Estimating Sparse Precision Matrix with Bayesian Regularization

Lingrui Gan, Naveen N. Narisetty, Feng Liang

Department of Statistics
University of Illinois at Urbana-Champaign

Table of Contents

Introduction

Problem Statement:

$$Y_1, \cdots, Y_n \overset{iid}{\sim} \mathsf{N}_p(0, \Theta^{-1}).$$

Denote $S = \frac{1}{n}\sum Y_i Y_i^t$ as the sample covariance matrix of the data, then the $\log$-likelihood is given by

$$l(\Theta) = \log L(\Theta) = \frac{n}{2}\Big(\log\det(\Theta) - \mathsf{tr}(S\Theta)\Big). \quad (1)$$

- Our target is to estimate with respect to Θ, the precision matrix.

Graphical Representation

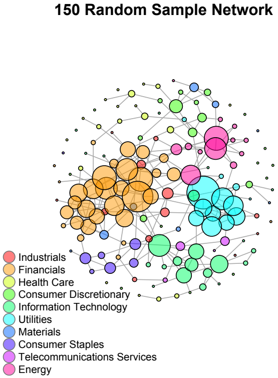
Well-known Fact:

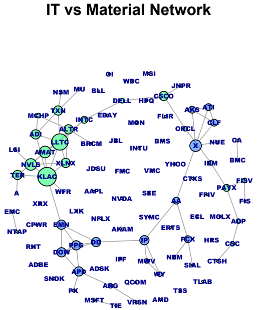
- Consider Undirected graph G=(V,E) with V is the vertex set and E is the edge set
- Edge (α, β) not exists $\Leftrightarrow \alpha \perp\!\!\!\perp \beta | V \setminus (\alpha, \beta) \Leftrightarrow \Theta_{\alpha,\beta} = 0$

Due to the relationship between precision matrix and graph, our problem of interest is often called **Gaussian Graphical Model**.

Examples: gene network in biology and financial network.

Financial Network[Gan and Liang, 2016]





To make this problem applicable under the high-dimensional scenario, assumptions need to be made.

Sparsity Assumption

The sparsity assumption is the most common and practical useful one [Dempster, 1972]. It assumes that the majority of the entries are zero, while only a few entries in Θ are non-zero.

Literature Review

Penalized Likelihood

Minimize the negative log-likelihood function with an element-wise penalty on the off-diagonal entries of Θ, i.e.,

$$\arg\min_{\Theta} \Big[-\frac{n}{2}\Big(\log\det(\Theta) - \mathrm{tr}(S\Theta) \Big) + \lambda \sum_{i<j} \mathrm{pen}(\theta_{ij}) \Big].$$

- The penalty function $\mathrm{pen}(\theta_{ij})$ is often taken to be L_1 [Yuan and Lin, 2007, Banerjee et al., 2008, Friedman et al., 2008],
- but SCAD is also been used [Fan et al., 2009].
- Asymptotic properties have been studied in [Rothman et al., 2008, Lam and Fan, 2009]

Literature Review

Regression

- Sparse regression model is estimated separately in each column of Θ.
  Implicitly, they are modeling with under the likelihood $\prod_i P(\mathbf{Y}[i,]|\mathbf{Y}[-i,])$, instead of $P(\mathbf{Y})$.[a]
  [Meinshausen and Bühlmann, 2006, Peng et al., 2009]

[a] Denote $\mathbf{Y} = (Y_1, \cdots, Y_n)$.

- Other work: [Liu et al., 2009, Ravikumar et al., 2011]; CLIME estimator[Cai et al., 2011];

Literature Review

Bayesian Regularization

- Several Bayesian approaches have also been proposed [Wang, 2012, Banerjee and Ghosal, 2015, Gan and Liang, 2016].
- However, Bayesian methods are not in wide use in this fields, because of the high computation cost of MCMC.

Table of Contents

Spike and Slab Prior

Double Exponential Spike and Slab Prior

The cornerstone of our Bayesian formulation is the following spike and slab prior on the off diagonal entries θ_{ij} $(i < j)$:

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

where $0 \le v_0 < v_1$ and r_{ij} for all i,j, follows

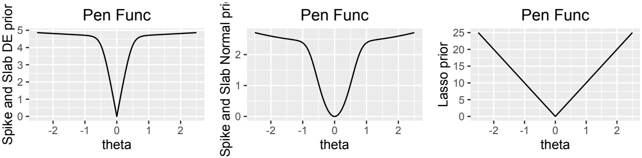
$$r_{ij} \sim \text{Bern}(\eta).$$

Penalized Likelihood Perspective

Bayesian Regularization Function

The "signal" indicator r_{ij} can be treated as latent and integrate it out, then we get the Bayesian regularization function:

$$\text{pen}(\theta_{ij}) = -\log \int \pi(\theta_{ij}|r_{ij})\pi(r_{ij}|\eta)dr_{ij}$$



Model Specification

$$Y_1, \cdots, Y_n|\Theta \overset{iid}{\sim} \mathsf{N}_p(0, \Theta^{-1}).$$

$$\theta_{ij} \sim \eta \mathsf{DE}(0, v_1) + (1-\eta)\mathsf{DE}(0, v_0) \quad i < j$$

$$\theta_{ji} = \theta_{ji}$$

$$\theta_{ii} \sim \mathsf{Ex}(\tau)$$

The full posterior distribution $\pi(\Theta, R|\mathbf{Y_{1:n}})$ is proportional to

$$f(\mathbf{Y_{1:n}}|\Theta)\Big(\prod_{i<j} \pi(\theta_{ij}|r_{ij})\pi(r_{ij}|\eta) \prod_{i} \pi(\theta_{ii}|\tau)\Big) \tag{2}$$

where $R_{p\times p}$ is a matrix with binary entries r_{ij}

Table of Contents

EM Algorithm

- We treated R as latent and derive an EM algorithm to obtain a maximum a posterior (MAP) estimate of Θ in the M-step and the posterior distribution of R in the E-step.
- The updating scheme is in the similar fashion with [Friedman et al., 2008], i.e. updating one column and one row at a time.

Algorithm 1 EM Algorithm

If $n < p$: **Initialize** W= sample covariance matrix S
Else: **Initialize** $W = S + \text{diag}(\frac{2\tau}{n}, ..., \frac{2\tau}{n})$
Initialize Θ=W^{-1}
repeat
  Update P with $\log \frac{p_{ij}}{1-p_{ij}} = \left(\log \frac{v_0}{v_1} + \log \frac{\eta}{1-\eta} - \frac{|\theta_{ij}|}{v_1} + \frac{|\theta_{ij}|}{v_0} \right)$
  for j in $1 : p$ **do**
    Move the j-th column and jth row to the end (implicitly), namely $\Theta_{11} = \Theta_{\backslash j \backslash j}$, $\theta_{12} = \theta_{\backslash jj}$, $\theta_{22} = \theta_{jj}$
    Save $W^0 = W$, $\Theta^0 = \Theta$
    Update w_{22} using $w_{22} \leftarrow s_{22} + \frac{2}{n}\tau$
    Update W_{12} using $w_{12} \leftarrow s_{12} + \frac{1}{nv_1} P_{12} \odot \text{sign}(\theta_{12}) + \frac{1}{nv_0}(1 - P_{12}) \odot \text{sign}(\theta_{12})$
    Update θ_{12} using $\theta_{12} \leftarrow -\frac{\Theta_{11} w_{12}}{w_{22}}$
    Update θ_{22} using $\theta_{22} \leftarrow \frac{1 - w_{12}^T \theta_{12}}{w_{22}}$
    Update Θ
    If $Q(\Theta|\Theta^0) \leq Q(\Theta^0|\Theta^0)$: $W \leftarrow W^0$, $\Theta \leftarrow \Theta^0$
  end for
until Converge
Return Θ, P

Our algorithm always ensures the symmetry and positive definiteness of the precision matrix estimation outputted.

Theorem

(Symmetry)
The estimate of Θ is always guaranteed to be symmetric.

Theorem

(Positive Definiteness)
If $\Theta^{(0)} > 0$, i.e the initial estimate of precision matrix is positive definite, $\Theta^{(t)} > 0$, $\forall t \geq 1$.

For the existing algorithms, the positive definiteness of the estimate usually doesn't hold [Mazumder and Hastie, 2012].

Table of Contents

Asymptotic

Theorem (Rate of Convergence)

Under the regularity conditions (A)-(B), if $n\sqrt{\frac{\log p}{n}} \preceq \frac{1}{v_0} \preceq n\sqrt{\frac{(p+s)\log p}{sn}}$, $\log(\frac{v_1}{v_0}) \succeq \frac{(p+s)\log p}{\sqrt{s}}$ *and* $\tau \preceq n(\sqrt{\frac{\log p}{n}})$, *then there exists a local minimizer* $\hat{\Theta}$, *which is positive definite and symmetric, and it satisfies*

$$\|\hat{\Theta} - \Theta_0\|_F^2 = O_p\{(p+s)\log p/n\}$$

Theorem (Selection Consistency)

Under the same conditions given in previous theorem and regularity conditions on the signal strength[a] *, for any constant $C > 0$, we have*

$$P(\max_{(i,j)\notin S_g} \log \frac{p_{ij}}{1-p_{ij}} < -C) \to 1 \quad (3)$$

and

$$P(\min_{(i,j)\in S_g} \log \frac{p_{ij}}{1-p_{ij}} > C) \to 1 \quad (4)$$

Consequently,

$$P(\hat{S}_g = S_g) \to 1 \quad (5)$$

[a]Denote S_g as the true signal set

Table of Contents

Simulation Studies

1. Model 1: An $AR(1)$ model with $w_{ii} = 1$, $w_{i,i-1} = w_{i-1,i} = 0.5$
2. Model 2: An $AR(2)$ model $w_{ii} = 1$, $w_{i,i-1} = w_{i-1,i} = 0.5$ and $w_{i,i-2} = w_{i-2,i} = 0.25$.
3. Model 3: A circle model with $w_{ii} = 2$, $w_{i,i-1} = w_{i-1,i} = 1$, and $w_{1,p} = w_{p,1}$=0.9
4. Model 4: Random Select Model.

For each model, three scenarios will be considered: *C*ase 1: $n = 100$, $p = 50$; *C*ase 2: $n = 200, p = 100$; *C*ase 3: $n = 100, p = 100$.

Metrics

Average Selection accuracy and L_2 distance between estimates and truths on 50 replications.

Table: Model1 AR(1)

	Fnorm	Specificity	Sensitivity	MCC
$n = 100, p = 50$				
GLasso	**2.058(0.080)**	0.478(0.039)	1(0)	0.188(0.015)
SPACE	9.763(0.133)	0.908(0.007)	1(0)	0.533(0.015)
Bayes EM	**2.143(0.401)**	0.997(0.004)	0.998(0.007)	**0.961(0.038)**
Sample	17.743(2.147)	NA	NA	NA
$n = 200, p = 100$				
	Fnorm	Specificity	Sensitivity	MCC
GLasso	**2.421(0.073)**	0.553(0.006)	1.000(0.000)	0.155(0.002)
SPACE	13.919(0.080)	0.936(0.009)	1.000(0.000)	0.478(0.035)
Bayes EM	3.716 (0.971)	0.998(0.003)	0.998(0.006)	**0.951(0.055)**
Sample	24.044(1.175)	NA	NA	NA
$n = 100, p = 100$				
	Fnorm	Specificity	Sensitivity	MCC
GLasso	**3.012(0.081)**	0.571(0.006)	1.000(0.000)	0.161(0.002)
SPACE	14.097(0.159)	0.940(0.010)	1.000(0.002)	0.491(0.037)
Bayes EM	**2.916(0.309)**	1.000(0.001)	1.000(0.001)	**0.990(0.018)**
Sample	NA	NA	NA	NA

Model 1

An $AR(1)$ model with $w_{ii} = 1$, $w_{i,i-1} = w_{i-1,i} = 0.5$

Table: Model2 AR(2)

	Fnorm	Specificity	Sensitivity	MCC
$n = 100, p = 50$				
GLasso	**3.361(0.240)**	0.479(0.056)	0.981(0.015)	0.251(0.028)
SPACE	5.903(0.070)	0.982(0.004)	0.608(0.038)	0.656(0.029)
Bayes EM	**3.256(0.276)**	0.988(0.008)	0.644(0.070)	**0.712(0.038)**
Sample	17.882(2.144)	NA	NA	NA
$n = 200, p = 100$				
	Fnorm	Specificity	Sensitivity	MCC
GLasso	4.315(0.073)	0.559(0.007)	0.998(0.003)	0.219(0.003)
SPACE	10.810(0.077)	0.991(0.001)	0.796(0.027)	0.784(0.019)
Bayes EM	**3.185(0.215)**	0.995(0.002)	0.867(0.029)	**0.864(0.023)**
Sample	24.273(1.269)	NA	NA	NA
$n = 100, p = 100$				
	Fnorm	Specificity	Sensitivity	MCC
GLasso	8.130(0.035)	0.901(0.007)	0.745(0.028)	0.382(0.017)
SPACE	9.819(0.083)	0.991(0.002)	0.566(0.025)	0.625(0.021)
Bayes EM	**6.552(0.308)**	0.998(0.004)	0.491(0.042)	**0.663(0.024)**
Sample	NA	NA	NA	NA

Model 2

An $AR(2)$ model $w_{ii} = 1$, $w_{i,i-1} = w_{i-1,i} = 0.5$ and $w_{i,i-2} = w_{i-2,i} = 0.25$.

Table: Model3 Circle Model

	Fnorm	Specificity	Sensitivity	MCC
$n = 100, p = 50$				
GLasso	4.319(0.174)	0.492(0.064)	1.000(0.000)	0.196(0.024)
SPACE	19.402(0.232)	0.930(0.006)	1.000(0.000)	0.595(0.019)
Bayes EM	**3.338(0.416)**	0.979(0.008)	1.000(0.003)	**0.812(0.053)**
Sample	35.509(4.291)	NA	NA	NA
$n = 200, p = 100$				
GLasso	**4.787(0.223)**	0.515(0.020)	1.000(0.000)	0.145(0.006)
SPACE	27.708(0.196)	0.971(0.009)	0.999(0.004)	0.645(0.066)
Bayes EM	6.541(1.548)	0.981(0.005)	1.000(0.000)	**0.717(0.047)**
Sample	48.105(2.354)	NA	NA	NA
$n = 100, p = 100$				
GLasso	**6.981(0.192)**	0.647(0.005)	1.000(0.000)	0.189(0.002)
SPACE	27.737(0.345)	0.975(0.010)	0.994(0.008)	**0.674(0.062)**
Bayes EM	**6.603(1.497)**	0.975(0.008)	1.000(0.000)	**0.673(0.064)**
Sample	NA	NA	NA	NA

Model 3

A circle model with $w_{ii} = 2$, $w_{i,i-1} = w_{i-1,i} = 1$, and $w_{1,p} = w_{p,1}$=0.9

Table: Model4 Random Select Model

	$n = 100, p = 50$			
	Fnorm	Specificity	Sensitivity	MCC
GLasso	**7.017(0.256)**	0.592(0.027)	0.839(0.042)	0.236(0.025)
SPACE	13.519(0.573)	0.999(0.001)	0.179(0.059)	0.390(0.071)
Bayes EM	**7.438(0.718)**	0.987(0.007)	0.477(0.053)	**0.563(0.048)**
Sample	17.232(1.971)	NA	NA	NA
	$n = 200, p = 100$			
	Fnorm	Specificity	Sensitivity	MCC
GLasso	8.597(0.164)	0.722(0.007)	0.891(0.019)	0.259(0.007)
SPACE	18.276(0.536)	0.999(0.000)	0.168(0.050)	0.371(0.059)
Bayes EM	**7.816(0.397)**	0.997(0.002)	0.498(0.039)	**0.644(0.019)**
Sample	23.433(1.065)	NA	NA	NA
	$n = 100, p = 100$			
	Fnorm	Specificity	Sensitivity	MCC
GLasso	11.85(0.900)	0.837(0.047)	0.720(0.049)	0.285(0.033)
SPACE	17.706(0.203)	1.000(0.000)	0.068(0.015)	0.236(0.028)
Bayes EM	**10.847(0.230)**	0.999(0.000)	0.286(0.019)	**0.498(0.023)**
Sample	NA	NA	NA	NA

Model 4

Random Select Model.
Specifically, the model generating process is:

1. Set the diagonal entry to be 1.
2. Randomly selected $1.5 \times p_n$ of the edges and set them to be random number uniform from $[0.4, 1] \cup [-1, -0.4]$.
3. First sum the absolute values of the off-diagonal entries, and then divide each off-diagonal entry by 1.1 fold of the sum
4. Average this rescaled matrix with its transpose to ensure symmetry.
5. Multiple each entry by σ^2, which set to be 3 here.

Telephone call center arrival data prediction

- Forecast the call arrival pattern from one call center in a major U.S. northeastern financial organization.
- The training set contains data for the first 205 days. The remaining 34 days are used for testing.
- In the testing set, the first 51 intervals are assumed observed and we will predict the last 51 intervals, using the following relationship:

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

Error Metric

To evaluate the prediction performance, we used the same criteria as [Fan et al., 2009], the average absolute forecast error (AAFE):

$$\mathsf{AAFE}_t = \frac{1}{34}\sum_{i=206}^{239} |\hat{y}_{it} - y_{it}|$$

where $\hat{y}_{it}$ and y_{it} are the predicted and observed values.

Telephone call center arrival data

From the results shown, our method has shown a significant improvement in prediction accuracy when compared with existing methods.

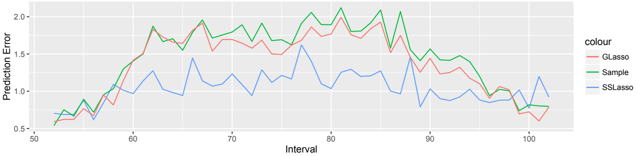


Figure: Prediction Error

	Average Prediction Error				
	Sample	Lasso	Adaptive Lasso	SCAD	SS Lasso
Average AAFE	1.46	1.39	1.34	1.31	**1.05**

Summary

1. We propose a Bayesian model, using Spike and Slab Prior, for Gaussian Graphical Model.
2. An EM algorithm is derived to achieve the fast computation.
3. Simultaneous estimation and selection consistency of our method is proved.
4. Empirical Studies have shown promising results.

References I

Banerjee, O., El Ghaoui, L., and d'Aspremont, A. (2008).
Model selection through sparse maximum likelihood estimation for multivariate gaussian or binary data.
The Journal of Machine Learning Research, 9:485–516.

Banerjee, S. and Ghosal, S. (2015).
Bayesian structure learning in graphical models.
Journal of Multivariate Analysis, 136:147–162.

Cai, T., Liu, W., and Luo, X. (2011).
A constrained ? 1 minimization approach to sparse precision matrix estimation.
Journal of the American Statistical Association, 106(494):594–607.

Dempster, A. P. (1972).
Covariance selection.
Biometrics, pages 157–175.

Fan, J., Feng, Y., and Wu, Y. (2009).
Network exploration via the adaptive lasso and scad penalties.
The Annals of Applied statistics, 3(2):521.

Friedman, J., Hastie, T., and Tibshirani, R. (2008).
Sparse inverse covariance estimation with the graphical lasso.
Biostatistics, 9(3):432–441.

References II

Gan, L. and Liang, F. (2016).
A bayesian em algorithm for graphical model selection.

Lam, C. and Fan, J. (2009).
Sparsistency and rates of convergence in large covariance matrix estimation.
Annals of Statistics, 37(6B):4254.

Liu, H., Lafferty, J., and Wasserman, L. (2009).
The nonparanormal: Semiparametric estimation of high dimensional undirected graphs.
Journal of Machine Learning Research, 10(Oct):2295–2328.

Mazumder, R. and Hastie, T. (2012).
The graphical lasso: New insights and alternatives.
Electronic journal of statistics, 6:2125.

Meinshausen, N. and Bühlmann, P. (2006).
High-dimensional graphs and variable selection with the lasso.
The Annals of Statistics, pages 1436–1462.

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.

References III

Ravikumar, P., Wainwright, M. J., Raskutti, G., Yu, B., et al. (2011).
High-dimensional covariance estimation by minimizing ?1-penalized log-determinant divergence.
Electronic Journal of Statistics, 5:935–980.

Rothman, A. J., Bickel, P. J., Levina, E., Zhu, J., et al. (2008).
Sparse permutation invariant covariance estimation.
Electronic Journal of Statistics, 2:494–515.

Wang, H. (2012).
Bayesian graphical lasso models and efficient posterior computation.
Bayesian Analysis, 7(4):867–886.

Yuan, M. and Lin, Y. (2007).
Model selection and estimation in the gaussian graphical model.
Biometrika, 94(1):19–35.