Bayesian Inference and Markov Chain Monte Carlo

November 2001

Peter Lenk

Peter Lenk is Associate Professor of Statistics and Marketing, The University of Michigan Business School, Ann Arbor, MI 48109-1234, Phone: 734–936–2619 and Fax: 734–936–0274, Emai: plenk@umich.edu

Contents

Chapter 1

Introduction

1.1 Goals

1. **Formulating Bayesian models,**

2. **Analyzing these models, and**

3. **Interpreting output from software programs.**

Participants need a working knowledge of:

- **Basic statistics,**
- **Probability distributions,**
- **Matrix notation, and**
- **Computational programming languages, such as FORTRAN, C, Pascal, Basic.**

1.2 Computer Programs

- **GAUSS will be used to demonstrate Bayesian computations.**
  Learning GAUSS is not a primary objective of the workshop.

- **WinBugs is a free, software program for Bayesian analysis.**
  If is fairly powerful and flexible with a sophisticated user interface.
  It is not user–friendly but has a number of examples.
  Download WinBUGS from http://www.mrc-bsu.cam.ac.uk/bugs.

1.3 Outline

1. **Foundations**

    - **Subjective Probability**
    - **Decision Theory**
    - **Large Sample Theory**

2. **Bayesian Inference**

    - **Basic concepts**
    - **Multivariate normal, gamma, and inverted gamma distributions**
    - **Three easy models:**

      (a) **Beta–Binomial**

      (b) **Conjugate Normal**

      (c) **Conjugate, Linear Regression**

3. Linear Regression

- **Markov chain Monte Carlo (MCMC)**
- **Numerical Integration**
- **Slice sampling**
- **Autoregressive errors**

4. Multivariate Regression

- **Multiple, dependent variables**
- **Matrix algebra facts**
- **Matrix normal, Wishart, and Inverted Wishart distributions**

5. **HB Regression: Interaction Model**

   - **Within–Subject Model:**
     Linear Regression
   - **Between–Subjects Model:**
     Multivariate Regression

6. **HB Regression: Mixture Model**

   - **Within–Subject Model: Linear Regression**
   - **Between–Subjects Model: Mixture Model**
   - **Uses "latent" variables.**

7. Revealed Preference Models

- **Categorical dependent variable:**
  - **Probit assumes normal errors.**
  - **Logit assumes extreme value errors.**
  - **Multivariate Probit: many 0/1 choices.**
- **Hastings–Metropolis algorithm, a general purpose method of generating random variables in MCMC.**

References

- Berger, James *Statistical Decision Theory and Bayesian Analysis*, Springer–Verlag, New York, 1985. Good for mathematical statistics.
- Bernardo, Jose and Adrian Smith *Bayesian Theory,* Wiley, New York, 1994. Delves into some advanced topics such as exchangeability, symmetry, and invariance. Only attempt it after knowing the material in this workshop.
- Blackwell, D. and M. A. Girshick, *Theory of Games and Statistical Decisions,* Dover, New York, 1954. A classic.
- Congdon, Peter, *Bayesian Statistical Modelling*, John Wiley & Sons, 2001. Very nice treatment.
- DeGroot, Morris *Optimal Statistical Decisions*, McGraw–Hill, New York, 1970. One of the best books on the subject ever. DeGroot elegant presentation illustrates profound points while using only basic math skills.
- Gelman, A.; J. Carlin, H. Stern, and D. Rubin *Bayesian Data Analysis,* Chapman & Hall/CRC, New York. 1995. A more modern approach. Lacks detail.
- Jeffreys, Harold, *Theory of Probability*, Oxford University Press, Oxford, 1961. (Originally published in 1939) Jeffreys was a truly original thinker.
- von Neumann, John and Oskar Morgenstern, *Theory of Games and Economic Behavior,* Princeton University Press, New Jersey, 1947. A classic in economics
- Savage, Leonard J. *The Foundations of Statistics*, Dover, New York, 1972. (Originally published in 1954) A monumental work.
- Zellner, Arnold *An Introduction to Bayesian Inference in Econometric*, John Wiley & Sons, New York, 1971. A fantastic resource.

Chapter 2

Foundations

Outline

1. **Objectives**

2. **Subjective Probability**

3. **Coherence**

4. **Decision Theory**

5. **Statistical Decision Problems**

6. **Large Sample Theory**

2.1 Objectives

1. **Introduce subjective probability and its foundations.**

2. **Describe decision theoretic approach to statistical inference.**

3. **State large sample approximations for posterior distributions.**

2.2 Subjective Probability

1. **Probability distributions encode the observer's beliefs about uncertain events.**

2. **Subjective probability is more general than the frequentist interpretation.**

3. **Frequentist interpretation is logically flawed. It relies on long-term behavior or infinite sequences and the strong law of large numbers. In turn, the strong law of large numbers relies on having probabilities, which leads to circular definitions.**

4. **Bayesians use frequentist information in updating their subjective beliefs.**

5. **Long-term frequencies or repeated sampling is not a valid concept in many situations.**

2.3 Coherence

<u>Let's Gamble</u>:

1. **You are the bookie. You quote betting odds $P(A)$, $P(B)$, . . . , on events A, B, ...**

2. **I am the gambler. I bet a stake S_A on event A. S_A can be positive or negative.**

3. **It costs me $S_A P(A)$ to play the game.**

4. **If A occurs, you pay me S_A, and I win $W = S_A(1 - P(A))$.**

5. **If A does not occur, you pay me 0, and I win $W = -S_A P(A)$.**

Coherence $\Leftrightarrow$ No Arbitrage

1. **You do not want to assign P to events so that I can make a series of wagers such that I will be a sure winner, regardless of the outcomes. That is, you should guard against presenting me with an arbitrage opportunity.**

2. **P is coherent if it is assigned in such a way that there is not arbitrage.**

3. **Coherence does not mean that your specification of P is good or will make you a lot of money, only that you cannot be a sure loser.**

DeFinetti's Coherence Theorem

Suppose that the collection $\mathcal{E}$ of events is an algebra:

- **The null event $\emptyset \in \mathcal{E}$.**
- **The certain event $\Omega \in \mathcal{E}$.**
- $A \in \mathcal{E}$ **and** $B \in \mathcal{E}$ **imply that**
  - $A \cap B \in \mathcal{E}$,
  - $A \cup B \in \mathcal{E}$, **and**
  - $A^c \in \mathcal{E}$.

Then there does not exist an arbitrage opportunity if and only if P is a probability function on $\mathcal{E}$:

1. $0 \leq P(A) \leq 1$.

2. If U is a certain event, then $P(U) = 1$.

3. $A \cap B = \emptyset \Rightarrow P(A \cup B) = P(A) + P(B)$.

Proof:

1. **If** A **occurs, I win** $W_1 = S_A[1 - P(A)]$

   If A^c **occurs, I win** $W_2 = -S_A P(A)$

   Coherence requires

$$\begin{aligned} W_1 W_2 &\leq 0 \\ (1 - P(A))P(A) &\geq 0 \\ 0 \leq P(A) &\leq 1 \end{aligned}$$

2. **If** U **is a certain event, my winnings are** $W = S_U[1 - P(U)]$**. If** $P(U) < 1$**, I can make** W **arbitrarily large.**

3. Consider three events: A, B, and $C = A \cup B$ where $A \cap B = \emptyset$. I bet S_A, S_B, and S_C.

- **If $A \cap B^c$ occurs, I win**

$$W_1 = S_A[1 - P(A)] - S_B P(B) + S_C[1 - P(C)].$$

- **If $A^c \cap B$ occurs, I win**

$$W_2 = -S_A P(A) + S_B[1 - P(B)] + S_C[1 - P(C)].$$

- **If C^c occurs, I win**

$$W_3 = -S_A P(A) - S_B P(B) - S_C P(C).$$

These bets results in a system of linear equations:

$$\begin{bmatrix} 1-P(A) & -P(B) & 1-P(C) \\ -P(A) & 1-P(B) & 1-P(C) \\ -P(A) & -P(B) & -P(C) \end{bmatrix} \begin{bmatrix} S_A \\ S_B \\ S_C \end{bmatrix} = \begin{bmatrix} W_1 \\ W_2 \\ W_3 \end{bmatrix}$$

$$RS = W$$

The above equation tells me what my possible winning will be.

You will be a sure loser if I can make W strictly positive.

If R^{-1} exists, I can find $S = R^{-1}W$ for any W.

Thus, you do not want R^{-1} to exist. Or

$$\begin{aligned} \det(R) &= 0 \\ P(A) + P(B) - P(C) &= 0 \end{aligned}$$

2.4 Decision Theory

Decision making under uncertainty.

Von Neumann and Morgenstern (1947) and Savage (1954).

1. **Elements of Decision Theory**

    - **Actions**

      What the decision maker can choose to do.

    - **States**

      What the decision maker cannot control & what is uncertain.

    - **Consequences**

      What the decision maker gets given an action and a realized state.

2. **Individual's Preference Structure on Actions**

3. Savage showed that if the preferences satisfy a set of axioms, then a *mathematician* can find:

- **a utility function on the set of consequences and**
- **a probability function on the states**

such that the ordering of actions based on expected utility agrees with the ordering according to the *individual's preferences*.

Offspring of Decision Theory

1. **Microeconomic theory is derived from decision theory.**

2. **Cognitive psychologist investigate whether or not people are "rational."**

3. **A branch of statistical inference sets parameter estimation in a decision theory context:**
   - **Actions: Choose values for parameters.**
   - **States: "True" parameter values.**
   - **Consequences: Loss function that measures estimation error.**

2.5 Statistical Decision Problems

DeGroot, M (1970) *Optimal Statistical Decisions*,McGraw–Hill, New York, pages 121–149.

1. **State space:** $\Omega = \{\omega\}$.

   In statistical inference, Ω **is the parameter space.**

2. **Decision space:** $D = \{d\}$.

   In statistical inference, d **is an estimator.**

3. R **is the space of all possible rewards** r**, which depend on** d **and** ω**:** $r = \sigma(\omega, d)$.

   The statistician selects d**; "nature" selects** ω**; payoff is** r.

4. P **is a probability distribution on** Ω.

   In statistical inference, P **is the prior or posterior distribution.**

5. Expected utility:

$$E[U(d)|P] = \int_\Omega U[\sigma(\omega, d)]dP(\omega).$$

6. Choose d which maximizes $E[U(d)|P]$.

7. Instead of utility, statisticians use loss:

$$L(\omega, d) = -U[\sigma(\omega, d)].$$

Without loss of generality, $L \geq 0$.

8. Risk or expected loss:

$$\rho(P, d) = \int_\Omega L(\omega, d)dP(\omega) = E[L(W, d)] < \infty.$$

where W is the random variable with distribution P for the unknown states.

Bayes Risk and Bayes Decisions

1. **Bayes Risk $\rho^*(P)$ is the greatest lower bound for the risks for all decisions:**

$$\rho^*(P) = \inf_{d \in D} \rho(P, d).$$

2. **Any decision d^* such that its risk is equal to the Bayes risk is called a "Bayes decision against the distribution P" or "Bayes rule":**

$$\rho(P, d^*) = \rho^*(P).$$

Example

1. Two–Point Parameter Space

- **Parameter Space:** $\Omega = \{0, 1\}$**.**
- **Probability:** $P(W = 0) = 1 - p$ **and** $P(W = 1) = p$**.**
- **Decision Space:** $D = \{d : 0 \le d \le 1\}$**.**
- **Loss function:**

$$L(\omega, d) = |w - d|^{\alpha} \textbf{ where } \alpha > 0 \textbf{ is an integer.}$$

- **Risk function:**

$$\begin{aligned} \rho(P, d) &= (1 - p)L(0, d) + pL(1, d) \\ &= (1 - p)d^{\alpha} + p(1 - d)^{\alpha} \end{aligned}$$

2. If $\alpha = 1$, the loss function is absolute error, and the Bayes decision is:

$$d^* = \begin{cases} 0 & \textbf{if } p < 0.5 \\ 1 & \textbf{if } p > 0.5 \\ \textbf{any } d & \textbf{if } p = 0.5 \end{cases}$$

and the Bayes risk is:

$$\rho^*(p) = \begin{cases} p & \textbf{if } p < 0.5 \\ 1 - p & \textbf{if } p > 0.5 \\ 0.5 & \textbf{if } p = 0.5 \end{cases}$$

If $D = \{d : 0 < d \leq 1\}$ and if $p < 0.5$, then no decision is the Bayes decision against p.

3. If $\alpha > 1$, then

$$\begin{aligned}\frac{\partial \rho(p,d)}{\partial d} &= (1-p)\alpha d^{\alpha-1} + p\alpha(1-d)^{\alpha-1} = 0 \\ d^* &= \left[1 + \left(\frac{1-p}{p}\right)^{\frac{1}{\alpha-1}}\right]^{-1}\end{aligned}$$

4. For squared-error loss ($\alpha = 2$):

$$\begin{aligned}\rho(p,d) &= d^2 - 2pd + p \\ d^* &= p \\ \rho^*(p) &= p(1-p)\end{aligned}$$

Admissible Decisions

1. **A decision d^* is admissible if there does not exist a decision d such that**

$$\begin{aligned} L(\omega, d) &\leq L(\omega, d^*) \text{ for all } \omega \\ L(\omega, d) &< L(\omega, d^*) \text{ for some } \omega \end{aligned}$$

2. **If such a d did exist, you definitely would not want to use d^*.**

3. **James–Stein**

   Under squared error loss, the sample mean is admissible estimator of the population mean in one or two dimensions. It is not admissible in three or more dimensions!

Complete Class Theorem

Consider finite parameter and decision spaces.

- **If p is strictly positive, then Bayes rules are admissible.**
- **If a decision rule is admissible, then there exists a prior distribution on the parameter space such that this decision is a Bayes rule.**

Bayes Rules Rule!

Using Sample Information

1. **Collect data X. Sample space $\mathcal{X}$.**

2. **Distribution of X given parameter ω:**

$$f(x|\omega)d\nu(x).$$

3. **Prior distribution of W:**

$$p(\omega)d\mu(\omega).$$

4. **Marginal distribution of X:**

$$f(x)d\nu(x) = \left[\int_\Omega f(x|\omega)p(\omega)d\mu(\omega)\right]d\nu(x)$$

5. **Posterior distribution of W given X:**

$$p(\omega|x)d\mu(\omega) = \left[\frac{f(x|\omega)p(\omega)}{f(x)}\right]d\mu(\omega).$$

   Note that

$$f(x|\omega)p(\omega) = p(\omega|x)f(x).$$

6. **Allow decisions to depend on observed x: $d(x)$.**

7. Risk function integrates loss over both W and X:

$$\begin{aligned}\rho(P,d) &= E\{L[W,d(X)]\}\\ &= \int_\Omega \left[\int_\mathcal{X} L[\omega,d(x)]f(x|\omega)d\nu(x)\right]p(\omega)d\mu(\omega).\end{aligned}$$

8. Interchange the order of integration:

$$\begin{aligned}\rho(P,d) &= \int_\Omega \left[\int_\mathcal{X} L[\omega,d(x)]f(x|\omega)d\nu(x)\right]p(\omega)d\mu(\omega)\\ &= \int_\mathcal{X} \left[\int_\Omega L[\omega,d(x)]p(\omega|x)d\mu(\omega)\right]f(x)d\nu(x)\\ &= \int_\mathcal{X} \left[\rho(P,d|x)\right]f(x)d\nu(x)\end{aligned}$$

9. $\rho(P,d|x) = E[L(W,d|x)]$ is the posterior risk of $d(X)$ or posterior expected loss.

10. Bayes Risk and Posterior Bayes Risk:

$$\inf_{d\in D}\rho(P,d) \;\geq\; \int_{\mathcal{X}} \underbrace{\left[\inf_{d\in D}\rho(P,d|x)\right]}_{\rho^*(P|x)} f(x)d\nu(x)$$

11. $\rho^*(P|x)$ **is the posterior Bayes risk against** P **given** X**.**

12. $d^*(X)$ **is the Bayes decision against** P **given** X **if:**

$$\rho(P,d^*(x)|x) = \rho^*(P|x).$$

Examples

1. Squared-error Loss:

$$
\begin{aligned}
L[\omega, d(x)] &= [\omega - d(x)]^2 \\
\rho(P, d(x)|x) &= \int_\Omega [\omega - d(x)]^2 p(\omega|x) d\mu(\omega) \\
&= E\left\{[\omega - d(x)]^2 |X\right\} \\
\frac{\partial \rho(P, d(x)|x)}{\partial d(x)} &= -2 \int_\Omega [\omega - d(x)] p(\omega|x) d\mu(\omega) = 0 \\
d^*(x) &= \int_\Omega \omega p(\omega|x) d\mu(\omega) = E(W|X)
\end{aligned}
$$

The posterior mean of W is posterior Bayes decision with respect to squared error loss. The posterior variance of W is the posterior Bayes risk.

2. Absolute-error Loss:

$$
\begin{aligned}
L[\omega, d(x)] &= |\omega - d(x)| \\
\rho(P, d(x)|x) &= \int_\Omega |\omega - d(x)| p(\omega|x) d\mu(\omega) \\
&= \int_{\omega<d} [d(x) - \omega] p(\omega|x) d\mu(\omega) \\
&+ \int_{\omega \geq d} [\omega - d(x)] p(\omega|x) d\mu(\omega) \\
\frac{\partial \rho(P, d(x)|x)}{\partial d(x)} &= P(W < d|x) - P(W \geq d|x) = 0 \\
P(W < d|x) &= 0.5
\end{aligned}
$$

The posterior median of W is the posterior Bayes decision with respect to absolute error loss.

3. Finite parameter and decision space.

- **Finite parameter space:**

  $\Omega = \{\omega_j, \text{ for } j = 1, \ldots, J\}$.

- **Decision space:** d_j **means select** ω_j.

- **Loss function:**

$$L(w_j, d_k) = \begin{cases} 0 & \textbf{if } j = k \\ c_{j,k} > 0 & \textbf{if } j \neq k \end{cases}$$

- **Prior probabilities:** $p_j = P(W = \omega_j)$.

- **Posterior probabilities:**

$$p_j(x) = P(W = \omega_j | x) = \frac{f(x|\omega_j)p_j}{f(x)}.$$

- **Posterior risk:**

$$\rho(P, d_k|x) = \sum_{j \neq k}^{J} c_{j,k} p_j(x).$$

- **Bayes decision rule:**

  Select w_i**, that is** $d^* = d_i$ **if**

$$\rho(P, d_i|x) \leq \rho(P, d_k|x) \text{ for all } k.$$

- **If the misclassification costs are equal:**

$$c_{j,k} = c > 0 \text{ for all } j \neq k,$$

  then the posterior risk is:

$$\rho(P, d_k|x) = c \sum_{j \neq k}^{J} p_j(x) = c[1 - p_k(x)].$$

  Select w_i **or** $d^* = d_i$ **if**

$$1 - p_i(x) \leq 1 - p_k(x) \text{ for all } k$$

  or $p_i(x) \geq p_k(x)$ **for all** k**.**

- **For equal misclassification costs, the Bayes rule is to select the parameter with maximal posterior probability.**

- **If the costs are equal and if the each parameter is equally likely: $p_i = p_j$, then the Bayes rule selects ω_i if $f(x|\omega_i) \geq f(x|\omega_k)$ for all k.**

- **Applications:**
  - **Bayesian model selection**

    Kass, R. E., and Raftery, A. E. (1995). Bayes Factors. *Journal of the American Statistical Association*, 90, 773–795.
  - **Discriminate analysis**

2.6 Large Sample Theory

Berger (1985), *Statistical Decision Theory and Bayesian Analysis*, Springer–Verlag, New York, page 225.

Assume:

1. **$\{X_i\}$ are i.i.d. given ω with density**

$$f(x|\omega) = \prod_{i=1}^{n} f(x_i|\omega).$$

2. **Prior: p.**

3. **Posterior:**

$$p_n(\omega|x) \propto \prod_{i=1}^{n} f(x_i|\omega)p(\omega).$$

4. **f and p are positive and twice differentiable near the maximum likelihood estimate $\hat{\omega}$ of ω.**

Then for large sample sizes n the posterior density p_n of ω can be approximated in the following four ways, in order of decreasing accuracy:

1. $p_n \approx N(\mu(x), V(x))$ **where** $\mu(x)$ **and** $V(x)$ **are the posterior mean and covariance matrix of** ω **given** x**.**

2. $p_n \approx N(\hat{\omega}^p, [I^p(x)]^{-1})$ **where**

$$\begin{aligned} \hat{\omega}^p &= \arg\max_{\omega} f(x|\omega)p(\omega) \\ I^p_{i,j}(x) &= -\left\{ \frac{\partial^2}{\partial\omega_i \partial\omega_j} \log[f(x|\omega)p(\omega)] \right\}_{\omega=\hat{\omega}^p} \end{aligned}$$

3. $p_n \approx N(\hat{\omega}, [\hat{I}(x)]^{-1})$ **where** $\hat{I}(x)$ **is the observed Fisher's information having** (i, j) **element:**

$$\hat{I}_{i,j}(x) = -\left\{\frac{\partial^2}{\partial\omega_i\partial\omega_j}\log[f(x|\omega)]\right\}_{\omega=\hat{\omega}}$$

4. $p_n \approx N(\hat{\omega}, [I(\hat{\omega})]^{-1})$ **where** $\hat{\omega}$ **is the maximum likelihood estimator of** ω**, and** $I(\omega)$ **is the expected Fisher's information matrix with** (i, j) **element:**

$$I_{i,j} = -nE_{X_1|\omega}\left\{\frac{\partial^2}{\partial\omega_i\partial\omega_j}\log[f(X_1|\omega)]\right\}$$

2.7 Summary

1. **Subjective Probability**

2. **Coherence**

   You can't lose, for sure.

3. **Decision Theory**

   Includes sample information, prior information, and costs.

4. **Complete Class Theorem**

   Bayes decisions are good decisions.

5. **Large Sample Theory**

   Truth is revealed.

Chapter 3

Bayesian Inference

Outline

1. **Objectives**

2. **Not So Simple Probability**

3. **Basically Bayes**

4. **Common Distributions**

5. **Beta–Binomial Model**

6. **Normal–Normal–Inverted Gamma Model**

7. **Conjugate Normal Regression**

3.1 Objectives

1. This chapter presents the "bare–bones" of Bayesian inference that we will need in later chapters.

2. After fixing notation and ideas, we will look at the three simplest models:

   (a) Beta–Binomial for 0/1 outcomes,

   (b) Conjugate Normal for continuous outcomes,

   (c) Conjugate Linear Regression.

3.2 Why Bayes?

1. **It provides a unified method for evaluating risk, making decisions under uncertainty, and updating beliefs in the light of new information.**

2. **Given that the model holds, it optimally uses information and accounts for all sources of uncertainty.**

3. **Bayes estimators have many attractive frequentist properties.**

4. **Bayes Rules! It can't be beat!**

3.3 Not So Simple Probability

1. **A random variable X has probability mass function (pmf) or probability density function (pdf) $[x]$ with:**

$$
\begin{aligned}
[x] &\geq 0 \\
\sum_x [x] &= 1 \textbf{ if } X \textbf{ is discrete} \\
\int_x [x]\, dx &= 1 \textbf{ if } X \textbf{ is continuous.}
\end{aligned}
$$

   I will use "$\int$" for "$\sum$" when X is discrete. I will not be consistent. I will call $[x]$ the "distribution of X."

2. **The probability that X is in set A is:**

$$P(X \in A) = \int_A [x]\, dx.$$

3. **$[x, y]$ is the joint pmf or pdf of two random variables X and Y.**

4. The marginal distribution of X is:

$$[x] = \int_y [x, y]\, dy.$$

5. The conditional distribution of Y given X is:

$$[y|x] = \frac{[x, y]}{\int_y [x, y]\, dy} = \frac{[x, y]}{[x]}.$$

Note:

$$[x, y] = [y|x][x] = [x|y][y].$$

6. Total Probability:

$$[x] = \int_y [x|y][y]\, dy.$$

Check:

$$\int_y [x|y][y]\, dy = \int_y \frac{[x, y]}{[y]}[y]\, dy = \int_y [x, y]\, dy.$$

7. Bayes Theorem:

$$[y|x] = \frac{[x|y][y]}{\int_y [x|y][y]\, dy} \propto [x|y][y].$$

Check:

$$[y|x] = \frac{[x, y]}{[x]} = \frac{[x|y][y]}{\int_y [x|y][y]\, dy}.$$

8. X **and** Y **are independent if:**

$$[x, y] = [x][y] \textbf{ or } [y|x] = [y] \textbf{ or } [x|y] = [x].$$

9. X **and** Y **are independent given** Z **if:**

$$[x, y|z] = [x|z][y|z] \textbf{ or } [y|x, z] = [y|z] \textbf{ or } [x|y, z] = [x|z].$$

Check:

$$[y|x, z] = \frac{[x, y, z]}{[x, z]} = \frac{[x, y|z][z]}{[x, z]} = \frac{[x|z][y|z][z]}{[x|z][z]}.$$

10. If X **and** Y **are independent given** Z**, then**

$$[x, y] = \int_z [x, y|z][z]\, dz = \int_z [x|z][y|z][z]\, dz.$$

Also,

$$[y|x] = \frac{[x, y]}{[x]} = \frac{\int_z [x|z][y|z][z]\, dz}{[x]} = \int_z [y|z][z|x]\, dz.$$

3.4 Basically Bayes

1. **Conditional distribution of the data given parameters:**

   - **Given the unknown parameter θ, the distribution of the data X_1, X_2, ..., X_n is:**

     $$[x_1, \ldots, x_n|\theta].$$

     θ and x_i may be multidimensional.

   - **A useful special case is when the observations are mutually independent given θ:**

     $$[x_1, \ldots, x_n|\theta] = \prod_{i=1}^{n} [x_i|\theta].$$

2. Likelihood Function of θ

$$L(\theta) = [x_1, \ldots, x_n|\theta].$$

3. Prior Distribution of θ is $[\theta]$.

4. Joint distribution of the data and θ is:

$$[x_1, \ldots, x_n, \theta] = [x_1, \ldots, x_n|\theta][\theta].$$

5. Marginal distribution of the data:

$$[x_1, \ldots, x_n] = \int_\theta [x_1, \ldots, x_n|\theta][\theta]\, d\theta$$

6. Posterior distribution of θ

$$
\begin{aligned}
[\theta|x_1, \ldots, x_n] &= \frac{[x_1, \ldots, x_n|\theta][\theta]}{[x_1, \ldots, x_n]} \\
&\propto [x_1, \ldots, x_n|\theta][\theta] \\
&\propto L(\theta)[\theta]
\end{aligned}
$$

7. Bayesian inference about θ is based on $[\theta|x_1, \ldots, x_n]$:

- **Posterior Mean $\Leftrightarrow$ Squared–Error Loss**
  Posterior Variance or Standard Deviation
- **Posterior Median $\Leftrightarrow$ Absolute–Error Loss**
  Posterior Absolute Error
- **Highest Posterior Density Intervals**

8. Predictive Distribution

- **Suppose that we observe x_1, . . . , x_n and that we want to describe the likely vales of future observations $X_{n+1}, \ldots, X_{n+m}$.**
- **The joint pdf or pmf for $X_1, \ldots, X_{n+m}$ is**

$$[x_1, \ldots, x_{n+m}].$$

- **The predictive pmf or pdf of $X_{n+1}, \ldots, X_{n+m}$ given the data x_1, . . . , x_n is:**

$$[x_{n+1}, \ldots, x_{n+m}|x_1, \ldots, x_n] = \frac{[x_1, \ldots, x_{n+m}]}{[x_1, \ldots, x_n]}.$$

- **If X_{n+1}, … X_{n+m} are independent of X_1, …, X_n given θ:**

$$[x_1, \ldots, x_{n+m}|\theta] = [x_1, \ldots, x_n|\theta][x_{n+1}, \ldots, x_{n+m}|\theta],$$

then the predictive pmf or pdf is:

$$[x_{n+1}, \ldots, x_{n+m}|x_1, \ldots, x_n] = \frac{[x_1, \ldots, x_{n+m}]}{[x_1, \ldots, x_n]}$$

$$= \frac{\int [x_1, \ldots, x_{n+m}|\theta][\theta]\, d\theta}{[x_1, \ldots, x_n]}$$

$$= \frac{\int [x_1, \ldots, x_n|\theta][x_{n+1}, \ldots, x_{n+m}|\theta][\theta]\, d\theta}{[x_1, \ldots, x_n]}$$

$$= \int [x_{n+1}, \ldots, x_{n+m}|\theta][\theta|x_1, \ldots, x_n]\, d\theta$$

- **Compare to the marginal pdf of $X_{n+1}, \ldots, X_{n+m}$:**

$$[x_{n+1}, \ldots, x_{n+m}] = \int [x_{n+1}, \ldots, x_{n+m}|\theta][\theta]\, d\theta.$$

3.5 Binomial and Beta Distributions

3.5.1 Binomial Distribution

1. **X has a binomial distribution with parameters θ and n if its pmf is:**

$$\begin{aligned}[x|\theta, n] &= B(x|\theta, n) \\ &= \begin{pmatrix} n \\ x \end{pmatrix} \theta^x (1-\theta)^{n-x}\end{aligned}$$

for $x = 0, 1, \ldots, n;$

$0 \leq \theta \leq 1;$ **and integer** $n > 0.$

2. **Moments:**

$$E(X|\theta, n) = \theta \textbf{ and } V(X|\theta, n) = n\theta(1-\theta).$$

3.5.2 Beta Distribution

1. **θ has a beta distribution with parameters α and β if its pdf is:**

$$
\begin{aligned}
[\theta] &= Beta(\theta|\alpha,\beta) \\
&= \frac{\Gamma(\alpha+\beta)}{\Gamma(\alpha)\Gamma(\beta)}\theta^{\alpha-1}(1-\theta)^{\beta-1}
\end{aligned}
$$

$$\textbf{for } 0 \le \theta \le 1; \quad \alpha > 0; \textbf{ and } \beta > 0$$

$$\Gamma(x) = \int_0^\infty y^{x-1}\exp(-y)\,dy$$

$$
\begin{aligned}
\Gamma(x+1) &= x\Gamma(x) \textbf{ and } \Gamma(n+1) = n! \textbf{ if } n \textbf{ is an integer} \\
\Gamma(0) &= 1;\ \Gamma(0.5) = \sqrt{\pi};\ \textbf{and } \Gamma(1) = 1
\end{aligned}
$$

$$\frac{\Gamma(\alpha)\Gamma(\beta)}{\Gamma(\alpha+\beta)} = \int_0^1 x^{\alpha-1}(1-x)^{\beta-1}\,dx$$

2. Moments:

$$
\begin{aligned}
E[\theta^u(1-\theta)^v] &= \frac{\Gamma(\alpha+\beta)}{\Gamma(\alpha)\Gamma(\beta)} \int_0^1 \theta^{u+\alpha-1}(1-\theta)^{v+\beta-1}\, d\theta \\
&= \left[\frac{\Gamma(\alpha+\beta)}{\Gamma(\alpha)\Gamma(\beta)}\right] \left[\frac{\Gamma(\alpha+u)\Gamma(\beta+v)}{\Gamma(\alpha+\beta+u+v)}\right]
\end{aligned}
$$

for $u > -\alpha$ **and** $v > -\beta$

$$
\begin{aligned}
E(\theta) &= \frac{\alpha}{\alpha+\beta} \textbf{ and } E(1-\theta) = \frac{\beta}{\alpha+\beta} \\
V(\theta) &= \frac{\alpha\beta}{(\alpha+\beta)^2(\alpha+\beta+1)} \\
&= \frac{E(\theta)E(1-\theta)}{\alpha+\beta+1}
\end{aligned}
$$

3.6 Beta–Binomial Model

1. Model

- **Given** θ **the observations** $X_1, \ldots, X_m$ **are mutually independent with** $B(x|\theta, 1)$ **pmf:**

$$[x|\theta] = \theta^x(1-\theta)^{1-x}$$

for $x = \mathbf{0}$ **or 1, and** $0 \le \theta \le 1$.

- **The conjugate prior distribution for** θ **is the beta distribution** $Beta(\theta|\alpha_0, \beta_0)$ **with pdf:**

$$[\theta] = \frac{\Gamma(\alpha_0 + \beta_0)}{\Gamma(\alpha_0)\Gamma(\beta_0)}\theta^{\alpha_0 - 1}(1-\theta)^{\beta_0 - 1}$$

2. The joint pmf of X_1, . . . , X_n given θ is:

$$[x_1, \ldots, x_n|\theta] = \theta^s(1-\theta)^{n-s}$$

where $s = \sum_{i=1}^n x_i$ is the number of ones in n trials.

3. The marginal distribution of X_1, . . . , X_n has pmf:

$$[x_1, \ldots, x_n] = \int \prod_{i=1}^{n} [x_i|\theta][\theta]\, d\theta$$

$$= \frac{\Gamma(\alpha_0+\beta_0)}{\Gamma(\alpha_0)\Gamma(\beta_0)} \int_0^1 \theta^{\alpha_0+\sum_{i=1}^n x_i}(1-\theta)^{\beta_0+n-\sum_{i=1}^n x_i}\, d\theta$$

$$= \left[\frac{\Gamma\left(\alpha_0+\beta_0\right)}{\Gamma\left(\alpha_0\right)\Gamma\left(\beta_0\right)}\right]\left[\frac{\Gamma\left(\alpha_0+\sum_{i=1}^n x_i\right)\Gamma\left(\beta_0+n-\sum_{i=1}^n x_i\right)}{\Gamma\left(\alpha_0+\beta_0+n\right)}\right]$$

Define

$$\begin{aligned}\alpha_n &= \alpha_0 + \sum_{i=1}^{n} x_i \\ \beta_n &= \beta_0 + n - \sum_{i=1}^{n} x_i\end{aligned}$$

4. Define $S = X_1 + \cdots + X_n$.

S **given** θ **is** $B(s|\theta, n)$**.**

The marginal distribution of S **is the Beta–Binomial distribution with pmf:**

$$[s] = BB(s|n, \alpha_0, \beta_0)$$

$$= \binom{n}{s} \left[\frac{\Gamma(\alpha_0 + \beta_0)}{\Gamma(\alpha_0)\,\Gamma(\beta_0)}\right] \left[\frac{\Gamma(\alpha_0 + s)\Gamma(\beta_0 + n - s)}{\Gamma(\alpha_0 + \beta_0 + n)}\right]$$

for $s = \mathbf{0}, \ldots, n$

Moments of S:

$$
\begin{aligned}
E(S) &= E_\theta[E(S|\theta)] \\
&= E_\theta(n\theta) \\
&= n\frac{\alpha_0}{\alpha_0+\beta_0}
\end{aligned}
$$

$$
\begin{aligned}
V(S) &= E_\theta(V(S|\theta)) + V_\theta(E(S|\theta)) \\
&= E_\theta(n\theta(1-\theta)) + V_\theta(n\theta) \\
&= n\frac{\alpha_0\beta_0}{(\alpha_0+\beta_0)(\alpha_0+\beta_0+1)} \\
&+ n^2\frac{E(\theta)[1-E(\theta)]}{\alpha_0+\beta_0+1} \\
&= nE(\theta)[1-E(\theta)]\left(\frac{\alpha_0+\beta_0+n}{\alpha_0+\beta_0+1}\right)
\end{aligned}
$$

"Extra Binomial Variation"

5. The posterior distribution of θ given S has pdf:

$$
\begin{aligned}
[\theta|x_1, \ldots, x_n] &\propto L(\theta)[\theta] \\
&\propto \theta^s(1-\theta)^{n-s}\theta^{\alpha_0-1}(1-\theta)^{\beta_0-1} \\
&= \frac{\Gamma(\alpha_n+\beta_n)}{\Gamma(\alpha_n)\Gamma(\beta_n)}\theta^{\alpha_n-1}(1-\theta)^{\beta_n-1} \\
&= Beta(\theta|\alpha_n, \beta_n) \\
\alpha_n &= \alpha_0 + s \\
\beta_n &= \beta_0 + n - s.
\end{aligned}
$$

Updating Parameters

Prior		**Posterior**
α_0	$\Rightarrow$	$\alpha_n = \alpha_0 + s$
β_0	$\Rightarrow$	$\beta_n = \beta_0 + n - s$

6. Squared Error Loss Estimator of θ:

$$\begin{aligned} E(\theta|s) &= \frac{\alpha_n}{\alpha_n + \beta_n} \\ &= \frac{\alpha_0 + s}{\alpha_0 + \beta_0 + n} \\ &= w\hat{\theta} + (1-w)E(\theta) \end{aligned}$$

$$\hat{\theta} = \frac{s}{n} \textbf{ and } w = \frac{n}{\alpha_0 + \beta_0 + n}.$$

The Bayes estimator is a convex combination of the MLE of θ and its prior mean. It "shrinks" the MLE towards the prior mean.

7. Posterior variance:

$$V(\theta|s) = \frac{E(\theta|s)[1 - E(\theta|s)]}{n + \alpha_0 + \beta_0 + 1}$$

8. The predictive distribution of X_{n+1}, ..., X_{n+m} given x_1, ..., x_n has pmf:

$$
\begin{aligned}
&[x_{n+1}, \ldots, x_{n+m} | x_1, \ldots, x_n] \\
&= \int_0^1 [x_{n+1}, \ldots, x_{n+m} | \theta][\theta | x_1, \ldots, x_n] \, d\theta \\
&= \int_0^1 \theta^{\sum_{i=n+1}^{n+m} x_i} (1-\theta)^{m - \sum_{i=n+1}^{n+m} x_i} \\
&\times \frac{\Gamma(\alpha_n + \beta_n)}{\Gamma(\alpha_n)\Gamma(\beta_n)} \theta^{\alpha_n - 1} (1-\theta)^{\beta_n - 1} \; d\theta \\
&= \left[\frac{\Gamma(\alpha_n + \beta_n)}{\Gamma(\alpha_n)\Gamma(\beta_n)}\right] \left[\frac{\Gamma\left(\alpha_n + \sum_{i=n+1}^{n+m} x_i\right)\Gamma\left(\beta_n + m - \sum_{i=n+1}^{n+m} x_i\right)}{\Gamma(\alpha_n + \beta_n + m)}\right]
\end{aligned}
$$

9. Define $T = X_{n+1} + \cdots + X_{n+m}$. The predictive distribution of T given S is $BB(t|m, \alpha_n, \beta_n)$.

3.7 Normal, Gamma, and T Distributions

3.7.1 Normal Distribution

1. **A random variable X has a normal distribution with mean μ, standard deviation σ, and pdf:**

$$[x|\mu,\sigma] = N(x|\mu,\sigma^2)$$

$$= \frac{1}{\sqrt{2\pi\sigma^2}}\exp\left[-\frac{(x-\mu)^2}{2\sigma^2}\right] \textbf{ for } -\infty < x < \infty.$$

2. **In GAUSS, to generate a $n \times m$ matrix of independent, normal random variables:**

$$X = mean + sigma * \mathbf{rndn}(n,m);$$

3.7.2 Multivariate Normal Distribution

1. **If a m dimensional random vector X has a multivariate normal distribution with mean vector μ and $m \times m$ positive definite covariance matrix Σ, then its density is given by:**

$$\begin{aligned}[x|\mu,\Sigma] &= N_m(x|\mu,\Sigma) \\ &= (2\pi)^{-\frac{m}{2}}|\Sigma|^{-\frac{1}{2}}\exp\left[-\frac{1}{2}(x-\mu)'\Sigma^{-1}(x-\mu)\right]\end{aligned}$$

2. **Mean and variance (covariance):**

$$\begin{aligned}E(X) &= \mu \\ V(X) &= E[(X-\mu)(X-\mu)'] = \Sigma.\end{aligned}$$

3. **Linear Functions:**

   If $Y = AX + b$ where A is a $n \times m$ matrix of rank $n \leq m$ and b is a n vector, then

$$[y] = N_n(y|A\mu + b, A\Sigma A').$$

4. Conditional normals:

$$[Y] = \begin{bmatrix} Y_1 \\ Y_2 \end{bmatrix} = N_M\left(Y \middle| \begin{bmatrix} \mu_1 \\ \mu_2 \end{bmatrix}, \begin{bmatrix} \Sigma_{11} & \Sigma_{12} \\ \Sigma_{21} & \Sigma_{22} \end{bmatrix}\right).$$

Then

$$[Y_2|Y_1] = N(Y_2|\mu_{2|1}, \Sigma_{2|1})$$

$$\mu_{2|1} = \mu_2 - \Sigma_{21}\Sigma_{11}^{-1}(Y_1 - \mu_1)$$

$$\Sigma_{2|1} = \Sigma_{22} - \Sigma_{21}\Sigma_{11}^{-1}\Sigma_{12}.$$

5. Generating Multivariate Normals:

Let C be the Cholesky decomposition of Σ.

C is an upper triangular matrix such that

$$C'C = \Sigma.$$

If Z is $N_m(z|0, I)$ where I is the identity matrix, then

$$X = \mu + C'Z$$

is $N_m(x|\mu, \Sigma)$.

In GAUSS,

$$\begin{aligned} C &= \mathbf{chol}(\Sigma); \\ X &= \mu + C'\mathbf{rndn}(m, 1); \end{aligned}$$

where rndn(r,c) returns a $r \times c$ matrix of independent, standard normal random variates.

3.7.3 Gamma Distribution

1. **A random variable X has a gamma distribution with pdf:**

$$\begin{aligned}
[x|\alpha,\beta] &= G(x|\alpha,\beta) \\
&= \frac{\beta^\alpha}{\Gamma(\alpha)} x^{\alpha-1}\exp(-\beta x) \\
&\quad \textbf{for } x > 0;\ \alpha > 0;\ \textbf{and } \beta > 0.
\end{aligned}$$

2. **The moments of a gamma distribution are:**

$$\begin{aligned}
E(X^k) &= \frac{\beta^\alpha}{\Gamma(\alpha)} \int_0^\infty x^{k+\alpha-1}\exp(-\beta x)\,dx \\
&= \frac{\beta^\alpha}{\Gamma(\alpha)} \frac{\Gamma(k+\alpha)}{\beta^{k+\alpha}} \\
&= \frac{\Gamma(\alpha+k)}{\Gamma(\alpha)\beta^k} \textbf{ for } k > -\alpha \\
E(X) &= \frac{\alpha}{\beta} \textbf{ and } V(X) = \frac{\alpha}{\beta^2} = E(X)\frac{1}{\beta}
\end{aligned}$$

3.7.4 Inverted Gamma Distribution

1. **Define $Y = 1/X$. Then Y has an Inverted Gamma distribution with density:**

$$
\begin{aligned}
[y|\alpha, \beta] &= IG(y|\alpha, \beta) \\
&= \frac{\beta^\alpha}{\Gamma(\alpha)} y^{-(\alpha+1)} \exp(-\beta/y) \textbf{ for } y > 0
\end{aligned}
$$

2. **Moments:**

$$
\begin{aligned}
E(Y^k) &= E(X^{-k}) = \beta^k \frac{\Gamma(\alpha - k)}{\Gamma(\alpha)} \textbf{ for } k < \alpha \\
E(Y) &= \frac{\beta}{\alpha - 1} \\
V(Y) &= \frac{\beta^2}{(\alpha-1)^2(\alpha-2)} = E(Y)^2 \frac{1}{\alpha - 2}
\end{aligned}
$$

Generating Gamma Random Deviates in GAUSS

1. If X is $G(x|\alpha,\beta)$, then $Y = cX$ is $G(y|\alpha,\beta/c)$.

2. $r \times c$ matrix of independent $G(x|\alpha,\beta)$:

$$X = \textbf{rndgam}(r, c, \alpha)/\beta;$$

3. $r \times c$ matrix of independent $IG(y|\alpha,\beta)$:

$$Y = \beta/\textbf{rndgam}(r, c, \alpha);$$

3.7.5 T–Distribution

Suppose that:

$$[x|m, w\sigma^2] \;=\; N\left(x|m, w\sigma^2\right) \textbf{ and } [\sigma^2] = IG\left(\sigma^2|\frac{r}{2}, \frac{s}{2}\right)$$

Marginal pdf of x has a T Distribution:

$$
\begin{aligned}
[x|m, w, r, s] &= \int_0^\infty [x|m, w\sigma^2][\sigma^2]\, d\sigma^2 \\
&= \left[\frac{1}{2\pi w}\right]^{\frac{1}{2}} \left[\frac{\left(\frac{s}{2}\right)^{\frac{r}{2}}}{\Gamma\left(\frac{r}{2}\right)}\right] \\
&\times \int_0^\infty \left(\sigma^2\right)^{-(r+3)/2} \exp\left\{-\left[\frac{(x-m)^2}{2w} + \frac{s}{2}\right]\frac{1}{\sigma^2}\right\} d\sigma^2 \\
&= \left[\frac{1}{2\pi w}\right]^{\frac{1}{2}} \left[\frac{\left(\frac{s}{2}\right)^{\frac{r}{2}}}{\Gamma\left(\frac{r}{2}\right)}\right] \left[\frac{\Gamma\left(\frac{r+1}{2}\right)}{\left[\frac{s}{2} + \frac{(x-m)^2}{2w}\right]^{(r+1)/2}}\right] \\
&= \left[\frac{1}{\pi s w}\right]^{\frac{1}{2}} \left[\frac{\Gamma\left(\frac{r+1}{2}\right)}{\Gamma\left(\frac{r}{2}\right)}\right] \left[1 + \frac{(x-m)^2}{sw}\right]^{-(r+1)/2} \\
&= T(x|m, w, r, s)
\end{aligned}
$$

3.7.6 Multivariate T–Distribution

Suppose that:

$$[x|m, W\sigma^2] \;=\; N_p(x|m, W\sigma^2) \textbf{ and } [\sigma^2] = IG\left(\sigma^2|\frac{r}{2}, \frac{s}{2}\right)$$

Integrate out σ^2:

$$[x|m, W, r, s] = \int_0^\infty [x|m, W\sigma^2][\sigma^2]\, d\sigma^2$$

$$= \left[\frac{1}{2\pi}\right]^{p/2} |W|^{-\frac{1}{2}} \left[\frac{\left(\frac{s}{2}\right)^{\frac{r}{2}}}{\Gamma\left(\frac{r}{2}\right)}\right]$$

$$\times \int_0^\infty \left(\sigma^2\right)^{-\left(\frac{r+p}{2}+1\right)} \exp\left\{-\left[\frac{(x-m)'W^{-1}(x-m)+s}{2}\right]\frac{1}{\sigma^2}\right\} d\sigma^2$$

$$= (\pi s)^{-\frac{p}{2}} |W|^{-\frac{1}{2}} \left[\frac{\Gamma\left(\frac{r+p}{2}\right)}{\Gamma\left(\frac{r}{2}\right)}\right]$$

$$\times \left[1 + \frac{1}{s}(x-m)'W^{-1}(x-m)\right]^{-\left(\frac{r+p}{2}\right)}$$

$$= T_p(x|m, W, r, s)$$

3.8 Normal–Normal–Inverted Gamma Model

1. **Conjugate Model:**

   - **Given μ and σ, the data X_1, X_2, ... are mutually independent from $N\left(x|\mu,\sigma^2\right)$.**

   - **μ given σ is $N\left(\mu|m_0,\frac{\sigma^2}{w_0}\right)$.**

   - **σ^2 is $IG\left(\sigma^2|\frac{r_0}{2},\frac{s_0}{2}\right)$.**

   Moments:

$$E(\sigma^2) = \frac{s_0}{r_0-2} \textbf{ and } V(\sigma^2) = \left(\frac{s_0}{r_0-2}\right)^2\left(\frac{2}{r_0-4}\right)$$

$$E(\mu) = m_0 \textbf{ and } V(\mu) = E(\sigma^2)/w_0$$

$$E(X) = m_0 \textbf{ and } V(X) = E(\sigma^2) + E(\sigma^2)/w_0$$

Prior for μ is scale invariant.

Suppose $Y = aX$ for some scalar a.

Define $\mu_Y = a\mu$, $m_{0,Y} = am_0$, and $\sigma_Y = |a|\sigma$. Then

$$\begin{aligned}
[y|\mu, \sigma^2] &= N(y|a\mu, a^2\sigma^2) \\
&= N(y|\mu_Y, \sigma_Y^2) \\
[\mu_Y|\sigma] &= N(\mu_Y|am_0, a^2\sigma^2/w_0) \\
&= N(\mu_Y|m_{0,Y}, \sigma_Y/w_0) \\
[\sigma_Y^2] &= IG(\sigma_Y^2|r_0/2, s_0/(2a^2))
\end{aligned}$$

2. Setting Prior Parameters.

- **Specify:**

$$e_0 = E(\sigma^2) \textbf{ and } v_0 = V(\sigma^2).$$

- **Solve for r_0 and s_0:**

$$\begin{aligned}
e_0 &= \frac{s_0}{r_0 - 2} \\
v_0 &= e_0^2 \left[\frac{2}{r_0 - 4}\right] \\
s_0 &= e_0[r_0 - 2]
\end{aligned}$$

$$\begin{aligned}
r_0 &= 2\frac{e_0^2}{v_0} + 2 \\
s_0 &= 2e_0 \left[\frac{e_0^2}{v_0} + 1\right]
\end{aligned}$$

- **m_0 is your prior guess at the mean of X.**
- **w_0 expresses your uncertainty about the mean of X. Small w_0 corresponds to large uncertainty, and large w_0 corresponds to high confidence. w_0 is called the "precision."**

3. Marginal Distribution of X:

$$[x] \;=\; \int_0^\infty \int_{-\infty}^\infty [x|\mu, \sigma^2][\mu|\sigma^2][\sigma^2]\, d\mu\, d\sigma^2$$

Integrate out μ:

$$[x|\sigma^2] = N\left(x|m_0, \left[1 + w_0^{-1}\right]\sigma^2\right).$$

Integrate out σ^2. Set w on page (75) to $1 + w_0^{-1}$.

$$\begin{aligned} [x] \;&=\; \int_0^\infty [x|\sigma^2][\sigma^2]\, d\sigma^2 \\ &=\; T\left(x|m_0, 1 + w_0^{-1}, r_0, s_0\right) \end{aligned}$$

4. Joint Distribution:

$$[x_1, \ldots, x_n, \mu, \sigma^2] = \prod_{i=1}^{n} [x_i|\mu, \sigma^2][\mu|\sigma^2][\sigma^2]$$

$$= \prod_{i=1}^{n} N\left(x_i|\mu, \sigma^2\right) N\left(\mu|m_0, \frac{\sigma^2}{w_0}\right) IG\left(\sigma^2|\frac{r_0}{2}, \frac{s_0}{2}\right)$$

$$= (2\pi)^{-(n+1)/2}\sqrt{w_0}\frac{\left(\frac{s_0}{2}\right)^{\frac{r_0}{2}}}{\Gamma\left(\frac{r_0}{2}\right)}\left(\sigma^2\right)^{-(n+r_0+3)/2}$$
$$\times \exp\left\{-\frac{1}{2\sigma^2}\left[\sum_{i=1}^{n}(x_i-\mu)^2 + w_0(\mu-m_0)^2 + s_0\right]\right\}$$

$$\propto \left(\sigma^2\right)^{-(n+r_0+3)/2}\exp\left\{-\frac{1}{2\sigma^2}\left[n(\mu-\bar{x}_n)^2 \right.\right.$$
$$\left.\left. + w_0(\mu-m_0)^2 + s_0 + SSE_n\right]\right\}$$

$$\bar{x}_n = n^{-1}\sum_{i=1}^{n} x_i \textbf{ and } SSE_n = \sum_{i=1}^{n}(x_i-\bar{x}_n)^2$$

Complete the squares in μ:

$$n(\mu - \bar{x}_n)^2 + w_0(\mu - m_0)^2$$

$$= (n + w_0)\mu^2 - 2(n\bar{x} + w_0 m_0)\mu + n\bar{x}^2 + w_0 m_0^2$$

$$= (n + w_0)\left[\mu^2 - 2\mu\left(\frac{n\bar{x} + w_0 m_0}{n + w_0}\right)\right] + n\bar{x}^2 + w_0 m_0^2$$

Define $m_n = \dfrac{n\bar{x} + w_0 m_0}{n + w_0}$ **and** $w_n = n + w_0$

$$= w_n(\mu - m_n)^2 - w_n m_n^2 + n\bar{x}^2 + w_0 m_0^2$$

$$= w_n(\mu - m_n)^2 + \left(\frac{n w_0}{n + w_0}\right)(\bar{x} - m_0)^2$$

Joint Distribution:

$$
\begin{aligned}
&[x_1, \ldots, x_n, \mu, \sigma^2] \\
&\propto \left(\sigma^2\right)^{-(n+r_0+3)/2} \exp\left\{-\frac{1}{2\sigma^2}\left[n(\mu - \bar{x}_n)^2\right.\right. \\
&\qquad\qquad\left.\left. + w_0(\mu - m_0)^2 + s_0 + SSE_n\right]\right\} \\
&\propto \left(\sigma^2\right)^{-\frac{1}{2}} \exp\left[-\frac{1}{2\sigma^2} w_n(\mu - m_n)^2\right] \\
&\times \left(\sigma^2\right)^{-(r_n+2)/2} \exp\left[-\frac{s_n}{2\sigma^2}\right] \\
&\propto N\left(\mu | m_n, \frac{\sigma^2}{w_n}\right) IG\left(\sigma^2 | \frac{r_n}{2}, \frac{s_n}{2}\right)
\end{aligned}
$$

$$
\begin{aligned}
r_n &= r_0 + n \\
s_n &= s_0 + SSE_n + \left(\frac{n w_0}{n + w_0}\right)(\bar{x} - m_0)^2
\end{aligned}
$$

5. Posterior Distributions

- **σ^2 given the data is $IG\left(\sigma^2|\frac{r_n}{2}, \frac{s_n}{2}\right)$.**

- **μ given σ^2 and the data is $N\left(\mu|m_n, \frac{\sigma^2}{w_n}\right)$.**

Updating:

Prior	Posterior

$$m_0 \;\Rightarrow\; m_n = \frac{n\bar{x} + w_0 m_0}{n + w_0}$$

$$w_0 \;\Rightarrow\; w_n = w_0 + n$$

$$r_0 \;\Rightarrow\; r_n = r_0 + n$$

$$s_0 \;\Rightarrow\; s_n = s_0 + SSE_n + \left(\frac{n w_0}{n + w_0}\right)(\bar{x} - m_0)^2$$

"Non–informative" Prior:

$$m_0 = w_0 = r_0 = s_0 = 0.$$

6. Predictive Distribution

$$
\begin{aligned}
&[x|x_1, \ldots, x_n] \\
&= \int_0^\infty \int_{-\infty}^\infty [x|\mu, \sigma^2][\mu|\sigma^2, x_1, \ldots, x_n][\sigma^2|x_1, \ldots, x_n]\, d\mu\, d\sigma^2 \\
&= \int_0^\infty \int_{-\infty}^\infty N\left(x|\mu, \sigma^2\right) \\
&\quad \times N\left(\mu|m_n, \frac{\sigma^2}{w_n}\right) IG\left(\sigma^2|\frac{r_n}{2}, \frac{s_n}{2}\right)\, d\mu\, d\sigma^2 \\
&= T\left(x|m_n, 1 + w_n^{-1}, r_n, s_n\right)
\end{aligned}
$$

3.9 Conjugate Normal Regression

1. Model:

$$
\begin{aligned}
Y &= X\beta + \epsilon \\
[\epsilon|\sigma^2] &= N_n(\epsilon|0, \sigma^2 I_n) \\
[y|\beta, \sigma^2] &= N_n(y|X\beta, \sigma^2 I_n)
\end{aligned}
$$

- **Y is a n–vector of dependent observations.**
- **X is a $n \times p$ design matrix.**

  Need not be full rank.
- **ϵ is a n–vector of error terms.**

2. Conjugate Priors:

$$\begin{aligned}[\beta|u_0, V_0, \sigma^2] &= N_p(\beta|u_0, \sigma^2 V_0) \\ [\sigma^2|r_0, s_0] &= IG\left(\sigma^2|\frac{r_0}{2}, \frac{s_0}{2}\right)\end{aligned}$$

Strange? If you change the scale of Y, then the scale of σ^2 changes, and your prior beliefs about β are the same.

3. Marginal Distribution of Y:

$$[y|u_0, V_0, r_0, s_0] = T_n(y|Xu_0, I_n + XV_0X', r_0, s_0).$$

4. Posterior Distributions:

$$[\beta|Y,\sigma^2] = N_p(\beta|u_n,\sigma^2 V_n)$$

$$[\sigma^2|Y] = IG\left(\sigma^2|\frac{r_n}{2},\frac{s_n}{2}\right)$$

$$\begin{aligned}
V_n &= \left(X'X + V_0^{-1}\right)^{-1} \\
u_n &= V_n\left(X'Y + V_0^{-1}u_0\right) \\
r_n &= r_0 + n \\
s_n &= s_0 + Y'Y + u_0'V_0^{-1}u_0 - u_n'V_n^{-1}u_n
\end{aligned}$$

5. The so-called "non-informative" prior sets:

$$u_0 = 0; \quad V_0 = 0; \;\; r_0 = 0; \;\; \textbf{and} \;\; s_0 = 0.$$

You need to expand and complete the square in β using matrices:

$$(y - X\beta)'(y - X\beta) + (\beta - u_0)V_0^{-1}(\beta - u_0)$$

$$= \beta'(X'X + V_0^{-1})\beta - 2\beta'(X'y + V_0^{-1}u_0) + C_0$$

$$= \beta' V_n^{-1}\beta - 2\beta' V_n^{-1}u_n + C_0$$

where C_0 is the appropriate constant, and V_n and u_n are defined on page (88).
Then add and subtract $u_n' V_n^{-1} u_n$ to the above equation to complete the square:

$$(y - X\beta)'(y - X\beta) + (\beta - u_0)V_0^{-1}(\beta - u_0)$$

$$= (\beta - u_n)' V_n^{-1}(\beta - u_n) + C_1$$

where C_1 is the appropriate constant.

6. If X has full rank, then the MLE of β is:

$$\hat{\beta} = (X'X)^{-1}X'Y.$$

The posterior mean of β is:

$$\begin{aligned} u_n &= \left(X'X + V_0^{-1}\right)^{-1}\left(X'X\hat{\beta} + V_0^{-1}u_0\right) \\ &= W_n\hat{\beta} + (I_n - W_n)u_0 \\ W_n &= \left(X'X + V_0^{-1}\right)^{-1} X'X \end{aligned}$$

- **u_n is a convex sum of the prior mean u_0 and the MLE $\hat{\beta}$.**
- **The weights depend on the prior variance $\sigma^2 V_0$ of β and the sampling variance $\sigma^2(X'X)^{-1}$ of $\hat{\beta}$.**
- **Under weak conditions W_n approaches I_n as n becomes large.**

7. Predictive distribution of $Y_f = X_f\beta + \epsilon$ where Y_f is m–vector:

$$[y_f|y, u_n, V_n, r_n, s_n] = T_m(y_f|X_f u_n, I_m + X_f V_n X_f', r_n, s_n).$$

8. Model Selection

- **Bayesian model selection is based on the decision theoretic development on page (37).**
- **Let ω_j indicate the model with design matrix X_j for $j = 1, \ldots, J$.**
- **X_j is a $n \times p_j$ design matrix.**
- **Priors for model j:**

$$
\begin{aligned}
[\beta|\sigma^2, \omega_j] &= N_{p_j}(\beta|u_{0,j}, V_{0,j}) \\
[\sigma^2|\omega_j] &= IG\left(\sigma^2|r_{0,j}/1, s_{0,j}/2\right) \\
q_j &= P(\omega_j).
\end{aligned}
$$

- **Posterior probability of model j:**

$$q_j(y) = [\omega_j|y] = \frac{[y|\omega_j]q_j}{\sum_{k=1}^{J}[y|\omega_k]q_k}$$

$$[y|\omega_j] = T_n(y|X_j u_{0,j}, I_n + X_j V_{0,j} X_j', r_{0,j}, s_{0,j})$$

3.10 Used Car Prices

Automobile Prices
Kelly Blue Book: kbb.com
Zipcode is 48109. Mid–level Trim Lines

Year	Miles (1000)	Camary	Accord	Taurus	Grand Prix	Intrepid
00/01	0	23,613	22,390	22,135	22,615	22,920
98	10	17,830	18,315	12,965	16,365	17,500
98	20	17,730	18,215	12,890	16,265	17,400
98	30	17,155	17,640	12,415	15,690	16,000
96	20	15,065	14,815	10,450	11,090	12,345
96	40	14,790	14,540	10,250	10,865	12,120
96	60	13,965	13,715	9,725	10,190	11,445
94	30	11,465	10,250	7,660	8,025	8,175
94	60	11,115	9,900	7,410	7,775	7,925
94	90	9,890	8,675	6,560	6,925	7,075

Dependent Variable:
Percent Difference from New = 100*[Price 00/01 - Price t]/[Price 00/01]

Priors:

$$[\beta|\sigma^2] \quad = \quad N(\beta|0, 100\sigma^2 I) \text{ and } [\sigma^2] = IG(\sigma^2|1, 1)$$

Estimated Models
Posterior STD are parentheses.

Model	$ln[Y]$	Constant	Age	Miles	Japan	Japan* Age	Error Variance
1	−168.7	11.59	8.47				52.30
		(2.85)	(0.66)				(11.28)
2	−192.5	27.61		0.45			128.47
		(3.28)		(0.07)			(27.71)
3	−200.1	50.21			−11.84		210.03
		(2.79)			(4.41)		(45.30)
4	−175.4	11.59	7.44	0.10			48.99
		(2.76)	(0.87)	(0.06)			(10.57)
5	−149.1	16.33	8.47		−11.84		18.63
		(1.78)	(0.39)		(1.31)		(4.02)
6	−154.1	15.84	8.59		−10.62	−0.31	18.57
		(2.19)	(0.51)		(3.47)	(0.80)	(4.00)
7	−190.0	32.35		0.45	−11.84		94.80
		(3.06)		(0.06)	(2.96)		(20.45)
8	−152.7	16.33	7.44	0.10	−11.84		15.32
		(1.62)	(0.49)	(0.03)	(1.19)		(3.30)
9	−157.7	15.84	7.56	0.10	−10.62	−0.31	15.26
		(1.99)	(0.57)	(0.03)	(3.14)	(0.73)	(3.29)

$\ln[Y]$ **is the natural logarithm of the marginal distribution of the data. See page 37 for using these quantities to pick the "best" model.**

- **If misclassification costs are unequal, select the model that minimized the expected loss.**
- **If misclassification costs are equal, select the model the maximized its posterior probability** $q_j(y)$**.**
- **If prior probabilities** q_j **are equal, select the model with the largest marginal distribution of** y**.**
- **The models do not have to be nested.**
- **You could use different transformations of** Y**, but you need to be careful about the priors.**

3.11 Summary

1. **Basic computations for Bayesian Analysis**
2. **Beta–Binomial Conjugate Family**
3. **Normal–Normal–Inverted Gamma Conjugate Family**
4. **Conjugate Normal Regression**
5. **Model Selection**

Chapter 4

Linear Regression

Outline

1. **Objectives**

2. **Linear Regression Model**

3. **Markov Chain Monte Carlo (MCMC)**

4. **Numerical Integration**

5. **Slice Sampling**

6. **Autocorrelated Errors**

4.1 Objectives

1. **The Bayesian analysis of linear regression is straightforward if one uses conjugate priors. In this chapter, we will use a non–conjugate model in order to introduce Markov chain Monte Carlo (MCMC), which is a numerical method for computing integrals. MCMC uses the structure of the statistical model (joint distributions are expressed as products of standard distributions) to simplify the analysis.**

2. **Any practical benefits for being a Bayesian in linear regression? Usually not. For moderate sample sizes, MLE & Bayes are approximately the same. If your design matrix is ill–conditioned, then Bayes estimates are more stable (ridge regression).**

3. **The chapter presents a brief discussion about numerical integration.**

4. **This chapter also presents “slice sampling,” which decomposes complex distributions into simpler ones.**

5. **We then analyze the autocorrelated error regression model using slice sampling.**

4.2 Model

1. **Linear regression model for observation i is**

$$y_i = x_i'\beta + \epsilon_i \text{ for } i = 1, \ldots n.$$

$$= \sum_{j=1}^{p} x_{i,j}\beta_p + \epsilon_i$$

where

- y_i **is the dependent variable for subject** i**.**
- x_i **is a** p **vector of independent variables.**
- **Usually,** $x_{i,1} = 1$**.**
- β **is a** p **vector of unknown regression coefficients.**
- **The error terms** $\{\epsilon_i\}$ **form a random sample from a normal distribution with mean 0 and variance** σ^2**.**

2. Matrix Model:

$$Y = X\beta + \epsilon$$

$$Y = \begin{bmatrix} y_1 \\ y_2 \\ \vdots \\ y_n \end{bmatrix}; \quad X = \begin{bmatrix} x_1' \\ x_2' \\ \vdots \\ x_n' \end{bmatrix}; \text{ and } \quad \epsilon = \begin{bmatrix} \epsilon_1 \\ \epsilon_2 \\ \vdots \\ \epsilon_n \end{bmatrix}.$$

- Y **is a** n **vector of dependent observations.**
- X **is the** $n \times p$ **design matrix.**
- β **is a** p **vector of unknown regression coefficients.**
- ϵ **is a** n **vector of random errors:**

$$[\epsilon] = N_n(\epsilon | 0, \sigma^2 I_n)$$

where I_n **is a** $n \times n$ **identity matrix.**

3. The density of Y is:

$$[Y|\beta,\sigma^2] = N_n(Y|X\beta,\sigma^2 I_n)$$

$$= (2\pi\sigma^2)^{-\frac{n}{2}}\exp\left[-\frac{1}{2\sigma^2}(Y-X\beta)'(Y-X\beta)\right].$$

4. If X has full rank, the maximum likelihood estimators are:

$$\hat{\beta} = (X'X)^{-1}X'Y$$

$$\hat{\sigma}^2 = \frac{1}{n}(Y-X\hat{\beta})'(Y-X\hat{\beta}).$$

In GAUSS:

```
bhat  =  invpd(x'x)*x'y;
s2hat =  (y - x*bhat)'(y- x*bhat)/n;
```

4.3 Prior Distributions

1. **β has a normal distribution with density:**

$$
\begin{aligned}
[\beta|u_0, V_0] &= N_p(\beta|u_0, V_0) \\
&= (2\pi)^{-\frac{p}{2}}|V_0|^{-\frac{1}{2}} \exp\left[-\frac{1}{2}(\beta - u_0)' V_0^{-1}(\beta - u_0)\right].
\end{aligned}
$$

I usually set $u_0 = 0$, and $V_0 = cI_p$ for large c.

2. **σ^2 has an Inverted Gamma distribution with pdf:**

$$
\begin{aligned}
[\sigma^2|r_0, s_0] &= IG(\sigma^2|r_0/2, s_0/2) \\
&= \frac{\left(\frac{s_0}{2}\right)^{\frac{r_0}{2}}}{\Gamma\left(\frac{r_0}{2}\right)} \left(\sigma^2\right)^{-\frac{r_0}{2}-1} \exp\left(-\frac{s_0}{2\sigma^2}\right).
\end{aligned}
$$

for $\sigma^2 > 0$.

I usually set r_0 and s_0 to very small positive numbers.

4.4 Bayesian Inference

1. Joint density:

$$
\begin{aligned}
[Y, \beta, \sigma^2] &= [Y|\beta, \sigma^2][\beta][\sigma^2] \\
\\
&= N_n(Y|X\beta, \sigma^2 I) \\
&\times \ N_p(\beta|u_0, V_0) \\
&\times \ IG(\sigma^2|r_0/2, s_0/2)
\end{aligned}
$$

2. Posterior Distribution of β and σ:

$$
\begin{aligned}
[\beta, \sigma^2 | Y] &= \frac{[Y, \beta, \sigma^2]}{\int\int [Y, \beta, \sigma^2]\, d\beta\, d\sigma^2} \\
&= \frac{[Y|\beta, \sigma^2][\beta][\sigma^2]}{\int\int [Y|\beta, \sigma^2][\beta][\sigma^2]\, d\beta\, d\sigma^2} \\
&\propto [Y|\beta, \sigma^2][\beta][\sigma^2]
\end{aligned}
$$

3. Predictive Distribution of Y_f

(a) Model for Y_f:

$$Y_f = X_f\beta + \epsilon_f,$$

or

$$[Y_f|\beta, \sigma^2] = N_m(Y_f|X_f\beta, \sigma^2 I).$$

Its predictive distribution is:

$$[Y_f|Y] = \int [Y_f|\beta, \sigma^2][\beta, \sigma^2|Y]\, d\beta\, d\sigma^2$$

(b) Predictive mean:

$$E(Y_f|Y) = X_f E(\beta|Y).$$

(c) Predictive variance:

$$\begin{aligned} V(Y_f|Y) &= E(V(Y_f|\beta, \sigma^2)|Y) + V(E(Y_f|\beta, \sigma^2)|Y) \\ &= E(\sigma^2|Y) + V(X_f\beta|Y) \\ &= E(\sigma^2|Y) + X_f V(\beta|Y) X_f' \end{aligned}$$

4.5 Markov Chain Monte Carlo

1. **Generate β and σ^2 from their posterior distribution.**

   (a) **Recursively generate from "full conditionals:"**

   $$[\beta|\sigma, Y] \textbf{ and } [\sigma^2|\beta, Y].$$

   - **Generate $\beta^{(i+1)}$ from**

     $$[\beta|\sigma^{(i)}, Y].$$

     item Generate $\sigma^{(i+1)}$ from

     $$[\sigma^2|\beta^{(i+1)}, Y].$$

   - **The sequence $\{\beta^{(i)}, \sigma^{(i)}\}$ forms a Markov chain such that the stationary distribution is the posterior distribution. That is, eventually the sequence will act as though they are random draws from $[\beta, \sigma|Y]$.**

(b) Joint density:

$$
\begin{aligned}
[Y, \beta, \sigma^2] &= [Y|\beta, \sigma^2][\beta][\sigma^2] \\
\\
&= N_n(Y|X\beta, \sigma^2 I) \\
&\times\ N_p(\beta|u_0, V_0) \\
&\times\ IG(\sigma^2|r_0/2, s_0/2)
\end{aligned}
$$

(c) Full conditional for β:

$$
\begin{aligned}
[\beta|Y,\sigma^2] &= \frac{[Y|\beta,\sigma^2][\beta][\sigma^2]}{\int[Y|\beta,\sigma^2][\beta][\sigma^2]d\beta} \\
&\propto [Y|\beta,\sigma^2][\beta] \\
&\propto N_n(Y|X\beta,\sigma^2 I)N_p(\beta|u_0,V_0) \\
&\propto \exp\left[-\frac{1}{2\sigma^2}(Y-X\beta)'(Y-X\beta)\right] \\
&\times \exp\left[-\frac{1}{2}(\beta-u_0)'V_0^{-1}(\beta-u_0)\right]
\end{aligned}
$$

Write this as a function of β.

Expand the squares in β.

$$\frac{1}{2\sigma^2}(Y - X\beta)'(Y - X\beta) =$$

$$\frac{1}{2}(\beta - u_0)'V_0^{-1}(\beta - u_0) =$$

Complete the squares in β:

Did you get

$$[\beta|Y, \sigma^2] = N_p(\beta|u_n, V_n)$$

with

$$V_n = \left(\frac{1}{\sigma^2} X'X + V_0^{-1}\right)^{-1}$$

$$u_n = V_n \left(\frac{1}{\sigma^2} X'Y + V_0^{-1} u_0\right)$$

(d) Full conditional for σ^2.

Write it as a function of σ^2.

Did you get

$$[\sigma^2|Y,\beta] = IG(\sigma^2|r_n/2, s_n/2)$$

with

$$\begin{aligned} r_n &= r_0 + n \\ s_n &= s_0 + (Y - X\beta)'(Y - X\beta) \end{aligned}$$

2. Use random iterates for inference.

Suppose you have generated a sequence of random deviates: $\{\beta^{(i)}, \sigma^{(i)}\}$ **for** $i = B + 1, \ldots, M$**.** **Blow-off the first** B **iterates (transitory period).**

(a) Point Estimates

Approximate posterior parameters by corresponding summary statistics from $\{\beta^{(i)}, \sigma^{(i)}\}$**.**

- **Posterior Mean ≈ Sample Means:**

$$E(\beta|Y) \approx \frac{1}{M - B} \sum_{i=B+1}^{M} \beta^{(i)}.$$

- **Posterior Median ≈ Sample Median.**
- **Posterior Standard Deviations ≈ Sample Standard Deviations.**
- **Posterior Covariance ≈ Sample Covariance.**

(b) Marginal Distributions

- **Make histograms based on $\{\beta^{(i)}, \sigma^{(i)}\}$.**
- **Better but more work:**

$$
\begin{aligned}
[\beta|Y] &= \int [\beta|Y,\sigma][\sigma|Y]\, d\sigma \\
&\approx \frac{1}{M-B} \sum_{i=B+1}^{M} N_p(\beta|u_n^{(i)}, V_n^{(i)}) \\
[\sigma^2|Y] &= \int [\sigma^2|Y,\beta][\beta|Y]\, d\beta \\
&\approx \frac{1}{M-B} \sum_{i=B+1}^{M} IG(\sigma^2|r_n^{(i)}/2, s_n^{(i)}/2)
\end{aligned}
$$

For example, fix a grid of values for σ^2. At each iteration of the MCMC, compute $IG(\sigma^2|r_n^{(i)}/2, s_n^{(i)}/2)$ density at each grid point σ^2. Then average these densities over the iterations.

(c) Predictive distributions:

- **Nice but lots of work:**

$$
\begin{aligned}
[Y_f|Y] &= \int [Y_f|\beta,\sigma^2][\beta,\sigma^2|Y]\, d\beta\, d\sigma^2 \\
&\approx \frac{1}{M-B} \sum_{i=B+1}^{M} N_m(Y_f|X_f\beta^{(i)}, \left(\sigma^{(i)}\right)^2 I)
\end{aligned}
$$

- **During or after MCMC, you could generate**

$$[Y_f|Y,\beta^{(i)},\sigma^{(i)}] = N_n(Y_f|X_f\beta^{(i)}, (\sigma^{(i)})^2 I)$$

and use $\{Y_f^{(i)}\}$ **anyway you want.**

- **Predictive mean:**

$$E(Y_f|Y) = X_f E(\beta|Y) \approx \frac{1}{M-B} \sum_{i=B+1}^{M} Y_f^{(i)}$$

- **Predictive variance:**

$$
\begin{aligned}
V(Y_f|Y) &= E(\sigma^2|Y) + X_f V(\beta|Y) X_f' \\
&\approx \textbf{ Sample Covariance of } \{Y_f^{(i)}\}.
\end{aligned}
$$

4.6 Example using Simulated Data

Generate 30 observations from:

$$Y = 2 - 1X_1 + 3X_2 + 0X_3 + \epsilon$$

where ϵ is from the normal distribution with mean 0 and standard deviation 2.

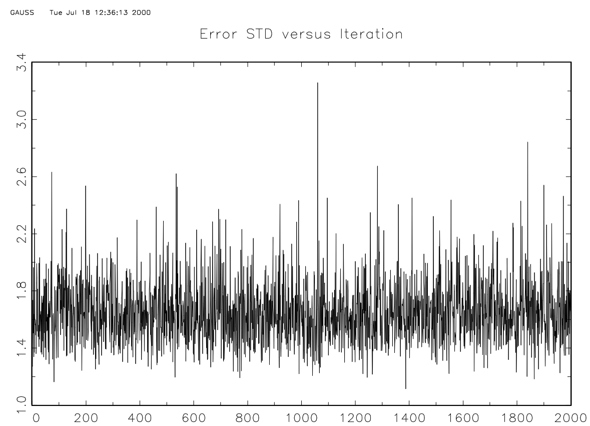
Priors:

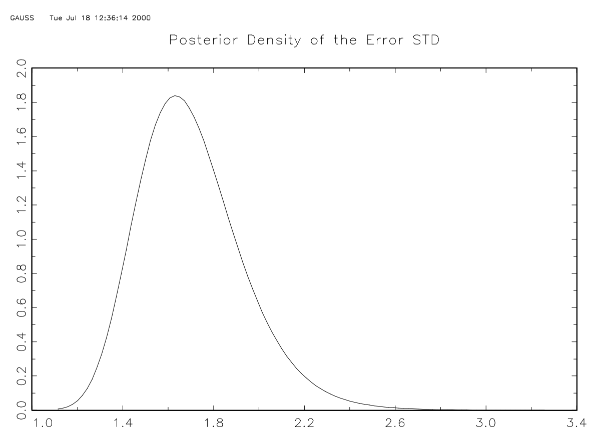
$$\begin{aligned} [\beta] &= N_4(\beta|0, 100I) \\ [\sigma^2] &= IG(\sigma^2|1, 1) \end{aligned}$$

```
Model
Y = X*beta + epsilon
Number of observations          =   30.00000
Number of independent variables =    3.00000 (excluding the intercept).
Summary Statistics
Variable        Mean       STD       MIN       MAX
X1          -0.07043   1.15915  -2.71526   2.11850
X2          -0.09506   1.14323  -2.46804   2.28341
X3           0.04298   0.76840  -1.82211   1.51536
Y            2.11729   3.53558  -5.30671   9.19462


R-Squared     =    0.79150
Multiple R    =    0.88966
MLE Error STD =    1.58728

Estimated Regression Coefficients
Variable         MLE  StdError
Constant     2.31687   0.29142
X1          -0.66872   0.28760
X2           2.86992   0.28536
X3           0.60799   0.39830
```

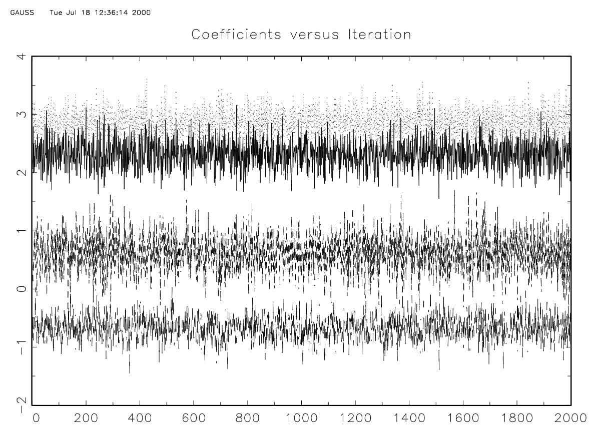
```
MCMC Analysis

Total number of MCMC iterations              = 2000
Number of iterations used in the analysis    = 1000
Number in transition period                  = 1000
Number of iterations between saved iterations =    0.00000


Bayes R-Square   =      0.79150
Bayes Multiple R =      0.88966

Error Standard Deviation
Posterior mean of sigma =    1.65428
Posterior STD  of sigma =    0.22341

Regression Coefficients
Variable    PostMean   PostSTD
Constant     2.31821   0.22730
X1          -0.66451   0.22825
X2           2.86502   0.23195
X3           0.60480   0.32320
```

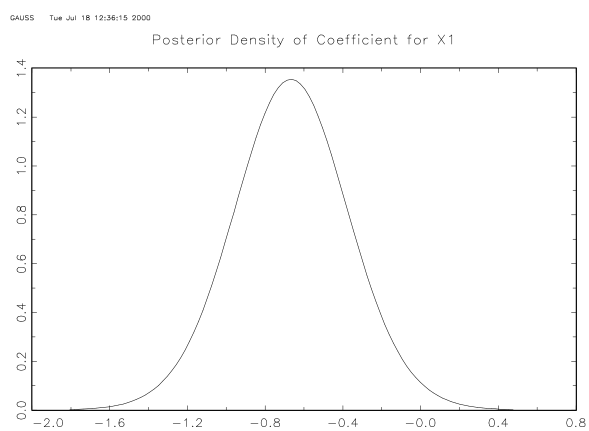
MCMC for Error STD

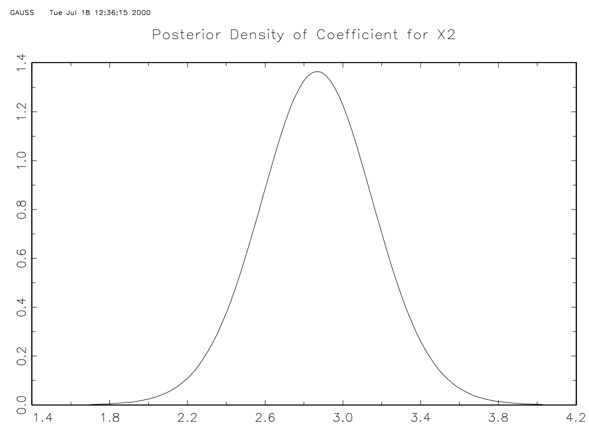


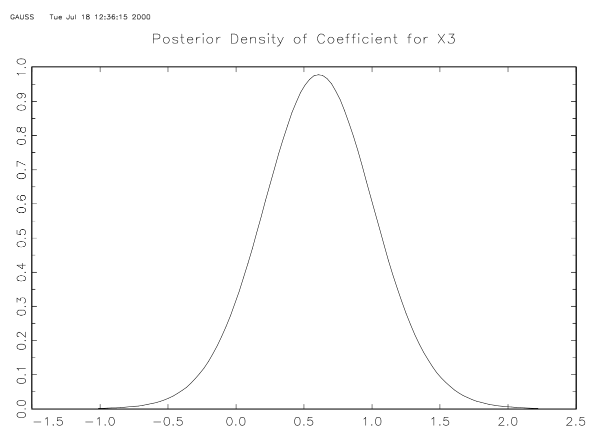


MCMC for Coefficients






GAUSS Tue Jul 18 12:36:15 2000
Posterior Density of Coefficient for X2
0.0
0.2
0.4
0.6
0.8
1.0
1.2
1.4
1.4
1.8
2.2
2.6
3.0
3.4
3.8
4.2


GAUSS Tue Jul 18 12:36:15 2000
Posterior Density of Coefficient for X3
0.0
0.1
0.2
0.3
0.4
0.5
0.6
0.7
0.8
0.9
1.0
−1.5
−1.0
−0.5
0.0
0.5
1.0
1.5
2.0
2.5

4.7 Numerical Integration

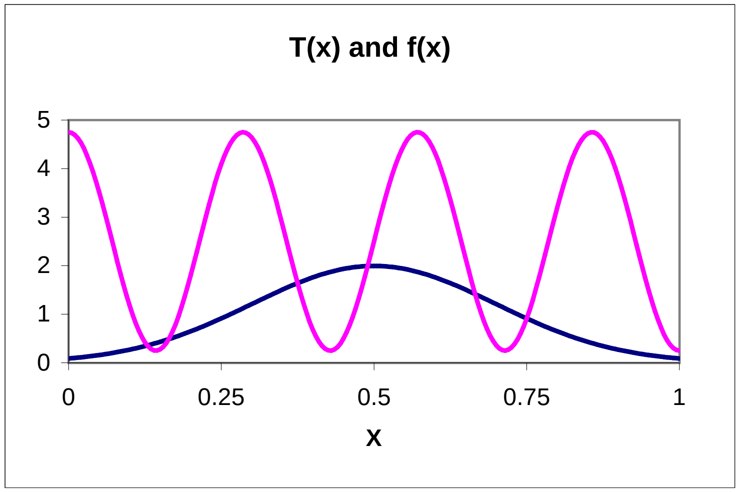
1. **Bayesian analysis requires the computation of integrals:**

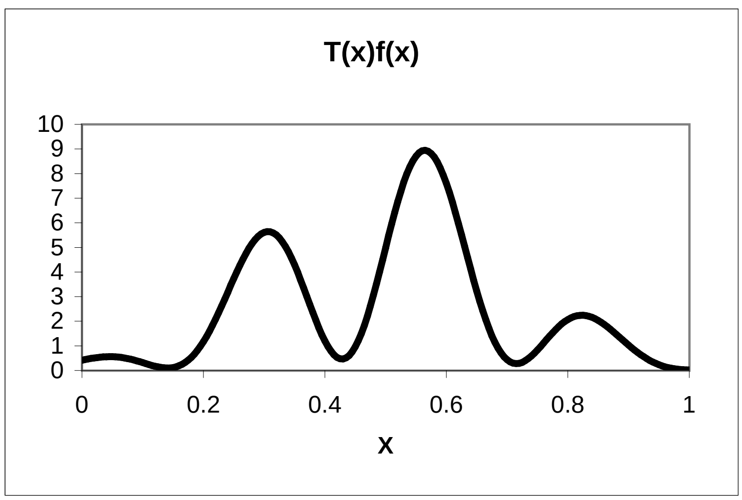
$$E[T(X)] = \int_{\mathcal{X}} T(x)f(x)dx.$$

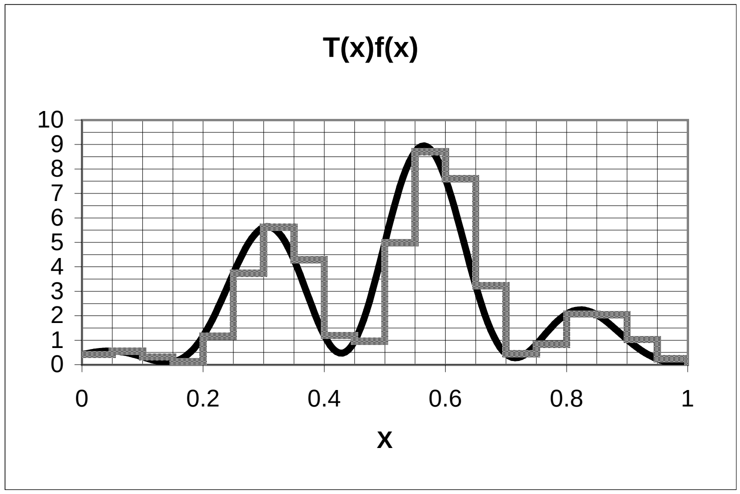
where X has density f, and T is a function.

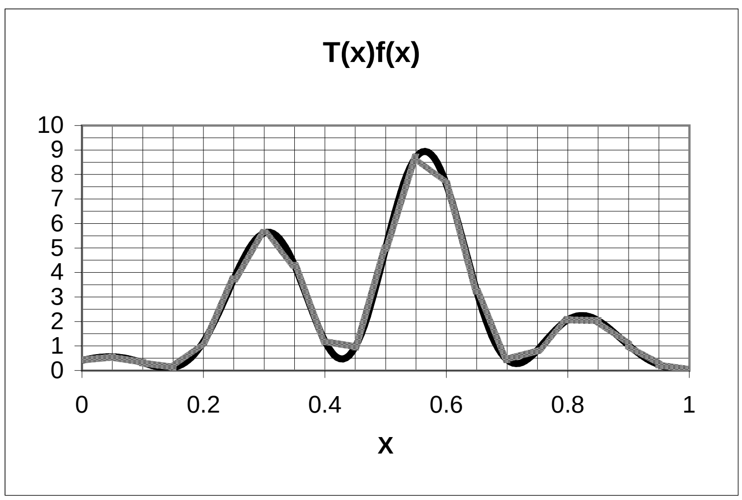
2. **Examples**

   - **Under squared-error loss, the Bayes rule is the posterior mean, and the Bayes risk is the posterior variance.**
   - **Under absolute-error loss, the Bayes rule is the posterior median.**
   - **Posterior distributions and the posterior probability of a model require the marginal distribution of the data, which integrates the likelihood with respect to the prior.**


T(x) and f(x)
5
4
3
2
1
0
0
0.25
0.5
0.75
1
X


T(x)f(x)
10
9
8
7
6
5
4
3
2
1
0
0
0.2
0.4
0.6
0.8
1
X


T(x)f(x)
10
9
8
7
6
5
4
3
2
1
0
0
0.2
0.4
0.6
0.8
1
X


T(x)f(x)
10
9
8
7
6
5
4
3
2
1
0
0
0.2
0.4
0.6
0.8
1
X

3. **Grid methods such as the trapezoid rule and Simpson's integration can achieve a high degree of accuracy with relative few functional evaluations of $T(x)f(x)$.**

4. **Downside of grid methods**

    - **You have to be really smart to make them work well.**
        - **You need to know the support of $T(x)f(x)$.**
        - **You need to know how wavy $T(x)f(x)$ is.**
    - **They do not scale-up to higher dimensions. The number of gird points increases geometrically with the number of dimensions.**

5. Monte Carlo

- **Suppose that you have a random number generator for f.**
- **Generate an iid sequence X_1, X_2, . . . , X_M.**
- **Approximate**

$$E[\widehat{T(X)}] = \frac{1}{M}\sum_{i=1}^{M} T(X_i).$$

  This converges to $E[T(X)]$ by the strong law of large numbers as M increases.

- **The accuracy of the approximation in root mean squared error is**

$$M^{-1/2}STD[T(X_1)].$$

6. **Upside of Monte Carlo**

   - **You do not have to be smart.**

     The researchers who developed f and its random number generator did all the hard thinking for you.

   - **It scales up to higher dimensions.**

     The rate of convergence is $M^{-1/2}$ regardless of the dimension. As you increase dimensions, the absolute accuracy declines, but the rate stays the same.

7. **Downside of Monte Carlo**

   - **In many applications, you do not have a random number generator for f.**

8. Importance sampling.

- **Suppose that you only know the function form of f:**

$$
\begin{aligned}
f(x) &= g(x)/c \\
c &= \int_{\mathcal{X}} g(x)dx.
\end{aligned}
$$

  Importance sampling does not require that you know c.

- **For example, the posterior density is:**

$$p(\theta|x) \propto f(x|\theta)p(\theta).$$

- **You would like to use Monte Carlo, but you have a random number generator for h, not f where h has the same support as f.**

- **Generate Y_1, Y_2, . . . Y_M iid from h.**

- **We need to approximate**

$$
\begin{aligned}
\int_{\mathcal{X}} T(x)f(x)dx &= \frac{\int_{\mathcal{X}} T(x)g(x)dx}{\int_{\mathcal{X}} g(x)dx} \\
&= \frac{\int_{\mathcal{X}} T(x)\frac{g(x)}{h(x)}h(x)dx}{\int_{\mathcal{X}} \frac{g(x)}{h(x)}h(x)dx} \\
E[\widehat{T(X)}] &= \frac{M^{-1}\sum_{i=1}^{M} T(Y_i)\frac{g(Y_i)}{h(Y_i)}}{M^{-1}\sum_{i=1}^{M} \frac{g(Y_i)}{h(Y_i)}} \\
&= \frac{M^{-1}\sum_{i=1}^{M} T(Y_i)W(Y_i)}{M^{-1}\sum_{i=1}^{M} W(Y_i)} \\
W(Y) &= g(Y)/h(Y)
\end{aligned}
$$

- **The strong law of large number applies if**

$$
\begin{aligned}
\int_{\mathcal{X}} T(x)^2\frac{g(x)^2}{h(x)}dx &< \infty \\
\int_{\mathcal{X}} \frac{g(x)^2}{h(x)}dx &< \infty.
\end{aligned}
$$

9. Upside of importance sampling.

- **Greatly extends the applicability of existing random number generators.**

10. Downside of importance sampling.

- **You need to be a little smart or else it will not converge very rapidly or at all.**
- h **should match** g **as well as possible.**
- **If the tails of** h **are smaller than that of** g**, the approximation may fail.**
- **If the tails of** h **are much larger than that of** g**, the approximation may be inaccurate.**
- **How do you know?**

11. Markov Chain Monte Carlo

- **Exploits the structure of Bayesian models.**
- **Simplifies complex posterior distributions by successive conditioning.**
- **Generate random deviates from a Markov chain such that the stationary distribution is the posterior distribution.**

12. Example:

- **Generate** (X, Y) **from the joint distribution** $f(x, y)$**.**
- **We do not have a random number generator for** $f(x, y)$**.**
- **We have random number generators for the conditionals** $g(x|y)$ **and** $h(y|x)$**.**
- **Recursively generating** $Y|X$ **and** $X|Y$**:**

$$
\begin{aligned}
[x_i|y_{i-1}] &= g(x_i|y_{i-1}) \\
[y_i|x_i] &= h(y_i|x_i)
\end{aligned}
$$

- **Why does it work?**

$$g(x) = \int_{\mathcal{Y}} g(x|s)h(s)ds$$

$$\begin{aligned}
h(y) &= \int_{\mathcal{X}} h(y|x)g(x)dx \\
&= \int_{\mathcal{X}} h(y|x)\left[\int_{\mathcal{Y}} g(x|s)h(s)ds\right]dx \\
&= \int_{\mathcal{Y}} \left[\int_{\mathcal{X}} h(y|x)g(x|s)dx\right]h(s)ds \\
&= \int_{\mathcal{Y}} h(s)K(s,y)ds
\end{aligned}$$

- **The marginal distribution of Y is the stationary distribution for the transition probability $K(s,y)$, which the probability that the Markov chain moves from s to y.**
- **The joint distribution of the pairs (X_i, Y_i) converges to $f(x,y)$.**

13. Upside of MCMC

- **Often it is very easy.**
- **Allows the analysis of very complex models.**

14. Downside of MCMC

- **Random deviates are not independent.**
- **It is more difficult to compute the numerical accuracy than in Monte Carlo.**
- **In complex models, the autocorrelation is very high. This means that the MCMC will have to run for a long time to obtain accurate approximations.**
- **There is a transition period before the random deviates start coming from the stationary distribution.**
- **There are diagnostics for the transition period, but all of them are flawed.**

4.8 Uniform Distribution

1. **X has the uniform distribution on θ_1 to θ_2 for $\theta_1 < \theta_2$ if its density is:**

$$\begin{aligned}[x] &= U(x|\theta_1, \theta_2) \\ &= \frac{1}{\theta_2 - \theta_1} \textbf{ for } \theta_1 < x < \theta_2\end{aligned}$$

2. **Moments**

$$\begin{aligned}E(X^k) &= \frac{1}{\theta_2 - \theta_1} \int_{\theta_1}^{\theta_2} x^k dx \\ &= \frac{\theta_2^{k+1} - \theta_1^{k+1}}{(k+1)(\theta_2 - \theta_1)}\end{aligned}$$

$$E(X) = \frac{1}{2}(\theta_1 + \theta_2) \textbf{ and } V(X) = \frac{1}{12}(\theta_2 - \theta_1)^2$$

3. **Generate X:**

$$x = (\theta_2 - \theta_1)u + \theta_1$$

where u is uniform on 0 to 1.

4.9 Slice Sampling

1. **Slice sampling is a method of decomposing complex distributions into simpler ones for random variable generation.**

2. **Suppose that the distribution you want to generate from has the form:**

$$[x] \propto g(x)h(x).$$

3. **Introduce an auxiliary random variable V so that the joint density of V and X is:**

$$[v, x] \propto I[0 < v < g(x)]h(x).$$

where I is the indicator function.

4. **Key concept:**

$$[x] = \int [v, x] dv \propto \left[\int_0^{g(x)} dv\right] h(x) = g(x)h(x).$$

5. Full conditional of V is:

$$
\begin{aligned}
[v|x] &\propto I[0 < v < g(x)] \\
&= U(v|0, g(x))
\end{aligned}
$$

So V is uniform from 0 to $g(x)$. Generate V:

$V = g(x)U$ **where U is uniform on 0 to 1**.

6. Full conditional of X is:

$[x|v] \propto h(x)$ **for x such that $v < g(x)$**

7. Generate X from the truncated distribution of h on the set $\{x|v < g(x)\}$:

- **if it is easy to invert $g(x)$ and**
- **if it is easy to generate from truncated density h.**

8. Univariate, truncated distributions.

<u>Fact</u>: If F is the cdf of X,

$$F(a) = P(X < a),$$

then $U = F(X)$ is uniform:

$$\begin{aligned} P(U < u) &= P(F(X) < u) = P(X < F^{-1}(u)) \\ &= F[F^{-1}(u)] = u \textbf{ for } 0 < u < 1. \end{aligned}$$

The cdf F of X is:

$$[x] \propto h(x)I(a < x < b)$$

$$F(x) = \frac{\int_a^x h(s)dx}{\int_a^b h(s)ds}$$

$$F(x) = \frac{H(x) - H(a)}{H(b) - H(a)}$$

where H is the cdf corresponding

to the density h.

Set $F(x) = u$ where u is uniform on 0 to 1.

Solve for x:

$$x = H^{-1}\left[uH(b) + (1-u)H(a)\right].$$

This works well if you can easily obtain H and H^{-1}.

- **Uniform**
- **Exponential**
- **Gamma (sometimes)**
- **Normal**

4.10 Autocorrelated Errors

4.10.1 Model

$$
\begin{aligned}
y_t &= x_t'\beta + \epsilon_t \text{ for } t = 1, \ldots, T \\
\epsilon_t &= \xi_t + \rho\epsilon_{t-1} \text{ for } t = 2, \ldots, T \\
\rho &\in (-1, 1) \\
[\xi_t] &= N(\xi_t|0, \sigma^2) \text{ for } t = 2, \ldots, T \\
[\epsilon_1] &= N\left(\epsilon_1|0, \frac{\sigma^2}{1-\rho^2}\right).
\end{aligned}
$$

1. **The innovations or "shocks,"** $\{\xi_2, \ldots, \xi_T\}$**, are iid.**

2. ϵ_1 **is independent of the** $\{\xi_t\}$ **and has a normal distribution with mean 0 and stationary variance** $\sigma^2/(1-\rho^2)$**.**

3. **Write the error terms as geometric series of past innovations:**

$$\begin{aligned}
\epsilon_2 &= \xi_2 + \rho\epsilon_1 \\
\epsilon_3 &= \xi_3 + \rho\epsilon_2 \\
&= \xi_3 + \rho\xi_2 + \rho^2\epsilon_1 \\
\epsilon_4 &= \xi_4 + \rho\epsilon_3 \\
&= \xi_4 + \rho\xi_3 + \rho^2\xi_2 + \rho^3\epsilon_3 \\
\epsilon_t &= \xi_t + \rho\epsilon_{t-1} \\
&= \xi_t + \rho\xi_{t-1} + \rho^2\xi_{t-2} + \cdots + \rho^{t-2}\xi_2 + \rho^{t-1}\epsilon_1
\end{aligned}$$

4. **The stationary variance makes all of the variances of the ϵ_t equal, say σ_ϵ^2:**

$$\begin{aligned}
V(\epsilon_t) &= \sigma^2 + \rho^2 V(\epsilon_{t-1}) \\
\sigma_\epsilon^2 &= \sigma^2 + \rho^2\sigma_\epsilon^2 \\
\sigma_\epsilon^2 &= \sigma^2/(1-\rho^2)
\end{aligned}$$

5. Using the geometric series expression for ϵ_t, the covariances are:

$$E(\epsilon_t \epsilon_{t+u}) = \sigma^2 \frac{\rho^u}{1-\rho^2} \textbf{ for } u > 0.$$

6. The variance–covariance matrix of the error terms is:

$$\begin{aligned}
E(\epsilon\epsilon') &= \Sigma \\
&= \frac{\sigma^2}{1-\rho^2}\Upsilon \\
\Upsilon &= \begin{bmatrix} 1 & \rho & \rho^2 & \dots & \rho^{t-1} \\ \rho & 1 & \rho & \dots & \rho^{t-2} \\ \vdots & & \ddots & & \vdots \\ \rho^{t-1} & \rho^{t-2} & \rho^{t-3} & \dots & 1 \end{bmatrix}
\end{aligned}$$

7. Υ is the familiar Topelitz matrix and has inverse and determinate:

$$\Upsilon^{-1} = \frac{1}{1-\rho^2}\begin{bmatrix} 1 & -\rho & 0 & 0 & \dots & 0 \\ -\rho & 1+\rho^2 & -\rho & 0 & \dots & 0 \\ 0 & -\rho & 1+\rho^2 & -\rho & \dots & 0 \\ \vdots & & \ddots & \ddots & \ddots & \vdots \\ 0 & \dots & & -\rho & 1+\rho^2 & -\rho \\ 0 & \dots & & 0 & -\rho & 1 \end{bmatrix}$$

$$\det(\Upsilon) = (1-\rho^2)^{T-1}$$

8. Define the residuals:

$$r_t = y_t - x_t'\beta.$$

Then the conditional distribution of

Y **given** β, σ, **and** ρ **is:**

$$[Y|\beta, \sigma, \rho] \propto \det(\Sigma)^{-1/2} \exp\left\{-\frac{1}{2}(Y - X\beta)'\Sigma^{-1}(Y - X\beta)\right\}$$

$$\propto \left(\frac{\sqrt{1-\rho^2}}{\sigma^T}\right) \exp\left\{-\frac{1}{2\sigma^2}\left[(1-\rho^2)r_1^2 + \sum_{t=2}^{T}(r_t - \rho r_{t-1})^2\right]\right\}$$

9. Note that:

$$r_t - \rho r_{t-1} = y_t - \rho y_{t-1} - (x_t - \rho x_{t-1})'\beta.$$

10. Define

$$
\begin{aligned}
y_1^* &= (1-\rho)y_1 \\
y_t^* &= y_t - \rho y_{t-1} \textbf{ for } t = 2, \ldots T \\
x_1^* &= (1-\rho)x_1 \\
x_t^* &= x_t - \rho x_{t-1} \\
Y^* &= \begin{bmatrix} y_1^* \\ y_2^* \\ \vdots \\ y_T^* \end{bmatrix} \textbf{ and } X^* = \begin{bmatrix} x_1^{*'} \\ x_2^{*'} \\ \vdots \\ x_T^{*'} \end{bmatrix}.
\end{aligned}
$$

11. The AR normal density is:

$$
[Y|\beta, \sigma, \rho] \propto \frac{\sqrt{1-\rho^2}}{\sigma^T} \exp\left\{-\frac{1}{2\sigma^2}(Y^* - X^*\beta)'(Y^* - X^*\beta)\right\}.
$$

12. The prior distributions are:

$$
\begin{aligned}
[\beta] &= N_p(\beta|u_0, V_0) \\
[\sigma^2] &= IG(\sigma^2|r_0/2, s_0/2) \\
[\rho] &= U(\rho|-1, 1).
\end{aligned}
$$

4.10.2 MCMC

1. Full Conditional for β:

$$
\begin{aligned}
[\beta|Y,\sigma,\rho] &\propto [Y|\beta,\sigma][\beta] \\
&= N_p(\beta|u_T,V_T) \\
V_T &= \left[X^{*'}X^*/\sigma^2 + V_0^{-1}\right]^{-1} \\
u_T &= V_T\left[X^{*'}Y^*/\sigma^2 + V_0^{-1}u_0\right]
\end{aligned}
$$

2. Full Conditional of σ:

$$
\begin{aligned}
[\sigma^2|Y,\beta,\rho] &\propto [Y|\beta,\sigma,\rho][\sigma] \\
&= IG(\sigma^2|r_T/2, s_T/2) \\
r_T &= r_0 + T \\
s_T &= s_0 + (Y^* - X^*\beta)'(Y^* - X^*\beta)
\end{aligned}
$$

3. Full Conditional of ρ:

$$
\begin{aligned}
[\rho|Y,\beta,\sigma] &\propto [Y|\beta,\sigma^2][\rho] \\
&\propto \exp\left\{-\frac{1}{2\sigma^2}\left[(1-\rho^2)r_1^2 + \sum_{t=2}^{T}(r_t - \rho r_{t-1})^2\right]\right\} \\
&\times \sqrt{1-\rho^2} I(-1<\rho<1) \\
&\propto \exp\left\{-\frac{1}{2\sigma^2}\left[\rho^2\left(\sum_{t=2}^{T} r_t^2\right) - 2\rho\left(\sum_{t=2}^{T} r_t r_{t-1}\right)\right]\right\} \\
&\times \sqrt{1-\rho^2} I(-1<\rho<1) \\
&\propto \sqrt{1-\rho^2} N(\rho|a,b^2) I(-1<\rho<1) \\
b^2 &= \sigma^2\left(\sum_{t=2}^{T} r_t^2\right)^{-1} \\
a &= \frac{\sum_{t=2}^{T} r_t r_{t-1}}{\sum_{t=2}^{T} r_t^2}
\end{aligned}
$$

Use slice sampling with

$$g(\rho) = \sqrt{1-\rho^2}$$
$$h(\rho) = N(\rho|a,b^2)I(-1<\rho<1)$$

- **Given** ρ**, generate** V **from a uniform on 0 to** $(1-\rho^2)^{1/2}$**:**

$$v = \sqrt{1-\rho^2}u \text{ **where u is uniform on 0 to 1**}.$$

- **Given** v**, find the region where**

$$v < \sqrt{1-\rho^2} \text{ **or** } -\sqrt{1-v^2} < \rho < \sqrt{1-v^2}.$$

- **Given** v**, generate** ρ **from a truncated normal:**

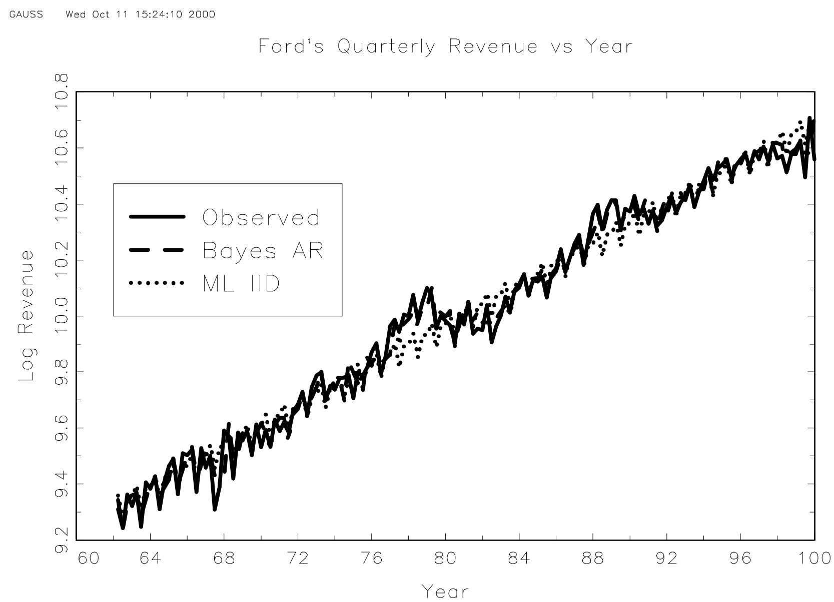
$$[\rho|v] \propto N(\rho|a,b^2)$$
$$\textbf{for} \quad \max(-1,-\sqrt{1-v^2}) < \rho < \min(1,\sqrt{1-v^2}).$$

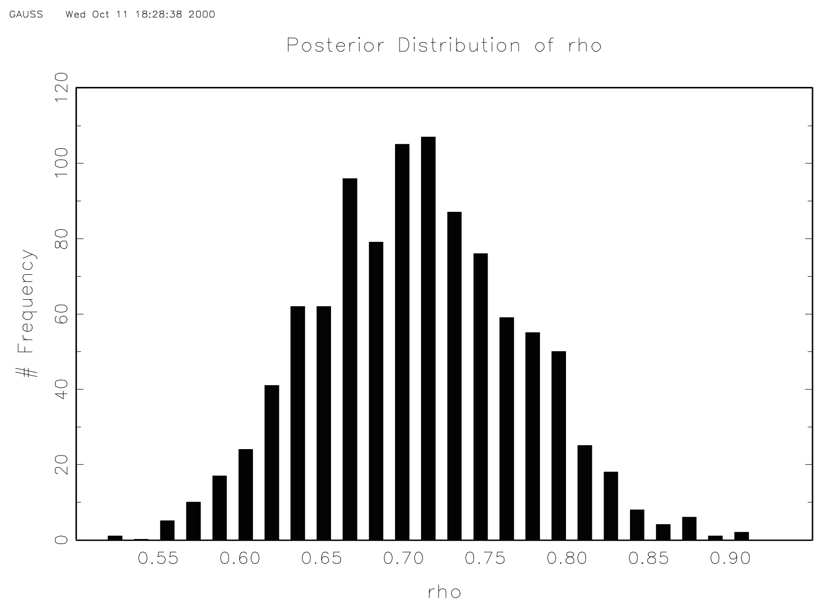
4.10.3 Quarterly Revenue

Data: Quarterly revenues for the Ford Motor Corp. from 1962 Q1 to 2000 Q1.

```
Linear Regression with Autocorrelated Errors

Y = X*beta + epsilon
epsilon_t = rho*epsilon_{t-1} + z_t

Number of observations          =  153.00000
Summary Statistics
Variable        Mean       STD       MIN       MAX
Year        81.00000  11.07785  62.00000 100.00000
Q1           0.25490   0.43724   0.00000   1.00000
Q2           0.24837   0.43348   0.00000   1.00000
Q3           0.24837   0.43348   0.00000   1.00000
Sales        9.99955   0.40603   9.24249  10.70802

---------------------------------------------------------------------------
MLE Analysis

R-Squared     =    0.98043
Multiple R    =    0.99017
One-Step Ahead Predictive RMSE =    0.05678

MLE Error STD =    0.05662

Estimated Regression Coefficients
Variable         MLE  StdError
Const        7.08147   0.03492
Year         0.03617   0.00041
Q1          -0.00484   0.01291
Q2           0.02564   0.01299
Q3          -0.06815   0.01299
```

```
MCMC Analysis

Total number of MCMC iterations                 = 2000.00000
Number of iterations used in the analysis       = 1000.00000
Number in transition period                     = 1000.00000
Number of iterations between saved iterations =     0.00000

Bayes R-Square   =      0.98042
Bayes Multiple R =      0.99016
One-Step Ahead Predictive RMSE without AR Correction   =    0.05697
One-Step Ahead Predictive RMSE Corrected for AR Errors =    0.04123

Error Standard Deviation
Posterior mean of sigma =    0.04197
Posterior STD  of sigma =    0.00246

Error Correlation
Posterior mean of rho =    0.71377
Posterior STD  of rho =    0.06288

Regression Coefficients
Variable    PostMean   PostSTD
Const        7.10347   0.09090
Year         0.03587   0.00111
Q1          -0.00405   0.00704
Q2           0.02403   0.00785
Q3          -0.06863   0.00691
```


GAUSS Wed Oct 11 15:24:10 2000
Ford's Quarterly Revenue vs Year
Log Revenue
9.2
9.4
9.6
9.8
10.0
10.2
10.4
10.6
10.8
Observed
Bayes AR
ML IID
60
64
68
72
76
80
84
88
92
96
100
Year


GAUSS Wed Oct 11 18:28:38 2000
Posterior Distribution of rho
Frequency
0
20
40
60
80
100
120
0.55
0.60
0.65
0.70
0.75
0.80
0.85
0.90
rho

4.11 Summary

1. **Non–conjugate Linear Regression Model**

2. **Markov Chain Monte Carlo**

3. **Slice sampling**

4. **Autoregressive errors**

References

- Damien, P., Wakefield, J. C., and Walker, S. (1999), "Gibbs sampling for Bayesian nonconjugate and hierarchical models using auxiliary variables," *Journal of the Royal Statistical Society, Series B*, Vol. 61 Part 2, 331–344.

- Gelfand, A. E. and Smith, A. F. M. (1990). "Sampling–Based Approaches to Calculating Marginal Densities," *Journal of the American Statistical Association,* 85, 398–409.

- Tanner, M. A. (1993). *Tools for Statistical Inference,* Lecture Notes in Statistics 67, New York: Springer–Verlag.

- Smith, A. F. M. and Roberts, G. O. (1993). "Bayesian Computation Via the Gibbs Sampler and Related Markov Chain Monte Carlo Methods," *Journal of the Royal Statistical Society Series B*, 55, 3–23.

Chapter 5

Multivariate Regression

Outline

1. **Objectives**
2. **Matrix Algebra**
3. **Distributions**
   - **Matrix Normal Distribution**
   - **Wishart Distribution**
   - **Inverted Wishart Distribution**
4. **Model**
5. **Prior Distributions**
6. **Full Conditionals**

5.1 Objectives

1. **Multivariate regression is an extension of linear regression. It requires advanced "book keeping" to keep track of the numbers. The advanced book keeping are some definitions and identities from matrix algebra. Its not hard, but if you were not aware of these identities, the statistics would become very tough.**

2. **The analysis of hierarchical Bayes models relies heavily on this chapter. One output of this chapter will be a subroutine that is frequently called for other models.**

3. **The multivariate model requires the matrix normal and Wishart and Inverted Wishart distributions.**

4. **The Wishart and Inverted Wishart distributions are the multivariate extensions of the Gamma and Inverted Gamma distributions. The Inverted Wishart distribution is used for the prior distribution of the covariance matrix.**

5.2 Matrix Algebra

1. **$A = (a_{ij})$ is a $m \times m$ matrix.**

   The trace of A is the sum of its diagonal elements:

$$\mathbf{tr}(A) = \sum_{i=1}^{m} a_{ii}.$$

2. **A is a $m \times n$ matrix with columns a_j**

$$A = [a_1\, a_2\, \cdots\, a_n].$$

   vec(A) is a $mn \times 1$ vector that stacks the columns of A:

$$\mathbf{vec}(A) = \begin{bmatrix} a_1 \\ a_2 \\ \vdots \\ a_n \end{bmatrix}.$$

   vec(A') stacks the rows of A.

3. **Gauss has the operators "vec(A)" that stacks the columns of A, and "vecr(A)" that stacks the rows of A.**

4. Kronecker Product or Direct Product

$A = (a_{ij})$ **is a** $p \times q$ **matrix.**

$B = (b_{ij})$ **is a** $r \times s$ **matrix.**

Their direct product is a $pr \times qs$ **matrix:**

$$A \otimes B = \begin{bmatrix} a_{11}B & a_{12}B & \dots & a_{1q}B \\ a_{21}B & a_{22}B & \dots & a_{2q}B \\ \vdots & \vdots & \ddots & \vdots \\ a_{p1}B & a_{p2}B & \dots & a_{pq}B \end{bmatrix}$$

5. Mini Facts about Direct Products

(a) $(aA) \otimes (bB) = ab(A \otimes B)$ **for scalars** a **and** b**.**

(b) $(A + B) \otimes C = A \otimes C + B \otimes C.$

(c) $(A \otimes B) \otimes C = A \otimes (B \otimes C).$

(d) $(A \otimes B)' = A' \otimes B'$**.**

(e) $(A \otimes B)(C \otimes D) = AC \otimes BD$

(f) $(A \otimes B)^{-1} = A^{-1} \otimes B^{-1}$

(g) If H and Q are both orthogonal matrices ($H' = H$ and $H'H = I$), then so is $H \otimes Q$.

(h) If A and B are both $m \times m$:

$$\mathbf{tr}(A \otimes B) = [\mathbf{tr}(A)][\mathbf{tr}(B)].$$

(i) If A is $m \times m$ and B is $n \times n$, then

$$|A \otimes B| = |A|^n |B|^m.$$

(j) A is $m \times m$ with latent roots a_1, . . . , a_m.

B is $n \times n$ with latent roots b_1, . . . , b_m.

The latent roots of $A \otimes B$ are $a_i b_j$

for $i = 1, \ldots, n$ and $j = 1, \ldots, m$.

(k) If A and B are positive definite, then so is $A \otimes B$.

(l) If B is $r \times m$; X is $m \times n$, and C is $n \times s$, then

$$\mathbf{vec}(BXC) = (C' \otimes B)\mathbf{vec}(X).$$

(m) If B is $k \times m$ and C is $m \times n$.

$$\begin{aligned} \mathbf{vec}(BC) &= (I_n \otimes B)\mathbf{vec}(C) \\ &= (C' \otimes I_k)\mathbf{vec}(B) \\ &= (C' \otimes B)\mathbf{vec}(I_m) \end{aligned}$$

(n) B is $k \times m$; C is $m \times n$, and D is $n \times k$.

$$\mathbf{tr}(BCD) = [\mathbf{vec}(B')]'(I_n \otimes C)\mathbf{vec}(D)$$

(o) For B, X, C, and D of the correct dimensions:

$$\begin{aligned} \mathbf{tr}(BX'CXD) &= [\mathbf{vec}(X)]'(B'D' \otimes C)\mathbf{vec}(X) \\ &= [\mathbf{vec}(X)]'(DB \otimes C')\mathbf{vec}(X) \end{aligned}$$

6. Example

- **There are n subjects.**
- **There are m measurements for each subject:**

  Y_i **is a m vector for $i = 1, \ldots, n$.**
- **The subjects are independent, and**

$$\begin{aligned} E(Y_i) &= \mu_i \\ V(Y_i) &= \Sigma. \end{aligned}$$

- **Define**

$$Y = \begin{bmatrix} Y_1' \\ Y_2' \\ \vdots \\ Y_n' \end{bmatrix} \textbf{ and } M = \begin{bmatrix} \mu_1' \\ \mu_2' \\ \vdots \\ \mu_n' \end{bmatrix}.$$

  - Y **is a** $n \times m$ **random matrix.**
  - **The rows of** Y **correspond to subjects.**
  - **The columns of** Y **correspond to variables.**
  - M **is a** $n \times m$ **matrix such that**

$$E(Y) = M.$$

- **The covariance matrix of** Y **is defined as:**

$$\begin{aligned} V(Y) &\equiv V[\mathbf{vec}(Y')] \\ &= E\{[\mathbf{vec}(Y') - \mathbf{vec}(M')][\mathbf{vec}(Y') - \mathbf{vec}(M')]'\} \\ &= I_n \otimes \Sigma \end{aligned}$$

5.3 Distributions

5.3.1 Matrix Normal Distribution

1. Y **is a** $n \times m$ **matrix. Usually,**

   - **Rows of** Y **correspond to subjects.**
   - **Columns of** Y **correspond to variables.**
   - Y_i **are the** m **measurements for subject** i**.**
   - $E(Y_i) = \mu_i$**.**
   - **Set**

$$Y = \begin{bmatrix} Y_1' \\ \vdots \\ Y_n' \end{bmatrix} \text{ and } M = \begin{bmatrix} \mu_1' \\ \vdots \\ \mu_n' \end{bmatrix}.$$

   - **See previous example.**

2. Special form for the covariance.

- **Let Σ be a $m \times m$ pds matrix.**
- **Let Φ be a $n \times n$ pds matrix.**

If

$$\begin{aligned} V(Y_i) &= \phi_{ii}\Sigma \\ Cov(Y_i, Y_j) &= E[(Y_i - \mu_i)(Y_j - \mu_j)'] = \phi_{ij}\Sigma, \end{aligned}$$

then

$$V(Y) \equiv V[\mathbf{vec}(Y')] = \Phi \otimes \Sigma.$$

If the subjects are mutually independent,

$$\Phi = I_n.$$

3. The matrix normal pdf for Y is:

$$[Y|M,\Phi,\Sigma] = N_{n\times m}(Y|M,\Phi,\Sigma)$$

$$= (2\pi)^{-\frac{mn}{2}}|\Phi|^{-\frac{m}{2}}|\Sigma|^{-\frac{n}{2}}$$

$$\times \exp\left\{-\frac{1}{2}\mathbf{tr}\left[\Sigma^{-1}(Y-M)'\Phi^{-1}(Y-M)\right]\right\}.$$

4. The matrix multivariate density can also be written by stacking the rows of Y. Define

$$Y^* = \mathbf{vec}(Y') \textbf{ and } M^* = \mathbf{vec}(M').$$

Then

$$[Y^*|M^*, \Phi, \Sigma] = N_{mn}(Y^*|M^*, \Phi \otimes \Sigma)$$

$$= (2\pi)^{-\frac{mn}{2}} |\Phi \otimes \Sigma|^{-\frac{1}{2}}$$

$$\times \exp\left\{-\frac{1}{2}(Y^* - M^*)'(\Phi \otimes \Sigma)^{-1}(Y^* - M^*)\right\}$$

5.

Are the two pdfs the same?

Use Mini–Fact (5i) on page (170):

$$|\Phi \otimes \Sigma|^{-\frac{1}{2}} = |\Phi|^{-\frac{m}{2}}|\Sigma|^{-\frac{n}{2}}.$$

Use Mini–Fact (5o) on page (171):

$$\begin{aligned}&\mathbf{tr}\left[\Sigma^{-1}(Y-M)'\Phi^{-1}(Y-M)I\right]\\ &= \ (Y^* - M^*)' \, (\Phi \otimes \Sigma)^{-1} \, (Y^* - M^*) \, .\end{aligned}$$

6. If $V(Y)$ is not $\Phi \otimes \Sigma$, then the matrix normal for Y is defined by vec(Y') being multivariate normal.

5.3.2 Wishart Distribution

(Arnold Zellner, *An Introduction to Bayesian Inference and Econometrics*, 1971, John Wiley & Sons, ISBN 0–471–98165–6)

1. X **is a** $m \times m$ **positive definite, symmetric matrix.**

2. X **has a Wishart distribution if its density is:**

$$
\begin{aligned}
[X|v, G] &= W_m(X|v, G) \\
&= k \frac{|X|^{(v-m-1)/2}}{|G|^{v/2}} \exp\left[-\frac{1}{2}\mathbf{tr}\left(G^{-1}X\right)\right] \\
k^{-1} &= 2^{vm/2}\pi^{m(m-1)/4}\prod_{i=1}^{m}\Gamma\left[(v+1-i)/2\right]
\end{aligned}
$$

for $v \geq m$, **and** G **is a positive definite, symmetric,** $m \times m$ **matrix.**

3. **The Wishart is the multivariate generalization of the Gamma distribution.**

4. **I will call v the degrees of freedom, and G the scale matrix.**

5. **Moments:**

$$\begin{aligned} E(X) &= vG \\ V(x_{ij}) &= v(g_{ij}^2 + g_{ii}g_{jj}) \\ Cov(x_{ij}, x_{kl}) &= v(g_{ik}g_{jl} + g_{il}g_{jk}) \end{aligned}$$

6. **If $z_1, \ldots, z_v$ are iid $N_m(z|0, \Sigma)$, then the distribution of**

$$X = \sum_{i=1}^{v} z_i z_i'$$

is $W_m(X|v, \Sigma)$.

$S = \frac{1}{v} \Sigma_{i=1}^{v} z_i z_i'$ **is $W_m(S|v, \Sigma/v)$.**

7. The Standard Wishart sets $G = I$.

8. If $Y = W_m(Y|v, I)$ and if $X = C'YC$, then

$$X = W_m(X|v, C'C).$$

9. Bartlett's Decomposition

(Brian Ripley, *Stochastic Simulation,* pp. 99–100, 1987, John Wiley & Sons, ISBN 0271-6356)

is used to generate the standard Wishart.

5.3.3 Inverted Wishart Distribution

1. **Y, a $m \times m$, positive definite, symmetric matrix has the inverted Wishart distribution with density:**

$$
\begin{aligned}
[Y|v, H] &= IW_m(v, H) \\
&= k \frac{|H|^{v/2}}{|Y|^{(v+m+1)/2}} \exp\left\{-\frac{1}{2}\mathbf{tr}\left(Y^{-1}H\right)\right\} \\
k^{-1} &= 2^{vm/2}\pi^{m(m-1)/4} \prod_{i=1}^{m} \Gamma\left[(v+1-i)/2\right]
\end{aligned}
$$

where $v \geq m$ and H is a $m \times m$, positive definite, symmetric matrix.

2. **If X is $W_m(X|v, G)$, then $Y = X^{-1}$ is $IW_m(Y|v, G^{-1})$.**

3. I wrote a subroutine in plbam.src that returns a Wishart and Inverted Wishart. Its calling statement is

$$\{\mathbf{w},\ \mathbf{wi}\ \} = \mathbf{Wishart(m,v,G);}$$

where

$$\begin{aligned}[\mathbf{w}] &= W_m(\mathbf{w}|v, G)\\ [\mathbf{wi}] &= IW_m(\mathbf{wi}|v, G^{-1}).\end{aligned}$$

5.4 Multivariate Regression Model

1. **For subject i:**

$$Y_i = B'x_i + \epsilon_i \text{ for } i = 1, \ldots, n$$

where

- **there are n subjects**
- **and m dependent observations for each subject;**
- Y_i **is a m vector for $i = 1, \ldots, n$;**
- x_i **is a k vector for $i = 1, \ldots, n$;**
- B **is a $k \times m$ matrix of regression coefficients;**
- $[\epsilon_i]$ **is $N_m(\epsilon_i|0, \Sigma)$;**
- **the error terms are mutually independent and independent of $\{x_i\}$.**

2. Define

$$Y = \begin{bmatrix} Y_1' \\ Y_2' \\ \vdots \\ Y_n' \end{bmatrix}; \quad X = \begin{bmatrix} x_1' \\ x_2' \\ \vdots \\ x_n' \end{bmatrix} \textbf{ and } U = \begin{bmatrix} \epsilon_1' \\ \epsilon_2' \\ \vdots \\ \epsilon_n' \end{bmatrix}.$$

- Y **is a** $n \times m$ **matrix.**
- **The rows correspond to subjects.**
- **The columns correspond to variables.**
- X **is the** $n \times k$ **design matrix.**
- U **is the** $n \times m$ **error matrix with**

$$\begin{aligned} E(U) &= 0 \\ V(U) &= V[\mathbf{vec}(U')] = I_n \otimes \Sigma \end{aligned}$$

3. The multivariate regression model is:

$$Y = XB + U.$$

The pdf of Y given B and Σ is:

$$[Y|B,\Sigma] = N_{n\times m}(Y|XB, I_n, \Sigma)$$

$$\propto \ |\Sigma|^{-\frac{n}{2}} \exp\left\{-\frac{1}{2}\mathbf{tr}\left[\Sigma^{-1}(Y-XB)'(Y-XB)\right]\right\}$$

This version is used in computing the full conditional of Σ.

4. Set $Y^* = \mathbf{vec}(Y')$ and $B^* = \mathbf{vec}(B')$.

Another representation can be derived from:

$$\begin{aligned}
\mathbf{vec}(Y') &= \mathbf{vec}(B'X') + \mathbf{vec}(U') \\
&= (X \otimes I_m)\mathbf{vec}(B') + \mathbf{vec}(U') \\
Y^* &= (X \otimes I_m)B^* + \epsilon^* \\
E(\epsilon^*) &= 0 \\
V(\epsilon^*) &= I_n \otimes \Sigma
\end{aligned}$$

The pdf of Y^* is:

$$[Y^*|B^*, \Sigma] = N_{nm}(Y^*|[X \otimes I_m]B^*, I_n \otimes \Sigma)$$

$$\propto \; |\Sigma|^{-\frac{n}{2}} \exp\left\{-\frac{1}{2}\left[Y^* - (X \otimes I_m)B^*\right]' \left(I_n \otimes \Sigma^{-1}\right) \left[Y^* - (X \otimes I_m)B^*\right]\right\}.$$

This version is used in computing

the full conditional of B or B^*.

5. If X has full rank, the MLEs are:

$$\hat{B} = (X'X)^{-1}X'Y$$

$$\hat{\Sigma} = \frac{1}{n}(Y - X\hat{B})'(Y - X\hat{B})$$

In Gauss,

```
bhat = invpd(x'x)*x'y;

sighat = (y-x*bhat)'(y-x*bhat)/n;
```

5.5 Conjugate Model

This section present the analysis of the multivariate normal model with conjugate prior distributions. The analysis has analytical expressions.

5.5.1 Conjugate Prior Distributions

The conjugate prior distribution is similar to that for linear regression: the prior distribution for the regression coefficients depend on the variance of the error terms.

$$
\begin{aligned}
[B|\Sigma] &= N_{k\times m}(B|U_0, V_0, \Sigma) \\
&\propto |V_0|^{-\frac{m}{2}}|\Sigma|^{-\frac{k}{2}} \exp\left\{-\frac{1}{2}\mathbf{tr}\left[\Sigma^{-1}(B-U_0)'V_0^{-1}(B-U_0)\right]\right\} \\
[\Sigma] &= IW_m(\Sigma|f_0, G_0) \\
&\propto |\Sigma|^{-\frac{(f_0+m+1)}{2}} \exp\left[-\frac{1}{2}\mathbf{tr}\left(\Sigma^{-1}G_0^{-1}\right)\right]
\end{aligned}
$$

5.5.2 Posterior Distributions

$$
\begin{aligned}
[B|Y,\Sigma] &= N_{k\times m}(B|U_n, V_n, \Sigma) \\
V_n &= \left(X'X + V_0^{-1}\right)^{-1} \\
U_n &= V_n\left(X'Y + V_0^{-1}U_0\right) \\
[\Sigma|Y] &= IG_m(\Sigma|f_n, G_n) \\
f_n &= f_0 + n \\
G_n &= G_0 + \left(Y'Y + U_0'V_0^{-1}U_0 - U_n'V_n^{-1}U_n\right)^{-1}
\end{aligned}
$$

The key computation is combining the traces from the likelihood and prior distribution for B:

$$
\begin{aligned}
&\mathbf{tr}\left[\Sigma^{-1}(Y - XB)'(Y - XB)\right] + \mathbf{tr}\left[\Sigma^{-1}(B - U_0)'V_0^{-1}(B - U_0)\right] \\
&= \mathbf{tr}\left[Y'Y + U_0'V_0^{-1}U_0 + B'\left(X'X + V_0^{-1}\right)B\right. \\
&- \left. B'\left(X'Y + V_0^{-1}U_0\right) - \left(Y'X + U_0'V_0^{-1}\right)B\right] \\
&= \mathbf{tr}\left[Y'Y + U_0'V_0^{-1}U_0 - U_n'V_n^{-1}U_n + (B - U_n)'V_n^{-1}(B - U_n)\right]
\end{aligned}
$$

If one needs to generate from these distributions, then generate Σ from an inverted Wishart. Given Σ generate B, which has compact code in a matrix based languages, such as GAUSS:

$$B = U_n + A'Z * D$$

where $A = \mathbf{chol}(V_n)$, the upper-triangular Cholesky decomposition as in Gauss; $D = \mathbf{chol}(\Sigma)$; and Z is a $k \times m$ matrxix of iid standard normal random deviates. Obviously, B will have the correct mean, U_n. A check on the covariance matrix gives:

$$\begin{aligned}
\mathbf{var}(A'ZD) &= \mathbf{var}[\mathbf{vec}(D'Z'A)] \\
&= A'A \otimes D'D \\
&= V_n \otimes \Sigma
\end{aligned}$$

5.6 Non-conjugate Model

We re-analyze the multivariate normal model without using conjugate prior distributions. In this case, one needs to use MCMC.

5.6.1 Prior Distributions

1. **$B^* = \mathbf{vec}(B')$ is $N_{km}(B^*|u_0, V_0)$:**

$$[B^*|u_0, V_0] \propto \exp\left\{-\frac{1}{2}(B^* - u_0)'V_0^{-1}(B^* - u_0)\right\}.$$

2. **Σ is $IW_m(\Sigma|f_0, G_0^{-1})$:**

$$[\Sigma|f_0, G_0] \propto |\Sigma|^{-(f_0+m+1)/2} \exp\left\{-\frac{1}{2}\mathbf{tr}\left(\Sigma^{-1}G_0^{-1}\right)\right\}.$$

5.6.2 Full Conditionals

1. **Generate B.**

   Define $Y^* = \mathbf{vec}(Y')$ and $B^* = \mathbf{vec}(B')$.

$$
\begin{aligned}
[B^*|Y^*,\Sigma] &\propto [Y^*|B^*,\Sigma][B^*] \\
\\
&\propto N_{nm}(Y^*|[X\otimes I_m]B^*, I_n\otimes\Sigma) \\
&\times N_{km}(B^*|u_0,V_0)
\end{aligned}
$$

- **Expand the squares in B^*, combine terms, and compete the squares.**
- **It works just like the linear regression model on page (88).**
- **Use Kronecker product algebra:**

$$(X \otimes I_m)'(I_n \otimes \Sigma^{-1}) = X' \otimes \Sigma^{-1}$$

$$(X \otimes I_m)'(I_n \otimes \Sigma^{-1})(X \otimes I_m) = X'X \otimes \Sigma^{-1}$$

Full Conditional of B^*:

$$[B^*|Y^*, \Sigma] = N_{km}(B^*|u_n, V_n)$$

$$V_n = \left[\left(X'X \otimes \Sigma^{-1}\right) + V_0^{-1}\right]^{-1}$$

$$u_n = V_n \left[\left(X' \otimes \Sigma^{-1}\right) Y^* + V_0^{-1} u_0\right]$$

2. Generate Σ.

$$\begin{aligned}
[\Sigma|Y,B] &\propto [Y|B,\Sigma][\Sigma] \\
\\
&\propto N_{n\times m}(Y|XB,I_n,\Sigma) \\
&\times IW_m(\Sigma|f_0,G_0^{-1})
\end{aligned}$$

Full conditional of Σ:

$$[\Sigma|Y,B] = IW_m(\Sigma|f_n, G_n^{-1})$$

$$f_n = f_0 + n$$

$$G_n^{-1} = G_0^{-1} + (Y - XB)'(Y - XB).$$

So, Σ^{-1} **is** $W_m(\Sigma^{-1}|f_n, G_n)$**.**

The calling statement in Gauss is:

```
{sigmai, sigma} = wishart(mvar,f0n,gn);
```

5.7 Summary

1. **Multivariate regression is a “easy” extension of multiple regression.**
2. **It requires some specialized matrix algebra to simplify the “book keeping.”**
3. **Other models heavily rely on components of multivariate regression.**

Chapter 6

HB Regression: Interaction Model

Outline

1. **Objectives**

2. **Model**

3. **Priors**

4. **Joint Distribution**

5. **Full Conditionals**

6. **Special Case: Common Design Matrix**

6.1 Objectives

1. **Hierarchical Bayes (HB) models allow for multiple sources of uncertainty.**

2. **Random effects models are a special case.**

3. **Simplest yet powerful case:**

   - **Within–Subject Model:**
     A linear regression model that relates covariates to individual–level regression coefficients.

   - **Between–Subject Model:**
     A multivariate regression model that describes the variation or heterogeneity in the individual–level coefficients across the population of customers.

4. Example:

- **All households have the same structural form for their sales response function.**
- **Households are allowed to have their own preferences and responses to the marketing mix. That is, they have household–level coefficients.**
- **The household–level coefficients may be related to demographics such as household income, family size, and age and education of head of household. E.g., high–income households are less price sensitive than low–income households, and older households are less sensitive to advertising than younger households.**

6.2 Model

1. **Within–subject model:**

$$Y_i = X_i\beta_i + \epsilon_i \textbf{ for } i = 1, \ldots, n$$

where

- **there are** n **subjects and**
- m_i **observations for subject** i**;**
- Y_i **is a** m_i **vector;**
- X_i **is a** $m_i \times p$ **design matrix;**
- β_i **is a** p **vector of individual–level regression coefficients; and**
- ϵ_i **is a** m_i **vector of error terms with pdf**

$$[\epsilon_i|\sigma_i] = N_{m_i}(\epsilon_i|0, \sigma^2 I_{m_i}).$$

2. Between–subjects model:

$$\beta_i = \Theta' z_i + \delta_i \text{ for } i = 1, \ldots, n$$

where

- z_i **is a** q **vector of covariates for subject** i**.**
- Θ **is a** $q \times p$ **matrix of regression coefficients.**
- δ_i **is a** p **vector of error terms with pdf:**

$$[\delta_i|\Lambda] = N_p(\delta_i|0, \Lambda).$$

The between–subjects model describes the heterogeneity in the subject–level coefficients across the population.

3. Matrix version of the between–subjects model:

$$B = Z\Theta + \Delta,$$

where

$$B = \begin{bmatrix} \beta_1' \\ \vdots \\ \beta_n' \end{bmatrix}, \ Z = \begin{bmatrix} z_1' \\ \vdots \\ z_n' \end{bmatrix} \text{ and } \Delta = \begin{bmatrix} \delta_1' \\ \vdots \\ \delta_n' \end{bmatrix}.$$

- B **is a** $n \times p$ **matrix of the individual–level coefficients.**
- Z **is a** $n \times q$ **matrix of covariates.**
- Θ **is a** $q \times p$ **matrix of regression coefficients.**
- Δ **is a** $n \times p$ **matrix of error terms with pdf:**

$$[\Delta|\Lambda] = N_{n\times p}(\Delta|0, I_n, \Lambda).$$

- **The pdf of** B **is:**

$$[B|\Theta, \Lambda] = N_{n\times p}(B|Z\Theta, I_n \otimes \Lambda).$$

- **This model is the same as the multivariate regression model on page (186).**

4. Why "Interaction" Model?

$$\begin{aligned} Y_i &= X_i\beta_i + \epsilon_i \\ \beta_i &= \Theta' z_i + \delta_i \end{aligned}$$

$$\begin{aligned} Y_i &= X_i(\Theta' z_i + \delta_i) + \epsilon_i \\ &= X_i\Theta' z_i + X_i\delta_i + \epsilon_i. \end{aligned}$$

$X_i\Theta' z_i$ is a m_i vector.

Use Mini–Fact (5l) on page (171):

$$X_i\Theta' z_i = \mathbf{vec}(X_i\Theta' z_i) = (z_i' \otimes X_i)\mathbf{vec}(\Theta').$$

Define:

$$X_i^* = z_i' \otimes X_i\textbf{;}\;\; \Theta^* = \mathbf{vec}(\Theta') \textbf{ and } \epsilon_i^* = X_i\delta_i + \epsilon_i.$$

Note that

$$[\epsilon_i^*] = N_{m_i}(\epsilon_i^*|0, \sigma^2 I_{m_i} + X_i\Lambda X_i').$$

Within–subject Model:

$$Y_i = X_i^* \Theta^* + \epsilon_i^*$$

where

- **The design matrix**

  $$X_i^* = z_i' \otimes X_i$$

  contains all of the cross products between the variables in X_i and z_i.

- Θ^* **is a pq vector of regression coefficients that do not depend on the subject.**

- ϵ_i^* **has a non-zero correlation structure:**

  $$[\epsilon_i^*] = N_{m_i}(\epsilon_i^* | 0, \sigma^2 I_{m_i} + X_i \Lambda X_i').$$

6.3 Priors

1. **The prior pdf for σ^2 is:**

$$[\sigma^2|r_0, s_0] = IG\left(\sigma^2|\frac{r_0}{2}, \frac{s_0}{2}\right).$$

2. **The prior pdf for $\Theta^* = \mathbf{vec}(\Theta')$ is:**

$$[\Theta^*|u_0, V_0] = N_{pq}(\Theta^*|u_0, V_0).$$

3. **The prior pdf for Λ is:**

$$[\Lambda|f_0, G_0^{-1}] = IW_m(\Lambda|f_0, G_0^{-1}).$$

6.4 Joint Distribution

$$\prod_{i=1}^{n} \left\{ [Y_i|\beta_i, \sigma^2][\beta_i|\Theta, \Lambda] \right\} [\sigma^2|r_0, s_0][\Theta^*|u_0, V_0][\Lambda|f_0, G_0] =$$

$$= \prod_{i=1}^{n} N_{m_i}(Y_i|X_i\beta_i, \sigma^2 I_{m_i}) N_{n\times p}(B|Z\Theta, I_n, \Lambda)$$

$$\times \; IG\left(\sigma^2|\frac{r_0}{2}, \frac{s_0}{2}\right) N_{pq}(\Theta^*|u_0, V_0) IW(\Lambda|f_0, G_0^{-1})$$

6.5 Full Conditionals

1. **Full Conditional for β_i.**

$$
\begin{aligned}
[\beta_i| \textbf{ Rest }] &\propto [Y_i|\beta_i, \sigma^2][\beta_i|\Theta, \Lambda] \\
&\propto N_{m_i}(Y_i|X_i\beta_i, \sigma^2 I_{m_i})N_p(\beta_i|\Theta' z_i, \Lambda)
\end{aligned}
$$

2. **Generate β_i.**

$$
[\beta_i| \textbf{ Rest }] = N_p(\beta_i|u_i, V_i)
$$

$$
V_i = \left(\frac{1}{\sigma^2}X_i'X_i + \Lambda^{-1}\right)^{-1}
$$

$$
u_i = V_i\left(\frac{1}{\sigma^2}X_i'Y_i + \Lambda^{-1}\Theta' z_i\right),
$$

which is similar to the full conditional of β on page (115) for the linear regression model.

3. Suppose X_i has full rank.

- **The MLE of β_i is:**

$$\hat{\beta}_i = (X_i' X_i)^{-1} X_i' Y_i.$$

- **The conditional, posterior mean of β_i is:**

$$
\begin{aligned}
& E(\beta_i | Y_i, \Theta, \sigma, \Lambda) \\
&= \left(\frac{1}{\sigma^2} X_i' X_i + \Lambda^{-1} \right)^{-1} \left(\frac{1}{\sigma^2} (X_i' X_i) \hat{\beta}_i + \Lambda^{-1} \Theta' z_i \right) \\
&= W \hat{\beta}_i + (I_p - W) \Theta' z_i
\end{aligned}
$$

$$W = \left(\frac{1}{\sigma^2} X_i' X_i + \Lambda^{-1} \right)^{-1} \left(\frac{1}{\sigma^2} X_i' X_i \right),$$

which is a convex combination of

– **within–subject MLE for β_i, and**

– **its between–subjects estimate $\Theta' z_i$.**

- **The Bayes estimator "shrinks" the individual–level MLE towards the between–subjects estimator.**
- **The amount of shrinkage depends on the relative precision of the two estimators.**
- **As m_i increases, $W \to I_p$ under mild conditions on X_i, so that the Bayes estimator puts more weight on the within–subject estimator and less on the between-subjects estimator.**

4. Full conditional for σ^2.

$$
\begin{aligned}
[\sigma^2 | \textbf{ Rest }] &\propto \prod_{i=1}^{n} [Y_i|\beta_i, \sigma^2][\sigma^2|r_0, s_0] \\
&\propto \prod_{i=1}^{n} N_{m_i}(Y_i|X_i\beta_i, \sigma^2 I_{m_i}) IG\left(\sigma^2|\frac{r_0}{2}, \frac{s_0}{2}\right).
\end{aligned}
$$

5. Generate σ^2

$$
\begin{aligned}
[\sigma^2 | \textbf{ Rest }] &= IG\left(\sigma^2|\frac{r_n}{2}, \frac{s_n}{2}\right) \\
r_n &= r_0 + \sum_{i=1}^{n} m_i \\
s_n &= s_0 + \sum_{i=1}^{n} (Y_i - X_i\beta_i)'(Y_i - X_i\beta_i),
\end{aligned}
$$

which is similar to the full conditional for σ^2 on page (117) for the linear regression model.

6. Full conditional for Θ.

$$
\begin{aligned}
[\Theta|\ \textbf{Rest}\ ] &\propto \prod_{i=1}^{n}[\beta_i|\Theta,\Lambda][\Theta] \\
&\propto N_{np}(B^*|[Z\otimes I_p]\Theta^*, I_n\otimes\Lambda)N_{pq}(\Theta^*|u_0,V_0)
\end{aligned}
$$

7. Generate $\Theta^* = \mathbf{vec}(\Theta')$.

$$
[\Theta^*|\ \textbf{Rest}\ ] = N_{pq}(\Theta^*|u_n,V_n)
$$

$$
V_n = \left[\left(Z'Z\otimes\Lambda^{-1}\right)+V_0^{-1}\right]^{-1}
$$

$$
u_n = V_n\left[\left(Z'\otimes\Lambda^{-1}\right)B^* + V_0^{-1}u_0\right],
$$

which is similar to the full conditional for B^* on page (195) for the multivariate regression model.

8. Full Conditional for Λ:

$$[\Lambda| \textbf{ Rest }] \propto \prod_{i=1}^{n}[\beta_i|\Theta, \Lambda][\Lambda|f_0, G_0]$$

$$\propto N_{n\times p}(B|Z\Theta, I_n, \Lambda)IW_m(\Lambda|f_0, G_0^{-1})$$

9. Generate Λ.

$$[\Lambda| \textbf{ Rest }] = IW_m(\Lambda|f_n, G_n^{-1})$$

$$f_n = f_0 + n$$

$$G_n^{-1} = G_0^{-1} + (B - Z\Theta)'(B - Z\Theta),$$

which is similar to the full conditional for Σ on page (197) for the multivariate regression model.

6.6 Common Design Matrix

1. $X_i = X$ **and** $m_i = m$ **for all** $i = 1, \ldots, n$**. Then**

$$Y_i = X\beta_i + \epsilon_i$$

$$Y = BX' + U$$

$$Y = \begin{bmatrix} Y_1' \\ \vdots \\ Y_n' \end{bmatrix}; \; B = \begin{bmatrix} \beta_1' \\ \vdots \\ \beta_n' \end{bmatrix} \textbf{; and } U = \begin{bmatrix} \epsilon_1' \\ \vdots \\ \epsilon_n' \end{bmatrix}.$$

2. Full conditional of B is:

$$[B|Y,\sigma,\Theta,\Lambda] \;=\; N_{n\times p}(B|\bar{B}, I_n, V)$$

$$V \;=\; \left(\frac{1}{\sigma^2}X'X + \Lambda^{-1}\right)^{-1}$$

$$\bar{B} \;=\; \left(YX + Z\Theta\Lambda^{-1}\right)V$$

3. B can be generated in Gauss in one line:

$$B = (Y * X + Z * \Theta * \Lambda^{-1}) * V + \mathbf{rndn(n,p)} * V^{\frac{1}{2}}$$

where

$$V^{\frac{1}{2}} = \mathbf{chol}(V).$$

4. Full conditional of σ^2

$$[\sigma^2|Y, B, \Theta, \Lambda] = IG\left(\sigma^2|\frac{r_n}{2}, \frac{s_n}{2}\right)$$

$$r_n = r_0 + nm$$

$$s_n = s_0 + \mathbf{tr}[(Y - BX')'(Y - BX')],$$

although this is an inefficient method of computing s_n.

6.7 Different Design Matrices

If each subject has a different design matrix, then the data structures become more complex.

1. **Stack the independent and dependent variables.**

$$\mathbf{ydata} = \begin{bmatrix} Y_1 \\ \vdots \\ Y_n \end{bmatrix} \text{ and } \mathbf{xdata} = \begin{bmatrix} X_1 \\ \vdots \\ X_n \end{bmatrix}.$$

2. **Use pointers to indicate the rows of xdata and ydata for each subject:**
   - **iptxy is a n by 2 matrix.**
   - **iptxy[i,1] = starting row for subject i.**
   - **iptxy[i,2] = ending row for subject i.**
   - **xi = xdata[iptxy[i,1]:iptxy[i,2],.];**
   - **yi = ydata[iptxy[i,1]:iptxy[i,2],.];**

6.8 Examples

6.8.1 Simulated Data

- **100 subjects**
- **10 observations per subject**
- **3 predictor X variables**
- **2 predictor Z variables.**
- **Common design matrix.**

- **Parameter Values:**

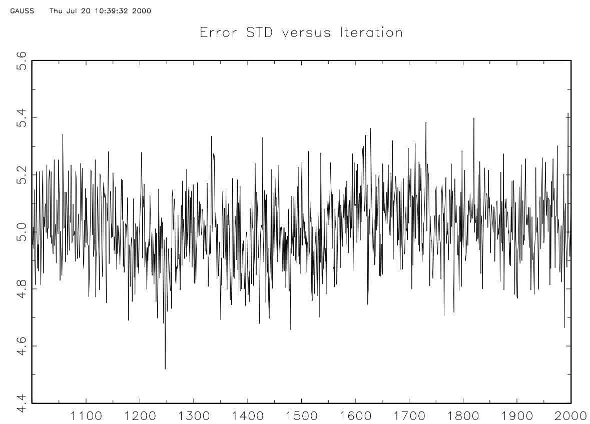
True Λ

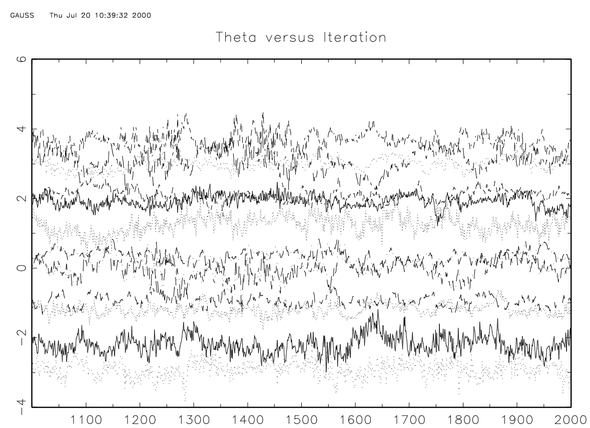
$$\left|\begin{array}{rrrr} \mathbf{0.250} & \mathbf{0.125} & \mathbf{-0.250} & \mathbf{0.0250} \\ \mathbf{0.125} & \mathbf{1.062} & \mathbf{-0.125} & \mathbf{1.512} \\ \mathbf{-0.250} & \mathbf{-0.125} & \mathbf{2.500} & \mathbf{-0.775} \\ \mathbf{0.0250} & \mathbf{1.512} & \mathbf{-0.775} & \mathbf{6.502} \end{array}\right|$$

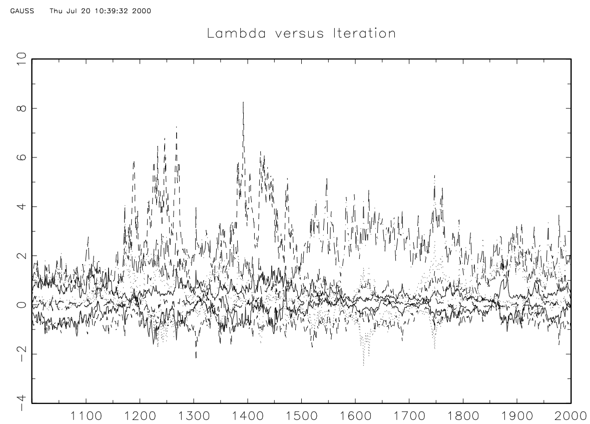
True Θ

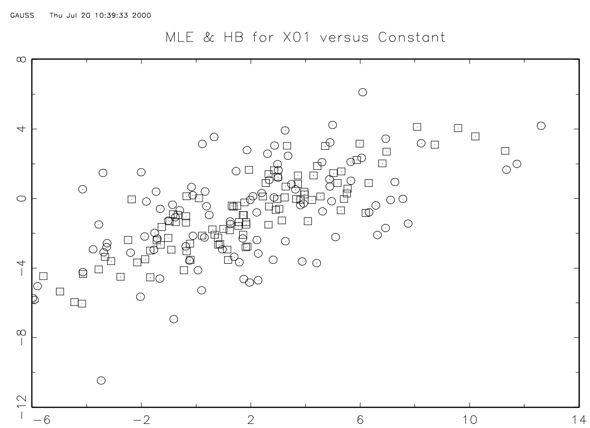
$$\left|\begin{array}{rrrr} \mathbf{2} & \mathbf{-1} & \mathbf{-3} & \mathbf{4} \\ \mathbf{-1} & \mathbf{0} & \mathbf{-2} & \mathbf{3} \\ \mathbf{3} & \mathbf{2} & \mathbf{1} & \mathbf{0} \end{array}\right|$$

True σ = 5


GAUSS Thu Jul 20 10:39:32 2000
Error STD versus Iteration
5.6
5.4
5.2
5.0
4.8
4.6
4.4
1100 1200 1300 1400 1500 1600 1700 1800 1900 2000


GAUSS Thu Jul 20 10:39:32 2000
Theta versus Iteration
6
4
2
0
−2
−4
1100 1200 1300 1400 1500 1600 1700 1800 1900 2000


GAUSS Thu Jul 20 10:39:32 2000
Lambda versus Iteration
10
8
6
4
2
0
−2
−4
1100
1200
1300
1400
1500
1600
1700
1800
1900
2000


GAUSS Thu Jul 20 10:39:33 2000
MLE & HB for X01 versus Constant
8
4
0
−4
−8
−12
−6
−2
2
6
10
14

```
MCMC Analysis

Total number of MCMC iterations              = 2000.00000
Number of iterations used in the analysis    = 1000.00000
Number in transition period                  = 1000.00000
Number of iterations between saved iterations =    0.00000

Number of subjects                  =  100.00000
Number of observations per subject =   10.00000
Number of dependent variables X    =    3.00000  (excluding intercept)
Number of dependent variables Z    =    2.00000  (excluding intercept)

Independent variables in first level equation:
Y_i = X*beta_i + epsilon_i

       Summary Statistics for X
Variable        Mean       STD       MIN       MAX
Constant     1.00000   0.00000   1.00000   1.00000
X01         -0.11050   1.27044  -2.12788   2.20340
X02         -0.46143   0.94077  -2.15512   0.98589
X03         -0.09866   0.65677  -1.03349   1.12585


Independent variables in second level equation:
beta_i = Theta*z_i + delta_i

      Summary Statistics for Z
Variable        Mean       STD       MIN       MAX
Constant     1.00000   0.00000   1.00000   1.00000
Z01          0.11417   1.02464  -2.84694   3.01263
Z02          0.02264   1.04794  -2.37418   2.44308


----------------------------------------------------------
```

```
Fit Measures:
HB Predictive Correlation (Mulitple R)     =    0.82457
HB R-Square                                =    0.67991
ML Predictive Correlation (Mulitple R)     =    0.89144
ML R-Square                                =    0.79466

Estimation of the error STD sigma
True Sigma     =    5.00000
MLE            =    3.80709
Posterior Mean =    5.01259
Posterior STD  =    0.12984

----------------------------------------------------------
Statistics for Individual-Level Regression Coefficients
True Beta
Variable        Mean       STD
Constant     1.88851   3.28475
X01         -0.98853   2.16424
X02         -3.06624   2.73293
X03          4.15631   4.21870

MLE of Beta
Variable     MeanMLE    StdMLE
Constant     1.84409   3.80061
X01         -0.84922   2.83973
X02         -3.16501   3.81054
X03          4.01565   4.61408

HB Estimates of Beta
Variable    PostMean   PostSTD
Constant     1.85342   3.42690
X01         -0.85958   2.33013
X02         -3.13697   2.95597
X03          3.98942   3.44017
-----------------------------------------------------------
```

```
Comparison of True Beta to Individual Level Estimates
Component    1.00000
Correlation between true and HB  =    0.98804
RMSE between true and HB         =    0.51925

Correlation between true and MLE =    0.87768
RMSE between true and MLE        =    1.81354

Component    2.00000
Correlation between true and HB  =    0.90406
RMSE between true and HB         =    0.97365

Correlation between true and MLE =    0.79723
RMSE between true and MLE        =    1.71425

Component    3.00000
Correlation between true and HB  =    0.88952
RMSE between true and HB         =    1.28619

Correlation between true and MLE =    0.77942
RMSE between true and MLE        =    2.38908

Component    4.00000
Correlation between true and HB  =    0.82364
RMSE between true and HB         =    2.39809

Correlation between true and MLE =    0.86105
RMSE between true and MLE        =    2.35154

-----------------------------------------------------------
```

```
HB Estimates of Theta
True Theta
            Constant       X01       X02       X03
Constant     2.00000  -1.00000  -3.00000   4.00000
Z01         -1.00000   0.00000  -2.00000   3.00000
Z02          3.00000   2.00000   1.00000   0.00000

Posterior Mean of Theta
            Constant       X01       X02       X03
Constant     1.92295  -0.94529  -2.91245   3.64369
Z01         -1.19826   0.33663  -2.19426   3.09412
Z02          2.99088   2.10593   1.25096  -0.02394

Posterior STD of Theta
            Constant       X01       X02       X03
Constant     0.41884   0.40341   0.49638   0.51457
Z01          0.40569   0.43122   0.51118   0.56754
Z02          0.41302   0.41631   0.50330   0.53965
-------------------------------------------------------
```

```
HB Estimate of Lambda
True Lambda
            Constant       X01       X02       X03
Constant     0.25000   0.12500  -0.25000   0.02500
X01          0.12500   1.06250  -0.12500   1.51250
X02         -0.25000  -0.12500   2.50000  -0.77500
X03          0.02500   1.51250  -0.77500   6.50250

Posterior Mean of Lambda
            Constant       X01       X02       X03
Constant     0.55059   0.02989  -0.65302  -0.29281
X01          0.02989   0.33414  -0.18733   0.16148
X02         -0.65302  -0.18733   2.02920   0.60925
X03         -0.29281   0.16148   0.60925   1.60154

Posterior STD of Lambda
            Constant       X01       X02       X03
Constant     0.28660   0.14160   0.30269   0.33139
X01          0.14160   0.25645   0.44023   0.46170
X02          0.30269   0.44023   0.81322   0.70251
X03          0.33139   0.46170   0.70251   1.43604
==================================================
```

6.8.2 Metric Conjoint Study

Lenk, DeSarbo, Green, and Young (1996) *Marketing Science*
MBA Computer Survey

Attributes and Their Level

A.	Telephone Service Hotline				H.	Color of Unit			
	–1	=	No			–1	=	Beige	
	1	=	Yes			1	=	Black	
B.	Amount of RAM				I.	Availability			
	–1	=	8 MB			–1	=	Mail order only	
	1	=	16 MB			1	=	Computer store only	
C.	Screen Size				J.	Warranty			
	–1	=	14 inch			–1	=	1 year	
	1	=	17 inch			1	=	3 year	
D.	CPU Speed				K.	Bundled Productivity Software			
	–1	=	50 MHz			–1	=	No	
	1	=	100 MHz			1	=	Yes	
E.	Hard Disk Size				L.	Money Back Guarantee			
	–1	=	340 MB			–1	=	None	
	1	=	730 MB			1	=	Up to 30 days	
F.	CD ROM/Multimedia				M.	Price			
	–1	=	No			–1	=	$2000	
	1	=	Yes			1	=	$3500	
G.	Cache								
	–1	=	128 KB						
	1	=	256 KB						

Y = Likelihood of purchase from 0 to 10.
HB model for σ_i^2 is IG.

Subject Level Covariates

FEMALE	=	0 if male and 1 if female
YEARS	=	Years of full–time work experience
OWN	=	1 if own or lease a microcomputer and 0 otherwise
TECH	=	1 if engineer, computer programmer or systems analysis 0 otherwise
APPLY	=	Number of categories of applications used with microcomputers
EXPERT	=	Sum of two self–evaluations. Each evaluation in on a five–point scale with 1 = Strongly Disagree, 3 = Neutral, and 5 = Strongly Agree. The first evaluation is, "When it comes to purchasing a microcomputer, I consider myself pretty knowledgeable about the microcomputer market." The second is, "When it comes to using a microcomputer, I consider myself pretty knowledgeable about microcomputers."

Number of subjects: 179
Number of calibration profiles per subject: 16
Number of validation profiles per subject: 4

Pooled Sample Aggregate Conjoint Analysis

R–Squared: 0.2437; Adjusted R–Squared: 0.2403
Standard Error of the Estimate: 2.439

Estimated Coefficients

	Variable	Coefficient	STD Error	T–Value	
Intercept		4.7301	0.0457	103.5541	**
A	Hotline	0.0946	0.0457	2.0715	*
B	RAM	0.3446	0.0457	7.5447	**
C	Screen Size	0.1924	0.0457	4.2119	**
D	CPU	0.3900	0.0457	8.5384	**
E	Hard Drive	0.1700	0.0457	3.7227	**
F	CD ROM	0.4920	0.0457	10.7705	**
G	Cache	0.0304	0.0457	0.6650	
H	Color	0.0262	0.0457	0.5733	
I	Availability	0.0772	0.0457	1.6893	
J	Warranty	0.1233	0.0457	2.6984	**
K	Software	0.1945	0.0457	4.2577	**
L	Guarantee	0.1114	0.0457	2.4385	*
M	Price	–1.1205	0.0457	–24.5298	**

ANOVA Table

Source	Sums of Squares	DF	Mean Square	F–Ratio
Regression	5488.019	13	392.0013	65.6**
Error	17030.347	2850	5.9756	
Total	22518.366	2863		

$^{*}p < 0.05$
$^{**}p < 0.01$

Sensitivity of Part–worths to Subject Level Covariates
(Posterior standard deviations are in parentheses.)

		Covariate						
	Variable	Intercept	FEMALE	YEARS	OWN	TECH	APPLY	EXPERT
Intercept		3.698**	–0.043	–0.111**	–0.158	–0.248	0.112*	0.167**
		(0.598)	(0.271)	(0.049)	(0.347)	(0.271)	(0.080)	(0.071)
A	Hotline	–0.047	0.226**	–0.002	–0.105	–0.019	–0.004	0.026*
		(0.195)	(0.087)	(0.016)	(0.115)	(0.084)	(0.025)	(0.023)
B	RAM	0.515**	–0.085	–0.003	0.139*	0.168*	0.043*	–0.065**
		(0.208)	(0.093)	(0.017)	(0.127)	(0.086)	(0.027)	(0.024)
C	Screen Size	0.058	–0.055	–0.009	0.044	0.109*	0.005	0.013
		(0.176)	(0.079)	(0.014)	(0.102)	(0.078)	(0.022)	(0.020)
D	CPU	–0.167	–0.101	–0.026*	0.158	0.171*	0.014	0.059*
		(0.279)	(0.131)	(0.023)	(0.172)	(0.127)	(0.038)	(0.033)
E	Hard Drive	0.013	–0.157*	–0.014	0.037	0.060	0.017	0.015
		(0.183)	(0.082)	(0.014)	(0.105)	(0.080)	(0.023)	(0.021)
F	CD ROM	0.591**	–0.164*	–0.010	–0.062	–0.075	0.015	0.001
		(0.251)	(0.113)	(0.020)	(0.148)	(0.107)	(0.033)	(0.029)
G	Cache	–0.266*	–0.043	–0.004	0.127*	0.019	–0.036*	0.049**
		(0.192)	(0.092)	(0.015)	(0.118)	(0.087)	(0.026)	(0.023)
H	Color	0.274*	–0.047	–0.004	0.017	–0.095*	–0.014	–0.019*
		(0.160)	(0.070)	(0.013)	(0.093)	(0.072)	(0.021)	(0.019)
I	Availability	0.157*	0.037	0.021*	0.138*	–0.097*	–0.011	–0.029*
		(0.156)	(0.068)	(0.013)	(0.092)	(0.070)	(0.021)	(0.018)
J	Warranty	–0.089	0.149*	0.024*	0.029	0.008	0.026*	–0.010
		(0.167)	(0.079)	(0.015)	(0.100)	(0.072)	(0.022)	(0.020)
K	Software	0.315*	0.009	–0.032**	–0.034	0.101*	0.010	–0.004
		(0.179)	(0.081)	(0.014)	(0.104)	(0.079)	(0.023)	(0.020)
L	Guarantee	0.023	0.031	0.025*	–0.117*	–0.081	0.013	0.004
		(0.185)	(0.085)	(0.015)	(0.107)	(0.081)	(0.025)	(0.022)
M	Price	–1.560**	0.385**	0.040*	–0.176	–0.064	0.001	0.041
		(0.398)	(0.173)	(0.031)	(0.233)	(0.170)	(0.052)	(0.047)

* The posterior mean is at least one posterior standard deviation from zero.
** The posterior mean is at least two posterior standard deviations from zero.

Validation Sample Performance Measures

				Market Shares			
Profiles	Cor$(Y, \hat{Y})$	RMSE_Y	Hit Rates	1	2	3	4
Individual-Level Ordinary Least Squares							
16	0.7152	1.998	0.637	0.115	0.099	0.325	0.462
Hierarchical Bayes							
16	0.7530	1.811	0.670	0.061	0.089	0.363	0.492
12	0.7425	1.851	0.687	0.039	0.078	0.335	0.548
8	0.7029	1.983	0.654	0.028	0.106	0.358	0.508
4	0.5877	2.262	0.587	0.028	0.045	0.285	0.643
Observed Market Shares				0.095	0.049	0.395	0.461

I randomly deleted profiles from the calibration sample. The individual–subject OLS estimates did not exist for everyone with only 12 profiles per subject. Using all 16 profiles, the HB predictions of the hold–out sample are better than the OLS. With only 8 profiles per subject, the HB predictions performed about as well as the OLS.

6.9 Stock Returns & Portfolio Analysis

Young and Lenk (1998) *Management Science*

Model

$$
\begin{aligned}
Y_i &= \beta_i X_i + \epsilon_i \\
\beta_i &= \Theta' z_{i,b} + \delta_{i,b} \\
\ln(\sigma^2) &= \psi' z_{i,s} + \delta_{i,s}
\end{aligned}
$$

1. **Response Variable: Monthly Returns**
2. **Predictor Variables:**
   - **Value weighted monthly returns of NYSE**
   - **Return for portfolio of lowest decile market value minus return for portfolio of highest decile market value on NYSE.**

3. Covariates:

- **Manufacturing 0/1**
- **Utility 0/1**
- **Finance 0/1**
- **Service 0/1**
- **Firm Size**

Data

1. **500 randomly selected securities.**
2. **19 four year intervals:**

   1955–1959, 1957–1961, . . . , 1991–1994
3. **First two years used for estimation: HB and Multiple Shrinkage (MS) Karolyi (1992)**
4. **Compare to OLS in second two years.**
5. **Form optimal portfolio using HB and MS.**

Parameter Estimates

1. **Utilities tend to have lower beta and idiosyncratic risk.**

2. **Larger firms have lower beta and idiosyncratic risk.**

3. **Firm size is strongly related with size sensitivity measure.**

MAE and Portfolio Certainty Equivalent

Out–of–sample Performance:

Estimate during first two years.

Compare to OLS during second two years.

19 time periods, 500 securities

		Mean	STD	# Wins
Intercept	HB	1.57	0.30	18
	MS	1.64	0.33	1
Beta	HB	0.41	0.07	18
	MS	0.43	0.07	1
Variance	HB	2.33	0.54	11
	MS	2.34	0.53	8
Certainty	HB	−49.58	30.56	14
Equivalent	MS	−78.56	51.76	5

Certainty Equivalent is a risk adjusted measure of portfolio performance: the bigger the better.

6.10 Summary

1. **Hierarchical models vastly extend "standard" statistical models.**

2. **They provide a fuller description of complex, multi–level data.**

3. **The interaction model uses a multivariate regression model to describe the variation in the parameters.**

Chapter 7

HB Regression: Mixture Model

Outline

1. **Objectives**
2. **Distributions**
3. **Model**
4. **Priors**
5. **“Latent” Variables**
6. **Joint Distribution**
7. **Full Conditionals**

7.1 Objective

1. **Distributions:**
   - **Multinomial Distribution**
   - **Dirichlet Distribution**
   - **Dirichlet–Multinomial Distribution**
   - **Ordered Dirichlet Distribution**
2. **Mixture Models**
   - **Unobserved segment membership.**
   - **Subjects within a segment are more homogeneous than subjects in different segments.**

3. "Latent" Variables

- **Introducing "latent" variables in the model can simplify MCMC.**
- **Idea is similar to that used in data imputation and EM.**
  - **Given "missing data," it is simple to generate "parameters."**
  - **Given "parameters" it is simple to generate "missing data."**

- **Bayesian inference treats all unknown quantities as random variables. It does not make a distinction between "missing data" and "parameters."**
- **Concept is not new to us. In linear regression, it is simple to generate β given σ and to generate σ given β.**
- **Now, we introduce "parameters" or "missing data" that are not explicitly part of the model specification.**
- **At an abstract level, these parameters correspond to dummy variables of integration.**

7.2 Distributions

7.2.1 Multinomial Distribution

1. **Define:**

   - K **vector of non–negative integers:**

$$N = (n_1, \ldots, n_K)'$$
$$n = n_1 + \cdots n_K.$$

   - K **vector of probabilities:**

$$\Psi = (\psi_1, \ldots, \psi_K)'$$

where $0 \leq \psi_k$ **and** $\psi_1 + \cdots + \psi_K = 1.$

2. **Example:**

   - n **is the total number of customers.**
   - n_k **is the number of customers in segment** k**.**
   - ψ_k **is the probability of segment** k**.**

3. N given Ψ has a multinomial distribution with pmf:

$$
\begin{aligned}
[N|\Psi] &= MN_K(N|\Psi) \\
&\equiv \begin{pmatrix} n \\ n_1\, n_2\, \cdots\, n_K \end{pmatrix} \prod_{k=1}^{K} \psi_k^{n_k} \\
&= n! \prod_{k=1}^{K} \frac{\psi_k^{n_k}}{n_k!}
\end{aligned}
$$

4. Moments:

$$E(n_k|\Psi) = n\psi_k$$

$$V(n_k|\Psi) = n\psi_k(1-\psi_k)$$

$$\mathbf{Cov}(n_j, n_k|\Psi) = -n\psi_j\psi_k$$

$$\mathbf{Cor}(n_j, n_k|\Psi) = -\left(\frac{\psi_j}{1-\psi_j}\right)^{\frac{1}{2}}\left(\frac{\psi_k}{1-\psi_k}\right)^{\frac{1}{2}}$$

5. We will need a special case where $n = 1$.

I have written a Gauss routine in plbam.src that returns a vector of segment memberships given a matrix of membership probabilities.

$$\mathbf{z = rndzmn(zprob);}$$

zprob is a nsub by K matrix of segment probabilities, and z is a nsub vector whose entries are 1 to K.

7.2.2 Dirichlet Distribution

1. **The Dirichlet distribution is the multivariate extension of the Beta distribution.**

2. **Let $\Psi = (\psi_1, \ldots, \psi_K)'$ be a K vector of probabilities:**

$$0 \leq \psi_k \textbf{ and } \sum_{k=1}^{K} \psi_k = 1.$$

3. **Let $W = (w_1, \ldots, w_K)'$ be a K vector of positive numbers, and define $w = w_1 + \cdots w_K$.**

4. **Ψ has a Dirichlet distribution with pdf:**

$$\begin{aligned} [\Psi|W] &= Dir_K(\Psi|W) \\ &= \frac{\Gamma(w)}{\prod_{k=1}^{K}\Gamma(w_k)} \prod_{k=1}^{K} \psi_k^{w_k-1} \end{aligned}$$

$$\textbf{for } 0 \leq \psi_k \textbf{ and } \sum_{k=1}^{K} \psi_k = 1$$

5. It can be derived from the following.

- **Let** X_k **be** $G(X_k|w_k, \beta)$**.**
- **Assume that** X_1**, . . . ,** X_K **are mutually independent.**
- **Define:**

$$\begin{aligned} \psi_k &= X_k/S \textbf{ for } k = 1, \ldots, K \\ S &= X_1 + \cdots X_K \end{aligned}$$

- Ψ **and** S **are independent.**
- $[S|W] = G(S|w, \beta)$**.**
- $[\Psi|W] = Dir_K(\Psi|W)$**.**

6. Moments

Let v_k be positive numbers, and $v = v_1 + \cdots + v_K$.

$$E\left(\prod_{k=1}^{K} \psi_k^{v_k}\right) = \left[\frac{\Gamma(w)}{\prod_{k=1}^{K} \Gamma(w_k)}\right] \left[\frac{\prod_{k=1}^{K} \Gamma(w_k + v_k)}{\Gamma(w + v)}\right]$$

$$E(\psi_k) = \frac{w_k}{w}$$

$$V(\psi_k) = \frac{1}{w+1} E(\psi_k)[1 - E(\psi_k)]$$

$$\mathbf{Cov}(\psi_j, \psi_k) = -\frac{1}{w+1} E(\psi_j) E(\psi_k)$$

$$\mathbf{Cor}(\psi_j, \psi_k) = -\left[\frac{E(\psi_j)}{1 - E(\psi_j)}\right]^{\frac{1}{2}} \left[\frac{E(\psi_k)}{1 - E(\psi_k)}\right]^{\frac{1}{2}}$$

7.2.3 Dirichlet–Multinomial

$$
\begin{aligned}
[N|W] &= \int_\Psi MN_K(N|\Psi)\, Dir_K(\Psi|W)\, d\Psi \\
&= n!\Gamma(w) \int_\Psi \prod_{k=1}^{K} \frac{\psi_k^{n_k+w_k-1}}{n_k!\Gamma(w_k)}\, d\Psi \\
&= n! \frac{\Gamma(w)}{\Gamma(n+w)} \prod_{k=1}^{K} \frac{\Gamma(n_k+w_k)}{n_k!\Gamma(w_k)}
\end{aligned}
$$

7.2.4 Ordered Dirichlet Distribution

1. **Ψ has the ordered Dirichlet Distribution with pdf:**

$$
\begin{aligned}
[\Psi|W] &= ODir_K(\Psi|W) \\
&\propto Dir_K(\Psi|W)I(0 \leq \psi_1 \leq \psi_2 \leq \cdots \leq \psi_K)
\end{aligned}
$$

2. **I have written a Gauss routine in plbam.src that generates ordered Dirichlet.**

{psi, xgam} = dirord(w, xgam),

where

- **w is the K vector of parameters.**
- **xgam is nsub by K matrix of ordered gamma random deviates. xgam is updated on each call of dirord. It needs to be initialized for the first call.**
- **psi is a nsub by K matrix of ordered Dirichlet probabilities.**

7.3 Model

1. **Within–Subject Model:**

$$Y_i = X_i\beta_i + \epsilon_i$$

where

- **there are** n **subjects and**
- m_i **observations for subject** i**;**
- Y_i **is a** m_i **vector;**
- X_i **is a** $m_i \times p$ **design matrix;**
- β_i **is a** p **vector of individual–level regression coefficients; and**
- ϵ_i **is a** m_i **vector of error terms with pdf**

$$[\epsilon_i|\sigma] = N_{m_i}(\epsilon_i|0, \sigma^2 I_{m_i}).$$

2. Between–Subjects Mixture Model:

$$[\beta_i|\Theta, \Lambda, K] = \sum_{k=1}^{K} \psi_k N_p(\beta_i|\theta_k, \Lambda_k).$$

where

- θ_k **is a** p **vector for** $k = 1, \ldots, K$.
- Λ_k **is a** $p \times p$ **pds covariance matrix for** $k = 1, \ldots, K$.
- $0 \leq \psi_1 < \psi_2 < \cdots < \psi_K$ **and** $\Sigma_{k=1}^{K} \psi_k = 1$.

3. Interpretation of the Mixture Model.

- **Each subject belongs to one of K segments.**
- **The distribution of parameter heterogeneity in segment k is:**

$$[\beta_i|k, \theta_k, \Lambda_k] = N_p(\beta_i|\theta_k, \Lambda_k).$$

- **Segment membership is unknown.**
- **The prior probability of belonging to segment k is ψ_k.**
- **In order the identify the model, the probabilities are ordered: the first segment is the smallest, and the last segment is the largest.**

4. Mixture Model for Y_i

- **Mixture model for β_i induces a mixture model for the marginal distribution of Y_i.**
- **Integrate β_i out of the model.**
- **Obtain:**

$$[Y_i|K,\Theta,\Lambda] = \sum_{k=1}^{K} \psi_k N_{m_i}(Y_i|X_i\theta_k, \sigma^2 I_{m_i} + X_i\Lambda_k X_i').$$

- **If subject i belongs to segment k, then**

$$\begin{aligned} Y_i &= X_i\theta_k + \epsilon_i(k) \\ V(\epsilon_i(k)) &= \sigma^2 I_{m_i} + X_i\Lambda_k X_i' \end{aligned}$$

- **Probability of belonging to segment k is ψ_k.**

7.4 Priors

$$
\begin{aligned}
[\sigma^2|r_0, s_0] &= IG\left(\sigma^2|\frac{r_0}{2}, \frac{s_0}{2}\right) \\
[\theta_k|u_0, V_0] &= N_p(\theta_k|u_0, V_0) \\
[\Lambda_k|f_0, G_0] &= IW_p(\Lambda_k|f_0, G_0^{-1}) \\
[\Psi|W_0] &= ODir_K(\Psi|W_0)
\end{aligned}
$$

7.5 “Latent” Variables

1. **In the MCMC we will introduce a segment membership variable for each subject.**

2. **If subject i belongs to segment k, define**

$$Z_i = k$$

$$[Z_i = k] = \psi_k.$$

3. **Given $Z_i = k$, the distribution of β_i is:**

$$[\beta_i | Z_i = k] = N_p(\beta_i | \theta_k, \Lambda_k).$$

7.6 Joint Distribution

1. Define the number of subjects in segment k:

$$n_k = \sum_{i=1}^{n} I(Z_i = k)$$

$$N = (n_1, \ldots, n_K)'$$

$$[N|\Psi] = MN_K(N|\Psi)$$

2. Joint distribution for the HB Mixture Model:

$$\prod_{i=1}^{n} \{[Y_i|\beta_i, \sigma][\beta_i|Z_i = k][Z_i = k]\} \prod_{k=1}^{K} \{[\theta_k][\Lambda_k]\} [\sigma][\Psi]$$

$$\begin{aligned}
= & \prod_{i=1}^{n} N_{m_i}(Y_i|X_i\beta_i, \sigma^2) \\
\times & \prod_{i=1}^{n} N_p(\beta_i|\theta_k, \Lambda_k) \\
\times & \prod_{k=1}^{K} N_p(\theta_k|u_0, V_0) IW_p(\Lambda_k|f_0, G_0^{-1}) \\
\times & MN_K(N|\Psi) ODir_K(\Psi|W_0) \\
\times & IG\left(\sigma^2|\frac{r_0}{2}, \frac{s_0}{2}\right)
\end{aligned}$$

7.7 Full Conditionals

1. **Given segment membership Z and Ψ, how do you generate β_i, θ_k, Λ_k, and σ^2?**

2. Given segment membership and the rest, generate Ψ:

$$
\begin{aligned}
[\Psi|\mathbf{Rest}] &= ODir_K(\Psi|W_n) \\
W_n &= W_0 + N
\end{aligned}
$$

3. Full conditional of Z_i:

$$
[Z_i = k|\mathbf{Rest}] = \frac{[\beta_i|Z_i = k]\psi_k}{\Sigma_{j=1}^{K}[\beta_i|Z_i = j]\psi_j}
$$

$$
= \frac{|\Lambda_k|^{-\frac{1}{2}} \exp\left\{-\frac{1}{2}(\beta_i - \theta_k)'\Lambda_k^{-1}(\beta_i - \theta_k)\right\}\psi_k}{\Sigma_{j=1}^{K}|\Lambda_j|^{-\frac{1}{2}} \exp\left\{-\frac{1}{2}(\beta_i - \theta_j)'\Lambda_j^{-1}(\beta_i - \theta_j)\right\}\psi_j}
$$

This corresponds to the posterior probability of subject i belonging to segment k given the parameters.

7.8 Simulated Data

- **200 subjects**
- **5 observations per subject**
- **1 predictor X variable**
- **Error STD $\sigma = 5$**
- **3 component model**
- **True means for components:**

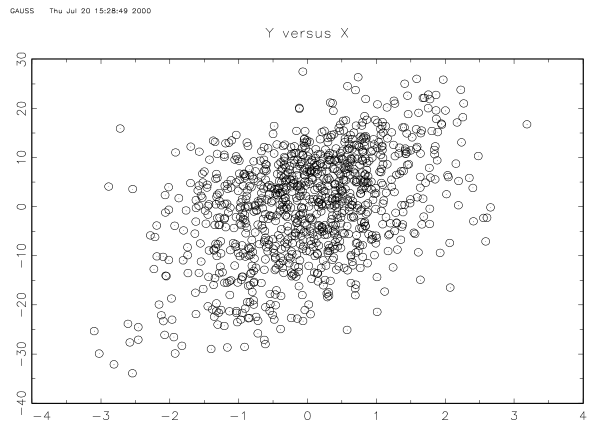
$$\theta_1 = \begin{bmatrix} 0 \\ 0 \end{bmatrix}; \quad \theta_2 = \begin{bmatrix} -10 \\ 7 \end{bmatrix}; \ \textbf{and } \theta_3 = \begin{bmatrix} 7 \\ 5 \end{bmatrix},$$

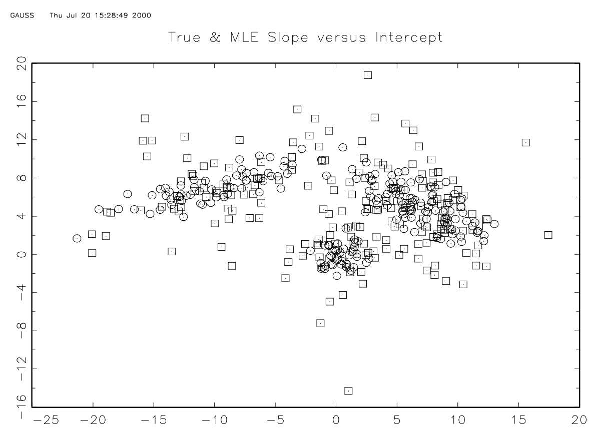
- **True variance matrices for components:**

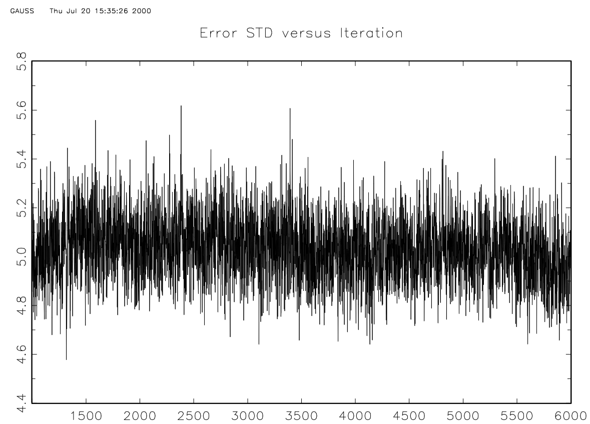
$$\Lambda_1 = \begin{bmatrix} 1 & 0 \\ 0 & 1 \end{bmatrix}; \quad \Lambda_2 = \begin{bmatrix} 25 & 9 \\ 9 & 4 \end{bmatrix}; \ \textbf{and } \Lambda_3 = \begin{bmatrix} 9 & -5 \\ -5 & 5 \end{bmatrix},$$

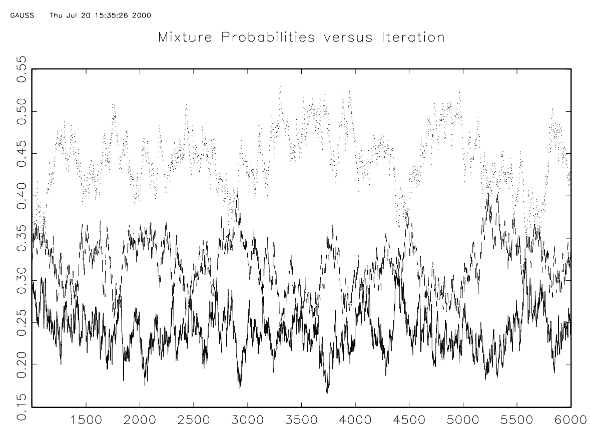
- **Mixture proportions:**

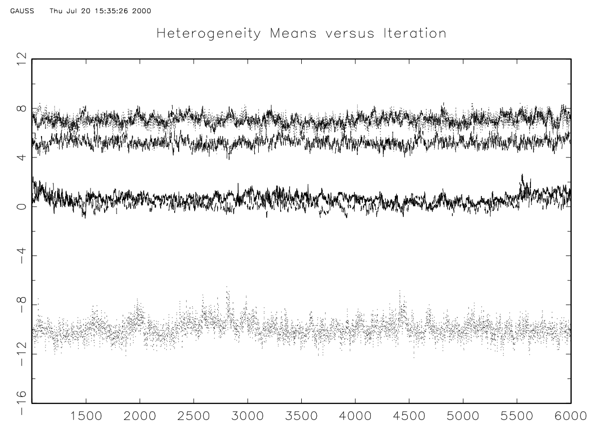
$$\psi_1 = 0.2, \quad \psi_2 = 0.3, \ \textbf{and } \psi_3 = 0.5$$

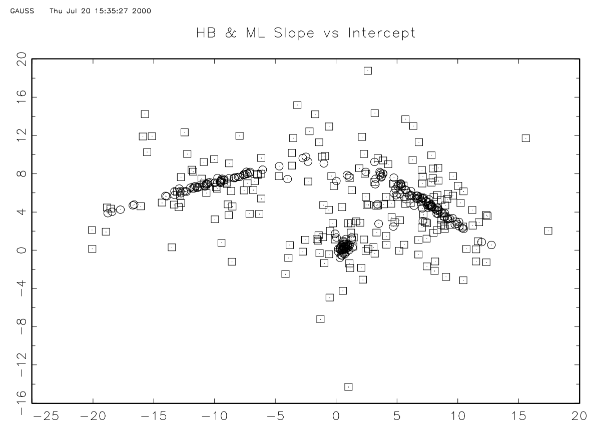

GAUSS Thu Jul 20 15:28:49 2000
Y versus X


GAUSS Thu Jul 20 15:28:49 2000
True & MLE Slope versus Intercept


GAUSS Thu Jul 20 15:35:26 2000
Error STD versus Iteration
4.4 4.6 4.8 5.0 5.2 5.4 5.6 5.8
1500 2000 2500 3000 3500 4000 4500 5000 5500 6000


GAUSS Thu Jul 20 15:35:26 2000
Mixture Probabilities versus Iteration
0.15 0.20 0.25 0.30 0.35 0.40 0.45 0.50 0.55
1500 2000 2500 3000 3500 4000 4500 5000 5500 6000


GAUSS Thu Jul 20 15:35:26 2000
Heterogeneity Means versus Iteration
12
8
4
0
−4
−8
−12
−16
1500 2000 2500 3000 3500 4000 4500 5000 5500 6000


GAUSS Thu Jul 20 15:35:27 2000
HB & ML Slope vs Intercept
20
16
12
8
4
0
−4
−8
−12
−16
−25 −20 −15 −10 −5 0 5 10 15 20

```
MCMC Analysis

Total number of MCMC iterations                = 6000.00000
Number of iterations used in the analysis      = 5000.00000
Number in transition period                    = 1000.00000
Number of iterations between saved iterations  =    0.00000

Number of subjects                   =  200.00000
Mean # of observations per subject   =    5.00000
STD  # of observations per subject   =    0.00000
MIN  # of observations per subject   =    5.00000
MAX  # of observations per subject   =    5.00000
Total number of observations         = 1000.00000
Number of independent variables X    =    1.00000  (excluding intercept)

Dependent variable is         Y

Independent variables in first level equation:
Y_i = X_i*beta_i + epsilon_i
Variable        Mean       STD       MIN       MAX
Constant     1.00000   0.00000   1.00000   1.00000
X 1          0.00654   0.99302  -3.09851   3.18555

--------------------------------------------------------
```

```
Statistics of Fit Measures for each Subject
Average Predictive Correlation (Muptiple R) =    0.58701
STD of Predictive Correlations              =    0.39932
Average R-Squared                           =    0.50324
STD of R-Squared                            =    0.32678
Average Error Standard Deviation            =    4.37049
STD of Error Standard Deviation             =    1.42166

---------------------------------------------------------
Estimation of the error STD sigma
True Sigma     =    5.00000
MLE            =    3.81725
Posterior Mean =    5.03351
Posterior STD  =    0.13035

---------------------------------------------------------
Comparison of True Beta to Individual Level Estimates
Variable is   Constant
Correlation between true and HB  =    0.96306
RMSE between true and HB         =    2.13864

Correlation between true and MLE =    0.95487
RMSE between true and MLE        =    2.43523

Variable is        X 1
Correlation between true and HB  =    0.72655
RMSE between true and HB         =    2.24838

Correlation between true and MLE =    0.65192
RMSE between true and MLE        =    3.41463
----------------------------------------------------------
```

Estimated Group Probabilities psi

	Group 1	Group 2	Group 3
True	0.20000	0.30000	0.50000
	Group 1	Group 2	Group 3
HB Mean	0.23917	0.31850	0.44233
HB STD	0.02444	0.03008	0.03519

Classification Rates:

True versus Maximum HB Posterior Probability

HB Group	True 1	True 2	True 3	Total
Group 1	38	3	7	48
Group 2	2	58	1	61
Group 3	3	1	87	91
Total	43	62	95	200

```
HB Estimates of Theta
True Theta
Variable     Group 1   Group 2   Group 3
Constant     0.00000 -10.00000   7.00000
X 1          0.00000   7.00000   5.00000

Posterior Mean of Theta
Variable     Group 1   Group 2   Group 3
Constant     0.65579  -9.96220   7.03063
X 1          0.28745   7.08078   5.23424

Posterior STD of Theta
Variable     Group 1   Group 2   Group 3
Constant     0.39694   0.73518   0.43332
X 1          0.48984   0.41366   0.38365

-------------------------------------------------------
```

```
HB Estimate of Lambda
True Lambda for group    1.00000
Variable    Constant       X 1
Constant     1.00000   0.00000
X 1          0.00000   1.00000

Posterior Mean of Lambda for group    1.00000
Variable    Constant       X 1
Constant     0.64158   0.18413
X 1          0.18413   1.19855

Posterior STD of Lambda for group    1.00000
Variable    Constant       X 1
Constant     0.62639   0.70570
X 1          0.70570   1.66026

-----------------------------------------------------
```

```
True Lambda for group    2.00000
Variable    Constant       X 1
Constant    25.00000   9.00000
X 1          9.00000   4.00000

Posterior Mean of Lambda for group    2.00000
Variable    Constant       X 1
Constant    17.01619   5.78113
X 1          5.78113   2.49008

Posterior STD of Lambda for group    2.00000
Variable    Constant       X 1
Constant     5.65565   2.08201
X 1          2.08201   1.24675

-------------------------------------------------------
```

```
True Lambda for group    3.00000
Variable    Constant       X 1
Constant     9.00000  -5.00000
X 1         -5.00000   5.00000

Posterior Mean of Lambda for group    3.00000
Variable    Constant       X 1
Constant     6.82231  -5.35847
X 1         -5.35847   5.01464

Posterior STD of Lambda for group    3.00000
Variable    Constant       X 1
Constant     2.44737   1.57009
X 1          1.57009   1.70303
```

7.9 Model Selection

- **Vary the number of components.**
- **Compute the posterior probability of the model.**
- **See page (37) for the decision theoretic basis for model selection. The different models correspond to different ω_i.**
- **Lenk and DeSarbo (2000) *Psychometrika* use the method of Gelfand and Dey (1994) *JRSSb* to select the model.**
- **For the model with K components, indicate all of the parameters by Ω_K.**

- **The marginal density of the data given K components is:**

$$
\begin{aligned}
f_K(Y) &= \int_{\Omega_K} f_K(Y|\Omega_K) p_K(\Omega_K) d\Omega_K \\
&= \left\{ E\left[\frac{g_K(\Omega_K)}{f_K(Y|\Omega_K) p_K(\Omega_K)} \right] \right\}^{-1}.
\end{aligned}
$$

  - f_K **is the density of the data given the parameters for model K.**
  - p_K **is the prior density of the parameters.**
  - g_K **is an arbitrary density on the support of Ω_K.**
  - **The expectation is with respect to the posterior distribution of Ω_K.**

- **The MCMC approximation is**

$$\tilde{f}_K(Y) = \left[\frac{1}{U-B}\sum_{u=B+1}^{U}\frac{g_K\left(\Omega_K^{(u)}\right)}{f_K\left(Y|\Omega_K^{(u)}\right)p_K\left(\Omega_K^{(u)}\right)}\right]^{-1}.$$

  - $\Omega_K^{(u)}$ **is the value of** Ω_K **on the iteration** u **of the Markov chain.**
  - **The last** $U-B$ **iterations of** U **iterations are used.**
  - **If** g_K **is the posterior density of** Ω_K**, then the approximation is exact.**
  - **One choice of** g_K **is multivariate normal for suitably transformed parameters. Estimate the mean and covariance matrix from the MCMC random deviates.**

7.10 Metric Conjoint Study

See page (229) for a description of the MBA computer survey.

Posterior probabilities of the number of components and predictive performance.

	Finite Mixture, Random Effects				Individual Level MLE	Latent Class
Number of Components	Four	Three	Two	One	—	Four
Probability[1]	0.083	0.382	0.426	0.109	—	—
Correlation[2]	0.783	0.782	0.782	0.778	0.732	0.683
RMSE[3]	1.724	1.726	1.728	1.742	1.948	4.048
Hit Rate[4]	0.700	0.705	0.711	0.689	0.626	0.380

[1] Posterior probability of the model.

[2] Correlation between observed and predicted responses for the validation data.

[3] Root mean squared error between observed and predicted responses for the validation data.

[4] Proportion of times correctly predicted the maximum in validation sample.

Means and variances of the regression coefficients within each class for the two component solution for the computer survey. Posterior standard errors are in parentheses.

	Means		Variances	
Mixing Probability	0.061	0.939	0.061	0.939
Intercept	4.931	4.662	0.437	2.252
	(0.343)	(0.117)	(0.404)	(0.250)
Hot Line	0.030	0.088	0.307	0.110
	(0.228)	(0.034)	(0.319)	(0.022)
RAM	0.389	0.309	0.363	0.124
	(0.214)	(0.036)	(0.258)	(0.023)
Screen	0.015	0.200	0.226	0.074
	(0.173)	(0.031)	(0.157)	(0.014)
CPU	1.143	0.337	1.869	0.222
	(0.417)	(0.046)	(1.250)	(0.047)
Hard Disk	0.674	0.125	1.843	0.070
	(0.352)	(0.031)	(1.314)	(0.015)
CD ROM	0.442	0.492	0.803	0.209
	(0.319)	(0.043)	(0.604)	(0.042)
Cache	0.162	0.044	0.259	0.117
	(0.188)	(0.035)	(0.187)	(0.023)
Store	0.163	0.082	0.204	0.052
	(0.162)	(0.028)	(0.140)	(0.011)
Warranty	0.128	0.097	0.282	0.063
	(0.185)	(0.030)	(0.192)	(0.013)
Software	0.131	0.195	0.197	0.080
	(0.160)	(0.031)	(0.131)	(0.016)
Guarantee	0.133	0.102	0.230	0.084
	(0.177)	(0.032)	(0.152)	(0.018)
Price	-0.459	-1.168	0.264	0.764
	(0.211)	(0.070)	(0.195)	(0.094)

7.11 Summary

1. **The mixture model describes the variation in the parameters with a mixture of multivariate normal distributions.**

2. **The interaction model describes this variation with a multivariate regression model.**

3. **Which one is better is an empirical issue.**

4. **A model that generalizes both is a mixture of multivariate regression models.**

References

- Diebolt, J., and Robert, C. P. (1994). Estimation of finite mixture distributions through Bayesian sampling. *Journal of the Royal Statistical Society, Series B*, 56, 362–375.
- Gelfand, A. E., and Dey, D. K. (1994). Bayesian model choice: asymptotics and exact calculations. *Journal of the Royal Statistical Society, Series B*, 56, 501–514.

Chapter 8

Revealed Preference Models

Outline

1. **Objectives**
2. **Random Utility Model**
3. **Probit Model**
4. **Logit Model**
5. **Hastings–Metropolis**
6. **Data Structures**
7. **Multivariate Probit Model**

8.1 Objectives

1. **Revealed preference models use random utilities.**

2. **Probit models assume that utilities are multivariate normal.**

3. **Probit MCMC generates latent, random utilities.**

4. **Logit models assume that the random utilities have extreme value distributions.**

5. **Logit MCMC uses the Hastings–Metropolis algorithm.**

6. **Hastings–Metropolis algorithm is a general purpose, flexible algorithm for generating random variables.**

8.2 Random Utility Model

1. **Utility for alternative m is:**

$$Y_{i,j,m} = x'_{i,j,m}\beta_i + \epsilon_{i,j,m}$$

$$i = 1, \ldots, n$$

$$j = 1, \ldots, J_i$$

$$m = 1, \ldots, M + 1$$

where

- **there are n subjects or customers,**
- **$M + 1$ alternatives in the choice set, and**
- **J_i choice occasions for subject i.**

2. Subject picks alternative k if

$$Y_{i,j,k} \geq Y_{i,j,m} \textbf{ for all } m.$$

3. The probability of selecting k is

$$P(Y_{i,j,k} \geq Y_{i,j,m} \textbf{ for all } m).$$

4. Statistical Models:

- $\{\epsilon_{i,j,m}\}$ **are Normal $\Rightarrow$ Probit Model.**
- $\{\epsilon_{i,j,m}\}$ **are Extreme Value $\Rightarrow$ Logit Model.**

$$[\epsilon] \;=\; \exp\{-\epsilon - e^{-\epsilon}\} \textbf{ for } -\infty < \epsilon < \infty$$

$$P(\epsilon \leq x) \;=\; \int_{-\infty}^{x} [\epsilon]\, d\epsilon = \exp\{-\exp(-x)\}$$

5. Revealed preference data:

$$C_{i,j} = k$$

if alternative k was selected by subject i on choice occasion j.

6. Model Identification

- **Alternative $M+1$ is the base alternative.**
- **Assume $Y_{i,j,M+1} = 0$.**
- **Measure independent variables relative to base alternative.**
- **Define:**

$$Y_{i,j} = \begin{bmatrix} Y_{i,j,1} \\ \vdots \\ Y_{i,j,M} \end{bmatrix}; \quad X_{i,j} = \begin{bmatrix} x'_{i,j,1} \\ \vdots \\ x'_{i,j,M} \end{bmatrix}, \textbf{ and } \epsilon_{i,j} = \begin{bmatrix} \epsilon_{i,j,1} \\ \vdots \\ \epsilon_{i,j,M} \end{bmatrix}.$$

- **Example**
  - **Four brands.**
  - **Independent variables are Price and Advertising.**

$$X_{i,j} = \begin{bmatrix} 1 & 0 & 0 & p_1 - p_4 & a_1 - a_4 \\ 0 & 1 & 0 & p_2 - p_4 & a_2 - a_4 \\ 0 & 0 & 1 & p_3 - p_4 & a_3 - a_4 \end{bmatrix}$$

and

$$\beta_i = \begin{bmatrix} \beta_{i,1} \\ \beta_{i,2} \\ \beta_{i,3} \\ \beta_{i,P} \\ \beta_{i,A} \end{bmatrix} \quad \begin{array}{l} \textbf{Brand Preference 1} \\ \textbf{Brand Preference 2} \\ \textbf{Brand Preference 3} \\ \textbf{Price Effect} \\ \textbf{Advertising Effect} \end{array}$$

8.3 Probit Model

1. Within–Subject Latent Utility Model:

$$Y_{i,j} = X_{i,j}\beta_i + \epsilon_{i,j}$$

where

$$\begin{aligned} [\epsilon_{i,j}] &= N_M(\epsilon_{i,j}|0, \Sigma) \\ \sigma_{M,M} &= 1. \end{aligned}$$

2. Between–Subjects Model:

$$B = Z\Theta + \Delta$$

where

$$[\Delta] = N_{n\times p}(\Delta|0, I_n, \Lambda).$$

and

$$B = \begin{bmatrix} \beta_1' \\ \vdots \\ \beta_n' \end{bmatrix} \textbf{ and } Z = \begin{bmatrix} z_1' \\ \vdots \\ z_n' \end{bmatrix}.$$

3. **Identification Trick. The model specifics that $\sigma_{M,M}$ is one. In this case, the inverted Wishart distribution is not appropriate for Σ. McCulloch and Rossi (1994) had a brilliant insight:**

    - **Ignore the constraint on $\sigma_{M,M}$.**
    - **Use the inverted Wishart prior for Σ. This model is not identified.**
    - **After generating random iterates in MCMC, divide Y, β, and Θ by $\sqrt{\sigma_{M,M}}$, and divide Σ and Λ by $\sigma_{M,M}$.**
    - **Next, compute posterior means, STDs, etc.**

4. Priors for Unidentified Model

(No constraint on $\sigma_{M,M}$):

$$[\Sigma] = IW_M(\Sigma|s_0, R_0^{-1})$$

$$[\mathbf{vec}(\Theta')] = N_{pq}(\mathbf{vec}(\Theta')|u_0, V_0)$$

$$[\Lambda] = IW_p(\Lambda|f_0, G_0^{-1})$$

5. Probit MCMC

(a) Joint pdf:

$$\prod_{i=1}^{n} \prod_{j=1}^{J_i} [C_{i,j}|\beta_i, \Sigma] \prod_{i=1}^{n} [\beta_i|\Theta, \Lambda][\Sigma][\Theta][\Lambda].$$

Introduce latent variables $Y_{i,j}$:

$$\prod_{i=1}^{n} \prod_{j=1}^{J_i} [C_{i,j}|Y_{i,j}][Y_{i,j}|\beta_i, \Sigma] \prod_{i=1}^{n} [\beta_i|\Theta, \Lambda][\Sigma][\Theta][\Lambda].$$

What is $[C_{i,j}|Y_{i,j}]$

- **when $C_{i,j} = k$ for $k \leq M$?**

- **when $C_{i,j} = M + 1$?**

(b) **Full conditional of $Y_{i,j}$:**

$$[Y_{i,j}|C_{i,j} = k, \mathbf{Rest}]$$
$$\propto \; N_M(Y_{i,j}|X_{i,j}\beta_i, \Sigma) I(Y_{i,j,k} \geq Y_{i,j,m} \text{ for all } m)$$

which is a truncated normal density.

- **Sequentially generate components of $Y_{i,j}$.**
- **Define**

$$Y_{i,j,-m} = (Y_{i,j,1}, \ldots, Y_{i,j,m-1}, Y_{i,j,m+1}, \ldots Y_{i,j,M})'.$$

- **Generate $Y_{i,j,m}$ given $Y_{i,j,-m}$.**
  - **See page (69) for the conditional normal distribution.**
  - **See page (143) for generating from truncated, univariate distributions.**

(c) Full conditional of β_i

$$[\beta_i|\mathbf{Rest}] = N_p(\beta_i|u_i, V_i)$$

$$V_i = \left(\sum_{j=1}^{J_i} X'_{i,j}\Sigma^{-1}X_{i,j} + \Lambda^{-1}\right)^{-1}$$

$$u_i = V_i\left(\sum_{j=1}^{J_i} X'_{i,j}\Sigma^{-1}Y_{i,j} + \Lambda^{-1}\Theta' z_i\right)$$

(d) Full conditional of Σ:

$$[\Sigma|\mathbf{Rest}] = IW_M(\Sigma|s_n, R_n^{-1})$$

$$s_n = s_0 + \sum_{i=1}^{n} J_i$$

$$R_n^{-1} = R_0^{-1} + \sum_{i=1}^{n}\sum_{j=1}^{J_i}(Y_{i,j} - X_{i,j}\beta_i)(Y_{i,j}^* - X_{i,j}^*\beta_i).'$$

6. **Post–MCMC Identification. On the iterations that you save for analysis, perform the following standardizations:**

   - $\Sigma \leftarrow \Sigma/\sigma_{M,M}$
   - $\Lambda \leftarrow \Lambda/\sigma_{M,M}$
   - $Y_{i,j} \leftarrow Y_{i,j}/\sqrt{\sigma_{M,M}}$
   - $\beta_i \leftarrow \beta_i/\sqrt{\sigma_{M,M}}$
   - $\Theta \leftarrow \Theta/\sqrt{\sigma_{M,M}}$

   The left arrows mean replace the left–hand–side with the right–hand–side.

8.3.1 Example

- **Study by McKinsey and IntelliQuest**
- **Conjoint Survey of Company Purchasers**
- **Profiles: Personal Computers**
- **316 Subjects**
- **3 Profiles per Choice Task + "None"**
- **8 Choice Sets per Person**
- **Different Design Matrices**

Attributes

1. **5 Brands**
2. **Performance or Speed:**

   Low, Average, High
3. **Channel:**

   Telephone, Retail Store, Onsite Sales Rep
4. **Warranty:**

   90 Day, 1 Year, 5 Year
5. **Service:**

   Ship back, Retail Store, Onsite
6. **Price:**

   Low, Med-Low, Med-High, High

Subject Level Covariates Z

1. **Expect to Pay:**

   Low, Average, High

2. **Buying Expertise:**

   Average, High

3. **Education:**

   HS, College Graduate, Advanced Graduate

4. **Gender**

5. **Company Size:**

   Small, Medium, Large

MCMC

1. **6000 total iterations**
2. **5000 initial iterations**
3. **1000 iterations used in the analysis**
4. **13 hours on a 430 MHz Pentium**

Posterior Means of β_i

	Mean	STD
Brand A	-3.50	3.51
Brand B	-1.23	4.43
Brand C	0.96	4.93
Brand D	-1.62	3.34
Brand E	-1.13	4.13
Slow	-1.83	2.53
Fast	24.82	6.69
Buy over Telephone	-1.48	1.77
Buy Onsite	-1.49	2.46
90 Day Warranty	-0.43	3.69
5 Year Warranty	-1.78	2.64
Ship back for Service	2.15	2.18
Onsite Service	2.22	2.77
Med-Low Price	-2.09	3.42
Med-High Price	-2.89	3.68
High Price	-3.00	5.34

Error Variance Σ

Posterior Mean

	Profile 1	Profile 2	Profile 3
Profile 1	1.24	0.68	0.32
Profile 2	0.68	2.17	0.43
Profile 3	0.33	0.43	1.00

Posterior STD

	Profile 1	Profile 2	Profile 3
Profile 1	0.51	0.44	0.20
Profile 2	0.45	1.19	0.30
Profile 3	0.20	0.30	0.00

Posterior Mean of Θ'
(Blank if |mean|/std < 2)

		Price		Expert	Education			Company	
	Constant	Low	High	Buyer	HS	Grad	Female	Small	Large
Brand									
A	-3.34	4.21			-2.84	-3.27	-4.05	3.44	
B			-3.32		-4.17	-8.84		2.01	
C	-2.49	-3.04	1.85	5.73	-4.13	-3.32	5.46	6.38	
D	-5.65		4.50					2.27	4.01
E		5.22				-6.45			
Speed									
Slow	-3.22			2.04					
Fast	19.43		4.16		5.69	2.31		3.76	
Channel									
Telephone		-2.62							
Onsite			-2.90	-2.00		2.39			
Warranty									
90 Day			-6.56		3.83				
5 Year			-3.50						
Service									
Ship	1.85					-2.03	-2.20		2.38
Onsite	3.20	-2.89		-2.21	1.76				
Price									
Med-Low		-1.96	-2.17			4.95		-4.64	-3.08
Med-High		-3.28		-1.66				-3.84	3.56
High	4.72	-5.71		-4.41	-4.77	2.06		-7.91	-1.79

Error Variance Λ for Second Equation

- **Posterior mean of the STDs were close to one.**
- **Correlation were small and not very informative**

8.4 Logit Model

1. Within–Subject Model:

$$P(C_{i,j} = k|\beta_i) = P(Y_{i,j,k} \geq Y_{i,j,m} \textbf{ for all } m)$$

$$= \frac{\exp(x'_{i,j,k}\beta_i)}{1 + \Sigma_{m=1}^{M} \exp(x'_{i,j,m}\beta_i)} \textbf{ if } k \leq M$$

$$= \frac{1}{1 + \Sigma_{m=1}^{M} \exp(x'_{i,j,m}\beta_i)} \textbf{ if } k = M + 1.$$

2. Between–Subjects Model:

$$B = Z\Theta + \Delta$$

$$[\Delta] = N_{n\times p}(\Delta|0, I_n, \Lambda)$$

3. Priors:

$$[\mathbf{vec}(\Theta')] = N_{pq}(\mathbf{vec}(\Theta')|u_0, V_0)$$

$$[\Lambda] = IW_p(\Lambda|f_0, G_0^{-1})$$

8.5 Hastings–Metropolis

Generate random variables X from a density that is proportional to f:

$$[x] \propto f(x).$$

These random deviates are X_1, X_2, ...

1. **Initialize X_1.**
2. **At iteration $i+1$, generate a candidate Y from a *jump* distribution:** $g_i(y|x_i)$
   - **Independence:**

     $$g(y|x) = g(y).$$

   - **Symmetric:**

     $$g(y|x) = g(x|y).$$

   - **Conditional normal:**

     $$g(y|x) = N(y|x, \Upsilon).$$

3. Set $X_{i+1} = Y$ on iteration $i+1$ with probability:

$$p(x_i, y) = \min\left\{\frac{f(y)g_i(x_i|y)}{f(x_i)g_i(y|x_i)}, 1\right\}.$$

4. Set $X_{i+1} = X_i$ on iteration i with probability:

$1 - p(x_i, y)$**.**

The resulting sequence is a Markov chain such that its stationary distribution is proportional to f.

Logit MCMC,

$$f(\beta_i) \;=\; L(\beta_i)N_p(\beta_i|\Theta' z_i, \Lambda)$$

$$g(y|x) \;=\; N_p(y|x, c^2 I_p)$$

1. **c^2 is selected by the user.**

2. Advice about c:

- **If c is too large, then Y tends to be far from X, and there will be too many rejects. That is, X is retained too often.**
- **If c is too small, then Y is too close to X. It will be accepted frequently, but the chain will move slowly through its sample space. That is, the auto correlation of the chain will be big.**
- **Some authors recommend c so that the proportion of acceptances is in the 30% to 40% range.**
- **c is similar to step size in some optimization routines.**

8.6 Data Structures

1. **$M + 1$ alternatives, which do not change.**

2. **J choice occasions per subject.**

3. **Each subject receives the same design matrix on choice j.**

4. **cdata contains the subjects' selections:**

$$\mathbf{cdata} = (C_{i,j}).$$

cdata is an $n \times J$ matrix.

5. **xdata stacks the design matrices:**

$$\textbf{xdata} = \begin{bmatrix} X_1 \\ \vdots \\ X_J \end{bmatrix}.$$

xdata is a $JM \times p$ **matrix.**

6. **iptx is a pointer into xdata that gives the design matrix for the choice sets. iptx is a** $J \times 2$ **matrix.**

7. **The design matrix for choice set j is:**

$$\textbf{xj = xdata[iptx[j,1]:iptx[j,2],.];}$$

8. **beta is a** $n \times p$ **matrix.**

9. **theta is a** $q \times p$ **matrix.**

10. **zdata is a** $n \times q$ **matrix.**

8.7 Scanner Panel Data

Allenby and Lenk (1994) *JASA*

Data

1. **735 households in Springfield, Missouri from 1986 to 1987.**
2. **Household Demographics:**
   - **Mean income = $30,800 and STD = $19,300.**
   - **Mean family size = 3 and STD = 1.25**

3. **Four brands of ketchup: Heinz, Hunt's, Del Monte, and House Brand.**

4. **Market Shares:**

   - **Heinz = 43.1%**
   - **Hunt's = 23.9%**
   - **House Brand = 22.4%**
   - **Del Monte = 10.6%**

5. **Marketing Mix**

Brand	% Time Display	% Time Feature	Mean Price	STD Price
Heinz	8.3	14.3	1.23	0.29
Huntz	10.3	5.7	1.27	0.26
Del Monte	5.4	1.2	1.28	0.24
House	8.5	5.2	0.78	0.11

Logistic Regression Model

1. **Choice probabilities:**

   Household i, purchase occasion t, brand j:

$$p_{i,t}(j) \;=\; \exp[y_{i,t}(j)] / \left\{ \sum_{k=1}^{m} \exp[y_{i,t}(k)] \right\}$$

2. Utilities:

$$y_{i,t}(j) = [\alpha_0(j) + \beta_{i,0}(j)] + x_{i,t}(j)'[\alpha_1 + \beta_{i,1}] + d'_{i,t}\alpha_2(j) + \epsilon_{i,t}(j)$$

- $x_{i,t}(j)$ **are the marketing variables.**
- $d_{i,t}$ **are the demographic variables.**
- α**'s are fixed effects.**
- β_i**'s are random effects** $N(0, \Lambda)$**.**
- $\epsilon_{i,t}$**'s are error terms.**
- **After adjusting for the marketing activity and the household's demographics, household's** i **preference for brand** j **on purchase occasion** t **is**

$$\alpha_0(j) + \beta_{i,0}(j) + \epsilon_{i,t}(j).$$

3. Autocorrelated Error Structure:

$$\epsilon_{i,t} = \Phi\epsilon_{i,t-1} + \zeta_{i,t}$$

- Φ **is a diagonal matrix with** $\phi(j)$ **on the diagonal.**
- **Each** $\phi(j)$ **is between –1 and 1.**
- $\{\zeta_{i,t}\}$ **are mutually independent and identically distributed from** $N_m(0, \Sigma)$**.**
- **The error terms that proceed the observation period,** $\{\epsilon_{i,0}\}$**, are mutually independent and identically distributed from** $N_m(0, C)$**.**

4. Identification:

$$\begin{aligned} \alpha_0(4) &= 0 \\ \beta_{i,0}(4) &= 0 \\ \alpha_2(4) &= 0 \end{aligned}$$

Need more than two choices.

Estimated Fixed Effects

Parameters		Chain 1	Chain 2	Chain 3
Fixed Effects	Heinz	1.988	1.956	1.903
Intercepts	Hunt's	1.675	1.672	1.683
	Del Monte	0.496	0.487	0.492
Fixed Effects	Heinz	1.394	1.404	1.411
Income	Hunt's	0.987	0.965	0.984
	Del Monte	0.987	0.984	0.970
Fixed Effects	Heinz	–1.227	–1.234	–1.219
Family Size	Hunt's	–0.537	–0.560	–0.542
	Del Monte	–0.621	–0.614	–0.618
Fixed Effects	Price	–6.637	–6.682	–6.571
Marketing	Display	2.235	2.289	2.301
Variables	Feature	2.087	2.176	2.063
Random Effects	Heinz	1.735	1.812	1.805
Intercepts	Hunt's	0.670	0.664	0.671
	Del Monte	1.201	1.159	1.250
Random Effects	Price	2.331	2.415	2.301
Marketing	Display	2.175	2.174	2.210
Variables	Feature	1.671	1.701	1.615
Error	Heinz	0.482	0.463	0.478
Variances	Hunt's	0.281	0.274	0.269
	Del Monte	0.219	0.251	0.235
	House Brand	1.017	1.093	1.040
Autocorrelation	Heinz	0.469	0.473	0.480
Coefficients	Hunt's	0.563	0.572	0.575
	Del Monte	0.430	0.418	0.452
	House Brand	0.969	0.972	0.971

Random Effects Covariance Matrix

Upper triangular matrices are correlations.
Standard deviations are in parentheses.

	Intercepts			Slopes		
	Heinz	Hunt's	Del Monte	Price	Display	Feature
Heinz	1.753 (0.532)	0.097	–0.388	0.413	–0.237	0.039
Hunt's	0.105 (0.237)	0.670 (0.206)	0.057	–0.223	–0.034	–0.132
Del Monte	–0.563 (0.248)	0.051 (0.283)	1.201 (0.505)	–0.497	0.230	–0.102
Price	0.834 (0.437)	–0.279 (0.534)	–0.831 (0.689)	2.331 (1.331)	–0.194	0.104
Display	–0.462 (0.382)	–0.041 (0.263)	0.371 (0.314)	–0.436 (0.599)	2.175 (0.573)	0.558
Feature	0.067 (0.378)	–0.140 (0.298)	–0.144 (0.330)	0.206 (0.792)	1.063 (0.403)	1.671 (0.453)

Error Covariance and Autocorrelation

	Error Covariances			
	Heinz	Hunt's	Del Monte	House Brand
Heinz	0.482 (0.196)	0.347	0.182	–0.571
Hunt's	0.128 (0.132)	0.281 (0.088)	0.118	–0.386
Del Monte	0.059 (0.099)	0.029 (0.071)	0.219 (0.075)	–0.154
House Brand	–0.400 (0.182)	–0.206 (0.135)	–0.073 (0.150)	1.017 (0.369)
	Autocorrelation Coefficients			
	0.469 (0.144)	0.563 (0.153)	0.430 (0.337)	0.969 (0.012)

Aggregate Market Response

$$\begin{aligned}\psi_{i,t}(j,k) &\equiv E\left[\frac{\partial p_{i,t}(j)}{\partial \log(\textbf{Price of Brand } k)}\right] \\ &= (\alpha_1 + \beta_{i,1})E\{p_{i,t}(j)\,[\delta_k(j) - p_{i,t}(j)]\}\end{aligned}$$

	Price Sensitivity			
	Heinz	Hunt's	Del Monte	House Brand
Heinz	–0.527	0.246	0.122	0.160
	(0.019)	(0.011)	(0.006)	(0.010)
Hunt's	0.246	–0.458	0.105	0.107
	(0.011)	(0.016)	(0.006)	(0.008)
Del	0.122	0.105	–0.314	0.087
Monte	(0.006)	(0.006)	(0.014)	(0.007)
House	0.160	0.107	0.087	–0.354
	(0.010)	(0.008)	(0.007)	(0.018)
	Display Sensitivity			
	Heinz	Hunt's	Del Monte	House Brand
	0.170	0.145	0.099	0.121
	(0.011)	(0.009)	(0.007)	(0.008)
	Feature Sensitivity			
	Heinz	Hunt's	Del Monte	House Brand
	0.163	0.133	0.082	0.110
	(0.013)	(0.010)	(0.008)	(0.009)

1. **Heinz has the largest aggregate market response to its own price changes (entries on the diagonal), followed by Hunt's, the House Brand, and Del Monte.**

2. **20% price reduction in the price of Heinz increases its choice share by 10.5%.**

3. **This 10.5% increase in choice share for Heinz from a 20% discount in its price comes at the expense of decreasing the choice share of Hunt's by 4.9%, of Del Monte by 2.4%, and of the House Brand by 3.2%.**

4. **The market share weighted mean choice probability increases by 14.6% for an in–store display and 13.5% for a feature advertisement.**

8.8 Multivariate Probit

1. **Pick/Don't Pick decision for many alternative**

2. **Person i, product j.**

3. $C_{i,j}$ **= 1 if j is selected, and 0 if it is not.**

4. **Random utility to person i for product j:**

$$Y_{i,j} = \mu_j + \epsilon_{i,j}.$$

5. **Pick j if $Y_{i,j} > 0$**

Constraints

1. **Error variances are 1.**
2. **Errors are correlated.**
3. **The covariance matric Σ for the error is a correlation matrix.**

MCMC

1. **Ignore the constraint on Σ during the MCMC.**

2. **Postprocess the iterates by:**

    - **Dividing μ_j by $sqrt(\sigma_{j,j}$.**

    - **Making Σ a correlation matrix.**

8.9 Summary

1. **The presentation of this chapter framed the models as a choice problem.**

2. **These models are also applicable to any situation that has nominal outcomes.**

3. **The logit and probit models assume different error structures. Which one to use? Good question, but it probably does not matter too much.**

References

- Albert, J. and Chib, S. (1993). "Bayesian Analysis of Binary and Polychotomous Response Data," *Journal of the American Statistical Association*, 88, 669–679.

- Chib, S. and Greenberg, E. (1995). Understanding the Metropolis–Hastings Algorithm. *The American Statistician,* 49, 327–335.

- Hastings, W. K. (1970). Monte Carlo sampling methods using Markov chains and their applications. *Biometrika,* 57, 97–109.

- Luce, R. D. (1959). *Individual Choice Behavior: A Theoretical Analysis,* New York: John Wiley & Sons.

- McCulloch, R. and Rossi, P. E. (1994). "An Exact Likelihood Analysis of the Multinomial Probit Model," *Journal of Econometrics,* 64, (1-2) 207-240.

- McFadden, D. (1974) Conditional Logit Analysis of Qualitative Choice Behavior. *Frontiers in Econometrics,* editor P. Zarembda, New York: Academic Press, pp. 105–142.

- Zeger, S. L., and Karim, M. R. (1991). Generalized linear models with random effects: a Gibbs sampling approach. *Journal of the American Statistical Association*, 86, 79–679.

Chapter 9

Summary

Foundations

- **Subjective Probability**
- **Coherence**
- **Decision Theory**
- **Complete Class Theorem**
- **Large Sample Theory**

Beta–Binomial & Conjugate Normal Models

- **Preliminaries**
  - **Binomial Distribution**
  - **Beta Distribution**
  - **Normal Distribution**
  - **Gamma and Inverted Gamma Distributions**
  - **T–Distribution**
- **Bayesian Inference**
  - **Joint Distribution**
  - **Marginal Distribution**
  - **Posterior Distribution**
  - **Predictive Distribution**

Linear Regression

- **Preliminaries**
  - **Multivariate Normal Distribution**
  - **Gamma and Inverted Gamma Distributions**
- **Bayesian Inference**
  - **Full Conditionals**
  - **MCMC**
- **Slice Sampling**
- **Autocorrelated Errors**

Multivariate Regression

- **Preliminaries**
  - **Matrix Facts**
  - **Matrix Normal Distribution**
  - **Wishart and Inverted Wishart Distributions**
- **Bayesian Inference**
  - **Full Conditionals**
  - **MCMC**

Hierarchical Bayes Regression: Interaction Model

- **Preliminaries**
  - **Application of multiple and multivariate regression**
- **Bayesian Inference**
  - **Within & Between Models**
  - **Full Conditionals**

Hierarchical Bayes Regression:

Mixture Model

- **Preliminaries**
  - **Multinomial Distribution**
  - **Dirichlet Distribution**
  - **Mixture Distributions**
- **Bayesian Inference**
  - **Latent Variables**

Probit Model

- **Preliminaries**
  - **Random Utility Model**
- **Bayesian Inference**
  - **Latent Variables**

Logit Model

- **Preliminaries**
  - **Extreme Value Distribution**
- **Bayesian Inference**
  - **Hastings–Metropolis Algorithm**

Conclusion

- **Good models include all major sources of uncertainty and variation.**
- **Bayesian inference explicitly account for these sources.**
- **MCMC has proven to be a flexible method of analyzing complex models.**
- **This course has presented the basic framework.**
- **As the complexity of your problems increase, you will want to go beyond the basics.**