Class Notes for Biostat 250C

Scribe: Elvis Cui, Instructor: Sudipto Banerjee

May 3, 2021

Contents

Preface

This lecture notes is based on the Biostat 250C course I took in 2021, spring.

Note that x and $\mathbf{x}$ can both represent a vector and X and $\mathbf{X}$ can both represent a matrix. Also, I switch between y and Y from time to time. Both $\mathcal{N}$ and $\mathcal{N}_p$ can represent a multivariate distribution. I hope they do not cause too much confusion and I am too lazy to modify them so I apologize.

Important homework rule: Any HW given in a week is due on or before the beginning of class of the Wednesday of the following week.

Finally, a quote from CR Rao,

$$\text{Statistics} = \text{Approximation} + \text{Optimization}$$

and another quote from George Box,

All models are wrong, some are useful.

1 Week 1

1.1 03-29

1.1.1 Conditioning

Let $X \sim F(\cdot)$ be a univariate random variable. Then $X_1, \cdots, X_n \sim_{\text{iid}} F(\cdot)$ is a sequence of realization from $F(\cdot)$. The question is, how to construct a **multivariate version** of $F(\cdot)$?

$$\mathbb{P}(X_1 \in A_1, \cdots, X_n \in A_n)$$

Elements of $X^T = (X_1, \cdots, X_n)$ are dependent! So we have to model dependency.
Recall $X \perp\!\!\!\perp Y$ if and only if

$$\mathbb{P}(X \in A, Y \in B) = \mathbb{P}(X \in A)\mathbb{P}(Y \in B)$$

for any Borel sets A and B. Using box notation, we write

$$[X, Y] = [Y|X][X] = [X][Y]$$

Where $[X|Y] = \mathbb{P}(X|Y)$ is the (regular) conditional probability of X given Y. Independence means

$$[X|Y] = [X] \text{ and } [Y|X] = [Y]$$

where

$$[X] = \int [X, Y], \ [Y] = \int [X, Y]$$

Thus,

$$[X|Y] = \frac{[X][Y|X]}{[Y]} \propto [X][Y|X] = [X, Y] \tag{1.1}$$

This is the famous **Bayes' theorem**. Suppose we are given $[X|Y]$ and $[Y|X]$, and suppose that $[X, Y]$ exists, i.e.,

$$\iint [X, Y] = 1$$

Is is possible to find $[X, Y]$ from $[X|Y]$ and $[Y|X]$?

$$\begin{aligned} \frac{[Y|X]}{[X|Y]} &= \frac{[Y]}{[X]} \\ \Rightarrow \int \frac{[Y|X]}{[X|Y]} &= \frac{1}{[X]} \\ \Rightarrow \qquad [X] &= \left(\int \frac{[Y|X]}{[X|Y]} \right)^{-1} \end{aligned}$$

Thus,

$$[X,Y] = [X][Y|X] = [Y|X]\left(\int \frac{[Y|X]}{[X|Y]}\right)^{-1} \tag{1.2}$$

What is the key?

$$[X] = \frac{[Y][X|Y]}{[Y|X]} \text{ or } [X] \propto \frac{[X|Y]}{[Y|X]}$$

So I can evaluate $[X]$ based on $[X|Y]$ and $[Y|X]$. We do NOT need integration explicitly.

Definition 1.1 (Conditional independence)**.** Let Z be a third random variable. Then X and Y are **conditionally independent** given Z iff

$$[X,Y|Z] = [X|Z][Y|Z]$$

and we write

$$X \perp\!\!\!\perp Y|Z$$

What about $[X|Y,Z]$?

$$[X|Y,Z] = [X|Z]$$

Because

$$\begin{aligned}[X,Y|Z] &= [Y|Z][X|Y,Z] \\ &=_{\text{def}} [Y|Z][X|Z]\end{aligned}$$

Lemma 1 (Decomposition property of conditional independence)**.** *Let W be a fourth random variable. Then*

$$X \perp\!\!\!\perp (Y,W)|Z \Rightarrow X \perp\!\!\!\perp Y|Z \wedge X \perp\!\!\!\perp W|Z \tag{1.3}$$

Proof.

$$\begin{aligned}[X,Y,W|Z] &= [X|Z][Y,W|X,Z] \\ &=_{\text{def}} [X|Z][Y,W|Z]\end{aligned}$$

Integrate or **marginalize** out W from both sides:

$$\begin{aligned}LHS &= \int [X,Y,W|Z] = [X,Y|Z] \\ RHS &= \int [X|Z][Y,W|Z] = [X|Z][Y|Z]\end{aligned}$$

Done! □

Lemma 2 (Sufficient and necessary condition, [1])**.**

$$(X \perp\!\!\!\perp W|Z,Y) \wedge (X \perp\!\!\!\perp Y|Z) \text{ iff } X \perp\!\!\!\perp (Y,W)|Z \tag{1.4}$$

For a proof, see homework.

Comment: If $[X,Y]$ exists, then it is defined uniquely by $[X|Y]$ and $[Y|X]$. But what if $[X,Y]$ does NOT exist?

$$p(x,y) = e^{-\frac{1}{2}(x-y)^2}$$

and we have

$$\iint p(x,y)dxdy = +\infty$$

But

$$[X|Y] \sim \mathcal{N}(Y,1),\ [Y|X] \sim \mathcal{N}(X,1)$$

1.2 03-31

1.2.1 More on conditioning

Thinking ahead: we saw how to recover $[X,Y]$ from $[X|Y]$ and $[Y|X]$. What if we have three variables X,Y and Z? Can we obtain $[X,Y,Z]$ from $[X|Y,Z],[Y|X,Z],[Z|X,Y]$? How do we think about it?

Think differently about the 2-variable case. Consider

$$x_0, y_0 \in \text{Supp}(x,y)$$
$$\text{s.t. } p(x_0,y_0) > 0$$

How to write $p(x,y)$ in terms of $p(x_0,y_0)$?

$$p(x,y) = \text{?}_{\text{something interesting}} \frac{p(y|x_0)}{p(y_0|x_0)} p(x_0,y_0) \tag{1.5}$$

Think about it!

1.2.2 Bayesian Conjugate Linear Regression

Recall $\beta \in \mathbb{R}^p$ and $y \in \mathbb{R}^n$,

$$y = X\beta + \epsilon$$
$$y_i = x_i^T \beta + \epsilon_i$$
$$\epsilon_i \sim_{\text{iid}} \mathcal{N}(0,\sigma^2)$$

Bayesian setting (BHM=Bayesian hierarchical model):

$$BHM \begin{cases} y_i|\beta,\sigma^2 \sim_{\text{ind}} \mathcal{N}(x_i^T\beta,\sigma^2) \\ \beta|\sigma^2 \sim \mathcal{N}(\mu_\beta,\sigma^2 V_\beta) \\ \sigma^2 \sim \text{IG}(a,b) \end{cases} \tag{1.6}$$

Note

$$[\beta,\sigma^2,y] = [\sigma^2][\beta|\sigma^2][y|\beta,\sigma^2] \tag{1.7}$$

What we seek:

$$[\text{ Unknown } | \text{ Known }]$$

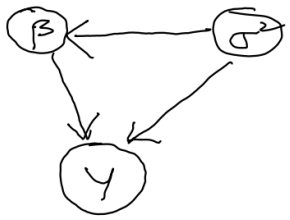
Assume μ_β, V_β, a, b are fixed and known.

$$\begin{aligned} \text{Unknown} &= \{\beta,\sigma^2\} \\ \text{Known} &= \{y\} \\ \text{Posterior distribution} &= [\beta,\sigma^2|y] = [\text{Unknown} \mid \text{Known}] \end{aligned}$$

Estimate of β : $\mathbb{E}(\beta|y)$, or median of $[\beta|y]$, or mode of $[\beta|y]$. Similarly for σ^2, $\mathbb{E}(\sigma^2|y)$ or median or mode of $[\sigma^2|y]$. It takes statisticians nearly 50 years to fully realize the power of this framework.

We use graphical model or directed acyclic graph (DAG), also known as Bayesian network, see figure 2.

Figure 1: DAG.



Your path that you are looking for: joint model for parameters and data.

$$\text{Joint posterior } \rightarrow_{\text{sampling}} \text{ Marginal posterior}$$

$$\begin{aligned} p(\theta_1,\cdots,\theta_p,y) &\rightarrow p(\theta_1,\cdots,\theta_p|y) \\ &\rightarrow p(\theta_i|y), i=1,\cdots p \end{aligned}$$

Embark on an mission to get $[\beta, \sigma^2|y]$.

$$[\beta, \sigma^2|y] \propto [\beta, \sigma^2][y|\beta, \sigma^2] = [\sigma^2][\beta|\sigma^2][y|\beta, \sigma^2]$$

and $[y]$ the marginal is absorbed into $\propto$.

$$p(\beta, \sigma^2|y) \propto \mathsf{IG}(\sigma^2|a, b) \times \mathcal{N}(\beta|\mu_\beta, \sigma^2 V_\beta) \times \mathcal{N}(y|X\beta, \sigma^2 V_y)$$

and V_y is known.

$$\mathsf{IG}(\sigma^2|a, b) \propto \left(\frac{1}{\sigma^2}\right)^{a+1} e^{-b/\sigma^2}$$

note that $\frac{1}{\sigma^2} \sim G(a, b)$ if $\sigma^2 \sim IG(a, b)$.

$$\begin{aligned}
\mathcal{N}_p(\beta|\mu_\beta, \sigma^2 V_\beta) &\propto \frac{1}{|\sigma^2 V_\beta|^{1/2}} \exp\left(-\frac{1}{2\sigma^2}(\beta - \mu_\beta)^T V_\beta^{-1}(\beta - \mu_\beta)\right) \\
&= (\sigma^2)^{-\frac{p}{2}} exp\left(-\frac{1}{2\sigma^2}(\beta - \mu_\beta)^T V_\beta^{-1}(\beta - \mu_\beta)\right) \\
\mathcal{N}_n(y|X\beta, \sigma^2 V_y) &\propto \frac{1}{|\sigma^2 V_y|^{1/2}} \exp\left(-\frac{1}{2\sigma^2}(y - X\beta)^T V_y^{-1}(y - X\beta)\right) \\
&= (\sigma^2)^{-\frac{n}{2}} \exp\left(-\frac{1}{2\sigma^2}(y - X\beta)^T V_y^{-1}(y - X\beta)\right)
\end{aligned}$$

note that $\det|\alpha A| = \alpha^m$ if A is a $m \times m$ matrix. Thus,

$$p(\beta, \sigma^2|y) \propto \left(\frac{1}{\sigma^2}\right)^{a+\frac{n}{2}+\frac{p}{2}+1} \exp\left(-\frac{1}{\sigma^2}\left[b + \frac{1}{2}Q(\beta)\right]\right) \tag{1.8}$$

where

$$Q(\beta) = (\beta - \mu_\beta)^T V_\beta^{-1}(\beta - \mu_\beta) + (y - X\beta)^T V_y^{-1}(y - X\beta)$$

Strategy: can we write

$$Q(\beta) = (\beta - m)^T M^{-1}(\beta - m) + \text{const.}$$

in this way?

Recall univariate completing the squares $ax^2 + bx + c = A(x - B)^2 + C$:

$$a \underbrace{(x^2 + \frac{b}{a}x)}_{(x+\frac{b}{2a})^2 - \frac{b^2}{4a^2}} + c = a(x + \frac{b}{2a})^2 + c - \frac{b^2}{4a}$$

Thus,

$$A = a,\ B = -\frac{b}{2a},\ C = c - \frac{b^2}{4a}$$

Alternative: write and compare coefficients:

$$ax^2 + bx + c = Ax^2 - 2ABx + B^2 + C$$

Thus,

$$A = a,\ 2AB = -b \ \Rightarrow\ B = -\frac{b}{2A} = -\frac{b}{2a}$$
$$B^2 + C = c \ \Rightarrow\ C = c - \frac{b^2}{4a}$$

Return to our case,

$$\begin{aligned}
Q(\beta) &= (\beta - \mu_\beta)^T V_\beta^{-1}(\beta - \mu_\beta) + (y - X\beta)^T V_y^{-1}(y - X\beta) \\
&= \beta^T V_\beta^{-1}\beta - 2\beta^T V_\beta^{-1}\mu_\beta + \mu_\beta^T V_\beta^{-1}\mu_\beta + \\
&\quad \beta^T X^T V_y^{-1} X\beta - 2\beta^T X^T V_y^{-1} y + y^T V_y^{-1} y \\
&= \beta^T \underbrace{\left(V_\beta^{-1} + X^T V_y^{-1} X\right)}_{M^{-1}} \beta - 2\beta^T \underbrace{\left(V_\beta^{-1}\mu_\beta + X^T V_y^{-1} y\right)}_{m} + \\
&\quad \underbrace{\mu_\beta^T V_\beta^{-1}\mu_\beta + y^T V_y^{-1} y}_{c} \\
&= \beta^T M^{-1}\beta - 2\beta^T m + c
\end{aligned}$$

Remarkably enough, watch this, Dr. Banerjee called this Mm-identity!

$$(\beta - Mm)^T M^{-1}(\beta - Mm) + c^* = \beta^T M^{-1}\beta - 2\beta^T m + \underbrace{m^T Mm + c^*}_{c}$$

Therefore,

$$\begin{aligned}
Q(\beta) &= \beta^T M^{-1}\beta - 2\beta^T m + c \\
&= (\beta - Mm)^T M^{-1}(\beta - Mm) + c - m^T Mm \\
&= (\beta - Mm)^T M^{-1}(\beta - Mm) + \\
&\quad \{\mu_\beta^T V_\beta^{-1}\mu_\beta + y^T V_y^{-1} y - m^T Mm\}
\end{aligned}$$

So,

$$Q(\beta) = (\beta - Mm)^T M^{-1} (\beta - Mm) + c^* \tag{1.9}$$

where

$$
\begin{aligned}
M^{-1} &= V_\beta^{-1} + X^T V_y^{-1} X \\
m &= V_\beta^{-1}\mu_\beta + X^T V_y y \\
c^* &= \mu_\beta^T V_\beta^{-1}\mu_\beta + y^T V_y^{-1} y - m^T M m
\end{aligned}
$$

Hence the posterior density is

$$
\begin{aligned}
p(\beta,\sigma^2|y) &\propto \left(\frac{1}{\sigma^2}\right)^{a+\frac{n}{2}+\frac{p}{2}+1} \exp\left\{-\frac{1}{\sigma^2}\left(b + \frac{c^*}{2} + \frac{1}{2}(\beta - Mm) M^{-1} (\beta - Mm)\right)\right\} \\
&\propto \underbrace{\left(\frac{1}{\sigma^2}\right)^{a+\frac{n}{2}+1} \exp\{-\frac{b + c^*/2}{\sigma^2}\}}_{f(\sigma^2|y)} \times \\
&\quad \underbrace{\left(\frac{1}{\sigma^2}\right)^{p/2} \exp\{-\frac{1}{2\sigma^2}(\beta - Mm)^T M^{-1} (\beta - Mm)\}}_{g(\beta|y,\sigma^2)} \qquad (1.10) \\
&\propto f(\sigma^2|y) \times g(\beta|y,\sigma^2)
\end{aligned}
$$

Thus,

$$
f(\sigma^2|y) \propto \mathsf{IG}(\sigma^2|a + \frac{n}{2}, b + \frac{c^*}{2}) \qquad (1.11)
$$

$$
g(\beta|y,\sigma^2) \propto \mathcal{N}(\beta|Mm, \sigma^2 M) \qquad (1.12)
$$

So we have factorized:

$$
p(\beta,\sigma^2|y) = p(\sigma^2|y)p(\beta|\sigma^2,y) = \mathsf{IG}(\sigma^2|a + \frac{n}{2}, b + \frac{c^*}{2}) \times \mathcal{N}(\beta|Mm, \sigma^2 M)
$$

Next class: sampling from the above.

1.3 Homework week 1

1. Given $X \perp\!\!\!\perp (Y,W)|Z$, prove $X \perp\!\!\!\perp W|Z$ using factorization and integration.
2. Given $X \perp\!\!\!\perp (Y,W)|Z$, prove $X \perp\!\!\!\perp Y|W,Z$.
3. (Contraction) Given

   $$(X \perp\!\!\!\perp W|Z,Y) \wedge (X \perp\!\!\!\perp Y|Z)$$

   Prove

   $$X \perp\!\!\!\perp (Y,W)|Z$$

2 Week 2

2.1 04-05

2.1.1 Bayesian conjugate linear regression

See slides. Highlights:

- Page 4, be careful about notations. Building blocks of multivariate statistics.
- Page 6, second equality should be "proportional to".
- Page 8 to 10: **Derive it ON YOUR OWN** (Mm formula).
- Page 11: composition sampling.
- Page 12: Automatic marginalization as a theoretical justification of composition sampling.
- Page 14: Sampling from predictive distribution.

2.1.2 Multivariate simulation

Two random variable case: X and Y. The joint density is $p(x, y)$.

Composition sampling or **mixture sampling**:

$$p_{X,Y}(x, y) = p_X(x)p_{Y|X}(y|x)$$

Sample

$$\begin{aligned} x_0 &\sim p_X(x) \\ y_0 &\sim p_{Y|X}(y|x_0) \\ (x_0, y_0) &\sim ? \end{aligned}$$

Where have I sampled y_0 from?

$$\mathbb{P}(x_0 \leq u, y_0 \leq v) = ?$$

Abuse of notation:

$$\begin{aligned} \mathbb{P}(x_0 = u, y_0 = v) &= \underbrace{\mathbb{P}(x_0 = u)}_{p_X(u)} \underbrace{\mathbb{P}(y_0 = v|x_0 = u)}_{p_{Y|X}(v|u)} \\ &= p_{X,Y}(u, v) \end{aligned} \tag{2.1}$$

Thus,

$$\begin{aligned}\mathbb{P}(y_0 = v) &= \int \mathbb{P}(x_0 = u, y_0 = v)du \\ &= \int p_{X,Y}(u, v)du \\ &= p_Y(v)\end{aligned}$$

What about $\mathbb{P}(x_0 = u)$? Immediately from our algorithm:

$$\mathbb{P}(x_0 = u) = p_X(u)$$

Generalization: **automatic marginalization**. Let

$$(y_1, y_2, \cdots, y_n) \sim p(y_1, y_2, \cdots, y_n)$$

If $(y_1, y_2, \cdots, y_n)$ is a sample from $p(y_1, y_2, \cdots, y_n)$, then each $y_i \sim p_i(y)$, where $p_i(y)$ is the marginal distribution of y_i. That is,

$$p_i(y) = \int p(u_1, u_2, \cdots, u_{i-1}, \underbrace{y}_{i^{th}}, u_{i+1}, \cdots, u_m)du_{(-i)}$$

which is an $(n-1)$-dimensional integral. **Mixture sampling**: sample in sequence

$$p(y_1), p(y_2|y_1), \cdots, p(y_i|y_1, \cdots, y_{i-1})$$

The resulting sample is from

$$p(y_1, y_2, \cdots y_{i-1}, y_i)$$

and

$$y_i \sim p_i(y) = \int p(y_1, \cdots, y_i)dy_1 \cdots dy_{i-1}$$

Take home message: sample so that avoid integral.

2.2 04-07

2.2.1 Multivariate simulation

Weak union: $X \perp\!\!\!\perp (Y, W)|Z \Rightarrow X \perp\!\!\!\perp Y|W, Z$ and $X \perp\!\!\!\perp W|Y, Z$. Take another look at

$$[X, Y, W|Z]$$

Assume $(X \perp\!\!\!\perp W|Y,Z) \wedge (X \perp\!\!\!\perp Y|Z)$,

$$\begin{aligned}[X,Y,W|Z] &= [X,Y|Z][W|X,Y,Z] \\ &= [X|Z][Y|Z]\underbrace{[W|X,Y,Z]}_{[W|Y,Z]} \\ &= [X|Z][W,Y|Z]\end{aligned}$$

Contraction: $(X \perp\!\!\!\perp W|Y,Z) \wedge (X \perp\!\!\!\perp Y|Z) \Rightarrow X \perp\!\!\!\perp (W,Y)|Z$.

What about **full conditionals**? Consider now 3 variables, X, Y and W.

$$[X \perp\!\!\!\perp Y|W] \wedge [X \perp\!\!\!\perp W|Y]$$

Can we say anything about $[X,Y]$ and $[X,W]$ (under regularity assumption)? **Peculiar** or **curious**:

$$[X|Y,W] = [X|W] = [X|Y]$$

The second part implies that $[X|Y]$ does not depend on Y and $[X|W]$ does not depend on W. Another look:

$$\begin{aligned}[X] &= \int [X,W] \\ &= \int [W][X|W] \\ &= \int [W][X|Y] \\ &= [X|Y]\int [W] \\ &= [X|Y] \qquad (2.2)\end{aligned}$$

Analogously,

$$X \perp\!\!\!\perp Y$$

More to 4 variables: X, Y, W and Z.

$$(X \perp\!\!\!\perp Y|W,Z) \wedge (X \perp\!\!\!\perp W|Y,Z)$$

Similarly,

$$[X|Y,W,Z] = [X|W,Z] = [X|Y,Z]$$

Thus,

$$\begin{aligned}[X|Z] &= \int [X, Y|Z] \\ &= \int [X|Y, Z][Y|Z] \\ &= \int [X|W, Z][Y|Z] \\ &= [X|W, Z] \end{aligned} \tag{2.3}$$

Analogously,

$$X \perp\!\!\!\perp W|Z \text{ and } X \perp\!\!\!\perp Y|Z$$

Hence,

$$X \perp\!\!\!\perp (W, Y)|Z$$

Earlier comment (2-variable case): $p(x, y)$ is a bivariate pmf or density. Suppose we fix (x_0, y_0) at some specific values for x and y (in their support). Can I write

$$p(x, y) = \text{[only in terms of conditionals] } p(x_0, y_0)$$

Algorithmic:

$$p(x_0, y_0) \to p(x_0, y) \to p(x, y)$$

How would this proceed if I can only compute $p(\cdot|\cdot)$ WITHOUT integration?

$$\begin{aligned} & p(x_0, y_0) = p(x_0)p(y_0|x_0) \\ \Rightarrow & \frac{p(x_0, y_0)}{p(y_0|x_0)} = p(x_0) \\ \Rightarrow & \frac{p(y|x_0)}{p(y_0|x_0)} p(x_0, y_0) = p(x_0)p(y|x_0) \\ \Rightarrow & \frac{p(y|x_0)}{p(y_0|x_0)} p(x_0, y_0) = p(x_0, y) \end{aligned}$$

Question:

$$(?) \times p(x_0, y) = p(x, y)$$

$$p(x_0, y) = p(y)p(x_0|y) \Rightarrow \frac{1}{p(x_0|y)} p(x_0, y) = p(y)$$

Thus,

$$\frac{p(x|y)}{p(x_0|y)} p(x_0, y) = p(y)p(x|y) = p(x, y)$$

So, let's go backward:

$$\begin{aligned} p(x,y) &= \frac{p(x|y)}{p(x_0|y)} p(x_0,y) \\ &= \frac{p(x|y)}{p(x_0|y)} \frac{p(y|x_0)}{p(y_0|x_0)} p(x_0,y_0) \end{aligned}$$

2.2.2 Logistic Issues

- Midterm: May 5^{th}, 2021.
- Final: Take-home.
- Midterm rule: 4 problems @ 25 points each.
  1. is an absolutely easy problem.
  2. moderately easy.
  3. moderately difficult.
  4. very challenging.

2.3 Homework week 2

1. (Brook's lemma) Repeat above algorithm for 3-variables.

   $$p(x,y,z) = \frac{?}{?} \times p(x_0,y_0,z_0)$$

   Note $\frac{?}{?}$ only depends upon $p(\cdot|\cdot,\cdot)$.

2. Recall notes from Bayesian linear regression:

   $$\begin{aligned} y|\beta,\sigma^2 &\sim \mathcal{N}(X\beta, \sigma^2 V_y) \\ \beta|\sigma^2 &\sim \mathcal{N}(\mu_\beta, \sigma^2 V_\beta) \\ \sigma^2 &\sim IG(a,b) \end{aligned}$$

   Unknown: $\{\beta, \sigma^2\}$. Joint density:

   $$p(\beta,\sigma^2,y) \propto p(\beta,\sigma^2) \cdot p(y|\beta,\sigma^2)$$

   Look at

   $$p(\sigma^2|y) \cdot p(\beta|\sigma^2,y) = p(\beta,\sigma^2|y) \propto p(\beta,\sigma^2) p(y|\beta,\sigma^2)$$

   This means

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

Show that we can derive the marginal posterior of σ^2 as

$$p(\sigma^2|y) \propto \frac{p(0,\sigma^2)p(y|0,\sigma^2)}{p(0|\sigma^2,y)} \tag{2.4}$$

by setting $\beta = 0$ (could set β to any value; 0 is convenient). You will get

$$p(\sigma^2|y) = IG(\sigma^2|a^*,b^*)$$

as in class notes.

3 Week 3

3.1 04-12

3.1.1 Bayesian regression revisited

Recall

$$y|\beta,\sigma^2 \sim \mathcal{N}(X\beta,\sigma^2 V_y)$$

$$\beta|\sigma^2 \sim \mathcal{N}(\mu_\beta,\sigma^2 V_\beta)$$

$$\sigma^2 \sim IG(a,b)$$

Then

$$p(\beta,\sigma^2|y) = IG(\sigma^2|a^*,b^*) \times \mathcal{N}(\beta|Mm,\sigma^2 M)$$

where

$$\begin{aligned} a^* &= a + \frac{n}{2} \\ b^* &= b + \frac{1}{2}\left(y^T V_y^{-1} y + \mu_\beta^T V_\beta^{-1}\mu_\beta - m^T M m\right) \\ m &= V_\beta^{-1}\mu_\beta + X^T V_y^{-1} y \\ M^{-1} &= V_\beta^{-1} + X^T V_y^{-1} X \end{aligned}$$

Sampling from $p(\beta,\sigma^2|y)$, how? Recall **method of mixtures** (composition sampling). In general: to simulate $\theta_1,\theta_2 \sim p(\theta_1,\theta_2)$,

1. $\theta_1^{(i)} \sim p(\theta_1)$
2. $\theta_2^{(i)} \sim p(\theta_2|\theta_1^{(i)})$

Then

$$\{(\theta_1^{(i)},\theta_2^{(i)})\}_{i=1}^N \sim p(\theta_1,\theta_2)$$

Posterior sampling (exact):

$$p(\theta_1, \theta_2|y) = \underbrace{p(\theta_1|y)p(\theta_2|\theta_1, y)}_{\text{available for sampling}}$$

For $(i$ in $1:M)$ {

1. $\theta_1^{(i)} \sim p(\theta_1|y)$
2. $\theta_2^{(i)} \sim p(\theta_2|\theta_1^{(i)}, y)$

}

Apply to: $p(\beta, \sigma^2|y) = \underbrace{IG(\sigma^2|a^*, b^*)}_{p(\sigma^2|y)} \underbrace{\mathcal{N}(\beta|Mm, \sigma^2 M)}_{p(\beta|\sigma^2, y)}$.

For $(i$ in $1:M)$ {

1. $\sigma^2_{(i)} \sim IG(a^*, b^*)$
2. $\beta_{(i)} \sim \mathcal{N}(Mm, \sigma^2_{(i)} M)$

} and

$$\{\beta_{(i)}, \sigma^2_{(i)}\}_{i=1}^M \sim p(\beta, \sigma^2|y)$$

Augmented linear model: useful for understanding complicated densities.

Motivation: $y \sim \mathcal{N}(\mu, \sigma^2)$ or

$$y = \mu + \epsilon, \epsilon \sim \mathcal{N}(0, \sigma^2)$$

This shows the connection/interplay between normal densities and the linear model.

$$y|\mu, \sigma^2 \sim \mathcal{N}(\mu, \sigma^2)$$

$$\mu|\sigma^2 \sim \mathcal{N}(\theta, \sigma^2\alpha)$$

Thus,

$$y = \mu + \epsilon;\ \epsilon \sim \mathcal{N}(0, \sigma^2)$$

$$\mu = \theta + \tilde{\epsilon};\ \tilde{\epsilon} \sim \mathcal{N}(0, \sigma^2\alpha)$$

$$\underbrace{\begin{bmatrix} y \\ \theta \end{bmatrix}}_{y^*} = \underbrace{\begin{bmatrix} 1 \\ 1 \end{bmatrix}}_{x^*} \mu + \underbrace{\begin{bmatrix} \epsilon \\ -\tilde{\epsilon} \end{bmatrix}}_{\epsilon^*}$$

This is the **augmented linear model**.

Bayesian regression:

$$y|\beta, \sigma^2 \sim \mathcal{N}(X\beta, \sigma^2 V_y)$$

$$\beta|\sigma^2 \sim \mathcal{N}(\mu_\beta, \sigma^2 V_\beta)$$

$$\sigma^2 \sim IG(a, b)$$

Then

$$\underbrace{\begin{bmatrix} y \\ \mu_\beta \end{bmatrix}}_{y^*} = \underbrace{\begin{bmatrix} X_{n\times p} \\ I_{p\times p} \end{bmatrix}}_{X^*} \beta + \underbrace{\begin{bmatrix} e_y \\ e_\beta \end{bmatrix}}_{e^*}$$

and

$$y^* = X^*\beta + e^*$$

$$e^* \sim \mathcal{N}\left(\begin{pmatrix} 0 \\ 0 \end{pmatrix}, \sigma^2 \underbrace{\begin{pmatrix} V_y & 0 \\ 0 & V_\beta \end{pmatrix}}_{V^*} \right)$$

So $e^* \sim \mathcal{N}(0, \sigma^2 V^*)$. In 250B territory (general linear model):

$$y^* = X^*\beta + e^*;\ e^* \sim \mathcal{N}(0, \sigma^2 V^*)$$

MLE of β is $\widehat{\beta}$ and it solves the (weighted) normal equation:

$$X_*^T V_*^{-1} X_* \beta = X_*^T V_*^{-1} y_*$$

$$\widehat{\beta} = (X_*^T V_*^{-1} X_*)^{-1} X_*^T V_*^{-1} y_*$$

Residual sums of squares:

$$\text{RSS}(y_*, X_*, V_*) = (y_* - X_*\widehat{\beta})^T V_*^{-1} (y_* - X_*\widehat{\beta})$$

Return to Bayes:

$$y_*|\sigma^2, \beta \sim \mathcal{N}(X_*\beta, \sigma^2 V_*)$$

$$\sigma^2 \sim IG(a, b)$$

Thus,

$$\begin{aligned} p(\beta, \sigma^2|y) &\propto p(\sigma^2) p(y_*|\beta, \sigma^2) \\ &\propto \text{IG}(\sigma^2|a, b) \times \mathcal{N}(y_*|X_*\beta, \sigma^2 V_*) \\ &= (\frac{1}{\sigma^2})^{a+1} e^{-b/\sigma^2} \times (\frac{1}{\sigma^2})^{\frac{n+p}{2}} e^{-\frac{1}{2\sigma^2} Q(\beta, y_*, X_*, V_*)} \end{aligned}$$

where

$$Q(\beta, y_*, X_*, V_*) = (y_* - X_*\beta)^T V_*^{-1} (y_* - X_*\beta)$$

Note that

$$\begin{aligned}
&(y_* - X_*\beta)^T V_*^{-1}(y_* - X_*\beta) \\
&= (y_* - X_*\widehat{\beta} + X_*\widehat{\beta} - X_*\beta)^T V_*^{-1}(y_* - X_*\widehat{\beta} + X_*\widehat{\beta} - X_*\beta) \\
&= (y_* - X_*\widehat{\beta} + X_*(\widehat{\beta} - \beta))^T V_*^{-1}(y_* - X_*\widehat{\beta} + X_*(\widehat{\beta} - \beta)) \\
&= (y_* - X_*\widehat{\beta})^T V_*^{-1}(y_* - X_*\widehat{\beta}) + 2(\widehat{\beta} - \beta)^T X_*^T V_*^{-1}(y_* - X_*\widehat{\beta}) + \\
&\quad (\beta - \widehat{\beta})^T X_*^T V_*^{-1} X_*(\beta - \widehat{\beta})
\end{aligned}$$

Choose $\widehat{\beta}$ such that

$$V_*^{-1/2} X_*(\widehat{\beta} - \beta) \perp\!\!\!\perp V_*^{-1/2}(y_* - X_*\widehat{\beta})$$

or choose $\widehat{\beta}$ such that

$$y_* - X_*\widehat{\beta} \perp\!\!\!\perp \mathcal{C}(V_*^{-1} X_*)$$

Thus $\widehat{\beta}$ must satisfy:

$$\underbrace{X_*^T V_*^{-1}(y_* - X_*\widehat{\beta}) = 0}_{\text{Solve for } \widehat{\beta}}$$

Back to the quadratic form:

$$\begin{aligned}
Q(\beta, y_*, X_*, V_*) &= (y_* - X_*\widehat{\beta})^T V_*^{-1}(y_* - X_*\widehat{\beta}) + (\beta - \widehat{\beta})^T X_*^T V_*^{-1} X_*(\beta - \widehat{\beta}) \\
&= \mathsf{RSS}(y_*, X_*, V_*) + (\beta - \widehat{\beta})^T X_*^T V_*^{-1} X_*(\beta - \widehat{\beta})
\end{aligned}$$

Thus,

$$p(\beta, \sigma^2|y) \propto (\frac{1}{\sigma^2})^{a+\frac{n}{2}+1} e^{-\frac{1}{\sigma^2}\{b+\frac{1}{2}\mathsf{RSS}(y_*, X_*, V_*)\}} \times (\frac{1}{\sigma^2})^{p/2} e^{-\frac{1}{2\sigma^2}(\beta-\widehat{\beta})^T X_*^T V_*^{-1} X_*(\beta-\widehat{\beta})}$$

$$\propto IG(\sigma^2|a^*, b^*) \times \mathcal{N}(\beta|\widehat{\beta}, \sigma^2(X_*^T V_*^{-1} X_*)^{-1})$$

where

$$\begin{aligned}
a^* &= a + \frac{n}{2} \\
b^* &= b + \frac{1}{2}\mathsf{RSS}(y_*, X_*, V_*) \\
\widehat{\beta} &: \text{ solves } X_*^T V_*^{-1} X_*\widehat{\beta} = X_*^T V_*^{-1} y_*
\end{aligned}$$

This exactly matches the posterior distribution we derived in class earlier.

3.2 04-14

3.2.1 Brook's lemma

Recap:

$$p(x,y,z) = \frac{p(x|y,z)}{p(x_0|y,z)} \frac{p(y|x_0,z)}{p(y_0|x_0,z)} \frac{p(z|x_0,y_0)}{p(z_0|x_0,y_0)} p(x_0,y_0,z_0)$$

Theorem 3.1 (Brook's Lemma / Hammersley-Clifford Theorem)**.** In general, suppose $x \in \mathbb{R}^n$, then

$$\begin{aligned} & p(x_1, x_2, \cdots, x_n) \\ &= \frac{p(x_1|x_2,\cdots,x_n)}{p(x_{10}|x_2,\cdots,x_n)} \frac{p(x_2|x_{10},x_3,\cdots,x_n)}{p(x_{20}|x_{10},x_3,\cdots,x_n)} \cdots \frac{p(x_n|x_{10},\cdots,x_{(n-1)0})}{p(x_{n0}|x_{10},\cdots,x_{(n-1)0})} p(x_{10},x_{20},\cdots,x_{n0}) \end{aligned} \tag{3.1}$$

In 1964, Brook proved this lemma, answering one of the key questions in 1920s and 1930s. Later this lemma was generalized by Hammersley and Clifford.

Message:

Complete set of full conditionals lets us to the joint. Need positivity conditions for $p(\cdot)$.

3.2.2 More on sampling

Back to Bayesian Linear Model: Bayesian predictions.

Recall

$$[\text{ Unknown | Known }]$$

Model: likelihood $p(y|\theta)$ and prior $p(\theta)$, y known and θ unknown. Prediction introduces a new unknown: $\widetilde{Y}$.

Bayesian:

$$[\text{ Unknown | Known }] \to [\theta, \widetilde{Y}|y] \propto \underbrace{[\theta|y]}_{\text{posterior}} \times \underbrace{[\widetilde{y}|\theta,y]}_{\text{cond pred}}$$

What we seek? Sample from $[\theta, \widetilde{y}|y]$. Devise an algorithm to sample from this!

For i in $1:M$, do:

- $\theta^{(i)} \sim [\theta|y]$.
- $\widetilde{y}^{(i)} \sim [\widetilde{y}|\theta^{(i)}, y]$.

Resulting samples $\{\theta^{(i)}, \widetilde{y}^{(i)}\}_{i=1}^M$ follow $[\theta, \widetilde{y}|y]$. Thus, $\{\widetilde{y}^{(i)}\}_{i=1}^M$ are samples from

$$\underbrace{[\,\widetilde{y} \mid y\,]}_{\text{posterior predictive}}$$

Separate posterior sampling from predictive sampling.

Posterior sampling: For i in $1:M$, do:

- $\theta^{(i)} \sim [\,\theta \mid y\,]$.

Predictive sampling: For i in $1:M$, do:

- $\widetilde{y}^{(i)} \sim [\,\widetilde{y} \mid \theta^{(i)}, y\,]$.

Note: if we replace $\theta^{(i)}$ by a point estimate, then we will get narrower confidence intervals but will not be able to address the uncertainty of θ.

Also recall posterior distributions are proportional to joint distributions.

Example 3.1.

$$p(\theta|y) \propto p(\theta)p(y|\theta)$$

$$[\theta|y] \propto [\theta][y|\theta]$$

$$[\theta, \widetilde{y}|y] \propto \underbrace{[\theta][y|\theta][\widetilde{y}|\theta, y]}_{[\theta,\widetilde{y},y]}$$

Building models:

$$[\theta], [y|\theta] \text{ and } [\widetilde{y}|\theta, y]$$

and

$$\widetilde{y} \perp\!\!\!\perp y|\theta$$

Example 3.2.

$$\begin{aligned} y|\beta, \sigma^2 &\sim \mathcal{N}(X\beta, \sigma^2 V_y) \\ \beta|\sigma^2 &\sim \mathcal{N}(\mu_\beta, \sigma^2 V_\beta) \\ \sigma^2 &\sim IG(a, b) \end{aligned}$$

We wish to predict $\widetilde{y}$ ($m \times 1$) with predictors $\widetilde{X}$ ($m \times p$). Cast this into

$$[\theta, \widetilde{y}|y] \propto [\theta, \widetilde{y}, y]$$

framework with $\theta = \{\beta, \sigma^2\}$.

$$[\theta, \widetilde{y}|y] \propto [\theta][y|\theta][\widetilde{y}|\theta, y] \propto [\theta|y][\widetilde{y}|\theta, y]$$

We already know how to sample from

$$[\theta|y] = [\beta, \sigma^2|y] \propto IG(\sigma^2|a^*, b^*) \times \mathcal{N}(\beta|Mm, \sigma^2 M)$$

where

$$a^* = a + \frac{n}{2}$$

$$b^* = b + \frac{1}{2}\{y^T V_y^{-1} y + \mu_\beta^T V_\beta^{-1} \mu_\beta - m^T M m\}$$

$$M^{-1} = V_\beta^{-1} + X^T V_y^{-1} X$$

$$m = V_\beta^{-1}\mu_\beta + X^T V_y^{-1} y$$

What remains is the model for $[\widetilde{y}|\theta, y]$. Freedom to model. Assume

$$\widetilde{y}|\theta \sim \mathcal{N}(\widetilde{X}\beta, \sigma^2\widetilde{V})$$

That is, $\widetilde{y} \perp\!\!\!\perp y \mid \theta$.

Given posterior samples $\{\beta^{(i)}, \sigma^2_{(i)}\}_{i=1}^M$: For i in $1:M$, do:

- $\widetilde{y}^{(i)} \sim \mathcal{N}(\widetilde{X}\beta^{(i)}, \sigma^2_{(i)}\widetilde{V})$.

Special case: $V_y = I_n$, i.e., classical linear regression.

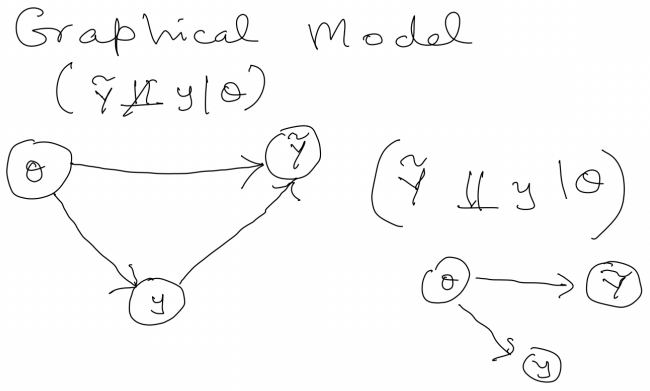
$$\begin{aligned} y|\beta, \sigma^2 &\sim \mathcal{N}(X\beta, \sigma^2 I_n) \\ \beta|\sigma^2 &\sim \mathcal{N}(\mu_\beta, \sigma^2 V_\beta) \\ \sigma^2 &\sim IG(a, b) \\ \widetilde{y}|\beta, \sigma^2 &\sim \mathcal{N}(\widetilde{X}\beta, \sigma^2 I_m) \end{aligned}$$

So predictions:

$$\widetilde{y}^{(i)} \sim \mathcal{N}(\widetilde{X}\beta^{(i)}, \sigma^2_{(i)} I_m)$$

and $\widetilde{y}^{(i)} \sim [\widetilde{y}|y]$.

Figure 2: Graphical model.



3.3 Homework week 3

1. Verify the following:
   - $X_*^T V_*^{-1} X_* = V_\beta^{-1} + X^T V_y^{-1} X$ (we call M^{-1}).
   - $\widehat{\beta} = Mm$, $m = V_\beta^{-1}\mu_\beta + X^T V_y^{-1} y$.
   - RSS$(y_*, X_*, V_*) = y^T V_y^{-1} y + \mu_\beta^T V_\beta^{-1}\mu_\beta - m^T M m$.

2. $p(y|x) = \mathcal{N}(y|x, 1)$ and $p(x|y) = \mathcal{N}(x|y, 1)$. Apply Brook's lemma to find a form (up to proportionality constant) of $p(x, y)$. What can you say about $p(x, y)$?

3. Consider the BHM,

$$y|\beta, \sigma^2 \sim \mathcal{N}(X\beta, \sigma^2 I_n)$$

$$\beta|\sigma^2 \sim \mathcal{N}(\mu_\beta, \sigma^2 V_\beta)$$

$$\sigma^2 \sim IG(a, b)$$

Assume

$$\widetilde{y}|\beta, \sigma^2, y \sim \mathcal{N}(\widetilde{X}\beta, \sigma^2 I_m)$$

and find what specifications for a, b and V_β^{-1} will yield:

- $p(\beta|\sigma^2, y) = \mathcal{N}(\widehat{\beta}, \sigma^2(X^TX)^{-1})$ where $\widehat{\beta}$ solves $X^TX\widehat{\beta} = X^Ty$.
- $\mathbb{E}(\sigma^2|y) = \widehat{\sigma}^2$ from classical analysis.
- $Var(\widetilde{y}|\sigma^2, y) =?$ Compare the classical.

4 Week 4

4.1 04-19

4.1.1 Sampling

Recall:

$$\text{Data (known): } y, X, \widetilde{X}$$

$$\text{Unknown: } \theta = \{\beta, \sigma^2\}, \widetilde{y}$$

$$p(\theta, \widetilde{y}|y) = p(\theta|y) \times p(\widetilde{y}|\theta, y)$$

Directly sample (draw i.i.d. samples) from $p(\theta, \widetilde{y}|y)$ using method of mixtures (composition):

$$\{\theta^{(i)}\}_{i=1}^M \sim p(\theta|y)$$
$$\{\widetilde{y}^{(i)}\}_{i=1}^M \sim p(\widetilde{y}|\theta^{(i)}, y) \tag{4.1}$$
$$\Rightarrow \{\widetilde{y}^{(i)}\}_{i=1}^M \sim p(\widetilde{y}|y) \tag{4.2}$$

Constrast with MCMC: MCMC will also target sampling from $p(\theta, \widetilde{y}|y)$ or $p(\theta|y) = p(\beta, \sigma^2|y)$.

Example 4.1 (Gibbs sampling)**.**

$$\beta^{(i)} \sim p(\beta|\sigma^2_{(i-1)}, y)$$
$$\sigma^2_{(i)} \sim p(\sigma^2|\beta^{(i-1)}, y) \tag{4.3}$$

Eventually

$$\underbrace{\{\beta^{(i)}, \sigma^2_{(i)}\}}_{\theta^{(i)}} \to_d p(\beta, \sigma^2|y)$$

Posterior predictive sampling: same as before:

$$\widetilde{y}^{(i)} \sim p(\widetilde{y}|\theta^{(i)}, y)$$

Alternative: include sampling of $\widetilde{y}$ in the Gibbs sampler:

$$\begin{aligned}
\beta^{(i)} &\sim p(\beta|\sigma^2_{(i-1)}, y, \widetilde{y}^{(i-1)}) \\
\sigma^2_{(i)} &\sim p(\sigma^2|\beta^{(i)}, y, \widetilde{y}^{(i-1)}) \\
\widetilde{y}^{(i)} &\sim p(\widetilde{y}|\beta^{(i)}, y, \sigma^2_{(i)})
\end{aligned} \tag{4.4}$$

Eventually

$$\{\beta^{(i)}, \sigma^2_{(i)}, \widetilde{y}^{(i)}\} \to_d p(\beta, \sigma^2, \widetilde{y}|y)$$

Example 4.2 (General Gibbs sampler)**.** Target distribution:

$$p(\theta_1, \theta_2, \cdots, \theta_p|y)$$

Algorithm: for i in $1:M$, do:

$$\begin{aligned}
\theta_1^{(i)} &\sim p(\theta_1|\theta_2^{(i-1)}, \cdots, \theta_p^{(i-1)}, y) \\
&\cdots\cdots \\
\theta_j^{(i)} &\sim p(\theta_j|\theta_1^{(i)}, \theta_2^{(i)}, \cdots, \theta_{j-1}^{(i)}, \theta_{j+1}^{(i-1)}, \theta_p^{(i-1)}, y)
\end{aligned} \tag{4.5}$$

4.1.2 Augmented linear model (with predictions)

Think of linear models:

$$y = X\beta + e_y, e_y \sim \mathcal{N}(0, \sigma^2 I_n) \tag{4.6}$$

$$\beta = \mu_\beta + e_\beta, e_\beta \sim \mathcal{N}(0, \sigma^2 V_\beta) \tag{4.7}$$

$$\widetilde{y} = \widetilde{X}\beta + e_{\widetilde{y}}, e_{\widetilde{y}} \sim \mathcal{N}(0, \sigma^2 I_m) \tag{4.8}$$

We have

$$\underbrace{\begin{bmatrix} y \\ \mu_\beta \\ 0 \end{bmatrix}}_{y_*} = \underbrace{\begin{bmatrix} X_{n\times p} & \mathbf{0}_{n\times m} \\ I_p & \mathbf{0}_{p\times n} \\ \widetilde{X}_{m\times p} & -I_m \end{bmatrix}}_{X_*} \underbrace{\begin{bmatrix} \beta \\ \widetilde{y} \end{bmatrix}}_{\gamma} + \underbrace{\begin{bmatrix} e_y \\ e_\beta \\ e_{\widetilde{y}} \end{bmatrix}}_{e_*} \tag{4.9}$$

Thus,

$$p(\gamma, \sigma^2|y_*) \propto p(\sigma^2|y_*) \times p(\gamma|\sigma^2, y_*)$$
$$p(\gamma, \sigma^2|y) \propto p(\sigma^2|y) \times p(\gamma|\sigma^2, y)$$

4.1.3 Sherman-Woodbury-Morrison (SWM) identity

Look at

$$\mathcal{N}(x|0, A) \times \mathcal{N}(y|Bx, D)$$

where A is $p \times p$, x is $p \times 1$, y is $n \times 1$, B is $n \times p$ and D is $n \times n$. Assume both A and D are p.s.d.,

$$p(y|x) = \mathcal{N}(y|Bx, D)$$
$$p(x) = \mathcal{N}(x|0, A)$$

Question: what is $p(y)$?

First approach: reformulate as linear models.

$$y = Bx + e_y, e_y \sim \mathcal{N}(0, D)$$

$$x = e_x, e_x \sim \mathcal{N}(0, A)$$

Thus,

$$y = Be_x + e_y$$

so y is normally distributed.

$$\mathbb{E}[y] = B\mathbb{E}[e_x] + \mathbb{E}[e_y] = 0$$
$$Var(y) = BVar(e_x)B^T + Var(e_y) = BAB^T + D \tag{4.10}$$

Thus,

$$p(y) = \mathcal{N}(y|0, D + BAB^T)$$

and $y \sim \mathcal{N}(0, V_y),\ V_y = D + BAB^T$.

Second approach: integrate out x.

$$p(y) = \int_{\mathbb{R}} p(x)p(y|x)dx$$

Thus,

$$\begin{aligned}
p(x,y) &= p(x) \times p(y|x) \\
&= \mathcal{N}(x|0,A) \times \mathcal{N}(y|Bx,D) \\
&\propto \frac{1}{|A|^{1/2}} e^{-\frac{1}{2}x^T A^{-1} x} \frac{1}{|D|^{1/2}} e^{-\frac{1}{2}(y-Bx)^T D^{-1}(y-Bx)} \\
&\propto \frac{1}{|A|^{1/2}|D|^{1/2}} e^{-\frac{1}{2}Q} \\
&\propto e^{-\frac{1}{2}Q}
\end{aligned}$$

What is Q? Key observation: $p(x,y) = p(x)p(y|x) = p(y)p(x|y)$. Then

$$\begin{aligned}
Q &= x^T A^{-1} x + (y-Bx)^T D^{-1}(y-Bx) \\
&= x^T A^{-1} x + y^T D^{-1} y - 2x^T B^T D^{-1} y + x^T B^T D^{-1} B x \\
&= x^T (A^{-1} + B^T D^{-1} B) x - 2x^T B^T D^{-1} y + y^T D^{-1} y \\
&= \underbrace{x^T M^{-1} x - 2x^T m}_{M^{-1}=A^{-1}+B^T D^{-1} B;\ m=B^T D^{-1} y} + y^T D^{-1} y \\
&= (x-Mm)^T M^{-1}(x-Mm) - m^T M m + y^T D^{-1} y
\end{aligned}$$

Thus,

$$\begin{aligned}
p(x,y) &\propto e^{-\frac{1}{2}(x-Mm)^T M^{-1}(x-Mm)} \times e^{-\frac{1}{2}\{y^T D^{-1} y - m^T M m\}} \\
&\propto \mathcal{N}(x|Mm,M) \times e^{-\frac{1}{2}Q^*(y)}
\end{aligned} \tag{4.11}$$

and

$$\begin{aligned}
p(x|y) &= \mathcal{N}(x|Mm,M) \\
p(y) &\propto e^{-\frac{1}{2}Q^*(y)}
\end{aligned}$$

where

$$\begin{aligned}
Q^*(y) &= y^T D^{-1} y - m^T M m \\
&= y^T D^{-1} y - y^T D^{-1} B (A^{-1} + B^T D^{-1} B)^{-1} B^T D^{-1} y \\
&= y^T \{D^{-1} - D^{-1} B (A^{-1} + B^T D^{-1} B)^{-1} B^T D^{-1}\} y
\end{aligned}$$

So

$$p(y) \propto e^{-\frac{1}{2}Q^*(y)} = \mathcal{N}(y|0,V_y)$$

Here $V_y^{-1} = D^{-1} - D^{-1} B (A^{-1} + B^T D^{-1} B)^{-1} B^T D^{-1}$.

From 1^{st} approach:

$$V_y = D + B^T AB$$

From 2^{nd} approach:

$$V_y^{-1} = D^{-1} - D^{-1}B(A^{-1} + B^T D^{-1}B)^{-1}B^T D^{-1}$$

In conclusion, we have **Sherman-Woodbury-Morrison formula**:

$$(D + B^T AB)^{-1} = D^{-1} - D^{-1}B(A^{-1} + B^T D^{-1}B)^{-1}B^T D^{-1} \tag{4.12}$$

4.1.4 Linear systems

$$\begin{bmatrix} A_{p\times p} & B_{p\times n} \\ C_{n\times p} & D_{n\times n} \end{bmatrix} \begin{bmatrix} X_{p\times r} \\ Y_{n\times r} \end{bmatrix} = \begin{bmatrix} E_{p\times r} \\ F_{n\times r} \end{bmatrix} \tag{4.13}$$

Thus,

$$\begin{bmatrix} I & 0 \\ -CA^{-1} & I \end{bmatrix} \begin{bmatrix} A & B \\ C & D \end{bmatrix} \begin{bmatrix} X \\ Y \end{bmatrix} = \begin{bmatrix} I & 0 \\ -CA^{-1} & I \end{bmatrix} \begin{bmatrix} E \\ F \end{bmatrix}$$
$$\Rightarrow \begin{bmatrix} A & B \\ 0 & D - CA^{-1}B \end{bmatrix} \begin{bmatrix} X \\ Y \end{bmatrix} = \begin{bmatrix} E \\ F - CA^{-1}E \end{bmatrix} \tag{4.14}$$

Next time: simplify $E = 0$ and $F = I$.

4.2 04-21

4.2.1 Analissa's transformation and Sherman-Woodbury-Morrison

Consider

$$\begin{bmatrix} A_{p\times p} & B_{p\times n} \\ C_{n\times p} & D_{n\times n} \end{bmatrix} \begin{bmatrix} X_{p\times n} \\ Y_{n\times n} \end{bmatrix} = \begin{bmatrix} \mathbf{0}_{p\times n} \\ \mathbf{I}_{n\times n} \end{bmatrix} \tag{4.15}$$

Gaussian block elimination:

$$\begin{bmatrix} I & 0 \\ -CA^{-1} & I \end{bmatrix} \begin{bmatrix} A & B \\ C & D \end{bmatrix} \begin{bmatrix} X \\ Y \end{bmatrix} = \begin{bmatrix} I & 0 \\ -CA^{-1} & I \end{bmatrix} \begin{bmatrix} 0 \\ I \end{bmatrix}$$
$$\Rightarrow \begin{bmatrix} A & B \\ 0 & D - CA^{-1}B \end{bmatrix} \begin{bmatrix} X \\ Y \end{bmatrix} = \begin{bmatrix} 0 \\ I \end{bmatrix} \tag{4.16}$$

Thus,

$$AX + BY = 0$$

$$(D - CA^{-1}B)Y = I$$

$$\Rightarrow$$

$$Y = (D - CA^{-1}B)^{-1}$$

Take another look @ the system:

$$AX + BY = 0$$
$$CX + DY = I$$

Rewrite (**Analissa's transformation**)

$$DY + CX = I$$
$$BY + AX = 0$$

Multiply both sides by

$$\begin{bmatrix} I & 0 \\ -BD^{-1} & I \end{bmatrix}$$

Thus,

$$DY + CX = I$$
$$(A - BD^{-1}C)X = -BD^{-1}$$

Hence

$$X = -(A - BD^{-1}C)^{-1}BD^{-1}$$
$$Y = D^{-1} + D^{-1}C(A - BD^{-1}C)^{-1}BD^{-1}$$

In conclusion,

$$(D - CA^{-1}B)^{-1} = D^{-1} + D^{-1}C(A - BD^{-1}C)^{-1}BD^{-1} \tag{4.17}$$

4.2.2 Homework from week 3

Recall

$$y|\beta, \sigma^2 \sim \mathcal{N}(X\beta, \sigma^2 I_n)$$
$$\beta|\sigma^2 \sim \mathcal{N}(\mu_\beta, \sigma^2 V_\beta)$$
$$\sigma^2 \sim IG(a, b)$$

Find a, b, V_β to reproduce classical analysis

$$\widetilde{y}|\beta, \sigma^2 \sim \mathcal{N}(\widetilde{X}\beta, \sigma^2 I_m)$$

Dr. Banerjee's solution:

$$p(\beta, \sigma^2|y) \propto IG(\sigma^2|a_*, b_*) \times \mathcal{N}(\beta|Mm, \sigma^2 M)$$

and

$$a_* = a + \frac{n}{2};\ b_* = b + \frac{1}{2}\text{RSS}(y, X_*, V_*)$$

$$M^{-1} = V_\beta^{-1} + X^T X;\ m = V_\beta^{-1}\mu_\beta + X^T y$$

If $V_\beta^{-1} = 0$, then

$$Mm = (X^T X)^{-1} X^T y = \widehat{\beta}$$

Thus,

$$\beta|\sigma^2, y \sim \mathcal{N}(\widehat{\beta}, \sigma^2 (X^T X)^{-1})$$

$$\frac{a^T(\beta - \widehat{\beta})}{\sigma\sqrt{a^T (X^T X)^{-1} a}} \sim \mathcal{N}(0, 1)\ \forall a \in \mathbb{R}^p$$

For $\sigma^2|y$, we have

$$\mathbb{E}(\sigma^2|y) = \frac{b_*}{a_* - 1} = \frac{b + \frac{1}{2}\text{RSS}}{a + \frac{n}{2} - 1}$$

If we want $\widehat{\sigma}^2$ to be unbiased, then

$$\widehat{\sigma}^2 = \frac{\text{RSS}}{n - p}$$

and

$$p(\beta, \sigma^2|y) \propto (\frac{1}{\sigma^2})^{a + \frac{n}{2} + 1 + \frac{p}{2}} \times e^{-\frac{1}{\sigma^2}(b + \frac{1}{2}\text{RSS})} \times e^{-\frac{1}{2\sigma^2}(\beta - Mm)^T M^{-1} (\beta - Mm)}$$

What is the relationship between χ^2 and Gamma?

$$\frac{(n-p)\widehat{\sigma}^2}{\sigma^2} \sim \chi^2_{n-p}$$

4.2.3 Normal-Inverse-Gamma distribution (NIG)

We say

$$\{\beta, \sigma^2\} \sim \text{NIG}(\mu_\beta, V_\beta, a, b)$$

if and only if

$$\sigma^2 \sim IG(a, b)$$

$$\beta|\sigma^2 \sim \mathcal{N}(\mu_\beta, \sigma^2 V_\beta)$$

and $\{\beta, \sigma^2\}$ is a $(p + 1)$ dimensional vector. Density:

$$\text{NIG}(\mu_\beta, V_\beta, a, b) = IG(\sigma^2|a, b) \times \mathcal{N}(\beta|\mu_\beta, \sigma^2 V_\beta)$$

Bayesian linear model:

$$y|\beta,\sigma^2 \sim \mathcal{N}(X\beta,\sigma^2 V_y)$$
$$(\beta,\sigma^2) \sim \text{NIG}(\mu_\beta, V_\beta, a, b)$$

From what have already seen:

$$\beta|\sigma^2, y \sim \mathcal{N}(Mm, \sigma^2 M)$$
$$\sigma|y \sim IG(a_*, b_*)$$

Thus,

$$\beta,\sigma^2|y \sim \text{NIG}(Mm, M, a_*, b_*) \tag{4.18}$$

4.2.4 Non-informative prior and classical analysis

Look @ (non-informative) prior distribution:

$$p(\beta,\sigma^2) \propto (\frac{1}{\sigma^2})^{a+\frac{p}{2}+1} \times \exp\left(-\frac{1}{\sigma^2}\{(\beta-\mu_\beta)^T V_\beta^{-1}(\beta-\mu_\beta)\}\right)$$

Suppose

$$V_\beta^{-1} = 0,\ a = -\frac{p}{2},\ b = 0$$

Then

$$p(\beta,\sigma^2) \propto \frac{1}{\sigma^2}$$

Posterior (assume $V_y = I$):

$$\text{NIG}(Mm, M, a_*, b_*)$$
$$M = (X^T X)^{-1};\ m = X^T y$$
$$a_* = a + \frac{n}{2} = \frac{n-p}{2}$$
$$b_* = \frac{1}{2}(n-p)\widehat{\sigma}^2$$

where $\widehat{\sigma}^2$ is an unbiased estimate $\frac{RSS}{n-p}$. Therefore,

$$\sigma^2|y \sim \text{IG}(a_*, b_*)$$

or

$$\sigma^2|y \sim \text{IG}(\frac{n-p}{2}, \frac{(n-p)\widehat{\sigma}^2}{2})$$
$$\frac{1}{\sigma^2}|y \sim \text{Ga}(\frac{n-p}{2}, \frac{(n-p)\widehat{\sigma}^2}{2})$$

and

$$\mathbb{E}(\sigma^2|y) = \frac{b_*}{a_* - 1} = \frac{\widehat{\sigma}^2}{1 - \frac{2}{n-p}} \tag{4.19}$$

4.3 Homework week 4

1. Find expressions for

   $$p(\gamma, \sigma^2|y_*) \propto p(\sigma^2|y_*) \times p(\gamma|\sigma^2, y_*)$$

   What is $p(\widetilde{y}|\sigma^2, y)$ (Hint: find from $p(\gamma|\sigma^2, y)$)?

2. Is there a lurking identity related to determinant? Hint: look @

   $$p(x, y) = \underbrace{p(x)p(y|x)}_{\frac{1}{|A|^{1/2}|D|^{1/2}}\cdots} = \underbrace{p(y)p(x|y)}_{\frac{1}{|V_y|^{1/2}|M|^{1/2}}\cdots}$$

   Then

   $$|D + BAB^T| = |A_{p\times p}|^{1/2}|D|^{1/2}|?|_{p\times p}$$

3. Find the matrix M in the following identity:

   $$|D - CA^{-1}B| = |D||A||M| \tag{4.20}$$

4. Find how χ^2_{n-p} is a special case of $Gamma(a, b)$.

5 Week 5

5.1 04-26

5.1.1 Determinant version of SWM

Block-elimination: Gaussian elimination.

1. row (i) $\leftarrow$ row(i) + α row(j) [linear addition]
2. row(i) $\leftarrow \alpha$ row(i) [scalar mult]
3. row(i) $\leftrightarrow$ row(j) [swap]

For the rest, see **Biostat250C Lec Notes** on goodnotes.

5.1.2 Application of SWM: linear mixed models

See **Biostat250C Lec Notes** on goodnotes.

5.2 04-28

5.2.1 Sequential Bayesian learning

Idea: single data point $y_1 \sim p(y|\theta)$,

$$\theta \sim p(\theta)$$

Posterior:

$$p(\theta|y_1) \propto p(\theta)p(y_1|\theta)$$

Second data point: $y_2 \sim p(y|\theta)$ and $y_2 \perp\!\!\!\perp y_1|\theta$. Update:

$$p(\theta|y_1, y_2) \propto p(\theta)p(y_1, y_2|\theta)$$

For the rest, see goodnotes.

5.3 Homework week 5

1. Let A be 3×3. Fill out a table of all permutations of $(1, 2, 3)$. Then find $\det(A)$.
2. Let A be $p \times p$ and D be $n \times n$. Suppose $|A| \neq 0$ and $|D| \neq 0$. Prove that

$$\det \begin{pmatrix} A & B \\ 0 & D \end{pmatrix} = \det A \det D$$

3. What is the SWM form for

$$|U + VWX|$$

   where U is $n \times n$, W is $p \times p$, V is $n \times p$ and X is $p \times n$ and $|U| \neq 0$, $|W| \neq 0$.

4. **Coding problem**:

   - Write an R function that will generate from an NIG(mu.beta, V.beta, a,b) distribution but such that the function will require the input of the precision — inverse of V.beta.
   - Using the function you have written in the above problem, write an R function that will generate samples from the posterior distribution of $\{\beta, \sigma^2\}$.
   - Next write the function to generate samples from the posterior distribution of $\{\beta, \sigma^2\}$ such that the function will only take in the following arguments: (i) an object that is the output of the lm() function for fitting linear models; (ii) prior mean of β; (iii) prior precision of β; (iv) prior shape and rate of σ^2; and (v) the number of samples to be drawn.
   - Apply your program in 2 and 3 to the data obtained from the file LinearModelExample.txt uploaded in Week 5.

Please try to write these functions without using any "for" loops as they unnecessarily slow down R programs. While not required, it is recommended that you prepare an Rmd program to produce an html file clearly explaining the distribution theory (from class) and the use of your R functions.

References

[1] Theodore Wilbur Anderson. An introduction to multivariate statistical analysis. Technical report, Wiley New York, 2003.