Network Spike-and-slab GLM

Boyi Guo & Nengjun Yi

Department of Biostatistics
University of Alabama at Birmingham

3/16/2021

Outlines

Outlines

- Motivation & Background
- Network Spike-and-slab GLM
  - Spike-and-slab prior review
  - Network spike-and-slab prior & GLM models
  - EM-based algorithms
- Numeric Studies

Motivation & Background

Motivation

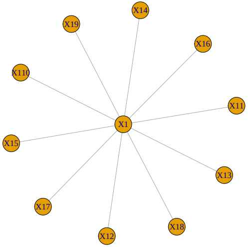
- High-dimensional data analysis (p » n) requires extra assumptions to be feasible
  - Sparsity
- Additional assumptions can make the inference more stable and accurate
  - External knowledge
- To integrate complicated external information of predictor structures
  - Grouping: gene pathway
  - Network: gene network, phylogenetic tree

Network Structure

- Un-directed weighted graphs
- Normalized similarity matrix D
  - Similar to adjacency matrix in essence
  - $d_{ij} \in [0, 1]$, $d_{ij} = 0$ means X_i and X_j are not related, $d_{ij} = 1$ means related
  - Strength of pairwise relationships *or* Degree of belief in pairwise relationships
  - We define $d_{ii} = 1$ for $i = 1, \ldots p$
  - Naive example: absolute values of correlation matrix

Network Structure Example

- "Hub-and-spike" structure (C. Li and Li 2008)



$$\boldsymbol{D} = \begin{bmatrix} 1 & 1 & 1 & \dots & 1 & 1 \\ 1 & 1 & 0 & \dots & 0 & 0 \\ 1 & 0 & 1 & \dots & 0 & 0 \\ \vdots & 0 & 0 & \ddots & 0 & 0 \\ 1 & 0 & 0 & \dots & 1 & 1 \\ 1 & 0 & 0 & \dots & 0 & 1 \end{bmatrix}$$

Past solutions

- Grouped analysis
  - Group spike-and-slab lasso GLM and Cox models (Tang et al. 2018, 2019)
- Network analysis
  - Network constrained penalty (C. Li and Li 2008)
  - L_γ-norm penalty (Pan, Xie, and Shen 2010)
  - Ising Prior (F. Li and Zhang 2010)
  - Markov random field (Monni and Li 2010)

Network Spike-and-slab GLM

Spike and Slab Priors

- First coined by Mitchell and Beauchamp (1988)
- A mixture distribution $(1-\gamma_i)f_{spike}(\beta_i) + \gamma_i f_{slab}(\beta_i)$
  - $f_{spike}(x)$: a spike density concentrate around 0 for small effects
  - $f_{slab}(x)$: a flat density for large effects
- Advantages:
  - Simultaneous variable selection and prediction
  - Robust estimation
- Examples:
  - Stochastic search variable selection (George and McCulloch 1993)
  - Spike-and-slab lasso (Ročková and George 2018)

Network Spike-and-slab GLM

- GLM Components

$$E(Y_i) = g^{-1}(\beta^T \boldsymbol{X}_i), i = 1 \dots n, Y_i \sim EF_{\beta,\phi}$$

- Network Spike and Slab Priors

$$\begin{aligned} \beta_j|\gamma_j, s_0, s_1 &\sim (1-\gamma_j)DE(\beta_j|0, s_0) + \gamma_j DE(\beta_j|0, s_1), s_0 < s_1 \\ \gamma_j|\theta_j &\sim Bin(\gamma_j|1, \theta_j) \\ \theta_j|\boldsymbol{D}, \boldsymbol{\theta}_{\{1:p\}/\{j\}} &\sim Beta(a_j, b_j), j = 1 \dots p \end{aligned}$$

$$a_j = \sum_{k \in \{1:p\}/\{j\}} d_{jk} * \theta_k, \qquad b_j = 1 - a_j$$

Computation algorithm

We develop two Expectation-Maximization (EM) based algorithms

- EM - Iterative weighted least square
  - Uncertainty inference
- EM - Coordinate descent algorithm
  - Sparse Solution

EM Algorithm Recap

EM algorithm is an iterative algorithm to find local maximum a posterori estimates

- Treat nuisance parameters as "missing values"
- Calculate the expected likelihood with respect to the missing values
- Maximize the expected likelihood
- Iterate the process until convergence

EM algorithms

- Treat γ as "missing"
- $E(\cdot) \equiv E_{\gamma|\Theta^{(t)}}(\cdot)$ where $\Theta^{(t)} \equiv (\beta^{(t)}, \boldsymbol{\theta}^{(t)}, \phi^{(t)})$

$$E\left[Q(\Theta)\right] = C + E\left[Q_1(\beta, \phi)\right] + E\left[Q_2(\boldsymbol{\theta})\right]$$

where

$$E\left[Q_1(\beta, \phi)\right] = \sum_{i=1}^{n} \log f(y_i|\beta, \phi) + \log f(\phi) + \sum_{j=1}^{p} |\beta_i| E\left[(1-\gamma_j)s_0 + \gamma_j s_1\right]^{-1}$$

$$E\left[Q_2(\boldsymbol{\theta})\right] = \sum_{j=1}^{p} \log(\frac{\theta_j}{1-\theta_j}) E\left[\gamma_j\right] + \sum_{j=1}^{p} \left[(a_j^{(t)} - 1)\log(\theta_j) + (b_j^{(t)} + 1 - 1)\log(1-\theta_j)\right]$$

EM algorithms (Cont.)

Using *Bayes' Theorem*, we can have

$$p_j^{(t)} \equiv f(\gamma_j = 1|\beta_j^{(t)}, \theta_j^{(t)}) = E_{\gamma|\boldsymbol{\Theta}^{(t)}}\gamma_j =$$

$$= \frac{f(\beta_j^{(t)}|\gamma_j = 1, s_1)f(\gamma_j = 1|\theta_j^{(t)})}{f(\beta_j^{(t)}|\gamma_j = 1, s_1)f(\gamma_j = 1|\theta_j^{(t)}) + f(\beta_j^{(t)}|\gamma_j = 0, s_1)f(\gamma_j = 0|\theta_j^{(t)})}$$

It is trivial to show that the expectations with respect to $\gamma_j|\beta_j, \theta_j$

$$\lambda_j^{(t)} \equiv E(\left[(1-\gamma_j)s_0 + \gamma_j s_1\right]^{-1})|\beta_j^{(t)}, \theta_j^{(t)}) = \frac{p_j^{(t)}}{s_1} + \frac{1-p_j^{(t)}}{s_0}$$

EM - Iterative Weighted Least Square

- Update β by maximizing the linear approximation to the normal likelihood using weighted least square (Yi and Ma 2012).

Equivalently, running the autmented weighted normal linear regression

$$z_* \approx N(X_*\beta, \phi\Sigma_*),$$

where $z_* = \begin{pmatrix} z \\ 0 \end{pmatrix}$, $X_* = \begin{pmatrix} X \\ I_{p+1} \end{pmatrix}$, $\Sigma_* = diag(w_1^{-1}, \ldots, w_n^{-1}, \tau_0^2/\phi, \ldots, \tau_p^2/\phi)$

- Update $\boldsymbol{\theta}$

EM - Coordinate Descent

- Update β, ϕ with coordinate descent algorithm (Friedman, Hastie, and Tibshirani 2010)

$$\{\beta^{(t+1)}, \phi^{(t+1)}\} = \text{argmax } Q_1(\beta, \phi | \beta^{(t)}, \boldsymbol{\theta}^{(t)}, \phi^{(t)})$$

- Update $\boldsymbol{\theta}$

$$\boldsymbol{\theta}^{(t+1)} = \text{argmax } Q_2(\boldsymbol{\theta} | \beta^{(t)}, \boldsymbol{\theta}^{(t)} \phi^{(t)})$$

$$\theta_j^{(t+1)} = \frac{a_j^{(t)} + p_j^{(t)} - 1}{a_j^{(t)} + b_j^{(t)} - 1}$$

Numeric Studies

Continous Outcome

The simulation follows C. Li and Li (2008); Pan, Xie, and Shen (2010)

- Number of Predictors: 2 settings: p=110/ p=1100
- Sample Size for Training Data: 100/1000
- Sample Size for Testing Data: 1000
- $\boldsymbol{X}$ has the hub-and-spike structure for every 11 predictors
- 5 settings of βs for various directions and magnitudes of signals
- Model

$$y = X\beta + N(0, \frac{1}{2}\sum_i \beta_i^2),$$

- Number of Iteration: 1000

Methods & Metrics

- Other methods of comparison
  - Least absolute shrinkage and selection operator (LASSO)
  - Elastic net regularization
  - Network constrained penalty (GRACE) (C. Li and Li 2008) via `glmgraph`[1]
  - Spike-and-slab GLM (Tang et al. 2017) via `BhGLM`
  - 10-fold cross-validation
- Metrics Evaluated
  - Sensitivity: Number of true positive predictors / Number of positive predictors
  - Specificity: Number of true negative predictors / Number of negative predictors
  - Prediction mean squared error + mean absolute error

[1] glmgraph: Graph-Constrained Regularization for Sparse Generalized Linear Models. R package version 1.0.3.

Simulation Results

An interactive R Shiny app to browser simulation result is available via https://boyiguo1.shinyapps.io/Network_Result_Browser/.

Selected Results

- Coefficients

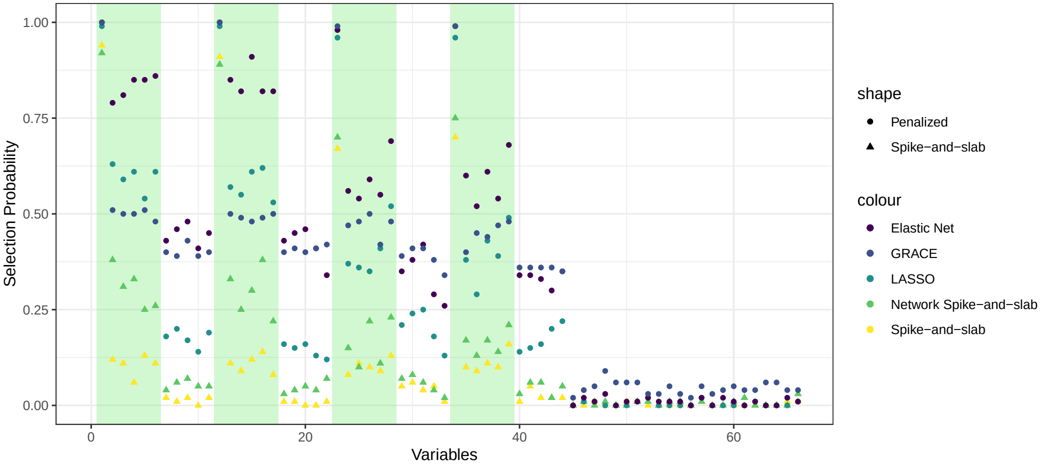
$$\beta = (5, \underbrace{\frac{5}{\sqrt{10}}, \ldots, \frac{5}{\sqrt{10}}}_{5}, \underbrace{0, \ldots, 0}_{5}, -5, \underbrace{-\frac{5}{\sqrt{10}}, \ldots, -\frac{5}{\sqrt{10}}}_{5}, \underbrace{0, \ldots, 0}_{5},$$
$$3, \underbrace{\frac{3}{\sqrt{10}}, \ldots, \frac{3}{\sqrt{10}}}_{5}, \underbrace{0, \ldots, 0}_{5}, -3, \underbrace{-\frac{3}{\sqrt{10}}, \ldots, -\frac{3}{\sqrt{10}}}_{5}, \underbrace{0, \ldots, 0}_{5},$$
$$0, \ldots, 0)^T$$

- $p = 1100, n_{train} = 100$

Performance

Method	MSE	MAE	Sensitivity	Specificity
LASSO	25.032	3.930	0.644	0.990
Elastic Net	26.002	4.013	0.547	0.995
GRACE	16.734	3.233	0.332	0.990
Spike-and-slab	19.261	3.447	0.788	0.983
Network Spike-and-slab	22.868	3.767	0.588	0.985

Variable Selection



Binary Outcome Results

- Coefficients

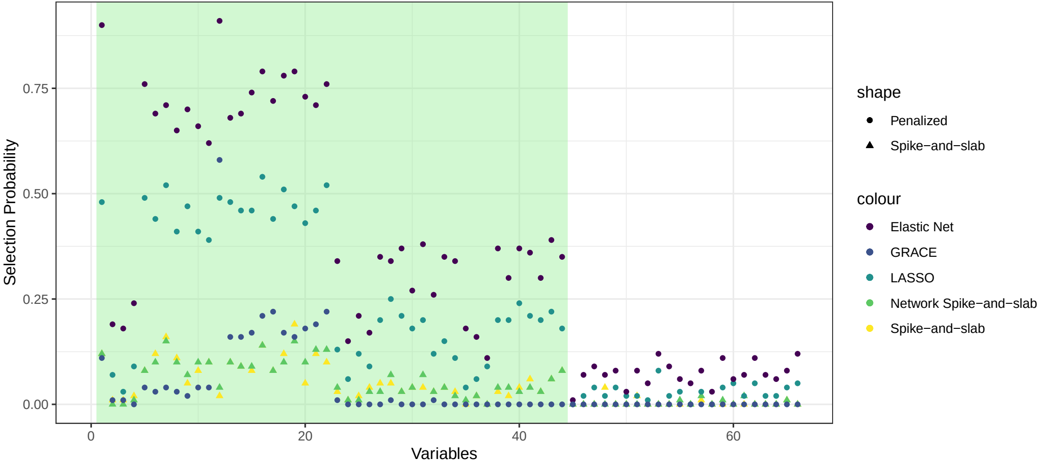
$$\beta = (1, \underbrace{-0.7, \ldots, -0.7}_{3}, \underbrace{0.7, \ldots, 0.7}_{7}, -1, \underbrace{0.7, \ldots, 0.7}_{3}, \underbrace{-0.7, \ldots, -0.7}_{7},$$
$$0.5, \underbrace{-0.3, \ldots, -0.3}_{3}, \underbrace{0.3, \ldots, 0.3}_{7}, -0.5, \underbrace{0.3, \ldots, 0.3}_{3}, \underbrace{-0.3, \ldots, -0.3}_{7},$$
$$0, \ldots, 0)^T$$

- $p = 110, n_{train} = 100$

Performance

Method	AUC	MSE	MAE	Sensitivity	Specificity
LASSO	0.918	0.127	0.290	0.891	0.672
Elastic Net	0.932	0.118	0.279	0.845	0.730
GRACE	0.745	0.247	0.494	0.800	0.617
Spike-and-slab	0.911	0.127	0.237	0.910	0.614
Network Spike-and-slab	0.910	0.128	0.237	0.924	0.615

Variable Selection



Conclusion

Conclusion

- Bayesian models for simultaneous variable selection and prediction
- Fast computing algorithms to establish maximum a posteriori estimates
- All-in-one analysis toolkit for GLM and Cox models
    - BhGLM via https://github.com/nyiuab/BhGLM
- Future steps
    - More simulations
    - Study prior set-up s_0, s_1
    - Improve package documentation

Ackowledgement

Dr. Nengjun Yi

Dr. D. Leann Long

References

References

Friedman, Jerome, Trevor Hastie, and Rob Tibshirani. 2010. "Regularization Paths for Generalized Linear Models via Coordinate Descent." *Journal of Statistical Software* 33 (1): 1.

George, Edward I., and Robert E. McCulloch. 1993. "Variable Selection via Gibbs Sampling." *Journal of the American Statistical Association* 88 (423): 881–89. https://doi.org/10.1080/01621459.1993.10476353.

Li, Caiyan, and Hongzhe Li. 2008. "Network-constrained regularization and variable selection for analysis of genomic data." *Bioinformatics* 24 (9): 1175–82. https://doi.org/10.1093/bioinformatics/btn081.

Li, Fan, and Nancy R Zhang. 2010. "Bayesian Variable Selection in Structured High-Dimensional Covariate Spaces With Applications in Genomics." *Journal of the American Statistical Association* 105 (491): 1202–14.