Model Selection and Variable Selection

Bayesian Methodology in Biostatistics (BST 249)

Jeffrey W. Miller

Department of Biostatistics
Harvard T.H. Chan School of Public Health

Outline

Introduction

Variable selection

Oxygen update example

Diabetes example

Bayesian variable selection

Oxygen uptake example, revisited

Gibbs sampling for Bayesian variable selection

Diabetes example, revisited

Outline

Introduction

Variable selection

- Oxygen update example
- Diabetes example

Bayesian variable selection

- Oxygen uptake example, revisited

Gibbs sampling for Bayesian variable selection

- Diabetes example, revisited

Introduction

- When using regression in practice, many possible variables could be included as covariates.
- Including too many variables can lead to poor prediction performance.
- Ideally, one would include only the variables that are informative about the outcome.
- But how can one determine which variables are informative?
- This is the variable selection problem.
- Variable selection is a particular type of model selection.

Bayesian model selection

- The standard Bayesian approach to model selection is to put a prior on models, and simply consider the posterior on models.

- This is sometimes referred to as “Bayesian model averaging”, rather than “model selection”.

- The “model averaging” terminology emphasizes the fact that we use the posterior over all models rather than selecting a single model.

- Thus, posterior expectations involve averaging over models.

Bayesian model selection

- Suppose we are considering models $\mathcal{M}_1, \mathcal{M}_2, \ldots$.

- Suppose model $\mathcal{M}_k$ has parameter θ_k.

- Suppose we have a prior on the model index $p(k)$, and a prior on parameters $p(\theta_k \mid k)$ for each model.

- Given data x, the posterior on models is then

$$
\begin{aligned}
p(k|x) &\propto p(x|k)\, p(k) \\
&= p(k) \int p(x \mid \theta_k,\, k)\, p(\theta_k \mid k)\, d\theta_k.
\end{aligned}
$$

- $p(x|k)$ is the marginal likelihood of model k. Incidentally, this justifies the term "marginal likelihood", since it is the likelihood function for k.

Outline

Introduction

Variable selection

- Oxygen update example
- Diabetes example

Bayesian variable selection

- Oxygen uptake example, revisited

Gibbs sampling for Bayesian variable selection

- Diabetes example, revisited

Variable selection

- Variable selection is a special case of model selection.
- Consider each subset of possible variables to be a different model.
- Represent this as follows: $z_j = 1$ if variable j is included, and $z_j = 0$ otherwise.
- Consider the linear regression model:

$$Y_i = z_1 \beta_1 x_{i1} + \cdots + z_p \beta_p x_{ip} + \varepsilon_i.$$

- Each possible vector $z = (z_1, \ldots, z_p) \in \{0,1\}^p$ represents a different model.
- So, for model selection, we are interested in $p(z \mid x, y)$.

Oxygen update example: Data

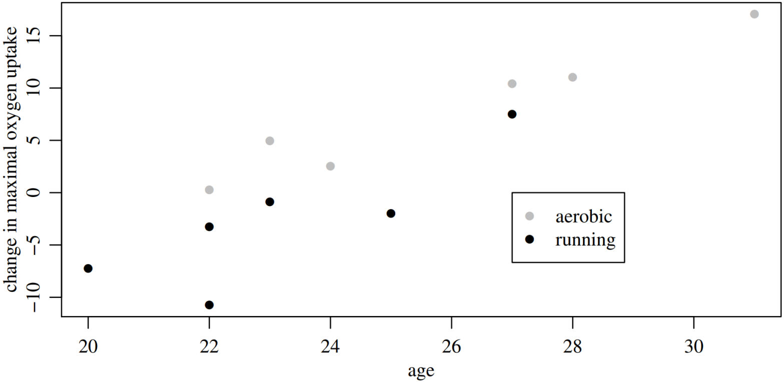
- Twelve healthy men were recruited to take part in a study on the effects of exercise. The men did not exercise regularly.

- Subjects were randomly divided into two groups of six.
  - Group 1 followed a 12-week running program.
  - Group 2 followed a 12-week step aerobics program.

- Maximum oxygen uptake (liters/minute) was measured while running on a treadmill, both before and after the program.

- Goal: Assess the effect of the running program on oxygen uptake.

Oxygen update example: Data



Change in maximal oxygen uptake as a function of age and exercise program.

Oxygen update example: Models considered

- Five models under consideration:

$$\mathrm{E}\big(Y \mid X,\, \beta,\, z = (1,0,0,0)\big) = \beta_1$$

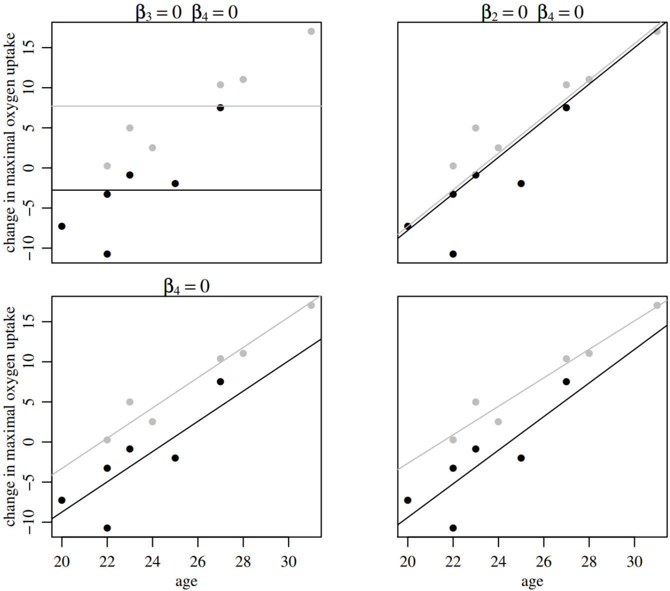
$$\mathrm{E}\big(Y \mid X,\, \beta,\, z = (1,1,0,0)\big) = \beta_1 + \beta_2 \text{ group}$$

$$\mathrm{E}\big(Y \mid X,\, \beta,\, z = (1,0,1,0)\big) = \beta_1 + \beta_3 \text{ age}$$

$$\mathrm{E}\big(Y \mid X,\, \beta,\, z = (1,1,1,0)\big) = \beta_1 + \beta_2 \text{ group} + \beta_3 \text{ age}$$

$$\begin{aligned} &\mathrm{E}\big(Y \mid X,\, \beta,\, z = (1,1,1,1)\big) \\ &\quad = \beta_1 + \beta_2 \text{ group} + \beta_3 \text{ age} + \beta_4 \text{ group} \times \text{age}. \end{aligned}$$

Oxygen update example: OLS for four models



Least squares regression lines for the oxygen uptake data, under four different models.

Oxygen update example: Motivation

- Visually, some of these models clearly seem better than others.

- Statistically, how much evidence is there for each model?

- A frequentist approach would employ hypothesis testing, but this gets complicated when there are many possible variables to include.

- It is common to use lasso, elastic net, or stepwise selection, but significance testing is not simple with these methods. Recent methods for “post-selection inference” have been developed to address this.

- The Bayesian approach is simply to consider the posterior on which variables to include.

Diabetes example: Data

- $n = 442$ patients, 10 binary measurements for each patient.

- Outcome: $y_i =$ quantitative measure of disease progression of patient i, one year after measurements.

- Goal: Predict y_i from the measurements.

- Baseline model: Linear regression with main effects as well as interactions between each pair of measurements.

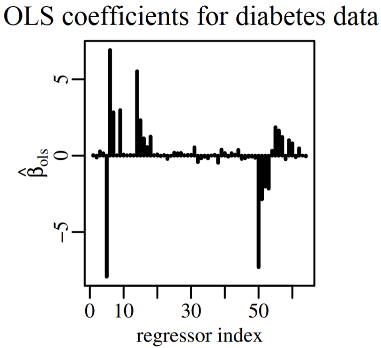
- The baseline model has $p = 64$ covariates, which are centered and scaled for interpretability.

- 342 patients used for training, 100 patients used for testing.

Diabetes example: Performance comparison

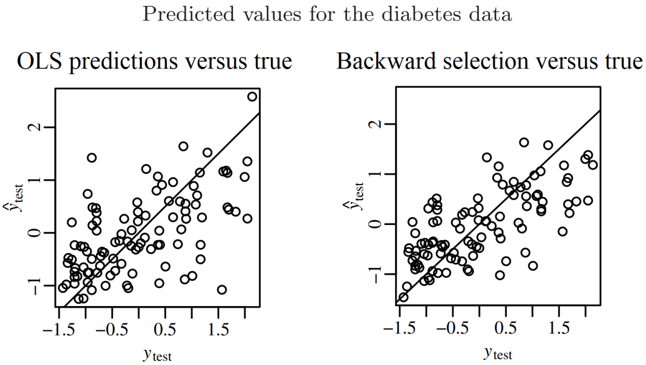
- Approach #1: Constant prediction $\hat{y}_i = c$. Test MSE $= 0.97$.
- Approach #2: OLS including all covariates. Test MSE $= 0.67$.
- Approach #3: OLS + backward selection. Test MSE $= 0.53$.
- Approach #4: Bayesian variable selection. Test MSE $= 0.45$.

Diabetes example: OLS coefficients



(figure from Hoff, 2009)

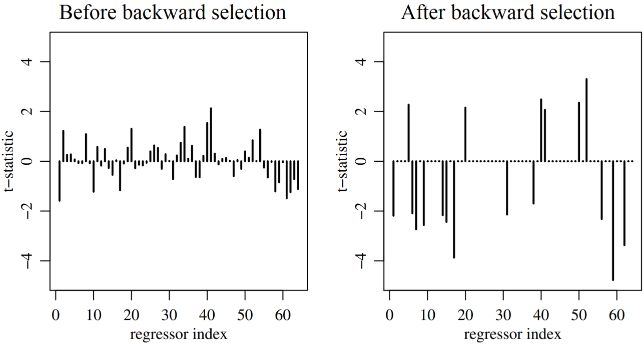
Diabetes example: OLS versus backward selection



(figure from Hoff, 2009)

- Backward selection starts with all variables included, and iteratively removes the least significant variable and refits, until all variables are significant. This leaves 20 variables.
- Backward selection helps improve prediction performance.

Diabetes example: Issue with backward selection



t-statistics for permuted diabetes data (none should be significant)

(figure from Hoff, 2009)

- Unfortunately, backward selection often indicates relationships even when there are none.
- Above, the y values were randomly permuted, breaking any dependencies with the x's.
- Backward selection incorrectly finds many "significant" coefficients.

Outline

Introduction

Variable selection

- Oxygen update example
- Diabetes example

Bayesian variable selection

- Oxygen uptake example, revisited

Gibbs sampling for Bayesian variable selection

- Diabetes example, revisited

Bayesian variable selection

- Consider the linear regression model:

  $$Y_i = z_1\beta_1 x_{i1} + \cdots + z_p\beta_p x_{ip} + \varepsilon_i$$

  where $\varepsilon_i \sim \mathcal{N}(0, \gamma^{-1})$.

- Prior on models: $z \sim p(z)$, e.g., $Z_1, \ldots, Z_p \overset{\text{iid}}{\sim} \text{Bernoulli}(\alpha)$.

- Prior on parameters for each model:

  $$\begin{aligned} \gamma &\sim \text{Gamma}(\tfrac{1}{2}\nu_0,\ \tfrac{1}{2}\nu_0\sigma_0^2) \\ \beta_z | X, z, \gamma &\sim \mathcal{N}(0,\ g(\gamma X_z^{\mathtt{T}} X_z)^{-1}) \end{aligned}$$

  where $\beta_z = (\beta_j : z_j = 1)$ contains the entries of β where $z_j = 1$, and likewise, X_z is the design matrix including only the columns j for which $z_j = 1$. If $z_j = 0$, then set β_j to any arbitary real value, e.g., $\beta_j = 0$.

- Thus, given z, this is a g-prior on β_z.

Bayesian variable selection: Posterior computation

- The posterior on models z, integrating out β and γ, is

$$p(z|X,y) \propto p(y|X,z)p(z).$$

- It turns out that $p(y|X,z)$ can be computed analytically (see Hoff 9.3.1):

$$\begin{aligned} p(y|X,z) &= \int p(y \mid \beta,\gamma,X,z)p(\beta \mid \gamma,X,z)p(\gamma)\,d\beta\,d\gamma \qquad (1)\\ &= \frac{1}{\pi^{n/2}(1+g)^{p_z/2}}\frac{\Gamma\big(\frac{1}{2}(\nu_0+n)\big)}{\Gamma\big(\frac{1}{2}\nu_0\big)}\frac{(\nu_0\sigma_0^2)^{\nu_0/2}}{(\nu_0\sigma_0^2+\mathrm{SSR}_g^z)^{(\nu_0+n)/2}} \end{aligned}$$

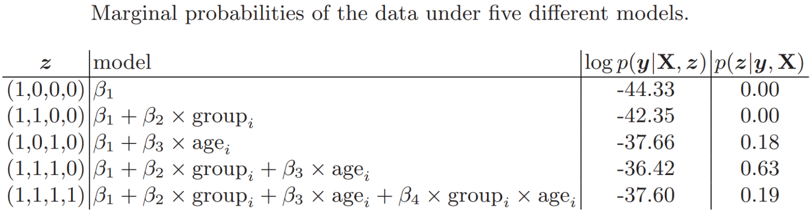
where $p_z = \sum_{j=1}^{p} z_j$ and

$$\mathrm{SSR}_g^z = y^{\mathtt{T}}\Big(I - \frac{g}{g+1}X_z(X_z^{\mathtt{T}}X_z)^{-1}X_z^{\mathtt{T}}\Big)y.$$

Bayesian variable selection: Prior settings

- A natural default is to use a unit information prior.
- Unit information prior: $g = n$, $\nu_0 = 1$, and $\sigma_0^2 = \hat{\sigma}^2_{\mathsf{mle}}$.
- For the prior on models, in these examples, we will consider a uniform prior $p(z) \propto 1$. Equivalently, $Z_j \overset{\text{iid}}{\sim} \text{Bernoulli}(1/2)$.
- However, to favor sparsity, it is common to use $Z_j \overset{\text{iid}}{\sim} \text{Bernoulli}(\alpha)$ where α is of order $1/p$.
- For instance, if $\alpha = c/p$ then (under the prior) the proportion of included coefficients is c.
- It is also common (and recommended) to integrate out a Beta prior on α, which can easily be done analytically since it is conjugate.

Oxygen uptake example: Posterior on models



Marginal probabilities of the data under five different models.

$\boldsymbol{z}$	model	$\log p(\boldsymbol{y}\|\mathbf{X}, \boldsymbol{z})$	$p(\boldsymbol{z}\|\boldsymbol{y}, \mathbf{X})$
(1,0,0,0)	β_1	-44.33	0.00
(1,1,0,0)	$\beta_1 + \beta_2 \times \text{group}_i$	-42.35	0.00
(1,0,1,0)	$\beta_1 + \beta_3 \times \text{age}_i$	-37.66	0.18
(1,1,1,0)	$\beta_1 + \beta_2 \times \text{group}_i + \beta_3 \times \text{age}_i$	-36.42	0.63
(1,1,1,1)	$\beta_1 + \beta_2 \times \text{group}_i + \beta_3 \times \text{age}_i + \beta_4 \times \text{group}_i \times \text{age}_i$	-37.60	0.19

(figure from Hoff, 2009)

- The posterior favors the model including age and the exercise group, but not the interaction.

- However, a sizable amount of posterior mass is also given to the other two models that include age.

- Thus, according to this analysis, the data provides some evidence that the type of exercise program has an effect, but it is not definitive.

Outline

Introduction

Variable selection

- Oxygen update example
- Diabetes example

Bayesian variable selection

- Oxygen uptake example, revisited

Gibbs sampling for Bayesian variable selection

- Diabetes example, revisited

Gibbs sampling for Bayesian variable selection

- In the oxygen uptake example, there are only 5 models under consideration, so we can easily analytically compute the posterior over all 5 models.
- However, when there are more variables, the number of models will be far too large to consider them all.
- If we consider all subsets of p variables, there are 2^p possible models.
- For instance, for the diabetes data, $p = 64$, so there are around 1.8×10^{19} models!
- Gibbs sampling is a common approach to doing approximate posterior inference for variable selection.

Gibbs sampling for Bayesian variable selection

- As mentioned earlier, one can integrate out β and γ to obtain an analytic expression for $p(y|X, z)$.

- We can use this expression to do Gibbs sampling directly on $p(z|X, y)$, since $p(z|X, y) \propto p(y|X, z)p(z)$.

- **Gibbs sampler algorithm for variable selection**
  - Initialize $z_1 = \cdots = z_p = 0$.
  - At each iteration, for each $j = 1, \dots, J$

    Update z_j by sampling from $p(z_j \mid X, y, z_{-j})$.
- Here, z_{-j} denotes all the entries of z except z_j.

- Note: Initializing $z_j = 0$ speeds up burn-in when p is very large and the posterior is concentrated on sparse z vectors.

Gibbs sampling for Bayesian variable selection

- The full conditional for z_j is

$$p(z_j \mid X, y, z_{-j}) \underset{z_j}{\propto} p(y|X,z)p(z)$$

  and $p(y|X,z)$ is given by Equation 1 (earlier in these slides).

- This can be written as

$$\begin{aligned} p(z_j = 1 \mid X, y, z_{-j}) &= r_j/(1+r_j) \\ p(z_j = 0 \mid X, y, z_{-j}) &= 1/(1+r_j) \end{aligned}$$

  where

$$r_j = r_j(z_{-j}) = \frac{p(y \mid X, z_{-j}, z_j{=}1)}{p(y \mid X, z_{-j}, z_j{=}0)} \frac{p(z_{-j}, z_j{=}1)}{p(z_{-j}, z_j{=}0)}.$$

Gibbs sampling for Bayesian variable selection

- A Gibbs sampler in which some of the variables have been integrated out is sometimes called a "collapsed Gibbs sampler".

- What if we want to get posterior samples of β and γ as well? It turns out that this is easy to do.

- First run the Gibbs sampler for T iterations to get samples $z^{(1)}, \ldots, z^{(T)}$ from $p(z|X, y)$.

- Then, for each $t = 1, \ldots, T$,
  1. Sample $\gamma^{(t)} \sim p(\gamma \mid X, y, z^{(t)})$.
  2. Sample $\beta^{(t)} \sim p(\beta \mid X, y, z^{(t)}, \gamma^{(t)})$.

- Sampling $\gamma^{(t)}$ and $\beta^{(t)}$ can be done exactly, using the formulas from the slides on g-priors for Bayesian linear regression.

Diabetes example: Posterior

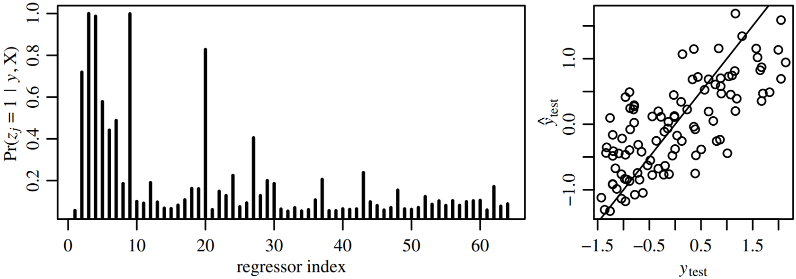
- Run sampler for $T = 10000$ iterations, to get $z^{(t)}, \gamma^{(t)}, \beta^{(t)}$ for $t = 1, \ldots, T$.

- Only a small fraction of the total number of models are ever visited by the sampler.

- However, if the sampler is performing reasonably well, then the set of models that are not visited should have small posterior probability.

- Further, the samples often provide a reasonable approximation to the marginal distribution of each z_j and β_j, even if the joint posterior on z, β is not very well approximated.

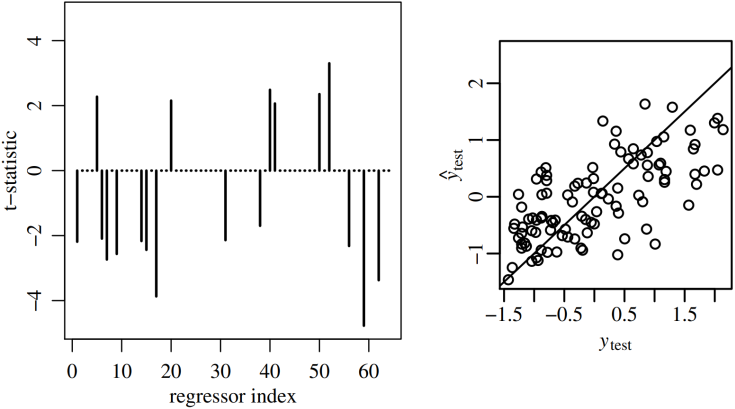
Diabetes example: Posterior



The first panel shows posterior probabilities that each coefficient is non-zero. The second panel shows y_{test} versus predictions based on the model averaged estimate of $\boldsymbol{\beta}$.

- Test MSE $= 0.45$ for Bayesian variable selection.
- Bayes selects quite different variables than backward selection.

Diabetes example: Compare with backward selection



- Test MSE $= 0.53$ for backward selection.

References and supplements

- Hoff, P. D. (2009). A first course in Bayesian statistical methods (Vol. 580). New York: Springer.
- George, E. I., & McCulloch, R. E. (1993). Variable selection via Gibbs sampling. Journal of the American Statistical Association, 88(423), 881-889.