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

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

[1]Dept. of Biostatistics
Erasmus MC, Rotterdam, The Netherlands

[2]L-BioStat
K.U. Leuven, Leuven, Belgium

Bayes 2013 Rotterdam

Outline

Outline

Classical variable selection

Two aims of variable selection: explanation and prediction

- Linear regression case: Prune model

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

- Formally: remove regressors for which β_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

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

- Formally: remove regressors for which β_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

Notation, concepts and principles of BVS

- Model notation: $K = 2^d$ models indexed by vectors γ
  - $\gamma = (\gamma_1, \ldots, \gamma_d)^T$: indicator vector of variables in model
  - $\boldsymbol{X}_\gamma$: design matrix
  - $\boldsymbol{\beta}_\gamma$: d_γ-dim regression vector
  - $\boldsymbol{\theta}_\gamma$: all parameters of model

Notation, concepts and principles of BVS

- Model notation: $K = 2^d$ models indexed by vectors γ
  - $\gamma = (\gamma_1, \ldots, \gamma_d)^T$: indicator vector of variables in model
  - $\boldsymbol{X}_\gamma$: design matrix
  - $\boldsymbol{\beta}_\gamma$: d_γ-dim regression vector
  - $\boldsymbol{\theta}_\gamma$: all parameters of model
- Bayesian hierarchical model:
  - Prior of model: $p(\gamma)$
  - Prior parameters: $p(\boldsymbol{\theta}_\gamma \mid \gamma)$
  - Model: $p(\boldsymbol{y} \mid \boldsymbol{\theta}_\gamma, \gamma)$

General principle BVS

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

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

with

$$p(\boldsymbol{y} \mid \gamma) = \int p(\boldsymbol{y} \mid \boldsymbol{\theta}_\gamma, \gamma)p(\boldsymbol{\theta}_\gamma \mid \gamma)\, d\boldsymbol{\theta}_\gamma$$

General principle BVS

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

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

with

$$p(\boldsymbol{y} \mid \gamma) = \int p(\boldsymbol{y} \mid \boldsymbol{\theta}_\gamma, \gamma)p(\boldsymbol{\theta}_\gamma \mid \gamma)\, d\boldsymbol{\theta}_\gamma$$

Bayesian principle:

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

Questions

1. What to take for prior probabilities $p(\gamma)$?
2. What priors for $p(\boldsymbol{\theta}_\gamma \mid \gamma)$ ($p(\boldsymbol{\beta}_\gamma \mid \gamma)$)?
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 π 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: $p(\gamma \mid \pi) = \prod \pi^{d_\gamma}(1-\pi)^{(d-d_\gamma)}, \ (\pi \in (0,1))$
  $\Rightarrow$ for π 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 γ
- Spike and slab:
  exploring the parameter and model space $\Rightarrow$ sampling $\boldsymbol{\theta}$ and γ
- Lasso: estimating $\boldsymbol{\theta}$ (shrinking $\boldsymbol{\beta}$)

Outline

MC^3 (Raftery et al. JASA 1997)

Concept

Given that $p(\gamma \mid \boldsymbol{y})$ (e.g. BIC approximation) has been computed:

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

MC^3 (Raftery et al. JASA 1997)

Concept

Given that $p(\gamma \mid \boldsymbol{y})$ (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 $p(\gamma \mid \boldsymbol{y})$ is easy/quick to compute and d/K not too large
- In second step θ must be sampled

MC^3 (Raftery et al. JASA 1997)

Algorithm

- Based on MCMC methods to sample from $p(\gamma \mid \boldsymbol{y})$
- MC^3: Model Composition using MCMC
  - MH-algorithm on space of models
  - Sample γ^* in neighborhood of γ by

$$q(\gamma^* \mid \gamma) = 1/d$$

  - Neighborhood: γ and γ^* differ in one position
  - MH acceptance probability:

$$\min\left(1, \frac{p(\gamma^* \mid \boldsymbol{y})}{p(\gamma \mid \boldsymbol{y})}\right)$$

SSVS (George & McCulloch, 1993)

Concept

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

- 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 $p(\beta, \sigma, \gamma \mid \boldsymbol{y})$:

- 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: $\gamma^{(1)}, \gamma^{(2)}, \ldots$

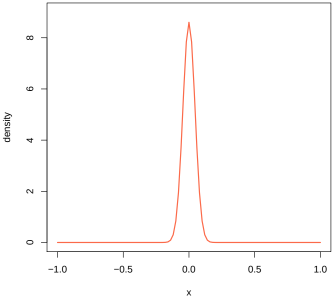
SSVS (George & McCulloch, 1993)

Algorithm

Stochastic Search Variable Selection

$$\beta_k|\gamma_k, c, \tau_k^2 \sim (1-\gamma_k)\mathrm{N}(0,\tau_k^2) + \gamma_k \mathrm{N}(0,\tau_k^2 c^2),$$
$$\gamma_k|\pi_k \sim \mathrm{Bernoulli}(\pi_k)$$

SPIKE



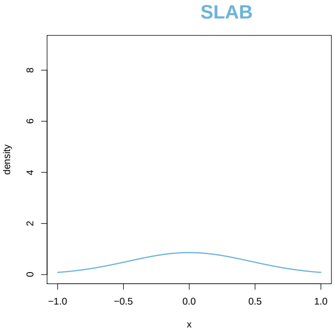
- Variable not in the model
  $\gamma_k = 0$
- Variable in the model
  $\gamma_k = 1$
- Calibration of hyper-parameters c, τ_k^2 needed

SSVS (George & McCulloch, 1993)

Algorithm

Stochastic Search Variable Selection

$$\beta_k|\gamma_k, c, \tau_k^2 \sim (1-\gamma_k)\mathrm{N}(0,\tau_k^2) + \gamma_k \mathrm{N}(0,\tau_k^2 c^2),$$
$$\gamma_k|\pi_k \sim \mathrm{Bernoulli}(\pi_k)$$



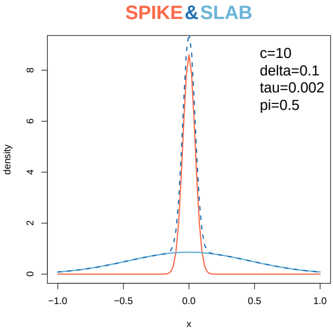
- Variable not in the model
  $$\gamma_k = 0$$
- Variable in the model
  $$\gamma_k = 1$$
- Calibration of hyper-parameters c, τ_k^2 needed

SSVS (George & McCulloch, 1993)

Algorithm

Stochastic Search Variable Selection

$$\beta_k|\gamma_k, c, \tau_k^2 \sim (1-\gamma_k)\mathrm{N}(0, \tau_k^2) + \gamma_k \mathrm{N}(0, \tau_k^2 c^2),$$
$$\gamma_k|\pi_k \sim \mathrm{Bernoulli}(\pi_k)$$



- Variable not in the model
  $\gamma_k = 0$
- Variable in the model
  $\gamma_k = 1$
- Calibration of hyper-parameters c, τ_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 $p(\beta_k \mid \boldsymbol{y})$ "spread far from zero"

SSVS (George & McCulloch, 1993)

Alternative spike and slab models

- Popular approach in genomic research
- Variants:
  - Conjugate version:

    $$\beta_k|\gamma_k, c, \tau_k^2 \sim (1-\gamma_k)\mathrm{N}(0, \sigma^2\tau_k^2) + \gamma_k\mathrm{N}(0, \sigma^2\tau_k^2 c^2)$$
  - 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 (Park & Casella, 2008)

Concept

Classical lasso:

- Minimize

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

- Differential shrinkage of the regression coefficients: some regression coefficients put to zero for λ large

Bayesian lasso (Park & Casella, 2008)

Concept

Classical lasso:

- Minimize

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

- Differential shrinkage of the regression coefficients: some regression coefficients put to zero for λ large

$\Rightarrow$ Do not select variables, but **shrink unimportant variables to zero**

Bayesian lasso (Park & Casella, 2008)

Concept

Classical lasso:

- Minimize

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

- Differential shrinkage of the regression coefficients: some regression coefficients put to zero for λ large

$\Rightarrow$ 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

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

- Hierarchical representation of prior structure:

$$\begin{aligned}
\beta_k \mid \sigma^2_{\beta_k} &\sim \mathrm{N}(0, \sigma^2_{\beta_k}),\ (k = 1, \ldots, d) \\
\sigma^2_{\beta_k} &= \sigma^2 \tau^2_k \\
\tau^2_k &\sim \frac{\lambda^2}{2} e^{-\lambda^2 \tau^2_k/2},\ (k = 1, \ldots, d) \\
\sigma^2 &\sim p(\sigma^2)
\end{aligned}$$

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**