518 (2013)
- Huser, R., Dombry, C., Ribatet, M., Genton, M.G.: Full likelihood inference for max-stable data. *Stat* **8**(1), e218 (2019)
- Kabluchko, Z., Schlather, M., De Haan, L., et al.: Stationary max-stable fields associated to negative definite functions. *The Annals of Probability* **37**(5), 2042–2065 (2009)
- Lindsay, B.G.: Composite likelihood methods. *Contemp. Math.* **80**(1), 221–39 (1988)
- Lindsay, B.G., Yi, G.Y., Sun, J.: Issues and strategies in the selection of composite likelihoods. *Stat. Sin.* **21**(1), 71 (2011)
- Padoan, S.A., Ribatet, M., Sisson, S.A.: Likelihood-based inference for max-stable processes. *J. Am. Stat. Assoc.* **105**(489), 263–277 (2010)
- Ribatet, M.: Spatialextremes: Modelling spatial extremes. <https://CRAN.R-project.org/package=SpatialExtremes>, R package version 2.0-2 (2015)
- Sang, H., Genton, M.G.: Tapered composite likelihood for spatial max-stable models. *Spatial Statistics* **8**, 86–103 (2014)
- Schlather, M., Tawn, J.A.: A dependence measure for multivariate and spatial extreme values: Properties and inference. *Biometrika* **90**(1), 139–156 (2003)
- Smith, R.I.C.H.A.R.D.L.: Max-stable processes and spatial extremes. Unpublished manuscript, University of Northern California (1990)
- Thibaud, E., Aalto, J., Cooley, D.S., Davison, A.C., Heikkinen, J., et al.: Bayesian inference for the brown–resnick process, with an application to extreme low temperatures. *The Annals of Applied Statistics* **10**(4), 2303–2324 (2016)
- Varin, C., Reid, N.M., Firth, D.: An overview of composite likelihood methods. *Stat. Sin.* **21**(1), 5–42 (2011)

Publisher's note Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.