LEARNING GAUSSIAN GRAPHICAL MODELS USING DISCRIMINATED HUB GRAPHICAL LASSO

Zhen Li, Jingtian Bai, Weilian Zhou

Department of Statistics, North Carolina State University

Objectives

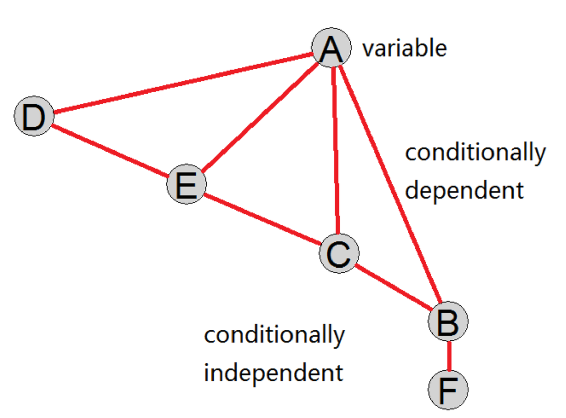
Learning underlying **stochastic dependency structures** among different factors for the data:

- n observations and p variables, $p > n$ possible
- Assuming $\boldsymbol{x}_1, \cdots, \boldsymbol{x}_n \overset{\text{iid}}{\sim} N_p(\mathbf{0}, \boldsymbol{\Sigma})$
- Domain knowledge of some dependency relationships incorporated

Hub Gaussian Graphical Model

Graphical Model: A set of multivariate joint distributions associated with a graph $G = (V, E)$

- V: vertex set, representing **variables**
- E: edge set, representing **conditional dependency**. X satisfies the **pairwise Markov property** if X_v and X_w are independent given $X_{V\setminus\{v,w\}}$ whenever $\{v, w\} \notin E$



Gaussian Graphical Model: Further assuming $\boldsymbol{x}_1, \cdots, \boldsymbol{x}_n \overset{\text{iid}}{\sim} N_p(\mathbf{0}, \boldsymbol{\Sigma})$, and $\boldsymbol{\Theta} = \boldsymbol{\Sigma}^{-1}$ is the **precision matrix**

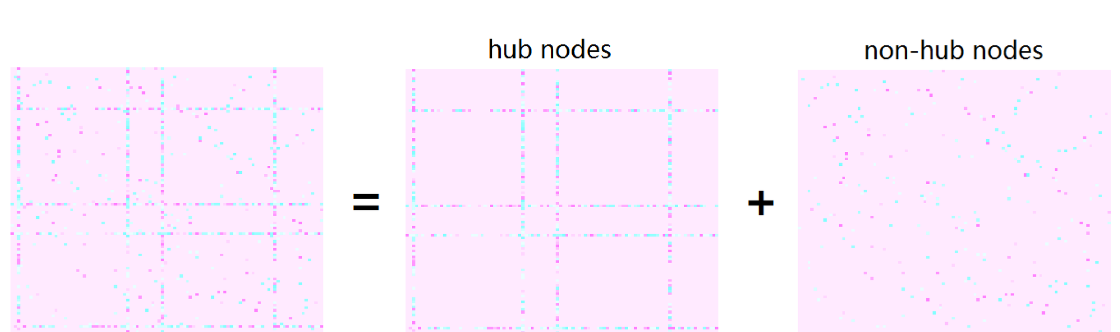
- The **MLE** maximizes

$$\ell(\boldsymbol{X}, \boldsymbol{\Theta}) = -\log\det\boldsymbol{\Theta} + \mathrm{trace}(\boldsymbol{S}\boldsymbol{\Theta})$$

$\boldsymbol{S}$ is the **empirical covariance matrix** of $\boldsymbol{X}$

- V_v and V_w $(v \neq w)$ are conditionally independent iff $\Theta_{vw} = 0$

Graphical Model with Hubs: Nodes that are connected to a very **substantial number of** other nodes in a graph



$$\boldsymbol{\Theta} = \boldsymbol{V} + \boldsymbol{V}^T + \boldsymbol{Z}$$

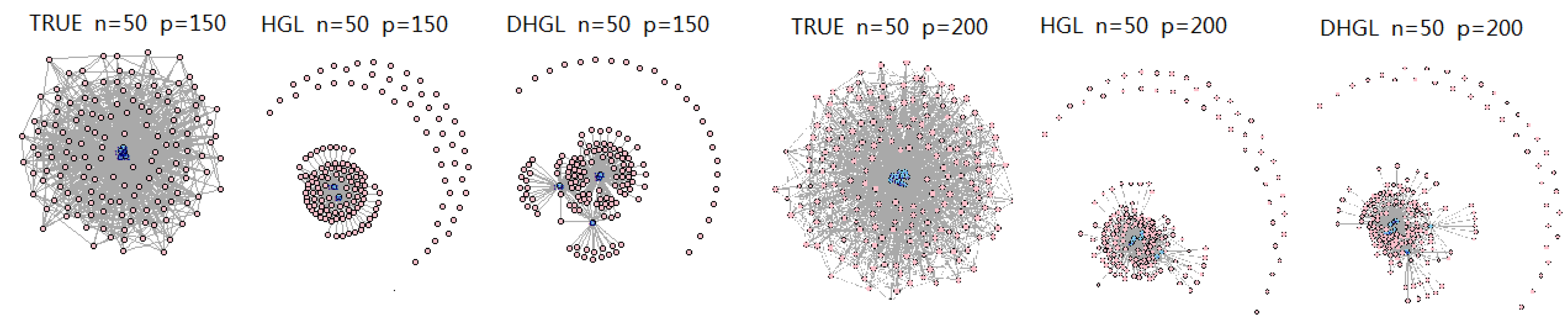
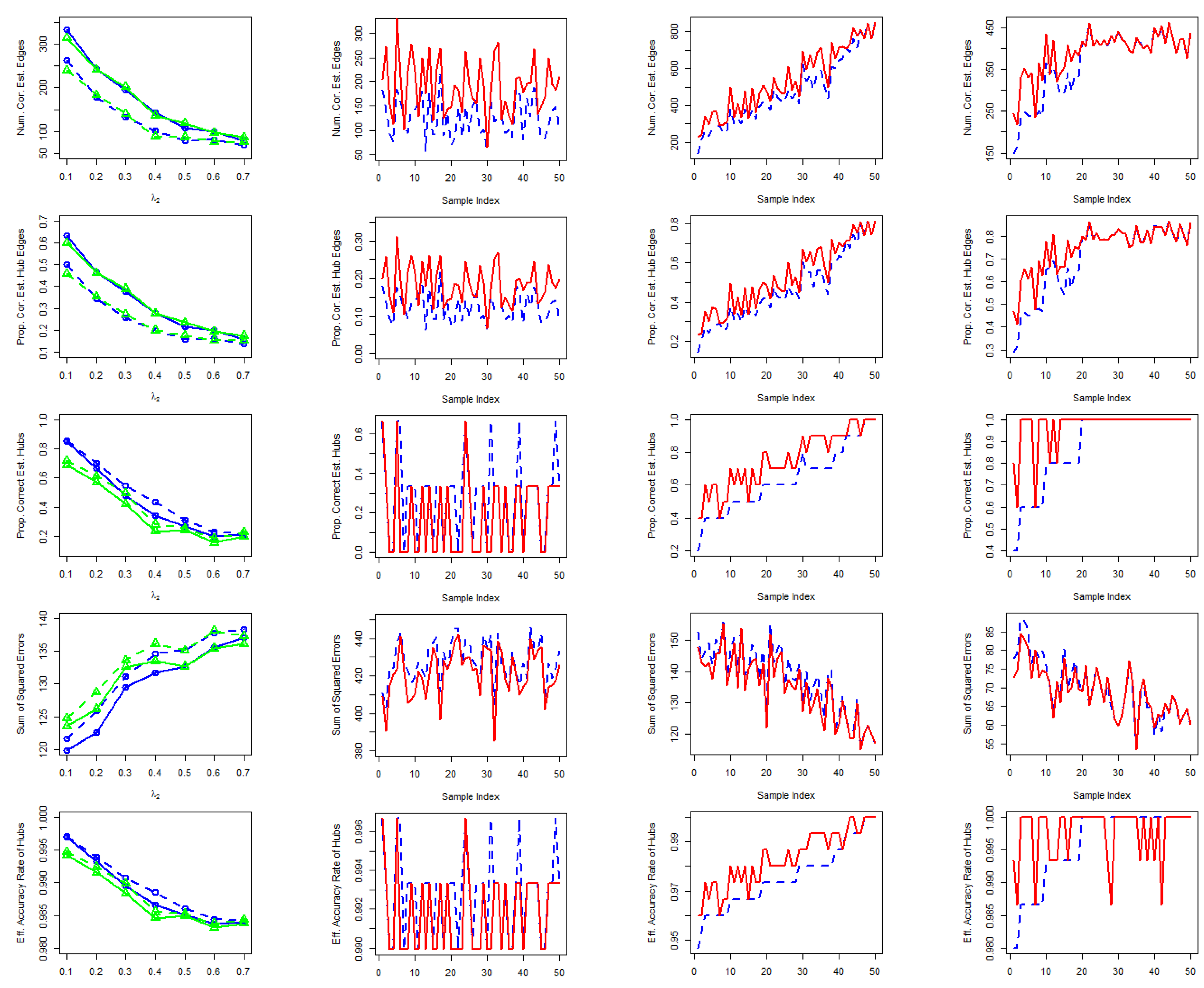


Figure 1: True and estimated graphs (hubs are blue), with $\mathcal{K}$ given (left three) and selected using Graphical Lasso (right three).

Discriminated Hub Graphical Lasso (DHGL)

$$\underset{\boldsymbol{\Theta}\in\mathcal{S},\boldsymbol{V},\boldsymbol{Z}}{\text{minimize}}\ \ell(\boldsymbol{X},\boldsymbol{\Theta}) + \lambda_1\|\mathbf{Z}-\mathrm{diag}(\mathbf{Z})\|_1 + \lambda_2\sum_{j\notin\mathcal{D}}\|(\mathbf{V}-\mathrm{diag}(\mathbf{V}))_j\|_1 + \lambda_3\sum_{j\notin\mathcal{D}}\|(\mathbf{V}-\mathrm{diag}(\mathbf{V}))_j\|_q$$
$$+\lambda_4\sum_{j\in\mathcal{D}}\|(\mathbf{V}-\mathrm{diag}(\mathbf{V}))_j\|_1 + \lambda_5\sum_{j\in\mathcal{D}}\|(\mathbf{V}-\mathrm{diag}(\mathbf{V}))_j\|_q$$
$$\text{subject to } \boldsymbol{\Theta} = \boldsymbol{V} + \boldsymbol{V}^T + \boldsymbol{Z};\ \mathcal{S} = \{\boldsymbol{\Theta} : \boldsymbol{\Theta} \succ 0 \text{ and } \boldsymbol{\Theta} = \boldsymbol{\Theta}^T\}$$



$\lambda_3 = 1$ (blue), 1.5 (green) | $p = 300$, $n = 100$ | $p = 150$, $n = 50$, $|\mathcal{H}| = 10$ | $p = 150$, $n = 50$, $|\mathcal{H}| = 5$

Figure 2: Measures of performances when some (left two) or no hubs are known (right two), using DHGL (solid) or HGL (dashed).

Computation

- Give **"loose conditions"** ($\lambda_4 \leq \lambda_2$, $\lambda_5 \leq \lambda_3$) to nodes in $\mathcal{D}$
- Reduce to HGL in Tan et al. (2014) when $\mathcal{D} = \varnothing$
- Use Alternating Direction Methods of Multipliers **(ADMM)** to solve the convex problem
- Computational complexity: $\boldsymbol{\mathcal{O}(p^3)}$ per iteration
- Select tunning parameters by minimizing a **BIC-type quantity**

DHGL with Known Hub Nodes

1. Use HGL to get the estimated hubs $\hat{\mathcal{H}}_{\text{HGL}}$.
2. Set $\mathcal{D} = \mathcal{K} \setminus \hat{\mathcal{H}}_{\text{HGL}}$, where $\mathcal{K}$ is set of known hubs.
3. If $\mathcal{D} \neq \varnothing$, use DHGL to estimate Θ and get the estimated hubs $\hat{\mathcal{H}}_{\text{DHGL}}$, where λ_1, λ_2, λ_3 remain the same values as in HGL and λ_4, λ_5 are selected using the BIC-type quantity. Then, set $\hat{\mathcal{H}} = \hat{\mathcal{H}}_{\text{HGL}} \cup \hat{\mathcal{H}}_{\text{DHGL}}$ as the set of estimated hubs. If $\mathcal{D} = \varnothing$, use the estimation in HGL directly.

DHGL without Known Hub Nodes

1. Use HGL to get the estimated hubs $\hat{\mathcal{H}}_{\text{HGL}}$.
2. Adjust regularization parameter λ of GL from large to small until $|\hat{\mathcal{H}}_{\text{GL},\lambda} \setminus \hat{\mathcal{H}}_{\text{HGL}}| > 0$ and $|\hat{\mathcal{H}}_{\text{GL},\lambda} \cup \hat{\mathcal{H}}_{\text{HGL}}| \leq \max\{|\hat{\mathcal{H}}_{\text{HGL}}| + a, b|\hat{\mathcal{H}}_{\text{HGL}}|\}$, where $a \in \boldsymbol{N}_+$, $b > 1$ but $b \approx 1$. $\hat{\mathcal{H}}_{\text{GL},\lambda}$ is the set of estimated hubs by GL with the parameter λ.
3. Set $\mathcal{D} = \hat{\mathcal{H}}_{\text{GL},\lambda} \setminus \hat{\mathcal{H}}_{\text{HGL}}$ which is non-empty.
4. Use DHGL to estimate $\boldsymbol{\Theta}$, where λ_1, λ_2, λ_3, λ_4 remain the same values as in HGL and λ_5 is selected using the BIC-type quantity.

Conclusion

- With some hubs known, DHGL outperforms HGL in estimating the precision matrix.
- Without known hubs, DHGL outperforms HGL given correct prior information, and rarely degenerates even if the prior information is incorrect.