3. A BRIEF REVIEW OF BAYESIAN INFERENCE

Most prediction methods based on "regularization" admit a Bayesian solution.

(If they do not, they are probably useless)

Crucial to understand Bayesianism when analyzing high-dimensional data

1

Rev. Thomas Bayes

1702 London, England
1761 Tunbridge Wells, Kent, England

1763. "An essay towards solving a problem in the doctrine of chances". *Philosophical Transactions of the Royal Society of London* **53**, 370-418.

Pierre-Simon Laplace

1749 Beaumont-en-Auge, France
1827 Paris, France

1774. "Mémoire sur la probabilité des causes par les événements". *Savants étranges* **6**, 621-656. *Oeuvres* **8**, 27-65

INVERSE PROBABILITY

HISTORICAL NOTES

- Karl Pearson (without knowing) used Bayes
- Fisher (likelihood, fiducial inference)
- Lack of admissibility of classical procedures (James-Stein)
- Revival: Neo-Bayesianism (Lindley, Box, Zellner)
- MCMC procedures (Metropolis, Geman and Geman)
- Bayesian methods in genetics: Haldane (1948), Dempfle (1977), Gianola and Fernando (1986)
- Explosion of Bayesianism in statistics: Gelfand and Smith (1990)
- Explosion in genetics as well

3

Bayesian methods in Genetics: today

-Classification of genotypes
-Molecular evolution
-Linkage mapping
-QTL cartography
-Genetic risk analysis
-Gaussian linear and non-linear models
: cross-sectional+ longitudinal univariate+ multivariate
-Generalized linear models
-Survival analysis
-Thick-tailed processes
-Mixtures
-Semi-parametrics
-Transcriptional analysis
-Structural equation modeling
-Bayesian proteomics with wavelets
-Methods for genomic selection
(the Bayesian Alphabet—A, B, C-pi, L… and more)
-Bayesian non-parametrics (Dirichlet process priors)

RED: animal breeders

4

THE BAYESIAN APPROACH IN A NUTSHELL

- All unknowns in statistical system treated as random
- Randomness reflects (typically) **subjective** uncertainty
- Can include as unknowns:
- The model (distribution, functional form)
- Its parameters (heritability, inbreeding coefficient)
- Genetic effects, number of QTL loci, marker effects
- Combine with what is known a priori with information from data: Bayesian learning
- Bayesian approach can also be used for developing predictors of future observations without taking inference too seriously

5

HOW DOES ONE DO THIS?

- Introduce a prior distribution for all unknowns (PRIOR)
- Define a distribution for the data under a certain model (LIKELIHOOD)
- Arrive at conditional distribution of all unknowns given data (POSTERIOR)
- Derive marginal or conditional posterior distributions of interest by standard probability theory
- Display summaries or entire distribution
- Interpret results probabilistically
- Example: the posterior probability of Ho is 8%

6

BAYES THEOREM: DISCRETE

- *M* disjoint hypotheses about some mechanism. Assign probabilities to the events "the hypothesis is true":

$$0 \le p(H_i) \le 1, \quad i = 1,2,\ldots,M,$$
$$p(H_i \cap H_j) = 0, \quad i \ne j,$$
$$\sum_{i=1}^{M} p(H_i) = 1.$$

Mutually exclusive

Exhaustive

7

- N observable effects. Given that a hypothesis holds, one observes events with probabilities

$$0 \le p(E_j|H_i) \le 1, \quad i = 1,2,\ldots,M, j = 1,2,\ldots,N,$$
$$p(E_j \cap E_{j'}|H_i) = 0, \quad j \ne j',$$
$$\sum_{j=1}^{N} p(E_j|H_i) = 1.$$

Probability distribution of events under hypothesis ("likelihood")

8

•Assume that events E and the hypotheses H have joint distribution:

$$p(H = H_i, E = E_j) = p(E_j|H_i)p(H_i)$$

•The conditional probability that a hypothesis holds, given the observed effects is:

Bayes theorem

Prior probability

$$p(H_i|E_j) = \frac{p(H=H_i, E=E_j)}{p(E_j)} = \frac{p(E_j|H_i)p(H_i)}{p(E_j)}$$

Posterior probability

Likelihood

Marginal distribution of data

$$p(E_j) = \sum_{i=1}^{M} p(E_j|H_i)p(H_i) = E_H[p(E_j|H_i)]$$

9

THE "PROPORTIONAL TO" REPRESENTATION:

$$p(H_i|E_j) = \frac{p(E_j|H_i)p(H_i)}{E_H[p(E_j|H_i)]}$$

$$\propto p(E_j|H_i)p(H_i).$$

The "Pac-Man" operator: eats anything that does not depend on H_i

48

BAYESIAN LEARNING: DISCRETE

•Let 2 bits of evidence accumulate. Then:

$$p(H_i|E'_{j'},E_j) = \frac{p(E'_{j'},E_j|H_i)p(H_i)}{E_H[p(E'_{j'},E_j|H_i)]}$$

$$\propto p(E'_{j'},E_j|H_i)p(H_i)$$

$$\propto p(E'_{j'}|E_j,H_i)p(E_j|H_i)p(H_i)$$

$$\propto p(E'_{j'}|E_j,H_i)p(H_i|E_j).$$

•Prior before bit 2 is posterior after bit 1

•If, given the hypothesis, the 2 bits of evidence are independent:

$$p(E'_{j'}|E_j,H_i)p(E_j|H_i) = p(E'_{j'}|H_i)p(E_j|H_i)$$

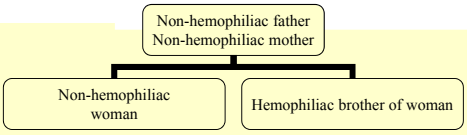
Conditional independence

49

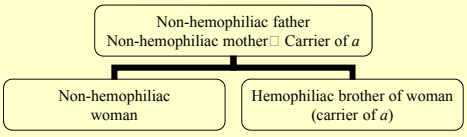
EXAMPLE OF DISCRETE PROBLEM: HEMOPHILIA IN HUMANS

(1995). Hemophilia is a genetic disease in humans. The locus responsible for its expression is located in the sex chromosomes (these are denoted as XX in women, and XY in men). The condition is observed in women only in double recessive individuals (aa), and in men that are carriers of the a allele in the X-chromosome.

DATA



IMPLICATION



Problem: Probability(woman is carrier)= Probability(θ=1) ?

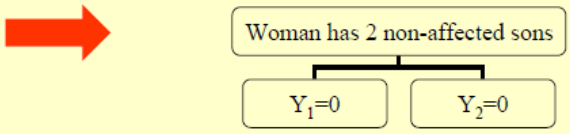
PRIOR IMPLIED BY THIS INFORMATION

(mother must be *Aa*)

$$\Pr(\theta = 1) = \Pr(\theta = 0) = \frac{1}{2}.$$

12

MORE DATA BECOME AVAILABLE



Given that $\theta = 1$, *the probability of the observed data is*

$$\Pr(Y_1 = 0, Y_2 = 0|\theta = 1)$$
$$= \Pr(Y_1 = 0|\theta = 1)\Pr(Y_2 = 0|\theta = 1) = \left(\frac{1}{2}\right)^2 = \frac{1}{4}.$$

On the other hand, if she is not a carrier $(\theta = 0)$,

$$\Pr(Y_1 = 0, Y_2 = 0|\theta = 0)$$
$$= \Pr(Y_1 = 0|\theta = 0)\Pr(Y_2 = 0|\theta = 0) = 1 \times 1 = 1,$$

THE DATA CONFER 4 TIMES MORE LIKELIHOOD TO NON-CARRIER HYPOTHESIS

POSTERIOR PROBABILITIES

$$\Pr(\theta = 1|Y_1 = 0, Y_2 = 0) = \frac{\Pr(\theta = 1)\Pr(Y_1 = 0, Y_2 = 0|\theta = 1)}{\Pr(Y_1 = 0, Y_2 = 0)}$$
$$= \frac{\Pr(\theta = 1)\Pr(Y_1 = 0, Y_2 = 0|\theta = 1)}{\sum_{i=0}^{1}\Pr(\theta = i)\Pr(Y_1 = 0, Y_2 = 0|\theta = i)}$$
$$= \frac{\frac{1}{2}\frac{1}{4}}{\frac{1}{2}1 + \frac{1}{2}\frac{1}{4}} = \frac{1}{5}$$

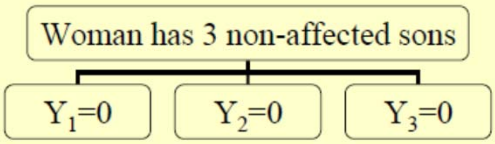
and

$$\Pr(\theta = 0|Y_1 = 0, Y_2 = 0) = 1 - \frac{1}{5} = \frac{4}{5}.$$

➔WE MOVED FROM **0.5: 0.5** TO **0.2: 0.8**
➔SHARPER STATE OF KNOWLEDGE BUT CANNOT RULE OUT HYPOTHESIS WOMAN IS A CARRIER
➔STILL UNCERTAINTY…MORE DATA NEEDED

14

EVEN MORE DATA BECOME AVAILABLE…



Using posterior as prior for new data, assuming conditional independence

$$\Pr(\theta = 1|Y_1 = 0, Y_2 = 0, Y_3 = 0) = \frac{\frac{1}{5}\Pr(Y_3 = 0|\theta = 1)}{\frac{1}{5}\Pr(Y_3 = 0|\theta = 1) + \frac{4}{5}\Pr(Y_3 = 0|\theta = 0)} = \frac{\frac{1}{5}\frac{1}{2}}{\frac{1}{5}\frac{1}{2} + \frac{4}{5}1} = \frac{1}{9}.$$

Using prior before Any progeny data, And combining all information

$$\Pr(\theta = 1|Y_1 = 0, Y_2 = 0, Y_3 = 0) = \frac{\frac{1}{2}\cdot\left(\frac{1}{2}\right)^3}{\frac{1}{2}\cdot\left(\frac{1}{2}\right)^3 + \frac{1}{2}\cdot(1)^3} = \frac{1}{9}.$$

WOMAN COULD STILL BE A CARRIER!

$$\Pr(\theta = 1|Y_1 = 0, Y_2 = 0, \ldots, Y_N = 0) = \frac{\Pr(\theta = 1)(\frac{1}{2})^n}{\Pr(\theta = 1)(\frac{1}{2})^n + \Pr(\theta = 0)1^n} = \frac{(\frac{1}{2})^n}{(\frac{1}{2})^n + \frac{\Pr(\theta=0)}{\Pr(\theta=1)}} = \frac{1}{1 + \frac{\Pr(\theta=0)}{\Pr(\theta=1)}2^n}$$

TENDS TO 0 AS n GOES TO INFINITY. HOWEVER,

$$\Pr(\theta = 1|Y_1 = 0, Y_2 = 0, \ldots, Y_N = 0, Y_{N+1} = 1) = 1$$

IF WOMAN HAS AT LEAST ONE HEMOPHILIAC SON.

54

BAYES THEOREM: CONTINUOUS

- Evidence is now given by a vector of observations y
- Hypothesis is a vector of unknowns θ
- A probability model M poses joint distribution $[\theta, y|M]$ with density

$$h(\theta, \mathbf{y}) = g(\theta)f(\mathbf{y}|\theta) = m(\mathbf{y})p(\theta|\mathbf{y})$$

- Assume that both the unknowns and the parameter are continuous-valued

55

BAYES THEOREM IN A NUTSHELL

$$p(\theta|\mathbf{y}) = \frac{g(\theta)f(\mathbf{y}|\theta)}{m(\mathbf{y})} \propto g(\theta)f(\mathbf{y}|\theta)$$

Prior density

Likelihood function

Marginal data density

Posterior density

18

•**Prior density** $g(\theta)$

•**Sampling density** (likelihood function when viewed as function of θ

$f(\mathbf{y}|\theta)$ Pac-man operator $L(\theta|\mathbf{y})$

•**Marginal density of the observations**

Prior must be **proper** for integral to exist

$$m(\mathbf{y}) = \int h(\theta,\mathbf{y})d\theta = \int f(\mathbf{y}|\theta)g(\theta)d\theta = E_\theta[f(\mathbf{y}|\theta)]$$

•**Posterior density**

$$p(\theta|\mathbf{y}) = \frac{g(\theta)f(\mathbf{y}|\theta)}{m(\mathbf{y})} \propto g(\theta)f(\mathbf{y}|\theta)$$

equivalently

$$p(\theta|\mathbf{y}) = \frac{g(\theta)L(\theta|\mathbf{y})}{\int g(\theta)L(\theta|\mathbf{y})d\theta}$$

56

BAYESIAN INFERENCE and MCMC
(can fit any model)

PRIOR **DATA** **POSTERIOR**

Most of the times the prior comes "out of the blue" 20

THUS: THE ANTI-BAYESIAN ARGUMENT…

PRIOR

DATA

POSTERIOR

Analysis with Prior 1: collecting more and more data….

Analysis with Prior 2: collecting more and more data….

Implications

- This is called “asymptotic domination” of the prior by the data (likelihood)
- For parameters on which there is a lot of information from the data, the prior matters little
- Prior may be influential in small samples; worthwhile to investigate sensitivity
- What is a small sample?
- Even if prior matters little, Bayesian approach allows to use probability theory to measure uncertainty

What about asymptotics in situations where n<<<p, and where there are strong non-linearities? Can one learn about marker effects?

24

POSTERIOR AS A STANDARDIZED LIKELIHOOD

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

Constant not involving parameters

likelihood

$$p(\boldsymbol{\theta}|\mathbf{y}) = \frac{kl(\boldsymbol{\theta}|\mathbf{y})p(\boldsymbol{\theta})}{p(\mathbf{y})}$$

If prior is flat:

$$p(\boldsymbol{\theta}|\mathbf{y}) \propto l(\boldsymbol{\theta}|\mathbf{y})$$

If likelihood is integrable

$$p(\boldsymbol{\theta}|\mathbf{y}) = \frac{l(\boldsymbol{\theta}|\mathbf{y})}{\int l(\boldsymbol{\theta}|\mathbf{y})d\boldsymbol{\theta}}$$

25

EXAMPLE: CONTINUOUS PROBLEM- INFERRING THE MEAN OF A NORMAL DISTRIBUTION WITH KNOWN VARIANCE

Sampling model → $y_1, y_2, \ldots, y_N \sim NIID(\mu, \sigma^2)$

$$p(y_1, y_2, \ldots, y_N|\mu, \sigma^2) = \prod_{i=1}^{N} \frac{1}{\sqrt{2\pi\sigma^2}} \exp\left[-\frac{1}{2\sigma^2}(y_i - \mu)^2\right]$$

$$= \left(\frac{1}{\sqrt{2\pi\sigma^2}}\right)^N \exp\left[-\frac{1}{2\sigma^2}\sum_{i=1}^{N}(y_i - \mu)^2\right]$$

$$= \left(\frac{1}{\sqrt{2\pi\sigma^2}}\right)^N \exp\left[-\frac{1}{2\sigma^2}\sum_{i=1}^{N}(y_i - \bar{y})^2 - \frac{N}{2\sigma^2}(\mu - \bar{y})^2\right]$$

Maximum likelihood estimator of μ $\hat{\mu} = \bar{y}$

Part conferring likelihood to μ

Frequentist distribution of ML estimator

$$\hat{\mu} \sim N\left(\mu, \frac{\sigma^2}{n}\right)$$

DISCUSS

59

Normal distribution with known variance: continued

$$p(\mu|y_1,y_2,\ldots,y_N,\sigma^2) \propto p(y_1,y_2,\ldots,y_N|\mu,\sigma^2)p(\mu)$$

Flat prior

$$p(\mu|y_1,y_2,\ldots,y_N,\sigma^2) \propto p(y_1,y_2,\ldots,y_N|\mu,\sigma^2)$$

$$p(\mu|y_1,y_2,\ldots,y_N,\sigma^2) \propto \prod_{i=1}^{N} \exp\left[-\frac{1}{2\sigma^2}(y_i-\mu)^2\right]$$

$$\propto \exp\left[-\frac{1}{2\sigma^2}\sum_{i=1}^{N}(y_i-\mu)^2\right]$$

$$\propto \exp\left[-\frac{1}{2\sigma^2}\sum_{i=1}^{N}(y_i-\bar{y})^2\right]\exp\left[-\frac{N}{2\sigma^2}(\mu-\bar{y})^2\right]$$

$$\propto \exp\left[-\frac{N}{2\sigma^2}(\mu-\bar{y})^2\right]$$ Kernel of density

$$p(\mu|\mathbf{y},\sigma^2) = \frac{\exp\left[-\frac{N}{2\sigma^2}(\mu-\bar{y})^2\right]}{\int \exp\left[-\frac{N}{2\sigma^2}(\mu-\bar{y})^2\right]d\mu}$$

$$= \frac{\exp\left[-\frac{N}{2\sigma^2}(\mu-\bar{y})^2\right]}{\sqrt{2\pi\frac{\sigma^2}{N}}}$$

$$\mu|\mathbf{y},\sigma^2 \sim N\left(\bar{y},\frac{\sigma^2}{N}\right)$$

27

Another Bayesian treatment: uniform prior for μ (all values in the *(a,b)* range are equally plausible, a priori)

$$p(\mu) = \frac{1}{b-a}$$ **Note**: only values in *(a,b)* are allowed

Posterior density

Known

$$p(\mu|y_1,y_2,\ldots,y_N,\sigma^2) = \frac{p(y_1,y_2,\ldots,y_N|\mu,\sigma^2)p(\mu)}{\int_a^b p(y_1,y_2,\ldots,y_N|\mu,\sigma^2)p(\mu)d\mu}$$

$$p(\mu|y_1,y_2,\ldots,y_N,\sigma^2) = \frac{\left(\frac{1}{\sqrt{2\pi\sigma^2}}\right)^N \exp\left[-\frac{1}{2\sigma^2}\sum_{i=1}^{N}(y_i-\bar{y})^2-\frac{N}{2\sigma^2}(\mu-\bar{y})^2\right]\frac{1}{b-a}}{\int_a^b \left(\frac{1}{\sqrt{2\pi\sigma^2}}\right)^N \exp\left[-\frac{1}{2\sigma^2}\sum_{i=1}^{N}(y_i-\bar{y})^2-\frac{N}{2\sigma^2}(\mu-\bar{y})^2\right]\frac{1}{b-a}d\mu}$$

28

Doing the algebra (hard way):

$$p(\mu|y_1,y_2,\ldots,y_N,\sigma^2) = \frac{\exp\left[-\frac{N}{2\sigma^2}(\mu-\overline{y})^2\right]}{\int_a^b \exp\left[-\frac{N}{2\sigma^2}(\mu-\overline{y})^2\right]d\mu}$$

$$\int_a^b \exp\left[-\frac{N}{2\sigma^2}(\mu-\overline{y})^2\right]d\mu = \int_a^b \left(\frac{1}{\sqrt{2\pi\frac{\sigma^2}{N}}}\right)\left(\frac{1}{\sqrt{2\pi\frac{\sigma^2}{N}}}\right)^{-1}\exp\left[-\frac{N}{2\sigma^2}(\mu-\overline{y})^2\right]d\mu$$

$$= \sqrt{\frac{N}{2\pi\sigma^2}}\int_a^b \left(\frac{1}{\sqrt{2\pi\frac{\sigma^2}{N}}}\right)\exp\left[-\frac{N}{2\sigma^2}(\mu-\overline{y})^2\right]d\mu$$

$$= \sqrt{\frac{N}{2\pi\sigma^2}}\left[\Phi\left(\frac{b-\overline{y}}{\sqrt{\frac{\sigma^2}{N}}}\right)-\Phi\left(\frac{a-\overline{y}}{\sqrt{\frac{\sigma^2}{N}}}\right)\right]$$

Note: -1

29

$$p(\mu|y_1,y_2,\ldots,y_N,\sigma^2) = \frac{\sqrt{\frac{1}{2\pi\frac{\sigma^2}{N}}}\exp\left[-\frac{N}{2\sigma^2}(\mu-\overline{y})^2\right]}{\left[\Phi\left(\frac{b-\overline{y}}{\sqrt{\frac{\sigma^2}{N}}}\right)-\Phi\left(\frac{a-\overline{y}}{\sqrt{\frac{\sigma^2}{N}}}\right)\right]}$$

$$\mu|\mathbf{y},\sigma^2 \sim TN_{(a,b)}\left(\overline{y},\frac{\sigma^2}{N}\right)$$

Posterior distribution is truncated normal

Parameters in the absence of truncation

30

Doing the algebra with Pac-Man (easy way):

Start all over

$$p(\mu|y_1,y_2,\ldots,y_N,\sigma^2) = \frac{\left(\frac{1}{\sqrt{2\pi\sigma^2}}\right)^N \exp\left[-\frac{1}{2\sigma^2}\sum_{i=1}^{N}(y_i-\bar{y})^2-\frac{N}{2\sigma^2}(\mu-\bar{y})^2\right]\frac{1}{b-a}}{\int_a^b\left(\frac{1}{\sqrt{2\pi\sigma^2}}\right)^N \exp\left[-\frac{1}{2\sigma^2}\sum_{i=1}^{N}(y_i-\bar{y})^2-\frac{N}{2\sigma^2}(\mu-\bar{y})^2\right]\frac{1}{b-a}d\mu}$$

Pac-Man is allowed to eat anything not depending on μ, i.e., eats all symbols or functions to the right of the "bar

➔Can eat the denominator, since μ is integrated out

➔Can eat several things in the numerator, leading to

$$p(\mu|y_1,y_2,\ldots,y_N,\sigma^2) \propto \exp\left[-\frac{N}{2\sigma^2}(\mu-\bar{y})^2\right]$$

Since only values in *(a,b)* are allowed➔ Posterior is TN 64

Example: infer additive genetic of an individual

- *n* measurements taken on single individual. No permanent environmental effects
- Model for *ith* measurement is $y_i = \mu + a + e_i$
- Assume $a \sim N(0, v_a)$ *and* $e \sim N(0, v_e)$ independent
- Conditional distribution of phenotype, given μ and a

$$[y|\mu, a, v_a, v_e] \sim N(\mu + a, v_e)$$

Vector of 1's

- Conditional distribution of *n* measurements

$$[\mathbf{y}|\mu, a, v_a, v_e] \sim N(\mathbf{1}\mu + \mathbf{1}a, \mathbf{I}v_e)$$

- Suppose variance components known
- Unknowns➔ *a* and μ Assume joint prior density:

$$p(\mu, a|\mu_0, v_0, v_a) = p(\mu|\mu_0, v_0)p(a|v_a) = N(\mu_0, v_0)N(0, v_a)$$

32

Example: continued

•Joint posterior density

$$p(\mu, a | y_1, y_2, \ldots, y_n, \mu_0, v_0, v_a, v_e)$$

$$\propto \prod_{i=1}^{n} p(y_i | \mu, a, v_e) p(a | v_a) p(\mu | \mu_0, v_0)$$

$$\propto \prod_{i=1}^{n} exp\left[-\frac{(y_i - \mu - a)^2}{2v_e} \right] exp\left[-\frac{a^2}{2v_a} \right] exp\left[-\frac{(\mu - \mu_0)^2}{2v_0} \right]$$

$$\propto exp\left[-\frac{\sum_{i=1}^{n}(y_i - \mu - a)^2}{2v_e} - \frac{a^2}{2v_a} \right] exp\left[-\frac{(\mu - \mu_0)^2}{2v_0} \right].$$

66

Formulae for combining quadratic forms

•SCALAR

$$M(z-m)^2 + B(z-b)^2 = (M+B)(z-c)^2 + \frac{MB}{M+B}(m-b)^2$$

$$c = (M+B)^{-1}(Mm + Bb)$$

Weighted ave. of m and b

•VECTORIAL

$$(z-m)'M(z-m) + (z-b)'B(z-b)$$

$$= (z-c)'(M+B)(z-c) + (m-b)'M(M+B)^{-1}B(m-b)$$

67

Example: Conditional posterior density of additive genetic effect

Reciprocal of variance

Mean

$$p(a|\mu,\mathbf{y},\mu_0,v_0,v_a,v_e)$$

$$\propto exp\left[-\frac{1}{2}\left(\frac{n}{v_e}+\frac{1}{v_a}\right)\left\{a-\left(\frac{n}{v_e}+\frac{1}{v_a}\right)^{-1}\left[\frac{n}{v_e}(\bar{y}-\mu)\right]\right\}^2\right]$$

- Normal!

➔ $$E(a|\mu,\mathbf{y},\mu_0,v_0,v_a,v_e)=\left(\frac{n}{v_e}+\frac{1}{v_a}\right)^{-1}\left[\frac{n}{v_e}(\bar{y}-\mu)\right]$$

$$=\frac{v_a}{v_a+\frac{v_e}{n}}(\bar{y}-\mu)$$

$$=\frac{n}{n+\frac{1-h^2}{h^2}}(\bar{y}-\mu)$$

"EBV"

"PEV"

➔ $$Var(a|\mu,\mathbf{y},\mu_0,v_0,v_a,v_e)=\left(\frac{n}{v_e}+\frac{1}{v_a}\right)^{-1}=v_e\left(n+\frac{1-h^2}{h^2}\right)^{-1}$$

35

EXAMPLE OF CONTINUOUS CASE
Inferring the Poisson parameter (ML)

N independent samples

$$p(y_1,y_2,\ldots,y_N|\lambda)=\frac{\lambda^{\sum y_i}e^{-N\lambda}}{\prod y_i!}$$ likelihood $$l(\lambda|\mathbf{y})\propto\lambda^{\sum y_i}e^{-N\lambda}$$

Log-likelihood

$$L(\lambda|\mathbf{y})=K+\sum y_i\log(\lambda)-N\lambda$$

$$\frac{dL(\lambda|\mathbf{y})}{d\lambda}=\frac{\sum y_i}{\lambda}-N$$

$$\boxed{MLE(\lambda)=\frac{\sum y_i}{N}}$$

$$-E\frac{d^2L(\lambda|\mathbf{y})}{(d\lambda)^2}=E\left(\frac{\sum y_i}{\lambda^2}\right)=\frac{N}{\lambda}$$

$$\boxed{\widehat{AsyVar}(\hat{\lambda})=\frac{\hat{\lambda}}{N}}$$

69

Inferring the Poisson parameter (Bayes)

Gamma prior

$$p(\lambda|\alpha,\beta) = \frac{1}{\beta^{\alpha}\Gamma(\alpha)}\lambda^{\alpha-1}\exp\left(-\frac{\lambda}{\beta}\right)$$

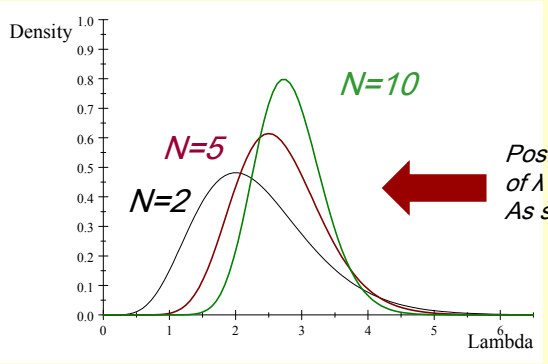
$$\begin{aligned} p(\lambda|\mathbf{y},\alpha,\beta) &\propto l(\lambda|\mathbf{y})p(\lambda|\alpha,\beta) \\ &\propto \lambda^{\sum y_i} e^{-N\lambda}\lambda^{\alpha-1}\exp\left(-\frac{\lambda}{\beta}\right) \\ &\propto \lambda^{\sum y_i+\alpha-1}\exp\left[-\left(N+\frac{1}{\beta}\right)\lambda\right] \\ &\propto \lambda^{N\bar{y}+\alpha-1}\exp\left[-\frac{\lambda}{\left(\frac{\beta}{N\beta+1}\right)}\right] \end{aligned}$$

Posterior is Gamma as well (Conjugacy)

$$\lambda|\mathbf{y},\alpha,\beta \sim Gamma\left(N\bar{y}+\alpha, \left(\frac{N\beta+1}{\beta}\right)^{-1}\right)$$

$$p(\lambda|\mathbf{y},\alpha,\beta) = \frac{\left(\frac{N\beta+1}{\beta}\right)^{N\bar{y}+\alpha}}{\Gamma(N\bar{y}+\alpha)}\lambda^{N\bar{y}+\alpha-1}e^{-\left(\frac{N\beta+1}{\beta}\right)\lambda}$$

Suppose the mean of the observations is 3 and that N=2,5,10. α=β=1. The posterior densities look like



Posterior density of λ tends to normal As sample size increas

$$E(\lambda|\alpha,\beta) = \alpha\beta; \quad Var(\lambda|\alpha,\beta) = \alpha\beta^2$$

$$E(\lambda|\mathbf{y},\alpha,\beta) = (N\bar{y}+\alpha)\left(\frac{\beta}{N\beta+1}\right)$$

$$= \left(\frac{N\beta}{N\beta+1}\right)\bar{y} + \left(\frac{1}{N\beta+1}\right)\alpha\beta$$

$$= \left(\frac{N}{N+\frac{1}{\beta}}\right)\bar{y} + \left(\frac{\frac{1}{\beta}}{N+\frac{1}{\beta}}\right)\alpha\beta$$

1) Weighted ave. of MLE and prior mean
2) When N goes to infinity, expectation tend to MLE.

$$Var(\lambda|\alpha,\beta) = (N\bar{y}+\alpha)\left(\frac{\beta}{N\beta+1}\right)^2$$

$$= (N\bar{y}+\alpha)\left(\frac{1}{N+\frac{1}{\beta}}\right)^2$$

$$= N\left(\frac{1}{N+\frac{1}{\beta}}\right)^2 MLE(\lambda) + \left(\frac{1}{N+\frac{1}{\beta}}\right)^2\alpha$$

$$\lim_{N\to\infty} Var(\lambda|\alpha,\beta) = \frac{MLE(\lambda)}{N}$$

Tends to AsyVar of MLE estimator

Side note: Independence versus conditional independence

Random cluster (e.g., family) effect

$$y_{ij} = \mu + s_i + e_{ij}$$

$$s_i \sim NIID(0,\sigma_s^2)$$

$$e_{ij} \sim NIID(0,\sigma_e^2)$$

$$s_i, e_{ij} \text{ independent } \forall\, i,j$$

$$y_{ij} \sim NID(\mu, \sigma_s^2+\sigma_e^2)$$

$$Cov(y_{ij}, y_{ij'}) = \sigma_s^2$$

$$\begin{bmatrix} y_{ij} \\ y_{ij'} \end{bmatrix} \sim N_2\left(\begin{bmatrix} \mu \\ \mu \end{bmatrix}, \begin{bmatrix} \sigma_s^2+\sigma_e^2 & \sigma_s^2 \\ \sigma_s^2 & \sigma_s^2+\sigma_e^2 \end{bmatrix}\right)$$

Pairs of members of same cluster have bivariate normal distribution:
➔ Observations from same cluster **are not** independent

$$y_{ij} = \mu + s_i + e_{ij}$$
$$s_i \sim NIID(0, \sigma_s^2)$$
$$e_{ij} \sim NIID(0, \sigma_e^2)$$
$$s_i, e_{ij} \text{ independent } \forall\, i,j$$

Conditional distribution

$$E(y_{ij}|s_i) = \mu + s_i$$
$$Var(y_{ij}|s_i) = \sigma_e^2$$
$$Cov(y_{ij}, y_{ij'}|s_i) = Cov(\mu + s_i + e_{ij}, \mu + s_i + e_{ij'}|s_i)$$
$$= Cov(e_{ij}, e_{ij'}|s_i) = Cov(e_{ij}, e_{ij'}) = 0$$

$$p(y_{ij}, y_{ij'}|s_i) = p(y_{ij}|s_i)p(y_{ij}|s_i) = N(\mu + s_i, \sigma_e^2)N(\mu + s_i, \sigma_e^2)$$

GIVEN THE CLUSTER EFFECT, OBSERVATIONS IN SAME CLUSTER **ARE** CONDITIONALLY INDEPENDENT!

41

Probability model of the observations **and** of the unobserved cluster effect

$$p(y_{ij}, y_{ij'}, s_i) = p(y_{ij}, y_{ij'}|s_i)p(s_i)$$
$$= p(y_{ij}|s_i)p(y_{ij}|s_i)p(s_i)$$

$$p(y_{ij}, y_{ij'}) = \int_{-\infty}^{\infty} p(y_{ij}|s_i)p(y_{ij}|s_i)p(s_i)ds_i$$

Process called "deconditioning"

$$= \int_{-\infty}^{\infty} N(\mu + s_i, \sigma_e^2)N(\mu + s_i, \sigma_e^2)N(0, \sigma_s^2)ds_i$$

$$= N_2\left(\begin{bmatrix} \mu \\ \mu \end{bmatrix}, \begin{bmatrix} \sigma_s^2 + \sigma_e^2 & \sigma_s^2 \\ \sigma_s^2 & \sigma_s^2 + \sigma_e^2 \end{bmatrix}\right)$$

MARGINAL DISTRIBUTIONS ARE OFTEN MORE COMPLEX

42

Conditional posterior distribution of the unobserved cluster effect, given the parameters

$$y_{ij} = \mu + s_i + e_{ij}$$

Joint distribution for $n_i=2$

Can be viewed as "population" or uncertainty distribution parameters

$$\begin{pmatrix} s_i \\ y_{i1} \\ y_{i2} \end{pmatrix} \sim N\left(\begin{bmatrix} 0 \\ \mu \\ \mu \end{bmatrix}, \begin{bmatrix} \sigma_s^2 & \sigma_s^2 & \sigma_s^2 \\ \sigma_s^2 & \sigma_s^2 + \sigma_e^2 & \sigma_s^2 \\ \sigma_s^2 & \sigma_s^2 & \sigma_s^2 + \sigma_e^2 \end{bmatrix} \right)$$

Conditional posterior distribution can be shown to be

$$s_i | y_{i1}, y_{i2} \sim N(\hat{s}, \hat{v}_s)$$

43

$$\hat{s} = E(s_i | y_{i1}, y_{i2}) = \frac{2}{2 + \frac{\sigma_e^2}{\sigma_s^2}} \left(\frac{y_{i1} + y_{i2}}{2} - \mu \right) = \frac{2}{2 + \frac{4-h^2}{h^2}} (\bar{y}_i - \mu)$$

$$\hat{v}_s = \frac{\sigma_e^2}{2 + \frac{\sigma_e^2}{\sigma_s^2}}$$

$$\hat{s} = \frac{1}{\frac{1}{\sigma_s^2} + \frac{n}{\sigma_e^2}} \left[\frac{1}{\sigma_s^2} 0 + \frac{n}{\sigma_e^2} (\bar{y}_i - \mu) \right]$$

Weighted average of data and of "prior"

$$= \frac{1}{\frac{1}{\sigma_s^2} + \frac{n}{\sigma_e^2}} \frac{n}{\sigma_e^2} (\bar{y}_i - \mu) = \frac{n}{n + \frac{\sigma_e^2}{\sigma_s^2}} (\bar{y}_i - \mu)$$

$$\hat{v}_s = \frac{1}{\frac{1}{\sigma_s^2} + \frac{n}{\sigma_e^2}} = \frac{\sigma_s^2 \sigma_e^2}{\sigma_e^2 + n\sigma_s^2}$$

$$= \frac{\sigma_e^2}{n + \frac{\sigma_e^2}{\sigma_s^2}}$$

44

- Suppose heritability is 0.1 $\sigma_e^2 = 1, \mu = 2.5$
- 2 unrelated sires with $n_1 = 4, \overline{y}_1 = 3$ and $n_2 = 8, \overline{y}_i = 2.95$

$$\hat{s}_1 = \frac{4}{4 + \frac{4-0.1}{0.1}}(3 - 2.5) = 4.651162790\,70 \times 10^{-2}$$

$$\hat{s}_2 = \frac{8}{8 + \frac{4-0.1}{0.1}}(2.95 - 2.5) = 7.659574468\,09 \times 10^{-2}$$

$$\hat{v}_1 = \frac{1}{4 + \frac{4-0.1}{0.1}} = 2.325581395\,35 \times 10^{-2}$$

$$\hat{v}_2 = \frac{1}{8 + \frac{4-0.1}{0.1}} = 2.127659574\,47 \times 10^{-2}$$

➔ *Progeny of sire 1 have better performance*
➔*Sire 2 has higher posterior mean (EBV)*
➔*Sire 2 has more "information" (less shrinkage)*

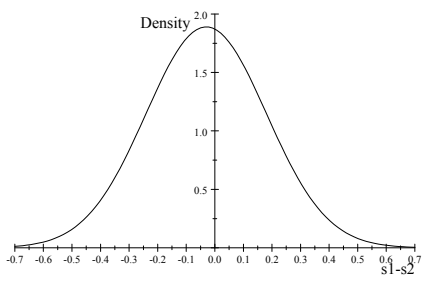
45

What is the ***strength of the evidence*** *that sire 2 is better than 1?*

1) Consider posterior distribution of sire 1-sire 2 . Because sires are unrelated and parameters are known, the 2 sires have independent conditional posterior distributions

$$s_1 - s_2 | \overline{y}_1, \overline{y}_2 \sim N\left(\hat{s_1} - \hat{s_2}, \hat{v}_1 + \hat{v}_2\right)$$

$$s_1 - s_2 | \overline{y}_1 = 3, \overline{y}_2 = 2.6 \sim N(-3.008411677\,39 \times 10^{-2}, 4.453240969\,82 \times 10^{-2})$$



Not enough difference to state that the 2 sires differ at all

46

2) Consider posterior distribution of sire 1/sire 2 . If sires not different this posterior should be centered at 1.

Problem!!! The distribution cannot be found analytically

$$s_1/s_2|\overline{y}_1, \overline{y}_2 \sim ??????$$

Solution:

$$s_1|\overline{y}_1 \sim N(s_1, \hat{v}_1)$$

$$s_2|\overline{y}_2 \sim N(s_2, \hat{v}_2)$$

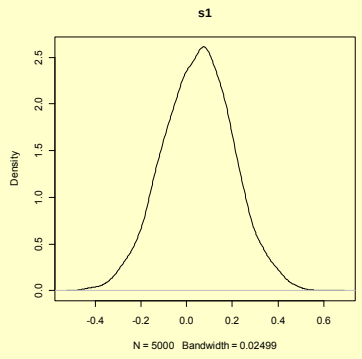
are independent distributions. Hence, we can sample the two random variable and form draws of

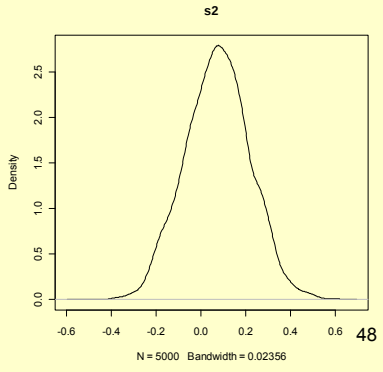
$$s_1/s_2|\overline{y}_1, \overline{y}_2$$

47

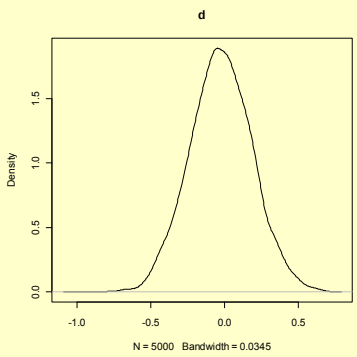
Sampling from the posterior distributions using R

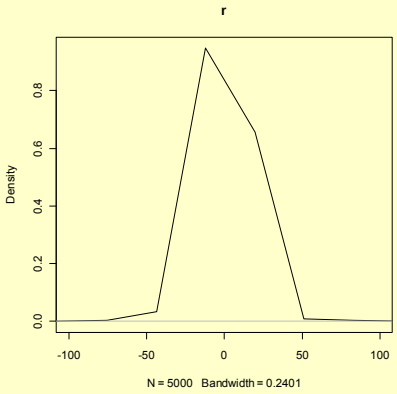
```
> s1<-rnorm(5000,0.0465,sqrt(0.0232))
> s2<-rnorm(5000,0.0766,sqrt(0.02128))
> d<-s1-s2
>r<-s1/s2
> plot(density(s1),main="s1")
> plot(density(s2),main="s2")
> plot(density(d),main="d")
> plot(density(r),main="r",xlim=c(-100,100))
```





48






```
> summary(d)
   Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
-0.99100 -0.16390 -0.02539 -0.02357 0.11820 0.68890
> mean(d)
[1] -0.02357277
> var(d)
[1] 0.04484314
```

Monte Carlo estimates of mean and variance of conditional distribution

```
> summary(r)
     Min.   1st Qu.    Median      Mean   3rd Qu.      Max.
-1339.0000    -0.7747     0.1940     3.6470     1.1890
14820.0000
> mean(r)
[1] 3.647075
> var(r)
[1] 46015.02
```

MUST BE CAREFUL HERE…

*ILLUSTRATION THAT WE CAN "HIT" THE TRUE POSTERIOR DISTRIBUTION OF **d** BY TAKING MORE SAMPLES*

$$s_1 - s_2 | \overline{y}_1, \overline{y}_2 \sim N\left(\widehat{s_1} - \widehat{s_2}, \hat{v}_1 + \hat{v}_2\right)$$

$$s_1 - s_2 | \overline{y}_1 = 3, \overline{y}_2 = 2.6 \sim N(-3.008411677739 \times 10^{-2}, 4.453240969682 \times 10^{-2})$$

1) #Samples=5000

```
> mean(d)
[1] -0.02357277
> var(d)
[1] 0.04484314
```

2) #Samples=20000

```
> mean(d)
[1] -0.03136750
> var(d)
[1] 0.04511715
```

3) #Samples=200000

```
> mean(d)
[1] -0.03037269
> var(d)
[1] 0.04457762
```

4) #Samples=1000000

```
> mean(d)
[1] -0.02993406
> var(d)
[1] 0.04450943
```

EXAMPLE: A SIMPLE BAYESIAN SURVIVAL ANALYSIS MODEL

51

BASICS

- Non-negative random variable *T*
- Typically: time-to-event
  - ➔ death
  - ➔ onset of disease
  - ➔ successful fertilization
  - ➔ failure of component
- Censoring (right) can occur. For *n* individuals (*i=1,2,…,n*) observe

$y_i = \min(t_i, v_i)$ ← "True" failure time / "Censoring point"

$$c_i = \begin{cases} 1 \text{ if } t_i \le v_i & \leftarrow \text{uncensored observation} \\ 0 \text{ if } t_i > v_i & \leftarrow \text{censored} \end{cases}$$

52

Distribution function of failure time

$$F(t) = \Pr(T \le t) = \int_0^t f(u)du$$

Survivor function

$$S(t) = 1 - F(t) = \Pr(t > t)$$

Hazard function

$$h(t) = \frac{f(t)}{S(t)} \Rightarrow f(t) = h(t)S(t)$$

$$\frac{d}{dt}\log[S(t)] = \frac{1}{S(t)}\frac{d}{dt}S(t)$$

$$= -\frac{f(t)}{S(t)} = -h(t)$$

53

$$h(t) = -\frac{d}{dt}\log[S(t)]$$

Integrated hazard

$$\int_0^t h(u)du = -\log[S(t)]$$

$$S(t) = \exp\left[-\int_0^t h(u)du\right]$$

Cumulative hazard

$$H(t) = \int_0^t h(u)du$$

Representation of density of failure times

$$f(t) = h(t)\exp\left[-\int_0^t h(u)du\right]$$

54

Parametric exponential model: homogeneous population

$$f(y_i|\lambda) = \lambda \exp(-\lambda y_i)$$

Exponential density

$$S(y_i|\lambda) = 1 - \int_0^{y_i} \lambda \exp(-\lambda u) du$$

$$= \exp(-\lambda y_i)$$

Survival function

Note that hazard (ratio between f and S) is constant= λ

Likelihoood (assuming conditional independence)

$$L(\lambda|\mathbf{y}) \propto \prod_{i=1}^{n} [\lambda \exp(-\lambda y_i)]^{c_i} [\exp(-\lambda y_i)]^{1-c_i}$$

$$\propto \lambda^{\sum_{i=1}^{n} c_i} \exp(-\lambda \sum_{i=1}^{n} y_i)$$

55

Put Gamma prior on λ

$$p(\lambda|\alpha_0, \lambda_0) \propto \lambda^{\alpha_0 - 1} \exp(-\lambda_0 \lambda)$$

Posterior of λ

$$p(\lambda|\mathbf{y}, \alpha_0, \lambda_0) \propto \lambda^{\sum_{i=1}^{n} c_i} \exp(-\lambda \sum_{i=1}^{n} y_i) \lambda^{\alpha_0 - 1} \exp(-\lambda_0 \lambda)$$

$$\propto \lambda^{\sum_{i=1}^{n} c_i + \alpha_0 - 1} \exp\left[-\lambda\left(\sum_{i=1}^{n} y_i + \lambda_0\right)\right]$$

The posterior distribution is Gamma, with parameters

$$\alpha_o^* = \alpha_0 + \sum_{i=1}^{n} c_i$$

$$\lambda_o^* = \lambda_0 + \sum_{i=1}^{n} y_i$$

56

Joint, Conditional and Marginal Posterior Distributions

•Put $\boldsymbol{\theta} = [\boldsymbol{\theta}_1', \boldsymbol{\theta}_2']'$ representing distinct features of models, (e.g., means and variances)

•Then, elicit a joint prior density

$$g(\boldsymbol{\theta}_1, \boldsymbol{\theta}_2) = g(\boldsymbol{\theta}_1|\boldsymbol{\theta}_2)g(\boldsymbol{\theta}_2) = g(\boldsymbol{\theta}_2|\boldsymbol{\theta}_1)g(\boldsymbol{\theta}_1)$$

where $g(\boldsymbol{\theta}_1)$ is the marginal prior and $g(\boldsymbol{\theta}_2|\boldsymbol{\theta}_1)$ is a conditional prior

•Joint posterior density is

$$p(\boldsymbol{\theta}_1, \boldsymbol{\theta}_2|\mathbf{y}) = \frac{L(\boldsymbol{\theta}_1, \boldsymbol{\theta}_2|\mathbf{y})g(\boldsymbol{\theta}_1, \boldsymbol{\theta}_2)}{\int\int L(\boldsymbol{\theta}_1, \boldsymbol{\theta}_2|\mathbf{y})g(\boldsymbol{\theta}_1, \boldsymbol{\theta}_2)d\boldsymbol{\theta}_1 d\boldsymbol{\theta}_2}$$

$$\propto L(\boldsymbol{\theta}_1, \boldsymbol{\theta}_2|\mathbf{y})g(\boldsymbol{\theta}_1, \boldsymbol{\theta}_2),$$

•Must decide which is the object of inference

•Joint, conditional or marginal posterior probability statements?

57

Marginal posterior densities

•Obtained directly from probability calculus as:

$$p(\boldsymbol{\theta}_1|\mathbf{y}) = \int p(\boldsymbol{\theta}_1, \boldsymbol{\theta}_2|\mathbf{y})d\boldsymbol{\theta}_2$$

$$p(\boldsymbol{\theta}_2|\mathbf{y}) = \int p(\boldsymbol{\theta}_1, \boldsymbol{\theta}_2|\mathbf{y})d\boldsymbol{\theta}_1$$

•Additional marginalizing may be needed if $\boldsymbol{\theta}_1 = [\boldsymbol{\theta}_{1A}', \boldsymbol{\theta}_{1B}']'$

$$p(\boldsymbol{\theta}_{1A}|\mathbf{y}) = \int\int p(\boldsymbol{\theta}_1, \boldsymbol{\theta}_2|\mathbf{y})d\boldsymbol{\theta}_{1B}d\boldsymbol{\theta}_2$$

$$= \int p(\boldsymbol{\theta}_1|\mathbf{y})d\boldsymbol{\theta}_{1B}.$$

58

Marginal posteriors are mixtures

•Let θ_2 be a "nuisance" parameter

$$\begin{aligned} p(\theta_1|\mathbf{y}) &= \int p(\theta_1,\theta_2|\mathbf{y})d\theta_2 \\ &= \int p(\theta_1|\theta_2,\mathbf{y})p(\theta_2|\mathbf{y})d\theta_2 \\ &= E_{\theta_2|\mathbf{y}}[p(\theta_1|\theta_2,\mathbf{y})], \end{aligned}$$

•Distribution $[\theta_1|\theta_2,\mathbf{y}]$ describes uncertainty when the nuisance parameter is known. — *Conditional posterior*

•Distribution $[\theta_2|\mathbf{y}]$: uncertainty about nuisance parameter — *Marginal posterior of nuisance*

•Distribution $[\theta_1|\mathbf{y}]$: uncertainty about primary parameter — *Marginal posterior of primary parameter*

59

Conditional posterior distributions

•By definition of conditional density:

$$p(\theta_1|\theta_2,\mathbf{y}) = \frac{p(\theta_1,\theta_2|\mathbf{y})}{p(\theta_2|\mathbf{y})}$$

•Here, one is interested in variation about θ_1 only

$$\begin{aligned} p(\theta_1|\theta_2,\mathbf{y}) &\propto p(\theta_1,\theta_2|\mathbf{y}) \\ &\propto L(\theta_1,\theta_2|\mathbf{y})p(\theta_1,\theta_2) \\ &\propto L(\theta_1,\theta_2|\mathbf{y})p(\theta_1|\theta_2) \\ &\propto L(\theta_1|\theta_2,\mathbf{y})p(\theta_1|\theta_2). \end{aligned}$$

•Identifying conditional posterior distributions: important for implementing MCMC methods (sampling from posteriors)

60

BAYESIAN LINEAR REGRESSION MODEL (normal distribution of residuals)

•MAKE DISTINCTION BETWEEN 2 SETS OF LOCATION PARAMETERS

Dummy variates

Treatment effects

regressions

$$\mathbf{y} = \mathbf{X}_1\boldsymbol{\beta}_1 + \mathbf{X}_2\boldsymbol{\beta}_2 + \mathbf{e}$$

regressors

$$= \begin{bmatrix} \mathbf{X}_1 & \mathbf{X}_2 \end{bmatrix} \begin{bmatrix} \boldsymbol{\beta}_1 \\ \boldsymbol{\beta}_2 \end{bmatrix} + \mathbf{e} = \mathbf{X}\boldsymbol{\beta} + \mathbf{e},$$

Maximum likelihood (also least-squares) estimator:

Vector of right-hand sides

$$\begin{bmatrix} \hat{\boldsymbol{\beta}}_1 \\ \hat{\boldsymbol{\beta}}_2 \end{bmatrix} = \mathbf{C}^{-1} \begin{bmatrix} \mathbf{X}_1'\mathbf{y} \\ \mathbf{X}_2'\mathbf{y} \end{bmatrix}$$

$$\mathbf{C} = \begin{bmatrix} \mathbf{X}_1'\mathbf{X}_1 & \mathbf{X}_1'\mathbf{X}_2 \\ \mathbf{X}_2'\mathbf{X}_1 & \mathbf{X}_2'\mathbf{X}_2 \end{bmatrix}$$

Coefficient matrix

LIKELIHOOD FUNCTION

•Under standard conditional independence

$$L(\boldsymbol{\beta}_1, \boldsymbol{\beta}_2, \sigma^2 | \mathbf{y}) \propto (\sigma^2)^{-\frac{n}{2}} \exp\left[-\frac{(\mathbf{y}-\mathbf{X}\boldsymbol{\beta})'(\mathbf{y}-\mathbf{X}\boldsymbol{\beta})}{2\sigma^2} \right]$$

•Decompose

$$(\mathbf{y} - \mathbf{X}_1\boldsymbol{\beta}_1 - \mathbf{X}_2\boldsymbol{\beta}_2)'(\mathbf{y} - \mathbf{X}_1\boldsymbol{\beta}_1 - \mathbf{X}_2\boldsymbol{\beta}_2) = S_e + S_\beta$$

$$S_e = \left(\mathbf{y} - \mathbf{X}_1\hat{\boldsymbol{\beta}}_1 - \mathbf{X}_2\hat{\boldsymbol{\beta}}_2\right)'\left(\mathbf{y} - \mathbf{X}_1\hat{\boldsymbol{\beta}}_1 - \mathbf{X}_2\hat{\boldsymbol{\beta}}_2\right)$$

Does not involve β

Involves β

$$S_\beta = \left[\left(\boldsymbol{\beta}_1 - \hat{\boldsymbol{\beta}}_1\right)' \quad \left(\boldsymbol{\beta}_2 - \hat{\boldsymbol{\beta}}_2\right)' \right] \mathbf{C} \begin{bmatrix} \boldsymbol{\beta}_1 - \hat{\boldsymbol{\beta}}_1 \\ \boldsymbol{\beta}_2 - \hat{\boldsymbol{\beta}}_2 \end{bmatrix}$$

Inference using improper priors

$$p(\boldsymbol{\beta}_1, \boldsymbol{\beta}_2, \sigma^2 | \mathbf{y}) \propto (\sigma^2)^{-\frac{n}{2}} \exp\left[-\frac{S_e + S_\beta}{2\sigma^2} \right]$$

Joint posterior is proportional to likelihood

a) Conditional posterior of coefficients, given variance

$$p(\boldsymbol{\beta}_1, \boldsymbol{\beta}_2 | \sigma^2, \mathbf{y}) \propto \exp\left[-\frac{S_\beta}{2\sigma^2} \right]$$

$$p(\boldsymbol{\beta}_1, \boldsymbol{\beta}_2 | \sigma^2, \mathbf{y}) \propto \exp\left[-\frac{\left[\left(\boldsymbol{\beta}_1 - \hat{\boldsymbol{\beta}}_1\right)' \quad \left(\boldsymbol{\beta}_2 - \hat{\boldsymbol{\beta}}_2\right)' \right] \mathbf{c} \begin{bmatrix} \boldsymbol{\beta}_1 - \hat{\boldsymbol{\beta}}_1 \\ \boldsymbol{\beta}_2 - \hat{\boldsymbol{\beta}}_2 \end{bmatrix}}{2\sigma^2} \right]$$

$$\begin{bmatrix} \boldsymbol{\beta}_1 \\ \boldsymbol{\beta}_2 \end{bmatrix} \Big| \sigma^2, \mathbf{y} \propto N\left(\begin{bmatrix} \hat{\boldsymbol{\beta}}_1 \\ \hat{\boldsymbol{\beta}}_2 \end{bmatrix}, \begin{bmatrix} \mathbf{X}_1'\mathbf{X}_1 & \mathbf{X}_1'\mathbf{X}_2 \\ \mathbf{X}_2'\mathbf{X}_1 & \mathbf{X}_2'\mathbf{X}_2 \end{bmatrix}^{-1} \sigma^2 \right)$$

b) Conditional posterior distribution of coefficients, given variance and other coefficients

$$\boldsymbol{\beta}_1 | \boldsymbol{\beta}_2, \sigma^2, \mathbf{y} \propto N\left(\tilde{\boldsymbol{\beta}}_1, (\mathbf{X}_1'\mathbf{X}_1)^{-1}\sigma^2\right)$$

$$\boldsymbol{\beta}_2 | \boldsymbol{\beta}_1, \sigma^2, \mathbf{y} \propto N\left(\tilde{\boldsymbol{\beta}}_2, (\mathbf{X}_2'\mathbf{X}_2)^{-1}\sigma^2\right)$$

"offset"

$$\tilde{\boldsymbol{\beta}}_i = (\mathbf{X}_i'\mathbf{X}_i)^{-1}\mathbf{X}_i'\left(\mathbf{y} - \mathbf{X}_j\boldsymbol{\beta}_j\right), \quad i = 1, 2, \; i \neq j.$$

c) Conditional posterior of individual coefficient, given the variance and all other coefficients

Normal, with mean and variance:

$$\tilde{\beta}_k = \frac{\mathbf{x}_k'(\mathbf{y} - \mathbf{X}_{-k}\boldsymbol{\beta}_{-k})}{\mathbf{x}_k'\mathbf{x}_k}$$

without parameter k

Column k of $\mathbf{X}$

$$Var(\beta_k | \boldsymbol{\beta}_{-k}, \sigma^2, \mathbf{y}) = \frac{\sigma^2}{\mathbf{x}_k'\mathbf{x}_k}$$

d) Conditional posterior of variance, given all coefficients

$$p(\sigma^2 | \boldsymbol{\beta}_1, \boldsymbol{\beta}_2, \mathbf{y}) \propto (\sigma^2)^{-\frac{n}{2}} \exp\left[-\frac{S_e + S_\beta}{2\sigma^2}\right]$$

$$p(\sigma^2 | \boldsymbol{\beta}_1, \boldsymbol{\beta}_2, \mathbf{y}) \propto (\sigma^2)^{-\left(\frac{n-2}{2}+1\right)} \exp\left[-\frac{S_e + S_\beta}{2\sigma^2}\right].$$

$$\sigma^2 | \boldsymbol{\beta}_1, \boldsymbol{\beta}_2, \mathbf{y} \sim (n-2)\left(\frac{S_e + S_\beta}{n-2}\right)\chi^{-2}_{n-2}$$

Curious loss of 2 d.f. (due to prior)

$$\Rightarrow \quad \sigma^2|\boldsymbol{\beta}_1,\boldsymbol{\beta}_2,\mathbf{y} \sim (n-2)\left(\tfrac{S_e+S_\beta}{n-2}\right)\chi^{-2}_{n-2}$$

$$E(\chi^{-2}_\nu) = \tfrac{1}{\nu-2};\; Var(\chi^{-2}_\nu) = \tfrac{2\nu^2}{(\nu-2)^2(\nu-4)}$$

$$E(\sigma^2|\boldsymbol{\beta}_1,\boldsymbol{\beta}_2,\mathbf{y}) = (n-2)\left(\frac{S_e+S_\beta}{n-2}\right)E(\chi^{-2}_{n-2})$$

$$\Rightarrow \quad = \frac{S_e+S_\beta}{n-4}$$

$$Var(\sigma^2|\boldsymbol{\beta}_1,\boldsymbol{\beta}_2,\mathbf{y}) = \left[(n-2)\left(\frac{S_e+S_\beta}{n-2}\right)\right]^2 \frac{2(n-2)^2}{(n-4)^2(n-6)}$$

$$\Rightarrow \quad = \frac{2(S_e+S_\beta)^2\,(n\text{-}2)^2}{(n-4)^2(n-6)}$$

MULTIVARIATE-t DISTRIBUTION

Let:

$$\mathbf{y}|\boldsymbol{\mu},\boldsymbol{\Sigma},w \sim N(\mathbf{y}|\boldsymbol{\mu},\tfrac{\boldsymbol{\Sigma}}{w})$$

and

$$w \sim Ga(\tfrac{\nu}{2},\tfrac{\nu}{2});\nu > 0$$

Joint density:

$$p(\mathbf{y},w|\boldsymbol{\mu},\boldsymbol{\Sigma},\nu) = p(\mathbf{y}|\boldsymbol{\mu},\boldsymbol{\Sigma},w)p(w|\nu)$$

$$= \left|2\pi\left(\frac{\boldsymbol{\Sigma}}{w}\right)\right|^{-\frac{1}{2}} \exp\left[-\frac{1}{2}(\mathbf{y}-\boldsymbol{\mu})'\left(\frac{\boldsymbol{\Sigma}}{w}\right)^{-1}(\mathbf{y}-\boldsymbol{\mu})\right]$$

$$\times \frac{(\nu/2)^{\frac{\nu}{2}}}{\Gamma(\nu/2)} w^{\frac{\nu}{2}-1} \exp\left[-\frac{\nu w}{2}\right].$$

Marginal density of **y:**

$$p(\mathbf{y}|\boldsymbol{\mu},\boldsymbol{\Sigma},\nu) = |2\pi\boldsymbol{\Sigma}|^{-\frac{1}{2}} \frac{(\nu/2)^{\frac{\nu}{2}}}{\Gamma(\nu/2)}$$

$$\times \int_0^\infty w^{\frac{n+\nu}{2}-1} \exp\left[-w\frac{(\mathbf{y}-\boldsymbol{\mu})'\boldsymbol{\Sigma}^{-1}(\mathbf{y}-\boldsymbol{\mu})+\nu}{2} \right] dw.$$

Integrand is kernel of $Ga\left(w|\frac{n+\nu}{2}, \frac{(\mathbf{y}-\boldsymbol{\mu})'\boldsymbol{\Sigma}^{-1}(\mathbf{y}-\boldsymbol{\mu})+\nu}{2}\right)$

$$\int_0^\infty w^{\frac{n+\nu}{2}-1} \exp\left[-w\frac{(\mathbf{y}-\boldsymbol{\mu})'\boldsymbol{\Sigma}^{-1}(\mathbf{y}-\boldsymbol{\mu})+\nu}{2} \right] dw$$

$$= \frac{\Gamma(\frac{n+\nu}{2})}{\left[\frac{(\mathbf{y}-\boldsymbol{\mu})'\boldsymbol{\Sigma}^{-1}(\mathbf{y}-\boldsymbol{\mu})+\nu}{2}\right]^{\frac{n+\nu}{2}}}.$$

Multivariate-t density:

Degrees of freedom

dimension

$$p(\mathbf{y}|\boldsymbol{\mu},\boldsymbol{\Sigma},\nu) = \frac{(\nu)^{\frac{\nu}{2}}\Gamma(\frac{n+\nu}{2})}{\Gamma(\frac{\nu}{2})|\pi\boldsymbol{\Sigma}|^{\frac{1}{2}}}\left[(\mathbf{y}-\boldsymbol{\mu})'\boldsymbol{\Sigma}^{-1}(\mathbf{y}-\boldsymbol{\mu})+\nu\right]^{-\frac{n+\nu}{2}}$$

$$= \boxed{\frac{\Gamma(\frac{n+\nu}{2})}{\Gamma(\frac{\nu}{2})|\nu\pi\boldsymbol{\Sigma}|^{\frac{1}{2}}}\left[1+\frac{(\mathbf{y}-\boldsymbol{\mu})'\boldsymbol{\Sigma}^{-1}(\mathbf{y}-\boldsymbol{\mu})}{\nu}\right]^{-\frac{n+\nu}{2}}}.$$

$$E(\mathbf{y}|\boldsymbol{\mu},\boldsymbol{\Sigma},\nu) = \boldsymbol{\mu}$$

"Scale matrix"

$$Var(\mathbf{y}|\boldsymbol{\mu},\boldsymbol{\Sigma},\nu) = \frac{\nu}{\nu-2}\boldsymbol{\Sigma}$$

All marginal and conditional distributions are multivariate or univariate t

Starting from

$$\begin{bmatrix} \mathbf{y}_1 \\ \mathbf{y}_2 \end{bmatrix} \Big| \boldsymbol{\mu}, \boldsymbol{\Sigma}, w \sim N\left(\begin{bmatrix} \boldsymbol{\mu}_1 \\ \boldsymbol{\mu}_2 \end{bmatrix}, \begin{bmatrix} \boldsymbol{\Sigma}_{11} & \boldsymbol{\Sigma}_{12} \\ \boldsymbol{\Sigma}_{21} & \boldsymbol{\Sigma}_{22} \end{bmatrix} \frac{1}{w} \right)$$

all marginal and conditional distributions are normal. Integration over

$$Ga\left(w | \frac{n+v}{2}, \frac{(\mathbf{y}-\boldsymbol{\mu})'\boldsymbol{\Sigma}^{-1}(\mathbf{y}-\boldsymbol{\mu})+v}{2} \right)$$

yields t-distributions. For example, the n_1 dimensional distribution $\mathbf{y}_1|\mathbf{y}_2, \boldsymbol{\mu}, \boldsymbol{\Sigma}, v$ has mean vector and covariance matrix

$$E(\mathbf{y}_1|\mathbf{y}_2, \boldsymbol{\mu}, \boldsymbol{\Sigma}, v) = \boldsymbol{\mu}_1 + \boldsymbol{\Sigma}_{12}(\boldsymbol{\Sigma}_{22})^{-1}(\mathbf{y}_2 - \boldsymbol{\mu}_2)$$

$$Var(\mathbf{y}_1|\mathbf{y}_2, \boldsymbol{\mu}, \boldsymbol{\Sigma}, v) = \frac{v}{v-2}\left[\boldsymbol{\Sigma}_{11} - \boldsymbol{\Sigma}_{12}(\boldsymbol{\Sigma}_{22})^{-1}\boldsymbol{\Sigma}_{21}\right]$$

Side note on the *t-distribution*

$$t = \frac{z}{\sqrt{\chi^2_v/v}}; \quad z \sim N(0,1)$$

$$E(t) = 0$$

$$Var(t) = \frac{v}{v-2}$$

$$f(t) = \frac{\Gamma\left(\frac{v+1}{2}\right)}{\sqrt{v\pi}\,\Gamma\left(\frac{v}{2}\right)}\left(1 + \frac{t^2}{v}\right)^{-\left(\frac{v+1}{2}\right)}$$

Model with t-distributed errors

$$y = \mu + St$$

$$E(y) = \mu$$

$$Var(y) = S^2\frac{v}{v-2}$$

$$t = \frac{y-\mu}{S}$$

$$\frac{dt}{dy} = \frac{1}{S}$$

$$f(y) = \frac{\Gamma\left(\frac{v+1}{2}\right)}{\sqrt{v\pi S^2}\,\Gamma\left(\frac{v}{2}\right)}\left[1 + \frac{(y-\mu)^2}{vS^2}\right]^{-\left(\frac{v+1}{2}\right)}$$

"degrees of freedom"

"scale"

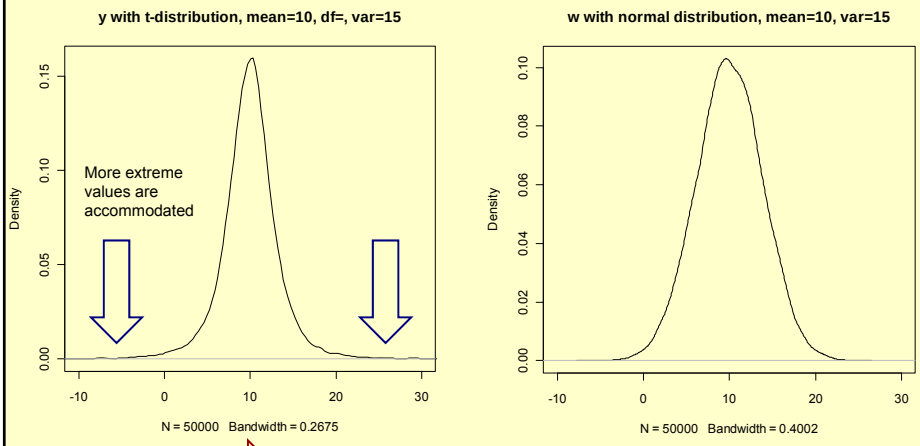
72

Simulate a t-distribution with mean 10, scale 5 and 3 d.f. The variance is 5 x 3/(3-2)= 15. We will compare with a normal distribution with Mean 10 and variance 15

```
> m<-10
> df<-3
> S<-sqrt(5)
> z<-rnorm(50000,0,1)
> chisq<-rchisq(50000,3)
> y<-10+S*z/sqrt(chisq/3)
> mean(z)
[1] 0.0008953519
> var(z)
[1] 1.009415
> summary(chisq)
> mean(chisq)
[1] 2.978861
> var(chisq)
[1] 5.896806
> mean(y)
[1] 9.999967
> var(y)
[1] 13.83878
```

73

The t-distribution accommodates more extreme values



Student-t

```
> summary(y)
   Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
-61.440   8.266  10.020  10.000  11.730 100.400
```

Normal

```
> summary(w)
   Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
 -6.509   7.436  10.010  10.020  12.620  25.300
```

74

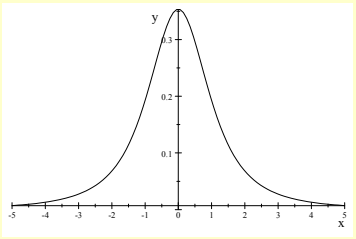
Curious t-distributions….

t-DISTRIBUTION WITH MEAN O AND SCALE 1

$$f(t) = \frac{\Gamma\left(\frac{\nu+1}{2}\right)}{\sqrt{\nu\pi}\,\Gamma\left(\frac{\nu}{2}\right)}\left(1+\frac{t^2}{\nu}\right)^{-(\nu+1)/2},$$

t-DISTRIBUTION WITH MEAN O AND SCALE 1 AND 2 df (infinite variance)

$$f(t) = \frac{\Gamma\left(\frac{3}{2}\right)}{\sqrt{2\pi}\,\Gamma(1)}\left(1+\frac{t^2}{2}\right)^{-\frac{3}{2}} = \frac{\Gamma\left(\frac{3}{2}\right)}{\sqrt{2\pi}\left(1+\frac{t^2}{2}\right)^{\frac{3}{2}}}$$



Proper distribution

$$0.886226925453\int_{-\infty}^{\infty}\frac{1}{\sqrt{2\pi}\left(1+\frac{t^2}{2}\right)^{\frac{3}{2}}}dt = 1.0$$

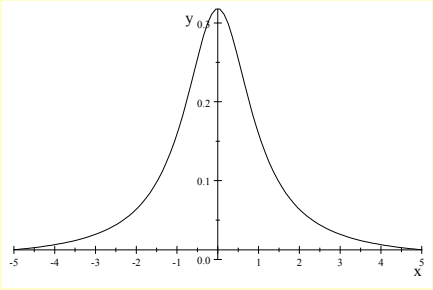
75

Cauchy distribution= t with 1 df

Does not have a mean or variance…

$$\frac{\Gamma(1)}{\sqrt{\pi}\,\Gamma\left(\frac{1}{2}\right)}\left(1+\frac{t^2}{2}\right)^{-1} = \frac{1}{\pi}(1+t^2)^{-1}$$

$$\int_{-\infty}^{\infty}\frac{1}{\pi}(1+t^2)^{-1}dt = 1$$




```
> cauchymillion<-rcauchy(1000000)
> range(cauchymillion)
[1] -1073944  2219368
> mean(cauchymillion)
[1] 2.614773
> median(cauchymillion)
[1] 7.042597e-05
> var(cauchymillion)
[1] 7410156
```

No meaning whatsoever

76

Marginal distribution of regression coefficients

Multivariate t

$$p(\boldsymbol{\beta}_1, \boldsymbol{\beta}_2|\mathbf{y}) \propto \int (\sigma^2)^{-\left(\frac{n-2}{2}+1\right)} \exp\left[-\frac{S_e + S_\beta}{2\sigma^2}\right] d\sigma^2$$

$$\propto (S_e + S_\beta)^{-\left(\frac{n-2}{2}\right)} \propto \left[1 + \frac{S_\beta}{(n-2-p_1-p_2)\frac{S_e}{(n-2-p_1-p_2)}}\right]^{-(n-2-p_1-p_2+p_1+p_2)/2},$$

$$S_\beta = \left[\left(\boldsymbol{\beta}_1 - \hat{\boldsymbol{\beta}}_1\right)' \left(\boldsymbol{\beta}_2 - \hat{\boldsymbol{\beta}}_2\right)'\right] \mathbf{C} \begin{bmatrix} \boldsymbol{\beta}_1 - \hat{\boldsymbol{\beta}}_1 \\ \boldsymbol{\beta}_2 - \hat{\boldsymbol{\beta}}_2 \end{bmatrix}$$

dimension

Degrees of freedom

$$n > p_1 + p_2 + 2$$

Mean vector

$$\hat{\boldsymbol{\beta}} = [\hat{\boldsymbol{\beta}}_1', \hat{\boldsymbol{\beta}}_2']'$$

Covariance matrix

$$Var(\boldsymbol{\beta}_1, \boldsymbol{\beta}_2|\mathbf{y}) = \frac{S_e}{(n-p_1-p_2-4)} \begin{bmatrix} \mathbf{X}_1'\mathbf{X}_1 & \mathbf{X}_1'\mathbf{X}_2 \\ \mathbf{X}_2'\mathbf{X}_1 & \mathbf{X}_2'\mathbf{X}_2 \end{bmatrix}^{-1}$$

Marginal distribution of variance

$$p(\sigma^2|\mathbf{y}) \propto (\sigma^2)^{-\frac{n}{2}} \exp\left[-\frac{S_e}{2\sigma^2}\right] \int\int \exp\left[-\frac{S_\beta}{2\sigma^2}\right] d\boldsymbol{\beta}_1 d\boldsymbol{\beta}_2$$

$$\int\int \exp\left[-\frac{\left[\left(\boldsymbol{\beta}_1 - \hat{\boldsymbol{\beta}}_1\right)' \left(\boldsymbol{\beta}_2 - \hat{\boldsymbol{\beta}}_2\right)'\right] \mathbf{C} \begin{bmatrix} \boldsymbol{\beta}_1 - \hat{\boldsymbol{\beta}}_1 \\ \boldsymbol{\beta}_2 - \hat{\boldsymbol{\beta}}_2 \end{bmatrix}}{2\sigma^2}\right] d\boldsymbol{\beta}_1 d\boldsymbol{\beta}_2$$

$$= (2\pi)^{\frac{p_1+p_2}{2}} \left|\mathbf{C}^{-1}\sigma^2\right|^{\frac{1}{2}}.$$

$$p(\sigma^2|\mathbf{y}) \propto (\sigma^2)^{-\left(\frac{n-p_1-p_2-2}{2}+1\right)} \exp\left(-\frac{S_e}{2\sigma^2}\right)$$

$$\sigma^2|\mathbf{y} \sim (n - p_1 - p_2 - 2)\frac{S_e}{(n - p_1 - p_2 - 2)}\chi^{-2}_{n-p_1-p_2-2}$$

$$E(\sigma^2|\mathbf{y}) = \frac{S_e}{n - p_1 - p_2 - 4},$$

$$Var(\sigma^2|\mathbf{y}) = \frac{2S_e^2}{(n-p_1-p_2-4)^2(n-p_1-p_2-6)}$$

Posterior distribution of residuals

$$e_i = y_i - \mathbf{x}_i'\boldsymbol{\beta}$$

univariate-t on $n - 2 - p_1 - p_2$ degrees of freedom

$$E(e_i|\mathbf{y}) = y_i - \mathbf{x}_i'\hat{\boldsymbol{\beta}}$$

$$Var(e_i|\mathbf{y}) = Var(\mathbf{x}_i'\boldsymbol{\beta}|\mathbf{y}) = \frac{S_e\mathbf{x}_i'\mathbf{C}^{-1}\mathbf{x}_i}{(n-p_1-p_2-4)}$$

Predictive Distributions

•Let $\mathbf{y}_f$ = unobserved vector of "future" or "missing" data. Then

•1) $p(\boldsymbol{\theta},\mathbf{y},\mathbf{y}_f) = p(\mathbf{y}_f|\boldsymbol{\theta},\mathbf{y})p(\boldsymbol{\theta},\mathbf{y})$
$= p(\mathbf{y}_f|\boldsymbol{\theta},\mathbf{y})p(\boldsymbol{\theta}|\mathbf{y})p(\mathbf{y})$

•2) $p(\boldsymbol{\theta},\mathbf{y}_f|\mathbf{y}) = p(\mathbf{y}_f|\boldsymbol{\theta},\mathbf{y})p(\boldsymbol{\theta}|\mathbf{y})$

•3) $p(\mathbf{y}_f|\mathbf{y}) = \int p(\mathbf{y}_f|\boldsymbol{\theta},\mathbf{y})p(\boldsymbol{\theta}|\mathbf{y})d\boldsymbol{\theta}$
$= E\ [p(\mathbf{y}_f|\boldsymbol{\theta},\mathbf{y})]$

Use for posterior predictive checks

Truncated
Censored
Missing data
Future
Data augmentation!

$p(\boldsymbol{\theta},\mathbf{y}) = \int p(\boldsymbol{\theta},\mathbf{y},\mathbf{y}_f)d\mathbf{y}_f$

•4) If, given the parameters, data are conditionally independent:

$$p(\mathbf{y}_f|\mathbf{y}) = \int p(\mathbf{y}_f|\boldsymbol{\theta})p(\boldsymbol{\theta}|\mathbf{y})d\boldsymbol{\theta}$$

Usual representation of posterior predictive density

81

A model for binary data

$y_i \sim Bernoulli(\theta)$ ← Probability of success

Assuming conditional independence → $p(y_1,y_2,\ldots,y_N|\theta) \propto \theta^x(1-\theta)^{N-x}$

Beta prior → $p(\theta) \propto \theta^{a-1}(1-\theta)^{b-1}$

$$p(\theta|\mathbf{y}) \propto \theta^x(1-\theta)^{N-x}\theta^{a-1}(1-\theta)^{b-1}$$
$$\propto \theta^{x+a-1}(1-\theta)^{N-x+b-1}$$

⇒ $\theta|\mathbf{y} \sim Beta(x+a, N-x+b)$

	Prior	Posterior
Distribution	Beta	Beta
Mean	$\frac{a}{a+b}$	$\frac{x+a}{N+a+b}$
Variance	$\frac{ab}{(a+b)^2(a+b+1)}$	$\frac{(x+a)(N-x+b)}{(N+a+b)^2(N+a+b+1)}$

82

Binary Data: Predictive distribution

Future data

Future number of Bernoulli trials

$$p(\mathbf{y}_f|\mathbf{y}) = \int \binom{N_f}{x_f} \theta^{x_f}(1-\theta)^{N_f-x_f} \frac{\theta^{x+a-1}(1-\theta)^{N-x+b-1}}{B(x+a, N-x+b)} d\theta$$

Future number of "successes"

Posterior

$$= \frac{\binom{N_f}{x_f}}{B(x+a, N-x+b)} \int \theta^{x_f+x+a-1}(1-\theta)^{N_f+N-x_f-x+b-1} d\theta$$

$$= \binom{N_f}{x_f} \frac{B(x_f+x+a, N_f+N-x_f-x+b)}{B(x+a, N-x+b)}$$

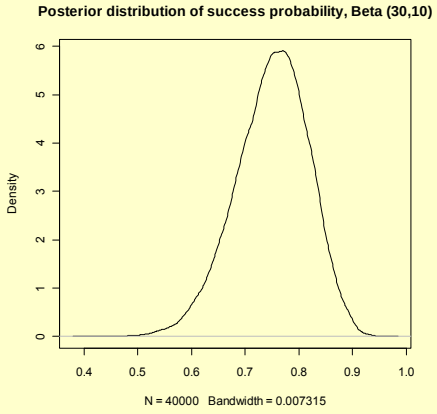
Beta-binomial distribution (discrete)

83

Simulating the predictive distribution

Suppose the posterior distribution of the success probability is Beta(30,10). We draw 40,000 samples and plot the posterior

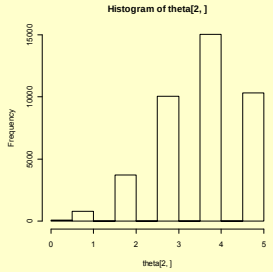
```
> #Bernoulli probability is prob~Beta(30,10)
> prob<-rbeta(40000,30,10)
> mean(prob)
[1] 0.7499961
> var(prob)
[1] 0.004579322
```

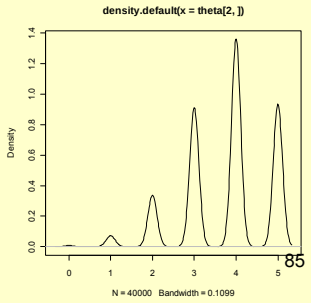


84

We construct the predictive distribution via composition sampling.
Simulate 40,000 binomial trials with n=5
#Simulation by composition

```
m<-2
r<-40000
theta<-matrix(0,m,r)

theta[1,]<-rbeta(40000,30,10)
for (i in 1:r) {theta[2,i]<-rbinom(1,5,theta[1,i])}

mean(theta[1,])
mean(theta[2,])
histcases<-hist(theta[2,])
plot<-density(theta[2,])
```

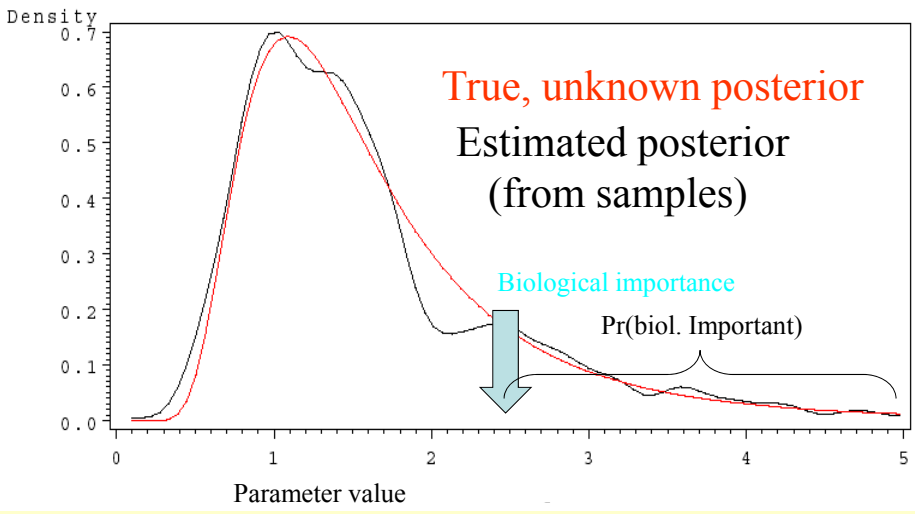





```
> mean(theta[1,])
[1] 0.7501773
> mean(theta[2,])
[1] 3.751525
```

85

Exact and estimated posterior densities
(most of the time we will not be able to derive the posterior, but may be able to sample from it)



Estimating a posterior expectation and variance from samples

Posterior Expectation: $E(\boldsymbol{\theta}|\mathbf{y}) = \int \boldsymbol{\theta} p(\boldsymbol{\theta}|\mathbf{y}) d\boldsymbol{\theta}$

- May be posterior is unknown or integral impossible to compute
- Samples available from $[\boldsymbol{\theta}|\mathbf{y}]$

  $\boldsymbol{\theta}^{(1)}, \boldsymbol{\theta}^{(2)}, \ldots, \boldsymbol{\theta}^{(S)}$
- Estimate integral as $\hat{E}(\boldsymbol{\theta}|\mathbf{y}) = \frac{1}{S}\sum_{i=1}^{S} \boldsymbol{\theta}^{(i)}$
- Monte Carlo Error $= \hat{E}(\boldsymbol{\theta}|\mathbf{y}) - E(\boldsymbol{\theta}|\mathbf{y})$

  $= \frac{1}{S}\sum_{i=1}^{S} \boldsymbol{\theta}^{(i)} - E(\boldsymbol{\theta}|\mathbf{y})$

  Goes to *0* as *S* tends to infinity

- Monte Carlo Variance of estimate of posterior mean

Measures variability to be expected if repeated sampling (each time *S* samples drawn) is done from the posterior

$$Var\left(\text{Monte Carlo Error}\right) = Var_{\boldsymbol{\theta}|\mathbf{y}}\left[\hat{E}(\boldsymbol{\theta}|\mathbf{y}) - E(\boldsymbol{\theta}|\mathbf{y})\right]$$

$$Var\left(\text{Monte Carlo Error}\right) = Var_{\boldsymbol{\theta}|\mathbf{y}}\left[\frac{1}{S}\sum_{i=1}^{S} \boldsymbol{\theta}^{(i)} - E(\boldsymbol{\theta}|\mathbf{y})\right]$$

$$= Var_{\boldsymbol{\theta}|\mathbf{y}}\left[\frac{1}{S}\sum_{i=1}^{S} \boldsymbol{\theta}^{(i)}\right]$$

$$Var(\text{MCE}) = \frac{1}{S^2}\left[\sum_{i=1}^{S} Var_{\theta|\mathbf{y}}(\boldsymbol{\theta}^{(i)}) + 2\sum\sum_{i<j} Cov_{\theta|\mathbf{y}}(\boldsymbol{\theta}^{(i)}, \boldsymbol{\theta}^{(i)})\right]$$

$$= \frac{1}{S^2}\left[\sum_{i=1}^{S} Var(\boldsymbol{\theta}|\mathbf{y}) + 2Var(\boldsymbol{\theta}|\mathbf{y})\sum\sum_{i<j} \rho_{ij}\right]$$

$$= \frac{Var(\boldsymbol{\theta}|\mathbf{y})}{S}\left(1 + \frac{2}{S}\sum\sum_{i<j} \rho_{ij}\right)$$

Null only if samples are independent

IF MARKOV CHAIN MONTE CARLO SAMPLING IS PRACTICED, SAMPLES ARE TYPICALLY SERIALLY CORRELATED

IMPORTANT TO EVALUATE AUTO-CORRELATIONS IN MCMC, TO ASSES MONTE CARLO ERROR

89

POSTERIOR PROBABILITIES

- Joint probabilities

$$\Pr(\boldsymbol{\theta} \in \Re|\mathbf{y}) = \int_{\Re} p(\boldsymbol{\theta}|\mathbf{y})d\boldsymbol{\theta}$$

- Marginal probabilities

$$\Pr(\boldsymbol{\theta}_1 \in \Re_1|\mathbf{y}) = \int_{\Re_1}\int_{\Theta_2} p(\boldsymbol{\theta}_2, \boldsymbol{\theta}_1|\mathbf{y})d\boldsymbol{\theta}$$

$$= \int_{\Re_1}\int_{\Theta_2} p(\boldsymbol{\theta}_1|\boldsymbol{\theta}_2, \mathbf{y})p(\boldsymbol{\theta}_2|\mathbf{y})d\boldsymbol{\theta}.$$

- Equivalently

$$\Pr(\boldsymbol{\theta}_1 \in \Re_1|\mathbf{y}) = E_{\boldsymbol{\theta}_2|\mathbf{y}}[\Pr(\boldsymbol{\theta}_1 \in \Re_1|\boldsymbol{\theta}_2, \mathbf{y})]$$

90

MONTE CARLO ESTIMATES OF POSTERIOR PROBABILITIES

- Suppose samples $\boldsymbol{\theta}_2^{(1)}, \boldsymbol{\theta}_2^{(1)}, \ldots, \boldsymbol{\theta}_2^{(m)}$ available from $[\boldsymbol{\theta}_2|\mathbf{y}]$

$$\frac{1}{m}\sum_{i=1}^{m} \Pr\left(\boldsymbol{\theta}_1 \in \Re_1 | \boldsymbol{\theta}_2^{(i)}, \mathbf{y}\right) \square$$

$$\Pr(\boldsymbol{\theta}_1 \in \Re_1 | \mathbf{y}) = \int_{\Theta_2} \left[\int_{\Re_1} p(\boldsymbol{\theta}_1 | \boldsymbol{\theta}_2, \mathbf{y}) d\boldsymbol{\theta}_1 \right] p(\boldsymbol{\theta}_2 | \mathbf{y}) d\boldsymbol{\theta}_2$$

- 1) Must be easy to sample from $[\boldsymbol{\theta}_2|\mathbf{y}]$
- 2) $\Pr(\boldsymbol{\theta}_1 \in \Re_1 | \boldsymbol{\theta}_2, \mathbf{y})$ must be available in closed form so it can be evaluated at each draw $\boldsymbol{\theta}_2^{(i)}$

91

Estimating the predictive density from posterior samples

$$p(\mathbf{y}_f|\mathbf{y}) = \int p(\mathbf{y}_f|\boldsymbol{\theta}) p(\boldsymbol{\theta}|\mathbf{y}) \mathbf{d}\boldsymbol{\theta}$$

Samples available from $[\boldsymbol{\theta}|\mathbf{y}]$

$\boldsymbol{\theta}^{(1)}, \boldsymbol{\theta}^{(2)}, \ldots, \boldsymbol{\theta}^{(S)}$

Density at point y_0

$$\hat{p}(y_f = y_0|\mathbf{y}) = \frac{1}{S}\sum_{i=1}^{S} p(y_f = y_0|\boldsymbol{\theta}^{(i)})$$

Called ergodic averaging: basis of the Monte Carlo method

92

METROPOLIS-HASTINGS ALGORITHM

…and derivatives

93

1. FORM OF ALGORITHM

1. Generate candidate $\boldsymbol{\theta}^*$ from proposal density $f(\boldsymbol{\theta}^*|\boldsymbol{\theta}^{[t-1]})$
2. Draw random number $U(0,1)$
3. Compute ratio

Posterior or conditional posterior

$$R = \frac{g(\boldsymbol{\theta}^*)/f(\boldsymbol{\theta}^*|\boldsymbol{\theta}^{[t-1]})}{g(\boldsymbol{\theta}^{[t-1]})/f(\boldsymbol{\theta}^{[t-1]}|\boldsymbol{\theta}^*)}$$

4. If $\begin{cases} U < \min(R,1) \text{ set } \boldsymbol{\theta}^{[t]} = \boldsymbol{\theta}^* \\ \boldsymbol{\theta}^{[t]} = \boldsymbol{\theta}^{[t-1]} \end{cases}$

Important: sample not rejected. Chain value is just repeated

Integration constant is not needed

$$R = \frac{cp(y|\boldsymbol{\theta}^*)p(\boldsymbol{\theta}^*)/f(\boldsymbol{\theta}^*|\boldsymbol{\theta}^{[t-1]})}{cp(y|\boldsymbol{\theta}^{[t-1]})p(\boldsymbol{\theta}^{[t-1]})/f(\boldsymbol{\theta}^{[t-1]}|\boldsymbol{\theta}^*)}$$

$$= \boxed{\frac{p(y|\boldsymbol{\theta}^*)p(\boldsymbol{\theta}^*)/f(\boldsymbol{\theta}^*|\boldsymbol{\theta}^{[t-1]})}{p(y|\boldsymbol{\theta}^{[t-1]})p(\boldsymbol{\theta}^{[t-1]})/f(\boldsymbol{\theta}^{[t-1]}|\boldsymbol{\theta}^*)}}$$

94

2. SPECIAL FORMS: USING THE POSTERIOR AS PROPOSAL

$$R = \frac{p(y|\boldsymbol{\theta}^*)p(\boldsymbol{\theta}^*)/f(\boldsymbol{\theta}^*|\boldsymbol{\theta}^{[t-1]})}{p(y|\boldsymbol{\theta}^{[t-1]})p(\boldsymbol{\theta}^{[t-1]})/f(\boldsymbol{\theta}^{[t-1]}|\boldsymbol{\theta}^*)}$$

$$= \frac{p(y|\boldsymbol{\theta}^*)p(\boldsymbol{\theta}^*)/[cp(y|\boldsymbol{\theta}^*)p(\boldsymbol{\theta}^*)]}{p(y|\boldsymbol{\theta}^{[t-1]})p(\boldsymbol{\theta}^{[t-1]})/[cp(y|\boldsymbol{\theta}^{[t-1]})p(\boldsymbol{\theta}^{[t-1]})]} = 1$$

If this were not so, one would have doubts…

95

3. SPECIAL FORMS: METROPOLIS ALGORITHM

Take a symmetric (in its arguments) proposal density:

$$f(\boldsymbol{\theta}^*|\boldsymbol{\theta}^{[t-1]}) = f(\boldsymbol{\theta}^{[t-1]}|\boldsymbol{\theta}^*)$$

Acceptance rate becomes

$$R = \frac{p(y|\boldsymbol{\theta}^*)p(\boldsymbol{\theta}^*)/f(\boldsymbol{\theta}^*|\boldsymbol{\theta}^{[t-1]})}{p(y|\boldsymbol{\theta}^{[t-1]})p(\boldsymbol{\theta}^{[t-1]})/f(\boldsymbol{\theta}^{[t-1]}|\boldsymbol{\theta}^*)}$$

$$= \boxed{\frac{p(y|\boldsymbol{\theta}^*)p(\boldsymbol{\theta}^*)}{p(y|\boldsymbol{\theta}^{[t-1]})p(\boldsymbol{\theta}^{[t-1]})}}$$

$\Rightarrow$ If $\begin{cases} U < \min(R, 1) \text{ set } \boldsymbol{\theta}^{[t]} = \boldsymbol{\theta}^* \\ \boldsymbol{\theta}^{[t]} = \boldsymbol{\theta}^{[t-1]} \end{cases}$

96

4. SPECIAL FORMS: INDEPENDENCE CHAIN METROPOLIS

Proposal density is independent of current state of the sequence of sampled values

$$f(\boldsymbol{\theta}^* | \boldsymbol{\theta}^{[t-1]}) = f(\boldsymbol{\theta}^*)$$

Acceptance ratio calculated as in standard Metropolis

$$R = \frac{p(y|\boldsymbol{\theta}^*)p(\boldsymbol{\theta}^*)/f(\boldsymbol{\theta}^*)}{p(y|\boldsymbol{\theta}^{[t-1]})p(\boldsymbol{\theta}^{[t-1]})/f(\boldsymbol{\theta}^*)} = \frac{p(y|\boldsymbol{\theta}^*)p(\boldsymbol{\theta}^*)}{p(y|\boldsymbol{\theta}^{[t-1]})p(\boldsymbol{\theta}^{[t-1]})}$$

97

6. SPECIAL FORMS: GIBBS SAMPLING (notation is not precise…)

$$f_i\left(\theta_i^* | \theta_i^{[t]}\right) = \pi\left(\theta_i^* | \theta_{-i}^{[t]}\right)$$
$$f_i\left(\theta_i^{[t]} | \theta_i^*\right) = \pi\left(\theta_{-i}^{[t]} | \theta_i^*\right)$$

⇒ Use fully conditionals as proposal distributions

By Bayes theorem ⇒

$$\pi\left(\theta_i^* | \theta_{-i}^{[t]}\right) = \frac{\pi\left(\theta_i^*, \theta_{-i}^{[t]}\right)}{\pi\left(\theta_{-i}^{[t]}\right)}$$
$$\pi\left(\theta_{-i}^{[t]} | \theta_i^*\right) = \frac{\pi\left(\theta_i^*, \theta_{-i}^{[t]}\right)}{\pi(\theta_i^*)}$$

MH acceptance ratio (recall that g(θ) is a posterior (or conditional posterior) that we do not recognize (or know how to sample from)

$$R = \frac{g(\boldsymbol{\theta}^*)/f(\boldsymbol{\theta}^* | \boldsymbol{\theta}^{[t-1]})}{g(\boldsymbol{\theta}^{[t-1]})/f(\boldsymbol{\theta}^{[t-1]} | \boldsymbol{\theta}^*)} == \frac{g(\boldsymbol{\theta}^*)}{g(\boldsymbol{\theta}^{[t-1]})} \times \frac{\pi\left(\theta_{-i}^{[t]} | \theta_i^*\right)}{\pi\left(\theta_i^* | \theta_{-i}^{[t]}\right)}$$

$$= \frac{g(\boldsymbol{\theta}^*)}{g(\boldsymbol{\theta}^{[t-1]})} \times \frac{\frac{\pi\left(\theta_i^*, \theta_{-i}^{[t]}\right)}{\pi(\theta_i^*)}}{\frac{\pi\left(\theta_i^*, \theta_{-i}^{[t]}\right)}{\pi\left(\theta_{-i}^{[t]}\right)}} = \frac{g(\boldsymbol{\theta}^*)}{g(\boldsymbol{\theta}^{[t-1]})} \times \frac{\pi\left(\theta_{-i}^{[t]}\right)}{\pi(\theta_i^*)} = 1$$

98

GIBBS SAMPLING: SPECIAL CASE OF MH: proposal always accepted

ILLUSTRATION OF DIRECT, COMPOSITION AND GIBBS SAMPLING

1) Assumptions and basic results

Suppose we wish to draw samples from

$$\begin{pmatrix}\theta_1\\ \theta_2\end{pmatrix} \sim N\left(\begin{bmatrix}0\\ 2\end{bmatrix}, \begin{bmatrix}1 & -\frac{3}{4}\times 1\times\frac{1}{2} = -0.375\\ -0.375 & \left(\frac{1}{2}\right)^2\end{bmatrix}\right)$$

$$\begin{aligned}E(\theta_1|\theta_2) &= E(\theta_1) + \frac{Cov(\theta_1,\theta_2)}{Var(\theta_2)}[\theta_2 - E(\theta_2)]\\ &= 1 + \frac{-0.375}{\left(\frac{1}{2}\right)^2}(\theta_2 - 2)\\ &= 4.0 - 1.5\theta_2\\ Var(\theta_1|\theta_2) &= Var(\theta_1)(1-\rho^2)\\ &= 1\times\left[1 - \left(-\frac{3}{4}\right)^2\right]\\ &= \frac{7}{16} = 0.4375\end{aligned}$$

$$\begin{aligned}E(\theta_2|\theta_1) &= E(\theta_2) + \frac{Cov(\theta_1,\theta_2)}{Var(\theta_1)}[\theta_1 - E(\theta_1)]\\ &= 2 + \frac{-0.375}{1}(\theta_1 - 0)\\ &= 2 - 0.375\theta_1\\ Var(\theta_2|\theta_1) &= Var(\theta_2)(1-\rho^2)\\ &= \left(\frac{1}{2}\right)^2\times\left[1 - \left(-\frac{3}{4}\right)^2\right]\\ &= \frac{7}{64} = 0.109375\end{aligned}$$

99

2) Direct sampling from the bivariate normal distribution

$$\begin{pmatrix}\theta_1\\ \theta_2\end{pmatrix} \sim N\left(\begin{bmatrix}0\\ 2\end{bmatrix}, \begin{bmatrix}1 & -\frac{3}{4}\times 1\times\frac{1}{2} = -0.375\\ -0.375 & \left(\frac{1}{2}\right)^2\end{bmatrix}\right)$$

$$\begin{aligned}\theta &= \begin{bmatrix}0\\ 2\end{bmatrix} + Cholesky\left(\begin{bmatrix}1 & -\frac{3}{4}\times 1\times\frac{1}{2} = -0.375\\ -0.375 & \left(\frac{1}{2}\right)^2\end{bmatrix}\right)\begin{bmatrix}z_1\\ z_2\end{bmatrix}\\ &= \begin{bmatrix}0\\ 2\end{bmatrix} + \begin{bmatrix}1.0 & 0\\ -0.375 & 0.330718913883\end{bmatrix}\begin{bmatrix}z_1\\ z_2\end{bmatrix}\\ &= \begin{bmatrix}z_1\\ 2 - 0.375z_1 + 0.330718913883z_2\end{bmatrix}\end{aligned}$$

100

```
> # Simulation of a bivariate normal- Direct
> mu<-matrix(c(0,2),2,1)
> mu
     [,1]
[1,]    0
[2,]    2
>
> V<-matrix(c(1.0,-0.375,-0.375,0.25),2,2)
> V
      [,1]   [,2]
[1,]  1.000 -0.375
[2,] -0.375  0.250
> C<-t(chol(V))
> C
      [,1]      [,2]
[1,]  1.000 0.0000000
[2,] -0.375 0.3307189
```

```
> m<-2
> r<-20000
>
> thetavec<-matrix(0,m,r)
> for (i in 1:r) {z<-matrix(rnorm(m),m,1)
+                 thetavec[,i]<-mu+C%*%z}
> mean(thetavec[1,])
[1] 0.003480938
> mean(thetavec[2,])
[1] 1.997486
> v1<-var(thetavec[1,])
> v2<-var(thetavec[2,])
> v12<-cov(thetavec[1,],thetavec[2,])
> Vsim<-matrix(c(v1,v12,v12,v2),2,2)
> v1
[1] 0.9859114
> v2
[1] 0.2473953
> Vsim
          [,1]       [,2]
[1,]  0.9859114 -0.3697481
[2,] -0.3697481  0.2473953
```

$$\begin{pmatrix} \theta_1 \\ \theta_2 \end{pmatrix} \sim N\left(\begin{bmatrix} 0 \\ 2 \end{bmatrix}, \begin{bmatrix} 1 & -\frac{3}{4} \times 1 \times \frac{1}{2} = -0.375 \\ -0.375 & \left(\frac{1}{2}\right)^2 \end{bmatrix} \right)$$

101

COMPOSITION OR CHAIN SAMPLING FROM A JOINT DISTRIBUTION

$$\begin{aligned} &\Pr(X_1 = x_1, X_2 = x_2, \ldots, X_N = x_N) \\ &= \Pr(X_1 = x_1) \times \Pr(X_2 = x_2 | X_1 = x_1) \times \Pr(X_3 = x_3 | X_1 = x_1, X_2 = x_2) \\ &\quad \ldots \times \Pr(X_N = x_N | X_1 = x_1, X_2 = x_2, \ldots, X_{N-1} = x_{N-1}) \end{aligned}$$

Then:

$$\mathbf{x}' = [x_1, x_2 \ldots, x_{N-1}, x_N]$$

Is a realization from joint distribution above

102

3) Sampling from the bivariate normal distribution using composition

```
> m<-2
> r<-20000
>
> thetavec<-matrix(0,m,r)
>
> thetavec[1,]<-rnorm(20000,0,1)
> for (i in 1:r) {thetavec[2,i]<-2-
0.375*thetavec[1,i]+sqrt(0.109375)*rnorm(1,0,1)}
> mean(thetavec[1,])
[1] -0.007715923
> mean(thetavec[2,])
[1] 2.004965
> v1<-var(thetavec[1,])
> v2<-var(thetavec[2,])
> v12<-cov(thetavec[1,],thetavec[2,])
> Vsim<-matrix(c(v1,v12,v12,v2),2,2)
> Vsim
          [,1]       [,2]
[1,]  1.0113990 -0.3806295
[2,] -0.3806295  0.2519804
>
```

$$\begin{pmatrix}\theta_1\\ \theta_2\end{pmatrix} \sim N\left(\begin{bmatrix}0\\ 2\end{bmatrix}, \begin{bmatrix}1 & -\frac{3}{4}\times 1\times\frac{1}{2} = -0.375\\ -0.375 & \left(\frac{1}{2}\right)^2\end{bmatrix}\right)$$

GIBBS SAMPLING

Want to sample from joint posterior

[A,B,C|DATA]

Sample is

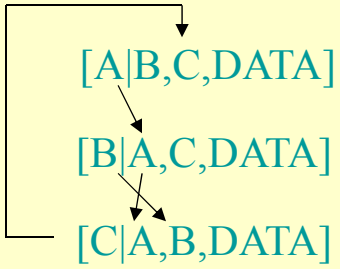
$[A^{(j)}, B^{(j)}, C^{(j)} | DATA]$

Each coordinate is a draw from marginal posterior

$[A^{(j)} | DATA]$ $[B^{(j)} | DATA]$ $[C^{(j)} | DATA]$

Gibbs sampling works as follows:

1) Form all fully conditional posteriors
2) Draw and update successively
3) Repeat a number of times without storing samples (burn-in)
4) Collect all subsequent samples, and thin them if needed for storage purposes



105

At the end of process:

j	A	B	C	
1	$A^{(1)}$	$B^{(1)}$	$C^{(1)}$	Discard first *t* samples as burn-in
2	$A^{(2)}$	$B^{(2)}$	$C^{(2)}$	
.	.	.	.	
t	$A^{(t)}$	$B^{(t)}$	$B^{(t)}$	
t+1	$A^{(t+1)}$	$B^{(t+1)}$	$B^{(t+1)}$	Keep subsequent *m* samples for Posterior analysis
.	.	.	.	
t+m	$A^{(t+m)}$	$B^{(t+m)}$	$B^{(t+m)}$	

106

Sampling from the bivariate normal distribution using the Gibbs sampler

Wish to draw samples from $\begin{pmatrix} \theta_1 \\ \theta_2 \end{pmatrix} \sim N\left(\begin{bmatrix} 0 \\ 2 \end{bmatrix}, \begin{bmatrix} 1 & -\frac{3}{4} \times 1 \times \frac{1}{2} = -0.375 \\ -0.375 & \left(\frac{1}{2}\right)^2 \end{bmatrix} \right)$

$$\theta_1|\theta_2 \sim N(4.0 - 1.5\theta_2, 0.109375)$$
$$\theta_2|\theta_1 \sim N(2 - 0.375\theta_1, 0.4375)$$

```
> # Simulation by Gibbs sampling
>
> L<-15000                    #Chain length
> burn<-1000                  #Length of burn-in
> thetavec<-matrix(0,L,2)     #Contains the chain
> mu1<-0                      #True mean of theta 1
> mu2<-2                      #True mean of theta 2
> sigma1<-1                   #True SD of theta 1
> sigma2<-0.5                 #True SD of theta 2
> rho<--0.75                  #True correlation
> s1<-sqrt(1-rho^2)*sigma1    #SD of cond. distribution of 1 given 2
> s2<-sqrt(1-rho^2)*sigma2    #SD of cond. distribution of 2 given 1
```

107

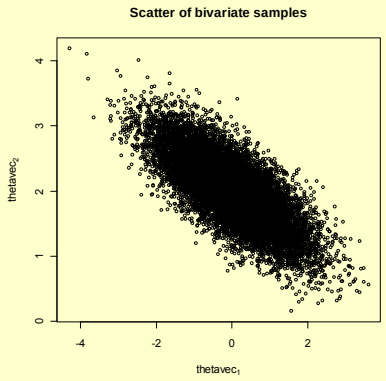
```
###Run the chain####

thetavec[1,]<-c(0,0)                #Initialize

for (i in 2:L) {                    #Open loop
  thetasamp2<-thetavec[i-1,2]       #Sample of theta2
  m1<-mu1+(rho*sigma1/sigma2)*(thetasamp2-mu2)
  thetavec[i,1]<-rnorm(1,m1,s1)
  thetasamp1<-thetavec[i,1]         #Sample of theta1

  m2<-mu2+(rho*sigma2/sigma1)*(thetasamp1-mu1)
  thetavec[i,2]<-rnorm(1,m2,s2)
  }                                 #Close loop

b<-burn+1
theta<-thetavec[b:L,]               #post-burn in samples
```

108

```
#Evaluate samples
colMeans(theta)
cov(theta)
cor(theta)
#Plot samples
plot(theta,main="Scatter of bivariate samples", cex=.5, xlab=bquote(thetavec[1]),
     ylab=bquote(thetavec[2]),ylim=range(theta[,2]))
```

$$\begin{pmatrix}\theta_1\\ \theta_2\end{pmatrix} \sim N\left(\begin{bmatrix}0\\ 2\end{bmatrix}, \begin{bmatrix}1 & -\frac{3}{4}\times 1\times\frac{1}{2} = -0.375\\ -0.375 & \left(\frac{1}{2}\right)^2\end{bmatrix}\right)$$

```
> colMeans(theta)
[1] -0.01379416  2.00553885
> cov(theta)
           [,1]       [,2]
[1,]  0.9910101 -0.3735499
[2,] -0.3735499  0.2506818
> cor(theta)
           [,1]       [,2]
[1,]  1.0000000 -0.7494595
[2,] -0.7494595  1.0000000
```



109

GIBBS SAMPLING IN A BETA-BINOMIAL MODEL: draw samples from the predictive distribution

Likelihood and prior

$$p(y_1, y_2, \ldots, y_n|\theta) = \prod_{i=1}^{n} \theta^{y_i}(1-\theta)^{1-y_i}$$

$$p(\theta) \propto \theta^{a-1}(1-\theta)^{b-1}$$

Posterior

$$p(\theta|\mathbf{y}) \propto \prod_{i=1}^{n} \theta^{y_i}(1-\theta)^{1-y_i}\theta^{a-1}(1-\theta)^{b-1}$$

$$\propto \theta^{\sum_{i=1}^{n} y_i + a - 1}(1-\theta)^{n - \sum_{i=1}^{n} y_i + b - 1}$$ (Beta density)

110

Joint density of parameter and future observation

$$p(y_f, \theta|\mathbf{y}) = p(y_f|\theta, \mathbf{y})p(\theta|\mathbf{y})$$

$$\propto \theta^{y_f}(1-\theta)^{1-y_f}\theta^{\sum_{i=1}^{n} y_i + a - 1}(1-\theta)^{n - \sum_{i=1}^{n} y_i + b - 1}$$

The fully conditionals

$$p(y_f|\theta, \mathbf{y}) = p(y_f|\theta)$$

$$p(\theta|y_f, \mathbf{y}) = Beta\left(y_f + \sum_{i=1}^{n} y_i + a, 1 - y_f + n - \sum_{i=1}^{n} y_i + b\right)$$

- Initialize the parameter
- Sample y_f from Binomial distribution with this parameter
- Sample parameter from Beta distribution with parameters as above
- Repeat many times, throw early draws (burn-in)

EXAMPLE: MH FOR A GLIM (Carlin and Louis, 2000)

Number of flour beetles killed after exposure to carbon disulphide

Dosage w_i	No. Killed y_i	No. Exposed n_i
1.6097	6	59
1.7242	13	60
1.7552	18	62
1.7842	28	56
1.8113	52	63
1.8369	53	59
1.8610	61	62
1.8839	60	60

Generalized logit model

$$\Pr(death|w) = h(w) = \left[\frac{\exp(x)}{1+\exp(x)}\right]^{m_1}$$

w_i = dose $i = 1, 2, \ldots, k$

$x = \frac{w - \mu}{\sigma}$

$m_1 > 0$

Unknown parameters

Priors

$$m_1 \sim Gamma(a_0, b_0) \propto m_1^{a_0 - 1}\exp\left(-\frac{m_1}{b_0}\right) \qquad \mu \sim N(c_0, d_0)$$

$$\sigma^2 \sim Inverse\ Gamma(e_0, f_0) \propto (\sigma^2)^{-(e_0+1)}\exp\left(-\frac{1}{f_0\sigma^2}\right)$$

112

Joint posterior

$$p(\mu,\sigma^2,m_1|\mathbf{y},a_0,b_0,c_0,d_0,e_0,f_0)$$

$$\propto \left\{\prod_{i=1}^{k}[h(w_i)]^{y_i}[1-h(w_i)]^{n_i-y_i}\right\}\exp\left[-\frac{(\mu-c_0)^2}{2d_0^2}\right]$$

$$\times(\sigma^2)^{-(e_0+1)}\exp\left(-\frac{1}{f_0\sigma^2}\right)m_1^{a_0-1}\exp\left(-\frac{m_1}{b_0}\right)$$

$$\propto \left\{\prod_{i=1}^{k}[h(w_i)]^{y_i}[1-h(w_i)]^{n_i-y_i}\right\}\frac{m_1^{a_0-1}}{(\sigma^2)^{(e_0+1)}}\exp\left[-\frac{(\mu-c_0)^2}{2d_0^2}-\frac{m_1}{b_0}-\frac{1}{f_0\sigma^2}\right]$$

Joint posterior is not recognizable…Use Metropolis-Hastings

113

Transform variables, to work on

$$\begin{bmatrix}\theta_1=\mu\\ \theta_2=\frac{1}{2}\log(\sigma^2)\\ \theta_3=\log(m_1)\end{bmatrix} <=> \begin{bmatrix}\mu=\theta_1\\ \sigma^2=\exp(2\theta_2)\\ m_1=\exp(\theta_3)\end{bmatrix}$$

$\Re^3$ so that Gaussian proposals can be used

$$J=\begin{bmatrix}\frac{\partial\mu}{\partial\theta_1} & \frac{\partial\mu}{\partial\theta_2} & \frac{\partial\mu}{\partial\theta_3}\\ \frac{\partial\sigma^2}{\partial\theta_1} & \frac{\partial\sigma^2}{\partial\theta_2} & \frac{\partial\sigma^2}{\partial\theta_3}\\ \frac{\partial m_1}{\partial\theta_1} & \frac{\partial m_1}{\partial\theta_2} & \frac{\partial m_1}{\partial\theta_3}\end{bmatrix}=\begin{bmatrix}1 & 0 & 0\\ 0 & 2\exp(2\theta_2) & 0\\ 0 & 0 & \exp(\theta_3)\end{bmatrix}$$

$$\rightarrow |J|=2\exp(2\theta_2+\theta_3)$$

114

New density= old density (evaluated at transformed variables) times Jacobian

$$p(\theta_1,\theta_2,\theta_3|\mathbf{y},a_0,b_0,c_0,d_0,e_0,f_0)$$
$$\propto \left\{\prod_{i=1}^{k}[h(w_i)]^{y_i}[1-h(w_i)]^{n_i-y_i}\right\}\frac{[\exp(\theta_3)]^{a_0-1}}{(\exp(2\theta_2))^{(e_0+1)}}$$
$$\times \exp\left[-\frac{(\theta_1-c_0)^2}{2d_0^2}-\frac{\exp(\theta_3)}{b_0}-\frac{1}{f_0\exp(2\theta_2)}\right]\exp(2\theta_2+\theta_3)$$

Collecting terms

$$p(\theta_1,\theta_2,\theta_3|\mathbf{y},a_0,b_0,c_0,d_0,e_0,f_0)$$
$$\propto \left\{\prod_{i=1}^{k}[h(w_i)]^{y_i}[1-h(w_i)]^{n_i-y_i}\right\}\exp(a_0\theta_3-2e_0\theta_2)$$
$$\times \exp\left[-\frac{(\theta_1-c_0)^2}{2d_0^2}-\frac{\exp(\theta_3)}{b_0}-\frac{1}{f_0\exp(2\theta_2)}\right]$$

POSTERIOR IS NOT RECOGNIZABLE…

115

Hyper-parameters: a_0= .25, b_0=4, c_0=2, d_0=10, e_0=2.000004, f_0=1000

1) Metropolis-Hastings proposal distribution used

$$\begin{bmatrix}\theta_1^*\\ \theta_2^*\\ \theta_3^*\end{bmatrix} \sim N\left(\begin{bmatrix}\theta_1^{[t-1]}\\ \theta_2^{[t-1]}\\ \theta_3^{[t-1]}\end{bmatrix}, \mathbf{D}=\begin{bmatrix}.00012 & 0 & 0\\ 0 & .033 & 0\\ 0 & 0 & .10\end{bmatrix}\right)$$

$$R = \frac{p(y|\boldsymbol{\theta}^*)p(\boldsymbol{\theta}^*)/f\left(\boldsymbol{\theta}^*|\boldsymbol{\theta}^{[t-1]}\right)}{p\left(y|\boldsymbol{\theta}^{[t-1]}\right)p\left(\boldsymbol{\theta}^{[t-1]}\right)/f\left(\boldsymbol{\theta}^{[t-1]}|\boldsymbol{\theta}^*\right)}$$

$$f(\boldsymbol{\theta}^*|\boldsymbol{\theta}^{[t-1]}) = \frac{1}{(2\pi)^3|\mathbf{D}|}\exp\left[-\frac{1}{2}(\boldsymbol{\theta}^*-\boldsymbol{\theta}^{[t-1]})'\mathbf{D}^{-1}(\boldsymbol{\theta}^*-\boldsymbol{\theta}^{[t-1]})\right]$$
$$f(\boldsymbol{\theta}^{[t-1]}|\boldsymbol{\theta}^*) = \frac{1}{(2\pi)^3|\mathbf{D}|}\exp\left[-\frac{1}{2}(\boldsymbol{\theta}^{[t-1]}-\boldsymbol{\theta}^*)'\mathbf{D}^{-1}(\boldsymbol{\theta}^{[t-1]}-\boldsymbol{\theta}^*)\right]$$
$$f(\boldsymbol{\theta}^*|\boldsymbol{\theta}^{[t-1]}) = f(\boldsymbol{\theta}^{[t-1]}|\boldsymbol{\theta}^*)$$

Symmetric: use METROPOLIS RATIO

116

$$R = \frac{p(y|\boldsymbol{\theta}^*)p(\boldsymbol{\theta}^*)}{p(y|\boldsymbol{\theta}^{[t-1]})p(\boldsymbol{\theta}^{[t-1]})}$$

$$= \frac{\left\{\prod_{i=1}^{k}[h^*(w_i)]^{y_i}[1-h^*(w_i)]^{n_i-y_i}\right\}\exp\left[a_0\theta_3^* - 2e_0\theta_2^* - \frac{(\theta_1^*-c_0)^2}{2d_0^2} - \frac{\exp(\theta_3^*)}{b_0} - \frac{1}{f_0\exp(2\theta_2^*)}\right]}{\left\{\prod_{i=1}^{k}[h^{[t-1]}(w_i)]^{y_i}[1-h^{[t-1]}(w_i)]^{n_i-y_i}\right\}\exp\left[a_0\theta_3^{[t-1]} - 2e_0\theta_2^{[t-1]} - \frac{\left(\theta_1^{[t-1]}-c_0\right)^2}{2d_0^2} - \frac{\exp\left(\theta_3^{[t-1]}\right)}{b_0} - \frac{1}{f_0\exp\left(2\theta_2^{[t-1]}\right)}\right]}$$

$$= \left\{\prod_{i=1}^{k}\left[\frac{h^*(w_i)}{h^{[t-1]}(w_i)}\right]^{y_i}\left[\frac{1-h^*(w_i)}{1-h^{[t-1]}(w_i)}\right]^{n_i-y_i}\right\}$$

$$\times \exp\left[a_0\left(\theta_3^* - \theta_3^{[t-1]}\right) - 2e_0\left(\theta_2^* - \theta_2^{[t-1]}\right) - \frac{(\theta_1^*-c_0)^2 - \left(\theta_1^{[t-1]}-c_0\right)^2}{2d_0^2} - \frac{\exp\left(\theta_3^* - \theta_3^{[t-1]}\right)}{b_0}\right]\frac{\exp\left[-\frac{1}{f_0\exp(2\theta_2^*)}\right]}{\exp\left[-\frac{1}{f_0\exp\left(2\theta_2^{[t-1]}\right)}\right]}$$

Numerical stability is improved by computing acceptance ratio as $\boxed{R = \exp[\log(R)]}$

$$\log(R) = \sum_{i=1}^{k}\left\{y_i\log\left[\frac{h^*(w_i)}{h^{[t-1]}(w_i)}\right] + (n_i - y_i)\log\left[\frac{1-h^*(w_i)}{1-h^{[t-1]}(w_i)}\right]\right\}$$
$$+ a_0\left(\theta_3^* - \theta_3^{[t-1]}\right) - 2e_0\left(\theta_2^* - \theta_2^{[t-1]}\right) - \frac{(\theta_1^*-c_0)^2 - \left(\theta_1^{[t-1]}-c_0\right)^2}{2d_0^2} - \frac{\exp\left(\theta_3^* - \theta_3^{[t-1]}\right)}{b_0}$$
$$+ \frac{1}{f_0}\left[\frac{1}{\exp\left(2\theta_2^{[t-1]}\right)} - \frac{1}{\exp(2\theta_2^*)}\right]$$

117

➔Three parallel chains run each with 10,000 iterations
➔ Burn-in= 2,000 in each chain
➔Histograms based on the (10,000-2,000)3= 24,000 sampled values
➔Autocorrelations and inter-correlations estimated from chain 2

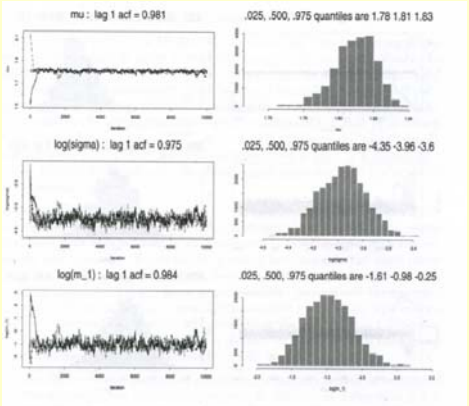


Figure 5.7 *Metropolis analysis of the flour beetle mortality data using a Gaussian proposal density with a diagonal $\tilde{\Sigma}$ matrix. Monitoring plots use three parallel chains, and histograms use all samples following iteration 2000. Overall Metropolis acceptance rate: 13.5%.*

-Chains mixed slowly (13.5% acceptance rate)
-High correlations between parameters:
-Makes sense to explore different proposal

$$\begin{bmatrix} 1 & -0.78 & -0.94 \\ -0.78 & 1 & 0.89 \\ & & 1 \end{bmatrix}$$

118

2) Metropolis-Hastings proposal distribution used

➔From first algorithm ,estimate posterior covariance matrix as $\hat{\Sigma} = \frac{1}{m}\sum_{j=1}^{m}(\theta^{(j)} - \overline{\theta})(\theta^{(j)} - \overline{\theta})'$

➔Use Gaussian proposal with covariance matrix (gave acceptance rate 27.3%)

$$\Psi = 2\hat{\Sigma} = \begin{bmatrix} 0.000292 & -003546 & -0.007856 \\ -003546 & 0.074733 & 0.117809 \\ -0.007856 & 0.117809 & 0.241551 \end{bmatrix}$$

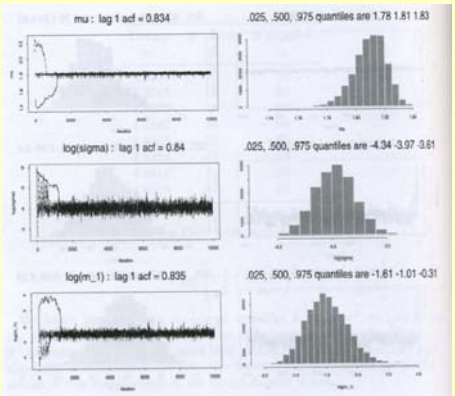


Figure 5.8 Metropolis analysis of the flour beetle mortality data using a Gaussian proposal density with a nondiagonal $\hat{\Sigma}$ matrix. Monitoring plots use three parallel chains, and histograms use all samples following iteration 2000. Overall Metropolis acceptance rate: 27.3%.

119

If fully conditionals are not recognizable, use other sampling methods

- Metropolis
- Metropolis-Hastings
- Acceptance/rejection sampling
- Importance sampling

These methods require an auxiliary distribution, which must be tuned, and calculation of an “acceptance probability”

120