Journal of Multivariate Analysis 202 (2024) 105300

Contents lists available at ScienceDirect

Journal of Multivariate Analysis

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

Latent model extreme value index estimation

Joni Virta [a,b,*], Niko Lietzén [a,b], Lauri Viitasaari [c], Pauliina Ilmonen [b]

[a] University of Turku, Finland
[b] Aalto University School of Science, Finland
[c] Aalto University School of Business, Finland

ARTICLE INFO

AMS 2020 subject classifications:
primary 62H12
secondary 62H25
62G32

Keywords:
Blind source separation
Hill estimator
Independent component analysis
Moment estimator
Tail index

ABSTRACT

We propose a novel strategy for multivariate extreme value index estimation. In applications such as finance, volatility and risk of multivariate time series are often driven by the same underlying factors. To estimate the latent risks, we apply a two-stage procedure. First, a set of independent latent series is estimated using a method of latent variable analysis. Then, univariate risk measures are estimated individually for the latent series. We provide conditions under which the effect of the latent model estimation to the asymptotic behavior of the risk estimators is negligible. Simulations illustrate the theory under both i.i.d. and dependent data, and an application into currency exchange rate data shows that the method is able to discover extreme behavior not found by component-wise analysis of the original series.

1. Introduction

Let $\mathbf{x}_1, \dots, \mathbf{x}_n$ be a sample of p-variate random vectors with possibly dependent distributions. For each observation, we assume the instantaneous latent variable model,

$$\mathbf{x}_i = f(\mathbf{z}_i), \quad i \in \{1, \dots, n\}, \tag{1}$$

where the latent p-variate random vectors $\mathbf{z}_1, \dots, \mathbf{z}_n$ are assumed to have independent components in the sense that if the vectors are gathered into a matrix $\mathbf{Z} = (\mathbf{z}_1, \dots, \mathbf{z}_n)$, then the rows of $\mathbf{Z}$ are mutually independent but arbitrary dependencies are allowed within the elements of individual rows. Furthermore, we assume $f : \mathbb{R}^p \to \mathbb{R}^p$ is a deterministic invertible linear mapping. Note that while no explicit noise term is present in (1), the general formulation still captures noisy latent models as well, as one or several of the p latent components can represent noise which is then combined with the other components (signals) by the function f.

The model (1) is used in independent component analysis and has applications in numerous fields such as in telecommunications, psychometrics, economics and finance [9,28]. The model provides a powerful alternative to standard multivariate modeling schemes as, after having estimated the latent vectors, the independence of their components implies that all subsequent modeling can be done univariately. This structural simplification leads to both smaller number of parameters to estimate and simplified interpretations for the components as no interactions between the series need to be acknowledged.

In this paper we focus on estimating the tail behavior of the latent variables in the model (1), evaluated through the extreme value indices of the corresponding distributions [10]. This is a natural goal to pursue in many financial and signal processing applications as the heaviness of the tails of a distribution is an indicator of an unstable and risky signal. For example, the independent component model has been applied in cashflow analysis and prediction of financial time series data [32,34,57], and in this context assessing the tail behavior of the obtained independent components could help identify common sources of financial risk. Similarly, the evaluation

* Corresponding author at: University of Turku, Finland.
E-mail address: joni.virta@utu.fi (J. Virta).

https://doi.org/10.1016/j.jmva.2024.105300
Received 7 August 2023; Received in revised form 13 February 2024; Accepted 13 February 2024
Available online 15 February 2024

of the extreme behavior of latent components could help identify the sources of abnormalities in applications such as biomedical imaging [49] or maritime vessel track analysis [52].

This objective can be reached in two steps. First, we estimate a mapping $\hat{f}^{-1}$ such that $\hat{f}^{-1}(\mathbf{x})$ equals the latent components up to order and scales (in latent component analysis, the order and scales of the latent components are usually neither of interest nor identifiable, see, e.g., [54]). Numerous techniques for obtaining consistent estimators under various types of data exist, see Section 4 for examples. Second, after having obtained the sample estimates $\hat{f}^{-1}(\mathbf{x}_1), \dots, \hat{f}^{-1}(\mathbf{x}_n)$ of the latent vectors, we use one of the several univariate extreme value index estimators presented in the literature [10] to assess the extreme behavior of the individual, now independent, components.

Note that, in contrast to the above, the standard approach in multivariate extreme value theory is to assess the extreme behavior component-wise for the observed multivariate signal itself [10]. Approaches which in some way acknowledge the multivariate structure of the data are much rarer. Some examples include: considering convex combinations of the component-wise estimators [14,31], extreme risk region estimation [3], estimation of the extreme value index of the generating variate of an underlying elliptical model [15,21], deriving limit theory for the estimation of component-wise extreme value indices under multivariate dependent data [23] and M-estimation under the assumption of a parametric form for the dependence structure [19]. However, most of these methods either involve complicated estimation or require strict distributional assumptions, making them less than ideal in practice. In comparison, our proposed two-step procedure is straightforward to apply and takes the multivariate form of the data into account in a natural way. The associated latent variable model is also very flexible, allowing different tail behaviors for the underlying independent components. Moreover, as illustrated by the currency exchange rate data example in Section 6, the method is able to find sources of risk that component-wise univariate extreme value index estimation completely misses.

1.1. Scope and structure of the paper

Of the two steps of our proposed method, we are primarily interested in the latter. That is, we focus on assessing the extreme behavior of the individual components in the independent component model (1). Throughout the article, we assume that there exists an estimator $\hat{f}^{-1}$ with the asymptotic linearization

$$\hat{\mathbf{z}}_i := \hat{f}^{-1}(\mathbf{x}_i) = \mathbf{z}_i + \hat{H}\mathbf{z}_i + \hat{\mathbf{r}}, \tag{2}$$

where the $p \times p$ -matrix $\hat{H} := \hat{H}_n = \mathcal{O}_p(c_n^{-1})$ and the p-vector $\hat{\mathbf{r}} := \hat{\mathbf{r}}_{i,n} = \mathcal{O}_p(c_n^{-1})$ (uniformly in i) for some rate $c_n \to \infty$ as $n \to \infty$. Here $\mathcal{O}_p(c_n^{-1})$ denotes the element-wise "convergence rate" in probability. For a precise definition, see Section 3. The form (2) is general and encompasses many popular estimators $\hat{f}^{-1}$ in the independent component analysis and blind source separation literature, see Section 4 for examples. Moreover, all our subsequent developments are based on the form (2), meaning, in particular, that the map f in the model (1) does not necessarily have to be linear, as long as the estimator $\hat{f}^{-1}$ satisfies (2). However, in this work we essentially restrict to linear maps as examples of non-linear estimators are still scarce in the independent component literature.

Assuming for now that an estimator $\hat{f}^{-1}$ exists in the sense of (2), our main objective is to estimate the extreme value indices of the components of the latent variables using $\hat{\mathbf{z}}_i$ as a proxy for $\mathbf{z}_i$, and to show that this approximation incurs no loss in asymptotic efficiency under a suitable set of assumptions. The strictest of these assumptions is related to the rate c_n, which turns out to directly control how heavy-tailed latent distributions we can consider. As an example (see Section 2 for more details), in the case where $c_n = n^\beta$ for some $\beta > 0$, the extreme value index of the heaviest latent component needs to be smaller than β (that is, it must have finite moments of order $1/\beta$) for the asymptotic behavior to be retained. In particular, in the standard case with $c_n = \sqrt{n}$ the existence of second moments is required.

A further complicating factor is that latent variable models such as (1) are well-known for not having fixed signs or scales for the latent components. That is, the vector $\mathbf{z}_i$ on the right-hand side of (2) actually corresponds in many models to the true latent vectors only up to the signs and scales of its components. In the standard usage of latent variable modeling this is most often acceptable, as our interest lies commonly not in the signs, but in the shapes of the distributions of the latent variables. Similarly, in the present context, the scale of the components is irrelevant as most commonly applied extreme value index estimators are scale-invariant. However, as risk is estimated from the tails of the components, knowing in which of the tails we are in is for our purposes of paramount importance, and we need a way of identifying the correct tail. A simple, but restrictive, solution would be to require that all the latent components have symmetric distributions. Instead, we choose to assess the extreme behavior of, not the latent components, but their absolute values. This rids us of the sign indeterminacy by "stacking" the two tails on top of each other. Since the absolute value inherits its tail behavior from the heavier of the two tails, this approach has the interpretation of us always looking at the heavier of the two tails. Moreover, as heavier tails correspond to larger risk, the use of absolute values can be seen as a conservative approach to tail behavior estimation.

The rest of the paper is organized as follows: Preliminaries on extreme value theory along with the popular extreme value index estimators, the Hill estimator and the moment estimator, are reviewed in Section 2. These extreme value index estimators are known to be consistent and asymptotically normal under mild technical conditions. In Section 3, we derive sufficient conditions ensuring that the asymptotic properties of the extreme value estimators are preserved when estimated using the proxy sample $\hat{f}^{-1}(\mathbf{x}_i)$. In Section 4, we consider two example cases of the general framework and discuss the particular assumptions needed to achieve the limiting results for the corresponding proxy samples. In Section 5, we present a large simulation study and a real data application into currency exchange rate data is considered in Section 6. Appendix A is devoted to auxiliary technical lemmas and the proofs of our main theorems. The supplementary Appendix B contains an auxiliary simulation (the same simulation study as conducted in the main text, but with i.i.d. observations instead of time dependent), and the supplementary Appendix C presents additional figures for the real data example.

2. Preliminaries on extreme value theory

In the following we provide a brief introduction to the topics in univariate extreme value theory that are most relevant to our objectives. See [10] and the references therein for more information.

Consider an i.i.d. random sample $\mathbf{y} = (y_1, \dots, y_n)$ from a univariate distribution F and the sample maximum $M_n = \max_{1 \le i \le n} y_i$. If there exists sequences of constants $a_n > 0$ and b_n such that $a_n M_n + b_n$ has a non-degenerate limiting distribution G, we say that G is an extreme value distribution of F. One of the fundamental results in extreme value theory is the Fisher–Tippett–Gnedenko theorem which identifies the class of distributions G.

Theorem 1 (*Fisher–Tippett–Gnedenko*). *The class of extreme value distributions is* $G_\gamma(ax + b)$ *with* $a > 0$ *and* $b \in \mathbb{R}$, *where*

$$G_\gamma(x) = \exp\left\{-(1+\gamma x)^{-1/\gamma}\right\}, \quad 1 + \gamma x > 0,$$

with $\gamma \in \mathbb{R}$ *and where for* $\gamma = 0$ *the right-hand side is interpreted as* $\exp(-e^{-x})$.

According to Theorem 1, the family of possible extreme value distributions has a remarkably simple form, parametrized by a single real number γ. If G_γ is the extreme value distribution of F, the distribution F is said to be in the domain of attraction of G_γ, and we write $F \in D(G_\gamma)$. The parameter γ is said to be the extreme value index of F. The parameter γ measures the thickness of the (right) tail of F and knowing its value leads to a complete characterization of the asymptotic tail behavior of F (up to linear transformations), allowing extrapolating probabilities beyond the observed dataset. Thus γ is a key ingredient in risk assessment.

It is widely accepted that distributions are divided into heavy and short tailed ones based on the sign of γ. More precisely, for $\gamma > 0$, the distributions $F \in D(G_\gamma)$ are called heavy tailed and belonging to the domain of attraction of the Fréchet distribution. Similarly, if $\gamma < 0$ and $F \in D(G_\gamma)$, then we say that F is short tailed and belongs to the domain of attraction of the Weibull distribution. Finally, if $F \in D(G_0)$, then F belongs to the domain of attraction of the Gumbel distribution. This corresponds to the border case between short and heavy tails, and includes, e.g., the case of a normal distribution.

One of the most commonly applied classical estimators of the extreme value index, suitable for $\gamma > 0$, is the Hill estimator introduced in [22],

$$\hat{\gamma}_H(\mathbf{y}) = \frac{1}{k_n} \sum_{m=0}^{k_n - 1} \ln \frac{(\mathbf{y})_{(n-m,n)}}{(\mathbf{y})_{(n-k_n,n)}},$$

where $(\mathbf{y})_{(n,n)} \ge \cdots \ge (\mathbf{y})_{(1,n)}$ are the order statistics of the sample $\mathbf{y}$, and $1 \le k_n \le n$ is a sequence of thresholds for the portion of observations that are considered to form the tail. Common choices for the threshold include, e.g., $k_n = \sqrt{n}$ and $k_n = \ln(n)$.

Another well-known estimator, which in turn is valid for any value of γ, is the moment estimator introduced in [13]. For that, we define the auxiliary quantities,

$$M_n^{(j)}(\mathbf{y}) = \frac{1}{k_n} \sum_{m=0}^{k_n - 1} \left\{ \ln \frac{(\mathbf{y})_{(n-m,n)}}{(\mathbf{y})_{(n-k_n,n)}} \right\}^j,$$

for $j \in \{1, 2, \dots\}$ The Hill estimator now corresponds directly to the choice $j = 1$, that is, $\hat{\gamma}_H(\mathbf{y}) = M_n^{(1)}(\mathbf{y})$, whereas the moment estimator $\hat{\gamma}_M(\mathbf{y})$ is built from both $M_n^{(1)}(\mathbf{y})$ and $M_n^{(2)}(\mathbf{y})$ as,

$$\hat{\gamma}_M(\mathbf{y}) = M_n^{(1)}(\mathbf{y}) + 1 - \frac{1}{2}\left[1 - \frac{\{M_n^{(1)}(\mathbf{y})\}^2}{M_n^{(2)}(\mathbf{y})}\right]^{-1}.$$

In the next section, both the Hill estimator and the moment estimator are used to estimate the extreme value indices of the absolute values of the latent components in (1).

3. Extreme value index estimation for latent variables

Recall from Section 1 that we consider an estimated sample $\hat{\mathbf{z}}_1, \dots, \hat{\mathbf{z}}_n$ of the not necessarily i.i.d. latent vectors $\mathbf{z}_1, \dots, \mathbf{z}_n$ satisfying

$$\hat{\mathbf{z}}_i = \mathbf{z}_i + \hat{H}\mathbf{z}_i + \hat{\mathbf{r}}, \tag{3}$$

where the $p \times p$ -matrix $\hat{H} := \hat{H}_n = \mathcal{O}_p(c_n^{-1})$ and the p-vector $\hat{\mathbf{r}} := \hat{\mathbf{r}}_{i,n} = \mathcal{O}_p(c_n^{-1})$, uniformly in i, for some rate $c_n \to \infty$ as $n \to \infty$. Here and throughout the paper, for an arbitrary real sequence g_n, the notation $X_n = \mathcal{O}_p(g_n)$ is used to denote that the family of random variables $g_n^{-1} X_n$ is bounded in probability. Similarly, we use other Landau notation, such as $o(1)$ to indicate convergence towards zero. With $\to_p$, we denote convergence in probability, and with $\rightsquigarrow$, we indicate weak convergence, i.e., convergence in distribution.

In order to provide insight on the model (3), we next compare it to various models from the literature and inspect whether or not they can be seen as special cases of (3). Consider first the so-called blind source separation model where the observed $\mathbf{x}_i$ and the latent $\mathbf{z}_i$ are related through $\mathbf{x}_i = \boldsymbol{\mu} + \Omega \mathbf{z}_i$, for some unknown parameters $\boldsymbol{\mu} \in \mathbb{R}^p, \Omega \in \mathbb{R}^{p \times p}$. Under various assumptions for the latent variables $\mathbf{z}_i$, this model covers a wide range of different frameworks, such as independent component analysis (ICA) and second-order source separation, see [9]. Later in Section 4 we show that, as soon as one has estimators $\hat{\boldsymbol{\mu}}, \hat{\Omega}$ for the two model parameters (possibly up to some finite set of transformations such as permutations), then the estimated latent variables $\hat{\mathbf{z}}_i = \hat{\Omega}^{-1}(\mathbf{x}_i - \hat{\boldsymbol{\mu}})$ satisfy (3) for a

sequence c_n^{-1} equal to the convergence rate of $\hat{\boldsymbol{\mu}}, \hat{\Omega}$ (or the slowest of these if they differ). See Section 4 and, in particular, (12) for more details.

As our second example, we consider the orthogonal factor model (OFM) (see, e.g., [30]) defined as $\mathbf{x}_i = \boldsymbol{\mu} + L\mathbf{z}_i + \boldsymbol{\epsilon}_i$ where $\boldsymbol{\mu} \in \mathbb{R}^p$, $L \in \mathbb{R}^{p\times m}$ are parameters (the mean and the matrix of loadings, respectively) and the vector of latent variables (i.e., common factors) $\mathbf{z}_i$ has length $m \leq p$ and is independent of the vector of specific factors $\boldsymbol{\epsilon}_i$ whose covariance matrix Ψ is assumed to be diagonal. Furthermore, it is taken that both $\mathbf{z}_i$ and $\boldsymbol{\epsilon}_i$ have zero mean and that the former has identity covariance matrix. Even under these conditions, the OFM is not identifiable as one can always replace L and $\mathbf{z}_i$ with LT and $T^\top \mathbf{z}_i$, respectively, for any orthogonal $m \times m$ matrix $\mathbf{T}$. This ambiguity is commonly solved through factor rotations which essentially fix T by requiring $T^\top \mathbf{z}_i$ to satisfy certain additional conditions. For the purposes of this example, assume next that T has been uniquely fixed via a factor rotation (and absorbed into L). This means that estimators $\hat{\boldsymbol{\mu}}$, $\hat{L}$, $\hat{\Psi}$ for the model parameters can be obtained [30]. Using them, a standard way of estimating the factors is as

$$\hat{\mathbf{z}}_i = (\hat{L}^\top \hat{\Psi}^{-1} \hat{L})^{-1} \hat{L}^\top \hat{\Psi}^{-1} (\mathbf{x}_i - \hat{\boldsymbol{\mu}}).$$

Arguing next as in (12) later on, we observe that the above can be written as

$$\hat{\mathbf{z}}_i = \mathbf{z}_i + (\hat{L}^\top \hat{\Psi}^{-1} \hat{L})^{-1} \hat{L}^\top \hat{\Psi}^{-1} (L - \hat{L}) \mathbf{z}_i + (\hat{L}^\top \hat{\Psi}^{-1} \hat{L})^{-1} \hat{L}^\top \hat{\Psi}^{-1} (\boldsymbol{\mu} - \hat{\boldsymbol{\mu}} + \boldsymbol{\epsilon}_i). \tag{4}$$

Denote next $\hat{H} = (\hat{L}^\top \hat{\Psi}^{-1} \hat{L})^{-1} \hat{L}^\top \hat{\Psi}^{-1} (L - \hat{L})$ and $\hat{\mathbf{r}} = \hat{\mathbf{r}}_{i,n} = (\hat{L}^\top \hat{\Psi}^{-1} \hat{L})^{-1} \hat{L}^\top \hat{\Psi}^{-1} (\boldsymbol{\mu} - \hat{\boldsymbol{\mu}} + \boldsymbol{\epsilon}_i)$, as in (3). Now, even if the estimators $\hat{\boldsymbol{\mu}}$, $\hat{L}$, $\hat{\Psi}$ all converge at some rate c_n^{-1}, the decomposition (4) is still not of the form (3). This is because the specific factors $\boldsymbol{\epsilon}_i$ have a fixed covariance matrix Ψ, meaning that the term $(\hat{L}^\top \hat{\Psi}^{-1} \hat{L})^{-1} \hat{L}^\top \hat{\Psi}^{-1} \boldsymbol{\epsilon}_i$ is of the order $\mathcal{O}_p(1)$ and does not converge to zero when $n \to \infty$. Consequently, the OFM does *not* fall in our framework. Essentially, this result boils down to the fact that the OFM incorporates additive noise (via $\boldsymbol{\epsilon}_i$), whereas the earlier blind source separation model incorporates noise internally, i.e., by assuming that a subset of the factors $\mathbf{z}_i$ themselves are noise, which gets mixed with the signal factors to produce the observations $\mathbf{x}_i$. See Section 4 for more details on the blind source separation model. In Section 7 we briefly discuss the extension of our results to accommodate also the OFM.

As a third example, we observe that (3) also includes the simplest kind of general data contamination model where, for each $k \in \{1, \dots, p\}$ and $i \in \{1, \dots, n\}$, we have

$$\hat{z}_{ik} = z_{ik} + r_{ik},$$

and $r_{ik} = \mathcal{O}_p(c_n^{-1})$ uniformly in i and k. Such a model can be recovered from (3) by simply plugging in $\hat{H} = 0$.

Moving now back to work under the general framework of decomposition (3), we emphasize that it is not important how the decomposition came to be (be it through a latent variable model or something else), but instead simply having the form (3) is sufficient. Regardless, the existence of this form for $\hat{\mathbf{z}}_i$ might naturally require additional assumptions on the data-generating process, depending on the method used to produce it. As an example of such assumption, if independent component analysis is used, then any Gaussian components in $\mathbf{z}_i$ are not estimable (unless there is only one of them), see Section 4 for further discussion.

We also note that in this section we take the dimensionality p to be fixed. We impose this requirement because many latent factor models in the literature (in particular, both models in Section 4) make this assumption and without it we do not obtain the "decomposition" (3) on which our main theoretical results rely.

A common assumption in extreme value literature as well as in the latent variable literature is to assume that each component z_{ik} of the true non-observable signals $\mathbf{z}_i$ is strictly stationary, and has a univariate marginal F_k, i.e., each observation z_{ik} has marginal distribution F_k, for all i. The standard case of i.i.d. observations with components drawn from different distributions is then a special case of this model. However, while our main examples arise from stationary series falling into the above setting, our main results do not even require stationarity of the components z^k (covering, for example, a case where the marginal distributions vary in time but share a common extreme value index). Thus, our results could in principle be applied to non-stationary source separation estimators such as in [8], assuming that (2) can be first established.

For technical purposes and in order to ensure that our logarithm-based estimators are well-defined, we assume that marginals do not have point-mass at zero. That is, in the general non-stationary case, we assume that, for each component $k \in \{1, \dots, p\}$, we have

$$\lim_{\delta \to 0} \inf_{i \geq 1} \Pr\left(|z_{ik}| \geq \delta\right) = 1. \tag{5}$$

In the sequel, (5) is always assumed, even if it is not explicitly stated. Note that (5) is a natural assumption and not very restrictive. First of all, (5) implies that $\Pr(z_{ik} = 0) = 0$ for all i and k. Moreover, in the case of equal marginals, (5) is equivalent to $\Pr(z_{ik} = 0) = 0$. In the general case, (5) excludes also the situations where the observations come from a sequence of distributions $F_{i,k}$ that approach a distribution having point mass at zero.

In the sequel, the notation $|\mathbf{z}^k|$ refers to the sample $|z_{1k}|, \dots, |z_{nk}|$ of the absolute values of the kth latent series and $|\mathbf{z}^k|_{(m,n)}$ denotes the mth smallest element of $|\mathbf{z}^k|$.

Throughout the article, we make the following assumption.

Assumption 1. For all $k \in \{1, \dots, p\}$, there exists deterministic sequences a_{nk}, b_{nk} for which the kth component z_{ik} of $\mathbf{z}_i$ satisfies

$$\frac{|\mathbf{z}^k|_{(n,n)} - b_{nk}}{a_{nk}} = \mathcal{O}_p(1).$$

We stress that Assumption 1 is very relaxed, and in the extreme value theory literature it is usually taken as granted, without explicitly stating it. Indeed, if the observations are independent with a distribution function F, then Assumption 1 follows immediately whenever $F \in D(G_\gamma)$, i.e., F is in the domain of attraction of some extreme value distribution G_γ. Thus, in the case of independent observations, discussing the extreme value index γ without Assumption 1 is not sensible. The reason for explicitly posing Assumption 1 is that it is, however, not obvious whether it holds when arbitrary dependence structures are allowed between the $z_{ik}, i \in \{1, \dots, n\}$. For example, Assumption 1 is valid for ARMA processes, further examples being given in Section 4.2.

The main contribution of this article is the derivation of sufficient conditions under which the asymptotic properties of the Hill and moment estimators are preserved under Model (3). Intuitively, one would expect that these asymptotic properties remain the same, provided that c_n^{-1} vanishes rapidly enough to compensate the growth of the sample maximum of the heaviest component. Theorems 2 and 3 below contain the precise statements of this heuristic argument. In the sequel, we use the notation $g_{nk} = \max\{a_{nk}, b_{nk}\}$.

Theorem 2. *Let Assumption 1 hold and assume that,*

$$\frac{\max_\ell \{g_{n\ell}\}}{c_n} = o(1). \tag{6}$$

Let $k \in \{1, \dots, p\}$ be fixed and let C_H and C_M be arbitrary constants (depending on k).

(i) *If $\hat{\gamma}_H(|\mathbf{z}^k|) \to_p C_H$, then $\hat{\gamma}_H(|\hat{\mathbf{z}}^k|) \to_p C_H$.*

(ii) *If $\hat{\gamma}_H(|\mathbf{z}^k|) \to_p C_H$, $\dfrac{\max_\ell \{g_{n\ell}\}}{c_n \hat{\gamma}_H(|\mathbf{z}^k|)} \to_p 0$, $\hat{\gamma}_M(|\mathbf{z}^k|) \to_p C_M$, then $\hat{\gamma}_M(|\hat{\mathbf{z}}^k|) \to_p C_M$.*

Note that in the above result it is not required that C_H and C_M are the correct extreme value indices — any constants suffice. Actually, we prove that the difference between estimators based on "true" signals and estimated signals converges to zero in probability. Now consequently, the part (i) of Theorem 2 simply states that whenever the Hill estimator based on the "true" latent signals converges towards some constant, then the Hill estimator based on the estimated latent signals converges towards the same constant. The reason behind our formulation is that usually, as is the case for independent observations, the Hill estimator converges towards $\max(0, \gamma)$, see [10], pp. 101. In other words, the Hill estimator vanishes for distributions that are not heavy tailed. For such distributions, one can then apply the moment estimator. Part (ii) of Theorem 2 says that whenever both, the Hill estimator and the moment estimator based on the true latent signals $|\mathbf{z}^k|$, converge towards any constants, then the Hill and the moment estimator based on the estimated latent signals converge towards the same constants. As in most cases the Hill estimator converges towards $\max(0, \gamma)$, and does not explode, part (ii) of Theorem 2 implies that asymptotic properties of the moment estimator are inherited to the estimated model as well. The extra condition in part (ii) concerns the case when the Hill estimator converges towards zero, $C_H = 0$, and ensures that this convergence is not too rapid in comparison to the growth of the heaviest tail. In many cases of interest, the convergence rate of the Hill estimator is $\sqrt{k_n}$. This leads to the same condition as in Theorem 3, and can be achieved by a suitable choice of k_n. Finally, we stress that an examination of the proof of Theorem 2 reveals that the item (ii) is valid as long as the Hill estimator does not tend to infinity. Thus one can safely apply the moment estimator for short tailed distributions under Model (3). Finally, we remark that while in the case of independent observations one can obtain better sufficient condition

$$\frac{\max_\ell \{g_{n\ell}\}}{c_n F_k^{\leftarrow}\left(1 - \frac{k_n}{n}\right)} = o(1),$$

where $F_k^{\leftarrow}$ denotes the quantile function of $|\mathbf{z}^k|$, this is not true for arbitrary dependent sequences. In the case of independent observations, note also that $\max_\ell \{g_{n\ell}\}$ corresponds to the heaviest component, cf. Lemma 1. However, this is either not true in general for dependent sequences.

In order to gain better understanding on the behavior of the estimators, we next consider their limiting distributions. The following result is similar in spirit to the work of [7] who detail conditions under which the pre-filtering of innovations does not affect the limiting distribution of the Hill estimator under a GARCH-model.

Theorem 3. *Let Assumption 1 hold and assume that,*

$$\frac{\sqrt{k_n} \max_\ell \{g_{n\ell}\}}{c_n} = o(1). \tag{7}$$

Let $k \in \{1, \dots, p\}$ be fixed and let $C_H, C_M, \mu_H, \mu_M, \sigma_H$, and σ_M be arbitrary constants (depending on k).

(i) *If $\sqrt{k_n}\{\hat{\gamma}_H(|\mathbf{z}^k|) - C_H\} \rightsquigarrow \mathcal{N}(\mu_H, \sigma_H^2)$, then*

$$\sqrt{k_n}\{\hat{\gamma}_H(|\hat{\mathbf{z}}^k|) - C_H\} \rightsquigarrow \mathcal{N}(\mu_H, \sigma_H^2).$$

(ii) *If $\hat{\gamma}_H(|\mathbf{z}^k|) \to_p C_H$, $\dfrac{\sqrt{k_n} \max_\ell \{g_{n\ell}\}}{c_n \hat{\gamma}_H(|\mathbf{z}^k|)} \to_p 0$, $\sqrt{k_n}\{\hat{\gamma}_M(|\mathbf{z}^k|) - C_M\} \rightsquigarrow \mathcal{N}(\mu_M, \sigma_M^2)$, then*

$$\sqrt{k_n}\{\hat{\gamma}_M(|\hat{\mathbf{z}}^k|) - C_M\} \rightsquigarrow \mathcal{N}(\mu_M, \sigma_M^2).$$

Remark 1. It is customary to analyze extreme value index estimators by using asymptotic results for empirical quantiles. In the i.i.d. case, for example, one can rely on Donsker-type functional limit theorem for the empirical quantiles which then yields a central limit theorem, e.g. for the Hill estimator (see [10]). However, in our general case the limiting process for the empirical quantiles is not necessary arising from the Brownian motion. (The limiting process might be, for example, fractional Brownian motion or a general Hermite process.) For this reason we prove Theorem 3 by studying the errors in the estimators directly. One could further analyze the asymptotic behavior of the extreme value index estimators separately under different limiting processes of the quantiles, but that is beyond the scope of this article.

In the above result, the constants μ_H, μ_M, σ_H, and σ_M can be computed explicitly in most cases, their exact values depending on the so-called second order conditions. For details, we refer to [10]. We also remark that in our proof, we could easily replace the convergence rate $\sqrt{k_n}$ with some other rate, or the limiting normal distribution with some other distribution. The underlying reason for the above formulation is that asymptotic results for extreme value index estimators where the rate is other than $\sqrt{k_n}$ or where the limiting distribution is not normal seem to be extremely scarce in the literature. The only example we are aware of is [12] where the limiting distribution is the standard Gumbel distribution and the rate is $\ln(k_n)$ (for some specific choices of k_n).

We end this section by discussing the strictness of the key conditions $\max_\ell\{g_{n\ell}\} = o(c_n)$ and $\sqrt{k_n}\max_\ell\{g_{n\ell}\} = o(c_n)$. These conditions state that the convergence rate c_n of the estimated latent sample to the true latent sample must be sufficiently fast compared both to $\sqrt{k_n}$, the square root of the tail threshold, and to $\max_\ell\{g_{n\ell}\}$, the heaviness of the heaviest of the latent components. Moreover, the rate k_n can be seen as a type of a tuning parameter. Choosing a faster growing k_n will make the Hill estimator converge more rapidly, but it will, at the same time, limit the range of distributions whose extreme value indices we can estimate in the first place.

To shed further light on these conditions, we consider an example. Assume that there exists at least one latent component belonging to the domain of attraction of the Fréchet distribution, i.e., $\max_\ell\{g_{n\ell}\} = \mathcal{O}(n^\gamma)$ for some $\gamma > 0$ (cf. Lemma 1 in Appendix A). Now, letting $k_n = n^\alpha$ for some $\alpha > 0$ and under the standard rate $c_n = \sqrt{n}$, we end up with the restriction $\gamma < (1-\alpha)/2$. Hence, taking a sufficiently small number k_n of upper order statistics, we see that extreme value index estimation is feasible as long as the heaviest Fréchet component among the latent variables has its extreme value index smaller than $1/2$, that is, all latent components have finite variance. Heavier components, i.e., ones without second moments, can be captured through estimators which yield faster convergence rates c_n than the usual $\sqrt{n}$ for the model estimation. Conversely, if the convergence rate c_n is slower than the usual $\sqrt{n}$ (see, e.g., [33]), then c_n might not be sufficient to compensate too heavy tails, and, e.g., assumptions on the existence of higher moments are required.

4. Example models

In this section, we illustrate the applicability of our main results by considering two popular example models: independent component model and stationary second order source separation model. The first of these is aimed for i.i.d. observations and the second one for time-dependent data. For simplicity, we consider only the Hill estimator, although the following analysis could be easily extended for the moment estimator as well (see Remark 2). Throughout, we assume that the heaviest component has index $\gamma > 0$, often implying that $\max_\ell\{g_{n\ell}\} = \mathcal{O}(n^\gamma)$, see the examples below. As the rate $c_n = \sqrt{n}$ is the best possible that one can usually expect, we also assume $\gamma < 1/2$. This ensures the square integrability of all of our random variables, which is also a minimum requirement for (3) to hold for the standard estimation procedures in our example models.

Let now $k \in \{1, \dots, p\}$ be fixed. We illustrate our results in cases where both Theorems 2 and 3 are applicable. Thus, in order to obtain limiting normality for the Hill estimator $\hat{\gamma}_H(|\mathbf{z}^k|)$ based on the true values $|\mathbf{z}^k|$, we impose a second order condition for the marginal distribution F of $|z^k|$. The distribution F is called second order regularly varying (with index γ) if there exists a positive or negative function A with the property $\lim_{t\to\infty} A(t) = 0$ such that, for all $x > 0$,

$$\lim_{t\to\infty} \frac{\frac{U(tx)}{U(t)} - x^\gamma}{A(t)} = x^\gamma \frac{x^\rho - 1}{\rho}, \tag{8}$$

holds for some real number $\rho \leq 0$. Here the function U is given by

$$U = \left(\frac{1}{1-F}\right)^{\leftarrow},$$

where $^{\leftarrow}$ denotes the left-continuous (pseudo-)inverse function, see, e.g., [10, Section 1.1.2]. Then, in the case of independent observations, the limiting normality,

$$\sqrt{k_n}\left\{\hat{\gamma}_H(|\mathbf{z}^k|) - \gamma\right\} \rightsquigarrow \mathcal{N}\left(\frac{\lambda}{1-\rho}, \sigma^2\right), \tag{9}$$

holds, provided that $\lim_{n\to\infty} \sqrt{k_n} A\left(n/k_n\right) = \lambda \in \mathbb{R}$. This leads to an upper bound on the rate at which k_n can grow. Similarly, conditions of Theorems 2 and 3 give upper bounds for the rate at which k_n can grow. Thus, we can obtain limiting normality (and consistency) by choosing a not-too-rapidly growing sequence k_n, at the cost of a slower rate of convergence. For details on the limiting normality of the Hill estimator in the case of i.i.d. observations, see [10], and in the case of stationary dependent observations, see [11] and the references therein.

4.1. Independent component model

In independent component analysis (ICA) the observed p-vectors $\mathbf{x}_1, \ldots, \mathbf{x}_n$ are assumed to be a random sample from the independent component (IC) model,

$$\mathbf{x} = \Omega\mathbf{z} + \boldsymbol{\mu}, \tag{10}$$

where the latent p-vector $\mathbf{z}$ has independent components, $\Omega \in \mathbb{R}^{p\times p}$ is invertible and $\boldsymbol{\mu} \in \mathbb{R}^p$ is a location parameter that can be taken to be the zero vector as all standard estimators involve centering of the data [28] . The objective in ICA is find an unmixing matrix $\Gamma \in \mathbb{R}^{p\times p}$, such that $\Gamma\mathbf{x}$ has independent components. Standard theory then shows that if at most one of the ICs is Gaussian, any such solution coincides with $\mathbf{z}$ up to scaling, order and signs of the components. The scales can be fixed by second order standardization of $\mathbf{z}$. This guarantees that all the solutions are of the form $\Gamma = PJ\Omega^{-1}$ where $P \in \mathbb{R}^{p\times p}$ is a permutation matrix and $J \in \mathbb{R}^{p\times p}$ is a sign-change matrix (diagonal matrix with diagonal entries equal to ± 1). In our approach in assessing extreme behavior, the sign ambiguity is of no concern, as we consider the absolute values of the source components. Moreover, the order of the components is irrelevant, as we anyway order them later based on their tail behavior.

It is simple to show that, in (10), every observed variable in $\mathbf{x}$ inherits the extreme value index of the most heavy-tailed component in $\mathbf{z}$ (unless Ω contains zeros in appropriate locations). Thus, by analyzing $\mathbf{x}$ directly, one will discover only the behavior of the most extreme latent variable. Whereas, by first estimating $\mathbf{z}$ with ICA, we can uncover also other behaviors besides the most extreme one. The equivalent holds also in a non-linear case, $\mathbf{x} = f(\mathbf{z})$, with the exception that the extreme value indices of $\mathbf{x}$ are further affected by the growth rates of the corresponding component functions of f at infinity. For example, if f_1 grows eventually as $x \mapsto x^2$, then this doubles the corresponding extreme value index.

Numerous estimators $\hat{\Gamma}$ of the unmixing matrix have been proposed and under suitable assumptions and standardizations, most estimators converge to Ω^{-1} at some rate $c_n \to \infty$, as $n \to \infty$, that is,

$$c_n\left(\hat{\Gamma}\Omega - I_p\right) = \mathcal{O}_p(1), \tag{11}$$

where we implicitly assume that the signs and order of the rows of $\hat{\Gamma}$ have been chosen appropriately. Typically, in the context of i.i.d. observations, we have $c_n = \sqrt{n}$ (implying that the heaviest Fréchet latent component should have $\gamma < 1/2$ for our Theorems 2 and 3 to apply). See [39,41] for several examples including FastICA, fourth order blind identification (FOBI) [5], and joint approximate diagonalization of eigenmatrices (JADE) [6]. Also the ICA-estimators based on the simultaneous diagonalization of two symmetrized scatter matrices [42,44] can be shown, using the techniques of [29], to have the rate $\sqrt{n}$, assuming that the applied symmetrized scatter matrices have the same convergence rate.

Assuming that $\hat{\Gamma}$ is of the form (11), the estimated latent vectors can be written as

$$\hat{\mathbf{z}}_i = \hat{\Gamma}\left(\mathbf{x}_i - \bar{\mathbf{x}}\right) = \hat{\Gamma}\Omega\left(\mathbf{z}_i - \bar{\mathbf{z}}\right) = \mathbf{z}_i + \left(\hat{\Gamma}\Omega - I_p\right)\mathbf{z}_i - \hat{\Gamma}\Omega\bar{\mathbf{z}}. \tag{12}$$

Writing now $\hat{H} := \hat{\Gamma}\Omega - I_p$ and $\hat{\mathbf{r}} := -\hat{\Gamma}\Omega\bar{\mathbf{z}}$, we observe that we have arrived to the form (3). By the assumption that $\max_\ell\{g_{n\ell}\} = \mathcal{O}(n^\gamma)$ with $\gamma < 1/2$, we observe that (6) is automatically valid, and (7) is valid for a suitably chosen sequence k_n. However, note that most standard ICA methods operate on higher-order information, making still stronger moment assumptions. For example, the $\sqrt{n}$-consistency of FOBI requires the existence of finite eighth moments of the latent variables [29] (along with the assumption that the kurtoses of the latent variables are distinct). By using squared FastICA with the hyperbolic tangent as the ICA-estimator [37], one can reduce the order of the moments required to achieve $\sqrt{n}$-consistency to four (assuming that certain moments of the latent variables are suitably distinct). Ultimately, by conducting ICA through a simultaneous diagonalization of two symmetrized robust $\sqrt{n}$-consistent scatter matrices, no finite moments are required at all. Note still that even in this case, the convergence rate $\sqrt{n}$ poses the restriction $\gamma < 1/2$.

While the aforementioned assumption of at most one Gaussian latent variable is standard in ICA, it can be weakened in our context. Namely, even if several of the latent variables were Gaussian, the remaining latent variables are still estimated consistently by the methods listed in Section 4.1 (and by their deflation-based versions, see Section 7). This means that our main results in Section 3 can be applied to the subset of estimated non-Gaussian latent variables to estimate their extreme value behavior. Note also that this restriction essentially does not cause us to lose any information since Gaussian variables have a non-interesting extreme value behavior (i.e., they have an extreme value index of zero).

Finally, we stress that under independent observations drawn from a second order regular varying heavy tailed distribution, the Hill estimator $\hat{\gamma}_H(|\mathbf{z}^k|)$ is consistent and asymptotically normal (see [10]), and while different technical assumptions are required for the classical ICA-estimators, none of them interfere with our assumption that guarantees the consistency and limiting normality of the Hill estimator. Thus, as a conclusion, we can safely apply Theorems 2 and 3.

Remark 2. In the above discussions we have considered only the Hill estimator. However, applying Theorem 2 or Theorem 3 for the moment estimator in the ICA-context is straightforward. Indeed, under second order regularly varying tails and independence, the Hill estimator always converges to $\max(0, \gamma)$, and the moment estimator is both consistent and asymptotically normal. Thus it suffices to check the extra condition $\sqrt{k_n}\max_\ell\{g_{n\ell}\}/\{c_n\hat{\gamma}_H(|\mathbf{z}^k|)\} \to_p 0$. However, even if $\gamma \leq 0$, $\sqrt{k_n}\hat{\gamma}_H(|\mathbf{z}^k|)$ converges (see the proof of Theorem 3.5.4 in [10]), and thus it suffices to choose the sequence k_n such that $k_n \max_\ell\{g_{n\ell}\}/c_n = o(1)$.

4.2. Second order source separation model

Our second example moves to the realm of signal processing and blind source separation (BSS). Like the IC model, also the second order BSS model is linear and based on the general location-scatter model. In the model, the observed run $\mathbf{x}_1, \dots, \mathbf{x}_n$ of a stationary p-variate time series is assumed to have the instantaneous latent representation,

$$\mathbf{x}_i = \Omega \mathbf{z}_i + \boldsymbol{\mu}, \quad i \in \{1, \dots, n\}, \tag{13}$$

where the latent p-variate time series $\mathbf{z}_i$ is stationary and has standardized uncorrelated components, and $\Omega \in \mathbb{R}^{p \times p}$ is of full rank. The location $\boldsymbol{\mu} \in \mathbb{R}^p$ is (by stationarity) trivial to estimate by using a standard average estimator, which provides a consistent estimator if the system is ergodic. Thus, for the sake of simplicity, it will be omitted in the following. Note also that the non-identifiability of signs and order holds in the BSS model as well. However, for our purposes this does not matter due to the reasons explained in Section 4.1.

One standard approach to estimate $\mathbf{z}_i$ is algorithm for multiple unknown signals extraction (AMUSE) [55] where the autocovariance matrices $\Sigma_\tau(\mathbf{x}_i) = \mathrm{E}(\mathbf{x}_i \mathbf{x}_{i+\tau}^\top)$ for $\tau \in \{0, \tau_0\}$ are diagonalized simultaneously. An extension of AMUSE that is less sensitive to the choice of τ_0 is the second order blind identification (SOBI) [1] algorithm where the autocovariance matrices $\Sigma_\tau(\mathbf{x}_i) = \mathrm{E}(\mathbf{x}_i \mathbf{x}_{i+\tau}^\top)$ over a chosen set of lags $\mathcal{T} = \{\tau_1, \dots, \tau_{|\mathcal{T}|}\}$ are jointly diagonalized. As in the IC-model, one would expect that the algorithm provides a consistent estimator $\hat{\Gamma}$ with some rate $c_n \to \infty$, as $n \to \infty$, that is,

$$c_n \left(\hat{\Gamma} \Omega - I_p \right) = \mathcal{O}_p(1), \tag{14}$$

where, again, suitable matrices P, J are implicitly assumed to be part of the estimator $\hat{\Gamma}$. It turns out that (14) holds true whenever

$$c_n \left\{ \hat{\Sigma}_\tau(\mathbf{x}_i) - \Sigma_\tau(\mathbf{x}_i) \right\} = \mathcal{O}_p(1), \quad \tau \in \mathcal{T}, \tag{15}$$

where $\hat{\Sigma}_\tau(\mathbf{x}_i)$ denotes the estimator of the autocovariance matrix $\Sigma_\tau(\mathbf{x}_i)$. The fact that (15) implies (14) is proved in the case of complex valued AMUSE in [33] with general rate c_n, and in the case of real valued SOBI (and its variants) in [35] for the rate $c_n = \sqrt{n}$ (both results assume the distinctness of specific autocovariances). It is also straightforward to check that the arguments of [35] apply with arbitrary rate function c_n. For examples with a general rate $c_n = n^\beta$, $0 < \beta < 1/2$, instead of the standard $\sqrt{n}$, we refer to [33].

Eq. (15) is the first key assumption on the rate of convergence for the autocovariance estimators, which on the other hand gives us our speed c_n. If c_n is non-standard, (6) gives us also the restriction $n^\gamma = o(c_n)$, limiting the possible values of γ. This can be seen as an interchange between moment assumptions and the speed at which the autocovariance estimators converge, as higher moments are required if the estimators converge slowly.

In order to make Theorem 3 applicable, we also require that the Hill estimator $\hat{\gamma}_H(|\mathbf{z}^k|)$ satisfies limiting normality (9). Compared to independent observations, the problem is much more subtle in the case of dependent sequences and one needs to pose extra assumptions in addition to the second order regularly varying condition (8). The extra assumptions are, roughly speaking, conditions that ensure the dependence to be weak enough so that the series "behaves" similarly as a series of independent observations. The precise definition of weak dependence or asymptotic independence varies in the literature. Usually asymptotic independence is encoded to mixing-conditions (for different notions of mixing-conditions and their relations, see the survey in [2]). It is known (see, e.g., [16,18]) that (9) holds provided that $|\mathbf{z}^k|$ forms a β-mixing stationary sequence such that some minor additional regularity conditions are met (see, e.g., conditions (a)–(c) of [11]). In particular, all these conditions are satisfied for the following sequences:

- m-dependent process and AR(1)-process [18,50,51],
- AR(p)-processes and MA(∞)-processes (with suitable assumptions on the coefficients) [17,46],
- MA(q)-processes [17,25,50,51],
- ARCH(1)-processes [17,18],
- GARCH-processes [16,53].

We emphasize that the above examples form a very large and applicable class of processes. For details and more information on the above examples, see also [11].

We now turn back to extreme value index estimation under the BSS model. In order to apply Theorem 2 or Theorem 3, it suffices to make sure that, for a given heavy tailed component $|\mathbf{z}^k|$, the above mentioned conditions guaranteeing the limiting normality (9) for the Hill estimator $\hat{\gamma}_H(|\mathbf{z}^k|)$ are satisfied. At the same time, one needs that (15) holds, with some rate c_n satisfying $n^\gamma = o(c_n)$. After that, it remains to choose k_n not increasing too rapidly so that (7) holds as well. We next explore the connection between the given assumptions. We first observe that conditions required to ensure (9) are solely on the dependence structure and distribution of the given component $|\mathbf{z}^k|$ of interest. At the same time, condition (15) considers the rate of convergence of autocovariance estimators for all components simultaneously. Note that β-mixing and assumptions (a)–(c) of [11] do not imply convergence of the autocovariance estimators (as the conditions do not even require existence of second moments). Conversely, convergence of the autocovariance estimators is related to the so-called ρ-mixing (see [2] for precise definition) which does not imply β-mixing. Thus, even the convergence rate of the autocovariance estimator of the component $|\mathbf{z}^k|$ does not provide any information regarding the

validity of conditions implying limiting normality (9) for the Hill estimator. This means that the assumptions do not contradict, and also that in practice, one has to verify (15) and the limiting normality of the Hill estimator separately.

Remark 3. If all components are ρ-mixing, then the slowest decay of the ρ-mixing coefficients gives us an upper bound for c_n. Moreover, if one poses a stronger mode of mixing, ϕ-mixing, for the sequence $|\mathbf{z}^k|$, then [2, p.112] the sequence is also both β- and ρ-mixing.

5. Simulations

In this section, we illustrate the tail index estimation under the second order source separation model of Section 4.2 via two simulation studies. We study the effect of the sample size n in Section 5.1 and the effect of the dimension p in Section 5.2. Appendix B in the supplementary material presents additional simulations for the independent component model in Section 4.1, with largely the same conclusions as in Section 5.1.

5.1. Sample size

In the first simulation study, we consider the $\mathbb{R}^3$-process $\mathbf{z}$, where the components are n-length realizations, i.e., time series, of the independent stochastic processes in $\tilde{\mathbf{z}} = (\mathtt{ARCH}(1), \mathtt{D}^{(1)}, \mathtt{D}^{(2)})^\top$. The first component of $\tilde{\mathbf{z}}$ is an ARCH(1)-process with the parameter vector $(\alpha_0, \alpha_1) = (1/4, (2^3\sqrt{2/\pi})^{-2/5})$. At time t, the second and third components are defined as $\mathtt{D}_t^{(1)} = (B_{t+1}^{(1)} - B_t^{(1)})^2 - 1$ and $\mathtt{D}_t^{(2)} = (B_{t+1}^{(2)} - B_t^{(2)})^2 - 1$, where $B^{(1)}$ denotes a fractional Brownian motion (fBm) with Hurst parameter $3/4$ and $B^{(2)}$ denotes a fBm with Hurst parameter $4/5$, such that $B^{(1)}$ are $B^{(2)}$ are mutually independent. For a comprehensive study on fBm, see, e.g., [43]. Out of the three components, the ARCH(1) process has the largest theoretical extreme value index $1/5$ [40, Theorem 2.1], and our objective is to estimate it. The remaining two components have extreme value indices equal to zero (this follows as both have unbounded support but finite moments of all order). We considered the sample sizes $n \in \{300, 10^3, 10^4, 10^5, 10^6, 10^7\}$ and the threshold sequence k_n was chosen to be $k_n = \lfloor n^{1/4} \rfloor$. For each sample size, the simulation was iterated 2000 times.

As a preliminary step, the simulated observations $\tilde{\mathbf{z}}_i$ were centered by subtracting the sample mean. Here, the centered observations are denoted as $\mathbf{z}_i$. In every iteration $h \in \{1, \dots, 2000\}$, we applied, for all $i \in \{1, \dots, n\}$, the linear transformation $\mathbf{x}_i = \Omega_h \mathbf{z}_i$, where the elements of the $\mathbb{R}^{3\times 3}$-matrix Ω_h were simulated independently, and separately in every iteration, from the univariate uniform distribution $\mathtt{unif}(-100, 100)$ (we let the matrix Ω_h vary between the iterations in order to cover a larger range of mixing matrices than just a few fixed ones). We then applied the AMUSE unmixing procedure with lag $\tau = 1$ to the mixed time series, using the implementation contained in the R-package JADE [38]. The existence of the limiting distribution of the AMUSE unmixing estimator requires finite fourth moments. Note that the ARCH(1) parameters α_0, α_1 are chosen such that the fourth moments exist for all components. We denote the absolute values of the AMUSE unmixed time series and the absolute values of the original centered time series as $|\hat{\mathbf{z}}|$ and $|\mathbf{z}|$, respectively.

Now, we have $\max_\ell(g_{n\ell}) = n^{1/5}$, which corresponds to the ARCH(1) process. The $\mathtt{D}^{(2)}$ process in the third component has the slowest rate of convergence, giving $c_n = n^{2/5}$, see [33]. Hereby, under our choice of $k_n = \lfloor n^{1/4} \rfloor$, we have that the assumptions required by Theorems 2 and 3 hold and, hence, for large sample sizes, the extreme value index estimates calculated from $|\hat{\mathbf{z}}|$ and $|\mathbf{z}|$ are expected to be close to each other.

We estimated the extreme value indices for every component from both $|\hat{\mathbf{z}}|$ and $|\mathbf{z}|$, using both the Hill estimator and the moment estimator. Note that both estimators produce three extreme value index estimates, one for each component. To capture the ARCH(1) component, we collected, in every simulation iteration, the largest of the three estimates, denoted in the following by $\hat{\gamma}(|\hat{\mathbf{z}}|) := \max\{\hat{\gamma}(|\hat{\mathbf{z}}^1|), \hat{\gamma}(|\hat{\mathbf{z}}^2|), \hat{\gamma}(|\hat{\mathbf{z}}^3|)\}$ and $\hat{\gamma}(|\mathbf{z}|) := \max\{\hat{\gamma}(|\mathbf{z}^1|), \hat{\gamma}(|\mathbf{z}^2|), \hat{\gamma}(|\mathbf{z}^3|)\}$ (this induces a slight bias to the results which is, however, rendered negligible with increasing n). The histograms of $\hat{\gamma}(|\hat{\mathbf{z}}|)$ and $\hat{\gamma}(|\mathbf{z}|)$ for sample sizes $n \in \{300, 10^3, 10^4\}$ are shown in Fig. 1, where the extreme value indices estimated from $|\hat{\mathbf{z}}|$ correspond to light blue color, and the extreme value index estimates calculated from the original $|\mathbf{z}|$ correspond to light red color. Dark blue color is used for the parts of the histograms that overlap and the dashed yellow vertical line represents the theoretical extreme value index value $\gamma = 1/5$. To keep the plots legible, values smaller than -2 are omitted from the figure; a total of 21 moment estimator estimates were smaller than -2.

In Fig. 1, already for the small sample size $n = 300$, the two histograms overlap significantly. Moreover, starting from $n = 1000$, the histograms are basically identical, showing that, as predicted by the theory, the effect of the BSS-step on the estimation of the extreme value indices is almost negligible. When comparing the Hill estimator and the moment estimator, Fig. 1 indicates that the variance of the moment estimator is larger, when compared to the Hill estimator. Since the moment estimator is known to have larger asymptotic variance than the Hill estimator for i.i.d. data [10], this behavior is not too surprising in our context either. In addition, the bias of the Hill estimator is visible in the histograms and seems to decrease as the sample size increases. The histograms corresponding to the sample sizes $10^5, 10^6$ and 10^7 have been omitted here, as they introduce no new information to the simulation study.

Fig. 2 illustrates the absolute differences, scaled with $\sqrt{k_n}$, between the estimates calculated from $|\mathbf{z}|$ and $|\hat{\mathbf{z}}|$. The red and blue curves represent the first and third empirical quartiles of the absolute differences, respectively, and the yellow curve is the corresponding sample median curve. The differences can be seen to converge to zero for both estimators, but the moment estimator requires larger sample sizes for this. That is, the quartile Q_3 for the Hill estimator is close to zero already with $n = 10^5$ and, conversely, the moment estimator quartile Q_3 requires samples of size $n = 10^7$ for achieving the same magnitude.

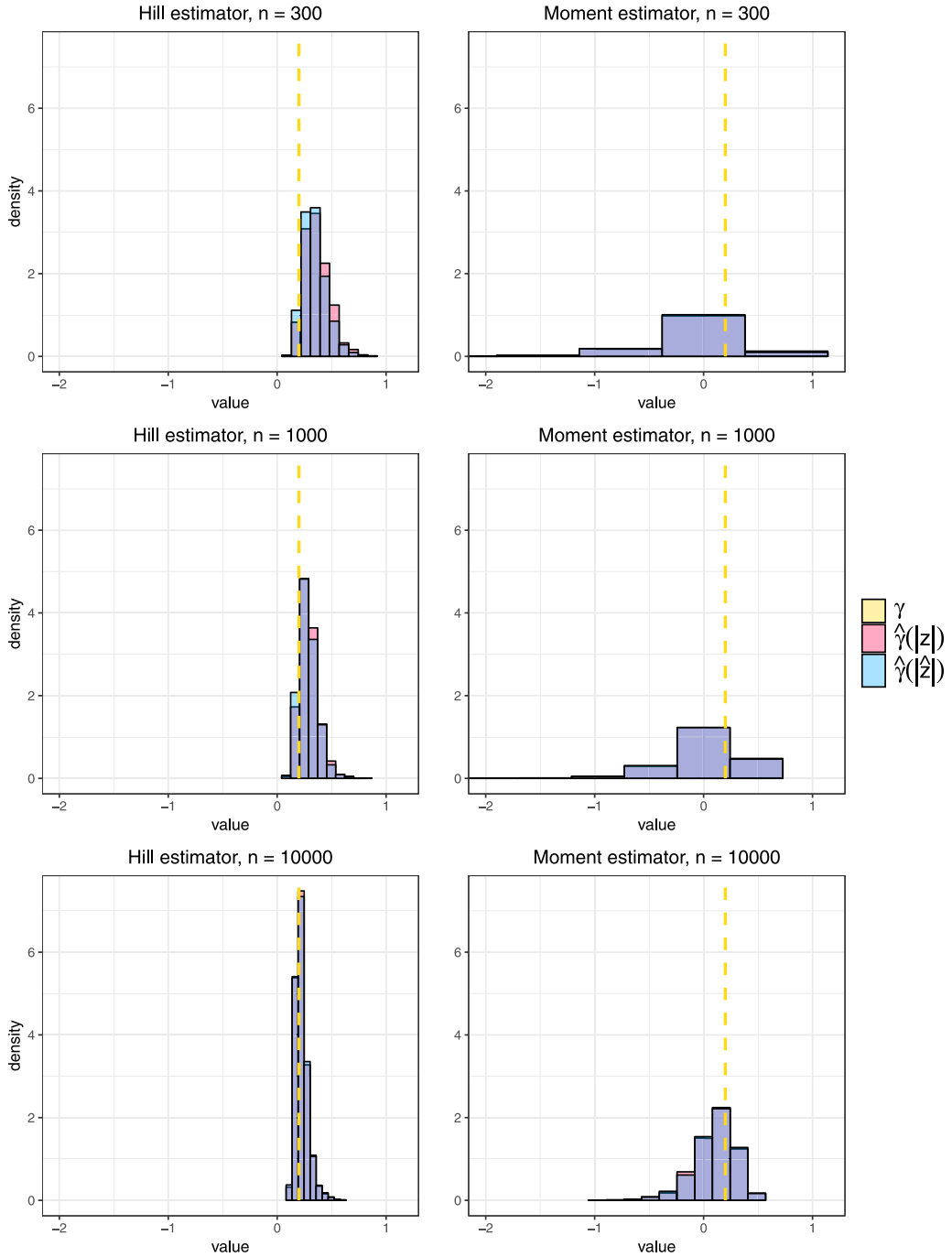

Fig. 1. Histograms of $\hat{\gamma}(|\mathbf{z}|)$ (light red) and $\hat{\gamma}(|\hat{\mathbf{z}}|)$ (light blue) in the simulation study with sample sizes 300, 1000 and 10 000. The dashed yellow vertical line is the theoretical extreme value index $\gamma = 1/5$. The dark blue color in the histograms represents the area, where the two histograms overlap. (For interpretation of the references to color in this figure legend, the reader is referred to the web version of this article.)

5.2. Dimension

In the second simulation study, we consider $\mathbb{R}^p$-processes $\mathbf{z}$, where the components are 10^5-length realizations of the independent processes in $\tilde{\mathbf{z}} = (\texttt{ARCH}(1), N^{(1)}, \dots, N^{(p-1)})^\top$. The ARCH(1) component is defined as in Section 5.1. The remaining $p - 1$ noise

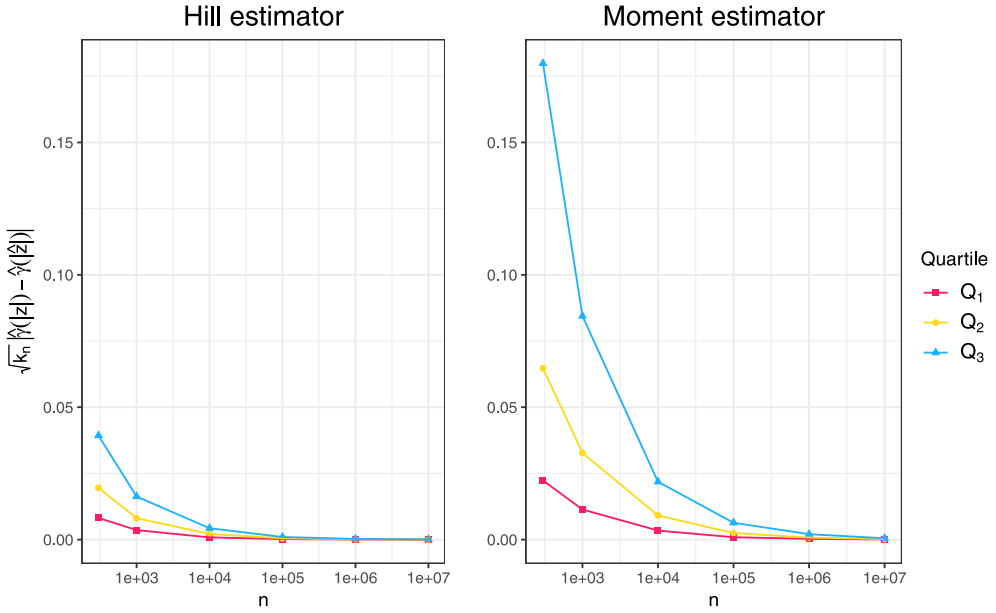

Fig. 2. The quartiles of $\sqrt{k_n}|\hat{\gamma}(|\mathbf{z}|) - \hat{\gamma}(|\hat{\mathbf{z}}|)|$, for the Hill estimator and the moment estimator as a function of the sample size n.

components $N^{(i)}$ are all zero mean AR(1)-ARCH(1) processes with AR parameter $(-1/4)$ and ARCH(1) parameters $(1/4,\ 2^{-1} \cdot \pi^{1/9} \cdot (4!)^{-2/9})$. The study is conducted as in Section 5.1 and the results are presented in Fig. 3. Here, the matrix Ω_h is a $p \times p$ matrix simulated using the same logic as before and $k_n = \lfloor \sqrt{10^5} \rfloor$.

In this study, the ARCH(1)-process is again the one with the heaviest tail and AMUSE aims to separate it from the $p - 1$ noise components. The estimators produce p estimates for every simulation round, and again the largest of them is collected. Note that, contrary to the first simulation study, the noise processes now also have a positive theoretical extreme value index γ.

From Fig. 3, we observe that the separation gets more difficult as the amount of noise components increases, making the average difference between the two estimates of $\hat{\gamma}$ grow with p. We also see that the increasing number of noise components has a more dramatic effect on the moment estimator quartiles, its behavior deteriorating faster as p grows. This effect is expected as the moment estimator is based on higher moments of the tail observations than the Hill estimator, making it, in general, more sensitive to noise.

6. Data example

Heavy-tailed distributions are encountered frequently in financial contexts [45]. Here, we consider extreme value index estimation for a currency exchange rate data available at Kaggle (https://www.kaggle.com/brunotly/foreign-exchange-rates-per-dollar-20002019/). The data consist of the daily exchange rates for 22 currencies against the U.S. dollar in the period of January 3rd, 2000 – December 31st, 2019. To showcase the proposed method's utility, we conduct a comparative analysis between it and the classic approach of estimating the extreme value indices of each observed series individually. As a pre-processing, we removed all dates having missing values for any of the variables, then transformed the observations to log-returns and finally standardized them to have unit variances, which is without loss of generality as both our extreme value index estimators are scale invariant. The resulting data set has 5015 observations of the 22-variate time series.

The full multivariate time series are visualized in Figs. C3 and C4 in the supplementary Appendix C. Several volatility spikes affecting both individual series and larger subsets of them are visible, the financial crisis of 2007–2008 being especially pronounced. A more accurate description of the most volatile periods is obtained from the top plot of Fig. 4, showing extreme value indices of the individual 22 series. Note that we have omitted the legend from Fig. 4 to keep the plot concise and, to differentiate between the individual currencies, the same series are given in separate plots in Figs. C5 and C6 in the supplementary Appendix C. The estimation in Fig. 4 was conducted by moving a window of length 300 days through each univariate series and estimating the extreme value index of the absolute value of the series in each window with the Hill estimator with the tail length $k_n = 30$. Experimentation (not shown here) revealed that the resulting estimates are rather robust to the choice of the parameter value. The x-axis values in the plot correspond to the middle days (150th days) of the windows. The Hill estimator plot reveals multiple time periods where various subsets of the series demonstrate simultaneous volatility (again, most notably the period of 2007–2008), giving plausibility to the hypothesis that there exist multiple latent series driving the extreme behavior of the observed series.

Before fitting a latent variable model to the data, we note that with a data set such as the current one, there is the possibility of conducting "expert knowledge" variable selection. For example, if one of the currencies is not floating against USD during the observation period, there is certain motivation to drop the corresponding variable out. However, such actions should, in general,

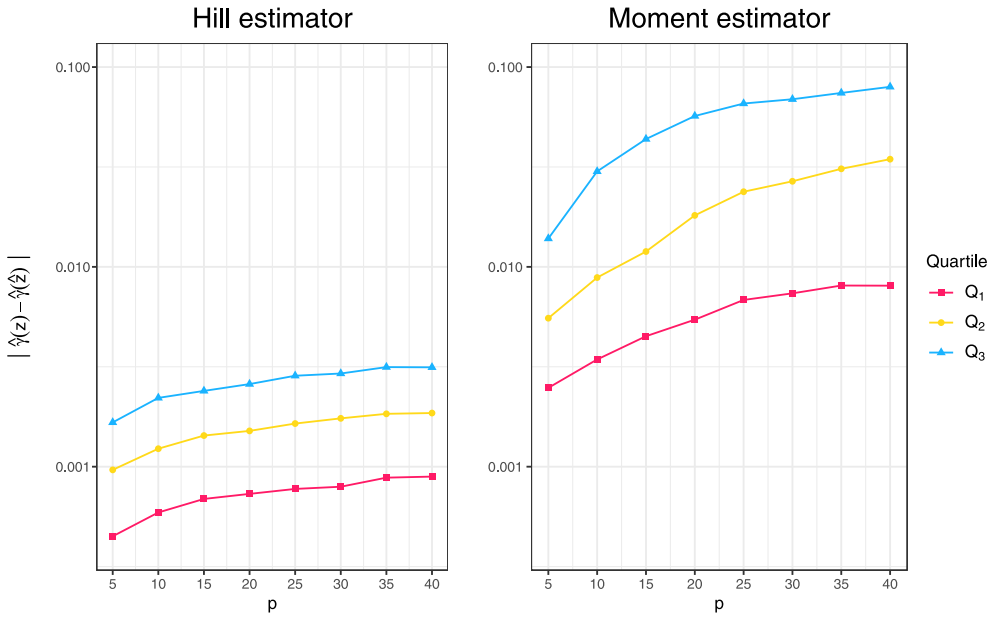

Fig. 3. The quartiles of $|\hat{\gamma}(|\mathbf{z}|) - \hat{\gamma}(|\hat{\mathbf{z}}|)|$, for the Hill estimator and the moment estimator as a function of the dimension p.

be avoided for two reasons: (i) On the surface, it is often very difficult to say whether or not a particular observed variable has a non-trivial contribution to the overall latent variable structure. In some sense, not having to make such choices (omitting variables) can even be seen as one of the main purposes of latent variable models; the estimated loadings will make the decisions for us. (ii) Due to the nature of ICA, if one or more of the p variables are fully independent of the others, such variables will form their individual independent components (with loading vectors proportional to the corresponding standard basis vectors). This means that the inclusion of "unrelated" variables in ICA has no effect (asymptotically) on the obtained results and their interpretation. And vice versa, if the omission of a variable (asymptotically) changes the results of ICA, then we know that the omitted variable was, in fact, related to the other variables. Hence, we next continue to estimate a latent variable model for the full observed series.

As a word of caution, we remark that the previous guidelines are posed strictly in the context of ICA (and blind source separation) models and may not hold in other latent variable models. For example, in a general factor model $\mathbf{x}_i = \boldsymbol{\mu} + L\mathbf{z}_i + \boldsymbol{\epsilon}_i$ where the p specific factors in $\boldsymbol{\epsilon}_i$ are allowed to be dependent of each other, dropping any observed variables is not advisable. This is because, even if none of the common factors $\mathbf{z}_i$ load onto a specific observed variable, say x_{ij}, then x_{ij} and $\mathbf{z}_i$ might still not be independent because of the dependency in the specific factors.

As the data are time-dependent, we use SOBI [1], a more fine-grained variant of the AMUSE-method described in Section 4.2, to estimate the latent variables using the lag set $\mathcal{T} = \{1, \dots, 12\}$. The main assumption when using SOBI is that the latent (and, hence, the observed) multivariate time series is covariance and autocovariance stationary (on the used lag set). In essence, this ensures that the sample (auto)covariance matrices used by SOBI estimate meaningful population quantities. From the plots of the data in Figs. C3 and C4 in the supplementary Appendix C, we see that most parts of the series are stable, but the few volatility spikes slightly violate the stationarity assumption. However, ICA and BSS methods, including SOBI, are known to be robust against moderate violations of the stationarity assumptions. For example, ICA and BBS are used in [4] to analyze fMRI signals, in [58] to estimate modes of damping vibratory systems, in [56] to detect outlying frames in a surveillance video, and, e.g., in [26, Section 7] to separate mixed audio signals. Despite each of these scenarios being non-stationary, the methods were shown to successfully tackle them and to identify practically meaningful latent variables. Motivated by this we next proceed with SOBI.

Denoting the original 22-variate series at time i by $\mathbf{x}_i$, the estimates of the centered latent series are given by $\hat{\mathbf{z}}_i = \hat{\Gamma}(\mathbf{x}_i - \bar{\mathbf{x}})$ where $\hat{\Gamma} \in \mathbb{R}^{22\times 22}$ is the unmixing matrix estimate given by SOBI. The estimated latent series are shown in Figs. C7 and C8 in the supplementary Appendix C, ordered in decreasing order w.r.t. their "importance" (sum of squared autocovariances). Inspecting the first few (hence, the most important) latent series shows that they each indeed correspond roughly to a single (or a few) time periods, exhibiting volatility during the given time period and staying relatively stable outside of it. For example, the first latent series seems to represent time periods in the beginning and the end of the measurement frame and the second to fourth latent series all correspond to the 2007–2008 period.

To get a clearer view, the bottom plot of Fig. 4 shows the extreme value index estimates of the absolute values of the five leading latent series, obtained using the same rolling window approach as in the top plot of Fig. 4. The most prominent feature in the plot is the steadily increasing risk in one of the latent series towards the end of the measurement period (the black curve in the bottom plot of Fig. 4). Interestingly, this behavior is not visible in the extreme value index estimates of the original series (top plot of Fig. 4).

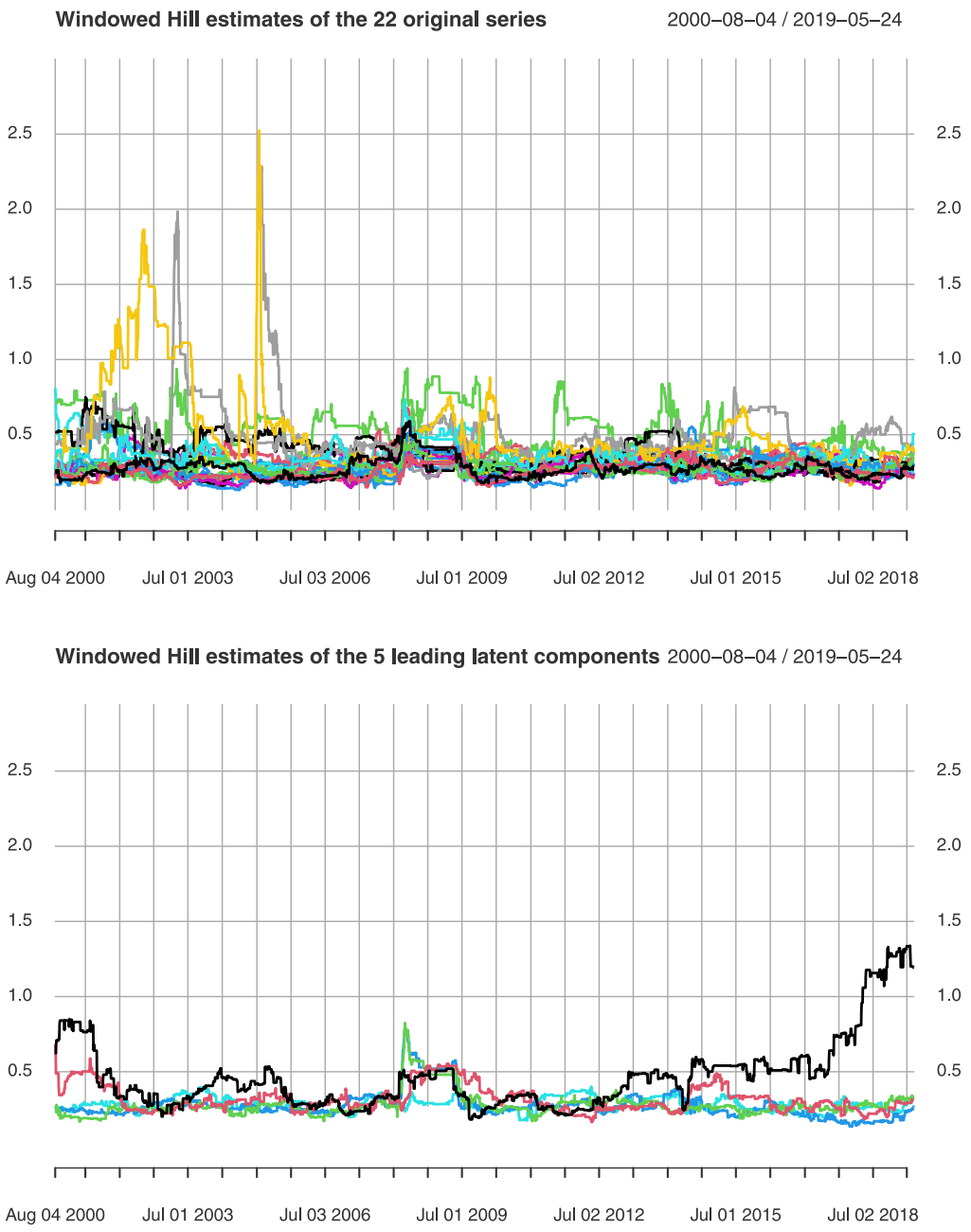

Fig. 4. The extreme value indices of the absolute values of the 22 observed log-return series $\mathbf{x}_i$ (top plot) and the five leading latent series (bottom plot), estimated with a rolling window of length 300 days. Hill estimator with the tail length $k_n = 30$ was used. The x-axis in the plot denotes the middle (150th) days of the windows.

Other standout features are the increased risk in several of the five leading latent series during both 2000–2001 and 2007–2008. The extreme value index estimates of all 22 latent series are shown individually in Figs. C9 and C10 in the supplementary Appendix C.

Finally, to study the connection between the latent factors and the observed series, we inspect the transformation matrix $\hat{\Gamma}$. Its elements provide the contributions, or loadings, of the original series to each of the latent series and a heatmap of it is given in Fig. 5. To improve readability, the heatmap shows only those elements of $\hat{\Gamma}$ which have absolute values greater than their 90th percentile. That is, the colored cells in Fig. 5 represent the 10% of the strongest effects among all loadings. The heatmap reveals, for example, the following: (i) The leading latent series (whose extreme value index series has the steady rise towards the end of 2019 in Fig. 4) is roughly a contrast between the exchange rates of EUR and DKK, leading us to conjecture that the component is related to the European debt crisis of 2009–2019. (ii) Two of the currencies, CNY and TWD, are such that they load (after the percentile

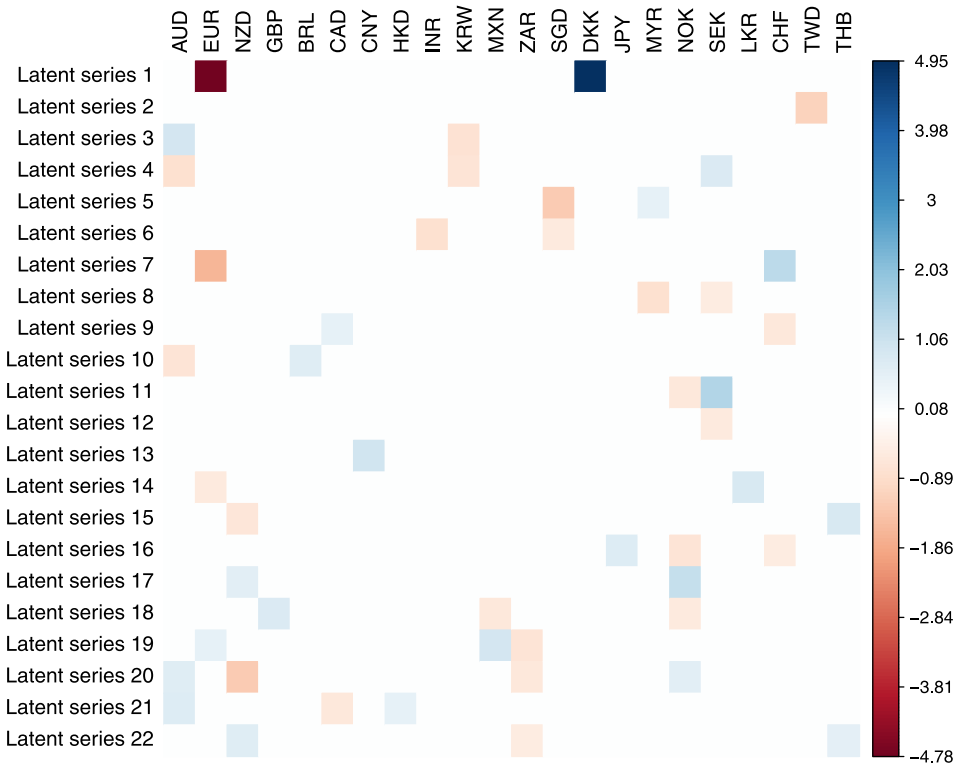

Fig. 5. Heatmap of the transformation matrix $\hat{\Gamma}$ containing the loadings of the original series (the columns) to the latent series (the rows). For clarity, all loadings with absolute values below the 90th percentile have been set to zero (indicated by the white color). (For interpretation of the references to color in this figure legend, the reader is referred to the web version of this article.)

thresholding) only to single latent series where no other currency loads. This implies, as described earlier, that the corresponding variables are more or less independent of the others and have little contribution on the general latent variable structure. Interestingly, both of these currencies also have a non-floating exchange rate, giving a partial explanation for this behavior.

We also further experimented on the data by (a) considering only a subset of the data set comprising the 4 most recent years, and (b) estimating the loadings using factor analysis with the varimax-rotation, instead of SOBI, as implemented in the function `fa` in the R-package `pscyh` [47]. In the interest of space, we only summarize these results here. The findings were: (i) Factor analysis was also able to find latent factors displaying high-risk behavior but the corresponding volatility spikes were markedly weaker than the ones found by SOBI. This is understandable as factor analysis targets different aspects of the data generating process than SOBI. (ii) The loadings of the leading factor for SOBI in the 4-year subset were very similar to their counterparts in the full data set, alleviating the problem of possible non-stationarity. The same was true also for factor analysis, whose first two factors approximately matched for the two time periods.

7. Conclusion

We studied the effect of a preliminary latent variable extraction on the estimation of the extreme value indices of the latent independent components. This approach to multivariate extreme value analysis is highly practical in the sense that it reduces the problem into several univariate extreme value problems, allowing the use of the standard extreme value machinery. Moreover, our asymptotic analysis revealed that, under reasonably mild conditions, the consistency and limiting normality of the Hill estimator and the moment estimator are preserved in this construction.

A natural question to pursue in the future is whether the conditions in Theorems 2 and 3 can be weakened (we only showed that they are sufficient). Some preliminary simulation (not shown here) indicates that this might indeed be the case. Moreover, the current work can likely be used to simplify the task of deriving similar results for other suitable estimators besides the Hill estimator and the moment estimator. This is because the perturbation bounds for tail observations given in Appendix A are not tied to any particular extreme value index estimator (indeed, they concern the latent variable estimation part of the model). As such, one only needs to derive for the new methods the analogues of the perturbation bounds for the actual extreme value index estimation step (as in the proof of Theorem 2).

In some applications the interest is more on the estimation of extreme quantiles (due to their easier interpretability) than on the extreme value index itself. However, as the limiting distributions of extreme value index estimators play a key role in the

asymptotics and estimation of extreme quantiles, see [10, Section 4], our theoretical results could possibly be extended to estimate also the extreme quantiles of the latent variables in the model (3), a task we will leave for future work.

Finally, we note that while Section 4 was written from the viewpoint of estimating the full mixing matrix Ω, not all of the produced latent components are usually interesting. Indeed, in applications it is often both practical and computationally efficient to assume that the dependency structure of the data is driven by a small set of latent variables. This case is often modeled in the ICA-literature by assuming that there are still a total of p latent variables, but that $p-d$ of them are pure noise (Gaussian or white noise) [48]. And, in fact, this paradigm can be incorporated into our existing theoretical framework, for example, as follows: Given a p-variate sample $\mathbf{x}_1, \dots, \mathbf{x}_n$ we use a method of sequential latent variable estimation to extract d-variate latent vectors $\hat{\mathbf{z}}_1^*, \dots, \hat{\mathbf{z}}_n^*$ where $d \ll p$. Examples of methods capable of this are deflation-based FastICA (for i.i.d. data) [27,41] and deflation-based SOBI (for time series data) [36]. These methods automatically extract the p most "interesting" (exhibiting highest non-Gaussianity and highest autocorrelation, respectively) latent components, leaving only the noise behind. In [36,41] it is shown that these methods produce $\sqrt{n}$-consistent estimates $\hat{\Gamma}^* \in \mathbb{R}^{d\times p}$ of a subset of rows of Ω^{-1}. Correspondingly, writing as in Section 4,

$$\hat{\mathbf{z}}_i^* = \hat{\Gamma}^* \left(\mathbf{x}_i - \bar{\mathbf{x}}\right) = \hat{\Gamma}^* \Omega \left(\mathbf{z}_i - \bar{\mathbf{z}}\right) = E_{d,p}\mathbf{z}_i + \left(\hat{\Gamma}\Omega - E_{d,p}\right)\mathbf{z}_i - \hat{\Gamma}^*\Omega\bar{\mathbf{z}},$$

where $E_{d,p} \in \mathbb{R}^{d\times p}$ contains the corresponding subset of rows of the $p \times p$ identity matrix. This shows that the vectors $\hat{\mathbf{z}}_i^*$ admit the decomposition (3) and all our main theoretical results apply component-wise to them. Such partial estimation can also be useful when some of the (non-interesting) latent variables are almost or fully non-identifiable, for example, through an almost singular mixing matrix or by having excess kurtosis close to 0 in ICA. In such cases, deflation-based methods can still estimate the identifiable components and our estimation guarantees for them are not affected by the non-identifiable components.

In Section 3 we demonstrated that the classical orthogonal factor model does not fall in the framework of the decomposition (3). However, some preliminary investigations show that results similar to Theorems 2 and 3 could possibly be derived for the OFM by assuming that the extreme value behaviors of the common factors $\mathbf{z}_i$ dominate those of the specific factors $\boldsymbol{\epsilon}_i$. Such results are, however, beyond the scope of the current work.

Finally, we remark that our main results in Section 3 are, in principle, extendable to high-dimensional scenarios where $p \equiv p_n$. Namely, we only ever use Eq. (3) on the level of individual components, meaning that if one can show that a subset of latent variables estimated with some high-dimensional methodology satisfies (3), then our main results continue to apply. However, establishing (3) for high-dimensional models is likely difficult since, even if it is posed on the level of individual components, the quantities $\hat{H}$ and $\hat{\mathbf{r}}$ can still depend on the full set of p_n variables, meaning that one needs to show that the desired convergence rates are not compromised by the growth of p_n. As such, we have left this for future work.

CRediT authorship contribution statement

Joni Virta: Methodology, Software, Writing – original draft, Writing – review & editing. **Niko Lietzén:** Methodology, Software, Writing – original draft, Writing – review & editing. **Lauri Viitasaari:** Conceptualization, Methodology, Supervision, Writing – original draft, Writing – review & editing. **Pauliina Ilmonen:** Conceptualization, Methodology, Supervision, Writing – original draft, Writing – review & editing.

Acknowledgments

All authors acknowledge the computational resources provided by the Aalto Science-IT project. Joni Virta gratefully acknowledges the financial support from the Academy of Finland, Finland (Grants 335077, 347501, 353769). Niko Lietzén gratefully acknowledges the financial support from the Emil Aaltonen Foundation, Finland (Grant 190135 N). Pauliina Ilmonen gratefully acknowledges support from the Academy of Finland, Finland, decision number 346308 (Centre of Excellence in Randomness and Structures). The authors are grateful to the Associate Editor and two anonymous Reviewers for their comments which helped greatly in improving the presentation and quality of the manuscript.

Appendix A. Proofs

We begin by deriving a collection of auxiliary lemmas. The overarching objective behind them is to establish the rate at which the quantity

$$||\hat{\mathbf{z}}^k|_{(n-m,n)}/|\mathbf{z}^k|_{(n-m,n)} - 1|$$

vanishes. We begin with the next result that allows, in the case of independent observations, us to consider the component with the heaviest tail as the conservative bound for the error. This translates, under independence, into $\max_l\{g_{nl}\}$ on our main theorems.

Lemma 1. *Let $F_0 \in D(G_{\gamma_0}), F_1 \in D(G_{\gamma_1}), F_2 \in D(G_{\gamma_2}), F_3 \in D(G_{\gamma_3})$ be distributions such that,*

$$\gamma_0 > \gamma_1 > \gamma_2 = 0 > \gamma_3.$$

For $k \in \{0,1,2,3\}$, put $g_{nk} = \max\{a_{nk}, b_{nk}\}$, where a_{nk}, b_{nk} are the normalizing sequences such that $\frac{y^k_{(n,n)} - b_{nk}}{a_{nk}} \rightsquigarrow G_\gamma$, where y^k follows F_k. Then

$$\frac{g_{nk}}{g_{n0}} = o(1), \quad k \in \{1,2,3\}.$$

Proof. Note first that since $\gamma_0 > 0$, the distribution F_0 is heavy tailed and belongs to the domain of attraction of the Fréchet distribution. Thus, by [20, Section 3.4], $g_{n0} = a_{n0} = n^{\gamma_0} L_0(n)$ where L_0 is a slowly varying function. Similarly $g_{n1} = n^{\gamma_1} L_1(n)$ for some slowly varying function L_1 and we have the claim for the value $k = 1$, that is,

$$\frac{g_{n1}}{g_{n0}} = n^{\gamma_1 - \gamma_0} \frac{L_1(n)}{L_0(n)} = o(1).$$

Similarly, the distribution F_3 is light tailed and belongs to the domain of attraction of the Weibull distribution. As such, by [20, Section 3.4], we have $g_{n3} = \max\{n^{\gamma_3} L_3(n), d\}$ for some slowly-varying function L_3 and constant d. Since $\gamma_3 < 0$, we have $g_{n3} = \mathcal{O}(1)$ and the claim for $k = 3$ follows from

$$\frac{g_{n3}}{g_{n0}} = \frac{\mathcal{O}(1)}{n^{\gamma_0} L_0(n)} = o(1).$$

It remains to prove the case $k = 2$ that corresponds to the border case $\gamma_2 = 0$. Now F_2 belongs to the domain of attraction of the Gumbel distribution and, by [20, Section 3.4], we have $g_{n2} = \max\{a(b_n), b_n\}$, where $a(b_n)$ is as in [20, Definition 3.3.18], $b_n = F_2^{\leftarrow}(1 - 1/n)$ and $F_2^{\leftarrow}$ is the quantile function. Let $y_F \leq \infty$ be the right endpoint of the distribution F_2. We consider two cases, $y_F < \infty$ and $y_F = \infty$, separately. In the former, $b_n \to y_F$ as $n \to \infty$ and by [20, Remark 2, Section 3.3] $a(b_n) \to 0$ as $n \to \infty$. Thus, for a large enough n, we have $g_{n2} = b_n \to y_F < \infty$ and

$$\frac{g_{n2}}{g_{n0}} = \frac{y_F + o(1)}{n^{\gamma_0} L_0(n)} = o(1).$$

For $y_F = \infty$, we have $b_n \to \infty$ and, by [20, Remark 1, Section 3.3], $a(b_n) = o(b_n)$. Thus, for a large enough n, we have $g_{n2} = b_n$ and

$$\frac{g_{n2}}{g_{n0}} = \frac{F_2^{\leftarrow}(1 - 1/n)}{n^{\gamma_0} L_0(n)}.$$

We continue by proof by contradiction, and assume that $F_2^{\leftarrow}(1 - 1/n)/\{n^{\gamma_0} L_0(n)\}$ does not converge to zero. Then there exists $\epsilon_0 > 0$ such that we can find an arbitrarily large n such that $F_2^{\leftarrow}(1 - 1/n) \geq \epsilon_0 n^{\gamma_0} L_0(n)$. It follows that $1 - F_2\left(\epsilon_0 n^{\gamma_0} L_0(n)\right) \geq 1/n$ and since L_0 is slowly varying, this further implies that $1 - F_2\left(c n^{\gamma_0}\right) \geq 1/n$ for some constant $c > 0$ and a large enough n. Since $\gamma_0 > 0$, this implies that F_2 is heavy tailed giving us the contradiction. This completes the proof for the case $k = 2$ as well. □

The next result shows that the denominator in $||\hat{\mathbf{z}}^k|_{(n-m,n)}/|\mathbf{z}^k|_{(n-m,n)} - 1|$ is negligible.

Lemma 2. *Let $(z_k), k \in \{1, \dots, n\}$ be an arbitrary sequence of non-negative random variables such that*

$$\lim_{\delta \to 0} \inf_{k \geq 1} \Pr\left(z_k \geq \delta\right) = 1. \tag{A.1}$$

Then, for any $\epsilon > 0$ and any intermediate sequence k_n, there exists $\delta > 0$ and N such that

$$\Pr(z_{(n-k_n,n)} < \delta) < \epsilon, \quad n \geq N.$$

Proof. Let

$$S_n(\delta) = \sum_{k=1}^{n} \mathbf{1}_{z_k \geq \delta}.$$

Then

$$\Pr(z_{(n-k_n,n)} < \delta) = \Pr(S_n(\delta) < k_n).$$

Indeed, $S_n(\delta) < k_n$ means that less than k_n of the values are above or equal to δ, which implies that k_n:th maximum of z is strictly less than δ. Vice versa, if $z_{(n-k_n,n)} < \delta$, then at most $k_n - 1$ of values z_k can be above or equal to δ. Thus it suffices to prove that for any $\epsilon > 0$, we can find N and δ such that for $n \geq N$ we have $\Pr(S_n(\delta) < k_n) < \epsilon$. Equivalently, we need to show

$$\Pr(S_n(\delta) \geq k_n) > 1 - \epsilon. \tag{A.2}$$

Denote $\overline{S}_n(\delta) = S_n(\delta)/n$. By (A.1), for any $\widetilde{\epsilon} > 0$ we can find $\delta > 0$ small enough such that

$$\mathrm{E}\overline{S}_n(\delta) = \frac{1}{n} \sum_{k=1}^{n} \Pr(z_k \geq \delta) > 1 - \widetilde{\epsilon} \tag{A.3}$$

uniformly in n. Together with $\overline{S}_n(\delta) \leq 1$ this gives us

$$\frac{\left\{\mathrm{E}\overline{S}_n(\delta)\right\}^2}{\mathrm{E}\overline{S}_n^2(\delta)} \geq (1 - \widetilde{\epsilon})^2$$

which holds for every n. Next we recall the Paley–Zygmund inequality which states that, for any random variable $Z \geq 0$ with finite variance and any number $\theta \in [0, 1]$, we have

$$\Pr\left(Z \geq \theta \mathrm{E}Z\right) \geq (1 - \theta)^2 \frac{(\mathrm{E}Z)^2}{\mathrm{E}Z^2}. \tag{A.4}$$

Since k_n is an intermediate sequence, we have, applying (A.3), that

$$\frac{k_n}{n\mathrm{E}\overline{S}_n(\delta)} \leq \frac{k_n}{n(1-\widetilde{\epsilon})} \leq 1$$

provided that n is large enough. Hence we may apply (A.4) with $Z = \overline{S}_n(\delta)$ and $\theta = k_n/\{n\mathrm{E}\overline{S}_n(\delta)\}$ to compute

$$\Pr(S_n(\delta) \geq k_n) = \Pr\left\{\overline{S}_n(\delta) \geq \frac{k_n}{n\mathrm{E}\overline{S}_n(\delta)}\mathrm{E}\overline{S}_n(\delta)\right\} \geq \left\{1 - \frac{k_n}{n\mathrm{E}\overline{S}_n(\delta)}\right\}^2 \frac{\left\{\mathrm{E}\overline{S}_n(\delta)\right\}^2}{\mathrm{E}\overline{S}_n^2(\delta)} \geq \left\{1 - \frac{k_n}{n(1-\widetilde{\epsilon})}\right\}^2 (1-\widetilde{\epsilon})^2.$$

This implies (A.2), since $\widetilde{\epsilon} > 0$ can be chosen arbitrarily and independently of n. This concludes the proof. □

The next two results allow us to deduce bounds for the difference between order statistics of $|\hat{\mathbf{z}}^k|$ and $|\mathbf{z}^k|$.

Lemma 3. *Let $\mathbf{a} = (a_1, \dots, a_n)$ and $\mathbf{b} = (b_1, \dots, b_n)$ satisfy $a_i \leq b_i$, for all $i \in \{1, \dots, n\}$. Then $(\mathbf{a})_{(k,n)} \leq (\mathbf{b})_{(k,n)}$, for all $k \in \{1, \dots, n\}$.*

Proof. By the definition, we have $b_i \leq (\mathbf{b})_{(k,n)}$ for k different indices i. Since for all such i we have $a_i \leq b_i$, the claim follows directly. □

Lemma 4. *Let $\mathbf{x} = (x_1, \dots, x_n)$ and $\boldsymbol{\epsilon} = (\epsilon_1, \dots, \epsilon_n)$ be arbitrary. Then for all $k \in \{1, \dots, n\}$, $\left||\mathbf{x}+\boldsymbol{\epsilon}|_{(k,n)} - |\mathbf{x}|_{(k,n)}\right| \leq |\boldsymbol{\epsilon}|_{(n,n)}$, where for a vector $\mathbf{a} = (a_1, \dots, a_n)$ the notation $|\mathbf{a}| \in \mathbb{R}^n$ refers to the vector of the element-wise absolute values of $\mathbf{a}$.*

Proof. Recall Weyl's inequality: if $R, S \in \mathbb{R}^{n\times n}$ are symmetric matrices and $\lambda_j(R)$ denotes the jth largest eigenvalue of the matrix R, $j \in \{1, \dots, n\}$, then

$$\lambda_{j+m}(R) + \lambda_{k-m}(S) \leq \lambda_j(R+S) \leq \lambda_{j-\ell}(R) + \lambda_{1+\ell}(S).$$

for all $\ell \in \{0, \dots, j-1\}$, $m \in \{0, \dots, k-j\}$, see [24].

Let $\mathrm{diag}(\mathbf{r}) \in \mathbb{R}^{n\times n}$ denote the diagonal matrix having the elements of the vector $\mathbf{r} = (r_1, \dots, r_n)$ as its diagonal elements. Then $(\mathbf{r})_{(k,n)} = \lambda_{n-k+1}[\mathrm{diag}(\mathbf{r})]$ and the right-hand side of Weyl's inequality with $j = n-k+1$ and $\ell = 0$ gives,

$$(\mathbf{r}+\mathbf{s})_{(k,n)} \leq (\mathbf{r})_{(k,n)} + (\mathbf{s})_{(n,n)}, \tag{A.5}$$

for any two vectors $\mathbf{r} = (r_1, \dots, r_n)$ and $\mathbf{s} = (s_1, \dots, s_n)$.

Eq. (A.5) with $\mathbf{r} = |\mathbf{x}|$ and $\mathbf{s} = |\boldsymbol{\epsilon}|$ in conjunction with the triangle inequality, $|x_i + \epsilon_i| \leq |x_i| + |\epsilon_i|$, and Lemma 3 allow us to estimate,

$$|\mathbf{x}+\boldsymbol{\epsilon}|_{(k,n)} - |\mathbf{x}|_{(k,n)} \leq (|\mathbf{x}| + |\boldsymbol{\epsilon}|)_{(k,n)} - |\mathbf{x}|_{(k,n)} \leq |\boldsymbol{\epsilon}|_{(n,n)},$$

giving the first half of the inequality. For the other half, we have by the same set of inequalities and the expansion $\mathbf{x} = \mathbf{x} + \boldsymbol{\epsilon} - \boldsymbol{\epsilon}$,

$$|\mathbf{x}|_{(k,n)} - |\mathbf{x}+\boldsymbol{\epsilon}|_{(k,n)} \leq (|\mathbf{x}+\boldsymbol{\epsilon}| + |\boldsymbol{\epsilon}|)_{(k,n)} - |\mathbf{x}+\boldsymbol{\epsilon}|_{(k,n)} \leq |\boldsymbol{\epsilon}|_{(n,n)},$$

where the second inequality is obtained by applying (A.5) to $\mathbf{r} = |\mathbf{x}+\boldsymbol{\epsilon}|$ and $\mathbf{s} = |\boldsymbol{\epsilon}|$. □

Combining previous results yields the following lemma that provides a crucial estimate for the proofs of our main theorems.

Lemma 5. *Let $k \in \{1, \dots, p\}$ be fixed. Then, under (3) and Assumption 1, we have*

$$\max_{0\leq m\leq k_n}\left|\frac{|\hat{\mathbf{z}}^k|_{(n-m,n)}}{|\mathbf{z}^k|_{(n-m,n)}} - 1\right| = \mathcal{O}_p\left(\frac{1}{c_n}\max_{\ell}\{g_{n\ell}\}\right).$$

Proof. The left-hand side of the claim equals

$$\max_{0\leq m\leq k_n}\left|\frac{|\hat{\mathbf{z}}^k|_{(n-m,n)} - |\mathbf{z}^k|_{(n-m,n)}}{|\mathbf{z}^k|_{(n-m,n)}}\right|, \tag{A.6}$$

where by (3) and Lemma 4 the numerator can be bounded by

$$\left||\hat{\mathbf{z}}^k|_{(n-m,n)} - |\mathbf{z}^k|_{(n-m,n)}\right| \leq \max_i\left\{\left|\sum_{j=1}^p \hat{h}_{kj}z_{ij} + \hat{r}_k\right|\right\} \leq \sum_{j=1}^p |\hat{h}_{kj}||\mathbf{z}^j|_{(n,n)} + |\hat{r}_k|,$$

where $\hat{h}_{kj} = \mathcal{O}_p(c_n^{-1})$, $j \in \{1, \dots, p\}$, and $\hat{r}_k = \mathcal{O}_p(c_n^{-1})$. Now, by Assumption 1, we have

$$\begin{aligned}\sum_{j=1}^p |\hat{h}_{kj}||\mathbf{z}^j|_{(n,n)} + |\hat{r}_k| &= \sum_{j=1}^p\left\{a_{nj}|\hat{h}_{kj}|\frac{|\mathbf{z}^j|_{(n,n)} - b_{nj}}{a_{nj}} + |\hat{h}_{kj}|b_{nj}\right\} + |\hat{r}_k| = \sum_{j=1}^p\left\{\frac{a_{nj}}{c_n}\mathcal{O}_p(1) + \frac{b_{nj}}{c_n}\mathcal{O}_p(1)\right\} + \mathcal{O}_p\left(\frac{1}{c_n}\right)\\ &= \sum_{j=1}^p \mathcal{O}_p\left(\frac{g_{nj}}{c_n}\right) + \mathcal{O}_p\left(\frac{1}{c_n}\right) = \mathcal{O}_p\left(\frac{1}{c_n}\max_{\ell}\{g_{n\ell}\}\right),\end{aligned}$$

where we have used the result that if one deterministic sequence eventually majorizes another, $r_n \leq s_n$, for all $n \geq N$, then any sequence of random variables x_n with $x_n = \mathcal{O}_p(r_n)$ has also $x_n = \mathcal{O}_p(s_n)$.

The previous bound holds uniformly in m. Thus

$$\max_{0\leq m\leq k_n}\left|\frac{|\hat{\mathbf{z}}^k|_{(n-m,n)}-|\mathbf{z}^k|_{(n-m,n)}}{|\mathbf{z}^k|_{(n-m,n)}}\right| = \mathcal{O}_p\left(\frac{1}{c_n}\max_{\ell}\{g_{n\ell}\}\right)\max_{0\leq m\leq k_n}\frac{1}{|\mathbf{z}^k|_{(n-m,n)}},$$

where $\max_{0\leq m\leq k_n}|\mathbf{z}^k|^{-1}_{(n-m,n)} = |\mathbf{z}^k|^{-1}_{(n-k_n,n)}$ is, by Lemma 2, of order $\mathcal{O}_p(1)$. This concludes the proof. □

Remark 4. In the case of independent observations, we have $|\mathbf{z}^k|^{-1}_{(n-k_n,n)} \sim \left[F_k^{\leftarrow}\left(1-\frac{k_n}{n}\right)\right]^{-1}$ which would yield slightly less restrictive condition. With our more conservative bound however, we can cover relatively arbitrary dependence structures.

Finally, we show the following result allowing us to handle logarithm in the estimators.

Lemma 6. *Let x_n be an arbitrary triangular array of random variables satisfying $\max_{0\leq m\leq d_n}|x_m| = \mathcal{O}_p(e_n)$ for some d_n and $e_n = o(1)$. Furthermore, let $g : (a,b) \mapsto \mathbb{R}$ with $-\infty \leq a < 0 < b \leq \infty$ be such that g is continuously differentiable at the neighborhood of 0. Then*

$$\max_{0\leq m\leq d_n}|g(x_m)-g(0)| = \mathcal{O}_p(e_n).$$

Proof. Let $\epsilon > 0$ be fixed. Then there exists $C > 0$ and N such that

$$\Pr\left(\frac{\max_{0\leq m\leq d_n}|x_m|}{e_n} > C\right) < \frac{\epsilon}{2}$$

for $n \geq N$. By assumptions, there exists $\delta > 0$ such that g is continuously differentiable on an open interval $(-\delta,\delta)$. Moreover, by continuity of g' we also have

$$\left(g'\right)^* = \sup_{-\frac{\delta}{2}\leq x\leq\frac{\delta}{2}}|g'(x)| < \infty.$$

Moreover, since $e_n = o(1)$ there exists N^* such that $e_nC \leq \delta/2$ for $n \geq N^*$. Thus, on the set $A_n = \{\max_{0\leq m\leq d_n}|x_m| \leq e_nC\}$ mean value theorem implies

$$\max_{0\leq m\leq d_n}|g(x_m)-g(0)| \leq \left(g'\right)^*\max_{0\leq m\leq d_n}|x_m|.$$

Let $n \geq \max(N,N^*)$ and put $\widetilde{C} = \left(g'\right)^* C$. We have

$$\begin{aligned}\Pr\left(\frac{\max_{0\leq m\leq d_n}|g(x_m)-g(0)|}{e_n} > \widetilde{C}\right) &= \Pr\left(A_n, \frac{\max_{0\leq m\leq d_n}|g(x_m)-g(0)|}{e_n} > \widetilde{C}\right) + \Pr\left(A_n^c, \frac{\max_{0\leq m\leq d_n}|g(x_m)-g(0)|}{e_n} > \widetilde{C}\right)\\ &\leq \Pr\left(A_n, \frac{\left(g'\right)^*\max_{0\leq m\leq d_n}|x_m|}{e_n} > \widetilde{C}\right) + \Pr\left(A_n^c\right) \leq \Pr\left(\frac{\max_{0\leq m\leq d_n}|x_m|}{e_n} > C\right) + \Pr\left(\frac{\max_{0\leq m\leq d_n}|x_m|}{e_n} > C\right) < \epsilon\end{aligned}$$

concluding the proof. □

We are now ready to prove Theorems 2 and 3, beginning with the former.

Proof of Theorem 2. Let $\mathbf{y} = (y_1,\dots,y_n) \geq 0$ and $\hat{\mathbf{y}} = (\hat{y}_1,\dots,\hat{y}_n) \geq 0$ be an arbitrary pair of samples that satisfy

$$\max_{0\leq m\leq k_n}\left|\frac{(\hat{\mathbf{y}})_{(n-m,n)}}{(\mathbf{y})_{(n-m,n)}}-1\right| = \mathcal{O}_p\left(h_n\right), \tag{A.7}$$

where $h_n = o(1)$.

Recall that the Hill estimator is given by

$$\hat{\gamma}_H(\mathbf{y}) = M_n^{(1)}(\mathbf{y}) = \frac{1}{k_n}\sum_{m=0}^{k_n-1}\ln\frac{(\mathbf{y})_{(n-m,n)}}{(\mathbf{y})_{(n-k_n,n)}},$$

where $k_n/n \to 0$, $k_n \to \infty$. In the proof, we use the short notation

$$\hat{w}_m := \frac{(\hat{\mathbf{y}})_{(n-m,n)}}{(\mathbf{y})_{(n-m,n)}} - 1.$$

We now have

$$\begin{aligned}\left|M_n^{(1)}(\hat{\mathbf{y}}) - M_n^{(1)}(\mathbf{y})\right| &= \left|\frac{1}{k_n}\sum_{m=0}^{k_n-1}\left\{\ln\frac{(\hat{\mathbf{y}})_{(n-m,n)}}{(\hat{\mathbf{y}})_{(n-k_n,n)}} - \ln\frac{(\mathbf{y})_{(n-m,n)}}{(\mathbf{y})_{(n-k_n,n)}}\right\}\right| = \left|\frac{1}{k_n}\sum_{m=0}^{k_n-1}\left[\ln(1+\hat{w}_m) - \ln(1+\hat{w}_{k_n})\right]\right|\\ &\leq \frac{1}{k_n}\sum_{m=0}^{k_n-1}\left|\ln(1+\hat{w}_m)\right| + \left|\ln(1+\hat{w}_{k_n})\right| \leq 2\max_{0\leq m\leq k_n}\left|\ln(1+\hat{w}_m)\right|.\end{aligned}$$

The assumptions of Lemma 6 are now satisfied for $x_n = \hat{w}_n$, $d_n = k_n$, $e_n = h_n$ and $g(x) = \ln(1+x)$, implying that $|M_n^{(1)}(\hat{\mathbf{y}}) - M_n^{(1)}(\mathbf{y})| = \mathcal{O}_p(h_n)$. Plugging in $\mathbf{y} = \mathbf{z}^k$ and $\hat{\mathbf{y}} = \hat{\mathbf{z}}^k$, and using Lemma 5, now give the convergence of the Hill estimator. For the moment estimator, recall that

$$\hat{\gamma}_M(\mathbf{y}) = M_n^{(1)}(\mathbf{y}) + 1 - \frac{1}{2}\left(1 - \frac{\{M_n^{(1)}(\mathbf{y})\}^2}{M_n^{(2)}(\mathbf{y})}\right)^{-1}.$$

By the first part of the proof, we have

$$|M_n^{(1)}(\hat{\mathbf{y}}) - M_n^{(1)}(\mathbf{y})| = \mathcal{O}_p(h_n). \tag{A.8}$$

It thus suffices to prove that

$$\left|\frac{\{M_n^{(1)}(\mathbf{y})\}^2}{M_n^{(2)}(\mathbf{y})} - \frac{\{M_n^{(1)}(\hat{\mathbf{y}})\}^2}{M_n^{(2)}(\hat{\mathbf{y}})}\right| = \mathcal{O}_p\left(\frac{h_n}{\hat{\gamma}_H(\mathbf{y})}\right). \tag{A.9}$$

Indeed, since $M_n^{(1)}(\mathbf{y}) = \hat{\gamma}_H(\mathbf{y})$ as a convergent sequence is uniformly tight, i.e., $\mathcal{O}_p(1)$, it follows from the convergence of $\hat{\gamma}_M(\mathbf{y})$ that

$$\left(1 - \frac{\{M_n^{(1)}(\mathbf{y})\}^2}{M_n^{(2)}(\mathbf{y})}\right)^{-1} = \mathcal{O}_p(1).$$

Then (A.9) together with the assumption $h_n/\hat{\gamma}_H(\mathbf{y}) \to_p 0$ implies that also

$$\left(1 - \frac{\{M_n^{(1)}(\hat{\mathbf{y}})\}^2}{M_n^{(2)}(\hat{\mathbf{y}})}\right)^{-1} = \mathcal{O}_p(1).$$

The claim then follows by using

$$(1-a)^{-1} - (1-b)^{-1} = \frac{a-b}{1-a}(1-b)^{-1}, \quad a, b \in (0,1)$$

with $a = \{M_n^{(1)}(\mathbf{y})\}^2/M_n^{(2)}(\mathbf{y})$ and $b = \{M_n^{(1)}(\hat{\mathbf{y}})\}^2/M_n^{(2)}(\hat{\mathbf{y}})$, leading to

$$\left|\hat{\gamma}_M(\hat{\mathbf{y}}) - \hat{\gamma}_M(\mathbf{y})\right| = \mathcal{O}_p\left(\frac{h_n}{\hat{\gamma}_H(\mathbf{y})}\right). \tag{A.10}$$

In order to prove (A.9) we write

$$\begin{aligned}\left|\frac{\{M_n^{(1)}(\mathbf{y})\}^2}{M_n^{(2)}(\mathbf{y})} - \frac{\{M_n^{(1)}(\hat{\mathbf{y}})\}^2}{M_n^{(2)}(\hat{\mathbf{y}})}\right| &\le \frac{1}{M_n^{(2)}(\mathbf{y})}\left|\{M_n^{(1)}(\mathbf{y})\}^2 - \{M_n^{(1)}(\hat{\mathbf{y}})\}^2\right| + \frac{\{M_n^{(1)}(\hat{\mathbf{y}})\}^2}{M_n^{(2)}(\hat{\mathbf{y}})M_n^{(2)}(\mathbf{y})}\left|M_n^{(2)}(\mathbf{y}) - M_n^{(2)}(\hat{\mathbf{y}})\right| \\ &=: I_1(n) + I_2(n).\end{aligned}$$

For the first term $I_1(n)$, we use $a^2 - b^2 = (a-b)(a+b)$ and (A.8) to get

$$\begin{aligned}\left|\{M_n^{(1)}(\mathbf{y})\}^2 - \{M_n^{(1)}(\hat{\mathbf{y}})\}^2\right| &= \left|M_n^{(1)}(\mathbf{y}) - M_n^{(1)}(\hat{\mathbf{y}})\right|\left|M_n^{(1)}(\mathbf{y}) + M_n^{(1)}(\hat{\mathbf{y}})\right| \le \left|M_n^{(1)}(\mathbf{y}) - M_n^{(1)}(\hat{\mathbf{y}})\right|^2 + 2\left|M_n^{(1)}(\mathbf{y}) - M_n^{(1)}(\hat{\mathbf{y}})\right| M_n^{(1)}(\mathbf{y}) \\ &= \mathcal{O}_p\left(h_n M_n^{(1)}(\mathbf{y})\right).\end{aligned}$$

Here we used also the fact that $h_n/M_n^{(1)}(\mathbf{y}) \to_p 0$. Moreover, by Cauchy–Schwarz inequality we have $\{M_n^{(1)}(\mathbf{y})\}^2 \le M_n^{(2)}(\mathbf{y})$. Thus we can estimate

$$I_1(n) = \frac{1}{M_n^{(2)}(\mathbf{y})} \le \frac{\left|\{M_n^{(1)}(\mathbf{y})\}^2 - \{M_n^{(1)}(\hat{\mathbf{y}})\}^2\right|}{[M_n^{(1)}(\mathbf{y})]^2}\mathcal{O}_p\left(h_n M_n^{(1)}(\mathbf{y})\right) = \mathcal{O}_p\left(\frac{h_n}{M_n^{(1)}(\mathbf{y})}\right)$$

which, by recalling that $\hat{\gamma}_H(\mathbf{y}) = M_n^{(1)}(\mathbf{y})$, gives the claim for the term $I_1(n)$. For the term $I_2(n)$, we apply $a^2 - b^2 = (a-b)(a+b)$ again yielding

$$\begin{aligned}\left|M_n^{(2)}(\hat{\mathbf{y}}) - M_n^{(2)}(\mathbf{y})\right| &= \left|\frac{1}{k_n}\sum_{m=0}^{k_n-1}\left[\left\{\ln\frac{(\hat{\mathbf{y}})_{(n-m,n)}}{(\hat{\mathbf{y}})_{(n-k_n,n)}}\right\}^2 - \left\{\ln\frac{(\mathbf{y})_{(n-m,n)}}{(\mathbf{y})_{(n-k_n,n)}}\right\}^2\right]\right| \\ &\le \frac{1}{k_n}\sum_{m=0}^{k_n-1}\left|\ln(1+\hat{w}_m) - \ln(1+\hat{w}_{k_n})\right|\left|\ln\frac{(\hat{\mathbf{y}})_{(n-m,n)}}{(\hat{\mathbf{y}})_{(n-k_n,n)}} + \ln\frac{(\mathbf{y})_{(n-m,n)}}{(\mathbf{y})_{(n-k_n,n)}}\right| \\ &\le \frac{2\max_{0\le m\le k_n}|\ln(1+\hat{w}_m)|}{k_n}\sum_{m=0}^{k_n-1}\left|\ln\frac{(\hat{\mathbf{y}})_{(n-m,n)}}{(\hat{\mathbf{y}})_{(n-k_n,n)}} + \ln\frac{(\mathbf{y})_{(n-m,n)}}{(\mathbf{y})_{(n-k_n,n)}}\right|.\end{aligned}$$

Here

$$\frac{1}{k_n}\sum_{m=0}^{k_n-1}\left|\ln\frac{(\hat{\mathbf{y}})_{(n-m,n)}}{(\hat{\mathbf{y}})_{(n-k_n,n)}} + \ln\frac{(\mathbf{y})_{(n-m,n)}}{(\mathbf{y})_{(n-k_n,n)}}\right| \le \frac{1}{k_n}\sum_{m=0}^{k_n-1}\left|\ln\frac{(\hat{\mathbf{y}})_{(n-m,n)}}{(\hat{\mathbf{y}})_{(n-k_n,n)}} - \ln\frac{(\mathbf{y})_{(n-m,n)}}{(\mathbf{y})_{(n-k_n,n)}}\right| + \frac{2}{k_n}\sum_{m=0}^{k_n-1}\ln\frac{(\mathbf{y})_{(n-m,n)}}{(\mathbf{y})_{(n-k_n,n)}}$$

$$\leq 2 \max_{0\leq m\leq k_n} \left|\ln(1+\hat{w}_m)\right| + 2M_n^{(1)}(\mathbf{y}).$$

Together with Lemma 6 this gives us

$$\left|M_n^{(2)}(\hat{\mathbf{y}}) - M_n^{(2)}(\mathbf{y})\right| \leq \mathcal{O}_p\left(h_n M_n^{(1)}(\mathbf{y})\right).$$

Applying Cauchy–Schwarz again to get $\{M_n^{(1)}(\hat{\mathbf{y}})\}^2 \leq M_n^{(2)}(\hat{\mathbf{y}})$ gives us

$$I_2(n) = \frac{\{M_n^{(1)}(\hat{\mathbf{y}})\}^2}{M_n^{(2)}(\hat{\mathbf{y}})M_n^{(2)}(\mathbf{y})}\left|M_n^{(2)}(\mathbf{y}) - M_n^{(2)}(\hat{\mathbf{y}})\right| \leq \frac{1}{\{M_n^{(1)}(\mathbf{y})\}^2}\left|M_n^{(2)}(\mathbf{y}) - M_n^{(2)}(\hat{\mathbf{y}})\right| = \mathcal{O}_p\left(\frac{h_n}{M_n^{(1)}(\mathbf{y})}\right).$$

Plugging in $\mathbf{y} = \mathbf{z}^k$ and $\hat{\mathbf{y}} = \hat{\mathbf{z}}^k$, and using Lemma 5, now give the convergence of the moment estimator. This completes the proof. □

Applying the above computations, the proof of Theorem 3 is now quite simple.

Proof of Theorem 3. We write

$$\sqrt{k_n}\left\{\hat{\gamma}_H(|\hat{\mathbf{z}}^1|) - C_H\right\} = \sqrt{k_n}\left\{\hat{\gamma}_H(|\hat{\mathbf{z}}^1|) - \hat{\gamma}_H(|\mathbf{z}^1|)\right\} + \sqrt{k_n}\left\{\hat{\gamma}_H(|\mathbf{z}^1|) - C_H\right\}.$$

The first claim now follows directly from (A.8). Similarly, the second claim follows directly from

$$\sqrt{k_n}\left\{\hat{\gamma}_M(|\hat{\mathbf{z}}^1|) - C_M\right\} = \sqrt{k_n}\left\{\hat{\gamma}_M(|\hat{\mathbf{z}}^1|) - \hat{\gamma}_M(|\mathbf{z}^1|)\right\} + \sqrt{k_n}\left\{\hat{\gamma}_M(|\mathbf{z}^1|) - C_M\right\}$$

together with (A.10). □

Appendix B. Supplementary data

Supplementary material related to this article can be found online at https://doi.org/10.1016/j.jmva.2024.105300.

References

[1] A. Belouchrani, K. Abed-Meraim, J.-F. Cardoso, E. Moulines, A blind source separation technique using second-order statistics, IEEE Trans. Signal Process. 45 (2) (1997) 434–444.
[2] R. Bradley, Basic properties of strong mixing conditions. A survey and some open questions, Probab. Surv. 2 (2005) 107–144.
[3] J.-J. Cai, J.H. Einmahl, L. De Haan, Estimation of extreme risk regions under multivariate regular variation, Ann. Statist. 39 (3) (2011) 1803–1826.
[4] V.D. Calhoun, T. Adali, G. Pearlson, J.J. Pekar, Spatial and temporal independent component analysis of functional MRI data containing a pair of task-related waveforms, Hum. Brain Mapp. 13 (1) (2001) 43–53.
[5] J.-F. Cardoso, Source separation using higher order moments, in: 1989 International Conference on Acoustics, Speech, and Signal Processing, IEEE, 1989, pp. 2109–2112.
[6] J.-F. Cardoso, A. Souloumiac, Blind beamforming for non-Gaussian signals, in: IEE Proceedings F - Radar and Signal Processing, vol. 140, (6) IET, 1993, pp. 362–370.
[7] N.H. Chan, S.-J. Deng, L. Peng, Z. Xia, Interval estimation of value-at-risk based on GARCH models with heavy-tailed innovations, J. Econometrics 137 (2) (2007) 556–576.
[8] S. Choi, A. Cichocki, Blind separation of nonstationary sources in noisy mixtures, Electron. Lett. 36 (9) (2000) 848–849.
[9] P. Comon, C. Jutten, Handbook of Blind Source Separation: Independent Component Analysis and Applications, Academic Press, New York, 2010.
[10] L. De Haan, A. Ferreira, Extreme Value Theory: An Introduction, Springer Science & Business Media, New York, 2006.
[11] L. De Haan, C. Mercadier, C. Zhou, Adapting extreme value statistics to financial time series: dealing with bias and serial dependence, Finance Stoch. 20 (2) (2016) 321–354.
[12] L. de Haan, S.I. Resnick, A simple asymptotic estimate for the index of a stable distribution, J. R. Stat. Soc. Ser. B Stat. Methodol. 42 (1) (1980) 83–87.
[13] A.L. Dekkers, J.H. Einmahl, L. De Haan, A moment estimator for the index of an extreme-value distribution, Ann. Statist. 17 (4) (1989) 1833–1855.
[14] A. Dematteo, S. Clémençon, On tail index estimation based on multivariate data, J. Nonparametr. Stat. 28 (1) (2016) 152–176.
[15] Y. Dominicy, P. Ilmonen, D. Veredas, Multivariate hill estimators, Internat. Statist. Rev. 85 (1) (2017) 108–142.
[16] H. Drees, Weighted approximations of tail processes for β-mixing random variables, Ann. Appl. Probab. 10 (4) (2000) 1274–1301.
[17] H. Drees, Tail empirical processes under mixing conditions, in: Empirical Process Techniques for Dependent Data, Birkhäuser Boston, Boston, 2002, pp. 325–342.
[18] H. Drees, Extreme quantile estimation for dependent data, with applications to finance, Bernoulli 9 (4) (2003) 617–657.
[19] J. Einmahl, A. Krajina, J. Segers, An M-estimator for tail dependence in arbitrary dimensions, Ann. Statist. 40 (3) (2012) 1764–1793.
[20] P. Embrechts, C. Klüppelberg, T. Mikosch, Modelling Extremal Events: For Insurance and Finance, vol. 33, Springer Science & Business Media, Berlin, 2013.
[21] M. Heikkilä, Y. Dominicy, P. Ilmonen, On multivariate separating hill estimator under estimated location and scatter, Statistics 53 (2) (2019) 301–320.
[22] B.M. Hill, A simple general approach to inference about the tail of a distribution, Ann. Statist. 3 (5) (1975) 1163–1174.
[23] Y. Hoga, Detecting tail risk differences in multivariate time series, J. Time Series Anal. 39 (5) (2018) 665–689.
[24] R.A. Horn, C.R. Johnson, Matrix Analysis, Cambridge University Press, New York, 2013.
[25] T. Hsing, On tail index estimation using dependent data, Ann. Statist. 19 (3) (1991) 1547–1569.
[26] A. Hyvärinen, J. Karhunen, E. Oja, Independent Component Analysis, Wiley, New York, 2001.
[27] A. Hyvärinen, E. Oja, A fast fixed-point algorithm for independent component analysis, Neural Comput. 9 (7) (1997) 1483–1492.
[28] A. Hyvärinen, E. Oja, Independent component analysis: algorithms and applications, Neural Netw. 13 (4–5) (2000) 411–430.
[29] P. Ilmonen, J. Nevalainen, H. Oja, Characteristics of multivariate distributions and the invariant coordinate system, Statist. Probab. Lett. 80 (23) (2010) 1844–1853.
[30] R.A. Johnson, D.W. Wichern, Applied Multivariate Statistical Analysis, Pearson Prentice Hall, Upper Saddle River, NJ, 2002.
[31] M. Kim, S. Lee, Estimation of the tail exponent of multivariate regular variation, Ann. Inst. Statist. Math. 69 (5) (2017) 945–968.
[32] K. Kiviluoto, E. Oja, Independent component analysis for parallel financial time series, in: ICONIP, vol. 2, 1998, pp. 895–898.

[33] N. Lietzén, L. Viitasaari, P. Ilmonen, Modeling temporally uncorrelated components of complex-valued stationary processes, Mod. Stoch. Theory Appl. 8 (4) (2021) 475–508.
[34] C.-J. Lu, T.-S. Lee, C.-C. Chiu, Financial time series forecasting using independent component analysis and support vector regression, Decis. Support Syst. 47 (2) (2009) 115–125.
[35] J. Miettinen, K. Illner, K. Nordhausen, H. Oja, S. Taskinen, F.J. Theis, Separation of uncorrelated stationary time series using autocovariance matrices, J. Time Series Anal. 37 (3) (2016) 337–354.
[36] J. Miettinen, K. Nordhausen, H. Oja, S. Taskinen, Deflation-based separation of uncorrelated stationary time series, J. Multivariate Anal. 123 (2014) 214–227.
[37] J. Miettinen, K. Nordhausen, H. Oja, S. Taskinen, J. Virta, The squared symmetric FastICA estimator, Signal Process. 131 (2017) 402–411.
[38] J. Miettinen, K. Nordhausen, S. Taskinen, Blind source separation based on joint diagonalization in R: The packages JADE and BSSasymp, J. Stat. Softw. 76 (2) (2017) 1–31.
[39] J. Miettinen, S. Taskinen, K. Nordhausen, H. Oja, Fourth moments and independent component analysis, Stat. Sci. 30 (3) (2015) 372–390.
[40] T. Mikosch, C. Stărică, Limit theory for the sample autocorrelations and extremes of a GARCH(1, 1) process, Ann. Statist. 28 (5) (2000) 1427–1451.
[41] K. Nordhausen, P. Ilmonen, A. Mandal, H. Oja, E. Ollila, Deflation-based FastICA reloaded, in: 2011 19th European Signal Processing Conference, IEEE, 2011, pp. 1854–1858.
[42] K. Nordhausen, H. Oja, E. Ollila, Robust independent component analysis based on two scatter matrices, Austrian J. Stat. 37 (1) (2008) 91–100.
[43] D. Nualart, Fractional Brownian motion: stochastic calculus and applications, in: International Congress of Mathematicians, vol. 3, European Mathematical Society, 2006, pp. 1541–1562.
[44] H. Oja, S. Sirkiä, J. Eriksson, Scatter matrices and independent component analysis, Austrian J. Stat. 35 (2&3) (2006) 175–189.
[45] S.T. Rachev, Handbook of Heavy Tailed Distributions in Finance: Handbooks in Finance, Elsevier, Amsterdam, 2003.
[46] S. Resnick, C. Stărică, Asymptotic behavior of Hill's estimator for autoregressive data, Commun. Stat. Stoch. Models 13 (4) (1997) 703–721.
[47] W. Revelle, psych: Procedures for Psychological, Psychometric, and Personality Research, Northwestern University, Evanston, Illinois, 2023, R package version 2.3.9.
[48] B.B. Risk, D.S. Matteson, D. Ruppert, Linear non-Gaussian component analysis via maximum likelihood, J. Amer. Statist. Assoc. 114 (525) (2019) 332–343.
[49] S.J. Roberts, Extreme value statistics for novelty detection in biomedical data processing, IEE Proc. Sci. Meas. Technol. 147 (6) (2000) 363–367.
[50] H. Rootzén, The Tail Empirical Process for Stationary Sequences, Technical Report 9, Chalmers University of Technology, 1995.
[51] H. Rootzén, Weak convergence of the tail empirical function for dependent sequences, Stochastic Process. Appl. 119 (2) (2009) 468–490.
[52] M. Smith, S. Reece, S. Roberts, I. Rezek, Online maritime abnormality detection using Gaussian processes and extreme value theory, in: 2012 IEEE 12th International Conference on Data Mining, IEEE, 2012, pp. 645–654.
[53] C. Stărică, On the Tail Empirical Process of Solutions of Stochastic Difference Equations, Technical Report 25, Chalmers University of Technology, 2000.
[54] L. Tong, R.-W. Liu, V.C. Soon, Y.-F. Huang, Indeterminacy and identifiability of blind identification, IEEE Trans. Circuits Syst. 38 (5) (1991) 499–509.
[55] L. Tong, V. Soon, Y. Huang, R. Liu, AMUSE: a new blind identification algorithm, in: Proceedings of IEEE International Symposium on Circuits and Systems, 1990, pp. 1784–1787.
[56] J. Virta, N. Lietzén, P. Ilmonen, K. Nordhausen, Fast tensorial JADE, Scand. J. Stat. 48 (1) (2021) 164–187.
[57] S. Yang, Z. Yi, Fast ICA for online cashflow analysis, in: International Symposium on Neural Networks, Springer, 2005, pp. 891–896.
[58] W. Zhou, D. Chelidze, Blind source separation based vibration mode identification, Mech. Syst. Signal Process. 21 (8) (2007) 3072–3087.