Bayesian Methods for Variable Selection with Applications to High-Dimensional Data

Part 1: Mixture Priors for Linear Settings

Marina Vannucci

Rice University, USA

ABS13-Italy
06/17-21/2013

Part 1: Mixture Priors for Linear Settings

- Linear regression models (univariate and multivariate responses)
- Extensions to categorical responses and survival outcomes
- Matlab code
- Examples from genomics/proteomics
- Bayesian models for integrative genomics (next part)

Regression Model

$$\mathbf{Y}_{n\times 1} = \mathbf{1}\alpha + \mathbf{X}_{n\times p}\boldsymbol{\beta}_{p\times 1} + \varepsilon, \quad \varepsilon \sim N(0, \sigma^2\mathbf{I})$$

Introduce latent variable $\boldsymbol{\gamma} = (\gamma_1, \ldots, \gamma_p)'$ to select variables

$$\begin{cases} \gamma_j = 1 & \text{if variable } j \text{ included in model} \\ \gamma_j = 0 & \text{otherwise} \end{cases}$$

Specify priors for model parameters:

$$\begin{aligned} \beta_j|\sigma^2 &\sim (1-\gamma_j)\delta_0(\beta_j) + \gamma_j N(0, \sigma^2 h_j) \\ \alpha|\sigma^2 &\sim N(\alpha_0, h_0\sigma^2) \\ \sigma^2 &\sim IG(\nu/2, \lambda/2) \\ p(\boldsymbol{\gamma}) &= \prod_{j=1}^{p} w^{\gamma_j}(1-w)^{1-\gamma_j}. \end{aligned}$$

where $\delta_0(\cdot)$ is the Dirac function.

Posterior Distribution

Combine data and prior information into a posterior distribution $\Rightarrow$ interest in posterior distribution

$$p(\boldsymbol{\gamma}|\mathbf{Y},\mathbf{X}) \propto p(\boldsymbol{\gamma}) \int f(\mathbf{Y}|\mathbf{X},\alpha,\boldsymbol{\beta},\boldsymbol{\sigma}) p(\alpha|\sigma) p(\boldsymbol{\beta}|\sigma,\boldsymbol{\gamma}) p(\sigma) d\alpha d\boldsymbol{\beta} d\sigma$$

$$p(\boldsymbol{\gamma}|\mathbf{Y},\mathbf{X}) \propto g(\boldsymbol{\gamma}) =$$

$$|\tilde{\mathbf{X}}'_{(\boldsymbol{\gamma})}\tilde{\mathbf{X}}_{(\boldsymbol{\gamma})}|^{-1/2}(\nu\lambda + \mathbf{S}^2_{\boldsymbol{\gamma}})^{-(n+\nu)/2} p(\boldsymbol{\gamma})$$

$$\tilde{\mathbf{X}}_{(\boldsymbol{\gamma})} = \begin{pmatrix} \mathbf{X}_{(\boldsymbol{\gamma})}\mathbf{H}^{\frac{1}{2}}_{(\boldsymbol{\gamma})} \\ I_{p_{\boldsymbol{\gamma}}} \end{pmatrix}, \quad \tilde{\mathbf{Y}} = \begin{pmatrix} \mathbf{Y} \\ 0 \end{pmatrix}$$

$$\mathbf{S}^2_{\boldsymbol{\gamma}} = \tilde{\mathbf{Y}}'\tilde{\mathbf{Y}} - \tilde{\mathbf{Y}}'\tilde{\mathbf{X}}_{(\boldsymbol{\gamma})}(\tilde{\mathbf{X}}'_{(\boldsymbol{\gamma})}\tilde{\mathbf{X}}_{(\boldsymbol{\gamma})})^{-1}\tilde{\mathbf{X}}'_{(\boldsymbol{\gamma})}\tilde{\mathbf{Y}}$$

the residual sum of squares from the least squares regression of $\tilde{\mathbf{Y}}$ on $\tilde{\mathbf{X}}_{(\boldsymbol{\gamma})}$. Fast updating schemes use Cholesky or QR decompositions with efficient algorithms to remove or add columns.

Model Fitting via MCMC

- With p variables there are 2^p different γ values. We use Metropolis as stochastic search.
- At each MCMC iteration we generate a candidate γ^{new} by randomly choosing one of these moves:
  - (i) **Add or Delete**: randomly choose one of the indices in γ^{old} and change its value.
  - (ii) **Swap**: choose independently and at random a 0 and a 1 in γ^{old} and switch their values.

  The proposed γ^{new} is accepted with probability

$$\min\left\{\frac{p(\gamma^{new}|\mathbf{X},\mathbf{Y})}{p(\gamma^{old}|\mathbf{X},\mathbf{Y})},1\right\}.$$

Posterior inference

The stochastic search results in a list of visited models $(\boldsymbol{\gamma}^{(0)}, \boldsymbol{\gamma}^{(1)}, \ldots)$ and their corresponding relative posterior probabilities

$$p(\boldsymbol{\gamma}^{(0)}|\mathbf{X}, \mathbf{Y}), p(\boldsymbol{\gamma}^{(1)}|\mathbf{X}, \mathbf{Y}) \ldots$$

Select variables:

- in the "best" models, i.e. the $\boldsymbol{\gamma}$'s with highest $p(\boldsymbol{\gamma}|\mathbf{X}, \mathbf{Y})$ or
- with largest marginal posterior probabilities

$$\begin{aligned} p(\gamma_j = 1|\mathbf{X}, \mathbf{Y}) &= \int p(\gamma_j = 1, \boldsymbol{\gamma}_{(-j)}|\mathbf{X}, \mathbf{Y}) d\boldsymbol{\gamma}_{(-j)} \\ &\approx \sum_{\boldsymbol{\gamma}:\gamma_j=1} p\left(\mathbf{Y}|\mathbf{X}, \boldsymbol{\gamma}^{(t)}\right) p(\gamma^{(t)}) \end{aligned}$$

or more simply by empirical frequencies in the MCMC output

$$p(\gamma_j = 1|\mathbf{X}, \mathbf{Y}) = E(\gamma_j = 1|\mathbf{X}, \mathbf{Y}) \approx \#\{\boldsymbol{\gamma}^{(t)} = 1\}$$

Multivariate Response

$$\mathbf{Y}_{n\times q} = \mathbf{1}\alpha' + \mathbf{X}_{n\times p}\mathbf{B}_{p\times q} + \mathbf{E}, \quad \mathbf{E}_i \sim N(0, \mathbf{\Sigma})$$

Variable selection via $\boldsymbol{\gamma}$ as

$$\boldsymbol{B}_j|\mathbf{\Sigma} \sim (1-\gamma_j)\mathcal{I}_0 + \gamma_j N(0, h_j\mathbf{\Sigma}),$$

with $\boldsymbol{B}_j$ the j-th row of $\boldsymbol{B}$ and $\mathcal{I}_0$ a vector of point masses at 0.

Need to work with matrix-variate distributions (Dawid, 1981):

$$\mathbf{Y} - \mathbf{1}\boldsymbol{\alpha}' - \mathbf{XB} \sim \mathcal{N}(\mathbf{I}_n, \mathbf{\Sigma})$$

$$\begin{aligned} \boldsymbol{\alpha} - \boldsymbol{\alpha}_0 &\sim \mathcal{N}(h_0, \mathbf{\Sigma}) \\ \mathbf{B}\boldsymbol{\gamma} - \mathbf{B}_0\boldsymbol{\gamma} &\sim \mathcal{N}(\mathbf{H}\boldsymbol{\gamma}, \mathbf{\Sigma}) \\ \mathbf{\Sigma} &\sim \mathcal{IW}(\delta, \mathbf{Q}). \end{aligned}$$

with $\mathcal{IW}$ an inverse-Wishart with parameters δ and $\mathbf{Q}$ to be specified.

Posterior Distribution

Combine data and prior information into a posterior distribution $\Rightarrow$ interest in posterior distribution

$$p(\boldsymbol{\gamma}|\mathbf{Y},\mathbf{X}) \propto p(\boldsymbol{\gamma}) \int f(\mathbf{Y}|\mathbf{X},\boldsymbol{\alpha},\mathbf{B},\Sigma) p(\boldsymbol{\alpha}|\Sigma) p(\mathbf{B}|\Sigma,\boldsymbol{\gamma}) p(\Sigma) d\boldsymbol{\alpha} d\mathbf{B} d\Sigma$$

$$p(\boldsymbol{\gamma}|\mathbf{Y},\mathbf{X}) \propto g(\boldsymbol{\gamma}) =$$

$$|\tilde{\mathbf{X}}'_{(\boldsymbol{\gamma})}\tilde{\mathbf{X}}_{(\boldsymbol{\gamma})}|^{-q/2} |\mathbf{Q}_{\boldsymbol{\gamma}}|^{-(n+\delta+q-1)/2} p(\boldsymbol{\gamma})$$

$$\tilde{\mathbf{X}}_{(\boldsymbol{\gamma})} = \begin{pmatrix} \mathbf{X}_{(\boldsymbol{\gamma})} \mathbf{H}^{\frac{1}{2}}_{(\boldsymbol{\gamma})} \\ I_{p_{\boldsymbol{\gamma}}} \end{pmatrix}, \quad \tilde{\mathbf{Y}} = \begin{pmatrix} \mathbf{Y} \\ 0 \end{pmatrix}$$

$$\mathbf{Q}_{\boldsymbol{\gamma}} = \mathbf{Q} + \tilde{\mathbf{Y}}'\tilde{\mathbf{Y}} - \tilde{\mathbf{Y}}'\tilde{\mathbf{X}}_{(\boldsymbol{\gamma})} (\tilde{\mathbf{X}}'_{(\boldsymbol{\gamma})}\tilde{\mathbf{X}}_{(\boldsymbol{\gamma})})^{-1} \tilde{\mathbf{X}}'_{(\boldsymbol{\gamma})}\tilde{\mathbf{Y}}$$

It can be calculated via *QR*-decomposition (Seber, ch.10, 1984). Use *qrdelete* and *qrinsert* algorithms to remove or add a column.

Prediction

Prediction of future Y^f given the corresponding $\mathbf{X}^f$ can be done:

- as posterior weighted average of model predictions (BMA)

$$p(Y^f|\mathbf{X},\mathbf{Y}) = \sum_{\gamma} p(\mathbf{Y}^f|\mathbf{X},\mathbf{Y},\gamma)p(\gamma|\mathbf{X},\mathbf{Y})$$

with $p(\mathbf{Y}^f|\mathbf{X},\mathbf{Y},\gamma)$ a matrix-variate T distribution with mean $\mathbf{X}^f\hat{\mathbf{B}}_\gamma$

$$\hat{Y}_f = \sum_{\gamma} \left(\mathbf{X}^f_\gamma \hat{\mathbf{B}}_\gamma\right) p(\gamma|\mathbf{X},\mathbf{Y})$$

$$\hat{\mathbf{B}}_\gamma = (\mathbf{X}'_\gamma\mathbf{X}_\gamma + \mathbf{H}_\gamma^{-1})^{-1}\mathbf{X}'_\gamma\mathbf{Y}$$

- as LS or Bayes predictions on single best models
- as LS or Bayes predictions with "threshold" models (eg, "median" model) obtained from estimated marginal probabilities of inclusion.

Prior Specification

Priors on $\boldsymbol{\alpha}$ and Σ vague and largely uninformative

$$\boldsymbol{\alpha}' - \boldsymbol{\alpha}_0' \sim \mathcal{N}(h, \boldsymbol{\Sigma}), \quad \boldsymbol{\alpha}_0 \equiv 0, h \to \infty,$$

$$\boldsymbol{\Sigma} \sim \mathcal{IW}(\delta, \mathbf{Q}), \quad \delta = 3, \mathbf{Q} = k\mathbf{I}$$

Choices for $H_{\boldsymbol{\gamma}}$:

- $\mathbf{H}_{\boldsymbol{\gamma}} = c * (\mathbf{X}_{\boldsymbol{\gamma}}'\mathbf{X}_{\boldsymbol{\gamma}})^{-1}$ (Zellner g-prior)
- $\mathbf{H}_{\boldsymbol{\gamma}} = c * diag(\mathbf{X}_{\boldsymbol{\gamma}}'\mathbf{X}_{\boldsymbol{\gamma}})^{-1}$
- $\mathbf{H}_{\boldsymbol{\gamma}} = c * I_{\boldsymbol{\gamma}}$

Choice of $w_j = p(\gamma_j = 1)$: $w_j = w, \quad w \sim Beta(a, b)$ (sparsity). Also, choices that reflect prior information (e.g., gene networks).

Advantages of Bayesian Approach

- Past and collateral information through priors
- $n << p$
- Rich modeling via Markov chain Monte Carlo (MCMC) (for p large)
- Optimal model averaging prediction
- Extends to multivariate response

Main References

- GEORGE, E.I. and MCCULLOCH, R.E. (1993). Variable Selection via Gibbs Sampling. *Journal of the American Statistical Association*, **88**, 881–889.
- GEORGE, E.I. and MCCULLOCH, R.E. (1997). Approaches for Bayesian Variable Selection. *Statistica Sinica*, **7**, 339–373.
- MADIGAN, D. and YORK, J. (1995). Bayesian Graphical Models for Discrete Data. *International Statistical Review*, **63**, 215–232
- BROWN, P.J., VANNUCCI, M. and FEARN, T. (1998). Multivariate Bayesian Variable Selection and Prediction. *Journal of the Royal Statistical Society, Series B*, **60**, 627–641.
- BROWN, P.J., VANNUCCI, M. and FEARN, T. (2002). Bayes model averaging with selection of regressors. *Journal of the Royal Statistical Society, Series B*, **64(3)**, 519–536.

Additional References

- Use of g-priors:
  LIANG, F., PAULO, R., MOLINA, G., CLYDE, M. and BERGER, J. (2008). Mixture of g priors for Bayes variable section. *Journal of the American Statistical Association*, **103**, 410-423.
- Improving MCMC mixing:
  BOTTOLO, L. and RICHARDSON, S. (2010). Evolutionary stochastic search for Bayesian Model Exploration. *Bayesian Analysis*, 5(3), 583-618. The authors propose an evolutionary Monte Carlo scheme combined with a parallel tempering approach that prevents the chain from getting stuck in local modes.
- Multiplicity:
  SCOTT, J. and BERGER, J. (2010). Bayes and empirical-Bayes multiplicity adjustment in the variable-selection problem. *The Annals of Statistics*, 38(5), 2587-2619. The marginal prior on γ contains a non-linear penalty which is a function of p and therefore, as p grows, with the number of true variables remaining fixed, the posterior distribution of w concentrates near 0.

Code from my Website

- bvsme_fast: Bayesian Variable Selection with fast form of QR updating
- Metropolis search
- gPrior or diagonal and non-diagonal selection prior
- Bernoulli priors or Beta-Binomial prior
- Predictions by LS, BMA and BMA with selection

http://stat.rice.edu/∼marina

Data augmentation techniques

- Binary response case.
- Basic idea: re-expression of discrete-data regression models as unobserved (latent) continuous data.
- Aids interpretation and allows convenient MCMC sampling.
- Used both for logistic and probit regression.
- Albert and Chib (1993) demonstrated an auxiliary variable approach to simplify binary probit regression.
- Introduce extra variables into model, y such that $z = g(y)$, g any non-decreasing function for interpretability.
- It can also be used for multinomial/ordinal data.

Probit link for binary outcome

The auxiliary variable formulation for binary outcomes assumes that a continuous latent variable Y_i exists such that

- The latent value Y_i is related to the binary z_i via

$$\begin{cases} z_i = 1 & \text{if } Y_i > 0 \\ z_i = 0 & \text{if } Y_i \leq 0 \end{cases}$$

- Associated with the i-th response, the values of p covariates $x_{i1}, \ldots, x_{ip}$ are observed.
- The latent value Y_i is related to the p covariates by the normal regression model

$$Y_i = x_{i1}\beta_1 + \ldots + x_{ip}\beta_p + \varepsilon_i \qquad \varepsilon_i \sim N(0,1)$$

[Define $y_i = \eta_i + \epsilon_i$, with $\epsilon_i \sim N(0,1)$ and $\eta_i = x_i'\beta$ (linear predictor - GLM)].

- Then we can show that

$$\begin{aligned} p(z_i = 1|\beta) &= p(z_i = 1|y_i > 0, \beta)p(y_i > 0|\beta) \\ &+ p(z_i = 1|y_i < 0, \beta)p(y_i < 0|\beta) \\ &= 1 \times p(y_i > 0|\beta) + 0 \times p(y_i < 0|\beta) \\ &= p(y_i - \eta_i > -\eta_i|\beta) \\ &= \Phi(\eta_i). \end{aligned}$$

  with $\eta_i = (x_{i1}\beta_1 + \ldots + x_{ip}\beta_p)$ and where $\Phi()$ is the cdf of a standard normal distribution.

- The latent values Y_i are viewed as additional parameters.
- Gibbs sampling can be used to obtain posterior draws of β and $Y = (Y_1, \ldots, Y_n)$.

If we specify, for example, a normal prior density for β

$$\beta \sim \mathcal{N}(0, S_0)$$

the full conditionals are

$$\begin{aligned}
\beta|Y, \boldsymbol{X}, z &\sim \mathcal{N}_p\left((\boldsymbol{X}^T\boldsymbol{X} + S_0^{-1})^{-1}\boldsymbol{X}^T Y, (\boldsymbol{X}^T\boldsymbol{X} + S_0^{-1})^{-1}\right) \\
Y_i|\beta, \boldsymbol{X}, z &\sim \begin{cases} \mathcal{N}(x_i\beta, 1) \cdot I\{Y_i > 0\} & \text{if } z_i = 1 \\ \mathcal{N}(x_i\beta, 1) \cdot I\{Y_i \leq 0\} & \text{if } z_i = 0 \end{cases}
\end{aligned}$$

Sampling from truncated normal density, $y \sim N(\mu, \sigma^2) \cdot I(a < y < b)$, via the inverse CDF transformation method:

1. Setting $u_1 = \Phi(a; \mu, \sigma^2)$ and $u_2 = \Phi(b; \mu, \sigma^2)$
2. Sampling $u \sim U(u_1, u_2)$
3. Setting $y = \Phi^{-1}(u; \mu, \sigma^2)$

Probit Models with Binary Response

- Response with $G = 2$ classes: $z_i \in \{0, 1\}$ associated with a set of p predictors $\mathbf{X}_i, i = 1, \ldots, n$.

- Data augmentation: Latent (unobserved) y_i linearly associated with the $\mathbf{X}_i$'s

$$y_i = \alpha + \mathbf{X}_i'\beta + \epsilon_i, \quad \epsilon_i \sim N(0, \sigma^2 = 1), \qquad i = 1, \ldots, n.$$

  with intercept α and coefficient vector $\beta_{p\times 1}$.

- Association

$$z_i = \begin{cases} 0 & \text{if } y_i < 0 \\ 1 & \text{if otherwise} \end{cases}$$

Probit Models with Multinomial Response

- Response with G classes: $z_i \in \{0, 1, ..., G-1\}$ associated with a set of p predictors $\mathbf{X}_i, i = 1, \cdots, n$ (gene expressions).

- Data augmentation: Latent (unobserved) vector $\mathbf{Y}_i$ linearly associated with the $\mathbf{X}_i$'s

$$\mathbf{Y}_i = \boldsymbol{\alpha}' + \mathbf{X}_i'\mathbf{B} + \mathbf{E}_i, \quad \mathbf{E}_i \sim N(0, \boldsymbol{\Sigma}), \qquad i = 1, \ldots, n.$$

  with intercepts $\boldsymbol{\alpha}_{(G-1)\times 1}$ and coefficient matrix $\mathbf{B}_{p\times(G-1)}$.

- Association

$$z_i = \begin{cases} 0 & \text{if } y_{ig} < 0 \text{ for each } g \\ g & \text{if } y_{ig} = \max\limits_{1\le g\le G-1} \{y_{ig}\} \end{cases}$$

Variable Selection

- We introduce a binary latent vector $\boldsymbol{\gamma}$ for variable selection

$$\begin{cases} \gamma_j = 1 & \text{if variable } j \text{ discriminate the samples} \\ \gamma_j = 0 & \text{otherwise} \end{cases}$$

- A mixture prior is placed on the jth row of $\mathbf{B}$, given $\boldsymbol{\gamma}$

$$\mathbf{B}_j \sim (1-\gamma_j)\mathcal{I}_0 + \gamma_j N(0, c\boldsymbol{\Sigma})$$

- Assume γ_j's are independent Bernoulli variables

- Combine data and priors into posterior $p(\boldsymbol{\gamma}|\mathbf{X}, \mathbf{Y})$. Inference is complicated because response variable is latent.

- $\boldsymbol{\Sigma} \sim IW(\delta; \mathbf{Q})$, $\boldsymbol{\alpha} \sim N(0, h_0\boldsymbol{\Sigma})$, large h_0.

MCMC Algorithm

We sample $(\boldsymbol{\gamma}, \mathbf{Y})$ by Metropolis within Gibbs

- Metropolis step to update $\boldsymbol{\gamma}$ from $[\boldsymbol{\gamma}|\mathbf{X}, \mathbf{Z}, \mathbf{Y}]$. We update $\boldsymbol{\gamma}^{(old)}$ to $\boldsymbol{\gamma}^{(new)}$ by:

  (a) **Add/delete**: randomly choose a γ_j and change its value.
  (b) **Swap**: randomly choose a 0 and a 1 in $\boldsymbol{\gamma}^{old}$ and switch values.

  The new candidate $\boldsymbol{\gamma}^{(new)}$ is accepted with probability

$$\min\{\frac{p(\boldsymbol{\gamma}^{(new)}|\mathbf{X}, \mathbf{Z}, \mathbf{Y})}{p(\boldsymbol{\gamma}^{(old)}|\mathbf{X}, \mathbf{Z}, \mathbf{Y})}, 1\}$$

- We sample $(\mathbf{Y}|\boldsymbol{\gamma}, \mathbf{X}, \mathbf{Z})$ from a *truncated* normal or t distribution with truncation based on $\mathbf{Z}$.

Posterior Inference

- Select variables that are in the "best" models

$$\widehat{\boldsymbol{\gamma}}* = \underset{1\leq t\leq M}{\text{argmax}} \left\{ p(\boldsymbol{\gamma}^{(t)}|\mathbf{X}, \mathbf{Z}, \widehat{\mathbf{Y}}) \right\}, \text{ with } \widehat{\mathbf{Y}} = \frac{1}{M}\sum_{t=1}^{M} \mathbf{Y}^{(t)}$$

- Select variables with largest marginal probabilities

$$p(\gamma_j = 1|\mathbf{X}, \mathbf{Z}, \hat{\mathbf{Y}})$$

- Predict future Y_f by a posterior predictive mean

$$\hat{Y}_f = \sum_{\boldsymbol{\gamma}} \hat{Y}_{f(\boldsymbol{\gamma})} \pi(\boldsymbol{\gamma}|\hat{\mathbf{Y}}, \mathbf{X}, \mathbf{Z})$$

with $Y_{f(\boldsymbol{\gamma})} = \mathbf{1}\tilde{\alpha}' + \mathbf{X}_{f(\boldsymbol{\gamma})}\tilde{\mathbf{B}}_{\boldsymbol{\gamma}}$ and $\tilde{\boldsymbol{\alpha}}$, $\tilde{\mathbf{B}}_{\boldsymbol{\gamma}}$ based on $\hat{\mathbf{Y}}$

Code from my website

- bvsme_prob: Bayesian Variable Selection for classification with fast form of QR updating
- binary/multinomial/ordinal response
- Metropolis search
- gPrior or diagonal and non-diagonal selection prior
- Bernoulli priors or Beta-Binomial prior
- Predictions by LS, BMA and BMA with selection

http://stat.rice.edu/∼marina

Logit Models

- More naturally interpretable in terms of odds ratios. Marginalization not possible.
- For binary data, a data augmented model is

$$z_i = \mathbf{x}_i'\boldsymbol{\beta} + \epsilon_i,$$

with ϵ_i a scale mixture of normals with marginal logistic,

$$\begin{aligned} \epsilon_i &\sim N(0, \lambda_i) \\ \lambda_i &= (2\psi_i)^2 \\ \psi_i &\sim KS, \end{aligned}$$

with *KS* the Kolmogorov-Smirnov distribution.
- Variable selection is achieved by imposing mixture priors on β_j's.
- Sampling schemes improve mixing by joint updates of correlated parameters, i.e, $(\boldsymbol{\gamma}, \boldsymbol{\beta})$ using a Metropolis-Hastings with proposal the full conditional of $\boldsymbol{\beta}$ and the add-delete-swap Metropolis for $\boldsymbol{\gamma}$. Also, $(\mathbf{z}, \boldsymbol{\lambda})$ from truncated logistics and rejection sampling.

Accelerated Failure Time models for Survival Outcomes

We use accelerated failure time (AFT) models

$$\log(T_i) = \alpha + \mathbf{X}_i'\beta + \varepsilon_i, \quad i = 1, \ldots, n.$$

Observe $y_i = \min(t_i, c_i)$ and $\delta_i = I\{t_i \leq c_i\}$, where c_i censoring time.

- We introduce augmented data $\mathbf{W} = (w_1, \ldots, w_n)'$ to impute the censored survival times

$$\begin{cases} w_i = \log(y_i) & \text{if } \delta_i = 1 \\ w_i > \log(y_i) & \text{if } \delta_i = 0 \end{cases}$$

- We consider different distributional assumptions for ε_i.

- Introduce latent vector $\boldsymbol{\gamma}$ for variable selection.
- MCMC steps consist of
  (1) Metropolis search to update $\boldsymbol{\gamma}$ from $f(\boldsymbol{\gamma}|\mathbf{X}, \mathbf{W})$.
  (2) Impute censored failure times, w_i with $\delta_i = 0$, from $f(w_i|\mathbf{W}_{-i}, \mathbf{X}, \boldsymbol{\gamma})$.
- Inference on variables based on $p(\gamma_j = 1|\mathbf{X}, \widetilde{\mathbf{W}})$ or $p(\boldsymbol{\gamma}|\mathbf{X}, \widetilde{\boldsymbol{W}})$.

- Prediction of survival time for future patients

$$\widehat{\mathbf{W}}_f = \sum_{\boldsymbol{\gamma}} \left(\mathbf{1}\widehat{\alpha}' + \mathbf{X}_{f(\boldsymbol{\gamma})}\widehat{\beta}_{\boldsymbol{\gamma}} \right) p(\boldsymbol{\gamma}|\mathbf{X}, \widetilde{\mathbf{W}}).$$

- Predictive survivor function

$$P(T_f > t|\mathbf{X}_f, \mathbf{X}, \widetilde{\mathbf{W}}) \approx \sum_{\boldsymbol{\gamma}} P\left(W > w|\mathbf{X}_f, \mathbf{X}, \widetilde{\mathbf{W}}, \boldsymbol{\gamma} \right) p(\boldsymbol{\gamma}|\mathbf{X}, \widetilde{\mathbf{W}}).$$

Code from my website

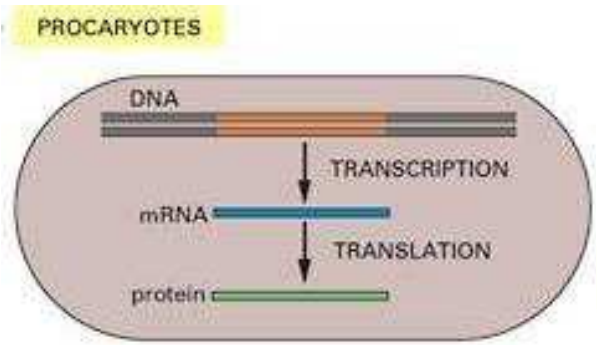
- bvsme_surv: Bayesian Variable Selection for AFT models with right censoring
- Metropolis search
- diagonal selection prior
- Bernoulli priors or Beta-Binomial prior

http://stat.rice.edu/∼marina

Main References

- ALBERT, J.H. and CHIB, S. (1993), "Bayesian Analysis of Binary and Polychotomous Response Data", *JASA*, **88(422)**, 669-679.
- SHA, N., VANNUCCI, M., TADESSE, M.G., BROWN, P.J., DRAGONI, I., DAVIES, N., ROBERTS, T. C., CONTESTABILE, A., SALMON, N., BUCKLEY, C. and FALCIANI, F. (2004). Bayesian variable selection in multinomial probit models to identify molecular signatures of disease stage, *Biometrics*, **60**, 812-819.
- HOLMES, C.C. and HELD, L. (2006). Bayesian auxiliary variable models for binary and multinomial regression. *Bayesian Analysis*, **1(1)**, 145-166. See also Comment by Ralf van der Lans (2011) *Bayesian Analysis*, **6(2)**, 353-355 and response from authors.
- SHA, N., TADESSE, M.G. and VANNUCCI, M. (2006). Bayesian variable selection for the analysis of microarray data with censored outcome. *Bioinformatics*, **22(18)**, 2262-2268.

Applications to High-Throughput Data: DNA microarrays



- DNA fragments arrayed on glass slide or chip
- Parallel quantification of thousands of genes in a single experiment
- Identify biomarkers for treatment strategies and diagnostic tools

Statistical Analyses

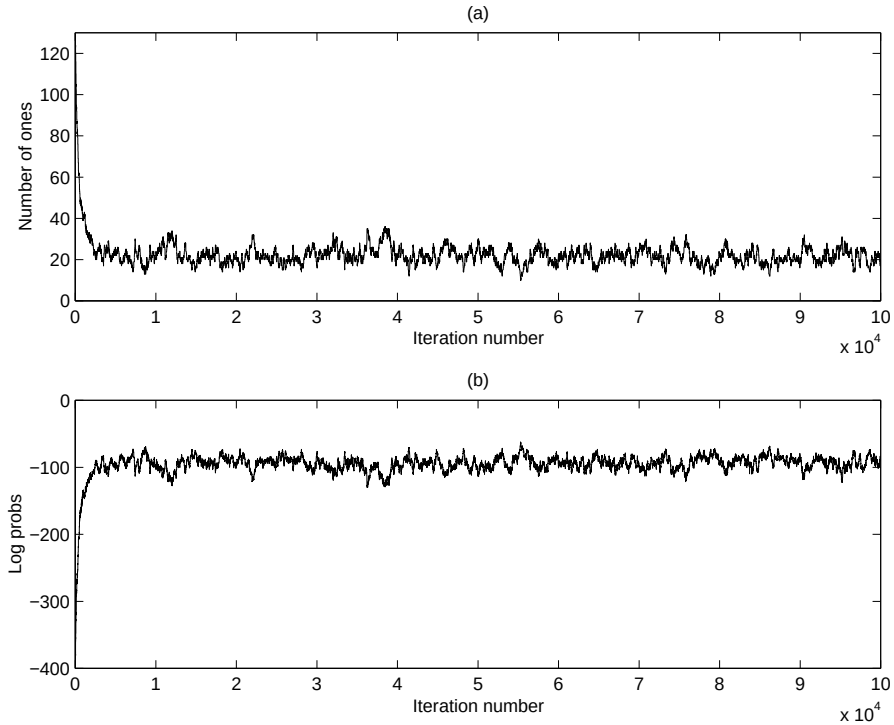
- Identification of differentially expressed genes (gene selection for sample classification)
- Discovery of subtypes of tissue/disease that respond differently to treatment (gene selection and sample clustering)
- Prediction of continuous responses (clinical outcome, survival time)
- The major challenge is the high-dimensionality of the data.

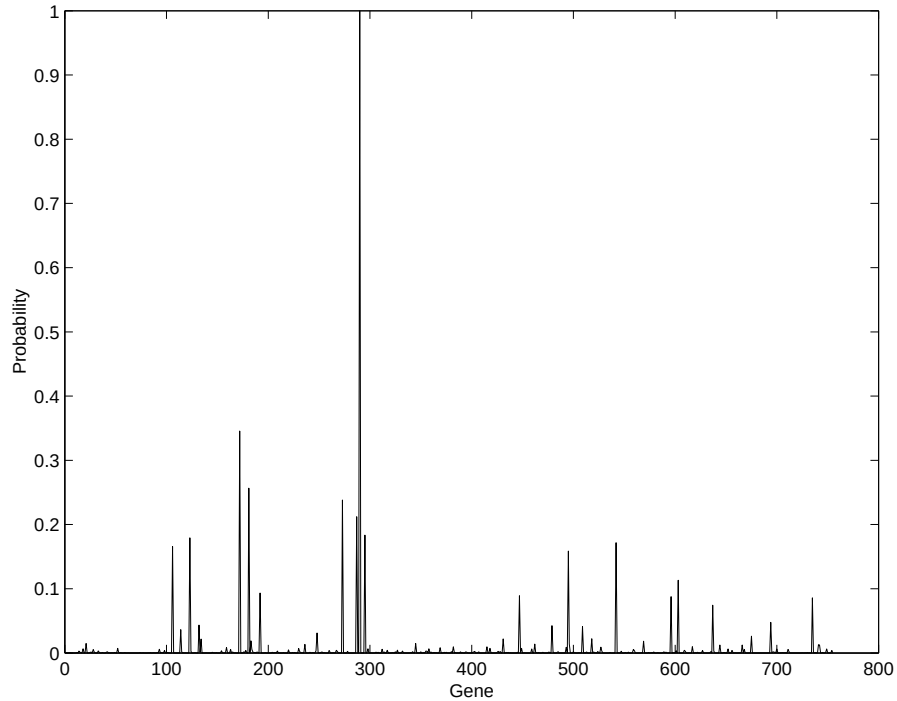
$$p \gg n$$

- Widely used approaches: t-test, ANOVA, Cox model on single genes (ignores joint effect of genes; multiple testing issue) or dimension reduction techniques, PCA, PLS (leads to linear combinations; cannot assess original variables). Lately emphasis on subset selection methods (LASSO, Bayesian models).

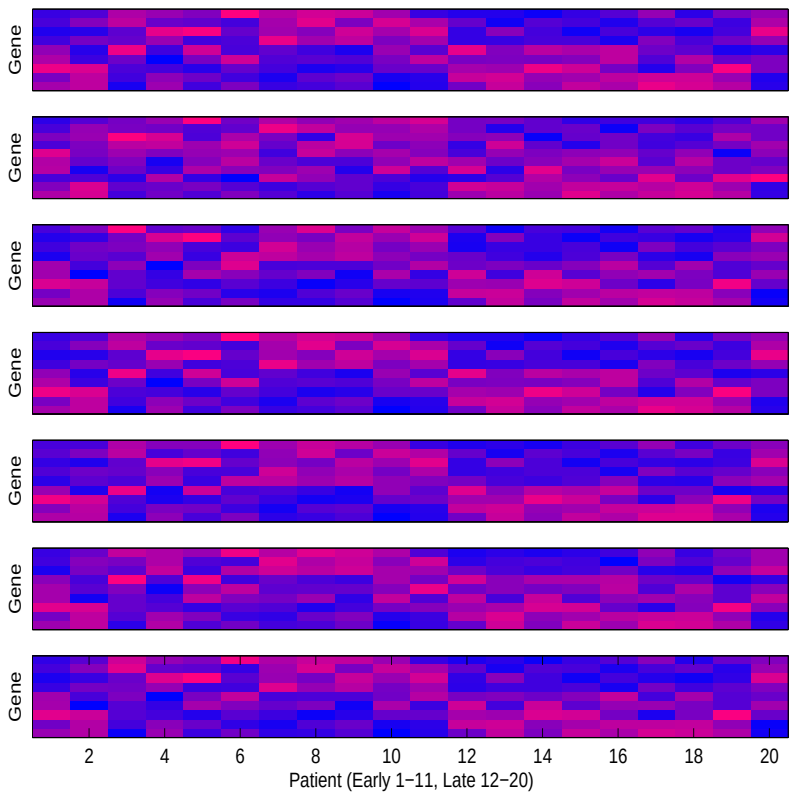
Identification of Biomarkers of Disease Stage

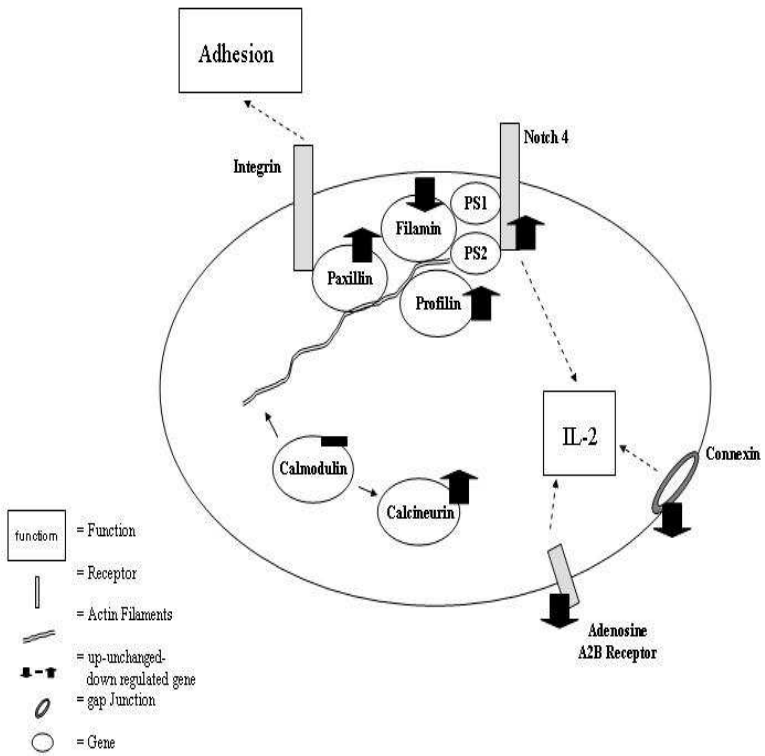
- Data consist of 11 early stage (duration less than 2 years) and 9 late stage (over 15 years) rheumatoid arthritis patients.
- mRNA samples extracted from peripheral blood and hybridized to custom-made cDNA arrays.
- 755 gene expressions. Logged and std-ed data
- Bernoulli prior with expected model size 10
- We ran six MCMC chains with very different starting γ vectors.


(a)
Number of ones
Iteration number
x 10^4
(b)
Log probs
Iteration number
x 10^4


Probability
Gene

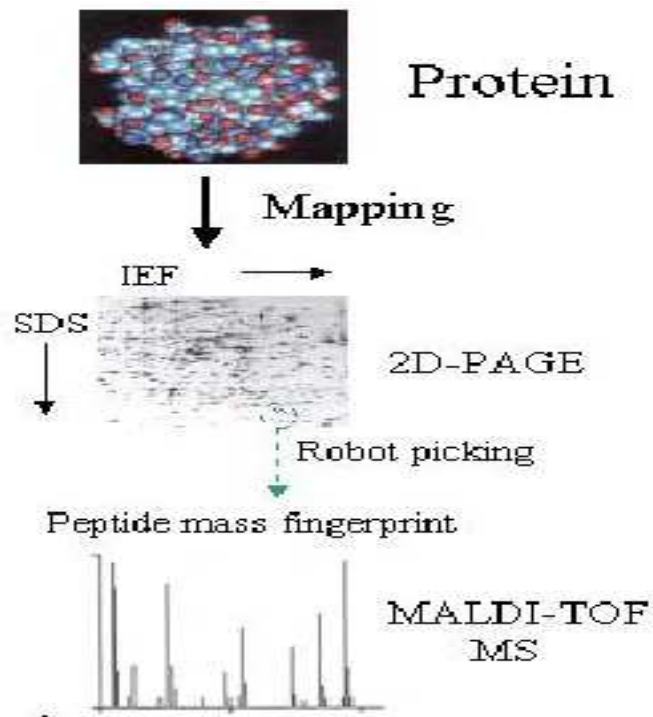
- Selected genes by best 10 models of each chain and of their union
- Small sets of functionally related genes involved in cytoskeleton remodeling and motility, and with lymphocytes' ability to respond to activation.
- .05(1/20) misclassification error.

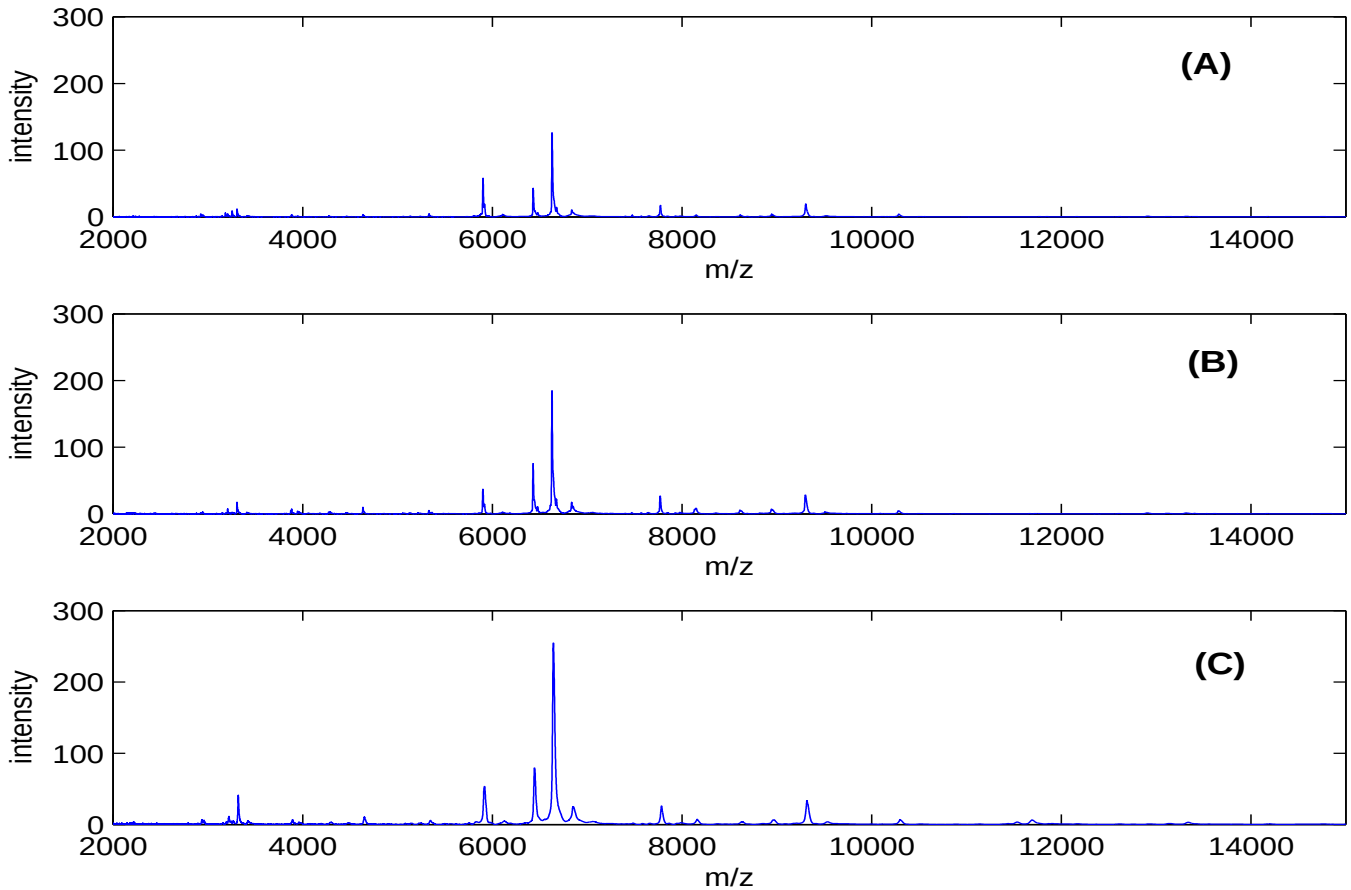




Peak Selection for Protein Mass spectra

- Cancer classification based on mass spectra at 15,000 *m/z* ratios.
- *x*-axis: ratio of weight of a molecule to its electrical charge (m/z), *y*-axis: intensity $\sim$ abundance of that molecule in the sample.
- Goal: identification of peaks (proteins or protein fragments) related to a clinical outcome or disease status





Serum spectra on 50 (10+11+29) subjects (SELDI-TOF). Ordinal response - tumor grade (ovarian cancer).

Data Processing

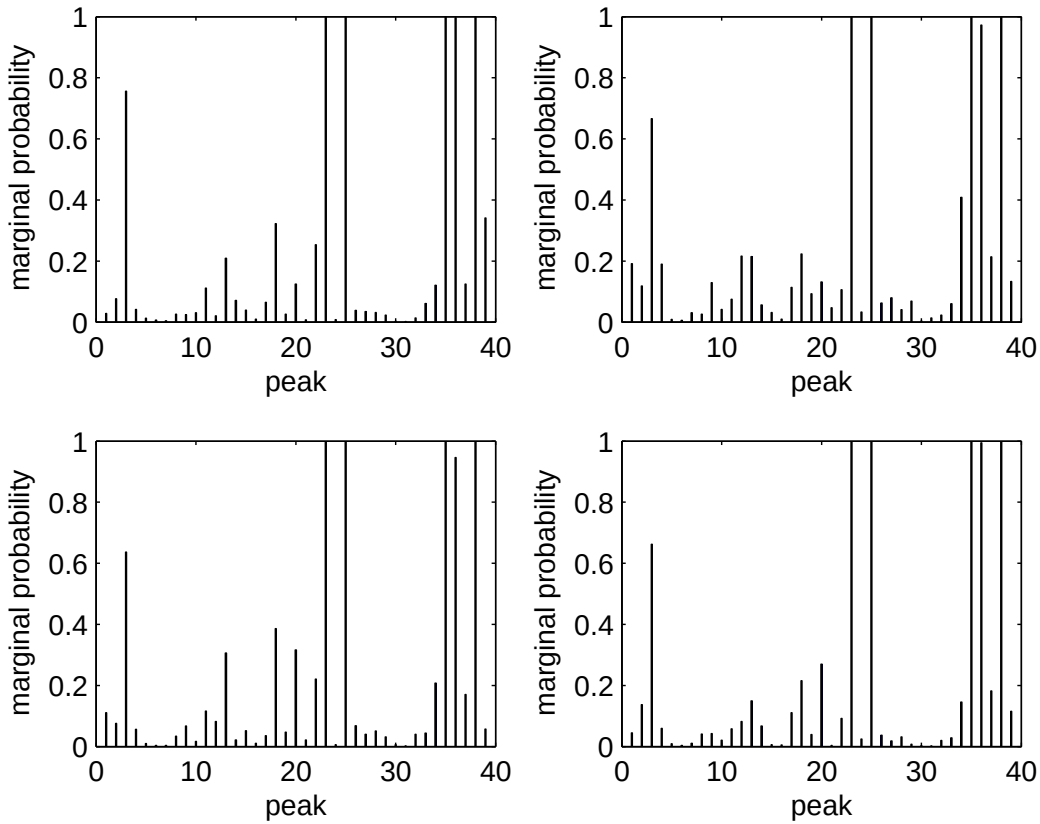
- Preprocessing:

  - Baseline subtraction
  - Denoising (often by wavelets)
  - Peak identification
  - Normalization
  - Alignment

- Analysis:

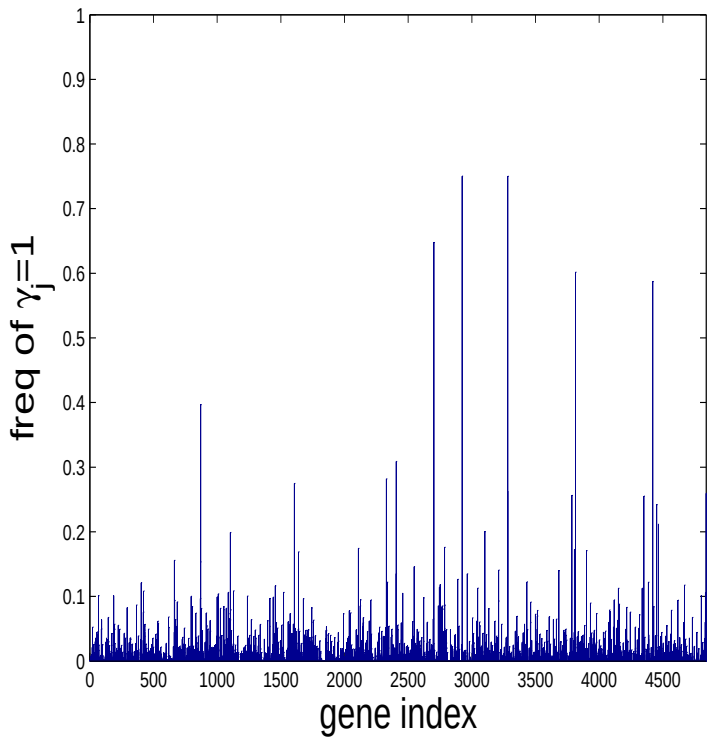
  - Model fitting
  - Validation

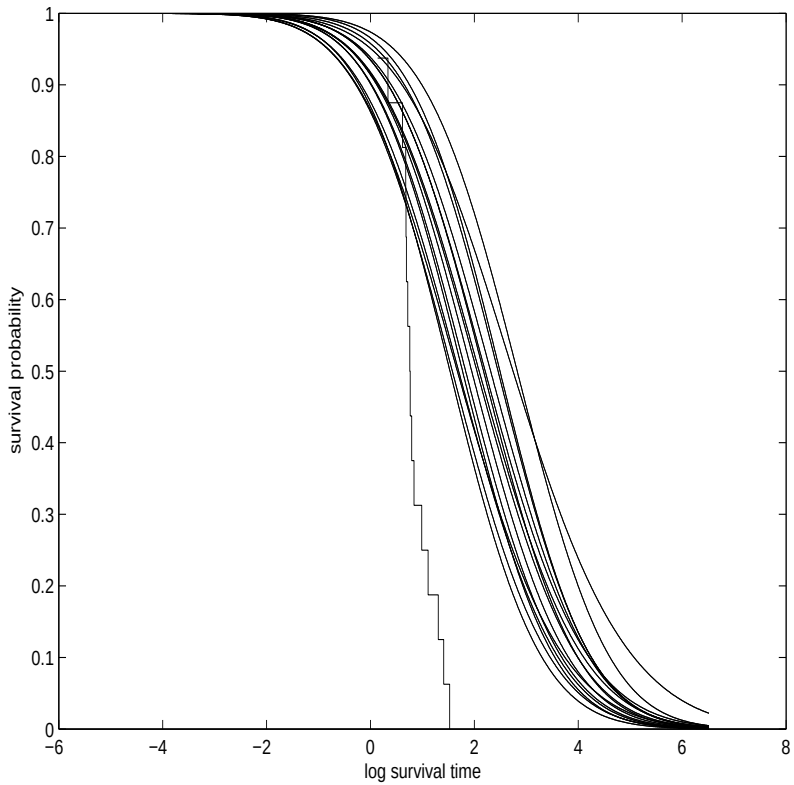
- Data processing results in 39 identified peaks.
- Probit model with Bayesian variable selection applied to 39 peaks.
- “Best” models with around 7 peptides (6 common).
- Misclassification errors $(2/10, 8/11, 9/29)$



Survival outcome: Case Study on Breast Cancer (van't Veer *et al.* (2002))

- Microarray data on 76 patients, 33 who developed distant metastases within 5 years and 43 who did not (censored).
- Training and test sets (38+38 patients). About 5,000 genes. MSE=1.9 (with 11 genes).





Main References

- SHA, N., VANNUCCI, M., TADESSE, M.G., BROWN, P.J., DRAGONI, I., DAVIES, N., ROBERTS, T. C., CONTESTABILE, A., SALMON, N., BUCKLEY, C. and FALCIANI, F. (2004). Bayesian variable selection in multinomial probit models to identify molecular signatures of disease stage, *Biometrics*, **60**, 812-819.
- KWON, D.W., TADESSE, M.G., SHA, N., PFEIFFER, R.M. and VANNUCCI, M. (2007). Identifying biomarkers from mass spectrometry data with ordinal outcome. *Cancer Informatics*, **3**, 19–28.
- SHA, N., TADESSE, M.G. and VANNUCCI, M. (2006). Bayesian variable selection for the analysis of microarray data with censored outcome. *Bioinformatics*, **22(18)**, 2262-2268.
- TADESSE, M.G., SHA, N., KIM, S. and VANNUCCI, M. (2006). Identification of biomarkers in classification and clustering of high-throughput data. In *Bayesian Inference for Gene Expression and Proteomics*, K. Anh-Do, P. Mueller and M. Vannucci (Eds). Cambridge University Press.