Advances in structured Bayesian factor models

Antonio Canale · antonio.canale@unipd.it

April 26, 2024 – *WU Vienna*

UNIVERSITÀ
DEGLI STUDI
DI PADOVA

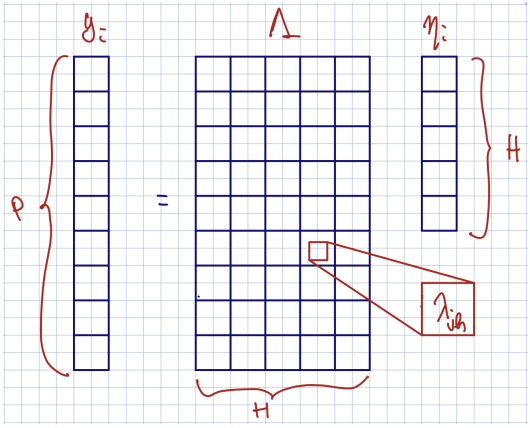
Outline of the presentation

1. Introduction
2. Structured Increasing Shrinkage priors
3. Flexible Multi Study Bayesian Factor Analysis
4. Illustrations: Finnish bird co-occurrence data
   - Single campaign
   - Repeated campaigns

Factor Models (FM)

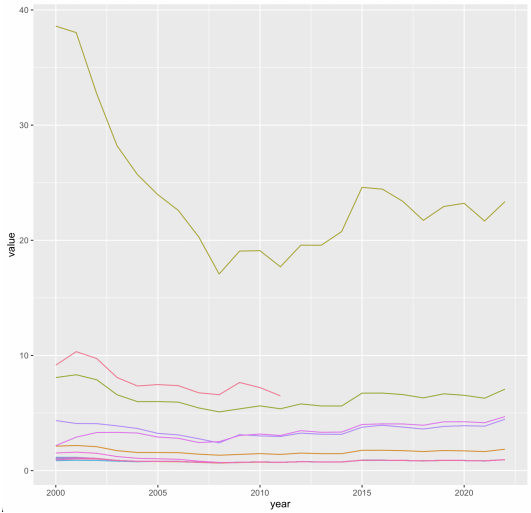
$$y_i = \Lambda \eta_i + \epsilon_i \qquad \epsilon_i \sim f_\epsilon, \qquad \eta_i \sim f_\eta,$$

- y_i: i-th p-variate random variable;
- Λ: $p \times H$ factor loadings matrix;
- η_i: i-th vector of H latent factors.



Example 1: economics and finance

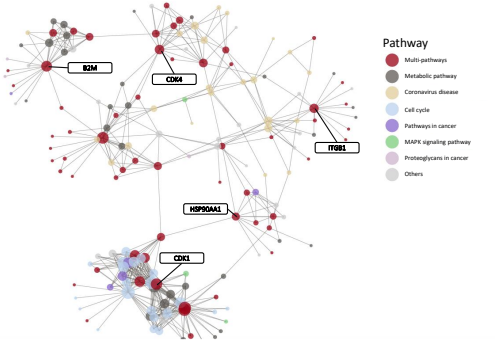
- (log) returns of p different assets/stocks or exchange rates may are observed through time
- many of these observations follows some latent "market trend"
- there may be blocks of assets in terms of market sectors/geographical areas



Exchange rates data. Source: data.oecd.org

Example 2: sc-RNA-Seq

- Single-cell gene expression data are often characterized by large matrices, where the number of cells $\ll$ than the number of genes;
- Matrix factorizations can reveal low-dimensional structures;
- these techniques can uncover new biological knowledge from diverse high-throughput omics data;



Source: Canale et al. (2023), Structured factorization for single-cell gene expression data, arXiv

Example 3: multi-study factor analysis

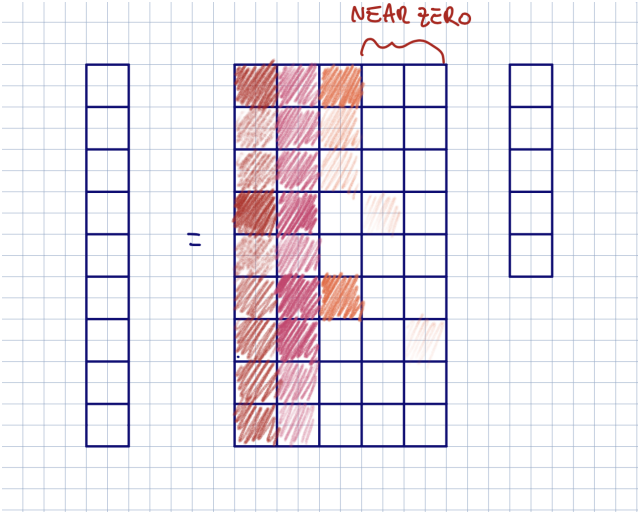
- analysis of critical disease are often carried out using systematic collections of data generated over time in different laboratories and/or hospitals
- multi-study factor analysis (De Vito et al., 2018) postulates the existence of common factors shared across multiple studies, and study-specific factors

Source: artificial image generated from Stable Diffusion

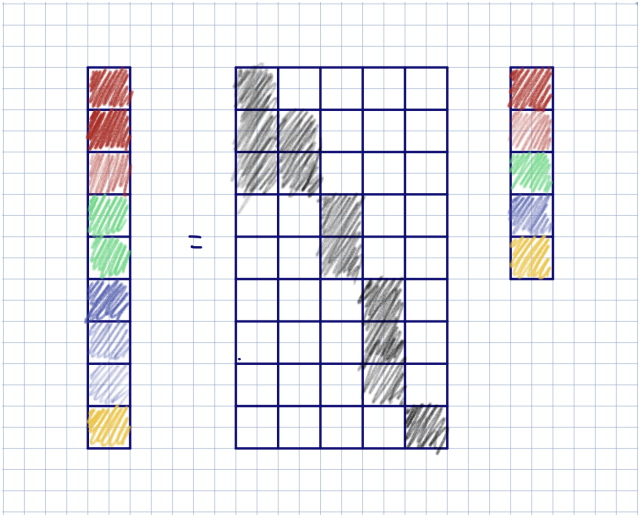
Interpretability and sparsity

- Most of the interest revolves around the concept of interpretability;
- Interpretation of factor models is assigning a meaning to the latent factors and then to their impact on the observed data;
- this is promoted by the concept of sparsity in many ways…

Sparsity for dimensionality reduction

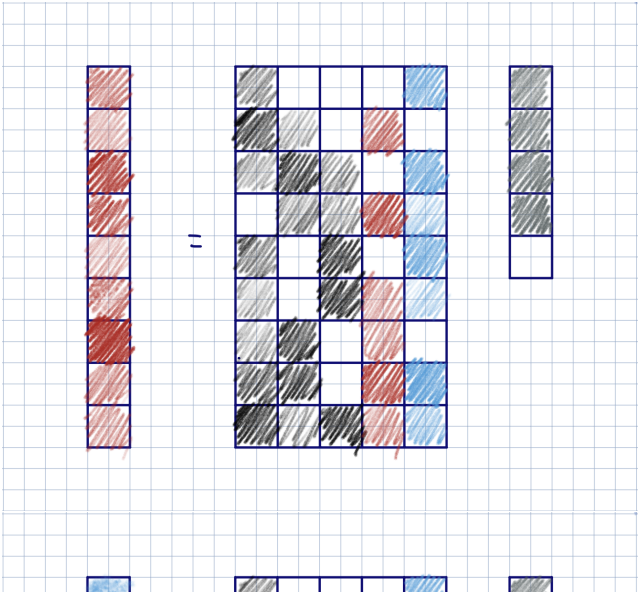


Sparsity for block structure



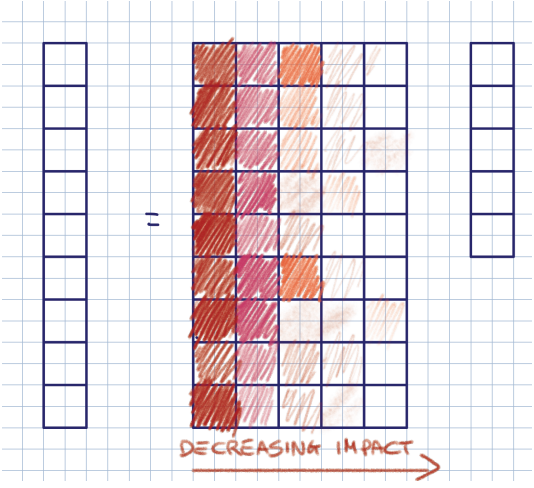
Interpretation of loadings matrix is strongly favored when each factor has an impact only on a small group of components of observed variables

Sparsity for multi-study



Bayesian (infinite) factor models

- Infinite factor models are Bayesian Nonparametric models for dimensionality reduction
- Batthacharia & Dunson (2011) first introduced this idea with the multiplicative gamma process (MGP)
- Key idea: there are infinitely many factors, with the impact of these factors decreasing with the factor index
- Similar in spirit is the cumulative shrinkage process (CUSP) by Legramanti et al. (2020) recently generalized by Frühwirth-Schnatter (2024)



Bayesian (infinite) factor models

- In the MGP

$$\lambda_{jh} \sim N(0, \phi_{jh}\tau_h), \quad \phi_{jh} \sim \text{Ga}(\nu/2, \nu/2), \quad \tau_h = \prod_{l=1}^{h} \delta_l, \quad \delta_l \sim \text{Gamma}$$

- In the (generalized) CUSP

$$\lambda_{jh} \sim N(0, \theta_{jh}), \quad \theta_{jh} \sim \text{spike-and-slab}(\pi_h), \quad \pi_h = \sum_{l=1}^{h} \omega_l, \quad \omega_h \sim \text{stick-breaking}$$

Bayesian (sparse) factor models

- Local sparsity is another key concept in Bayesia FM
- Priors for each λ_{jh} can be broadly divided into
  - Continuous shrinkage prior: Zhao et al. (2016), Rockova & George (2017), Kastner (2019)
  - Discrete shrinkage priors (spike & slab): West (2003), Carvalho et al. (2008), Conti et al. (2014), Kaufmann & Schumacher (2019) and Frühwirth-Schnatter et al. (2024)

Structured sparsity

- In this talk we will present some recent contributions that deal with sparsity in an informed way;
- Postulating the existence of covariates and metacovariates we define suitable continuous shrinkage priors;
- Exploiting the information contained in such variables, we define regression models on the prior scale parameters thus promoting sparsity in a structured way

Outline

1. Introduction
2. Structured Increasing Shrinkage priors
3. Flexible Multi Study Bayesian Factor Analysis
4. Illustrations: Finnish bird co-occurrence data
   - Single campaign
   - Repeated campaigns

Outline

Generalized Infinite Factorizations

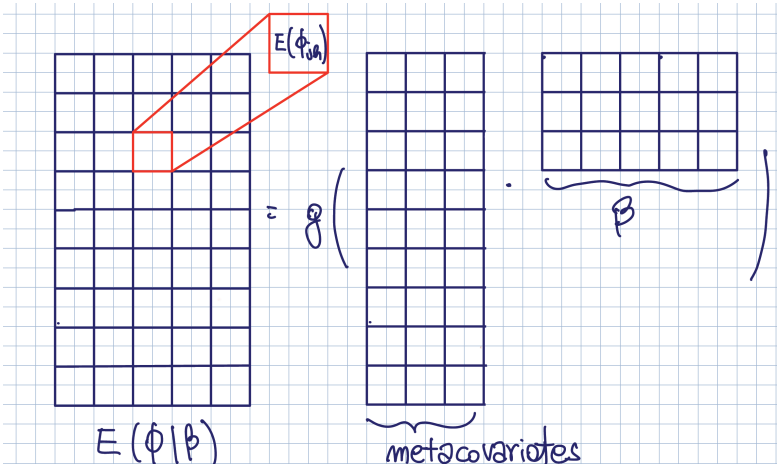
$$\lambda_{jh} \mid \theta_{jh} \sim N(0, \theta_{jh})$$

$$\theta_{jh} = \tau_0 \, \gamma_h \, \phi_{jh}$$

- $\tau_0 \sim f_{\tau_0}$: global scale;
- $\gamma_h \sim f_{\gamma_h}$: column scale;
- $\phi_{jh} \sim f_{\phi_j}$: local scale. That depends on meta covariates: $E(\phi_{jh}) = g(x_j^\top \beta_h)$

Schiavon, Canale, Dunson (2022)

Exogenous information about the sparsity structure



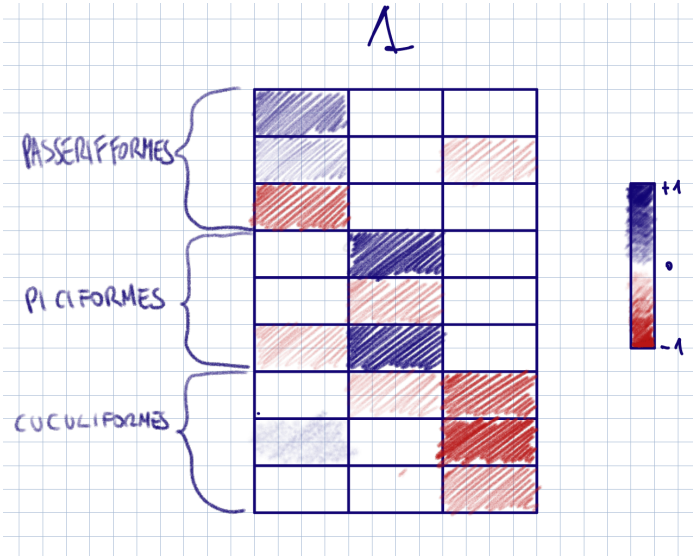
$$E(\phi_{jh} \mid \beta_h) = g(x_j^\top \beta_h), \qquad \beta_h = (\beta_{1h}, \ldots, \beta_{qh})^\top, \qquad \beta_{mh} \sim f_\beta$$

Bird species occurrence example (1)

- **y**: occurrence of **p species** in **n** different **environments**;
- η: **H** latent **factors**;
- **Λ**: **impact of the latent factors** on the species occurrence;
- **x**: q **species characteristics** (taxonomy, size, migratory strategy…), providing similarities between different species.

Considering *x* indicating the phylogenetic order of each species.
If the *h*-th factor does not impact the occurrence of the species *j* ($\lambda_{jh} = 0$), it could not even impact the other species *s* belonging to the same order of *j* ($\lambda_{sh} = 0$).

Bird species occurrence example (2)



Theoretical prior properties

We define desirable properties for the GIF class including

- Increasing shrinkage ($\text{var}(\lambda_{jh}) < \text{var}(\lambda_{j(h-1)})$ for any h)
- Robustness to large signals (not overshrinking)
- Asymptotic increasing sparsity (for $p \to \infty$ the sparsity rate increases)

We provide conditions for the properties to hold.

Structured Increasing Shrinkage prior

$$\lambda_{jh} \mid \theta_{jh} \sim \mathrm{N}(0, \theta_{jh}) \qquad \theta_{jh} = \tau_0 \, \gamma_h \, \phi_{jh}$$

General GIF equations

$$\tau_0 = 1, \quad \gamma_h = \vartheta_h \rho_h, \quad \vartheta_h^{-1} \sim \mathrm{Ga}(a_\theta, b_\theta),$$

Power law tail column scale

$$\rho_h = \mathrm{Ber}\left(1 - \pi_h\right), \quad \pi_h \sim \mathrm{CUSP}(\alpha)$$

Incr. shrinkage

$$\phi_{jh} \mid \beta_h \sim \mathrm{Ber}\{\mathrm{logit}(X_j^\top \beta_h)\} \quad \beta_h \sim \mathrm{N}_q(0, \sigma_\beta^2 I_q),$$

Meta covariates impacting the sparsity pattern

Simulation scenarios

- We compare the performance of our proposal with current approaches (Bhattacharya & Dunson, 2011; Legramanti et al., 2020)
- Scenarios:
  1. increasing shrinkage FM (no local sparsity;
  2. locally sparse FM (no increasing shrinkage);
  3. 1 + 2;
  4. 1 + 2 + metacovariate-dependence in sparsity
- Performance measures: LPML, posterior mean of k (estimated number of columns of Λ), MSE of Ω

Results (1)

	(p, k)	MGP		CUSP		SIS	
		$Q_{0.5}$	IQR	$Q_{0.5}$	IQR	$Q_{0.5}$	IQR
LPML	(16,4)	-28.68	0.42	-28.68	0.43	-28.65	0.41
	(32,8)	-60.08	0.45	-60.09	0.45	-60.07	0.49
	(64,12)	-117.68	0.56	-117.75	0.53	-117.88	0.56
	(128,16)	-225.04	1.04	-225.13	1.04	-228.76	1.47
$E(H_a \mid y)$	(16,4)	8.17	1.44	4.00	0.00	4.00	0.00
	(32,8)	10.68	0.33	8.00	0.00	8.00	0.00
	(64,12)	14.16	1.09	12.00	0.00	12.00	0.00
	(128,16)	17.03	0.47	16.00	0.00	18.00	0.02

Figure: LPML and estimated latent dimension (k) in Scenario 1 —worst case for the proposed method

Results (2)

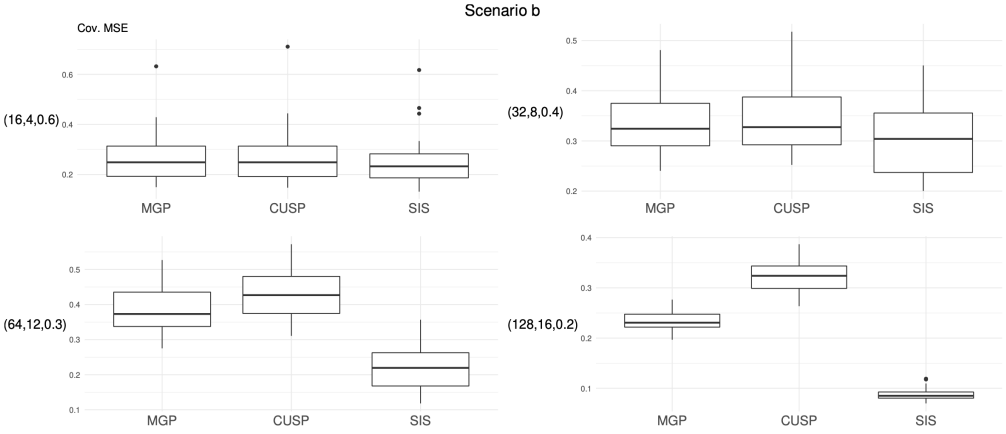


Figure: MSE for Ω for different combination of (p, k, s)

Outline

Multi study factor analysis

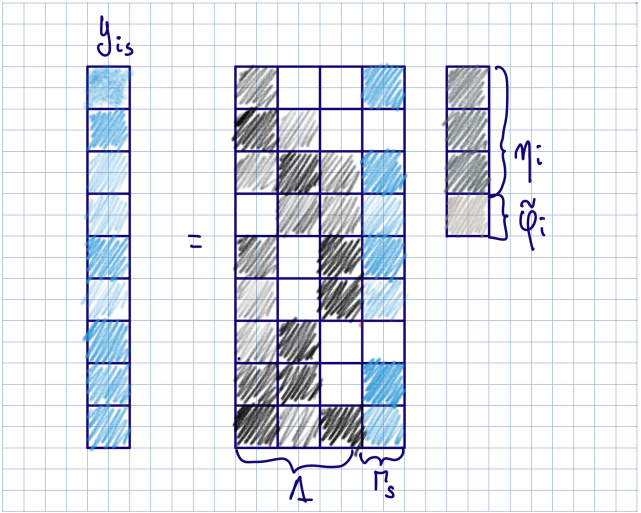
- Multi-study factor analysis (MSFA) assumes the existence of both shared latent factors and study-specific latent factors
- This approach has been introduced by De Vito et al. (2019) and later extended by De Vito et al. (2021) Grabski et al. (2023), and De Vito & Avalos-Pacheco (2023)
- Specifically

$$y_{is} = \Lambda \eta_{is} + \Gamma_s \tilde{\varphi}_{is} + \epsilon_{is}. \tag{1}$$

  where Γ_s is a (study-specific) factor loading matrix of dimension $p \times k_s$, with $k_s \ll p$ possibly different in each study, and $\tilde{\varphi}_{is}$ its corresponding latent factor.
- The resulting marginal distribution of y_{is} is Gaussian with covariance

$$\Omega_s = \Lambda\Lambda^\top + \Gamma_s\Gamma_s^\top + \Sigma_s.$$

MSFA graphically



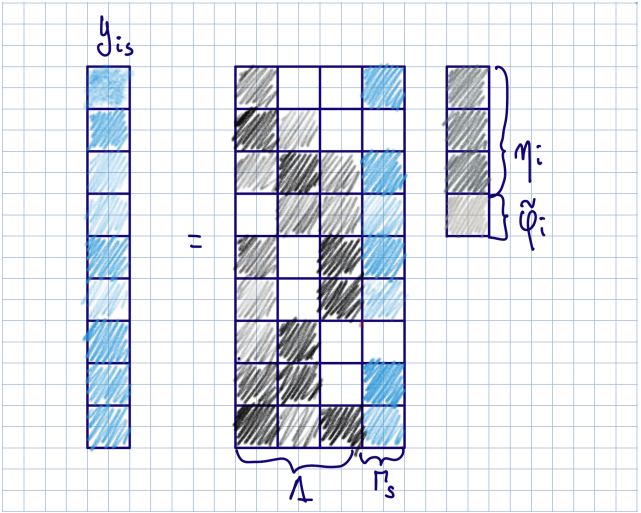
Rethinking the MSFA

- Rewrite the MSFA as

$$y_{is} = \Lambda \eta_{is} + \Gamma \varphi_{is} + \epsilon_{is}. \tag{2}$$

- Here $\Gamma = (\Gamma_1, \ldots, \Gamma_S)$ binds all the study-specific factor loading matrices into a $p \times k$ matrix with $k = \sum_{s=1}^{S} k_s$
- φ_{is} is a k-dimensional augmented vector containing the original $\tilde{\varphi}_{is}$ framed with suitable pattern of zeroes.

Rethinking the MSFA graphically



Structured sparsity for the MSFA

- MSFA permits precisely S study-specific loading matrices Γ_s
- practical scenarios often present more complex situations:
  - two or more studies may present high homogeneity, potentially sharing identical or nearly identical latent representations
  - some studies may involve a highly heterogeneous group of subjects, possibly leading to two or more sub-populations displaying distinct latent representations
- Structured sparsity helps in solving these issues

Bortolato and Canale (2024?)

Flexible MSFA

- For the shared latent factors, we specify

$$\eta_{ih} \sim N(0, \theta_{ih}\tau_h^{\eta}),$$

- For the study specific latent factors, similarly to what we did for Λ previously

$$\varphi_{ih} \sim N(0, \psi_{ih}(x_i)\tau_h^{\varphi}).$$

- In particular, we assume the dependence between the scale parameters of the group-specific latent factors and the group indicator x_i, as follows:

$$\psi_{ih}(x_i) \sim \text{Ber}\{\tilde{\psi}_{ih}\}, \text{ with } \tilde{\psi}_{ih} = \text{logit}^{-1}(x_i^{\top}\beta_h^{\varphi}).$$

- Both τ_h^{η} and τ_h^{φ} are related to (two) CUSP
- For the shared and study-specific loadings

$$\lambda_{jh} \sim N(0, \zeta_h^{\lambda}),\ \zeta_h^{\lambda} \sim \text{I-Ga}(a_{\lambda}, b_{\lambda}),\ \gamma_{jh} \sim N(0, \zeta_h^{\lambda}),\ \zeta_h^{\gamma} \sim \text{I-Ga}(a_{\gamma}, b_{\gamma}).$$

Simulations: scenarios

We simulate data under the following scenarios

- *Scenario 1 — Correct specification*: $S = 3$ groups with sample size $n_1 = n_2 = n_3$, $d = 2$ active shared factors, and $k = 3$ ($[1 + 1 + 1]$ for the groups) specific factors
- *Scenario 2 — Homogeneity between groups*: While we provide $S = 3$ groups, the structure of latent factors is homogeneous among all the studies, i.e. $k = 0$.
- *Scenario 3 — Latent Heterogeneity*: we do not provide the group labels but the data are generated as in Scenario 1.
- *Scenario 4 — Mixed situations*: There exist groups but $k \neq S$

Simulations: metrics & competitors

- We evaluate the performance of the proposed flexible multi-study factor mode (flexMSFA) with the approach proposed by De Vito et al. (2021) (MSFA) and that of Gabski et al. (2023) (Tetris)
- For our method only, we evaluate the ability in discovering the group structure.
- To compare the relative performance of each competitor, we measure the adequacy of the reconstructed the variances due to the shared loadings, and the global variance for each group.
- To this end we compute the *RV* coefficient (Abdi, 2007)defined for two symmetric positive definite matrices S, T as

$$RV_{S,T} = \frac{Tr(S^\top T)}{\sqrt{Tr(S^\top S)Tr(T^\top T)}} \in (0,1),$$

with higher values associated to stronger similarity.

Results: group discovering

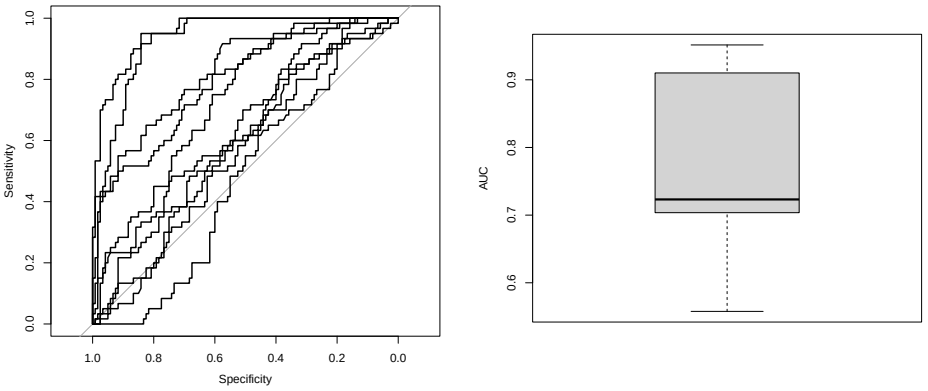


Figure: ROC curve for correct identification of the non-zero patterns in the study specific latent factors (left) and boxplot of the related AUC

Results: variance matrices reconstruction

Scenario	Method	d	k	Ω_1	Ω_2	Ω_3	$\Lambda\Lambda^\top$
1	flexMSFA	2	3	0.90	0.91	0.91	0.95
	25%	2	3	0.82	0.86	0.80	0.93
	75%	2	3	0.96	0.95	0.93	0.98
1	MSFA	2	-	0.43	0.29	0.29	0.17
	25%	2	-	0.27	0.25	0.21	0.06
	75%	2	-	0.48	0.33	0.40	0.35
1	TETRIS	2	0	0.92	0.88	0.90	0.79
	25%	2	0	0.84	0.82	0.70	0.72
	75%	2	0.75	0.95	0.95	0.92	0.90
2	flexMSFA	2	2	0.82	0.82	0.86	0.92
	25%	2	2	0.74	0.70	0.77	0.90
	75%	2	2	0.92	0.93	0.93	0.95
3	flexMSFA	5	0	0.86	0.86	0.86	0.84
	25%	5	0	0.79	0.79	0.79	0.77
	75%	5	0	0.88	0.88	0.88	0.85
4 (a)	flexMSFA	2	3	0.92	0.85	0.90	0.96
	25%	2	3	0.84	0.80	0.78	0.91
	75%	2	3	0.95	0.93	0.95	0.98
4 (b)	flexMSFA	2	3	0.81	0.83	0.83	0.89
	25%	2	3	0.76	0.71	0.76	0.88
	75%	2	3	0.90	0.92	0.93	0.94

Table: Configuration $n > p$, performance of flexMSFA, MSFA, TETRIS based on RV metric on several simulation scenarios.

Outline

Finnish bird co-occurrence data

Data on the co-occurrences of 50 birds species (columns) in Finland in 200 locations (rows) in 5 different sampling campaigns

Illustration 1

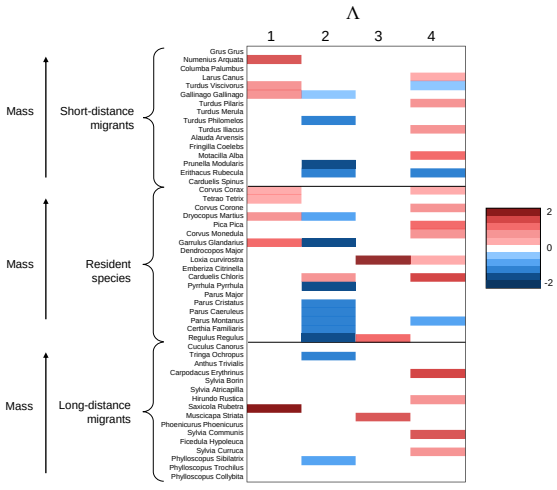
We first focus on a single sampling campaign

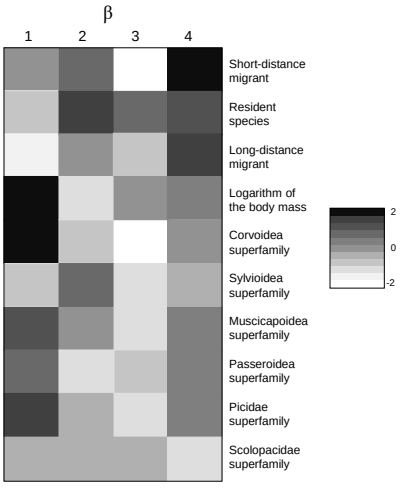
- $\boldsymbol{y}$: $n \times p$ binary matrix of occurrence of **p species** in **n** different **environments**.
- $\boldsymbol{w}$: $n \times c$ **covariate matrix** including habitat type and the 'spring temperature'.
- $\boldsymbol{x}$: $p \times q$ **meta covariate matrix** including **species traits**: the species log body mass, the species migratory strategy and species superfamily.

$$y_{ij} = \mathbb{1}(z_{ij} > 0), \quad z_{ij} = w_i^T \mu_j + \epsilon_{ij}, \quad \epsilon_i = (\epsilon_{i1}, \ldots, \epsilon_{ip})^T \sim N_p(0, \Lambda\Lambda^T + I_p),$$

- $\boldsymbol{z}$: $n \times p$ underlying continuous matrix.
- Λ: loadings matrix with structured increasing shrinkage prior such that the species traits x can impact the covariance structure across species.

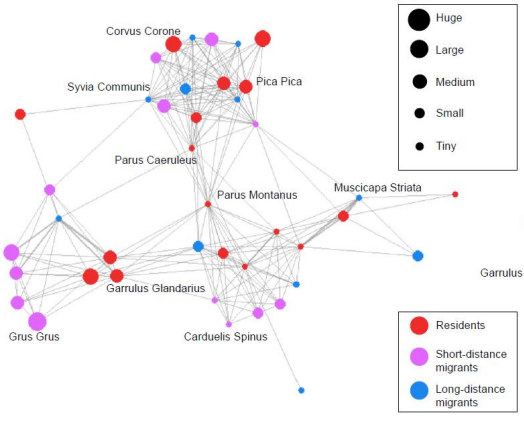
Posterior mode of Loadings





Graph representation of covariance structure

Structured Increasing Shrinkage prior



Multiplicative Gamma Process

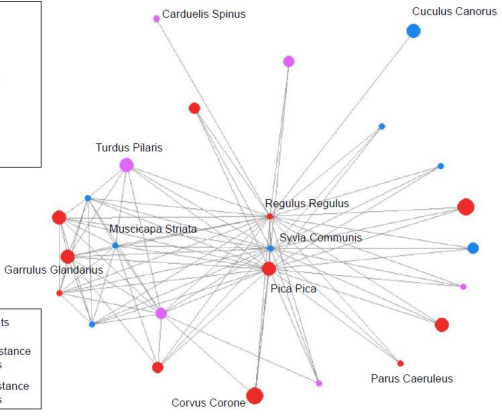


Illustration 2

We now analyze the whole dataset considering the different sampling campaigns.

- The locations are considered as groups in a multi-study framework
- $\boldsymbol{y}$: $n \times p$ binary matrix of occurrence of **p species** in **n** different **years**.
- $\boldsymbol{S} = 200$: number of sites
- we do not use neither the covariates or the metacovariates (only the sampling campaign indicator) but …
- The 200 locations can be clustered into 5 different types of location: Urban, Broadleleaved forests, Coniferous forests, Open, and Wetlands.

Posterior of the group specific factors

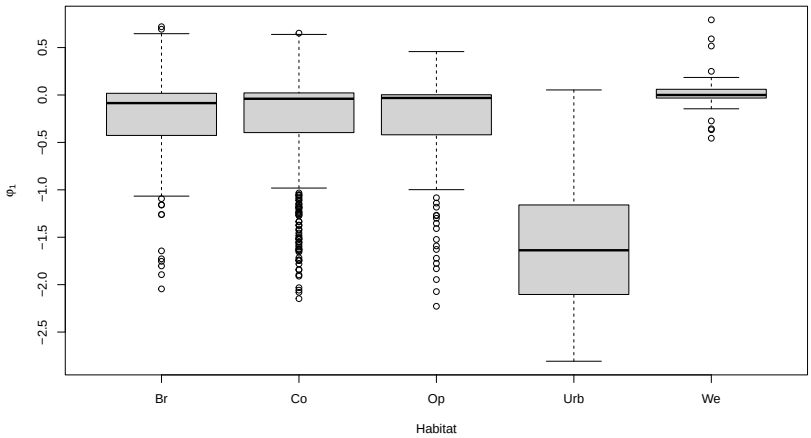


Figure: Estimates of the specific factor φ_1 and type of habitat in the sites.

Wrapping up & essential references

- Sparsity is useful as dimensionality reduction and to promote interpretability
- Bayesian shrinkage prixors can be equipped with regression-like dependence from covariates and metacovariates
- We called this approach structured shrinkage and exploit it in several contexts/directions:
  - Schiavon, L., Canale, A., & Dunson, D. B. (2022). *Generalized infinite factorization models.* Biometrika
  - Schiavon, L., Nipoti, B., & Canale, A. (2024). *Accelerated structured matrix factorization*. JCGS
  - Canale, A., Galtarossa, L., Risso, D., Schiavon, L, & Toto, G. (2023), *Structured factorization for single-cell gene expression data*, arXiv preprint arXiv:2305.11669
  - Bortolato, E. Canale, A., (2024?), *Flexible multi-study factor analysis*, (in preparation)

Joint work with

Lorenzo Schiavon
(University Ca' Foscari)

Elena Bortolato
(University of Padova)

David Dunson
(Duke University)

Thank you for your attention!

University of Padova …

Università degli Studi di Padova

…since 1222

References

- Abdi, H. (2007). Rv coefficient and congruence coefficient. Encyclopedia of measurement and statistics, 849, 853.
- Avalos-Pacheco, A., Rossell, D., & Savage, R. S. (2022). Heterogeneous large datasets integration using b ayesian factor regression. Bayesian Analysis, 17(1), 33–66
- Bhattacharya, A. & Dunson, D. B. (2011). Sparse bayesian infinite factor models. Biometrika, 98(2), 291–306.
- De Vito, R. & Avalos-Pacheco, A. (2023). Multi-study factor regression model: an application in nutritional epidemiology. arXiv preprint arXiv:2304.13077.
- De Vito, R., Bellio, R., Trippa, L., & Parmigiani, G. (2019). Multi-study factor analysis. Biometrics, 75(1), 337–346.
- De Vito, R., Bellio, R., Trippa, L., & Parmigiani, G. (2021). Bayesian multistudy factor analysis for high-throughput biological data. The Annals of Applied Statistics, 15(4), 1723–1741.
- Frühwirth-Schnatter, S. (2023). Generalized cumulative shrinkage process priors with applications to sparse Bayesian factor analysis. Philosophical Transactions of the Royal Society A, 381(2247), 20220148.
- Grabski, I. N., De Vito, R., Trippa, L., & Parmigiani, G. (2023). Bayesian combinatorial multistudy factor analysis. The annals of applied statistics, 17(3), 2212.
- Legramanti, S., Durante, D., & Dunson, D. B. (2020). Bayesian cumulative shrinkage for infinite factorizations. Biometrika, 107(3), 745–752