Day 5: Bayesian Modelling in R – Part 4
Bayesian variable selection using BAS

Ioannis Ntzoufras

Associate Professor in Statistics, e-mail: ntzoufras@aueb.gr

Bayesian Variable Selection Tutorial
table of contents (1)

1. Prior predictive distributions as measures of model comparison: Posterior model odds and Bayes factors
2. Sensitivity of the posterior model probabilities: The Lindley–Bartlett paradox
3. Prior distributions for variable selection in GLM (G-prior, Hyper-g prior)
4. Bayesian variable Selection (Inclusion probabilities, MAP model, Median probability Model)
5. BAS Package

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.

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.

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

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

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

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

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

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

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.

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

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

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_θ^2/n instead of σ_θ^2.
- The specification of σ_θ^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

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

Marginal likelihood is analytically available

Main problem ⇒ 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

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

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

$c^2=n$ ⇒ 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*)

$$\boldsymbol{\beta}_m | m \sim N\left(\widehat{\boldsymbol{\beta}}_m, n[\mathcal{I}(\widehat{\boldsymbol{\beta}}_m)]^{-1}\right)$$

MLEs

Observed Fisher information matrix

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 ***y****=0 and $\mathbf{X}_m^*=\mathbf{X}_m$ ⇒ 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

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

Bayesian Variable Selection Tutorial

3. Priors for Bayesian Variable Selection in GLM

Some comments

Normal priors ⇒ ridge regression type of shrinkage

Double exponential priors ⇒ 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* $\boldsymbol{y}$* *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 $\boldsymbol{\gamma}=(\gamma_1, \gamma_2, \ldots, \gamma_p)$ [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.

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 *$f(\theta_m|\mathbf{y},m)$ evaluated at the posterior mode*

Works reasonably well for GLMs.

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)

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).

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 $\Rightarrow$ 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 $\Rightarrow$ 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 $m^{(t)}$!!!

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

$I(\cdot)$: 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 [MC^3] (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 $\boldsymbol{\gamma}=(\gamma_1, \gamma_2, ..., \gamma_p)$ [George & McCulloch, 1993, *JASA*]

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

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

Use binary system and calculate m using the equation

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

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.*)

Bayesian Variable Selection Tutorial

7. Mixtures of g-priors for regression models

A simple idea:

use g-prior with a hyper-prior on g

First prior with this idea:

Zellner and Siow (1980) $\Rightarrow$ multivariate Cauchy priors

Cauchy can be written as a scale mixture of normals

$$\pi(\boldsymbol{\beta}_\gamma|\phi) \propto \int N\left(\boldsymbol{\beta}_\gamma \,\middle|\, \mathbf{0}, \frac{g}{\phi}(\mathbf{X}_\gamma^T\mathbf{X}_\gamma)^{-1}\right)\pi(g)\,dg,$$

with

$$\pi(g) = \frac{(n/2)^{1/2}}{\Gamma(1/2)} g^{-3/2} e^{-n/(2g)}.$$

i.e. is like putting an

inverse Gamma(1/2, n/2) hyper-prior on g

Bayesian Variable Selection Tutorial

7. Mixtures of g-priors for regression models

Hyper-g prior

- Covariates are centered
- Intercept is treated separated from covariate effects $\mathcal{M}_\gamma: \quad \boldsymbol{\mu} = \mathbf{1}_n\alpha + \mathbf{X}_\gamma\boldsymbol{\beta}_\gamma$
- Improper (Jeffreys) prior is placed on the intercept and the error variance $p(\alpha, \phi|\mathcal{M}_\gamma) = \frac{1}{\phi},$
- g-structure is placed only on covariate effects $\boldsymbol{\beta}_\gamma|\phi, \mathcal{M}_\gamma \sim \mathrm{N}\left(\mathbf{0}, \frac{g}{\phi}(\mathbf{X}_\gamma^T\mathbf{X}_\gamma)^{-1}\right)$
- Hyper prior on g $\pi(g) = \frac{a-2}{2}(1+g)^{-a/2}, \qquad g > 0$

Bayesian Variable Selection Tutorial

7. Mixtures of g-priors for regression models

Hyper-g prior

w ~ Beta(1, a/2-1) with mean 2/a,

variance=4(a-2)/[a2(a+1)]

w=g/(g+1) is the shrinkage parameter

$$\Rightarrow \quad \pi(g) = \frac{a-2}{2}(1+g)^{-a/2}, \qquad g > 0$$

- ✓ proper for a>2
- ✓ a=2 reference prior and Jeffreys prior
- ✓ posterior is proper also for 1<a ≤2 but BFs are not available
- ✓ Proposed prior values: 2<a ≤4

Bayesian Variable Selection Tutorial

7. Mixtures of g-priors for regression models

Hyper-g prior properties

- ✓ Extremely robust for $2<a \leq 4$
- ✓ Lindley's paradox eventually appears for $a \rightarrow 2$
  [since $w \rightarrow 1$ and $g \rightarrow$ infinity]
- ✓ Increases uncertainty in model space resulting in
  ⇒[Extremely] Small probabilities for MAP and best models
  ⇒ Inclusion probabilities of non-influential covariates are inflated towards 0.5
  [while the remaining do not change]

Bayesian Variable Selection Tutorial

7. Mixtures of g-priors for regression models

Hyper-g prior computational properties

- ✓ Closed form expression for the posterior of g
- ✓ Closed form expresion for the posterior mean of g
- ✓ Posterior means are readily available for betas
  (= posterior mean of $w \times$ MLEs)
- ✓ Marginal likelihoods are not available
  [a normalizing constant from the improper prior on intercept & error term is not computable]
- ✓ BFs and posterior model probabilities are available
  [the unknown normalizing constant is common to all models]

Bayesian Variable Selection Tutorial

7. Mixtures of g-priors for regression models

Hyper-g prior computational properties

- ✓ Posterior samples can be obtained using a simple Monte Carlo scheme
  [generate g ~ strange univariate distribution, (a,b)~Normal, σ2~IG]
- ✓ Variable selection can be obtained by running a MC^3 type algorithm where
  [generate g ~ strange univariate distribution, γ's using MC^3moves]
- ✓ Pairwise BF's can be obtained by calculating one dimensional integrals (e.g. using the function integrate in R).

Bayesian Variable Selection Tutorial

7. Mixtures of g-priors for regression models

Hyper-g prior – Extension to GLMs

Daniel Sabanés Bové & Leonhard Held (2011). Hyper-g Priors for Generalized Linear Models, *Bayesian Analysis*, **6**, 387-410.

http://ba.stat.cmu.edu/journal/2011/vol06/issue03/sabanes.pdf

Bayesian Variable Selection Tutorial

7. Mixtures of g-priors for regression models

Hyper-g prior – Running the method

BAS package

- ✓ Implements g-prior, Zellner and Siow and hyper-g
- ✓ Uses full enumeration for small spaces and adaptive sampling (Clyde, Ghosh, and Littman, 2011, *Stats & Computing*)
- ✓ Very fast and easy to use
- ✓ GLMs => using Laplace Approximations

51

Bayesian Variable Selection Tutorial

8. The BAS package

Hyper-g prior – Running the method using BAS

Dellaportas et al. (2002) Simulated data (revisited 1)

- p=15 covariates simulated from N(0,1)
- n=50, 2^{15}= 32,768 models
- Independent Xs so MCMC easy to implement
- True model $Y_i \sim N(X_{i4} + X_{i5}, (2.5)^2)$ for $i = 1, 2, \ldots, 50$

52

Bayesian Variable Selection Tutorial

8. The BAS package

Number of models to sample.
If NULL, BAS will evaluate all for p ≤ 25

```
bas.lm(formula, data, n.models=NULL,
    prior="ZS-null",
    alpha=NULL,
    modelprior=uniform(),
    initprobs="Uniform",
    method="BAS",
    update=NULL,
    bestmodel = NULL,
    bestmarg = NULL, prob.local = 0.0, prob.rw=0.5,
    Burnin.iterations = NULL,
    MCMC.iterations = NULL,
    lambda = NULL, delta = 0.025)
```

Prior distribution for β.
Choices: "AIC", "BIC", "g-prior",
"ZS-null", "ZS-full",
"hyper-g", "hyper-g-laplace",
"EB-local", and "EB-global"

α for hyper-g
recommended value=3 or any in (2,4),
g for g-prior (g=n – UIP)

53

Bayesian Variable Selection Tutorial

8. The BAS package

```
bas.lm(formula, data, n.models=NULL,
    prior="ZS-null",
    alpha=NULL,
    modelprior=uniform(),
    initprobs="Uniform",
    method="BAS",
    update=NULL,
    bestmodel = NULL,
    bestmarg = NULL, prob.local = 0.0, prob.rw=0.5,
    Burnin.iterations = NULL,
    MCMC.iterations = NULL,
    lambda = NULL, delta = 0.025)
```

Prior on model space.
Choices: uniform (all models same prob)
Bernoulli(probs=.5)
beta.binomial(alpha=1.0, beta=1.0)
[puts beta hyper-prior on var. incl. probs leads to beta-binomial on model size default = uniform on p]

Number of iterations to discard in MCMC options
Number of iterations to run MCMC when MCMC options are used

54

Bayesian Variable Selection Tutorial

8. The BAS package

```
library(BAS)
res1<-bas.lm( y~., data=ex1, prior='hyper-g', alpha=3 )
res1
Call:
bas.lm(formula = y ~ ., data = ex1, prior = "hyper-g", alpha = 3)

 Marginal Posterior Inclusion Probabilities:
Intercept        X1        X2        X3        X4        X5        X6
   1.0000    0.2799    0.2939    0.2815    0.9610    0.9989    0.3280
       X7        X8        X9       X10       X11       X12       X13
   0.2896    0.2981    0.3443    0.3100    0.3579    0.6952    0.2982
      X14       X15
   0.2899    0.4559
```

Bayesian Variable Selection Tutorial

8. The BAS package

```
names(res1)
 [1] "probne0"      "which"        "logmarg"      "postprobs"
     "priorprobs"   "sampleprobs"  "mse"          "ols"
     "ols.se"       "shrinkage"    "size"         "R2"
     "namesx"       "n"            "prior"        "modelprior"
     "alpha"        "n.models"     "n.vars"       "Y"          "X"
     "mean.x"       "call"
```

- probne0 = Posterior variable inclusion probabilities
- which = list with the included vars for each model
- logmarg = vector with the log-marginal values
- mse = σ for each model (vector)
- ols, ols.se = ordinary least square estimates and the corresponding st.errors (list with one vector for each model)
- R2 = R^2 for each model (vector)
- shrinkage = vector of posterior means of g/(g+1) [different for each model] – in g-prior the same for all models and not mean

Bayesian Variable Selection Tutorial

8. The BAS package

>summary(res1, 10)

```
> summary(res1, 10)
      Intercept X1 X2 X3 X4 X5 X6 X7 X8 X9 X10 X11 X12 X13 X14 X15        BF PostProbs     R2 dim   logmarg
 [1,]         1  0  0  0  1  1  0  0  0  0   0   0   1   0   0   0 1.0000000    0.0115 0.5227   4 10.976638
 [2,]         1  0  0  0  1  1  0  0  0  0   0   0   1   0   0   1 0.5854705    0.0067 0.5408   5 10.441299
 [3,]         1  0  0  0  1  1  0  0  0  0   0   0   0   0   0   0 0.5695763    0.0065 0.4752   3 10.413776
 [4,]         1  0  0  0  1  1  0  0  0  0   0   1   1   0   0   0 0.4681850    0.0054 0.5356   5 10.217746
 [5,]         1  0  0  0  1  1  0  0  0  1   0   0   1   0   0   0 0.3371152    0.0039 0.5279   5  9.889308
 [6,]         1  0  0  0  1  1  0  0  0  0   1   0   1   0   0   0 0.3339303    0.0038 0.5276   5  9.879815
 [7,]         1  0  0  0  1  1  1  0  0  0   0   0   1   0   0   0 0.3229228    0.0037 0.5268   5  9.846296
 [8,]         1  0  0  0  1  1  0  0  0  0   0   0   1   1   0   0 0.3099909    0.0036 0.5259   5  9.805426
 [9,]         1  0  0  0  1  1  0  1  0  0   0   0   1   0   0   0 0.3008688    0.0035 0.5251   5  9.775557
[10,]         1  0  1  0  1  1  0  0  0  0   0   0   1   0   0   0 0.2969264    0.0034 0.5248   5  9.762367
>
```


Bayesian Variable Selection Tutorial

8. The BAS package

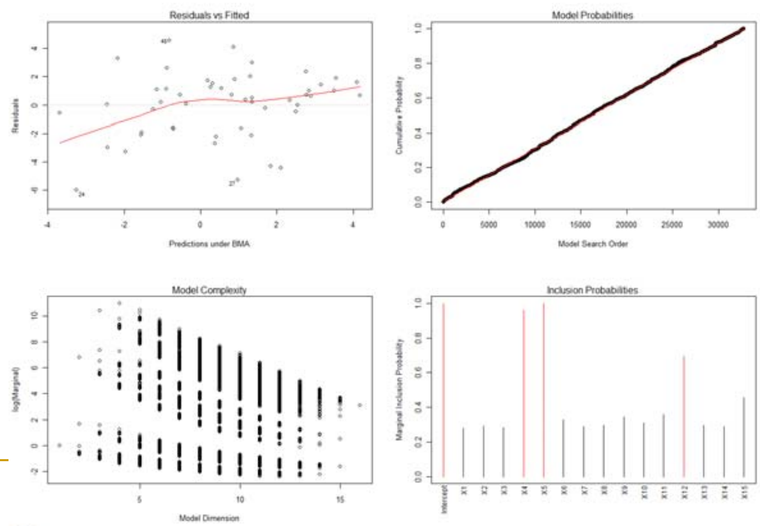
>coef(res1)

```
Marginal Posterior Summaries of Coefficients:
           post mean  post SD   post p(B != 0)
Intercept   0.599294  0.348005  1.000000
X1          0.001596  0.176758  0.279912
X2         -0.035679  0.223080  0.293890
X3          0.009791  0.181924  0.281518
X4          0.974055  0.377372  0.961014
X5          1.440046  0.297106  0.998908
X6          0.077719  0.227118  0.328045
X7         -0.030877  0.188952  0.289637
X8          0.048296  0.229017  0.298116
X9         -0.093995  0.244189  0.344321
X10        -0.068747  0.255255  0.309953
X11        -0.102938  0.244242  0.357932
X12        -0.639829  0.578404  0.695182
X13        -0.045241  0.208098  0.298160
X14         0.030421  0.220113  0.289938
X15        -0.180793  0.288275  0.455901
```

Extract conditional posterior means and standard deviations, marginal posterior means and standard deviations, posterior probabilities, and marginal inclusions probabilities under Bayesian Model Averaging from an object of class BMA

Bayesian Variable Selection Tutorial

8. The BAS package



59

Bayesian Variable Selection Tutorial

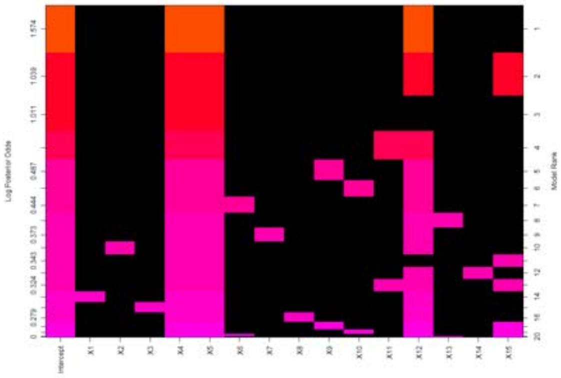
8. The BAS package

```
>predict(object, newdata, top=NULL, ...)
    BMA prediction based on top models
>fitted(object, type="HPM", top=NULL, ...)
    fitted values based on
    HPM = highest probability model (or map)
    MPM = median probability model of Barbieri & Berger
    BMA = Bayesian model averaging based on top models
>eplogprob(lm.obj, thresh=.5, max = 0.99, int=TRUE)
    rough approximation of inclusion probabilities from
    p-values (Sellke, Bayarri and Berger, 2001) BF(p) =
    -e p log(p)
    p( γ=1 | data ) ≈ 1/(1 + BF(p))
>update(object, newprior, alpha=NULL, ...)
    update object with different prior on β
```

Bayesian Variable Selection Tutorial

8. The BAS package

```
>image(res1, top.models=20, intensity=TRUE, prob=TRUE,
    log=TRUE, rotate=TRUE, color="rainbow",
    subset=NULL, offset=.75, digits=3,
    vlas=2,plas=0,rlas=0, ...)
```



Bayesian Variable Selection Tutorial

8.1. Implementing hyper-g in WinBUGS

Comparison with WinBUGS (g-prior of Liang etal)

Node statistics

node	mean	sd
g[1]	0.1257	0.3315
g[2]	0.1387	0.3456
g[3]	0.1247	0.3304
g[4]	0.9719	0.1654
g[5]	1.0	0.0
g[6]	0.1656	0.3717
g[7]	0.1261	0.332
g[8]	0.1319	0.3383
g[9]	0.1657	0.3718
g[10]	0.1387	0.3456
g[11]	0.1873	0.3901
g[12]	0.5881	0.4922
g[13]	0.1346	0.3413
g[14]	0.1317	0.3382
g[15]	0.2599	0.4386
pmodel[1]	0.06314	0.2432
pmodel[2]	0.07657	0.2659

```
> coef(res2)

 Marginal Posterior Summaries of Coefficients:
            post mean   post SD   post p(B != 0)
Intercept    0.599294  0.344610   1.000000
X1           0.003345  0.126433   0.124563
X2          -0.013867  0.161411   0.132140
X3           0.007888  0.131240   0.126307
X4           1.151358  0.387495   0.975430
X5           1.679273  0.305133   0.999967
X6           0.041779  0.175736   0.157807
X7          -0.021892  0.144361   0.135966
X8           0.018042  0.158313   0.131546
X9          -0.047732  0.185875   0.165307
X10         -0.035394  0.191088   0.144826
X11         -0.064665  0.204041   0.190233
X12         -0.629013  0.647779   0.592987
X13         -0.024481  0.153329   0.138017
X14          0.015289  0.155662   0.130470
X15         -0.121004  0.264046   0.270164
```

Bayesian Variable Selection Tutorial

8.1. Implementing hyper-g in WinBUGS

Comparison with WinBUGS (g-prior of Liang etal)

Node statistics

node	me
g[1]	0.12
g[2]	0.13
g[3]	0.12
g[4]	0.97
g[5]	1.0
g[6]	0.16
g[7]	0.12
g[8]	0.13
g[9]	0.16
g[10]	0.13
g[11]	0.18
g[12]	0.58
g[13]	0.13
g[14]	0.13
g[15]	0.25
pmodel[1]	0.06
pmodel[2]	0.07

node	mean
g[1]	0.1257
g[2]	0.1387
g[3]	0.1247
g[4]	0.9719
g[5]	1.0
g[6]	0.1656
g[7]	0.1261
g[8]	0.1319
g[9]	0.1657
g[10]	0.1387
g[11]	0.1873
g[12]	0.5881
g[13]	0.1346
g[14]	0.1317
g[15]	0.2599

```
> coef(re
post p(B != 0)
 1.000000
 0.124563
 0.132140
 0.126307
 0.975430
 0.999967
 0.157807
 0.135966
 0.131546
 0.165307
 0.144826
 0.190233
 0.592987
 0.138017
 0.130470
 0.270164
```

```
es of Coefficients:
SD     post p(B != 0)
4610     1.000000
6433     0.124563
1411     0.132140
1240     0.126307
7495     0.975430
5133     0.999967
5736     0.157807
4361     0.135966
8313     0.131546
5875     0.165307
1088     0.144826
4041     0.190233
7779     0.592987
3329     0.138017
5662     0.130470
4046     0.270164
```

Bayesian Variable Selection Tutorial

8.1. Implementing hyper-g in WinBUGS

Comparison with WinBUGS (g-prior of Liang etal)

Node statistics

```
> summary(res2)
```

	Intercept	X1	X2	X3	X4	X5	X6	X7	X8	X9	X10	X11	X12	X13	X14	X15	BF	PostProbs
[1,]	1	0	0	0	1	1	0	0	0	0	0	0	1	0	0	0	1.0000000	0.0758
[2,]	1	0	0	0	1	1	0	0	0	0	0	0	0	0	0	0	0.7643910	0.0579
[3,]	1	0	0	0	1	1	0	0	0	0	0	0	1	0	0	1	0.3472619	0.0263
[4,]	1	0	0	0	1	1	[illegible]	[illegible]	[illegible]	[illegible]	0	1	1	0	0	0	0.2665822	0.0202
[5,]	1	0	0	0	1	1					0	0	0	0	0	1	0.2442257	0.0185

MC error

pmodel[1]	0.06314	0.2432	0.00349	0.0	0.0	1.0	4001	7000
pmodel[2]	0.07657	0.2659	0.004335	0.0	0.0	1.0	4001	7000
pmodel[1]	0.05738	0.2326	0.001667	0.0	0.0	1.0	4001	37000
pmodel[2]	0.07424	0.2622	0.001816	0.0	0.0	1.0	4001	37000
g[12]	0.5881	0.4922	0.008462	0.0	1.0	1.0	4001	7000
g[13]	0.1346	0.3413	0.006309	0.0	0.0	1.0	4001	7000
g[14]	0.1317	0.3382	0.004261	0.0	0.0	1.0	4001	7000
g[15]	0.2599	0.4386	0.006655	0.0	0.0	1.0	4001	7000
pmodel[1]	0.06314	0.2432	0.00349	0.0	0.0	1.0	4001	7000
pmodel[2]	0.07657	0.2659	0.004335	0.0	0.0	1.0	4001	7000

Bayesian Variable Selection Tutorial

8.1. Implementing hyper-g in WinBUGS

Comparison with WinBUGS (hyper-g prior with a=3)

Node statistics

node	mean	MC error	post p(B != 0)	
g[1]	0.2817	0.005282	1.000000	0.2817
g[2]	0.2861	0.004893	0.279912	0.2861
g[3]	0.2869	0.005552	0.293890	0.2869
g[4]	0.9651	0.003451	0.281518	0.9651
g[5]	0.9993	4.324E-4	0.961014	0.9993
g[6]	0.2798	0.02826	0.998908	0.2798
g[7]	0.2817	0.006293	0.328045	0.2817
g[8]	0.2947	0.005244	0.289637	0.2947
g[9]	0.3499	0.007866	0.298116	0.3499
g[10]	0.3031	0.00522	0.344321	0.3031
g[11]	0.3555	0.005548	0.309953	0.3555
g[12]	0.6984	0.005719	0.357932	0.6984
g[13]	0.3159	0.008983	0.695182	0.3159
g[14]	0.2887	0.005414	0.298160	0.2887
g[15]	0.4453	0.006509	0.289938	0.4453
			0.455901	

Bayesian Variable Selection Tutorial

8.1. Implementing hyper-g in WinBUGS

Comparison with WinBUGS (hyper-g prior with a=3)

BMA estimates of shrinkage weight (w) and g parameter

Node statistics

node	mean	sd	MC error	2.5%	median	97.5%	start	sample
w	0.9394	0.03445	4.946E-4	0.8584	0.9451	0.99	4001	10000

```
> sum(res3$postprobs*res1$shrinkage)
[1] 0.9390098
```

Node statistics

node	mean	sd	MC error	2.5%	median	97.5%	start	sample
G	26.0	42.51	0.4788	6.061	17.22	98.69	4001	10000

```
> w<-res3$shrinkage
> w[w==1] <- 1-exp(-10)
> g<-w/(1-w)
> sum(res3$postprobs*g)
[1] 16.31460
```

66

Bayesian Variable Selection Tutorial

8.1. Implementing hyper-g in WinBUGS

Comparison with WinBUGS (hyper-g prior with a=3)

Posterior model probabilities for true and MAP

Node statistics

node	mean	sd	MC error	2.5%	median	97.5%	start	sample
pmodel[1]	0.0233	0.1509	0.001832	0.0	0.0	0.0	4001	10000
pmodel[2]	0.0328	0.1781	0.002429	0.0	0.0	1.0	4001	10000

```
> summary(res3)
     Intercept X1 X2 X3 X4 X5 X6 X7 X8 X9 X10 X11 X12 X13 X14 X15        BF PostProbs
[1,]         1  0  0  0  1  1  0  0  0  0   0   0   1   0   0   0 1.0000000    0.0324
[2,]         1  0  0  0  1  1  0  0  0  0   0   0   0   0   0   0 0.6370245    0.0206
```


Bayesian Variable Selection Tutorial

8.1. Implementing hyper-g in WinBUGS

Comparison with WinBUGS (Zellner and Siow prior)

Node statistics

node	mean	MC error	post p(B != 0)	
			1.000000	
g[1]	0.1936	0.004326	0.196936	0.1936
g[2]	0.2028	0.003998	0.209879	0.2028
g[3]	0.1942	0.004358	0.198653	0.1942
g[4]	0.9779	0.003131	0.978460	0.9779
g[5]	1.0	1.0E-12	0.999965	1.0
g[6]	0.2603	0.02607	0.242996	0.2603
g[7]	0.2033	0.004993	0.208557	0.2033
g[8]	0.2107	0.004358	0.210968	0.2107
g[9]	0.2509	0.006402	0.257782	0.2509
g[10]	0.2229	0.00485	0.226958	0.2229
g[11]	0.2715	0.004963	0.280207	0.2715
g[12]	0.6731	0.006764	0.678978	0.6731
g[13]	0.2114	0.006564	0.215078	0.2114
g[14]	0.2049	0.004487	0.206660	0.2049
g[15]	0.38	0.00687	0.379183	0.38

Bayesian Variable Selection Tutorial

8.1. Implementing hyper-g in WinBUGS

Comparison with WinBUGS (Zellner and Siow prior)

BMA estimates of shrinkage weight (w) and g parameter

Node statistics

node	mean	sd	MC error	2.5%	median	97.5%	start	sample
w	0.832	0.1044	0.002442	0.5701	0.8546	0.9683	4001	10000

```
> sum(res1$postprobs*res1$shrinkage)
[1] 0.8278971
```

Node statistics

node	mean	sd	MC error	2.5%	median	97.5%	start	sample
G	8.795	22.2	0.2725	1.326	5.88	30.51	4001	10000

```
> w<-res1$shrinkage
> w[w==1] <- 1-exp(-10)
> g<-w/(1-w)
> sum(res1$postprobs*g)
[1] 5.257485
```


Bayesian Variable Selection Tutorial

8.1. Implementing hyper-g in WinBUGS

Comparison with WinBUGS (Zellner and Siow prior)

Posterior model probabilities for true and MAP

Node statistics

node	mean	sd	MC error	2.5%	median	97.5%	start	sample
pmodel[1]	0.0051	0.07123	8.703E-4	0.0	0.0	0.0	4001	10000
pmodel[2]	0.0111	0.1048	0.001302	0.0	0.0	0.0	4001	10000

```
> summary(res1)
     Intercept X1 X2 X3 X4 X5 X6 X7 X8 X9 X10 X11 X12 X13 X14 X15        BF PostProbs
[1,]         1  0  0  0  1  1  0  0  0  0   0   0   1   0   0   0 1.0000000    0.0115
[2,]         1  0  0  0  1  1  0  0  0  0   0   0   1   0   0   1 0.5854705    0.0067
```

Bayesian Variable Selection Tutorial

9. 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
- **PROBLEM OF THE DECADE:** *Large p – small n problem*
  How to handle problems with large number of covariates and small number of observations

2009-2013, by Ioannis Ntzoufras
Department of Statistics, AUEB