# Bayesian variable selection with a focus on the analysis of genomic data - Part II

*Veronika Ročková*[1] *Emmanuel Lesaffre*[1,2]

[1]Erasmus University Rotterdam [2]KU Leuven, Belgium

22$^{nd}$ May 2013

Bayes 2013

# Challenges in High-dimensional Data

- **"High-dimensionality"**: growing data dimension along with the sample size, where $p_n > n$
- Important attributes of statistical procedures:
  - accuracy of inference
  - computational tractability
- **Challenges** in high-dimensional data
  - What is the sensible *threshold of dimensionality* to apply a statistical procedure?
  - Characterization of *optimality attributes*
  - Development of *reliable inferential tools*
  - *"Low assumptions in high dimensions"*

Erasmus MC
University Medical Center Rotterdam

# High Dimensions $\sim$ Strong Assumptions

- What is the convenient dimensionality of the parametrization?
- A crucial assumption is the one of sparsity
  - *parametrization using only a few coefficients*
  - *often in line with biological intuition*
- *Sparsity central to the implementation of variable selection!*
- **Bayesian variable selection**: natural incorporation of the prior knowledge on the pattern of sparsity

# Computational Aspects

- Penalized likelihood methods: (LASSO, SCAD...)
  - *easy for convex penalties*
  - *non-convex penalties*

    (Fan and Li (2009), Hunter and Li (2005))
- Bayesian shrinkage methods: (Bayesian LASSO...)
  - *MCMC with block updates*
  - *MAP estimation using EM algorithm*

    (Griffin and Brown (2005), Rockova and Lesaffre (2013))
- Bayesian variable selection: (spike and slab, SSVS)
  - *MCMC*

    (George and McCulloch (1993); Hans et al. (2010))
  - *EM algorithm for posterior model mode detection*

    (Rockova and George (2012))

# SSVS Setup

- Assume $\boldsymbol{Y} \sim \mathrm{N}_n(\alpha + \boldsymbol{X}\beta, \sigma^2 \mathrm{I}_n)$, interest in $p > n$
- Binary variable selection indicators $\gamma = (\gamma_1, \ldots, \gamma_p)'$, where $\gamma_i = 0$ if $\beta_i$ is "small" and $\gamma_i = 1$ if $\beta_i$ is "large"
- Conjugate "spike and slab" prior on regression coefficients

$$\pi(\beta_i \mid \sigma, \gamma) = \mathrm{N}(0, \sigma^2[(1-\gamma_i)v_0 + \gamma_i v_1]),$$

$\gamma_i = 0$: Spike variance $\sigma^2 v_0$ small
$\gamma_i = 1$: Slab variance $\sigma^2 v_1$ large

- Prior distribution for the variance $\pi(\sigma^2 \mid \gamma) = \mathrm{IG}(\nu/2, \nu\lambda/2)$
- Uniform improper prior on the intercept $\alpha \rightsquigarrow$ margined out

# SSVS Setup

- Assume $\boldsymbol{Y} \sim \mathrm{N}_n(\alpha + \boldsymbol{X}\boldsymbol{\beta}, \sigma^2 \mathrm{I}_n)$, interest in $p > n$
- Binary variable selection indicators $\boldsymbol{\gamma} = (\gamma_1, \ldots, \gamma_p)'$, where $\gamma_i = 0$ if $\beta_i$ is “small" and $\gamma_i = 1$ if $\beta_i$ is “large"
- Conjugate “spike and slab" prior on regression coefficients

$$\pi(\beta_i \mid \sigma, \gamma) = \mathrm{N}(0, \sigma^2[(1-\gamma_i)v_0 + \gamma_i v_1]),$$

$\gamma_i = 0$: Spike variance $\sigma^2 v_0$ small
$\gamma_i = 1$: Slab variance $\sigma^2 v_1$ large

- Prior distribution for the variance $\pi(\sigma^2 \mid \gamma) = \mathrm{IG}(\nu/2, \nu\lambda/2)$
- Uniform improper prior on the intercept $\alpha \rightsquigarrow$ margined out

# SSVS Setup

- Assume $\boldsymbol{Y} \sim \mathrm{N}_n(\alpha + \boldsymbol{X}\boldsymbol{\beta}, \sigma^2 \mathrm{I}_n)$, interest in $p > n$
- Binary variable selection indicators $\boldsymbol{\gamma} = (\gamma_1, \ldots, \gamma_p)'$, where $\gamma_i = 0$ if $\beta_i$ is “small" and $\gamma_i = 1$ if $\beta_i$ is “large"
- Conjugate “spike and slab" prior on regression coefficients

$$\pi(\beta_i \mid \sigma, \boldsymbol{\gamma}) = \mathrm{N}(0, \sigma^2[(1-\gamma_i)v_0 + \gamma_i v_1]),$$

$\gamma_i = 0$: Spike variance $\sigma^2 v_0$ small
$\gamma_i = 1$: Slab variance $\sigma^2 v_1$ large

- Prior distribution for the variance $\pi(\sigma^2 \mid \gamma) = \mathrm{IG}(\nu/2, \nu\lambda/2)$
- Uniform improper prior on the intercept $\alpha \rightsquigarrow$ margined out

# SSVS Setup

- Assume $\boldsymbol{Y} \sim \mathrm{N}_n(\alpha + \boldsymbol{X}\boldsymbol{\beta}, \sigma^2 \mathrm{I}_n)$, interest in $p > n$
- Binary variable selection indicators $\boldsymbol{\gamma} = (\gamma_1, \ldots, \gamma_p)'$, where $\gamma_i = 0$ if $\beta_i$ is “small" and $\gamma_i = 1$ if $\beta_i$ is “large"
- Conjugate “spike and slab" prior on regression coefficients

$$\pi(\beta_i \mid \sigma, \boldsymbol{\gamma}) = \mathrm{N}(0, \sigma^2[(1 - \gamma_i)v_0 + \gamma_i v_1]),$$

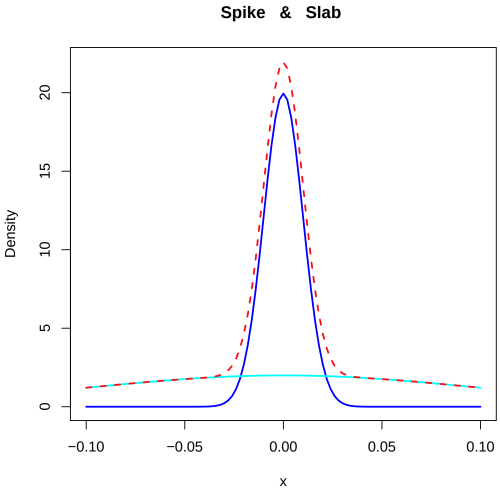
$\gamma_i = 0$: Spike variance $\sigma^2 v_0$ small
$\gamma_i = 1$: Slab variance $\sigma^2 v_1$ large

- Prior distribution for the variance $\pi(\sigma^2 \mid \gamma) = \mathrm{IG}(\nu/2, \nu\lambda/2)$
- Uniform improper prior on the intercept $\alpha \rightsquigarrow$ margined out

Erasmus MC
University Medical Center Rotterdam

# SSVS Setup

Spike and slab prior for $v_0 = 0.01, v_1 = 0.1$



# Stochastic Model Search

- Favored selection criteria based on $\pi(\gamma | \boldsymbol{Y})$
  (a) *Highest posterior probability model*

$$\text{argmax}_{\gamma} \pi(\gamma \mid \boldsymbol{Y})$$

  (b) *Median probability model*: select variables with

$$\text{P}(\gamma_i = 1 \mid \boldsymbol{Y}) > 0.5$$

- MCMC stochastic search algorithms attempt to find these
  - *SSVS* (George and McCulloch (1993))
  - *ESS* (Botollo and Richardson (2010))
  - *SSS* (Hans et al. (2010))
- Slow and inefficient, especially when $p$ is large.
- Is there a better way?

Erasmus MC
University Medical Center Rotterdam

# Deterministic Model Search

Rockova and George (2012) propose an EM model search algorithm

(1) "$\pi(\gamma \mid \boldsymbol{Y}) \leftrightarrow \pi(\beta \mid \boldsymbol{Y})$"

- $\rightsquigarrow$ High posterior modes of $\pi(\gamma \mid \boldsymbol{Y})$ can be located by thresholding small coefficient estimates of associated high posterior modes of $\pi(\beta \mid \boldsymbol{Y})$
- $\rightsquigarrow$ Modes of the posterior $\pi(\beta \mid \boldsymbol{Y})$ can be found deterministically

(2) *"Spike-and-slab Regularization Diagram"*

- $\rightsquigarrow$ Obtain modal estimates for a sequence of mixture priors with increasing $v_0 > 0$
- $\rightsquigarrow$ Depicts evolution and gradual sparsification of selected subsets

# Deterministic Model Search

Rockova and George (2012) propose an EM model search algorithm

(1) “$\pi(\gamma \mid \boldsymbol{Y}) \leftrightarrow \pi(\beta \mid \boldsymbol{Y})$”

- $\rightsquigarrow$ High posterior modes of $\pi(\gamma \mid \boldsymbol{Y})$ can be located by thresholding small coefficient estimates of associated high posterior modes of $\pi(\beta \mid \boldsymbol{Y})$
- $\rightsquigarrow$ Modes of the posterior $\pi(\beta|\boldsymbol{Y})$ can be found deterministically

(2) *“Spike-and-slab Regularization Diagram"*

- $\rightsquigarrow$ Obtain modal estimates for a sequence of mixture priors with increasing $v_0 > 0$
- $\rightsquigarrow$ Depicts evolution and gradual sparsification of selected subsets

SSVS $\rightarrow$ EMVS

# Key Ingredients

(1) *Conjugacy*

- $\rightsquigarrow$ Allows analytical simplifications for the EM algorithm
- $\rightsquigarrow$ Enables computation of posterior model probabilities

(2) *Use of both* $v_0 > 0$ *and* $v_0 = 0$

- $\rightsquigarrow$ $v_0 > 0$: Feasible closed form EM algorithm
- $\rightsquigarrow$ $v_0 > 0$: Spike distribution absorbs small coefficients
- $\rightsquigarrow$ $v_0 = 0$: Correct posterior for candidate model evaluation

(3) $\pi(\gamma|\boldsymbol{\theta})$ *Flexibility*

- $\rightsquigarrow$ Allows incorporation of covariate pattern information

# EMVS Algorithm

GOAL To locate posterior mode

$$\text{argmax}_{\boldsymbol{\beta},\boldsymbol{\theta},\sigma} \log \pi(\boldsymbol{\beta}, \boldsymbol{\theta}, \sigma^2 \mid \boldsymbol{y}) \tag{1}$$

IDEA Solve this via EM by treating $\gamma$ as “missing data" and focusing on

$$\log \pi(\boldsymbol{\beta}, \boldsymbol{\theta}, \sigma^2, \gamma \mid \boldsymbol{y})$$

## *E-step*

Compute conditional expectation of “log complete data posterior":

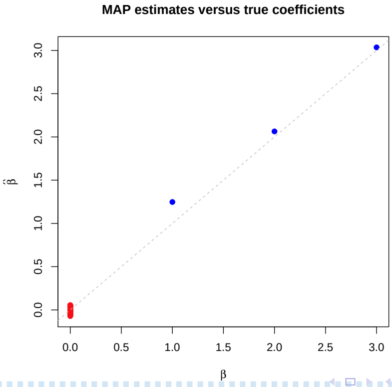
$$Q\left(\boldsymbol{\beta}, \boldsymbol{\theta}, \sigma \mid \boldsymbol{\beta}^{(k)}, \boldsymbol{\theta}^{(k)}, \sigma^{(k)}\right) = \mathsf{E}_{\gamma|\cdot}\left[\log \pi(\boldsymbol{\beta}, \boldsymbol{\theta}, \sigma, \gamma|\boldsymbol{y}) \mid \boldsymbol{\beta}^{(k)}, \boldsymbol{\theta}^{(k)}, \sigma^{(k)}, \boldsymbol{y}\right]$$

## *M-step*

Maximize $Q\left(\boldsymbol{\beta}, \boldsymbol{\theta}, \sigma \mid \boldsymbol{\beta}^{(k)}, \boldsymbol{\theta}^{(k)}, \sigma^{(k)}\right)$ to get $(\boldsymbol{\beta}^{(k+1)}, \boldsymbol{\theta}^{(k+1)}, \sigma^{(k+1)})$

# Simple Implementation

- Assume $n = 100, p = 1\,000, \beta = (1, 2, 3, 0, \ldots, 0)'$
- Rows in $\boldsymbol{X}$ sampled from $\mathrm{N}_p(\mathbf{0}, \Sigma)$, where $\Sigma = (0.6^{|i-j|})_{i,j=1}^{p}$
- $\boldsymbol{Y} = \boldsymbol{X}\beta + \varepsilon$, where $\varepsilon \sim \mathrm{N}_n(\mathbf{0}, \sigma^2\,\mathrm{I}_p)$ and $\sigma^2 = 3$
- For now fixed $v_0 = 1$ and $v_1 = 1\,000$, $\theta \sim \mathrm{Beta}(1, 1)$



# Simple Implementation

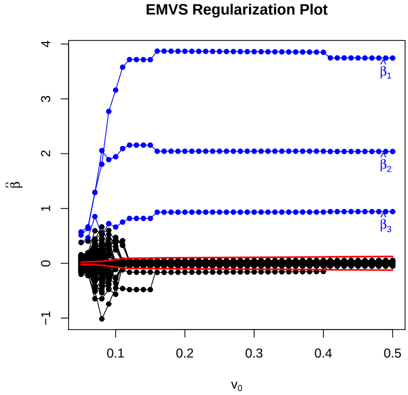
- Subset selection for fixed $v_0$:
  (a) *Based on conditional inclusion probabilities*
  Select $X_i$ when $\mathrm{P}(\gamma_i = 1|\widehat{\boldsymbol{\beta}}, \widehat{\sigma}, \widehat{\boldsymbol{\theta}}) > 0.5$
  (b) *Based on modal estimates* $\widehat{\boldsymbol{\beta}}$ (equivalent to (a))
  Select $X_i$ when $|\widehat{\beta}_i| > \mu_{v_0,v_1,\widehat{\sigma}} = \widehat{\sigma}\sqrt{2v_0 \log(\omega_i\, c)\, c^2/(c^2-1)}$
  with $c^2 = v_1/v_0$ and $\omega_i = [1 - \mathrm{P}(\gamma_i = 1 \mid \widehat{\boldsymbol{\theta}})]/\mathrm{P}(\gamma_i = 1 \mid \widehat{\boldsymbol{\theta}})$
- We can consider a grid $V$ of values $v_0$ and $\forall v_0 \in V$ determine an active set $\mathcal{S}_{v_0} = \{1 \leq i \leq p : |\widehat{\beta}_i| > \mu_{v_0,v_1,\widehat{\sigma}}\}$
- For each active set we evaluate $\gamma$ using

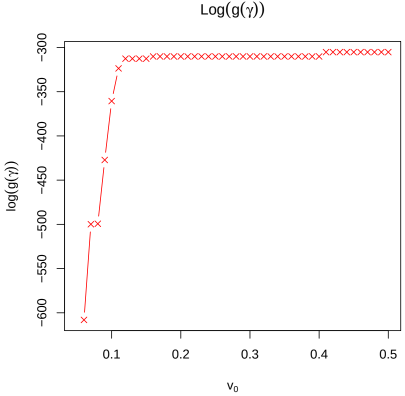
$$g_{v_0}(\gamma) = C\, p(\gamma|\boldsymbol{Y})$$

assuming that $v_0 = 0$ (correct submodel evaluation)

# Simple Implementation Continued

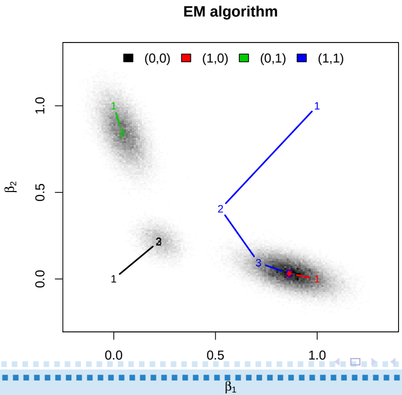
- [Regularization plot](#) for a grid of values $v_0$
- Starting values $\beta^{(0)} = \mathbf{1}_p, \sigma^{(0)} = 1$ and $\theta^{(0)} = 0.5$
- Increasing $v_0$ absorbs smaller coefficients



# Multimodality Issues

- EM algorithm guarantees monotonical convergence towards at least a local maximum
- Prone to entrapment around local modes
- Posterior from conjugate model with two correlated predictors, $\beta = (1, 0)'$, $v_1 = 1\,000$ and $v_0 = 0.005$, $\widehat{\beta}_{MLE} = (0.52, 0.4)'$



# Deterministic Annealing

- Maximize a tempered version of the objective function: for $0 < t < 1$

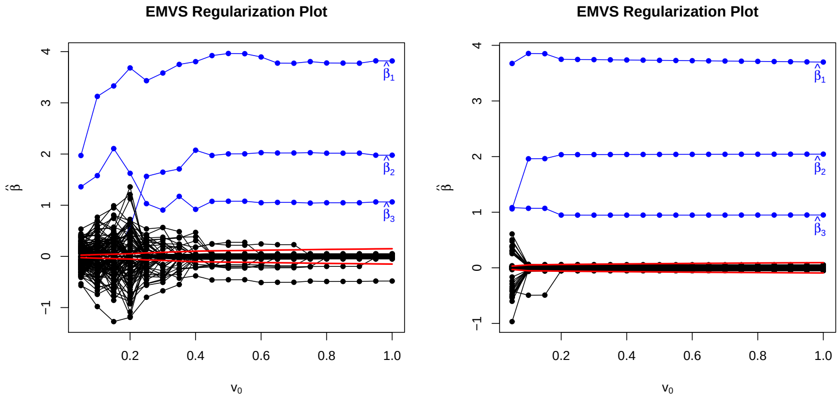
$$H_t(\boldsymbol{\beta}, \boldsymbol{\theta}, \sigma) = \frac{1}{t} \log \sum_{\gamma} \pi(\boldsymbol{\beta}, \boldsymbol{\theta}, \sigma, \gamma \mid \boldsymbol{y})^t \tag{2}$$

- Temperature $1/t$ regulates the degree of separation between multiple modes
- Small values $t$ smooth the function to have only one mode
- Consider temperature ladder $1/t_1 < 1/t_2 < \cdots < 1/t_T$
- Solutions at lower temperature can be used as starting points for computation at higher temperature

# Simple Implementation Continued

- [Regularization plot] for a grid of values $v_0$, $v_1 = 1\,000$
- Randomly generated starting values $\beta^{(0)} \sim \mathrm{N}_p(\mathbf{0}, \mathrm{I})$

Temperature 1 | Temperature 10



# Structured Model Priors

⇝ Variable selection indicators assigned prior distribution $\pi(\gamma \mid \boldsymbol{\theta})$

(a) *Beta-binomial prior* (George and McCulloch (1993))

$$\pi(\gamma \mid \theta) = \theta^{\sum \gamma_i}(1-\theta)^{p-\sum \gamma_i} \quad \text{with} \quad \theta \sim B(a, b)$$

(b) *Logistic regression product prior* (Stingo et al. (2010))

$$\pi(\gamma \mid \boldsymbol{\theta}) = \prod_{i=1}^{p} \left( \frac{\exp(\boldsymbol{Z}_i'\boldsymbol{\theta})}{1+\exp(\boldsymbol{Z}_i'\boldsymbol{\theta})} \right)^{\gamma_i} \left( \frac{1}{1+\exp(\boldsymbol{Z}_i'\boldsymbol{\theta})} \right)^{1-\gamma_i}$$

(c) *Markov random field prior* (Li and Zhang (2010))

$$\pi(\gamma \mid \boldsymbol{\theta}) = \exp\left[\boldsymbol{\theta}_1'\gamma + \gamma'\boldsymbol{\theta}_2\gamma - \psi(\boldsymbol{\theta}_1, \boldsymbol{\theta}_2)\right]$$

# Simulated Example with Structured Covariates

- Assume $p = 99$ covariates cluster within three non-overlapping groups: $\boldsymbol{Z} = [\boldsymbol{z}_1, \boldsymbol{z}_2, \boldsymbol{z}_3]$, where $z_{ij} = \mathbb{I}_{[33\,(i-1)+1;\,33\,i]}(j)$
- Within group correlation 0.8, between group correlation 0
- Responses $\boldsymbol{Y} \sim \mathrm{N}(\boldsymbol{X}\beta, \sigma^2 \mathrm{I}_n)$ with $\beta = 2 \times \mathbb{I}_{[1;33]}(i)$, $n = 100$ and $\sigma^2 = 5$
- EMVS with *(a) Beta-binomial prior, (b) logistic regression prior, (c) MRF prior*

  Settings for model exploration

  - $v_0 \in \{0.01 + k \times 0.5 : 0 \leq k \leq 10\}$
  - $v_1$ assigned prior (6) with $a = 0.5$ and $b = 250$

  Settings for model evaluation

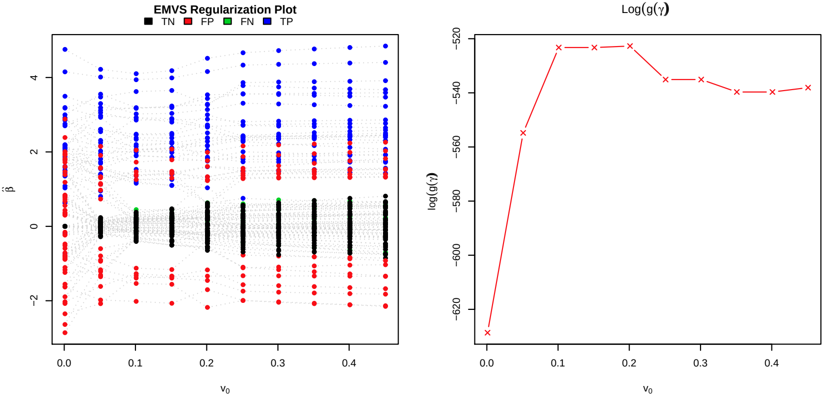
  - $v_1$ fixed to 1 000
  - Uniform beta-binomial for the $g$-function

# Simulated Example with Structured Covariates

(a) Beta-binomial prior, where $\theta \sim \mathrm{B}(1, 1)$

Best visited model:

25 *true positives* together with 11 *false negatives* and 8 *false positives*

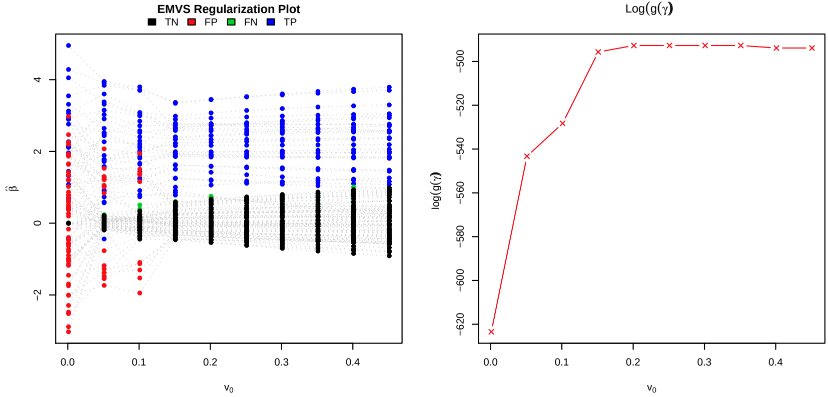


# Simulated Example with Structured Covariates

(b) Logistic regression prior $\boldsymbol{\theta} \sim \pi(\boldsymbol{\theta})$ in (3) with $a = b = 1$

Best visited model:

28 *true positives* together with 5 *false negatives* and 0 *false positives*

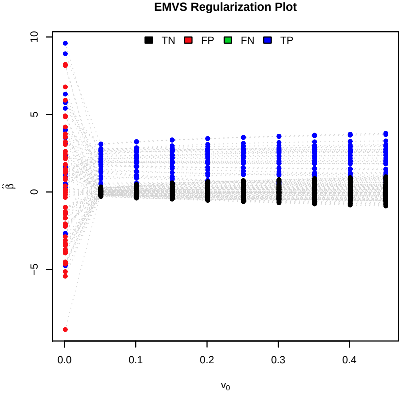


# Simulated Example with Structured Covariates

(c) MRF prior, $\theta$ fixed to the phase transition point, $\boldsymbol{\theta}_2 = (\mathbf{1}_{33\times 33} - \mathrm{I}_{33}) \otimes \mathrm{I}_3$

Best visited model:

33 *true positives* together with 0 *false negatives* and 0 *false positives*

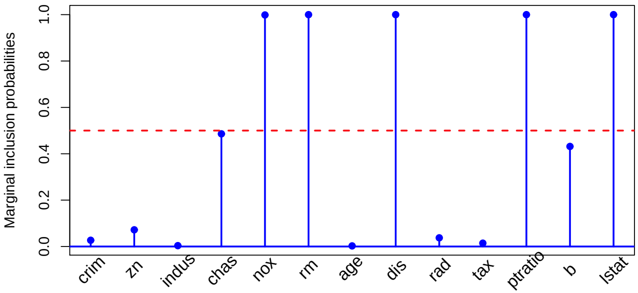


# Boston Housing Data: Application "$p < n$"

Predicting median price of homes in Boston on the basis of 13 predictors, $n = 506$

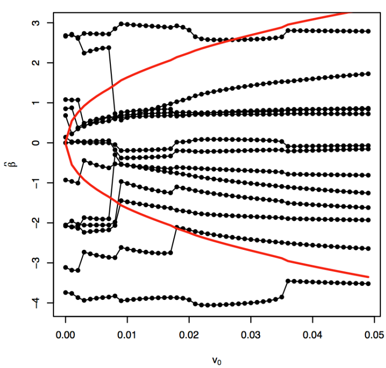
(1) Conjugate SSVS Gibbs sampler (CPU 21s)
$v_0 = 0.01, v_1 = 1\,000, \theta \sim B(1, 1)$, 10 000 iterations
Median probability model includes $\{5, 6, 8, 11, 13\}$

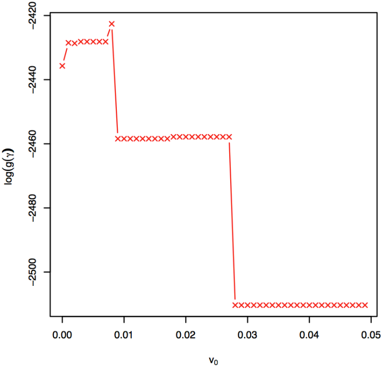
(2) Exhaustive evaluation of posterior model probabilities (CPU 6s)
$v_0 = 0, v_1 = 1\,000$ and $\theta \sim B(1, 1)$



# Boston Housing Data: Application "$p < n$"

(3) EMVS algorithm (CPU 0.53s)
$v_1 = 1\,000$, $v_0 \in \{10^{-6} + k \times 0.001; 1 \leq k \leq 50\}$, $\theta \sim B(1,1)$
$\beta^{(0)} \sim \mathrm{N}_{13}(\mathbf{0}, 10 \times \mathrm{I}_{13})$, best model contains $\{5, 6, 8, 11, 13\}$



# Detecting DNA Binding Motifs Using EMVS

- Spellman (1998) describe a yeast experiment to identify TF binding sites associated with cell cycle
- 800 genes found to have a periodic expression pattern across 2 cell cycles
- Another 800 that do not show any differential pattern across time selected as a reference
- Response vector $\boldsymbol{Y} = (Y_1, \ldots, Y_{1\,600})'$ summarizes expression of 1 600 genes over time

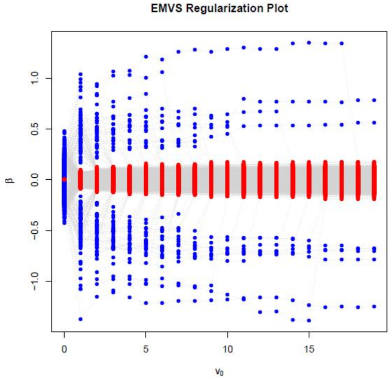
? *Can we explain the gene expression pattern by the occurrence of shared regulatory motifs?*

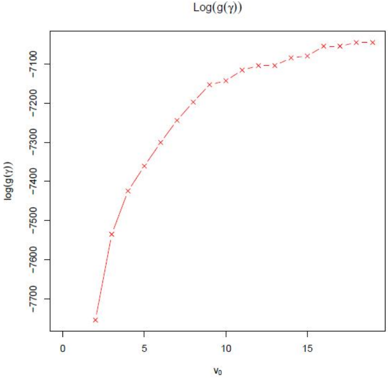
# Detecting DNA Binding Motifs Using EMVS

- Promoter regions of each genes screened for DNA motifs
- Motif $\equiv$ word of length 7 consisting of letters$\{A, G, T, C\}$ (altogether $4^7/2 = 8\,192$ motifs)
- Regression matrix $\boldsymbol{X}_{1\,600 \times 8\,192}$ contains counts of occurrences of each motif in the promoter region of each gene
- The motifs lie on a network with similar motifs being the neighbors
- For instance, *ACCTGTC* and *TCCTGTC* differ by only one letter $\rightsquigarrow$ they are connected on a graph
- Similar motifs assumed to attract the same TFs $\rightsquigarrow$ influence gene expression in a similar way

# (a) Beta-binomial Model

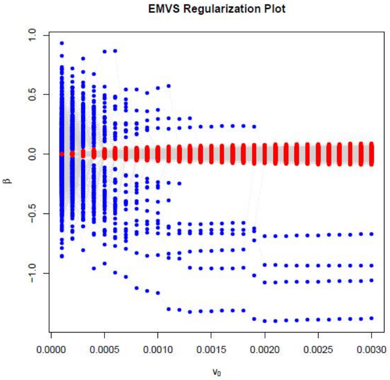
- $v_0 \in \{0.001 + k \times 1 : 0 \leq k \leq 20\}$
- $v_1$: random in EM with $a_{v_1} = 0.5$, $b_{v_1} = 250$, fixed to 1 000 in $g_0(\cdot)$
- $\beta^{(0)}$ according to (5) with $v_0 = 1$ and $v_1 = 1\,000$

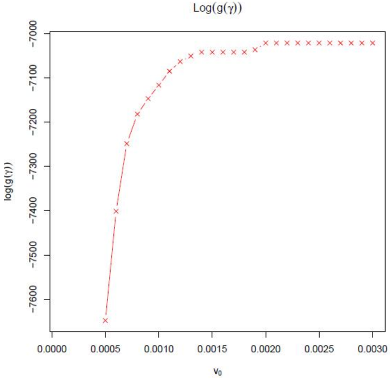


# (b) MRF Model

- $v_0 \in \{10^{-5} + k \times 10^{-5} : 0 \leq k \leq 30\}$
- $v_1$ and $\beta^0$ as in (a)
- Prior $\pi(\theta)$ located in the phase transition region



# Results

| 18 Selected Motifs | | 7 Selected Motifs | | |
|---|---|---|---|---|
| (a) | (b) | (a) | (b) | Known |
| *GACGCGT*[1] | *GACGCGT*[1] | *GACGCGT*[1] | *GACGCGT*[1] | × |
| *TACGCGT*[1] | *TACGCGT*[1] | | *TACGCGT*[1] | × |
| *TTCGCGT*[1] | *TTCGCGT*[1] | *TTCGCGT*[1] | *TTCGCGT*[1] | × |
| | *TTACGCG*[2] | | | |
| *TTTCGCG*[2] | *TTTCGCG*[2] | *TTTCGCG*[2] | *TTTCGCG*[2] | × |
| | *TGACGCG*[2] | | | |
| *TTAGCAG* | | | | |
| *ACGCGTT* | *ACGCGTT* | *ACGCGTT* | *ACGCGTT* | |
| *CCGCTTG* | *CCGCTTG* | | | |
| *CCGTCCT* | *CCGTCCT* | | | |
| *CGCGTTT* | *CGCGTTT* | *CGCGTTT* | *CGCGTTT* | |
| *CGTCCCT* | *CGTCCCT* | | | |
| *CTGATGG* | *CTGATGG* | | | |
| *GAATTAT* | *GAATTAT* | | | |
| *GACAGGT* | | | | |
| *GCCATTT* | *GCCATTT* | | | |
| | *GCGTTTT* | | | |
| *GGACGAT* | *GGACGAT* | *GGACGAT* | | × |
| *GTCCTCT* | | | | |
| *TACACAG* | *TACACAG* | | | × |
| *TTTATCG* | *TTTATCG* | *TTTATCG* | *TTTATCG* | |

Known motifs found in the SCPD (Sacharomyces Cerevisiae Pomoter Database)

# Summary

- We develop a rapid deterministic method based on EM algorithm as an alternative to stochastic model search
- Regularization diagram combined with rigorous model evaluation enable simultaneous exploration and evaluation of candidate models
- EMVS framework encompasses situations with structured covariates
- Heavy-tailed slab distributions can be considered to alleviate over-shrinkage
- Extensions to multivariate/factor analytic models possible (Rockova and Lesaffre (2013))

Erasmus MC
University Medical Center Rotterdam

Thank you!

Erasmus MC
University Medical Center Rotterdam

# References


Rockova, V. and George, E. (2012)
EMVS: The EM Approach to Bayesian Variable Selection
Under revision for Journal of the American Statistical Association

George, E. and McCulloch, R. (1993)
Variable Selection Via Gibbs Sampling
Journal of the American Statistical Association (88) 881-889

Li, F. and Zhang, N. R. (2010)
Bayesian Variable Selection in Structured High-dimensional Covariate Spaces with Applications in Genomics
Journal of the American Statistical Association (105) 1978-2002

Rockova, V. and Lesaffre, E. (2013)
Bayesian Sparse Factor Regression Approach to Genomic Data Integration
To appear in Proceedings of the 28th IWSM

Rockova, V. and Lesaffre, E. (2012)
Incorporating Grouping in Bayesian Variable Selection with Applications in Genomics
Under revision for Bayesian Analysis

Stingo, F., Chen, Y., Vannucci, M., Barrier, M., and Mirkes, P. (2010)
A Bayesian Graphical Modeling Approach to MicroRNA Regulatory Networks
Annals of Applied Statistics (4) 2024-2048


Erasmus MC
University Medical Center Rotterdam

# EMVS Algorithm: a Closer Look

Objective function:

$$Q\left(\boldsymbol{\beta},\boldsymbol{\theta},\sigma \mid \boldsymbol{\beta}^{(k)},\boldsymbol{\theta}^{(k)},\sigma^{(k)}\right) = C(\gamma) + Q_1\left(\boldsymbol{\beta},\sigma \mid \boldsymbol{\beta}^{(k)},\boldsymbol{\theta}^{(k)},\sigma^{(k)}\right) + Q_2\left(\boldsymbol{\theta} \mid \boldsymbol{\beta}^{(k)},\boldsymbol{\theta}^{(k)},\sigma^{(k)}\right),$$

where

$$Q_1\left(\boldsymbol{\beta},\sigma|\boldsymbol{\beta}^{(k)},\boldsymbol{\theta}^{(k)},\sigma^{(k)}\right) = -\frac{(\boldsymbol{Y}-\boldsymbol{X}\beta)'(\boldsymbol{Y}-\boldsymbol{X}\beta)}{2\sigma^2} - \frac{n+p+\nu}{2}\log(\sigma^2) - \frac{\nu\lambda}{2\sigma^2} - \frac{1}{2\sigma^2}\sum_{i=1}^{p}\beta_i^2\,\mathrm{E}_{\gamma|\cdot}\left[\frac{1}{v_0(1-\gamma_i)+v_1\gamma_i}\right],$$

$$Q_2\left(\boldsymbol{\theta}|\boldsymbol{\beta}^{(k)},\boldsymbol{\theta}^{(k)},\sigma^{(k)}\right) = \mathrm{E}_{\gamma|\cdot}\log\pi(\gamma|\boldsymbol{\theta}) + \log\pi(\boldsymbol{\theta}),$$

and $\mathrm{E}_{\gamma|\cdot}(\cdot)$ denotes the conditional expectation $\mathrm{E}_{\gamma|\boldsymbol{\beta}^{(k)},\boldsymbol{\theta}^{(k)},\sigma^{(k)},\boldsymbol{y}}(\cdot)$

Erasmus MC
University Medical Center Rotterdam

# EMVS Algorithm (Beta-binomial Case)

Objective function:

$$Q\left(\boldsymbol{\beta}, \boldsymbol{\theta}, \sigma \mid \boldsymbol{\beta}^{(k)}, \boldsymbol{\theta}^{(k)}, \sigma^{(k)}\right) = C(\gamma) + Q_1\left(\boldsymbol{\beta}, \sigma \mid \boldsymbol{\beta}^{(k)}, \boldsymbol{\theta}^{(k)}, \sigma^{(k)}\right) + Q_2\left(\boldsymbol{\theta} \mid \boldsymbol{\beta}^{(k)}, \boldsymbol{\theta}^{(k)}, \sigma^{(k)}\right),$$

where

$$Q_1\left(\boldsymbol{\beta}, \sigma | \boldsymbol{\beta}^{(k)}, \boldsymbol{\theta}^{(k)}, \sigma^{(k)}\right) = -\frac{(\boldsymbol{Y} - \boldsymbol{X}\beta)'(\boldsymbol{Y} - \boldsymbol{X}\beta)}{2\sigma^2} - \frac{n+p+\nu}{2}\log(\sigma^2) - \frac{\nu\lambda}{2\sigma^2} - \frac{1}{2\sigma^2}\sum_{i=1}^{p}\beta_i^2 \, \mathrm{E}_{\gamma|\cdot}\left[\frac{1}{v_0(1-\gamma_i) + v_1\gamma_i}\right],$$

$$Q_2\left(\boldsymbol{\theta} | \boldsymbol{\beta}^{(k)}, \boldsymbol{\theta}^{(k)}, \sigma^{(k)}\right) = \sum_{i=1}^{p}\log\left(\frac{\theta}{1-\theta}\right)\mathrm{E}_{\gamma|\cdot}\gamma_i + (a-1)\log\theta + (b+p-1)\log(1-\theta),$$

and $\mathrm{E}_{\gamma|\cdot}(\cdot)$ denotes the conditional expectation $\mathrm{E}_{\gamma|\boldsymbol{\beta}^{(k)},\boldsymbol{\theta}^{(k)},\sigma^{(k)},\boldsymbol{y}}(\cdot)$.

# E-step (Beta-binomial Case)

Variables $\gamma$ depend on the data $\boldsymbol{Y}$ only through the current estimates $\beta^{(k)}$.

We have:

$$(1) \quad \mathrm{E}_{\gamma|\cdot}\gamma_i = \mathrm{P}(\gamma_i = 1|\boldsymbol{\beta}^{(k)}, \theta^{(k)}, \sigma^{(k)}, \boldsymbol{y}) = \mathrm{P}(\gamma_i = 1|\boldsymbol{\beta}^{(k)}, \theta^{(k)}, \sigma^{(k)}) = p_i^\star,$$

where

$$p_i^\star = \frac{\pi(\beta_i^{(k)} \mid \sigma^{(k)}, \gamma_i = 1)\mathrm{P}(\gamma_i = 1 \mid \theta^{(k)})}{\pi(\beta_i^{(k)} \mid \sigma^{(k)}, \gamma_i = 1)\mathrm{P}(\gamma_i = 1 \mid \theta^{(k)}) + \pi(\beta_i^{(k)} \mid \sigma^{(k)}, \gamma_i = 0)\mathrm{P}(\gamma_i = 0 \mid \theta^{(k)})}$$

are the mixing proportions when fitting a Gaussian mixture model via EM.

$$(2) \quad \mathrm{E}_{\gamma|\cdot}\left[\frac{1}{v_0(1-\gamma_i) + v_1\gamma_i}\right] = \frac{\mathrm{E}_{\gamma|\cdot}(1-\gamma_i)}{v_0} + \frac{\mathrm{E}_{\gamma|\cdot}\gamma_i}{v_1} = \frac{1-p_i^\star}{v_0} + \frac{p_i^\star}{v_1} = d_i^\star$$

Erasmus MC
University Medical Center Rotterdam

# M-step (Beta-binomial Case)

(1) Update $\beta^{(k+1)}$ Closed form ridge regression solution

$$\beta^{(k+1)} = (\boldsymbol{X}'\boldsymbol{X} + \boldsymbol{D}^{\star})^{-1}\boldsymbol{X}'\boldsymbol{Y}, \quad D^{\star} = \mathrm{diag}\{d_1^{\star}, \ldots, d_p^{\star}\}$$

$\rightsquigarrow$ For $p > n$, use Woodbury-Sherman formula to get

$$\beta^{(k+1)} = \left[\boldsymbol{D}^{\star-1} - \boldsymbol{D}^{\star-1}\boldsymbol{X}'\left(\mathrm{I}_{n\times n} + \boldsymbol{X}\boldsymbol{D}^{\star-1}\boldsymbol{X}'\right)^{-1}\boldsymbol{X}\boldsymbol{D}^{\star-1}\right]\boldsymbol{X}'\boldsymbol{y}$$

(2) Update $\sigma^{(k+1)}$ Closed form

$$\sigma^{(k+1)} = \sqrt{\frac{|\boldsymbol{Y} - \boldsymbol{X}\beta^{(k+1)}|_{l_2} + |\boldsymbol{D}^{\star 1/2}\beta^{(k+1)}|_{l_2} + \eta\lambda}{n + p + \eta}}.$$

(2) Update $\theta^{(k+1)}$ Closed form

$$\theta^{(k+1)} = \frac{\sum_{i=1}^{p} p_i^* + a - 1}{a + b + p - 2}$$

# EMVS for Structured Priors

(a) *Logistic regression product prior*

$$Q_2\left(\boldsymbol{\theta}|\boldsymbol{\beta}^{(k)}, \boldsymbol{\theta}^{(k)}, \sigma^{(k)}\right) = \sum_{i=1}^{p}\{\boldsymbol{Z}_i'\boldsymbol{\theta}\, \mathrm{E}_{\gamma|\cdot}\gamma_i - \log[1+\exp(\boldsymbol{Z}_i'\boldsymbol{\theta})]\} + \sum_{j=1}^{q}\log\pi(\theta_j),$$

Beta distribution on the inverse logistic transformation of $\theta_j$

$$\pi(\theta_j) = \frac{1}{B(a,b)}\left[\frac{\exp(\theta_j)}{1+\exp(\theta_j)}\right]^a\left[\frac{1}{1+\exp(\theta_j)}\right]^b \tag{3}$$

(b) *MRF prior with $\boldsymbol{\theta}_1 = \theta(1,\dots,1)'$*, where $\theta \sim \pi(\theta)$ in (3)

$$Q_2(\theta\,|\,\boldsymbol{\beta}^{(k)}, \boldsymbol{\theta}^{(k)}, \sigma^{(k)}) = \theta\left(\sum_{i=1}^{p}\mathrm{E}_{\gamma|\cdot}\gamma_i + a\right) + \psi(\theta, \boldsymbol{\theta}_2) - (a+b)\log[1+\exp(\theta)].$$

- $\leadsto$ $\mathrm{E}_{\gamma|\cdot}\gamma_i$ complicated due to dependence between $\gamma_i'$s
- $\leadsto$ $\psi(\theta, \boldsymbol{\theta}_2)$ not in closed form

Erasmus MC
University Medical Center Rotterdam

# EMVS for Structured Priors

(a) *Logistic regression product prior*

$$Q_2\left(\boldsymbol{\theta}|\boldsymbol{\beta}^{(k)},\boldsymbol{\theta}^{(k)},\sigma^{(k)}\right)=\sum_{i=1}^{p}\left\{\boldsymbol{Z}_i'\boldsymbol{\theta}\,\mathrm{E}_{\boldsymbol{\gamma}|\cdot}\gamma_i-\log[1+\exp(\boldsymbol{Z}_i'\boldsymbol{\theta})]\right\}+\sum_{j=1}^{q}\log\pi(\theta_j),$$

Beta distribution on the inverse logistic transformation of $\theta_j$

$$\pi(\theta_j)=\frac{1}{B(a,b)}\left[\frac{\exp(\theta_j)}{1+\exp(\theta_j)}\right]^a\left[\frac{1}{1+\exp(\theta_j)}\right]^b \tag{3}$$

(b) *MRF prior with $\boldsymbol{\theta}_1=\theta(1,\ldots,1)'$*, where $\theta\sim\pi(\theta)$ in (3)

$$Q_2(\theta\,|\,\boldsymbol{\beta}^{(k)},\boldsymbol{\theta}^{(k)},\sigma^{(k)})=\theta\left(\sum_{i=1}^{p}\mathrm{E}_{\boldsymbol{\gamma}|\cdot}\gamma_i+a\right)+\psi(\theta,\boldsymbol{\theta}_2)-(a+b)\log[1+\exp(\theta)].$$

- $\rightsquigarrow$ $\mathrm{E}_{\boldsymbol{\gamma}|\cdot}\gamma_i$ complicated due to dependence between $\gamma_i's$
- $\rightsquigarrow$ $\psi(\theta,\boldsymbol{\theta}_2)$ not in closed form

Erasmus MC
University Medical Center Rotterdam

# Approximated E-step (MRF Prior)

Conditional posterior distribution $\pi(\gamma \mid \boldsymbol{\beta}^{(k)}, \theta^{(k)}, \sigma^{(k)}, \boldsymbol{y})$ proportional to

$$\exp\left[\left(\frac{1}{2}\log\left(v_0/v_1\right)\mathbf{1}' - \frac{v_0 - v_1}{2\sigma^{(k)2}v_1 v_0}\boldsymbol{\beta}^{(k)'}\mathrm{diag}\{\beta_i^{(k)}\}_{i=1}^{p} + \theta^{(k)}\mathbf{1}'\right)\gamma + \gamma'\boldsymbol{\theta}_2\gamma\right].$$

Markov random field distribution $\mathrm{MRF}(\boldsymbol{\theta}^{\star}, \boldsymbol{\theta}_2)$

- $\rightsquigarrow$ Expectation $\mathrm{E}(\gamma \mid \boldsymbol{\beta}^{(k)}, \theta^{(k)}, \sigma^{(k)}, \boldsymbol{y}) = \frac{\partial\psi(\boldsymbol{\theta},\boldsymbol{\theta}_2)}{\partial\boldsymbol{\theta}}|_{\boldsymbol{\theta}=\boldsymbol{\theta}^{\star}}$ not analytically tractable
- $\rightsquigarrow$ Mean field approximation $\rightsquigarrow$ iteratively solving

$$\widehat{\mu}_i = \frac{\exp(\theta_i^{\star} + \sum_{j\neq i}\theta_{ij}\widehat{\mu}_j)}{1 + \exp(\theta_i^{\star} + \sum_{j\neq i}\theta_{ij}\widehat{\mu}_j)}, \quad 1 \leq i \leq p$$

- $\rightsquigarrow$ We obtain approximations to the mixing proportions

$$\widehat{\mu}_i = \mathrm{P}(\gamma_i = 1 \mid \boldsymbol{\beta}^{(k)}, \theta^{(k)}, \sigma^{(k)}, \boldsymbol{y})$$

Erasmus MC
University Medical Center Rotterdam

# Approximated E-step (MRF Prior)

Conditional posterior distribution $\pi(\gamma \mid \boldsymbol{\beta}^{(k)}, \theta^{(k)}, \sigma^{(k)}, \boldsymbol{y})$ proportional to

$$\exp\left[\left(\frac{1}{2}\log\left(v_0/v_1\right)\mathbf{1}' - \frac{v_0 - v_1}{2\sigma^{(k)2} v_1 v_0}\boldsymbol{\beta}^{(k)'}\mathrm{diag}\{\beta_i^{(k)}\}_{i=1}^{p} + \theta^{(k)}\mathbf{1}'\right)\gamma + \gamma'\boldsymbol{\theta}_2\gamma\right].$$

Markov random field distribution $\mathrm{MRF}(\boldsymbol{\theta}^{\star}, \boldsymbol{\theta}_2)$

- ⇝ Expectation $\mathrm{E}(\gamma \mid \boldsymbol{\beta}^{(k)}, \theta^{(k)}, \sigma^{(k)}, \boldsymbol{y}) = \frac{\partial\psi(\boldsymbol{\theta},\boldsymbol{\theta}_2)}{\partial\boldsymbol{\theta}}|_{\boldsymbol{\theta}=\boldsymbol{\theta}^{\star}}$ not analytically tractable
- ⇝ Mean field approximation ⇝ iteratively solving

$$\widehat{\mu}_i = \frac{\exp(\theta_i^{\star} + \sum_{j\neq i}\theta_{ij}\widehat{\mu}_j)}{1 + \exp(\theta_i^{\star} + \sum_{j\neq i}\theta_{ij}\widehat{\mu}_j)}, \quad 1 \leq i \leq p$$

- ⇝ We obtain approximations to the mixing proportions

$$\widehat{\mu}_i = \mathrm{P}(\gamma_i = 1 \mid \boldsymbol{\beta}^{(k)}, \theta^{(k)}, \sigma^{(k)}, \boldsymbol{y})$$

Erasmus MC
University Medical Center Rotterdam

# Approximated M-step (MRF Prior)

- Updates for $\beta$ and $\sigma$ the same as before
- Update $\theta$

$$\theta^{(k+1)} = \text{argmax}_{\theta \in \mathbb{R}} \left\{ \theta \left( \sum_{i=1}^{p} \widehat{\mu}_i^\star + a \right) + \psi(\theta, \boldsymbol{\theta}_2) - (a+b) \log[1 + \exp(\theta)] \right\}.$$

Mean field approximation to the partition function $\psi(\theta, \boldsymbol{\theta}_2)$

$$\psi(\theta, \boldsymbol{\theta}_2) \approx \theta \sum_{i=1}^{p} \mu_i + \boldsymbol{\mu}' \boldsymbol{\theta}_2 \boldsymbol{\mu} - \psi^\star(\boldsymbol{\mu}), \qquad (4)$$

where $\mu_i = \mathrm{E}_{\theta, \boldsymbol{\theta}_2}(\gamma_i)$ and $\psi^\star(\cdot)$ denotes the conjugate dual function

$$\psi^\star(\boldsymbol{\mu}) = \sum_{i=1}^{p} [\mu_i \log \mu_i + (1 - \mu_i) \log(1 - \mu_i)]$$

Erasmus MC
University Medical Center Rotterdam

# Deterministic Annealing EM (DAEM)

DAEM algorithm to maximize (2)

(1) Logistic/beta-binomial prior:

Mixing proportions replaced by

$$p^{\star}_{i,t} = \frac{\pi(\beta_i^{(k)} \mid \sigma^{(k)}, \gamma_i = 1)^t \mathrm{P}(\gamma_i = 1 \mid \boldsymbol{b}^{(k)})^t}{\pi(\beta_i^{(k)} \mid \sigma^{(k)}, \gamma_i = 1)^t \mathrm{P}(\gamma_i = 1 \mid \boldsymbol{b}^{(k)})^t + \pi(\beta_i^{(k)} \mid \sigma^{(k)}, \gamma_i = 0)^t \mathrm{P}(\gamma_i = 0 \mid \boldsymbol{b}^{(k)})^t},$$

$\rightsquigarrow$ Limiting behavior $p^{\star}_{i,t} \to 0.5$ as $t \to 0$ suggests a starting vector

$$\widehat{\boldsymbol{\beta}}_{t=0} = \left[\boldsymbol{X}'\boldsymbol{X} + \frac{v_0 + v_1}{2 v_0 v_1} \boldsymbol{I}_p\right]^{-1} \boldsymbol{X}'\boldsymbol{y} \tag{5}$$

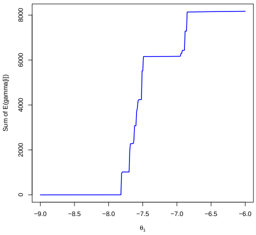
(2) MRF prior:

MF approximation to evaluate expectation of $\mathrm{MRF}(t\,\theta^{\star}, t\,\boldsymbol{\theta}_2)$

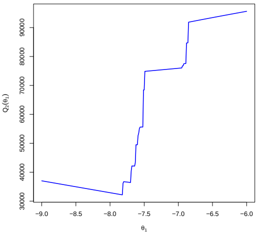
Erasmus MC
University Medical Center Rotterdam

# Phase Transition

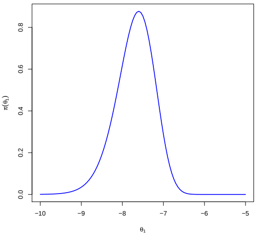
$\sum_{i=1}^{p} \mathrm{E}_{\theta,\boldsymbol{\theta}_2}\gamma_i$




$Q_2^{MRF}(\theta)$




$\pi(\theta)$




Erasmus MC
University Medical Center Rotterdam

# Heavy-tailed Alternative

- Gaussian slab density may overshrink $\rightsquigarrow$ put prior on $v_1$ to induce heavier tails
- Prior suggested in the "g-prior" context (Cui and George (2008), Maruyama and George (2011))

$$p(v_1) = \frac{v_1^b (1+v_1)^{-a-b-2}}{\mathrm{B}(a+1, b+1)} \mathrm{I}_{(0,\infty)}(v_1) \qquad (6)$$

- Implied marginal prior distribution

$$\pi(\beta_i | v_0, \sigma, \gamma) = (1-\gamma_i)\mathrm{N}(0, \sigma^2 v_0) + \gamma_i \widetilde{\pi}_{a,b,\sigma}(\beta_i),$$

where

$$\widetilde{\pi}_{a,b,\sigma}(\beta_i) \propto \frac{\exp\left(\frac{\beta_i^2}{4\sigma^2}\right)}{\sqrt{2\pi\sigma^2}} \left(\frac{\beta_i^2}{2\sigma^2}\right)^{\frac{b}{2}-\frac{1}{4}} W_{-a-\frac{b}{2}-\frac{5}{4}, -\frac{b}{2}-\frac{1}{4}} \left(\frac{\beta_i^2}{2\sigma^2}\right)$$

# Heavy-tailed Alternative

- Integrating out $v_1$ complicates the tractability of the M-step
- We treat $v_1$ as an additional unknown parameter
- Assuming prior (6) we update $v_1$ at the $k$-th iteration according to

$$v_1^{(k+1)} = \text{argmax}_{v_1} \left\{ -\frac{|\boldsymbol{P}^{\star 1/2}\boldsymbol{\beta}|_{l_2}}{2\sigma^{(k+1)}} \frac{1}{v_1} + \left( b - \sum_{i=1}^{p} \frac{p_i^\star}{2} \right) \log(v_1) - (a+b+2)\log(1+v_1) \right\},$$

where $\boldsymbol{P}^\star = \mathrm{diag}\{p_1^\star, \ldots, p_p^\star\}$

- EM algorithm remains unchanged, just with updates based on the current value $v_1^{(k)}$