# Poisson PCA for matrix count data

**J. Virta**[1] A. Artemiou[2]

[1]University of Turku

[2]University of Limassol

Statistical Days 2024

# Reference

This talk is based on the paper

- **Virta, J. and Artemiou, A.** (2023). Poisson PCA for matrix count data. *Pattern Recognition*, 138:109401.

# Table of Contents

# Abundance matrix data

- Let $X_1, \ldots, X_n \in \mathbb{R}^{p_1 \times p_2}$ be a sample of abundance matrices for $n$ species.
  - $n =$ number of animal species,
  - $p_1 =$ number of areas,
  - $p_2 =$ number of time periods.

## Abundance matrix for the $i$th species

$$X_i = \begin{array}{c} \\ \text{Area}_1 \\ \vdots \\ \text{Area}_{p_1} \end{array} \begin{array}{c} \begin{array}{ccc} \text{Time}_1 & \cdots & \text{Time}_{p_2} \end{array} \\ \begin{pmatrix} x_{i,11} & \cdots & x_{i,1p_2} \\ \vdots & \ddots & \vdots \\ x_{i,p_1 1} & \cdots & x_{i,p_1 p_2} \end{pmatrix} \end{array}$$

# Vectorization

- One way to model $X_1, \ldots, X_n$ is to vectorize them and use standard multivariate models.

$$X_i \in \mathbb{R}^{p_1 \times p_2} \quad \mapsto \quad \mathrm{vec}(X_i) = \begin{pmatrix} x_{i,11} \\ x_{i,21} \\ \vdots \\ x_{i,p_1 p_2} \end{pmatrix} \in \mathbb{R}^{p_1 p_2}.$$

- This (a) loses the matrix row-column structure, and (b) leads to high-dimensional vectors.

# Row-column paradigm

- Often the rows and columns of $X_i$ are modeled separately
- E.g., instead of working with the linear combinations

$$\beta^\top \mathrm{vec}(X_i), \quad \beta \in \mathbb{R}^{p_1 p_2},$$

we work with the linear combinations

$$\beta_1^\top X_i \beta_2 \quad \beta_1 \in \mathbb{R}^{p_1}, \beta_2 \in \mathbb{R}^{p_2}.$$

- The "weight" of an element is the combined weight of its row and column.

# Row-column paradigm, cont.

- In row-column modeling, the number of parameters drops from $p_1 p_2$ to $p_1 + p_2$.
- In many applications, it captures the essential structure of the data, see the references in [Virta and Artemiou, 2023].

# Table of Contents

# Matrix-variate normal distribution

- We want to fit a factor model to the abundance data.
- A classical option is the matrix-variate normal distribution [Gupta and Nagar, 2018],

$$X_i \sim \mathcal{N}_{p_1 \times p_2}(\mu, \Sigma_1, \Sigma_2),$$

  - $\mu =$ the mean matrix,
  - $\Sigma_1 =$ the row covariance matrix,
  - $\Sigma_2 =$ the column covariance matrix.
- Not a natural choice for count-valued data.

# Poisson factor model

## Poisson-Normal mixture

We assume a hierarchical model for the $X_i$:

$$\begin{cases} Z_i \sim \mathcal{N}_{d_1 \times d_2}(0, \Lambda_1, \Lambda_2) \\ X_i \mid Z_i \sim \mathrm{Po}_{p_1 \times p_2}\{\exp(\mu + U_1 Z_i U_2^\top)\}, \end{cases}$$

- $Z_i$ = the latent factor matrix of the $i$th species.
- $\Lambda_1, \Lambda_2$ = diagonal matrices giving the importances of the row factors and column factors,
- $U_1, U_2$ = the row and column loadings,
- $\mu$ = mean shift.

# Poisson factor model, cont.

$Z_i$

$$X_i \mid Z_i \sim \mathrm{Po}_{p_1 \times p_2}\{\exp(\mu + U_1 Z_i U_2^\top)\}$$

$$\begin{pmatrix} \cdot & \cdot \\ \cdot & \cdot \\ \cdot & \cdot \end{pmatrix}$$

$$\begin{pmatrix} \cdot & \cdot & \cdot & \cdot & \cdot & \cdot & \cdot \\ \cdot & \cdot & \cdot & \cdot & \cdot & \cdot & \cdot \\ \cdot & \cdot & \cdot & \cdot & \cdot & \cdot & \cdot \\ \cdot & \cdot & \cdot & \cdot & \cdot & \cdot & \cdot \\ \cdot & \cdot & \cdot & \cdot & \cdot & \cdot & \cdot \end{pmatrix}$$

# Poisson factor model, cont.

- The model uses an exponential "link function" between continuous latent variables and discete observed data.
- Compared to GLM, we do not observe the $Z_i$

- To fit the model to the data, we have to
  1. estimate the parameters,
  2. estimate the dimensions $d_1, d_2$,
  3. predict the latent factor matrices.
- Usually $p_1 \gg d_1, p_2 \gg d_2$, leading to considerable dimension reduction.

# Vector Poisson-Normal mixture

- If $p_2 = 1$, we obtain a vector Poisson-Normal mixture that was originally proposed by [Aitchison and Ho, 1989] under the title of Poisson-Lognormal model.
- The model was later studied by [Hall et al., 2011, Kenney et al., 2021].

# Parameter estimation

- The method of moments yields closed-form solutions for the model parameters.

$$\theta := (\mu, U_1, \Lambda_1, U_2, \Lambda_2).$$

- For example, the left loadings $U_1$ and scale $\Lambda_1$ can be estimated from the eigendecomposition of the matrix $S_1$ defined as,

$$s_{1,jk} := \frac{1}{p_2} \sum_{\ell=1}^{p_2} \log \left\{ \frac{\mathrm{E}(x_{j\ell} x_{k\ell})}{\mathrm{E}(x_{j\ell})\mathrm{E}(x_{k\ell})} \right\},$$

$$s_{1,jj} := \frac{1}{p_2} \sum_{\ell=1}^{p_2} \log \left[ \frac{\mathrm{E}\{x_{j\ell}(x_{j\ell} - 1)\}}{\{\mathrm{E}(x_{j\ell})\}^2} \right],$$

# Dimension estimation

We estimate the latent dimensions $d_1, d_2$ using *predictor augmentation* [Luo and Li, 2020]:

- The dimension $d_1$ equals the rank of the matrix $S_1$.
- We augment the observed matrices with noise,

$$X_i^* = \begin{pmatrix} X_i \\ R_i \end{pmatrix},$$

  where $R_i$ have iid $\mathrm{Poisson}(1)$-elements.
- By comparing the sample estimates $S_{n1}$ and $S_{n1}^*$ it is possible to identify where the $d_1$-dimension signal "ends".

# Augmentation curve

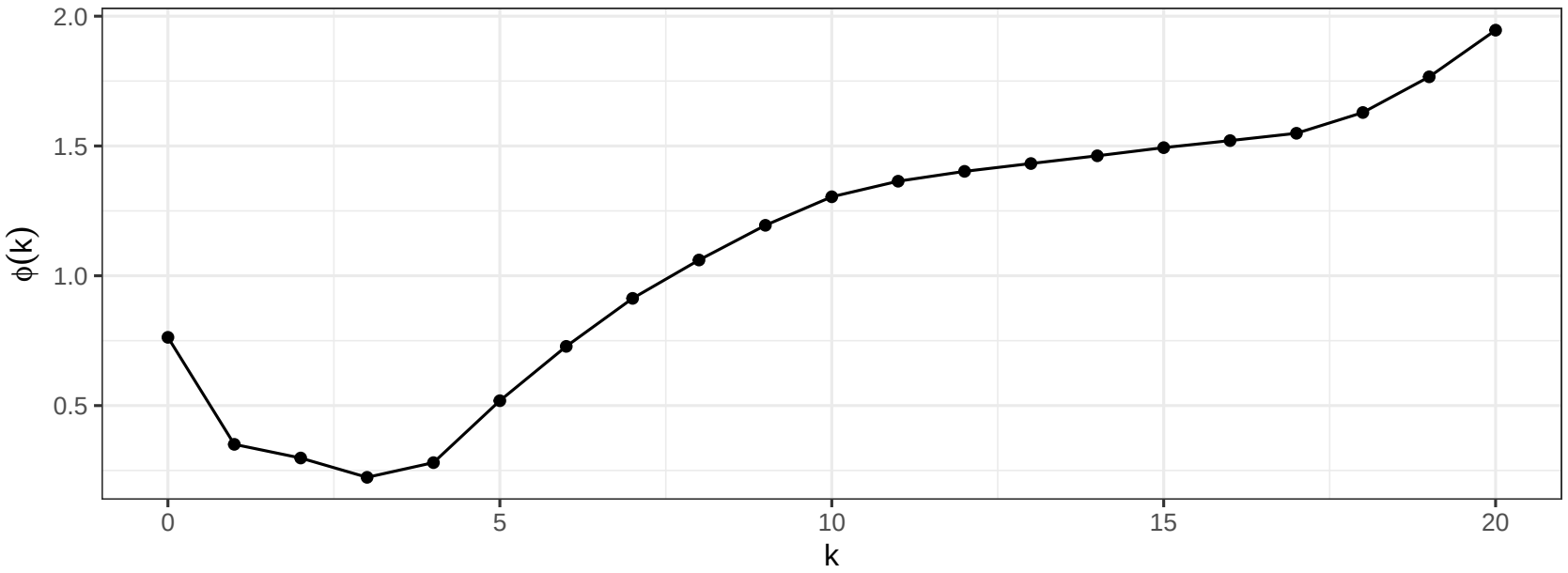



Figure: The minimum of the augmentation curve is achieved at $d_1 = 3$.

# Latent factor prediction

We predict $Z_i$ as the mode of the conditional distribution $Z_i \mid X_i$

Gradient descent can be used for finding the mode as $Z_i \mid X_i$ has a log-concave unimodal density function.

# Table of Contents

# Data description

- Abundance data available at `https://github.com/rfrelat/Multivariate2D3D`.
- The data consists of abundances of a total of $n = 65$ fish species
  - in seven areas (RA 1 – RA 7), $p_1 = 7$,
  - during 6 time periods $(1985 - 1989, \ldots, 2005 - 2009, 2010 - 2015)$, $p_2 = 6$.

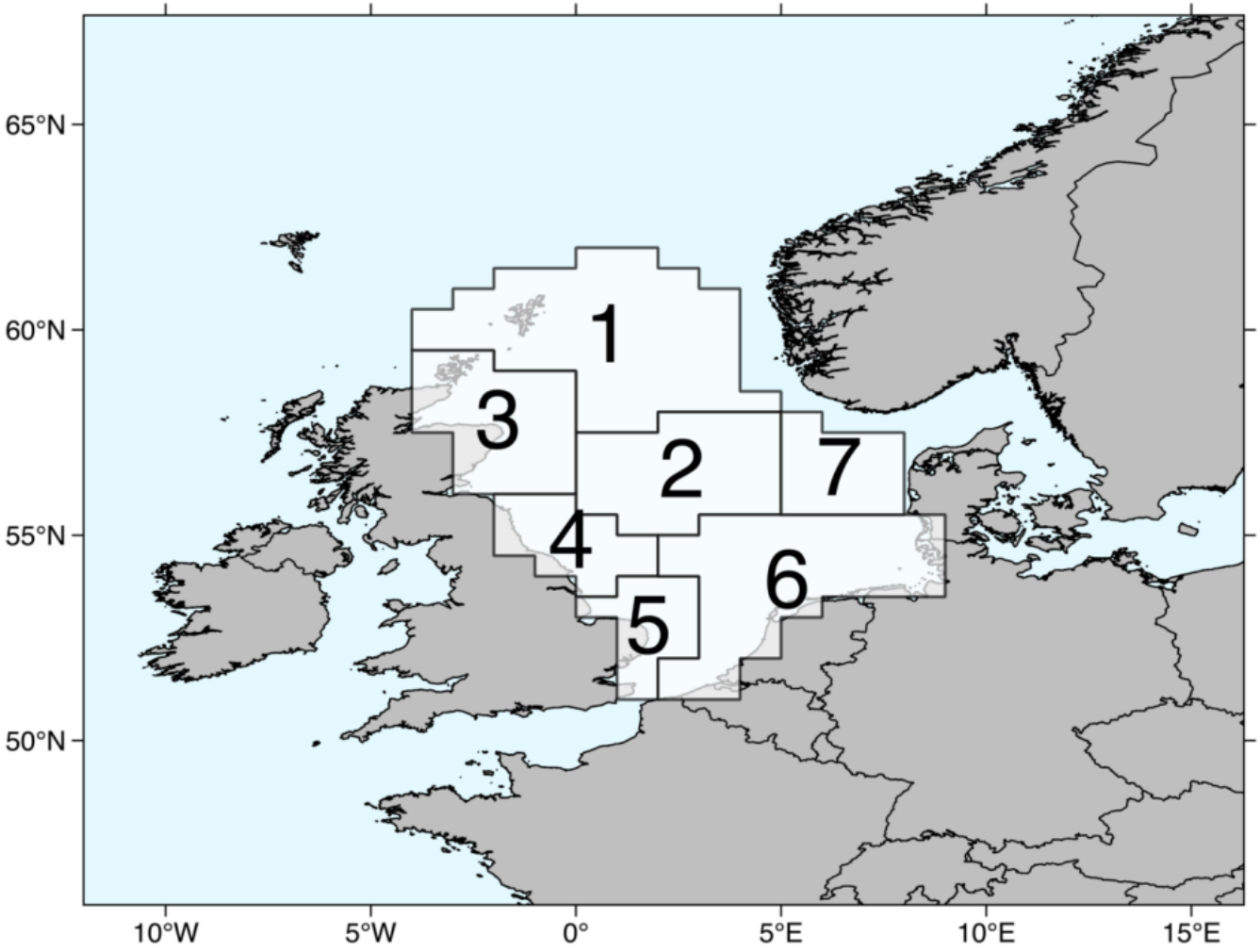



Figure: The seven areas in the study. Image from [Frelat et al., 2017].

# Dimension estimation

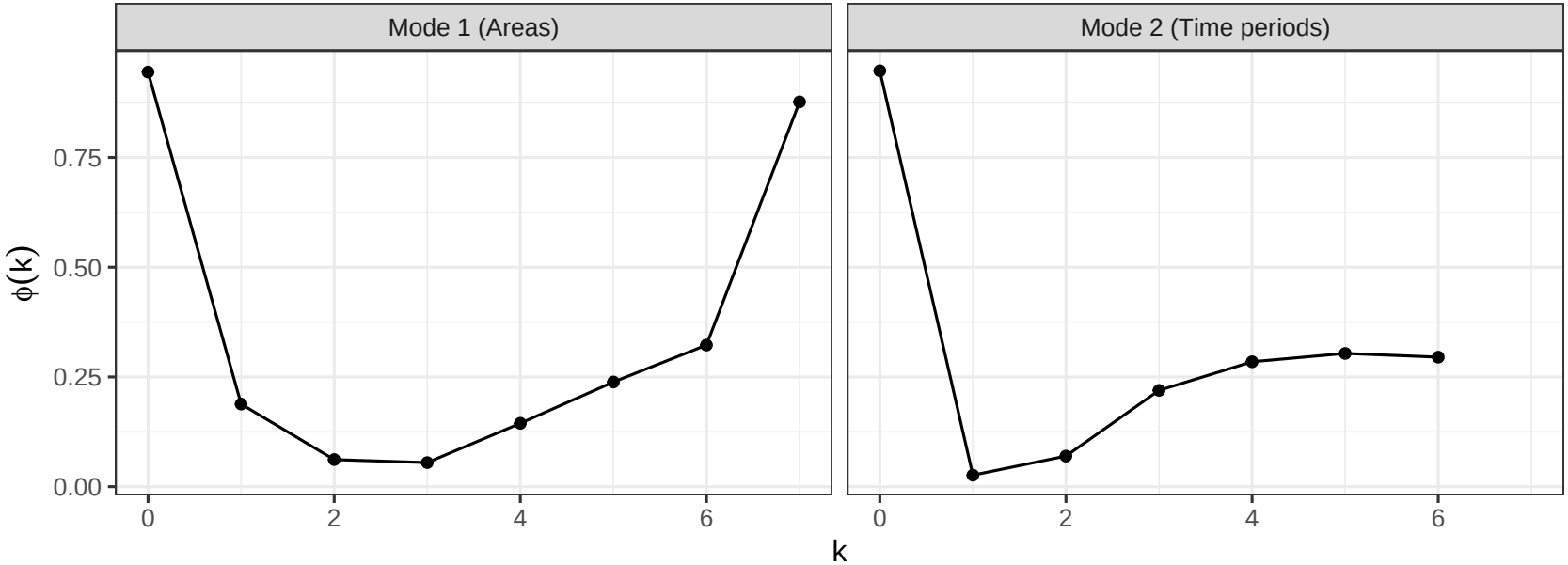



Figure: The dimensions are estimated as $d_1 = 3$ and $d_2 = 1$.

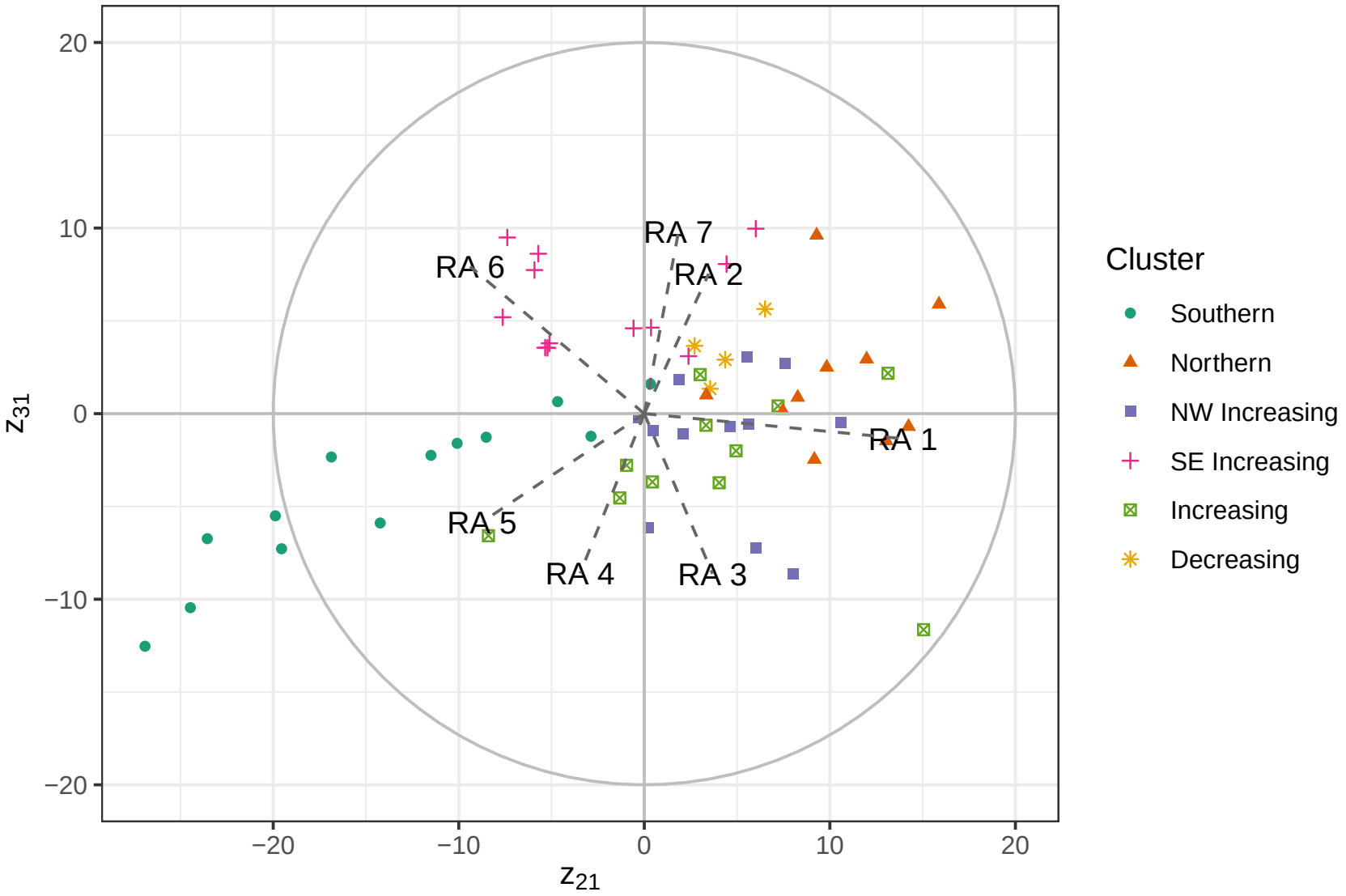
# Latent components

- We estimate $(3 \times 1)$-sized latent matrices,

$$Z_i = \begin{pmatrix} z_{i,11} \\ z_{i,21} \\ z_{i,31} \end{pmatrix}.$$

- $z_{i,11}$ measures the overall abundances of the species.
- We plot $z_{i,21}, z_{i,31}$ with the row loadings into a biplot.
- We color the species according to the six biologically meaningful clusters [Frelat et al., 2017] identified in the data.

# Biplot


RA 1
RA 2
RA 3
RA 4
RA 5
RA 6
RA 7
$z_{31}$
$z_{21}$
20
10
0
−10
−20
Cluster
Southern
Northern
NW Increasing
SE Increasing
Increasing
Decreasing

# Table of Contents

# Possible next steps

- Zero-inflated variant for sparse matrix count data?
- Bernoulli-Normal mixture for binary matrix data?

Thank you for your attention!

# References I


Aitchison, J. and Ho, C. (1989).
The multivariate Poisson-log normal distribution.
*Biometrika*, 76(4):643–653.

Frelat, R., Lindegren, M., Denker, T. S., Floeter, J., Fock, H. O., Sguotti, C., Stäbler, M., Otto, S. A., and Möllmann, C. (2017).
Community ecology in 3D: Tensor decomposition reveals spatio-temporal dynamics of large ecological communities.
*PloS one*, 12(11):e0188205.

Gupta, A. K. and Nagar, D. K. (2018).
*Matrix Variate Distributions*, volume 104.
CRC Press.

# References II


Hall, P., Ormerod, J. T., and Wand, M. P. (2011).
Theory of Gaussian variational approximation for a Poisson mixed model.
*Statistica Sinica*, pages 369–389.

Kenney, T., Gu, H., and Huang, T. (2021).
Poisson PCA: Poisson measurement error corrected PCA, with application to microbiome data.
*Biometrics*.

Luo, W. and Li, B. (2020).
On order determination by predictor augmentation.
*Biometrika*.

Virta, J. and Artemiou, A. (2023).
Poisson PCA for matrix count data.
*Pattern Recognition*, 138:109401.

# Table of Contents

# Efficiency study

- We simulated samples of $4 \times 3$ matrices from the Poisson factor model and estimated its parameters using three different methods:
  - Our proposal.
  - Vectorizing and method of moments [Aitchison and Ho, 1989].
  - Vectorizing and MLE with variational inference [Hall et al., 2011].
- Average errors over 1000 replicates for various sample sizes and covariance structures are shown on the following slide.

# Efficiency plot

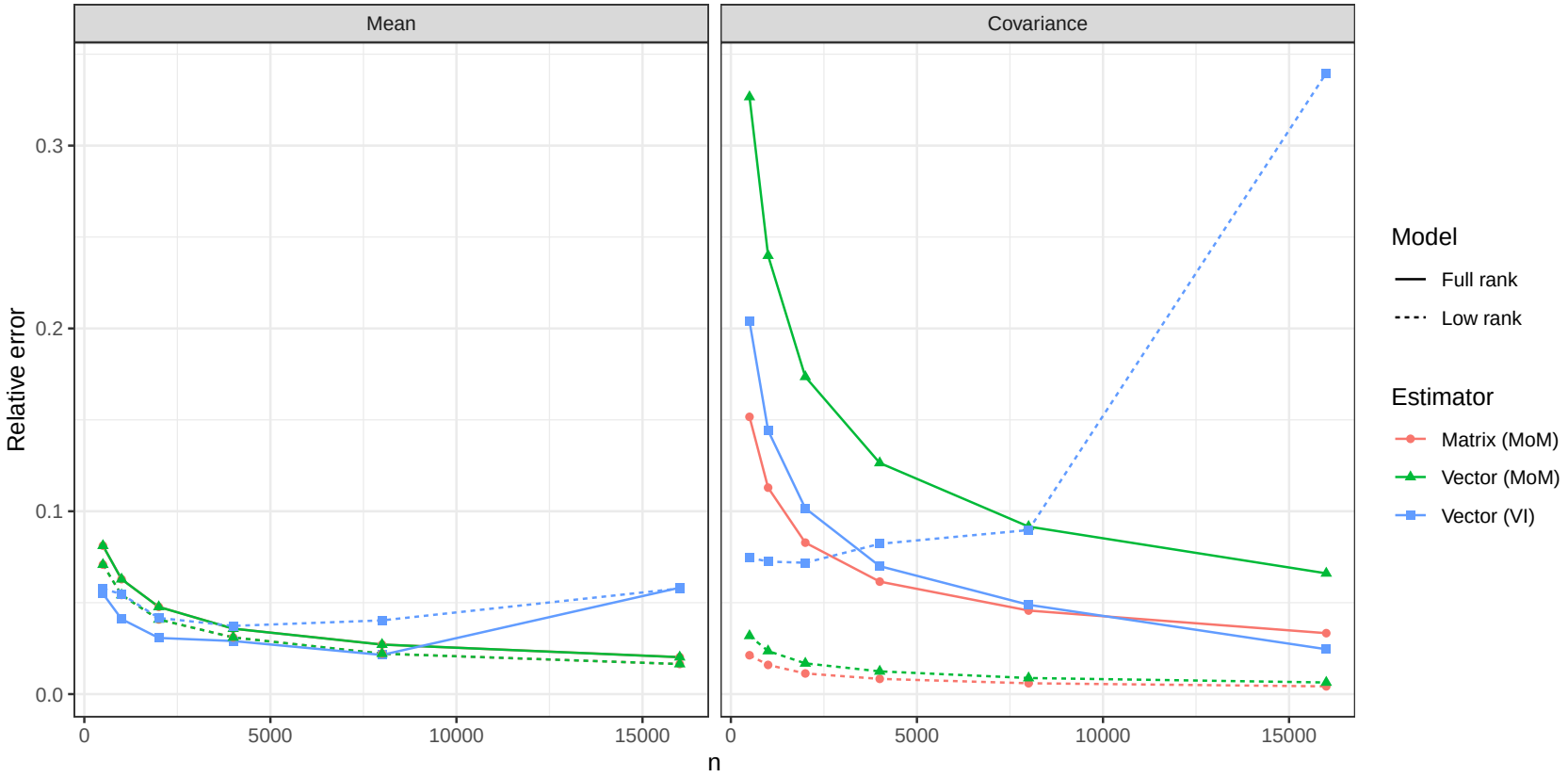



Figure: Average estimation errors of the three methods.