A PENALIZED LEAST-SQUARES ESTIMATOR FOR EXTREME-VALUE MODELS WITH MULTIPLE EXTREME DIRECTIONS

Anas Mourahib, Anna Kiriliouk, Johan Segers

REPRINT | 2026 / 22

UCLouvain

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

ISBA

Voie du Roman Pays 20 - L1.04.01

B-1348 Louvain-la-Neuve

Email : lidam-library@uclouvain.be

https://uclouvain.be/en/research-institutes/lidam/isba/publication

Computational Statistics and Data Analysis 225 (2027) 108460

Contents lists available at ScienceDirect

Computational Statistics and Data Analysis

journal homepage: www.elsevier.com/locate/csda

A penalized least-squares estimator for extreme-value models with multiple extreme directions

Anas Mourahib [a,*], Anna Kiriliouk [b], Johan Segers [c]

[a] Technical University of Eindhoven, Department of Mathematics and Computer Science, Eindhoven, the Netherlands
[b] UCLouvain, LIDAM/ISBA, Louvain-la-Neuve, Belgium
[c] KU Leuven, Department of mathematics, Leuven, and UCLouvain, LIDAM/ISBA, Louvain-la-Neuve, Belgium

ARTICLE INFO

Keywords:
Extreme directions
Max-linear model
Penalized least-squares
Stable tail dependence function
Threshold exceedances

ABSTRACT

Estimating the parameters of max-stable parametric models poses significant challenges, particularly when some parameters lie on the boundary of the parameter space. This situation arises when a subset of variables exhibits extreme values simultaneously, while the remaining variables do not—a phenomenon commonly referred to as an extreme direction. A novel estimator is proposed for the parameters of a general parametric mixture model, incorporating a threshold exceedances approach based on a pseudo-norm penalization. The latter plays a crucial role in accurately identifying parameters at the boundary of the parameter space. Additionally, the estimator comes with a data-driven algorithm to detect groups of variables corresponding to extreme directions. The performance of the estimator is assessed in terms of both parameter estimation and the identification of extreme directions through extensive simulation studies. Finally, the method is applied to two real-world datasets: discharge measurements at stations along the Danube river, and financial portfolio losses from stocks listed on the NYSE, AMEX, and NASDAQ. In both applications, the sets of variables that can become large simultaneously are identified.

1. Introduction

Consider a random vector $\boldsymbol{X} = (X_1, \dots, X_d)$ of risk factors and suppose the focus is on rare events. In this regime, neither parametric nor nonparametric methods are satisfactory: parametric models rely on distributional assumptions that can be dominated by the bulk of the data, and thus tend to underestimate extreme outcomes, while nonparametric techniques demand large sample size, which are often unavailable when studying extremes. Extreme value theory offers a rigorous framework for extrapolating beyond observed levels and has found applications in finance, hydrology, weather forecasting, and other fields.

Univariate extreme value theory, that is, when $d = 1$, has been extensively studied (De Haan and Ferreira, 2006; Beirlant et al., 2004; Coles, 2001) and modeling rare events in this case is done via a finite-dimensional parametric family of distributions. Multivariate modeling of $d \geqslant 2$ risk factors is a more challenging topic, since risks may exhibit some sort of dependence. The goal is not only to model rare events of each risk factor independently but to capture the extremal dependence between these risk factors. Two main approaches appear in this context: the block-maxima approach and the threshold exceedances approach. In the first approach, standardized componentwise maxima of $\boldsymbol{X}$ are asymptotically modeled using a max-stable random vector $\boldsymbol{Z} = (Z_1, \dots, Z_d)$. In the second approach, asymptotically, observations where at least one risk factor is extreme are modeled using a multivariate (generalized) Pareto random vector $\boldsymbol{Y} = (Y_1, \dots, Y_d)$ (Rootzén and Tajvidi, 2006).

* Corresponding author.
E-mail addresses: a.mourahib@tue.nl (A. Mourahib), anna.kiriliouk@uclouvain.be (A. Kiriliouk), jjjsegers@kuleuven.be (J. Segers).

https://doi.org/10.1016/j.csda.2026.108460
Received 22 August 2025; Received in revised form 3 April 2026; Accepted 27 July 2026
Available online 28 July 2026

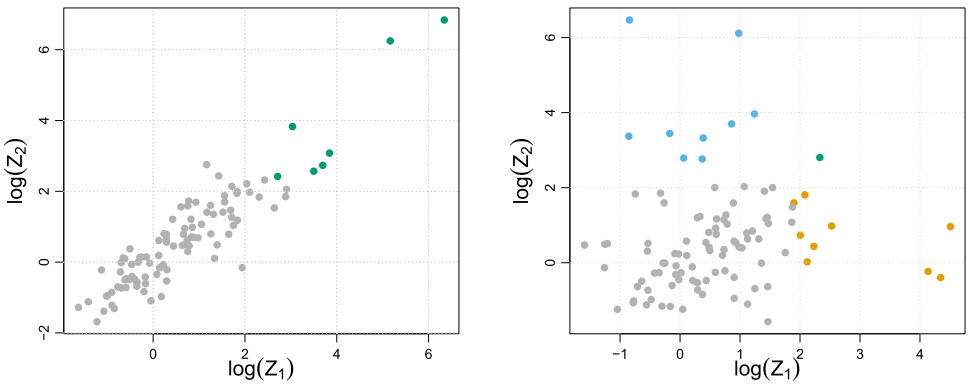

Fig. 1. A sample of size $n = 100$ from a max-stable Hüsler–Reiss model (left) and a max-stable mixture Hüsler–Reiss model (right). Points in the extreme direction $\{1, 2\}$ are shown in green, those in $\{1\}$ in orange, and those in $\{2\}$ in blue. (For interpretation of the references to colour in this figure legend, the reader is referred to the web version of this article.)

For both approaches, many parametric families have been proposed; popular examples include the Hüsler-Reiss model (Hüsler and Reiss, 1989) and the logistic model (Tawn, 1990), the latter being a special case of the scaled extremal Dirichlet model (Belzile and Nešlehová, 2017). Estimation methods for such models are numerous, for example, through censored likelihood (Padoan et al., 2010), weighted least squares (Einmahl et al., 2018), or based on neural networks (Lenzi et al., 2023; Richards et al., 2024).

A limitation of many parametric models is that they enforce all components of $\boldsymbol{X}$ to be extreme simultaneously. Fig. 1 (left) shows a log-transformed simulation from a max-stable Hüsler–Reiss model. Observations are close to the diagonal line $x = y$, suggesting that both components are large simultaneously. However, in practice, it may happen that only certain subsets $J \subseteq \{1, \dots, d\}$ of variables are jointly extreme while the others are not. Such a subset J is called an *extreme direction* (Mourahib et al., 2024; Simpson et al., 2020).

Multiple extreme directions can be accommodated in the max-linear model (Fougères et al., 2013; Einmahl et al., 2012) and the asymmetric logistic model (Tawn, 1990). However, the first model has a singular distribution (neither discrete nor absolutely continuous), making statistical inference complicated, while the second one may be too specific to represent a broad class of max-stable dependence structures. Moreover, existing estimation methods for these models are not designed to identify extreme directions, as this requires reliably detecting parameters on the boundary of the parameter space.

To capture multiple extreme directions, Mourahib et al. (2024) introduce the *mixture model* that extends the classic max-linear model and can represent *any* max-stable distribution. It is identified via:

- a $(d \times r)$ coefficient matrix A with entries in $[0, 1]$, whose columns correspond to distinct extreme directions of $\boldsymbol{X}$. Zero entries in A indicate which variables are not contained in a given extreme direction (Mourahib et al., 2024, Proposition 4.5);
- a second set of parameters that governs the dependence structure *within* each extreme direction.

A simulated sample from the mixture model is shown in Fig. 1 (right), where we can see that it is capable of not only representing the extreme direction $\{1, 2\}$ (points in black), but also $\{1\}$ and $\{2\}$ (points in orange and blue, respectively).

A natural question is how to estimate the model parameters. The weighted least-squares estimator of Einmahl et al. (2018) requires the true parameters to be in the interior of the parameter space. In contrast, the mixture model's reliance on zero entries in A pushes parameters to the boundary of the space, violating this condition. We propose an estimator based on threshold exceedances for the parameters of max-stable distributions that accommodate arbitrary collections of extreme directions, building on the least squares estimator introduced in Einmahl et al. (2018). A pseudo-norm L_p with $p \leqslant 1$ (see Section 3) is used to penalize the coefficient matrix A of the mixture model, which leads to zero entries in A and thus multiple extreme directions, in particular those different from $\{1, \dots, d\}$. The proposed penalized least squares estimator can simultaneously identify extreme directions *and* estimate the mixture model parameters. This sets our approach apart from the existing literature, which focuses exclusively on identifying extreme directions (Goix et al., 2017; Simpson et al., 2020; Meyer and Wintenberger, 2024).

A common criticism of extreme-value models possessing asymptotic dependence is their difficulty in handling situations where not all variables are extreme, that is, when the extreme directions differ from the full set $\{1, \dots, d\}$. This issue has motivated alternative approaches such as conditional extremes (Heffernan and Tawn, 2004) and geometric extremes (Wadsworth and Campbell, 2024). We address this limitation via our algorithm that can detect multiple extreme directions, including those involving only subsets of variables. We compare our approach with the DAMEX algorithm from Goix et al. (2017), which considers a subset $J \subseteq \{1, \dots, d\}$ to define an extreme direction if the corresponding cone $\mathcal{E}_J = \{\boldsymbol{x} \geqslant 0 : x_j > 0 \text{ iff } j \in J\}$ has positive mass under an empirical estimate of the exponent measure (Goix et al., 2017, Eq.20). However, as noted in Wadsworth and Campbell (2024), at finite levels, data rarely lie exactly on such cones when $J \neq \{1, \dots, d\}$, which can lead to missed directions. In contrast, our algorithm uses tuning parameters that can be selected via cross-validation, improving robustness and practical applicability.

The paper is structured as follows. Section 2 provides essential background on extreme value theory, focusing in particular on the notions of extreme directions (Simpson et al., 2020), the mixture model (Mourahib et al., 2024) and the least-squares estima-

tor (Einmahl et al., 2018). In Section 3, we introduce a penalized estimator for the mixture model, prove its consistency, and develop a data-driven algorithm to identify the extreme directions present in a given sample. The simulation study in Section 4 illustrates that the choice of the tuning parameters in our estimation procedure effectively reduces to the choice of a single penalization parameter, which we determine via cross-validation. Moreover, with this choice of tuning parameter, we show that our estimation procedure gives good results both in terms of estimation of the mixture model and extreme directions identification; in particular, it outperforms the DAMEX algorithm. Finally, in Section 5, we illustrate the practical relevance of our approach by applying the algorithm, first to river discharge data across stations from the Danube basin, and second to financial industry portfolios constructed from stocks listed on the NYSE, AMEX, and NASDAQ.

Notation. Let d be a positive integer and write $D = \{1, \dots, d\}$. Throughout, bold symbols will refer to multivariate quantities. Let $\mathbf{0}$ and $\mathbf{1}$ denote vectors of zeroes and ones, respectively, of a dimension clear from the context. Let $\boldsymbol{x} = (x_1, \dots, x_d)$ denote a d-dimensional vector and, for a non-empty set $J \subseteq \{1, \dots, d\}$, write $\boldsymbol{x}_J = (x_j)_{j \in J}$. Mathematical operations on vectors such as addition, multiplication and comparison are considered component-wise. For two vectors $\boldsymbol{a}$, $\boldsymbol{b}$, the set $[\boldsymbol{a}, \boldsymbol{b}]$ will denote the Cartesian product $\prod_{j=1}^{d} [a_j, b_j]$ and so forth for $(\boldsymbol{a}, \boldsymbol{b})$, $[\boldsymbol{a}, \boldsymbol{b})$, $(\boldsymbol{a}, \boldsymbol{b}]$. For two sets A, B, we define their symmetric difference by $A \triangle B = (A \setminus B) \cup (B \setminus A)$. For two non-negative integers d, r, let $[0,1]^{d \times r}$ denote the set of $d \times r$ matrices with entries in $[0,1]$. The arrow $\rightsquigarrow$ denotes convergence in distribution (weak convergence). For two vectors $\boldsymbol{x}$ and $\boldsymbol{y}$ in $\mathbb{R}^d$, define the *lexicographic order* $\leqslant_{\text{lex}}$ as

$$\boldsymbol{x} \leqslant_{\text{lex}} \boldsymbol{y} \quad \text{if and only if} \quad x_j < y_j \text{ for the first index } j \in D \text{ for which } x_j \neq y_j. \tag{1.1}$$

The lexicographic order is a total order in $\mathbb{R}^d$, meaning that either $\boldsymbol{x} \leqslant_{\text{lex}} \boldsymbol{y}$ or $\boldsymbol{y} \leqslant_{\text{lex}} \boldsymbol{x}$. For a set S, let $S^n = S \times \cdots \times S$ (n times). Finally, let $\mathbb{E} := [0, \infty)^d \setminus \{\mathbf{0}\}$.

Implementation. The source code used to conduct the experiments and generate the results in this paper is available at https://github.com/AnasMourahib/Penalized_least-squares_estimator, providing full transparency and supporting reproducibility.

2. Background

2.1. Multivariate extreme value distributions

For $i = 1, \dots, n$, let $\boldsymbol{X}_i = \left(X_{i1}, \dots, X_{id}\right)$ be iid copies of $\boldsymbol{X}$, a random vector on $\mathbb{R}^d$ with joint cumulative distribution function (cdf) F and margins $F_1, \dots, F_d$. We say that F is in the (max-)domain of attraction of a *max-stable* distribution G if there exist sequences $\boldsymbol{a}_n \in (0, \infty)^d$ and $\boldsymbol{b}_n \in \mathbb{R}^d$ such that

$$\frac{\max_{i=1,\dots,n} \boldsymbol{X}_i - \boldsymbol{b}_n}{\boldsymbol{a}_n} \rightsquigarrow G, \qquad n \to \infty.$$

The margins of G, denoted $G_1, \dots, G_d$, are univariate extreme-value distributions themselves, and any max-stable distribution G admits the representation

$$G(\boldsymbol{x}) = \exp\left\{-\ell\left(-\ln G_1(x_1), \dots, -\ln G_d(x_d)\right)\right\},$$

where $\ell : [0, \infty)^d \to [0, \infty)$ is the *stable tail dependence function (stdf)*. Then the distribution of $\boldsymbol{X}^* = (X_1^*, \dots, X_d^*)$ with $X_j^* = 1/\{1 - F_j(X_j)\}$ for $j \in D$ is in the domain of attraction of $G^*(\boldsymbol{x}) = \exp\{-\ell\{1/x_1, \dots, 1/x_d\}\}$, a max-stable distribution with unit-Fréchet margins.

The stable tail dependence function can be retrieved via

$$\ell(\boldsymbol{x}) = \lim_{t \downarrow 0} t^{-1} \, \mathsf{P}\left[F_1(X_1) \geqslant 1 - tx_1 \text{ or } \dots \text{ or } F_d(X_d) \geqslant 1 - tx_d\right]. \tag{2.1}$$

The function ℓ is convex and homogeneous in the sense that for $c \geqslant 0$ and $\boldsymbol{x} \in [0, \infty)^d$, we have $\ell(c\boldsymbol{x}) = c\ell(\boldsymbol{x})$. Moreover, unit-Fréchet margins of G^* imply that $\ell(0, \dots, 0, x_j, 0, \dots, 0) = x_j$ for $x_j \geqslant 0$ and $j \in D$. Finally, ℓ can be bounded by

$$\max(x_1, \dots, x_d) \leqslant \ell(x) \leqslant x_1 + \cdots + x_d, \tag{2.2}$$

where the lower bound corresponds to perfect tail dependence and the upper bound to asymptotic independence.

Suppose that $\boldsymbol{Z} = (Z_1, \dots, Z_d)$ follows a max-stable distribution with unit-Fréchet margins. For non-empty $J \subseteq D$, let ℓ_J be the stdf associated with $\boldsymbol{Z}_J$, that is,

$$\ell_J(\boldsymbol{x}_J) = \ell(\tilde{\boldsymbol{x}}), \qquad \boldsymbol{x}_J \in [0, \infty)^{|J|},$$

with $\tilde{\boldsymbol{x}} = \sum_{j \in J} x_j \boldsymbol{e}_j$ and $\boldsymbol{e}_j$ the j-th canonical unit vector in $\mathbb{R}^d$. The *extremal coefficients* (Schlather and Tawn, 2002) are defined via $\zeta_J := \ell(\sum_{j \in J} \boldsymbol{e}_j)$, for sets $J \subseteq D$ with $|J| \geqslant 2$. From (2.2), we get that $1 \leqslant \zeta_J \leqslant |J|$. The value ζ_J can be interpreted as the effective number of tail independent variables among $\boldsymbol{Z}_J$.

The *exponent measure* μ of G^* is the unique Borel measure concentrated on $\mathbb{E} = [0, \infty)^d \setminus \{\mathbf{0}\}$ such that

$$\mu\left(\mathbb{E} \setminus \left[\mathbf{0}, 1/\boldsymbol{x}\right]\right) = \ell(\boldsymbol{x}), \qquad \boldsymbol{x} \in [\mathbf{0}, \infty)^d.$$

The measure μ is finite on Borel sets that are bounded away from the origin. Moreover, $\mu(\mathbb{E} \setminus [\mathbf{0}, \boldsymbol{y}]) = \infty$ as soon as $y_j = 0$ for some $j \in D$.

2.2. Extreme directions and the mixture model

For non-empty $J \subseteq D$, define $\mathbb{E}_J := \{\boldsymbol{x} \in \mathbb{E} : \ x_j > 0 \text{ iff } j \in J\}$. Then $\{\mathbb{E}_J : \ \emptyset \neq J \subseteq D\}$ forms a partition of $\mathbb{E}$. Let $\boldsymbol{X}^*$ be again a d-dimensional random vector in the domain of attraction of a max-stable distribution with unit-Fréchet margins and exponent measure μ. The following definition was introduced in Simpson et al. (2020) in order to identify *extreme directions* of $\boldsymbol{X}^*$. Intuitively, an extreme direction is a group of components of $\boldsymbol{X}^*$ that may take on large values simultaneously while the other components are of smaller order.

Definition 2.1. The set J is said to be an extreme direction of $\boldsymbol{X}^*$ or μ if $\mu(\mathbb{E}_J) > 0$.

Note that μ has no mass outside $\bigcup_{J \in \mathcal{R}(\mu)} \mathbb{E}_J$, where $\mathcal{R}(\mu)$ is the collection of extreme directions of μ. Further, even if a certain set $J \subseteq D$ is not an extreme direction of $\boldsymbol{X}^*$, it is still possible that J is a superset or a subset of an extreme direction. Finally, note that every margin j appears in at least one extreme direction $J \in \mathcal{R}(\mu)$.

Many standard max-stable models are only appropriate for data exhibiting a single extreme direction $J = D$. Notable exceptions are the asymmetric logistic (Tawn, 1988) and the max-linear (Wang and Stoev, 2011; Einmahl et al., 2012) models. Mourahib et al. (2024) introduce the *mixture model*, a smoothed version of the max-linear model, which allows any collection of non-empty subsets of D that together covers D, so that no element of D is left out, to be the extreme directions of μ. The mixture model is key to modeling data with non-standard dependence structures as it covers *all* max-stable distributions.

Definition 2.2 (Mourahib et al., 2024)**.**
Let r be a positive integer and write $R = \{1, \dots, r\}$. Let $\boldsymbol{Z}_{\cdot 1}, \dots, \boldsymbol{Z}_{\cdot r}$ be independent random vectors in $\mathbb{R}^d$. For all $s \in R$, suppose that $\boldsymbol{Z}_{\cdot s} = (Z_{1s}, \dots, Z_{ds})$ are max-stable random vectors with unit-Fréchet margins, stdf $\ell^{(s)}$ and a single extreme direction D. Let $A = (a_{js})_{j \in D; s \in R}$ be an element of Θ_A, the set of $d \times r$ matrices with elements a_{js} in $[0, 1]$, unit row sums $\sum_{s=1}^{r} a_{js} = 1$ for all $j \in D$ and positive column sums $\sum_{j=1}^{d} a_{js} > 0$ for all $s \in R$. The d-variate random vector

$$\boldsymbol{M} = (M_1, \dots, M_d) = \left(\max_{s \in R} \{a_{1s} Z_{1s}\}, \dots, \max_{s \in R} \{a_{ds} Z_{ds}\} \right), \tag{2.3}$$

will be called the $d \times r$ *mixture model* with coefficient matrix A and factors $\boldsymbol{Z}_{\cdot 1}, \dots, \boldsymbol{Z}_{\cdot r}$.

Remark 2.3. Note that two different coefficient matrices, containing the same columns but in different orders, will lead to the same mixture model $\boldsymbol{M}$ if the distributions of $\boldsymbol{Z}_{\cdot 1}, \dots, \boldsymbol{Z}_{\cdot r}$ are identical. To ensure identifiability, we will always assume that the columns of A are decreasingly ordered according to the lexicographic order (1.1).

In (2.3), for a given $j \in D$, only those $s \in R$ that satisfy $a_{js} > 0$ matter. Let

$$J_s := \{j \in D : \ a_{js} > 0\}, \tag{2.4}$$

denote the s-th *signature* of the coefficient matrix A.

Theorem 2.4 (Stephenson, 2003; Mourahib et al., 2024)**.** *The random vector $\boldsymbol{M}$ in (2.3) follows a max-stable distribution with unit-Fréchet margins and stdf*

$$\ell(\boldsymbol{x}) = \sum_{s \in R} \ell_{J_s}^{(s)} \left((a_{js} x_j)_{j \in J_s} \right), \qquad \boldsymbol{x} \in [0, \infty)^d. \tag{2.5}$$

Zero entries of the coefficient matrix A allow some groups of components of the random vector $\boldsymbol{M}$ to constitute an extreme direction. Mourahib et al. (2024, Proposition 4.5) show that the set of extreme directions of the mixture model is given by the signatures of its coefficient matrix A. The model is general in the sense that any max-stable distribution with unit-Fréchet margins is the cdf of a mixture model whose coefficient matrix has mutually different signatures (Mourahib et al., 2024, Theorem 4.8).

Example 2.5 (Mixture logistic model)**.** For $s \in R$, suppose that the stdf associated with $\boldsymbol{Z}_{J_s, s}$ is

$$\ell_{J_s}^{(s)}(x_1, \dots, x_{|J_s|}) = \left(x_1^{1/\alpha_s} + \cdots + x_{|J_s|}^{1/\alpha_s} \right)^{\alpha_s},$$

for $\boldsymbol{x} \in [0, \infty)^{|J_s|}$, where $0 < \alpha_s < 1$. The stdf of the *mixture logistic model* $\boldsymbol{M}$ is

$$\ell(\boldsymbol{x}) = \sum_{s \in R} \left\{ \sum_{j \in J_s} (a_{js} x_j)^{1/\alpha_s} \right\}^{\alpha_s}, \qquad \boldsymbol{x} \in [0, \infty)^d.$$

In the rest of the paper, we will assume that each extreme direction only corresponds to a single column of the coefficient matrix A, i.e., we will assume that the signatures of A are mutually different. Under this assumption, the mixture logistic model coincides with the asymmetric logistic model.

Example 2.6 (Mixture Hüsler–Reiss model)**.** Let $s \in R$. For $|J_s| > 1$, suppose that $\ell_{J_s}^{(s)}$ is the Hüsler–Reiss stdf associated with the $(|J_s| \times |J_s|)$ variogram matrix $\Gamma^{(s)}$; see, e.g., Huser and Davison (2013). That is,

$$\ell_{J_s}^{(s)}(\boldsymbol{x}) = \sum_{j \in J_s} x_j \, \Phi_{|J_s|-1}\left(\boldsymbol{\eta}^{j,(s)}(\boldsymbol{x}); \Sigma^{j,(s)} \right), \qquad \boldsymbol{x} \in [0, \infty)^{J_s},$$

where for $j \in J_s$, $\boldsymbol{\eta}^{j,(s)}(\boldsymbol{x}) := \left\{ \ln(x_j/x_t) + \Gamma^{(s)}_{jj}/2 \right\}_{t \in J_s; t \neq j}$ where by convention, $0/0$ is set to ∞, and where the $\left(|J_s \setminus \{j\}| \times |J_s \setminus \{j\}|\right)$ matrix $\Sigma^{j,(s)}$ is defined by

$$\Sigma^{j,(s)} := \frac{1}{2}\left\{ \Gamma^{(s)}_{jt} + \Gamma^{(s)}_{jt'} - \Gamma^{(s)}_{tt'} \right\}_{t,t' \in J_s \setminus \{j\}},$$

and finally, where $\Phi_m(\cdot\,; \Sigma)$ is the cdf of the centered m-variate normal distribution with covariance matrix Σ. The stdf of the mixture model $\boldsymbol{M}$ is

$$\ell(\boldsymbol{x}) = \sum_{s \in R} \left\{ \sum_{j \in J_s} a_{js} x_j \, \Phi_{|J_s|-1}\left(\boldsymbol{\eta}^{j,(s)}\left(\left(a_{js} x_j\right)_{j \in J_s} \right); \Sigma^{j,(s)} \right) \right\}, \qquad \boldsymbol{x} \in [0,\infty)^d.$$

2.3. Least squares estimator of tail dependence

Let again $\boldsymbol{X}_1, \ldots, \boldsymbol{X}_n$ denote iid random vectors in the domain of attraction of a max-stable distribution with stdf ℓ. We will first recall how to estimate ℓ non-parametrically. Let $k = k_n \in (0, n]$ be an intermediate sequence such that

$$k \to \infty \quad \text{and} \quad k/n \to 0, \quad \text{as } n \to \infty. \tag{2.6}$$

The *tail fraction* k/n determines the fraction of the data that is considered as extreme and that will be used for inference. A standard non-parametric estimator of the stdf is

$$\hat{\ell}_{n,k}(\boldsymbol{x}) = \frac{1}{k} \sum_{i=1}^{n} \mathbf{1}\left\{ R^n_{i1} > n + 1/2 - k x_1 \text{ or } \ldots \text{ or } R^n_{id} > n + 1/2 - k x_d \right\}, \tag{2.7}$$

where R^n_{ij} denotes the rank of X_{ij} among $X_{1j}, \ldots, X_{nj}$ for $j = 1, \ldots, d$. Note that (2.7) is a minor modification of the standard empirical stdf (Drees and Huang, 1998). Alternatives are for example the bias-corrected estimator proposed in Beirlant et al. (2016) or a smooth estimator based on the empirical beta copula (Kiriliouk et al., 2018).

Suppose now that ℓ belongs to a parametric family $\{\ell(\cdot\,; \boldsymbol{\theta}) : \boldsymbol{\theta} \in \Theta\}$, with $\Theta \subseteq \mathbb{R}^{p_\Theta}$ for some integer $p_\Theta \geqslant 1$. Furthermore, assume the existence of a unique parameter $\boldsymbol{\theta}_0 \in \Theta$ such that $\ell(\boldsymbol{x}) = \ell(\boldsymbol{x}; \boldsymbol{\theta}_0)$ for all $\boldsymbol{x} \in [0,\infty)^d$. Einmahl et al. (2018) propose a continuous updating weighted least-squares estimator for $\boldsymbol{\theta}_0$, which, in its simplest form, reduces to the ordinary least-squares estimator

$$\hat{\boldsymbol{\theta}}_{n,k} := \operatorname*{arg\,min}_{\boldsymbol{\theta} \in \Theta} \sum_{m=1}^{q} \left(\hat{\ell}_{n,k}(\boldsymbol{c}_m) - \ell(\boldsymbol{c}_m; \boldsymbol{\theta}) \right)^2. \tag{2.8}$$

Here, $\boldsymbol{c}_1, \ldots, \boldsymbol{c}_q \in [0,\infty)^d$, where $\boldsymbol{c}_m = (c_{m1}, \ldots, c_{md})$ for $m = 1, \ldots, q$ are the q points in which ℓ and $\hat{\ell}_{n,k}$ are evaluated. These grid points $\boldsymbol{c}_1, \ldots, \boldsymbol{c}_q$ must be chosen such that the mapping $L : \Theta \to \mathbb{R}^q$, $\boldsymbol{\theta} \mapsto \{\ell(\boldsymbol{c}_m; \boldsymbol{\theta}),\ m = 1, \ldots, q\}$, is one-to-one, ensuring that the parameter vector $\boldsymbol{\theta}$ is identifiable from $\{\ell(\boldsymbol{c}_m; \boldsymbol{\theta}), m = 1, \ldots, q\}$; see (Einmahl et al., 2018, Section 2.2). The estimator (2.8) is easy to compute and shows good performance for many well-known max-stable models (Einmahl et al., 2018, Section 3). However, it is not suitable for the mixture model, where some parameters may lie on the boundary of the parameter space. An extension is proposed in Kiriliouk (2020), but unfortunately, it is based on conditions that may be difficult to verify.

3. Estimation of the mixture model

In Section 3.1, we will start by introducing our estimation procedure for the simplified setting where the number of extreme directions, r, is assumed to be known. The case where r is unknown will be discussed in Section 3.2.

3.1. Known number of extreme directions

As in Definition 2.2, let $\boldsymbol{M}$ be the $d \times r$ mixture model with coefficient matrix A and factors $\boldsymbol{Z}_{\cdot 1}, \ldots, \boldsymbol{Z}_{\cdot r}$ with stdfs $\ell^{(1)}, \ldots, \ell^{(r)}$. Recall $R = \{1, \ldots, r\}$. From now on, suppose that the stdfs $\ell^{(1)}, \ldots, \ell^{(r)}$ belong to a common parametric family $\{\ell(\cdot\,; \boldsymbol{\theta}_Z) : \boldsymbol{\theta}_Z \in \Theta_Z\}$, with $\Theta_Z \subseteq \mathbb{R}^v$ for some integer $v \geqslant 1$, and that for each $s \in R$, there exists a unique parameter vector $\boldsymbol{\theta}^{(s)}_Z \in \Theta_Z$ such that $\ell^{(s)} = \ell(\cdot\,; \boldsymbol{\theta}^{(s)}_Z)$. In what follows, for simplicity, we assume that the dependence parameters across the columns are identical, meaning that all $\boldsymbol{\theta}^{(s)}_Z$, $s \in R$, are equal to a common $\boldsymbol{\theta}_Z \in \Theta_Z$. However, all the results remain valid in the general case.

For $\boldsymbol{\theta} = (A, \boldsymbol{\theta}_Z) \in [0,1]^{d \times r} \times \Theta_Z$, define $\ell(\cdot\,; \boldsymbol{\theta}) : [0,\infty)^d \to [0,\infty)$ by

$$\ell(\boldsymbol{x}, \boldsymbol{\theta}) = \sum_{s \in R} \ell_{J_s}\left(\left(a_{js} x_j\right)_{j \in J_s}; \boldsymbol{\theta}_Z \right), \qquad \boldsymbol{x} \in [0,\infty)^d.$$

Recall $\Theta_A \subset [0,1]^{d \times r}$ in Definition 2.2. If $A \in \Theta_A$, then $\ell(\cdot\,, \boldsymbol{\theta})$ is the stdf of the mixture model $\boldsymbol{M}$ in Theorem 2.4, but $\ell(\cdot\,, \boldsymbol{\theta})$ is defined for $A \in [0,1]^{d \times r} \setminus \Theta_A$ too. We use the extended parameter space $[0,1]^{d \times r}$ in the preliminary procedure (3.1), and we standardize to unit row sums later in (3.2).

The *preliminary non-standardized penalized least-squares (PNPLS) estimator* is defined as

$$\hat{\boldsymbol{\theta}}_{\mathrm{pr}} = \left(\hat{A}_{\mathrm{pr}}, \hat{\boldsymbol{\theta}}_{Z,\mathrm{pr}}\right) = \operatorname*{arg\,min}_{\boldsymbol{\theta} = (A, \boldsymbol{\theta}_Z) \in [0,1]^{d \times r} \times \Theta_Z} \left\{ \sum_{m=1}^{q} \left(\hat{\ell}_{n,k}(\boldsymbol{c}_m) - \ell(\boldsymbol{c}_m; \boldsymbol{\theta}) \right)^2 + \lambda \mathcal{P}(A) \right\}, \tag{3.1}$$

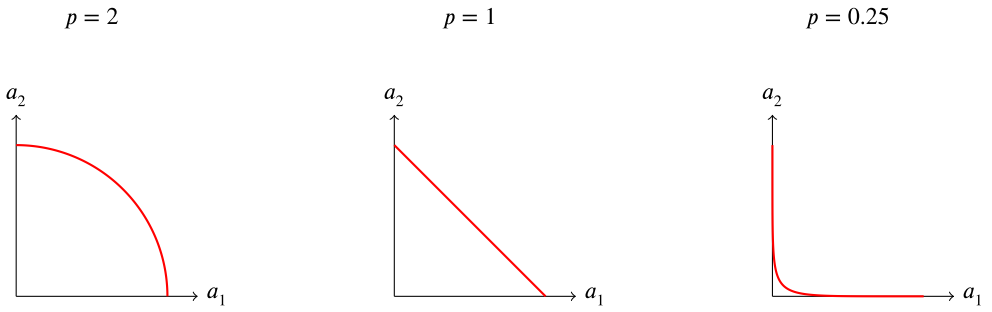

Fig. 2. Bivariate contours of $(\Sigma_s a_s^p)^{1/p} = 1$ for given values of p.

where $\lambda > 0$ is called the *penalization parameter* and where the penalty is defined as $\mathcal{P}(A) = \sum_{j\in D}(\sum_{s\in R} a_{js}^p)^{1/p}$ for some *penalization exponent* $p \in (0, 1]$.

For $\lambda = 0$, the PNPLS estimator is equal to the ordinary least-squares estimator defined in (2.8), provided A is restricted to lie in Θ_A. The penalization term $\lambda\mathcal{P}(A)$ in (3.1) is needed to find zero entries in the coefficient matrix A. We restrict our attention to penalization exponent $p \leqslant 1$. Geometrically, using a penalization $\mathcal{P}(A)$ with $p \leqslant 1$ promotes sparsity because its level sets are spiky and tend to intersect the loss contours at sparse solutions, similar to the well-known lasso ($p = 1$). This is in contrast to the case $p > 1$, illustrated in Fig. 2. In Remark 3.3, we discuss some additional arguments for the choice $p \leqslant 1$ related to the minimization algorithm.

The minimization problem in (3.1) is non-convex and is not guaranteed to give a global minimum. The parameter vector $\boldsymbol{\theta}_Z$, corresponding to the latent factors $\boldsymbol{Z}_{\cdot 1}, \ldots, \boldsymbol{Z}_{\cdot r}$, may be particularly difficult to estimate. A straightforward way to improve the PNPLS estimator is as follows. After obtaining $\hat{\boldsymbol{\theta}}_{\text{pr}}$, we follow-up with a two-step estimation procedure, where each step minimizes (3.1) with respect to either $\boldsymbol{\theta}_Z \in \Theta_Z$ or $A \in \Theta_A$. More precisely,

1. let $\hat{A}_{\text{new}}$ be the partial solution of (3.1) w.r.t. A, where $\boldsymbol{\theta}_Z$ is fixed to $\hat{\boldsymbol{\theta}}_{Z,\text{pr}}$;
2. write $\hat{A}_{\text{new}} = (\hat{a}_{js,\text{new}})_{j\in D, s\in R}$ and standardize to ensure unit row sums,

$$\hat{a}_{js} = \hat{a}_{js,\text{new}} \Big/ \sum_{l=1,\ldots,r} \hat{a}_{jl,\text{new}}, \qquad j \in D, s \in R. \tag{3.2}$$

   This yields $\hat{A} \in [0, 1]^{d\times r}$ with unit row sums;
3. let $\hat{\boldsymbol{\theta}}_Z$ be the partial solution of (3.1) w.r.t. $\boldsymbol{\theta}_Z$, with $\lambda = 0$, where A is fixed to $\hat{A}$.

The resulting estimate $\hat{\boldsymbol{\theta}} = (\hat{A}, \hat{\boldsymbol{\theta}}_Z)$ is called the *penalized least-squares (PLS) estimator*. Note that numerically, we find that $\hat{A}_{\text{new}} \neq \hat{A}_{\text{pr}}$. In fact, $\hat{A}_{\text{pr}}$ is not a global minimum. The two steps of the estimation procedure can also be iteratively alternated until a predefined stopping criterion is met. However, as we will demonstrate in the simulation studies in Section 4, performing these steps only once is typically sufficient to achieve satisfactory results in practice. The full estimation procedure is explained in Algorithm 1.

Remark 3.1. Since A has unit row sums by definition, one can argue that the minimization problem (3.1) can be reduced to a lower-dimensional formulation where $\Theta_A = [0, 1]^{d\times(r-1)}$ by eliminating the column $\boldsymbol{a}_{\cdot r}$ and imposing the constraint $a_{j1} + \cdots + a_{jr-1} \leqslant 1$, for each $j \in D$. However, such an approach penalizes only the first $r - 1$ columns of the matrix A, leading to an asymmetric treatment of the parameters.

Remark 3.2. By Theorem 2.4, the extreme directions of a mixture model correspond to the set of signatures of its coefficient matrix A. For $s \in R$, the sets $\hat{J}_s = \{j \in D : \hat{a}_{js} > 0\}$ thus explicitly provide estimates of the extreme directions associated with the data.

Minimization algorithm and choice of p. To solve the minimization problem (3.1), we use the bounded limited-memory modification of the BFGS quasi-Newton method (`L-BFGS-B` algorithm in `R`), see (Byrd et al., 1995). This iterative method has the update step

$$\boldsymbol{\theta}_{m+1} = \boldsymbol{\theta}_m - \gamma_m H_m^{-1} \nabla L(\boldsymbol{\theta}_m), \qquad m \geqslant 0, \tag{3.4}$$

where L is the loss function in (3.1), $\gamma_m > 0$ is the step-size, H_m is a $(d \times r + v) \times (d \times r + v)$ approximate curvature positive definite matrix capturing second-order information based on the gradients observed during optimization, and $\nabla L(\boldsymbol{\theta}_m)$ is the gradient of L evaluated at $\boldsymbol{\theta}_m$. At each step, the gradient is approximated using a finite-difference method.

Remark 3.3. Another motivation for choosing the penalization exponent $p \in (0, 1]$ is that the update direction in (3.4) is opposite to the gradient. For a fixed parameter a_{js}, the penalization term in (3.1) then contributes to the gradient proportionally to

$$T(p) = a_{js}^{p-1} \left(\sum_{t\in R} a_{jt}^p \right)^{1/p-1}. \tag{3.5}$$

Recall that $p < 1$ and thus $p - 1 < 0$. Consequently, if the true value of a_{js} is zero, then as the corresponding coefficient approaches zero at some step of the minimization algorithm L-BFGS-B, the term $T(p)$ in Eq. (3.5) diverges to $+\infty$. As a result, the gradient also diverges to $+\infty$, and since the update direction in (3.4) is opposite to the gradient, this leads to a solution where $\hat{a}_{js} = 0$.

Algorithm 1 Estimation of the parameters of a mixture model for known r.

Input: integers d, $r \geqslant 1$, points $\boldsymbol{c}_1, \ldots, \boldsymbol{c}_q \in [0,\infty)^d$, sample $\mathcal{S} = \boldsymbol{X}_1, \ldots, \boldsymbol{X}_n \in [0,\infty)^d$, sample size $n \in \mathbb{N}$, integer $k \leqslant n$, reals $0 < p \leqslant 1$ and $\lambda > 0$.

Output: penalized least-squares estimator $\hat{\boldsymbol{\theta}}$.

1: Compute the empirical stdf estimates $\left(\hat{\ell}_{n,k}(\boldsymbol{c}_m)\right)_{m=1,\ldots,q}$ based on the sample $\mathcal{S}$.
2: **Preliminary non-standardized penalized least-squares estimator**: Compute the parameter estimate $\hat{\boldsymbol{\theta}}_{\mathrm{pr}} = (\hat{A}_{\mathrm{pr}}, \hat{\boldsymbol{\theta}}_{Z,\mathrm{pr}}) \in \Theta_A \times \Theta_Z$ as the solution of (3.1) based on empirical stdf estimates calculated in Step 1.
3: **Step 1. Update the coefficient matrix**: Compute $\hat{A}_{\mathrm{new}}$ as the solution of (3.1) with $\boldsymbol{\theta}_Z$ fixed at $\hat{\boldsymbol{\theta}}_{Z,\mathrm{pr}}$ and using the starting value $\hat{A}_{\mathrm{pr}}$.
4: **Step 2. Standardization**: Apply the transformation (3.2) to $\hat{A}_{\mathrm{new}}$ to ensure unit row sums, leading to a matrix $\hat{A}$.
5: **Step 3. Update the dependence parameter**: Compute $\hat{\boldsymbol{\theta}}_{Z,\mathrm{new}}$ as the solution of (3.1) with $\lambda = 0$, A fixed at $\hat{A}$ and using the starting value $\hat{\boldsymbol{\theta}}_{Z,\mathrm{pr}}$.
6: **Standardized penalized least-squares estimator:** Define

$$\hat{\boldsymbol{\theta}} := (\hat{A}, \hat{\boldsymbol{\theta}}_Z) = (\hat{A}_{\mathrm{new}}, \hat{\boldsymbol{\theta}}_{Z,\mathrm{new}}). \tag{3.3}$$

7: **return** $(\hat{\boldsymbol{\theta}})$.

Choice of tuning parameters. There are three tuning parameters for the estimation procedure in Algorithm 1: the tail fraction k/n, the penalization exponent p and the penalization parameter λ. First, the choice of the tail fraction is a classical trade-off between bias and variance; a smaller k/n or larger k/n introduces more variance or more bias, respectively, for the empirical stdf in (2.7). Second, for $0 < p_1 \leqslant p_2 < 1$, we have $T(p_1)/T(p_2)$ explodes to ∞ when a_{js} is closer to zero, with T defined in (3.5). The same reasoning after (3.5) shows that a smaller penalization exponent yields stronger penalization. Different choices of p will be illustrated in Section 4. Finally, the penalization parameter λ will be chosen using the K-fold cross validation in Algorithm 2.

Algorithm 2 K-fold cross validation score for a penalization parameter λ.

Input: integers d, $r \geqslant 1$, points $\boldsymbol{c}_1, \ldots, \boldsymbol{c}_q \in [0,\infty)^d$, sample $\mathcal{S} = \boldsymbol{X}_1, \ldots, \boldsymbol{X}_n \in [0,\infty)^d$, sample size $n \in \mathbb{N}$, integer $k \leqslant n$, reals $0 < p \leqslant 1$ and $\lambda > 0$.

Output: cross validation score associated to the penalization parameter λ.

1: Split the sample $\mathcal{S}$ into K (approximately) equal-sized folds $\mathcal{S}_1, \ldots, \mathcal{S}_K$.
2: **for** each fold $i \in \{1, \ldots, K\}$ **do**
3: **Validation set:** set $\mathcal{S}_i$ as the validation set.
4: **Training set:** set $\mathcal{S} \setminus \mathcal{S}_i$ (all other folds) as the training set.
5: Let $\hat{\boldsymbol{\theta}}^{(-i)}$ be the output of Algorithm 1 with parameters d, r, points $\boldsymbol{c}_1, \ldots, \boldsymbol{c}_q$, sample $\mathcal{S} \setminus \mathcal{S}_i$, sample size $\lfloor n(K-1)/k \rfloor$, integer $\lfloor k(K-1)/K \rfloor$, reals p and λ.
6: Calculate the empirical stdf estimates $\left(\hat{\ell}^{(i)}_{\lfloor n/K \rfloor, \lfloor k/K \rfloor}(\boldsymbol{c}_m)\right)_{m=1,\ldots,q}$ using the validation set.
7: Evaluate the trained model on the validation set using the i-th score

$$\mathrm{CV}^{(-i)}(\lambda) = \sum_{m=1}^{q} \left(\hat{\ell}^{(i)}_{\lfloor n/K \rfloor, \lfloor k/K \rfloor}(\boldsymbol{c}_m) - \ell\left(\boldsymbol{c}_m; \hat{\boldsymbol{\theta}}^{(-i)}\right)\right)^2.$$

8: **end for**
9: Compute the average cross validation score along all folds via

$$\mathrm{CV}(\lambda) = \frac{1}{K} \sum_{i=1}^{K} \mathrm{CV}^{(-i)}(\lambda).$$

10: **return** $\mathrm{CV}(\lambda)$.

The penalized least squares estimator is consistent when we restrict the search space in the optimization problem in (3.1) to a compact subset K of the parameter space $\Theta_A \times \Theta_Z$ that contains the true parameter. By continuity of the objective function, a global minimum always exists on K. The minimizer converges in probability to the true parameter under an identifiability condition. Recall the empirical stdf $\hat{\ell}_{n,k}$ in (2.7).

Theorem 3.4. *Let $\boldsymbol{X}_1, \ldots, \boldsymbol{X}_n$ be iid random vectors on $\mathbb{R}^d$ with continuous marginal distribution and stdf $\ell = \ell(\cdot, \boldsymbol{\theta}_0) \in \{\ell(\cdot, \boldsymbol{\theta}) : \boldsymbol{\theta} \in \Theta_A \times \Theta_Z\}$, following a parametrically specified mixture model as in the beginning of Section 3.1. Assume the points $\boldsymbol{c}_1, \ldots, \boldsymbol{c}_m \in [0,\infty)^d$ are such that the map $\boldsymbol{\theta} \mapsto (\ell(\boldsymbol{c}_m, \boldsymbol{\theta}))_{m=1}^q$ is continuous and injective on $\Theta_A \times \Theta_Z$. Let $K \subset \Theta_A \times \Theta_Z$ be compact and contain the true*

parameter $\theta_0 = (A_0, \theta_{Z,0})$ *and define*

$$\hat{\theta}_K = \underset{\theta \in K}{\arg\min} \left\{ \sum_{m=1}^{q} \left(\hat{\ell}_{n,k}(\boldsymbol{c}_m) - \ell(\boldsymbol{c}_m; \theta) \right)^2 + \lambda \mathcal{P}(A) \right\}.$$

for $k = k_n \in (0, n]$ *and* $\lambda = \lambda_n \geqslant 0$. *Then* $\hat{\theta}_K \to \theta_0$ *in probability as soon as* $\lambda \to 0$, $k \to \infty$ *and* $k/n \to 0$ *as* $n \to \infty$.

The proof of Theorem 3.4 is given in Appendix A.

3.2. Identifying the extreme directions

In practice, the number of extreme directions r is unknown. Using the same assumptions and notation as in Section 3.1, we now estimate θ without taking any prior knowledge of the number of columns, r, of A into account. Our proposed method is iterative: at each step $t \geqslant 1$, we fix the number of columns of A to t and compute the standardized penalized least-squares estimator $\hat{\theta} = (\hat{A}, \hat{\theta}_Z) \in \Theta_A \times \Theta_Z$ using Algorithm 1. We then iterate to $t+1$ as long as all columns of $\hat{A}$ remain non-null. The method is presented in Algorithm 3.

Algorithm 3 Estimation of the extreme directions associated with a data sample.

Input: integer $d \geqslant 1$, points $\boldsymbol{c}_1, \ldots, \boldsymbol{c}_q \in [0, \infty)^d$, sample $\mathcal{S} = \boldsymbol{X}_1, \ldots, \boldsymbol{X}_n \in [0, \infty)^d$, sample size $n \in \mathbb{N}$, integer $k \leqslant n$, reals $0 < p \leqslant 1$ and $\lambda > 0$.

Output: extreme directions associated with the sample $\mathcal{S}$.

1: Set $t = 1$.
2: Let $\hat{\theta}_t = (\hat{A}, \hat{\theta}_Z)$ be the output of Algorithm 1 with parameters d, t, points $\boldsymbol{c}_1, \ldots, \boldsymbol{c}_q$, sample $\mathcal{S}$, sample size n, integer $k \leqslant n$, reals p and λ.
3: Compute the signatures $\hat{\boldsymbol{J}}_s = \{ j \in D : \hat{a}_{js} > 0 \}$ for $s = 1, \ldots, t$ and let $\hat{\mathcal{J}} = \{\hat{\boldsymbol{J}}_1, \ldots, \hat{\boldsymbol{J}}_t\}$.
4: **while** $\varnothing \notin \hat{\mathcal{J}}$ **do**
5: Iterate $t \leftarrow t + 1$.
6: Repeat Steps 2 and 3.
7: **end while**
8: Remove $\{\varnothing\}$ from $\hat{\mathcal{J}}$, that is, update $\hat{\mathcal{J}} \leftarrow \hat{\mathcal{J}} \setminus \{\varnothing\}$.
9: **return** $\hat{\mathcal{J}}$.

The DAMEX algorithm introduced by Goix et al. (2017, Algorithm 1) can also be used to identify extreme directions. It relies on three key parameters: a threshold beyond which data are considered extreme, analogous to the tail fraction k/n used in Algorithm 1; a tolerance parameter that heuristically defines the region associated with an extreme direction; and a mass threshold, which determines whether a subset $J \subseteq V$ with sufficiently large mass is considered an extreme direction. As explained in Goix et al. (2017), in a supervised setting, these parameters are selected via cross-validation. As we will detail in the simulation study in Section 4, this is also the case for Algorithm 1: the choice of the tail fraction involves a trade-off between bias and variance. In contrast, the choice of the penalization exponent p is less critical, provided that the penalization parameter λ is appropriately tuned-this being achieved, as outlined in Algorithm 2, through K-fold cross validation.

Our procedure is tailored to max-stable mixture models, motivated by Mourahib et al. (2024, Theorem 4.8), who show that every max-stable distribution with unit-Fréchet margins can be represented as a max-stable mixture model. In practice, the factors $\boldsymbol{Z}_{\cdot 1}, \ldots, \boldsymbol{Z}_{\cdot t}$ are assumed to follow a parametric model, ensuring that each factor has a single extreme direction D. While this could be seen as a limitation of our method, it has little impact on the identification of extreme directions. Indeed, for max-stable mixture models, the set of extreme directions is determined solely by the matrix A, and does not depend on the specific parametric form imposed on the factors. This claim will be supported by simulation results in Section 4.

4. Simulation study

In this section, we conduct a simulation study to evaluate the estimation procedure introduced in Section 3. We begin by outlining the setup used throughout the simulations in Section 4.1. First we consider the setting where the number of extreme directions is known, as described in Algorithm 1 and detailed in Section 4.2. We then turn to the more general case in which the number of extreme directions is unknown, corresponding to Algorithm 3, and discussed in Section 4.3.

4.1. Preliminaries

Coefficient matrix. Throughout the simulation study, we fix the dimension $d = 4$ and the coefficient matrix

$$A = \begin{pmatrix} 1/3 & 1/3 & 1/3 & 0 \\ 1/2 & 0 & 0 & 1/2 \\ 0 & 3/4 & 1/4 & 0 \\ 0 & 1/2 & 0 & 1/2 \end{pmatrix}. \tag{4.1}$$

Domain of attraction. We simulate samples $\boldsymbol{X}_1, \dots, \boldsymbol{X}_n$ of size n in the domain of attraction of a mixture model $\boldsymbol{M}$ as defined in (2.3). To do so, we simply perturb $\boldsymbol{M}$ by adding a lighter-tailed noise. More precisely,

$$\boldsymbol{X}_i = \boldsymbol{M}_i + \boldsymbol{N}_i, \qquad i = 1, \dots, n, \tag{4.2}$$

where for $i = 1, \dots, n$, $\boldsymbol{M}_i \in \mathbb{R}^d$ follows a mixture model and $\boldsymbol{N}_i \in \mathbb{R}^d$ has independent centered Gaussian margins with variance $\sigma^2 > 0$. Algorithm 4 in Appendix B outlines the procedure for simulating from a mixture model, relying on the generation of the max-stable random vectors $\boldsymbol{Z}_{\cdot 1}, \dots, \boldsymbol{Z}_{\cdot r}$. This is feasible for most popular models; see (Dombry et al., 2016, Algorithm 1) or the `R` package `mev` (Belzile et al., 2024). Furthermore, code for Algorithm 4 is freely accessible for both the mixture logistic model and the mixture Hüsler–Reiss model; see https://github.com/AnasMourahib/Penalized_least-squares_estimator/blob/main/Functions/Simulation_mixture_model.R for the `R` implementation of Algorithm 4 for the mixture logistic and the mixture Hüsler–Reiss models.

Starting points in the minimization algorithm. To calculate the PNPLS estimator, we need to provide good starting values for $\boldsymbol{\theta} = (A, \boldsymbol{\theta}_Z)$. Suppose that the number of extreme directions is fixed to t, and therefore $A \in [0,1]^{d \times t}$. Consider a d-variate sample $\boldsymbol{X}_1, \dots, \boldsymbol{X}_n$ and write $\boldsymbol{X}_i = (X_{i1}, \dots, X_{id})$, for $i = 1, \dots, n$. We determine a starting value for A by applying the k-means clustering algorithm with t clusters to the unit-Pareto transformed data,

$$\left(\frac{n}{n+1-R_{i1}}, \dots, \frac{n}{n+1-R_{id}} \right), \qquad i = 1, \dots, n,$$

where for $i = 1, \dots, n$ and $j = 1, \dots, d$, R_{ij} is the rank of X_{ij} among $X_{1j}, \dots, X_{nj}$. We retain only the top 10% of observations with the highest sum of transformed values.

Starting values for the dependence parameters $\boldsymbol{\theta}_Z$ will be assigned randomly, depending on the selected model. As in Section 3, recall that for simplicity, we suppose that the dependence parameters across the columns of A are the same.

Grid points $\boldsymbol{c}_1, \dots, \boldsymbol{c}_q$. Both the parametric and the empirical stdf are evaluated in $\boldsymbol{c}_1, \dots, \boldsymbol{c}_q$ during the minimization (3.1). A sufficiently large number of points q ensures reliable estimation, and does not lead to computational difficulties even with a larger dimension: the evaluation of the empirical stable tail dependence function is fast, as it does not enter the optimization problem. We use the `R` package `tailDepFun` (Kiriliouk, 2016), we select points that can be generated in a 4-dimensional space:

- **for the mixture logistic fit:** using values $\{1/4, 1/3, 1/2, 3/4, 1\}$ where each $\boldsymbol{c}_m$ has either two or three non-zero components. This gives $q = 650$ points;
- **for the mixture Hüsler–Reiss fit:** using values $\{1/6, 1/8, 1/4, 1/3, 2/3, 3/4, 1\}$, where each $\boldsymbol{c}_m$ has two non-zero components. This gives $q = 384$ points.

For the mixture Hüsler–Reiss model, the grid includes more points with only two non-zero components than does the mixture logistic model. This choice is motivated by the fact that, in the former, the dependence coefficients $\Gamma_{i,j}$ of the variogram matrix must be estimated for each pair (i, j), while for the latter, the dependence coefficient α is the same for all pairs.

Performance assessment. We simulate $N = 100$ samples as in (4.2) and for each sample, we obtain the PLS estimator $\hat{\boldsymbol{\theta}}^{(l)} = (\hat{A}^{(l)}, \hat{\boldsymbol{\theta}}_Z^{(l)})$, $l = 1, \dots, N$, using Algorithm 1. We focus on the mixture logistic model, for which $\boldsymbol{\theta}_Z \in (0,1)$, and the mixture Hüsler–Reiss model, for which $\boldsymbol{\theta}_Z \in (0, \infty)^{d(d-1)/2}$.

First, we assess the performance of our estimator in terms of the identification of extreme directions. We use a score, denoted "ED−S", which represents the Jaccard distance and computes the fraction of misidentified extreme directions,

$$\text{ED}-\text{S} = \frac{1}{N} \sum_{l=1}^{N} \text{D}(\hat{\mathcal{J}}_l, \mathcal{J}), \qquad \text{D}(\hat{\mathcal{J}}_l, \mathcal{J}) = 1 - \frac{|\hat{\mathcal{J}}_l \cap \mathcal{J}|}{|\hat{\mathcal{J}}_l \cup \mathcal{J}|} = \frac{|\hat{\mathcal{J}}_l \triangle \mathcal{J}|}{|\hat{\mathcal{J}}_l \cup \mathcal{J}|} \tag{4.3}$$

where $\mathcal{J}$ and $\hat{\mathcal{J}}_l$ denote the signatures of A and $\hat{A}^{(l)}$, respectively, as defined in (2.4), and Δ denotes the symmetric difference between two sets as defined in Section 1. Second, we assess the performance of our estimator by measuring its proximity to the true parameter $\boldsymbol{\theta} = (A, \boldsymbol{\theta}_Z)$ of the mixture model. In the following, for $l = 1, \dots, N$, we reorder the columns of $\hat{A}^{(l)}$ so as to minimize the difference $\sum_{j \in D} \sum_{s \in R} (\hat{a}_{js}^{(l)} - a_{js})^2$.

For simplicity, we continue to denote the resulting matrix by $\hat{A}^{(l)}$. To construct a robust performance metric and ensure that the marginal components of our estimator are equally scaled, we mitigate the disproportionate influence of their magnitudes. This is achieved by normalizing each coefficient of $\hat{\boldsymbol{\theta}}$ using the maximum value computed over the N estimates corresponding to that component. More precisely, let $\hat{m}_A = (\hat{m}_{A,js})_{j \in D,\, s \in R}$ denote the $(d \times r)$ matrix where the (j,s)-th entry represents the maximum value calculated based on the coefficients $\hat{a}_{js}^{(l)}$ obtained from the matrices $\hat{A}^{(l)}$, for $l = 1, \dots, N$. Similarly, recall from Section 3 the dimension v of the dependence parameter vector $\hat{\boldsymbol{\theta}}_Z$. Let $\hat{\boldsymbol{m}}_Z = (\hat{m}_{Z,t})_{t=1,\dots,v}$ denote the v-dimensional vector where the t-th entry represents the maximum value calculated based on the coefficients $\hat{\theta}_t^{(l)}$ obtained from the vectors $\hat{\boldsymbol{\theta}}_Z^{(l)}$, for $l = 1, \dots, N$. We define a standardized version of the mean-squared error as

$$\text{SMSE} = \frac{1}{N} \sum_{l=1}^{N} \left[\text{diff}_{m_A}\left(\hat{A}^{(l)}, A\right) + \text{diff}_{m_Z}\left(\hat{\boldsymbol{\theta}}_Z^{(l)}, \boldsymbol{\theta}_Z\right) \right], \quad \text{where} \tag{4.4}$$

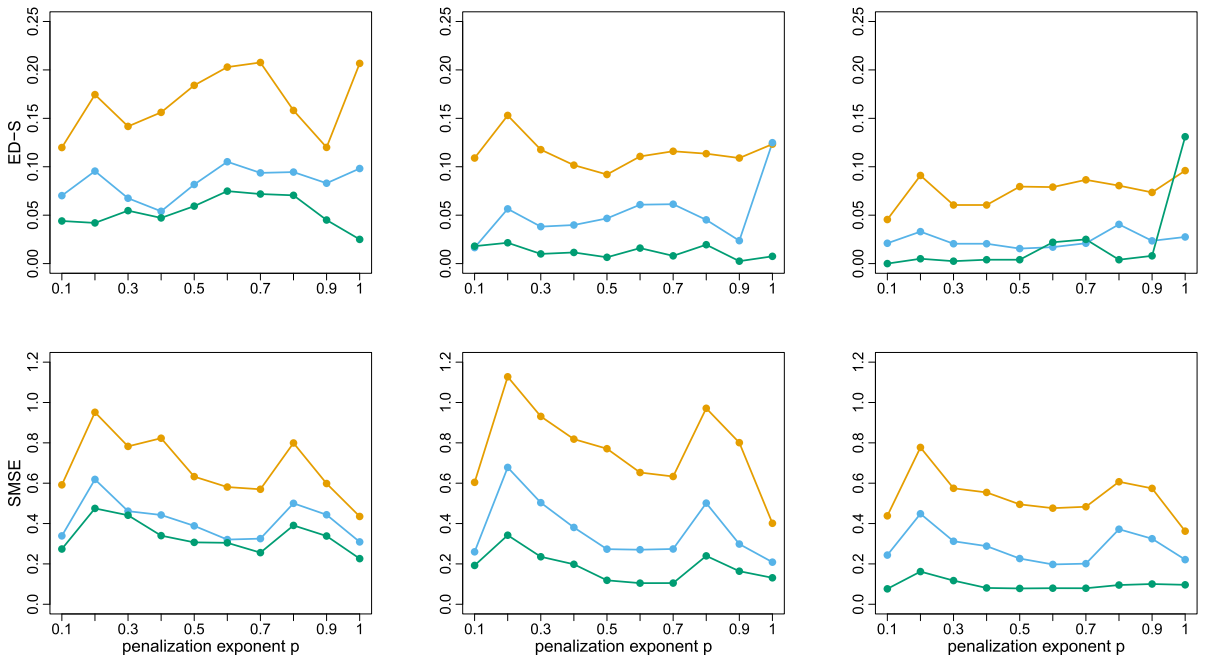

Fig. 3. Scores ED−S (top) and SMSE (bottom) defined in (4.3) and (4.4), respectively, as functions of the penalization exponent p, for the mixture logistic model perturbed as in (4.2) and sample sizes $n = 1\,000$ (left), $n = 2\,000$ (center), and $n = 3\,000$ (right). Curves correspond to varying tail fractions, with approximately $k/n = 0.01$ (orange), $k/n = 0.02$ (blue), and $k/n = 0.05$ (green). (For interpretation of the references to colour in this figure legend, the reader is referred to the web version of this article.)

$$\mathrm{diff}_A\left(\hat{A}^{(l)}, A\right) = \sum_{j\in D}\sum_{s\in R}\frac{\left(\hat{a}_{js}^{(l)} - a_{js}\right)^2}{\hat{m}_{A,js}^2}, \quad \mathrm{diff}_Z\left(\hat{\theta}_Z^{(l)}, \theta_Z\right) = \sum_{t=1}^{v}\frac{\left(\hat{\theta}_{Z,t}^{(l)} - \theta_{Z,t}\right)^2}{\hat{m}_{Z,t}^2}. \tag{4.5}$$

Note that the maximum of a coefficient across N estimates is zero if and only if all estimates of that coefficient are equal to zero. This occurs only for the null coefficients of A. In such cases, by convention, we define $0/0 = 0$.

4.2. Known number of columns

We consider the mixture logistic model in Example 2.5 and the mixture Hüsler–Reiss model in Example 2.6, both with coefficient matrix A in (4.1) and

1. dependence parameter $\alpha := \alpha_1 = \ldots = \alpha_4 = 0.25$ for the mixture logistic model,
2. variogram matrix $\Gamma := \Gamma^{(1)} = \ldots = \Gamma^{(4)}$ with all off-diagonal elements equal to 1 for the mixture Hüsler–Reiss model.

Note that for the mixture Hüsler–Reiss model, we do not suppose that all the off-diagonal elements of the variogram matrix are equal in the estimation procedure, meaning that $\Theta_Z \subseteq (0,\infty)^6$ in (3.1).

In this section, we suppose that the number of extreme directions, $r = 4$, is known. We generate $N = 100$ samples of size $n \in \{1\,000, 2\,000, 3\,000\}$ in the domain of attraction of the mixture logistic model and of the mixture Hüsler–Reiss model as in (4.2), both with a noise variance set to $\sigma^2 = 9$.

Effect of the penalization exponent. We start by studying the effect of the penalization exponent p used in the minimization problem (3.1) on the scores ED−S in (4.3) and SMSE in (4.4). For the tail fraction, we set $k/n = n^c/n$ with $c = \{0.4, 0.5, 0.6\}$. With sample sizes $n \in \{1\,000, 2\,000, 3\,000\}$, this corresponds approximately to $k/n \in \{0.01, 0.02, 0.05\}$. Recall that we choose the penalization parameter λ in (3.1) based on a 10-fold cross validation (Algorithm 2) over a well chosen grid. We use an adaptive penalization parameter grid, meaning that for each value of the penalization exponent p, we have a distinct grid of penalization parameters λ.

Fig. 3 displays the scores ED−S and SMSE as functions of the penalization exponent p. Results are shown for the mixture logistic fit. Qualitatively similar results arise under the mixture Hüsler–Reiss fit. The effect of p is not important as long as the penalization parameter λ in (3.1) is well chosen. In fact, for $0 < p_1 < p_2 < 1$ and a coefficient a of A, the term T in (3.5), which governs the opposite direction step in the minimization algorithm, satisfies $T(p_1)/T(p_2) \to \infty$, whenever the parameter $a \to 0$. It follows that the penalization in (3.1) is stronger with smaller penalization exponent. This explains the use of an adaptive grid of penalization parameters, where the values of the grid associated with p_1 are smaller than the ones used for p_2.

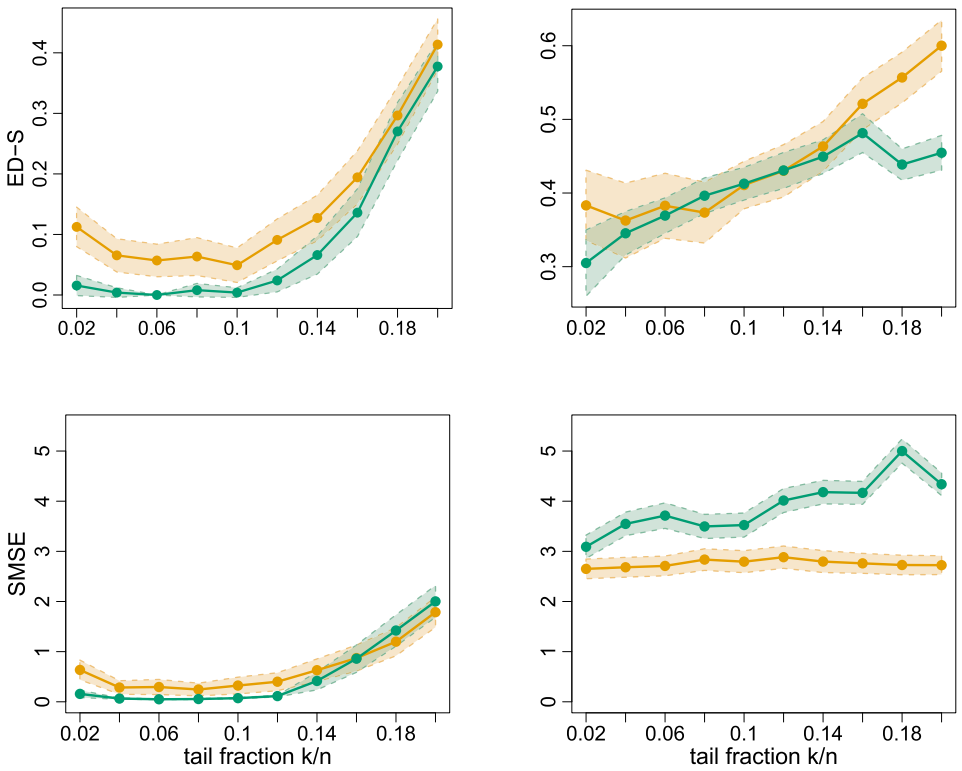

Fig. 4. Scores ED−S (top) and SMSE (bottom) defined in Equations (4.3) and (4.4) from the paper, respectively, as functions of the tail fraction k/n. The shaded regions represent pointwise 95% confidence bands, constructed from the empirical 2.5% and 97.5% quantiles of the two scores. Results are shown for the mixture logistic model (left) and the mixture Hüsler–Reiss model (right), both perturbed by the addition of independent multivariate Gaussian noise with independent margins. The penalization exponent is fixed at $p = .4$. The curves correspond to sample sizes $n = 1\,000$ (orange), and $n = 3\,000$ (green). (For interpretation of the references to colour in this figure legend, the reader is referred to the web version of this article.)

Effect of the tail fraction. We study the effect of the tail fraction k/n used in (3.1) on the scores ED−S in (4.3) and SMSE in (4.4). Fig. 3 shows that the value of the penalization exponent p is not important as long as the penalization parameter λ is well chosen using 10-fold cross validation. Therefore, we set $p = .4$ and perform a 10-fold cross validation to choose λ, where for each value p, we use an adapted grid.

Fig. 4 illustrates the behavior of the scores ED−S and SMSE as functions of the tail fraction k/n for a mixture logistic and a mixture Hüsler–Reiss fit, respectively.

For the mixture logistic model, both scores tend to improve toward the middle of the tail fraction grid. Additionally, the 95% confidence bands for both scores become narrower in this region. This pattern reflects a bias-variance trade-off: a smaller tail fraction k/n results in higher variance but lower bias, as fewer points are treated as extreme observations. In contrast, a larger k/n leads to lower variance but increased bias, since more points-possibly not truly extreme-are included in the estimation. We also observe that as the sample size increases, the optimal choice of the tail fraction k/n-that is, the optimal proportion of points considered as extreme-tends to decrease. This is intuitive: with a larger sample size, a smaller proportion of extreme observations is sufficient to achieve accurate estimation.

The score ED−S for the mixture Hüsler–Reiss fit behaves similarly to that of the mixture logistic fit. In contrast, the score SMSE does not reflect the bias-variance trade-off effect observed for the mixture logistic fit. We can also see that this score is higher for the mixture Hüsler–Reiss model. This makes sense, since the dependence parameter of the mixture Hüsler–Reiss model is of dimension $\upsilon = 6$, while the dependence parameter of the mixture logistic model is of dimension $\upsilon = 1$. This makes the estimation of the dependence parameter for the mixture Hüsler–Reiss model more complicated.

We now discuss the estimation of the variogram matrix Γ of the mixture Hüsler–Reiss model. Fig. 5 shows the boxplots of the upper off-diagonal entries of Γ, except for the entry Γ_{23}; since no signature of the matrix A in (4.1) contains the pair $(2, 3)$, the coefficient Γ_{23} is not identifiable. More precisely, Γ_{23} does not contribute to the stdf in (2.5) associated to the corresponding mixture model.

4.3. Unknown number of columns

We now suppose that the number of extreme directions is unknown and identify the extreme directions using Algorithm 3. Based on Fig. 3, the choice of the penalization parameter p is not important as long as the penalization parameter λ is well chosen using

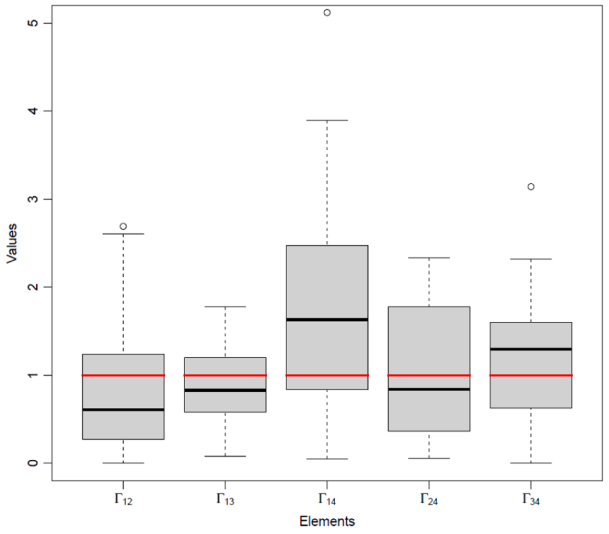

Fig. 5. Boxplots of the estimated coefficients Γ_{ij} from the variogram matrix Γ for all pairs (i, j) included in some signature J of A. The true value 1 for each coefficient is presented with a horizontal red line. The tail fraction k/n is set to 0.06, the penalization exponent to $p = .4$ and the sample size to $n = 1\,000$. (For interpretation of the references to colour in this figure legend, the reader is referred to the web version of this article.)

K-fold cross validation as explained in Algorithm 2. Therefore, we simply choose $p = .4$ and use a 10-fold cross validation procedure to choose λ from the grid associated with $p = .4$. Based on Fig. 4, we choose $k/n \in \{0.04, 0.06, 0.08\}$. We simulate independent samples $\boldsymbol{L}_1, \ldots, \boldsymbol{L}_n \in \mathbb{R}^d$ from the mixture logistic model described in Point 1, and samples $\boldsymbol{Y}_1, \ldots, \boldsymbol{Y}_n \in \mathbb{R}^d$ from the mixture Hüsler–Reiss model introduced in Point 2. As a first step, both models are perturbed by adding independent centered Gaussian noise, as defined in (4.2). In a second, more challenging experiment, we perturb both models by adding positively associated Gaussian noise with Fréchet margins. More precisely,

$$\boldsymbol{X}_i = \boldsymbol{M}_i + \boldsymbol{U}_i, \qquad i = 1, \ldots, n, \tag{4.6}$$

where, for each $i = 1, \ldots, n$, $\boldsymbol{M}_i$ denotes either $\boldsymbol{L}_i$ or $\boldsymbol{Y}_i$, and $\boldsymbol{U}_i \in \mathbb{R}^d$ is a random vector with unit Fréchet margins and a Gaussian copula exhibiting positive dependence, with correlation matrix R such that R_{st} is chosen randomly from $[0, 0.1]$ for two pairs $s \neq t$ from $\{1, \ldots, d\}$. We apply Algorithm 3 with the following two settings:

1. **Correct model specification.** Fit a mixture logistic model to $\boldsymbol{X}_1, \ldots, \boldsymbol{X}_n$ and a mixture Hüsler–Reiss model to $\boldsymbol{Y}_1, \ldots, \boldsymbol{Y}_n$;
2. **Misspecified model.** Fit a mixture logistic model to $\boldsymbol{Y}_1, \ldots, \boldsymbol{Y}_n$ and a mixture Hüsler–Reiss model to $\boldsymbol{X}_1, \ldots, \boldsymbol{X}_n$.

We study the effect of the sample size n on the score ED−S defined in (4.3). Fig. 6 shows the results obtained when applying both Algorithm 3 with a mixture logistic fit and the DAMEX algorithm from Goix et al. (2017, Algorithm 1). We start by presenting the results for Algorithm 3. Similar qualitative behavior is observed under the mixture Hüsler–Reiss fit; therefore, we report only the results corresponding to the mixture logistic fit. For the model in (4.2), both the correctly specified and misspecified model fits exhibit good performance, even for relatively small sample sizes, with satisfactory results from $n = 500$ onward. In contrast, for the model perturbed as in (4.6), performance shows a slight improvement as the sample size increases, even though it remains inferior to that observed for Model (4.2). This behavior is expected. Indeed, for heavy-tailed random variables, the sum of two independent components is asymptotically equivalent to their maximum. Consequently, the random vector $\boldsymbol{X}$ in (4.6) is asymptotically equivalent to a mixture model with extreme directions $\mathcal{J} = \mathcal{J}_A \cup \mathcal{S}_4$, where $\mathcal{J}_A$ are the extreme directions of the matrix A in (4.1) and $\mathcal{S}_4 = \{\{1\}, \{2\}, \{3\}, \{4\}\}$. Compared to Model (4.2), this formulation involves a larger number of parameters to estimate, which explains its comparatively weaker performance. For the DAMEX algorithm, we follow the tuning recommendations in Goix et al. (2017, Remark 7), setting $\epsilon = 0.01$ and $k = n^{1/2}$. For the parameter μ, we consider a grid of values $\{0.01, \ldots, 0.1\}$ in steps of 0.01 and select the value that minimizes the ED−S score. Overall, Algorithm 3 yields better results than DAMEX. More specifically, for Model (4.2), Algorithm 3 starts recovering the exact extreme directions associated to the matrix A from a sample size of $n = 1500$ in both the correct model and the misspecified model. In contrast, when using a sample size of the same order for DAMEX, the results were not satisfactory. Therefore, we consider a larger sample size of order 10^5, as also adopted in the simulation study of Goix et al. (2017, Section 5). We observe that the DAMEX consistently identifies the full set $\{1, 2, 3, 4\}$ as an extreme direction, even though this is not supported by the matrix A in (4.1). Moreover for Model (4.6), the DAMEX does not identify the singletons as extreme directions.

Computational complexity. In practice, changing the sample size n or the tail fraction k/n does not significantly affect the runtime of the algorithm. Typically, it runs in only a few seconds on a standard computer with 6 cores. The first computationally demanding part of Algorithm 3 is the K-fold cross-validation step used to select the optimal penalization parameter λ from a grid of candidate values, denoted Λ. For each $\lambda \in \Lambda$, this requires:

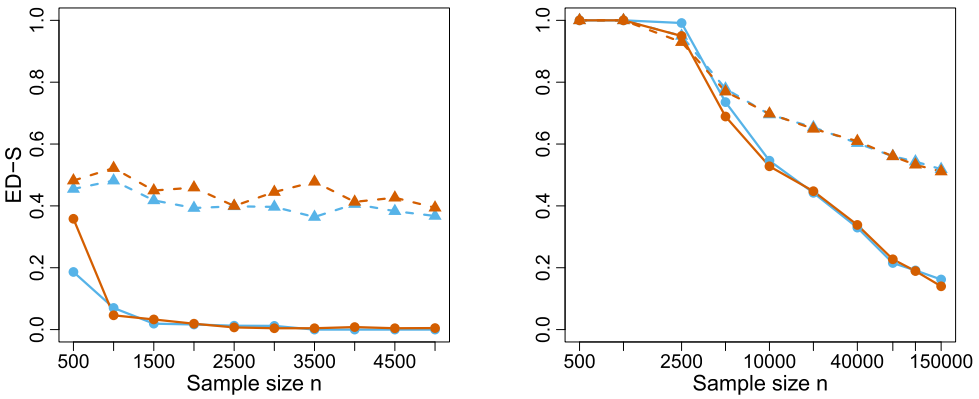

Fig. 6. The score ED−S, defined in (4.3), for simulations from the mixture logistic model (blue) and the mixture Hüsler–Reiss model (dark orange). Solid lines correspond to data perturbed as in (4.2), while dashed lines represent data perturbed as in (4.6). Results on the left are obtained using Algorithm 3, with a mixture logistic fit using $p = .4$ and a tail fraction of $k/n = 0.08$. Results on the right are obtained using the DAMEX algorithm from Goix et al. (2017, Algorithm 1), with tuning parameters $k = n^{1/2}$ and $\epsilon = 0.01$, as recommended in Goix et al. (2017, Remark 7). Note that the sample size used for the DAMEX algorithm reaches the order of 10^5, whereas the sample size used for Algorithm 3 is of order 10^3. (For interpretation of the references to colour in this figure legend, the reader is referred to the web version of this article.)

a) Splitting the data into K folds.
b) For each fold, fit the estimator on the training set and compute the cross-validation error on the held-out fold. This involves minimizing the loss function in Eq. (3.1) using a quasi-Newton method, which requires evaluating the loss multiple times to approximate the gradient and the Hessian.

To reduce the computational burden, we rely on two strategies. First, we use parallel computing to distribute the values of Λ across multiple cores. Second, we implemented the loss function in `C` in order to speed up its evaluation. The second computationally demanding part of Algorithm 3 is the dimension d. In our experiments, using a supercomputer with 15 CPU cores and 15 GB of total RAM, with dimension $d = 4$, sample size $n = 1000$, tail fraction $k/n = 0.08$, and a grid Λ containing 20 values, the full cross-validation procedure takes approximately 3 minutes.

5. Applications

5.1. River discharge data

We illustrate our method on river discharge data from 31 stations along the Danube River. The dataset consists of daily river discharge measurements in the summer months between 1960 and 2010. To mitigate temporal dependence, the data were declustered in Engelke et al. (2024), resulting in approximately seven to ten observations per year and yielding $n = 428$ observations in total. The declustered dataset is available in the `GraphicalExtremes` package in `R`; see (Engelke et al., 2024).

To reduce the dimensionality and lower the computational cost of identifying extreme directions, we apply a spatial clustering technique. Specifically, we partition the stations into K clusters using the Partitioning Around Medoids (PAM) algorithm (Kaufman and Rousseeuw, 2009; Bernard et al., 2013), adapted to threshold exceedances as in Kiriliouk and Naveau (2020). As a pseudo-distance between two stations s and t, we use

$$\widehat{\mathrm{dist}}_{st} = \frac{1 - \hat{\chi}_{st}}{2 \cdot (3 - \hat{\chi}_{st})}, \qquad \hat{\chi}_{st} := 2 - \hat{\ell}_{n,k,st}(1,1), \tag{5.1}$$

where k/n represents the tail fraction and $\hat{\ell}_{n,k,st}$ the pairwise empirical stdf in (2.7) for the margin $\{s,t\}$.

On the one hand, the number $\mathcal{K}$ of clusters balances model simplicity (fewer clusters) and intra-cluster homogeneity. Various methods exist for selecting an appropriate value of $\mathcal{K}$, such as minimizing within-cluster inertia (Kaufman and Rousseeuw, 2009). On the other hand, as explained in Section 4, the choice of the tail fraction k/n is a trade-off between bias and variance. Here, we simply choose $\mathcal{K} = 5$ and $q = 0.05$ which gives spatially coherent results as displayed in Fig. 7. The cluster centers are given then by the five stations $\{7, 18, 24, 27, 30\}$.

Consider the $d = \mathcal{K} = 5$ stations of interest to be the cluster centers. The data $\boldsymbol{X}_i = (X_{i1}, \ldots, X_{id})$ representing the time series of river discharge over the five stations are assumed to be independent and identically distributed for $i = 1, \ldots, n$ and we will identify their associated extreme directions using Algorithm 3.

Choice of tuning parameters. Before estimating the extreme directions associated with the five central stations, we first address the choice of the tuning parameters: the tail fraction k/n, the penalization exponent p, the penalization coefficient λ and the grid points $\boldsymbol{c}_1, \ldots, \boldsymbol{c}_q \in [0, \infty)^5$. As noted in the simulation study, the choice of the penalization exponent p has limited impact provided that

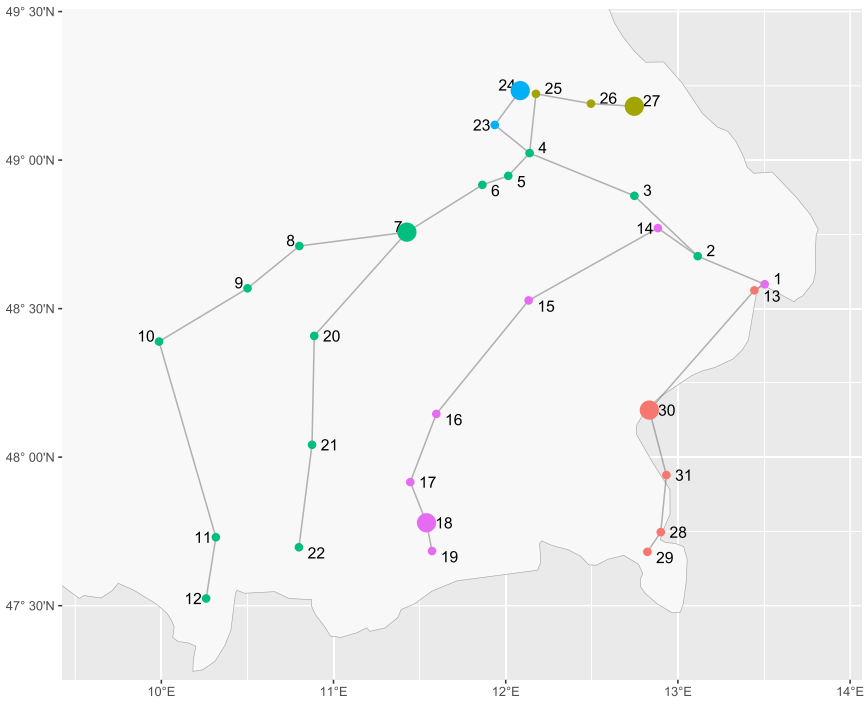

Fig. 7. 31 stations from the Danube dataset clustered into $\mathcal{K} = 5$ groups using an adapted version of the PAM algorithm with the pseudo-distance $\widehat{\text{dist}}$ in (5.1) with tail fraction $k/n = 0.1$. Stations with the same color belong to the same cluster, and within each cluster, the station represented by a larger-sized shape indicates the cluster center.

the penalization parameter λ is appropriately selected. We simply choose $p = .4$, although, in terms of results, other values of the penalization exponent yield similar findings. Guided by the results in Section 4 and given the sample size, $n = 428$, we set the tail fraction to $k/n = 0.1$. The penalization parameter λ is selected via 5-fold cross-validation procedure as described in Algorithm 2 over a grid of values. This grid is constructed in a greedy iterative way such that the optimal value λ^* lies roughly in the middle of the range: we start with a grid $\{0.1, 0.2, \dots, 10\}$, ranging from 0.1 to 10 in steps of 0.1, we then find the optimal value with respect to the 5-fold cross-validation, and construct a new grid centered around this value. We repeat this process until results are stable. This yields a grid of values, $\{0.01, 0.02, \dots, 1\}$, ranging from 0.01 to 1 in steps of 0.01. Note that given the limited sample size, $n = 428$, we perform 5-fold cross-validation, instead of 10-fold cross-validation, as in Section 4, where the sample size was larger. Finally, we select points that can be generated in a 5-dimensional space using values $\{1/4,\ 1/3,\ 1/2,\ 3/4,\ 1\}$ where each point satisfies the condition of having either two or three non-zero components. This yields $q = 1500$ evaluation points.

Results. At each step of Algorithm 3, we fit a mixture logistic model to the five-dimensional river discharge dataset. While in the simulation study, the initial dependence parameter α of the mixture logistic model was randomly initialized within the interval $[0, 1]$, here we set a fixed initial value $\alpha = 0.5$. The results are summarized in Table 1 (left). The algorithm identifies five extreme directions, which appear spatially consistent with the stations indicated on the map in Fig. 7.

As indicated in Table 1 (left), each extreme direction is associated with a weight, see (Mourahib et al., 2024, Proposition 5.1). Specifically, for a mixture logistic model with column coefficients $(\boldsymbol{a}_{\cdot J} : J \in \mathcal{J})$, where $\mathcal{J}$ is the set of extreme directions (equivalently, the signature of the model) and a dependence parameter $\alpha \in [0, 1]$, the scalar

$$w_{\tilde{J}} = \frac{\left(\sum_{j \in \tilde{J}} a_{j\tilde{J}}^{1/\alpha}\right)^{\alpha}}{\sum_{J \in \mathcal{J}} \left(\sum_{j \in J} a_{jJ}^{1/\alpha}\right)^{\alpha}} \tag{5.2}$$

can be interpreted as the weight corresponding to the extreme direction $\tilde{J} \in \mathcal{J}$.

Results make sense given the hydrological nature of the dataset. The standard extreme direction $\{7, 18, 24, 27, 30\}$ can be explained by the fact that extreme river discharges are often caused by large-scale precipitation during the summer months; see (Böhm and Wetzel, 2006). The other non-standard extreme directions can be divided into two clusters. The first cluster corresponds to the extreme directions $\{24\}$, $\{24, 27\}$, and $\{7, 24\}$, with stations 7, 24, and 27 located along Bavarian rivers. Stations 24 and 27 are geographically close and often experience the same weather systems. However, the presence of the extreme direction $\{7, 24\}$ is surprising, since the two stations receive water from different sources; see (Engelke et al., 2025, Fig.9). This might explain why the extreme direction $\{7, 24\}$ is detected with the lowest weight. The second cluster corresponds to the extreme direction $\{7, 18, 30\}$, with stations 7, 18,

and 30 located along Alpine tributaries (southern Germany and western Austria). This can be explained by the fact that these three stations receive water from the same region, flowing from south to north; see (Engelke et al., 2025, Fig.9).

Comparison to DAMEX and goodness of fit. The DAMEX algorithm introduced in Goix et al. (2017, Algorithm 1) identifies only the extreme direction $\{7, 18, 24, 27, 30\}$, even when varying its tuning parameters. This extreme direction was also identified by our algorithm with the largest weight (see Table 1 (left)). Our extreme direction identification method thus appears capable of detecting non-standard extreme-direction structures. However, it is computationally more expensive than the DAMEX algorithm.

To assess the fit of the mixture logistic model to the Danube dataset, we compare the empirical extremal correlation $\hat{\chi}_{st}$ in (5.1) with the estimated extremal correlation for each pair of stations. For a mixture logistic model with estimated matrix $\hat{A} \in [0,1]^{d\times r}$ and dependence coefficient $\hat{\alpha}$, the extremal correlation between two margins s and t is

$$\hat{\chi}_{st,\,\mathrm{est}} = 2 - \sum_{k=1}^{r} \left(\hat{a}_{sk}^{1/\hat{\alpha}} + \hat{a}_{tk}^{1/\hat{\alpha}} \right)^{\hat{\alpha}}. \tag{5.3}$$

Fig. 8 (left) compares the empirical and fitted extremal correlations. Here we use $p = .4$ and $k/n = 0.1$, noting that other choices of tuning parameters yield qualitatively similar results. The agreement is strong, as most points lie close to the line $x = y$. Note that the evaluation process is entirely separate from the model fitting.

5.2. Financial industry portfolios

We consider the value-weighted returns for $d = 10$ industry portfolios, downloaded from the Kenneth French Data Library[1]: "1 = Nondurables", "2 = Durables", "3 = Manufacturing", "4 = Energy", "5 = HiTech", "6 = Telecom", "7 = Shops", "8 = Health", "9 = Utilities", "10 = Other". The individual stocks that make up the five industry portfolios are taken from all listed firms on the NYSE, AMEX, and NASDAQ. Data is available between July 1926 and March 2025, leading to a sample of size $n = 51\,922$.

Data pre-processing. We convert the standard returns to log-returns and multiply by (-1) since we are interested in extreme *losses*. Let $\boldsymbol{X}_t = (X_{t1}, \ldots, X_{td})$ denote the negative log-returns of the five portfolios at time $t = 1, \ldots, n$. In order to obtain a time series without temporal dependence, we fit a GARCH$(1,1)$ model marginally. More precisely, for each portfolio $j = 1, \ldots, 10$,

$$X_{tj} = \mu_{tj} + \sigma_{tj}\varepsilon_{tj}, \qquad t = 1, \ldots, n, \tag{5.4}$$

where σ_{tj}^2 is the conditional variance (volatility), and $\varepsilon_{tj} \overset{\text{iid}}{\sim} \mathcal{N}(0,1)$. For more details about the GARCH model and its application to financial data, see for example (Francq and Zakoian, 2019). We continue with the fitted residuals

$$\hat{\varepsilon}_{tj} = \frac{X_{tj} - \hat{\mu}_{tj}}{\hat{\sigma}_{tj}}, \qquad t = 1, \ldots, n.$$

We apply the Ljung–Box test for serial correlation to each of the $d = 10$ marginal residual series $\hat{\varepsilon}_{1j}, \ldots, \hat{\varepsilon}_{nj}$, using lags 1 through 10. In every case, the test fails to reject the null hypothesis of no autocorrelation, suggesting that the residuals are serially uncorrelated.

Choice of tuning parameters. Similarly to Section 5.1, we set the penalization exponent to $p = .4$. Other values of the penalization exponent yielded similar findings. The sample size ($n = 51\,922$) of the industry portfolios dataset is much larger than the one of the Danube dataset ($n = 428$) in Section 5.1. Therefore, in this section, we set the tail fraction k/n to a smaller value $k/n = 0.05$. The penalization parameter λ is selected via the 10-fold cross-validation procedure described in Algorithm 2 over a grid of values. As explained in Section 5.1, this grid is chosen in a greedy iterative way until results stabilize. This yields a grid of values, $\{0.001, 0.002, \ldots, 0.02\}$, ranging from 0.001 to 0.02 in steps of 0.001. Finally, we select points that can be generated in a 10-dimensional space using values $\{0, 1/3, 1/2, 1\}$ where each point has two non-zero components which yields $q = 405$ evaluation points.

Results. As in Section 5.1, at each step of Algorithm 3 we fit a mixture logistic model (Example 2.5) to the five industry portfolio and set the initial value for the dependence parameter to $\alpha = 0.5$. The results are summarized in Table 1 (right). Unlike the DAMEX algorithm, our approach identifies multiple extreme directions, including non-standard ones. Among these, the standard extreme direction $\{1, \ldots, 10\}$ receives the highest weight, corresponding to the unique extreme direction detected by DAMEX, even when varying its tuning parameters. Again, this emphasizes that our extreme direction identification method is capable of detecting sparser extreme-direction structures. The standard extreme direction $\{1, \ldots, 10\}$ corresponds to the one assigned the highest weight, which is intuitive given that the portfolios are composed of stocks from U.S.-based exchanges (NYSE, AMEX, and NASDAQ). As such, losses are not confined to a single sector but rather spread across the entire U.S. economy. Some of the other extreme directions can also be meaningfully interpreted. For instance, "1 = NonDurables", includes stocks from sectors such as Food and Tobacco, while "10 = Other", comprises sectors like Transportation and Entertainment; these two belong to the same extreme direction, suggesting similar tail-risk behavior. Likewise, "6 = Telecom" (including Telephone and Television Transmission), and "7 = Shops" (including Repair Shops), also fall within the same extreme direction. However, some extreme directions are more challenging to interpret without

[1] See https://mba.tuck.dartmouth.edu/pages/faculty/ken.french/Data_Library/det_10_ind_port.html for more details about the data and the definition of each portfolio.

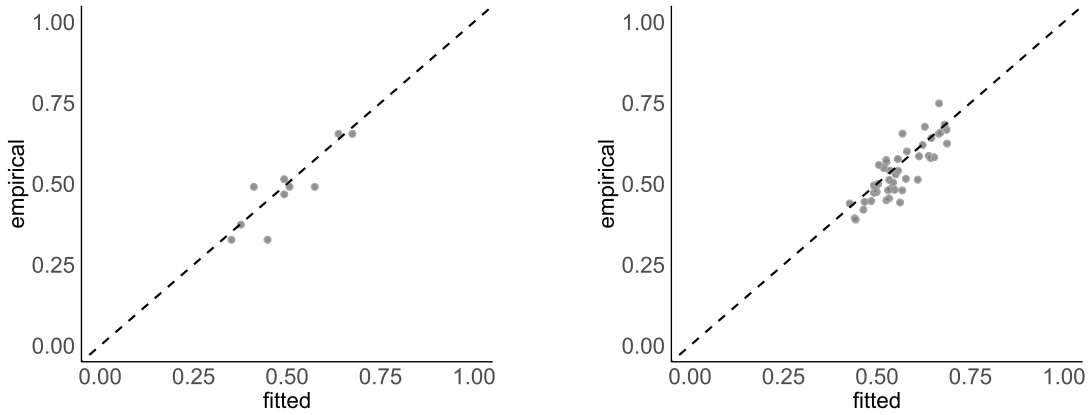

Fig. 8. Empirical (5.1) versus fitted (5.3) extremal correlations for each pair of margins. Left: results from the Danube dataset (Section 5.1) with $k/n = q = 0.1$. Right: industry portfolios (Section 5.2) with $k/n = q = 0.05$. In both applications, the penalization exponent is fixed at $p = .4$.

Table 1
Extreme directions identified by Algorithm 3 for the Danube river discharge data at five stations, discussed in Section 5.1 (left), and the ten industry portfolios: "1 = Nondurables", "2 = Durables", "3 = Manufacturing", "4 = Energy", "5 = HiTech", "6 = Telecom", "7 = Shops", "8 = Health", "9 = Utilities", "10 = Other", discussed in Section 5.2 (right), along with their associated weights as in (5.2).

Extreme directions	Weights
$\{7, 18, 24, 27, 30\}$	0.49
$\{7, 18, 30\}$	0.17
$\{24\}$	0.15
$\{24, 27\}$	0.14
$\{7, 24\}$	0.06

Extreme directions	Weights
$\{1, \ldots, 10\}$	0.52
$\{7, 8\}$	0.10
$\{6\}$	0.09
$\{1, 9, 10\}$	0.09
$\{4\}$	0.08
$\{1, 2, 3, 6, 7\}$	0.07
$\{2\}$	0.02

deeper economic insight. This includes singleton directions such as "2 = Durables", 4 = Energy", "6 = Telecom", which may reflect more nuanced or isolated sector-specific behaviors. Finally, the standard extreme direction $\{1, \ldots, 10\}$ prevents the existence of a negative dependence structure between every pair of distinct sectors.

To assess the goodness-of-fit of the mixture logistic model to the industry portfolios dataset, we compute for each pair (s, t) both the empirical extremal correlation in (5.1) and the fitted extremal correlation in (5.3). Fig. 8 (right) compares the empirical and fitted extremal correlations, computed both with $p = .4$ and $k/n = 0.05$. Other choices of tuning parameters yielded qualitatively similar results. The agreement is strong, as most points lie close to the line $x = y$.

6. Conclusion

Given a vector of d risk factors, extreme directions refer to groups of factors $J \subseteq \{1, \ldots, d\}$ that are large simultaneously while those outside this group are not. Mourahib et al. (2024) propose a construction method of multivariate generalized Pareto distributions, or equivalently max-stable distributions, with multiple extreme directions. Extreme directions are a consequence of parameters of the mixture model being on the border of the parameter space (Mourahib et al., 2024, Proposition 4.5). Parameter estimation is therefore a challenging topic. We proposed a penalized least-squares estimator (3.1) based on the estimator from Einmahl et al. (2018). We prove its consistency and propose a data-driven algorithm for extreme-directions identification. Using a simulation study, we show that although the algorithm relies on a parametric assumption for the mixture model (e.g., mixture Hüsler–Reiss graphical model or mixture logistic model), the specific choice of the model is not crucial. In addition, the algorithm involves three tuning parameters;

again through simulations, we demonstrate that the tuning essentially reduces to selecting two parameters: one that controls the proportion of data considered as extreme observations and the other that controls the penalization strength, chosen using K-fold cross-validation. We illustrated our algorithm on two real datasets, determining the sets of variables that form extreme directions. We compare our approach with the DAMEX algorithm in Goix et al. (2017) and show that our algorithm appears capable of detecting non-standard extreme directions.

Acknowledgment

The research of Anas Mourahib was supported financially by the *Fonds de la Recherche Scientifique - FNRS*, Belgium (grant number T.0203.21). Computational resources have been provided by the supercomputing facilities of the Université catholique de Louvain (CISM/UCL) and the Consortium des Équipements de Calcul Intensif en Fédération Wallonie Bruxelles (CÉCI) funded by the Fond de la Recherche Scientifique de Belgique (F.R.S.-FNRS) under convention 2.5020.11 and by the Walloon Region.

The authors thank Sebastian Engelke for insightful discussions on the interpretation of the river discharge data analysis in Section 5.1, Michel Denuit for his input on the interpretation of the portfolio data in Section 5.2, and Anne Sabourin for generously sharing her code and providing a clear explanation of the DAMEX algorithm.

Appendix A. Proof of Theorem 3.4

Proof of Theorem 3.4. By the consistency of the empirical stable tail dependence function (Drees and Huang, 1998) as $k \to \infty$ and $k/n \to \infty$, we have $\hat{\ell}_{n,k}(\boldsymbol{c}_m) \to \ell(\boldsymbol{c}_m, \boldsymbol{\theta}_0)$ in probability for all $m \in \{1, \dots, q\}$ as $n \to \infty$. In addition, $0 \leqslant \ell(\boldsymbol{c}_m; \boldsymbol{\theta}) \leqslant \sum_{j=1}^{d} c_{mj}$ for all $m \in \{1, \dots, q\}$ and $\boldsymbol{\theta} \in \Theta$, and $\mathcal{P}(A) \leqslant d \cdot r$ for all $A \in [0,1]^{d \times r}$. Since $\lambda \to 0$, it follows that

$$\Delta_{n,k} := \sup_{\boldsymbol{\theta} \in \Theta} \left| \mathrm{Loss}_{n,k}(\boldsymbol{\theta}) - \mathrm{Loss}(\boldsymbol{\theta}) \right| \to 0 \quad \text{in probability}$$

as $n \to \infty$, where

$$\mathrm{Loss}_{n,k}(\boldsymbol{\theta}) = \sum_{m=1}^{q} \left(\hat{\ell}_{n,k}(\boldsymbol{c}_m) - \ell(\boldsymbol{c}_m; \boldsymbol{\theta}) \right)^2 + \lambda \mathcal{P}(A),$$

$$\mathrm{Loss}(\boldsymbol{\theta}) = \sum_{m=1}^{q} \left(\ell(\boldsymbol{c}_m; \boldsymbol{\theta}_0) - \ell(\boldsymbol{c}_m; \boldsymbol{\theta}) \right)^2.$$

Clearly, $\mathrm{Loss}(\boldsymbol{\theta}_0) = 0$, while, by the injectivity assumption, $\mathrm{Loss}(\boldsymbol{\theta}) > 0$ as soon as $\boldsymbol{\theta} \neq \boldsymbol{\theta}_0$. Let $\epsilon > 0$ and write $K_\epsilon = \left\{ \boldsymbol{\theta} \in K : \|\boldsymbol{\theta} - \boldsymbol{\theta}_0\| \geqslant \epsilon \right\}$. By compactness of K (and thus of K_ϵ), we find

$$\inf_{\boldsymbol{\theta} \in K_\epsilon} \mathrm{Loss}(\boldsymbol{\theta}) =: \delta(\epsilon) > 0.$$

On the event $\{\Delta_{n,k} < \delta(\epsilon)/2\}$, we have

$$\forall \boldsymbol{\theta} \in K_\epsilon : \qquad \mathrm{Loss}_{n,k}(\boldsymbol{\theta}) > \mathrm{Loss}(\boldsymbol{\theta}) - \delta(\epsilon)/2 \geqslant \delta(\epsilon)/2 > 0 = \mathrm{Loss}(\boldsymbol{\theta}_0).$$

It follows that, on the event $\{\Delta_{n,k} < \delta(\epsilon)/2\}$ we must have $\|\hat{\boldsymbol{\theta}}_K - \boldsymbol{\theta}_0\| < \epsilon$. Since the probability of this event tends to zero and since $\epsilon > 0$ is arbitrary, the claim follows. □

Appendix B. Mixture model simulation algorithm

Algorithm 4 Simulation of a mixture model.

Input: integers $d, r \geqslant 1$, matrix $A = \left(a_{jk}\right)_{j \leqslant d, k \leqslant r} \in [0,1]^{d \times r}$ and max-stable random vectors $\boldsymbol{Z}_{\cdot 1}, \dots, \boldsymbol{Z}_{\cdot r} \in \mathbb{R}^d$ as in Definition 2.2.
▷ For simplicity, we consider the vector $\boldsymbol{Z}_{\cdot k} = (Z_{1k}, \dots, Z_{dk}) \in \mathbb{R}^d$ for each $k = 1, \dots, r$. However, as discussed after Definition 2.2, only the sub-vector $\boldsymbol{Z}_{J_k, k}$ is relevant.

Output: mixture logistic model $\boldsymbol{M} \in \mathbb{R}^d$ in (2.3) with coefficient matrix A and factors $\boldsymbol{Z}_{\cdot 1}, \dots, \boldsymbol{Z}_{\cdot r}$.

Initialize a vector $\boldsymbol{M} = (0, \dots, 0) \in \mathbb{R}^d$
Initialize a list $\tilde{Z} = \left\{ \tilde{\boldsymbol{Z}}_1, \dots, \tilde{\boldsymbol{Z}}_r \right\}$, where each $\tilde{\boldsymbol{Z}}_k \in \mathbb{R}^d$ is initialized as a zero vector.
for each $k \in \{1, \dots, r\}$ **do**
    Define $\boldsymbol{J}_k$ as in (2.4), the k-th signature of the coefficient matrix A.
    Set $\tilde{Z}_{k,j}$, the j-th component of $\tilde{\boldsymbol{Z}}_k$, to $a_{jk} Z_{jk}$, for every $j \in J_k$.
end for
Compute $M_j = \max \left\{ \tilde{Z}_{j1}, \dots, \tilde{Z}_{jr} \right\}$, for each $j \in \{1, \dots, d\}$.
return $\boldsymbol{M}$

References

Beirlant, J., Escobar-Bach, M., Goegebeur, Y., Guillou, A., 2016. Bias-corrected estimation of stable tail dependence function. J. Multivar. Anal. 143, 453–466.
Beirlant, J., Goegebeur, Y., Segers, J., Teugels, J.L., 2004. Statistics of extremes: theory and applications. John Wiley & Sons.
Belzile, L., Wadsworth, J.L., Northrop, P.J., Grimshaw, S.D., Zhang, J., Stephens, M.A., Owen, A.B., 2024. Package 'mev'.
Belzile, L.R., Nešlehová, J.G., 2017. Extremal attractors of Liouville copulas. J. Multivar. Anal. 160, 68–92.
Bernard, E., Naveau, P., Vrac, M., Mestre, O., 2013. Clustering of maxima: spatial dependencies among heavy rainfall in France. J. Clim. 26 (20), 7929–7937.
Böhm, O., Wetzel, K.-F., 2006. Flood history of the Danube tributaries lech and isar in the alpine foreland of Germany. Hydrol. Sci. J. 51 (5), 784–798.
Byrd, R.H., Lu, P., Nocedal, J., Zhu, C., 1995. A limited memory algorithm for bound constrained optimization. SIAM J. Sci. Comput. 16 (5), 1190–1208.
Coles, S., 2001. An introduction to statistical modeling of extreme values. Springer.
De Haan, L., Ferreira, A., 2006. Extreme value theory: an introduction. Springer.
Dombry, C., Engelke, S., Oesting, M., 2016. Exact simulation of max-stable processes. Biometrika 103 (2), 303–317.
Drees, H., Huang, X., 1998. Best attainable rates of convergence for estimators of the stable tail dependence function. J. Multivar. Anal. 64 (1), 25–46.
Einmahl, J. H.J., Kiriliouk, A., Segers, J., 2018. A continuous updating weighted least squares estimator of tail dependence in high dimensions. Extremes 21, 205–233.
Einmahl, J. H.J., Krajina, A., Segers, J., 2012. An M-estimator for tail dependence in arbitrary dimensions. Ann. Stat. 40 (3), 1764–1793.
Engelke, S., Gnecco, N., Röttger, F., 2025. Extremes of structural causal models. Available at https://arxiv.org/abs/2503.06536.
Engelke, S., Hitz, A.S., Gnecco, N., Hentschel, M., 2024. Package 'graphicalExtremes'.
Fougères, A.-L., Mercadier, C., Nolan, J.P., 2013. Dense classes of multivariate extreme value distributions. J. Multivar. Anal. 116, 109–129.
Francq, C., Zakoian, J.-M., 2019. GARCH models: structure, statistical inference and financial applications. John Wiley & Sons.
Goix, N., Sabourin, A., Clémençon, S., 2017. Sparse representation of multivariate extremes with applications to anomaly detection. J. Multivar. Anal. 161, 12–31.
Heffernan, J.E., Tawn, J.A., 2004. A conditional approach for multivariate extreme values (with discussion). J. R. Stat. Soc. B Stat. Methodol. 66 (3), 497–546.
Huser, R., Davison, A.C., 2013. Composite likelihood estimation for the Brown–Resnick process. Biometrika 100 (2), 511–518.
Hüsler, J., Reiss, R.-D., 1989. Maxima of normal random vectors: between independence and complete dependence. Stat. Probab. Lett. 7 (4), 283–286.
Kaufman, L., Rousseeuw, P.J., 2009. Finding groups in data: an introduction to cluster analysis. John Wiley & Sons.
Kiriliouk, A., 2016. Vignette for the tailDepFun package.
Kiriliouk, A., 2020. Hypothesis testing for tail dependence parameters on the boundary of the parameter space. Econom. Stat. 16, 121–135.
Kiriliouk, A., Naveau, P., 2020. Climate extreme event attribution using multivariate peaks-over-thresholds modeling and counterfactual theory. Ann. Appl. Stat.
Kiriliouk, A., Segers, J., Tafakori, L., 2018. An estimator of the stable tail dependence function based on the empirical beta copula. Extremes 21, 581–600.
Lenzi, A., Bessac, J., Rudi, J., Stein, M.L., 2023. Neural networks for parameter estimation in intractable models. Comput. Stat. Data Anal. 185, 107762.
Meyer, N., Wintenberger, O., 2024. Multivariate sparse clustering for extremes. J. Am. Stat. Assoc. 119 (547), 1911–1922.
Mourahib, A., Kiriliouk, A., Segers, J., 2024. Multivariate generalized Pareto distributions along extreme directions. Extremes, 1–34.
Padoan, S.A., Ribatet, M., Sisson, S.A., 2010. Likelihood-based inference for max-stable processes. J. Am. Stat. Assoc. 105 (489), 263–277.
Richards, J., Sainsbury-Dale, M., Zammit-Mangion, A., Huser, R., 2024. Neural bayes estimators for censored inference with peaks-over-threshold models. J. Mach. Learn. Res. 25 (390), 1–49.
Rootzén, H., Tajvidi, N., 2006. Multivariate generalized Pareto distributions. Bernoulli 12 (5), 917–930.
Schlather, M., Tawn, J., 2002. Inequalities for the extremal coefficients of multivariate extreme value distributions. Extremes 5, 87–102.
Simpson, E.S., Wadsworth, J.L., Tawn, J.A., 2020. Determining the dependence structure of multivariate extremes. Biometrika 107 (3), 513–532.
Stephenson, A., 2003. Simulating multivariate extreme value distributions of logistic type. Extremes 6 (1), 49–59.
Tawn, J.A., 1988. Bivariate extreme value theory: models and estimation. Biometrika 75 (3), 397–415.
Tawn, J.A., 1990. Modelling multivariate extreme value distributions. Biometrika 77 (2), 245–253.
Wadsworth, J.L., Campbell, R., 2024. Statistical inference for multivariate extremes via a geometric approach. J. R. Stat. Soc. B Stat. Methodol. 86 (5), 1243–1265.
Wang, Y., Stoev, S.A., 2011. Conditional sampling for spectrally discrete max-stable random fields. Adv. Appl. Probab. 43 (2), 461–483.