Model-based mixed-type PCA

Lauri Heinonen & Joni Virta

University of Turku

Nordstat 2026, Helsinki

Topic

- We present a method for principal component analysis on a mixed-type data.
- By mixed-type data we mean that the variables are of continuous, binary, integer-valued or positive continuous type.
- We use the method of moments to estimate the covariance matrix of combinations of latent components and then use regular PCA to solve for the latent components.
- The manuscript will be submitted soon and made available in ArXiv.

Outline

- Underlying probability model
- Estimating the loadings
- Estimating the scores
- Adding sparsity
- Results from a simulation and a real data example

Probability model

- Observed variables have a appropriate exponential family distributions.
- We assume that the latent principal components have a Gaussian distribution.
- Parameters of the distributions come from linear combinations of the principal components through appropriate link functions.
- Our model is in a sense extension of probabilistic PCA (Tipping and Bishop, 1999) to mixed-type data.
- Similar kind of exponential family based methods for finding latent variables with mixed-type data earlier proposed by, for example, Collins et al. (2001) and Liu et al. (2023).

Probability model

Let $Z \sim \mathrm{N}_d(0, \Lambda)$ be a multivariate normal random vector with diagonal covariance matrix. Let $W := \mu + UZ$ for a vector $\mu \in \mathbb{R}^p$ and matrix $U \in \mathbb{R}^{p \times d}$ with orthonormal columns and partition W into W_1, W_2, W_3, W_4. Assume further that the observed data X depends on the vector W as follows:

$$
\begin{aligned}
X_{1j}|Z &\sim \mathrm{N}(W_{1j}, \tau_j^2), && (X_{1j} \in \mathbb{R}),\\
X_{2j}|Z &\sim \mathrm{Exp}(e^{W_{2j}}), && (X_{2j} \in \mathbb{R}_+),\\
X_{3j}|Z &\sim \mathrm{Poi}(e^{W_{3j}}), && (X_{3j} \in \mathbb{N}_0),\\
X_{4j}|Z &\sim \mathrm{Ber}(\Phi(W_{4j})), && (X_{4j} \in \{0, 1\}),
\end{aligned}
$$

where Φ is the CDF on standard normal distribution. We assume that all $\tau_j^2 > 0$ and that all elements of X are conditionally independent conditional on Z.

Probability model

- Let μ_i be an element of the mean vector μ corresponding to variable of type i.
- Let $\Sigma := \mathrm{Cov}(W) = U\Lambda U'$ be the covariance matrix of the latent combinations W.
  - Estimating this matrix is our goal!
- Let σ_i^2 a diagonal element of Σ corresponding to a variable of the type i.
- Let σ_{ij} be a non-diagonal element of Σ corresponding to variables of types i and j, with $\sigma_{ii'}$ denoting to two different variables of type i.

We have to make an additional assumption, so that the parameters can be identified: Each of the Bernoulli-variables satisfies

$$\frac{\mu_4}{\sqrt{1+\sigma_4^2}} = s\sqrt{\sigma_4^2},$$

for some sign $s \in \{-1, 1\}$.

Estimating the covariance matrix

Using the method of moments we get the following estimates

$$\hat{\mu}_1 := \overline{X_1}$$

$$\hat{\mu}_2 := -\frac{1}{2}\log 2 - 2\log \overline{X_2} + \frac{1}{2}\log \overline{X_2^2}$$

$$\hat{\mu}_3 := 2\log \overline{X_3} - \frac{1}{2}\log\left(\overline{X_3^2} - \overline{X_3}\right)$$

$$\hat{\mu}_4 := \Phi^{-1}(\overline{X_4})\sqrt{1 + \left\{\Phi^{-1}(\overline{X_4})\right\}^2}$$

$$\hat{\sigma}_1^2 := \overline{X_1^2} - \overline{X_1}^2 - \tau_j^2$$

$$\hat{\sigma}_2^2 := -\log 2 + \log \overline{X_2^2} - 2\log \overline{X_2}$$

$$\hat{\sigma}_3^2 := \log\left(\overline{X_3^2} - \overline{X_3}\right) - 2\log \overline{X_3}$$

$$\hat{\sigma}_4^2 := \left\{\Phi^{-1}(\overline{X_4})\right\}^2$$

$$\hat{\sigma}_{11'} := \overline{X_1 X_{1'}} - \overline{X_1}\ \overline{X_{1'}}$$

$$\hat{\sigma}_{22'} := \log \overline{X_2 X_{2'}} - \log \overline{X_2} - \log \overline{X_{2'}}$$

$$\hat{\sigma}_{33'} := \log \overline{X_3 X_{3'}} - \log \overline{X_3} - \log \overline{X_{3'}}$$

$$\hat{\sigma}_{12} := \frac{\overline{X_1}\ \overline{X_2} - \overline{X_1 X_2}}{\overline{X_2}}$$

$$\hat{\sigma}_{13} := \frac{\overline{X_1 X_3} - \overline{X_1}\ \overline{X_3}}{\overline{X_3}}$$

$$\hat{\sigma}_{14} := (\overline{X_1 X_4} - \overline{X_1}\ \overline{X_4})\frac{\sqrt{1 + \left\{\Phi^{-1}(\overline{X_4})\right\}^2}}{\varphi\left(\Phi^{-1}(\overline{X_4})\right)}$$

$$\hat{\sigma}_{23} := \log \overline{X_2} + \log \overline{X_3} - \log \overline{X_2 X_3}$$

$$\hat{\sigma}_{24} := \sqrt{1 + \left\{\Phi^{-1}(\overline{X_4})\right\}^2}\left\{\Phi^{-1}(\overline{X_4})) - \Phi^{-1}\left(\frac{\overline{X_2 X_4}}{\overline{X_2}}\right)\right\}$$

$$\hat{\sigma}_{34} := \sqrt{1 + \left\{\Phi^{-1}(\overline{X_4})\right\}^2}\left\{\Phi^{-1}\left(\frac{\overline{X_3 X_4}}{\overline{X_3}}\right) - \Phi^{-1}(\overline{X_4}))\right\}$$

Estimating the covariance matrix

To estimate the missing Bernoulli cross-moments, a natural estimator is $\hat{\sigma}_{44}$ defined implicitly as

$$\Phi_2\left(\Phi^{-1}(\overline{X_4}), \Phi^{-1}(\overline{X_{4'}}); \frac{\hat{\sigma}_{44'}}{\sqrt{1+\{\Phi^{-1}(\overline{X_4})\}^2}\sqrt{1+\{\Phi^{-1}(\overline{X_{4'}})\}^2}}\right) - \overline{X_4 X_{4'}} = 0,$$

where $\Phi_2(y_1, y_2; \rho)$ is the cumulative distribution function of the bivariate normal distribution with zero means, unit variances and correlation equal to ρ.

MTPCA and sparsity

- When we know the covariance matrix of the latent combinations W of the principal components Z, we can do regular PCA to find the components.
- Sparsity can be added by using Sparse PCA (Zou, Hastie and Tibshirani, 2006).
- For this, we use the SICS algorithm (Heinonen and Virta, 2026).

Estimating the variance parameter

- For Gaussian observed variables we assume a probability model $X_{1j}|Z \sim \mathrm{N}(W_{1j}, \tau_j^2)$.
- Usually we assume τ_j to be the same for all variables j (after standardization).
- Assume that the real number of principal components d is smaller than the number of Gaussian components p_1. Then the variances of (regular) principal components, calculated using only Gaussian variables, are, in decreasing order, $(\lambda_1 + \tau, \ldots, \lambda_d + \tau, \tau, \ldots, \tau)$, so the estimate for τ is the average of the last $p_1 - d$ elements.

Estimating the PC scores

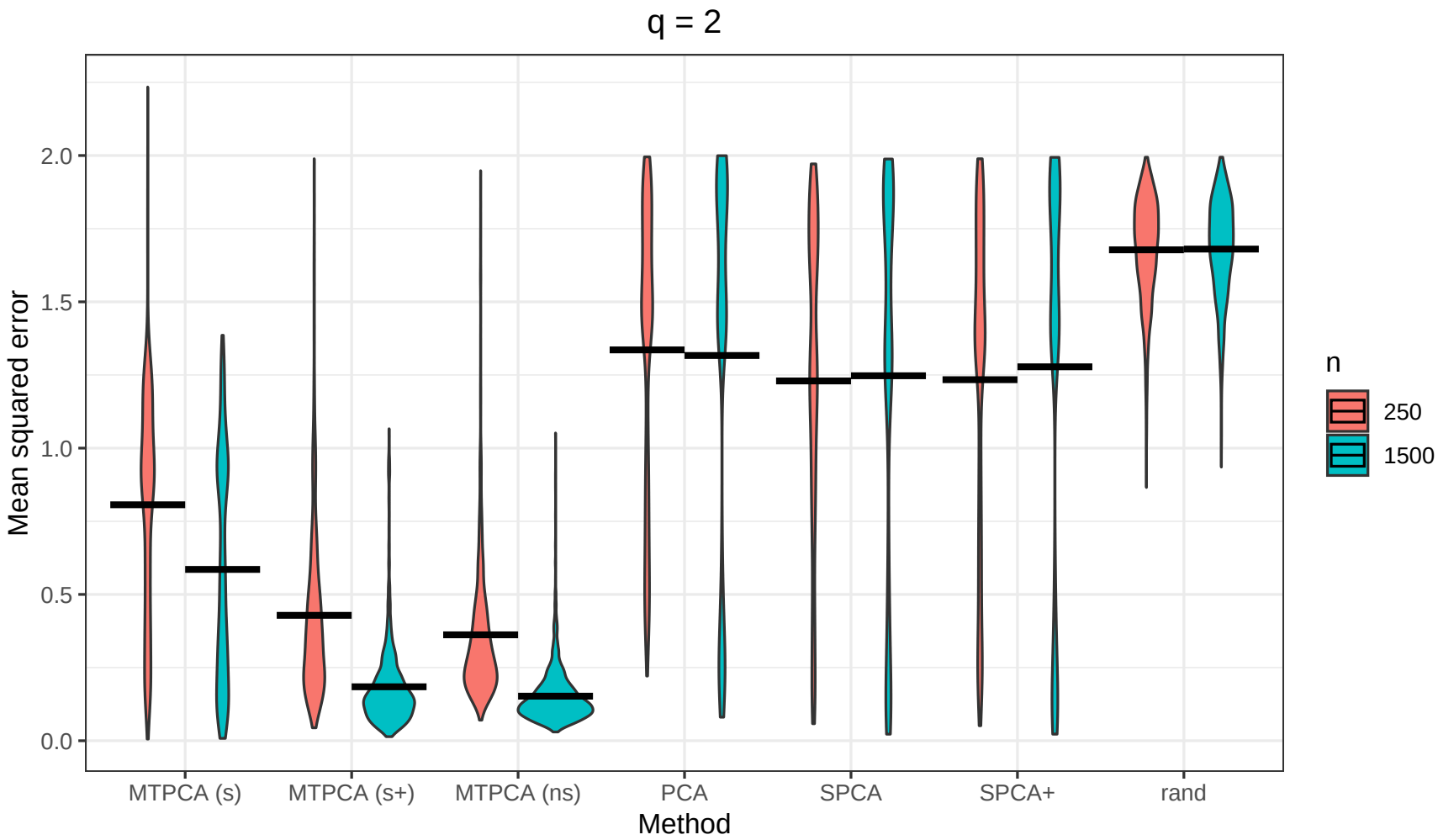
- In PCA we are interested in two things, the loadings $\hat{U}$ and the principal component scores $\hat{Z}$.
- First we estimate the loadings $\hat{U}$, component variances $\hat{\Lambda}$ and the mean vector $\hat{\mu}$ using the method of moments.
- Then we can estimate the scores $\hat{Z}$ by maximizing the conditional log-density $\log f_{Z|X}(Z \mid X, \hat{U}, \hat{\Lambda}, \hat{\mu})$, where X stands for the observations.
- In case of regular PCA model, this approach of finding the mode of the conditional log-density leads to the regular principal component scores.
- The function $\log f_{Z|X}$ is strictly concave[1] and therefore the solution is unique.

[1] If the matrix Λ is not singular

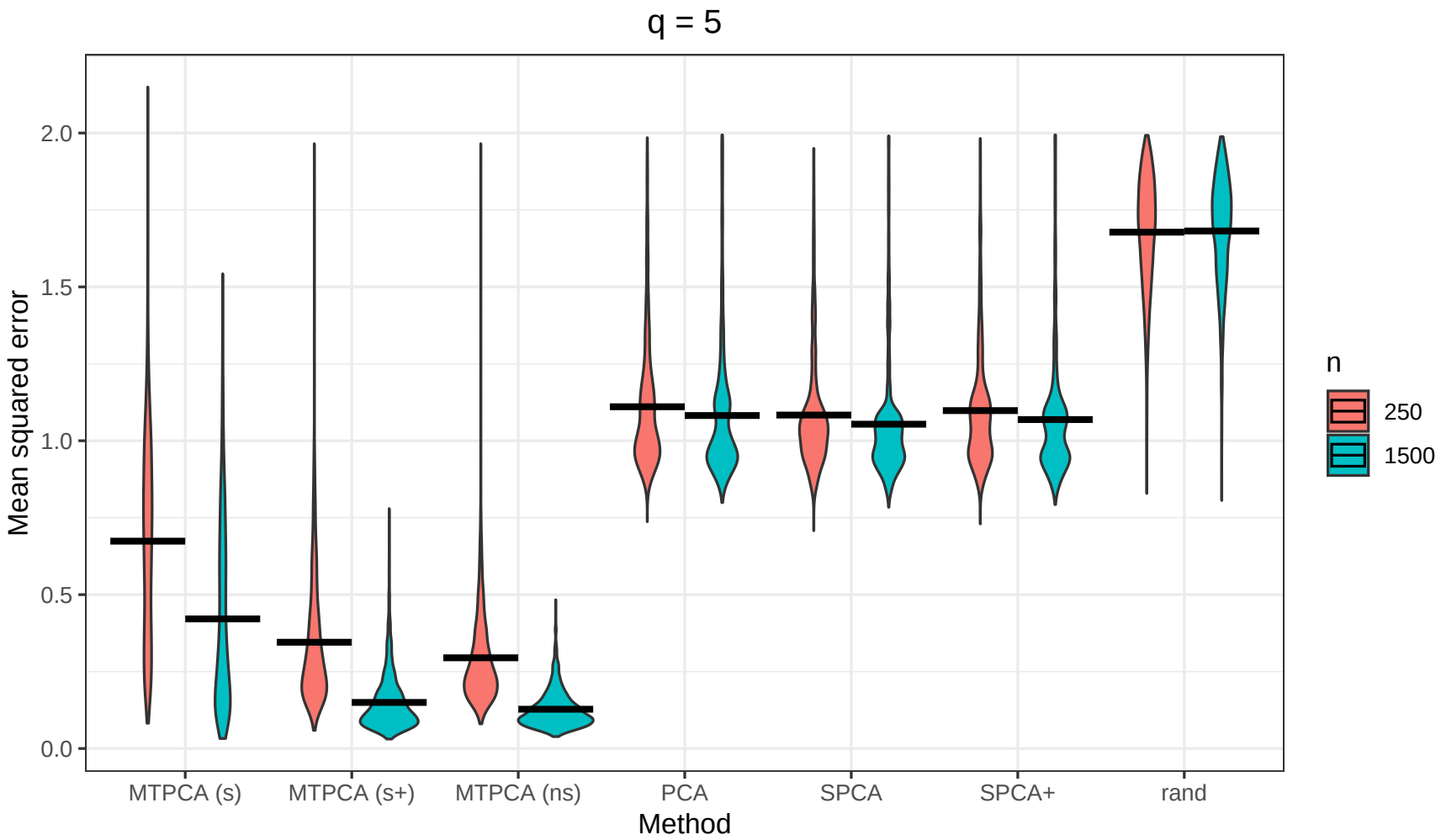
Simulation

- We had a simulation where we simulated data from our probability model and compared MTPCA and regular PCA.
- We used two sample sizes $n = 250, 1500$, two real amounts of non-zero coefficients $q = 2, 5$ (when number of variables is $p = 8$) and three levels of sparsity $r = q$, $r = q + 2$, $r = p$.
- We measure the Frobenius norm between first two components of the real and estimated loading matrix.

Simulation


q = 2
Mean squared error
0.0
0.5
1.0
1.5
2.0
MTPCA (s)
MTPCA (s+)
MTPCA (ns)
PCA
SPCA
SPCA+
rand
Method
n
250
1500

Simulation


q = 5
Mean squared error
2.0
1.5
1.0
0.5
0.0
MTPCA (s)
MTPCA (s+)
MTPCA (ns)
PCA
SPCA
SPCA+
rand
Method
n
250
1500

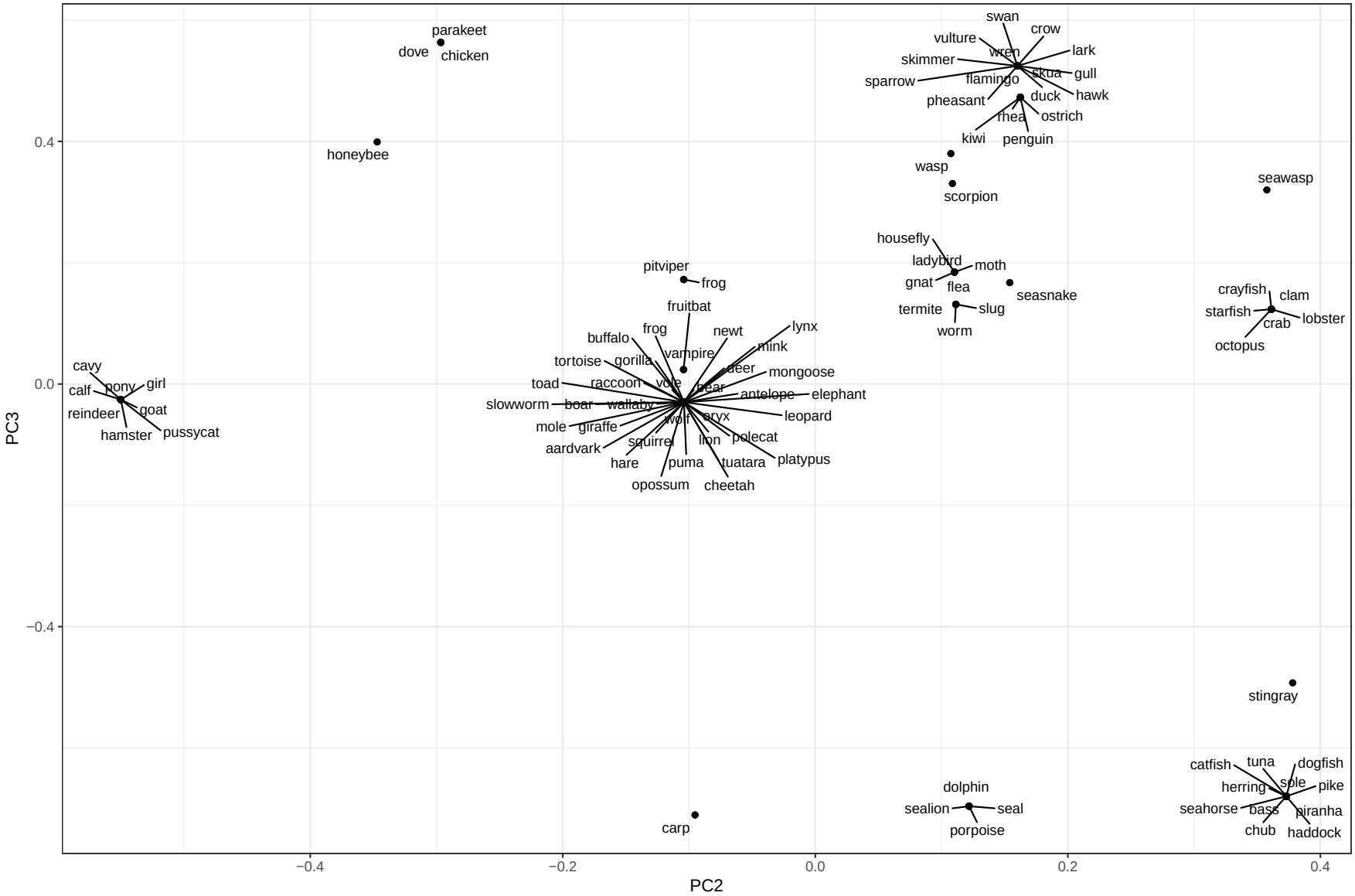
Zoo example

- We use MTPCA on the Zoo dataset (Forsyth, 1990)
- The data cosists of $n = 101$ animals and $p = 15$ binary variables that describe the presence or absence of certain feature (is predator, is venomous, etc).

Zoo example - loadings

variable	PC1	PC2	PC3
hair	0	0	0
feathers	0.14	0.45	0.58
eggs	0	0	0
milk	0	0	0
airborne	−0.099	0	0.059
aquatic	0	0	0
predator	0	0	0
toothed	0	0	0
backbone	0	−0.32	−0.17
breathes	0.037	−0.40	0
venomous	−0.98	0	0.17
fins	−0.078	0.36	−0.78
tail	0	0	0
domestic	0	−0.63	0
catsize	0	0	0

Zoo example - scores



Advantages of MTPCA

- MTPCA is so similar to regular PCA that all normal techniques and concepts work nicely (scree plots, loadings and scores...)
  - of these, we demonstrate sparsity.
- We have a proper probability model from latent components to observations.
- We have explicit formulas for almost everything.

References

- Michael Collins, Sanjoy Dasgupta, and Robert E Schapire. 2001. *A generalization of principal components analysis to the expo- nential family*. In Advances in Neural Information Processing Systems, Vol 14
- Wei Liu, Huazhen Lin, Shurong Zheng and Jin Liu. *Gen- eralized Factor Model for Ultra-High Dimensional Correlated Variables with Mixed Types*. Journal of the American Statistical Association 118, no. 542 (2023): 1385–1401.
- Michael E Tipping and Christopher M Bishop. *Probabilistic principal component analysis*. Journal of the Royal Statistical Society Series B: Statistical Methodology 61, no. 3 (1999): 611-622
- Hui Zou, Trevor Hastie and Robert Tibshirani. *Sparse principal component analysis*. Journal of Computational and Graphical Statistics 15, no. 2 (2006): 265-286
- Lauri Heinonen and Joni Virta. *A method for sparse and robust independent component analysis*. Journal of Multivariate Analysis 213 (2026): 105587
- Richard Forsyth 1990. *Zoo*. UCI Machine Learning Repository, DOI: https://doi.org/10.24432/C5R59V.