Probabilistic PCA as a tool in hierarchical Bayesian modelling

Pierre Gloaguen, Eric Parent

Institut Henri Poincaré, 23rd of September, 2022

Probabilistic Principal component analysis

Gaussian model for a $n \times p$ data-set

- Data set $\mathbf{Y}$ having n rows and p columns;
- Rows of $\mathbf{Y}$ are supposed to be $i.i.d.$ samples;
- The focus is made in the dependance between columns:
  - Matrix distribution point of view:

$$Y \sim \mathcal{MN}(\mathbf{0}, I_n, \Omega).$$

  - Rowwise Gaussian vector point of view:

$$Y_k \overset{ind}{\sim} \mathcal{N}_p(0, \Omega), \ 1 \leq k \leq n$$

- We want to give a low rank structure to Ω;

Probabilistic PCA

- Latent variable point of view:
- For each $1 \leq k \leq n$, there exists a latent variable $Z_k \sim \mathcal{N}_q(0, I_q)$, and a matrix of loadings $\Lambda \in \mathcal{M}_{p \times q}$ such that:

$$Y_k = \Lambda Z_k + E_k \;\; E_k \sim \mathcal{N}_p \left(\mathbf{0}, \Sigma := \begin{pmatrix} \sigma_1^2 & 0 & \cdots & 0 \\ 0 & \sigma_2^2 & \cdots & 0 \\ \vdots & \cdots & \ddots & \vdots \\ 0 & \cdots & \cdots & \sigma_p^2 \end{pmatrix} \right).$$

- This results in $Y_k \sim \mathcal{N}_p \left(\mathbf{0}, \Omega = \Lambda\Lambda^T + \Sigma \right)$, thus, Ω is structured as a matrix of rank $q < p$ plus a diagonal matrix.
- Link with PCA:

$$\mathbf{Y} = \mathbf{Z}\Lambda^T + \mathbf{E}.$$

Bayesian probabilistic PCA

- Choosing the priors to:
  - Have a nice conjugate scheme;
  - Penalize high rank matrices for Λ;
  - Focus on the *multiplicative gamma process shrinkage prior* of Bhattacharya and Dunson (2011)

Priors of Bhattacharya and Dunson (2011)

- **Variance priors** on $\Sigma = \text{diag}(\sigma_j^2), 1 \leq j \leq p$ Inverse Gamma: Standard
- **Latent variables priors** on $Z_k, 1 \leq k \leq p$: $\mathcal{N}_p(0, I_p)$ Standard
- **Loading priors** on Λ: *multiplicative gamma process shrinkage prior*.

Priors of Bhattacharya and Dunson (2011)

- **Variance priors** on $\Sigma = \text{diag}(\sigma_j^2), 1 \leq j \leq p$ Inverse Gamma: Standard
- **Latent variables priors** on $Z_k, 1 \leq k \leq p$: $\mathcal{N}_p(0, I_p)$ Standard
- **Loading priors** on Λ: *multiplicative gamma process shrinkage prior*.

Shrinkage prior

- Λ is a $p \times q$ matrix;
- Idea: Penalize matrices whose last columns have too big values;

Priors of Bhattacharya and Dunson (2011)

- **Variance priors** on $\Sigma = \text{diag}(\sigma_j^2), 1 \leq j \leq p$ Inverse Gamma: Standard
- **Latent variables priors** on $Z_k, 1 \leq k \leq p$: $\mathcal{N}_p(0, I_p)$ Standard
- **Loading priors** on Λ: *multiplicative gamma process shrinkage prior*.

Shrinkage prior

- Λ is a $p \times q$ matrix;
- Idea: Penalize matrices whose last columns have too big values;
- Let, for $1 \leq j \leq p$ and $1 \leq h \leq q$, $\phi_{j,h} \overset{\text{ind}}{\sim} \mathcal{G}\text{amma}\left(\frac{3}{2}, \frac{3}{2}\right)$.
- Let, for $1 \leq h \leq q$, $\delta_h \overset{\text{ind}}{\sim} \mathcal{G}\text{amma}(\alpha, 1)$ such that $\alpha > 1$ (thus $\mathbb{E}[\delta_h] > 1$);
- Then set as prior:

$$\Lambda_{j,h} | \phi_{j,h}, \delta_{1:h} \overset{\text{ind}}{\sim} \mathcal{N}\left(0, \phi_{j,h}^{-1} \prod_{\ell=1}^{h} \delta_\ell^{-1}\right).$$

- When h increases (for last columns), the variance tends to collapse to 0.
- Remains the prior over α: Non informative, greater than 1.

Posterior sampling

Gibbs sampling

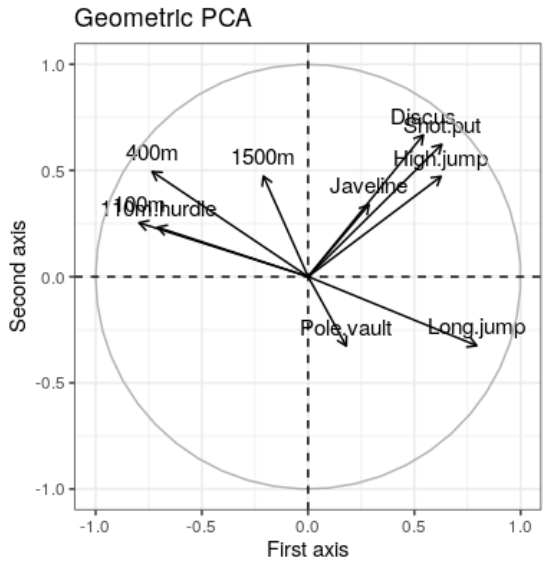
- The joint posterior distribution has no closed form;
- Samples can be obtained via Gibbs sampling, as all conditional distributions are Normal-Gamma conjugations;
- All, except the one for α! $\Rightarrow$ Metropolis-Hastings;

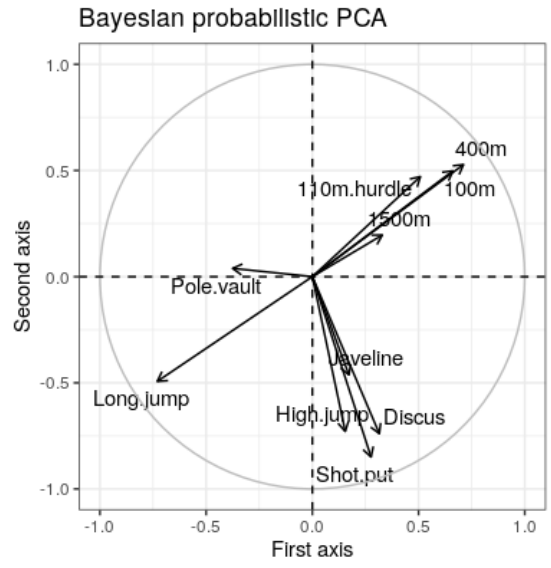
Variational inference

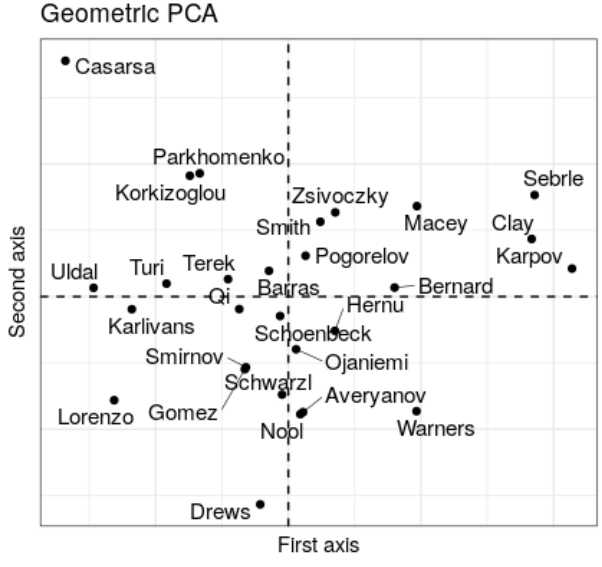
- One can approximate the posterior distribution by a product of marginal distributions (mean field family);
- The inference problem boils down to find the best distributions among this family;
- The optimization problem can be solve iteratively with an explicit gradient ascent algorithm (Coordinate ascent variational inference);

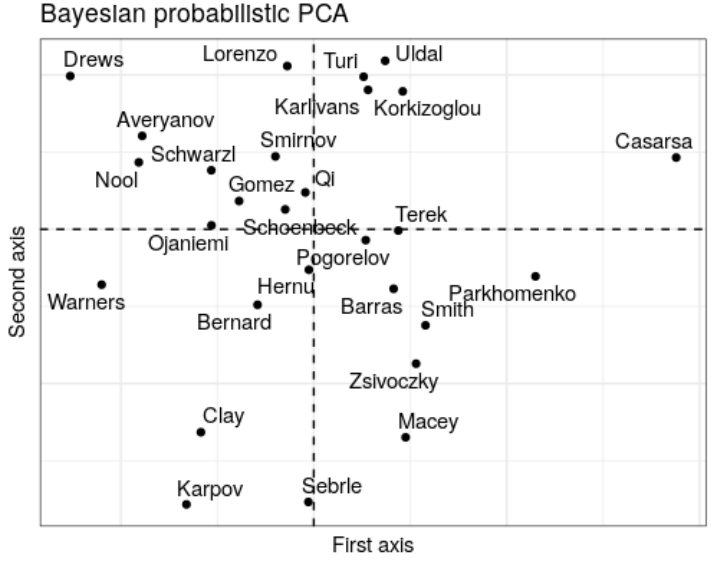
Does it give the same results as standard PCA?

Data set, performance of $n = 28$ athletes at Olympic games in decathlon ($p = 10$ variables).









Inclusion in a Bayesian hierarchical model

Context: Joint species distribution modelling

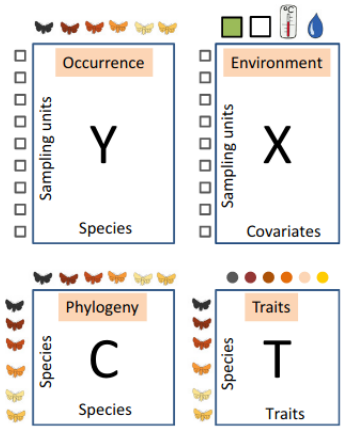


Figure 1: From Ovaskainen et al, 2020

- **Y** Matrix of **species counts**, over **sampling units** having:
  - n rows (*sampling units*)
  - p columns (*species*);
- **X** Matrix of **sampling units covariates**, having:
  - n rows (*sampling units*)
  - k columns (*covariates*);
- **C** Matrix of **species phylogeny**, having:
  - p rows (*species*)
  - p columns (*species*);
- **T** Matrix of species **species traits**, having:
  - p rows (*species*)
  - n_t columns (*traits*);

Question: Can we establish a statistical link between **Y** and **X**, **C**, **T**?

A classical statistical approach

- **Y** is a matrix of **counts** $\Rightarrow$ **Poisson distribution**;

$$\mathbf{Y} \sim \mathcal{P}\text{oisson}(\exp(\mathbf{L}))$$

  where **L** is a matrix having the same dimensions as **Y** (the exponential is taken entrywise).
- **L** will be a **linear predictor**;
- **X** and **T** are seen as **features** (explanatory variables);
- **C** is seen as a **correlation matrix**;

Model on the linear predictor (Ovaskainen and Abrego (2020))

- **L** is a matrix $n \times p$ (# of sites $\times$ # of species), modelling the **intensity** of presence of species per unit;
- We suppose it is random, with **Normal** distribution;
- A Matrix Normal random variable is characterized by:
  - Its expected value (mean intensity) **M**;
  - Its covariance between rows (sites) Σ_{sites} (matrix $n \times n$);
  - Its covariance between columns (species) $\Sigma_{species}$ (matrix $p \times p$);

$$\mathbf{L} \sim \mathcal{MN}\left(\mathbf{M}, \Sigma_{sites}, \Sigma_{species}\right)$$

Model on the linear predictor (Ovaskainen and Abrego (2020))

- **L** is a matrix $n \times p$ (# of sites $\times$ # of species), modelling the **intensity** of presence of species per unit;
- We suppose it is random, with **Normal** distribution;
- A Matrix Normal random variable is characterized by:
  - Its expected value (mean intensity) **M**;
  - Its covariance between rows (sites) Σ_{sites} (matrix $n \times n$);
  - Its covariance between columns (species) $\Sigma_{species}$ (matrix $p \times p$);

$$\mathbf{L} \sim \mathcal{MN}(\mathbf{M}, \Sigma_{sites}, \Sigma_{species})$$

Model on **M**, the expected intensity

- The expected intensity is linked to environment covariates **X**:

$$\mathbf{M} = \mathbf{X}\beta$$

  Where β is a, **unknown** $k \times p$ (# of covariates $\times$ # of species) matrix giving the unknown **response of species to environnement**.
- Each species is then characterized by a vector of response to environnement: **its niche**.

Model over the niches

- The matrix β stacks the niches of species (vector of responses to environnement);
- We assume that:
  - The **traits** might affect this response to environment (i.e. similar traits lead to similar niche);
  - The response to environment might be correlated between species, because of **phylogeny**.

Model over the niches

- The matrix β stacks the niches of species (vector of responses to environnement);
- We assume that:
  - The **traits** might affect this response to environment (i.e. similar traits lead to similar niche);
  - The response to environment might be correlated between species, because of **phylogeny**.

Formally, β is assumed to be a Matrix Normal random variable such that:

$$\beta \sim \mathcal{MN}\left(\mathbf{\Gamma T}', \mathbf{V}, \rho\mathbf{C} + (1-\rho)\mathbf{I}_{n_s}\right)$$

- $\mathbf{\Gamma}$ is a $k \times n_t$ ($\#$ of covariates $\times$ $\#$ of traits) describing how the response to environment is structured by the traits; **Do the species niches are correlated to species traits?**
- $\mathbf{V}$ models the covariance between rows of β, i.e. between response to different covariates;
- $0 \leq \rho \leq 1$ is the importance weight of **phylogeny** in the columns correlation of β.

Random effects on the presence intensity

- So far:

$$
\begin{array}{lcll}
\mathbf{Y} & \sim & \mathcal{P}\text{oisson}(\exp(\mathbf{L})) & \text{Abundance distribution} \\
\mathbf{L} & \sim & \mathcal{MN}\left(\mathbf{X}\beta, \Sigma_{sites}, \Sigma_{species}\right) & \text{Model for the presence intensities} \\
\beta & \sim & \mathcal{MN}\left(\mathbf{\Gamma T}', \mathbf{V}, \rho\mathbf{C} + (1-\rho)\mathbf{I}_{n_s}\right) & \text{Model for the niches}
\end{array}
$$

- What about Σ_{sites} (the covariance between intensities in sampling units)?
  - Spatial block structure;
- What about $\Sigma_{species}$ (the covariance between intensities of species)?
  - If environment explained all, residual species intensities would be independant!
  - However, some species **coocurence** might remain!
  - We will write (as in Chiquet, Mariadassou, and Robin (2018)):

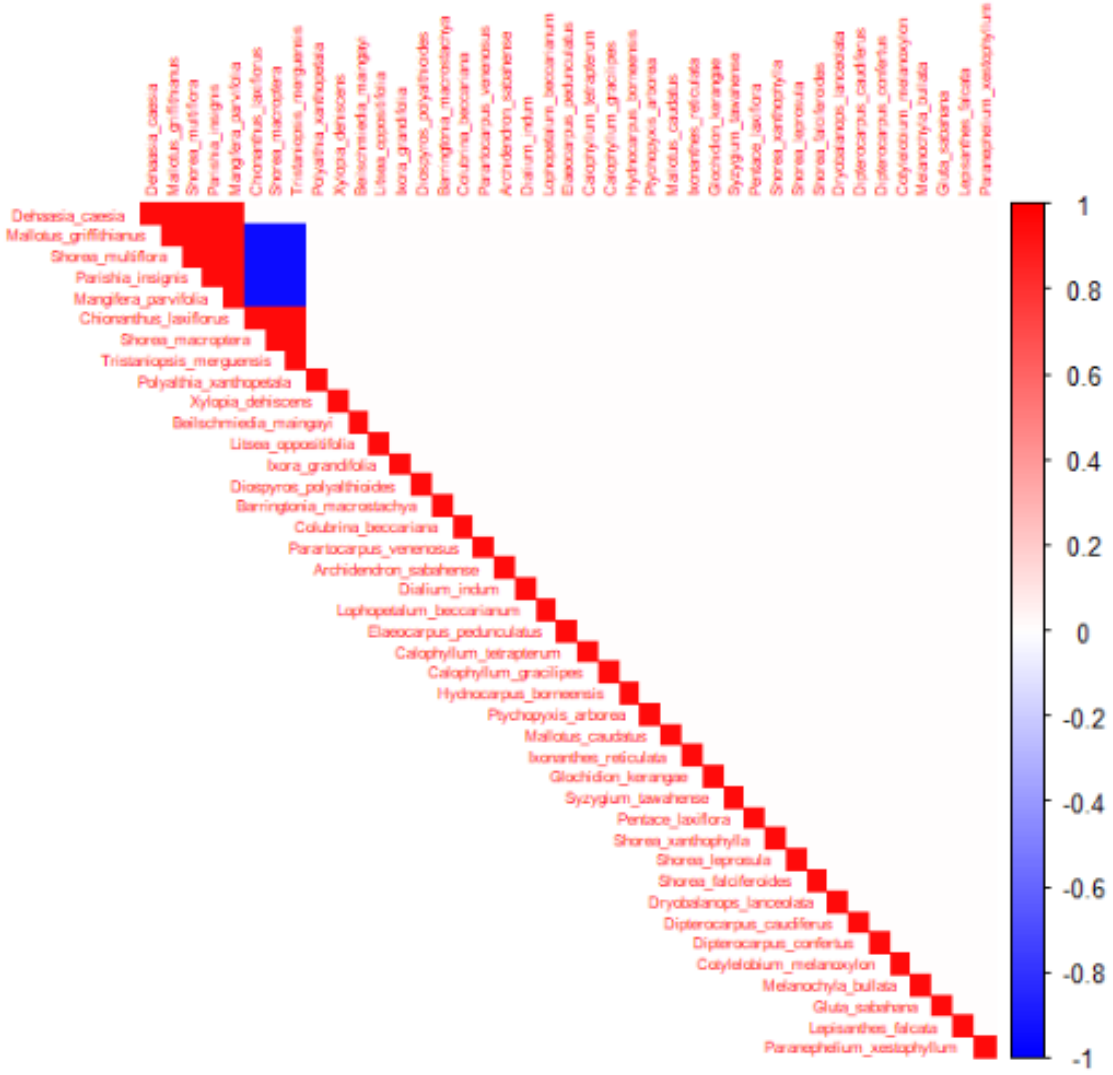
$$\Sigma_{species} = \text{diag}\boldsymbol{\sigma}^2 + \overbrace{\mathbf{\Omega}}^{\text{Coocurence matrix}},$$

where $\mathbf{\Omega} = \Lambda\Lambda^T$

First results

- Modelling on 41 (relatively) abundant species spanning a wide range of traits;
- Focus on 180 sites where soil chemistry was measured;
- Choosing as environment variables:
  - the soil type (**qualitative** with 3 levels, Alluvial, Heath, Sandstone);
  - Available phosphorus;
  - Available exchangeable cations;

Residual co-occurrence in predicted presence intensities?



- 2 blocks of species having intra co-occurrence and inter coavoidance;

Conclusions and perspectives

Conclusions

- PCA as probabilistic model with latent variable is a powerful tool for covariance modelling;
- Existing priors are well suited for efficient conjugate inference;
- In joint species distribution modelling, PPCA can help to identify residual co-occurence;

Perspectives

- Turn the full JSDM model into a variational framework;
- So far, there is a latent variable per site species (900×500);
- Next goal: convert Eric to amortized inference with machine learning methods (in the spirit of VAEs)!

References

Bhattacharya, Anirban, and David B Dunson. 2011. "Sparse Bayesian Infinite Factor Models." *Biometrika*, 291–306.

Chiquet, Julien, Mahendra Mariadassou, and Stéphane Robin. 2018. "Variational Inference for Probabilistic Poisson PCA." *The Annals of Applied Statistics* 12 (4): 2674–98.

Ovaskainen, Otso, and Nerea Abrego. 2020. *Joint Species Distribution Modelling: With Applications in R*. Cambridge University Press.