Adaptive Partition Factor Analysis

Antonio Canale · antonio.canale@unipd.it

Psicostat · January 16, 2026

Joint Work with Elena Bortolato, UPF, Barcelona

Università degli Studi di Padova

General settings

- we observe a p-dimensional vector of item responses (binary / ordinal) to p psychological questionnaire items
- observations are grouped in S independent studies (or cohorts), with $s = 1, \ldots, S$
- in each study we have multiple participants, for a total of n individuals overall ($i = 1, \ldots, n$)
- data y_{is} is the p-dimensional response vector for individual i in study s
- goal: infer latent psychological dimensions (factors) that are comparable across studies

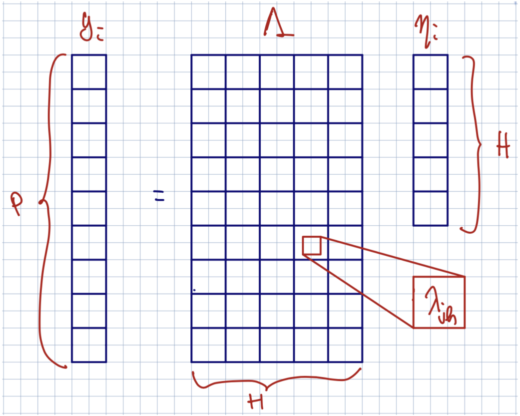
Multi-study setup

- we consider this data in a "multi-study" fashion
- other examples:
  - medical studies performed in differnt hospitals
  - genomics studies performed with differet technological platforms
  - ...
- we are going to follow a factor model approach, and specifically a Multi-Study Factor Analysis (MSFA, De Vito et al. 2017) approach
- MSFA extensions and generalizations:
  - Roy et al. (2021) proposed a perturbed factor analysis that focuses on inferring the shared structure while making use of subject-specific perturbations
  - Grabski et al. (2023) proposed a model allowing for partially-shared latent factors
  - Chandra et al. (2024) proposed a class of subspace factor models with appealing identifiability properties

Factor Analysis (FA)

$$y_i = \Lambda \eta_i + \epsilon_i$$

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



Interpretability and sparsity

- Most of the interest of FA 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
- In this talk I will exploit sparsity within a MSFA framework

Multi study factor analysis

- Multi-study factor analysis (MSFA) assumes the existence of both shared latent factors and study-specific latent factors
- Specifically

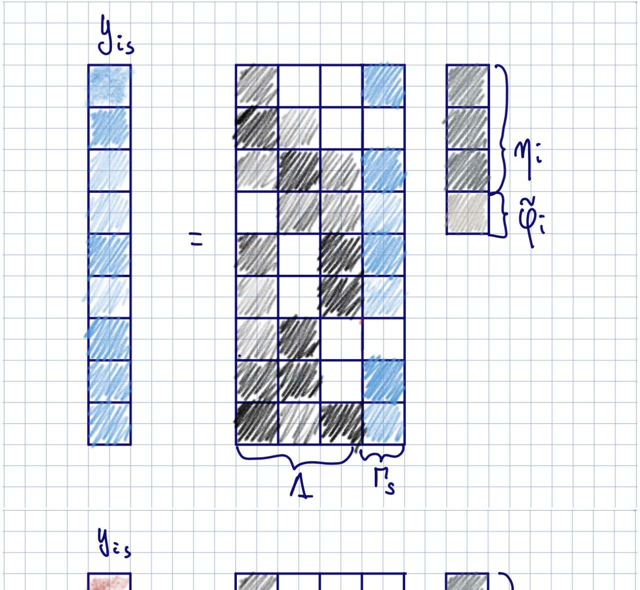
$$y_{is} = \Lambda\eta_{is} + \Gamma_s\varphi_{is} + \epsilon_{is} \tag{1}$$

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 φ_{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



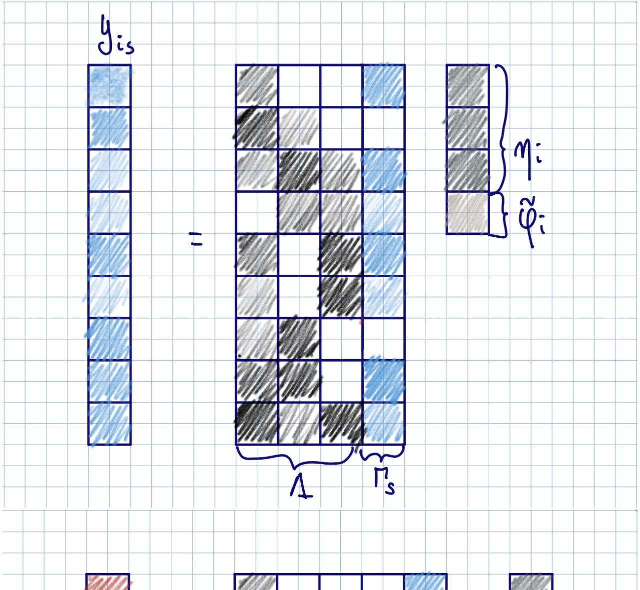
Rethinking the MSFA

- Rewrite the MSFA as

$$y_i = \Lambda\eta_i + \Gamma\varphi_i + \epsilon_i. \tag{2}$$

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

Rethinking the MSFA graphically



Fixed partition vs Adaptive partition

- 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
- An adaptive partition that accounts for the above situations (and beyond) would be useful!

From factor analysis to neural networks

- As customary in factor analysis marginalize out the latent factors $\eta_i \sim N(0, I_p)$,

$$y_i \sim N(\Gamma \varphi_i, \Lambda\Lambda^\top + \Sigma). \tag{3}$$

- Let x_i be a dummy variable scharacterizing the study to which unit i belongs
- The h-th element of the vector φ_i is $\varphi_{ih} = \tilde{\varphi}_{ih}\psi_{ih}$ where $\tilde{\varphi}_{ih}$ is a continuous random variable and $\psi_{ih} = f_h(x_i)$ with f_h a deterministic activation function.
- for MSFA $f_h(x_i) = x_i^\top 1_S$ where 1_S is a S dimensional vector of ones

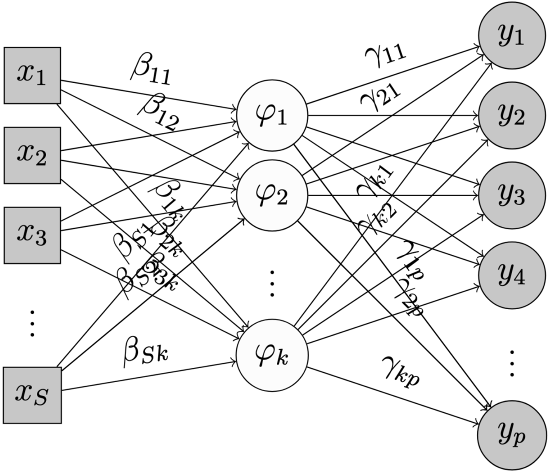
From factor analysis to neural networks

- What about incorporating the information contained in the x_i's in a more flexible manner?
- For example with

$$\psi_{ih} = f_h(x_i^\top \beta_h). \quad (4)$$

- Which leads to a specific single layer neural network where
  - the x_is are the input variables,
  - y_i are output variables,
  - φ_is are the nodes of the hidden layer
  - f_h are the activation functions
  - β_h are the weights between the input and the hidden layer,
  - Γ are the weights between the hidden layer and the outcome

From factor analysis to neural networks



Adaptive partition Factor Analysis (APAFA)

- We encode the study-specific membership into categorical the variables x_i
- The study-specific latent factors are assumed

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

- We then assume a dependence between the scale parameters of the group-specific latent factors and the group indicator x_i, as follows:

$$\psi_{ih}(x_i) \sim \text{Ber}\{\text{logit}^{-1}(x_i^\top \beta)\}.$$

APAFA benefits

- The shrinkage prior on the elements $\varphi_{ih}(x_i)$ enables and promotes, yet does not mandate, the sparse representation of MSFA
- Wide range of scenarios including:
  - two or more studies exhibit high homogeneity and share nearly identical latent representations,
  - some studies involve highly heterogeneous groups of subjects, potentially resulting in two or more sub-populations with distinct latent structures,
  - any combinations of the above.

Time for discussion?

Other (more technical) stuff...

1. Introduction
2. From MSFA to APAFA
3. Theoretical properties
4. Empirical performance
   - Simulation Study
   - Illustration: Finnish birds co-occurrence data
5. References

Identifiability

- Factor loadings are not identifiable due to rotations, i.e .

$$\Lambda\Lambda^T = \tilde{\Lambda}\tilde{\Lambda}^T, \quad \text{if} \quad \tilde{\Lambda} = \Lambda P, \quad \text{and} \quad PP^T = I_k$$

- In the multi-study context and, specifically, in

$$y_i \sim N(0, \Omega_i), \quad \Omega_i = \Lambda\Lambda^\top + \Gamma \text{diag}\{\psi_i\}\Gamma^\top + \Sigma$$

we may additionally face “information switching”

Identifiability of APAFA: information switching

Università degli Studi di Padova

Definition (Information Switching)

Let S_n the number of distinct groups, i.e. $S_n = |\cup_{i=1}^{n} \Omega_i|$. Denote with Ψ the $n \times k$ matrix that stacks in distinct rows all $\psi_i = (\psi_{i1}, \ldots, \psi_{ik})$ and with Ψ_h its generic column with, $\Psi_h \neq 1_n$ for all $h = 1, \ldots, k$. Let Ω_s^* and ψ_s^* $(s = 1, \ldots, S_n)$ be the distinct values of Ω_i and ψ_i, respectively. Similarly, let $W_s^* = \Omega_s^* - \Lambda\Lambda^\top - \Sigma = \Gamma \mathrm{diag}\{\psi_s^*\}\Gamma^\top$. The model suffers from information switching if there exist $\tilde{\Gamma} \neq \Gamma$ and $\tilde{\Psi} \neq \Psi$ such that $W_s^* = \tilde{\Gamma}\mathrm{diag}\{\tilde{\psi}_s^*\}\tilde{\Gamma}$ for all s, with $\tilde{\Psi}_h = 1_n$ for at least one h.

Theorem

If $\Psi_h \neq 1_n$ for all $h \in \{1, \ldots, k\}$ and Γ is of full column rank k with $k < p(p+1)/2$, then the model is resistant to information switching.

Identifiability of APAFA: rotational invariance

Definition (Non-replicable Sparsity Pattern Condition)

All columns of Ψ^*, where Ψ^* is the $S_n \times k$ matrix stacking all the distinct ψ_s^*, are different.

Theorem

Let $\Gamma \in \mathbb{R}^{p \times k}$ be a real matrix with full column rank k. If $P \in \mathcal{O}_k$ and $\Gamma' = \Gamma P$ is a rotation of the specific factors, under the Non-replicable Sparsity Pattern Condition, then for each $s = 1, \dots, S_n$

$$\Lambda\Lambda^\top + \Gamma\Psi_s^*\Gamma^\top + \Sigma = \Lambda\Lambda^\top + \Gamma' diag(\psi_s^*)\Gamma'^\top + \Sigma,$$

if and only if P is a permutation matrix.

Outline

1 Introduction

2 From MSFA to APAFA

3 Theoretical properties

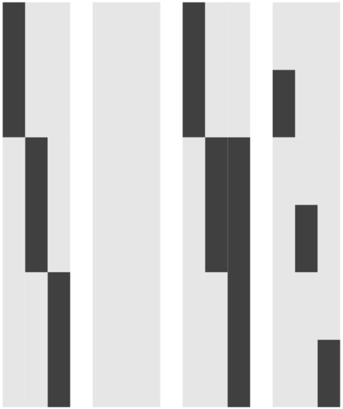
4 Empirical performance
- Simulation Study
- Illustration: Finnish birds co-occurrence data

5 References

Simulations: scenarios

We simulate data under the following scenarios

- *Scenario A — 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 A* — Latent Heterogeneity*: we do not provide the group labels but the data are generated as in Scenario A.
- *Scenario B — Homogeneity between groups*: While we provide $S = 3$ groups, the structure of latent factors is homogeneous among all the studies, i.e. $k = 0$.
- *Scenarios C and D – Mixed situations*: There exist groups but $k \neq S$



(A, A*) **(B)** **(C)** **(D)**

Simulations: metrics & competitors

- We evaluate the performance of APAFA with the approach proposed 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 of each group.

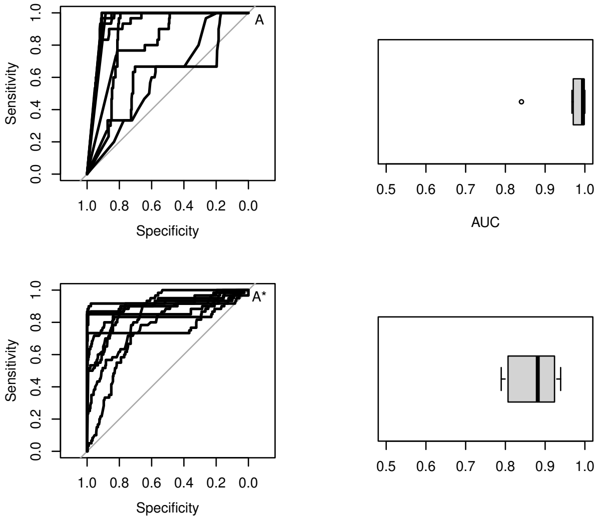
Results: variance matrices reconstruction

Table: Monte Carlo average (and interquartile range) of the posterior mean number of factors and RV coefficients for Ω_1, Ω_2, and Ω_3 on several simulation scenarios (the higher the better).

Scen	Method	d	k	Ω_1	Ω_2	Ω_3
A	APAFA	3.00 (1.00)	3.01 (0.23)	0.90 (0.07)	0.88 (0.07)	0.89 (0.05)
	TETRIS	3.67 (1.26)	4.91 (0.89)	0.92 (0.07)	0.93 (0.09)	0.88 (0.08)
A^*	APAFA	4.00 (1.00)	3.00 (1.00)	0.69 (0.13)	0.78 (0.08)	0.75 (0.08)
B	APAFA	3.00 (0.00)	0.00 (0.00)	0.94 (0.04)	0.94 (0.04)	0.94 (0.04)
	TETRIS	3.00 (0.00)	0.00 (0.00)	0.90 (0.12)	0.92 (0.05)	0.92 (0.09)
C	APAFA	3.00 (0.00)	3.00 (0.05)	0.89 (0.05)	0.78 (0.04)	0.92 (0.02)
	TETRIS	3.00 (1.00)	2.00 (1.01)	0.72 (0.09)	0.75 (0.08)	0.79 (0.05)
D	APAFA	3.00 (0.00)	3.00 (0.12)	0.91 (0.03)	0.88 (0.03)	0.90 (0.04)

Results: group discovering

Università degli Studi di Padova



Finnish bird co-occurrence data

- we have a p-dimensional (binary) vector measuring the co-occurrence if p species of birds
- observations are grouped in $S = 200$ different locations, with $s = 1, \ldots, S$.
- In each location we have repeated sampling campaigns for a total of n observation in total ($i = 1, \ldots, n$)
- data y_{is} is the p-dimensional measurement for location s at sampling campaign i.

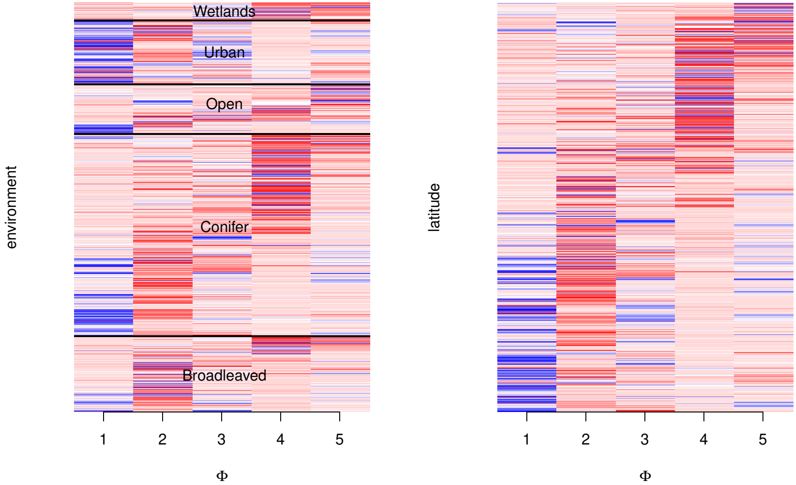
Finnish bird co-occurrence data

- 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 **sampling campaigns**.
- We model species presence or absence using the multivariate probit regression model,

$$y_{ij} = 1(z_{ij} > 0), \quad z_i = \Lambda \eta_i + \Gamma \varphi_i + \epsilon_i, \quad \epsilon_i \sim N(0, \Sigma).$$

- $\boldsymbol{S} = 200$: number of sites
- we do not use any information but the sampling campaign indicator. However ...
- The 200 locations can be clustered into 5 different types of location: Urban, Broadleaved forests, Coniferous forests, Open, and Wetlands.

Posterior estimate of specific factors



Wrapping up & essential references

UNIVERSITÀ DEGLI STUDI DI PADOVA

- The concept of sparsity has been used to generalize the MSFA model
- The model not only enjoys appealing theoretical properties but shares many characteristics of neural networks
- APAFA exploits the broader concept of structured shrinkage:
  - 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. (202+), *Structured factorization for single-cell gene expression data*, arXiv:2305.11669, minor revision submitted
  - Bortolato, E., Canale, A., (202+), *Adaptive Partition Factor Analysis* , arXiv:2410.18939, minor revision submitted
  - Canale, A., Schiavon, L., Stolf, F. (202+), *Identifiable Sparse Bayesian Factorizations via Meta Regression*, Submitted

References

- Avalos-Pacheco, A., Rossell, D., & Savage, R. S. (2022). Heterogeneous large datasets integration using Bayesian factor regression. Bayesian Analysis, 17(1), 33–66
- Bhattacharya, A. & Dunson, D. B. (2011). Sparse Bayesian infinite factor models. Biometrika, 98(2), 291–306.
- Chandra, N. K., Dunson, D. B. & Xu, J. (2024), Inferring covariance structure from multiple data sources via subspace factor analysis, JASA
- De Vito, R. & Avalos-Pacheco, A. (2025). Multi-study factor regression model: an application in nutritional epidemiology, Statistics in Medicine
- 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