Fully Bayesian Robit Regression with Heavy-tailed Priors for Selection in High-Dimensional Features with Grouping Structure

Longhai Li

Department of Mathematics and Statistics
University of Saskatchewan
Saskatoon, SK, CANADA

Presented at Wuhan University
December 23, 2016

Acknowledgements

- This is a joint work with my Ph.D graduate Lai Jiang and Prof. Weixin Yao at UCR.
- Thanks to the funding support from NSERC and CFI of Canada.
- Thanks to Prof. Yuanshan Wu for hosting my visit to Wuhan University.

Outline

Section 1

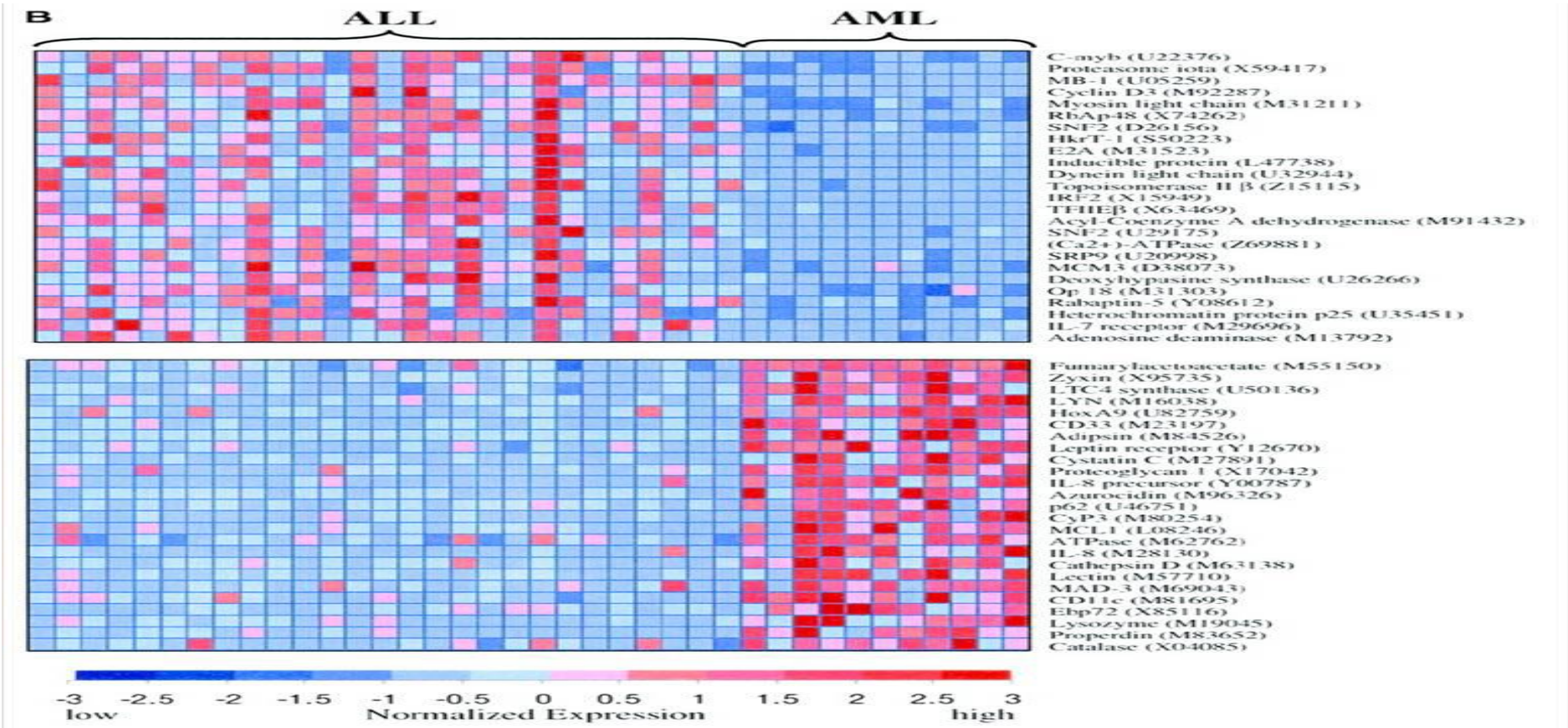
Introduction

Subsection 1

Research Problem

A Genomic Example

High-throughput biotechnologies can measure the expression levels of all genes. We are interested in discovering the genes related to a complex disease.



Formulation of the Statistical Problem

- Response: y is a categorical variable indicating types of disease.
- Features (variables): $x = \{x_j | j = 1, \ldots, p)$ contains measurements of p features, such as gene expressions.
- The goal is to identify a subset of features that are related to the response y.

Challenges of Selecting from High-dimensional Features

- High dimensionality
  The number of features, p, is greatly larger than n. We face severe overfitting problems. In other words, it is difficult to identify the signal from a vast amount of noise.

- Grouping structure
  High-throughput data have grouping structures. For example, a group of genes may have similar expression levels. We face severe collinearity problem.

- An anaogy to understand the chanlleges: *looking for needles from hay*.

Subsection 2

Literature Review

Univariate Screening Methods (GWAS)

Examples:

- t-test, F-test, Significance Analysis of Microarray (SAM).
- Diagonal Linear Discriminant Analysis (DLDA).
- Prediction Analysis of Microarrays method (PAM).

Problems:

- Generally speaking, univariate screening methods ignore the join effects between genes. We believe that many complex diseases or traits are *polygenetic*.
- Many marginally correlated features come as the top list.
- Features with weak marginal correlation will be omitted. They may be very useful!

Penalized Likelihood Methods

- **LASSO**: Least Absolute Shrinkage and Selection Operator (LASSO) minimizes a likelihood function of β penalized by L_1 norm (absolute shrinkage) to enforce sparse solutions:

$$\hat{\beta} = \text{argmin}_{\beta} \left\{ -\sum_{i=1}^{n} \log(P(y_i|\boldsymbol{x}_i, \beta)) + \lambda \sum_{j=1}^{p} |\beta_j| \right\} \tag{1}$$

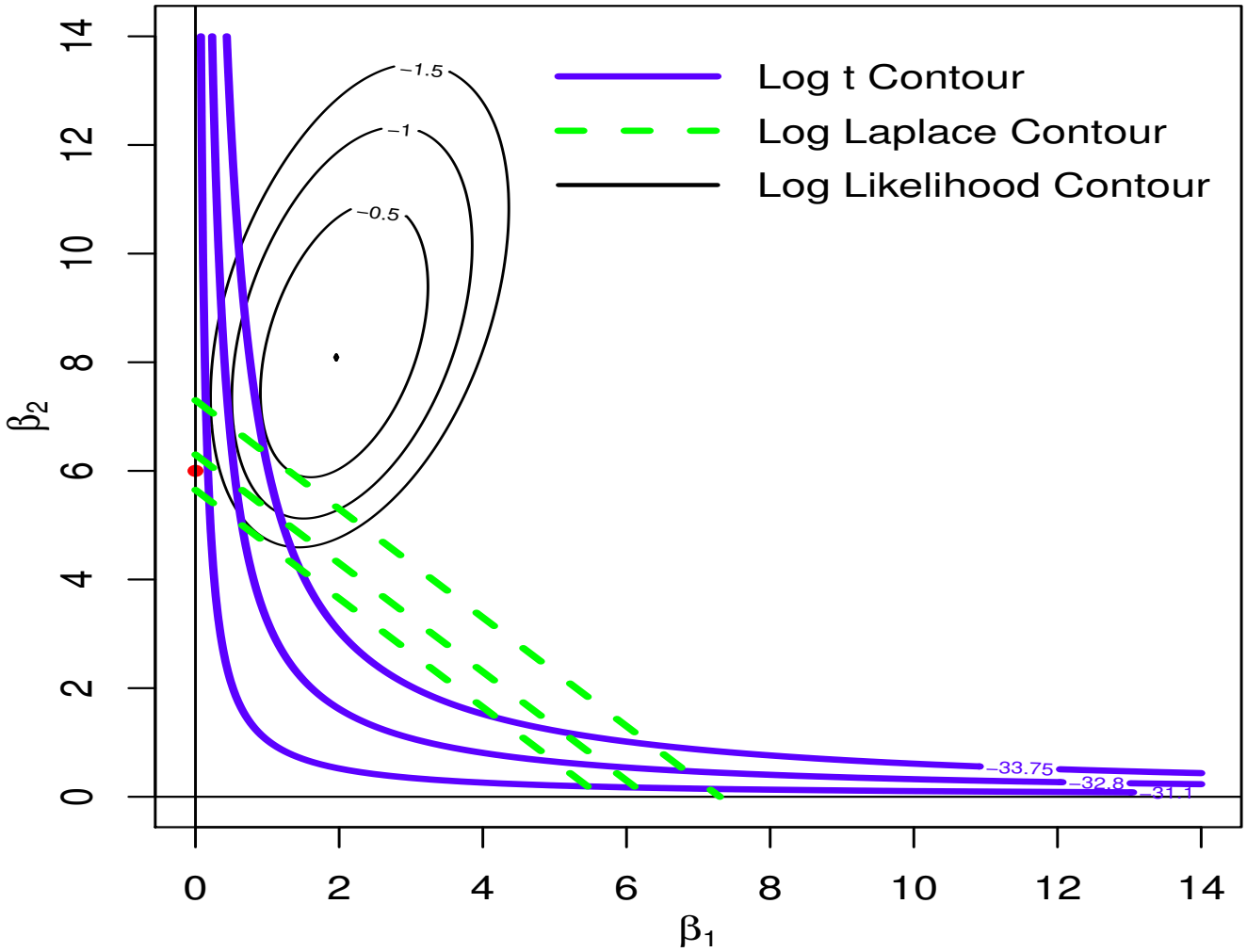
- LASSO solution is not sparse enough when $p >> n$.
- **Non-convex penalization**

$$\hat{\beta} = \text{argmin}_{\beta} \left\{ -\sum_{i=1}^{n} \log(P(y_i|\boldsymbol{x}_i, \beta)) + \sum_{j=1}^{p} t_{\alpha}(\beta_j) \right\} \tag{2}$$

where $t_{\alpha}(\beta)$ is a non-convex penalty function, for example, SCAD, log t density, horseshoe, NEG priors, and MCP.

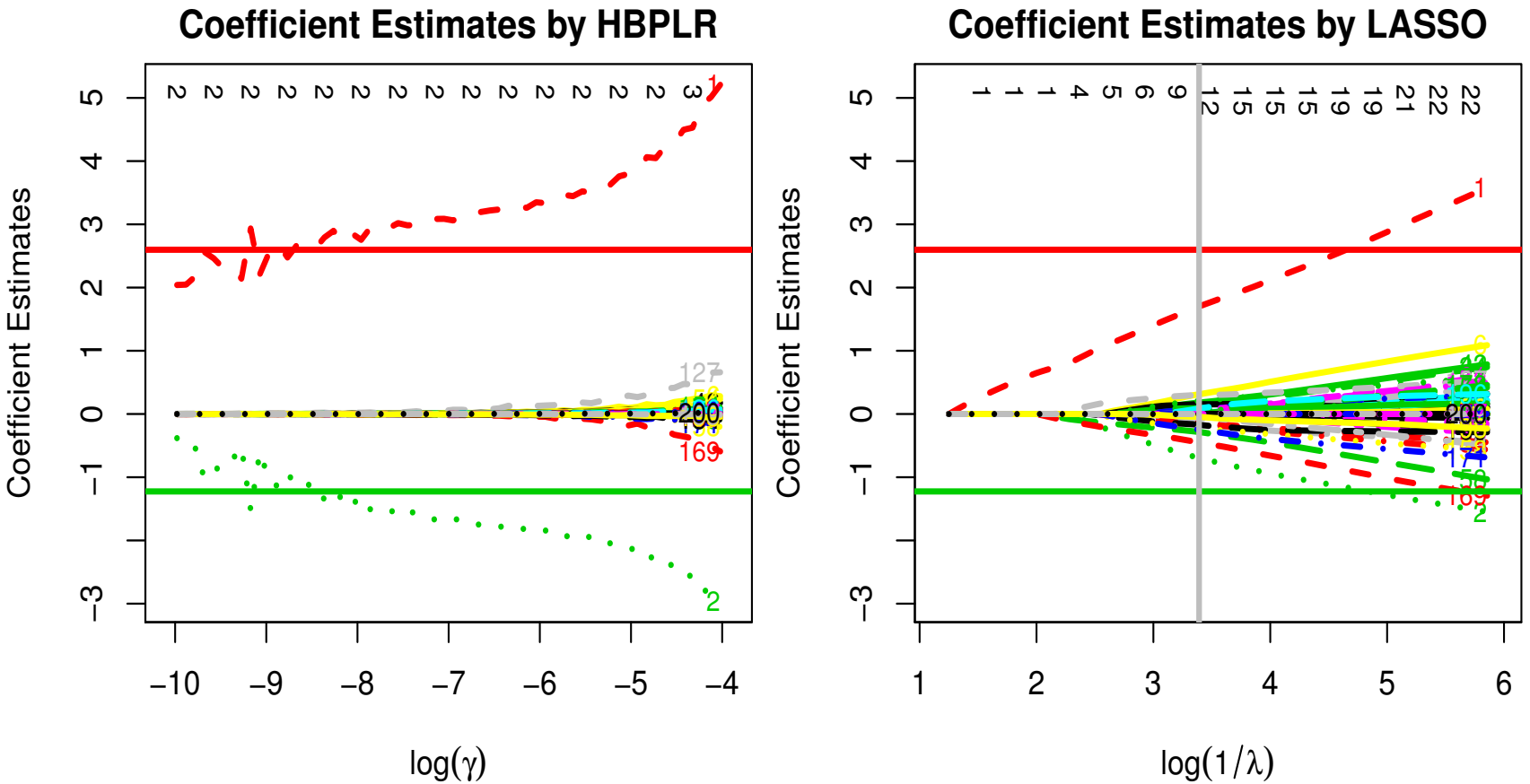
Hyper-LASSO Sparsity of Non-convex Penalty (I)

Compared to LASSO, non-convex penalty can more aggressively shrink small coefficients towards 0, while retaining large coefficients.



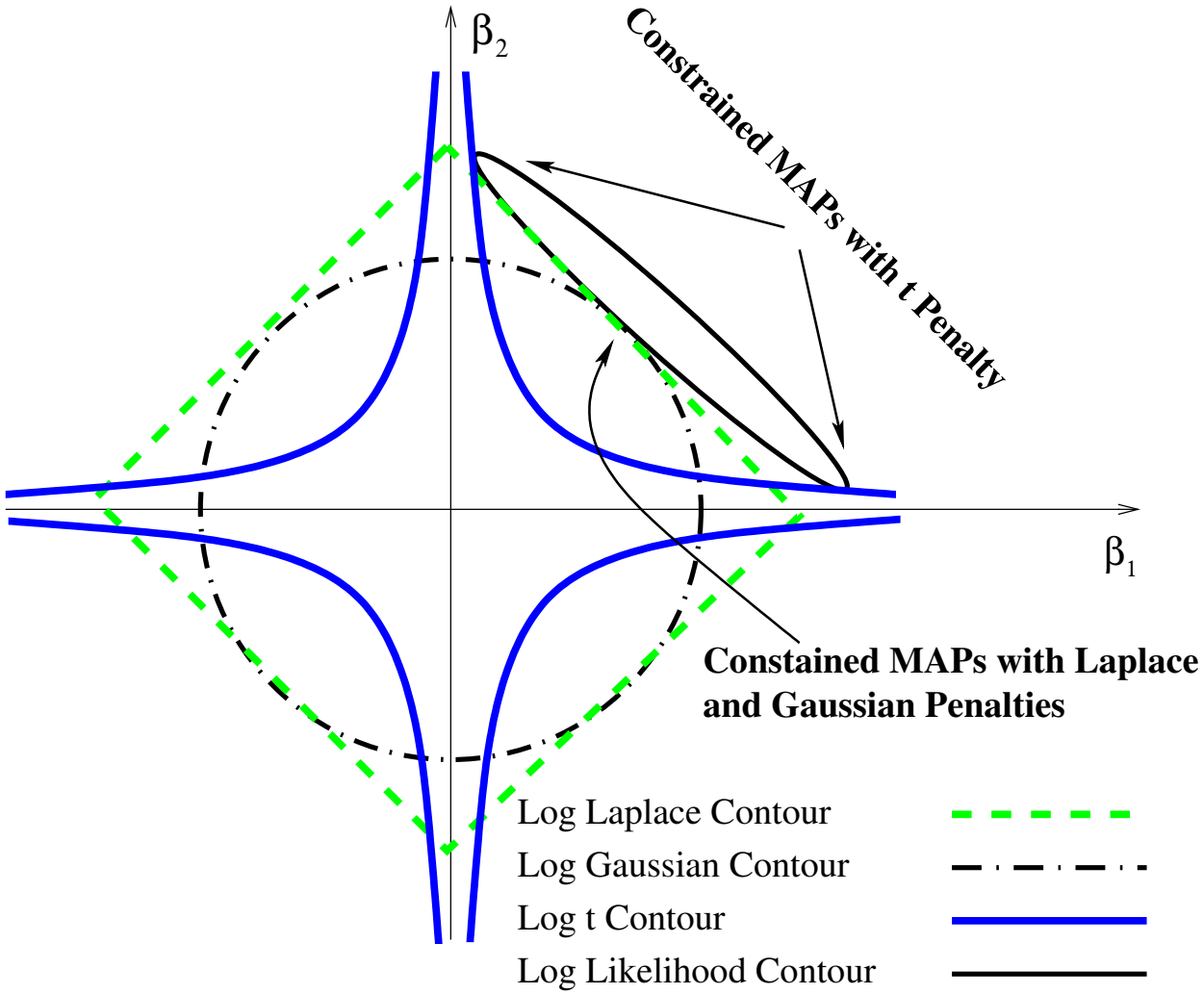
Hyper-LASSO Sparsity of Non-convex Penalty (II)

Comparing the solution path of LASSO and non-convex penalization:



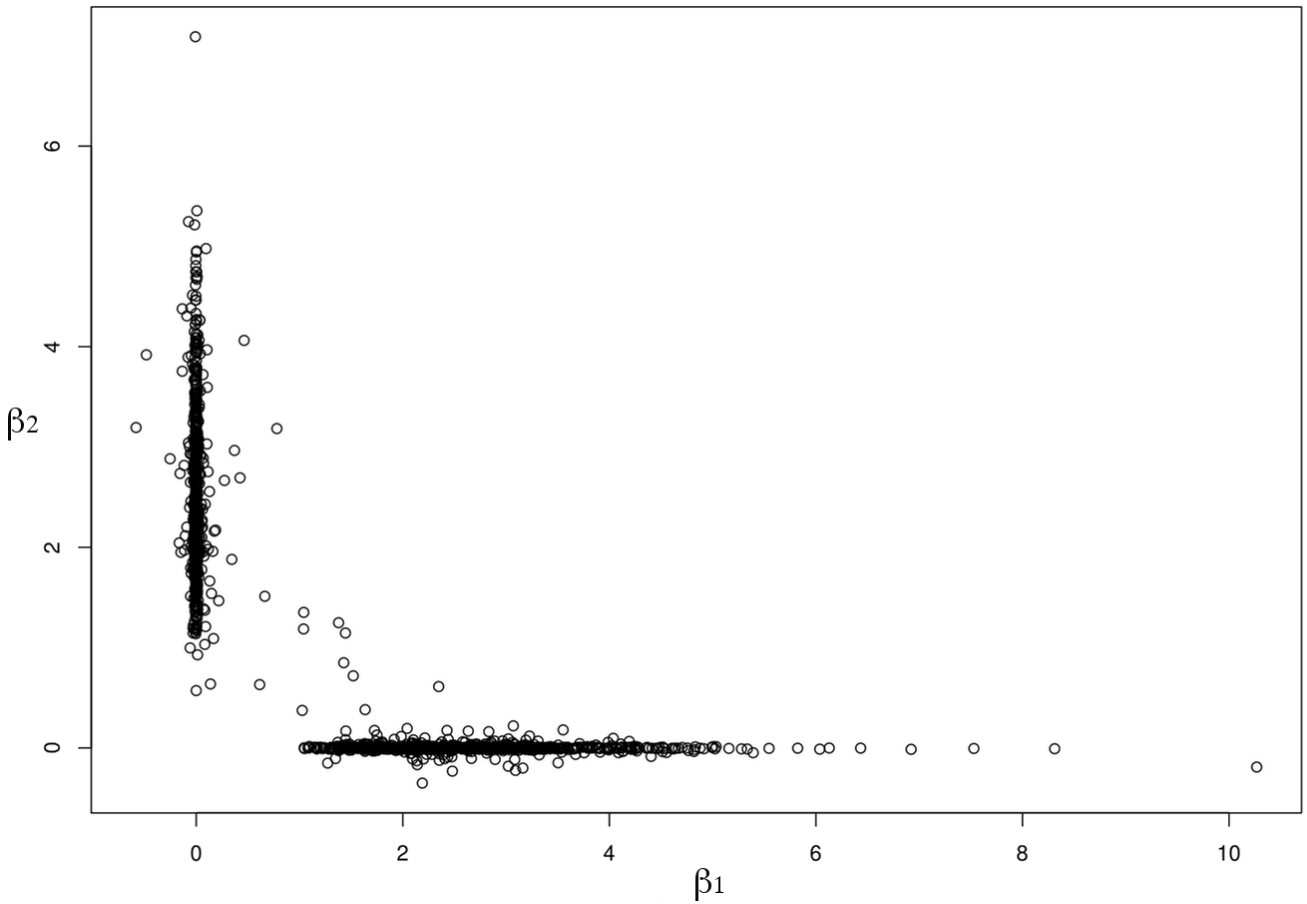
Separation of Correlated Features (I)

Non-convex penalty can separate the coefficients of correlated features into different modes.



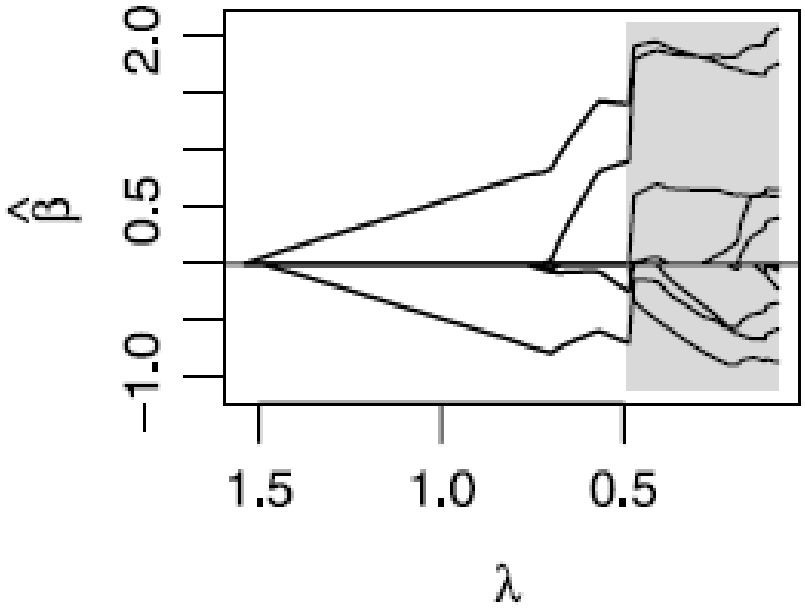
Separation of Correlated Features (II)

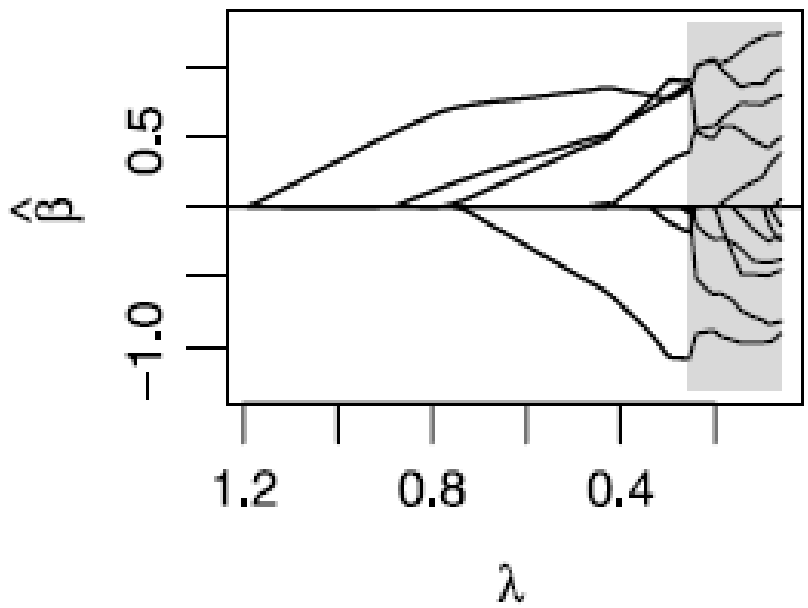
Sample representation of a non-convex penalized likelihood of two coefficients for two highly correlated features:



Difficulty of Optimizing Non-complex Penalized Likelihood

Unfortunately, non-convex penalized likelihood has many modes. Although good theoretical properties of the global mode of non-convex penalized likelihood have been established, practitioners have been reluctant to embrace these methods for good reason: non-convex penalized likelihoods are difficult to optimize and often produce unstable solutions (Breheny and Huang, 2011, *AOAS*).





Other Methods Using the Grouping Structure in Features

- **Group LASSO** (GL): Group LASSO uses a penalty function to enforce similarity of coefficients within a group:

$$LL(\beta|x_i, y_i) = -\sum_{i=1}^{n} \log(P(y_i|\boldsymbol{x}_i, \boldsymbol{\beta})) + \lambda \sum_{l=1}^{L} \sqrt{p_l}||\beta_l||_2, \qquad (3)$$

- Group LASSO achieves better prediction accuracy, but makes selection within group harder
- **Two-stage selection**: For example, supervised group LASSO (SGL) applies LASSO to each feature group first then fit LASSO with all the features selected from each group.
- SGL cannot consider joint effect of features across group.
- Both GL and two-stage selection require a pre-defined grouping structure.

Our Propoal: Explore Multi-mode Posterior with MCMC

- Markov chain Monte Carlo sampling (MCMC) can travel across many modes of the non-convex penalized likelihood. We propose to use a sophisticated HMC based method to explore the posterior of Robit model assigned with a class of heavy-tailed priors—t distribution with moderate degree freedom (such as 1, corresponding to Cauchy distribution) and small scale.

- We then divide MCMC samples from many modes to find multiple feature subsets.

- We will refer to our method as fully Bayesian Robit with heavy-tailed priors, or **FBRHT** to be short.

Section 2

Fully Bayesian Robit Regression with Heavy-tailed Priors

Robit Models with Heavy-tailed Priors

- The **heavy-tailed Robit model** can be written as:

$$
\begin{aligned}
P(y_i|x_i,\beta) &= T_{\alpha_0,\omega_0}(x_i\beta)^{y_i}(1-T_{\alpha_0,\omega_0}(x_i\beta))^{1-y_i}, && (4)\\
\beta_j|\lambda_j &\sim N(0,\lambda_j), && (5)\\
\lambda_j &\overset{\text{iid}}{\sim} \text{Inverse-Gamma}(\alpha_1/2,\alpha_1\omega_1/2), && (6)
\end{aligned}
$$

  where $T_{\alpha_0,\omega_0}(\cdot)$ is the CDF of t-distribution with degree freedom α_0 and scale parameter $\sqrt{\omega_0}$
- Integrating λ_j way in (5) and (6), we have

$$\beta_j \sim T(\alpha_1,\omega_1)$$

- We fix $\alpha_1 = 1, \omega_1 = \exp(-10)$ after a seires of simulation studies.
- We fix $\alpha_0 = 1, \omega_0 = 0.5$ so that $T_{\alpha_0,\omega_0}(\cdot)$ is similar to logistic regression but is more robust to outliers.

Restricted Gibbs Sampling with HMC

Given a previous state for $(\boldsymbol{\beta}, \lambda)$, we iteratively obtain a new state denoted by $(\hat{\boldsymbol{\beta}}, \hat{\lambda})$ as follows:

Step 1: For each j, draw a new $\hat{\lambda}_j$ from the conditional distribution $f(\lambda_j|\boldsymbol{\beta}_j)$, that is,

$$\hat{\lambda}_j \sim \text{Inverse-Gamma}\left(\frac{\alpha_1 + 1}{2}, \frac{\alpha_1\omega_1 + \beta_j^2}{2}\right).$$

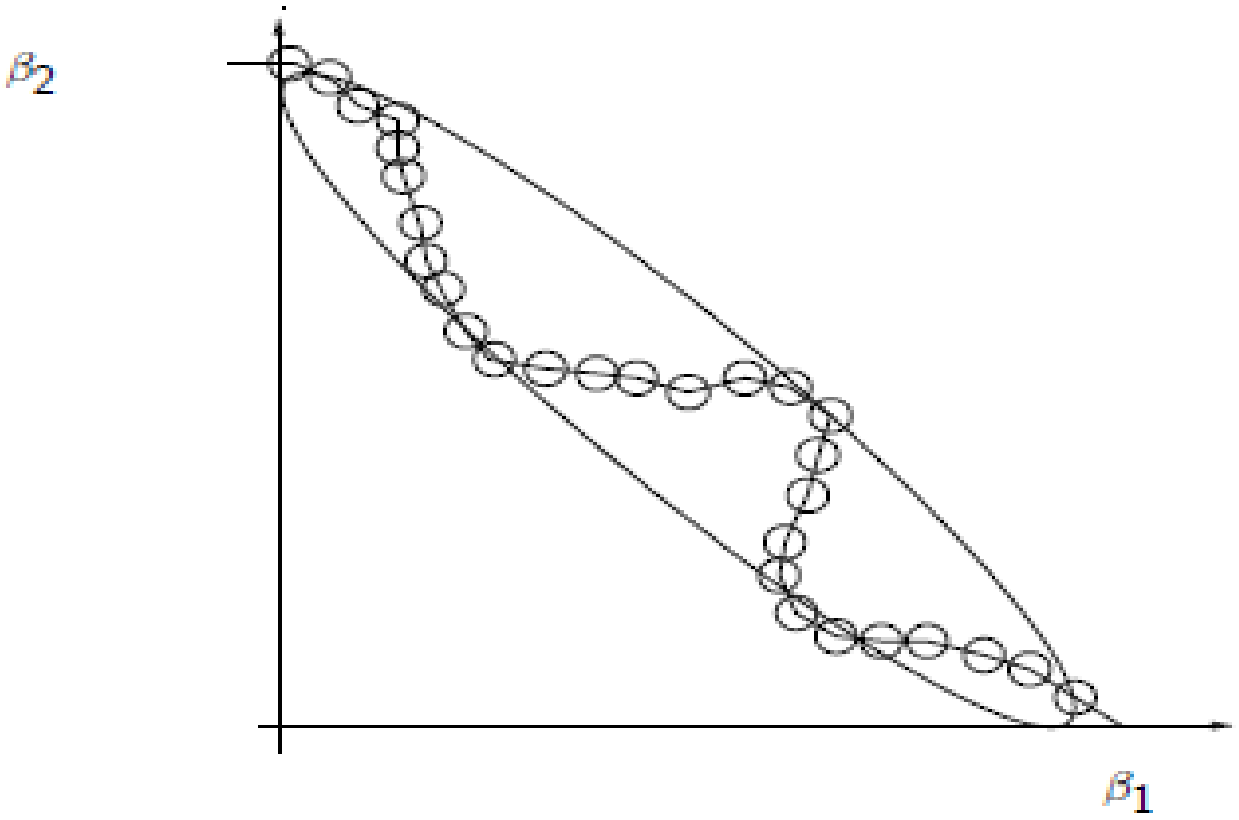
Step 2: Determine a subset of features to update in next step:

$$U = \{j|\hat{\lambda}_j > \eta\}$$

Step 3: Update $\boldsymbol{\beta}_U = \{\beta_j|j \in U\}$ by applying **Hamiltonian Monte Carlo (HMC)** to the conditional distribution of $\boldsymbol{\beta}_U$. Other β_j are unchanged.

Exploration of Many Modes with HMC

The major advantage of HMC is that it can travel efficiently according to the least constrained direction when there are strong correlations between features. As a result HMC is helpful to move from one posterior mode to another one. A schematic demonstration of this property is shown below:



Two Stage Sampling

For interpretation purpose, we implement two-stage sampling for robit models when the number of features p is large (such as thousands):

Stage 1: We first apply the sampling scheme to the dataset with all p features.

Stage 2: Choose only the top 100 features with the largest mean values in MCMC samples. We apply the sampling scheme to the dataset with only the 100 features selected in stage 1.

Divide MCMC Samples to Find Feature Subsets

Our scheme is described as follows:

Step 1: Truncating small coefficients in each MCMC sample indexed by i We set $I_{j,i} = 1$ if $|\beta_{j,i}| > 0.1 \times \max\{|\beta_{1,i}|, \ldots, |\beta_{p,i}|\}$, and $I_{j,i} = 0$ otherwise.

Step 2: Further discard the features with overall low frequency in step 1. We calculate $f_j = \frac{1}{R} \sum_{i=1}^{R} I_{j,i}$. We will discard a feature j if f_j is smaller than a pre-defined threshold, such as 5%.

Step 3: Find a list of feature subset by search unique columns in I. Each unique column in I represents a different feature subset.

Feature Subset Evaluation Using Cross-validation

- Obtain Predictive Probabilities:
  Given a selected feature subset S we apply LOOCV (Leave-one-out Cross Validation) to dataset $(Y, X_{1:n,S})$ to obtain the predictive probabilities $\hat{p}_i(y_i)$ using penalized logistic regression (bayesglm).
- Assessment Criteria:
  - ER (error rate):

$$\text{ER} = \frac{1}{n}\sum_{i=1}^{n} I(\hat{y}_i \neq y_i), \tag{7}$$

  - Average minus log-probability on observed y_i (information criterion):

$$\text{AMLP} = -\frac{1}{n}\sum_{i=1}^{n} \log(\hat{p}_i(y_i)). \tag{8}$$

  - AUC value: area under the curve of ROC.

Section 3

Simulation Studies

Subsection 1

A Toy Example of Correlated Features

Data Generating Scheme

The mean of feature x_j in class c is denoted by μ_j^c, for $c = 0, 1$. We fix

$$\mu_j^0 = 0, \mu_j^1 = 2 \text{ for } j = 1, 2.$$

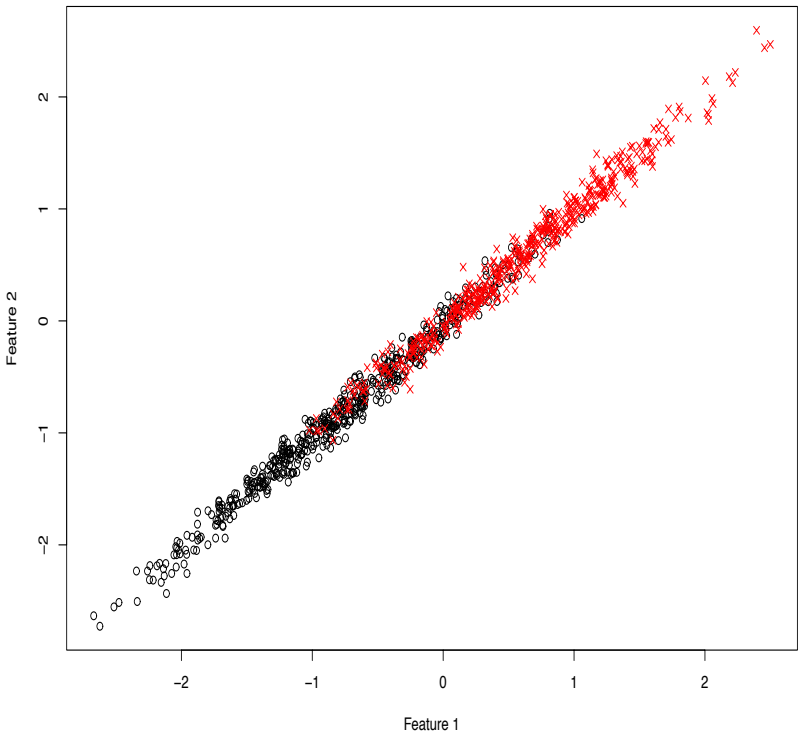
The response y_i and feature values $\boldsymbol{x}_i = (x_{i1}, x_{i2})$ for each case i are generated as follows:

$$P(y_i = c) = 1/2, \text{ for } c = 0, 1, \tag{9}$$

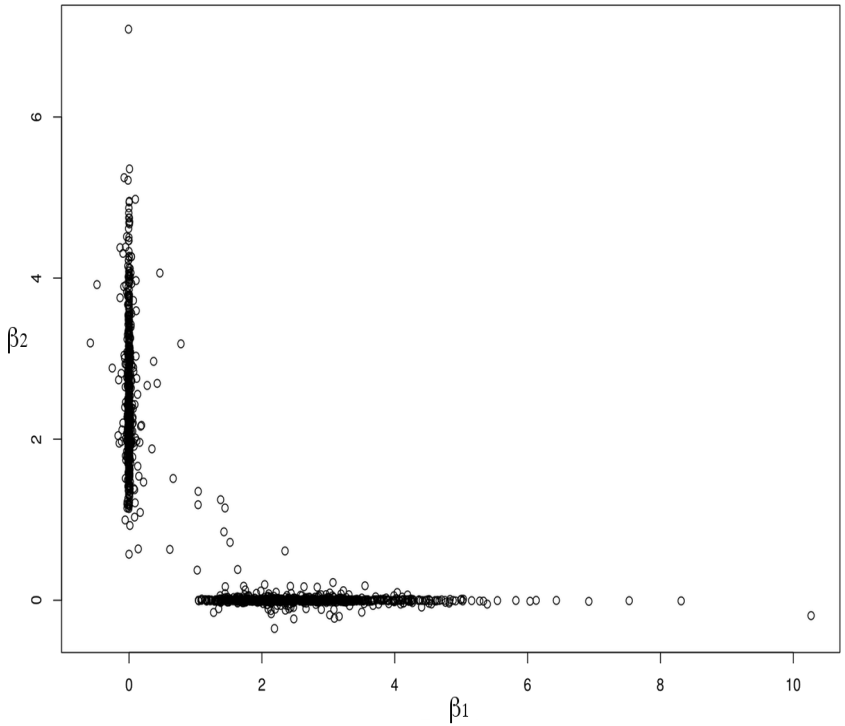
$$z_i \sim N(0, 1), \ \epsilon_{ij} \sim N(0, 1^2), \text{ for } j = 1, 2 \tag{10}$$

$$x_{ij} = \mu_j^{y_i} + z_i + 0.1\epsilon_{ij}, \text{ for } j = 1, 2. \tag{11}$$

Demonstration of within-group selection



(a) Scatterplot of x_1 and x_2



(b) MCMC samples of (β_1, β_2)

Feature Subset Selection and Prediction Results

(a) Feature subsets selected by FBRHT

fsubsets	freqs	coefs	AMLP	ER	AUC
1	0.56	2.62	0.37	0.185	0.91
2	0.42	2.58	0.37	0.180	0.91
1,2	0.02	0.67, 1.94	0.37	0.178	0.91

(b) LASSO, PLR (`bayesglm`) and Random Forest (RF)

Method	coefs	AMLP	ER	AUC
LASSO	1.15, 1.27	0.37	0.184	0.91
RF	1.26, 1.26	Inf	0.219	0.88
PLR	24.63, 24.53	0.37	0.184	0.91

Subsection 2

Simulation Studies Using Datasets with Group Structure

Data Generating Scheme

Each dataset has $p = 2000$ features and $n = 1200$ cases, 200 of which are used as training cases and the other 1000 cases are used as test cases. With $z_{ij}, \epsilon_{ij}, e_i$ generated from $N(0,1)$, we generate the feature values x_{ij} for $i = 1, ..., n, j = 1, ..., p$ in four groups and the class label y_i as follows:

$$x_{il} = z_{i1} + 0.5\epsilon_{il}, i = 1, ..., n, l = 1, ..., 50, \textbf{(Group 1)} \quad (12)$$

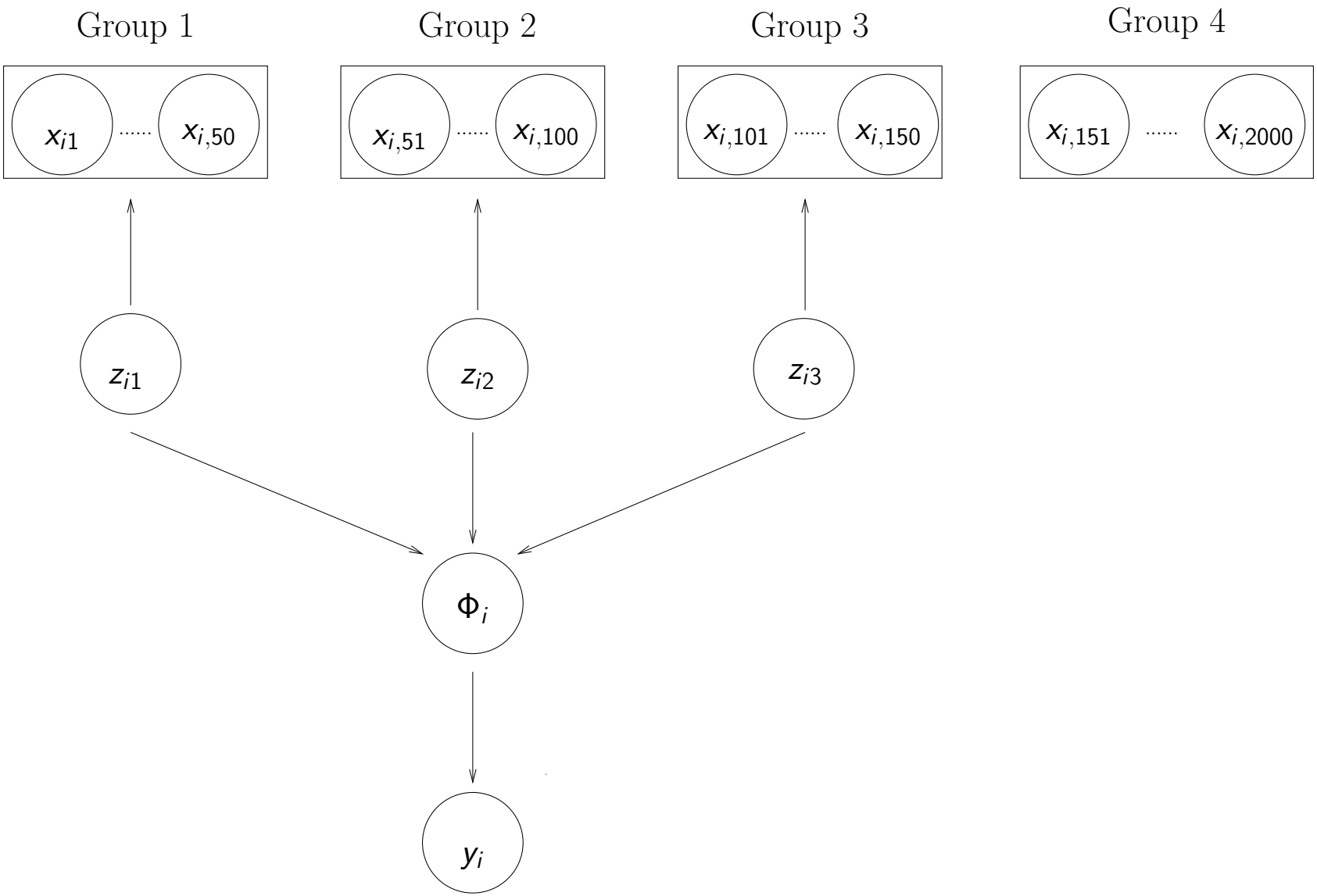
$$x_{im} = z_{i2} + 0.5\epsilon_{im}, i = 1, ..., n, m = 51, ..., 100, \textbf{(Group 2)} \quad (13)$$

$$x_{ik} = z_{i3} + 0.5\epsilon_{ik}, i = 1, ..., n, k = 101, ..., 150, \textbf{(Group 3)} \quad (14)$$

$$x_{ij} \sim N(0,1), i = 1, ..., n, j = 151, ..., 2000, \textbf{(Group 4)} \quad (15)$$

$$y_i = 1 \text{ if } (z_{i1} + z_{i2} + z_{i3})/\sqrt{3} + 0.1e_i > 0; = 0 \text{ otherwise.} \quad (16)$$

A Graphical Representation (Structural Equations)



Feature Subsets Selected by FBRHT

Table 1: "fsubsets" gives I.D. of features in each subset, "cvAMLP" - "cvAUC" are cross-validatory predictive power measures of each feature subset.

	fsubsets	freqs	cvAMLP	cvER	cvAUC
1	1,57,140	0.22	0.13	0.09	0.99
2	1,51,140	0.11	0.13	0.08	0.99
3	16,57,140	0.10	0.14	0.08	0.99
4	1,51,101	0.09	0.14	0.08	0.99
5	12,57	0.04	0.41	0.39	0.89
⋮					

Comparison of Top 3 Features

Table 2: Comparison of out-of-sample predictive power of different subsets containing 3 features on a dataset with independent groups of features. The predictive measures are obtained by applying `bayesglm` to make predictions for the test cases.

Method	fsubsets	AMLP	ER	AUC
FBRHTtop	1,57,140	0.22	0.10	0.97
FBRHTopt	1,57,140	0.22	0.10	0.97
LASSO	16,57,61	0.46	0.22	0.87
GL	16,32,57	0.44	0.20	0.88
SGL	16,138,140	0.47	0.24	0.86
RF	28,50,67	0.46	0.22	0.86
PLR	12,32,218	0.63	0.34	0.72

Comparing with Complete Subsets of Other Methods

Table 3: Comparison of feature selection and out-of-sample prediction performance of different methods. The number of features are counted by thresholding the absolute coefficients by 0.1 times the maximum.

(a) Numbers of selected features in respective group

	FBRHTtop	FBRHTopt	FBRHTavg	LASSO	GL	SGL	RF	PLR
Group 1	1	1	-	6	49	7	49	50
Group 2	1	1	-	5	50	10	49	50
Group 3	1	1	-	6	50	6	48	50
Group 4	0	0	-	13	341	12	14	1305
Total	3	3	$\leq$100	30	490	35	160	1455

(b) Out-of-sample predictive performance

ER	0.10	0.10	0.06	0.09	0.07	0.10	0.08	0.08
AMLP	0.22	0.22	0.15	0.21	0.22	0.24	0.38	0.18
AUC	0.97	0.97	0.99	0.97	0.99	0.97	0.98	0.98

Replicated Results on 100 Datasets

Table 4: Comparison of feature selection and out-of-sample prediction performance of different methods by averaging over 100 datasets with independent group of features.

(a) Numbers of selected features in respective group

	FBRHTtop	FBRHTopt	FBRHTavg	LASSO	GL	SGL	RF	PLR
Group 1	1.00	1.07	-	6.01	49.95	6.20	47.82	50.00
Group 2	1.00	1.08	-	6.00	49.94	5.99	47.48	50.00
Group 3	1.00	1.06	-	5.94	49.95	6.04	48.34	50.00
Group 4	0.00	0.19	-	14.44	401.33	8.83	3.78	1297.68
Total	3.00	3.40	$\leq$100	32.39	551.17	27.06	147.42	1447.68

(b) Out-of-sample predictive performance

ER	0.05	0.06	0.04	0.08	0.06	0.08	0.10	0.08
AMLP	0.15	0.16	0.12	0.21	0.20	0.20	0.39	0.17
AUC	0.99	0.99	0.99	0.98	0.99	0.98	0.97	0.98

Section 4

Real Data Analysis

Subsection 1

Analysis of Breast Cancer Methylation Data

Breast Cancer Methylation Data

- The DNA methylation level of each sample was measured with Human Methylation27 DNA Analysis BeadChip (GPL8490), which includes 27,578 probes. log2 transformation was applied to the original ratio of the methylation level.
- Response of interest
  - 48 samples with moderately estrogen receptor-positive (ER+).
  - 53 samples with receptor-negative (ER-).
  - There are effective treatment methods (e.g. tamoxifen) for ER+ patients since cancer cells with ER+ depends on estrogen to grow.
- We select top p=5000 genes (out of 27,578) with SAM for classification model based analysis.
- We are interested in finding DNA locations whose methylation levels can predict ER+ and ER-.

Comparison of Subsets of Top Features

Table 5: LOOCV predictive measures of feature subsets found from Breast Cancer Data.

(a) Feature subsets given by FBRHT

	fsubsets	freqs	cvAMLP	cvER	cvAUC
1	23,77	0.05	0.21	9/101	0.98
2	77,554	0.03	0.25	11/101	0.96
3	1,366,1795	0.02	0.11	4/101	0.99
4	23,77,1587	0.02	0.16	6/101	0.99
5	1,1526	0.02	0.23	12/101	0.96

(b) Feature subsets given by other methods

Method	fsubsets	cvAMLP	cvER	cvAUC
LASSO	25,266,614	0.27	10/101	0.95
GL	2256,1795,266	0.52	21/101	0.82
SGL	266,2256,1756	0.51	25/101	0.83
RF	10,8,103	0.32	13/101	0.93
PLR	1,2256,4832	0.27	12/101	0.95

Comparing with Complete Subsets of Other Methods

Table 6: Comparison of out-of-sample predictive performance on Breast Cancer Data.

	FBRHTopt	FBRHTtop	FBRHTavg	LASSO	GL	SGL	RF	PLR
No. of Genes	2.98	2.02	≤ 100	39.57	2209.73	36.62	187.63	2667.47
ER$\times$101	9	21	10	8	9	10	10	12
AMLP	0.33	0.51	0.33	0.28	0.27	0.42	0.34	0.33
AUC	0.96	0.88	0.91	0.94	0.94	0.95	0.93	0.94

Subsection 2

Analysis of Leukemia Microarray Data

Leukemia Microarray Data

- Gene expression data (with Swegene Human DNA microarray platform) on leukemia patients was collected at Lund University Hospital and Linkoping University Hospital (Andersson et. al., 2007). We downloaded the dataset from GEO (GSE7186).
- Response of interest
  - 98 samples with acute lymphoblastic leukemia (ALL)
  - 23 samples with acute myeloid leukemia (AML)
- We preprocessed the data using SAM and kept $p = 5000$ genes for the following analysis.

Comparison of Top Features

Table 7: LOOCV predictive measures of feature subsets found from Acute Leukaemia Data.

(a) Feature subsets given by FBRHT

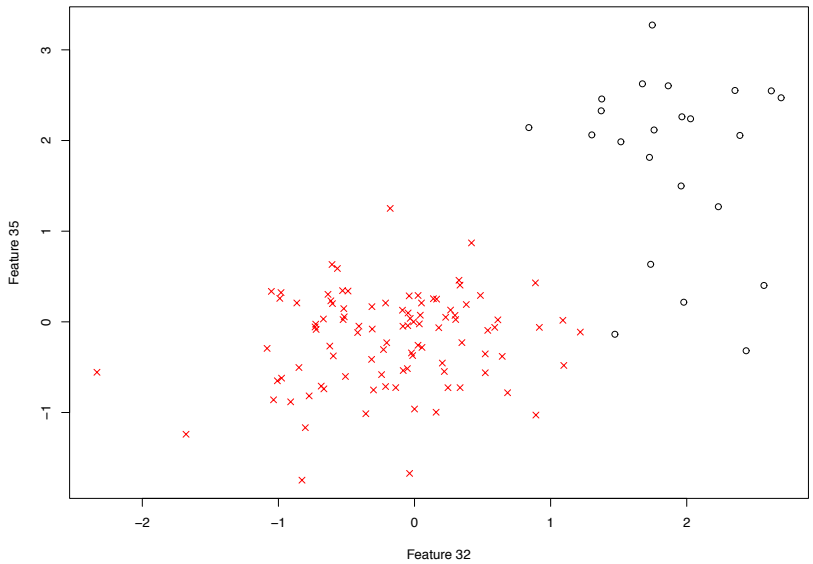
	fsubsets	freqs	cvAMLP	cvER	cvAUC
1	32	0.38	0.06	2/121	1.00
2	30	0.18	0.07	4/121	0.99
3	36	0.09	0.09	2/121	0.99
7	30,35	0.02	0.03	1/121	1.00
8	32,35	0.02	0.03	1/121	1.00

(b) Feature subsets given by other methods

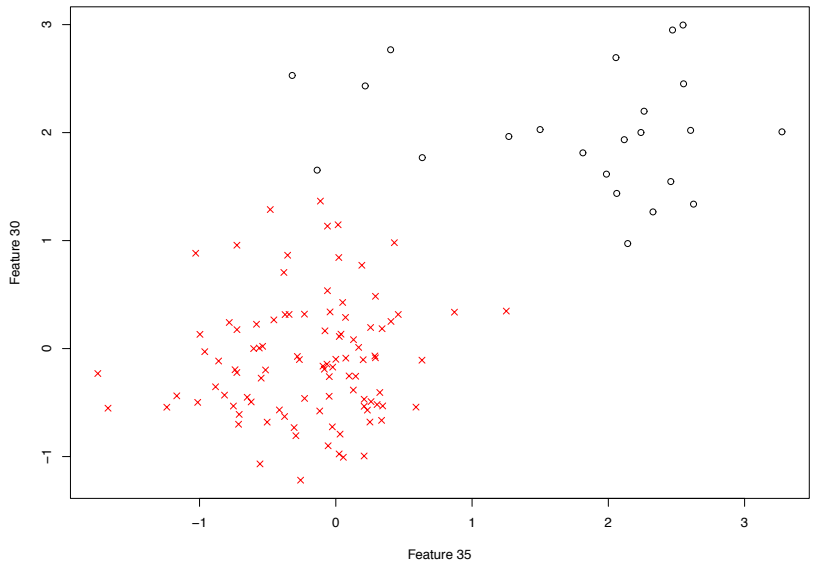
Method	fsubsets	cvAMLP	cvER	cvAUC
LASSO	32,35	0.03	1/121	1.00
GL	35,115	0.15	4/121	0.95
SGL	115,35	0.13	4/121	0.96
RF	36,28	0.07	4/121	1.00
PLR	1,5794	0.20	12/121	0.96

Scatterplots of Top Features

Figure 1: Scatterplots of two feature subsets found from the leukamia data.



(a) Gene subset (32, 35).



(b) Gene subset (30, 35).

Comparison with Complete Subsets of Other Methods

Table 8: Comparison of out-of-sample predictive performance on Breast Cancer Data.

	FBRHTtop	FBRHTopt	FBRHTavg	LASSO	GL	SGL	RF	PLR
No. of Genes	1.00	1.95	≤ 100	26.43	2783.26	50.34	149.33	3484.88
ER$\times$121	3	5	2	1	0	2	2	10
AMLP	0.07	0.09	0.09	0.04	0.05	0.03	0.12	0.34
AUC	1.00	0.99	1.00	1.00	1.00	1.00	1.00	1.00

Section 5

Conlusions and Discussions

Conclusions

- FBRHT makes selection within groups automatically *without* a pre-specified grouping structure. Meanwhile, the joint effects of features from different groups can also be considered.
- FBRHT finds succinct feature subsets, which are much easier to interpret or comprehend based on existing biological knowledge, and easier for further experimental investigations.
- The succinct feature subsets found by FBRHT have comparable predictive power as other much larger feature subsets found by other methods.
- The multiple feature subsets found by FBRHT provide multiple explanations of the associations for scientists to further explore.

Future Work

- Applications to real high-throughput data analysis. I am involved in a huge CFREF project which aims to "design" crops for helping feed the world through transformative innovations in agriculture and food production. Many high-throughput data arise in this project.
- Improve the method for intepreting MCMC samples. A very interesting method is the "reference" approach.
- Extensions of fully Bayesian methods with heavy-tailed priors to other models.
- Fitting structural equaiton modelling to high-throughput data.