

# Module 11: Linear Regression, the g-prior, and model selection

Rebecca C. Steorts

Hoff, Chapter 9

# Agenda

- ▶ Review of the Multivariate setup
- ▶ The g-prior
- ▶ Application to diabetes (Hoff, 9.2)
- ▶ Bayesian model selection (g-prior)
- ▶ Bayesian model averaging

# Notation

- ▶  $X_{n \times p}$ : regression features or covariates (design matrix)
- ▶  $\mathbf{x}_{p \times 1}$ :  $i$ th row vector of the regression covariates
- ▶  $\mathbf{y}_{n \times 1}$ : response variable (vector)
- ▶  $\beta_{p \times 1}$ : vector of regression coefficients

## Multivariate Setup

Let's assume that we have data points  $(x_i, y_i)$  available for all  $i = 1, \dots, n$ .

- $\mathbf{y}$  is the response variable

$$\mathbf{y} = \begin{pmatrix} y_1 \\ y_2 \\ \vdots \\ y_n \end{pmatrix}_{n \times 1}$$

- $\mathbf{x}_i$  is the  $i$ th row of the design matrix  $\mathbf{X}_{n \times p}$ .

Consider the regression coefficients

$$\boldsymbol{\beta} = \begin{pmatrix} \beta_1 \\ \beta_2 \\ \vdots \\ \beta_p \end{pmatrix}_{p \times 1}$$

## Multivariate Setup

$$\mathbf{y} \mid \mathbf{X}, \beta, \sigma^2 \sim MVN(\mathbf{X}\beta, \sigma^2 \mathbf{I})$$

$$\beta \sim MVN(\beta_0, \Sigma_o)$$

Recall the posterior can be shown to be

$$\beta \mid \mathbf{y}, \mathbf{X}, \sigma^2 \sim MVN(\beta_n, \Sigma_n)$$

where

$$\beta_n = E[\beta \mid \mathbf{y}, \mathbf{X}, \sigma^2] = (\Sigma_o^{-1} + (\mathbf{X}^T \mathbf{X})^{-1} / \sigma^2)^{-1} (\Sigma_o^{-1} \beta_0 + \mathbf{X}^T \mathbf{y} / \sigma^2)$$

$$\Sigma_n = \text{Var}[\beta \mid \mathbf{y}, \mathbf{X}, \sigma^2] = (\Sigma_o^{-1} + (\mathbf{X}^T \mathbf{X})^{-1} / \sigma^2)^{-1}$$

How do we specify **a prior on**  $\beta_0$  and  $\Sigma_o$ ?

# The g-prior

To do the *least amount of calculus*, we can put a *g-prior* on  $\beta$ .

The g-prior on  $\beta$  has the following form:

$$\beta \mid \mathbf{X}, \sigma^2 \sim MVN(0, g \sigma^2 (\mathbf{X}^T \mathbf{X})^{-1}),$$

where  $g$  is a constant, such as  $g = n$ .

The prior also happens to be invariant and is widely studied for regression problems (Zellner, 1986).

We will find that

1.  $g$  shrinks the coefficients and can prevent overfitting to the data
2. if  $g = n$ , then as  $n$  increases, inference approximates that using  $\hat{\beta}_{ols}$

## The g-prior

Under the g-prior, it follows that

$$\beta_n = E[\beta \mid \mathbf{y}, \mathbf{X}, \sigma^2] \quad (1)$$

$$= \left( \frac{\mathbf{X}^T \mathbf{X}}{g\sigma^2} + \frac{\mathbf{X}^T \mathbf{X}}{\sigma^2} \right)^{-1} \frac{\mathbf{X}^T \mathbf{y}}{\sigma^2} \quad (2)$$

$$= \frac{g}{g+1} (\mathbf{X}^T \mathbf{X})^{-1} \mathbf{X}^T \mathbf{y} = \frac{g}{g+1} \hat{\beta}_{ols} \quad (3)$$

$$\Sigma_n = \text{Var}[\beta \mid \mathbf{y}, \mathbf{X}, \sigma^2] \quad (4)$$

$$= \left( \frac{\mathbf{X}^T \mathbf{X}}{g\sigma^2} + \frac{\mathbf{X}^T \mathbf{X}}{\sigma^2} \right)^{-1} = \frac{g}{g+1} \sigma^2 (\mathbf{X}^T \mathbf{X})^{-1} \quad (5)$$

$$= \frac{g}{g+1} \text{Var}[\hat{\beta}_{ols}] \quad (6)$$

## Variance component $\sigma^2$

What about a prior on  $1/\sigma^2 = \lambda$

$$y \mid X, \beta, \sigma^2 \sim \text{MVN}(X\beta, \sigma^2 I) \quad (7)$$

$$1/\sigma^2 = \lambda \sim \text{Gamma}(\nu_0/2, \nu_0\sigma_0^2/2) \quad (8)$$

Then the posterior can be shown to be

$$p(\sigma^2 \mid y, X) \sim \text{InverseGamma}([\nu_0 + n]/2, [\nu_0\sigma_0^2 + SSR_g]/2).$$

1

where  $SSR_g$  is somewhat complicated (see Hoff for details, p. 158).

---

<sup>1</sup>This is left as an exercise to be done on your own.



## Variance component $\sigma^2$

The joint distribution can be written as

$$p(\sigma^2, \beta \mid y, X) = p(\sigma^2 \mid y, X) \times p(\beta \mid y, X, \sigma^2)$$

Goal: simulate  $(\sigma^2, \beta) \sim p(\sigma^2, \beta \mid y, X)$

Starting value  $(\beta_0, \sigma_0^2)$

1. Simulate

$$\sigma^2 \sim p(\sigma^2 \mid y, X)$$

Give us  $(\sigma_1^2, \beta_0)$

2. Use this updated value of  $\sigma_1^2$  to simulate

$$\beta \sim p(\beta \mid y, X, \sigma_1^2)$$

Gives us  $(\sigma_1^2, \beta_1)^2$

Run the sampler for  $S$  iterations.

<sup>2</sup>Here,  $\beta_1$  should not be confused as the first component of the vector of  $\beta$  as here we are taking a draw from the Gibbs sampler.

## Application to diabetes (Exercise 9.2, part a)

Suppose we have data on health-related variables of a population of 532 women.

Our goal is to model the conditional distribution of glucose level (glu) as a linear combination of the other variables, excluding the variable diabetes.<sup>3</sup>

---

<sup>3</sup>See Exercise 7.6 for the data description.

## Model specification

$$y \mid X, \beta, \sigma^2 \sim MVN(X\beta, \lambda^{-1}I) \quad (9)$$

$$\beta \mid \lambda \sim MVN(0, g\lambda^{-1}(X^T X)^{-1}) \quad (10)$$

$$\lambda \sim \text{Gamma}(\nu_0/2, \nu_0\sigma_o^2/2) \quad (11)$$

## Regression model on the g-prior

Fit a regression model using the g-prior with  $g = n$ ,  $\nu_0 = 2$  and  $\sigma_0^2 = 1$ . Obtain posterior confidence intervals for all of the parameters.

## Regression model on the g-prior

Section 9.2.2 (Hoff) shows that under the  $g$  prior,  $p(\sigma^2 \mid \mathbf{y}, \mathbf{X})$  and  $p(\boldsymbol{\beta} \mid \mathbf{y}, \mathbf{X}, \sigma^2)$  are inverse gamma and multivariate normal distributions respectively.

## Regression model on the g-prior

Therefore samples from the joint posterior  $p(\sigma^2, \beta \mid \mathbf{y}, \mathbf{X}, \sigma^2)$  can be made with a Monte Carlo approximation.

We first center and scale all the variables so that there is no need to include an intercept in the model.

## Regression model on the g-prior

```
#library(knitr)  
#rm(list=ls())  
azd_data = read.table("azdiabetes.dat", header = TRUE)  
head(azd_data)
```

##	npreg	glu	bp	skin	bmi	ped	age	diabetes
## 1	5	86	68	28	30.2	0.364	24	No
## 2	7	195	70	33	25.1	0.163	55	Yes
## 3	5	77	82	41	35.8	0.156	35	No
## 4	0	165	76	43	47.9	0.259	26	No
## 5	0	107	60	25	26.4	0.133	23	No
## 6	5	97	76	27	35.6	0.378	52	Yes

## Regression model on the g-prior

```
y = azd_data$glu  
# remove glu and diabetes  
X = as.matrix(azd_data[,c(-2,-8)])  
head(X)
```

##		npreg	bp	skin	bmi	ped	age
##	[1,]	5	68	28	30.2	0.364	24
##	[2,]	7	70	33	25.1	0.163	55
##	[3,]	5	82	41	35.8	0.156	35
##	[4,]	0	76	43	47.9	0.259	26
##	[5,]	0	60	25	26.4	0.133	23
##	[6,]	5	76	27	35.6	0.378	52



# Standardization

```
# standardize data to have mean 0 and variance 1  
ys = scale(y)  
Xs = scale(X)  
n = dim(Xs)[1]  
p = dim(Xs)[2]
```

# Hyper-parameters

```
g = n  
nu0 = 2  
s20 = 1
```

## Intermediate Matrices

```
# intermediate matrices  
Hg = (g/(g+1)) * Xs %*% solve(t(Xs) %*% Xs) %*% t(Xs)  
SSRg = t(ys) %*% ( diag(1,nrow=n) - Hg ) %*% ys
```

# Monte carlo

```
# number of posterior samples
```

```
S = 1000
```

```
# generate posteriors
```

```
# we know that the sigma2 is
```

```
# an updated inverse gamma
```

```
s2 = 1/rgamma(S, (nu0+n)/2, (nu0*s20 + SSRg)/2)
```

```
head(s2)
```

```
## [1] 0.8757964 0.8670699 0.8333923 0.8522915 0.8581030 0.
```

## Monte carlo

```
# updated posterior mean and variance  
# from using the g-prior (see slide 7)  
Vb = g*solve(t(Xs) %*% Xs)/(g+1)  
Eb = Vb %*% t(Xs) %*% ys  
E = matrix(rnorm(S*p, 0, sqrt(s2)),S,p)  
# use cholesky factorization  
beta_s = t( t(E %*% chol(Vb)) + c(Eb))  
  
# transform coefficients to the original scale  
sd_X = apply(X,2,sd)  
Beta_a = sweep(beta_s,2,sd_X,FUN = "/")
```

# The 95% posterior confidence intervals

```
# 95% credible interval
```

```
(Beta_CIa = apply(Beta_a, 2, quantile, c(0.025, 0.975)))
```

```
##           npreg           bp           skin           bmi           ped
## 2.5% -0.05269071 -0.0002984302 -0.002941644 0.004704214 0.1080635
## 97.5% 0.01234342 0.0144919397 0.016274524 0.036106047 0.5562314
##           age
## 2.5% 0.01467044
## 97.5% 0.03460817
```

```
#kable(data.frame(Beta_CIa))
```

# Model selection

- ▶ Often we have a large number of covariates.
- ▶ Using all of them induces poor statistical performance.
- ▶ How can we reduce the covariates and have good inference and prediction?
- ▶ Common method: Backwards and stepwise regression (slow).

## Model selection

Suppose that we believe some of the regression coefficients are 0.

Come up with a prior distribution that reflects the probability of this occurring.

Consider

$$y_i = z_1 b_1 x_{i,1} + \dots z_p b_p x_{i,p},$$

where  $b_p$  is a real number and  $z_j$  indicate which regression coefficients are nonzero.

Note:  $\beta_j = b_j \times z_j$ .



# Bayesian model selection

Bayesian model selection works by obtaining a posterior distribution for  $\mathbf{z}$ .

Assume a prior  $p(\mathbf{z})$ .

Then

$$p(\mathbf{z} \mid \mathbf{Y}, \mathbf{X}) = \frac{p(\mathbf{z})p(\mathbf{Y} \mid \mathbf{X}, \mathbf{z})}{\sum_{\mathbf{z}} p(\mathbf{z})p(\mathbf{Y} \mid \mathbf{X}, \mathbf{z})}$$

## Bayesian model selection

Suppose we want to compare two models  $z_a$  and  $z_b$ . Consider

$$\text{odds}(z_a, z_b \mid \mathbf{Y}, \mathbf{X}) = \frac{p(z_a \mid \mathbf{Y}, \mathbf{X})}{p(z_b \mid \mathbf{Y}, \mathbf{X})} = \frac{p(z_a)}{p(z_b)} \times \frac{p(\mathbf{Y} \mid \mathbf{X}, z_a)}{p(\mathbf{Y} \mid \mathbf{X}, z_b)}$$

This is posterior odds = prior odds  $\times$  “Bayes factor”

“Bayes factor”: how much do the data favor model  $z_a$  over model  $z_b$

To obtain a posterior distribution over models, we must compute  $p(\mathbf{Y} \mid \mathbf{X}, z)$  for *each* model under consideration.

# Bayesian model selection

We must compute

$$p(y \mid X, \mathbf{z}) = \int \int p(y, \beta, \sigma^2, \mid X, \mathbf{z}) d\beta d\sigma^2 \quad (12)$$

$$\propto \int \int p(y \mid X, \mathbf{z}) p(\beta \mid X, \mathbf{z}) p(\sigma^2) d\beta d\sigma^2. \quad (13)$$

To do the *least amount of calculus*, we can put a *g-prior* on  $\beta$

$$\beta \mid X, \mathbf{z} \sim \text{MVN}(0, g \sigma^2 (X_z^T X_z)^{-1}).$$

## Back to the g-prior

Given the g-prior

$$\beta \mid X, \mathbf{z} \sim MVN(0, g \sigma^2 (X_z^T X_z)^{-1}),$$

$p(y \mid X, \mathbf{z})$  can be worked out in closed form (details p. 165).

Go through the details on your own.

## Back to the g-prior

This results in being able to compute

$$\frac{p(y \mid X, \mathbf{z}_a)}{p(y \mid X, \mathbf{z}_b)} = (1 + n)^{(p_{z_b} - p_{z_a})/2} \times \left( \frac{s_{z_a}^2}{s_{z_b}^2} \right)^{1/2} \quad (14)$$

$$\times \left( \frac{s_{z_b}^2 + SSR_g^{z_b}}{s_{z_a}^2 + SSR_g^{z_a}} \right)^{(n+1)/2} \quad (15)$$

We have a ratio of the marginal probabilities, giving us a balance between model complexity and model fit.

Suppose  $p_{z_b}$  is large compared to  $p_{z_a}$ .

This causes a penalization of model  $z_b$

Note that a large value of  $SSR_g^{z_a}$  compared with  $SSR_g^{z_b}$  will penalize model  $z_a$ .

# Bayesian Model Averaging

Suppose that we have an estimate of  $\beta$  from which we can make predictions.

We may also want a list of relatively high probability models. We can use a Markov chain to search through the space of models for values of  $z$  with high posterior probability.

## Bayesian model averaging

Suppose  $p$  is large. Then  $2^p$  models to consider.

Instead let's use a Gibbs sampler to search through the space of models for values where  $\mathbf{z}$  has a high posterior probability.

Generate a new value of  $\mathbf{z}$  via

$$p(z_j \mid \mathbf{Y}, \mathbf{X}, \mathbf{z}_{-j}).$$

The full conditional that  $z_j = 1$  can be written as  $o_j / (o_j + 1)$ .

$$o_j = \frac{p(z_j = 1 \mid \mathbf{Y}, \mathbf{X}, \mathbf{z}_{-j})}{p(z_j = 0 \mid \mathbf{Y}, \mathbf{X}, \mathbf{z}_{-j})} \quad (16)$$

$$= \frac{p(z_j = 1)p(\mathbf{Y} \mid \mathbf{X}, \mathbf{z}_{-j}, z_j = 1)}{p(z_j = 0)p(\mathbf{Y} \mid \mathbf{X}, \mathbf{z}_{-j}, z_j = 0)} \quad (17)$$

# Bayesian model averaging

Note: we may also want to obtain posterior samples of  $\beta$  and  $\sigma^2$ .

Using the conditional distributions from Section 9.2, we can sample from these directly.

The Gibbs sampling scheme requires using Section 9.2 and 9.3 (covered in lab).



# Bayesian model averaging

$$\begin{array}{ccccc} \mathbf{z}^{(s)} & \longrightarrow & \sigma^{2(s)} & \longrightarrow & \boldsymbol{\beta}^{(s)} \\ \downarrow & & & & \\ \mathbf{z}^{(s+1)} & \longrightarrow & \sigma^{2(s+1)} & \longrightarrow & \boldsymbol{\beta}^{(s+1)} \end{array}$$

Figure 1: Start with  $\mathbf{z}^{(s)}$ . Then in random order update  $z_j$  from its full conditional.

# Bayesian model averaging

Generate

$$\{\mathbf{z}^{(s+1)}, \sigma^{2(s+1)}, \boldsymbol{\beta}^{(s+1)}\} :$$

1. Set  $\mathbf{z} = \mathbf{z}^{(s)}$
2. For  $j \in \{1, \dots, p\}$  in random order, replace  $z_j$  with a sample from

$$p(z_j \mid \mathbf{z}_{-j}, \mathbf{Y}, \mathbf{X})$$

3. Set  $\mathbf{z}^{(s+1)} = \mathbf{z}$
4. Sample  $\sigma^{2(s+1)} \sim p(\sigma^2 \mid \mathbf{z}^{(s+1)}, \mathbf{Y}, \mathbf{X})$
5. Sample  $\boldsymbol{\beta}^{(s+1)} \sim p(\boldsymbol{\beta} \mid \mathbf{z}^{(s+1)}, \sigma^{2(s+1)}, \mathbf{Y}, \mathbf{X})$

## Back to diabetes data (Exercise 9.2, b)

Let's perform Bayesian model averaging (as described in Section 9.3)

Obtain  $P(\beta_j \neq 0 \mid y)$  as well as posterior confidence intervals for all of the parameters. Compare our results to that in part (a.)

## Back to diabetes data (Exercise 9.2, b)

The following function `lpy.X` calculates the log of  $p(\mathbf{y} \mid \mathbf{X})$ , which we will use in implementing the Gibbs sampler for Bayesian model averaging.

```
## a function to compute the marginal probability
lpy.X <- function(y, X, g=length(y), nu0=1,
                  s20=try(summary(lm(y~ -1+X))$sigma^2, silent=TRUE)){
  n = dim(X)[1]
  p = dim(X)[2]
  if (p==0) { Hg = 0; s20 = mean(y^2)}
  if (p>0){ Hg = (g/(g+1)) * X %>% solve(t(X) %>% X) %>% t(X) }
  SSRg = t(y) %>% ( diag(1, nrow=n) - Hg ) %>% y -
    .5*( n*log(pi) + p*log(1+g) + (nu0+n)*log(nu0*s20 + SSRg) -
      nu0*log(nu0*s20)) +
  lgamma( (nu0+n)/2 ) - lgamma(nu0/2)
}
```

## Back to diabetes data (Exercise 9.2, b)

Let  $\mathbf{z}$  be the random binary vector of variable indicators. Generating samples of  $p(\mathbf{z}, \sigma^2, \beta)$  from the joint posterior distribution is achieved with the following steps:

1. For  $j \in \{1, \dots, p\}$  in random order, draw  $z_j$  from  $p(z_j \mid \mathbf{z}_{-j}, \mathbf{y}, \mathbf{X})$ .
2. Sample  $\sigma^2 \sim p(\sigma^2 \mid \mathbf{z}, \mathbf{y}, \mathbf{X})$ .
3. Sample  $\beta \sim p(\beta \mid \mathbf{z}, \sigma^2, \mathbf{y}, \mathbf{X})$ .

## MCMC setup

```
g = n
nu0 = 1 # unit information prior
z = rep(1, p)
# picking a starting value for the marginal probability
lpy.c = lpy.X(ys, Xs[,z==1,drop=FALSE])
S = 10
Z = matrix(NA, S, p)
B = matrix(0, S, p)
```

# Gibbs sampler

```
## Gibbs sampler
for(s in 1:S){
  # if(s %% 100 ==0) {print(s)}
  # sample z
  for (j in sample(1:p)){
    zp = z
    zp[j] = 1 - zp[j]
    lpy.p = lpy.X(ys,Xs[, zp==1, drop=FALSE])
    r = (lpy.p - lpy.c) * (-1)^(zp[j]==0)
    zp[j] = rbinom(1, 1, 1/(1+exp(-r)))
    if(z[j] == zp[j]) {lpy.c = lpy.p}
  }
  Z[s,] = z

  # sample s2
  pm = sum(z==1) # number of nonzero variables in the model
  if (pm==0){
    Hg = 0
    s20 = mean(y^2)
  }
  if (pm>0){
    Hg = (g/(g+1)) * Xs[,z==1,drop=F] %*% solve(t(Xs[,z==1,drop=F]) %*%
      Xs[,z==1,drop=F]) %*% t(Xs[,z==1,drop=F])
    # estimated residual variance from OLS
    s20=summary(lm(ys ~ -1+Xs[,z==1,drop=F]))$sigma^2
  }

  SSRg = t(ys) %*% ( diag(1,nrow=n) - Hg ) %*% ys
  s2 = 1/rgamma(1, (nu0+n)/2, (nu0*s20 + SSRg)/2)
  # Sigma2[s] = s2

  # sample beta
  Vb = g * solve(t(Xs[,z==1,drop=F]) %*% Xs[,z==1,drop=F])/(g+1)
  Eb = Vb %*% t(Xs[,z==1,drop=F]) %*% ys

  E = rnorm(p, 0, sqrt(s2))
  beta.z = E %*% chol(Vb) + c(Eb)
```

# Results

The posterior probability  $\Pr(\beta_j \neq 0 \mid \mathbf{y})$  is listed below for each predictor. Clearly all predictors are highly relevant to the response.

```
pprob_Z = apply(Z,2,mean)
pprob_Z = data.frame(matrix(pprob_Z,nr=1,nc=p))
names(pprob_Z) = names(azd_data[c(-2,-8)])
row.names(pprob_Z) = 'posterior including probability'
```

```
pprob_Z
```

```
##                                npreg bp skin bmi ped age
## posterior including probability    1  1    1  1  1  1
```

```
#kable(pprob_Z)
```



## Results

The 95% posterior confidence intervals for all the parameters from Bayesian model averaging are listed below. The results are similar to those in part (a) because all the predictors are included in each iteration of Gibbs sampler all the time.

```
# transform coefficients to the original scale
#sd_X = apply(X,2,sd)
Beta_b = sweep(B,2,sd_X,FUN = "/")

# 95% credible interval
(Beta_CIb = apply(Beta_b, 2, quantile, c(0.025, 0.975)))
```

```
##           [,1]      [,2]      [,3]      [,4]      [,5]
## 2.5% -0.046783559 0.003124393 0.001945646 0.008172701 0.2445281
## 97.5% 0.008861632 0.012332815 0.012511617 0.027131571 0.4482633
##           [,6]
## 2.5% 0.01229628
## 97.5% 0.03405374
```

# Preparation for Final Exam

- ▶ April 29, 9 am – noon, Old Chem 116 PM
- ▶ Closed note, closed book
- ▶ No calculators

# Material

- ▶ The exam will be cumulative
- ▶ There will be an emphasis on what we covered after Exam II
- ▶ You must take the final exam to pass the course
- ▶ No make up exams will be given

# Material after Exam II

(Modules 8 – 11)

- ▶ Gibbs sampling
- ▶ Data augmentation (latent variable models)
- ▶ Data augmentation with an application to censored values
- ▶ Gaussian mixture models (with latent variables)
- ▶ Entity resolution lecture (more latent variable modeling)
- ▶ Multivariate distributions and multivariate modeling
- ▶ Missing data
- ▶ Linear regression
- ▶ The g-prior
- ▶ Model selection

You have many practice exercises posted on this material and the cumulative material with solutions.

## Schedule Before Final Exam

- ▶ Thursday, April 16th: Special lecture by Neil Marchant on how latent variables models are used for entity resolution.
- ▶ Tuesday, April 23: Prof. Steorts will hold class at the usual time in case there are questions that students wish to go over with me relating to the final exam material.
- ▶ Wednesday, April 24: Labs will be an extra OH.
- ▶ All regular OH by TAs and Professor Steorts will be held through Sunday April 28th.
- ▶ If you're looking for an old exam, please see Prof. Steorts after class this Thursday or next week in OH.
- ▶ Prof Steorts will be happy to give anyone their current class ranking after homeworks are calculated next week in OH or near the end of class.

# Resources for the Final Exam

- ▶ There are many practice problems I recommend that are lab exercises.  
<https://github.com/resteorts/modern-bayes/tree/master/labs>
- ▶ There are many practice exercises that I have posted here to help prepare you for the final exam. <https://github.com/resteorts/modern-bayes/tree/master/exercises>