Bayesian Regularization with an Application to Gaussian Graphical Models

Lingrui Gan

Department of Statistics
University of Illinois at Urbana-Champaign

Joint Work with Feng Liang and Naveen N. Narisetty

Table of Contents

Data Challenges in Modern Applications

- In modern applications in science and engineering:
  - Models are always with large size of parameters (high-dimensional models).
  - Examples: graphical models, deep learning models.

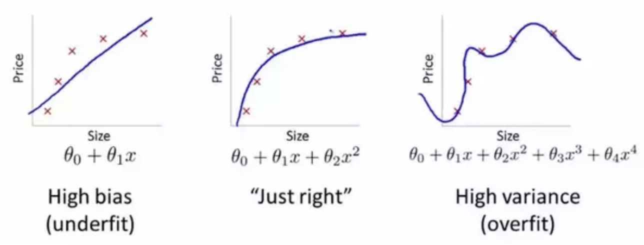


Figure: When model complexity is high, it is easy to get over-fitted.

How to avoid over-fitting?

- Regularization.

Picture Source: https://www.quora.com/What-is-the-best-way-to-explain-the-bias-variance-trade-off-in-layman

Regularization

Why Regularization?

With regularization, we

- incorporate prior subject knowledge, e.g., structure, smoothness.
- stabilize the estimates.

Prior desired property when we have a high-dimensional model:

- Can we recover a low-dimensional structure within the high-dimensional parameter space that can represent the data?

High dimensional model estimation (sparse estimation) is challenging both theoretically and algorithmically.

Optimal Behavior in High Dimensional Model Estimation:

1. Accuracy in estimation.
2. Fast computation.
3. Optimality of theoretical properties.
   (Optimal rate of convergence in estimation error & Structure recovery)
4. Ability to quantify model uncertainty.

Ideally, we want to achieve all the optimal behaviors simultaneously.

High dimensional model estimation (sparse estimation) is challenging both theoretically and algorithmically.

Optimal Behavior in High Dimensional Model Estimation:

1. Accuracy in estimation.
2. Fast computation.
3. Optimality of theoretical properties.
   (Optimal rate of convergence in estimation error & Structure recovery)
4. **Ability to quantify model uncertainty.**

Model uncertainty is particularly important in automated decision areas,such as medical diagnosis and self-driving cars where the safety of AI mechanisms is critical.

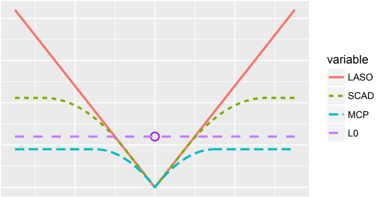
Review on Regularization Approaches: Penalized Likelihood

Penalized Likelihood Framework

Let Z be samples drawn from a distribution, the regularization framework have the following form:

$$\underbrace{\hat{\theta}_\lambda}_{\text{Estimate}} \in \arg\min_{\theta \in \Omega} \Big\{ \underbrace{-\log f(\theta; Z)}_{\text{Loss function}} + \underbrace{\mathcal{R}_\lambda(\theta)}_{\text{Penalty function}} \Big\}$$

Popular forms of $\mathcal{R}_\lambda(\theta)$ include:



Review on Regularization Approaches: Bayesian Approaches

Bayesian Framework

Suppose $\boldsymbol{\theta}$ follows a prior distribution $\pi(\boldsymbol{\theta})$ and data given on parameter $\boldsymbol{\theta}$ is generated from $f(\cdot)$:

$$\boldsymbol{\theta} \sim \pi(\cdot),$$

$$Data|\boldsymbol{\theta} \sim f(\cdot).$$

Goal: estimate $\boldsymbol{\theta}$ through the posterior distribution $\boldsymbol{\theta}|Data$.

- Pros: modeling is flexible; model uncertainty is naturally quantified; parameter inference is flexible, e.g., from posterior mode (MAP estimator), mean.
- Cons: to explore the whole posterior distribution (through MCMC methods), computation cost is high.

Connection between Bayesian and Penalized Likelihood Perspective

Estimating the MAP estimate of $\boldsymbol{\theta}$ is equivalent to minimizing the following objective function :

$$\underset{\boldsymbol{\theta}}{\operatorname{argmin}} \mathcal{L}(\boldsymbol{\theta}),$$

where

$$\begin{aligned}
\mathcal{L}(\boldsymbol{\theta}) &= -\log \pi(\boldsymbol{\theta}|Data) \\
&= -\log \left(f(Data|\boldsymbol{\theta})\pi(\boldsymbol{\theta}) / \int f(Data|\boldsymbol{\theta})\pi(\boldsymbol{\theta}) d\boldsymbol{\theta} \right) \\
&= -\log f(\boldsymbol{\theta}; Data) - \log \pi(\boldsymbol{\theta}) + \log \int f(\boldsymbol{\theta}; Data)\pi(\boldsymbol{\theta}) d\boldsymbol{\theta} \\
&\propto -\log f(\boldsymbol{\theta}; Data) + (\underbrace{-\log \pi(\boldsymbol{\theta})}_{\text{Bayesian-induced penalty}}).
\end{aligned}$$

Spike and Slab Prior: Uncover the Sparse Structure

Suppose the parameter $\boldsymbol{\theta} = [\theta_1, ..., \theta_p]$.

Spike and Slab Lasso Prior

The cornerstone of our Bayesian formulation for sparse estimation is the following spike and slab prior on θ_i:

$$\begin{cases} \theta_i \mid r_i = 0 & \sim & f_1(\cdot) \Leftarrow \text{ spike part,} \\ \theta_i \mid r_i = 1 & \sim & f_2(\cdot) \Leftarrow \text{ slab part.} \end{cases}$$

where r_i follows

$$r_i \sim \text{Bern}(\eta).$$

Different forms of the spike and slab priors

- The original spike and slab prior:
  the spike part is a point mass and the slab part is a uniform distribution [Mitchell and Beauchamp, 1988].
- Spike and slab normal prior:
  the spike and slab parts are all Gaussian distributions and variance for the slab prior is larger [George and McCulloch, 1997, Ishwaran and Rao, 2005].
- Spike and slab Lasso prior:
  the spike and slab parts are all Laplace distributions and variance for the slab prior is larger
  [Ročková and George, 2014, Ročková, 2016, Ročková and George, 2016].

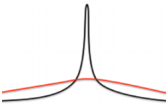


Figure: An illustration of spike and slab prior

Bayesian Regularization Framework

Specification

Our model formulation is given by:

$$Data|\boldsymbol{\theta} \overset{iid}{\sim} f(\cdot).$$

$$\theta_i \sim \eta \text{LP}(0, v_1) + (1-\eta)\text{LP}(0, v_0).$$

Goal:

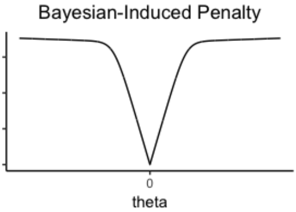
- Maximum a posteriori (MAP) estimate of $\boldsymbol{\theta}$.
- The posterior inclusion probability of $r_i|\cdot$.

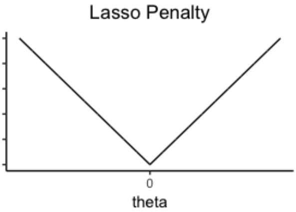
Bayesian Regularization Penalty

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

$$\begin{aligned}\text{pen}_{SS}(\theta_i) &= -\log \int \pi(\theta_i|r_i)\pi(r_i|\eta)dr_i \\ &= -\log\left[\left(\frac{\eta}{2v_1}\right)e^{-\frac{|\theta_i|}{v_1}} + \left(\frac{1-\eta}{2v_0}\right)e^{-\frac{|\theta_i|}{v_0}}\right],\end{aligned}$$

which is a **non-convex penalty**.





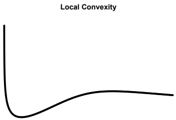
Pros & Cons of Non-convex Penalties

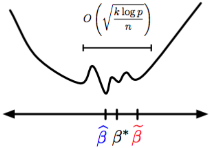
- Pros: lead to desired shrinkage and selection behavior.
- Cons: could bring additional computation and theoretical challenges because the objective function could be non-convex.

Our Findings

If we constrain the parameter to be considered in a reasonable large space:

- for Gaussian Graphical Model(discussed later), our optimization is strongly convex with a unique optimal.
- for Gaussian conditional random field, estimation error for all stationary points are bounded.





Right Figure Credit to: https://www.math.wustl.edu/~kuffner/WHOA-PSI-2/LohSlides.pdf

Efficient Computation

- Utilizing the mixture distribution structure of the prior, we propose an efficient EM algorithm that computes MAP estimate and posterior probabilities simultaneously.
- The computation cost is the same as computing the state-of-the-art Lasso estimator for the same model.

Table of Contents

Gaussian Graphical Model

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

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

- Our target is to estimate Θ and also obtain an estimate of its support.
- It is challenging particularly in high dimensional settings, e.g., when $p > n$, the sample covariance matrix is even not invertible.

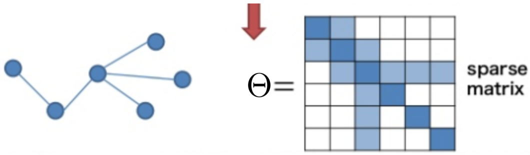
Gaussian Graphical Model

A Well-known Fact:

- Consider Undirected graph G=(V,E) with V is the vertex set and E is the edge set.
- When Y is multivariate Gaussian, no edge between $(Y^{(i)}, Y^{(j)})$ $\Leftrightarrow Y^{(i)} \perp\!\!\!\perp Y^{(j)} | Y^{-(i,j)} \Leftrightarrow \theta_{i,j} = 0.$[a]

[a] $Y^{(i)}$ is the i-th entry in Y. $Y = (Y^{(1)}, ..., Y^{(p)})$.

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



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.,

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

- The penalty function $\mathrm{pen}(\theta_{ij})$ is often taken to be Lasso [Yuan and Lin, 2007, Banerjee et al., 2008, Friedman et al., 2008],
- SCAD, which is a non-convex penalty, has also been used [Fan et al., 2009].
- Theoretical results: estimation errors in Frobenius norm have been studied in [Rothman et al., 2008, Lam and Fan, 2009]; support recovery and estimation errors in ℓ_∞ norm have been studied in [Ravikumar et al., 2011, Loh and Wainwright, 2014].

Literature Review

Motivation for the Sparse Regression Framework

For $1 \leq i \leq p$, $Y^{(i)}$ is expressed as $Y^{(i)} = \sum_{j \neq i} \beta_{ij} Y^{(j)} + \epsilon_i$ such that ϵ_i is uncorrelated with $Y^{-(i)}$ if and only if $\beta_{ij} = -(\theta_{ij}/\theta_{ii})$. Moreover, for such defined β_{ij}, $var(\epsilon_i) = (1/\theta_{ii})$, $cov(\epsilon_i, \epsilon_j) = \theta_{ij}/(\theta_{ii}\theta_{jj})$.

Regression

- In sparse regression framework, every $Y^{(i)}$ is regressed on the other variables $Y^{-(i)}$ and the coefficients are estimated jointly in a sparse way.
- Implicitly, they are modeling with under the likelihood $\prod_i P(Y^{(i)}|Y^{-(i)})$, instead of $P(Y)$.[a] [Meinshausen and Bühlmann, 2006, Peng et al., 2009]

[a] Denote $Y = (Y^{(1)}, \cdots, Y^{(p)})$.

Literature Review

Other Work

CLIME estimator [Cai et al., 2011]:
Let $\hat{\Theta}$ be the solution set of the following optimization problem:

$$\min ||\Theta||_1 \text{ subject to:}$$
$$|S\Theta - I| \leq \lambda.$$

Find matrix with the smallest ℓ_1 norm within a local neighborhood around the inverse sample covariance matrix.

Literature Review

Bayesian Methods

- Several Bayesian approaches have also been proposed.
- Prior specification:
  Laplace priors [Wang, 2012]; G-Wishart priors [Carvalho and Scott, 2009, Dobra et al., 2011, Wang and Li, 2012, Mohammadi et al., 2015]; Mixture prior distributions that have a point-mass and a Laplace distribution [Banerjee and Ghosal, 2015].
- Pros: a natural way to quantify uncertainty.
- Cons: slow in computation because of the high computational cost of sampling methods.

Our contributions

1. Propose a new approach for precision matrix estimation using the Bayesian regularization framework.
2. With mild conditions, we show the optimal estimation error rate in the ℓ_∞ norm and selection consistency under both exponential and polynomial tail distributions.
3. A fast EM algorithm which produces the MAP estimate of the precision matrix and (approximate) posterior probabilities on all the edges is proposed.

Model Specification

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

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

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

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

Our target is the MAP estimate of Θ and the posterior inclusion probability of $r_{ij}|\cdot$. We restrict the support of the parameter Θ in the posterior to satisfy $\|\Theta\|_2 \leq B$.

- The multivariate Gaussian distribution is a "working" likelihood for our inference. Performance is also guaranteed for the observations with non-Gaussian distributions including those with polynomial tails or exponential tails (more on this later).

Penalized Likelihood Perspective

This is equivalent to minimizing the following objective function under the **constraint** $\|\Theta\|_2 \leq B$ and $\Theta \succ 0$:

$$\begin{aligned}\mathcal{L}(\Theta) &= -\log \pi(\Theta | Y_1, \cdots, Y_n) \\ &= \frac{n}{2}\Big(\text{tr}(S\Theta) - \log\det(\Theta)\Big) + \sum_{i<j} \text{pen}_{SS}(\theta_{ij}) + \sum_{i} \text{pen}_1(\theta_{ii})\end{aligned}$$

where $\text{pen}_1(\theta) = \tau|\theta|$.

We call our method **BAGUS**, short for Bayesian Regularization for Graphical Models with Unequal Shrinkage.


BAGUS!

Posterior Maximization and Local Convexity

Pros & Cons of Non-convex Penalties

- Pros: lead to desired shrinkage and selection behavior.
- Cons: could bring additional computation challenges and may have multiple local optima as the objective function could be no longer convex .

Theorem (Local Convexity)

If $B \leq (2nv_0)^{\frac{1}{2}}$*, then* $\min_{\Theta \succ 0, \|\Theta\|_2 \leq B} \mathcal{L}(\Theta)$ *is a strictly convex problem.*

Even though the penalty is non-convex, we are dealing with convex optimization and it results in a unique MAP estimate.

Assumptions

(A1) $\lambda_{\max}(\Theta^0) \leq 1/k_1 < \infty$ or equivalently $0 < k_1 \leq \lambda_{\min}(\Sigma^0)$.

(A2) The minimal "signal" entry satisfies $\min_{(i,j)\in S_g} |\theta_{ij}^0| \geq K_0\sqrt{\frac{\log p}{n}}$, where $K_0 > 0$ is a sufficiently large constant not depending on n.

Rate of Convergence

Assume condition (A1) holds. For any pre-defined constants $C_3 > 0$, $\tau_0 > 0$, when the exponential tail (C1) or the polynomial tail (C2) condition holds. Assume that:

i) v_0, v_1, η, and τ satisfied certain conditions (shown in Appendix);

ii) the spectral norm B satisfies $\frac{1}{k_1} + 2d(C_1 + C_3)M_{\Gamma^0}\sqrt{\frac{\log p}{n}} < B < (2nv_0)^{\frac{1}{2}}$,

iii) the sample size n satisfies $\sqrt{n} \geq M\sqrt{\log p}$,

Theorem (Estimation Accuracy in Entrywise ℓ_∞ Norm)

Then, the MAP estimator $\tilde{\Theta}$ satisfies

$$\|\tilde{\Theta} - \Theta^0\|_\infty \leq 2(C_1 + C_3)M_{\Gamma^0}\sqrt{\frac{\log p}{n}}.$$

with probability greater than $1 - \delta_1$, where $\delta_1 = 2p^{-\tau_0}$ when condition (C1) holds, and $\delta_1 = O(n^{-\delta_0/8} + p^{-\tau_0/2})$ when condition (C2) holds.

Theorem (Selection Consistency)

Assume the same conditions in previous Theorem and condition (A2) with the following restriction:

$$\epsilon_0 < \frac{1}{\log p} \log \left(\frac{v_1(1-\eta)}{v_0 \eta} \right) < (C_4 - C_3)\big(K_0 - 2(C_1 + C_3)K_{\Gamma^0}\big)$$

for some arbitrary small constant $\epsilon_0 > 0$*. Then, for any* T *such that* $0 < T < 1$*, we have*

$$\mathbb{P}\left(\hat{S}_0 = S_0\right) \to 1.$$

Comparison with Existing Results

- **Graphical Lasso [Ravikumar et al., 2011]:**
  Graphical Lasso assumes the relatively restrictive irrepresentable condition, $|||\Gamma_{S_g^c S_g}\Gamma^{-1}_{S_g S_g}|||_\infty \leq 1-\alpha$.
- Under the polynomial tail condition, the rate of convergence for Graphical Lasso is $O_p\left(\sqrt{\frac{p^c}{n}}\right)$, slower than our rate $O_p\left(\sqrt{\frac{\log p}{n}}\right)$.
- **CLIME [Cai et al., 2011] :**
  CLIME assumes the boundedness of $|||\Theta^0|||_1$, which is strictly stronger than our condition on the largest eigenvalue.
- **Non-convex Penalties like SCAD, MCP [Loh and Wainwright, 2014] :**
- They require beta-min condition (minimal signal strength to be greater than some threshold) for the results on estimation error.
- Their results are only available for sub-Gaussian distributions.

EM Algorithm

- We treated r_{ij} 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_{ij}, denoted as p_{ij}, in the E-step.
- E-step: compute the posterior distribution of r_{ij}.
- M-step: optimize the following optimization problem:

$$\underset{\Theta \succ 0, ||\Theta||_2 \leq B}{\operatorname{argmin}} \left(\mathcal{L}(\Theta) + \sum_{i,j} \lambda(\theta_{ij}) |\theta_{ij}| \right), \tag{2}$$

where $\lambda(\theta_{ij}) = \frac{p_{ij}}{v_1} + \frac{1-p_{ij}}{v_0}$.

M-step

- The updating scheme is in the fashion of updating one column and one row at a time. Similar strategy has been used in [Friedman et al., 2008] and [Mazumder and Hastie, 2012].
- Without loss of generality, we describe the updating rule for the last column of Θ while fixing the others.

We list the following equalities from $W\Theta = \mathbf{I}_p$ which will be used in our algorithm:

$$\begin{bmatrix} W_{11} & w_{12} \\ \cdot & w_{22} \end{bmatrix} = \begin{bmatrix} \Theta_{11}^{-1} + \frac{\Theta_{11}^{-1}\theta_{12}\theta_{12}^T\Theta_{11}^{-1}}{\theta_{22}-\theta_{12}^T\Theta_{11}^{-1}\theta_{12}} & -\frac{\Theta_{11}^{-1}\theta_{12}}{\theta_{22}-\theta_{12}^T\Theta_{11}^{-1}\theta_{12}} \\ \cdot & \frac{1}{\theta_{22}-\theta_{12}^T\Theta_{11}^{-1}\theta_{12}} \end{bmatrix}. \quad (3)$$

M-step

Given Θ_{11}, we will update the last column $(\theta_{12}, \theta_{22})$. To do that, we set the subgradient of Q with respect to $(\theta_{12}, \theta_{22})$ to zero.
First take the subgradient of Q with respect to θ_{22}:

$$\frac{\partial Q}{\partial \theta_{22}} = \frac{n}{2} \frac{1}{\theta_{22} - \theta_{12}^T \Theta_{11}^{-1} \theta_{12}} - \frac{n}{2} (s_{22} + \tau) = 0. \quad (4)$$

Due to Equations (3) and (4), we have

$$w_{22} = \frac{1}{\theta_{22} - \theta_{12}^T \Theta_{11}^{-1} \theta_{12}} = s_{22} + \frac{2}{n}\tau,$$

which leads to the following update for θ_{22}:

$$\theta_{22} \leftarrow \frac{1}{w_{22}} + \theta_{12}^T \Theta_{11}^{-1} \theta_{12}. \quad (5)$$

M-step

Next take the subgradient of Q with respect to θ_{12}:

$$\begin{aligned}\frac{\partial Q}{\partial \theta_{12}} =& \frac{n}{2}\left(\frac{-2\Theta_{11}^{-1}\theta_{12}}{\theta_{22}-\theta_{12}^T\Theta_{11}^{-1}\theta_{12}} - 2s_{12}\right) - \left(\frac{1}{v_1}p_{12} + \frac{1}{v_0}(1-p_{12})\right) \odot \mathsf{sign}(\theta_{12}) \\ =& n(-\Theta_{11}^{-1}\theta_{12}w_{22} - s_{12}) - \left(\frac{1}{v_1}p_{12} + \frac{1}{v_0}(1-p_{12})\right) \odot \mathsf{sign}(\theta_{12}) = 0,\end{aligned} \tag{6}$$

where $\odot$ denotes element-wise multiplication. The second line of (6) is due to the identities in (3).

To update θ_{12}, we solve the following stationary equation with coordinate descent, under the constraint $\|\Theta\|_2 \leq B$:

$$ns_{12} + nw_{22}\Theta_{11}^{-1}\theta_{12} + \left(\frac{1}{v_1}P_{12} + \frac{1}{v_0}(1 - P_{12})\right) \odot \text{sign}(\theta_{12}) = 0. \quad (7)$$

Algorithm 1 Coordinate Descent for θ_{12}

1: **Initialize** θ_{12} from the previous iteration as the starting point.
2: **repeat**
3: **for** j in $1 : (p-1)$ **do**
4: Solve the following equation for θ_{12_j}:

$$ns_{12_j} + nw_{22}{\Theta_{11}^{-1}}_{j,\setminus j}\theta_{12\setminus j} + nw_{22}{\Theta_{11}^{-1}}_{j,j}\theta_{12_j} + \left[\left(\frac{1}{v_1}P_{12} + \frac{1}{v_0}(1 - P_{12})\right) \odot \text{sign}(\theta_{12})\right]_j$$

5: **end for**
6: **until** Converge or Max Iterations Reached.
7: If $\|\Theta\|_2 > B$: **Return** θ_{12} from the previous iteration
8: Else: **Return** θ_{12}

Algorithm 2 BAGUS

Initialize $W = \Theta = \mathbf{I}$
repeat
  Update P with each entry p_{ij} updated as $\log \frac{p_{ij}}{1-p_{ij}} \leftarrow \Big(\log \frac{v_0}{v_1} + \log \frac{\eta}{1-\eta} - \frac{|\theta_{ij}^{(t)}|}{v_1} + \frac{|\theta_{ij}^{(t)}|}{v_0} \Big)$.
  for j in $1 : p$ **do**
    Move the j-th column and j-th row to the end (implicitly), namely $\Theta_{11} := \Theta_{-j-j}$, $\theta_{12} := \theta_{-jj}$, $\theta_{22} := \theta_{jj}$
    Update w_{22} using $w_{22} \leftarrow s_{22} + \frac{2}{n}\tau$
    Update θ_{12} by solving (7) with Coordinate Descent for θ_{12}.
    Update θ_{22} using $\theta_{22} \leftarrow \frac{1}{w_{22}} + \theta_{12}^T \Theta_{11}^{-1} \theta_{12}$.
    Update W **using (3).**
  end for
until Converge
Return Θ, P

Properties of the Algorithm

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

Theorem (Positive Definiteness & Symmetry)

- *The estimate of Θ is always guaranteed to be symmetric.*
- *If $\Theta^{(0)} > 0$, i.e the initial estimate of precision matrix is positive definite, then $\Theta^{(t)} > 0$, $\forall t \geq 1$.*

For well-known algorithms including Graphical Lasso [Friedman et al., 2008], SPACE [Peng et al., 2009], CLIME [Cai et al., 2011], the positive definiteness is not guaranteed [Mazumder and Hastie, 2012].

Simulation Studies

1. Model 1: An star model with $w_{ii} = 1$ and $w_{1,i} = 1$, $w_{i,1} = 1/\sqrt{p}$ for $i \neq 1$.
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 Edge Model.

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

Metrics

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

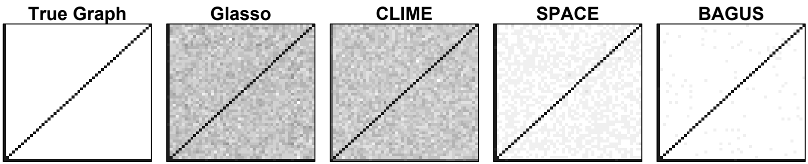


Figure: Average of the estimated precision matrices for the model with the star structure

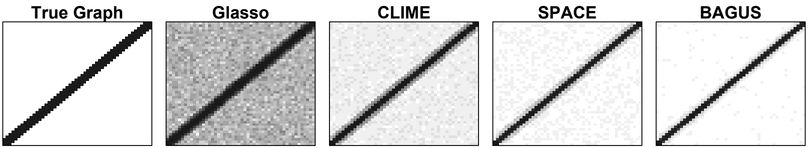


Figure: Average of the estimated precision matrices for the model with the AR(2) structure

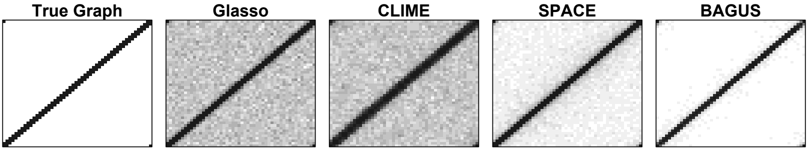


Figure: Average of the estimated precision matrices for the model with the **circle structure**

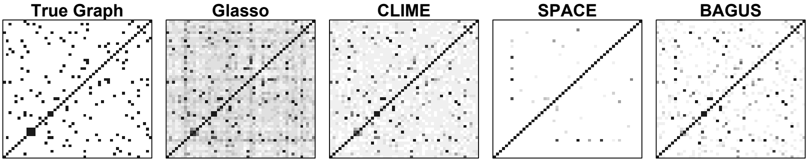


Figure: Average of the estimated precision matrices for the model with the **random structure**

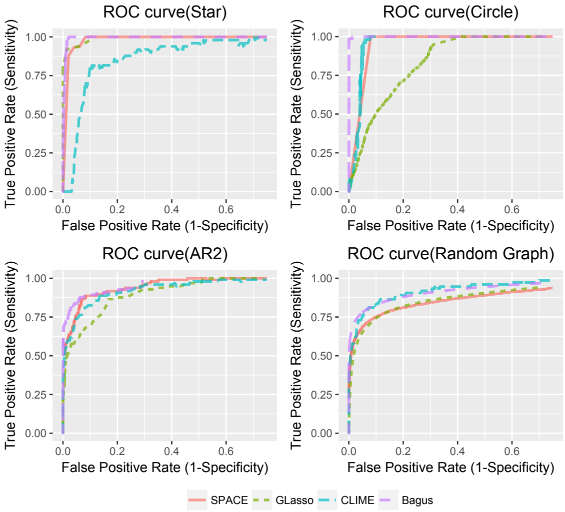


Figure: ROC Curves for different methods and different data generating models with $p = 50$.

Table: Model 1: Star

		$n = 100, p = 50$		
	Fnorm	Specificity	Sensitivity	MCC
GLasso	2.301(0.126)	0.687(0.015)	0.998(0.004)	0.339(0.011)
CLIME	3.387(0.401)	0.452(0.051)	0.971(0.023)	0.168(0.021)
SPACE	2.978(0.244)	0.972(0.039)	1.000(0.003)	0.824(0.163)
BAGUS	**1.053(0.107)**	1.000(0.000)	1.000(0.000)	**1.000(0.000)**
		$n = 100, p = 100$		
	Fnorm	Specificity	Sensitivity	MCC
GLasso	4.219(0.118)	0.715(0.007)	0.989(0.008)	0.260(0.005)
CLIME	4.818(0.449)	0.998(0.004)	0.336(0.000)	0.131(0.067)
SPACE	3.207(0.311)	0.987(0.022)	0.996(0.024)	0.842(0.162)
BAGUS	**1.499(0.138)**	1.000(0.000)	1.000(0.000)	**1.000(0.000)**
		$n = 100, p = 200$		
	Fnorm	Specificity	Sensitivity	MCC
GLasso	3.028(0.068)	0.947(0.003)	0.999(0.002)	0.389(0.009)
CLIME	5.595(0.528)	0.978(0.018)	0.000(0.000)	-0.014(0.006)
SPACE	3.735(0.294)	0.985(0.007)	1.000(0.000)	0.656(0.138)
BAGUS	**2.006(0.100)**	1.000(0.000)	1.000(0.001)	**1.000(0.001)**

Table: Model 2: $AR(2)$

$n = 100, p = 50$				
	Fnorm	Specificity	Sensitivity	MCC
GLasso	**3.361(0.240)**	0.479(0.056)	0.981(0.015)	0.251(0.028)
CLIME	3.758(0.381)	0.822(0.054)	0.906(0.039)	0.472(0.053)
SPACE	5.903(0.070)	0.982(0.004)	0.608(0.038)	0.656(0.029)
BAGUS	3.671(0.291)	0.997(0.002)	0.551(0.032)	**0.707(0.025)**
$n = 100, p = 100$				
	Fnorm	Specificity	Sensitivity	MCC
GLasso	8.130(0.035)	0.901(0.007)	0.745(0.028)	0.382(0.017)
CLIME	5.595(1.578)	0.837(0.075)	0.821(0.191)	0.371(0.085)
SPACE	9.819(0.083)	0.991(0.002)	0.566(0.025)	0.625(0.021)
BAGUS	**5.330(0.369)**	0.998(0.001)	0.549(0.018)	**0.707(0.022)**
$n = 100, p = 200$				
	Fnorm	Specificity	Sensitivity	MCC
GLasso	11.728(0.045)	0.990(0.001)	0.478(0.017)	0.481(0.014)
CLIME	11.552(0.382)	0.989(0.004)	0.580(0.031)	0.539(0.028)
SPACE	13.696(0.079)	0.995(0.000)	0.518(0.018)	0.588(0.013)
BAGUS	**8.214(0.548)**	0.998(0.001)	0.543(0.015)	**0.677(0.027)**

Table: Model 3: Circle

		$n = 100, p = 50$		
	Fnorm	Specificity	Sensitivity	MCC
GLasso	**4.319(0.174)**	0.492(0.064)	1.000(0.000)	0.196(0.024)
CLIME	5.785(0.440)	0.555(0.026)	1.000(0.000)	0.221(0.010)
SPACE	19.402(0.232)	0.930(0.006)	1.000(0.000)	0.595(0.019)
BAGUS	**4.253(0.578)**	0.993(0.004)	0.964(0.029)	**0.903(0.049)**
		$n = 100, p = 100$		
	Fnorm	Specificity	Sensitivity	MCC
GLasso	6.981(0.192)	0.647(0.005)	1.000(0.000)	0.189(0.002)
CLIME	19.282(2.802)	0.224(0.226)	0.995(0.015)	0.069(0.058)
SPACE	27.737(0.345)	0.975(0.010)	0.994(0.008)	0.674(0.062)
BAGUS	**6.012(0.513)**	0.996(0.002)	0.957(0.032)	**0.895(0.055)**
		$n = 100, p = 200$		
	Fnorm	Specificity	Sensitivity	MCC
GLasso	**7.664(0.209)**	0.752(0.003)	1.000(0.000)	0.172(0.001)
CLIME	33.009(0.535)	0.857(0.154)	0.769(0.167)	0.209(0.052)
SPACE	32.142(0.832)	0.981(0.012)	0.783(0.212)	0.485(0.129)
BAGUS	10.378(1.001)	0.995(0.001)	0.886(0.033)	**0.752(0.028)**

Table: Model 4: Random Graph

		$n = 100, p = 50$		
	Fnorm	Specificity	Sensitivity	MCC
GLasso	7.017(0.256)	0.877(0.010)	0.766(0.039)	0.417(0.027)
CLIME	11.347(0.452)	0.971(0.012)	0.614(0.068)	0.572(0.042)
SPACE	12.278(0.183)	1.000(0.000)	0.073(0.031)	0.257(0.051)
BAGUS	**5.811(0.357)**	0.999(0.001)	0.443(0.032)	**0.637(0.027)**
		$n = 100, p = 100$		
	Fnorm	Specificity	Sensitivity	MCC
GLasso	11.851(0.900)	0.837(0.047)	0.720(0.049)	0.285(0.033)
CLIME	12.649(1.587)	0.735(0.153)	0.761(0.120)	0.243(0.123)
SPACE	17.706(0.203)	1.000(0.000)	0.068(0.015)	0.236(0.028)
BAGUS	**8.754(0.366)**	0.999(0.001)	0.400(0.022)	**0.598(0.022)**
		$n = 100, p = 200$		
	Fnorm	Specificity	Sensitivity	MCC
GLasso	15.054(0.356)	0.951(0.012)	0.633(0.029)	0.307(0.017)
CLIME	23.568(0.954)	0.993(0.004)	0.469(0.048)	0.492(0.038)
SPACE	24.997(0.213)	0.999(0.000)	0.090(0.014)	0.221(0.024)
BAGUS	**13.096(0.522)**	0.999(0.000)	0.382(0.050)	**0.565(0.032)**

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.

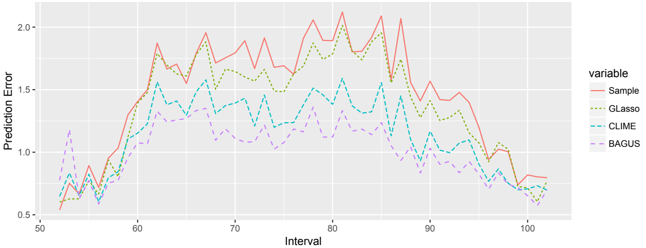


Figure: Prediction Error for the call center cata: AAFE_t on Y axis and t on X axis.

	Average Prediction Error					
	Sample	GLasso	Adaptive Lasso	SCAD	CLIME	BAGUS
Average AAFE	1.46	1.38	1.34	1.31	1.14	**1.00**

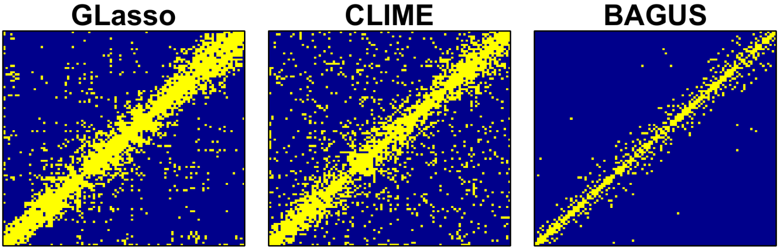


Figure: Sparsity structures estimated for different methods for the call center data

- The estimated structure from BAGUS is the most sparse one.
- Even with a sparse model, average prediction error for BAGUS is the smallest.

Conclusion

1. Propose a new approach for precision matrix estimation, named BAGUS, with Bayesian Regularization.
2. Both numerically and theoretically, the Bayesian regularization method we proposed works very well.

Table of Contents

Conclusion

- We have observed promising results of Bayesian regularization in various models we studies.
- Hope its success demonstrated in our work will motivate further interest in this direction.

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 l1 minimization approach to sparse precision matrix estimation.
Journal of the American Statistical Association, 106(494):594–607.

Carvalho, C. M. and Scott, J. G. (2009).
Objective bayesian model selection in gaussian graphical models.
Biometrika, 96(3):497–512.

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

Dobra, A., Lenkoski, A., and Rodriguez, A. (2011).
Bayesian inference for general gaussian graphical models with application to multivariate lattice data.
Journal of the American Statistical Association, 106(496):1418–1433.

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.

References II

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

George, E. I. and McCulloch, R. E. (1997).
Approaches for Bayesian variable selection.
Statistica Sinica, pages 339–373.

Ishwaran, H. and Rao, J. S. (2005).
Spike and slab variable selection: Frequentist and Bayesian strategies.
Annals of Statistics, 33:730–773.

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

Loh, P.-L. and Wainwright, M. J. (2014).
Support recovery without incoherence: A case for nonconvex regularization.
arXiv preprint arXiv:1412.5632.

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.

References III

Mitchell, T. J. and Beauchamp, J. J. (1988).
Bayesian variable selection in linear regression.
Journal of the American Statistical Association, 83(404):1023–1032.

Mohammadi, A., Wit, E. C., et al. (2015).
Bayesian structure learning in sparse Gaussian graphical models.
Bayesian Analysis, 10:109–138.

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.

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.

Ročková, V. and George, E. I. (2014).
EMVS: The EM approach to Bayesian variable selection.
Journal of the American Statistical Association, 109(506):828–846.

Ročková, V. and George, E. I. (2016).
The spike-and-slab lasso.
Journal of the American Statistical Association, (just-accepted).

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.

References IV

Ročková, V. (2016).
Bayesian estimation of sparse signals with a continuous spike-and-slab prior.
Annals of Statistics, (just-accepted).

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

Wang, H. and Li, S. (2012).
Efficient Gaussian graphical model determination under G-Wishart prior distributions.
Electronic Journal of Statistics, 6:168–198.

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

Table of Contents

The following theorem gives estimation accuracy under the entrywise ℓ_∞ norm. In particular, the following theorem implies that with an appropriate choice of v_0, v_1, η and τ, we could achieve the $O_p\left(\sqrt{\frac{\log p}{n}}\right)$ error rate for distributions with an exponential or a polynomial tail.

Notation

- For a $p \times p$ matrix $A = [a_{ij}]$, we denote its spectral norm by $\|A\|_2 = \lambda_{\max}(A)$, $|||A|||_\infty = \max_{1 \le j \le q} \sum_{i=1}^{p} |a_{ij}|$.
- Let $\Theta^0 = [\theta^0_{ij}]$ and $\Sigma^0 = [\sigma^0_{ij}]$ denote the true precision matrix and covariance matrix.
- Let $S^0 = \{(i,j) : \theta^0_{ij} \neq 0\}$ denote the index set of all nonzero entries in Θ^0 and S^{0^c} is its complement.
- Define $M_{\Sigma^0} = \left|\left|\left|\Sigma^0\right|\right|\right|_\infty$.
- Define $\Gamma = \Theta^{-1} \otimes \Theta^{-1}$ as the Hessian matrix of $g := -\log\det(\Theta)$. We further denote $M_{\Gamma^0} = \left|\left|\left|{\Gamma^0}^{-1}_{S^0 S^0}\right|\right|\right|_\infty = \left|\left|\left|(\Theta^0 \otimes \Theta^0)_{S^0 S^0}\right|\right|\right|_\infty$.
- Define the column sparsity $d = \max_{i=1,2,\ldots,p} card\{j : \theta^0_{ij} \neq 0\}$ and the off-diagonal sparsity $s = card(S^0) - p$, where $card$ denotes the cardinality of the set in its argument.

Assumptions

(A1) $\lambda_{\max}(\Theta^0) \le 1/k_1 < \infty$ or equivalently $0 < k_1 \le \lambda_{\min}(\Sigma^0)$.

(A2) The minimal "signal" entry satisfies $\min_{(i,j)\in S_g} |\theta^0_{ij}| \ge K_0 \sqrt{\frac{\log p}{n}}$, where $K_0 > 0$ is a sufficiently large constant not depending on n.

Tail Conditions

(C1) Exponential tail condition: Suppose that there exists some $0 < \eta_1 < 1/4$ such that $\frac{\log p}{n} < \eta_1$ and

$$Ee^{tY^{(j)^2}} \le K \text{ for all } |t| \le \eta_1, \text{ for all } j = 1, \dots, p$$

where K is a bounded constant.

(C2) Polynomial tail condition: Suppose that for some γ, $c_1 > 0$, $p \le c_1 n^\gamma$, and for some $\delta_0 > 0$,

$$E|Y^{(j)}|^{4\gamma+4+\delta_0} \le K, \text{ for all } j = 1, \dots, p.$$

Theorem

(Estimation accuracy in entrywise ℓ_∞ norm)
Assume condition (A1) holds. For any pre-defined constants $C_3 > 0$, $\tau_0 > 0$, define $C_1 = \eta_1^{-1}(2 + \tau_0 + \eta_1^{-1} K^2)$ when the exponential tail condition (C1) holds, and $C_1 = \sqrt{(\theta^0_{\max} + 1)(4 + \tau_0)}$ when the polynomial tail condition (C2) holds. Assume that
i) the prior hyper-parameters v_0, v_1, η, and τ satisfy

$$\begin{cases} \frac{1}{nv_1} = C_3\sqrt{\frac{\log p}{n}}(1-\varepsilon_1), \quad \frac{1}{nv_0} > C_4\sqrt{\frac{\log p}{n}} \\ \frac{v_1^2(1-\eta)}{v_0^2\eta} \le p^{\varepsilon}, \quad \text{and} \quad \tau \le C_3 \frac{n}{2}\sqrt{\frac{\log p}{n}} \end{cases} \tag{8}$$

for some constants $\varepsilon_1 > 0$, $C_4 > C_3$ and some sufficiently small ε,
ii) the spectral norm B satisfies $\frac{1}{k_1} + 2d(C_1 + C_3)M_{\Gamma^0}\sqrt{\frac{\log p}{n}} < B < (2nv_0)^{\frac{1}{2}}$, and
iii) the sample size n satisfies $\sqrt{n} \ge M\sqrt{\log p}$,
where $M = \max\left\{2d(C_1 + C_3)M_{\Gamma^0}\max\left(3M_{\Sigma^0}, 3M_{\Gamma^0}M_{\Sigma^0}{}^3, \frac{2}{k_1^2}\right), \frac{2C_3\varepsilon_1}{k_1^2}\right\}$.
Then, the MAP estimator $\tilde{\Theta}$ satisfies

$$\|\tilde{\Theta} - \Theta^0\|_\infty \le 2(C_1 + C_3)M_{\Gamma^0}\sqrt{\frac{\log p}{n}}. \tag{9}$$

with probability greater than $1 - \delta_1$, where $\delta_1 = 2p^{-\tau_0}$ when condition (C1) holds, and $\delta_1 = O(n^{-\delta_0/8} + p^{-\tau_0/2})$ when condition (C2) holds.