Electronic Journal of Statistics
Vol. 19 (2025) 5695–5740
ISSN: 1935-7524
https://doi.org/10.1214/25-EJS2469

Information criteria for the number of directions of extremes in high-dimensional data

Lucas Butsch[1] **and Vicky Fasen-Hartmann**[1]

[1] *Institute of Stochastics, Karlsruhe Institute of Technology,*
e-mail: lucas.butsch@kit.edu; vicky.fasen@kit.edu

Abstract: In multivariate extreme value analysis, the estimation of the dependence structure in extremes is demanding, especially in the context of high-dimensional data. Therefore, a common approach is to reduce the model dimension by considering only the directions in which extreme values occur. In this paper, we use the concept of sparse regular variation recently introduced by Meyer and Wintenberger [28] to derive information criteria for the number of directions in which extreme events occur, such as a Bayesian information criterion (BIC), a mean-squared error-based information criterion (MSEIC), and a quasi-Akaike information criterion (QAIC) based on the Gaussian likelihood function. As is typical in extreme value analysis, a challenging task is the choice of the number k_n of observations used for the estimation. Therefore, for all information criteria, we present a two-step procedure to estimate both the number of directions of extremes and an optimal choice of k_n. We prove that the AIC of Meyer and Wintenberger [29] and the MSEIC are inconsistent information criteria for the number of extreme directions whereas the BIC and the QAIC are consistent information criteria. Finally, the performance of the different information criteria is compared in a simulation study and applied on wind speed data.

MSC2020 subject classifications: Primary 62G32, 62H30; secondary 62F07, 62H12.
Keywords and phrases: AIC, BIC, consistency, extreme directions, extreme value statistics, information criteria, multivariate regular variation, sparse regular variation.

Received September 2024.

1. Introduction

Multivariate extreme value statistics analyses the probabilities of joint extreme events in multivariate data with a wide range of applications, such as finance, insurance, meteorology, hydrology and, more generally, environmental risks due to the influence of climate change. This is a challenging task, especially for high-dimensional data, where modern research combines knowledge from extreme value theory with multivariate statistics and machine learning.

Multivariate regular variation is a classical concept for modeling multivariate extremes (Falk [19], Resnick [32, 33]). Suppose $\boldsymbol{X} \in \mathbb{R}_+^d$ is a d-dimensional random vector and there exists an index $\alpha > 0$ (tail index) and a measure S on the unit sphere $\mathbb{S}_+^{d-1} := \{\boldsymbol{x} \in \mathbb{R}_+^d : \|\boldsymbol{x}\| = 1\}$ (spectral measure) such that

$$\mathbb{P}\left(\frac{\|\boldsymbol{X}\|}{t} > r, \frac{\boldsymbol{X}}{\|\boldsymbol{X}\|} \in A \middle| \|\boldsymbol{X}\| > t\right) \longrightarrow r^{-\alpha} S(A), \quad t \to \infty, \tag{1.1}$$

for all $r > 0$ and all Borel sets $A \subset \mathbb{S}_+^{d-1}$ with $S(\partial A) = 0$, then $\boldsymbol{X}$ is called *multivariate regularly varying* of index α. The spectral measure S contains the information about the

dependence structure in the extremes of $\boldsymbol{X}$ and therefore a particular goal is the determination of S. However, in high-dimensional data sets where d is large, this can be challenging and computationally intensive because the dependence structure in the extremes is usually complex. In the case of high dimensions, the spectral measure is often sparse and has support in a lower-dimensional subspace. Therefore, a standard approach from multivariate statistics is to first apply a dimension reduction method to find the support of S and then to estimate S, which drastically reduces the computational time and the quality of the estimation.

The literature on dimension reduction methods for multivariate extremes using statistical learning methods has grown rapidly in recent years. Starting with Chautru [8] who first applies a principal component analysis (PCA) and then a cluster analysis with spherical k-means to the spectral measure of a multivariate regularly varying random vector to find a group of variables that are jointly extreme. The reconstruction error of PCA is then analyzed in Drees and Sabourin [13] and recently, Clémençon, Huet and Sabourin [10] extend the PCA approach to Hibert-valued regularly varying random objects, whereas Avella-Medina, Davis and Samorodnitsky [2] use with kernel PCA a nonlinear generalization of PCA. In addition, Cooley and Thibaud [11], Rohrbeck and Cooley [34] apply a PCA to the tail pairwise dependence matrix. The unsupervised learning approach of using spherical k-means, a variant of k-means, for cluster analysis in extreme observations was taken up in Avella Medina, Davis and Samorodnitsky [1], Bernard et al. [3], Fomichov and Ivanovs [20], Janßen and Wan [26]. The topic of this paper is support identification of the spectral measure, and the related literature is Goix, Sabourin and Clémençon [23], Jalalzai and Leluc [25], Meyer and Wintenberger [29], Simpson, Wadsworth and Tawn [36]. A completely different line of research to represent the sparsity structure in multivariate models are graphical models as, e.g., Engelke and Hitz [15], Engelke and Volgushev [17], Engelke et al. [18], Gissibl and Klüppelberg [21], Gissibl, Klüppelberg and Lauritzen [22], to name only a few. A very nice overview of recent advances in probabilistic and statistical aspects of sparse structures in extremes is given in Engelke and Ivanovs [16].

The support of S can be identified by the disjoint partition of the unit sphere $\mathbb{S}_+^{d-1}$ into sets of the form

$$C_\beta \coloneqq \{\boldsymbol{x} \in \mathbb{S}_+^{d-1} : x_i > 0 \text{ for } i \in \beta, x_i = 0 \text{ for } i \notin \beta\} \subseteq \mathbb{S}_+^{d-1}, \quad \beta \subset \{1, \dots, d\}. \tag{1.2}$$

Knowing $S(C_\beta)$ for all $\beta \subseteq \{1, \dots, d\}$ allows us to draw conclusions about the support of S and the directions of the extremes. Of course, $S(C_\beta) > 0$ implies that the components in the set β are jointly extreme, we have an extreme event in the direction β. However, the disjoint partition of $\mathbb{S}_+^{d-1}$ consists of $2^d - 1$ sets so it is huge for large values of d, and estimating $S(C_\beta)$ is non-trivial. On the one hand, $C_\beta = \partial C_\beta$ and therefore the interior of C_β is the empty set. As a consequence, if $S(C_\beta) > 0$ then the convergence in (1.1) for $A = C_\beta$ does not necessarily hold. On the other hand, if $\boldsymbol{X}$ has a continuous distribution there are empirically no observations in the set C_β. Therefore, the empirical estimator for $S(C_\beta)$ based on (1.1) is not consistent and useful anymore. To avoid this problem, the support detection algorithm DAMEX (Detecting Anomalies among Multivariate EXtremes) of Goix, Sabourin and Clémençon [23] works with truncated ε-cones to generate continuity sets that approximate the sets in (1.2), and Simpson, Wadsworth and Tawn [36] use the concept of hidden regular variation on a collection of nonstandard subcones of $[0, \infty]^d \backslash \{0\}$.

A completely different approach to mitigate this problem is proposed in Meyer and Wintenberger [28, 29] by introducing the concept of sparse regular variation, which is equivalent to regular variation under some mild assumptions (see Section 2 for a definition). The main difference between regular variation and sparse regular variation is that the self-normalization $\boldsymbol{X}/\|\boldsymbol{X}\|$ in (1.1) is replaced by the Euclidean projection $\pi(\boldsymbol{X}/t)$ of $\boldsymbol{X}/t$ for large $t > 0$, where the Euclidean projection $\pi : \mathbb{R}_+^d \to \mathbb{S}_+^{d-1}$ is defined as in Duchi et al. [14] as $\pi(\boldsymbol{v}) = \arg\min_{\boldsymbol{w}\in\mathbb{R}_+^d:\|w\|_1=1} \|\boldsymbol{w} - \boldsymbol{v}\|_2^2$. The advantage of this approach is that $\pi(\boldsymbol{X}/t)$ usually has more zero entries than $\boldsymbol{X}/\|\boldsymbol{X}\|$ and therefore, is more sparsely populated and advantageous when only a few components are extreme together, as in a high-dimensional setting. Since their empirical estimator for the number of extreme directions in the sparse regularly varying model is biased, indeed overestimates the true number of directions, they develop an Akaike Information Criterion (AIC) consisting of two steps. In the first step, they estimate the number of extreme directions by the AIC for *bias selection*, but as usual, in extreme value theory, the estimation depends on the chosen threshold that goes into the estimation; the observations above this threshold determine the extreme observations. Therefore, they extend the AIC for *bias selection* to an AIC for *threshold selection*, where the threshold is also estimated. What is really special is that they were able to develop a method to estimate the number of extreme directions and the threshold at the same time, both of which are very challenging tasks on their own. But as we prove in Theorem 3.1, the AIC for bias selection is not a weakly consistent information criterion, as is often the case for Akaike's information criteria, and so we develop alternatives. Consistency is examined only for *bias selection* and not for *threshold selection*, because there is no "true" threshold. Here, we have the well-known bias-variance tradeoff: If the threshold is chosen too high, there are not enough extreme observations leading to a high variance, and if it is too low, non-extreme observations lead to a bias in the estimation.

In this paper we use the approach of Meyer and Wintenberger [29] of sparse regular variation and propose three different information criteria to estimate the number of extreme directions and the choice of the threshold, the BIC, QAIC and MSEIC for *bias selection* and *threshold selection*, which are particularly suitable for high dimensional data with a sparsity structure in the extreme behavior. Thus, we develop procedures to estimate the number of extreme directions and the optimal choice of the threshold at the same time. The application of these information criteria is very simple in practice and not computationally intensive. Besides the AIC, the Bayesian Information Criterion (BIC), which goes back to Schwarz [35], is the most popular in practice and tries to select the model with the highest posterior probability. The statistical model behind our BIC is the same as that of the AIC in [29], where we fit a multinomial model to the number of extreme observations in the subspaces C_β and derive an asymptotic upper bound on the posterior likelihood, which then defines the BIC. In contrast, the QAIC for Quasi-Akaike Information Criterion approximates the Kullback-Leibler divergence of the true model and a Gaussian model, rather than a multinomial model as used in the AIC and BIC, respectively. The advantage of BIC and QAIC over AIC is that they are consistent information criteria for *bias selection*. Finally, the third method, MSEIC, stands for mean-squared error information criteria, because we approximate the mean-squared error (MSE) of the relative number of extreme observations and the true probabilities of extremes in the different subspaces C_β. Although MSEIC is not consistent for bias selection, it performs extremely well in all simulations.

Structure of the paper

The paper is organized as follows. In Section 2 we properly define extreme directions based on the concept of sparse regular variation and introduce consistent and asymptotically normally distributed estimators for the probabilities of the extreme directions as in Meyer and Wintenberger [29]. We also present statistical models for some of our information criteria. The main results of the paper are derived in Sections 3 to 5. In Section 3, we first introduce the QAIC for bias selection and threshold selection following the framework of Akaike information criteria, which aims to minimize the expected Kullback-Leibler (KL) divergence, here applied to a Gaussian likelihood function. We prove that, unlike the AIC proposed by Meyer and Wintenberger [29], the QAIC for bias selection is a consistent information criterion. In Section 4, we develop the MSEIC and finally, in Section 5, the BIC for both bias selection and threshold selection. In addition, we demonstrate in these sections that the BIC is a consistent information criterion for bias selection, whereas the MSEIC is not consistent. Moreover, we compare all information criteria in a simulation study in Section 6 and apply them to extreme wind data from the Republic of Ireland in Section 7. Finally, we draw some conclusions in Section 8. The main proofs of the paper are moved to the appendix, while the proofs of some auxiliary results can be found in the Supplementary Material A [5].

Notation

In this paper, we use the following notation. For a vector $\boldsymbol{x} = (x_1, \dots, x_d)^\top \in \mathbb{R}^d$ and a set $I \subset \{1, \dots, d\}$ we write $\boldsymbol{x}_I \in \mathbb{R}^{|I|}$ for $(x_i)_{i \in I}$ and $\operatorname{diag}(\boldsymbol{x}) \in \mathbb{R}^{d \times d}$ for a diagonal matrix with the components of $\boldsymbol{x}$ on the diagonal. Furthermore, $\boldsymbol{I}_d \in \mathbb{R}^{d \times d}$ is the identity matrix, $\boldsymbol{0}_d \coloneqq (0, \dots, 0)^\top \in \mathbb{R}^d$ is the zero vector and $\boldsymbol{1}_d \coloneqq (1, \dots, 1)^\top \in \mathbb{R}^d$ is the vector containing only 1. Moreover, $\|\boldsymbol{x}\| \coloneqq \|\boldsymbol{x}\|_1$ is the L_1-norm and $\|\boldsymbol{x}\|_2$ is the Euclidean norm for $\boldsymbol{x} \in \mathbb{R}^d$. The unit sphere $\mathbb{S}_+^{d-1} = \{\boldsymbol{x} \in [0, \infty)^d : x_1 + \cdots + x_d = 1\}$ is defined with respect to the L_1-norm. For $a \in \mathbb{R}$, $\boldsymbol{x}, \boldsymbol{y} \in \mathbb{R}^d$ operations as $\boldsymbol{x}^a$, $\sqrt{\boldsymbol{x}}$ and $\boldsymbol{x} \cdot \boldsymbol{y}$ are meant component-wise. The gradient of a function $f : \mathbb{R}^d \mapsto \mathbb{R}^k$ is written as $\nabla f(\boldsymbol{x}) \in \mathbb{R}^{k \times d}$ for $\boldsymbol{x} \in \mathbb{R}^d$ and the partial derivative with respect to the i-the component x_i of $\boldsymbol{x} = (x_1, \dots, x_d)^\top$ is $\frac{\partial}{\partial x_i} f(\boldsymbol{x})$. By $|a|$ we denote the absolute value of a real number a and by $|A|$ the cardinality of a set A, but the meaning should be clear from the context. In addition, $\mathcal{P}_d$ is the power set of the set $\{1, \dots, d\}$ and $\mathcal{P}_d^* \coloneqq \mathcal{P}_d \setminus \emptyset$. Finally, $\xrightarrow{\mathcal{D}}$ is the notation for convergence in distribution and $\xrightarrow{\mathbb{P}}$ is the notation for convergence in probability.

2. Preliminaries

This section addresses the main concepts of the paper which are based on Meyer and Wintenberger [28, 29]. We start with an introduction into sparse regular variation and then derive a proper definition of *extreme direction* in Section 2.1. The challenging task in the statistical inference of extreme directions is the detection of the *bias directions* which are rigorously defined and motivated in Section 2.2. Then, in Section 2.3, we give an overview on the statistical inference of the empirical estimator of the probabilities of extreme directions and the assumptions of the present paper. Finally, in Section 2.4, we present statistical models on which the information criteria are based.

2.1. Sparse regular variation and extreme directions

First, we introduce the concept of sparse regular variation with the Euclidean projection $\pi : \mathbb{R}_+^d \to \mathbb{S}_+^{d-1}$ defined as $\pi(\boldsymbol{v}) = \arg\min_{\boldsymbol{w} \in \mathbb{R}_+^d : \|w\|_1 = 1} \|\boldsymbol{w} - \boldsymbol{v}\|_2^2$.

Definition 2.1. An $\mathbb{R}_+^d$-valued random vector $\boldsymbol{X}$ is called *sparse regular varying*, if a $\mathbb{S}_+^{d-1}$-valued random vector $\boldsymbol{Z}$ and a non degenerate random variable R exist such that

$$\mathbb{P}\left(\frac{\|\boldsymbol{X}\|}{t} > r, \pi\left(\frac{\boldsymbol{X}}{t}\right) \in A \,\middle|\, \|\boldsymbol{X}\| > t\right) \to \mathbb{P}(R > r, \boldsymbol{Z} \in A), \quad t \to \infty,$$

for all $r > 0$ and all Borel sets $A \subset \mathbb{S}_+^{d-1}$ with $\mathbb{P}(\boldsymbol{Z} \in \partial A) = 0$.

Remark 2.2.

(a) Note that R is Pareto(α)-distributed for an $\alpha > 0$ and models the radial part, whereas the $\mathbb{S}_+^{d-1}$-valued random vector $\boldsymbol{Z}$ corresponds to the angular part. Therefore, we write briefly $\boldsymbol{X} \in \mathrm{SRV}(\alpha, \boldsymbol{Z})$.

(b) The concept of sparse regular variation introduced by Meyer and Wintenberger [28] is currently limited to random vectors in the positive orthant. A corresponding theory for $\mathbb{R}^d$-valued random vectors has not yet been developed. Consequently, in this paper, we also restrict our analysis to random vectors in the positive orthant, which aligns with ours and many other applications.

A proper definition of extreme direction is now the following, where we use the notation that $\mathcal{P}_d$ is the power set of the set $\{1, \ldots, d\}$ and $\mathcal{P}_d^* := \mathcal{P}_d \setminus \emptyset$.

Definition 2.3. A direction $\beta \in \mathcal{P}_d^*$ is an *extreme direction*, if $\mathbb{P}(\boldsymbol{Z} \in C_\beta) > 0$. The set of all extreme directions is denoted as

$$S(\boldsymbol{Z}) := \{\beta \in \mathcal{P}_d^* : \mathbb{P}(\boldsymbol{Z} \in C_\beta) > 0\} \quad \text{with} \quad s^* := |S(\boldsymbol{Z})|.$$

Remark 2.4.

(a) The use of the L_1-projection leads to a sparse representation, in the sense that under π more components are projected to zero compared to the normalization $\boldsymbol{v} \mapsto \boldsymbol{v}/\|\boldsymbol{v}\|$. Therefore, it is not surprising that according to Meyer and Wintenberger [28, Theorem 2], $S(C_\beta) > 0$ implies $\mathbb{P}(\boldsymbol{Z} \in C_\beta) > 0$ for $\beta \in \mathcal{P}_d^*$. Thus, an extreme direction under regular variation is as well an extreme direction under sparse regular variation but the opposite does not necessarily hold. However, the maximal directions under regular variation and sparse regular variation are equivalent, such that we do not lose much information on the support of S under sparse regular variation. Note that a direction $\beta \in \mathcal{P}_d^*$ is called a maximal direction of the regularly varying random vector $\boldsymbol{X}$ if $\mathbb{P}(\boldsymbol{\Theta} \in C_\beta) > 0$ and $\mathbb{P}(\boldsymbol{\Theta} \in C_{\beta'}) = 0$ for all $\beta \subset \beta' \in \mathcal{P}_d^*$. In the case of sparse regular variation, the definition of a maximal direction is analogous, except that the random vector $\boldsymbol{\Theta}$ is replaced by $\boldsymbol{Z}$.

(b) Since the preimages $\pi^{-1}(C_\beta)$ are sets with positive Lebesgue measure, the sets C_β are continuity sets of $\mathbb{P}(\boldsymbol{Z} \in \cdot)$. Finally, from Meyer and Wintenberger [28, Proposition 2] we know that

$$\mathbb{P}(\pi(\boldsymbol{X}/t) \in C_\beta \mid \|\boldsymbol{X}\| > t) \longrightarrow \mathbb{P}(\boldsymbol{Z} \in C_\beta), \quad \text{as } t \to \infty,$$

so that $\mathbb{P}(\boldsymbol{Z} \in C_\beta)$ can be estimated empirically in contrast to $S(C_\beta)$.

The aim of the paper is to estimate s^*, the number of extreme directions under sparse regular variation, through the use of information criteria.

2.2. Bias directions

A major challenge for the estimation of the extreme directions is that the empirical estimators of the probabilities $\mathbb{P}(\boldsymbol{Z} \in C_\beta)$, $\beta \in \mathcal{P}_d^*$, detect more extremal directions than there are true extremal directions, which we call *bias directions*. To understand the idea of bias directions better we require some further notation. Suppose $\|\boldsymbol{X}_{(1,n)}\| \geq \cdots \geq \|\boldsymbol{X}_{(n,n)}\|$ is the order statistic of $\|\boldsymbol{X}_1\|, \ldots, \|\boldsymbol{X}_n\|$ and the number of extreme observations used for the estimations is denoted by $k_n \in \mathbb{N}$, whereas we assume that $k_n \to \infty$ as $n \to \infty$. Suppose that there exists a sequence of high thresholds $u_n > 0$ for $n \in \mathbb{N}$ such that $k_n/n \sim \mathbb{P}(\|\boldsymbol{X}\| > u_n)$ and $u_n \to \infty$ as $n \to \infty$. Due to Meyer and Wintenberger [29, Proposition 1] the empirical estimator

$$\frac{T_n(C_\beta, k_n)}{k_n} \coloneqq \frac{1}{k_n} \sum_{j=1}^{n} \mathbb{1}\left\{\pi(\boldsymbol{X}_j/\|\boldsymbol{X}_{(k_n+1,n)}\|) \in C_\beta, \|\boldsymbol{X}_j\| > \|\boldsymbol{X}_{(k_n+1,n)}\|\right\},$$

of the probability

$$p(C_\beta) \coloneqq \mathbb{P}(\boldsymbol{Z} \in C_\beta) = \lim_{n\to\infty} \mathbb{P}(\pi(\boldsymbol{X}/u_n) \in C_\beta \mid \|\boldsymbol{X}\| > u_n) \tag{2.1}$$

is a consistent estimator, so that the empirical observed set of extreme directions is

$$\widehat{\mathcal{S}}_n(\boldsymbol{Z}) \coloneqq \{\beta \in \mathcal{P}_d^* : T_n(C_\beta, k_n) > 0\}.$$

To be able to relate the true set of extreme directions $\mathcal{S}(\boldsymbol{Z})$ with the empirically estimated set of extreme directions, we define the set

$$\mathcal{R} \coloneqq \{\beta \in \mathcal{P}_d^* : \lim_{n\to\infty} k_n p_n(C_\beta) = \infty\} \qquad \text{and} \qquad r \coloneqq |\mathcal{R}|,$$

where $\mathcal{R}$ depends on the chosen sequence $(k_n)_{n\in\mathbb{N}}$, which we neglect for the ease of notation, and

$$p_n(C_\beta) \coloneqq \mathbb{P}(\pi(\boldsymbol{X}/u_n) \in C_\beta \mid \|\boldsymbol{X}\| > u_n).$$

Of course, $\beta \in \mathcal{S}(\boldsymbol{Z})$ implies $k_n p_n(C_\beta) \to \infty$ such that trivially, $\mathcal{S}(\boldsymbol{Z}) \subseteq \mathcal{R}$ and $s^* \leq r$. Under the Assumption HRV, a shorthand for hidden regular variation, we can say more about the relations of these sets.

Assumption HRV. *For every $\beta \in \mathcal{P}_d^*$ we define the cone*

$$\mathbb{C}_\beta \coloneqq \left\{\boldsymbol{x} = (x_1, \ldots, x_d)^\top \in \mathbb{R}_+^d : \sum_{j\in\beta} (x_j - \max_{i\in\beta^c} x_i) \geq 0\right\} \subseteq \mathbb{R}_+^d$$

and suppose that the random vector $\boldsymbol{X}$ is multivariate regular varying on $\mathbb{R}_+^d \setminus \mathbb{C}_\beta$ with tail index $\alpha(\beta)$ and exponent measure μ_β satisfying

$$\mu_\beta\left(\left\{\boldsymbol{x} = (x_1, \ldots, x_d)^\top \in \mathbb{R}_+^d : \max_{i\in\beta} x_i < 1, \min_{i\in\beta^c} x_i \geq 1\right\}\right) > 0.$$

A conclusion from Meyer and Wintenberger [29, Proposition 2] is then that under Assumption HRV even

$$\lim_{n\to\infty} \mathbb{P}(\mathcal{S}(\boldsymbol{Z}) \subseteq \mathcal{R} \subseteq \widehat{\mathcal{S}}_n(\boldsymbol{Z})) = 1 \tag{2.2}$$

holds. Thus, the empirical estimator tends to overestimate the set of extreme directions (but does not underestimate it asymptotically). On the one hand, for n large and $\beta \in \mathcal{P}_d^*$ with $T_n(C_\beta, k_n) = 0$ this means that β is not an extreme direction. But on the other hand, for n large there might be a $\beta \in \mathcal{P}_d^*$ with $T_n(C_\beta, k_n) > 0$ which is not an extreme direction; a mathematical more rigorous interpretation is given in Meyer and Wintenberger [29]. Such a direction is referred to as a *bias direction*. The main challenge is to identify these bias directions.

Remark 2.5. There exists as well a stronger statement than (2.2). Suppose additionally that $\lim_{n\to\infty} k_n p_n(\beta) = 0$ for all $\beta \in \mathcal{P}_d^* \setminus \mathcal{R}$. A conclusion of Meyer and Wintenberger [29, Lemma 1] is then that $\lim_{n\to\infty} \mathbb{P}(T_n(C_\beta, k_n) = 0) = 1$ for all $\beta \in \mathcal{P}_d^* \setminus \mathcal{R}$ and hence,

$$\lim_{n\to\infty} \mathbb{P}(\mathcal{S}(\boldsymbol{Z}) \subseteq \mathcal{R} = \widehat{\mathcal{S}}_n(\boldsymbol{Z})) = 1.$$

In particular, this means that $\widehat{r}_n := |\mathcal{S}_n(\boldsymbol{Z})| \xrightarrow{\mathbb{P}} r$ as $n \to \infty$.

2.3. Statistical inference for the probabilities of extreme directions

The general assumptions of the present paper are motivated by the statistical inference of the probabilities of extreme directions as derived in Meyer and Wintenberger [29]. To understand the statistical inference and hence, the assumptions, we have to enumerate the $\beta \in \mathcal{P}_d^*$ in the following way with $p(C_\beta)$ as defined in (2.1):

$$\begin{aligned}
\beta_1 &:= \operatorname*{arg\,max}_{\beta\in\mathcal{P}_d^*} p(C_\beta),\\
\beta_2 &:= \operatorname*{arg\,max}_{\beta\in\mathcal{P}_d^*\setminus\{\beta_1\}} p(C_\beta),\\
&\vdots\\
\beta_{s^*} &:= \operatorname*{arg\,max}_{\beta\in\mathcal{P}_d^*\setminus\{\beta_1,\ldots,\beta_{s^*-1}\}} p(C_\beta),
\end{aligned}$$

where the remaining $\beta_{s^*+1}, \ldots, \beta_{2^d-1}$ with $p(C_{\beta_j}) = 0$, $j = s^* + 1, \ldots, 2^d - 1$, are ordered in an arbitrary but fixed order such that $\beta_j \in \mathcal{R}$ for $j = s^* + 1, \ldots, r$. We write briefly for $j = 1, \ldots, 2^d - 1$,

$$\begin{aligned}
p_j &:= p(C_{\beta_j}), & p_{n,j} &:= p_n(C_{\beta_j}) := \mathbb{P}(\pi(\boldsymbol{X}/u_n) \in C_{\beta_j} \mid \|\boldsymbol{X}\| > u_n),\\
\mathcal{T}_{n,j} &:= \mathcal{T}_n(C_{\beta_j}), & T_{n,j}(k_n) &:= T_n(C_{\beta_j}, k_n),
\end{aligned}$$

where

$$\frac{\mathcal{T}_n(C_\beta)}{k_n} := \frac{1}{k_n}\sum_{j=1}^{n} \mathbb{1}\{\pi(\boldsymbol{X}_j/u_n) \in C_\beta, \|\boldsymbol{X}_j\| > u_n\}.$$

Finally, we define the associated vectors

$$\begin{aligned}
\boldsymbol{p} &:= (p_1, \ldots, p_r)^\top, & \boldsymbol{p}_n &:= (p_{n,1}, \ldots, p_{n,r})^\top,\\
\boldsymbol{\mathcal{T}}_n &:= (\mathcal{T}_{n,1}, \ldots, \mathcal{T}_{n,r})^\top, & \boldsymbol{T}_n(k_n) &:= (T_{n,1}(k_n), \ldots, T_{n,r}(k_n))^\top.
\end{aligned}$$

In the next theorem, we summarize the asymptotic behavior of these estimators as derived in Meyer and Wintenberger [29, Theorem 1 and Proposition 3].

Proposition 2.6. *Suppose Assumption HRV holds and the sequence $(k_n)_{n\in\mathbb{N}}$ in $\mathbb{N}$ with $k_n \to \infty$ and $k_n/n \to 0$ satisfies $\mathcal{R} = \widehat{\mathcal{S}}_n(\boldsymbol{Z})$ almost surely for all n large enough. Furthermore, assume that for some $\tau > 0$ and any $j = 1, \dots, r$ as $n \to \infty$,*

$$\sup_{r\in[\frac{1}{1+\tau},1+\tau]} \sqrt{\frac{k_n}{p_{n,j}}}\left|\frac{n}{k_n}\mathbb{P}(\boldsymbol{X}/u_n \in \{\boldsymbol{x} \in \mathbb{R}_+^d : r\|\boldsymbol{x}\| > 1, \pi(r\boldsymbol{x}) \in C_{\beta_j}\}) - r^{\alpha(\beta_j)}p_{n,j}\right| \to 0.$$

(a) Then, as $n \to \infty$,

$$\sqrt{k_n}\mathrm{diag}(\boldsymbol{p}_n)^{-1/2}\left(\frac{\boldsymbol{T}_n}{k_n} - \boldsymbol{p}_n\right) \xrightarrow{\mathcal{D}} \mathcal{N}_r(\boldsymbol{0}_r, \boldsymbol{I}_r).$$

(b) If additionally $\sqrt{k_n}(p_{n,j} - p_j) \to 0$ as $n \to \infty$ and $j = 1, \dots, r$, then as $n \to \infty$,

$$\sqrt{k_n}\mathrm{diag}(\boldsymbol{p}_n)^{-1/2}\left(\frac{\boldsymbol{T}_n(k_n)}{k_n} - \boldsymbol{p}_n\right) \xrightarrow{\mathcal{D}} \left(\boldsymbol{I}_r - \sqrt{\boldsymbol{p}}\cdot\sqrt{\boldsymbol{p}}^{\top}\right)\mathcal{N}_r(\boldsymbol{0}_r, \boldsymbol{I}_r).$$

Motivated by this result we define for any $n \in \mathbb{N}$

$$\boldsymbol{p}_n^* := (p_{n,1}, \dots, p_{n,s^*}, \rho_n, \dots, \rho_n)^{\top} \in \mathbb{R}^r \quad \text{with} \quad \rho_n := \frac{1}{r - s^*}\sum_{j=s^*+1}^{r} p_{n,j}$$

and suppose the following assumption throughout the paper.

Assumption A.

(A1) Suppose $(k_n)_{n\in\mathbb{N}}$ is a sequence in $\mathbb{N}$ with $k_n \to \infty$ and $k_n/n \to 0$. Furthermore $\mathcal{R} = \widehat{\mathcal{S}}_n(\boldsymbol{Z})$ almost surely for all n large enough, which implies $r = |\mathcal{R}| = |\widehat{\mathcal{S}}_n(\boldsymbol{Z})| \geq s^*$ almost surely for all n large enough.

(A2) $T_{n,1}(k_n) \geq T_{n,2}(k_n) \geq \cdots \geq T_{n,r}(k_n)$ almost surely for all n large enough.

(A3) Suppose that as $n \to \infty$,

$$\sqrt{k_n}\mathrm{diag}(\boldsymbol{p}_n^*)^{-1/2}\left(\frac{\boldsymbol{T}_n(k_n)}{k_n} - \boldsymbol{p}_n^*\right) \xrightarrow{\mathcal{D}} \left(\boldsymbol{I}_r - \sqrt{\boldsymbol{p}}\cdot\sqrt{\boldsymbol{p}}^{\top}\right)\mathcal{N}_r(\boldsymbol{0}_r, \boldsymbol{I}_r).$$

(A4) Suppose that as $n \to \infty$,

$$\sqrt{k_n}\mathrm{diag}(\boldsymbol{p}_n^*)^{-1/2}\left(\frac{\boldsymbol{T}_n}{k_n} - \boldsymbol{p}_n^*\right) \xrightarrow{\mathcal{D}} \mathcal{N}_r(\boldsymbol{0}_r, \boldsymbol{I}_r).$$

Remark 2.7.

(a) A justification of Assumption (A1) is given in Remark 2.5, where a sufficient criterion for $\lim_{n\to\infty}\mathbb{P}(\mathcal{R} = \widehat{\mathcal{S}}_n(\boldsymbol{Z})) = 1$ is stated. Assumption (A1) is particularly useful for modelling purposes, as can be seen in the derivation of the AIC in Meyer and Wintenberger [29], and from other statements in that paper such as Proposition 2.6 above. If Assumption (A1) is not made, then the consistency results in this paper can be obtained by replacing r with $\widehat{r}_n := |\mathcal{S}_n(\boldsymbol{Z})|$ and assuming $\sqrt{k_n\rho_n}(\widehat{r}_n - r) \xrightarrow{\mathbb{P}} 0$ (cf. Remark 3.2 and Remark 3.9).

(b) Assumption (A2) is motivated by the fact that we have $\boldsymbol{T}_n(k_n)/k_n \xrightarrow{\mathbb{P}} \boldsymbol{p}$ and thus, for n sufficiently large $\boldsymbol{T}_n(k_n)$ is ordered by size with probability close to 1 because $\boldsymbol{p}$ is ordered by size.

(c) The assumptions (A3) and (A4) are not strong, in the case $p_{n,s^*+1} = \ldots = p_{n,r} = \rho_n$, Proposition 2.6 gives a sufficient criteria for (A3) or (A4) to hold.

The following lemma is a direct consequence of Assumption A.

Lemma 2.8. *Suppose Assumption A holds. Then the following statements are valid.*

(a) $\rho_n \to 0$ *and* $\rho_n k_n \to \infty$ *as* $n \to \infty$.

(b) For $j = 1, \ldots, s^*$ *and* $n \to \infty$,

$$\frac{T_{n,j}(k_n)}{k_n p_{n,j}} \xrightarrow{\mathbb{P}} 1 \quad \text{and} \quad \frac{\mathcal{T}_{n,j}}{k_n p_{n,j}} \xrightarrow{\mathbb{P}} 1.$$

(c) For $j = s^* + 1, \ldots, r$ *and* $n \to \infty$,

$$\frac{T_{n,j}(k_n)}{k_n \rho_n} \xrightarrow{\mathbb{P}} 1 \quad \text{and} \quad \frac{T_{n,j}(k_n)}{k_n} \xrightarrow{\mathbb{P}} 0,$$

and similarly,

$$\frac{\mathcal{T}_{n,j}}{k_n \rho_n} \xrightarrow{\mathbb{P}} 1 \quad \text{and} \quad \frac{\mathcal{T}_{n,j}}{k_n} \xrightarrow{\mathbb{P}} 0.$$

2.4. Statistical models

A challenging task in extreme value theory is the optimal choice of k_n, the number of extreme observations used for the estimation procedure. Therefore, we follow a two-step procedure as motivated in Meyer and Wintenberger [29]. In the first step, we fix k_n and estimate the relevant extreme directions $\beta \in \mathcal{S}(\boldsymbol{Z})$ and separate them from the so-called bias directions $\beta \in \widehat{\mathcal{S}}_n(\boldsymbol{Z}) \setminus \mathcal{S}(\boldsymbol{Z})$ using some information criteria. Therefore this step is called *bias selection*. In the second step, we estimate the threshold k_n, this step is therefore named *threshold selection*. In the following subsections, we present some statistical models for the *bias selection* in Section 2.4.1 and the statistical models for the *threshold selection* in Section 2.4.2.

2.4.1. The local model for the bias selection

Due to Assumption (A1) with $r = |\widehat{\mathcal{S}}_n(\boldsymbol{Z})|$ the random vector $\boldsymbol{T}_n(k_n)$ is multinomial distributed with k_n repetitions and unknown r-dimensional probability vector $\boldsymbol{p}_{n,k_n}$ which converges as $n \to \infty$ to $\boldsymbol{p}$. To detect the bias directions and hence, to estimate s^*, the idea is now to fit for any $s \in \{1, \ldots, r\}$ a multinomial distribution from the class $\{\text{Mult}(k_n, \boldsymbol{A}_s(\widetilde{\boldsymbol{p}}^s)) : \widetilde{\boldsymbol{p}}^s \in \Theta_s\}$ where $\boldsymbol{A}_s : \mathbb{R}^s \to \mathbb{R}^r$ is defined as

$$\boldsymbol{A}_s(\widetilde{\boldsymbol{p}}^s) = \left(\widetilde{p}_1^s, \ldots, \widetilde{p}_s^s, \frac{1 - \sum_{j=1}^s \widetilde{p}_j^s}{r-s}, \ldots, \frac{1 - \sum_{j=1}^s \widetilde{p}_j^s}{r-s} \right)^\top$$

and the parameter space Θ_s is defined as

$$\Theta_s := \left\{ \widetilde{\boldsymbol{p}}^s = (\widetilde{p}_1^s, \ldots, \widetilde{p}_s^s) \in (0,1)^s : \widetilde{p}_1^s \geq \cdots \geq \widetilde{p}_s^s, \sum_{j=1}^s \widetilde{p}_j^s < 1 \right\},$$

which reflects that there are $r-s$ bias directions. Finally, we define

$$\widetilde{\rho}^s := \frac{1-\sum_{j=1}^s \widetilde{p}_j^s}{r-s} \in (0,1) \quad \text{for } \widetilde{\boldsymbol{p}}^s \in \Theta_s.$$

We summarize this in the following model.

MODEL $M_{k_n}^s$: *The family of multinomial distributions* $\{\mathrm{Mult}(k_n, \boldsymbol{A}_s(\widetilde{\boldsymbol{p}}^s)) : \widetilde{\boldsymbol{p}}^s \in \Theta_s\}$ *with likelihood function*

$$L_{M_{k_n}^s}(\widetilde{\boldsymbol{p}}^s \mid \boldsymbol{T}_n(k_n)) = \frac{k_n!}{\prod_{j=1}^r T_{n,j}(k_n)!} \prod_{j=1}^s (\widetilde{p}_j^s)^{T_{n,j}(k_n)} \prod_{j=s+1}^r (\widetilde{\rho}^s)^{T_{n,j}(k_n)}$$

and log-likelihood function

$$\log L_{M_{k_n}^s}(\widetilde{\boldsymbol{p}}^s \mid \boldsymbol{T}_n(k_n)) = \log(k_n!) - \sum_{j=1}^r \log(T_{n,j}(k_n)!) + \sum_{j=1}^s T_{n,j}(k_n) \log(\widetilde{p}_j^s) + \log(\widetilde{\rho}^s) \sum_{j=s+1}^r T_{n,j}(k_n) \tag{2.3}$$

is called Model $M_{k_n}^s$.

Now, an information criterion aims to find the Model $M_{k_n}^s$ from $s \in \{1,\dots,r\}$ which best fits the distribution of $\boldsymbol{T}_n(k_n)$ and results in an estimator $\widehat{s}_n$ for s^*. Then, for a given estimator $\widehat{s}_n$ of s^* we estimate the probability vector $\boldsymbol{p}$ by

$$\widehat{\boldsymbol{p}}_{n,*}^{\widehat{s}_n} := \left(\frac{\widehat{p}_{n,1}^{\widehat{s}_n}}{\sum_{j=1}^{\widehat{s}_n} \widehat{p}_{n,j}^{\widehat{s}_n}}, \dots, \frac{\widehat{p}_{n,\widehat{s}_n}^{\widehat{s}_n}}{\sum_{j=1}^{\widehat{s}_n} \widehat{p}_{n,j}^{\widehat{s}_n}}, 0, \dots, 0 \right)^\top, \tag{2.4}$$

where

$$\widehat{\boldsymbol{p}}_n^s := (\widehat{p}_{n,1}^s, \dots, \widehat{p}_{n,s}^s)^\top := \left(\frac{T_{n,1}(k_n)}{k_n}, \dots, \frac{T_{n,s}(k_n)}{k_n} \right)^\top \tag{2.5}$$

is the maximum likelihood estimator (MLE) of the multinomial model $M_{k_n}^s$ (see Meyer and Wintenberger [29], Section 4.1). Finally, we define

$$\widehat{\rho}_n^s := \frac{1}{r-s}\left(1 - \sum_{j=1}^s \widehat{p}_{n,j}^s\right) = \frac{\sum_{j=s+1}^r T_{n,j}(k_n)}{(r-s)k_n}$$

as estimator for $\widetilde{\rho}^s$.

2.4.2. *The global model for the threshold* k_n

Next, we extend the previous model and assume that $k_n \in \mathbb{N}$ is not fixed anymore, it has additionally to be estimated. For this task, we use all observations $\boldsymbol{X}_1, \dots, \boldsymbol{X}_n$ and not only the k_n largest observations. We consider an artificial random vector $\boldsymbol{T}_n' = (T_{n,1}', \dots, T_{n,2^d}')^\top$ in $\mathbb{R}^{2^d}$ which includes extreme and non-extreme observations, where the $2^d - 1$ components $T_{n,1}', \dots, T_{n,2^{d-1}}'$ count the number of extreme observations in the subsets $C_{\beta_1}, \dots, C_{\beta_{2^d-1}}$.

The 2^d-th component $T'_{n,2^d}$ counts the number of non-extreme values and is $\mathrm{Bin}(n, 1-q_n)$-distributed for some $q_n \in (0,1)$. To be more precise we assume that $\boldsymbol{T}'_n \sim \mathrm{Mult}(n, \boldsymbol{p}'_n)$ with

$$\boldsymbol{p}'_n = (q_n p'_{n,1}, \dots, q_n p'_{n,2^{d-1}}, 1-q_n)$$

and the conditional distribution given $T'_{n,2^d} = n - k_n$ satisfies

$$\mathbb{P}_{(T'_{n,1},\dots,T'_{n,2^d-1})|T'_{n,2^d}=n-k_n} = \mathbb{P}_{(T_{n,1}(k_n),\dots,T_{n,2^d-1}(k_n))}. \tag{2.6}$$

The idea of this assumption is that if we have k_n extreme observations (and hence, $n-k_n$ non-extreme observations), then the distribution of the extreme directions $(T'_{n,1}, \dots, T'_{n,2^d-1})$ in the global model is the same as that of the local model $(T_{n,1}(k_n), \dots, T_{n,2^d-1}(k_n))$ with threshold k_n.

Now, the approach to detect the bias directions and the threshold k_n is similar to the previous section. We fit a multinomial distribution from the class $\{\mathrm{Mult}(n, \boldsymbol{A}'_s(\widetilde{\boldsymbol{p}}'^s)) : \widetilde{\boldsymbol{p}}'^s \in \Theta'_s\}$ to the artificial random vector $\boldsymbol{T}'_n$ where $\boldsymbol{A}'_s : \mathbb{R}^{s+1} \to \mathbb{R}^{2^d}$ is defined as

$$\boldsymbol{A}'_s(\widetilde{\boldsymbol{p}}'^s) = (q'^s \widetilde{p}'^s_1, \dots, q'^s \widetilde{p}'^s_s, \underbrace{q'^s \frac{1-\sum_{j=1}^s \widetilde{p}'^s_j}{r-s}, \dots, q'^s \frac{1-\sum_{j=1}^s \widetilde{p}'^s_j}{r-s}}_{r-s}, \underbrace{0, \dots, 0}_{2^d-r-1}, 1-q'^s)^\top$$

and the parameter space Θ'_s is

$$\Theta'_s := \left\{ \widetilde{\boldsymbol{p}}'^s = (\widetilde{p}'^s_1, \dots, \widetilde{p}'^s_s, q'^s) \in (0,1)^{s+1} : \widetilde{p}'^s_1 \geq \dots \geq \widetilde{p}'^s_s, \sum_{j=1}^s \widetilde{p}'^s_j < 1 \right\} = \Theta_s \times (0,1).$$

Finally, we define

$$\widetilde{\rho}'^s := \frac{1-\sum_{j=1}^s \widetilde{p}'^s_j}{r-s} \quad \text{for } \widetilde{\boldsymbol{p}}'^s \in \Theta'_s.$$

This ends in the following model.

MODEL M'^s_n: *The family of multinomial distributions* $\{\mathrm{Mult}(n, \boldsymbol{A}'_s(\widetilde{\boldsymbol{p}}'^s)) : \widetilde{\boldsymbol{p}}'^s \in \Theta'_s\}$ *with log-likelihood function*

$$\begin{aligned} \log L_{M'^s_n}(\widetilde{\boldsymbol{p}}'^s \mid \boldsymbol{T}'_n) = \log(n!) - \sum_{j=1}^{2^d} \log(T'_{n,j}!) + \sum_{j=1}^{s} T_{n,j} \log(\widetilde{q}\widetilde{p}^s_j) \\ + \left(\sum_{j=s+1}^{2^d-1} T'_{n,j} \right) \log(\widetilde{q}\widetilde{\rho}^s) + T'_{n,2^d} \log(1-\widetilde{q}) \end{aligned} \tag{2.7}$$

is called Model M'^s_n.

To link the global model with the local model we require further assumptions.

Assumption B.

(B1) Suppose $T'_{n,2^d}$ *and* $\boldsymbol{T}_n$ *are independent, and for* $j = 1, \dots, r$ *we have as* $n \to \infty$,

$$\mathbb{E}\left[\frac{1}{n - T'_{n,2^d}} T'_{n,j} | T'_{n,2^d}\right] = \mathbb{E}\left[\frac{1}{k_n} T_{n,j}(k_n)\right] + o_{\mathbb{P}}(1).$$

(B2) Suppose for $j = 1, \dots, r$ *we have as* $n \to \infty$,

$$\mathbb{E}\left[\frac{1}{(n - T'_{n,2^d})^2} (T'_{n,j})^2 | T'_{n,2^d}\right] = \mathbb{E}\left[\frac{1}{k_n^2} (T_{n,j}(k_n))^2\right] + o_{\mathbb{P}}(1).$$

(B3) There exist constants $K_1, K_2 \in (0, \infty)$ *such that*

$$K_1 < \liminf_{n\to\infty} \frac{nq_n}{k_n} \leq \limsup_{n\to\infty} \frac{nq_n}{k_n} < K_2.$$

Due to the Assumptions (B1) and (B2) the first and second moment of the relative number of extreme observations in the global model and the local model behave similarly. The last Assumption (B3) gives a connection between the asymptotic behavior of q_n and k_n. In particular, it implies $k_n = O(nq_n)$ as $n \to \infty$.

3. Quasi-Akaike information criterion

In the following, we propose an information criterion inspired by the Akaike information criterion and therefore, we refer to as *quasi-Akaike information criterion* (QAIC). Unlike the approach of Meyer and Wintenberger [29], which is based on the likelihood function of a multinomial distribution, our method employs the Gaussian distribution. More specifically, the Akaike information criterion (AIC) introduced by Meyer and Wintenberger [29] for selecting the number of extreme directions is motivated by minimizing the expected Kullback-Leibler (KL) divergence between the true distribution of $\boldsymbol{T}_n(k_n)$ and the multinomial distribution $\mathrm{Mult}(k_n, \widehat{\boldsymbol{p}}_n^s)$ where $\widehat{\boldsymbol{p}}_n^s$ is the MLE given in (2.5). The AIC is defined as

$$\mathrm{AIC}_{k_n}(s) \coloneqq -\log L_{M_{k_n}^s}(\widehat{\boldsymbol{p}}_n^s \,|\, \boldsymbol{T}_n(k_n)) + s, \quad s = 1, \dots, r, \tag{3.1}$$

for fixed k_n. The number s^* of extreme directions is then estimated via

$$\widehat{s}_n = \operatorname*{arg\,min}_{s=1,\dots,r} \mathrm{AIC}_{k_n}(s).$$

However, a limitation of the AIC is that it is not a weakly consistent information criterion which is typically expected in a fixed-dimensional setting as $n \to \infty$ and $d \in \mathbb{N}$ (see Burnham and Anderson [4], Claeskens [9]).

Theorem 3.1. *Suppose Assumption A holds. Then*

$$\lim_{n\to\infty} \mathbb{P}(\mathrm{AIC}_{k_n}(s) > \mathrm{AIC}_{k_n}(s^*)) \begin{cases} < 1 & \text{for } s > s^*, \\ = 1 & \text{for } s < s^*. \end{cases}$$

A key conclusion of Theorem 3.1 is that the AIC has asymptotically a non-vanishing probability of overestimating s^* and hence, it is not a weakly consistent information criterion. The proof of Theorem 3.1, along with all proofs of this section, is relegated to Appendix A.1.

Remark 3.2. Suppose Assumption (A1) is replaced by the condition $\sqrt{k_n\rho_n}(\widehat{r}_n - r) \xrightarrow{\mathbb{P}} 0$ and that the AIC is defined using $\widehat{r}_n$ instead of r. Then

$$\sqrt{k_n\rho_n}\sum_{j=r+1}^{\widehat{r}_n}\left(\frac{T_{n,j}(k_n)}{\rho_n k_n} - 1\right) = o_{\mathbb{P}}(1)$$

and hence, if we follow the proof of Theorem 3.1, we see that the consistency result remains true for this modified AIC, which is finally used in practice.

In contrast, the main advantage of the QAIC, which we introduce next, is that it is a weakly consistent information criterion.

3.1. Quasi Akaike information criterion for the number of directions s

The reason behind employing the likelihood function of a Gaussian distribution for the QAIC is that due to Assumption A the asymptotic behavior as $n \to \infty$,

$$\sqrt{k_n}\mathrm{diag}(\boldsymbol{p}_n^*)^{-1/2}\left(\frac{\boldsymbol{\mathcal{T}}_n}{k_n} - \boldsymbol{p}_n^*\right) \xrightarrow{\mathcal{D}} \mathcal{N}_r(\mathbf{0}_r, \boldsymbol{I}_r)$$

holds, i.e. the asymptotic distribution of $\boldsymbol{\mathcal{T}}_n$ is similar to the distribution of a r-variate normal distribution with mean $k_n\boldsymbol{p}_n^*$ and covariance matrix $k_n\mathrm{diag}(\boldsymbol{p}_n^*)$. Therefore, the idea is to calculate the expected Kullback-Leibler divergence of the true distribution $\mathbb{P}_{\boldsymbol{\mathcal{T}}_n}$ of $\boldsymbol{\mathcal{T}}_n$ with density f_* and the normal distribution $\mathcal{N}_r(k_n\underline{\boldsymbol{A}}_s(\underline{\widetilde{\boldsymbol{p}}}^s), k_n\mathrm{diag}(\underline{\boldsymbol{A}}_s(\underline{\widetilde{\boldsymbol{p}}}^s)))$, $\underline{\widetilde{\boldsymbol{p}}}^s = (\underline{\widetilde{p}}_1^s, \ldots, \underline{\widetilde{p}}_s^s, \underline{\widetilde{\rho}}^s) \in \mathbb{R}_+^{s+1}$, where $\underline{\boldsymbol{A}}_s : \mathbb{R}_+^{s+1} \to \mathbb{R}_+^r$ is defined as

$$\underline{\boldsymbol{A}}_s(z) = \left(z_1, \ldots, z_s, z_{s+1}, \ldots, z_{s+1}\right)^\top.$$

The likelihood function of $\mathcal{N}_r(k_n\underline{\boldsymbol{A}}_s(\underline{\widetilde{\boldsymbol{p}}}^s), k_n\mathrm{diag}(\underline{\boldsymbol{A}}_s(\underline{\widetilde{\boldsymbol{p}}}^s)))$ is denoted by $L_{\mathcal{N}_r}(\underline{\widetilde{\boldsymbol{p}}}^s|\boldsymbol{\mathcal{T}}_n)$. For $\underline{\widetilde{\boldsymbol{p}}}^s$ we use the estimator

$$\begin{aligned}
&\underline{\widehat{\boldsymbol{p}}}_n^s(\widetilde{\boldsymbol{\mathcal{T}}}_n) \coloneqq (\underline{\widehat{p}}_{n,1}^s(\widetilde{\boldsymbol{\mathcal{T}}}_n), \ldots, \underline{\widehat{p}}_{n,s}^s(\widetilde{\boldsymbol{\mathcal{T}}}_n), \underline{\widehat{\rho}}_n^s(\widetilde{\boldsymbol{\mathcal{T}}}_n))^\top \in \mathbb{R}_+^{s+1} \qquad \text{with} \\
&\underline{\widehat{p}}_{n,j}^s(\widetilde{\boldsymbol{\mathcal{T}}}_n) \coloneqq \frac{\widetilde{\mathcal{T}}_{n,j}}{k_n}, \quad j = 1, \ldots, s, \qquad \underline{\widehat{\rho}}_n^s(\widetilde{\boldsymbol{\mathcal{T}}}_n) \coloneqq \frac{1}{r-s}\sum_{j=s+1}^{r}\frac{\widetilde{\mathcal{T}}_{n,j}}{k_n}
\end{aligned} \tag{3.2}$$

where $\widetilde{\boldsymbol{\mathcal{T}}}_n$ is an i.i.d. copy of $\boldsymbol{\mathcal{T}}_n$.

Remark 3.3. It might happen that $\sum_{j=1}^s \underline{\widehat{p}}_{n,j}^s(\widetilde{\boldsymbol{\mathcal{T}}}_n) + (r-s)\underline{\widehat{\rho}}_n^s(\widetilde{\boldsymbol{\mathcal{T}}}_n) \neq 1$. In this case, we have that $\underline{\boldsymbol{A}}_s(\underline{\widehat{\boldsymbol{p}}}_n^s(\widetilde{\boldsymbol{\mathcal{T}}}_n))$ is in general not a probability vector and $(\underline{\widehat{p}}_{n,1}^s(\widetilde{\boldsymbol{\mathcal{T}}}_n), \ldots, \underline{\widehat{p}}_{n,s}^s(\widetilde{\boldsymbol{\mathcal{T}}}_n)) \notin \Theta_s$. But due to Assumption (A4) it holds as $n \to \infty$ that,

$$\frac{\underline{\widehat{p}}_{n,j}^s(\widetilde{\boldsymbol{\mathcal{T}}}_n)}{p_{n,j}} \xrightarrow{\mathbb{P}} 1 \quad \text{and} \quad \frac{\underline{\widehat{\rho}}_n^s(\widetilde{\boldsymbol{\mathcal{T}}}_n)}{\frac{1}{r-s}\sum_{j=s+1}^r p_{n,j}} \xrightarrow{\mathbb{P}} 1,$$

such that $\lim_{n\to\infty}\mathbb{P}((\underline{\widehat{p}}_{n,1}^s(\widetilde{\boldsymbol{\mathcal{T}}}_n), \ldots, \underline{\widehat{p}}_{n,s}^s(\widetilde{\boldsymbol{\mathcal{T}}}_n)) \in \Theta_s) = 1$.

In summary, we calculate

$$\begin{aligned}&\mathbb{E}\left[\mathrm{KL}(\mathbb{P}_{\mathcal{T}_n}, \mathcal{N}_r(k_n \underline{\boldsymbol{A}}_s(\underline{\widetilde{\boldsymbol{p}}}^s), k_n \mathrm{diag}(\underline{\widetilde{\boldsymbol{p}}}^s)))|_{\underline{\widetilde{\boldsymbol{p}}}^s = \underline{\widehat{\boldsymbol{p}}}^s_n(\widetilde{\mathcal{T}}_n)}\right] \\ &\quad = \mathbb{E}\left[\log f_*(\mathcal{T}_n)\right] - \mathbb{E}\left[\log\left(L_{\mathcal{N}_r}(\underline{\widehat{\boldsymbol{p}}}^s_n(\widetilde{\mathcal{T}}_n)|\mathcal{T}_n)\right)\right]. \end{aligned} \tag{3.3}$$

Remark 3.4. The AIC is based on the multinomial distribution whereas the QAIC is based on the multivariate normal distribution. Although it seems at first view that both approaches are different they are related due to local limit theorems for the multinomial distribution as given in Ouimet [30].

Next, we derive an auxiliary result that helps to approximate the second term in (3.3) for $s \geq s^*$.

Proposition 3.5. *Suppose Assumption A holds and $s \geq s^*$. Furthermore, let $\widetilde{\mathcal{T}}_n$ be an independent and identically distributed copy of $\mathcal{T}_n$, and let $\underline{\widehat{\boldsymbol{p}}}^s_n(\widetilde{\mathcal{T}}_n)$ be the estimator in (3.2) and similarly we define $\underline{\widehat{\boldsymbol{p}}}^s_n(\mathcal{T}_n)$. Then there exists a random variable Y with $\mathbb{E}[Y] = 0$ such that as $n \to \infty$,*

$$\begin{aligned}&\log L_{\mathcal{N}_r}(\underline{\widehat{\boldsymbol{p}}}^s_n(\widetilde{\mathcal{T}}_n) \mid \mathcal{T}_n) + \frac{1}{2} r \log(2\pi) + \frac{1}{2} r \log(k_n) \\ &\quad + \frac{1}{2}\sum_{j=1}^{s} \log(\widehat{p}^s_{n,j}(\mathcal{T}_n)) + \frac{1}{2}(r-s)\log(\widehat{\rho}^s_n(\mathcal{T}_n)) + \frac{r+s+1}{2} \xrightarrow{\mathcal{D}} Y. \end{aligned}$$

Therefore, for $s \geq s^*$ we approximate the second term in (3.3) by

$$\begin{aligned}&-\mathbb{E}\left[\log L_{\mathcal{N}_r}(\underline{\widehat{\boldsymbol{p}}}^s_n(\widetilde{\mathcal{T}}_n) \mid \mathcal{T}_n)\right] \\ &\approx \frac{1}{2}\mathbb{E}\left[r\log(2\pi) + r\log(k_n) + \sum_{j=1}^{s}\log(\widehat{p}^s_{n,j}(\mathcal{T}_n)) + (r-s)\log(\widehat{\rho}^s_n(\mathcal{T}_n)) + r + s + 1\right]\end{aligned}$$

and neglect the expectation. The first term $\mathbb{E}\left[\log f_*(\mathcal{T}_n)\right]$ in (3.3) and the +1 do not influence the choice of the model, therefore we skip them. This leads to the following definition of the theoretic quasi-information criterion for $s \geq s^*$,

$$\mathrm{QAIC}'_{k_n}(s) := r\log(2\pi) + r\log(k_n) + \sum_{j=1}^{s}\log(\widehat{p}^s_{n,j}(\mathcal{T}_n)) + (r-s)\log(\widehat{\rho}^s_n(\mathcal{T}_n)) + r + s.$$

If $s < s^*$ this information criterion works as well since

$$\begin{aligned}&\sum_{j=1}^{s}\log(\widehat{p}^s_{n,j}(\mathcal{T}_n)) + (r-s)\log(\widehat{\rho}^s_n(\mathcal{T}_n)) \\ &\qquad \xrightarrow{\mathbb{P}} \sum_{j=1}^{s}\log(p_j) + (r-s)\log\left(\frac{\sum_{j=s+1}^{s^*} p_j}{r-s}\right) > -\infty\end{aligned}$$

and for $s > s^*$ we have

$$\sum_{j=1}^{s}\log(\widehat{p}^s_{n,j}(\mathcal{T}_n)) + (r-s)\log(\widehat{\rho}^s_n(\mathcal{T}_n)) \xrightarrow{\mathbb{P}} -\infty.$$

Therefore, the information criterion does not select $s < s^*$.

Moreover, since

$$\sum_{j=1}^{s} \log(\widehat{\underline{p}}^s_{n,j}(\mathcal{T}_n)) + (r-s)\log(\widehat{\underline{\rho}}^s_n(\mathcal{T}_n)) - \sum_{j=1}^{s} \log(\widehat{\underline{p}}^s_{n,j}(\boldsymbol{T}_n(k_n))) + -(r-s)\log(\widehat{\underline{\rho}}^s_n(\boldsymbol{T}_n(k_n))) \xrightarrow{\mathbb{P}} 0$$

the choice between estimator $\widehat{\underline{\boldsymbol{p}}}^s_n(\mathcal{T}_n)$ or $\widehat{\boldsymbol{p}}^s_n = \widehat{\boldsymbol{p}}^s_n(\boldsymbol{T}_n(k_n)) \in \Theta_s$ with $\widehat{\rho}^s_n = \widehat{\rho}^s_n(\boldsymbol{T}_n(k_n))$ does not significantly change the outcome, so either can be used. Since in applications u_n and hence, $\widehat{\underline{\boldsymbol{p}}}^s_n(\mathcal{T}_n)$ is unknown, we finally define the information criterion based on the estimators $\widehat{\boldsymbol{p}}^s_n$ and $\widehat{\rho}^s_n$.

Definition 3.6. For the number of extreme directions s with fixed k_n the *quasi Akaike information criterion* (QAIC) is defined as

$$\mathrm{QAIC}_{k_n}(s) \coloneqq r\log(2\pi) + r\log(k_n) + \sum_{j=1}^{s} \log(\widehat{p}^s_{n,j}) + (r-s)\log(\widehat{\rho}^s_n) + r + s$$

for $s = 1, \ldots, r$ and an estimator for s^* is $\widehat{s}_n \coloneqq \arg\min_{1 \le s \le r} \mathrm{QAIC}_{k_n}(s)$.

Remark 3.7.

(a) During the derivation of the QAIC we assumed that r is constant and hence, it should not influence the optimal value of the QAIC. However, the simulation study shows that in applications r has a significant impact on the performance of the QAIC because in practice r depends on k_n.

(b) The derivation of a QAIC with an estimator based on the likelihood function of the normal distribution $L_{\mathcal{N}_r}$ is possible with similar results but leads to a more elaborate and longer calculation. In this case, the estimator is given by

$$\widehat{p}^G_{n,j} = \frac{-1}{2k_n} + \sqrt{\frac{1}{4k_n^2} + \frac{T_{n,j}(k_n)^2}{k_n^2}}, \quad j = 1, \ldots, s,$$

$$\widehat{\rho}^G_n = \frac{-1}{2k_n} + \sqrt{\frac{1}{4k_n^2} + \frac{1}{r-s}\sum_{j=s+1}^{r} \frac{T_{n,j}(k_n)^2}{k_n^2}}.$$

The performance of both approaches is similar and therefore only QAIC is included in the simulation study.

Theorem 3.8. *Suppose Assumption A holds. Then*

$$\lim_{n\to\infty} \mathbb{P}(\mathrm{QAIC}_{k_n}(s) - \mathrm{QAIC}_{k_n}(s^*) > 0) = 1 \quad \text{for } s \neq s^*.$$

Compared to the AIC, the QAIC has the advantage that it is weakly consistent for fixed k_n in contrast to the AIC.

Remark 3.9. Suppose Assumption (A1) is replaced by the condition $\widehat{r}_n \xrightarrow{\mathbb{P}} r$ and that the QAIC is defined using $\widehat{r}_n$ instead of r. Then the consistency result remains true for this modified QAIC. Note that here a weaker condition is used as for the AIC in Remark 3.2, where we required $\sqrt{k_n\rho_n}(\widehat{r}_n - r) \xrightarrow{\mathbb{P}} 0$.

3.2. Quasi Akaike information criterion for the threshold k_n

For the QAIC for the threshold k_n we follow the definition of the global model for the AIC in Meyer and Wintenberger [29] which is defined as

$$\mathrm{AIC}_{n,s}(k_n) \coloneqq \frac{\mathrm{AIC}_{k_n}(s)}{k_n} + \frac{k_n}{n}$$

with $\mathrm{AIC}_{k_n}(s)$ as in (3.1). However, since we consider two times the negative likelihood instead of just the negative likelihood we include additionally the factor $1/2$ and obtain the following information criterion.

Definition 3.10. For the number of exceedances k_n the *quasi-Akaike information criterion* (QAIC) *for the threshold* k_n for the Model $M_n'^s$ is defined as

$$\begin{aligned}\mathrm{QAIC}_{n,s}(k_n) &\coloneqq \frac{\mathrm{QAIC}_{k_n}(s)}{2k_n} + \frac{k_n}{n} \\ &= \frac{r\log(2\pi) + r\log(k_n) + \sum_{j=1}^{s}\log(\widehat{p}^{\,s}_{n,j}) + (r-s)\log(\widehat{\rho}^{\,s}_{n,j}) + r + s}{2k_n} + \frac{k_n}{n}\end{aligned}$$

for $k_n = 1,\dots,n$ with estimator $\widehat{k}_n \coloneqq \arg\min_{k_n\in K}\{\min_{1\le s\le r}\mathrm{QAIC}_{n,s}(k_n)\}$ for $K \subset \{1,\dots,n\}$.

Remark 3.11. An interpretation of this information criterion is as follows. The division by k_n can be seen as a weight, which is assigned to a pair (s,k_n). Therefore, when k_n is large, the weight of the corresponding model gets smaller. Also, k_n/n corresponds to the relative proportion of extreme observations and acts as a penalty for increasing k_n.

4. Mean squared error information criterion

Next, we explore an information criterion based on the mean squared error (MSE) for both the number of directions s in Section 4.1 as well as for the threshold k_n in Section 4.2, which performs in particular well for a small number of observations. The proofs of this section are moved to Appendix A.2.

4.1. Mean squared error information criterion for the number of extreme directions s

The basic idea of the AIC is to minimize the Kullback-Leibler distance of the true distribution and a parametric family of distributions. This minimum is approximated by the expected Kullback-Leibler distance of the true distribution and the estimated distribution as is done in (3.3). In the following, we use the same ideas but instead of using the Kullback Leibler distance we use the normalized mean-squared error (MSE) of the parameter estimator and find an approximation of

$$\mathrm{MSE}_{k_n}(s) \coloneqq \mathbb{E}\left[\ell^2\big(\underline{\widehat{\boldsymbol{p}}}^{\,s}_n(\widetilde{\boldsymbol{T}}_n(k_n))|\boldsymbol{T}_n\big)\right] \tag{4.1}$$

instead of $\mathbb{E}\left[\log L_{\mathcal{N}_r}(\underline{\widehat{\boldsymbol{p}}}^{\,s}_n(\widetilde{\boldsymbol{\mathcal{T}}}_n)\,|\,\boldsymbol{T}_n(k_n))\right]$ as is done in (3.3), where $\widetilde{\boldsymbol{T}}_n(k_n)$ is an independent and identically distributed copy of $\boldsymbol{T}_n(k_n)$ and

$$\ell^2\big(\underline{\widetilde{\boldsymbol{p}}}^{\,s}|\boldsymbol{T}_n(k_n)\big) \coloneqq \left\|\sqrt{k_n}\,\mathrm{diag}\big(\underline{\boldsymbol{A}}_s(\underline{\widetilde{\boldsymbol{p}}}^{\,s})\big)^{-1/2}\left(\frac{\boldsymbol{T}_n(k_n)}{k_n} - \underline{\boldsymbol{A}}_s(\underline{\widetilde{\boldsymbol{p}}}^{\,s})\right)\right\|_2^2$$

$$= \sum_{j=1}^{s} \frac{k_n}{\widetilde{\underline{p}}_j^s} \left(\frac{T_{n,j(k_n)}}{k_n} - \widetilde{\underline{p}}_j^s \right)^2 + \frac{k_n}{\widetilde{\underline{\rho}}^s} \sum_{j=s+1}^{r} \left(\frac{T_{n,j}(k_n)}{k_n} - \widetilde{\underline{\rho}}^s \right)^2$$

for $\widetilde{\underline{\boldsymbol{p}}}^s = (\widetilde{\underline{p}}_1^s, \ldots, \widetilde{\underline{p}}_s^s, \widetilde{\underline{\rho}}^s) \in \mathbb{R}_+^{s+1}$. Note, if in Assumption (A3) not only the weak convergence but also the componentwise L_1 convergence holds, then

$$\lim_{n\to\infty} \mathbb{E}\left[\ell^2\big((p_{n,1}, \ldots, p_{n,s^*}, \rho_n) | \boldsymbol{T}_n(k_n)\big)\right] = r - 1$$

which motivates this approach. First, we derive an auxiliary result that helps to approximate $\ell^2\big(\widehat{\underline{\boldsymbol{p}}}_n^s(\widetilde{\boldsymbol{T}}_n(k_n)) | \boldsymbol{T}_n(k_n)\big)$.

Theorem 4.1. *Suppose Assumption A holds and $s \geq s^*$. Furthermore, let $\widetilde{\boldsymbol{T}}_n(k_n)$ be an independent and identically distributed copy of $\boldsymbol{T}_n(k_n)$, and let $\widehat{\underline{\boldsymbol{p}}}_n^s(\widetilde{\boldsymbol{T}}_n(k_n))$ be the estimator in (3.2). Similarly, we define $\widehat{\underline{\boldsymbol{p}}}_n^s(\boldsymbol{T}_n(k_n))$. Then there exists a random variable Y with $\mathbb{E}[Y] = 0$ such that as $n \to \infty$,*

$$\ell^2\big(\widehat{\underline{\boldsymbol{p}}}_n^s(\widetilde{\boldsymbol{T}}_n(k_n)) | \boldsymbol{T}_n(k_n)\big) - \frac{k_n}{\widehat{\underline{\rho}}_n^s(\boldsymbol{T}_n(k_n))} \sum_{j=s+1}^{r} \left(\frac{T_{n,j}(k_n)}{k_n} - \widehat{\underline{\rho}}_n^s(\boldsymbol{T}_n(k_n)) \right)^2 - 2s \xrightarrow{\mathcal{D}} Y.$$

Therefore, for $s \geq s^*$ we approximate (4.1) by

$$\mathrm{MSE}_{k_n}(s) \approx \mathbb{E}\left[\frac{k_n}{\widehat{\underline{\rho}}_n^s(\boldsymbol{T}_n(k_n))} \sum_{j=s+1}^{r} \left(\frac{T_{n,j}(k_n)}{k_n} - \widehat{\underline{\rho}}_n^s(\boldsymbol{T}_n(k_n)) \right)^2 + 2s \right].$$

Analogously to Section 3, we neglect the expectation, which leads to the following information criterion.

Definition 4.2. For the number of extreme directions s with fixed k_n the *mean squared error information criterion* (MSEIC) is defined as

$$\mathrm{MSEIC}_{k_n}(s) := \frac{k_n}{\sum_{l=s+1}^{r} \frac{T_{n,l}(k_n)}{k_n(r-s)}} \sum_{j=s+1}^{r} \left(\frac{T_{n,j}(k_n)}{k_n} - \sum_{i=s+1}^{r} \frac{T_{n,i}(k_n)}{k_n(r-s)} \right)^2 + 2s,$$

for $s = 1, \ldots, r-1$ with $\mathrm{MSEIC}_{k_n}(r) := 2r$. An estimator for s^* is defined by $\widehat{s}_n := \arg\min_{1 \leq s \leq r} \mathrm{MSEIC}_{k_n}(s)$.

Theorem 4.3. *Suppose Assumption A holds. Then*

$$\lim_{n\to\infty} \mathbb{P}(\mathrm{MSEIC}_{k_n}(s) > \mathrm{MSEIC}_{k_n}(s^*)) \begin{cases} < 1 & \text{for } s > s^*, \\ = 1 & \text{for } s < s^*. \end{cases}$$

In particular, for $s < s^*$ this information criterion is consistent, but unfortunately not for $s > s^*$. However, this is not surprising because the basic ideas are related to the AIC which is also not a consistent information criterion. However, the simulation study in Section 6 shows that MSEIC performs extremely good in practice.

4.2. Mean squared error information criterion for the threshold k_n

Now, we extend the information criterion MSEIC to choose the optimal threshold k_n. Therefore, we use not only our knowledge about the extreme observations but also our knowledge of the non-extreme observations, similarly to the global model $M_n'^s$, only that there is no distributional assumption. As before we assume here that $\boldsymbol{T}'_{n,\{1,\dots,r\}}$ pertains the information about the observed extreme directions and $T'_{n,2^d}$ the non-extreme observations, where $T'_{n,2^d}$ is assumed to be binomially distributed. The MSE information criterion for the threshold k_n is then defined as weighted MSE

$$\begin{aligned}\mathrm{MSE}'^s_n :=& \mathbb{E}\left[q'\mathbb{E}\left[\left\|\sqrt{n-T'_{n,2^d}}\,\mathrm{diag}(\boldsymbol{p}')^{-1/2}\left(\frac{\boldsymbol{T}'_{n,\{1,\dots,r\}}}{n-T'_{n,2^d}}-\boldsymbol{p}'\right)\right\|_2^2\right]\Big|_{\boldsymbol{p}'=\widehat{\boldsymbol{p}}_n(\frac{\widetilde{\boldsymbol{T}}_n(k_n)}{k_n}),q'=\frac{k_n}{n}}\right]\\ &+\mathbb{E}\left[(1-q')\mathbb{E}\left[\left\|\sqrt{n}(q'(1-q'))^{-1/2}\left(\frac{T'_{n,2^d}}{n}-(1-q')\right)\right\|_2^2\right]\Big|_{q'=\frac{k_n}{n}}\right]\end{aligned} \tag{4.2}$$

with weight q' for the estimation of the probabilities of extreme directions and weight $(1-q')$ for the estimation of the probability of non-extremes. Since we want to make statements about the optimal choice of k_n which models the number of extreme directions, the weight in the estimation of the probabilities of the extreme directions is chosen higher. A connection between the MSE information criterion for the threshold k_n and the MSE information criterion for the number of extreme directions s exists through the following theorem.

Theorem 4.4. *Suppose Assumptions (B1), (B2) and* $k_n(1-\frac{nq_n}{k_n})^2 \to 0$ *as* $n\to\infty$. *Then*

$$\mathrm{MSE}'^s_n = q_n\left(\mathrm{MSE}_{k_n}(s)+\frac{n}{k_n}+no\left(\frac{1}{nq_n}\right)\right).$$

Since q_n is not influenced by k_n and $o((nq_n)^{-1})$ is of a smaller order than $1/k_n$ by Assumption (B3), we neglect q_n and the last term. Consequently, we define the following information criterion.

Definition 4.5. For the number of exceedances k_n the *mean squared error information criterion (*MSEIC*) for the threshold* k_n for the Model $M_n'^s$ is defined as

$$\mathrm{MSEIC}_{n,s}(k_n) := \mathrm{MSEIC}_{k_n}(s)+\frac{n}{k_n},\quad k_n=1,\dots,n,$$

with estimator $\widehat{k}_n := \arg\min_{k_n\in K}\{\min_{1\le s\le r}\mathrm{MSEIC}_{n,s}(k_n)\}$ for $K\subset\{1,\dots,n\}$.

Remark 4.6. The general structure of this threshold information criterion differs from the other derived information criteria for the threshold selection as

$$\mathrm{AIC}_{n,s}(k_n)=\frac{\mathrm{AIC}_{k_n}(s)}{k_n}+\frac{k_n}{n}\quad\text{and}\quad \mathrm{QAIC}_{n,s}(k_n)=\frac{\mathrm{QAIC}_{k_n}(s)}{2k_n}+\frac{k_n}{n}.$$

Therefore, we performed a simulation study with the criterion $\mathrm{MSEIC}_{k_n}(s)/k_n+k_n/n$, defined analog to $\mathrm{AIC}_{n,s}(k_n)$. The simulation study confirms that this choice of information criteria is not the suitable choice. The result is not surprising, since MSEIC is not based on a likelihood-based approach.

5. Bayesian information criterion

In addition to the AIC, the Bayesian information criterion (BIC) introduced in Schwarz [35] is the most popular one. The basic idea of the BIC is to find the model with the highest posterior probability given the data. First, we derive a BIC for s in Section 5.1 and then for k_n in Section 5.2. The proofs of this section can be found in Appendix A.3.

5.1. Bayesian information criterion for the number of extreme directions s

In the following, we derive a BIC for s by bounding the posterior probability as in Cavanaugh and Neath [7]. Therefore, we assume throughout this section Model $M_{k_n}^s$ and use the following notation. Let $\mathbb{Q}$ be a discrete prior distribution over the set of models $\{M_{k_n}^s : s = 1, \dots, r\}$, $g(\cdot \mid M_{k_n}^s)$ be the prior density over the parameter space Θ_s given Model $M_{k_n}^s$, $L_{M_{k_n}^s}(\cdot \mid \boldsymbol{T}_n(k_n))$ be the likelihood function of Model $M_{k_n}^s$ if we observe $\boldsymbol{T}_n(k_n)$ and f be the (unknown) marginal probability of $\boldsymbol{T}_n(k_n)$. Given the data $\boldsymbol{T}_n(k_n)$ the goal is to determine the Model $M_{k_n}^s$ with the highest posterior probability $\mathbb{P}(M_{k_n}^s | \boldsymbol{T}_n(k_n))$ for $s = 1, \dots, r$. Therefore, note that Bayes Theorem yields for the posterior density for $M_{k_n}^s$ and $\widetilde{\boldsymbol{p}}^s$

$$h((M_{k_n}^s, \widetilde{\boldsymbol{p}}^s) \mid \boldsymbol{T}_n(k_n)) = \frac{L_{M_{k_n}^s}(\widetilde{\boldsymbol{p}}^s \mid \boldsymbol{T}_n(k_n)) g(\widetilde{\boldsymbol{p}}^s \mid M_{k_n}^s)\, \mathbb{Q}(M_{k_n}^s)}{f(\boldsymbol{T}_n(k_n))}.$$

Hence, the posterior probability for $M_{k_n}^s$ is

$$\mathbb{P}(M_{k_n}^s | \boldsymbol{T}_n(k_n)) = \frac{\mathbb{Q}(M_{k_n}^s) \int_{\Theta_s} L_{M_{k_n}^s}(\widetilde{\boldsymbol{p}}^s \mid \boldsymbol{T}_n(k_n)) g(\widetilde{\boldsymbol{p}}^s \mid M_{k_n}^s)\, \mathrm{d}\widetilde{\boldsymbol{p}}^s}{f(\boldsymbol{T}_n(k_n))}.$$

Consequently maximizing the posterior probability is equivalent to minimizing

$$\begin{aligned} -2 \log \mathbb{P}(M_{k_n}^s | \boldsymbol{T}_n(k_n)) =& 2 \log f(\boldsymbol{T}_n(k_n)) - 2 \log \mathbb{Q}(M_{k_n}^s) \\ & - 2 \log\Big(\int_{\Theta_s} L_{M_{k_n}^s}(\widetilde{\boldsymbol{p}}^s \mid \boldsymbol{T}_n(k_n)) g(\widetilde{\boldsymbol{p}}^s \mid M_{k_n}^s)\, \mathrm{d}\widetilde{\boldsymbol{p}}^s \Big). \end{aligned} \tag{5.1}$$

For the derivation of the BIC, we require further assumptions.

Assumption C. *For any $s \in \{1, \dots, r\}$ we assume the following:*

(C1) There exist constants $0 < b \leq B < \infty$ such that the prior density $g(\cdot \mid M_{k_n}^s)$ on Θ_s satisfies

$$b \leq g(\widetilde{\boldsymbol{p}}^s \mid M_{k_n}^s) \leq B \quad \text{for all } \widetilde{\boldsymbol{p}}^s \in \Theta_s.$$

(C2) The prior distribution $\mathbb{Q}$ is a uniform distribution on $\{M_{k_n}^s : s = 1, \dots, r\}$, i.e. $\mathbb{Q}(M_{k_n}^s) = \frac{1}{r}$ for $s = 1, \dots, r$.

(C3) $k_n \rho_n^{5/3} \to \infty$ and $k_n \rho_n^2 \to 0$.

Remark 5.1.

(a) Both Assumptions (C1) and (C2) are assumptions on prior distributions, and they reflect that we have no prior information in advance. The lower bound of Assumption (C1) can be relaxed since we require only a lower bound in the neighborhood of $\widehat{p}_n^s$. However, it has been omitted in this paper for the sake of brevity.

(b) The assumption on the uniform distribution on the set of all possible models in (C2) is an uninformative prior distribution where all models have the same probability. Thus, the term $-2\log \mathbb{Q}(M_{k_n}^s) = 2\log r$ in (5.1) is independent of s and has, from a theoretical point of view, no influence on the information criterion. Of course, it is possible to use a prior distribution depending on s but then the BIC receives an additional penalty term.

(c) The assumption $k_n \rho_n^{5/3} \to \infty$ in (C3) ensures that ρ_n does not converge to zero too quickly.

The next theorem gives an upper bound for

$$-2\log \mathbb{E}_{g_s}[L_{M_{k_n}^s}(\widetilde{\boldsymbol{p}}^s \mid \boldsymbol{T}_n(k_n))] \coloneqq -2\log \int_{\Theta_s} L_{M_{k_n}^s}(\widetilde{\boldsymbol{p}}^s \mid \boldsymbol{T}_n(k_n)) g(\widetilde{\boldsymbol{p}}^s \mid M_{k_n}^s)\, \mathrm{d}\widetilde{\boldsymbol{p}}^s,$$

whereby $\mathbb{E}_{g_s}$ denotes the conditional expectation regarding the prior density $g(\cdot \mid M_{k_n}^s)$ on Θ_s. This results then in an upper bound for the negative log posterior probability of the s-th Model $M_{k_n}^s$ given $\boldsymbol{T}_n(k_n)$.

Theorem 5.2. *Suppose Assumptions A, (C3) and (C1) hold. Then the inequality*

$$\begin{aligned}&-2\log \mathbb{E}_{g_s}[L_{M_{k_n}^s}(\widetilde{\boldsymbol{p}}^s \mid \boldsymbol{T}_n(k_n))]\\ &\quad\le -2\log L_{M_{k_n}^s}(\widehat{\boldsymbol{p}}_n^s \mid \boldsymbol{T}_n(k_n)) - s\log(2\pi) + 2s\log\left(k_n\sqrt{\frac{r}{r-s}}\right) - 2\log b + o_{\mathbb{P}}(1)\end{aligned}$$

as $n \to \infty$ holds.

Plugging in Assumption (C2) and the upper bound in Theorem 5.2 in (5.1) results in

$$\begin{aligned}&-2\log \mathbb{P}(M_{k_n}^s(k_n) \mid \boldsymbol{T}_n(k_n))\\ &\quad= 2\log f(\boldsymbol{T}_n(k_n)) + 2\log r - 2\log \mathbb{E}_{g_s}[L_{M_{k_n}^s}(\widetilde{\boldsymbol{p}}^s \mid \boldsymbol{T}_n(k_n))]\\ &\quad\le -2\log L_{M_{k_n}^s}(\widehat{\boldsymbol{p}}_n^s \mid \boldsymbol{T}_n(k_n)) - s\log(2\pi) + 2s\log\left(k_n\sqrt{\frac{r}{r-s}}\right)\\ &\qquad + 2\log f(\boldsymbol{T}_n(k_n)) - 2\log b + 2\log r + o_{\mathbb{P}}(1).\end{aligned}$$

This motivates the definition of the following information criterion, where we neglect the terms $2\log f(\boldsymbol{T}_n(k_n)) - 2\log b + 2\log r$ as they are not influenced by s.

Definition 5.3. For the number of extreme directions s with fixed k_n the *Bayesian information criterion concerning the upper bound* (BICU) is defined as

$$\mathrm{BICU}_{k_n}(s) \coloneqq -2\log L_{M_{k_n}^s}(\widehat{\boldsymbol{p}}_n^s \mid \boldsymbol{T}_n(k_n)) + 2s\log(k_n) + s\log\left(\frac{r}{2\pi(r-s)}\right),$$

for $s = 1, \dots, r-1$ and an estimator for s^* is $\widehat{s}_n \coloneqq \arg\min_{1\le s\le r-1} \mathrm{BICU}_{k_n}(s)$.

Motivated by the BICU, which is based on the largest eigenvalue $\lambda_{n,1}$ from Lemma A.8, we define a BIC based on a lower bound for the posterior distribution by using the smallest eigenvalue $\lambda_{n,2} = k_n / T_{n,1}(k_n)$ from Lemma A.8.

Definition 5.4. For the number of extreme directions s with fixed k_n the *Bayesian information criterion concerning the lower bound* (BICL) for Model $M_{k_n}^s$ is defined as

$$\mathrm{BICL}_{k_n}(s) \coloneqq -2\log L_{M_{k_n}^s}(\widehat{\boldsymbol{p}}_n^s \mid \boldsymbol{T}_n(k_n)) + s\log(k_n) + s\log\left(\frac{k_n}{2\pi T_{n,1}(k_n)}\right),\ s = 1, \dots, r,$$

and an estimator for s^* is $\widehat{s}_n \coloneqq \arg\min_{1\le s\le r} \mathrm{BICL}_{k_n}(s)$.

Theorem 5.5. *Suppose Assumption A holds. Then*

(a) $\lim_{n\to\infty} \mathbb{P}(\mathrm{BICU}_{k_n}(s) > \mathrm{BICU}_{k_n}(s^*)) = 1 \quad$ *for* $s \neq s^*$,
(b) $\lim_{n\to\infty} \mathbb{P}(\mathrm{BICL}_{k_n}(s) > \mathrm{BICL}_{k_n}(s^*)) = 1 \quad$ *for* $s \neq s^*$.

Thus, in contrast to the AIC criterion, both information criteria are weakly consistent and select asymptotically with probability 1 the true Model $M_{k_n}^{s^*}$. This is also a typical property of Bayesian information criteria (see Burnham and Anderson [4], Claeskens [9]).

5.2. Bayesian information criterion for the threshold k_n

In the following, we determine an upper bound for the posterior probability of the global Model $M_n'^s$ analog to the previous Section 5.1 using the following assumptions.

Assumption D. *Suppose the following statements hold.*

(D1) There exist constants $0 < b' \leq B' < \infty$ *such that the prior density* $g'(\cdot \mid M_n'^s)$ *on* Θ_s' *satisfies*

$$b' \leq g'(\widetilde{\boldsymbol{p}}'^s | M_n'^s) \leq B' \quad \text{for all } \widetilde{\boldsymbol{p}}'^s \in \Theta_s'.$$

(D2) The prior distribution $\mathbb{Q}'$ *is a uniform distribution on* $\{M_n'^s : s = 1, \ldots, r\}$, *i.e.* $\mathbb{Q}'(M_n'^s) = \frac{1}{r}$ *for* $s = 1, \ldots, r$.

(D3) $\lim_{n\to\infty} nq_n^{5/3} = \infty$ *and* $\lim_{n\to\infty} nq_n^2 = 0$.

(D4) For $\mathbb{E}_\lambda[L_{M^s_{n-T'_{n,2^d}}}(\widetilde{\boldsymbol{p}}^s \mid \boldsymbol{T}'_{n,\{1,\ldots,r\}})] := \int_{\Theta_s} L_{M^s_{n-T'_{n,2^d}}}(\widetilde{\boldsymbol{p}}^s \mid \boldsymbol{T}'_{n,\{1,\ldots,r\}})\, \mathrm{d}\widetilde{\boldsymbol{p}}^s$ *the following upper bound*

$$\begin{aligned}
&\mathbb{E}\Big[-2\log\mathbb{E}_\lambda[L_{M^s_{n-T'_{n,2^d}}}(\widetilde{\boldsymbol{p}}^s \mid \boldsymbol{T}'_{n,\{1,\ldots,r\}})]\Big] \\
&\quad \leq \mathbb{E}\Big[\mathbb{E}\Big[-2\log L_{M^s_{n-T'_{n,2^d}}}(\widehat{\boldsymbol{p}}^s_n(\boldsymbol{T}'_{n,\{1,\ldots,r\}}) \mid \boldsymbol{T}'_{n,\{1,\ldots,r\}})\Big| T'_{n,2^d}\Big]\Big] \\
&\qquad + 2s\mathbb{E}\Big[\log\Big((n - T'_{n,2^d})\sqrt{\frac{r}{r-s}}\Big)\Big] - s\log(2\pi) + o(1)
\end{aligned}$$

holds.

Remark 5.6.

(a) Assumptions (D1) and (D2) in the global model correspond to the Assumptions (C1) and (C2) in the local model. Assumption (D3) is the counterpart to Assumption (C3) for the binomial part of the likelihood function in the global model.
(b) Assumption (D3) ensures a suitable convergence rate of q_n and implies $nq_n \to \infty$. For example $q_n := n^{-11/20}$ fulfills the conditions of Assumption (D3).
(c) Assumption (C3) for the local model is required for the proof of Theorem 5.2. Assumption (D4) for the global model is motivated from Theorem 5.2 and (2.6). Because we then obtain directly

$$\mathbb{E}\Big[-2\log\mathbb{E}_\lambda[L_{M^s_{n-T'_{n,2^d}}}(\widetilde{\boldsymbol{p}}^s \mid \boldsymbol{T}'_{n,\{1,\ldots,r\}})] | T'_{n,2^d} = k_n\Big]$$

$$\leq \mathbb{E}\Big[-2\log L_{M^s_{n-T'_{n,2^d}}}(\widehat{\boldsymbol{p}}^s_n(\boldsymbol{T}'_{n,\{1,\dots,r\}}) \mid \boldsymbol{T}'_{n,\{1,\dots,r\}}) \Big| T'_{n,2^d} = k_n \Big]$$
$$+ 2s\mathbb{E}\Big[\log\left((n - T'_{n,2^d}) \sqrt{\frac{r}{r-s}} \right) | T'_{n,2^d} = k_n \Big] - s\log(2\pi) + o(1)$$

for k_n satisfying the assumptions of the previous section and $T'_{n,1} \geq T'_{n,2} \geq \ldots \geq T'_{n,r}$. Assumption (D4) for the global model is only a slightly stronger assumption than Assumption (C3) for the local model.

In analogy to Section 5.1, the goal is to derive asymptotic bounds for $-2\log \mathbb{P}(M'^s_n | \boldsymbol{T}'_n)$ which we obtain through upper bounds for

$$-2\log \mathbb{E}_{g'_s}[L_{M'^s_n}(\widetilde{\boldsymbol{p}}'^s | \boldsymbol{T}'_n)] \coloneqq -2\log\left\{ \int_{\Theta'_s} L_{M'^s_n}(\widetilde{\boldsymbol{p}}'^s \mid \boldsymbol{T}'_n) \cdot g'(\widetilde{\boldsymbol{p}}'^s | M'^s_n)\, \mathrm{d}\widetilde{\boldsymbol{p}}'^s \right\}, \tag{5.2}$$

where $\mathbb{E}_{g'_s}$ denotes the conditional expectation with respect to the prior density $g'(\cdot | M'^s_n)$ on Θ'_s given Model M'^s_n.

Theorem 5.7. *Under Assumptions (B1), (B3) and D the asymptotic upper bound as $n \to \infty$,*

$$-2\mathbb{E}\big[\log \mathbb{E}_{g'_s}[L_{M'^s_n}(\widetilde{\boldsymbol{p}}'^s \mid \boldsymbol{T}'_n)] \big]$$
$$\leq nq_n \left(-2\frac{\mathbb{E}[\log L_{M^s_{k_n}}(\widehat{\boldsymbol{p}}^s_n \mid \boldsymbol{T}_n(k_n))]}{k_n} + 2\frac{s}{nq_n}\log\left(k_n \sqrt{\frac{r}{2\pi(r-s)}} \right) + \frac{2\log(n)}{nq_n} + C \right)$$

holds, where $C > 0$ is a constant independent of s and n.

Compared to Theorem 5.2 in the previous section, we take additionally the expectation in Theorem 5.7 to achieve a connection between the global model and the local model.

Theorem 5.7 motivates the definition of the following information criterion, where the expectation is omitted, the inequality is divided by nq_n and the term $2\log(b')/(nq_n)$ as well as C are neglected as they are either constant concerning s or converge to zero uniformly.

Definition 5.8. For the number of exceedances k_n the *Bayesian information criterion concerning the upper bound (*BICU*) for the threshold k_n* for Model M'^s_n is defined as

$$\mathrm{BICU}_{n,s}(k_n) \coloneqq \frac{-2\log L_{M^s_{k_n}}(\widehat{\boldsymbol{p}}^s_n \mid \boldsymbol{T}_n(k_n)) + 2s\log(k_n) + s\log\left(\frac{r}{2\pi(r-s)}\right)}{k_n} + \frac{\log(n^2)}{k_n}$$
$$= \frac{\mathrm{BICU}_{k_n}(s)}{k_n} + \frac{\log(n^2)}{k_n},$$

for $k_n = 1, \ldots, n$, with estimator $\widehat{k}_n \coloneqq \arg\min_{k_n \in K}\{\min_{1 \leq s \leq r} \mathrm{BICU}_{n,s}(k_n)\}$ for $K \subset \{1, \ldots, n\}$ for k_n.

Similarly to Definition 5.4 we also define the Bayesian information criterion based on the lower bound for the threshold k_n.

Definition 5.9. For the number of exceedances k_n the *Bayesian information criterion concerning the lower bound (*BICL*) for the threshold k_n* for Model M'^s_n is defined as

$$\mathrm{BICL}_{n,s}(k_n) \coloneqq \frac{\mathrm{BICL}_{k_n}(s)}{k_n} + \frac{\log(n^2)}{k_n}, \quad k_n = 1, \ldots, n,$$

with estimator $\widehat{k}_n \coloneqq \arg\min_{k_n \in K}\{\min_{1 \leq s \leq r} \mathrm{BICL}_{n,s}(k_n)\}$ for $K \subset \{1, \ldots, n\}$.

6. Simulation study

In this section, we compare the performance of the different information criteria through a simulation study. Therefore, we simulate n times a multivariate regularly varying random vector $\boldsymbol{X}$ of dimension d. For the distribution of $\boldsymbol{X}$, we distinguish two cases: Either $\boldsymbol{X}$ exhibits asymptotic independence (Section 6.2) or asymptotic dependence (Section 6.3); these examples can be found in Meyer and Wintenberger [29] as well. In both examples, we estimate the parameter s^* based on the n observations with the different information criteria: AIC, BICU, BICL, MSEIC and QAIC, and then estimate the probability vector $\boldsymbol{p} = (p_1, \dots, p_{s^*}, 0, \dots, 0)^\top$ by $\widehat{\boldsymbol{p}}_{n,*}^{\widehat{s}_n}$ given in (2.4). For comparison, we run simulations for the local model with $k_n = 0.05 \cdot n$ and for the global model with an estimated k_n. Since r is not known we use the estimator

$$\widehat{r}_n = |\widehat{\mathcal{S}}_n(\boldsymbol{Z})| = |\{\beta \in \mathcal{P}_d^* : T_n(C_\beta, k_n) > 0\}|$$

at this point. In total, we conducted 500 repetitions with sample sizes $n = 1000, 5000, 10000, 20000$. The code for the following simulation study is available in the Supplementary Material B [6] and at https://gitlab.kit.edu/projects/164856.

6.1. Error measures

To quantify the discrepancy between the true distribution p and the estimated distribution $\widehat{\boldsymbol{p}}_{n,*}^{\widehat{s}_n}$ in (2.4) we use different measures. We start with the Hellinger distance, which is for discrete probability measures $\mathbb{P}$ and $\mathbb{Q}$ with probabilities $p_1, \dots, p_m$ and $q_1, \dots, q_m$ for $m \in \mathbb{N}$ given by $H(P, Q) := \frac{1}{\sqrt{2}} \|\boldsymbol{p} - \boldsymbol{q}\|_2$ where $\boldsymbol{p} = (p_1, \dots, p_m)^\top$ and $\boldsymbol{q} = (q_1, \dots, q_m)^\top$. Since our primary goal is the identification of the relevant directions s^*, we employ alternative measures. These measures evaluate the validity of a detected direction, without considering the weight assigned to it.

To be more precise, the confusion matrix visualizes the performance of an information criterion. Suppose an information criterion gives $\widehat{s}$ as an estimator for the number s^* of true directions of $2^d - 1$ possible directions. Then we define the confusion matrix for the different information criteria (IC)

	Theoretic direction	No theoretic direction	#Directions
IC detects direction	True positive (TP)	False positive (FP)	$\widehat{s}$
IC detects no direction	False negative (FN)	True negative (TN)	$2^d - 1 - \widehat{s}$
#Directions	s^*	$2^d - 1 - s^*$	$2^d - 1$

and as error measures

$$\text{Accuracy Error} := 1 - \frac{\text{TP} + \text{TN}}{\text{TP} + \text{TN} + \text{FP} + \text{FN}} = \frac{\text{FP} + \text{FN}}{2^d - 1},$$
$$F_1 \text{ Error} := 1 - \frac{2\text{TP}}{2\text{TP} + \text{FP} + \text{FN}} = 1 - \frac{2\text{TP}}{s^* + \widehat{s}},$$

which reflects the errors. If we take 1 − Accuracy Error and 1 − F_1 Error, respectively, we obtain the original definition in Powers [31] such that our error measures are negatively oriented and a lower value is better. The Accuracy Error measures the relative number of false classified directions, whereas the F_1 Error is the harmonic mean based on the precision and

the recall. Note, that the precision error is the relative amount of actual theoretical directions to the number of detected directions whereas the recall gives the proportion of theoretical directions.

6.2. Asymptotic tail independent model

In the first example, we consider d-dimensional i.i.d. random vectors whose spectral measure only concentrates on the axis. To define their distribution, we assume that $\boldsymbol{H} = (h_{ij})_{1\le i,j\le d} \in \mathbb{R}^{d\times d}$ with $h_{ij} \overset{\text{i.i.d.}}{\sim} \mathcal{U}((0,1))$ and

$$\boldsymbol{\Sigma} \coloneqq \operatorname{diag}(h_{11}^{-1/2},\dots,h_{dd}^{-1/2}) \cdot \boldsymbol{H}^\top \cdot \boldsymbol{H} \cdot \operatorname{diag}(h_{11}^{-1/2},\dots,h_{dd}^{-1/2}).$$

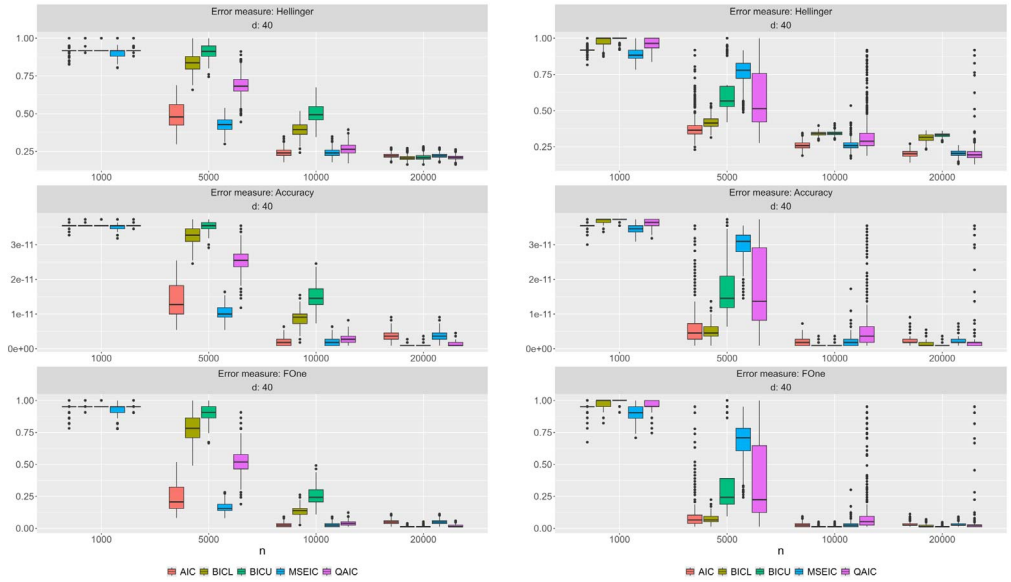
Note that $\boldsymbol{\Sigma}_{ii} = 1$, $i = 1,\dots,d$ and $\boldsymbol{\Sigma}_{ij} < 1$, $i \neq j$. Suppose now $\boldsymbol{Y} = (Y_1,\dots,Y_d) \sim \mathcal{N}_d(\mathbf{0}_d, \boldsymbol{\Sigma})$ under the condition of $\boldsymbol{\Sigma}$ whose components have, by construction, as marginal distribution the standard normal distribution Φ. It is well known that the multivariate normal distribution with correlations smaller than 1 exhibits pairwise asymptotic independence (Resnick [32], Corollary 5.28). Now, let $\boldsymbol{Y}^1,\dots,\boldsymbol{Y}^n$ be an i.i.d. sequence of random vectors with distribution $\boldsymbol{Y}$ and define the i.i.d. random vectors $\boldsymbol{X}^i = (X_1^i,\dots,X_d^i)^\top \in \mathbb{R}_+^d$, $i = 1,\dots,n$, as

$$X_j^i \coloneqq \frac{1}{1-\Phi(Y_j^i)}, \quad 1 \le j \le d,$$

which are regularly varying with tail index $\alpha = 1$ and exhibit pairwise asymptotic independence so that the extreme directions are the $s^* = d$ axes. For our simulation study, we assume now that $d = s^* = 40$; the results are presented in Figure 1, on the left hand side for the local model with $k_n = 0.05 \cdot n$ and on the right hand side for the global model. In the local model we see that for small values of n, as $n = 5000$ and $n = 10000$, the AIC and MSEIC perform better than the other information criteria, while for $n = 10000$ the QAIC performs only slightly worse than the AIC and the MSEIC. But this changes for $n = 20000$: When evaluating the Accuracy Error and the F_1 Error the BIC and the QAIC outperform the AIC and MSEIC. It even seems that the Accuracy Error and F_1 Error of the AIC and MSEIC increase, suggesting a tendency toward overfitting, which is in agreement with the theoretical results that the AIC and MSEIC are overfitting with a positive probability (Theorem 3.1 and Theorem 4.3), whereas the QAIC and BIC are consistent (Theorem 3.8 and Theorem 5.5). If we compare the simulation results for the local model (left part of Figure 1) with the results for the global model (right part of Figure 1), we realize that for $n = 5000$ and 10000 the global model of the AIC and BIC performs better than their corresponding local models, whereas the global model of the QAIC is, on average, better than its local version, it has many outliers with the tendency to overfit.

6.3. Asymptotic dependent model

Next, we present an additional simulation study for a model with asymptotic dependence which can also be found in Meyer and Wintenberger [27]. Consequently not only directions



(a) Local model with $k_n/n = 0.05$ (b) Global model

Fig 1. *Simulations for asymptotically independent data with $s^* = d = 40$ directions of extremes: In the top row we use as error measure the Hellinger distance, in the middle row the Accuracy Error and in the bottom row the F_1 Error, which are plotted against the sample size n on the x-axis.*

with $|\beta| = 1$ are relevant. Let $\boldsymbol{X}$ be an $\mathbb{R}^d$ valued random vector and $d_1, d_2, d_3 \in \mathbb{N} \cup \{0\}$, such that

$$d \geq d_1 + 2d_2 + 3d_3.$$

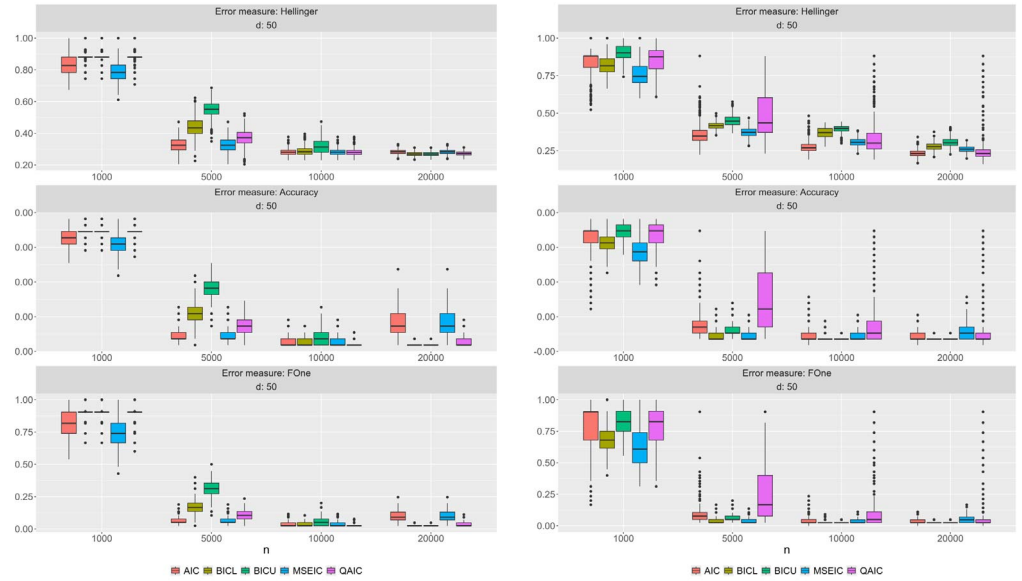
The parameters d_1, d_2, d_3 specify the number of one, two, and three-dimensional directions. The marginal distributions of $\boldsymbol{X}$ are defined by

$$\begin{aligned}
X_j &\sim \text{Pareto}(1), && j = 1, \ldots, d_1, \\
(X_j, X_{j+1}) &\sim (\text{Pareto}(1), X_j + \text{Pareto}(2)), && j = d_1 + 1, d_1 + 3, \ldots, d_1 + 2 \cdot d_2 - 1, \\
(X_j, X_{j+1}, X_{j+2}) &\sim (\text{Pareto}(1), X_j + \text{Pareto}(2), X_j + \text{Pareto}(2)), \\
& && j = d_1 + 2 \cdot d_2 + 1, d_1 + 2 \cdot d_2 + 4, \ldots, d_1 + 2 \cdot d_2 + 3 \cdot d_3 - 2, \\
X_j &\sim \text{Pareto}(2), && j = d_1 + 2 \cdot d_2 + 3 \cdot d_3, \ldots, d.
\end{aligned}$$

The random vector $\boldsymbol{Z}$ in Definition 2.1 puts mass on the sets

$$\begin{aligned}
&C_{\{1\}}, \ldots, C_{\{d_1\}}, \\
&C_{\{d_1+1, d_1+2\}}, \ldots, C_{\{d_1+2\cdot d_2-1, d_1+2\cdot d_2\}}, \\
&C_{\{d_1+2\cdot d_2+1, d_1+2\cdot d_2+2, d_1+2\cdot d_2+3\}}, \ldots, C_{\{d_1+2\cdot d_2+3\cdot d_3-2, d_1+2\cdot d_2+3\cdot d_3-1, d_1+2\cdot d_2+3\cdot d_3\}}.
\end{aligned}$$

In total, there are $d_1 + d_2 + d_3$ directions with probability mass, and the goal is again to identify these directions. For the simulation study in Figure 2 we chose $d_1 = 10, d_2 = d_3 = 5$ and $d = 50$ resulting in $s^* = 20$ extreme directions. The plots show similar features as for the asymptotic independent case in Section 6.2 (cf. Figure 1).



(a) Local model with $k_n/n = 0.05$ (b) Global model

Fig 2. *Simulations for asymptotic dependent data with* $s^* = 20$ *directions of extremes and* $d = 50$*: In the top row we use as an error measure the Hellinger distance, in the middle row the Accuracy Error and in the bottom row the* F_1 *Error, which are plotted on the* y*-axis against the sample size* n *on the* x*-axis.*

7. Application to real-world data

In this section, we examine the dependence structure of extreme wind speeds using the same example as Meyer and Wintenberger [29]. For this purpose, the daily average wind speed at 12 synoptic meteorological stations in the Republic of Ireland from 1961 until 1978 with $n = 6574$ observations are considered. The data was subject to Haslett and Raftery [24] and taken from StatLib - Datasets Archive [37]. To what extent dependencies exist, that are not due to the geographical proximity, will be analyzed in the following. The locations of the stations are shown in Figure 4 and consist of: Belmullet (BEL), Birr (BIR), Claremorris (CLA), Clones (CLO), Dublin (DUB), Kilkenny (KIL), Malin Head (MAL), Mullingar (MUL), Roche's Pt. (RPT), Rosslare (ROS), Shannon (SHA) and Valentia (VAL). For the preprocessing, we use the same Hill estimator $\widehat{\alpha} = 10.7$ as Meyer and Wintenberger [29]. We considered values of k_n between 33 and 1183.

The values of the estimators for k_n, k_n/n, and s^* are presented in Table 1.

Table 1
Estimators for the wind speed data set based on the different information criteria

IC	$\widehat{k}$	$\widehat{k}/n$	$\widehat{s}$
AIC	460	0.07	11
BICU	1118	0.17	12
BICL	1118	0.17	13
MSEIC	230	0.03	9
QAIC	592	0.09	11

The number of extreme observations k_n varies between 230 and 1118, which corresponds to 3% to 17% of the data. However, the information criteria reported between 9 and 13 number of extreme directions, which is not a large range compared to the choice of k_n. On the left-hand side of Figure 3, the values of the information criteria are plotted against the threshold k_n, while on the right-hand side, the number of estimated directions is mapped as well against k_n. The vertical lines indicate the minimum of the information criteria. It appears that for the number s of extremal directions, there is a more distinct plateau around the optimal value $\widehat{k}_n$ for BICU, MSEIC and QAIC compared to AIC and BICL.

Fig 3. *On the left-hand side in the figure the value of the information criteria (IC) and on the right-hand side, the number s of extremal directions is plotted against k_n. The values of the IC are scaled, such that they start at* 1. *The vertical lines indicate the minimum value of the information criteria.*

A graphic of the Republic of Ireland is given in Figure 4, where the black dots highlight the different locations of the stations. Colored diamonds close to a station are markers for estimated extreme wind speeds at that station based on an information criterion. All information criteria only identify stations on the coast as extreme, all inland stations have non-extreme wind speeds. AIC missed one station on the coast, which is Valentia located more than 130 km away from the other stations. MSEIC, QAIC, BICU and BICL recovered the same maximal clusters and missed the coastal stations Shannon and Dublin. The first station, Shannon, is connected to the ocean but nearly 40 kilometers away from the open sea. The second station, Dublin, is oriented towards the Irish Sea, rather than the Atlantic Ocean. All information criteria identified Belmullet, Mullingar, Rosslare and Roche's Pt., and four out of five information criteria also recognized Valentia.

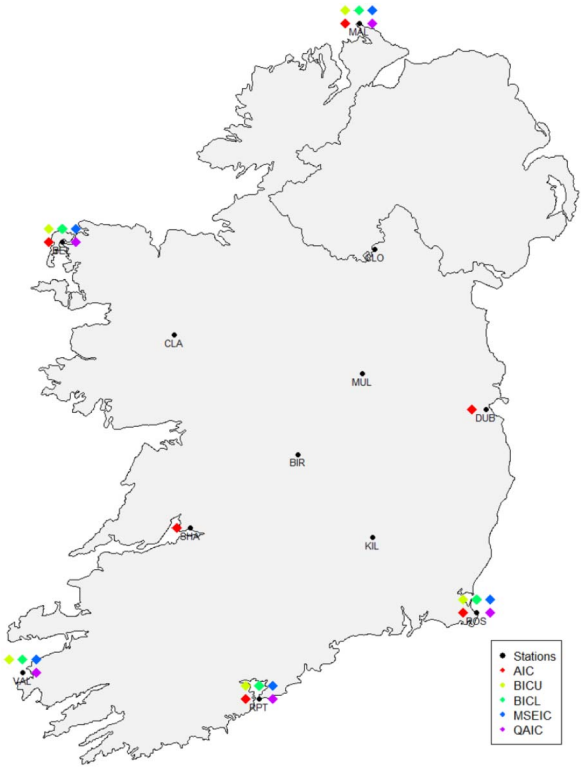


Fig 4. *Maximal subsets recovered by the information criteria of the daily average wind speed.*

8. Conclusion

In this paper, we developed three different information criteria for both the number of extreme directions s^* as well as for the choice of the optimal threshold k_n. Where the BIC is based on a Bayesian approach for a multinomial model in analogy to the AIC of Meyer and Wintenberger [29], the QAIC uses the ideas of an Akaike information criterion, but it is based on a Gaussian likelihood function in comparison to the AIC. In contrast, for MSEIC no likelihood assumption is necessary; it uses the MSE. The advantage of BICU, BICL and QAIC is that they are weakly consistent information criteria for the number of extreme directions s^*, where AIC and MSEIC tend to overestimate s^* for large sample sizes, which we slightly see in the simulation study of the local models for large n but for small n the MSEIC performs extraordinarily well. All information criteria performed quite well, none is particularly superior in all situations. Finally, the information criteria were successfully applied to a real-world data set, where MSEIC, QAIC, BICU and BICL detected the same extreme clusters. In practice, we estimate, of course, r by $\widehat{r}_n = |\widehat{\mathcal{S}}_n(\boldsymbol{Z})|$ and plug this estimate in the information criteria. In this setup, all the consistency results in the paper remain true if we additionally assume that $\sqrt{k_n \rho_n}(\widehat{r}_n - r) \xrightarrow{\mathbb{P}} 0$ as $n \to \infty$. However, the motivation for the definitions of the information criteria is much clearer when it is assumed that $r = |\widehat{\mathcal{S}}_n(\boldsymbol{Z})|$ is deterministic and independent of n.

Appendix A: Proofs

A.1. Proofs of Section 3

A.1.1. Proof of Theorem 3.1

Proof of Theorem 3.1.
Step 1: Suppose $s > s^*$. By the definition of the AIC and the log-likelihood function in (2.3) it follows that

$$\begin{aligned}
&\mathrm{AIC}_{k_n}(s) - \mathrm{AIC}_{k_n}(s^*) \\
&\quad = -\log L_{M_{k_n}^s}(\widehat{\boldsymbol{p}}_n^s \mid \boldsymbol{T}_n(k_n)) + s + \log L_{M_{k_n}^{s^*}}(\widehat{\boldsymbol{p}}_n^{s^*} \mid \boldsymbol{T}_n(k_n)) - s^* \\
&\quad = -\sum_{j=s^*+1}^{s} T_{n,j}(k_n) \log\left(\frac{T_{n,j}(k_n)}{k_n}\right) - \log\left(\frac{1}{r-s}\sum_{j=s+1}^{r}\frac{T_{n,j}(k_n)}{k_n}\right)\sum_{i=s+1}^{r} T_{n,i}(k_n) \\
&\qquad + \log\left(\frac{1}{r-s^*}\sum_{j=s^*+1}^{r}\frac{T_{n,j}(k_n)}{k_n}\right)\sum_{i=s^*+1}^{r} T_{n,i}(k_n) + (s-s^*),
\end{aligned} \tag{A.1}$$

where we used that $s > s^*$. Inserting the alternative representation

$$T_{n,j}(k_n) = k_n\rho_n + \sqrt{k_n\rho_n}Y_{n,j}$$

where

$$Y_{n,j} \coloneqq \sqrt{k_n\rho_n}\left(\frac{T_{n,j}(k_n)}{\rho_n k_n} - 1\right), \quad j = s^*+1, \dots, r,$$

gives that

$$\begin{aligned}
&\mathrm{AIC}_{k_n}(s) - \mathrm{AIC}_{k_n}(s^*) \\
&\quad = -\sum_{j=s^*+1}^{s}(k_n\rho_n + \sqrt{k_n\rho_n}Y_{n,j})\log\left(1 + \frac{1}{\sqrt{k_n\rho_n}}Y_{n,j}\right) \\
&\qquad - \log\left(1 + \frac{1}{r-s}\sum_{j=s+1}^{r}\frac{1}{\sqrt{k_n\rho_n}}Y_{n,j}\right)\sum_{i=s+1}^{r}(k_n\rho_n + \sqrt{k_n\rho_n}Y_{n,i}) \\
&\qquad + \log\left(1 + \frac{1}{r-s^*}\sum_{j=s^*+1}^{r}\frac{1}{\sqrt{k_n\rho_n}}Y_{n,j}\right)\sum_{i=s^*+1}^{r}(k_n\rho_n + \sqrt{k_n\rho_n}Y_{n,i}) \\
&\qquad + (s-s^*).
\end{aligned} \tag{A.2}$$

For the asymptotic behavior we apply Assumption (A3) which results in

$$(Y_{n,s^*+1}, \dots, Y_{n,r}) \xrightarrow{\mathcal{D}} (Y_{s^*+1}, \dots, Y_r) =: \boldsymbol{Y} \sim \mathcal{N}_{r-s^*}(\boldsymbol{0}_{r-s}, \boldsymbol{I}_{r-s^*}), \quad n \to \infty, \tag{A.3}$$

and thus,

$$Y_{n,i} = O_{\mathbb{P}}(1) \quad \text{for} \quad i = s^*+1, \dots, r.$$

This and the Taylor expansion of the logarithm

$$\log(1+x) = x - \frac{1}{2}x^2 + O(x^3), \quad x \to 0,$$

we insert in (A.2) such that

$$\begin{aligned}
&\mathrm{AIC}_{k_n}(s) - \mathrm{AIC}_{k_n}(s^*) \\
&\quad = -\sum_{j=s^*+1}^{s} (k_n\rho_n + \sqrt{k_n\rho_n}Y_{n,j}) \left(\frac{1}{\sqrt{k_n\rho_n}} Y_{n,j} - \frac{1}{2}\frac{1}{k_n\rho_n} Y_{n,j}^2 \right) \\
&\qquad - \left(\frac{1}{r-s} \sum_{j=s+1}^{r} \frac{1}{\sqrt{k_n\rho_n}} Y_{n,j} - \frac{1}{2}\Big(\frac{1}{r-s} \sum_{j=s+1}^{r} \frac{1}{\sqrt{k_n\rho_n}} Y_{n,j} \Big)^2 \right) \\
&\qquad\qquad \cdot \sum_{i=s+1}^{r} (k_n\rho_n + \sqrt{k_n\rho_n}Y_{n,i}) \\
&\qquad + \left(\frac{1}{r-s^*} \sum_{j=s^*+1}^{r} \frac{1}{\sqrt{k_n\rho_n}} Y_{n,j} - \frac{1}{2}\Big(\frac{1}{r-s^*} \sum_{j=s^*+1}^{r} \frac{1}{\sqrt{k_n\rho_n}} Y_{n,j} \Big)^2 \right) \\
&\qquad\qquad \cdot \sum_{i=s^*+1}^{r} (k_n\rho_n + \sqrt{k_n\rho_n}Y_{n,i}) \\
&\qquad + (s - s^*) + O_{\mathbb{P}}((k_n\rho_n)^{-1/2}).
\end{aligned}$$

Since $k_n\rho_n \to \infty$ (Lemma 2.8(a)) we receive

$$\begin{aligned}
\mathrm{AIC}_{k_n}(s) - \mathrm{AIC}_{k_n}(s^*) = &-\frac{1}{2}\sum_{j=s^*+1}^{s} Y_{n,j}^2 - \frac{1}{2(r-s)} \left(\sum_{j=s+1}^{r} Y_{n,j} \right)^2 \\
&+ \frac{1}{2(r-s^*)} \left(\sum_{j=s^*+1}^{r} Y_{n,j} \right)^2 + (s - s^*) + o_{\mathbb{P}}(1).
\end{aligned}$$

Due to (A.3) and the continuous mapping theorem we finally obtain as $n \to \infty$,

$$\begin{aligned}
\mathrm{AIC}_{k_n}(s) - \mathrm{AIC}_{k_n}(s^*) \xrightarrow{\mathcal{D}} &-\frac{1}{2}\sum_{j=s^*+1}^{s} Y_j^2 - \frac{1}{2(r-s)} \left(\sum_{j=s+1}^{r} Y_j \right)^2 \\
&+ \frac{1}{2(r-s^*)} \left(\sum_{j=s^*+1}^{r} Y_j \right)^2 + (s - s^*). \qquad \text{(A.4)}
\end{aligned}$$

Obviously,

$$\mathbb{P}\left(-\frac{1}{2}\sum_{j=s^*+1}^{s} Y_j^2 - \frac{\left(\sum_{j=s+1}^{r} Y_j\right)^2}{2(r-s)} + \frac{\left(\sum_{j=s^*+1}^{r} Y_j\right)^2}{2(r-s^*)} + s - s^* < 0 \right) > 0.$$

Step 2: Suppose $s < s^*$. We obtain analog to (A.1) that

$$\mathrm{AIC}_{k_n}(s) - \mathrm{AIC}_{k_n}(s^*)$$

$$= \sum_{j=s+1}^{s^*} T_{n,j}(k_n) \log\left(\frac{T_{n,j}(k_n)}{k_n}\right) - \log\left(\frac{1}{r-s}\sum_{j=s+1}^{r}\frac{T_{n,j}(k_n)}{k_n}\right)\sum_{i=s+1}^{r} T_{n,i}(k_n)$$
$$+ \log\left(\frac{1}{r-s^*}\sum_{j=s^*+1}^{r}\frac{T_{n,j}(k_n)}{k_n}\right)\sum_{i=s^*+1}^{r} T_{n,i}(k_n) + (s-s^*). \tag{A.5}$$

A direct consequence of $T_{n,j}(k_n)/k_n \xrightarrow{\mathbb{P}} 0$ for $j = s^*+1,\dots,r$ (Lemma 2.8(c)) and $\lim_{x\to 0} x\log(x) = 0$ is that

$$\log\left(\frac{1}{r-s^*}\sum_{j=s^*+1}^{r}\frac{T_{n,j}(k_n)}{k_n}\right)\sum_{i=s^*+1}^{r}\frac{T_{n,i}(k_n)}{k_n} \xrightarrow{\mathbb{P}} 0.$$

Furthermore, Lemma 2.8(b) yields $T_{n,j}(k_n)/k_n \xrightarrow{\mathbb{P}} p_j > 0$ for $j = 1,\dots,s^*$ and thus, as $n\to\infty$,

$$\sum_{i=s+1}^{s^*}\frac{T_{n,i}(k_n)}{k_n}\log\left(\frac{T_{n,i}(k_n)}{k_n}\right) - \log\left(\frac{1}{r-s}\sum_{j=s+1}^{r}\frac{T_{n,j}(k_n)}{k_n}\right)\sum_{i=s+1}^{r}\frac{T_{n,i}(k_n)}{k_n}$$
$$\xrightarrow{\mathcal{D}} \sum_{i=s+1}^{s^*} p_i\left(\log(p_i) - \log\left(\frac{1}{r-s}\sum_{j=s+1}^{s^*} p_j\right)\right), \tag{A.6}$$

while we used $p_i = 0$ for $s^* \le i \le r$. Next, we apply the log sum inequality (Cover [12], Theorem 2.7.1) to the limit of (A.6) and receive

$$\sum_{i=s+1}^{s^*} p_i\left(\log(p_i) - \log\left(\frac{1}{r-s}\sum_{j=s+1}^{s^*} p_j\right)\right) = \sum_{i=s+1}^{s^*} p_i \log\left(\frac{p_i}{\frac{1}{r-s}\sum_{j=s+1}^{s^*} p_j}\right)$$
$$\ge \left(\sum_{i=s+1}^{s^*} p_i\right)\log\left(\frac{\sum_{i=s+1}^{s^*} p_i}{\frac{s^*-s}{r-s}\sum_{j=s+1}^{s^*} p_j}\right) = \left(\sum_{i=s+1}^{s^*} p_i\right)\log\left(\frac{r-s}{s^*-s}\right) > 0, \tag{A.7}$$

since $r > s^*$. Dividing (A.5) by k_n and using (A.6) and (A.7) gives

$$\frac{1}{k_n}\left(\mathrm{AIC}_{k_n}(s) - \mathrm{AIC}_{k_n}(s^*)\right) \xrightarrow{\mathcal{D}} \sum_{i=s+1}^{s^*} p_i\left(\log(p_i) - \log\left(\frac{1}{r-s}\sum_{j=s+1}^{s^*} p_j\right)\right) > 0,$$

and thus, the assertion follows. □

A.1.2. Proof of Proposition 3.5

Before we are able to present the proof of Proposition 3.5 we require some auxiliary lemmata whose proofs are moved to Section 1 of the Supplementary Material A [5]. In the following, we work with the r-dimensional multivariate normal distribution $\mathcal{N}_r(k_n\underline{\boldsymbol{A}}_s(\underline{\widetilde{\boldsymbol{p}}}^s), k_n\mathrm{diag}(\underline{\boldsymbol{A}}_s(\underline{\widetilde{\boldsymbol{p}}}^s)))$, $\underline{\widetilde{\boldsymbol{p}}}^s \in \mathbb{R}_+^{s+1}$, which has a negative log-likelihood function

$$-2\log L_{\mathcal{N}_r}(\underline{\widetilde{\boldsymbol{p}}}^s \mid \mathcal{T}_n) = r\log(2\pi) + r\log(k_n) + \sum_{j=1}^{s}\log(\underline{\widetilde{p}}_j^s) + (r-s)\log(\underline{\widetilde{\rho}}^s)$$

$$+ k_n\Bigg(\sum_{j=1}^{s}\frac{1}{\widetilde{\underline{p}}_j^s}\Big(\frac{\mathcal{T}_{n,j}}{k_n}-\widetilde{\underline{p}}_j^s\Big)^2+\sum_{j=s+1}^{r}\frac{1}{\widetilde{\underline{\rho}}^s}\Big(\frac{\mathcal{T}_{n,j}}{k_n}-\widetilde{\underline{\rho}}^s\Big)^2\Bigg).$$

Lemma A.1. *Suppose the assumptions of Proposition 3.5 hold and $\widehat{\underline{p}}_n^s(\mathcal{T}_n)$ is defined analog to $\widehat{\underline{p}}_n^s(\widetilde{\mathcal{T}}_n)$ in (3.2). Then as $n\to\infty$,*

$$\boldsymbol{Y}_n := \sqrt{k_n}\operatorname{diag}\Big(p_{n,1},\dots,p_{n,s},\frac{\rho_n}{(r-s)},\rho_n,\dots,\rho_n\Big)^{-1/2}\begin{pmatrix}\Big(\widehat{\underline{p}}_n^s(\widetilde{\mathcal{T}}_n)-\widehat{\underline{p}}_n^s(\mathcal{T}_n)\Big)\\ \Big(\frac{\mathcal{T}_{n,s+1}}{k_n}-\widehat{\underline{\rho}}_n^s(\mathcal{T}_n)\Big)\\ \vdots\\ \Big(\frac{\mathcal{T}_{n,r}}{k_n}-\widehat{\underline{\rho}}_n^s(\mathcal{T}_n)\Big)\end{pmatrix}$$

$$\xrightarrow{\mathcal{D}} \mathcal{N}_{r+1}\left(\mathbf{0}_{r+1},\Sigma\right),$$

where

$$\Sigma := \begin{pmatrix}2\boldsymbol{I}_{s+1} & \mathbf{0}_{s\times(r-s)}\\ \mathbf{0}_{(r-s)\times(s+1)} & \boldsymbol{I}_{r-s}-\frac{\mathbf{1}_{r-s}\mathbf{1}_{r-s}^\top}{r-s}\end{pmatrix}.$$

Lemma A.2. *Suppose the assumptions of Proposition 3.5 hold and $\widehat{\underline{p}}_n^s(\mathcal{T}_n)$ is defined analog to $\widehat{\underline{p}}_n^s(\widetilde{\mathcal{T}}_n)$ in (3.2).*

(a) Then as $n\to\infty$,

$$\nabla\log L_{\mathcal{N}_r}(\widehat{\underline{p}}_n^s(\mathcal{T}_n)\mid\mathcal{T}_n)(\widehat{\underline{p}}_n^s(\widetilde{\mathcal{T}}_n)-\widehat{\underline{p}}_n^s(\mathcal{T}_n))\xrightarrow{\mathbb{P}}0.$$

(b) Suppose $\bar{\boldsymbol{p}}_n := (\bar{p}_{n,1},\dots,\bar{p}_{n,s},\bar{\rho}_n)^\top\in\mathbb{R}_+^{s+1}$ satisfies

$$\|\bar{\boldsymbol{p}}_n-\widehat{\underline{p}}_n^s(\mathcal{T}_n)\|\le\|\widehat{\underline{p}}_n^s(\widetilde{\mathcal{T}}_n)-\widehat{\underline{p}}_n^s(\mathcal{T}_n)\|,\quad n\in\mathbb{N}.$$

Then as $n\to\infty$,

$$(\widehat{\underline{p}}_n^s(\widetilde{\mathcal{T}}_n)-\widehat{\underline{p}}_n^s(\mathcal{T}_n))^\top\Big(\nabla^2\log L_{\mathcal{N}_r}(\bar{\boldsymbol{p}}_n\mid\mathcal{T}_n)+k_n\big(\operatorname{diag}(p_{n,1},\dots,p_{n,s},\rho_n/(r-s))^{-1}\big)\Big)$$
$$\cdot(\widehat{\underline{p}}_n^s(\widetilde{\mathcal{T}}_n)-\widehat{\underline{p}}_n^s(\mathcal{T}_n))\xrightarrow{\mathbb{P}}0.$$

Proof of Proposition 3.5. Using a Taylor expansion of $\log L_{\mathcal{N}_r}(\widehat{\underline{p}}_n^s(\widetilde{\mathcal{T}}_n)\mid\mathcal{T}_n)$ around $\widehat{\underline{p}}_n^s(\mathcal{T}_n)$ yields the existence of a random vector $\bar{\boldsymbol{p}}_n := (\bar{p}_{n,1},\dots,\bar{p}_{n,s},\bar{\rho}_n)^\top$ with

$$\|\bar{\boldsymbol{p}}_n-\widehat{\underline{p}}_n^s(\mathcal{T}_n)\|\le\|\widehat{\underline{p}}_n^s(\widetilde{\mathcal{T}}_n)-\widehat{\underline{p}}_n^s(\mathcal{T}_n)\|$$

such that

$$\begin{aligned}&\log L_{\mathcal{N}_r}(\widehat{\underline{p}}_n^s(\widetilde{\mathcal{T}}_n)\mid\mathcal{T}_n)\\ &\quad=\log L_{\mathcal{N}_r}(\widehat{\underline{p}}_n^s(\mathcal{T}_n)\mid\mathcal{T}_n)+\nabla\log L_{\mathcal{N}_r}(\widehat{\underline{p}}_n^s(\mathcal{T}_n)\mid\mathcal{T}_n)\big(\widehat{\underline{p}}_n^s(\widetilde{\mathcal{T}}_n)-\widehat{\underline{p}}_n^s(\mathcal{T}_n)\big)\\ &\qquad+\frac{1}{2}\big(\widehat{\underline{p}}_n^s(\widetilde{\mathcal{T}}_n)-\widehat{\underline{p}}_n^s(\mathcal{T}_n)\big)^\top\nabla^2\log L_{\mathcal{N}_r}(\bar{\boldsymbol{p}}_n\mid\mathcal{T}_n)\big(\widehat{\underline{p}}_n^s(\widetilde{\mathcal{T}}_n)-\widehat{\underline{p}}_n^s(\mathcal{T}_n)\big).\end{aligned}$$

Applying Lemma A.2 (b) gives

$$\log L_{\mathcal{N}_r}(\widehat{\underline{p}}_n^s(\widetilde{\mathcal{T}}_n)\mid\mathcal{T}_n)$$

$$
\begin{aligned}
&= \log L_{\mathcal{N}_r}(\underline{\widehat{\boldsymbol{p}}}_n^s(\boldsymbol{\mathcal{T}}_n) \mid \boldsymbol{\mathcal{T}}_n) \\
&\quad - \frac{1}{2}\big(\underline{\widehat{\boldsymbol{p}}}_n^s(\widetilde{\boldsymbol{\mathcal{T}}}_n) - \underline{\widehat{\boldsymbol{p}}}_n^s(\boldsymbol{\mathcal{T}}_n)\big)^\top k_n \mathrm{diag}(p_{n,1}, \ldots, p_{n,s}, \rho_n/(r-s))^{-1}\big(\underline{\widehat{\boldsymbol{p}}}_n^s(\widetilde{\boldsymbol{\mathcal{T}}}_n) - \underline{\widehat{\boldsymbol{p}}}_n^s(\boldsymbol{\mathcal{T}}_n)\big) \\
&\quad + o_{\mathbb{P}}(1).
\end{aligned}
$$

Inserting the definition of $\log L_{\mathcal{N}_r}(\underline{\widehat{\boldsymbol{p}}}_n^s(\boldsymbol{\mathcal{T}}_n) \mid \boldsymbol{\mathcal{T}}_n)$ and $\underline{\widehat{p}}_{n,j}^s(\boldsymbol{\mathcal{T}}_n) = \frac{\mathcal{T}_{n,j}}{k_n}$, $j = 1, \ldots, s$, yield

$$
\begin{aligned}
&\log L_{\mathcal{N}_r}(\underline{\widehat{\boldsymbol{p}}}_n^s(\widetilde{\boldsymbol{\mathcal{T}}}_n) \mid \boldsymbol{\mathcal{T}}_n) \\
&\quad = -\frac{1}{2} r \log(2\pi k_n) - \frac{1}{2}\sum_{j=1}^{s} \log(\underline{\widehat{p}}_{n,j}^s(\boldsymbol{\mathcal{T}}_n)) - \frac{1}{2}(r-s)\log(\underline{\widehat{\rho}}_n^s(\boldsymbol{\mathcal{T}}_n)) \\
&\qquad - \frac{1}{2} k_n \sum_{j=s+1}^{r} \frac{1}{\underline{\widehat{\rho}}_n^s(\boldsymbol{\mathcal{T}}_n)} \left(\frac{\mathcal{T}_{n,j}}{k_n} - \underline{\widehat{\rho}}_n^s(\boldsymbol{\mathcal{T}}_n)\right)^2 \\
&\qquad - \frac{1}{2}\big(\underline{\widehat{\boldsymbol{p}}}_n^s(\widetilde{\boldsymbol{\mathcal{T}}}_n) - \underline{\widehat{\boldsymbol{p}}}_n^s(\boldsymbol{\mathcal{T}}_n)\big)^\top k_n \mathrm{diag}(p_{n,1}, \ldots, p_{n,s}, \rho_n/(r-s))^{-1}\big(\underline{\widehat{\boldsymbol{p}}}_n^s(\widetilde{\boldsymbol{\mathcal{T}}}_n) - \underline{\widehat{\boldsymbol{p}}}_n^s(\boldsymbol{\mathcal{T}}_n)\big) \\
&\qquad + o_{\mathbb{P}}(1).
\end{aligned}
$$

Next, we move some terms on the right-hand side and use $\rho_n/\underline{\widehat{\rho}}_n^s(\boldsymbol{\mathcal{T}}_n) \xrightarrow{\mathbb{P}} 1$ (cf. Lemma 2.8 (c) and the assumption $s \geq s^*$) and $\boldsymbol{Y}_n$ as defined in Lemma A.1, which result in

$$
\begin{aligned}
&\log L_{\mathcal{N}_r}(\underline{\widehat{\boldsymbol{p}}}_n^s(\widetilde{\boldsymbol{\mathcal{T}}}_n) \mid \boldsymbol{\mathcal{T}}_n) + \frac{1}{2} r \log(2\pi k_n) + \frac{1}{2}\sum_{j=1}^{s} \log(\underline{\widehat{p}}_{n,j}^s(\boldsymbol{\mathcal{T}}_n)) + \frac{1}{2}(r-s)\log(\underline{\widehat{\rho}}_n^s(\boldsymbol{\mathcal{T}}_n)) \\
&\quad = -\frac{1}{2} k_n \sum_{j=s+1}^{r} \frac{1}{\rho_n} \left(\frac{\mathcal{T}_{n,j}}{k_n} - \underline{\widehat{\rho}}_n^s(\boldsymbol{\mathcal{T}}_n)\right)^2 \\
&\qquad - \frac{1}{2}\big(\underline{\widehat{\boldsymbol{p}}}_n^s(\widetilde{\boldsymbol{\mathcal{T}}}_n) - \underline{\widehat{\boldsymbol{p}}}_n^s(\boldsymbol{\mathcal{T}}_n)\big)^\top \cdot k_n \mathrm{diag}(p_{n,1}, \ldots, p_{n,s}, \rho_n/(r-s))^{-1}\big(\underline{\widehat{\boldsymbol{p}}}_n^s(\widetilde{\boldsymbol{\mathcal{T}}}_n) - \underline{\widehat{\boldsymbol{p}}}_n^s(\boldsymbol{\mathcal{T}}_n)\big) \\
&\qquad + o_{\mathbb{P}}(1) \\
&\quad = -\frac{1}{2}\boldsymbol{Y}_n^\top \boldsymbol{Y}_n + o_{\mathbb{P}}(1) \\
&\xrightarrow{\mathcal{D}} -\frac{1}{2}\boldsymbol{Y}^\top \boldsymbol{Y}
\end{aligned}
$$

by Lemma A.1, where $\boldsymbol{Y} \sim \mathcal{N}_{r+1}(\boldsymbol{0}_{r+1}, \Sigma)$. Since

$$
\mathbb{E}[-\frac{1}{2}\boldsymbol{Y}^\top \boldsymbol{Y}] = -\frac{1}{2}\mathbb{E}[\mathrm{trace}(\boldsymbol{Y}^\top \boldsymbol{Y})] = -\frac{1}{2}\mathrm{trace}(\Sigma) = -\frac{r+s+1}{2}
$$

the assertion follows. □

A.1.3. Proof of Theorem 3.8

Proof of Theorem 3.8.
Step 1: Suppose $s < s^*$. We have $\widehat{p}_{n,j}^s = \widehat{p}_{n,j}^{s^*}$ for $j = 1, \ldots, s$ and due to Lemma 2.8(b,c) we have as $n \to \infty$,

$$
\widehat{p}_{n,j}^{s^*} \xrightarrow{\mathbb{P}} p_j > 0, \quad j = 1, \ldots, s^*,
$$

and similarly $\widehat{\rho}_n^s \xrightarrow{\mathbb{P}} \frac{1}{r-s}\sum_{j=s+1}^{s^*} p_j > 0$ as well as $\widehat{\rho}_n^{s^*} \xrightarrow{\mathbb{P}} 0$. Thus,

$$-\sum_{j=s+1}^{s^*} \log(\widehat{p}_{n,j}^{s^*}) + (r-s)\log(\widehat{\rho}_n^s) \xrightarrow{\mathbb{P}} -\sum_{j=s+1}^{s^*} \log(p_j) + (r-s)\log\Big(\frac{1}{r-s}\sum_{j=s+1}^{s^*} p_j\Big)$$

and $\log(\widehat{\rho}_n^{s^*}) \xrightarrow{\mathbb{P}} -\infty$. Therefore, we have as $n \to \infty$,

$$\begin{aligned}&\mathrm{QAIC}_{k_n}(s) - \mathrm{QAIC}_{k_n}(s^*)\\ &\quad= -\sum_{j=s+1}^{s^*} \log(\widehat{p}_{n,j}^{s^*}) + (r-s)\log(\widehat{\rho}_n^s) - (r-s^*)\log(\widehat{\rho}_n^{s^*}) + (s-s^*)\\ &\quad\xrightarrow{\mathbb{P}} \infty.\end{aligned}$$

Step 2: Suppose $s > s^*$. In this case, we have by Lemma 2.8(b,c) that

$$\frac{\widehat{p}_{n,j}^s}{\rho_n} \xrightarrow{\mathbb{P}} 1, \quad j = s^*+1, \dots, s,$$

and similarly $\widehat{\rho}_n^s/\rho_n \xrightarrow{\mathbb{P}} 1$ as well as $\widehat{\rho}_n^{s^*}/\rho_n \xrightarrow{\mathbb{P}} 1$. Hence, with the continuous mapping theorem we receive $\log(\widehat{p}_{n,j}^s/\rho_n) \xrightarrow{\mathbb{P}} 0$ for $j = s^*+1, \dots, s$, $\log(\widehat{\rho}_n^{s^*}/\rho_n) \xrightarrow{\mathbb{P}} 0$ and $\log(\widehat{\rho}_n^s/\rho_n) \xrightarrow{\mathbb{P}} 0$ as $n \to \infty$. Thus, as $n \to \infty$,

$$\begin{aligned}&\mathrm{QAIC}_{k_n}(s) - \mathrm{QAIC}_{k_n}(s^*)\\ &\quad= \sum_{j=s^*+1}^{s} \log\left(\frac{\widehat{p}_{n,j}^s}{\rho_n}\right) + (r-s)\log\left(\frac{\widehat{\rho}_n^s}{\rho_n}\right) - (r-s^*)\log\left(\frac{\widehat{\rho}_n^{s^*}}{\rho_n}\right) + (s-s^*)\\ &\quad\xrightarrow{\mathbb{P}} s - s^* > 0,\end{aligned}$$

which gives the statement. □

A.2. *Proofs of Section 4*

A.2.1. *Proof of Theorem 4.1*

The proof of Theorem 4.1 is similar to the proof of Proposition 3.5. In the first step, we start to calculate the Jacobian vector of $\ell^2\big(\underline{\widetilde{\boldsymbol{p}}}^s | \boldsymbol{T}_n(k_n)\big)$ for $\underline{\widetilde{\boldsymbol{p}}}^s = (\underline{\widetilde{p}}_1^s, \dots \underline{\widetilde{p}}_s^s, \underline{\widetilde{\rho}}^s) \in \mathbb{R}_+^{s+1}$, which is

$$\nabla\ell^2\big(\underline{\widetilde{\boldsymbol{p}}}^s \,|\, \boldsymbol{T}_n(k_n)\big) = k_n\left(\frac{(\underline{\widetilde{p}}_1^s)^2 - \frac{T_{n,1}(k_n)^2}{k_n^2}}{(\underline{\widetilde{p}}_1^s)^2}, \dots, \frac{(\underline{\widetilde{p}}_s^s)^2 - \frac{T_{n,s}(k_n)^2}{k_n^2}}{(\underline{\widetilde{p}}_s^s)^2}, \sum_{j=s+1}^{r} \frac{(\underline{\widetilde{\rho}}^s)^2 - \frac{T_{n,j}(k_n)^2}{k_n^2}}{(\underline{\widetilde{\rho}}^s)^2}\right)$$

and the Hessian matrix is

$$\nabla^2\ell^2\big(\underline{\widetilde{\boldsymbol{p}}}^s \,|\, \boldsymbol{T}_n(k_n)\big) = 2\mathrm{diag}\left(\frac{T_{n,1}(k_n)^2}{k_n(\underline{\widetilde{p}}_1^s)^3}, \dots, \frac{T_{n,s}(k_n)^2}{k_n(\underline{\widetilde{p}}_s^s)^3}, \sum_{j=s+1}^{r} \frac{T_{n,j}(k_n)^2}{k_n(\underline{\widetilde{\rho}}^s)^3}\right).$$

Analog to Lemma A.1 and Lemma A.2 we get the following results.

Lemma A.3. *Suppose Assumption A holds, $s \geq s^*$ and $\widehat{\underline{\boldsymbol{p}}}_n^s(\boldsymbol{T}_n(k_n))$ is defined analogously to $\widehat{\underline{\boldsymbol{p}}}_n^s(\widetilde{\boldsymbol{T}}_n(k_n))$ in (3.2). Then as $n \to \infty$,*

$$\begin{aligned}\mathbf{U}_n &:= \sqrt{k_n}\operatorname{diag}\left(p_{n,1},\dots,p_{n,s},\frac{\rho_n}{(r-s)}\right)^{-1/2}\left(\widehat{\underline{\boldsymbol{p}}}_n^s(\widetilde{\boldsymbol{T}}_n(k_n)) - \widehat{\underline{\boldsymbol{p}}}_n^s(\boldsymbol{T}_n(k_n))\right)\\ &\xrightarrow{\mathcal{D}} \mathcal{N}_{s+1}\left(\mathbf{0}_{s+1}, \begin{pmatrix} 2(\boldsymbol{I}_s - \sqrt{\boldsymbol{p}_{\{1,\dots,s\}}}\sqrt{\boldsymbol{p}_{\{1,\dots,s\}}}^\top) & \mathbf{0}_s \\ \mathbf{0}_s^\top & 2\end{pmatrix}\right).\end{aligned}$$

Lemma A.4. *Suppose Assumption A holds, $s \geq s^*$ and $\widehat{\underline{\boldsymbol{p}}}_n^s(\boldsymbol{T}_n(k_n))$ is defined analogously to $\widehat{\underline{\boldsymbol{p}}}_n^s(\widetilde{\boldsymbol{T}}_n(k_n))$ in (3.2).*

(a) Then as $n \to \infty$,

$$\nabla\ell^2\big(\widehat{\underline{\boldsymbol{p}}}_n^s(\boldsymbol{T}_n(k_n)) \,|\, \boldsymbol{T}_n(k_n)\big)\big(\widehat{\underline{\boldsymbol{p}}}_n^s(\widetilde{\boldsymbol{T}}_n(k_n)) - \widehat{\underline{\boldsymbol{p}}}_n^s(\boldsymbol{T}_n(k_n))\big) \xrightarrow{\mathbb{P}} 0.$$

(b) Suppose $\bar{\boldsymbol{p}}_n := (\bar{p}_{n,1},\dots,\bar{p}_{n,s},\bar{\rho}_n)^\top \in \mathbb{R}_+^{s+1}$ satisfies

$$\|\bar{\boldsymbol{p}}_n - \widehat{\underline{\boldsymbol{p}}}_n^s(\boldsymbol{T}_n(k_n))\| \leq \|\widehat{\underline{\boldsymbol{p}}}_n^s(\widetilde{\boldsymbol{T}}_n(k_n)) - \widehat{\underline{\boldsymbol{p}}}_n^s(\boldsymbol{T}_n(k_n))\|, \quad n \in \mathbb{N}.$$

Then as $n \to \infty$,

$$\begin{aligned}&\big(\widehat{\underline{\boldsymbol{p}}}_n^s(\widetilde{\boldsymbol{T}}_n(k_n)) - \widehat{\underline{\boldsymbol{p}}}_n^s(\boldsymbol{T}_n(k_n))\big)^\top\\ &\qquad\cdot\left(\nabla^2\ell^2\big(\bar{\boldsymbol{p}}_n \,|\, \boldsymbol{T}_n(k_n)\big) - 2k_n\operatorname{diag}\left(p_{n,1},\dots,p_{n,s},\frac{\rho_n}{(r-s)}\right)^{-1}\right)\\ &\qquad\cdot\big(\widehat{\underline{\boldsymbol{p}}}_n^s(\widetilde{\boldsymbol{T}}_n(k_n)) - \widehat{\underline{\boldsymbol{p}}}_n^s(\boldsymbol{T}_n(k_n))\big) \xrightarrow{\mathbb{P}} 0.\end{aligned}$$

Proof of Theorem 4.1. Using a Taylor expansion of $\ell^2(\widehat{\underline{\boldsymbol{p}}}_n^s(\widetilde{\boldsymbol{T}}_n(k_n)) \,|\, \boldsymbol{T}_n(k_n))$ at $\widehat{\underline{\boldsymbol{p}}}_n^s(\boldsymbol{T}_n(k_n))$ yields the existence of a random vector $\bar{\boldsymbol{p}}_n := (\bar{p}_{n,1},\dots,\bar{p}_{n,s},\bar{\rho}_n)^\top$ with

$$\|\bar{\boldsymbol{p}}_n - \widehat{\underline{\boldsymbol{p}}}_n^s(\boldsymbol{T}_n(k_n))\| \leq \|\widehat{\underline{\boldsymbol{p}}}_n^s(\widetilde{\boldsymbol{T}}_n(k_n)) - \widehat{\underline{\boldsymbol{p}}}_n^s(\boldsymbol{T}_n(k_n))\|$$

such that

$$\begin{aligned}&\ell^2(\widehat{\underline{\boldsymbol{p}}}_n^s(\widetilde{\boldsymbol{T}}_n(k_n)) \,|\, \boldsymbol{T}_n(k_n))\\ &= \ell^2(\widehat{\underline{\boldsymbol{p}}}_n^s(\boldsymbol{T}_n(k_n)) \,|\, \boldsymbol{T}_n(k_n)) + \nabla\ell^2(\widehat{\underline{\boldsymbol{p}}}_n^s(\boldsymbol{T}_n(k_n)) \,|\, \boldsymbol{T}_n(k_n))\big(\widehat{\underline{\boldsymbol{p}}}_n^s(\widetilde{\boldsymbol{T}}_n(k_n)) - \widehat{\underline{\boldsymbol{p}}}_n^s(\boldsymbol{T}_n(k_n))\big)\\ &\quad + \frac{1}{2}\big(\widehat{\underline{\boldsymbol{p}}}_n^s(\widetilde{\boldsymbol{T}}_n(k_n)) - \widehat{\underline{\boldsymbol{p}}}_n^s(\boldsymbol{T}_n(k_n))\big)^\top\nabla^2\ell^2\big(\bar{\boldsymbol{p}}_n \,|\, \boldsymbol{T}_n(k_n)\big)\big(\widehat{\underline{\boldsymbol{p}}}_n^s(\widetilde{\boldsymbol{T}}_n(k_n)) - \widehat{\underline{\boldsymbol{p}}}_n^s(\boldsymbol{T}_n(k_n))\big).\end{aligned}$$

Applying Lemma A.4 gives

$$\begin{aligned}&\ell^2(\widehat{\underline{\boldsymbol{p}}}_n^s(\widetilde{\boldsymbol{T}}_n(k_n)) \,|\, \boldsymbol{T}_n(k_n))\\ &= \ell^2(\widehat{\underline{\boldsymbol{p}}}_n^s(\boldsymbol{T}_n(k_n)) \,|\, \boldsymbol{T}_n(k_n)) + \big(\widehat{\underline{\boldsymbol{p}}}_n^s(\widetilde{\boldsymbol{T}}_n(k_n)) - \widehat{\underline{\boldsymbol{p}}}_n^s(\boldsymbol{T}_n(k_n))\big)^\top\\ &\qquad\cdot k_n\operatorname{diag}(p_{n,1},\dots,p_{n,s},\rho_n/(r-s))^{-1}\big(\widehat{\underline{\boldsymbol{p}}}_n^s(\widetilde{\boldsymbol{T}}_n(k_n)) - \widehat{\underline{\boldsymbol{p}}}_n^s(\boldsymbol{T}_n(k_n))\big) + o_{\mathbb{P}}(1)\\ &= k_n\sum_{j=s+1}^{r}\frac{1}{\widehat{\rho}_n^s(\boldsymbol{T}_n(k_n))}\left(\frac{T_{n,j}(k_n)}{k_n} - \widehat{\underline{\rho}}_n^s(\boldsymbol{T}_n(k_n))\right)^2 + \big(\widehat{\underline{\boldsymbol{p}}}_n^s(\widetilde{\boldsymbol{T}}_n(k_n)) - \widehat{\underline{\boldsymbol{p}}}_n^s(\boldsymbol{T}_n(k_n))\big)^\top\\ &\qquad\cdot k_n\operatorname{diag}(p_{n,1},\dots,p_{n,s},\rho_n/(r-s))^{-1}\big(\widehat{\underline{\boldsymbol{p}}}_n^s(\widetilde{\boldsymbol{T}}_n(k_n)) - \widehat{\underline{\boldsymbol{p}}}_n^s(\boldsymbol{T}_n(k_n))\big) + o_{\mathbb{P}}(1).\end{aligned}$$

Next, we move some terms on the right-hand side and use Lemma A.3, which result in

$$\begin{aligned}
&\ell^2(\widehat{\underline{\boldsymbol{p}}}_n^s(\widetilde{\boldsymbol{T}}_n(k_n)) \,|\, \boldsymbol{T}_n(k_n)) - k_n \sum_{j=s+1}^{r} \frac{1}{\rho_n}\left(\frac{T_{n,j}(k_n)}{k_n} - \widehat{\underline{\rho}}_n^s(\boldsymbol{T}_n(k_n))\right)^2 \\
&= \left(\widehat{\underline{\boldsymbol{p}}}_n^s(\widetilde{\boldsymbol{T}}_n(k_n)) - \widehat{\underline{\boldsymbol{p}}}_n^s(\boldsymbol{T}_n(k_n))\right)^\top k_n \mathrm{diag}(p_{n,1},\dots,p_{n,s},\rho_n/(r-s))^{-1} \\
&\quad \cdot\left(\widehat{\underline{\boldsymbol{p}}}_n^s(\widetilde{\boldsymbol{T}}_n(k_n)) - \widehat{\underline{\boldsymbol{p}}}_n^s(\boldsymbol{T}_n(k_n))\right) + o_{\mathbb{P}}(1) \\
&= \mathbf{U}_n^\top \mathbf{U}_n + o_{\mathbb{P}}(1) \\
&\xrightarrow{\mathcal{D}} \mathbf{U}^\top \mathbf{U},
\end{aligned}$$

as $n \to \infty$, where $\mathbf{U}^\top\mathbf{U} \sim 2\chi_s^2$. Since $\mathbb{E}[\mathbf{U}^\top\mathbf{U}] = 2s$ the assertion follows. □

A.2.2. *Proof of Theorem 4.3*

Proof of Theorem 4.3.

Step 1: Suppose $s < s^*$. An application of Lemma 2.8(b,c) gives on the one hand,

$$\begin{aligned}
&\frac{1}{\sum_{l=s+1}^{r} \frac{T_{n,l}(k_n)}{k_n(r-s)}} \sum_{j=s+1}^{r}\left(\frac{T_{n,j}(k_n)}{k_n} - \sum_{i=s+1}^{r} \frac{T_{n,i}(k_n)}{k_n(r-s)}\right)^2 \\
&\qquad \xrightarrow{\mathbb{P}} \frac{1}{\sum_{l=s+1}^{s^*} \frac{p_l}{r-s}} \sum_{j=s+1}^{s^*}\left(p_j - \sum_{i=s+1}^{s^*} \frac{p_i}{r-s}\right)^2,
\end{aligned}$$

where we already applied that $p_j = 0$ for $j = s^*, \dots, r$. Moreover,

$$p_{s+1} - \sum_{i=s+1}^{s^*} \frac{p_i}{r-s} \geq p_{s+1} - \frac{s^*-s}{r-s} p_{s+1} = \frac{r-s^*}{r-s} p_{s+1} > 0.$$

Hence,

$$\frac{k_n}{\sum_{l=s+1}^{r} \frac{T_{n,l}(k_n)}{k_n(r-s)}} \sum_{j=s+1}^{r}\left(\frac{T_{n,j}(k_n)}{k_n} - \sum_{i=s+1}^{r} \frac{T_{n,i}(k_n)}{k_n(r-s)}\right)^2 \xrightarrow{\mathbb{P}} \infty. \tag{A.8}$$

On the other hand, define

$$\boldsymbol{V}_n := \sqrt{k_n\rho_n}\left(\frac{\boldsymbol{T}_{n,\{s^*+1,\dots,r\}}(k_n)}{\rho_n k_n} - \mathbf{1}_{r-s^*}\right) \quad \text{and} \quad \boldsymbol{V} \sim \mathcal{N}_{r-s^*}(\mathbf{0}_{r-s^*}, \boldsymbol{I}_{r-s^*}).$$

By Assumption (A3) we have $\boldsymbol{V}_n \xrightarrow{\mathcal{D}} \boldsymbol{V}$. Furthermore, since $T_{n,l}(k_n)/(k_n\rho_n) \xrightarrow{\mathbb{P}} 1$ for $l = s^*+1, \dots, r$ by Lemma 2.8(c), it follows that

$$\begin{aligned}
&\frac{\rho_n}{\sum_{l=s^*+1}^{r} \frac{T_{n,l}(k_n)}{k_n(r-s^*)}} \frac{k_n}{\rho_n} \sum_{j=s^*+1}^{r}\left(\frac{T_{n,j}(k_n)}{k_n} - \sum_{i=s^*+1}^{r} \frac{T_{n,i}(k_n)}{k_n(r-s^*)}\right)^2 \\
&= \underbrace{\frac{\rho_n}{\sum_{l=s^*+1}^{r} \frac{T_{n,l}(k_n)}{k_n(r-s^*)}}}_{\xrightarrow{\mathbb{P}} 1} \underbrace{\boldsymbol{V}_n^\top\left(\boldsymbol{I}_{r-s^*} - \frac{1}{r-s^*}\mathbf{1}_{r-s^*}\mathbf{1}_{r-s^*}^\top\right)^\top\left(\boldsymbol{I}_{r-s^*} - \frac{1}{r-s^*}\mathbf{1}_{r-s^*}\mathbf{1}_{r-s^*}^\top\right)\boldsymbol{V}_n}_{\xrightarrow{\mathcal{D}} \chi^2_{r-s^*-1}}
\end{aligned}$$

$$\xrightarrow{\mathcal{D}} \chi^2_{r-s^*-1} = O_{\mathbb{P}}(1). \tag{A.9}$$

Combining (A.8) and (A.9) yields

$$\begin{aligned}
&\mathrm{MSEIC}_{k_n}(s) - \mathrm{MSEIC}_{k_n}(s^*) \\
&= \quad 2(s-s^*) + \frac{k_n}{\sum_{l=s+1}^r \frac{T_{n,l}(k_n)}{k_n(r-s)}} \sum_{j=s+1}^r \left(\frac{T_{n,j}(k_n)}{k_n} - \sum_{i=s+1}^r \frac{T_{n,i}(k_n)}{k_n(r-s)} \right)^2 \\
&\quad - \frac{k_n}{\sum_{l=s^*+1}^r \frac{T_{n,l}(k_n)}{k_n(r-s^*)}} \sum_{j=s^*+1}^r \left(\frac{T_{n,j}(k_n)}{k_n} - \sum_{i=s^*+1}^r \frac{T_{n,i}(k_n)}{k_n(r-s^*)} \right)^2 \\
&\xrightarrow{\mathbb{P}} \quad \infty.
\end{aligned}$$

Step 2: Suppose $s > s^*$. An application of (A.9) and Lemma 2.8(c) yield

$$\begin{aligned}
&\frac{k_n}{\sum_{l=s^*+1}^r \frac{T_{n,l}(k_n)}{k_n(r-s^*)}} \sum_{j=s^*+1}^r \left(\frac{T_{n,j}(k_n)}{k_n} - \sum_{i=s^*+1}^r \frac{T_{n,i}(k_n)}{k_n(r-s^*)} \right)^2 \\
&\qquad - \frac{k_n}{\rho_n} \sum_{j=s^*+1}^r \left(\frac{T_{n,j}(k_n)}{k_n} - \sum_{i=s^*+1}^r \frac{T_{n,i}(k_n)}{k_n(r-s^*)} \right)^2 \\
&\qquad = \underbrace{\left(\frac{\rho_n}{\sum_{l=s^*+1}^r \frac{T_{n,l}(k_n)}{k_n(r-s^*)}} - 1 \right)}_{\xrightarrow{\mathbb{P}} 0} \underbrace{\frac{k_n}{\rho_n} \sum_{j=s^*+1}^r \left(\frac{T_{n,j}(k_n)}{k_n} - \sum_{i=s^*+1}^r \frac{T_{n,i}(k_n)}{k_n(r-s^*)} \right)^2}_{\xrightarrow{\mathcal{D}} \chi^2_{r-s^*-1} \text{ by (A.9)}} \\
&\qquad = o_{\mathbb{P}}(1).
\end{aligned} \tag{A.10}$$

Since $s > s^*$ the analog holds when s^* is replaced by s. Using $\boldsymbol{V}_n = (V_{n,s^*+1}, \ldots, V_{n,r})^\top$ defined as above, we have the representation $\frac{T_{n,j}(k_n)}{k_n \rho_n} = \frac{1}{\sqrt{k_n \rho_n}} V_{n,j} + 1$. Thus, when inserting the definition of MSEIC we get with (A.10) that

$$\begin{aligned}
&\mathrm{MSEIC}_{k_n}(s) - \mathrm{MSEIC}_{k_n}(s^*) \\
&= 2(s-s^*) + \frac{k_n}{\rho_n} \sum_{j=s+1}^r \left(\frac{T_{n,j}(k_n)}{k_n} - \sum_{i=s+1}^r \frac{T_{n,i}(k_n)}{k_n(r-s)} \right)^2 \\
&\quad - \frac{k_n}{\rho_n} \sum_{j=s^*+1}^r \left(\frac{T_{n,j}(k_n)}{k_n} - \sum_{i=s^*+1}^r \frac{T_{n,i}(k_n)}{k_n(r-s^*)} \right)^2 + o_{\mathbb{P}}(1) \\
&= 2(s-s^*) + \sum_{j=s+1}^r \left\{ \sqrt{k_n \rho_n} \left(\frac{T_{n,j}(k_n)}{k_n \rho_n} - 1 \right) - \frac{\sqrt{k_n \rho_n}}{(r-s)} \sum_{i=s+1}^r \left(\frac{T_{n,i}(k_n)}{k_n \rho_n} - 1 \right) \right\}^2 \\
&\quad - \sum_{j=s^*+1}^r \left\{ \sqrt{k_n \rho_n} \left(\frac{T_{n,j}(k_n)}{k_n \rho_n} - 1 \right) - \frac{\sqrt{k_n \rho_n}}{(r-s^*)} \sum_{i=s^*+1}^r \left(\frac{T_{n,i}(k_n)}{k_n \rho_n} - 1 \right) \right\}^2 + o_{\mathbb{P}}(1) \\
&= 2(s-s^*) + \sum_{j=s+1}^r \left\{ V_{n,j} - \frac{1}{(r-s)} \sum_{i=s+1}^r V_{n,i} \right\}^2
\end{aligned}$$

$$- \sum_{j=s^*+1}^{r} \left\{ V_{n,j} - \frac{1}{(r-s^*)} \sum_{i=s^*+1}^{r} V_{n,i} \right\}^2 + o_{\mathbb{P}}(1)$$

$$\xrightarrow{\mathcal{D}} 2(s-s^*) + \sum_{j=s+1}^{r} \left\{ V_j - \frac{1}{(r-s)} \sum_{i=s+1}^{r} V_i \right\}^2 - \sum_{j=s^*+1}^{r} \left\{ V_j - \frac{1}{(r-s^*)} \sum_{i=s^*+1}^{r} V_i \right\}^2 .$$

Similar to the proof of Theorem 3.1, there exists a positive probability that the right-hand side is positive. Hence, the assertion follows. □

A.2.3. Proof of Theorem 4.4

Before we are able to present the proof of Theorem 4.4 we require some auxiliary lemmata whose proofs are moved to Section 2 in the Supplementary Material A [5].

Lemma A.5. *Suppose assumptions (B1) and (B2) hold. Then for $\boldsymbol{p}' \in \mathbb{R}_+^r$ the asymptotic behavior*

$$\mathbb{E}\left[\left\| \sqrt{n - T'_{n,2^d}} \operatorname{diag}(\boldsymbol{p}')^{-1/2} \left(\frac{\boldsymbol{T}'_{n,\{1,\dots,r\}}}{n - T'_{n,2^d}} - \boldsymbol{p}' \right) \right\|_2^2 \right] = nq_n \left(\frac{1}{k_n} \mathbb{E}[\ell^2(\boldsymbol{p}' | \boldsymbol{T}_n(k_n))] + o\left(\frac{1}{nq_n} \right) \right)$$

as $n \to \infty$ holds.

Lemma A.6. *For $q' \in (0,1)$ the equality*

$$\mathbb{E}\left[\left\| \sqrt{n}(q'(1-q'))^{-1/2} \left(\frac{T'_{n,2^d}}{n} - (1-q') \right) \right\|_2^2 \right] = nq_n \left(\frac{(1-q_n)}{nq'(1-q')} + \frac{(q'-q_n)^2}{q_n q'(1-q')} \right)$$

holds.

Proof of Theorem 4.4. For $q' \in (0,1)$ and $\boldsymbol{p}' \in \mathbb{R}_+^r$ we have as a consequence of Lemmas A.5 and A.6, that

$$\begin{aligned} & q' \mathbb{E}\left[\left\| \sqrt{n - T'_{n,2^d}} \operatorname{diag}(\boldsymbol{p}')^{-1/2} \left(\frac{\boldsymbol{T}'_{n,\{1,\dots,r\}}}{n - T'_{n,2^d}} - \boldsymbol{p}' \right) \right\|_2^2 \right] \\ & + (1-q') \mathbb{E}\left[\left\| \sqrt{n}(q'(1-q'))^{-1/2} \left(\frac{T'_{n,2^d}}{n} - (1-q') \right) \right\|_2^2 \right] \\ & = nq_n \left(\frac{q'}{k_n} \mathbb{E}[\ell^2(\boldsymbol{p}' | \boldsymbol{T}_n(k_n))] + o\left(\frac{q'}{nq_n} \right) \right) + nq_n \left(\frac{(1-q_n)}{nq'} + \frac{(q'-q_n)^2}{q_n q'} \right). \end{aligned}$$

Therefore, it follows that

$$\mathbb{E}\left[q' \mathbb{E}\left[\left\| \sqrt{n - T'_{n,2^d}} \operatorname{diag}(\boldsymbol{p}')^{-1/2} \left(\frac{\boldsymbol{T}'_{n,\{1,\dots,r\}}}{n - T'_{n,2^d}} - \boldsymbol{p}' \right) \right\|_2^2 \right] \Big|_{\boldsymbol{p}' = \widehat{\boldsymbol{p}}_n\left(\frac{\widetilde{\boldsymbol{T}}_n(k_n)}{k_n}\right), q' = \frac{k_n}{n}} \right]$$

$$
\begin{aligned}
&+\mathbb{E}\left[(1-q')\mathbb{E}\left[\left\|\sqrt{n}(q'(1-q'))^{-1/2}\left(\frac{T'_{n,2^d}}{n}-(1-q')\right)\right\|_2^2\right]\Big|_{q'=\frac{k_n}{n}}\right]\\
&=nq_n\mathbb{E}\left[\left(\frac{q'}{k_n}\mathbb{E}[\ell^2(\boldsymbol{p}'|\boldsymbol{T}_n(k_n))]+o\left(\frac{q'}{nq_n}\right)\right)\Big|_{\boldsymbol{p}'=\widehat{\boldsymbol{p}}_n(\frac{\widetilde{\boldsymbol{T}}_n(k_n)}{k_n}),q'=\frac{k_n}{n}}\right]\\
&\quad+nq_n\mathbb{E}\left[\left(\frac{(1-q_n)}{nq'}+\frac{(q'-q_n)^2}{q_nq'}\right)\Big|_{q'=\frac{k_n}{n}}\right]\\
&=nq_n\left(\frac{1}{n}\,\mathrm{MSE}_{k_n}(s)+\frac{(1-q_n)}{k_n}+\frac{(\frac{k_n}{n}-q_n)^2}{q_n\frac{k_n}{n}}+o\left(n^{-1}\right)\right).
\end{aligned}
$$

Due to the asymptotic behavior as $n\to\infty$,

$$
\frac{(\frac{k_n}{n}-q_n)^2}{q_n\frac{k_n}{n}}+o\left(n^{-1}\right)=\frac{k_n(1-\frac{nq_n}{k_n})^2}{nq_n}+o\left(n^{-1}\right)=o((nq_n)^{-1})+o\left(n^{-1}\right)=o((nq_n)^{-1}),
$$

where we used the additional assumption $k_n(1-\frac{nq_n}{k_n})^2\to 0$ as $n\to\infty$, we can conclude the statement. □

A.3. Proofs of Section 5

A.3.1. Proof of Theorem 5.2

In the next two lemmata, we derive auxiliary results used for the derivation of an upper bound of the posterior probability $\mathbb{P}(M^s_{k_n}|\boldsymbol{T}_n(k_n))$. First, in Lemma A.7, we give a Taylor approximation of the log-likelihood function $\log(L_{M^s_{k_n}}(\,\cdot\,|\boldsymbol{T}_n(k_n)))$ of Model $M^s_{k_n}$, and second, in Lemma A.8, we present boundaries for the eigenvalues of the Hessian of the log-likelihood function; the proofs of these auxiliary results are included in Section 3.1 of the Supplementary Material A [5]. Finally, for the proof of the upper bound of the log-posterior distribution in Theorem 5.2 we combine these two results.

Lemma A.7. *Let the assumptions of Theorem 5.2 hold. Define the ball*

$$
U_{\varepsilon_{n,\gamma}}(\widehat{\boldsymbol{p}}^s_n):=\{\widetilde{\boldsymbol{p}}^s\in\Theta_s:\|\widetilde{\boldsymbol{p}}^s-\widehat{\boldsymbol{p}}^s_n\|_2<\varepsilon_{n,\gamma}\}
$$

with radius $\varepsilon_{n,\gamma}:=(\rho_n)^\gamma/2$ *for* $\gamma\geq 4/3$ *around* $\widehat{\boldsymbol{p}}^s_n$. *Then the following statement holds*

$$
\begin{aligned}
\sup_{\widetilde{\boldsymbol{p}}^s\in U_{\varepsilon_{n,\gamma}}(\widehat{\boldsymbol{p}}^s_n)}\Big|&\log L_{M^s_{k_n}}(\widetilde{\boldsymbol{p}}^s\,|\,\boldsymbol{T}_n(k_n))-\log L_{M^s_{k_n}}(\widehat{\boldsymbol{p}}^s_n\,|\,\boldsymbol{T}_n(k_n))\\
&-\frac{1}{2}(\widetilde{\boldsymbol{p}}^s-\widehat{\boldsymbol{p}}^s_n)^\top\nabla^2\log L_{M^s_{k_n}}(\widehat{\boldsymbol{p}}^s_n\,|\,\boldsymbol{T}_n(k_n))(\widetilde{\boldsymbol{p}}^s-\widehat{\boldsymbol{p}}^s_n)\Big|=o_{\mathbb{P}}(1).
\end{aligned}
$$

Lemma A.8. *Let the assumptions of Theorem 5.2 hold. Define* $\lambda_{n,2}:=k_n/T_{n,1}(k_n)$ *and* $\lambda_{n,1}:=k_n/T_{n,s}(k_n)+sk_n/\sum_{j=s+1}^r T_{n,j}$. *For* $\widetilde{\boldsymbol{p}}^s\in\Theta_s$ *we have on the one hand,*

$$
\lambda_{n,2}(\widetilde{\boldsymbol{p}}^s-\widehat{\boldsymbol{p}}^s_n)^\top(\widetilde{\boldsymbol{p}}^s-\widehat{\boldsymbol{p}}^s_n)\leq(\widetilde{\boldsymbol{p}}^s-\widehat{\boldsymbol{p}}^s_n)^\top\frac{-1}{k_n}\nabla^2\log L_{M^s_{k_n}}(\widehat{\boldsymbol{p}}^s_n\,|\,\boldsymbol{T}_n(k_n))(\widetilde{\boldsymbol{p}}^s-\widehat{\boldsymbol{p}}^s_n)\quad\mathbb{P}\text{-a.s.}
$$

and on the other hand,

$$
\lambda_{n,1}(\widetilde{\boldsymbol{p}}^s-\widehat{\boldsymbol{p}}^s_n)^\top(\widetilde{\boldsymbol{p}}^s-\widehat{\boldsymbol{p}}^s_n)\geq(\widetilde{\boldsymbol{p}}^s-\widehat{\boldsymbol{p}}^s_n)^\top\frac{-1}{k_n}\nabla^2\log L_{M^s_{k_n}}(\widehat{\boldsymbol{p}}^s_n\,|\,\boldsymbol{T}_n(k_n))(\widetilde{\boldsymbol{p}}^s-\widehat{\boldsymbol{p}}^s_n)\quad\mathbb{P}\text{-a.s.}
$$

Proof of Theorem 5.2. In the following let $\gamma = 4/3$ and $\varepsilon_n \coloneqq \varepsilon_{n,4/3} = (\rho_n)^{4/3}/2$. An application of Lemma A.7, Lemma A.8 and Assumption (C1) give

$$\begin{aligned}
&-2\log\mathbb{E}_{g_s}[L_{M^s_{k_n}}(\widetilde{\boldsymbol{p}}^s \mid \boldsymbol{T}_n(k_n))] \\
&\quad\le -2\log\int_{U_{\varepsilon_n}(\widehat{\boldsymbol{p}}^s_n)} L_{M^s_{k_n}}(\widetilde{\boldsymbol{p}}^s \mid \boldsymbol{T}_n(k_n))\,\mathrm{d}\widetilde{\boldsymbol{p}}^s - 2\log b \\
&\quad\le -2\log L_{M^s_{k_n}}(\widehat{\boldsymbol{p}}^s_n \mid \boldsymbol{T}_n(k_n)) \\
&\qquad -2\log\int_{U_{\varepsilon_n}(\widehat{\boldsymbol{p}}^s_n)} \exp\left\{\frac{-k_n}{2}(\widetilde{\boldsymbol{p}}^s-\widehat{\boldsymbol{p}}^s_n)^\top \frac{-1}{k_n}\nabla^2 \log L_{M^s_{k_n}}(\widehat{\boldsymbol{p}}^s_n \mid \boldsymbol{T}_n(k_n))(\widetilde{\boldsymbol{p}}^s-\widehat{\boldsymbol{p}}^s_n)\right\}\mathrm{d}\widetilde{\boldsymbol{p}}^s \\
&\qquad -2\log b + o_{\mathbb{P}}(1) \\
&\quad\le -2\log L_{M^s_{k_n}}(\widehat{\boldsymbol{p}}^s_n \mid \boldsymbol{T}_n(k_n)) - s\log(2\pi) + s\log(k_n\lambda_{n,1}) - 2\log b \\
&\qquad -2\log\int_{U_{\varepsilon_n}(\widehat{\boldsymbol{p}}^s_n)} \left(\frac{k_n\lambda_{n,1}}{2\pi}\right)^{s/2} \exp\left\{\frac{-1}{2}\frac{(\widetilde{\boldsymbol{p}}^s-\widehat{\boldsymbol{p}}^s_n)^\top(\widetilde{\boldsymbol{p}}^s-\widehat{\boldsymbol{p}}^s_n)}{1/(k_n\lambda_{n,1})}\right\}\mathrm{d}\widetilde{\boldsymbol{p}}^s + o_{\mathbb{P}}(1). \qquad \text{(A.11)}
\end{aligned}$$

The integrand is a s-dimensional Gaussian density with expectation vector $\widehat{\boldsymbol{p}}^s_n$ and covariance matrix $(k_n\lambda_{n,1})^{-1}\boldsymbol{I}_s$. Furthermore, due to the definition of $\lambda_{n,1}$, Assumption (C3) and Lemma 2.8, the asymptotic behavior

$$0 \le k_n\lambda_{n,1}\varepsilon_n^2 = \underbrace{\frac{k_n(\rho_n)^{5/3}}{4}}_{\to\infty} \underbrace{\left(\frac{k_n\rho_n}{T_{n,s}(k_n)} + \frac{sk_n\rho_n}{\sum_{j=s+1}^r T_{n,j}}\right)}_{\overset{\text{Lemma 2.8}}{\to} \mathbb{1}_{\{s\ge s^*\}} + \frac{s}{r-\max(s,s^*)}} \xrightarrow{\mathbb{P}} \infty \qquad \text{(A.12)}$$

holds in probability. Let $\boldsymbol{N} \sim \mathcal{N}_s(\boldsymbol{0}_s, \boldsymbol{I}_s)$. Since $\|\boldsymbol{N}\|_2^2 \sim \chi^2_s$ the Markov inequality yields

$$\begin{aligned}
&\int_{U_{\varepsilon_n}(\widehat{\boldsymbol{p}}^s_n)} \left(\frac{k_n\lambda_{n,1}}{2\pi}\right)^{s/2} \exp\left\{\frac{-1}{2}\frac{(\widetilde{\boldsymbol{p}}^s-\widehat{\boldsymbol{p}}^s_n)^\top(\widetilde{\boldsymbol{p}}^s-\widehat{\boldsymbol{p}}^s_n)}{1/(k_n\lambda_{n,1})}\right\}\mathrm{d}\widetilde{\boldsymbol{p}}^s \\
&\qquad = \mathbb{P}\left(\widehat{\boldsymbol{p}}^s_n + \frac{1}{\sqrt{(k_n\lambda_{n,1})}}\boldsymbol{N} \in U_{\varepsilon_n}(\widehat{\boldsymbol{p}}^s_n)\,\middle|\,\boldsymbol{T}_n(k_n)\right) \\
&\qquad = 1 - \mathbb{P}\left(\|\boldsymbol{N}\|_2^2 \ge k_n\lambda_{n,1}\varepsilon_n^2\middle|\,\boldsymbol{T}_n(k_n)\right) \\
&\qquad \ge 1 - \frac{s}{k_n\lambda_{n,1}\varepsilon_n^2} \to 1,
\end{aligned}$$

as $n \to \infty$ almost surely, where we used in the last step (A.12). Thus,

$$-2\log\int_{U_{\varepsilon_n}(\widehat{\boldsymbol{p}}^s_n)} \left(\frac{k_n\lambda_{n,1}}{2\pi}\right)^{s/2} \exp\left\{\frac{-1}{2}\frac{(\widetilde{\boldsymbol{p}}^s-\widehat{\boldsymbol{p}}^s_n)^\top(\widetilde{\boldsymbol{p}}^s-\widehat{\boldsymbol{p}}^s_n)}{1/(k_n\lambda_{n,1})}\right\}\mathrm{d}\widetilde{\boldsymbol{p}}^s = o_{\mathbb{P}}(1). \qquad \text{(A.13)}$$

Inserting (A.13) into (A.11) gives then

$$\begin{aligned}
&-2\log\mathbb{E}_{g_s}[L_{M^s_{k_n}}(\widetilde{\boldsymbol{p}}^s \mid \boldsymbol{T}_n(k_n))] \\
&\qquad \le -2\log L_{M^s_{k_n}}(\widehat{\boldsymbol{p}}^s_n \mid \boldsymbol{T}_n(k_n)) - s\log(2\pi) + s\log(k_n\lambda_{n,1}) - 2\log b + o_{\mathbb{P}}(1).
\end{aligned}$$

Since $T_{n,j}(k_n) \ge 1$ for $j = 1, \dots, s$, we receive the upper bound

$$\lambda_{n,1} = \left(\frac{k_n}{T_{n,s}(k_n)} + \frac{sk_n}{\sum_{j=s+1}^r T_{n,j}(k_n)}\right) \le k_n\left(1 + \frac{s}{r-s}\right) = k_n\frac{r}{r-s},$$

and finally,

$$-2\log\mathbb{E}_{g_s}\left[L_{M^s_{k_n}}(\widetilde{\boldsymbol{p}}^s \mid \boldsymbol{T}_n(k_n))\right]$$
$$\leq -2\log L_{M^s_{k_n}}(\widehat{\boldsymbol{p}}^s_n \mid \boldsymbol{T}_n(k_n)) - s\log(2\pi) + 2s\log\left(k_n\sqrt{\frac{r}{r-s}}\right) - 2\log b + o_{\mathbb{P}}(1),$$

which is the statement. □

A.3.2. Proof of Theorem 5.5

Proof of Theorem 5.5.
(a) Note that

$$\mathrm{BICU}_{k_n}(s) = 2\,\mathrm{AIC}_{k_n}(s) - 2s + 2s\log(k_n) + s\log\left(\frac{r}{2\pi(r-s)}\right).$$

We consider now the different cases $s > s^*$ and $s < s^*$ separately.
Step 1: Suppose $s > s^*$. We receive with (A.4) that

$$\begin{aligned}
&\mathrm{BICU}_{k_n}(s) - \mathrm{BICU}_{k_n}(s^*)\\
&\quad = 2\,\mathrm{AIC}_{k_n}(s) - 2s + 2s\log(k_n) + s\log\left(\frac{r}{2\pi(r-s)}\right)\\
&\qquad - 2\,\mathrm{AIC}_{k_n}(s^*) + 2s^* - 2s^*\log(k_n) - s^*\log\left(\frac{r}{2\pi(r-s^*)}\right)\\
&\quad = 2(s-s^*)\log(k_n) + O_{\mathbb{P}}(1).
\end{aligned}$$

Dividing the last equation by $\log(k_n)$ results in

$$\frac{\mathrm{BICU}_{k_n}(s) - \mathrm{BICU}_{k_n}(s^*)}{\log(k_n)} \xrightarrow{\mathbb{P}} 2(s-s^*) > 0,$$

where we used $\log(k_n) \to \infty$.
Step 2: Suppose $s < s^*$. Here we have as in the proof of Theorem 3.1 and due to $\log(k_n)/k_n \to 0$ that

$$\begin{aligned}
&\frac{\mathrm{BICU}_{k_n}(s) - \mathrm{BICU}_{k_n}(s^*)}{k_n}\\
&\quad = 2\frac{\mathrm{AIC}_{k_n}(s) - \mathrm{AIC}_{k_n}(s^*)}{k_n} + \frac{-2s + 2s\log(k_n) + s\log\left(\frac{r}{2\pi(r-s)}\right)}{k_n}\\
&\qquad + \frac{2s^* - 2s^*\log(k_n) - s^*\log\left(\frac{r}{2\pi(r-s^*)}\right)}{k_n}\\
&\quad \xrightarrow{\mathcal{D}} 2\sum_{i=s+1}^{s^*} p_i\left(\log(p_i) - \log\left(\frac{1}{r-s}\sum_{j=s+1}^{s^*} p_j\right)\right) > 0,
\end{aligned}$$

and thus, the assertion follows.
(b) Again, note that

$$\mathrm{BICL}_{k_n}(s) = 2\,\mathrm{AIC}_{k_n}(s) - 2s + s\log(k_n) + s\log\left(\frac{k_n}{2\pi T_{n,1}(k_n)}\right).$$

By a calculation analog to part (a), the BICL is also consistent since $s \log\left(\frac{k_n}{2\pi T_{n,1}(k_n)}\right) \xrightarrow{\mathbb{P}} s\log\left(\frac{1}{2\pi p_1}\right) > 0$ as $n \to \infty$. □

A.3.3. *Proof of Theorem 5.7*

First, we derive some auxiliary results before we prove Theorem 5.7. Therefore, note that due (2.7) (cf. Equation (1.23) in the Supplementary Material of Meyer and Wintenberger [29]) and $\sum_{j=1}^{2^d-1} T'_{n,j} = n - T'_{n,2^d}$, the likelihood function of Model M'^s_n can be written as

$$L_{M'^s_n}(\widetilde{\boldsymbol{p}}'^s \mid \boldsymbol{T}'_n) = L_{M^s_{n-T'_{n,2^d}}}(\widetilde{\boldsymbol{p}}^s \mid \boldsymbol{T}'_{n,\{1,\dots,r\}}) \cdot L_{\mathrm{Bin}_n}(1-\widetilde{q} \mid T'_{n,2^d}), \tag{A.14}$$

for $\widetilde{\boldsymbol{p}}'^s = (\widetilde{\boldsymbol{p}}^s, \widetilde{q}) \in \Theta'_s = \Theta_s \times (0,1)$, where

$$L_{\mathrm{Bin}_n}(1-\widetilde{q} \mid T'_{n,2^d}) := \left(\binom{n}{T'_{n,2^d}}\right)(1-\widetilde{q})^{T'_{n,2^d}} \widetilde{q}^{\,n-T'_{n,2^d}} \tag{A.15}$$

is the likelihood function of the binomial model. Next, we define the following expectations with respect to the Lebesgue measure λ. Let

$$\begin{aligned} \mathbb{E}_\lambda[L_{M^s_{n-T'_{n,2^d}}}(\widetilde{\boldsymbol{p}}^s \mid \boldsymbol{T}'_{n,\{1,\dots,r\}})] &:= \int_{\Theta_s} L_{M^s_{n-T'_{n,2^d}}}(\widetilde{\boldsymbol{p}}^s \mid \boldsymbol{T}'_{n,\{1,\dots,r\}})\,\mathrm{d}\widetilde{\boldsymbol{p}}^s, \\ \mathbb{E}_\lambda[L_{\mathrm{Bin}_n}(1-\widetilde{q} \mid T'_{n,2^d})] &:= \int_{(0,1)} L_{\mathrm{Bin}_n}(1-\widetilde{q} \mid T'_{n,2^d})\,\mathrm{d}\widetilde{q}. \end{aligned} \tag{A.16}$$

Then taking the expectation and logarithm in (A.14) results under Assumption (D1) in

$$\begin{aligned} &-2\log \mathbb{E}_{g'_s}[L_{M'^s_n}(\widetilde{\boldsymbol{p}}'^s \mid \boldsymbol{T}'_n)] \\ &\quad \le -2\log b' - 2\log\left\{\int_{\Theta_s\times(0,1)} L_{M^s_{n-T'_{n,2^d}}}(\widetilde{\boldsymbol{p}}^s \mid \boldsymbol{T}'_{n,\{1,\dots,r\}}) \cdot L_{\mathrm{Bin}_n}(1-\widetilde{q} \mid T'_{n,2^d})\,\mathrm{d}(\widetilde{\boldsymbol{p}}^s, \widetilde{q})\right\} \\ &\quad = -2\log b' - 2\log\left\{\int_{\Theta_s} L_{M^s_{n-T'_{n,2^d}}}(\widetilde{\boldsymbol{p}}^s \mid \boldsymbol{T}'_{n,\{1,\dots,r\}})\,\mathrm{d}\widetilde{\boldsymbol{p}}^s \cdot \int_{(0,1)} L_{\mathrm{Bin}_n}(1-\widetilde{q} \mid T'_{n,2^d})\,\mathrm{d}\widetilde{q}\right\} \\ &\quad = -2\log b' - 2\log \mathbb{E}_\lambda[L_{M^s_{n-T'_{n,2^d}}}(\widetilde{\boldsymbol{p}}^s \mid \boldsymbol{T}'_{n,\{1,\dots,r\}})] - 2\log \mathbb{E}_\lambda[L_{\mathrm{Bin}_n}(1-\widetilde{q} \mid T'_{n,2^d})]. \end{aligned} \tag{A.17}$$

In the following two auxiliary lemmata, we determine upper bounds for the expectation of both summands.

Proposition A.9. *Under Assumptions (B1), (B3) and (D4) the asymptotic upper bound as* $n \to \infty$,

$$\begin{aligned} &-2\mathbb{E}\left[\log \mathbb{E}_\lambda[L_{M^s_{n-T'_{n,2^d}}}(\widetilde{\boldsymbol{p}}^s \mid \boldsymbol{T}'_{n,\{1,\dots,r\}})]\right] \\ &\quad \le -2\mathbb{E}\left[\log\left((n-T'_{n,2^d})!\right) - (n-T'_{n,2^d})\left(\log(n-T'_{n,2^d}) - 1\right)\right] \\ &\qquad - 2\frac{nq_n}{k_n}\mathbb{E}[\log L_{M^s_{k_n}}(\widehat{\boldsymbol{p}}^s_n(\boldsymbol{T}_n(k_n)) \mid \boldsymbol{T}_n(k_n))] + 2s\log\left(k_n\sqrt{\frac{r}{2\pi(r-s)}}\right) + C\log(nq_n), \end{aligned}$$

for a constant $C > 0$ *independent of* s *and* n, *holds.*

Proposition A.10. *Suppose Assumptions (D3) and (B3) hold. The expectation of the binomial likelihood satisfies as $n \to \infty$ the inequality*

$$-2\mathbb{E}[\log \mathbb{E}_\lambda[L_{\mathrm{Bin}_n}(1-\widetilde{q}\,|\,T'_{n,2^d})]] \le -2\log(n!) + 2\mathbb{E}[\log((n-T'_{n,2^d})!)] + 2\mathbb{E}[\log(T'_{n,2^d}!)] - 2nq_n\log(k_n/n) + 2\log(n) + Cnq_n,$$

for a constant $C > 0$ independent of s and n.

Proof of Theorem 5.7. For the ease of notation we define $x\log x$ as zero if $x = 0$. Inserting the bounds derived in Proposition A.9 with constant C_1 and Proposition A.10 with constant C_2 into (A.17) gives for sufficiently large n that

$$\begin{aligned}
&-2\mathbb{E}[\log \mathbb{E}_{g'_s}[L_{M'^s_n}(\widetilde{\boldsymbol{p}}'^s\,|\,\boldsymbol{T}'_n)]] + 2\log b' \\
&\quad\le -2\mathbb{E}[\log \mathbb{E}_\lambda[L_{M^s_{n-T'_{n,2^d}}}(\widetilde{\boldsymbol{p}}^s\,|\,\boldsymbol{T}'_{n,\{1,\dots,r\}})]] - 2\mathbb{E}[\log \mathbb{E}_\lambda[L_{\mathrm{Bin}_n}(1-\widetilde{q}\,|\,T'_{n,2^d})]] \\
&\quad\le -2\mathbb{E}\Big[\log\big((n-T'_{n,2^d})!\big) - (n-T'_{n,2^d})\Big(\log(n-T'_{n,2^d}) - 1\Big)\Big] \\
&\qquad - 2\frac{nq_n}{k_n}\mathbb{E}[\log L_{M^s_{k_n}}(\widehat{\boldsymbol{p}}^s_n\,|\,\boldsymbol{T}_n(k_n))] + 2s\log\left(k_n\sqrt{\frac{r}{2\pi(r-s)}}\right) \\
&\qquad - 2\log(n!) + 2\mathbb{E}[\log((n-T'_{n,2^d})!)] + 2\mathbb{E}[\log(T'_{n,2^d}!)] \\
&\qquad - 2nq_n\log(k_n/n) + 2\log(n) + (C_1+C_2)nq_n \\
&\quad= \Big\{-2\log(n!) + 2\mathbb{E}\Big[(n-T'_{n,2^d})\Big(\log(n-T'_{n,2^d}) - 1\Big)\Big] + 2\mathbb{E}[\log(T'_{n,2^d}!)]\Big\} \\
&\qquad + \Big\{-2\frac{nq_n}{k_n}\mathbb{E}[\log L_{M^s_{k_n}}(\widehat{\boldsymbol{p}}^s_n\,|\,\boldsymbol{T}_n(k_n))] + 2s\log\left(k_n\sqrt{\frac{r}{2\pi(r-s)}}\right) \\
&\qquad\quad - 2nq_n\log(k_n/n) + 2\log(n)\Big\} + (C_1+C_2)nq_n \\
&\quad=: I_{n,1} + I_{n,2} + (C_1+C_2)nq_n. \qquad \text{(A.18)}
\end{aligned}$$

Next, we simplify $I_{n,1}$. Therefore, we use the following calculation. Let B be a positive random variable with finite positive variance. For $u > 0$ and $x > 0$ we the inequality $\log(x/u) \le x/u - 1$ holds, which is equivalent to $x\log(x) \le x^2/u + x\log(u) - x$. Then we have

$$\mathbb{E}[B\log(B)] \le \frac{\mathbb{E}[B^2]}{u} + \mathbb{E}[B]\log(u) - \mathbb{E}[B],$$

and in particular for $u = \mathbb{E}[B^2]/\mathbb{E}[B]$ we receive

$$\mathbb{E}[B\log(B)] \le \mathbb{E}[B]\log(\mathbb{E}[B^2]/\mathbb{E}[B]).$$

Since $\mathbb{E}[T'_{n,2^d}\mathbb{1}\{T'_{n,2^d} > 0\}] = \mathbb{E}[T'_{n,2^d}] = n(1-q_n)$ and $\mathbb{E}[T'^2_{n,2^d}\mathbb{1}\{T'_{n,2^d} > 0\}] = \mathbb{E}[T'^2_{n,2^d}] = nq_n(1-q_n)$ the previous inequality gives

$$\begin{aligned}
&\mathbb{E}[T'_{n,2^d}\log(T'_{n,2^d})\mathbb{1}\{T'_{n,2^d} > 0\}] \\
&\qquad\le n(1-q_n)\log\left(\frac{n^2(1-q_n)^2 + nq_n(1-q_n)}{n(1-q_n)}\right) \\
&\qquad= n(1-q_n)\log\left(n(1-q_n) + q_n\right) \\
&\qquad= n(1-q_n)\log\left(n(1-q_n)\right) + n(1-q_n)\log\left(\frac{n(1-q_n)+q_n}{n(1-q_n)}\right)
\end{aligned}$$

$$\leq n(1-q_n)\log\left(n(1-q_n)\right) + C_3 \tag{A.19}$$

for a constant $C_3 > 0$ independent of s and n. Furthermore, we use the inequality

$$n\log n - n < \log(n!) < n\log n - n + \log n + 1 \tag{A.20}$$

to derive a bound for $\mathbb{E}[\log(T'_{n,2^d}!)]$. Hence, using the upper bound (A.20), (A.19) and applying Jensen inequality we receive that

$$\begin{aligned}
&\mathbb{E}[\log(T'_{n,2^d}!)] \\
&\quad = \mathbb{E}[\log(T'_{n,2^d}!)\mathbb{1}\{T'_{n,2^d} > 0\}] \\
&\quad \leq \mathbb{E}[T'_{n,2^d}\log(T'_{n,2^d})\mathbb{1}\{T'_{n,2^d} > 0\}] - \mathbb{E}[T'_{n,2^d}\mathbb{1}\{T'_{n,2^d} > 0\}] + \mathbb{E}[\log(T'_{n,2^d}\mathbb{1}\{T'_{n,2^d} > 0\})] \\
&\quad \leq n(1-q_n)\log\left(n(1-q_n)\right) - n(1-q_n) + \log\left(n(1-q_n)\right) + C_4
\end{aligned}$$

for a constant $C_4 > 0$ independent of s and n. Additionally to the last inequality, we obtain by (A.20) and (A.19) (for $n - T'_{n,2^d}$ instead of $T'_{n,2^d}$ and q_n instead of $1-q_n$, respectively) that

$$\begin{aligned}
I_{n,1} &= -2\log(n!) + 2\mathbb{E}\left[(n - T'_{n,2^d})\left(\log(n - T'_{n,2^d}) - 1\right)\right] + 2\mathbb{E}[\log(T'_{n,2^d}!)] \\
&< -2n\log(n) + 2n + 2nq_n\log(nq_n) - 2nq_n + 2n(1-q_n)\log(n(1-q_n)) \\
&\quad - 2n(1-q_n) + 2\log(n(1-q_n)) + C_5 \\
&= [-2n\log(n) + 2nq_n\log(n) + 2n(1-q_n)\log(n(1-q_n))] \\
&\quad + [2nq_n\log(q_n) + 2\log(n(1-q_n)) + C_5] \\
&\leq 2nq_n\log(q_n) + 2\log(n) + C_5
\end{aligned} \tag{A.21}$$

for some constant $C_5 > 0$ independent of s and n holds, where we used that the bracket in the second last equation is negative.

Combining (A.18) and (A.21) ends up with

$$\begin{aligned}
&-2\mathbb{E}[\log \mathbb{E}_{g'_s}[L_{M_n'^s}(\widetilde{\boldsymbol{p}}'^s \mid \boldsymbol{T}'_n)]] \\
&\leq I_{n,1} + I_{n,2} - 2\log b' + C_2 nq_n \\
&\leq -2\frac{nq_n}{k_n}\mathbb{E}[\log L_{M^s_{k_n}}(\widehat{\boldsymbol{p}}^s_n \mid \boldsymbol{T}_n(k_n))] + 2s\log\left(k_n\sqrt{\frac{r}{2\pi(r-s)}}\right) \\
&\quad + nq_n\left(2\log\left(\frac{nq_n}{k_n}\right) + \frac{2\log(n)}{nq_n}\right) + nq_n \max_{i=1,\dots,5} C_i \\
&= 2nq_n\left[-\frac{\mathbb{E}[\log L_{M^s_{k_n}}(\widehat{\boldsymbol{p}}^s_n \mid \boldsymbol{T}_n(k_n))]}{k_n} + \frac{s}{nq_n}\log\left(k_n\sqrt{\frac{r}{2\pi(r-s)}}\right) + \frac{\log(n)}{nq_n}\right] + Cnq_n,
\end{aligned}$$

for a constant $C > 0$ independent of s and n. □

Supplementary Material

Supplementary Material A: Proofs of auxiliary results
(doi: 10.1214/25-EJS2469A; .pdf). This supplementary material contains proofs of some auxiliary results of this paper, as well as an additional simulation study.

Supplementary Material B: R-codes
(doi: 10.1214/25-EJS2469B; .zip). The zip archive contains all R-codes used in this paper.

References

[1] Avella Medina, M., Davis, R. A. and Samorodnitsky, G. (2024). Spectral learning of multivariate extremes. *J. Mach. Learn. Res.* **25**. MR4749776

[2] Avella-Medina, M., Davis, R. A. and Samorodnitsky, G. (2025). Insights into Kernel PCA with Application to Multivariate Extremes. *SIAM J. Math. Data Sci.* **7** 777–801. MR4916117

[3] Bernard, E., Naveau, P., Vrac, M. and Mestre, O. (2013). Clustering of Maxima: Spatial Dependencies among Heavy Rainfall in France. *Journal of Climate* **26** 7929–7937.

[4] Burnham, K. P. and Anderson, D. R. (1998). *Model selection and inference: a practical information theoretic approach*. Springer, New York. MR1919620

[5] Butsch, L. and Fasen-Hartmann, V. (2025a). Supplementary Material A: Proofs of auxiliary results for "Information criteria for the number of directions of extremes in high-dimensional data". *Electron. J. Stat.*. https://doi.org/10.1214/25-EJS2469A.

[6] Butsch, L. and Fasen-Hartmann, V. (2025b). Supplementary Material B: R-codes for "Information criteria for the number of directions of extremes in high-dimensional data". *Electron. J. Stat.*. https://doi.org/10.1214/25-EJS2469B.

[7] Cavanaugh, J. E. and Neath, A. A. (1999). Generalizing the derivation of the Schwarz information criterion. *Comm. Statist. Theory Methods* **28** 49–66. MR1669504

[8] Chautru, E. (2015). Dimension reduction in multivariate extreme value analysis. *Electron. J. Stat.* **9** 383–418. MR3323204

[9] Claeskens, G. (2016). Statistical Model Choice. *Annu. Rev. Stat. Appl.* **3** 233–256.

[10] Clémençon, S., Huet, N. and Sabourin, A. (2024). Regular variation in Hilbert spaces and principal component analysis for functional extremes. *Stochastic Process. Appl.* **174** Paper No. 104375, 22. MR4745556

[11] Cooley, D. and Thibaud, E. (2019). Decompositions of dependence for high-dimensional extremes. *Biometrika* **106** 587–604. MR3992391

[12] Cover, T. M. (2006). *Elements of information theory*. Wiley-Interscience. MR2239987

[13] Drees, H. and Sabourin, A. (2021). Principal component analysis for multivariate extremes. *Electron. J. Stat.* **15** 908–943. MR4255291

[14] Duchi, J., Shalev-Shwartz, S., Singer, Y. and Chandra, T. (2008). Efficient projections onto the L_1 ball for learning in high dimensions. *Proceedings of the 25th International Conference on Machine Learning* 272–279. MR2177896

[15] Engelke, S. and Hitz, A. S. (2020). Graphical models for extremes. *J. R. Stat. Soc. Ser. B. Stat. Methodol.* **82** 871–932. With discussions. MR4136498

[16] Engelke, S. and Ivanovs, J. (2021). Sparse structures for multivariate extremes. *Annu. Rev. Stat. Appl.* **8** 241–270. MR4243547

[17] Engelke, S. and Volgushev, S. (2022). Structure learning for extremal tree models. *J. R. Stat. Soc. Ser. B. Stat. Methodol.* **84** 2055–2087. MR4515566

[18] Engelke, S., Hentschel, M., Lalancette, M. and Röttger, F. (2024). Graphical models for multivariate extremes. https://arxiv.org/abs/2402.02187.

[19] Falk, M. (2019). *Multivariate extreme value theory and D-norms*. *Springer Series in Operations Research and Financial Engineering*. Springer, Cham. MR3890054

[20] FOMICHOV, V. and IVANOVS, J. (2023). Spherical clustering in detection of groups of concomitant extremes. *Biometrika* **110** 135–153. MR4565448
[21] GISSIBL, N. and KLÜPPELBERG, C. (2018). Max-linear models on directed acyclic graphs. *Bernoulli* **24** 2693–2720. MR3779699
[22] GISSIBL, N., KLÜPPELBERG, C. and LAURITZEN, S. (2021). Identifiability and estimation of recursive max-linear models. *Scand. J. Stat.* **48** 188–211. MR4233170
[23] GOIX, N., SABOURIN, A. and CLÉMENÇON, S. (2017). Sparse representation of multivariate extremes with applications to anomaly detection. *J. Multivariate Anal.* **161** 12–31. MR3698112
[24] HASLETT, J. and RAFTERY, A. E. (1989). Space-Time Modelling with Long-Memory Dependence: Assessing Ireland's Wind Power Resource. *J. R. Stat. Soc. Ser. C. Appl. Stat.* **38** 1–50.
[25] JALALZAI, H. and LELUC, R. (2021). Feature Clustering for Support Identification in Extreme Regions. In *Proceedings of the 38th International Conference on Machine Learning* (M. MEILA and T. ZHANG, eds.). *Proceedings of Machine Learning Research* **139** 4733–4743. PMLR.
[26] JANSSEN, A. and WAN, P. (2020). k-means clustering of extremes. *Electron. J. Stat.* **14** 1211–1233. MR4071364
[27] MEYER, N. and WINTENBERGER, O. (2020). Tail inference for high-dimensional data. *arXiv:2007.11848v1*.
[28] MEYER, N. and WINTENBERGER, O. (2021). Sparse regular variation. *Adv. in Appl. Probab.* **53** 1115–1148. MR4342579
[29] MEYER, N. and WINTENBERGER, O. (2023). Multivariate Sparse Clustering for Extremes. *J. Amer. Statist. Assoc.* 1–12. MR4797911
[30] OUIMET, F. (2021). A precise local limit theorem for the multinomial distribution and some applications. *J. Statist. Plann. Inference* **215** 218–233. MR4249129
[31] POWERS, D. (2008). Evaluation: From Precision, Recall and F-Factor to ROC, Informedness, Markedness and Correlation. *Mach. Learn. Technol.* **2**.
[32] RESNICK, S. I. (1987). *Extreme Values, Regular Variation, and Point Processes*. Springer. MR0900810
[33] RESNICK, S. I. (2007). *Heavy-Tail Phenomena: Probabilistic and Statistical Modeling*. Springer. MR2271424
[34] ROHRBECK, C. and COOLEY, D. (2023). Simulating flood event sets using extremal principal components. *Ann. Appl. Stat.* **17** 1333–1352. MR4582715
[35] SCHWARZ, G. (1978). Estimating the dimension of a model. *Ann. Statist.* **6** 461–464. MR0468014
[36] SIMPSON, E. S., WADSWORTH, J. L. and TAWN, J. A. (2020). Determining the dependence structure of multivariate extremes. *Biometrika* **107** 513–532. MR4138974
[37] STATLIB - DATASETS ARCHIVE (2023). Wind. `http://lib.stat.cmu.edu/datasets/wind.data`, accessed October 26, 2023.