Bayesian Modeling for Gaussian Conditional Random Fields

Lingrui Gan

Department of Statistics
University of Illinois at Urbana-Champaign

Joint work with Naveen Naidu Narisetty, Feng Liang

Table of Contents

Uncover the dependence structures between high-dimensional vectors

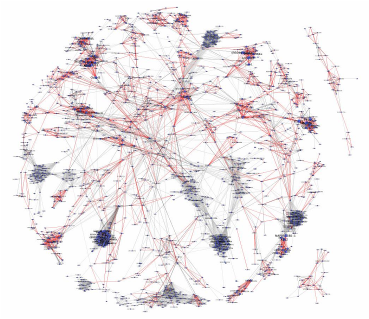


Image source: http://www.john.ranola.org/

- One of the canonical statistical problems is to understand the dependence structure between the variables of interest.

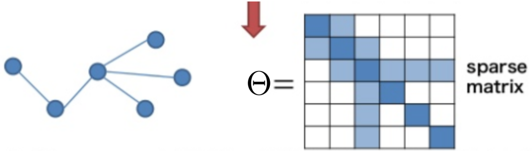
Gaussian graphical model and its limitations

A common tool we use is called Gaussian graphical model.

Gaussian graphical model

- $$Y = [Y_1, ..., Y_p] \overset{iid}{\sim} \mathsf{N}_p(0, \Theta^{-1}).$$
- $Y_i \perp\!\!\!\perp Y_j | Y_{[-i,-j]} \Leftrightarrow \Theta_{i,j} = 0 \Leftrightarrow$ No edge between (Y_i, Y_j).
- **Limitation:** We can only model the dependences within response Y.

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



What if we also have a set of covariates X?

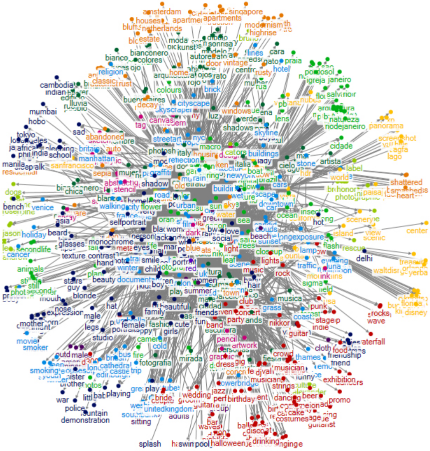


Figure: Flickr Tag Network

What if we also have a set of covariates X?

Scenario

- Gene expression data: one is interested in modeling genetic outcomes given biomarkers.
- Financial data (S&P 500): one is interested in modeling current asset prices given historical prices in portfolio analysis.
- Flickr data (MIRFlickr25k): one is interested in modeling scores of images with their text annotations.
- News data (RCV1-v2): one is interested in modeling the patterns of Reuters newswire stories given categories (Topics, Industries and Region).

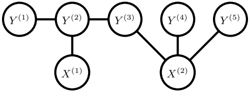
...

In these settings, we also care about the dependences between X and Y!

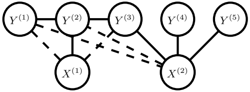
Existing approaches can lead to inappropriate dependences

Can we use some existing remedies?

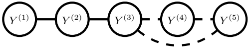
- GGM on (X,Y).
  Cons: computationally very expensive; large error.
- GGM on Y only. / Multivariate regression for $Y|X$. $\Rightarrow$ Inappropriate dependences.



(a) True graph.



(b) Graph of the dependence structure based on B and Λ from multivariate regression.



(c) Graph of the dependence structure from the marginal Gaussian graphical model on Y only.

Gaussian conditional random field

Model formulation

To address this setting, we propose to use the Gaussian conditional random field model in the following manner:

$$p(Y \mid X, \Lambda, \Theta) \propto \exp\left\{-\frac{1}{2}Y^T\Lambda Y - X^T\Theta Y\right\}, \tag{1}$$

where Λ is a $p \times p$ positive definite and symmetric matrix and $\Theta \in \mathbb{R}^{q\times p}$ is a matrix of dimension $q \times p$.

$$\begin{aligned} \Theta_{ij} = 0 \quad &\Longleftrightarrow \quad X^{(i)} \perp\!\!\!\perp Y^{(j)} \mid X^{-(i)}, Y^{-(j)}, \\ \Lambda_{ij} = 0 \quad &\Longleftrightarrow \quad Y^{(i)} \perp\!\!\!\perp Y^{(j)} \mid X, Y^{-(i,j)}, \end{aligned}$$

Literature review

- Gaussian conditional random field (GCRF) model with ℓ_1 penalty has been recently considered by several researchers [Sohn and Kim, 2012, Yuan and Zhang, 2014, Wytock and Kolter, 2013].
- Theoretical results on estimation accuracy have been established by [Yuan and Zhang, 2014] in Frobenius norm and by [Wytock and Kolter, 2013] in ℓ_∞ norm.

Our contributions

- We propose a Bayesian regularization method for the Gaussian conditional random field estimation.
- The optimal rate of convergence and sparsistent of our estimate is established under mild conditions. Our theoretical results are stronger than the ones on the Gaussian conditional random field with ℓ_1 penalty from [Yuan and Zhang, 2014] and [Wytock and Kolter, 2013].
- An efficient EM algorithm based on a second-order approximation method is proposed for computation.
- Our simulation studies and real application on asset return predictions demonstrate that the proposed Bayesian regularization approach provides a better performance compared to alternative methods.

Table of Contents

Bayesian regularization formulation

Prior specification

Our formulation is based on the following spike and slab Lasso prior [George and McCulloch, 1993, Ročková and George, 2014, Ročková, 2016, Ročková and George, 2016, Gan et al., 2018]:

$$\begin{cases} \pi_{\mathsf{SS}}(\theta)|r = 1 \sim \mathsf{DE}(\theta; v_1) \\ \pi_{\mathsf{SS}}(\theta)|r = 0 \sim \mathsf{DE}(\theta; v_0), \end{cases} \tag{2}$$

$$r \sim \mathsf{Bern}(\eta).$$

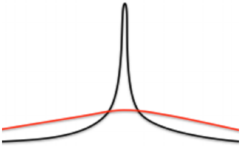


Figure: An illustration of the spike and slab prior.

- We place the spike and slab Lasso prior on all the entries of Θ and the upper triangular entries of Λ (due to symmetry), and place a Uniform prior on the diagonal entries of Λ:

$$\pi(\Theta, \Lambda) = \Big[\prod_{i,j} \pi_{\mathsf{SS}}(\Theta_{ij})\Big] \times \Big[\prod_{i<j} \pi_{\mathsf{SS}}(\Lambda_{ij})\Big] \times \Big[\prod_{i} \pi_{\mathsf{Unif}}(\Lambda_{ii})\Big].$$

- The support of the joint prior distribution is on the set $\{(\Theta, \Lambda) : \Lambda \succ 0, \|\Theta\|_1 + \|\Lambda\|_1 \leq R\}$, where $\Lambda \succ 0$ means that the matrix Λ is restricted to be positive definite.

We estimate (Θ, Λ) using the posterior mode.

MAP estimate

MAP estimate

Finding the MAP estimator of (Θ, Λ) is equivalent to solving the following optimization problem

$$\underset{\Lambda \succ 0, \Theta, \|\Theta\|_1 + ||\Lambda||_1 \le R}{\arg\min} L(\Theta, \Lambda), \tag{3}$$

where the negative log posterior can be written as

$$L(\Theta, \Lambda) = -\ell(\Theta, \Lambda) + \sum_{i<j} \text{pen}_{\text{SS}}(\Theta_{ij}) + \sum_{i,j} \text{pen}_{\text{SS}}(\Lambda_{ij}), \tag{4}$$

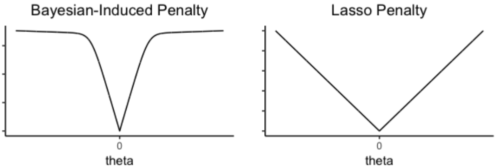
where $\ell(\cdot)$ is the $\log$-likelihood function:

$$\ell(\Theta, \Lambda) = \frac{n}{2}\Big(\log\det(\Lambda) - \text{tr}(S_{yy}\Lambda + 2S_{xy}\Theta + \Lambda^{-1}\Theta^T S_{xx}\Theta)\Big), \tag{5}$$

The spike and slab Lasso penalty

The Bayesian induced penalty $\text{pen}_{SS}(\cdot)$ is a **non-convex** penalty that takes the following form:

$$\text{pen}_{SS}(\theta) = -\log\left(\frac{\eta}{2v_1}e^{-\frac{|\theta|}{v_1}} + \frac{1-\eta}{2v_0}e^{-\frac{|\theta|}{v_0}}\right). \tag{6}$$



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

With our formulation, we found:

- estimation error for all stationary points are bounded in Frobenius norm.
- at least one stationary point is bounded in ℓ_∞ norm.

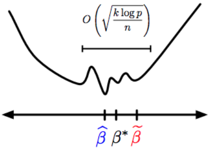


Figure Credit to: https://www.math.wustl.edu/ kuffner/WHOA-PSI-2/LohSlides.pdf

Table of Contents

Assumption

Assumption 1: In our theoretical analysis, we assume that the covariate vector X is from a random design with covariance matrix Σ_{xx}^0 and satisfies the following s_0-sparse restricted isometry property condition:

$$\begin{cases} \inf\left(\frac{u^T S_{xx} u}{u^T \Sigma_{xx}^0 u} : u \neq 0, ||u||_0 \leq s_0\right) \geq 0.5, \\ \sup\left(\frac{u^T S_{xx} u}{u^T \Sigma_{xx}^0 u} : u \neq 0, ||u||_0 \leq s_0\right) \leq 1.5, \\ \frac{\lambda_{\max}[(\Theta^0)^T S_{xx} \Theta^0]}{\lambda_{\max}[(\Theta^0)^T \Sigma_{xx}^0 \Theta^0]} \leq 1.4. \end{cases}$$

The same assumption is used in [Yuan and Zhang, 2014] for analyzing Gaussian conditional random field with the ℓ_1 penalty and is also frequently used in compressed sensing. It is also well known [Candes and Tao, 2007] that this condition holds with high probability when X is sub-Gaussian and n is sufficiently large.

Theorem (Rate of convergence for all stationary points)

Assume that Assumption 1 holds with $s_0 = |S_0| + \lceil 4(\rho_2/\rho_1)|S_0| \rceil$ *and that* X *is sub-Gaussian. If* $R \leq \frac{k_1}{6C_0}\sqrt{\frac{n}{c_0 \log(p+q)}}$ *and if the prior parameters satisfy*

$$\begin{cases} \frac{3}{4nv_0} < k_1, \\ 24 \max(C_1, k_1)\sqrt{\frac{c_0 \log(p+q)}{n}} \leq \frac{\lambda}{n} \leq C_0\sqrt{\frac{c_0 \log(p+q)}{n}}, \end{cases} \tag{7}$$

where C_0 *is some sufficiently large constant. Then for any stationary point* $\hat{\Phi}$ *of* (3)*, when the sample size* $n \geq 2c_0 \log(p+q)$ *for sufficiently large constant* $c_0 > 0$*, we have*

$$||\hat{\Phi} - \Phi^0||_1 \leq c_3|S_0|\sqrt{\frac{c_0 \log(p+q)}{n}}, ||\hat{\Phi} - \Phi^0||_F \leq c_4\sqrt{\frac{c_0|S_0| \log(p+q)}{n}},$$

with probability at least $1 - c_1 \exp(-c_2 \log(pq))$.

Theorem (Sparsistency for all stationary points)

Under the conditions given in Theorem 1, for all the local minimizers $\hat{\Phi}$ of (3), if $||\hat{\Phi} - \Phi^0||_2^2 = O_p(\eta_n)$ for a sequence $\eta_n \to 0$ and if $\sqrt{\log(p+q)/n + \eta_n} = O(\lambda/n)$, then with probability converging to 1, $\hat{\Phi}_{ij} = 0$ for all $(i,j) \in S_0^c$.

We present two scenarios making use of the inequalities $||\hat{\Phi} - \Phi||_F^2/p \leq ||\hat{\Phi} - \Phi||_2^2 \leq ||\hat{\Phi} - \Phi||_F^2$, and provide a sufficient condition on the sparsity level in each scenario to achieve sparsistency.

- When $||\hat{\Phi} - \Phi||_2^2 = ||\hat{\Phi} - \Phi||_F^2 = O_p\left(\frac{|S_0|}{n}\lambda\right)$, $|S_0| = O(1)$ (worst scenario).
- When $||\hat{\Phi} - \Phi||_2^2 = ||\hat{\Phi} - \Phi||_F^2/p = O_p\left(\frac{|S_0|}{np}\lambda\right)$, $|S_0| = O(p)$.

Theorem (Faster rate of convergence for a local optimum)

Assume that Assumption 1 holds with $s_0 = |S_0| + \lceil 4(\rho_2/\rho_1)|S_0| \rceil$ *and that* X *is sub-Gaussian. If* (i) *the prior hyper-parameters* v_0, v_1, η *satisfy:*

$$\begin{cases} \frac{1}{nv_1} < C_L \sqrt{\frac{c_0 \log(p+q)}{n}}, \frac{1}{nv_0} > C_R \sqrt{\frac{c_0 \log(p+q)}{n}}, \\ \frac{v_1^2(1-\eta)}{v_0^2 \eta} \leq (p+q)^{\epsilon}, \end{cases} \tag{8}$$

for some constants $C_R > C_L$ *and some sufficiently small* $\epsilon > 0$*,*
(ii) *the matrix norm bound* R *satisfies* $|S_0|r + ||\Phi^0||_1 < R$*, and*
(iii) *the sample size* n *satisfies* $\sqrt{n} \geq M\sqrt{c_0 \log(p+q)}$*, where*

$$M = \max\left\{2(2C_1 + C_L)c_{\Gamma^0} \max\left\{3c_{\Sigma^0} d, 3708 d^2 c_{\Gamma^0}^2 c_{\Sigma^0}^4 \rho_2\right\}, \frac{2C_L}{nk_1}, d\right\},$$

then for sufficiently large constant $c_0 > 0$*, there exists a local minimizer* $\tilde{\Phi}$ *such that*

$$||\tilde{\Phi} - \Phi^0||_\infty < 2(2C_1 + C_L)c_{\Gamma^0}\sqrt{\frac{c_0 \log(p+q)}{n}} \tag{9}$$

with probability at least $1 - c_1 \exp(-c_2 \log(pq))$*.*

Theorem (Sparsistency for the estimator)

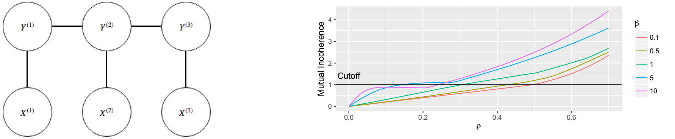
For the estimator $\tilde{\Phi}$ *in Theorem 3,* $\tilde{\Phi}_{ij} = 0$ *for all* $(i, j) \in S_0^c$ *with probability converging to* 1*.*

Comparison with existing results

- Gaussian conditional random field model with the ℓ_1 penalty: Under mutual incoherence conditions, i.e., $|||H_{S_0^c S_0}(H_{S_0 S_0})^{-1}|||_\infty < 1$, [Wytock and Kolter, 2013] showed that the convergence rate in element-wise ℓ_∞ norm for the Gaussian conditional random field model with the Lasso penalty is of the same order as ours. Such a condition is quite restrictive and often is too ideal to be true.

Consider a simple Markov chain Gaussian conditional random field model in Figure 4(a), with

$$\Lambda^0 = \begin{bmatrix} 1 & \rho & 0 \\ \rho & 1 & \rho \\ 0 & \rho & 1 \end{bmatrix}, \quad \Theta^0 = \begin{bmatrix} \rho\beta & 0 & 0 \\ 0 & \rho\beta & 0 \\ 0 & 0 & \rho\beta \end{bmatrix}.$$



(a) Visualization for the Markov chain (b) Mutual incoherence is violated as ρ

Table of Contents

EM Algorithm

- We treat r_{ij} as latent and derive an EM algorithm to obtain the MAP estimator of Θ
- E-step: compute the conditional posterior distribution of r_{ij}.
- M-step: optimize the following optimization problem:

$$\underset{\Lambda \succ 0, ||\Theta||_1 + ||\Lambda|| \le R}{\operatorname{argmin}} \left(\ell(\Theta, \Lambda) + \sum_{i,j} \lambda(\theta_{ij})|\theta_{ij}| + + \sum_{i,j} \lambda(\Lambda_{ij})|\Lambda_{ij}| \right), \quad (10)$$

where $\lambda(\cdot) = \frac{p_{ij}}{v_1} + \frac{1-p_{ij}}{v_0}$ and p_{ij} is the expectation of r_{ij} from E-step.

- Let Φ denotes $\begin{bmatrix} \Lambda \\ \Theta \end{bmatrix}$, we iteratively approximate $\ell(\Phi + \Delta)$ with its second-order Taylor expansion $g(\Delta)$ on Φ, and then solve the following optimization problem using coordinate descent for all the coordinates once:

$$\hat{\Delta} = \arg\max_{\Delta} \left(g(\Delta) - \sum_{i,j} \lambda(\Phi_{ij}) |\Phi_{ij} + \Delta_{ij}| \right). \quad (11)$$

Table of Contents

Simulation studies

Simulation Set-up

1. We generate X from a zero-mean multivariate Gaussian distribution with *dense* precision matrix $\Theta^0_{xx} = 0.5(J + I)$, where J is the matrix of ones.
2. The precision matrix Λ^0 is generated as a random graph similar to the set-up of the random graph in [Peng et al., 2009].
   - We first generate the entries in the precision matrix following the distribution of $S \times B \times U_1$, where $(S + 1)/2 \sim \text{Bern}(0.5)$, $B \sim \text{Bern}(0.1)$, $U_1 \sim \text{Uniform}(1, 2)$, and the three random variables are independent.
   - We then rescale the non-zero elements to assure positive definiteness of Λ.

We consider the following forms of true Θ^0:

1. Model 1 (Random Graph): $\Theta^0 \sim S \times B \times U_2$, where S and B are random variables as defined before, and independent of $U_2 \sim$ Uniform$(0.5, 1)$.
2. Model 2 (Banded Model 1): for i-th row of Θ^0, $(i-1)/\lfloor q/p \rfloor + 1$-th element is generated from $S \times B \times U_2$.
3. Model 3 (Banded Model 2): the i-th row of Θ^0 is of probability 0.1 to be non-zero and probability 0.9 to be all zero; when the i-th row of the Θ^0 is non-zero, its entries are generated with the distribution of $S \times B \times U_2$, where $(S+1)/2 \sim$ Bern(0.5), $B \sim$ Bern(0.1), and $U_2 \sim$ Uniform$(0.5, 1)$.

Methods in Comparisons

1. Gaussian conditional random field model with Lasso regularization,
2. Graphical Lasso [Friedman et al., 2008] jointly for (X, Y),
3. CAPME, a covariate adjusted Graphical model proposed by [Cai et al., 2012].

Simulation Studies

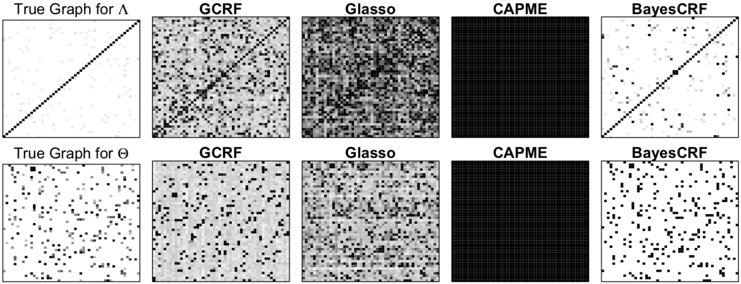


Figure: Estimates for Random Graph.

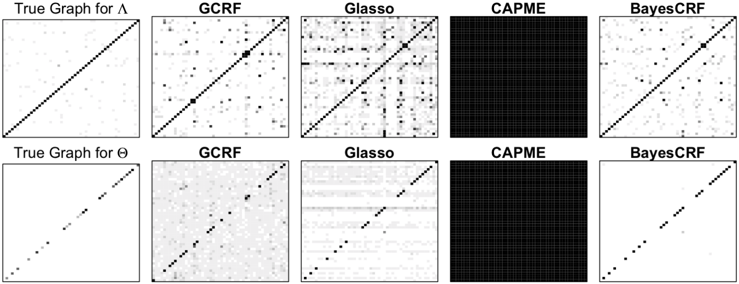


Figure: Estimates for Banded Model 1.

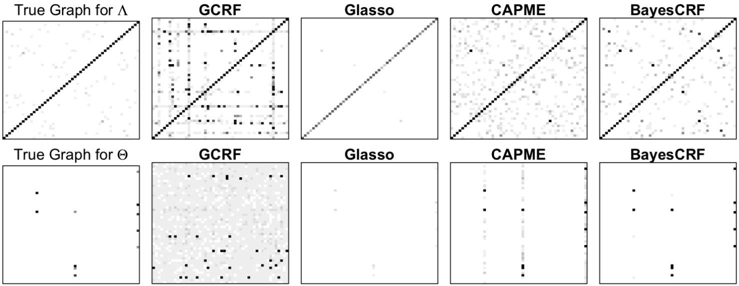


Figure: Estimates for Banded Model 2.

Table: Random Graph

	$n = 100, q = 50, p = 50$			$n = 100, q = 100, p = 50$		
	MCC	Fnorm	Test Error	MCC	Fnorm	Test Error
GLasso	0.263(0.039)	10.606(0.735)	2.001(0.296)	0.375(0.013)	17.767(0.061)	4.922(0.181)
CAPME	-0.025(0.001)	46.965(5.653)	2.442(0.125)	-0.020(0.010)	51.674(5.724)	3.934(0.199)
GCRF	0.360(0.0181)	6.901(0.344)	1.446(0.036)	0.481(0.011)	11.709(0.360)	1.652(0.039)
BayesCRF	**0.608(0.010)**	**6.012(0.149)**	**1.390(0.031)**	**0.711(0.006)**	**11.088(0.154)**	**1.560(0.041)**
	$n = 100, q = 200, p = 50$			$n = 100, q = 500, p = 50$		
	MCC	Fnorm	Test Error	MCC	Fnorm	Test Error
GLasso	0.337(0.007)	25.472(0.004)	8.180(0.154)	0.180(0.004)	38.747(0.004)	10.366(0.310)
CAPME	-0.015(0.008)	21.532(0.544)	5.433(0.205)	0.000(0.008)	37.889(0.155)	10.086(0.329)
GCRF	0.411(0.008)	22.213(0.338)	**3.142(0.071)**	**0.270(0.012)**	38.963(0.018)	21.706(3.835)
BayesCRF	**0.517(0.036)**	**21.075(0.242)**	3.484(0.601)	0.186(0.008)	**37.127(0.110)**	**7.142(1.341)**

Table: Banded Model 1

	$n = 100, q = 50, p = 50$			$n = 100, q = 100, p = 50$		
	MCC	Fnorm	Test Error	MCC	Fnorm	Test Error
GLasso	0.330(0.022)	4.223(0.040)	1.279(0.032)	0.314(0.015)	5.316(0.035)	1.390(0.035)
CAPME	-0.037(0.001)	30.346(2.709)	1.455(0.046)	-0.036(0.012)	43.642(3.320)	1.696(0.046)
GCRF	0.130(0.020)	3.050(0.110)	**1.250(0.028)**	0.216(0.021)	3.595(0.194)	**1.309(0.031)**
BayesCRF	**0.409(0.026)**	**2.498(0.094)**	1.278(0.032)	**0.452(0.024)**	**2.453(0.077)**	1.335(0.031)
	$n = 100, q = 200, p = 50$			$n = 100, q = 500, p = 50$		
	MCC	Fnorm	Test Error	MCC	Fnorm	Test Error
GLasso	0.394(0.012)	9.118(0.015)	2.051(0.053)	0.304(0.046)	12.684(0.162)	2.777(0.187)
CAPME	-0.033(0.010)	63.073(6.914)	2.294(0.069)	0.071(0.004)	13.735(1.546)	2.232(0.060)
GCRF	0.361(0.015)	5.369(0.228)	1.489(0.031)	0.412(0.011)	8.628(0.333)	1.665(0.041)
BayesCRF	**0.606(0.015)**	**3.163(0.110)**	**1.431(0.032)**	**0.674(0.011)**	**6.297(0.143)**	**1.555(0.035)**

Table: Banded Model 2

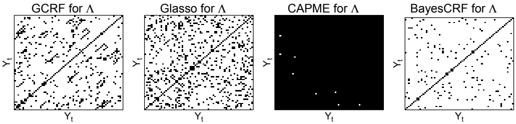
	$n = 100, q = 50, p = 50$			$n = 100, q = 100, p = 50$		
	MCC	Fnorm	Test Error	MCC	Fnorm	Test Error
GLasso	0.262(0.017)	3.763(0.047)	1.191(0.031)	0.278(0.015)	5.294(0.031)	1.342(0.030)
CAPME	-0.037(0.000)	27.884(2.113)	1.362(0.044)	-0.035(0.011)	43.030(3.666)	1.658(0.062)
GCRF	0.131(0.023)	3.827(0.136)	**1.215(0.026)**	0.164(0.023)	4.435(0.122)	**1.260(0.027)**
BayesCRF	**0.322(0.026)**	**2.725(0.092)**	1.238(0.031)	**0.392(0.021)**	**2.873(0.106)**	1.316(0.030)
	$n = 100, q = 200, p = 50$			$n = 100, q = 500, p = 50$		
	MCC	Fnorm	Test Error	MCC	Fnorm	Test Error
GLasso	0.326(0.022)	8.489(0.182)	1.775(0.067)	0.255(0.005)	12.543(0.011)	2.577(0.072)
CAPME	-0.034(0.010)	67.937(6.744)	2.066(0.086)	0.109(0.005)	12.534(0.905)	2.166(0.075)
GCRF	0.263(0.017)	6.468(0.119)	**1.379(0.036)**	0.383(0.012)	10.182(0.173)	1.666(0.042)
BayesCRF	**0.476(0.016)**	**3.566(0.097)**	1.386(0.030)	**0.634(0.012)**	**6.372(0.142)**	**1.550(0.038)**

Real application: asset return predictions

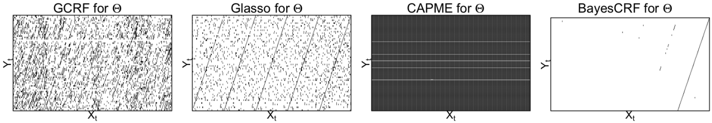
- Weekly price data of S&P 500 stocks for 265 consecutive weeks from March 10, 2003 to March, 24, 2008 collected by [Pfaff, 2016].
- We screen out all the stocks with extremely low or high marginal variance and keep 67 stocks that vary modestly, i.e., stocks with a variance between 25 and 40. All the stock prices are $\log$ transformed.
- We apply all the methods on the first 212 days to estimate Φ and make predictions on the remaining 53 days.

We want to uncover the insights on the dependency between the prices of different stocks and between their previous prices, and make good predictions. The average prediction errors are evaluated by:

$$\overline{Err} = \frac{1}{49} \sum_{t=213}^{265} ||Y_t - \hat{Y}_t||_2.$$



(a) Estimates for the precision matrix Λ for the asset return data.



(b) Estimates for Θ for the asset return data.

BayesCRF	GCRF	CAPME	Glasso
0.910(0.384)	3.817(0.468)	1.443(0.442)	1.250(0.495)

Table: Average Prediction Error for Asset Return Prediction

Conclusion

- Propose a new approach for Gaussian conditional random field estimation using Bayesian Regularization.
- Both numerically and theoretically, the Bayesian regularization method we proposed works very well.
- Hope the success demonstrated in our work will motivate further interest in using Bayesian Regularization.

References I

Cai, T. T., Li, H., Liu, W., and Xie, J. (2012).
Covariate-adjusted precision matrix estimation with an application in genetical genomics.
Biometrika, 100(1):139–156.

Candes, E. and Tao, T. (2007).
The dantzig selector: Statistical estimation when p is much larger than n.
The Annals of Statistics, pages 2313–2351.

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

Gan, L., Narisetty, N. N., and Liang, F. (2018).
Bayesian regularization for graphical models with unequal shrinkage.
Journal of the American Statistical Association, (just-accepted).

George, E. I. and McCulloch, R. E. (1993).
Variable selection via Gibbs sampling.
Journal of the American Statistical Association, 88:881–889.

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.

Pfaff, B. (2016).
Financial Risk Modelling and Portfolio Optimisation with R.
John Wiley & Sons, Ltd, London, 2nd edition.

References II

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

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).

Sohn, K.-A. and Kim, S. (2012).
Joint estimation of structured sparsity and output structure in multiple-output regression via inverse-covariance regularization.
In *AISTATS*, pages 1081–1089.

Wytock, M. and Kolter, Z. (2013).
Sparse Gaussian conditional random fields: Algorithms, theory, and application to energy forecasting.
In *Proceedings of the 30th International Conference on Machine Learning (ICML-13)*, pages 1265–1273.

Yuan, X.-T. and Zhang, T. (2014).
Partial Gaussian graphical model estimation.
IEEE Transactions on Information Theory, 60(3):1673–1687.