
# Bayesian Shrinkage for Structure Learning

**Jyotishka Datta**

May 2, 2024

Virginia Tech

# Key Idea

- We propose a new **MAP estimation scheme** in Gaussian graphical models under a general class of **completely monotone priors**, including the graphical horseshoe.
- The algorithmic procedure is an **Local Linear Approximation scheme**, and equivalence between LLA and MAP estimates is established by showing the complete monotonicity of the prior density, and then appealing to existing results in Zou and Li [2008].
- The resultant estimate is **naturally sparse**, owing to the **soft thresholding step** in LLA, aiding graph structure learning.

Received: 23 November 2023 | Revised: 8 January 2024 | Accepted: 24 March 2024

DOI: 10.1002/sta4.682

ORIGINAL ARTICLE

WILEY

# Maximum a posteriori estimation in graphical models using local linear approximation

Ksheera Sagar[1] | Jyotishka Datta[2] | Sayantan Banerjee[3] | Anindya Bhadra[1]

[1]Department of Statistics, Purdue University, West Lafayette, Indiana, USA

[2]Department of Statistics, Virginia Tech, Blacksburg, Virginia, USA

[3]Operations Management and Quantitative Techniques Area, Indian Institute of Management Indore, Indore, Madhya Pradesh, India

**Correspondence**
Sayantan Banerjee, Operations Management and Quantitative Techniques Area, Indian Institute of Management Indore, J-212, Academic Block, Prabandh Shikhar, Rau-Pithampur Road, Indore, Madhya Pradesh 453 556, India.
Email: sayantanb@iimidr.ac.in

**Funding information**
U.S. National Science Foundation, Grant/Award Number: DMS-2014371

## Abstract

Sparse structure learning in high-dimensional Gaussian graphical models is an important problem in multivariate statistical inference, since the sparsity pattern naturally encodes the conditional independence relationship among variables. However, maximum a posteriori (MAP) estimation is challenging under hierarchical prior models, and traditional numerical optimization routines or expectation–maximization algorithms are difficult to implement. To this end, our contribution is a novel local linear approximation scheme that circumvents this issue using a very simple computational algorithm. Most importantly, the condition under which our algorithm is guaranteed to converge to the MAP estimate is explicitly stated and is shown to cover a broad class of completely monotone priors, including the graphical horseshoe. Further, the resulting MAP estimate is shown to be sparse and consistent in the $\ell_2$-norm. Numerical results validate the speed, scalability and statistical performance of the proposed method.

**KEYWORDS**
complete monotonicity, graph structure learning, graphical horseshoe prior, precision matrix estimation

# Outline of My Talk

## Global-Local Shrinkage for Sparse and Structured Data

1. Precision matrix estimation
2. Horseshoe-like prior
3. Optimal Properties & Computation.

Part II: New Directions

1. New Laplace Mixture Representation.
2. EM via LLA.
3. Results and Remarks.

# Precision matrix estimation

# Gaussian Graphical Model i

- Gaussian graphical model (GGM): fundamental building block for network estimation because of the ease of interpretation.
- Both Bayesian and frequentist approaches to this, but difficult to obtain good Bayesian and frequentist properties *under the same prior–penalty dual*, complicating justification.
- Goal: fully Bayesian inference on $\boldsymbol{\Omega}$, we need a **suitable sparsity-favoring prior** that also results in a penalty function with good frequentist properties.

# Gaussian Graphical Model ii

- $\mathbf{X}^{(n)} = (\mathbf{X}_1, \ldots, \mathbf{X}_n)^T \sim \mathcal{N}_p(\mathbf{0}, \mathbf{\Sigma})$.
- The corresponding precision matrix: $\mathbf{\Omega} = ((\omega_{ij}))$ is defined as $\mathbf{\Omega} = \mathbf{\Sigma}^{-1}$.
- No edge between $X_i$ and $X_j \leftrightarrow \Omega_{ij} = 0$.
- Assume that $\mathbf{\Omega}$ is sparse: the number of non-zero off-diagonal elements is small.
- We need a suitable prior to adapt to sparsity in $G$.

# Horseshoe Regularization

- Horseshoe prior:

$$\theta_i \mid \lambda_i \sim \mathcal{N}\left(0, \lambda_i^2 \tau^2\right);$$

$$\underbrace{\lambda_i}_{\text{local}} \overset{\text{ind}}{\sim} \mathcal{C}^+(0,1), \underbrace{\tau}_{\text{global}} \sim \mathcal{C}^+(0,\sigma) \quad \text{(Heavy-tailed prior)}$$

# Horseshoe Regularization

- Horseshoe prior:

$$\theta_i \mid \lambda_i \sim \mathcal{N}\left(0, \lambda_i^2 \tau^2\right);$$

$$\underbrace{\lambda_i}_{\text{local}} \overset{\text{ind}}{\sim} \mathcal{C}^+(0,1), \underbrace{\tau}_{\text{global}} \sim \mathcal{C}^+(0,\sigma) \quad \text{(Heavy-tailed prior)}$$

- $p(\theta)$ not analytically tractable !

$$\frac{1}{\tau(2\pi)^{3/2}} \log\left(1+\frac{4\tau^2}{\theta^2}\right) < p_{HS}(\theta \mid \tau) < \frac{2}{\tau(2\pi)^{3/2}} \log\left(1+\frac{2\tau^2}{\theta^2}\right),$$

- Hindrance in learning via EM-type algorithms.

# Horseshoe Regularization

- Horseshoe prior:

$$\theta_i \mid \lambda_i \sim \mathcal{N}\left(0, \lambda_i^2 \tau^2\right);$$

$$\underbrace{\lambda_i}_{\text{local}} \overset{\text{ind}}{\sim} \mathcal{C}^+(0,1), \underbrace{\tau}_{\text{global}} \sim \mathcal{C}^+(0,\sigma) \quad \text{(Heavy-tailed prior)}$$

- $p(\theta)$ not analytically tractable !

$$\frac{1}{\tau(2\pi)^{3/2}} \log\left(1+\frac{4\tau^2}{\theta^2}\right) < p_{HS}(\theta \mid \tau) < \frac{2}{\tau(2\pi)^{3/2}} \log\left(1+\frac{2\tau^2}{\theta^2}\right),$$

- Hindrance in learning via EM-type algorithms.
- Solution: normalize the tight bounds: 'horseshoe-like' [Bhadra et al., 2017].

$$p_{\widetilde{HS}}(\theta \mid a) = \frac{1}{2\pi a^{1/2}} \log\left(1+\frac{a}{\theta^2}\right).$$

# Normal Scale Mixture Representation!

- Frullani's identity [Jeffreys and Swirles, 1972, pp. 406–407]:

$$\int_0^\infty \frac{f(ax) - f(bx)}{x} dx = \{f(0) - f(\infty)\} \log(b/a),$$

- $f(x) = \exp(-x)$ yields a latent variable representation:

$$\frac{1}{2\pi a^{1/2}} \log\left(1 + \frac{a}{\theta^2}\right) = \int_0^\infty \exp\left(-\frac{u\theta^2}{a}\right) \frac{(1 - e^{-u})}{2\pi a^{1/2} u} du$$

- Normal scale mixture:

$$(\theta \mid u, a) \overset{ind}{\sim} \mathcal{N}\left(0, \frac{a}{2u}\right), \; p(u) = \frac{1 - e^{-u}}{2\pi^{1/2} u^{3/2}}$$

- Extended for graphical models [Sagar et al., 2024, EJS].

# Graphical Horseshoe-Like prior

- For the fully Bayesian model, the element-wise prior specification induced by the horseshoe-like prior is:

$$\omega_{ij} \mid a \quad \sim \quad \pi(\omega_{ij} \mid a),\ 1 \leq i < j \leq p; \qquad \omega_{ii} \propto 1,\ 1 \leq i \leq p,$$

  where $\pi(\omega_{ij} \mid a)$ is the density of the horseshoe-like prior.
- The horseshoe-like prior above can be expressed as a Gaussian scale-mixture [Bhadra et al., 2017], thus giving us a GL shrinkage prior:

$$\omega_{ij} \mid \nu_{ij}, a \sim \mathcal{N}\left(0, \frac{a}{2\nu_{ij}}\right),\ \pi(\nu_{ij}) \sim \frac{1 - \exp(-\nu_{ij})}{2\pi^{1/2}\nu_{ij}^{3/2}}. \qquad (1)$$

- Only $\nu_{ij}$ is considered to be latent and the global scale parameter $a$ is considered to be fixed.
- Can estimate $a$ by the effective model size approach of Piironen and Vehtari [2017] to avoid it collapsing to zero.

# MAP Estimation

- Gaussian scale mixture representation $\Rightarrow$ Expectation Conditional Maximization (ECM) for MAP estimation.
- For updating the elements of the precision matrix, uses the **coordinate descent technique** of Wang [2014].
- Posterior contraction rate of the precision matrix $\boldsymbol{\Omega}$ around the true precision matrix $\boldsymbol{\Omega}_0$ with respect to the Frobenius norm under the graphical horseshoe-like prior.
- We make certain assumptions on the true precision matrix, the dimension and sparsity, and the prior space.
- We can also show that the MAP estimator of $\boldsymbol{\Omega}$, given by $\hat{\boldsymbol{\Omega}}^{\text{MAP}}$ is consistent, in the sense that

$$\|\hat{\boldsymbol{\Omega}}^{\text{MAP}} - \boldsymbol{\Omega}_0\|_2 = O_P(\epsilon_n),$$

where $\epsilon_n = n^{-1/2}(p+s)^{1/2}(\log p)^{1/2}$ is the same posterior convergence rate as the full posterior.

▸ Posterior concentration rate

# Graphical HSlike

- Graphical Horseshoe-Like is tight approximation of the Graphical Horseshoe.
- Key advantage: Gaussian scale mixture: allows for MCMC + EM and LLA algorithms.
- Can be interpreted as non-convex penalty (horseshoe-like).

# Graphical HSlike

- Graphical Horseshoe-Like is tight approximation of the Graphical Horseshoe.
- Key advantage: Gaussian scale mixture: allows for MCMC + EM and LLA algorithms.
- Can be interpreted as non-convex penalty (horseshoe-like).
- Now we'll show: **it is possible to obtain MAP estimate under the original graphical horseshoe model** as well.

# Graphical HSlike

- Graphical Horseshoe-Like is tight approximation of the Graphical Horseshoe.
- Key advantage: Gaussian scale mixture: allows for MCMC + EM and LLA algorithms.
- Can be interpreted as non-convex penalty (horseshoe-like).
- Now we'll show: **it is possible to obtain MAP estimate under the original graphical horseshoe model** as well.
- We bypass the **unavailability of the marginal model** by taking an indirect route by first establishing the complete monotonicity of $p(\omega_{ij})$ and using this to establish the **equivalence between LLA and MAP estimates**.

# Remark about EM

- The difficulty in implementing the usual EM algorithm for the horseshoe or its graphical version comes from the need to treat the **local shrinkage parameters as latent variables, and not as tuning parameters** that could be plugged in with their conditional modes.
- The latter approach evaluates $\text{argmax}_\Omega\, p(\Omega \mid Y, \dagger)$ where $\dagger$ denotes the local modes for the other shrinkage parameters ($\lambda_{ij}$s in case of horseshoe), but not $\text{argmax}_\Omega\, p(\Omega \mid Y)$, which is what we seek.
- An LLA approach also has the advantage of being portable to many other shrinkage priors with the same challenges: a difficult EM due to multiple levels of hierarchy of the local shrinkage parameters but a straightforward LLA.
- Note that the availability of full conditionals is sufficient for a Gibbs sampler or iterated conditional modes of Besag [1986], but not for EM.

# MAP Estimation using Local Linear Approximation

# LLA

- Introduce novel technique for MAP estimation in graphical models using the local linear approximation (LLA) algorithm Zou and Li [2008]
- Explicitly identify the conditions under which the LLA is guaranteed to find the MAP estimate.
- Complete monotonicity of the prior density on the elements of $\Omega$, which includes a few well-known priors:
  1. Graphical horseshoe Li et al. [2019]
  2. Laplace scale mixture of gamma Garrigues and Olshausen [2010],
  3. Special cases of power exponential scale mixtures Giri and Rao [2016], Zhang et al. [2012] and
  4. Horseshoe-like [Bhadra et al., 2021].

# Horseshoe is a Laplace scale mixture

♣ Sagar and Bhadra [2022] established the following properties of the horseshoe density:

1. In addition to a normal scale mixture, the horseshoe density can also be expressed as a Laplace scale mixture, which, in turn, establishes that the density is completely monotone (via Bernstein—Widder theorem [Widder, 1946]).
2. Using a result of Bochner [1955] yields that the induced penalty function (negative of logarithm of the horseshoe density) admits a completely monotone derivative.
3. Combining 1), 2) above and a result from Zou and Li [2008] for completely monotone densities, LLA is equivalent to EM, for a sparse MAP estimation under the horseshoe.

# Key results from [Sagar and Bhadra, 2022]

**Proposition**

*The marginal horseshoe density for a scalar random variable admits the following representation as a Laplace mixture:*

$$p_{HS}(x) = \frac{2}{\pi\sqrt{\pi\tau}} \int_0^\infty \exp\left(-u\frac{|x|}{\tau}\right) D_+\left(\frac{u}{\sqrt{2}}\right) du,$$

*where* $D_+(z) = \exp(z^2) \int_0^z \exp(t^2)\, dt$, $z > 0$ *is the Dawson function.*

**Proposition**

*The first and second derivatives of the penalty function induced by the horseshoe density,* $pen(|x|) = -\log p_{HS}(|x|)$, *are bounded.*

# Consequence of Laplace Mixture i

- Propositions 1 and 2 imply that

$$\text{pen}(\Omega) = \sum_{i \neq j} \text{pen}(|\omega_{ij}|) = -\sum_{i \neq j} \log p_{HS}(|\omega_{ij}|),$$

  is strongly concave and has a completely monotone derivative ... which implies that EM and LLA are equivalent.

- First order Taylor approximations of $\text{pen}(|\omega_{ij}|)$:

$$\text{pen}(\Omega) \approx \sum_{i \neq j} \text{pen}\left(|\omega_{ij}^{(t)}|\right) + \sum_{i \neq j} \text{pen}'\left(|\omega_{ij}^{(t)}|\right) (|\omega_{ij}| - |\omega_{ij}^{(t)}|),$$

- Thus, the optimization problem reduces to:

$$\Omega^{(t+1)} = \underset{\Omega}{\text{argmin}} \left( \ell(\Omega;\, Y) + \sum_{i \neq j} \text{pen}'\left(|\omega_{ij}^{(t)}|\right) |\omega_{ij}| \right). \qquad (2)$$

# Consequence of Laplace Mixture ii

- Solving for $\Omega^{(t+1)}$ is analogous to solving for $\Omega^{(t+1)}$ in a graphical lasso problem Friedman et al. [2008].
- For simplicity, let $G_{ij}^{(t)} = \text{pen}'\left(|\omega_{ij}^{(t)}|\right)$. Using Proposition 1:

$$G_{ij}^{(t)} = \frac{\int_0^\infty \frac{u}{\tau} \exp\left(-\frac{|\omega_{ij}^{(t)}|}{\tau} u\right) D_+\left(\frac{u}{\sqrt{2}}\right) du}{\int_0^\infty \exp\left(-\frac{|\omega_{ij}^{(t)}|}{\tau} u\right) D_+\left(\frac{u}{\sqrt{2}}\right) du}. \tag{3}$$

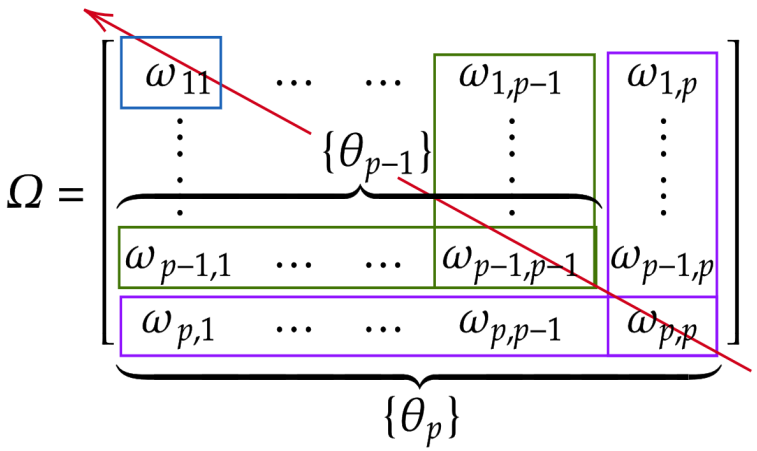
- $G_{ij}^{(t)}$ acts as the tuning parameter for the graphical lasso problem.
- The ratio of integrals in $G_{ij}^{(t)}$, in (3) can be computed in a computationally efficiently manner as simple Riemann sums using the rational approximation of the Dawson function Lether [1997] and having the values of $D_+(u)$ pre-computed and stored on a large enough and fine grid of $u$.

# Coordinate Descent Approach [Wang, 2014] i

- Apply the decomposition:

$$\boldsymbol{\Omega}_{p\times p} = \begin{bmatrix} \boldsymbol{\Omega}_{(p-1)\times(p-1)} & \boldsymbol{\omega}_{\bullet,p} \\ \boldsymbol{\omega}_{\bullet,p}^{T} & \omega_{pp} \end{bmatrix}.$$

- Let $\boldsymbol{\theta}_p = (\boldsymbol{\omega}_{\bullet,p}\,,\, \omega_{pp})$ be a vector of length $p$ denoting the last column of $\boldsymbol{\Omega}_{p\times p}$ and $z$ be collection of all latent variables.




- where $\Omega_{11}, s_{11}$ are $(q-1)\times(q-1)$ matrices comprising the first $(q-1)$ rows and columns of $\Omega, S$ respectively.

# Coordinate Descent Approach [Wang, 2014] ii

- Similarly, $\Omega_{12}, s_{12}$ are $(q-1)$ vectors comprising the off-diagonals in the $q^{th}$ columns of $\Omega, nS$; and $\Omega_{22} = \omega_{qq}$, a scalar.
- Define:

$$\gamma = \Omega_{22} - \Omega_{12}^T \Omega_{11}^{-1} \Omega_{12}, \quad \beta = \Omega_{12}.$$

- Then each term in Equation (2) can be simplified as follows:

$$\begin{aligned} \log|\Omega| &= \log(\gamma) + c_1, \\ \text{tr}(nS\Omega) &= 2s_{12}^T\beta + s_{22}\gamma + s_{22}\beta^T\Omega_{11}^{-1}\beta + c_2, \\ \sum_{i\neq j} G_{ij}^{(t)}|\omega_{ij}| &= 2G_{\bullet q}^{(t)}|\beta| + c_3, \end{aligned}$$

where $G_{\bullet q}^{(t)} = (G_{1p}^{(t)}, \ldots, G_{q-1,\,q}^{(t)})$, $\beta = (\omega_{1q}, \ldots, \omega_{q-1,\,q})^T$, $|\beta| = (|\omega_{1q}|, \ldots, |\omega_{q-1,q}|)^T$ and $c_1, c_2, c_3$ are independent of $\beta$ and $\gamma$.

# Coordinate Descent Approach [Wang, 2014] iii

- With these simplifications, the objective function (2) in terms of $\beta$, $\gamma$ can be expressed as:

$$-\frac{n}{2}\log(\gamma)+\frac{1}{2}\left[2s_{12}^{T}\beta+s_{22}\gamma+s_{22}\beta^{T}\left(\Omega_{11}^{(t)}\right)^{-1}\beta\right]+2G_{\bullet q}^{(t)}|\beta|. \tag{4}$$

- Minimizing the above function with respect to the entries of $\beta$ is analogous to solving a lasso problem.
- Cannot update all entries of $\beta$ at once, we update each entry one at a time by conditional minimization.

# Consistency of the MAP estimator

**Theorem**

*The MAP estimator of $\Omega$, denoted by $\hat{\Omega}^{\text{MAP}}$ is consistent, in the sense that,*

$$\|\hat{\Omega}^{\text{MAP}} - \Omega_0\|_2 = O_P(\epsilon_n),$$

*where $\Omega_0$ is the true precision matrix, $\epsilon_n = n^{-1/2}(q+s)^{1/2}(\log q)^{1/2}$ is the posterior convergence rate and $s$ is the number of non-zero off-diagonal elements in $\Omega_0$.*

# Numerical results

# Numerical results

- LLA of graphical horseshoe penalty under Laplace (LLA (l)) and half-Cauchy (LLA (c)) mixtures,
- full Bayes MCMC estimate GHS prior Li et al. [2019]
- frequentist graphical lasso Friedman et al. [2008], with penalized (GL1) and unpenalized (GL2) diagonal elements.
- Two problem dimensions $(n, q) = \{(120, 100), (120, 200)\}$ and
- Two structures of precision matrix: 'hubs', 'random'

# Hubs structure

**Table 1:** Comparison of results for competing procedures. $n = 120$, $q = 100$ and Hubs structure.

| | LLA (l) | LLA (c) | GHS | GL1 | GL2 |
|---|---|---|---|---|---|
| Stein's loss | 3.873 | 3.898 | 5.1 | 5.255 | 6.328 |
| | (0.379) | (0.389) | (0.454) | (0.263) | (0.414) |
| F norm | 2.235 | 2.246 | 2.547 | 3.018 | 3.432 |
| | (0.099) | (0.098) | (0.128) | (0.091) | (0.112) |
| TPR | 0.967 | 0.967 | 0.871 | 0.995 | 0.986 |
| | (0.024) | (0.025) | (0.04) | (0.007) | (0.017) |
| FPR | 0.032 | 0.032 | 0.003 | 0.101 | 0.045 |
| | (0.012) | (0.012) | (0.001) | (0.016) | (0.008) |
| MCC | 0.592 | 0.593 | 0.848 | 0.373 | 0.523 |
| | (0.063) | (0.064) | (0.028) | (0.027) | (0.0391) |
| Time (s) | 3.04 | 3.09 | 250.12 | 1.59 | 1.65 |

# Random Structure

**Table 2:** Comparison of results for competing procedures. $n = 120$, $q = 100$ and Random structure.

| | LLA (l) | LLA (c) | GHS | GL1 | GL2 |
|---|---|---|---|---|---|
| Stein's loss | 2.432 | 2.428 | 2.173 | 5.245 | 6.785 |
| | (0.296) | (0.288) | (0.28) | (0.254) | (0.464) |
| F norm | 1.997 | 1.981 | 1.961 | 3.348 | 4.084 |
| | (0.132) | (0.132) | (0.144) | (0.115) | (0.143) |
| TPR | 0.874 | 0.881 | 0.82 | 0.951 | 0.882 |
| | (0.05) | (0.049) | (0.043) | (0.03) | (0.038) |
| FPR | 0.019 | 0.021 | 0.0005 | 0.101 | 0.045 |
| | (0.009) | (0.009) | (0.0003) | (0.013) | (0.007) |
| MCC | 0.481 | 0.464 | 0.868 | 0.232 | 0.321 |
| | (0.078) | (0.068) | (0.032) | (0.018) | (0.024) |
| Time (s) | 6.53 | 6.56 | 253.41 | 4.14 | 4.61 |

# Analysis of TCGA Proteomics Data

# Analysis of TCGA Proteomics Data i

- Proteomics data from the The Cancer Genome Atlas (TCGA) project streamlined and pre-processed by Ha et al. [2018].
- Expression levels of $q = 50$ proteins for $n = 250$ patients with lung squamous cell carcinoma.
- Split the data into $Y_{\text{train}}$ and $Y_{\text{test}}$, then estimate $\hat{\Omega}^{\text{MAP}}$ using $Y_{\text{train}}$ for all the competing procedures and also estimate the posterior mean of samples of $\Omega$ in the case of fully Bayesian graphical horseshoe (GHS).
- Compare partial prediction loss on the test data $Y_{\text{test}}$, defined as:

$$\{\sum_{i=1}^{p} ||Y_j + \sum_{j \in \{1,\ldots,q\} \setminus i} Y_k \hat{\omega}_{ji}/\hat{\omega}_{ii}||^2\}^{1/2}$$

and sparsity *i.e.*, proportion of zeros in the lower or upper triangle of the precision matrix estimate.

# Analysis of TCGA Proteomics Data ii

**Table 3:** Comparison of average (sd) of prediction norm on $Y_{\text{test}}$, and estimates of sparsity for the competing procedures. Best performers in bold.

| | **LLA (I)** | **LLA (c)** | **GHS** | **GL1** | **GL2** |
|---|---|---|---|---|---|
| Loss | 41.98 (0.66) | 41.88 (0.65) | **33.63** (0.734) | 73.09 (4.58) | 71.62 (4.03) |
| Sparsity | **0.81** (0.003) | 0.81 (0.01) | 0.69 (0.01) | 0.40 (0.02) | 0.46 (0.03) |
| Time (s) | 8.69 | 9.53 | 68.45 | 4.78 | 5.2 |

- MCMC-based GHS has the lowest average prediction loss.
- Estimates obtained by local linear approximation of the graphical horseshoe: LLA (I), LLA(c), also perform well compared to the graphical lasso procedures.
- Moreover, the estimates LLA (I) and LLA(c) are much sparser compared to the alternatives.
- Note that GHS does not produce exact zeroes, and we use credible intervals.

# Summary of Numerical Results

- Estimates LLA (I) and LLA (c) perform similarly.
- In general, they also have the lowest or comparable Stein's loss, F norm and the second best FPR, MCC among the competing procedures.
- It can also be seen that the LLA based methods are computationally several orders of magnitude faster than fully Bayesian MCMC of Li et al. [2019], and
- LLAs result in better statistical estimates compared to the graphical lasso.

# Summary and Scopes

- We present an LLA method for precision matrix estimation with the graphical horseshoe.
- We use a Laplace or a Cauchy mixture instead of a Gaussian scale mixture.
- Can port this to other hierarchical shrinkage priors.
- Scopes:
  1. Establish multiple testing optimality.
  2. Multiple graphical models.
  3. Extend beyond Gaussian set-up, e.g. functional.

# References

# References (for this talk)


- "**Maximum a Posteriori Estimation in Graphical Models Using Local Linear Approximation.**" Sagar, K., Datta, J., Banerjee, S., & Bhadra, A. (2024). *STAT*.
- **Graphical horseshoe-like prior:** Sagar, Ksheera, Banerjee, S., **Datta, J.**, and Bhadra, A.. (2024). "Precision Matrix Estimation under the Horseshoe-like Prior-Penalty Dual." *Electronic Journal of Statistics*.
- **Graphical evidence.** Bhadra, A., Sagar, K., Banerjee, S., & Datta, J. (2022). *Preprint.* arXiv preprint arXiv:2205.01016.
- **Graphical horseshoe:** 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.


**Matlab codes:**
**https://github.com/sagarknk/GHS-LLA-codes**

# References (General global-Local)


- Bhadra, A., **Datta, J.**, Li, Y., Polson, N. G., & Willard, B. (2019). Prediction risk for global-local shrinkage regression. **20 (78)**, 1-39, Journal of Machine Learning Research. arXiv:1605.04796.
- Bhadra, A., **Datta, J.**, Polson, N. G., & Willard, B. T. (2019). Lasso Meets Horseshoe: A Survey. **34(3)**, 405-427. Statistical Science.
- Bhadra, **Datta**, Li and Polson (2019). "Horseshoe Regularization for Machine Learning in Complex and Deep Models". *Published*, *International Statistical Review. Discussed paper* [preprint].
- Bhadra, **Datta**, Polson, and Willard (2019), (*alphabetical), "Global-local mixtures - A Unifying Framework".*Accepted*, *Sankhya A*.
- Bhadra, A., **Datta, J.**, Polson, N. G., & Willard, B. (2017). The horseshoe+ estimator of ultra-sparse signals. Bayesian Analysis, 12(4), 1105-1131.
- **Datta, J.**, & Dunson, D. B. (2016). Bayesian inference on quasi-sparse count data. Biometrika, 103(4), 971-983.
- Bhadra, A., **Datta, J.**, Polson, N. G., & Willard, B. (2016). Default Bayesian analysis with global-local shrinkage priors. Biometrika, 103(4), 955-969.
- **Datta, J.**, & Ghosh, J. K. (2013). Asymptotic properties of Bayes risk for the horseshoe prior. Bayesian Analysis, 8(1), 111-132.
- Li, **Datta**, Craig, and Bhadra, (2020+). "Joint Mean-Covariance Estimation via the Horseshoe with an Application in Genomic Data Analysis". *submitted*. [preprint].

# Thank you!

# Supplementary: Graphical HS

# Estimation procedure

Expectation-Conditional Maximization (ECM) algorithm for estimating the posterior mode (MAP estimator).

- **E Step:** The conditional expectation of the latent variable $\nu_{ij}, 1 \leq i < j \leq p$, at current iteration $(t)$ is:

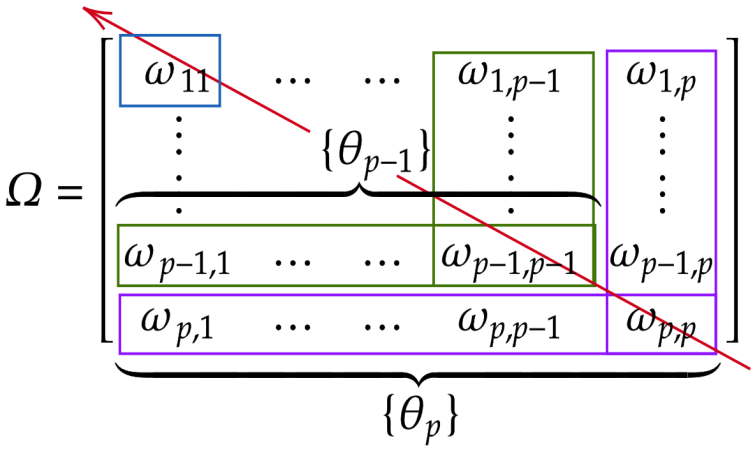
$$\nu_{ij}^{(t)} = \mathbb{E}(\nu_{ij} \mid \omega_{ij}^{(t)}, a) = \left( \log \left( 1 + \frac{a}{(\omega_{ij}^{(t)})^2} \right) \right)^{-1} \frac{a^2}{\left( (\omega_{ij}^{(t)})^2 + a \right) \left( (\omega_{ij}^{(t)})^2 \right)}.$$

# Estimation procedure: CM Step

- We use **coordinate descent** for the conditional maximization step.
- **CM Step:** First we divide the precision matrix $\boldsymbol{\Omega}$ and the sample covariance matrix $\boldsymbol{S}$ into blocks as follows:

$$\boldsymbol{\Omega} = \begin{bmatrix} \boldsymbol{\Omega}_{11} & \boldsymbol{\Omega}_{12} \\ \boldsymbol{\Omega}_{12}^T & \Omega_{22} \end{bmatrix}, \quad \boldsymbol{S} = \begin{bmatrix} \boldsymbol{S}_{11} & \boldsymbol{S}_{12} \\ \boldsymbol{S}_{12}^T & S_{22} \end{bmatrix}.$$

- Define $\gamma = \Omega_{22} - \boldsymbol{\Omega}_{12}^T \boldsymbol{\Omega}_{11}^{-1} \boldsymbol{\Omega}_{12}$ and $\boldsymbol{\beta} = \boldsymbol{\Omega}_{12}$.



# Estimation procedure: CM Step

- Updating the $p^{th}$ column:

$$\log|\boldsymbol{\Omega}| = \log(\gamma) + c_1,$$
$$\text{tr}(\boldsymbol{S}\boldsymbol{\Omega}) = 2\boldsymbol{S}_{12}^T\boldsymbol{\beta} + S_{22}\gamma + S_{22}\boldsymbol{\beta}^T\boldsymbol{\Omega}_{11}^{-1}\boldsymbol{\beta} + c_2,$$
$$\sum_{i,j:i<j} \frac{\nu_{ij}^{(t)}}{a} \cdot \omega_{ij}^2 = \beta^T\boldsymbol{\Lambda}^{(t)}\boldsymbol{\beta} + c_3,$$
$$\boldsymbol{\Lambda}^{(t)} = \frac{1}{a}\text{diag}\left(\nu_{1p}^{(t)}, \nu_{2p}^{(t)}, \ldots, \nu_{p-1,p}^{(t)}\right),$$

- $c_1,\ c_2,\ c_3$ are constants independent of $\boldsymbol{\beta},\ \gamma$.
- The log-posterior with the transformed variables is given by:

$$\mathcal{L} \propto \frac{n}{2}\log(\gamma) - \frac{1}{2}\left(2\boldsymbol{S}_{12}^T\boldsymbol{\beta} + S_{22}\gamma + S_{22}\boldsymbol{\beta}^T\boldsymbol{\Omega}_{11}^{-1}\boldsymbol{\beta}\right) - \boldsymbol{\beta}^T\boldsymbol{\Lambda}^{(t)}\boldsymbol{\beta}.$$

# Estimation procedure: CM Step

- Maximizing the above over $\boldsymbol{\beta}, \gamma$ gives the required update as:

$$\hat{\gamma} = \frac{n}{S_{22}}, \quad \hat{\boldsymbol{\beta}} = -(S_{22}\boldsymbol{\Omega}_{11}^{-1} + 2 \cdot \boldsymbol{\Lambda}^{(t)})^{-1}\boldsymbol{S}_{12}^{T}.$$

- The $p$th column update of the precision matrix for the next iteration $(t+1)$ becomes:

$$\hat{\boldsymbol{\Omega}}_{12}^{(t+1)} = \hat{\boldsymbol{\beta}}, \quad \hat{\boldsymbol{\Omega}}_{12}^{T(t+1)} = \hat{\boldsymbol{\beta}}^{T}, \quad \hat{\Omega}_{22}^{(t+1)} = \hat{\gamma} + \hat{\boldsymbol{\beta}}^{T}\boldsymbol{\Omega}_{11}^{-1}\hat{\boldsymbol{\beta}}$$

- We repeat the above steps for the remaining $(p-1)$ columns to complete the CM Step updates for $\boldsymbol{\Omega}$, until convergence to the MAP estimator $\hat{\boldsymbol{\Omega}}^{\mathrm{MAP}}$.

# Posterior sampling

- Posterior sampling strategy: [Bhadra et al., 2017] + [Li et al., 2017].
- Reparametrize: $2\nu_{ij} \mapsto t_{ij}^2$ and $a \mapsto \tau^2$:

$$\omega_{ij} \mid \nu_{ij}, \tau \sim \mathcal{N}\left(0, \tau^2 / t_{ij}^2\right), \ \pi(t_{ij}) = \frac{1 - \exp\left(-t_{ij}^2/2\right)}{(2\pi)^{1/2} t_{ij}^2}, \ t_{ij} \in \mathbb{R}.$$

$\pi(t_{ij})$ is the **slash normal density** $(\phi(0) - \phi(t))/t^2$.

# Posterior sampling

- Posterior sampling strategy: [Bhadra et al., 2017] + [Li et al., 2017].
- Reparametrize: $2\nu_{ij} \mapsto t_{ij}^2$ and $a \mapsto \tau^2$:

$$\omega_{ij} \mid \nu_{ij}, \tau \sim \mathcal{N}\left(0, \tau^2 / t_{ij}^2\right), \ \pi(t_{ij}) = \frac{1 - \exp\left(-t_{ij}^2/2\right)}{(2\pi)^{1/2} t_{ij}^2}, \ t_{ij} \in \mathbb{R}.$$

  $\pi(t_{ij})$ is the **slash normal density** $(\phi(0) - \phi(t))/t^2$.
- Slash itself is a normal scale mixture with Pareto mixing distribution:

$$t_{ij} \mid r_{ij} \sim \mathcal{N}(0, r_{ij}), \ r_{ij} \sim \text{Pareto}\,(1/2)\,.$$

- Blocked Gibbs sampler $\sim$ Graphical horseshoe sampler of Li et al. [2017].

# Theoretical guarantees

**Assumption**

*The actual dimension $p$ satisfies the condition $p = n^b$, $b \in (0, 1)$, and the effective dimension $p + s$ satisfies $(p + s) \log p / n = o(1)$.*

**Assumption**

*The true precision matrix $\mathbf{\Omega}_0$ belongs to the parameter space given by*

$$\begin{aligned} \mathcal{U}(\varepsilon_0, s) \quad = \quad & \{\mathbf{\Omega} \in \mathcal{M}_p^+ : \sum_{1 \le i < j \le p} \mathbb{1}(\omega_{ij} \neq 0) \le s, \\ & 0 < \varepsilon_0^{-1} \le \mathrm{eig}_1(\mathbf{\Omega}_0) \le \cdots \le \mathrm{eig}_p(\mathbf{\Omega}_0) \le \varepsilon_0 < \infty\}. \end{aligned}$$

Back to MAP

# Theoretical guarantees

**Assumption**

*The bound* $[L^{-1}, L]$ *on the eigenvalues of* $\mathbf{\Omega}$ *satisfies* $L > \varepsilon_0$*, or, in other words,* $\varepsilon_0 = cL$*, for some* $c \in (0, 1)$.

**Assumption**

*The global shrinkage parameter a satisfies the condition,*
$a^{1/2} < n^{-1/2} p^{-b_1}$*, for some constant* $b_1 > 0$.

# Posterior Convergence Rate

**Theorem**
*Let* $\mathbf{X}^{(n)} = (\mathbf{X}_1, \ldots, \mathbf{X}_n)^T$ *be a random sample from a p-dimensional normal distribution with mean* $\mathbf{0}$ *and covariance matrix* $\mathbf{\Sigma}_0 = \mathbf{\Omega}_0^{-1}$, *where* $\mathbf{\Omega}_0 \in \mathcal{U}(\varepsilon_0, s)$. *Consider the Graphical Horseshoe-like prior specification. Under the assumptions on the prior as described above, the posterior distribution of* $\mathbf{\Omega}$ *satisfies*

$$\mathbb{E}_0 \left[ P\{ \|\mathbf{\Omega} - \mathbf{\Omega}_0\|_2 > M\epsilon_n \mid \mathbf{X}^{(n)} \} \right] \to 0,$$

*for* $\epsilon_n = n^{-1/2}(p+s)^{1/2}(\log p)^{1/2}$ *and a sufficiently large constant* $M > 0$.

# MAP estimator

- We can prove that the extended real-valued penalty function $pen_a(x) = -\log\log(1 + a/x^2)$, $a > 0$, is strongly concave, and hence strictly concave, for all $x \in dom(pen_a)$, separately for $x > 0$ and $x < 0$.
- Strict concavity of penalty function guarantees that the LLA algorithm will satisfy an ascent property, that is, $Q(\mathbf{\Omega}^{(t+1)}) > Q(\mathbf{\Omega}^{(t)})$.

**Theorem**

*Under the conditions of Theorem 8, the MAP estimator of* $\mathbf{\Omega}$*, given by* $\hat{\mathbf{\Omega}}^{\text{MAP}}$ *is consistent, in the sense that*

$$\|\hat{\mathbf{\Omega}}^{\text{MAP}} - \mathbf{\Omega}_0\|_2 = O_P(\epsilon_n),$$

*where* $\epsilon_n$ *is the posterior convergence rate as defined in Theorem 8.*

- Converges to the true precision matrix $\mathbf{\Omega}_0$ at the same rate as the posterior convergence rate in the Frobenius norm.

# Simulation: selected

*Hubs.* The rows/columns are partitioned into $K$ disjoint groups $G_1, \ldots, G_K$. The off-diagonal entries $\omega_{ij}^0$ are set to 0.25 if $i \neq j$ and $i, j \in G_k$ for $k = 1, \ldots, K$. In our simulations we consider $p/10$ groups with equal number of elements in each group.

**Table 4:** 50 data sets generated with precision matrix $\mathbf{\Omega}_0$, where $n = 120$ and $p = 100$. Candidates: frequentist graphical lasso with penalized diagonal elements (GL1) and with unpenalized diagonal elements (GL2), graphical SCAD (GSCAD), Bayesian graphical lasso (BGL), the graphical horseshoe (GHS), graphical horseshoe-like ECM (ECM) and graphical horseshoe-like MCMC (MCMC).

| | Hubs<br>90 nonzero pairs out of 4950<br>nonzero elements = 0.25 | | | | | | |
|---|---|---|---|---|---|---|---|
| | GL1 | GL2 | GSCAD | BGL | GHS | ECM | MCMC |
| Stein's loss | 5.255 | 6.328 | 5.213 | 43.042 | 5.101 | **4.22** | 5.310 |
| | (0.263) | (0.414) | (0.261) | (0.802) | (0.455) | (0.369) | (0.485) |
| F norm | 3.018 | 3.432 | 3.003 | 4.295 | 2.544 | **2.415** | 2.687 |
| | (0.091) | (0.112) | (0.093) | (0.156) | (0.126) | (0.103) | (0.141) |
| TPR | .995 | .986 | **.998** | .995 | .872 | 0.985 | 0.754 |
| | (.007) | (.017) | (.002) | (.008) | (.04) | (.014) | (0.004) |
| FPR | .101 | .045 | .983 | .186 | **.003** | .062 | **0.003** |
| | (.016) | (.008) | (.012) | (.007) | (.001) | (0.005) | (0.001) |
| MCC | 0.373 | 0.523 | 0.016 | 0.27 | **0.85** | 0.458 | 0.775 |
| | (.027) | (.039) | (.006) | (.006) | (.027) | (.015) | (.033) |

# RPPA Data

Reverse Phase Protein Array (RPPA) data of 33 patients with lymphoid neoplasm "Diffuse Large B-cell Lymphoma" to infer the protein interaction network. The data set consists of protein expressions for 67 genes across 12 pathways for all patients.

**Table 5:** Percentage of zeros (% Sparsity) and number of non-zero entries (NNZ) in the lower triangle of the precision matrix estimate of RPPA data for the competing approaches.

| | MCMC | ECM | GHS | BGL | GL1 | GL2 | GSCAD |
|---|---|---|---|---|---|---|---|
| % Sparsity | 95.79 | 88.6 | 91.59 | 73.72 | 69.88 | 73.67 | $9.06 \times 10^{-4}$ |
| NNZ | 93 | 252 | 186 | 581 | 666 | 582 | 2209 |

# RPPA Data

We note that the GHS-LIKE-MCMC gives the sparsest estimate, almost 4% sparser than the GHS. This is consistent with prior studies that found robust gene networks are typically sparse [Leclerc, 2008]. As in the simulations, GSCAD performs the worst.

**Table 6:** Percentage of zeros (% Sparsity) and number of non-zero entries (NNZ) in the lower triangle of the precision matrix estimate of RPPA data for the competing approaches.

| | MCMC | ECM | GHS | BGL | GL1 | GL2 | GSCAD |
|---|---|---|---|---|---|---|---|
| % Sparsity | 95.79 | 88.6 | 91.59 | 73.72 | 69.88 | 73.67 | $9.06 \times 10^{-4}$ |
| NNZ | 93 | 252 | 186 | 581 | 666 | 582 | 2209 |

# PRECISE comparison

To compare with a prior analysis of the same data set, we use the PRECISE framework of Ha et al. [2018]. This method can infer directed edges, but we ignore the directionality since we are interested in interactions and not causation. The proportions of edges in the estimates that 'agree' and 'do not agree' with the edges inferred using PRECISE framework are presented below:

**Table 7:** Proportion of edges that 'agree' (AE) and 'do not agree' (NE) with the edges inferred using the PRECISE framework.

| | MCMC | ECM | GHS | BGL | GL1 | GL2 | GSCAD |
|---|---|---|---|---|---|---|---|
| AE | 0.238 | 0.412 | 0.325 | 0.575 | 0.638 | 0.6 | 1 |
| NE | 0.034 | 0.101 | 0.074 | 0.247 | 0.284 | 0.247 | 0.984 |

# Resources for Horseshoe Prior

# Learning $\tau$

1. Maximum marginal likelihood estimator (MMLE)
2. Full Bayes estimator: half-Cauchy prior truncated to the interval $[1/n, 1]$.
3. Cross-validation.
4. By studying the prior for $m_{\text{eff}} = \sum_{i=1}^{n}(1-\kappa_i)$ [Piironen and Vehtari, 2016]

- MMLE beats simple thresholding:

$$\hat{\tau}_s(c_1, c_2) = \max\left\{\frac{\sum_{i=1}^{n}\mathbf{1}\{|y_i| \geq \sqrt{c_1 \log(n)}\}}{c_2 n}, \frac{1}{n}\right\} .$$

- Empirical Bayes estimate of $\tau$ can replace a full Bayes estimate of $\tau$.
- Caution to prevent the estimator from getting too close to zero.

# Computation for Horseshoe

1. MCMC : block-updating $\boldsymbol{\theta}$, $\boldsymbol{\lambda}$ and $\tau$ using either a Gibbs or parameter expansion or slice sampling strategy.
2. Makalic and Schmidt [2016]: Inverse-gamma scale mixture for Gibbs sampling scheme for horseshoe and horseshoe+ prior for linear regression and logistic and negative binomial regression.
3. Hahn et al. [2016]: Elliptical slice sampler – wins over Gibbs strategies!
4. Bhattacharya et al. [2016]: Gaussian sampling alternative to the naïve Cholesky decomposition to reduce the computational burden from $O(p^3)$ to $O(n^2 p)$.

# Implementation

**Table 8:** Implementations of Horseshoe and Other Shrinkage Priors

| Implementation (Package/URL) | Authors |
|---|---|
| R package: `monomvn` | Gramacy et al. [2010] |
| R code in paper | Scott [2010] |
| R package: `horseshoe` | van der Pas et al. [2016] |
| R package: `fastHorseshoe` | Hahn et al. [2016] |
| MATLAB code | Bhattacharya et al. [2016] |
| GPU accelerated Gibbs sampling | Terenin et al. [2016] |
| `bayesreg` + MATLAB code in paper | Makalic and Schmidt [2016] |
| MATLAB code | Johndrow and Orenstein [2017] |