Pattern Recognition 138 (2023) 109401

Contents lists available at ScienceDirect

Pattern Recognition

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

Poisson PCA for matrix count data

Joni Virta[a,*], Andreas Artemiou[b]

[a] Department of Mathematics and Statistics, University of Turku, Finland
[b] School of Mathematics, Cardiff University, Wales, United Kingdom

ARTICLE INFO

Article history:
Received 28 October 2022
Revised 4 January 2023
Accepted 5 February 2023
Available online 6 February 2023

2020 MSC:
Primary 62H12
Secondary 62F12

Keywords:
Discrete data
Kronecker model
Matrix normal distribution
Poisson log-normal distribution

ABSTRACT

We develop a dimension reduction framework for data consisting of matrices of counts. Our model is based on the assumption of existence of a small amount of independent normal latent variables that drive the dependency structure of the observed data, and can be seen as the exact discrete analogue of a contaminated low-rank matrix normal model. We derive estimators for the model parameters and establish their limiting normality. An extension of a recent proposal from the literature is used to estimate the latent dimension of the model. The method is shown to outperform both its vectorization-based competitors and matrix methods assuming the continuity of the data distribution in analysing simulated data and real world abundance data.

1. Introduction

Modern applications typically see data with structures going significantly beyond the traditional "n observations of p continuous variables" framework. In this work, our focus is on data where the observations $X_i = (x_{i,jk})$, $i = 1, \dots, n$, $j = 1, \dots, p_1$, $k = 1, \dots, p_2$, are $p_1 \times p_2$ matrices of non-negative counts. Such data appear naturally, for example, in the analysis of publication data ($x_{i,jk}$ describes the word count of the jth word for the ith author at the kth venue) [13], abundance studies ($x_{i,jk}$ describes the abundance of the ith species in the jth location at the kth time peroid) [9] and in the analysis of dyadic events ($x_{i,jk}$ describes the number of actions initiated by the jth actor targeting the kth actor during the ith time period) [32]. A common aspect to all these applications is that the involved data sets are usually both large in size and inherently complex. As such, a natural first step in their analysis is dimension reduction, which both helps reduce their size and allows interpreting the data through the discovered latent variables. The development of a natural framework for the dimension reduction of matrix-valued count data is thus the objective of the current work.

In order for us to succeed in this task, the developed methods have to naturally accommodate the two main features of the data: the matrix structure and the discreteness of the observations. Ignoring the latter of these for a second, the recent decade has seen an ever-increasing amount of standard multivariate statistical methods being generalized to allow for matrix-valued data. The majority of these extensions uses the so-called *Kronecker approach* to modelling which we next exemplify in the simple context of a linear latent factor model.

Given a $p_1 \times p_2$ random matrix X, the "naive" approach to latent factor modeling would be to vectorize X to obtain the $(p_1 p_2)$-dimensional vector $\mathrm{vec}(X)$ (vec stacks the columns of its input to a long vector) and assume, for example, that

$$\mathrm{vec}(X) = \mu_0 + U_0 z + \varepsilon_0, \tag{1}$$

where $\mu_0 \in \mathbb{R}^{p_1 p_2}$, $U_0 \in \mathbb{R}^{p_1 p_2 \times d}$ are unknown parameters and the d-dimensional random vector z and the $(p_1 p_2)$-dimensional random vector ε_0 signify the latent signal and the noise, respectively. Whereas, under the Kronecker approach, we would instead preserve the matrix structure of X and assume that,

$$X = \mu + U_1 Z U_2^\top + \varepsilon, \tag{2}$$

where $\mu \in \mathbb{R}^{p_1 \times p_2}$, $U_1 \in \mathbb{R}^{p_1 \times d_1}$, $U_2 \in \mathbb{R}^{p_2 \times d_2}$ are unknown parameters and the random $d_1 \times d_2$ matrix Z and the random $p_1 \times p_2$ matrix ε represent the latent signal and the noise, respectively. To understand the relationship between the two approaches, we apply vectorization vec to the model (2) and use the formula $\mathrm{vec}(AYB^\top) = (B \otimes A)\mathrm{vec}(Y)$ to reveal that any X obeying the matrix model (2) also admits the vectorial form (1) with $U_0 = U_2 \otimes U_1$ and $d = d_1 d_2$. Indeed, under the assumption of the Kronecker structure $U_0 = U_2 \otimes U_1$, the models (1) and (2) are exactly equivalent,

* Corresponding author.
E-mail address: joni.virta@utu.fi (J. Virta).

https://doi.org/10.1016/j.patcog.2023.109401

meaning that the collection of all matrix models of the form (2) constitutes a strict subset of the collection of all vector models (1) (where U_0 is allowed to be arbitrary).

From matrix form (2) we see that U_1 determines the dependency structure of the rows of X and U_2 governs the dependencies within the columns of X. Besides facilitating the previous natural intepretation, the structural assumption $U_0 = U_2 \otimes U_1$ also works as a form of *regularization* that helps avoid overfitting (see Section 4.3 for an example in the context of real world abundance data). Indeed, the loading matrices U_1, U_2 in model (2) have a total of $p_1 d_1 + p_2 d_2$ parameters whereas the unstructured U_0 in the vector model has $p_1 p_2 d_1 d_2$ parameters, a significantly larger amount already for moderately large dimensions.

The Kronecker approach has been used to great success in developing matrix versions of, e.g., principal component analysis (PCA) [6,14,40], independent component analysis [38] and sufficient dimension reduction [7,21]. Consequently, the Kronecker approach will also be our tool of choice. Note that the same idea can also be seen to underlie several tensor decompositions, such as the higher order singular value decomposition (HOSVD) and the Tucker decomposition, see, e.g., [5,17].

Moving on to the count aspect of our data, the error ε in (2) is typically Gaussian, implying that the model is not suitable for count data. Moreover, even if the error variable was taken to be discrete, some heavy constraints would need to be placed to the model parameters to guarantee that X contains only non-negative integers, making the approach rather unnatural. As such, we will instead derive a discrete analogy for (2). Such discrete extensions of latent variable models have been intensively studied in the case of vector-valued data and the most popular approach is perhaps that of *exponential family PCA*, where a latent variable model is assumed for the canonical parameter of an exponential family distribution, see [4,19,22] for a general treatment and [33,34] for sparse extensions. Another major framework for the modelling of count data, and the one on which we base our extension of (2), is the Poisson log-normal (PLN) distribution where the data are taken to be conditionally Poisson distributed with the mean parameters following the log-normal distribution. Introduced originally in [1] (but without the dimension reduction context), the PLN model has since been studied from different viewpoints, see [3,11,15], in particular from the perspective of variational inference. The model can also be seen as a special case of the compound Poisson factor model studied in [39].

The primary contributions of this work are as follows:

- We extend the PLN model to matrix-valued data using the Kronecker approach, the obtained model retaining both the interpretability and the parsimonity of the Gaussian Kronecker model (2).
- We develop natural closed-form estimators for the model parameters using the method of moments, studying also their asymptotic behavior. As far as we are aware, large-sample results for PLN models have been proposed earlier only in [11] (in the vectorial case) and even then in a restrictive context requiring repeated measurements for each observational unit.
- Using the parameter estimates we (i) form predictions for the latent variables and, (ii) derive an estimator for the latent model dimensionality, based on the recently proposed idea of predictor augmentation [24].
- We establish the practical superiority of the proposed method to several competitors, including the (vectorial) estimator of [3] based on variational infernce, in both simulations and an application to real world abundance data.

Furthermore, we note that Bayesian models targeting the same type of data (matrices of counts) have been earlier studied in the literature [13,23] but, as far as we are aware, ours is the first frequentist approach.

The manuscript is organized as follows. In Section 2 we discuss briefly some notation and recall the matrix normal distribution that plays a key role in defining our model. Section 3 discusses our model along with the estimation of its parameters, the latent variables and the latent dimensions. In Section 4 we study the finite-sample properties of the parameter and dimension estimators using simulated data and, additionally, present an application to matrix-valued abundance data. In Section 5, we finally close with some discussion.

2. Notation and some preliminaries

Throughout the manuscript, the subscripts 1 and 2 are used to refer to the left-hand and the right-hand sides of the model, respectively, as in (2). The $p_1 \times p_2$ matrix-variate normal distribution with mean $\mu \in \mathbb{R}^{p_1 \times p_2}$ and invertible left and right covariance matrices $\Sigma_1 \in \mathbb{R}^{p_1 \times p_1}$, $\Sigma_2 \in \mathbb{R}^{p_2 \times p_2}$ is denoted by $\mathcal{N}_{p_1 \times p_2}(\mu, \Sigma_1, \Sigma_2)$. That is, if $X \sim \mathcal{N}_{p_1 \times p_2}(\mu, \Sigma_1, \Sigma_2)$ then X has the density function $f_X : \mathbb{R}^{p_1 \times p_2} \to \mathbb{R}$ defined as,

$$f_X(X) = \frac{1}{(2\pi)^{-p_1 p_2/2} |\Sigma_1|^{p_2/2} |\Sigma_2|^{p_1/2}} \exp\left[-\frac{1}{2} \mathrm{tr}\left\{ \Sigma_1^{-1} (X - \mu) \Sigma_2^{-1} (X - \mu)^\top \right\} \right],$$

see, for example, [10]. In a notable special case where the covariance matrices Σ_1 and Σ_2 are diagonal matrices, the elements of X are mutually independent and the variance of the (j,k)th element equals the product $\sigma_{1,jj}\sigma_{2,kk}$ of the corresponding diagonal elements. Note also that the parameters Σ_1, Σ_2 are defined only up to the scaling $(\Sigma_1, \Sigma_2) \mapsto (t\Sigma_1, t^{-1}\Sigma_2)$ for any $t > 0$ (we get rid of this non-identifiability in the next section via a suitable reparametrization).

Some of our results are more naturally formulated in terms of the (column) vectorizations of the related matrices. We use the convention that the vectorization of the observed matrix X is denoted using the lower case $x := \mathrm{vec}(X)$ (and similarly for the latent matrix Z). A property we will repeatedly use without explicit mention is $\mathrm{vec}(AYB^\top) = (B \otimes A)\mathrm{vec}(Y)$, valid for all matrices A, B, Y with appropriate dimensions for the multiplication $AYB^\top$ to be well-defined.

3. Matrix Poisson PCA

3.1. Low-rank normal and Poisson models

To motivate our proposed model for the dimension reduction of matrix count data, we first briefly review the analogous low-rank model for Gaussian data. Namely, assume that the observed $p_1 \times p_2$ random matrix X is generated as

$$X = \mu + U_1 Z U_2^\top + \varepsilon, \tag{3}$$

where $\mu \in \mathbb{R}^{p_1 \times p_2}$, $Z \sim \mathcal{N}_{d_1 \times d_2}(0, \tau\Lambda_1, \tau\Lambda_2)$, $\tau > 0$, $\Lambda_1 \in \mathbb{R}^{d_1 \times d_1}$ and $\Lambda_2 \in \mathbb{R}^{d_2 \times d_2}$ are positive-definite diagonal matrices satisfying $\mathrm{tr}(\Lambda_1) = p_1$ and $\mathrm{tr}(\Lambda_2) = p_2$ for some $d_1 < p_1$, $d_2 < p_2$, the error $\varepsilon \sim \mathcal{N}_{p_1 \times p_2}(0, \sigma I_{p_1}, \sigma I_{p_2})$ for some $\sigma > 0$ and I_p denotes the $p \times p$ identity matrix. Furthermore, the random matrices Z and ε are assumed to be mutually independent. Alternatively, the same model can be written by requiring that $Z \sim \mathcal{N}_{d_1 \times d_2}(0, \tau\Lambda_1, \tau\Lambda_2)$ and

$$X \mid Z \sim \mathcal{N}_{p_1 \times p_2}(\mu + U_1 Z U_2^\top, \sigma I_{p_1}, \sigma I_{p_2}) \tag{4}$$

In practice, the objective underlying this model is, given a sample from the distribution of X, to estimate the "loading matrices" U_1 and U_2 along with the corresponding latent matrices Z, achieving

dimension reduction in the process. Naturally, the latent dimensions d_1, d_2 are usually unknown in practice and have to be estimated as well. As is common in the dimension reduction literature we separate this problem from the estimation of the other parameters and the latent components and assume, for now, that d_1, d_2 are known. Their estimation is then tackled later in Section 3.6. Finally, we note that the parameters U_1, U_2 can, without loss of generality, be assumed to have orthonormal columns. This is because any transformation of the form $U_1 \mapsto U_1 A_1$ can be absorbed to the latent variables Z (similarly for U_2). This property of model (3) is in analogy to standard PCA where the loading matrix is also taken to be orthogonal.

The mean matrix μ in (3) can be estimated through $\mathrm{E}(X)$ and the covariance parameters are standardly estimated using the higher-order singular value decomposition (HOSVD) [5], also known as 2D^{2}PCA [40], where the matrix U_1 is found through the eigenvectors of the *left covariance matrix* $\mathrm{Cov}_1(X) := (1/p_2)\mathrm{E}[\{X - \mathrm{E}(X)\}\{X - E(X)\}^\top]$. Namely, under model (3), we have

$$\mathrm{Cov}_1(X) = \tau^2 U_1 \Lambda_1 U_1^\top + \sigma^2 I_{p_1},$$

showing that the leading eigenvectors of $\mathrm{Cov}_1(X)$ serve as an estimator for U_1 (or, in case of non-simple eigenvalues, for the corresponding subspace). The right-hand side matrix U_2 can be determined similarly by first transposing the observations, after which an estimate for the latent variables is obtained as $U_1^\top (X - \mu) U_2 = Z + \varepsilon_0$ where $\varepsilon_0 \sim \mathcal{N}_{d_1 \times d_2}(0, \sigma I_{d_1}, \sigma I_{d_2})$. Note that a noisy estimate is indeed the best we can do since the original observations are themselves contaminated with ε.

We base our matrix count model on the above ideas, similarly assuming the existence of a matrix of mutually independent normal latent variables $Z \sim \mathcal{N}_{d_1 \times d_2}(0, \tau\Lambda_1, \tau\Lambda_2)$ where the covariance parameters satisfy the trace constraints $\mathrm{tr}(\Lambda_1) = p_1$ and $\mathrm{tr}(\Lambda_2) = p_2$. Conditional on Z the observed $p_1 \times p_2$ matrix X of counts is assumed to satisfy

$$X \mid Z \sim \mathrm{Po}_{p_1 \times p_2}\{\exp(\mu + U_1 Z U_2^\top)\}, \tag{5}$$

where the parameters μ, U_1, U_2 are as in (4), the exponential function is applied element-wise and the notation $\mathrm{Po}_{p_1 \times p_2}(M)$, $M \in \mathbb{R}^{p_1 \times p_2}$, refers to a distribution on $p_1 \times p_2$ matrices having independent elements and whose (j,k)th component is Poisson-distributed with the mean m_{jk}. Since the specification (5) is conditional on Z, this leads to the elements of the observed matrix X being dependent, where the magnitude and exact type of the dependency is controlled by the model parameters and the moments of Z. Comparison to (4) now reveals that the proposed model (5) is indeed a straightforward count analogy of the Gaussian model where the exponential map plays the same role as the inverse link function in log-linear models. The identical structure of the latent variables $\mu + U_1 Z U_2^\top$ in models (4) and (5) also means that our earlier discussion about the orthonormality of U_1, U_2 in the model (4) applies in the Poisson model as well.

For the remainder of this manuscript, we assume that we have observed a random sample $X_1, \ldots, X_n$ from the model (5). Our objectives are then three-fold: (i) We first derive root-n consistent estimates for the model parameters. Of these, especially of interest are the loading matrices U_1, U_2 which describe how the elements of the observations X_i depend on the corresponding latent variables in Z_i. (ii) Given the parameter estimates, we establish estimators for the latent dimensions d_1, d_2. The estimators are based on a recent idea for using predictor augmentation for rank estimation [24]. (iii) Finally, given the previous information, we estimate the latent matrices Z_i themselves. Note that the estimates are again necessarily noisy as error is introduced to model (5) through the Poisson sampling.

In the special case where $p_2 = 1$ (and our observations are vectors), model (5) reduces to a multivariate Poisson log-normal distribution (meaning that the observations are conditionally Poisson-variate with log-normal mean parameters), that was first proposed in [1] for modelling multivariate count data. However, [1] did not consider the model from the viewpoint of dimension reduction, meaning that our results on the estimation of the latent variables and their dimension are novel also for the case $p_2 = 1$.

3.2. Parameter estimation

The model (5) has a total of six parameters to estimate, μ, U_1, U_2, τ^2, Λ_1 and Λ_2. The model being fully parametric, a natural approach to their estimation would be maximum likelihood, as was done with the vectorial version of the model in [1,11,15]. However, the marginal density of X in the model involves an integral lacking a closed-form solution which, besides complicating the parameter estimation, would also make studying the asymptotic properties of the estimators very difficult. Hence, we base our subsequent estimators on the method of moments which, conveniently, yields analytical solutions with tractable asymptotic behavior. In the vectorial case, $p_2 = 1$, our proposed method-of-moments estimators reduce to known quantities that were used in [1,15] as initial values for an iterative maximum likelihood procedure for the fitting of a PLN model to vector data.

For $j, k = 1, \ldots, p_1$, $j \neq k$, we define the following quantities

$$\begin{aligned} s_{1,jk} &:= \frac{1}{p_2} \sum_{\ell=1}^{p_2} \log\left\{ \frac{\mathrm{E}(x_{j\ell} x_{k\ell})}{\mathrm{E}(x_{j\ell})\mathrm{E}(x_{k\ell})} \right\}, \\ s_{1,jj} &:= \frac{1}{p_2} \sum_{\ell=1}^{p_2} \log\left[\frac{\mathrm{E}\{x_{j\ell}(x_{j\ell} - 1)\}}{\{\mathrm{E}(x_{j\ell})\}^2} \right], \end{aligned} \tag{6}$$

along with their "right-hand side" variants, defined for $j, k = 1, \ldots, p_2$, $j \neq k$,

$$\begin{aligned} s_{2,jk} &:= \frac{1}{p_1} \sum_{\ell=1}^{p_1} \log\left\{ \frac{\mathrm{E}(x_{\ell j} x_{\ell k})}{\mathrm{E}(x_{\ell j})\mathrm{E}(x_{\ell k})} \right\}, \\ s_{2,jj} &:= \frac{1}{p_1} \sum_{\ell=1}^{p_1} \log\left[\frac{\mathrm{E}\{x_{\ell j}(x_{\ell j} - 1)\}}{\{\mathrm{E}(x_{\ell j})\}^2} \right]. \end{aligned}$$

Let S_1 be the $p_1 \times p_1$ matrix having the $s_{1,jk}$ as its elements and analogously for S_2. Note that in [1,15], the matrix S_2 was not needed as in the vectorial case with $p_2 = 1$ the right-hand side model parameters $U_2, \Lambda_2 \in \mathbb{R}^{1 \times 1}$ can be absorbed to the left-hand side of the model. Recall that $\tau > 0$ denotes the joint scaling parameter of the latent covariance matrices in the model (5). Then the following holds.

Lemma 1. *Under model (5), we have*

$$S_1 = \tau^2 U_1 \Lambda_1 U_1^\top, \quad S_2 = \tau^2 U_2 \Lambda_2 U_2^\top.$$

Lemma 1 shows that $\mathrm{tr}(S_1)/(2p_1) + \mathrm{tr}(S_2)/(2p_2) = \tau^2$, allowing the estimation of τ^2 through S_1 and S_2. Consequently, matrices U_1 and Λ_1 can be estimated through the leading d_1 eigenvectors and eigenvalues of S_1/τ^2, respectively (implying that S_1 plays the role of the matrix $\mathrm{Cov}_1(X)$ in the Poisson model), and U_2, Λ_2 can be obtained similarly from S_2. This leaves us just with the mean parameter μ which, while a nuisance parameter in the Gaussian model (3), may in the Poisson model be of independent interest, being part of the latent variables $\mu + U_1 Z U_2^\top$. To estimate it, we use the relationship

$$\mu_{jk} = 2 \log \mathrm{E}(x_{jk}) - \frac{1}{2} \log \mathrm{E}\{x_{jk}(x_{jk} - 1)\}$$

following from the proof of Lemma 1.

3.3. Asymptotic normality of the estimators

We next turn to the asymptotic properties of the estimators. Given a random sample $X_1, \dots, X_n$ from the model (5), let S_{n1} denote the sample version of the matrix S_1 (that is, with elements as in (6) but with the expected values replaced by sample means) and define S_{n2} analogously. Finally, define $t_n^2 := \mathrm{tr}(S_{n1})/(2p_1) + \mathrm{tr}(S_{n2})/(2p_2)$ and let the elements of the $p_1 \times p_2$ matrix M_n be $m_{n,jk} := 2\log\{(1/n)\sum_{i=1}^n x_{i,jk}\} - (1/2)\log\{(1/n)\sum_{i=1}^n x_{i,jk}(x_{i,jk}-1)\}$. As our main result of the section, we show that the concatenation $(\mathrm{vec}(S_{n1}), \mathrm{vec}(S_{n2}), t_n^2, \mathrm{vec}(M_n))$ of the vectorizations of the estimators converges in distribution to a multivariate normal distribution. We do this by deriving asymptotic linearizations for them in terms of the first two moments of the sample, which we gather into the following expression:

$$g_n := \frac{1}{\sqrt{n}} \sum_{i=1}^n \begin{pmatrix} \mathrm{vec}\{\mathrm{X_i} \otimes \mathrm{X_i} - \mathrm{E}(\mathrm{X} \otimes \mathrm{X})\} \\ \mathrm{vec}\{X_i - \mathrm{E}(X)\}. \end{pmatrix} \tag{7}$$

Application of the standard CLT reveals that g_n has a limiting normal distribution and its limiting covariance matrix Θ can be straightforwardly (if tediously) derived by computing all pair-wise covariances between the elements of $\mathrm{vec}(X \otimes X)$ and $\mathrm{vec}(X)$, in the same manner as carried out in Lemma 1. However, our calculations revealed that Θ has a particularly complicated form and each of its elements essentially has to be computed individually. This is rather expected as our matrix PLN model does not enjoy any form of invariance properties on the sample level, in the sense as possessed, e.g., by the matrix normal distribution. As such, we view that presenting the full form of Θ here would carry no additional insight, and we have refrained from including it in the paper, instead suggesting to approximate Θ numerically, if needed. This can be achieved by generating a large amount of replications of g_n in (7) where the X_i are taken to follow the model (5), with either estimated or known true values for the parameters. To further exemplify the complexity of Θ, in Appendix A we have derived the asymptotic variance of S_{n1} in the simple special case with $p_1 = p_2 = 1$, demonstrating its rather unintuitive form.

To avoid notational overload in Theorem 1 below, we give the linearizations explicitly only for S_{n1} and M_n. The equivalent expression for S_{n2} can be obtained by applying the formulas for S_{n1} to the transposed sample $X_i^\top$ and for $t_n^2 = \mathrm{tr}(S_{n1})/(2p_1) + \mathrm{tr}(S_{n2})/(2p_2)$ by simply summing over the scaled diagonal elements of S_{n1} and S_{n2}. In Theorem 1 e_j refers to the jth standard basis vector and $f_{jk} := (e_j \otimes e_k)$.

Theorem 1.

(i) *(Off-diagonal elements of S_{n1}) For $j \neq k$, we have,*

$$\sqrt{n}(S_{n1} - S_1)_{jk} = h_{jk}{}^\top g_n + o_p(1),$$

where, for $j \neq k$,

$$h_{jk}^\top := \left(\frac{1}{p_2} \sum_{\ell=1}^{p_2} \frac{(f_{\ell\ell} \otimes f_{kj})^\top}{f_{kj}^\top \mathrm{E}(X \otimes X) f_{\ell\ell}}, -\frac{1}{p_2} \sum_{\ell=1}^{p_2} \left\{ \frac{f_{\ell j}}{e_j^\top \mathrm{E}(X) e_\ell} + \frac{f_{\ell k}}{e_k^\top \mathrm{E}(X) e_\ell} \right\}^\top \right).$$

(ii) *(Diagonal elements of S_{n1}) For $j = k$, we have,*

$$\sqrt{n}(S_{n1} - S_1)_{jj} = h_{jj}{}^\top g_n + o_p(1),$$

where,

$$h_{jj}^\top := \left(\frac{1}{p_2} \sum_{\ell=1}^{p_2} \frac{(f_{\ell\ell} \otimes f_{jj})^\top}{b_{j\ell}}, -\frac{1}{p_2} \sum_{\ell=1}^{p_2} \left\{ \frac{1}{b_{j\ell}} + \frac{2}{e_j^\top \mathrm{E}(X) e_\ell} \right\} f_{\ell j}^\top \right),$$

and $b_{j\ell} := f_{jj}{}^\top \mathrm{E}(X \otimes X) f_{\ell\ell} - e_j{}^\top \mathrm{E}(X) e_\ell$.

(iii) *(The elements of M_n) For all j, k, we have,*

$$\sqrt{n}(M_n - \mu)_{jk} = a_{jk}{}^\top g_n + o_p(1),$$

where,

$$a_{jk}^\top := \left(-\frac{(f_{kk} \otimes f_{jj})^\top}{2b_{jk}}, \left\{ \frac{1}{2b_{jk}} + \frac{2}{e_j{}^\top \mathrm{E}(X) e_k} \right\} f_{kj}^\top \right).$$

Theorem 1 shows that each of the estimators has an asymptotic expansion as a linear combination of the vector g_n. Thus the vector $(\mathrm{vec}(S_{n1}), \mathrm{vec}(S_{n2}), t_n^2, \mathrm{vec}(M_n))$ converges in distribution to a multivariate normal distribution whose covariance matrix is of the form $H\Theta H^\top$ where Θ is the limiting covariance matrix of g_n and H is the matrix containing the coefficients of the asymptotic linearizations from Theorem 1. Equivalent results for the estimators of the eigenelements $U_1, U_2, \Lambda_1 \Lambda_2$ now follow with standard asymptotic techniques [8,35].

3.4. Interpretation of S_1 and S_2

We conclude the section by providing interpretations for the matrices S_1 and S_2. In the special case when $p_2 = 1$ (making the observation X simply a p_1-variate vector x), the matrix S_1 has its (j,k)th off-diagonal element and its (j,j)th diagonal element equal to

$$\log\left\{ \frac{\mathrm{E}(x_j x_k)}{\mathrm{E}(x_j)\mathrm{E}(x_k)} \right\} \quad \text{and} \quad \log\left[\frac{\mathrm{E}\{x_j(x_j-1)\}}{\{\mathrm{E}(x_j)\}^2} \right] \tag{8}$$

respectively, showing that S_1 may be viewed as a count data analogue of the ordinary covariance matrix. That is, instead of additive centering by the mean, we conduct the multiplicative centering $x_j \mapsto x_j/\mathrm{E}(x_j)$ and, instead of raw moments, we use the factorial moments.

In the literature on elliptical distributions, a commonly used family of alternatives to the covariance matrix are known as scatter functionals, defined as any *affine equivariant* mappings $F \mapsto S(F)$ of a p-variate distribution F to the space of positive semi-definite matrices. By affine equivariance, it is meant that, for any invertible matrix $A \in \mathbb{R}^{p\times p}$ and any $b \in \mathbb{R}^p$, the scatter functional satisfies $S(F_{A,b}) = AS(F)A^\top$ where $F_{A,b}$ is the distribution of the random vector $Ax + b$ and $x \sim F$, see, for example, [36] for examples and references on scatter functionals in the context of dimension reduction.

Now, being also an alternative of sorts to the covariance matrix, it is of interest to see whether the current matrix S_1 possesses any similar properties as scatter functionals. For a random p-variate count vector x, it is seen from (8) that the matrix S_1 is invariant under the transformations $x \mapsto Dx$ where $D \in \mathbb{R}^{p\times p}$ is an arbitrary diagonal matrix with positive diagonal elements. Hence, we observe that S_1 does not actually measure the "scatter", or scale, of x but rather some higher order property ("shape"). Inspection also reveals that if the jth and kth element of x are independent, the corresponding off-diagonal element of S_1 vanishes, which is known in the context of scatter functionals as the (element-wise) independence property [28].

Take next the elements of x to be i.i.d. from various standard count data distributions: If $x_1 \sim \mathrm{Po}(\lambda)$, we have $s_{1,11} = 0$ (and, consequently, $S_1 = 0$), showing that Poisson-distribution is viewed as being pure noise by the matrix S_1 (this observation will be used in Section 3.6 to estimate the latent dimension d). If $x_1 \sim \mathrm{NegBin}(r, p)$ (the negative binomial with success probability p and stopping after the rth failure), then $s_{1,11} = \log(1 + 1/r)$. If $x \sim \mathrm{Bin}(n, p)$, we get $s_{1,11} = \log(1 - 1/n)$, showing, in particular, that S_1 is not necessarily positive semi-definite. A common thread behind the previous cases is that in all three the value of S_1 is independent of a subset of the involved parameters (taken to the extreme with the Poisson-distribution). Additionally, the signs of the diagonal elements of S_1

correspond in each case with the presence of overdispersion in the distributions (negative/positive sign being linked with underdispersion/overdispersion). Hence, instead of being a measure of scale, it seems more fitting to view the matrix S_1 as measuring the amount of overdispersion in the data. However, this analogy is not perfect, as, e.g., for the negative binomial distribution the severity of the overdispersion (in the classical sense) depends on the parameter p, to which S_1 is invariant.

3.5. Latent component estimation

For convenience, the results of this section are formulated in terms of the column vectorizations x and z of the matrices X and Z. Additionally, we denote $m := \text{vec}(\mu)$, $\Lambda := \Lambda_2 \otimes \Lambda_1$, $U := U_2 \otimes U_1$, $p := p_1 p_2$ and $d := d_1 d_2$. Recall from Section 3.1 that in the matrix normal model (4), a natural estimator for the latent variables, or principal components (PCs), z is obtained straightforwardly as $U^\top(x - m)$. The simple, linear form of the estimator can be seen to follow from the fact that the observed and the latent matrices belong to the same distributional family, a property that does not hold for the Poisson model (5). However, we can still draw an analogy with the normal model by observing that the linear estimate $U^\top(x - m)$ admits a characterization as the mode of the conditional distribution of the scaled latent components $(I_d + \sigma^2\tau^{-2}\Lambda^{-1})z$ given the observation x.

Lemma 2. *Under the normal model (4), we have*

$$(I_d + \sigma^2\tau^{-2}\Lambda^{-1})z \mid x \sim \mathcal{N}_d\{U^\top(x - m), \sigma^2(I_d + \sigma^2\tau^{-2}\Lambda^{-1})\}.$$

Guided by Lemma 2, we estimate the principal components z in the Poisson model (5) analogously as the mode of the conditional distribution of z given x (we do not incorporate the scaling matrix $I_d + \sigma^2\tau^{-2}\Lambda^{-1}$ as σ^2 has no analogue in the Poisson model and, besides, scale is often anyway seen as a nuisance in dimension reduction). Unlike in the normal model, the resulting conditional distribution does not belong to any standard distributional family, but its mode can still be estimated efficiently through numerical maximization of a concave objective function, as shown in the next theorem. Similar approaches have been used earlier for count data in [16,22]. In the sequel, we denote by $\ell(z|x) := \log f_{z|x}(z|x)$ the logarithmic density function of the conditional distribution.

Theorem 2. *The logarithmic conditional density $\ell(z|x)$ satisfies the following.*

i) *For a constant C not depending on z,*

$$\ell(z|x) = C + x^\top U z - 1^\top \exp(m + Uz) - \frac{1}{2\tau^2} z^\top \Lambda^{-1} z,$$

where $1 \in \mathbb{R}^p$ is a vector of ones and the exponential function is applied element-wise.

ii) *For all $x \in \mathbb{R}^p$, the function $z \mapsto \ell(z|x)$ is strictly concave and admits a unique maximum in $\mathbb{R}^d$.*

Denote the gradient and the Hessian matrix of the map $z \mapsto \ell(z|x)$ as $g(z|x)$ and $H(z|x)$, respectively, the exact forms of which are given in the proof of Theorem 2 in Appendix B. Furthermore, given the estimates of the model parameters from Section 3.2, denote by $\ell_n(z|x)$, $g_n(z|x)$ and $H_n(z|x)$ the logarithmic conditional density, gradient and Hessian, respectively, with the parameter estimates plugged in. The d-variate latent vector z_i corresponding to a vectorized observation x_i can now be estimated as the unique maximizer of $z \mapsto \ell_n(z|x_i)$ using the standard Newton-Raphson method, Theorem 2 guaranteeing its convergence. Recall finally that the model (5) assumes the principal components to have zero mean. Hence, as the final step in their estimation, we still center the estimated sample PCs $z_1, \dots, z_n \in \mathbb{R}^d$.

We end the section with a collection of remarks: (i) If two or more diagonal elements of Λ_1 are equal then the related eigenvectors in U_1 are not uniquely defined (even up to their signs). In such a case, also the corresponding principal components are non-uniquely defined, exhibiting rotational indeterminacy similar to what one encounters in classical factor models. However, we did not find this an issue in practical scenarios where eigenvalues are typically distinct enough up to some numerical precision. (ii) Interestingly, Theorem 2 reveals that the estimate of the latent variables z_i depends on the observed data x_i only through the projection $U_n{}^\top x_i$ where $U_n := U_{n2} \otimes U_{n1}$ and the $p_1 \times d_1$ matrix U_{n1} contains any first d_1 eigenvectors of S_{n1} as its columns (and similarly for U_{n2}). This is somewhat surprising as, based on the formulation of the model (5), one would expect the matrix U_n to act linearly only with Z, and not with X (to which it has a non-linear functional dependency). (iii) Finally, while we viewed the Gaussian estimate $U^\top(x - m)$ above as the mode of the conditional normal distribution in Lemma 2, it is, naturally, also the mean of the same distribution. Thus, an alternative strategy in the Poisson model would be to base the estimates of the latent components on the conditional means of the random vector z given the observations x_i. However, while equally valid (and heuristic) as the taken viewpoint, relying on the mean would lead to an intractable integral requiring numerical approximation, leading us to favor the mode approach with its concave optimization problem.

3.6. Dimension estimation

We next develop an estimator for the latent dimension d_1 using the recently proposed idea of predictor augmentation [24]. By the symmetry of model (5), an estimator for d_2 is obtained exactly analogously after the transposition of the observations.

In predictor augmentation, artificially generated noise is concatenated to the observations in order to reveal the cut-off point from positive values to zero in the spectrum of the matrix of interest. More precisely, given a random sample $X_1, \dots, X_n$ from the model (5), fix a positive integer $r_1 \in \mathbb{N}^+$ and let $X_1^*, \dots, X_n^*$ be the augmented sample where $X_i^* = (X_i{}^\top, R_i{}^\top)^\top$ and the elements of the $r_1 \times p_2$ matrices R_i are sampled i.i.d. from the Poisson distribution with the rate parameter $\lambda = 1$, $i = 1, \dots, n$. Letting S_{n1}^* denote the equivalent of the matrix S_{n1} but computed from the augmented sample, techniques similar to the ones used in Lemma 1 and Theorem 1 show that

$$S_{n1}^* = \begin{pmatrix} \tau^2 U_1 \Lambda_1 U_1{}^\top & 0 \\ 0 & 0 \end{pmatrix} + \mathcal{O}_p(1/\sqrt{n}). \tag{9}$$

Let now the r_1-dimensional vectors $\beta_{n11}, \dots, \beta_{n1(p_1+r_1)}$ contain the final r_1 elements of any set of orthogonal eigenvectors of S_{n1}^* (that is, β_{n11} contains the last r_1 entries of an eigenvector corresponding to the first eigenvalue of S_{n1}^* etc.) Now, for $k \leq d_1$, we expect the norms $\|\beta_{n1k}\|$ to be close to zero as the corresponding eigenspaces are, in the limit $n \to \infty$, concentrated fully on the subspace spanned by the d_1 columns of the $(p_1 + r_1) \times d_1$ matrix $(U_1{}^\top, 0)^\top$. On the other hand, for $k > d_1$, there is no reason for $\|\beta_{n1k}\|$ to be small as the final $p_1 + r_1 - d_1$ eigenvalues of the limiting matrix in (9) are all equal to zero and, hence, the corresponding eigenvectors should not favor any direction (in the null space). For a rigorous presentation of this concept, along with more details on the full procedure, see [24]. In predictor augmentation, this information provided by the eigenvectors is further supplemented by the eigenvalues $\lambda_{n11} \geq \cdots \geq \lambda_{n1(p_1+r_1)}$ of the matrix S_{n1}^* to define the objective function $\phi_{n1} : \{0, \dots, p_1\} \to \mathbb{R}$

$$\phi_{n1}(k) = \sum_{j=0}^{k} \|\beta_{n1j}\|^2 + \frac{\lambda_{n1(k+1)}}{1 + \sum_{j=1}^{k+1} \lambda_{n1j}}, \tag{10}$$

where we define $\|\beta_{n10}\|$ to be equal to zero. Note that the second term of (10) corresponds essentially to a scaled version of the scree plot used commonly in PCA. By the earlier discussion, we expect the first term of $\phi_{n1}(k)$ to be small for $k \leq d_1$, whereas, the second term takes (for large enough n) small values for $k \geq d_1$ (i.e., at the indices corresponding to the zero limit eigenvalues). Consequently, we take as our estimate of d_1 the value k at which ϕ_{n1} is minimized,

$$d_{n1} := \text{argmin}_{k\in\{0,\dots,p\}}\phi_{n1}(k).$$

To increase the stability of the estimate for small n, [24] further advocated independently carrying out the augmentation procedure s_1 times and replacing $\|\beta_{n1j}\|^2$ in (10) with its mean over the s_1 replicates. Similarly, we also replace the eigenvalues λ_{n1j} with their means over the replicates (although this was not done in [24]). The purpose of this modification is to reduce the variability in their estimation, even though experimentation (not shown here) reveals that the variability of individual eigenvalues over the replicates is typically much smaller than the variability between the different eigenvalues corresponding to the signal and noise. The procedure has two tuning parameters, r_1 and s_1, the latter of which directly reduces variation in the results and, hence, we suggest to use large values for it in practice. Discussion regarding the choice of the value of r_1 is given later in Section 4.

Our simulation results in Section 4 suggest that d_{n1} could be a consistent estimator of d_1 as $n \to \infty$, but proving this turned out to be less than straightforward. Namely, [24] give sufficient conditions under which estimators such as d_{n1} are consistent for the true dimension. Our scenario is easily checked to satisfy these assumptions apart from one, i.e., the requirement (12) in [24] that the sequence of augmented matrices S^*_{n1} is in a specific sense contiguous to Lebesgue measure. Now, a standard way of showing this would be to establish that $\sqrt{n}(S^*_{n1} - S^*_1)$, where S^*_1 is the limit of S^*_{n1}, converges in total variation to a non-singular normal distribution. However, by the classical result of Prohorov, see, e.g., Theorem 2.2 in [2], the convergence in total variation happens in the central limit theorem if and only if the corresponding sample estimators have non-trivial absolutely continuous components. Naturally, this is not the case with our count data model, implying that alternative strategies must be sought and, hence, we leave this question for future study.

We next illustrate the dimension determination procedure in a simple special case of the model (5) with $p_2 = 1$, meaning that the observations can be treated as p_1-dimensional vectors and it is sufficient to consider dimension estimation for the left-hand side of the model only (recall that our results are novel also in this special case). This, and all the examples to follow, were implemented in the language R [30]. The data set `microbialdata` in the R-package `gllvm` [26] consists of the abundances of 985 bacteria species measured at $n = 56$ soil sample sites located either in Austria, Finland or Norway. For the purposes of this demonstration, we limit our attention to the subset of the $p_1 = 20$ most abundant species, defined as the ones having the least proportions of zero counts over all 56 sites. We now apply the proposed dimension determination procedure to this subset of the data with the choices $r_1 = \lceil p_1/5 \rceil = 4$ and $s_1 = 100$. The value of r_1 was chosen for its good success in the simulation results of [24] and for s_1 we chose a large enough value to increase the stability of the estimate in the presence of the rather low sample size. The resulting function ϕ_{1n} is plotted in Fig. 1 and leads to the estimate $d_{n1} = 3$. To assess whether the corresponding latent variables are meaningful, we estimate the model parameters, followed by the values of the first three latent variables for each of the $n = 56$ sites using the method of Section 3.5. For ease of presentation, we limit ourselves to the first two latent variables whose scatter plot is presented in Fig. 2 (the colored plot markers), overlaid with the loadings of the first two latent variables, i.e., the first two columns of U_{n1} (the numbers connected with dashed lines to the origin). The numbering of the $p_1 = 20$ species corresponds to their indexing in a specific taxonomy. The sites of the three regions appear to be rather well-separated in the first two latent variables and, to verify this finding, we fit a generalized linear Poisson latent variable model [27] between the abundances and the region variable using the function `gllvm` in the R-package `gllvm`. Based on the coefficients estimates of the model, the sites in Austria are the most (the least) associated with the bacteria species 8, 4 (52, 184), the sites in Finland are the most (the least) associated with the species 7, 1 (4, 13) and the sites in Norway are the most (the least) associated with the species 64, 1242 (8, 70). Comparison of the previous with Fig. 2 now reveals that the same pattern is indeed rather accurately reflected in the positioning of the loadings and the sites in our "biplot", showing that the latent variables managed to capture essential biological information. Further illustration of the methodology in the general case $p_2 > 1$ are given in Subsection 4.3.

4. Examples

4.1. Dimension estimation in a simulation

We next study the performance of the augmentation procedure in estimating the latent dimension. We take as our competitor the same augmentation estimator but applied to the Gaussian model (3). This is essentially achieved by replacing the matrix S_1 in the augmentation procedure with $\text{Cov}_1(X)$, see Section 3.1, and similarly for S_2. This estimator, which can be seen as a "naive" data type ignoring approach to matrix-valued count data has been recently studied in the context of image data in [31], see their work for more details.

We use two different sample sizes, $n = 100, 500$, and two different observed dimensions, either $(p_1, p_2) = (10, 5)$ ("Low dimension") or $(p_1, p_2) = (50, 25)$ ("High dimension"). In each case we take the mean μ to be the zero matrix of appropriate size. Two different models for the covariance parameters are considered: In Model 1, the first dimension is of rank one, $S_1 = \tau^2 U_1 \Lambda_1 U_1^\top = 1_{p_1} 1_{p_1}^\top$ (1_{p_1} is the p_1-dimensional vector with all elements equal to one), and the second dimension is of rank five, $S_2 = \tau^2 U_2 \Lambda_2 U_2^\top = W E_5 W^\top$, where W is a uniformly random $p_2 \times p_2$ orthogonal matrix (drawn separately for every iteration of the study) and $E_5 \in \mathbb{R}^{p_2 \times p_2}$ is a diagonal matrix with its first five diagonal elements equal to one and the rest of them zero. In Model 2, both dimensions are taken to have rank five, with the covariance parameters being created analogously to that of the second dimension in Model 1. Data from every combination of the parameters and models are simulated 200 times and for each replicate we estimate the two dimensions with seven different approaches. These include our proposed augmentation approach with the numbers of augmentations $(r_1, r_2) = (p_1, p_2), (\lceil p_1/2 \rceil, \lceil p_2/2 \rceil), (\lceil p_1/5 \rceil, \lceil p_2/5 \rceil), (\lceil p_1/10 \rceil, \lceil p_2/10 \rceil), (1, 1)$ (denoted in the following by A1, A2, A3, A4, A5 respectively), where in each case we take the numbers of repetitions to be $s_1 = s_2 = 5$. In addition, we consider the Gaussian augmentation procedure as described in [31] and implemented in the R-package `tensorBSS` [37], and having either $r = 1$ or $r = 5$ augmentations (denoted in the following by G1, G2), with both cases using $s = 5$ repetitions.

The rounded percentages of correctly estimated left and right dimensions over the 200 replicates are shown in Table 1 where L and R refer to the left-hand side dimension d_1 and the right-hand side dimension d_2, respectively. Comparison of the methods A1 – A5 shows that, overall, the best results are obtained with small amounts of augmentations (r_1, r_2). The differences occur mostly in the high-dimensional version of Model 2 whose dimensions

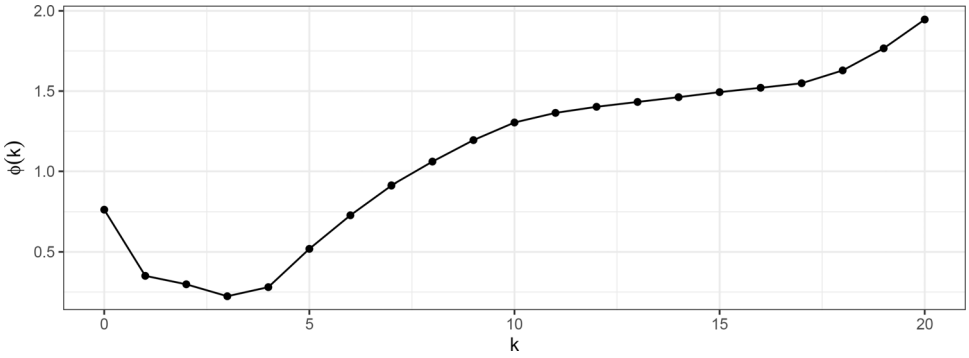

Fig. 1. Plot of the map $k \mapsto \phi_{n1}(k)$ for the abundance data set. The minimum value is achieved at $k = 3$.

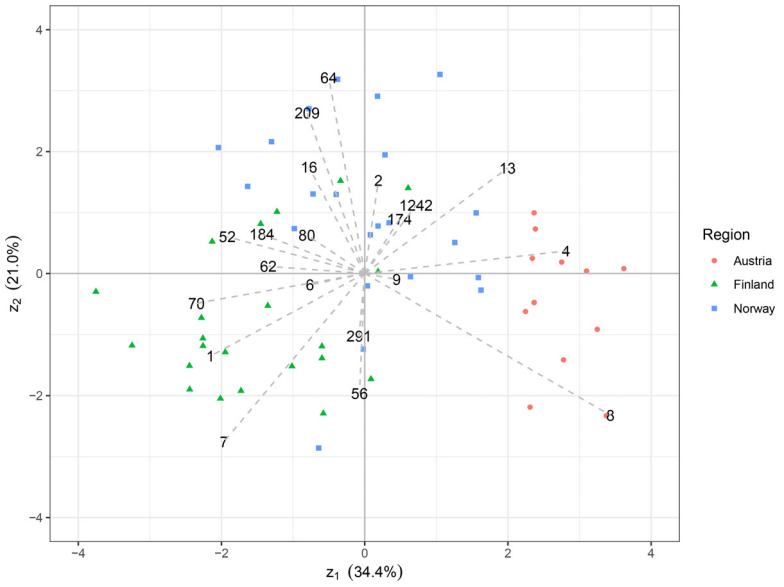

Fig. 2. Scatter plot of the first two latent components for the abundance data set (the colored plot markers), overlaid with the corresponding loadings (the numbers connected with dashed lines to the origin). The numbering of the species represents their indexing in a specific taxonomy. The percentages on the axes denote the corresponding explained proportions of variance (ratios of the individual eigenvalues of S_{n1} to their sum).

Table 1
Results of the dimension estimation study. The numbers refer to the percentages of correctly estimated dimensions in the different combinations of models and parameters. The results for the left-hand side dimension d_1 are denoted by L whereas R signifies the estimates for the right-hand side dimension d_2.

			G1		G2		A1		A2		A3		A4		A5	
Dim.	Model	n	L	R	L	R	L	R	L	R	L	R	L	R	L	R
Low	1	100	98	96	100	83	100	76	100	95	100	100	100	100	100	100
		500	100	100	100	100	100	100	100	100	100	100	100	100	100	100
Low	2	100	26	0	0	0	64	76	100	99	100	100	98	100	100	100
		500	84	0	46	0	100	100	100	100	100	100	100	100	100	100
High	1	100	87	23	100	86	100	100	100	100	100	100	100	100	100	99
		500	98	2	100	34	100	100	100	100	100	100	100	100	100	100
High	2	100	10	35	8	0	0	0	0	0	0	0	0	0	38	8
		500	58	88	100	98	0	0	0	0	32	0	98	44	90	95

none of A1–A5 is able to consistently estimate for $n = 100$, but, after increasing the sample size to $n = 500$, A5 achieves there almost perfect results, making it the most consistent of the methods. Also, interestingly, for $n = 100$, the method A5 has more difficulties in estimating d_2 (R) than d_1 (L) even though we have $p_1 > p_2$ and $d_1 = d_2$. Turning our attention to the Gaussian augmentations G1, G2, we observe that they work very consistently under some settings (low-dimensional Model 1) and badly underperform under some (low-dimensional Model 2). But most interestingly, G2 achieves the best performance out of all seven methods in the high-dimensional Model 2 with $n = 500$. Thus, while the Gaussian approach appears to be too unreliable to be used in practical situations of count data, it clearly does work extremely well in some specific situations, warranting more research in the future.

Recall that the augmentation procedure is based on appending independent Po(1)-variates to the original observed matrices. Nevertheless, there is nothing special in the rate parameter value $\lambda = 1$ since, as discussed in Section 3.4, all Poisson-distributions Po(λ), $\lambda > 0$ are seen as noise by our estimators and could be used in the augmentation in place of Po(1). Indeed, the same zero-structure is obtained for the augmented S^*_{n1} in (9) in the limit for all values of λ, meaning that the exact choice of λ is irrelevant on the

Table 2
Results of the dimension estimation study when using $Po(10)$-distribution to augment. The numbers refer to the percentages of correctly estimated dimensions in the different combinations of models and parameters. The results for the left-hand side dimension d_1 are denoted by L whereas R signifies the estimates for the right-hand side dimension d_2.

			G1		G2		A1		A2		A3		A4		A5	
Dim.	Model	n	L	R	L	R	L	R	L	R	L	R	L	R	L	R
Low	1	100	98	97	100	82	100	100	100	100	100	100	98	100	98	100
		500	98	100	100	100	100	100	100	100	100	100	100	100	98	100
Low	2	100	37	0	1	0	100	100	98	100	89	100	72	100	69	100
		500	88	0	51	0	100	100	100	100	97	100	82	100	84	100
High	1	100	89	26	100	86	100	100	100	100	100	97	92	84	2	30
		500	99	2	100	36	100	100	100	100	100	99	98	97	38	66
High	2	100	16	38	9	2	0	0	14	0	34	34	0	44	0	6
		500	55	86	100	99	100	80	100	100	69	84	12	54	0	4

population level. However, the choice of λ might still affect the finite-sample performance of the augmentation procedure and, to investigate this, we reran the current simulation by using Po(10) in place of Po(1).

The obtained results are shown in Table 2 and mostly coincide with Table 1, but with one interesting difference: The larger value $\lambda = 10$ seems to give overall better results when combined with a larger number of augmentations (A1, A2, A3) whereas the smaller value $\lambda = 1$ works the best when coupled with a smaller number of augmentations (A4, A5). In most settings this effect is rather small, but under the high-dimensional variant of Model 2 the differences in estimation accuracy are actually quite drastic. Namely, A2 with Po(10)-augmentation achieves there perfect accuracy for $n = 500$, whereas Po(1)-augmentation completely fails in the same scenario.

We conclude this subsection by discussing the impact of the tuning parameters on the augmentation procedure, based on the previous sets of results. In low-dimensional scenarios where the sample size is much larger than the dimensionality (first four rows of Tables 1 and 2), it appears that the choice of (r_1, r_2) (and the mean λ of the augmentation distribution $\text{Po}(\lambda)$) has little effect on the results. This observation is perfectly in line with the heuristic arguments in Section 3.6, in which the exact values of the tuning parameters are (asymptotically) irrelevant. Moreover, the same conclusion was reached also in [24] in the context of augmentation of vector data, see their Table 2. As such, for simplicity and in order to minimize computational time, we suggest using Po(1)-distribution with the number of augmentations $r_1 = r_2 = 1$ in low-dimensional scenarios.

The story is quite different in high-dimensional scenarios where the sample size and the dimensionality are of comparable magnitudes. As can be observed on the last four rows of Tables 1 and 2, in such cases the exact choices of the tuning parameters can have much greater impact on the results. The underlying reason for this is that the asymptotic arguments in Section 3.6 are in general not valid in high-dimensional regimes. This phenomenon is analogous to the problem of consistently estimating the eigenstructure of the covariance matrix which is notoriously complicated for high-dimensional data due to eigenvalue phase transition and similar effects, see [18] for a review. A full solution to the current problem would thus involve the theoretical study of the high-dimensional asymptotics of the augmentation estimator, which we leave for future work due to the complicated mathematics involved in it. In absence of such results, as a practical rule of thumb, we advise to conduct the augmentation with several different values of the tuning parameters, in order to assess the sensitivity of the results. Finally, we note that the same instability of the augmentation procedure in the context of high-dimensional data is visible also in Table 3 in [24].

4.2. Efficiency study

Next we study the finite-sample behavior of the estimators S_{n1}, S_{n2}, M_n of the model parameters. For this we generate 4×3 observations $X_1, \dots, X_n$ from the model (5) either with $S_1 = \tau^2 U_1 \Lambda_1 U_1^\top = I_4$, $S_2 = \tau^2 U_2 \Lambda_2 U_2^\top = I_3$ (the full rank model) or with $U_1 \Lambda_1 U_1^\top = 1_4 1_4^\top$, $U_2 \Lambda_2 U_2^\top = 1_3 1_3^\top$ (the low-rank model) where 1_p denotes the p-dimensional vector consisting solely of ones. In both cases we take $\mu = 1_4 1_3^\top$. We consider a total of six different sample sizes, $n = 500, 1000, 2000, 4000, 8000, 16000$, and independently replicate each combination of the previous simulation settings 1000 times.

As competitors to the proposed method, we employ estimators of the vectorial version of the PLN model. This is feasible since, as discussed in Section 1, our matrix PLN model (5) reduces to the vector PLN model when vectorized, $\text{vec}(X) \mid \text{vec}(Z) \sim \text{Po}_{p_1 p_2}[\exp\{\text{vec}(\mu) + (U_2 \otimes U_1)\text{vec}(Z)\}]$, where $\text{vec}(Z) \sim \mathcal{N}_{d_1 d_2}(0, \tau\Lambda_2 \otimes \tau\Lambda_1)$. As such, any estimator of the vector PLN model can be used to estimate the parameters of the matrix PLN model also (but with a loss of information, see below). For example, letting S_n be the equivalent of the method-of-moments estimator S_{n1} in the vector PLN model, then, reasoning as in Section 3.3, we observe that $S_n/\text{tr}(S_n)$ converges in the limit to the matrix $A := (U_2 \Lambda_2 U_2^\top / p_2) \otimes (U_1 \Lambda_1 U_1^\top / p_1)$. Besides the method-of-moments estimator, we also include the variational inference (VI) estimator of the vector PLN model implemented in the R-package `PLNmodels` [3]. To allow comparisons to our proposed estimators, we combine the matrix PLN estimators S_{n1} and S_{n2} to $\{S_{n2}/\text{tr}(S_{n2}) \otimes S_{n1}/\text{tr}(S_{n1})\}$ which also estimates the quantity A. To evaluate the accuracy of the estimators, we compute for each the relative differences, e.g., $\|S_n/\text{tr}(S_n) - A\|_F / \|A\|_F$, where $\|\cdot\|_F$ denotes the Frobenius norm. For the mean parameter μ, we similarly compute the relative differences between its estimators and true value.

As the estimators of the vector PLN model completely ignore the natural matricial covariance structure of the model, we expect them to give worse results in the estimation of the combined covariance parameter A. One purpose of this study is then to quantify the severity of this loss in efficiency.

The resulting average relative errors over 1000 replications are shown as functions of the sample size in Fig. 3. Note that in estimating the mean (left panel of the figure), both our proposed estimator (the red line with circles) and its vectorization-based counterpart (the green line with triangles) have overlapping lines. This is because the two estimators are actually the same as the Kronecker structure manifests only in the covariance part of the model (5). Based on Fig. 3, we make the following observations: (i) For the mean parameter estimation, the competing estimator based on variational inference (VI, blue line with the squares) starts off better but actually gets worse with increasing n, our pro-

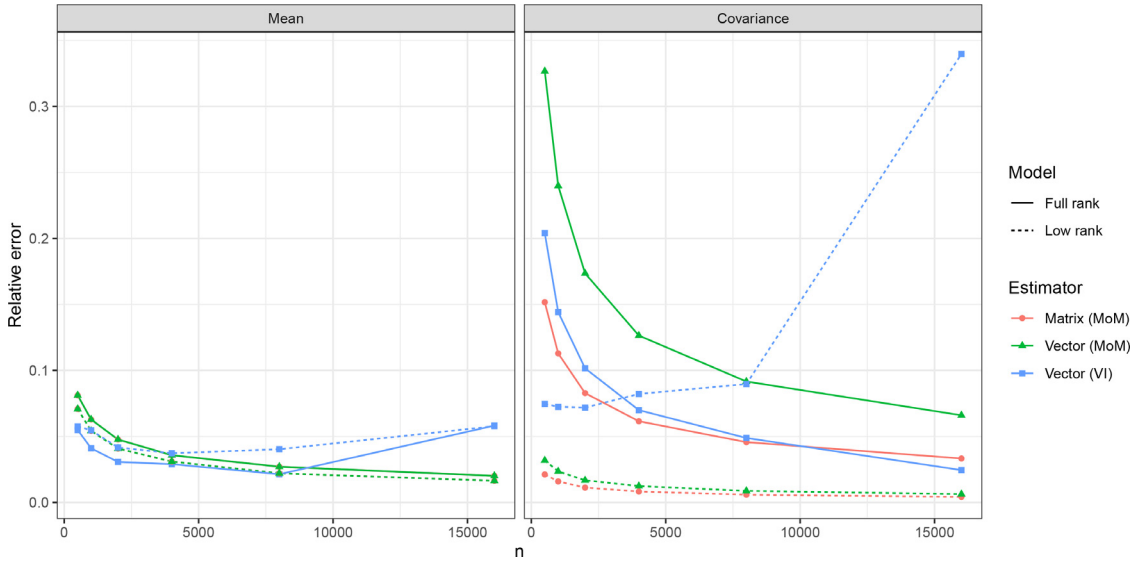

Fig. 3. The average relative errors as a function of the sample size in the efficiency study. As described in the text, in the left panel the lines for "Matrix (MoM)" (red) and "Vector (MoM)" (green) are exactly overlapping.. (For interpretation of the references to colour in this figure legend, the reader is referred to the web version of this article.)

posed estimator eventually overtaking it. (ii) There is little difference between the two models with respect to the mean estimation, which is not a surprise as the models differ only in their covariance structures (full rank vs. low rank). (iii) Regarding the covariance estimation in the full rank model, our proposed matrix method and the VI estimator have close to equal performance but for larger sample sizes the latter is slightly preferable. (iv) In the low rank model, our proposed estimator has the best performance and the VI estimator breaks down as n grows. Closer inspection (not shown here) reveals that when the covariance matrix has a low-rank structure the VI estimator suffers from a systematic and severe underestimation of the off-diagonal elements of the covariance matrix A, but estimates its diagonal elements very efficiently. Consequently, while this behavior can possibly be fixed with a suitable bias correction, in its current form the VI estimator cannot really be recommended in dimension reduction scenarios where one assumes the data to exhibit a low-rank structure.

4.3. Real data example

Following in the spirit of our preliminary example in Section 3.6, we next apply the proposed method to matrix-valued abundance data used earlier in [9], and available in https://www.github.com/rfrelat/Multivariate2D3D. The data consists of the relative abundances (rounded to nearest integer) of a total of $n = 65$ fish species in seven different so-called roundfish areas (RA 1 – RA 7) in the North Sea, studied during the years 1985–2015 which we further divided into 6 time periods $(1985-1989, \dots, 2005-2009, 2010-2015)$. Thus, for the ith species, we have the 7×6 matrix X_i whose (j,k)th element tells the relative abundance of that particular species in the area RA j during the kth period.

In [9], six biologically meaningful clusters (*Southern, Northern, NW Increasing, SE Increasing, Increasing* and *Decreasing*) were identified among the 65 species in the data using the combination of principal tensor analysis [20] and hierarchical clustering. As one of the primary practical objectives of dimension reduction is the discovery of structure (such as groups) in data, it seems reasonable to require that any successful method for reducing the dimension of the current data should be able to detect the previous six clusters, that were indeed in [9] deemed biologically internally consistent.

With the previous in mind, we next estimate our Poisson model for the data, starting with dimension estimation using the augmentation procedure of Section 3.6. We used Po(1)-distribution to generate the augmentations along with the tuning parameter values $(r_1, r_2) = (1, 1)$ and $(s_1, s_2) = (100, 100)$, and experimentation (not shown here) revealed that the resulting estimates are not too sensitive to these choices. The resulting augmentation curves are shown in Fig. 4 (the left panel corresponding to the areas and the right panel to the time periods) and let us conclude that the latent dimension of the time periods is clearly one. For the areas, while the minimum of the curve is achieved at three dimensions, also two seems to be a reasonable option. In order not to lose any information, we retain a total of three principal components, $z_{i,11}, z_{i,21}, z_{i,31}$, estimated with the algorithm in Section 3.5 (of which three observations failed to converge due to numerical overflow and are not shown in the subsequent plots). Examination of the corresponding loading vectors (the columns of U_{n1} and U_{n2}) then reveals that in both modes the first loading vector has roughly constant elements, indicating that the corresponding PC $z_{i,11}$ simply measures the overall abundances of the species (the absolute correlation between $z_{i,11}$ and the average abundances of the species in the 42 area-time combinations is 0.55).

As our interests lie deeper than in the aggregate abundances, we next ignore the PC $z_{i,11}$ and plot the remaining two, $z_{i,21}$ and $z_{i,31}$, in a bivariate scatter plot, depicted in Fig. 5. The coloring/shapes in the plot correspond to the six clusters identified in [9] and we observe that they are indeed well-separated in the plot, with the exception of *NW Increasing* and *Increasing*[9]. actually remark that the *NW Increasing* is a "very heterogeneous cluster" and, by Fig. 3 in [9], if their hierarchical clustering had been stopped at five instead of six clusters, it is precisely the clusters *NW Increasing* and *Increasing* that would have been joined next. Thus, we conclude that the principal components in Fig. 5 have quite successfully managed to capture the group structure of the data. Overlaid in Fig. 5 as dashed lines are also the corresponding area loadings (given by the second and third column of U_{n1}) for the seven roundfish areas. Comparison to Fig. 4 in [9] reveals that these rather accurately capture the division of the clusters in the seven areas (for example, the *Southern* cluster is heavily concentrated in RA 5, as suggested by the aligning of the corresponding group and dashed line in our Fig. 5). We also observe that the loadings of the areas manage to capture some geographical information, as, after reflecting w.r.t. the x-axis and rotating clock-wise by 90 degrees, the loading map in Fig. 5 matches approximately

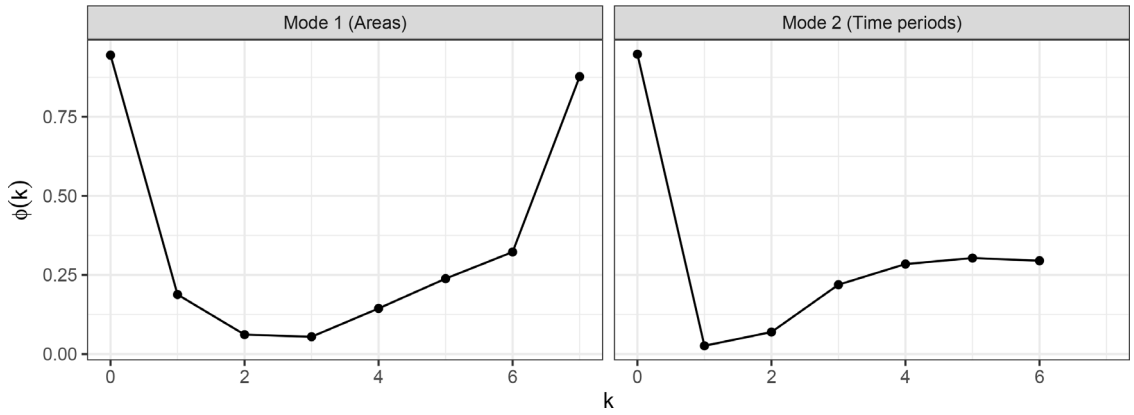

Fig. 4. The two augmentation estimator curves for the matrix-valued abundance data. The left panel corresponds to the area dimension and the right one to the time dimension.

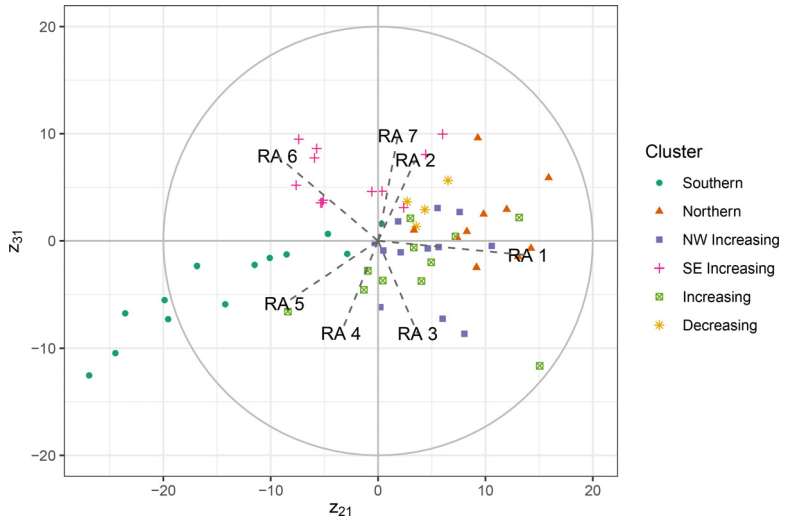

Fig. 5. The scatter plot of the principal components $z_{i,21}$ and $z_{i,31}$ for the matrix-valued abundance data. Overlaid as dashed lines are the corresponding loadings of the seven roundfish areas and the coloring corresponds to the clustering in [9].

with the actual map of the seven areas in the Northern sea, see Fig. 1 in [9]. Note that we have not included the loadings of the time dimension in Fig. 5 as it is one-dimensional for the PCs (all of $z_{i,11}$, $z_{i,21}$, $z_{i,31}$ have the same column coordinate). Besides, the corresponding dimension was already earlier deemed as uninteresting.

Before concluding, we still briefly compare the obtained results to those of two natural competitors: our proposed Poisson PCA procedure applied to vectorized observations (that is, each 7×6 matrix X_i is replaced with a 42-dimensional vector x_i) and the popular (Gaussian) tensor decomposition known as *higher order singular value decomposition* (HOSVD) [5] and implemented as the function `tPCA` in the R-package `tensorBSS` [37]. The augmentation plot (not shown here) for the vectorial version of our proposed method reveals that the latent dimension is three. Similarly to the matrix model, the first PC has again almost constant loadings for all 42 variables and in Fig. 6 we have visualized the second and the third PC. The plot clearly manages to separate the *Southern* and *SE Increasing* clusters but the remaining four are left more or less overlapping. In addition, incorporation of any loading information to Fig. 6 would be difficult as we lost the distinction between the row and column variables in the vectorization. Note also that, as described in the introduction, our proposed matrix-model is actually a submodel of its vectorial counterpart and, thus, any PCs found under the matrix model are also possible to discover under the vector model. Hence, the fact that the matrix model actually performed better than its more general vector version leads us to conclude that the "regularization" offered by the former was indeed beneficial in practice.

Finally, we applied HOSVD, i.e., the Gaussian alternative of our proposed model, to the data. The corresponding Gaussian augmentation plots [31] are shown in Fig. 7 and advocate using either 2 or 3 area components and 1 or 2 time components. However, inspection of the resulting 6 latent components $z_{i,11}, z_{i,21}, z_{i,31}, z_{i,12}, z_{i,22}, z_{i,32}$ reveals that each of them is heavily dominated by a set of six outliers. As an example, we have shown the scatter plot of $(z_{i,11}, z_{i,21})$ in Fig. 8, demonstrating this behavior. Closer examination reveals that these six species ("*Eutrigla gurnardus*", "*Hippoglossoides platessoides*", "*Limanda limanda*", "*Melanogrammus aeglefinus*", "*Merlangius merlangus*" and "*Trisopterus esmarkii*") are precisely those that have the largest average abundances over the 42 area-time combinations. Moreover, these six species also have the largest standard deviations of the abundances among all the species. This observation indicates that the Gaussian method, which assumes constant variation irrespective of the mean, is unable to accommodate the dispersion of the count data which typically increases with the observation size. Hence, we refrain from interpreting the results of the Gaussian method further. We conclude with two remarks: (i) Our proposed Poisson method does not consider the previous six species, which

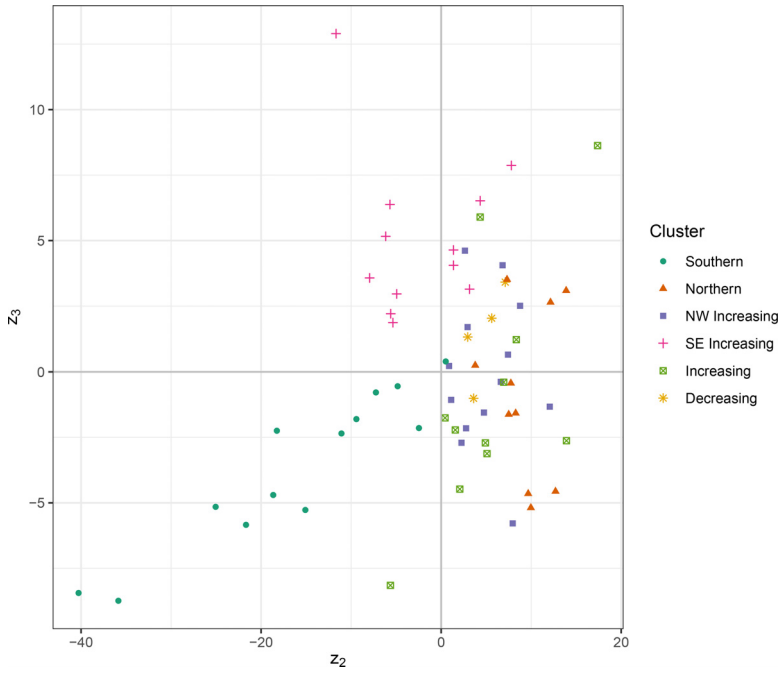

Fig. 6. The scatter plot of the second and third PCs extracted from the vectorized matrix abundance data. The coloring corresponds to the clustering in [9].

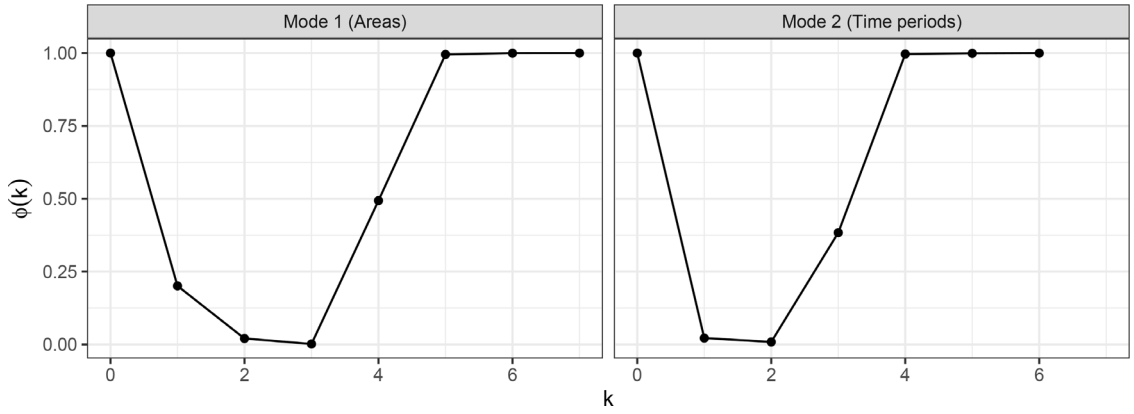

Fig. 7. The two Gaussian augmentation estimator curves for the matrix-valued abundance data. The left panel corresponds to the area dimension and the right one to the time dimension.

all belong to the cluster "NW Increasing", as outliers, as can be seen in Fig. 5 where no member of this cluster stands out particularly. (ii) Removing the six outlying species and rerunning the Gaussian HOSVD method is of no help as in that case a new small set of observations again dominates the components.

5. Discussion

In this work, we proposed a latent variable model for data where the observations are matrices of counts. We used the method of moments to obtain parameter estimators which, while unlikely to be the optimal choice (or equivalent to the intractable maximum likelihood estimators), are nevertheless natural and admit closed-form solutions, allowing fast computation. We estimated the latent principal components through their conditional modes via a concave maximization problem. Finally, we also proposed an efficient procedure for estimating the latent dimensions of the data.

A natural continuation to the current work would be to extend the results to apply also to count-valued tensors (higher-order counterparts of matrices). Indeed, we expect that this could be rather straightforwardly carried out through the concept of tensor flattening, see [38], for a detailed derivation of a particular dimension reduction procedure, first for matrices and then to general tensors through the use of flattening. However, in the current work, we decided to limit ourselves to matrix data as: (i) examples of higher-order tensorial count data are still rather rare and, more importantly, (ii) the presentation of the theory is considerably less notationally intensive in the matrix case (cf. the two approaches in [38]).

As a second future extension, adding some kind of a zero-inflation mechanism to the model would be warranted since count data often exhibit more zero observations than the standard discrete probability models predict. One possible approach would be to replace the Poisson distribution in model (5) with its zero-inflated counterpart such that the probability matrix of the zero-inflation has a low rank structure, see [25] for a similar idea in the context of missing data.

Another possible direction would be to consider a binary equivalent of our proposed model, obtained by replacing the conditional

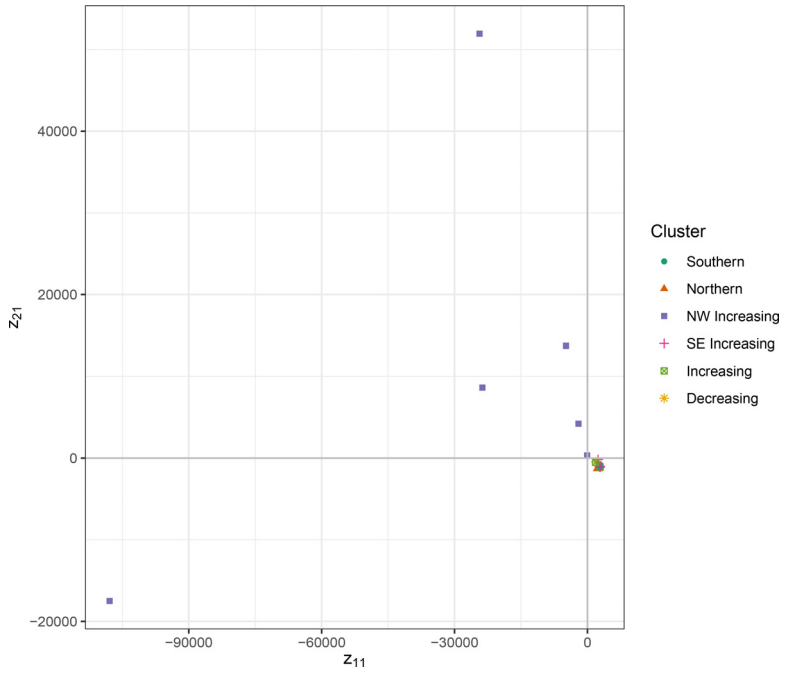

Fig. 8. The scatter plot of the principal components $z_{i,11}$ and $z_{i,21}$ for the matrix-valued abundance data extracted with the Gaussian HOSVD method.

distribution (5) with

$$X \mid Z \sim \text{Ber}_{p_1 \times p_2}\{\Phi(\mu + U_1 Z U_2^\top)\}, \tag{11}$$

where Ber denotes the Bernoulli distribution and the CDF Φ of the standard normal distribution is used as a "link function". Model (11) would offer a natural approach to the dimension reduction of a sample of binary matrices. However, while analogous in form to our proposed Poisson model, a completely different set of estimators would be required for the parameters of (11). Moreover, some preliminary investigation reveals that already the method of moments estimator for the parameters $U_1, U_2, \Lambda_1, \Lambda_2$ of the binary model (11) leads to rather intractable calculations involving Owen's T-function [29]. The vectorial version, $p_2 = 1$, of this model was proposed and illustrated (but not studied further) in [12].

Finally, as observed in Section 4, the Gaussian version of the augmentation procedure turned out to perform remarkably well in dimension estimation under the Poisson model, even though it quite severely violates the model assumptions. This interesting fact thus also warrants more study.

Declaration of Competing Interest

The authors declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

Data availability

Data will be made available on request.

Acknowledgments

The work of JV was supported by the Academy of Finland (Grants 335077, 347501 and 353769). The work was supported by LMS Research in Pairs (Grant #41821) awarded to AA to initiate the collaboration with JV.

Appendix A. Limiting distribution of S_1 for $p_1 = p_2 = 1$

For convenience, we introduce the following notation more suited to the current one-dimensional case: Let $z \sim \mathcal{N}(0,1)$ and let $x \mid z \sim \text{Po}\{\exp(\mu + \sigma z)\}$. Moreover, with $x_1, \dots, x_n$ denoting a random sample from the previous model, we let $m_{n1} := (1/n)\sum_{i=1}^n x_i$, $m_{n2} := (1/n)\sum_{i=1}^n x_i(x_i - 1)$, $m_1 := E(x)$ and $m_2 := E\{x(x-1)\}$. Consequently, our objective is to find the limiting distribution of

$$h_n := \sqrt{n}\left\{\log\left(\frac{m_{n2}}{m_{n1}^2}\right) - \log\left(\frac{m_2}{m_1^2}\right)\right\}.$$

By the CLT, the limiting distribution of $\sqrt{n}(m_{n1} - m_1, m_{n2} - m_2)^\top$ is $\mathcal{N}_2(0, \Sigma)$, where

$$\Sigma := \begin{pmatrix} \text{Var(x)} & \text{Cov}\{x, x(x-1)\} \\ \text{Cov}\{x, x(x-1)\} & \text{Var}\{x(x-1)\} \end{pmatrix}.$$

Arguing as in the proof of Lemma 1 and using the fact that the jth factorial moment of the Poisson distribution is equal to the jth power of its mean, we find that the jth factorial moment $m_j := \text{E}\{x(x-1)\cdots(x-j+1)\}$ of x satisfies $m_j = \exp\{j\mu + (1/2)j^2\sigma^2\}$. The first four factorial and regular moments $t_j := \text{E}(x^j)$ have the relationships $t_1 = m_1$, $t_2 = m_1 + m_2$, $t_3 = m_1 + 3m_2 + m_3$ and $t_4 = m_1 + 7m_2 + 6m_3 + m_4$, implying that

$$\Sigma = \begin{pmatrix} m_1 + m_2 - m_1^2 & 2m_2 + m_3 - m_1 m_2 \\ 2m_2 + m_3 - m_1 m_2 & 2m_2 + 4m_3 + m_4 - m_2^2 \end{pmatrix}.$$

As the gradient of the map $(a_1, a_2) \mapsto \log(a_2/a_1^2)$ is $(-2/a_1, 1/a_2)$, the delta method entails that the limiting distribution of h_n is normal with zero mean and the variance

$$(-2m_1^{-1}, m_2^{-1})\Sigma(-2m_1^{-1}, m_2^{-1})^\top = -4m_1^{-1} + 4m_2 m_1^{-2} - 4m_3 m_1^{-1} m_2^{-1} + 2m_2^{-1} + 4m_3 m_2^{-2} + m_4 m_2^{-2} - 1.$$

Plugging in the values $m_j = \exp\{j\mu + (1/2)j^2\sigma^2\}$ now reveals that the asymptotic variance does not indeed simplify any further.

Appendix B. Proofs

Proof of Lemma 1.. It is sufficient to show the claim for S_1, as the result for S_2 follows instantly after transposition of the model.

By the law of total expectation,

$$\mathrm{E}(x_{j\ell}) = \mathrm{E}\{\mathrm{E}(x_{j\ell} \mid Z)\} = \mathrm{E}\{\exp(\mu_{j\ell} + {u_{1,j}}^\top Z u_{2,\ell})\},$$

where $u_{a,j} \in \mathbb{R}^{d_a}$ denotes the jth row of U_a, for $a = 1, 2$. The distribution of ${u_{1,j}}^\top Z u_{2,\ell}$ is $\mathcal{N}(0, \tau^2 {u_{1,j}}^\top \Lambda_1 u_{1,j} {u_{2,\ell}}^\top \Lambda_2 u_{2,\ell})$, showing that

$$\mathrm{E}(x_{j\ell}) = \exp\{\mu_{j\ell} + (1/2)\tau^2 {u_{1,j}}^\top \Lambda_1 u_{1,j} {u_{2,\ell}}^\top \Lambda_2 u_{2,\ell}\}.$$

One can similarly establish that

$$\mathrm{E}\{x_{j\ell}(x_{j\ell} - 1)\} = \exp(2\mu_{j\ell} + 2\tau^2 {u_{1,j}}^\top \Lambda_1 u_{1,j} {u_{2,\ell}}^\top \Lambda_2 u_{2,\ell}),$$

and that

$$\begin{aligned}\mathrm{E}(x_{j\ell}x_{k\ell}) = \exp\{&\mu_{j\ell} + \mu_{k\ell} + (1/2)\tau^2 (u_{1,j} + u_{1,k})^\top \\ &\times \Lambda_1 (u_{1,j} + u_{1,k}) {u_{2,\ell}}^\top \Lambda_2 u_{2,\ell}\},\end{aligned}$$

where the latter result assumes $j \neq k$ and uses the conditional independence of $x_{j\ell}$ and $x_{k\ell}$ given Z. Plugging now in to the definitions of $s_{1,jk}$ and $s_{1,jj}$ in (6) gives

$$\begin{aligned}s_{1,jk} &= \frac{1}{p_2}\sum_{\ell=1}^{p_2} \tau^2 {u_{1,j}}^\top \Lambda_1 u_{1,k} {u_{2,\ell}}^\top \Lambda_2 u_{2,\ell} \\ &= \frac{1}{p_2}\mathrm{tr}(U_2 \Lambda_2 U_2^\top)\tau^2 {u_{1,j}}^\top \Lambda_1 u_{1,k} = \tau^2 {u_{1,j}}^\top \Lambda_1 u_{1,k},\end{aligned}$$

showing the claim for the off-diagonal elements $s_{1,jk}$. A similar plug-in gives the analogous result also for the diagonal elements $s_{1,jj}$. □

Proof of Theorem 1.. Fix $j, \ell = 1, \dots, p_2$. A first-order Taylor expansion around $\mathrm{E}(x_{j\ell})$ shows that,

$$\begin{aligned}&\sqrt{n}\left\{\log\left(\frac{1}{n}\sum_{i=1}^n x_{i,j\ell}\right) - \log \mathrm{E}(x_{j\ell})\right\} \\ =&\frac{\sqrt{n}}{{e_j}^\top \mathrm{E}(X) e_\ell} {e_j}^\top \left\{\frac{1}{n}\sum_{i=1}^n X_i - \mathrm{E}(X)\right\} e_\ell + o_p(1) \\ =&\frac{1}{{e_j}^\top \mathrm{E}(X) e_\ell} {f_{\ell j}}^\top \sqrt{n}(\bar{m} - m) + o_p(1),\end{aligned}$$

where we use the notation $\bar{m} := (1/n)\sum_{i=1}^n \mathrm{vec}(X_i)$ and $m := \mathrm{E}\{\mathrm{vec}(X)\}$. The same steps can be used to show that,

$$\begin{aligned}&\sqrt{n}\left\{\log\left(\frac{1}{n}\sum_{i=1}^n x_{i,j\ell}x_{i,k\ell}\right) - \log \mathrm{E}(x_{j\ell}x_{k\ell})\right\} \\ =&\frac{1}{{f_{kj}}^\top \mathrm{E}(X \otimes X) f_{\ell\ell}} (f_{\ell\ell} \otimes f_{kj})^\top \sqrt{n}(\bar{c} - c) + o_p(1),\end{aligned}$$

where $\bar{c} := (1/n)\sum_{i=1}^n \mathrm{vec}(X_i \otimes X_i)$ and $c := \mathrm{E}\{\mathrm{vec}(X \otimes X)\}$. Similarly, we get that,

$$\begin{aligned}&\sqrt{n}\left[\log\left\{\frac{1}{n}\sum_{i=1}^n x_{i,j\ell}(x_{i,j\ell} - 1)\right\} - \log \mathrm{E}\{x_{j\ell}(x_{j\ell} - 1)\}\right] \\ =&\frac{1}{b_{j\ell}}(f_{\ell\ell} \otimes f_{jj})^\top \sqrt{n}(\bar{c} - c) - \frac{1}{b_{j\ell}} {f_{\ell j}}^\top \sqrt{n}(\bar{m} - m) + o_p(1),\end{aligned}$$

where $b_{j\ell}$ is as in the statement of the theorem.

We now observe that the elements of $\sqrt{n}(S_{n1} - S_1)$ and $\sqrt{n}(M_n - \mu)$ are linear functions of the previous three linearizations of the logarithmic first and second sample moments of X. Thus, all three claims of the theorem follow after collecting the coefficients corresponding to $\sqrt{n}(\bar{c} - c)$ and $\sqrt{n}(\bar{m} - m)$ into the vectors h_{jk}, h_{jj} and a_{jk}. □

Proof of Lemma 2.. The vectorization of the model (3) reads,

$$x = m + Uz + \mathrm{vec}(\varepsilon),$$

where $z \sim \mathcal{N}_d(0, \tau^2\Lambda)$ is independent of $\mathrm{vec}(\varepsilon) \sim \mathcal{N}_p(0, \sigma^2 I_p)$. Consequently, the joint distribution of $(z^\top, x^\top)^\top$ is

$$\begin{pmatrix} z \\ x \end{pmatrix} \sim \mathcal{N}_{d+p}\left\{\begin{pmatrix} 0 \\ m \end{pmatrix}, \begin{pmatrix} \tau^2\Lambda & \tau^2\Lambda U^\top \\ \tau^2 U\Lambda & \tau^2 U\Lambda U^\top + \sigma^2 I_p \end{pmatrix}\right\}.$$

Now, $(\tau^2\Lambda U^\top)(\tau^2 U\Lambda U^\top + \sigma^2 I_p)^{-1} = \tau^2\Lambda(\tau^2\Lambda + \sigma^2 I_d)^{-1}U^\top$, and by the standard properties of Gaussian conditional distributions,

$$z \mid x \sim \mathcal{N}_d\{a(x), B\},$$

where $a(x) := \tau^2\Lambda(\tau^2\Lambda + \sigma^2 I_d)^{-1}U^\top(x - m)$ and $B := \{I_d - \tau^2\Lambda(\tau^2\Lambda + \sigma^2 I_d)^{-1}\}\tau^2\Lambda$. The result now follows. □

Proof of Theorem 2.. The Bayes rule implies that,

$$f_{z|x}(z|x) = C_0 f_{x|z}(x \mid z) f_z(z), \tag{B.1}$$

for some positive constant C_0 not depending on z. Then, indexing the elements of x as x_j, $j = 1, \dots, p$, we have,

$$\log f_{x|z}(x \mid z) = -\sum_{j=1}^p \log x_j! + x^\top h(z) - 1^\top \exp\{h(z)\},$$

where the exponential function is taken element-wise and we denote $h(z) := m + Uz$. Additionally,

$$\log f_z(z) = -\frac{d}{2}\log 2\pi - \frac{1}{2}\tau^{2d}\log|\Lambda| - \frac{1}{2\tau^2} z^\top \Lambda^{-1} z.$$

Plugging the previous two formulas into (B.1) now yields result *i)*.

To see *ii)*, we fix x and first establish the strict concavity by showing that the Hessian of $z \mapsto \ell(z|x)$ is negative definite. The gradient of the map is

$$\nabla \ell(z|x) = U^\top x - U^\top \exp\{h(z)\} - \tau^{-2}\Lambda^{-1} z,$$

giving further the Hessian,

$$\nabla^2 \ell(z|x) = -U^\top \mathrm{diag}[\exp\{h(z)\}]U - \tau^{-2}\Lambda^{-1}.$$

The diagonal matrices $\mathrm{diag}[\exp\{h(z)\}]$ and Λ^{-1} have strictly positive diagonal elements (and $\tau > 0$), making the Hessian negative definite and implying that $\ell(z|x)$ is indeed strictly concave in z, for all x.

We next show that $z \mapsto \ell(z|x)$ is coercive in the sense that, for all sequences $z_n \in \mathbb{R}^d$ such that $\|z_n\| \to \infty$, we have $\ell(z_n|x) \to -\infty$. Combined with the continuity of $z \mapsto \ell(z|x)$, the coercivity will then imply that the function admits at least one global maximizer, and strict concavity then guarantees that the maximizer is unique.

Let thus $z_n \in \mathbb{R}^d$ be such that $\|z_n\| \to \infty$ and write $\alpha_n := \|z_n\| \to \infty$ and $u_n := z_n/\|z_n\|$. Then,

$$\begin{aligned}\ell(z_n|x) &= C + \alpha_n x^\top U u_n - 1^\top \exp\{h(\alpha_n u_n)\} - \frac{\alpha_n^2}{2\tau^2} u_n^\top \Lambda^{-1} u_n \\ &\leq C + \alpha_n x^\top U u_n - \frac{\alpha_n^2}{2\tau^2} u_n^\top \Lambda^{-1} u_n \\ &\leq C + \alpha_n \|U^\top x\| - \frac{\alpha_n^2}{2\tau^2}\phi_d(\Lambda^{-1}),\end{aligned}$$

where $\phi_d(\Lambda^{-1}) > 0$ denotes the smallest eigenvalue of the positive definite matrix Λ^{-1}. The derived upper bound is dominated by the quadratic term, thus guaranteeing that $\ell(z_n|x) \to -\infty$, as desired. □

References

[1] J. Aitchison, C. Ho, The multivariate Poisson-log normal distribution, Biometrika 76 (4) (1989) 643–653.
[2] V. Bally, L. Caramellino, Asymptotic development for the CLT in total variation distance, Bernoulli 22 (4) (2016) 2442–2485.
[3] J. Chiquet, M. Mariadassou, S. Robin, Variational inference for probabilistic Poisson PCA, Annals of Applied Statistics 12 (4) (2018) 2674–2698.
[4] M. Collins, S. Dasgupta, R.E. Schapire, A generalization of principal components analysis to the exponential family, in: NIPS 2001, volume 13, 2001, p. 23.
[5] L. De Lathauwer, B. De Moor, J. Vandewalle, A multilinear singular value decomposition, SIAM J. Matrix Anal. Appl. 21 (4) (2000) 1253–1278.
[6] S. Ding, R.D. Cook, Dimension folding PCA and PFC for matrix-valued predictors, Stat Sin 24 (1) (2014) 463–492.
[7] S. Ding, R.D. Cook, Tensor sliced inverse regression, J Multivar Anal 133 (2015) 216–231.
[8] M.L. Eaton, D.E. Tyler, On Wielandt's inequality and its application to the asymptotic distribution of the eigenvalues of a random symmetric matrix, Ann Stat 19 (1) (1991) 260–271.
[9] R. Frelat, M. Lindegren, T.S. Denker, J. Floeter, H.O. Fock, C. Sguotti, M. Stäbler, S.A. Otto, C. Möllmann, Community ecology in 3D: tensor decomposition reveals spatio-temporal dynamics of large ecological communities, PLoS ONE 12 (11) (2017) e0188205.
[10] A.K. Gupta, D.K. Nagar, Matrix Variate Distributions, volume 104, CRC Press, 2018.
[11] P. Hall, J.T. Ormerod, M.P. Wand, Theory of Gaussian variational approximation for a Poisson mixed model, Stat Sin 21 (1) (2011) 369–389.
[12] M. Hartmann, Extending Owen's integral table and a new multivariate bernoulli distribution, arXiv preprint arXiv:1704.04736 (2017).
[13] C. Hu, P. Rai, C. Chen, M. Harding, L. Carin, Scalable Bayesian non-negative tensor factorization for massive count data, in: Joint European Conference on Machine Learning and Knowledge Discovery in Databases, Springer, 2015, pp. 53–70.
[14] H. Hung, P. Wu, I. Tu, S. Huang, On multilinear principal component analysis of order-two tensors, Biometrika 99 (3) (2012) 569–583.
[15] R. Izsák, Maximum likelihood fitting of the poisson lognormal distribution, Environ Ecol Stat 15 (2) (2008) 143–156.
[16] T. Kenney, H. Gu, T. Huang, Poisson PCA: poisson measurement error corrected PCA, with application to microbiome data, Biometrics (2019).
[17] T.G. Kolda, B.W. Bader, Tensor decompositions and applications, SIAM Rev. 51 (3) (2009) 455–500.
[18] C. Lam, High-dimensional covariance matrix estimation, Wiley Interdiscip. Rev. Comput. Stat. 12 (2) (2020) e1485.
[19] A.J. Landgraf, Generalized principal component analysis: dimensionality reduction through the projection of natural parameters, Ohio State University, 2015 Ph.D. Thesis.
[20] D. Leibovici, R. Sabatier, A singular value decomposition of a *k*-way array for a principal component analysis of multiway data, PTA-*k*, Linear Algebra Appl 269 (1–3) (1998) 307–329.
[21] B. Li, M.K. Kim, N. Altman, On dimension folding of matrix-or array-valued statistical objects, Ann Stat 38 (2) (2010) 1094–1121.
[22] J. Li, D. Tao, Simple exponential family PCA, in: Proceedings of the Thirteenth International Conference on Artificial Intelligence and Statistics, JMLR Workshop and Conference Proceedings, 2010, pp. 453–460.
[23] S. Liu, Z. Chen, X. Li, Time-semantic-aware Poisson tensor factorization approach for scalable hotel recommendation, Inf Sci (Ny) 504 (2019) 422–434.
[24] W. Luo, B. Li, On order determination by predictor augmentation, Biometrika 108 (3) (2020) 557–574.
[25] X. Mao, R.K. Wong, S.X. Chen, Matrix completion under low-rank missing mechanism, arXiv preprint arXiv:1812.07813 (2018).
[26] J. Niku, W. Brooks, R. Herliansyah, F.K. Hui, S. Taskinen, D.I. Warton, B. van der Veen, GLLVM: Generalized Linear Latent Variable Models, 2020. R package version 1.2.3, https://CRAN.R-project.org/package=gllvm.
[27] J. Niku, D.I. Warton, F.K. Hui, S. Taskinen, Generalized linear latent variable models for multivariate count and biomass data in ecology, Journal of Agricultural, Biological and Environmental Statistics 22 (4) (2017) 498–522.
[28] K. Nordhausen, D.E. Tyler, A cautionary note on robust covariance plug-in methods, Biometrika 102 (3) (2015) 573–588.
[29] D.B. Owen, A table of normal integrals, Communications in Statistics - Simulation and Computation 9 (4) (1980) 389–419.
[30] R Core Team, R: A Language and Environment for Statistical Computing, R Foundation for Statistical Computing, 2020. Vienna, Austria https://www.R-project.org/.
[31] U. Radojičić, N. Lietzén, K. Nordhausen, J. Virta, On estimating the latent dimension in two-dimensional PCA, in: Proceedings of the 12 International Symposium on Image and Signal Processing and Analysis (ISPA 2021), 2021, pp. 16–22.
[32] A. Schein, J. Paisley, D.M. Blei, H. Wallach, Bayesian Poisson tensor factorization for inferring multilateral relations from sparse dyadic event counts, in: Proceedings of the 21th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, 2015, pp. 1045–1054.
[33] L. Smallman, A. Artemiou, J. Morgan, Sparse generalised principal component analysis, Pattern Recognit 83 (2018) 443–455.
[34] L. Smallman, W. Underwood, A. Artemiou, Simple Poisson PCA: an algorithm for (sparse) feature extraction with simultaneous dimension determination, Comput Stat 35 (2020) 559–577.
[35] D.E. Tyler, Asymptotic inference for eigenvectors, Ann Stat 9 (4) (1981) 725–736.
[36] D.E. Tyler, F. Critchley, L. Dümbgen, H. Oja, Invariant co-ordinate selection, Journal of the Royal Statistical Society: Series B (Statistical Methodology) 71 (3) (2009) 549–592.
[37] J. Virta, C.L. Koesner, B. Li, K. Nordhausen, H. Oja, tensorBSS: Blind Source Separation Methods for Tensor-Valued Observations, 2016. R package version 0.3.8, https://www.CRAN.R-project.org/package=tensorBSS.
[38] J. Virta, B. Li, K. Nordhausen, H. Oja, Independent component analysis for tensor-valued data, J Multivar Anal 162 (2017) 172–192.
[39] M. Wedel, U. Böckenholt, W.A. Kamakura, Factor models for multivariate count data, J Multivar Anal 87 (2) (2003) 356–369.
[40] D. Zhang, Z.-H. Zhou, $(2D)^2$PCA: Two-directional two-dimensional PCA for efficient face representation and recognition, Neurocomputing 69 (1–3) (2005) 224–231.

Joni Virta received his Ph.D. from University of Turku, Finland, in 2018, on the topic of independent component analysis. Currently he is working as an Academy Research Fellow / Assistant Professor in University of Turku as a PI of the Academy of Finland project "Robust non-linear multivariate methods". His main research interests include dimension reduction and asymptotics, especially in the context of non-standard forms of data.

Andreas Artemiou is a Reader in Statistics at the School of Mathematics in Cardiff University. He obtained his B.Sc. in 2005 from University of Cyprus and his M.Sc. in 2008 and Ph.D. in 2010 from Pennsylvania State University. His research interests include supervised and unsupervised dimension reduction methodology, statistical and machine learning, kernel methods and applications. His research has been funded by National Science Foundation, the London Mathematical Society, the GW4 Network and the Wellcome Trust.