Bayesian inference on high-dimensional Seemingly Unrelated Regressions

Alex Lewin and Marco Banterle
Brunel University London
April 2017
Genetic studies aiming at identifying association between point mutations (SNPs) and **multivariate** phenotypes:
- gene expression measurements
- metabolomics data
- protein concentrations
- ...

- Looking for sparse variable selection
- Take into account data correlations
Motivation

Multivariate Data

Aim: identify which of the $p$ variables in $X$ are significantly associated with the outcomes in $Y$.
Case study: mQTL discovery in the North Finland Birth Cohort study (NFBC)

- The NFBC66 is a cohort of 12000 adults followed since 1966
  - Collection of data at age 31 years, including clinical data and blood samples
  - DNA extracted leading to a sample of 5746 adults genotyped across genome (~ 300,000 SNPs)
  - Metabolite lipid profile quantified from serum samples by NMR, giving measures of 137 metabolites
- Question of interest is the discovery of genetic markers associated with metabolite regulation of lipids
- These responses are highly structured, with strong correlations
Correlations in the mQTL data set

Y: Metabolite correlations

X: SNP correlations

(only subset plotted)
Multivariate Regression Model with Variable Selection

Frame the problem as a multivariate linear regression model:

\[ Y_{n \times q} = X_{n \times p} B_{p \times q} + E_{n \times p} \]

or equivalently:

\[ Y \sim \mathcal{MN}(XB, I_n, R) \]

- Sparse associations: set most elements of B to zero
- Correlated outcomes: allow non-diagonal residual covariance matrix R
Variable selection performed through binary matrix $\Gamma \ (p \times q)$

$$
\gamma_{jk} = \begin{cases} 
1 & \text{outcome } k \text{ associated with predictor } j \\
0 & \text{else}
\end{cases}
$$

Sparsity prior $\gamma_{jk} \sim Bern(\omega_{jk}), \quad \omega_{jk} \sim Beta()$

Predictor $X_j$ only appears in a regression if $\gamma_{jk}$ is 1.
Bayesian Setting

- Very high dimensional data
  - $p \approx 10^4$ to $10^6$ variables in $X$
  - $q$ ranging from 1 to $10^4$ variables in $Y$
  - Around $n = 5000$ observations

- Focus on **Sparse Bayesian Variable Selection** (sparse BVS)
  - Minimise arbitrary tuning of model.
  - Provides the *posterior probability of association* for each predictor and each response.

- Bayesian model averaging
  - Explore space of $2^p$ models
  - Marginal posterior inclusion probabilities are model averages.
SUR model

Seemingly Unrelated Regressions (SUR) model:

\[
y_k = X_{\gamma_k} \beta_{\gamma_k} + \epsilon_k \quad \text{for } k = 1, \cdots, q
\]

- \( \text{Cov}[\epsilon_k, \epsilon_l] = R_{kl} \neq 0 \implies \) Outcomes do not naturally separate.
- Different variables selected for each outcome: \( \gamma_k, k = 1, \cdots, q \).

So: vectorise model:

\[
\text{vec}(y_1, y_2, \cdots y_q) \sim \mathcal{N}(\text{vec}(X_{\gamma_1} \beta_{\gamma_1}, X_{\gamma_2} \beta_{\gamma_2}, \cdots, X_{\gamma_q} \beta_{\gamma_q}), R \otimes \mathbb{I}_n)
\]

Covariance matrix is not block diagonal.
SUR model

Priors:

\[ \text{vec}(\beta_1, \beta_2, \cdots, \beta_q) | \gamma_1, \gamma_2, \cdots, \gamma_q \sim \mathcal{N}(0, (I_q \otimes W) \text{vec}(\gamma_1, \gamma_2, \cdots, \gamma_q)) \]

\[ R \sim \mathcal{IW}(\nu, M) \]

\[ \gamma_{kj} \sim \text{Bern}(\omega_k), \quad \omega_k \sim \text{Beta}(\cdot) \]

- \( W \) is a \( p \times p \) matrix (constant, or g-prior \( g(X^T X)^{-1} \)); different rows/columns selected for different outcomes.
- Not conjugate, cannot integrate out \( \beta \) and \( R \).
SUR model

Can calculate posterior full conditionals for $\beta_k$ and $R$ → Gibbs sampler for $\gamma_k, \beta_k$ and $R$.

However, this is computationally prohibitive due to calculation of

$$\left( \mathbf{X}' (R^{-1} \otimes I_n) \mathbf{X} \right)^{-1}$$

where

$$\mathbf{X} = \begin{pmatrix}
X_{\gamma_1} & 0 & \cdots & 0 \\
0 & X_{\gamma_2} & \cdots & 0 \\
0 & \cdots & \cdots & 0 \\
0 & \cdots & 0 & X_{\gamma_q}
\end{pmatrix}$$

an $nq \times (d_1 + d_2 + \cdots + d_q)$ matrix, which changes every MCMC iteration.
New work: computation for the SUR model

Idea from Zellner and Ando (2010): decompose the Likelihood:

\[
\begin{align*}
y_1 &= X_{\gamma_1} \beta_{\gamma_1} + \varepsilon_1 \\
y_2 &= X_{\gamma_2} \beta_{\gamma_2} + \rho_{21} (y_1 - X_{\gamma_1} \beta_{\gamma_1}) + \varepsilon_2 \\
&\vdots \\
y_k &= X_{\gamma_k} \beta_{\gamma_k} + \sum_{l<k} \rho_{kl} (y_l - X_{\gamma_l} \beta_{\gamma_l}) + \varepsilon_k
\end{align*}
\]

with \(\mathbb{E}[\varepsilon_k, \varepsilon_l] = \begin{cases} 0 & k \neq l \\ \sigma_k^2 \mathbb{I}_n & k = l \end{cases}\)

So Likelihood separates across separate responses.

Reparametrisation is \( R \leftrightarrow \{ \sigma_k^2, \rho_{kl} \} \)
New work: priors in the transformed space

We aim to decompose into product over responses, as with Likelihood.

Beta’s straightforward:

$$\prod_k \mathcal{N}(\beta_k | \gamma_k, W)$$

Covariance matrix:

$$R \sim \mathcal{IW}(\nu, M)$$

becomes

$$\prod_k \mathcal{N}(\{\rho_{k1}, \cdots, \rho_{k,k-1}\} | \sigma_k^2, M) \times \mathcal{IG}(\sigma_k^2 | \{\rho_{k1}, \cdots, \rho_{k,k-1}\}, \nu, M)$$
New: posterior conditionals in transformed space

Covariance matrix

In transformed space,

\[
\prod_k \mathcal{N}(\{\rho_{k1}, \cdots, \}\mid \sigma_k^2, M, Y, B, \Gamma) \times \mathcal{IG}(\sigma_k^2\mid \{\rho_{k1}, \cdots, \}, \nu, M, Y, B, \Gamma)
\]

So MCMC updates for \(R\) parameters factorise over responses.

Betas

MCMC for \(B\) not so straightforward: Zellner and Ando used simplified factorisation + Gibbs resampling
New: posterior conditionals in transformed space

We found the correct full conditionals for $B$ (does decompose):

$$
\beta_{\gamma_k} \mid \ldots \sim \mathcal{N} \left( W_k \times X^t_{\gamma_k} \tilde{y}_k, W_k \right) \quad \text{(k = 1, \ldots ,q)}
$$

where

$$
W_k = \left( X^t_{\gamma_k} X_{\gamma_k} \left( \frac{1}{\sigma_k^2} + \sum_{l>k} \frac{\rho_{lk}^2}{\sigma_l^2} \right) + W_{\gamma_k}^{-1} \right)^{-1},
$$

$$
\tilde{y}_k = \text{function of } Y, B, \Gamma \text{ across responses}
$$
New work: computation for the SUR model

- Main point: we have got rid of the $\left( X^t (R^{-1} \otimes I_n) X \right)^{-1}$ big matrix calculation.

- Using ESS (evolutionary stochastic search) algorithm (Bottolo et al.) to explore space of $\Gamma$ variable selection parameters.

These in R; in C++ will be greater relative speedup.
Simulated data

- $n=100$
- $q=30$, $p=30$
- with correlated residuals
Simulated data

True Gamma matrix

HESS Gamma
Full heat map

SUR Gamma
Full heat map
mQTL analysis of NFBC data

After quality control,
\( n = 4023 \) people
\( q = 103 \) metabolites
\( p = 9172 \) SNPs on chromosome 16
Evidence of enhanced linkage for Chromosome 16
Summary

- Bayesian SUR model with sparsity prior to perform variable selection for multiple responses.
- Estimating the residual covariance matrix increases the accuracy of the variable selection.
- We have extended the Zellner and Ando method to obtain directly the correct posteriors.
- Computational speed-up → model can be used on large genomic data sets.
Thank you!

- Sylvia Richardson
- Leonardo Bottolo
- Marjo-Riitta Jarvelin
- Habib Saadi
- Marc Chadeau-Hyam

**Lewin A et al. (2015)**
MT-HESS: an efficient Bayesian approach for simultaneous association detection in OMICS datasets, with application to eQTL mapping in multiple tissues, *Bioinformatics* 10.1093.

**Bottolo L, Chadeau-Hyam, M et al. (2013)**

**Bhadra A and Mallick BK (2013)**
Joint high-dimensional Bayesian variable and covariance selection with an application to eQTL analysis, *Biometrics* 10.1111.

**Bottolo L, Petretto, E et al. (2011)**
Bayesian detection of expression quantitative trait loci hot spots, *Genetics* 189:1449–1459.
Model-based clustering for multivariate categorical data

Michael Fop, Keith Smart and Brendan Murphy

6th IBS Channel Network Conference
A study to investigate the use of a mechanisms-based classification of musculoskeletal pain in clinical practice.

The aim of the study was to assess the discriminative power of the taxonomy of pain in Nociceptive, Peripheral Neuropathic and Central Sensitization for low-back disorders.

There are $N = 464$ patients who were assessed according to a list of 36 binary clinical indicators (“Present”/“Absent”).

Some of the indicators carry the same information about the pain categories, thus the interest here is to select a subset of most relevant clinical criteria, performing a partition of the patients.

Does the partition of the patients agree with the clinical taxonomy?
The motivating example shows the need for:

- **Clustering:** Can we establish the existence of subgroups? How can we characterize these subgroups?

- **Variable Selection:** Can we use a subset of the variables to distinguish the subgroups?
Denote the $N \times M$ data matrix by $X$.

The $n$th observation is denoted by $X_n$.

Model-based clustering assumes that $X_n$ arises from a finite mixture model.

Assuming $G$ classes (components)

$$p(X_n|\tau, \theta, G) = \sum_{g=1}^{G} \tau_g p(X_n|\theta_g).$$

$\tau_g$ are mixture weights.

$p(X_n|\theta_g)$ is the component distribution.
Latent Class Analysis (LCA) model

- Latent Class Analysis (LCA) is a model for clustering categorical data.
- Let \( X_n = (X_{n1}, X_{n2}, \ldots, X_{nM}) \) where \( X_{nm} \) takes a value from \( \{1, 2, \ldots, C_m\} \).
- In LCA we assume that there is local independence between variables, so that if we knew \( X_n \) was in class \( g \) we could write it's density as

\[
p(X_n | \theta_g) = \prod_{m=1}^{M} \prod_{c=1}^{C_m} \theta^{I(X_{nm}=c)}_{gm},
\]

where \( \{\theta_{gm1}, \ldots, \theta_{gmC_m}\} \) give the probabilities of observing the categories \( \{1, \ldots, C_m\} \) in variable \( m \)

- \( \theta_g \) will characterize and embody the differences between groups
LCA model (general)

- Model likelihood of the form,

\[
p(X_n | \theta, \tau, G) = \sum_{g=1}^{G} \tau_g \prod_{m=1}^{M} \prod_{c=1}^{C_m} \theta^{I(X_{nm}=c)}_{gmc}.
\]

- More convenient to work with completed data

- Augment data with class labels \( Z_n = (Z_{n1}, Z_{n2}, \ldots, Z_{nG}) \) where

\[
Z_{ng} = \begin{cases} 
1 & \text{if observation } n \text{ belongs to group } g \\
0 & \text{otherwise.}
\end{cases}
\]

- Then we can write down completed data likelihood for an observation

\[
p(X_n, Z_n | \theta, \tau, G) = \prod_{g=1}^{G} \left\{ \tau_g \prod_{m=1}^{M} \prod_{c=1}^{C_m} \theta^{I(X_{nm}=c)}_{gmc} \right\}^{Z_{ng}}.
\]
LCA model (general)

- Estimation by EM algorithm or VB (see BayesLCA package)

- Note that $G$ must be chosen in advance; possible to discriminate the best $G$ for the data using information criteria (e.g., BIC)

- Bayesian approaches:
  - Pandolfi, Bartolucci and Friel (2014) use reversible jump to get posterior probability for $G$.
  - White, Wyse and Murphy (2016) use a collapsed Gibbs sampler and incorporate variable selection.
Back Pain Data: LCA Results
# Back Pain Data: LCA Clustering

<table>
<thead>
<tr>
<th></th>
<th>1</th>
<th>2</th>
<th>3</th>
<th>4</th>
<th>5</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>Central Sensitization</strong></td>
<td>48</td>
<td>1</td>
<td>5</td>
<td>0</td>
<td>41</td>
</tr>
<tr>
<td><strong>Nociceptive</strong></td>
<td>0</td>
<td>10</td>
<td>96</td>
<td>126</td>
<td>3</td>
</tr>
<tr>
<td><strong>Peripheral Neuropathic</strong></td>
<td>0</td>
<td>89</td>
<td>1</td>
<td>1</td>
<td>4</td>
</tr>
</tbody>
</table>
Dean & Raftery (2010) proposed a greedy stepwise variable selection algorithm for LCA. The observation vector $X_n$ is partitioned as

$$X_n = (X_n^C, X_n^P, X_n^O)$$

where

- $X_n^C$ are the current clustering variables.
- $X_n^P$ is proposed to be added to the clustering variables.
- $X_n^O$ are the other variables.
Two competing models are compared:

- $\mathcal{M}_1$ assumes that the proposed variable has clustering structure.
- $\mathcal{M}_2^*$ assumes that the proposed variable has no clustering structure.

This framework reduces the independence assumption of the previously described approach.
Local Independence (A Problem?)

- When analyzing the back pain data, we achieved very little data reduction.
- In fact, only one variable was labeled as non-clustering.
- An explanation for this is the *local independence* assumption in the model.
- Suppose we have two variables that are highly dependent and both exhibit clustering.
- The variable selection method will include both variables in the model, even if one variable contains no extra clustering information.
It is unrealistic to assume that $X^C_n$ and $X^P_n$ are conditionally independent in $M^*_2$.

We propose replacing $M^*_2$ with a different model.

$M_1$ assumes that the proposed variable has clustering structure.

$M_2$ assumes that the proposed variable has no clustering structure beyond that explained by the clustering variables.
We propose a stepwise search algorithm to find an *optimal* set of variables for clustering.

The algorithm involves the following steps:

- **Add:** Add a variable to the current clustering variables.
- **Remove:** Remove a variable from the current clustering variables.
- **Swap:** Swap a proposed variable with one already in the clustering variables.

Model selection is implemented using the Bayesian Information Criterion (BIC).
The proposed model was applied to the back pain data:

<table>
<thead>
<tr>
<th>Variables</th>
<th>N. latent classes</th>
<th>BIC</th>
<th>ARI</th>
</tr>
</thead>
<tbody>
<tr>
<td>All</td>
<td>5</td>
<td>-12582.62</td>
<td>0.50</td>
</tr>
<tr>
<td>All</td>
<td>3*</td>
<td>-12763.81</td>
<td>0.82</td>
</tr>
<tr>
<td>35 Criteria</td>
<td>5</td>
<td>-12116.32</td>
<td>0.50</td>
</tr>
<tr>
<td>35 Criteria</td>
<td>3*</td>
<td>-12305.67</td>
<td>0.80</td>
</tr>
<tr>
<td>11 Criteria</td>
<td>3</td>
<td>-3965.24</td>
<td>0.75</td>
</tr>
</tbody>
</table>

The new model achieves much greater data reduction.
## Algorithm Run

<table>
<thead>
<tr>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>Remove Crit.5</td>
<td>-122.2</td>
<td>Accepted</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>2</td>
<td>Remove Crit.23</td>
<td>-126.3</td>
<td>Accepted</td>
<td>Swap Crit.22 with Crit.5</td>
<td>-73.2</td>
<td>Rejected</td>
</tr>
<tr>
<td>3</td>
<td>Remove Crit.38</td>
<td>-109.0</td>
<td>Accepted</td>
<td>Swap Crit.25 with Crit.5</td>
<td>-81.5</td>
<td>Rejected</td>
</tr>
<tr>
<td>4</td>
<td>Remove Crit.4</td>
<td>-103.5</td>
<td>Accepted</td>
<td>Swap Crit.2 with Crit.38</td>
<td>-98.6</td>
<td>Rejected</td>
</tr>
<tr>
<td>5</td>
<td>Remove Crit.1</td>
<td>-78.3</td>
<td>Accepted</td>
<td>Swap Crit.29 with Crit.4</td>
<td>-23.1</td>
<td>Rejected</td>
</tr>
<tr>
<td>6</td>
<td>Remove Crit.29</td>
<td>-73.2</td>
<td>Accepted</td>
<td>Swap Crit.12 with Crit.1</td>
<td>2.7</td>
<td>Accepted</td>
</tr>
<tr>
<td>7</td>
<td>Remove Crit.1</td>
<td>-73.5</td>
<td>Accepted</td>
<td>Swap Crit.26 with Crit.29</td>
<td>3.2</td>
<td>Accepted</td>
</tr>
<tr>
<td>8</td>
<td>Remove Crit.29</td>
<td>-66.8</td>
<td>Accepted</td>
<td>Swap Crit.18 with Crit.12</td>
<td>-10.2</td>
<td>Rejected</td>
</tr>
<tr>
<td>9</td>
<td>Remove Crit.35</td>
<td>-63.0</td>
<td>Accepted</td>
<td>Swap Crit.7 with Crit.29</td>
<td>-9.0</td>
<td>Rejected</td>
</tr>
<tr>
<td>10</td>
<td>Remove Crit.7</td>
<td>-59.6</td>
<td>Accepted</td>
<td>Swap Crit.11 with Crit.35</td>
<td>-7.6</td>
<td>Rejected</td>
</tr>
<tr>
<td>11</td>
<td>Remove Crit.10</td>
<td>-62.9</td>
<td>Accepted</td>
<td>Swap Crit.8 with Crit.7</td>
<td>-76.1</td>
<td>Rejected</td>
</tr>
<tr>
<td>12</td>
<td>Remove Crit.11</td>
<td>-50.4</td>
<td>Accepted</td>
<td>Swap Crit.16 with Crit.10</td>
<td>6.8</td>
<td>Accepted</td>
</tr>
<tr>
<td>13</td>
<td>Remove Crit.8</td>
<td>-54.5</td>
<td>Accepted</td>
<td>Swap Crit.10 with Crit.16</td>
<td>-32.0</td>
<td>Rejected</td>
</tr>
<tr>
<td>14</td>
<td>Remove Crit.3</td>
<td>-44.2</td>
<td>Accepted</td>
<td>Swap Crit.31 with Crit.16</td>
<td>-9.5</td>
<td>Rejected</td>
</tr>
<tr>
<td>15</td>
<td>Remove Crit.31</td>
<td>-33.2</td>
<td>Accepted</td>
<td>Swap Crit.18 with Crit.16</td>
<td>-22.7</td>
<td>Rejected</td>
</tr>
<tr>
<td>16</td>
<td>Remove Crit.22</td>
<td>-30.9</td>
<td>Accepted</td>
<td>Swap Crit.24 with Crit.23</td>
<td>-1.7</td>
<td>Rejected</td>
</tr>
<tr>
<td>17</td>
<td>Remove Crit.14</td>
<td>-22.7</td>
<td>Accepted</td>
<td>Swap Crit.32 with Crit.31</td>
<td>-5.0</td>
<td>Rejected</td>
</tr>
<tr>
<td>18</td>
<td>Remove Crit.32</td>
<td>-19.2</td>
<td>Accepted</td>
<td>Swap Crit.37 with Crit.14</td>
<td>-8.0</td>
<td>Rejected</td>
</tr>
<tr>
<td>19</td>
<td>Remove Crit.10</td>
<td>-35.4</td>
<td>Accepted</td>
<td>Swap Crit.9 with Crit.3</td>
<td>-1.3</td>
<td>Rejected</td>
</tr>
<tr>
<td>20</td>
<td>Remove Crit.24</td>
<td>-17.6</td>
<td>Accepted</td>
<td>Swap Crit.30 with Crit.8</td>
<td>15.7</td>
<td>Accepted</td>
</tr>
<tr>
<td>21</td>
<td>Remove Crit.34</td>
<td>-15.7</td>
<td>Accepted</td>
<td>Swap Crit.37 with Crit.1</td>
<td>-0.7</td>
<td>Rejected</td>
</tr>
<tr>
<td>22</td>
<td>Remove Crit.25</td>
<td>-13.7</td>
<td>Accepted</td>
<td>Swap Crit.36 with Crit.1</td>
<td>3.3</td>
<td>Accepted</td>
</tr>
<tr>
<td>23</td>
<td>Remove Crit.18</td>
<td>-10.5</td>
<td>Accepted</td>
<td>Swap Crit.1 with Crit.31</td>
<td>8.5</td>
<td>Accepted</td>
</tr>
<tr>
<td>24</td>
<td>Remove Crit.27</td>
<td>-13.7</td>
<td>Accepted</td>
<td>Swap Crit.6 with Crit.26</td>
<td>6.1</td>
<td>Accepted</td>
</tr>
<tr>
<td>25</td>
<td>Remove Crit.31</td>
<td>-1.3</td>
<td>Accepted</td>
<td>Swap Crit.20 with Crit.6</td>
<td>5.6</td>
<td>Accepted</td>
</tr>
<tr>
<td>26</td>
<td>Remove Crit.37</td>
<td>1.4</td>
<td>Rejected</td>
<td>Swap Crit.6 with Crit.5</td>
<td>-3.1</td>
<td>Accepted</td>
</tr>
<tr>
<td>27</td>
<td>Remove Crit.5</td>
<td>0.4</td>
<td>Rejected</td>
<td>Swap Crit.37 with Crit.20</td>
<td>4.0</td>
<td>Rejected</td>
</tr>
</tbody>
</table>
The clustering closely follows the clinical taxonomy.

<table>
<thead>
<tr>
<th></th>
<th>Class 1</th>
<th>Class 2</th>
<th>Class 3</th>
</tr>
</thead>
<tbody>
<tr>
<td>Nociceptive</td>
<td>210</td>
<td>21</td>
<td>4</td>
</tr>
<tr>
<td>Peripheral Neuropathic</td>
<td>5</td>
<td>88</td>
<td>2</td>
</tr>
<tr>
<td>Central Sensitiization</td>
<td>3</td>
<td>3</td>
<td>89</td>
</tr>
</tbody>
</table>

It is not unusual for patients diagnosed as Nociceptive may have Peripheral Neuropathic aspects to their back pain.
The selected variables exhibit strong clustering across the three groups.
The chosen variables have the following descriptions.

<table>
<thead>
<tr>
<th>Crit.</th>
<th>Description</th>
<th>Class 1</th>
<th>Class 2</th>
<th>Class 3</th>
</tr>
</thead>
<tbody>
<tr>
<td>2</td>
<td>Pain associated to trauma, pathologic process or dysfunction</td>
<td>0.94</td>
<td>0.90</td>
<td>0.04</td>
</tr>
<tr>
<td>5</td>
<td>Usually intermittent and sharp with movement/mechanical provocation</td>
<td>0.94</td>
<td>0.84</td>
<td>0.24</td>
</tr>
<tr>
<td>8</td>
<td>Pain localized to the area of injury/dysfunction</td>
<td>0.97</td>
<td>0.50</td>
<td>0.31</td>
</tr>
<tr>
<td>9</td>
<td>Pain referred in a dermatomal or cutaneous distribution</td>
<td>0.06</td>
<td>1.00</td>
<td>0.11</td>
</tr>
<tr>
<td>13</td>
<td>Disproportionate, nonmechanical, unpredictable pattern of pain</td>
<td>0.01</td>
<td>0.00</td>
<td>0.91</td>
</tr>
<tr>
<td>15</td>
<td>Pain in association with other dysesthesias</td>
<td>0.03</td>
<td>0.51</td>
<td>0.34</td>
</tr>
<tr>
<td>19</td>
<td>Night pain/disturbed sleep</td>
<td>0.34</td>
<td>0.70</td>
<td>0.86</td>
</tr>
<tr>
<td>26</td>
<td>Pain in association with high levels of functional disability</td>
<td>0.07</td>
<td>0.36</td>
<td>0.79</td>
</tr>
<tr>
<td>28</td>
<td>Clear, consistent and proportionate pattern of pain</td>
<td>0.97</td>
<td>0.94</td>
<td>0.07</td>
</tr>
<tr>
<td>33</td>
<td>Diffuse/nonanatomic areas of pain/tenderness on palpation</td>
<td>0.03</td>
<td>0.01</td>
<td>0.73</td>
</tr>
<tr>
<td>37</td>
<td>Pain/symptom provocation on palpation of relevant neural tissues</td>
<td>0.07</td>
<td>0.57</td>
<td>0.19</td>
</tr>
</tbody>
</table>
Discarded Variables

- Many of the discarded variables are related with the clustering variables.

- These are not clustering variables because they don’t exhibit clustering *beyond* what can be explained by the clustering variables.
Summary

- Model-based approaches to clustering and variable selection achieve excellent performance.
- Removing independence assumptions in the model achieves improved variable selection. Care needed interpreting the chosen/discarded variables.
Simulation 1

\[
\begin{align*}
  &Z \\
  & \downarrow \\
  & X_1 \quad X_2 \quad X_3 \quad X_4 \\
  & \downarrow \quad \downarrow \quad \downarrow \quad \downarrow \\
  & X_5 \quad X_6 \quad X_7 \quad X_8 \\
  & \downarrow \quad \downarrow \quad \downarrow \quad \downarrow \\
  & X_9 \quad X_{10} \quad X_{11} \quad X_{12}
\end{align*}
\]
Simulation 1 Results
Simulation 2
Simulation 2 Results
Exploring the dependence structure between categorical variables: Benefits and limitations of using variable selection within Bayesian clustering

Michail Papathomas

6th IBS Channel Network Conference - Hasselt - 2017
Main collaborators in the development of the clustering approach [profile regression, based on the Dirichlet process]:

- Silvia Liverani
- David Hastie
- John Molitor
- Sylvia Richardson

Work on the relation between clustering and log-linear modelling with Sylvia Richardson
Assume observations from $P$ categorical variables $\{x_1, \ldots, x_P\}$. For example,

<table>
<thead>
<tr>
<th>Subject</th>
<th>Smoking (X)</th>
<th>Drinking (Y)</th>
</tr>
</thead>
<tbody>
<tr>
<td>John</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td>Mary</td>
<td>0</td>
<td>1</td>
</tr>
<tr>
<td>Jim</td>
<td>1</td>
<td>1</td>
</tr>
</tbody>
</table>

The resulting data can be arranged as counts in a $P$-way contingency table. For example,

**Table 1:** Smoking (X) and Drinking (Y).

<table>
<thead>
<tr>
<th>Smoking (X)</th>
<th>Drinking (Y)</th>
<th></th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td>No (0)</td>
<td>Yes (1)</td>
<td></td>
<td></td>
</tr>
<tr>
<td>No (0)</td>
<td>456</td>
<td>44</td>
<td></td>
</tr>
<tr>
<td>Yes (1)</td>
<td>583</td>
<td>911</td>
<td></td>
</tr>
</tbody>
</table>
Denote the cell counts as $n_l$, $l = 1, \ldots, n$.

A Poisson distribution is assumed for the counts so that $E(n_l) = \mu_l$.

A Poisson log-linear model $\log(\mu) = X_{DM}\lambda$ is a GLM that relates the expected counts to the variables.

<table>
<thead>
<tr>
<th>Table 1: Smoking (X) and drinking (Y).</th>
</tr>
</thead>
<tbody>
<tr>
<td>Smoking (X)</td>
</tr>
<tr>
<td>No (0)</td>
</tr>
<tr>
<td>No (0)</td>
</tr>
<tr>
<td>Yes (1)</td>
</tr>
</tbody>
</table>
Motivation

Dependence structure and log-linear modelling

Sometimes, the dependence structure between \( \{x_1, \ldots, x_P\} \) (marginal and conditional independence) can be inferred by the form of the log-linear model. For example,

\[
\log(\mu_{ij}) = \lambda + \lambda_i^X + \lambda_j^Y + \lambda_{ij}^{XY}
\]

implies that \( X \) and \( Y \) are dependent.

\[
\log(\mu_{ij}) = \lambda + \lambda_i^X + \lambda_j^Y
\]

implies marginal independence.

In practice, for more than 3 variables, interpreting the dependence structure by looking at the presence/absence of interaction terms becomes too difficult.
Motivation

Dependence structure and log-linear modelling

Joint probabilities for the categorical variables are obtained using the parameters of the log-linear model. In our simple example,

$$P(X = i, Y = j) = \frac{\mu_{i,j}}{\sum_{i',j'=0,1} \mu_{i',j'}}.$$ 

Then, for instance,

$$P(\text{smoker, drinker}) = P(X = 1, Y = 1) = \frac{\mu_{1,1}}{\mu_{0,0} + \mu_{1,0} + \mu_{0,1} + \mu_{1,1}} = \frac{\exp(\lambda + \lambda_X^1 + \lambda_Y^1 + \lambda_{XY}^{1,1})}{\exp(\lambda) + \exp(\lambda + \lambda_X^1) + \ldots}.$$ 

- So, in principle, we could fully explore the variables’ dependence structure by calculating many joint probabilities and considering the laws of probability
- or, even better, by considering graphical log-linear models
Graphical models

- They allow to visualize and build complex dependence structures for the covariates under consideration.
- Undirected Graphs and Directed Acyclic Graphs allow to define and write complex joint distributions through factorization and conditional independence; Lauritzen (2011).
- Neighborhoods of models are easily defined, and it is straightforward to move in the space of models by adding, removing or replacing edges.
Example of a graphical model
Problems with Log-linear modelling for detecting interactions

When it is of interest to detect interactions between covariates, linear regression modelling may become problematic.

- In a classical setting, fitting linear models with many parameters sometimes requires an impractically large vector of observations for valid inferences (Burton et al., IJE, 2009). Also, identifiability and collinearity problems are often present.

- In Bayesian model comparison, the space of models becomes vast, and model search algorithms like the Reversible Jump approach (Green, Bka 1995) require an impractically large number of iterations before they converge (Dobra and Massam, St Meth 2010).
Bayesian clustering with the Dirichlet process

- Partitions the subjects into groups according to their profile
- Flexible Bayesian clustering
- Uncertainty with regard to the clustering is evaluated
- Post-processing leads to tractable output
Notation

- Consider categorical variables $x_p$, $p = 1, ..., P$.
- For individual $i$, denote the variable/covariate profile by $x_i = (x_{i1}, ..., x_{iP})$.
- For example,
  - $x_{1}$: smokes, does not smoke
  - $x_{2}$: drinks, does not drink
  - $x_{3}$: exercises, does not exercise
- Now, for instance, $x_i=$(smokes, drinks, does not exercise).
Notation

For individual $i$

- $z_i = c$ allocates subject, $i$, to cluster $c$.
- $\phi^c_p(x)$ the probability that variable $x_p = x$, for $z_i = c$.
- Given $z_i = c$, $x_p$ has a multinomial distribution with cluster specific parameters $\phi^c_p = [\phi^c_p(1), \ldots, \phi^c_p(M_p)]$
- A priori, $\phi^c_p \sim \text{Dirichlet}(\lambda_1, \ldots, \lambda_{M_p})$
- $\psi_c$ denotes the probability that a subject is assigned to cluster $c$. 

(University of St Andrews)
Bayesian clustering with the Dirichlet process

Statistical Framework

For $\phi = \{\phi^c_p, c \in N, p = 1, \ldots, P\}$,

- ‘stick-breaking’ prior on the allocation weights $\psi_c$
- $x_1, \ldots, x_P$ are assumed independent given the clustering allocation and parameters...
- and calculating joint probabilities for the categorical variables becomes easy!
- $Pr(x_i | z, \phi) = \prod_{p=1}^{P} \phi^z_i (x_{ip})$ for $i = 1, 2, \ldots, n$.

This implies

$$Pr(x_i | \phi, \psi) = \sum_{c=1}^{\infty} Pr(Z_i = c | \psi) \prod_{p=1}^{P} Pr(x_{ip} | Z_i = c) = \sum_{c=1}^{\infty} \psi_c \prod_{p=1}^{P} \phi^c_p(x_{ip}).$$
Using 0-1 variable selection switches

- Identify covariates that contribute more than others to the formation of clusters. [Tadesse et al. (2005), Chung and Dunson (2009); Papathomas et al. (2012)]

- Cluster specific binary indicators, $\gamma^c_p$, so that $\gamma^c_p = 1$ when covariate $x_p$ is important for allocating subjects to cluster $c$; otherwise $\gamma^c_p = 0$.

- Prior for switches: given $\rho_p$, $\gamma^c_p \sim \text{Bernoulli}(1, \rho_p)$.

- We consider a sparsity inducing prior for $\rho$ with an atom at zero:

$$\rho_p \sim 1_{\{w_p=0\}} \delta_0(\rho_p) + 1_{\{w_p=1\}} \text{Beta}(\alpha_\rho, \beta_\rho)$$

where $w_j \sim \text{Bernoulli}(0.5)$.

Similar to Chung and Dunson (JASA, 2009), but in their set up, covariate observations contribute to the likelihood through a regression model. In our case, covariate observations contribute directly to the likelihood, and we introduce $\pi_p(x)$. 
Example - Using the R package PReMiuM; Liverani et al. (2015, JSS)

- Simulated observations from 10000 subjects, recording...
- 10 binary variables, say \{SMO, DRI, EXE, D, E, ..., H, I, J\}.
Example
Table 2: Cluster profiles. In parenthesis the number of subjects typically allocated to each group.

<table>
<thead>
<tr>
<th></th>
<th>Simulation 1</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>SMO</td>
</tr>
<tr>
<td>Median($\rho_p$)</td>
<td>0.36</td>
</tr>
<tr>
<td>Group 1 (5465)</td>
<td>$&gt;&lt;$</td>
</tr>
<tr>
<td>Group 2 (3159)</td>
<td>$&lt;&gt;$</td>
</tr>
<tr>
<td>Group 3 (1376)</td>
<td>00</td>
</tr>
</tbody>
</table>
It is not clear how clustering output translates to interactions in a log-linear regression modelling framework.

Can we assist the process of comparing a large number of linear models with the clustering variable selection results?

The important aspect of a model that combines clustering and variable selection is that covariates are not chosen in accordance with size of marginal effect. They are selected because they combine to create distinct groups of subjects. Consequently, we expect that this type of modelling should be able to inform on interactions in a linear model setting.
Theoretical results

**Theorem 1:** Consider random variables $x_p$ and $x_q$, $1 \leq p, q \leq P$, $p \neq q$. If $\sum_{c=1}^{C} \gamma^c_p \times \gamma^c_q = 0$ then $x_p$ and $x_q$ are independent.

**Theorem 2:** Consider a set of random variables $\{x_1, \ldots, x_P\}$. If, for some $p \in \{1, \ldots, P\}$, $\sum_{c=1}^{C} \gamma^c_p \times \gamma^c_q = 0$, for all $q \neq p$, then $x_p$ is independent of $\{x_1, \ldots, x_P\} \setminus x_p$.

**Proofs:** See Papathomas and Richardson (2016, JSPI). Note that the converse is not true.

The previous Theorems imply the following Corollary,

**Corollary:** Consider covariate $x_p$. If $\sum_{c=1}^{C} \gamma^c_p = 0$ then $x_p$ is independent from all other covariates.
Theoretical results

Therefore,

- if the selection probability $\rho_p$ for $x_p$ is zero or close to zero, something that implies that $\sum_{c=1}^{C} \gamma_p^c$ is also zero or close to zero, we can assume that $x_p$ is independent from all other covariates.

- Assuming that our interest lies in exploring interactions, to reduce the dimensionality of the problem when fitting linear models to sparse contingency tables, $x_p$ could be removed from the analysis.
Simulated data using graphical log-linear models
Construction of matrix $T_\gamma$

- For iteration $i_t$ and for each cluster $c$ with more than one subject, form matrix $T^{c,i_t}$, so that element $(p_1, p_2), 1 \leq p_1 < p_2 \leq P$ is either zero or one, and equal to $\gamma^c_{p_1}(i_t) \times \gamma^c_{p_2}(i_t)$. All other matrix cells are empty.

- Sum up all matrices $T^{c,i_t}$, weighing by cluster size, to create an information matrix $T_\gamma$,

$$T_\gamma = \sum_{i_t} \sum_c n_{c,i_t} \times T^{c,i_t}.$$  

where $n_{c,i_t}$ is the size of cluster $c$ at iteration $i_t$. Therefore, $T_\gamma$ is a straightforward summary of all $T^{c,i_t}$ matrices into one, with small clusters contributing less to this summary.

- For ease of interpretation reweight the elements of $T_\gamma$ so that the maximum element is one, $T_\gamma = (\max\{T_\gamma\})^{-1} \times T_\gamma$.

Matrix $T_\gamma$ is constructed in such a manner so that if element $t_\gamma(p_1, p_2), 1 \leq p_1 < p_2 \leq P$, is close to zero, this implies that an edge between $x.p_1$ and $x.p_2$ is not likely to be present in a highly supported graphical model.
Example. Simulation 1

- Simulated observations from 10000 subjects.
- 10 binary categorical variables \( \{A, B, C, \ldots, H, I, J\} \) are observed.
- Observations are simulated in accordance with the log-linear model

\[
\log(\mu) = \lambda + \lambda^A + \lambda^B + \lambda^C + \ldots + \lambda^J + \lambda^{AB} + \lambda^{BC} + \lambda^{CD} + \lambda^{DA} + \lambda^{HI} + \lambda^{IJ} + \lambda^{HJ} + \lambda^{HIJ}
\]

- (For more details see Papathomas and Richardson (2016)).
Example. Simulation 1

\[
\log(\mu) = \lambda + \lambda^A + \lambda^B + \lambda^C + ... + \lambda^J + \lambda^{AB} + \lambda^{BC} + \lambda^{CD} + \lambda^{DA} + \lambda^{HI} + \lambda^{IJ} + \lambda^{HJ} + \lambda^{HIJ}
\]
Example. Simulation 1

\[ \log(\mu) = \lambda + \lambda^A + \lambda^B + \lambda^C + \ldots + \lambda^J + \lambda^{AB} + \lambda^{BC} + \lambda^{CD} + \lambda^{DA} + \lambda^{HI} + \lambda^{IJ} + \lambda^{HJ} + \lambda^{HIJ} \]

Table 2: Cluster profiles. In parenthesis the number of subjects typically allocated to each group.

<table>
<thead>
<tr>
<th></th>
<th>A</th>
<th>B</th>
<th>C</th>
<th>D</th>
<th>E</th>
<th>F</th>
<th>G</th>
<th>H</th>
<th>I</th>
<th>J</th>
</tr>
</thead>
<tbody>
<tr>
<td>Median((\rho_P))</td>
<td>0.36</td>
<td>0.78</td>
<td>0.32</td>
<td>0.75</td>
<td>0.06</td>
<td>0.05</td>
<td>0.00</td>
<td>0.48</td>
<td>0.57</td>
<td>0.50</td>
</tr>
<tr>
<td>Group 1 (5465)</td>
<td>&gt;&gt;</td>
<td>&lt;&lt;</td>
<td>00</td>
<td>&lt;&lt;</td>
<td>00</td>
<td>00</td>
<td>00</td>
<td>&gt;&gt;</td>
<td>&gt;&gt;</td>
<td>&gt;&gt;</td>
</tr>
<tr>
<td>Group 2 (3159)</td>
<td>&lt;&lt;</td>
<td>&gt;&gt;</td>
<td>00</td>
<td>&gt;&gt;</td>
<td>00</td>
<td>00</td>
<td>00</td>
<td>&gt;&gt;</td>
<td>&gt;&gt;</td>
<td>&gt;&gt;</td>
</tr>
<tr>
<td>Group 3 (1376)</td>
<td>00</td>
<td>&gt;&gt;</td>
<td>00</td>
<td>00</td>
<td>00</td>
<td>00</td>
<td>00</td>
<td>00</td>
<td>&gt;&gt;</td>
<td>&gt;&gt;</td>
</tr>
</tbody>
</table>


Example. Simulation 1

$T_{\gamma}^{sim1} = \begin{pmatrix}
A & B & C & D & E & F & G & H & I & J \\
A & .52 & .08 & .50 & .04 & .02 & .02 & .20 & .27 & .15 \\
B & .45 & 1 & .06 & .04 & .03 & .47 & .64 & .47 \\
C & .45 & .02 & .02 & .009 & .12 & .23 & .16 \\
D & .06 & .04 & .03 & .45 & .65 & .48 \\
E & .003 & .003 & .03 & .04 & .03 \\
F & .002 & .02 & .03 & .03 \\
G & .02 & .02 & .02 \\
H & 0.61 & .56 \\
I & .74
\end{pmatrix}$
**Example. Simulation 1**

Table 3: Mixing performance of samplers. Median of iterations to best model is calculated after 30 runs of the reversible jump MCMC chain. First and third quartiles are given in parentheses. PDV denotes the unrefined model search strategy adopted in Papathomas et al (2011b). See Figure 2 for the highest posterior probability model.

<table>
<thead>
<tr>
<th>Scenario</th>
<th>Acceptance rate as a percentage</th>
<th>Iterations (median) to highest posterior probability model</th>
<th>Posterior probability for highest probability model</th>
</tr>
</thead>
<tbody>
<tr>
<td>(a) Uniformly random (PDV)</td>
<td>5.1</td>
<td>590 (452,821)</td>
<td>0.55</td>
</tr>
<tr>
<td>(b) Cluster specific</td>
<td>3.8</td>
<td>247 (164,369)</td>
<td>0.55</td>
</tr>
<tr>
<td>(c) Combined (30%,10%)</td>
<td>5.3</td>
<td>540 (290,674)</td>
<td>0.53</td>
</tr>
<tr>
<td>(d) Combined (20%,20%)</td>
<td>4.9</td>
<td>403 (312,493)</td>
<td>0.55</td>
</tr>
</tbody>
</table>
Real data example

- 30 single nucleotide polymorphisms (SNPs) in chromosomes 6 and 15.
  (Data from 4260 subjects in a genome-wide association study of lung cancer presented in Hung et al. (2008).)

- 12 SNPs were indicated as important by variable selection within clustering.
  (Two from chromosome 15 and ten from chromosome 6.)

- SNPs were highly correlated. 3 SNPs included in the competing log-linear graphical models as representatives.
  rs8034191 from chromosome 15 and \{rs4324798,rs1950081\} from chromosome 6.

- Also include age, gender and smoking status in the competing log-linear graphical models, to search for gene-environment interactions.
Reducing the number of SNPs from 30 to 12, and then to 3, allows for the use of reversible jump MCMC to compare competing graphical models. The $2^{33}$ contingency table would be too sparse with the vast majority of cells equal to zero.

The highest posterior probability model ($P=0.8$) is

‘SNP1+SNP2+SNP3+AGE*GENDER*SMOKING’

which does not support the presence of gene-gene or gene-environment interactions.
Real data example

‘SNP1 + SNP2 + SNP3 + AGE * GENDER * SMOKING’

Table 6: Cluster profiles. In parenthesis the number of subjects typically allocated to each group.

<table>
<thead>
<tr>
<th>Genetic-environmental data (GE)</th>
<th>rs8034191 (A)</th>
<th>rs4324798 (B)</th>
<th>rs1950081 (C)</th>
<th>age (D)</th>
<th>gender (E)</th>
<th>smoking (F)</th>
</tr>
</thead>
<tbody>
<tr>
<td>Median((\rho_p))</td>
<td>0.01</td>
<td>0.00</td>
<td>0.10</td>
<td>0.92</td>
<td>0.82</td>
<td>0.85</td>
</tr>
<tr>
<td>Cluster 1 (2222)</td>
<td>00</td>
<td>00</td>
<td>00</td>
<td>&gt;&gt;</td>
<td>&gt;&gt;</td>
<td>&lt;&lt;</td>
</tr>
<tr>
<td>Cluster 2 (2059)</td>
<td>00</td>
<td>00</td>
<td>00</td>
<td>&lt; &gt;</td>
<td>&lt; &gt;</td>
<td>&gt;&gt;</td>
</tr>
</tbody>
</table>
Real data example

’SNP1 + SNP2 + SNP3 + AGE * GENDER * SMOKING’

\[ T^\gamma_{\text{Real data}} = \begin{pmatrix}
  S1 & S2 & S3 & AGE & GEN & SM \\
  S1 & 0.002 & 0.01 & 0.06 & 0.06 & 0.06 \\
  S2 & 0.001 & 0.02 & 0.02 & 0.02 & \\
  S3 & 0.09 & 0.07 & 0.08 & \\
  AGE & 1 & .98 & \\
  GEN & & & .88
\end{pmatrix} \]
Real data example

**Table 7:** Mixing performance of samplers. Median of iterations to best model is calculated after 300 runs of the reversible jump MCMC chain. First and third quartiles are given in parentheses. PDV denotes the unrefined model search strategy adopted in Papathomas et al (2011b).

<table>
<thead>
<tr>
<th>Genetic-environmental data [including important (characterized as such by clustering) representative SNPs]</th>
<th>Acceptance rate as a percentage</th>
<th>Iterations (median) to highest posterior probability model</th>
<th>Posterior probability for highest probability model ‘A+B+C+DEF’</th>
</tr>
</thead>
<tbody>
<tr>
<td>(a) Uniformly random</td>
<td>6.3</td>
<td>564 (257,1205)</td>
<td>0.53</td>
</tr>
<tr>
<td>(b) Cluster specific</td>
<td>8.4</td>
<td>196 (83,443)</td>
<td>0.51</td>
</tr>
<tr>
<td>(c) Combined (30%,10%)</td>
<td>6.9</td>
<td>310 (147,670)</td>
<td>0.51</td>
</tr>
<tr>
<td>(d) Combined (20%,20%)</td>
<td>7.5</td>
<td>235 (91,516)</td>
<td>0.52</td>
</tr>
</tbody>
</table>
Relevant work on latent class structures and log-linear modelling (Johndrow et al., 2014)

- Johndrow et al. (2014) consider standard and novel latent class structures. The DP is a special case.
- Its rank is defined as the minimum number of clusters required to describe the joint probability tensor for the categorical covariates.
- Bounds are derived for the rank, in relation to the number and structure of the interactions in a weakly hierarchical log-linear model.
- A massive reduction in the upper bound of the rank is shown, under a sparse log-linear model.
- The rank of the latent structure depends only on variables that are not marginally independent.
- A straightforward application gives that an upper bound of the rank corresponding to simulation 1 is $2^7$, rather than the default $2^9$. The upper bound corresponding to simulation 5 is $2^8$, rather than the default $2^{99}$. 

(University of St Andrews)
Relevant work on latent class structures and log-linear modelling (Zhou et al., 2015)

- Zhou et al. (2015) also utilize the idea that marginally independent variables reduce the dimensionality of the problem.
- A PARAFAC factorization is adopted, which can be viewed as a more general representation of the Dirichlet process.
- Dimensionality reduction is achieved with the sparse PARAFAC (sp-PARAFAC) formulation, where marginal independence is modelled with quantities of similar nature to the $\pi_p(x)$.
- The focus is in providing expressions for parameters of the log-linear models, assessing the level of shrinkage, and the convergence of the probability tensor induced by sp-PARAFAC to the true probability tensor.
- The prior formulation for detecting marginally independent covariates and reducing dimensionality is also different in the two approaches.
- Different objectives, as we focus on accelerating log-linear model selection with the Reversible Jump by utilizing the clustering process.
Summary

- The advantage in utilizing variable selection within partitioning to inform log-linear model selection is mostly pertinent to marginal independence.
- For sparse contingency tables, this information can lead to the substantial reduction of the number of covariates considered, making the exploration of the model space feasible.
- Informing the model search algorithm with $T_\gamma$ often improves the efficiency of the search. Marginal independence is not always detected, because the converse of the Theorems does not hold.
- Importantly, using $T_\gamma$ to assist the model search never resulted in a worse algorithm, compared to the standard model search approach in Papathomas et al. (2011b).
Further work

- Sparse contingency tables and conditional independence
- Improved model search algorithms
- Sparse contingency tables and identifiability
Some References

Fast sampling with Gaussian scale-mixture priors in high-dimensional regression

Bani K Mallick
(joint with Anirban Bhattacharya & Antik Charaborty)
Department of Statistics, Texas A&M University

April 13, 2017
Outline

- High Dimensional Regression
- Global-Local scale Mixtures of Gaussians
- Efficient sampling from structures multivariate Gaussian distributions
- Applications
High dimensional regression with \( p \) covariates and sample size \( n \)

\( p \) is much larger than \( n \)

Sparsity through continuous shrinkage priors

Need efficient computational scheme for \( p \) greater than \( n \) problems
$\mathbf{Y} = \mathbf{X}\beta + \epsilon, \epsilon \sim \mathcal{N}(0, \sigma^2 I_n)$

item $\mathbf{X} : n \times p$ matrix of covariates where $p$ potentially much larger than $n$

In such setting, one expects $\beta$, the vector of regression coefficients to be sparse

Global-local sparse prior on $\beta$
Global-local Prior

- $\beta_j | \lambda_j, \tau, \sigma \sim N(0, \lambda_j^2 \tau^2 \sigma^2), (j = 1, \ldots, p)$
- $\lambda_j \sim f, \tau \sim g, \sigma \sim h$
- $f, g, h$ are densities supported on $(0, \infty)$
- $\lambda_j^2$: local variance component works for scale mixing
- $\tau^2$: global variance component (like the regularization parameter in the penalized likelihood formulation)
Create different sparsity priors by choosing $f$ the distribution of $\lambda_j$

- Student-t [Tipping 2001]: $f$ is inverse-gamma
- Double-Exponential [Park and Casella, 2008]: $f$ exponential
- Normal/Jeffreys [Bae and Mallick, 2004]: Jeffrey’s Prior
- Horseshoe Prior [Carvalho et al., 2010]: Half-Cauchy
Horseshoe Priors

- Flat Cauchy like tails allow strong signals to remain large (un-shrunk)
- Tall spike at the origin provides severe shrinkage for the zero elements of $\beta$
- Due to scale mixture of Gaussian formulation most of the conditional distributions are in explicit form
The conditional posterior of $\beta$ given $\lambda, \tau$ and $\sigma$ is Gaussian:

$$\beta \mid y, \lambda, \tau, \sigma \sim \mathcal{N}(A^{-1}X^Ty, \sigma^2 A^{-1}), \quad A = (X^TX + D^{-1}),$$

where $D = \tau^2 \text{diag}(\lambda_1^2, \ldots, \lambda_p^2)$.

- Need to invert $A$ which is $p \times p$.
- Covariance no longer diagonal.
- The design matrix $X$ distorts the prior geometry.
- Need to sample from a high-dimensional Gaussian per iteration.
- The $p$ local scale parameters $\lambda_j$ have conditionally independent posteriors: $\lambda = (\lambda_1, \ldots, \lambda_p)^T$ updated in a block.
Computational Difficulties

- A standard algorithm to sample from Gaussian distributions can be found in (Rue, 2001)
- It avoids inverting $A$ and instead performs a Cholesky decomposition of $A$ and a series of linear system solutions to generate samples
- This is efficient for moderate values of $p$, obtaining a Cholesky decomposition of $A$ at each MCMC step
A be an $n \times n$ Symmetric, Positive Definite Matrix

Cholesky Decomposition: $L$ is a lower triangular matrix where $L_{ii} > 0$ and $A = LL^T$

Algorithm Solving $Ax = b$ where $A > 0$

(i) Compute the Cholesky decomposition $A = LL^T$.
(ii) Solve $Lv = b$.
(iii) Solve $L^Tx = v$.
(iv) Return $x$

As $x = (L^{-1})^Tv = L^T(L^{-1}b) = (LL^T)^{-1}b = A^{-1}b$

The solutions in (ii) and (iii) can be done through forward and backward substitution due to the triangular nature of $L$
Computational Difficulties

- It becomes highly expensive for large $p$
- One cannot resort to precomputing the Cholesky factors since the matrix $D$ changes from one iteration to the other
- The resulting computational bottleneck obscures the computational advantages of global-local priors when $p$ is large
General scheme

- Given $X \in \mathbb{R}^{n \times p}$, $D \in \mathbb{R}^{p \times p}$ p.d. & $Y \in \mathbb{R}^{n \times 1}$ with $p \gg n$
- Without Loss of Generality assume $\sigma = 1$
- Goal: sample from $N_p(\mu, \Sigma)$ with

$$\Sigma = (X^T X + D^{-1})^{-1}, \quad \mu = \Sigma X^T Y.$$ 

- $D$ need not be diagonal as long as $D^{-1}$ is easy to calculate
- Precision $Q = \Sigma^{-1} = (X^T X + D^{-1})$ and $b = X^T Y$, write $\mu = Q^{-1} b$. 
Rue (2001): sampling from $\mathcal{N}(\mu = Q^{-1}b, \Sigma = Q^{-1})$ with density

**Algorithm Rue (2001)**

(i) Compute the Cholesky decomposition $Q = LL^T$.
(ii) Solve $Lv = b$.
(iii) Solve $L^Tm = v$.
(iv) Solve $L^Tw = z$, where $z \sim \mathcal{N}(0, I_p)$.
(vi) Set $\theta = m + w$. Then, $\theta \sim \mathcal{N}(\mu, \Sigma)$.

Original motivation: GMRFs where $Q$ is banded.

Cholesky in step (i) can be computed very fast.
Present setting: \( Q = (X^T X + D^{-1}) \) & \( b = X^T Y \).

**Algorithm Rue (2001)**

(i) Compute the Cholesky decomposition \((X^T X + D^{-1}) = LL^T\).
(ii) Solve \( Lv = X^T Y \).
(iii) Solve \( L^T m = v \).
(iv) Solve \( L^T w = z \), where \( z \sim N(0, I_p) \).
(vi) Set \( \theta = m + w \). Then, \( \theta \sim N(\mu, \Sigma) \).

- \((X^T X + D^{-1})\) is a \( p \times p \) dense matrix.
- Cholesky decomposition costly in present setting.
- Overall complexity \( O(p^3) \).
Our proposal

- Sample from $\mathcal{N}_p(\mu, \Sigma)$ with

$$\Sigma = (X^TX + D^{-1})^{-1}, \quad \mu = \Sigma X^TY.$$

- Woodbury matrix identity,

$$\Sigma = (X^TX + D^{-1})^{-1} = D - DX^T(XDX^T + I_n)^{-1}XD.$$

- Can show $\mu = DX^T(XDX^T + I_n)^{-1}Y$.

- Sample $\eta \sim \mathcal{N}(0, \Sigma)$ & set $\theta = \mu + \eta$.

- Data augmentation - sample an $(n + p)$ dimensional quantity.
Our proposal

- $P \in \mathbb{R}^{n \times n}$ and $R \in \mathbb{R}^{p \times p}$ are invertible and $S \in \mathbb{R}^{n \times p}$.

$$
\begin{pmatrix}
P & S \\
S^T & R
\end{pmatrix} =
\begin{pmatrix}
I_n & 0 \\
S^T P^{-1} & I_p
\end{pmatrix}
\begin{pmatrix}
P & 0 \\
0 & R - S^T P^{-1} S
\end{pmatrix}
\begin{pmatrix}
I_n & P^{-1} S \\
0 & I_p
\end{pmatrix}
$$

- $\Omega$ can be recognized as a covariance matrix.

- Set $P = (XDX^T + I_n)$, $S = XD$ and $R = D$.

- $\Omega$ can be recognized as a covariance matrix.

- Let $u \sim N_p(0, D)$, $\delta \sim N_n(0, I_n)$ independent and $v = Xu + \delta$.

- Set $\zeta = (v^T, u^T)^T \in \mathbb{R}^{n+p}$. Then, $\zeta \sim N(0, \Omega)$.

- Upshot: easy to sample from $N_{n+p}(0, \Omega)$. 
Our proposal

- \( P \in \mathbb{R}^{n \times n} \) and \( R \in \mathbb{R}^{p \times p} \) are invertible and \( S \in \mathbb{R}^{n \times p} \).

\[
\begin{pmatrix}
P & S \\
S^T & R
\end{pmatrix}
= \begin{pmatrix}
I_n & 0 \\
S^T P^{-1} & I_p
\end{pmatrix}
\begin{pmatrix}
P & 0 \\
0 & \Gamma
\end{pmatrix}
\begin{pmatrix}
I_n & P^{-1} S \\
0 & I_p
\end{pmatrix}
\]

\( \Omega = \begin{pmatrix} L & \Gamma \end{pmatrix} \begin{pmatrix} L^T \\ \Gamma \end{pmatrix} \)

- Set \( P = (XDX^T + I_n) \), \( S = XD \) and \( R = D \).

- \( \Omega \) can be recognized as a covariance matrix.

- Let \( u \sim N_p(0, D), \delta \sim N_n(0, I_n) \) independent and \( v = Xu + \delta \).

- Set \( \zeta = (v^T, u^T)^T \in \mathbb{R}^{n+p} \). Then, \( \zeta \sim N(0, \Omega) \).

- Upshot: easy to sample from \( N_{n+p}(0, \Omega) \).
Our proposal

- $P \in \mathbb{R}^{n \times n}$ and $R \in \mathbb{R}^{p \times p}$ are invertible and $S \in \mathbb{R}^{n \times p}$.

\[
\left( \begin{array}{cc}
P & S \\
S^T & R
\end{array} \right) \Omega = \left( \begin{array}{cc}
I_n & 0 \\
S^T P^{-1} & I_p
\end{array} \right) \left( \begin{array}{cc}
P & 0 \\
0 & R - S^T P^{-1} S
\end{array} \right) \left( \begin{array}{cc}
I_n & P^{-1} S \\
0 & I_p
\end{array} \right)
\]

- Set $P = (XDX^T + I_n)$, $S = XD$ and $R = D$.

- $\Omega$ can be recognized as a covariance matrix.

- Let $u \sim N_p(0, D)$, $\delta \sim N_n(0, I_n)$ independent and $v = Xu + \delta$.

- Set $\zeta = (v^T, u^T)^T \in \mathbb{R}^{n+p}$. Then, $\zeta \sim N(0, \Omega)$.

- Upshot: easy to sample from $N_{n+p}(0, \Omega)$. 
Our proposal

- $P \in \mathbb{R}^{n \times n}$ and $R \in \mathbb{R}^{p \times p}$ are invertible and $S \in \mathbb{R}^{n \times p}$.

\[
\begin{pmatrix}
P & S \\ S^T & R
\end{pmatrix} = \begin{pmatrix}
I_n & 0 \\ S^TP^{-1} & I_p
\end{pmatrix} \begin{pmatrix}
P & 0 \\ 0 & R - S^TP^{-1}S
\end{pmatrix} \begin{pmatrix}
I_n & P^{-1}S \\ 0 & I_p
\end{pmatrix}
\]

- $\Omega$ can be recognized as a covariance matrix.

- Set $P = (XDX^T + I_n)$, $S = XD$ and $R = D$.

- $\Omega$ can be recognized as a covariance matrix.

- Let $u \sim N_p(0, D)$, $\delta \sim N_n(0, I_n)$ independent and $v = Xu + \delta$.

- Set $\zeta = (v^T, u^T)^T \in \mathbb{R}^{n+p}$. Then, $\zeta \sim N(0, \Omega)$.

- Upshot: easy to sample from $N_{n+p}(0, \Omega)$. 
Our proposal

- $P \in \mathbb{R}^{n \times n}$ and $R \in \mathbb{R}^{p \times p}$ are invertible and $S \in \mathbb{R}^{n \times p}$.

$$
\begin{pmatrix}
P & S \\
S^T & R
\end{pmatrix}
= 
\begin{pmatrix}
I_n & 0 \\
S^TP^{-1} & I_p
\end{pmatrix}
\begin{pmatrix}
P & 0 \\
0 & R - S^TP^{-1}S
\end{pmatrix}
\begin{pmatrix}
I_n & P^{-1}S \\
0 & I_p
\end{pmatrix}
= \Omega
$$

- Set $P = (XDX^T + I_n)$, $S = XD$ and $R = D$.

- $\Omega$ can be recognized as a covariance matrix.

- Let $u \sim \mathcal{N}_p(0, D)$, $\delta \sim \mathcal{N}_n(0, I_n)$ independent and $v = Xu + \delta$.

- Set $\zeta = (v^T, u^T)^T \in \mathbb{R}^{n+p}$. Then, $\zeta \sim \mathcal{N}(0, \Omega)$.

- Upshot: easy to sample from $\mathcal{N}_{n+p}(0, \Omega)$. 
Our proposal

- \( P \in \mathbb{R}^{n \times n} \) and \( R \in \mathbb{R}^{p \times p} \) are invertible and \( S \in \mathbb{R}^{n \times p} \).

\[
\begin{pmatrix}
P & S \\ S^T R
\end{pmatrix} = \begin{pmatrix}
I_n & 0 \\ S^T P^{-1} & I_p
\end{pmatrix} \begin{pmatrix}
P & 0 \\ 0 & R - S^T P^{-1} S
\end{pmatrix} \begin{pmatrix}
I_n & P^{-1} S \\ 0 & I_p
\end{pmatrix}
\]

\( \Omega = L \Gamma L^T \)

- Set \( P = (XDX^T + I_n) \), \( S = XD \) and \( R = D \).

- \( \Omega \) can be recognized as a covariance matrix.

- Let \( u \sim \mathcal{N}_p(0, D) \), \( \delta \sim \mathcal{N}_n(0, I_n) \) independent and \( v = Xu + \delta \).

- Set \( \zeta = (v^T, u^T)^T \in \mathbb{R}^{n+p} \). Then, \( \zeta \sim \mathcal{N}(0, \Omega) \).

- Upshot: easy to sample from \( \mathcal{N}_{n+p}(0, \Omega) \).
Our proposal

- $P \in \mathbb{R}^{n \times n}$ and $R \in \mathbb{R}^{p \times p}$ are invertible and $S \in \mathbb{R}^{n \times p}$.

$$
\begin{pmatrix}
P & S \\
S^T & R
\end{pmatrix} =
\begin{pmatrix}
I_n & 0 \\
S^T P^{-1} & I_p
\end{pmatrix}
\begin{pmatrix}
P & 0 \\
0 & R - S^T P^{-1} S
\end{pmatrix}
\begin{pmatrix}
I_n & P^{-1} S \\
0 & I_p
\end{pmatrix}
\Omega =
\begin{pmatrix}
L & \Gamma \\
\Gamma^T & L^T
\end{pmatrix}
$$

- Recall $P = (XD \Phi^T + I_n)$, $S = XD$ and $R = D$.
- Easy to sample from $N(0, \Omega)$.
- Key observation:

$$
R - S^T P^{-1} S = D - DX^T(XDX^T + I_n)^{-1} XD = \Sigma.
$$

- The lower $p \times p$ diagonal block of $\Gamma$ is $\Sigma$. 

Our proposal

- \( P \in \mathbb{R}^{n \times n} \) and \( R \in \mathbb{R}^{p \times p} \) are invertible and \( S \in \mathbb{R}^{n \times p} \)

\[
\begin{pmatrix}
P & S \\
S^T & R
\end{pmatrix}
= \begin{pmatrix}
I_n & 0 \\
S^T P^{-1} & I_p
\end{pmatrix}
\begin{pmatrix}
P & 0 \\
0 & R - S^T P^{-1} S
\end{pmatrix}
\begin{pmatrix}
I_n & P^{-1} S \\
0 & I_p
\end{pmatrix}
\]

\[
\Omega = L \Gamma L^T
\]

- Write \( \Gamma = L^{-1} \Omega (L^{-1})^T \).

- \( L \) lower triangular, so \( L^{-1} \) easy to calculate

\[
L^{-1} = \begin{pmatrix}
I_n & 0 \\
-S^T P^{-1} & I_p
\end{pmatrix}
\]

- Summary: sample \( \zeta \sim \mathcal{N}(0, \Omega) \), set \( \zeta^* = L^{-1} \zeta \) and collect the last \( p \) entries of \( \zeta^* \), say \( \eta \).
Our proposal

- \( P \in \mathbb{R}^{n \times n} \) and \( R \in \mathbb{R}^{p \times p} \) are invertible and \( S \in \mathbb{R}^{n \times p} \)

\[
\begin{align*}
\begin{pmatrix} P & S \\ S^T & R \end{pmatrix} &= \begin{pmatrix} I_n & 0 \\ S^T P^{-1} & I_p \end{pmatrix} \begin{pmatrix} P & 0 \\ 0 & R - S^T P^{-1} S \end{pmatrix} \begin{pmatrix} I_n & P^{-1} S \\ 0 & I_p \end{pmatrix} \\
\Omega &= \begin{pmatrix} I_{n0} & 0 \\ 0 & I_{p0} \end{pmatrix}
\end{align*}
\]

- Write \( \Gamma = L^{-1} \Omega (L^{-1})^T \).
  - \( L \) lower triangular, so \( L^{-1} \) easy to calculate

\[
L^{-1} = \begin{pmatrix} I_n & 0 \\ -S^T P^{-1} & I_p \end{pmatrix}
\]

- Summary: sample \( \zeta \sim \mathcal{N}(0, \Omega) \), set \( \zeta^* = L^{-1} \zeta \) and collect the last \( p \) entries of \( \zeta^* \), say \( \eta \).
Our proposal

- \( P \in \mathbb{R}^{n \times n} \) and \( R \in \mathbb{R}^{p \times p} \) are invertible and \( S \in \mathbb{R}^{n \times p} \)

\[
\begin{pmatrix}
P & S \\
S^T & R
\end{pmatrix}
= \begin{pmatrix}
I_n & 0 \\
S^T P^{-1} & I_p
\end{pmatrix}
\begin{pmatrix}
P & 0 \\
0 & R - S^T P^{-1} S
\end{pmatrix}
\begin{pmatrix}
I_n & P^{-1} S \\
0 & I_p
\end{pmatrix}
\]

\( \Omega = \left( L \Gamma L^T \right) \)

- Write \( \Gamma = L^{-1} \Omega (L^{-1})^T \).
- \( L \) lower triangular, so \( L^{-1} \) easy to calculate

\[
L^{-1} = \begin{pmatrix}
I_n & 0 \\
-S^T P^{-1} & I_p
\end{pmatrix}
\]

- Summary: sample \( \zeta \sim \mathcal{N}(0, \Omega) \), set \( \zeta^* = L^{-1} \zeta \) and collect the last \( p \) entries of \( \zeta^* \), say \( \eta \).
Our proposal

- $P \in \mathbb{R}^{n \times n}$ and $R \in \mathbb{R}^{p \times p}$ are invertible and $S \in \mathbb{R}^{n \times p}$

$$
\begin{pmatrix}
P & S \\
S^T & R
\end{pmatrix}
= 
\begin{pmatrix}
I_n & 0 \\
S^T P^{-1} & I_p
\end{pmatrix}
\begin{pmatrix}
P & 0 \\
0 & R - S^T P^{-1} S
\end{pmatrix}
\begin{pmatrix}
I_n & P^{-1} S \\
0 & I_p
\end{pmatrix}
$$

- Write $\Gamma = L^{-1} \Omega (L^{-1})^T$.
- $L$ lower triangular, so $L^{-1}$ easy to calculate

$$
L^{-1} = 
\begin{pmatrix}
I_n & 0 \\
-S^T P^{-1} & I_p
\end{pmatrix}
$$

- Summary: sample $\zeta \sim \mathcal{N}(0, \Omega)$, set $\zeta^* = L^{-1} \zeta$ and collect the last $p$ entries of $\zeta^*$, say $\eta$. 

---

**[Back to top]**
Recall $P = (XD\Phi^T + I_n)$, $S = XD$ and $R = D$.

$\eta \sim N(0, \Sigma)$

$\eta = u - S^T P^{-1} v$

$\mu = DX^T (XD\Phi^T + I_n)^{-1} Y$

$\mu = S^T P^{-1} Y$.

Finally $\theta = \mu + \eta$

$\theta = u + S^T P^{-1} (Y - v)$
Proposed algorithm

Algorithm Proposed algorithm.

(i) Sample \( u \sim \mathcal{N}(0, D) \) and \( \delta \sim \mathcal{N}(0, I_n) \) independently.
(ii) Set \( v = Xu + \delta \).
(iii) Solve \((XDX^T + I_n)w = (Y - v)\).
(iv) Set \( \theta = u + DX^Tw \).

Overall complexity: \( O(n^2p) \) if \( p > n \).
In \( p \gg n \) settings, reduction from cubic to linear in \( p \).
**Table**: Absolute time (in seconds) to run 6000 iterations of the Gibbs sampler reported for the two algorithms for chosen values of $p$.

<table>
<thead>
<tr>
<th>$p$</th>
<th>Time proposed</th>
<th>Time old</th>
</tr>
</thead>
<tbody>
<tr>
<td>200</td>
<td>5.50</td>
<td>6.05</td>
</tr>
<tr>
<td>500</td>
<td>8.79</td>
<td>31.03</td>
</tr>
<tr>
<td>1000</td>
<td>12.83</td>
<td>160.92</td>
</tr>
<tr>
<td>2000</td>
<td>20.04</td>
<td>944.78</td>
</tr>
<tr>
<td>3000</td>
<td>27.60</td>
<td>2616.80</td>
</tr>
<tr>
<td>4000</td>
<td>35.76</td>
<td>5775.70</td>
</tr>
<tr>
<td>5000</td>
<td>43.99</td>
<td>11314.28</td>
</tr>
</tbody>
</table>

Big gains when $p$ large.
Replicated simulation study with horseshoe prior.

\[ n = 200 \ & p = 5000 \]. True \( \beta_0 \) has 5 non-zero entries.

Two signal strengths:
(i) weak - \( \beta_{0S} = \pm (0.75, 1, 1.25, 1.5, 1.75) \)
(ii) moderate - \( \beta_{0S} = \pm (1.5, 1.75, 2, 2.25, 2.5) \)

Two types of design matrix:
(i) Independent - \( X_j \) i.i.d. \( N(0, I_p) \)
(ii) compound symmetry - \( X_j \) i.i.d. \( N(0, \Sigma) \), \( \Sigma_{jj'} = 0.5 + 0.5 \delta_{jj'} \)

100 data sets generated.

Compare horseshoe with MCP, SCAD.
Simulation Results

Boxplots of $\ell_1$, $\ell_2$ and prediction error across 100 simulation replicates. HSme and HSm respectively denote posterior point wise median and mean for the horeshoe prior. True $\beta_0$ is 5-sparse with non-zero entries $\pm \{1.5, 1.75, 2, 2.25, 2.5\}$. Top row: $\Sigma = I_p$ (independent). Bottom row: $\Sigma_{jj} = 1$, $\Sigma_{jj'} = 0.5, j \neq j'$ (compound symmetry).
Same setting as in Fig 23. True $\beta_0$ is 5-sparse with non-zero entries $\pm \{0.75, 1, 1.25, 1.5, 1.75\}$.

Provide confidence sets for coefficients of subsets of variables.

Bayesian variable selection by post-processing MCMC output with one-group priors (Bondell & Reich, 2012; Hahn & Carvalho, 2015; Li & Pati, 2015)

Used in conjunction with shrinkage priors.
<table>
<thead>
<tr>
<th></th>
<th>p 500</th>
<th></th>
<th>p 1000</th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>Design</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td>Independent</td>
<td>Comp Symm</td>
<td>Independent</td>
</tr>
<tr>
<td></td>
<td></td>
<td>HS LASSO SS</td>
<td>HS LASSO SS</td>
<td>HS LASSO SS</td>
</tr>
<tr>
<td>Signal Coverage</td>
<td>93.0</td>
<td>75.12</td>
<td>82.3</td>
<td>95.0</td>
</tr>
<tr>
<td>Signal Length</td>
<td>42</td>
<td>46</td>
<td>41</td>
<td>85</td>
</tr>
<tr>
<td>Noise Coverage</td>
<td>100.0</td>
<td>99.8</td>
<td>99.10</td>
<td>100.0</td>
</tr>
<tr>
<td>Noise Length</td>
<td>2</td>
<td>43</td>
<td>40</td>
<td>4</td>
</tr>
</tbody>
</table>

Frequentist coverages (%) and $100 \times$ lengths of pointwise 95% intervals. Average coverages and lengths are reported after averaging across all signal variables (rows 1 and 2) and noise variables (rows 3 and 4). Subscripts denote standard errors (%) for coverages. LASSO and SS respectively stand for the methods in van De Geer et. al. (2013) and Javanmard & Montanari (2014). The intervals for the horseshoe (HS) are the symmetric posterior credible intervals.

The data portal has DNA and reverse phase protein arrays (RPPA) expression data with several other clinical variables e.g. survival time of the patients.

Integrative analysis is possible as we have quantitative protein expression data over large cohorts of well characterized TCGA patient tumors, with linked DNA and RNA analyses.
Classification Analysis

- We have integrated the RPPA data with the DNA expression data for our analysis.
- Toward the goal of classification we consider two types of Lung tumors data viz. Lung adenocarcinoma (LUAD) and Lung squamous cell carcinoma (LUSC).
- After some clean up of the raw data we end up with having 179 proteins and 76 genes which are present in both the tumors. The tumors have 55 and 45 samples respectively.
- Our goal is to classify the tumor groups with the help of genes and protein together and simultaneously select the important genes and proteins.
Suppose $y_i \in \{0, 1\}$ and $X_i \in \mathbb{R}^p$ denote the dichotomous tumor class and the associated genes and proteins for the $i^{th}$ subject respectively.

We assume $P[y_i = 1 \mid X_i, \beta] = \Phi(X_i^T \beta)$ where $\Phi(\cdot)$ is the cdf of the standard normal variable.

We adopt the latent variable formulation of the above model following Chib & Greenberg (1998):

$$z_i = X_i^T \beta + \epsilon, \quad \epsilon \sim N(0, 1) \quad (1)$$

$$y_i = \begin{cases} 1 & \text{iff } z_i > 0 \\ 0 & \text{o.w} \end{cases} \quad (2)$$

Given a posterior of $\beta$ predictions can be done using the posterior predictive distribution.
To select the genes and proteins that are important to the tumor classification we use the Horseshoe shrinkage prior from Carvalho et al. (2009) on the coefficients $\beta$.

Specifically, $\pi(\beta_j | \lambda_j, \tau) \sim N(0, \lambda_j^2 \tau^2)$ and $\lambda_j, \tau \sim C^+(0, 1)$ for $j = 1 \ldots p$.

The latent variable formulation enables simple conditional Gibb’s steps for posterior computation.

However, we need to employ some post processing scheme to select important variables due to the continuous nature of the prior.
Let $\bar{\beta}$ denote the posterior mean of $\beta$.

The posterior predictive loss $|| X\bar{\beta} - X\beta ||_{2}^{2}$ can be minimized subject to a $l_1$ constraint to select important variables.

Decision theoretic justification of such a post processing step can be found in Hahn & Carvalho (2015).

We minimize the following objective function to select important variables:

$$
\beta_p = \beta || X\bar{\beta} - X\beta ||_{2}^{2} + \sum_{j=1}^{p} \lambda_j | \beta_j | .
$$

Following Chakraborty et. al. (2016) we use $\lambda_j = 1/\bar{\beta}_j^2$. 

(3)
To compare our results we choose the Lasso penalty (Tibshirani, 1995) with a logistic link function from the R package ncvreg/glmnet.

We report the selected model size and misclassification rate for both methods.

For the Horseshoe probit model the misclassification rate was 5% and selected model size was 5: the selected genes are GSKA-3 and ERBB3, and proteins are PI3KP110ALPHA, MYOSINIIA_pS1943, and ANNEXIN1.

For the LASSO logistic model they were 5% and 32 respectively.

The variables selected by the Horseshoe probit model were also selected by the LASSO logistic model.
The effect of the GSKA-3 gene as one of the Hypoxia-inducible factors resulting in solid cancer was established in Gort et al. (2008).

The gene ERBB3 has been seen to be have direct impact on lung cancer. See Sitharaman & Anderson (2008).

A recent work by Elkabets et al. (2013) studies in detail the effect of the PI3KP110ALPHA gene on breast cancer and lung cancer.

Wong et al. (2012) developed treatments for the suppression of the lung cancer cell tumor markers related to ANNEXIN1 gene.
Survival Analysis for pan-cancer data

- One of the fundamental interest is to establish the relationship of survival of the patients on different proteins
- Data from different kind of tumors
- As sample size may not be very large in each group so Bayesian hierarchical model will be useful to borrow strength
- Furthermore, to deal with high dimensionality, we require either penalized approach (frequentist statistics) or shrinkage based approach (Bayesian statistics)
- We apply Bayesian techniques with f Horseshoe priors
We consider the Kidney tumors: Kidney Chromophobe (KICH) 63 samples, Kidney renal clear cell carcinoma (KIRC) 150 samples, Kidney renal papillary cell carcinoma (KIRP) 124 samples

All the tumors have 189 proteins
Typically, survival outcomes have two variables: time \((t)\) and censored indicator \((\delta)\) which allows the data to represent either censored or not.

- \(i\): Patient, \(j\): Protein, \(k\): Tumor type

We fit an Accelerated Failure rate (AFT) model:

\[
\log(t_{ik}) = \sum_{j=1}^{p} x_{ijk} \beta_{jk} + \epsilon_{ik}
\]

where \(t_{ik}\) is the survival time of \(i\)-th patient who has the \(k\)-th cancer, other symbols have their usual meaning, and \(\epsilon_i\) are iid \(N(0, \sigma^2)\).

For Bayesian MCMC we impute the censored data \(w_i\)

\[
\begin{align*}
    w_{ik} &= \log t_{ik} & \text{if } t_{ik} \text{ is event time} \\
    w_{ik} &> \log t_{ik} & \text{if } t_{ik} \text{ is right censored}
\end{align*}
\]
We can carry out a regular shrinkage analysis as in linear models. Here we adopt global local Horseshoe prior and fit individual regression model for each kind of tumor

$$\log t_{ik} | \beta_{jk}, \sigma^2 \sim N \left( \sum_{j=1}^{p} x_{ijk} \beta_{jk}, \sigma^2_k \right)$$

$$\beta_{jk} | \lambda_{jk}, \tau, \sigma^2 \sim N(0, \lambda_{jk}^2 \tau^2 \sigma^2_k)$$

$$\lambda_{jk} \sim C^+(0, 1)$$

$$\tau \sim C^+(0, 1)$$

$$\pi(\sigma^2_k) \sim \frac{1}{\sigma^2_k}$$
Estimation of Protein Effects
Previous plot shows that due to shrinkage power of Horseshoe prior and due to absence of sufficient number of samples in each Kidney tumor group almost all of the proteins turn out to be insignificant in explaining the survival curve.

So we decide to fit a pan cancer model.

While fitting this model we make use of the idea of borrowing strength across cancers by allowing the prior distributions of the parameters accordingly.
Pan Cancer Model

\[
\log t_{ik} | \beta_{jk}, \sigma^2 \sim N \left( \sum_{j=1}^{p} x_{ijk} \beta_{jk}, \sigma^2 \right) \\
\beta_{jk} | \lambda_{jk}, \tau, \sigma^2 \sim N(b_j, \lambda_{jk}^2 \tau^2 \sigma^2) \\
\lambda_{jk} | \tau, \sigma^2 \sim C^+(0, 1) \\
\tau | \sigma^2 \sim C^+(0, 1) I(0, 1) \\
\pi(\sigma^2) \sim \frac{1}{\sigma^2} \\
b_j \sim N(0, \sigma_b^2), \quad j = 1, \ldots, p
\]

Then,

\[
\text{Corr}(\beta_{jk}, \beta_{jk'}) = \frac{\sigma_b^2}{(\lambda_{jk}^2 \tau^2 \sigma^2 + \sigma_b^2)^{\frac{1}{2}} (\lambda_{jk'}^2 \tau^2 \sigma^2 + \sigma_b^2)^{\frac{1}{2}}}
\]
Estimation

Graphs showing the relationship between beta post corr group and proteins for KICH, KIRC, and KIRP.
Some of the proteins which are significant for KIRC tumor group are BAK, CRAF_pS338, GAB2, HER3_pY1298, PCADHERIN, PCNA, RAD51, FOXO3A_pS318S321, SF2, DIRAS3

The effects of these proteins have been discussed in the literature e.g. see Adams et al (2012) GAB2 - A Scaffolding Protein in Cancer, Molecular Cancer Research

The significant proteins for other tumor groups could be found in the similar fashion