Multiple Graphical horseshoe estimator for modeling correlated precision matrices

Claudio Busatto

Department of Statistical Sciences, University of Padua

SISBayes 2025 - Padova

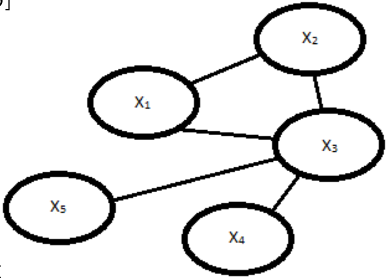
Multiple Graphical horseshoe

- **Goal**: joint analysis of complex correlated (biological) networks
- **Statistical challenge**: infer the connections between variables by exploiting the similarity between groups
- **Model approach**:
  - Bayesian multiple graphical models
  - Non-decomposable GGM
  - Multivariate horseshoe prior
  - "Cut"-model for posterior edge selection
- **Reference**: Busatto C. & Stingo F.C. (2025). *Inference of multiple high-dimensional networks with the Graphical Horseshoe prior*. Journal of Computational and Graphical Statistics, 1-12

Graphical networks

- **Graphs**
  - Represent a collection of variables $\mathbf{x} = [x_1, \ldots, x_p]$ with a set of vertex $\mathcal{V} = \{1, \ldots, p\}$
  - Encode conditional independence by a set of edges $\mathcal{E}$: for any pair of vertices i and j it yields

$$(i,j) \notin \mathcal{E} \iff x_i \perp\!\!\!\perp x_j \mid \mathbf{x}_{\mathcal{V}\setminus\{i,j\}}$$



- **Example**: x_1 and x_5 are conditionally independent
- **Graphical models**: use graph structure to model the dependencies among a set of variables
- **When p increases the percentage of decomposable graphs decreases**

Gaussian graphical models

- **Likelihood**: multivariate normal assumption

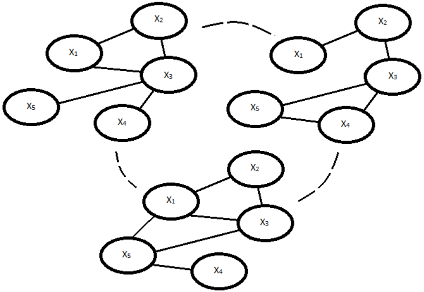
$$\mathbf{x}_s \mid \boldsymbol{\Omega} \sim \mathcal{N}_p\left(\mathbf{0}, \boldsymbol{\Omega}^{-1}\right), \quad s = 1, \dots, n$$

  - $\mathbf{x}_s$ is the p-dimensional vector of observed data for sample s
  - $\boldsymbol{\Omega} = \boldsymbol{\Sigma}^{-1}$ is the precision matrix, inverse of covariance matrix
  - $\omega_{ij} = 0$ if variables i and j are conditionally independent

- **Goal**: infer precision matrix $\boldsymbol{\Omega}$

- **Fundamental assumption**: true precision matrix $\boldsymbol{\Omega}$ is sparse

- **Multiple graphical models**
  - Graphical model for each group over a common set of variables
  - Improve performances by exploiting the similarity among groups



- **Likelihood**: multivariate normal assumption for each group

$$\mathbf{x}_{sk} \mid \boldsymbol{\Omega} \sim \mathcal{N}_p\left(\mathbf{0}, \boldsymbol{\Omega}_k^{-1}\right), \quad \begin{array}{l} s = 1, \ldots, n_k \\ k = 1, \ldots, K \end{array}$$

  - $\mathbf{x}_{sk}$ is the p-dimensional vector of observed data for sample s of group k

Hierarchical Bayesian model

- **Prior distribution for $\boldsymbol{\Omega}$** (Li et al., 2019; Peterson et al., 2020)

$$\pi\left(\omega_{jj}^{k}\right) \propto 1, \quad j = 1, \ldots, p, \quad k = 1, \ldots, K$$

$$\pi\left(\boldsymbol{\Omega}_1, \ldots, \boldsymbol{\Omega}_K \mid \boldsymbol{\Psi}_{ij} : i < j\right) \propto \prod_{i<j} \mathcal{N}_K\left(\boldsymbol{\omega}_{ij} \mid \mathbf{0}, \boldsymbol{\Psi}_{ij}\right) \cdot \mathbb{I}_{\left(\boldsymbol{\Omega}_1, \ldots, \boldsymbol{\Omega}_K \in \mathbb{M}^+\right)}$$

  - $\boldsymbol{\omega}_{ij}$ is the K-dimensional vector of entries ω_{ij}^{k}
  - Decomposition $\boldsymbol{\Psi}_{ij} = \boldsymbol{\Delta}_{ij}\mathbf{R}\boldsymbol{\Delta}_{ij}$ (Barnard et al., 2000)
  - $\boldsymbol{\Delta}_{ij} = \mathrm{diag}\left(\tau_1\lambda_{ij,1}, \ldots, \tau_K\lambda_{ij,K}\right)$, where $\boldsymbol{\tau}$ is the vector of global shrinkage parameters and $\lambda_{ij,k}$ are the local shrinkage parameters
  - $\mathbf{R} \in \mathbb{M}_{+}^{K \times K}$ valid correlation matrix

- **Prior distribution for R** (Peterson et al., 2020)

$$\pi\left(\mathbf{R}\right) \propto 1 \cdot \mathbb{I}_{\left(\mathbf{R} \in \mathbb{M}^+\right)}.$$

Hierarchical Bayesian model

- **Prior distribution for λ and τ:** Horseshoe prior (Carvalho et al., 2010)

$$\lambda_{ij,k} \sim \mathcal{HC}^{+}(0,1)$$
$$\tau_k \sim \mathcal{HC}^{+}(0,1)$$

  - $\mathcal{HC}^{+}(0,1)$ denotes the positive half-Cauchy distribution
  - In linear regression: $\mathbb{E}\left(\beta_j^{HS} \mid \lambda_j, \tau, \sigma^2, \mathbf{y}, \mathbf{X}\right) = (1-\kappa_j)\beta_j^{OLS}$
  - Data augmentation scheme: if $\lambda^2 \mid \eta \sim \mathcal{IG}\left(1/2, 1/\eta\right)$ and $\eta \sim \mathcal{IG}\left(1/2, 1\right)$, then $\lambda \sim \mathcal{HC}^{+}\left(0,1\right)$ (Makalic & Schmidt, 2016)

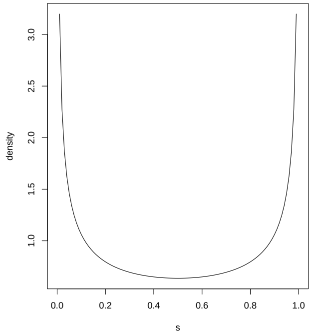


Figure 1: HS shrinkage

- **If $K = 1$ the model reduces to Graphical horseshoe** (Li et al., 2019)

Posterior inference

- **Algorithm**
  - At each iteration parameters are sampled from full-conditionals
- **Sample Ω**
  - Precision matrices are updated following Wang (2015): column-wise update
  - Involve inversion of p covariance matrices
- **Sample λ and τ**
  - Shared amount of local and global shrinkage
  - **Problem**: intractable full-conditional distribution
- **Sample R**
  - MH step based on parameter expansion method from Liu & Daniels (2006)
  - Allows to sample all at once from an inverse-Wishart distribution

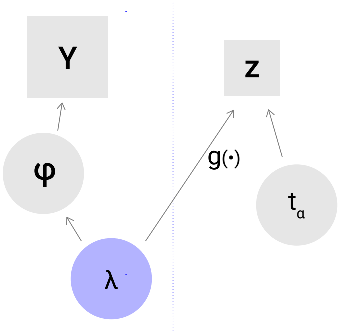
Posterior edge selection

- **MPM model** (Barbieri & Berger, 2004):
  - Includes only those edges with marginal posterior probability greater than $1/2$
  - Threshold equal to $1/2$ often is not optimal
- **Cut-model** (Zigler et al., 2013):
  - Quasi-Bayesian method
  - Cut function can be seen as a modularization of the model
  - Prevents model feedback, which can negatively affect the performances of the model
  - **Goal**: estimation of the optimal threshold t_α

Posterior edge selection

- **Prior distributions**:
  - $z_{ij} \mid \kappa_{ij}, t_\alpha \sim \mathcal{B}\left(q_{ij}^{\alpha}\right)$, where
    $q_{ij}^{\alpha} = \Phi\left(\kappa_{ij} - t_\alpha\right)$
  - $t_\alpha \sim \text{Beta}\left(a, b\right)$

- **Posterior update**:
  - Parameters z_{ij}
    are updated with a Gibbs step
  - Threshold t_α is updated with a MH step



Posterior edge selection

- Cut-models are an approximation of the full-Bayesian conditional distribution
- The algorithm often fails to converge to a well-defined distribution: possible solution is the tempered cut algorithm (Plummer, 2015)
- Whereas cut-models can outperform full-Bayesian models in terms of performance and computational efficiency, doubts still arise about the validity and the convergence of such models, which should be tested through different sampling methods.

Simulation results: $n = 50$, $p = 100$

$n = 50$ $p = 100$	*Independent* ($p_0 = 0.039$)							*Coupled* ($p_0 = 0.036$)						
	Acc	MCC	TPR	FPR	AUC	Fr Loss	$\hat{p}_0$	Acc	MCC	TPR	FPR	AUC	Fr Loss	$\hat{p}_0$
mGHS_{MPM}	0.634 (0.037)	0.135 (0.022)	0.709 (0.031)	0.369 (0.038)	0.749 (0.027)	20.762 (1.454)	0.382	0.555 (0.027)	0.073 (0.023)	**0.642** (0.041)	0.448 (0.027)	0.656 (0.041)	17.429 (1.343)	0.454
mGHS_{t^α}	0.906 (0.195)	0.302 (0.087)	0.245 (0.209)	0.067 (0.211)	0.749 (0.027)	20.762 (1.454)	0.074	0.884 (0.229)	0.144 (0.066)	0.144 (0.256)	0.089 (0.247)	0.656 (0.041)	17.429 (1.343)	0.091
GHS_{MPM}	0.666 (0.022)	0.153 (0.018)	**0.716** (0.029)	0.336 (0.022)	**0.766** (0.022)	20.747 (1.533)	0.351	0.568 (0.028)	0.077 (0.022)	0.640 (0.040)	0.434 (0.029)	0.660 (0.038)	17.191 (1.139)	0.440
fJGL	0.932 (0.010)	**0.313** (0.025)	0.447 (0.048)	0.048 (0.012)	0.700 (0.020)	**20.069** (0.969)	0.064	0.953 (0.009)	**0.235** (0.032)	0.218 (0.070)	0.019 (0.011)	0.599 (0.030)	**16.181** (0.645)	0.026
gJGL	0.930 (0.010)	0.311 (0.026)	0.450 (0.048)	0.049 (0.012)	0.700 (0.020)	20.127 (1.032)	0.066	0.953 (0.009)	0.231 (0.031)	0.216 (0.069)	0.019 (0.011)	0.598 (0.030)	16.653 (0.710)	0.026
$\text{GemBAG}_{\text{MPM}}$	**0.962** (0.001)	0.176 (0.043)	0.047 (0.023)	**0.001** (0.001)	0.683 (0.070)	23.540 (2.114)	0.039	**0.965** (0.001)	0.143 (0.042)	0.031 (0.014)	**0.000** (0.000)	**0.712** (0.041)	16.908 (0.883)	0.001

	P2020 ($p_0 = 0.037$)							*Full Dependence* ($p_0 = 0.037$)						
	Acc	MCC	TPR	FPR	AUC	Fr Loss	$\hat{p}_0$	Acc	MCC	TPR	FPR	AUC	Fr Loss	$\hat{p}_0$
mGHS_{MPM}	0.632 (0.028)	0.166 (0.017)	**0.804** (0.027)	0.374 (0.029)	0.827 (0.021)	18.770 (2.030)	0.388	0.566 (0.025)	0.094 (0.021)	**0.684** (0.039)	0.438 (0.025)	0.695 (0.035)	17.031 (1.596)	0.446
mGHS_{t^α}	0.901 (0.185)	0.449 (0.133)	0.451 (0.195)	0.082 (0.199)	0.827 (0.021)	18.770 (2.030)	0.095	0.890 (0.203)	0.198 (0.076)	0.192 (0.243)	0.083 (0.220)	0.695 (0.035)	17.031 (1.596)	0.087
GHS_{MPM}	0.711 (0.017)	0.179 (0.016)	0.728 (0.026)	0.290 (0.018)	0.798 (0.017)	20.423 (1.599)	0.314	0.560 (0.026)	0.072 (0.021)	0.632 (0.037)	0.443 (0.027)	0.647 (0.036)	17.037 (1.070)	0.451
fJGL	0.934 (0.011)	0.393 (0.028)	0.589 (0.042)	0.052 (0.012)	0.768 (0.017)	18.076 (0.969)	0.073	0.956 (0.006)	0.254 (0.043)	0.210 (0.079)	0.015 (0.009)	0.597 (0.036)	**15.792** (0.833)	0.022
gJGL	0.921 (0.012)	0.330 (0.024)	0.541 (0.041)	0.064 (0.013)	0.738 (0.017)	19.699 (0.949)	0.082	0.955 (0.006)	0.228 (0.037)	0.187 (0.071)	0.015 (0.009)	0.586 (0.032)	15.526 (0.763)	0.019
$\text{GemBAG}_{\text{MPM}}$	**0.975** (0.002)	**0.551** (0.036)	0.322 (0.049)	**0.000** (0.000)	**0.867** (0.013)	**15.755** (1.217)	0.037	**0.965** (0.001)	**0.267** (0.034)	0.079 (0.018)	**0.000** (0.000)	**0.802** (0.043)	16.268 (0.700)	0.003

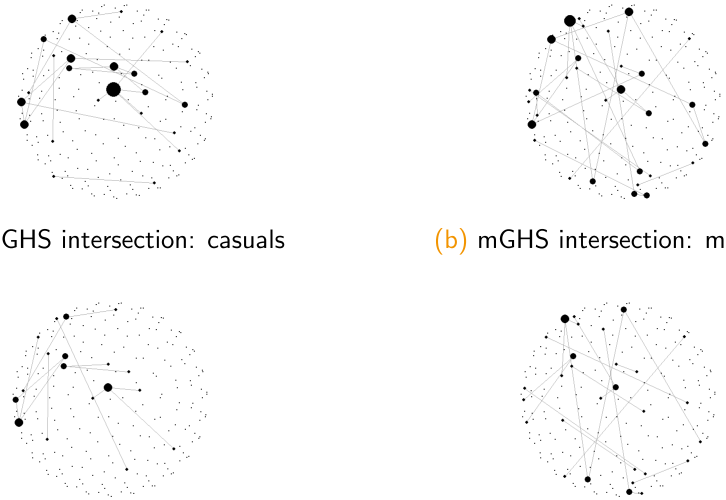
Real data application: a bike-sharing dataset

- Capital Bikeshare system data (Zhu & Foygel Barber, 2015; Yang et al. 2021): contains records of bike rentals in a bicycle sharing system during years 2016, 2017, 2018
- y_{ij}^{c} and y_{ij}^{m} denote the number of registered casual and member trips initiated at station j on day i, respectively
- Data are corrected for seasonal trend, marginally standardized and transformed with the Yeo-Johnson transformation to better approximate a gaussian distribution

Real data application: a bike-sharing dataset

- $K = 6, n = 364, p = 239$
- The method is tested against the ordinary graphical horseshoe applied to each single group separately (Li et al., 2019), the GemBAG (Yang et al., 2021) and the jointGHS (Lingjaerde et al., 2020)
- Prediction performance with *average absolute forecast error* across groups (AAFE; Fan et al., 2009): GemBAG and jointGHS provide the sparsest solutions and the worst predictions (mAAFE $= 0.602$ and mAAFE $= 0.604$, respectively), mGHS and GHS with full graphs are the best predictive methods with mAAFE $= 0.581$ and mAAFE $= 0.586$, respectively

Bike-sharing dataset: intersection of networks across the years



(a) mGHS intersection: casuals

(b) mGHS intersection: members

(c) GHS intersection: casuals

(d) GHS intersection: members

Bike-sharing dataset: estimated correlation matrix

- Estimated correlation matrix **R**:

$$
\begin{array}{cccccc|l}
\text{casual 2016} & \text{casual 2017} & \text{casual 2018} & \text{member 2016} & \text{member 2017} & \text{member 2018} & \\
\end{array}
$$

$$
\left(\begin{array}{cccccc}
1.000 & 0.969 & 0.893 & 0.479 & 0.515 & 0.483 \\
0.969 & 1.000 & 0.958 & 0.518 & 0.562 & 0.526 \\
0.893 & 0.958 & 1.000 & 0.461 & 0.502 & 0.475 \\
0.479 & 0.518 & 0.461 & 1.000 & 0.984 & 0.971 \\
0.515 & 0.562 & 0.502 & 0.984 & 1.000 & 0.980 \\
0.483 & 0.526 & 0.475 & 0.971 & 0.980 & 1.000
\end{array}\right)
\begin{array}{l}
\text{casual 2016} \\
\text{casual 2017} \\
\text{casual 2018} \\
\text{member 2016} \\
\text{member 2017} \\
\text{member 2018}
\end{array}
$$

- Member rides show a more regular behaviour across the years

Real data application: a genomic dataset

- Contains records of $p = 381$ genes' expression in primary CD14$^+$ monocytes after healthy individuals are exposed to different immune conditions (Lingjaerde et al., 2024).
- $K = 4$ groups: $n_{unstim} = 413$ individuals with unstimulated cells, $n_{IFN} = 366$ individuals stimulated with interferon-γ inflammation proxy and $n_{LPS2h} = 260$ and $n_{LPS24h} = 321$ individuals with a duration of 2 and 24 hours of lipopolysaccharide stimulation proxy, respectively.

Genomic dataset: numbers of connections for top hub-genes

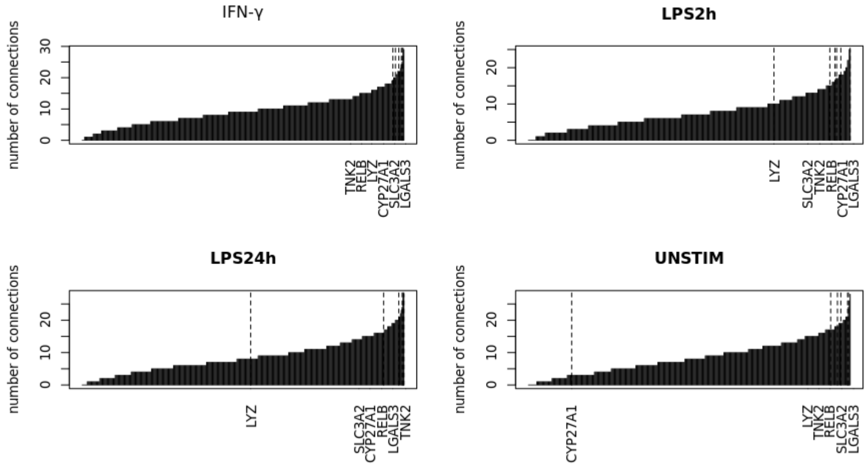


Figure 3: Number of estimated connections for each gene with mGHS in the four conditions.

Selected references

Barbieri M. M. & Berger J. O. (2004). *Optimal predictive model selection*, The Annals of Statistics, 32(3):870-897

Barnard J., McCulloch R. & Meng X. L. (2000). *Modeling covariance matrices in terms of standard deviations and correlations, with application to shrinkage*, Statistics Sinica, 10(4), 1281-1311

Carvalho C. M., Polson N. G. & Scott J. G (2010), *The horseshoe estimator for sparse signals*, Biometrika, 97(2), 465-480

Fan J., Feng Y., & Wu Y. (2009). *Network exploration via the adaptive Lasso and SCAD penalties*, The Annals of Applied Statistics, 3(2):521-541

George E. I. & McCulloch R. E. (1993). *Variable selection via Gibbs sampling*, Journal of the American Statistical Association, 88(423):881-889

Li Y., Craig B. A. & Bhadra A. (2019). *The Graphical Horseshoe estimator for inverse covariance matrices*, Journal of Computational and Graphical Statistics, 28(3), 747-757

Lingjaerde C., Fairfax B. P., Richardson S., and Ruffieux H. (2024). *Scalable multiple network inference with the joint graphical horseshoe*, The Annals of Applied Statistics, 18(3), 1899-1923

Liu X. & Daniels M. (2006). *A new algorithm for simulating a correlation matrix based on parameter expansion and reparameterization*, Journal of Computational and Graphical Statistics, 15, 897-914

Selected references

Makalic E. & Schmidt D. F. (2016). *A simple sampler for the horseshoe estimator*, IEEE Signal Processing Letters, 23(1):179-182

Peterson C. B., Osborne N., Stingo F.C., Bourgeat P., Doecke J. D. & Vannucci M. (2020). *Bayesian modeling of multiple structural connectivity networks during the progression of Alzheimer's disease*, Biometrics, 76(4), 1120-1132

Plummer M. (2015). *Cuts in Bayesian Graphical models*, Statistics and Computing, page 37-43

Wang H. (2015). *Scaling it up: stochastic search structure learning in Graphical models*. Bayesian Analysis, 10(2), 351-377

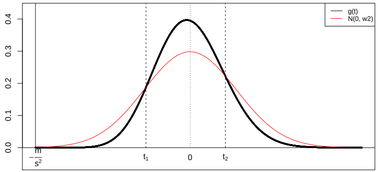
Yang X., Gan L., Narisetty N., & Liang F. (2021). *GemBAG: group estimation of multiple Bayesian Graphical models*, Journal of Machine Learning Research, 22

Zhu Y. & Foygel Barber R. (2015). *The log-shift penalty for adaptive estimation of multiple Gaussian Graphical models*, In Proceedings of the Eighteenth International Conference on Artificial Intelligence and Statistics, pages 1153-1161

Zigler C., Watts K., Yeh R., Wang Y., Coull B., & Dominici F. (2013). *Model feedback in Bayesian propensity score estimation*, Biometrics, 69

Sampling λ and τ

- **Target density**: $f(x) \propto x^{-\frac{c}{2}} e^{-\frac{a^2}{x} + \frac{b}{\sqrt{x}}}$, $c > 0$
- **Solution**: sample variable y from distribution $g(y) \propto y^{c-3} e^{-a^2 y^2 + by}$ and apply the transformation $x = 1/y^2$



- **No efficient hat function available**: standard rejection sampling cannot be applied
- **Modified rejection sampling** (Ahrens & Dieter, 1982):
  - The method draws samples from the actual target distribution
  - Acceptance probability is 1
  - Usually more efficient than the standard rejection sampling

Sampling results

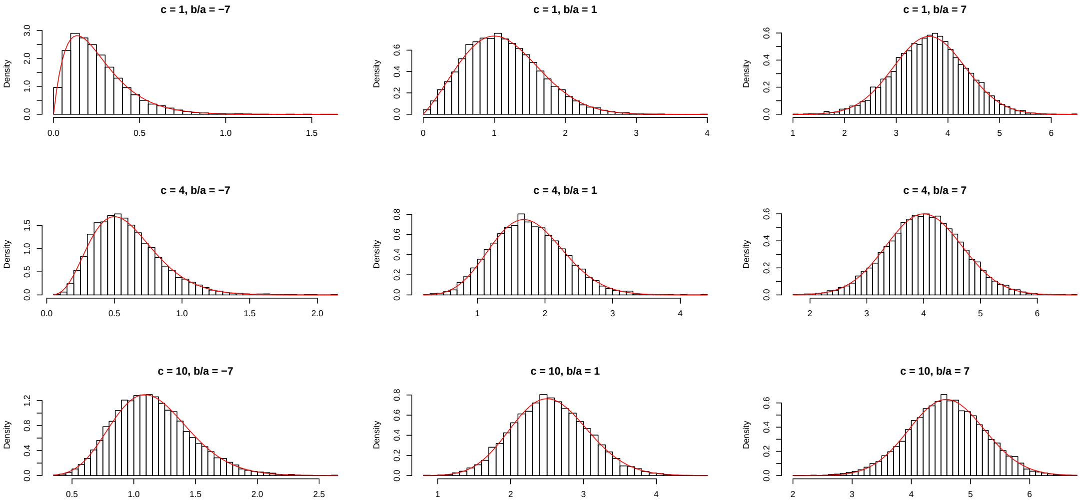


Figure 4: histogram of $n = 10000$ values from $g(y)$ and true density function (red line).