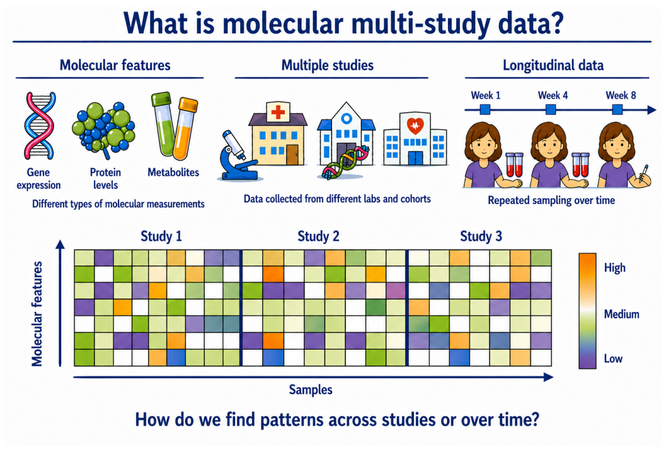
IMPERIAL

Multi-study Factor Regression Biclustering

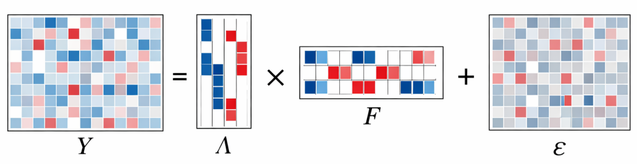
Margarita Grushanina[1]
Leonardo Bottolo[2]
Verena Zuber[1]

High-dimensional Molecular Data



- **Gene expression (RNA-seq, scRNA-seq)** — expression levels of thousands of genes per sample
- **Proteomics** — protein abundance via mass spectrometry
- **Metabolomics** — small-molecule concentrations reflecting cellular processes
- **Epigenomics** — DNA methylation or chromatin accessibility
- **Multi-omics** — integration of multiple molecular layers

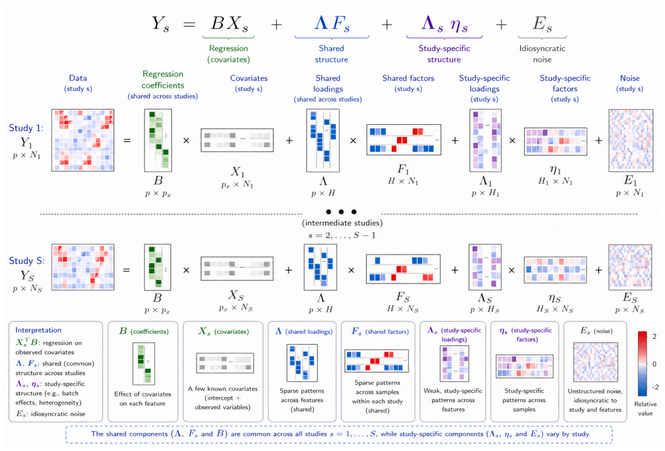
Biclustering with Bayesian Sparse Factor Models



Gao et al. (2016)

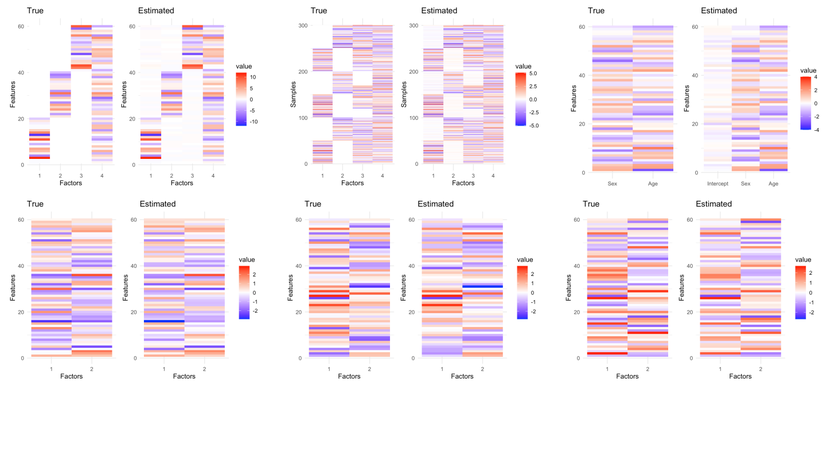
- Identifies **overlapping groups of features and samples** that exhibit similar patterns
- Uses **sparsity-inducing priors** on both loadings and factor scores to select relevant patterns
- Enables **interpretable biclusters**, where only subsets of features are active in subsets of samples
- Existing approaches often **ignore multi-study structure** and **batch effects**

We propose: Multi-study Factor Regression Biclustering (MSFRB)

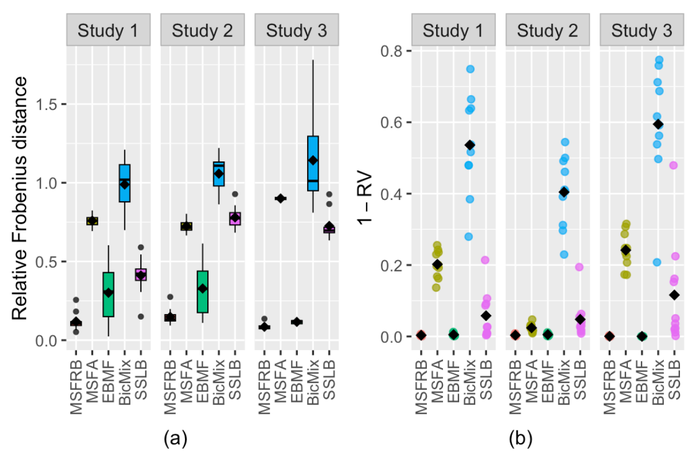


- **Unified framework:** Combines multi-study factor analysis [1], regression [2], and biclustering to model shared and study-specific structure with covariate adjustment
- **Sparse and interpretable:** Uses sparse ESP prior [3] to uncover clear feature–sample patterns (biclusters)
- **Better integration:** Separates common, study-specific, and covariate-driven variation for more informative analysis

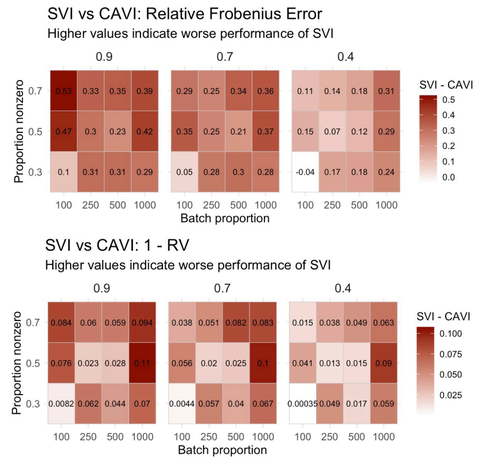
Results



Pattern recovery on simulated data



Comparative performance on simulated data



Performance comparison CAVI vs. SVI

Sparse Exchangeable Shrinkage Process (ESB) Prior

ESP prior on factor loadings

$$\lambda_{ih} \mid \theta_h, \phi_{ih}, \kappa_\Lambda \sim \mathcal{N}(0, \theta_h \phi_{ih} \kappa_\Lambda)$$
$$\theta_h \mid \tau_h \sim \tau_h \mathcal{G}^{-1}(a_\Lambda, b_\Lambda) + (1-\tau_h)\mathcal{G}^{-1}(a_0^\Lambda, b_0^\Lambda)$$
$$\tau_h \sim \mathcal{B}\left(\frac{\alpha_\Lambda}{H}, 1\right)$$
$$\phi_{ih} \sim \mathcal{G}^{-1}(a_\phi, \beta_{i\phi}), \quad \beta_{i\phi} \sim \mathcal{G}(c, d)$$
$$\kappa_\Lambda \sim \mathcal{G}^{-1}(a^\kappa, b^\kappa)$$

ESP prior on factor scores

$$f_{hn} \mid \upsilon_h, \zeta_{hn}, \kappa_F \sim \mathcal{N}(0, \upsilon_h \zeta_{hn} \kappa_F)$$
$$\upsilon_h \mid \pi_h \sim \pi_h \mathcal{G}^{-1}(a_F, b_F) + (1-\pi_h)\mathcal{G}^{-1}(a_0^F, b_0^F)$$
$$\pi_h \sim \mathcal{B}\left(\frac{\alpha_F}{H}, 1\right)$$
$$\zeta_{hn} \sim \mathcal{G}^{-1}(a_\zeta, \beta_{\zeta n}), \quad \beta_{\zeta n} \sim \mathcal{G}(e, w)$$
$$\kappa_F \sim \mathcal{G}^{-1}(c^\kappa, d^\kappa)$$

active factor selection (θ_h, υ_h) global shrinkage (κ_Λ, κ_F) element-wise sparsity (ϕ_{ih}, ζ_{hn})

Identification via unique support

Problem: Factor models are not identifiable due to **label switching** and **information sharing** across components
Idea: Identify factors via their **unique support** across features and samples

$$B_h = G_h \otimes T_h \quad \Rightarrow \quad U_h = B_h \cdot \left(1 - \sum_{h' \neq h} B_{h'} - \sum_{s,r} B_r^{(s)}\right)$$

- G_h: active features, T_h: active samples
- B_h: support of factor h
- U_h: **unique support** not explained by other factors

Identification: a factor is identifiable if $U_h \neq 0$
Key insight: sparsity induced by the ESP prior naturally promotes unique support, making identification likely in practice

Variational inference

- Developed **coordinate ascent (CAVI)** and **stochastic (SVI)** algorithms for scalable inference in the MSFRB model
- Derived **mean-field updates for the ESP prior**, handling its **hierarchical mixture (spike-and-slab) structure**
- Introduced **variational responsibilities** for latent inclusion variables, enabling **continuous factor selection**
- Jointly infers **shared and study-specific factors, loadings, and regression effects** within a unified framework

For a simple ESP prior:

$$\lambda_{ih} \mid \theta_h \sim \mathcal{N}(0, \theta_h), \ \theta_h \mid \gamma_h \sim \gamma_h \mathcal{G}^{-1}(a_\Lambda, b_\Lambda) + (1-\gamma_h)\mathcal{G}^{-1}(a_0, b_0), \ \gamma_h \mid \tau_h \sim \text{Bernoulli}(\tau_h), \ \tau_h \sim \mathcal{B}(\tfrac{\alpha_\Lambda}{H}, 1).$$

Mean-field:

$$q(\gamma_h) = \text{Bernoulli}(r_h), \ q(\tau_h) = \mathcal{B}(\tilde{a}_{\tau h}, \hat{b}_{\tau h}).$$

$$r_h = (1 + \exp(s_h^{\text{spike}} - s_h^{\text{slab}}))^{-1}, \text{ with}$$

$$s_h^{\text{slab}} = E_q[\log \tau_h] + E_{q_{\text{slab}}}[\log p(\theta_h \mid \gamma_h = 1) + \Sigma_{i=1}^p \log p(\lambda_{ih} \mid \theta_h)],$$

$$s_h^{\text{spike}} = E_q[\log(1-\tau_h)] + E_{q_{\text{spike}}}[\log p(\theta_h \mid \gamma_h = 0) + \Sigma_{i=1}^p \log p(\lambda_{ih} \mid \theta_h)].$$

Responsibilities enable continuous selection of active factors under the ESP prior.

Affiliations
[1] Department of Epidemiology and Biostatistics, School of Public Health, Imperial College London
[2] Department of Genomic Medicine, University of Cambridge

References
[1] De Vito et al. (2021), Ann. Appl. Stat.; [2] De Vito and Avalos-Pacheco (2025), Stat. Med.; [3] Frühwirth-Schnatter (2023), Phil. Trans. R. Soc. A.; [4] Gao et al. (2016), PLoS Comput. Biol.

UK Research and Innovation