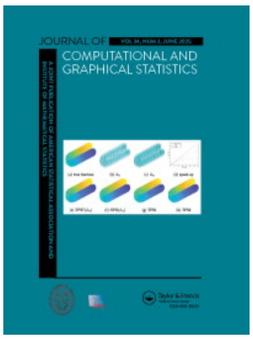

Taylor & Francis
Taylor & Francis Group

Journal of Computational and Graphical Statistics

ISSN: 1061-8600 (Print) 1537-2715 (Online) Journal homepage: www.tandfonline.com/journals/ucgs20

Order Determination for Tensor-Valued Observations Using Data Augmentation

Una Radojičić, Niko Lietzén, Klaus Nordhausen & Joni Virta

To cite this article: Una Radojičić, Niko Lietzén, Klaus Nordhausen & Joni Virta (24 Jun 2025): Order Determination for Tensor-Valued Observations Using Data Augmentation, Journal of Computational and Graphical Statistics, DOI: 10.1080/10618600.2025.2500977

To link to this article: https://doi.org/10.1080/10618600.2025.2500977

© 2025 The Author(s). Published with license by Taylor & Francis Group, LLC.

View supplementary material

Published online: 24 Jun 2025.

Submit your article to this journal

Article views: 324

View related articles

View Crossmark data

Full Terms & Conditions of access and use can be found at
https://www.tandfonline.com/action/journalInformation?journalCode=ucgs20

JOURNAL OF COMPUTATIONAL AND GRAPHICAL STATISTICS
2025, VOL. 00, NO. 0, 1–11
https://doi.org/10.1080/10618600.2025.2500977

Taylor & Francis
Taylor & Francis Group

OPEN ACCESS

Check for updates

Order Determination for Tensor-Valued Observations Using Data Augmentation

Una Radojičić[a], Niko Lietzén[b], Klaus Nordhausen[c,d], and Joni Virta[b]

[a]Institute of Statistics and Mathematical Methods in Economics, Vienna University of Technology, Vienna, Austria; [b]Department of Mathematics and Statistics, University of Turku, Turku, Finland; [c]Department of Mathematics and Statistics, University of Helsinki, Helsinki, Finland; [d]Department of Mathematics and Statistics, University of Jyväskylä, Jyväskylä, Finland

ABSTRACT

Tensor-valued data benefit greatly from dimension reduction as the reduction in size is exponential in the number of modes. To achieve maximal reduction without loss of information, our objective in this work is to provide an automated procedure for the optimal selection of reduced dimensionality. Our approach combines a recently proposed data augmentation procedure with the higher-order singular value decomposition (HOSVD) in a tensorially natural way. We give theoretical guidelines on how to choose the tuning parameters and further inspect their influence in a simulation study. As our primary result, we show that the procedure consistently estimates the true latent dimensions under a noisy tensor model, both at the population and sample levels. Additionally, we propose a bootstrap-based alternative to the augmentation estimator. Simulations are used to demonstrate the estimation accuracy of the two methods under various settings. Supplementary materials for this article are available online.

ARTICLE HISTORY
Received October 2023
Accepted April 2025

KEYWORDS
Augmentation; Dimension reduction; HOSVD; Order determination; Scree plot

1. Introduction

1.1. Tensors and Image Data

Tensors offer a convenient framework for the modeling of different types of image data (Aja-Fernández et al. 2009; Jouni 2021). For example, a collection of gray-scale images can be viewed as a sample of second-order tensors (i.e., matrices) where the tensor elements correspond to the intensities of the pixels. A set of color images can be represented as a sample of third-order tensors where the third mode has dimensionality equal to 3 and collects the color information through RGB values of the pixels. Still, higher-dimensional tensors are obtained if we sample multiple images per subject (possibly in the form of a video).

The number of elements in a tensor grows exponentially in its order. Therefore, as soon as our images have even a reasonably large resolution, the resulting data consists of a huge number of variables. Consider, for instance, the *butterfly* dataset[1] that will serve as our running example throughout the article; see the top row of Figure 1 for five images in the dataset. For simplicity, we use in this article a subsample of the full data, consisting of $n = 882$ RGB images of butterflies of various species with the resolution 224×224. As such, the number of variables per image is 150,528. However, as any two neighboring pixels in an image are typically highly correlated, the "signal dimension" of the butterfly data, and image data in general, is likely much smaller than the total number of observed variables.

The standard approach to remove the redundant information and noise from the data is through low-rank decomposition where the images are approximated as products of low-rank tensors (Aja-Fernández et al. 2009; Inoue 2016). As with any dimension reduction method, a key problem in this procedure is the selection of the reduced rank/order/dimension (we use all three terms interchangeably in the sequel). We want to keep the order small not to incorporate noise in the decomposition but, at the same time, enough components should be included to ensure accurate capturing of the signal. As the main contribution of the current work, we develop a method for the automatic determination of the order, separately in each mode. Reconstructions of the five butterfly images based on the dimensionalities chosen by our method are shown in the bottom row of Figure 1. For further details on this example, see Section 5.2. We develop our method under a specific statistical model and show that asymptotically, when the number of images n generated from the model grows without bounds, we are guaranteed to recover the true rank of the data.

Our main idea is based on the combination of a tensor decomposition known as *higher-order singular valued decomposition* (HOSVD) (De Lathauwer, De Moor, and Vandewalle 2000) and a method of order determination known as *predictor augmentation* (Luo and Li 2021). To set up the framework, the next two subsections briefly review the relevant literature, and, afterward, in Section 1.4 we highlight the contributions of this work in comparison to the literature. In the same section, we also discuss the connection to the conference paper Radojičić et al. (2021), of which this work is an extended version.

CONTACT Una Radojičić *una.radojicic@tuwien.ac.at* Institute of Statistics & Mathematical Methods in Economics, Vienna University of Technology, Wiedner Hauptstr. 8-10, 1040 Vienna, Austria.

Supplementary materials for this article are available online. Please go to *www.tandfonline.com/r/JCGS*.

[1] Available at *https://www.kaggle.com/datasets/gpiosenka/butterfly-images40-species*

© 2025 The Author(s). Published with license by Taylor & Francis Group, LLC.
This is an Open Access article distributed under the terms of the Creative Commons Attribution License (*http://creativecommons.org/licenses/by/4.0/*), which permits unrestricted use, distribution, and reproduction in any medium, provided the original work is properly cited. The terms on which this article has been published allow the posting of the Accepted Manuscript in a repository by the author(s) or with their consent.

Figure 1. Five images from the *butterfly* dataset (top row) and the corresponding reconstructed images using compressed, $103 \times 111 \times 3$-dimensional core tensors (bottom row).

1.2. Tensor Decompositions

The term tensor decomposition refers to a class of algorithms that approximate a tensor as a sum or product of tensors/matrices of lower rank. The low-rank structure is usually thought to represent the signal/information of the original data tensor, whereas, if we choose the rank properly, the difference between the original tensor and its low-rank approximation is pure noise. A comprehensive review of tensor decompositions is given by Kolda and Bader (2009) and, for some recent uses of tensor decompositions in the context of image data, see Inoue (2016), Zhang et al. (2020), and Lou and Cheung (2019).

Our decomposition of choice in this work is the higher-order singular value decomposition (De Lathauwer, De Moor, and Vandewalle 2000). In HOSVD, an mth order tensor $\mathcal{A} \in \mathbb{R}^{p_1 \times \cdots \times p_m}$ is approximated as $\mathcal{A} \approx \mathcal{B} \times_{i=1}^{m} \mathbf{U}_i$, a multilinear product of a low-dimensional *core tensor* $\mathcal{B}$ and the matrices $\mathbf{U}_1, \ldots, \mathbf{U}_m$ having orthonormal columns. The matrices $\mathbf{U}_i$ are in HOSVD estimated through SVDs of flattenings of the tensor $\mathcal{A}$, in turn allowing the estimation of the core. In practice, the optimal dimensions of the core tensor are usually not known a priori and have to be estimated, which is the main objective of the current work. We discuss HOSVD more closely in Section 2 in conjunction with our model of choice, using a specific "statistical" version of HOSVD known as $(2D)^2$PCA (Zhang and Zhou 2005).

Besides HOSVD, another classical tensor decomposition is the Tucker decomposition (Tucker 1966), which also seeks an approximation of the previous form but uses a different algorithm for estimating the loading matrices and the core. However, here we have chosen to work with HOSVD as it has a closed-form solution, which allows studying the theoretical properties of our proposed estimator of the latent dimensionality.

Finally, we note that the standard way of decomposing images to low-rank components is principal component analysis (PCA) (Jolliffe 2002). However, as standard PCA operates on vector data and not tensors, this approach requires the vectorization of the data, causing them to lose their natural structure. Whereas, tensor decompositions retain the row-column structure of image data, making them a natural choice in the current context.

1.3. Order Determination

The problem of choosing the (in some sense) optimal rank in tensor decompositions, or in dimension reduction in general, is known as order determination.

The order determination literature can be roughly divided into two: (a) methods targeted for specific parametric/semi-parametric models and, (b) more general methods that estimate the rank of a matrix from its noisy estimate (under some regularity conditions).

Category (1) is by far the more popular one and usually achieves the estimation using the asymptotic properties of eigenvalues of certain matrices, see, for example, Schott (2006) and Nordhausen, Oja, and Tyler (2022) for PCA and Bura and Yang (2011) and Zhu, Miao, and (2006) for dimension reduction in the context of regression (sufficient dimension reduction). Whereas, methods belonging to category (2) tend to be based on bootstrapping and related procedures, see Ye and Weiss (2003) and Luo and Li (2016) and, in particular, Luo and Li (2021) whose augmentation procedure serves as a starting point for the current work.

We note that all of the previous references targeted exclusively vector-valued data and order determination in the context of tensor-valued data is still rare in the literature. Moreover, aside from Radojičić et al. (2021), to our best knowledge, automated order determination in the context of matrix-variate (gray-scale image) data has been studied only by Tu, Huang, and Hsieh (2019) who use Stein's unbiased risk estimation for the task. Also, a simpler problem of selecting the dimension when the amount of retained variance is pre-determined is discussed in Hung et al. (2012). Thus, there is still much work to do in the order determination of tensor data and the current work is a step toward this direction.

1.4. Relation to Previous Work

In this work, we propose an automatic procedure for the selection of the dimensionality in the HOSVD decomposition of tensorial data. Our primary contributions in relation to earlier literature are the following.

(i) To our best knowledge, ours is the first-order determination procedure for general tensor-valued data with statistical guarantees.
(ii) Unlike Luo and Li (2021) who introduced the augmentation procedure in vector-valued context, we prove the validity of the procedure also on the population level, see Corollary 1.
(iii) We derive the full limiting distribution for the norms of augmented parts of the noise eigenvectors, see Theorem 2. This is in strict contrast to Luo and Li (2021, sec. 6) who, in an analogous context (PCA) for vector-valued data, based their work on the weaker claim that the norms are non-negligible in probability. This improvement both allows us to get quantitative results concerning the eigenvectors and sheds some light on the interplay of the augmentation procedure with the true data dimension.
(iv) To accompany the augmentation estimator, we also propose an alternative estimator of the latent dimensionality based on the "ladle" procedure (Luo and Li 2016).

Compared to the conference paper Radojičić et al. (2021), which presented versions of the augmentation and ladle procedures for matrix-valued data and of which the current work is an extended version, we go beyond them in the following respects:

(i) We allow for general tensor-valued data, not just matrices.
(ii) We give estimation guarantees for the augmentation both on the population and the sample level (unlike Radojičić et al. (2021) who did not consider theoretical properties).

1.5. Organization of the Article

The rest of the article is organized as follows. In Section 2 we introduce the statistical framework along with HOSVD for the proposed model. The proposed augmentation estimator as well as its theoretical properties are discussed in Section 3, separately on the population and sample levels. Section 4 covers the alternative, bootstrap-based ladle estimator. We conclude by evaluating the performances of the proposed estimators in Section 5 and by discussing future work in Section 6. A summary of tensor notation is given in Appendix A and the proofs of all technical results are collected in Appendix B in the supplement.

2. Model

We use standard tensor notation throughout the manuscript: the calligraphy font $\mathcal{A}$ refers to individual tensors, the k-unfolding of a tensor $\mathcal{A}$ is denoted by $\mathcal{A}_k$ and the k-mode multiplication of a tensor $\mathcal{A}$ by a matrix $\mathbf{A}_k$ is denoted by $\mathcal{A} \times_k \mathbf{A}_k$, etc. For readers unfamiliar with tensor notation a summary is given in Section A of the supplement; see also Kolda and Bader (2009).

We start by defining an appropriate statistical framework. Let $\mathcal{X}^1, \ldots, \mathcal{X}^n \in \mathbb{R}^{p_1 \times \cdots \times p_m}$ be an observed set of tensors of order $m \in \mathbb{N}$ drawn independently from the model

$$\mathcal{X} = \mathcal{M} + \mathcal{Z} \times_{k=1}^{m} \mathbf{U}_k + \mathcal{E}, \tag{1}$$

where $\mathcal{M} \in \mathbb{R}^{p_1 \times \cdots \times p_m}$ is the mean tensor, $\mathbf{U}_k \in \mathbb{R}^{p_k \times d_k}$, $k = 1, \ldots, m$, are unknown mixing matrices with orthonormal columns and $\mathcal{Z} \in \mathbb{R}^{d_1 \times \cdots \times d_m}$ is a *core tensor* of order m with zero mean, finite second moment, $\mathbb{E}(\|\mathcal{X}\|_F^2) < \infty$, and dimensions $d_k \leq p_k$, $k = 1, \ldots, m$. Furthermore, the additive noise $\mathcal{E}$ is assumed to follow a tensor spherical distribution, that is, $\mathcal{E} \times_{k=1}^{m} \mathbf{V}_k \sim \mathcal{E}$, for all orthogonal matrices $\mathbf{V}_k \in \mathbb{R}^{p_k \times p_k}$, $k = 1, \ldots, m$, where the symbol $\sim$ means "is equal in distribution to". Additionally, we make the assumption that for all k-flattenings $\mathcal{Z}_k \in \mathbb{R}^{d_k \times \prod_{j \neq k} d_j}$ of $\mathcal{Z}$ the matrix $\mathbb{E}(\mathcal{Z}_k \mathcal{Z}_k') \in \mathbb{R}^{d_k \times d_k}$ is positive definite. This assumption is made precisely for the identifiability of the latent dimensions $(d_1, \ldots, d_m)$. A schematic representation of the model for $m = 3$ is given in Figure S1 in the Supplement. A special case of Model 1 for $m = 2$ was studied in Radojičić et al. (2021).

Our proposed approach is based on a statistical formulation of HOSVD known as (2D)^{2}PCA (Zhang and Zhou 2005). For this, consider the k-flattenings of Model 1

$$\mathcal{X}_k = \mathcal{M}_k + \mathbf{U}_k \mathcal{Z}_k (\mathbf{U}_{-k}^{\otimes})' + \mathcal{E}_k, \tag{2}$$

where for $k = 1, \ldots, m$, $\mathcal{X}_k, \mathcal{M}_k, \mathcal{E}_k \in \mathbb{R}^{p_k \times \rho_k}$, $\rho_k = \prod_{i \neq k} p_i$, $\mathcal{Z}_k \in \mathbb{R}^{d_k \times \prod_{i \neq k} d_i}$ and $\mathbf{U}_{-k}^{\otimes} := \mathbf{U}_{k+1} \otimes \cdots \otimes \mathbf{U}_m \otimes \mathbf{U}_1 \otimes \cdots \otimes \mathbf{U}_{k-1}$. Then, the HOSVD solution to Model 1 is $(\mathcal{X} - \mathcal{M}) \times_{k=1}^{m} \mathbf{V}_k'$, where the columns of $\mathbf{V}_k \in \mathbb{R}^{p_k \times d_k}$, $k = 1, \ldots, m$, are the first d_k eigenvectors of the matrix $\mathbb{E}\{(\mathcal{X}_k - \mathcal{M}_k)(\mathcal{X}_k - \mathcal{M}_k)'\}$. If, for a fixed k, the matrix $\mathbb{E}(\mathcal{Z}_k \mathcal{Z}_k')$ is a diagonal matrix with distinct diagonal elements, then $\mathbf{V}_k$ equals the mixing matrix $\mathbf{U}_k$ up to a permutation and sign change of columns. However, even if that is not the case, the column space of $\mathbf{V}_k$ coincides with that of $\mathbf{U}_k$.

It is worth mentioning that even though Model 1 assumes that the core tensor $\mathcal{Z}$ is mixed by matrices $\mathbf{U}_1, \ldots, \mathbf{U}_m$ with *orthonormal* columns, this is without loss of generality since the singular values and the right singular vectors of a general mixing matrix can be absorbed into the core. This, however, has the drawback of making the core tensor identifiable only up to multiplications by invertible matrices from each mode. Nevertheless, the latent dimensionalities d_k are identifiable in this case as well, and therefore we tolerate this ambiguity. Furthermore, this reveals why the proposed method is also a valid pre-processing step for computationally more involved linear feature extraction procedures.

Having estimated (in the previous sense) the parameters $\mathbf{U}_k$, we next develop our augmentation estimator for the latent dimensionality.

3. Augmentation Estimator

3.1. Population-Level Methodology

Next, we describe our basic strategy behind estimating the latent dimensions $d_1, \ldots, d_m$ using data augmentation. The approach can be seen as a tensorial extension of the method introduced by Luo and Li (2021). Let $\mathcal{X}_k$ be the k-flattening (2). Then, as $\mathbf{U}_{-k}^{\otimes}$ has orthonormal columns, we have

$$\mathbb{E}\{(\mathcal{X}_k - \mathcal{M}_k)(\mathcal{X}_k - \mathcal{M}_k)'\} = \mathbf{U}_k \mathbb{E}(\mathcal{Z}_k \mathcal{Z}_k') \mathbf{U}_k' + \mathbb{E}(\mathcal{E}_k \mathcal{E}_k').$$

Lemma 1. Let $\mathcal{E} \in \mathbb{R}^{p_1 \times \cdots \times p_m}$ be a random tensor with tensor spherical distribution. Then, for $k = 1, \ldots, m$, $\mathbb{E}(\mathcal{E}_k \mathcal{E}_k') = \sigma_k^2 \mathbf{I}_{p_k}$, for some $\sigma_k^2 > 0$, where $\mathcal{E}_k$ is the k-flattening of $\mathcal{E}$.

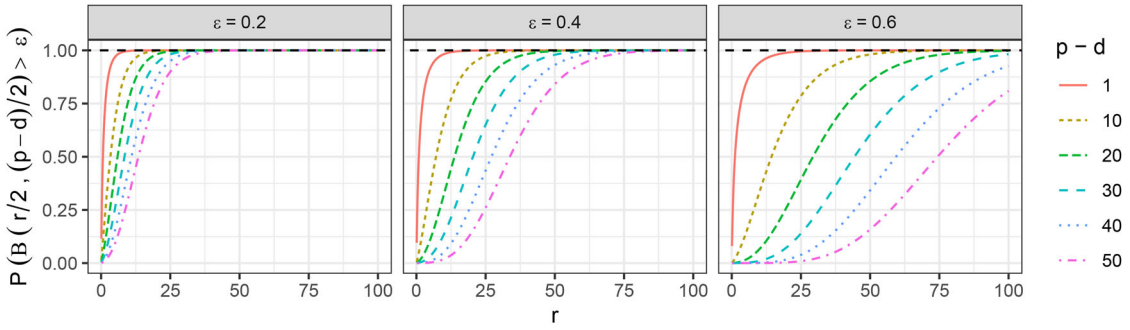

Figure 2. The curves represent the probability that a random variable from Beta$(r/2, (p-d)/2)$-distribution takes a value larger than $\varepsilon > 0$, as a function of the parameter r, for various values of $p - d$ and ε.

Using Lemma 1 we obtain that $\mathbb{E}\{(\mathcal{X}_k - \mathcal{M}_k)(\mathcal{X}_k - \mathcal{M}_k)'\} = \mathbf{U}_k\mathbb{E}(\mathcal{Z}_k\mathcal{Z}_k')\mathbf{U}_k' + \sigma_k^2\mathbf{I}_{p_k}$, and, since $\operatorname{rank}(\mathbf{U}_k\mathbb{E}(\mathcal{Z}_k\mathcal{Z}_k')\mathbf{U}_k') = d_k$, the problem of estimating d_k boils down to the problem of estimating the rank of the matrix $\mathbb{E}\{(\mathcal{X}_k-\mathcal{M}_k)(\mathcal{X}_k-\mathcal{M}_k)'\}-\sigma_k^2\mathbf{I}_{p_k}$. A naive way would be to inspect the scree plot of the eigenvalues of the sample estimate of $\mathbb{E}\{(\mathcal{X}_k - \mathcal{M}_k)(\mathcal{X}_k - \mathcal{M}_k)'\}$ and search for an *elbow*, a point at which the eigenvalues start to even off. However, it is often very difficult and subjective to find such a point. Therefore, we supplement the scree plot with additional information extracted from the eigenvectors of an appropriately constructed matrix, a task in which we employ the augmentation technique demonstrated in Radojičić et al. (2021), and initially introduced by Luo and Li (2021) for vectors.

More precisely, for fixed $k = 1, \ldots, m$ and $r_k \in \mathbb{N}$, we define $\mathbf{X}_S \in \mathbb{R}^{r_k \times \rho_k}$ to be a random matrix with iid entries having the distribution $\mathcal{N}(0, \sigma_k^2/\rho_k)$, implying that $\mathbb{E}(\mathbf{X}_S) = \mathbf{0}$ and $\mathbb{E}(\mathbf{X}_S\mathbf{X}_S') = \sigma_k^2\mathbf{I}_{r_k}$. We then augment (concatenate) the centered k-flattening of $\mathcal{X}$ with $\mathbf{X}_S$ to obtain the $(p_k + r_k) \times \rho_k$ matrix $\mathbf{X}_k^* = ((\mathcal{X}_k - \mathcal{M}_k)', \mathbf{X}_S')'$ that satisfies,

$$\mathbb{E}\{\mathbf{X}_k^*(\mathbf{X}_k^*)'\} = \begin{pmatrix} \mathbf{U}_k\mathbb{E}(\mathcal{Z}_k\mathcal{Z}_k')\mathbf{U}_k' & \mathbf{0} \\ \mathbf{0} & \mathbf{0} \end{pmatrix} + \sigma_k^2\mathbf{I}_{p_k+r_k}. \tag{3}$$

We further define $\mathbf{M}_k^* := \mathbb{E}\{\mathbf{X}_k^*(\mathbf{X}_k^*)'\} - \sigma_k^2\mathbf{I}_{p_k+r_k}$. Then $\mathbf{M}_k^*$ and $\mathbb{E}(\mathcal{Z}_k\mathcal{Z}_k')$ both have the same rank d_k and also the same positive eigenvalues $\lambda_{k,1} \geq \lambda_{k,2} \geq \cdots \geq \lambda_{k,d_k} > 0$.

Let next $\boldsymbol{\beta}_{k,i}^* = (\boldsymbol{\beta}_{k,i}', \boldsymbol{\beta}_{k,i,S}')' \in \mathbb{R}^{p_k+r_k}$, $i = 1, \ldots, p_k + r_k$, be any eigenvector of $\mathbf{M}_k^*$ corresponding to its ith largest eigenvalue, where we call the r_k-dimensional subvector $\boldsymbol{\beta}_{k,i,S}$ its *augmented part*. Then, for $i \leq d_k$, we have $\mathbf{M}_k^*\boldsymbol{\beta}_{k,i}^* = (\boldsymbol{\beta}_{k,i}'\mathbf{U}_k\mathbb{E}(\mathcal{Z}_k\mathcal{Z}_k')\mathbf{U}_k', \mathbf{0}')' = \lambda_{k,i}(\boldsymbol{\beta}_{k,i}', \boldsymbol{\beta}_{k,i,S}')'$, implying that $\boldsymbol{\beta}_{k,i,S} = \mathbf{0}$ for $i = 1, \ldots d_k$. Not only does the equivalent fail for the eigenvectors belonging to a zero eigenvalue ($i > d_k$), but the following theorem shows that for $r_k \to \infty$, exactly the opposite happens.

Prior to stating the theorem, we discuss the form and the arbitrariness of the zero-eigenvalue eigenvectors of $\mathbf{M}_k^*$. Let $\mathbf{B}_{k,0}^* \in \mathbb{R}^{(p_k+r_k)\times(p_k+r_k-d_k)}$ denote a matrix that contains any orthonormal basis of the null space of $\mathbf{M}_k^*$ as its columns (the below result is invariant to the exact choice of this basis). Then, for $i > d_k$, any eigenvector $\boldsymbol{\beta}_{k,i}^*$ lies in the null space of $\mathbf{M}_k^*$ and is thus of the form $\boldsymbol{\beta}_{k,i}^* = \mathbf{B}_{k,0}^*\mathbf{a}$ for some unit length vector $\mathbf{a} \in \mathbb{R}^{p_k+r_k-d_k}$.

The following theorem then shows that the norm of the augmented part of a randomly chosen zero-eigenvalue eigenvector follows a specific beta distribution.

Theorem 1. Fix $i = (d_k + 1), \ldots, (p_k + r_k)$ and let $\boldsymbol{\beta}_{k,i}^* = (\boldsymbol{\beta}_{k,i}', \boldsymbol{\beta}_{k,i,S}')' \in \mathbb{R}^{p_k+r_k}$ be of the form $\boldsymbol{\beta}_{k,i}^* = \mathbf{B}_{k,0}^*\mathbf{a}$ where $\mathbf{a}$ is drawn uniformly from the unit sphere in $\mathbb{R}^{p_k+r_k-d_k}$. Then $\|\boldsymbol{\beta}_{k,i,S}\|^2 \sim \text{Beta}\{r_k/2, (p_k - d_k)/2\}$.

Figure 2 illustrates the tail probabilities of the augmented parts of randomly chosen eigenvectors (in the sense of Theorem 1) belonging to a zero eigenvalue, as a function of r_k. Note that, in practice, we would like the sample analogs of the quantities $\|\boldsymbol{\beta}_{k,i,S}\|^2$, $i = d_k + 1, \ldots, p_k$ to be as large as possible to be able to distinguish the transition from signal to noise. Based on Figure 2 this can be achieved by using large values of r_k. However, this matter turns out to be more complicated in a finite-sample case where increasing r_k with n held fixed might lead to high-dimensional phenomena, see Section 6.

Given the distributional result in Theorem 1, the following properties of the augmented parts of the null eigenvectors of $\mathbf{M}_k^*$ now straightforwardly follow.

Corollary 1. Under the conditions of Theorem 1, the following hold.

(i) Let ε_n be any sequence of positive real numbers such that $\varepsilon_n \to 0$ as $n \to \infty$. Then, $\mathbb{P}(\|\boldsymbol{\beta}_{k,i,S}\|^2 > \varepsilon_n) \to 1$ as $n \to \infty$.
(ii) For fixed p_k, d_k and for every $\varepsilon > 0$, $\mathbb{P}(\|\boldsymbol{\beta}_{k,i,S}\|^2 \geq 1-\varepsilon) \to 1$, as $r_k \to \infty$.
(iii) In a high-dimensional regime, if $p_k - d_k = o(r_k)$, then for every $\varepsilon > 0$, $\mathbb{P}(\|\boldsymbol{\beta}_{k,i,S}\|^2 \geq 1 - \varepsilon) \to 1$, as $r_k \to \infty$.

Corollary 1 indicates that for r_k large enough, the norms of the augmented parts of the zero-eigenvalue eigenvectors get arbitrary close to 1, thus, explaining (on the population level) the behavior observed in Radojičić et al. (2021), where, when the number of augmentations was increased, the function $\hat{f}_k$ that captures the information from the eigenvectors by accumulating the norms of their augmented parts acted as a linear function of the dimension j, for $j > d_1$; see Figure 5 in Radojičić et al. (2021) for more insight.

The previous properties of the augmented parts of the eigenvectors serve as the basis for the construction of the augmentation estimator. As the successful estimation of the noise variance σ_k^2 played a key role above, we therefore discuss it next in more detail.

3.2. Estimation of the Noise Variance

Let $\mathcal{X}^1, \ldots, \mathcal{X}^n$ be an iid sample from Model 1 and let $\bar{\mathcal{X}}$ be the corresponding sample mean. Let $\hat{\sigma}^2_{k,1} \geq \cdots \geq \hat{\sigma}^2_{k,p_k}$ be the eigenvalues of $(1/n)\sum_{i=1}^n (\mathcal{X}_{k,i}-\bar{\mathcal{X}}_k)(\mathcal{X}_{k,i}-\bar{\mathcal{X}}_k)'$ and denote by $\sigma^2_{k,1} \geq \cdots \geq \sigma^2_{k,p_k}$ the eigenvalues of $\mathbb{E}\left\{(\mathcal{X}_k - \mathcal{M}_k)(\mathcal{X}_k - \mathcal{M}_k)'\right\}$. Due to the independence of the signal and the noise, $\sigma^2_{k,i} = \lambda_{k,i} + \sigma^2_k$ for $i = 1, \ldots, d_k$, and $\sigma^2_{k,i} = \sigma^2_k$, for $i = d_k + 1, \ldots, p_k$, where $\lambda_{k,1}, \ldots, \lambda_{k,d_k}$ are the eigenvalues of $\mathbb{E}(\mathcal{Z}_k\mathcal{Z}_k')$. This, together with Corollary 2 in Section 3.4 implies how we can justly use $\hat{\sigma}^2_{k,d_k+1}, \ldots, \hat{\sigma}^2_{k,p_k}$ to construct a consistent estimator of the noise variance σ^2_k of the kth mode. However, since it is mostly the case that one wishes to estimate the latent dimensions in all modes, the following discussion allows us to construct a pooled estimator of the noise variance using all modes.

As $\mathcal{E}_k$ is left spherical for $k = 1, \ldots m$, we have $\mathbb{E}(\mathcal{E}_k\mathcal{E}_k') = \sigma^2_k \mathbf{I}_{p_k}$, $\sigma^2_k > 0$, and

$$\sigma^2_k = \mathbb{E}(\mathcal{E}_k\mathcal{E}_k')_{i,i} = \sum_{j=1}^{\rho_k} \mathbb{E}(\mathcal{E}^2_{k,(i,j)}). \tag{4}$$

If we sum the identity (4) over all $i = 1, \ldots, p_k$, we obtain $p_k\sigma^2_k = \sum_{i=1}^{p_k}\sum_{j=1}^{\rho_k}\mathbb{E}(\mathcal{E}^2_{k,(i,j)}) = \mathbb{E}\|\mathcal{E}\|^2_{\mathrm{F}}$, thus, implying the relationship

$$p_1\sigma^2_1 = p_2\sigma^2_2 = \cdots = p_m\sigma^2_m, \tag{5}$$

between the noise variances of the k-flattenings $\mathcal{E}_k$ of the noise tensor $\mathcal{E}$.

Define now $S_k := \{\frac{p_i}{p_k}\sigma^2_{i,j} : i = 1, \ldots, m, j = 1, \ldots, p_i\}$ to be the set of eigenvalues from all modes in the "scale" of the kth mode. Similarly, define $\hat{S}_k := \{\frac{p_i}{p_k}\hat{\sigma}^2_{i,j} : i = 1, \ldots, m, j = 1, \ldots, p_i\}$, to be the sample counterpart of S_k. Lemma 2 in Section 3.4 shows that under certain assumptions on the compressibility of the data, quantiles as well as means of suitable tails of $\hat{S}_k$ are consistent estimators of σ^2_k. Naturally, once the noise variance of the kth mode has been estimated, we obtain estimates $\hat{\sigma}^2_i$, $i \neq k$, for the noise variances of the remaining modes simply by scaling $\hat{\sigma}^2_i = (p_k/p_i)\hat{\sigma}^2_k$, $i \neq k$.

Remark 1. To further clarify the scaling constants p_i/p_k, $i = 1, \ldots, m$, used in the estimation of the noise variance in the kth mode, consider a scenario where the entries of $\mathcal{E}$ are uncorrelated with zero mean and variance $\delta^2 > 0$. Then, for $k = 1, \ldots, m$, $\mathrm{E}(\mathcal{E}_k\mathcal{E}_k') = \sum_{i=1}^{\rho_k}\delta^2\mathbf{I}_{p_k} = \rho_k\delta^2\mathbf{I}_{p_k}$, thus, showing that the noise variance accumulates with the number of columns.

3.3. Sample-Level Estimation

We are now equipped to define the augmentation estimator for the estimation of the kth latent dimension d_k. Let $\mathbf{X}_{1,S}, \ldots, \mathbf{X}_{n,S}$ be a sample of iid $r_k \times \rho_k$ matrices with elements drawn from the standard normal distribution $\mathcal{N}(0,1)$. We define the augmented k-flattenings of the observations $\mathcal{X}^1, \ldots, \mathcal{X}^n$ as the $(p_k+r_k)\times\rho_k$ matrices $\mathbf{X}^*_{i,k} := ((\mathcal{X}^i_k)', \hat{\sigma}_k\mathbf{X}'_{i,S})'$, $i = 1, \ldots, n$, where $\hat{\sigma}^2_k$ is any consistent estimator of the noise variance σ^2_k, see Lemma 2 in Section 3.4 for examples. Let further $\bar{\mathbf{X}}^*_k$ be the sample mean of the obtained augmented sample. A sample estimate $\hat{\mathbf{M}}^*_k$ of the matrix $\mathbf{M}^*_k$ is then

$$\hat{\mathbf{M}}^*_k = \frac{1}{n}\sum_{i=1}^{n}(\mathbf{X}^*_{i,k} - \bar{\mathbf{X}}^*_k)(\mathbf{X}^*_{i,k} - \bar{\mathbf{X}}^*_k)' - \hat{\sigma}^2_k\mathbf{I}_{p_k+r_k}, \tag{6}$$

whose first p_k eigenvectors we denote in the following by $\hat{\boldsymbol{\beta}}^*_{k,1}, \ldots, \hat{\boldsymbol{\beta}}^*_{k,p_k}$. Mimicking Luo and Li (2021) and Radojičić et al. (2021), we define the normalized scree plot curve,

$$\hat{\Phi}_k : \{0, 1, \ldots, p_k\} \to \mathbb{R}, \quad \hat{\Phi}_k(l) = \hat{\lambda}_{k,l+1} \Big/ \left(\sum_{i=1}^{l+1}\hat{\lambda}_{k,i} + 1\right),$$

where $(\hat{\lambda}_{k,1}, \ldots, \hat{\lambda}_{k,p_k}) := (\hat{\sigma}^2_{k,1} - \hat{\sigma}^2_k, \ldots, \hat{\sigma}^2_{k,p_k} - \hat{\sigma}^2_k)$, and we take $\hat{\lambda}_{k,p_k+1} := 0$. However, as the values $\hat{\sigma}^2_{k,i} - \hat{\sigma}^2_k$ are not necessarily nonnegative (unlike their population counterparts, the eigenvalues of $\mathbb{E}(\mathcal{Z}_k\mathcal{Z}_k')$), we suggest using $\hat{\lambda}_{k,i} = \max\{\hat{\sigma}^2_{k,i} - \hat{\sigma}^2_k, 0\}$, $i = 1, \ldots, p_k$, instead. In any case, one should proceed with caution as very negative values of $\hat{\sigma}^2_{k,i} - \hat{\sigma}^2_k$ indicate a possible overestimation of σ^2_k. Lemma 2 in Section 3.4 lists a number of consistent estimators of the noise variance with possibly different behaviors. Thus, in Remark 2 we discuss the effect of misestimation of the noise variance for the presented procedure.

Remark 2. Let $\mathbf{X}_S \in \mathbb{R}^{r_k\times\rho_k}$ be the augmented submatrix for the kth flattening $\mathcal{X}_k$, having independent $\mathcal{N}(0, \sigma^2_S/\rho_k)$-elements, where $\sigma^2_S > 0$ is now understood to be the (fixed) estimated value of σ^2_k. Furthermore, $\mathbf{M}^*_k$ equals

$$\mathbb{E}\{(\mathbf{X}^*_k - \mathbb{E}(\mathbf{X}^*_k))(\mathbf{X}^*_k - \mathbb{E}(\mathbf{X}^*_k))'\} - \sigma^2_S\mathbf{I}_{p_k+r_k} = \begin{pmatrix} \mathbf{U}_k\{\mathrm{E}(\mathcal{Z}_k\mathcal{Z}_k') + (\sigma^2_k - \sigma^2_S)\mathbf{I}_{p_k}\}\mathbf{U}_k' & \mathbf{0} \\ \mathbf{0} & \mathbf{0} \end{pmatrix},$$

and the eigenvalues of $\mathbf{M}^*_k$ are $\lambda_{k,i} + (\sigma^2_k - \sigma^2_S)$, $i = 1, \ldots, p_k$, where $\lambda_{k,i} = 0$ for $i > d_k$, in addition to the r_k zero eigenvalues corresponding to the lower right block. In practice, since the eigenvalues of $\mathbf{M}^*_k$ serve as the estimators of the eigenvalues of the positive-definite matrix $\mathbb{E}(\mathcal{Z}_k\mathcal{Z}_k')$, as discussed in Section 3.3, we replace $\lambda_{k,i} + (\sigma^2_k - \sigma^2_S)$ with $\max\{0, \lambda_{k,i} + (\sigma^2_k - \sigma^2_S)\}$, $i = 1, \ldots, p_k$, to avoid negative values. Let now $\sigma^2_S = \sigma^2_k + \delta$, where $0 \leq \delta < \lambda_{k,d_k}$ and $\delta > 0$ corresponds to the amount of overestimation of σ^2_k. Then, $\max\{0, \lambda_{k,i} + (\sigma^2_k - \sigma^2_S)\} = \lambda_{k,i} - \delta$, $i = 1, \ldots, d_k$, and $\max\{0, \lambda_{k,i} + (\sigma^2_k - \sigma^2_S)\} = 0$, for $i > d_k$, that is, the thresholding retains the rank d_k while shifting nontrivial eigenvalues by $-\delta$.

Remark 2 shows that the method is robust toward slight overestimation of the noise variance, and this is directly related to the thresholding of the eigenvalues of $\hat{\mathbf{M}}^*_k$ from below by 0. The "allowed" amount of overestimation equals the smallest nontrivial eigenvalue of $\mathbb{E}(\mathcal{Z}_k\mathcal{Z}_k')$. Though Remark 2 explains the effect of the overestimation of the noise variance at the population level, an approximation to the phenomenon holds in the sample case.

Moving back to the estimation of the order d_k, additional information about it can now be obtained by using the eigenvectors of $\mathbf{M}^*_k$. To reduce the effect of randomness in the augmentation, the augmentation procedure is conducted independently s_k times, and we compute the eigenvectors of $\hat{\mathbf{M}}^*_k$ for each replicate. For $j = 1, \ldots, s_k$, we denote by $\hat{\boldsymbol{\beta}}^j_{k,i,S}$ the augmented part of

Algorithm 1: Augmentation estimator for the dimension d_k of the kth mode

Input: $\mathcal{X}^1, \ldots, \mathcal{X}^n \in \mathbb{R}^{p_1 \times \cdots \times p_m}$, centered sample of tensors;

Set the row dimension $r_k > 0$;
Set the number of augmented replicates $s_k > 0$;
Calculate $\hat{\mathbf{M}}_k = \frac{1}{n}\sum_{i=1}^{n} \mathcal{X}_k^i \mathcal{X}_k^{i\prime}$, for $k = 1, \ldots, m$;
Calculate the estimate $\hat{\sigma}_k^2$ of the noise variance based on the pooled set of scaled eigenvalues of $\hat{\mathbf{M}}_k$, $\hat{S}_k = \{(p_i/p_k)\hat{\sigma}_{i,j_i}^2 : i = 1, \ldots, m, j_i = 1, \ldots, p_i\}$.
Compute $\hat{\lambda}_{k,i} = \max\{\hat{\sigma}_{k,i}^2 - \hat{\sigma}_k^2, 0\}$;
for $i \leftarrow 1$ **to** n **do**
  for $j \leftarrow 1$ **to** s_k **do**
    Generate an $r_k \times \rho_k$ matrix $\mathbf{X}_{i,S}^j$, with entries drawn iid from $\mathcal{N}(0,1)$;
    Define the augmented ith observation as $\mathbf{X}_i^{*j} = (\mathcal{X}_k^{i\prime}, \hat{\sigma}_k \mathbf{X}_{i,S}^{j\prime})'$;
for $j \leftarrow 1$ **to** s_k **do**
  Compute the eigendecomposition of the jth replicated matrix $\hat{\mathbf{M}}_k^{*j} = \frac{1}{n}\sum_{i=1}^{n} \mathbf{X}_i^{*j}\mathbf{X}_i^{*j\prime} - \hat{\sigma}_k^2 \mathbf{I}_{p_k+r_k}$.
  Let $\hat{\boldsymbol{\beta}}_{k,i,S}^j$ be the augmented part of the ith eigenvector of $\hat{\mathbf{M}}_k^{*j}$;
Compute $\hat{g}_k(j) = \hat{\Phi}_k(j) + \sum_{i=0}^{j} \hat{f}_k(i)$, where $\hat{\boldsymbol{\beta}}_{k,0,S}^j = \mathbf{0}$ and $\hat{\lambda}_{k,p_k+1} = 0$;
Return $\hat{d}_k = \operatorname{argmin}\{\hat{g}_k(i) : i = 0, \ldots, p_k\}$;

the ith eigenvector of the matrix $\hat{\mathbf{M}}_k^{*j}$ in the jth replicate. The eigenvector information is then captured by the function

$$\hat{f}_k : \{0, 1, \ldots, p_k\} \to \mathbb{R}, \quad \hat{f}_k(i) = \frac{1}{s_k}\sum_{j=1}^{s_k} \|\hat{\boldsymbol{\beta}}_{k,i,S}^j\|^2,$$

where $\hat{\boldsymbol{\beta}}_{k,0,S}^j := \mathbf{0}$. Finally, we combine the eigenvalue information captured by $\hat{\Phi}_k$ and the eigenvector information in $\hat{f}_k$ to form the final objective function $\hat{g}_k : \{0, 1, \ldots, p_k\} \to \mathbb{R}$,

$$\hat{g}_k(j) = \hat{\Phi}_k(j) + \sum_{i=0}^{j} \hat{f}_k(i), \tag{7}$$

whose minimizer $\hat{d}_k$ is taken to be the estimator of the latent dimension d_k. This definition of $\hat{d}_k$ is intuitively clear as, assuming that $d_k > 0$, for any $i < d_k$ the eigenvalue part $\hat{\Phi}_k(i)$ of (7) is large, while the eigenvector part $\hat{f}_k(i)$ is small. For $i > d_k$, the opposite happens and the eigenvalue part is small while the eigenvector part is large, due to Corollary 3 in Section 3.4. At the true dimension $i = d_k$ both parts are small, thus implying that the sum curve $\hat{g}_k$ is (at the population level) minimized precisely at $i = d_k$. In the extreme noise case where $d_k = 0$, the eigenvalue part in (7) is always negligible, while again due to Corollary 3, the eigenvector part is always large, except for $i = 0$, in which case it vanishes, causing the minimum to occur at $i = 0$. An algorithm for the augmentation estimator is given in Algorithm 1 and the augmentation process is visualized in Figure 3.

3.4. Asymptotic Properties of the Augmentation Estimator

Corollary 2 gives the asymptotic behavior of the eigenvalue part $\hat{\Phi}_k$ of $\hat{g}_k$, justifying its use.

Corollary 2. Let $\lambda_{k,i}$ and $\hat{\lambda}_{k,i}$, $i = 1, \ldots, p_k + r_k$ be the eigenvalues of $\mathbf{M}_k^*$ and $\hat{\mathbf{M}}_k^*$, respectively. Then, $\hat{\lambda}_{k,i} \to_P \lambda_{k,i}$ for $i = 1, \ldots, p_k + r_k$, as $n \to \infty$.

Based on Corollary 2, Lemma 2, lists a number of consistent estimators of σ_k^2.

Lemma 2. Let $\hat{\sigma}_{k,q}^2$, $q \in (0,1)$, be the qth sample quantile of $\hat{S}_k$ and $\bar{\sigma}_{k,q}^2$, $q \in (0,1)$, be the mean of those elements of $\hat{S}_k$ that are smaller than or equal to $\hat{\sigma}_{k,q}^2$.

(i) If $d_1 + \cdots + d_m < (1-q)(p_1 + \cdots + p_m)$, then $\hat{\sigma}_{k,q_1}^2$ and $\bar{\sigma}_{k,q_1}^2$, for any $q_1 \le q$, are consistent estimators of σ_k^2.
(ii) If $d_1 + \cdots + d_m < p_1 + \cdots + p_m$, then $\min\{\hat{S}_k\}$ is a consistent estimator of σ_k^2.

Theorem 2 below illustrates the behavior of the norms of the augmented parts of the eigenvectors on the sample level, under the normality of the additive noise $\mathcal{E}$, and shows that (i) for $i \le d_k$ the norms are negligible and (ii) this is not the case for $i > d_k$.

Assumption 1. The additive noise $\mathcal{E}$ in Model (1) has iid Gaussian entries.

Theorem 2. Let $\hat{\boldsymbol{\beta}}_{k,i}^* = (\hat{\boldsymbol{\beta}}_{k,i,1}', \hat{\boldsymbol{\beta}}_{k,i,S}')'$, $i = 1, \ldots, p_k + r_k$ be any set of eigenvectors of $\hat{\mathbf{M}}_k^*$, where $\hat{\boldsymbol{\beta}}_{k,i,S} \in \mathbb{R}^{r_k}$ is the augmented part of the ith eigenvector of $\hat{\mathbf{M}}_k^*$. Then,

(i) $\|\hat{\boldsymbol{\beta}}_{k,i,S}\|^2 = o_P(1)$, $i \le d_k$.
(ii) If also Assumption (1) holds, then $\|\hat{\boldsymbol{\beta}}_{k,i,S}\|^2 \rightsquigarrow \text{Beta}\{r_k/2, (p_k - d_k)/2\}$ for $i > d_k$.

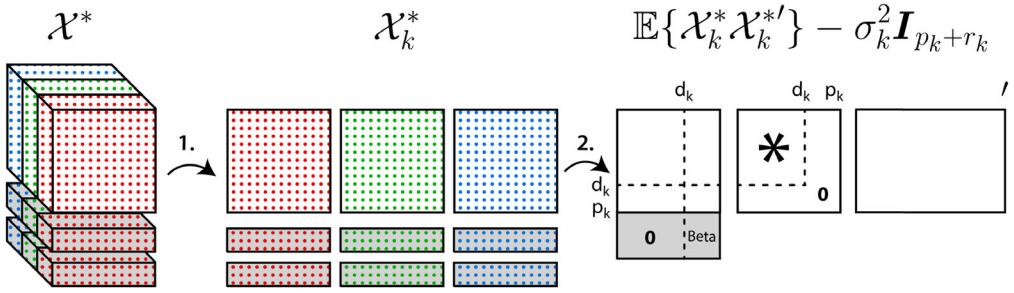

Figure 3. Representation of the augmentation process in the case $m = 3$. The parts of the tensors and matrices corresponding to the augmentation have a gray background. Step 1 represents the flattening along mode k and Step 2 the subsequent computation of the scatter in that mode along with the corresponding eigenvalue-eigenvector decomposition.

Corollary 3 further illustrates the behavior of the norms $\|\hat{\boldsymbol{\beta}}_{k,i,S}\|$ for $i > d_k$.

Corollary 3. Let $\hat{\mathbf{M}}_k^*$ be as defined above and let $\hat{\boldsymbol{\beta}}_{k,i}^* = (\hat{\boldsymbol{\beta}}_{k,i,1}', \hat{\boldsymbol{\beta}}_{k,i,S}')' \in \mathbb{R}^{p_k+r_k}$, $i = 1, \ldots, p_k + r_k$, be an eigenvector of $\hat{\mathbf{M}}_k^*$ corresponding to its ith eigenvalue. Then, under Assumption (1), for $i > d_k$ and for every $\varepsilon > 0$, $\lim_{\varepsilon \to 0^+} \mathbb{P}(\|\hat{\boldsymbol{\beta}}_{k,i,S}\|^2 > \varepsilon) \to 1$, as $n \to \infty$.

Interestingly, the limiting distribution of $\|\hat{\boldsymbol{\beta}}_{k,i,S}\|^2$, $i > d_k$, in Theorem 2 does not depend directly on the dimension p_k, but rather on the "amount of noise" $p_k - d_k$ in the kth mode. Thus, the more noise components there are in the kth mode, the more difficult it is to differentiate between the signal and the noise eigenvectors using the norms of the corresponding augmented parts; see Figure 2 for more insight.

Finally, the next theorem proves the validity of the method, in the sense of the consistency of the estimated dimensions, under the assumption of normal noise.

Theorem 3. Let Assumption (1) be satisfied and let $\hat{d}_k$ be the estimator of the unknown dimension d_k defined in (7). Then $\lim_{n\to\infty} \mathbb{P}(\hat{d}_k = d_k) = 1$, for $k = 1, \ldots, m$.

Remark 3. Inspecting the proof of Theorem 2 reveals that Assumption 1 on the normality of the noise tensor $\mathcal{E}$ ensures that the joint distribution of the noise part of the kth flattening $\mathcal{E}_k$ and the augmented noise is invariant under orthogonal transformations. This invariance allows us to deduce that the norm of the latter part of the zero-eigenvalue eigenvectors of the row-covariance matrix of $\mathbf{X}_k^*$ follows a beta distribution. In fact, (Luo and Li 2021, Theorem 1) implicitly assumes a similar condition, as discussed in (Luo and Li 2021, Lemma 3) and the subsequent discussion. In Theorem 2 of Luo and Li (2021), a related result is established under weaker assumptions. However, these weaker assumptions make it more challenging to determine the exact limiting distribution of the eigenvector norms—a critical piece of information for understanding the method's performance and tuning its parameters. Additional simulation results in Section D.1 in the supplement indicate that the method is rather robust against violations of Assumption 1.

4. Bootstrap-based Ladle Estimator

As a competitor to the augmentation strategy presented in Section 3.3, we introduce a generalization of the bootstrap-based "ladle"-technique for extracting information from the eigenvectors of the variation matrix presented in Luo and Li (2016) for vector-valued observations. The general idea is to use bootstrap resampling to approximate the variation of the span of the first k eigenvectors of the corresponding sample scatter matrix, where the high variation of the span indicates that the chosen eigenvectors belong to the same eigenspace, that is, that the difference between the corresponding eigenvalues is small, see Ye and Weiss (2003). The information from the eigenvectors is then combined with the one from the eigenvalues of the variation matrix, as in the augmentation estimator. As in Luo and Li (2016), we refer to this composite estimator as the bootstrap ladle estimator, since a ladle-shaped curve is obtained when the bootstrap ladle estimator is plotted as a function of the dimension. Extending the bootstrap ladle estimator of Luo and Li (2016) to tensor observations, we obtain estimators for the orders d_k, $k = 1, \ldots, m$, where the information contained in the eigenvalues is extracted similarly as in Algorithm 1.

More precisely, for centered, independent realizations $\{\mathcal{X}^1, \ldots, \mathcal{X}^n\}$ of a zero-mean tensor from Model (1), let $\hat{\mathbf{M}}_k$ and $\hat{\sigma}_{k,i}^2$, $k = 1, \ldots, m$, $i = 1, \ldots, p_k$, be defined as in Algorithm 1, and let $\hat{\mathbf{B}}_{j,k}$ be a matrix containing any first j eigenvectors of $\hat{\mathbf{M}}_k$. We define further $\hat{\phi}_{k,\mathrm{B}} : \{0, 1, \ldots, p_k - 1\} \to \mathbb{R}$, with $\hat{\phi}_{k,\mathrm{B}}(j) = \hat{\sigma}_{k,j+1}^2 / \left(\sum_{i=1}^{p_k-1} \hat{\sigma}_{k,i}^2 + 1\right)$, observing that the value of $\hat{\phi}_{k,\mathrm{B}}(j)$ is large for $j < d_k$ and small, but not zero, for $j \geq d_k$, for the reasons presented in Section 3.3.

As mentioned earlier, the eigenvalue information is in ladle supplemented with information taken from the eigenvectors of $\hat{\mathbf{M}}_k$ using a bootstrap technique. For the ith centered bootstrap sample, $\{\mathcal{X}_i^{1*}, \ldots, \mathcal{X}_i^{n*}\}$, let $\mathbf{M}_k^{i*} = \frac{1}{n}\sum_{j=1}^n \mathcal{X}_{i,k}^{j*}(\mathcal{X}_{i,k}^{j*})'$ be the scatter matrix of the k-flattening of the ith bootstrap sample. Furthermore, let $\mathbf{B}_{j,k}^{i*} = (\boldsymbol{\beta}_{1,k}^{i*}, \ldots, \boldsymbol{\beta}_{j,k}^{i*})$ be a matrix of any first j eigenvectors, belonging to the j largest eigenvalues of $\mathbf{M}_k^{i*}$, $j = 1, \ldots, p_k - 1$. Define further $\hat{f}_{k,\mathrm{B}} : \{0, 1, \ldots, p_k - 1\} \to \mathbb{R}$, with $\hat{f}_{k,\mathrm{B}}(0) := 0$ and $\hat{f}_{k,\mathrm{B}}(j) = \frac{1}{s_k}\sum_{i=1}^{s_k} \left(1 - |\det(\hat{\mathbf{B}}_{j,k}' \mathbf{B}_{j,k}^{i*})|\right)$, for $j > 0$, where s_k is the number of bootstrap samples. Note that $\hat{f}_{k,\mathrm{B}}(p_k) = 0$, since $\hat{\mathbf{B}}_{j,k}$ and $\mathbf{B}_{j,k}^{i*}$ both span the same space

$\mathbb{R}^{p_k}$. The bootstrap ladle estimator $\hat{d}_k$ of d_k is then defined as the minimizer of $\hat{g}_{k,\mathrm{B}} : \{0, 1, \ldots, p_k - 1\} \to \mathbb{R}$, where

$$\hat{g}_{k,\mathrm{B}}(j) = \hat{\phi}_{k,\mathrm{B}}(j) + \hat{f}_{k,\mathrm{B}}(j) / \left(\sum_{i=1}^{p_k - 1} \hat{f}_{k,\mathrm{B}}(i) + 1 \right),$$

see Algorithm S1 in Section C of the supplement for a detailed algorithm.

As the value of the objective function $\hat{g}_{k,\mathrm{B}}$ is not defined at p_k, it is assumed that compression is possible in every mode, which is different from the augmentation estimator where compression in at least one mode is compulsory. Luo and Li (2016) argue that if p_k is large (e.g., $p_k > 10$), the normalizing constants for the eigenvalue and eigenvector parts of the objective function should be replaced by $\sum_{i=1}^{\lfloor p_k / \log(p_k) \rfloor} \hat{\sigma}_{k,i}^2 + 1$ and $\sum_{i=1}^{\lfloor p_k / \log(p_k) \rfloor} \hat{f}_{k,\mathrm{B}}(i) + 1$, respectively. The reason is that if p_k is very large, the normalization constant of the eigenvector part of $\hat{g}_{k,\mathrm{B}}$ increases and weights down the bootstrap part compared to $\hat{\phi}_{k,\mathrm{B}}$. Therefore, it is usually beneficial to optimize $\hat{g}_{k,\mathrm{B}}$ only up to some $q_k < p_k$, and $q_k = \lfloor p_k / \log(p_k) \rfloor$ seems, according to Luo and Li (2016), justifiable in many applications.

A clear advantage of the bootstrap ladle estimator over the augmentation one is that no noise variance estimation is required and one can use the matrix $\hat{\mathbf{M}}_k$ directly to draw inference on the order d_k. However, as shown in a simulation study in Radojičić et al. (2021) for the case of matrix-valued data, the ladle estimator is computationally more demanding and less accurate. As both methods are based on flattening, the computational disadvantages translate directly also to the general tensor case. The accuracy differences will be discussed in the simulation study in Section 5.1, while, due to its computational complexity, the ladle estimator will be omitted from the real data analysis in Section 5.2. Further detailed intuition on the ladle estimator can be found in Luo and Li (2016).

5. Numerical Results

The analyses were conducted using R (R Core Team 2020) with the packages ICtest (Nordhausen et al. 2021), MixMatrix (Thompson 2019) and tensorBSS (Virta et al. 2021). To obtain the augmentation and ladle estimates of the latent dimensions we use Algorithms 1 and S1, respectively, which are available in the package tensorBSS (Virta et al. 2021).

5.1. Simulation Study

In the simulation study, the data are generated from the model

$$\mathcal{X} = \mathcal{Z} \times_{i=1}^{3} \mathbf{U}_i + \sigma \mathcal{E}, \tag{8}$$

where $\mathcal{E}$ has iid $\mathcal{N}(0, 1)$ entries and $\mathcal{Z} = \mathcal{Z}_0 \times_{i=1}^{3} \mathbf{A}_i$, for $\mathcal{Z}_0$ with iid $t(5)$ entries and regular matrices $\mathbf{A}_i \in \mathbb{R}^{d_i \times d_i}$, $i = 1, 2, 3$. $\mathbf{U}_i \in \mathbb{R}^{p_i \times d_i}$, $i = 1, 2, 3$, have orthonormal columns, where $d_1 = 3$, $d_2 = 5$, $d_3 = 10$, $p_1 = 5$, $p_2 = 15$, $p_3 = 20$.

We consider three different values for the noise variance $\sigma^2 = 0.1, 0.5, 1$. Thus,

$$\mathbb{E}(\mathcal{Z}_k \mathcal{Z}_k') = \tfrac{5}{3} \mathbf{A}_k \mathbf{A}_k' \prod_{i \neq k} \mathrm{tr}(\mathbf{A}_i \mathbf{A}_i'), \; \mathbb{E}(\mathcal{E}_k \mathcal{E}_k') = \sigma^2 \prod_{i \neq k} p_i \mathbf{I}_{p_k}. \tag{9}$$

Table 1. Eigenvalues of $\mathbb{E}(\mathcal{E}_k \mathcal{E}_k')$ and SNR for Model (8), $k = 1, 2, 3$.

	Eigenvalues			SNR		
σ^2	Mode 1	Mode 2	Mode 3	Mode 1	Mode 2	Mode 3
0.1	30	10	7.5	0.162	0.756	0.650
0.5	150	50	37.5	0.006	0.030	0.026
1.0	300	100	75.0	0.002	0.007	0.006

The eigenvalues of $\mathbb{E}(\mathcal{Z}_k \mathcal{Z}_k')$ are for mode 1 {5.75, 12.93, 22.99}, for mode 2 {5.39, 5.94, 8.41, 9.81, 12.12} and for mode 3 {2.74, 3.02, 3.31, 3.62, 3.94, 4.28, 4.63, 4.99, 5.37, 5.76}. The eigenvalues of $\mathbb{E}(\mathcal{E}_k \mathcal{E}_k')$, for $k = 1, 2, 3$ and $\sigma^2 \in \{0.1, 0.5, 1\}$, along with the resulting signal-to-noise ratios (SNR) for Model (8) are given in Table 1 where we define $\mathrm{SNR}_k = \|\mathbb{E}(\mathcal{Z}_k \mathcal{Z}_k')\|^2 / \|\mathbb{E}(\mathcal{E}_k \mathcal{E}_k')\|^2$ to be the ratio of the total variations of the signal and the noise components in the kth mode. Note that for a fixed mode $k = 1, 2, 3$, all eigenvalues of $\mathbb{E}(\mathcal{E}_k \mathcal{E}_k')$ are equal. The table shows that the SNR is quite small in all settings and the simulation settings are quite challenging.

For each of the 3 variance values, we simulate 1000 datasets of size $n = 1000$ from Model (8). The invertible matrices specifying the covariance structure of the core are of the form $\mathbf{A}_i = \mathbf{W}_i \mathbf{D}_i \mathbf{W}_i'$, where $\mathbf{W}_i \in \mathbb{R}^{d_i \times d_i}$, $i = 1, \ldots, 3$, are randomly generated orthogonal matrices and $\mathbf{D}_1 \approx \mathrm{diag}(1.857, 2.785, 3.714)$, $\mathbf{D}_2 \approx \mathrm{diag}(1.797, 1.887, 2.247, 2.427, 2.696)$ and $\mathbf{D}_3 \approx 1.282\,\mathrm{diag}(1.05, 1.1, \ldots, 1.45)$. Furthermore, the mixing matrices $\mathbf{U}_i \in \mathbb{R}^{p_i \times d_i}$ are taken to be the first d_i columns of random $p_i \times p_i$ orthogonal matrices.

The augmentation estimator has two tuning parameters, s_k and r_k, for the estimation of the latent dimensions d_i. Additionally, one has to choose the estimator of the noise variance. It is natural to choose the number of replications s_k to be large to reduce variation. Based on the simulations in Radojičić et al. (2021), we choose $s_k = 50$, $k = 1, 2, 3$, as a compromise between stability and computation time and since the choice of s_k was in Radojičić et al. (2021) deemed less crucial than the choice of r_k. In the vector case, Luo and Li (2021) propose $r \approx p_1/5$ which in light of our results at the population level, see Corollary 1, seems insufficiently large. This effect is visualized in Figure 2 which also shows how, in practice, the choice of r_k might be guided by the difference between the signal dimension and the full data dimension. However, the optimal choice of r_k should be investigated further, most likely in the high-dimensional framework, and is beyond the scope of this article.

To cover a wide range of values, we consider in the simulation study for each mode the values $r_k \in \{1, 5, 10, 25, 50\}$. For the estimation of the noise variance in the matrix case, Radojičić et al. (2021) considered different estimators with the "largest" being the median estimator which also turned out to give the best performance. To evaluate if this is the case in the current scenario as well, we consider the quantile-based estimators, $\hat{\sigma}_{k,q}^2$ (see Lemma 2), for the quantiles $q \in \{0, 0.1, \ldots, 0.6\}$. As a competitor, we use the bootstrap-based ladle estimator described in Algorithm S1, with the number of bootstrap samples $s_k = 200$ and where we ignored the if-condition $p_k \leq 10$ and always used $q_k = p_k - 1$.

The results of the simulation study are presented in Figure 4, where, in each sub-figure, columns correspond to the row

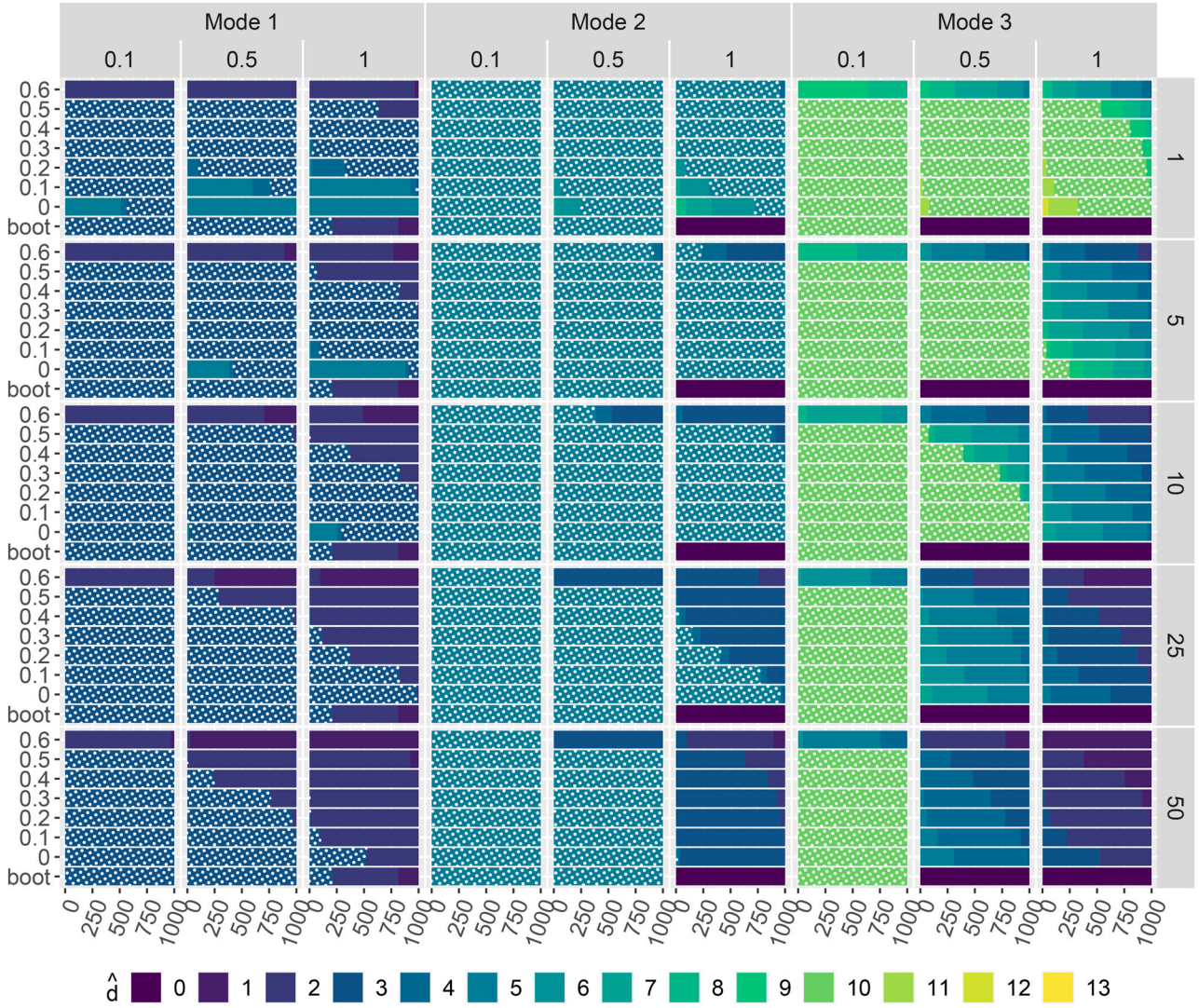

Figure 4. Frequencies of estimated latent dimensions in Model 8 based on 1000 repetitions, for different estimators: ladle (boot) and the augmentation based, where the numeric values on the *y*-axis give the quantile used for estimating noise variance $\hat{\sigma}^2_{k,q}$. The true latent dimensions are $d_1 = 3$, $d_2 = 5$, and $d_3 = 10$, and are always marked with a white dot pattern. For *k*th mode, panel rows correspond to r_k and columns to σ^2_k. Note that the estimates for ladle (method boot) are the same for all r_k.

dimensions r_k, $k = 1, 2, 3$, of the augmentation matrices, while the rows specify the standardized noise variance.

The results show that for the smallest value of the noise variance $\sigma^2 = 0.1$, the correct signal dimension is obtained in all modes and for all values of r_k if the quantile index in the noise variance estimation is not too extreme. However, when SNR is decreased the number of augmented components r_k and the quantile used for the noise variance estimation become of more relevance, especially for the modes with large values of p_k. Counter-intuitively to our theory, too large values of r_k and noise variance lead to underestimation of the signal dimension. Recall, however, that in these simulations the SNR is in all cases quite low showing that all estimators perform rather well. Nevertheless, we suggest not to use too drastic r_k, restricting to values such as $r_k = 10$ or $r_k = 25$ and using 0.2 or 0.3 quantile level for the noise variance estimation. The results also show that if p_k is small and the SNR is not too low, the ladle works very well, but it also suffers the most from deviations from these optimal settings and then always estimates the signal dimension as zero.

The disagreement between Corollary 1 and the simulation results with respect to increasing r_k, especially when p_k is large, might be due to the fact that our asymptotic results assume a fixed dimension and growing sample size n. Whereas, in practice, n is fixed and, therefore, if $p_k + r_k$ is relatively large, the estimation of $\sigma^2_1, \ldots, \sigma^2_m$ will be subject to the well-known high-dimensional bias (Johnstone and Paul 2018), quantified by the celebrated Marchenko–Pastur law (Götze and Tikhomirov 2004). Results of a higher-dimensional simulation study presented in Section D.2 of supplement align with the upper claim. However, since all of the $2m$ parameters, $p_1, \ldots, p_m, r_1, \ldots, r_m$, contribute to this bias, its theoretical analysis requires careful studying of their interactions and will as such be left for future research. Finally, additional simulation results where we explore the behavior of the method under violations of Assumption 1 are presented in the supplementary material, section D.1.

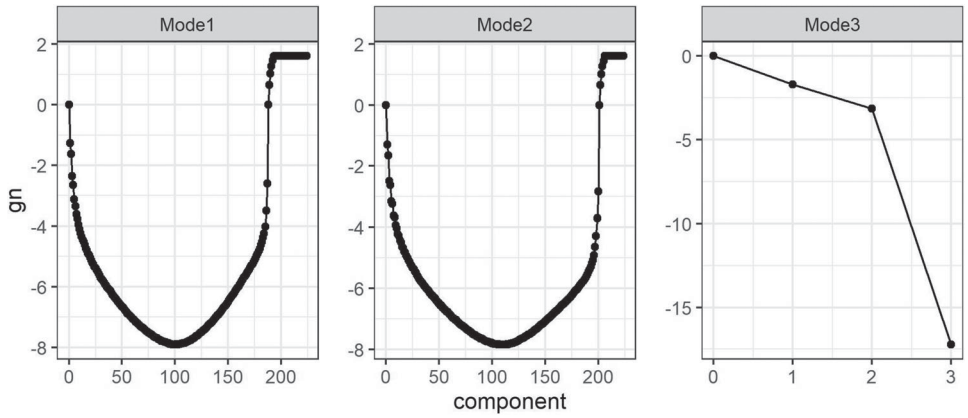

Figure 5. Logarithmized objective function $\hat{g}_k$ for the augmentation estimator using $r_k = 5, s_k = 50, k = 1, 2, 3$, calculated for the *butterfly* dataset.

5.2. Example

To illustrate the augmentation estimator we considered 882 224×224 RGB images of butterflies from various species, mostly taken in natural habitats. We assume that the SNR of the data is larger than in the previous simulation and therefore use $s_k = 50$, $r_k = 5$ and estimate the noise variance using the 30% quantile.

Based on these values, Figure 5 visualizes the graph of our augmentation objective function $\hat{g}_k$ in (7) for each mode on a logarithmic scale. The estimated signal dimension, obtained as the modewise minima, is (103, 111, 3). The fact that d_3 is estimated to be 3 indicates that there is a lot of information in the data contained in all three colors, a fact that indeed is plausible since butterflies are greatly characterized by the color of their wings. Figure 1 illustrates five selected original (top) and reconstructed (bottom) images of butterflies and differences between the two are indeed only visible when zooming in.

6. Discussion

Data that have a natural tensor representation, like images or video, are increasingly common and often have large dimensionality. In such cases, it is often assumed that the data are noisy and that a representation using smaller tensors is sufficient to capture the information content of the data. A key question is then to decide on the dimensions of these smaller signal tensors. We tackled this problem in the framework of a tensorial PCA, also known as HOSVD, and suggested an automated procedure for the order determination by extending the works of Luo and Li (2021) and Radojičić et al. (2021). The properties of the novel estimator were rigorously derived and its usefulness was demonstrated using simulated and real data.

We also observed that, for finite n, increasing the number r_k of augmented rows did not fully agree with the theory, which might be related to the Marchenko-Pastur law. This will be investigated in future research under a high-dimensional framework where we will also consider different norms when computing the objective criterion. Furthermore, we plan to derive also a hypothesis test that would allow inference about the signal tensor dimension.

Our theoretical results are based on assuming that the additive noise $\mathcal{E}$ follows a tensor spherical distribution. However, our results continue to hold also when this assumption is replaced with the following condition.

Assumption 2. The elements of the additive noise $\mathcal{E}$ are iid.

The proofs of our results under Assumption 2 are analogous to the current ones and have thus been omitted, but the main idea is that Assumption 2 preserves both the "low rank + identity" form in (3) and the noise variance identity (5), on which the theory is based on. Note that, as long as the Gaussianity Assumption 1 is retained, also Theorems 2, 3, and Corollary 3 trivially continue to hold under Assumption 2, since Assumption 2 covers also the tensor-variate standard normal distribution.

We further note that Assumption 2 is not actually weaker than the sphericity assumption, but rather *parallel* to it. That is, both assumptions involve a family of distributions satisfying one key property of the tensor-variate standard Gaussian distribution (orthogonal invariance or the iid-ness of the elements). In our view, the original sphericity assumption is the more natural one in the context of image data, since it allows the noises in two neighboring pixels to correlate and, as such, the main text is written from its viewpoint.

Finally, an alternative, natural way to conduct the augmentation would be to, instead of augmenting each mode at a time, simultaneously augment all m modes of the data. This would lead to augmented observations of size $(p_1+r_1) \times \cdots \times (p_m+r_m)$ whose flattenings could be used to compute the equivalents, say $\hat{\mathbf{M}}_{k,\text{alt}}$, of the sample estimates $\hat{\mathbf{M}}_k$ in (6). However, this augmentation approach is not advocated since, as shown by a simple calculation, the two matrices satisfy the relationship $\hat{\mathbf{M}}_{k,\text{alt}} = \hat{\mathbf{M}}_k + \mathbf{W}_k$ where $\mathbf{W}_k$ is a $(p_1 + r_1) \times (p_1 + r_1)$ Wishart noise matrix. Thus, this alternative augmentation strategy would amount to computing a noisy version of our preferred estimator, leading to lower finite-sample efficiency.

Thus far augmentation-based order determination has been developed for various multivariate methods predominantly in the vector setting (Luo and Li 2021; Radojičić and Nordhausen 2024). However, we anticipate that the augmentation procedure can be extended to apply also to other tensorial methods besides HOSVD, particularly when these methods operate mode-wise, as is the case with, for example, tensorial FOBI (Virta et al. 2017) and tensorial SIR (Ding and Cook 2015). These exten-

sions are expected to be more straightforward than the current work since, unlike these methods, PCA/HOSVD requires the estimation of a nuisance parameter (the noise variance). Details of these extensions will be explored in future research.

Supplementary Materials

Supplementary materials are available online and contain additional details on tensor notation, proofs, additional simulation results, and the R-script for reproducing the results of the Simulation study in Section 5.1, as well as the results shown in Section 5.2.

Acknowledgments

The authors would like to thank two anonymous referees, whose insightful comments greatly improved the quality of the article. The authors acknowledge TU Wien Bibliothek for financial support through its Open Access Funding Programme.

Disclosure Statement

No potential conflict of interest was reported by the authors.

Funding

The work of UR was supported by the Austrian Science Fund (Grant 10.55776/I5799). The work of NL, KN, and JV was supported by the Research Council of Finland (Grants 321968, 335077, 347501, 353769, 363261).

ORCID

Una Radojičić http://orcid.org/0000-0003-0329-0595
Klaus Nordhausen http://orcid.org/0000-0002-3758-8501

References

Aja-Fernández, S., de Luis Garcia, R., Tao, D., and Li, X. (2009), *Tensors in Image Processing and Computer Vision*, London: Springer. [1]

Bura, E., and Yang, J. (2011), "Dimension Estimation in Sufficient Dimension Reduction: A Unifying Approach," *Journal of Multivariate Analysis*, 102, 130–142. [2]

De Lathauwer, L., De Moor, B., and Vandewalle, J. (2000), "A Multilinear Singular Value Decomposition," *SIAM Journal on Matrix Analysis and Applications*, 21, 1253–1278. [1,2]

Ding, S., and Cook, R. D. (2015), "Tensor Sliced Inverse Regression," *Journal of Multivariate Analysis*, 133, 216–231. [10]

Götze, F., and Tikhomirov, A. (2004), "Rate of Convergence in Probability to the Marchenko-Pastur Law," *Bernoulli*, 10, 503–548. [9]

Hung, H., Wu, P., Tu, I., and Huang, S. (2012), "On Multilinear Principal Component Analysis of Order-Two Tensors," *Biometrika*, 99, 569–583. [2]

Inoue, K. (2016), *Generalized Tensor PCA and Its Applications to Image Analysis*, pp. 51–71, Tokyo: Springer. [1,2]

Johnstone, I. M., and Paul, D. (2018), "PCA in High Dimensions: An Orientation," *Proceedings of the IEEE*, 106, 1277–1292. [9]

Jolliffe, I. (2002), *Principal Component Analysis*, New York, NY: Springer. [2]

Jouni, M. (2021), "Image Analysis Based on Tensor Representations," Theses, Université Grenoble Alpes. [1]

Kolda, T., and Bader, B. (2009), "Tensor Decompositions and Applications," *SIAM Review*, 51, 455–500. [2,3]

Lou, J., and Cheung, Y.-M. (2019), "Robust Low-Rank Tensor Minimization via a New Tensor Spectral k-Support Norm," *IEEE Transactions on Image Processing*, 29, 2314–2327. [2]

Luo, W., and Li, B. (2016), "Combining Eigenvalues and Variation of Eigenvectors for Order Determination," *Biometrika*, 103, 875–887. [2,3,7,8]

——— (2021), "On Order Determination by Predictor Augmentation," *Biometrika*, 108, 557–574. [1,2,3,4,5,7,8,10]

Nordhausen, K., Oja, H., and Tyler, D. (2022), "Asymptotic and Bootstrap Tests for Subspace Dimension," *Journal of Multivariate Analysis*, 188, 104830. [2]

Nordhausen, K., Oja, H., Tyler, D. E., and Virta, J. (2021), *ICtest: Estimating and Testing the Number of Interesting Components in Linear Dimension Reduction*, R package version 0.3-3. [8]

R Core Team. (2020), *R: A Language and Environment for Statistical Computing*, Vienna, Austria: R Foundation for Statistical Computing. [8]

Radojičić, U., and Nordhausen, K. (2024), "Order Determination in Second-Order Source Separation Models Using Data Augmentation," in *Combining, Modelling and Analyzing Imprecision, Randomness and Dependence*, eds. J. Ansari, S. Fuchs, W. Trutschnig, M. A. Lubiano, M. Á. Gil, P. Grzegorzewski, and O. Hryniewicz, pp. 371–379, Cham: Springer. [10]

Radojičić, U., Lietzen, N., Nordhausen, K., and Virta, J. (2021), "Dimension Estimation in Two-Dimensional PCA," in *2021 12th International Symposium on Image and Signal Processing and Analysis (ISPA)*, pp. 16–22. [1,2,3,4,5,8,10]

Schott, J. R. (2006), "A High-Dimensional Test for the Equality of the Smallest Eigenvalues of a Covariance Matrix," *Journal of Multivariate Analysis*, 97, 827–843. [2]

Thompson, G. (2019), *MixMatrix: Classification with Matrix Variate Normal and t Distributions*, R package version 0.2.4. [8]

Tu, I.-P., Huang, S.-Y., and Hsieh, D.-N. (2019), "The Generalized Degrees of Freedom of Multilinear Principal Component Analysis," *Journal of Multivariate Analysis*, 173, 26–37. [2]

Tucker, L. R. (1966), "Some Mathematical Notes on Three-Mode Factor Analysis," *Psychometrika*, 31, 279–311. [2]

Virta, J., Koesner, C. L., Li, B., Nordhausen, K., Oja, H., and Radojicic, U. (2021), *tensorBSS: Blind Source Separation Methods for Tensor-Valued Observations*, R package version 0.3.8. [8]

Virta, J., Li, B., Nordhausen, K., and Oja, H. (2017), "Independent Component Analysis for Tensor-Valued Data," *Journal of Multivariate Analysis*, 162, 172–192. [10]

Ye, Z., and Weiss, R. E. (2003), "Using the Bootstrap to Select One of a New Class of Dimension Reduction Methods," *Journal of the American Statistical Association*, 98, 968–979. [2,7]

Zhang, D., and Zhou, Z.-H. (2005), "$(2D)^2$PCA: Two-Directional Two-Dimensional PCA for Efficient Face Representation and Recognition," *Neurocomputing*, 69, 224–231. [2,3]

Zhang, M., Gao, Y., Sun, C., and Blumenstein, M. (2020), "Robust Tensor Decomposition for Image Representation based on Generalized Correntropy," *IEEE Transactions on Image Processing*, 30, 150–162. [2]

Zhu, L., Miao, B., and Peng, H. (2006), "On Sliced Inverse Regression with High-Dimensional Covariates," *Journal of the American Statistical Association*, 101, 630–643. [2]