Bayesian Variable Selection

ISA SHORT COURSES
"MCMC, WinBUGS and Bayesian Model Selection"
5–6 December 2011

Ioannis Ntzoufras
Associate Professor
Department of Statistics
Athens University of Economics and Business

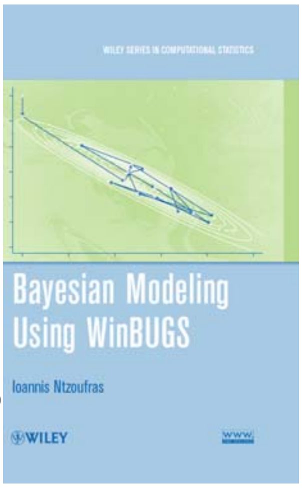
Bayesian Variable Selection Tutorial

The following presentation is based on chapter 11 of my book

"**Bayesian Modeling Using WinBUGS**"

Material extra to this book/chapter will be highlighted with an asterisk on the header

- It is introductory
- There are some tricks of how to use WinBUGS for variable selection
- It is not exhaustive (since there are a lot of methods around the last years)
- Priors are not reviewed thoroughly (difficult and important subject with a lot of ongoing research)



Bayesian Variable Selection Tutorial

table of contents (1)

Bayesian Variable Selection Tutorial

table of contents (2)

Bayesian Variable Selection Tutorial
Introduction

What is Model Selection?

- Evaluation of performance of scientific scenarios and
- Selection of the 'best'.

'Best' Model?

- The 'best' performed model is totally subjective
- Different procedures (or scientists) support different scientific theories, scenarios and models.

6/12/2011@ University College Dublin
ISA short courses
Ioannis Ntzoufras
Bayesian Variable Selection – An Introductory Tutorial
5

Bayesian Variable Selection Tutorial
Introduction

Two **MAJOR** principles:

1. *Goodness of Fit*
   How close is theory [model] to reality [data]
2. *Parsimony*
   Simplicity of theory;
   In stats: Economy in parameters.

6/12/2011@ University College Dublin
ISA short courses
Ioannis Ntzoufras
Bayesian Variable Selection – An Introductory Tutorial
6

Bayesian Variable Selection Tutorial
Introduction

Available Model/Variable Selection Methods

- ***Classical Model Selection:*** based on Significance tests and stepwise model search methods (Forward Strategy, Backward Elimination, Stepwise Procedures)
- ***Bayesian Model Selection/Comparison***
  - Posterior odds and model probabilities – BMA – BIC
  - Utility measures
  - Predictive measures
  - Deviance Information Criterion (DIC)
- ***Information Criteria***: BIC, AIC, other.

Bayesian Variable Selection Tutorial
Introduction

Disadvantages of Classical Stepwise Procedures

- Large datasets ⇒ small p-values even if the hypothesized model is plausible.
- Stepwise methods are sequential application of simple significance tests ⇒ Exact significance level cannot be calculated (Freedman, 1983, Am.Stat.).
- The maximum *F*-to-enter statistic 'is not even remotely like an F-distribution' (Miller, 1984, JRSSA).
- ***The selection of a single model ignores model uncertainty*** *(This is avoided in Bayesian theory via the Bayesian Model Averaging – BMA)*
- We can ***compare only nested models***.
- Different procedures or starting from different models ⇒ Different selected models. (stepwise procedures are sub-optimal)

Bayesian Variable Selection Tutorial

1. Posterior model odds and Bayes factors

Comparison of models m_1 and m_2 (or hypotheses H_1 and H_2) is performed via the *posterior model probabilities* $f(m_k|\mathbf{y})$ and their corresponding ratio

$$PO_{12} = \underbrace{\frac{f(m_1|\mathbf{y})}{f(m_2|\mathbf{y})}}_{\text{PO}_{12}\text{: Posterior model odds of model } m_1 \text{ vs. } m_2} = \underbrace{\frac{f(\mathbf{y}|m_1)}{f(\mathbf{y}|m_2)}}_{\text{B}_{12}\text{: Bayes Factor of model } m_1 \text{ vs. } m_2} \times \underbrace{\frac{f(m_1)}{f(m_2)}}_{\text{Prior Model Odds of } m_1 \text{ vs. } m_2}$$

6/12/2011@ University College Dublin
ISA short courses
Ioannis Ntzoufras
Bayesian Variable Selection – An Introductory Tutorial
9

Bayesian Variable Selection Tutorial

1. Posterior model odds and Bayes factors

- *m* Model indicator of model *m*
- *f(m)*: Prior model probability of *m*
- *f(m | **y**)*: Posterior model probability of *m*
- *f(**y** | m)*: Marginal likelihood of model *m* (or prior predictive distribution of model *m*) given by

$$f(\boldsymbol{y}|m) = \int f(\boldsymbol{y}|\boldsymbol{\theta}_m, m) f(\boldsymbol{\theta}_m|m) d\boldsymbol{\theta}_m$$

6/12/2011@ University College Dublin
ISA short courses
Ioannis Ntzoufras
Bayesian Variable Selection – An Introductory Tutorial
10

Bayesian Variable Selection Tutorial

1. Posterior model odds and Bayes factors

Marginal likelihood of model m

$$f(\boldsymbol{y}|m) = \int f(\boldsymbol{y}|\boldsymbol{\theta}_m, m) f(\boldsymbol{\theta}_m|m) d\boldsymbol{\theta}_m$$

Likelihood — Prior under model m

$\boldsymbol{\theta}_m$: Parameter vector of model m

THE ABOVE INTERGAL:

- Is analytically available when conjugate priors are used
- Computation is hard in 99,9% of the remaining cases

Bayesian Variable Selection Tutorial

1. Posterior model odds and Bayes factors

Table 11.1 Bayes factor interpretation according to Kass and Raftery (1995)

$\log(B_{10})$	B_{10}	Evidence against H_0
0 – 1	1 – 3	Negligible
1 – 3	3 – 20	Positive
3 – 5	20 – 150	Strong
> 5	> 150	Very strong

Bayesian Variable Selection Tutorial

1. Posterior model odds and Bayes factors

Bayesian Model Averaging

- Do not select a single model but a group of 'good' models (or all)
- *Incorporate uncertainty by weighting inferences* by their posterior model probabilities
  - Adjust predictions (and inference) according to the observed model uncertainty.
  - Average over all conditional model specific posterior distributions weighted by their posterior model probabilities.
- Base predictions on all models under consideration (or a group of good models) and therefore account for model uncertainty.
- The predictive distribution of a quantity Δ is given by

$$f(\Delta|\underline{y}) = \sum_{m \in \mathcal{M}} f(\Delta|m, \underline{y}) f(m|\underline{y})$$

6/12/2011@ University College Dublin
ISA short courses
Ioannis Ntzoufras
Bayesian Variable Selection – An Introductory Tutorial
13

Bayesian Variable Selection Tutorial

1. Posterior model odds and Bayes factors

Bayesian Model Averaging

- Reviews on Bayesian model averaging
  - Hoeting *et al.* (1999, *Stat.Science*)
  - Wasserman (2000, *J.Math.Psych.*)
- BMA has better predictive ability evaluated by the logarithmic scoring rule

  [Madigan and Raftery (1994, *JASA*), Kass and Raftery (1995, *JASA*) and Raftery *et al.* (1997, *JASA*)]
- Used frequently by Econometricians for prediction.

6/12/2011@ University College Dublin
ISA short courses
Ioannis Ntzoufras
Bayesian Variable Selection – An Introductory Tutorial
14

Bayesian Variable Selection Tutorial

1. Posterior model odds and Bayes factors

GOOD NEWS

Advantages of Bayesian methods

- Efficient Model Search via MCMC methods
- Automatic selection of the 'best' model (after specifying the model and the method of estimation)
- Posterior model probabilities are comparable across models and have a more straightforward interpretation
- Allows for model uncertainty via selecting a class of 'good' models with close posterior model probabilities
- Can compare non-nested models

Bayesian Variable Selection Tutorial

1. Posterior model odds and Bayes factors

BAD NEWS

Main Disadvantage of Bayesian methods

- Sensitivity of posterior model probabilities and Bayes factors on prior (Lindley-Bartlett Paradox).

  [a lot of ongoing research on this area]

Other disadvantages of Bayesian methods

- Computation of marginal likelihood is hard (but feasible)
- Model search may be demanding computationally especially when the model space is large
- Setting up an algorithm for the above is a **PAPER** and sometimes a good one.

Bayesian Variable Selection Tutorial

2. The Lindley – Bartlett Paradox

Let us consider the comparison of Lindley (1957, Bka).

H_0: $Y_i \sim N(\theta_0, \sigma^2)$, with θ_0, σ^2 known

versus

H_1: $Y_i \sim N(\theta \neq \theta_0, \sigma^2)$, with σ^2 known and θ unknown to be estimated.

m_0 (model under H_0) does not have any parameters!

m_1(model under H_1) has θ parameter!

PRIOR: $\theta | m_1 \sim N(\theta_0, \sigma_\theta^2)$

Bayesian Variable Selection Tutorial

2. The Lindley – Bartlett Paradox

$$f(\boldsymbol{y}|m_0) = \left(2\pi\sigma^2\right)^{-n/2} \exp\left(-\frac{1}{2\sigma^2}\sum_{i=1}^{n}(y_i-\theta_0)^2\right)$$

$$f(\boldsymbol{y}|m_1) = \int \left(2\pi\sigma^2\right)^{-n/2} e^{-\frac{1}{2\sigma^2}\sum\limits_{i=1}^{n}(y_i-\theta)^2} \left(2\pi\sigma_\theta^2\right)^{-1/2} e^{-\frac{1}{2\sigma_\theta^2}(\theta-\theta_0)^2} d\theta$$

$$= \left(2\pi\sigma^2\right)^{-n/2} \left(\frac{\sigma^2}{n\sigma_\theta^2+\sigma^2}\right)^{1/2} \exp\left\{\frac{1}{2\sigma^2}\left[\sum_{i=1}^{n}(y_i-\overline{y})^2 + \frac{n(\overline{y}-\theta_0)^2}{1+n\sigma_\theta^2/\sigma^2}\right]\right\}$$

Bayesian Variable Selection Tutorial

2. The Lindley – Bartlett Paradox

$$PO_{01} = \frac{f(H_0)}{f(H_1)}\sqrt{1+n\frac{\sigma_\theta^2}{\sigma^2}}\exp\left\{-\frac{1}{2\sigma^2}\left[\sum_{i=1}^{n}(y_i-\theta_0)^2-\sum_{i=1}^{n}(y_i-\overline{y})^2-\frac{n(\overline{y}-\theta_0)^2}{1+n\sigma_\theta^2/\sigma^2}\right]\right\}$$

Lindley considered samples at the border of significance for α=q ⇒ $\overline{y}=\theta_0 \pm z_{q/2}\sigma/\sqrt{n}$.

Then

$$PO_{01} = \frac{f(H_0)}{f(H_1)}\sqrt{1+n\frac{\sigma_\theta^2}{\sigma^2}}\exp\left\{-\frac{1}{2}\frac{n\sigma_\theta^2}{n\sigma_\theta^2+\sigma^2}Z_{q/2}^2\right\}$$

Bayesian Variable Selection Tutorial

2. The Lindley – Bartlett Paradox

Posterior odds at the limit of significance for a=q

$$PO_{01} = \frac{f(H_0)}{f(H_1)}\sqrt{1+n\frac{\sigma_\theta^2}{\sigma^2}}\exp\left\{-\frac{1}{2}\frac{n\sigma_\theta^2}{n\sigma_\theta^2+\sigma^2}Z_{q/2}^2\right\}$$

- *It is the posterior model odds when classical significance tests cannot decide.*
- ***Depends on n**: for $n \rightarrow \infty$, $PO_{01} \rightarrow \infty$*
- ***Depends on prior variance σ_θ^2**: for $\sigma_\theta^2 \rightarrow \infty$, $PO_{01} \rightarrow \infty$* [Bartlett, 1957, bka]
- In both the above cases, Bayesian methods support the simplest model while classical methods cannot decide

Bayesian Variable Selection Tutorial

2. The Lindley – Bartlett Paradox

$$PO_{01} = \frac{f(H_0)}{f(H_1)}\sqrt{1+n\frac{\sigma_\theta^2}{\sigma^2}}\exp\left\{-\frac{1}{2\sigma^2}\left[\sum_{i=1}^{n}(y_i-\theta_0)^2-\sum_{i=1}^{n}(y_i-\overline{y})^2-\frac{n(\overline{y}-\theta_0)^2}{1+n\sigma_\theta^2/\sigma^2}\right]\right\}$$

The same behavior is true for the general PO

- ***Depends on n:*** *for $n \to \infty$, $PO_{01} \to \infty$ (support H_0)*
- ***Depends on prior variance σ_θ^2:*** *for $\sigma_\theta^2 \to \infty$, $PO_{01} \to \infty$*
- While ***classical methods*** for $n \to \infty$, significance tests reject the simplest hypothesis H_0
- The term is used for any case where classical and Bayesian methods support different models or hypotheses.

Bayesian Variable Selection Tutorial

2. The Lindley – Bartlett Paradox

- The sensitivity on sample size n can be eliminated by setting prior variance to depend on n i.e. use σ_0^2/n instead of σ_0^2.
- The specification of σ_0^2 remains hard since in non-informative cases
  - must be large to avoid prior bias within each model and
  - Not large enough to activate the Lindley-Bartlett paradox and fully support the simplest model.
- The same problem appears in any model selection problem and it is more evident in nested model comparisons.

Bayesian Variable Selection Tutorial

2. The Lindley – Bartlett Paradox

As an extension of this behavior

improper priors cannot be used

since the Bayes factor will depend on the ratio of the undetermined normalizing constants

For an improper prior $\pi(\theta)$

Actual prior $\Rightarrow f(\boldsymbol{\theta}_m|m) = C_m \pi(\theta) \propto \pi(\theta)$

$$B_{01} = \frac{C_{m_0}}{C_{m_1}} \times \frac{\int f(\boldsymbol{y}|\boldsymbol{\theta}_{m_0}, m_0)\pi(\boldsymbol{\theta}_{m_0})d\boldsymbol{\theta}_{m_0}}{\int f(\boldsymbol{y}|\boldsymbol{\theta}_{m_1}, m_1)\pi(\boldsymbol{\theta}_{m_1})d\boldsymbol{\theta}_{m_1}}$$

Bayesian Variable Selection Tutorial

3. Priors for Bayesian Variable Selection in GLM

Normal models

Normal – Inverse Gamma (NIG) conjugate Prior

$$f(\boldsymbol{\beta}_m|\sigma^2, m) \sim N\left(\boldsymbol{\mu}_{\beta_m}, c^2\boldsymbol{V}_m\sigma^2\right) \;\; f(\sigma^2) \sim \text{IG}(a, b)$$

Marginal likelihood is analytically available

Main problem $\Rightarrow$ specification of $c^2\mathbf{V}_m$

Bayesian Variable Selection Tutorial

3. Priors for Bayesian Variable Selection in GLM

Normal models

Zellner's g-prior (Zellner, 1986)

NIG with

$$\boldsymbol{\mu}=\mathbf{0} \text{ and } \mathbf{V}_m=c^2\,(\mathbf{X}_m^T\mathbf{X}_m)^{-1}$$

$g=c^2$ in the original work of Zellner

$c^2=n \Rightarrow$ unit information prior (Kass and Wasserman, 1995, *JASA*)

See Fernandez *et al.* (2000, *J.Econom.*) for selection of g/c^2

Can be extended for GLMs

Bayesian Variable Selection Tutorial

3. Priors for Bayesian Variable Selection in GLM

Unit information prior (Kass and Wasserman, 1995, *JASA*)

$$\boldsymbol{\beta}_m|m \sim N\left(\underbrace{\widehat{\boldsymbol{\beta}}_m}_{\text{MLEs}},\, n[\underbrace{\mathcal{I}(\widehat{\boldsymbol{\beta}}_m)}_{\text{Observed Fisher information matrix}}]^{-1}\right)$$

Information equal to one data point

Uses data but minimally. It is still empirical.

Behavior approximately equal to BIC

Bayesian Variable Selection Tutorial
3. Priors for Bayesian Variable Selection in GLM

Unit information Empirical prior

Can built an empirical prior of unit information prior by using independent normal priors

$$\beta_j \sim N(\widetilde{\beta}_j, n\widetilde{\sigma}_j^2)$$

Posterior mean from full model

Posterior variance from full model

Will be ok when no correlated variables are included

Can be used as a yardstick

Bayesian Variable Selection Tutorial
3. Priors for Bayesian Variable Selection in GLM

Power prior and imaginary data

(Ibrahim and Chen, 2000, *Stat.Sci.*, Chen et al. 2000, *JSPI*)

$$f(\boldsymbol{\theta}_m|m) \propto f(\boldsymbol{y}^*|\boldsymbol{\theta}_m, m)^{1/c^2}$$

y*: imaginary data

c^2 : controls the weight given to imaginary data

c^2=n : accounts for one data point (Unit info prior)

Pre-prior can be also used ⇒ posterior using ***y****=prior for ***y.***

Bayesian Variable Selection Tutorial

3. Priors for Bayesian Variable Selection in GLM

Power prior and imaginary data

Normal models

$$f(\boldsymbol{\beta}_m|\sigma^2, \boldsymbol{y}^*, \boldsymbol{X}^*_m, m) \sim N\left(\widehat{\boldsymbol{\beta}}^*_m, c^2(\boldsymbol{X}^{*T}_m \boldsymbol{X}^*_m)^{-1}\sigma^2\right)$$

For $\boldsymbol{y}^*=0$ and $\mathbf{X}_m^*=\mathbf{X}_m \Rightarrow$ Zellner's g-prior

Other GLMs

Similar arguments can be used.

The distribution is approximately normal

(see for binary in Fouskakis *et al.* 2009, *Ann.Appl.Stats*)

Bayesian Variable Selection Tutorial

3. Priors for Bayesian Variable Selection in GLM

Zellner and Siow (1980) Priors

$\beta \sim$ Cauchy prior

Mean and variance similar to Zellner's g-prior

Mixtures of Zellner's g-priors *Liang et al. (2008, JASA)*

- Putting prior on g
- $\pi(g) = \frac{a-2}{2}(1+g)^{-a/2}, \quad g > 0$ $\Rightarrow$ Cauchy (Z-S prior)
- $\frac{g}{1+g} \sim \text{Beta}\left(1, \frac{a}{2}-1\right)$ $\Rightarrow$prior on shrinkage factor
  $\Rightarrow$$2<\alpha<4$ (α=3,4)

Bayesian Variable Selection Tutorial

3. Priors for Bayesian Variable Selection in GLM

Some comments

Normal priors $\Rightarrow$ ridge regression type of shrinkage

Double exponential priors $\Rightarrow$ LASSO regression type of shrinkage and penalization

Multivariate structure is important

Bayesian Variable Selection Tutorial

3. Priors for Bayesian Variable Selection in GLM

Intrinsic Priors

(Berger and Perrichi, 1996, JASA)

Priors that give approximately the same results as the Intrinsic Bayes Factor

IBF => BF after using a minimal training sample to build prior information within each model

AIBF => arithmetic IBF average over all possible training samples

Intrinsic Prior can use improper priors. Avoids Lindley-Bartlett paradox

Difficult to calculate

Bayesian Variable Selection Tutorial

3. Priors for Bayesian Variable Selection in GLM

Expected Posterior Priors

(Perez & Berger, 2002, Bka)

- *The posterior given some imaginary data* ***y**** *is averaged over all possible data configurations taken from the prior predictive distribution of a reference model* m_0.

$$f(\boldsymbol{\theta}_m|m) = \int f(\boldsymbol{\theta}_m|\boldsymbol{y}^*, m) f(\boldsymbol{y}^*|m_0) d\boldsymbol{y}^*$$

- Intrinsic prior of Berger & Perrichi (1996) = Expected Posterior prior
- Nice interpretation.
- Related with power prior via the use of imaginary data

Bayesian Variable Selection Tutorial

3. Priors for Bayesian Variable Selection in GLM

Priors on models

- Uniform on model space

$$f(m) = \frac{1}{|\mathcal{M}|} \propto 1$$

- A-priori penalizing for the model dimension

$$f(m) \propto \exp(-d_m F/2)$$

Bayesian Variable Selection Tutorial

3. Priors for Bayesian Variable Selection in GLM

Priors on variable indicators

Substitute *m* by $\gamma=(\gamma_1, \gamma_2, \ldots, \gamma_p)$ [George & McCulloch, 1993, *JASA*]

$\gamma_j \Rightarrow$ binary indicator $=1$ if X_j in the model
$=0$ if X_j out of the model

- Uniform on *m* $\Rightarrow f(\gamma_j)\sim$Bernoulli(1/2)
  Gives a-priori more weight to models with dimension p/2
- $f(\gamma_j)\sim$Bernoulli(π) and put beta hyper-prior on π.

6/12/2011@ University College Dublin
ISA short courses
Ioannis Ntzoufras
Bayesian Variable Selection – An Introductory Tutorial
35

Bayesian Variable Selection Tutorial

4. Computation of the marginal likelihood

Laplace Approximation

$$f(\boldsymbol{y}|m) \approx (2\pi)^{d_m/2}|\widetilde{\boldsymbol{\Sigma}}_m|^{1/2} f(\boldsymbol{y}|\widetilde{\boldsymbol{\theta}}_m, m) f(\widetilde{\boldsymbol{\theta}}_m|m)$$

$\widetilde{\boldsymbol{\theta}}_m$: the Posterior mode

$$\widetilde{\boldsymbol{\Sigma}}_m = \left(\mathbf{H}_m(\widetilde{\boldsymbol{\theta}}_m)\right)^{-1}$$

$\mathbf{H}_m(\widetilde{\boldsymbol{\theta}}_m)$: minus the second derivative of log $f(\theta_m|\boldsymbol{y},m)$ *evaluated at the posterior mode*

Works reasonably well for GLMs.

6/12/2011@ University College Dublin
ISA short courses
Ioannis Ntzoufras
Bayesian Variable Selection – An Introductory Tutorial
36

Bayesian Variable Selection Tutorial

4. Computation of the marginal likelihood

Laplace – Metropolis Estimator

[Raftery (1996, MCMC in Practice) & Lewis and Raftery (1997, JASA)]

The posterior mode can be substituted by the posterior mean or median (estimated from an MCMC output)

The approximate posterior variance can be estimated from an MCMC output.

ASSUMPTION: Posterior is symmetric (or close)

6/12/2011@ University College Dublin
ISA short courses
Ioannis Ntzoufras
Bayesian Variable Selection – An Introductory Tutorial
37

Bayesian Variable Selection Tutorial

4. Computation of the marginal likelihood

MONTE CARLO/MCMC ESTIMATORS

- Sampling from the prior – a naive Monte Carlo estimator
- Sampling from the posterior: The harmonic mean estimator (Kass and Raftery, 1995, *JASA*)
- Importance sampling estimators (Newton and Raftery, 1994)
- Bridge sampling estimators (Meng and Wong, 1996, *Stat.Sin.*),
- Chib's marginal likelihood estimator (Chib, 1995, *JASA*) and estimator via the Metropolis-Hastings output (Chib and Jeliazkov, 2001, *JASA*)
- Power Posteriors estimator (Friel and Pettit, 2008, *JRSSB*)
- Estimator via Gaussian Copula (Nott et al., 2009, Technical Report).

6/12/2011@ University College Dublin
ISA short courses
Ioannis Ntzoufras
Bayesian Variable Selection – An Introductory Tutorial
38

Bayesian Variable Selection Tutorial

4. Computation of the marginal likelihood

Disadvantages of MONTE CARLO/MCMC Estimators

- Need to obtain (one or more) samples from the posterior (or prior or other distributions) for every model.
- If the model space under consideration is large then evaluation of all models is impossible.
- Recommended only if the model space is small.

Bayesian Variable Selection Tutorial

5. MCMC algorithms for Bayesian Model Selection

Trans-dimensional MCMC methods $\Rightarrow$ extensions of usual MCMC methods

They solve both problems of

1) Calculation of the posterior model probabilities (and indirectly the marginal likelihood computation)
2) Model search especially when the model is large

Bayesian Variable Selection Tutorial

5. MCMC algorithms for Bayesian Model Selection

Trans-dimensional MCMC methods ⇒ extensions of usual MCMC methods

Good News – Advantages

1) Automatic after setting up the algorithm
2) Accurately traces best models and explores the model space
3) Posterior odds of best models can be estimated accurately
4) BMA can be directly applied
5) Obtain posterior distributions of both parameters and models

Bayesian Variable Selection Tutorial

5. MCMC algorithms for Bayesian Model Selection

Trans-dimensional MCMC methods ⇒ extensions of usual MCMC methods

Disadvantages

1) Need extensive computational resources
2) Experience on MCMC
3) Patience
4) Careful selection of proposals
5) Not accurate estimation of the marginal likelihood since focus is given on the estimation of posterior model probabilities (and odds)
6) Automatically cut-offs 'bad' models with low posterior probabilities
7) Over-estimates the probabilities of best models when the model space is large
8) Model exploration might demand extremely complicated algorithms when the model space is complicated (e.g. when collinear variables are involved).

Bayesian Variable Selection Tutorial

5. MCMC algorithms for Bayesian Model Selection

Notation

- m: model indicator for model m.
- $\boldsymbol{\theta}_m$: Parameter vector of model m.
  - Normal regression models $\Rightarrow$ $\boldsymbol{\theta}_m = (\boldsymbol{\beta}_m, \sigma^2)$.
  - In other GLMs (usually) $\Rightarrow$ $\boldsymbol{\theta}_m = \boldsymbol{\beta}_m$.

    $\boldsymbol{\beta}_m$ $\Rightarrow$ parameters involved in the linear predictor of a GLM.
- T: total number of iterations in an MCMC sample.
- $\theta^{(t)}$: value of θ generated at t iteration of the MCMC algorithm.

Bayesian Variable Selection Tutorial

5. MCMC algorithms for Bayesian Model Selection

Some details

- Generate a sample $(m^{(t)}, \boldsymbol{\theta}_m^{(t)},\ t = 1, \ldots, T)$ using an MCMC algorithm.
- Estimate posterior model probabilities by

Actually a frequency tabulation of $m^{(t)}$!!!

$$\widehat{f}(m|\boldsymbol{y}) = \frac{1}{T}\sum_{t=1}^{T} I(m^{(t)} = m) \qquad m \in \mathcal{M}$$

$I(\cdot)$: Indicator function; $\mathcal{M}$ is the set of models under consideration.

- Estimate $f(\boldsymbol{\theta}_m|m, \boldsymbol{y})$ using the sample $(\boldsymbol{\theta}_m^{(t)}$ for $m^{(t)} = m)$. This is available for 'best' models with samples large enough to be able to estimate the corresponding posterior distributions.

Bayesian Variable Selection Tutorial

5. MCMC algorithms for Bayesian Model Selection

What to report

1) **MAP model** – Maximum a-posteriori model: model with highest estimated posterior probability.
2) **Highest Probability Models**: Set a threshold and report the best model.
3) Report **Posterior Odds or Bayes Factors** (PO/BF) in comparison to MAP model (do not depend on the size of model space)
4) **Threshold** ⇒ difficult to be specified in terms of posterior probabilities (depends on the problem and the size of model space)

   ⇒ Use PO/BF interpretation to define the threshold for best models reported. For example report all models with PO<3 ("evidence in favor of better model which does not worth more than a bare mention") when compared to MAP.
5) When model uncertainty is large, select a group of good models and apply BMA (for example select the ones close to MAP with PO<3).

Bayesian Variable Selection Tutorial

5. MCMC algorithms for Bayesian Model Selection

General Model Selection Algorithms

- Markov chain Monte Carlo model composition [MC³] (Madigan and York, 1995, *Int.Stat.Review*).
- Reversible jump MCMC (Green, 1995, *Bka*).
- Carlin and Chib (1995, *JRSSB*) Gibbs sampler.

Variable selection samplers

- Stochastic Search Variable Selection [SSVS] (George & McCulloch, 1993, *JASA*).
- Kuo and Mallick (1998, *Sankya B*) Gibbs sampler.
- Gibbs Variable Selection (Dellaportas et al., 2002, *Stat. & Comp.*).

Bayesian Variable Selection Tutorial

6. Gibbs based methods for Bayesian variable selection

Substitute *m* by $\gamma=(\gamma_1, \gamma_2, \ldots, \gamma_p)$ [George & McCulloch, 1993, *JASA*]

γ_j $\Rightarrow$ binary indicator =1 if X_j in the model

=0 if X_j out of the model

m $\leftrightarrow$ γ **:** one-to-one relation between m and γ in variable selection problems.

Use binary system and calculate *m* using the equation

$$m = 1 + \sum_{j=1}^{p} \gamma_j 2^{j-1}$$

6/12/2011@ University College Dublin
ISA short courses

Ioannis Ntzoufras
Bayesian Variable Selection – An Introductory Tutorial

47

Bayesian Variable Selection Tutorial

6. Gibbs based methods for Bayesian variable selection

Important detail: In each MCMC iteration update all gammas (using random scan) $\Rightarrow$ big jumps in model space

What to report – (additional for variable selection)

1) **Posterior variable inclusion probabilities:** $f(\gamma_j=1|\mathbf{y})$ estimated by

$$\widehat{f}(\gamma_j = 1|\boldsymbol{y}) = \frac{1}{T}\sum_{t=1}^{T} I(\gamma_j^{(t)} = 1)$$

Means of $\gamma_j^{(t)}$!!!

2) **Median Probability (MP) Model.**

$\Rightarrow$ Model including variables with $f(\gamma_j=1|y) > 0.5$

$\Rightarrow$ Has better predictive performance than MAP model under certain conditions (Barbieri & Berger, 2004, *Ann.Stat.*)

6/12/2011@ University College Dublin
ISA short courses

Ioannis Ntzoufras
Bayesian Variable Selection – An Introductory Tutorial

48

Bayesian Variable Selection Tutorial

6.1 Stochastic Search Variable Selection (SSVS)

George & McCullogh (1993, JASA)

- Originally for Normal models and then applied in other GLM type models.
- Also popular in genetics and models for microarray data.
- $m \Rightarrow \gamma=(\gamma_1, \gamma_2, \ldots, \gamma_p)$
- γ_j :binary indicators for variable inclusion
- $\gamma_j=1$ in ($\gamma_j=0$ out) of the model

Bayesian Variable Selection Tutorial

6.1 Stochastic Search Variable Selection (SSVS)

Main Characteristic

The dimension of the model parameter vector is constant

(does not depend on model/variable indicator $\boldsymbol{\gamma}$)

- The model likelihood is the same $f(\boldsymbol{y}|\boldsymbol{\beta})$ for all models.
- $\Rightarrow$ So we can apply usual MCMC

Bayesian Variable Selection Tutorial

6.1 Stochastic Search Variable Selection (SSVS)

Main Characteristic

The dimension of the model parameter vector is constant.

How is this achieved?

- When the variable is not important (γ_j=0),
  $\Rightarrow$ posterior of β_j takes values very close to zero (instead of β_j=0)
  To do this $\Rightarrow$ very strong prior centered to zero when γ_j=0
- Linear predictor (in GLM) $\Rightarrow$ $\boldsymbol{\eta}=\mathbf{X}\boldsymbol{\beta}$
  (refers to the full model with all covariates in the model)

Bayesian Variable Selection Tutorial

6.1 Stochastic Search Variable Selection (SSVS)

Prior specification

- When variable is important

$$\boldsymbol{\beta}_j|\gamma_j=1 \sim N(0, \boldsymbol{\Sigma}_j)$$

- When variable is not important

$$\boldsymbol{\beta}_j|\gamma_j=0 \sim N(0, k_j^{-2}\boldsymbol{\Sigma}_j)$$

- k_j is large enough to ensure that the posterior will be also close to zero.

Bayesian Variable Selection Tutorial

6.1 Stochastic Search Variable Selection (SSVS)

Prior $\Rightarrow$ Mixture of two normals

$$\boldsymbol{\beta}_j|\gamma_j \sim \gamma_j N(0, \boldsymbol{\Sigma}_j) + (1-\gamma_j)N(0, k_j^{-2}\boldsymbol{\Sigma}_j)$$

Specification of k_j is described in the original paper.

For Poisson log-linear models; see Ntzoufras et el. (2000, *JSCS*).

Bayesian Variable Selection Tutorial

6.1 Stochastic Search Variable Selection (SSVS)

The algorithm (simple Gibbs sampler)

- Update β_j from

$$f(\boldsymbol{\beta}_j|\boldsymbol{y},\boldsymbol{\gamma},\boldsymbol{\beta}_{\setminus j}) \propto f(\boldsymbol{y}|\boldsymbol{\gamma},\boldsymbol{\beta})f(\boldsymbol{\beta}_j|\gamma_j)$$

- Update γ_j from a Bernoulli with $p = O_j/(1+O_j)$

$$O_j = \frac{f(\gamma_j=1|\boldsymbol{y},\boldsymbol{\gamma}_{\setminus j},\boldsymbol{\beta})}{f(\gamma_j=0|\boldsymbol{y},\boldsymbol{\gamma}_{\setminus j},\boldsymbol{\beta})} = \boxed{\frac{f(\boldsymbol{\beta}|\gamma_j=1,\boldsymbol{\gamma}_{\setminus j})}{f(\boldsymbol{\beta}|\gamma_j=0,\boldsymbol{\gamma}_{\setminus j})}}\boxed{\frac{f(\gamma_j=1,\boldsymbol{\gamma}_{\setminus j})}{f(\gamma_j=0,\boldsymbol{\gamma}_{\setminus j})}}$$

Variable selection Step does not depend on likelihood but only on the prior

Prior ratio for parameters

Prior model odds

Bayesian Variable Selection Tutorial

6.1 Stochastic Search Variable Selection (SSVS)

The algorithm (simple Gibbs sampler)

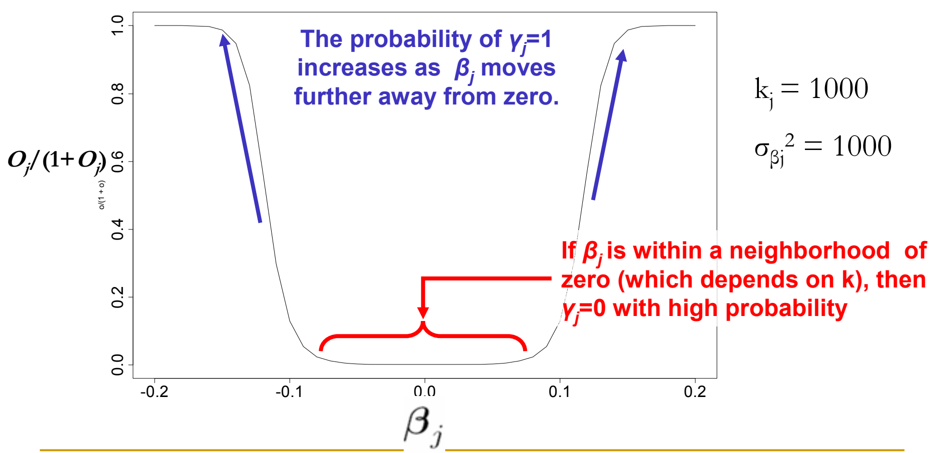
$$O_j = \frac{f(\gamma_j = 1|\boldsymbol{y}, \boldsymbol{\gamma}_{\setminus j}, \boldsymbol{\beta})}{f(\gamma_j = 0|\boldsymbol{y}, \boldsymbol{\gamma}_{\setminus j}, \boldsymbol{\beta})} = \frac{f(\boldsymbol{\beta}|\gamma_j = 1, \boldsymbol{\gamma}_{\setminus j})}{f(\boldsymbol{\beta}|\gamma_j = 0, \boldsymbol{\gamma}_{\setminus j})} \frac{f(\gamma_j = 1, \boldsymbol{\gamma}_{\setminus j})}{f(\gamma_j = 0, \boldsymbol{\gamma}_{\setminus j})}$$

$$= \frac{1}{k_j} \exp\left(-\frac{1}{2} \frac{1 - k_j^2}{\sigma^2_{\beta_j}} \right)$$

When
- **all covariates are numeric or binary**
- **all models have the same prior probability**
- **independent priors are used as earlier**

Bayesian Variable Selection Tutorial

6.1 Stochastic Search Variable Selection (SSVS)



Bayesian Variable Selection Tutorial

6.1 Stochastic Search Variable Selection (SSVS)

Advantages

- The same likelihood for all models
- Simple Gibbs Sampler
- Easy to adopt for any GLM

Bayesian Variable Selection Tutorial

6.1 Stochastic Search Variable Selection (SSVS)

Disadvantages

- Results are not exactly the same as in usual variable selection (with β_j=0). Tend to be similar as $k_j \rightarrow \infty$
- Selection of k_j may be difficult.
- MCMC not flexible.
- when k_j too large
  - ⇒ MCMC not mobile in model space
  - ⇒ Overflows are observed in MCMC when k_j too large
- Independent priors my cause strange behavior especially when X are collinear

Bayesian Variable Selection Tutorial

6.2 Kuo & Mallick (KM) Sampler

Kuo & Mallick (1998, Stat. Sinica)
Unconditional (on model) prior distribution

Main Characteristics

- Model Dimension is constant
- Likelihood depends on γ

  How? Via the linear predictor $\eta = \sum_{j=0}^{p} \gamma_j \boldsymbol{X}_j \beta_j$
- Prior is specified only for the full model

Bayesian Variable Selection Tutorial

6.2 Kuo & Mallick (KM) Sampler

Main Characteristics – Prior unconditional on γ

$$f(\boldsymbol{\beta},\gamma) = f(\boldsymbol{\beta})f(\gamma) = f(\boldsymbol{\beta}_\gamma|\boldsymbol{\beta}_{\setminus\gamma})f(\boldsymbol{\beta}_{\setminus\gamma})f(\gamma)$$

$\boldsymbol{\beta}_\gamma$ **Parameter vector of model γ (i.e. β_j with γ_j=1)**

$\boldsymbol{\beta}_{\setminus\gamma}$ **Parameters for variables not included in model γ (i.e. β_j with γ_j=0)**

Actual Prior: $f(\boldsymbol{\beta}_\gamma) = \int f(\boldsymbol{\beta}_\gamma, \boldsymbol{\beta}_{\setminus\gamma}) d\boldsymbol{\beta}_{\setminus\gamma}$

Bayesian Variable Selection Tutorial

6.2 Kuo & Mallick (KM) Sampler

The algorithm (Gibbs sampler)

- Update β_j from

Posterior

$$f(\boldsymbol{\beta}_j|\boldsymbol{y},\boldsymbol{\gamma},\boldsymbol{\beta}_{\setminus j}) \propto \begin{cases} \boxed{f(\boldsymbol{y}|\boldsymbol{\gamma},\boldsymbol{\beta})f(\boldsymbol{\beta}_j|\boldsymbol{\beta}_{\setminus j}) \quad \gamma_j = 1} \\ \boxed{f(\boldsymbol{\beta}_j|\boldsymbol{\beta}_{\setminus j}) \quad \gamma_j = 0} \end{cases}$$

Conditional prior

"Proposes" values for β_j when X_j *is not included in the model*

Bayesian Variable Selection Tutorial

6.2 Kuo & Mallick (KM) Sampler

The algorithm (Gibbs sampler)

- Update γ_j from a Bernoulli with $p= O_j/(1+O_j)$

$$O_j = \frac{f(\gamma_j = 1|\boldsymbol{y},\boldsymbol{\gamma}_{\setminus j},\boldsymbol{\beta})}{f(\gamma_j = 0|\boldsymbol{y},\boldsymbol{\gamma}_{\setminus j},\boldsymbol{\beta})} = \boxed{\frac{f(\boldsymbol{y}|\gamma_j = 1,\boldsymbol{\gamma}_{\setminus j},\boldsymbol{\beta})}{f(\boldsymbol{y}|\gamma_j = 0,\boldsymbol{\gamma}_{\setminus j},\boldsymbol{\beta})}}\boxed{\frac{f(\gamma_j = 1,\boldsymbol{\gamma}_{\setminus j})}{f(\gamma_j = 0,\boldsymbol{\gamma}_{\setminus j})}}$$

Variable selection Step does not depend on the prior but only on the likelihood

Likelihood ratio

Prior model odds

If current β_j is close to the conditional MLE we include the variable with high probability (close to 1)

If current β_j is close to zero we exclude the variable with probability ½

Bayesian Variable Selection Tutorial

6.2 Kuo & Mallick (KM) Sampler

Advantages

- Simple Gibbs Sampler
- Need to specify only the prior of the full model
- Multivariate priors on β can be used without any problem
- Easy to adopt for any GLM
- Works reasonably well for GLM

Bayesian Variable Selection Tutorial

6.2 Kuo & Mallick (KM) Sampler

Disadvantages

- Selection of the prior for the full model may result to strange priors for each model
- MCMC not flexible.
- Does not work efficiently when collinear variables exist

Bayesian Variable Selection Tutorial

6.3 Gibbs variable Selection (GVS)

Dellaportas et al. (2002) Statistics & Computing

Natural hybrid of SSVS and KM sampler

Main Characteristics

- Same likelihood as in KM Sampler
- Prior depends on model structure

$$\eta = \sum_{j=0}^{p} \gamma_j \boldsymbol{X}_j \boldsymbol{\beta}_j$$

6/12/2011@ University College Dublin
ISA short courses
Ioannis Ntzoufras
Bayesian Variable Selection – An Introductory Tutorial
65

Bayesian Variable Selection Tutorial

6.3 Gibbs variable Selection (GVS)

Prior Structure

$$f(\boldsymbol{\beta}, \boldsymbol{\gamma}) = \underbrace{\boxed{f(\boldsymbol{\beta}_{\gamma}|\boldsymbol{\gamma})}}_{\text{Actual Prior}} \underbrace{\boxed{f(\boldsymbol{\beta}_{\setminus\gamma}|\boldsymbol{\beta}_{\gamma}, \boldsymbol{\gamma})}}_{\text{“Pseudo prior”!!!}} \underbrace{\boxed{f(\boldsymbol{\gamma})}}_{\text{Prior of model } \boldsymbol{\gamma}}$$

6/12/2011@ University College Dublin
ISA short courses
Ioannis Ntzoufras
Bayesian Variable Selection – An Introductory Tutorial
66

Bayesian Variable Selection Tutorial

6.3 Gibbs variable Selection (GVS)

Prior Structure

$$f(\boldsymbol{\beta},\gamma) = f(\boldsymbol{\beta}_\gamma|\gamma)\boxed{f(\boldsymbol{\beta}_{\setminus\gamma}|\boldsymbol{\beta}_\gamma,\gamma)}f(\gamma)$$

Why Pseudo prior???

Does not affect posterior of $f(\boldsymbol{\beta}_\gamma|\boldsymbol{y},\gamma)$

Since the likelihood does not depend on $\boldsymbol{\beta}_{\setminus\gamma}$

Why is it useful

Since it does not appears the posterior can be defined at our convenience

Can be used to make the algorithm more efficient

Bayesian Variable Selection Tutorial

6.3 Gibbs variable Selection (GVS)

The algorithm (Gibbs sampler)

- Update β_j from

Posterior for β_j included in model

$$\begin{aligned} f(\boldsymbol{\beta}_\gamma|\boldsymbol{\beta}_{\setminus\gamma},\gamma,\boldsymbol{y}) &\propto \boxed{f(\boldsymbol{y}|\boldsymbol{\beta},\gamma)f(\boldsymbol{\beta}_\gamma|\gamma)f(\boldsymbol{\beta}_{\setminus\gamma}|\boldsymbol{\beta}_\gamma,\gamma)} \\ f(\boldsymbol{\beta}_{\setminus\gamma}|\boldsymbol{\beta}_\gamma,\gamma,\boldsymbol{y}) &\propto \boxed{f(\boldsymbol{\beta}_{\setminus\gamma}|\boldsymbol{\beta}_\gamma,\gamma),} \end{aligned}$$

Pseudo prior

"Proposes" values for β_j when X_j *is not included in the model*

Bayesian Variable Selection Tutorial

6.3 Gibbs variable Selection (GVS)

The algorithm (Gibbs sampler)

- Update β_j from

$$
\begin{aligned}
f(\boldsymbol{\beta}_{\gamma}|\boldsymbol{\beta}_{\setminus\gamma},\gamma,\boldsymbol{y}) &\propto f(\boldsymbol{y}|\boldsymbol{\beta},\gamma)f(\boldsymbol{\beta}_{\gamma}|\gamma)\cancel{f(\boldsymbol{\beta}_{\setminus\gamma}|\boldsymbol{\beta}_{\gamma},\gamma)}\\
f(\boldsymbol{\beta}_{\setminus\gamma}|\boldsymbol{\beta}_{\gamma},\gamma,\boldsymbol{y}) &\propto f(\boldsymbol{\beta}_{\setminus\gamma}|\cancel{\boldsymbol{\beta}_{\gamma}},\gamma),
\end{aligned}
$$

CONVENIENT/PLAUSIBLE ASSUMPTION

$$f(\boldsymbol{\beta}_{\setminus\gamma}|\boldsymbol{\beta}_{\gamma},\gamma) = f(\boldsymbol{\beta}_{\setminus\gamma}|\gamma)$$

Bayesian Variable Selection Tutorial

6.3 Gibbs variable Selection (GVS)

The algorithm (Gibbs sampler)

- Update γ_j from a Bernoulli with $p = O_j/(1+O_j)$

$$O_j = \frac{f(\gamma_j = 1|\gamma_{\setminus j},\boldsymbol{\beta},\boldsymbol{y})}{f(\gamma_j = 0|\gamma_{\setminus j},\boldsymbol{\beta},\boldsymbol{y})}$$

Bayesian Variable Selection Tutorial

6.3 Gibbs variable Selection (GVS)

The algorithm (Gibbs sampler)

- Update γ_j from a Bernoulli with $p = O_j/(1+O_j)$

$$O_j = \boxed{\frac{f(\boldsymbol{y}|\boldsymbol{\beta},\gamma_j=1,\boldsymbol{\gamma}_{\setminus j})}{f(\boldsymbol{y}|\boldsymbol{\beta},\gamma_j=0,\boldsymbol{\gamma}_{\setminus j})}} \boxed{\frac{f(\boldsymbol{\beta}|\gamma_j=1,\boldsymbol{\gamma}_{\setminus j})}{f(\boldsymbol{\beta}|\gamma_j=0,\boldsymbol{\gamma}_{\setminus j})}} \boxed{\frac{f(\gamma_j=1,\boldsymbol{\gamma}_{\setminus j})}{f(\gamma_j=0,\boldsymbol{\gamma}_{\setminus j})}}$$

Likelihood ratio

Prior over Proposal

Prior model odds

Variable selection step depends on both LR (as in KM) and Prior ratio (as in SSVS)

The prior ratio naturally decomposes to prior vs. pseudo prior ratio if independent priors+proposal are used.

6/12/2011@ University College Dublin
ISA short courses

Ioannis Ntzoufras
Bayesian Variable Selection – An Introductory Tutorial

71

Bayesian Variable Selection Tutorial

6.3 Gibbs variable Selection (GVS)

The algorithm (Gibbs sampler)

- Update γ_j from a Bernoulli with $p = O_j/(1+O_j)$

$$O_j = \frac{f(\boldsymbol{y}|\boldsymbol{\beta},\gamma_j=1,\boldsymbol{\gamma}_{\setminus j})}{f(\boldsymbol{y}|\boldsymbol{\beta},\gamma_j=0,\boldsymbol{\gamma}_{\setminus j})} \boxed{\frac{f(\beta_j|\gamma_j=1)}{f(\beta_j|\gamma_j=0)}} \frac{f(\gamma_j=1,\boldsymbol{\gamma}_{\setminus j})}{f(\gamma_j=0,\boldsymbol{\gamma}_{\setminus j})}$$

The prior ratio naturally decomposes to prior vs. pseudo prior ratio if independent priors+proposal are used i.e.

$$f(\boldsymbol{\beta}|\boldsymbol{\gamma}) = \prod_{j=1}^{p} f(\beta_j|\gamma_j)$$ with (for example)

$$f(\boldsymbol{\beta}_j|\boldsymbol{\gamma}_j) = \gamma_j N(0,\boldsymbol{\Sigma}_j) + (1-\gamma_j)N(\overline{\mu}_j,\overline{\boldsymbol{S}}_j)$$

72

Bayesian Variable Selection Tutorial

6.3 Gibbs variable Selection (GVS)

Selection of Pseudo prior (1/3)

1) Automatic – conditional MLE.

$$f(\boldsymbol{\beta}_{\setminus\gamma}|\boldsymbol{\beta}_{\gamma},\boldsymbol{\gamma})$$

1. Calculate residuals of current model
2. Fit MLE without constant
3. Use estimate and error of this model as mean and variance of the conditional proposal.

Maximization in (2) is trivial for normal model since one variable is involved.

For GLMs, also normal approximation can be used.

Bayesian Variable Selection Tutorial

6.3 Gibbs variable Selection (GVS)

Selection of Pseudo prior (2/3)

2) Independent pseudo priors from the full model

$$f(\boldsymbol{\beta}_{\setminus\gamma}|\boldsymbol{\beta}_{\gamma},\boldsymbol{\gamma}) = \prod^{p} f(\beta_j|\gamma_j=0)^{1-\gamma_j}$$

$$\beta_j|\gamma_j = 0 \sim N(\boxed{\overline{\mu}_j, \overline{\boldsymbol{S}}_j})$$

THIS WORKS REASONABLY WELL WHEN NO HIGHLY CORRELATED VARIABLES ARE INVOLVED

Pseudo prior parameter can be taken by an MCMC pilot run of the full model

Bayesian Variable Selection Tutorial

6.3 Gibbs variable Selection (GVS)

Selection of Pseudo prior (3/3)

3) Independent SSVS type pseudo priors

$$f(\boldsymbol{\beta}_{\setminus\gamma}|\boldsymbol{\beta}_{\gamma},\boldsymbol{\gamma}) = \prod_{j=1}^{p} f(\beta_j|\gamma_j=0)^{1-\gamma_j}$$

$$\beta_j|\gamma_j = 0 \sim N(0, \boxed{k_j^{-2}\boldsymbol{\Sigma}_j})$$

Pseudo prior is equal to the prior with variance deinflated by a factor k_j as in SSVS

When posterior of β_j *is close to zero* $\Rightarrow$ like SSVS$\Rightarrow$ prior ratio dominates the variable selection step

When posterior of β_j *is away from zero*$\Rightarrow$ *even small deviations from zero will make the LR to support the inclusion of* β_j in the model

EMPIRICAL RESULTS SHOW THAT THIS WORKS REASONABLY WELL IN MOST CASES

Bayesian Variable Selection Tutorial

6.3 Gibbs variable Selection (GVS)

Advantages

1. Relatively simple
2. Combines advantages of SSVS+KM samplers
3. Allows for multivariate priors
4. Flexible since pseudo priors can be defined
5. Efficient due to the use of pseudo priors
6. Searches the space efficiently in no collinear covariates are involved with pseudo prior of type 2.
7. Pseudoprior of type 1 can improve the model search and it is also "automatic"

Bayesian Variable Selection Tutorial

6.3 Gibbs variable Selection (GVS)

Disadvantages

1. Specification of pseudopriors
2. May not so efficient when collinear variables are involved

Bayesian Variable Selection Tutorial

6.5 Comparison of Gibbs based methods

Method	η	O_j		
		PSR_j	LR_j	PR_j
SSVS	$\boldsymbol{X\beta}$			$\surd$
KM	$\sum \gamma_j \boldsymbol{X}_j \boldsymbol{\beta}_j$		$\surd$	
GVS	$\sum \gamma_j \boldsymbol{X}_j \boldsymbol{\beta}_j$	$\surd$	$\surd$	$\surd$

Key: PSR = Pseudoprior Ratio; LR = Likelihood Ratio; PR = Prior Density Ratio; SSVS = stochastic search variable selection; KM = Kuo–Mallick method; GVS = Gibbs variable selection.

Bayesian Variable Selection Tutorial

7. Bayesian variable selection Using WinBUGS

All Gibbs based methods can be implemented in WinBUGS

See

Dellaportas, P., Forster, J.J. and Ntzoufras, I. (2000). Bayesian Variable Selection Using the Gibbs Sampler. *Generalized Linear Models: A Bayesian Perspective* (D. K. Dey, S. Ghosh, and B. Mallick, eds.). New York: Marcel Dekker, 271 – 286.

Ntzoufras, I. (2002). Gibbs Variable Selection Using BUGS. *Journal of Statistical Software*, Volume 7, Issue 7.

Ntzoufras, I. (2009). *Bayesian Modeling Using WinBUGS*. Wiley Series in Computational Statistics, Hoboken, USA.

Bayesian Variable Selection Tutorial

7. Bayesian variable selection Using WinBUGS

Dellaportas et al. (2002) Simulated data

- p=15 simulated N(0,1) covariates
- 2^{15}= 32,768
- n=50
- Independent Xs so MCMC easy to implement
- True model

$$Y_i \sim N(X_{i4} + X_{i5}, (2.5)^2) \text{ for } i = 1, 2, \ldots, 50$$

Bayesian Variable Selection Tutorial

7. Bayesian variable selection Using WinBUGS

Three prior set-ups

1) Prior used in Dellaportas *et al.* (2002) $\beta_j \sim N(0, 100)$

$$\sigma^2 \sim Inv.Gamma(0.01, 0.01)$$

2) Zellner's g-prior with g=c^2=n

3) Empirical Bayes independent prior distribution accounting for approximately one data point.

$$\beta_j \sim N(\widetilde{\beta}_j, n\widetilde{S}^2_{\beta_j})$$

posterior mean of β_j *posterior* variance of β_j

Bayesian Variable Selection Tutorial

7. Bayesian variable selection Using WinBUGS

Model Likelihood

$$Y_i \sim N(\mu_i, \sigma^2)$$
$$\mu_i = \gamma_0\beta_0 + \beta_1\gamma_1 X_{i1} + \cdots + \beta_{15}\gamma_{15}X_{i15}$$
$$\text{for } i = 1, 2, \ldots, n$$

```
for (i in 1:n){
    y[i] ~ dnorm( mu[i], tau)
    mu[i] <- gamma0*beta0 + x[i,1]*gamma[1]*beta[1] + ...
                          + x[i,15]*gamma[15]*beta[15]
}
```

Bayesian Variable Selection Tutorial

7. Bayesian variable selection Using WinBUGS

Model Likelihood

$$Y_i \sim N(\mu_i, \sigma^2)$$
$$\mu_i = \gamma_0\beta_0 + \beta_1\gamma_1 X_{i1} + \cdots + \beta_{15}\gamma_{15}X_{i15}$$
$$\boxed{\text{for } i = 1, 2, \ldots, n}$$

```
for (i in 1:n){
    y[i] ~ dnorm( mu[i], tau)
    mu[i] <- gamma0*beta0 + x[i,1]*gamma[1]*beta[1] + ...
                    + x[i,15]*gamma[15]*beta[15]
}
```

6/12/2011@ University College Dublin
ISA short courses

Ioannis Ntzoufras
Bayesian Variable Selection – An Introductory Tutorial

Bayesian Variable Selection Tutorial

7. Bayesian variable selection Using WinBUGS

Model Likelihood

alternative expression for linear predictor

Calculate all (actual) betas by setting $gb_j = \gamma_j \beta_j$ for j=1,…, p

Use inprod to calculate the sum $\sum_{j=1}^{p} \beta_j \gamma_j X_{ij}$ involved the linear predictor

```
for (j in 1:p){ gb[j] <- gamma[j]*beta[j] }
for (i in 1:n){
        y[i] ~ dnorm( mu[i], tau )
        mu[i] <- gamma0 * beta0 + inprod( x[i,1:p], gb[1:p] )
}
```

6/12/2011@ University College Dublin
ISA short courses

Ioannis Ntzoufras
Bayesian Variable Selection – An Introductory Tutorial

Bayesian Variable Selection Tutorial

7. Bayesian variable selection Using WinBUGS

Model Likelihood

Sum can be used instead

gb = stores the actual values of parameters while beta has also proposed values (which are non-sense for inference)

The inprod is convenient when p is large

When multivariate prior is used then β_0 must be also included in a single coefficient vector

Bayesian Variable Selection Tutorial

7. Bayesian variable selection Using WinBUGS

Prior on variable indicators

$$\gamma_j \sim Bernoulli(1/2) \text{ for } j = 0, 1, \ldots, p$$

If the constant is always included in the model, then

`gamma0 <- 1.0`

```
gamma0~dbern(0.5)
for (j in 1:p){ gamma[j]~dbern(0.5)   }
```

Bayesian Variable Selection Tutorial

7. Bayesian variable selection Using WinBUGS

Prior & pseudoprior specification

$$\beta_j \sim \gamma_j N(0,100) + (1-\gamma_j) N(\overline{\mu}_j, \overline{\sigma}^2_{\beta_j})$$

$$\begin{aligned} \beta_j &\sim \gamma_j N(\mu_j, \tau_{\beta_j}^{-1}) \\ \mu_j &= \gamma_j 0 + (1-\gamma_j)\overline{\mu}_j \\ \tau_{\beta_j} &= \gamma_j/100 + (1-\gamma_j)/\overline{\sigma}^2_{\beta_j} \end{aligned}$$

```
for (j in 1:p){
    beta[j] ~ dnorm( mb[j], taub[j])
    mb[j] <- (1-gamma0) * prop.mean.beta[j]
    taub[j]<- gamma[j]*0.01 + (1-gamma[j])/pow(prop.sd.beta[j],2)
}
```

Bayesian Variable Selection Tutorial

7. Bayesian variable selection Using WinBUGS

Prior & pseudoprior specification

$$\begin{aligned} \beta_j &\sim \gamma_j N(\mu_j, \tau_{\beta_j}^{-1}) \\ \mu_j &= \gamma_j 0 + (1-\gamma_j)\overline{\mu}_j \\ \tau_{\beta_j} &= \gamma_j/100 + (1-\gamma_j)/\overline{\sigma}^2_{\beta_j} \end{aligned}$$

$$\tau \sim Gamma(0.01, 0.01) \Leftrightarrow \sigma^2 \sim Inv.Gamma(0.01, 0.01)$$

```
gamma0~dbern(0.5)
for (j in 1:p){ gamma[j]~dbern(0.5)  }
beta0 ~ dnorm( mb0, taub0)
mb0 <- (1-gamma0) * prop.mean.beta0
taub0 <- gamma0*0.01 + (1-gamma0) /pow(prop.sd.beta0 ,2)
for (j in 1:p){
    beta[j] ~ dnorm( mb[j], taub[j])
    mb[j] <- (1-gamma0) * prop.mean.beta[j]
    taub[j]<- gamma[j]*0.01 + (1-gamma[j])/pow(prop.sd.beta[j],2)
}
tau~dgamma( 0.01, 0.01)
```

Bayesian Variable Selection Tutorial

7. Bayesian variable selection Using WinBUGS

Multivariate prior specification

$$
\begin{aligned}
\boldsymbol{\beta}_\gamma &\sim N(\mathbf{0}, n(\boldsymbol{X}_\gamma^T \boldsymbol{X}_\gamma)^{-1}) \\
presicion &= (\boldsymbol{X}_\gamma^T \boldsymbol{X}_\gamma)/n \\
&\Rightarrow \text{submatrix of } (\boldsymbol{X}^T \boldsymbol{X})/n \\
\beta_j | \gamma_j = 0 &\sim N(\overline{\mu}_j, \overline{\sigma}^2_{\beta_j})
\end{aligned}
$$

```
B[1:(p+1)] ~ dmnorm( mean.beta[1:(p+1)], T[ 1:(p+1), 1:(p+1)] )
tau~dgamma( 0.01, 0.01)
for (j in 1:(p+1) ){ mean.beta[j] <-  (1-g[j])*prop.mean.beta[j] }
for (j in 1:(p+1) ){ for (k in 1:(p+1) ){
    T[j,k] < - g[j]*g[k]*tau*XTX[j,k]/n
             + ( 1- g[j]*g[k])*equals(j,k)*pow(prop.sd.beta[k],-2)
}}
```

Bayesian Variable Selection Tutorial

7. Bayesian variable selection Using WinBUGS

Multivariate prior specification

$$
\begin{pmatrix} \boldsymbol{\beta}_\gamma \\ \boldsymbol{\beta}_{\setminus\gamma} \end{pmatrix} \sim N\left(\begin{pmatrix} \mathbf{0} \\ (\overline{\mu}_j \text{ for } \gamma_j = 0) \end{pmatrix}, \begin{pmatrix} (\boldsymbol{X}_\gamma^T \boldsymbol{X}_\gamma)/n & \mathbf{0} \\ \mathbf{0} & \text{diag}(\overline{\sigma}^2_{\beta_j} \text{ for } \gamma_j = 0) \end{pmatrix}^{-1} \right)
$$

```
B[1:(p+1)] ~ dmnorm( mean.beta[1:(p+1)], T[ 1:(p+1), 1:(p+1)] )
tau~dgamma( 0.01, 0.01)
for (j in 1:(p+1) ){ mean.beta[j] <-  (1-g[j])*prop.mean.beta[j] }
for (j in 1:(p+1) ){ for (k in 1:(p+1) ){
    T[j,k] < - g[j]*g[k]*tau*XTX[j,k]/n
             + ( 1- g[j]*g[k])*equals(j,k)*pow(prop.sd.beta[k],-2)
}}
```

Bayesian Variable Selection Tutorial

7. Bayesian variable selection Using WinBUGS

Multivariate prior specification

$$\begin{aligned}\boldsymbol{\beta} &\sim N(\boldsymbol{\mu}, \boldsymbol{T}^{-1})\\ \mu_j &= \gamma_j 0 + (1-\gamma_j)\overline{\mu}_j\\ \boldsymbol{T} &= \text{precision matrix}\\ T_{jk} &= \begin{cases}\boldsymbol{\Sigma}_{jk}^{-1} & \gamma_j=\gamma_k=1\\ \overline{\sigma}_j^{-2} & j=k \text{ and } \gamma_j=0\\ 0 & otherwize\end{cases}\end{aligned}$$

```
B[1:(p+1)] ~ dmnorm( mean.beta[1:(p+1)], T[ 1:(p+1), 1:(p+1)] )
tau~dgamma( 0.01, 0.01)
for (j in 1:(p+1) ){ mean.beta[j] <-  (1-g[j])*prop.mean.beta[j] }
for (j in 1:(p+1) ){ for (k in 1:(p+1) ){
    T[j,k] < - g[j]*g[k]*tau*XTX[j,k]/n
             + ( 1- g[j]*g[k])*equals(j,k)*pow(prop.sd.beta[k],-2)
}}
```

Bayesian Variable Selection Tutorial

7. Bayesian variable selection Using WinBUGS

Multivariate prior specification

$$\begin{aligned}\boldsymbol{\beta} &\sim N(\boldsymbol{\mu}, \boldsymbol{T}^{-1})\\ \mu_j &= \gamma_j 0 + (1-\gamma_j)\overline{\mu}_j\\ \boldsymbol{T} &= \text{precision matrix}\\ T_{jk} &= \begin{cases}\boldsymbol{\Sigma}_{jk}^{-1} & \gamma_j=\gamma_k=1\\ \overline{\sigma}_j^{-2} & j=k \text{ and } \gamma_j=0\\ 0 & otherwize\end{cases}\end{aligned}$$

```
B[1:(p+1)] ~ dmnorm( mean.beta[1:(p+1)], T[ 1:(p+1), 1:(p+1)] )
tau~dgamma( 0.01, 0.01)
for (j in 1:(p+1) ){ mean.beta[j] <-  (1-g[j])*prop.mean.beta[j] }
for (j in 1:(p+1) ){ for (k in 1:(p+1) ){
    T[j,k] < - g[j]*g[k]*tau*XTX[j,k]/n
             + ( 1- g[j]*g[k])*equals(j,k)*pow(prop.sd.beta[k],-2)
}}
```

Bayesian Variable Selection Tutorial

7. Bayesian variable selection Using WinBUGS

Multivariate prior specification

Similarly can be adjusted for any multivariate prior

Bayesian Variable Selection Tutorial

7. Bayesian variable selection Using WinBUGS

Model indicator

Use equation $m = 1 + \sum_{j=1}^{p} \gamma_j 2^{j-1}$

```
for (j in 1:p){ mindex[j] <- pow(2,j)}
model <- gamma0 + inprod( gamma[], mindex[] )
```

Model probabilities

In our example: TRUE model $\Rightarrow$ $2^4+2^5=48$

Best/2nd best model $\Rightarrow$ $2^4+2^5+2^{12}=4144$

```
pmodel[1] <- equals( model, pow(2,4)+pow(2,5) )
pmodel[2] <- equals( model, pow(2,4)+pow(2,5)+pow(2,12) )
```

Bayesian Variable Selection Tutorial

7. Bayesian variable selection Using WinBUGS

Results – Posterior inclusion probabilities

	Prior 1[a]		Prior 2[b]		Prior 3[c]	
	$f(\gamma_j=1\|\boldsymbol{y})$	MC error	$f(\gamma_j=1\|\boldsymbol{y})$	MC error	$f(\gamma_j=1\|\boldsymbol{y})$	MC error
γ_0	0.042	0.0045	0.134	0.0025	0.039	0.0016
γ_1	0.031	0.0012	0.128	0.0025	0.106	0.0022
γ_2	0.039	0.0014	0.136	0.0027	0.113	0.0023
γ_3	0.033	0.0013	0.127	0.0024	0.101	0.0018
γ_4	0.970	0.0024	0.992	0.0001	0.990	0.0001
γ_5	0.999	0.0001	1.000	0.0000	1.000	0.0001
γ_6	0.046	0.0016	0.155	0.0028	0.128	0.0025
γ_7	0.037	0.0015	0.138	0.0028	0.117	0.0023
γ_8	0.041	0.0015	0.133	0.0023	0.105	0.0025
γ_9	0.044	0.0014	0.168	0.0027	0.138	0.0027
γ_{10}	0.043	0.0015	0.141	0.0029	0.115	0.0021
γ_{11}	0.048	0.0015	0.184	0.0030	0.147	0.0027
γ_{12}	0.338	0.0033	0.615	0.0040	0.545	0.0034
γ_{13}	0.038	0.0014	0.137	0.0024	0.106	0.0024
γ_{14}	0.042	0.0014	0.137	0.0023	0.104	0.0021
γ_{15}	0.076	0.0019	0.277	0.0037	0.243	0.0032

95

Bayesian Variable Selection Tutorial

7. Bayesian variable selection Using WinBUGS

Rank	m	m_k	Model	$f(m\|\boldsymbol{y})$	$PO_{m_1\,m_k}$
			Prior 1[a]		
1	48	m_1	X_4+X_5	0.3664	1.00
2	4,144	m_2	$X_4+X_5+X_{12}$	0.1854	1.98
3	32,816	m_3	$X_4+X_5+X_{15}$	0.0292	12.55
4	560	m_4	$X_4+X_5+X_9$	0.0196	18.69
5	112	m_5	$X_4+X_5+X_6$	0.0178	20.58
6	16,432	m_6	$X_4+X_5+X_{14}$	0.0176	20.82
7	2,096	m_7	$X_4+X_5+X_{11}$	0.0172	21.30
8	1,072	m_8	$X_4+X_5+X_{10}$	0.0157	23.34
9	8,240	m_9	$X_4+X_5+X_{13}$	0.0150	24.43
10	49	m_{10}	$X_0+X_4+X_5$	0.0149	24.59

ISA short courses *Bayesian Variable Selection – An Introductory Tutorial* 96

Bayesian Variable Selection Tutorial

7. Bayesian variable selection Using WinBUGS

Rank	m	m_k	Model	$f(m\|\boldsymbol{y})$	$\mathrm{PO}_{m_1\, m_k}$
			Prior 1[a]		
1	48	m_1	$X_4 + X_5$	0.3664	1.00
2	4,144	m_2	$X_4 + X_5 + X_{12}$	0.1854	1.98
3	32,816	m_3	$X_4 + X_5 + X_{15}$	0.0292	12.55
4	560	m_4	$X_4 + X_5 + X_9$	0.0196	18.69
5	112	m_5	$X_4 + X_5 + X_6$	0.0178	20.58
6	16,432	m_6	$X_4 + X_5 + X_{14}$	0.0176	20.82
7	2,096	m_7	$X_4 + X_5 + X_{11}$	0.0172	21.30
8	1,072	m_8	$X_4 + X_5 + X_{10}$	0.0157	23.34
9	8,240	m_9	$X_4 + X_5 + X_{13}$	0.0150	24.43
10	49	m_{10}	$X_0 + X_4 + X_5$	0.0149	24.59

ISA short courses *Bayesian Variable Selection – An Introductory Tutorial* 97

Bayesian Variable Selection Tutorial

7. Bayesian variable selection Using WinBUGS

Rank	m	m_k	Model	$f(m\|\boldsymbol{y})$	$\mathrm{PO}_{m_1\, m_k}$
			Prior 2[b]		
1	4,144	m_2	$X_4 + X_5 + X_{12}$	0.0679	0.67
2	48	m_1	$X_4 + X_5$	0.0453	1.00
3	36,912	m_{11}	$X_4 + X_5 + X_{12} + X_{15}$	0.0252	1.80
4	6,192	m_{12}	$X_4 + X_5 + X_{11} + X_{12}$	0.0176	2.57
5	32,816	m_3	$X_4 + X_5 + X_{15}$	0.0158	2.87
6	4,208	m_{13}	$X_4 + X_5 + X_6 + X_{12}$	0.0118	3.84
7	12,336	m_{14}	$X_4 + X_5 + X_{12} + X_{13}$	0.0116	3.91
8	4,656	m_{15}	$X_4 + X_5 + X_9 + X_{12}$	0.0115	3.94
9	4,272	m_{16}	$X_4 + X_5 + X_7 + X_{12}$	0.0114	3.97
10	5,168	m_{17}	$X_4 + X_5 + X_{10} + X_{12}$	0.0112	4.04

ISA short courses *Bayesian Variable Selection – An Introductory Tutorial* 98

Bayesian Variable Selection Tutorial

7. Bayesian variable selection Using WinBUGS

Rank	m	m_k	Model	$f(m\|\boldsymbol{y})$	$\text{PO}_{m_1\, m_k}$
			Prior 3[c]		
1	4,144	m_2	$X_4 + X_5 + X_{12}$	0.1014	0.88
2	48	m_1	$X_4 + X_5$	0.0896	1.00
3	36,912	m_{11}	$X_4 + X_5 + X_{12} + X_{15}$	0.0312	2.87
4	32,816	m_3	$X_4 + X_5 + X_{15}$	0.0277	3.23
5	6,192	m_{12}	$X_4 + X_5 + X_{11} + X_{12}$	0.0207	4.33
6	4,656	m_{15}	$X_4 + X_5 + X_9 + X_{12}$	0.0151	5.93
7	560	m_4	$X_4 + X_5 + X_9$	0.0142	6.31
8	5,168	m_{17}	$X_4 + X_5 + X_{10} + X_{12}$	0.0138	6.49
9	4,208	m_{13}	$X_4 + X_5 + X_6 + X_{12}$	0.0136	6.59
10	112	m_5	$X_4 + X_5 + X_6$	0.0133	6.74

ISA short courses *Bayesian Variable Selection – An Introductory Tutorial* 99

Bayesian Variable Selection Tutorial

7. Bayesian variable selection Using WinBUGS

Posterior model odds in reduced space

(variables with posterior inclusion prob > 0.2)

Vars 4, 5, 12 and 15 (in prior set-ups 2 & 3)

		Posterior model probability		
Model	m	Prior 1[a]	Prior 2[b]	Prior 3[c]
$X_4 + X_5$	4	0.6505	0.2987	0.3503
$X_4 + X_5 + X_{12}$	8	0.3265	0.4338	0.4118
X_5	3	0.0127	0.0017	0.0013
$X_5 + X_{12}$	7	0.0102	0.0025	0.0035
$X_4 + X_5 + X_{15}$	12	–	0.1032	0.1055
$X_4 + X_5 + X_{12} + X_{15}$	16	–	0.1568	0.1239

100

Bayesian Variable Selection Tutorial

7. Bayesian variable selection Using WinBUGS

Posterior model odds in reduced space

(variables with posterior inclusion prob > 0.2)

		Posterior model odds[d]		
Model	m	Prior 1[a]	Prior 2[b]	Prior 3[c]
$X_4 + X_5$	4	1.00	1.00	1.00
$X_4 + X_5 + X_{12}$	8	1.99	0.69	0.85
X_5	3	51.22	175.71	269.46
$X_5 + X_{12}$	7	63.77	119.48	100.09
$X_4 + X_5 + X_{15}$	12	–	2.89	3.32
$X_4 + X_5 + X_{12} + X_{15}$	16	–	1.90	2.83

[d]Each model is compared with the true model $X_4 + X_5$

Bayesian Variable Selection Tutorial

7. Bayesian variable selection Using WinBUGS

SSVS

Change in the likelihood (do not use gammas)

Change in the prior (use the first prior set-up illustrated in GVS with independent Normal priors)

KM

Change in the prior (use directly a multivariate prior on β)

For details see Dellaportas et al. (2000, *BGLM*) and Ntzoufras (2002, *JSS*)

Bayesian Variable Selection Tutorial

8. Model Search when the marginal likelihood is available

Markov Chain Monte Carlo Model Composition (MC^3)

Madigan and York (1995, Int. Stat. Rev.) for graphical models

Characteristics

- Marginal likelihood must be analytically available
- Can be used for model search if the model space is large

  e.g. variable selection with

  100 covariates $\Rightarrow$ **2^{100}=1.27×10^{30}** possible models

  30 covariates $\Rightarrow$ **2^{30} = 1,073,741,824** (≈1 billion) models

Bayesian Variable Selection Tutorial

8. Model Search when the marginal likelihood is available (MC^3)

The Algorithm

If the current model is *m* (or γ)

- Propose new model *m'* with probability *j(m,m')* which is the probability of proposing model *m'* when we are currently in model *m*.

  Usually *j(m,m')* is restricted to a neighborhood of the current model (e.g. adding/deleting one variable).
- Accept the proposed move with probability

$$\alpha = \min\left(1, \frac{f(m'|\boldsymbol{y})j(m',m)}{f(m|\boldsymbol{y})j(m',m)}\right) = \min\left(1, PO_{m',m}\frac{j(m',m)}{j(m',m)}\right)$$

Bayesian Variable Selection Tutorial

8.1 MC³ for variable selection

Variations of the algorithm

- Usual nb(m) → change the status of one covariate (i.e. add/delete)

  $j(m,m')$ = P(select one covariate)P(change the selected covariate)
- Use $\boldsymbol{\gamma}$ instead of m.
- Update all γ_j using random scan

  $j(m,m') \rightarrow j(\gamma_j, \gamma_j')$
- Gibbs variant can be used instead (see Smith & Kohn, 1996, *J. Econometrics*)
- But setting $j(\gamma_j, \gamma_j'=1-\gamma_j)=1$ [i.e. always propose to change] is optimal according to Liu (1996, *Bka*)

Bayesian Variable Selection Tutorial

8.1 MC³ for variable selection

Proposed Algorithm

If the current model is $\boldsymbol{\gamma}$

1. For j=1,…,p (order can be set by a random permutation)

   a. Set $\gamma_j'=1-\gamma_j$ *(i.e. propose to change status with prob. 1)*

   b. Accept the proposed move with probability

$$\alpha = \min\left(1, \frac{f(\gamma_j', \gamma_{\setminus j} | \boldsymbol{y})}{f(\gamma_j, \gamma_{\setminus j} | \boldsymbol{y})}\right)$$

2. Save the current model status $\boldsymbol{\gamma}$
3. Return to 1 until the required number of iterations is achieved.

Bayesian Variable Selection Tutorial

8.1 MC[3] for variable selection

Normal models

- With conjugate NIG Prior

$$f(\boldsymbol{\beta}_m|\sigma^2, m) \sim N\left(\boldsymbol{\mu}_{\beta_m}, c^2 \boldsymbol{V}_m \sigma^2\right) \;\; f(\sigma^2) \sim \mathrm{IG}(a, b)$$

- Posterior model probability

Marginal likelihood = Multivariate Student

parameters in linear predictor

$$f(m|\boldsymbol{y}) = \boxed{constant \times c^{\boxed{-P_m}} \left(\frac{|\widetilde{\boldsymbol{\Sigma}}_m|}{|\boldsymbol{V}_m|}\right)^{1/2} \left(b + \frac{SS_m}{2}\right)^{-a-n/2}} f(m)$$

$$\mathrm{SS}_m = \boldsymbol{y}^T\boldsymbol{y} - \widetilde{\boldsymbol{\beta}}_m^T \widetilde{\boldsymbol{\Sigma}}_m^{-1} \widetilde{\boldsymbol{\beta}}_m + c^{-2} \boldsymbol{\mu}_{\beta_m}^T \boldsymbol{V}_m^{-1} \boldsymbol{\mu}_{\beta_m}$$

Posterior Sum of Squares

$$\widetilde{\boldsymbol{\beta}}_m = \widetilde{\boldsymbol{\Sigma}}_m \left(\boldsymbol{X}_m^T \boldsymbol{X}_m \widehat{\boldsymbol{\beta}}_m + c^{-2} \boldsymbol{V}_m^{-1} \boldsymbol{\mu}_{\beta_m}\right),$$

Posterior mean

$$\widetilde{\boldsymbol{\Sigma}}_m^{-1} = \boldsymbol{X}_m^T \boldsymbol{X}_m + c^{-2} \boldsymbol{V}_m^{-1},$$

Proportional to posterior precision

107

Bayesian Variable Selection Tutorial

8.1 MC[3] for variable selection

Normal models

- With conjugate NIG Prior – Zellner's g-prior

$$f(\boldsymbol{\beta}_m|\sigma^2, m) \sim N(\boldsymbol{0}, c^2 (\boldsymbol{X}_m^T \boldsymbol{X})^{-1} \sigma^2)$$

- Posterior model probability

$$f(m|\boldsymbol{y}) = constant \times (c^2+1)^{-P_m/2} \left(b + \frac{SS_m}{2}\right)^{-a-n/2} f(m)$$

$$SS_m = \boldsymbol{y}^T\boldsymbol{y} - \frac{c^2}{c^2+1} \widehat{\boldsymbol{\beta}}_m^T (\boldsymbol{X}_m^T \boldsymbol{X}_m) \widehat{\boldsymbol{\beta}}_m$$

Posterior Sum of Squares

6/12/2011@ University College Dublin
ISA short courses

Ioannis Ntzoufras
Bayesian Variable Selection – An Introductory Tutorial

108

Bayesian Variable Selection Tutorial

8.1 MC³ for variable selection

Normal models

- With conjugate NIG Prior – Zellner's g-prior
- For g large and a=b=0

$$SS_m \approx \boldsymbol{y}^T\boldsymbol{y} - \widehat{\boldsymbol{\beta}}_m^T(\boldsymbol{X}_m^T\boldsymbol{X}_m)\widehat{\boldsymbol{\beta}}_m = RSS_m$$ Residual Sum of Squares

$$-2\log f(m|\boldsymbol{y}) = constant + n\log(RSS_m) + P_m\log(c^2+1) - 2\log f(m)$$

DOES THIS REMINDS YOU SOMETHING?

$$BIC_m = -2\log LR + d_m \log n$$ (in our notation) ⇒

$$BIC_m = constant + n\log RSS_m + P_m \log n$$

Bayesian Variable Selection Tutorial

8.1 MC³ for variable selection

Normal models

- With conjugate NIG Prior – Zellner's g-prior
- For **g=n large** and **a=b=0**

  we end up to BIC

  n is substituted by n+1 because we have information equal to n+1 data points due to the unit information prior (see Kass and Wasserman, 1995, *JASA*)

Bayesian Variable Selection Tutorial

8.1 MC^3 for variable selection

Normal Models

See Hoeting et al. (1996, CSDA) Raftery et al. (1997, JASA)

Other GLM

- Use Laplace approximation which works reasonably well for these models
- See Raftery (1996, *Bka*)

R package

BMA by Raftery, Hoeting, Volinsky, Painter & Yeung

http://cran.r-project.org/web/packages/BMA/index.html

Bayesian Variable Selection Tutorial

9. Reversible Jump MCMC (RJMCMC)

Characteristics

- Metropolis type algorithm – adjusted to account for comparing models with different dimensions
- Fashionable
- Flexible – can handle almost all types of models
- Extremely difficult to apply in some cases
- Generally easy to apply in variable selection

Bayesian Variable Selection Tutorial

9. Reversible Jump MCMC (RJMCMC)

The algorithm

- Propose a new model m' with probability $j(m, m')$.
- Generate $\boldsymbol{u}$ from a specified proposal density $q(\boldsymbol{u}|\boldsymbol{\beta}_m, m, m')$.
- Set $(\boldsymbol{\beta}'_{(m')}, \boldsymbol{u}') = h_{m,m'}(\boldsymbol{\beta}_m, \boldsymbol{u})$
- Accept the proposed move to model m' with probability

$$\alpha = \min\left(1, \frac{f(\boldsymbol{y}|\boldsymbol{\beta}'_{m'}, m')f(\boldsymbol{\beta}'_{m'}|m')f(m')j(m', m)q(\boldsymbol{u}'|\boldsymbol{\beta}'_{m'}, m', m)}{f(\boldsymbol{y}|\boldsymbol{\beta}_m, m)f(\boldsymbol{\beta}_m|m)f(m)j(m, m')q(\boldsymbol{u}|\boldsymbol{\beta}_m, m, m')}\left|\frac{\partial h(\boldsymbol{\beta}_m, \boldsymbol{u})}{\partial(\boldsymbol{\beta}_m, \boldsymbol{u})}\right|\right)$$

Bayesian Variable Selection Tutorial

9.1 RJMCMC for Variable Selection

- Select randomly a variable j and propose it to change.
- If $\gamma_j = 0 \rightarrow \gamma'_j = 1$ (include a new covariate X_j in the model)
  - Propose $\beta'_j \sim q(\beta_j|\boldsymbol{\beta}_\gamma)$
  - Set $\beta'_k = \beta_k$ for all k with $\gamma_k = 1$ $(k \neq j)$.
  - Accept the proposed move with probability

$$\alpha = \min(1, O_j) \text{ with } O_j = \frac{f(\boldsymbol{y}|\boldsymbol{\beta}'_{\gamma'}, \gamma')f(\boldsymbol{\beta}'_{\gamma'}|\gamma')f(\gamma')}{f(\boldsymbol{y}|\boldsymbol{\beta}_\gamma, \gamma)f(\boldsymbol{\beta}_\gamma|\gamma)f(\gamma)q(\beta_j|\boldsymbol{\beta}_\gamma)}.$$

- If $\gamma_j = 1 \rightarrow \gamma'_j = 0$ (exclude a covariate X_j from the model)
  - Set $\beta'_k = \beta_k$ for all k with $\gamma'_k = 1$ $(k \neq j)$.
  - Accept the proposed move with probability

$$\alpha = \min(1, O_j^{-1}).$$

Bayesian Variable Selection Tutorial
9.1 RJMCMC for Variable Selection

Characteristics

- Jacobian equal to one
- Proposal parameters **u** are equal to the additional coefficients needed.
- **u'** is not needed (to achieve equality of dimensions)
- Very simple to use
- Efficient when no highly correlated/collinear covariates exist
- Proposals can be defined as in GVS

Bayesian Variable Selection Tutorial
9.1 RJMCMC for Variable Selection

RJMCMC+GVS

If we Metropolize GVS ⇒RJMCMC step with the proposal = pseudo-prior when

$$f(\boldsymbol{\beta}_{\setminus\gamma}|\boldsymbol{\beta}_{\gamma},\boldsymbol{\gamma}) = \prod_{j=1}^{p} f(\beta_j|\boldsymbol{\beta}_{\gamma},\boldsymbol{\gamma})^{1-\gamma_j}$$

Bayesian Variable Selection Tutorial

9.2 Independence sampler

- Propose a new model m' with probability $j(m, m')$.
- Generate all the parameters of the new model from a proposal $q(\boldsymbol{\beta}'_{m'}|m')$ which does not depends on the current model m.
- Accept the proposed move to model m' with probability

$$\alpha = \min\left(1, \frac{f(\boldsymbol{y}|\boldsymbol{\beta}'_{m'}, m')f(\boldsymbol{\beta}'_{m'}|m')f(m')j(m', m)q(\boldsymbol{\beta}_m|m)}{f(\boldsymbol{y}|\boldsymbol{\beta}_m, m)f(\boldsymbol{\beta}_m|m)f(m)j(m, m')q(\boldsymbol{\beta}'_{m'}|m')}\right).$$

Bayesian Variable Selection Tutorial

9.2 Independence sampler

Characteristics

- Jacobian equal to one
- Proposal parameters $\boldsymbol{u}=\boldsymbol{\beta}'_{\gamma'}$ and $\boldsymbol{u}'=\boldsymbol{\beta}_{\gamma}$
- Simple to use
- Efficient when no highly correlated/collinear covariates exist
- Proposals should be close to the corresponding posteriors
- Efficient approximations or MLEs can be used for proposals

Bayesian Variable Selection Tutorial

9.2 Independence sampler

Carlin+Chib and RJMCMC

Metropolize model selection step of Carlin and Chib (1995) method ⇒ Independence RJMCMC

MC³ and RJMCMC

RJMCMC proposals of β_m = *posterior distributions* ⇒ MC³

see Dellaportas et al. (2002, *Stat. & Comp.*)

Bayesian Variable Selection Tutorial

9.3 RJMCMC in WinBUGS

WiNBUGS **`jump`** interface

- recently developed by Dave Lunn
- For variable selection & spline models.

Available at the WinBUGS development site:
http://www.winbugs-development.org.uk/main.html

see Lunn *et al.* (2008, *Stat. & Comp.*)
Lunn *et al.* (2006, *Gen. Epidem.*) and
the interface's manual and examples.

Bayesian Variable Selection Tutorial

10. More advanced methods for variable selection

Population based RJMCMC
    generate multiple chains with different temperature and mobility
    see Jasra et al. (2007, *JASA*)
Moves based on genetic algorithms
    Exchange, Crossover, Snooker Jumps, Partitioning
    see Jasra et al. (2007, *Bka*), Goswami & Liu (2007, *Stat. & Comp.*)
Moves based on spatial moves on the model space
    see Nott & Green (2004, *JCGS*, Normal models)
    Nott & Leonte (2004, *JCGS*, GLMs)
Adaptive sampling
    Nott & Kohn (2005, *Bka*)

Bayesian Variable Selection Tutorial

11. Bayesian Lasso and Shrinkage Methods

- They became quite popular during the last decade
- They try to over-shrink small coefficients and leave unaffected (as much as possible) large ones
- It is actually a different use of priors (Double exponential for Lasso and more general for extensions)
- Posterior mode = Lasso estimates

Bayesian Variable Selection Tutorial

11. Bayesian Lasso and Shrinkage Methods

Advantages

- Lasso estimates (i.e. posterior mode) are set equal to zero for small (non-important) coefficients
- So it can be directly implemented on the full model

Disadvantages

- Posterior means and medians (which are more frequently used in Bayesian framework) do not have this "nice" property
- Do not quantify model uncertainty

Bayesian Variable Selection Tutorial

11. Bayesian Lasso and Shrinkage Methods

- They do not solve the Bartlett-Lindley Paradox
- Therefore it is not easy to define the shrinkage parameter (proportional to the prior precision)
- Use of Hyper-priors leading to extensions

Bayesian Variable Selection Tutorial

11. Bayesian Lasso and Shrinkage Methods

Some key-note references

[The original publication introducing lasso]

- **Tibshirani (1996)**. Regression shrinkage and selection via the lasso. *JRSSB*, **58**:267–288.

[Bayesian lasso]

- **Yuan and Lin (2005)**. Efficient empirical Bayes variable selection and estimation in linear models. *JASA*, **100**:1215–1225.
- **Park & Casella (2008)**. The Bayesian lasso. *JASA*, **103**:681–687.
- **Hans (2009)**. Bayesian Lasso regression. *Biometrika*, **96**:835–845.
- **Hans (2010)**. Model uncertainty and variable selection in Bayesian lasso regression. *Statistics and Computing*, **20**:221–229.

Bayesian Variable Selection Tutorial

11. Bayesian Lasso and Shrinkage Methods

Some key-note references (cont.)

[Shrinkage methods and extensions of lasso]

- **Carvalho, Polson & Scott (2010)**. The horseshoe estimator for sparse signal. *Biometrika*, **97**:465–480.
- **Griffin & Brown (2010)**. Inference with normal-gamma prior distributions in regression problems. *Bayesian Analysis*, **5**:171–188.
- **Scheipl (2010)**. Normal-mixture-of-inverse-gamma priors for Bayesian regularization and model selection in structured additive regression models. *Technical Report*; available at http://epub.ub.uni-muenchen.de/11785/.

Bayesian Variable Selection Tutorial

11. Bayesian Lasso and Shrinkage Methods

Our work on the topic

- **Lykou & Ntzoufras (2011).** On Bayesian Lasso Variable Selection and the Specification of the Shrinkage Parameter. (submitted);

  available at http://stat-athens.aueb.gr/~jbn/papers/paper25.htm (R code is also available)

  [focuses on variable selection using lasso and the selection of λ and its hyper-prior]
- **Lykou & Ntzoufras (2011).** WinBUGS: A Tutorial, *WIREs Computational Statistics* **3**, 385–396.

  [the illustration is an implementation of Bayesian Lasso (including variable selection) on WinBUGS]

Bayesian Variable Selection Tutorial

11. Bayesian Lasso and Shrinkage Methods

R Packages

- lars for common lasso variable selection
- spikeSlabGAM: implements the approach of Scheipl (2010)
- Monomvn: implements the approach of Griffin and Brown (2010) and the Horse-shoe prior of Carvalho et al. (2010)

  [on the full model (no direct variable selection) and using variable selection via RJMCMC]

Bayesian Variable Selection Tutorial

12. Other methods

(Described in the book)

- Using posterior predictive densities for model evaluation
  - Posterior Bayes Factors (Aitkin, *JRSSB*, 1991)
  - Pseudo Bayes Factors (Geisser and Eddy, 1979, *JASA*)
  - Negative cross-validatory log-likelihood (Spiegelhalter et al., 1996*a*, p. 42)
- Information criteria (BIC, AIC, Other)
- DIC - Deviance Information Criterion

  Spiegelhalter *et al.* (2002, *RSSB*)

  Stepwise procedure using WiNBUGS

Bayesian Variable Selection Tutorial

12. Other methods

(not in the book)

- Fractional Bayes Factor
- Intrinsic Bayes Factor

Bayesian Variable Selection Tutorial

12. Other issues

TO ADD

- Using posterior predictive densities for model evaluation
  Estimation from an MCMC output, simple example in WinBUGS
- Information criteria (BIC, AIC, other)
- Deviance Information Criterion
  Stepwise method in WinBUGS
- Calculation of penalized deviance measures from the MCMC output
- Implementation in WinBUGS

Bayesian Variable Selection Tutorial

13. Closing remarks

- Variable selection is a wide topic (this presentation is not exhaustive – just a introduction)
- Posterior odds – Bayes Factors are the main measures
- BMA is also important tool
- Be careful on the prior specification
- Start from Gibbs methods
- Try to use RJMCMC in the 2nd step (more efficient and more fashionable => i.e. publication in a good journal)
- **PROBLEM OF THE DECADE:** *Large p – small n problem*
  How to handle problems with large number of covariates and small number of observations