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

Emmanuel Lesaffre\textsuperscript{1,2} \hspace{1cm} Veronika Ročková\textsuperscript{1}

\textsuperscript{1}Dept. of Biostatistics
Erasmus MC, Rotterdam, The Netherlands

\textsuperscript{2}L-BioStat
K.U. Leuven, Leuven, Belgium

Bayes 2013 Rotterdam
Outline

1. Introduction
2. Bayesian variable selection
3. BVS approaches
Outline

1. Introduction
2. Bayesian variable selection
3. BVS approaches
Classical variable selection

Two aims of variable selection: explanation and prediction

- Linear regression case: Prune model

  \[ y_i = \alpha + \sum_{k=1}^{d} \beta_k x_{ki} + \varepsilon_i, \quad (i = 1, \ldots, n) \]

- Formally: remove regressors for which \( \beta_k \) equal to zero
- Compromise between bias and variance
Classical variable selection

Two aims of variable selection: **explanation** and **prediction**

- **Linear regression case**: **Prune model**
  \[ y_i = \alpha + \sum_{k=1}^{d} \beta_k x_{ki} + \varepsilon_i, \quad (i = 1, \ldots, n) \]

- Formally: **remove** regressors for which \( \beta_k \) equal to zero
- Compromise between bias and variance
- Also referred to as **subset selection techniques**
- Focus on observational studies
Classical variable selection

Automated variable selection: all subsets and stepwise selection

- All subsets: challenging when $d$ large $\Rightarrow 2^d$ models
- Stepwise selection based on search algorithm & stopping criterion

Issues:
- No guarantee that best model is found
- No clear interpretation of significance of selected regressors
- Select one best model? Or base inference on many good models?
Classical variable selection

Automated variable selection: all subsets and stepwise selection

- **All subsets**: challenging when $d$ large $\Rightarrow 2^d$ models
- **Stepwise selection** based on search algorithm & stopping criterion

**Issues:**
- No guarantee that best model is found
- No clear interpretation of significance of selected regressors
- Select one best model? Or base inference on many good models?

**Alternative**: statistical model based on substantive knowledge

Often at least a(n initial) selection is needed (genomics, proteomics,...)
Bayesian variable selection (BVS)

- Bayesian variable selection based on:
  - Searching for most probable models (using model probability)
  - Parameter estimation rather than hypothesis testing

- Issues:
  - Partly the same as for classical variable selection
  - Computationally more demanding

- But: substantive knowledge can be implemented via the prior
Outline

1. Introduction
2. Bayesian variable selection
3. BVS approaches
Notation, concepts and principles of BVS

- **Model notation**: $K = 2^d$ models indexed by vectors $\gamma$
  - $\gamma = (\gamma_1, \ldots, \gamma_d)^T$: indicator vector of variables in model
  - $X_\gamma$: design matrix
  - $\beta_\gamma$: $d_\gamma$-dim regression vector
  - $\theta_\gamma$: all parameters of model
Notation, concepts and principles of BVS

- **Model notation**: $K = 2^d$ models indexed by vectors $\gamma$
  - $\gamma = (\gamma_1, \ldots, \gamma_d)^T$: indicator vector of variables in model
  - $X_\gamma$: design matrix
  - $\beta_\gamma$: $d_\gamma$-dim regression vector
  - $\theta_\gamma$: all parameters of model

- **Bayesian hierarchical model**:
  - Prior of model: $p(\gamma)$
  - Prior parameters: $p(\theta_\gamma \mid \gamma)$
  - Model: $p(y \mid \theta_\gamma, \gamma)$
General principle BVS

Computation of posterior model probabilities $p(\gamma \mid y)$:

$$p(\gamma \mid y) = \frac{p(y \mid \gamma)p(\gamma)}{\sum_{j=1}^{K} p(y \mid \gamma_j)p(\gamma_j)}$$

with

$$p(y \mid \gamma) = \int p(y \mid \theta_{\gamma}, \gamma)p(\theta_{\gamma} \mid \gamma) \, d\theta_{\gamma}$$
General principle BVS

Computation of posterior model probabilities $p(\gamma \mid y)$:

$$p(\gamma \mid y) = \frac{p(y \mid \gamma)p(\gamma)}{\sum_{j=1}^{K} p(y \mid \gamma_j)p(\gamma_j)}$$

with

$$p(y \mid \gamma) = \int p(y \mid \theta_{\gamma}, \gamma)p(\theta_{\gamma} \mid \gamma) \, d\theta_{\gamma}$$

Bayesian principle:

Pick model(s) with largest $p(\gamma \mid y)$
(maximum a posteriori (MAP) model)
Questions

1. What to take for prior probabilities $p(\gamma)$?

2. What priors for $p(\theta_\gamma | \gamma)$ ($p(\beta_\gamma | \gamma)$)?

3. For $K$ large: What search strategies can be implemented to quickly find the most promising models?
Model priors

- **Equal probabilities**: $p(\gamma) = 1/2^d$
  $\Rightarrow$ $d/2$-sized models are a priori preferred

- **Independence prior**: $p(\gamma \mid \pi) = \prod \pi^{d_\gamma} (1 - \pi)^{(d-d_\gamma)}$, ($\pi \in (0, 1)$)
  $\Rightarrow$ for $\pi$ small yields sparse models

- **Dependence prior**: $p(\gamma) = \frac{1}{d+1} \binom{d}{d_\gamma}^{-1}$
  $\Rightarrow$ uniform probability on size of model

...
Model priors

- **Equal probabilities**: \( p(\gamma) = 1/2^d \)
  \( \Rightarrow \) \( d/2 \)-sized models are a priori preferred

- **Independence prior**: \( p(\gamma | \pi) = \prod \pi^{d_\gamma} (1 - \pi)^{(d-d_\gamma)}, (\pi \in (0, 1)) \)
  \( \Rightarrow \) for \( \pi \) small yields sparse models

- **Dependence prior**: \( p(\gamma) = \frac{1}{d+1} \binom{d}{d_\gamma}^{-1} \)
  \( \Rightarrow \) uniform probability on size of model

- …

- Model prior **can steer the variable selection process** and be based on substantive knowledge (2nd part of talk)
Approaches

- $MC^3$: exploring the model space $\Rightarrow$ sampling $\gamma$
- Spike and slab:
  - exploring the parameter and model space $\Rightarrow$ sampling $\theta$ and $\gamma$
- Lasso: estimating $\theta$ (shrinking $\beta$)
Outline

1. Introduction

2. Bayesian variable selection

3. BVS approaches
   - Sampling model space
   - Sampling model and parameter space
   - Estimating the regression parameters
Given that $p(\gamma \mid y)$ (e.g. BIC approximation) has been computed:

- Sample in space of models
- Search for the best model(s)
- Result: chain $\gamma^{(1)}, \gamma^{(2)}, \ldots$
Given that $p(\gamma \mid y)$ (e.g. BIC approximation) has been computed:

- Sample in space of models
- Search for the best model(s)
- Result: chain $\gamma^{(1)}, \gamma^{(2)}, \ldots$

Rather model selection than variable selection

Possible if $p(\gamma \mid y)$ is easy/quick to compute and $d/K$ not too large

In second step $\theta$ must be sampled
**MC³** (Raftery et al. JASA 1997)

**Algorithm**

- Based on MCMC methods to sample from $p(\gamma \mid y)$

- **MC³**: Model Composition using MCMC
  - MH-algorithm on space of models
  - Sample $\gamma^*$ in neighborhood of $\gamma$ by
    $$q(\gamma^* \mid \gamma) = 1/d$$
  - Neighborhood: $\gamma$ and $\gamma^*$ differ in one position
  - MH acceptance probability:
    $$\min\left(1, \frac{p(\gamma^* \mid y)}{p(\gamma \mid y)}\right)$$
**SSVS** (George & McCulloch, 1993)

**Concept**

Exploration of $p(\beta, \sigma, \gamma \mid y)$:

- **Mitchell and Beauchamp (1988): spike and slab prior**
  - **Spike**: Dirac at 0 expressing $\beta_k = 0$
  - **Slab**: Uniform prior expressing $\beta_k \neq 0$
**SSVS** (George & McCulloch, 1993)

**Concept**

Exploration of \( p(\beta, \sigma, \gamma \mid y) \):

- **Mitchell and Beauchamp (1988):** spike and slab prior
  - **Spike:** Dirac at 0 expressing \( \beta_k = 0 \)
  - **Slab:** Uniform prior expressing \( \beta_k \neq 0 \)

- **George and McCulloch (1993):** SSVS
  - **Spike:** Normal around 0 with small variance expressing \( \beta_k = 0 \)
  - **Slab:** Normal around 0 with big variance expressing \( \beta_k \neq 0 \)

- **Result:** chain \( \beta^{(1)}, \sigma^{(1)}, \gamma^{(1)}, \beta^{(2)}, \sigma^{(2)}, \gamma^{(2)}, \ldots \)

- **Yields subchain:** \( \gamma^{(1)}, \gamma^{(2)}, \ldots \)
**Stochastic Search Variable Selection**

\[
\beta_k | \gamma_k, \mathbf{c}, \tau_k^2 \sim (1 - \gamma_k) \mathcal{N}(0, \tau_k^2) + \gamma_k \mathcal{N}(0, \tau_k^2 \mathbf{c}^2),
\]

\[
\gamma_k | \pi_k \sim \text{Bernoulli}(\pi_k)
\]

\[\Rightarrow \text{Variable not in the model} \quad \gamma_k = 0\]

\[\Rightarrow \text{Variable in the model} \quad \gamma_k = 1\]

\[\Rightarrow \text{Calibration of hyper-parameters } c, \tau_k^2 \text{ needed}\]
*BVS approaches*

**SSVS** (George & McCulloch, 1993)

**Algorithm**

Stochastic Search Variable Selection

\[ \beta_k | \gamma_k, c, \tau_k^2 \sim (1 - \gamma_k)N(0, \tau_k^2) + \gamma_k N(0, \tau_k^2 c^2), \]

\[ \gamma_k | \pi_k \sim \text{Bernoulli}(\pi_k) \]

\[ \Rightarrow \text{Variable not in the model} \quad \gamma_k = 0 \]

\[ \Rightarrow \text{Variable in the model} \quad \gamma_k = 1 \]

\[ \Rightarrow \text{Calibration of hyper-parameters } c, \tau_k^2 \text{ needed} \]
SSVS (George & McCulloch, 1993)

Algorithm

Stochastic Search Variable Selection

\[
\begin{align*}
\beta_k | \gamma_k, c, \tau_k^2 &\sim (1 - \gamma_k) N(0, \tau_k^2) + \gamma_k N(0, \tau_k^2 c^2), \\
\gamma_k | \pi_k &\sim \text{Bernoulli}(\pi_k)
\end{align*}
\]

⇝ Variable not in the model
\(\gamma_k = 0\)

⇝ Variable in the model
\(\gamma_k = 1\)

⇝ Calibration of hyper-parameters \(c, \tau_k^2\) needed
SSVS (George & McCulloch, 1993)

Inference for variable selection

- **Highest posterior model (HPM):** Select a model that has been visited most often
**SSVS** (George & McCulloch, 1993)

Inference for variable selection

- **Highest posterior model (HPM):**
  Select a model that has been visited most often

- **Median probability model (MPM):**
  Select variables that appear at least in 50% of visited models
SSVS (George & McCulloch, 1993)

Inference for variable selection

- **Highest posterior model (HPM):** Select a model that has been visited most often.
- **Median probability model (MPM):** Select variables that appear at least in 50% of visited models.
- **Hard shrinkage:** Select variables with $p(\beta_k \mid y)$ “spread far from zero.”
SSVS (George & McCulloch, 1993)

Alternative spike and slab models

- Popular approach in genomic research

- Variants:
  - Conjugate version:
    \[
    \beta_k | \gamma_k, c, \tau_k^2 \sim (1 - \gamma_k)N(0, \sigma^2 \tau_k^2) + \gamma_k N(0, \sigma^2 \tau_k c^2)
    \]
  - SSVS2: spike normal replaced by Dirac
  - NMIG: Normal mixture of inverse gammas (Ishrawan & Rao, 2005)
  - ...
Alternative BVS approaches

- Reversible Jump MCMC (RJMCMC)
- Combinations of SSVS, $MC^3$, RJMCMC, etc.
- ...
Alternative BVS approaches

- Reversible Jump MCMC (RJMCMC)
- Combinations of SSVS, $MC^3$, RJMCMC, etc.
- ... 

MCMC-based approaches are computationally involved
Especially when $d >> n$ as e.g. in genomics
Bayesian lasso \textsuperscript{(Park & Casella, 2008)}

Concept

Classical lasso:

- Minimize

\[
(y - X\beta)^T(y - X\beta) + \lambda \sum_{k=1}^{d} |\beta_k|
\]

- Differential shrinkage of the regression coefficients: some regression coefficients put to zero for $\lambda$ large
Bayesian lasso (Park & Casella, 2008)

Concept

Classical lasso:

- Minimize

\[(y - X\beta)^T(y - X\beta) + \lambda \sum_{k=1}^{d} |\beta_k|\]

- Differential shrinkage of the regression coefficients: some regression coefficients put to zero for \(\lambda\) large

⇒ Do not select variables, but \textit{shrink unimportant variables to zero}
Bayesian lasso (Park & Casella, 2008)

Concept

Classical lasso:

- Minimize

\[(y - X\beta)^T(y - X\beta) + \lambda \sum_{k=1}^{d} |\beta_k|\]

- Differential shrinkage of the regression coefficients: some regression coefficients put to zero for \(\lambda\) large

⇒ Do not select variables, but shrink unimportant variables to zero

Bayesian lasso: take Laplace prior

\[p(\beta) = \prod_{k=1}^{d} \frac{\lambda}{2} e^{-\lambda|\beta_k|}\]
Bayesian lasso (Park & Casella, 2008)

Hierarchical representation

- **Take conditional** Laplace prior for regression coefficients

\[
p(\beta | \sigma^2) = \prod_{k=1}^{d} \frac{\lambda}{2\sigma} e^{-\lambda |\beta_k|/\sigma}
\]

- **Hierarchical representation of prior structure:**

\[
\begin{align*}
\beta_k | \sigma_{\beta_k}^2 & \sim N(0, \sigma_{\beta_k}^2), \ (k = 1, \ldots, d) \\
\sigma_{\beta_k}^2 & = \sigma^2 \tau_k^2 \\
\tau_k^2 & \sim \frac{\lambda^2}{2} e^{-\lambda^2 \tau_k^2/2}, \ (k = 1, \ldots, d) \\
\sigma^2 & \sim p(\sigma^2)
\end{align*}
\]
Bayesian lasso (Park & Casella, 2008)

Variations

Classical and Bayesian lasso:

- **Adaptive lasso**: more differential shrinkage
- **Fused lasso**: regressors have natural ordering
- **Grouped lasso**: take grouping of regressors into account
- **Elastic net**: compromise between lasso and ridge
- **Adaptive elastic net**: adaptive version of elastic net
- ...
End part I
The many regressors case

When $d >> n$:

- Most methods break down
- Many ad hoc combinations of existing approaches have been suggested
- **Still computationally prohibitive**