Karlsruhe Institute of Technology

Dimension Reduction of High-Dimensional Extremes using Information Criteria

Zur Erlangung des akademischen Grades eines

DOKTORS DER NATURWISSENSCHAFTEN

von der KIT-Fakultät für Mathematik des
Karlsruher Instituts für Technologie (KIT)
genehmigte

DISSERTATION

von

Lucas Butsch, M.Sc.

Tag der mündlichen Prüfung:	21. Januar 2026
Referentin:	Prof. Dr. Vicky Fasen-Hartmann
Korreferent:	JProf. Dr. Marco Oesting

To my grandfather.

Acknowledgments

First, I would like to thank Prof. Dr. Vicky Fasen-Hartmann. Over the past few years, I have been able to turn to her at any time, and she has always made time for me. Whenever I got stuck, she gave me the right advice at the right time. The many discussions, her support, and her commitment have contributed significantly to the creation of this work, which would not exist in this form otherwise. I would also like to thank her for her excellent and understanding supervision.

Furthermore, I would like to thank JProf. Dr. Marco Oesting for agreeing to act as a referee.

I would like to thank my current and former colleagues for the great atmosphere over the last few years and the conversations during coffee and lunch breaks, in the hallway and on walks. Many thanks to Florian, Franz, and Luca for proofreading.

My thanks also go to my friends and especially my parents, who have always supported me and stood behind me.

Lastly, I would like to thank my wife Kristina, who has accompanied me through all the ups and downs and enriches my life every day.

Prior Publications and further Declarations

Parts of all chapters in this thesis are direct quotes from articles published in peer-reviewed journals. The quotes appear throughout the thesis without being explicitly marked in order to improve readability. Specifically, Chapter 3 and most of Appendix A consist of

- Butsch, L. and Fasen-Hartmann, V. (2025). Information criteria for the number of directions of extremes in high-dimensional data. *Electron. J. Stat.* **19**(2): 5695-5740.
- Butsch, L. and Fasen-Hartmann, V. (2026). Statistical inference for extremal directions in high-dimensional spaces. Preprint available at: arxiv: 2603.26618.

and parts are also contained in

- Butsch, L. (2024, joint work with Fasen-Hartmann, V.). Information criteria for the number of directions of extremes in high-dimensional data. Oberwolfach Rep. 21, no. 3, pp. 2179–2250. Available at: Oberwolfach Rep. 21 (2024), no. 3.

Furthermore, Chapter 4 consists of

- Butsch, L. and Fasen-Hartmann, V. (2025). Estimation of the number of principal components in high-dimensional multivariate extremes. *Scand. J. Stat.* **52**(4); 2270-2312.

Most parts of Chapter 5 are also taken from these preprints. Some results from the preprints appear in other chapters of this thesis, as results have been merged and rearranged to ensure a consistent thesis.

This thesis was written in accordance with the KIT guidelines on the use of generative AI at KIT. By 1.d) of this guideline, the use of generative AI should be documented transparently whenever appropriate and necessary. The author considers all immediate influences of generative AI on this thesis to belong to this category.

In this context, the following AI models have been used: ChatGPT by OpenAI, DeepL by DeepL SE and Gemini by Google. Parts of the text were proofread and checked for typos and grammatical errors.

Contents

CHAPTER 1

INTRODUCTION

Extreme Value Theory (EVT) gained attention in the 20th century, initially within the field of civil engineering (Coles, 2001). Early applications focused on challenges such as determining the height of dams required to withstand a once-in-a-century flood. It was also applied to model other meteorological events, including extreme wind speeds, heat waves, wave heights, river flow rates and precipitation. More recently, EVT has been applied in other domains, including the financial industry, to analyze insurance losses, stock market shocks, and operational risk. A key obstacle in estimating a once-in-a-century event, however, is the uncertainty regarding its potential magnitude, as such an event may not have occurred during the observation period.

Therefore, a specialty of EVT is its focus on only the most extreme observations to make inferences about the far tail of a distribution. Hence, EVT can be viewed as an extrapolation technique where conclusions about unobserved extreme levels are drawn from a small number of extreme realizations. In many practical applications, such as those mentioned above, interest extends beyond univariate extremes, e.g., the height of a dam at a single location or the loss associated with a single asset, to multivariate settings involving multiple spatial locations or assets. For instance, to better understand the occurrence of extreme concentrations of the greenhouse gas ozone and the associated meteorological factors, an extreme value study was conducted in Russell et al. (2016) investigating the dependence structure of these factors on extreme ozone levels. The dependence structure of extremes was also analyzed in other fields, such as the study of high sea levels (Tawn, 1992), precipitation (Jiang et al., 2020) and in finance (Hilal et al., 2014). This necessitates the use of multivariate extreme value theory (MEVT), which accounts for the dependence structure among components of multivariate data, as such dependencies can critically influence the occurrence and frequency of joint extreme events.

A classical concept for modeling multivariate extremes is *multivariate regular variation* (Resnick, 1987, 2007; Falk, 2019). A d-dimensional random vector $\boldsymbol{X} \in \mathbb{R}^d$ is called *multivariate regularly varying* with index $\alpha > 0$ (tail index) if there exists 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 $S(\partial A) = 0$. The $\mathbb{S}^{d-1}$-valued random vector $\boldsymbol{\Theta}$ with distribution $\mathbb{P}(\boldsymbol{\Theta} \in \cdot) = S(\cdot)$ is called *spectral vector*. The *spectral measure*

S models the directional components and contains information about the dependence structure in the extremes of $\boldsymbol{X}$. Therefore, a primary goal is the estimation of S. Note that for the norm $\|\cdot\|$ we subsequently use either the L_1 norm, defined by $\|\boldsymbol{x}\|_1 \coloneqq \sum_{j=1}^{d} |x_j|$ for $\boldsymbol{x} = (x_1, \ldots, x_d) \in \mathbb{R}^d$, or the Euclidean ($L_2$) norm, defined by $\|\boldsymbol{x}\|_2 \coloneqq \sqrt{\sum_{j=1}^{d} x_j^2}$ for $\boldsymbol{x} = (x_1, \ldots, x_d) \in \mathbb{R}^d$. Furthermore, in what follows, the term dimension refers to the (ambient) dimension d of the space $\mathbb{R}^d$.

Although the estimation of the spectral measure S for bivariate data has been addressed in previous work, e.g. Einmahl et al. (2001); Einmahl and Segers (2009), the extension to arbitrary dimensions presents challenges, as mentioned in Clémençon et al. (2023).

Additionally, as data become more widely available in modern applications, both the sample size and the dimensionality of the observations may increase proportionally. The increasing of the dimension leads to the well-known curse of dimensionality (Bellman, 1957), which significantly complicates the estimation of the spectral measure S. In the setting of MEVT, this challenge is further amplified because inference is based only on extreme observations, which are by definition scarce.

However, in high-dimensional regimes, the spectral measure often exhibits sparsity and is concentrated on a lower-dimensional subspace. Consequently, a common strategy in multivariate statistics is to perform dimension reduction prior to estimation: first, one identifies the approximate support of S, and second, one estimates S restricted to this subspace. This approach can substantially improve both computational efficiency and estimation accuracy.

This leads to the main research question of the thesis:

(Q) *How can the dimension of extreme data be reliably reduced in high dimensions?*

This question touches on several topics. First, we are dealing with extreme data, for which, as noted, observations are scarce. Second, the task of "reducing the dimension" paves the way for estimating the low-dimensional support of the spectral measure. To accomplish this, we use information criteria, which are tools for model selection. The criteria we consider include the Akaike information criterion (AIC; Akaike, 1974) and the Bayesian information criterion (BIC; Schwarz, 1978). Furthermore, our objective for the dimension reduction is to estimate the extremal dependence structure reliably, which, in this framework, means that the information criteria are consistent. Finally, we assume that the dimension is large and may even tend to infinity. Consequently, we consider not only the classical large-sample setting (fixed dimension, sample size tending to infinity) but also the high-dimensional regime, where both the sample size and dimension diverge. In the latter case, the asymptotic behavior of the information criteria differs substantially from that in the fixed-dimensional setting.

The literature on dimension reduction methods for multivariate extremes using statistical learning methods has grown rapidly in recent years. For example, Chautru (2015) identified groups of jointly extreme variables by applying Principal Component Analysis (PCA)

followed by cluster analysis with spherical k-means to the spectral measure of a multivariate regularly varying random vector. Chiapino et al. (2019) used the joint tail dependence coefficient for this task. In Jalalzai and Leluc (2021), the empirical risk between the data and the output of a representation function evaluated on the data is minimized over extreme regions.

Another approach to reduce the dimension of an $\mathbb{R}_+^d$-valued random vector (i.e. a real-valued random vector with positive entries) involves the estimation of S using the L_1 norm. The support of S can be identified by the disjoint partition of the (positive) unit simplex $\mathbb{S}_+^{d-1} \coloneqq \{\boldsymbol{x} \in \mathbb{R}_+^d : \ \|\boldsymbol{x}\|_1 \coloneqq \sum_{j=1}^d |x_j| = 1\}$ with respect to the L_1 norm into sets of the form

$$C_\beta \coloneqq \{\boldsymbol{x} = (x_1, \ldots, x_d) \in \mathbb{S}_+^{d-1} \colon x_i > 0 \text{ for } i \in \beta, x_i = 0 \text{ for } i \notin \beta\} \subseteq \mathbb{S}_+^{d-1}, \ \beta \subset \{1, \ldots, d\}. \tag{1.2}$$

Knowing $S(C_\beta)$ for all $\beta \subseteq \{1, \ldots, d\}$ allows us to draw conclusions about the support of S and the directions of extreme events. The sets $\beta \in \mathcal{P}_d^*$, each corresponding to C_β, will henceforth be referred to as directions. If $S(C_\beta) > 0$, it implies that the components in the set β are jointly extreme and thus we have an extreme event in the direction β. The dimension is then reduced by separately considering the entries of $\boldsymbol{X}$ belonging to maximal directions, where a direction $\beta \in \mathcal{P}_d^*$ is maximal for $\boldsymbol{\Theta}$, if

$$\mathbb{P}(\boldsymbol{\Theta} \in C_\beta) > 0 \text{ and } \mathbb{P}(\boldsymbol{\Theta} \in C_{\beta'}) = 0 \ \text{ for all } \beta' \supsetneq \beta.$$

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. The underlying assumption is that the support of $\boldsymbol{\Theta}$ is that s^*, the number of directions $\beta \subset \{1, \ldots, d\}$ with $S(C_\beta) > 0$, is small, i.e. $s^* \ll 2^d - 1$.

However, complications arise for the estimation. On the one hand, for $\beta \subset \{1, \ldots, d\}$ with $|\beta| < d$ holds $C_\beta = \partial C_\beta$, which implies 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, empirically there are no observations in the set C_β. Therefore, the empirical estimator for $S(C_\beta)$ based on (1.1) is not consistent and is not useful. To avoid this problem, numerous approaches have been suggested, such as the support detection algorithm DAMEX (Detecting Anomalies among Multivariate EXtremes) of Goix et al. (2017) which works with truncated ε-cones to generate continuity sets that approximate the sets in (1.2). Simpson et al. (2020) use the concept of hidden regular variation on a collection of nonstandard subcones of $[0, \infty]^d \backslash \{0\}$.

An alternative solution to mitigate this problem is proposed in Meyer and Wintenberger (2021, 2023) by introducing the concept of sparse regular variation, which is equivalent to regular variation under some mild assumptions (see Section 3.1 for a definition). The main difference between regular variation and sparse regular variation is that the self-normalization $\boldsymbol{X}/\|\boldsymbol{X}\|_1$ in (1.1) is replaced by the Euclidean projection $\pi(\boldsymbol{X}/t)$ of $\boldsymbol{X}/t$

onto $\mathbb{S}_+^{d-1}$ for large $t > 0$. The Euclidean projection $\pi : \mathbb{R}_+^d \to \mathbb{S}_+^{d-1}$ is defined (and implemented as in Duchi et al., 2008) by

$$\pi(\boldsymbol{v}) = \underset{\boldsymbol{w} \in \mathbb{S}_+^{d-1}}{\arg\min} \|\boldsymbol{w} - \boldsymbol{v}\|_2^2. \tag{1.3}$$

The advantage of this projection is that the resulting vector $\pi(\boldsymbol{X}/t)$ is typically sparser, i.e. has more zero entries than $\boldsymbol{X}/\|\boldsymbol{X}\|_1$. This is particularly advantageous when only a few components are jointly extreme, as in a high-dimensional setting. However, the empirical estimator of Meyer and Wintenberger (2021) for the number of extreme directions, i.e. directions in which extreme events occur, in the sparse regularly varying model is biased and overestimates the true number of directions. To correct this, Meyer and Wintenberger (2021, 2023) developed an *Akaike Information Criterion* (AIC) consisting of two steps. The first step uses an AIC for *bias selection* to estimate the number of extreme directions. The second step extends this to an AIC for *threshold selection* to simultaneously estimate the threshold.

An alternative line of research applies *Principal Component Analysis* (PCA) to extremes. For an $\mathbb{R}^d$-valued regularly varying random vector $\boldsymbol{X}$, the dependence structure of the extremes is captured by the spectral vector $\boldsymbol{\Theta}$ and its covariance matrix is denoted as $\mathrm{Cov}(\boldsymbol{\Theta})$. Drees and Sabourin (2021) and Drees (2025) established a mathematical framework for PCA for the empirical covariance estimator of $\mathrm{Cov}(\boldsymbol{\Theta})$ by analyzing the squared reconstruction error, the excess risk and the asymptotic behavior.

Recently, Clémençon et al. (2024) extended the PCA approach to Hilbert-valued regularly varying random objects, whereas Avella-Medina et al. (2025) used kernel PCA, a nonlinear generalization of PCA. In addition, Cooley and Thibaud (2019) introduced decompositions to analyze tail dependence, summarized in the tail dependence matrix to which PCA was then applied. Furthermore, Rohrbeck and Cooley (2023) derived extreme principal components of the tail dependence matrix to model river flow in northern England and southern Scotland. This also enables them to generate synthetic data for the river flow. Wan (2026) analyzed the extremal dependence structure of random vectors projected on a $(d-1)$-dimensional hyperplane and subsequently applied PCA. Variants of k-means, an unsupervised clustering method, have also been applied to extreme observations. For example, Janßen and Wan (2020) analyzed spherical k-means for extreme observations and investigated it further for a max-linear model. In Fomichov and Ivanovs (2023) a different cost function was used, leading to k-principal-component clustering and Avella Medina et al. (2024) used a random k-nearest neighbor graph applied to a linear factor model to estimate the spectral measure. In Bernard et al. (2013) the Partitioning Around Medoids algorithm was proposed and compared to classical k-means based on french precipitation data.

Outside the extreme value framework, PCA has been analyzed in the high-dimensional setting. For example, the spiked covariance model, introduced by Johnstone (2001), is a

widely used model for PCA and similar applications in high-dimensional statistics. The model is used in Fujikoshi and Sakurai (2016) and Bai et al. (2018) for PCA in the fixed- and high-dimensional cases, respectively. In Bai and Yao (2012) convergence of the empirical eigenvalues is analyzed. Additionally, Jiang et al. (2023) applied the model to high-dimensional compositional data and Johnstone and Yang (2018) considered a spiked model to analyze the sample correlation matrix. Furthermore, the spiked covariance model has found applications in a variety of fields, including speech recognition, wireless communication, and statistical learning as mentioned in Paul (2007). The spiked covariance model was originally proposed for a Gaussian distribution with covariance matrix $\mathbf{\Sigma}$ whose eigenvalues $\lambda_1, \ldots, \lambda_d$ satisfy

$$\lambda_1 \geq \lambda_2 \geq \cdots \geq \lambda_{p^*} > \lambda_{p^*+1} = \cdots = \lambda_d =: \lambda > 0. \tag{1.4}$$

The idea behind the model is that we have p^* leading (or spiked) eigenvalues, i.e. eigenvalues which are strictly larger than λ. Hence, the Gaussian distribution concentrates in the directions of eigenvectors corresponding to these leading eigenvalues. Note that p^* differs from s^*. The motivation behind it is that s^* from the aforementioned approach is the number of relevant directions and lies between 1 and $2^d - 1$, whereas p^* is the number of leading eigenvalues, which lies between 1 and d. Generally, we assume that both s^* and p^* are unknown and relatively small compared to $2^d - 1$ and d, respectively.

For non-extreme data, information criteria to estimate p^* were investigated in Fujikoshi and Sakurai (2016) for Gaussian random vectors in a large-sample asymptotic framework and in Bai et al. (2018) for general data in the high-dimensional case. These criteria are motivated by a Gaussian likelihood function. Since in a Gaussian spiked covariance model the log-likelihood function can be written as a functional of the empirical eigenvalues both the AIC and BIC are defined as functionals of these empirical eigenvalues.

However, research on the number of significant principal components (the so-called *dimensionality*) in PCA for multivariate extremes has been limited. In Drees and Sabourin (2021), the dimension was estimated by examining empirical risk plots. An alternative approach is to analyze the scree plot, which is the plot of the empirical eigenvalues, in search of an "elbow" as a cutoff point, indicating a minimal variation in the empirical eigenvalues after this point. However, a big challenge in extreme value theory is the choice of the threshold t, which defines extreme observations as those whose norm exceeds t. Changing this threshold also changes the number of extreme observations and, consequently, the estimates of the empirical eigenvalues. Thus, a change of the threshold results in a different scree plot and possibly a different elbow. One of the few works that developed a statistical method for estimating the dimensionality is that of Drees (2025). The proposed method is based on asymptotic results for the reconstruction error of the projections and is a testing problem, but a disadvantage is the dependence on several tuning parameters.

Another approach to data clustering is proposed in the recent work of Chen et al. (2025), where data from multiple datasets are clustered into groups based on the value of the

tail index. In this setting, the dimension of the data is also allowed to approach ∞. Furthermore, the recent overview by Clémençon and Sabourin (2025) provides an in-depth look at the intersection of machine learning and extreme value theory.

A completely different line of research to represent the sparsity structure in multivariate models are graphical models as in, e.g., Engelke and Hitz (2020); Engelke and Volgushev (2022); Engelke et al. (2024); Gissibl et al. (2021); Gissibl and Klüppelberg (2018), to name only a few. A method for fitting the flexible Hüssler-Reiss distribution, popular for its use in graphical models, to high-dimensional data is presented in Lederer and Oesting (2024). A comprehensive overview of recent advances in probabilistic and statistical aspects of sparse structures in extremes is given in Engelke and Ivanovs (2021).

OUR CONTRIBUTIONS

In this thesis, we propose and analyze methods for identifying lower-dimensional structures in the support of the spectral measure. To address question **(Q)**, we consider two different techniques for dimension reduction. The first induces sparsity via the projection π in (1.3), which is motivated by a *Least Absolute Shrinkage and Selection Operator* (Lasso) approach. We call this approach the sparse regular variation (SRV) approach. The dimension reduction is then achieved by considering maximal directions. The second achieves a dimension reduction through Principal Component Analysis, which is based on a spiked covariance model similar to (1.4). We call this approach the Principal Component Analysis (PCA) approach. The two settings differ in the assumptions made about the support of S. In the first, the SRV approach, it is assumed that $\boldsymbol{X}$ is $\mathbb{R}^d_+$-valued, the L_1 norm is used and the support of S is sparse, meaning that only a few C_β have positive probability mass. In the second, the PCA approach, it is assumed that $\boldsymbol{X}$ is $\mathbb{R}^d$-valued, the L_2 norm is used and the support of S concentrates on a low-dimensional subspace of the unit sphere $\mathbb{S}^{d-1}$, with the remaining components regarded as noise. Hence, both settings rely on an underlying signal-noise model, and we aim to distinguish between signal and noise by employing information criteria to estimate the true signal.

In the following, we give an overview of our contributions in both settings and the connection to question **(Q)**.

In the first setting, we work with the partition given in (1.2) and want to find all directions β such that $S(C_\beta) > 0$. In the context of question **(Q)** the dimension is reduced by finding these directions with positive probability mass and hence maximal directions. Here we focus on deriving information criteria and analyze their consistency in the fixed-dimensional and the high-dimensional setting. The derivation is done for the fixed-dimensional setting, where we first derive a *Bayesian Information Criterion* (BIC), which goes back to Schwarz (1978). Besides this classical and widely-used information criterion, we also develop two novel information criteria. Firstly, we consider the *quasi-Akaike information criterion* (QAIC). Similarly to the AIC, the QAIC minimizes the Kullback-Leibler divergence between the

true distribution and the distribution of the model. However, we assume a Gaussian model, in contrast to the multinomial model used for the AIC. The idea behind this is motivated by the asymptotic normality of the random vector of counts of each direction (number of observations contained in C_β). Secondly, the *mean squared error information criterion* (MSEIC) is the other novel information criterion, where we approximate the weighted mean squared error (MSE) of the relative number of extreme observations and the true probabilities of the directions. We show that if the dimension d is fixed, the AIC for bias selection in Meyer and Wintenberger (2023) is not a weakly consistent information criterion, as opposed to the reliability required in question **(Q)**. The similarly derived MSEIC is also not consistent. In contrast, the BIC and the QAIC are consistent information criteria. Furthermore, we apply the information criteria to the high-dimensional case, where it is assumed that the dimension d depends on the sample size n such that $d = d_n \to \infty$ as $n \to \infty$. In this setting, we show that the AIC, BIC, QAIC and MSEIC are consistent. In addition, we also give information criteria to estimate the number of extremes k_n for fixed d.

In the second setting, we work with a spiked covariance model (1.4) as the covariance structure for $\mathbf{\Theta}$ and the dimension reduction of question **(Q)** is done by applying a PCA procedure to $\mathbf{\Theta}$. Hence, a low-dimensional representation of the data is achieved by estimating the number of relevant eigenvalues and the principal components. Motivated by the asymptotic normality of the empirical estimator of $\mathbf{\Theta}$, we adapt the AIC and BIC derived in Fujikoshi and Sakurai (2016) for the fixed-dimensional case and in Bai et al. (2018) for the high-dimensional case to use them with $\mathbf{\Theta}$ and estimate p^*. In the fixed-dimensional case, we use the asymptotic normality of the empirical eigenvalues to analyze the consistency of the information criteria. We show that the BIC is consistent and the AIC is not consistent. In the high-dimensional case, the self-normalization of $\mathbf{\Theta}$ induces dependencies that require new theoretical developments. By combining results from random matrix theory and compositional data analysis, we prove that the spectral distribution of the empirical eigenvalues converges. This allows us to show that both AIC and BIC are consistent information criteria. In addition, the information criteria are applicable when the dimension is larger than the number of extreme observations. To the best of our knowledge, this is the first approach to develop consistent information criteria for the dimensionality p^* of PCA in high-dimensional multivariate extremes. The only other comparable information criteria, those of Meyer and Wintenberger (2023) and from the SRV approach, use the concept of sparse regular variation to construct sparsity in the data, which is a different assumption from that underlying this PCA approach.

In summary, we recap the key contributions of our work, emphasizing its significance and potential impact.

- In response to question **(Q)**, we consider two distinct approaches to reduce the dimension. The advantage is that the assumed structure of the support of $\mathbf{\Theta}$ is different and therefore a wide variety of dependence structures is covered.

- In both settings, we introduce information criteria which are consistent and allow us to estimate the extremal dependence structure. Since the properties of the information criteria differ, it is possible to choose one that suits specific needs, e.g., preferring a larger model over one that is too small, or vice versa.
- Among the information criteria considered are the novel QAIC and MSEIC. The underlying idea of these criteria can be extended to other scenarios in which asymptotic normality holds, thereby offering substantial potential for applications beyond the present setting.
- We obtain supplementary results in both the fixed- and high-dimensional regimes, contributing new theory to MEVT and to random matrix theory.
- Finally, we validate our methods with a simulation study and apply them to real-world examples. The real-world examples include a low-dimensional case, where wind speed data is analyzed, and a high-dimensional case, where the data consists of precipitation measurements and the number of extremes is smaller than the dimension of the space.

Outline of the thesis

This thesis is structured as follows. In Chapter 2 we introduce the main concepts on which the subsequent results are based. We start in Section 2.1 with univariate and multivariate regular variation. As already stated above, multivariate regular variation is a key framework in multivariate EVT and allows us to analyze the dependence structure. In addition, we introduce the Hill estimator for the extremal index and describe how the marginal distribution of data can be transformed, and we also discuss the notions of asymptotic (full) dependence and asymptotic independence. In Section 2.2 we move to the high-dimensional setting and present the approaches to reduce the dimension and induce sparsity, based on SRV and PCA. Then we give an introduction to the model selection via information criteria in Section 2.3. The most widely used information criteria, the AIC and the BIC, are introduced and the mean squared error is presented. In Chapter 3 we begin with the introduction of the so-called relevant directions and bias directions in Section 3.1. The underlying model is introduced as well. Then, in Section 3.2 the first information criteria are presented. In Section 3.2.1 we define the QAIC and show further that the QAIC is consistent in contrast to the AIC. The information criterion based on the mean squared error, which is derived similarly to the AIC and is also inconsistent, is analyzed in Section 3.2.2. The final information criteria in this chapter, the BIC and its consistency, are treated in Section 3.2.3. The high-dimensional setting is treated in Section 3.3, where we first define bias directions for this setting in Section 3.3.1 and then analyze the numerical behavior of these directions in Section 3.3.2. In Section 3.3.3 the information criteria are adapted to the high-dimensional setting and the consistency is analyzed. Moving on to Chapter 4, the PCA approach is given and the underlying

model is detailed. Further, Section 4.1 lays the foundation for analyzing the behavior of empirical eigenvalues of a covariance matrix for fixed dimension (Section 4.1.1) and in high dimensions (Section 4.1.2), where we introduce the Marčenko-Pastur law and draw the connection to the empirical eigenvalues. The asymptotic behavior of the eigenvalues for the fixed-dimensional and high-dimensional case for extremes is derived in Section 4.2. For both cases, we introduce the AIC and BIC in Section 4.3 and Section 4.4. Along the way, we also analyze the consistency properties of the information criteria. Following up on the theoretical results, the simulation study is presented in Chapter 5. In Section 5.1 we present the results for the SRV approach and in Section 5.2 the results for the PCA approach. We apply each method to a real-world dataset in Section 5.3: wind speed measurements from the Republic of Ireland, representing the fixed-dimensional setting for the first approach, and precipitation measurements from Germany, representing the high-dimensional setting. Finally, in Chapter 6 we come to a conclusion and refer to question **(Q)**. Furthermore, we reflect on the new research questions which have arisen. In Appendix A auxiliary results are stated.

CHAPTER 2

FUNDAMENTALS

This chapter presents the fundamental concepts on which this thesis is built. We start in Section 2.1 with the introduction of extreme value theory and regular variation. Then, moving to the high-dimensional setting, in Section 2.2 the procedures used to induce sparsity or reduce the dimension, given by SRV and PCA, respectively, are presented. Finally, in Section 2.3, we introduce the information criteria used for model selection.

2.1. Extreme Value Theory

First, we introduce univariate regular variation in Section 2.1.1, before we present multivariate regular variation and the polar decomposition in Section 2.1.2. In the last part, Section 2.1.3, the Hill estimator and a method to transform the marginal distribution as a data preprocessing step are given.

2.1.1. Univariate regular variation

Regular variation is a concept for modeling the behavior of functions, where regularly varying functions behave similarly to power functions. We start with the introduction of a regularly varying function before we proceed with regularly varying random variables.

Definition 2.1. A function $f : \mathbb{R}_+ \to \mathbb{R}_+$ is regularly varying (at ∞) with index $\alpha \in \mathbb{R}$, when for $x > 0$

$$\lim_{t\to\infty} \frac{f(xt)}{f(t)} = x^\alpha.$$

For $\alpha = 0$ the function is called slowly varying. Examples of slowly varying functions are $\log(1 + x)$ and $\log(\log(e + x))$. Furthermore, due to the characterization theorem (cf. Bingham et al., 1989, Theorem 1.4.1), a regularly varying function f with index α can be represented as $f(x) = L(x)x^\alpha$, where L is a slowly varying function.

Next, we consider regular variation of random variables which is defined using the cumulative distribution function.

Definition 2.2. Let X be an $\mathbb{R}_+$-valued random variable with cumulative distribution function F. Then X is called regularly varying with (tail) index α if $\overline{F} \coloneqq 1 - F$ is regularly varying with index $-\alpha$. In this case, we write $X \in \mathrm{RV}(\alpha)$.

Remark 2.3.

(a) A random variable is regularly varying if and only if the distribution function F is in the domain of attraction of the Fréchet distribution (Embrechts et al., 1997, Theorem 3.3.7).

(b) Univariate regular variation is introduced for $\mathbb{R}_+$-valued random variables because the theory of multivariate regular variation for a vector $\boldsymbol{X} \in \mathbb{R}^d$ is based on the regular variation of its norm, $\|\boldsymbol{X}\|$, which is an $\mathbb{R}_+$-valued random variable. Univariate regular variation can also be defined for $\mathbb{R}$-valued random variables.

(c) If $X \in \mathrm{RV}(\alpha)$ with $\alpha \geq 0$, then (Kulik and Soulier, 2020, Proposition 1.4.6)

$$\mathbb{E}[|X|^\beta] < \infty \quad \text{for } \beta < \alpha$$

and

$$\mathbb{E}[|X|^\beta] = \infty \quad \text{for } \beta > \alpha,$$

which means that the moments only exist up to the order of the index α.

There are equivalent conditions for regular variation of positive random variables, which we present in the next theorem. A sequence of measures (μ_n) on the Borel-σ-field of $(0, \infty]$ converges vaguely to μ on $(0, \infty]$, denoted by $\mu_n \xrightarrow{v} \mu$, if

$$\int_{(0,\infty]} f \, \mathrm{d}\mu_n \to \int_{(0,\infty]} f \, \mathrm{d}\mu$$

as $n \to \infty$ for every non-negative continuous function f with compact support. In the following $M_+(0, \infty]$ is the space of all non-negative Radon measures on the σ-field of $(0, \infty]$, i.e. the set of all non-negative measures which are finite on compact subsets of $(0, \infty]$.

Theorem 2.4 (Theorem 3.6, Resnick, 2007)**.** *Let X be an $\mathbb{R}_+$-valued random variable with cumulative distribution function F and $\overline{F} = 1 - F$. Then, the following statements are equivalent.*

(a) $\overline{F}$ is regularly varying with index $-\alpha$ for $\alpha > 0$.

(b) There exists a sequence a_n with $a_n \to \infty$ as $n \to \infty$ such that

$$\lim_{n\to\infty} n\overline{F}(a_n x) = x^{-\alpha}, \quad x > 0.$$

(c) There exists a sequence a_n with $a_n \to \infty$ as $n \to \infty$ such that

$$\mu_n(\cdot) := n\mathbb{P}\left(\frac{X}{a_n} \in \cdot\right) \xrightarrow{v} \mu_\alpha(\cdot),$$

in $M_+(0, \infty]$ and $\mu_\alpha\big((x, \infty]\big) = x^{-\alpha}$.

Remark 2.5. If one of the conditions is fulfilled, one may choose (cf. Resnick, 2007, Remark 3.3)

$$a(t) = F^{\leftarrow}\left(1 - \frac{1}{t}\right)$$

and $a_n = a(n)$, where $F^{\leftarrow}(x) \coloneqq \inf\{z \in \mathbb{R} : F(z) \geq x\}$ is the generalized inverse of F.

2.1.2. Multivariate regular variation

Until now, we only considered univariate random variables. However, the concept of regular variation can also be transferred to the multivariate case, as introduced in (1.1). For the definition of multivariate regular variation we follow Chapter 3 in Mikosch and Wintenberger (2024), where it is not required that every component is regularly varying; rather, the norm of the vector is regularly varying.

Definition 2.6. Let $\boldsymbol{X}$ be an $\mathbb{R}^d$-valued random vector, where $\|\boldsymbol{X}\|$ is regularly varying with index $\alpha > 0$ such that there exists a sequence $a_n > 0$ with $n\mathbb{P}(\|\boldsymbol{X}\| > a_n) \to 1$ as $n \to \infty$. Then, $\boldsymbol{X}$ is called regularly varying if a non-null Radon measure μ on the Borel σ-field of $\mathbb{R}^d_{\mathbf{0}} \coloneqq \mathbb{R}^d \setminus \{\mathbf{0}\}$ exists, such that as $n \to \infty$

$$\mu_n(A) \coloneqq n\mathbb{P}\left(\frac{\boldsymbol{X}}{a_n} \in A\right) \to \mu(A)$$

for every μ-continuity set A, which means that μ_n converges vaguely to μ (cf. Portmanteau theorem in Resnick, 2007, Theorem 3.2).

Further, it is possible to define regular variation in spherical coordinates, which is convenient when a separation between the radial component and the directional component of a regularly varying random vector is desired.

Proposition 2.7 (Proposition 3.2.20, Mikosch and Wintenberger, 2024)**.** *Let $\boldsymbol{X}$ be an $\mathbb{R}^d$-valued random vector. Then $\boldsymbol{X}$ is regularly varying with index $\alpha > 0$ and non-null Radon measure μ on $\mathbb{R}^d_{\mathbf{0}}$ if and only if one of the following conditions holds.*

(a) *(i) $\|\boldsymbol{X}\|$ is regularly varying with (tail) index $\alpha > 0$.*

(ii) As $x \to \infty$ the following vague convergence holds

$$\frac{\mathbb{P}(\boldsymbol{X}/x \in \cdot)}{\mathbb{P}(\|\boldsymbol{X}\| > x)} \xrightarrow{v} \mu(\cdot).$$

(b) *(i) $\|\boldsymbol{X}\|$ is regularly varying with (tail) index $\alpha > 0$.*

(ii) There exists an $\mathbb{S}^{d-1}$-valued random vector $\boldsymbol{\Theta}$ such that as $x \to \infty$,

$$\mathbb{P}\left(\frac{\boldsymbol{X}}{\|\boldsymbol{X}\|} \in \cdot \,\middle|\, \|\boldsymbol{X}\| > x\right) \xrightarrow{w} \mathbb{P}(\boldsymbol{\Theta} \in \cdot).$$

(c) The weak convergence

$$\mathbb{P}\left(\left(\frac{\|\boldsymbol{X}\|}{x}, \frac{\boldsymbol{X}}{\|\boldsymbol{X}\|}\right) \in \cdot \,\middle|\, \|\boldsymbol{X}\| > x\right) \xrightarrow{w} \mathbb{P}((Y, \boldsymbol{\Theta}) \in \cdot)$$

holds as $x \to \infty$, where $Y \sim \text{Pareto}(\alpha)$ is independent of the $\mathbb{S}^{d-1}$-valued random vector $\boldsymbol{\Theta}$.

(d) The weak convergence

$$\mathbb{P}\left(\frac{\boldsymbol{X}}{x} \in \cdot \,\middle|\, \|\boldsymbol{X}\| > x\right) \xrightarrow{w} \mathbb{P}(Y\boldsymbol{\Theta} \in \cdot)$$

holds as $x \to \infty$, where $Y \sim \text{Pareto}(\alpha)$ is independent of the $\mathbb{S}^{d-1}$-valued random vector $\boldsymbol{\Theta}$.

We write $\boldsymbol{X} \in RV(\alpha, \boldsymbol{\Theta})$.

The random vector $\boldsymbol{\Theta}$ (or the distribution of $\boldsymbol{\Theta}$) is called spectral vector (or spectral measure) of $\boldsymbol{X}$. In later chapters, we work with different norms for $\boldsymbol{\Theta}$. Recall that we mainly use the L_1 norm, given by $\|\boldsymbol{x}\|_1 = \sum_{j=1}^d |x_j|$ for $\boldsymbol{x} = (x_1, \ldots, x_d) \in \mathbb{R}^d$, and the Euclidean norm, given by $\|\boldsymbol{x}\|_2 = \sqrt{\sum_{j=1}^d x_j^2}$ for $\boldsymbol{x} = (x_1, \ldots, x_d) \in \mathbb{R}^d$. The distribution of $\boldsymbol{\Theta}$ depends on the choice of the norm. Nevertheless, it is possible to transform the distribution of the spectral vector from one norm to another. For the following result, we index $\boldsymbol{\Theta}$ by the norm used.

Proposition 2.8 (Proposition 3.2.22, Mikosch and Wintenberger, 2024)**.** *Suppose $\boldsymbol{X} \in RV(\alpha, \boldsymbol{\Theta}_{\|\cdot\|})$, where $\boldsymbol{\Theta}_{\|\cdot\|} \in \mathbb{S}_{\|\cdot\|}^{d-1} = \{\boldsymbol{x} \in \mathbb{R}^d : \|\boldsymbol{x}\| = 1\}$. For another norm $\|\cdot\|_0$, the spectral measure $\boldsymbol{\Theta}_{\|\cdot\|_0}$ of $\boldsymbol{X}$ on $\mathbb{S}_{\|\cdot\|_0}^{d-1} = \{\boldsymbol{x} \in \mathbb{R}^d : \|\boldsymbol{x}\|_0 = 1\}$ is given by*

$$\mathbb{P}(\boldsymbol{\Theta}_{\|\cdot\|_0} \in \cdot) = \mathbb{E}\left[\frac{\|\boldsymbol{\Theta}_{\|\cdot\|}\|_0^\alpha}{\mathbb{E}[\|\boldsymbol{\Theta}_{\|\cdot\|}\|_0^\alpha]} \mathbb{1}\left\{\frac{\boldsymbol{\Theta}_{\|\cdot\|}}{\|\boldsymbol{\Theta}_{\|\cdot\|}\|_0} \in \cdot\right\}\right].$$

In the following, we omit the indexation with the norm, as the choice will be clear from the context.

The spectral vector can be seen as the direction of $\boldsymbol{X}$ given that $\|\boldsymbol{X}\|$ is sufficiently large. Hence, in $\boldsymbol{\Theta}$ the extremal dependence structure of $\boldsymbol{X}$ is decoded and various choices for $\boldsymbol{\Theta}$ are possible. Next, we present asymptotic independence and asymptotic full dependence as possible dependence structures of a regularly varying random vector $\boldsymbol{X}$ and its spectral vector. While multiple notions of asymptotic dependence (cf. Das and Fasen-Hartmann, 2025) exist, we adopt the definitions from Resnick (2007).

Definition 2.9 (Asymptotic independence)**.** An $\mathbb{R}_+^d$-valued, regularly varying random vector $\boldsymbol{X}$ with index α possesses asymptotic independence if the limit measure μ from

Definition 2.6 satisfies

$$\mu(\,\mathrm{d}x_1, \ldots, \mathrm{d}x_d) = \sum_{j=1}^{d} \delta_0(\,\mathrm{d}x_1) \otimes \cdots \otimes \delta_0(\,\mathrm{d}x_{j-1}) \otimes \mu_\alpha(\,\mathrm{d}x_j) \otimes \delta_0(\,\mathrm{d}x_{j+1}) \otimes \cdots \otimes \delta_0(\,\mathrm{d}x_d),$$

where δ_0 is the Dirac measure in 0 and μ_α is defined in Theorem 2.4. Hence, for $\boldsymbol{x} = (x_1, \ldots, x_d) > 0$ we have

$$\mu([\mathbf{0}, \boldsymbol{x}]^c) = \sum_{j=1}^{d} x_j^{-\alpha}.$$

Hence, for a random vector $\boldsymbol{X}$ exhibiting asymptotic independence, the mass of the limit measure μ is concentrated on the axis (cf. Resnick, 2007) and the spectral vector $\boldsymbol{\Theta}$ takes only the standard unit vectors as values $\mathbb{P}$-a.s.

Definition 2.10 (Asymptotic full dependence)**.** An $\mathbb{R}_+^d$-valued, regularly varying random vector $\boldsymbol{X}$ with index α possesses asymptotic full dependence if the limit measure μ from Definition 2.6 satisfies for $\boldsymbol{x} > 0$

$$\mu([\mathbf{0}, \boldsymbol{x}]^c) = \Big(\min_{j=1,\ldots,d} x_j\Big)^{-\alpha}.$$

If $\boldsymbol{X}$ exhibits asymptotic full dependence, then the components of $\boldsymbol{X}$ are strongly dependent in their joint upper tail.

2.1.3. Statistical Methods

When working with real-world data, it can generally not be assumed that the marginal distributions are identical or have 1 as tail index. For example, in Section 5.3.2 we consider precipitation data, where the magnitude of an extreme realization depends on the geographic location and a value that is extreme at one station may be a non-extreme observation at another. However, we introduce in this section a method for estimating the tail index α and a method to transform the marginal distribution.

Hill estimator

For a regularly varying random variable X with index α, one of the most widely used estimators of α is the Hill estimator (Hill, 1975). Note that the Hill estimator estimates $1/\alpha$ and not α directly. Let $X_1, \ldots, X_n$ be an i.i.d. sample from the distribution of X, and denote by $X_{(1,n)} \geq \cdots \geq X_{(n,n)}$ the corresponding order statistics. Then for $k = 1, \ldots, n-1$ the Hill estimator is defined by

$$H_{k,n} \coloneqq \frac{1}{k} \sum_{i=1}^{k} \log \frac{X_{(i,n)}}{X_{(k+1,n)}}.$$

The Hill estimator is a consistent estimator, as shown in the following theorem.

Theorem 2.11 (Mason, 1982)**.** *Suppose that $X_1, X_2, \ldots$ is an i.i.d. sequence of regularly varying random variables with index $\alpha > 0$ and (k_n) is a sequence in $\mathbb{N}$ such that $k_n \to \infty$ and $k_n/n \to 0$ as $n \to \infty$. Then follows that*

$$H_{k_n,n} \xrightarrow{\mathbb{P}} \alpha^{-1}.$$

Note that if X is regularly varying with index α, then it follows by direct calculation that X^α is regularly varying with index 1.

MARGINAL TRANSFORMATION

To analyze extremal dependence independently of the marginal distributions, it is standard practice to transform each marginal distribution to a common, heavy-tailed distribution, such as the unit Fréchet distribution. This is achieved using a transformation similar to the probability integral transformation. In the following, we present the procedure for a univariate sample; this method can also be applied to each marginal of a multivariate sample.

Suppose that X is a random variable with continuous distribution function F and generalized inverse $F^{\leftarrow}$. Then for $x \in (0,1)$ holds by the definition of the generalized inverse that

$$\mathbb{P}(F(X) \leq x) = \mathbb{P}(X \leq F^{\leftarrow}(x)) = F(F^{\leftarrow}(x)) = x.$$

Hence, $F(X)$ is uniformly distributed on $(0,1)$. On the other hand, if G is a distribution function with generalized inverse $G^{\leftarrow}$ and $U \sim \mathcal{U}(0,1)$, then $G^{\leftarrow}(U)$ follows the distribution of G. Therefore, we can transform X to the distribution G via the transformation $G^{\leftarrow}(F(X))$.

In practice, F is unknown and is estimated by the empirical distribution function, which allows us to transform a data sample $X_1, \ldots, X_n$ to the unit Fréchet distribution with distribution function $G(y) = \exp(-y^{-1})$ for $y > 0$. The generalized inverse is $G^{\leftarrow}(u) = -(\log u)^{-1}$. Then, the transformed sample $Y_1, \ldots, Y_n$ is calculated as

$$Y_i := -\Big(\log\Big(\frac{1}{n+1}\sum_{j=1}^{n} \mathbb{1}\{X_j \leq X_i\}\Big)\Big)^{-1}.$$

The scaling factor is chosen as $n+1$ instead of n to avoid taking the logarithm of 1 for the largest observation, which would result in a division by zero. Note that

$$\frac{1}{n+1}\sum_{j=1}^{n} \mathbb{1}\{X_j \leq X_i\} = \frac{R_i}{n+1},$$

where R_i is the rank of X_i in the sample $X_1, \ldots, X_n$. This allows us to calculate Y_i using

the ranks of $X_1, \dots, X_n$ by

$$Y_i = -\Big(\log\Big(\frac{R_i}{n+1}\Big)\Big)^{-1}.$$

2.2. Sparsity and Dimension reduction for extremes

As mentioned in Chapter 1, the estimation of the support of the spectral measure S is challenging when the dimension is high. In this section, we will present the underlying framework of the SRV and PCA approach from Chapter 1 in more detail. First, we introduce the concept of the Least Absolute Shrinkage and Selection Operator (Lasso) and sparse regular variation in Section 2.2.1. Then, in Section 2.2.2 we present the approach based on Principal Component Analysis (PCA).

2.2.1. Sparse regular variation

In this section, we introduce the concept of sparse regular variation which goes back to Meyer and Wintenberger (2021, 2023). The main idea is to use the Euclidean projection instead of the self-normalization for regular variation to mitigate the issue mentioned in Chapter 1 with the sets C_β using $\boldsymbol{\Theta}$. The idea of this method is based on the Least Absolute Shrinkage and Selection Operator (Lasso). Lasso originated in the context of regression models to set coefficients to zero, where it is a regularization method that restricts the L_1 norm of the coefficients. It was first introduced by Tibshirani (1996) and is now a widely used method. The regularization with the L_2 norm, called Ridge regression, does not have this property (see Hastie et al., 2015). The same idea can be applied to project a vector onto the L_1 simplex instead of using the self-normalization $\boldsymbol{X}/\|\boldsymbol{X}\|_1$ in Proposition 2.7 (cf. Meyer and Wintenberger, 2021) and thus setting some components to zero, i.e. to induce sparsity.

The motivation for this approach arises when one analyzes the support of the spectral vector $\boldsymbol{\Theta}$ as given in (1.1) to determine directions in which extreme events occur. For this, the simplex $\mathbb{S}_+^{d-1}$ is partitioned into the sets C_β as defined in (1.2) and each partition is separately addressed. Note that we assume that $\boldsymbol{X}$ is an $\mathbb{R}_+^d$-valued regularly varying random variable throughout this section and the L_1 norm is used.

For the analysis of the dependence structure of the tail of $\boldsymbol{X}$, we recall the notion of a maximal direction for $\boldsymbol{\Theta}$ from Chapter 1.

Definition 2.12 (Maximal direction for $\boldsymbol{\Theta}$)**.** A direction $\beta \in \mathcal{P}_d^*$ is maximal for $\boldsymbol{\Theta}$, if

$$\mathbb{P}(\boldsymbol{\Theta} \in C_\beta) > 0 \text{ and } \mathbb{P}(\boldsymbol{\Theta} \in C_{\beta'}) = 0 \text{ for all } \beta' \supsetneq \beta.$$

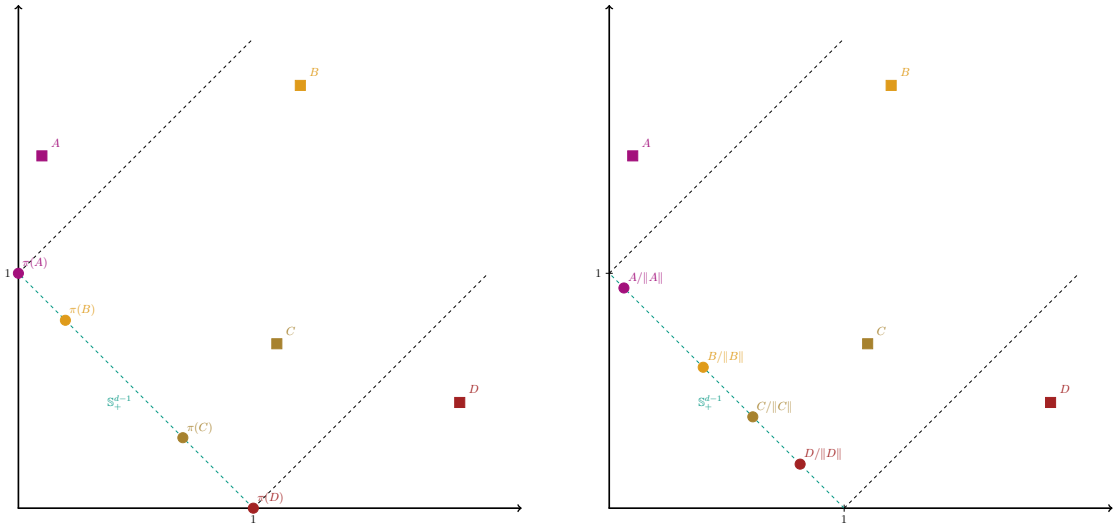
However, as mentioned in Chapter 1 there arises a problem, when dealing with these maximal directions for $\boldsymbol{\Theta}$. The sets C_β for $\beta \neq \{1, \dots d\}$ are lower dimensional sets in $\mathbb{R}^d$. As a consequence, when $\mathbb{P}(\boldsymbol{\Theta} \in C_\beta) > 0$ holds, then C_β is not a continuity set of the

spectral measure.

To mitigate this issue, in Meyer and Wintenberger (2021) the self-normalization of $\boldsymbol{X}$, i.e. $\boldsymbol{X}/\|\boldsymbol{X}\|_1$, was replaced by the Euclidean projection $\pi(\boldsymbol{X}/t)$ of $\boldsymbol{X}/t$ onto $\mathbb{S}_+^{d-1}$. The Euclidean projection π_z onto the L_1 simplex (cf. Duchi et al., 2008) is for a vector $\boldsymbol{v} \in \mathbb{R}_+^d$ and a $z > 0$ defined by

$$\begin{aligned} \pi_z : \mathbb{R}_+^d \to \mathbb{S}_+^{d-1}(z) &:= \{\boldsymbol{x} \in \mathbb{R}_+^d : \|\boldsymbol{x}\|_1 = z\} \\ \boldsymbol{v} \mapsto \boldsymbol{w} = \underset{\boldsymbol{w} \in \mathbb{S}_+^{d-1}(z)}{\arg\min} &\|\boldsymbol{w} - \boldsymbol{v}\|_2^2, \end{aligned}$$

where $\boldsymbol{w} \in \mathbb{S}_+^{d-1}(z)$. For $z = 1$ we write $\pi := \pi_1$ as in (1.3).



(a) Euclidean projection

(b) Self-normalization

Figure 2.1.: *Projection of four points onto the simplex* $\mathbb{S}_+^{d-1}$ *by using the Euclidean projection on the left hand side and self-normalization on the right hand side.*

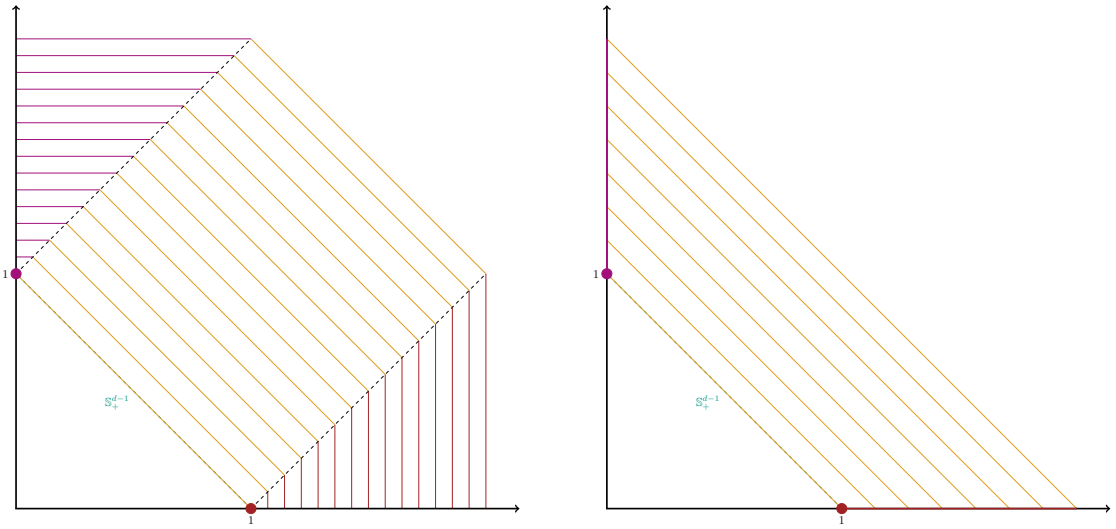
In Figure 2.1 we see the difference between the Euclidean projection and the self-normalization. The number of zero components of the points B and C and their projections does not change, since both projections lie in the interior of the simplex. However, the number of zero components of the points A and D and their projections differ, since the Euclidean projection sets some components to zero. When we evaluate the preimage of the subsimplex in Figure 2.2 this becomes even more evident.

This lets us now define sparse regular variation (see Meyer and Wintenberger, 2021).

Definition 2.13. 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

$$\mathbb{P}\left(\frac{\|\boldsymbol{X}\|_1}{t} > r, \pi\left(\frac{\boldsymbol{X}}{t}\right) \in A \,\middle|\, \|\boldsymbol{X}\|_1 > t\right) \to \mathbb{P}(R > r, \boldsymbol{Z} \in A), \quad t \to \infty,$$

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



(a) Euclidean projection (b) Self-normalization

Figure 2.2.: *Preimages of the Euclidean projection on the left hand side and self-normalization on the right hand side.*

Remark 2.14. (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 $\boldsymbol{X} \in \mathrm{SRV}(\alpha, \boldsymbol{Z})$. However, R and $\boldsymbol{Z}$ are not independent.

(b) The concept of sparse regular variation introduced by Meyer and Wintenberger (2021) 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 setting, we also restrict our analysis to random vectors in the positive orthant.

(c) 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 (2021, Proposition 2) we know that

$$\mathbb{P}(\pi(\boldsymbol{X}/t) \in C_\beta \,|\, \|\boldsymbol{X}\|_1 > t) \longrightarrow \mathbb{P}(\boldsymbol{Z} \in C_\beta), \quad \text{as } t \to \infty,$$

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

As shown in Meyer and Wintenberger (2021), sparse regular variation and regular variation are under suitable conditions equivalent. For the equivalence of them, the following sets and functions are introduced. Let

$$\begin{aligned}
A_{\boldsymbol{x}} &:= \{\boldsymbol{u} \in \mathbb{S}_+^{d-1} : \boldsymbol{u} \geq \boldsymbol{x}\}, \text{ for } \boldsymbol{x} \in B_+^d(0,1) := \{\boldsymbol{x} \in \mathbb{R}_+^d : \|\boldsymbol{x}\|_1 \leq 1\}, \\
\chi_\beta &:= \{\boldsymbol{x} \in B_+^d(0,1) : \boldsymbol{x}_\beta > \boldsymbol{0}_\beta, \boldsymbol{x}_{\beta^c} = \boldsymbol{0}_{\beta^c}\}, \text{ for } \beta \in \mathcal{P}_d^*, \\
\chi_\beta^0 &:= \{\boldsymbol{x} \in B_+^d(0,1) : \boldsymbol{x}_{\beta^c} = \boldsymbol{0}_{\beta^c}\}, \text{ for } \beta \in \mathcal{P}_d^*.
\end{aligned}$$

By λ_β we denote the Lebesgue measure on χ_β. For $\beta, \gamma \in \mathcal{P}_d^*$ with $\gamma \supset \beta$ define

$$G_\beta(\boldsymbol{x}) \coloneqq \mathbb{P}(\boldsymbol{Z}_\beta > \boldsymbol{x}_\beta, \boldsymbol{Z}_{\beta^c} \leq x_{\beta^c}), \text{ for } \boldsymbol{x}_\beta \in B_+^{|\beta|}(0,1), \boldsymbol{x}_{\beta^c} \in B_+^{|\beta^c|}(0,1)$$

and

$$H_{\beta,\gamma}(\boldsymbol{u}, v, w) = \mathbb{P}(\phi_\gamma(\boldsymbol{Z})_\beta \geq \boldsymbol{u}, \min_{j \in \gamma \setminus \beta} \phi_\gamma(\boldsymbol{Z})_j > v, \max_{j \in \gamma^c} \phi_\gamma(\boldsymbol{Z})_j \leq w)$$

for $\boldsymbol{u} \in B_+^{|\beta|}(0,1), v, w \in [0,1]$, where $\phi_\gamma : \mathbb{S}_+^{d-1} \to \mathbb{R}_+^d, \boldsymbol{u} \mapsto \phi_\gamma(\boldsymbol{u})$ is defined by

$$\phi_\gamma(\boldsymbol{u})_j = \begin{cases} \boldsymbol{u}_j + \frac{\|\boldsymbol{u}_{\gamma^c}\|_1}{|\gamma|}, & \text{for } j \in \gamma, \\ \boldsymbol{u}_j + \frac{\|\boldsymbol{u}_{\gamma^c \setminus \{j\}}\|_1}{|\gamma|+1}, & \text{for } j \in \gamma^c. \end{cases}$$

Additionally, we need to following assumption on $H_{\beta,\gamma}$.

Assumption RV. For all $\beta, \gamma \in \mathcal{P}_d^*$ with $\gamma \supset \beta$ and for λ_β-a.s. all $x \in \chi_\beta$, the function $H_{\beta,\gamma}$ is continuously differentiable at $(\boldsymbol{x}_\beta, 0, 0)$ with value $\mathrm{d}H_{\beta,\gamma}(\boldsymbol{x}_\beta, 0, 0)$.

The equivalence of regular variation and sparse regular variation results from the next Theorem Theorem 1, (Meyer and Wintenberger, 2021, Theorem 1).

Theorem 2.15 (Theorem 1, Meyer and Wintenberger, , 2021)**.**

(a) If $\boldsymbol{X} \in RV(\alpha, \boldsymbol{\Theta})$, then $\boldsymbol{X} \in SRV(\alpha, \boldsymbol{Z})$ with $\boldsymbol{Z} = \pi(Y\boldsymbol{\Theta})$ and

$$G_\beta(\boldsymbol{x}) = \mathbb{E}\Bigg[\Bigg(1 \wedge \min_{j \in \beta_+} \left(\frac{|\beta|\boldsymbol{\Theta}_j - |\boldsymbol{\Theta}_\beta|}{|\beta|\boldsymbol{x}_j - 1}\right)_+^\alpha \wedge \min_{j \in \beta^c} \left(|\boldsymbol{\Theta}_\beta| - |\beta|\boldsymbol{\Theta}_j\right)_+^\alpha - \max_{j \in \beta_-} \left(\frac{|\beta|\boldsymbol{\Theta}_j - |\boldsymbol{\Theta}_\beta|}{|\beta|\boldsymbol{x}_j - 1}\right)_+^\alpha\Bigg)_+\Bigg]$$

for all $\boldsymbol{x} \in \chi_\beta^0$ such that $\boldsymbol{x}_j \neq \frac{1}{|\beta|}$ for all $j \in \beta$ and with $\beta_+ \coloneqq \{j \in \beta : \boldsymbol{x}_j > \frac{1}{|\beta|}\}$ and $\beta_- \coloneqq \{j \in \beta : \boldsymbol{x}_j < \frac{1}{|\beta|}\}$.

(b) If $\boldsymbol{X} \in SRV(\alpha, \boldsymbol{Z})$ and Assumption RV holds then $\boldsymbol{X} \in RV(\alpha, \boldsymbol{\Theta})$, where $\boldsymbol{\Theta}$ satisfies

$$\mathbb{P}(\boldsymbol{\Theta} \in A_{\boldsymbol{x}}) = \mathbb{P}(\boldsymbol{Z} \in A_{\boldsymbol{x}}) + \frac{1}{\alpha} \sum_{\gamma \supset \beta} \mathrm{d}H_{\beta,\gamma}(\boldsymbol{x}_\beta, 0, 0) \left(\boldsymbol{x}_\beta - \frac{1}{|\gamma|}, \frac{-1}{|\gamma|}, \frac{-1}{|\gamma|+1}\right)^\top$$

for $\beta \in \mathcal{P}_d^*$ and λ_β-almost surely all $\boldsymbol{x} \in \chi_\beta$.

Hence, when a random vector $\boldsymbol{X}$ is regularly varying, we can also assume that under suitable assumptions that $\boldsymbol{X}$ is also sparse regularly varying. In Chapter 3 we will use the concept of sparse regular variation in more depth.

2.2.2. Principal component analysis

Alternatively to SRV, we use in this section another widely used method for dimension reduction namely Principal Component Analysis (PCA). We introduce a PCA approach for extremes, which is motivated by Drees and Sabourin (2021) and Drees (2025).

In PCA (see Muirhead, 1982) the set of principal components is used as a set of coordinates to represent the data as linear combinations of them. For d-dimensional data, there exist d principal components, which correspond to the eigenvectors of the covariance matrix. To achieve dimension reduction, only the principal components corresponding to directions with large variance, reflected by a large value of the corresponding eigenvalue, are used. The number of principal components to keep is later determined by information criteria.

PCA cannot be applied directly to a vector $\boldsymbol{X} \in \mathrm{RV}(\alpha, \boldsymbol{\Theta})$, since by Remark 2.3 (c) the second and even the first moments of the components do not have to exist. However, the spectral vector $\boldsymbol{\Theta}$ lives on the sphere and has norm equal to 1 and hence all moments exist. The application of PCA to $\boldsymbol{\Theta}$ is the subject of Drees and Sabourin (2021). Let $\boldsymbol{\Sigma} \coloneqq \mathrm{Cov}(\boldsymbol{\Theta})$ with spectral decomposition

$$\boldsymbol{\Sigma} = \boldsymbol{O}\boldsymbol{\Lambda}\boldsymbol{O}^\top,$$

where $\boldsymbol{O} \coloneqq (\boldsymbol{O}_1, \ldots, \boldsymbol{O}_d) \in \mathbb{R}^{d\times d}$ is an orthogonal matrix and $\boldsymbol{\Lambda} \coloneqq \mathrm{diag}(\lambda_1, \ldots, \lambda_d) \in \mathbb{R}^{d\times d}$ consists of the eigenvalues $\lambda_1 \geq \lambda_2 \geq \ldots \geq \lambda_d$ of $\boldsymbol{\Sigma}$. The principal components of $\boldsymbol{\Theta}$ are then given by

$$T_j = \boldsymbol{O}_j^\top(\boldsymbol{\Theta} - \mathbb{E}[\boldsymbol{\Theta}]), \quad j = 1, \ldots, d.$$

When we represent the data with only the first p eigenvectors $\boldsymbol{O}_1, \ldots, \boldsymbol{O}_p$, with $p \leq d$, we get

$$(T_1, \ldots, T_p) = (\boldsymbol{O}_1, \ldots, \boldsymbol{O}_p)^\top(\boldsymbol{\Theta} - \mathbb{E}[\boldsymbol{\Theta}]),$$

which is a low-dimensional representation of $\boldsymbol{\Theta} - \mathbb{E}[\boldsymbol{\Theta}]$. This transformation is optimal (see Seber, 1984) in the sense that the subspace $V_p = \mathrm{span}(\boldsymbol{O}_1, \ldots, \boldsymbol{O}_p)$ minimizes the risk $\mathbb{R}_\infty(V)$ (or the expected reconstruction error) defined as

$$R_\infty(V) \coloneqq \mathbb{E}[\|\Pi_V(\boldsymbol{\Theta}) - \boldsymbol{\Theta}\|^2] = \mathbb{E}[\|\Pi_{V^\perp}(\boldsymbol{\Theta})\|^2],$$

where Π_V is the projection of $\boldsymbol{\Theta}$ onto V, over all linear subspaces V with $\dim(V) = p$. The projection Π_V is estimated using the empirical covariance matrix of an i.i.d. sample $\boldsymbol{\Theta}_1, \ldots, \boldsymbol{\Theta}_n$ of $\boldsymbol{\Theta}$ given by

$$\widehat{\boldsymbol{\Sigma}}_n \coloneqq \frac{1}{n}\sum_{j=1}^{n}\left(\boldsymbol{\Theta}_j - \frac{1}{n}\sum_{i=1}^{n}\boldsymbol{\Theta}_i\right)\left(\boldsymbol{\Theta}_j - \frac{1}{n}\sum_{i=1}^{n}\boldsymbol{\Theta}_i\right)^\top$$

with eigenvalues $\widehat{\lambda}_1, \ldots, \widehat{\lambda}_d$.

In Drees and Sabourin (2021) the asymptotic properties of the empirical risk are derived and analyzed. Suppose $\boldsymbol{X}_1, \ldots, \boldsymbol{X}_n$ is an i.i.d. sequence following the distribution of $\boldsymbol{X}$. For $\widehat{\boldsymbol{\Theta}}_{n,i} \coloneqq \boldsymbol{X}_i/\|\boldsymbol{X}_i\|\mathbb{1}\{\|\boldsymbol{X}_i\| > \|\boldsymbol{X}_{(k_n+1,n)}\|\}$, $i = 1, \ldots, n, k_n \to \infty, k_n/n \to 0$ the empirical

counterpart of $\mathbb{R}_\infty(V)$ is defined as

$$\widehat{R}_{n,k_n}(V) \coloneqq \frac{1}{k_n}\sum_{i=1}^{n} \|\Pi_{V^\perp}(\widehat{\mathbf{\Theta}}_{n,i})\|^2.$$

Then Drees and Sabourin (2021) showed the consistency of the empirical risk, i.e. $\widehat{R}_{n,k_n}(V) \to R_\infty(V)$ as $n \to \infty$. In addition, the consistency of the empirical risk minimizer is presented. This is shown by the convergence to zero of the distance between the minimizing subspace with dimension p of $\widehat{R}_{n,k_n}(V)$ and $R_\infty(V)$ under a specific metric. Furthermore, risk bounds are derived for the empirical risk. In Drees (2025) the analysis of the risk was further extended and a multiple testing procedure is introduced to estimate the true subspace.

2.3. Model Selection and information criteria

The selection of a suitable model is an important step in statistics and extreme value theory. As stated in (Burnham and Anderson, 1998, p. 5), the question is 'What model to use?'. We first need to choose a model that fits our data before we can explore it. The tools used in this thesis for selecting a fitting model are information criteria. In other works in the field of extreme value theory information criteria have also been used. For example, in the context of graphical models for extremes Engelke and Hitz (2020) used information criteria to fit a multivariate Pareto-model and in Kiriliouk et al. (2019) information criteria were used for fitting multivariate generalized Pareto distributions. As a result, in the first part of this section we present the general idea of an information criterion and the definition of consistency. The section concludes with examples of information criteria such as the AIC in Section 2.3.1, the BIC in Section 2.3.2 and the mean squared error in Section 2.3.3. The application of the information criteria in the extreme setting is made in Chapter 3 and Chapter 4.

Next, we introduce the general concept of information criteria. For this, let $M_1, \dots, M_L$ be a family of models. We assume that we have a random variable Y which is distributed according to an unknown and unique model M_{l^*} for $1 \leq l^* \leq L$. Then, given realizations $Y_1, \dots, Y_n$ of Y, the goal is to determine which model is the true model. In general, an information criterion quantifies the fit of the observation Y to a possible model. In the following, we denote by $IC_n(l)$ an arbitrary information criterion for the l-th model M_l, $1 \leq l \leq L$ given $Y_1, \dots, Y_n$. We follow the convention that information criteria are negatively oriented and a lower value of an information criterion indicates a better fit. Thus, the idea of an information criterion is that it should be small when the correct model is chosen and large for all other models.

In the literature, there is a large number of different information criteria which are based on different ideas or refinements of other information criteria. In this work, we use the Akaike information criterion (AIC) of Akaike (1974), the Bayesian information criterion

(BIC) or the Schwarz information criterion of Schwarz (1978) and an information criterion based on the mean squared error. Other information criteria are the deviance information criterion (cf. Spiegelhalter et al., 2002), the Watanabe-Akaike information criterion (cf. Watanabe, 2010), the Takeuchi information criterion (cf. Burnham and Anderson, 1998), the Hannan-Quinn information criterion cf. (cf. Hannan and Quinn, 1979) and many more.

Next, we define consistency for information criteria.

Definition 2.16. Let M_{l^*} be the true model and $IC_n(l)$ an information criterion for the l-th model and $n \in \mathbb{N}$. The information criterion $IC_n(l)$ is called consistent if

$$\lim_{n\to\infty} \mathbb{P}((IC_n(l) - IC_n(l^*)) > 0) = 1 \text{ for } l \neq l^*.$$

In the following, we give a brief summary of the Akaike information criterion, the Bayesian information criterion and the mean squared error.

2.3.1. Akaike Information Criterion

The AIC is an estimator for the expected Kullback-Leibler divergence, evaluated at the maximum likelihood estimator. For a distribution P with density p and a family of distributions $M_l(\boldsymbol{\vartheta}_l)$ with density $q_{\boldsymbol{\vartheta}_l}$ for $\boldsymbol{\vartheta}_l \in \Theta_l$, where Θ_l is some parameter space and $l = 1, \ldots, L, L \in \mathbb{N}$ with respect to a measure μ, the Kullback-Leibler divergence is defined by

$$KL(P, M_l(\boldsymbol{\vartheta}_l)) = \int p \log\left(\frac{p}{q_{\boldsymbol{\vartheta}_l}}\right) \mathrm{d}\mu = \mathbb{E}\left[\log\left(\frac{p(Y)}{q_{\boldsymbol{\vartheta}_l}(Y)}\right)\right]$$

with $Y \sim P$.

The target is to determine the optimal model by minimizing the Kullback-Leibler divergence.

The approach of Akaike (Akaike, 1974) was first to insert the MLE for $\boldsymbol{\vartheta}_l$ and then to perform a bias correction. Hence, we approximate the Kullback-Leibler divergence by inserting an estimator $\widehat{\boldsymbol{\vartheta}}_l(\widetilde{Y}) \in \Theta_l$ for $\boldsymbol{\vartheta}_l$ of $\widetilde{Y} \sim P$ independent of data $Y \sim P$

$$\mathbb{E}_{\widetilde{Y}}[KL(P, M_l(\widehat{\boldsymbol{\vartheta}}_l(\widetilde{Y})))] = \mathbb{E}_{\widetilde{Y}}\left[\mathbb{E}_Y\left[\log\left(p(Y)\right)\right]\right] - \mathbb{E}_{\widetilde{Y}}\left[\mathbb{E}_Y\left[\log\left(q_{\widehat{\boldsymbol{\vartheta}}_l(\widetilde{Y})}(Y)\right)\right]\right].$$

Then we approximate the last term

$$\mathbb{E}_{\widetilde{Y}}\left[\mathbb{E}_Y\left[\log\left(q_{\widehat{\boldsymbol{\vartheta}}_l(\widetilde{Y})}(Y)\right)\right]\right]$$

with

$$\log\left(q_{\widehat{\boldsymbol{\vartheta}}_l(Y)}(Y)\right) - D_l$$

for a suitable constant $D_l > 0$. Since $\mathbb{E}_{\widetilde{Y}}\left[\mathbb{E}_Y\left[\log\left(p(Y)\right)\right]\right]$ is not influenced by the chosen model, the Akaike information criterion is defined as

$$-2\log\left(q_{\widehat{\boldsymbol{\vartheta}}_l(Y)}(Y)\right) + 2D_l.$$

The multiplication by -2 ensures that the information criterion is negatively oriented and it cancels the $1/2$ factor of a Gaussian log-likelihood function.

2.3.2. Bayesian Information Criterion

Alongside the AIC, the Bayesian information criterion (Schwarz, 1978) is one of the most popular information criteria in practice. The idea behind it is to select the model with the highest posterior probability.

Let data Y_n with true density f and a family of models M_l with parameter vector $\boldsymbol{\vartheta}_l$, $l = 1, \ldots, L$ be given. We denote the likelihood function of $\boldsymbol{\vartheta}_l$ given Y_n and the model M_l by $L_{M_l}(\boldsymbol{\vartheta}_l \,|\, Y_n)$, the probability for the l-th model M_l by $\pi(M_l)$ and by $g(\boldsymbol{\vartheta}_l \,|\, M_l)$ the prior distribution of $\boldsymbol{\vartheta}_l$ given M_l. The conditional density h of M_l and $\boldsymbol{\vartheta}_l$ given Y_n is with Bayes' theorem given by

$$h((M_l, \boldsymbol{\vartheta}_l) \,|\, Y_n) = \frac{\pi(M_l) g(\boldsymbol{\vartheta}_l \,|\, M_l) L(Y_n \,|\, (M_l, \boldsymbol{\vartheta}_l))}{f(Y_n)}.$$

The BIC aims to determine a model M_l that is suitable for the given data Y_n by maximizing the posterior probability of the model M_l given the data Y_n

$$\mathbb{P}(M_l \,|\, Y_n) = \int h((M_l, \boldsymbol{\vartheta}_l) \,|\, Y_n)\, \mathrm{d}\boldsymbol{\vartheta}_l$$

or equivalently, by minimizing

$$-2\log\mathbb{P}(M_l \,|\, Y_n) = -2\log\pi(M_l) - 2\log\int g(\boldsymbol{\vartheta}_l \,|\, M_l) L(Y_n \,|\, (M_l, \boldsymbol{\vartheta}_l))\, \mathrm{d}\boldsymbol{\vartheta}_l + 2\log f(Y_n).$$

In Cavanaugh and Neath (1999) it is shown that under certain assumptions it is possible to asymptotically bound the term

$$-2\log\int g(\boldsymbol{\vartheta}_l \,|\, M_l) L(Y_n \,|\, (M_l, \boldsymbol{\vartheta}_l))\, \mathrm{d}\boldsymbol{\vartheta}_l$$

by the lower bound

$$-2\log L(Y_n \,|\, (M_l, \widehat{\boldsymbol{\vartheta}}_l)) + D_l \log(n) + R_2(M_l)$$

and the upper bound

$$-2\log L(Y_n \,|\, (M_l, \widehat{\boldsymbol{\vartheta}}_l)) + D_l \log(n) + R_1(M_l)$$

where $R_1(M_l)$ and $R_2(M_l)$ are constant with respect to n and $R_2(M_l) < R_1(M_l)$, D_l is a constant depending on the model M_l and $\widehat{\boldsymbol{\vartheta}}_l$ is the MLE.

Then an information criterion based on these bounds can be defined by

$$-2 \log L(Y_n \mid (M_l, \widehat{\boldsymbol{\vartheta}}_l)) + D_l \log(n).$$

Remark 2.17. In the fixed-dimensional, large sample size setting (i.e. $d \in \mathbb{N}, n \to \infty$), the AIC is usually not consistent while the BIC is consistent (see Burnham and Anderson, 1998, Section 2.8.2 and Claeskens, 2016, Section 2.2.1). In the high-dimensional, large sample size setting, the AIC can be consistent (e.g. Bai et al., 2018).

2.3.3. Mean Squared Error

A criterion for measuring the performance of a model or an estimator is the mean squared error. In general, if $\widehat{\boldsymbol{\vartheta}} \in \mathbb{R}^d$ is an estimator for a true parameter $\boldsymbol{\vartheta} \in \mathbb{R}^d$ then the mean squared error is defined as

$$\mathbb{E}\left[\|\widehat{\boldsymbol{\vartheta}} - \boldsymbol{\vartheta}\|_2^2\right].$$

The mean squared error can be motivated by a normal distribution. If $\widehat{\boldsymbol{\vartheta}} \sim \mathcal{N}_d(\boldsymbol{\vartheta}, I_d)$, then follows for the likelihood function

$$L_{\mathcal{N}_d}(\boldsymbol{\vartheta} \mid \boldsymbol{x}) = (2\pi)^{-d/2} \exp\left(-\frac{1}{2}(\widehat{\boldsymbol{\vartheta}} - \boldsymbol{\vartheta})^\top (\widehat{\boldsymbol{\vartheta}} - \boldsymbol{\vartheta})\right)$$

of $\widehat{\boldsymbol{\vartheta}}$ that

$$\begin{aligned} -2\mathbb{E}\left[\log L_{\mathcal{N}_d}(\boldsymbol{\vartheta} \mid \widehat{\boldsymbol{\vartheta}})\right] &= d \log(2\pi) + \mathbb{E}\left[(\widehat{\boldsymbol{\vartheta}} - \boldsymbol{\vartheta})^\top (\widehat{\boldsymbol{\vartheta}} - \boldsymbol{\vartheta})\right] \\ &= d \log(2\pi) + \mathbb{E}\left[\|\widehat{\boldsymbol{\vartheta}} - \boldsymbol{\vartheta}\|_2^2\right]. \end{aligned}$$

Hence, minimizing the mean squared error is equivalent to maximizing the expected likelihood function. A lower value of the mean squared error indicates a better fit of the estimator $\widehat{\boldsymbol{\vartheta}} \in \mathbb{R}^d$ to the true parameter $\boldsymbol{\vartheta} \in \mathbb{R}^d$. Thus, the mean squared error is always positive and can be seen as a risk function. Further, it is possible to decompose the mean squared error into a bias and variance term

$$\begin{aligned} \mathbb{E}\|\widehat{\boldsymbol{\vartheta}} - \boldsymbol{\vartheta}\|_2^2 &= \mathbb{E}\Big[(\widehat{\boldsymbol{\vartheta}} - \boldsymbol{\vartheta})^\top (\widehat{\boldsymbol{\vartheta}} - \boldsymbol{\vartheta})\Big] \\ &= \operatorname{tr}\left(\mathbb{E}\Big[(\widehat{\boldsymbol{\vartheta}} - \boldsymbol{\vartheta})(\widehat{\boldsymbol{\vartheta}} - \boldsymbol{\vartheta})^\top\Big]\right) \\ &= \operatorname{tr}\left(\mathbb{E}\Big[\Big(\widehat{\boldsymbol{\vartheta}} - \mathbb{E}[\widehat{\boldsymbol{\vartheta}}] - (\boldsymbol{\vartheta} - \mathbb{E}[\widehat{\boldsymbol{\vartheta}}])\Big)\Big(\widehat{\boldsymbol{\vartheta}} - \mathbb{E}[\widehat{\boldsymbol{\vartheta}}] - (\boldsymbol{\vartheta} - \mathbb{E}[\widehat{\boldsymbol{\vartheta}}])\Big)^T\Big]\right) \\ &= \operatorname{tr}\left(\mathbb{E}\Big[\Big(\widehat{\boldsymbol{\vartheta}} - \mathbb{E}[\widehat{\boldsymbol{\vartheta}}]\Big)\Big(\widehat{\boldsymbol{\vartheta}} - \mathbb{E}[\widehat{\boldsymbol{\vartheta}}]\Big)^T\Big] + \mathbb{E}\Big[(\widehat{\boldsymbol{\vartheta}} - \boldsymbol{\vartheta})\Big]\mathbb{E}\Big[(\widehat{\boldsymbol{\vartheta}} - \boldsymbol{\vartheta})\Big]^\top\right) \end{aligned}$$

$$= \operatorname{tr}\left(\operatorname{Cov}(\widehat{\boldsymbol{\vartheta}})\right) + \|\mathbb{E}[\widehat{\boldsymbol{\vartheta}}] - \boldsymbol{\vartheta}\|_2^2.$$

Therefore, minimizing the mean squared error results in a bias-variance trade-off.

CHAPTER 3

Directions of extremes

In this chapter, we use the SRV approach, which is based on sparse regular variation introduced in Meyer and Wintenberger (2021, 2023) (cf. Section 2.2.1) and propose three different information criteria to estimate the number of extreme directions in both fixed and high-dimensional settings. The Bayesian information criterion (BIC), the quasi-Akaike information criterion (QAIC) and the mean squared error information criterion (MSEIC) are derived. They are particularly suitable for high-dimensional data with a sparsity structure in the extremes, while the application of these information criteria is very simple in practice and is not computationally intensive. The information criteria are first derived for the fixed-dimensional case and then subsequently adapted to the high-dimensional case. For the fixed-dimensional case, we also develop procedures to estimate the optimal threshold simultaneously to the estimation of the number of extreme directions.

Throughout this chapter, we assume that $\boldsymbol{X}$ is an $\mathbb{R}_+^d$-valued random vector, which is sparse regularly varying (cf. Definition 2.13). The goal is to determine all directions $\beta \subset \{1, \dots, d\}$ such that $\mathbb{P}(\boldsymbol{Z} \in C_\beta) > 0$, where $\boldsymbol{Z}$ is the limit vector from sparse regular variation (Definition 2.13) and C_β is defined in (1.2). Knowledge of these directions allows us to obtain maximal directions and hence, a dimension reduction as mentioned in **(Q)**.

The statistical model behind the proposed BIC is the same as that of the AIC in Meyer and Wintenberger (2023). 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 approximates the Kullback-Leibler divergence between the true model and a Gaussian model, rather than a multinomial model as used in the AIC and BIC, respectively. The advantage of the BIC and QAIC in the fixed-dimensional case over the AIC is that they are consistent information criteria for *bias selection*. In this setting, bias selection refers to estimating the number of extreme directions and thereby separating them from the bias directions introduced later. Finally, the third method, the MSEIC (mean squared error information criteria) is so named because it is based on an approximate of the mean squared error (MSE) between the relative number of extreme observations and the true probabilities of extremes in the different subspaces C_β. Although the MSEIC is not consistent for bias selection in the fixed-dimensional case, it performs remarkably well in all simulations. However, the consistency results differ in the high-dimensional case, to which we adapt the information criteria. While the BIC remains consistent, the AIC, MSEIC and QAIC are also consistent. For the fixed-dimensional case, we further expand the information criteria for the *threshold selection*, where the threshold

is estimated simultaneously.

The chapter is organized as follows. In Section 3.1 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 (2023). We also present statistical models for our information criteria. The main results of the chapter for the fixed-dimensional case are derived in Sections 3.2.1 to 3.2.3. In Section 3.2.1, we first introduce the QAIC for bias selection and threshold selection following the framework of an Akaike information criterion, which aims to minimize the expected Kullback-Leibler (KL) divergence, here applied to a Gaussian likelihood function. Unlike the AIC proposed by Meyer and Wintenberger (2023), we prove that the QAIC for bias selection is a consistent information criterion. In Section 3.2.2, we develop the MSEIC and finally, in Section 3.2.3, 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. Then, in Section 3.3 the information criteria for bias selection are adapted to the high-dimensional case and conditions for the consistency of all information criteria are derived. The main proofs of the chapter are moved to Section 3.4, while the proofs of some auxiliary results can be found in Appendix A. Note that most parts of this chapter consist of Butsch and Fasen-Hartmann (2025b, 2026).

3.1. PRELIMINARIES

This section addresses the main concepts of this chapter, which are based on Meyer and Wintenberger (2021, 2023). Recall that sparse regular variation was introduced in Section 2.2.1. We start with the introduction of an *extreme direction* for sparse regular variation in Section 3.1.1. 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 3.1.2. Then, in Section 3.1.3, we give an overview of the statistical inference of the empirical estimator of the probabilities of extreme directions and the assumptions of the present chapter. Finally, in Section 3.1.4, we present statistical models on which the information criteria are based. Note that throughout this section, we assume that $d \in \mathbb{N}$ is fixed.

3.1.1. SPARSE REGULAR VARIATION AND EXTREME DIRECTIONS

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 $\{1, \dots, d\}$ and $\mathcal{P}_d^* \coloneqq \mathcal{P}_d \setminus \emptyset$.

Definition 3.1. 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

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

In this context, we reduce the dimension by only considering the corresponding entries of $\boldsymbol{X}$ for each maximal directions $\beta \in \mathcal{S}(\boldsymbol{Z})$. The aim of this chapter is to estimate s^*, the number of extreme directions under sparse regular variation, by using information criteria in order to estimate the maximal directions.

Remark 3.2. The use of the Euclidean projection leads to a sparse representation, in the sense that under π more components are projected to zero compared to the normalization $\boldsymbol{v} \mapsto \boldsymbol{v}/\|\boldsymbol{v}\|_1$ as mentioned in Section 2.2.1. Therefore, it is not surprising that according to Meyer and Wintenberger (2021, 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.

3.1.2. 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 better understand the idea of bias directions, we require some further notation. Suppose $\|\boldsymbol{X}_{(1,n)}\|_1 \geq \cdots \geq \|\boldsymbol{X}_{(n,n)}\|_1$ is the order statistic of $\|\boldsymbol{X}_1\|_1, \ldots, \|\boldsymbol{X}_n\|_1$ 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 $k_n/n \sim \mathbb{P}(\|\boldsymbol{X}\|_1 > u_n)$ and $u_n \to \infty$ as $n \to \infty$. Due to Meyer and Wintenberger (2023, Proposition 1) the empirical estimator

$$\frac{T_n(C_\beta, k_n)}{k_n} \coloneqq \frac{1}{k_n} \sum_{j=1}^{n} \mathbb{1}\left\{\pi(\boldsymbol{X}_j/\|\boldsymbol{X}_{(k_n+1,n)}\|_1) \in C_\beta, \|\boldsymbol{X}_j\|_1 > \|\boldsymbol{X}_{(k_n+1,n)}\|_1\right\},$$

of the probability

$$p(C_\beta) \coloneqq \mathbb{P}(\boldsymbol{Z} \in C_\beta) = \lim_{n\to\infty} \mathbb{P}(\pi(\boldsymbol{X}/u_n) \in C_\beta \mid \|\boldsymbol{X}\|_1 > u_n) \tag{3.1}$$

is a consistent estimator, so that the set of empirically observed 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 ease of notation, and

$$p_n(C_\beta) \;\coloneqq\; \mathbb{P}(\pi(\boldsymbol{X}/u_n) \in C_\beta \mid \|\boldsymbol{X}\|_1 > u_n).$$

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 (2023, Proposition 2) is then that under Assumption HRV even

$$\lim_{n\to\infty} \mathbb{P}(\mathcal{S}(\boldsymbol{Z}) \subseteq \mathcal{R} \subseteq \widehat{\mathcal{S}}_n(\boldsymbol{Z})) = 1 \tag{3.2}$$

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 $T_n(C_\beta, k_n) > 0$ that is not an extreme direction; a mathematical more rigorous interpretation is given in Meyer and Wintenberger (2023). Such a direction is referred to as a bias direction as mentioned above. The main challenge is to identify these bias directions.

Remark 3.3. There exists as well a stronger statement than (3.2). 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 (2023, 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 $\widehat{s}_n \coloneqq |\mathcal{R}| = |\widehat{\mathcal{S}}_n(\boldsymbol{Z})| \xrightarrow{\mathbb{P}} r$ as $n \to \infty$.

3.1.3. Statistical inference for the probabilities of extreme directions

The general assumptions of the present chapter are motivated by the statistical inference of the probabilities of extreme directions as derived in Meyer and Wintenberger (2023). 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 (3.1):

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

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 $j = 1, \ldots, 2^d - 1$,

$$\begin{aligned}
p_j &\coloneqq p(C_{\beta_j}), & p_{n,j} &\coloneqq p_n(C_{\beta_j}) \coloneqq \mathbb{P}(\pi(\boldsymbol{X}/u_n) \in C_{\beta_j} \mid \|\boldsymbol{X}\|_1 > u_n), \\
\mathcal{T}_{n,j} &\coloneqq \mathcal{T}_n(C_{\beta_j}), & T_{n,j}(k_n) &\coloneqq T_n(C_{\beta_j}, k_n),
\end{aligned}$$

where

$$\frac{\mathcal{T}_n(C_\beta)}{k_n} \coloneqq \frac{1}{k_n} \sum_{j=1}^{n} \mathbb{1}\{\pi(\boldsymbol{X}_j/u_n) \in C_\beta, \|\boldsymbol{X}_j\|_1 > u_n\}.$$

Finally, we define the associated vectors

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

In the next theorem, we summarize the asymptotic behavior of these estimators as derived in Meyer and Wintenberger (2023, Theorem 1 and Proposition 3).

Proposition 3.4. *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 $j = 1, \ldots, r$ as $n \to \infty$,*

$$\sup_{r \in [\frac{1}{1+\tau}, 1+\tau]} \sqrt{\frac{k_n}{p_{n,j}}} \Big| \frac{n}{k_n} \mathbb{P}(\boldsymbol{X}/u_n \in \{\boldsymbol{x} \in \mathbb{R}_+^d : r\|\boldsymbol{x}\|_1 > 1, \pi(r\boldsymbol{x}) \in C_{\beta_j}\}) - r^{\alpha(\beta_j)} p_{n,j} \Big| \to 0.$$

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

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

(b) If additionally $\sqrt{k_n}(p_{n,j} - p_j) \to 0$ *as* $n \to \infty$ *and* $j = 1, \ldots, r$*, 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 $n \in \mathbb{N}$

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

and suppose the following assumption throughout the chapter.

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 $n \to \infty$*,*

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

Remark 3.5.

(a) A justification of Assumption (A1) is given in Remark 3.3, 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 modeling purposes, as can be seen in the derivation of the AIC in Meyer and Wintenberger (2023), and from other statements in that chapter such as Proposition 3.4 above. If Assumption (A1) is not made, then the consistency results in this chapter can be obtained by replacing r with $\widehat{s}_n := |\mathcal{S}_n(\boldsymbol{Z})|$ and assuming $\sqrt{k_n\rho_n}(\widehat{s}_n - r) \xrightarrow{\mathbb{P}} 0$ (cf. Remark 3.8 and Remark 3.15).

(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 3.4 gives a sufficient criterion for (A3) or (A4) to hold.

The following lemma is a direct consequence of Assumption A.

Lemma 3.6. *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 $j = 1, \dots, s^*$ *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 $j = s^* + 1, \dots, r$ *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,

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

3.1.4. 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 (2023). 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* and the statistical models for the *threshold selection*.

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^* and answer question **(Q)**, the idea is now to fit for any $s \in \{1, \dots, r\}$ a multinomial distribution from the class $\{\mathrm{Mult}(k_n, \boldsymbol{A}_s(\widetilde{\boldsymbol{p}}^s)) : \widetilde{\boldsymbol{p}}^s \in \Theta_s\}$ where $\boldsymbol{A}_s : \mathbb{R}^s \to \mathbb{R}^r$ is defined as

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

and the parameter space Θ_s is defined as

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

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

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

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*

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

and log-likelihood function

$$\begin{aligned}\log L_{M_{k_n}^s}(\widetilde{\boldsymbol{p}}^s \,|\, \boldsymbol{T}_n(k_n)) = \log(k_n!) - \sum_{j=1}^r \log(T_{n,j}(k_n)!) + \sum_{j=1}^s T_{n,j}(k_n)\log(\widetilde{p}_j^s) \\ + \log(\widetilde{\rho}^s) \sum_{j=s+1}^r T_{n,j}(k_n)\end{aligned} \tag{3.3}$$

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

$$\widehat{\boldsymbol{p}}_{n,*}^{\widehat{s}_n^*} \coloneqq \Bigg(\frac{\widehat{p}_{n,1}^{\widehat{s}_n^*}}{\sum_{j=1}^{\widehat{s}_n^*} \widehat{p}_{n,j}^{\widehat{s}_n^*}}, \dots, \frac{\widehat{p}_{n,\widehat{s}_n^*}^{\widehat{s}_n^*}}{\sum_{j=1}^{\widehat{s}_n^*} \widehat{p}_{n,j}^{\widehat{s}_n^*}}, 0, \dots, 0 \Bigg)^\top, \tag{3.4}$$

where

$$\widehat{\boldsymbol{p}}_n^s \coloneqq (\widehat{p}_{n,1}^s, \dots, \widehat{p}_{n,s}^s)^\top \coloneqq \Big(\frac{T_{n,1}(k_n)}{k_n}, \dots, \frac{T_{n,s}(k_n)}{k_n} \Big)^\top \tag{3.5}$$

is the maximum likelihood estimator (MLE) of the multinomial model $M_{k_n}^s$ (see Meyer and Wintenberger, 2023, Section 4.1). Finally, we define

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

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

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 the 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 $q_n\in(0,1)$. Note, that the vector $\boldsymbol{T}'_n$ is not directly observable because we do not know which observations are extreme. To be more precise we assume that $\boldsymbol{T}'_n\sim\mathrm{Mult}(n,\boldsymbol{p}'_n)$ with

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

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

$$\mathbb{P}_{(T'_{n,1},\dots,T'_{n,2^d-1})|T'_{n,2^d}=n-k_n}=\mathbb{P}_{(T_{n,1}(k_n),\dots,T_{n,2^d-1}(k_n))}. \tag{3.6}$$

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 non-observable 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

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

Finally, we define

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

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*

$$\begin{aligned}\log L_{M'^s_n}(\widetilde{\boldsymbol{p}}'^s\,|\,\boldsymbol{T}'_n)=&\log(n!)-\sum_{j=1}^{2^d}\log(T'_{n,j}!)+\sum_{j=1}^s T_{n,j}\log(\widetilde{q}\widetilde{p}^s_j)\\&+\left(\sum_{j=s+1}^{2^d-1}T'_{n,j}\right)\log(\widetilde{q}\widetilde{\rho}^s)+T'_{n,2^d}\log(1-\widetilde{q})\end{aligned} \tag{3.7}$$

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 $n \to \infty$.

3.2. Information criteria in the fixed-dimensional case

In this section we derive the information criteria for the fixed-dimensional case, i.e. we assume that $d \in \mathbb{N}$ is fixed. We start with the QAIC in Section 3.2.1, where we also analyze the consistency of the AIC from Meyer and Wintenberger (2023). Then, in Section 3.2.2 and Section 3.2.3 we derive the MSEIC and the BIC, respectively.

3.2.1. 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 (2023), 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 (2023) 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 $\text{Mult}(k_n, \widehat{\boldsymbol{p}}_n^s)$ where $\widehat{\boldsymbol{p}}_n^s$ is the MLE given in (3.5). The AIC is defined as

$$\text{AIC}_{k_n}(s) \coloneqq -\log L_{M_{k_n}^s}(\widehat{\boldsymbol{p}}_n^s \,|\, \boldsymbol{T}_n(k_n)) + s, \quad s = 1, \dots, r, \tag{3.8}$$

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 $d \in \mathbb{N}$ (cf. Remark 2.17).

Theorem 3.7. *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 3.7 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 3.7, along with all proofs of this section, is relegated to Section 3.4.1.

Remark 3.8. Suppose Assumption (A1) is replaced by the condition $\sqrt{k_n\rho_n}(\widehat{s}_n - r) \xrightarrow{\mathbb{P}} 0$ and that the AIC is defined using $\widehat{s}_n$ instead of r. Then

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

and hence, if we follow the proof of Theorem 3.7, 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.

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 an 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 $\mathcal{N}_r(k_n\boldsymbol{B}_s(\underline{\widetilde{\boldsymbol{p}}}^s), k_n\,\mathrm{diag}(\boldsymbol{B}_s(\underline{\widetilde{\boldsymbol{p}}}^s))), \underline{\widetilde{\boldsymbol{p}}}^s = (\underline{\widetilde{p}}_1^s, \dots, \underline{\widetilde{p}}_s^s, \underline{\widetilde{\rho}}^s) \in \mathbb{R}_+^{s+1}$, where $\boldsymbol{B}_s : \mathbb{R}_+^{s+1} \to \mathbb{R}_+^r$ is defined as

$$\boldsymbol{B}_s(\boldsymbol{z}) = \Big(z_1, \dots, z_s, z_{s+1}, \dots, z_{s+1}\Big)^\top.$$

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

$$\begin{aligned} \widehat{\underline{\boldsymbol{p}}}_n^s(\widetilde{\boldsymbol{\mathcal{T}}}_n) &:= (\widehat{\underline{p}}_{n,1}^s(\widetilde{\boldsymbol{\mathcal{T}}}_n), \dots, \widehat{\underline{p}}_{n,s}^s(\widetilde{\boldsymbol{\mathcal{T}}}_n), \widehat{\underline{\rho}}_n^s(\widetilde{\boldsymbol{\mathcal{T}}}_n))^\top \in \mathbb{R}_+^{s+1} \qquad \text{with} \\ \widehat{\underline{p}}_{n,j}^s(\widetilde{\boldsymbol{\mathcal{T}}}_n) &:= \frac{\widetilde{\mathcal{T}}_{n,j}}{k_n}, \quad j = 1, \dots, s, \qquad \widehat{\underline{\rho}}_n^s(\widetilde{\boldsymbol{\mathcal{T}}}_n) := \frac{1}{r-s} \sum_{j=s+1}^{r} \frac{\widetilde{\mathcal{T}}_{n,j}}{k_n} \end{aligned} \tag{3.9}$$

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

Remark 3.9. It might happen that $\sum_{j=1}^s \widehat{\underline{p}}_{n,j}^s(\widetilde{\boldsymbol{\mathcal{T}}}_n) + (r-s)\widehat{\underline{\rho}}_n^s(\widetilde{\boldsymbol{\mathcal{T}}}_n) \neq 1$. In this case, $\boldsymbol{B}_s(\widehat{\underline{\boldsymbol{p}}}_n^s(\widetilde{\boldsymbol{\mathcal{T}}}_n))$ is in general not a probability vector and $(\widehat{\underline{p}}_{n,1}^s(\widetilde{\boldsymbol{\mathcal{T}}}_n), \dots, \widehat{\underline{p}}_{n,s}^s(\widetilde{\boldsymbol{\mathcal{T}}}_n)) \notin \Theta_s$. But due to Assumption (A4) we have as $n \to \infty$,

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

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

In summary, we calculate

$$\begin{aligned} &\mathbb{E}\left[\mathrm{KL}(\mathbb{P}_{\boldsymbol{\mathcal{T}}_n}, \mathcal{N}_r(k_n \boldsymbol{B}_s(\widetilde{\underline{\boldsymbol{p}}}^s), k_n \operatorname{diag}(\widetilde{\underline{\boldsymbol{p}}}^s)))|_{\widetilde{\underline{\boldsymbol{p}}}^s = \widehat{\underline{\boldsymbol{p}}}_n^s(\widetilde{\boldsymbol{\mathcal{T}}}_n)}\right] \\ &\qquad = \mathbb{E}\left[\log f_*(\boldsymbol{\mathcal{T}}_n)\right] - \mathbb{E}\left[\log\left(L_{\mathcal{N}_r}(\widehat{\underline{\boldsymbol{p}}}_n^s(\widetilde{\boldsymbol{\mathcal{T}}}_n) | \boldsymbol{\mathcal{T}}_n)\right)\right]. \end{aligned} \tag{3.10}$$

Remark 3.10. 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 (2021).

Next, we derive an auxiliary result that helps to approximate the second term in (3.10) for $s \geq s^*$.

Proposition 3.11. *Suppose Assumption A holds and $s \geq s^*$. Furthermore, let $\widetilde{\boldsymbol{\mathcal{T}}}_n$ be an independent and identically distributed copy of $\boldsymbol{\mathcal{T}}_n$, and let $\widehat{\underline{\boldsymbol{p}}}_n^s(\widetilde{\boldsymbol{\mathcal{T}}}_n)$ be the estimator in* (3.9) *and similarly we define $\widehat{\underline{\boldsymbol{p}}}_n^s(\boldsymbol{\mathcal{T}}_n)$. Then there exists a random variable Y with $\mathbb{E}[Y] = 0$ such that as $n \to \infty$,*

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

Therefore, for $s \geq s^*$ we approximate the second term in (3.10) by

$$-\mathbb{E}\left[\log L_{\mathcal{N}_r}(\widehat{\underline{\boldsymbol{p}}}_n^s(\widetilde{\boldsymbol{\mathcal{T}}}_n) \,|\, \boldsymbol{\mathcal{T}}_n)\right]$$

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

and neglect the expectation. The first term $\mathbb{E}\left[\log f_*(\boldsymbol{\mathcal{T}}_n)\right]$ in (3.10) 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 $s \geq s^*$,

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

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

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

and for $s > s^*$ we have

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

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

Moreover, since

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

the choice between estimator $\widehat{\underline{\boldsymbol{p}}}^s_n(\boldsymbol{\mathcal{T}}_n)$ 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, $\widehat{\underline{\boldsymbol{p}}}^s_n(\boldsymbol{\mathcal{T}}_n)$ is unknown, we finally define the information criterion based on the estimators $\widehat{\boldsymbol{p}}^s_n$ and $\widehat{\rho}^s_n$.

Definition 3.12. 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 $\widehat{s}^*_n \coloneqq \arg\min_{1\leq s\leq r}\mathrm{QAIC}_{k_n}(s)$.

Remark 3.13.

(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

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

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

Theorem 3.14. *Suppose Assumption A holds. Then*

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

In view of question **(Q)**, the QAIC has the advantage that it is weakly consistent for fixed k_n in contrast to the AIC.

Remark 3.15. Suppose Assumption (A1) is replaced by the condition $\widehat{s}_n \xrightarrow{\mathbb{P}} r$ and that the QAIC is defined using $\widehat{s}_n$ 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 3.8, where we required $\sqrt{k_n\rho_n}(\widehat{s}_n - r) \xrightarrow{\mathbb{P}} 0$.

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 (2023) which is defined as

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

with $\mathrm{AIC}_{k_n}(s)$ as in (3.8). 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 3.16. 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

$$\mathrm{QAIC}_{n,s}(k_n) := \frac{\mathrm{QAIC}_{k_n}(s)}{2k_n} + \frac{k_n}{n}$$

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

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

Remark 3.17. 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.

3.2.2. 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 as well as for the threshold k_n, which performs in particular well for a small number of observations. The proofs of this section are moved to Section 3.4.2.

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 (3.10). 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

$$\mathrm{MSE}_{k_n}(s) := \mathbb{E}\left[\ell^2(\widehat{\underline{\boldsymbol{p}}}_n^s(\widetilde{\boldsymbol{T}}_n(k_n))|\boldsymbol{T}_n)\right] \tag{3.11}$$

instead of $\mathbb{E}\left[\log L_{\mathcal{N}_r}(\widehat{\underline{\boldsymbol{p}}}_n^s(\widetilde{\boldsymbol{\mathcal{T}}}_n) \,|\, \boldsymbol{T}_n(k_n))\right]$ as is done in (3.10), where $\widetilde{\boldsymbol{T}}_n(k_n)$ is an independent and identically distributed copy of $\boldsymbol{T}_n(k_n)$ and

$$\begin{aligned}\ell^2(\widetilde{\underline{\boldsymbol{p}}}^s|\boldsymbol{T}_n(k_n)) &:= \left\| \sqrt{k_n}\,\mathrm{diag}(\boldsymbol{B}_s(\widetilde{\underline{\boldsymbol{p}}}^s))^{-1/2}\left(\frac{\boldsymbol{T}_n(k_n)}{k_n} - \boldsymbol{B}_s(\widetilde{\underline{\boldsymbol{p}}}^s)\right)\right\|_2^2 \\ &= \sum_{j=1}^{s}\frac{k_n}{\widetilde{\underline{p}}_j^s}\left(\frac{T_{n,j(k_n)}}{k_n} - \widetilde{\underline{p}}_j^s\right)^2 + \frac{k_n}{\widetilde{\underline{\rho}}^s}\sum_{j=s+1}^{r}\left(\frac{T_{n,j}(k_n)}{k_n} - \widetilde{\underline{\rho}}^s\right)^2\end{aligned}$$

for $\widetilde{\underline{\boldsymbol{p}}}^s = (\widetilde{\underline{p}}_1^s, \dots, \widetilde{\underline{p}}_s^s, \widetilde{\underline{\rho}}^s) \in \mathbb{R}_+^{s+1}$. Note, if in Assumption (A3) not only the weak convergence but also the componentwise L_1 convergence holds, then the approach is motivated by $\lim_{n\to\infty}\mathbb{E}\left[\ell^2((p_{n,1}, \dots, p_{n,s^*}, \rho_n)|\boldsymbol{T}_n(k_n))\right] = r - 1$. First, we derive an auxiliary result that helps to approximate $\ell^2(\widehat{\underline{\boldsymbol{p}}}_n^s(\widetilde{\boldsymbol{T}}_n(k_n))|\boldsymbol{T}_n(k_n))$.

Theorem 3.18. *Suppose Assumption A holds and $s \ge s^*$. 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 (3.9). *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* $n \to \infty$,

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

Therefore, for $s \geq s^*$ we approximate (3.11) 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 3.2.1, we neglect the expectation, which leads to the following information criterion.

Definition 3.19. 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 3.20. *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 5.1 shows that MSEIC performs extremely good in practice.

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 $\boldsymbol{T}'_{n,\{1,\ldots,r\}}$ 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

$$\begin{aligned}\mathrm{MSE}_n'^s :=& \mathbb{E}\left[q'\mathbb{E}\left[\left\|\sqrt{n-T'_{n,2^d}}\,\mathrm{diag}(\boldsymbol{p}')^{-1/2}\left(\frac{\boldsymbol{T}'_{n,\{1,\dots,r\}}}{n-T'_{n,2^d}}-\boldsymbol{p}'\right)\right\|_2^2\right]\Big|_{\boldsymbol{p}'=\widehat{\boldsymbol{p}}_n(\frac{\widetilde{\boldsymbol{T}}_n(k_n)}{k_n}),q'=\frac{k_n}{n}}\right] \\ &+\mathbb{E}\left[(1-q')\mathbb{E}\left[\left\|\sqrt{n}(q'(1-q'))^{-1/2}\left(\frac{T'_{n,2^d}}{n}-(1-q')\right)\right\|_2^2\right]\Big|_{q'=\frac{k_n}{n}}\right] \qquad (3.12)\end{aligned}$$

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

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

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 3.22. 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 3.23. 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.

3.2.3. BAYESIAN INFORMATION CRITERION

The basic idea of the Bayesian information criterion (BIC) is to find the model with the highest posterior probability given the data. First, we derive a BIC for s and then for k_n. The proofs of this section can be found in Section 3.4.3.

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 (1999). 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 $\{M_{k_n}^s : s = 1, \dots, r\}$, $g(\cdot \,|\, M_{k_n}^s)$ be the prior density over the parameter space Θ_s given Model $M_{k_n}^s$, $L_{M_{k_n}^s}(\cdot \,|\, \boldsymbol{T}_n(k_n))$ 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 $\widetilde{\boldsymbol{p}}^s$

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

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

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

Consequently maximizing the posterior probability is equivalent to minimizing

$$\begin{aligned} -2 \log \mathbb{P}(M_{k_n}^s|\boldsymbol{T}_n(k_n)) =& 2 \log f(\boldsymbol{T}_n(k_n)) - 2 \log \mathbb{Q}(M_{k_n}^s) \\ & - 2 \log\Big(\int_{\Theta_s} L_{M_{k_n}^s}(\widetilde{\boldsymbol{p}}^s \,|\, \boldsymbol{T}_n(k_n)) g(\widetilde{\boldsymbol{p}}^s \,|\, M_{k_n}^s) \, \mathrm{d}\widetilde{\boldsymbol{p}}^s \Big). \end{aligned} \tag{3.13}$$

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 $g(\cdot \,|\, M_{k_n}^s)$ on Θ_s satisfies

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

(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 3.24.

(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 chapter 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 (3.13) 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 3.25. *Suppose Assumptions A, (C1) and (C3) hold. Then the inequality*

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

as $n \to \infty$ holds.

Plugging in Assumption (C2) and the upper bound in Theorem 3.25 in (3.13) results in

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

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

Definition 3.26. For the number of extreme directions s with fixed k_n the *Bayesian information criterion concerning the upper bound* (BICU) is defined as

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

for $s = 1, \ldots, r-1$ 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 3.50, 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 3.50.

Definition 3.27. 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

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

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

Theorem 3.28. *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, 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 (cf. Remark 2.17).

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 using the following assumptions.

Assumption D. *Suppose the following statements hold.*

(D1) There exist constants $0 < b' \le B' < \infty$ *such that the prior density* $g'(\cdot \,|\, M_n'^s)$ *on* $\boldsymbol{\Theta}_s'$ *satisfies*

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

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

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

holds.

Remark 3.29.

(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 $q_n \coloneqq n^{-11/20}$ fulfills the conditions of Assumption (D3).

(c) Assumption (C3) for the local model is required for the proof of Theorem 3.25. Assumption (D4) for the global model is motivated from Theorem 3.25 and (3.6). Because we then obtain directly

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

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

In analogy to the BIC for the local model, 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

$$-2\log\mathbb{E}_{g'_s}[L_{M'^s_n}(\widetilde{\boldsymbol{p}}'^s|\boldsymbol{T}'_n)] \coloneqq -2\log\left\{\int_{\Theta'_s} L_{M'^s_n}(\widetilde{\boldsymbol{p}}'^s \,|\, \boldsymbol{T}'_n)\cdot g'(\widetilde{\boldsymbol{p}}'^s|M'^s_n)\,\mathrm{d}\widetilde{\boldsymbol{p}}'^s\right\}, \tag{3.14}$$

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

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

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

Compared to Theorem 3.25 in the previous section, we take additionally the expectation in Theorem 3.30 to achieve a connection between the global model and the local model.

Theorem 3.30 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 3.31. 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

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

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 3.27 we also define the Bayesian information criterion based on the lower bound for the threshold k_n.

Definition 3.32. 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 $K \subset \{1, \ldots, n\}$.

Remark 3.33. Note that in practice, we estimate, of course, r by $\widehat{s}_n = |\widehat{\mathcal{S}}_n(\boldsymbol{Z})|$ and plug this estimate into the information criteria. In this setup, all the consistency results in this section remain true if we additionally assume $\sqrt{k_n\rho_n}(\widehat{s}_n - r) \xrightarrow{\mathbb{P}} 0$ 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.

3.3. High-dimensional setting

Until now, we considered the setting where the dimension d is fixed. In the following, we analyze the high-dimensional case, where it is assumed that the dimension is allowed to grow with the sample size n. Since the number of possible bias directions increases, we also let the number of both bias and observed directions increase. However, the number of relevant directions is assumed to remain constant. In this section, the information criteria from the fixed-dimensional model are adapted to this high-dimensional setting. This requires adapting the previous definitions for the high-dimensional setting, although the underlying motivation is unchanged.

This section is structured as follows. In Section 3.3.1 we introduce the underlying model and define bias directions in the high-dimensional setting. In addition, we analyze the (numerical) behavior of the bias directions in a small simulation study in Section 3.3.2. Finally, in Section 3.3.3, the information criteria are adapted to the high-dimensional setting. The proofs of this section are moved to Section 3.4.5.

3.3.1. Bias directions in high dimensions

For the high-dimensional case, we expand the model from the fixed-dimensional case used in Section 3.1.4. Let $\boldsymbol{X}^{(n)}$ be an $\mathbb{R}_+^{d_n}$-valued random vector, where $d_n \to \infty$. Similarly to the fixed-dimensional case, we assume that $\boldsymbol{X}^{(n)}$ exhibits some extreme behavior but the number of directions in which the extremes occur is fixed and relatively small compared to the growing dimension d_n. As before, the major challenge in estimating the extreme directions is that we tend to identify more extremal directions than actually exist. To better understand the idea of these directions in high dimensions, we require some further notation. Let $\boldsymbol{X}_1^{(n)}, \boldsymbol{X}_2^{(n)}, \dots$ be a sequence of i.i.d. random vectors following the distribution of $\boldsymbol{X}^{(n)}$. Suppose $\|\boldsymbol{X}_{(1,n)}^{(n)}\|_1 \geq \cdots \geq \|\boldsymbol{X}_{(n,n)}^{(n)}\|_1$ is the order statistic of $\|\boldsymbol{X}_1^{(n)}\|_1, \dots, \|\boldsymbol{X}_n^{(n)}\|_1$ and the number of extreme observations used for estimation is denoted by $k_n \in \mathbb{N}$, whereas we assume that $k_n \to \infty$ and $k_n/n \to 0$ as $n \to \infty$. Then we define for $\beta \in \mathcal{P}_{\mathbb{N}}^* \coloneqq \mathcal{P}(\mathbb{N}) \setminus \{\emptyset\}$,

$$T^{(n)}(\beta, k_n) \coloneqq \sum_{j=1}^{n} \mathbb{1}\left\{ \left(\pi\left(\frac{\boldsymbol{X}_j^{(n)}}{\|\boldsymbol{X}_{(k_n+1,n)}^{(n)}\|_1} \right), 0, 0, \dots \right) \in C_\beta, \|\boldsymbol{X}_j^{(n)}\|_1 > \|\boldsymbol{X}_{(k_n+1,n)}^{(n)}\|_1 \right\}.$$

Note that if $\beta \notin \mathcal{P}_{d_n}^*$, then $T^{(n)}(\beta, k_n) = 0$.

The directions $\beta \in \mathcal{P}_{\mathbb{N}}^*$ are partitioned and named as follows using that

$$\sum_{\beta \in \mathcal{P}_{\mathbb{N}}^*} T^{(n)}(\beta, k_n) = k_n.$$

Definition 3.34.

(a) A direction $\beta \in \mathcal{P}_{\mathbb{N}}^*$ is called a *relevant direction* if $\liminf_{n\to\infty} T^{(n)}(\beta, k_n)/k_n > 0$ a.s.

The number of relevant directions is defined by

$$s^* := |\{\beta \in \mathcal{P}^*_{\mathbb{N}} : \liminf_{n\to\infty} T^{(n)}(\beta, k_n)/k_n > 0\}|.$$

(b) A direction $\beta \in \mathcal{P}^*_{\mathbb{N}}$ is called a *bias direction of type (i)* if $T^{(n)}(\beta, k_n)/k_n \xrightarrow{\mathbb{P}} 0$ and $T^{(n)}(\beta, k_n) \xrightarrow{\mathbb{P}} \infty$.

(c) A direction $\beta \in \mathcal{P}^*_{\mathbb{N}}$ is called a *bias direction of type (ii)* if $T^{(n)}(\beta, k_n)/k_n \xrightarrow{\mathbb{P}} 0$ and $\limsup_{n\to\infty} T^{(n)}(\beta, k_n) < \infty$ a.s.

(d) The *set of observed directions* is defined as

$$\widehat{\mathcal{S}}_n := \{\beta \in \mathcal{P}^*_{\mathbb{N}} : T^{(n)}(\beta, k_n) > 0\}$$

and the *number of observed directions* as

$$\widehat{s}_n := |\widehat{\mathcal{S}}_n|.$$

Remark 3.35.

(a) For large n, the support of $\boldsymbol{X}^{(n)}$ given that $\|\boldsymbol{X}^{(n)}\|_1$ is large, becomes more concentrated on the space spanned by the relevant directions.

(b) As we are working with finite n we are not able to observe only relevant directions but also bias directions, which act as noise. The number of observations in each bias direction grows at a slower rate than k_n, but it can still go to infinity.

Furthermore, we order the values of $T^{(n)}(\beta, k_n)$ for $\beta \in \mathcal{P}^*_{\mathbb{N}}$ such that $T^{(n)}_i(k_n)$ corresponds to the direction with the i-th largest number of observations. The ordered vector containing all directions is denoted by

$$\boldsymbol{T}^{(n)}(k_n) := (T^{(n)}_1(k_n), \dots, T^{(n)}_{2^{d_n}-1}(k_n))^\top.$$

In the following, we fix k_n and estimate s^*, the number of relevant directions using some information criteria.

For this, we make the following assumptions.

Assumption E.

(E1) Suppose that all relevant directions $\beta \in \mathcal{P}^*_{\mathbb{N}}$ satisfy $|\beta| < \infty$ and for the number of relevant directions holds

$$s^* < \infty.$$

(E2) As $n \to \infty$ holds $\widehat{s}_n \xrightarrow{\mathbb{P}} \infty$ and $\widehat{s}_n/k_n \xrightarrow{\mathbb{P}} c \in [0, 1)$.

(E3) Suppose $j \leq s^*$. Then

$$\frac{T_j^{(n)}(k_n)}{k_n} \xrightarrow{\mathbb{P}} p'_j \in (0,1).$$

(E4) Suppose there exists a $\mu \geq 1$ such that

$$\frac{1}{\widehat{s}_n} \sum_{j=s^*+1}^{\widehat{s}_n} T_j^{(n)}(k_n) \xrightarrow{\mathbb{P}} \mu.$$

(E5) Suppose there exists a $q \geq 1$ such that

$$T_{s^*+1}^{(n)}(k_n) \xrightarrow{\mathbb{P}} q\mu.$$

Remark 3.36.

(a) Similarly to the fixed-dimensional case (cf. Section 3.1 and Meyer and Wintenberger, 2023), we assume by Assumption (E1) that there is a fixed number of relevant directions. Furthermore, since the relevant directions are finite, we are able to observe relevant directions empirically.

(b) Note, since $\widehat{s}_n \leq \sum_{j=1}^{\widehat{s}_n} T_j^{(n)}(k_n) \leq k_n$, we get that $\widehat{s}_n/k_n$ is bounded and then Assumption (E2) guarantees that the ratio converges. Due to this assumption, we have a clear differentiation from the fixed-dimensional setting, where $\widehat{s}_n$ is assumed to be fixed.

(c) The motivation for Assumption (E3) is based on the case for finite d, where $T^{(n)}(\beta, k_n) \xrightarrow{\mathbb{P}} p(C_\beta)$ for $\beta \in \mathcal{P}_d^*$, by Proposition 3.4. However, since $c > 0$ is possible, p'_j may not be equal to $p(C_\beta)$ for the corresponding β to $T_j^{(n)}(k_n)$.

(d) If $c > 0$ then by Assumption (E2) as $n \to \infty$

$$1 \leq \frac{1}{\widehat{s}_n} \sum_{j=s^*+1}^{\widehat{s}_n} T_j^{(n)}(k_n) \leq \frac{k_n}{\widehat{s}_n} \xrightarrow{\mathbb{P}} \frac{1}{c} \in (1, \infty).$$

Since the sequence of the weighted sum is bounded, we assume (E4).

(e) Assumption (E5) guarantees that there are no bias directions of type (i). If $T_{s^*+1}^{(n)}(k_n)$ diverges, we cannot expect the information criteria to be consistent as presented later (cf. Theorem 3.53).

(f) If $c > 0$ and $p' \coloneqq \sum_{j=1}^{s^*} p'_j$, then $\mu = (1 - p')/c$.

3.3.2. EMPIRICAL OBSERVATIONS

For an initial analysis of the behavior of the bias directions and the number of observed directions $\widehat{s}_n$, we conduct a simulation study. We run each of the following settings with 100 repetitions, where we consider d_n-dimensional i.i.d. random vectors whose spectral measure only concentrates on the first s^* axes and fill the remaining entries with i.i.d. random variables, which act as noise. To define their distribution, we assume that $\boldsymbol{H} = (h_{ij})_{1\leq i,j\leq s^*} \in \mathbb{R}^{s^*\times s^*}$ with $h_{ij} \overset{\text{i.i.d.}}{\sim} \mathcal{U}((0,1))$ and

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

Note that $\boldsymbol{\Sigma}_{ii} = 1$, $i = 1,\dots,s^*$ and $\boldsymbol{\Sigma}_{ij} < 1$, $i \neq j$. Suppose now $\boldsymbol{Y} = (Y_1,\dots,Y_{s^*}) \sim \mathcal{N}_{s^*}(\boldsymbol{0}_{s^*}, \boldsymbol{\Sigma})$ 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 (Corollary 5.28 Resnick, 1987). Now, let $\boldsymbol{Y}^1,\dots,\boldsymbol{Y}^n, i = 1,\dots,n$ be an i.i.d. sequence of random vectors with distribution $\boldsymbol{Y}$ and $U_1^i,\dots,U_{d_n-s^*}^i$ be the absolute value of an independent sequence of independent normally distributed random variables. Then we define the i.i.d. random vectors $\boldsymbol{X}^i = (X_1^i,\dots,X_{d_n}^i)^\top \in \mathbb{R}_+^{d_n}$, $i = 1,\dots,n$, as

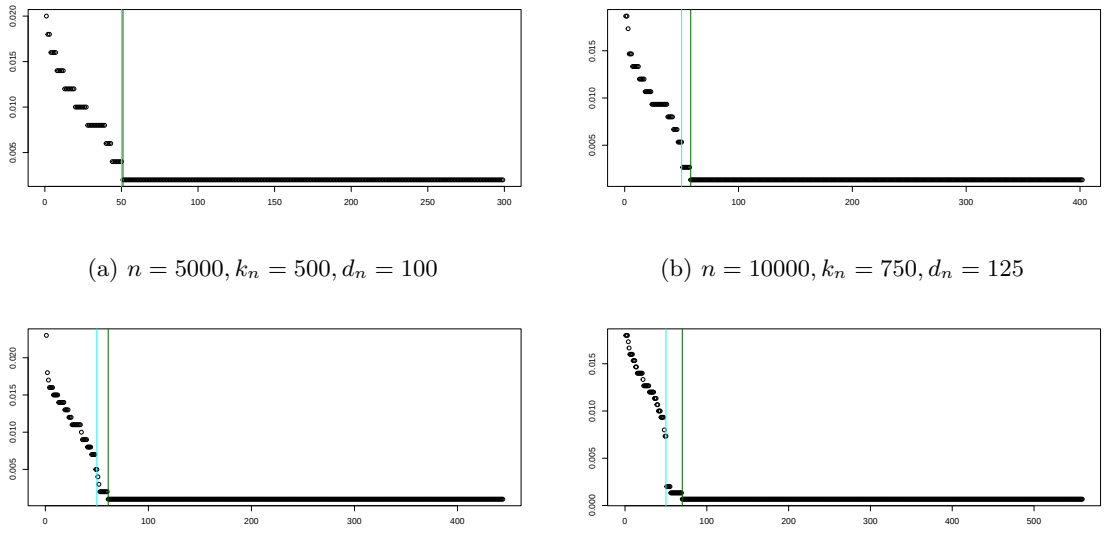
$$X_j^i \coloneqq \begin{cases} \frac{1}{1-\Phi(Y_j^i)}, & 1 \leq j \leq s^*, \\ U_{j-s^*}^i, & s^*+1 \leq j \leq d_n. \end{cases}$$

In the simulation, the ratio k_n/n is decreasing and we consider $s^* = 50$.

n	k_n	d_n	$\widehat{s}_n$	$\widehat{s}_n/k_n$
5000	500	100	309.99	0.62
10000	750	125	407.5	0.54
20000	1000	150	454.32	0.45
40000	1500	175	582.37	0.39

Table 3.1.: *Average of the realizations of $\widehat{s}_n$ and $\widehat{s}_n/k_n$ over 100 realizations.*

In Figure 3.1 are the entries of $\boldsymbol{T}^{(n)}(k_n)/k_n$ mapped for different values of n and d. We observe that the number of observed directions increases, but the number of occurrences for each direction is nearly constant to 1 for the last directions. From Table 3.1 it is evident, while k_n increases also $\widehat{s}_n$ increases nearly proportionally. Another observation from the plot is that not only the number of entries after the green line increases but also the number of entries between the cyan line and the green line. In addition, the gap between the smallest relevant index and the first bias index widens as n increases, indicating that the order of the number of observations differs.



(a) $n = 5000, k_n = 500, d_n = 100$

(b) $n = 10000, k_n = 750, d_n = 125$

(c) $n = 20000, k_n = 1000, d_n = 150$

(d) $n = 40000, k_n = 1500, d_n = 175$

Figure 3.1.: *Mapping of one realization of the entries of* $\boldsymbol{T}^{(n)}(k_n)/k_n$*. The cyan line indicates the* 50*-th entry and the green line the first entry of* $\boldsymbol{T}^{(n)}(k_n)$ *equal to* 1*. In this example,* 50 *directions are relevant.*

3.3.3. Information criteria in the high-dimensional case

In the following, we introduce the information criteria for the high-dimensional case and present the conditions for consistency. In the high-dimensional case, we do not evaluate all possible models, but restrict the set of candidate models by q_n with $q_n \leq 2^{d_n} - 1$. For an information criterion IC, an estimator for s^* is

$$\widehat{s}_n^* := \underset{1 \leq s \leq q_n}{\arg\min}\, IC_{k_n}(s).$$

Akaike information criterion

We recall that the Akaike information criterion for the fixed-dimensional case introduced by Meyer and Wintenberger (2023) (cf. (3.8)) is motivated by minimizing the expected Kullback-Leibler (KL) divergence between the true distribution of $\boldsymbol{T}^{(n)}(k_n)$ and a multinomial distribution. In the following, the definition of the AIC is adopted from the fixed-dimensional case and adjusted to the high-dimensional case by exchanging r with $\widehat{s}_n$. The *Akaike information criterion* (AIC°) for the high-dimensional case is defined as

$$\begin{aligned}\mathrm{AIC}_{k_n}^{\circ}(s) := &- \log(k_n!) + \sum_{j=1}^{\widehat{s}_n} \log(T_j^{(n)}(k_n)!) - \sum_{j=1}^{s} T_j^{(n)}(k_n) \log\left(\frac{T_j^{(n)}(k_n)}{k_n}\right) \\ &- \log\left(\frac{1}{k_n(\widehat{s}_n - s)} \sum_{j=s+1}^{\widehat{s}_n} T_j^{(n)}(k_n)\right) \sum_{j=s+1}^{\widehat{s}_n} T_j^{(n)}(k_n) + s,\ s = 1, \ldots, q_n.\end{aligned}$$

In Section 3.2.1 we showed that the AIC is not consistent for fixed d, which is a typical property of an AIC in the large sample size and fixed dimensional case (cf. Remark 2.17). In the high-dimensional case, the AIC is consistent as shown in Bai et al. (2018) or later in Chapter 4. In the present setting, we can also establish the consistency of the AIC, provided that there is no bias direction of type (i).

Theorem 3.37. *Suppose Assumption E holds and let* $q_n/\sqrt{\widehat{s}_n} = o_{\mathbb{P}}(1)$. *Then, the AIC is weakly consistent, i.e.*

$$\lim_{n\to\infty} \mathbb{P}\Big(\operatorname*{arg\,min}_{1\le s<q_n} \mathrm{AIC}^{\circ}_{k_n}(s) = s^* \Big) = 1$$

if and only if

$$g_{\mathrm{AIC}^{\circ}}(q,\mu) \coloneqq q(1-\log(q)) - 1 + \frac{1}{\mu} > 0.$$

Remark 3.38. Since in Figure 3.1 the entries of $\boldsymbol{T}^{(n)}(k_n)$ between the cyan line and the green line do not appear to differ by a large margin from the values after the green line, we only consider bias directions of type (ii) in Theorem 3.37 and assume that the set of bias directions of type (i) is empty by imposing Assumption (E5). This setting is also in line with the consistency results derived in Bai et al. (2018), where the strength of the noise, i.e. the value of the non-spiked eigenvalues in the mentioned paper, is assumed to be constant. If bias directions of type (i) are present, we have shown in Section 3.4.4 that the AIC is not consistent.

QUASI-AKAIKE INFORMATION CRITERION

The quasi-Akaike information criterion from Section 3.2.1 is inspired by the Akaike information criterion by minimizing the KL divergence. However, in contrast to the approach of Meyer and Wintenberger (2023), which is based on the likelihood function of a multinomial distribution, the approach in Section 3.2.1 uses the Gaussian distribution for the derivation of the QAIC. Similarly to the AIC, the QAIC is adapted to the high-dimensional setting as follows.

For the number of extreme directions s with fixed k_n the *quasi-Akaike information criterion* (QAIC$^{\circ}$) for the high-dimensional case is defined as

$$\mathrm{QAIC}^{\circ}_{k_n}(s) \coloneqq \widehat{s}_n \log(2\pi) + \widehat{s}_n \log(k_n) + \sum_{j=1}^{s} \log\Big(\frac{T_j^{(n)}(k_n)}{k_n}\Big) + (\widehat{s}_n - s)\log(\widehat{\rho}_n^s) + \widehat{s}_n + s$$

for $s = 1, \dots, q_n$, where $\widehat{\rho}_n^s \coloneqq 1/(\widehat{s}_n - s)\sum_{j=s+1}^{\widehat{s}_n} T_j^{(n)}(k_n)/k_n$.

In the fixed-dimensional case the QAIC is consistent (cf. Section 3.2.1) and in the high-dimensional case we also have the consistency of the QAIC, as shown in the next theorem.

Theorem 3.39. *Suppose Assumption E holds and let* $q_n/\sqrt{\widehat{s}_n} = o_{\mathbb{P}}(1)$. *Then the* QAIC$^\circ$ *is consistent, i.e.*

$$\lim_{n\to\infty} \mathbb{P}\Big(\operatorname*{arg\,min}_{1\leq s<q_n} \mathrm{QAIC}^\circ_{k_n}(s) = s^* \Big) = 1$$

if and only if

$$g_{\mathrm{QAIC}^\circ}(q) \coloneqq \log(q) - q + 2 > 0.$$

The condition $\log(q) - q + 2 > 0$ is similar to the condition for the AIC$^\circ$ in Theorem 3.37, where q, the spread of the noise, is not allowed to be too large.

Mean squared error information criterion

The basic idea of the AIC is to minimize the Kullback-Leibler distance of the true distribution and a parametric family of distributions (cf. Section 2.3.1). This minimum is approximated by the expected Kullback-Leibler distance of the true distribution and the estimated distribution. In Section 3.2.2 the same ideas were used but instead of using the KL distance, we used the normalized mean squared error (MSE) of the parameter estimator to derive the MSEIC. The mean squared error information criterion (MSEIC$^\circ$) for the high-dimensional case is defined as

$$\mathrm{MSEIC}^\circ_{k_n}(s) \coloneqq \frac{k_n}{\sum_{l=s+1}^{\widehat{s}_n} \frac{T_l^{(n)}(k_n)}{k_n(\widehat{s}_n - s)}} \sum_{j=s+1}^{\widehat{s}_n} \left(\frac{T_j^{(n)}(k_n)}{k_n} - \sum_{i=s+1}^{\widehat{s}_n} \frac{T_i^{(n)}(k_n)}{k_n(\widehat{s}_n - s)} \right)^2 + 2s,$$

for $s = 1, \ldots, q_n$.

Similarly to the AIC, we showed in Section 3.2.2 that the MSEIC is not consistent in the fixed-dimensional case but it is in the high-dimensional case.

Theorem 3.40. *Suppose Assumption E holds and let* $q_n/\sqrt{\widehat{s}_n} = o_{\mathbb{P}}(1)$. *If*

$$g_{\mathrm{MSEIC}^\circ}(q, \mu) \coloneqq 2 - (q-1)^2\mu > 0$$

is satisfied, then

$$\lim_{n\to\infty} \mathbb{P}\Big(\operatorname*{arg\,min}_{1\leq s<q_n} \mathrm{MSEIC}^\circ_{k_n}(s) = s^* \Big) = 1.$$

Bayesian information criterion

The idea of the BIC is to determine the Model s with the highest posterior probability $\mathbb{P}(s|\boldsymbol{T}^{(n)}(k_n))$ given the data $\boldsymbol{T}^{(n)}(k_n)$ for $s = 1, \ldots, \widehat{s}_n$ (cf. Section 2.3.2). In Section 3.2.3 we derived the BIC by bounding the posterior probability as in Cavanaugh and Neath (1999). The *Bayesian information criterion concerning the upper bound* (BICU$^\circ$) for the

high-dimensional case is defined as

$$\begin{aligned}\mathrm{BICU}^{\circ}_{k_n}(s) \coloneqq &-2\log(k_n!) + 2\sum_{j=1}^{\widehat{s}_n}\log(T_j^{(n)}(k_n)!) - 2\sum_{j=1}^{s}T_j^{(n)}(k_n)\log\left(\frac{T_j^{(n)}(k_n)}{k_n}\right)\\ &- 2\log\left(\frac{1}{k_n(\widehat{s}_n - s)}\sum_{j=s+1}^{\widehat{s}_n}T_j^{(n)}(k_n)\right)\sum_{j=s+1}^{\widehat{s}_n}T_j^{(n)}(k_n) + 2s\log(k_n)\\ &+ s\log\left(\frac{\widehat{s}_n}{2\pi(\widehat{s}_n - s)}\right)\end{aligned}$$

for $s = 1, \dots, q_n$ and the *Bayesian information criterion concerning the lower bound* (BICL$^{\circ}$) for the high-dimensional case is defined as

$$\begin{aligned}\mathrm{BICL}^{\circ}_{k_n}(s) \coloneqq &-2\log(k_n!) + 2\sum_{j=1}^{\widehat{s}_n}\log(T_j^{(n)}(k_n)!) - 2\sum_{j=1}^{s}T_j^{(n)}(k_n)\log\left(\frac{T_j^{(n)}(k_n)}{k_n}\right)\\ &- 2\log\left(\frac{1}{k_n(\widehat{s}_n - s)}\sum_{j=s+1}^{\widehat{s}_n}T_j^{(n)}(k_n)\right)\sum_{j=s+1}^{\widehat{s}_n}T_j^{(n)}(k_n) + s\log(k_n)\\ &+ s\log\left(\frac{k_n}{2\pi T_1^{(n)}(k_n)}\right)\end{aligned}$$

for $s = 1, \dots, q_n$.

In the fixed-dimensional case the BIC is consistent (cf. Section 3.2.3) and we can also show that it is consistent in the high-dimensional setting.

Theorem 3.41. *Suppose Assumption E holds and let* $q_n/\sqrt{\widehat{s}_n} = o_{\mathbb{P}}(1)$. *Then holds that*

(a) $\lim_{n\to\infty}\mathbb{P}\Big(\operatorname*{arg\,min}_{1\le s<q_n}\mathrm{BICU}^{\circ}_{k_n}(s) = s^*\Big) = 1,$

(b) $\lim_{n\to\infty}\mathbb{P}\Big(\operatorname*{arg\,min}_{1\le s<q_n}\mathrm{BICL}^{\circ}_{k_n}(s) = s^*\Big) = 1.$

This result is also in line with the consistency results in Bai et al. (2018), where the BIC is consistent in the high-dimensional setting when the strength of the signal (i.e. the leading eigenvalues or in our case $T_1^{(n)}(k_n), \dots, T_{s^*}^{(n)}(k_n)$) go to infinity.

Remark 3.42. Note that the numerical implementation of the information criteria derived in Section 3.2 for the local model and in this section does not differ, therefore we denote the information criteria for the simulation study only by AIC, BIC, QAIC and MSEIC.

COMPARISON OF POSSIBLE VALUES OF q AND μ FOR THE CONSISTENCY

In the following, we plot $g_{\mathrm{AIC}^{\circ}}(q,\mu), g_{\mathrm{QAIC}^{\circ}}(q)$ and $g_{\mathrm{MSEIC}^{\circ}}(q,\mu)$ as functions in q for fixed μ. If the value of one of the functions is positive, then (q,μ) or q is a feasible point for the consistency of the AIC$^{\circ}$, QAIC$^{\circ}$ or MSEIC$^{\circ}$, respectively.

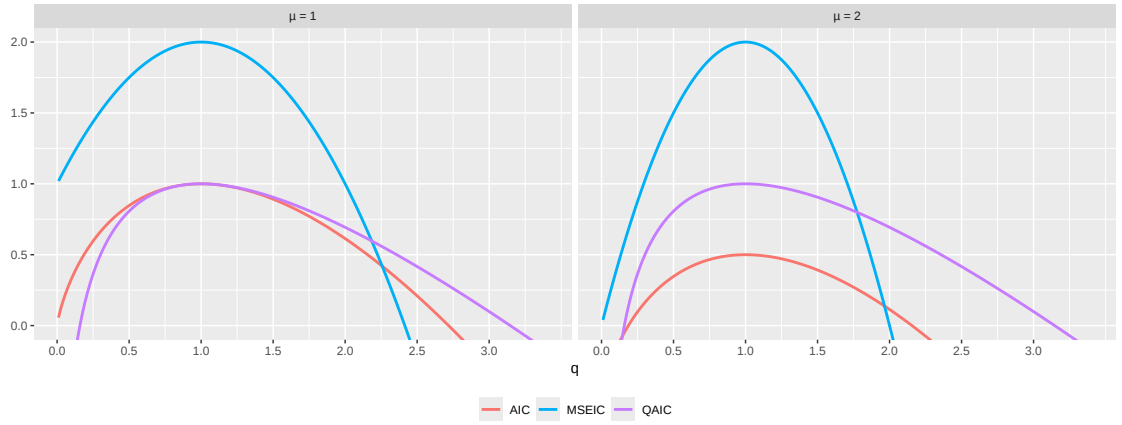


Figure 3.2.: *Plot of the functions in q for the consistency of $g_{\mathrm{AIC}^\circ}(q,\mu)$, $g_{\mathrm{QAIC}^\circ}(q)$ and $g_{\mathrm{MSEIC}^\circ}(q,\mu)$ for $\mu = 1$ and $\mu = 2$.*

In Figure 3.2 the functions for the consistency of the information criteria are displayed as functions of q for $\mu = 1$ and $\mu = 2$. The QAIC$^\circ$ has the largest range of positive values. Additionally, the function of the QAIC$^\circ$ does not depend on μ, while the range for the consistency of the AIC$^\circ$ and MSEIC$^\circ$ depends on the value of μ. As μ increases, the interval in which the functions for the AIC$^\circ$ and MSEIC$^\circ$ are positive shrinks. Note that the height of the functions has no influence on the theoretical consistency results. For the consistency, it is only necessary that the functions are positive.

3.4. Proofs

3.4.1. Proofs of Section 3.2.1

Proof of Theorem 3.7

Proof of Theorem 3.7.

Step 1: Suppose $s > s^*$. By the definition of the AIC and the log-likelihood function in (3.3) it follows that

$$\begin{aligned}
&\mathrm{AIC}_{k_n}(s) - \mathrm{AIC}_{k_n}(s^*) \\
&\quad = -\log L_{M^s_{k_n}}(\widehat{\boldsymbol{p}}^{\,s}_n \,|\, \boldsymbol{T}_n(k_n)) + s + \log L_{M^{s^*}_{k_n}}(\widehat{\boldsymbol{p}}^{\,s^*}_n \,|\, \boldsymbol{T}_n(k_n)) - s^* \\
&\quad = -\sum_{j=s^*+1}^{s} T_{n,j}(k_n) \log\left(\frac{T_{n,j}(k_n)}{k_n}\right) - \log\left(\frac{1}{r-s}\sum_{j=s+1}^{r}\frac{T_{n,j}(k_n)}{k_n}\right)\sum_{i=s+1}^{r} T_{n,i}(k_n) \\
&\qquad + \log\left(\frac{1}{r-s^*}\sum_{j=s^*+1}^{r}\frac{T_{n,j}(k_n)}{k_n}\right)\sum_{i=s^*+1}^{r} T_{n,i}(k_n) + (s-s^*), \qquad (3.15)
\end{aligned}$$

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

$$\begin{aligned}
&\mathrm{AIC}_{k_n}(s) - \mathrm{AIC}_{k_n}(s^*) \\
&\quad = -\sum_{j=s^*+1}^{s} (k_n\rho_n + \sqrt{k_n\rho_n}Y_{n,j}) \log\left(1 + \frac{1}{\sqrt{k_n\rho_n}}Y_{n,j}\right) \\
&\qquad - \log\left(1 + \frac{1}{r-s}\sum_{j=s+1}^{r}\frac{1}{\sqrt{k_n\rho_n}}Y_{n,j}\right)\sum_{i=s+1}^{r}(k_n\rho_n + \sqrt{k_n\rho_n}Y_{n,i}) \\
&\qquad + \log\left(1 + \frac{1}{r-s^*}\sum_{j=s^*+1}^{r}\frac{1}{\sqrt{k_n\rho_n}}Y_{n,j}\right)\sum_{i=s^*+1}^{r}(k_n\rho_n + \sqrt{k_n\rho_n}Y_{n,i}) \\
&\qquad + (s - s^*).
\end{aligned} \tag{3.16}$$

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

$$(Y_{n,s^*+1}, \dots, Y_{n,r}) \xrightarrow{\mathcal{D}} (Y_{s^*+1}, \dots, Y_r) =: \boldsymbol{Y} \sim \mathcal{N}_{r-s^*}(\mathbf{0}_{r-s}, \boldsymbol{I}_{r-s^*}), \quad n \to \infty, \tag{3.17}$$

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 (3.16) such that

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

$$+\Big(\frac{1}{r-s^*}\sum_{j=s^*+1}^{r}\frac{1}{\sqrt{k_n\rho_n}}Y_{n,j}-\frac{1}{2}\Big(\frac{1}{r-s^*}\sum_{j=s^*+1}^{r}\frac{1}{\sqrt{k_n\rho_n}}Y_{n,j}\Big)^2\Big)\cdot\sum_{i=s^*+1}^{r}(k_n\rho_n+\sqrt{k_n\rho_n}Y_{n,i})$$
$$+(s-s^*)+O_{\mathbb{P}}((k_n\rho_n)^{-1/2}).$$

Since $k_n\rho_n\to\infty$ (Lemma 3.6(a)) we receive

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

Due to (3.17) and the continuous mapping theorem we finally obtain as $n\to\infty$,

$$\mathrm{AIC}_{k_n}(s)-\mathrm{AIC}_{k_n}(s^*)\xrightarrow{\mathcal{D}}-\frac{1}{2}\sum_{j=s^*+1}^{s}Y_j^2-\frac{1}{2(r-s)}\Big(\sum_{j=s+1}^{r}Y_j\Big)^2+\frac{1}{2(r-s^*)}\Big(\sum_{j=s^*+1}^{r}Y_j\Big)^2+(s-s^*).\tag{3.18}$$

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 (3.15) that

$$\begin{aligned}&\mathrm{AIC}_{k_n}(s)-\mathrm{AIC}_{k_n}(s^*)\\&\quad=\sum_{j=s+1}^{s^*}T_{n,j}(k_n)\log\left(\frac{T_{n,j}(k_n)}{k_n}\right)-\log\left(\frac{1}{r-s}\sum_{j=s+1}^{r}\frac{T_{n,j}(k_n)}{k_n}\right)\sum_{i=s+1}^{r}T_{n,i}(k_n)\\&\qquad+\log\left(\frac{1}{r-s^*}\sum_{j=s^*+1}^{r}\frac{T_{n,j}(k_n)}{k_n}\right)\sum_{i=s^*+1}^{r}T_{n,i}(k_n)+(s-s^*).\end{aligned}\tag{3.19}$$

A direct consequence of $T_{n,j}(k_n)/k_n\xrightarrow{\mathbb{P}}0$ for $j=s^*+1,\dots,r$ (Lemma 3.6(c)) and $\lim_{x\to0}x\log(x)=0$ 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 3.6(b) yields $T_{n,j}(k_n)/k_n\xrightarrow{\mathbb{P}}p_j>0$ for $j=1,\dots,s^*$ and thus, as

$n \to \infty$,

$$\sum_{i=s+1}^{s^*} \frac{T_{n,i}(k_n)}{k_n} \log\left(\frac{T_{n,i}(k_n)}{k_n}\right) - \log\left(\frac{1}{r-s}\sum_{j=s+1}^{r} \frac{T_{n,j}(k_n)}{k_n}\right) \sum_{i=s+1}^{r} \frac{T_{n,i}(k_n)}{k_n}$$
$$\xrightarrow{\mathcal{D}} \sum_{i=s+1}^{s^*} p_i \left(\log(p_i) - \log\left(\frac{1}{r-s}\sum_{j=s+1}^{s^*} p_j\right)\right), \tag{3.20}$$

while we used $p_i = 0$ for $s^* \leq i \leq r$. Next, we apply the log sum inequality (Cover, 2006, Theorem 2.7.1) to the limit of (3.20) and receive

$$\sum_{i=s+1}^{s^*} p_i \left(\log(p_i) - \log\left(\frac{1}{r-s}\sum_{j=s+1}^{s^*} p_j\right)\right) = \sum_{i=s+1}^{s^*} p_i \log\left(\frac{p_i}{\frac{1}{r-s}\sum_{j=s+1}^{s^*} p_j}\right)$$
$$\geq \left(\sum_{i=s+1}^{s^*} p_i\right) \log\left(\frac{\sum_{i=s+1}^{s^*} p_i}{\frac{s^*-s}{r-s}\sum_{j=s+1}^{s^*} p_j}\right) = \left(\sum_{i=s+1}^{s^*} p_i\right) \log\left(\frac{r-s}{s^*-s}\right) > 0, \tag{3.21}$$

since $r > s^*$. Dividing (3.19) by k_n and using (3.20) and (3.21) gives

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

and thus, the assertion follows. □

PROOF OF PROPOSITION 3.11

Before we are able to present the proof of Proposition 3.11 we require some auxiliary lemmata whose proofs are moved to Appendix A.1 of the Supplementary Material. In the following, we work with the r-dimensional multivariate normal distribution $\mathcal{N}_r(k_n \boldsymbol{B}_s(\underline{\widetilde{\boldsymbol{p}}}^s), k_n \operatorname{diag}(\boldsymbol{B}_s(\underline{\widetilde{\boldsymbol{p}}}^s)))$, $\underline{\widetilde{\boldsymbol{p}}}^s \in \mathbb{R}_+^{s+1}$, which has a negative log-likelihood function

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

Lemma 3.43. *Suppose the assumptions of Proposition 3.11 hold and $\underline{\widehat{\boldsymbol{p}}}_n^s(\boldsymbol{\mathcal{T}}_n)$ is defined analog to $\underline{\widehat{\boldsymbol{p}}}_n^s(\widetilde{\boldsymbol{\mathcal{T}}}_n)$ in (3.9). Then as $n \to \infty$,*

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

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

where

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

Lemma 3.44. *Suppose the assumptions of Proposition 3.11 hold and* $\widehat{\underline{\boldsymbol{p}}}_n^s(\boldsymbol{\mathcal{T}}_n)$ *is defined analog to* $\widehat{\underline{\boldsymbol{p}}}_n^s(\widetilde{\boldsymbol{\mathcal{T}}}_n)$ *in* (3.9).

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

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

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

$$\|\bar{\boldsymbol{p}}_n - \widehat{\underline{\boldsymbol{p}}}_n^s(\boldsymbol{\mathcal{T}}_n)\|_1 \le \|\widehat{\underline{\boldsymbol{p}}}_n^s(\widetilde{\boldsymbol{\mathcal{T}}}_n) - \widehat{\underline{\boldsymbol{p}}}_n^s(\boldsymbol{\mathcal{T}}_n)\|_1, \quad n \in \mathbb{N}.$$

Then as $n \to \infty$,

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

Proof of Proposition 3.11. Using a Taylor expansion of $\log L_{\mathcal{N}_r}(\widehat{\underline{\boldsymbol{p}}}_n^s(\widetilde{\boldsymbol{\mathcal{T}}}_n) \,|\, \boldsymbol{\mathcal{T}}_n)$ around $\widehat{\underline{\boldsymbol{p}}}_n^s(\boldsymbol{\mathcal{T}}_n)$ yields the existence of a random vector $\bar{\boldsymbol{p}}_n \coloneqq (\bar{p}_{n,1}, \ldots, \bar{p}_{n,s}, \bar{\rho}_n)^\top$ with

$$\|\bar{\boldsymbol{p}}_n - \widehat{\underline{\boldsymbol{p}}}_n^s(\boldsymbol{\mathcal{T}}_n)\|_1 \le \|\widehat{\underline{\boldsymbol{p}}}_n^s(\widetilde{\boldsymbol{\mathcal{T}}}_n) - \widehat{\underline{\boldsymbol{p}}}_n^s(\boldsymbol{\mathcal{T}}_n)\|_1$$

such that

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

Applying Lemma 3.44 (b) gives

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

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

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

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 3.6 (c) and the assumption $s \geq s^*$) and $\boldsymbol{Y}_n$ as defined in Lemma 3.43, which result in

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

by Lemma 3.43, where $\boldsymbol{Y} \sim \mathcal{N}_{r+1}(\boldsymbol{0}_{r+1}, \Sigma)$. Since

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

the assertion follows.

□

Proof of Theorem 3.14

Proof of Theorem 3.14.

Step 1: Suppose $s < s^*$. We have $\widehat{p}_{n,j}^s = \widehat{p}_{n,j}^{s^*}$ for $j = 1, \dots, s$ and due to Lemma 3.6 (b,c) we have as $n \to \infty$,

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

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 $n \to \infty$,

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

Step 2: Suppose $s > s^*$. In this case, we have by Lemma 3.6 (b,c) that

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

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 $j = s^*+1, \ldots, s$, $\log(\widehat{\rho}_n^{s^*}/\rho_n) \xrightarrow{\mathbb{P}} 0$ 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. □

3.4.2. Proofs of Section 3.2.2

Proof of Theorem 3.18

The proof of Theorem 3.18 is similar to the proof of Proposition 3.11. In the first step, we start to calculate the Jacobian vector of $\ell^2(\underline{\widetilde{p}}^s | \boldsymbol{T}_n(k_n))$ for $\underline{\widetilde{p}}^s = (\underline{\widetilde{p}}_1^s, \ldots \underline{\widetilde{p}}_s^s, \underline{\widetilde{\rho}}^s) \in \mathbb{R}_+^{s+1}$, which is

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

and the Hessian matrix is

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

Analog to Lemma 3.43 and Lemma 3.44 we get the following results.

Lemma 3.45. *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 (3.9). Then as $n \to \infty$,*

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

Lemma 3.46. *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 (3.9).*

(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

$$\|\bar{\boldsymbol{p}}_n - \widehat{\underline{\boldsymbol{p}}}_n^s(\boldsymbol{T}_n(k_n))\|_1 \leq \|\widehat{\underline{\boldsymbol{p}}}_n^s(\widetilde{\boldsymbol{T}}_n(k_n)) - \widehat{\underline{\boldsymbol{p}}}_n^s(\boldsymbol{T}_n(k_n))\|_1, \quad n \in \mathbb{N}.$$

Then as $n \to \infty$,

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

Proof of Theorem 3.18. Using a Taylor expansion of $\ell^2(\widehat{\underline{\boldsymbol{p}}}_n^s(\widetilde{\boldsymbol{T}}_n(k_n)) \,|\, \boldsymbol{T}_n(k_n))$ around $\widehat{\underline{\boldsymbol{p}}}_n^s(\boldsymbol{T}_n(k_n))$ 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

$$\|\bar{\boldsymbol{p}}_n - \widehat{\underline{\boldsymbol{p}}}_n^s(\boldsymbol{T}_n(k_n))\|_1 \leq \|\widehat{\underline{\boldsymbol{p}}}_n^s(\widetilde{\boldsymbol{T}}_n(k_n)) - \widehat{\underline{\boldsymbol{p}}}_n^s(\boldsymbol{T}_n(k_n))\|_1$$

such that

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

Applying Lemma 3.46 gives

$$\ell^2(\widehat{\underline{\boldsymbol{p}}}_n^s(\widetilde{\boldsymbol{T}}_n(k_n)) \,|\, \boldsymbol{T}_n(k_n))$$

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

Next, we move some terms on the right-hand side and use Lemma 3.45, which result in

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

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. □

Proof of Theorem 3.20

Proof of Theorem 3.20.

Step 1: Suppose $s < s^*$. An application of Lemma 3.6(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,

$$
\frac{k_n}{\sum_{l=s+1}^{r} \frac{T_{n,l}(k_n)}{k_n(r-s)}} \sum_{j=s+1}^{r} \left(\frac{T_{n,j}(k_n)}{k_n} - \sum_{i=s+1}^{r} \frac{T_{n,i}(k_n)}{k_n(r-s)} \right)^2 \xrightarrow{\mathbb{P}} \infty. \tag{3.22}
$$

On the other hand, define

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

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 $l = s^*+1, \ldots, r$ by Lemma 3.6(c), it follows that

$$\begin{aligned}
&\frac{\rho_n}{\sum_{l=s^*+1}^r \frac{T_{n,l}(k_n)}{k_n(r-s^*)}} \frac{k_n}{\rho_n} \sum_{j=s^*+1}^r \left(\frac{T_{n,j}(k_n)}{k_n} - \sum_{i=s^*+1}^r \frac{T_{n,i}(k_n)}{k_n(r-s^*)} \right)^2 \\
&= \underbrace{\frac{\rho_n}{\sum_{l=s^*+1}^r \frac{T_{n,l}(k_n)}{k_n(r-s^*)}}}_{\xrightarrow{\mathbb{P}} 1} \underbrace{\boldsymbol{V}_n^\top (\boldsymbol{I}_{r-s^*} - \frac{1}{r-s^*} \mathbf{1}_{r-s^*} \mathbf{1}_{r-s^*}^\top)^\top (\boldsymbol{I}_{r-s^*} - \frac{1}{r-s^*} \mathbf{1}_{r-s^*} \mathbf{1}_{r-s^*}^\top) \boldsymbol{V}_n}_{\xrightarrow{\mathcal{D}} \chi^2_{r-s^*-1}} \\
&\xrightarrow{\mathcal{D}} \chi^2_{r-s^*-1} = O_{\mathbb{P}}(1). \qquad (3.23)
\end{aligned}$$

Combining (3.22) and (3.23) yields

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

Step 2: Suppose $s > s^*$. An application of (3.23) and Lemma 3.6(c) yield

$$\begin{aligned}
&\frac{k_n}{\sum_{l=s^*+1}^r \frac{T_{n,l}(k_n)}{k_n(r-s^*)}} \sum_{j=s^*+1}^r \left(\frac{T_{n,j}(k_n)}{k_n} - \sum_{i=s^*+1}^r \frac{T_{n,i}(k_n)}{k_n(r-s^*)} \right)^2 \\
&\qquad - \frac{k_n}{\rho_n} \sum_{j=s^*+1}^r \left(\frac{T_{n,j}(k_n)}{k_n} - \sum_{i=s^*+1}^r \frac{T_{n,i}(k_n)}{k_n(r-s^*)} \right)^2 \\
&= \underbrace{\left(\frac{\rho_n}{\sum_{l=s^*+1}^r \frac{T_{n,l}(k_n)}{k_n(r-s^*)}} - 1 \right)}_{\xrightarrow{\mathbb{P}} 0} \underbrace{\frac{k_n}{\rho_n} \sum_{j=s^*+1}^r \left(\frac{T_{n,j}(k_n)}{k_n} - \sum_{i=s^*+1}^r \frac{T_{n,i}(k_n)}{k_n(r-s^*)} \right)^2}_{\xrightarrow{\mathcal{D}} \chi^2_{r-s^*-1} \text{ by (3.23)}} \\
&= o_{\mathbb{P}}(1). \qquad (3.24)
\end{aligned}$$

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 (3.24) that

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

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

Similar to the proof of Theorem 3.7, there exists a positive probability that the right-hand side is positive. Hence, the assertion follows. □

Proof of Theorem 3.21

Before we are able to present the proof of Theorem 3.21 we require some auxiliary lemmata whose proofs are moved to Appendix A.2 in the Supplementary Material.

Lemma 3.47. *Suppose Assumptions (B1) and (B2) hold. Then for $\boldsymbol{p}' \in \mathbb{R}_+^r$ the asymptotic behavior*

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

as $n \to \infty$ holds.

Lemma 3.48. *For $q' \in (0,1)$ the equality*

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

holds.

Proof of Theorem 3.21. For $q' \in (0,1)$ and $\boldsymbol{p}' \in \mathbb{R}_+^r$ we have as a consequence of Lem-

mas 3.47 and 3.48, that

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

Therefore, it follows that

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

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

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

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.

□

3.4.3. Proofs of Section 3.2.3

Proof of Theorem 3.25

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 3.49, 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 3.50, we present boundaries for the eigenvalues of the Hessian of the log-likelihood function; the proofs of these auxiliary results are included in Appendix A.3.1 of the Supplementary Material. Finally, for the proof of the upper bound of the log-posterior distribution in Theorem 3.25 we combine these two results.

Lemma 3.49. *Let the assumptions of Theorem 3.25 hold. Define the ball*

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

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

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

Lemma 3.50. *Let the assumptions of Theorem 3.25 hold. Define* $\lambda_{n,2} \coloneqq k_n/T_{n,1}(k_n)$ *and* $\lambda_{n,1} \coloneqq k_n/T_{n,s}(k_n) + sk_n/\sum_{j=s+1}^r T_{n,j}$. *For* $\widetilde{\boldsymbol{p}}^s \in \Theta_s$ *we have on the one hand,*

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

and on the other hand,

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

Proof of Theorem 3.25. In the following let $\gamma = 4/3$ and $\varepsilon_n \coloneqq \varepsilon_{n,4/3} = (\rho_n)^{4/3}/2$. An application of Lemma 3.49, Lemma 3.50 and Assumption (C1) give

$$\begin{aligned}
&-2\log \mathbb{E}_{g_s}[L_{M_{k_n}^s}(\widetilde{\boldsymbol{p}}^s \,|\, \boldsymbol{T}_n(k_n))] \\
&\quad \leq -2\log \int_{U_{\varepsilon_n}(\widehat{\boldsymbol{p}}_n^s)} L_{M_{k_n}^s}(\widetilde{\boldsymbol{p}}^s \,|\, \boldsymbol{T}_n(k_n))\, \mathrm{d}\widetilde{\boldsymbol{p}}^s - 2\log b \\
&\quad \leq -2\log L_{M_{k_n}^s}(\widehat{\boldsymbol{p}}_n^s \,|\, \boldsymbol{T}_n(k_n)) \\
&\qquad -2\log \int_{U_{\varepsilon_n}(\widehat{\boldsymbol{p}}_n^s)} \exp\Big\{\frac{-k_n}{2}(\widetilde{\boldsymbol{p}}^s - \widehat{\boldsymbol{p}}_n^s)^\top \frac{-1}{k_n} \nabla^2 \log L_{M_{k_n}^s}(\widehat{\boldsymbol{p}}_n^s \,|\, \boldsymbol{T}_n(k_n))(\widetilde{\boldsymbol{p}}^s - \widehat{\boldsymbol{p}}_n^s)\Big\} \mathrm{d}\widetilde{\boldsymbol{p}}^s \\
&\qquad -2\log b + o_{\mathbb{P}}(1) \\
&\quad \leq -2\log L_{M_{k_n}^s}(\widehat{\boldsymbol{p}}_n^s \,|\, \boldsymbol{T}_n(k_n)) - s\log(2\pi) + s\log(k_n\lambda_{n,1}) - 2\log b \\
&\qquad -2\log \int_{U_{\varepsilon_n}(\widehat{\boldsymbol{p}}_n^s)} \left(\frac{k_n\lambda_{n,1}}{2\pi}\right)^{s/2} \exp\left\{\frac{-1}{2}\frac{(\widetilde{\boldsymbol{p}}^s - \widehat{\boldsymbol{p}}_n^s)^\top(\widetilde{\boldsymbol{p}}^s - \widehat{\boldsymbol{p}}_n^s)}{1/(k_n\lambda_{n,1})}\right\} \mathrm{d}\widetilde{\boldsymbol{p}}^s + o_{\mathbb{P}}(1). \quad (3.25)
\end{aligned}$$

The integrand is a s-dimensional Gaussian density with expectation vector $\widehat{\boldsymbol{p}}_n^s$ 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 3.6, the asymptotic behavior

$$0 \leq k_n\lambda_{n,1}\varepsilon_n^2 = \underbrace{\frac{k_n(\rho_n)^{5/3}}{4}}_{\to\infty} \underbrace{\left(\frac{k_n\rho_n}{T_{n,s}(k_n)} + \frac{sk_n\rho_n}{\sum_{j=s+1}^r T_{n,j}}\right)}_{\overset{\text{Lemma 3.6}}{\to} \mathbb{1}_{\{s \geq s^*\}} + \frac{s}{r - \max(s,s^*)}} \xrightarrow{\mathbb{P}} \infty \quad (3.26)$$

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

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

as $n \to \infty$ almost surely, where we used in the last step (3.26). Thus,

$$-2\log\int_{U_{\varepsilon_n}(\widehat{\boldsymbol{p}}_n^s)} \left(\frac{k_n\lambda_{n,1}}{2\pi}\right)^{s/2} \exp\left\{\frac{-1}{2}\frac{(\widetilde{\boldsymbol{p}}^s - \widehat{\boldsymbol{p}}_n^s)^\top(\widetilde{\boldsymbol{p}}^s - \widehat{\boldsymbol{p}}_n^s)}{1/(k_n\lambda_{n,1})}\right\} \mathrm{d}\widetilde{\boldsymbol{p}}^s = o_{\mathbb{P}}(1). \tag{3.27}$$

Inserting (3.27) into (3.25) gives then

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

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

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

and finally,

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

which is the statement. □

PROOF OF THEOREM 3.28

Proof of Theorem 3.28.
(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 (3.18) that

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

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

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 3.7 and due to $\log(k_n)/k_n \to 0$ that

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

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 $n \to \infty$. □

Proof of Theorem 3.30

First, we derive some auxiliary results before we prove Theorem 3.30. Therefore, note that due (3.7) (cf. Equation (1.23) in the Supplementary Material of Meyer and Wintenberger, 2023) 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

$$
L_{M'^s_n}(\widetilde{\boldsymbol{p}}'^s \,|\, \boldsymbol{T}'_n) = L_{M^s_{n-T'_{n,2^d}}}(\widetilde{\boldsymbol{p}}^s \,|\, \boldsymbol{T}'_{n,\{1,\dots,r\}}) \cdot L_{\mathrm{Bin}_n}(1-\widetilde{q}\,|\,T'_{n,2^d}), \tag{3.28}
$$

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

$$L_{\mathrm{Bin}_n}(1-\widetilde{q}\,|\,T'_{n,2^d}) := \left(\binom{n}{T'_{n,2^d}}\right)(1-\widetilde{q})^{T'_{n,2^d}}\widetilde{q}^{\,n-T'_{n,2^d}} \tag{3.29}$$

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

$$\begin{aligned}\mathbb{E}_\lambda[L_{M^s_{n-T'_{n,2^d}}}(\widetilde{\boldsymbol{p}}^s\,|\,\boldsymbol{T}'_{n,\{1,\dots,r\}})] &:= \textstyle\int_{\Theta_s} L_{M^s_{n-T'_{n,2^d}}}(\widetilde{\boldsymbol{p}}^s\,|\,\boldsymbol{T}'_{n,\{1,\dots,r\}})\,\mathrm{d}\widetilde{\boldsymbol{p}}^s,\\ \mathbb{E}_\lambda[L_{\mathrm{Bin}_n}(1-\widetilde{q}\,|\,T'_{n,2^d})] &:= \textstyle\int_{(0,1)} L_{\mathrm{Bin}_n}(1-\widetilde{q}\,|\,T'_{n,2^d})\,\mathrm{d}\widetilde{q}.\end{aligned} \tag{3.30}$$

Then taking the expectation and logarithm in (3.28) results under Assumption (D1) in

$$\begin{aligned}&-2\log\mathbb{E}_{g'_s}[L_{M'^s_n}(\widetilde{\boldsymbol{p}}'^s\,|\,\boldsymbol{T}'_n)]\\ &\quad\le -2\log b' - 2\log\Big\{\int_{\Theta_s\times(0,1)} L_{M^s_{n-T'_{n,2^d}}}(\widetilde{\boldsymbol{p}}^s\,|\,\boldsymbol{T}'_{n,\{1,\dots,r\}})\cdot L_{\mathrm{Bin}_n}(1-\widetilde{q}\,|\,T'_{n,2^d})\,\mathrm{d}(\widetilde{\boldsymbol{p}}^s,\widetilde{q})\Big\}\\ &\quad= -2\log b' - 2\log\Big\{\int_{\Theta_s} L_{M^s_{n-T'_{n,2^d}}}(\widetilde{\boldsymbol{p}}^s\,|\,\boldsymbol{T}'_{n,\{1,\dots,r\}})\,\mathrm{d}\widetilde{\boldsymbol{p}}^s\cdot\int_{(0,1)} L_{\mathrm{Bin}_n}(1-\widetilde{q}\,|\,T'_{n,2^d})\,\mathrm{d}\widetilde{q}\Big\}\\ &\quad= -2\log b' - 2\log\mathbb{E}_\lambda[L_{M^s_{n-T'_{n,2^d}}}(\widetilde{\boldsymbol{p}}^s\,|\,\boldsymbol{T}'_{n,\{1,\dots,r\}})] - 2\log\mathbb{E}_\lambda[L_{\mathrm{Bin}_n}(1-\widetilde{q}\,|\,T'_{n,2^d})].\end{aligned} \tag{3.31}$$

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

Proposition 3.51. *Under Assumptions (B1), (B3) and (D4) the asymptotic upper bound as $n\to\infty$,*

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

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

Proposition 3.52. *Suppose Assumptions (B3) and (D3) hold. The expectation of the binomial likelihood satisfies as $n\to\infty$ the inequality*

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

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

Proof of Theorem 3.30. For the ease of notation we define $x\log x$ as zero if $x=0$. Inserting the bounds derived in Proposition 3.51 with constant C_1 and Proposition 3.52 with constant

C_2 into (3.31) gives for sufficiently large n that

$$\begin{aligned}
&-2\mathbb{E}[\log \mathbb{E}_{g'_s}[L_{M'^s_n}(\widetilde{\boldsymbol{p}}'^s \,|\, \boldsymbol{T}'_n)]] + 2\log b' \\
&\quad\le -2\mathbb{E}[\log \mathbb{E}_\lambda[L_{M^s_{n-T'_{n,2^d}}}(\widetilde{\boldsymbol{p}}^s \,|\, \boldsymbol{T}'_{n,\{1,\dots,r\}})]] - 2\mathbb{E}[\log \mathbb{E}_\lambda[L_{\mathrm{Bin}_n}(1-\widetilde{q}\,|\, T'_{n,2^d})]] \\
&\quad\le -2\mathbb{E}\big[\log\big((n-T'_{n,2^d})!\big) - (n-T'_{n,2^d})\Big(\log(n-T'_{n,2^d})-1\Big)\big] \\
&\qquad - 2\frac{nq_n}{k_n}\mathbb{E}[\log L_{M^s_{k_n}}(\widehat{\boldsymbol{p}}^s_n \,|\, \boldsymbol{T}_n(k_n))] + 2s\log\Big(k_n\sqrt{\frac{r}{2\pi(r-s)}}\Big) \\
&\qquad - 2\log(n!) + 2\mathbb{E}[\log((n-T'_{n,2^d})!)] + 2\mathbb{E}[\log(T'_{n,2^d}!)] \\
&\qquad - 2nq_n\log(k_n/n) + 2\log(n) + (C_1+C_2)nq_n \\
&\quad= \Big\{-2\log(n!) + 2\mathbb{E}\big[(n-T'_{n,2^d})\Big(\log(n-T'_{n,2^d})-1\Big)\big] + 2\mathbb{E}[\log(T'_{n,2^d}!)]\Big\} \\
&\qquad + \Big\{-2\frac{nq_n}{k_n}\mathbb{E}[\log L_{M^s_{k_n}}(\widehat{\boldsymbol{p}}^s_n \,|\, \boldsymbol{T}_n(k_n))] + 2s\log\Big(k_n\sqrt{\frac{r}{2\pi(r-s)}}\Big) \\
&\qquad\quad - 2nq_n\log(k_n/n) + 2\log(n)\Big\} + (C_1+C_2)nq_n \\
&\quad=: I_{n,1} + I_{n,2} + (C_1+C_2)nq_n. \qquad (3.32)
\end{aligned}$$

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 $x > 0$ 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

$$\begin{aligned}
&\mathbb{E}[T'_{n,2^d}\log(T'_{n,2^d})\mathbb{1}\{T'_{n,2^d}>0\}] \\
&\qquad\le n(1-q_n)\log\left(\frac{n^2(1-q_n)^2 + nq_n(1-q_n)}{n(1-q_n)}\right) \\
&\qquad= n(1-q_n)\log\left(n(1-q_n)+q_n\right) \\
&\qquad= n(1-q_n)\log\left(n(1-q_n)\right) + n(1-q_n)\log\left(\frac{n(1-q_n)+q_n}{n(1-q_n)}\right) \\
&\qquad\le n(1-q_n)\log\left(n(1-q_n)\right) + C_3 \qquad (3.33)
\end{aligned}$$

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

$$n \log n - n < \log(n!) < n \log n - n + \log n + 1 \tag{3.34}$$

to derive a bound for $\mathbb{E}[\log(T'_{n,2^d}!)]$. Hence, using the upper bound (3.34), (3.33) 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 (3.34) and (3.33) (for $n - T'_{n,2^d}$ instead of $T'_{n,2^d}$ and q_n instead of $1 - q_n$, respectively) that

$$\begin{aligned}
I_{n,1} &= -2\log(n!) + 2\mathbb{E}[(n - T'_{n,2^d})\left(\log(n - T'_{n,2^d}) - 1\right)] + 2\mathbb{E}[\log(T'_{n,2^d}!)] \\
&< -2n\log(n) + 2n + 2nq_n\log(nq_n) - 2nq_n + 2n(1-q_n)\log(n(1-q_n)) \\
&\quad - 2n(1-q_n) + 2\log(n(1-q_n)) + C_5 \\
&= [-2n\log(n) + 2nq_n\log(n) + 2n(1-q_n)\log(n(1-q_n))] \\
&\quad + [2nq_n\log(q_n) + 2\log(n(1-q_n)) + C_5] \\
&\leq 2nq_n\log(q_n) + 2\log(n) + C_5
\end{aligned} \tag{3.35}$$

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 (3.32) and (3.35) ends up with

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

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

3.4.4. Proofs of Section 3.3.1

In this section we show that the AIC is not consistent if there are bias directions of type (i). For this we work under the previous assumptions for $c \in (0,1)$ when $T^{(n)}_{s^*+1}(k_n) \xrightarrow{\mathbb{P}} \infty$ and $T^{(n)}_{s^*+2}(k_n) = O_{\mathbb{P}}(1)$. The case in which more entries go to infinity then goes in a similar way.

Theorem 3.53. *Suppose Assumption (E1), (E2) and (E4) are satisfied, $c \in (0,1)$, $T^{(n)}_{s^*+1}(k_n) \xrightarrow{\mathbb{P}} \infty$, $T^{(n)}_{s^*+1}(k_n)/k_n \xrightarrow{\mathbb{P}} 0$ and $T^{(n)}_{s^*+2}(k_n) = O_{\mathbb{P}}(1)$. Then*

$$\lim_{n\to\infty} \mathbb{P}\Big(\mathrm{AIC}^{\circ}_{k_n}(s^*+1) > \mathrm{AIC}^{\circ}_{k_n}(s^*) \Big) = 0$$

holds, i.e., the AIC° *is not consistent.*

Proof of Theorem 3.53. By the definition of the AIC° it follows that

$$\begin{aligned}
&\mathrm{AIC}^{\circ}_{k_n}(s^*+1) - \mathrm{AIC}^{\circ}_{k_n}(s^*) \\
&\qquad = -\sum_{j=s^*+1}^{s^*+1} T^{(n)}_j(k_n) \log\left(\frac{T^{(n)}_j(k_n)}{k_n}\right) \\
&\qquad\quad - \log\left(\frac{1}{\widehat{s}_n - s^* - 1} \sum_{i=s^*+2}^{\widehat{s}_n} \frac{T^{(n)}_i(k_n)}{k_n}\right) \sum_{i=s^*+2}^{\widehat{s}_n} T^{(n)}_i(k_n) \\
&\qquad\quad + \log\left(\frac{1}{\widehat{s}_n - s^*} \sum_{j=s^*+1}^{\widehat{s}_n} \frac{T^{(n)}_j(k_n)}{k_n}\right) \sum_{i=s^*+1}^{\widehat{s}_n} T^{(n)}_i(k_n) + (s^*+1-s^*) \\
&\qquad = T^{(n)}_{s^*+1}(k_n) \left(-\log\left(T^{(n)}_{s^*+1}(k_n)\right) + \log\left(\frac{1}{\widehat{s}_n - s^*} \sum_{i=s^*+1}^{\widehat{s}_n} T^{(n)}_i(k_n)\right)\right) \\
&\qquad\quad + \log\left(\frac{\widehat{s}_n - s^* - 1}{\widehat{s}_n - s^*} \frac{\sum_{j=s^*+1}^{\widehat{s}_n} T^{(n)}_j(k_n)}{\sum_{i=s^*+2}^{\widehat{s}_n} T^{(n)}_i(k_n)}\right) \sum_{i=s^*+2}^{\widehat{s}_n} T^{(n)}_i(k_n) + 1, \qquad (3.36)
\end{aligned}$$

where we used that $s^*+1 > s^*$. By Assumption (E4) follows

$$\frac{1}{\widehat{s}_n - s^* + 1} \sum_{j=s^*+1}^{\widehat{s}_n} T^{(n)}_j(k_n) \xrightarrow{\mathbb{P}} \mu,$$

which gives in combination with a Taylor expansion of the logarithm that

$$\begin{aligned}
&\log\left(\frac{\widehat{s}_n - s^* - 1}{\widehat{s}_n - s^*} \frac{\sum_{j=s^*+1}^{\widehat{s}_n} T^{(n)}_j(k_n)}{\sum_{i=s^*+2}^{\widehat{s}_n} T^{(n)}_i(k_n)}\right) \sum_{i=s^*+2}^{\widehat{s}_n} T^{(n)}_i(k_n) \\
&\quad = \left(\frac{\widehat{s}_n - s^* - 1}{\widehat{s}_n - s^*} \frac{\sum_{j=s^*+1}^{\widehat{s}_n} T^{(n)}_j(k_n)}{\sum_{i=s^*+2}^{\widehat{s}_n} T^{(n)}_i(k_n)} - 1\right) \sum_{i=s^*+2}^{\widehat{s}_n} T^{(n)}_i(k_n) + O_{\mathbb{P}}\left(\frac{(T^{(n)}_{s^*+1}(k_n))^2}{\widehat{s}_n}\right) \\
&\quad = \frac{\widehat{s}_n - s^* - 1}{\widehat{s}_n - s^*} \sum_{j=s^*+1}^{\widehat{s}_n} T^{(n)}_j(k_n) - \sum_{i=s^*+2}^{\widehat{s}_n} T^{(n)}_i(k_n) + O_{\mathbb{P}}\left(\frac{(T^{(n)}_{s^*+1}(k_n))^2}{\widehat{s}_n}\right)
\end{aligned}$$

$$= \frac{\widehat{s}_n - s^* - 1}{\widehat{s}_n - s^*} T^{(n)}_{s^*+1}(k_n) - \frac{1}{\widehat{s}_n - s^*} \sum_{i=s^*+2}^{\widehat{s}_n} T^{(n)}_i(k_n) + O_{\mathbb{P}}\left(\frac{(T^{(n)}_{s^*+1}(k_n))^2}{\widehat{s}_n}\right)$$
$$= \frac{\widehat{s}_n - s^* - 1}{\widehat{s}_n - s^*} T^{(n)}_{s^*+1}(k_n) - \mu + O_{\mathbb{P}}\left(\frac{(T^{(n)}_{s^*+1}(k_n))^2}{\widehat{s}_n}\right),$$

where the rate of the $O_{\mathbb{P}}$ term is a result of the quadratic error term of the Taylor expansion given by

$$\left(\frac{\widehat{s}_n - s^* - 1}{\widehat{s}_n - s^*} \frac{\sum_{j=s^*+1}^{\widehat{s}_n} T^{(n)}_j(k_n)}{\sum_{i=s^*+2}^{\widehat{s}_n} T^{(n)}_i(k_n)} - 1\right)^2 \sum_{i=s^*+2}^{\widehat{s}_n} T^{(n)}_i(k_n)$$
$$= \left(\frac{\widehat{s}_n - s^* - 1}{\widehat{s}_n - s^*} \sum_{j=s^*+1}^{\widehat{s}_n} T^{(n)}_j(k_n) - \sum_{i=s^*+2}^{\widehat{s}_n} T^{(n)}_i(k_n)\right)^2 \left(\sum_{i=s^*+2}^{\widehat{s}_n} T^{(n)}_i(k_n)\right)^{-1}$$
$$= \frac{1}{\widehat{s}_n} \left(T^{(n)}_{s^*+1}(k_n) - \frac{1}{\widehat{s}_n - s^*} \sum_{j=s^*+1}^{\widehat{s}_n} T^{(n)}_j(k_n)\right)^2 \left(\frac{1}{\widehat{s}_n} \sum_{i=s^*+2}^{\widehat{s}_n} T^{(n)}_i(k_n)\right)^{-1}.$$

Since by Assumption (E4) we have $1/(\widehat{s}_n - s^*) \sum_{j=s^*+1}^{\widehat{s}_n} T^{(n)}_j(k_n) \xrightarrow{\mathbb{P}} \mu$ and $1/\widehat{s}_n \sum_{i=s^*+2}^{\widehat{s}_n} T^{(n)}_i(k_n) \xrightarrow{\mathbb{P}} \mu$, we get the order of the error term as above. Inserting the Taylor expansion into (3.36) gives

$$\begin{aligned}
&\mathrm{AIC}^{\circ}_{k_n}(s^*+1) - \mathrm{AIC}^{\circ}_{k_n}(s^*) \\
&= T^{(n)}_{s^*+1}(k_n) \left(-\log\left(T^{(n)}_{s^*+1}(k_n)\right) + \log\left(\frac{1}{\widehat{s}_n - s^*} \sum_{i=s^*+1}^{\widehat{s}_n} T^{(n)}_i(k_n)\right)\right) \\
&\quad + \frac{\widehat{s}_n - s^* - 1}{\widehat{s}_n - s^*} T^{(n)}_{s^*+1}(k_n) - \mu + 1 + O_{\mathbb{P}}\left(\frac{(T^{(n)}_{s^*+1}(k_n))^2}{\widehat{s}_n}\right) \\
&= T^{(n)}_{s^*+1}(k_n) \Bigg\{ -\log\left(T^{(n)}_{s^*+1}(k_n)\right) + \log\left(\frac{1}{\widehat{s}_n - s^*} \sum_{i=s^*+1}^{\widehat{s}_n} T^{(n)}_i(k_n)\right) + \frac{\widehat{s}_n - s^* - 1}{\widehat{s}_n - s^*} \\
&\quad + O_{\mathbb{P}}\left(\frac{T^{(n)}_{s^*+1}(k_n)}{\widehat{s}_n}\right)\Bigg\} - \mu + 1.
\end{aligned}$$

Note that $T^{(n)}_{s^*+1}(k_n)/\widehat{s}_n \xrightarrow{\mathbb{P}} 0$ since $\widehat{s}_n/k_n \to c \in (0,1)$ and $T^{(n)}_{s^*+1}(k_n) = o_{\mathbb{P}}(k_n)$. Then follows from $T^{(n)}_{s^*+1}(k_n) \xrightarrow{\mathbb{P}} \infty, T^{(n)}_{s^*+1}(k_n)/\widehat{s}_n \xrightarrow{\mathbb{P}} 0$ and

$$\frac{1}{\widehat{s}_n - s^*} \sum_{i=s^*+1}^{\widehat{s}_n} T^{(n)}_i(k_n) \xrightarrow{\mathbb{P}} \mu$$

that

$$\mathrm{AIC}^{\circ}_{k_n}(s^*+1) - \mathrm{AIC}^{\circ}_{k_n}(s^*) \xrightarrow{\mathbb{P}} -\infty.$$

□

3.4.5. Proofs of Section 3.3.3

Proof of Theorem 3.37

Step 1: Suppose $s^* < s_n < q_n$ and $q_n/\sqrt{\widehat{s}_n} = o_{\mathbb{P}}(1)$. By the definition of the AIC°

$$\begin{aligned}
&\mathrm{AIC}^\circ_{k_n}(s_n) - \mathrm{AIC}^\circ_{k_n}(s^*) \\
&\quad = -\sum_{j=s^*+1}^{s_n} T_j^{(n)}(k_n) \log\Big(\frac{T_j^{(n)}(k_n)}{k_n}\Big) - \log\Big(\frac{1}{\widehat{s}_n - s_n} \sum_{i=s_n+1}^{\widehat{s}_n} \frac{T_i^{(n)}(k_n)}{k_n}\Big) \sum_{i=s_n+1}^{\widehat{s}_n} T_i^{(n)}(k_n) \\
&\qquad + \log\Big(\frac{1}{\widehat{s}_n - s^*} \sum_{j=s^*+1}^{\widehat{s}_n} \frac{T_j^{(n)}(k_n)}{k_n}\Big) \sum_{i=s^*+1}^{\widehat{s}_n} T_i^{(n)}(k_n) + (s_n - s^*).
\end{aligned}$$

A Taylor expansion of the logarithm around 1 combined with

$$\frac{\sum_{j=s^*+1}^{\widehat{s}_n} T_j^{(n)}(k_n)}{\sum_{j=s_n+1}^{\widehat{s}_n} T_j^{(n)}(k_n)} \xrightarrow{\mathbb{P}} 1,$$

which follows from Assumption (E4) and $s_n/\widehat{s}_n \xrightarrow{\mathbb{P}} 0$ as $n \to \infty$, result in

$$\begin{aligned}
&\mathrm{AIC}^\circ_{k_n}(s_n) - \mathrm{AIC}^\circ_{k_n}(s^*) \\
&\quad = \sum_{j=s^*+1}^{s_n} T_j^{(n)}(k_n) \log\left(\frac{1}{\widehat{s}_n - s^*} \sum_{i=s^*+1}^{\widehat{s}_n} \frac{T_i^{(n)}(k_n)}{T_j^{(n)}(k_n)}\right) \\
&\qquad + \frac{(\widehat{s}_n - s^*) + (s^* - s_n)}{\widehat{s}_n - s^*} \sum_{j=s^*+1}^{s_n} T_j^{(n)}(k_n) + 2\frac{s^* - s_n}{\widehat{s}_n - s^*} \sum_{j=s_n+1}^{\widehat{s}_n} T_j^{(n)}(k_n) \\
&\qquad + (s_n - s^*) + o_{\mathbb{P}}(1) \\
&\quad = \sum_{j=s^*+1}^{s_n} T_j^{(n)}(k_n) \left\{\log\left(\frac{1}{\widehat{s}_n - s^*} \sum_{i=s^*+1}^{\widehat{s}_n} \frac{T_i^{(n)}(k_n)}{T_j^{(n)}(k_n)}\right) + 1\right\} \\
&\qquad + \frac{s^* - s_n}{\widehat{s}_n - s^*} \sum_{j=s^*+1}^{s_n} T_j^{(n)}(k_n) + \frac{s^* - s_n}{\widehat{s}_n - s^*} \sum_{j=s_n+1}^{\widehat{s}_n} T_j^{(n)}(k_n) + (s_n - s^*) + o_{\mathbb{P}}(1). \quad (3.37)
\end{aligned}$$

Note that the error term of the Taylor expansion of the logarithm is similarly to the proof of Theorem 3.53 indeed $o_{\mathbb{P}}(1)$ since ${s_n}^2 (T_{s^*}^{(n)}(k_n))^2 \,/\widehat{s}_n \xrightarrow{\mathbb{P}} 0$ for $s^* + 1 \le s_n \le q_n$. By Assumptions (E4) and (E5) follows for $s^* + 1 \le j_n \le q_n$ that

$$\left| T_j^{(n)}(k_n) \log\left(\frac{1}{\widehat{s}_n - s^*} \sum_{i=s^*+1}^{\widehat{s}_n} \frac{T_i^{(n)}(k_n)}{\mu}\right)\right|$$

$$\begin{aligned}
&\leq \left|T_{s^*+1}^{(n)}(k_n)\right| \left|\log\left(\frac{1}{\widehat{s}_n - s^*}\sum_{i=s^*+1}^{\widehat{s}_n}\frac{T_i^{(n)}(k_n)}{\mu}\right)\right| \\
&\xrightarrow{\mathbb{P}} q\mu\left|\log\left(\frac{\mu}{\mu}\right)\right| = 0,
\end{aligned}$$

which allows us to replace $T_j^{(n)}(k_n)\log\left(\frac{1}{\widehat{s}_n - s^*}\sum_{i=s^*+1}^{\widehat{s}_n}\frac{T_i^{(n)}(k_n)}{T_j^{(n)}(k_n)}\right)$ by $T_j^{(n)}(k_n)\log\left(\frac{\mu}{T_j^{(n)}(k_n)}\right)$ and an additional $o_{\mathbb{P}}(1)$ term. Moreover, Assumption (E5) with $s^* < s_n \leq q_n$ for $q_n/\sqrt{\widehat{s}_n} = o_{\mathbb{P}}(1)$ yields that

$$\left|\frac{s^* - s_n}{\widehat{s}_n - s^*}\sum_{j=s^*+1}^{s_n} T_j^{(n)}(k_n)\right| \leq \frac{(s_n - s^*)^2}{\widehat{s}_n - s^*}T_{s^*+1}^{(n)}(k_n) \leq \frac{(q_n - s^*)^2}{\widehat{s}_n - s^*}T_{s^*+1}^{(n)}(k_n) \xrightarrow{\mathbb{P}} 0.$$

Hence,

$$\begin{aligned}
&\mathrm{AIC}_{k_n}^{\circ}(s_n) - \mathrm{AIC}_{k_n}^{\circ}(s^*) \\
&\quad = \sum_{j=s^*+1}^{s_n}\left\{T_j^{(n)}(k_n)\left(\log\left(\frac{\mu}{T_j^{(n)}(k_n)}\right) + 1\right) + 1 - \frac{1}{\widehat{s}_n - s^*}\sum_{i=s_n+1}^{\widehat{s}_n} T_i^{(n)}(k_n)\right\} + o_{\mathbb{P}}(1).
\end{aligned}$$

Since the function $f(x) = x\left(\log\left(\frac{\mu}{x}\right) + 1\right)$ with derivative $f'(x) = \log(\mu/x)$ for $x > \mu$ is decreasing in x and $\liminf_{n\to\infty} T_{q_n}^{(n)}(k_n) \geq \mu$ $\mathbb{P}$-a.s. (because otherwise we receive a contradiction to Assumption (E4)), an application of Assumptions (E4) and (E5) with $T_{s^*+1}^{(n)}(k_n) \geq T_{s^*+2}^{(n)}(k_n) \geq \cdots \geq T_{s_n}^{(n)}(k_n)$ yields that

$$\begin{aligned}
&\frac{\mathrm{AIC}_{k_n}^{\circ}(s_n) - \mathrm{AIC}_{k_n}^{\circ}(s^*)}{s_n - s^*} \\
&\quad \geq \left\{T_{s^*+1}^{(n)}(k_n)\left(\log\left(\frac{\mu}{T_{s^*+1}^{(n)}(k_n)}\right) + 1\right) - \frac{1}{\widehat{s}_n - s^*}\sum_{i=s_n+1}^{\widehat{s}_n} T_i^{(n)}(k_n) + 1\right\} + o_{\mathbb{P}}(1) \\
&\quad \xrightarrow{\mathbb{P}} q\mu\left(\log\left(\frac{1}{q}\right) + 1\right) - \mu + 1 = \mu\Big(q(1 - \log(q)) - 1 + \frac{1}{\mu}\Big).
\end{aligned}$$

By assumption the right-hand side is positive and the assertion follows. This condition is also a necessary condition, since if it is not satisfied, we do not have consistency against the $(s^* + 1)$-th model.

Step 2: Suppose $s < s^*$. We obtain analogously to the other step that

$$\begin{aligned}
&\frac{\mathrm{AIC}_{k_n}^{\circ}(s) - \mathrm{AIC}_{k_n}^{\circ}(s^*)}{k_n} \\
&\quad = \sum_{j=s+1}^{s^*}\frac{T_j^{(n)}(k_n)}{k_n}\log\left(\frac{T_j^{(n)}(k_n)}{k_n}\right) - \log\left(\frac{1}{\widehat{s}_n - s}\sum_{j=s+1}^{\widehat{s}_n}\frac{T_j^{(n)}(k_n)}{k_n}\right)\sum_{i=s+1}^{\widehat{s}_n}\frac{T_i^{(n)}(k_n)}{k_n} \\
&\qquad + \log\left(\frac{1}{\widehat{s}_n - s^*}\sum_{j=s^*+1}^{\widehat{s}_n}\frac{T_j^{(n)}(k_n)}{k_n}\right)\sum_{i=s^*+1}^{\widehat{s}_n}\frac{T_i^{(n)}(k_n)}{k_n} + \frac{(s - s^*)}{k_n}
\end{aligned}$$

$$
\begin{aligned}
&= \sum_{j=s+1}^{s^*} \frac{T_j^{(n)}(k_n)}{k_n} \log\left(\frac{T_j^{(n)}(k_n)}{k_n}\right) - \log\left(\frac{1}{\widehat{s}_n - s} \sum_{j=s+1}^{\widehat{s}_n} \frac{T_j^{(n)}(k_n)}{k_n}\right) \sum_{i=s+1}^{s^*} \frac{T_i^{(n)}(k_n)}{k_n} \\
&\quad + \log\left(\frac{\widehat{s}_n - s}{\widehat{s}_n - s^*} \frac{\sum_{j=s^*+1}^{\widehat{s}_n} T_j^{(n)}(k_n)}{\sum_{i=s+1}^{\widehat{s}_n} T_i^{(n)}(k_n)}\right) \sum_{i=s^*+1}^{\widehat{s}_n} \frac{T_i^{(n)}(k_n)}{k_n} + \frac{(s-s^*)}{k_n} \\
&= E_{n,1} + E_{n,2} + E_{n,3} + \frac{(s-s^*)}{k_n}. \qquad (3.38)
\end{aligned}
$$

Next, we derive the behavior of each of the terms in (3.38). By Assumption (E3) follows

$$
E_{n,1} = \sum_{j=s+1}^{s^*} \frac{T_j^{(n)}(k_n)}{k_n} \log\left(\frac{T_j^{(n)}(k_n)}{k_n}\right) \xrightarrow{\mathbb{P}} \sum_{j=s+1}^{s^*} p'_j \log\left(p'_j\right).
$$

Further, by Assumptions (E2), (E3) and (E4) we have that

$$
\sum_{j=s+1}^{\widehat{s}_n} \frac{T_j^{(n)}(k_n)}{k_n} = \sum_{j=s+1}^{s^*} \frac{T_j^{(n)}(k_n)}{k_n} + \frac{\widehat{s}_n}{k_n} \sum_{j=s^*+1}^{\widehat{s}_n} \frac{T_j^{(n)}(k_n)}{\widehat{s}_n} \xrightarrow{\mathbb{P}} \sum_{j=s+1}^{s^*} p'_j + c\mu
$$

and therefore

$$
-\log\left(\sum_{j=s+1}^{\widehat{s}_n} \frac{T_j^{(n)}(k_n)}{k_n}\right) \sum_{i=s+1}^{s^*} \frac{T_i^{(n)}(k_n)}{k_n} \xrightarrow{\mathbb{P}} -\log\left(\sum_{j=s+1}^{s^*} p'_j + c\mu\right)\left(\sum_{i=s+1}^{s^*} p'_i\right)
$$

as well as

$$
-\log\left(\frac{1}{\widehat{s}_n - s}\right) \sum_{i=s+1}^{s^*} \frac{T_i^{(n)}(k_n)}{k_n} = \log\left(\widehat{s}_n - s\right) \sum_{i=s+1}^{s^*} \frac{T_i^{(n)}(k_n)}{k_n} \xrightarrow{\mathbb{P}} \infty
$$

such that

$$
E_{n,2} = -\log\left(\frac{1}{\widehat{s}_n - s} \sum_{j=s+1}^{\widehat{s}_n} \frac{T_j^{(n)}(k_n)}{k_n}\right) \sum_{i=s+1}^{s^*} \frac{T_i^{(n)}(k_n)}{k_n} \xrightarrow{\mathbb{P}} \infty.
$$

Similarly, from Assumptions (E2), (E3) and (E4) follows for $c > 0$

$$
\begin{aligned}
E_{n,3} &= \log\left(\frac{\widehat{s}_n - s}{\widehat{s}_n - s^*} \frac{\sum_{j=s^*+1}^{\widehat{s}_n} T_j^{(n)}(k_n)}{\sum_{i=s+1}^{\widehat{s}_n} T_i^{(n)}(k_n)}\right) \sum_{i=s^*+1}^{\widehat{s}_n} \frac{T_i^{(n)}(k_n)}{k_n} \\
&\xrightarrow{\mathbb{P}} \log\left(\frac{\mu}{c^{-1}\sum_{i=s+1}^{s^*} p'_i + \mu}\right) c\mu.
\end{aligned}
$$

For $c = 0$ we get with the same assumptions that

$$
E_{n,3} = \log\left(\frac{\widehat{s}_n}{k_n}\left(\frac{\widehat{s}_n - s}{\widehat{s}_n} \frac{\frac{1}{\widehat{s}_n - s^*} \sum_{j=s^*+1}^{\widehat{s}_n} T_j^{(n)}(k_n)}{\frac{1}{k_n} \sum_{i=s+1}^{\widehat{s}_n} T_i^{(n)}(k_n)}\right)\right) \frac{\widehat{s}_n}{k_n} \sum_{i=s^*+1}^{\widehat{s}_n} \frac{T_i^{(n)}(k_n)}{\widehat{s}_n}
$$

$$\xrightarrow{\mathbb{P}} \quad 0,$$

since $x\log(x) \to 0$ for $x \searrow 0$. In summary, as $(s-s^*)/k_n \to 0$ as $n \to \infty$, we get

$$\frac{\mathrm{AIC}^{\circ}_{k_n}(s) - \mathrm{AIC}^{\circ}_{k_n}(s^*)}{k_n} \xrightarrow{\mathbb{P}} \infty.$$

PROOF OF THEOREM 3.39

Step 1: Suppose $s^* < s_n < q_n$ and $q_n/\sqrt{\widehat{s}_n} = o_{\mathbb{P}}(1)$. In this case,

$$\begin{aligned} &\mathrm{QAIC}^{\circ}_{k_n}(s_n) - \mathrm{QAIC}^{\circ}_{k_n}(s^*) \qquad (3.39)\\ &= \sum_{j=s^*+1}^{s_n} \log\left(T_j^{(n)}(k_n)\right) + (\widehat{s}_n - s_n)\log\left(\frac{\widehat{\rho}_n^{s_n}}{\widehat{\rho}_n^{s^*}}\right) - (s_n - s^*)\log\left(k_n\widehat{\rho}_n^{s^*}\right) + (s_n - s^*). \end{aligned}$$

In the following, we derive an alternative representation of the second summand by using a Taylor expansion of the logarithm. Therefore, note that

$$\begin{aligned} \left(\frac{\widehat{\rho}_n^{s_n}}{\widehat{\rho}_n^{s^*}} - 1\right) &= \frac{1}{k_n\widehat{\rho}_n^{s^*}}(k_n\widehat{\rho}_n^{s_n} - k_n\widehat{\rho}_n^{s^*})\\ &= \frac{1}{k_n\widehat{\rho}_n^{s^*}}\left(\left(\frac{s_n - s^*}{(\widehat{s}_n - s_n)(\widehat{s}_n - s^*)}\right)\sum_{j=s_n+1}^{\widehat{s}_n} T_j^{(n)}(k_n) - \frac{1}{\widehat{s}_n - s_n}\sum_{j=s^*+1}^{s_n} T_j^{(n)}(k_n)\right)\\ &= O_{\mathbb{P}}(s_n/\widehat{s}_n) = o_{\mathbb{P}}(1), \qquad (3.40) \end{aligned}$$

which justifies a Taylor expansion, and the behavior of the error term

$$(\widehat{s}_n - s_n)\left(\frac{\widehat{\rho}_n^{s_n}}{\widehat{\rho}_n^{s^*}} - 1\right)^2 = O_{\mathbb{P}}(s_n^2/\widehat{s}_n) = o_{\mathbb{P}}(1).$$

By Assumption (E4) and $T_j^{(n)}(k_n) = O_{\mathbb{P}}(1)$ for $j = s^*+1, \ldots, s$ due to Assumption (E5) and by using the Taylor expansion we have as $n \to \infty$ that

$$\begin{aligned} &(\widehat{s}_n - s_n)\log\left(\frac{\widehat{\rho}_n^{s_n}}{\widehat{\rho}_n^{s^*}}\right)\\ &= (\widehat{s}_n - s_n)\left(\frac{\widehat{\rho}_n^{s_n}}{\widehat{\rho}_n^{s^*}} - 1\right) + o_{\mathbb{P}}(1)\\ &= -(\widehat{s}_n - s_n)\frac{\widehat{s}_n - s^*}{\widehat{s}_n - s_n}\frac{\sum_{j=s^*+1}^{s_n} T_j^{(n)}(k_n)}{\sum_{j=s^*+1}^{\widehat{s}_n} T_j^{(n)}(k_n)} + (\widehat{s}_n - s_n)\left(\frac{\widehat{s}_n - s^*}{\widehat{s}_n - s_n} - 1\right) + o_{\mathbb{P}}(1)\\ &= -\sum_{j=s^*+1}^{s_n}\frac{T_j^{(n)}(k_n)}{\mu} + s_n - s^* + o_{\mathbb{P}}(1). \end{aligned}$$

Then, we get

$$\begin{aligned}
&\mathrm{QAIC}^{\circ}_{k_n}(s_n) - \mathrm{QAIC}^{\circ}_{k_n}(s^*) \\
&= \sum_{j=s^*+1}^{s_n} \log\left(T_j^{(n)}(k_n)\right) - \sum_{j=s^*+1}^{s_n} \frac{T_j^{(n)}(k_n)}{\mu} + s_n - s^* - (s_n - s^*)\log(\mu) + s_n - s^* + o_{\mathbb{P}}(1) \\
&= \sum_{j=s^*+1}^{s_n} \left(\log\left(\frac{T_j^{(n)}(k_n)}{\mu}\right) - \frac{T_j^{(n)}(k_n)}{\mu} + 2 \right) + o_{\mathbb{P}}(1).
\end{aligned}$$

Since the function $\log(x) - x$ is monotone decreasing in x for $x \geq 1$ and the $T_j^{(n)}(k_n)$ are also decreasing in j and lower bounded by 1, follows by Assumption (E5)

$$\begin{aligned}
\frac{\mathrm{QAIC}^{\circ}_{k_n}(s_n) - \mathrm{QAIC}^{\circ}_{k_n}(s^*)}{s_n} &\geq \frac{s_n - s^*}{s_n}\Big(\log\left(\frac{T_{s^*+1}^{(n)}(k_n)}{\mu}\right) - \frac{T_{s^*+1}^{(n)}(k_n)}{\mu} + 2\Big) + o_{\mathbb{P}}(1) \\
&\geq \frac{1}{s^*+1}\Big(\log(q) - q + 2\Big) + o_{\mathbb{P}}(1),
\end{aligned}$$

which yields the consistency. Note that this condition is also a necessary condition, since if it is not satisfied, we do not have consistency against the (s^*+1)-th model.

Step 2: Suppose $s < s^*$. Due to Assumption (E3) we have as $n \to \infty$,

$$\frac{T_j^{(n)}(k_n)}{k_n} \xrightarrow{\mathbb{P}} p'_j > 0, \quad j = 1, \ldots, s^*,$$

and with Assumptions (E2), (E3) and (E4) follows

$$\widehat{s}_n \widehat{\rho}_n^s = \frac{\widehat{s}_n}{\widehat{s}_n - s} \sum_{j=s+1}^{s^*} \frac{T_j^{(n)}(k_n)}{k_n} + \frac{\widehat{s}_n}{k_n} \sum_{j=s^*+1}^{\widehat{s}_n} \frac{T_j^{(n)}(k_n)}{\widehat{s}_n - s} \quad \xrightarrow{\mathbb{P}} \quad \sum_{j=s+1}^{s^*} p'_j + c\mu,$$

as well as

$$\widehat{s}_n \widehat{\rho}_n^{s^*} \xrightarrow{\mathbb{P}} c\mu.$$

Hence, if $c = 0$ then

$$-\log(\widehat{s}_n \widehat{\rho}_n^{s^*}) \xrightarrow{\mathbb{P}} \infty. \tag{3.41}$$

Therefore, we have as $n \to \infty$ for $c > 0$,

$$\begin{aligned}
&\frac{\mathrm{QAIC}^{\circ}_{k_n}(s) - \mathrm{QAIC}^{\circ}_{k_n}(s^*)}{\widehat{s}_n} \\
&= \quad - \sum_{j=s+1}^{s^*} \frac{1}{\widehat{s}_n} \log\left(\widehat{s}_n \frac{T_j^{(n)}(k_n)}{k_n}\right) + \frac{\widehat{s}_n - s}{\widehat{s}_n} \log(\widehat{s}_n \widehat{\rho}_n^s) - \frac{\widehat{s}_n - s^*}{\widehat{s}_n} \log(\widehat{s}_n \widehat{\rho}_n^{s^*}) + \frac{s - s^*}{\widehat{s}_n}
\end{aligned}$$

$$
\begin{aligned}
&= -\sum_{j=s+1}^{s^*} \frac{\log(\widehat{s}_n) + \log\left(\frac{T_j^{(n)}(k_n)}{k_n}\right)}{\widehat{s}_n} + \frac{\widehat{s}_n - s}{\widehat{s}_n}\log(\widehat{s}_n\widehat{\rho}_n^{s}) - \frac{\widehat{s}_n - s^*}{\widehat{s}_n}\log(\widehat{s}_n\widehat{\rho}_n^{s^*}) + \frac{s - s^*}{\widehat{s}_n} \\
&\xrightarrow{\mathbb{P}} \log\left(\frac{\sum_{j=s+1}^{s^*} p_j' + c\mu}{c\mu}\right) > 0.
\end{aligned}
$$

If $c = 0$, then the difference converges to ∞ in probability by (3.41).

PROOF OF THEOREM 3.40

Step 1: Suppose $s^* < s_n < q_n$ and $q_n/\sqrt{\widehat{s}_n} = o_{\mathbb{P}}(1)$. Since $\widehat{\rho}_n^{s} \le \widehat{\rho}_n^{s^*}$, it follows that

$$
\begin{aligned}
&\mathrm{MSEIC}^{\circ}_{k_n}(s_n) - \mathrm{MSEIC}^{\circ}_{k_n}(s^*) \\
&\quad = 2(s_n - s^*) + \frac{k_n}{\widehat{\rho}_n^{s_n}} \sum_{j=s_n+1}^{\widehat{s}_n} \left(\frac{T_j^{(n)}(k_n)}{k_n} - \widehat{\rho}_n^{s_n}\right)^2 - \frac{k_n}{\widehat{\rho}_n^{s^*}} \sum_{j=s^*+1}^{\widehat{s}_n} \left(\frac{T_j^{(n)}(k_n)}{k_n} - \widehat{\rho}_n^{s^*}\right)^2 \\
&\quad \ge 2(s_n - s^*) + \frac{1}{k_n\widehat{\rho}_n^{s^*}} \left(\sum_{j=s_n+1}^{\widehat{s}_n} \left(T_j^{(n)}(k_n) - k_n\widehat{\rho}_n^{s_n}\right)^2 - \sum_{j=s^*+1}^{\widehat{s}_n} \left(T_j^{(n)}(k_n) - k_n\widehat{\rho}_n^{s^*}\right)^2 \right).
\end{aligned} \tag{3.42}
$$

By Assumption (E4) and $T_j^{(n)}(k_n) \le T_{s^*+1}^{(n)}(k_n) = O_{\mathbb{P}}(1)$ for $j > s^*$ due to Assumption (E5), we get

$$
\begin{aligned}
\frac{(\widehat{s}_n - s^*)k_n}{s_n - s^*}(\widehat{\rho}_n^{s^*} - \widehat{\rho}_n^{s_n}) &= \frac{1}{s_n - s^*} \sum_{j=s^*+1}^{\widehat{s}_n} T_j^{(n)}(k_n) - \frac{s_n - s^*}{(\widehat{s}_n - s_n)(s_n - s^*)} \sum_{j=s_n+1}^{\widehat{s}_n} T_j^{(n)}(k_n) \\
&= \frac{1}{s_n - s^*} \sum_{j=s^*+1}^{s_n} T_j^{(n)}(k_n) + \frac{s^* - s_n}{s_n - s^*} \frac{1}{\widehat{s}_n - s_n} \sum_{j=s_n+1}^{\widehat{s}_n} T_j^{(n)}(k_n) \\
&= O_{\mathbb{P}}(1).
\end{aligned} \tag{3.43}
$$

Hence, with $k_n\widehat{\rho}_n^{s^*} \xrightarrow{\mathbb{P}} \mu$ and $k_n\widehat{\rho}_n^{s_n} \xrightarrow{\mathbb{P}} \mu$ by Assumption (E4) we get

$$
\left(2\frac{1}{\widehat{s}_n} \sum_{j=s^*+1}^{\widehat{s}_n} T_j^{(n)}(k_n) - k_n(\widehat{\rho}_n^{s^*} + \widehat{\rho}_n^{s})\right) = k_n\widehat{\rho}_n^{s^*} - k_n\widehat{\rho}_n^{s} \xrightarrow{\mathbb{P}} \mu - \mu = 0. \tag{3.44}
$$

From this it follows with (3.43) that

$$
2 \sum_{j=s^*+1}^{\widehat{s}_n} T_j^{(n)}(k_n)k_n(\widehat{\rho}_n^{s^*} - \widehat{\rho}_n^{s}) + (\widehat{s}_n - s^*)k_n^2((\widehat{\rho}_n^{s})^2 - (\widehat{\rho}_n^{s^*})^2)
$$

$$
\begin{aligned}
&= \widehat{s}_n k_n(\widehat{\rho}_n^{s^*} - \widehat{\rho}_n^{s}) \left(2\frac{1}{\widehat{s}_n} \sum_{j=s^*+1}^{\widehat{s}_n} T_j^{(n)}(k_n) - k_n(\widehat{\rho}_n^{s^*} + \widehat{\rho}_n^{s}) \right) \\
&\xrightarrow{\mathbb{P}} 0. \qquad (3.45)
\end{aligned}
$$

For the last term in (3.42), we have

$$
\begin{aligned}
&\sum_{j=s+1}^{\widehat{s}_n} \left(T_j^{(n)}(k_n) - k_n\widehat{\rho}_n^{s} \right)^2 - \sum_{j=s^*+1}^{\widehat{s}_n} \left(T_j^{(n)}(k_n) - k_n\widehat{\rho}_n^{s^*} \right)^2 \\
&\quad = -\sum_{j=s^*+1}^{s} T_j^{(n)}(k_n)^2 + 2k_n\widehat{\rho}_n^{s} \sum_{j=s^*+1}^{s} T_j^{(n)}(k_n) + 2\sum_{j=s^*+1}^{\widehat{s}_n} T_j^{(n)}(k_n) k_n(\widehat{\rho}_n^{s^*} - \widehat{\rho}_n^{s}) \\
&\qquad + (s^* - s)k_n^2(\widehat{\rho}_n^{s})^2 + (\widehat{s}_n - s^*)k_n^2((\widehat{\rho}_n^{s})^2 - (\widehat{\rho}_n^{s^*})^2) \\
&\quad \geq -\sum_{j=s^*+1}^{s} \left(T_{s^*+1}^{(n)}(k_n) - k_n\widehat{\rho}_n^{s} \right)^2 + 2\sum_{j=s^*+1}^{\widehat{s}_n} T_j^{(n)}(k_n) k_n(\widehat{\rho}_n^{s^*} - \widehat{\rho}_n^{s}) \\
&\qquad + (\widehat{s}_n - s^*)k_n^2((\widehat{\rho}_n^{s})^2 - (\widehat{\rho}_n^{s^*})^2),
\end{aligned}
$$

where we used that $0 < T_j^{(n)}(k_n) - k_n\widehat{\rho}_n^{s_n} < T_{s^*+1}^{(n)}(k_n) - k_n\widehat{\rho}_n^{s_n}$ for $j = s^*+1, \ldots, s_n$. Further, as $T_{s^*+1}^{(n)}(k_n) - k_n\widehat{\rho}_n^{s_n} \xrightarrow{\mathbb{P}} q\mu - \mu$ for $j = s^*+1, \ldots, s_n$ by Assumption (E4) and Assumption (E5), it follows from (3.45) that

$$
\begin{aligned}
&\frac{1}{s_n - s^*} \Big(-\sum_{j=s^*+1}^{s_n} \big(T_{s^*+1}^{(n)}(k_n) - k_n\widehat{\rho}_n^{s_n} \big)^2 + 2\sum_{j=s^*+1}^{\widehat{s}_n} T_j^{(n)}(k_n) k_n(\widehat{\rho}_n^{s^*} - \widehat{\rho}_n^{s_n}) \\
&\qquad\qquad + (\widehat{s}_n - s^*)k_n^2((\widehat{\rho}_n^{s_n})^2 - (\widehat{\rho}_n^{s^*})^2) \Big) \\
&\qquad \xrightarrow{\mathbb{P}} -(q-1)^2\mu^2.
\end{aligned}
$$

In summary with $k_n\widehat{\rho}_n^{s^*} \xrightarrow{\mathbb{P}} \mu$ by Assumption (E4), (3.42) then yields

$$
\begin{aligned}
&\frac{\mathrm{MSEIC}^{\circ}_{k_n}(s_n) - \mathrm{MSEIC}^{\circ}_{k_n}(s^*)}{s_n - s^*} \\
&\geq \quad 2 + \frac{1}{k_n\widehat{\rho}_n^{s^*}(s_n - s^*)} \Big(-\sum_{j=s^*+1}^{s} \left(T_{s^*+1}^{(n)}(k_n) - k_n\widehat{\rho}_n^{s} \right)^2 \\
&\qquad +2\sum_{j=s^*+1}^{\widehat{s}_n} T_j^{(n)}(k_n) k_n(\widehat{\rho}_n^{s^*} - \widehat{\rho}_n^{s}) + (\widehat{s}_n - s^*)k_n^2((\widehat{\rho}_n^{s})^2 - (\widehat{\rho}_n^{s^*})^2) \Big) \\
&\xrightarrow{\mathbb{P}} \quad 2 - \frac{1}{\mu}(q-1)^2\mu^2 = 2 - (q-1)^2\mu.
\end{aligned}
$$

Step 2: Suppose $s < s^*$. Analogously to the proof of Theorem 3.39 we get by Assumptions (E2), (E3) and (E4) that

$$\widehat{s}_n \widehat{\rho}_n^s = \frac{\widehat{s}_n}{\widehat{s}_n - s} \sum_{j=s+1}^{s^*} \frac{T_j^{(n)}(k_n)}{k_n} + \frac{\widehat{s}_n}{k_n} \sum_{j=s+1}^{\widehat{s}_n} \frac{T_j^{(n)}(k_n)}{\widehat{s}_n - s} \quad \xrightarrow{\mathbb{P}} \quad \sum_{j=s+1}^{s^*} p_j' + c\mu,$$

$$\widehat{\rho}_n^s \xrightarrow{\mathbb{P}} 0$$

as well as

$$\widehat{s}_n \widehat{\rho}_n^{s^*} \xrightarrow{\mathbb{P}} c\mu.$$

With $T_j^{(n)}(k_n)/k_n \xrightarrow{\mathbb{P}} p_j'$ by Assumption (E3) and $T_j^{(n)}(k_n) = O_{\mathbb{P}}(1)$ due to Assumption (E5) follows

$$\begin{aligned}
&\frac{1}{\widehat{\rho}_n^s \widehat{s}_n} \sum_{j=s+1}^{\widehat{s}_n} \left(\frac{T_j^{(n)}(k_n)}{k_n} - \widehat{\rho}_n^s \right)^2 \\
&= \quad \frac{1}{\widehat{\rho}_n^s \widehat{s}_n} \left(\sum_{j=s+1}^{\widehat{s}_n} \left(\frac{T_j^{(n)}(k_n)}{k_n} \right)^2 - 2\widehat{\rho}_n^s \sum_{j=s+1}^{\widehat{s}_n} \frac{T_j^{(n)}(k_n)}{k_n} + \widehat{s}_n (\widehat{\rho}_n^s)^2 \right) \\
&\xrightarrow{\mathbb{P}} \quad \left(\sum_{j=s+1}^{s^*} p_j' + c\mu \right)^{-1} \sum_{j=s+1}^{s^*} (p_j')^2.
\end{aligned}$$

Additionally, Assumption (E5) gives that $T_j^{(n)}(k_n) \leq T_{s^*+1}^{(n)}(k_n) = O_{\mathbb{P}}(1)$ and since $k_n \widehat{\rho}_n^{s^*} \xrightarrow{\mathbb{P}} \mu$ by Assumption (E4) we have that

$$0 \leq \frac{1}{\widehat{\rho}_n^{s^*} \widehat{s}_n} \sum_{j=s^*+1}^{\widehat{s}_n} \frac{T_j^{(n)}(k_n)^2}{k_n^2} \leq \frac{1}{\widehat{\rho}_n^{s^*} k_n} \frac{\widehat{s}_n - s^*}{k_n \widehat{s}_n} T_{s^*+1}^{(n)}(k_n)^2 \xrightarrow{\mathbb{P}} 0$$

and hence, it follows that

$$\begin{aligned}
&\frac{1}{\widehat{\rho}_n^{s^*} \widehat{s}_n} \sum_{j=s^*+1}^{\widehat{s}_n} \left(\frac{T_j^{(n)}(k_n)}{k_n} - \widehat{\rho}_n^{s^*} \right)^2 \\
&= \quad \frac{1}{\widehat{\rho}_n^{s^*} \widehat{s}_n} \sum_{j=s^*+1}^{\widehat{s}_n} \frac{T_j^{(n)}(k_n)^2}{k_n^2} - 2\frac{1}{\widehat{s}_n} \sum_{j=s^*+1}^{\widehat{s}_n} \frac{T_j^{(n)}(k_n)}{k_n} + \frac{1}{\widehat{s}_n} (\widehat{\rho}_n^{s^*})^2 \\
&\xrightarrow{\mathbb{P}} \quad 0.
\end{aligned}$$

Therefore,

$$\frac{\mathrm{MSEIC}_{k_n}^{\circ}(s) - \mathrm{MSEIC}_{k_n}^{\circ}(s^*)}{k_n \widehat{s}_n}$$

$$= 2\frac{s-s^*}{k_n\widehat{s}_n} + \frac{1}{\widehat{s}_n\widehat{\rho}_n^s}\sum_{j=s+1}^{\widehat{s}_n}\left(\frac{T_j^{(n)}(k_n)}{\widehat{s}_n} - \widehat{\rho}_n^s\right)^2 - \frac{1}{\widehat{\rho}_n^{s^*}\widehat{s}_n}\sum_{j=s^*+1}^{\widehat{s}_n}\left(\frac{T_j^{(n)}(k_n)}{k_n} - \widehat{\rho}_n^{s^*}\right)^2$$
$$\xrightarrow{\mathbb{P}} \left(\sum_{j=s+1}^{s^*} p_j' + c\mu\right)^{-1}\sum_{j=s+1}^{s^*}(p_j')^2 > 0.$$

Proof of Theorem 3.41

(a) Note that

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

We now consider the different cases $s > s^*$ and $s < s^*$ separately.

Step 1: Suppose $s^* < s_n < q_n$ and $q_n/\sqrt{\widehat{s}_n} = o_{\mathbb{P}}(1)$. We receive similarly to Step 1 in the proof of Theorem 3.37 that

$$\begin{aligned}
&\frac{\mathrm{BICU}_{k_n}^{\circ}(s_n) - \mathrm{BICU}_{k_n}^{\circ}(s^*)}{(s_n - s^*)\log(k_n)} \\
&= \frac{2}{(s_n - s^*)\log(k_n)} \\
&\quad \cdot \sum_{j=s^*+1}^{s_n}\left\{T_j^{(n)}(k_n)\left(\log\left(\frac{\mu}{T_j^{(n)}(k_n)}\right) + 1\right) - \frac{1}{\widehat{s}_n - s^*}\sum_{i=s_n+1}^{\widehat{s}_n} T_i^{(n)}(k_n)\right\} \\
&\quad + 2 + \frac{s_n}{(s_n - s^*)\log(k_n)}\log\left(\frac{\widehat{s}_n}{2\pi(\widehat{s}_n - s_n)}\right) - \frac{s^*}{(s_n - s^*)\log(k_n)}\log\left(\frac{\widehat{s}_n}{2\pi(\widehat{s}_n - s^*)}\right) \\
&\quad + o_{\mathbb{P}}(\log(k_n)^{-1}) \\
&\xrightarrow{\mathbb{P}} 2 > 0.
\end{aligned}$$

Step 2: Suppose $s < s^*$. Here we have as in Step 2 in the proof of Theorem 3.37 and due to $\log(k_n)/k_n \to 0$ that

$$\begin{aligned}
&\frac{\mathrm{BICU}_{k_n}^{\circ}(s) - \mathrm{BICU}_{k_n}^{\circ}(s^*)}{k_n} \\
&= 2\frac{\mathrm{AIC}_{k_n}^{\circ}(s) - \mathrm{AIC}_{k_n}^{\circ}(s^*)}{k_n} + \frac{-2s + 2s\log(k_n) + s\log\left(\frac{\widehat{s}_n}{2\pi(\widehat{s}_n - s)}\right)}{k_n} \\
&\quad + \frac{2s^* - 2s^*\log(k_n) - s^*\log\left(\frac{\widehat{s}_n}{2\pi(\widehat{s}_n - s^*)}\right)}{k_n} \\
&\xrightarrow{\mathbb{P}} \infty,
\end{aligned}$$

and thus, the assertion follows.

(b) Again, note that

$$\mathrm{BICL}^{\circ}_{k_n}(s) = 2\,\mathrm{AIC}^{\circ}_{k_n}(s) - 2s + s\log(k_n) + s\log\left(\frac{k_n}{2\pi T_1^{(n)}(k_n)}\right).$$

By a calculation analog to part (a), the BICL° is also consistent since $\log\left(\frac{k_n}{2\pi T_1^{(n)}(k_n)}\right) \xrightarrow{\mathbb{P}} \log\left(\frac{1}{2\pi p_1'}\right) > 0$ as $n \to \infty$.

CHAPTER 4

PCA FOR MULTIVARIATE EXTREMES

In this chapter, we use the PCA approach for the dimension reduction of multivariate extreme data. We propose two information criteria, an Akaike information criterion and a Bayesian information criterion for both the fixed- and the high-dimensional cases. In the fixed-dimensional case we assume that the dimension is fixed and the number of extremes as well as the sample size go to infinity. In the high-dimensional case, the dimension is assumed to grow proportionally to the number of extremes. Notably, in the high-dimensional case, it is possible that the number of extremes is smaller than the dimension.

In the following, we consider a regularly varying random vector $\boldsymbol{X}$ with spectral vector $\boldsymbol{\Theta}$ and, similarly to Section 2.2.2, apply Principal Component Analysis to $\boldsymbol{\Theta}$. The underlying framework is the spiked covariance model (cf. (1.4)), which we adapt to the extreme case as follows.

SPIKED COVARIANCE MODEL: *The eigenvalues* $\lambda_1, \dots, \lambda_d$ *of* $\boldsymbol{\Sigma} = \mathrm{Cov}(\boldsymbol{\Theta})$ *satisfy*

$$\lambda_1 \geq \lambda_2 \geq \cdots \geq \lambda_{p^*} > \lambda_{p^*+1} = \cdots = \lambda_{d-1} =: \lambda > 0. \tag{4.1}$$

The smallest eigenvalue λ_d is not considered here to avoid numerical instability, as it can be equal to 0. Generally, for this approach we assume that the random vector $\boldsymbol{X}$ and therefore also $\boldsymbol{\Theta}$ are $\mathbb{R}^d$-valued. However, if $\boldsymbol{X}$ takes values only in $\mathbb{R}^d_+$ such that $\boldsymbol{\Theta}$ is concentrated on the positive orthant, the components of $\boldsymbol{\Theta}$ are linearly dependent. In this case, the value of the last entry of $\boldsymbol{\Theta}$ can be inferred from the first $d-1$ entries and hence $\lambda_d = 0$. The number of spiked eigenvalues p^*, i.e., the number of eigenvalues larger than λ, allows us to estimate the lower dimensionality for the dimension reduction in **(Q)**.

In the following, we develop an AIC and a BIC to estimate p^*. The main goal of this chapter is to derive necessary and sufficient conditions for our AIC and BIC to be consistent and thus to reliably reduce the dimension as in question **(Q)**. For this purpose, we differentiate between two cases where n observations are available and k_n of these are extreme. The first is the classic large sample size and fixed-dimension case, where $n \to \infty$ and the dimension d is fixed. As is typical for such information criteria (cf. Remark 2.17), we find that the BIC is consistent, whereas the AIC is not consistent. In the second case, we assume that $d = d_n$ depends on n and $d_n/k_n \to c > 0$ as $n \to \infty$.

In this case, the empirical eigenvalues are no longer consistent estimators of the theoretical eigenvalues (cf. Section 4.1.2). Therefore, we require methods from random matrix theory to derive the asymptotic properties of the empirical eigenvalues, which are the components of the AIC and BIC. For high-dimensional i.i.d. data with finite fourth moments, we

know from Theorem 4.3 that the *empirical spectral distribution function* converges to the Marčenko-Pastur law, which describes the bulk distribution of the empirical eigenvalues. The spiked covariance model (4.1) extends the Marčenko-Pastur framework by allowing for a small number of spiked eigenvalues corresponding to relevant dimensions for the PCA. In the context of this chapter, we also derive the asymptotic properties of the empirical eigenvalues of $\mathbf{\Sigma}$, whose bulk converges to the Marčenko-Pastur distribution, and use these results for the investigation of the consistency of our information criteria.

This chapter is organized as follows. In Section 4.1, we give a short summary of results for the asymptotic behavior of empirical eigenvalues for non-extreme data. THen, in Section 4.2, we properly define the empirical eigenvalues $\widehat{\lambda}_{n,1}, \ldots, \widehat{\lambda}_{n,d}$ of $\mathbf{\Sigma}$, which are the main components in the definition of the information criteria. In addition, we explore the asymptotic properties of the empirical eigenvalues, where in the high-dimensional case we restrict our study to a parametric family of distributions, the so-called *directional model*. The subjects of Section 4.3 are the AIC and the BIC for estimating the location p^* of the spiked eigenvalue in the fixed-dimensional case, where Section 4.4 explores the high-dimensional case when $d_n/k_n \to c > 0$ as $n \to \infty$. We will examine the case $0 < c < 1$ and $c > 1$ separately in Section 4.4.1 and Section 4.4.2, respectively. In both cases, we derive sufficient criteria for the AIC and the BIC to be weakly consistent. The proofs for the results presented in this chapter are provided in Section 4.5. Note that most parts of this chapter consist of Butsch and Fasen-Hartmann (2025a). Throughout this section, we use the L_2 norm $\|\cdot\|_2$.

4.1. PRELIMINARIES

In this section we give a brief overview of the results for the asymptotic behavior of empirical eigenvalues for non-extreme data, where we present the results for the fixed-dimensional setting in Section 4.1.1 and for the high-dimensional setting in Section 4.1.2. We consider the settings where the effective sample size is given by k_n, as we are working in the subsequent sections with the number of extremes given by k_n.

4.1.1. FIXED-DIMENSIONAL CASE

For non-extreme data, the asymptotic distribution of the eigenvalues of the empirical covariance matrix was analyzed in the Gaussian case in Anderson (1963).

Theorem 4.1 (Theorem 1, Anderson, 1963 and Lemma 1, Fujikoshi and Sakurai, 2016)**.** *Suppose that*

$$\widehat{\mathbf{\Psi}}_n \coloneqq \frac{1}{k_n} \sum_{j=1}^{k_n} \left(\boldsymbol{U}_j - \frac{1}{k_n} \sum_{i=1}^{k_n} \boldsymbol{U}_i \right) \left(\boldsymbol{U}_j - \frac{1}{k_n} \sum_{i=1}^{k_n} \boldsymbol{U}_i \right)^\top$$

is an empirical covariance matrix derived from a normally distributed sample $\boldsymbol{U}_1, \ldots, \boldsymbol{U}_{k_n}$ with sample size k_n, where the covariance matrix $\mathbf{\Psi}$ has eigenvalues $\zeta_1 > \zeta_2 > \cdots > \zeta_p >$

$\zeta_{p+1} = \cdots = \zeta_d = \zeta > 0$. *Let the standardized sample eigenvalues for* $j = 1, \ldots, d$ *be defined by*

$$\widetilde{\zeta}_{n,j} \coloneqq \sqrt{k_n}(\widehat{\zeta}_j - \zeta_j).$$

Then the following results hold.

(a) The limiting distributions of $\widetilde{\zeta}_{n,1}, \ldots, \widetilde{\zeta}_{n,p}, \{\widetilde{\zeta}_{n,p+1}, \ldots, \widetilde{\zeta}_{n,d}\}$ *are independent.*

(b) For $j = 1, \ldots, p$ *we have as* $n \to \infty$

$$\widetilde{\zeta}_{n,j} \xrightarrow{\mathcal{D}} \mathcal{N}(0, 2\zeta_j^2).$$

(c) The limiting distribution of $\widetilde{\zeta}_{n,p+1}, \ldots, \widetilde{\zeta}_{n,d}$ *for* $\ell_{p+1}, \ldots, \ell_d$ *has joint density proportional to*

$$\frac{(d-p)}{\zeta^{(d-p)(d-p+1)/2}} e^{-\sum_{i=1}^{d-p} \ell_{i+p}^2/(4\zeta^2)} \prod_{\substack{i,j=1,\\ i\neq j}}^{d-p} (\ell_{i+p} - \ell_{j+p})\, \mathbb{1}\{\ell_{p+1} > \cdots > \ell_d\}.$$

A generalization of this result beyond the normal distribution in $\mathbb{R}^d$ can be found in Dauxois et al. (1982), where random functions in a separable Hilbert space are considered. Note that we directly get the consistency of the empirical eigenvalues in this setting in contrast to the high-dimensional case presented next.

4.1.2. High-dimensional case

In EVT one works only with a fraction of the data, which corresponds to the most extreme observations, and as a result the sample size is vastly reduced. Therefore, it is possible that the dimension of the data is proportional to the sample size. In this case, the classical limiting theorems are not valid in general, as shown in this section. Instead, we are in the realm of random matrix theory, where the dimension of the data d_n is supposed to be proportional to the effective sample size k_n. In this section we introduce the Marčenko-Pastur law, which is then generalized by the spiked covariance model.

Before we introduce the Marčenko-Pastur (MP) law, we give the necessary notation. For this, let $\boldsymbol{A}^{(n)} \in \mathbb{R}^{d_n \times d_n}$ be a matrix with eigenvalues $\lambda_1(\boldsymbol{A}^{(n)}), \ldots, \lambda_{d_n}(\boldsymbol{A}^{(n)})$. Then we define the so-called empirical spectral distribution (ESD) (Bai and Silverstein, 2010, p. 5) of $\boldsymbol{A}^{(n)}$ by

$$F^{\boldsymbol{A}^{(n)}}(x) \coloneqq \frac{1}{d_n} \sum_{j=1}^{d_n} \mathbb{1}\{\lambda_j(\boldsymbol{A}^{(n)}) \leq x\}, \quad x \in \mathbb{R},$$

and if the pointwise limit as $n \to \infty$ of the ESD exists, the resulting function is called limiting spectral distribution (LSD) of $\boldsymbol{A}^{(n)}$.

MARČENKO-PASTUR LAW

For the introduction of the MP law we do not assume that the random variables are heavy-tailed. The connection between EVT and the MP law is established later. We begin by presenting the definition of the distribution and subsequently describe its generation as the limiting spectral distribution of a sequence of matrices.

Definition 4.2. The Marčenko-Pastur (MP) law with distribution function F_c for $c > 0$ has density function

$$f_c(x) = \begin{cases} \frac{1}{2\pi x c}\sqrt{((1+\sqrt{c})^2 - x)(x - (1-\sqrt{c})^2)}, & x \in ((1-\sqrt{c})^2, (1+\sqrt{c})^2), \\ 0, & \text{otherwise}, \end{cases}$$

and point mass $1 - 1/c$ at 0 if $c > 1$.

Suppose that $\boldsymbol{V}^{(n)} = (V_1, \ldots, V_{d_n})^\top \in \mathbb{R}^{d_n}$ is a centered random vector consisting of i.i.d. symmetric components with variance 1. Let $\boldsymbol{V}_1^{(n)}, \ldots, \boldsymbol{V}_{k_n}^{(n)}$ be a sample of i.i.d. random vectors following the distribution of $\boldsymbol{V}^{(n)}$. Then we define the sample covariance matrix by

$$\boldsymbol{S}^{(n)} \coloneqq \frac{1}{k_n}\sum_{i=1}^{k_n}(\boldsymbol{V}_i^{(n)} - \overline{\mathbf{V}}^{(n)})(\boldsymbol{V}_i^{(n)} - \overline{\mathbf{V}}^{(n)})^\top,$$

where $\overline{\mathbf{V}}^{(n)} \coloneqq \frac{1}{k_n}\sum_{i=1}^{k_n}\boldsymbol{V}_i^{(n)}$. Note that the covariance matrix of $\boldsymbol{V}^{(n)}$ is given by $\boldsymbol{I}_{d_n}$.

Theorem 4.3 (Theorem 3.6, Bai and Silverstein, 2010)**.** *Let $\boldsymbol{V}_1^{(n)}, \ldots, \boldsymbol{V}_{k_n}^{(n)}$ be a sample of i.i.d. random vectors following the distribution of $\boldsymbol{V}^{(n)}$. If $d_n/k_n \to c > 0$ as $n \to \infty$, then*

$$\mathbb{P}(\lim_{n\to\infty} F^{\boldsymbol{S}^{(n)}}(x) = F_c(x) \text{ for all } x \in \mathcal{C}(F_c)) = 1,$$

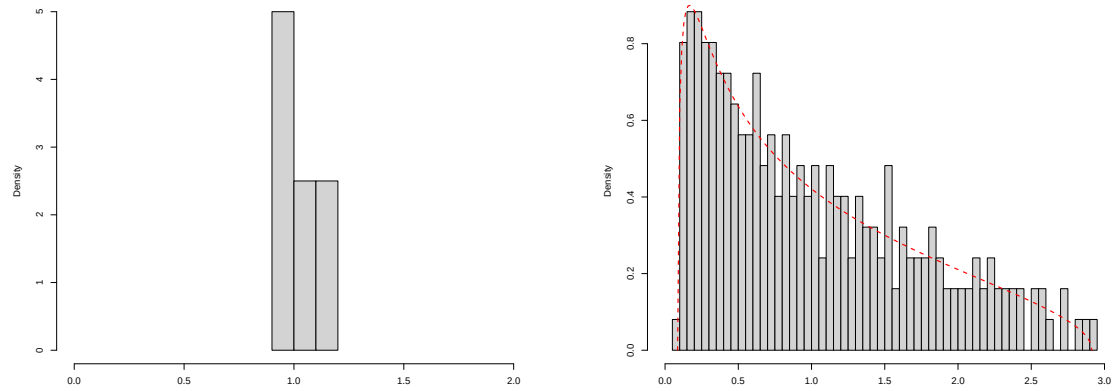
where F_c is defined as in Definition 4.2 and $\mathcal{C}(F_c)$ is the set of all continuity points of F_c.

Remark 4.4. Note that a rank-1-update such as $\overline{\mathbf{V}}^{(n)}$ has no influence on the LSD. Indeed, by Bai and Silverstein (2010, Theorem A.44) holds for matrices $\boldsymbol{A}^{(n)}, \boldsymbol{B}^{(n)} \in \mathbb{R}^{d_n \times k_n}$ that

$$\|F^{\boldsymbol{A}^{(n)}\boldsymbol{A}^{(n)\top}} - F^{\boldsymbol{B}^{(n)}\boldsymbol{B}^{(n)\top}}\| \leq \frac{1}{d_n}\operatorname{rank}(\boldsymbol{A}^{(n)} - \boldsymbol{B}^{(n)})$$

which gives the desired result.

A consequence of Theorem 4.3 is that the empirical eigenvalues do not converge in probability to their theoretical counterparts and, therefore, are not consistent. As illustrated in Figure 4.1 for a standard multivariate normal distribution, the left-hand plot shows the low-dimensional case, where the estimator is consistent and concentrated near the true value 1. On the right-hand side we see the high-dimensional case, where the histogram of the empirical eigenvalues approaches the MP law (denoted by the red line). This is one major



(a) Low-dimensional case (b) High-dimensional case

Figure 4.1.: *Histogram of empirical eigenvalues of a standard multivariate normal distribution for the low-dimensional case on the left hand side ($d = 5, k_n = 500, c = 0.01$) and the high-dimensional case on the right hand side ($d = 250, k_n = 500, c = 0.5$), where $c = d/k_n$. The red line is the density of the MP law.*

difference between the fixed-dimensional case, where empirical eigenvalues are consistent estimators, and the high-dimensional case. Clearly, results from the fixed-dimensional case cannot be taken for granted, and different methods are needed.

Spiked covariance model

For the Marčenko-Pastur law it is assumed that the eigenvalues of the covariance matrix of $\boldsymbol{V}^{(n)}$ are all equal to 1. The case where some eigenvalues are larger than 1 was the topic of Johnstone (2001). In that work, the spiked covariance model was introduced, where for a fixed $p \in \mathbb{N}$ with $p \leq d_n$ the eigenvalues of the non-negative definite symmetric matrix $\boldsymbol{\Gamma}^{(n)}$ are similar to (1.4) given by

$$\lambda_1 > \lambda_2 > \cdots > \lambda_p > \lambda_{p+1} = \cdots = \lambda_{d_n} = 1.$$

The asymptotic behavior of the empirical eigenvalues was analyzed in Baik and Silverstein (2006) and is introduced next. Suppose that $\boldsymbol{V}_1^{(n)}, \ldots, \boldsymbol{V}_{k_n}^{(n)}$ is a sample of i.i.d. random vectors following the distribution of $\boldsymbol{V}^{(n)}$ defined as in Section 4.1.2, where the components of $\boldsymbol{V}^{(n)}$ have second moment equal to 1 and the fourth moment is finite. Let

$$\boldsymbol{\Upsilon}^{(n)} := \frac{1}{k_n} \sum_{i=1}^{k_n} \boldsymbol{\Gamma}^{(n)1/2} \boldsymbol{V}_i^{(n)} \boldsymbol{V}_i^{(n)\top} \boldsymbol{\Gamma}^{(n)1/2}$$

be the sample covariance matrix with eigenvalues $\lambda_1(\boldsymbol{\Upsilon}^{(n)}), \ldots, \lambda_{d_n}(\boldsymbol{\Upsilon}^{(n)})$. Then, for $d_n/k_n \to c > 0$ the following results hold for the cases $c < 1$ and $c > 1$.

Theorem 4.5 (Theorem 1.1, Baik and Silverstein, 2006)**.** *Suppose that* $\lim_{n\to\infty} d_n/k_n = c \in (0,1)$.

(a) If $\lambda_j > 1+\sqrt{c}$ *then as* $n \to \infty$ *a.s.*

$$\lambda_j(\boldsymbol{\Upsilon}^{(n)}) \to \lambda_j + \frac{c\lambda_j}{\lambda_j - 1}.$$

(b) For $j = \arg\min\{i \in \{1,\dots,p+1\} : \lambda_i < 1+\sqrt{c}\}$ *holds as* $n \to \infty$ *a.s.*

$$\lambda_j(\boldsymbol{\Upsilon}^{(n)}) \to (1+\sqrt{c})^2.$$

(c) As $n \to \infty$ *a.s. holds*

$$\lambda_{d_n}(\boldsymbol{\Upsilon}^{(n)}) \to (1-\sqrt{c})^2.$$

Note that $j = \arg\min\{i \in \{1,\dots,p+1\} : \lambda_i < 1+\sqrt{c}\}$ is the index of the largest non-spiked eigenvalue, i.e. the largest eigenvalue smaller than $1+\sqrt{c}$.

Theorem 4.6 (Theorem 1.2 Baik and Silverstein, 2006)**.** *Suppose that* $\lim_{n\to\infty} d_n/k_n = c > 1$.

(a) If $\lambda_j > 1+\sqrt{c}$ *then as* $n \to \infty$ *a.s.*

$$\lambda_j(\boldsymbol{\Upsilon}^{(n)}) \to \lambda_j + \frac{c\lambda_j}{\lambda_j - 1}.$$

(b) For $j = \arg\min\{i \in \{1,\dots,p+1\} : \lambda_i < 1+\sqrt{c}\}$ *holds as* $n \to \infty$ *a.s.*

$$\lambda_j(\boldsymbol{\Upsilon}^{(n)}) \to (1+\sqrt{c})^2.$$

(c) As $n \to \infty$ *a.s. holds*

$$\lambda_{k_n}(\boldsymbol{\Upsilon}^{(n)}) \to (1-\sqrt{c})^2$$

and $\lambda_{k_n+1}(\boldsymbol{\Upsilon}^{(n)}), \dots, \lambda_{d_n}(\boldsymbol{\Upsilon}^{(n)}) = 0$ *for all* $n \in \mathbb{N}$.

From these results, we see that a phase transition occurs, depending on whether the eigenvalues are greater or less than $1+\sqrt{c}$. If an eigenvalue is smaller than $1+\sqrt{c}$, we say that the eigenvalue is in the bulk, and then the limit of the empirical eigenvalues lies $\mathbb{P}$-a.s. in the support of the MP law. On the other hand, all empirical eigenvalues corresponding to eigenvalues larger than $1+\sqrt{c}$, which we call spiked eigenvalues in the following, remain spiked in the limit, by which we mean that they are separated from the support of the MP law.

Remark 4.7. Note that Theorem 4.3 remains valid when exchanging $\boldsymbol{S}^{(n)}$ by $\boldsymbol{\Upsilon}^{(n)}$ since as $n \to \infty$ it holds $\mathbb{P}$-a.s. that

$$\frac{1}{d_n} \sum_{j=1}^{p} \lambda_j(\boldsymbol{\Upsilon}^{(n)}) \to 0.$$

Building on these results, we derive in the following sections asymptotic results for the high-dimensional setting for extremal data.

4.2. Asymptotics of the empirical eigenvalues of $\mathbf{\Sigma}$

The information criteria AIC and BIC of this chapter are defined by the empirical eigenvalues $\widehat{\lambda}_{n,1}, \dots, \widehat{\lambda}_{n,d}$ of $\mathbf{\Sigma}$. Therefore, in the first step, in Section 4.2.1, we define and explore the empirical eigenvalues and their asymptotic properties in the fixed-dimensional case, and then, in Section 4.2.2, in the high-dimensional case. With the knowledge of the asymptotic behavior of the empirical eigenvalues, we will be able to derive the asymptotic behavior of the AIC and the BIC in Section 4.3 and Section 4.4. The proofs of this section are moved to Section 4.5.1.

4.2.1. Fixed-dimensional case

In the case where the dimension d is fixed, we consider the following model.

Model S.

(S1) Let $\boldsymbol{X}, \boldsymbol{X}_1, \boldsymbol{X}_2, \dots$ be a sequence of i.i.d. regularly varying random vectors with tail index $\alpha > 0$ and spectral vector $\mathbf{\Theta}$.

(S2) The eigenvalues $\lambda_1, \dots, \lambda_d$ of $\mathbf{\Sigma} = \operatorname{Cov}(\mathbf{\Theta})$ satisfy

$$\lambda_1 \geq \lambda_2 \geq \cdots \geq \lambda_{p^*} > \lambda_{p^*+1} = \cdots = \lambda_{d-1} =: \lambda > 0.$$

(S3) Let $(k_n)_{n\in\mathbb{N}}$ be a sequence in $\mathbb{N}$ with $k_n \to \infty$ and $k_n/n \to 0$ for $n \to \infty$.

(S4) Suppose $(u_n)_{n\in\mathbb{N}}$ is a positive sequence such that for $n \to \infty$, $n\mathbb{P}(\|\boldsymbol{X}\|_2 > u_n)/k_n \to 1$ and

$$\sup_{x\in[\frac{1}{1+\tau},1+\tau]} \sqrt{k_n} \left\| \frac{n}{k_n} \mathbb{E}\left[\begin{pmatrix} \frac{\operatorname{vec}(\boldsymbol{X}\boldsymbol{X}^\top)}{\|\boldsymbol{X}\|_2^2} \\ 1 \end{pmatrix} \mathbb{1}\{x\|\boldsymbol{X}\|_2 > u_n\} \right] - x^\alpha \begin{pmatrix} \operatorname{vec}(\mathbb{E}[\mathbf{\Theta}\mathbf{\Theta}^\top]) \\ 1 \end{pmatrix} \right\|_2 \to 0.$$

The spiked covariance model enters through assumption (S2), where p^* is the number of relevant eigenvalues. The last assumption (S4) is a technical assumption that we require for some proofs (cf. Remark 4.9). Uniform convergence is required in order to replace the threshold u_n with the order statistic $\|\boldsymbol{X}_{(k_n+1,n)}\|_2$. Ultimately, this condition is an

assumption on the slowly varying function of the tail distribution of $\|\boldsymbol{X}\|_2$ and it is an assumption on the growth rate of k_n. Under these model assumptions, the empirical estimator for $\boldsymbol{\Theta}$ is defined as

$$\widehat{\boldsymbol{\Theta}}_n \coloneqq \frac{1}{k_n}\sum_{i=1}^{n} \frac{\boldsymbol{X}_i}{\|\boldsymbol{X}_i\|_2}\mathbb{1}\{\|\boldsymbol{X}_i\|_2 > \|\boldsymbol{X}_{(k_n+1,n)}\|_2\},$$

and hence, the empirical covariance matrix $\widehat{\boldsymbol{\Sigma}}_n$ of $\boldsymbol{\Sigma}$ is

$$\widehat{\boldsymbol{\Sigma}}_n \coloneqq \frac{1}{k_n}\sum_{j=1}^{n} \left(\frac{\boldsymbol{X}_j}{\|\boldsymbol{X}_j\|_2} - \widehat{\boldsymbol{\Theta}}_n\right)\left(\frac{\boldsymbol{X}_j}{\|\boldsymbol{X}_j\|_2} - \widehat{\boldsymbol{\Theta}}_n\right)^{\top} \mathbb{1}\{\|\boldsymbol{X}_j\|_2 > \|\boldsymbol{X}_{(k_n+1,n)}\|_2\} \tag{4.2}$$

with eigenvalues $\widehat{\lambda}_{n,1}, \dots, \widehat{\lambda}_{n,d}$ where $n \in \mathbb{N}$ is the number of observations and $\boldsymbol{X}_{(k_n+1,n)}$ denotes the observation with the (k_n+1)-th largest norm. Both the AIC and the BIC for the estimation of p^* will be defined by the empirical eigenvalues $\widehat{\lambda}_{n,1}, \dots, \widehat{\lambda}_{n,d}$. Therefore, it is important to know the asymptotic behavior. We start to derive the asymptotic behavior of the empirical covariance matrix $\widehat{\boldsymbol{\Sigma}}_n$ in the next proposition and use this to derive the asymptotic behavior of the empirical eigenvalues.

Proposition 4.8. *Let Model S be given. Then as $n \to \infty$,*

$$\sqrt{k_n}(\widehat{\boldsymbol{\Sigma}}_n - \boldsymbol{\Sigma}) \xrightarrow{\mathcal{D}} \boldsymbol{S},$$

where $\operatorname{vec}(\boldsymbol{S})$ *follows a centered normal distribution with covariance matrix*

$$\operatorname{Cov}(\operatorname{vec}((\boldsymbol{\Theta} - \mathbb{E}[\boldsymbol{\Theta}])(\boldsymbol{\Theta} - \mathbb{E}[\boldsymbol{\Theta}])^{\top})).$$

Remark 4.9. In the bivariate case and for $h : \mathbb{R}^2 \mapsto \mathbb{R}$ defined as $h(x, y) = xy$, the asymptotic distribution of

$$\frac{1}{k_n}\sum_{i=1}^{n} h\left(\frac{\boldsymbol{X}_i}{\|\boldsymbol{X}_i\|_2}\right)\mathbb{1}\{\|\boldsymbol{X}_i\|_2 > \|\boldsymbol{X}_{(k_n+1,n)}\|_2\}$$

was derived in Larsson and Resnick (2012, Theorem 1). The techniques of the proof can be generalized and applied to $\operatorname{vec}(\widehat{\boldsymbol{\Sigma}}_n)$ with the technical assumption (S4), and therefore, the proof of Proposition 4.8 is omitted. Note that if $\|\boldsymbol{\theta}\|_2 = 1$ for $\boldsymbol{\theta} \in \mathbb{R}^d$ then $\|\operatorname{vec}(\boldsymbol{\theta}\boldsymbol{\theta}^{\top})\|_2 = 1$ and higher moments of $\boldsymbol{\Theta}$ exist, since $\boldsymbol{\Theta}$ is bounded. A complementary result on the asymptotic behavior of the empirical covariance matrix is also given in the recent publication Drees (2025, Theorem 2.1).

Now, we are able to present the asymptotic distribution of the empirical eigenvalues.

Theorem 4.10. *Let Model S be given.*

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

$$(\widehat{\lambda}_{n,1}, \dots, \widehat{\lambda}_{n,d-1}) = (\lambda_1, \dots, \lambda_{d-1}) + O_{\mathbb{P}}(1/\sqrt{k_n}),$$

(b) and

$$\sqrt{k_n}\big((\widehat{\lambda}_{n,p^*+1}, \dots, \widehat{\lambda}_{n,d-1}) - \lambda \mathbf{1}_{d-p^*-1}\big) \xrightarrow{\mathcal{D}} \boldsymbol{M},$$

where the entries of the random vector $\boldsymbol{M} \in \mathbb{R}^{d-p^*-1}$ are the $(d-p^*-1)$ largest eigenvalues of $\boldsymbol{P}_\lambda \boldsymbol{S} \boldsymbol{P}_\lambda$ in decreasing order, $\boldsymbol{S}$ is defined as in Proposition 4.8 and $\boldsymbol{P}_\lambda \in \mathbb{R}^{d\times d}$ is the orthogonal projection onto the space spanned by the eigenvectors with respect to the eigenvalue λ of $\boldsymbol{\Sigma}$.

4.2.2. Directional model in the high-dimensional case

In the high-dimensional setting, where $d = d_n$ depends on n and $d_n \to \infty$ as $n \to \infty$, we restrict our studies to a parametric family of distributions, the so-called *directional model*. A directional model has the advantage that the underlying random vectors have an independent radial and directional component, but still the covariance matrix of the spectral vector has a spiked structure. The explicit definition of a directional model is the following.

Directional Model: *Suppose for any $n \in \mathbb{N}$ that*

$$\boldsymbol{\Gamma}^{(n)} := \begin{pmatrix} \boldsymbol{\Gamma}_n & \mathbf{0}_{p^*\times d_n} \\ \mathbf{0}_{d_n\times p^*} & \boldsymbol{I}_{d_n-p^*} \end{pmatrix} \in \mathbb{R}^{d_n\times d_n}, \tag{4.3}$$

where $\boldsymbol{\Gamma}_n \in \mathbb{R}^{p^*\times p^*}$ is a covariance matrix with eigenvalues

$$\xi_{n,1} \geq \dots \geq \xi_{n,p^*} > 1,$$

$\boldsymbol{V}^{(n)} = (V_1, \dots, V_{d_n})^\top \in \mathbb{R}^{d_n}$ is a centered random vector consisting of i.i.d. symmetric components with variance 1 and finite fourth moment and Z is a standard Fréchet distributed random variable. Then the sequence of random vectors $(\boldsymbol{X}^{(n)})_{n\in\mathbb{N}}$ with

$$\boldsymbol{X}^{(n)} := \frac{{\boldsymbol{\Gamma}^{(n)}}^{1/2}\boldsymbol{V}^{(n)}}{\|{\boldsymbol{\Gamma}^{(n)}}^{1/2}\boldsymbol{V}^{(n)}\|_2} \cdot Z \in \mathbb{R}^{d_n},$$

follows the so-called directional model.

Due to construction, we see directly that the directional component

$$\boldsymbol{\Theta}^{(n)} := \frac{\boldsymbol{X}^{(n)}}{\|\boldsymbol{X}^{(n)}\|_2} = \frac{{\boldsymbol{\Gamma}^{(n)}}^{1/2}\boldsymbol{V}^{(n)}}{\|{\boldsymbol{\Gamma}^{(n)}}^{1/2}\boldsymbol{V}^{(n)}\|_2},$$

of $\boldsymbol{X}^{(n)}$ is independent of the radial component $\|\boldsymbol{X}^{(n)}\|_2 = Z$, and additionally, $\boldsymbol{\Theta}^{(n)}$ is the spectral vector of the multivariate regularly varying random vector $\boldsymbol{X}^{(n)}$ of index 1. Thus, the dependence structure of $\boldsymbol{X}^{(n)}$ is completely determined by $\boldsymbol{\Theta}^{(n)}$.

Remark 4.11.

(a) In high-dimensional models it is necessary to specify the model as we have done with the directional model, because due to the increase in dimensionality, the empirical covariance matrix and even the covariance matrix do not converge and hence, it will be impossible to get any kind of limit result without assuming some structure on the underlying random vector $\boldsymbol{X}^{(n)}$. Our model assumption results on the one hand, in a spiked covariance model for $\mathrm{Cov}(\boldsymbol{\Theta}^{(n)})$ where p^*, the location of the smallest eigenvalue bigger than 1, is independent of n and fixed (see Lemma 4.12 for more details). On the other hand, it implies the independence of the directional and radial parts of $\boldsymbol{X}^{(n)}$, so that the order statistic of an i.i.d. sequence $\|\boldsymbol{X}_1^{(n)}\|_2, \ldots, \|\boldsymbol{X}_n^{(n)}\|_2$ is reflected by the order statistic of the i.i.d. sequence of radial parts $Z_1, \ldots, Z_n$.

(b) The directional model is inspired by several models in the non-extreme world. For instance, Bai et al. (2018) employed a model of the form ${\boldsymbol{\Gamma}^{(n)}}^{1/2}\boldsymbol{V}^{(n)}$, while Jiang et al. (2023) considered a model of the form $\boldsymbol{V}^{(n)}/\|\boldsymbol{V}^{(n)}\|_1$. The first model is not suitable for statistical inference of extremes because the radial and directional parts are not independent, whereas the second model is not suitable either, as the covariance matrix only has eigenvalues equal to 1 if the components V_i are symmetric. Therefore, our model $\boldsymbol{X}^{(n)} = {\boldsymbol{\Gamma}^{(n)}}^{1/2}\boldsymbol{V}^{(n)}/\|{\boldsymbol{\Gamma}^{(n)}}^{1/2}\boldsymbol{V}^{(n)}\|_2 \cdot Z$ combines both approaches: The factor ${\boldsymbol{\Gamma}^{(n)}}^{1/2}\boldsymbol{V}^{(n)}$ serves as a latent vector for the directional component $\boldsymbol{\Theta}^{(n)}$, which captures the spiked covariance structure. Normalizing it by $\|{\boldsymbol{\Gamma}^{(n)}}^{1/2}\boldsymbol{V}^{(n)}\|_2$ ensures that the norm of $\boldsymbol{X}^{(n)}$ is solely determined by Z facilitating the calculation of the order statistics of an i.i.d. sequence with distribution $\|\boldsymbol{X}_1^{(n)}\|_2$ by the order statistics of the radial parts.

(c) Although $\mathrm{Cov}(\boldsymbol{\Theta}^{(n)})$ is a spiked covariance model with p^* leading eigenvalues (see Lemma 4.12) the support of the distribution of $\boldsymbol{\Theta}^{(n)}$ might have a higher dimension than p^*. However, if ξ_{n,p^*} is large, then the support of $\boldsymbol{\Theta}^{(n)}$ is more concentrated on the p^*-dimensional subspace generated by the leading eigenvalues.

One special case for $\boldsymbol{\Theta}^{(n)}$ is the angular central Gaussian distribution, which is obtained by using a Gaussian distribution for the i.i.d. entries of $\boldsymbol{V}^{(n)}$. The density of the angular central Gaussian distribution $\boldsymbol{\Theta}^{(n)}$ is given by (Tyler, 1987) as

$$f_{\boldsymbol{\Theta}^{(n)}}(\boldsymbol{\theta} \mid \boldsymbol{\Gamma}^{(n)}) = \frac{2\pi^{\frac{d_n}{2}}}{\Gamma(\frac{d_n}{2})}\det(\boldsymbol{\Gamma}^{(n)})^{-1/2}(\boldsymbol{\theta}^\top {\boldsymbol{\Gamma}^{(n)}}^{-1}\boldsymbol{\theta})^{-d_n/2}, \quad \boldsymbol{\theta} \in \{\boldsymbol{x} \in \mathbb{R}^{d_n} : \|\boldsymbol{x}\|_2 = 1\}$$

where $\det(\cdot)$ is the determinant and $\Gamma(\cdot)$ is the Gamma function.

(d) Scaling of $\boldsymbol{V}^{(n)}$ has no influence on the distribution of $\boldsymbol{X}^{(n)}$, therefore setting the variance of V_i to 1 is no restriction.

(e) The *empirical spectral distribution* (see Section 4.1.2) of $\mathbf{\Gamma}^{(n)}$ is defined as

$$F^{\mathbf{\Gamma}^{(n)}}(x) = \frac{1}{d_n}\sum_{i=1}^{d_n} \mathbb{1}\{\xi_{n,i} \leq x\}, \quad x \in \mathbb{R},$$

and the *limiting spectral distribution* (LSD) of $\mathbf{\Gamma}^{(n)}$ is the Dirac measure δ_1, since

$$\lim_{n\to\infty} F^{\mathbf{\Gamma}^{(n)}}(x) = \lim_{n\to\infty} \frac{1}{d_n}\sum_{j=1}^{p} \mathbb{1}\{\xi_{n,j} \leq x\} + \frac{d_n - p}{d_n}\mathbb{1}\{1 \leq x\} = \mathbb{1}\{1 \leq x\}, \quad x \in \mathbb{R}.$$

In the following, we denote the covariance matrix of $\mathbf{\Theta}^{(n)}$ as

$$\mathbf{\Sigma}^{(n)} \coloneqq \mathrm{Cov}(\mathbf{\Theta}^{(n)})$$

whereas $\mathbf{\Gamma}^{(n)}$ is the covariance matrix of the non-standardized directional component ${\mathbf{\Gamma}^{(n)}}^{1/2}\boldsymbol{V}^{(n)}$. Not only $\mathbf{\Gamma}_n$ has the eigenvalues $\xi_{n,1}, \ldots, \xi_{n,p^*}$ but $\mathbf{\Gamma}^{(n)}$ has likewise these eigenvalues. Additionally, $\mathbf{\Gamma}^{(n)}$ has $(d_n - p^*)$-times the eigenvalue 1 which we denote as well as $\xi_{n,p^*+1}, \ldots, \xi_{n,d_n}$. We are still in the setup of the last section because not only the eigenvalues of $\mathbf{\Gamma}^{(n)}$ satisfy the spiked covariance structure

$$\xi_{n,1} \geq \cdots \geq \xi_{n,p^*} > 1 = \xi_{n,p^*+1} = \cdots = \xi_{n,d_n}$$

in (4.1) but as well the eigenvalues of $\mathbf{\Sigma}^{(n)}$ satisfy the structure in (4.1) although $\mathbf{\Sigma}^{(n)}$ has different eigenvalue than $\mathbf{\Gamma}^{(n)}$.

Lemma 4.12. *Suppose $(\boldsymbol{X}^{(n)})_{n\in\mathbb{N}}$ follows the directional model and $\lambda_{n,1} \geq \cdots \geq \lambda_{n,d_n}$ are the ordered eigenvalues of $\mathbf{\Sigma}^{(n)} = \mathrm{Cov}(\mathbf{\Theta}^{(n)})$. Then*

$$\lambda_{n,p^*} > \lambda_{n,p^*+1} = \cdots = \lambda_{n,d_n}.$$

Hence, there is a spike after the p^*-th eigenvalue λ_{n,p^*} of $\mathbf{\Sigma}^{(n)}$ and the eigenvalues λ_{n,p^*+1}, $\ldots, \lambda_{n,d_n-1}$ are all equal, as required in the definition of the spiked covariance model in (4.1). Further, the dimension of the low-dimensional structure is the same as for the fixed dimensional case. We summarize the model in the following.

Model D.

(D1) Let $\boldsymbol{X}^{(n)}, \boldsymbol{X}_1^{(n)}, \boldsymbol{X}_2^{(n)}, \ldots, \boldsymbol{X}_n^{(n)}$ be an i.i.d. sequence of d_n-dimensional random vectors satisfying the Directional Model with $\mathbb{E}[|V_1|^4] < \infty$.

(D2) The ordered eigenvalues $\xi_{n,1} \geq \cdots \geq \xi_{n,d_n}$ of $\mathbf{\Gamma}^{(n)}$ in (4.3) satisfy

$$\xi_{n,1} \geq \cdots \geq \xi_{n,p^*} > 1 = \xi_{n,p^*+1} = \cdots = \xi_{n,d_n},$$

whereas the ordered eigenvalues of $\mathbf{\Sigma}^{(n)}$ are denoted by $\lambda_{n,1} \geq \cdots \geq \lambda_{n,d_n}$.

(D3) Let $(k_n)_{n\in\mathbb{N}}$ be a sequence in $\mathbb{N}$ with $k_n \to \infty$, $k_n/n \to 0$ and

$$d_n/k_n \to c > 0, \quad \text{as } n \to \infty.$$

Remark 4.13.

(a) The assumption $d_n/k_n \to c > 0$ as $n \to \infty$ guarantees that the dimension d_n increases with a rate similar to the number of extremes k_n.

(b) Due to Lemma 4.12, Model D is also a spiked covariance model but it is a special type of spiked covariance, namely, a directional model, where the dimensionality parameter p^* is independent of n although the dimension d_n depends on n.

(c) Eigenvalues which are larger than $1 + \sqrt{c}$, are *called distant spiked eigenvalues*, whereby the asymptotic behavior of the corresponding empirical eigenvalues changes if they are above or below $1 + \sqrt{c}$ (cf. Section 4.1.2). Due to Silverstein and Choi (1995, Theorem 4.1 and Theorem 4.2), the assumption $\xi_{n,p^*} > 1 + \sqrt{c}$ is equivalent to $\varphi_c'(\xi_{n,p^*}) > 0$ where

$$\varphi_c(x) \coloneqq x\left(1 + c\int \frac{t}{x-t}\,\mathrm{d}\delta_1(t)\right) = x\left(1 + \frac{c}{x-1}\right). \tag{4.4}$$

(d) The convergence in (D3) is analogous to the assumption $\widehat{s}_n/k_n \xrightarrow{\mathbb{P}} c \in [0,1)$ in Section 3.3. In both settings we assume that the dimension of the underlying space and the number of categories of the multinomial distribution, respectively, grow proportionally to k_n.

Analog to (4.2) we define the $d_n \times d_n$ empirical covariance matrix of $\mathbf{\Sigma}^{(n)}$ as

$$\widehat{\mathbf{\Sigma}}^{(n)} \coloneqq \frac{1}{k_n}\sum_{j=1}^{n}\left(\frac{\boldsymbol{X}_j^{(n)}}{\|\boldsymbol{X}_j^{(n)}\|_2} - \widehat{\mathbf{\Theta}}^{(n)}\right)\cdot\left(\frac{\boldsymbol{X}_j^{(n)}}{\|\boldsymbol{X}_j^{(n)}\|_2} - \widehat{\mathbf{\Theta}}^{(n)}\right)^{\top} \mathbb{1}\{\|\boldsymbol{X}_j^{(n)}\|_2 > \|\boldsymbol{X}_{(k_n+1,n)}^{(n)}\|_2\}, \tag{4.5}$$

with eigenvalues $\widehat{\lambda}_{n,1}, \ldots, \widehat{\lambda}_{n,d_n}$, where

$$\widehat{\mathbf{\Theta}}^{(n)} \coloneqq \frac{1}{k_n}\sum_{i=1}^{n}\frac{\boldsymbol{X}_i^{(n)}}{\|\boldsymbol{X}_i^{(n)}\|_2}\mathbb{1}\{\|\boldsymbol{X}_i^{(n)}\|_2 > \|\boldsymbol{X}_{(k_n+1,n)}^{(n)}\|_2\}.$$

In contrast to the empirical covariance matrix $\widehat{\mathbf{\Sigma}}_n$ in (4.2) with a fixed dimension $d \times d$, the dimension of the empirical covariance matrix $\widehat{\mathbf{\Sigma}}^{(n)}$ in (4.5) is $d_n \times d_n$ and hence, growing in n.

Let us first present the asymptotic distribution of the eigenvalue $\widehat{\lambda}_{n,1}, \ldots, \widehat{\lambda}_{n,d_n}$ of $\widehat{\mathbf{\Sigma}}^{(n)}$ under the constraint that $\mathbf{\Gamma}_n$ and its eigenvalues $\xi_{n,1}, \ldots, \xi_{n,p^*}$ are converging, and afterwards when $\xi_{n,p^*} \to \infty$. Recall that the MP law F_c (cf. Definition 4.2) has density

$$f_c(x) = \begin{cases} \frac{1}{2\pi x c}\sqrt{((1+\sqrt{c})^2 - x)(x - (1-\sqrt{c})^2)}, & x \in ((1-\sqrt{c})^2, (1+\sqrt{c})^2), \\ 0, & \text{otherwise}, \end{cases}$$

and point mass $1 - 1/c$ at 0 if $c > 1$.

Theorem 4.14. *Let Model D be given. Suppose that $\mathbf{\Gamma}_n \to \mathbf{\Gamma}$ and $(\xi_{n,1}, \ldots, \xi_{n,p^*}) \to (\xi_1, \ldots, \xi_{p^*})$ as $n \to \infty$ with $\xi_{p^*} > 1 + \sqrt{c}$.*

(a) Let $i \in \{1, \ldots, p^*\}$. Then the asymptotic behavior

$$d_n \widehat{\lambda}_{n,i} \xrightarrow{\mathbb{P}} \varphi_c(\xi_i), \qquad \text{as } n \to \infty$$

holds, where φ_c is defined as in (4.4).

(b) Let $(i_n(\alpha))_{n\in\mathbb{N}}$ be a sequence in $\mathbb{N}$ with $i_n(\alpha) > p^*$ and $i_n(\alpha)/d_n \to \alpha \in [0,1]$ for any $\alpha \in (0,1)$. Then

$$\sup_{\alpha\in(0,1)} \left| d_n \widehat{\lambda}_{n,i_n(\alpha)} - F_c^{\leftarrow}(1-\alpha) \right| \xrightarrow{\mathbb{P}} 0, \qquad \text{as } n \to \infty,$$

where $F_c^{\leftarrow}$ is the generalized inverse of F_c. In particular, if $(q_n)_{n\in\mathbb{N}}$ is a sequence in $\mathbb{N}$ with $q_n = o(d_n)$ and $q_n > p^*$, then $d_n \widehat{\lambda}_{n,q_n} \xrightarrow{\mathbb{P}} (1+\sqrt{c})^2$.

(c) Suppose $0 < c \leq 1$ and $(q_n)_{n\in\mathbb{N}}$ is a sequence in $\mathbb{N}$ with $q_n = o(d_n)$ as $n \to \infty$. Then as $n \to \infty$,

$$\frac{1}{d_n - q_n} \sum_{i=q_n+1}^{d_n} d_n \widehat{\lambda}_{n,i} \xrightarrow{\mathbb{P}} 1.$$

(d) Suppose $c > 1$ and $(q_n)_{n\in\mathbb{N}}$ is a sequence in $\mathbb{N}$ with $q_n = o(d_n)$ as $n \to \infty$. Then as $n \to \infty$,

$$\frac{1}{k_n - q_n} \sum_{i=q_n+1}^{k_n} d_n \widehat{\lambda}_{n,i} \xrightarrow{\mathbb{P}} c.$$

So far we have assumed that the first p^* eigenvalues $\xi_{n,1}, \ldots, \xi_{n,p^*}$ of $\mathbf{\Gamma}^{(n)}$ are bounded. Alternatively, it is also possible to suppose that $\xi_{n,p^*} \to \infty$ as $n \to \infty$.

Theorem 4.15. *Let Model D be given. Suppose $\xi_{n,p^*} \to \infty$ and $\xi_{n,1} = o(d_n^{1/2})$ as $n \to \infty$.*

(a) Let $i \in \{1, \ldots, p^*\}$. Then the asymptotic behavior

$$d_n \widehat{\lambda}_{n,i} / \xi_{n,i} \xrightarrow{\mathbb{P}} 1, \qquad \text{as } n \to \infty$$

holds.

(b) Let $(i_n(\alpha))_{n\in\mathbb{N}}$ *be a sequence in* $\mathbb{N}$ *with* $i_n(\alpha) > p^*$ *and* $i_n(\alpha)/d_n \to \alpha \in [0,1]$ *for any* $\alpha \in (0,1)$*. Then*

$$\sup_{\alpha\in(0,1)} \left| d_n \widehat{\lambda}_{n,i_n(\alpha)} - F_c^{\leftarrow}(1-\alpha) \right| \xrightarrow{\mathbb{P}} 0, \qquad \text{as } n \to \infty,$$

where $F_c^{\leftarrow}$ *is defined as in Theorem 4.14. In particular, if* $(q_n)_{n\in\mathbb{N}}$ *is a sequence in* $\mathbb{N}$ *with* $q_n = o(d_n)$ *and* $q_n > p^*$*, then* $d_n \widehat{\lambda}_{n,q_n} \xrightarrow{\mathbb{P}} (1+\sqrt{c})^2$*.*

(c) Suppose $0 < c \leq 1$ *and* $(q_n)_{n\in\mathbb{N}}$ *is a sequence in* $\mathbb{N}$ *with* $q_n = o(d_n)$ *as* $n \to \infty$*. Then as* $n \to \infty$*,*

$$\frac{1}{d_n - q_n} \sum_{i=q_n+1}^{d_n} d_n \widehat{\lambda}_{n,i} \xrightarrow{\mathbb{P}} 1.$$

(d) Suppose $c > 1$ *and* $(q_n)_{n\in\mathbb{N}}$ *is a sequence in* $\mathbb{N}$ *with* $q_n = o(d_n)$ *as* $n \to \infty$*. Then as* $n \to \infty$*,*

$$\frac{1}{k_n - q_n} \sum_{i=q_n+1}^{k_n} d_n \widehat{\lambda}_{n,i} \xrightarrow{\mathbb{P}} c.$$

(e) Suppose $0 < c < 1$ *and let* $i \in \{1, \ldots, p^*\}$*. Then as* $n \to \infty$*,*

$$\frac{d_n \widehat{\lambda}_{n,i}}{\frac{1}{d_n - i} \sum_{j=i+1}^{d_n} d_n \widehat{\lambda}_{n,j}} \xrightarrow{\mathbb{P}} \infty.$$

Remark 4.16. The assumption $\xi_{n,1} = o(d_n^{1/2})$ as $n \to \infty$ guarantees that the largest eigenvalue grows sufficiently slowly compared to the dimension d_n. When all moments of V_1 exist this assumption can be relaxed to $\xi_{n,1} = o(d_n^{\beta})$ as $n \to \infty$ for any $\beta < 1$ due to Remark 4.33.

4.3. Information criteria in the fixed-dimensional case

The aim of this section is to derive estimators for p^*, the location of the spike in the eigenvalues of $\mathbf{\Sigma} = \mathrm{Cov}(\mathbf{\Theta})$, which defines the dimensionality of the PCA, by exploiting information criteria. In the context of PCA for Gaussian data, an Akaike information criteria (AIC) and a Bayesian information criteria (BIC) were developed in Fujikoshi and Sakurai (2016) and the consistency was analyzed in the high-dimensional case for general data in Bai et al. (2018). In this chapter, we adopt these information criteria and study their statistical properties. In this section, we start with the fixed-dimensional case and give the proper definitions of the information criteria under Model S. The proofs of this section are moved to Section 4.5.2.

Definition 4.17. Suppose $\widehat{\lambda}_{n,1}, \ldots, \widehat{\lambda}_{n,d-1}$ are the empirical eigenvalues of $\widehat{\mathbf{\Sigma}}_n$ as defined in (4.2).

(a) The *Akaike information criterion* (AIC) for the fixed-dimensional case is defined as

$$\begin{aligned}\mathrm{AIC}_{k_n}(p) :=& k_n \sum_{i=1}^{p} \log(\widehat{\lambda}_{n,i}) + k_n(d-1-p)\log\Big(\frac{1}{d-1-p}\sum_{j=p+1}^{d-1}\widehat{\lambda}_{n,j}\Big) \\ &+ k_n \log\Big(\frac{k_n-1}{k_n}\Big)^{d-1} + k_n(d-1)(\log(2\pi)+1) \\ &+ 2(p+1)(d-p/2),\end{aligned}$$

for $p = 1, \ldots, d-2$ and an estimator for p^* is $\widehat{p}_n := \arg\min_{1\le p\le d-2} \mathrm{AIC}_{k_n}(p)$.

(b) The *Bayesian information criterion* (BIC) for the fixed-dimensional case is defined as

$$\begin{aligned}\mathrm{BIC}_{k_n}(p) :=& k_n \sum_{i=1}^{p} \log(\widehat{\lambda}_{n,i}) + k_n(d-1-p)\log\Big(\frac{1}{d-1-p}\sum_{j=p+1}^{d-1}\widehat{\lambda}_{n,j}\Big) \\ &+ k_n \log\Big(\frac{k_n-1}{k_n}\Big)^{d-1} + k_n(d-1)(\log(2\pi)+1) \\ &+ \log(k_n)(p+1)(d-\frac{p}{2}/2),\end{aligned}$$

for $p = 1, \ldots, d-2$ and an estimator for p^* is $\widehat{p}_n := \arg\min_{1\le p\le d-2} \mathrm{BIC}_{k_n}(p)$.

Remark 4.18.

(a) The penalty $(p+1)(d-p/2) = (p+1)d - p(1+p)/2$ arises as it is the number of parameters that define a $(d-1)$-dimensional normal distribution with an arbitrary mean vector and covariance matrix following the p-th spiked covariance model (cf. Fujikoshi and Sakurai, 2016, Section 2). As baseline model, we take a $(d-1)$-dimensional normal distribution instead of a d-dimensional normal distribution because $\boldsymbol{\Theta}$ is a random vector on the unit sphere and hence, potentially the first $(d-1)$ components already determine the last component. In summary, we use a modified version of the AIC and the BIC of Fujikoshi and Sakurai (2016) by replacing d with $d-1$ and dropping the last empirical eigenvalue $\widehat{\lambda}_{n,d}$.

(b) The AIC and BIC are invariant to scaling of the eigenvalues. Consequently, scaling the sample covariance matrix $\widehat{\boldsymbol{\Sigma}}_n$, or equivalently the eigenvalues $\widehat{\lambda}_{n,1}, \ldots, \widehat{\lambda}_{n,d-1}$, does not affect the point at which the information criteria achieve their minimum.

Next, we check the consistency of the AIC and the BIC. First, we present the result for the AIC and second for the BIC.

Theorem 4.19. *Let Model S be given and $\boldsymbol{M}$ be the limit vector in Theorem 4.10. Then*

$$\lim_{n\to\infty} \mathbb{P}(\mathrm{AIC}_{k_n}(p) > \mathrm{AIC}_{k_n}(p^*)) = \begin{cases} \mathbb{P}(g_p(\boldsymbol{M}) > 0) & \text{for } p > p^*, \\ 1 & \text{for } p < p^*, \end{cases}$$

where

$$g_p(\boldsymbol{m}) \coloneqq -\frac{1}{2}\sum_{i=p^*+1}^{p} m_i^2 - \frac{1}{2(d-1-p)}\left(\sum_{j=p+1}^{d-1} m_j\right)^2 + \frac{1}{2(d-1-p^*)}\left(\sum_{j=p^*+1}^{d-1} m_j\right)^2 - (d-p-2)(d-p+1) + (d-p^*-2)(d-p^*+1)$$

for $\boldsymbol{m} = (m_1, \ldots, m_d) \in \mathbb{R}^d$.

Remark 4.20. Under some technical assumptions on the distribution of $\boldsymbol{\Theta}$, it is possible to state a density for $\boldsymbol{M}$ (cf. Davis, 1977) and derive that $\mathbb{P}(g_p(\boldsymbol{M}) > 0) < 1$. For the present chapter, it is sufficient to give an example such that the AIC is not consistent.

Example 4.21. We assume the Directional Model, where $\boldsymbol{\Gamma}^{(n)} := \boldsymbol{\Gamma} := \operatorname{diag}(9, 4, 4, 1)$, $\boldsymbol{V}^{(n)} := \boldsymbol{V} := (V_1, V_2, V_3, V_4)^\top$ with $V_i \sim \mathcal{U}(\{-1, 1\})$, $i = 1, \ldots, 4$, $Z \sim$ Fréchet(1) and the dimension $d = 4$ is fixed. Then, we have $\|\boldsymbol{\Gamma}^{(n)^{1/2}}\boldsymbol{V}^{(n)}\|_2 = \|\boldsymbol{\Gamma}\boldsymbol{V}\|_2 = \sqrt{9+4+4+1} = \sqrt{18}$ and $\boldsymbol{\Theta}^{(n)} := \boldsymbol{\Theta} := (3V_1, 2V_2, 2V_3, V_4)/\sqrt{18}$. We have $\mathbb{E}[\boldsymbol{\Theta}] = \boldsymbol{0}_4$, $\boldsymbol{\Sigma} = \boldsymbol{\Gamma}/18$, where the eigenvalues of $\boldsymbol{\Sigma}$ are $(1/2, 2/9, 2/9, 1/18)$ and the corresponding eigenvectors are the unit vectors $\boldsymbol{e}_1, \ldots, \boldsymbol{e}_4 \in \mathbb{R}^4$. Consequently, the spiked covariance assumption is satisfied with $\lambda = 2/9, d = 4$ and $p^* = 1$.

In the following, we calculate the probability $\mathbb{P}(g_2(\boldsymbol{M}) < 0)$ by first determining the asymptotic distribution of $\boldsymbol{M}$. An application of Theorem 4.10 (b) yields

$$\sqrt{k_n}((\widehat{\lambda}_{n,2}, \widehat{\lambda}_{n,3}) - (\lambda, \lambda)) \xrightarrow{\mathcal{D}} (M_2, M_3)$$

in $\mathbb{R}^2$ is the joint distribution of the decreasingly ordered non-null eigenvalues of

$$\boldsymbol{P}_\lambda \boldsymbol{S} \boldsymbol{P}_\lambda = \begin{pmatrix} 0 & 0 & 0 & 0 \\ 0 & 1 & 0 & 0 \\ 0 & 0 & 1 & 0 \\ 0 & 0 & 0 & 0 \end{pmatrix} \begin{pmatrix} S_{11} & S_{12} & S_{13} & S_{14} \\ S_{12} & S_{22} & S_{23} & S_{24} \\ S_{13} & S_{23} & S_{33} & S_{34} \\ S_{14} & S_{24} & S_{34} & S_{44} \end{pmatrix} \begin{pmatrix} 0 & 0 & 0 & 0 \\ 0 & 1 & 0 & 0 \\ 0 & 0 & 1 & 0 \\ 0 & 0 & 0 & 0 \end{pmatrix} = \begin{pmatrix} 0 & 0 & 0 & 0 \\ 0 & S_{22} & S_{23} & 0 \\ 0 & S_{23} & S_{33} & 0 \\ 0 & 0 & 0 & 0 \end{pmatrix},$$

where $\operatorname{vec}(\boldsymbol{S})$ follows a centered multivariate normal distribution with covariance $\operatorname{Cov}(\operatorname{vec}(\boldsymbol{\Theta} - \mathbb{E}[\boldsymbol{\Theta}])(\boldsymbol{\Theta} - \mathbb{E}[\boldsymbol{\Theta}])^\top))$ and $\boldsymbol{P}_\lambda := (\boldsymbol{e}_2, \boldsymbol{e}_3) \cdot (\boldsymbol{e}_2, \boldsymbol{e}_3)^\top \in \mathbb{R}^{4\times 4}$ is the projection onto the 2-dimensional eigenspace of the orthonormal eigenvectors $\boldsymbol{e}_2, \boldsymbol{e}_3$ corresponding to $\lambda = 2/9$. Since $\mathbb{V}(S_{22}) = \mathbb{E}[\Theta_2^4] - (\mathbb{E}[\Theta_2^2])^2 = 0$ and $\mathbb{V}(S_{33}) = 0$, the distributions of S_{22} and S_{33} are degenerate with expectation zero. By the symmetry of $\boldsymbol{P}_\lambda \boldsymbol{S} \boldsymbol{P}_\lambda$, the non-null eigenvalues of the matrix $\boldsymbol{P}_\lambda \boldsymbol{S} \boldsymbol{P}_\lambda$ can be calculated directly and are given by

$$M_2 = S_{23} \qquad \text{and} \qquad M_3 = -S_{23}.$$

Next, since $(d - p^* - 2)(d - p^* + 1) - (d - p - 2)(d - p + 1) = 4$ for $p = 2$ and $p^* = 1$, the

inequality $g_2(\boldsymbol{M}) < 0$ is equivalent to

$$4 < \frac{1}{2}M_2^2 + \frac{1}{2}M_3^2 - \frac{1}{4}(M_2 + M_3)^2 = S_{23}^2.$$

Due to the definition of $\boldsymbol{S}$, the distribution of S_{23} is Gaussian with expectation zero and $\mathbb{V}(S_{23}) = \mathbb{E}[\Theta_2^2\Theta_3^2] = 1$ so that $\mathbb{P}(g_2(\boldsymbol{M}) < 0) > 0$.

In contrast to the AIC, the BIC is a weakly consistent information criterion and selects the true dimension p^* with probability converging to 1 as stated in the next theorem.

Theorem 4.22. *Let Model S be given. Then*

$$\lim_{n\to\infty} \mathbb{P}(\mathrm{BIC}_{k_n}(p) > \mathrm{BIC}_{k_n}(p^*)) = 1 \quad \textit{for } p \neq p^*.$$

The non-consistency of the AIC and the consistency of the BIC are typical for these information criteria in the fixed-dimensional case similar to the results in Chapter 3. In the high-dimensional case, the asymptotic properties, derived in the next section, differ. In view of question **(Q)**, we can conclude that in theory the BIC reliably reduced the dimension.

4.4. Information criteria in the high-dimensional case

The topic in this section is information criteria in the high-dimensional case of Model D, where $d = d_n$ depends on n and $d_n/k_n \to c > 0$ as $n \to \infty$. For the definition of the information criteria and the asymptotic properties, we need to differentiate between the cases $c < 1$ and $c > 1$. The reason behind it is that if $d_n > k_n$, the last $d_n - k_n$ empirical eigenvalues of $\widehat{\boldsymbol{\Sigma}}^{(n)}$ are equal to zero, i.e. $\widehat{\lambda}_{n,k_n+1} = \cdots = \widehat{\lambda}_{n,d_n} = 0$. Therefore, in Section 4.4.1, we analyze the information criteria for $0 < c < 1$ and in Section 4.4.2 for $c > 1$. The proofs of this section are provided in Section 4.5.3.

4.4.1. Information criteria for $0 < c < 1$

In the case $0 < c < 1$, the definition of the information criteria are similar to the fixed-dimensional setting but we would like to point out that in the high-dimensional setting, we do not necessarily evaluate the information criteria at all possible values $1, \ldots, d_n - 1$ but rather restrict to $1, \ldots, q_n$ with $q_n \leq d_n$. The number q_n is called the number of candidate dimensions.

Definition 4.23. Suppose $\widehat{\lambda}_{n,1}, \ldots, \widehat{\lambda}_{n,d_n-1}$ are the empirical eigenvalues of $\widehat{\boldsymbol{\Sigma}}^{(n)}$ as defined in (4.5) and let $q_n \leq d_n - 2$.

(a) The *Akaike information criterion* (AIC°) for the high-dimensional case with $d_n < k_n$

is defined as

$$\begin{aligned}\mathrm{AIC}^{\circ}_{k_n}(p) \coloneqq \sum_{i=1}^{p} \log(\widehat{\lambda}_{n,i}) + (d_n - 1 - p) \log\Big(\frac{1}{d_n - 1 - p} \sum_{j=p+1}^{d_n - 1} \widehat{\lambda}_{n,j}\Big) \\ + \log\left(\frac{k_n - 1}{k_n}\right)^{d_n - 1} + (d_n - 1)(\log(2\pi) + 1) + \frac{(p+1)(2d_n - p)}{k_n},\end{aligned}$$

for $p = 1, \ldots, d_n - 2$ and an estimator for p^* is $\widehat{p}_n \coloneqq \arg\min_{1 \le p \le q_n} \mathrm{AIC}^{\circ}_{k_n}(p)$.

(b) The *Bayesian information criterion* (BIC°) for the high-dimensional case with $d_n < k_n$ is defined as

$$\begin{aligned}\mathrm{BIC}^{\circ}_{k_n}(p) \coloneqq \sum_{i=1}^{p} \log(\widehat{\lambda}_{n,i}) + (d_n - 1 - p) \log\Big(\frac{1}{d_n - 1 - p} \sum_{j=p+1}^{d_n - 1} \widehat{\lambda}_{n,j}\Big) \\ + \log\left(\frac{k_n - 1}{k_n}\right)^{d_n - 1} + (d_n - 1)(\log(2\pi) + 1) \\ + \log(k_n) \frac{(p+1)(d_n - p/2)}{k_n},\end{aligned}$$

for $p = 1, \ldots, d_n - 2$ and an estimator for p^* is $\widehat{p}_n \coloneqq \arg\min_{1 \le p \le q_n} \mathrm{BIC}^{\circ}_{k_n}(p)$.

In the next theorem, we present sufficient assumptions for the AIC° to be weakly consistent, i.e.,

$$\lim_{n \to \infty} \mathbb{P}\Big(\operatorname*{arg\,min}_{1 \le p < q_n} \mathrm{AIC}^{\circ}_{k_n}(p) = p^* \Big) = 1$$

and afterwards for the BIC°.

Theorem 4.24. *Let Model D with $0 < c < 1$ be given and let the number q_n of candidate dimensions satisfy $q_n = o(d_n)$ as $n \to \infty$.*

(a) Suppose $\mathbf{\Gamma}_n \to \mathbf{\Gamma}$ and $(\xi_{n,1}, \ldots, \xi_{n,p^*}) \to (\xi_1, \ldots, \xi_{p^*})$ as $n \to \infty$ with $\xi_{p^*} > 1 + \sqrt{c}$. If the gap condition

$$\varphi_c(\xi_{p^*}) - 1 - \log(\varphi_c(\xi_{p^*})) - 2c > 0 \tag{4.6}$$

with φ_c as defined in (4.4) holds, then the AIC° is weakly consistent.

(b) Suppose $\mathbf{\Gamma}_n \to \mathbf{\Gamma}$ and $(\xi_{n,1}, \ldots, \xi_{n,p^*}) \to (\xi_1, \ldots, \xi_{p^*})$ as $n \to \infty$ with $\xi_{p^*} > 1 + \sqrt{c}$. If the gap condition (4.6) does not hold, then

$$\lim_{n \to \infty} \mathbb{P}\Big(\min_{1 \le p < p^*} \Big\{ \mathrm{AIC}^{\circ}_{k_n}(p) - \mathrm{AIC}^{\circ}_{k_n}(p^*) \Big\} > 0 \Big) < 1$$

and the AIC° is not weakly consistent.

(c) Suppose $\xi_{n,p^*} \to \infty$ and $\xi_{n,1} = o(d_n^{1/2})$ as $n \to \infty$. Then the AIC° is weakly consistent.

Remark 4.25.

(a) The division of AIC° by k_n in contrast to the AIC has no influence in applications, as it does not affect the location of the minimum of the information criteria for a fixed sample size n. As a result, in the simulation study, the minima of AIC and AIC° coincide, and we do not need to distinguish between these criteria. The division by k_n in the definition of AIC°, as in Bai et al. (2018), ensures that the limit of the information criteria exists.

(b) The gap condition (4.6) was introduced in Bai et al. (2018) and it also guarantees that the gap between ξ_{p^*} and the non-dominant eigenvalues is sufficiently large.

In the following theorem, consistency criteria for the BIC° are stated, which are slightly different from the results for the AIC°.

Theorem 4.26. *Let Model D with $0 < c < 1$ be given. Suppose that either*

$$\mathbf{\Gamma}_n \to \mathbf{\Gamma} \text{ such that } (\xi_{n,1}, \ldots, \xi_{n,p^*}) \to (\xi_1, \ldots, \xi_{p^*}) \text{ as } n \to \infty \text{ with } \xi_{p^*} > 1 + \sqrt{c},$$

or

$$\xi_{n,p^*} \to \infty \quad \text{and} \quad \xi_{n,1} = o(d_n^{1/2}) \quad \text{as } n \to \infty.$$

(a) If $\xi_{n,p^*} / \log(d_n) \to 0$ as $n \to \infty$, then

$$\lim_{n\to\infty} \mathbb{P}\Big(\min_{1 \le p < p^*} \Big\{ \mathrm{BIC}^\circ_{k_n}(p) - \mathrm{BIC}^\circ_{k_n}(p^*) \Big\} > 0 \Big) < 1$$

and the BIC° is not weakly consistent.

(b) If $\xi_{n,p^*} / \log(d_n) \to \infty$ as $n \to \infty$, then the BIC° is weakly consistent.

Remark 4.27.

(a) When the gap condition is fulfilled, the AIC° is weakly consistent whereas the consistency of the BIC° depends on the properties of ξ_{n,p^*}. The BIC° and, if the gap condition is violated, the AIC°, tends to underestimate the number of significant principal components. A similar result was also obtained by Bai et al. (2020) for multivariate linear regressions in high dimensions.

(b) The consistency of the AIC° and BIC° in the high-dimensional case is opposite to the fixed-dimensional case. Specifically, while the AIC may not be consistent and the BIC is consistent in the fixed-dimensional setting, the opposite behavior is observed in the high-dimensional setting, similar to Chapter 3. Moreover, in Theorem 4.24 (b), we have

$$\lim_{n\to\infty} \mathbb{P}\Big(\min_{1 \le p < p^*} \Big\{ \mathrm{AIC}^\circ_{k_n}(p) - \mathrm{AIC}^\circ_{k_n}(p^*) \Big\} > 0 \Big) < 1,$$

which is opposite to the fixed-dimensional case, where the AIC tends to overestimate rather than underestimate the number of principal components.

(c) The case $c = 1$ is excluded from the consideration due to potential complications with the asymptotic behavior of the eigenvalues (cf. Bai et al., 2018, Section 4). While Theorem 4.14 and Theorem 4.15 are valid for $c = 1$, issues arise with the convergence of ratios of quantiles of the Marčenko-Pastur law in Bai et al. (2018, Lemma 2.3) when $q_n = o(d_n)$ is not assumed. If $q_n = o(d_n)$ is assumed, then the results for $0 < c < 1$ also apply to $c = 1$. Additionally, the support of the Marčenko-Pastur law for $c = 1$ is given by the interval $(0, 4)$, which can lead to empirical eigenvalues close to zero, causing numerical problems when calculating the logarithm of the empirical eigenvalues.

(d) If $\lim_{n\to\infty} \xi_{n,p^*}/\log(d_n) \in (0, \infty)$ further assumptions are needed to assess the consistency of the BIC°.

(e) In view of the high-dimensional case in Chapter 3, if $\xi_{n,p^*} \to \infty$, we are in a similar setting to $T_{n,s^*}(k_n) \xrightarrow{\mathbb{P}} \infty$ where both AIC and BIC are consistent.

4.4.2. INFORMATION CRITERIA FOR $c > 1$

For the case $c > 1$ we have to adapt the information criteria. Therefore, we follow the definition of the AIC and the BIC in Bai et al. (2018), which leads to the following definition.

Definition 4.28. Suppose $\widehat{\lambda}_{n,1}, \ldots, \widehat{\lambda}_{n,d_n-1}$ are the empirical eigenvalues of $\widehat{\mathbf{\Sigma}}^{(n)}$ as defined in (4.5) and let $q_n \leq k_n - 2$.

(a) The *Akaike information criterion* (AIC*) for the high-dimensional case with $d_n > k_n$ is defined as

$$\begin{aligned}\mathrm{AIC}^*_{k_n}(p) \coloneqq \sum_{i=1}^{p} \log(\widehat{\lambda}_{n,i}) + (k_n - 1 - p) \log\Big(\frac{1}{k_n - 1 - p} \sum_{j=p+1}^{k_n-1} \widehat{\lambda}_{n,j}\Big) \\ + \log\left(\frac{d_n - 1}{d_n}\right)^{k_n-1} + (k_n - 1)(\log(2\pi) + 1) + \frac{(p+1)(2k_n - p)}{d_n},\end{aligned}$$

for $p = 1, \ldots, k_n - 2$ and an estimator for p^* is $\widehat{p}_n \coloneqq \arg\min_{1\leq p\leq q_n} \mathrm{AIC}^*_{k_n}(p)$.

(b) The *Bayesian information criterion* (BIC*) for the high-dimensional case with $d_n > k_n$ is defined as

$$\begin{aligned}\mathrm{BIC}^*_{k_n}(p) \coloneqq \sum_{i=1}^{p} \log(\widehat{\lambda}_{n,i}) + (k_n - 1 - p) \log\Big(\frac{1}{k_n - 1 - p} \sum_{j=p+1}^{k_n-1} \widehat{\lambda}_{n,j}\Big) \\ + \log\left(\frac{d_n - 1}{d_n}\right)^{k_n-1} + (k_n - 1)(\log(2\pi) + 1) \\ + \log(d_n)\frac{(p+1)(k_n - p/2)}{d_n},\end{aligned}$$

for $p = 1, \ldots, k_n - 2$ and an estimator for p^* is $\widehat{p}_n := \arg\min_{1 \leq p \leq q_n} \mathrm{BIC}^*_{k_n}(p)$.

For the consistency analysis of the AIC^* and BIC^* we use the same definition for weakly consistent as for the AIC° in Section 4.4.1.

Theorem 4.29. *Let Model D with $c > 1$ be given and let the number q_n of candidate dimensions satisfy $q_n = o(d_n)$ as $n \to \infty$.*

(a) Suppose $\mathbf{\Gamma}_n \to \mathbf{\Gamma}$ and $(\xi_{n,1}, \ldots, \xi_{n,p^*}) \to (\xi_1, \ldots, \xi_{p^*})$ as $n \to \infty$ with $\xi_{p^*} > 1 + \sqrt{c}$. If the modified gap condition

$$\frac{\varphi_c(\xi_{n,p^*})}{c} - 1 - \log\left(\frac{\varphi_c(\xi_{n,p^*})}{c}\right) - \frac{2}{c} > 0 \tag{4.7}$$

with φ_c as defined in (4.4) holds, then the AIC^* is weakly consistent.

(b) Suppose $\mathbf{\Gamma}_n \to \mathbf{\Gamma}$ and $(\xi_{n,1}, \ldots, \xi_{n,p^*}) \to (\xi_1, \ldots, \xi_{p^*})$ as $n \to \infty$ with $\xi_{p^*} > 1 + \sqrt{c}$. If the modified gap condition (4.7) does not hold, then the AIC^* is not weakly consistent.

(c) Suppose that $\xi_{n,p^*} \to \infty$ and $\xi_{n,1} = o(d_n^{1/2})$ as $n \to \infty$. Then the AIC^* is weakly consistent.

Theorem 4.30. *Let Model D with $c > 1$ be given. Suppose that either*

$$\mathbf{\Gamma}_n \to \mathbf{\Gamma} \text{ such that } (\xi_{n,1}, \ldots, \xi_{n,p^*}) \to (\xi_1, \ldots, \xi_{p^*}) \text{ as } n \to \infty \text{ with } \xi_{p^*} > 1 + \sqrt{c},$$

or

$$\xi_{n,p^*} \to \infty \quad \text{and} \quad \xi_{n,1} = o(d_n^{1/2}) \quad \text{as } n \to \infty.$$

(a) If $\xi_{n,p^*} / \log(d_n) \to 0$ as $n \to \infty$, then the BIC^* is not weakly consistent.

(b) If $\xi_{n,p^*} / \log(d_n) \to \infty$ as $n \to \infty$, then the BIC^* is weakly consistent.

Remark 4.31.

(a) The AIC^* is weakly consistent when the gap condition is fulfilled and not consistent otherwise, whereas the consistency of the BIC^* depends on the asymptotic behavior of ξ_{n,p^*}. The results are identical to the case $0 < c < 1$.

(b) Since the last $(d_n - k_n)$ eigenvalues of $\widehat{\mathbf{\Sigma}}^{(n)}$ are equal to 0, additional simulation studies showed that if the dimension d_n is sufficiently large, setting some eigenvalues of $\mathbf{\Sigma}^{(n)}$ to zero has no big influence on the performance of the AIC^* and BIC^*. However, when $c < 1$, the zero eigenvalues do influence the performance of the AIC and BIC. In such cases, we recommend first projecting the data onto a lower-dimensional space to ensure that the zero eigenvalues have no impact on the analysis.

4.5. PROOFS

4.5.1. PROOFS OF SECTION 4.2

Proof of Theorem 4.10.

(a) We use Theorem A.46 in Bai and Silverstein (2010), which states that for Hermitian matrices $\boldsymbol{A}, \boldsymbol{B} \in \mathbb{R}^{d\times d}$ with eigenvalues $\lambda_i(\boldsymbol{A})$ and $\lambda_i(\boldsymbol{B})$, $i = 1, \ldots, d$, the inequality

$$\max_{i=1,\ldots,d} |\lambda_i(\boldsymbol{A}) - \lambda_i(\boldsymbol{B})| \leq \|\boldsymbol{A} - \boldsymbol{B}\|_2 \tag{4.8}$$

holds. A conclusion from Proposition 4.8 is that $\sqrt{k_n}(\widehat{\boldsymbol{\Sigma}}_n - \boldsymbol{\Sigma}) = O_{\mathbb{P}}(1)$ and therefore (4.8) yields

$$(\widehat{\lambda}_{n,1}, \ldots, \widehat{\lambda}_{n,d-1}) = (\lambda_1, \ldots, \lambda_{d-1}) + O_{\mathbb{P}}(1/\sqrt{k_n}).$$

(b) The result corresponds to Dauxois et al. (1982, Proposition 8), which is based on a similar convergence as Proposition 4.8.

□

Proof of Lemma 4.12. Note that

$$\boldsymbol{\Sigma}^{(n)} = \operatorname{Cov}(\boldsymbol{\Theta}^{(n)}) = {\boldsymbol{\Gamma}^{(n)}}^{1/2} \operatorname{Cov}\left(\frac{\boldsymbol{V}^{(n)}}{\|{\boldsymbol{\Gamma}^{(n)}}^{1/2}\boldsymbol{V}^{(n)}\|_2}\right) {\boldsymbol{\Gamma}^{(n)}}^{1/2}.$$

Utilizing the spectral decomposition $\boldsymbol{\Gamma}^{(n)} = \boldsymbol{W}^{(n)}\boldsymbol{D}^{(n)}{\boldsymbol{W}^{(n)}}^{\top}$, where $\boldsymbol{W}^{(n)} = (\boldsymbol{W}_1^{(n)}, \ldots, \boldsymbol{W}_{d_n}^{(n)})$ is a $d_n \times d_n$-dimensional orthogonal matrix and

$$\boldsymbol{D}^{(n)} \coloneqq \operatorname{diag}(\overline{\boldsymbol{D}}_n, \boldsymbol{I}_{d_n - p^*}) \coloneqq \operatorname{diag}(\xi_{n,1}, \ldots, \xi_{n,p^*}, 1, \ldots, 1) \in \mathbb{R}^{d_n \times d_n}$$

is a diagonal matrix consisting of the eigenvalues of $\boldsymbol{\Gamma}^{(n)}$, we receive with $\|\boldsymbol{W}^{(n)}\boldsymbol{x}\|_2 = \|\boldsymbol{x}\|_2$ for $\boldsymbol{x} \in \mathbb{R}^{d_n}$ that

$$\boldsymbol{\Sigma}^{(n)} = \operatorname{Cov}\left(\frac{{\boldsymbol{\Gamma}^{(n)}}^{1/2}\boldsymbol{V}^{(n)}}{\|{\boldsymbol{\Gamma}^{(n)}}^{1/2}\boldsymbol{V}^{(n)}\|_2}\right) = \boldsymbol{W}^{(n)} \operatorname{Cov}\left(\frac{{\boldsymbol{D}^{(n)}}^{1/2}\boldsymbol{V}^{(n)}}{\|{\boldsymbol{D}^{(n)}}^{1/2}\boldsymbol{V}^{(n)}\|_2}\right) {\boldsymbol{W}^{(n)}}^{\top}.$$

Hence, the matrices

$$\operatorname{Cov}\left(\frac{{\boldsymbol{\Gamma}^{(n)}}^{1/2}\boldsymbol{V}^{(n)}}{\|{\boldsymbol{\Gamma}^{(n)}}^{1/2}\boldsymbol{V}^{(n)}\|_2}\right) \quad \text{and} \quad \operatorname{Cov}\left(\frac{{\boldsymbol{D}^{(n)}}^{1/2}\boldsymbol{V}^{(n)}}{\|{\boldsymbol{D}^{(n)}}^{1/2}\boldsymbol{V}^{(n)}\|_2}\right)$$

are similar and share the same eigenvalues (Horn and Johnson, 2013, Theorem 1.3.22).

Therefore, we assume in the following w.l.o.g. that $\mathbf{\Gamma}^{(n)} = \boldsymbol{D}^{(n)}$ and hence,

$$\mathbf{\Sigma}^{(n)} = \mathrm{Cov}\left(\frac{\boldsymbol{D}^{(n)^{1/2}}\boldsymbol{V}^{(n)}}{\|\boldsymbol{D}^{(n)^{1/2}}\boldsymbol{V}^{(n)}\|_2}\right)$$

is a diagonal matrix. Indeed, since $\boldsymbol{D}^{(n)}$ is a diagonal matrix and $V_1, \ldots, V_{d_n}$ are symmetric and i.i.d., the components of $\boldsymbol{D}^{(n)^{1/2}}\boldsymbol{V}^{(n)}/\|\boldsymbol{D}^{(n)^{1/2}}\boldsymbol{V}^{(n)}\|_2$ are uncorrelated. Further, the eigenvalues of $\mathbf{\Sigma}^{(n)}$ are the diagonal entries

$$\mathrm{diag}(\mathbf{\Sigma}^{(n)})_i = \mathbb{E}\left[\frac{\xi_{n,i}V_i^2}{\|\mathbf{\Gamma}^{(n)^{1/2}}\boldsymbol{V}^{(n)}\|_2^2}\right] = \mathbb{E}\left[\frac{\xi_{n,i}V_i^2}{\sum_{j=1}^{p^*}\xi_{n,j}V_j^2 + \sum_{j=p^*+1}^{d_n}V_j^2}\right], \quad i = 1, \ldots, p^*$$

and

$$\mathrm{diag}(\mathbf{\Sigma}^{(n)})_i = \mathrm{diag}(\mathbf{\Sigma}^{(n)})_{d_n} = \mathbb{E}\left[\frac{V_{d_n}^2}{\sum_{j=1}^{p^*}\xi_{n,j}V_j^2 + \sum_{j=p^*+1}^{d_n}V_j^2}\right], \quad i = p^*+1, \ldots, d_n,$$

which has multiplicity $(d_n - p^*)$. For $1 \leq i \leq p^*$ and $l > p^*$, the function

$$\frac{\xi V_i^2 - V_l^2}{\xi V_i^2 + \sum_{\substack{j=1\\ j\neq i}}^{p^*}\xi_{n,j}V_j^2 + \sum_{j=p^*+1}^{d_n}V_j^2}$$

is strictly increasing function in ξ since the derivative in ξ is strictly positive. A conclusion is then for $1 \leq i \leq p^*$ with $\xi_{n,i} > 1$ and $l > p^*$ that

$$\begin{aligned}\mathrm{diag}(\mathbf{\Sigma}^{(n)})_i - \mathrm{diag}(\mathbf{\Sigma}^{(n)})_l &= \mathbb{E}\left[\frac{\xi_{n,i}V_i^2 - V_l^2}{\sum_{j=1}^{p^*}\xi_{n,p^*}V_j^2 + \sum_{j=p^*+1}^{d_n}V_j^2}\right] \\ &> \mathbb{E}\left[\frac{V_i^2 - V_l^2}{\sum_{\substack{j=1\\ j\neq i}}^{p^*}\xi_{n,p^*}V_j^2 + V_i^2 + \sum_{j=p^*+1}^{d_n}V_j^2}\right] = 0.\end{aligned}$$

Therefore, we receive that the first p^* diagonal entries of $\mathbf{\Sigma}^{(n)}$ correspond to the p^* largest eigenvalues of $\mathbf{\Sigma}^{(n)}$ namely $\mathrm{diag}(\mathbf{\Sigma}^{(n)})_1$, …, $\mathrm{diag}(\mathbf{\Sigma}^{(n)})_{p^*}$ and the remaining $(d_n - p^*)$ eigenvalues are strictly smaller and identical to $\mathrm{diag}(\mathbf{\Sigma}^{(n)})_{d_n}$. □

Proof of Theorem 4.14

For the proof of Theorem 4.14 we combine ideas for the spiked covariance model from Johnstone and Yang (2018) and for compositional data from Jiang et al. (2023). First, we derive an alternative representation for $\widehat{\mathbf{\Sigma}}^{(n)}$ in (4.5). As a consequence of the independence between the radial components $Z_1, \ldots, Z_n$ and the directional components

$\boldsymbol{X}_1^{(n)}/\|\boldsymbol{X}_1^{(n)}\|_2, \dots, \boldsymbol{X}_{k_n}^{(n)}/\|\boldsymbol{X}_{k_n}^{(n)}\|_2$ we obtain

$$\widehat{\boldsymbol{\Theta}}^{(n)} = \sum_{j=1}^{n} \frac{\boldsymbol{\Gamma}^{(n)1/2}\boldsymbol{V}_j^{(n)}}{\|\boldsymbol{\Gamma}^{(n)1/2}\boldsymbol{V}_j^{(n)}\|_2} \mathbb{1}\{Z_j > Z_{(k_n+1,n)}\} \stackrel{\mathcal{D}}{=} \sum_{j=1}^{k_n} \frac{\boldsymbol{X}_j^{(n)}}{\|\boldsymbol{X}_j^{(n)}\|_2},$$

and similarly

$$\widehat{\boldsymbol{\Sigma}}^{(n)\prime} := \frac{1}{k_n}\sum_{j=1}^{k_n}\left(\frac{\boldsymbol{X}_j^{(n)}}{\|\boldsymbol{X}_j^{(n)}\|_2} - \frac{1}{k_n}\sum_{i=1}^{k_n}\frac{\boldsymbol{X}_i^{(n)}}{\|\boldsymbol{X}_i^{(n)}\|_2}\right)\left(\frac{\boldsymbol{X}_j^{(n)}}{\|\boldsymbol{X}_j^{(n)}\|_2} - \frac{1}{k_n}\sum_{i=1}^{k_n}\frac{\boldsymbol{X}_i^{(n)}}{\|\boldsymbol{X}_i^{(n)}\|_2}\right)^\top \stackrel{\mathcal{D}}{=} \widehat{\boldsymbol{\Sigma}}^{(n)}. \tag{4.9}$$

The eigenvalues of $\widehat{\boldsymbol{\Sigma}}^{(n)\prime}$ are denoted by $\widehat{\lambda}'_{n,1}, \dots, \widehat{\lambda}'_{n,d_n}$ and due to (4.9) we receive that

$$(\widehat{\lambda}'_{n,1}, \dots, \widehat{\lambda}'_{n,d_n}) \stackrel{\mathcal{D}}{=} (\widehat{\lambda}_{n,1}, \dots, \widehat{\lambda}_{n,d_n}). \tag{4.10}$$

Thus, to prove Theorem 4.14 it suffices to derive the asymptotic behavior of $(\widehat{\lambda}'_{n,1}, \dots, \widehat{\lambda}'_{n,d_n})$, which relies on the spectral analysis of the empirical covariance matrix of $\boldsymbol{\Gamma}^{(n)1/2}\boldsymbol{V}^{(n)}$. Therefore, assume that $\boldsymbol{V}_1^{(n)}, \dots, \boldsymbol{V}_{k_n}^{(n)}$ is an i.i.d. sequence with distribution $\boldsymbol{V}^{(n)}$, i.e. $\boldsymbol{V}_i^{(n)} \in \mathbb{R}^{d_n}$ has i.i.d. entries with mean 0 and variance 1. Then we define the sequence of matrices

$$\begin{aligned}\boldsymbol{\Upsilon}^{(n)} := \frac{1}{k_n}\sum_{i=1}^{k_n}&\left(\boldsymbol{\Gamma}^{(n)1/2}\boldsymbol{V}_i^{(n)} - \frac{1}{k_n}\sum_{j=1}^{k_n}\boldsymbol{\Gamma}^{(n)1/2}\boldsymbol{V}_j^{(n)}\right)\\ &\cdot\left(\boldsymbol{\Gamma}^{(n)1/2}\boldsymbol{V}_i^{(n)} - \frac{1}{k_n}\sum_{j=1}^{k_n}\boldsymbol{\Gamma}^{(n)1/2}\boldsymbol{V}_j^{(n)}\right)^\top, \quad n \in \mathbb{N},\end{aligned} \tag{4.11}$$

whose eigenvalues are denoted by $\widehat{\xi}_{n,1} > \cdots > \widehat{\xi}_{n,d_n} > 0$. The aim now is to write $\widehat{\boldsymbol{\Sigma}}^{(n)\prime}$ and $\boldsymbol{\Upsilon}^{(n)}$ as matrix products. Therefore, define

$$\mathcal{V}^{(n)} := (\boldsymbol{V}_1^{(n)}, \dots, \boldsymbol{V}_{k_n}^{(n)}) \in \mathbb{R}^{d_n \times k_n}$$

and

$$\boldsymbol{T}^{(n)} := \operatorname{diag}(\|\boldsymbol{\Gamma}^{(n)1/2}\boldsymbol{V}_1^{(n)}\|_2^{-1}, \dots, \|\boldsymbol{\Gamma}^{(n)1/2}\boldsymbol{V}_{k_n}^{(n)}\|_2^{-1}) \in \mathbb{R}^{k_n \times k_n},$$

which allows us to write

$$\left(\frac{\boldsymbol{X}_1^{(n)}}{\|\boldsymbol{X}_1^{(n)}\|_2}, \dots, \frac{\boldsymbol{X}_{k_n}^{(n)}}{\|\boldsymbol{X}_{k_n}^{(n)}\|_2}\right)^\top = \boldsymbol{\Gamma}^{(n)1/2}\mathcal{V}^{(n)}\boldsymbol{T}^{(n)}.$$

Finally, with the projection matrix $\boldsymbol{P}^{(n)} \coloneqq (\boldsymbol{I}_{k_n} - \mathbf{1}_{k_n}\mathbf{1}_{k_n}^\top/k_n)$, the matrices $\widehat{\boldsymbol{\Sigma}}^{(n)\prime}$ and $\boldsymbol{\Upsilon}^{(n)}$, as defined in (4.9) and (4.11), can be written as

$$\begin{aligned}\widehat{\boldsymbol{\Sigma}}^{(n)\prime} &= \frac{1}{k_n}(\boldsymbol{\Gamma}^{(n)1/2}\mathcal{V}^{(n)}\boldsymbol{T}^{(n)}\boldsymbol{P}^{(n)})(\boldsymbol{\Gamma}^{(n)1/2}\mathcal{V}^{(n)}\boldsymbol{T}^{(n)}\boldsymbol{P}^{(n)})^\top,\\ \boldsymbol{\Upsilon}^{(n)} &= \frac{1}{k_n}(\boldsymbol{\Gamma}^{(n)1/2}\mathcal{V}^{(n)}\boldsymbol{P}^{(n)})(\boldsymbol{\Gamma}^{(n)1/2}\mathcal{V}^{(n)}\boldsymbol{P}^{(n)})^\top.\end{aligned} \tag{4.12}$$

In the following theorem the connection between the eigenvalues $\widehat{\xi}_{n,i}$ and $d_n\widehat{\lambda}'_{n,i}$ is derived.

Theorem 4.32. *Let Model D be given. Suppose that $\boldsymbol{\Gamma}_n \to \boldsymbol{\Gamma}$ and $(\xi_{n,1},\dots,\xi_{n,p^*}) \to (\xi_1,\dots,\xi_{p^*})$ as $n\to\infty$. If $\widehat{\xi}_{n,1},\dots,\widehat{\xi}_{n,d_n}$ denote the eigenvalues of $\boldsymbol{\Upsilon}^{(n)}$ in (4.11) and $\widehat{\lambda}'_{n,1},\dots,\widehat{\lambda}'_{n,d_n}$ denote the eigenvalues of $\widehat{\boldsymbol{\Sigma}}^{(n)\prime}$ in (4.9), then as $n\to\infty$,*

$$\max_{1\le i\le d_n} |\widehat{\xi}_{n,i} - d_n\widehat{\lambda}'_{n,i}| \xrightarrow{\mathbb{P}} 0.$$

Proof of Theorem 4.32. Due to Theorem A.46 in Bai and Silverstein (2010) and the sub-multiplicativity of the spectral norm we receive that

$$\begin{aligned}\max_{1\le i\le d_n}\left|\sqrt{\widehat{\xi}_{n,i}} - \sqrt{d_n\widehat{\lambda}'_{n,i}}\right| &\le \left\|\frac{\sqrt{d_n}\boldsymbol{\Gamma}^{(n)1/2}\mathcal{V}^{(n)}\boldsymbol{T}^{(n)}\boldsymbol{P}^{(n)}}{\sqrt{k_n}} - \frac{\boldsymbol{\Gamma}^{(n)1/2}\mathcal{V}^{(n)}\boldsymbol{P}^{(n)}}{\sqrt{k_n}}\right\|_2\\ &\le \left\|\boldsymbol{P}^{(n)}\right\|_2 \cdot \left\|\sqrt{d_n}\boldsymbol{T}^{(n)} - \boldsymbol{I}_{k_n}\right\|_2 \cdot \left\|\frac{\boldsymbol{\Gamma}^{(n)1/2}\mathcal{V}^{(n)}}{\sqrt{k_n}}\right\|_2\\ &= \left\|\sqrt{d_n}\boldsymbol{T}^{(n)} - \boldsymbol{I}_{k_n}\right\|_2 \cdot \left\|\frac{\boldsymbol{\Gamma}^{(n)1/2}\mathcal{V}^{(n)}}{\sqrt{k_n}}\right\|_2 =: J_n \cdot H_n,\end{aligned} \tag{4.13}$$

where we used that the spectral norm of $\boldsymbol{P}^{(n)}$ is bounded by 1, because the only eigenvalues of $\boldsymbol{P}^{(n)}$ are 1 and 0 as $\boldsymbol{P}^{(n)}$ is a projection matrix.

Step 1. First, we show that J_n in (4.13) converges to 0 in probability. Therefore, we use the partitioning of the random vector $\boldsymbol{\Gamma}^{(n)1/2}\boldsymbol{V}_j^{(n)}$ into the first p^* dependent entries and the remaining $d_n - p^*$ independent entries

$$\boldsymbol{\Gamma}^{(n)1/2}\boldsymbol{V}_j^{(n)} = \begin{pmatrix}\boldsymbol{\Gamma}_n^{1/2}\boldsymbol{V}_{j,\{1,\dots,p^*\}}^{(n)}\\ \boldsymbol{V}_{j,\{p^*+1,\dots,d_n\}}^{(n)}\end{pmatrix} =: \begin{pmatrix}(U_{j,1}^{(n)},\dots,U_{j,p^*}^{(n)})^\top\\ (V_{j,(p^*+1)},\dots,V_{j,d_n})^\top\end{pmatrix}.$$

The eigenvalues of $(\sqrt{d_n}\boldsymbol{T}^{(n)} - \boldsymbol{I}_{k_n})$ correspond to the diagonal entries. Since we apply the spectral norm, we receive that

$$J_n^{1/2} = \left\|\sqrt{d_n}\boldsymbol{T}^{(n)} - \boldsymbol{I}_{k_n}\right\|_2^{1/2} = \max_{1\le i\le k_n}\left|\frac{\sqrt{d_n}}{\left(\sum_{l=1}^{p^*} {U_{i,l}^{(n)}}^2 + \sum_{l=p^*+1}^{d_n} {V_{i,l}}^2\right)^{1/2}} - 1\right|.$$

On the one hand, by $\mathbb{E}[{V_{i,j}}^2] = 1, d_n/k_n \to c > 0$ and Bai and Yin (1993, Lemma 2) we

obtain that as $n \to \infty$

$$\max_{1\le i\le k_n}\left|\frac{\sum_{l=p^*+1}^{d_n} {V_{i,l}}^2}{d_n}-1\right| \xrightarrow{\mathbb{P}\text{-a.s.}} 0.$$

On the other hand, for $1 \le i \le k_n$,

$$\sum_{l=1}^{p^*} {U_{i,l}^{(n)}}^2 = \|\mathbf{\Gamma}_n^{1/2}\boldsymbol{V}_{i,\{1,\dots,p^*\}}^{(n)}\|_2^2 \le \|\mathbf{\Gamma}_n^{1/2}\|_2^2\|\boldsymbol{V}_{i,\{1,\dots,p^*\}}^{(n)}\|_2^2 = \xi_{n,1}\sum_{l=1}^{p^*} {V_{i,l}}^2.$$

Since the second moment of V_1^2 exists, we can conclude from Markov inequality for $\varepsilon > 0$

$$\begin{aligned}\mathbb{P}\left(\frac{\xi_{n,1}}{d_n}\max_{1\le i\le k_n}\left|\sum_{l=1}^{p^*} {V_{i,l}}^2\right| > \varepsilon\right) &\le \sum_{i=1}^{k_n}\mathbb{P}\left(\left|\sum_{l=1}^{p^*} {V_{i,l}}^2\right| > \frac{d_n}{\xi_{n,1}}\varepsilon\right)\\ &= k_n\mathbb{P}\left(\left|\sum_{l=1}^{p^*} {V_{1,l}}^2\right| > \frac{d_n}{\xi_{n,1}}\varepsilon\right)\\ &\le k_n\frac{\xi_{n,1}^2}{d_n^2\varepsilon^2}\mathbb{E}\left|\sum_{l=1}^{p^*} {V_{1,l}}^2\right|^2, \end{aligned} \tag{4.14}$$

where the right-hand side converges to 0 as $n \to \infty$, since $k_n/d_n \to c^{-1}$ and $\xi_{n,1}^2/d_n \to 0$ as $n \to \infty$. Therefore, we get

$$\max_{1\le i\le k_n}\left|\frac{\sum_{l=1}^{p^*} {U_{i,l}^{(n)}}^2}{d_n}\right| \le \frac{\xi_{n,1}}{d_n}\max_{1\le i\le k_n}\left|\sum_{l=1}^{p^*} {V_{i,l}}^2\right| \xrightarrow{\mathbb{P}} 0.$$

To summarize,

$$\begin{aligned}&\max_{1\le i\le k_n}\left|\left(\frac{\sum_{l=1}^{p^*} {U_{i,l}^{(n)}}^2 + \sum_{l=p^*+1}^{d_n} {V_{i,l}}^2}{d_n}\right)-1\right|\\ &\quad\le \max_{1\le i\le k_n}\left|\left(\frac{\sum_{l=1}^{p^*} {U_{i,l}^{(n)}}^2}{d_n}\right)\right| + \max_{1\le i\le k_n}\left|\left(\frac{\sum_{l=p^*+1}^{d_n} {V_{i,l}}^2}{d_n}\right)-1\right| \xrightarrow{\mathbb{P}} 0.\end{aligned}$$

Finally, by the mean value theorem the inequality

$$\left|1-1/\sqrt{x}\right| \le 2\left|x-1\right| \quad \text{for } x > \frac{1}{2}$$

holds and hence, as $n \to \infty$,

$$J_n^{1/2} = \max_{1\le i\le k_n}\left|1-\left(\frac{1}{d_n}\sum_{l=1}^{p^*} {U_{i,l}^{(n)}}^2 + \frac{1}{d_n}\sum_{l=p^*+1}^{d_n} {V_{i,l}}^2\right)^{-1/2}\right| \xrightarrow{\mathbb{P}} 0. \tag{4.15}$$

Step 2. Next, we show that H_n in (4.13) is $\mathbb{P}$-a.s. bounded. By Yin et al. (1988, Theorem 3.1) (cf. Bai and Silverstein, 2010, Theorem 5.8)

$$H_n = \left\| \frac{\mathbf{\Gamma}^{(n)^{1/2}} \mathcal{V}^{(n)}}{\sqrt{k_n}} \right\|_2 \leq \left\| \mathbf{\Gamma}^{(n)^{1/2}} \right\|_2 \cdot \left\| \frac{\mathcal{V}^{(n)}}{\sqrt{k_n}} \right\|_2^2 = \xi_{n,1} \frac{\lambda_{\max}(\mathcal{V}^{(n)^\top} \mathcal{V}^{(n)})}{k_n} \xrightarrow{\mathbb{P}\text{-a.s.}} \xi_1$$

as $n \to \infty$, where $\lambda_{\max}(\cdot)$ denotes the largest eigenvalue of a matrix.

Finally, a combination of (4.13), Step 1 and Step 2 result in the statement. □

Remark 4.33. For the convergence of the right-hand side of (4.14) and hence, (4.15) to zero, it is not necessary that $\xi_{n,1}$ is bounded; it is sufficient that $\xi_{n,1} = o(\sqrt{d_n})$ as $n \to \infty$. But if all moments of V_1 exist, it is even sufficient to assume that $\xi_{n,1} = o(d_n^\beta)$ as $n \to \infty$ for some $\beta < 1$. Indeed, we get analog to (4.14) for $\varepsilon > 0$ that

$$\mathbb{P}\left(\frac{\xi_{n,1}}{d_n} \max_{1 \leq i \leq k_n} \left| \sum_{l=1}^{p^*} {V_{i,l}}^2 \right| > \varepsilon \right) \leq k_n \frac{\xi_{n,1}^{1/(1-\beta)}}{d_n^{1/(1-\beta)}} \varepsilon^{1/(1-\beta)} \mathbb{E}\left[\left| \sum_{l=1}^{p^*} {V_{1,l}}^2 \right|^{1/(1-\beta)} \right] \to 0$$

as $n \to \infty$, since $k_n/d_n \to c$ and ${\xi_{n,1}}^{(1-\beta)^{-1}} / d_n^{(1-\beta)^{-1}-1} = (\xi_{n,1}/d_n^\beta)^{1/(1-\beta)} = o(1)$ as $n \to \infty$.

Next, we repeat results on the asymptotic distribution of the eigenvalues of $\mathbf{\Upsilon}^{(n)}$ which is mainly based on Bai and Yao (2012) and Bai et al. (2018).

Lemma 4.34. *Let Model D be given. Suppose that $\mathbf{\Gamma}_n \to \mathbf{\Gamma}$ and $(\xi_{n,1}, \ldots, \xi_{n,p^*}) \to (\xi_1, \ldots, \xi_{p^*})$ as $n \to \infty$ with $\xi_{p^*} > 1 + \sqrt{c}$. Then the following statements hold.*

(a) If $1 \leq i \leq p^*$ (i.e. $\xi_i > 1 + \sqrt{c}$), then $\widehat{\xi}_{n,i} \xrightarrow{\mathbb{P}\text{-a.s.}} \varphi_c(\xi_i)$ as $n \to \infty$.

(b) Define $l^* := 0$ if $c \leq 1$ and $l^* := 1 - c^{-1}$ if $c > 1$. Then

$$\lim_{n \to \infty} \sup_{\alpha \in (l^*, 1)} \left| F^{\mathbf{\Upsilon}^{(n)} \leftarrow}(\alpha) - F_c^{\leftarrow}(\alpha) \right| = 0 \quad \mathbb{P}\text{-a.s.},$$

where $F^{\mathbf{\Upsilon}^{(n)} \leftarrow}$ is the generalized inverse of the empirical spectral distribution function of $\mathbf{\Upsilon}^{(n)}$ and $F_c(x)$ is defined as in Theorem 4.14.

(c) If $i_n(\alpha) > p^*$ (i.e. $\xi_{i_n(\alpha)} = 1$) and $i_n(\alpha)/d_n \to \alpha \in (0,1)$ as $n \to \infty$, then

$$\sup_{\alpha \in (0,1)} \left| \widehat{\xi}_{n, i_n(\alpha)} - F_c^{\leftarrow}(1-\alpha) \right| \xrightarrow{\mathbb{P}\text{-a.s.}} 0, \qquad \text{as } n \to \infty.$$

In particular, if $(q_n)_{n \in \mathbb{N}}$ is a sequence in $\mathbb{N}$ with $q_n = o(d_n)$ as $n \to \infty$ and $q_n > p^*$, then $\widehat{\xi}_{n,q_n} \xrightarrow{\mathbb{P}\text{-a.s.}} (1 + \sqrt{c})^2$.

(d) Suppose $0 < c \leq 1$ and $(q_n)_{n \in \mathbb{N}}$ is a sequence in $\mathbb{N}$ with $q_n = o(d_n)$ as $n \to \infty$. Then

as $n \to \infty$,

$$\frac{1}{d_n - q_n} \sum_{i=q_n+1}^{d_n} \widehat{\xi}_{n,i} \xrightarrow{\mathbb{P}\text{-a.s.}} 1$$

and for $q_n > p^*$ *we receive that* $\widehat{\xi}_{n,q_n} \xrightarrow{\mathbb{P}\text{-a.s.}} (1+\sqrt{c})^2$.

(e) Suppose $c > 1$ *and* $(q_n)_{n\in\mathbb{N}}$ *is a sequence in* $\mathbb{N}$ *with* $q_n = o(d_n)$ *as* $n \to \infty$. *Then as* $n \to \infty$,

$$\frac{1}{k_n - q_n} \sum_{i=q_n+1}^{k_n} \widehat{\xi}_{n,i} \xrightarrow{\mathbb{P}\text{-a.s.}} c$$

and for $q_n > p^*$ *we receive that* $\widehat{\xi}_{n,q_n} \xrightarrow{\mathbb{P}\text{-a.s.}} (1+\sqrt{c})^2$.

Proof.

(a) When the eigenvalues $(\xi_{n,1}, \ldots, \xi_{n,p^*}) = (\xi_1, \ldots, \xi_{p^*})$ do not depend on n, (a) goes back to Bai and Yao (2012, Theorem 4.1) (cf. Bai et al., 2018, Lemma 2.1). In the case $\mathbf{\Gamma}_n \to \mathbf{\Gamma}$ and $(\xi_{n,1}, \ldots, \xi_{n,p^*}) \to (\xi_1, \ldots, \xi_{p^*})$ as $n \to \infty$ the assertion also holds because by Bai and Silverstein (2010, Theorem A.46) it can be shown with the same arguments as before that

$$\max_{1\le i\le p^*} \left| \sqrt{\widehat{\xi}_{n,i}(\mathbf{\Gamma}_n)} - \sqrt{\widehat{\xi}_{n,i}(\mathbf{\Gamma})} \right| \le \|\mathbf{\Gamma}_n - \mathbf{\Gamma}\|_2 \left\| \frac{\mathcal{Y}^{(n)}}{\sqrt{k_n}} \right\|_2 \|\boldsymbol{P}^{(n)}\|_2 \xrightarrow{\mathbb{P}\text{-a.s.}} 0,$$

where $\widehat{\xi}_{n,i}(\mathbf{\Gamma}_n)$ and $\widehat{\xi}_{n,i}(\mathbf{\Gamma})$ is the empirical eigenvalue when $\mathbf{\Gamma}_n$ and $\mathbf{\Gamma}$, respectively is used.

(b) The second part is similar to Bai and Yao (2012, Theorem 4.1) however, the wording is not clear and therefore we prefer to include the proper statement and proof here. Note, if $d_n/k_n \to c > 0$ as $n \to \infty$, then for almost all $\omega \in \Omega$, $F^{\mathbf{\Upsilon}^{(n)}}(\omega)$ converges in distribution to F_c (cf. Bai et al., 2018, p. 1054 and Silverstein, 1995, Theorem 1.1). This means that there exists a set $\Omega_0 \in \mathcal{F}$ with $\mathbb{P}(\Omega_0) = 1$ and for any $\omega \in \Omega_0$ and any continuity point $x \in \mathbb{R}$ of F_c,

$$\lim_{n\to\infty} F^{\mathbf{\Upsilon}^{(n)}}(x, \omega) = F_c(x).$$

Since the distribution function F_c is continuous on the interval $I := ((1-\sqrt{c})^2, (1+\sqrt{c})^2)$, a conclusion of Polya's Theorem is the uniform convergence

$$\lim_{n\to\infty} \sup_{x\in I} \left| F^{\mathbf{\Upsilon}^{(n)}}(x, \omega) - F_c(x) \right| = 0,$$

which implies by de Haan and Ferreira (2006, Lemma 1.1.1) and again Polya's

Theorem as well as the uniform convergence of the quantile function

$$\lim_{n\to\infty} \sup_{\alpha\in(l^*,1)} \left| F^{\boldsymbol{\Upsilon}^{(n)}\leftarrow}(\alpha,\omega) - F_c^{\leftarrow}(\alpha) \right| = 0.$$

(c) Since $\widehat{\xi}_{n,i_n(\alpha)} = F^{\boldsymbol{\Upsilon}^{(n)}\leftarrow}(1 - i_n(\alpha)/d_n)$ the statement follows directly from (b).

(d) Due to (b), we receive that

$$\begin{aligned} \frac{1}{d_n - q_n} \sum_{i=q_n+1}^{d_n} \widehat{\xi}_{n,i} &= \frac{d_n}{d_n - q_n} \int_0^{1-\frac{q_n}{d_n}} F^{\boldsymbol{\Upsilon}^{(n)}\leftarrow}(1-\alpha)\, d\alpha \\ &\xrightarrow{\mathbb{P}\text{-a.s.}} 1 \cdot \int_0^1 F_c^{\leftarrow}(1-\alpha)\, d\alpha = 1. \end{aligned}$$

(e) Similarly to (d) we have

$$\begin{aligned} \frac{1}{k_n - q_n} \sum_{i=q_n+1}^{k_n} \widehat{\xi}_{n,i} &= \frac{d_n}{k_n - q_n} \int_{1-\frac{k_n}{d_n}}^{1-\frac{q_n}{d_n}} F^{\boldsymbol{\Upsilon}^{(n)}\leftarrow}(1-\alpha)\, d\alpha \\ &\xrightarrow{\mathbb{P}\text{-a.s.}} c \cdot \int_{1-c^{-1}}^1 F_c^{\leftarrow}(1-\alpha)\, d\alpha = c. \end{aligned}$$

□

Finally, we have all the auxiliary results for the proof of Theorem 4.14.

Proof of Theorem 4.14. (a) An assumption is that $\xi_i > 1 + \sqrt{c}$ and hence, ξ_i is a distant spiked eigenvalue for $i = 1, \ldots, p^*$. A conclusion of Lemma 4.34 (a) is then that $\widehat{\xi}_{n,i} \xrightarrow{\mathbb{P}\text{-a.s.}} \varphi_c(\xi_i)$. Combined with Theorem 4.32 we receive that $d_n \widehat{\lambda}'_{n,i} \xrightarrow{\mathbb{P}} \varphi_c(\xi_i)$ as $n \to \infty$. Due to (4.10), the identical distribution of $\widehat{\lambda}'_{n,i}$ and $\widehat{\lambda}_{n,i}$, we obtain the final statement, $d_n \widehat{\lambda}_{n,i} \xrightarrow{\mathbb{P}} \varphi_c(\xi_i)$ as $n \to \infty$.

Similarly as in (a), the statements (b)-(d) are combinations of Lemma 4.34, Theorem 4.32 and (4.10). □

Proof of Theorem 4.15

For the proof of Theorem 4.15, Theorem 4.32 is not useful and an adapted version does not exist. Therefore, the approach is slightly different. First, the next lemma gives the asymptotic distribution of the eigenvalues of $\widehat{\boldsymbol{\Sigma}}^{(n)\prime}$, which is then used for the proof of Theorem 4.15.

Lemma 4.35. *Let Model D with $\xi_{n,p^*} \to \infty$ and $\xi_{n,1} = o(d_n^{1/2})$ as $n \to \infty$ be given. If $i \in \{1, \ldots, p^*\}$ then*

$$\frac{d_n \widehat{\lambda}'_{n,i}}{\xi_{n,i}} \xrightarrow{\mathbb{P}} 1 \quad \text{as } n \to \infty.$$

Proof. We proceed similarly to the proof of Bai et al. (2018, Lemma 2.2) and use the spectral decomposition of $\boldsymbol{\Gamma}^{(n)}$. Let $\boldsymbol{\Gamma}^{(n)} = \boldsymbol{W}^{(n)}\boldsymbol{D}^{(n)}\boldsymbol{W}^{(n)\top}$, where $\boldsymbol{W}^{(n)} = (\boldsymbol{W}_1^{(n)}, \ldots, \boldsymbol{W}_{d_n}^{(n)})$ is a $(d_n \times d_n)$-dimensional orthogonal matrix and $\boldsymbol{D}^{(n)} := \operatorname{diag}(\overline{\boldsymbol{D}}_n, \boldsymbol{I}_{d_n - p^*})$
$:= \operatorname{diag}(\xi_{n,1}, \ldots, \xi_{n,p^*}, 1, \ldots, 1) \in \mathbb{R}^{d_n \times d_n}$ consists of the eigenvalues of $\boldsymbol{\Gamma}^{(n)}$. Then with representation (4.12) and

$$\boldsymbol{A}^{(n)} := \mathcal{V}^{(n)}\boldsymbol{T}^{(n)}\boldsymbol{P}^{(n)}\boldsymbol{T}^{(n)}\mathcal{V}^{(n)\top} \tag{4.16}$$

we receive

$$\begin{aligned}\widehat{\boldsymbol{\Sigma}}^{(n)\prime} &= \frac{1}{k_n}\boldsymbol{W}^{(n)}\boldsymbol{D}^{(n)1/2}\boldsymbol{W}^{(n)\top}\mathcal{V}^{(n)}\boldsymbol{T}^{(n)}\boldsymbol{P}^{(n)}\boldsymbol{T}^{(n)}\mathcal{V}^{(n)\top}\boldsymbol{W}^{(n)}\boldsymbol{D}^{(n)1/2}\boldsymbol{W}^{(n)\top} \\ &= \frac{1}{k_n}\boldsymbol{W}^{(n)}\boldsymbol{D}^{(n)1/2}\boldsymbol{W}^{(n)\top}\boldsymbol{A}^{(n)}\boldsymbol{W}^{(n)}\boldsymbol{D}^{(n)1/2}\boldsymbol{W}^{(n)\top}. \end{aligned} \tag{4.17}$$

Further, the eigenvectors are partitioned into the first p^* and the remaining eigenvectors by defining $\overline{\boldsymbol{W}}^{(n)} = (\boldsymbol{W}_1^{(n)}, \ldots, \boldsymbol{W}_{p^*}^{(n)})$ in $\mathbb{R}^{d_n \times p^*}$and $\widetilde{\boldsymbol{W}}^{(n)} = (\boldsymbol{W}_{p^*+1}^{(n)}, \ldots, \boldsymbol{W}_{d_n}^{(n)})$ in $\mathbb{R}^{d_n \times (d_n - p^*)}$ such that

$$\widehat{\boldsymbol{\Sigma}}^{(n)\prime} = \frac{1}{k_n}\boldsymbol{W}^{(n)}\begin{pmatrix} \overline{\boldsymbol{D}}_n^{1/2}\overline{\boldsymbol{W}}^{(n)\top}\boldsymbol{A}^{(n)}\overline{\boldsymbol{W}}^{(n)}\overline{\boldsymbol{D}}_n^{1/2} & \overline{\boldsymbol{D}}_n^{1/2}\overline{\boldsymbol{W}}^{(n)\top}\boldsymbol{A}^{(n)}\widetilde{\boldsymbol{W}}^{(n)} \\ \widetilde{\boldsymbol{W}}^{(n)\top}\boldsymbol{A}^{(n)}\overline{\boldsymbol{W}}^{(n)}\overline{\boldsymbol{D}}_n^{1/2} & \widetilde{\boldsymbol{W}}^{(n)\top}\boldsymbol{A}^{(n)}\widetilde{\boldsymbol{W}}^{(n)} \end{pmatrix}\boldsymbol{W}^{(n)\top}.$$

Similarly, we receive with (4.12) and

$$\boldsymbol{B}^{(n)} := \mathcal{V}^{(n)}\boldsymbol{P}^{(n)}\mathcal{V}^{(n)\top} \tag{4.18}$$

that

$$\boldsymbol{\Upsilon}^{(n)} = \frac{1}{k_n}\boldsymbol{W}^{(n)}\begin{pmatrix} \overline{\boldsymbol{D}}_n^{1/2}\overline{\boldsymbol{W}}^{(n)\top}\boldsymbol{B}^{(n)}\overline{\boldsymbol{W}}^{(n)}\overline{\boldsymbol{D}}_n^{1/2} & \overline{\boldsymbol{D}}_n^{1/2}\overline{\boldsymbol{W}}^{(n)\top}\boldsymbol{B}^{(n)}\widetilde{\boldsymbol{W}}^{(n)} \\ \widetilde{\boldsymbol{W}}^{(n)\top}\boldsymbol{B}^{(n)}\overline{\boldsymbol{W}}^{(n)}\overline{\boldsymbol{D}}_n^{1/2} & \widetilde{\boldsymbol{W}}^{(n)\top}\boldsymbol{B}^{(n)}\widetilde{\boldsymbol{W}}^{(n)} \end{pmatrix}\boldsymbol{W}^{(n)\top}.$$

Let $i \in \{1, \ldots, p^*\}$. The Courant-Fischer min-max theorem (Horn and Johnson, 2013, Theorem 4.2.6) gives

$$\frac{d_n\widehat{\lambda}'_{n,i}}{\xi_{n,i}} = \frac{d_n}{\xi_{n,i}} \inf_{\boldsymbol{v}_1,\ldots,\boldsymbol{v}_{i-1}\in\mathbb{R}^{d_n}} \sup_{\boldsymbol{w}\perp\boldsymbol{v}_1,\ldots,\boldsymbol{v}_{i-1},\|\boldsymbol{w}\|_2=1} \boldsymbol{w}^\top\widehat{\boldsymbol{\Sigma}}^{(n)\prime}\boldsymbol{w}. \tag{4.19}$$

The proof is split into two parts, wherein we establish that $d_n\widehat{\lambda}'_{n,i}/\xi_{n,i}$ is bounded below and above by a random variable which converges in probability to 1 as $n \to \infty$.

Step 1: First, we derive a lower bound of (4.19) which converges in probability to 1. Therefore, note for arbitrary $\boldsymbol{u}_j \in \mathbb{R}^{d_n}$ with $\|\boldsymbol{u}_j\|_2 = 1$ for $1 \leq j \leq p^*$, Bai et al. (2018,

Lemma A.2) yields that as $n \to \infty$,

$$\max_{1 \leq j \leq p^*} \left| \boldsymbol{u}_j^\top \frac{\boldsymbol{B}^{(n)}}{k_n} \boldsymbol{u}_j - 1 \right| \xrightarrow{\mathbb{P}\text{-a.s.}} 0, \tag{4.20}$$

where $\boldsymbol{B}^{(n)}$ is defined as in (4.18). Now, let $\boldsymbol{A}^{(n)}$ be defined as in (4.16). Then

$$\begin{aligned}
&\left| \boldsymbol{u}_j^\top \left(\frac{\boldsymbol{B}^{(n)}}{k_n} - \frac{d_n \boldsymbol{A}^{(n)}}{k_n} \right) \boldsymbol{u}_j \right| \\
&\quad \leq \left\| \frac{\boldsymbol{B}^{(n)}}{k_n} - \frac{d_n \boldsymbol{A}^{(n)}}{k_n} \right\|_2 \\
&\quad = \frac{1}{k_n} \left\| \mathcal{V}^{(n)} \left(\boldsymbol{P}^{(n)} - d_n \boldsymbol{T}^{(n)} \boldsymbol{P}^{(n)} \boldsymbol{T}^{(n)\top} \right) \mathcal{V}^{(n)\top} \right\|_2 \\
&\quad \leq \frac{1}{k_n} \|\mathcal{V}^{(n)}\|_2^2 \left\| \left(\boldsymbol{P}^{(n)} - d_n \boldsymbol{T}^{(n)} \boldsymbol{P}^{(n)} + d_n \boldsymbol{T}^{(n)} \boldsymbol{P}^{(n)} - d_n \boldsymbol{T}^{(n)} \boldsymbol{P}^{(n)} \boldsymbol{T}^{(n)\top} \right) \right\|_2 \\
&\quad \leq \frac{1}{k_n} \|\mathcal{V}^{(n)}\|_2^2 \|\boldsymbol{P}^{(n)}\|_2 \left(1 + \|\sqrt{d_n} \boldsymbol{T}^{(n)}\|_2 \right) \left\| \sqrt{d_n} \boldsymbol{T}^{(n)} - \boldsymbol{I}_{k_n} \right\|_2 .
\end{aligned}$$

On the one hand, Yin et al. (1988, Theorem 3.1) implies that

$$\frac{1}{k_n} \|\mathcal{V}^{(n)}\|_2^2 = \left(\frac{1}{\sqrt{k_n}} \|\mathcal{V}^{(n)}\|_2 \right)^2 \xrightarrow{\mathbb{P}\text{-a.s.}} (1 + \sqrt{c})^2.$$

On the other hand, since $\xi_{n,1} = o(d_n^{1/2})$ as $n \to \infty$, a conclusion of Remark 4.33 is that $\|\sqrt{d_n} \boldsymbol{T}^{(n)} - \boldsymbol{I}_{k_n}\|_2 \xrightarrow{\mathbb{P}} 0$ and $\|\sqrt{d_n} \boldsymbol{T}^{(n)}\|_2 \leq \|\boldsymbol{I}_{k_n}\|_2 + \|\sqrt{d_n} \boldsymbol{T}^{(n)} - \boldsymbol{I}_{k_n}\|_2 \xrightarrow{\mathbb{P}} 1$. In summary, as $n \to \infty$

$$\left| \boldsymbol{u}_j^\top \left(\frac{\boldsymbol{B}^{(n)}}{k_n} - \frac{d_n \boldsymbol{A}^{(n)}}{k_n} \right) \boldsymbol{u}_j \right| \leq \left\| \frac{\boldsymbol{B}^{(n)}}{k_n} - \frac{d_n \boldsymbol{A}^{(n)}}{k_n} \right\|_2 \xrightarrow{\mathbb{P}} 0, \tag{4.21}$$

and finally, using (4.20) we have as well

$$\max_{1 \leq j \leq p^*} \left| \boldsymbol{u}_j^\top \frac{d_n \boldsymbol{A}^{(n)}}{k_n} \boldsymbol{u}_j - 1 \right| \xrightarrow{\mathbb{P}} 0. \tag{4.22}$$

Further, for arbitrary vectors $\boldsymbol{v}_1, \ldots, \boldsymbol{v}_{i-1} \in \mathbb{R}^{d_n}$ we take a vector $\boldsymbol{w}_{\boldsymbol{v}} = \sum_{j=1}^i a_j \boldsymbol{W}_j^{(n)}$ orthogonal to $\boldsymbol{v}_1, \ldots, \boldsymbol{v}_{i-1}$ with $\sum_{j=1}^i a_j^2 = 1$ and hence, $\|\boldsymbol{w}_{\boldsymbol{v}}\|_2 = 1$. Since $\boldsymbol{W}^{(n)}$ is an orthogonal matrix, we receive with representation (4.17) that

$$\begin{aligned}
&\frac{d_n}{\xi_{n,i}} \boldsymbol{w}_{\boldsymbol{v}}^\top \widehat{\boldsymbol{\Sigma}}^{(n)\prime} \boldsymbol{w}_{\boldsymbol{v}} \\
&\quad = \frac{d_n}{\xi_{n,i}} \sum_{j,l=1}^i a_j a_l \boldsymbol{W}_j^{(n)\top} \boldsymbol{W}^{(n)} \boldsymbol{D}^{(n)1/2} \boldsymbol{W}^{(n)\top} \frac{\boldsymbol{A}^{(n)}}{k_n} \boldsymbol{W}^{(n)} \boldsymbol{D}^{(n)1/2} \boldsymbol{W}^{(n)\top} \boldsymbol{W}_l^{(n)} \\
&\quad = \sum_{j=1}^i a_j^2 \frac{\xi_{n,j}}{\xi_{n,i}} \boldsymbol{W}_j^{(n)\top} \frac{d_n \boldsymbol{A}^{(n)}}{k_n} \boldsymbol{W}_j^{(n)}.
\end{aligned}$$

A conclusion of (4.19), $\|\boldsymbol{W}_j^{(n)}\|_2 = 1$ and (4.22) is then

$$\begin{aligned}\frac{d_n\widehat{\lambda}'_{n,i}}{\xi_{n,i}} &\geq \inf_{\boldsymbol{v}_1,\dots,\boldsymbol{v}_{i-1}\in\mathbb{R}^{d_n}} \frac{d_n}{\xi_{n,i}}\boldsymbol{w}_{\boldsymbol{v}}^\top \widehat{\boldsymbol{\Sigma}}^{(n)\prime}\boldsymbol{w}_{\boldsymbol{v}} \\ &\geq \inf_{\boldsymbol{a}\in\mathbb{R}^i:\sum_{j=1}^i a_j^2=1} \sum_{j=1}^i a_j^2 \boldsymbol{W}_j^{(n)\top}\frac{d_n\boldsymbol{A}^{(n)}}{k_n}\boldsymbol{W}_j^{(n)} \\ &\geq 1 - \max_{1\leq j\leq p^*}\left|\boldsymbol{W}_j^{(n)\top}\frac{d_n\boldsymbol{A}^{(n)}}{k_n}\boldsymbol{W}_j^{(n)} - 1\right| \xrightarrow{\mathbb{P}} 1\end{aligned}$$

as $n\to\infty$.

Step 2: Next, we derive an upper bound for (4.19) which converges in probability to 1. Therefore, note that

$$\begin{aligned}\frac{d_n\widehat{\lambda}'_{n,i}}{\xi_{n,i}} &= \frac{d_n}{\xi_{n,i}} \inf_{\boldsymbol{v}_1,\dots,\boldsymbol{v}_{i-1}\in\mathbb{R}^{d_n}} \sup_{\boldsymbol{w}\perp\boldsymbol{v}_1,\dots,\boldsymbol{v}_{i-1},\|\boldsymbol{w}\|_2=1} \boldsymbol{w}^\top\widehat{\boldsymbol{\Sigma}}^{(n)\prime}\boldsymbol{w} \\ &\leq \frac{d_n}{\xi_{n,i}} \sup_{\boldsymbol{w}\perp\boldsymbol{W}_1^{(n)},\dots,\boldsymbol{W}_{i-1}^{(n)},\|\boldsymbol{w}\|_2=1} \boldsymbol{w}^\top\widehat{\boldsymbol{\Sigma}}^{(n)\prime}\boldsymbol{w}.\end{aligned}$$

Since $\boldsymbol{W}_l^{(n)} \perp \boldsymbol{W}_j^{(n)}$ for $l\neq j$ we can write a vector $\boldsymbol{w}\perp\boldsymbol{W}_1^{(n)},\dots,\boldsymbol{W}_{i-1}^{(n)}$ with $\|\boldsymbol{w}\|_2=1$ as

$$\boldsymbol{w} = c^2\boldsymbol{u} + (1-c^2)\boldsymbol{v},$$

where $c\in[0,1], \boldsymbol{u} = \sum_{j=i}^{p^*} a_j\boldsymbol{W}_j^{(n)} = \overline{\boldsymbol{W}}^{(n)}\boldsymbol{a}, \|\boldsymbol{a}\|_2 = \sum_{j=i}^{p^*} a_j^2 = 1$ and $\boldsymbol{v} = \sum_{j=p^*+1}^{d_n} b_j\boldsymbol{W}_j^{(n)} = \widetilde{\boldsymbol{W}}^{(n)}\boldsymbol{b}$ satisfying $\sum_{j=p^*+1}^{d_n} b_j^2 = 1$. Recall that $\widetilde{\boldsymbol{W}}^{(n)\top}\widehat{\boldsymbol{\Sigma}}^{(n)\prime}\widetilde{\boldsymbol{W}}^{(n)} = \widetilde{\boldsymbol{W}}^{(n)\top}\frac{\boldsymbol{A}^{(n)}}{k_n}\widetilde{\boldsymbol{W}}^{(n)}$. Then,

$$\begin{aligned}&\frac{d_n}{\xi_{n,i}} \sup_{\boldsymbol{w}\perp\boldsymbol{W}_1^{(n)},\dots,\boldsymbol{W}_{i-1}^{(n)},\|\boldsymbol{w}\|_2=1} \boldsymbol{w}^\top\widehat{\boldsymbol{\Sigma}}^{(n)\prime}\boldsymbol{w} \\ &\quad\leq \frac{d_n}{\xi_{n,i}} \sup_{c\in[0,1]}\Bigg\{c^2 \sup_{\substack{\boldsymbol{a}\in\mathbb{R}^{p^*-i+1},\\ \|\boldsymbol{a}\|_2=1}} \boldsymbol{a}^\top\overline{\boldsymbol{W}}^{(n)\top}\widehat{\boldsymbol{\Sigma}}^{(n)\prime}\overline{\boldsymbol{W}}^{(n)}\boldsymbol{a} \\ &\qquad\qquad + (1-c^2)\sup_{\substack{\boldsymbol{b}\in\mathbb{R}^{d-p^*},\\ \|\boldsymbol{b}\|_2=1}} \boldsymbol{b}^\top\widetilde{\boldsymbol{W}}^{(n)\top}\widehat{\boldsymbol{\Sigma}}^{(n)\prime}\widetilde{\boldsymbol{W}}^{(n)}\boldsymbol{b}\Bigg\} \\ &\quad= \frac{d_n}{\xi_{n,i}} \sup_{c\in[0,1]}\Bigg\{c^2 \sup_{\substack{\boldsymbol{a}\in\mathbb{R}^{p^*-i},\\ \|\boldsymbol{a}\|_2=1}} \sum_{j=i}^{p^*} a_j^2\xi_{n,j}\boldsymbol{W}_j^{(n)\top}\frac{\boldsymbol{A}^{(n)}}{k_n}\boldsymbol{W}_j^{(n)} \\ &\qquad\qquad + (1-c^2)\|\widetilde{\boldsymbol{W}}^{(n)\top}\frac{\boldsymbol{A}^{(n)}}{k_n}\widetilde{\boldsymbol{W}}^{(n)}\|_2\Bigg\} \\ &\quad\leq \frac{d_n}{\xi_{n,i}} \sup_{c\in[0,1]}\Bigg\{c^2 \sup_{\substack{\boldsymbol{a}\in\mathbb{R}^{p^*-i},\\ \|\boldsymbol{a}\|_2=1}} \sum_{j=i}^{p^*} a_j^2\xi_{n,i}\boldsymbol{W}_j^{(n)\top}\frac{\boldsymbol{A}^{(n)}}{k_n}\boldsymbol{W}_j^{(n)}\end{aligned}$$

$$+ (1-c^2)\|\widetilde{\boldsymbol{W}}^{(n)\top}\frac{\boldsymbol{A}^{(n)}}{k_n}\widetilde{\boldsymbol{W}}^{(n)}\|_2\Big\}$$

$$\leq \sup_{c\in[0,1]}\Big\{c^2 \sup_{\substack{\boldsymbol{a}\in\mathbb{R}^{p^*-i+1},\\ \|\boldsymbol{a}\|_2=1}} \sum_{j=i}^{p^*} a_j^2 \boldsymbol{W}_j^{(n)\top}\frac{d_n\boldsymbol{A}^{(n)}}{k_n}\boldsymbol{W}_j^{(n)}$$

$$+ (1-c^2)\frac{d_n}{\xi_{n,i}}\left\|\widetilde{\boldsymbol{W}}^{(n)\top}\frac{\boldsymbol{A}^{(n)}}{k_n}\widetilde{\boldsymbol{W}}^{(n)}\right\|_2\Big\}.$$

Note that (4.21) and $\widetilde{\boldsymbol{W}}^{(n)}$ being a orthogonal matrix imply that

$$\left\|\widetilde{\boldsymbol{W}}^{(n)\top}\left(\frac{d_n\boldsymbol{A}^{(n)}}{k_n}-\frac{\boldsymbol{B}^{(n)}}{k_n}\right)\widetilde{\boldsymbol{W}}^{(n)}\right\|_2 \leq \left\|\frac{d_n\boldsymbol{A}^{(n)}}{k_n}-\frac{\boldsymbol{B}^{(n)}}{k_n}\right\|_2 \xrightarrow{\mathbb{P}} 0.$$

We then conclude from $\|\widetilde{\boldsymbol{W}}^{(n)\top}\frac{\boldsymbol{B}^{(n)}}{k_n}\widetilde{\boldsymbol{W}}^{(n)}\|_2 \xrightarrow{\mathbb{P}} (1+\sqrt{c})^2$ (cf. proof of Lemma 2.2 (i) Bai et al., 2018) and $\xi_{n,i}\to\infty$ that as $n\to\infty$,

$$\frac{1}{\xi_{n,i}}\left\|\widetilde{\boldsymbol{W}}^{(n)\top}\frac{d_n\boldsymbol{A}^{(n)}}{k_n}\widetilde{\boldsymbol{W}}^{(n)}\right\|_2 \xrightarrow{\mathbb{P}} 0.$$

Additionally, with $\boldsymbol{W}_j^{(n)\top}\frac{d_n\boldsymbol{A}^{(n)}}{k_n}\boldsymbol{W}_j^{(n)} \xrightarrow{\mathbb{P}} 1$ for $j=i+1,\ldots,p^*$ by (4.22) we get,

$$\frac{d_n\widehat{\lambda}'_{n,i}}{\xi_{n,i}} \leq \frac{d_n}{\xi_{n,i}} \sup_{\substack{\boldsymbol{w}\perp\boldsymbol{W}_1^{(n)},\ldots,\boldsymbol{W}_{i-1}^{(n)},\\ \|\boldsymbol{w}\|_2=1}} \boldsymbol{w}^\top\widehat{\boldsymbol{\Sigma}}^{(n)\prime}\boldsymbol{w} \xrightarrow{\mathbb{P}} \sup_{c\in[0,1]}\left\{c^2 \sup_{\sum_{j=i}^{p^*}a_j^2=1}\sum_{j=i}^{p^*}a_j^2\right\} = 1$$

as $n\to\infty$, which proves Step 2. □

Lemma 4.36. *Let Model D with $\xi_{n,p^*}\to\infty$ and $\xi_{n,1}=o(d_n^{1/2})$ as $n\to\infty$ be given. Then as $n\to\infty$,*

$$\sup_{x\in((1-\sqrt{c})^2,(1+\sqrt{c})^2)}\left|F^{d_n\widehat{\boldsymbol{\Sigma}}^{(n)\prime}}(x)-F_c(x)\right| \xrightarrow{\mathbb{P}} 0,$$

where $F^{d_n\widehat{\boldsymbol{\Sigma}}^{(n)\prime}}$ is the empirical spectral distribution function of $d_n\widehat{\boldsymbol{\Sigma}}^{(n)\prime}$ and $F_c(x)$ is defined as in Theorem 4.14.

Proof. For ease of notation define the interval $I := ((1-\sqrt{c})^2,(1+\sqrt{c})^2)$.
Let $F^{\widetilde{\boldsymbol{W}}^{(n)\top}\boldsymbol{B}^{(n)}\widetilde{\boldsymbol{W}}^{(n)}/k_n}$ and $F^{\widetilde{\boldsymbol{W}}^{(n)\top}d_n\boldsymbol{A}^{(n)}\widetilde{\boldsymbol{W}}^{(n)}/k_n}$ be the empirical spectral distribution function of $\widetilde{\boldsymbol{W}}^{(n)\top}\boldsymbol{B}^{(n)}\widetilde{\boldsymbol{W}}^{(n)}/k_n$ and $\widetilde{\boldsymbol{W}}^{(n)\top}d_n\boldsymbol{A}^{(n)}\widetilde{\boldsymbol{W}}^{(n)}/k_n$, respectively. Due to (4.21), it follows by Bai and Silverstein (2010, Theorem A.45) that as $n\to\infty$,

$$\sup_{x\in I}\left|F^{\widetilde{\boldsymbol{W}}^{(n)\top}\frac{\boldsymbol{B}^{(n)}}{k_n}\widetilde{\boldsymbol{W}}^{(n)}}(x)-F^{\widetilde{\boldsymbol{W}}^{(n)\top}\frac{d_n\boldsymbol{A}^{(n)}}{k_n}\widetilde{\boldsymbol{W}}^{(n)}}(x)\right| \xrightarrow{\mathbb{P}} 0.$$

By Silverstein (1995, Theorem 1.1) and Bai and Silverstein (2010, Theorem A.44) combined

with $\operatorname{rank}(\boldsymbol{I} - \boldsymbol{P}^{(n)}) = \operatorname{rank}(\frac{1}{k_n}\mathbf{1}_{k_n}\mathbf{1}_{k_n}^\top) = 1$ there exists a set $\Omega_0 \in \mathcal{F}$ with $\mathbb{P}(\Omega_0) = 1$ so that for any $\omega \in \Omega_0$ the convergence

$$\lim_{n\to\infty} \sup_{x\in I} \left| F^{\widetilde{\boldsymbol{W}}^{(n)\top} \frac{\boldsymbol{B}^{(n)}}{k_n} \widetilde{\boldsymbol{W}}^{(n)}}(x,\omega) - F_c(x) \right| = 0$$

holds which ends in

$$\sup_{x\in I} \left| F^{\widetilde{\boldsymbol{W}}^{(n)\top} \frac{d_n\boldsymbol{A}^{(n)}}{k_n} \widetilde{\boldsymbol{W}}^{(n)}}(x) - F_c(x) \right| \xrightarrow{\mathbb{P}} 0. \tag{4.23}$$

Since the matrices ${\boldsymbol{W}^{(n)}}^\top d_n \widehat{\boldsymbol{\Sigma}}^{(n)\prime} \boldsymbol{W}^{(n)}$ and $d_n \widehat{\boldsymbol{\Sigma}}^{(n)\prime}$ share the same eigenvalues $d_n\widehat{\lambda}'_{n,p^*+1}$, $\ldots, d_n\widehat{\lambda}'_{n,d_n}$, we get for any $i \in \{p^*+1, \ldots, d_n - p^*\}$ with the interlacing theorem for the eigenvalues (Horn and Johnson, 2013, Theorem 4.3.28) $\mathbb{P}$-a.s. that

$$\begin{aligned} \lambda_i\left(\widetilde{\boldsymbol{W}}^{(n)\top} \frac{d_n\boldsymbol{A}^{(n)}}{k_n} \widetilde{\boldsymbol{W}}^{(n)}\right) &\geq \lambda_{p^*+i}\left({\boldsymbol{W}^{(n)}}^\top d_n \widehat{\boldsymbol{\Sigma}}^{(n)\prime} \boldsymbol{W}^{(n)}\right) \\ &= d_n\widehat{\lambda}'_{n,p^*+i} \\ &\geq \lambda_{p^*+i}\left(\widetilde{\boldsymbol{W}}^{(n)\top} \frac{d_n\boldsymbol{A}^{(n)}}{k_n} \widetilde{\boldsymbol{W}}^{(n)}\right). \end{aligned} \tag{4.24}$$

Therefore, due to (4.23) and (4.24),

$$\begin{aligned} \sup_{x\in I} \left| F^{d_n\widehat{\boldsymbol{\Sigma}}^{(n)\prime}}(x) - F_c(x) \right| &= \sup_{x\in I} \left| \frac{1}{d_n} \sum_{i=1}^{d_n} \mathbb{1}\left\{ d_n\widehat{\lambda}'_{n,i} \leq x \right\} - F_c(x) \right| \\ &\leq \sup_{x\in I} \left| F^{\widetilde{\boldsymbol{W}}^{(n)\top} \frac{d_n\boldsymbol{A}^{(n)}}{k_n} \widetilde{\boldsymbol{W}}^{(n)}}(x) - F_c(x) \right| + \frac{4p^*}{d_n} \xrightarrow{\mathbb{P}} 0, \end{aligned}$$

which is the statement. □

Proof of Theorem 4.15. The proof of Theorem 4.15 (a)-(d) follows with the same arguments as the proof of Lemma 4.34 using only Lemma 4.35 and Lemma 4.36 in combination with $\widehat{\boldsymbol{\Sigma}}^{(n)} \overset{\mathcal{D}}{=} \widehat{\boldsymbol{\Sigma}}^{(n)\prime}$ (cf. (4.9)). Only the proof (e) remains. Therefore, note that for $i < p^*$ the asymptotic behavior $\frac{d_n\widehat{\lambda}_{n,i}}{\xi_{n,i}} \xrightarrow{\mathbb{P}} 1$ and $\frac{1}{d_n - i}\sum_{j=p^*+1}^{d_n} \frac{d_n\widehat{\lambda}_{n,j}}{\xi_{n,i}} \xrightarrow{\mathbb{P}} 0$ as $n \to \infty$ hold by (a) and (d), respectively. Hence,

$$\begin{aligned} \frac{d_n\widehat{\lambda}_{n,i}}{\frac{1}{d_n-i}\sum_{j=i+1}^{d_n} d_n\widehat{\lambda}_{n,j}} &= \frac{d_n\widehat{\lambda}_{n,i}}{\frac{1}{d_n-i}\sum_{j=i+1}^{p^*} d_n\widehat{\lambda}_{n,j} + \frac{1}{d_n-i}\sum_{j=p^*+1}^{d_n} d_n\widehat{\lambda}_{n,j}} \\ &\geq \frac{\frac{d_n\widehat{\lambda}_{n,i}}{\xi_{n,i}}}{\frac{p^*-i}{d_n-i}\frac{d_n\widehat{\lambda}_{n,i}}{\xi_{n,i}} + \frac{1}{d_n-i}\sum_{j=p^*+1}^{d_n} \frac{d_n\widehat{\lambda}_{n,j}}{\xi_{n,i}}} \xrightarrow{\mathbb{P}} \infty, \end{aligned}$$

which shows (e). □

4.5.2. Proofs of Section 4.3

Proof of Theorem 4.19. Since by Remark 4.18 (b) the AIC is scale invariant and hence, we assume w.l.o.g. that $\lambda = 1$.
Step 1: Suppose $p > p^*$. Note

$$2(p+1)(d-p/2) - (d-1)(d+2) = -(d-p-2)(d-p+1).$$

By the definition of the AIC we obtain

$$\begin{aligned}\mathrm{AIC}_{k_n}(p) - \mathrm{AIC}_{k_n}(p^*) = k_n \sum_{i=p^*+1}^{p} \log(\widehat{\lambda}_{n,i}) + k_n(d-1-p)\log\left(\frac{1}{d-1-p}\sum_{j=p+1}^{d-1}\widehat{\lambda}_{n,j}\right)\\ - k_n(d-1-p^*)\log\left(\frac{1}{d-1-p^*}\sum_{j=p^*+1}^{d-1}\widehat{\lambda}_{n,j}\right)\\ - (d-p-2)(d-p+1) + (d-p^*-2)(d-p^*+1),\end{aligned}$$

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

$$(\widehat{\lambda}_{n,p^*+1}, \ldots, \widehat{\lambda}_{n,d})^\top = \mathbf{1}_{d-p^*} + \frac{1}{\sqrt{k_n}}\boldsymbol{M}_n,$$

where

$$\boldsymbol{M}_n \coloneqq \sqrt{k_n}((\widehat{\lambda}_{n,p^*+1}, \ldots, \widehat{\lambda}_{n,d})^\top - \mathbf{1}_{d-p^*}),$$

gives that

$$\begin{aligned}\mathrm{AIC}_{k_n}(p) - \mathrm{AIC}_{k_n}(p^*) = k_n \sum_{i=p^*+1}^{p} \log\left(1 + \frac{1}{\sqrt{k_n}}M_{n,i}\right)\\ + k_n(d-1-p)\log\left(1 + \frac{1}{d-1-p}\sum_{j=p+1}^{d-1}\frac{1}{\sqrt{k_n}}M_{n,j}\right)\\ - k_n(d-1-p^*)\log\left(1 + \frac{1}{d-1-p^*}\sum_{j=p^*+1}^{d-1}\frac{1}{\sqrt{k_n}}M_{n,j}\right)\\ - (d-p-2)(d-p+1) + (d-p^*-2)(d-p^*+1).\end{aligned}$$

Furthermore, $\boldsymbol{M}_n = O_{\mathbb{P}}(1)$ due to Theorem 4.10 (b). Additionally the Taylor expansion of the logarithm as $x \to 0$,

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

gives that

$$\begin{aligned}&\mathrm{AIC}_{k_n}(p) - \mathrm{AIC}_{k_n}(p^*)\\ &\quad = k_n \sum_{i=p^*+1}^{p}\left(\frac{1}{\sqrt{k_n}}M_{n,i} - \frac{1}{2}\frac{1}{k_n}M_{n,i}^2 + O_{\mathbb{P}}(k_n^{-3/2})\right)\end{aligned}$$

$$
\begin{aligned}
&+ k_n \left(\sum_{j=p+1}^{d-1} \frac{1}{\sqrt{k_n}} M_{n,j} - \frac{1}{2(d-1-p)} \left(\sum_{j=p+1}^{d-1} \frac{1}{\sqrt{k_n}} M_{n,j} \right)^2 + O_{\mathbb{P}}(k_n^{-3/2}) \right) \\
&- k_n \left(\sum_{j=p^*+1}^{d-1} \frac{1}{\sqrt{k_n}} M_{n,j} - \frac{1}{2(d-1-p^*)} \left(\sum_{j=p^*+1}^{d-1} \frac{1}{\sqrt{k_n}} M_{n,j} \right)^2 + O_{\mathbb{P}}(k_n^{-3/2}) \right) \\
&- (d-p-2)(d-p+1) + (d-p^*-2)(d-p^*+1) \\
&= k_n \sum_{i=p^*+1}^{p} \left(-\frac{1}{2} \frac{1}{k_n} M_{n,i}^2 + O_{\mathbb{P}}(k_n^{-3/2}) \right) \\
&+ k_n \left(-\frac{1}{2(d-1-p)} \left(\sum_{j=p+1}^{d-1} \frac{1}{\sqrt{k_n}} M_{n,j} \right)^2 + O_{\mathbb{P}}(k_n^{-3/2}) \right) \\
&- k_n \left(-\frac{1}{2(d-1-p^*)} \left(\sum_{j=p^*+1}^{d-1} \frac{1}{\sqrt{k_n}} M_{n,j} \right)^2 + O_{\mathbb{P}}(k_n^{-3/2}) \right) \\
&- (d-p-2)(d-p+1) + \frac{1}{2}(d-p^*-2)(d-p^*+1) \\
&= -\frac{1}{2} \sum_{i=p^*+1}^{p} M_{n,i}^2 - \frac{1}{2(d-1-p)} \left(\sum_{j=p+1}^{d-1} M_{n,j} \right)^2 + \frac{1}{2(d-1-p^*)} \left(\sum_{j=p^*+1}^{d-1} M_{n,j} \right)^2 \\
&- (d-p-2)(d-p+1) + (d-p^*-2)(d-p^*+1) + O_{\mathbb{P}}(k_n^{-1/2}).
\end{aligned}
$$

An application of Theorem 4.10 (b) gives then

$$
\begin{aligned}
&\mathrm{AIC}_{k_n}(p) - \mathrm{AIC}_{k_n}(p^*) \qquad (4.25) \\
&\xrightarrow{\mathcal{D}} -\frac{1}{2} \sum_{i=p^*+1}^{p} M_i^2 - \frac{1}{2(d-1-p)} \left(\sum_{j=p+1}^{d-1} M_j \right)^2 + \frac{1}{2(d-1-p^*)} \left(\sum_{j=p^*+1}^{d-1} M_j \right)^2 \\
&\quad - (d-p-2)(d-p+1) + (d-p^*-2)(d-p^*+1). \qquad (4.26)
\end{aligned}
$$

Hence, the assertion follows.

Step 2: Suppose $p < p^*$. Again by the definition of the AIC we receive

$$
\begin{aligned}
\frac{\mathrm{AIC}_{k_n}(p) - \mathrm{AIC}_{k_n}(p^*)}{k_n} &= -\sum_{j=p+1}^{p^*} \log(\widehat{\lambda}_{n,j}) + (d-1-p) \log \left(\frac{1}{d-1-p} \sum_{j=p+1}^{d-1} \widehat{\lambda}_{n,j} \right) \\
&\quad - (d-1-p^*) \log \left(\frac{1}{d-1-p^*} \sum_{j=p^*+1}^{d-1} \widehat{\lambda}_{n,j} \right) \\
&\quad - \frac{(d-p-2)(d-p+1) + (d-p^*-2)(d-p^*+1)}{k_n}.
\end{aligned}
$$

Due to Theorem 4.10 (a), $\widehat{\lambda}_{n,i} \xrightarrow{\mathbb{P}} \lambda_i$ for $i = 1, \dots, d-1$ holds and therefore,

$$\frac{\mathrm{AIC}_{k_n}(p) - \mathrm{AIC}_{k_n}(p^*)}{k_n} \quad \xrightarrow{\mathbb{P}} \quad - \sum_{j=p+1}^{p^*} \log(\lambda_j) + (d-1-p) \log\Big(\frac{1}{d-1-p} \sum_{j=p+1}^{d-1} \lambda_j\Big)$$

Using $\lambda_i = 1$ for $i = p^*+1, \dots, d-1$ we get

$$\begin{aligned}
&- \sum_{j=p+1}^{p^*} \log(\lambda_j) + (d-1-p) \log\Big(\frac{1}{d-1-p} \sum_{j=p+1}^{d-1} \lambda_j\Big) \\
&\qquad = - \sum_{j=p+1}^{d} \log(\lambda_j) + (d-1-p) \log\Big(\frac{1}{d-1-p} \sum_{j=p+1}^{d-1} \lambda_j\Big) \\
&\qquad = - \log\left(\frac{\prod_{j=p+1}^{d} \lambda_j}{\big(\frac{1}{d-1-p} \sum_{j=p+1}^{d-1} \lambda_j\big)^{(d-1-p)}} \right) > 0,
\end{aligned}$$

due to the inequality of arithmetic and geometric means (Uchida, 2008) which says that

$$\frac{\big(\prod_{j=p+1}^{d} \lambda_j\big)^{1/(d-1-p)}}{\frac{1}{d-1-p} \sum_{j=p+1}^{d-1} \lambda_j} < 1. \qquad \square$$

Proof of Theorem 4.22.

Step 1: Suppose $p > p^*$. Due to (4.25) we receive

$$\begin{aligned}
&\mathrm{BIC}_{k_n}(p) - \mathrm{BIC}_{k_n}(p^*) \\
&\qquad = \frac{\log(k_n)}{2}(d-p^*-2)(d-p^*+1) - \frac{\log(k_n)}{2}(d-p-2)(d-p+1) + O_{\mathbb{P}}(1).
\end{aligned}$$

A division by $\log(k_n)$ provides

$$\frac{\mathrm{BIC}_{k_n}(p) - \mathrm{BIC}_{k_n}(p^*)}{\log(k_n)} \xrightarrow{\mathbb{P}} \frac{1}{2}(d-p^*-2)(d-p^*+1) - \frac{1}{2}(d-p-2)(d-p+1),$$

which is strictly positive.

Step 2: Suppose $p < p^*$. Since as $n \to \infty$,

$$\begin{aligned}
&\frac{\mathrm{BIC}_{k_n}(p) - \mathrm{BIC}_{k_n}(p^*)}{k_n} - \frac{\mathrm{AIC}_{k_n}(p) - \mathrm{AIC}_{k_n}(p^*)}{k_n} \\
&\qquad = \frac{\log(k_n) - 2}{k_n} \left(\frac{(d-p^*-2)(d-p^*+1)}{2} - \frac{(d-p-2)(d-p+1)}{2} \right) \to 0,
\end{aligned}$$

the statement follows from Theorem 4.19.

$\square$

4.5.3. PROOFS OF SECTION 4.4

Proof of Theorem 4.24. Note, as stated in Remark 4.18, the information criteria are scale invariant and hence

$$\mathrm{AIC}^{\circ}_{k_n}(p_n; \widehat{\lambda}_{n,1}, \ldots, \widehat{\lambda}_{n,d_n-1}) =: \mathrm{AIC}^{\circ}_{k_n}(p_n) = \mathrm{AIC}^{\circ}_{k_n}(p_n; d_n\widehat{\lambda}_{n,1}, \ldots, d_n\widehat{\lambda}'_{n,d_n-1}).$$

Due to Theorem 4.14 for (a,b) and Theorem 4.15 for (c), the proof of Bai et al. (2018, Theorem 3.1) for $\widehat{\xi}_{n,1}, \ldots, \widehat{\xi}_{n,d_n-1}$ can be carried out step by step for $d_n\widehat{\lambda}_{n,1}, \ldots, d_n\widehat{\lambda}_{n,d_n-1}$. The only difference is that there we have almost sure convergence and here we have convergence in probability. □

Proof of Theorem 4.26. Due to the scale invariance of the $\mathrm{BIC}^{\circ}_{k_n}(p^*)$, $\log(d_n)/\log(k_n) \to 1$ as $n \to \infty$, Theorem 4.14 and Theorem 4.15, the proof of Bai et al. (2018, Theorem 3.2) for $\widehat{\xi}_{n,1}, \ldots, \widehat{\xi}_{n,d_n-1}$ can be carried out step by step for $d_n\widehat{\lambda}_{n,1}, \ldots, d_n\widehat{\lambda}_{n,d_n-1}$. □

Proof of Theorem 4.29. Due to the scale invariance of the AIC^*, Theorem 4.14 and Theorem 4.15, the proof of Bai et al. (2018, Theorem 3.3) for $\widehat{\xi}_{n,1}, \ldots, \widehat{\xi}_{n,d_n-1}$ can be carried out step by step for $d_n\widehat{\lambda}_{n,1}, \ldots, d_n\widehat{\lambda}_{n,d_n-1}$. □

Proof of Theorem 4.30. Due to the scale invariance of the BIC^*, $\log(d_n)/\log(k_n) \to 1$ as $n \to \infty$, Theorem 4.14 and Theorem 4.15, the proof of Bai et al. (2018, Theorem 3.4) for $\widehat{\xi}_{n,1}, \ldots, \widehat{\xi}_{n,d_n-1}$ can be carried out step by step for $d_n\widehat{\lambda}_{n,1}, \ldots, d_n\widehat{\lambda}_{n,d_n-1}$. □

CHAPTER 5

Simulation study

Until now, we considered question **(Q)** from a theoretic viewpoint. In this chapter, we compare the performance of the different information criteria through simulation studies and real-world data. For the simulation study, we start with the SRV approach from Chapter 3 in Section 5.1. In the second part of this chapter, Section 5.2 we employ the PCA approach from Chapter 4. Then, we apply them to real-world data in Section 5.3.

All calculations are performed in R (R Core Team, 2025) with RStudio (Posit team, 2025). Parts of the code for the following simulation study are available at:
`https://gitlab.kit.edu/projects/164856`,
`https://gitlab.kit.edu/projects/178647`.

5.1. Simulation study: Directions of extremes

In this section, we consider the information criteria from Section 3.2. We simulate n times a multivariate regularly varying random vector $\boldsymbol{X}$ of dimension d.

In Section 5.1.1 we give a brief overview of the error measures used. For the distribution of $\boldsymbol{X}$, we distinguish two cases: Either $\boldsymbol{X}$ exhibits asymptotic independence (Section 5.1.2) or asymptotic dependence (Sections 5.1.3 and 5.1.4); these examples can also be found in Meyer and Wintenberger (2023). In the first two examples, we also consider the high-dimensional case, where the dimension is large. In all 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 (3.4). For comparison, we run simulations for the local model with a fixed k_n and for the global model with an estimated k_n. Since r is not known, we use the estimator

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

at this point also for the fixed-dimensional case. In total, we conducted 500 repetitions for each setting.

5.1.1. Error measures

To quantify the discrepancy between the true distribution $\boldsymbol{p}$ as defined in Section 3.1.3 and the estimated distribution $\widehat{\boldsymbol{p}}_{n,*}^{\widehat{s}_n^*}$ in (3.4) we use different measures. We start with the Hellinger distance, which is for two probability measures $\mathbb{P}$ and $\mathbb{Q}$ with densities p and q

defined by

$$H(P,Q) \coloneqq \sqrt{\frac{1}{2}\int \left(\sqrt{p(x)} - \sqrt{q(x)}\right)^2 \mathrm{d}x} = \sqrt{1 - \int \sqrt{p(x)q(x)}\,\mathrm{d}x}.$$

Since $\left(\sqrt{p(x)} - \sqrt{q(x)}\right)^2$ and $\sqrt{p(x)q(x)}$ are always positive, we obtain 0 and 1 as lower and upper bounds of the Hellinger distance. Thus, it follows that $H(P,Q) \in [0,1]$. For discrete probability measures $\mathbb{P}$ and $\mathbb{Q}$ with probabilities $p_1, \ldots, p_m$ and $q_1, \ldots, q_m$ for $m \in \mathbb{N}$ the Hellinger distance is given by $H(P,Q) \coloneqq \frac{1}{\sqrt{2}}\|\sqrt{\boldsymbol{p}} - \sqrt{\boldsymbol{q}}\|_2$ where $\boldsymbol{p} = (p_1, \ldots, p_m)^\top$ and $\boldsymbol{q} = (q_1, \ldots, q_m)^\top$. The Hellinger distance between $\mathbb{P}$ and $\mathbb{Q}$ is equal to zero if and only if $p = q$ almost sure. On the other hand, $H(P,Q)$ is equal to 1 if and only if $p \cdot q = 0$ almost sure.

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 that an information criterion gives $\widehat{s}_n^*$ 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)	$\widehat{s}_n^*$
IC detects no direction	False negative (FN)	True negative (TN)	$2^d - 1 - \widehat{s}_n^*$
#Directions	s^*	$2^d - 1 - s^*$	$2^d - 1$

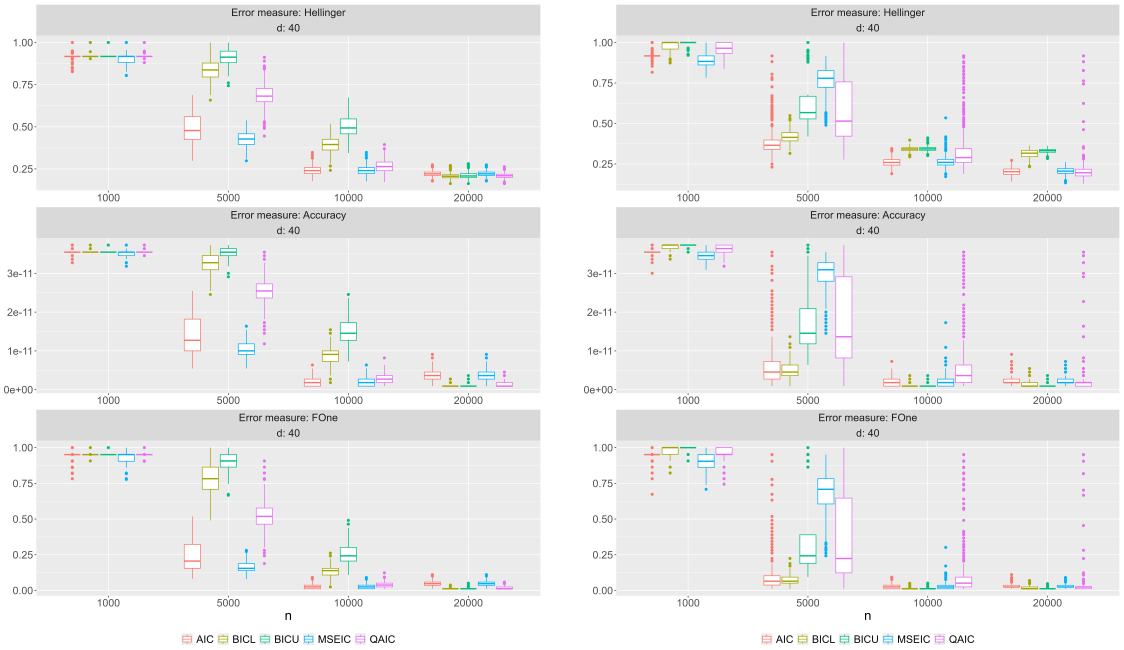
and as error measures

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

which reflect the errors. If we take 1 − Accuracy Error and $1 - F_1$ Error, respectively, we obtain the original definition of Accuracy and F_1 as in Powers (2008) 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.

5.1.2. Asymptotically tail independent model

In the first example, we consider d-dimensional i.i.d. random vectors whose spectral measure only concentrates on the axes. For the distribution of the random vectors, we use the model



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

(b) Global model

Figure 5.1.: *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.*

introduced in Section 3.3.2 and set $d = s^* = 40$, i.e. there are no normally distributed random variables added.

The results are presented in Figure 5.1, 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 the 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 positive probability (Theorem 3.7 and Theorem 3.20), whereas the QAIC and BIC are consistent (Theorem 3.14 and Theorem 3.28). If we compare the simulation results for the local model (left part of Figure 5.1) with the results for the global model (right part of Figure 5.1), 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.

For the high-dimensional example, we assume that $k_n/n = 0.1$ is fixed and we consider $s^* = 50, 75, 100$ and $d_n = s^*, s^* + 25, s^* + 50$ leading to $2^{50} - 1 \approx 10^{15.05}$ to $2^{150} - 1 \approx 10^{45.15}$ potential directions. Since s^* is relatively large compared to $\sqrt{k_n}$ and k_n, we selected

$q_n = k_n$. The Hellinger distance is plotted in Figure 5.2, where a lower value is better and the number of estimated dimensions in Figure 5.3, where a value closer to s^* is better.

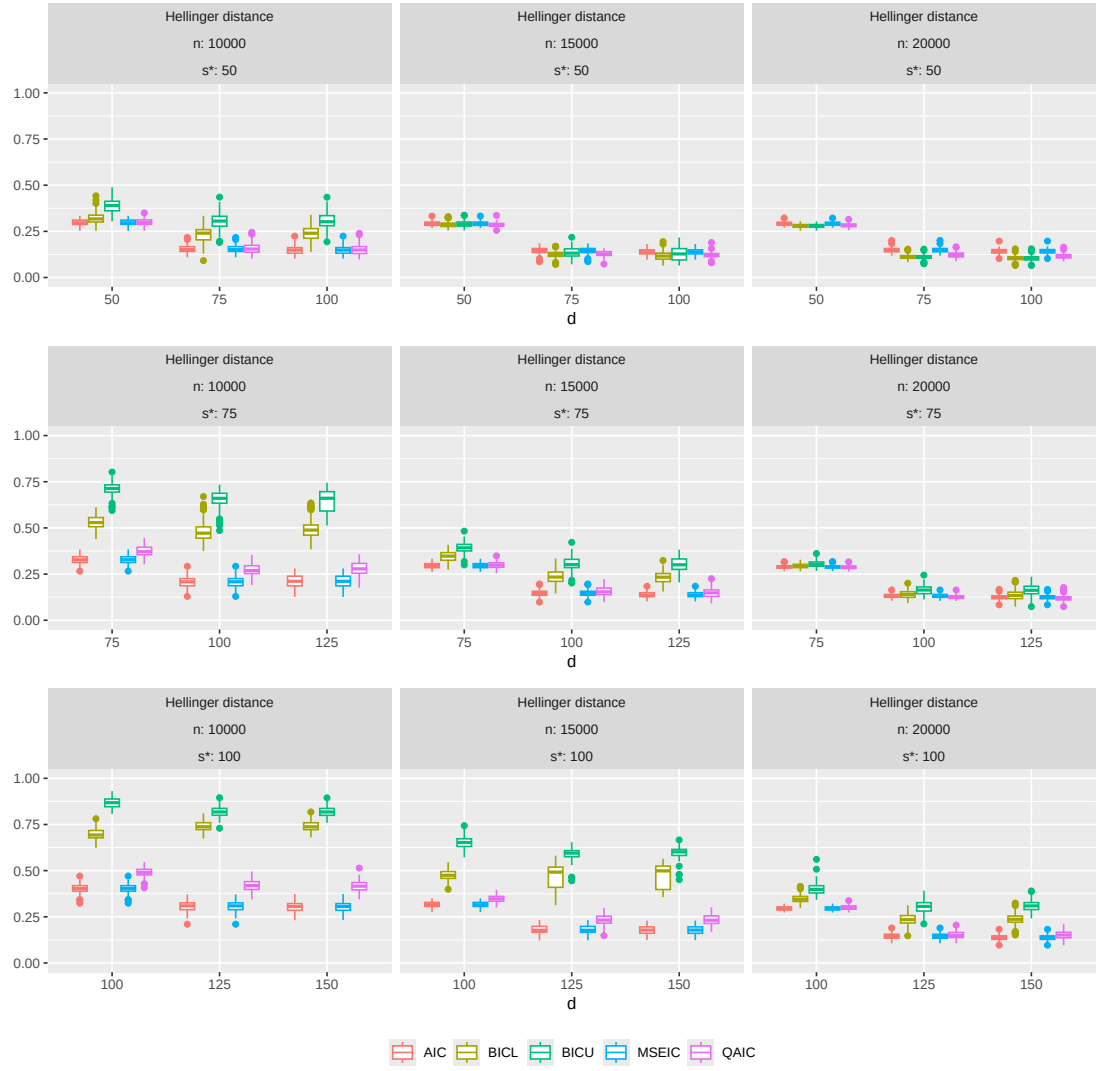


Figure 5.2.: *Simulations for asymptotically independent data and Normal noise with* $s^* = 50, 75, 100$ *and* $d_n = s^*, s^* + 25, s^* + 50$*: The Hellinger distance is plotted against the dimension on the* x*-axis.*

In Figure 5.2 we see that, generally, the estimation improves when n becomes larger. The performance of the BIC declines with an increase in dimension. This is also visible in Figure 5.3, where the BIC tends to underestimate the number of relevant directions heavily when the dimension is large. For the AIC it is different, where in the low-dimensional case the AIC has the tendency to overestimate the number of relevant directions and not in the high-dimensional case. After an investigation, it became evident that the underestimation by the BIC results from the stronger penalization, which overpowers the log-likelihood term for small n.

In Table 5.1 the ratio $\widehat{s}_n/k_n$ is displayed. The values of the ratio are between 0.43 and

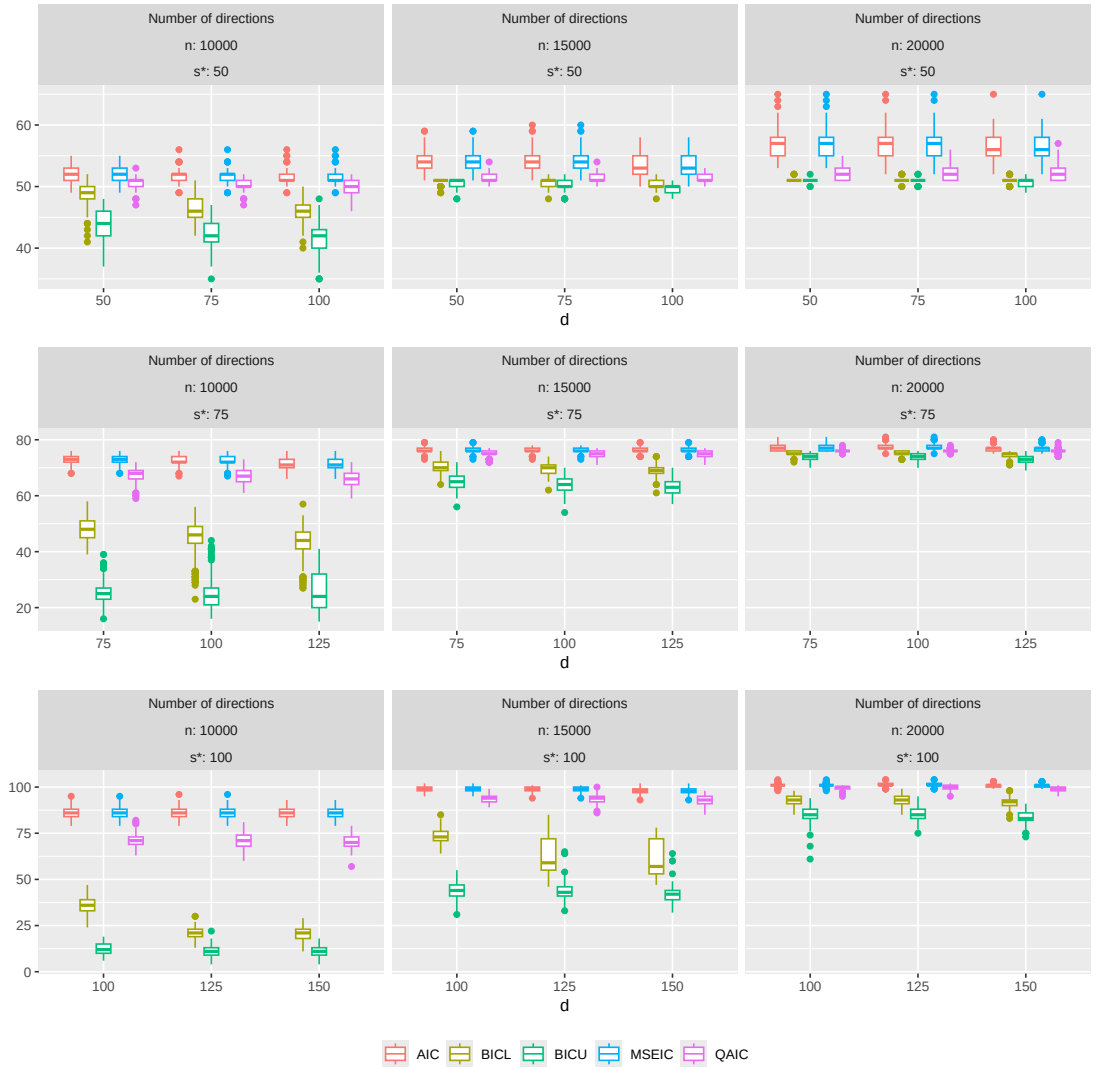


Figure 5.3.: *Simulations for asymptotically independent data and Normal noise with* $s^* = 50, 75, 100$ *and* $d_n = s^*, s^* + 25, s^* + 50$*: The estimated number of relevant directions is plotted against the dimension on the* x*-axis.*

0.624 and hence not at the border of the interval $(0, 1)$. For increasing n and d the ratio is relatively stable, which is in agreement with the assumption that $\widehat{s}_n/k_n$ converges.

5.1.3. 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 (2020). Consequently, not only directions 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

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

s^*	d	$n = 10000$	$n = 15000$	$n = 20000$
50	50	0.466	0.447	0.430
	75	0.553	0.527	0.516
	100	0.563	0.542	0.528
75	75	0.508	0.477	0.466
	100	0.584	0.556	0.542
	125	0.595	0.564	0.554
100	100	0.535	0.504	0.485
	125	0.612	0.579	0.562
	150	0.624	0.591	0.569

Table 5.1.: *Mean of the ratio $\widehat{s}_n/k_n$ for asymptotically independent data and Normal noise with $s^* = 50, 75, 100$ and $d = s^*, s^* + 25, s^* + 50$.*

The parameters d_1, d_2, d_3 specify the number of one-, two-, and three-dimensional directions. In the following, we denote by Exp(1) the exponential distribution with parameter 1. The marginal distributions of $\boldsymbol{X}$ are defined by

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

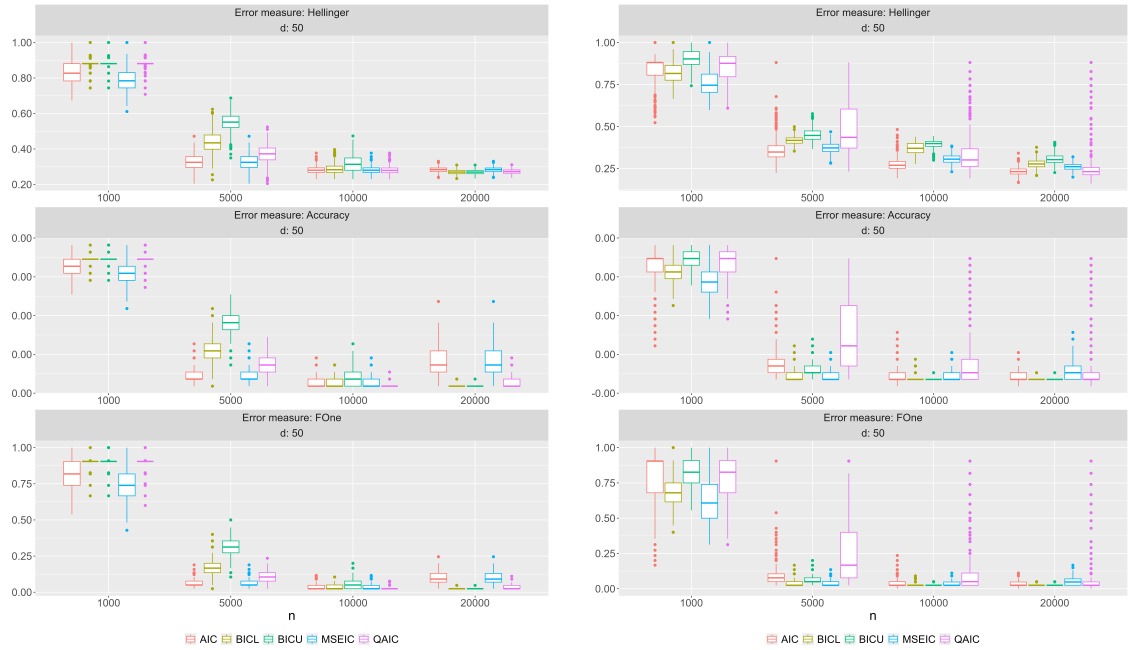
The random vector $\boldsymbol{Z}$ from Definition 2.13 puts mass on the sets

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

In total, there are $s^* = d_1 + d_2 + d_3$ directions with probability mass, and the goal is again to identify these directions. For the simulation study in Figure 5.4 we chose $k_n/n = 0.05$ for the fixed case, $d_1 = 10, d_2 = d_3 = 5$ and $d = 50$ resulting in $s^* = 20$ extreme directions. The plots show similar features as for the asymptotic independent case in Section 5.1.2 (cf. Figure 5.1).

For the high-dimensional asymptotic dependent simulation study, we chose $d_1 = 10, d_2 = 0, 15, 30$, $d_3 = 5$ and $d = 50, 100$ and 200 resulting in $s^* = 20, 35, 50$ extreme directions. Furthermore, $k_n/n = 0.05$ is fixed.

In Figure 5.5, the Hellinger distance of the AIC and the BIC decreases for larger n. However, in Figure 5.6 we see a similar behavior of the AIC and BIC as in the asymptotic independent case in Section 5.1.2 (cf. Figures 5.2 and 5.3), where the AIC tends to



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

(b) Global model

Figure 5.4.: *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.*

overestimate s^* when the dimension is low and the BIC underestimates s^* in this case.

The behavior of the ratio $\widehat{s}_n/k_n$ in Table 5.2 is similar to before and in agreement with the assumption of the convergence of it.

5.1.4. Max-mixture model

We explore an additional simulation study for the max-mixture model of Simpson et al. (2020), which exhibits asymptotic dependence. For $\beta \in \mathcal{P}_d^*$ and $d = 5$ suppose $\boldsymbol{F}_\beta = (F_{\beta,j})_{j\in\beta}$ is a $|\beta|$-dimensional random vector with Fréchet(1) distributed margins and the following dependence structure. First, $\boldsymbol{F}_{\{1,2\}}$ and $\boldsymbol{F}_{\{4,5\}}$ have a bivariate Gaussian copula with correlation parameter $\rho = 0.25$. On the other hand, $\boldsymbol{F}_{\{1,2,3\}}, \boldsymbol{F}_{\{3,4,5\}}$ and $\boldsymbol{F}_{\{1,2,3,4,5\}}$ have a three-dimensional and five-dimensional extreme value logistic copula, respectively, with dependence parameter ϑ. Then the regularly varying vector $\boldsymbol{X} \in \mathbb{R}^5$ of index -1 is defined as

$$\boldsymbol{X} := (X_1, \dots, X_5)^\top := \begin{pmatrix} \max\left\{\frac{5}{7}F_{\{1,2\},1},\ \frac{1}{7}F_{\{1,2,3\},1},\ \frac{1}{7}F_{\{1,2,3,4,5\},1}\right\} \\ \max\left\{\frac{5}{7}F_{\{1,2\},2},\ \frac{1}{7}F_{\{1,2,3\},2},\ \frac{1}{7}F_{\{1,2,3,4,5\},2}\right\} \\ \max\left\{\frac{3}{7}F_{\{1,2,3\},3},\ \frac{3}{7}F_{\{3,4,5\},3},\ \frac{1}{7}F_{\{1,2,3,4,5\},3}\right\} \\ \max\left\{\frac{5}{7}F_{\{4,5\},4},\ \frac{1}{7}F_{\{3,4,5\},4},\ \frac{1}{7}F_{\{1,2,3,4,5\},4}\right\} \\ \max\left\{\frac{5}{7}F_{\{4,5\},5},\ \frac{1}{7}F_{\{3,4,5\},5},\ \frac{1}{7}F_{\{1,2,3,4,5\},5}\right\} \end{pmatrix}.$$

Figure 5.5.: *Simulations for asymptotic dependent data with* $s^* = 20, s^* = 35$ *and* $s^* = 50$ *relevant directions,* $d = 100, d = 200$ *and* $d = 300$*, Exponential noise and as error measure the Hellinger distance, which is plotted on the y-axis against the dimension* d *on the* x*-axis.*

Since the Gaussian copula exhibits pairwise asymptotic independence, the random vector $\boldsymbol{\Theta}$ puts mass on the cones $C_{\{1\}}$, $C_{\{2\}}$, $C_{\{4\}}$, $C_{\{5\}}$, $C_{\{1,2,3\}}$, $C_{\{3,4,5\}}$, $C_{\{1,2,3,4,5\}}$ and by the choice of the scaling factors, each cone has the same probability. However, the distribution of $\boldsymbol{Z}$ is not discrete and we need to estimate the support of $\boldsymbol{Z}$ via a Monte-Carlo simulation, where we use the implementation of Meyer and Wintenberger (2023). Since the choice of weights is fixed for dimension $d = 5$ we only consider this example in the fixed-dimensional setting.

The simulation results of this 5-dimensional model with $s^* = 7$ are presented in Figure 5.7. In this simulation study, the dependence parameter ϑ takes values 0.1, 0.5 and 0.9 and the sample sizes are $n = 1000$, 5000, 10000 and 20000. As before, we conduct 500 repetitions. We report only the Hellinger distance, as the Accuracy error and the F_1 error are not informative in this context. This is because, in the Monte Carlo simulation used to estimate the probabilities of the cones (which are not known explicitly), all $2^5 - 1 = 31$ possible

Figure 5.6.: *Simulations for asymptotic dependent data with* $s^* = 20, s^* = 35$ *and* $s^* = 50$ *relevant directions,* $d = 100, d = 200$ *and* $d = 300$*, Exponential noise and the number of estimated relevant directions, which is plotted on the* y*-axis against the dimension* d *on the* x*-axis.*

cones were detected and thus classified as a relevant direction. The figure shows similar patterns across all information criteria. In particular, as the sample size n increases, the performance improves. The dependence parameter ϑ does not appear to have a strong impact on the information criteria. However, for $n = 1000$, the Hellinger distance tends to be smaller when ϑ is higher, suggesting a potential influence at smaller sample sizes.

5.2. Simulation study: PCA for multivariate extremes

In this section, we present the simulation study for the approach in Chapter 4. Again, we simulate n times a regularly varying random vector $\boldsymbol{X}$ of dimension d_n. For the distribution of $\boldsymbol{X}$, we distinguish three models. First, in Section 5.2.1 we use the directional model and in Section 5.2.2, we extend the directional model by adding an additional noise term. Finally, the model in Section 5.2.3 exhibits asymptotic dependence but differs from

s^*	d	$n = 10000$	$n = 15000$	$n = 20000$
20	100	0.389	0.373	0.364
	200	0.457	0.442	0.428
	300	0.512	0.491	0.483
35	100	0.411	0.387	0.375
	200	0.449	0.426	0.412
	300	0.477	0.455	0.441
50	100	0.437	0.404	0.384
	200	0.466	0.430	0.418
	300	0.488	0.452	0.435

Table 5.2.: *Mean of the ratio* $\widehat{s}_n/k_n$ *for asymptotic dependent data with* $s^* = 20, s^* = 35$ *and* $s^* = 50$ *relevant directions,* $d = 100, d = 200$ *and* $d = 300$.

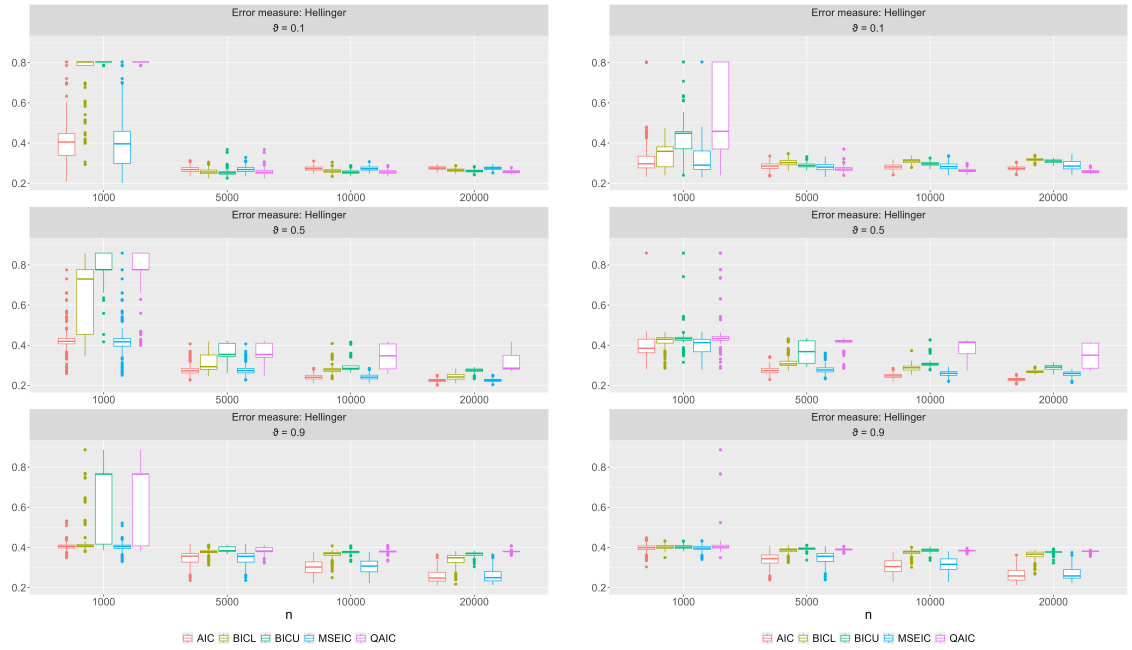
the directional model. In all models, we estimate the parameter p^* by $\widehat{p}_n$ based on n observations. We run the simulations with 500 repetitions. Throughout these examples, $c = d_n/k_n$. When $c < 1$ we use the AIC and the BIC, and if $c > 1$ we use the AIC* and the BIC*. If for some c, k_n is greater than n, we set $k_n = n$.

Note that we do not directly compare the approaches from Chapter 3 and Chapter 4, as the requirements for the support of the noise are vastly different. In Chapter 3 we assumed that there are only a few directions and in Chapter 4 we assumed that we have mass on every axis.

5.2.1. DIRECTIONAL MODEL

First, we consider the directional model with $p^* = 10$ as introduced in Section 4.2.2. On the one hand, we investigate the fixed-dimensional case with $d = 20$ and on the other hand the high-dimensional case with $d = 100$, 200 and 300. For comparison, we run simulations with sample sizes $n = 1000$, 5000, 10000. The matrix $\mathbf{\Gamma}_n$ from (4.3) is a diagonal matrix and the deterministic eigenvalues $\xi_{n,1}, \ldots, \xi_{n,p^*}$ are all equal to λ^*, which is chosen to be larger than 1 and to satisfy the distant spiked eigenvalue condition $\lambda^* > 1 + \sqrt{c}$. The entries of $\boldsymbol{V}^{(n)}$ are i.i.d. standard normally distributed.

The results for $d = 20$ are presented in Figure 5.8. The estimator $\widehat{p}_n$ of both information criteria gets closer to the true value $p^* = 10$ if k_n increases. For $n = 1000$ and $k_n/n = 0.01$, we have $k_n = 10 < d = 20$ and therefore we use the AIC* and BIC*. Both information criteria underestimate p^*, which is expected as the number of extreme observations k_n equals p^*. In all other cases, the AIC and BIC are used. For $k_n/n \geq 0.05$ and $\lambda^* = 3$, the AIC either estimates p^* or shows more outliers above p^*. Overall, the AIC performs better when λ^* or k_n increases. The BIC estimates the true value of p^* or underestimates p^*, where the number of cases with underestimation decreases when λ^* or k_n grows. This is also intuitive: for a higher value of λ^*, the spike is more pronounced. In comparison to the AIC, the estimates of the BIC have, in general, fewer fluctuations and outliers.



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

(b) Global model

Figure 5.7.: *Simulations for the max-mixture model with $s^* = 7$ directions of extremes and $d = 5$: From top to the bottom, the dependence parameter increases from $\vartheta = 0.1$, $\vartheta = 0.5$ to $\vartheta = 0.9$. The Hellinger distance is plotted against the sample size n on the x-axis.*

For the high-dimensional case $d \geq 100$, Figure 5.9 depicts the simulation results. Note that for $\lambda^* = 3$ the gap condition is satisfied when $c < 1$, and for $\lambda^* = 5$ and 20 for all c. It should also be noted that for fixed n and d_n but increasing c, the number of extreme observations k_n decreases, leading to a smaller sample size. The AIC and AIC* both profit from an increase in dimension and λ^*. Overall, the estimates of both criteria get better for a larger dimension. In comparison to Figure 5.8, we see that the AIC* has a tendency to underestimate p^* for $\lambda^* \leq 5$, $c = 2$ and $c = 3$, which is consistent with Theorem 4.29. The estimates $\widehat{p}_n$ of the AIC and AIC* are closer to p^* in comparison to the BIC and BIC* as soon as the gap condition is fulfilled. When the gap condition is not satisfied, the information criteria underestimate p^*, where for $c \geq 0.5$ the BIC and BIC* only give usable results for $\lambda^* = 20$. For $c > 1$ the BIC* shows underestimation in all cases.

5.2.2. Directional model with noise

In this example, we again consider the directional model with the same choice of distributions as in Section 5.2.1, but additionally, we add noise. As noise, we use the d-dimensional random vector

$$\varepsilon \sim \left|\mathcal{N}_d\Big(\mathbf{0}_d, \frac{100}{d}\boldsymbol{I}_d\Big)\right|,$$

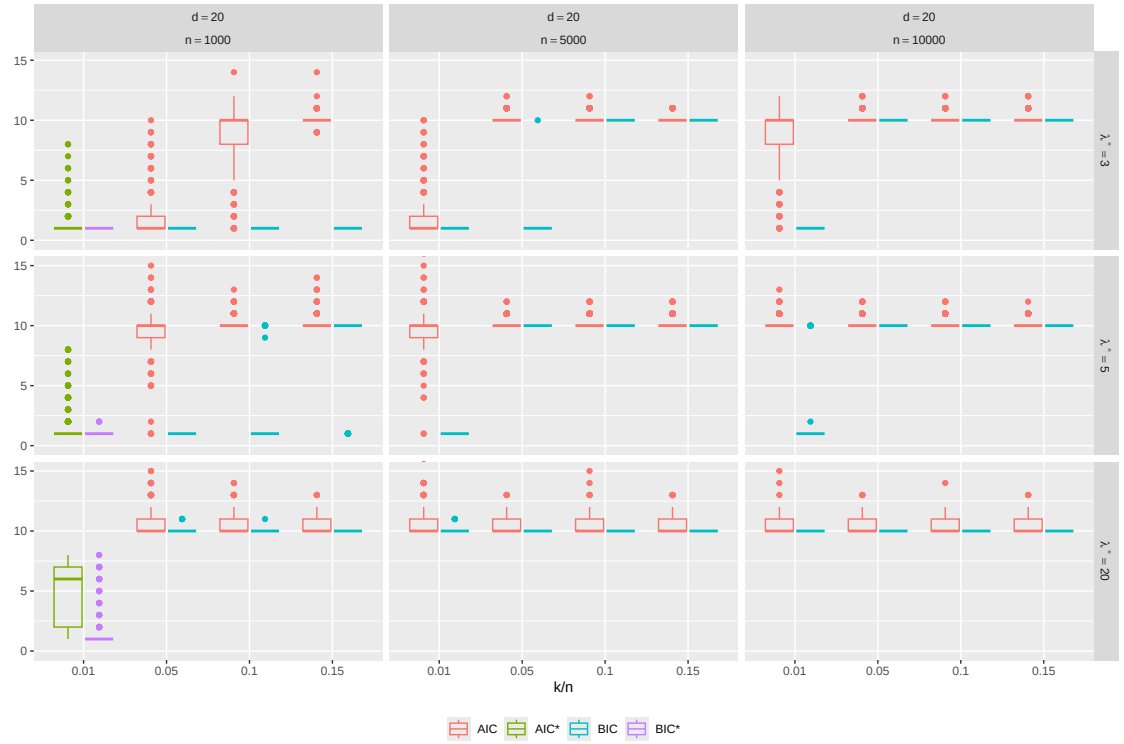


Figure 5.8.: *Simulations for the directional model with* $p^* = 10$ *and dimension* $d = 20$*: From left to the right the sample size increases from* $n = 1000$*,* $n = 5000$ *to* $n = 10000$*. From top to bottom, the value of the relevant eigenvalues increases from* $\lambda^* = 3$*,* $\lambda^* = 5$ *to* $\lambda^* = 20$*. In every subplot the ratio* k_n/n *increases from left to right from* 0.01*,* 0.05*,* 0.1 *to* 0.15*. The box plots show the estimator* $\widehat{p}_n$ *for* $p^* = 10$ *of the* AIC *and* BIC*.*

where the absolute value is entry-wise. Due to the scaling of the covariance matrix by $100/d$ the variance of the norm of $\boldsymbol{\varepsilon}$ converges as $d \to \infty$ to $100/\sqrt{2}$ (see Lemma A.3). Then, we construct the regularly varying random vector

$$\boldsymbol{X}^{(n)} = \frac{\boldsymbol{\Gamma}^{(n)^{1/2}} \boldsymbol{V}^{(n)}}{\|\boldsymbol{\Gamma}^{(n)^{1/2}} \boldsymbol{V}^{(n)}\|} \cdot Z + \boldsymbol{\varepsilon} \in \mathbb{R}^{d_n},$$

where $\boldsymbol{\Gamma}^{(n)}, \boldsymbol{V}^{(n)}$ and Z are defined as in Section 5.2.1 and $\boldsymbol{\varepsilon}$ is given as above.

The results are shown for $d = 20$ in Figure 5.10. In general, the results are similar to Figure 5.8, but with a greater deviation from the true value p^*. In most cases (e.g. $n = 5000, 10000$, $k_n/n \geq 0.05$ and $\lambda^* \geq 5$), the information criteria estimated $\widehat{p}_n = 11$ relevant eigenvalues, therefore identifying not only the 10 dominant eigenvalues but also including the noise. The noise leads to more fluctuation of the AIC estimates, especially to overestimation of p^*. For the BIC there are cases (e.g. $n = 1000$, $\lambda^* \leq 5$ and $k_n/n = 0.15$), where the BIC estimates $\widehat{p}_n = 1$ instead $p^* = 10$ and without noise the estimate is concentrated near $p^* = 10$. The AIC does not show this behavior. The influence of the noise decreases for larger λ^*, resulting in a larger spike.

Figure 5.11 provides a visualization of the results in the high-dimensional cases $d = 100$,

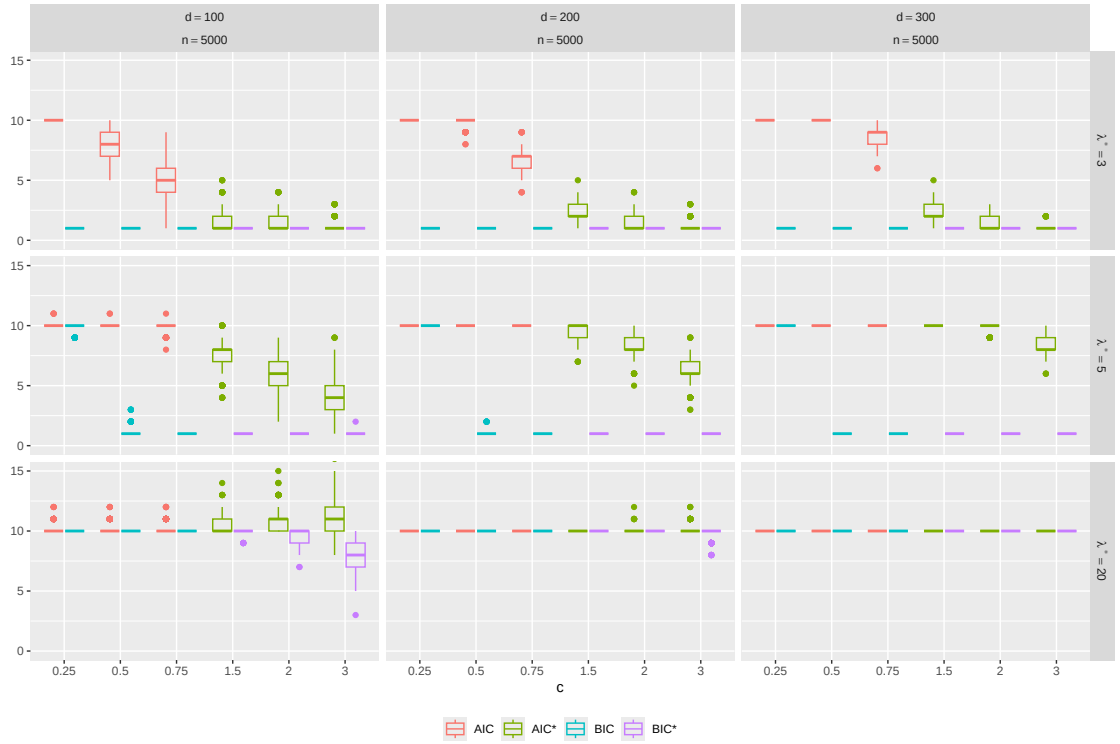


Figure 5.9.: *Simulations for the directional factor model with* $p^* = 10$ *and sample size* $n = 5000$*: From left to the right the dimension increases from* $d = 100$*,* $d = 200$ *to* $d = 300$*. From top to bottom, the value of the relevant eigenvalues increases from* $\lambda^* = 3$*,* $\lambda^* = 5$ *to* $\lambda^* = 20$*. In every subplot the ratio* $c = d/k_n$ *increases from left to right from* $c = 0.25$*,* $c = 0.5$*,* $c = 0.75$*,* $c = 1.5$*,* $c = 2$ *to* $c = 3$*. The box plot shows the estimator* $\widehat{p}_n$ *for* $p^* = 10$ *for the different information criteria.*

$d = 200$ and $d = 300$. We see that the effect of the noise is similar to the low-dimensional case. The overall fluctuation increases compared to the simulation without noise in Figure 5.9. The information criteria estimate the noise as an additional direction, for example, when $\lambda^* = 20$ and $d = 300$.

5.2.3. Spiked angular Gaussian model

In this section, we consider the contaminated spiked angular Gaussian model, which can also be found in Avella-Medina et al. (2025). For $1 \leq p^* \leq d$ we define the regularly varying random vector

$$\boldsymbol{X} = \boldsymbol{N}Z \in \mathbb{R}^d,$$

where Z is a univariate standard Fréchet distributed random variable, $\boldsymbol{N}$ follows a d-dimensional centered normal distribution with covariance matrix

$$\boldsymbol{H} := \sum_{i=1}^{p^*} \lambda_i \boldsymbol{v}_i \boldsymbol{v}_i^\top + \lambda \boldsymbol{I}_d,$$

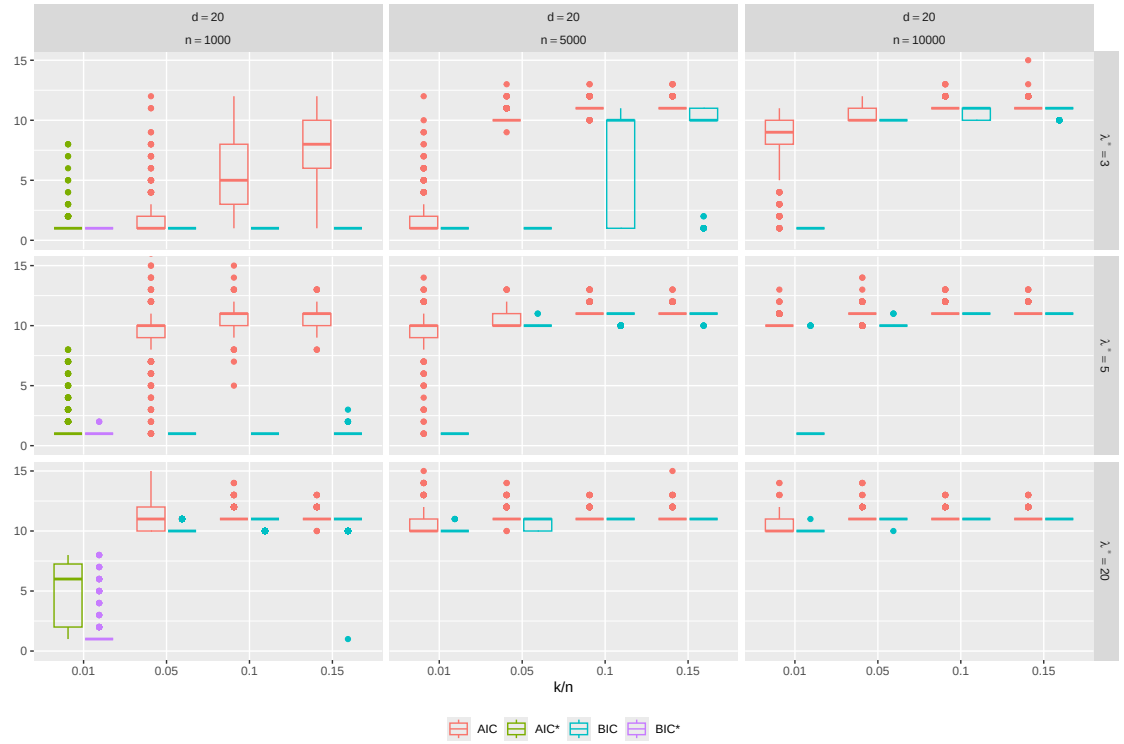


Figure 5.10.: *Simulations for noisy directional factor model with* $p^* = 10$ *and dimension* $d = 20$*: From left to the right the sample size increases from* $n = 1000$, $n = 5000$ *to* $n = 10000$. *From top to bottom, the value of the relevant eigenvalues increases from* $\lambda^* = 3$, $\lambda^* = 5$ *to* $\lambda^* = 20$. *In every subplot the ratio* k_n/n *increases from left to right from* 0.01, 0.05, 0.1 *to* 0.15. *The box plots show the estimator* $\widehat{p}_n$ *for* $p^* = 10$ *for the* AIC *and* BIC.

where $\boldsymbol{v}_i$, $i = 1, \ldots, p^*$ are orthogonal vectors and $\lambda_1 \geq \lambda_2 \geq \cdots \geq \lambda_{p^*} > \lambda = \cdots = \lambda > 0$. Note that the distribution of $\boldsymbol{X}$ differs from the directional model in Section 5.2.1, as the normal distribution is not standardized when $\boldsymbol{X}$ is generated. The spectral vector arising from $\boldsymbol{X}$ concentrates on a p-dimensional subspace and is given by (see Avella-Medina et al., 2025)

$$\mathbb{P}(\boldsymbol{\Theta} \in \cdot) = \frac{\mathbb{E}[\|\boldsymbol{N}\| \delta_{\boldsymbol{N}/\|\boldsymbol{N}\|}(\cdot)]}{\mathbb{E}[\|\boldsymbol{N}\|]}.$$

For the comparison, we run simulations with sample size $n = 5000$, dimension $d = 100$, $d = 300$ to $d = 900$. The matrix $\boldsymbol{H}$ is fixed for each sample but is initially randomly generated for the simulation, where the eigenvectors $\boldsymbol{v}_i$ are generated following the approach of Mezzadri (2007) with the R package pracma and the eigenvalues $\lambda_1 = \cdots = \lambda_{10} = 20$ are equal to 20, $p^* = 10$ and the last eigenvalue λ varies; we analyze the behavior of the information criteria when λ approaches 0 and thus the spiked covariance assumption is closer to being violated. Therefore, we compare the results for $\lambda = 0.01$, $\lambda = 0.1$, and $\lambda = 1$.

The results are illustrated in Figure 5.12. It is evident that, when the gap is sufficiently

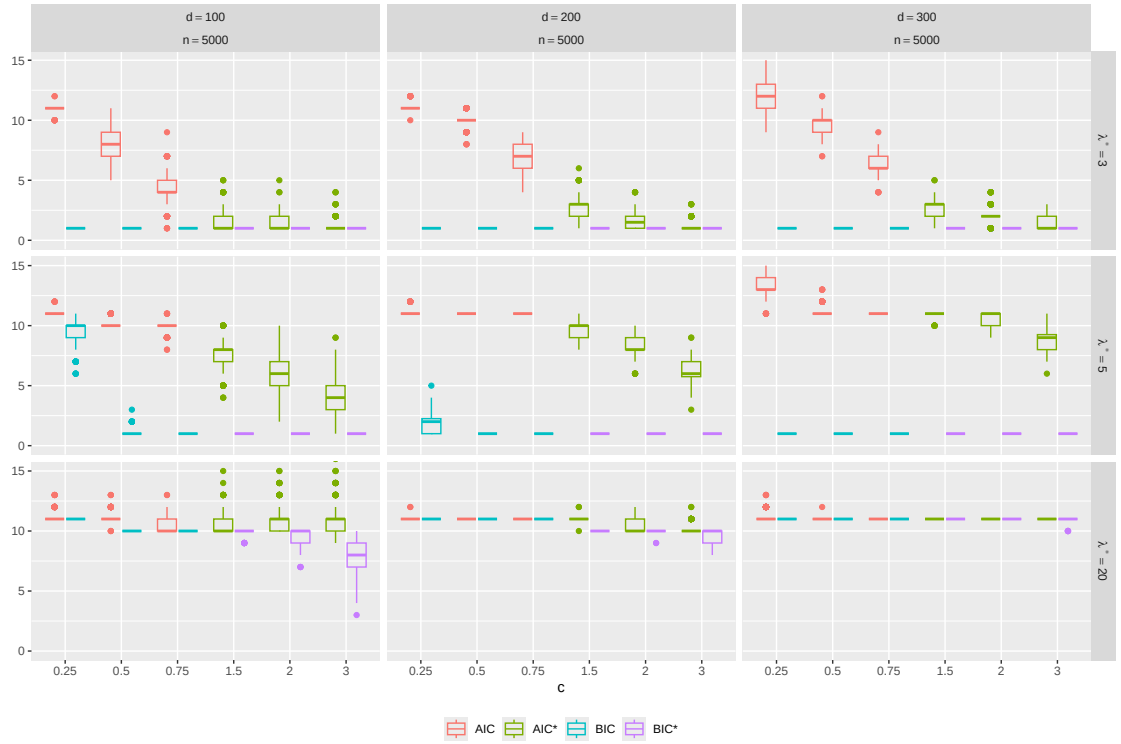


Figure 5.11.: *Simulations for the noisy directional factor model with* $p^* = 10$ *and sample size* $n = 5000$*: From left to the right the dimension increases from* $d = 100$*,* $d = 200$ *to* $d = 300$*. From top to bottom, the value of the relevant eigenvalues increases from* $\lambda^* = 3$*,* $\lambda^* = 5$ *to* $\lambda^* = 20$*. In every subplot the ratio* $c = d/k_n$ *increases from left to right from* $c = 0.25$*,* $c = 0.5$*,* $c = 0.75$*,* $c = 1.5$*,* $c = 2$ *to* $c = 3$*. The box plot shows the estimator* $\widehat{p}_n$ *for* $p^* = 10$ *for the different information criteria.*

large, then the BIC and BIC* are less affected by a small eigenvalue λ than the AIC and AIC*. The smaller λ is chosen, the larger the overestimation of the AIC and AIC* is, whereby for $d = 900$ and $\lambda = 0.01, 0.1$ the AIC* overestimates p^* more than AIC. When $\lambda = 1$ and $d \geq 300$ the performance of all criteria is nearly identical.

5.3. Application to real-world data

In this section, we evaluate the performance of the proposed information criteria on two real-world datasets. The first, concerning wind speed measurements (see Section 5.3.1), is analyzed using the methodology developed in Chapter 3 and corresponds to a fixed-dimensional setting. The second dataset, comprising precipitation records from Germany (see Section 5.3.2), is examined via the techniques introduced in Chapter 4, and can be considered to be in a high-dimensional regime in which the dimension exceeds the sample size.

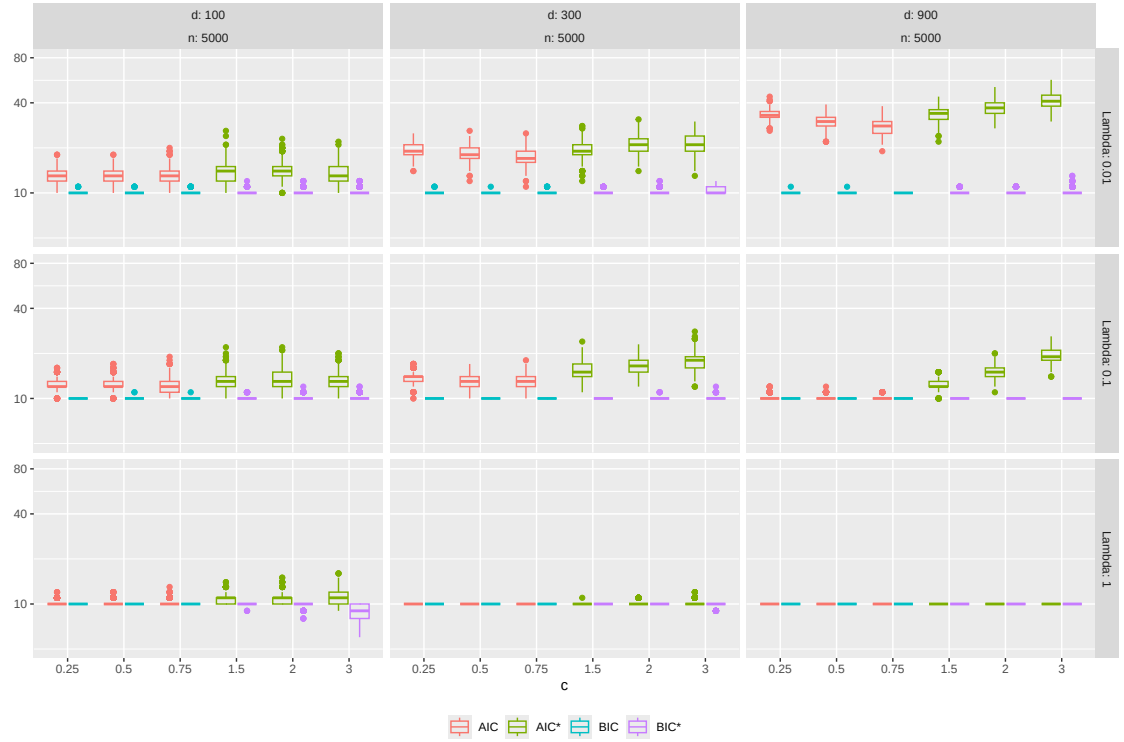


Figure 5.12.: *Simulations for spiked angular Gaussian data with* $p^* = 10$*: From left to the right the dimension increases from* $d = 100$*,* $d = 300$ *to* $d = 900$*. From top to bottom, the value of* λ *increases from* $\lambda = 0.01$*,* $\lambda = 0.1$ *to* $\lambda = 1$*. In every subplot the ratio* $c = d/k_n$ *increases from left to right from* $c = 0.25$*,* $c = 0.5$*,* $c = 0.75$*,* $c = 1.5$*,* $c = 2$ *to* $c = 3$*. The box plot is log-scaled and shows the estimator* $\widehat{p}_n$ *for the different information criteria.*

5.3.1. APPLICATION TO WIND SPEED DATA

In the following, we examine the dependence structure of extreme wind speeds using the same example as Meyer and Wintenberger (2023) by applying the method from Chapter 3. For this purpose, the daily average wind speeds 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 (1989) and taken from StatLib - Datasets Archive (2023). 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 5.14 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 preprocessing, we use the same Hill estimator $\widehat{\alpha} = 10.7$ (see Section 2.1.3) as Meyer and Wintenberger (2023). 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 5.3.

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 the number of

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

Table 5.3.: *Estimators for the wind speed dataset based on the different information criteria.*

extreme directions between 9 and 13, which is not a large range compared to the choice of k_n. On the left-hand side of Figure 5.13, 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 also plotted 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.

Figure 5.13.: *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 5.14, where the black dots highlight the different locations of the stations. Colored diamonds close to a station are markers

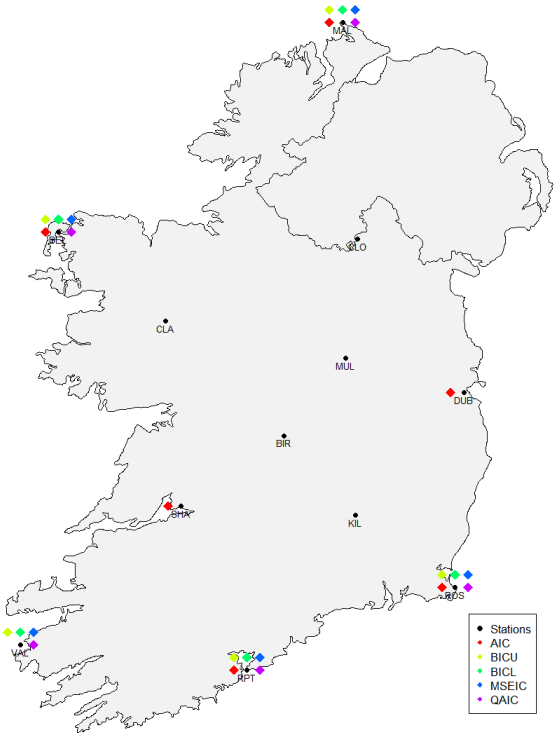


Figure 5.14.: *Maximal subsets recovered by the information criteria of the daily average wind speed.*

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 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 is nearly 40 km 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.

5.3.2. Application to precipitation data

In the following, the information criteria developed in Chapter 4 are applied to precipitation data in Germany taken from DWD Climate Data Center (CDC) (1951 - 2022). The dataset consists of daily station observations of the precipitation height for Germany between January 1, 1951 and March 31, 2022 at $d = 500$ stations. The stations are marked by black dots in Figure 5.15. The data is preprocessed to include only observations from January, February and March, and transformed to standard Fréchet margins (see Section 2.1.3). After data cleaning, the resultant dataset contains $n = 2546$ observations, each with precipitation records from at least one station. In Figure 5.15 we see the stations of the empirical eigenvectors $\widehat{\boldsymbol{v}}_i = (v_1^{(i)}, \ldots, v_d^{(i)})^\top$, where $v_j^{(i)} \geq 0.6 v_{(1)}^{(i)}$, $i = 1, \ldots, 5$, of the 5

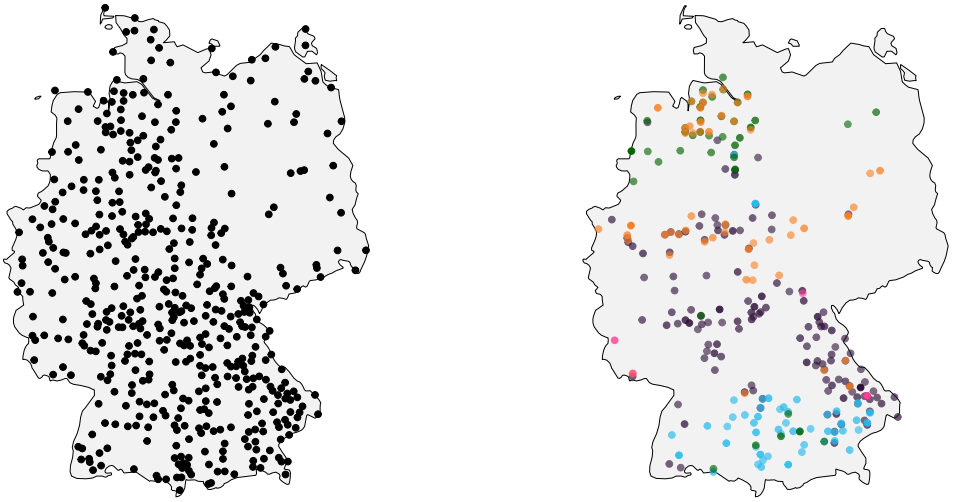


Figure 5.15.: *Left figure: Map of Germany with all stations highlighted by black dots. Right figure: Map of Germany with the most extreme stations of the empirical eigenvectors of the five largest empirical eigenvalues, colored by eigenvectors.*

largest empirical eigenvalues $\widehat{\lambda}_{n,1}, \ldots, \widehat{\lambda}_{n,5}$ if $k_n = 76$; the stations of each eigenvector are colored differently.

We consider 1% to 15% of the data as extreme, corresponding to 25 to 382 observations. In these cases $d > k_n$ and, therefore, we assume to be in the high-dimensional setting with $c > 1$ and apply AIC* and BIC* from Definition 4.28. The number of candidate models q_n for the AIC* is chosen as $d/2 = 250$ to account for the assumption of Theorem 4.29.

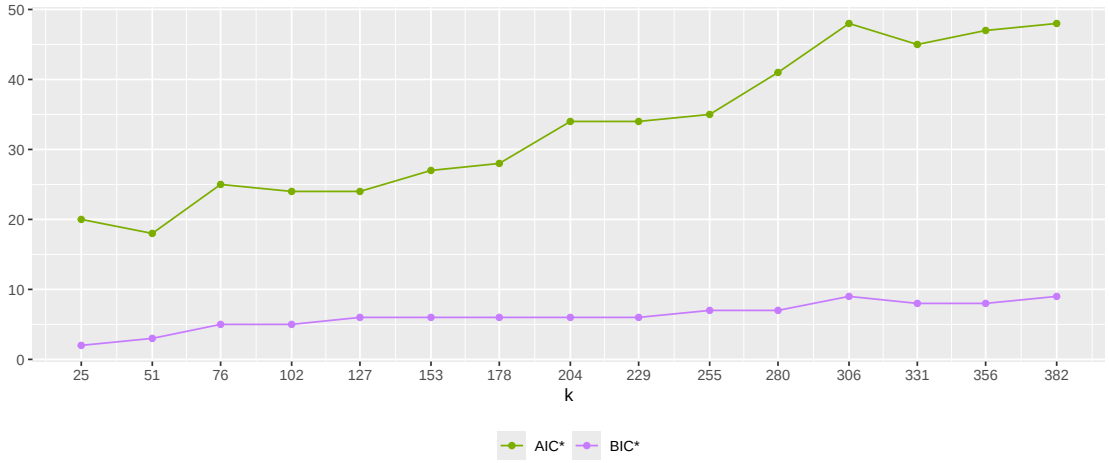


Figure 5.16.: *The estimated number $\widehat{p}_n$ of significant eigenvalues determined by* AIC* *and* BIC* *plotted against k_n.*

Figure 5.16 shows the number of estimated significant eigenvalues $\widehat{p}_n$ mapped against k_n. The estimates using AIC* stabilize between $k_n = 76$ and $k_n = 178$, ranging from 24 to 28, before further increasing. In contrast, BIC* stabilizes for k_n between 76 and 229, with values of 5 and 6. Even for $k_n \geq 255$, the BIC* remains between 7 and 9, whereas

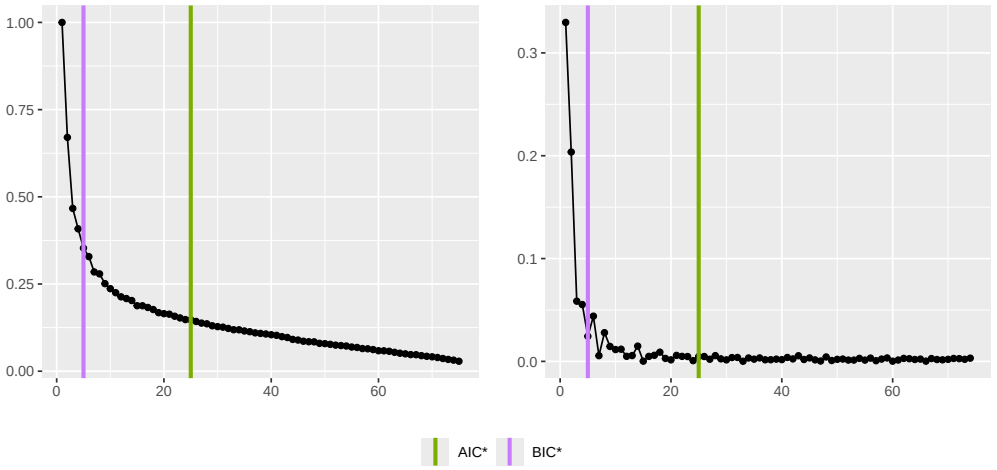


Figure 5.17.: *For $k_n = 76$, on the left hand side the scaled ordered empirical eigenvalues $\widehat{\lambda}_{n,i}/\widehat{\lambda}_{n,1}$, $i = 1, \ldots, 75$ and on the right hand side the differences of the ordered empirical eigenvalues divided by the value of the largest eigenvalue $(\widehat{\lambda}_{n,i} - \widehat{\lambda}_{n,i+1})/\widehat{\lambda}_{n,1}$, $i = 1, \ldots, 75$ are plotted. The vertical line indicates the* AIC* *estimator $\widehat{p}_n = 25$ and the* BIC* *estimator $\widehat{p}_n = 5$.*

AIC* continues to increase. This difference between the estimates is aligned with the heavier penalty imposed by BIC*, which leads to smaller estimates compared to AIC*. These estimates reduce the dimensionality of $d = 500$ by factors of 20 and 100, respectively. For comparison of these different estimates, the scaled empirical eigenvalues $\widehat{\lambda}_{n,i}/\widehat{\lambda}_{n,1}$, $i = 1, \ldots, 75$, are plotted in the left picture of Figure 5.17. At first view, they seem not to be constant after some point, contradicting the spiked covariance assumption. But in a spiked covariance model with

$$\lambda_1 > \lambda_2 > \cdots > \lambda_{p^*} > \lambda_{p^*+1} = \cdots = \lambda_{d-1} =: \lambda > 0.$$

we have

$$\frac{\lambda_i - \lambda_{i+1}}{\lambda_1} > 0 \quad \text{for } i = 1, \ldots, p^* \qquad \text{and} \qquad \frac{\lambda_i - \lambda_{i+1}}{\lambda_1} = 0 \quad \text{for } i = p^* + 1, \ldots, d.$$

Therefore, the scaled increments $(\widehat{\lambda}_{n,i} - \widehat{\lambda}_{n,i+1})/\widehat{\lambda}_{n,1}$ $i = 1, \ldots, 75$, of the empirical eigenvalues are plotted in the right picture of Figure 5.17. We realize that after some point these increments are nearly constant zero, indicating that these are non-spiked eigenvalues. The information criteria seem to estimate the point where these increments are constant zero because in the interval $[5, 24]$, which is spanned by our estimators, this happens. In the non-extreme value setting, Hung et al. (2022) analyzed a dataset on the habitual diet of the human gut microbiome, in which the empirical eigenvalues and the estimators of the information criteria, displayed in Hung et al. (2022, Fig. 10 (c)), have a similar behavior to our empirical eigenvalues in our Figure 5.17.

CHAPTER 6

Conclusion and outlook

In this thesis, we developed and analyzed information criteria to the reduce the dimension of extreme data, thereby answering question **(Q)**. The information criteria, derived from two different approaches, achieve a dimension reduction in high-dimensional settings where the dimension can grow with the sample size.

In Chapter 3, we developed three different information criteria for the number of extreme directions s^* based on the concept of sparse regular variation. In the fixed-dimensional setting, we derived information criteria for the choice of the optimal threshold k_n. We first considered the QAIC, which follows the same ideas as an Akaike information criterion but is based on a Gaussian likelihood function in contrast to the multinomial model of the AIC of Meyer and Wintenberger (2023). We showed that the AIC is not consistent, but the QAIC is consistent in the fixed-dimensional setting. Next, we introduced the MSEIC, where the idea of minimizing the mean squared error as a risk function is analogous to the AIC, which minimizes the Kullback–Leibler divergence, but without requiring a likelihood assumption. However, similar to the AIC, it is not consistent. The last information criterion in this chapter is the BIC, which is based on a Bayesian approach for a multinomial model analogous to the AIC of Meyer and Wintenberger (2023). The advantage of BICU and BICL in the fixed-dimensional case is that they are weakly consistent information criteria for the number of extreme directions s^*. Then, we moved to the high-dimensional setting and adapted all four information criteria for this setting. Under suitable assumptions, we showed that all information criteria are consistent, provided that for the AIC, QAIC and MSEIC the fluctuation of the number of observations across bias directions is not too large. In the simulation study of the local model for large n and fixed dimension, we observed that AIC and MSEIC tend to overestimate s^* for large sample sizes, but for small n the MSEIC performs extraordinarily well. This changes in the high-dimensional setting, where the information criteria do not have the tendency to overestimate and the BIC even underestimates s^* if the dimension is large. Overall, all information criteria performed quite well, none was particularly superior in all situations.

In Chapter 4, we proposed information criteria based on the AIC and BIC for Gaussian random vectors to detect the number p^* of significant principal components in multivariate extremes, which corresponds to the location of the spike in the eigenvalues of the covariance matrix of the spectral measure. We used a spiked covariance model for the eigenvalues of $\mathbf{\Sigma}$, where ξ_{p^*} is the smallest spiked eigenvalue. Our analysis encompassed both the classical large-sample setting and the high-dimensional setting, which has become increasingly

relevant for extreme value theory in today's applications. We established the consistency of the BIC in the large-sample setting, where we derived some new results on the asymptotic properties of the empirical eigenvalues of $\mathbf{\Sigma}$ under the spiked covariance model. In the high-dimensional setting, we employed methods from random matrix theory to validate the consistency of the AIC and the BIC when working with the directional model. In particular, we derived the behavior of the empirical eigenvalues by connecting them to the empirical eigenvalues for the underlying latent vector. In the simulation study, we also observed that in the fixed-dimensional case the AIC has the tendency to overestimate the true dimension if the sample size n is large. In contrast, in the high-dimensional case, the BIC underestimates the true dimension when the sample size n is low.

Furthermore, the performance of the information criteria was validated through a simulation study. Since we used two different methods to reduce the dimension, it is not possible to directly compare the information criteria of the two methods. However, in general, the information criteria for both methods worked and generally showed similar behavior and consistency results. For example, under some assumptions, the BIC is consistent in both settings and the AIC is only consistent in the high-dimensional case. Similarly, the simulation studies for both methods showed that the AIC tends to overestimate the true dimension if the sample size is large and the dimension is small, compared to the BIC, which underestimates the true dimension if the sample size is small and the dimension is high. These results are also in accordance with the literature, as mentioned before.

In view of question **(Q)**, we derived information criteria to consistently reduce the dimension of data following the SRV or PCA approach, even in high-dimensional settings where the dimension tends to infinity. Along the way, we derived new theory for the fixed- and high-dimensional settings. In addition, the information criteria were successfully applied to real-world datasets. Therefore, we have developed methods that allow practitioners to select from a variety of information criteria depending on the suspected dependence structure in order to address the problem of dimension reduction for extreme data.

Although we have thoroughly addressed all aspects of question **(Q)** using multiple methods, new research questions have emerged. Next, we present open questions, which could be part of further research.

- In the high-dimensional setting for the spiked covariance model, we assume that ξ_{p^*} is a distant spiked eigenvalue in the sense that $\xi_{p^*} > 1 + \sqrt{c}$ and $c = \lim_{n\to\infty} d_n/k_n > 0$. For applications, these eigenvalue assumptions are restrictive, as we see in our data example in Figure 5.17, where the empirical eigenvalues decrease but do not stabilize at some point. Therefore, it would be worthwhile to explore more general eigenvalue structures of the covariance matrix of $\mathbf{\Sigma}$ such as $\xi_{p^*} > 1 + \sqrt{c}$ and $\xi_{p^*+1} < 1 + \sqrt{c}$ where all eigenvalues ξ_j for $j = p^* + 1, \ldots, d_n$ are in a neighborhood of 1 or 0.

- In Chapter 4, the case $c = 0$ is not covered and could be further investigated. We suspect that the AIC and the BIC, as defined here, are inconsistent even when a

type of *gap condition* like (4.6) is satisfied. Such a condition relates to the distance between the smallest eigenvalue bigger than 1, denoted by ξ_{n,p^*}, and the following eigenvalue, denoted by ξ_{n,p^*+1}.

- Additionally, as a starting point for this line of research on PCA for high-dimensional extremes, the consistency results of the information criteria were based on the assumption that the underlying model is a directional model, similar to multivariate statistics, where the first results were derived for Gaussian models with a special covariance structure. Of course, it would also be interesting to explore generalizations or alternatives to the directional model.

- The optimal choice of k_n is nontrivial, as it has a direct influence on the estimates of the information criteria. For the choice of k_n further research is needed in the high-dimensional setting of Section 3.3 and for the PCA approach in Chapter 4, as discussed for the SRV approach (Chapter 3 and Meyer and Wintenberger, 2023).

- There are other methods to reduce the dimension of data, which are established in the non-extreme field of statistics. For example, Autoencoders (cf. Goodfellow et al., 2016, Chapter 14), which are neural networks that encode data to a lower-dimensional space and then decode it to get back to the original space. One could also estimate the true dimension of the encoded space. Another method is to apply self-organizing maps (Kohonen, 2001), which can be used to derive low-dimensional representations.

- In this thesis, we considered the setting in which both the effective sample size and the dimension grow proportionally at most. Another possibility is the high-dimensional, low-sample-size case, where the dimension grows while the sample size remains fixed. For example, Jung and Marron (2009) analyzed the consistency of PCA in this framework. In a similar way, PCA could be studied for extremes in situations where the dimension increases much faster than the sample size.

In conclusion, we proposed several solutions to question **(Q)** and many opportunities for further research have arisen.

Bibliography

Akaike, H. (1974). A new look at the statistical model identification, *IEEE Trans. Automatic Control* **AC-19**: 716–723.

Anderson, T. W. (1963). Asymptotic theory for principal component analysis, *Ann. Math. Statist.* **34**: 122–148.

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

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**(2): 777–801.

Bai, Z., Choi, K. P. and Fujikoshi, Y. (2018). Consistency of AIC and BIC in estimating the number of significant components in high-dimensional principal component analysis, *Ann. Statist.* **46**(3): 1050–1076.

Bai, Z. D. and Yin, Y. Q. (1993). Limit of the smallest eigenvalue of a large-dimensional sample covariance matrix, *Ann. Probab.* **21**(3): 1275–1294.

Bai, Z., Fujikoshi, Y. and Hu, J. (2020). Strong consistency of the AIC, BIC, C_p and KOO methods in high-dimensional multivariate linear regression, *arXiv: 1810.12609* .

Bai, Z. and Silverstein, J. W. (2010). *Spectral analysis of large dimensional random matrices*, Springer.

Bai, Z. and Yao, J. (2012). On sample eigenvalues in a generalized spiked population model, *J. Multivariate Anal.* **106**: 167–177.

Baik, J. and Silverstein, J. W. (2006). Eigenvalues of large sample covariance matrices of spiked population models, *J. Multivariate Anal.* **97**(6): 1382–1408.

Bellman, R. (1957). *Dynamic programming*, Princeton University Press, Princeton.

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

Bingham, N. H., Goldie, C. M. and Teugels, J. L. (1989). *Regular variation*, Vol. 27 of *Encyclopedia of Mathematics and its Applications*, Cambridge University Press.

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

Butsch, L. and Fasen-Hartmann, V. (2025a). Estimation of the number of principal components in high-dimensional multivariate extremes, *Scand. J. Stat.* **52**(4): 2270–2313.

Butsch, L. and Fasen-Hartmann, V. (2025b). Information criteria for the number of directions of extremes in high-dimensional data, *Electron. J. Statist.* **19**(2): 5695 – 5740.

Butsch, L. and Fasen-Hartmann, V. (2026). Statistical inference for extremal directions in high-dimensional spaces, *arXiv: 2603.26618* .

Cavanaugh, J. E. and Neath, A. A. (1999). Generalizing the derivation of the Schwarz information criterion, *Comm. Statist. Theory Methods* **28**(1): 49–66.

Chautru, E. (2015). Dimension reduction in multivariate extreme value analysis, *Electron. J. Stat.* **9**(1): 383–418.

Chen, L., Oesting, M. and Zhou, C. (2025). Clustering tails in high dimension, *arXiv: 2506.19414* .

Chiapino, M., Sabourin, A. and Segers, J. (2019). Identifying groups of variables with the potential of being large simultaneously, *Extremes* **22**(2): 193–222.

Claeskens, G. (2016). Statistical model choice, *Annu. Rev. Stat. Appl.* **3**(1): 233–256.

Clémençon, S., Huet, N. and Sabourin, A. (2024). Regular variation in Hilbert spaces and principal component analysis for functional extremes, *Stochastic Process. Appl.* **174**: 104375.

Clémençon, S., Jalalzai, H., Lhaut, S., Sabourin, A. and Segers, J. (2023). Concentration bounds for the empirical angular measure with statistical learning applications, *Bernoulli* **29**(4): 2797–2827.

Clémençon, S. and Sabourin, A. (2025). Weak signals and heavy tails: Machine-learning meets extreme value theory, *arXiv: 2504.06984* .

Coles, S. (2001). *An introduction to statistical modeling of extreme values*, Springer-Verlag London, Ltd.

Cooley, D. and Thibaud, E. (2019). Decompositions of dependence for high-dimensional extremes, *Biometrika* **106**(3): 587–604.

Cover, T. M. (2006). *Elements of information theory*, A Wiley-Interscience publication, Wiley-Interscience, Hoboken, NJ.

Das, B. and Fasen-Hartmann, V. (2025). Asymptotic independence in higher dimensions and its implications on risk management, *Can J Statistics* **e70036**.

Dauxois, J., Pousse, A. and Romain, Y. (1982). Asymptotic theory for the principal component analysis of a vector random function: some applications to statistical inference, *J. Multivariate Anal.* **12**(1): 136–154.

Davis, A. W. (1977). Asymptotic theory for principal component analysis: non-normal case, *Austral. J. Statist.* **19**(3): 206–212.

de Haan, L. and Ferreira, A. (2006). *Extreme Value Theory: An Introduction*, Springer, New York.

Drees, H. (2025). Asymptotic behavior of principal component projections for multivariate extremes, *arXiv: 2503.22296* .

Drees, H. and Sabourin, A. (2021). Principal component analysis for multivariate extremes, *Electron. J. Stat.* **15**(1): 908–943.

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* pp. 272–279.

DWD Climate Data Center (CDC) (1951 - 2022). Daily station observations precipitation height in mm for Germany, version v21.3, last accessed: May 03, 2023.

Einmahl, J. H. J., de Haan, L. and Piterbarg, V. I. (2001). Nonparametric estimation of the spectral measure of an extreme value distribution, *Ann. Statist.* **29**(5): 1401–1423.

Einmahl, J. H. J. and Segers, J. (2009). Maximum empirical likelihood estimation of the spectral measure of an extreme-value distribution, *Ann. Statist.* **37**(5B): 2953–2989.

Elezović, N., Giordano, C. and Pečarić, J. (2000). The best bounds in Gautschi's inequality, *Math. Inequal. Appl.* **3**(2): 239–252.

Embrechts, P., Klüppelberg, C. and Mikosch, T. (1997). *Modelling extremal events*, Springer-Verlag.

Engelke, S., Hentschel, M., Lalancette, M. and Röttger, F. (2024). Graphical models for multivariate extremes, *arXiv: 2402.02187* .

Engelke, S. and Hitz, A. S. (2020). Graphical models for extremes, *J. R. Stat. Soc. Ser. B. Stat. Methodol.* **82**(4): 871–932.

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

Engelke, S. and Volgushev, S. (2022). Structure learning for extremal tree models, *J. R. Stat. Soc. Ser. B. Stat. Methodol.* **84**(5): 2055–2087.

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

Fomichov, V. and Ivanovs, J. (2023). Spherical clustering in detection of groups of concomitant extremes, *Biometrika* **110**(1): 135–153.

Fujikoshi, Y. and Sakurai, T. (2016). Some properties of estimation criteria for dimensionality in principal component analysis, *AJMMS* **35**(2): 133–142.

Fujikoshi, Y., Ulyanov, V. V. and Shimizu, R. (2010). *Multivariate statistics : high-dimensional and large-sample approximations*, Wiley series in probability and statistics, Hoboken, N.J, Wiley.

Gissibl, N. and Klüppelberg, C. (2018). Max-linear models on directed acyclic graphs, *Bernoulli* **24**(4A): 2693–2720.

Gissibl, N., Klüppelberg, C. and Lauritzen, S. (2021). Identifiability and estimation of recursive max-linear models, *Scand. J. Stat.* **48**(1): 188–211.

Goix, N., Sabourin, A. and Clémençon, S. (2017). Sparse representation of multivariate extremes with applications to anomaly detection, *J. Multivariate Anal.* **161**: 12–31.

Goodfellow, I., Bengio, Y. and Courville, A. (2016). *Deep learning*, MIT Press.

Hannan, E. J. and Quinn, B. G. (1979). The determination of the order of an autoregression, *J. Roy. Statist. Soc. Ser. B* **41**(2): 190–195.

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): 1–50.

Hastie, T., Tibshirani, R. and Wainwright, M. (2015). *Statistical learning with sparsity*, CRC Press, Boca Raton.

Hilal, S., Poon, S.-H. and Tawn, J. (2014). Portfolio risk assessment using multivariate extreme value methods, *Extremes* **17**(4): 531–556.

Hill, B. M. (1975). A simple general approach to inference about the tail of a distribution, *Ann. Statist.* **3**(5): 1163–1174.

Hille, E. (1964). *Analysis*, Vol. 1:, Blaisdell Publ. Co, Waltham, Mass.

Hogg, R. V., McKean, J. W. and Craig, A. T. (2005). *Introduction to mathematical statistics*, 6. edn, Pearson Prentice Hall.

Horn, R. A. and Johnson, C. R. (2013). *Matrix analysis*, 2. edn, Cambridge Univ. Press.

Hung, H., Huang, S.-Y. and Ing, C.-K. (2022). A generalized information criterion for high-dimensional PCA rank selection, *Statist. Papers* **63**(4): 1295–1321.

Jalalzai, H. and Leluc, R. (2021). Feature clustering for support identification in extreme regions, *Proceedings of the 38th International Conference on Machine Learning*, Vol. 139 of *Proc. Mach. Learn. Res.*, PMLR, pp. 4733–4743.

Janßen, A. and Wan, P. (2020). k-means clustering of extremes, *Electron. J. Stat.* **14**(1): 1211–1233.

Jiang, Q., Qiu, J. and Li, Z. (2023). On eigenvalues of sample covariance matrices based on high dimensional compositional data, *arXiv: 2312.14420* .

Jiang, Y., Cooley, D. and Wehner, M. F. (2020). Principal component analysis for extremes and application to U.S. precipitation, *J. Clim.* **33**(15): pp. 6441–6451.

Johnstone, I. M. (2001). On the distribution of the largest eigenvalue in principal components analysis, *Ann. Statist.* **29**(2): 295–327.

Johnstone, I. M. and Yang, J. (2018). Notes on asymptotics of sample eigenstructure for spiked covariance models with non-gaussian data, *arXiv: 1810.10427* .

Jung, S. and Marron, J. S. (2009). PCA consistency in high dimension, low sample size context, *Ann. Statist.* **37**(6B): 4104–4130.

Kiriliouk, A., Rootzén, H., Segers, J. and Wadsworth, J. L. (2019). Peaks over thresholds modeling with multivariate generalized Pareto distributions, *Technometrics* **61**(1): 123–135.

Kohonen, T. (2001). *Self-organizing maps*, Vol. 30, third edn, Springer-Verlag.

Kulik, R. and Soulier, P. (2020). *Heavy-tailed time series*, Springer.

Larsson, M. and Resnick, S. I. (2012). Extremal dependence measure and extremogram: the regularly varying case, *Extremes* **15**(2): 231–256.

Lederer, J. and Oesting, M. (2024). Extremes in high dimensions: Methods and scalable algorithms, *arXiv: 2303.04258* .

Mason, D. M. (1982). Laws of large numbers for sums of extreme values, *Ann. Probab.* **10**(3): 754–764.

Meyer, N. and Wintenberger, O. (2020). Tail inference for high-dimensional data, *arXiv: 2007.11848v1* .

Meyer, N. and Wintenberger, O. (2021). Sparse regular variation, *Adv. in Appl. Probab.* **53**(4): 1115–1148.

Meyer, N. and Wintenberger, O. (2023). Multivariate sparse clustering for extremes, *J. Amer. Statist. Assoc.* **0**(0): 1–12.

Mezzadri, F. (2007). How to generate random matrices from the classical compact groups, *Notices Amer. Math. Soc.* **54**(5): 592–604.

Mikosch, T. and Wintenberger, O. (2024). *Extreme value theory for time series—models with power-law tails*, Springer.

Muirhead, R. J. (1982). *Aspects of multivariate statistical theory*, John Wiley & Sons, Inc.

Ouimet, F. (2021). A precise local limit theorem for the multinomial distribution and some applications, *J. Statist. Plann. Inference* **215**: 218–233.

Paul, D. (2007). Asymptotics of sample eigenstructure for a large dimensional spiked covariance model, *Statist. Sinica* **17**(4): 1617–1642.

Posit team (2025). *RStudio: Integrated Development Environment for R*, Posit Software, PBC, Boston, MA.

Powers, D. (2008). Evaluation: From precision, recall and f-factor to roc, informedness, markedness and correlation, *Mach. Learn. Technol.* **2**.

R Core Team (2025). *R: A Language and Environment for Statistical Computing*, R Foundation for Statistical Computing, Vienna, Austria.

Resnick, S. (1987). *Extreme Values, Regular Variation, and Point Processes*, Springer.

Resnick, S. (2007). *Heavy-Tail Phenomena: Probabilistic and Statistical Modeling*, Springer.

Rohrbeck, C. and Cooley, D. (2023). Simulating flood event sets using extremal principal components, *Ann. Appl. Stat.* **17**(2): 1333–1352.

Russell, B. T., Cooley, D. S., Porter, W. C., Reich, B. J. and Heald, C. L. (2016). Data mining to investigate the meteorological drivers for extreme ground level ozone events, *Ann. Appl. Stat.* **10**(3): 1673–1698.

Schwarz, G. (1978). Estimating the dimension of a model, *Ann. Statist.* **6**(2): 461–464.

Seber, G. A. F. (1984). *Multivariate observations*, John Wiley & Sons, Inc.

Silverstein, J. W. (1995). Strong convergence of the empirical distribution of eigenvalues of large-dimensional random matrices, *J. Multivariate Anal.* **55**(2): 331–339.

Silverstein, J. W. and Choi, S.-I. (1995). Analysis of the limiting spectral distribution of large-dimensional random matrices, *J. Multivariate Anal.* **54**(2): 295–309.

Simpson, E. S., Wadsworth, J. L. and Tawn, J. A. (2020). Determining the dependence structure of multivariate extremes, *Biometrika* **107**(3): 513–532.

Spiegelhalter, D. J., Best, N. G., Carlin, B. P. and van der Linde, A. (2002). Bayesian measures of model complexity and fit, *J. R. Stat. Soc. Ser. B Stat. Methodol.* **64**(4): 583–639.

StatLib - Datasets Archive (2023). Wind, `http://lib.stat.cmu.edu/datasets/wind.data`, accessed October 26, 2023.

Tawn, J. A. (1992). Estimating probabilities of extreme sea-levels, *J. R. Stat. Soc. Ser. C. Appl. Stat.* **41**(1): 77–93.

Tibshirani, R. (1996). Regression shrinkage and selection via the lasso, *J. Roy. Statist. Soc. Ser. B* **58**(1): 267–288.

Tyler, D. E. (1987). Statistical analysis for the angular central Gaussian distribution on the sphere, *Biometrika* **74**(3): 579–589.

Uchida, Y. (2008). A simple proof of the geometric-arithmetic mean inequality, *J. Inequal. Pure Appl. Math.* **9**(2).

Wan, P. (2026). Characterizing extremal dependence on a hyperplane, *Biometrika* .

Watanabe, S. (2010). Asymptotic equivalence of Bayes cross validation and widely applicable information criterion in singular learning theory, *J. Mach. Learn. Res.* **11**: 3571–3594.

Yin, Y. Q., Bai, Z. D. and Krishnaiah, P. R. (1988). On the limit of the largest eigenvalue of the large-dimensional sample covariance matrix, *Probab. Theory Related Fields* **78**(4): 509–521.

APPENDIX A

Auxiliary results

A.1. Auxiliary results for the quasi-Akaike information criterion

In this section, we present supplementary results for Section 3.2.1. Note that most parts of this chapter consist of Butsch and Fasen-Hartmann (2025b).

A.1.1. Proof of Lemma 3.43

Lemma 3.43. Suppose the assumptions of Proposition 3.11 hold and $\widehat{\underline{\boldsymbol{p}}}_n^s(\boldsymbol{\mathcal{T}}_n)$ is defined analog to $\widehat{\underline{\boldsymbol{p}}}_n^s(\widetilde{\boldsymbol{\mathcal{T}}}_n)$ in (3.9). Then as $n \to \infty$,

$$\begin{aligned}\boldsymbol{Y}_n :=& \sqrt{k_n}\,\mathrm{diag}(p_{n,1},\ldots,p_{n,s},\frac{\rho_n}{(r-s)},\rho_n,\ldots,\rho_n)^{-1/2}\begin{pmatrix}\left(\widehat{\underline{\boldsymbol{p}}}_n^s(\widetilde{\boldsymbol{\mathcal{T}}}_n)-\widehat{\underline{\boldsymbol{p}}}_n^s(\boldsymbol{\mathcal{T}}_n)\right)\\ \left(\frac{\mathcal{T}_{n,s+1}}{k_n}-\widehat{\underline{\rho}}_n^s(\boldsymbol{\mathcal{T}}_n)\right)\\ \vdots\\ \left(\frac{\mathcal{T}_{n,r}}{k_n}-\widehat{\underline{\rho}}_n^s(\boldsymbol{\mathcal{T}}_n)\right)\end{pmatrix}\\ &\xrightarrow{\mathcal{D}} \mathcal{N}_{r+1}\left(\boldsymbol{0}_{r+1},\Sigma\right),\end{aligned}$$

where

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

Proof. From Assumption (A4) and the continuous mapping theorem we receive that

$$\begin{aligned}&\begin{pmatrix} \boldsymbol{I}_s & \boldsymbol{0}_{s\times(r-s)} \\ \boldsymbol{0}_s^\top & \frac{\boldsymbol{1}_{r-s}^\top}{\sqrt{r-s}} \\ \boldsymbol{0}_{(r-s)\times s} & \boldsymbol{I}_{r-s}-\frac{\boldsymbol{1}_{r-s}\boldsymbol{1}_{r-s}^\top}{r-s}\end{pmatrix}\sqrt{k_n}\,\mathrm{diag}(\boldsymbol{p}_n^*)^{-1/2}\left(\frac{\boldsymbol{\mathcal{T}}_n}{k_n}-\boldsymbol{p}_n^*\right)\\ &\qquad\xrightarrow{\mathcal{D}} \mathcal{N}_{r+1}\left(\boldsymbol{0}_{r+1},\begin{pmatrix}\boldsymbol{I}_{s+1} & \boldsymbol{0}_{(s+1)\times(r-s)}\\ \boldsymbol{0}_{(r-s)\times(s+1)} & \boldsymbol{I}_{r-s}-\frac{\boldsymbol{1}_{r-s}\boldsymbol{1}_{r-s}^\top}{r-s}\end{pmatrix}\right).\end{aligned}$$

Finally, it follows from the independence of $\boldsymbol{\mathcal{T}}_n$ and $\widetilde{\boldsymbol{\mathcal{T}}}_n$ as well as $p_{n,j}/\rho_n \to 1,\ j > s^*$, by

Assumption A, that as $n \to \infty$,

$$\begin{aligned}\boldsymbol{Y}_n = &\begin{pmatrix} \boldsymbol{I}_s & \boldsymbol{0}_{s\times(r-s)} \\ \boldsymbol{0}_s^\top & \frac{\boldsymbol{1}_{r-s}^\top}{\sqrt{r-s}} \\ \boldsymbol{0}_{(r-s)\times s} & \boldsymbol{I}_{r-s} - \frac{\boldsymbol{1}_{r-s}\boldsymbol{1}_{r-s}^\top}{r-s} \end{pmatrix} \sqrt{k_n}\,\mathrm{diag}(\boldsymbol{p}_n^*)^{-1/2}\left(\frac{\boldsymbol{\mathcal{T}}_n}{k_n} - \boldsymbol{p}_n^*\right) \\ &- \begin{pmatrix} \boldsymbol{I}_s & \boldsymbol{0}_{s\times(r-s)} \\ \boldsymbol{0}_s^\top & \frac{\boldsymbol{1}_{r-s}^\top}{\sqrt{r-s}} \\ \boldsymbol{0}_{(r-s)\times s} & \boldsymbol{0}_{(r-s)\times(r-s)} \end{pmatrix} \sqrt{k_n}\,\mathrm{diag}(\boldsymbol{p}_n^*)^{-1/2}\left(\frac{\widetilde{\boldsymbol{\mathcal{T}}}_n}{k_n} - \boldsymbol{p}_n^*\right) + o_{\mathbb{P}}(1) \\ &\xrightarrow{\mathcal{D}} \mathcal{N}_{r+1}(\boldsymbol{0}_{r+1}, \Sigma).\end{aligned}$$

□

A.1.2. Proof of Lemma 3.44

Lemma 3.44. Suppose the assumptions of Proposition 3.11 hold and $\widehat{\underline{\boldsymbol{p}}}_n^s(\boldsymbol{\mathcal{T}}_n)$ is defined analog to $\widehat{\underline{\boldsymbol{p}}}_n^s(\widetilde{\boldsymbol{\mathcal{T}}}_n)$ in (3.9).

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

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

(b) Suppose $\bar{\boldsymbol{p}}_n \coloneqq (\bar{p}_{n,1}, \ldots, \bar{p}_{n,s}, \bar{\rho}_n)^\top$ satisfies

$$\|\bar{\boldsymbol{p}}_n - \widehat{\underline{\boldsymbol{p}}}_n^s(\boldsymbol{\mathcal{T}}_n)\|_1 \le \|\widehat{\underline{\boldsymbol{p}}}_n^s(\widetilde{\boldsymbol{\mathcal{T}}}_n) - \widehat{\underline{\boldsymbol{p}}}_n^s(\boldsymbol{\mathcal{T}}_n)\|_1, \quad n \in \mathbb{N}.$$

Then as $n \to \infty$,

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

Proof. (a) The derivatives of the log-likelihood function are

$$\frac{\partial}{\partial \widetilde{\underline{p}}_j^s} \log L_{\mathcal{N}_r}(\widetilde{\underline{\boldsymbol{p}}}_n \,|\, \boldsymbol{\mathcal{T}}_n) = -\frac{1}{2\widetilde{\underline{p}}_j^s} - \frac{k_n}{2}\frac{(\widetilde{\underline{p}}_j^s)^2 - \frac{\mathcal{T}_{n,j}^2}{k_n^2}}{(\widetilde{\underline{p}}_j^s)^2}, \qquad j = 1, \ldots, s$$

and

$$\frac{\partial}{\partial \widetilde{\underline{\rho}}^s} \log L_{\mathcal{N}_r}(\widetilde{\underline{\boldsymbol{p}}}^s \,|\, \boldsymbol{\mathcal{T}}_n) = -\frac{(r-s)}{2\widetilde{\underline{\rho}}^s} - \frac{k_n}{2}\sum_{j=s+1}^{r}\frac{(\widetilde{\underline{\rho}}^s)^2 - \frac{\mathcal{T}_{n,j}^2}{k_n^2}}{(\widetilde{\underline{\rho}}^s)^2}, \quad \widetilde{\underline{\boldsymbol{p}}}^s \in \mathbb{R}_+^{s+1}.$$

Hence,

$$
\begin{aligned}
&\nabla \log L_{\mathcal{N}_r}(\widehat{\underline{\boldsymbol{p}}}_n^s(\boldsymbol{\mathcal{T}}_n) \,|\, \boldsymbol{\mathcal{T}}_n)(\widehat{\underline{\boldsymbol{p}}}_n^s(\widetilde{\boldsymbol{\mathcal{T}}}_n) - \widehat{\underline{\boldsymbol{p}}}_n^s(\boldsymbol{\mathcal{T}}_n)) \\
&\quad = \sum_{j=1}^{s} \left(-\frac{1}{2} \frac{1}{\widehat{\underline{p}}_{n,j}^s(\boldsymbol{\mathcal{T}}_n)} - \frac{k_n}{2} \frac{(\widehat{\underline{p}}_{n,j}^s(\boldsymbol{\mathcal{T}}_n))^2 - \frac{\mathcal{T}_{n,j}^2}{k_n^2}}{(\widehat{\underline{p}}_{n,j}^s)^2} \right) (\widehat{\underline{p}}_{n,j}^s(\widetilde{\boldsymbol{\mathcal{T}}}_n) - \widehat{\underline{p}}_{n,j}^s(\boldsymbol{\mathcal{T}}_n)) \\
&\qquad - \left(\frac{(r-s)}{2\widehat{\underline{\rho}}_n^s(\boldsymbol{\mathcal{T}}_n)} + \frac{k_n}{2} \sum_{i=s+1}^{r} \frac{(\widehat{\underline{\rho}}_n^s(\boldsymbol{\mathcal{T}}_n))^2 - \frac{\mathcal{T}_{n,i}^2}{k_n^2}}{(\widehat{\underline{\rho}}_n^s(\boldsymbol{\mathcal{T}}_n))^2} \right) (\widehat{\underline{\rho}}_n^s(\widetilde{\boldsymbol{\mathcal{T}}}_n) - \widehat{\underline{\rho}}_n^s(\boldsymbol{\mathcal{T}}_n)) \\
&\quad = -\frac{1}{2} \sum_{j=1}^{s} \frac{(\widehat{\underline{p}}_{n,j}^s(\widetilde{\boldsymbol{\mathcal{T}}}_n) - \widehat{\underline{p}}_{n,j}^s(\boldsymbol{\mathcal{T}}_n))}{\widehat{\underline{p}}_{n,j}^s(\boldsymbol{\mathcal{T}}_n)} - \frac{(r-s)}{2} \frac{(\widehat{\underline{\rho}}_n^s(\widetilde{\boldsymbol{\mathcal{T}}}_n) - \widehat{\underline{\rho}}_n^s(\boldsymbol{\mathcal{T}}_n))}{\widehat{\underline{\rho}}_n^s(\boldsymbol{\mathcal{T}}_n)} \\
&\qquad - \frac{k_n}{2} \left(\widehat{\underline{\rho}}_n^s(\widetilde{\boldsymbol{\mathcal{T}}}_n) - \widehat{\underline{\rho}}_n^s(\boldsymbol{\mathcal{T}}_n) \right) \sum_{i=s+1}^{r} \frac{(\widehat{\underline{\rho}}_n^s(\boldsymbol{\mathcal{T}}_n))^2 - \frac{\mathcal{T}_{n,i}^2}{k_n^2}}{(\widehat{\underline{\rho}}_n^s(\boldsymbol{\mathcal{T}}_n))^2} \\
&\quad =: I_{n,1} + I_{n,2} + I_{n,3}.
\end{aligned}
$$

First, note that $I_{n,1} = o_{\mathbb{P}}(1) = I_{n,2}$ due to Lemma 3.43 and $\sqrt{k_n \rho_n} \to \infty$. Therefore, it remains to investigate $I_{n,3}$. We define the function $g : \mathbb{R}^{r-s} \to \mathbb{R}$ as $g(\boldsymbol{x}) \coloneqq (r-s)^2 \frac{\boldsymbol{x}^\top \boldsymbol{x}}{(\mathbf{1}_{r-s}^\top \boldsymbol{x})^2}$ with Jacobian vector

$$
\nabla g(\boldsymbol{x}) = 2(r-s)^2 \Big(\frac{\boldsymbol{x}^\top}{(\mathbf{1}_{r-s}^\top \boldsymbol{x})^2} - \frac{\boldsymbol{x}^\top \boldsymbol{x} \mathbf{1}_{r-s}^\top}{(\mathbf{1}_{r-s}^\top \boldsymbol{x})^3} \Big) \quad \text{for } \boldsymbol{x} \in \mathbb{R}^{r-s}.
$$

Then, $g(\mathbf{1}_{r-s}) = r-s$ and $\nabla g(\mathbf{1}_{r-s}) = \mathbf{0}_{r-s}^\top$. From Assumption (A4) we already get the asymptotic behavior

$$
\sqrt{k_n \rho_n} \left(\frac{\boldsymbol{\mathcal{T}}_{n,\{s+1,\dots,r\}}}{\rho_n k_n} - \mathbf{1}_{r-s} \right) \xrightarrow{\mathcal{D}} \mathcal{N}_{r-s}(\mathbf{0}_{r-s}, \boldsymbol{I}_{r-s}).
$$

Then an application of the delta-method yields

$$
\sqrt{k_n \rho_n} \left(g\Big(\frac{\boldsymbol{\mathcal{T}}_{n,\{s+1,\dots,r\}}(k_n)}{\rho_n k_n} \Big) - g(\mathbf{1}_{r-s}) \right) \xrightarrow{\mathbb{P}} 0
$$

or equivalently

$$
\begin{aligned}
\sqrt{k_n \rho_n} \left(\frac{\sum_{j=s+1}^{r} \frac{\mathcal{T}_{n,j}^2}{\rho_n^2 k_n^2}}{\left(\sum_{j=s+1}^{r} \frac{\mathcal{T}_{n,j}}{\rho_n k_n (r-s)} \right)^2} - (r-s) \right) &= \sqrt{k_n \rho_n} \Big(\sum_{j=s+1}^{r} \frac{\frac{\mathcal{T}_{n,j}^2}{k_n^2}}{(\widehat{\underline{\rho}}_n^s(\mathcal{T}_n))^2} - (r-s) \Big) \\
&= o_{\mathbb{P}}(1).
\end{aligned}
$$

On the other hand, Lemma 3.43 implies that

$$\sqrt{\frac{k_n}{\rho_n}}(\widehat{\underline{\rho}}_n^s(\widetilde{\boldsymbol{\mathcal{T}}}_n) - \widehat{\underline{\rho}}_n^s(\boldsymbol{\mathcal{T}}_n)) = O_{\mathbb{P}}(1).$$

Finally, this results in

$$\begin{aligned} I_{n,3} &= -\frac{k_n}{2}(\widehat{\underline{\rho}}_n^s(\widetilde{\boldsymbol{\mathcal{T}}}_n) - \widehat{\underline{\rho}}_n^s(\boldsymbol{\mathcal{T}}_n)) \sum_{j=s+1}^{r} \frac{(\widehat{\underline{\rho}}_n^s(\boldsymbol{\mathcal{T}}_n))^2 - \frac{\mathcal{T}_{n,j}^2}{k_n^2}}{(\widehat{\underline{\rho}}_n^s(\boldsymbol{\mathcal{T}}_n))^2} \\ &= \frac{1}{2}\sqrt{\frac{k_n}{\rho_n}}(\widehat{\underline{\rho}}_n^s(\widetilde{\boldsymbol{\mathcal{T}}}_n) - \widehat{\underline{\rho}}_n^s(\boldsymbol{\mathcal{T}}_n))\sqrt{k_n\rho_n}\left(\sum_{j=s+1}^{r} \frac{\frac{\mathcal{T}_{n,j}^2}{k_n^2}}{(\widehat{\underline{\rho}}_n^s(\boldsymbol{\mathcal{T}}_n))^2} - (r-s)\right) = o_{\mathbb{P}}(1). \end{aligned}$$

(b) The Hessian matrix of the log-likelihood function is

$$\begin{aligned} \nabla^2 \log L_{\mathcal{N}_r}(\widetilde{\underline{\boldsymbol{p}}}^s \,|\, \boldsymbol{\mathcal{T}}_n) = \operatorname{diag}\Bigg(&\frac{1}{2(\widetilde{\underline{p}}_1^s)^2} - k_n\frac{\mathcal{T}_{n,1}^2}{k_n^2}\frac{1}{(\widetilde{\underline{p}}_1^s)^3}, \dots, \frac{1}{2(\widetilde{\underline{p}}_s^s)^2} - k_n\frac{\mathcal{T}_{n,s}^2}{k_n^2}\frac{1}{(\widetilde{\underline{p}}_s^s)^3}, \\ &\frac{(r-s)}{2(\widetilde{\underline{\rho}}^s)^2} - k_n\sum_{j=s+1}^{r}\frac{\mathcal{T}_{n,j}^2}{k_n^2}\frac{1}{(\widetilde{\underline{\rho}}^s)^3}\Bigg), \quad \widetilde{\underline{\boldsymbol{p}}}^s \in \mathbb{R}_+^{s+1}. \end{aligned}$$

Let $\bar{\boldsymbol{p}}_n \coloneqq (\bar{p}_{n,1}, \dots, \bar{p}_{n,s}, \bar{\rho}_n)^\top$ with $\|\bar{\boldsymbol{p}}_n - \widehat{\underline{\boldsymbol{p}}}_n^s(\boldsymbol{\mathcal{T}}_n)\|_1 \le \|\widehat{\underline{\boldsymbol{p}}}_n^s(\widetilde{\boldsymbol{\mathcal{T}}}_n) - \widehat{\underline{\boldsymbol{p}}}_n^s(\boldsymbol{\mathcal{T}}_n)\|_1$. Then,

$$\begin{aligned} &\nabla^2 \log L_{\mathcal{N}_r}(\bar{\boldsymbol{p}}_n \,|\, \boldsymbol{\mathcal{T}}_n) + k_n \operatorname{diag}(p_{n,1}, \dots, p_{n,s}, \rho_n/(r-s))^{-1} \\ &\quad= \operatorname{diag}\Big(\frac{1}{2\bar{p}_{n,1}^2} - k_n\frac{\mathcal{T}_{n,1}^2}{k_n^2}\frac{1}{\bar{p}_{n,1}^3} + \frac{k_n}{p_{n,1}}, \dots, \frac{1}{2\bar{p}_{n,s}^2} - k_n\frac{\mathcal{T}_{n,s}^2}{k_n^2}\frac{1}{\bar{p}_{n,s}^3} + \frac{k_n}{p_{n,s}}, \\ &\qquad\frac{(r-s)}{2}\frac{1}{\bar{\rho}_n^2} - k_n\sum_{j=s+1}^{r}\frac{\mathcal{T}_{n,j}^2}{k_n^2}\frac{1}{\bar{\rho}_n^3} + \frac{k_n(r-s)}{\rho_n}\Big) \\ &\quad=: \operatorname{diag}(B_n(1), \dots, B_n(s), B_n(s+1)). \end{aligned}$$

Since $\bar{p}_{n,j}/p_{n,j} \xrightarrow{\mathbb{P}} 1,\ j = 1, \dots, s$, we receive for the entries $B_n(j),\ j = 1, \dots, s$ that

$$\frac{p_{n,j}}{k_n}B_n(j) = \frac{p_{n,j}}{k_n}\left(\frac{1}{2\bar{p}_{n,j}^2} - k_n\frac{\mathcal{T}_{n,j}^2}{k_n^2}\frac{1}{\bar{p}_{n,j}^3} + \frac{k_n}{p_{n,j}}\right) = \frac{p_{n,j}}{2k_n\bar{p}_{n,j}^2} - \frac{\mathcal{T}_{n,j}^2 p_{n,j}}{k_n^2\bar{p}_{n,j}^3} + 1 \xrightarrow{\mathbb{P}} 0. \tag{A.1}$$

Similarly we receive with $\rho_n k_n \to \infty$ and $\bar{\rho}_n/\rho_n \xrightarrow{\mathbb{P}} 1$ for the entry $B_n(s+1)$ that

$$\frac{\rho_n}{k_n}B_n(s+1) = \frac{\rho_n}{k_n}\left(\frac{(r-s)}{2}\frac{1}{\bar{\rho}_n^2} - k_n\sum_{j=s+1}^{r}\frac{\mathcal{T}_{n,j}^2}{k_n^2}\frac{1}{\bar{\rho}_n^3} + \frac{k_n(r-s)}{\rho_n}\right) \xrightarrow{\mathbb{P}} 0. \tag{A.2}$$

Additionally, due to Lemma 3.43 we have as $n \to \infty$,

$$\sqrt{\frac{k_n}{p_{n,j}}}(\widehat{\underline{p}}^s_{n,j}(\widetilde{\boldsymbol{\mathcal{T}}}_n) - \widehat{\underline{p}}^s_{n,j}(\boldsymbol{\mathcal{T}}_n)) = O_{\mathbb{P}}(1) \quad \text{and} \quad \sqrt{\frac{k_n}{\rho_n}}(\widehat{\underline{\rho}}^s_n(\widetilde{\boldsymbol{\mathcal{T}}}_n) - \widehat{\underline{\rho}}^s_n(\boldsymbol{\mathcal{T}}_n)) = O_{\mathbb{P}}(1). \tag{A.3}$$

Therefore, Slutzky's lemma, (A.1), (A.2) and (A.3) yield

$$\begin{aligned}
&(\widehat{\underline{\boldsymbol{p}}}^s_n(\widetilde{\boldsymbol{\mathcal{T}}}_n) - \widehat{\underline{\boldsymbol{p}}}^s_n(\boldsymbol{\mathcal{T}}_n))^\top \operatorname{diag}(B_n(1), \dots, B_n(s), B_n(s+1))(\widehat{\underline{\boldsymbol{p}}}^s_n(\widetilde{\boldsymbol{\mathcal{T}}}_n) - \widehat{\underline{\boldsymbol{p}}}^s_n(\boldsymbol{\mathcal{T}}_n)) \\
&\quad = \sum_{j=1}^{s} \left(\sqrt{\frac{k_n}{p_{n,j}}}(\widehat{\underline{p}}^s_{n,j}(\widetilde{\boldsymbol{\mathcal{T}}}_n) - \widehat{\underline{p}}^s_{n,j}(\boldsymbol{\mathcal{T}}_n))\right)^2 \left(\frac{p_{n,j}}{k_n} B_n(j)\right) \\
&\qquad + \left(\sqrt{\frac{k_n}{\rho_n}}(\widehat{\underline{\rho}}^s_n(\widetilde{\boldsymbol{\mathcal{T}}}_n) - \widehat{\underline{\rho}}^s_n(\boldsymbol{\mathcal{T}}_n))\right)^2 \left(\frac{\rho_n}{k_n} B_n(s+1)\right) \\
&\quad \xrightarrow{\mathbb{P}} 0,
\end{aligned}$$

as $n \to \infty$, the statement.

□

A.2. Auxiliary results for the mean squared error information criterion

In this section, we present supplementary results for Section 3.2.2.

A.2.1. Proof of Lemma 3.47

Lemma 3.47. Suppose Assumptions (B1) and (B2) hold. Then for $\boldsymbol{p}' \in \mathbb{R}^r_+$ the asymptotic behavior

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

as $n \to \infty$ holds.

Proof. Under the Assumptions (B1) and (B2) we get

$$\begin{aligned}
&\mathbb{E}\left[\left(\frac{T'_{n,j}}{(n - T'_{n,2^d})} - p'_j\right)^2 \Big| T'_{n,2^d}\right] \\
&\quad = \mathbb{E}\left[\frac{(T'_{n,j})^2}{(n - T'_{n,2^d})^2} - 2p'_j\frac{T'_{n,j}}{(n - T'_{n,2^d})} + (p'_j)^2\Big| T'_{n,2^d}\right]
\end{aligned}$$

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

Hence,

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

□

A.2.2. Proof of Lemma 3.48

Lemma 3.48. For $q' \in (0,1)$ the equality

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

holds.

Proof. A straightforward calculation gives with

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

the statement. □

A.3. Auxiliary results for the Bayesian information criterion

In this section, we present supplementary results for Section 3.2.3.

A.3.1. Proof of Lemma 3.49 and Lemma 3.50

First, we provide the proofs of the auxiliary results in this subsection.

Lemma 3.49. Let the assumptions of Theorem 3.25 hold. Define the ball

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

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

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

Proof. First, we apply a multivariate Taylor expansion to the log-likelihood function $\log L_{M_{k_n}^s}(\cdot \mid \boldsymbol{T}_n(k_n))$ around the MLE $\widehat{\boldsymbol{p}}_n^s$ at $\widetilde{\boldsymbol{p}}^s$ analog to Lemma 2 of Meyer and Wintenberger (2023) (based on a generalization of Cauchy's Mean Value Theorem see Hille, 1964) which gives the existence of a constant $\theta_n \in (0,1)$ such that

$$\begin{aligned}
&\log L_{M_{k_n}^s}(\widetilde{\boldsymbol{p}}^s \,|\, \boldsymbol{T}_n(k_n)) \\
&\quad= \log L_{M_{k_n}^s}(\widehat{\boldsymbol{p}}_n^s \,|\, \boldsymbol{T}_n(k_n)) + (\widetilde{\boldsymbol{p}}^s - \widehat{\boldsymbol{p}}_n^s)^\top \nabla \log L_{M_{k_n}^s}(\widehat{\boldsymbol{p}}_n^s \,|\, \boldsymbol{T}_n(k_n)) \\
&\qquad+ \frac{1}{2}(\widetilde{\boldsymbol{p}}^s - \widehat{\boldsymbol{p}}_n^s)^\top \nabla^2 \log L_{M_{k_n}^s}(\theta_n \widehat{\boldsymbol{p}}_n^s + (1-\theta_n)\widetilde{\boldsymbol{p}}^s \,|\, \boldsymbol{T}_n(k_n))(\widetilde{\boldsymbol{p}}^s - \widehat{\boldsymbol{p}}_n^s) \\
&\quad= \log L_{M_{k_n}^s}(\widehat{\boldsymbol{p}}_n^s \,|\, \boldsymbol{T}_n(k_n)) \\
&\qquad+ \frac{(\widetilde{\boldsymbol{p}}^s - \widehat{\boldsymbol{p}}_n^s)^\top}{2} \nabla^2 \log L_{M_{k_n}^s}(\theta_n \widehat{\boldsymbol{p}}_n^s + (1-\theta_n)\widetilde{\boldsymbol{p}}^s \,|\, \boldsymbol{T}_n(k_n))(\widetilde{\boldsymbol{p}}^s - \widehat{\boldsymbol{p}}_n^s).
\end{aligned}$$

Thus, we receive

$$\begin{aligned}
&\Big| \log L_{M_{k_n}^s}(\widetilde{\boldsymbol{p}}^s \,|\, \boldsymbol{T}_n(k_n)) - \log L_{M_{k_n}^s}(\widehat{\boldsymbol{p}}_n^s \,|\, \boldsymbol{T}_n(k_n)) \\
&\qquad - \frac{1}{2}(\widetilde{\boldsymbol{p}}^s - \widehat{\boldsymbol{p}}_n^s)^\top \nabla^2 \log L_{M_{k_n}^s}(\widehat{\boldsymbol{p}}_n^s \,|\, \boldsymbol{T}_n(k_n))(\widetilde{\boldsymbol{p}}^s - \widehat{\boldsymbol{p}}_n^s) \Big| \\
&= \frac{1}{2} \Big| (\widetilde{\boldsymbol{p}}^s - \widehat{\boldsymbol{p}}_n^s)^\top \Big(\nabla^2 \log L_{M_{k_n}^s}(\theta_n \widehat{\boldsymbol{p}}_n^s + (1-\theta_n)\widetilde{\boldsymbol{p}}^s \,|\, \boldsymbol{T}_n(k_n)) \\
&\qquad - \nabla^2 \log L_{M_{k_n}^s}(\widetilde{\boldsymbol{p}}^s \,|\, \boldsymbol{T}_n(k_n)) \Big) (\widetilde{\boldsymbol{p}}^s - \widehat{\boldsymbol{p}}_n^s) \Big|. \qquad \text{(A.4)}
\end{aligned}$$

Therefore, to prove the statement, we show that the right side is $o_{\mathbb{P}}(1)$. Inserting the

derivatives of the log-likelihood function

$$\nabla \log L_{M^s_{k_n}}(\widetilde{\boldsymbol{p}}^s \,|\, \boldsymbol{T}_n(k_n)) = \begin{pmatrix} \frac{T_{n,1}}{\widetilde{p}^s_1} - \frac{\sum_{j=s+1}^{2^d-1} T_{n,j}}{1-\sum_{j=1}^{s}\widetilde{p}^s_j} \\ \vdots \\ \frac{T_{n,s}}{\widetilde{p}^s_s} - \frac{\sum_{j=s+1}^{2^d-1} T_{n,j}}{1-\sum_{j=1}^{s}\widetilde{p}^s_j} \end{pmatrix}, \tag{A.5}$$

$$\nabla^2 \log L_{M^s_{k_n}}(\widetilde{\boldsymbol{p}}^s \,|\, \boldsymbol{T}_n(k_n)) = -\operatorname{diag}\Big(\frac{T_{n,1}(k_n)}{(\widetilde{p}^s_1)^2}, \dots, \frac{T_{n,s}(k_n)}{(\widetilde{p}^s_s)^2}\Big) - \frac{\sum_{j=s+1}^{r} T_{n,j}}{\big(1-\sum_{j=1}^{s}\widetilde{p}^s_j\big)^2} \cdot \mathbf{1}_s \cdot \mathbf{1}_s^\top,$$

and applying the triangle inequality yields

$$\begin{aligned}
&\Big|(\widetilde{\boldsymbol{p}}^s - \widehat{\boldsymbol{p}}^s_n)^\top \Big(\nabla^2 \log L_{M^s_{k_n}}(\theta_n \widehat{\boldsymbol{p}}^s_n + (1-\theta_n)\widetilde{\boldsymbol{p}}^s \,|\, \boldsymbol{T}_n(k_n)) \\
&\qquad\qquad - \nabla^2 \log L_{M^s_{k_n}}(\widehat{\boldsymbol{p}}^s_n \,|\, \boldsymbol{T}_n(k_n)) \Big)(\widetilde{\boldsymbol{p}}^s - \widehat{\boldsymbol{p}}^s_n)\Big| \\
&\le \Big|(\widetilde{\boldsymbol{p}}^s - \widehat{\boldsymbol{p}}^s_n)^\top \Big\{ \operatorname{diag}\Big(\frac{T_{n,1}(k_n)}{(\theta_n \widehat{p}^s_{n,1} + (1-\theta_n)\widetilde{p}^s_1)^2}, \dots, \frac{T_{n,s}(k_n)}{(\theta_n \widehat{p}^s_{n,s} + (1-\theta_n)\widetilde{p}^s_s)^2}\Big) \\
&\qquad\qquad - \operatorname{diag}\Big(\frac{T_{n,1}(k_n)}{(\widehat{p}^s_{n,1})^2}, \dots, \frac{T_{n,s}(k_n)}{(\widehat{p}^s_{n,s})^2}\Big) \Big\}(\widetilde{\boldsymbol{p}}^s - \widehat{\boldsymbol{p}}^s_n)\Big| \\
&\quad + \Big|(\widetilde{\boldsymbol{p}}^s - \widehat{\boldsymbol{p}}^s_n)^\top \Big\{ \frac{\sum_{j=s+1}^{r} T_{n,j}}{\Big(1-\sum_{j=1}^{s}(\theta_n \widehat{p}^s_{n,j} + (1-\theta_n)\widetilde{p}^s_j)\Big)^2} \\
&\qquad\qquad - \frac{\sum_{j=s+1}^{r} T_{n,j}}{\Big(1-\sum_{j=1}^{s}\widehat{p}^s_{n,j}\Big)^2} \Big\} \cdot \mathbf{1}_s \cdot \mathbf{1}_s^\top (\widetilde{\boldsymbol{p}}^s - \widehat{\boldsymbol{p}}^s_n)\Big| \\
&=: I_1(\widetilde{\boldsymbol{p}}^s) + I_2(\widetilde{\boldsymbol{p}}^s).
\end{aligned}$$

In the following we only show that $I_2(\widetilde{\boldsymbol{p}}^s)$ is uniformly $o_{\mathbb{P}}(1)$; the calculation for $I_1(\widetilde{\boldsymbol{p}}^s)$ is similar but with a faster rate, since $p_j > 0,\ j = 1, \dots, s^*$ and $\rho_n \to 0$. Therefore, an application of the mean value theorem to the function $x \mapsto 1/x^2$ yields

$$\begin{aligned}
I_2(\widetilde{\boldsymbol{p}}^s) &= k_n \|\widetilde{\boldsymbol{p}}^s - \widehat{\boldsymbol{p}}^s_n\|_1^2 \Big| \frac{\sum_{j=s+1}^{r} \frac{T_{n,j}(k_n)}{k_n}}{\Big(1-\sum_{j=1}^{s}(\theta_n \widehat{p}^s_{n,j} + (1-\theta_n)\widetilde{p}^s_j)\Big)^2} - \frac{\sum_{j=s+1}^{r} \frac{T_{n,j}(k_n)}{k_n}}{\Big(1-\sum_{j=1}^{s}\widehat{p}^s_{n,j}\Big)^2} \Big| \\
&= k_n \|\widetilde{\boldsymbol{p}}^s - \widehat{\boldsymbol{p}}^s_n\|_1^2 (r-s)\widehat{\rho}^s_n \Big| \frac{1}{(\theta_n \widehat{\rho}^s_n + (1-\theta_n)\widetilde{\rho}^s)^2} - \frac{1}{(\widehat{\rho}^s_n)^2} \Big| \\
&\le k_n \|\widetilde{\boldsymbol{p}}^s - \widehat{\boldsymbol{p}}^s_n\|_1^2 (r-s)\widehat{\rho}^s_n \frac{2(1-\theta_n)|\widehat{\rho}^s_n - \widetilde{\rho}^s|}{\min(|\widehat{\rho}^s_n|, |\widetilde{\rho}^s|)^3}
\end{aligned} \tag{A.6}$$

Since $\widetilde{\boldsymbol{p}}^s \in U_{\varepsilon_{n,\gamma}}(\widehat{\boldsymbol{p}}^s_n)$ and $\widehat{\rho}^s_n = O_{\mathbb{P}}(\rho_n)$ we obtain

$$\sup_{\widetilde{\boldsymbol{p}}^s \in U_{\varepsilon_{n,\gamma}}(\widehat{\boldsymbol{p}}^s_n)} I_2(\widetilde{\boldsymbol{p}}^s) = O_{\mathbb{P}}(k_n \varepsilon^3_{n,\gamma} \rho_n^{-2}).$$

Finally, $\varepsilon_{n,\gamma} = (\rho_n)^\gamma/2$ and $k_n(\rho_n)^{3\gamma-2} \leq k_n(\rho_n)^2 \to 0$ (due to Assumption (C3)) which results in the uniform convergence of $\sup_{\widetilde{\boldsymbol{p}}^s \in U_{\varepsilon_{n,\gamma}}(\widehat{\boldsymbol{p}}_n^s)} I_2(\widetilde{\boldsymbol{p}}^s) \xrightarrow{\mathbb{P}} 0$ and the statement follows. □

Next, we derive boundaries for the eigenvalues of the second-order derivative of the log-likelihood function.

Lemma 3.50. Let the assumptions of Theorem 3.25 hold. Define $\lambda_{n,2} \coloneqq \frac{k_n}{T_{n,1}(k_n)}$ and $\lambda_{n,1} \coloneqq \frac{k_n}{T_{n,s}(k_n)} + \frac{sk_n}{\sum_{j=s+1}^r T_{n,j}(k_n)}$. For $\widetilde{\boldsymbol{p}}^s \in \Theta_s$ we have on the one hand,

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

and on the other hand,

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

Proof. Let $\widetilde{\boldsymbol{p}}^s \in \Theta_s$. Inserting the MLE $\widehat{\boldsymbol{p}}_n^s$ in the second order derivative in (A.5) yields

$$\begin{aligned}\frac{-1}{k_n}\nabla^2 \log L_{M_{k_n}^s}(\widehat{\boldsymbol{p}}_n^s \,|\, \boldsymbol{T}_n(k_n)) &= \operatorname{diag}\Big(\frac{k_n}{T_{n,1}(k_n)}, \ldots, \frac{k_n}{T_{n,s}(k_n)}\Big) + \frac{k_n}{\sum_{j=s+1}^r T_{n,j}(k_n)} \cdot \mathbf{1}_s \cdot \mathbf{1}_s^\top \\ &=: M_n + N_n.\end{aligned}$$

The eigenvalues of M_n and N_n are

$$\mu_i = \frac{k_n}{T_{n,s-i+1}(k_n)}, \quad i = 1, \ldots, s,$$

and

$$\nu_1 = \frac{sk_n}{T_{n,s}(k_n)} \quad \text{and} \quad \nu_i = 0, \quad i = 2, \ldots, s,$$

respectively. By $\lambda_1, \ldots, \lambda_s$ with $\lambda_1 \geq \cdots \geq \lambda_s$ we denote the ordered eigenvalues of $M_n + N_n$. Then Weyl's inequality (cf. Horn and Johnson, 2013, p. 239, Theorem 4.3.1) and Assumption (A2) yield

$$\lambda_{n,2} = \frac{k_n}{T_{n,1}(k_n)} = \mu_s \leq \lambda_s \leq \lambda_1 \leq \mu_1 + \nu_1 = \frac{k_n}{T_{n,s}(k_n)} + \frac{sk_n}{\sum_{j=s+1}^r T_{n,j}(k_n)} = \lambda_{n,1}. \quad \text{(A.7)}$$

An application of A.2.5 in Fujikoshi et al. (2010) and inequality (A.7) give then with

$$\begin{aligned}\lambda_{n,2}(\widetilde{\boldsymbol{p}}^s - \widehat{\boldsymbol{p}}_n^s)^\top(\widetilde{\boldsymbol{p}}^s - \widehat{\boldsymbol{p}}_n^s) &\leq (\widetilde{\boldsymbol{p}}^s - \widehat{\boldsymbol{p}}_n^s)^\top \frac{-1}{k_n}\nabla^2 \log L_{M_{k_n}^s}(\widehat{\boldsymbol{p}}_n^s \,|\, \boldsymbol{T}_n(k_n))(\widetilde{\boldsymbol{p}}^s - \widehat{\boldsymbol{p}}_n^s) \\ &\leq \lambda_{n,1}(\widetilde{\boldsymbol{p}}^s - \widehat{\boldsymbol{p}}_n^s)^\top(\widetilde{\boldsymbol{p}}^s - \widehat{\boldsymbol{p}}_n^s)\end{aligned}$$

the statement. □

A.3.2. PROOF OF PROPOSITION 3.51

Proposition 3.51. Under Assumptions (B1), (B3) and (D4) the asymptotic upper bound as $n \to \infty$,

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

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

Proof. Assumption (B3) says that

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

First, we find an upper bound for the first term. Therefore, note that for $j = 1, \dots, s$ the equality

$$\begin{aligned}
&\mathbb{E}\Big[T'_{n,j}\log\big(\widehat{p}^s_{n,j}(\boldsymbol{T}'_{n,\{1,\dots,r\}})\big)\Big|T'_{n,2^d}\Big] \\
&\qquad= \mathbb{E}\Big[T'_{n,j}\log\Big(\frac{T'_{n,j}}{n-T'_{n,2^d}}\Big)\Big|T'_{n,2^d}\Big] \\
&\qquad= \mathbb{E}[T'_{n,j}\log(T'_{n,j})|T'_{n,2^d}] - \mathbb{E}[T'_{n,j}\log(n-T'_{n,2^d})|T'_{n,2^d}]
\end{aligned}$$

holds. An application of (3.33) in the first step (which holds as well in analog form for $T'_{n,j}$) and Assumption (B1) in the second step give then

$$\begin{aligned}
&= \mathbb{E}[T'_{n,j}|T'_{n,2^d}]\log(\mathbb{E}[T'_{n,j}|T'_{n,2^d}]) - \log(n-T'_{n,2^d})\mathbb{E}[T'_{n,j}|T'_{n,2^d}] + O_{\mathbb{P}}(1) \\
&= \frac{n-T'_{n,2^d}}{k_n}\mathbb{E}\Big[T_{n,j}(k_n)\log\Big(\frac{\mathbb{E}[T_{n,j}(k_n)]}{k_n}\Big)\Big] + \frac{n-T'_{n,2^d}}{k_n}\mathbb{E}[T_{n,j}(k_n)]\log(n-T'_{n,2^d}) \\
&\quad - \frac{n-T'_{n,2^d}}{k_n}\mathbb{E}\Big[T_{n,j}(k_n)\Big]\log(n-T'_{n,2^d}) + O_{\mathbb{P}}(1) \\
&= \frac{n-T'_{n,2^d}}{k_n}\mathbb{E}\Big[T_{n,j}(k_n)\log\Big(\frac{\mathbb{E}[T_{n,j}(k_n)]}{k_n}\Big)\Big] + O_{\mathbb{P}}(1),
\end{aligned}$$

where we used in the calculations as well that $(n - T'_{n,2^d})/k_n = O_{\mathbb{P}}(1)$ due Assumption (B3). Finally, we apply again (3.33) to receive

$$\begin{aligned} &= \frac{n - T'_{n,2^d}}{k_n} \mathbb{E}\Big[T_{n,j}(k_n) \log\Big(\frac{T_{n,j}(k_n)}{k_n}\Big)\Big] + O_{\mathbb{P}}(1) \\ &= \frac{n - T'_{n,2^d}}{k_n} \mathbb{E}\Big[T_{n,j}(k_n) \log\big(\widehat{p}^s_{n,j}\big)\Big] + O_{\mathbb{P}}(1). \end{aligned} \tag{A.8}$$

Similarly, we obtain as well

$$\begin{aligned} &\sum_{j=s+1}^{r} \mathbb{E}\Big[T'_{n,j} \log\big(\widehat{\rho}^s_n(\boldsymbol{T}'_{n,\{1,\dots,r\}})\big)\Big|T'_{n,2^d}\Big] \\ &\qquad = \frac{n - T'_{n,2^d}}{k_n} \sum_{j=s+1}^{r} \mathbb{E}\Big[T_{n,j}(k_n) \log\big(\widehat{\rho}^s_n\big)\Big] + O_{\mathbb{P}}(1). \end{aligned} \tag{A.9}$$

A consequence of the log-likelihood function (cf. (3.3)), (A.8) and (A.9) is then

$$\begin{aligned} &\mathbb{E}\Big[-2 \log L_{M^s_{n-T'_{n,2^d}}}(\widehat{\boldsymbol{p}}^s_n(\boldsymbol{T}'_{n,\{1,\dots,r\}}) \,|\, \boldsymbol{T}'_{n,\{1,\dots,r\}})\Big|T'_{n,2^d}\Big] \\ &\quad = -2 \log\big((n - T'_{n,2^d})!\big) + 2\sum_{j=1}^{r} \mathbb{E}[\log(T'_{n,j}!)|T'_{n,2^d}] + 2\frac{n - T'_{n,2^d}}{k_n} \log(k_n!) \\ &\qquad - 2\frac{n - T'_{n,2^d}}{k_n} \log(k_n!) - 2\frac{n - T'_{n,2^d}}{k_n} \sum_{j=1}^{s} \mathbb{E}\Big[T_{n,j}(k_n) \log\big(\widehat{p}^s_{n,j}\big)\Big] \\ &\qquad - 2\frac{n - T'_{n,2^d}}{k_n} \sum_{j=s+1}^{r} \mathbb{E}\Big[T_{n,j}(k_n) \log\big(\widehat{\rho}^s_n\big)\Big] + O_{\mathbb{P}}(1). \end{aligned} \tag{A.10}$$

By the last equality on page 28 in Meyer and Wintenberger (2023) and $\sum_{j=1}^{r} T_{n,j}(k_n) = k_n$ we receive that

$$\begin{aligned} &\sum_{j=1}^{r} \mathbb{E}[\log(T'_{n,j}!)|T'_{n,2^d}] \\ &\quad \leq \frac{n - T'_{n,2^d}}{k_n} \sum_{j=1}^{r} \mathbb{E}[\log(T_{n,j}(k_n)!)] + (n - T'_{n,2^d}) \log\Big(\frac{n - T'_{n,2^d}}{k_n}\Big) + C_1 \log(n - T'_{n,2^d}) \end{aligned} \tag{A.11}$$

and

$$\begin{aligned} &2(n - T'_{n,2^d}) \log\Big(\frac{n - T'_{n,2^d}}{k_n}\Big) + 2\frac{n - T'_{n,2^d}}{k_n} \log(k_n!) \\ &\qquad \leq 2(n - T'_{n,2^d})(\log(n - T'_{n,2^d}) - 1) + C_2 \log(n - T'_{n,2^d}), \end{aligned} \tag{A.12}$$

for some constants $C_1, C_2 > 0$ independent of s and n.

Plugging then (A.11) into (A.10) yields

$$\begin{aligned}
&\mathbb{E}\Big[-2\log L_{M^s_{n-T'_{n,2^d}}}(\widehat{\boldsymbol{p}}^s_n(\boldsymbol{T}'_{n,\{1,\dots,r\}}) \,|\, \boldsymbol{T}'_{n,\{1,\dots,r\}}) \Big| T'_{n,2^d}\Big] \\
&\quad \le -2\log\left((n-T'_{n,2^d})!\right) + 2(n-T'_{n,2^d})\log\Big(\frac{n-T'_{n,2^d}}{k_n}\Big) + 2\frac{n-T'_{n,2^d}}{k_n}\log(k_n!) \\
&\qquad - 2\frac{n-T'_{n,2^d}}{k_n}\Bigg\{ \log(k_n!) - \sum_{j=1}^{r}\mathbb{E}[\log(T_{n,j}(k_n)!)] + \sum_{j=1}^{s}\mathbb{E}\Big[T_{n,j}(k_n)\log\left(\widehat{p}^s_{n,j}\right)\Big] \\
&\qquad\qquad + \sum_{j=s+1}^{r}\mathbb{E}\Big[T_{n,j}(k_n)\log\left(\widehat{\rho}^s_n\right)\Big]\Bigg\} + C_1\log(n-T'_{n,2^d}) \\
&\quad = -2\log\left((n-T'_{n,2^d})!\right) + 2(n-T'_{n,2^d})\log\Big(\frac{n-T'_{n,2^d}}{k_n}\Big) + 2\frac{n-T'_{n,2^d}}{k_n}\log(k_n!) \\
&\qquad - 2\frac{n-T'_{n,2^d}}{k_n}\mathbb{E}[\log L_{M^s_{k_n}}(\widehat{\boldsymbol{p}}^s_n(\boldsymbol{T}_n(k_n)) \,|\, \boldsymbol{T}_n(k_n))] + C_1\log(n-T'_{n,2^d}),
\end{aligned}$$

and using inequality (A.12) gives then

$$\begin{aligned}
&\le -2\log\left((n-T'_{n,2^d})!\right) + 2(n-T'_{n,2^d})(\log(n-T'_{n,2^d})-1) \\
&\quad - 2\frac{n-T'_{n,2^d}}{k_n}\mathbb{E}[\log L_{M^s_{k_n}}(\widehat{\boldsymbol{p}}^s_n(\boldsymbol{T}_n(k_n)) \,|\, \boldsymbol{T}_n(k_n))] + C_3\log(n-T'_{n,2^d}).
\end{aligned}$$

Finally, Assumption (B3), the last upper bound and Jensen's inequality result in

$$\begin{aligned}
&\mathbb{E}\Big[-2\log\mathbb{E}_\lambda[L_{M^s_{n-T'_{n,2^d}}}(\widetilde{\boldsymbol{p}}^s \,|\, \boldsymbol{T}'_{n,\{1,\dots,r\}})]\Big] \\
&\quad \le \ -2\mathbb{E}\Big[\log\left((n-T'_{n,2^d})!\right) - (n-T'_{n,2^d})\Big(\log(n-T'_{n,2^d})-1\Big)\Big] \\
&\qquad - 2\mathbb{E}\Big[\frac{n-T'_{n,2^d}}{k_n}\mathbb{E}[\log L_{M^s_{k_n}}(\widehat{\boldsymbol{p}}^s_n(\boldsymbol{T}_n(k_n)) \,|\, \boldsymbol{T}_n(k_n))]\Big] \\
&\qquad + 2s\mathbb{E}\Big[\log\Big(\Big(n-T'_{n,2^d}\Big)\sqrt{\frac{r}{2\pi(r-s)}}\Big)\Big] + C_3\log(nq_n) \\
&\quad \le \ -2\mathbb{E}\Big[\log\left((n-T'_{n,2^d})!\right) - (n-T'_{n,2^d})\Big(\log(n-T'_{n,2^d})-1\Big)\Big] \\
&\qquad - 2\frac{nq_n}{k_n}\mathbb{E}[\log L_{M^s_{k_n}}(\widehat{\boldsymbol{p}}^s_n(\boldsymbol{T}_n(k_n)) \,|\, \boldsymbol{T}_n(k_n))] + 2s\log\Big(k_n\sqrt{\frac{r}{2\pi(r-s)}}\Big) \\
&\qquad + C\log(nq_n),
\end{aligned}$$

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

A.3.3. Proof of Proposition 3.52

The target of this section is to prove Proposition 3.52. Note that for the estimator $\widehat{q}_n = (n - T'_{n,2^d})/n$ for q_n holds under Assumption (D3) that $\widehat{q}_n/q_n \xrightarrow{\mathbb{P}} 1$, since $\mathbb{E}[\widehat{q}_n/q_n] = 1$

and $\mathbb{V}(\widehat{q}_n/q_n) = nq_n(1-q_n)/(n^2q_n^2) \to 0$.

Lemma A.1. *Let $U'_{\varepsilon'_{n,\gamma}}(\widehat{q}_n) \coloneqq \{q \in (0,1) : |q - \widehat{q}_n| < \varepsilon'_{n,\gamma}\}$, whereby we choose $\varepsilon'_{n,\gamma} \coloneqq q_n^\gamma$ for $\gamma \geq 4/3$. Then under Assumption (D3)*

$$\sup_{\widetilde{q}_n \in U'_{\varepsilon'_{n,\gamma}}(\widehat{q}_n)} \left| \log L_{Bin_n}(1-\widetilde{q}_n \,|\, T'_{n,2^d}) - \log L_{Bin_n}(1-\widehat{q}_n \,|\, T'_{n,2^d}) + \frac{n}{2}\frac{(\widetilde{q}_n - \widehat{q}_n)^2}{\widehat{q}_n} \right| = o_{\mathbb{P}}(1).$$

Proof. First, we expand the log-likelihood function around the MLE $\widehat{q}_n$. The derivatives of $\log L_{Bin_n}(1-q \,|\, T'_{n,2^d})$ are given by

$$\begin{aligned}
\frac{\partial}{\partial q} \log L_{Bin_n}(1-q \,|\, T'_{n,2^d}) &= \frac{-T'_{n,2^d}}{1-q} + \frac{n - T'_{n,2^d}}{q}, \\
\frac{\partial^2}{(\partial q)^2} \log L_{Bin_n}(1-q \,|\, T'_{n,2^d}) &= \frac{-T'_{n,2^d}}{(1-q)^2} - \frac{n - T'_{n,2^d}}{q^2}.
\end{aligned}$$

Applying a Taylor expansion onto $\log L_{Bin_n}(1-q \,|\, T'_{n,2^d})$ at q_n provides for $\widetilde{q}_n \in U'_{\varepsilon'_{n,\gamma}}(\widehat{q}_n)$ that there exists some $c_n \in (0,1)$ such that

$$\begin{aligned}
&\log L_{Bin_n}(1-\widetilde{q}_n \,|\, T'_{n,2^d}) \\
&= \log L_{Bin_n}(1-\widehat{q}_n \,|\, T'_{n,2^d}) + (\widetilde{q}_n - \widehat{q}_n)\frac{\partial}{\partial q} \log L_{Bin_n}(1-\widehat{q}_n \,|\, T'_{n,2^d}) \\
&\qquad + \frac{1}{2}(\widetilde{q}_n - \widehat{q}_n)^2 \frac{\partial^2}{(\partial q)^2} \log L_{Bin_n}(1 - c_n\widehat{q}_n - (1-c_n)\widetilde{q}_n \,|\, T'_{n,2^d}) \\
&= \log L_{Bin_n}(1-\widehat{q}_n \,|\, T'_{n,2^d}) + \frac{n}{2}(\widetilde{q}_n - \widehat{q}_n)^2 \frac{\partial^2}{(\partial q)^2} \frac{1}{n} \log L_{Bin_n}(1 - c_n\widehat{q}_n - (1-c_n)\widetilde{q}_n \,|\, T'_{n,2^d}),
\end{aligned} \tag{A.13}$$

where we used that the MLE $\widehat{q}_n$ is a root of the first derivative.

Then we have

$$\begin{aligned}
&\left| \log L_{Bin_n}(1-\widetilde{q}_n \,|\, T'_{n,2^d}) - \log L_{Bin_n}(1-\widehat{q}_n \,|\, T'_{n,2^d}) + \frac{n}{2}\frac{(\widetilde{q}_n - \widehat{q}_n)^2}{\widehat{q}_n} \right| \\
&= \left| \frac{n}{2}(\widetilde{q}_n - \widehat{q}_n)^2 \frac{\partial^2}{(\partial q)^2} \frac{1}{n} \log L_{Bin_n}(1 - c_n\widehat{q}_n - (1-c_n)\widetilde{q}_n \,|\, T'_{n,2^d}) + \frac{n}{2}\frac{(\widetilde{q}_n - \widehat{q}_n)^2}{\widehat{q}_n} \right| \\
&= \frac{n}{2}(\widetilde{q}_n - \widehat{q}_n)^2 \left| - \frac{1-\widehat{q}_n}{(1 - c_n\widehat{q}_n - (1-c_n)\widetilde{q}_n)^2} - \frac{\widehat{q}_n}{(c_n\widehat{q}_n + (1-c_n)\widetilde{q}_n)^2} + \frac{1}{\widehat{q}_n} \right| \\
&\leq \frac{n}{2}\varepsilon'^2_{n,\gamma} \left| \frac{1-\widehat{q}_n}{(1 - c_n\widehat{q}_n - (1-c_n)\widetilde{q}_n)^2} \right| + \frac{n}{2}\varepsilon'^2_{n,\gamma} \left| \frac{1}{\widehat{q}_n} - \frac{\widehat{q}_n}{(c_n\widehat{q}_n + (1-c_n)\widetilde{q}_n)^2} \right|
\end{aligned}$$

and using that $\frac{n}{2}\varepsilon'^2_{n,\gamma} \le nq_n^2/2 \to 0$ by Assumption (D3) as well as

$$\left|\frac{1-\widehat{q}_n}{(1-c_n\widehat{q}_n-(1-c_n)\widetilde{q}_n)^2}\right| \le \left|(1-\widehat{q}_n)/(1-c_n\widehat{q}_n-(1-c_n)(\widehat{q}_n-\varepsilon))^2\right| \xrightarrow{\mathbb{P}} 1$$

gives

$$\begin{aligned}
&= \frac{n}{2}\varepsilon'^2_{n,\gamma}\frac{\widehat{q}_n}{(c_n\widehat{q}_n+(1-c_n)\widetilde{q}_n)^2}\left|1-\frac{(c_n\widehat{q}_n+(1-c_n)\widetilde{q}_n)}{\widehat{q}_n}\right|\cdot\left(1+\frac{(c_n\widehat{q}_n+(1-c_n)\widetilde{q}_n)}{\widehat{q}_n}\right)+o_{\mathbb{P}}(1)\\
&= \frac{n}{2}\varepsilon'^2_{n,\gamma}\frac{1}{(c_n\widehat{q}_n+(1-c_n)\widetilde{q}_n)^2}(1-c_n)|\widehat{q}_n-\widetilde{q}_n|\cdot\left(1+\frac{(c_n\widehat{q}_n+(1-c_n)\widetilde{q}_n)}{\widehat{q}_n}\right)+o_{\mathbb{P}}(1)\\
&\le \frac{n}{2}\varepsilon'^2_{n,\gamma}\frac{1}{(\widehat{q}_n+(1-c_n)\varepsilon'_{n,\gamma})^2}(1-c_n)\varepsilon'_{n,\gamma}\cdot\underbrace{\left(1+\frac{(\widehat{q}_n+(1-c_n)\varepsilon'_{n,\gamma})}{\widehat{q}_n}\right)}_{\to 2}+o_{\mathbb{P}}(1)\\
&= O_{\mathbb{P}}(n\varepsilon'^3_{n,\gamma}/q_n^2) = O_{\mathbb{P}}(nq_n^{3\gamma-2}) = o_{\mathbb{P}}(1),
\end{aligned}$$

where we used that $\gamma \ge 4/3$ as well as the consistency of $\widehat{q}_n$ for q_n and thus $\varepsilon'^3_{n,\gamma} = q_n^{3\gamma} \le q_n^4$ and $nq_n^2 \to 0$ by Assumption (D3). Thus the assertion follows.

□

Lemma A.2. *Under Assumption (D3) we have for sufficiently large n that*

$$\begin{aligned}
&-2\log \mathbb{E}_\lambda[L_{\mathrm{Bin}_n}(1-\widetilde{q}\,|\,T'_{n,2^d})]\\
&\qquad \le -2\log L_{\mathrm{Bin}_n}(1-\widehat{q}_n\,|\,T'_{n,2^d}) - \log(2\pi) + \log(n/\widehat{q}_n) + o_{\mathbb{P}}(1),
\end{aligned}$$

where $\widehat{q}_n \coloneqq (n-T'_{n,2^d})/n$ is an estimator for q_n. The expectation of the $o_{\mathbb{P}}(1)$ term is of order $o(1)$.

Proof. It follows from Lemma A.1 that

$$\log L_{Bin_n}(1-\widetilde{q}_n\,|\,T'_{n,2^d}) = \log L_{Bin_n}(1-\widehat{q}_n\,|\,T'_{n,2^d}) - \frac{n}{2}\frac{(\widetilde{q}_n-\widehat{q}_n)^2}{\widehat{q}_n} + o_{\mathbb{P}}(1)$$

uniformly for $\widetilde{q}_n \in U'_{\varepsilon'_{n,\gamma}}(\widehat{q}_n)$ and $\varepsilon'_{n,\gamma} = q_n^\gamma$ for $\gamma \ge 4/3$.

Due to the uniformity of the limit in the last equation we receive similarly to (3.25)

$$\begin{aligned}
&\mathbb{E}_\lambda[L_{Bin_n}(1-q\,|\,T'_{n,2^d})]\\
&\quad \ge \int_{U'_{\varepsilon'_{n,\gamma}}(\widehat{q}_n)} \exp\left(\log L_{Bin_n}(1-q\,|\,T'_{n,2^d})\right)\mathrm{d}q\\
&\quad = L_{Bin_n}(1-\widehat{q}_n\,|\,T'_{n,2^d})\int_{U'_{\varepsilon'_{n,\gamma}}(\widehat{q}_n)} \exp\left(-\frac{n}{2}\frac{(q-\widehat{q}_n)^2}{\widehat{q}_n}\right)\mathrm{d}q\cdot(1+o_{\mathbb{P}}(1))
\end{aligned}$$

and thus,

$$-2\log \mathbb{E}_q[L_{Bin_n}(1-q\,|\,T'_{n,2^d})]$$

$$
\begin{aligned}
&\leq -2\log L_{Bin_n}(1-\widehat{q}_n \,|\, T'_{n,2^d}) - 2\log \int_{U'_{\varepsilon'_{n,\gamma}}(\widehat{q}_n)} \exp\left(-\frac{n}{2}\frac{(q-\widehat{q}_n)^2}{\widehat{q}_n}\right) \mathrm{d}q + o_{\mathbb{P}}(1) \\
&= -2\log L_{Bin_n}(1-\widehat{q}_n \,|\, T'_{n,2^d}) \\
&\quad - 2\log \int_{U'_{\varepsilon'_{n,\gamma}}(\widehat{q}_n)} \sqrt{\frac{2\pi}{n/\widehat{q}_n}}\sqrt{\frac{n/\widehat{q}_n}{2\pi}} \exp\left(-\frac{1}{2}\frac{(q-\widehat{q}_n)^2}{\widehat{q}_n/n}\right) \mathrm{d}q + o_{\mathbb{P}}(1) \\
&= -2\log L_{Bin_n}(1-\widehat{q}_n \,|\, T'_{n,2^d}) - 2\log \sqrt{\frac{2\pi}{n/\widehat{q}_n}} \\
&\quad - 2\log \int_{U'_{\varepsilon'_{n,\gamma}}(\widehat{q}_n)} \sqrt{\frac{n/\widehat{q}_n}{2\pi}} \exp\left(-\frac{1}{2}\frac{(q-\widehat{q}_n)^2}{\widehat{q}_n/n}\right) \mathrm{d}q + o_{\mathbb{P}}(1).
\end{aligned}
$$

For the next step we need $\frac{n}{q_n}\varepsilon'^2_{n,\gamma} = nq_n^{2\gamma-1} \to \infty$. Since $\gamma \geq 4/3$ and $nq_n^{5/3} \to \infty$ by Assumption (D3) we define $\varepsilon'_n \coloneqq \varepsilon'_{n,4/3}$ and have $2\gamma - 1 = 5/3$. Due to Assumption (D3) and $\widehat{q}_n/q_n \xrightarrow{\mathbb{P}} 1$ we obtain similarly to (3.27) that

$$
\log \int_{U'_{\varepsilon'_n}(\widehat{q}_n)} \sqrt{\frac{n/\widehat{q}_n}{2\pi}} \exp\left(-\frac{1}{2}\frac{(q-\widehat{q}_n)^2}{\widehat{q}_n/n}\right) \mathrm{d}q = o_{\mathbb{P}}(1).
$$

Thus the assertion follows.

□

Proposition 3.52. Suppose Assumptions (B3) and (D3) hold. The expectation of the binomial likelihood satisfies as $n \to \infty$ the inequality

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

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

Proof. Without loss of generality, we assume in the following that the constant $C > 0$, which is independent of s and n, is chosen sufficiently large such that the following inequalities hold.

Under Assumption (B3), we are allowed to use the second equation on page 31 in the proof of Lemma 6 in Meyer and Wintenberger (2023)

$$
\begin{aligned}
&\mathbb{E}[\log L_{\mathrm{Bin}_n}(1-q_n \,|\, T'_{n,2^d})] \\
&\quad = \mathbb{E}[\log L_{\mathrm{Bin}_n}(1-\frac{k_n}{n} \,|\, T'_{n,2^d})] + \left(\frac{k_n}{n} - q_n\right)\left(\frac{nq_n}{k_n/n} - \frac{n(1-q_n)}{1-k_n/n}\right).
\end{aligned}
$$

A combination with the asymptotic expansion in the last equation on page 31 in the proof

of Lemma 6 in Meyer and Wintenberger (2023)

$$\frac{nq_n}{k_n/n} - \frac{n(1-q_n)}{1-k_n/n} = \left(q_n - \frac{k_n}{n}\right)\frac{n}{k_n/n} + O(k_n),$$

gives then

$$\mathbb{E}[\log L_{\mathrm{Bin}_n}(1-q_n \,|\, T'_{n,2^d})] = \mathbb{E}[\log L_{\mathrm{Bin}_n}(1-\frac{k_n}{n} \,|\, T'_{n,2^d})] - \left(\frac{k_n}{n} - q_n\right)^2 \frac{n}{k_n/n} + O(k_n).$$

By Assumption (B3) follows the existence of a positive constant $C_1 > 0$ such that

$$\mathbb{E}[\log L_{\mathrm{Bin}_n}(1-q_n \,|\, T'_{n,2^d})] \geq \mathbb{E}[\log L_{\mathrm{Bin}_n}(1-\frac{k_n}{n} \,|\, T'_{n,2^d})] - \left(\frac{k_n}{n} - q_n\right)^2 \frac{n}{k_n/n} - C_1 nq_n.$$

Since $nq_n \to \infty$ and for $\widehat{q}_n := (n - T'_{n,2^d})/n$ we have

$$\mathbb{E}[\log L_{\mathrm{Bin}_n}(1-q_n \,|\, T'_{n,2^d})] - \mathbb{E}[\log L_{\mathrm{Bin}_n}(1-\widehat{q}_n \,|\, T'_{n,2^d})] \to 0,$$

as $n \to \infty$, it follows the existence of a constant $C_2 > 0$ such that

$$\mathbb{E}[\log L_{\mathrm{Bin}_n}(1-\widehat{q}_n \,|\, T'_{n,2^d})] \geq \mathbb{E}[\log L_{\mathrm{Bin}_n}(1-\frac{k_n}{n} \,|\, T'_{n,2^d})] - \left(\frac{k_n}{n} - q_n\right)^2 \frac{n}{k_n/n} - C_2 nq_n.$$

A combination of Lemma A.2 and the equation above gives the existence of a constant $C_3 > 0$ such that

$$\begin{aligned}
&-2\mathbb{E}[\log \mathbb{E}_\lambda[L_{\mathrm{Bin}_n}(1-\widetilde{q} \,|\, T'_{n,2^d})]] \\
&\quad \leq -2\mathbb{E}[\log L_{\mathrm{Bin}_n}(1-\widehat{q}_n \,|\, T'_{n,2^d})] - \log(2\pi) + \mathbb{E}\left[\log\left(\frac{n}{\widehat{q}_n}\right)\right] + o(1) \\
&\quad \leq -2\mathbb{E}[\log L_{\mathrm{Bin}_n}(1-\frac{k_n}{n} \,|\, T'_{n,2^d})] + 2\Big(\frac{k_n}{n} - q_n\Big)^2 \frac{n}{k_n/n} + \mathbb{E}\Big[\log\left(\frac{n}{\widehat{q}_n}\right)\Big] + C_3 nq_n. \quad \text{(A.14)}
\end{aligned}$$

Inserting

$$\begin{aligned}
&\mathbb{E}[\log L_{\mathrm{Bin}_n}(1-\frac{k_n}{n} \,|\, T'_{n,2^d})] \\
&= \log(n!) - \mathbb{E}[\log((n-T'_{n,2^d})!)] - \mathbb{E}[\log(T'_{n,2^d}!)] + n(1-q_n)\log\left(1-\frac{k_n}{n}\right) + nq_n \log\left(\frac{k_n}{n}\right)
\end{aligned}$$

into (A.14) yields

$$\begin{aligned}
&-2\mathbb{E}[\log \mathbb{E}_\lambda[L_{\mathrm{Bin}_n}(1-\widetilde{q} \,|\, T'_{n,2^d})]] \\
&\quad \leq -2\log(n!) + 2\mathbb{E}[\log((n-T'_{n,2^d})!)] + 2\mathbb{E}[\log(T'_{n,2^d}!)] - 2n(1-q_n)\log\left(1-\frac{k_n}{n}\right) \\
&\qquad - 2nq_n \log\left(\frac{k_n}{n}\right) + 2\left(\frac{k_n}{n} - q_n\right)^2 \frac{n}{\frac{k_n}{n}} + \mathbb{E}\left[\log\left(\frac{n}{\widehat{q}_n}\right)\right] + C_3 nq_n. \quad \text{(A.15)}
\end{aligned}$$

We have by Assumption (B3) that $-\log(1-k_n/n) \leq C_4 q_n$ for some $C_4 > 0$ and thus,

$$-2n(1-q_n)\log(1-k_n/n) + 2\left(\frac{k_n}{n} - q_n\right)^2 \frac{n}{k_n/n} \leq C_5 n q_n \tag{A.16}$$

for some $C_5 > 0$.

Finally, we use for $B \sim \mathrm{Bin}(n, p_n)$ with $np_n \to \infty$ a Taylor expansion and the Chernoff inequality resulting in the existence of a positive constant $C > 0$ such that

$$\log(\mathbb{E}[B]) - C \leq \mathbb{E}[\log(B)\mathbb{1}\{B > 0\}].$$

But due to Assumption (D3) we know that $n\widehat{q}_n = n - T'_{n,2^d} \sim \mathrm{Bin}(n, q_n)$ with $nq_n \to \infty$ such that

$$\mathbb{E}\left[\log\left(\frac{n}{\widehat{q}_n}\right)\right] \leq \log\left(\frac{n}{q_n}\right) + C_6 \leq 2\log(n) + C_6 \tag{A.17}$$

for some constant $C_6 > 0$. Hence, the statement follows from (A.15)-(A.17). □

A.4. Auxiliary results for Section 5.2.1

Lemma A.3. *Let*

$$\boldsymbol{\varepsilon}_d \sim \left|\mathcal{N}_d\Big(\mathbf{0}_d, \frac{100}{d}\boldsymbol{I}_d\Big)\right|,$$

where the absolute value is entry-wise. Then

$$\lim_{d\to\infty} \mathbb{V}(\|\boldsymbol{\varepsilon}_d\|_2) = 100/\sqrt{2}.$$

Proof. Indeed, since $\|\boldsymbol{\varepsilon}_d\|_2^2 \sim 100/d \cdot \chi_d^2$, where χ_d^2 is a chi-square distribution with d degrees of freedom, the formula for the moments of a chi-square distribution (cf. Hogg et al., 2005, Theorem 3.3.2) gives

$$\mathbb{V}(\|\boldsymbol{\varepsilon}_d\|_2) = \mathbb{E}[\|\boldsymbol{\varepsilon}_d\|_2^2] - (\mathbb{E}[\|\boldsymbol{\varepsilon}_d\|_2])^2 = \frac{100}{d}\Big(d - \Big(\frac{\sqrt{2}\Gamma((d+1)/2)}{\Gamma(d/2)}\Big)^2\Big).$$

Further by Gautschi's inequality (cf. Elezović et al., 2000, p. 1) we have

$$\left(\frac{d-1}{2}\right)^{1/2} \leq \frac{\Gamma((d+1)/2)}{\Gamma(d/2)} \leq \left(\frac{d-1}{2} + 1\right)^{1/2}$$

and therefore

$$\frac{100}{\sqrt{2}}\frac{d-1}{d} = \frac{100}{d}\sqrt{2}\Big(d - \frac{d-1}{2} - 1\Big) \leq \mathbb{V}(\|\boldsymbol{\varepsilon}_d\|_2) \leq \frac{100}{d}\sqrt{2}\Big(d - \frac{d-1}{2}\Big) = \frac{100}{\sqrt{2}}\frac{d+1}{d}.$$

Letting $d \to \infty$ on the left and on the right-hand side gives the statement. □

Notation

Symbols

$\xrightarrow{\mathbb{P}\text{-a.s.}}$	almost sure (a.s.) convergence
$\boldsymbol{I}_d$	identity matrix in $\mathbb{R}^{d\times d}$
$\boldsymbol{x}^a, \sqrt{\boldsymbol{x}}, \boldsymbol{x}\cdot\boldsymbol{y}$	operations for $a\in\mathbb{R}$ and $\boldsymbol{x},\boldsymbol{y}\in\mathbb{R}^d$ that are meant component-wise
Cov	covariance (matrix)
$\xrightarrow{\mathcal{D}}$	convergence in distribution
$\operatorname{diag}(\boldsymbol{x})$	a diagonal matrix in $\mathbb{R}^{d\times d}$ with the components of a vector $\boldsymbol{x}\in\mathbb{R}^d$ on the diagonal
$\mathbb{E}$	expectation (vector)
$\frac{\partial}{\partial x_i} f(\boldsymbol{x})$	partial derivative of a function f with respect to the i-th component of the vector $\boldsymbol{x}$
$\mathbf{0}_d$	zero vector $(0,\dots,0)^\top$ in $\mathbb{R}^d$
$\mathbf{1}_d$	vector of ones $(1,\dots,1)^\top$ in $\mathbb{R}^d$
$\mathcal{C}(f)$	set of continuity points of a function f
$\mathcal{P}_d, \mathcal{P}(\mathbb{N})$	power set of $\{1,\dots,d\}$ and power set of $\mathbb{N}$
$\mathbb{N}$	set of natural numbers
$\nabla f(\boldsymbol{x})$	gradient of a function $f:\mathbb{R}^d\mapsto\mathbb{R}^k$, represented as a matrix in $\mathbb{R}^{k\times d}$
$\mathbb{P}$	probability
∂A	boundary of a set A
$\xrightarrow{\mathbb{P}}$	convergence in probability
$\mathcal{P}_d^*, \mathcal{P}_{\mathbb{N}}^*$	power set of $\{1,\dots,d\}$ and $\mathbb{N}$ excluding the empty set, i.e, $\mathcal{P}_d\setminus\{\emptyset\}$ and $\mathcal{P}(\mathbb{N})\setminus\{\emptyset\}$
π	Euclidean projection onto the simplex $\mathbb{S}_+^{d-1}$
$\mathbb{R}$, $\mathbb{R}_+$	set of (positive) real numbers
$\mathbb{R}^d$, $\mathbb{R}_+^d$	set of (positive) real d-dimensional vectors
$\mathbb{S}^{d-1}$	unit sphere with respect to the L_2-norm, i.e, $\{\boldsymbol{x}\in\mathbb{R}^d:\ \|\boldsymbol{x}\|_2=1\}$
$\mathbb{S}_+^{d-1}$	simplex with respect to the L_1-norm, i.e, $\{\boldsymbol{x}\in\mathbb{R}_+^d:\ \|\boldsymbol{x}\|_1=1\}$

$\mathbb{V}$	variance
$\xrightarrow{v}$	vague convergence
$\|\boldsymbol{x}\|_1$	L_1-norm, $\sum_{j=1}^{d} \|x_j\|$, for a vector $\boldsymbol{x} \in \mathbb{R}^d$
$\|\boldsymbol{x}\|_2$	Euclidean (L_2) norm, $\sqrt{\sum_{j=1}^{d} x_j^2}$, for a vector $\boldsymbol{x} \in \mathbb{R}^d$
$\xrightarrow{w}$	weak convergence

List of abbreviations

(M)EVT	(multivariate) extreme value theory
AIC	Akaike information criterion
BIC	Bayesian information criterion
ESD	empirical spectral distribution
i.i.d.	independent and identically distributed
Lasso	Least Absolute Shrinkage and Selection Operator
LSD	limiting spectral distribution
MLE	maximum likelihood estimator
MP law	Marčenko-Pastur law
MSE	mean squared error
MSEIC	mean squared error information criterion
PCA	Principal Component Analysis
QAIC	quasi-Akaike information criterion
RV	regular variation
SRV	sparse regular variation