Joint high-dimensional Bayesian variable and covariance selection with an application to eQTL analysis

Anindya Bhadra

Purdue University

October 10, 2012

Overview

- Variable and (inverse) covariance selections have been well-studied separately in high-dimensional problems.
- However, "joint" selection (or estimation) have not been studied until recently.
- We formulate a Bayesian technique and apply it to the analysis of expression quantitative trait loci (eQTL) analysis.
- Joint work with Bani K. Mallick, Texas A&M University.

Problem Formulation

- n = Sample size.
- 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 A: For the same n individuals, we might try to see how their SNP genotype (X) affect their gene expressions (Y).
- Example B: For the same n individuals with cancer, we might try to see how their microRNA (X) affect their mRNA (Y) expressions.
- I have worked on A; I plan to begin work on B.

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, $q = 100$.

Basics of variable and covariance selection

- When p and q are larger than n, it becomes necessary to determine a sparse set of predictors and inverse covariance matrix elements.
- Variable selection: Find out the important predictors.
  - Typical assumption: Errors are i.i.d (i.e., $\mathbf{\Sigma}_{q\times q} = \sigma^2 \mathbf{I}_q$).
- Covariance selection: Find out the important inverse covariance matrix elements.
  - For Gaussian models: $\mathbf{\Sigma}^{-1}_{i,j} = 0 \iff Y_i \perp Y_j | \text{rest}$.
  - Typical assumption: No covariates (i.e., $\mathbf{B}_{p\times q} = 0$).
- We do a joint selection. This is being done only recently.

Previous Work in variable selection

- Variable selection with i.i.d errors.
- Frequentist: Lasso (Tibshirani, 1996, JRSSB) and its various extensions using ℓ_1 penalty.
- Bayesian: Stochastic Search Variable Selection (George and McCulloch, 1997, JASA) and its extensions using sparsity prior.

Previous Work in covariance selection and estimation

- (Inverse) Covariance selection in Gaussian graphical model with zero mean.
- Frequentist: Meinshausen and Bühlmann (2006, Ann. Stat.), Graphical Lasso (Friedman et al, 2008, Biostatistics), Bickel and Levina (2008, Ann. Stat.) etc.
- Bayesian: Carvalho and West (2007, Biometrika) etc. primarily using hyper-inverse Wishart type of priors.

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) make a frequentist attempt at joint modeling with the MRCE approach. (essentially an iterative approach with alternating lasso() and glasso() steps).
- Yin and Li (2011, Ann. Appl. Stat.) apply a similar approach to gene expression and SNP data.
- Bhadra and Mallick (Biometrics, under revision) take a Bayesian approach.

Model conditional on indicators

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

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

- Dimension of $\mathbf{X}_{\gamma} = n \times p_{\boldsymbol{\gamma}}$; dimension of $\mathbf{B}_{\boldsymbol{\gamma},\mathbf{G}} = p_{\boldsymbol{\gamma}} \times q$; dimension of $\boldsymbol{\Sigma}_{\mathbf{G}} = q \times q$.

- $\gamma_i = 1 \Rightarrow \mathbf{B}_{i,\cdot} \neq 0$; $p_{\boldsymbol{\gamma}} = \sum_{i=1}^{p} \gamma_i$.

- $\mathbf{G}$ is a decomposable graph where $\mathbf{G}_{i,j} = 1 \Rightarrow \boldsymbol{\Sigma}^{-1}_{i,j} \neq 0$ with $i \neq j$; $i, j = 1, \ldots, q$.

Model conditional on indicators: Toy example

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

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

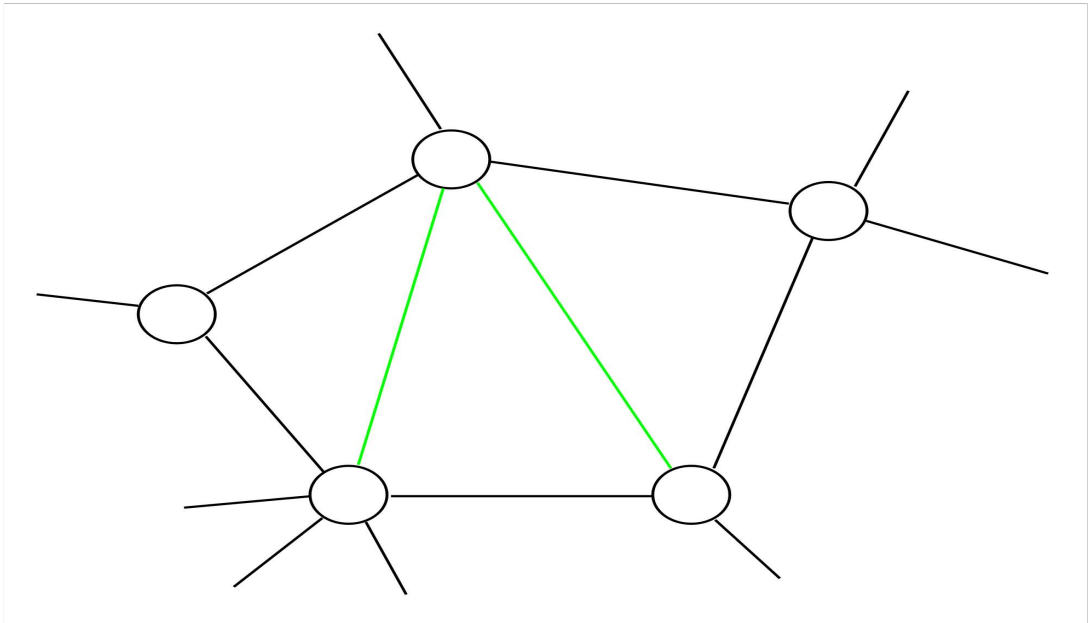
- For example, say $p = q = 4$. Then $\gamma = (1, 0, 1, 0)$ means only the first and the third predictors are important.

- 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

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

Mariginalization of $B_{\gamma,\mathbf{G}}$ and $\mathbf{\Sigma_G}$

- Remember from the last slide

$$\begin{aligned}
\epsilon &\sim \mathrm{MN}_{n\times q}(\mathbf{0}, \mathbf{I}_n, \mathbf{\Sigma_G}), \\
\mathbf{B}_{\gamma,\mathbf{G}}|\gamma, \Sigma_{\mathbf{G}} &\sim \mathrm{MN}_{p_\gamma\times q}(\mathbf{0}, c\mathbf{I}_{p_\gamma}, \mathbf{\Sigma_G}). \\
\Rightarrow \mathbf{X}_\gamma\mathbf{B}_{\gamma,\mathbf{G}}|\gamma, \mathbf{\Sigma_G} &\sim \mathrm{MN}_{n\times q}(0, c(\mathbf{X}_\gamma\mathbf{X}_\gamma'), \mathbf{\Sigma_G}). \\
\Rightarrow \mathbf{Y}|\gamma, \mathbf{\Sigma_G} &\sim \mathrm{MN}_{n\times q}(0, \mathbf{I}_n + c(\mathbf{X}_\gamma\mathbf{X}_\gamma'), \mathbf{\Sigma_G}).
\end{aligned}$$

- Define $\mathbf{T} = \mathbf{AY}$ where $\mathbf{AA}' = (\mathbf{I}_n + c(\mathbf{X}_\gamma\mathbf{X}_\gamma'))^{-1}$.

$$\begin{aligned}
\Rightarrow \mathbf{T}|\gamma, \mathbf{\Sigma_G} &\sim \mathrm{MN}_{n\times q}(\mathbf{0}, \mathbf{I}_n, \mathbf{\Sigma_G}). \\
\mathbf{\Sigma_G}|\mathbf{G} &\sim \mathrm{HIW}_{\mathbf{G}}(b, d\mathbf{I}_q). \\
\Rightarrow \mathbf{T}|\gamma, \mathbf{G} &\sim \mathrm{HMT}_{\mathbf{G}}(b, \mathbf{I}_n, d\mathbf{I}_q).
\end{aligned}$$

The marginalized model

- After the marginalization of $\mathbf{B}_{\gamma,\mathbf{G}}$ and $\mathbf{\Sigma}_{\mathbf{G}}$, the resultant distribution is a "hyper matrix t".
- 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 $\mathbf{G}$).
- Terms that were integrated out can always be sampled at the posterior, since we are working in a conjugate framework.

MCMC for γ given $\mathbf{G}$ and $\mathbf{T}$

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 $\mathbf{T}$

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

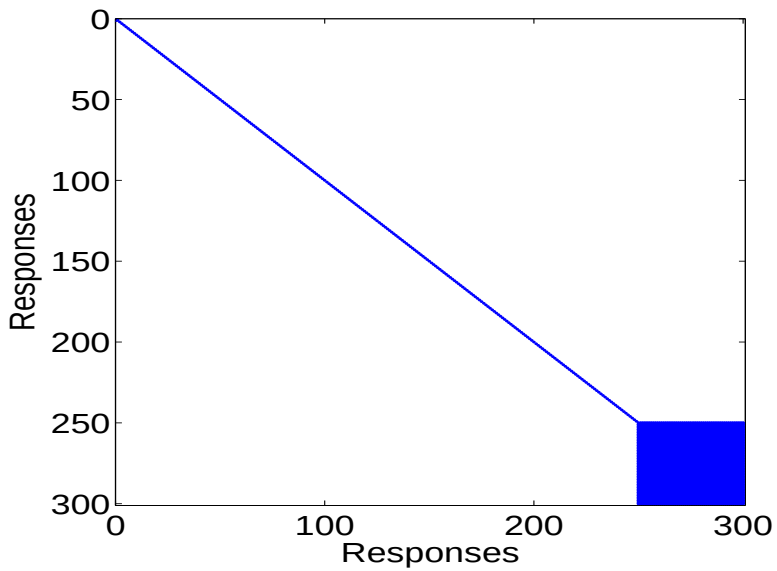
$$r(\mathbf{G}, \mathbf{G}^*) \quad = \quad \min \left\{ 1, \frac{f(\mathbf{t}|\mathbf{G}^*, \gamma) p(\mathbf{G}^*) q(\mathbf{G}|\mathbf{G}^*)}{f(\mathbf{t}|\mathbf{G}, \gamma) p(\mathbf{G}) q(\mathbf{G}^*|\mathbf{G})} \right\}.$$

Regeneration of $\mathbf{B}_{\gamma,\mathbf{G}}$ in the posterior

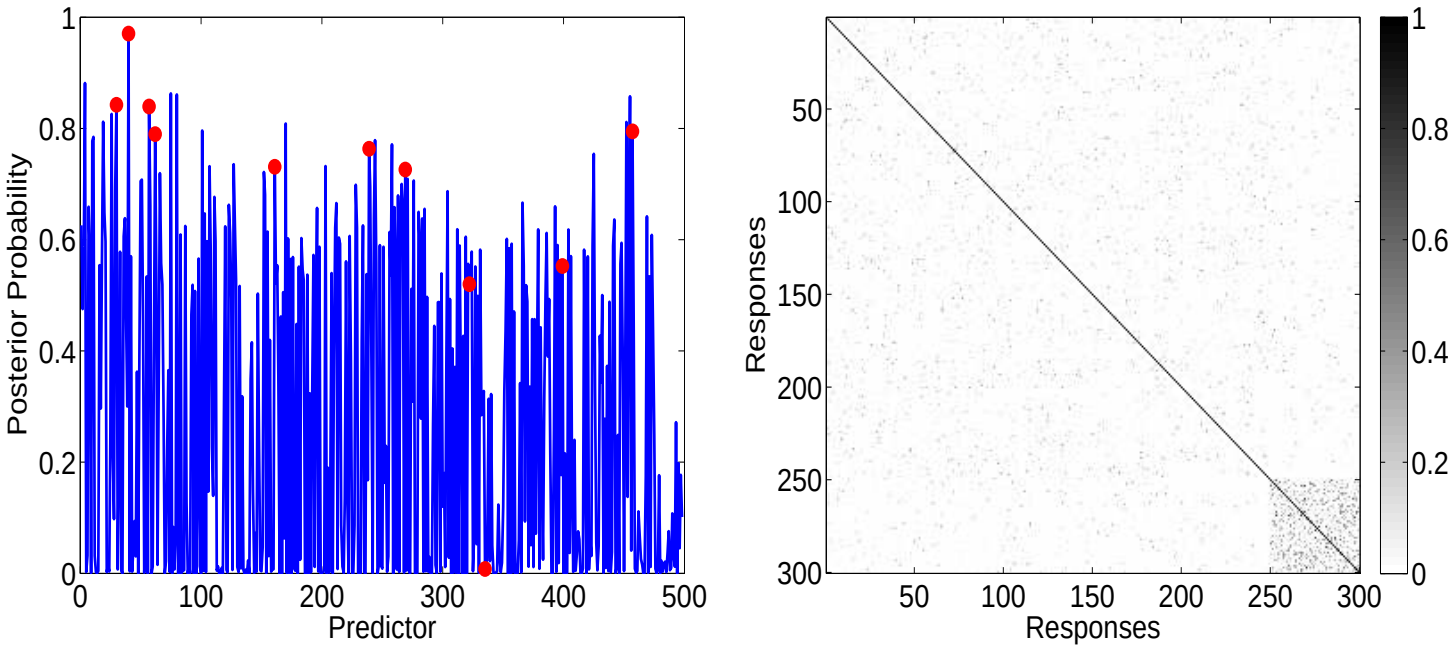
- $\mathbf{B}_{\gamma,\mathbf{G}}$ is the $p_\gamma \times q$ matrix of regression coefficients.
- By marginalizing it out we lose the association between the SNPs and expression levels necessary for an eQTL analysis.
- However, due to the conjugate structure, can be regenerated in the posterior conditional on $\hat{\gamma}$ and $\hat{\mathbf{G}}$.
- Generate $\mathbf{\Sigma}_G|\mathbf{Y}, \mathbf{B}_{\gamma,\mathbf{G}}, \gamma, G$ from $\mathrm{HIW}_G\{b+n, d\mathbf{I}_q + (\mathbf{Y} - \mathbf{X}_\gamma \mathbf{B}_{\gamma,\mathbf{G}})'(\mathbf{Y} - \mathbf{X}_\gamma \mathbf{B}_{\gamma,\mathbf{G}})\}$.
- Generate $\mathbf{B}_{\gamma,\mathbf{G}}|\mathbf{Y}, \mathbf{\Sigma}_G, \gamma, G$ from $\mathrm{MN}_{p_\gamma \times q}\{(\mathbf{X}'_\gamma \mathbf{X}_\gamma + c^{-1}\mathbf{I}_{p_\gamma})^{-1}\mathbf{X}'_\gamma \mathbf{Y}, (\mathbf{X}'_\gamma \mathbf{X}_\gamma + c^{-1}\mathbf{I}_{p_\gamma})^{-1}, \mathbf{\Sigma}_G\}$.

Simulation study 1

- We choose $p = 498$, $q = 300$ and $n = 120$.
- The eleven true predictors are $\{30, 40, 57, 62, 161, 239, 269, 322, 335, 399, 457\}$.
- True adjacency matrix for $\mathbf{G}$ is shown below.

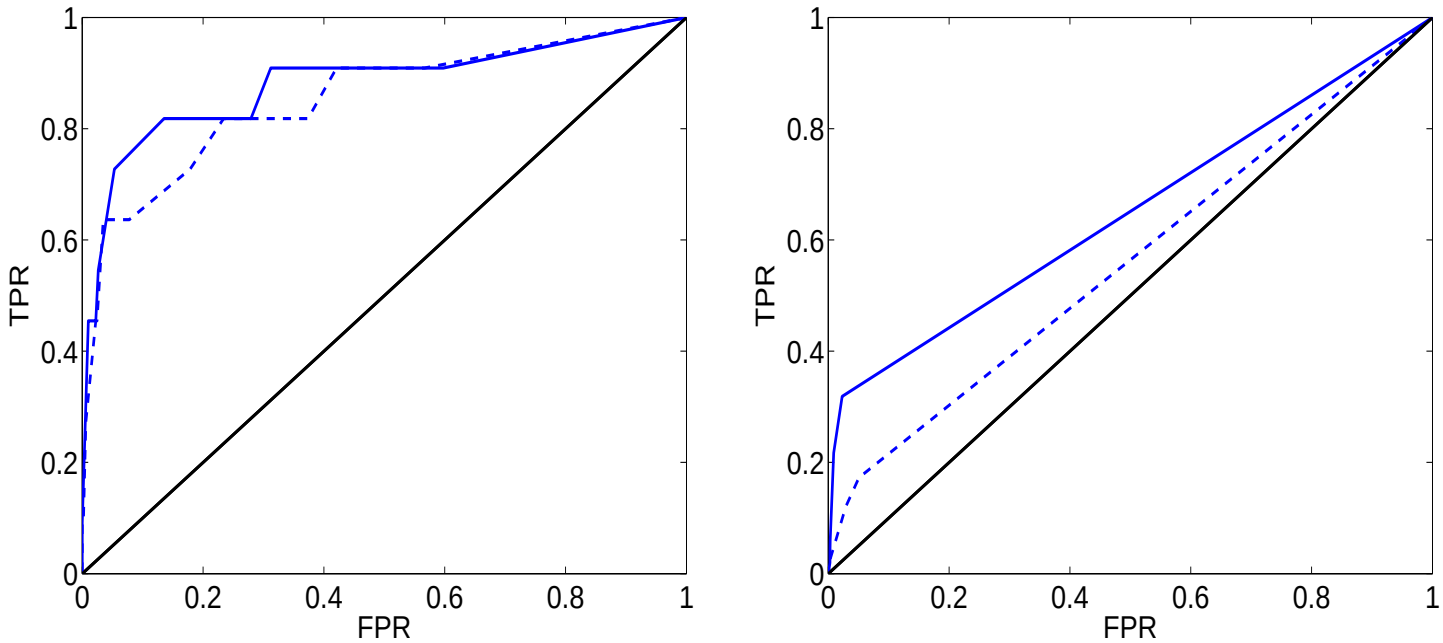


Results: Posterior probabilites



- Left: Posterior probabilities for γ, true variables circled in red.
- Right: Posterior probabilities for $\mathbf{G}$, compare with true graph.

Results: Does joint selection help over individual selection of variables and covariances?



- Left: ROC curve for γ, solid line: joint estimation, broken line: diagonal graph.
- Right: ROC curve for $\mathbf{G}$, solid line: joint estimation, broken line: zero mean model.

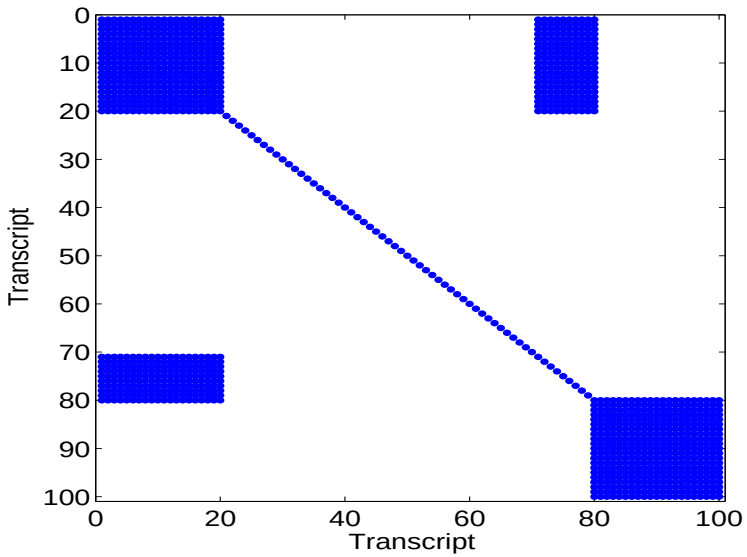
Simulation study 2

- We choose $p = 498$, $q = 100$ and $n = 120$.
- Consider 3 true predictors $\{30, 161, 239\}$. Associations between predictors and responses are generated according to following table:

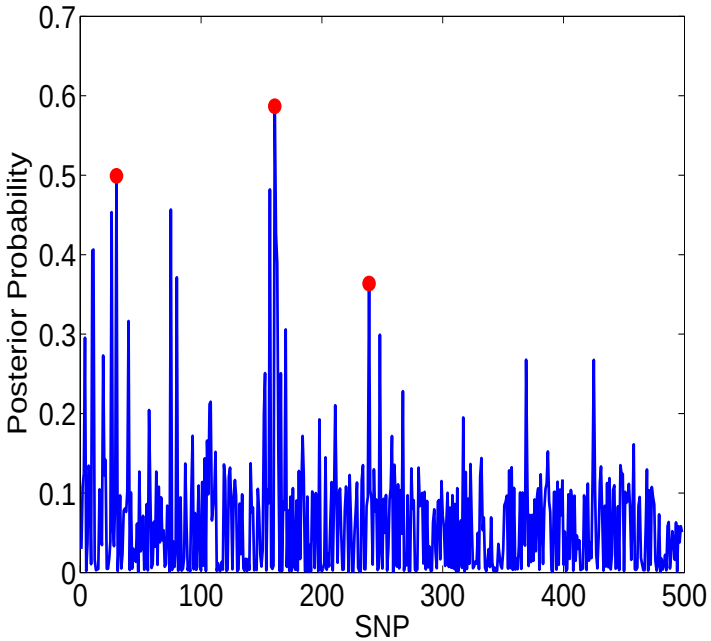
SNP ($\tilde{p}$)	Transcript ($\tilde{q}_p$)
30	1-20, 71-80
161	17-20
239	1-20, 71-80

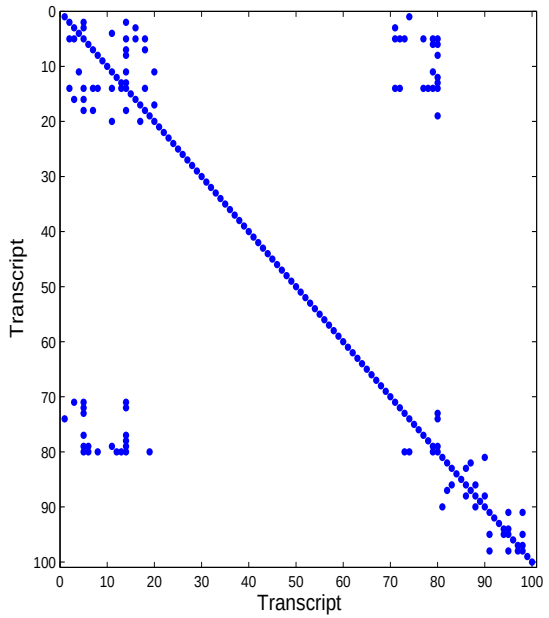
- Corresponding elements of $\mathbf{B}$ have sd 0.3.
- Rest of the responses are simulated from noise with sd 0.1.

Simulation study 2: The true graph



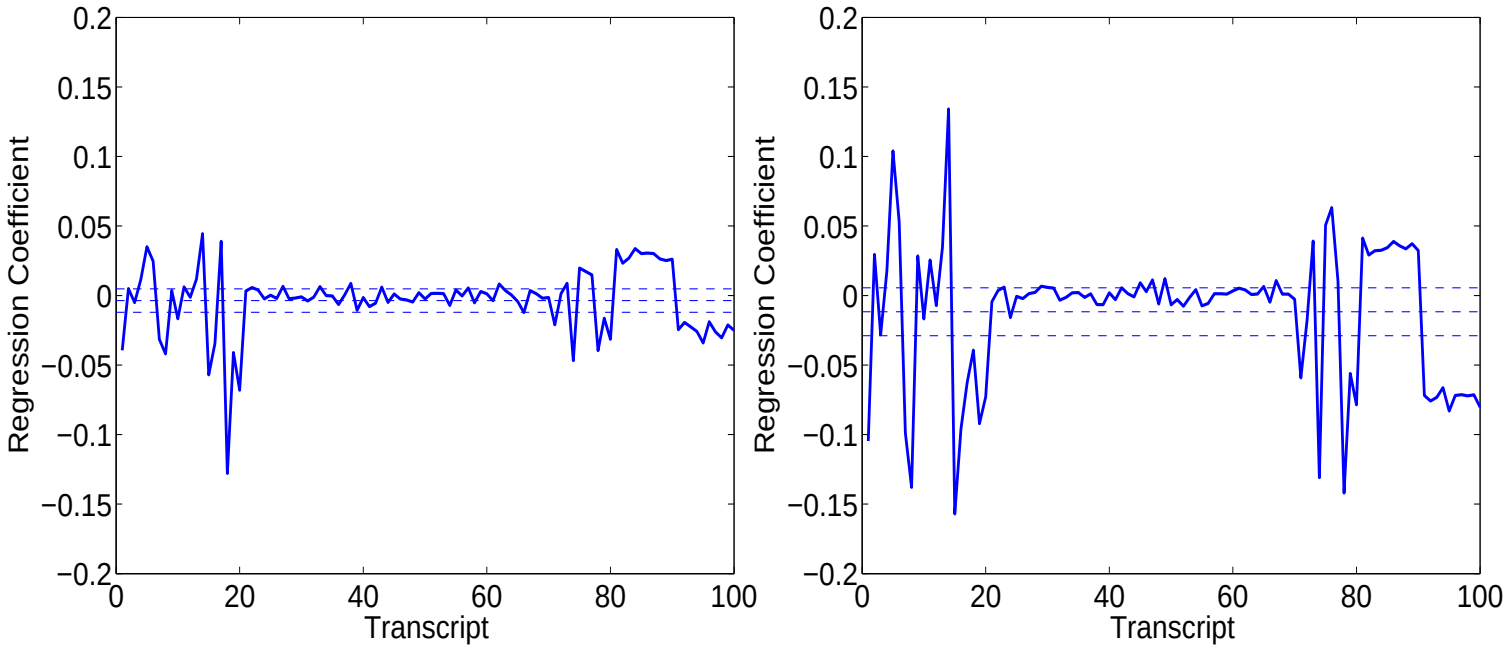
Results: Posterior probabilites





- Left: Posterior probabilities for γ, true variables circled in red.
- Right: Posterior probabilities for $\mathbf{G}$, with a cutoff on the posterior probabilities of edge inclusion set to 0.4

Results: Association analysis between SNPs and transcripts



- Left: Association of SNP 161 with all the 100 transcripts, showing enhanced association for transcripts 17-20.
- Right: association of SNP 239 with all the 100 transcripts, showing enhanced association for transcripts 1-20 and 71-80.

eQTL Analysis

- Essentially, this is a regression problem where $\mathbf{X}$ = An $n \times p$ matrix of SNPs (Single Neucleotide Polymorphisms) and $\mathbf{Y}$ = An $n \times q$ matrix of gene expression data, for the same set of n individuals.
- An eQTL analysis tries to infer the $p \times q$ matrix $\mathbf{B}$, trying to associate genetic variability to the gene expressions.
- It's long been known that the genes are a part of a regulatory/interaction nework.
- Statistically speaking, it is unreasonable to assume independence among the q traits.

Application to human eQTL analysis

- n = 60 unrelated individuals of Northern and Western European ancestry from Utah (CEU).
- SNP data publicly available from International Hapmap project (`http://hapmart.hapmap.org`).
- A total of p = 3125 SNPs found on 5' UTR of mRNA with minor allele frequency ≥ 0.1
- Gene expression data are also publicly available from the Sanger Institute website (`ftp://ftp.sanger.ac.uk/pub/genevar`).
- We work with q = 100 most variable transcripts out of a total of 47293.

Results

- Controlling for FDR at 5% level yields 8 globally significant SNPs and 38 non-zero inverse covariance matrix elements.
- Yields a total of 43 significant associations.
- Chen et al. (2008, Bioinformatics) detected a slightly higher number of associations by considering both 3' and 5' UTRs simultaneously.
- Yields a total of 55 significant edges.

Open questions and current investigations

- Could the technique be extended to more flexible models, e.g. models that can handle a nonlinear mean function?
- Is it possible to show simultaneous variable and graph selection consistency?
- What about non-Bayesian approaches?

References

1. **Bhadra, A.** and Mallick, B. K. (2012). Joint high-dimensional Bayesian variable and covariance selection with an application to eQTL analysis. *(under revision, Biometrics)*

2. Dawid, A. P. and Lauritzen, S. L. (1993). Hyper Markov laws in the statistical analysis of decomposable graphical models. *(Ann. Statist. 21, 1272 - 1317)*

3. Lauritzen, S. L. (1996). Graphical Models. *(Oxford University Press)*