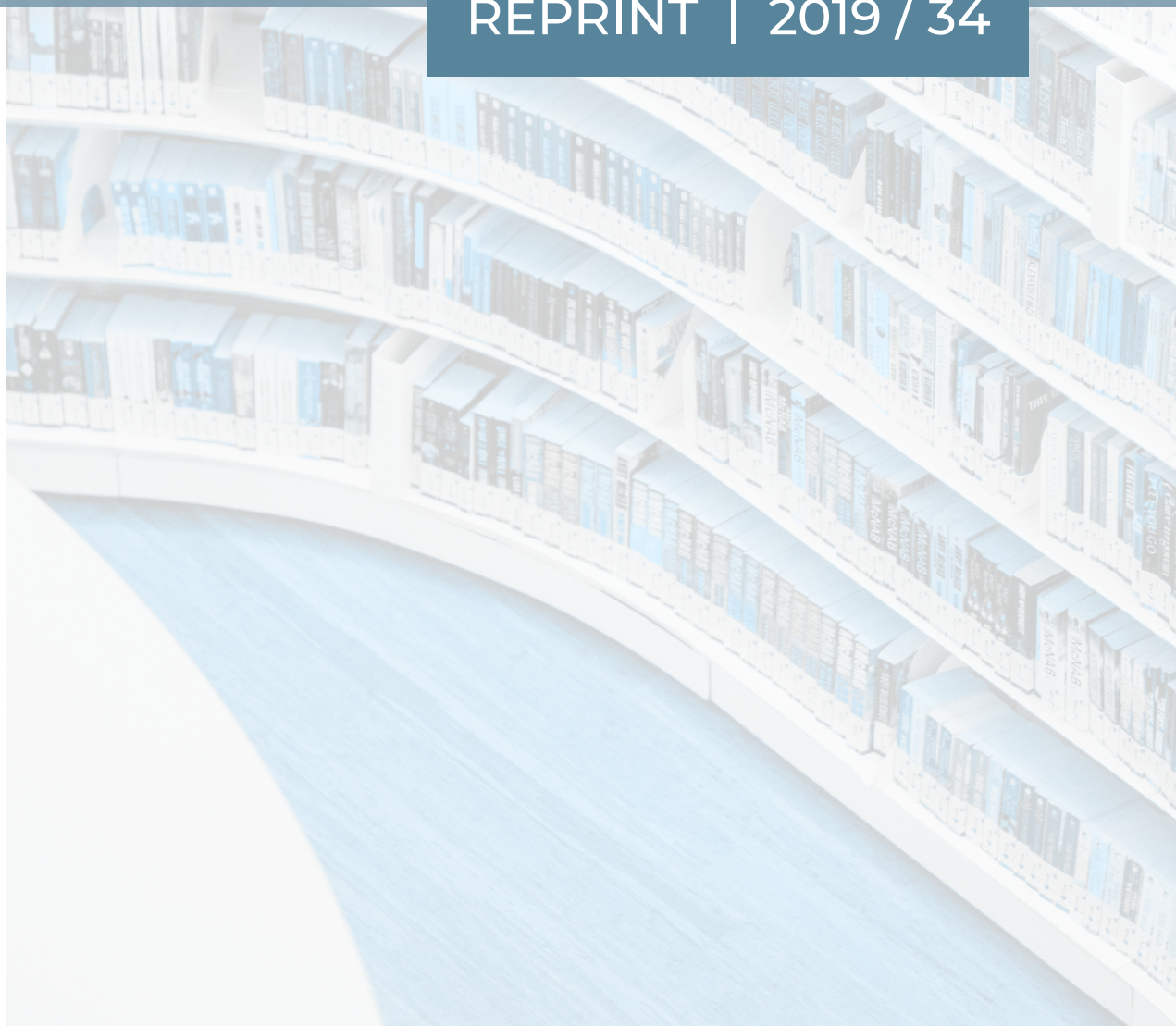
# BAYESIAN MODEL AVERAGING OVER TREE-BASED DEPENDENCE STRUCTURES FOR MULTIVARIATE EXTREMES

Vettori, Sabrina; Huser, Raphaël, Segers Johan and Marc G. Genton

REPRINT | 2019 / 34



UCLouvain

ISBA | LIDAM
Louvain Institute of Data Analysis and Modeling in economics and statistics

# Bayesian model averaging over tree-based dependence structures for multivariate extremes

Sabrina Vettori[1], Raphaël Huser[1], Johan Segers[2] and Marc G. Genton[1]

February 24, 2019

## Abstract

Describing the complex dependence structure of extreme phenomena is particularly challenging. To tackle this issue we develop a novel statistical method that describes extremal dependence taking advantage of the inherent tree-based dependence structure of the max-stable nested logistic distribution, and that identifies possible clusters of extreme variables using reversible jump Markov chain Monte Carlo techniques. Parsimonious representations are achieved when clusters of extreme variables are found to be completely independent. Moreover, we significantly decrease the computational complexity of full likelihood inference by deriving a recursive formula for the likelihood function of the nested logistic model. The method's performance is verified through extensive simulation experiments which also compare different likelihood procedures. The new methodology is used to investigate the dependence relationships between extreme concentrations of multiple pollutants in California and how these concentrations are related to extreme weather conditions. Overall, we show that our approach allows for the representation of complex extremal dependence structures and has valid applications in multivariate data analysis, such as air pollution monitoring, where it can guide policymaking.

**Keywords**: Air pollution; Extreme event; Fast likelihood inference; Nested logistic model; Reversible jump Markov chain Monte Carlo.

[1] King Abdullah University of Science and Technology (KAUST), Computer, Electrical and Mathematical Science and Engineering Division (CEMSE) Thuwal 23955-6900, Saudi Arabia.
E-mails: sabrina.vettori@kaust.edu.sa, raphael.huser@kaust.edu.sa, marc.genton@kaust.edu.sa.

[2] Université catholique de Louvain, Institut de Statistique, Biostatistique et Sciences Actuarielles (ISBA) Louvain-la-Neuve B-1348, Belgium.
E-mail: johan.segers@uclouvain.be.

# 1 Introduction

Estimating the probabilities associated with multivariate extreme phenomena beyond the original observation range requires flexible, yet interpretable, models with sound theoretical underpinning, that can be efficiently fitted to the data. However, these criteria are exponentially more difficult to meet as dimensionality increases; see, e.g., the review papers by Davison *et al.* (2012) and Davison and Huser (2015). For these reasons the notion of "high-dimensionality" in the field of extremes typically invokes far smaller scales than those considered in standard statistics. Indeed, most applications of multivariate Extreme-Value Theory have focused chiefly on relatively low-dimensional cases, although much higher dimensions have been handled in structured, spatial, and spatio-temporal settings (Wadsworth and Tawn, 2012; Huser and Davison, 2014). Moreover, the analysis of extreme events is also very challenging because of the intrinsic lack of data; extreme datasets comprise only the largest observations in a sample, and therefore are generally characterised by small sample sizes, i.e., few temporal replicates.

Recent literature has been focusing on developing innovative statistical methods to simplify the complex dependence structure of multivariate extremes in moderate or large dimensions. Chautru (2015) proposed a non-parametric technique able to identify possibly overlapping clusters of asymptotically dependent extreme variables, thus reducing the complexity of the initial problem. Another approach, proposed by Bernard *et al.* (2013), consists in a clustering technique for block maxima observed at different spatial locations; the new method uses the $F$-madogram "distance" in the $K$-means algorithm. Moreover, Cooley and Thibaud (2018) summarised tail dependence in high dimensions by the decomposition of a matrix of pairwise dependence metrics.

In this work, we contribute to this developing research field by proposing an efficient methodology to describe parsimoniously the dependence between (unstructured) multivariate maxima. The technique may also be adapted to threshold exceedances. Our approach relies on the nested logistic distribution (McFadden, 1978; Tawn, 1990; Coles and Tawn, 1991; Stephenson, 2003), which is max-stable, and thus, asymptotically justified for the description of the whole multivari-

ate extremal dependence structure (de Haan and Resnick, 1977; de Haan, 1984). In particular, thanks to its underlying tree structure, the nested logistic model can describe the dependence within and between distinct clusters of variables using logistic distributions (Gumbel, 1960a,b).

Likelihood-based inference for max-stable distributions is known to be computationally burdensome, especially in high dimensions as the full likelihood is numerically intractable (Castruccio *et al.*, 2016; Bienvenüe and Robert, 2017). One classical solution that dramatically reduces the computational burden but leads to a loss of efficiency consists in using composite likelihood techniques (Varin and Vidoni, 2005; Padoan *et al.*, 2010; Genton *et al.*, 2011; Davison and Gholamrezaee, 2012; Huser and Davison, 2013; Castruccio *et al.*, 2016). The simplified likelihood procedure proposed by Stephenson and Tawn (2005) leads to more efficient inference both computationally and statistically, but may be severely biased in certain cases. Apart from its interpretability, the nested logistic model is also appealing for its inference properties. Here, we derive a recursive formula for the likelihood of the nested logistic model. The formula greatly simplifies computations when dealing with multivariate extremes in moderate or high dimensions without causing any efficiency loss or introducing any additional bias.

Few authors have so far implemented a fully Bayesian inference method for the estimation of extremal dependence; see, e.g., Guillotte *et al.* (2011), Ribatet *et al.* (2012), Reich and Shaby (2012), Sabourin *et al.* (2013), Sabourin and Naveau (2014), Shaby (2014) and Thibaud *et al.* (2016). Our method is based on the reversible jump Markov chain Monte Carlo (RJ-MCMC) algorithm (Green, 1995), which allows us to sample from the posterior distribution of the nested logistic model parameters and simultaneously identify the most likely tree structures (i.e., clustering configurations) governing the extremal dependence. Moreover, instead of providing just one tree estimate, our methodology provides a natural measure of model uncertainty, which is given by the posterior probabilities associated to the tree structure. Model uncertainty is taken into account for predictions using Bayesian model averaging, i.e., we construct the posterior predictive distributions as mixtures of the model-specific distributions weighted by the associ-

ated posterior probabilities. For more details on Bayesian model averaging see, e.g., Hoeting *et al.* (1999). Therefore, we are able to describe complex extremal dependence structures using a mixture of clustering configurations, transcending the partial exchangeability limitations of the nested logistic model while maintaining a moderate number of parameters.

In Section 2, we describe the general Extreme-Value Theory framework and our modeling approach based on the nested logistic model. In Section 3, we introduce our new recursive likelihood formula, as well as alternative likelihood formulations. In Section 4, we describe the proposed RJ-MCMC algorithm, and in Section 5 we verify its performance through an extensive simulation study. In Section 6, we fit our model to multivariate air pollution data to investigate the dependence relations between extreme concentration of air pollutants and to understand how these relate to extreme weather conditions. Finally, we discuss these results in Section 7 with an outlook towards future research.

# 2 Modelling extremal dependence

## 2.1 Extreme-Value Theory framework

Suppose that $\mathbf{Y}_t = (Y_{t;1}, \ldots, Y_{t;D})^\top$, $t = 1, 2, \ldots$, is a time series of independent and identically distributed (i.i.d.) copies of the $D$-dimensional random vector $\mathbf{Y}$, representing $D$ variables of interest having a common distribution function (d.f.) $F$ with margins $F_d(y_d), d = 1, \ldots, D$, and let $\mathbf{M}_n = (M_{n;1}, \ldots, M_{n;D})^\top = (\max_{1\leq t\leq n} Y_{t;1}, \ldots, \max_{1\leq t\leq n} Y_{t;D})^\top$ denote the vector of multivariate componentwise maxima computed over temporal blocks of $n$ observations. We assume that, as $n \to \infty$, for some sequences of vectors $\mathbf{a}_n = (a_{n;1}, \ldots, a_{n;D})^\top \in \mathbb{R}_+^D$ and $\mathbf{b}_n = (b_{n;1}, \ldots, b_{n;D})^\top \in \mathbb{R}^D$, the renormalised vector of componentwise maxima $\mathbf{M}_n^* = \mathbf{a}_n^{-1}(\mathbf{M}_n - \mathbf{b}_n)$ converges in distribution to the random vector $\mathbf{Z}$, with limiting $D$-dimensional extreme-value d.f. $G$ and non-degenerate margins; that is, the distribution $F$ belongs to the max-domain of attraction (MDA) of $G$. Upon marginal standardisation, we may assume unit-Fréchet margins, i.e., $G_d(z_d) = \exp(-1/z_d)$ for

$z_d > 0$, $d = 1, \ldots, D$. The joint distribution of the random vector $\mathbf{Z}$ may be written as

$$\mathrm{P}(\mathbf{Z} \leq \mathbf{z}) = G(\mathbf{z}) = \exp\{-V(\mathbf{z})\}, \qquad V(\mathbf{z}) = \int_{S_D} \max_{1 \leq d \leq D} \left(\frac{\omega_d}{z_d}\right) \mathrm{d}H(\boldsymbol{\omega}), \qquad \mathbf{z} \in \mathbb{R}_+^D, \quad (1)$$

where $V(\mathbf{z})$, called exponent function, is homogeneous of order $-1$, i.e., $V(s\mathbf{z}) = s^{-1}V(\mathbf{z})$ for all $s > 0$, and $H$ is a finite spectral measure on the unit simplex $S_D = \{\boldsymbol{\omega} \in [0,1]^D \colon \sum_{d=1}^D \omega_d = 1\}$ for $\boldsymbol{\omega} = (\omega_1, \ldots, \omega_D)^\top$, satisfying the mean constraints $\int_{S_D} \omega_d \mathrm{d}H(\boldsymbol{\omega}) = 1$, $d = 1, \ldots, D$ (de Haan and Resnick, 1977; de Haan, 1984). Extreme-value distributions (1) are max-stable in the sense that $G^s(\mathbf{z}) = G(s\mathbf{z})$ for all $s > 0$. For more details, see, e.g., Davison and Huser (2015).

A useful summary of the dependence strength between multivariate extremes is the extremal coefficient, proposed by Smith (1990) and defined as $\theta_D = V(\mathbf{1}) \in [1, D]$, where $\mathbf{1}$ is a vector of ones. The coefficient $\theta_D$ decreases as the dependence strength between the margins increases, with $\theta_D = 1$ and $\theta_D = D$ corresponding to complete dependence and independence respectively.

## 2.2 The logistic and nested logistic models

Among max-stable distributions, the oldest and simplest one is the logistic model (Gumbel, 1960a,b), characterised by the exponent function

$$V_l(\mathbf{z} \mid \alpha_0) = \left(\sum_{d=1}^D z_d^{-1/\alpha_0}\right)^{\alpha_0}, \quad \alpha_0 \in (0, 1]. \quad (2)$$

The parameter $\alpha_0$ summarises the dependence strength between the extreme observations $\mathbf{z} = (z_1, \ldots, z_D)^\top \in \mathbb{R}_+^D$. In particular, the cases of complete dependence and independence between the margins correspond to $\alpha_0 \to 0$ and $\alpha_0 = 1$, respectively. Therefore, the logistic model summarises the extremal dependence structure by only one parameter $\alpha_0$, and its components are exchangeable. The extremal coefficient for the logistic distribution is $\theta_D = D^{\alpha_0}$.

Generalising the idea of the logistic model, McFadden (1978) and Tawn (1990), see also Coles and Tawn (1991), proposed the nested logistic distribution, a more flexible multivariate max-stable model that maintains the logistic model's simplicity and interpretability. The nested

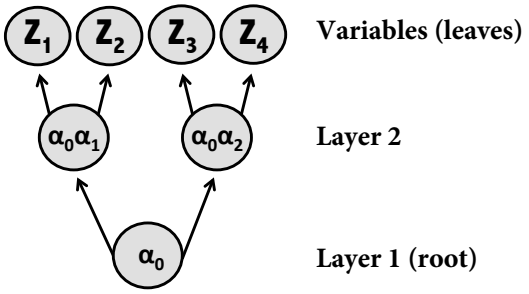



Figure 1: Example of simple two-layer tree structure, summarising the extremal dependence of the vector $\mathbf{Z} = (Z_1, Z_2, Z_3, Z_4)^\top$, where the dependence between the variables $(Z_i, Z_j)^\top$, $i \in \{1, 2\}$, $j \in \{3, 4\}$ is summarised by a logistic distribution with parameter $\alpha_0$, and the dependencies within the pairs of variables $(Z_1, Z_2)^\top$ and $(Z_3, Z_4)^\top$ are summarised by logistic distributions with parameters $\alpha_0\alpha_1$ and $\alpha_0\alpha_2$, respectively.

logistic model implies that the vector $\mathbf{z}$, containing the extreme observations, is split into $K$ homogeneous clusters and it describes the dependence within and between these clusters using the logistic distribution defined in (1) based on (2). It is defined by the exponent function

$$V_{nl}(\mathbf{z} \mid \boldsymbol{\alpha}) = V_l\left\{V_l(\mathbf{z}_1 \mid \alpha_0\alpha_1)^{-1}, \ldots, V_l(\mathbf{z}_K \mid \alpha_0\alpha_K)^{-1} \mid \alpha_0\right\} = \left\{\sum_{k=1}^{K}\left(\sum_{i_k=1}^{D_k} z_{k;i_k}^{-\frac{1}{\alpha_0\alpha_k}}\right)^{\alpha_k}\right\}^{\alpha_0}, \quad (3)$$

where $V_l(\mathbf{z}_k \mid \alpha_0\alpha_k)$ is the logistic exponent function in (2) of the sub-vector $\mathbf{z}_k = (z_{k;1}, \ldots, z_{k;D_k})^\top$ comprising extreme observations belonging to the $k^{\text{th}}$ cluster of dimension $D_k \in \{1, \ldots, D\}$, $k = 1, \ldots, K$, with $D = \sum_{k=1}^{K} D_k$, and where $\boldsymbol{\alpha} = (\alpha_0, \alpha_1, \ldots, \alpha_K)^\top \in (0, 1]^{K+1}$ are the dependence parameters. More precisely, the parameter $\alpha_0$ summarises the dependence strength between the clusters and the product of the parameters $\alpha_0\alpha_k$ summarises the dependence strength within the $k^{\text{th}}$ cluster. The extremal coefficient, which quantifies the effective number of independent variables among the $D$ variables, is equal to $\theta_D = \left(\sum_{k=1}^{K} D_k^{\alpha_k}\right)^{\alpha_0}$ for the nested logistic distribution. The hierarchical structure of the nested logistic model can be represented with a tree, as illustrated in Figure 1. In practice, more complex situations may arise with more clusters and, possibly, an arbitrary number of layers (Stephenson, 2003). For simplicity, in this paper, we limit ourselves to dependence structures with only two layers (i.e., one nesting level).

The nested logistic model can also be defined in terms of nested Archimedean copulas or nested Gumbel copulas (Okhrin *et al.*, 2009; Hofert and Pham, 2013). There exist already a number of papers on inference on such copulas, both parametric and non-parametric, see, e.g.,

Okhrin *et al.* (2013), Segers and Uyttendaele (2014) and Górecki *et al.* (2016).

## 2.3 Bayesian model averaging over tree-based dependence structures

Although the nested logistic distribution defined in Section 2.2 is more flexible than the logistic model, it is still fairly rigid in practice because it assumes that the variables are partially exchangeable, and that any two variables in different clusters have the *same* logistic distribution with dependence parameter $\alpha_0$. To overcome these limitations, we here embed the nested logistic model into the Bayesian framework and assume that the underlying tree structure is random.

Let $\mathcal{G}$ be the set of all possible two-layer trees with $D$ terminal nodes (i.e., leaves), and $\mathcal{T} \in \mathcal{G}$ denote a specific tree (e.g., as in Figure 1). Splitting the observed data $\mathbf{Y}_t = (Y_{t;1}, \ldots, Y_{t;D})^\top$, $t = 1, \ldots, M$, into $N$ blocks with an equal number of observations $n$ (assuming that $M = Nn$), we first extract componentwise maxima data $\mathbf{m}_i = (m_{i;1}, \ldots, m_{i;D})^\top$, $i = 1, \ldots, N$. Then, after transforming the maxima to the unit Fréchet scale, we assume the following hierarchical model:

$$\begin{aligned} \mathbf{m}_i \mid \mathcal{T}, \boldsymbol{\alpha} &\overset{\text{i.i.d.}}{\sim} \text{NestedLogistic}(\boldsymbol{\alpha}; \mathcal{T}) \\ \boldsymbol{\alpha} \mid \mathcal{T} &\sim \pi(\boldsymbol{\alpha}; \mathcal{T}) \\ \mathcal{T} &\sim \pi(\mathcal{T}), \end{aligned}$$

where the model parameters $\boldsymbol{\alpha} = (\alpha_0, \alpha_1, \ldots, \alpha_K)^\top \in (0, 1]^{K+1}$ have independent prior distributions, i.e., $\pi(\boldsymbol{\alpha}; \mathcal{T}) = \prod_{k=0}^{K} \pi(\alpha_k)$. To allow for independence within and between the clusters of variables, we choose the prior distribution for each dependence parameter $\alpha_k$ as a mixture between a point mass at $\alpha_k = 1$ (the upper boundary of the domain of definition), and a uniform distribution in $[0, 1]$. Specifically, for each $k = 0, \ldots, K$, we set $\pi(\alpha_k) = 0.5\delta_1 + 0.5\text{Unif}(0, 1)$, where $\delta_1$ is the Dirac delta function at 1. This allows us to reduce model complexity if certain parameters are found to be exactly equal to one. Notice that the parameter vector $\boldsymbol{\alpha}$ is intrinsically linked to the tree $\mathcal{T}$; in particular, the number of model parameters (i.e., the dimension of the vector $\boldsymbol{\alpha}$) is equal to $K + 1$, the number of non-terminal nodes of the underlying tree

structure $\mathcal{T}$. Finally, we assume that $\mathcal{T}$ has a discrete uniform prior on the (finite) set of all admissible trees, $\mathcal{G}$, i.e., $\pi(\mathcal{T}) = 1/|\mathcal{G}|$. Thus, a priori, all clustering configurations are equally likely. Every fixed tree $\mathcal{T}$ leads to a different nested logistic model for the data with its own vector of parameters $\boldsymbol{\alpha}$. Integrating out $\mathcal{T}$ therefore amounts to *averaging* different nested logistic models, which introduces partial asymmetries.

To fit this complicated model, we develop a reversible jump Markov chain Monte Carlo (RJ-MCMC) algorithm (Green, 1995), which allows us to sample from the posterior distribution of the underlying tree structure $\mathcal{T}$ and the model parameters $\boldsymbol{\alpha}$ (of varying dimension), jumping across different nested logistic models. The resulting RJ-MCMC output can then be used to perform model averaging by marginalizing over $\mathcal{T}$. To implement such an algorithm, the likelihood function of the nested logistic model plays a key role. In Section 3, we show how it can be computed efficiently, and in Section 4, we describe our RJ-MCMC algorithm in more details.

# 3 Full likelihood

## 3.1 General full likelihood formula

Assuming that the vectors of componentwise maxima $\mathbf{m}_i = (m_{i;1}, \ldots, m_{i;D})^\top$, $i = 1, \ldots, N$, follow a multivariate extreme-value distribution with unit Fréchet margins as in (1), the full likelihood function may be expressed in general form as

$$L(\boldsymbol{\alpha} \mid \mathbf{m}_1, \ldots, \mathbf{m}_N) = \prod_{i=1}^{N} \exp\{-V(\mathbf{m}_i \mid \boldsymbol{\alpha})\} \left\{ \sum_{E \in \mathcal{E}} \prod_{S \in E} -\dot{V}_S(\mathbf{m}_i \mid \boldsymbol{\alpha}) \right\}, \tag{4}$$

where $\boldsymbol{\alpha}$ is the vector of unknown dependence parameters, $\mathcal{E}$ denotes the collection of all partitions of $\mathcal{D} = \{1, \ldots, D\}$, $\dot{V}_S$ denotes the partial derivative of the function $V$ in (1) with respect to the variables whose indices lie in $S \subset \mathcal{D}$; see, e.g., Huser *et al.* (2016, 2019). The cardinality of $\mathcal{E}$ equals $B_D$, the Bell number of order $D$ (Graham *et al.*, 1988), meaning that the number of terms to be computed and the storage space required for evaluating the likelihood grow super-exponentially with the dimension $D$ (Castruccio *et al.*, 2016). In the next sections, we

propose two different ways to significantly speed-up computations in high dimensions: either by exploiting the specific structure of the nested logistic model (Section 3.2), or by using the additional information about occurrence times of maxima (Section 3.3).

## 3.2 Recursive full likelihood formula

Similarly to the recursive full likelihood formula for the logistic model derived by Shi (1995), we here develop a recursive formula to efficiently compute the full likelihood function (4) for the nested logistic model (3). Let $\boldsymbol{m}_{i;k;1:D_k}$ be the sub-vector of maxima belonging to the $k^{\text{th}}$ cluster of dimension $D_k$. Recalling that the nested logistic likelihood depends on a particular tree structure $\mathcal{T}$, the recursive likelihood formula may be written as

$$L(\boldsymbol{\alpha} \mid \mathbf{m}_1, \dots, \mathbf{m}_N; \mathcal{T}) = \prod_{i=1}^{N} \exp\{-V_{nl}(\mathbf{m}_i \mid \boldsymbol{\alpha})\} \prod_{i_1=1}^{D_1} m_{i;1;i_1}^{-\frac{1}{\alpha_0\alpha_1}-1} \cdots \prod_{i_K=1}^{D_K} m_{i;K;i_K}^{-\frac{1}{\alpha_0\alpha_k}-1} \sum_{i_1=1}^{D_1} \cdots \sum_{i_K=1}^{D_K} \times \sum_{j=1}^{\sum_{k=1}^{K} i_k} \beta_{i_1;\dots;i_K;j}^{(D_1;\dots;D_K)} V_l(\mathbf{m}_{i;1;1:D_1}^{1/\alpha_0} \mid \alpha_1)^{i_1-\frac{D_1}{\alpha_1}} \cdots V_l(\mathbf{m}_{i;K;1:D_K}^{1/\alpha_0} \mid \alpha_k)^{i_K-\frac{D_K}{\alpha_k}} V_{nl}(\mathbf{m}_i \mid \boldsymbol{\alpha})^{j-\frac{\sum_{k=1}^{K} i_k}{\alpha_0}}, \tag{5}$$

where $V_{nl}(\mathbf{m}_i \mid \boldsymbol{\alpha})$ is defined in (3) and $V_l(\mathbf{m}_{i;k;1:D_k}^{1/\alpha_0} \mid \alpha_k) = \left(\sum_{i_k=1}^{D_k} m_{i;k;i_k}^{-1/\alpha_0\alpha_k}\right)^{\alpha_k}$, $i = 1, \dots, N$, $k = 1, \dots, K$. The coefficients $\beta_{j_1;\dots;j_K;j}^{(D_1;\dots;D_K)}$ can be computed recursively in explicit form reducing the computational complexity to $\mathcal{O}\left(\sum_{k=1}^{K} (D_1 + \cdots + D_k)\, D_1 \cdots D_{k-1} D_k^2\right) \ll B_D$; the proof of (5) and the expression for $\beta_{j_1;\dots;j_K;j}^{(D_1;\dots;D_K)}$ are given in Appendix A.3. When all the clusters are of equal size, *i.e.,* $D_k = D_1$ for all $k = 2, \dots, K$, the complexity is $\mathcal{O}\left(K \sum_{k=1}^{K} D_1^{k+2}\right)$. Therefore, the computational time needed to compute the full likelihood using the recursive formula is polynomial in terms of the number of variables $D$, but exponential-linear in terms of the number of clusters $K$, suggesting that in practice it might be convenient to prevent $K$ from being large when $D$ is large. The formula (5) holds for the nested logistic model (3) constructed from an underlying two-layer tree structure. Recursive formulas for more complex tree structures with additional layers may be derived similarly, but they have a higher computational complexity.

### 3.3 Stephenson–Tawn likelihood

Stephenson and Tawn (2005) developed a simplified likelihood approach for multivariate extreme-value data, which uses additional information about the occurrence times of maxima. Precisely, using the same notation as in Section 3.1, for each $i = 1, \ldots, N$, let $E_i \in \mathcal{E}$ denote the partition of $\mathcal{D} = \{1, \ldots, D\}$ grouping the maxima $m_{i;1}, \ldots, m_{i;D}$ with identical occurrence times. For instance, for $D = 3$, $E_i = \{1, \{2, 3\}\}$ indicates that, in the $i^{\text{th}}$ block, the maxima $m_{i;2}$ and $m_{i;3}$ occurred simultaneously, but separately from $m_{i;1}$. The Stephenson–Tawn likelihood is

$$L\{\boldsymbol{\alpha} \mid (\mathbf{m}_1, E_1), \ldots, (\mathbf{m}_N, E_N)\} = \prod_{i=1}^{N} \exp\{-V(\mathbf{m}_i \mid \boldsymbol{\alpha})\} \left\{ \prod_{S \in E_i} -\dot{V}_S(\mathbf{m}_i \mid \boldsymbol{\alpha}) \right\}, \tag{6}$$

where $\boldsymbol{\alpha}$ is the vector of unknown parameters and $\dot{V}_S$ denotes the partial derivative of $V$ in (1) with respect to the variables in $S$. Therefore, using the partitions $E_i$, $i = 1, \ldots, N$ leads to a more efficient inference, both computationally and statistically, which comes at the price of introducing estimation bias; see Huser *et al.* (2016), Thibaud *et al.* (2016) and our simulations in Section 5. The Stephenson–Tawn likelihood replaces the *asymptotic* partitions with the observed, *sub-asymptotic* ones, which might be misspecified and cause bias for finite block sizes $n$. Wadsworth (2015) proposed a bias reduction technique, which remains intensive in case of strong dependence or large dimensions. To further speed up the computation of (6), we derived a recursive formula to compute the partial derivatives of the nested logistic model exponent function. Let $\boldsymbol{m}_{i;k;1:d_k}$ be the sub-vector of the first $1 \leq d_k \leq D_k$ components of the maxima vector belonging to the $k^{\text{th}}$ cluster. Partial derivatives of the nested logistic exponent function $V_{nl}$ with respect to the variables $\boldsymbol{m}_{i;k;1:d_k}$, $k = 1, \ldots, \kappa$, within $1 \leq \kappa \leq K$ clusters, may be expressed as

$$\begin{aligned} \frac{\partial^{\sum_{k=1}^{\kappa} d_k} V_{nl}(\mathbf{m}_i \mid \boldsymbol{\alpha})}{\partial \prod_{k=1}^{\kappa} \mathbf{m}_{i;k;1:d_k}} &= \prod_{i_1=1}^{d_1} m_{i;1;i_1}^{-\frac{1}{\alpha_0 \alpha_1} - 1} \cdots \prod_{i_\kappa=1}^{d_\kappa} m_{i;\kappa;i_\kappa}^{-\frac{1}{\alpha_0 \alpha_\kappa} - 1} \\ &\times \sum_{i_1=1}^{d_1} \cdots \sum_{i_\kappa=1}^{d_\kappa} \gamma_{i_1;\ldots;i_\kappa}^{(d_1;\ldots;d_\kappa)} V_l(\mathbf{m}_{i;1;1:d_1}^{1/\alpha_0} \mid \alpha_1)^{i_1 - \frac{d_1}{\alpha_1}} \cdots V_l(\mathbf{m}_{i;\kappa;1:d_\kappa}^{1/\alpha_0} \mid \alpha_\kappa)^{i_\kappa - \frac{d_\kappa}{\alpha_\kappa}} V_{nl}(\mathbf{m}_i \mid \boldsymbol{\alpha})^{1 - \frac{\sum_{k=1}^{\kappa} i_k}{\alpha_0}}, \end{aligned} \tag{7}$$

which can then be plugged into (6) leading to the likelihood $L\{\boldsymbol{\alpha} \mid (\mathbf{m}_1, E_1), \ldots, (\mathbf{m}_N, E_N); \mathcal{T}\}$ (now emphasizing that it depends on a tree structure $\mathcal{T}$). The total complexity for computing the recursive coefficients $\gamma_{i_1;\ldots;i_\kappa}^{(d_1;\ldots;d_\kappa)}$ is $\mathcal{O}\left(\sum_{k=1}^{\kappa} d_1 \cdots d_{k-1} d_k^2\right)$. If all clusters are of equal size, i.e., $d_k = d_1$, for all $k = 2, \ldots, \kappa$, then $\mathcal{O}\left(\sum_{k=1}^{\kappa} d_1^{k+1}\right)$. The proof of (7) and the expression for $\gamma_{i_1;\ldots;i_\kappa}^{(d_1;\ldots;d_\kappa)}$ are given in the Supplementary Material.

# 4 Inference by reversible jump Markov chain Monte Carlo

## 4.1 General considerations

To fit the nested logistic model with an unknown tree structure $\mathcal{T}$, we exploit the reversible jump MCMC algorithm (Green, 1995) (RJ-MCMC). This algorithm allows us to sample from the posterior distribution of the model parameters $\boldsymbol{\alpha} = (\alpha_0, \ldots, \alpha_K)^\top \in (0, 1]^{K+1}$ (of possibly varying dimensions) and the tree $\mathcal{T} \in \mathcal{G}$ itself, and thus, as a by-product, it can identify the most likely tree structures (i.e., clustering configurations) that best describe the componentwise maxima data. Moreover, instead of describing the data with a single, fixed tree structure, our approach accounts for model uncertainty and allows to improve the model's predictive performance using Bayesian model averaging. In particular, the final posterior predictive distributions are obtained as an average of the posterior distributions associated with each of the nested logistic models (i.e., each of the trees) considered by the RJ-MCMC algorithm weighted by their posterior probabilities, calculated as the proportion of time the chain has spent on each of the trees.

The joint posterior distribution $\pi(\boldsymbol{\alpha}, \mathcal{T} \mid \mathbf{m}_1, \ldots, \mathbf{m}_N)$ is proportional to

$$\begin{aligned} \pi(\boldsymbol{\alpha}, \mathcal{T} \mid \mathbf{m}_1, \ldots, \mathbf{m}_N) &\propto L(\boldsymbol{\alpha} \mid \mathbf{m}_1, \ldots, \mathbf{m}_N; \mathcal{T}) \pi(\boldsymbol{\alpha}; \mathcal{T}) \pi(\mathcal{T}) \\ &\propto L(\boldsymbol{\alpha} \mid \mathbf{m}_1, \ldots, \mathbf{m}_N; \mathcal{T}) \prod_{k=0}^{K} \pi(\alpha_k), \end{aligned} \tag{8}$$

where $L(\boldsymbol{\alpha} \mid \mathbf{m}_1, \ldots, \mathbf{m}_N; \mathcal{T})$ is either the full (recursive) likelihood (5) or the Stephenson–Tawn likelihood (6) computed using (7) (where the partitions $E_i$, $i = 1, \ldots, N$ have been dropped here for simplicity), $\pi(\boldsymbol{\alpha}; \mathcal{T}) = \prod_{k=0}^{K} \pi(\alpha_k)$ is the prior defined in Section 2.3 and $\pi(\mathcal{T}) = 1/|\mathcal{G}|$ is the

discrete uniform prior on the space of two-layer tree structures, $\mathcal{G}$; recall Section 2.3.

In our algorithm, we build an irreducible and ergodic Markov chain on the domain of admissible combinations of parameters $\boldsymbol{\alpha}$ and trees $\mathcal{T}$ (i.e., the product space $(0,1]^{K+1} \times \mathcal{G}$), in a way to ensure proper convergence to the joint posterior distribution (8). At each iteration, we first sequentially update each parameter $\alpha_k \in (0,1]$ *conditional* on the current tree structure $\mathcal{T}^c$ and the other model parameters, and then we update the tree $\mathcal{T}$ given the current parameter vector $\boldsymbol{\alpha}^c$. In Section 4.2, we introduce the Metropolis–Hastings (Hastings, 1970) algorithm that we use to update the model parameters $\boldsymbol{\alpha}$ conditional on the tree. In Section 4.3, we explain how to move across tree structures $\mathcal{T}$ given the model parameters.

## 4.2 Updating the model parameters using Metropolis–Hastings

We here explain how to update the model parameters $\boldsymbol{\alpha} = (\alpha_0, \alpha_1, \ldots, \alpha_K)^\top$ *conditional* on the current tree structure $\mathcal{T}^c$. At each MCMC iteration, we sequentially propose candidate parameter values $\alpha_k^*$ for each parameter $\alpha_k$ based on a proposal distribution $q(\alpha_k^* \mid \alpha_k^c)$, where $\alpha_k^c$ denotes the current value of the parameter $\alpha_k$. As we choose the prior $\pi(\alpha_k)$ as a mixture between a point mass at 1 and a uniform distribution in $[0,1]$ (recall Section 2.3), the proposal distribution $q$ also needs to reflect this feature, in order to explore the whole space of admissible parameters. Therefore, we define the proposal distribution as follows

$$q(\alpha_k^* \mid \alpha_k^c) = \begin{cases} 0.5\delta_1 + 0.5p(\alpha_k^* \mid \alpha_k^c), & \text{if } \alpha_k^c \neq 1 \text{ and } 1 \in [\alpha_k^c - \varepsilon_k, \alpha_k^c + \varepsilon_k]; \\ p(\alpha_k^* \mid \alpha_k^c), & \text{if } \alpha_k^c = 1 \text{ or } 1 \notin [\alpha_k^c - \varepsilon_k, \alpha_k^c + \varepsilon_k], \end{cases}$$

where $\delta_1$ denotes the Dirac delta function at 1, $\varepsilon_k$ is a tuning parameter, and $p(\alpha_k^* \mid \alpha_k^c)$ is the density of a uniform random variable with boundaries $\{\max(0, \alpha_k^c - \varepsilon_k), \min(\alpha_k^c + \varepsilon_k, 1)\}$. The parameter $\alpha_k^*$ is then accepted with probability

$$\min\left\{\frac{\pi(\boldsymbol{\alpha}^*, \mathcal{T}^c \mid \mathbf{m}_1, \ldots, \mathbf{m}_N) q(\alpha_k^c \mid \alpha_k^*)}{\pi(\boldsymbol{\alpha}^c, \mathcal{T}^c \mid \mathbf{m}_1, \ldots, \mathbf{m}_N) q(\alpha_k^* \mid \alpha_k^c)}, 1\right\} = \min\left\{\frac{L(\boldsymbol{\alpha}^* \mid \mathbf{m}_1, \ldots, \mathbf{m}_N; \mathcal{T}^c) \pi(\alpha_k^*) q(\alpha_k^c \mid \alpha_k^*)}{L(\boldsymbol{\alpha}^c \mid \mathbf{m}_1, \ldots, \mathbf{m}_N; \mathcal{T}^c) \pi(\alpha_k^c) q(\alpha_k^* \mid \alpha_k^c)}, 1\right\}.$$

If accepted, then $\alpha_k^c$ is replaced with $\alpha_k^*$, otherwise it keeps its current value. Adaptive methodologies for choosing the tuning parameters $\varepsilon_k$ allow for efficient simulations even in high dimensions. We update $\varepsilon_k$ every 100 iterations during the burn-in, ensuring that the acceptance rate remains between 20% and 50% to guarantee well-mixing properties of the chain. Then, we restart the algorithm with fixed $\varepsilon_k$ values. More details are available in the Supplementary Material.

### 4.3 Moving across tree structures using reversible jump MCMC

Any proposed transition from the current tree $\mathcal{T}^c$, with parameters $\boldsymbol{\alpha}^c$ of dimension $\dim(\boldsymbol{\alpha}^c)$, to a proposed tree $\mathcal{T}^*$, with parameters $\boldsymbol{\alpha}^*$ of dimension $\dim(\boldsymbol{\alpha}^*)$, must be reversible. Therefore, if the transition involves a dimension change, a random auxiliary variable $\mathbf{u}^c$ is generated such that, when moving from the current state $\mathbf{x}^c = (\boldsymbol{\alpha}^c, \mathbf{u}^c)$ to the proposed state $\mathbf{x}^* = (\boldsymbol{\alpha}^*, \mathbf{u}^*)$, the dimension matching condition $\dim(\mathbf{x}^c) = \dim(\mathbf{x}^*)$ is satisfied and the mapping $g_{\mathbf{x}^c \to \mathbf{x}^*}$ is a bijection (Green, 1995). Throughout, we assume a uniform prior distribution on the space of valid two-layer tree structures, although more efficient algorithms with informative priors might be designed. The proposed transition is accepted based on the acceptance probability

$$\min\left\{\frac{\pi(\boldsymbol{\alpha}^*, \mathcal{T}^* \mid \mathbf{m}_1, \ldots, \mathbf{m}_N) q(\mathbf{u}^c, \boldsymbol{\alpha}^c \mid \mathbf{u}^*, \boldsymbol{\alpha}^*) \pi_{\mathbf{x}^c \to \mathbf{x}^*}}{\pi(\boldsymbol{\alpha}^c, \mathcal{T}^c \mid \mathbf{m}_1, \ldots, \mathbf{m}_N) q(\mathbf{u}^*, \boldsymbol{\alpha}^* \mid \mathbf{u}^c, \boldsymbol{\alpha}^c) \pi_{\mathbf{x}^* \to \mathbf{x}^c}} \left|\frac{\partial(\boldsymbol{\alpha}^*, \mathbf{u}^*)}{\partial(\boldsymbol{\alpha}^c, \mathbf{u}^c)}\right|, 1\right\},$$

where $\mathbf{u}^*$ corresponds to the auxiliary variable generated to meet the reversibility condition, $\pi(\boldsymbol{\alpha}^*, \mathcal{T}^* \mid \mathbf{m}_1, \ldots, \mathbf{m}_N)$ is the posterior distribution evaluated at the state $\mathbf{x}^*$ (i.e., for the tree $\mathcal{T}^*$), $q(\mathbf{u}^*, \boldsymbol{\alpha}^* \mid \mathbf{u}^c, \boldsymbol{\alpha}^c)$ is the proposal distribution for moving from the state $\mathbf{x}^c$ to the state $\mathbf{x}^*$ and $\pi_{\mathbf{x}^c \to \mathbf{x}^*}$ is the probability of choosing such a move type (i.e., from the current tree $\mathcal{T}^c$ to the proposed tree $\mathcal{T}^*$); $\mathbf{u}^c$, $\pi(\boldsymbol{\alpha}^c, \mathcal{T}^c \mid \mathbf{m}_1, \ldots, \mathbf{m}_N)$, $q(\mathbf{u}^c, \boldsymbol{\alpha}^c \mid \mathbf{u}^*, \boldsymbol{\alpha}^*)$ and $\pi_{\mathbf{x}^* \to \mathbf{x}^c}$ correspond to the reverse move counterparts and $\left|\frac{\partial(\boldsymbol{\alpha}^*, \mathbf{u}^*)}{\partial(\boldsymbol{\alpha}^c, \mathbf{u}^c)}\right|$ is the Jacobian of the transformation $g_{\mathbf{x}^c \to \mathbf{x}^*}$. In practice, there might be no need to generate any auxiliary variable in one direction or the other.

We implement three different move types within the reversible jump MCMC algorithm in order to explore the whole space of admissible trees. The moves are defined below and represented

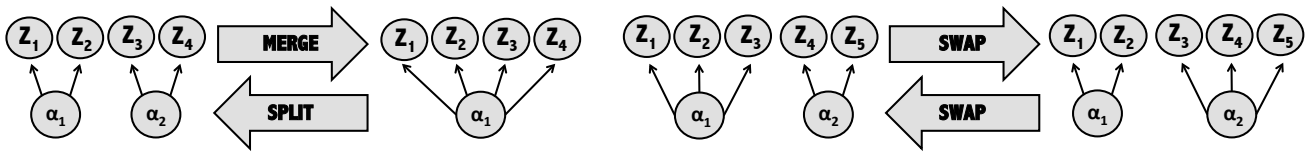



Figure 2: Illustration of the reversible split, merge and swap moves implemented in the TM-MCMC algorithm.

in Figure 2. At each iteration of the reversible jump MCMC algorithm, each move type is chosen with probability $\pi_{\mathbf{x}^c \to \mathbf{x}^*} = 1/3$. Given that our approach aims at describing extremal dependence using a mixture of tree structures, we henceforth call it the Tree Mixture (TM)-MCMC algorithm.

**Split Move:** The split move consists in splitting an existing cluster at random. It is the inverse of the merge move defined below, and the transition implies a dimension change. More precisely, the current state $\mathbf{x}^c = (\alpha_1^c, u^c)^\top$ is mapped to the proposed state as $\mathbf{x}^* = (\alpha_1^* = \alpha_1^c + u^c, \alpha_2^* = \alpha_1^c - u^c)^\top$, where $u^c$ is an auxiliary variable. The proposal distribution for the split move $q(u^c, \alpha_1^c \mid \alpha_1^*, \alpha_2^*)$ is defined by assuming that $u^c$ is a uniform random variable with boundaries chosen such that $\{\max(0, \alpha_1^c - \eta) \leq \alpha_1^*, \alpha_2^* \leq \min(\alpha_1^c + \eta, 1)\}$, for some $\eta \in [0, 1]$. The constant $\eta$ should, in practice, be chosen in order to ensure that the chain is well-mixing; in our case we fix $\eta = 0.4$. Here, the Jacobian reduces to $\left|\frac{\partial(\alpha_1^*, \alpha_2^*)}{\partial(\alpha_1^c, u^c)}\right| = 2$.

**Merge Move:** The merge move consists in merging two clusters at random. It is the inverse of the split move, implying a dimension change transition from the state $\mathbf{x}^c = (\alpha_1^c, \alpha_2^c)^\top$ to $\mathbf{x}^* = (\alpha_1^* = (\alpha_1^c + \alpha_2^c)/2, u^* = (\alpha_1^c - \alpha_2^c)/2)^\top$. In this case there is no need to generate any auxiliary random variable and the Jacobian reduces to $\left|\frac{\partial(\alpha_1^*, u^*)}{\partial(\alpha_1^c, \alpha_2^c)}\right| = \frac{1}{2}$.

**Swap Move:** The swap move consists in exchanging some variables from one cluster to another cluster at random. This move type is self-reversible because transitioning from the current state $\mathbf{x}^c = (\alpha_1^c, \alpha_2^c)^\top$ to the proposed state $\mathbf{x}^* = (\alpha_1^*, \alpha_2^*)^\top$ does not involve any parameter dimension change and therefore the Jacobian is simply equal to one.

In our implementation of this algorithm, the initial tree configuration groups all variables in the same cluster, so that the order in which the variables appear in the dataset does not matter. More details on the algorithm are available in the Supplementary Material.

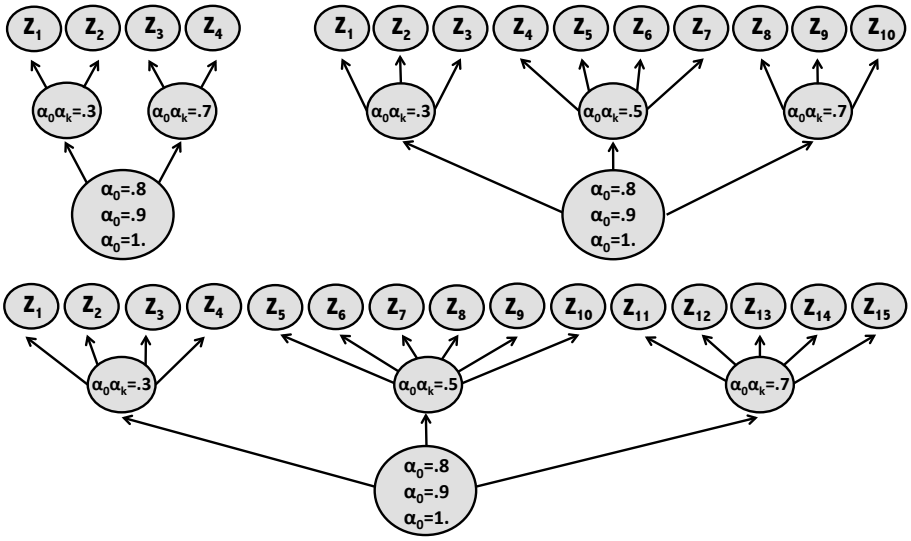



Figure 3: Dependence structure configurations used for the data simulation when $D = 4$ (top left), $D = 10$ (top right) and $D = 15$ (bottom).

# 5 Numerical experiments

## 5.1 Simulation settings

The TM-MCMC algorithm performance is tested on 1000 independent simulated datasets of various dimensions $D = 4, 10, 15$, imposing the dependence structures represented in Figure 3. It is important to bear in mind that multivariate extremes applications are normally carried out for relatively low dimensions. Here we simulate data up to dimensions $D = 15$, which is considered fairly large in this type of applications. The data are generated from: the nested logistic distribution (3) as explained by Stephenson (2003); a nested Archimedean copula with unit Fréchet margins whose distribution is in the max-domain of attraction (MDA) of the nested logistic distribution; and the Student-$t$ copula with unit Fréchet margins, whose limit max-stable distribution is not the nested logistic model, but rather the extremal $t$ model (Opitz, 2013), thus implying a more severe model misspecification. More specifically, data in the MDA of the nested logistic distribution are sampled using the outer power nested Clayton copula with one nesting level and $K$ child copulas (clusters), i.e.,

$$C(\mathbf{u}) = C\{C(\mathbf{u}_1; \varphi_1), \ldots, C(\mathbf{u}_K; \varphi_K); \varphi_0\}, \tag{9}$$

where $\mathbf{u} = (\mathbf{u}_1^\top, \ldots, \mathbf{u}_K^\top)^\top \in [0,1]^D$ and $\varphi_k = (1-t)^{-1/\theta_k}$, $k = 0, 1, \ldots, K$, are Archimedean generators of the Clayton's family (Hofert and Pham, 2013). The parameters $\theta_k \in [1, \infty)$ fulfill the sufficient nesting condition if $\theta_0 \leq \theta_k$, $k = 1, \ldots, K$ and are fixed according to $\theta_0 = 1/\alpha_0$ and $\theta_k = 1/(\alpha_0\alpha_k)$ with $\alpha_0, \alpha_k$, $k = 1, 2, 3$, specified in Figure 3. Sampling from such copulas is explained by Hofert (2011). More details about sampling nested Archimedean copulas using `R` can be found in Hofert and Martin (2011). Student-$t$ data are generated from the multivariate Student-$t$ distribution with 10 degrees of freedom, zero mean vector and covariance matrix $\mathbf{\Sigma}_{D\times D}(\rho_0)$ as specified below:

$$\mathbf{\Sigma}_{4\times4}(\rho_0) = \begin{bmatrix} \mathbf{A}_2(\rho_1) & \rho_0\mathbf{1}_{2\times2} \\ \rho_0\mathbf{1}_{2\times2} & \mathbf{A}_2(\rho_3) \end{bmatrix},$$

$$\mathbf{\Sigma}_{10\times10}(\rho_0) = \begin{bmatrix} \mathbf{A}_3(\rho_1) & \rho_0\mathbf{1}_{3\times4} & \rho_0\mathbf{1}_{3\times3} \\ \rho_0\mathbf{1}_{4\times3} & \mathbf{A}_4(\rho_2) & \rho_0\mathbf{1}_{4\times3} \\ \rho_0\mathbf{1}_{3\times3} & \rho_0\mathbf{1}_{3\times4} & \mathbf{A}_3(\rho_3) \end{bmatrix}, \mathbf{\Sigma}_{15\times15}(\rho_0) = \begin{bmatrix} \mathbf{A}_4(\rho_1) & \rho_0\mathbf{1}_{4\times6} & \rho_0\mathbf{1}_{4\times5} \\ \rho_0\mathbf{1}_{6\times4} & \mathbf{A}_6(\rho_2) & \rho_0\mathbf{1}_{6\times5} \\ \rho_0\mathbf{1}_{5\times4} & \rho_0\mathbf{1}_{5\times6} & \mathbf{A}_5(\rho_3) \end{bmatrix},$$

where $\mathbf{1}_{n\times m}$ is the $n\times m$ matrix of ones and $\mathbf{A}_n(\rho_k) = (1-\rho_k)\mathbf{I}_n + \rho_k\mathbf{1}_{n\times m}$, $k = 1, 2, 3$, with $\mathbf{I}_n$ being the $n \times n$ identify matrix and $\rho_k = 0.98, 0.94, 0.86$. The simulation is repeated for different values of $\rho_0 = 0.77, 0.62, 0$. The coefficients $\rho_0$ and $\rho_k$, $k = 1, 2, 3$, are chosen such that the pairwise extremal coefficients $\theta_2$ for the limiting extremal $t$ distribution match the extremal coefficients calculated for the nested logistic model, with respective parameters $\alpha_0$ and $\alpha_0\alpha_k$, $k = 1, 2, 3$, specified in Figure 3, except for $\rho_0 = 0$. Indeed, for $\rho_0 = 0$, $\theta_2 = 1.99$, only approximately 2, which corresponds to the complete independence case of $\alpha_0 = 1$. The simulation parameters used in the numerical experiments are summarised in Table 1. For the results presented below in Section 5.2, we identify for simplicity the Student-$t$ correlation parameters, $\rho_0, \rho_k$, to their analogue in terms of the nested logistic distribution, $\alpha_0, \alpha_0\alpha_k$.

For each simulation setting under model misspecification we generate datasets of sample sizes $M = 10000, 20000, 40000$ and extract $N = 100, 200, 400$ componentwise maxima with blocks of size $n = 100$, respectively. The simulation results are obtained considering $R = 15000$ iterations, including $R/5 = 3000$ burn-in iterations. We do not show simulation results for data sam-

Table 1: Dependence parameters used in the simulation experiments. From left to right, the coefficients of $\alpha_0$ and $\rho_0$ represent the cases from weak dependence to complete or near independence between the clusters, and the coefficients $\alpha_0\alpha_k$ and $\rho_k$ represent the cases of strong, mild and weak dependence within the clusters, respectively.

| | MDA data | | | | Student-$t$ data | | | |
|---|---|---|---|---|---|---|---|---|
| $D=4$ | $\alpha_0 = 0.8, 0.9, 1$ | $\alpha_0\alpha_k = 0.3$ | $\alpha_0\alpha_k = 0.7$ | | $\rho_0 = 0.77, 0.62, 0$ | $\rho_k = 0.98$ | $\rho_k = 0.86$ | |
| $D=10$ | $\alpha_0 = 0.8, 0.9, 1$ | $\alpha_0\alpha_k = 0.3$ | $\alpha_0\alpha_k = 0.5$ | $\alpha_0\alpha_k = 0.7$ | $\rho_0 = 0.77, 0.62, 0$ | $\rho_k = 0.98$ | $\rho_k = 0.94$ | $\rho_k = 0.86$ |
| $D=15$ | $\alpha_0 = 0.8, 0.9, 1$ | $\alpha_0\alpha_k = 0.3$ | $\alpha_0\alpha_k = 0.5$ | $\alpha_0\alpha_k = 0.7$ | $\rho_0 = 0.77, 0.62, 0$ | $\rho_k = 0.98$ | $\rho_k = 0.94$ | $\rho_k = 0.86$ |

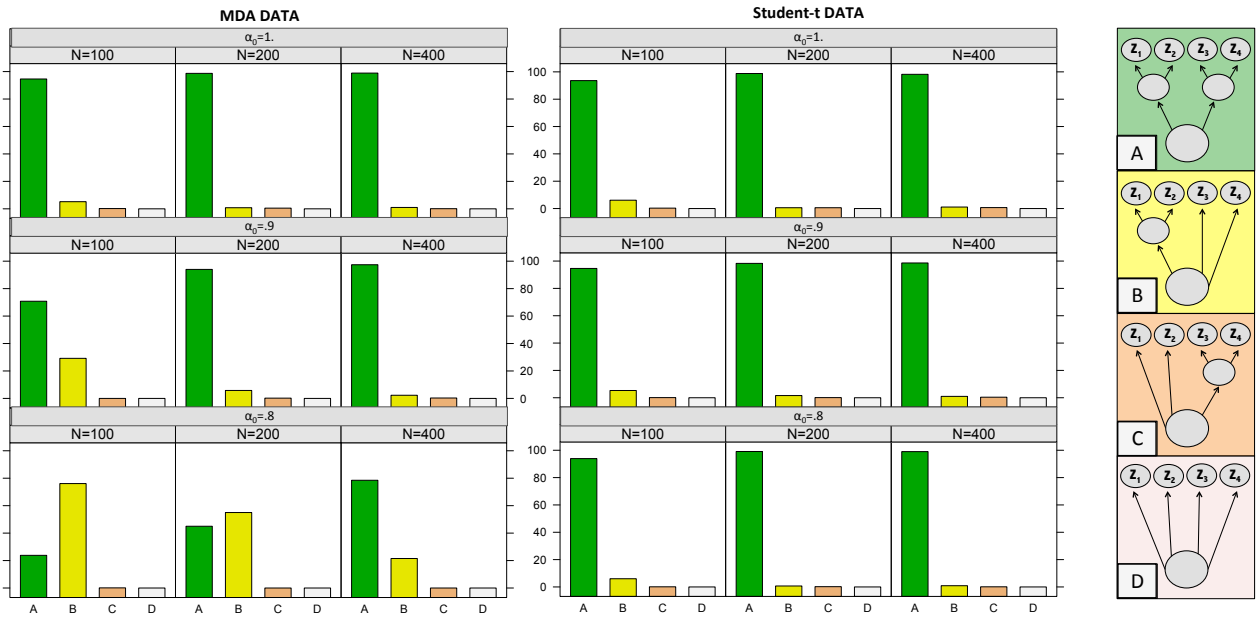



Figure 4: The histograms (left) report the proportion of times that a specific tree structure (right) is identified as the most likely one across $B = 1000$ TM-MCMC algorithm chains after $R = 15000$ iterations for data generated from the copula in (9) and from the Student-$t$ copula with $D = 4$ and tree structure comprising two clusters of variables (see the top-left tree in Figure 3). The recursive formula (5) was used for the likelihood evaluation. The correct tree structure is tree A (green).

pled from the nested logistic distribution (3) as they lead to the same conclusion as simulation experiments conducted for data simulated from the nested outer power Clayton copula (9).

## 5.2 Summary of simulation results

The histograms in Figure 4 indicate the proportion of times that each tree structure (A, B, C and D) is identified as the most likely clustering configuration representing the data across $B = 1000$ TM-MCMC algorithm chains, when the recursive likelihood formula (5) is used in each likelihood evaluation, considering data generated from both the copula in (9) and the Student-$t$ distribution. Generally, our method correctly identifies the true tree structure, represented by

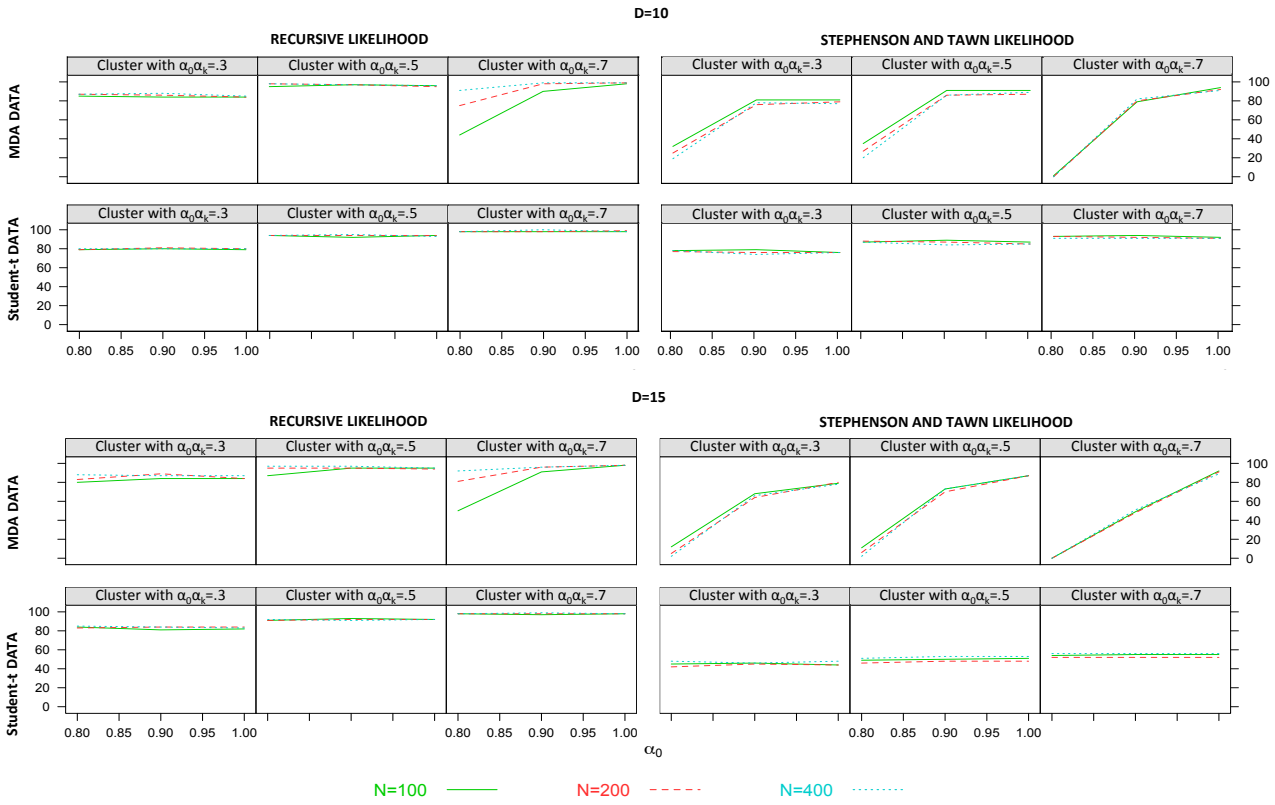



Figure 5: The true positive rate calculated for each cluster is plotted against different values of the parameter $\alpha_0$ (abscissa) for different sample sizes $N$ (colours) obtained by applying the TM-MCMC algorithm with $R = 15000$ iterations over $B = 1000$ replicates for data generated from the copula in (9) and from the Student-$t$ copula with $D = 10, 15$ and tree structure comprising three clusters of variables (see the bottom tree in Figure 3) implementing either the recursive formula (5) (left) or the Stephenson–Tawn (6) likelihood (right) using (7), respectively.

tree A (green), as the most likely clustering configuration of componentwise maxima describing the data dependence structure. Moreover, the method identifies at least the cluster characterised by strong dependence strength in tree B (yellow) across all chains. When the data are generated from the copula (9), the method's performance improves as the value of the $\alpha_0$ parameter and the sample size increase; this means that the between-cluster dependence strength plays an important role in identifying the true dependence structure. When data are simulated from the Student-$t$ distribution, the "true" tree structure is mostly identified by our method for any sample sizes, suggesting that our inference approach also works well in misspecified settings.

Figure 5 displays the proportion of times the true componentwise maxima partition was identified by applying the TM-MCMC algorithm, also called true positive rates, when using either the recursive formula (5) or the Stephenson–Tawn likelihood (6). The true positive rates

are plotted against different values of the parameter $\alpha_0$ for sample sizes $N = 100, 200, 400$. The data are simulated from the copula (9) and from the Student-$t$ distribution with $D = 10, 15$. In accordance with the findings for $D = 4$, the method performs reasonably well, identifying all clusters more than 80% of the time for all sample sizes when using the recursive likelihood. The effect of misspecification is more apparent in larger dimensions, but remains quite weak overall. In contrast, the true positive rate is generally lower for all clusters when the Stephenson–Tawn likelihood is implemented, and it decreases as the between-cluster dependence increases.

The computational time generally ranges from a few seconds to a few minutes depending on the running time for individual likelihood calculations. In the simplest case with $D = 4$ and $N = 100$, a single likelihood evaluation takes around 0.05 seconds using either the recursive likelihood formula or the Stephenson–Tawn likelihood formula, whereas in the most complicated case with $D = 15$ and $N = 400$ a single likelihood evaluation takes around 12 seconds using the recursive likelihood and less than 2 seconds using the approximate Stephenson–Tawn likelihood.

# 6 Application to air pollution data

## 6.1 Air pollutants

Air pollution has various negative effects on human health, ranging from respiratory illnesses to premature death, see, e.g., Peden (2001), Brunekreef and Holgate (2002) and Kampa and Castanas (2008), and causes serious global environmental issues such as global warming and ozone depletion, see, e.g., Murphy *et al.* (1999) and Seinfeld and Pandis (2016). Generally, regional air quality is affected by topography and weather, but also by emission sources. Ground level ozone ($O_3$) is formed by chemical reactions of volatile organic compounds (VOCs) and nitrogen dioxide ($NO_2$), in the presence of heat and sunlight (Jacob and Winner, 2009). $NO_2$, like its precursor nitric oxide (NO), and carbon monoxide (CO) are mostly created by the combustion of fossil fuels in power plants and automobile engines (Cofala *et al.*, 2007). VOCs include a large variety of both natural and artificial chemical species, such as methane and isoprene.

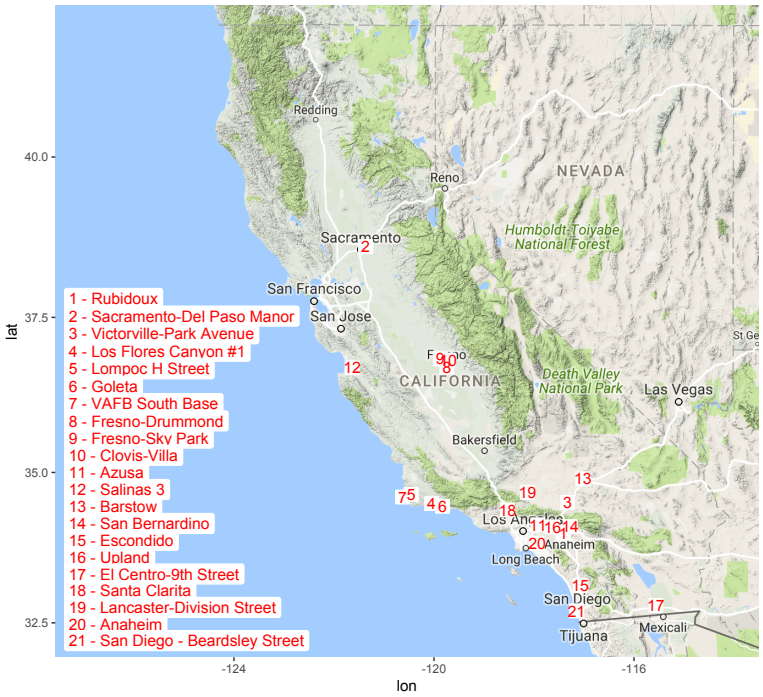



Figure 6: Map of California with the 21 sites under study indicated by numbers.

Several international institutions and agencies, including the United States Environmental Protection Agency (U.S. EPA), have traditionally focused on controlling the emissions of each of the most dangerous air pollutants, also called criteria pollutants, e.g., $O_3$, $NO_2$ and CO, separately. However, long-term and peak exposures to multiple air pollutants simultaneously have been demonstrated to cause serious health problems (Dominici *et al.*, 2010; Johns *et al.*, 2012) and so policy-makers worldwide, including the U.S. EPA, are now keen on moving towards a multi-pollutant approach to quantify air pollution risks (Johns *et al.*, 2012). There is also growing interest in studying the health effects of the interaction between ozone and temperature, see, e.g., Kahle *et al.* (2015), particularly given that temperatures are expected to rise in the coming decades. In this work, we investigate the dependence relationships between extreme concentrations of air pollutants and extreme weather conditions through max-stable distributions using the TM-MCMC algorithm described in the previous sections. In particular, we focus on daily observations available from the Air Quality System (AQS) database on the EPA website `http://aqsdr1.epa.gov/aqsweb/aqstmp/airdata/download_files.html` collected at 21 sites

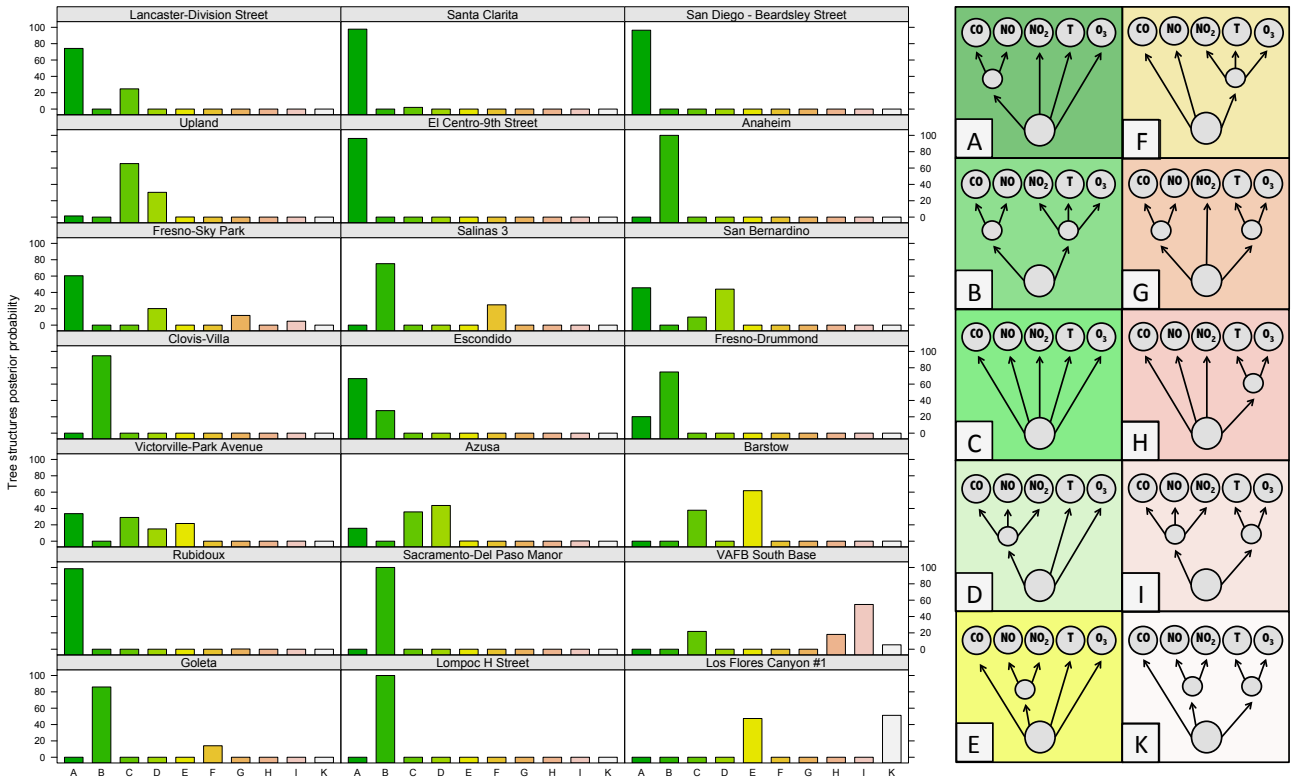



Figure 7: A different letter is associated to each of the most frequent tree structures (right) identified by the TM-MCMC algorithm after $R = 15000$ iterations and burn-in $R/5 = 3000$. The histograms (left) report the posterior probability associated to each tree for each site under study calculated according to the number of times each tree appears in the algorithm chain.

across the state of California, which is one of the most populated and polluted areas of the US (American Lung Association, 2017). The site locations can be visualised in Figure 6. Details on the data preprocessing and marginal estimations are available in the Supplementary Material.

## 6.2 Multivariate analysis of California air pollution extremes

Considering daily observations of the variables CO, NO, $NO_2$, $O_3$ and temperature (T) collected from January 2004 to December 2015, we derive at least 100 multivariate monthly maxima for each site in Figure 6 to which we apply the TM-MCMC algorithm with $R = 15000$ iterations, and burn-in $R/5 = 3000$. The most frequent dependence structures identified by our method and the corresponding posterior probabilities obtained for each of the sites under study are represented in Figure 7. In particular, the tree structures that appear most often across the chains are trees A and B, considered at 16 sites. Tree A includes only the CO-NO cluster, while tree B has two

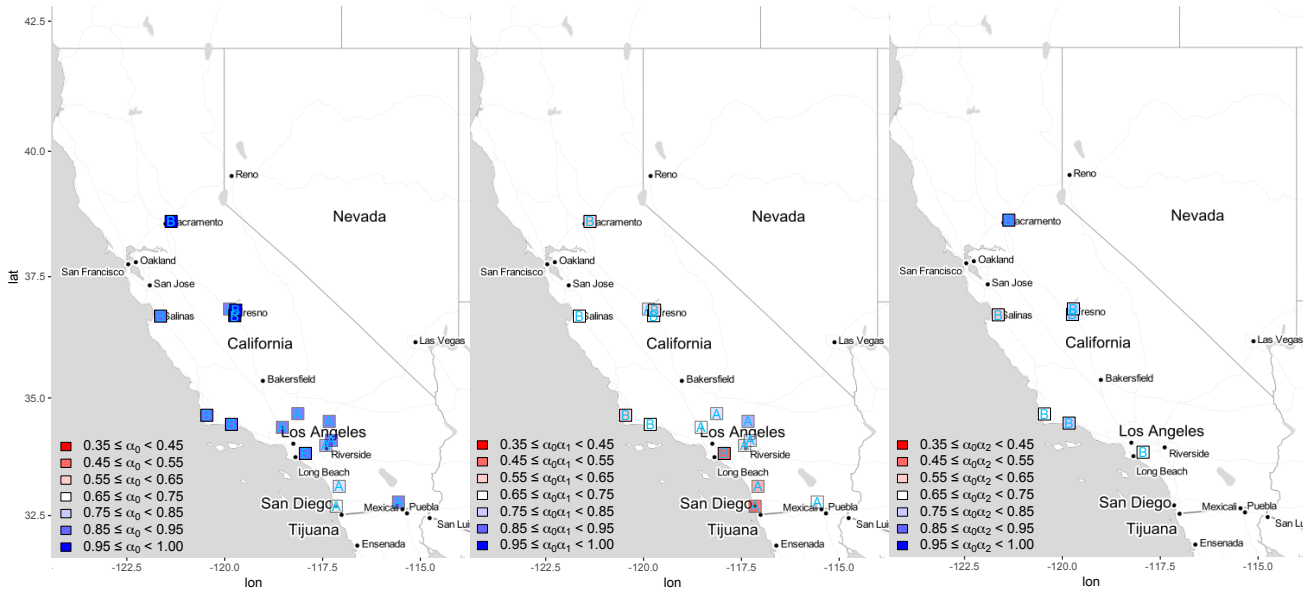



Figure 8: A colour scale is used to represent the point estimates of the nested logistic model (3) parameters computed as the median of the sub-chain corresponding to the most likely tree after burn-in. The parameter $\alpha_0$ represents the dependence between the clusters (left map) and the parameters $\alpha_0\alpha_1$, $\alpha_0\alpha_2$ represent the dependence within the clusters CO-NO (central map) and $O_3$-$NO_2$-T (right map), respectively.

clusters, respectively grouping the maxima of NO and CO and the maxima of $NO_2$ and $O_3$ with T. While the data dependence structure can be represented by only tree B at the Clovis-Villa site, for instance the TM-MCMC algorithm suggests to represent the extremal dependence structure at Victorville-Park Avenue using a mixture of the trees A, C, D and E.

Figure 8 represents the point estimates for the nested logistic model parameters $\boldsymbol{\alpha}$, considering the tree structures A and B in Figure 7. The values of the parameter $\alpha_0$ represented in the left map are generally close to $\alpha_0 = 0.7$ for the southern sites, whereas $\alpha_0$ tends to be close to 1 for the northern sites, indicating that most of the identified clusters can be treated independently. This suggests that the original observations from these distinct clusters are asymptotically independent, which can arise for example when they are weakly dependent in the bulk, but not in the upper tail, or when they are negatively associated. As expected, the dependence strength within the clusters is stronger. The maxima of NO and CO are found to be particularly dependent in areas characterised by heavy traffic, especially between the cities of Anaheim and Long Beach and close to Los Angeles and San Diego. The dependence strength between the maxima of $NO_2$, $O_3$ and T is generally mild or weak for the inland sites, but increases for the sites near the coast.

The Air Quality Index (AQI) is a measure typically used by the U.S. EPA to communicate the level of air pollution to the public. When multiple criteria pollutants are measured at the same location, the EPA reports the largest AQI value as a measure of air quality; for more details see, e.g., Airnow (2016). The AQI is divided into different categories that indicate different levels of health concern. Using our methodology, we are able to sample from the posterior predictive distribution of the AQI values for the maxima of the criteria pollutants CO, $O_3$ and $NO_2$, taking into account the uncertainty associated with the trees summarising the dependence relations between these pollutants, meteorological parameters and other air pollutants using Bayesian model averaging. In Figure 9 we display high $p$-quantiles with probabilities ranging from $p = 0.5$ to $p = 0.996$, considering August 2006 and August 2014 as baselines, computed for the smallest, the average and the largest AQI monthly maxima for CO, $O_3$ and $NO_2$. The projections and credible bands for the AQI of CO, $O_3$ and $NO_2$ were computed as explained in U.S. EPA (2016) on the basis of 500 new datasets simulated from the nested logistic distribution with the dependence parameters obtained by 500 independent TM-MCMC algorithm runs with $R = 50000$ and burn-in $R/5$. Notice that the values of $p = 1 - 1/12 \approx 0.917$ and $p = 1 - 1/(12 \times 20) \approx 0.996$ roughly correspond to 1 and 20 year-return levels, respectively, under stationary conditions. The AQI categories are represented by different colours. For comparison purposes, we also computed high quantiles for the site named Victorville-Park Avenue. Since we estimate the dependence structure separately from the margins, the 95% bootstrap credible intervals solely reflect the model uncertainty associated with the clustering, illustrating the ability of the TM-MCMC algorithm to realistically account for the model uncertainty in predictions. In practice, it might be better to adopt a fully Bayesian approach aggregating the margins and dependence structure uncertainties. Interestingly, the 95% bootstrap credible bands are very narrow in the case of the Clovis-Villa site. Indeed, from Figure 7, a single tree structure (tree B) is sufficient to represent the whole dependence structure. On the other hand, as shown in Figure 7, the site Victorville-Park Avenue is characterised by a higher model uncertainty, which reflects on

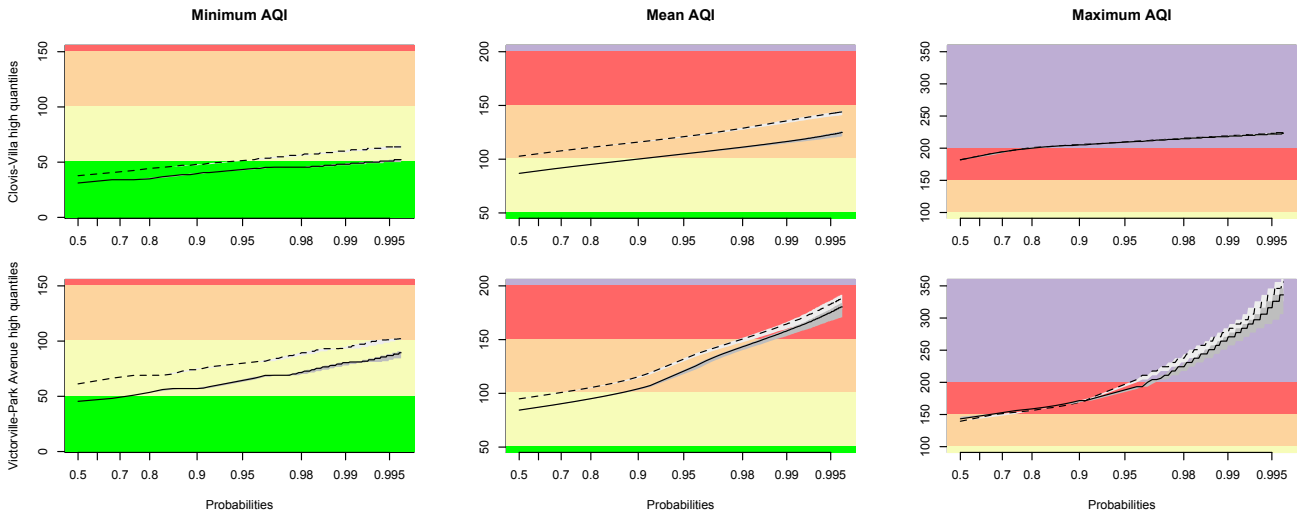



Figure 9: High $p$-quantiles $z_p$ computed for the minimum (left), the average (center) and the maximum (right) AQI setting August 2006 (dashed lines) or August 2014 (solid lines) as baselines with 95% bootstrap credible bands (grey areas) for CO, $O_3$ and $NO_2$ indexes, obtained for the Clovis-Villa (top) and Victorville-Park Avenue (bottom) site applying the TM-MCMC algorithm after $R = 50000$ iterations and burn-in $R/5$ iterations. AQI categories: 0-50 satisfactory (green); 51-100 acceptable (yellow); 101-150 unhealthy for sensitive groups (orange); 151-200 unhealthy (red); >200 very unhealthy (purple). Probabilities are displayed on a Gumbel scale, i.e., $z_p$ is plotted against $-\log\{-\log(p)\}$.

much wider credible bands for the high quantiles in Figure 9. The high $p$-quantiles projections calculated based on August 2006 are generally larger than the high quantiles based on August 2014. For 2014, the minimum AQI exceeds the satisfactory level of 50 with probability about 0.5% at Clovis-Villa and around 20% at Victorville-Park Avenue, i.e., once every 15-20 years and 3 times per year on average, respectively, under stationarity. Moreover, the average AQI generally lies within the unhealthy category only for sensitive groups of people in Clovis-Villa whereas at Victorville-Park Avenue it is expected to exceed this category with probability about 1.5%, so roughly once every 5 years on average. The maximum AQI high quantiles generally lie within the very unhealthy category, indicating that at least one of the criteria pollutants under study exceeds this most critical threshold with probability about 15% at Clovis-Villa and around 2.5% at Victorville-Park Avenue, and therefore 2 times per year and once every 3 or 4 years on average, respectively.

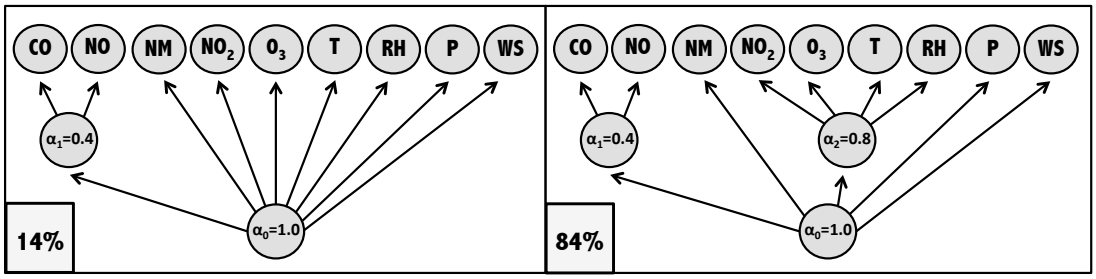



Figure 10: Dependence structures identified by the TM-MCMC algorithm after $R = 15000$ iterations and the associated posterior probability for Clovis-Villa.

## 6.3 Analysis of Clovis-Villa air pollution extremes

We now include in our analysis the monthly maxima of relative humidity (RH), barometric pressure (BP) and wind speed (WS), together with concentration maxima of non-methane organic compounds (NM), which are essentially VOCs without methane. For illustrative purposes, we focus on the site named Clovis-Villa, located in Fresno, for which we were able to derive 54 multivariate monthly maxima, collected from January 2006 to December 2014. The two tree structures identified by the TM-MCMC algorithm are illustrated in Figure 10. In accordance with previous findings, our method groups the maxima of NO and CO through 98% of the chain, and the maxima of $O_3$, $NO_2$ and T, now combined with RH, through 84% of the chain.

The algorithm output is provided in Figure 11. The sub-chains seem stationary from the trace-plots on the left panels and the autocorrelation functions on the right panels hint towards good mixing properties. In particular, the posterior median of the parameter $\alpha_0$ is exactly 1, indicating that the two clusters of variables, as well as the other single variables, can be treated independently. Therefore, our methodology allows us to significantly reduce the data complexity by describing their extremal dependence structure using only three parameters, or even two if $\alpha_0 = 1$, for the nine variables considered here. In contrast, fitting the bivariate logistic model to all possible pairs of variables yields 36 estimated dependence parameters, as shown in Figure 12. The bivariate fits suggest a relation of strong dependence between the maxima of NO and CO and of moderate dependence between the maxima of $O_3$, $NO_2$, T and RH, as indicated by the point

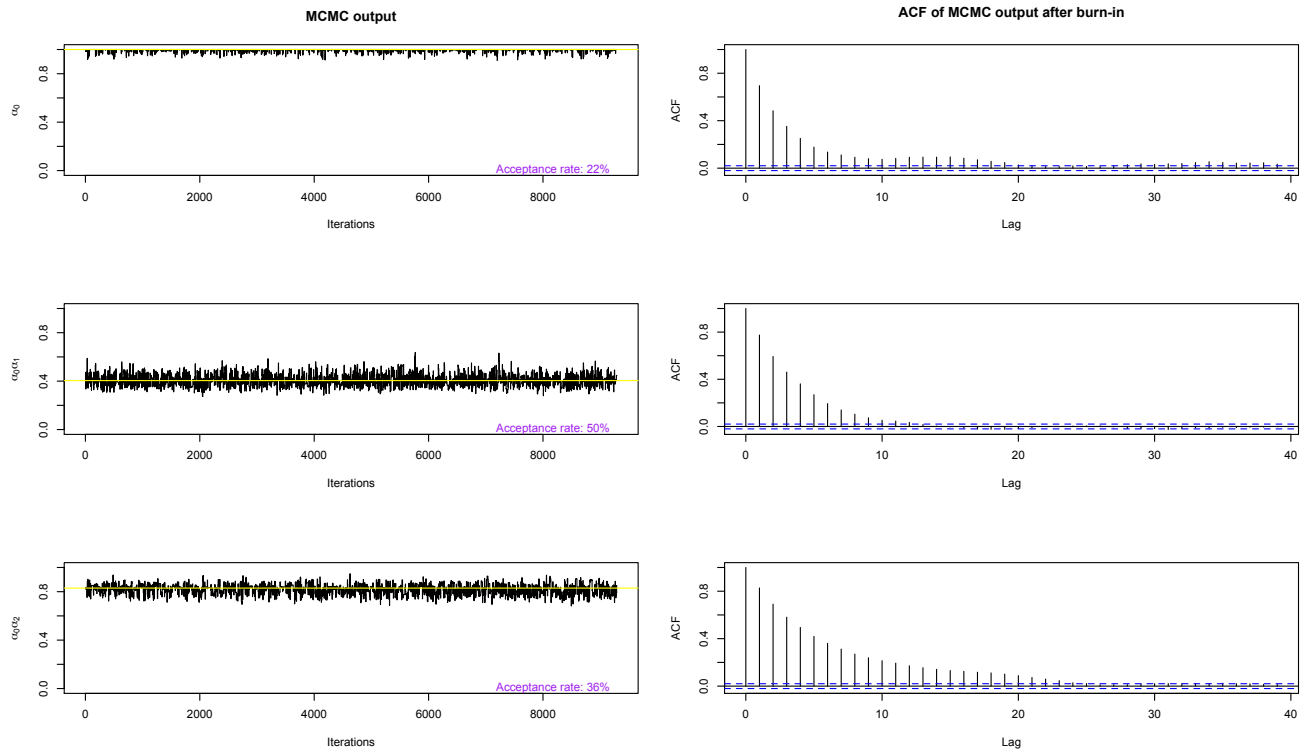



Figure 11: Trace-plots (left) and the autocorrelation functions (right) for the sub-chain corresponding to the tree structure represented in Figure 10, with posterior probability of 84% obtained by applying the TM-MCMC algorithm after $R = 15000$ iterations after a burn-in of $R/5$ iterations. The posterior medians are represented by yellow lines.

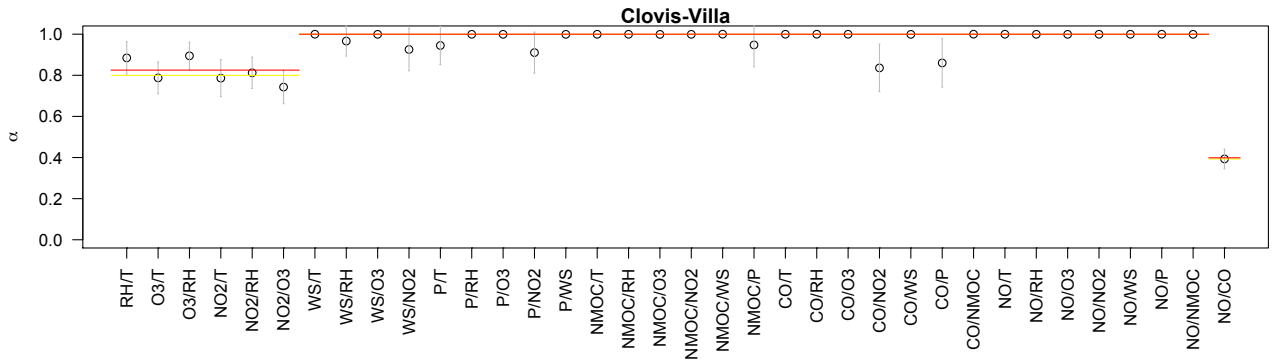



Figure 12: Parameters estimates (circles) and 95% credible intervals (vertical segments) resulting from logistic model bivariate fits for each pair of pollutants and meteorological parameters. The red lines represent the median of the parameter estimates and the yellow lines the posterior medians obtained from the multivariate fit using the TM-MCMC algorithm assuming the tree in the right panel of Figure 10, after $R = 15000$ iterations and burn-in $R/5$ iterations.

estimates obtained from the algorithm output. Interestingly, the joint fits match the bivariate fits almost perfectly, except for a few pairs of variables with high variability.

The Gelman–Rubin statistics (Gelman and Rubin, 1992), plotted in Figure 13, provide a

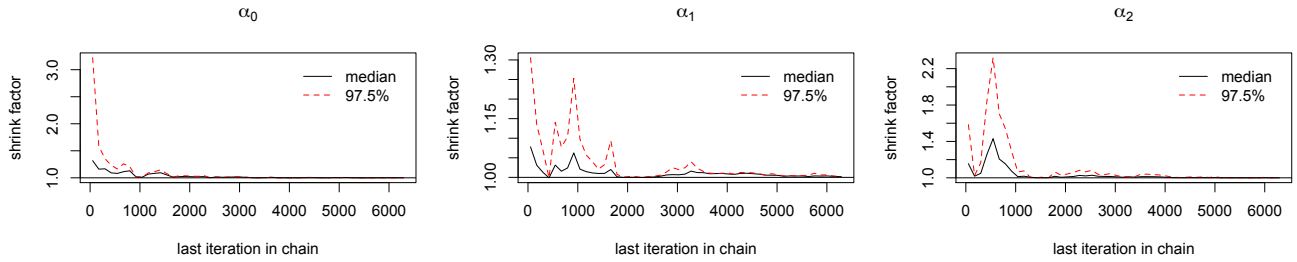



Figure 13: Gelman-Rubin statistics (Gelman and Rubin, 1992) comparing two parallel sub-chains obtained from two independent runs of the TM-MCMC algorithm for $R = 15000$ iterations and burn-in $R/5$.

convergence diagnostic measure based on the fact that if two different runs of the same model have converged, we expect the respective sub-chains to be similar to one another and the Gelman–Rubin statistics should tend to one as we increase the number of iterations. Since this is the case in Figure 13, we conclude that there is no significant difference between the variance within and between the two algorithm sub-chains for each of the dependence parameters.

# 7 Discussion

In order to describe the complex dependence relationships between extremes variables, we proposed a novel technique based on the hierarchical structure of the nested logistic model and Bayesian computational tools. Our methodology has the advantage of allowing parameter estimation and model selection to be done simultaneously, thus providing a Bayesian assessment of the model uncertainty. Numerical experiments suggest that, most of the time, the method is able to find the true dependence structure from the data, even in case of model misspecification.

The TM-MCMC algorithm was applied to air pollution concentration maxima collected at 21 sites across California. A particularly strong dependence relation between the maxima of CO and NO was found for most of the sites located between or within big cities, possibly because both pollutants are released by motor vehicles. Moreover, many of the sites close to coastal areas present a strong to mild dependence relation between the maxima of $O_3$ and $NO_2$ and high temperatures. This is expected as $O_3$ generally forms in chemical reactions that depend on the presence of $NO_2$ and heat. Fitting the bivariate logistic model to all possible pairs of variables

for the Clovis-Villa site in Fresno yields similar conclusions, verifying that, despite its simplicity, the nested logistic model was flexible enough to provide a good estimation of the overall data dependence structure. The AQI high quantiles indicate that air quality is generally expected to exceed the healthy threshold within the next two decades under stationarity. Further efforts to reduce emissions seem necessary in order to protect public health, especially given that longer-term climatic changes projections predict rising temperatures. Our method could be used for obtaining air quality measures that take into account the extremes of multiple pollutants and the public health risks associated with their exposure.

When fitting the nested logistic model, its within-cluster exchangeability may be seen as a limitation. However, embedding the nested logistic distribution within a Bayesian framework has the great advantage of describing the data dependence structure using a mixture of trees, allowing therefore to capture complex dependence relationships between variables. Indeed, our method is not only able to appropriately describe data dependence structures represented by homogeneous clusters, e.g., at the Clovis-Villa site, but also much more complicated dependencies, which often arise in applications, such as the situation at the Victorville-Park Avenue.

Our method is coded in `C++`, integrated within `R` using the package `Rcpp`. The code is available in the Supplementary Material and on the first author's website `https://extstat.kaust.edu.sa/Pages/Sabrina_Vettori.aspx`. The computational efficiency remains fairly moderate and limited to a few seconds in the settings we have considered. In order to further improve the computational efficiency in high dimensions and facilitate convergence, a guided procedure could be implemented for selecting the move type at each iteration of the algorithm and additional moves might be implemented.

A similar procedure might be applied to threshold exceedances based on the original series of observations, instead of on maxima, by implementing a censored recursive likelihood function.

Our method is based on the nested logistic model, whose hierarchical structure is less appropriate to describe spatial extremes. In Vettori *et al.* (2018), the nested logistic model is extended

to the multivariate spatial framework, allowing for much larger dimensions $D$.

Because of its max-stability property, the nested logistic model assumes asymptotic dependence between the margins, which might lead to overestimation of joint-tail probabilities under asymptotic independence (Davison *et al.*, 2013). Multivariate models for asymptotic independence were proposed among others by Ledford and Tawn (1996) and Heffernan and Tawn (2004); see de Carvalho and Ramos (2012) for a review. More recently, there has been an increasing research effort in bridging asymptotic dependence and independence in spatial (Huser *et al.*, 2017a,b; Huser and Wadsworth, 2018) and multivariate (Wadsworth *et al.*, 2017) settings, although the latter mostly focus on the bivariate case. When it is not clear whether extreme data are asymptotically dependent or independent, Davison *et al.* (2013) suggest using max-stable models since the latter provide more conservative bounds for probabilities of concurrent extremes. In this paper, we analysed multivariate, a priori unstructured, maxima data, and we therefore chose to work with max-stable models. It would be interesting to generalise our method to handle asymptotic independence scenarios.

# References


Airnow (2016) Air Quality Index (AQI) Basics. Research Triangle Park, N.C.: U.S. Environmental Protection Agency, Office of Air Quality Planning and Standards, 2000. https://airnow.gov/index.cfm?action=aqibasics.aqi.

American Lung Association (2017) *State of the air 2017*.

Bernard, E., Naveau, P., Vrac, M. and Mestre, O. (2013) Clustering of maxima: Spatial dependencies among heavy rainfall in france. *Journal of Climate* **26**(20), 7929–7937.

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

Brunekreef, B. and Holgate, S. T. (2002) Air pollution and health. *The Lancet* **360**(9341), 1233 –1242.

de Carvalho, M. and Ramos, A. (2012) Bivariate extreme statistics, II. *Revstat* **10**, 83–107.

Castruccio, S., Huser, R. and Genton, M. G. (2016) High-order composite likelihood inference for max-stable distributions and processes. *Journal of Computational and Graphical Statistics* **25**(4), 1212–1229.

Chautru, E. (2015) Dimension reduction in multivariate extreme value analysis. *Electronic Journal of Statistics* **9**, 383–418.

Cofala, J., Amann, M., Klimont, Z., Kupiainen, K. and Hoglund-Isaksson, L. (2007) Scenarios of global anthropogenic emissions of air pollutants and methane until 2030. *Atmospheric Environment* **41**(38), 8486–8499.

Coles, S. G. and Tawn, J. A. (1991) Modelling extreme multivariate events. *Journal of the Royal Statistical Society. Series B* **53**(2), 377–392.

Cooley, D. and Thibaud, E. (2018) Decompositions of dependence for high-dimensional extremes. arXiv preprint 1612.07190.

Davison, A. C. and Gholamrezaee, M. M. (2012) Geostatistics of extremes. *Proceedings of the Royal Society A: Mathematical, Physical and Engineering Sciences* **468**, 581–608.

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

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

Davison, A. C., Padoan, S. A. and Ribatet, M. (2012) Statistical modeling of spatial extremes. *Statistical Science* **27**(2), 161–186.

Dominici, F., Peng, R. D., Barr, C. D. and Bell, M. L. (2010) Protecting human health from air pollution: shifting from a single-pollutant to a multi-pollutant approach. *Epidemiology* **21**(2), 187–194.

Gelman, A. and Rubin, D. B. (1992) Inference from iterative simulation using multiple sequences. *Statistical Science* **7**, 457–511.

Genton, M. G., Ma, Y. and Sang, H. (2011) On the likelihood function of Gaussian max-stable processes. *Biometrika* **98**(2), 481–488.

Górecki, J., Hofert, M. and Holeňa, M. (2016) An approach to structure determination and estimation of hierarchical Archimedean copulas and its application to Bayesian classification. *Journal of Intelligent Information Systems* **46**(1), 21–59.

Graham, R., Knuth, D. E. and Patashnik, O. (1988) *Concrete Mathematics*. Reading MA: Addison-Wesley.

Green, P. J. (1995) Reversible jump Markov chain Monte Carlo computation and Bayesian model determination. *Biometrika* **82**, 711–732.

Guillotte, S., Perron, F. and Segers, J. (2011) Non-parametric Bayesian inference on bivariate extremes. *Journal of the Royal Statistical Society. Series B (Statistical Methodology)* **73**(3), 377–406.

Gumbel, E. J. (1960a) Distributions des valeurs extrêmes en plusieurs dimensions. *Publications de l'Institute de Statistique de l'Université de Paris* **9**(294), 171–173.

Gumbel, E. J. (1960b) Bivariate exponential distributions. *Journal of the American Statistical Association* **55**, 698–707.

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

de Haan, L. and Resnick, S. I. (1977) Limit theory for multivariate sample extremes. *Zeitschrift für Wahrscheinlichkeitstheorie und Verwandte Gebiete* **40**, 317–337.

Hastings, W. K. (1970) Monte Carlo sampling methods using Markov chains and their applications. *Biometrika* **57**(1), 97–109.

Heffernan, J. E. and Tawn, J. A. (2004) A conditional approach for multivariate extreme values (with discussion). *Journal of the Royal Statistical Society: Series B (Statistical Methodology)* **66**, 497–546.

Hoeting, J. A., Madigan, D., E., R. A. and Volinsky, C. T. (1999) Bayesian model averaging: A tutorial. *Statistical Science* **14**(4), 382–401.

Hofert, M. (2011) Efficiently sampling nested Archimedean copulas. *Computational Statistics and Data Analysis* **55**(1), 57–70.

Hofert, M. and Martin, M. (2011) Nested Archimedean copulas meet R : The nacopula package. *Journal of Statistical Software* **39**(9), 1–20.

Hofert, M. and Pham, D. (2013) Densities of nested Archimedean copulas. *Journal of Multivariate Analysis* **118**, 37–52.

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

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

Huser, R., Davison, A. C. and Genton, M. G. (2016) Likelihood estimators for multivariate extremes. *Extremes* **19**(1), 79–103.

Huser, R., Dombry, C., Ribatet, M. and Genton, M. G. (2019) Full likelihood inference for max-stable data. *Stat* **8**, e218.

Huser, R., Opitz, T. and Thibaud, E. (2017a) Bridging asymptotic independence and dependence in spatial extremes using Gaussian scale mixtures. *Spatial Statistics.* **21**, 166–186.

Huser, R., Opitz, T. and Thibaud, E. (2017b) Penultimate modeling of spatial extremes: statistical inference for max-infinitely divisible processes. arXiv preprint arXiv:1801.02946.

Huser, R. and Wadsworth, J. L. (2018) Modeling spatial processes with unknown extremal dependence class. *Journal of the American Statistical Association* To appear.

Jacob, D. and Winner, D. (2009) Effect of climate change on air quality. *Atmospheric Environment* **43**(1), 51–63.

Johns, D. O., Stanek, L. W., Walker, K., Benromdhane, S., Hubbell, B., Ross, M., Devlin, R. B., Costa, D. L. and Greenbaum, D. S. (2012) Practical advancement of multipollutant scientific and risk assessment approaches for ambient air pollution. *Environmental Health Perspectives* **120**(9), 1238–1242.

Kahle, J. J., Neas, L. M., Devlin, R. B., Case, M. W., Schmitt, M. T., Madden, M. C. and Diaz-Sanchez, D. (2015) Interaction effects of temperature and ozone on lung function and markers of systemic inflammation, coagulation, and fibrinolysis: A crossover study of healthy young volunteers. *Environ Health Perspect* **123**(4), 310–316.

Kampa, M. and Castanas, E. (2008) Human health effects of air pollution. *Environmental Pollution* **151**(2), 362–367.

Ledford, A. and Tawn, J. A. (1996) Statistics for near independence in multivariate extreme values. *Biometrika* **83**, 169–187.

McFadden, D. L. (1978) Modeling the choice of residental location. In *Spatial Interaction Theory and Planning Models (eds A. Karlquist et al.)*, pp. 75–96. North-Holland, Amsterdam.

Murphy, J. J., Delucchi, M. A., McCubbin, D. R. and Kim, H. J. (1999) The cost of crop damage caused by ozone air pollution from motor vehicles. *Journal of Environmental Management* **55**(4), 273–289.

Okhrin, O., Okhrin, Y. and Schmid, W. (2009) *Properties of Hierarchical Archimedean Copulas*. Humboldt University, Berlin, Germany.

Okhrin, O., Okhrin, Y. and Schmid, W. (2013) On the structure and estimation of hierarchical Archimedean copulas. *Journal of Econometrics* **173**(2), 189 – 204.

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. and Sisson, S. A. (2010) Likelihood-based inference for max-stable processes. *Journal of the American Statistical Association* **105**(489), 263–277.

Peden, D. B. (2001) Air pollution in asthma: effect of pollutants on airway inflammation. *Annals of Allergy, Asthma and Immunology* **87**(6), 12–17.

Reich, B. J. and Shaby, B. A. (2012) A hierarchical max-stable spatial model for extreme precipitation. *Annals of Applied Statistics* **6**(4), 1430–1451.

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

Sabourin, A. and Naveau, P. (2014) Bayesian Dirichlet mixture model for multivariate extremes: A re-parametrization. *Computational Statistics and Data Analysis* **71**, 542–567.

Sabourin, A., Naveau, P. and Fougères, A.-L. (2013) Bayesian model averaging for multivariate extremes. *Extremes* **16**(3), 325–350.

Segers, J. and Uyttendaele, N. (2014) Nonparametric estimation of the tree structure of a nested Archimedean copula. *Computational Statistics and Data Analysis* **72**, 190 – 204.

Seinfeld, J. H. and Pandis, S. N. (2016) *Atmospheric Chemistry and Physics: From Air Pollution to Climate Change*. John Wiley and Sons, Ltd. Hoboken, New Jersey.

Shaby, B. A. (2014) The open-faced sandwich adjustment for MCMC using estimating functions. *Journal of Computational and Graphical Statistics* **23**, 853–876.

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. University of North Carolina, Chapel Hill, U.S., Unpublished manuscript.

Stephenson, A. (2003) Simulating multivariate extreme value distributions of logistic type. *Extremes* **6**(1), 49–59.

Stephenson, A. and Tawn, J. (2005) Exploiting occurence times in likelihood inference for componetwise maxima. *Biometrika* **92**(1), 213–227.

Tawn, J. A. (1990) Modelling multivariate extreme value distributions. *Biometrika* **77**(2), 245–253.

Thibaud, E., J. Aalto, D. S. C., Davison, A. C. and Heikkinen, J. (2016) Bayesian inference for the Brown-Resnick process, with an application to extreme low temperatures. *Annals of Applied Statistics.* **10**(4), 2303–2324.

U.S. EPA (2016) Technical Assistance Document for the Reporting of Daily Air Quality—the Air Quality Index (AQI). https://www3.epa.gov/airnow/aqi-technical-assistance-document-may2016.pdf.

Varin, B. C. and Vidoni, P. (2005) A note on composite likelihood inference and model selection. *Biometrika* **92**(3), 519–528.

Vettori, S., Huser, R. and Genton, M. G. (2018) Bayesian modeling of air pollution extremes using nested multivariate max-stable processes. arXiv preprint 1804.04588.

Wadsworth, J. L. (2015) On the occurrence times of componentwise maxima and bias in likelihood inference for multivariate max-stable distributions. *Biometrika* **102**(3), 705–711.

Wadsworth, J. L. and Tawn, J. A. (2012) Dependence modelling for spatial extremes. *Biometrika* **99**(2), 253–272.

Wadsworth, J. L., Tawn, J. A. and Davison, A. C. (2017) Modelling across extremal dependence classes. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)* **79**(1), 149–175.

# A Recursive formulas for the nested logistic distribution

## A.1 Notation

For simplicity, we denote the distribution of the nested logistic model as $G = \exp(-V)$, where

$$V = \left( \sum_{k=1}^{K} V_k \right)^{\alpha_0}, \qquad V_k = \left\{ \sum_{i_k=1}^{D_k} z_{k;i_k}^{-1/(\alpha_0 \alpha_k)} \right\}^{\alpha_k}, \qquad k = 1, 2, \ldots, K,$$

with dependence parameters $0 < \alpha_0, \alpha_1, \ldots, \alpha_K \leq 1$, and where, for clarity, we have omitted the function arguments. Moreover, we use the following vector notation:

$$\begin{aligned}
\boldsymbol{z}_k &= (z_{k;1}, \ldots, z_{k;D_k})^\top && \textit{All variables in cluster } k = 1, \ldots, K; \\
\boldsymbol{z}_{k,1:d_k} &= (z_{k;1}, \ldots, z_{k;d_k})^\top && \textit{Sub-vector of the first } d_k \textit{ variables in cluster } k = 1, \ldots, K; \\
\boldsymbol{i}_{1:\kappa} &= (i_1, \ldots, i_\kappa)^\top && \textit{Vector of } \kappa \textit{ indices.}
\end{aligned}$$

## A.2 Preliminary results

From the definition in Section A.1, we deduce the following derivatives:

$$\frac{\partial V_k}{\partial z_{k;i_k}} = \alpha_k \left\{ \sum_{i_k=1}^{D_k} z_{k;i_k}^{-1/(\alpha_0 \alpha_k)} \right\}^{\alpha_k - 1} \times \frac{-1}{\alpha_0 \alpha_k} z_{k;i_k}^{-1/(\alpha_0 \alpha_k) - 1} = -\frac{1}{\alpha_0} z_{k;i_k}^{-1/(\alpha_0 \alpha_k) - 1} V_k^{1 - 1/\alpha_k}; \quad (10)$$

$$\frac{\partial V}{\partial z_{k;i_k}} = \alpha_0 \left( \sum_{k=1}^{K} V_k \right)^{\alpha_0 - 1} \frac{\partial V_k}{\partial z_{k;i_k}} = -z_{k;i_k}^{-1/(\alpha_0 \alpha_k)-1} V_k^{1-1/\alpha_k} V^{1-1/\alpha_0}; \tag{11}$$

$$\frac{\partial G}{\partial z_{k;i_k}} = -\frac{\partial V}{\partial z_{k;i_k}} \exp(-V) = z_{k;i_k}^{-1/(\alpha_0 \alpha_k)-1} G V_k^{1-1/\alpha_k} V^{1-1/\alpha_0}. \tag{12}$$

## A.3 Partial derivatives of $G$

The distribution $G$ is a function of $D = \sum_{k=1}^{K} D_k$ variables, namely $\boldsymbol{z}_1 = (z_{1;1}, \ldots, z_{1;D_1})^\top, \ldots, \boldsymbol{z}_K = (z_{K;1}, \ldots, z_{K;D_K})^\top$. The partial derivative of $G$ with respect to any subset of variables $\boldsymbol{z}_{1,1:d_1}, \ldots, \boldsymbol{z}_{\kappa,1:d_\kappa}$ in $1 \leq \kappa \leq K$ clusters of dimensions $1 \leq d_k \leq D_k$, $k = 1, \ldots, \kappa$, may be expressed as

$$\frac{\partial^{\sum_{k=1}^{\kappa} d_k} G}{\partial \prod_{k=1}^{\kappa} \mathbf{z}_{k;1:d_k}} = G \prod_{i_1=1}^{d_1} z_{1;i_1}^{-\frac{1}{\alpha_0 \alpha_1}-1} \cdots \prod_{i_\kappa=1}^{d_\kappa} z_{\kappa;i_\kappa}^{-\frac{1}{\alpha_0 \alpha_\kappa}-1} \sum_{i_1=1}^{d_1} \cdots \sum_{i_\kappa=1}^{d_\kappa} \sum_{j=1}^{\sum_{k=1}^{\kappa} i_k} \beta_{i_1;\ldots;i_\kappa;j}^{(d_1;\ldots;d_\kappa)} V_1^{i_1 - \frac{d_1}{\alpha_1}} \cdots V_\kappa^{i_\kappa - \frac{d_\kappa}{\alpha_k}} V^{j - \frac{\sum_{k=1}^{\kappa} i_k}{\alpha_0}} \tag{13}$$

where the coefficients $\beta_{i_1;\ldots;i_\kappa;j}^{(d_1;\ldots;d_\kappa)}$ can be computed recursively as demonstrated below. By specialising the expression in (13) to $\kappa = K$ clusters and $d_1 = D_1, \ldots, d_K = D_K$ variables, we obtain the density, or full likelihood, for one replicate.

**Proof** Equation (13) may be proven by double induction over $\kappa \in \{1, \ldots, K\}$ and $d_k \in \{1, \ldots, D_k\}$, $k = 1, \ldots, \kappa$. The proof also naturally provides a constructive approach to the recursive computation of the coefficients $\beta_{i_1;\ldots;i_\kappa;j}^{(d_1;\ldots;d_\kappa)}$. More precisely, we demonstrate the following four steps:

1. (13) holds for $\kappa = 1$ and $d_1 = 1$;
2. If (13) holds for $\kappa = 1$ and $d_1 \in \{1, \ldots, D_1 - 1\}$, then it also holds for $d_1 \mapsto d_1 + 1$;
3. If (13) holds for $\kappa \in \{1, \ldots, K - 1\}$, then it also holds for $\kappa \mapsto \kappa + 1$ with $d_{\kappa+1} = 1$;
4. If (13) holds for $\kappa \in \{1, \ldots, K\}$ and $d_\kappa \in \{1, \ldots, D_{\kappa-1}\}$, then it also holds for $d_\kappa \mapsto d_\kappa + 1$.

**Step 1.** From (12) we have

$$\frac{\partial G}{\partial z_{1;1}} = z_{1;1}^{-1/(\alpha_0 \alpha_1)-1} G V_1^{1-1/\alpha_1} V^{1-1/\alpha_0},$$

which proves the first step by setting $\beta_{1;1}^{(1)} = 1$.

**Step 2.** Assuming that (13) holds for $\kappa = 1$ and $d_1 \in \{1, \dots, D_1 - 1\}$, and using (10), (11) and (12), we obtain

$$\frac{\partial^{d_1+1} G}{\partial \mathbf{z}_{1;1:d_1+1}} = \prod_{i_1=1}^{d_1} z_{1;i_1}^{-\frac{1}{\alpha_0\alpha_1}-1} \left\{ \frac{\partial G}{z_{1;d_1+1}} \sum_{i_1=1}^{d_1}\sum_{j=1}^{i_1} \beta_{i_1;j}^{(d_1)} V_1^{i_1-\frac{d_1}{\alpha_1}} V^{j-\frac{i_1}{\alpha_0}} + G \sum_{i_1=1}^{d_1}\sum_{j=1}^{i_1} \beta_{i_1;j}^{(d_1)} \frac{\partial V_1^{i_1-\frac{d_1}{\alpha_1}}}{z_{1;d_1+1}} V^{j-\frac{i_1}{\alpha_0}} \right.$$
$$\left. + G \sum_{i_1=1}^{d_1}\sum_{j=1}^{i_1} \beta_{i_1;j}^{(d_1)} V_1^{i_1-\frac{d_1}{\alpha_1}} \frac{\partial V^{j-\frac{i_1}{\alpha_0}}}{z_{1;d_1+1}} \right\}$$

$$= G \prod_{i_1=1}^{d_1+1} z_{1;i_1}^{-\frac{1}{\alpha_0\alpha_1}-1} \left\{ \sum_{i_1=1}^{d_1}\sum_{j=1}^{i_1} \beta_{i_1;j}^{(d_1)} V_1^{(i_1+1)-\frac{(d_1+1)}{\alpha_1}} V^{(l+1)-\frac{(i_1+1)}{\alpha_0}} \right.$$
$$- \sum_{i_1=1}^{d_1}\sum_{j=1}^{i_1} \beta_{i_1;j}^{(d_1)} \left(i_1 - \frac{d_1}{\alpha_1}\right) \left(\frac{1}{\alpha_0}\right) V_1^{i_1-\frac{d_1+1}{\alpha_1}} V^{j-\frac{i_1}{\alpha_0}} - \left. \sum_{i_1=1}^{d_1}\sum_{j=1}^{i_1} \beta_{i_1;j}^{(d_1)} \left(l - \frac{i_1}{\alpha_0}\right) V_1^{(i_1+1)-\frac{d_1+1}{\alpha_1}} V^{j-\frac{(i_1+1)}{\alpha_0}} \right\}$$
$$= G \prod_{i_1=1}^{d_1+1} z_{1;i_1}^{-\frac{1}{\alpha_0\alpha_1}-1} \left\{ \sum_{i_1=2}^{d_1+1}\sum_{l=2}^{i_1} \beta_{i_1-1;j-1}^{(d_1)} V_1^{i_1-\frac{(d_1+1)}{\alpha_1}} V^{j-\frac{i_1}{\alpha_0}} - \sum_{i_1=1}^{d_1}\sum_{j=1}^{i_1} \beta_{i_1;j}^{(d_1)} \left(i_1 - \frac{d_1}{\alpha_1}\right) \left(\frac{1}{\alpha_0}\right) V_1^{i_1-\frac{d_1+1}{\alpha_1}} V^{j-\frac{i_1}{\alpha_0}} \right.$$
$$\left. - \sum_{i_1=2}^{d_1+1}\sum_{j=1}^{i_1} \beta_{i_1-1;j}^{(d_1)} \left(l - \frac{i_1-1}{\alpha_0}\right) V_1^{i_1-\frac{d_1+1}{\alpha_1}} V^{j-\frac{i_1}{\alpha_0}} \right\} = G \prod_{i_1=1}^{d_1+1} z_{1;i_1}^{-\frac{1}{\alpha_0\alpha_1}-1} \sum_{i_1=1}^{d_1+1}\sum_{j=1}^{i_1} \beta_{i_1;j}^{(d_1+1)} V_1^{i_1-\frac{d_1+1}{\alpha_1}} V^{j-\frac{i_1}{\alpha_0}},$$

where

$$\beta_{i_1;j}^{(d_1+1)} = \beta_{i_1-1;j-1}^{(d_1)} - \frac{1}{\alpha_0}\left(i_1 - \frac{d_1}{\alpha_1}\right)\beta_{i_1;j}^{(d_1)} - \left(l - \frac{i_1-1}{\alpha_0}\right)\beta_{i_1-1;j}^{(d_1)}, \quad 1 \le j \le i_1 \le d_1 + 1, \tag{14}$$

with

$$\beta_{i_1;j}^{(d_1)} = 0, \qquad \text{for all} \quad i_1 \notin \{1, \dots, d_1\}, \quad \text{or} \quad j \notin \{1, \dots, i_1\}. \tag{15}$$

Hence, (13) holds by induction for $\kappa = 1$ and any $1 \le d_1 \le D_1$, and the recursive formula to compute coefficients $\beta_{i_1;j}^{(d_1)}$ is given by (14) and (15) with the initial condition $\beta_{1;1}^{(1)} = 1$.

**Step 3.** Assuming that (13) holds for $\kappa \in \{1,\dots,K-1\}$, we obtain

$$\begin{aligned}
&\frac{\partial^{1+\sum_{k=1}^{\kappa} d_k} G}{\partial \prod_{k=1}^{\kappa} \mathbf{z}_{k;1:d_k} \partial z_{\kappa+1;1}} = \prod_{i_1=1}^{d_1} z_{1;i_1}^{-\frac{1}{\alpha_0\alpha_1}-1} \cdots \prod_{i_\kappa=1}^{d_\kappa} z_{\kappa;i_\kappa}^{-\frac{1}{\alpha_0\alpha_\kappa}-1} \left\{ \frac{\partial G}{\partial z_{\kappa+1;1}} \sum_{i_1=1}^{d_1} \cdots \sum_{i_\kappa=1}^{d_\kappa} V_1^{i_1-\frac{d_1}{\alpha_1}} \cdots V_k^{i_\kappa-\frac{d_\kappa}{\alpha_k}} \right. \\
&\left. \times \sum_{j=1}^{i_1+\cdots+i_\kappa} \beta_{i_1;\dots;i_\kappa;j}^{(d_1;\dots;d_\kappa)} V^{j-\frac{i_1+\cdots+i_\kappa}{\alpha_0}} + G \sum_{i_1=1}^{d_1} \cdots \sum_{i_\kappa=1}^{d_\kappa} \sum_{j=1}^{i_1+\cdots+i_\kappa} \beta_{i_1;\dots;i_\kappa;j}^{(d_1;\dots;d_\kappa)} V_1^{i_1-\frac{d_1}{\alpha_1}} \cdots V_k^{i_\kappa-\frac{d_\kappa}{\alpha_k}} \frac{\partial V^{j-\frac{i_1+\cdots+i_\kappa}{\alpha_0}}}{\partial z_{\kappa+1;1}} \right\} \\
&= G \prod_{i_1=1}^{d_1} z_{1;i_1}^{-\frac{1}{\alpha_0\alpha_1}-1} \cdots \prod_{i_\kappa=1}^{d_\kappa} z_{\kappa;i_\kappa}^{-\frac{1}{\alpha_0\alpha_\kappa}-1} (z_{\kappa+1;1})^{-\frac{1}{\alpha_0\alpha_{\kappa+1}}-1} \sum_{i_1=1}^{d_1} \cdots \sum_{i_\kappa=1}^{d_\kappa} V_1^{i_1-\frac{d_1}{\alpha_1}} \cdots V_k^{i_\kappa-\frac{d_\kappa}{\alpha_k}} \left\{ \sum_{l=2}^{i_1+\cdots+i_\kappa+1} \beta_{i_1;\dots;i_\kappa;j-1}^{(d_1;\dots;d_\kappa)} \right. \\
&\left. \times V_{\kappa+1}^{1-\frac{1}{\alpha_{\kappa+1}}} V^{j-\frac{i_1+\cdots+i_\kappa+1}{\alpha_0}} - \sum_{j=1}^{i_1+\cdots+i_\kappa} \beta_{i_1;\dots;i_\kappa;j}^{(d_1;\dots;d_\kappa)} \left( l - \frac{i_1+\cdots+i_\kappa}{\alpha_0} \right) V_{\kappa+1}^{1-\frac{1}{\alpha_{\kappa+1}}} V^{j-\frac{i_1+\cdots+i_\kappa+1}{\alpha_0}} \right\} \\
&= G \prod_{i_1=1}^{d_1} z_{1;i_1}^{-\frac{1}{\alpha_0\alpha_1}-1} \cdots \prod_{i_\kappa=1}^{d_\kappa} z_{\kappa;i_\kappa}^{-\frac{1}{\alpha_0\alpha_\kappa}-1} (z_{\kappa+1;1})^{-\frac{1}{\alpha_0\alpha_{\kappa+1}}-1} \\
&\times \sum_{i_1=1}^{d_1} \cdots \sum_{i_\kappa=1}^{d_\kappa} \sum_{j=1}^{i_1+\cdots+i_\kappa+1} \beta_{i_1;\dots;i_\kappa;1;j}^{(d_1;\dots;d_\kappa;1)} V_1^{i_1-\frac{d_1}{\alpha_1}} \cdots V_k^{i_\kappa-\frac{d_\kappa}{\alpha_k}} V_{\kappa+1}^{1-\frac{1}{\alpha_{\kappa+1}}} V^{j-\frac{i_1+\cdots+i_\kappa+1}{\alpha_0}},
\end{aligned}$$

where

$$\beta_{i_1;\dots;i_\kappa;1;j}^{(d_1;\dots;d_\kappa;1)} = \beta_{i_1;\dots;i_\kappa;j-1}^{(d_1;\dots;d_\kappa)} - \left( l - \frac{i_1+\cdots+i_\kappa}{\alpha_0} \right) \beta_{i_1;\dots;i_\kappa;j}^{(d_1;\dots;d_\kappa)}, 1 \le j \le \sum_{k=1}^{\kappa} i_k + 1, 1 \le i_k \le d_k,\ k = 1,\dots,\kappa, \tag{16}$$

with

$$\beta_{i_1;\dots;i_\kappa;1;j}^{(d_1;\dots;d_\kappa;1)} = 0, \quad \text{for all} \quad i_k \notin \{1,\dots,d_k\},\ k = 1,\dots,\kappa, \quad \text{or} \quad l \notin \{1,\dots,i_1+\cdots+i_\kappa\}. \tag{17}$$

Hence, (13) holds by induction for $\kappa \mapsto \kappa+1$ with $d_{\kappa+1} = 1$ and the recursive formula to compute coefficients $\beta_{i_1;\dots;i_\kappa,1,l}^{(d_1;\dots;d_\kappa,1)}$ is given by (16) and (17).

**Step 4.** Assuming that (13) holds for $\kappa \in \{1,\dots,K\}$ and $d_\kappa \in \{1,\dots,D_\kappa - 1\}$, we obtain

$$\frac{\partial^{1+\sum_{k=1}^{\kappa} d_k} G}{\partial \prod_{k=1}^{\kappa} \mathbf{z}_{k;1:d_k} \partial z_{\kappa;d_\kappa+1}} = \prod_{i_1=1}^{d_1} z_{1;i_1}^{-\frac{1}{\alpha_0\alpha_1}-1} \cdots \prod_{i_\kappa=1}^{d_\kappa} z_{\kappa;i_\kappa}^{-\frac{1}{\alpha_0\alpha_\kappa}-1} \left\{ \frac{\partial G}{\partial z_{\kappa;d_\kappa+1}} \sum_{i_1=1}^{d_1} \cdots \sum_{i_\kappa=1}^{d_\kappa} \sum_{j=1}^{i_1+\cdots+i_\kappa} \beta_{i_1;\dots;i_\kappa;j}^{(d_1;\dots;d_\kappa)} \right.$$

$$
\begin{aligned}
&\times V_1^{i_1-\frac{d_1}{\alpha_1}}\cdots V_\kappa^{i_\kappa-\frac{d_\kappa}{\alpha_k}}V^{j-\frac{i_1+\cdots+i_\kappa}{\alpha_0}}+G\sum_{i_1=1}^{d_1}\cdots\sum_{i_\kappa=1}^{d_\kappa}\sum_{j=1}^{i_1+\cdots+i_\kappa}\beta_{i_1;\ldots;i_\kappa;j}^{(d_1;\ldots;d_\kappa)}V_1^{i_1-\frac{d_1}{\alpha_1}}\cdots\frac{\partial V_\kappa^{i_\kappa-\frac{d_\kappa}{\alpha_\kappa}}}{\partial z_{\kappa;d_\kappa+1}}V^{j-\frac{i_1+\cdots+i_\kappa}{\alpha_0}}\\
&+G\sum_{i_1=1}^{d_1}\cdots\sum_{i_\kappa=1}^{d_\kappa}\sum_{j=1}^{i_1+\cdots+i_\kappa}\beta_{i_1;\ldots;i_\kappa;j}^{(d_1;\ldots;d_\kappa)}V_1^{i_1-\frac{d_1}{\alpha_1}}\cdots V_\kappa^{i_\kappa-\frac{d_\kappa}{\alpha_\kappa}}\frac{\partial V^{j-\frac{i_1+\cdots+i_\kappa}{\alpha_0}}}{\partial z_{\kappa;d_\kappa+1}}\Bigg\}\\
&=G\prod_{i_1=1}^{d_1}z_{1;i_1}^{-\frac{1}{\alpha_0\alpha_1}-1}\cdots\prod_{i_\kappa=1}^{d_\kappa+1}z_{\kappa;i_\kappa}^{-\frac{1}{\alpha_0\alpha_\kappa}-1}\Bigg\{\sum_{i_1=1}^{d_1}\cdots\sum_{i_\kappa=2}^{d_\kappa+1}\sum_{l=2}^{i_1+\cdots+i_\kappa}\beta_{i_1;\ldots;i_\kappa-1;j-1}^{(d_1;\ldots;d_\kappa)}V_1^{i_1-\frac{d_1}{\alpha_1}}\cdots V_\kappa^{i_\kappa-\frac{d_\kappa+1}{\alpha_\kappa}}V^{j-\frac{i_1+\cdots+i_\kappa}{\alpha_0}}\\
&-\sum_{i_1=1}^{d_1}\cdots\sum_{i_\kappa=1}^{d_\kappa}\sum_{j=1}^{i_1+\cdots+i_\kappa}\beta_{i_1;\ldots;i_\kappa;j}^{(d_1;\ldots;d_\kappa)}V_1^{i_1-\frac{d_1}{\alpha_1}}\cdots\left(i_\kappa-\frac{d_\kappa}{\alpha_\kappa}\right)\left(\frac{1}{\alpha_0}\right)V_\kappa^{i_\kappa-\frac{d_\kappa+1}{\alpha_\kappa}}V^{j-\frac{i_1+\cdots+i_\kappa}{\alpha_0}}\\
&-\sum_{i_1=1}^{d_1}\cdots\sum_{i_\kappa=2}^{d_\kappa+1}\sum_{j=1}^{i_1+\cdots+i_\kappa-1}\beta_{i_1;\ldots;i_\kappa-1;j}^{(d_1;\ldots;d_\kappa)}V_1^{i_1-\frac{d_1}{\alpha_1}}\cdots V_\kappa^{i_\kappa-\frac{d_\kappa+1}{\alpha_\kappa}}\left(l-\frac{i_1+\cdots+i_\kappa-1}{\alpha_0}\right)V^{j-\frac{i_1+\cdots+i_\kappa}{\alpha_0}}\Bigg\}\\
&=G\prod_{i_1=1}^{d_1}z_{1;i_1}^{-\frac{1}{\alpha_0\alpha_1}-1}\cdots\prod_{i_\kappa=1}^{d_\kappa+1}z_{\kappa;i_\kappa}^{-\frac{1}{\alpha_0\alpha_\kappa}-1}\sum_{i_1=1}^{d_1}\cdots\sum_{i_\kappa=1}^{d_\kappa+1}\sum_{j=1}^{i_1+\cdots+i_\kappa}\beta_{i_1;\ldots;i_\kappa;j}^{(d_1;\ldots;d_\kappa+1)}V_1^{i_1-\frac{d_1}{\alpha_1}}\cdots V_k^{i_\kappa-\frac{d_\kappa+1}{\alpha_k}}V^{j-\frac{i_1+\cdots+i_\kappa}{\alpha_0}}
\end{aligned}
$$

where

$$
\beta_{i_1;\ldots;i_\kappa;j}^{(d_1;\ldots;d_\kappa+1)}=\beta_{i_1;\ldots;i_\kappa-1;j-1}^{(d_1;\ldots;d_\kappa)}-\frac{1}{\alpha_0}\left(i_\kappa-\frac{d_\kappa}{\alpha_\kappa}\right)\beta_{i_1;\ldots;i_\kappa;j}^{(d_1;\ldots;d_\kappa)}-\left(l-\frac{i_1+\cdots+i_\kappa-1}{\alpha_0}\right)\beta_{i_1;\ldots;i_\kappa-1;j}^{(d_1;\ldots;d_\kappa)},\quad(18)
$$

if $1\le j\le 1+\sum_{k=1}^{\kappa}i_k, i_k\le d_k,\ k=1,\ldots,\kappa$, with

$$
\beta_{i_1;\ldots;i_\kappa;j}^{(d_1;\ldots;d_\kappa+1)}=0,\quad\text{for all}\quad i_k\notin\{1,\ldots,d_k\},\ k=1,\ldots,\kappa,\quad\text{or}\quad j\notin\{1,\ldots,\sum_{k=1}^{\kappa}i_k\}.\qquad(19)
$$

Hence, (13) holds by induction for $\in\{1,\ldots,K\}$ with $d_\kappa\mapsto d_\kappa+1$, and the recursive formula to compute coefficients $\beta_{i_1;\ldots;i_\kappa;j}^{(d_1;\ldots;d_\kappa+1)}$ is given by (18) and (19).

## A.4 Complexity

If $\kappa=1$, the number of coefficients $\beta_{i_1;j}^{(d_1)}$ to be computed recursively in (13) is

$$
\sum_{h=1}^{d_1}\left(\sum_{i_1=1}^{h}\sum_{j=1}^{i_1}1\right)=\sum_{h=1}^{d_1}\frac{h(h+1)}{2}=\frac{1}{2}\left(\sum_{h=1}^{d_1}h^2+\sum_{h=1}^{d_1}h\right)=\frac{d_1(d_1+1)(d_1+2)}{6},
$$

which implies that the complexity is $\mathcal{O}(d_1^3)$. For $1 \leq \kappa \leq K$ clusters of size $d_1, \ldots, d_\kappa$, the number of coefficients $\beta_{i_1;\ldots;i_\kappa;j}^{(d_1;\ldots;d_\kappa)}$ to be computed in (13) is

$$\begin{aligned}
&\sum_{h=1}^{d_\kappa} \left( \sum_{i_1=1}^{d_1} \cdots \sum_{i_{\kappa-1}=1}^{d_{\kappa-1}} \sum_{i_\kappa=1}^{h} \sum_{j=1}^{i_1+\cdots+i_\kappa} 1 \right) = \sum_{h=1}^{d_\kappa} \sum_{i_\kappa=1}^{h} \sum_{i_1=1}^{d_1} \cdots \sum_{i_{\kappa-1}=1}^{d_{\kappa-1}} (i_1 + \cdots + i_\kappa) \\
&= \sum_{h=1}^{d_\kappa} \sum_{i_\kappa=1}^{h} \left\{ \frac{d_1(d_1+1)}{2} d_1 \cdots d_{\kappa-1} + \cdots + d_1 \cdots d_{\kappa-2} \frac{d_{\kappa-1}(d_{\kappa-1}+1)}{2} + d_1 \cdots d_{\kappa-1} i_\kappa \right\} \\
&= \left\{ \frac{d_1(d_1+1)}{2} d_2 \cdots d_{\kappa-1} \frac{d_\kappa(d_\kappa+1)}{2} \right\} + \cdots + \left\{ d_1 \cdots d_{\kappa-2} \frac{d_{\kappa-1}(d_{\kappa-1}+1)}{2} \frac{d_\kappa(d_\kappa+1)}{2} \right\} \\
&\quad + \left\{ d_1 \cdots d_{\kappa-1} \frac{d_\kappa(d_\kappa+1)(d_\kappa+2)}{6} \right\},
\end{aligned}$$

which implies that the total complexity for the computation of the coefficients $\beta_{i_1;\ldots;i_\kappa;j}^{(d_1;\ldots;d_\kappa)}$ in (13) is

$$\mathcal{O}\left( \sum_{k=1}^{\kappa} (d_1 + \cdots + d_k) d_1 \cdots d_{k-1} d_k^2 \right).$$

If the cluster size is the same for all clusters, i.e., $d_k = d_1$, for all $k = 2, \ldots, \kappa$, then we have $\mathcal{O}\left(\kappa \sum_{k=1}^{\kappa} d_1^{k+2}\right)$.