Two-stage Bayesian variable selection for linear models using I-priors

Haziq Jamil
Supervisors: Dr. Wicher Bergsma & Prof. Irini Moustaki

Social Statistics (Year 2)
London School of Economics & Political Science

18 November 2015

Postgraduate Research Seminar

Outline

1. Introduction
   - Bayesian linear regression
   - Types of priors
2. ASIDE: Regression modelling using I-priors
   - Linear regression and motivation for I-priors
   - Examples of I-prior models
   - I-prior summary
3. Bayesian variable selection
   - Introduction
   - Bayesian variable selection methods
4. Using I-priors in Bayesian variable selection
   - Variable selection with I-priors
   - Simulation study
   - Two-stage procedure
   - g-priors
5. Summary

Bayesian linear regression

- Consider a linear regression model for n observations on p variables:

$$\begin{aligned} \mathbf{y} &= \boldsymbol{\alpha} + \mathbf{X}\boldsymbol{\beta} + \boldsymbol{\epsilon} \\ \boldsymbol{\epsilon} &\sim \mathrm{N}(\mathbf{0}, \psi^{-1}\mathbf{I}_n) \end{aligned} \tag{1}$$

  where $\boldsymbol{\alpha} = \alpha\mathbf{1}_n$.

- The OLS estimate for $\boldsymbol{\beta}$ is $\hat{\boldsymbol{\beta}} = (\mathbf{X}^T\mathbf{X})^{-1}\mathbf{X}^T\mathbf{y}$.

- The Bayesian approach supplements the data with additional information in the form of prior beliefs about the parameters:
  - $\alpha \sim \mathrm{N}(a, b)$
  - $\boldsymbol{\beta} \sim \mathrm{N}(\mathbf{c}, \mathbf{D})$
  - $\psi \sim \Gamma(e, f)$

- Inference on the parameters $\boldsymbol{\Theta}$ is through the posterior

$$f(\boldsymbol{\Theta}|\mathbf{y}) \propto \overbrace{f(\mathbf{y}|\boldsymbol{\Theta})}^{\text{likelihood}} \times \overbrace{f(\boldsymbol{\Theta})}^{\text{prior}}$$

Types of priors and the I-prior

- Priors can either be pure beliefs (subjective) or chosen according to some principle (objective). Either way, they can also be
  - Informative - has an impact on the results
  - Uninformative - provides little or vague information
  - Improper - may not be a proper distribution

Types of priors and the I-prior

- Priors can either be pure beliefs (subjective) or chosen according to some principle (objective). Either way, they can also be
  - Informative - has an impact on the results
  - Uninformative - provides little or vague information
  - Improper - may not be a proper distribution

- **I-priors (for regression coefficients)**
  An I-prior on $\boldsymbol{\beta}$ for the linear model in (1) is a distribution on $\boldsymbol{\beta}$ such that its covariance matrix is the Fisher information of $\boldsymbol{\beta}$. Also, assign a "best guess" on the prior mean, e.g. $\boldsymbol{\beta}_0 = \mathbf{0}$.

Types of priors and the I-prior

- Priors can either be pure beliefs (subjective) or chosen according to some principle (objective). Either way, they can also be
  - Informative - has an impact on the results
  - Uninformative - provides little or vague information
  - Improper - may not be a proper distribution

- **I-priors (for regression coefficients)**
  An I-prior on $\boldsymbol{\beta}$ for the linear model in (1) is a distribution on $\boldsymbol{\beta}$ such that its covariance matrix is the Fisher information of $\boldsymbol{\beta}$. Also, assign a "best guess" on the prior mean, e.g. $\boldsymbol{\beta}_0 = \mathbf{0}$.

- An objective and information theoretic prior for linear models with an intuitive appeal:

$$\uparrow \textit{Fisher information} \Rightarrow \uparrow \textit{variance} \Rightarrow \downarrow \textit{influence of prior mean.}$$

The I-prior linear regression model

$$\mathbf{y} = \boldsymbol{\alpha} + \mathbf{X}\beta + \boldsymbol{\epsilon}$$
$$\boldsymbol{\epsilon} \sim \text{N}(\mathbf{0}, \psi^{-1}\mathbf{I}_n)$$

- We know from linear regression theory that $I[\beta] = \psi\mathbf{X}^T\mathbf{X}$.

The I-prior linear regression model

$$\mathbf{y} = \boldsymbol{\alpha} + \mathbf{X}\boldsymbol{\beta} + \boldsymbol{\epsilon}$$
$$\boldsymbol{\epsilon} \sim \mathrm{N}(\mathbf{0}, \psi^{-1}\mathbf{I}_n)$$

- We know from linear regression theory that $I[\beta] = \psi\mathbf{X}^T\mathbf{X}$. An I-prior on $\boldsymbol{\beta}$ is then
$$\boldsymbol{\beta} \sim \mathrm{N}(\mathbf{0}, \lambda^2\psi\mathbf{X}^T\mathbf{X}).$$
- λ is introduced to resolve the scale of measurements of $\mathbf{X}$.
- Assumption: *All variables are measured on the same scale, or at least standardised. More on this later...*

The I-prior linear regression model

$$\mathbf{y} = \boldsymbol{\alpha} + \mathbf{X}\boldsymbol{\beta} + \boldsymbol{\epsilon}$$
$$\boldsymbol{\epsilon} \sim \mathrm{N}(\mathbf{0}, \psi^{-1}\mathbf{I}_n)$$

- We know from linear regression theory that $I[\beta] = \psi \mathbf{X}^T \mathbf{X}$. An I-prior on $\boldsymbol{\beta}$ is then
$$\boldsymbol{\beta} \sim \mathrm{N}(\mathbf{0}, \lambda^2 \psi \mathbf{X}^T \mathbf{X}).$$
- λ is introduced to resolve the scale of measurements of $\mathbf{X}$.
- Assumption: *All variables are measured on the same scale, or at least standardised. More on this later...*
- To complete the Bayesian model specification, set priors on the intercept and precision
$$\alpha \sim \mathrm{N}(0, 1000)$$
$$\psi \sim \Gamma(0.001, 0.001).$$

1 Introduction

2 ASIDE: Regression modelling using I-priors

3 Bayesian variable selection

4 Using I-priors in Bayesian variable selection

5 Summary

Linear regression

- Definition **(The linear regression model)**

$$y_i = f(x_i) + \epsilon_i$$

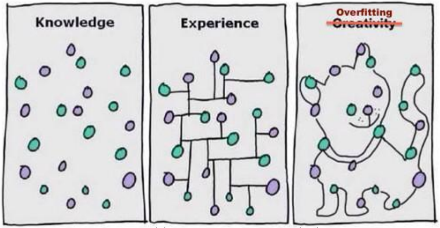
$$\begin{aligned} y_i \in \mathbb{R}, &\ \text{real-valued observations} \\ x_i \in \mathcal{X}, &\ \text{a set of characteristics for unit } i \\ f \in \mathcal{F}, &\ \text{a vector space of functions over the set } \mathcal{X} \\ (\epsilon_1, \ldots, \epsilon_n) &\sim \mathrm{N}(\mathbf{0}, \mathbf{\Psi}^{-1}) \\ i &= 1, \ldots, n \end{aligned} \tag{2}$$

Note: For iid observations, $\mathbf{\Psi} = \psi \mathbf{I}_n$. In general, $\mathbf{\Psi} = (\psi_{ij})$.

Motivation for I-priors: The issue of overfitting

- When dimensionality is large, maximum likelihood overfits. Solutions:
  - Dimension reduction
  - Random effects models
  - Regularization

  ...all require additional assumptions.
- I-priors require no assumptions other than those pertaining to the model of interest.
- But we do need a structural requirement for $\mathcal{F}$ in the form of an inner-product space (reproducing kernel Hilbert/Krein space).



credits: http://blog.sciencenet.cn/u/jerrycueb

Inner products

Functional vector spaces

Kernel methods

Reproducing kernels

Hilbert spaces

Gaussian random vectors

I-prior theory

Fisher Information

Krein spaces

Means of random functions

Feature maps

Variances of random functions

Moore-Aronszajn Theorem

Random functions

Definitions & theorem

- Theorem **(Gaussian I-priors)** [Bergsma, 2014]
  For the linear regression model (2), let $\mathcal{F}$ be the RKKS with kernel $h : \mathcal{X} \times \mathcal{X} \to \mathbb{R}$. Then, assuming it exists, the Fisher information for f is given by

$$I[f](x_i, x_i') = \sum_{k=1}^{n} \sum_{l=1}^{n} \psi_{kl} h(x_i, x_k) h(x_i', x_l).$$

  Let π be a Gaussian I-prior on f with prior mean f_0 and variance $I[f]$. Then π is called an I-prior for f, and a random vector $f \sim \pi$ has the random effect representation

$$f(x_i) = f_0(x_i) + \sum_{k=1}^{n} h(x_i, x_k) w_k$$

$$(w_1, \dots, w_n) \sim N(\mathbf{0}, \mathbf{\Psi}).$$

Back to the (standard) linear regression model

$$\mathbf{y} = \boldsymbol{\alpha} + \mathbf{X}\boldsymbol{\beta} + \boldsymbol{\epsilon}$$

$$\boldsymbol{\epsilon} \sim \mathrm{N}(\mathbf{0}, \psi^{-1}\mathbf{I}_n)$$

$$\boldsymbol{\beta} \sim \mathrm{N}(\boldsymbol{\beta}_0, \lambda^2\psi\mathbf{X}^T\mathbf{X})$$

Back to the (standard) linear regression model

$$\mathbf{y} = \overbrace{\boldsymbol{\alpha} + \mathbf{X}\boldsymbol{\beta}}^{\mathbf{f}} + \boldsymbol{\epsilon}$$
$$\boldsymbol{\epsilon} \sim \mathrm{N}(\mathbf{0}, \psi^{-1}\mathbf{I}_n)$$
$$\boldsymbol{\beta} \sim \mathrm{N}(\boldsymbol{\beta}_0, \lambda^2\psi\mathbf{X}^T\mathbf{X})$$

Equivalently,

$$\boldsymbol{\beta} = \boldsymbol{\beta}_0 + \lambda\mathbf{X}^T\mathbf{w}$$
$$\mathbf{w} \sim \mathrm{N}(\mathbf{0}, \psi\mathbf{I}_n).$$

Thus, an I-prior on $\mathbf{f}$ is

$$\mathbf{f} = \overbrace{\alpha + \mathbf{X}\boldsymbol{\beta}_0}^{\mathbf{f}_0} + \overbrace{\lambda\mathbf{X}\mathbf{X}^T\mathbf{w}}^{\mathbf{H}_\lambda\mathbf{w}}$$
$$\mathbf{w} \sim N(\mathbf{0}, \psi\mathbf{I}_n).$$

Toolbox of RKHS/RKKS

- Choose different $\{\mathcal{F}, h\}$ to suit type of data to model.

$\mathcal{X} = \{x_i\}$	Characteristic/Uses	Vector space $\mathcal{F}$	Kernel $h(x_i, x_k)$
Nominal	1) Categorical covariates; 2) In a multilevel setting, x_i = group no. of unit i.	Pearson	$\frac{\mathbb{I}[x_i = x_k]}{p_i} - 1$ where $p_i = \mathbb{P}[X = x_i]$
Real	As in classical regression, x_i = real-valued covariate associated with unit i.	Canonical	$x_i x_k$
Real	As in (1-dim) smoothing, x_i = data point associated with observation y_i.	Fractional Brownian Motion (FBM)	$\lvert x_i\rvert^{2\gamma} + \lvert x_k\rvert^{2\gamma} - \lvert x_i - x_k\rvert^{2\gamma}$ with $\gamma \in (0, 1)$
Nominal + Real	Used for random intercept/slope modelling.	ANOVA	Pearson + Canonical kernels

Example: Simple linear regression

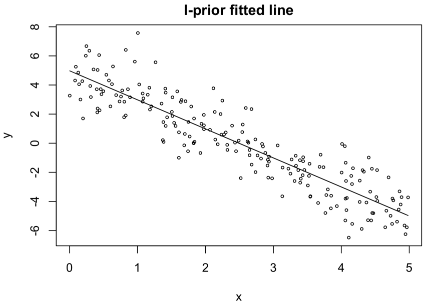
Classical model

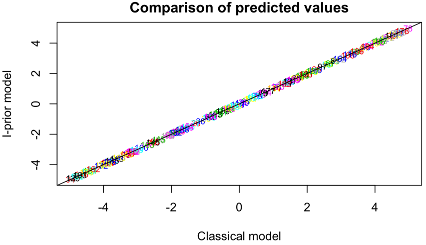
$$y_i = \beta_0 + \beta_1 x_i + \epsilon_i$$
$$\epsilon_i \sim N(0, \sigma^2)$$

I-prior model

$$y_i = \alpha + \sum_{k=1}^{n} h_\lambda(x_i, x_k) w_k + \epsilon_i$$
$$\epsilon_i \sim N(0, \psi^{-1})$$
$$w_i \sim N(0, \psi)$$

h_λ is the Canonical kernel





MSE(classical) = 1.770 MSE(I-prior) = 1.770

Example: 1-dimensional smoothing

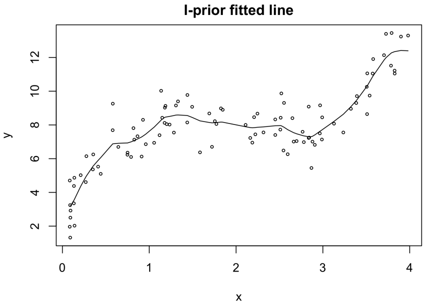
Classical model

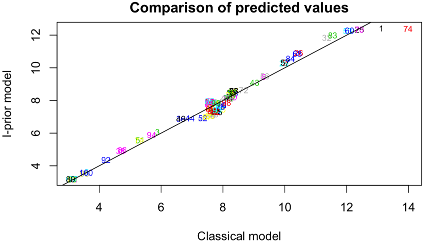
$$y_i = \beta_0 + \beta_1 x_i + \beta_2 x_i^2 + \beta_3 x_i^3$$
$$\epsilon_i \sim N(0, \sigma^2)$$

I-prior model

$$y_i = \alpha + \sum_{k=1}^{n} h_{\lambda,\gamma}(x_i, x_k) w_k + \epsilon_i$$
$$\epsilon_i \sim N(0, \psi^{-1})$$
$$w_i \sim N(0, \psi)$$

$h_{\lambda,\gamma}$ is the FBM kernel





MSE(classical) = 0.987 MSE(I-prior) = 0.836

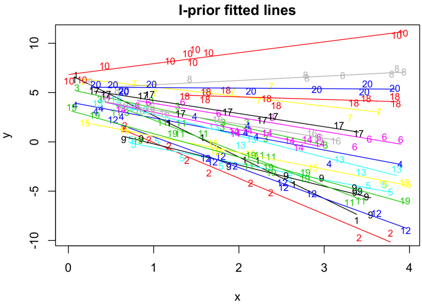
Example: Multilevel modelling

Classical model

$$y_{ij} = \beta_{0j} + \beta_{1j}x_{ij} + \epsilon_{ij}$$

$$\begin{pmatrix}\beta_{0j}\\ \beta_{1j}\end{pmatrix} \sim N\left(\begin{pmatrix}\beta_0\\ \beta_1\end{pmatrix}, \begin{pmatrix}\phi_0 & \phi_{01}\\ \phi_{01} & \phi_1\end{pmatrix} \right)$$

$$\epsilon_{ij} \sim N(0, \sigma^2)$$



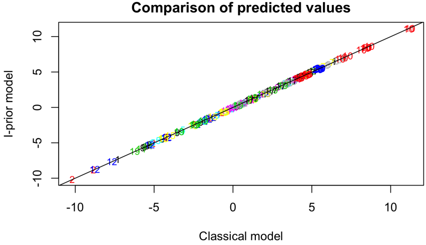
I-prior model

$$y_i = \alpha + \sum_{k=1}^{n} h_\lambda(x_i, x_k)w_k + \epsilon_i$$

$$\epsilon_i \sim N(0, \psi^{-1})$$

$$w_i \sim N(0, \psi)$$

h_λ is the ANOVA kernel



MSE(classical) = 0.227 MSE(I-prior) = 0.226

I-prior summary

- The I-prior methodology is a modelling technique that guards against overfitting linear models when dimensionality is large relative to sample size, with advantages such as
  - Model parsimony
  - Requires no additional assumptions
  - Simpler estimation (EM algorithm)
- Many models shown to work with using I-priors such as multiple regression, smoothing models, random effects models and growth curve models.
- Areas of research include
  - Extension to GLMs
  - Structural Equation Models
  - Models with structured error covariances
- Key idea: *Fisher information as the covariance matrix for priors*.

1 Introduction

2 ASIDE: Regression modelling using I-priors

3 Bayesian variable selection

4 Using I-priors in Bayesian variable selection

5 Summary

Model selection criteria

- Would like to search the entire model space to find the "best" model based on a certain criterion.
- Many methods for model selection criteria... (adjusted) R^2, AIC, BIC, Mallow's C_p, (k-fold) cross-validation, posterior model odds, Bayes factors, etc.
- When a large set of models to be compared, most tasks can be computationally prohibitive or even unfeasible.

Bayesian model evaluation

- It is believed that a set of data $\mathbf{Y}$ has been generated from the pdf $f(\mathbf{y}|m_k, \mathbf{\Theta}_k)$, where m_k is one of a set of $M = \{m_1, \ldots, m_K\}$ models.

Bayesian model evaluation

- It is believed that a set of data $\mathbf{Y}$ has been generated from the pdf $f(\mathbf{y}|m_k, \mathbf{\Theta}_k)$, where m_k is one of a set of $M = \{m_1, \dots, m_K\}$ models.
- As Bayesians do...
  - Assign priors $f(\mathbf{\Theta}_k|m_k)$ and $f(m_k)$
  - Compute the posterior

$$\begin{aligned} f(m_k|\mathbf{y}) &\propto f(\mathbf{y}|m_k)f(m_k) \\ &\propto \int f(\mathbf{y}|m_k, \mathbf{\Theta}_k)f(\mathbf{\Theta}_k|m_k)\,\mathrm{d}\mathbf{\Theta}_k\, f(m_k) \end{aligned}$$

  - Choose m_k with highest posterior probability

Bayesian model evaluation

- It is believed that a set of data $\mathbf{Y}$ has been generated from the pdf $f(\mathbf{y}|m_k, \mathbf{\Theta}_k)$, where m_k is one of a set of $M = \{m_1, \ldots, m_K\}$ models.

- As Bayesians do...
  - Assign priors $f(\mathbf{\Theta}_k|m_k)$ and $f(m_k)$
  - Compute the posterior

$$\begin{aligned} f(m_k|\mathbf{y}) &\propto f(\mathbf{y}|m_k)f(m_k) \\ &\propto \int f(\mathbf{y}|m_k, \mathbf{\Theta}_k)f(\mathbf{\Theta}_k|m_k)\,\mathrm{d}\mathbf{\Theta}_k\, f(m_k) \end{aligned}$$

  - Choose m_k with highest posterior probability

- Use MCMC methods to sample from the posterior when
  - the integral in the posterior is not analytically tractable; and/or
  - the model space is too large to make any calculation of the posterior for all models unfeasible.

Bayesian variable selection

- Consider again the linear model in (1)

$$y_i = \alpha + \beta_1 X_{i,1} + \cdots + \beta_p X_{i,p} + \epsilon_i$$
$$\epsilon_i \sim \mathrm{N}(0, \psi^{-1}) \text{ iid}$$
$$i = 1, \ldots, n$$

- A model is a subset of variables $\{\tilde{X}_1, \ldots, \tilde{X}_q\}$ from $\{X_1, \ldots, X_p\}$. There are 2^p models to consider.

Bayesian variable selection

- Consider again the linear model in (1)

$$y_i = \alpha + \beta_1 X_{i,1} + \cdots + \beta_p X_{i,p} + \epsilon_i$$
$$\epsilon_i \sim \mathrm{N}(0, \psi^{-1}) \text{ iid}$$
$$i = 1, \ldots, n$$

- A model is a subset of variables $\{\tilde{X}_1, \ldots, \tilde{X}_q\}$ from $\{X_1, \ldots, X_p\}$. There are 2^p models to consider.

- Index each of these 2^p models by the vector

$$\boldsymbol{\gamma} = (\gamma_1, \ldots, \gamma_p)$$

where $\gamma_j = 1$ if X_j is selected, and 0 otherwise.

Bayesian variable selection

- Consider again the linear model in (1)

$$y_i = \alpha + \beta_1 X_{i,1} + \cdots + \beta_p X_{i,p} + \epsilon_i$$
$$\epsilon_i \sim \mathrm{N}(0, \psi^{-1}) \text{ iid}$$
$$i = 1, \ldots, n$$

- A model is a subset of variables $\{\tilde{X}_1, \ldots, \tilde{X}_q\}$ from $\{X_1, \ldots, X_p\}$. There are 2^p models to consider.

- Index each of these 2^p models by the vector

$$\boldsymbol{\gamma} = (\gamma_1, \ldots, \gamma_p)$$

where $\gamma_j = 1$ if X_j is selected, and 0 otherwise.

- Assign priors $f(\boldsymbol{\gamma})$, and also $f(\boldsymbol{\beta}, \psi | \boldsymbol{\gamma})$. Interested in two things:
  - Posterior inclusion probabilities $\mathbb{P}[\gamma_j = 1 | \mathbf{y}]$ for variable X_j.
  - Posterior model probabilities $\mathbb{P}[\boldsymbol{\gamma} = \boldsymbol{\gamma}_k | \mathbf{y}]$ for model $\boldsymbol{\gamma}_k$.

1 George and McCulloch's (1993) Stochastic Search Variable Selection [SSVS]

$$y_i = \alpha + \beta_1 X_{i,1} + \cdots + \beta_p X_{i,p} + \epsilon_i$$
$$\epsilon_i \sim \text{N}(0, \psi^{-1}) \text{ iid}$$

Priors on $\boldsymbol{\beta}$ and $\boldsymbol{\gamma}$

$$\beta_j | \gamma_j \sim \gamma_j \text{N}(0, c_j^2 t_j^2) + (1 - \gamma_j)\text{N}(0, t_j^2)$$
$$\gamma_j \sim \text{Bern}(p_j)$$

- t_j and c_j are tuning parameters.
  - Suggested values are $\big(\text{SE}(\hat{\beta}_j)/t_j, c_j\big) = (1, 5)$, $(1, 10)$, $(10, 100)$, or $(10, 500)$.
  - $\text{SE}(\hat{\beta}_j) = \sqrt{\hat{\psi}^{-1}(\mathbf{X}^\top \mathbf{X})_{jj}}$ under the full model.

2 Kuo and Mallick's (1998) sampler [KM]

$$y_i = \alpha + \gamma_1 \beta_1 X_{i,1} + \cdots + \gamma_p \beta_p X_{i,p} + \epsilon_i$$
$$\epsilon_i \sim \text{N}(0, \psi^{-1}) \text{ iid}$$

Priors on $\boldsymbol{\beta}$ and $\boldsymbol{\gamma}$

$$\beta_j \sim \text{N}(b_j, d_j^2)$$
$$\gamma_j \sim \text{Bern}(p_j)$$

- Choices for b_j and d_j reflect prior beliefs on $\boldsymbol{\beta}$.
- In the absence of prior information
  - Choose $b_j = 0$
  - Standardise the $\mathbf{X}$ variables, and choose $d_j = d$ such that $1/2 \leq d \leq 4$

3 Dellaportas et. al. (2002) Gibbs Variable Selection [GVS]

$$y_i = \alpha + \gamma_1\beta_1 X_{i,1} + \cdots + \gamma_p\beta_p X_{i,p} + \epsilon_i$$
$$\epsilon_i \sim \mathrm{N}(0, \psi^{-1}) \text{ iid}$$

Priors on $\boldsymbol{\beta}$ and $\boldsymbol{\gamma}$

$$\beta_j|\gamma_j \sim \gamma_j \mathrm{N}(b_j, d_j^2) + (1-\gamma_j)\mathrm{N}(u_j, s_j^2)$$
$$\gamma_j \sim \mathrm{Bern}(p_j)$$

- u_j and s_j are tuning parameters. Choices include
  - $u_j = \hat{\beta}_j$, the OLS estimates, and correspondingly $s_j^2 = \widehat{\mathrm{Var}}(\hat{\beta}_j)$.
  - $u_j = 0$ and $s_j^2 \propto d_j^2$, but kept low.
- As before, we can choose $b_j = 0$ and $d_j = d$ with large d (after standardising $\mathbf{X}$) if no prior information.

Priors

- Priors for $\beta_1, \ldots, \beta_p$

SSVS $\beta_j|\gamma_j \sim \gamma_j \mathrm{N}(0, 500^2 \cdot \widehat{\mathrm{Var}}(\hat\beta_j)/10^2) + (1-\gamma_j)\mathrm{N}(0, 500^2)$

KM $\beta_j \sim \mathrm{N}(0, 4^2)$

GVS $\beta_j|\gamma_j \sim \gamma_j \mathrm{N}(0, 10^2) + (1-\gamma_j)\mathrm{N}\big(\hat\beta_j, \widehat{\mathrm{Var}}(\hat\beta_j)\big)$

- Priors for $\gamma_1, \ldots, \gamma_p$
  - $\gamma_j \sim \mathrm{Bern}(1/2)$
  - This shows our indifference between any choice of variables

- Priors for other parameters
  - $\alpha \sim \mathrm{N}(0, 1000)$
  - $\psi \sim \Gamma(0.001, 0.001)$
  - Not too bothered about estimating these - just let the data take care of them

Simulated example

- Simple variable selection problem with $p = 5$ and $n = 50$.
  - Draw $\mathbf{X}_1, \dots, \mathbf{X}_5 \sim \text{N}(\mathbf{0}, \mathbf{I}_{50})$.
  - Generate response variables $\mathbf{Y} = \mathbf{X}_4 + \mathbf{X}_5 + \epsilon$.
  - ϵ drawn from $\text{N}(\mathbf{0}, 2^2\mathbf{I}_{50})$.

Simulated example

- Simple variable selection problem with $p = 5$ and $n = 50$.
  - Draw $\mathbf{X}_1, \dots, \mathbf{X}_5 \sim \text{N}(\mathbf{0}, \mathbf{I}_{50})$.
  - Generate response variables $\mathbf{Y} = \mathbf{X}_4 + \mathbf{X}_5 + \epsilon$.
  - ϵ drawn from $\text{N}(\mathbf{0}, 2^2\mathbf{I}_{50})$.
- Simulation results for 10,000 MCMC samples

	SSVS		KM		GVS	
	$\widehat{P}[\gamma_j = 1 \mid \mathbf{y}]$	S.E.	$\widehat{P}[\gamma_j = 1 \mid \mathbf{y}]$	S.E.	$\widehat{P}[\gamma_j = 1 \mid \mathbf{y}]$	S.E.
γ_1	0.03	0.01	0.03	0.01	0.03	0.01
γ_2	0.16	0.04	0.10	0.01	0.11	0.01
γ_3	0.02	0.01	0.03	0.01	0.03	0.01
γ_4	0.80	0.07	0.84	0.02	0.87	0.01
γ_5	0.78	0.08	0.95	0.01	0.93	0.01

Rank	Model	Prob.	Odds	Model	Prob.	Odds	Model	Prob.	Odds
1	$X_4 + X_5$	0.63	1.00	$X_4 + X_5$	0.72	1.00	$X_4 + X_5$	0.73	1.00
2	X_2	0.09	7.16	X_5	0.10	7.32	X_5	0.07	10.6
3	$X_2 + X_5$	0.04	18.4	X_4	0.08	7.78	$X_2 + X_4 + X_5$	0.04	18.0

1 Introduction

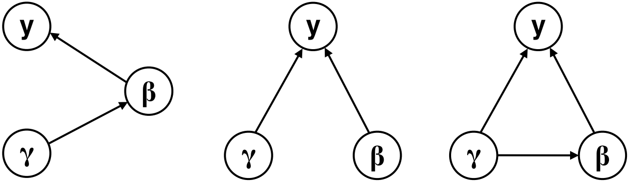
2 ASIDE: Regression modelling using I-priors

3 Bayesian variable selection

4 Using I-priors in Bayesian variable selection

5 Summary

Comparison between the methods



$f(\mathbf{y}|\boldsymbol{\beta})f(\boldsymbol{\beta}|\boldsymbol{\gamma})f(\boldsymbol{\gamma})$ $f(\mathbf{y}|\boldsymbol{\gamma},\boldsymbol{\beta})f(\boldsymbol{\gamma})f(\boldsymbol{\beta})$ $f(\mathbf{y}|\boldsymbol{\gamma},\boldsymbol{\beta})f(\boldsymbol{\beta}|\boldsymbol{\gamma})f(\boldsymbol{\gamma})$

	SSVS	**KM**	**GVS**
Parameter space	Retains original	Does not retain original	
Tuning parameters	Many	None	Some
Priors for $\boldsymbol{\beta}$	$\boldsymbol{\beta}|\boldsymbol{\gamma} \sim \text{N}(\mathbf{0}, \mathbf{R}_\gamma \mathbf{D} \mathbf{R}_\gamma)$ $\mathbf{D} = \mathbf{I}_p$ $\mathbf{R}_\gamma = \text{diag}(a_j t_j)$ $a_j = (1-\gamma_j) + \gamma_j c_j$	$\boldsymbol{\beta} \sim \text{N}(\mathbf{0}, \mathbf{D})$ $\mathbf{D} = d^2 \mathbf{I}_p$	$\boldsymbol{\beta}|\boldsymbol{\gamma} \sim \text{N}(\boldsymbol{\mu}, \boldsymbol{\Sigma})$ $\boldsymbol{\mu}_j = (1-\gamma_j)u_j$ $\boldsymbol{\Sigma}_{jk} = \gamma_j\gamma_k(d^2\mathbf{I}_p)_{jk}$ $+(1-\gamma_j\gamma_k)\mathbf{1}_{[j=k]}s_j^2$

- All three are inefficient when there are strong correlations in the data.

Using I-priors in Bayesian variable selection

- Opportunity to use I-prior in each of the three methods. Simply replace $\mathbf{D} = \lambda^2 \psi \mathbf{X}^\top \mathbf{X}$.
- The thought here is to replicate the correlations in the data into the prior covariance matrix of $\boldsymbol{\beta}$.
- Which method works best with I-priors?
  - SSVS is unappealing due to the many tuning (hyper)parameters.
  - GVS is similar to KM, but designed to make the sampling more efficient. This was not seen in our simulations.
  - KM seems the simplest, “hands-free” method.

4 The KM I-prior model [I-prior]

$$y_i = \alpha + \gamma_1 \beta_1 X_{i,1} + \cdots + \gamma_p \beta_p X_{i,p} + \epsilon_i$$
$$\epsilon_i \sim \mathrm{N}(0, \psi^{-1})$$
$$i = 1, \ldots, n$$

Priors

$$\boldsymbol{\beta} \sim \mathrm{N}(\mathbf{0}, \lambda^2 \psi \mathbf{X}^\top \mathbf{X})$$
$$\gamma_1, \ldots, \gamma_p \sim \mathrm{Bern}(1/2)$$
$$\alpha \sim \mathrm{N}(0, 1000)$$
$$\psi \sim \Gamma(0.001, 0.001)$$
$$1/\lambda^2 \sim \Gamma(0.001, 0.001)$$

Simulation study

- Variable selection problem with $p = 100$ and $n = 150$ with artificial pairwise correlations between variables.
  - Draw $\mathbf{Z}_1, \dots, \mathbf{Z}_{100} \sim \text{N}(\mathbf{0}, \mathbf{I}_{150})$.
  - Draw $\mathbf{U} \sim \text{N}(\mathbf{0}, \mathbf{I}_{150})$.
  - Let $\mathbf{X}_j = \mathbf{Z}_j + \mathbf{U}$. This induces pairwise correlations of about 0.5.
  - Generate response variables $\mathbf{Y} = \mathbf{X}\boldsymbol{\beta}_{true} + \boldsymbol{\epsilon}$.
  - $\boldsymbol{\epsilon}$ drawn from $\text{N}(\mathbf{0}, 2^2\mathbf{I}_{150})$.

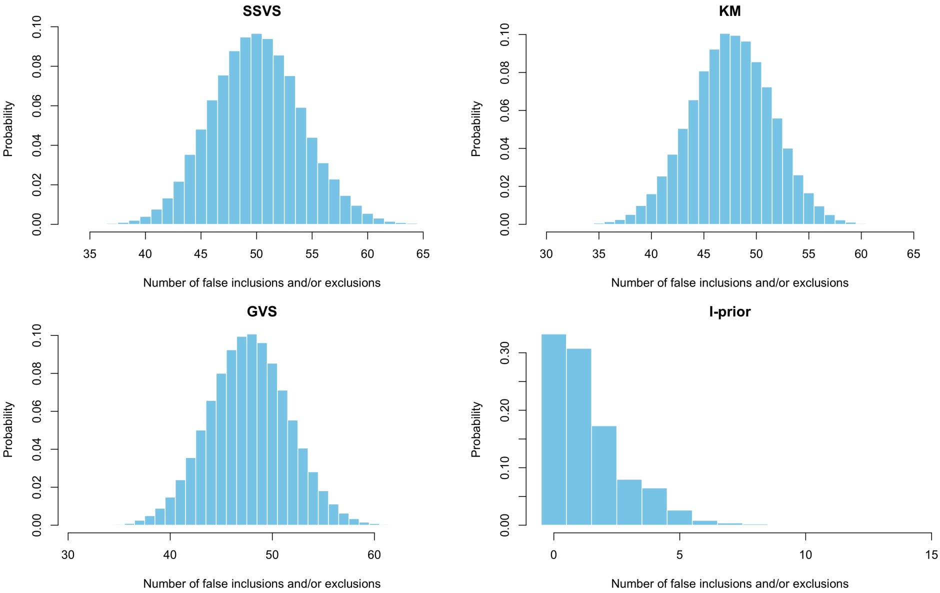
- Let $\boldsymbol{\beta}_{true} = (\boldsymbol{\beta}_{-k}, \boldsymbol{\beta}_k)$, where
  - $\boldsymbol{\beta}_{-k} = (\beta_1, \dots, \beta_k) = (0, \dots, 0)$; and
  - $\boldsymbol{\beta}_k = (\beta_{k+1}, \dots, \beta_{100}) = (1, \dots, 1)$.

  In other words, only variables X_{k+1} to X_{100} are used.

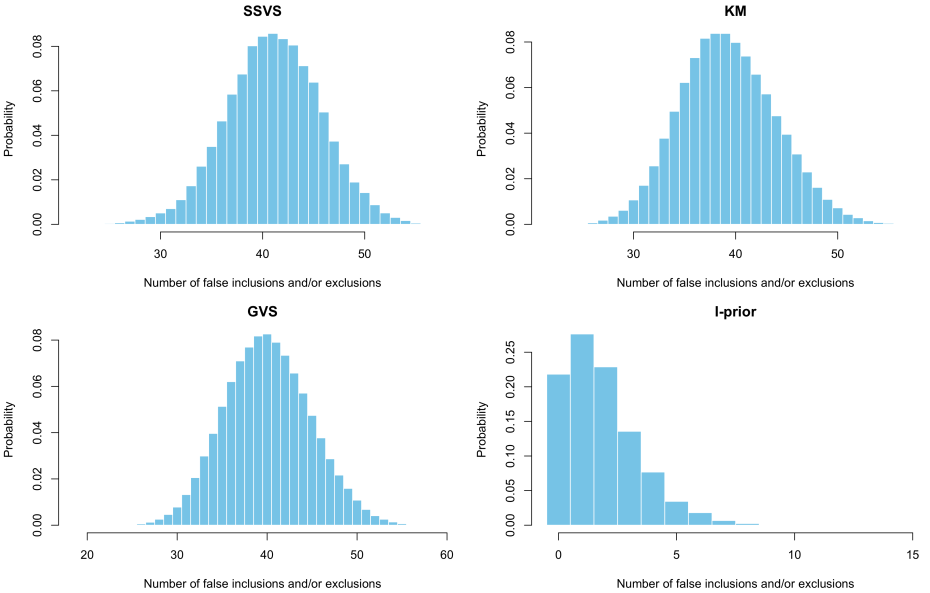
- The value of k is varied between 10, 25, 50, 75 and 90.

- 10,000 MCMC samples obtained for each scenario. Interested in how many false choices the models make. Each experiment repeated 10 times and results averaged.

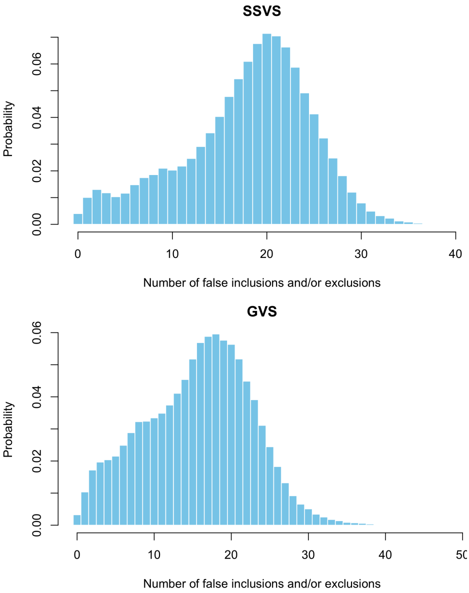
Scenario A (10,90)



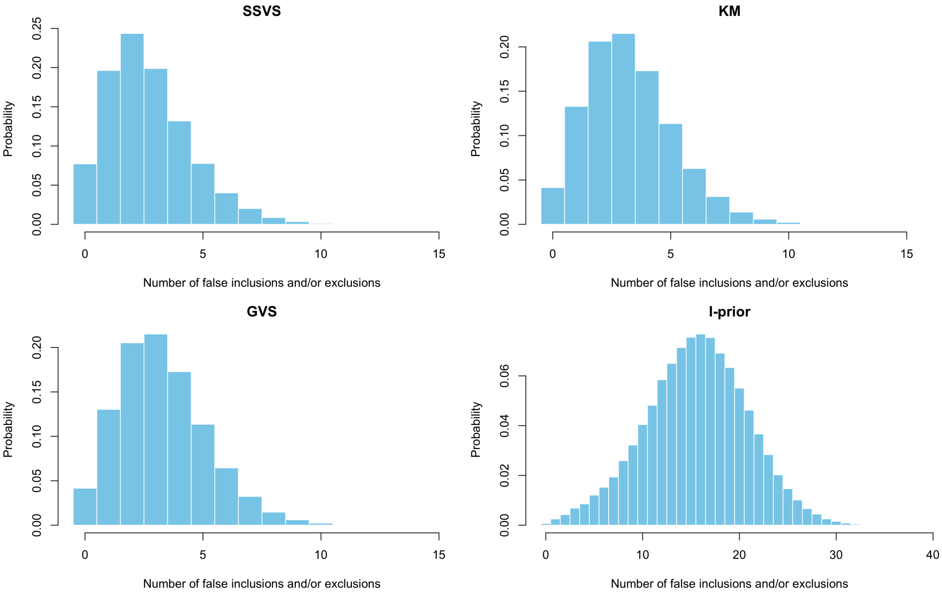
Scenario B (25,75)



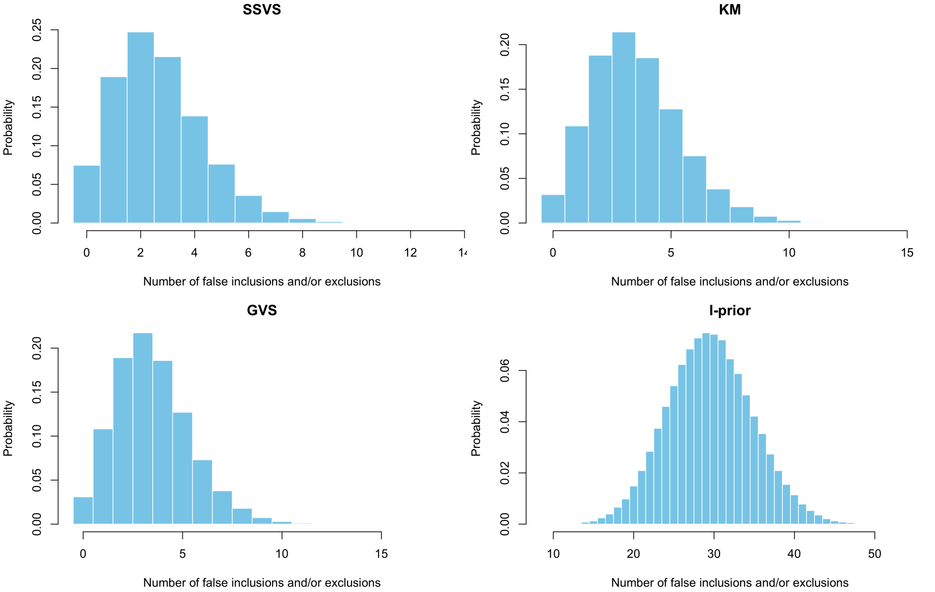
Scenario C (50,50)



Scenario D (75,25)



Scenario E (90,10)



Motivation for two-stage procedure

- I-priors performs very well when there are a lot of non-zero betas.
  - Strength comes from the Fisher information.
  - However, we can't expect I-priors to do well when few non-zero betas.
  - A lot of information becomes unnecessary and muddles the actual useful information.

Need an objective way to trim and reduce the variable space.

Motivation for two-stage procedure

- I-priors performs very well when there are a lot of non-zero betas.
  - Strength comes from the Fisher information.
  - However, we can't expect I-priors to do well when few non-zero betas.
  - A lot of information becomes unnecessary and muddles the actual useful information.

  Need an objective way to trim and reduce the variable space.

- Two stage process
  - 1st Run the model. Keep only variables with posterior inclusion probabilities greater or equal to 0.5.
  - 2nd Re-run the model on the set of reduced variables.

Motivation for two-stage procedure

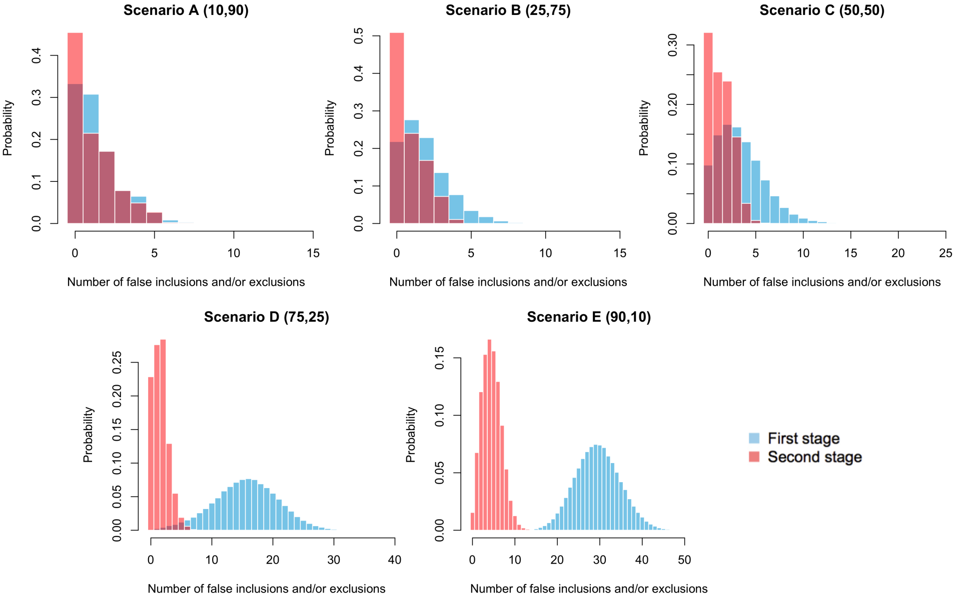
- I-priors performs very well when there are a lot of non-zero betas.
  - Strength comes from the Fisher information.
  - However, we can't expect I-priors to do well when few non-zero betas.
  - A lot of information becomes unnecessary and muddles the actual useful information.

  Need an objective way to trim and reduce the variable space.

- Two stage process
  - 1st Run the model. Keep only variables with posterior inclusion probabilities greater or equal to 0.5.
  - 2nd Re-run the model on the set of reduced variables.

- Barbieri and Berger (2004) showed that keeping such variables results in the most optimally predictive model being selected.

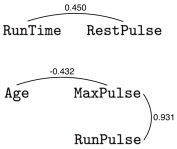
Two-stage simulation results



Real world applications (1)

- Factors affecting aerobic fitness ($n = 30$ and $p = 6$) [Kuo and Mallick, 1998].
    - Response variable `Oxygen`, a measurement of oxygen uptake rate in mL/kg body weight per minute.
    - Covariates: `Age`, `Weight`, `RunTime`, `RestPulse`, `RunPulse`, `MaxPulse`.

	Full model	**I-prior**	**Forward sel.**	**Back elim.**
`Intercept`	104.2 (0.00)	80.8 (0.00)	103.3 (0.00)	98.6 (0.00)
`Age`	-0.24 (0.03)		-0.25 (0.02)	-0.21 (0.05)
`Weight`	-0.08 (0.15)		-0.08 (0.15)	
`RunTime`	-2.59 (0.00)	-2.97 (0.00)	-2.64 (0.00)	-2.75 (0.00)
`RestPulse`	-0.02 (0.72)			
`RunPulse`	-0.38 (0.00)	-0.38 (0.01)	-0.39 (0.00)	-0.36 (0.01)
`MaxPulse`	0.32 (0.03)	0.36 (0.02)	0.32 (0.03)	0.28 (0.05)
C_p	7.0	7.7	5.1	5.3
AIC	56.8	58.5	54.9	55.6
5-CV RMSE	2.59	2.71	2.50	2.54



sample correlations

Real world applications (2)

- Effects of air pollution on mortality in a US metropolitan area ($n = 60$ and $p = 15$) [McDonald and Schwing, 1973].
  - Response variable `Mortality`, a total age adjusted mortality rate.
  - Pollution potential data for `HC`, `NOx` and `SO2`.
  - Environmental considerations are `Rain`, `JanTemp`, `JulTemp` and `Humid`.
  - Socioeconomic considerations are `Over65`, `Popn`, `Educ`, `Hous`, `Dens`, `NonW`, `WhiteCollar` and `Poor`.

		Full model	I-prior	Min C_p	Back elim.
Environmental & demographic variables selected		All	`Rain`, `JanTemp`, `JulTemp`, `Humid`, `Over65`, `Popn`, `Hous`, `NonW`, `Poor`	`Rain`, `JanTemp`, `JulTemp`, `Educ`, `NonW`	`JanTemp`, `Educ`, `NonW`
Pollution effect	`HC`	✗	✗	✗	✓ $\beta = -0.98$
	`NOx`	✗	✗	✗	✓ $\beta = 1.99$
	`SO2`	✗	✓ $\beta = 0.33$	✓ $\beta = 0.26$	✗
C_p		16.0	13.4	3.6	8.7
AIC		439.8	439.2	429.0	435.0
5-CV RMSE		50.6	41.6	37.1	38.6

Has this been done before...? A look at g-priors

- g-priors [Zellner, 1986] for linear regression coefficients has covariance matrix proportional to the inverse Fisher information

$$\boldsymbol{\beta} \sim \mathrm{N}\big(\mathbf{0}, g(\mathbf{X}^\top\mathbf{X})^{-1}\big)$$

- Popular choice of prior in Bayesian variable selection
  - "*...use of* $\propto (\mathbf{X}^T\mathbf{X})^{-1}$ *tends to replicate design correlation*" [George and McCulloch, 1993]
  - "The choice of $\propto (\mathbf{X}^\top\mathbf{X})^{-1}$ serves to replicate the covariance structure of the likelihood" [Chipman et. al., 2001]
  - Used as an informative prior for variable selection problems, e.g. gene selection [Lee et. al., 2002]

Why g-priors shouldn't work

- The intuition is wrong.

  $\uparrow$ *Fisher information* $\Rightarrow \downarrow$ *variance* $\Rightarrow \uparrow$ *influence of prior zero mean*

Why g-priors shouldn't work

- The intuition is wrong.

  $\uparrow$ *Fisher information* $\Rightarrow \downarrow$ *variance* $\Rightarrow \uparrow$ *influence of prior zero mean*

- It is equivalent to using an independent prior on decorrelated data.

$$\mathbf{y} = \boldsymbol{\alpha} + \mathbf{X}\boldsymbol{\beta} + \boldsymbol{\epsilon}$$
$$\boldsymbol{\epsilon} \sim \mathrm{N}(\mathbf{0}, \psi^{-1}\mathbf{I}_n)$$
$$\boldsymbol{\beta} \sim \mathrm{N}\big(\mathbf{0}, g(\mathbf{X}^\top\mathbf{X})^{-1}\big)$$

Why g-priors shouldn't work

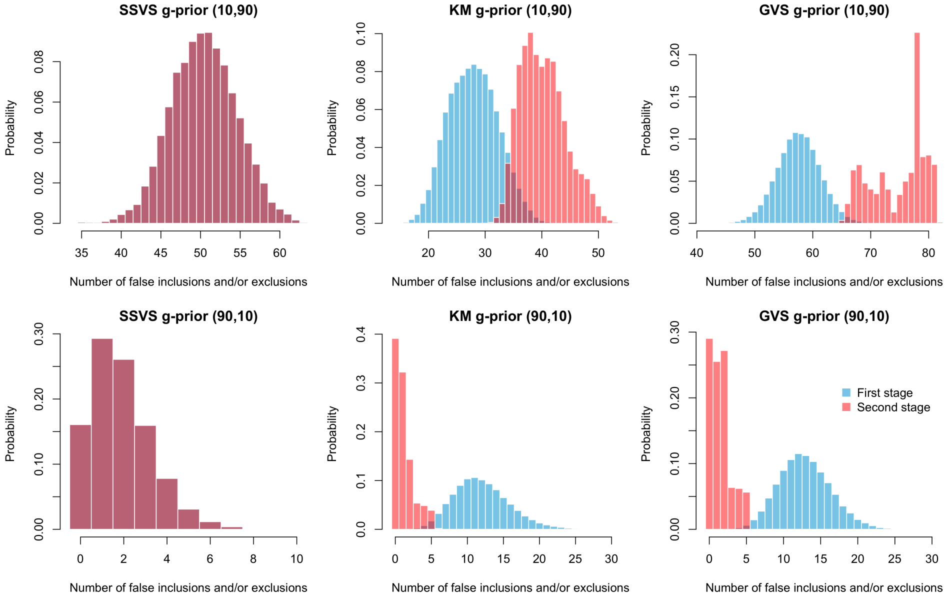
- The intuition is wrong.

  $\uparrow$ *Fisher information* $\Rightarrow \downarrow$ *variance* $\Rightarrow \uparrow$ *influence of prior zero mean*

- It is equivalent to using an independent prior on decorrelated data.

$$\left.\begin{array}{c} \mathbf{y} = \boldsymbol{\alpha} + \mathbf{X}\beta + \boldsymbol{\epsilon} \\ \boldsymbol{\epsilon} \sim \mathrm{N}(\mathbf{0}, \psi^{-1}\mathbf{I}_n) \\ \beta \sim \mathrm{N}\big(\mathbf{0}, g(\mathbf{X}^\top\mathbf{X})^{-1}\big) \end{array}\right\} \iff \left\{\begin{array}{c} \mathbf{y} = \boldsymbol{\alpha} + \tilde{\mathbf{X}}\tilde{\beta} + \boldsymbol{\epsilon} \\ \boldsymbol{\epsilon} \sim \mathrm{N}(\mathbf{0}, \psi^{-1}\mathbf{I}_n) \\ \tilde{\mathbf{X}} = \mathbf{X}(\mathbf{X}^\top\mathbf{X})^{-1/2} \\ \tilde{\beta} = (\mathbf{X}^\top\mathbf{X})^{1/2}\beta \\ \tilde{\beta} \sim \mathrm{N}(\mathbf{0}, g^2\mathbf{I}) \end{array}\right.$$

g-prior results



1 Introduction

2 ASIDE: Regression modelling using I-priors

3 Bayesian variable selection

4 Using I-priors in Bayesian variable selection

5 Summary

Summary

- Variable selection under a Bayesian approach reduces to a problem of parameter estimation (i.e. γ).
- At the outset, wanted to find a simple and automatic way of running a variable selection model even in cases with (strong) collinearity.
- Our small contribution was to introduce an information theoretic prior in a two-stage approach. Good simulation results, but not very convincing in real world applications (yet).
- Things to do:
  - Find a way to accommodate individual scaling parameters for each variable - akin to original I-prior modelling.
  - Write own Gibbs sampling code for $p << n$ case.
  - Try this out on generalised linear models.

Some things I'd like to share

- Running WinBUGS in non-Windows environment using JAGS (in R!).
  - JAGS is able to estimate Bayesian models using MH/Gibbs sampling.
  - If model is not too complex, convenient compared to writing own code.

Some things I'd like to share

- Running WinBUGS in non-Windows environment using JAGS (in `R`!).
  - JAGS is able to estimate Bayesian models using MH/Gibbs sampling.
  - If model is not too complex, convenient compared to writing own code.
- Parallelize your `R` code to save (lots of) time.
  - Run 4 chains of length 2,500 on each quad core instead of 10,000 on a single core. This cuts runtime in quarter.
  - I used `clusterApply()` {`snow`}.
  - There is also a parallelized JAGS function in `R2jags`.

Some things I'd like to share

- Running WinBUGS in non-Windows environment using JAGS (in `R`!).
  - JAGS is able to estimate Bayesian models using MH/Gibbs sampling.
  - If model is not too complex, convenient compared to writing own code.
- Parallelize your `R` code to save (lots of) time.
  - Run 4 chains of length 2,500 on each quad core instead of 10,000 on a single core. This cuts runtime in quarter.
  - I used `clusterApply()` {`snow`}.
  - There is also a parallelized JAGS function in `R2jags`.
- `R` code bug diagnostic: `traceback()`.
  - For those "where's this error coming from?!" moments

Some things I'd like to share

- Running WinBUGS in non-Windows environment using JAGS (in `R`!).
  - JAGS is able to estimate Bayesian models using MH/Gibbs sampling.
  - If model is not too complex, convenient compared to writing own code.
- Parallelize your `R` code to save (lots of) time.
  - Run 4 chains of length 2,500 on each quad core instead of 10,000 on a single core. This cuts runtime in quarter.
  - I used `clusterApply()` {`snow`}.
  - There is also a parallelized JAGS function in `R2jags`.
- `R` code bug diagnostic: `traceback()`.
  - For those "where's this error coming from?!" moments
- `R` textual progress bars: `create_progress_bar()` {`plyr`}.
  - `0%|=========================                   |100%`

Some things I'd like to share

- Running WinBUGS in non-Windows environment using JAGS (in `R`!).
    - JAGS is able to estimate Bayesian models using MH/Gibbs sampling.
    - If model is not too complex, convenient compared to writing own code.
- Parallelize your `R` code to save (lots of) time.
    - Run 4 chains of length 2,500 on each quad core instead of 10,000 on a single core. This cuts runtime in quarter.
    - I used `clusterApply()` {`snow`}.
    - There is also a parallelized JAGS function in `R2jags`.
- `R` code bug diagnostic: `traceback()`.
    - For those "where's this error coming from?!" moments
- `R` textual progress bars: `create_progress_bar()` {`plyr`}.
    - `0%|=========================                  |100%`
- If you work with matrices a lot, check out The Matrix Cookbook [Petersen and Pedersen, 2012].

End

Thank you!