Bayesian Variable Selection for Linear Models

Haziq Jamil

PhD (LSE), MSc (LSE), BSc MMORSE (Warw)

Assistant Professor in Statistics
Faculty of Science, UBD

13 November 2019

UBDSBE Seminar

https://haziqj.ml/talk/sbe-bvs/

Outline

The linear regression model

- For $i = 1, \dots, n$, consider the multiple regression model

$$y_i = \alpha + \sum_{j=1}^{p} x_{ij}\beta_j + \epsilon_i$$
$$\epsilon_i \overset{\text{iid}}{\sim} \mathrm{N}(0, \sigma^2) \tag{1}$$

The linear regression model

- For $i = 1, \ldots, n$, consider the multiple regression model

$$y_i = \alpha + \sum_{j=1}^{p} x_{ij}\beta_j + \epsilon_i \qquad \epsilon_i \overset{\text{iid}}{\sim} \mathrm{N}(0, \sigma^2) \tag{1}$$

- Without loss of generality, assume covariates are standardised.

The linear regression model

- For $i = 1, \ldots, n$, consider the multiple regression model

$$y_i = \alpha + \sum_{j=1}^{p} x_{ij}\beta_j + \epsilon_i \qquad (1)$$
$$\epsilon_i \overset{\text{iid}}{\sim} \text{N}(0, \sigma^2)$$

- Without loss of generality, assume covariates are standardised.
- The OLS estimate for $\boldsymbol{\beta} = (\beta_1, \ldots, \beta_p)^\top$ is

$$\hat{\boldsymbol{\beta}} = (\mathbf{X}^\top\mathbf{X})^{-1}\mathbf{X}^\top\mathbf{y},$$

where $\mathbf{X} = (x_{ij})_{i=1:n, j=1:p}$ and $\mathbf{y} = (y_i)_{i=1:n}$.

The linear regression model

- For $i = 1, \ldots, n$, consider the multiple regression model

$$y_i = \alpha + \sum_{j=1}^{p} x_{ij}\beta_j + \epsilon_i \qquad (1)$$
$$\epsilon_i \overset{\text{iid}}{\sim} \mathrm{N}(0, \sigma^2)$$

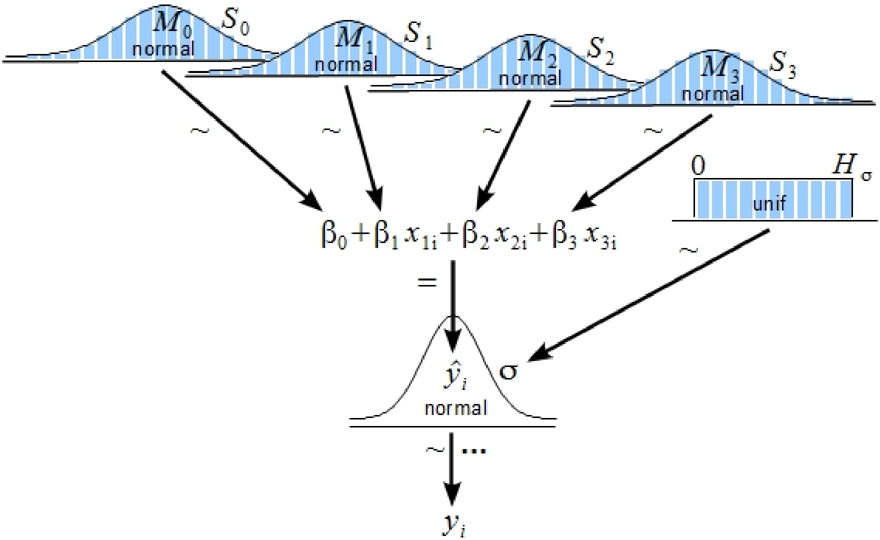
- Without loss of generality, assume covariates are standardised.
- The OLS estimate for $\boldsymbol\beta = (\beta_1, \ldots, \beta_p)^\top$ is

$$\hat{\boldsymbol\beta} = (\mathbf{X}^\top \mathbf{X})^{-1}\mathbf{X}^\top \mathbf{y},$$

where $\mathbf{X} = (x_{ij})_{i=1:n, j=1:p}$ and $\mathbf{y} = (y_i)_{i=1:n}$.

- This corresponds to the maximum likelihood (ML) estimator.

Bayesian linear regression



Bayesian linear regression

- The Bayesian approach supplements the data with additional information in the form of prior beliefs about the parameters:
  - $\boldsymbol{\beta} \sim \mathrm{N}_p(\mathbf{b}, \sigma^2\mathbf{B})$
  - $\sigma^2 \sim \Gamma^{-1}(c, d)$

Bayesian linear regression

- The Bayesian approach supplements the data with additional information in the form of prior beliefs about the parameters:
  - $\boldsymbol{\beta} \sim \mathrm{N}_p(\mathbf{b}, \sigma^2\mathbf{B})$
  - $\sigma^2 \sim \Gamma^{-1}(c, d)$

- Inference on the parameters $\boldsymbol{\Theta} = \{\alpha, \boldsymbol{\beta}, \sigma^2\}$ is done via the posterior

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

Bayesian linear regression

- The Bayesian approach supplements the data with additional information in the form of prior beliefs about the parameters:
  - $\boldsymbol\beta \sim \text{N}_p(\mathbf{b}, \sigma^2\mathbf{B})$
  - $\sigma^2 \sim \Gamma^{-1}(c, d)$

- Inference on the parameters $\boldsymbol\Theta = \{\alpha, \boldsymbol\beta, \sigma^2\}$ is done via the posterior

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

- In particular, the posterior distribution for $\boldsymbol\beta$ is $\text{N}_p(\tilde{\boldsymbol\beta}, \sigma^2\tilde{\text{B}})$, where

$$\tilde{\boldsymbol\beta} = (\mathbf{X}^\top\mathbf{X} + \mathbf{B}^{-1})^{-1}(\mathbf{B}^{-1}\mathbf{b} + \mathbf{X}^\top\mathbf{y}) \tag{2}$$

and

$$\tilde{\text{B}} = (\mathbf{X}^\top\mathbf{X} + \mathbf{B}^{-1})^{-1} \tag{3}$$

Motivation

- In a statistical model, which variables $X_1, \ldots, X_p$ are most important in explaining response variable y?

Motivation

- In a statistical model, which variables $X_1, \ldots, X_p$ are most important in explaining response variable y?
  - **Mortality and air pollution data ($n = 60, p = 15$).** Which of three pollutants (HC, NO_x, SO_2) contributes to mortality rate is a US metropolitan area? McDonald and Schwing (1973).

Motivation

- In a statistical model, which variables $X_1, \ldots, X_p$ are most important in explaining response variable y?
  - **Mortality and air pollution data ($n = 60, p = 15$).** Which of three pollutants (HC, NO_x, SO_2) contributes to mortality rate is a US metropolitan area? McDonald and Schwing (1973).
  - **Molecular inversion probes in breast cancer study ($n = 971, p = 16,253$).** Genomic partitioning to identify genes significantly associated with clinically relevant subtypes of breast cancer. Zhang et al. (2014).

Motivation

- In a statistical model, which variables $X_1, \ldots, X_p$ are most important in explaining response variable y?
  - **Mortality and air pollution data ($n = 60, p = 15$).** Which of three pollutants (HC, NO_x, SO_2) contributes to mortality rate is a US metropolitan area? McDonald and Schwing (1973).
  - **Molecular inversion probes in breast cancer study ($n = 971, p = 16,253$).** Genomic partitioning to identify genes significantly associated with clinically relevant subtypes of breast cancer. Zhang et al. (2014).
  - **Nowcasting economic time series data ($n = 100, p = 151$).** Using search engine query data as predictors for consumer sentiment and gun sales. Scott and Varian (2015).

Motivation

- In a statistical model, which variables $X_1, \ldots, X_p$ are most important in explaining response variable y?
  - **Mortality and air pollution data ($n = 60, p = 15$).** Which of three pollutants (HC, NO_x, SO_2) contributes to mortality rate is a US metropolitan area? McDonald and Schwing (1973).
  - **Molecular inversion probes in breast cancer study ($n = 971, p = 16,253$).** Genomic partitioning to identify genes significantly associated with clinically relevant subtypes of breast cancer. Zhang et al. (2014).
  - **Nowcasting economic time series data ($n = 100, p = 151$).** Using search engine query data as predictors for consumer sentiment and gun sales. Scott and Varian (2015).
- Premise: Too many covariates/predictors, not all are useful and/or unable to fit everything in the model.

Model selection

- For linear models, a model is defined to be a subset of variables from $\{X_1, \ldots, X_p\}$ which is included in the regression.
- Goal is to infer which model, from the set of all possible models $\mathcal{M} = \{M_1, \ldots, M_{2^p}\}$, is behind the true data generative process.

Model selection

- For linear models, a model is defined to be a subset of variables from $\{X_1, \ldots, X_p\}$ which is included in the regression.
- Goal is to infer which model, from the set of all possible models $\mathcal{M} = \{M_1, \ldots, M_{2^p}\}$, is behind the true data generative process.
- Broadly, model selection can be classified into three categories:
  - **Criterion-based model comparison.**
    Using some predictive-based (R^2, k-CV MSEP, C_p, etc.) or likelihood-based criterion (likelihood, AIC, BIC, etc.), models are compared pairwise. Used in conjunction with stepwise procedures.

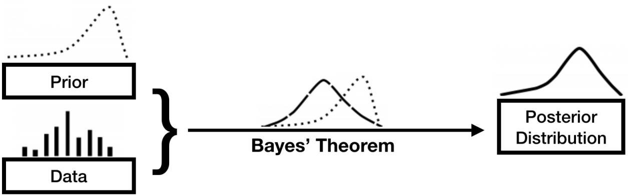
Model selection

- For linear models, a model is defined to be a subset of variables from $\{X_1, \dots, X_p\}$ which is included in the regression.
- Goal is to infer which model, from the set of all possible models $\mathcal{M} = \{M_1, \dots, M_{2^p}\}$, is behind the true data generative process.
- Broadly, model selection can be classified into three categories:
  - **Criterion-based model comparison.**
    Using some predictive-based (R^2, k-CV MSEP, C_p, etc.) or likelihood-based criterion (likelihood, AIC, BIC, etc.), models are compared pairwise. Used in conjunction with stepwise procedures.
  - **Shrinkage/regularisation.**
    Regularise the linear system of equations to induce sparsity (ridge regression, Lasso, elastic nets, etc.). Includes Bayesian priors on $\boldsymbol{\beta}$.

Model selection

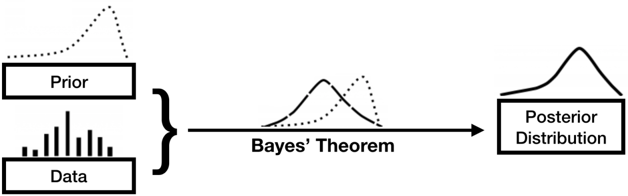
- For linear models, a model is defined to be a subset of variables from $\{X_1, \ldots, X_p\}$ which is included in the regression.
- Goal is to infer which model, from the set of all possible models $\mathcal{M} = \{M_1, \ldots, M_{2^p}\}$, is behind the true data generative process.
- Broadly, model selection can be classified into three categories:
  - **Criterion-based model comparison.**
    Using some predictive-based (R^2, k-CV MSEP, C_p, etc.) or likelihood-based criterion (likelihood, AIC, BIC, etc.), models are compared pairwise. Used in conjunction with stepwise procedures.
  - **Shrinkage/regularisation.**
    Regularise the linear system of equations to induce sparsity (ridge regression, Lasso, elastic nets, etc.). Includes Bayesian priors on $\boldsymbol{\beta}$.
  - **Bayesian approach.**
    A priori assign probabilities over the set of models, and obtain posterior model probabilities.

Bayesian model selection advantages



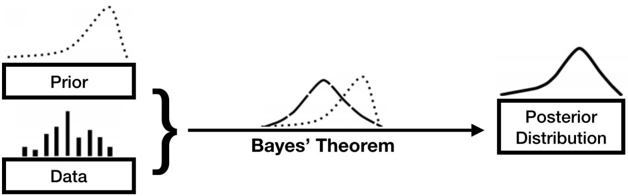
- Discern which model was likeliest to have been behind the data generative process of the observed responses: *highest* probability model, *median* probability model, etc.

Bayesian model selection advantages



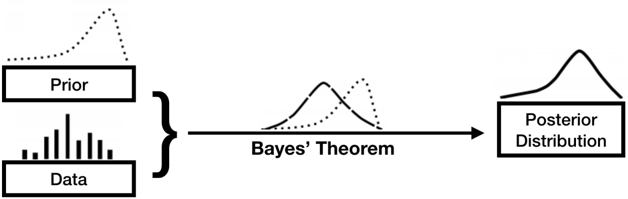
- Discern which model was likeliest to have been behind the data generative process of the observed responses: *highest* probability model, *median* probability model, etc.
- Able to quantify the amount of times a variable "enters" the likeliest of models: *posterior inclusion probabilities*.

Bayesian model selection advantages



- Discern which model was likeliest to have been behind the data generative process of the observed responses: *highest* probability model, *median* probability model, etc.
- Able to quantify the amount of times a variable "enters" the likeliest of models: *posterior inclusion probabilities*.
- At the same time, regression coefficients β are estimated as well.

Bayesian model selection advantages



- Discern which model was likeliest to have been behind the data generative process of the observed responses: *highest* probability model, *median* probability model, etc.
- Able to quantify the amount of times a variable "enters" the likeliest of models: *posterior inclusion probabilities*.
- At the same time, regression coefficients β are estimated as well.
- Ability to combine several (or all!) competing models for inference: Bayesian model averaging (Hoeting et al., 1999).

Bayesian model selection advantages (cont.)

We can use Markov chain Monte Carlo (MCMC) methods to overcome intractability of enumerating all 2^p model probabilities.

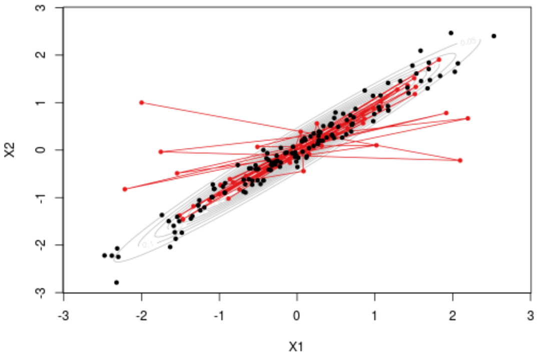
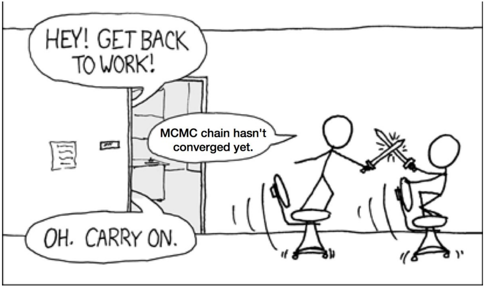


Figure: https://haziqj.shinyapps.io/hmc2/

MCMC is a stochastic method of obtaining random samples from a target posterior distribution.

Bayesian model selection criticisms

- Dancing around the issue–isn't the point *really* to do inference on β? If so, why not simply regularise?
- MCMC is slow and may mix poorly, especially for complex models with many predictors or samples.
- Bayesian approach may require a lot of tuning parameters that need to be set correctly for each unique application.



Kuo & Mallick (KM) model

- Index each of the 2^p possible models by the use of indicator variables $\gamma_j \in \{0, 1\}$ for each variable X_j, $j = 1, \ldots, p$.

Kuo & Mallick (KM) model

- Index each of the 2^p possible models by the use of indicator variables $\gamma_j \in \{0, 1\}$ for each variable X_j, $j = 1, \ldots, p$.
- So $\gamma_j = 1$ if X_j is selected, and 0 otherwise.

Kuo & Mallick (KM) model

- Index each of the 2^p possible models by the use of indicator variables $\gamma_j \in \{0, 1\}$ for each variable X_j, $j = 1, \ldots, p$.
- So $\gamma_j = 1$ if X_j is selected, and 0 otherwise.
  - $\boldsymbol{\gamma} = (0, \ldots, 0)$: Intercept only model.

Kuo & Mallick (KM) model

- Index each of the 2^p possible models by the use of indicator variables $\gamma_j \in \{0, 1\}$ for each variable X_j, $j = 1, \ldots, p$.
- So $\gamma_j = 1$ if X_j is selected, and 0 otherwise.
  - $\boldsymbol{\gamma} = (0, \ldots, 0)$: Intercept only model.
  - $\boldsymbol{\gamma} = (1, \ldots, 1)$: Full model.

Kuo & Mallick (KM) model

- Index each of the 2^p possible models by the use of indicator variables $\gamma_j \in \{0,1\}$ for each variable X_j, $j = 1, \dots, p$.
- So $\gamma_j = 1$ if X_j is selected, and 0 otherwise.
  - $\boldsymbol{\gamma} = (0, \dots, 0)$: Intercept only model.
  - $\boldsymbol{\gamma} = (1, \dots, 1)$: Full model.
  - $\boldsymbol{\gamma} = (0, 1, 0 \dots, 0)$: Model with X_2 only.

Kuo & Mallick (KM) model

- Index each of the 2^p possible models by the use of indicator variables $\gamma_j \in \{0,1\}$ for each variable X_j, $j = 1, \dots, p$.
- So $\gamma_j = 1$ if X_j is selected, and 0 otherwise.
  - $\boldsymbol{\gamma} = (0, \dots, 0)$: Intercept only model.
  - $\boldsymbol{\gamma} = (1, \dots, 1)$: Full model.
  - $\boldsymbol{\gamma} = (0, 1, 0 \dots, 0)$: Model with X_2 only.
- Following Kuo and Mallick (1998),

$$
\begin{aligned}
y_i &= \alpha + \sum_{j=1}^p x_{ij}\gamma_j\beta_j + \epsilon_i \\
\epsilon_i &\overset{\text{iid}}{\sim} \mathrm{N}(0, \sigma^2)
\end{aligned}
\tag{4}
$$

Kuo & Mallick (KM) model

- Index each of the 2^p possible models by the use of indicator variables $\gamma_j \in \{0,1\}$ for each variable X_j, $j = 1,\ldots,p$.
- So $\gamma_j = 1$ if X_j is selected, and 0 otherwise.
  - $\boldsymbol{\gamma} = (0,\ldots,0)$: Intercept only model.
  - $\boldsymbol{\gamma} = (1,\ldots,1)$: Full model.
  - $\boldsymbol{\gamma} = (0,1,0\ldots,0)$: Model with X_2 only.
- Following Kuo and Mallick (1998),

$$\begin{gathered} y_i = \alpha + \sum_{j=1}^p x_{ij}\gamma_j\beta_j + \epsilon_i \\ \epsilon_i \overset{\text{iid}}{\sim} \mathrm{N}(0,\sigma^2) \end{gathered} \tag{4}$$

- Note that we do not consider the intercept to be selectable.

Priors for γ_k

- Independent Bernoulli priors are specified for the model indicators

$$\gamma_j = \begin{cases} 1 & \text{w.p. } \pi_j \\ 0 & \text{w.p. } 1 - \pi_j \end{cases} \tag{5}$$

Priors for γ_k

- Independent Bernoulli priors are specified for the model indicators

$$\gamma_j = \begin{cases} 1 & \text{w.p. } \pi_j \\ 0 & \text{w.p. } 1 - \pi_j \end{cases} \tag{5}$$

- Choices for π_j include
  - $\pi_j = 0.5, \forall j$. This choice reflects equally likely probabilities that any variable/model may be chosen.

Priors for γ_k

- Independent Bernoulli priors are specified for the model indicators

$$\gamma_j = \begin{cases} 1 & \text{w.p. } \pi_j \\ 0 & \text{w.p. } 1 - \pi_j \end{cases} \tag{5}$$

- Choices for π_j include
  - $\pi_j = 0.5, \forall j$. This choice reflects equally likely probabilities that any variable/model may be chosen.
  - Alternatively, adjust each $\pi_j \in [0, 1]$ individually according so some subjective belief.

Priors for γ_k

- Independent Bernoulli priors are specified for the model indicators

$$\gamma_j = \begin{cases} 1 & \text{w.p. } \pi_j \\ 0 & \text{w.p. } 1 - \pi_j \end{cases} \tag{5}$$

- Choices for π_j include
  - $\pi_j = 0.5, \forall j$. This choice reflects equally likely probabilities that any variable/model may be chosen.
  - Alternatively, adjust each $\pi_j \in [0, 1]$ individually according so some subjective belief.
  - Hyperprior on π_j, e.g. $\pi_j \sim \text{Unif}(0, 1)$ or $\pi_j \sim \text{Beta}(1/2, 1, 2)$.

Priors for β

- What's nice about the KM model is that the regression coefficients are set to be independent of the indicator variables.

Priors for $\boldsymbol{\beta}$

- What's nice about the KM model is that the regression coefficients are set to be independent of the indicator variables.
- As such, any usual prior choice of $\boldsymbol{\beta}$ may be used, including
  - The independent prior $\boldsymbol{\beta} \sim \mathrm{N}_p(\mathbf{0}, b^2\mathbf{I}_p)$ for some choice of b (e.g. $b = 10$).

Priors for $\boldsymbol{\beta}$

- What's nice about the KM model is that the regression coefficients are set to be independent of the indicator variables.
- As such, any usual prior choice of $\boldsymbol{\beta}$ may be used, including
  - The independent prior $\boldsymbol{\beta} \sim \mathrm{N}_p(\mathbf{0}, b^2\mathbf{I}_p)$ for some choice of b (e.g. $b = 10$).
  - The g-prior $\boldsymbol{\beta}|\sigma^2 \sim \mathrm{N}_p(\mathbf{0}, g\sigma^2(\mathbf{X}^\top\mathbf{X})^{-1})$ for some value of g either chosen a priori or estimated. Zellner (1986).

Priors for $\boldsymbol{\beta}$

- What's nice about the KM model is that the regression coefficients are set to be independent of the indicator variables.
- As such, any usual prior choice of $\boldsymbol{\beta}$ may be used, including
  - The independent prior $\boldsymbol{\beta} \sim \mathrm{N}_p(\mathbf{0}, b^2\mathbf{I}_p)$ for some choice of b (e.g. $b = 10$).
  - The g-prior $\boldsymbol{\beta}|\sigma^2 \sim \mathrm{N}_p(\mathbf{0}, g\sigma^2(\mathbf{X}^\top\mathbf{X})^{-1})$ for some value of g either chosen a priori or estimated. Zellner (1986).
  - The I-prior $\boldsymbol{\beta}|\sigma^2, \kappa \sim \mathrm{N}_p(\mathbf{0}, \kappa\sigma^2\mathbf{X}^\top\mathbf{X})$ where κ is another scale parameter to be estimated. HJ (2018b).

Priors for $\boldsymbol{\beta}$

- What's nice about the KM model is that the regression coefficients are set to be independent of the indicator variables.
- As such, any usual prior choice of $\boldsymbol{\beta}$ may be used, including
  - The independent prior $\boldsymbol{\beta} \sim \mathrm{N}_p(\mathbf{0}, b^2\mathbf{I}_p)$ for some choice of b (e.g. $b = 10$).
  - The g-prior $\boldsymbol{\beta}|\sigma^2 \sim \mathrm{N}_p(\mathbf{0}, g\sigma^2(\mathbf{X}^\top\mathbf{X})^{-1})$ for some value of g either chosen a priori or estimated. Zellner (1986).
  - The I-prior $\boldsymbol{\beta}|\sigma^2, \kappa \sim \mathrm{N}_p(\mathbf{0}, \kappa\sigma^2\mathbf{X}^\top\mathbf{X})$ where κ is another scale parameter to be estimated. HJ (2018b).
- Typically we want to choose prior choices which maintain conjugacy to the normal regression model.

Priors for $\boldsymbol\beta$ (cont.)

- Note that the Fisher information for $\boldsymbol\beta$ is $I(\boldsymbol\beta) = \sigma^2 \mathbf{X}^\top \mathbf{X}$. This is a measure of the amount of information that the data carries about the unknown parameter $\boldsymbol\beta$.
  - The I-prior has variance proportional to the Fisher information (more data driven).
  - Whereas, the g-prior has variance inversely proportional to the Fisher information.
- The g-prior is a popular choice in model selection due to the algebraic simplifications in the posterior distributions (efficient computations).

$$\tilde{\boldsymbol\beta} = \big(\mathbf{X}^\top \mathbf{X} + \mathbf{B}^{-1}\big)^{-1} \big(\mathbf{B}^{-1}\mathbf{b} + \mathbf{X}^\top \mathbf{y}\big)$$

- The I-prior works well in the presence of multicollinearity (HJ, 2018b).

Priors for $\boldsymbol{\beta}$ (cont.)

- Write $\boldsymbol{\theta} = (\gamma_1\beta_1, \dots, \gamma_p\beta_p)^\top$. Then, the prior on $\boldsymbol{\theta}$ is

$$\boldsymbol{\theta}|\boldsymbol{\gamma} \sim \begin{cases} \mathrm{N}_p(\mathbf{0}, \mathbf{V}_\beta) & \text{w.p. } p(\boldsymbol{\gamma}) \\ \mathbf{0} & \text{w.p. } 1 - p(\boldsymbol{\gamma}) \end{cases} \tag{6}$$

- This is the so-called "spike-and-slab" prior for linear regression models (Mitchell and Beauchamp, 1988; Geweke, 1996).

Priors for $\boldsymbol{\beta}$ (cont.)

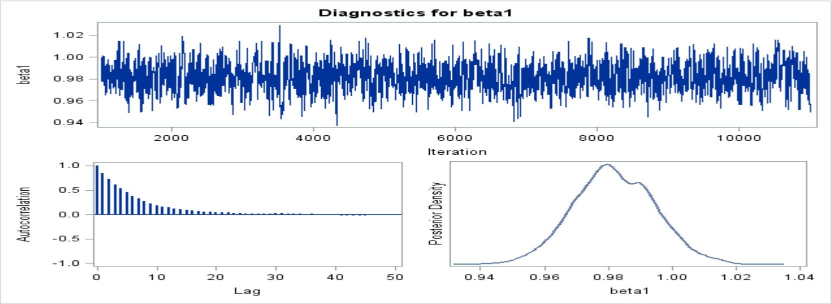
- Write $\boldsymbol{\theta} = (\gamma_1\beta_1, \dots, \gamma_p\beta_p)^\top$. Then, the prior on $\boldsymbol{\theta}$ is

$$\boldsymbol{\theta}|\boldsymbol{\gamma} \sim \begin{cases} \mathrm{N}_p(\mathbf{0}, \mathbf{V}_\beta) & \text{w.p. } p(\boldsymbol{\gamma}) \\ \mathbf{0} & \text{w.p. } 1 - p(\boldsymbol{\gamma}) \end{cases} \tag{6}$$

- This is the so-called "spike-and-slab" prior for linear regression models (Mitchell and Beauchamp, 1988; Geweke, 1996).
- The posterior distribution will also be a mixture of a point mass at zero and a normal density.
  - Regression coefficients are assigned zero values with positive probability.
  - Inference on $\boldsymbol{\theta}$ (the "model-averaged" regression coefficients) is of interest, as these coefficients will have incorporated model uncertainty.

Estimation

- Random samples are drawn from the posterior distributions using Gibbs sampling.
- This can easily be implemented in software such as JAGS.
- R packages exist e.g. `BAS` (Clyde, 2018), `ipriorBVS` (HJ, 2018a).



JAGS model

```
model{
  for (j in 1:p) { gb[j] <- gamma[j] * beta[j] }
  for (i in 1:n) {
    y[i] ~ dnorm(mu[i], psi)
    mu[i] <- alpha + inprod(X[i, 1:p], gb[1:p])
  }

  # Priors
  psi ~ dgamma(0.001, 0.001)
  for (j in 1:p) { gamma[j] ~ dbern(0.5)}
  beta[1:p] ~ dmnorm(rep(0,p), 1/100 * ident_mat)
}

#data# y, X, n, p
#monitor# gamma, alpha, gb, psi
```

ipriorBVS R package

```
(mod <- ipriorBVS(y ~ X, dat))
##              PIP     1     2     3     4     5
## X.1        1.000     x     x     x     x     x
## X.2        0.840     x     x     x     x     x
## X.3        0.568           x     x
## X.4        0.524                 x     x     x
## X.5        0.644     x     x                 x
## X.6        0.294
## X.7        0.480                 x
## X.8        0.238
## PMP              0.061 0.048 0.041 0.040 0.037
## BF               1.000 0.785 0.662 0.648 0.604
## Deviance         93.76 92.07 91.42 96.29 94.16
```

ipriorBVS R package (cont.)

```
coef(mod)
##                PIP   Mean   S.D.   2.5%  97.5%
## (Intercept) 1.000 -0.128  0.459 -1.069  0.739
## X.1         1.000  2.707  0.636  1.588  4.053
## X.2         0.840  1.547  0.878  0.000  2.787
## X.3         0.568  0.607  0.705  0.000  2.119
## X.4         0.524  0.468  0.585 -0.002  1.727
## X.5         0.644  0.858  0.903 -0.110  2.672
## X.6         0.294 -0.158  0.399 -1.273  0.324
## X.7         0.480  0.373  0.523 -0.054  1.582
## X.8         0.238  0.054  0.246 -0.333  0.815
```

Mortality and air pollution data

- Data from McDonald and Schwing (1973) ($n = 60, p = 15$)
  - $y =$ age-adjusted mortality rate.
  - Of interest: Effects of HC, NO_x, SO_2.
  - Environmental variables: precipitation, humidity, temperature.
  - Socioeconomic variables: population density, household size, education, elderly, ethnicity, income.

Mortality and air pollution data

- Data from McDonald and Schwing (1973) ($n = 60, p = 15$)
  - $y =$ age-adjusted mortality rate.
  - Of interest: Effects of HC, NO_x, SO_2.
  - Environmental variables: precipitation, humidity, temperature.
  - Socioeconomic variables: population density, household size, education, elderly, ethnicity, income.
- Clear need to perform model selection, as none of the pollutants were deemed significant in the full model.

Mortality and air pollution data

- Data from McDonald and Schwing (1973) ($n = 60, p = 15$)
  - y = age-adjusted mortality rate.
  - Of interest: Effects of HC, NO_x, SO_2.
  - Environmental variables: precipitation, humidity, temperature.
  - Socioeconomic variables: population density, household size, education, elderly, ethnicity, income.
- Clear need to perform model selection, as none of the pollutants were deemed significant in the full model.
- Comparative approaches
  - Variable elimination using Mallow's C_p as a criterion.
  - Shrinkage (ridge regression).
  - I-prior BVS model.

Results

	OLS	Min. C_p	Ridge	I-prior
Environmental factors				
Precipitation	0.306 (0.14)	0.247 (0.07)	0.230 (0.07)	0.254 (0.12)
Relative humidity	0.009 (0.10)			
January temperature	-0.318 (0.18)	-0.164 (0.06)	-0.172 (0.06)	-0.195 (0.11)
July temperature	-0.237 (0.15)	-0.073 (0.07)		
Demographic factors				
Population density	0.084 (0.09)		0.091 (0.06)	
Household size	-0.232 (0.15)			
Education	-0.233 (0.16)	-0.190 (0.06)	-0.171 (0.07)	-0.151 (0.12)
Sound housing units	-0.052 (0.15)			
Age >65 years	-0.213 (0.20)			
Non-white	0.640 (0.19)	0.481 (0.07)	0.462 (0.07)	0.517 (0.10)
White collar	-0.014 (0.12)			
Income <$3,000	-0.009 (0.22)			
Pollution potential				
HC	-0.979 (0.72)			
NO_x	0.983 (0.75)			
SO_2	0.090 (0.15)	0.255 (0.06)	0.232 (0.06)	0.302 (0.09)
R^2	0.764	0.541	0.553	0.676

Conclusion

- Miller (2002) writes:

  Many statisticians view model selection as "unclean" and "distasteful"... terms such as "fishing expeditions", "torturing the data until they confess", "data mining", and others are used as descriptions of these practices.

Conclusion

- Miller (2002) writes:

  Many statisticians view model selection as "unclean" and "distasteful"... terms such as "fishing expeditions", "torturing the data until they confess", "data mining", and others are used as descriptions of these practices.

- My view: variable selection as an exploratory approach is certainly justified. Further, there is often a genuine need to know the most reasonable, parsimonious and interpretable model.

Conclusion

- Miller (2002) writes:

  Many statisticians view model selection as "unclean" and "distasteful"... terms such as "fishing expeditions", "torturing the data until they confess", "data mining", and others are used as descriptions of these practices.

- My view: variable selection as an exploratory approach is certainly justified. Further, there is often a genuine need to know the most reasonable, parsimonious and interpretable model.

- BVS reduces the the problem of model search into one of estimation–it simultaneously shrinks and select predictors, thereby incorporating model uncertainty.

End

Thank you!

Slides are made available at: https://haziqj.ml/talk/sbe-bvs/

References I

- Clyde, M. (2018). *BAS: Bayesian Variable Selection and Model Averaging using BayesianAdaptive Sampling*. R package version 1.5.3.
- Geweke, J. (1996). "Variable Selection and Model Comparison in Regression". In: *Bayesian Statistics 5*. Proceedings of the Fifth Valencia International Meeting. Ed. by J. M. Bernardo, J. O. Berger, A. P. Dawid, and A. F. M. Smith. Oxford University Press. ISBN: 978-0-19-852356-7.
- Hoeting, J. A., D. Madigan, A. E. Raftery, and C. T. Volinsky (1999). "Bayesian Model Averaging: A Tutorial". *Statistical science* 14.4, pp. 382–401. DOI: 10.1214/ss/1009212519.
- Jamil, H. (2018a). ipriorBVS*: Bayesian Variable Selection using I-priors*. R package version 0.1.1. URL: https://github.com/haziqj/ipriorBVS.
- Kuo, L. and B. Mallick (1998). "Variable selection for regression models". *Sankhyā: The Indian Journal of Statistics, Series B* 60.1, pp. 65–81.

References II

McDonald, G. C. and R. C. Schwing (1973). "Instabilities of Regression Estimates Relating Air Pollution to Mortality". *Technometrics* 15.3, pp. 463–481. DOI: 10.2307/1266852.

Miller, A. (2002). *Subset Selection in Regression*. Chapman & Hall/CRC. ISBN: 978-1-58488-171-1.

Mitchell, T. J. and J. J. Beauchamp (1988). "Bayesian Variable Selection in Linear Regression". *Journal of the American Statistical Association* 83.404, pp. 1023–1032. DOI: 10.2307/2290129.

Jamil, H. (Oct. 2018b). "Regression modelling using priors depending on Fisher information covariance kernels (I-priors)". PhD thesis. London School of Economics and Political Science.

References III

Scott, S. L. and H. R. Varian (Apr. 2015). "Bayesian Variable Selection for Nowcasting Economic Time Series". In: Goldfarb, A., S. M. Greenstein, and C. E. Tucker. *Economic Analysis of the Digital Economy*. University of Chicago Press, pp. 119–135. DOI: 10.7208/chicago/9780226206981.003.0004.

Zellner, A. (1986). "On Assessing Prior Distributions and Bayesian Regression Analysis with g-Prior Distributions". In: *Bayesian Inference and Decision Techniques: Essays in Honor of Bruno de Finetti*. New York: Elsevier, pp. 233–243.

Zhang, L., V. Baladandayuthapani, B. K. Mallick, G. C. Manyam, P. A. Thompson, M. L. Bondy, and K.-A. Do (2014). "Bayesian hierarchical structured variable selection methods with application to molecular inversion probe studies in breast cancer". *Journal of the Royal Statistical Society: Series C (Applied Statistics)* 63.4, pp. 595–620.