UMassAmherst

School of Public Health
& Health Sciences
Biostatistics and Epidemiology

# Graphical Modeling of Multi-Study Data with Multi-Study Factor Analysis

Katherine H. Shutta[1], Roberta De Vito[2*], Raji Balasubramanian[1*]

[1] Department of Biostatistics and Epidemiology, University of Massachusetts Amherst [2] Department of Biostatistics and Data Science Initiative, Brown University *Equal contribution.

## Abstract

- Multi-study factor analysis (MSFA) is a method for performing factor analysis on measurements obtained across multiple studies [1].
- MSFA identifies a set of latent factors shared among studies and a set of latent factors that are specific to each study.
- Gaussian graphical models (GGMs) are networks of nodes corresponding to variables and weighted edges corresponding to partial correlations between variables.
- GGM estimation provides scope for network-level analysis of conditional dependencies present in a set of variables.
- In this work, we leverage the MSFA framework to estimate shared and study-specific GGMs.

## Multi-Study Factor Analysis (MSFA)

Multi-Study Factor Analysis (MSFA) is a form of factor analysis that estimates *shared* factors and *study-specific* factors:

$$\mathbf{x}_{is} = \Phi\mathbf{f}_{is} + \Lambda_s\mathbf{l}_{is} + \mathbf{e}_{is} \tag{1}$$

- $\mathbf{x}_{is}$ is a $P \times 1$ vector of measured variables for the $i^{th}$ observation in study $s$
- $\Phi$ is a $P \times K$ matrix of shared loadings
- $\mathbf{f}_{is}$ is a $K \times 1$ vector of shared factors, assumed to be $\sim N(0,1)$
- $\Lambda_s$ is a $P \times J_s$ matrix called of study-specific loadings
- $\mathbf{l}_{is}$ is a $J_s \times 1$ vector of study-specific factors, assumed to be $\sim N(0,1)$
- $\mathbf{e}_{is}$ is the residual error, with $\mathbf{e}_i \sim N_P(\mathbf{0}_P, \Psi)$ where $\Psi = diag(\psi_1, \ldots, \psi_P)$ is a diagonal matrix

Under these assumptions, the data are distributed as:

$$\mathbf{x}_{is} \sim MVN(\mathbf{0}_P, \Sigma_s = \Phi\Phi^T + \Lambda_s\Lambda_s^T + \Psi_s \tag{2}$$

## Gaussian Graphical Models (GGMs)

- Gaussian graphical models (GGMs, also known as partial correlation networks) are a commonly used graphical model for multivariate normal data [2]. In a GGM, nodes correspond to variables (e.g., genes, proteins, metabolites) and edges correspond to partial correlations between variables.
- MSFA assumes multivariate normal data, so data analyzed with this methods can also be analyzed with correlation networks and GGMs.
- For multivariate normal data with covariance matrix $\Sigma$, GGM estimation takes advantage of the following relationship between the partial correlation $\omega_{i,j|X_{-i,-j}}$ and the precision matrix $\Theta = \Sigma^{-1}$ ([3]):
$$\omega_{i,j|X_{-i,-j}} = -\frac{\theta ij}{\sqrt{\theta_{ii}}\sqrt{\theta_{jj}}} \tag{3}$$
- Edges in a GGM therefore correspond to nonzero entries of $\Theta$, and edge weights are calculated from $\Theta$ with the formula above
- Edges indicate conditional dependence, where the conditioning is on the state of the rest of the network nodes
- Conditional dependence, i.e., edges, between two nodes means that the observed association between two nodes cannot be explained through associations with any of the other nodes in the network.

## Covariance Matrix Decomposition with MSFA-X

Figure 1:Under the MSFA-X model formulation, the covariance matrix of the data in the $s^{th}$ study can be decomposed as $\Sigma_s = \Phi\Phi^T + \Lambda_s\Lambda_s^T + \Gamma + \mathbf{H}_s$. Here, we show the decomposition for simulated data. Our goal is to recover this decomposition for a given input dataset.

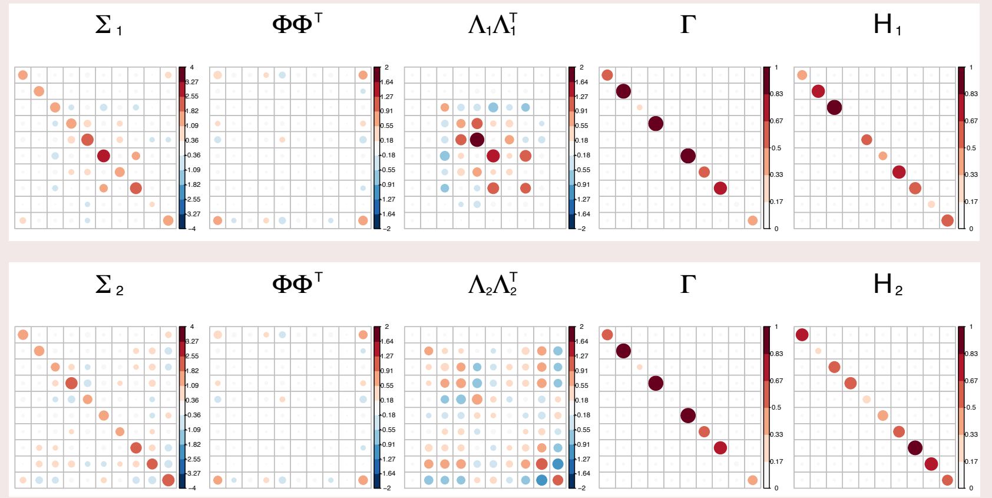


Figure 2:The shared and study-specific precision matrices are shown here. Left, $(\Phi\Phi^T + \Gamma)^{-1}$; center, $(\Lambda_1\Lambda_1^T + \mathbf{H}_1)^{-1}$; right, $(\Lambda_2\Lambda_2^T + \mathbf{H}_2)^{-1}$. Estimation of these matrices allows us to estimated shared and study-specific GGMs. Red indicates positive values; blue, negative.

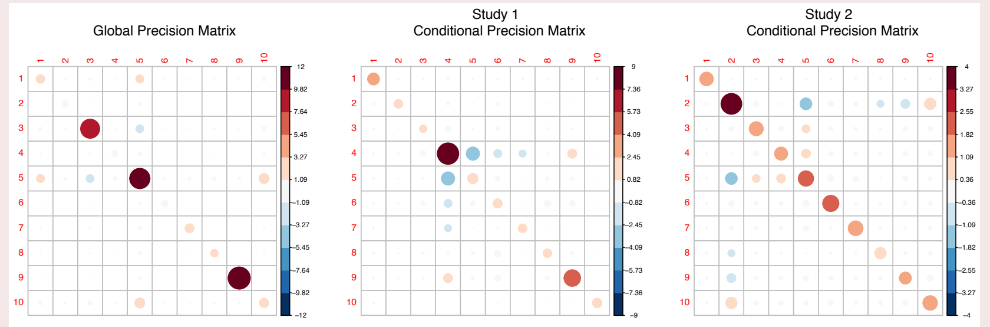
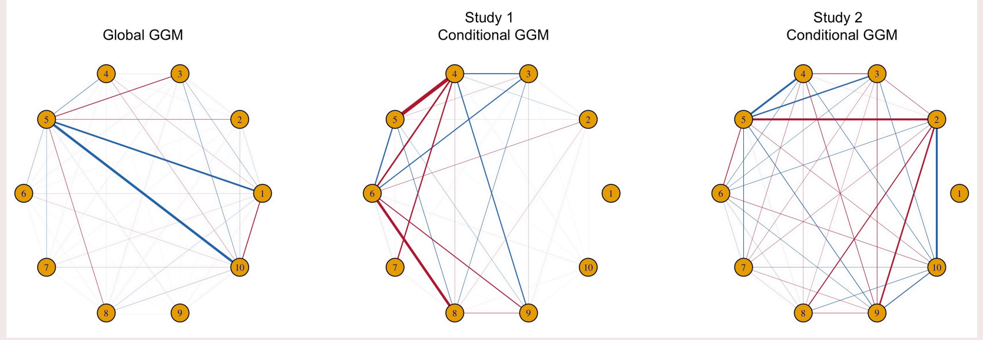


Figure 3:Shared and study-specific GGMs for study 1 and study 2 in the above example. Edge width is proportional to magnitude of partial correlation between two variables. Red indicates positive values; blue, negative. Note that the sign of the partial correlation is opposite of the sign of the corresponding precision matrix entry, so red edges in the network match blue entries in the precision matrix and vice-versa.



## ECM Algorithm for MSFA-X

```
Initialize $\theta_0 = (\Phi_0, \Lambda_1^0, \ldots, \Lambda_S^0, \Gamma_0, \mathbf{H}_1^0, \ldots, \mathbf{H}_S^0)$
Specify nIt = number of iterations
for $t$ in 1:nIt do
    E-step: calculate $Q(\theta) = \mathbb{E}[\ell_c(\theta|\mathbf{x}, \theta^{(t)})]$
    CM-1 step: $\Phi^{t+1} \leftarrow argmax_\Phi Q(\theta|\Lambda_1^t, \ldots, \Lambda_S^t, \Gamma^t, \mathbf{H}_1^t, \ldots, \mathbf{H}_S^t)$
    for s in 1:S do
        CM-2 step: $\Lambda_s^{t+1} \leftarrow argmax_{\Lambda_s} Q(\theta|\Phi^{(t+1)}, \Gamma^t, \mathbf{H}_1^t, \ldots, \mathbf{H}_S^t)$
    end for
    CM-3 step: $\Gamma^{t+1} \leftarrow argmax_\Gamma Q(\theta|\Phi^{t+1}, \Lambda_1^{t+1}, \ldots, \Lambda_S^{t+1}, \mathbf{H}_1^t, \ldots, \mathbf{H}_S^t)$
    for s in 1:S do
        CM-4 step: $\mathbf{H}_s^{t+1} \leftarrow argmax_{\mathbf{H}_s} Q(\theta|\Phi^{(t+1)}, \Lambda_1^{(t+1)}, \ldots, \Lambda_S^{t+1}, \Gamma^{t+1})$
    end for
end for
```

## Extended Multi-Study Factor Analysis (MSFA-X)

Suppose we reformulate the MSFA model as:

$$\mathbf{x}_{is} = \Phi\mathbf{f}_{is} + \Lambda_s\mathbf{l}_{is} + \mathbf{g}_{is} + \mathbf{h}_{is} \tag{4}$$

where we model two sources of noise, an overall noise $\mathbf{g}_{is}$ and a study-specific noise $\mathbf{h}_{is}$, each distributed as follows:

- $\mathbf{g}_{is} \sim N_P(0, \Gamma)$
- $\Gamma = diag(\gamma_1, \ldots, \gamma_P)$
- $\mathbf{h}_{is} \sim N_P(0, \mathbf{H}_s)$
- $\mathbf{H}_s = diag(\eta_{1s}, \ldots, \eta_{Ps})$

Under these assumptions, the data are distributed as:

$$\mathbf{x}_{is} \sim MVN(\mathbf{0}_P, \Sigma_s = \Phi\Phi^T + \Lambda_s\Lambda_s^T + \Gamma + \mathbf{H}_s) \tag{5}$$

## How MSFA-X Estimates Shared and Study-Specific GGMs

MSFA does not allow estimation of a shared and study-specific covariance or precision matrix. The decomposition $\Sigma_s = \Phi\Phi^T + \Lambda_s\Lambda_s^T + \Psi_s$ cannot be split into shared and study-specific components, since the noise term $\Psi_s$ contains shared and study-specific noise. MSFA-X decomposes the term $\Psi_s$ into a shared noise component $\Gamma$ and a study-specific noise component $\mathbf{H}_s$ (see Figure 1). This decomposition gives the following conditional distributions:

- **Study-specific**
  - Conditional covariance: $Cov(\mathbf{x}_{is}|\mathbf{f}_{is}, \mathbf{g}_{is}) = \Lambda_s\Lambda_s^T + \mathbf{H}_s$
  - Conditional precision: $\Theta_{\mathbf{x}_{is}|\mathbf{f}_{is},\mathbf{g}_{is}} = (\Lambda_s\Lambda_s^T + \mathbf{H}_s)^{-1}$
- **Shared**
  - Conditional covariance: $Cov(\mathbf{x}_{is}|\mathbf{l}_{is}, \mathbf{h}_{is}) = \Phi\Phi^T + \Gamma$
  - Conditional precision: $\Theta_{\mathbf{x}_{is}|\mathbf{l}_{is},\mathbf{h}_{is}} = (\Phi\Phi^T + \Gamma)^{-1}$

These conditional precision matrices are used to construct shared and study-specific GGMs using the relationship in Equation (3). An example is shown in Figure 2 and Figure 3.

## Parameter Estimation with Expectation-Conditional Maximization (ECM)

- To estimate the parameters in the MSFA-X model, we build on the work of [4], [5], and [1] in applying a variant of the expectation-maximization algorithm.
- To simultaneously maximize our objective function over all parameters in the model is not computationally feasible. However, coupling an E-step with a sequence of conditional M-steps is a solution to this problem ([5]).
- We describe this algorithm at left.

## Simulation Study Design

The ability of the original ECM algorithm to correctly recover the shared and study-specific factor loadings has been previously demonstrated via simulation study in [1]. We performed a comparable simulation study with our extended algorithm, assessing recovery of $\Phi, \Lambda_s, \Psi_s, \Gamma$, and $\mathbf{H}_s$, $s = 1 \ldots S$. Starting parameters were:

- $S = 2$ studies
- $k = 1$ shared latent factor
- $j_1 = j_2 = 1$ study-specific factor per study
- $n_1 = n_2 = 2000$ samples per study
- Nonzero entries of $\Phi \sim Unif(-1, 1)$
- Nonzero entries of $\Lambda_s \sim Unif(-1, 1), s = 1, \ldots, S$
- $diag(\Gamma) \sim Unif(1, 2)$
- $diag(\mathbf{H}_s \sim Unif(1, 2), s = 1, \ldots, S$

## Simulation Study Results

Figure 4:An example of estimated GGMs (left), true GGMs (center), and error (calculated as estimated - true partial correlations; right) for one iteration of the simulation study. The network structure of the true GGM is largely recovered using the MSFA-X algorithm.

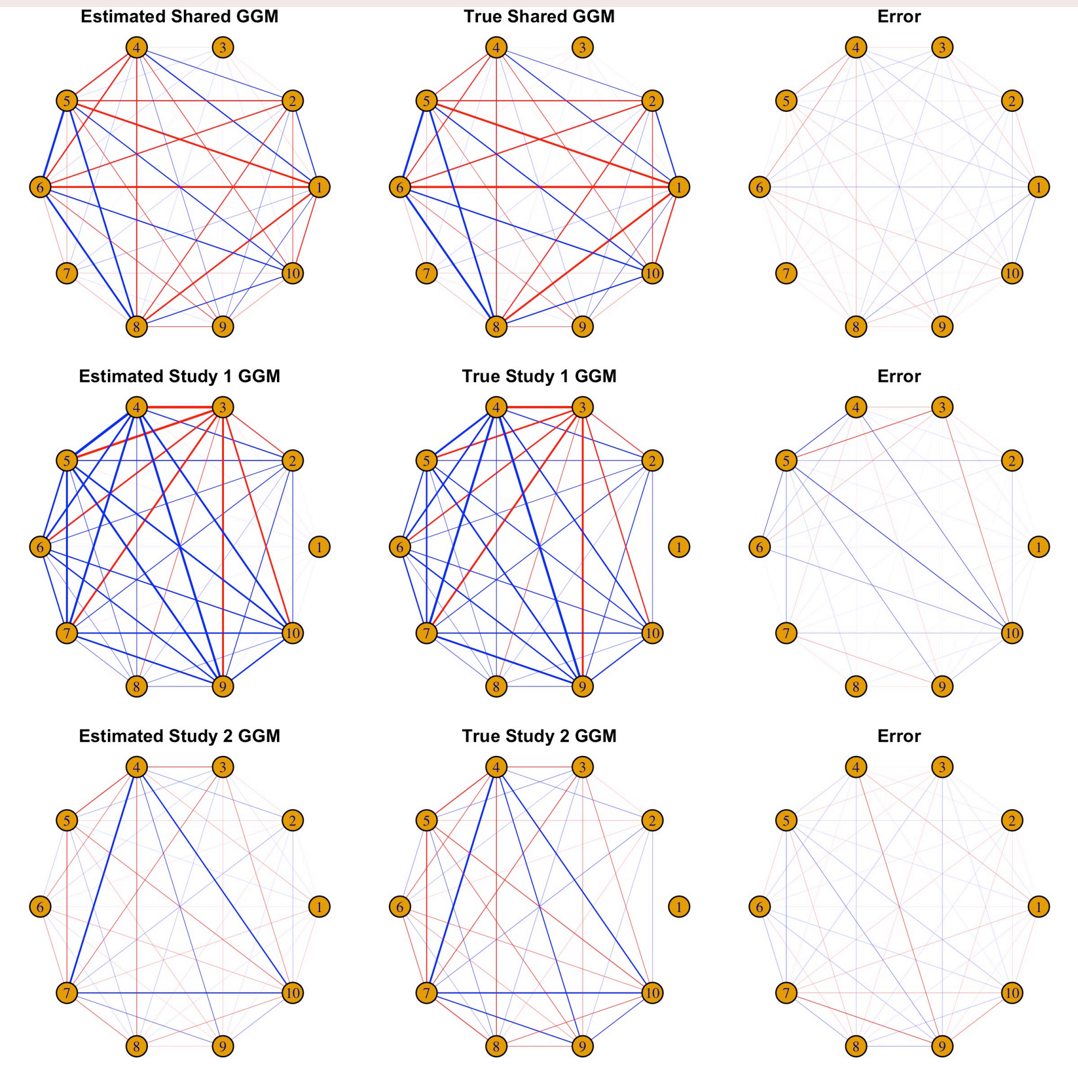
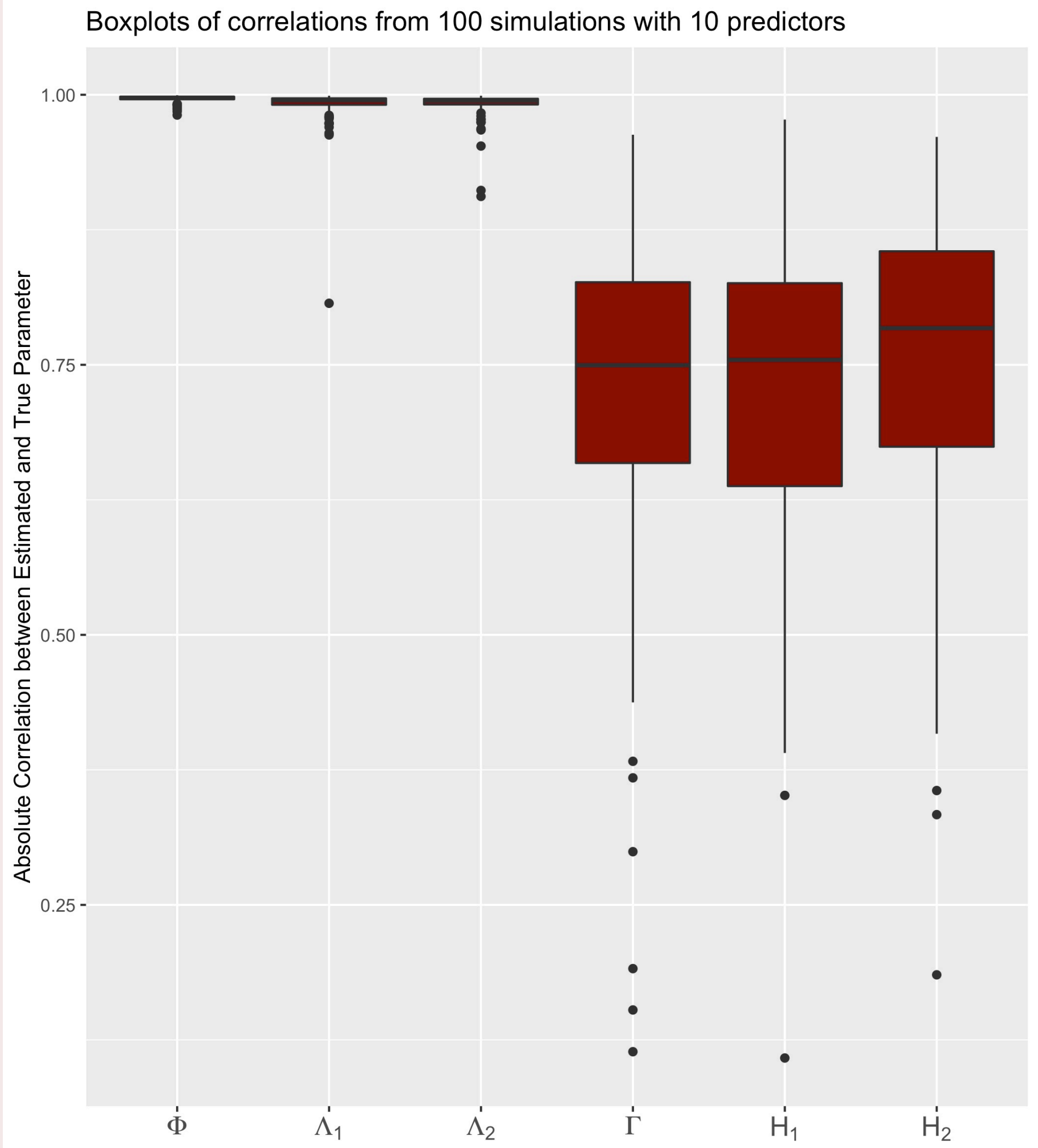


Figure 5:Boxplots show the correlation between the true parameter values and the estimated parameter values over 100 iterations using 10 predictors, starting the algorithm with the factor analysis method described in [1]. While some parameters are well-estimated, others show a modest correlation.



## Discussion of Simulation Study Results

- With the setting above, we ran 100 simulations. An example of the results of one simulation is shown in Figure 4. To assess how well true parameters are recovered from the simulated data, we measured the absolute value of the correlation between true and estimated parameters. (Absolute value is used because of sign indeterminacy in factor analysis.)
- The factor loading parameters $\Phi$ and $\Lambda_s$ are estimated very well; the noise parameters $\Gamma$ and $\mathbf{H}_s$ show modest correlation with the true values.
- We hypothesize that there may be a region of possible solutions for $\Gamma$ and $H_s$ rather than a unique solution, leading to the lower absolute correlation observed for these parameters vs. that observed for $\Phi$ and $\Lambda_s$. This is an important area of further exploration.
- The algorithm does not always converge correctly for certain starting values or simulation settings.

## Future Work

Our studies demonstrate that this approach has potential to correctly estimate the shared and study-specific noise parameters $\Gamma$ and $\mathbf{H}_s$, but that further improvements are needed, including:

- Assess identifiability of the solution for $\Gamma$ and $H_s$; explore more thorough simulation settings and understand convergence problems better.
- Improve algorithm speed; solving for $\Gamma$ requires solving $p$ polynomial equations per iteration, which is time-consuming and not feasible for high-dimensional data.
- Test algorithm on real multi-study data to determine feasibility of estimation on practical values for sample size $n$ and number of predictors $p$.

## References

[1] Roberta De Vito, Ruggero Bellio, Lorenzo Trippa, and Giovanni Parmigiani.
Multi-study factor analysis.
*Biometrics*, 75(1):337–346, 2019.

[2] Caroline Uhler.
Gaussian graphical models: an algebraic and geometric perspective.
*arXiv preprint arXiv:1707.04345*, 2017.

[3] Steffen L Lauritzen.
*Graphical models*, volume 17.
Clarendon Press, 1996.

[4] Donald B Rubin and Dorothy T Thayer.
Em algorithms for ml factor analysis.
*Psychometrika*, 47(1):69–76, 1982.

[5] Xiao-Li Meng and Donald B Rubin.
Maximum likelihood estimation via the ecm algorithm: A general framework.
*Biometrika*, 80(2):267–278, 1993.

## Acknowledgements

Research reported in this poster was supported by the National Institutes of Health under award number NIH R01HL122241.

## Contact Information

- Correspondence: Kate Hoff Shutta, kshutta@umass.edu
- Balasubramanian Lab Site: https://raji-lab.github.io/
- De Vito Lab Site: https://rdevito.github.io/web/