Bayesian Variable Selection in High-Dimensional Regressions with Correlated Noise

Anindya Bhadra
bhadra@purdue.edu
www.stat.purdue.edu/∼bhadra

Purdue University

July 13, 2014

Overview

- We consider a "multiple predictors, multiple responses" regression problem where the error terms may be correlated.
- Zellner (1962) discusses at length the consequences of ignoring the error covariance while performing regression.
- Many high-dimensional applications in genomics fall in this framework. For example: predictors could be copy number variations (CNV) and responses could be gene (mRNA) expressions.
- We formulate a Bayesian "joint" estimation technique of CNV-mRNA association and mRNA-mRNA interaction network.

Problem Formulation

- n = Number of humans.
- X = An $n \times p$ matrix of predictors
- Y = An $n \times q$ matrix of responses
- We would like to regress Y on X.
- Example (CNV-mRNA interaction in Breast Cancer): For n individuals with breast cancer, we analyze how CNVs (X) affect their mRNA expressions (Y).

Problem Formulation

- Consider the linear Gaussian regression model:

$$\begin{aligned} \mathbf{Y}_{n\times q} &= \mathbf{X}_{n\times p}\mathbf{B}_{p\times q} + \boldsymbol{\epsilon}_{n\times q}, \\ \boldsymbol{\epsilon}_{n\times q} &\sim \mathrm{MN}_{n\times q}(\mathbf{0}, \mathbf{I}_n, \boldsymbol{\Sigma}_{q\times q}), \\ \text{i.e.,}\mathrm{Vec}(\boldsymbol{\epsilon}_{n\times q}) &\sim \mathrm{N}_{nq}(\mathbf{0}, \mathbf{I}_n \otimes \boldsymbol{\Sigma}_{q\times q}). \end{aligned}$$

- The unknowns are $\mathbf{B}_{p\times q}$ and $\Sigma_{q\times q}$.

- The dimensions are pq and $q(q+1)/2$. Often much larger than n.

- Typical values: $n = 100$, $p = 500$ to 3000, $q = 100$.

Joint modeling of mean and covariance for Seemingly Unrelated Regression

- In a Seemingly Unrelated Regression setting, one might be interested in modeling "both" the mean and the covariance structure.
- Rothman et al. (2010, JCGS) and Yin and Li (2011, Ann. Appl. Stat.) make a frequentist attempt at joint modeling with the MRCE approach. (essentially an iterative approach with alternating lasso() and glasso() steps).
- Other approaches include the CAPME (Biometrika, 2013) and CLIME (`arXiv:1102.2233`) methods of Cai et al.
- Bhadra and Mallick (Biometrics, 2013) take a Bayesian approach.

Model conditional on indicators: Toy example

- Consider the model conditional upon indicators γ and $\mathbf{G}$.

$$\mathbf{Y} = \mathbf{X}_{\gamma}\mathbf{B}_{\gamma,\mathbf{G}} + \epsilon; \quad \epsilon \sim \mathrm{MN}(\mathbf{0}, \mathbf{I}_n, \mathbf{\Sigma}_{\mathbf{G}}).$$

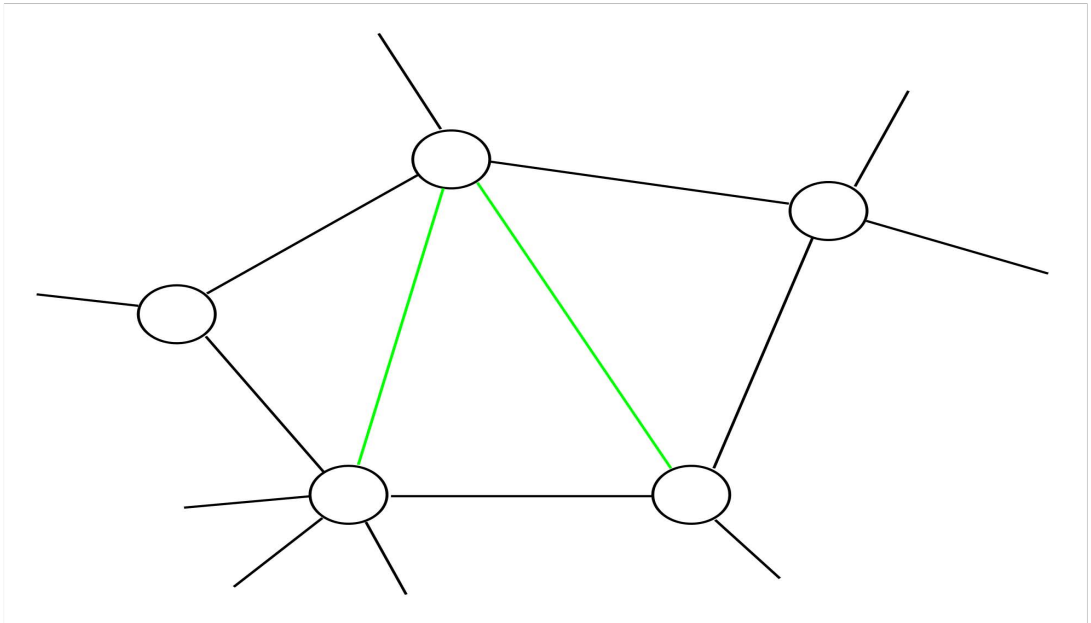
- For example, say $p = q = 4$. Then $\gamma = (1, 0, 1, 0)$ means only the first and the third predictors are important.

- Let's say $\mathbf{G}$ is:

$$\begin{bmatrix} 1 & 1 & 0 & 0 \\ 1 & 1 & 0 & 0 \\ 0 & 0 & 1 & 0 \\ 0 & 0 & 0 & 1 \end{bmatrix}$$

This means $\mathbf{\Sigma}^{-1}_{1,2} \neq 0$, the other off-diagonal terms are 0.

Decomposable (or triangulated) graphs



- No chordless cycle of length ≥ 3.
- Cliques (i.e., the connected components) and separators (i.e., the parts in common between two cliques) can be found in polynomial time (NP-complete for general graphs).
- The overall density splits as:
  $f(y) = \prod_{j=1}^{k} f(y_{C_j}) / \prod_{j=2}^{k} f(y_{S_j})$.

Bayesian hierarchical model

$$
\begin{aligned}
(\mathbf{Y} - \mathbf{X}_{\boldsymbol{\gamma}}\mathbf{B}_{\boldsymbol{\gamma},\mathbf{G}})|\mathbf{B}_{\boldsymbol{\gamma},\mathbf{G}}, \boldsymbol{\Sigma}_{\mathbf{G}} &\sim \mathrm{MN}_{n\times q}(\mathbf{0}, \mathbf{I}_n, \boldsymbol{\Sigma}_{\mathbf{G}}),\\
\mathbf{B}_{\boldsymbol{\gamma},\mathbf{G}}|\boldsymbol{\gamma}, \Sigma_{\mathbf{G}} &\sim \mathrm{MN}_{p_{\boldsymbol{\gamma}}\times q}(\mathbf{0}, c\mathbf{I}_{p_{\boldsymbol{\gamma}}}, \boldsymbol{\Sigma}_{\mathbf{G}}),\\
\boldsymbol{\Sigma}_{\mathbf{G}}|\mathbf{G} &\sim \mathrm{HIW}_{\mathbf{G}}(b, d\mathbf{I}_q),\\
\gamma_i &\overset{\text{i.i.d}}{\sim} \mathrm{Ber}(w_\gamma) \text{ for } i = 1, \ldots, p,\\
\mathrm{G}_k &\overset{\text{i.i.d}}{\sim} \mathrm{Ber}(w_G) \text{ for } k = 1, \ldots, q(q-1)/2,\\
w_\gamma, w_G &\sim \mathrm{Uniform}(0, 1).
\end{aligned}
$$

The marginalized model (Bhadra and Mallick, 2013)

- After the marginalization of $\mathbf{B}_{\gamma,\mathbf{G}}$ and $\mathbf{\Sigma}_{\mathbf{G}}$, the resultant distribution is a "hyper matrix t".
- Define $\mathbf{T} = \mathbf{AY}$ where $\mathbf{AA}' = (\mathbf{I}_n + c(\mathbf{X}_\gamma \mathbf{X}_\gamma'))^{-1}$. Then

$$\mathbf{T}|\gamma, \mathbf{G} \quad \sim \quad \mathrm{HMT}_{\mathbf{G}}(b, \mathbf{I}_n, d\mathbf{I}_q).$$

- This is a special type of "t-distribution" whose density splits over cliques and separators, given the graph.
- The marginalization has now resulted in a collapsed Gibbs sampler: need to sample only two quantities (γ and $\mathbf{G}$) instead of four ($\mathbf{B}_{\gamma,\mathbf{G}}$, $\mathbf{\Sigma}_{\mathbf{G}}$, γ and $\mathbf{G}$).

MCMC for γ given $\mathbf{G}$ and $\mathbf{T}$ (Bhadra and Mallick, 2013)

1. Given the current γ, propose γ^* by either (a) changing a non-zero entry in γ to zero with probability $(1 - \alpha_\gamma)$ or (b) changing a zero entry in γ to one, with probability α_γ.

2. Calculate $f(\mathbf{t}|\gamma^*, \mathbf{G})$ and $f(\mathbf{t}|\gamma, \mathbf{G})$ where f denotes the HMT density.

3. Jump from γ to γ^* with probability

$$r(\gamma, \gamma^*) = \min\left\{1, \frac{f(\mathbf{t}|\gamma^*, \mathbf{G})p(\gamma^*)q(\gamma|\gamma^*)}{f(\mathbf{t}|\gamma, \mathbf{G})p(\gamma)q(\gamma^*|\gamma)}\right\}.$$

MCMC for $\mathbf{G}$ given γ and $\mathbf{T}$ (Bhadra and Mallick, 2013)

1. Given the current $\mathbf{G}$, propose $\mathbf{G}^*$ by either (a) changing a non-zero edge in $\mathbf{G}$ to zero with probability $(1 - \alpha_G)$ or (b) changing a zero entry in $\mathbf{G}$ to one, with probability α_G.

2. Calculate $f(\mathbf{t}|\gamma, \mathbf{G}^*)$ and $f(\mathbf{t}|\gamma, \mathbf{G})$ where f denotes the HMT density.

3. Jump from $\mathbf{G}$ to $\mathbf{G}^*$ with probability

$$r(\mathbf{G}, \mathbf{G}^*) = \min\left\{1, \frac{f(\mathbf{t}|\mathbf{G}^*, \gamma)p(\mathbf{G}^*)q(\mathbf{G}|\mathbf{G}^*)}{f(\mathbf{t}|\mathbf{G}, \gamma)p(\mathbf{G})q(\mathbf{G}^*|\mathbf{G})}\right\}.$$

The special case of tree-structured graphs

- The hyper-matrix t-density has this form:

$$f(\mathbf{t}^n_{C_j}) \quad = \quad \text{Const.} \times [\det\{\mathbf{I}_n + (\mathbf{t}^n_{C_j})(\mathbf{t}^n_{C_j})'/d\}]^{-(b+n+|C_j|-1)/2},$$

and the overall density

$$f(\mathbf{t}^n) \quad = \quad \frac{\prod_{j=1}^k f(\mathbf{t}^n_{C_j})}{\prod_{j=2}^k f(\mathbf{t}^n_{S_j})}.$$

- "Trees" are a special case of decomposable graphs, where no cycles are allowed.
- The "cliques" are the edges and the "separators" are single nodes.

The special case of tree-structured graphs

- Let t be a symmetric matrix of non-negative weights for all pairs of distinct variables and zeros on the diagonal.
- Let $\mathscr{C}$ be the set of all possible spanning trees over vertex set $\mathcal{V}$. Then

$$P(\mathcal{G} \in \mathscr{C}) = \frac{1}{Z(t)} \prod_{\{i,j\} \in \mathcal{G}} t_{ij}$$

$$\text{where } Z(t) = \sum_{\mathcal{G} \in \mathscr{C}} \prod_{\{i,j\} \in \mathcal{G}} t_{ij}$$

The Matrix tree theorem (Meila and Jaakkola, 2006)

$Z(t)$ is equal to the determinant $|L^*(t)|$, with matrix $L^*(t)$ representing the first (q-1) rows and columns of the matrix $L(t)$ given by:

$$L_{ij}(t) = L_{ji}(t) = \begin{cases} -t_{ij} & i,j \in \mathcal{V}, i \neq j; \\ \sum_{j \in \mathcal{V}} t_{ij} & i,j \in \mathcal{V}, i = j. \end{cases} \tag{1}$$

Evaluating a determinant with $(q-1)$ rows has complexity $O(q^3)$.

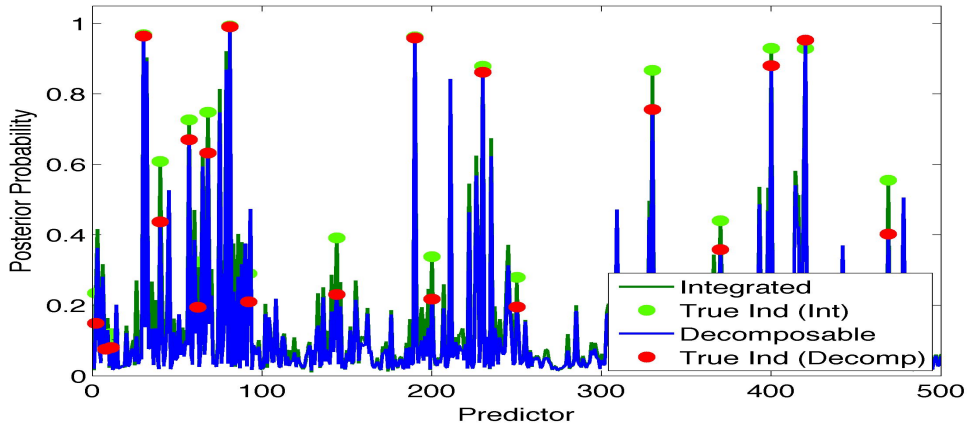
The special case of tree-structured graphs

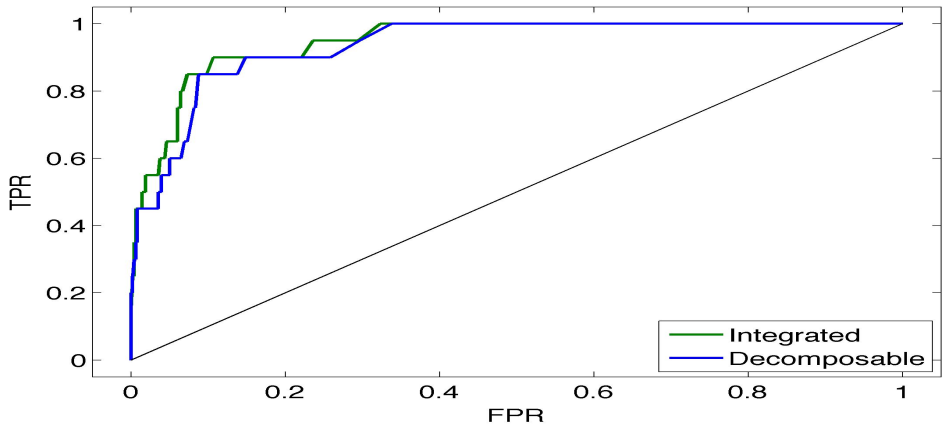
- Evaluating the normalizing constant of the density is difficult for most graphs, including decomposable graphs (non-polynomial time: $O(q^{q-2})$).
- However, the graph theoretic result known as Kirchoff's theorem, or Matrix-tree theorem (Meila and Jaakkola, 2006) provides a $O(q^3)$ algorithm for trees.
- Allows us to mix over the tree structure and this can capture a rich class of graphs.
- Instead of drawing $\gamma|\mathbf{G}, T$ and $\mathbf{G}|\gamma, T$ one can simply draw $\gamma|\tilde{T}$, where $\tilde{T}$ is the new marginal data distribution after integrating out the graph.

Simulation study

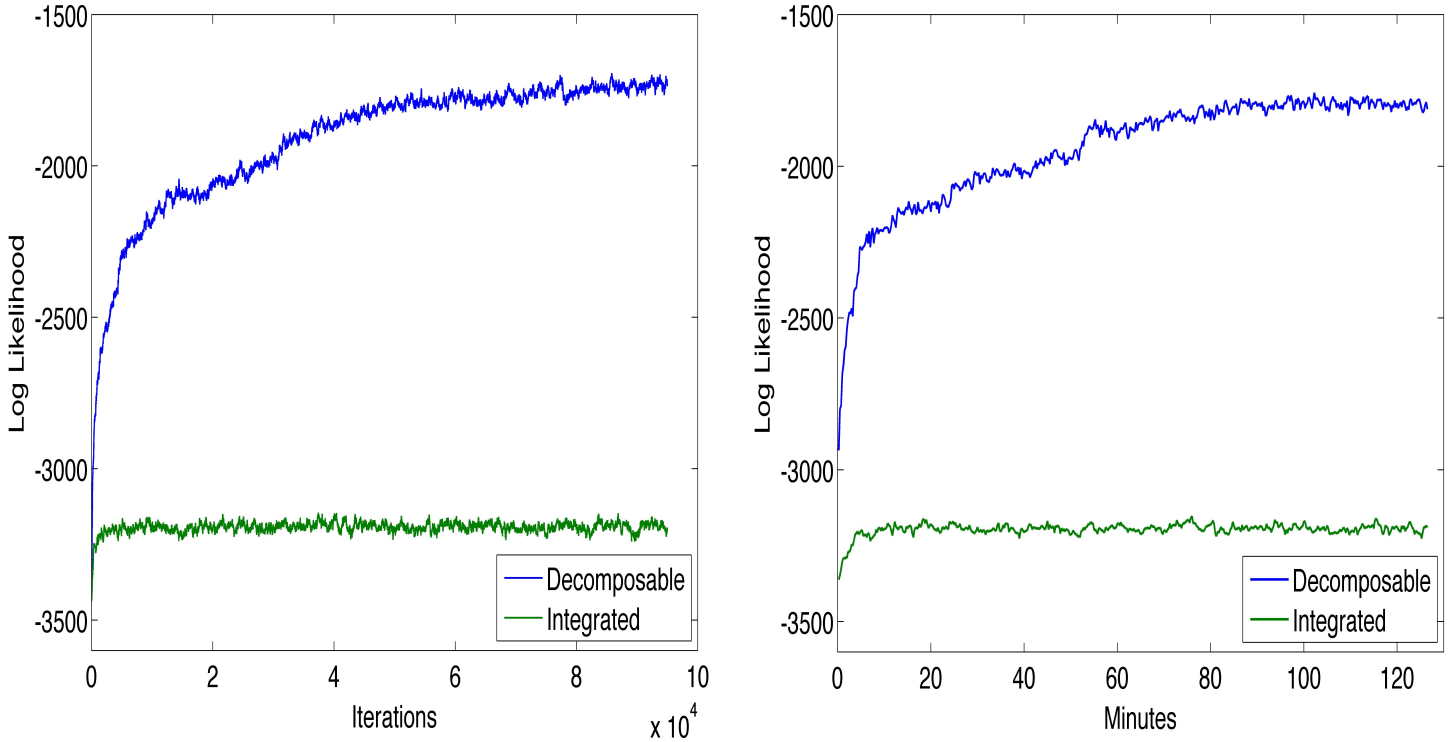
- We choose $p = 498$, $q = 50$ and $n = 100$.
- We chose the dimension of true predictors, $p_\gamma = 20$.
- The true adjacency matrix $\mathbf{G}$ was chosen to be decomposable.
- The $n \times p$ predictor matrix $\mathbf{X}$ was simulated from multivariate Normal. We tried both uncorrelated columns and correlated columns in X (a banded correlation structure of width 10 and maximum correlation of 0.8).

Results: Correlated predictors case





MCMC diagnostics



- The likelihood values are not directly comparable.
- However, mixing over trees helps MCMC convergence.

Analysis of a breast cancer data set

- We chose the breast cancer data analyzed by Peng et al. (2009).
- We have $n = 172$ breast tumor samples. For each sample, we have available $p = 384$ CNAs and $q = 654$ gene expression levels.
- Since all genes are known to be breast cancer related, we chose a subset of $\tilde{q} = 50$ genes that showed most variability.

Analysis of a breast cancer data set

Chromosomal location of selected feature (a)	Start base pair (b)	End base pair (c)	Posterior (d) (Tree)	Posterior (e) (Decomp.)
11q13.3-11q13.4	68415321	70200416	1.00	1.00
15q11.2-15q11.2	18786757	19949603	1.00	1.00
16p13.3-16p11.2	37048	32796294	1.00	0.94
17q12-17q12	31424350	31562438	1.00	1.00
17q12-17q12	34082227	34670370	1.00	1.00
17q12-17q12	34811630	34811630	1.00	1.00
17q12-17q12	34944071	35154416	1.00	1.00
17q12-17q21.1	35167500	35428880	1.00	1.00
17q21.1-17q21.2	35493689	35699243	1.00	1.00
17q21.2-17q21.2	36037494	36923525	1.00	0.66
3q29-3q29	199171511	199171511	1.00	1.00
23p22.33-23p11.3	2725527	46830187	0.99	0.89
10p15.3-10p12.1	288292	27260145	0.98	0.61
17q21.2-17q21.2	35724970	35724970	0.96	1.00
10q22.2-10q22.2	76790556	77072436	0.96	0.93
10q21.3-10q22.2	69353349	75083656	0.95	0.17
17q25.3-17q25.3	76816671	78649094	0.89	0.60

Major findings

- Several CNAs in 17q12 (location of BRCA1) and 17q21.1-17q21.2 (location of ERBB2) are selected.
- CNAs identified as having significant trans-effects by Peng et al. are in blue. The Bayesian methods select them with posterior probability 1.
- Differences bigger than 0.25 between the two Bayesian methods are highlighted in red.