BAGUS: A Bayesian algorithm for Gaussian Graphical Model with Unequal Shrinkage

Presenter: Lingrui Gan, joint work with Naveen N. Narisetty and Feng Liang
Department of Statistics, University of Illinois at Urbana-Champaign

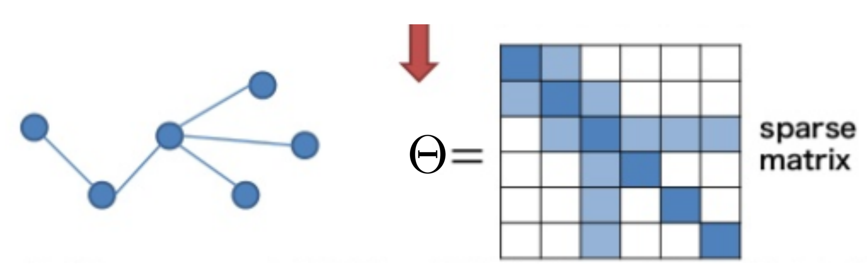
ABSTRACT

We consider a Bayesian framework for estimating a high-dimensional sparse precision matrix, in which adaptive shrinkage and sparsity are induced by a mixture of Laplace priors. For fast and efficient computation, an EM algorithm is proposed to obtain the MAP estimate of the precision matrix and posterior probabilities on the edges of the underlying sparse structure. Besides discussing our formulation from the Bayesian standpoint, we investigate the MAP estimate from a penalized likelihood perspective that gives rise to a new non-convex penalty. Optimal error rates for estimation consistency along with selection consistency for sparse structure recovery are shown for the unique global optimizer under mild conditions. Through extensive simulation studies and real applications, we have demonstrated the fine performance of our method compared with the existing alternatives.

INTRODUCTION

Gaussian Graphical Model:

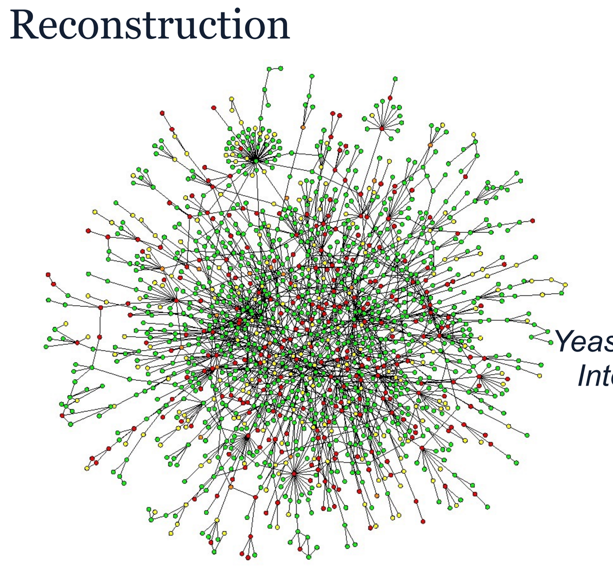
$$Y \sim N_p(0, \Theta^{-1})$$



Our goal is to estimate the high dimensional sparse inverse covariance matrix (precision matrix).

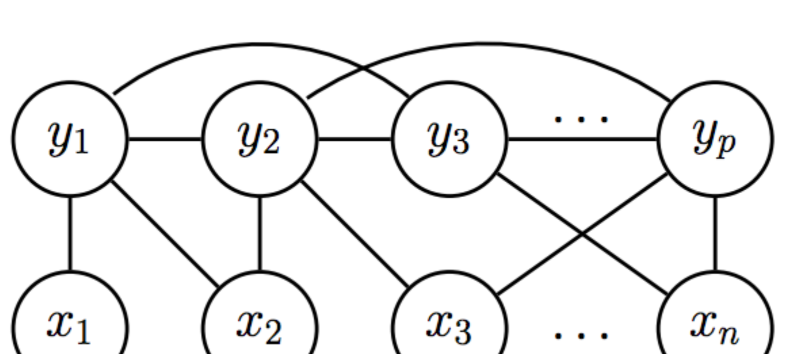
Applications:

1. Network Reconstruction



Example: Yeast Protein-Protein Interaction Network

2. Predictive Modeling



Example: Conditional Gaussian Graphical Model (Matt Wytock, Zico Kolter ICML 2014)

MODEL FORMULATION

Hierarchical Bayesian Model Formulation:
Prior distributions on Θ satisfying $\Theta \succ 0$ and $||\Theta||_2 \leq Ra$:

$$\begin{cases} \theta_{ij} \mid r_{ij} = 0 \sim \mathrm{DE}(0, v_0), \\ \theta_{ij} \mid r_{ij} = 1 \sim \mathrm{DE}(0, v_1), \end{cases}$$

$$r_{ij} \mid \eta \sim \mathrm{Bern}(\eta).$$

$$\theta_{ii} \mid \tau \sim \mathrm{Ex}(\tau)$$

Our target is to get the MAP estimate, or equivalently minimize the following objective function under constraints $\Theta \succ 0$ and $||\Theta||_2 \leq Ra$:

$$L(\Theta) = \frac{n}{2}\Big(\mathrm{tr}(S\Theta) - \log\det(\Theta)\Big) - \sum_{i<i} \log\left[\Big(\frac{\eta}{2v_1}\Big)e^{-\frac{|\theta_{ij}|}{v_1}} + \Big(\frac{1-\eta}{2v_0}\Big)e^{-\frac{|\theta_{ij}|}{v_0}}\Big)\right] + \sum_i \tau|\theta_{ii}|$$

THEORETICAL GUARANTEES

Details are omitted due to the space limit. Please refer to our manuscript for details.

Theorem 1. If $Ra < (\frac{nv_0}{2})^{\frac{1}{2}}$, $\arg\min_{\Theta \succ 0, ||\Theta||_2 \leq Ra} L(\Theta)$ is strictly convex.

Theorem 2[Estimation Error Bound].

If $||\Theta^0||_2$ is upper bounded, Y follows an exponential or polynomial tail distribution, with a proper tuning of v_0, v_1 and τ, let sample size satisfies $n \geq Md\log p$, where M is some constant and $d := \max_{i=1,2,\ldots,p} |\{j|\Theta^0_{ij} \neq 0\}|$, then there exists a unique global minimizer with an estimation error $O_p(\sqrt{\log p/n})$ in ℓ_1 norm.

Exponential-Tail Distributions Examples: Gaussian, any bounded random variable (e.g., Bernoulli, multinomial) or any finite mixture of them.

Theorem 3[Structure Recovery Consistency].
If the minimal signal strength is larger than $K\sqrt{\log p/n}$ for some large enough K, our model could recover the true structure (signal or noise) correctly with probability converging to 1.

Highlights of the theorems: 1:The incoherence condition for the results of Lasso methods are not required. 2:Very mild assumptions on the true precision matrix.

COMPUTATION: AN EM ALGORITHM

Algorithm 1 Bagus

Initialize $W = \Theta = \mathrm{I}$

repeat

Update P with each entry p_{ij} updated as $\log \frac{p_{ij}}{1-p_{ij}} \leftarrow \Big(\log \frac{v_0}{v_1} + \log \frac{\eta}{1-\eta} - \frac{|\theta^{(t)}_{ij}|}{v_1} + \frac{|\theta^{(t)}_{ij}|}{v_0}\Big)$.

for j in $1:p$ **do**

Move the j-th column and j-th row to the end (implicitly), namely $\Theta_{11} := \Theta_{\backslash j \backslash j}$, $\theta_{12} := \theta_{\backslash jj}$, $\theta_{22} := \theta_{jj}$

Update w_{22} using $w_{22} \leftarrow s_{22} + \frac{2}{n}\tau$

Update θ_{12} by solving $ns_{12} + nw_{22}\Theta_{11}^{-1}\theta_{12} + \Big(\frac{1}{v_1}P_{12} + \frac{1}{v_0}(1 - P_{12})\Big) \odot \mathrm{sign}(\theta_{12}) = 0$ with coordinate descent under the constraint $||\Theta||_F \leq Ra$.

Update θ_{22} using $\theta_{22} \leftarrow \frac{1}{w_{22}} + \theta_{12}^T\Theta_{11}^{-1}\theta_{12}$.

Update W using

$$\begin{bmatrix} W_{11} & w_{12} \\ w_{12}^T & w_{22} \end{bmatrix} \leftarrow \begin{bmatrix} \Theta_{11}^{-1} + \frac{\Theta_{11}^{-1}\theta_{12}\theta_{12}^T\Theta_{11}^{-1}}{\theta_{22} - \theta_{12}^T\Theta_{11}^{-1}\theta_{12}} & -\frac{\Theta_{11}^{-1}\theta_{12}}{\theta_{22} - \theta_{12}^T\Theta_{11}^{-1}\theta_{12}} \\ -\frac{\Theta_{11}^{-1}\theta_{12}}{\theta_{22} - \theta_{12}^T\Theta_{11}^{-1}\theta_{12}} & \frac{1}{\theta_{22} - \theta_{12}^T\Theta_{11}^{-1}\theta_{12}} \end{bmatrix}$$

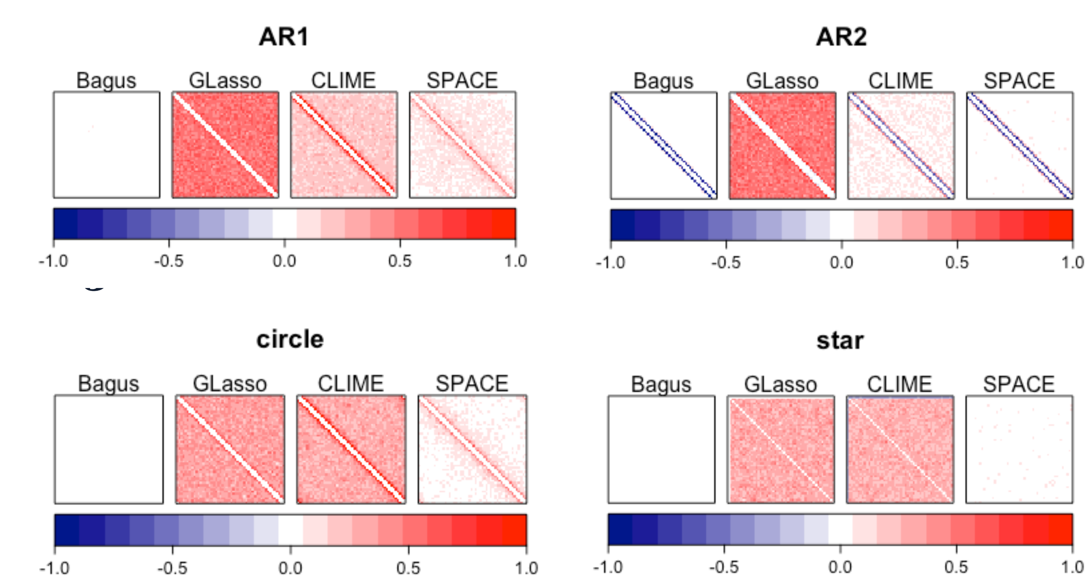
end for

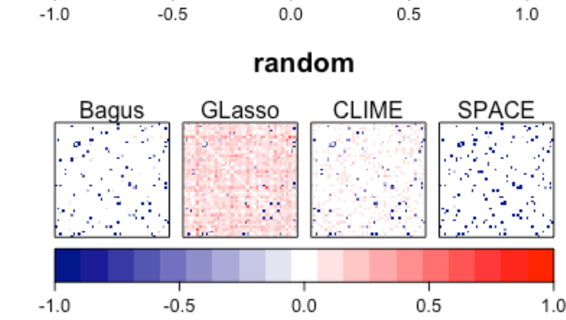
until Converge

Return Θ, P

SIMULATION RESULTS

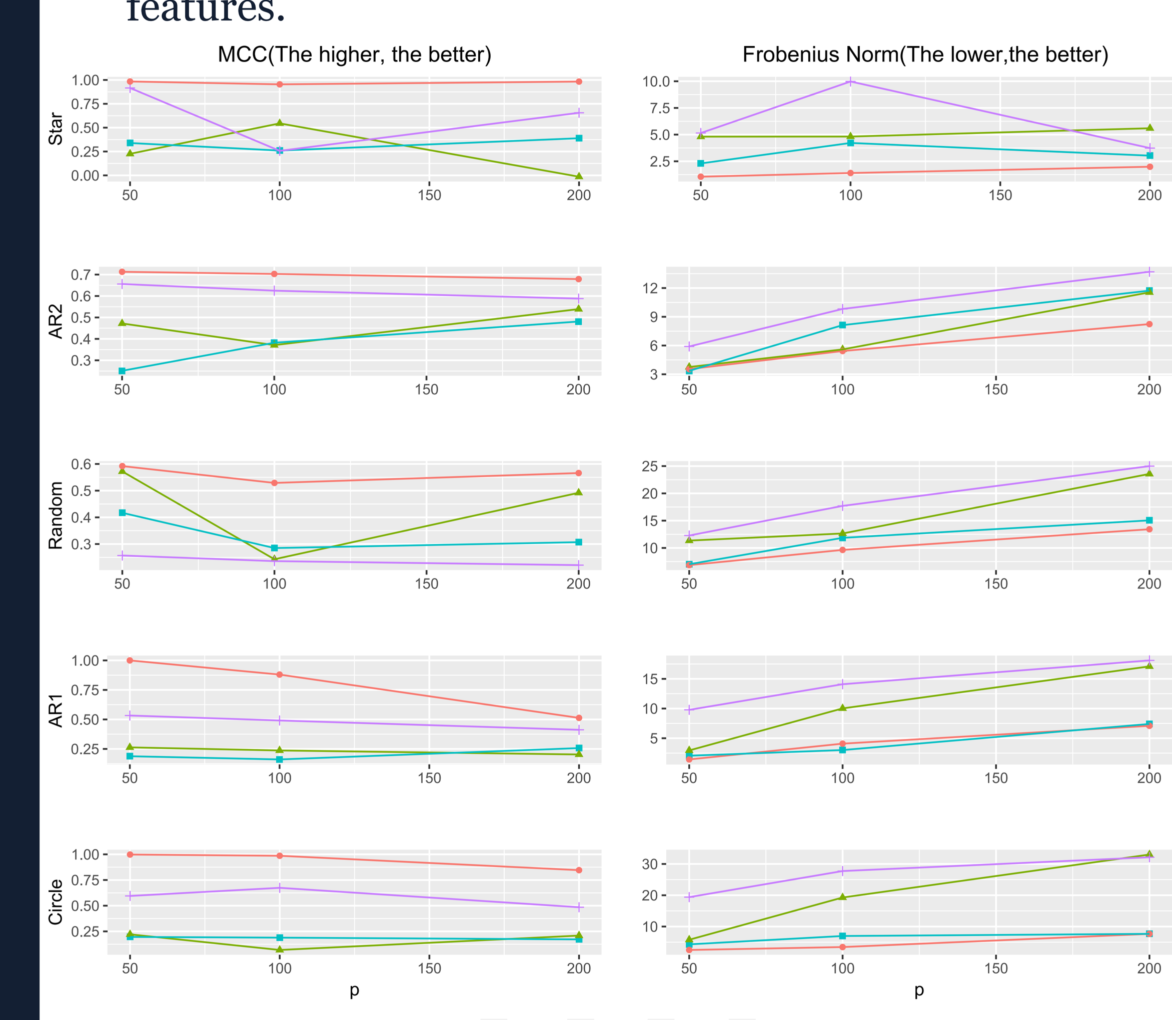
Models:
1: AR(1) Model; 2: AR(2) Model;
3: Circle Model; 4: Star Model; 5: Random Network
Scenarios Considered:
sample size n=100; variable size p=50,100,200;
Evaluation Metrics:
1. Structure Recovery Accuracy:
MCC(Matthews correlation coefficient)
2. Estimation Accuracy:
L2 Distance between the truth and the estimation
Experimental Results:





The differences between the estimated and true structures (p=50, 50 replicates average)
Red points: Select noise; Blue points: Miss signal

1. Our Bagus algorithm performs very well in structure recovery, although it could miss some signals sometimes.
2. SPACE is the second best performing algorithm, which behaves similar to Bagus but sometimes selects more noise features..
3. Graphical Lasso and CLIME tend to select more noise features.



Performance Measures for Each Cases; Our Bagus method always perform at the top , both in structure recovery and estimation.

PREDICT CALL ARRIVAL PATTERNS @A TELEPHONE CALL CENTER

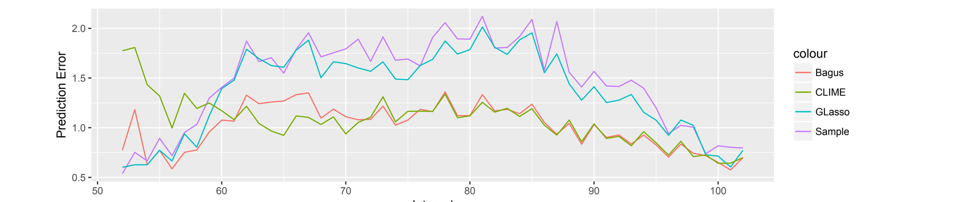
Dataset: Call arrival pattern of a telephone call center in a major U.S. northeastern financial organization. The data was collected every day in 2002 from 7 AM till midnight, except 6 days when the data collecting machine is out of order.

Method: Models are trained on the first 205 days and the remaining 34 days are used for evaluation. In the testing set, the first 51 intervals are assumed observed and we predict the last 51 intervals, using:

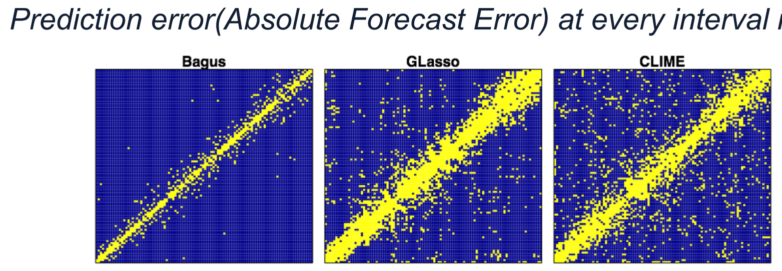
$$f(Y_{2i}|Y_{1i}) = \mathrm{N}(u_2 - \Theta_{22}^{-1}\Theta_{21}(Y_{1i} - u_1), \Theta_{22}^{-1})$$

Average Prediction Error

	Sample	GLasso	Adaptive Lasso	SCAD	CLIME	Bagus
Average AAFE	1.46	1.38	1.34	1.31	1.08	**1.00**



Prediction error(Absolute Forecast Error) at every interval in the test set.

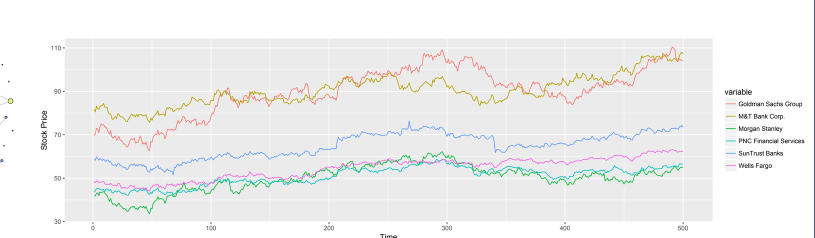


Precision matrix estimated from different methods.

S&P 500 STOCK DATA

Dataset: The closing prices of 452 stocks in S&P 500 companies from 2003 to 2008 (1258 trading days). Stocks are categorized into 10 separate sectors, e.g. Information Technology, Financials, according to the industry the corresponding companies are in.

Random Sampled 150 Stocks Network



A snapshot of stocks detected as dependent with SunTrust Banks from our model.

CITATIONS

1.Friedman, J., Hastie, T., and Tibshirani, R. (2008). Sparse inverse covariance estimation with the graphical lasso. Biostatistics, 9(3):432–441.
2.Peng, J., Wang, P., Zhou, N., and Zhu, J. (2009). Partial correlation estimation by joint sparse regression models. Journal of the American Statistical Association, 104(486):735– 746.
3.Cai, T., Liu, W., and Luo, X. (2011). A constrained l1 minimization approach to sparse precision matrix estimation. Journal of the American Statistical Association, 106(494):594–607.
4.Ravikumar, P., Wainwright, M. J., Raskutti, G., and Yu, B. (2011). High-dimensional covariance estimation by minimizing l_1-penalized log-determinant divergence. Electronic Journal of Statistics, 5:935–980.
5.Loh, P.-L. and Wainwright, M. J. (2015). Regularized m-estimators with nonconvexity: Statistical and algorithmic theory for local optima. Journal of Machine Learning Research, 16:559–616.

ILLINOIS
UNIVERSITY OF ILLINOIS AT URBANA-CHAMPAIGN