STA 602 - Intro to Bayesian Statistics

Lecture 17

Li Ma

Duke University

The regression problem

- So far when we have multivariate observations, we have attempted to model the joint distribution of all the variables.
- An alternative strategy is to model the conditional distribution of some of the variables given the others.
- The most common situation is to model the conditional distribution of one variable of interest y, often called the *response* or *outcome* variable, given the other variables, called the *covariates* or *predictors*.
- So the general regression problem is the modeling of the conditional distribution:

When $x = \mathbb{1}_n$, regression → unknown mean problem

$$p(y|\mathbf{x}).$$

- On the spectrum of model flexibility, one extreme is the Gaussian linear model and the other extreme is the nonparametric density regression.

Gaussian linear regression

- The model for the conditional distribution of the outcome is

$$y|\mathbf{x}, \boldsymbol{\beta}, \sigma^2 \sim \mathrm{N}(\mathbf{x}'\boldsymbol{\beta}, \sigma^2)$$

  where $\boldsymbol{\beta} = (\beta_0, \beta_1, \beta_2, \ldots, \beta_p)'$ are the so-called regression coefficients for the covariates $\mathbf{x} = (1, x_0, x_1, \ldots, x_p)'$
- When there are multiple observations $(y_1, \mathbf{x}_1), (y_2, \mathbf{x}_2), \ldots, (y_n, \mathbf{x}_n)$, we often adopt the independent assumption (conditional independence).

$$y_i \,|\, \mathbf{x}_i, \beta, \sigma^2 \overset{\text{ind}}{\sim} \mathrm{N}(\mathbf{x}_i'\beta, \sigma^2).$$

- Such conditional independence assumption corresponds to the notion of *conditional exchangeability*—that is, the order of the observations don't matter once we are given the values of the covariates.

An alternative representation

- An equivalent way to write down the Gaussian linear regression model is

$$y_i = \mathbf{x}_i'\beta + \varepsilon_i \quad \text{where } \varepsilon_i \overset{\text{iid}}{\sim} \mathrm{N}(0, \sigma^2).$$

- At the end of the day, a model is a set of assumptions. Both representation convey the same set of assumptions:
- y_i given $\mathbf{x}_i$ is independent Gaussian.
- Its conditional mean taken the linear form

$$\mathrm{E}(y_i|\mathbf{x}_i) = \mathbf{x}_i'\beta.$$

- The conditional variance of y_i given $\mathbf{x}_i$ are all equal to σ^2.
- Each of these assumptions can be relaxed leading to a more general version of the regression model.

Matrix representation

- One can write the model in matrix form

$$\mathbf{y} \sim \mathrm{N}(\mathbf{X}\boldsymbol{\beta}, \sigma^2 I)$$

where

$$\mathbf{y}_{n\times 1} = \begin{pmatrix} y_1 \\ y_2 \\ \vdots \\ y_n \end{pmatrix} \quad \mathbf{X}_{n\times(p+1)} = \begin{pmatrix} \mathbf{x}_1' \\ \mathbf{x}_2' \\ \vdots \\ \mathbf{x}_n' \end{pmatrix} = \begin{pmatrix} 1 & x_{11} & x_{12} & \cdots & x_{1p} \\ 1 & x_{21} & x_{22} & \cdots & x_{2p} \\ \vdots & \vdots & \vdots & \vdots & \vdots \\ 1 & x_{n1} & x_{n2} & \cdots & x_{np} \end{pmatrix}$$

and $I_{(p+1)\times(p+1)}$ is the $(p+1)\times(p+1)$ identity matrix.

- Or equivalently

$$\mathbf{y} = \mathbf{X}\boldsymbol{\beta} + \boldsymbol{\varepsilon}$$

where

$$\boldsymbol{\varepsilon}_n = \begin{pmatrix} \varepsilon_1 \\ \varepsilon_2 \\ \vdots \\ \varepsilon_n \end{pmatrix} \sim \mathrm{N}(0, \sigma^2 I).$$

Bayesian inference on Gaussian linear regression

- Bayesian inference can proceed as usual after we specify priors for the unknown parameters.
- The only difference is that the model is specified in terms of $\mathbf{y}$ *given* $\mathbf{X}$, and so we always condition on the value of $\mathbf{X}$ in our of our likelihood, prior, and posterior.
- As such, for simplicity usually we don't explicitly write in our equations given $\mathbf{X}$, but that should be assumed throughout this lecture.

The likelihood

- The joint probability of $\mathbf{y}$ given the parameters is simply that for a multivariate normal as usual

$$p(\mathbf{y}|\boldsymbol{\beta},\sigma^2) = (2\pi)^{-\frac{n}{2}}|\sigma^2 I|^{-\frac{1}{2}} e^{-\frac{1}{2}(\mathbf{y}-\mathbf{X}\boldsymbol{\beta})'(\sigma^2 I)^{-1}(\mathbf{y}-\mathbf{X}\boldsymbol{\beta})}$$
$$\propto (\sigma^2)^{-\frac{n}{2}} e^{-\frac{(\mathbf{y}-\mathbf{X}\boldsymbol{\beta})'(\mathbf{y}-\mathbf{X}\boldsymbol{\beta})}{2\sigma^2}}.$$

$(y-x\beta)^T\Sigma^{-1}(y-x\beta) = \beta^T x^T\Sigma^{-1}x\beta - 2\beta^T x^T\Sigma^{-1}y^T + C.$

- This could be attained using the equation perspective as well by taking products over n independent observations.

$$p(\mathbf{y}_1,\ldots,\mathbf{y}_n|\boldsymbol{\beta},\sigma^2) = \prod_{i=1}^{n}(2\pi\sigma^2)^{-\frac{1}{2}} e^{-\frac{(y_i-\mathbf{x}_i'\boldsymbol{\beta})^2}{2\sigma^2}}$$
$$\propto (\sigma^2)^{-\frac{n}{2}} e^{-\frac{\sum_i(y_i-\mathbf{x}_i'\boldsymbol{\beta})^2}{2\sigma^2}}.$$

- However, the matrix representation is more concise and convenient, and readily generalizable (e.g., correlated errors). So we shall stick with it.

Classical inference

- One can find the MLEs for $\boldsymbol{\beta}$ and σ^2 by maximizing the likelihood with respect to those parameters. This can be done by—(i) first maximizing the likelihood w.r.t. $\boldsymbol{\beta}$ for fixed σ, then (ii) maximize over σ.
- Maximizing the likelihood w.r.t. $\boldsymbol{\beta}$ for fixed σ boils down to minimizing the "sum of squared residuals":

$$\text{SSR}(\boldsymbol{\beta}) = (\mathbf{y} - \mathbf{X}\boldsymbol{\beta})'(\mathbf{y} - \mathbf{X}\boldsymbol{\beta}) = \sum_{i=1}^{n} (y_i - \mathbf{x}_i'\boldsymbol{\beta})^2.$$

- The solution is called the ordinary least squares (OLS) estimator

$$\hat{\boldsymbol{\beta}}_{ols} = \text{argmin}_{\boldsymbol{\beta}} \, \text{SSR}(\boldsymbol{\beta}),$$

  which exists when $\mathbf{X}$ has rank $p+1$, which can be solved by

$$\frac{d}{d\boldsymbol{\beta}} \text{SSR}(\boldsymbol{\beta}) = 2\mathbf{X}'(\mathbf{y} - \mathbf{X}\boldsymbol{\beta}) = \mathbf{0} \Rightarrow \hat{\boldsymbol{\beta}}_{ols} = (\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'\mathbf{y}.$$

- $\hat{\boldsymbol{\beta}}$ is BLUE (best *linear unbiased* estimator), where "best" is in the sense of achieving the smallest MSE.

A semi-conjugate prior specification

- It is easy to recognize that the likelihood has a quadratic function in terms of $\boldsymbol{\beta}$ in the exponent.
- So one can imagine that a multivariate normal prior on $\boldsymbol{\beta}$ will lead to conjugate full conditionals of $\boldsymbol{\beta}$.
- So we use

$$\boldsymbol{\beta} \sim \mathrm{N}(\boldsymbol{\beta}_0, \Sigma_0)$$

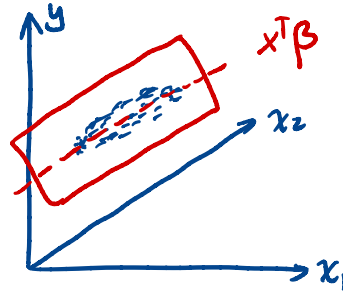
- For σ^2, we can use the usual inverse-Gamma prior

$$\sigma^2 \sim \mathrm{IG}(\nu_0/2, \nu_0\sigma_0^2/2).$$

- Also, we assume prior independence

$$p(\boldsymbol{\beta}, \sigma^2) = p(\boldsymbol{\beta})p(\sigma^2).$$

The full conditionals



- The full conditional of $\boldsymbol{\beta}$ is then

$$p(\boldsymbol{\beta}|\mathbf{y},\sigma^2) \propto p(\mathbf{y}|\boldsymbol{\beta},\sigma^2)p(\boldsymbol{\beta})$$

$(X^TX)_{kk} = \sum_{i=1}^{n} x_{ik}^2$

$$\propto e^{-\frac{1}{2\sigma^2}(\mathbf{y}-\mathbf{X}\boldsymbol{\beta})'(\mathbf{y}-\mathbf{X}\boldsymbol{\beta})} \cdot e^{-\frac{1}{2}(\boldsymbol{\beta}-\boldsymbol{\beta}_0)'\Sigma_0^{-1}(\boldsymbol{\beta}-\boldsymbol{\beta}_0)}$$

Large norm → small beta → small variance

$$\propto e^{-\frac{1}{2}\left[\boldsymbol{\beta}'\left(\frac{\mathbf{X}'\mathbf{X}}{\sigma^2}\right)\boldsymbol{\beta}-2\boldsymbol{\beta}'\left(\frac{\mathbf{X}'\mathbf{y}}{\sigma^2}\right)\right]} \cdot e^{-\frac{1}{2}(\boldsymbol{\beta}'\Sigma_0^{-1}\boldsymbol{\beta}-2\boldsymbol{\beta}'\Sigma^{-1}\boldsymbol{\beta}_0)}$$

Off diagonal → related covariates → small precision → large variance.

$$\propto e^{-\frac{1}{2}\boldsymbol{\beta}'\left(\frac{\mathbf{X}'\mathbf{X}}{\sigma^2}+\Sigma_0^{-1}\right)\boldsymbol{\beta}+\boldsymbol{\beta}'\left(\frac{\mathbf{X}'\mathbf{y}}{\sigma^2}+\Sigma_0^{-1}\boldsymbol{\beta}_0\right)}$$

* If X^TX is non-invertible, then the posterior largely depend on prior.

which is

$$\mathrm{N}(\boldsymbol{\beta}_n, \Sigma_n)$$

* using Bayesian shrinkage to protect from overfitting. "ridge estimator"

where When $X = \mathbb{1}_n$, $X^TX = n$

$$\Sigma_n^{-1} = \frac{\mathbf{X}'\mathbf{X}}{\sigma^2} + \Sigma_0^{-1} \quad \text{and} \quad \boldsymbol{\beta}_n = \Sigma_n\left(\frac{\mathbf{X}'\mathbf{y}}{\sigma^2} + \Sigma_0^{-1}\boldsymbol{\beta}_0\right)$$

- Note that

$$\boldsymbol{\beta}_n = \Sigma_n\left(\frac{\mathbf{X}'\mathbf{X}}{\sigma^2}\underbrace{(\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'\mathbf{y}}_{\hat{\boldsymbol{\beta}}_{OLS}} + \Sigma_0^{-1}\boldsymbol{\beta}_0\right) = \Sigma_n\left(\frac{\mathbf{X}'\mathbf{X}}{\sigma^2}\hat{\boldsymbol{\beta}}_{ols} + \Sigma_0^{-1}\boldsymbol{\beta}_0\right)$$

which is a weighted average between $\hat{\boldsymbol{\beta}}_{ols}$ and the prior mean $\boldsymbol{\beta}_0$.

The full conditionals

- The full conditional for σ^2 is then (letting $\gamma = 1/\sigma^2$)

$$\begin{aligned} p(\gamma|\mathbf{y},\boldsymbol{\beta}) &\propto p(\mathbf{y}|\boldsymbol{\beta},\gamma)p(\gamma) \\ &\propto \gamma^{n/2} e^{-\frac{\gamma}{2}(\mathbf{Y}-\mathbf{X}\boldsymbol{\beta})'(\mathbf{Y}-\mathbf{X}\boldsymbol{\beta})} \cdot \gamma^{\frac{\nu_0}{2}-1} e^{-\frac{\nu_0\sigma_0^2}{2}\gamma} \\ &\propto \gamma^{\frac{\nu_0+n}{2}-1} e^{-\frac{\gamma}{2}\left(\nu_0\sigma_0^2+(\mathbf{Y}-\mathbf{X}\boldsymbol{\beta})'(\mathbf{Y}-\mathbf{X}\boldsymbol{\beta})\right)} \end{aligned}$$

  which is

$$\gamma|\mathbf{y},\boldsymbol{\beta} \sim \text{Gamma}\left(\frac{\nu_0+n}{2}, \frac{\nu_0\sigma_0^2+\text{SSR}(\boldsymbol{\beta})}{2}\right).$$

  or

$$\sigma^2|\mathbf{y},\boldsymbol{\beta} \sim \text{IG}\left(\frac{\nu_0+n}{2}, \frac{\nu_0\sigma_0^2+\text{SSR}(\boldsymbol{\beta})}{2}\right).$$

- Given the full conditionals, Gibbs sampling can proceed as usual.

The invarance principle

If we specify $\beta|\sigma^2 \sim N(\beta_0, \sigma^2 \cdot \Sigma_0)$, then we'll get conjugate posterior.

- One often would want to ensure that our inference is invariant with respect to transforms of the covariates.
- That is, if we happen to be use a new set of covariates

$$\tilde{\mathbf{X}} = \mathbf{X}H$$

$\tilde{X}\tilde{\beta} = XH\tilde{\beta} = X\beta \Rightarrow H\tilde{\beta} = \beta$

Just like Jeffrey's prior.

for some invertible matrix H. Then $\tilde{\mathbf{X}}\tilde{\boldsymbol{\beta}} = \mathbf{X}\boldsymbol{\beta}$ where $\tilde{\boldsymbol{\beta}} = H^{-1}\boldsymbol{\beta}$.

- The way we specify a prior on $\boldsymbol{\beta}$ should induce the same prior had we been specifying the prior on $\tilde{\boldsymbol{\beta}}$.
- This is satsified by letting

$\tilde{\beta} = H^{-1}\beta \sim N(0, H^{-1} \cdot k(X^TX)^{-1}(H^{-1})^T)$

$$\boldsymbol{\beta} \sim \mathrm{N}(\mathbf{0}, k(\mathbf{X}'\mathbf{X})^{-1}).$$

$= N(0, k \cdot (H^T X^T X H)^{-1})$

$= N(0, k \cdot (\tilde{X}^T \tilde{X})^{-1})$

- Verify this by change of variable formula:

$$\boldsymbol{\beta} \sim \mathrm{N}(\mathbf{0}, k(\mathbf{X}'\mathbf{X})^{-1}) \quad \Leftrightarrow \quad \tilde{\boldsymbol{\beta}} \sim \mathrm{N}(\mathbf{0}, k(\tilde{\mathbf{X}}'\tilde{\mathbf{X}})^{-1}).$$

Zellner's g-prior

- In addition to the invariance consideration, we can also make the prior covariance for $\boldsymbol{\beta}$ proportional to σ^2, which makes the prior fully conjugate:

$$\boldsymbol{\beta} \mid \sigma^2 \sim \mathrm{N}(\mathbf{0}, g\sigma^2(\mathbf{X}'\mathbf{X})^{-1}).$$

- This prior is fully conjugate and invariant with respect to transforms of the covariates.
- It is called Zellner's *g-prior*.
- Conjugacy here is nice but not as important in making the posterior simple as it allows the *exact evaluation of the normalizing constant* in Bayes theorem, which will allow us to carry out *model selection*.

The posterior under the g-prior

- The full conditional of $\boldsymbol{\beta}$ is derived the same way as before

$$\boldsymbol{\beta} \,|\, \mathbf{y}, \sigma^2 \sim \mathrm{N}(\boldsymbol{\beta}_n, \Sigma_n)$$

and now with $\boldsymbol{\beta}_0 = \mathbf{0}$ and $\Sigma_0 = g\sigma^2(\mathbf{X}'\mathbf{X})^{-1}$,

$$\Sigma_n^{-1} = \frac{\mathbf{X}'\mathbf{X}}{\sigma^2} + \Sigma_0^{-1} = \frac{\mathbf{X}'\mathbf{X}}{\sigma^2} + \frac{\mathbf{X}'\mathbf{X}}{g\sigma^2} = \left(1 + \frac{1}{g}\right)\frac{\mathbf{X}'\mathbf{X}}{\sigma^2}.$$

and

larger g → larger prior variance → weaker prior belief → less shrinkage

$$\boldsymbol{\beta}_n = \Sigma_n\left(\frac{\mathbf{X}'\mathbf{X}}{\sigma^2}\hat{\boldsymbol{\beta}}_{ols} + \Sigma_0^{-1}\boldsymbol{\beta}_0\right) = \frac{g}{1+g}\hat{\boldsymbol{\beta}}_{ols} = \frac{g}{1+g}(\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'\mathbf{y}.$$

- So the parameter g plays the role of a shrinkage parameter.
- To incorporate adaptive shrinkage, place a prior on g—e.g., the hyper-g prior (Liang et al 2008).

The posterior under the g-prior

- Under the g-prior, due to conjugacy, we can actually derive the marginal posterior of σ^2 in closed form. (The details are given in the textbook.)
- Specifically,

$$\sigma^2 \,|\, \mathbf{y} \sim \text{IG}\left(\frac{\nu_n}{2}, \frac{\nu_n \sigma_n^2}{2}\right)$$

where

$$\nu_n = \nu_0 + n \quad \text{and} \quad \nu_n \sigma_n^2 = \nu_0 \sigma_0^2 + \text{SSR}_g$$

with

$$\text{SSR}_g = \mathbf{y}' \left(I - \frac{g}{1+g} \mathbf{X}(\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'\right) \mathbf{y}.$$

- In contrast, note that

$$\text{SSR}(\hat{\boldsymbol{\beta}}_{ols}) = \mathbf{y}' \left(I - \mathbf{X}(\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'\right) \mathbf{y}.$$

- Due to the shrinkage effect of the g-prior, the SSR_g is now larger than $\text{SSR}(\hat{\boldsymbol{\beta}}_{ols})$ but their differences diminish as $g \to \infty$.

Unit-information prior

- Another common prior on $\boldsymbol{\beta}$ that are both invariate w.r.t. linear transforms of $\mathbf{X}$ and are conjugate is the so-called *unit-information* prior.
- It takes the form

$$\boldsymbol{\beta} \sim \mathrm{N}(\mathbf{0}, n\sigma^2(\mathbf{X}'\mathbf{X})^{-1}).$$

- It is essentially the g-prior (as if $g = n$) but in standard g-prior, the parameter g is fixed, not changing with sample size.
- It is a weak prior that essentially quantifies the prior knowledge as that from a single prior observation.

The marginal likelihood of a model

- So far in applying Bayesian theorem, we have not used the normalizing constant $p(\mathbf{x})$ in the denominator

$$p(\boldsymbol{\theta}|\mathbf{x}) = \frac{p(\mathbf{x}|\boldsymbol{\theta})p(\boldsymbol{\theta})}{p(\mathbf{x})}.$$

- This quantity $p(\mathbf{x})$ is called the *marginal likelihood*. It is given by

$p(x|M_1), p(x|M_2), \ldots$

$$p(\mathbf{x}) = \int p(\mathbf{x}|\boldsymbol{\theta})p(\boldsymbol{\theta})d\boldsymbol{\theta} = \frac{p(x|\theta)\cdot p(\theta)}{p(\theta|x)}$$

  which involves an integral hard to evaluate when $\boldsymbol{\theta}$ is multivariate.
- It is the "marginal' ' likleihood as it is the likelihood averaged over the prior on the parameters.
- In some settings, such as conjugate families, this integral can actually be computed in closed form.
- For example, if we are able to derive exactly what $p(\boldsymbol{\theta}|\mathbf{x})$ is, then

$$p(\mathbf{x}) = \frac{P(\mathbf{x}|\boldsymbol{\theta})p(\boldsymbol{\theta})}{p(\boldsymbol{\theta}|\mathbf{x})}.$$

- The marginal likelihood is a measure for model fit.

Bayesian hypothesis testing and model selection

- Consider two competive models $\mathscr{M}_1$ and $\mathscr{M}_2$.
  - Model 1 involes parameters $\boldsymbol{\theta}_1$, sampling model $p(\mathbf{x}|\boldsymbol{\theta}_1)$ and prior $p(\boldsymbol{\theta}_1)$.
  - Model 2 involes parameters $\boldsymbol{\theta}_2$, sampling model $p(\mathbf{x}|\boldsymbol{\theta}_2)$ and prior $p(\boldsymbol{\theta}_2)$.
- How do we evaluate the fit of each model to the data?
- If we place prior probability on these two models $p(\mathscr{M}_1)$ and $p(\mathscr{M}_2)$. What is the posterior probability of these two models?

$$p(x|M_1) = \int p(x|\theta, M_1)\, p(\theta|M_1)\, d\theta$$
$$p(x|M_2) = \int p(x|\theta, M_2) \cdot p(\theta|M_2)\, d\theta$$

More general model comparison

- We can place prior probability on a collection of models.
- By Bayes theorem, the posterior probability for a model $\mathscr{M}$—speficied by sampling model $p(\mathbf{x}|\boldsymbol{\theta},\mathscr{M})$ and prior $p(\boldsymbol{\theta})$—is

$$p(\mathscr{M}|\mathbf{x}) \propto p(\mathbf{x}|\mathscr{M})p(\mathscr{M}).$$

  where

$$p(\mathbf{x}|\mathscr{M}) = \int p(\mathbf{x}|\boldsymbol{\theta},\mathscr{M})p(\boldsymbol{\theta}|\mathscr{M})d\boldsymbol{\theta}$$

  is exactly the marginal likelihood under model $\mathscr{M}$!
- In particular, for comparing a pair of models $\mathscr{M}_1$ and $\mathscr{M}_2$,

$$\underbrace{\frac{p(\mathscr{M}_1|\mathbf{x})}{p(\mathscr{M}_2|\mathbf{x})}}_{\text{the posterior odds}} = \underbrace{\frac{p(\mathbf{x}|\mathscr{M}_1)}{p(\mathbf{x}|\mathscr{M}_2)}}_{\text{the Bayes factor}} \cdot \underbrace{\frac{p(\mathscr{M}_1)}{p(\mathscr{M}_2)}}_{\text{the prior odds}}.$$

BF: characterizing the evidence from data.

- The *Bayes factor* (BF) for comparing a pair of models is simply the ratio of the marginal likelihoods.

When the number of models is massive

- Note that the posterior probability of the models

$$p(\mathscr{M}|\mathbf{x}) = \frac{p(\mathbf{x}|\mathscr{M})p(\mathscr{M})}{\sum_{\mathscr{M}} p(\mathbf{x}|\mathscr{M})p(\mathscr{M})}$$

  for which the normalizing constant requires enumeration over all models under consideration.
- When the collection of models being considered is massive, this normalizing constant cannot be computed exactly.
- Computational strategy
  - MCMC: design a Markov chain on the space of models with the desired stationary distribution.
  - Alternative stragegies: e.g., sequential importance sampling (SIS).

Bayesian model averaging for prediction

- When there are multiple models of consideration, which one do we use for prediction?
- A frequenstist will do this in two steps:
  - Select a *single* model based on some model selection procedure.
  - Make a prediction based on the model chosen.
- But, this ignores the uncertainty in the model selection step and places too much importance on a single model which might not be the best model for some observations.
- Bayesian model averaging (BMA):
  - Let us compute the average prediction with respect to the posterior distribution on the competitive models

$$\mathrm{E}(x_{n+1}|\mathbf{x}_n) = \sum_{\mathscr{M}} \mathrm{E}(x_{n+1}|\mathbf{x}_n, \mathscr{M}) p(\mathscr{M}|\mathbf{x}_n).$$

$$\iint \mathbb{E}(x_{n+1} \mid x_n, \theta, m) \cdot p(\theta \mid m, x_n) \cdot p(m \mid x_n) \, d\theta \, dm.$$

Carrying out BMA

* When inner integral is not analytically:
Draw $(\theta^{(t)}, \mathcal{M}^{(t)}) \sim p(\theta^{(t)}, \mathcal{M}^{(t)} | x_n)$
Then $\frac{1}{S}\sum_{t=1}^{S} \mathbb{E}(x_{n+1} | x_n, \theta^{(t)}, \mathcal{M}^{(t)}) \to \mathbb{E}(x_{n+1} | x_n)$

- To carry out such average prediction, we need to
  - either be able to compute $p(\mathscr{M}|\mathbf{x}_n)$, in which case, we must know the normalizing constant. (Why?)
  - or be able to carry out Monte Carlo by sampling from the posterior model distribution $p(\mathscr{M}|\mathbf{x}_n)$. Suppose we can draw a collection of models

$$\mathscr{M}^{(1)}, \mathscr{M}^{(2)}, \ldots, \mathscr{M}^{(S)}$$

from $p(\mathscr{M}|\mathbf{x}_n)$ independently or by other strategies such as MCMC. Then

$$\frac{1}{S}\sum_i \mathrm{E}(x_{n+1}|\mathbf{x}_n, \mathscr{M}^{(i)}) \to \mathrm{E}(x_{n+1}|\mathbf{x}_n)$$

by law of large number.

Application in the Bayesian linear model with g-priors

- Let's get back to the Bayesian linear regression model with g-priors.
- Due to the conjugacy of the g-prior, we are actually able to evaluate the marginal likelihood of the given model. (Details are in the textbook pp.164-165.)
- Specifically,

$$
\begin{aligned}
p(\mathbf{y}) &= \int p(\mathbf{y}|\boldsymbol{\beta},\sigma)p(\boldsymbol{\beta}|\sigma)p(\sigma)d\boldsymbol{\beta}d\sigma \\
&= \int p(\mathbf{y}|\sigma)p(\sigma)d\sigma \\
&= \pi^{-n/2}\frac{\Gamma(\nu_n/2)}{\Gamma(\nu_0/2)}(1+g)^{-\frac{p}{2}}\frac{(\nu_0\sigma_0^2)^{\nu_0/2}}{(\nu_n\sigma_n^2)^{\nu_n/2}}
\end{aligned}
$$

where there are a total of p covariates, that is, when $\mathbf{X}$ is $n \times (p+1)$.

The model space

- Typical model selection for linear models involves the selection of covariates to be included in the model.
- Suppose there are a total of p covariates.
- Each linear model can be represented by a sequence of indicators such as $\boldsymbol{z} = (z_1, \ldots, z_p) \in \{0,1\}^p$ such that $z_i = 1$ if and only if the ith covariate is in the model.
- The covariate matrix $\mathbf{X}_{\boldsymbol{z}}$ now depends on the model $\boldsymbol{z}$ and so do the set of coefficients $\boldsymbol{\beta}_{\boldsymbol{z}}$.

Bayesian model selection and model averaging

- Suppose we adopt the g-prior for $\boldsymbol{\beta}_{\boldsymbol{z}}$ under the model $\boldsymbol{z}$ and the IG prior for σ^2 as before.
- Then we can compare the fit of two models $\boldsymbol{z}_1$ and $\boldsymbol{z}_2$ by the Bayes factor

$$\mathrm{BF}_{\boldsymbol{z}_1,\boldsymbol{z}_2} = \frac{p(\mathbf{y}|\boldsymbol{z}_1)}{p(\mathbf{y}|\boldsymbol{z}_2)}.$$

- Moreover, if we place a prior on the space of $\boldsymbol{z}$, then by Bayes theorem the model posterior is

$$p(\boldsymbol{z}|\mathbf{y}) \propto p(\mathbf{y}|\boldsymbol{z})p(\boldsymbol{z}).$$

- When there are a small number of covariates, e.g., ≤ 20, we can enumerate the space of all models to compute the normalizing constant and get this posterior probaiblity exactly.
- This will allow us to carry out BMA exactly.

When the model space is large

- When there are more than 25 covariates, the model space gets so large that enumeration of the model space becomes comptuationally infeasible.
- How to sample from $p(\boldsymbol{z}|\mathbf{y})$?
- By Gibbs sampling!
- Let's consider $\boldsymbol{z} = (z_1, \ldots, z_p)$ as a p-dimensional parameter.
- Gibbs sampling corresponds to updating each of the p indicator from their corresponding full conditional.
- So we just need to figure out the full conditional for each z_i

$$p(z_i|\boldsymbol{z}_{-i}, \mathbf{y})$$

where $\boldsymbol{z}_{-i}$ represents the model index vector excluding its ith element.

- But because

$$p(z_i|\boldsymbol{z}_{-i},\mathbf{y}) \propto p(\mathbf{y}|\boldsymbol{z}_{-i},z_i)p(z_i|\boldsymbol{z}_{-i})p(\boldsymbol{z}_{-i}),$$

we have the full conditional odds for z_i

$$\begin{aligned} o_i = \frac{p(z_i=1|\boldsymbol{z}_{-i},\mathbf{y})}{p(z_i=0|\boldsymbol{z}_{-i},\mathbf{y})} &= \frac{p(\mathbf{y}|\boldsymbol{z}_{-i},z_i=1)p(z_i=1|\boldsymbol{z}_{-i})p(\boldsymbol{z}_{-i})}{p(\mathbf{y}|\boldsymbol{z}_{-i},z_i=0)p(z_i=0|\boldsymbol{z}_{-i})p(\boldsymbol{z}_{-i})} \\ &= \frac{p(\mathbf{y}|\boldsymbol{z}_{-i},z_i=1)p(z_i=1|\boldsymbol{z}_{-i})}{p(\mathbf{y}|\boldsymbol{z}_{-i},z_i=0)p(z_i=0|\boldsymbol{z}_{-i})} \\ &= \underbrace{\frac{p(\mathbf{y}|\boldsymbol{z}_{-i},z_i=1)}{p(\mathbf{y}|\boldsymbol{z}_{-i},z_i=0)}}_{\text{Bayes factor}} \cdot \underbrace{\frac{p(z_i=1|\boldsymbol{z}_{-i})}{p(z_i=0|\boldsymbol{z}_{-i})}}_{\text{Prior conditional odds}} \end{aligned}$$

- Thus

$$z_i|\boldsymbol{z}_{-i},\mathbf{y} \sim \text{Bernoulli}\left(\frac{o_i}{1+o_i}\right).$$

- So we just need to iteratively sample each z_i from the above full conditional to do Gibbs sampling.

Posterior inclusion probability

- An application of BMA is to estimate the probability for $z_i = 1$ that is for the ith covariate to be selected in the model.
- Note that

posterior marginal inclusion probability.

$$\mathrm{P}(X_i \text{ is included} \,|\, \mathbf{y}) = P(z_i = 1 | \mathbf{y}) = \mathrm{E}(z_i | \mathbf{y}).$$

- This corresponds to applying BMA to "predict' ' the value of z_i.
- Based on an MCMC sample from the Gibbs sampler, after discarding burn-ins

$$\boldsymbol{z}^{(1)}, \boldsymbol{z}^{(2)}, \ldots, \boldsymbol{z}^{(T)},$$

we can compute the BMA estimate

$$\frac{1}{S} \sum_t z_i^{(t)} \to \mathrm{P}\left(z_i = 1 \,|\, \mathbf{y}\right).$$

BMA for new observations

- To predict the value of a new set of observations $\mathbf{y}_{new}$ given covariate matrix $\mathbf{X}_{new}$ (again given), we again can use BMA.

$$\mathrm{E}(\mathbf{y}_{new}|\mathbf{y}_n) = \sum_{\boldsymbol{z}} \mathrm{E}(\mathbf{y}_{new}|\mathbf{y}_n, \boldsymbol{z}) p(\boldsymbol{z}|\mathbf{y}_n) \approx \frac{1}{S} \sum_{t} \mathrm{E}(\mathbf{y}_{new}|\mathbf{y}_n, \boldsymbol{z}^{(t)}).$$

- Now for each sampled model $\boldsymbol{z}^{(t)}$, the predictive value

$$\mathrm{E}(\mathbf{y}_{new}|\mathbf{y}_n, \boldsymbol{z}^{(t)}) = \int \mathrm{E}(\mathbf{y}_{new}|\mathbf{y}_n, \boldsymbol{\beta}, \sigma^2, \boldsymbol{z}^{(t)}) p(\boldsymbol{\beta}, \sigma^2|\boldsymbol{z}^{(t)}). \; d\beta \, d\sigma^2$$

(evaluate both inner integral & outter integral).

BMA for new observations

- One can attempt to evaluate this in closed form. But we can also evaluate this by standard Monte Carlo (not MCMC). Because we have a fully conjugate posterior for $(\boldsymbol{\beta}, \sigma^2)$ given the model, we can draw R independent samples from $p(\boldsymbol{\beta}, \sigma^2 | \boldsymbol{z}^{(t)})$

$$(\boldsymbol{\beta}^{(t,1)}, \sigma^{2\,(t,1)}), (\boldsymbol{\beta}^{(t,2)}, \sigma^{2\,(t,2)}), \ldots, (\boldsymbol{\beta}^{(t,R)}, \sigma^{2\,(t,R)})$$

$$\mathrm{E}(\mathbf{y}_{new} | \mathbf{y}_n, \boldsymbol{z}^{(t)}) \approx \frac{1}{R} \sum_{r=1}^{R} \mathbf{X}_{new} \boldsymbol{\beta}^{(t,r)}.$$

- The textbook proposes to draw only one sample (i.e., $R = 1$) from the posterior of $(\boldsymbol{\beta}, \sigma^2)$ so the index r is no longer needed.
- This is fine but we can also draw more samples.

Example: Diabetes data set

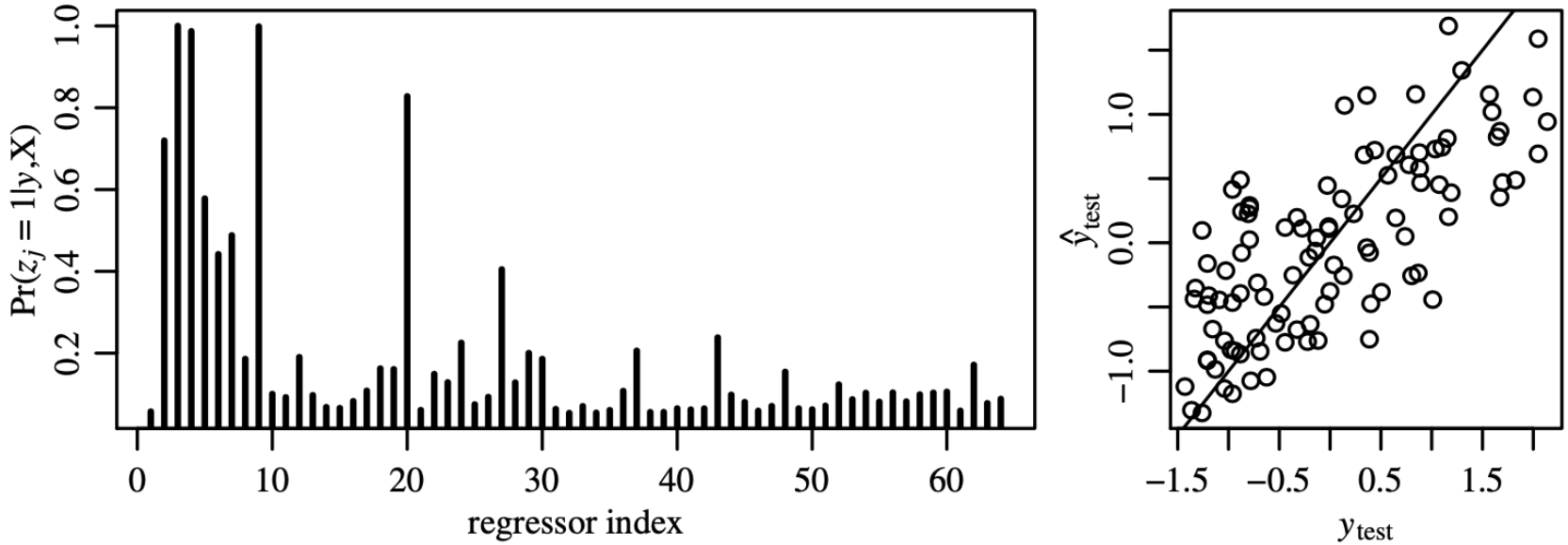


Fig. 9.7. The first panel shows posterior probabilities that each coefficient is non-zero. The second panel shows y_{test} versus predictions based on the model averaged estimate of $\boldsymbol{\beta}$.