Check for updates

Received: 28 May 2025 | Revised: 1 September 2025 | Accepted: 1 October 2025

DOI: 10.1111/sjos.70026

ORIGINAL ARTICLE

Scandinavian Journal of Statistics

Estimation of the number of principal components in high-dimensional multivariate extremes

Lucas Butsch | **Vicky Fasen-Hartmann**

Institute of Stochastics, Karlsruhe Institute of Technology, Germany

Correspondence
Vicky Fasen-Hartmann, Institute of Stochastics, Karlsruhe Institute of Technology, Englerstr. 2 76131 Karlsruhe, Germany.
Email: vicky.fasen@kit.edu

Abstract

For multivariate regularly random vectors of dimension d, the dependence structure of the extremes is modeled by the so-called angular measure. When the dimension d is high, estimating the angular measure is challenging because of its complexity. In this paper, we use Principal Component Analysis (PCA) as a method for dimension reduction and estimate the number of significant principal components of the empirical covariance matrix of the angular measure under the assumption of a spiked covariance structure. Therefore, we develop Akaike Information Criteria (AIC) and Bayesian Information Criteria (BIC) to estimate the location of the spiked eigenvalue of the covariance matrix, reflecting the number of significant components, and explore these information criteria on consistency. On the one hand, we investigate the case where the dimension d is fixed, and on the other hand, where the dimension d converges to ∞ under different high-dimensional scenarios. When the dimension d is fixed, we establish that the AIC is not consistent, whereas the BIC is weakly consistent. In high-dimensional contexts, we utilize methods from random matrix theory to establish sufficient conditions ensuring the consistency of the AIC and BIC.

This is an open access article under the terms of the Creative Commons Attribution License, which permits use, distribution and reproduction in any medium, provided the original work is properly cited.
© 2025 The Author(s). *Scandinavian Journal of Statistics* published by John Wiley & Sons Ltd on behalf of The Board of the Foundation of the Scandinavian Journal of Statistics.

Finally, the performance of the different information criteria is compared in a simulation study and applied to high-dimensional precipitation data.

KEYWORDS

AIC, BIC, dimension reduction, high-dimensional, multivariate extremes, multivariate regular variation, PCA, spiked model

1 | INTRODUCTION

In multivariate extreme value theory, extremes occur per se rarely so that the dimensionality of the data in fields such as finance, insurance, meteorology, hydrology, and, more broadly, environmental risk assessment often approaches or exceeds the number of extreme observations which is a big challenge in the statistical analysis of complex and high-dimensional data. Therefore, a standard approach from multivariate statistics is to apply a dimension reduction method to reduce the model complexity and circumvent the curse of dimensionality. A very nice overview of different methods for constructing sparsity in high-dimensional multivariate extremes is given in Engelke and Ivanovs (2021), including PCA, spherical k-means, graphical models, and sparse regular variation, to mention a few.

In multivariate statistics, PCA is a widely used method for dimension reduction, data visualization, clustering and feature extraction (Anderson, 2003; Muirhead, 1982). In recent years, the literature on implementing PCA for high-dimensional and complex data of multivariate extremes to construct some sparsity in the data has grown rapidly (Chautru, 2015; Clémençon et al., 2024; Cooley & Thibaud, 2019; Drees & Sabourin, 2021; Rohrbeck & Cooley, 2023; Drees, 2025; Wan, 2024). A classical concept of multivariate extreme value theory is multivariate regular variation (Falk, 2019; Resnick, 1987, 2007). A d-dimensional random vector $\boldsymbol{X}$ is *multivariate regularly varying*of index $\alpha > 0$ if there exists a random vector $\boldsymbol{\Theta}$ on the unit sphere such that

$$\mathbb{P}\left(\frac{\|\boldsymbol{X}\|}{t} > r, \frac{\boldsymbol{X}}{\|\boldsymbol{X}\|} \in \cdot \,\middle|\, \|\boldsymbol{X}\| > t\right) \xrightarrow{D} r^{-\alpha}\mathbb{P}(\boldsymbol{\Theta} \in \cdot), \quad t \to \infty,$$

for all $r > 0$. The dependence structure of the extremes of $\boldsymbol{X}$ is modeled in the spectral vector $\boldsymbol{\Theta}$ whose distribution is also called *angular measure* and its covariance matrix is denoted as $\boldsymbol{\Sigma} = \mathrm{Cov}(\boldsymbol{\Theta})$. Drees (2025) and Drees and Sabourin (2021) set the mathematical framework for PCA for the empirical covariance estimator of $\boldsymbol{\Sigma}$ by analyzing the squared reconstruction error, the excess risk and their asymptotic behavior. However, until now, the research on the number of significant principal components, the so-called *dimensionality*, in PCA for multivariate extremes, is limited. In Drees and Sabourin (2021), the dimension was estimated through the examination of 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. But a big challenge in extreme value theory is the choice of the threshold t, which defines the extreme observations as the data whose norm is above t. Changing this threshold also changes the number of extreme observations and the estimates of the empirical eigenvalues. Thus, a change of the threshold results in a different scree plot and possibly also in a different elbow. Given the ambiguity and uncertainty of both procedures, as well as the need for case-by-case evaluation, a mathematically based approach is necessary. One

14679469, 0, Downloaded from https://onlinelibrary.wiley.com/doi/10.1111/sjos.70026 by Karlsruher Institut Für Technologie, Wiley Online Library on [22/10/2025]. See the Terms and Conditions (https://onlinelibrary.wiley.com/terms-and-conditions) on Wiley Online Library for rules of use; OA articles are governed by the applicable Creative Commons License

14679469, 0, Downloaded from https://onlinelibrary.wiley.com/doi/10.1111/sjos.70026 by Karlsruher Institut Für Technologie, Wiley Online Library on [22/10/2025]. See the Terms and Conditions (https://onlinelibrary.wiley.com/terms-and-conditions) on Wiley Online Library for rules of use; OA articles are governed by the applicable Creative Commons License

of the few works that developed a statistical method for estimating the dimensionality is given in Drees (2025), whose method is based on asymptotic results for the reconstruction error of the projections and is a kind of testing problem with the disadvantage that it depends on different tuning parameters.

In this paper, we propose information criteria to estimate the number of significant principal components in multivariate extremes modeled through the covariance matrix $\mathbf{\Sigma}$ of $\mathbf{\Theta}$. Therefore, we combine an approach of Bai et al. (2018) and Jiang et al. (2023) from high-dimensional statistics with methods from extreme value theory (de Haan and Ferreira, 2006; Resnick, 1987, 2007) and random matrix theory (Bai & Silverstein, 2010). We assume a *spiked covariance model* for the covariance matrix $\mathbf{\Sigma}$ which goes back to Johnstone (2001) and is widely used in high-dimensional statistics (Bai et al., 2018; Bai & Yao 2012; Fujikoshi & Sakurai 2016; Jiang et al., 2023; Johnstone & Yang, 2018) with applications in various fields, for example, speech recognition, wireless communication, and statistical learning as mentioned in Paul (2007).

Spiked Covariance Model: *The eigenvalues* $\lambda_1,\dots,\lambda_d$ *of* $\mathbf{\Sigma}$ *satisfy*

$$\lambda_1 \geq \lambda_2 \geq \cdots \geq \lambda_{p^*} > \lambda_{p^*+1} = \cdots = \lambda_{d-1} =: \lambda > 0. \tag{1.1}$$

The smallest eigenvalue λ_d is not considered to avoid numerical instability as it can be equal to 0, if, for example, $\mathbf{\Theta}$ is concentrated on the subspace of the unit sphere with non-negative values. The main objective of this article is to develop an estimator $\widehat{p}_n$ for the unknown *dimensionality parameter* p^* of leading eigenvalues of $\mathbf{\Sigma}$ through information criteria; *leading eigenvalues* are defined to be all eigenvalues that are greater than λ. When d is relatively large compared to $\widehat{p}_n$, a useful lower-dimensional representation can be obtained by projecting the data on the $\widehat{p}_n$-dimensional subspace spanned by the empirical eigenvectors of the largest $\widehat{p}_n$ empirical eigenvalues $\widehat{\lambda}_{n,1} \geq \dots \geq \widehat{\lambda}_{n,\widehat{p}_n}$. This lower-dimensional representation allows for more extensive and in-depth analyses of the dependence structure in the extremes.

The estimation of the dimensionality p^* in PCA is explored in this paper using two information criteria: the Akaike Information Criterion (AIC) and the Bayesian Information Criterion (BIC). Similar information criteria 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. Both criteria are motivated by a Gaussian likelihood function, although our underlying model for $\mathbf{\Theta}$ is not Gaussian. Since in a Gaussian spiked covariance model with k_n observations, 2 times the negative log likelihood function can be written as a functional of the empirical eigenvalues in the form

$$k_n \sum_{i=1}^{p^*} \log(\widehat{\lambda}_{n,i}) + k_n(d-p^*) \log\left(\sum_{j=p^*}^{d} \frac{\widehat{\lambda}_{n,j}}{d-p^*} \right) + k_n \log\left(\frac{k_n - 1}{k_n} \right)^d + k_n d(\log(2\pi) + 1),$$

both the AIC and BIC are defined as functionals of the empirical eigenvalues $\widehat{\lambda}_{n,1} \geq \dots \geq \widehat{\lambda}_{n,d}$. In order to align with the extreme value setting that follows, where we have n observations but only k_n of these are extreme, we have denoted the number of observations as k_n instead of n. In the classical Gaussian setting, $k_n = n$.

The main goal of this paper is to derive necessary and sufficient conditions for our AIC and BIC to be consistent. Therefore, we require methods from random matrix theory to derive the asymptotic properties of the empirical eigenvalues, which are the basic components of the AIC and BIC.

For this purpose, we differentiate between two cases when n observations are available and of these k_n are extreme. The first is the classic large sample size and fixed-dimension case, where $n \to \infty$ and the dimension d are fixed. As is typical for such information criteria, we find that the BIC is consistent, whereas the AIC is, in general, not consistent. In the second case, we assume that $d = d_n$ also depends on n and $d_n/k_n \to c > 0$ as $n \to \infty$. In this case, the empirical eigenvalues are not consistent estimators for the eigenvalues anymore. For high-dimensional i.i.d. data with finite fourth moments, it is well-known that the *empirical spectral distribution function* converges to the Marčenko–Pastur law (Marčenko & Pastur, 1967), which describes the bulk distribution of the empirical eigenvalues. The spiked covariance model, introduced by Johnstone (2001), extends the Marčenko-Pastur framework by adding a small number of spiked eigenvalues corresponding to relevant dimension for the PCA. In the context of this paper, we derive as well the asymptotic properties of the empirical eigenvalues of $\mathbf{\Sigma}$ with the Marčenko–Pastur distribution in the limit and use it for the investigation of the consistency of our information criteria. To the best of our knowledge, this paper is the first one to develop consistent information criteria for the dimensionality p^* of the PCA in high-dimensional multivariate extremes. The only other information criteria of Meyer and Wintenberger (2023) and Butsch and Fasen-Hartmann (2025) use the concept of sparse regular variation to construct sparsity in the data, in contrast to PCA.

1.1 | Structure of the paper

This paper is organized as follows: In Section 2, we properly define the empirical eigenvalues $\widehat{\lambda}_{n,1},\dots,\widehat{\lambda}_{n,d}$ 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 3 are the AIC and the BIC for estimating the location p^* of the spiked eigenvalue in the fixed-dimensional case, where Section 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 4.1 and Section 4.2, respectively. In both cases, we derive sufficient criteria for the AIC and the BIC to be weakly consistent. In a simulation study in Section 5, we compare the different information criteria and apply them to precipitation data in Section 6. Finally, we state a conclusion in Section 7. The proofs for the results presented in this paper are provided in the Appendix.

1.2 | Notation

Throughout the paper, we use the following notation and assume that all random variables are defined on the same probability space $(\Omega, \mathcal{F}, \mathbb{P})$. First of all, $\|\boldsymbol{x}\|$ is the Euclidean norm for $\boldsymbol{x} \in \mathbb{R}^d$ and $\|\boldsymbol{A}\|$ is the spectral norm for matrices $\boldsymbol{A} \in \mathbb{R}^{d\times d}$. The matrix $\boldsymbol{I}_d \in \mathbb{R}^{d\times d}$ is the identity matrix, $\boldsymbol{e}_i$ is the i-th unit vector with 1 at the i-th entry and 0 else, $\mathbf{0}_d := (0,\dots,0)^\top \in \mathbb{R}^d$ is the zero vector and $\mathbf{1}_d := (1,\dots,1)^\top \in \mathbb{R}^d$ is the vector containing only 1. For a vector $\boldsymbol{x} \in \mathbb{R}^d$, we write $\operatorname{diag}(\boldsymbol{x}) \in \mathbb{R}^{d\times d}$ for a diagonal matrix with the components of $\boldsymbol{x}$ on the diagonal and for $\boldsymbol{A} = (\boldsymbol{a}^{(1)},\dots,\boldsymbol{a}^{(d)}) \in \mathbb{R}^{d\times d}$ the operator $\operatorname{vec}(\boldsymbol{A}) \in \mathbb{R}^{d^2}$ stacks the columns of $\boldsymbol{A}$ in a vector such that $\operatorname{vec}(\boldsymbol{A}) = (\boldsymbol{a}^{(1)^\top},\dots,\boldsymbol{a}^{(d)^\top})^\top$ and $\lambda_i(\boldsymbol{A})$ denotes the i-th largest eigenvalue of $\boldsymbol{A}$. If $\boldsymbol{B} \in \mathbb{R}^{d\times d}$ and $\boldsymbol{A} = \boldsymbol{B}^2 \in \mathbb{R}^{d\times d}$ then $\boldsymbol{A}^{1/2} := \boldsymbol{B}$ denotes the square root of a matrix. A sequence of matrices $\boldsymbol{A}_1, \boldsymbol{A}_2, \dots \in \mathbb{R}^{d\times d}$ with fixed dimension d is denoted by $(\boldsymbol{A}_n)_{n\in\mathbb{N}}$ and if the dimensions $d = d_n$

14679469, 0, Downloaded from https://onlinelibrary.wiley.com/doi/10.1111/sjos.70026 by Karlsruher Institut Für Technologie, Wiley Online Library on [22/10/2025]. See the Terms and Conditions (https://onlinelibrary.wiley.com/terms-and-conditions) on Wiley Online Library for rules of use; OA articles are governed by the applicable Creative Commons License

depends on n, we write $(\boldsymbol{A}^{(n)})_{n\in\mathbb{N}}$, where $\boldsymbol{A}^{(n)} \in \mathbb{R}^{d_n\times d_n}$. For a univariate distribution function F the function $F^{\leftarrow} : (0,1) \to \mathbb{R}$ with $p \mapsto \inf\{x \in \mathbb{R} : F(x) \geq p\}$ is the generalized inverse of F. By $\delta_{\boldsymbol{x}}$, we denote the Dirac measure in $\boldsymbol{x} \in \mathbb{R}^d$. Finally, $\xrightarrow{D}$ is the notation for convergence in distribution, $\xrightarrow{\mathbb{P}}$ is the notation for convergence in probability and $\xrightarrow{\mathbb{P}\text{-a.s.}}$ is the notation for almost sure convergence.

2 | ASYMPTOTIC BEHAVIOR OF THE EMPIRICAL EIGENVALUES OF $\boldsymbol{\Sigma}$

The information criteria AIC and BIC of this paper are defined by the empirical eigenvalues $\widehat{\lambda}_{n,1},\dots,\widehat{\lambda}_{n,d}$ of $\boldsymbol{\Sigma}$. Therefore, in the first step, in Section 2.1, we define and explore the empirical eigenvalues and their asymptotic properties in the fixed-dimensional case, and then, in Section 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 3 and Section 4. The proofs of this section are moved to Appendix A.

2.1 | Fixed-dimensional case

In the case where the dimension d is fixed, we consider the following spiked covariance 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 $\boldsymbol{\Theta}$.

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

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

(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 $n \to \infty$, $n\mathbb{P}(\|\boldsymbol{X}\| > u_n)/k_n \to 1$ and

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

as $n \to \infty$.

The last assumption (S4) is a technical assumption that we require for some proofs (cf. Remark 1). Uniform convergence is required in order to replace the threshold u_n with the order statistic $\|\boldsymbol{X}_{(k_n+1,n)}\|$. Ultimately, this condition is an assumption on the slowly varying function of the tail distribution of $\|\boldsymbol{X}\|$, 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

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

14679469, 0, Downloaded from https://onlinelibrary.wiley.com/doi/10.1111/sjos.70026 by Karlsruher Institut Für Technologie, Wiley Online Library on [22/10/2025]. See the Terms and Conditions (https://onlinelibrary.wiley.com/terms-and-conditions) on Wiley Online Library for rules of use; OA articles are governed by the applicable Creative Commons License

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

$$\widehat{\boldsymbol{\Sigma}}_n := \frac{1}{k_n}\sum_{j=1}^{n}\left(\frac{\boldsymbol{X}_j}{\|\boldsymbol{X}_j\|} - \widehat{\boldsymbol{\Theta}}_n\right)\left(\frac{\boldsymbol{X}_j}{\|\boldsymbol{X}_j\|} - \widehat{\boldsymbol{\Theta}}_n\right)^{\top} \mathbb{1}\{\|\boldsymbol{X}_j\| > \|\boldsymbol{X}_{(k_n+1,n)}\|\} \tag{2.1}$$

with eigenvalues $\widehat{\lambda}_{n,1} \geq \ldots \geq \widehat{\lambda}_{n,d}$ 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 information criteria for the estimation of p^* will be defined by the empirical eigenvalues $\widehat{\lambda}_{n,1},\ldots,\widehat{\lambda}_{n,d}$. 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 1. *Let Model 1 be given. Then as* $n \to \infty$,

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

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

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

Remark 1. In the bivariate case and for $h : \mathbb{R}^2 \mapsto \mathbb{R}$ defined as $h(x,y) = xy$, the asymptotic distribution of

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

was derived in Larsson and Resnick (2012, Theorem 1). The techniques of the proof can be generalized and applied to $\mathrm{vec}(\widehat{\boldsymbol{\Sigma}}_n)$ with the technical assumption (S4), and therefore, the proof of Proposition 1 is omitted. Note that if $\|\boldsymbol{\theta}\| = 1$ for $\boldsymbol{\theta} \in \mathbb{R}^d$ then $\|\mathrm{vec}(\boldsymbol{\theta}\boldsymbol{\theta}^{\top})\| = 1$ 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 1. *Let Model 1 be given.*

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

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

(*b*) *and*

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

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 1 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* $\boldsymbol{\Sigma}$.

14679469, 0, Downloaded from https://onlinelibrary.wiley.com/doi/10.1111/sjos.70026 by Karlsruher Institut Für Technologie, Wiley Online Library on [22/10/2025]. See the Terms and Conditions (https://onlinelibrary.wiley.com/terms-and-conditions) on Wiley Online Library for rules of use; OA articles are governed by the applicable Creative Commons License

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 $\mathbf{\Sigma}^{(n)}$ has a spiked structure. The explicit definition of a directional model is as follows:

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

$$\mathbf{\Gamma}^{(n)} := \begin{pmatrix} \mathbf{\Gamma}_n & \mathbf{0}_{p^* \times d_n} \\ \mathbf{0}_{d_n \times p^*} & \boldsymbol{I}_{d_n - p^*} \end{pmatrix} \in \mathbb{R}^{d_n \times d_n}, \tag{2.2}$$

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

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

$\boldsymbol{V}^{(n)} = (V_1, \ldots, V_{d_n})^\top \in \mathbb{R}^{d_n}$ is a centered random vector consisting of i.i.d. symmetric components with variance 1 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

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

follows the so-called directional model.

Due to construction, we see directly that the directional component

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

of $\boldsymbol{X}^{(n)}$ is independent of the radial component $\|\boldsymbol{X}^{(n)}\| = Z$, and additionally, $\mathbf{\Theta}^{(n)}$ 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 $\mathbf{\Theta}^{(n)}$.

Remark 2.

(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 $\mathrm{Cov}(\mathbf{\Theta}^{(n)})$ where p^*, the location of the smallest eigenvalue bigger than 1, is independent of n and fixed (see Lemma 1 for more details). On the other hand, it implies the independence of the directional and the radial part of $\boldsymbol{X}^{(n)}$, so that the order statistic of an i.i.d. sequence $\|\boldsymbol{X}_1^{(n)}\|, \ldots, \|\boldsymbol{X}_n^{(n)}\|$ is reflected by the order statistic of the i.i.d. sequence of radial parts $Z_1, \ldots, Z_n$.

14679469, 0, Downloaded from https://onlinelibrary.wiley.com/doi/10.1111/sjos.70026 by Karlsruher Institut Für Technologie, Wiley Online Library on [22/10/2025]. See the Terms and Conditions (https://onlinelibrary.wiley.com/terms-and-conditions) on Wiley Online Library for rules of use; OA articles are governed by the applicable Creative Commons License

(b) The directional model is inspired by several models in the nonextreme world. For instance, Bai et al. (2018) employed a model of the form $\mathbf{\Gamma}^{(n)^{1/2}}\boldsymbol{V}^{(n)}$, 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 $\boldsymbol{X}^{(n)} = \mathbf{\Gamma}^{(n)^{1/2}}\boldsymbol{V}^{(n)}/\|\mathbf{\Gamma}^{(n)^{1/2}}\boldsymbol{V}^{(n)}\| \cdot Z$ combines both approaches: The factor $\mathbf{\Gamma}^{(n)^{1/2}}\boldsymbol{V}^{(n)}$ serves as a latent vector for the directional component $\mathbf{\Theta}^{(n)}$, which captures the spiked covariance structure. Normalizing it by $\|\mathbf{\Gamma}^{(n)^{1/2}}\boldsymbol{V}^{(n)}\|$ 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 $\|\boldsymbol{X}_1^{(n)}\|$ by the order statistics of the radial parts.

(c) Although $\mathrm{Cov}(\mathbf{\Theta}^{(n)})$ is a spiked covariance model with p^* leading eigenvalues (see Lemma 1) the support of the distribution of $\mathbf{\Theta}^{(n)}$ might have a higher dimension than p^*. However, if ξ_{n,p^*} is large, then the support of $\mathbf{\Theta}^{(n)}$ is more concentrated on the p^*-dimensional subspace generated by the leading eigenvalues.
One special case for $\mathbf{\Theta}^{(n)}$ 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 $\mathbf{\Theta}^{(n)}$ is given by Tyler (1987) as

$$f_{\mathbf{\Theta}^{(n)}}(\boldsymbol{\theta}|\mathbf{\Gamma}^{(n)}) = \frac{2\pi^{\frac{d_n}{2}}}{\Gamma(\frac{d_n}{2})} \det(\mathbf{\Gamma}^{(n)})^{-1/2}(\boldsymbol{\theta}^\top \mathbf{\Gamma}^{(n)^{-1}}\boldsymbol{\theta})^{-d_n/2}, \quad \boldsymbol{\theta} \in \{\boldsymbol{x} \in \mathbb{R}^{d_n} : \|\boldsymbol{x}\| = 1\}$$

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* (Bai & Silverstein, 2010, p. 5) of $\mathbf{\Gamma}^{(n)}$ is defined as

$$F^{\mathbf{\Gamma}^{(n)}}(x) = \frac{1}{d_n}\sum_{i=1}^{d_n} \mathbb{1}\{\xi_{n,i} \le x\}, \quad x \in \mathbb{R},$$

and the *limiting spectral distribution* (LSD) of $\mathbf{\Gamma}^{(n)}$ is the Dirac measure δ_1, since

$$\lim_{n\to\infty} F^{\mathbf{\Gamma}^{(n)}}(x) = \lim_{n\to\infty} \frac{1}{d_n}\sum_{j=1}^{p^*} \mathbb{1}\{\xi_{n,j} \le x\} + \frac{d_n - p^*}{d_n}\mathbb{1}\{1 \le x\} = \mathbb{1}\{1 \le x\}, \quad x \in \mathbb{R}.$$

In the following, we denote the covariance matrix of $\mathbf{\Theta}^{(n)}$ as

$$\mathbf{\Sigma}^{(n)} := \mathrm{Cov}(\mathbf{\Theta}^{(n)})$$

whereas $\mathbf{\Gamma}^{(n)}$ is the covariance matrix of the nonstandardized directional component $\mathbf{\Gamma}^{(n)^{1/2}}\boldsymbol{V}^{(n)}$. Not only $\mathbf{\Gamma}_n$ has the eigenvalues $\xi_{n,1},\dots,\xi_{n,p^*}$ 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 $\xi_{n,p^*+1},\dots,\xi_{n,d_n}$. We are still in the setup of the last section because not only the eigenvalues of $\mathbf{\Gamma}^{(n)}$ satisfy the spiked covariance

14679469, 0, Downloaded from https://onlinelibrary.wiley.com/doi/10.1111/sjos.70026 by Karlsruher Institut Für Technologie, Wiley Online Library on [22/10/2025]. See the Terms and Conditions (https://onlinelibrary.wiley.com/terms-and-conditions) on Wiley Online Library for rules of use; OA articles are governed by the applicable Creative Commons License

14679469, 0, Downloaded from https://onlinelibrary.wiley.com/doi/10.1111/sjos.70026 by Karlsruher Institut Für Technologie, Wiley Online Library on [22/10/2025]. See the Terms and Conditions (https://onlinelibrary.wiley.com/terms-and-conditions) on Wiley Online Library for rules of use; OA articles are governed by the applicable Creative Commons License

structure

$$\xi_{n,1} \geq \cdots \geq \xi_{n,p^*} > 1 = \xi_{n,p^*+1} = \cdots = \xi_{n,d_n}$$

in (1.1) but as well the eigenvalues of $\boldsymbol{\Sigma}^{(n)}$ satisfy the structure in (1.1) although $\boldsymbol{\Sigma}^{(n)}$ has different eigenvalue than $\boldsymbol{\Gamma}^{(n)}$.

Lemma 1. *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 $\boldsymbol{\Sigma}^{(n)} = \mathrm{Cov}(\boldsymbol{\Theta}^{(n)})$. 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 $\boldsymbol{\Sigma}^{(n)}$ and the eigenvalues λ_{n,p^*+1}, …, λ_{n,d_n-1} are all equal, as required in the definition of the spiked covariance model in (1.1). We summarize the model as follows:

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 $\boldsymbol{\Gamma}^{(n)}$ in (2.2) 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 $\boldsymbol{\Sigma}^{(n)}$ 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 3.

(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 1, Model 1 is also a spiked covariance model but it is a special type of spiked covariance model, 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 $1 + \sqrt{c}$; see the following theorem: 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

$$\varphi_c(x) := x\left(1 + c\int \frac{t}{x-t}\,\mathrm{d}\delta_1(t)\right) = x\left(1 + \frac{c}{x-1}\right). \tag{2.3}$$

Analog to (2.1) we define the $d_n \times d_n$ empirical covariance matrix of $\boldsymbol{\Sigma}^{(n)}$ as

$$\widehat{\boldsymbol{\Sigma}}^{(n)} := \frac{1}{k_n}\sum_{j=1}^{n}\left(\frac{\boldsymbol{X}_j^{(n)}}{\|\boldsymbol{X}_j^{(n)}\|} - \widehat{\boldsymbol{\Theta}}^{(n)}\right)\cdot\left(\frac{\boldsymbol{X}_j^{(n)}}{\|\boldsymbol{X}_j^{(n)}\|} - \widehat{\boldsymbol{\Theta}}^{(n)}\right)^{\top} \mathbb{1}\{\|\boldsymbol{X}_i^{(n)}\| > \|\boldsymbol{X}_{(k_n+1,n)}^{(n)}\|\}, \tag{2.4}$$

with eigenvalues $\widehat{\lambda}_{n,1} \geq \ldots \geq \widehat{\lambda}_{n,d_n}$, where

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

In contrast to the empirical covariance matrix $\widehat{\boldsymbol{\Sigma}}_n$ in (2.1) with a fixed dimension $d \times d$, the dimension of the empirical covariance matrix $\widehat{\boldsymbol{\Sigma}}^{(n)}$ in (2.4) 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 $\widehat{\boldsymbol{\Sigma}}^{(n)}$ under the constraint that $\boldsymbol{\Gamma}_n$ and its eigenvalues $\xi_{n,1},\ldots,\xi_{n,p^*}$ are converging, and afterward when $\xi_{n,p^*} \to \infty$.

Theorem 2. *Let Model 1 be given. Suppose that* $\boldsymbol{\Gamma}_n \to \boldsymbol{\Gamma}$ *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*

$$d_n \widehat{\lambda}_{n,i} \xrightarrow{\mathbb{P}} \varphi_c(\xi_i), \quad \text{as } n \to \infty$$

holds, where φ_c *is defined as in* (2.3).

(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*

$$\sup_{\alpha\in(0,1)} \left| d_n \widehat{\lambda}_{n,i_n(\alpha)} - F_c^{\leftarrow}(1-\alpha) \right| \xrightarrow{\mathbb{P}} 0, \quad \text{as } n \to \infty,$$

where $F_c^{\leftarrow}$ *is the generalized inverse of* F_c *with 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* $c > 1$. *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* $n \to \infty$,

$$\frac{1}{k_n - q_n} \sum_{i=q_n+1}^{k_n} d_n \widehat{\lambda}_{n,i} \xrightarrow{\mathbb{P}} c.$$

14679469, 0, Downloaded from https://onlinelibrary.wiley.com/doi/10.1111/sjos.70026 by Karlsruher Institut Für Technologie, Wiley Online Library on [22/10/2025]. See the Terms and Conditions (https://onlinelibrary.wiley.com/terms-and-conditions) on Wiley Online Library for rules of use; OA articles are governed by the applicable Creative Commons License

Remark 4. The limiting spectral distribution F_c is called Marčenko–Pastur law after the authors of Marčenko and Pastur (1967) and plays an important role in random matrix theory (cf. Bai & Silverstein, 2010). Marčenko and Pastur (1967) first derived for random matrices with i.i.d. components the asymptotic distribution of the eigenvalues of the empirical covariance matrix when the sample size and the dimension tend to infinity, which differs from the classical statistical setting with fixed dimension.

So far, we have assumed that the first p^* eigenvalues $\xi_{n,1},\dots,\xi_{n,p^*}$ of $\mathbf{\Gamma}^{(n)}$ are bounded. Alternatively, it is also possible to suppose that $\xi_{n,p^*} \to \infty$ as $n \to \infty$.

Theorem 3. *Let Model 1 be given. Suppose* $\xi_{n,p^*} \to \infty$ *and* $\xi_{n,1} = o(d_n^{1/2})$ *as* $n \to \infty$.

(a) *Let* $i \in \{1,\dots,p^*\}$. *Then, the asymptotic behavior*

$$d_n \widehat{\lambda}_{n,i}/\xi_{n,i} \xrightarrow{\mathbb{P}} 1, \quad \text{as } n \to \infty$$

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*

$$\sup_{\alpha\in(0,1)} \left| d_n \widehat{\lambda}_{n,i_n(\alpha)} - F_c^{\leftarrow}(1-\alpha) \right| \xrightarrow{\mathbb{P}} 0, \quad \text{as } n \to \infty,$$

where $F_c^{\leftarrow}$ *is defined as in Theorem 2. 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* $0 < c \le 1$ *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* $i \in \{1,\dots,p^*\}$. *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 5. 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 11.

14679469, 0, Downloaded from https://onlinelibrary.wiley.com/doi/10.1111/sjos.70026 by Karlsruher Institut Für Technologie, Wiley Online Library on [22/10/2025]. See the Terms and Conditions (https://onlinelibrary.wiley.com/terms-and-conditions) on Wiley Online Library for rules of use; OA articles are governed by the applicable Creative Commons License

3 | INFORMATION CRITERIA FOR THE NUMBER OF PRINCIPAL COMPONENTS IN THE FIXED-DIMENSIONAL CASE

The aim of the paper 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 AIC and a BIC was developed in Fujikoshi and Sakurai (2016) and the consistency in the high-dimensional case for general data was analyzed in Bai et al. (2018). The AIC (Akaike, 1974) is based on minimizing the Kullback-Leibler divergence between the true distribution and the model, and the BIC (Schwarz, 1978) maximizes the posterior probability. In this paper, we adopt these information criteria and study their statistical properties. We start in this section with the fixed-dimensional case and give the proper definitions of the information criteria under Model 1. The proofs of this section are moved to Appendix B.

Definition 1. Suppose $\widehat{\lambda}_{n,1} \geq \ldots \geq \widehat{\lambda}_{n,d-1}$ are the empirical eigenvalues of $\widehat{\mathbf{\Sigma}}_n$ as defined in (2.1).

(a) The *AIC* for the fixed-dimensional case is defined as

$$\mathrm{AIC}_{k_n}(p) := k_n \sum_{i=1}^{p} \log(\widehat{\lambda}_{n,i}) + k_n(d-1-p)\log\left(\frac{1}{d-1-p}\sum_{j=p+1}^{d-1}\widehat{\lambda}_{n,j}\right) + k_n \log\left(\frac{k_n-1}{k_n}\right)^{d-1} + k_n(d-1)(\log(2\pi)+1) + 2(p+1)(d-p/2),$$

for $p = 1,\ldots,d-2$ and an estimator for p^* is $\widehat{p}_n := \arg\min_{1\leq p\leq d-2} \mathrm{AIC}_{k_n}(p)$.

(b) The *BIC* for the fixed-dimensional case is defined as

$$\mathrm{BIC}_{k_n}(p) := k_n \sum_{i=1}^{p} \log(\widehat{\lambda}_{n,i}) + k_n(d-1-p)\log\left(\frac{1}{d-1-p}\sum_{j=p+1}^{d-1}\widehat{\lambda}_{n,j}\right) + k_n \log\left(\frac{k_n-1}{k_n}\right)^{d-1} + k_n(d-1)(\log(2\pi)+1) + \log(k_n)(p+1)(d-p/2),$$

for $p = 1,\ldots,d-2$ and an estimator for p^* is $\widehat{p}_n := \arg\min_{1\leq p\leq d-2} \mathrm{BIC}_{k_n}(p)$.

Remark 6.

(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 & Sakurai, 2016, Section 2). As baseline model, we take a $(d-1)$-dimensional normal distribution instead of a d-dimensional normal distribution because $\mathbf{\Theta}$ is a random vector on the unit sphere and hence 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 $\widehat{\lambda}_{n,d}$.

14679469, 0, Downloaded from https://onlinelibrary.wiley.com/doi/10.1111/sjos.70026 by Karlsruher Institut Für Technologie, Wiley Online Library on [22/10/2025]. See the Terms and Conditions (https://onlinelibrary.wiley.com/terms-and-conditions) on Wiley Online Library for rules of use; OA articles are governed by the applicable Creative Commons License

(b) The AIC and BIC are invariant to scaling of the eigenvalues. Consequently, scaling the sample covariance matrix $\widehat{\mathbf{\Sigma}}_n$, or equivalently the eigenvalues $\widehat{\lambda}_{n,1},\dots,\widehat{\lambda}_{n,d-1}$, 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 BIC, which estimates the true parameter p^* with a probability converging to 1.

Theorem 4. *Let Model 1 be given. Then*

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

In contrast to the BIC, the AIC is not a weakly consistent information criterion.

Theorem 5. *Let Model 1 be given and $\boldsymbol{M}$ be the limit vector in Theorem 1. 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

$$g_p(\boldsymbol{m}) := -\frac{1}{2}\sum_{i=p^*+1}^{p} m_i^2 - \frac{1}{2(d-1-p)}\left(\sum_{j=p+1}^{d-1} m_j\right)^2 + \frac{1}{2(d-1-p^*)}\left(\sum_{j=p^*+1}^{d-1} m_j\right)^2 - (d-p-2)(d-p+1) + (d-p^*-2)(d-p^*+1)$$

for $\boldsymbol{m} = (m_1,\dots,m_d) \in \mathbb{R}^d$.

Remark 7.

(a) Under some technical assumptions on the distribution of $\mathbf{\Theta}$, 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 this paper, it is sufficient to give an example such that the AIC is not consistent.

(b) Suppose $(\boldsymbol{X}^{(n)})_{n\in\mathbb{N}}$ follows the directional model with $\mathbf{\Gamma}^{(n)} := \mathbf{\Gamma} := \mathrm{diag}(9,4,4,1)$, $\boldsymbol{V}^{(n)} := \boldsymbol{V} := (V_1,V_2,V_3,V_4)^\top$, where $V_i \sim \mathcal{U}(\{-1,1\})$ is uniformly distributed, $i = 1,\dots,4$, $Z \sim$ Fréechet(1) and the dimension $d = 4$ is fixed. Then, we have $\mathbb{P}(g_2(\boldsymbol{M}) < 0) > 0$. The detailed calculations have been moved to Appendix B.

The inconsistency of the AIC and the consistency of the BIC are typical for these information criteria in the fixed-dimensional case (cf. Burnham & Anderson, 1998, Section 2.8.2) and Claeskens (2016, Section 2.2.1). In the high-dimensional case, the asymptotic properties differ.

4 | INFORMATION CRITERIA FOR THE NUMBER OF PRINCIPAL COMPONENTS IN THE HIGH-DIMENSIONAL CASE

The topic in this section is information criteria in the high-dimensional case of Model 1, 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

14679469, 0, Downloaded from https://onlinelibrary.wiley.com/doi/10.1111/sjos.70026 by Karlsruher Institut Für Technologie, Wiley Online Library on [22/10/2025]. See the Terms and Conditions (https://onlinelibrary.wiley.com/terms-and-conditions) on Wiley Online Library for rules of use; OA articles are governed by the applicable Creative Commons License

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 $\widehat{\mathbf{\Sigma}}^{(n)}$ are equal to zero, that is, $\widehat{\lambda}_{n,k_n+1} = \cdots = \widehat{\lambda}_{n,d_n} = 0$. Therefore, in Section 4.1, we analyze the information criteria for $0 < c < 1$ and in Section 4.2 for $c > 1$. The proofs of this section are provided in Appendix C.

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 $1,\dots,d_n - 1$ but rather restrict to $1,\dots,q_n$ with $q_n \leq d_n$. The number q_n is called the number of *candidate dimensions*.

Definition 2. Suppose $\widehat{\lambda}_{n,1} \geq \dots \geq \widehat{\lambda}_{n,d_n-1}$ are the empirical eigenvalues of $\widehat{\mathbf{\Sigma}}^{(n)}$ as defined in (2.4) and let $q_n \leq d_n - 2$.

(a) The *AIC* for the high-dimensional case with $d_n < k_n$ is defined as

$$\mathrm{AIC}^{\circ}_{k_n}(p) := \sum_{i=1}^{p} \log(\widehat{\lambda}_{n,i}) + (d_n - 1 - p)\log\left(\frac{1}{d_n - 1 - p}\sum_{j=p+1}^{d_n-1}\widehat{\lambda}_{n,j}\right) + \log\left(\frac{k_n - 1}{k_n}\right)^{d_n-1} + (d_n - 1)(\log(2\pi) + 1) + \frac{(p+1)(2d_n - p)}{k_n},$$

for $p = 1,\dots,d_n - 2$ and an estimator for p^* is $\widehat{p}_n := \arg\min_{1\leq p\leq q_n} \mathrm{AIC}^{\circ}_{k_n}(p)$.

(b) The *BIC* for the high-dimensional case with $d_n < k_n$ is defined as

$$\mathrm{BIC}^{\circ}_{k_n}(p) := \sum_{i=1}^{p} \log(\widehat{\lambda}_{n,i}) + (d_n - 1 - p)\log\left(\frac{1}{d_n - 1 - p}\sum_{j=p+1}^{d_n-1}\widehat{\lambda}_{n,j}\right) + \log\left(\frac{k_n - 1}{k_n}\right)^{d_n-1} + (d_n - 1)(\log(2\pi) + 1) + \log(k_n)\frac{(p+1)(d_n - p/2)}{k_n},$$

for $p = 1,\dots,d_n - 2$ and an estimator for p^* is $\widehat{p}_n := \arg\min_{1\leq p\leq q_n} \mathrm{BIC}^{\circ}_{k_n}(p)$.

In the next theorem, we present sufficient assumptions for the AIC° to be weakly consistent, that is,

$$\lim_{n\to\infty} \mathbb{P}\left(\arg\min_{1\leq p<q_n} \mathrm{AIC}^{\circ}_{k_n}(p) = p^*\right) = 1$$

and afterward for the BIC°.

Theorem 6. *Let Model 1 with $0 < c < 1$ be given and let the number q_n of candidate dimensions satisfy $q_n = o(d_n)$ as $n \to \infty$.*

14679469, 0, Downloaded from https://onlinelibrary.wiley.com/doi/10.1111/sjos.70026 by Karlsruher Institut Für Technologie, Wiley Online Library on [22/10/2025]. See the Terms and Conditions (https://onlinelibrary.wiley.com/terms-and-conditions) on Wiley Online Library for rules of use; OA articles are governed by the applicable Creative Commons License

(a) *Suppose* $\mathbf{\Gamma}_n \to \mathbf{\Gamma}$ *and* $(\xi_{n,1},\dots,\xi_{n,p^*}) \to (\xi_1,\dots,\xi_{p^*})$ *as* $n \to \infty$ *with* $\xi_{p^*} > 1+\sqrt{c}$. *If the gap condition*

$$\varphi_c(\xi_{p^*}) - 1 - \log(\varphi_c(\xi_{p^*})) - 2c > 0 \tag{4.1}$$

with φ_c *as defined in* (2.3) *holds, then the* AIC° *is weakly consistent*.

(b) *Suppose* $\mathbf{\Gamma}_n \to \mathbf{\Gamma}$ *and* $(\xi_{n,1},\dots,\xi_{n,p^*}) \to (\xi_1,\dots,\xi_{p^*})$ *as* $n \to \infty$ *with* $\xi_{p^*} > 1+\sqrt{c}$. *If the gap condition* (4.1) *does not hold, then*

$$\lim_{n\to\infty} \mathbb{P}\left(\min_{1\le p<p^*} \left\{ \mathrm{AIC}^{\circ}_{k_n}(p) - \mathrm{AIC}^{\circ}_{k_n}(p^*) \right\} > 0 \right) < 1$$

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 8.

(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 (4.1) was introduced in Bai et al. (2018), and it also guarantees that the gap between ξ_{p^*} and the nonleading 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 7. *Let Model 1 with* $0 < c < 1$ *be given. Suppose that either*

$$\mathbf{\Gamma}_n \to \mathbf{\Gamma} \text{ such that } (\xi_{n,1},\dots,\xi_{n,p^*}) \to (\xi_1,\dots,\xi_{p^*}) \text{ as } n \to \infty \text{ with } \xi_{p^*} > 1+\sqrt{c},$$

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*

$$\lim_{n\to\infty} \mathbb{P}\left(\min_{1\le p<p^*} \left\{ \mathrm{BIC}^{\circ}_{k_n}(p) - \mathrm{BIC}^{\circ}_{k_n}(p^*) \right\} > 0 \right) < 1$$

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*.

14679469, 0, Downloaded from https://onlinelibrary.wiley.com/doi/10.1111/sjos.70026 by Karlsruher Institut Für Technologie, Wiley Online Library on [22/10/2025]. See the Terms and Conditions (https://onlinelibrary.wiley.com/terms-and-conditions) on Wiley Online Library for rules of use; OA articles are governed by the applicable Creative Commons License

Remark 9.

(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. Moreover, in Theorem 6 (b), we have

$$\lim_{n\to\infty} \mathbb{P}\left(\min_{1\le p<p^*} \left\{ \mathrm{AIC}^{\circ}_{k_n}(p) - \mathrm{AIC}^{\circ}_{k_n}(p^*) \right\} > 0 \right) < 1,$$

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 (see Bai et al., 2018, Section 4). While Theorem 2 and Theorem 3 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°.

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 3. Suppose $\widehat{\lambda}_{n,1} \ge \ldots \ge \widehat{\lambda}_{n,d_n-1}$ are the empirical eigenvalues of $\widehat{\mathbf{\Sigma}}^{(n)}$ as defined in (2.4) and let $q_n \le k_n - 2$.

(a) The *AIC* for the high-dimensional case with $d_n > k_n$ is defined as

$$\begin{aligned}\mathrm{AIC}^{*}_{k_n}(p) := &\sum_{i=1}^{p} \log(\widehat{\lambda}_{n,i}) + (k_n - 1 - p) \log\left(\frac{1}{k_n - 1 - p} \sum_{j=p+1}^{k_n-1} \widehat{\lambda}_{n,j} \right) \\ &+ \log\left(\frac{d_n - 1}{d_n} \right)^{k_n-1} + (k_n - 1)(\log(2\pi) + 1) + \frac{(p+1)(2k_n - p)}{d_n},\end{aligned}$$

for $p = 1, \ldots, k_n - 2$ and an estimator for p^* is $\widehat{p}_n := \arg\min_{1\le p\le q_n} \mathrm{AIC}^{*}_{k_n}(p)$.

14679469, 0, Downloaded from https://onlinelibrary.wiley.com/doi/10.1111/sjos.70026 by Karlsruher Institut Für Technologie, Wiley Online Library on [22/10/2025]. See the Terms and Conditions (https://onlinelibrary.wiley.com/terms-and-conditions) on Wiley Online Library for rules of use; OA articles are governed by the applicable Creative Commons License

(b) The *BIC* for the high-dimensional case with $d_n > k_n$ is defined as

$$\begin{aligned}\mathrm{BIC}^*_{k_n}(p) := &\sum_{i=1}^{p} \log(\widehat{\lambda}_{n,i}) + (k_n - 1 - p)\log\left(\frac{1}{k_n - 1 - p}\sum_{j=p+1}^{k_n-1}\widehat{\lambda}_{n,j}\right)\\ &+ \log\left(\frac{d_n - 1}{d_n}\right)^{k_n-1} + (k_n - 1)(\log(2\pi) + 1)\\ &+ \log(d_n)\frac{(p+1)(k_n - p/2)}{d_n},\end{aligned}$$

for $p = 1,\ldots,k_n - 2$ and an estimator for p^* is $\widehat{p}_n := \arg\min_{1\le p\le q_n} \mathrm{BIC}^*_{k_n}(p)$.

For the consistency analysis of the AIC* and BIC*, we use the same definition for weakly consistent as for the AIC° in Section 4.1.

Theorem 8. *Let Model 1 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*

$$\frac{\varphi_c(\xi_{n,p^*})}{c} - 1 - \log\left(\frac{\varphi_c(\xi_{n,p^*})}{c}\right) - \frac{2}{c} > 0 \tag{4.2}$$

with φ_c as defined in (2.3) 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 (4.2) 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 9. *Let Model 1 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 10.

(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 $0 < c < 1$.

14679469, 0, Downloaded from https://onlinelibrary.wiley.com/doi/10.1111/sjos.70026 by Karlsruher Institut Für Technologie, Wiley Online Library on [22/10/2025]. See the Terms and Conditions (https://onlinelibrary.wiley.com/terms-and-conditions) on Wiley Online Library for rules of use; OA articles are governed by the applicable Creative Commons License

(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.

5 | SIMULATION STUDY

In this section, we compare the performance of the different information criteria through a simulation study. In the following, we simulate n times a multivariate regularly varying random vector $\boldsymbol{X}$ of dimension d_n. For the distribution of $\boldsymbol{X}$, we distinguish three models. First, in Section 5.1, we use the directional model and in Section 5.2, we extend the directional model by adding an additional noise term. Finally, the model in Section 5.3 exhibits asymptotic dependence but differs from 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 larger than n, we set $k_n = n$. The code for the simulations is available at https://gitlab.kit.edu/projects/178647.

5.1 | Directional model

First, we consider the directional model (1) with $p^* = 10$ as introduced in Section 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, 10,000. The matrix $\mathbf{\Gamma}_n$ from (2.2) is a fixed diagonal matrix and the 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 1. 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 becomes smaller 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.

For the high-dimensional case $d \geq 100$, Figure 2 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

14679469, 0, Downloaded from https://onlinelibrary.wiley.com/doi/10.1111/sjos.70026 by Karlsruher Institut Für Technologie, Wiley Online Library on [22/10/2025]. See the Terms and Conditions (https://onlinelibrary.wiley.com/terms-and-conditions) on Wiley Online Library for rules of use; OA articles are governed by the applicable Creative Commons License

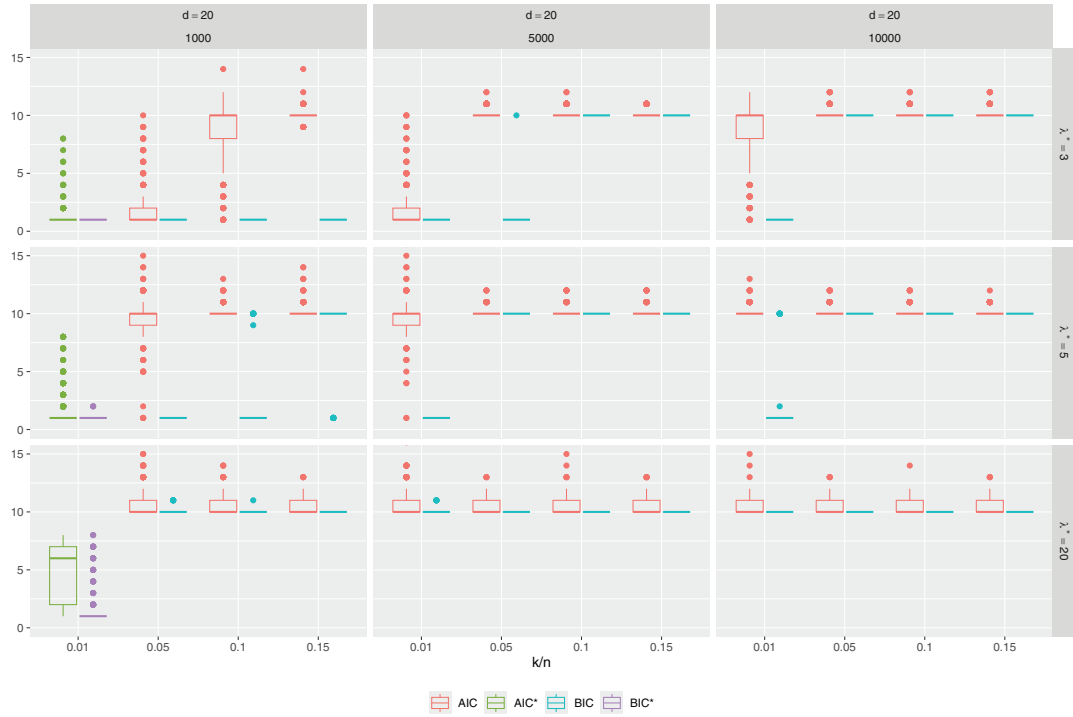

FIGURE 1 Simulations for the directional model with $p^* = 10$ and dimension $d = 20$: From left to the right the sample size increases from $n = 1000$, $n = 5000$ to $n = 10,000$. From top to bottom, the value of the leading eigenvalues increases from $\lambda^* = 3$, $\lambda^* = 5$ to $\lambda^* = 20$. 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.

comparison to Figure 1, we see that the AIC* has the tendency to underestimate p^* for $\lambda^* \leq 5$, $c = 2$ and $c = 3$, which is consistent with Theorem 8. 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 | Directional model with noise

In this example, we consider again the directional model (1) with the same choice of distributions as in Section 5.1, but additionally, we add noise. As noise, we use the d-dimensional random vector

$$\boldsymbol{\varepsilon} \sim \left| \mathcal{N}_d\left(\mathbf{0}_d, \frac{100}{d}\boldsymbol{I}_d\right)\right|,$$

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 5). Then, we construct the regularly varying random vector

14679469, 0, Downloaded from https://onlinelibrary.wiley.com/doi/10.1111/sjos.70026 by Karlsruher Institut Für Technologie, Wiley Online Library on [22/10/2025]. See the Terms and Conditions (https://onlinelibrary.wiley.com/terms-and-conditions) on Wiley Online Library for rules of use; OA articles are governed by the applicable Creative Commons License

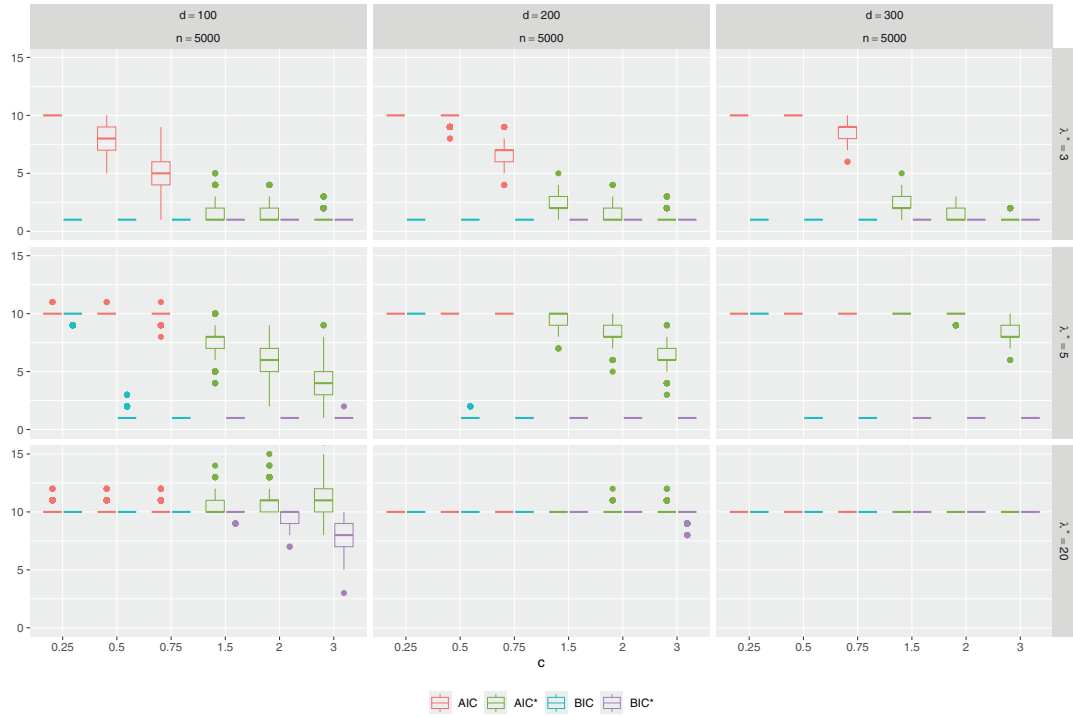

FIGURE 2 Simulations for the directional model with $p^* = 10$ and sample size $n = 5000$: From left to the right the dimension increases from $d = 100$, $d = 200$ to $d = 300$. From top to bottom, the value of the leading eigenvalues increases from $\lambda^* = 3$, $\lambda^* = 5$ to $\lambda^* = 20$. 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.

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

where $\boldsymbol{\Gamma}^{(n)}$, $\boldsymbol{V}^{(n)}$ and Z are defined as in Section 5.1 and $\boldsymbol{\varepsilon}$ is given as above.

The results are shown for $d = 20$ in Figure 3. Overall, the results are similar to Figure 1, but with more deviation from the true value p^*. In most cases (e.g., $n = 5000, 10,000$, $k_n/n \geq 0.05$ and $\lambda^* \geq 5$), the information criteria estimated $\widehat{p}_n = 11$ leading eigenvalues, therefore identifying not only the 10 leading eigenvalues but also 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 4 provides a visualization of the results in the high-dimensional cases $d = 100$, 200, and 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 2. The information criteria estimate the noise as an additional direction, for example, when $\lambda^* = 20$ and $d = 300$.

14679469, 0, Downloaded from https://onlinelibrary.wiley.com/doi/10.1111/sjos.70026 by Karlsruher Institut Für Technologie, Wiley Online Library on [22/10/2025]. See the Terms and Conditions (https://onlinelibrary.wiley.com/terms-and-conditions) on Wiley Online Library for rules of use; OA articles are governed by the applicable Creative Commons License

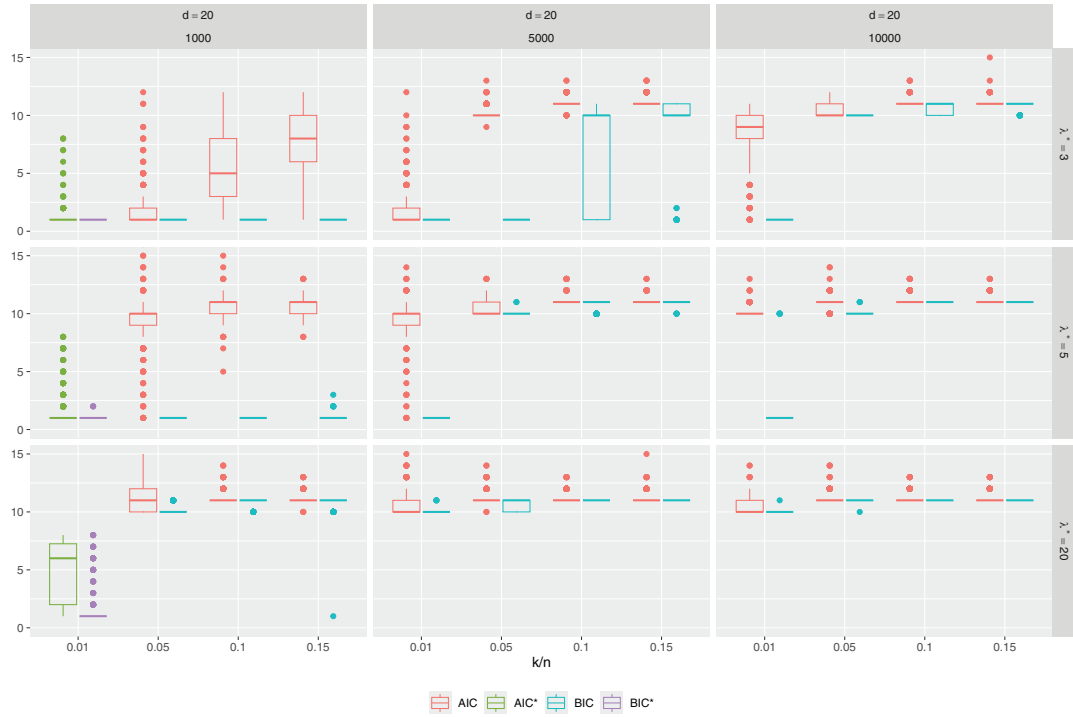

FIGURE 3 Simulations for the noisy directional model with $p^* = 10$ and $d = 20$: From left to the right, the sample size increases from $n = 1000$, $n = 5000$ to $n = 10,000$. From top to bottom, the value of the leading eigenvalues increases from $\lambda^* = 3$, $\lambda^* = 5$ to $\lambda^* = 20$. 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.

5.3 | 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

$$\boldsymbol{H} := \sum_{i=1}^{p^*} \lambda_i \boldsymbol{v}_i \boldsymbol{v}_i^\top + \lambda \boldsymbol{I}_d,$$

where $\boldsymbol{v}_i$, $i = 1,\dots,p^*$ 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 5.1, since 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)

$$\mathbb{P}(\boldsymbol{\Theta} \in \cdot) = \frac{\mathbb{E}[\|\boldsymbol{N}\| \delta_{\boldsymbol{N}/\|\boldsymbol{N}\|}(\cdot)]}{\mathbb{E}[\|\boldsymbol{N}\|]}.$$

14679469, 0, Downloaded from https://onlinelibrary.wiley.com/doi/10.1111/sjos.70026 by Karlsruher Institut Für Technologie, Wiley Online Library on [22/10/2025]. See the Terms and Conditions (https://onlinelibrary.wiley.com/terms-and-conditions) on Wiley Online Library for rules of use; OA articles are governed by the applicable Creative Commons License

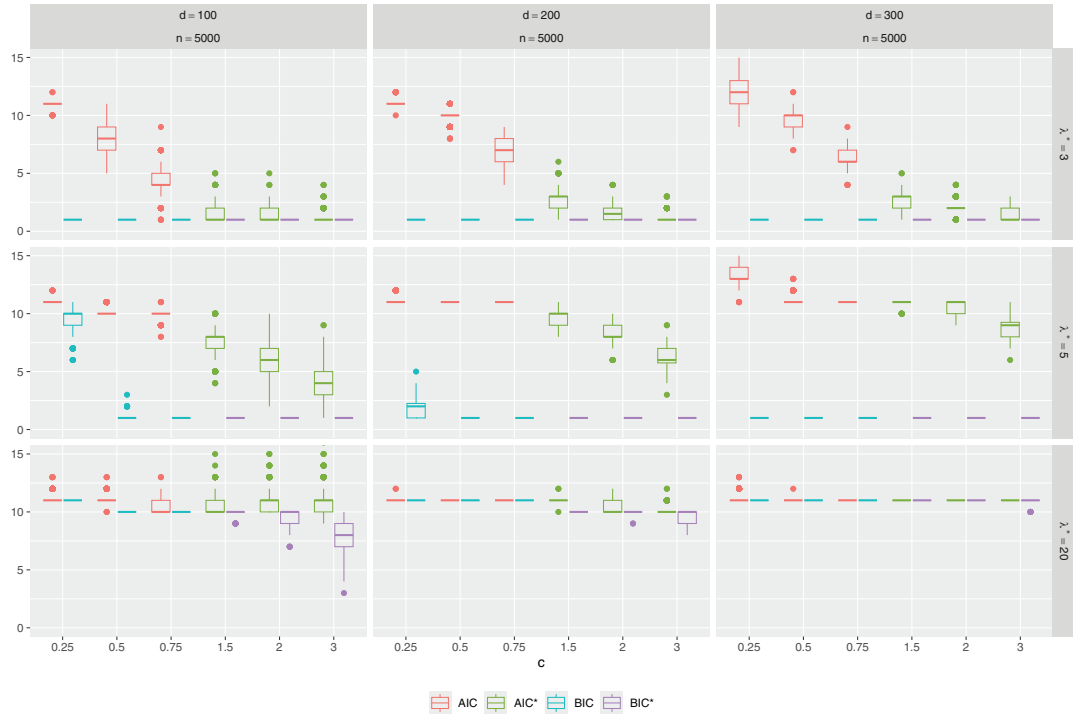

FIGURE 4 Simulations for the noisy directional model with $p^* = 10$ and sample size $n = 5000$: From left to the right, the dimension increases from $d = 100$, $d = 200$ to $d = 300$. From top to bottom, the value of the leading eigenvalues increases from $\lambda^* = 3$, $\lambda^* = 5$ to $\lambda^* = 20$. 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 $\hat{p}_n$ for $p^* = 10$ for the different information criteria.

For the comparison, we run simulations with sample size $n = 5000$ and dimension $d = 100,\ 300$ to 900. The matrix $\boldsymbol{H}$ is fixed for each sample but is initially randomly generated for the simulation, where 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 gets closer to 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 eigenvectors $\boldsymbol{v}_i$ are generated with the R package pracma.

The results are illustrated in Figure 5. It is evident that, when the gap is sufficiently 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.

6 | APPLICATION TO PRECIPITATION DATA

In this section, the information criteria are applied to precipitation data in Germany taken from DWD Climate Data Center (CDC) (1951–2022). The data set 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 6. The data is preprocessed to include only observations from January, February and March, and transformed to standard

14679469, 0, Downloaded from https://onlinelibrary.wiley.com/doi/10.1111/sjos.70026 by Karlsruher Institut Für Technologie, Wiley Online Library on [22/10/2025]. See the Terms and Conditions (https://onlinelibrary.wiley.com/terms-and-conditions) on Wiley Online Library for rules of use; OA articles are governed by the applicable Creative Commons License

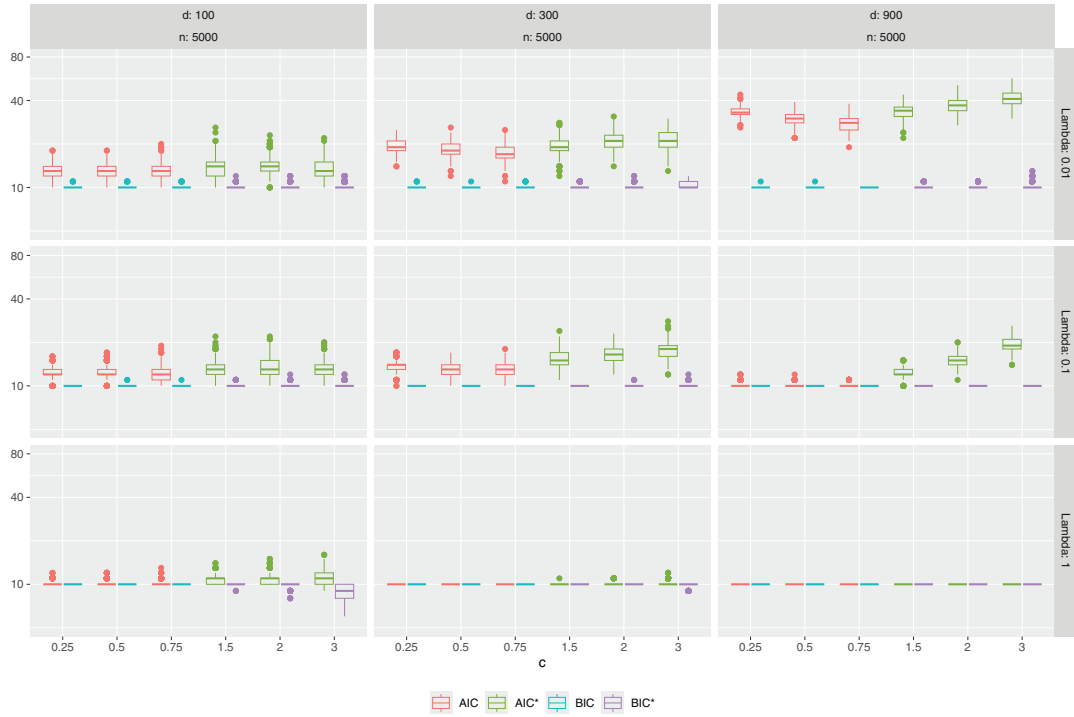

FIGURE 5 Simulations for the spiked angular Gaussian model with $p^* = 10$: From left to the right the dimension increases from $d = 100$, $d = 300$ to $d = 900$. From top to bottom, the value of λ increases from $\lambda = 0.01$, $\lambda = 0.1$ to $\lambda = 1$. In every subplot the ratio $c = d/k_n$ 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.

Fréchet margins. After data cleaning, the resultant dataset contains $n = 2546$ observations, each with precipitation records from at least one station. In Figure 6 we see the stations of the empirical eigenvectors $\widehat{\mathbf{v}}_i = (v_1^{(i)}, \ldots, v_d^{(i)})^\top$, where $v_j^{(i)} \geq 0.6 v_{(1)}^{(i)}$, $i = 1, \ldots, 5$, of the 5 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 the AIC* and BIC* from Definition 3. The number of candidate models q_n for the AIC* is chosen as $d/2 = 250$ to account for the assumption of Theorem 8.

Figure 7 shows the number of estimated leading 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 increasing further. In contrast, the 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 the AIC* continues to increase. This difference between the estimates aligns with the heavier penalty imposed by the BIC*, which leads to smaller estimates compared to the 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 8. At first view, they seem not to be constant after some point, contradicting the spiked covariance assumption. But in a spiked covariance model with

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

14679469, 0, Downloaded from https://onlinelibrary.wiley.com/doi/10.1111/sjos.70026 by Karlsruher Institut Für Technologie, Wiley Online Library on [22/10/2025]. See the Terms and Conditions (https://onlinelibrary.wiley.com/terms-and-conditions) on Wiley Online Library for rules of use; OA articles are governed by the applicable Creative Commons License

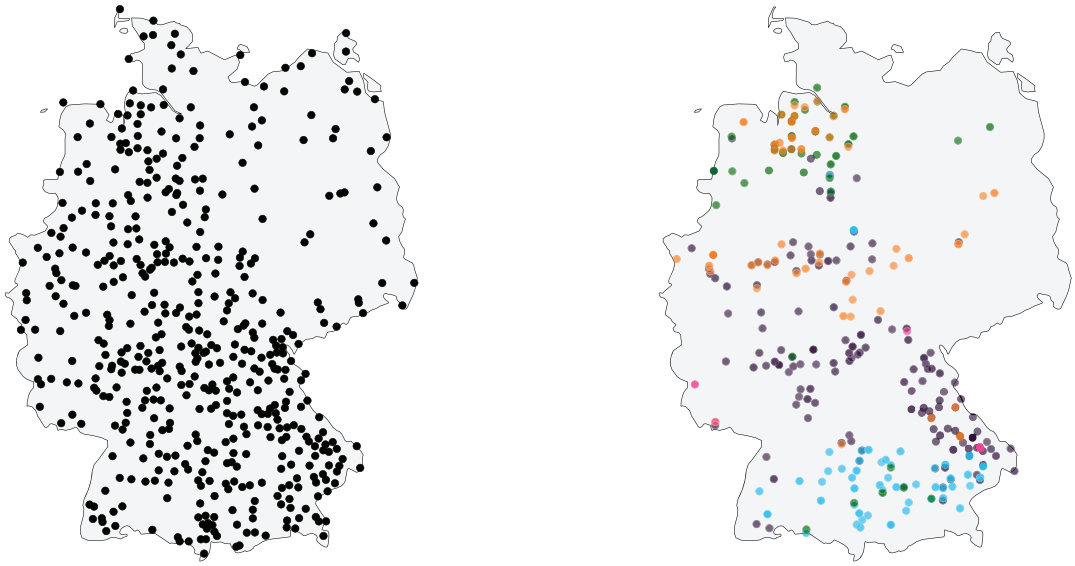

FIGURE 6 Left figure: Map of Germany with all stations highlighted by black dots. Right figure: Map of Germany with the most extreme stations of the empirical eigenvectors of the five largest empirical eigenvalues, colored by eigenvectors.

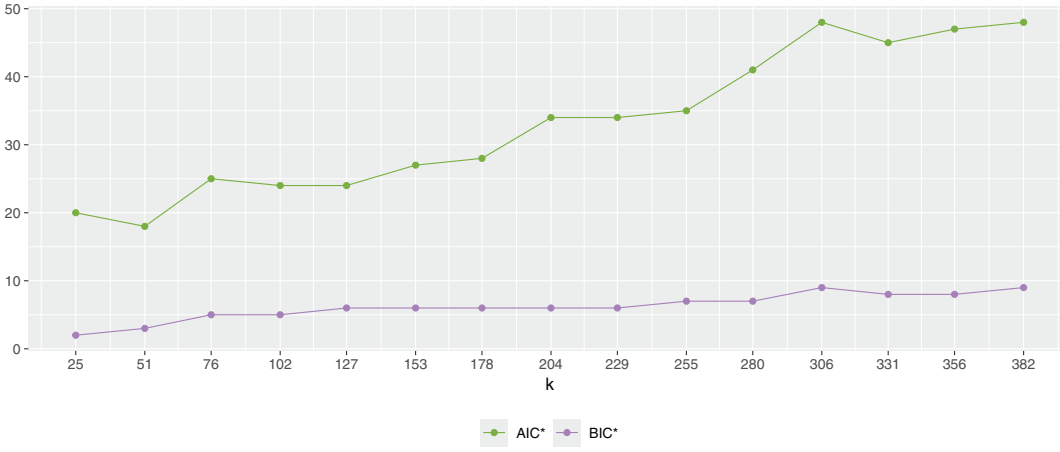

FIGURE 7 The estimated number $\widehat{p}_n$ of leading eigenvalues determined by AIC* and BIC* plotted against k_n.

we have

$$\frac{\lambda_i - \lambda_{i+1}}{\lambda_1} > 0 \quad \text{for } i = 1,\dots,p^* \qquad \text{and} \qquad \frac{\lambda_i - \lambda_{i+1}}{\lambda_1} = 0 \quad \text{for } i = p^*+1,\dots,d.$$

Therefore, the scaled increments $(\widehat{\lambda}_{n,i} - \widehat{\lambda}_{n,i+1})/\widehat{\lambda}_{n,1}$ $i = 1,\dots,75$, of the empirical eigenvalues are plotted in the right picture of Figure 8. We realize that after some point, these increments are nearly constant zero, indicating that these are nonspiked eigenvalues. The information criteria

14679469, 0, Downloaded from https://onlinelibrary.wiley.com/doi/10.1111/sjos.70026 by Karlsruher Institut Für Technologie, Wiley Online Library on [22/10/2025]. See the Terms and Conditions (https://onlinelibrary.wiley.com/terms-and-conditions) on Wiley Online Library for rules of use; OA articles are governed by the applicable Creative Commons License

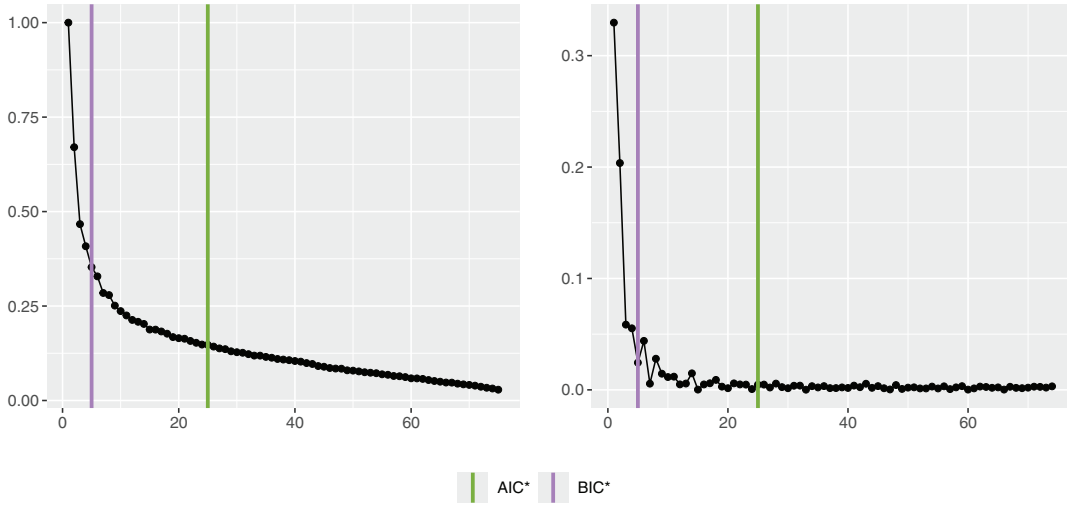

FIGURE 8 For $k_n = 76$, on the left hand side the scaled ordered empirical eigenvalues $\widehat{\lambda}_{n,i}/\widehat{\lambda}_{n,1}$, $i = 1,\ldots,75$ and on the right-hand side the differences of the ordered empirical eigenvalues divided by the value of the largest eigenvalue $(\widehat{\lambda}_{n,i} - \widehat{\lambda}_{n,i+1})/\widehat{\lambda}_{n,1}$, $i = 1,\ldots,75$ are plotted. The vertical lines indicate the AIC* estimator $\widehat{p}_n = 25$ and the BIC* estimator $\widehat{p}_n = 5$.

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 nonextreme 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 8.

7 | CONCLUSION

The paper 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 angular measure. 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 and sufficient criteria for the AIC and the BIC to be consistent in the high-dimensional setting of a directional model. The results of this paper are in accordance with the results in the nonextreme world. For the proofs we derived some new results on the asymptotic properties of the empirical eigenvalues of $\mathbf{\Sigma}$ in both the large-dimensional case, but, in particular, in the high-dimensional case using methods from random matrix theory. The performance of the information criteria was further validated through a simulation study and a real-world example.

The case $c = 0$ is not covered in this paper because we suspect that the AIC and the BIC, as defined here, are inconsistent even when a type of *gap condition* like (4.1) is satisfied. Such a condition relates to the distance between the smallest eigenvalue bigger than 1, denoted by ξ_{n,p^*}, and the preceding eigenvalue, denoted by ξ_{n,p^*+1}. From a practical point of view, we also believe

14679469, 0, Downloaded from https://onlinelibrary.wiley.com/doi/10.1111/sjos.70026 by Karlsruher Institut Für Technologie, Wiley Online Library on [22/10/2025]. See the Terms and Conditions (https://onlinelibrary.wiley.com/terms-and-conditions) on Wiley Online Library for rules of use; OA articles are governed by the applicable Creative Commons License

that in the context of multivariate extremes, the case $c = 0$ is not realistic because, usually, k_n, the number of extremes, is small, and therefore, d_n/k_n will be large.

The paper focused on eigenvalues of $\mathbf{\Sigma}$ that satisfy the spiked covariance structure in (1.1), where 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 8, where the empirical eigenvalues decrease but do not stabilize at some point. Therefore, it is worth exploring more general eigenvalue structures of the covariance matrix to estimate the number of significant components of $\mathbf{\Sigma}$, such as, for example, $\xi_{p^*} > 1 + \sqrt{c}$ 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.

Additionally, as a starting point of 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.

Finally, we would like to point out that changing k_n changes not only the AIC and the BIC estimators $\widehat{p}_n$, but also the empirical eigenvalues and hence, the scree plot as in Figure 8. Therefore, the optimal choice of k_n is nontrivial in this context, and some further research, as discussed in Butsch and Fasen-Hartmann (2025) and Meyer and Wintenberger (2023), for the choice of k_n is needed.

CONFLICT OF INTEREST STATEMENT

None of the authors have a conflict of interest to disclose.

DATA AVAILABILITY STATEMENT

The data that support the findings of this study are openly available in gitlab at https://gitlab.kit.edu/projects/178647.

ACKNOWLEDGMENT

Open Access funding enabled and organized by Projekt DEAL.

ORCID

Vicky Fasen-Hartmann https://orcid.org/0000-0002-5758-1999

REFERENCES

Akaike, H. (1974). A new look at the statistical model identification. *IEEE Transactions on Automatic Control*, *AC-19*, 716–723.

Anderson, T. W. (2003). *An introduction to multivariate statistical analysis* (3rd ed.). Wiley-Interscience.

Avella-Medina, M., Davis, R. A., & Samorodnitsky, G. (2025). Insights into kernel PCA with application to multivariate extremes. *SIAM Journal on Mathematics of Data Science*, *7*(2), 777–801.

Bai, Z., Choi, K. P., & Fujikoshi, Y. (2018). Consistency of AIC and BIC in estimating the number of significant components in high-dimensional principal component analysis. *Annals of Statistics*, *46*(3), 1050–1076.

Bai, Z., Fujikoshi, Y., & 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., & Silverstein, J. W. (2010). *Spectral analysis of large dimensional random matrices*. Springer.

Bai, Z., & Yao, J. (2012). On sample eigenvalues in a generalized spiked population model. *Journal of Multivariate Analysis*, *106*, 167–177.

Bai, Z., & Yin, Y. Q. (1993). Limit of the smallest eigenvalue of a large-dimensional sample covariance matrix. *The Annals of Probability*, *21*(3), 1275–1294.

14679469, 0, Downloaded from https://onlinelibrary.wiley.com/doi/10.1111/sjos.70026 by Karlsruher Institut Für Technologie, Wiley Online Library on [22/10/2025]. See the Terms and Conditions (https://onlinelibrary.wiley.com/terms-and-conditions) on Wiley Online Library for rules of use; OA articles are governed by the applicable Creative Commons License

Burnham, K. P., & Anderson, D. R. (1998). *Model selection and inference: A practical information theoretic approach*. Springer.

Butsch, L., & Fasen-Hartmann, V. (2025). Information criteria for the number of directions of extremes in high-dimensional data. arxiv: 2409.10174.

Chautru, E. (2015). Dimension reduction in multivariate extreme value analysis. *Electronic Journal of Statistics*, *9*(1), 383–418.

Claeskens, G. (2016). Statistical model choice. *Annual Review of Statistics and Its Application*, *3*, 233–256.

Clémençon, S., Huet, N., & Sabourin, A. (2024). Regular variation in Hilbert spaces and principal component analysis for functional extremes. *Stochastic Processes and their Applications*, *174*, 104375.

Cooley, D., & Thibaud, E. (2019). Decompositions of dependence for high-dimensional extremes. *Biometrika*, *106*(3), 587–604.

Dauxois, J., Pousse, A., & Romain, Y. (1982). Asymptotic theory for the principal component analysis of a vector random function: Some applications to statistical inference. *Journal of Multivariate Analysis*, *12*(1), 136–154.

Davis, A. W. (1977). Asymptotic theory for principal component analysis: Non-normal case. *The Australian Journal of Statistics*, *19*(3), 206–212.

de Haan, L., & Ferreira, A. (2006). *Extreme value theory: An introduction*. Springer.

Drees, H. (2025). Asymptotic behavior of principal component projections for multivariate extremes. arxiv: 2503.22296.

Drees, H., & Sabourin, A. (2021). Principal component analysis for multivariate extremes. *Electronic Journal of Statistics*, *15*(1), 908–943.

DWD-Climate-Data-Center-(CDC). (2022). Daily station observations precipitation height in mm for Germany, version v21.3, last accessed: May 03, 2023.

Elezović, N., Giordano, C., & Pecarić, J. (2000). The best bounds in Gautschi's inequality. *Mathematical Inequalities & Applications*, *3*(2), 239–252.

Engelke, S., & Ivanovs, J. (2021). Sparse structures for multivariate extremes. *Annual Review of Statistics and Its Application*, *8*, 241–270.

Falk, M. (2019). *Multivariate extreme value theory and D-norms*. Springer.

Fujikoshi, Y., & Sakurai, T. (2016). Some properties of estimation criteria for dimensionality in principal component analysis. *AJMMS*, *35*(2), 133–142.

Hogg, R. V., McKean, J. W., & Craig, A. T. (2005). *Introduction to mathematical statistics* (6th ed.). Pearson Prentice Hall.

Horn, R. A., & Johnson, C. R. (2013). *Matrix analysis* (2nd ed.). Cambridge University Press.

Hung, H., Huang, S.-Y., & Ing, C.-K. (2022). A generalized information criterion for high-dimensional PCA rank selection. *Statistical Papers*, *63*(4), 1295–1321.

Jiang, Q., Qiu, J., & Li, Z. (2023). On eigenvalues of sample covariance matrices based on high dimensional compositional data. arXiv:2312.14420.

Johnstone, I. M. (2001). On the distribution of the largest eigenvalue in principal components analysis. *The Annals of Probability*, *29*(2), 295–327.

Johnstone, I. M., & Yang, J. (2018). Notes on asymptotics of sample eigenstructure for spiked covariance models with non-Gaussian data. arXiv: 1810.10427.

Larsson, M., & Resnick, S. I. (2012). Extremal dependence measure and extremogram: The regularly varying case. *Extremes*, *15*(2), 231–256.

Marčenko, V. A., & Pastur, L. A. (1967). Distribution of eigenvalues for some sets of random matrices. *Matematicheskii Sbornik*, *1*(4), 457.

Meyer, N., & Wintenberger, O. (2023). Multivariate sparse clustering for extremes. *Journal of the American Statistical Association*, 1–12.

Muirhead, R. J. (1982). *Aspects of multivariate statistical theory*. John Wiley & Sons, Inc.

Paul, D. (2007). Asymptotics of sample eigenstructure for a large dimensional spiked covariance model. *Statistica Sinica*, *17*(4), 1617–1642.

Resnick, S. (1987). *Extreme values, regular variation, and point processes*. Springer.

Resnick, S. (2007). *Heavy-tail phenomena: Probabilistic and statistical modeling*. Springer.

Rohrbeck, C., & Cooley, D. (2023). Simulating flood event sets using extremal principal components. *The Annals of Applied Statistics*, *17*(2), 1333–1352.

14679469, 0, Downloaded from https://onlinelibrary.wiley.com/doi/10.1111/sjos.70026 by Karlsruher Institut Für Technologie, Wiley Online Library on [22/10/2025]. See the Terms and Conditions (https://onlinelibrary.wiley.com/terms-and-conditions) on Wiley Online Library for rules of use; OA articles are governed by the applicable Creative Commons License

Schwarz, G. (1978). Estimating the dimension of a model. *The Annals of Applied Statistics*, *6*(2), 461–464.
Silverstein, J. W. (1995). Strong convergence of the empirical distribution of eigenvalues of large-dimensional random matrices. *Journal of Multivariate Analysis*, *55*(2), 331–339.
Silverstein, J. W., & Choi, S.-I. (1995). Analysis of the limiting spectral distribution of large-dimensional random matrices. *Journal of Multivariate Analysis*, *54*(2), 295–309.
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. *Journal of Inequalities in Pure and Applied Mathematics*, *9*(2), 2.
Wan, P. (2024). Characterizing extremal dependence on a hyperplane. arxiv:2411.00573.
Yin, Y. Q., Bai, Z. D., & Krishnaiah, P. R. (1988). On the limit of the largest eigenvalue of the large-dimensional sample covariance matrix. *Probability Theory and Related Fields*, *78*(4), 509–521.

How to cite this article: Butsch, L., & Fasen-Hartmann, V. (2025). Estimation of the number of principal components in high-dimensional multivariate extremes. *Scandinavian Journal of Statistics*, 1–44. https://doi.org/10.1111/sjos.70026

APPENDIX A. PROOFS OF SECTION 2

Proof of Lemma 1. Note that

$$\boldsymbol{\Sigma}^{(n)} = \mathrm{Cov}(\boldsymbol{\Theta}^{(n)}) = \boldsymbol{\Gamma}^{(n)1/2}\mathrm{Cov}\left(\frac{\boldsymbol{V}^{(n)}}{\|\boldsymbol{\Gamma}^{(n)1/2}\boldsymbol{V}^{(n)}\|}\right)\boldsymbol{\Gamma}^{(n)1/2}.$$

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

$$\boldsymbol{D}^{(n)} := \mathrm{diag}(\overline{\boldsymbol{D}}_n, \boldsymbol{I}_{d_n-p^*}) := \mathrm{diag}(\xi_{n,1},\ldots,\xi_{n,p^*},1,\ldots,1) \in \mathbb{R}^{d_n\times d_n}$$

is a diagonal matrix consisting of the eigenvalues of $\boldsymbol{\Gamma}^{(n)}$, we receive with $\|\boldsymbol{W}^{(n)}\boldsymbol{x}\| = \|\boldsymbol{x}\|$ for $\boldsymbol{x} \in \mathbb{R}^{d_n}$ that

$$\boldsymbol{\Sigma}^{(n)} = \mathrm{Cov}\left(\frac{\boldsymbol{\Gamma}^{(n)1/2}\boldsymbol{V}^{(n)}}{\|\boldsymbol{\Gamma}^{(n)1/2}\boldsymbol{V}^{(n)}\|}\right) = \boldsymbol{W}^{(n)}\mathrm{Cov}\left(\frac{\boldsymbol{D}^{(n)1/2}\boldsymbol{V}^{(n)}}{\|\boldsymbol{D}^{(n)1/2}\boldsymbol{V}^{(n)}\|}\right)\boldsymbol{W}^{(n)\top}.$$

Hence, the matrices

$$\mathrm{Cov}\left(\frac{\boldsymbol{\Gamma}^{(n)1/2}\boldsymbol{V}^{(n)}}{\|\boldsymbol{\Gamma}^{(n)1/2}\boldsymbol{V}^{(n)}\|}\right) \quad \text{and} \quad \mathrm{Cov}\left(\frac{\boldsymbol{D}^{(n)1/2}\boldsymbol{V}^{(n)}}{\|\boldsymbol{D}^{(n)1/2}\boldsymbol{V}^{(n)}\|}\right)$$

are similar and share the same eigenvalues (Horn & Johnson, 2013, Theorem 1.3.22). Therefore, we assume in the following w.l.o.g. that $\boldsymbol{\Gamma}^{(n)} = \boldsymbol{D}^{(n)}$ and hence,

$$\boldsymbol{\Sigma}^{(n)} = \mathrm{Cov}\left(\frac{\boldsymbol{D}^{(n)1/2}\boldsymbol{V}^{(n)}}{\|\boldsymbol{D}^{(n)1/2}\boldsymbol{V}^{(n)}\|}\right)$$

14679469, 0, Downloaded from https://onlinelibrary.wiley.com/doi/10.1111/sjos.70026 by Karlsruher Institut Für Technologie, Wiley Online Library on [22/10/2025]. See the Terms and Conditions (https://onlinelibrary.wiley.com/terms-and-conditions) on Wiley Online Library for rules of use; OA articles are governed by the applicable Creative Commons License

is a diagonal matrix. Indeed, since $\boldsymbol{D}^{(n)}$ is a diagonal matrix and $V_1,\dots,V_{d_n}$ are symmetric and i.i.d., the components of $\boldsymbol{D}^{(n)1/2}\boldsymbol{V}^{(n)}/\|\boldsymbol{D}^{(n)1/2}\boldsymbol{V}^{(n)}\|$ are uncorrelated. Further, the eigenvalues of $\boldsymbol{\Sigma}^{(n)}$ are the diagonal entries

$$\operatorname{diag}(\boldsymbol{\Sigma}^{(n)})_i = \mathbb{E}\left[\frac{\xi_{n,i}V_i^2}{\|\boldsymbol{\Gamma}^{(n)1/2}\boldsymbol{V}^{(n)}\|^2}\right] = \mathbb{E}\left[\frac{\xi_{n,i}V_i^2}{\sum_{j=1}^{p^*}\xi_{n,j}V_j^2 + \sum_{j=p^*+1}^{d_n}V_j^2}\right], \quad i = 1,\dots,p^*$$

and

$$\operatorname{diag}(\boldsymbol{\Sigma}^{(n)})_i = \operatorname{diag}(\boldsymbol{\Sigma}^{(n)})_{d_n} = \mathbb{E}\left[\frac{V_{d_n}^2}{\sum_{j=1}^{p^*}\xi_{n,j}V_j^2 + \sum_{j=p^*+1}^{d_n}V_j^2}\right], \quad i = p^*+1,\dots,d_n,$$

which has multiplicity $(d_n - p^*)$. For $1 \le i \le p^*$ and $l > p^*$, the function

$$\frac{\xi V_i^2 - V_l^2}{\xi V_i^2 + \sum_{j1=j\neq i}^{p^*}\xi_{n,j}V_j^2 + \sum_{j=p^*+1}^{d_n}V_j^2}$$

is a strictly increasing function in ξ since the derivative in ξ is strictly positive. A conclusion is then for $1 \le i \le p^*$ with $\xi_{n,i} > 1$ and $l > p^*$ that

$$\begin{aligned}\operatorname{diag}(\boldsymbol{\Sigma}^{(n)})_i - \operatorname{diag}(\boldsymbol{\Sigma}^{(n)})_l &= \mathbb{E}\left[\frac{\xi_{n,i}V_i^2 - V_l^2}{\sum_{j=1}^{p^*}\xi_{n,p^*}V_j^2 + \sum_{j=p^*+1}^{d_n}V_j^2}\right] \\ &> \mathbb{E}\left[\frac{V_i^2 - V_l^2}{\sum_{\substack{j=1\\ j\neq i}}^{p^*}\xi_{n,p^*}V_j^2 + V_i^2 + \sum_{j=p^*+1}^{d_n}V_j^2}\right] = 0.\end{aligned}$$

Therefore, we receive that the first p^* diagonal entries of $\boldsymbol{\Sigma}^{(n)}$ correspond to the p^* largest eigenvalues of $\boldsymbol{\Sigma}^{(n)}$ namely $\operatorname{diag}(\boldsymbol{\Sigma}^{(n)})_1, \dots, \operatorname{diag}(\boldsymbol{\Sigma}^{(n)})_{p^*}$ and the remaining $(d_n - p^*)$ eigenvalues are strictly smaller and identical to $\operatorname{diag}(\boldsymbol{\Sigma}^{(n)})_{d_n}$. ∎

Proof of Theorem 1.

(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 $\lambda_i(\boldsymbol{B})$, $i = 1,\dots,d$, the inequality

$$\max_{i=1,\dots,d} |\lambda_i(\boldsymbol{A}) - \lambda_i(\boldsymbol{B})| \le \|\boldsymbol{A} - \boldsymbol{B}\| \tag{A1}$$

holds. A conclusion from Proposition 1 is that $\sqrt{k_n}(\widehat{\boldsymbol{\Sigma}}_n - \boldsymbol{\Sigma}) = O_{\mathbb{P}}(1)$ and therefore (A1) yields

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

(b) The result corresponds to Dauxois et al. (1982, Proposition 8), which is based on a similar convergence as Proposition 1. ∎

14679469, 0, Downloaded from https://onlinelibrary.wiley.com/doi/10.1111/sjos.70026 by Karlsruher Institut Für Technologie, Wiley Online Library on [22/10/2025]. See the Terms and Conditions (https://onlinelibrary.wiley.com/terms-and-conditions) on Wiley Online Library for rules of use; OA articles are governed by the applicable Creative Commons License

A.1 Proof of Theorem 2

For the proof of Theorem 2, 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 $\widehat{\boldsymbol{\Sigma}}^{(n)}$ in (2.4).

As a consequence of the independence between the radial components $Z_1,\dots,Z_n$ and the directional components $\boldsymbol{X}_1^{(n)}/\|\boldsymbol{X}_1^{(n)}\|,\dots,\boldsymbol{X}_{k_n}^{(n)}/\|\boldsymbol{X}_{k_n}^{(n)}\|$, we obtain

$$\widehat{\boldsymbol{\Theta}}^{(n)} = \sum_{j=1}^{n} \frac{\boldsymbol{\Gamma}^{(n)1/2}\boldsymbol{V}_j^{(n)}}{\|\boldsymbol{\Gamma}^{(n)1/2}\boldsymbol{V}_j^{(n)}\|} \mathbb{1}\{Z_j > Z_{(k_n+1,n)}\} \overset{D}{=} \sum_{j=1}^{k_n} \frac{\boldsymbol{X}_j^{(n)}}{\|\boldsymbol{X}_j^{(n)}\|},$$

and similarly

$$\widehat{\boldsymbol{\Sigma}}^{(n)\prime} := \frac{1}{k_n}\sum_{j=1}^{k_n}\left(\frac{\boldsymbol{X}_j^{(n)}}{\|\boldsymbol{X}_j^{(n)}\|} - \frac{1}{k_n}\sum_{i=1}^{k_n}\frac{\boldsymbol{X}_i^{(n)}}{\|\boldsymbol{X}_i^{(n)}\|}\right)\left(\frac{\boldsymbol{X}_j^{(n)}}{\|\boldsymbol{X}_j^{(n)}\|} - \frac{1}{k_n}\sum_{i=1}^{k_n}\frac{\boldsymbol{X}_i^{(n)}}{\|\boldsymbol{X}_i^{(n)}\|}\right)^{\top} \overset{D}{=} \widehat{\boldsymbol{\Sigma}}^{(n)}. \tag{A2}$$

The eigenvalues of $\widehat{\boldsymbol{\Sigma}}^{(n)\prime}$ are denoted by $\widehat{\lambda}'_{n,1} \geq \dots \geq \widehat{\lambda}'_{n,d_n}$ and due to (A2) we receive that

$$(\widehat{\lambda}'_{n,1},\dots,\widehat{\lambda}'_{n,d_n}) \overset{D}{=} (\widehat{\lambda}_{n,1},\dots,\widehat{\lambda}_{n,d_n}). \tag{A3}$$

Thus, to prove Theorem 2, 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 $\boldsymbol{V}^{(n)}$, that is, $\boldsymbol{V}_i^{(n)} \in \mathbb{R}^{d_n}$ has i.i.d. entries with mean 0 and variance 1. Then we define the sequence of matrices

$$\boldsymbol{\Upsilon}^{(n)} := \frac{1}{k_n}\sum_{i=1}^{k_n}\left(\boldsymbol{\Gamma}^{(n)1/2}\boldsymbol{V}_i^{(n)} - \frac{1}{k_n}\sum_{j=1}^{k_n}\boldsymbol{\Gamma}^{(n)1/2}\boldsymbol{V}_j^{(n)}\right)\cdot\left(\boldsymbol{\Gamma}^{(n)1/2}\boldsymbol{V}_i^{(n)} - \frac{1}{k_n}\sum_{j=1}^{k_n}\boldsymbol{\Gamma}^{(n)1/2}\boldsymbol{V}_j^{(n)}\right)^{\top}, \quad n \in \mathbb{N}, \tag{A4}$$

whose eigenvalues are denoted by $\widehat{\xi}_{n,1} > \dots > \widehat{\xi}_{n,d_n} > 0$. 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

$$\boldsymbol{T}^{(n)} := \operatorname{diag}(\|\boldsymbol{\Gamma}^{(n)1/2}\boldsymbol{V}_1^{(n)}\|^{-1},\dots,\|\boldsymbol{\Gamma}^{(n)1/2}\boldsymbol{V}_{k_n}^{(n)}\|^{-1}) \in \mathbb{R}^{k_n\times k_n},$$

which allows us to write

$$\left(\frac{\boldsymbol{X}_1^{(n)}}{\|\boldsymbol{X}_1^{(n)}\|},\dots,\frac{\boldsymbol{X}_{k_n}^{(n)}}{\|\boldsymbol{X}_{k_n}^{(n)}\|}\right)^{\top} = \boldsymbol{\Gamma}^{(n)1/2}\mathcal{V}^{(n)}\boldsymbol{T}^{(n)}.$$

14679469, 0, Downloaded from https://onlinelibrary.wiley.com/doi/10.1111/sjos.70026 by Karlsruher Institut Für Technologie, Wiley Online Library on [22/10/2025]. See the Terms and Conditions (https://onlinelibrary.wiley.com/terms-and-conditions) on Wiley Online Library for rules of use; OA articles are governed by the applicable Creative Commons License

Finally, with the projection matrix $\boldsymbol{P}^{(n)} := (\boldsymbol{I}_{k_n} - \mathbf{1}_{k_n}\mathbf{1}_{k_n}^\top/k_n)$, the matrices $\widehat{\boldsymbol{\Sigma}}^{(n)\prime}$ and $\boldsymbol{\Upsilon}^{(n)}$, as defined in (A2) and (A4), can be written as

$$\begin{aligned}\widehat{\boldsymbol{\Sigma}}^{(n)\prime} &= \frac{1}{k_n}(\boldsymbol{\Gamma}^{(n)1/2}\boldsymbol{\mathcal{Y}}^{(n)}\boldsymbol{T}^{(n)}\boldsymbol{P}^{(n)})(\boldsymbol{\Gamma}^{(n)1/2}\boldsymbol{\mathcal{Y}}^{(n)}\boldsymbol{T}^{(n)}\boldsymbol{P}^{(n)})^\top,\\ \boldsymbol{\Upsilon}^{(n)} &= \frac{1}{k_n}(\boldsymbol{\Gamma}^{(n)1/2}\boldsymbol{\mathcal{Y}}^{(n)}\boldsymbol{P}^{(n)})(\boldsymbol{\Gamma}^{(n)1/2}\boldsymbol{\mathcal{Y}}^{(n)}\boldsymbol{P}^{(n)})^\top.\end{aligned} \tag{A5}$$

In the following theorem, the connection between the eigenvalues $\widehat{\xi}_{n,i}$ and $d_n\widehat{\lambda}'_{n,i}$ is derived.

Theorem 10. *Let Model 1 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* $\hat{\xi}_{n,1} \geq \dots \geq \hat{\xi}_{n,d_n}$ *denote the eigenvalues of* $\boldsymbol{\Upsilon}^{(n)}$ *in (A4) and* $\widehat{\lambda}'_{n,1} \geq \dots \geq \widehat{\lambda}'_{n,d_n}$ *denote the eigenvalues of* $\widehat{\boldsymbol{\Sigma}}^{(n)\prime}$ *in (A2), then as* $n\to\infty$,

$$\max_{1\leq i\leq d_n}\left|\widehat{\xi}_{n,i} - d_n\widehat{\lambda}'_{n,i}\right| \xrightarrow{\mathbb{P}} 0.$$

Proof of Theorem 10. Due to Theorem A.46 in Bai and Silverstein (2010) and the submultiplicativity of the spectral norm, we receive that

$$\begin{aligned}\max_{1\leq i\leq d_n}\left|\sqrt{\widehat{\xi}_{n,i}} - \sqrt{d_n\widehat{\lambda}'_{n,i}}\right| &\leq \left\|\frac{\sqrt{d_n}\boldsymbol{\Gamma}^{(n)1/2}\boldsymbol{\mathcal{Y}}^{(n)}\boldsymbol{T}^{(n)}\boldsymbol{P}^{(n)}}{\sqrt{k_n}} - \frac{\boldsymbol{\Gamma}^{(n)1/2}\boldsymbol{\mathcal{Y}}^{(n)}\boldsymbol{P}^{(n)}}{\sqrt{k_n}}\right\|\\ &\leq \left\|\boldsymbol{P}^{(n)}\right\|\cdot\left\|\sqrt{d_n}\boldsymbol{T}^{(n)} - \boldsymbol{I}_{k_n}\right\|\cdot\left\|\frac{\boldsymbol{\Gamma}^{(n)1/2}\boldsymbol{\mathcal{Y}}^{(n)}}{\sqrt{k_n}}\right\|\\ &= \left\|\sqrt{d_n}\boldsymbol{T}^{(n)} - \boldsymbol{I}_{k_n}\right\|\cdot\left\|\frac{\boldsymbol{\Gamma}^{(n)1/2}\boldsymbol{\mathcal{Y}}^{(n)}}{\sqrt{k_n}}\right\| =: J_n\cdot H_n,\end{aligned} \tag{A6}$$

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 (A6) 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

$$\boldsymbol{\Gamma}^{(n)1/2}\boldsymbol{V}_j^{(n)} = \begin{pmatrix}\boldsymbol{\Gamma}_n^{1/2}\boldsymbol{V}^{(n)}_{j,\{1,\dots,p^*\}}\\ \boldsymbol{V}^{(n)}_{j,\{p^*+1,\dots,d_n\}}\end{pmatrix} =: \begin{pmatrix}(U^{(n)}_{j,1},\dots,U^{(n)}_{j,p^*})^\top\\ (V_{j,(p^*+1)},\dots,V_{j,d_n})^\top\end{pmatrix}.$$

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

$$J_n^{1/2} = \left\|\sqrt{d_n}\boldsymbol{T}^{(n)} - \boldsymbol{I}_{k_n}\right\|^{1/2} = \max_{1\leq i\leq k_n}\left|\frac{\sqrt{d_n}}{\left(\sum_{l=1}^{p^*}U^{(n)2}_{i,l} + \sum_{l=p^*+1}^{d_n}V^2_{i,l}\right)^{1/2}} - 1\right|.$$

14679469, 0, Downloaded from https://onlinelibrary.wiley.com/doi/10.1111/sjos.70026 by Karlsruher Institut Für Technologie, Wiley Online Library on [22/10/2025]. See the Terms and Conditions (https://onlinelibrary.wiley.com/terms-and-conditions) on Wiley Online Library for rules of use; OA articles are governed by the applicable Creative Commons License

On the one hand, by $\mathbb{E}[V_{i,l}^2] = 1$, $d_n/k_n \to c > 0$ 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 $1 \le i \le k_n$,

$$\sum_{l=1}^{p^*} U_{i,l}^{(n)2} = \|\boldsymbol{\Gamma}_n^{1/2}\boldsymbol{V}_{i,\{1,..,p^*\}}^{(n)}\|^2 \le \|\boldsymbol{\Gamma}_n^{1/2}\|^2\|\boldsymbol{V}_{i,\{1,..,p^*\}}^{(n)}\|^2 = \xi_{n,1}\sum_{l=1}^{p^*} V_{i,l}^2.$$

Since the second moment of V_1^2 exists, we can conclude from Markov's inequality for $\varepsilon > 0$

$$\begin{aligned}\mathbb{P}\left(\frac{\xi_{n,1}}{d_n}\max_{1\le i\le k_n}\left|\sum_{l=1}^{p^*} V_{i,l}^2\right| > \varepsilon\right) &\le \sum_{i=1}^{k_n}\mathbb{P}\left(\left|\sum_{l=1}^{p^*} V_{i,l}^2\right| > \frac{d_n}{\xi_{n,1}}\varepsilon\right)\\ &= k_n\mathbb{P}\left(\left|\sum_{l=1}^{p^*} V_{1,l}^2\right| > \frac{d_n}{\xi_{n,1}}\varepsilon\right)\\ &\le k_n\frac{\xi_{n,1}^2}{d_n^2\varepsilon^2}\mathbb{E}\left|\sum_{l=1}^{p^*} V_{1,l}^2\right|^2, \end{aligned} \tag{A7}$$

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,

$$\begin{aligned}\max_{1\le i\le k_n}\left|\left(\frac{\sum_{l=1}^{p^*} U_{i,l}^{(n)2} + \sum_{l=p^*+1}^{d_n} V_{i,l}^2}{d_n}\right) - 1\right| &\le \max_{1\le i\le k_n}\left|\left(\frac{\sum_{l=1}^{p^*} U_{i,l}^{(n)2}}{d_n}\right)\right|\\ &+ \max_{1\le i\le k_n}\left|\left(\frac{\sum_{l=p^*+1}^{d_n} V_{i,l}^2}{d_n}\right) - 1\right| \xrightarrow{\mathbb{P}} 0.\end{aligned}$$

Finally, by the mean value theorem the inequality

$$\left|1 - 1/\sqrt{x}\right| \le 2|x-1| \quad \text{for } x > \frac{1}{2}$$

holds and hence, as $n \to \infty$,

$$J_n^{1/2} = \max_{1\le i\le k_n}\left|1 - \left(\frac{1}{d_n}\sum_{l=1}^{p^*} U_{i,l}^{(n)2} + \frac{1}{d_n}\sum_{l=p^*+1}^{d_n} V_{i,l}^2\right)^{-1/2}\right| \xrightarrow{\mathbb{P}} 0. \tag{A8}$$

14679469, 0, Downloaded from https://onlinelibrary.wiley.com/doi/10.1111/sjos.70026 by Karlsruher Institut Für Technologie, Wiley Online Library on [22/10/2025]. See the Terms and Conditions (https://onlinelibrary.wiley.com/terms-and-conditions) on Wiley Online Library for rules of use; OA articles are governed by the applicable Creative Commons License

Step 2. Next, we show that H_n in (A6) is $\mathbb{P}$-a.s. bounded. By Yin et al. (1988, Theorem 3.1) (cf. Bai & Silverstein, 2010, Theorem 5.8)

$$H_n = \left\| \frac{\mathbf{\Gamma}^{(n)1/2} \mathcal{Y}^{(n)}}{\sqrt{k_n}} \right\| \leq \left\| \mathbf{\Gamma}^{(n)1/2} \right\| \cdot \left\| \frac{\mathcal{Y}^{(n)}}{\sqrt{k_n}} \right\|^2 = \xi_{n,1} \frac{\lambda_{\max}(\mathcal{Y}^{(n)\top} \mathcal{Y}^{(n)})}{k_n} \xrightarrow{\mathbb{P}\text{-a.s.}} \xi_1$$

as $n \to \infty$, where $\lambda_{\max}(\cdot)$ denotes the largest eigenvalue of a matrix.

Finally, a combination of (A6), Step 1 and Step 2 result in the statement. ∎

Remark 11. For the convergence of the right-hand side of (A7) and hence, (A8) 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 (A8) 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 will repeat results on the asymptotic distribution of the eigenvalues of $\mathbf{\Upsilon}^{(n)}$, which are mainly based on Bai and Yao (2012) and Bai et al. (2018).

Lemma 2. *Let Model 1 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 2.*

(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*

$$\sup_{\alpha \in (0,1)} \left| \widehat{\xi}_{n,i_n(\alpha)} - F_c^{\leftarrow}(1-\alpha) \right| \xrightarrow{\mathbb{P}\text{-a.s.}} 0, \quad \text{as } n \to \infty.$$

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* $\widehat{\xi}_{n,q_n} \xrightarrow{\mathbb{P}\text{-a.s.}} (1 + \sqrt{c})^2$.

14679469, 0, Downloaded from https://onlinelibrary.wiley.com/doi/10.1111/sjos.70026 by Karlsruher Institut Für Technologie, Wiley Online Library on [22/10/2025]. See the Terms and Conditions (https://onlinelibrary.wiley.com/terms-and-conditions) on Wiley Online Library for rules of use; OA articles are governed by the applicable Creative Commons License

(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 $(\xi_{n,1},\dots,\xi_{n,p^*}) = (\xi_1,\dots,\xi_{p^*})$ 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 $(\xi_{n,1},\dots,\xi_{n,p^*}) \to (\xi_1,\dots,\xi_{p^*})$ as $n \to \infty$ the assertion also holds because by Bai and Silverstein (2010, Theorem A.46) and similar arguments as before it can be shown that

$$\max_{1\le i\le p^*} \left| \sqrt{\widehat{\xi}_{n,i}(\mathbf{\Gamma}_n)} - \sqrt{\widehat{\xi}_{n,i}(\mathbf{\Gamma})} \right| \le \|\mathbf{\Gamma}_n - \mathbf{\Gamma}\| \left\| \frac{\mathcal{Y}^{(n)}}{\sqrt{k_n}} \right\| \|\boldsymbol{P}^{(n)}\| \xrightarrow{\mathbb{P}\text{-a.s.}} 0,$$

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), 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

$$\lim_{n\to\infty} \sup_{\alpha\in(l^*,1)} \left| F^{\mathbf{\Upsilon}^{(n)}\leftarrow}(\alpha,\omega) - F_c^{\leftarrow}(\alpha) \right| = 0.$$

(c) Since $\widehat{\xi}_{n,i_n(\alpha)} = F^{\mathbf{\Upsilon}^{(n)}\leftarrow}(1 - i_n(\alpha)/d_n)$ the statement follows directly from (b).

14679469, 0, Downloaded from https://onlinelibrary.wiley.com/doi/10.1111/sjos.70026 by Karlsruher Institut Für Technologie, Wiley Online Library on [22/10/2025]. See the Terms and Conditions (https://onlinelibrary.wiley.com/terms-and-conditions) on Wiley Online Library for rules of use; OA articles are governed by the applicable Creative Commons License

(d) Due to (b), we receive that

$$\frac{1}{d_n - q_n}\sum_{i=q_n+1}^{d_n}\widehat{\xi}_{n,i} = \frac{d_n}{d_n - q_n}\int_0^{1-\frac{q_n}{d_n}} F^{\mathbf{Y}^{(n)}\,\leftarrow}(1-\alpha)\,d\alpha \xrightarrow{\mathbb{P}\text{-a.s.}} 1\cdot\int_0^1 F_c^{\leftarrow}(1-\alpha)\,d\alpha = 1.$$

(e) Similar to (d), we have

$$\frac{1}{k_n - q_n}\sum_{i=q_n+1}^{k_n}\hat{\xi}_{n,i} = \frac{d_n}{k_n - q_n}\int_{1-\frac{k_n}{d_n}}^{1-\frac{q_n}{d_n}} F^{\mathbf{Y}^{(n)}\,\leftarrow}(1-\alpha)\,d\alpha \xrightarrow{\mathbb{P}\text{-a.s.}} c\int_{1-c^{-1}}^1 F_c^{\leftarrow}(1-\alpha)\,d\alpha = c.$$

■

Finally, we have all auxiliary results for the proof of Theorem 2.

Proof of Theorem 2 (a). An assumption is that $\xi_i > 1+\sqrt{c}$ and hence, ξ_i is a distant spiked eigenvalue for $i = 1,\dots,p^*$. A conclusion of Lemma 2(a) is then that $\widehat{\xi}_{n,i} \xrightarrow{\mathbb{P}\text{-a.s.}} \varphi_c(\xi_i)$. Combined with Theorem 10, we receive that $d_n\widehat{\lambda}'_{n,i} \xrightarrow{\mathbb{P}} \varphi_c(\xi_i)$ as $n\to\infty$. Due to (A3), 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 2, Theorem 10 and (A3). ■

A.2 Proof of Theorem 3

For the proof of Theorem 3, Theorem 10 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 from $\widehat{\mathbf{\Sigma}}^{(n)\prime}$, which is then used for the proof of Theorem 3.

Lemma 3. *Let Model 1 with* $\xi_{n,p^*}\to\infty$ *and* $\xi_{n,1} = o(d_n^{1/2})$ *as* $n\to\infty$ *be given. If* $i\in\{1,\dots,p^*\}$ *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 $\mathbf{\Gamma}^{(n)}$. Let $\mathbf{\Gamma}^{(n)} = \boldsymbol{W}^{(n)}\boldsymbol{D}^{(n)}\boldsymbol{W}^{(n)\top}$, where $\boldsymbol{W}^{(n)} = (\boldsymbol{W}_1^{(n)},\dots,\boldsymbol{W}_{d_n}^{(n)})$ is a $(d_n\times d_n)$-dimensional orthogonal matrix and $\boldsymbol{D}^{(n)} := \operatorname{diag}(\overline{\boldsymbol{D}}_n, \boldsymbol{I}_{d_n-p^*}) := \operatorname{diag}(\xi_{n,1},\dots,\xi_{n,p^*},1,\dots,1)\in\mathbb{R}^{d_n\times d_n}$ consists of the eigenvalues of $\mathbf{\Gamma}^{(n)}$. Then with representation (A5) and

$$\boldsymbol{A}^{(n)} := \mathcal{Y}^{(n)}\boldsymbol{T}^{(n)}\boldsymbol{P}^{(n)}\boldsymbol{T}^{(n)}\mathcal{Y}^{(n)\top} \tag{A9}$$

we receive

$$\begin{aligned}\widehat{\mathbf{\Sigma}}^{(n)\prime} &= \frac{1}{k_n}\boldsymbol{W}^{(n)}\boldsymbol{D}^{(n)1/2}\boldsymbol{W}^{(n)\top}\mathcal{Y}^{(n)}\boldsymbol{T}^{(n)}\boldsymbol{P}^{(n)}\boldsymbol{T}^{(n)}\mathcal{Y}^{(n)\top}\boldsymbol{W}^{(n)}\boldsymbol{D}^{(n)1/2}\boldsymbol{W}^{(n)\top}\\ &= \frac{1}{k_n}\boldsymbol{W}^{(n)}\boldsymbol{D}^{(n)1/2}\boldsymbol{W}^{(n)\top}\boldsymbol{A}^{(n)}\boldsymbol{W}^{(n)}\boldsymbol{D}^{(n)1/2}\boldsymbol{W}^{(n)\top}.\end{aligned} \tag{A10}$$

14679469, 0, Downloaded from https://onlinelibrary.wiley.com/doi/10.1111/sjos.70026 by Karlsruher Institut Für Technologie, Wiley Online Library on [22/10/2025]. See the Terms and Conditions (https://onlinelibrary.wiley.com/terms-and-conditions) on Wiley Online Library for rules of use; OA articles are governed by the applicable Creative Commons License

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 (A5) and

$$\boldsymbol{B}^{(n)} := \mathcal{V}^{(n)} \boldsymbol{P}^{(n)} \mathcal{V}^{(n)\top} \tag{A11}$$

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 & Johnson, 2013, Theorem 4.2.6) gives

$$\frac{d_n \widehat{\lambda}'_{n,i}}{\xi_{n,i}} = \frac{d_n}{\xi_{n,i}} \inf_{\boldsymbol{v}_1, \ldots, \boldsymbol{v}_{i-1} \in \mathbb{R}^{d_n}} \sup_{\boldsymbol{w} \perp \boldsymbol{v}_1, \ldots, \boldsymbol{v}_{i-1}, \|\boldsymbol{w}\|=1} \boldsymbol{w}^\top \widehat{\boldsymbol{\Sigma}}^{(n)\prime} \boldsymbol{w}. \tag{A12}$$

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 (A13), which converges in probability to 1. Therefore, note for arbitrary $\boldsymbol{u}_j \in \mathbb{R}^{d_n}$ with $\|\boldsymbol{u}_j\| = 1$ for $1 \le j \le p^*$, Bai et al. (2018, Lemma A.2) yields that as $n \to \infty$,

$$\max_{1 \le j \le p^*} \left| \boldsymbol{u}_j^\top \frac{\boldsymbol{B}^{(n)}}{k_n} \boldsymbol{u}_j - 1 \right| \xrightarrow{\mathbb{P}\text{-a.s.}} 0, \tag{A13}$$

where $\boldsymbol{B}^{(n)}$ is defined as in (A12). Now, let $\boldsymbol{A}^{(n)}$ be defined as in (A10). Then

$$\begin{aligned} \left| \boldsymbol{u}_j^\top \left(\frac{\boldsymbol{B}^{(n)}}{k_n} - \frac{d_n \boldsymbol{A}^{(n)}}{k_n} \right) \boldsymbol{u}_j \right| &\le \left\| \frac{\boldsymbol{B}^{(n)}}{k_n} - \frac{d_n \boldsymbol{A}^{(n)}}{k_n} \right\| \\ &= \frac{1}{k_n} \left\| \mathcal{V}^{(n)} \left(\boldsymbol{P}^{(n)} - d_n \boldsymbol{T}^{(n)} \boldsymbol{P}^{(n)} \boldsymbol{T}^{(n)\top} \right) \mathcal{V}^{(n)\top} \right\| \\ &\le \frac{1}{k_n} \|\mathcal{V}^{(n)}\|^2 \|\boldsymbol{P}^{(n)}\| \left(1 + \|\sqrt{d_n} \boldsymbol{T}^{(n)}\| \right) \left\| \sqrt{d_n} \boldsymbol{T}^{(n)} - \boldsymbol{I}_{k_n} \right\|. \end{aligned}$$

On the one hand, Yin et al. (1988, Theorem 3.1) implies that

$$\frac{1}{k_n} \|\mathcal{V}^{(n)}\|^2 = \left(\frac{1}{\sqrt{k_n}} \|\mathcal{V}^{(n)}\| \right)^2 \xrightarrow{\mathbb{P}\text{-a.s.}} (1 + \sqrt{c})^2.$$

14679469, 0, Downloaded from https://onlinelibrary.wiley.com/doi/10.1111/sjos.70026 by Karlsruher Institut Für Technologie, Wiley Online Library on [22/10/2025]. See the Terms and Conditions (https://onlinelibrary.wiley.com/terms-and-conditions) on Wiley Online Library for rules of use; OA articles are governed by the applicable Creative Commons License

On the other hand, since $\xi_{n,1} = o(d_n^{1/2})$ as $n \to \infty$, a conclusion of Remark 11 is that $\|\sqrt{d_n}\boldsymbol{T}^{(n)} - \boldsymbol{I}_{k_n}\| \xrightarrow{\mathbb{P}} 0$ and $\|\sqrt{d_n}\boldsymbol{T}^{(n)}\| \le \|\boldsymbol{I}_{k_n}\| + \|\sqrt{d_n}\boldsymbol{T}^{(n)} - \boldsymbol{I}_{k_n}\| \xrightarrow{\mathbb{P}} 1$. In summary, as $n \to \infty$

$$\left|\boldsymbol{u}_j^\top \left(\frac{\boldsymbol{B}^{(n)}}{k_n} - \frac{d_n\boldsymbol{A}^{(n)}}{k_n}\right)\boldsymbol{u}_j\right| \le \left\|\frac{\boldsymbol{B}^{(n)}}{k_n} - \frac{d_n\boldsymbol{A}^{(n)}}{k_n}\right\| \xrightarrow{\mathbb{P}} 0, \tag{A14}$$

and finally, using (A14) we have as well

$$\max_{1\le j\le p^*}\left|\boldsymbol{u}_j^\top \frac{d_n\boldsymbol{A}^{(n)}}{k_n}\boldsymbol{u}_j - 1\right| \xrightarrow{\mathbb{P}} 0. \tag{A15}$$

Further, for arbitrary vectors $\boldsymbol{v}_1,\dots,\boldsymbol{v}_{i-1} \in \mathbb{R}^{d_n}$ we take a vector $\boldsymbol{w}_{\boldsymbol{v}} = \sum_{j=1}^i a_j \boldsymbol{W}_j^{(n)}$ orthogonal to $\boldsymbol{v}_1,\dots,\boldsymbol{v}_{i-1}$ with $\sum_{j=1}^i a_j^2 = 1$ and hence, $\|\boldsymbol{w}_{\boldsymbol{v}}\| = 1$. Since $\boldsymbol{W}^{(n)}$ is an orthogonal matrix, we receive with representation (A10) that

$$\begin{aligned}\frac{d_n}{\xi_{n,i}}\boldsymbol{w}_{\boldsymbol{v}}^\top \widehat{\boldsymbol{\Sigma}}^{(n)\prime}\boldsymbol{w}_{\boldsymbol{v}} &= \frac{d_n}{\xi_{n,i}}\sum_{j,l=1}^i a_j a_l \boldsymbol{W}_j^{(n)\top}\boldsymbol{W}^{(n)}\boldsymbol{D}^{(n)1/2}\boldsymbol{W}^{(n)\top}\frac{\boldsymbol{A}^{(n)}}{k_n}\boldsymbol{W}^{(n)}\boldsymbol{D}^{(n)1/2}\boldsymbol{W}^{(n)\top}\boldsymbol{W}_l^{(n)}\\ &= \sum_{j=1}^i a_j^2 \frac{\xi_{n,j}}{\xi_{n,i}}\boldsymbol{W}_j^{(n)\top}\frac{d_n\boldsymbol{A}^{(n)}}{k_n}\boldsymbol{W}_j^{(n)}.\end{aligned}$$

A conclusion of (A12), $\|\boldsymbol{W}_j^{(n)}\| = 1$ and (A15) is then

$$\begin{aligned}\frac{d_n\widehat{\lambda}'_{n,i}}{\xi_{n,i}} &\ge \inf_{\boldsymbol{v}_1,\dots,\boldsymbol{v}_{i-1}\in\mathbb{R}^{d_n}}\frac{d_n}{\xi_{n,i}}\boldsymbol{w}_{\boldsymbol{v}}^\top\widehat{\boldsymbol{\Sigma}}^{(n)\prime}\boldsymbol{w}_{\boldsymbol{v}} \ge \inf_{\boldsymbol{a}\in\mathbb{R}^i:\sum_{j=1}^i a_j^2=1}\sum_{j=1}^i a_j^2\boldsymbol{W}_j^{(n)\top}\frac{d_n\boldsymbol{A}^{(n)}}{k_n}\boldsymbol{W}_j^{(n)}\\ &\ge 1 - \max_{1\le j\le p^*}\left|\boldsymbol{W}_j^{(n)\top}\frac{d_n\boldsymbol{A}^{(n)}}{k_n}\boldsymbol{W}_j^{(n)} - 1\right| \xrightarrow{\mathbb{P}} 1\end{aligned}$$

as $n \to \infty$.

Step 2: Next, we derive an upper bound for (A12) which converges in probability to 1. Therefore, note that

$$\frac{d_n\widehat{\lambda}'_{n,i}}{\xi_{n,i}} = \frac{d_n}{\xi_{n,i}}\inf_{\boldsymbol{v}_1,\dots,\boldsymbol{v}_{i-1}\in\mathbb{R}^{d_n}}\sup_{\boldsymbol{w}\perp\boldsymbol{v}_1,\dots,\boldsymbol{v}_{i-1},\|\boldsymbol{w}\|=1}\boldsymbol{w}^\top\widehat{\boldsymbol{\Sigma}}^{(n)\prime}\boldsymbol{w} \le \frac{d_n}{\xi_{n,i}}\sup_{\boldsymbol{w}\perp\boldsymbol{W}_1^{(n)},\dots,\boldsymbol{W}_{i-1}^{(n)},\|\boldsymbol{w}\|=1}\boldsymbol{w}^\top\widehat{\boldsymbol{\Sigma}}^{(n)\prime}\boldsymbol{w}.$$

Since $\boldsymbol{W}_l^{(n)} \perp \boldsymbol{W}_j^{(n)}$ for $l \ne j$ we can write a vector $\boldsymbol{w}\perp\boldsymbol{W}_1^{(n)},\dots,\boldsymbol{W}_{i-1}^{(n)}$ with $\|\boldsymbol{w}\| = 1$ as

$$\boldsymbol{w} = c^2\boldsymbol{u} + (1-c^2)\boldsymbol{v},$$

where $c \in [0,1]$, $\boldsymbol{u} = \sum_{j=i}^{p^*} a_j\boldsymbol{W}_j^{(n)} = \overline{\boldsymbol{W}}^{(n)}\boldsymbol{a}$, $\|\boldsymbol{a}\| = \sum_{j=i}^{p^*} a_j^2 = 1$ 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,

14679469, 0, Downloaded from https://onlinelibrary.wiley.com/doi/10.1111/sjos.70026 by Karlsruher Institut Für Technologie, Wiley Online Library on [22/10/2025]. See the Terms and Conditions (https://onlinelibrary.wiley.com/terms-and-conditions) on Wiley Online Library for rules of use; OA articles are governed by the applicable Creative Commons License

$$\frac{d_n}{\xi_{n,i}} \sup_{\boldsymbol{w}\perp \boldsymbol{W}_1^{(n)},\ldots,\boldsymbol{W}_{i-1}^{(n)},\|\boldsymbol{w}\|=1} \boldsymbol{w}^\top \widehat{\boldsymbol{\Sigma}}^{(n)\prime}\boldsymbol{w}$$

$$\leq \frac{d_n}{\xi_{n,i}} \sup_{c\in[0,1]} \left\{ c^2 \sup_{\substack{\boldsymbol{a}\in\mathbb{R}^{p^*-i+1},\\ \|\boldsymbol{a}\|=1}} \boldsymbol{a}^\top \overline{\boldsymbol{W}}^{(n)\top} \widehat{\boldsymbol{\Sigma}}^{(n)\prime} \overline{\boldsymbol{W}}^{(n)} \boldsymbol{a} + (1-c^2) \sup_{\substack{\boldsymbol{b}\in\mathbb{R}^{d-p^*},\\ \|\boldsymbol{b}\|=1}} \boldsymbol{b}^\top \widetilde{\boldsymbol{W}}^{(n)\top} \widehat{\boldsymbol{\Sigma}}^{(n)\prime} \widetilde{\boldsymbol{W}}^{(n)} \boldsymbol{b} \right\}$$

$$\leq \sup_{c\in[0,1]} \left\{ c^2 \sup_{\substack{\boldsymbol{a}\in\mathbb{R}^{p^*-i+1},\\ \|\boldsymbol{a}\|=1}} \sum_{j=i}^{p^*} a_j^2 \boldsymbol{W}_j^{(n)\top} \frac{d_n \boldsymbol{A}^{(n)}}{k_n} \boldsymbol{W}_j^{(n)} + (1-c^2) \frac{d_n}{\xi_{n,i}} \left\| \widetilde{\boldsymbol{W}}^{(n)\top} \frac{\boldsymbol{A}^{(n)}}{k_n} \widetilde{\boldsymbol{W}}^{(n)} \right\| \right\}.$$

Note that (A14) and $\widetilde{\boldsymbol{W}}^{(n)}$ being an orthogonal matrix imply that

$$\left\| \widetilde{\boldsymbol{W}}^{(n)\top} \left(\frac{d_n \boldsymbol{A}^{(n)}}{k_n} - \frac{\boldsymbol{B}^{(n)}}{k_n} \right) \widetilde{\boldsymbol{W}}^{(n)} \right\| \leq \left\| \frac{d_n \boldsymbol{A}^{(n)}}{k_n} - \frac{\boldsymbol{B}^{(n)}}{k_n} \right\| \xrightarrow{\mathbb{P}} 0.$$

We then conclude from $\|\widetilde{\boldsymbol{W}}^{(n)\top} \frac{\boldsymbol{B}^{(n)}}{k_n} \widetilde{\boldsymbol{W}}^{(n)}\| \xrightarrow{\mathbb{P}} (1+\sqrt{c})^2$ (cf. proof of Bai et al., 2018, Lemma 2.2 (i)) and $\xi_{n,i} \to \infty$ that as $n \to \infty$,

$$\frac{1}{\xi_{n,i}} \left\| \widetilde{\boldsymbol{W}}^{(n)\top} \frac{d_n \boldsymbol{A}^{(n)}}{k_n} \widetilde{\boldsymbol{W}}^{(n)} \right\| \xrightarrow{\mathbb{P}} 0.$$

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 (A15) we get,

$$\frac{d_n \widehat{\lambda}'_{n,i}}{\xi_{n,i}} \leq \frac{d_n}{\xi_{n,i}} \sup_{\substack{\boldsymbol{w}\perp \boldsymbol{W}_1^{(n)},\ldots \boldsymbol{W}_{i-1}^{(n)},\\ \|\boldsymbol{w}\|=1}} \boldsymbol{w}^\top \widehat{\boldsymbol{\Sigma}}^{(n)\prime} \boldsymbol{w} \xrightarrow{\mathbb{P}} \sup_{c\in[0,1]} \left\{ c^2 \sup_{\substack{p\\ j=i}} \sum_{j=i}^{p^*} a_j^2 1 = \sum_{j=i}^{p^*} a_j^2 \right\} = 1$$

as $n \to \infty$, which proves Step 2. ∎

Lemma 4. *Let Model 1 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 2.

Proof. For the ease of notation, we 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 (A14), it follows by Bai and Silverstein (2010, Theorem A.45) that as $n \to \infty$,

$$\sup_{x\in I} \left| F^{\widetilde{\boldsymbol{W}}^{(n)\top} \frac{\boldsymbol{B}^{(n)}}{k_n} \widetilde{\boldsymbol{W}}^{(n)}}(x) - F^{\widetilde{\boldsymbol{W}}^{(n)\top} \frac{d_n \boldsymbol{A}^{(n)}}{k_n} \widetilde{\boldsymbol{W}}^{(n)}}(x) \right| \xrightarrow{\mathbb{P}} 0.$$

14679469, 0, Downloaded from https://onlinelibrary.wiley.com/doi/10.1111/sjos.70026 by Karlsruher Institut Für Technologie, Wiley Online Library on [22/10/2025]. See the Terms and Conditions (https://onlinelibrary.wiley.com/terms-and-conditions) on Wiley Online Library for rules of use; OA articles are governed by the applicable Creative Commons License

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

$$\sup_{x\in I}\left|F^{\widetilde{\boldsymbol{W}}^{(n)\top}\frac{d_n\boldsymbol{A}^{(n)}}{k_n}\widetilde{\boldsymbol{W}}^{(n)}}(x) - F_c(x)\right| \xrightarrow{\mathbb{P}} 0. \tag{A16}$$

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 eigenvalues (Horn & Johnson, 2013, Theorem 4.3.28) that $\mathbb{P}$-a.s.

$$\begin{aligned}\lambda_i\left(\widetilde{\boldsymbol{W}}^{(n)\top}\frac{d_n\boldsymbol{A}^{(n)}}{k_n}\widetilde{\boldsymbol{W}}^{(n)}\right) &\geq \lambda_{p^*+i}\left(\boldsymbol{W}^{(n)\top} d_n\widehat{\boldsymbol{\Sigma}}^{(n)\prime}\boldsymbol{W}^{(n)}\right) = d_n\widehat{\lambda}'_{n,p^*+i}\\ &\geq \lambda_{p^*+i}\left(\widetilde{\boldsymbol{W}}^{(n)\top}\frac{d_n\boldsymbol{A}^{(n)}}{k_n}\widetilde{\boldsymbol{W}}^{(n)}\right).\end{aligned} \tag{A17}$$

Therefore, due to (A16) and (A17),

$$\begin{aligned}\sup_{x\in I}\left|F^{d_n\widehat{\boldsymbol{\Sigma}}^{(n)\prime}}(x) - F_c(x)\right| &= \sup_{x\in I}\left|\frac{1}{d_n}\sum_{i=1}^{d_n}\mathbb{1}\left\{d_n\widehat{\lambda}'_{n,i}\leq x\right\} - F_c(x)\right|\\ &\leq \sup_{x\in I}\left|F^{\widetilde{\boldsymbol{W}}^{(n)\top}\frac{d_n\boldsymbol{A}^{(n)}}{k_n}\widetilde{\boldsymbol{W}}^{(n)}}(x) - F_c(x)\right| + \frac{4p^*}{d_n} \xrightarrow{\mathbb{P}} 0,\end{aligned}$$

which is the statement. ∎

Proof of Theorem 3. The proof of Theorem 3 (a)-(d) follows with the same arguments as the proof of Lemma 2 using only Lemma 3 and Lemma 4 in combination with $\widehat{\boldsymbol{\Sigma}}^{(n)} \overset{D}{=} \widehat{\boldsymbol{\Sigma}}^{(n)\prime}$ (cf. (A2)). 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). ∎

14679469, 0, Downloaded from https://onlinelibrary.wiley.com/doi/10.1111/sjos.70026 by Karlsruher Institut Für Technologie, Wiley Online Library on [22/10/2025]. See the Terms and Conditions (https://onlinelibrary.wiley.com/terms-and-conditions) on Wiley Online Library for rules of use; OA articles are governed by the applicable Creative Commons License

APPENDIX B. PROOFS OF SECTION 3

Proof of Theorem 4. Since by Remark 6 (b) the BIC is scale invariant, we assume w.l.o.g. that $\lambda = 1$.

Step 1: Suppose $p > p^*$. Note

$$\log(k_n)(p+1)(d-p/2) - \frac{\log(k_n)}{2}(d-1)(d+2) = -\frac{\log(k_n)}{2}(d-p-2)(d-p+1).$$

By the definition of the BIC we obtain

$$\begin{aligned}\mathrm{BIC}_{k_n}(p) - \mathrm{BIC}_{k_n}(p^*) = {} & k_n \sum_{i=p^*+1}^{p} \log(\widehat{\lambda}_{n,i}) + k_n(d-1-p)\log\left(\frac{1}{d-1-p}\sum_{j=p+1}^{d-1}\widehat{\lambda}_{n,j}\right) \\ & - k_n(d-1-p^*)\log\left(\frac{1}{d-1-p^*}\sum_{j=p^*+1}^{d-1}\widehat{\lambda}_{n,j}\right) \\ & - \frac{\log(k_n)}{2}(d-p-2)(d-p+1) \\ & + \frac{\log(k_n)}{2}(d-p^*-2)(d-p^*+1),\end{aligned}$$

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

$$\boldsymbol{M}_n := \sqrt{k_n}((\widehat{\lambda}_{n,p^*+1},\ldots,\widehat{\lambda}_{n,d})^\top - \mathbf{1}_{d-p^*}),$$

gives that

$$\begin{aligned}\mathrm{BIC}_{k_n}(p) - \mathrm{BIC}_{k_n}(p^*) = {} & k_n \sum_{i=p^*+1}^{p} \log\left(1 + \frac{1}{\sqrt{k_n}}M_{n,i}\right) \\ & + k_n(d-1-p)\log\left(1 + \frac{1}{d-1-p}\sum_{j=p+1}^{d-1}\frac{1}{\sqrt{k_n}}M_{n,j}\right) \\ & - k_n(d-1-p^*)\log\left(1 + \frac{1}{d-1-p^*}\sum_{j=p^*+1}^{d-1}\frac{1}{\sqrt{k_n}}M_{n,j}\right) \\ & - \frac{\log(k_n)}{2}(d-p-2)(d-p+1) \\ & + \frac{\log(k_n)}{2}(d-p^*-2)(d-p^*+1).\end{aligned}$$

Furthermore, $\boldsymbol{M}_n = O_{\mathbb{P}}(1)$ due to Theorem 1 (b). Additionally, the Taylor expansion of the logarithm as $x \to 0$,

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

14679469, 0, Downloaded from https://onlinelibrary.wiley.com/doi/10.1111/sjos.70026 by Karlsruher Institut Für Technologie, Wiley Online Library on [22/10/2025]. See the Terms and Conditions (https://onlinelibrary.wiley.com/terms-and-conditions) on Wiley Online Library for rules of use; OA articles are governed by the applicable Creative Commons License

gives that

$$\begin{aligned}
&\mathrm{BIC}_{k_n}(p) - \mathrm{BIC}_{k_n}(p^*) \\
&= k_n \sum_{i=p^*+1}^{p} \left(\frac{1}{\sqrt{k_n}} M_{n,i} - \frac{1}{2} \frac{1}{k_n} M_{n,i}^2 + O_{\mathbb{P}}(k_n^{-3/2}) \right) \\
&\quad + k_n \left(\sum_{j=p+1}^{d-1} \frac{1}{\sqrt{k_n}} M_{n,j} - \frac{1}{2(d-1-p)} \left(\sum_{j=p+1}^{d-1} \frac{1}{\sqrt{k_n}} M_{n,j} \right)^2 + O_{\mathbb{P}}(k_n^{-3/2}) \right) \\
&\quad - k_n \left(\sum_{j=p^*+1}^{d-1} \frac{1}{\sqrt{k_n}} M_{n,j} - \frac{1}{2(d-1-p^*)} \left(\sum_{j=p^*+1}^{d-1} \frac{1}{\sqrt{k_n}} M_{n,j} \right)^2 + O_{\mathbb{P}}(k_n^{-3/2}) \right) \\
&\quad - \frac{\log(k_n)}{2}(d-p-2)(d-p+1) + \frac{\log(k_n)}{2}(d-p^*-2)(d-p^*+1) \\
&= -\frac{1}{2} \sum_{i=p^*+1}^{p} M_{n,i}^2 - \frac{1}{2(d-1-p)} \left(\sum_{j=p+1}^{d-1} M_{n,j} \right)^2 + \frac{1}{2(d-1-p^*)} \left(\sum_{j=p^*+1}^{d-1} M_{n,j} \right)^2 \\
&\quad - \frac{\log(k_n)}{2}(d-p-2)(d-p+1) + \frac{\log(k_n)}{2}(d-p^*-2)(d-p^*+1) \\
&\quad + O_{\mathbb{P}}(k_n^{-1/2}). \qquad \text{(B1)}
\end{aligned}$$

An application of Theorem 1 (b) gives then

$$\begin{aligned}
&-\frac{1}{2} \sum_{i=p^*+1}^{p} M_{n,i}^2 - \frac{1}{2(d-1-p)} \left(\sum_{j=p+1}^{d-1} M_{n,j} \right)^2 + \frac{1}{2(d-1-p^*)} \left(\sum_{j=p^*+1}^{d-1} M_{n,j} \right)^2 \\
&\xrightarrow{D} -\frac{1}{2} \sum_{i=p^*+1}^{p} M_i^2 - \frac{1}{2(d-1-p)} \left(\sum_{j=p+1}^{d-1} M_j \right)^2 + \frac{1}{2(d-1-p^*)} \left(\sum_{j=p^*+1}^{d-1} M_j \right)^2.
\end{aligned}$$

A division of (B1) 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. Hence, the assertion follows.

Step 2: Suppose $p < p^*$. Again by the definition of the BIC we receive

$$\begin{aligned}
\frac{\mathrm{BIC}_{k_n}(p) - \mathrm{BIC}_{k_n}(p^*)}{k_n} &= -\sum_{j=p+1}^{p^*} \log(\widehat{\lambda}_{n,j}) + (d-1-p) \log\left(\frac{1}{d-1-p} \sum_{j=p+1}^{d-1} \widehat{\lambda}_{n,j} \right) \\
&\quad - (d-1-p^*) \log\left(\frac{1}{d-1-p^*} \sum_{j=p^*+1}^{d-1} \widehat{\lambda}_{n,j} \right) \\
&\quad - \log(k_n) \frac{(d-p-2)(d-p+1) + (d-p^*-2)(d-p^*+1)}{2k_n}.
\end{aligned}$$

14679469, 0, Downloaded from https://onlinelibrary.wiley.com/doi/10.1111/sjos.70026 by Karlsruher Institut Für Technologie, Wiley Online Library on [22/10/2025]. See the Terms and Conditions (https://onlinelibrary.wiley.com/terms-and-conditions) on Wiley Online Library for rules of use; OA articles are governed by the applicable Creative Commons License

Due to Theorem 1 (a), $\widehat{\lambda}_{n,i} \xrightarrow{\mathbb{P}} \lambda_i$ for $i = 1,\dots,d-1$ holds and therefore,

$$\begin{aligned}\frac{\mathrm{BIC}_{k_n}(p) - \mathrm{BIC}_{k_n}(p^*)}{k_n} &\xrightarrow{\mathbb{P}} -\sum_{j=p+1}^{p^*} \log(\lambda_j) + (d-1-p)\log\left(\frac{1}{d-1-p}\sum_{j=p+1}^{d-1}\lambda_j\right)\\ &= -\sum_{j=p+1}^{d} \log(\lambda_j) + (d-1-p)\log\left(\frac{1}{d-1-p}\sum_{j=p+1}^{d-1}\lambda_j\right)\\ &= -\log\left(\frac{\prod_{j=p+1}^{d}\lambda_j}{\left(\frac{1}{d-1-p}\sum_{j=p+1}^{d-1}\lambda_j\right)^{(d-1-p)}}\right) > 0,\end{aligned}$$

due to the inequality of arithmetic and geometric means (Uchida, 2008), which says that

$$\frac{\left(\prod_{j=p+1}^{d}\lambda_j\right)^{1/(d-1-p)}}{\frac{1}{d-1-p}\sum_{j=p+1}^{d-1}\lambda_j} < 1.$$

■

Proof of Theorem 5. *Step 1*: Suppose $p > p^*$. Analogously to (B1), we receive

$$\begin{aligned}\mathrm{AIC}_{k_n}(p) - \mathrm{AIC}_{k_n}(p^*) = &-\frac{1}{2}\sum_{i=p^*+1}^{p} M_{n,i}^2 - \frac{1}{2(d-1-p)}\left(\sum_{j=p+1}^{d-1} M_{n,j}\right)^2\\ &+ \frac{1}{2(d-1-p^*)}\left(\sum_{j=p^*+1}^{d-1} M_{n,j}\right)^2 - (d-p-2)(d-p+1)\\ &+ (d-p^*-2)(d-p^*+1) + O_{\mathbb{P}}(k_n^{-1/2}).\end{aligned}$$

Again, an application of Theorem 1 (b) gives then

$$\begin{aligned}\mathrm{AIC}_{k_n}(p) - \mathrm{AIC}_{k_n}(p^*) \xrightarrow{\mathcal{D}} &-\frac{1}{2}\sum_{i=p^*+1}^{p} M_i^2 - \frac{1}{2(d-1-p)}\left(\sum_{j=p+1}^{d-1} M_j\right)^2\\ &+ \frac{1}{2(d-1-p^*)}\left(\sum_{j=p^*+1}^{d-1} M_j\right)^2\\ &- (d-p-2)(d-p+1) + (d-p^*-2)(d-p^*+1)\end{aligned}$$

which is the statement.

Step 2: Suppose $p < p^*$. Since as $n \to \infty$,

$$\begin{aligned}&\frac{\mathrm{AIC}_{k_n}(p) - \mathrm{AIC}_{k_n}(p^*)}{k_n} - \frac{\mathrm{BIC}_{k_n}(p) - \mathrm{BIC}_{k_n}(p^*)}{k_n}\\ &= \frac{2-\log(k_n)}{k_n}\left(\frac{(d-p^*-2)(d-p^*+1)}{2} - \frac{(d-p-2)(d-p+1)}{2}\right) \to 0,\end{aligned}$$

the statement follows from Theorem 5. ■

14679469, 0, Downloaded from https://onlinelibrary.wiley.com/doi/10.1111/sjos.70026 by Karlsruher Institut Für Technologie, Wiley Online Library on [22/10/2025]. See the Terms and Conditions (https://onlinelibrary.wiley.com/terms-and-conditions) on Wiley Online Library for rules of use; OA articles are governed by the applicable Creative Commons License

Proof of Remark 7(b). Under the assumptions, $\|\boldsymbol{\Gamma}^{(n)^{1/2}}\boldsymbol{V}^{(n)}\| = \|\boldsymbol{\Gamma}\boldsymbol{V}\| = \sqrt{9+4+4+1} = \sqrt{18}$ and $\boldsymbol{\Theta}^{(n)} = \boldsymbol{\Theta} = (3V_1, 2V_2, 2V_3, V_4)/\sqrt{18}$. Further $\mathbb{E}[\boldsymbol{\Theta}] = \mathbf{0}_4$ and $\boldsymbol{\Sigma} = \boldsymbol{\Gamma}/18$ hold, where the eigenvalues of $\boldsymbol{\Sigma}$ are $(1/2, 2/9, 2/9, 1/18)$ and the corresponding eigenvectors are the unit vectors $\boldsymbol{e}_1,\dots,\boldsymbol{e}_4 \in \mathbb{R}^4$. 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 1 (b) yields

$$\sqrt{k_n}\Big((\widehat{\lambda}_{n,2}, \widehat{\lambda}_{n,3}) - (\lambda, \lambda)\Big) \xrightarrow{\mathcal{D}} (M_2, M_3)$$

in $\mathbb{R}^2$ where (M_2, M_3) is the joint distribution of the decreasingly ordered nonzero 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 $\mathrm{vec}(\boldsymbol{S})$ follows a centered multivariate normal distribution with covariance $\mathrm{Cov}(\mathrm{vec}((\boldsymbol{\Theta} - \mathbb{E}[\boldsymbol{\Theta}])(\boldsymbol{\Theta} - \mathbb{E}[\boldsymbol{\Theta}])^\top))$ 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 $\mathbb{VAR}(S_{22}) = \mathbb{E}[\Theta_2^4] - (\mathbb{E}[\Theta_2^2])^2 = 0$ and $\mathbb{VAR}(S_{33}) = 0$, 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 nonzero 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 $\mathbb{VAR}(S_{23}) = \mathbb{E}[\Theta_2^2\Theta_3^2] = 1$ so that $\mathbb{P}(g_2(\boldsymbol{M}) < 0) > 0$. ∎

APPENDIX C. PROOFS OF SECTION 4

Proof of Theorem 6. Note, as stated in Remark 6, the information criteria are scale invariant and hence

$$\mathrm{AIC}^\circ_{k_n}(p_n; \widehat{\lambda}_{n,1},\dots,\widehat{\lambda}_{n,d_n-1}) =: \mathrm{AIC}^\circ_{k_n}(p_n) = \mathrm{AIC}^\circ_{k_n}(p_n; d_n\widehat{\lambda}_{n,1},\dots,d_n\widehat{\lambda}'_{n,d_n-1}).$$

Due to Theorem 2 for (a,b) and Theorem 3 for (c), the proof of Bai et al. (2018, Theorem 3.1) for $\widehat{\xi}_{n,1},\dots,\widehat{\xi}_{n,d_n-1}$ can be carried out step by step for $d_n\widehat{\lambda}_{n,1},\dots,d_n\widehat{\lambda}_{n,d_n-1}$.

14679469, 0, Downloaded from https://onlinelibrary.wiley.com/doi/10.1111/sjos.70026 by Karlsruher Institut Für Technologie, Wiley Online Library on [22/10/2025]. See the Terms and Conditions (https://onlinelibrary.wiley.com/terms-and-conditions) on Wiley Online Library for rules of use; OA articles are governed by the applicable Creative Commons License

The only difference is that there we have almost sure convergence and here we have convergence in probability. ∎

Proof of Theorem 7. 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 2 and Theorem 3, the proof of Bai et al. (2018, Theorem 3.2) for $\widehat{\xi}_{n,1},\dots,\widehat{\xi}_{n,d_n-1}$ can be carried out step by step for $d_n\widehat{\lambda}_{n,1},\dots,d_n\widehat{\lambda}_{n,d_n-1}$. ∎

Proof of Theorem 8. Due to the scale invariance of the AIC^*, Theorem 2 and Theorem 3, the proof of Bai et al. (2018, Theorem 3.3) for $\widehat{\xi}_{n,1},\dots,\widehat{\xi}_{n,d_n-1}$ can be carried out step by step for $d_n\widehat{\lambda}_{n,1},\dots,d_n\widehat{\lambda}_{n,d_n-1}$. ∎

Proof of Theorem 9. Due to the scale invariance of the BIC^*, $\log(d_n)/\log(k_n) \to 1$ as $n \to \infty$, Theorem 2 and Theorem 3, the proof of Bai et al. (2018, Theorem 3.4) for $\widehat{\xi}_{n,1},\dots,\widehat{\xi}_{n,d_n-1}$ can be carried out step by step for $d_n\widehat{\lambda}_{n,1},\dots,d_n\widehat{\lambda}_{n,d_n-1}$. ∎

APPENDIX D. PROOFS OF SECTION 5

Lemma 5. *Let*

$$\boldsymbol{\varepsilon}_d \sim \left|\mathcal{N}_d\left(\mathbf{0}_d, \frac{100}{d}\boldsymbol{I}_d\right)\right|,$$

where the absolute value is entry-wise. Then

$$\lim_{d\to\infty} \mathbb{VAR}(\|\boldsymbol{\varepsilon}_d\|) = 100/\sqrt{2}.$$

Proof. Indeed, since $\|\boldsymbol{\varepsilon}_d\|^2 \sim 100/d \cdot \chi^2_d$, where χ^2_d is a chi-square distribution with d degrees of freedom, the formula for the moments of a chi-square distribution (cf. Theorem 3.3.2 in Hogg et al., 2005) gives

$$\mathbb{VAR}(\|\boldsymbol{\varepsilon}_d\|) = \mathbb{E}[\|\boldsymbol{\varepsilon}_d\|^2] - (\mathbb{E}[\|\boldsymbol{\varepsilon}_d\|])^2 = \frac{100}{d}\left(d - \left(\frac{\sqrt{2}\Gamma((d+1)/2)}{\Gamma(d/2)}\right)^2\right).$$

Further by Gautschi's inequality (cf. Elezović et al., 2000, p. 1) we have

$$\left(\frac{d-1}{2}\right)^{1/2} \le \frac{\Gamma((d+1)/2)}{\Gamma(d/2)} \le \left(\frac{d-1}{2}+1\right)^{1/2}$$

and therefore

$$\begin{aligned}\sqrt{2}\cdot 100\frac{d-1}{2d} &= \frac{100}{d}\sqrt{2}\left(d - \frac{d-1}{2} - 1\right)\\ &\le \mathbb{VAR}(\|\boldsymbol{\varepsilon}_d\|) \le \frac{100}{d}\sqrt{2}\left(d - \frac{d-1}{2}\right) = \sqrt{2}\cdot 100\frac{d+1}{2d}.\end{aligned}$$

Letting $d \to \infty$ on the left and on the right-hand side gives the statement. ∎

14679469, 0, Downloaded from https://onlinelibrary.wiley.com/doi/10.1111/sjos.70026 by Karlsruher Institut Für Technologie, Wiley Online Library on [22/10/2025]. See the Terms and Conditions (https://onlinelibrary.wiley.com/terms-and-conditions) on Wiley Online Library for rules of use; OA articles are governed by the applicable Creative Commons License