Lecture 17
Bayesian Econometrics

1

Bayesian Econometrics: Introduction

• Idea: We are not estimating a parameter value, θ, but rather updating (changing) our subjective beliefs about θ.

• The centerpiece of the Bayesian methodology is Bayes theorem:

$$P(A|B) = P(A \cap B)/P(B) = P(B|A)\, P(A)/P(B).$$

• Think of B as "something known" –say, the data- and A as "something unknown" –e.g., the coefficients of a model.

• Our interest: Value of the parameters (θ), given the data (y).

• Reversing Bayes's theorem, we write the joint probability of θ and y:

$$P(\theta \cap y) = P(y|\theta)\, P(\theta)$$

Bayesian Econometrics: Introduction

• Then, we write: $P(\theta|y) = P(y|\theta)\, P(\theta)/P(y)$ (*Bayesian learning*)

• For estimation, we can ignore the term P(y) (a *normalizing constant*), since it does not depend on the parameters. Then, we can write:

$$P(\theta|y) \propto P(y|\theta) \times P(\theta)$$

• Terminology:

- $P(y|\theta)$: Density of the data, y, given the parameters, θ. Called the *likelihood function*. (I'll give you a value for θ, you should see y.)
- $P(\theta)$: *Prior* density of the parameters. Prior belief of the researcher.
- $P(\theta|y)$: *Posterior* density of the parameters, given the data. (A mixture of the prior and the "current information" from the data.)

Note: Posterior is proportional to likelihood times prior.

Bayesian Econometrics: Introduction

• The typical problem in Bayesian statistics involves obtaining the posterior distribution:

$$P(\theta|y) \propto P(y|\theta) \times P(\theta)$$

To get $P(\theta|y)$, we need:
- The likelihood, $P(y|\theta)$, will be assumed to be known. The likelihood carries all the current information about the parameters and the data.
- The prior, $P(\theta)$, will be also known. Q: Where does it come from?

Note: The posterior distribution embodies all that is "believed" about the model:

$$\text{Posterior} = f(\text{Model}|\text{Data})$$
$$= \text{Likelihood}(\theta,\text{Data}) \times \text{Prior}(\theta) \,/\, P(\text{Data})$$

Bayesian Econometrics: Introduction

• We want to get $P(\theta|y) \propto P(y|\theta) \times P(\theta)$. There are two ways to proceed to estimate $P(\theta|y)$:

(1) Pick $P(\theta)$ and $P(y|\theta)$ in such a manner that $P(\theta|y)$ can be analytically derived. This is the "old" way.

(2) Numerical approach. Randomly draw from $P(\theta|y)$ and, then, analyze the ED for θ. This is the modern way.

• Note: Nothing controversial about Bayes' theorem. For RVs with known pdfs, it is a *fact* of probability theory. But, the controvery starts when we model unkown pdfs and "*update*" them based on data.

Good Intro Reference (with references): "Introduction to Bayesian Econometrics and Decision Theory" by Karsten T. Hansen (2002).

Bayes' Theorem: Summary of Terminology

• Recall Bayes' Theorem:

$$P(\theta|y) = \frac{P(y|\theta)\,P(\theta)}{P(y)}$$

- $P(\theta)$: *Prior probability* about parameter θ.

- $P(y|\theta)$: Probability of observing the data, y, conditioning on θ. This conditional probability is called the *likelihood* –i.e., probability of event y will be the outcome of the experiment depends on θ.

- $P(\theta|y)$: *Posterior probability* -i.e., probability assigned to θ, after y is observed.

- $P(y)$: Marginal probability of y. This the prior probability of witnessing the data y under all possible scenarios for θ, and it depends on the prior probabilities given to each θ. (A normalizing constant from an estimation point of view.)

Bayes' Theorem: Example

Example: Player's skills evaluation in sports.

S: Event that the player has good skills (& be recruited by the team).

T: Formal tryout performance (say, good or bad).

After seeing videos and scouting reports and using her previous experience, the coach forms a personal belief about the player's skills. This initial belief is the *prior,* P(*S*).

After the formal tryout performance, the coach (event *T*) updates her prior beliefs. This update is the *posterior:*

$$\mathrm{P}(S|T) = \frac{\mathrm{P}(T \mid S)\,\mathrm{P}(S)}{\mathrm{P}(T)}$$

Bayes' Theorem: Example

Example: Player's skills evaluation in sports.

- P(*S*): Coach's personal estimate of the probability that the player has enough skills to be drafted –i.e., a good player- , based on evidence *other than* the tryout. (Say, .40.)
- P(*T*=*good*| *S*): Probability of seeing a good tryout performance if the player is actually good. (Say, .80.)
- *T* is related to *S*:

  P(*T*=*good*| *S (good player)*) = .80

  P(*T*=*good*| S^C *(bad player)*) = .20
- After the tryout, the coach updates her beliefs : P(*S* | *T*=*good*) becomes our new prior. That is:

$$\mathrm{P}(S|T=good) = \frac{\mathrm{P}(T=good \mid S)\,\mathrm{P}(S)}{\mathrm{P}(T=good)} = \frac{.80\text{x}.40}{.80\text{x}.40+.20\text{x}.60} = .7272$$

Bayesian Econometrics: Sequential Learning

• Consider the following data from N=50 Bernoulli trials:

00100100000101110000101000100000000000011000010100

If θ is the probability of a "1" at any one trial then the likelihood of any sequence of s trials containing y ones is

$$p(y|\theta) = \theta^{y} (1-\theta)^{s-y}$$

Let the prior be a uniform: p(θ)=1. Then, after 5 trials the posterior is:

$$p(\theta|y) \propto \theta (1-\theta)^{4} \times 1 = \theta (1-\theta)^{4}$$

and after 10 trials the posterior is

$$p(\theta|y) \propto \theta (1-\theta)^{4} \times \theta (1-\theta)^{4} = \theta^{2} (1-\theta)^{8}$$

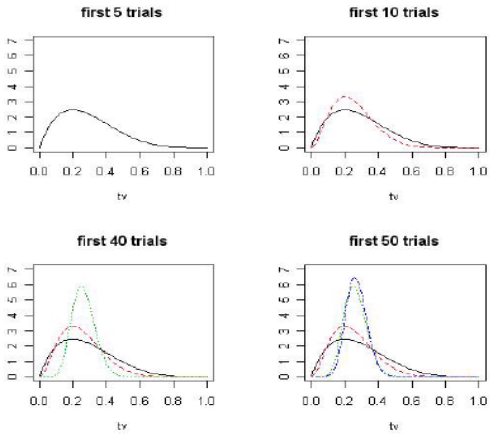
and after 40 trials the posterior is

$$p(\theta|y) \propto \theta^{8} (1-\theta)^{22} \times \theta^{2} (1-\theta)^{8} = \theta^{10} (1-\theta)^{30}$$

and after 50 trials the posterior is

$$p(\theta|y) \propto \theta^{4} (1-\theta)^{6} \times \theta^{10} (1-\theta)^{30} = \theta^{14} (1-\theta)^{36}$$

Bayesian Econometrics: Sequential Learning



• Notes:

- The previous posterior becomes the new prior.
- Beliefs tend to become more concentrated as N increases.
- Posteriors seem to look more normal as N increases.

Likelihood

• It represents the probability of observing the data, *y*, conditioning on θ. It is also called *sampling model*.

Example: Suppose the data follows a binomial distribution with probability of succcess θ. That is, $Y_1, Y_2, ..., Y_T \sim$ *i.i.d. Bin*$(1,\theta)$.

Then, the likelihood is

$$L(\mathbf{y}\,|\,\theta) = \theta^{\sum_i y_i}(1-\theta)^{T-\sum_i y_i}$$

Note: In this binomial case, it can be shown that the sum of successes, $\Sigma_i y_i$, is a sufficient statistic for θ and $p(y_1, y_2, ..., y_T\,|\,\theta)$. Moreover, $\Sigma_i y_i$ follows a *Bin(T,* θ*)*. These results are to be used later.

Likelihood: Normal

• Suppose $Y_i \sim$ *i.i.d.* $N(\theta, \sigma^2)$, then the likelihood is:

$$L(\mathbf{y}\,|\,\theta,\sigma^2) = (1/2\pi\sigma^2)^{T/2}\exp\{-\frac{1}{2\sigma^2}\sum_i (Y_i-\theta)^2\}$$

• There is a useful factorization, when $Y_i \sim$ *i.i.d.* $N(\theta, \sigma^2)$, which uses:

$$\sum_i (Y_i-\theta)^2 = \sum_i [(Y_i-\bar{Y})-(\theta-\bar{Y})]^2 = \sum_i (Y_i-\bar{Y})^2 + T(\theta-\bar{Y})^2 = (T-1)s^2 + T(\theta-\bar{Y})^2$$

where s^2 = sample variance. Then, the likelihood can be written as:

$$L(\mathbf{y}\,|\,\theta,\sigma^2) = (1/2\pi\sigma^2)^{T/2}\exp\{-\frac{1}{2\sigma^2}[(T-1)s^2 + T(\theta-\bar{Y})^2]\}$$

Note: Bayesians work with $h = 1/\sigma^2$, which is called "*precision*." A gamma prior is usually assumed for *h*. Then,

$$L(\mathbf{y}\,|\,\theta,\sigma^2) \;\propto\; (h)^{T/2}\exp\{-\frac{h}{2}[(T-1)s^2 + T(\theta-\bar{Y})^2]\}$$

$$\propto (h)^{T/2}\exp\{-\frac{h}{2}(T-1)s^2\}\ \text{x}\ \exp\{-\frac{Th}{2}(\theta-\bar{Y})^2\}$$

Priors

• A prior represents the (prior) belief of the researcher about θ, before seeing the data (**X,y**). These prior subjective probability beliefs about the value of θ are summarized with the *prior distribution*, P(θ).

Example: Suppose $Y_1, Y_2, ..., Y_T \sim$ *i.i.d. Bin*(1,θ). We know that $\Sigma_i y_i \sim Bin(T, \theta)$. Suppose we observe {Y=s}. Suppose from our prior information, we assume $\theta \sim$ Beta(α,β). That is,

$$p(\theta) = \frac{\Gamma(\alpha+\beta)}{\Gamma(\alpha)\Gamma(\beta)} \theta^{\alpha-1}(1-\theta)^{\beta-1}$$

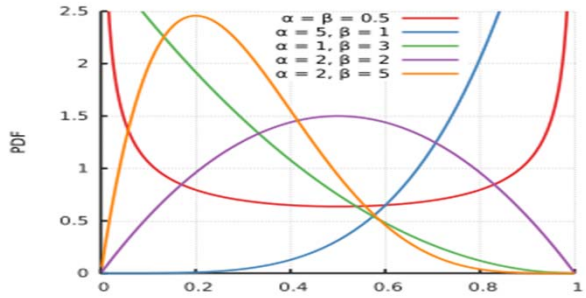
But, we could have assumed something different. For example, our prior information for θ tells us that all subintervals of [0,1] with the same length also have the same probability:

$$P(a \le \theta \le b) = P(a+c \le \theta \le b+c) \quad \text{for } 0 \le a < b < b+c \le 1,$$

which leads to a uniform for θ ⇒ P(θ)=1 for all θ ∈ [0,1].

Aside: The Beta Distribution

• Beta's pdf: $p(\theta) = \frac{\Gamma(\alpha+\beta)}{\Gamma(\alpha)\Gamma(\beta)} \theta^{\alpha-1}(1-\theta)^{\beta-1}$



• E[θ] = α/(β+α)

Var[θ] = α β/[(β+α)^2 (α +β+1)] = E[θ] (1-E[θ])/(α+β+1)

• When α & β are high, the Beta distribution can be approximated by a Normal.

Priors: Improper and Proper

• We can have *Improper* and *Proper* priors.

$$\text{Prob}\left(\theta_i|y\right)=\frac{\text{Prob}\left(y\mid\theta_i\right)\text{Prob}\left(\theta_i\right)}{\sum_j \text{Prob}\left(y\mid\theta_j\right)\text{Prob}\left(\theta_j\right)}$$

If we multiply $P(\theta_i)$ and $P(\theta_j)$ by a constant, the posterior probabilities will still integrate to 1 and be a proper distribution. But, now the priors do not integrate to 1. They are no longer proper.

• When this happens, the prior is called an *improper prior*. However, the posterior pdf need not be a proper pdf if the prior is improper.

"*Improper priors are not true pdfs, but if we pretend that they are, we will compute posterior pdfs that approximate the posteriors that we would have obtained using proper conjugate priors with extreme values of the prior hyperparameters*," from Degroot and Schervish's (2011) textbook.

Priors: Informative and Non-informative

• In a previous example, we assumed a prior $P(S)$ –i.e., a coach's prior belief about a player's skills, before tryouts.

• This is the Achilles heel of Bayesian statistics: Where do they came from?

• Priors can have many forms. We usually divide them in *non-informative* and *informative* priors for estimation of parameters

– Non-informative priors: There is a total lack of prior belief in the Bayesian estimator. The estimator becomes a function of the likelihood only.

– Informative prior: Some prior information enters the estimator. The estimator mixes the information in the likelihood with the prior information.

Priors: Informative and Non-informative

• Many statistitians like non-informative priors. Usual justification: "*Let the data speak for itself.*" According to this view, priors should play a small role in the posterior distribution.

• Non-informative priors can be called *diffuse, vague, flat, reference priors*.

• Uniform (*flat*) priors are usually taken as non-informative. There may be, however, other "less informative" priors.

• A formal definition of a non-informative prior is given by Jeffreys (1946).

• In general, with a lot of data the choice of flat priors should not matter, but when there is not a lot of data the choice of prior matters.

Priors: Informative and Non-informative

Example: Supposee we have *i.i.d.* Normal data, $Y_i \sim$ *i.i.d.* $N(\theta, \sigma^2)$. Assume σ^2 is known. We want to learn about θ, that is, we want to get $P(\theta|y)$. We need a prior for θ.

We assume a normal prior for θ: $P(\theta) \sim N(\theta_0, \sigma_0^2)$.

- θ_0 is our *best guess* for θ, before seeing y.

- σ_0^2 states the confidence in our prior. Small σ_0^2 shows big confidence. It is common to relate σ_0^2 to σ^2, say $\sigma_0^2 = \text{sqrt}\{\sigma^2 M\}$.

This prior gives us some flexibility. Depending on σ_0^2, this prior can be informative or diffuse. A small σ_0^2 represents the case of an informative prior. As σ_0^2 increases, the prior becomes more diffuse.

Q: Where do we get θ_0, σ_0^2? Previous data sets/a priori information?

Priors: Diffuse Prior - Example

Example: Suppose $Y_1, Y_2, ..., Y_T \sim$ *i.i.d. Bin*$(1,\theta)$. We know that $\Sigma_i y_i \sim Bin(T, \theta)$. Suppose we observe $\{Y=s\}$. Our prior information is not very good, it points towards a diffuse prior.

We formalize this information with a uniform distribution: $P(\theta)=1$ for all $\theta \in [0,1]$.

Detail for later: We can think of the Uniform as a special case of the Beta. Recall that the Beta(α,β) pdf is given by:

$$p(\theta) = \frac{\Gamma(\alpha+\beta)}{\Gamma(\alpha)\Gamma(\beta)}\theta^{\alpha-1}(1-\theta)^{\beta-1}$$

Then, setting $\alpha=1$ and $\beta=1$, delivers $P(\theta)=1$.

Priors: Jeffreys' Non-informative Prior

• Jeffreys (1946) provides a definition of non-informative priors, based on one-to-one transformations of the parameters.

• Jeffreys' general principle is that any rule for determining the prior pdf should yield an equivalent posterior if applied to the transformed parameters. The posterior should be invariant to the prior.

• Jeffreys' principle leads to defining the non-informative prior as $\Rightarrow$ $P(\theta) \propto [I(\theta)]^{1/2}$, where $I(\theta)$ is the Fisher information for θ:

$$I(\theta) = E\left[\left(\frac{\partial \log P(\theta)}{\partial \theta}\right)^2\right] = -E\left[\left(\frac{\partial^2}{\partial \theta^2}\log P(\theta)\right)^2\right]$$

If we take $[I(\theta)]^{1/2}$ as our prior, we call it the Jeffreys' prior for the likelihood $P(y|\theta)$.

Priors: Jeffreys' Non-informative Prior

• **Example:** Suppose $Y_1, ..., Y_T \sim$ *i.i.d. Bin*$(1,\theta)$. Then,

$\Sigma_i y_i = s \sim Bin(T,\theta)$, with a log-likelihood:

$$\log P(s \mid \theta) = c + s \log \theta + (T - s)\log(1 - \theta)$$

Then,

$$I[\theta] = -E\left[\left(\frac{\partial^2}{\partial \theta^2}\log P(\theta)\right)^2\right] = \frac{T}{\theta(1-\theta)}$$

Jeffreys' prior: $P(\theta) \propto \theta^{-1/2}(1-\theta)^{-1/2} \Longrightarrow$ a Beta(1/2,1/2).

Q: Non-informative? The uniform used before is a Beta(1,1). You can check later that Jeffreys' prior gives a lower weight to the prior information in the posterior. In this sense, it is "less informative."

Priors: Conjugate Priors

• When the posterior distributions $P(\theta \mid y)$ are in the same family $\mathcal{F}$ as the prior probability distributions, $P(\theta)$, the prior and posterior are then called *conjugate distributions*.

• Formally, let $P(\theta) \in \mathcal{F} \Rightarrow P(\theta \mid y) \in \mathcal{F}$. Then, $\mathcal{F}$ is *conjugate prior* for likelihood model $P(y \mid \theta)$.

• **Examples**:

- The beta distribution conjugates to itself (or *self-conjugate*) with respect to the Binomial likelihood.
- The normal family is conjugate to itself with respect to a normal likelihood function.

• Good! We know a lot about the normal and beta distributions.

Priors: Conjugate Priors

• Another good results: We can also generate values from these distributions with R (or other programs, like Matlab, Gauss, etc.). For example, *rbeta* and *rnorm* do the trick in R for the beta and normal distributions.

• Conjugate priors help to produce tractable posteriors.

• Q: What happens when we do not have conjugacy? We may have to deal with complicated posteriors –i.e., not easy to analytically integrate. In these cases, we will rely on numerical solutions.

Priors: Conjugate Priors - Example

Example: Suppose $Y_1, ..., Y_T \sim$ *i.i.d. Bin*(1,θ). Then, $\Sigma_i y_i \sim Bin(T, \theta)$. Suppose we observe {Y=s}. We assume $\theta \sim$ Beta(α,β). That is,

$$p(\theta) = \frac{\Gamma(\alpha+\beta)}{\Gamma(\alpha)\Gamma(\beta)}\theta^{\alpha-1}(1-\theta)^{\beta-1}$$

Then, the posterior is:

$$p(\theta \mid s) = \frac{\binom{T}{s}\theta^{s}(1-\theta)^{T-s} \times \frac{\Gamma(\alpha+\beta)}{\Gamma(\alpha)\Gamma(\beta)}\theta^{\alpha-1}(1-\theta)^{\beta-1}}{P(s)} \propto \theta^{s+\alpha-1}(1-\theta)^{T-s+\beta-1}$$

which looks, ignoring constants, like a Beta(s+α, *T*+β-s).

Note: When α & β are high, the Beta distribution can be approximated by a Normal. If a previous data set/prior info implies a mean and variance, they can be used to get the prior (α,β) values.

Priors: Hierarchical Models

• Bayesian methods can be effective in dealing with problems with a large number of parameters. In these cases, it is convenient to think about the prior for a vector parameters in stages.

• Suppose that $\boldsymbol{\theta}=(\theta_1, \theta_2, ..., \theta_K)$ and $\boldsymbol{\lambda}$ is a another parameter vector, of lower dimension than $\boldsymbol{\theta}$. $\boldsymbol{\lambda}$ maybe a parameter of a prior or a random quantity. The prior $p(\boldsymbol{\theta})$ can be derived in stages:

$$p(\boldsymbol{\theta},\lambda) = p(\boldsymbol{\theta}\,|\,\boldsymbol{\lambda})p(\boldsymbol{\lambda}).$$

Then,

$$p(\boldsymbol{\theta}) = \int p(\boldsymbol{\theta}\,|\,\boldsymbol{\lambda})p(\boldsymbol{\lambda})\, d\boldsymbol{\lambda}$$

We can write the joint as:

$$p(\boldsymbol{\theta},\boldsymbol{\lambda},\mathbf{y}) = p(\mathbf{y}\,|\,\boldsymbol{\theta},\boldsymbol{\lambda})\, p(\boldsymbol{\theta}\,|\,\boldsymbol{\lambda})p(\boldsymbol{\lambda}).$$

Priors: Hierarchical Models

• We can think of the joint $p(\boldsymbol{\theta},\boldsymbol{\lambda},\mathbf{y})$ as the result of a Hierarchical (or "*Multilevel*") Model:

$$p(\boldsymbol{\theta},\boldsymbol{\lambda},\mathbf{y}) = p(\mathbf{y}\,|\,\boldsymbol{\theta},\boldsymbol{\lambda})\, p(\boldsymbol{\theta},\lambda) = p(\mathbf{y}\,|\,\boldsymbol{\theta},\boldsymbol{\lambda})\, p(\boldsymbol{\theta}\,|\,\boldsymbol{\lambda})p(\boldsymbol{\lambda}).$$

The prior $p(\boldsymbol{\theta},\lambda)$ is decomposed using a prior for the prior, $p(\boldsymbol{\lambda})$, a *hyperprior*. Under this interpretation, we call $\boldsymbol{\lambda}$ *a hyperparameter*.

• Hierarchical models can be very useful, since it is often easier to work with conditional models than full joint models.

Example: In many stochastic volatility models, we estimate the time-varying variance (H_t) along with other parameters (θ). We write the joint as:

$$f(H_t,\theta) \propto f(Y_t|H_t)f(H_t|,\theta)f(\theta)$$

Priors: Hierarchical Models - Example

• Supposee we have *i.i.d.* Normal data, $Y_i \sim N(\theta, \sigma^2)$. We want to learn about (θ, σ^2) or, using $h = 1/\sigma^2$, $\boldsymbol{\varphi} = (\theta, h)$. That is, we want to get $P(\boldsymbol{\varphi} \mid y)$. We need a joint prior for $\boldsymbol{\varphi}$.

It can be easier to work with $P(\boldsymbol{\varphi}) = P(\theta \mid \sigma^{-2})\, P(\sigma^{-2})$.

For $\theta \mid \sigma^{-2}$, we assume $P(\theta \mid \sigma^{-2}) \sim N(\theta_0, \sigma_0^2)$, where $\sigma_0^2 = \text{sqrt}\{\sigma^2 M\}$.

For σ^2, we assume an inverse gamma (IG). Then, for $h = \sigma^{-2}$, we have a gamma distribution, which is function of (α,λ):

$$f(x;\alpha,\lambda) = \frac{\lambda^{\alpha}}{\Gamma(\alpha)}\, x^{\alpha-1}\, e^{-\lambda x} \Rightarrow f(h) = \frac{(\Phi/2)^{T/2}}{\Gamma(T/2)}\, (\sigma^{-2})^{(\frac{T}{2}-1)}\, e^{-(\Phi/2)\sigma^{-2}}$$

where $\alpha = T/2$ and $\lambda = 1/(2\eta^2) = \Phi/2$ are usual priors (η^2 is related to the variance of the T $N(0, \eta^2)$ variables we are implicitly adding).

Priors: Hierarchical Models - Example

Then, the joint prior, $P(\boldsymbol{\varphi})$ can be written as:

$$f(\theta,\sigma^{-2}) = (2\pi\sigma^2 M)^{-1/2} \exp\{-\frac{(\theta-\theta_0)^2}{2\sigma^2 M}\} \text{ x } \frac{(\Phi/2)^{T/2}}{\Gamma(T/2)}\, (\sigma^{-2})^{(T/2-1)}\, e^{-(\Phi/2)\sigma^{-2}}$$

Priors: Inverse Gamma for σ^2

• The usual prior for σ^2 is the *inverse-gamma* (IG). Recall that if X has a $\Gamma(\alpha,\lambda)$ distribution, then $1/X$ has an IG distribution with parameters α (shape) and λ^{-1} (scale). That is:

$$f(x;\alpha,\lambda) = \frac{\lambda^{\alpha}}{\Gamma(\alpha)}\ x^{-\alpha-1}\ e^{-(\lambda/x)} \qquad x>0.$$

• Then, $h=1/\sigma^2$ is distributed as $\Gamma(\alpha,\lambda)$:

$$f(x=\sigma^{-2};\alpha,\lambda) = \frac{\lambda^{\alpha}}{\Gamma(\alpha)}\ x^{\alpha-1}\ e^{-\lambda x} \qquad x>0.$$

• Q: Why do we choose an IG prior for σ^2?

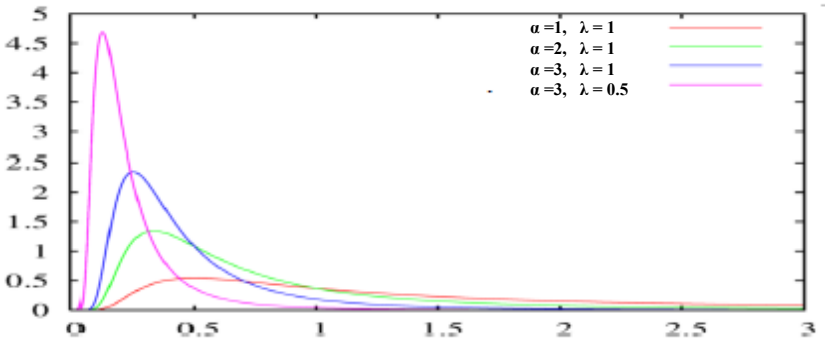
(1) $p(\sigma^2) = 0$ for $\sigma^2 < 0$.

(2) Flexible shapes for different values for α,λ –recall, when $\alpha = \nu/2$ and $\lambda = ½$, the gamma distribution becomes the χ^2_ν.

(3) *Conjugate prior* $\Rightarrow$ the posterior of $\sigma^{-2}|\mathbf{X}$ will also be $\Gamma(\alpha^*,\lambda^*)$.

Aside: The Inverse Gamma Distribution

• IG's pdf: $f(x;\alpha,\lambda) = \frac{\lambda^{\alpha}}{\Gamma(\alpha)}\ x^{-\alpha-1}\ e^{-(\lambda/x)} \qquad x>0.$



• Mean[x] = $\lambda/(\alpha-1)$ $(\alpha>1)$.
Var[x] = $\lambda^2/[(\alpha-1)(\alpha-2)]$ $(\alpha>2)$.

• A multivariate generalization of the IG distribution is the *inverse-Wishart* (IW) distribution.

Prior Information: Intuition

• (From Jim Hamilton.) Assume CLM with k=1. A student says: *"There is a 95% probability that β is between $b \pm 1.96\ sqrt\{\sigma^2 (\boldsymbol{X}'\boldsymbol{X})^{-1}\}$."*

A classical statistician says: *"No! β is a population parameter. It either equals 1.5 or it doesn't. There is no probability statement about β."*
"What is true is that if we use this procedure to construct an interval in thousands of different samples, in 95% of those samples, our interval will contain the true β."

• OK. Then, we ask the classical statistician:
- "Do you know the true β?" "No."
- "Choose between these options. Option A: I give you $5 now. Option B: I give you $10 if the true β is in the interval between 2.5 and 3.5." "I'll take the $5, thank you."

Prior Information: Intuition

• OK. Then, we ask the classical statistician, again:
- "Good. But, how about these? Option A: I give you $5 now. Option B: I give you $10 if the true β is between -8.0 and +13.9."
"OK, I'll take option B."

• Finally, we complicate the options a bit:
- "Option A: I generate a uniform number between 0 and 1. If the number is less than π, I give you $5.
Option B: I give you $5 if the true β is in the interval (2.0, 4.0). The value of π is 0.2"
"Option B."
- "How about if π = 0.8?"
"Option A."

Prior Information: Intuition

• Under certain axioms of rational choice, there will exist a unique π^*, such that he chooses Option A if $\pi > \pi^*$, and Option B otherwise. Consider π^* as the statistician's subjective probability.

• We can think of π^* as the statistician's subjective probability that β is in the interval (2.0, 4.0).

Posterior

• The goal is to say something about our subjective beliefs about θ; say, the mean θ, after seeing the data (**y**). We characterize this with the *posterior distribution*:

$$P(\theta|y) = P(y|\theta)\, P(\theta)/P(y)$$

• The posterior is the basis of Bayesian estimation. It takes into account the data (say, **y** & **X**) and our prior distribution (say, θ_0).

• $P(\theta|y)$ is a pdf. It is common to describe it with the usual classical measures. For example: the mean, median, variance, etc. Since they are functions of the data, they are *Bayesian estimators*.

• Under a quadratic loss function, it can be shown that the posterior mean, $E[\theta|y]$, is the *optimal Bayesian estimator* of θ.

Posterior: Optimal Estimator

• We assume a loss function, $g(\theta,\hat{\theta})$, where $\hat{\theta}$ is an estimate. Let $\hat{\theta}$ solve the minization problem:

$$\min_{\theta} \int_{\Theta} g(\theta,\hat{\theta})\, p(\theta \mid Y)\, d\theta \qquad \text{where } \theta \in \Theta$$

Different loss functions, produce different optimal estimators.

Example: Quadratic loss function with scalar case: $g(\theta,\hat{\theta}) = (\theta - \hat{\theta})^2$

$$\begin{aligned} E_{\theta|y}[\theta - \hat{\theta}]^2 &= E_{\theta|y}[\theta - E[\theta|y] + E[\theta|y] - \hat{\theta}]^2 \\ &= E_{\theta|y}[(\theta - E[\theta|y])^2] + (E[\theta|y] - \hat{\theta})^2 + \\ &+ 2\, E_{\theta|y}[(\theta - E[\theta|y])]\ (E[\theta|y] - \hat{\theta})] \\ &= E_{\theta|y}[(\theta - E[\theta|y])^2] + (E[\theta|y] - \hat{\theta})^2 \end{aligned}$$

which is minimized at $\hat{\theta} = E[\theta|y]$.

• Similar calculations for $g(\theta,\hat{\theta}) = |\theta - \hat{\theta}|$ produce the median as $\hat{\theta}$.

Posterior: Example – Binomial-Uniform

Example: Data: $Y_1, Y_2, ..., Y_T \sim$ *i.i.d. Bin*$(1,\theta)$. Then, $\Sigma_i\, y_i \sim$ *Bin*(T, θ). We observe $\{Y=s\}$.

• **Likelihood**: $$L(Y = s \mid \theta) = \binom{T}{s} \theta^s (1-\theta)^{T-s}$$

• **Prior**. For $\theta \sim$Unif[0,1]. That is, $P(\theta)=1$ for all $\theta \in [0,1]$.

• **Posterior**. Likelihood x Prior:

$$p(\theta \mid s) = \frac{\binom{T}{s} \theta^s (1-\theta)^{T-s} \times 1}{P(s)} = c(s)\theta^s (1-\theta)^{T-s}$$

where $c(s)$ is a constant independent of θ. We recongnize $P(\theta|Y=s)$ as a Beta (up to a constant).

Posterior: Example – Binomial-Uniform

Example (continuation):

We can derive c(s) since P(θ|y) should integrate to 1. To recover the constant we use:

$$\int_0^1 \theta^{\alpha}(1-\theta)^{\beta-1}d\theta = \frac{\Gamma(\alpha)\Gamma(\beta)}{\Gamma(\alpha+\beta)}$$

Then,

$$p(\theta \mid s) = \frac{\Gamma(T+2)}{\Gamma(s+1)\Gamma(T-s+1)}\theta^{s}(1-\theta)^{T-s} = Beta(s+1, T-s+1)$$

Note: Uniform prior + Bernoulli/Binomial likelihood => Beta posterior.

Posterior: Presentation of Results

• P(θ|y) is a pdf. For the simple case, the one parameter θ, it can be graphed. But, if **θ** is a vector of many parameters, the multivariate pdf cannot be presented in a graph of it.

• It is common to present measures analogous to classical point estimates and confidence intervals ("*credibility intervals,*" also C.I.).

For example:

(1) $E(\theta \mid y) = \int \theta\, p(\theta \mid y)\, d\theta$ -- posterior mean

(2) $Var(\theta \mid y) = E(\theta^2 \mid y) - \{E(\theta \mid y)\}^2$ -- posterior variance

(3) $p(k_1 > \theta > k_2 \mid y) = \int_{k1>\theta>k2} p(\theta \mid y)\, d\theta$ -- C.I.

• In general, it is not possible to evaluate these integrals analytically. We rely on numerical methods.

Posterior: Presentation of Results - Example

Example: In the Binomial-Uniform previous example, we obtained the posterior $P(\theta|y=s)=\text{Beta}(s+1, T-s+1)$.

From this Beta posterior, we can calculate the usual descriptive statistics:

$$E[\theta|y] = \alpha/(\beta+\alpha) =(s+1)/[(T-s+1)+(s+1)] = (s+1)/(T+2)$$

$$\text{Var}[\theta|y] = \alpha\,\beta/[(\beta+\alpha)^2\,(\alpha+\beta+1)] = E[\theta|y]\,(1-E[\theta|y])/(\alpha+\beta+1) =$$
$$= (s+1)(T-s+1)]/[(T+2)^2\,(T+3)]$$

Posterior: Presentation of Results - Example

Example (continuation): Suppose we have a sample of $T=25$ adults with MBA degrees, with $s=15$ of them trading stocks.

That is, we have a Beta(16,11) posterior. We can easily calculate the posterior mean, the posterior variance and CI {0.1,0.4}:

$$E[\theta|s=15] = \alpha/(\alpha+\beta) = 16/27 = 59.27$$
$$\text{Var}[\theta|s=15] = \alpha\beta/[(\beta+\alpha)^2(\alpha+\beta+1)] = 16*11/[(27)^2*(28)]= 0.00862$$
$$P_{\theta|s}\,(0.1>\theta>0.4|s=15) = 0.02166.$$

Posterior: Hypothesis Testing

• In the context of C.I., we calculate the probability of θ being in some interval. This allows for some easy hypothesis tests.

For example, we are interested in testing H_0: θ>0 against H_1: θ≤0. We can test H_0 by computing $P_{\theta|y=s}(\theta>0)$ and check if it is lower than some small level α. If it is lower, we reject H_0: θ>0 in favor of H_1.

Example: In the Binomial-Uniform model, we derive the posterior $P(\theta|y=s)$ as Beta(s+1, T+1-s). Suppose we are interested in testing H_0: $\theta \leq 0.3$ against H_1: $\theta > 0.3$. Suppose T=25 and s=15.

Then,

$$\text{Beta}(\theta \leq 0.3 \,|\, 16,11) = .00085 \text{ (too small!)}$$

$$\Rightarrow \text{reject } H_0\text{: } \theta \leq 0.3 \text{ in favor of } H_1\text{: } \theta > 0.3.$$

Posterior: Example – Binomial-Beta

Example: Same binomial data as before. We observe {Y=s}.

• **Prior**. We change. Now, we assume $\theta \sim \text{Beta}(\alpha,\beta)$. That is,

$$p(\theta) = \frac{\Gamma(\alpha+\beta)}{\Gamma(\alpha)\Gamma(\beta)}\theta^{\alpha-1}(1-\theta)^{\beta-1}$$

• **Posterior**. Likelihood x Prior (ignoring constants):

$$p(\theta|\mathbf{y}=s) \propto \binom{T}{s}\theta^{s}(1-\theta)^{T-s} \times \frac{\Gamma(\alpha+\beta)}{\Gamma(\alpha)\Gamma(\beta)}\theta^{\alpha-1}(1-\theta)^{\beta-1} \propto \theta^{\alpha+s-1}(1-\theta)^{\beta+T-s-1}$$

which is a Beta(s+α, T+β-s). (Not a surprise, we used a *conjugate prior*!) We can write the usual descriptive moments; for example:

$$E[\theta|y=s] = (\alpha+s)/(\alpha+s+\beta+T-s) = (\alpha+s)/(\alpha+\beta+T).$$

Remark: We think of the Binomial-Uniform model as a special case of the Binomial-Beta model, with the Unif[0,1] as a Beta(α=1, β=1).

Posterior: Combining Information

• In the Binomial-Beta model, the posterior P(θ|y) is:

$$p(\theta \mid \mathbf{y}=s) \propto \theta^{\alpha+s-1}(1-\theta)^{\beta+T-s-1}$$

• The posterior P(θ|y) combines prior information (α, β) and data (*T*,*s*), which can be seen by writing E[θ|y=*s*] as:

$$E[\theta \mid s] = \frac{\alpha+\beta}{(\alpha+\beta+T)} \times \text{Pior Expectation} + \frac{T}{(\alpha+\beta+T)} \times \text{Data Average}$$

Usually, α is thought of "the prior number of 1's;" while β is thought of as "the prior number of 0's" ($\Longrightarrow$ α+β ≈ "prior sample size.") Then, the prior expectation is [α/(α+β)].

• Role of *T:*
As *T* grows => E[θ|y=*s*] ≈ *s*/*T* $\Longrightarrow$ Data dominates.
Similar for the variance: As *T* grows $\Longrightarrow$ Var[θ|y=*s*] ≈ s/T^2 * [1-(*s*/*T*)]

Posterior: Constants

• In the previous example, we derive the posterior for θ in a "Binomial-Beta model," ignoring constants:

$$p(\theta \mid \mathbf{y}=s) \propto \theta^{\alpha+s-1}(1-\theta)^{\beta+T-s-1}$$

• To be a well-defined Beta pdf –i.e., integrates to 1-, we find the constant of proportionality as we did for the Binomial-Uniform case:

$$\frac{\Gamma(\alpha+\beta+T)}{\Gamma(\alpha+s)\Gamma(\beta+T-s)}$$

• Bayesians use this trick to recognize posteriors. That is, once you recognize that the posterior distribution is proportional to a known probability density, then it must be identical to that density.

Note: The constant of proportionality must be constant with respect to θ.

Posterior: Example - Normal-Normal

• **Likelihood**. We have *i.i.d.* normal data: $Y_i \sim N(\theta, \sigma^2)$. Then:

$$L(\theta \mid \mathbf{y}, \sigma^2) \quad \alpha \ (h)^{T/2} \exp\{-\frac{h}{2}\sum_i (Y_i - \theta)^2\}$$

• **Priors**. We need a joint prior: $f(\theta, \sigma^2)$. In the Normal-Normal model, we assume σ^2 known (usually, we work with $h=1/\sigma^2$). Thus, we only specify a normal prior for θ: $f(\theta) \sim N(\theta_0, \sigma_0^2)$.

• σ_0^2 states the confidence in our prior. Small σ_0^2 shows confidence.

• In realistic applications, we add a prior for $f(\sigma^2)$. Usually, an IG.

• **Posterior**. Likelihood x Prior:

$$f(\theta \mid \mathbf{y}, \sigma^2) \alpha \ \ (h)^{T/2} \exp\{-\frac{h}{2}\sum_i (Y_i - \theta)^2\} \text{ x } \frac{1}{2\sigma_0} \exp\{-\frac{(\theta-\theta_0)^2}{2\sigma_0^2}\}$$

Posterior: Example - Normal-Normal

• Or using the Likelihood factorization:

$$f(\theta \mid \mathbf{y}, \sigma^2) \ \propto \ (h)^{T/2} \exp\{-\frac{h}{2}[(T-1)s^2 + T(\theta-\bar{Y})^2]\} \text{x} \frac{1}{2\sigma_0} \exp\{-\frac{(\theta-\theta_0)^2}{2\sigma_0^2}\}$$

$$\propto (h)^{T/2} \exp\{-\frac{h}{2}(T-1)s^2\} \text{ x } \exp\{-\frac{Th}{2}(\theta-\bar{Y})^2\} \text{x} \frac{1}{2\sigma_0} \exp\{-\frac{(\theta-\theta_0)^2}{2\sigma_0^2}\}$$

$$\propto (\frac{1}{\sigma})^{T/2} \frac{1}{2\sigma_0} \exp\{-\frac{1}{2\sigma^2}(T-1)s^2\} \text{ x } \exp\{-\frac{T}{2\sigma^2}(\theta-\bar{Y})^2 - \frac{(\theta-\theta_0)^2}{2\sigma_0^2}\}$$

• A little bit of algebra, using:

$$a(x-b)^2 + c(x-d)^2 = (a+c)(x-\frac{ab+cd}{a+c})^2 + \frac{ac}{a+c}(b-d)^2$$

we get for the 2nd expression inside the exponential:

$$\frac{T}{\sigma^2}(\theta-\bar{Y})^2 + \frac{(\theta-\theta_0)^2}{\sigma_0^2} = [\mathrm{T}/\sigma^2 + 1/\sigma_0^2](\theta-\bar{\theta})^2 + \frac{1}{\sigma_0^2+\sigma^2/T}(\bar{Y}-\theta_0)^2$$

Posterior: Normal-Normal

$$\frac{T}{\sigma^2}(\theta-\bar{Y})^2+\frac{(\theta-\theta_0)^2}{\sigma_0^2}=\frac{1}{\bar{\sigma}^2}(\theta-\bar{\theta})^2+\frac{1}{\sigma_0^2+\sigma^2/T}(\bar{Y}-\theta_0)^2$$

where $\bar{\theta}=\dfrac{(\mathrm{T}/\sigma^2)\bar{Y}+(1/\sigma_0^2)\theta_0}{\mathrm{T}/\sigma^2+1/\sigma_0^2}$ & $\bar{\sigma}^2=\dfrac{1}{\mathrm{T}/\sigma^2+1/\sigma_0^2}$

• Since we only need to include the terms in θ, then:

$$f(\theta|\mathbf{y},\sigma^2)\propto(\frac{1}{\sigma})^{T/2}\frac{1}{2\sigma_0}\exp\{-\frac{1}{2\sigma^2}(T-1)s^2\}\,\mathrm{x}\,\exp\{-\frac{1}{2\bar{\sigma}^2}(\theta-\bar{\theta})^2-\frac{1}{2(\sigma_0^2+\sigma^2/T)}(\bar{Y}-\theta_0)^2\}$$

$$\propto\exp\{-\frac{1}{2\bar{\sigma}^2}(\theta-\bar{\theta})^2\}$$

That is, the posterior is: $N(\bar{\theta},\bar{\sigma}^2)$

• The posterior mean, $\bar{\theta}$, is the Bayesian estimator. It takes into account the data (**y**) and our prior distribution. It is a weighted average of our prior θ_0 and $\bar{Y}$.

Posterior: Bayesian Learning

• Update formula for *θ:* $\bar{\theta}=\dfrac{(\mathrm{T}/\sigma^2)\bar{Y}+(1/\sigma_0^2)\theta_0}{\mathrm{T}/\sigma^2+1/\sigma_0^2}=\omega\bar{Y}+(1-\omega)\theta_0$

where $\omega=\dfrac{(\mathrm{T}/\sigma^2)}{\mathrm{T}/\sigma^2+1/\sigma_0^2}=\dfrac{\sigma_0^2}{\sigma_0^2+\sigma^2/\mathrm{T}}$

• The posterior mean is a weighted average of the usual estimator and the prior mean, θ_0.

Results:
- As $\mathrm{T}\to\infty$, the posterior mean $\bar{\theta}$ converges to $\bar{Y}$.
- As $\sigma_0^2\to\infty$, our prior information is worthless.
- As $\sigma_0^2\to 0$, complete certainty about our prior information.

This result can be interpreted as *Bayesian learning*, where we combine our prior with the observed data. Our prior gets updated! The extent of the update will depend on our prior distribution.

Posterior: Bayesian Learning

- As more information is known or released, the prior keeps changing.

Example: In R.

```
bayesian_updating <- function(data,mu_0,sigma2_0,plot=FALSE) {
  require("ggplot2")
T = length(data)                          # length of data
xbar = mean(data)                         # mean of data
sigma2 = sd(data)^2                       # variance of data

# Likelihood (Normal)
xx <- seq(xbar-2*sqrt(sigma2), xbar+2*sqrt(sigma2),sqrt(sigma2)/40)
yy <- 1/(sqrt(2*pi*sigma2/T))*exp(-1/(2 *sigma2/T)*(xx - xbar)^2 )
# yy <- 1/(xbar+4*sqrt(sigma2)-xbar+4*sqrt(sigma2))
df_likelihood <- data.frame(xx,yy,1)      # store data
type <- 1
df1 <- data.frame(xx,yy,type)

# Prior (Normal)
xx <- seq(mu_0-4*sqrt(sigma2_0), mu_0+4*sqrt(sigma2_0),(sqrt(sigma2_0)/40))
yy = 1/(sqrt(2*pi*sigma2_0))*exp(-1/(2 *sigma2_0)*(xx - mu_0)^2)
type <- 2
df2 <- rbind(df1,data.frame(xx,yy,type))
```

Posterior: Bayesian Learning

Example (continuation):

```
# Posterior
omega <- sigma2_0/(sigma2_0 + sigma2/T)
pom = omega * xbar + (1-omega)*mu_0                    # posterier mean
pov = 1/(T/sigma2 + 1/sigma2_0)                        # posterior variance
xx = seq(pom -4*sqrt(pov), pom + 4*sqrt(pov),(sqrt(pov)/40))
yy = 1/(sqrt(2 * pi * pov))*exp(-1./(2 *pov)* (xx - pom)^2 )
type <- 3
df3 <- rbind(df2,data.frame(xx,yy,type))
df3$type <- factor(df3$type,levels=c(1,2,3),
                   labels = c("Likelihood", "Prior", "Posterior"))

if(plot==TRUE){
  return(ggplot(data=df3, aes(x=xx, y=yy, group=type, colour=type))
         + ylab("Density")
         + xlab("x")
         + ggtitle("Bayesian updating")
         + geom_line()+theme(legend.title=element_blank()))
    } else {
    Nor <- matrix(c(pom,pov), nrow=1, ncol=2, byrow = TRUE)
    return(Nor)
    }
}
```

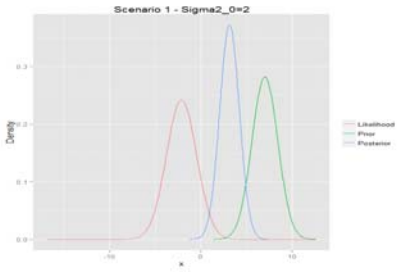
Posterior: Bayesian Learning

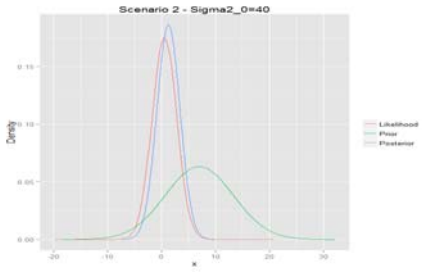
Example (continuation):

```
dat <- 5*rnorm(20,0,sqrt(2))                  # generate normal data T= 20, mean=0, var=50
                                              # xbar = -2.117, σ² = 54.27

# Scenario 1 – Precise prior (θ0=7, σ0² =2)
df <- bayesian_updating(dat,7,2,plot=TRUE)    # priors mu_0=7, sigma2_0=2
df                                            # ω = .4243, pom = 3.1314, pov = 1.1514

# Scenario 2 – Difusse prior (θ0=7, σ0² =40)
df <- bayesian_updating(dat,7,40,plot=TRUE)   # priors mu_0=7, sigma2_0=40
df                                            # ω = .9365, pom = -1.5382, pov = 2.5411
```





Posterior: James-Stein Estimator

- Let $x_t \sim N(\mu_t, \sigma^2)$ for t=1,2,...., T. Then, let MLE (also OLS) be $\hat{\mu}_t$.
- Let $m_1, m_2, ..., m_T$ be any numbers.
- Define

$$S = \Sigma_t (x_t - m_t)^2$$
$$\theta = 1 - [(T-2)\sigma^2/S]$$
$$m_i^* = \theta\, \hat{\mu}_t + (1-\theta)\, m_i$$

- **Theorem**: Under the previous assumptions,

$$E[\Sigma_t (\mu_t - m_t^*)^2] < E[\Sigma_t (\mu_t - \hat{\mu}_t)^2]$$

Remark: Some kind of shrinkage can always reduce the MSE relative to OLS/MLE.

Note: The Bayes estimator is the posterior mean of θ. This is a *shrinkage* estimator.

Predictive Posterior

• The posterior distribution of θ is obtained, after the data y is observed, by Bayes' Theorem::

$$P(\theta|y) \propto P(y|\theta) \times P(\theta)$$

Suppose we have a new set of observations, z, independent of y given θ. That is,

$$P(z,y|\theta) = P(z|\theta) \times P(y|\theta)$$

Then,

$$P(z|y) = \int P(z,\theta|y)\, d\theta = \int P(z|\theta,y)\, P(\theta|y)\, d\theta$$
$$= \int P(z|\theta)\, P(\theta|y)\, d\theta = E_{\theta|y}\,[P(z|\theta)]$$

$P(z|y)$ is the *predictive posterior distribution,* the distribution of new (unobserved) observations. It is equal to the conditional (over the posterior of $\theta|y$) expected value of the distribution of the new data, z.

Predictive Posterior: Example 1

Example: Player's skills evaluation in sports.
Suppose the player is drafted. Before the debut, the coach observes his performance in practices. Let Z be the performance in practices (again, good or bad). Suppose Z depends on *S* as given below:
$P(Z=good|S) = .95$
$P(Z=good|S^C) = .10$
(We have previously determined: $P(S|T=g) = 0.72727$.)

Using this information, the coach can compute predictive posterior of Z, given *T*. For example, the coach can calculate the probability of observing *Z=bad*, given *T=good:*

$P(Z=b|T=g) = P(Z=b|T=g, S^C)\, P(S^C|T=g) + P(Z=b|T=g,S)\, P(S|T=g)$
$= P(Z=b|S^C)\, P(S^C|T=g) + P(Z=b|S)\, P(S|T=g)$
$= .90 \times 0.27273 + .05 \times 0.72727 = .28182$

Note: Z and *T* are conditionally independent.

Predictive Posterior: Example 2

Example: We have $Y_1, ..., Y_{T=25} \sim$ *i.i.d. Bin*$(1,\theta)$. Let $\Sigma_i y_i = s$. We derive the predictive posterior of new data, Y*, as:

$$P(Y^*=1 \mid y_1, ..., y_T) = E[\theta \mid y=s] = (\alpha+s)/(\alpha+\beta+T)$$
$$P(Y^*=0 \mid y_1, ..., y_T) = 1 - P(Y^*=1 \mid y=s] = (\beta+T-s)/(\alpha+\beta+T)$$

Suppose we assume $\alpha=\beta=1$, $s=15$ and $T=25$. Then,

$$P(Y^*=1 \mid s) = 16/27 = 0.5926$$

Note: A Jeffreys' prior –i.e., a Beta(.5,.5)– is slightly less informative!

Remark: The predictive distribution does not depend upon unknown quantities. It depends on prior information and the observed data. The observed data gives us information about the new data, Y*.

Multivariate Models: Multivariate Normal

• So far, our models have been univariate models. Suppose, we are interested in the correlation between mutual fund returns. For this we need a multivariate setting.

• **Likelihood**: the most popular likelihood is the Multivariate normal model (MVN). We say **Y**, a *k*-dimensional data vector, has a MVN distribution if its sampling pdf is:

$$p(\boldsymbol{y} \mid \boldsymbol{\mu}, \boldsymbol{\Sigma}) = \frac{1}{(2\pi)^{n/2} |\boldsymbol{\Sigma}|^{1/2}} e^{-\frac{1}{2}(\boldsymbol{y}-\boldsymbol{\mu})'\boldsymbol{\Sigma}^{-1}(\boldsymbol{y}-\boldsymbol{\mu})}.$$

where $\boldsymbol{\mu}$ is the mean and $\boldsymbol{\Sigma}$ is the covariance matrix. Or

$$\mathbf{Y} \sim N_p(\boldsymbol{\mu}, \boldsymbol{\Sigma}).$$

• Recall a property of a MVN: The marginal distribution of each variable is also normal: $y_j \sim N(\mu_j, \sigma^2_j)$.

Multivariate Models: MVN – Prior for μ

• **Prior**: Following the univariate models and intuition, we propose a MVN prior for $\boldsymbol{\mu}$:

$$p(\boldsymbol{\mu}) \sim N_k(\boldsymbol{\mu}_0, \boldsymbol{\Lambda}_0).$$

where $\boldsymbol{\mu}_0$ is the prior mean and $\boldsymbol{\Lambda}_0$ is the prior covariance matrix of $\boldsymbol{\mu}$. We can write the prior as:

$$p(\mu) \propto e^{-\frac{1}{2}\mu' A_0 \mu + \mu' b_0}$$

where $\mathbf{A}_0 = \boldsymbol{\Lambda}_0^{-1}$ and $\boldsymbol{b}_0 = \boldsymbol{\Lambda}_0^{-1}\, \boldsymbol{\mu}_0$. (=> $\boldsymbol{\Lambda}_0 = \mathbf{A}_0^{-1}$ & $\boldsymbol{\mu}_0 = \mathbf{A}_0^{-1}\, \boldsymbol{b}_0$).

• Note that using a similar algebra and under the *i.i.d.* sampling model, we can write the joint likelihood as:

$$p(y_1,\ldots,y_N \mid \mu, \Sigma) \propto e^{-\frac{1}{2}\mu' A_1 \mu + \mu' b_1}.$$

where $\boldsymbol{A}_1 = N\,\boldsymbol{\Sigma}^{-1}$ and $\boldsymbol{b}_1 = N\,\boldsymbol{\Sigma}^{-1}\,\bar{\mathbf{y}}$.

Multivariate Models: MVN – P($\boldsymbol{\mu} \mid \mathbf{y}_1, \ldots, \mathbf{y}_N, \boldsymbol{\Sigma}$)

• **Posterior**: Likelihood x Prior. Then, the (conditional) posterior:

$$p(\mu \mid y_1,\ldots,y_N,\Sigma) \propto e^{-\frac{1}{2}\mu' A_1 \mu + \mu' b_1} \times e^{-\frac{1}{2}\mu' A_0 \mu + \mu' b_0} = e^{-\frac{1}{2}\mu' A_N \mu + \mu' b_N}$$

where $\mathbf{A}_N = \mathbf{A}_0 + \mathbf{A}_1 = \boldsymbol{\Lambda}_0^{-1} + N\,\boldsymbol{\Sigma}^{-1}$.

$\boldsymbol{b}_N = \boldsymbol{b}_0 + \boldsymbol{b}_1 = \boldsymbol{\Lambda}_0^{-1}\,\boldsymbol{\mu}_0 + N\,\boldsymbol{\Sigma}^{-1}\,\bar{\mathbf{y}}$.

A MVN with mean $\mathbf{A}_N^{-1}\,\boldsymbol{b}_N$ and covariance $\mathbf{A}_N^{-1}$. That is,

$\text{Cov}[\boldsymbol{\mu} \mid \mathbf{y}_1, \ldots, \mathbf{y}_N, \boldsymbol{\Sigma}] = \boldsymbol{\Lambda}_N = \mathbf{A}_N^{-1} = (\boldsymbol{\Lambda}_0^{-1} + N\,\boldsymbol{\Sigma}^{-1})^{-1}$

$\text{E}[\boldsymbol{\mu} \mid \mathbf{y}_1, \ldots, \mathbf{y}_N, \boldsymbol{\Sigma}] = \boldsymbol{\mu}_N = \mathbf{A}_N^{-1}\,\boldsymbol{b}_N = \boldsymbol{\Lambda}_N\,(\boldsymbol{\Lambda}_0^{-1}\,\boldsymbol{\mu}_0 + N\,\boldsymbol{\Sigma}^{-1}\,\bar{\mathbf{y}})$

• Similar to the univariate case: The posterior precision ($\mathbf{A}_N$) is the sum of the prior precision and data precision. The posterior expectation is a weighted average of the prior expectation and the sample mean.

Multivariate Models: MVN – Wishart PDF

• The results are conditional on $\mathbf{\Sigma}$. In general, we are also interested in learning about $\mathbf{\Sigma}$. Thus, we need a prior for $\mathbf{\Sigma}$ (a kxk symmetric pd matrix). We base our results on the multivariate version of the gamma distribution, the Wishart distribution.

• Similar to a gamma pdf, the Wishart pdf is a (semi-)conjugate prior for the precision matrix $\mathbf{\Sigma}^{-1}$. Then, the conjugate prior for $\mathbf{\Sigma}$ is the inverse-Wishart (IW).

• Conditions for $\mathbf{\Sigma} \sim \text{IW}(v_0, \mathbf{S_0}^{-1})$ distribution (with v_0 a positve integer, called *degrees of freedom*, and $\mathbf{S_0}$ a kxk symmetric pd matrix):

- Sample: $\mathbf{z}_1, ..., \mathbf{z}_{v0} \sim i.i.d.\ N_k\ (\mathbf{0}, \mathbf{S_0}^{-1})$
- $\mathbf{Z}'\mathbf{Z} = \Sigma_{i=1 \text{ to } v0}\ \mathbf{z}_i\ \mathbf{z}_i' \quad \Rightarrow \mathbf{\Sigma} = (\mathbf{Z}'\mathbf{Z})^{-1}$

Then, $\mathbf{\Sigma}^{-1} \sim \text{W}(v_0, \mathbf{S_0})$.

Multivariate Models: MVN – IW Prior for Σ

• The prior density for $\mathbf{\Sigma}$, an $\text{IW}(v_0, \mathbf{S_0}^{-1})$, is:

$$p(\mathbf{\Sigma}) \propto |\mathbf{\Sigma}|^{-\frac{1}{2}(v_0+k+1)} \times e^{-\frac{1}{2}tr(\mathbf{S_0}\mathbf{\Sigma}^{-1})}$$

• Properties:

$\text{E}[\mathbf{\Sigma}^{-1}] = v_0\ \mathbf{S_0}^{-1}$

$\text{E}[\mathbf{\Sigma}] = \mathbf{S_0}/(v_0\text{-}k\text{-}1)$

• Q: What are good values for v_0 & $\mathbf{S_0}$?

A: The larger v_0, the stronger the prior beliefs. For example, if we are confident that the true $\mathbf{\Sigma}$ is near some value, $\mathbf{\Sigma}_0$, then choose v_0 large and set $\mathbf{S_0}=(v_0\text{-}k\text{-}1)\ \mathbf{\Sigma}_0$ (the distribution is tightly centered around $\mathbf{\Sigma}_0$).

Vague IW priors, which make the correlations uniform, tend to associate large absolute correlations and large SDs. Potential problem!

Multivariate Models: MVN – IW Prior for Σ

• The prior density for $\boldsymbol{\Sigma}$, an IW(v_0, $\mathbf{S_0}^{-1}$), is:

$$p(\boldsymbol{y}_1,\ldots,\boldsymbol{y}_N \mid \boldsymbol{\mu},\boldsymbol{\Sigma}) \propto |\boldsymbol{\Sigma}|^{-\frac{N}{2}} \, e^{-\frac{1}{2}\sum_{i=1}^{n}(\boldsymbol{y}_i-\boldsymbol{\mu})'\boldsymbol{\Sigma}^{-1}(\boldsymbol{y}_i-\boldsymbol{\mu})}$$
$$\propto |\boldsymbol{\Sigma}|^{-\frac{N}{2}} \, e^{-\frac{1}{2}tr(\boldsymbol{S}_{\boldsymbol{\mu}}\boldsymbol{\Sigma}^{-1})}$$

where $\boldsymbol{S}_{\boldsymbol{\mu}} = \Sigma_{i=1 \text{ to } N} (\mathbf{y}_i - \boldsymbol{\mu})(\mathbf{y}_i - \boldsymbol{\mu})'$ is the RSS matrix for the vectors $\mathbf{y}_1$,..., $\mathbf{y}_N$, if the population mean is presumed to be $\boldsymbol{\mu}$.

To get the above result, we use the following property of traces:

$$\sum_{i=1}^{n} \boldsymbol{x}_i' \boldsymbol{A} \boldsymbol{x}_i = tr(\boldsymbol{XAX}') = tr(\boldsymbol{X}'\boldsymbol{XA})$$

Multivariate Models: MVN – P($\boldsymbol{\Sigma} \mid \mathbf{y}_1$,..., $\mathbf{y}_N$, $\boldsymbol{\mu}$)

• Now, we can derive the conditional posterior for $\boldsymbol{\Sigma}$:

$$p(\boldsymbol{\Sigma} \mid \boldsymbol{y}_1,\ldots,\boldsymbol{y}_N,\boldsymbol{\mu}) \propto \left\{ |\boldsymbol{\Sigma}|^{-\frac{N}{2}} \, e^{-\frac{1}{2}tr(\boldsymbol{S}_{\boldsymbol{\mu}}\boldsymbol{\Sigma}^{-1})} \right\} \times \left\{ |\boldsymbol{\Sigma}|^{-\frac{1}{2}(v_0+k+1)} \times e^{-\frac{1}{2}tr(\boldsymbol{S}_0\boldsymbol{\Sigma}^{-1})} \right\}$$
$$= |\boldsymbol{\Sigma}|^{-\frac{1}{2}(N+v_0+k+1)} \, e^{-\frac{1}{2}tr([\boldsymbol{S}_0+\boldsymbol{S}_{\boldsymbol{\mu}}]\boldsymbol{\Sigma}^{-1})}.$$

which looks like a IW(v_N, $\mathbf{S_N}^{-1}$), where $v_N = N + v_0$ and $\mathbf{S_N} = \mathbf{S_0} + \boldsymbol{S}_{\boldsymbol{\mu}}$. Similar to the results for $\boldsymbol{\mu}$, the posterior combines prior and data information. Then,

$$E[\boldsymbol{\Sigma} \mid \mathbf{y}_1 ,\ldots, \mathbf{y}_N, \boldsymbol{\mu}] = (\mathbf{S_0} + \boldsymbol{S}_{\boldsymbol{\mu}})/(N+v_0-k-1)$$

• We got the full conditional posteriors of $\boldsymbol{\mu}$ and $\boldsymbol{\Sigma}$. Later, we will go over a numerical method (Gibbs sampler) that easily estimates the joint density.

Multivariate Models: Alternative Prior for Σ

• Barnard, McCulloch and Meng (2000) present a workaround to avoid the problem of using a vague IW prior. They propose an alternative to the IW prior, based on a decomposition of $\boldsymbol{\Sigma}$:

$$\boldsymbol{\Sigma} = \text{diag}(\boldsymbol{S})\boldsymbol{R}\,\text{diag}(\boldsymbol{S}),$$

where $\boldsymbol{S}$ is the $k\times 1$ vector of SDs, diag($\boldsymbol{S}$) is the diagonal matrix with diagonal elements $\boldsymbol{S}$, and $\boldsymbol{R}$ is the $k\times k$ correlation matrix.

• A hierchical prior structure is used:

$$p(\boldsymbol{S},\boldsymbol{R}) = p(\boldsymbol{R}|\boldsymbol{S})\, p(\boldsymbol{S}).$$

• Then, impose a prior for $\boldsymbol{S}$, for example, an independent log normal –i.e., log($\boldsymbol{S}$)~N($\boldsymbol{\xi}$, $\boldsymbol{\Lambda}$)– and impose a diffuse prior on $\boldsymbol{R}$, for example, a uniform.

Bayesian vs. Classical: Review

• The goal of a classical statistician is getting a point estimate for the unknown fixed population parameter θ, say using OLS.

These point estimates will be used to test hypothesis about a model, make predictions and/or to make decisions –say, consumer choices, monetary policy, portfolio allocation, etc.

• In the Bayesian world, θ is unknown, but it is not fixed. A Bayesian statistician is interested in a distribution, the posterior distribution, P(θ|y); not a point estimate.

"*Estimation*:" Examination of the characteristics of P(θ|y):

- Moments (mean, variance, and other moments)
- Intervals containing specified probabilities

Bayesian vs. Classical: Review

• The posterior distribution will be incorporated in tests of hypothesis and/or decisions.

In general, a Bayesian statistician does not separate the problem of how to estimate parameters from how to use the estimates.

• In practice, classical and Bayesian inferences are often very similar.

• There are theoretical results under which both worlds produce the same results. For example, in large samples, under a uniform prior, the posterior mean will be approximately equal to the MLE.

Bayesian vs. Classical: Interpretation

• In practice, classical and Bayesian inferences and concepts are often similar. But, they have different interpretations.

• Likelihood function

- In classical statistics, the likelihood is the density of the observed data conditioned on the parameters.
  - Inference based on the likelihood is usually "maximum likelihood."
- In Bayesian statistics, the likelihood is a function of the parameters and the data that forms the basis for inference – not really a probability distribution.
  - The likelihood embodies the current information about the parameters and the data.

Bayesian vs. Classical: Interpretation

• Confidence Intervals (C.I.)

- In a regular parametric model, the classical C.I. around MLEs –for example, $b \pm 1.96$ sqrt$\{s^2 (\mathbf{X}' \mathbf{X})^{-1}\}$-- has the property that whatever the true value of the parameter is, with probability 0.95 the confidence interval covers the true value, β.

- This classical C.I. can also be also interpreted as an approximate Bayesian probability credibility interval. That is, conditional on the data and given a range of prior distributions, the posterior probability that the parameter lies in the C.I. is approximately 0.95.

• The formal statement of this remarkable result is known as the Bernstein-Von Mises theorem.

Bayesian vs. Classical: Bernstein-Von Mises Theorem

• Bernstein-Von Mises theorem:

- The posterior distribution converges to normal with covariance matrix equal to $1/T$ times the information matrix --same as classical MLE.

Note: The distribution that is converging is the posterior, not the sampling distribution of the estimator of the posterior mean.

- The posterior mean (empirical) converges to the mode of the likelihood function --same as the MLE. A proper prior disappears asymptotically.

- Asymptotic sampling distribution of the posterior mean is the same as that of the MLE.

Bayesian vs. Classical: Bernstein-Von Mises Theorem

• That is, in large samples, the choice of a prior distribution is not important in the sense that the information in the prior distribution gets dominated by the sample information.

That is, unless your prior beliefs are so strong that they cannot be overturned by evidence, at some point the evidence in the data outweights any prior beliefs you might have started out with.

• There are important cases where this result does not hold, typically when convergence to the limit distribution is not uniform, such as unit roots. In these cases, there are differences between both methods.

Linear Model: Classical Setup

• Consider the simple linear model:

$$y_t = X_t\beta + \varepsilon_t, \qquad \varepsilon_t \mid X_t, y_t \sim N(0, \sigma^2)$$

To simplify derivations, assume **X** is fixed. We want to estimate β.

• Classical OLS (MLE=MM) estimation

$$\mathbf{b} = (\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'\mathbf{y} \text{ and } \quad \mathbf{b} \mid \mathbf{X} \sim N(\beta, \sigma^2 (\mathbf{X}'\mathbf{X})^{-1})$$

- The estimate of σ^2 is $s^2 = (\mathbf{y} - \mathbf{X}\mathbf{b})'(\mathbf{y} - \mathbf{X}\mathbf{b})/(T-k)$

- The uncertainty about **b** is summarized by the regression coefficients standard errors –i.e., the diagonal of the matrix: $\text{Var}(\mathbf{b} \mid \mathbf{X}) = s^2(\mathbf{X}'\mathbf{X})^{-1}$.

• Testing: If V_{kk} is the k-th diagonal element of $\text{Var}(\mathbf{b} \mid \mathbf{X})$, then

$(b_k' - 0)/(sV_{kk}^{1/2}) = t_{T-k}$ --the basis for hypothesis tests.

Linear Model: Bayesian Setup

• For the normal linear model, we assume $f(y_i | \mu_i, \sigma^2)$:

$$y_t \sim N(\mu_t, \sigma^2) \quad \text{for } t = 1,..., T$$

where $\mu_t = \beta_0 + \beta_1 X_{1t} + \ldots + \beta_k X_{kt} = \mathbf{X}_t \boldsymbol{\beta}$

Bayesian goal: Get the posterior distribution of the parameters $(\boldsymbol{\beta}, \sigma^2)$.

• By Bayes' Theorem, we know that this is simply:

$$f(\boldsymbol{\beta}, \sigma^2 | \mathbf{y}, \mathbf{X}) \propto \prod_i f(y_i | \mu_i, \sigma^2) \times f(\boldsymbol{\beta}, \sigma^2)$$

=> we need to choose a prior distribution for $f(\boldsymbol{\beta}, \sigma^2)$.

• To simplify derivations, we assume **X** is fixed.

Linear Model: Likelihood

• In our linear model $y_t = \mathbf{X}_t \beta + \varepsilon_t$, with $\varepsilon_t \sim$ *i.i.d.* $N(0,\sigma^2)$. Then,

$$f(\mathbf{y} | \mathbf{X}, \boldsymbol{\beta}, \sigma^2) = (2\pi\sigma^2)^{-T/2} \exp\{-\frac{(\boldsymbol{y} - \boldsymbol{X\beta})'(\boldsymbol{y} - \boldsymbol{X\beta})}{2\sigma^2}\}$$

$$\propto (\frac{1}{\sigma^2})^{-T/2} \exp\{-\frac{1}{2\sigma^2}[\boldsymbol{y}'\boldsymbol{y} - 2\boldsymbol{\beta}'\boldsymbol{X}'\boldsymbol{y} + \boldsymbol{\beta}'\mathbf{X}'\mathbf{X}\boldsymbol{\beta}]\}$$

• Recall that we can write: $\mathbf{y} - \mathbf{X}\beta = (\mathbf{y} - \mathbf{X}b) - \mathbf{X}(\beta - b)$

$\Rightarrow$ TSS $= (\mathbf{y}-\mathbf{X}b)'(\mathbf{y}-\mathbf{X}b) + (\beta - b)'\mathbf{X}'\mathbf{X}(\beta - b) - 2(\beta - b)'\mathbf{X}'(\mathbf{y}-\mathbf{X}\beta)$
$= \upsilon s^2 + (\beta - b)'\mathbf{X}'\mathbf{X}(\beta - b)$

where $s^2 = RSS/(T-k) = (\mathbf{y}-\mathbf{X}b)'(\mathbf{y}-\mathbf{X}b)/(T-k)$; and $\upsilon = (T-k)$.

Linear Model: Likelihood

• The likelihood can be factorized as:

$$f(\mathbf{y} \mid \mathbf{X}, \boldsymbol{\beta}, \sigma^2) = (1/2\pi)^{T/2}\left(\frac{1}{\sigma^2}\right)^{k/2} \exp\{-\frac{1}{2\sigma^2}(\boldsymbol{\beta}-\boldsymbol{b})'\mathbf{X}'\mathbf{X}(\boldsymbol{\beta}-\boldsymbol{b})\} \text{x} \left(\frac{1}{\sigma^2}\right)^{\upsilon/2} \exp\{-\frac{\upsilon s^2}{2\sigma^2}\}$$

$$\propto (h)^{k/2} \exp\{-\frac{h}{2}(\boldsymbol{\beta}-\boldsymbol{b})'\mathbf{X}'\mathbf{X}(\boldsymbol{\beta}-\boldsymbol{b})\} \text{ x } (h)^{\upsilon/2} \exp\{-\frac{h\upsilon s^2}{2}\}$$

where $h=1/\sigma^2$.

• The likelihood can be written as a product of a normal and a density of form $f(\theta) = \varkappa\, \theta^{-\lambda} \exp\{-\lambda/\theta\}$. This is an *inverted gamma* (IG) distribution.

Linear Model: Prior Distribution for β

• The likelihood points towards a MVN for β as conjugate. Then, $f(\beta) \sim N(\boldsymbol{m}, \boldsymbol{\Sigma})$:

$$f(\boldsymbol{\beta}) = \frac{1}{(2\pi)^{-1/2}} |\boldsymbol{\Sigma}|^{-1/2} \exp\{-\frac{1}{2}(\boldsymbol{\beta}-\boldsymbol{m})'\boldsymbol{\Sigma}^{-1}(\boldsymbol{\beta}-\boldsymbol{m})\}$$

- $\boldsymbol{m}$ is our *best guess* for β, before seeing **y** and **X**.

- $\boldsymbol{\Sigma}$ measures the confidence in our guess. It is common to relate $\boldsymbol{\Sigma}$ to σ^2, say $\boldsymbol{\Sigma} = \{\sigma^2\mathbf{Q}\}$.

• This assumption for $f(\beta)$ gives us some flexibility: Depending on $\boldsymbol{\Sigma}$, this prior can be informative (small $\boldsymbol{\Sigma}$) or diffuse (big $\boldsymbol{\Sigma}$). In addition, it is a conjugate prior.

• But, we could have assumed a different prior distribution, say a uniform. Remember, priors are the Achilles heel of Bayesian statistics.

Linear Model: Prior Distribution for *h*

• The usual prior for σ^2 is the IG. Then, $h=1/\sigma^2$ is distributed as $\Gamma(\alpha_0,\lambda_0)$:

$$f(x=\sigma^{-2};\alpha_0,\lambda_0)=\frac{\lambda_0^{\alpha_0}}{\Gamma(\alpha_0)}\ x^{\alpha_0-1}\ e^{-\lambda_0 x} \qquad x>0.$$

• Usual values for (α,λ): $\alpha_0=T/2$ and $\lambda_0=1/(2\eta^2)=\Phi/2$, where η^2 is related to the variance of the T $N(0, \eta^2)$ variables we are implicitly adding. You may recognize this parameterization of the gamma as a non-central χ_T^2 distribution. Then,

$$f(\sigma^{-2})=\frac{(\Phi/2)^{T/2}}{\Gamma(T/2)}\ (\sigma^{-2})^{(\frac{T}{2}-1)}\ e^{-(\Phi/2)\sigma^{-2}}$$

Linear Model: Joint Prior Distribution for θ

• We have $\theta=(\beta,\sigma^2)$. We need the joint prior $P(\theta)$ along with the likelihood, $P(y|\theta)$, to obtain the posterior $P(\theta|y)$.

In this case, we can write $P(\theta) = P(\beta|\sigma^{-2})\ P(\sigma^{-2})$, ignoring constants:

$$f(\beta,\sigma^{-2})\propto e^{\{-\frac{1}{2}(\beta-m)'\Sigma^{-1}(\beta-m)\}}\times\ h^{\alpha_0-1}\ e^{-\lambda_0 h}$$

Then, we write the posterior as usual: $P(\theta|y) \propto P(y|\theta)\ P(\theta)$.

Linear Model: Assumptions

• So far, we have made the following assumptions:

- Data is *i.i.d.* Normal: $y_t \sim N(\mu_t, \sigma^2)$ for $t = 1,..., T$
- DGP for μ_t is known: $\mu_t = \beta_0 + \beta_1 X_{1t} + \ldots + \beta_k X_{kt} = \mathbf{X}_t \beta$
- $\mathbf{X}$ is fixed.
- Prior distributions: $h=1/\sigma^2 \sim \Gamma(\alpha_0,\lambda_0)$ & $\beta \sim N(\boldsymbol{m}, \boldsymbol{\Sigma})$.

Note: A subtle point regarding this Bayesian regression setup. A full Bayesian model includes a distribution for $\mathbf{X}$, $f(\mathbf{X}|\Psi)$. Thus, we have a joint likelihood $f(\mathbf{y},\mathbf{X}|\Psi,\beta,\sigma)$ and joint prior $f(\Psi,\beta,\sigma)$.

A key assumption of this linear model is that $f(\mathbf{y}|\mathbf{X}, \beta, \sigma)$ and $f(\boldsymbol{X}|\Psi)$ are independent in their priors. Then, the posterior factors into:

$$f(\Psi, \beta, \sigma|\mathbf{y},\mathbf{X}) = f(\beta, \sigma|\mathbf{y},\mathbf{X})\, f(\Psi|\mathbf{y},\mathbf{X})$$
$$\Rightarrow f(\beta, \sigma|\mathbf{y},\mathbf{X}) \propto f(\beta, \sigma)\, f(\mathbf{y}|\ \beta, \sigma,\mathbf{X})$$

Linear Model: Joint Posterior Distribution for θ

• Now, we are ready to write posterior as usual: Likelihood x Prior.

$$f(\theta|\mathbf{y},\mathbf{X}) \propto h^{k/2} e^{\{-\frac{h}{2}(\beta-b)'\mathbf{X}'\mathbf{X}(\beta-b)\}} \times h^{\upsilon/2} e^{-\frac{h\upsilon s^2}{2}} \times e^{\{-\frac{1}{2}(\beta-m)'\Sigma^{-1}(\beta-m)\}} \times h^{\alpha_0-1} e^{-\lambda_0 h}$$

• Then, we need the likelihood and the prior:

$$f(\theta|\mathbf{y},\mathbf{X}) \propto h^{k/2} e^{-\frac{1}{2}\{h(\beta-b)'\mathbf{X}'\mathbf{X}(\beta-b)+(\beta-m)'\boldsymbol{\Sigma}^{-1}(\beta-m)\}} \times h^{(\upsilon+\alpha_0)/2-1} e^{-h\{\frac{\upsilon s^2}{2}+\lambda_0\}}$$

which we do not recognize as a standard distribution –i.e., a "*complicated posterior.*" This posterior does not lead to convenient expressions for the marginals of β and h.

Linear Model: Conditional Posteriors

• When facing complicated posteriors, we usually rely on numerical methods to say something about $P(\boldsymbol{\theta}|\mathbf{y})$. A popular numerical method, the Gibbs Sampler, uses the conditional posteriors.

• In our setting, it is easy to get the analytical expressions for the *conditional posteriors* $f(\boldsymbol{\beta}|\mathbf{y},\mathbf{X})$ and $f(h|\mathbf{y},\mathbf{X})$.

• First, we derive $f(\boldsymbol{\beta}|\mathbf{y},\mathbf{X}, h)$. Again, to get the conditional posteriors, we use: Likelihood x Prior, but with a conditional prior $f(\boldsymbol{\beta}|h)$.

$$f(\boldsymbol{\beta}|\mathbf{y},\mathbf{X},\sigma^2) \propto h^{T/2}\exp\{-\frac{1}{2\sigma^2}[\boldsymbol{y}'\boldsymbol{y}-2\boldsymbol{\beta}'\boldsymbol{X}'\boldsymbol{y}+\boldsymbol{\beta}'\mathbf{X}'\mathbf{X}\boldsymbol{\beta}]\}$$
$$\times \exp\{-\frac{1}{2}(\boldsymbol{\beta}-\boldsymbol{m})'\boldsymbol{\Sigma}^{-1}(\boldsymbol{\beta}-\boldsymbol{m})\}$$

Linear Model: Conditional Posterior $f(\boldsymbol{\beta}|\mathbf{y},\mathbf{X}, h)$

• A little bit of algebra delivers:

$$f(\boldsymbol{\beta}|\mathbf{y},\mathbf{X},\sigma^2) \propto h^{T/2}\exp\{-[\frac{\boldsymbol{y}'\boldsymbol{y}}{2\sigma^2}-\boldsymbol{\beta}'(\frac{\boldsymbol{X}'\boldsymbol{y}}{\sigma^2}+\boldsymbol{\Sigma}^{-1}\boldsymbol{m})+\frac{1}{2}\boldsymbol{\beta}'(\frac{\boldsymbol{X}'\boldsymbol{X}}{\sigma^2}+\boldsymbol{\Sigma}^{-1})\boldsymbol{\beta}]\}$$
$$\propto \exp\{\boldsymbol{\beta}'(\frac{\boldsymbol{X}'\boldsymbol{y}}{\sigma^2}+\boldsymbol{\Sigma}^{-1}\boldsymbol{m})-\frac{1}{2}\boldsymbol{\beta}'(\frac{\boldsymbol{X}'\boldsymbol{X}}{\sigma^2}+\boldsymbol{\Sigma}^{-1})\boldsymbol{\beta}]\}$$

which from our previous MVN model, we recognize as proportional to an MVN with:

$$\boldsymbol{\Sigma}_n = Var[\boldsymbol{\beta}|\mathbf{y},\mathbf{X},\sigma^2] = \left(\frac{\boldsymbol{X}'\boldsymbol{X}}{\sigma^2}+\boldsymbol{\Sigma}^{-1}\right)^{-1}$$

$$\boldsymbol{m}_n = \boldsymbol{E}[\boldsymbol{\beta}|\mathbf{y},\mathbf{X},\sigma^2] = \boldsymbol{\Sigma}_n(\frac{\boldsymbol{X}'\boldsymbol{y}}{\sigma^2}+\boldsymbol{\Sigma}^{-1}\boldsymbol{m}) = \boldsymbol{\Sigma}_n(\frac{\boldsymbol{X}'\boldsymbol{X}\boldsymbol{b}}{\sigma^2}+\boldsymbol{\Sigma}^{-1}\boldsymbol{m})$$

• That is, $f(\boldsymbol{\beta}|\mathbf{y},\mathbf{X}, h)$ is $N_k(\mathbf{m}_n,\boldsymbol{\Sigma}_n)$.

Linear Model: Conditional Posterior $f(\beta \mid y, X, h)$

$$f(\boldsymbol{\beta} \mid \boldsymbol{y}, \boldsymbol{X}, \boldsymbol{\sigma}^2) \propto \boldsymbol{h}^{1/2} \mid \mathbf{A}^{-1} + \boldsymbol{h}(\mathbf{X}'\mathbf{X}) \mid^{-1/2} \exp\{-\frac{\boldsymbol{h}}{2}(\boldsymbol{\beta} - \boldsymbol{m}_n)'(\boldsymbol{h}(\mathbf{X}'\mathbf{X}) + \boldsymbol{A}^{-1})(\boldsymbol{\beta} - \boldsymbol{m}_n)\}$$

where $\boldsymbol{m}_n = (\boldsymbol{\Sigma}^{-1} + h(\mathbf{X'X}))^{-1}(\boldsymbol{\Sigma}^{-1} m + h\,(\mathbf{X'X})\,\mathbf{b})$.

In other words, the pdf of β, conditioning on the data, is normal with mean m_n and variance matrix $(h\,(\mathbf{X'X}) + \boldsymbol{\Sigma}^{-1})^{-1}$.

• Similar work for $f(h \mid \mathbf{y}, \mathbf{X}, \beta)$ delivers a gamma distribution. (Do it!).

Linear Model: Bayesian Learning

• The mean $\boldsymbol{m}_n$ takes into account the data (**X** and **y**) and our prior distribution. It is a weighted average of our prior m and **b** (OLS):

$$\boldsymbol{m}_n = (\boldsymbol{\Sigma}^{-1} + h(\mathbf{X'X}))^{-1}(\boldsymbol{\Sigma}^{-1} \boldsymbol{m} + h\,(\mathbf{X'X})\,\mathbf{b}).$$

• Bayesian learning: We combine prior information ($\boldsymbol{\Sigma}$, $\boldsymbol{m}$) with the data (**X**, **b**). As more information is known, we update our beliefs!

• If our prior distribution is very diffuse (say, the elements of $\boldsymbol{\Sigma}$ are large), our prior, $\boldsymbol{m}$, will have a lower weight.

As prior becomes more diffuse, $\boldsymbol{m}_n \rightarrow \mathbf{b}$ (prior info is worthless).
As prior becomes more certain, $\boldsymbol{m}_n \rightarrow \boldsymbol{m}$ (prior dominates).

• Note that with a diffuse prior, we can say now:
"*Having seen the data, there is a 95% probability that β is in the interval* $\mathbf{b} \pm 1.96\ \text{sqrt}\{\sigma^2\,(\mathbf{X'}\,\mathbf{X})^{-1}\}$."

Linear Model: Remarks

• We get a normal conditional posterior, a nice recognizable distribution, because we made clever distributional assumtions: We assumed an *i.i.d.* normal distribution for $(\mathbf{y}|\mathbf{X}, \sigma^2)$, and for β we chose a normal prior distribution ($\Rightarrow$ the normal (*conjugate*) prior was a very convenient choice).

• We can do similar calculations when we impose another prior. But, the results would change.

• If not exact results are possible, numerical solutions will be used.

Linear Model: Remarks

• When we setup our probability model, we are implicitly conditioning on a model, call it H, which represents our beliefs about the data-generating process. Thus,

$$f(\beta,\sigma|\mathbf{y},\mathbf{X},H) \propto f(\beta,\sigma|H)\, f(\mathbf{y}|\beta,\sigma,\mathbf{X},H)$$

It is important to keep in mind that our inferences are dependent on H.

• This is also true for the classical perspective, where results can be dependent on the choice of likelihood function, covariates, etc.

Linear Model: Interpretation of Priors

• Suppose we had an earlier sample, {**y**′,**X**′}, of *T'* observations, which are independent of the current sample, {**y**,**X**}.

• The OLS estimate based on all information available is:

$$b^* = \left(\sum_{t=1}^{T} x_t x_t' + \sum_{t=1}^{T'} x_t' x_t' \right)^{-1} \left(\sum_{t=1}^{T} x_t y_t' + \sum_{t=1}^{T'} x_t' y_t' \right)$$

and the variance is

$$Var[b^*] = \sigma^2 \left(\sum_{t=1}^{T} x_t x_t' + \sum_{t=1}^{T'} x_t' x_t' \right)^{-1}$$

• Let *m* be the OLS estimate based on the prior sample {**y**′,**X**′}:

$$m = \left(\sum_{t=1}^{T'} x_t' x_t' \right)^{-1} \left(\sum_{t=1}^{T'} x_t' y_t' \right) \quad \text{and} \quad Var[m] = \sigma^2 \left(\sum_{t=1}^{T'} x_t' x_t' \right)^{-1} = \sigma^2 A$$

Linear Model: Interpretation of Priors

• Then,

$$b^* = \left(\sum_{t=1}^{T} x_t x_t' + \sum_{t=1}^{T'} x_t' x_t' \right)^{-1} \left(\sum_{t=1}^{T} x_t y_t' + \sum_{t=1}^{T'} x_t' y_t' \right)$$

$$= \left(\sum_{t=1}^{T} x_t x_t' + A^{-1} \right)^{-1} \left(\sum_{t=1}^{T} x_t y_t' + A^{-1} m \right)$$

• This is the same formula for the posterior mean *m**.

• Thus, the question is what priors should we use?

• There are a lot of publications, using the same data. To form priors, we cannot use the results of previous research, if we are not going to use a correlated sample!

The Linear Regression Model – Example 1

• Let's go over the multivariate linear model. Now, we impose a diffuse uniform prior for $\theta=(\beta, h)$. Say, $f(\beta, h) \propto h^{-1}$.

Now, $f(\theta \mid y, \mathbf{X}) \propto h^{T/2} \exp\{-\frac{h}{2}[\upsilon s^2 + (\beta - b)'\mathbf{X}'\mathbf{X}(\beta - b)]\} \times h^{-1}$

• If we are interested in β, we can integrate out the nuisance parameter h to get the marginal posterior of $f(\beta \mid \mathbf{y}, \mathbf{X})$:

$$f(\beta \mid y, \mathbf{X}) \propto \int h^{T/2-1} \exp\{-\frac{h}{2}[\upsilon s^2 + (\beta - b)'\mathbf{X}'\mathbf{X}(\beta - b)]\}\, dh$$
$$\propto [1 + \frac{(\beta - b)'\mathbf{X}'\mathbf{X}(\beta - b)}{\upsilon s^2}]^{-T/2}$$

where we use the following integral result ($\Gamma(s,x)$: the incomplete Γ):

$$\int x^a \exp\{-xb\}\, dx = b^{-a-1}\, [\Gamma(a+1) - \Gamma(a+1, b)]$$

The Linear Regression Model – Example 1

• The marginal posterior $f(\beta \mid y, \mathbf{X}) \propto [1 + \frac{(\beta - b)'\mathbf{X}'\mathbf{X}(\beta - b)}{\upsilon s^2}]^{-T/2}$

is the kernel of a multivariate *t distribution*. That is,

$$f(\beta \mid y, \mathbf{X}) = t_v(\beta \mid b, s^2(\mathbf{X}'\mathbf{X})^{-1})$$

Note: This is the equivalent to the repeated sample distribution of **b**.

• Similarly, we can get $f(h \mid \mathbf{y}, \mathbf{X})$ by integrating out β:

$$f(h \mid y, \mathbf{X}) \propto \int h^{T/2-1} \exp\{-\frac{h}{2}[\upsilon s^2 + (\beta - b)'\mathbf{X}'\mathbf{X}(\beta - b)]\}\, d\beta$$
$$\propto h^{T/2-1} \exp\{-\frac{h}{2}\upsilon s^2\} \int \exp\{-\frac{h}{2}[(\beta - b)'\mathbf{X}'\mathbf{X}(\beta - b)]\}\, d\beta$$
$$\propto h^{\upsilon/2-1} \exp\{-\frac{h}{2}\upsilon s^2\}$$

which is the kernel of a $\Gamma(\alpha,\lambda)$ distribution, with $\alpha = \upsilon/2$ and $\lambda = \upsilon s^2/2$.

The Linear Regression Model – Example 1

• The mean of a gamma distribution is α/λ. Then,

$$E[h|\mathbf{y},\mathbf{X}] = [\upsilon/2]/[\upsilon s^2/2]=1/s^2.$$

• Now, we interpret the prior $f(\boldsymbol{\beta}, h) \propto h^{-1}$ as non-informative: The marginal posterior distributions have properties closely resembling the corresponding repeated sample distributions.

The Linear Regression Model – Example 2

• Let's go over the multivariate linear model. Now, we impose a diffuse uniform prior for β and an inverse gamma for σ^2.

Likelihood

$$L(\beta,\sigma^2|y,X)=[2\pi\sigma^2]^{-n/2}e^{-[(1/(2\sigma^2))(y-X\beta)'(y-X\beta)]}$$

Transformation using d=(N-K) and $s^2 = (1/d)(\mathbf{y} - \mathbf{Xb})'(\mathbf{y} - \mathbf{Xb})$

$$\left(-\frac{1}{2\sigma^2}\right)(\mathbf{y} - \mathbf{X\beta})'(\mathbf{y} - \mathbf{X\beta}) = \left(-\frac{1}{2}ds^2\right)\left(\frac{1}{\sigma^2}\right) - \frac{1}{2}(\boldsymbol{\beta} - \mathbf{b})'\left(\frac{1}{\sigma^2}\mathbf{X}'\mathbf{X}\right)(\boldsymbol{\beta} - \mathbf{b})$$

Diffuse uniform prior for $\boldsymbol{\beta}$, conjugate gamma prior for σ^2

Joint Posterior

$$f(\boldsymbol{\beta}, \sigma^2 \mid \mathbf{y}, \mathbf{X}) \propto \frac{[ds^2]^{v+2}}{\Gamma(d+2)}\left[\frac{1}{\sigma^2}\right]^{d+1} e^{-ds^2(1/\sigma^2)}[2\pi]^{-K/2} \mid \sigma^2(\mathbf{X}'\mathbf{X})^{-1}|^{-1/2}$$

$$\times \exp\{-(1/2)(\boldsymbol{\beta} - \mathbf{b})'[\sigma^2(\mathbf{X}'\mathbf{X})^{-1}]^{-1}(\boldsymbol{\beta} - \mathbf{b})\}$$

The Linear Regression Model – Example 2

• From the joint posterior, we can get the marginal posterior for β.

After integrating σ^2 out of the joint posterior:

$$f(\boldsymbol{\beta} \mid \mathbf{y}, \mathbf{X}) \propto \frac{[ds^2]^{\nu+2}\Gamma(d + K/2)}{\Gamma(d+2)}[2\pi]^{-K/2} \mid \mathbf{X}'\mathbf{X} \mid^{-1/2} \over [ds^2 + \tfrac{1}{2}(\boldsymbol{\beta} - \mathbf{b})'\mathbf{X}'\mathbf{X}(\boldsymbol{\beta} - \mathbf{b})]^{d+K/2}}.$$

Multivariate t with mean $\mathbf{b}$ and variance matrix $\frac{\text{n-K}}{n-K-2}[s^2(\mathbf{X}'\mathbf{X})^{-1}]$

The Bayesian 'estimator' equals the MLE. Of course; the prior was noninformative. The only information available is in the likelihood.

Presentation of Results

• $P(\theta|y)$ is a pdf. For the simple case, the one parameter θ, it can be graphed. But, if $\boldsymbol{\theta}$ is a vector, the multivariate pdf cannot be graphed.

• It is common to present measures analogous to classical point estimates and CIs. For example:

(1) $E(\theta|y) = \int \theta\, p(\theta|y)\, d\theta$ -- posterior mean

(2) $Var(\theta|y) = E(\theta^2|y) - \{E(\theta|y)\}^2$ -- posterior variance

(3) $p(k_1>\theta>k_2|y) = \int_{k1>\theta>k2} p(\theta|y)\, d\theta$ -- C.I.

• In many cases, it is not possible to evaluate these integrals analytically. Typically, we rely on numerical methods to approximate the integral as a (weighted) sum:

$$I = \int_a^b f(\theta) d\theta \approx \sum_{i=1}^{n} w_i f(\theta_i)$$

Presentation of Results: MC Integration

• In the Math Review, we covered different numerical integration methods (trapezoid rule, Gaussin quadrature, etc), where we pick the θ_i's and the w_i's in some fixed (deterministic) way.

• In this section, we will use Monte Carlo (MC) methods to integrate. MC Integration is based on selecting θ_i's randomly (from some pdf).

Example: We can compute the expected value of a Beta(3,3):

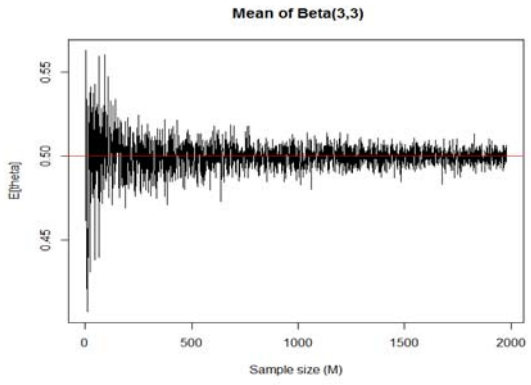
$$E(\theta) = \int \frac{\Gamma(\alpha+\beta)}{\Gamma(\alpha)\Gamma(\beta)} \theta^{\alpha-1}(1-\theta)^{\beta-1} d\theta = \frac{\alpha}{\alpha+\beta} \quad \Rightarrow E(\theta) = \frac{1}{2}$$

or via Monte Carlo methods (R Code):

```
M <- 10000
beta.sims <- rbeta(M, 3, 3)
sum(beta.sims)/M
[1] 0.4981763
```

Presentation of Results: MC Integration

Q: What is the advantage of MC methods? The LLN tells us that the MC approximation is a consistent (simulation) estimator of the value population value E[θ]. The following traceplot illustrates the point:



Note: The CLT can be used too!

MC Integration

• Obviously, we will not use MC methods to get the mean and variance of a Beta(3,3)! It will be used when we face integrals that involve complicated posteriors.

Example: Suppose Y ~ N(θ,1) and we have a Cauchy (0,1) prior. That is, θ ~ Ca(0,1). Then,

$$p(\theta \mid y) = \frac{1}{\sqrt{2\pi}} e^{-\frac{1}{2}(y-\theta)^2} \times \frac{1}{\pi(1+\theta^2)} = \frac{1}{\pi^{3/2}\sqrt{2}} \frac{e^{-\frac{1}{2}(y-\theta)^2}}{(1+\theta^2)}$$

$$\propto \frac{e^{-\frac{1}{2}(y-\theta)^2}}{(1+\theta^2)}$$

which we do not recognize as a known distribution. Suppose we are interested in E[h(θ|y)]. MC Integration can compute this integral.

MC Integration: Plug-in estimator

• Idea: We start with a posterior, P(θ|y): $\pi(\theta) = P(y|\theta)\, P(\theta)/P(y)$.

We want to get moments of some function of θ, say

$$E_\pi[h(\theta)] = \int h(\theta)\pi(\theta)d\theta.$$

• If $\theta^{(M)} = \{\theta^1, \theta^2, ..., \theta^M\}$ is an *i.i.d.* random sample from π(θ), then

$$\bar{h}_{MC} = \frac{1}{M}\sum_{m=1}^{M} h(\theta^m) \to E_\pi[h(\theta)] \qquad \text{as } M \to \infty.$$

• The h_{MC} average over θ is called the *plug-in estimator* for $E_\pi[h(\theta)]$. Note that when $h(\theta) = \theta$, we get the mean; when $h(\theta) = [\theta - E(\theta)]^2$, we get the variance, etc.

• Using the plug-in estimator, we can approximate almost any aspect of the posterior to arbitrary accuracy, with large enough *M*.

MC Integration: MC Standard Errors

• We can get MC standard errors to evaluate the accuracy of approximations to the posterior mean.

• Let $\bar{\theta}$ be the sample mean of the M MC samples. Then, the CLT approximate $\bar{\theta} \sim N(\theta, Var[\theta|y]/M)$. We approximate the $Var[\theta|y]/T$:

$$\hat{\sigma}^2 = \frac{1}{M-1}\sum_{m=1}^{M}(\theta^m - \bar{\theta})^2 \rightarrow Var[\theta|y_1,...,y_T]$$

Then, MC SE = $\sqrt{\hat{\sigma}^2/M}$. We can select M to give us a desired precision relative to the posterior moment we are interested.

MC Integration: MC Standard Errors

Example: We generate a MC sample of size $M = 200$ with $\bar{\theta} = .78$ and $\hat{\sigma}^2 = 0.35$. Then, the approximate MC SE is given by:

MC SE = sqrt[0.35/200] = 0.0418.

We can do a 95% C.I. for the posterior mean of θ:

$[.78 \pm 1.96 * .0418]$

• If we want the difference between $E[\theta|y]$ and its MC estimate to be less than 0.005 with high probability, we need to increase M such that

$1.96*$ sqrt$[0.35/M] < .005 \quad \Rightarrow M > 53{,}782$

Note: The plug-in estimator may have a large variance (MC error). In these cases, a very large M is needed.

MC Integration: Sampling Problems

• MC integration relies on being able to draw from $P(\theta|y)$. To do this, we need $P(\theta|y)$ to be a pdf that is represented by a standard library function, which allows us to get draws, say *rnorm* or *rbeta* in R.

• Q: What happens when $P(\theta|y)$ is not in the library?

A: There are several methods to work around this situation. For example, the method of inversion (based on the probability integral transformation) and the usual Bayesian tool, Markov chain Monte Carlo, or MCMC (coming soon).

• There are also MC methods to calculate posterior quantities of interest without the need to draw directly from the posterior. For example, *importance sampling* (IS).

MC Integration: Importance Sampling (IS)

• We want to calculate the (posterior) expectation:

$$E_\pi[h(\theta)] = \int h(\theta)\pi(\theta)d\theta.$$

It can be easier to compute this integral by sampling from another pdf, q(.), an *importance function,* also called a *proposal function.* Then,

$$E_\pi[h(\theta)] = \int \frac{h(\theta)\pi(\theta)}{q(\theta)} q(\theta)d\theta.$$

If $\theta^{(M)}=\{\theta^1, \theta^2, ..., \theta^M\}$ is a random sample from $q(\theta)$, then

$$\bar{h}_{IS} = \frac{1}{M}\sum_{m=1}^{M}\frac{\pi(\theta^m)h(\theta^m)}{q(\theta^m)} = \frac{1}{M}\sum_{m=1}^{M} w(\theta^m)h(\theta^m) \to E_\pi[h(\theta)] \qquad \text{as } M \to \infty.$$

where $w(\theta^m)=\pi(\theta^m)/q(\theta^m)$ is called *importance weight.* These weights give more *importance* to some θ^m than to others!

• The *IS estimator* –i.e., the weighted sum– approximates $E_\pi[h(\theta)]$.

MC Integration: IS - Remarks

• In principle, any proposal $q(\theta)$ can be used. But, some $q(\theta)$' s are more efficient than others. Choosing $q(\theta)$ close to the target, $\pi(\theta)$, works well (may be *"optimal,"* by reducing the variance of the MC estimator).

This variance reduction property of the IS estimator may be appealing over the *plug-in estimator*.

• Heavy-tailed q(.), relative to $\pi(\theta)$, are very efficient. The weights for thinner-tailed q(.) will be dominated by large $|\theta^m|$.

• IS can be turned into "importance sampling resampling" by using an additional resampling step based on the weights.

MC Integration: IS - Example

Example: We want to use importance sampling (IS) to evaluate the integral $x^{-1/2}$ over the range [0,1]:

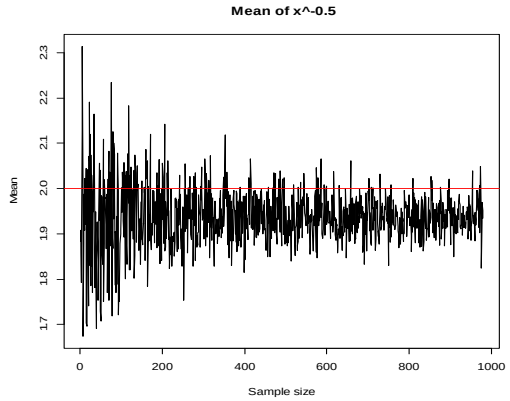
```
# Without Importance Sampling
set.seed(90)
M =1,000
lambda = 3
X <- runif(M,0.001,1)
h <- X^(-0.5)                                   # h(x)
c( mean(h), var(h) )

# Importance sampling Monte Carlo with an exponential
w <- function(x)   dunif(x, 0.001, 1)/dexp(x,rate=lambda) * pexp(1, rate=lambda)      # [pi(x)/q(x)]
h_f <- function(x)   x^(-0.5)                                                         # h(x)
X <- rexp(M,rate=lambda)
X <- X[X<=1]
Y.h <- w(X)*h_f(X)
c(mean(Y.h), var(Y.h))
```

Note: Make sure that q(x) is a well defined pdf –i.e., it integrates to 1. This is why above we use q(x)= dexp(x,lambda)/ pexp(1,lambda).

MC Integration: IS - Example

Example (continuation): Below, we plot the mean as a function of the sample size, *M*.



Note: After *M*=400, the mean stabilizes close to 2. A graph like this can be used to evaluate/determine *M* in MC integration.

MC Integration: IS – Importance weights

• If $\pi(\theta)$ is improper, $w(\theta^m)$ is normalized by $\sum_m w_m(\theta)$ (*normalized w*'s.)

Example: Suppose Y~N(θ,1) and we use a Cauchy (0,1) prior. That is, θ ~ Ca(0,1). Then,

$$p(\theta \mid y) \propto \frac{e^{-\frac{1}{2}(y-\theta)^2}}{(1+\theta^2)}$$

We set $q(\theta)$ as N(y,1). Then, the importance weights are given by:

$$w(\theta) = \frac{\pi(\theta)}{q(\theta)} = \frac{e^{-\frac{1}{2}(y-\theta)^2}}{(1+\theta^2)} \times \frac{1}{\frac{1}{\sqrt{2\pi}} e^{-\frac{1}{2}(\theta-y)^2}} = \frac{\sqrt{2\pi}}{(1+\theta^2)}$$

which we need to normalize:

$$\tilde{w}(\theta^m) = \frac{w(\theta^m)}{\sum_m q(\theta^m)} = \frac{\frac{\sqrt{2\pi}}{(1+\theta^{m\,2})}}{\sum_m \frac{\sqrt{2\pi}}{(1+\theta^{m\,2})}}$$

MC Integration: IS – Importance weights

Example (continuation): Now, we can calculate $E[\theta|y] = 0.5584926$.

Code in R

```
> M = 1000
> y = 1 # Data
> pi_th = function(theta,y) {
+ post_out = exp(-(y-theta)^2/2)/(1+theta^2)
+ return(post_out)
+ }
>
> # Draw theta from N(y,1)
> theta = rnorm(M,y,1)
>
> # Determine weights & post expectation
> w <- sqrt(2*pi)/(1+theta^2)
> w_norm <- w/sum(w)
>
> h = function(theta) theta
> sum(w_norm*h(theta))
[1] 0.5584926
```

Numerical Methods

• Q: Do we need to restrict our choices of prior distributions to these conjugate families? No. The posterior distributions are well defined irrespective of conjugacy. Conjugacy only simplifies computations.

• Outside the conjugate families, we rely on numerical methods for calculating posterior moments. In these cases, $P(\theta|y)$ is not recognized as a (analytical) pdf where we can sample from using a standard library function.

• Another situation where analytical posteriors are difficult to obtain is when the model is no longer linear.

• What do we do in these situations? We simulate the behavior of the model.

Numerical Methods: Simulation

• It is possible to do Bayesian estimation and inference over parameters in these situations, for example, with a nonlinear model.

• Steps:

1. Parameterize the model
2. Propose the likelihood conditioned on the parameters
3. Propose the priors – joint prior for all model parameters
4. As usual, the posterior is proportional to likelihood times prior. (Usually, it requires conjugate priors to be tractable. But, very likely, complicated $P(\theta|y)$.)
5. Sample –i.e., draw observations- from the posterior to study its characteristics.

Q: How do we draw? MCMC.

Numerical Methods: MCMC

• Sampling from the joint posterior $P(\boldsymbol{\theta}|y)$ may be difficult or impossible. For example, in the CLM, assuming a normal prior for β, and an inverse-gamma prior for σ^2, we get a complicated joint posterior distribution for (β, σ^2).

• To do simulation based estimation, we need joint draws on (β, σ^2). But, if $P(\boldsymbol{\theta}|y)$ is complicated $\Rightarrow$ we cannot easily draw from it.

• For these situations, many methods have been developed that make the process easier, including *Gibbs sampling*, *Data Augmentation*, and the *Metropolis-Hastings (MH) algorithm*.

• All three are examples of Markov Chain-Monte Carlo (MCMC) methods.

Numerical Methods: MCMC Preliminaries

• Monte Carlo (first MC): A simulation. We take quantities of interest of a distribution from simulated draws from the distribution.

Example: Monte Carlo integration

We have a posterior distribution $p(\theta|y)$ that we want to take quantities of interest from. We can evaluate the integral analytically, I:

$$I = \int h(\theta)\, p(\theta)\, d\theta$$

where $h(\theta)$ is some function of θ.

But when $p(\theta|y)$ is complicated, we will approximate the integral via MC Integration using the *plug-in estimator*, obtained by simulating M values from $p(\theta|y)$ and calculating:

$$I_M = \Sigma\, h(\theta)/M$$

Numerical Methods: MCMC Preliminaries

• From, the LLN, the MC approximation I_M is a consistent (simulation) estimator of the true value I. That is, $I_M \rightarrow I$, as $M \rightarrow \infty$.

Q: But, to apply the LLN we need independence. What happens if we cannot generate independent draws?

• Suppose we want to draw from our posterior $p(\theta|y)$, but we cannot sample independent draws from it. But, we may be able to sample draws from $p(\theta|y)$ that are "slightly" dependent.

If we can sample slightly dependent draws using a *Markov chain*, then we can still find quantities of interests from those draws.

Numerical Methods: MCMC Preliminaries

• Monte Carlo (first MC): A simulation.

• Markov Chain (the second MC): A stochastic process in which future states are independent of past states given the present state.

• Recall that a stochastic process is a consecutive set of random quantities defined on some known state space, Θ.

- Θ: our parameter space
- Consecutive implies a time component, indexed by t.

• A draw θ_t describes the state at time (iteration) t. The next draw θ^{t+1} is dependent *only* on θ^t. This is because of the *Markov property*:

$$p(\theta^{t+1} \mid \theta^t) = p(\theta^{t+1} \mid \theta^t, \theta^{t-1}, \theta^{t-2}, \ldots \theta^1)$$

Numerical Methods: MCMC Preliminaries

• The state of a Markov chain (MC) is a random variable indexed by *t*, say, θ_t. The state distribution is the distribution of θ_t, $p_t(\theta)$.

A stationary distribution of the chain is a distribution π such that, if

$$p_t(\theta) = \pi \qquad \Rightarrow p_{t+s}(\theta) = \pi \quad \text{for all } s.$$

• Under certain conditions a chain will have the following properties:

- A unique stationary distribution.
- Converge to that stationary distribution π as $t \to \infty$.
- Ergodic. That is, averages of successive realizations of θ will converge to their expectations with respect to π.

A lot of research has been devoted to establish the *certain conditions*.

MCMC – Ergodicity (P. Lam)

• Usual certain conditions for ergodicity

The Markov chain is *aperiodic*, *irreducible* (it is possible to go from any state to any other state), and *positive recurrent* (eventually, we expect to return to a state in a finite amount of time).

Ergodic Theorem

• Let $\theta^{(M)}=\{\theta^1, \theta^2, \theta^3, ..., \theta^M\}$ be M values from a Markov chain that is *aperiodic*, *irreducible*, and *positive recurrent* –i.e., chain is ergodic-, and $E[g(\theta)] < \infty$. Then, with probability 1:

$$\Sigma\, g(\theta)/M \rightarrow \int_\Theta g(\theta)\, \pi(\theta)\, d\theta$$

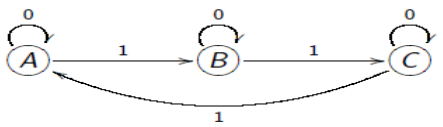
This is the Markov chain analog to the SLLN. It allows us to ignore the dependence between draws of the Markov chain when we calculate quantities of interest from the draws (like MC Integration).

MCMC - Ergodicity (P. Lam)

• Aperiodicity

A Markov chain is aperiodic if the only length of time for which the chain repeats some cycle of values is the trivial case with cycle length equal to one.

Let A, B, and C denote the states (analogous to the possible values of θ) in a 3-state Markov chain. The following chain is periodic with period 3, where the period is the number of steps that it takes to return to a certain state.



As long as the chain is not repeating an identical cycle, then the chain is aperiodic.

MCMC - Irreducibility and Stationarity

- Irreducibility

A Markov chain is irreducible if there no absorbing states or states in which the chain gets trapped.

- If $p_{ij} > 0$ (strictly positive) for all *i,j*, then the chain is irreducible

and there exists a *stationary distribution*, π, such that

$$\lim_{t\to\infty} \pi_0 \mathbf{P}^t = \pi$$

and

$$\pi \mathbf{P} = \pi.$$

Since the elements of **P** are all positive and each row sums to one, the maximum eigenvalue of $\mathbf{P}^T$ is one and π is determined by the corresponding eigenvector, $\mathbf{R}_1$, and the corresponding row vector from the inverse of the matrix for eigenvectors, $\mathbf{R}^{-1}$.

MCMC - Irreducibility and Stationarity

Proof: Using a singular value decomposition:

$$\mathbf{P} = \mathbf{R} \mathbf{\Lambda} \mathbf{R}^{-1}$$

where **R** is a matrix of eigenvectors and $\mathbf{\Lambda}$ is a diagonal matrix of corresponding eigenvalues, λ.

Recall: $(\mathbf{P}^T)^m = \mathbf{R} \mathbf{\Lambda}^m \mathbf{R}^{-1}$, since $(\mathbf{P}^T)^m = \mathbf{R} \mathbf{\Lambda} \mathbf{R}^{-1} \ldots \mathbf{R} \mathbf{\Lambda} \mathbf{R}^{-1} = \mathbf{R} \mathbf{\Lambda} \mathbf{R}^{-1}$.

• Then, the long-run steady-state is determined by $\max\{\lambda(\mathbf{P}^T)\}$ and in the direction of the corresponding vector from $\mathbf{R}^{-1}$(if the remaining λ's < 1 then $\lambda^m \to 0$ and their corresponding inverse eigenvectors' influence on direction dies out).

MCMC - Irreducibility and Stationarity

Since $\max\{\lambda(\mathbf{P}^T)\}=1$, with a large M $\Rightarrow \pi_0 P^t \rightarrow \pi_0 \times 1 = \pi$.

That is, after many iterations the Markov chain produces draws from a stationary distribution if the chain is irreducible.

MCMC: Markov Chain - Example

• A chain is characterized by its *transition kernel* whose elements provide the conditional probabilities of θ^{t+1} given the values of θ^t.

• The kernel is denoted by P(x,y). (The rows add up to 1.)

Example: Employees at t = 0 are distributed over two plants A & B

$$\pi_0' = [A_0 \quad B_0] = [100 \quad 100]$$

The employees stay and move between A & B according to P

$$P = \begin{bmatrix} P_{AA} & P_{AB} \\ P_{BA} & P_{BB} \end{bmatrix} = \begin{bmatrix} .7 & .3 \\ .4 & .6 \end{bmatrix}$$

At t = 1, the number of employees at A & B is given by:

$$\pi_1' = [A_1 \quad B_1] = \pi_0' P = [A_0 \quad B_0] \begin{bmatrix} P_{AA} & P_{AB} \\ P_{BA} & P_{BB} \end{bmatrix} = [100 \quad 100] \begin{bmatrix} .7 & .3 \\ .4 & .6 \end{bmatrix}$$

$$= [.7*100 + .4*100, \quad .3*100 + .6*100] = [110 \quad 90]$$

MCMC: Markov Chain - Example

At t = 2,

$$\begin{bmatrix} \boldsymbol{A}_1 & \boldsymbol{B}_1 \end{bmatrix} = \boldsymbol{\pi}_0' \boldsymbol{P} = \begin{bmatrix} \boldsymbol{A}_0 & \boldsymbol{B}_0 \end{bmatrix} \begin{bmatrix} P_{AA} & P_{AB} \\ P_{BA} & P_{BB} \end{bmatrix} = \begin{bmatrix} 110 & 90 \end{bmatrix}$$

$$\begin{bmatrix} \boldsymbol{A}_2 & \boldsymbol{B}_2 \end{bmatrix} = \boldsymbol{\pi}_0' \boldsymbol{P}^2 = \begin{bmatrix} \boldsymbol{A}_0 & \boldsymbol{B}_0 \end{bmatrix} \begin{bmatrix} P_{AA} & P_{AB} \\ P_{BA} & P_{BB} \end{bmatrix} \begin{bmatrix} P_{AA} & P_{AB} \\ P_{BA} & P_{BB} \end{bmatrix}$$

$$= \begin{bmatrix} 100 & 100 \end{bmatrix} \begin{bmatrix} .7 & .3 \\ .4 & .6 \end{bmatrix}^2 = \begin{bmatrix} 110 & 90 \end{bmatrix} \begin{bmatrix} .7 & .3 \\ .4 & .6 \end{bmatrix}$$

$$= \begin{bmatrix} 113 & 87 \end{bmatrix}$$

$\vdots$

After t = m years : $\boldsymbol{\pi}_k' = \begin{bmatrix} \boldsymbol{A}_m & \boldsymbol{B}_m \end{bmatrix} = \boldsymbol{\pi}_0' \boldsymbol{P}^m$

Note: Recall that under "certain conditions," as $t \to \infty$, P^t converges to the stationary distribution. That is, $\pi = \pi P$.

MCMC: General Idea

• We construct a chain, or sequence of values, $\theta^0, \theta^1, \ldots,$ such that for large m, θ^m can be viewed as a draw from the posterior distribution of θ, $p(\theta|\mathbf{X})$, given the data $\mathbf{X}=\{X_1, \ldots, X_N\}$.

• This is implemented through an algorithm that, given a current value of the parameter vector θ^m, and given $\mathbf{X}$, draws a new value θ^{m+1} from a distribution $f(.)$ indexed by θ^m and the data:

$$\theta^{m+1} \sim f(\theta | \theta^m, \mathbf{X})$$

• We do this in a way that if the original θ^m came from the posterior distribution, then so does θ^{m+1}. That is,

$$\theta^m | \mathbf{X} \sim p(\theta|\mathbf{X}), \quad \Rightarrow \quad \theta^{m+1} | \mathbf{X} \sim p(\theta | \mathbf{X}).$$

MCMC: General Idea

• In many cases, irrespective of where we start, that is, irrespective of θ^0, as $m \to \infty$, it will be the case that the distribution of the parameter conditional only on the data, **X**, converges to the posterior distribution as $m \to \infty$:

$$\theta^m | \mathbf{X} \xrightarrow{d} p(\theta | \mathbf{X}),$$

• Then just pick a θ^0 and approximate the mean and standard deviation of the posterior distribution as:

$$E[\theta | \mathbf{X}] = 1/(M - M^* + 1) \Sigma_{m=M^*...M} \theta^m$$
$$Var(\theta | \mathbf{X}) = 1/(M - M^* + 1) \Sigma_{m=M^*...M} \{\theta^m - E(\theta | \mathbf{X})\}^2$$

• Usually, the first M^*-1 iterations are discarded to let the algorithm converge to the stationary distribution without the influence of starting values, θ^0 (*burn in*).

MCMC: Burn-in (P. Lam)

• As a matter of practice, most people throw out a certain number of the first draws, $\{\theta^1, \theta^2, \theta^3, ..., \theta^{M^*}\}$, known as the *burn-in*. This is to make our draws closer to the stationary distribution and less dependent on the starting point.

• Think of it as a method to pick initial values.

• However, it is unclear how much we should burn-in since our draws are all slightly dependent and we do not know exactly when convergence occurs.

• Not a lot of theory about it.

MCMC: Thinning the Chain (P. Lam)

• In order to break the dependence between draws in the Markov chain, some have suggested only keeping every *d*th draw of the chain. That is, we keep $\theta^M = \{\theta^d, \theta^{2d}, \theta^{3d}, ... \theta^{Md}\}$

• This is known as thinning.

- Pros:
  - We may get a little closer to *i.i.d.* draws.
  - It saves memory since you only store a fraction of the draws.
- Cons:
  - It is unnecessary with ergodic theorem.
  - Shown to increase the variance of your MC estimates.

MCMC - Remarks

• In classical stats, we usually focus on finding the stationary distribution, given a Markov chain.

• MCMC methods turn the theory around: The invariant density is known (maybe up to a constant multiple) –it is the target density, $\pi(.)$, from which samples are desired–, but the transition kernel is unknown.

• To generate samples from $\pi(.)$, the methods find and utilize a transition kernel $P(x, y)$ whose Mth iteration converges to $\pi(.)$ for large M.

MCMC - Remarks

• The process is started at an arbitrary x and iterated a large number of times. After this, the distribution of the observations generated from the simulation is approximately the target distribution.

• Then, the problem is to find an appropriate $P(x, y)$ that works!

• Once we have a Markov chain that has converged to the stationary distribution, then the draws in our chain appear to be like draws from $p(\theta|y)$, so it seems like we should be able to use Monte Carlo Integration methods to find quantities of interest.

• Our draws are not independent, which we required for MC Integration to work (remember LLN). For dependent draws, we rely on the Ergodic Theorem.

MCMC - Remarks

• Our draws are dependent, the autocorrelation in the chain can be a problem for the MCMC estimators.

• Compared to MC estimators (MC simulations are the *"gold standard"*, since the draws are independent), MCMC estimators tend to have a higher variance and, then, worse approximations.

MCMC: Gibbs Sampling

• When we sample directly from the *conditional posterior distributions*, the algorithm is known as *Gibbs Sampling* (GS).

• The Gibbs sampler partitions the vector of parameters θ into two (or more) blocks or parts, say $\theta = (\theta_1, \theta_2, \theta_3)$. Instead of sampling θ^{m+1} directly from the (known) joint conditional distribution of

$$f(\theta \mid \theta^m; \mathbf{X}),$$

it may be easier to sample θ from the (known) *full conditional distributions*, $p(\theta_j \mid \theta_{-j}^m; \mathbf{X})$: $(\theta_{-j} = \theta_i^m, \theta_1^m)$

- first sample θ_1^{m+1} from $p(\theta_1 \mid \theta_2^m, \theta_3^m; \mathbf{X})$.
- then sample θ_2^{m+1} from $p(\theta_2 \mid \theta_1^{m+1}, \theta_3^k; \mathbf{X})$.
- then sample θ_3^{m+1} from $p(\theta_3 \mid \theta_1^{m+1}, \theta_2^{m+1}; \mathbf{X})$.

• It is clear that if θ^k is from the posterior distribution, then so is θ^{m+1}.

MCMC: Gibbs Sampling (P. Lam)

• Q: How can we know the joint distribution simply by knowing the full conditional distributions?

A: The Hammersley-Clifford Theorem shows that we can write the joint density, p(x, y), in terms of the conditionals p(x|y) and p(y|x).

• Then, how do we figure out the full conditionals?

Suppose we have a posterior $p(\theta|y)$. To calculate the full conditionals for each θ, do the following:

1. Write out the full posterior ignoring constants of proportionality.
2. Pick a block of parameters (say, θ_1) and drop everything that doesn't depend on θ_{-1}.
3. Figure out the normalizing constant (and, thus, the full conditional distribution $p(\theta_1 \mid \theta_{-1}, y)$).
4. Repeat steps 2 and 3 for all parameters.

MCMC: Gibbs Sampling - Steps

Example: In the previous MVN model, we derived the full conditional posteriors for $\boldsymbol{\mu}$ & $\boldsymbol{\Sigma}$:

$$p(\boldsymbol{\mu} \mid \boldsymbol{y}_1,\ldots,\boldsymbol{y}_N,\boldsymbol{\Sigma}) \propto e^{-\frac{1}{2}\boldsymbol{\mu}'\boldsymbol{A}_N\boldsymbol{\mu}+\boldsymbol{\mu}'\boldsymbol{b}_N}$$

$$p(\boldsymbol{\Sigma} \mid \boldsymbol{y}_1,\ldots,\boldsymbol{y}_N,\boldsymbol{\mu}) \propto |\boldsymbol{\Sigma}|^{-\frac{1}{2}(N+v_0+k+1)} e^{-\frac{1}{2}tr([\boldsymbol{S}_0+\boldsymbol{S}_{\boldsymbol{\mu}}]\boldsymbol{\Sigma}^{-1})}.$$

• Now, we draw $\boldsymbol{\theta}=(\boldsymbol{\mu}, \boldsymbol{\Sigma})$ using a GS. GS steps:

Step 1: Start with an arbitrary starting value $\theta^0 =(\boldsymbol{\mu}^0, \boldsymbol{\Sigma}^0)$. Set prior values: $\boldsymbol{\mu}_0$, Λ_0,v_0, S_0.

Step 2: Generate a sequence of θ's, following:

- Sample $\boldsymbol{\mu}^{m+1}$ from $p(\boldsymbol{\mu}| \boldsymbol{\Sigma}^m; \mathbf{X}) \sim$ Normal($\boldsymbol{\mu}_N$, Λ_N)
- Sample $\boldsymbol{\Sigma}^{m+1}$ from $p(\boldsymbol{\Sigma} | \boldsymbol{\mu}^{m+1}; \mathbf{X}) \sim$ IW(v_N, S_N^{-1})

Step 3: Repeat Step 2 for m= 1, 2, ... M.

MCMC: GS – MVN Example

Example (continuation): We are interested in the monthly correlation between the returns of IBM and DIS from 1990-2016. That is, we want to lear about $\boldsymbol{\Sigma}$.

• Priors. Suppose we have data on Kellog and SPY, which we use to set up the priors:

$\boldsymbol{\mu}_0$ = (.0066, 0065)

Λ_0 = S_0 = rbind(c(.0017, .00065),c(.00065, .0034))

v_0= K+2 = 4

• Data

$\bar{y}$ = (0.0089, 0.01)

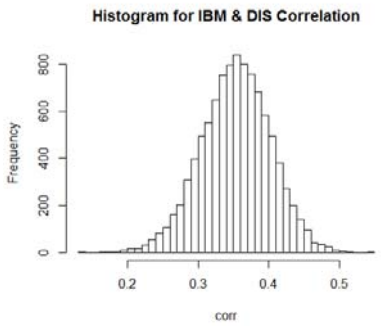
$\boldsymbol{\Sigma}$= rbind(c(.0061, .002),c(.002, .0054))

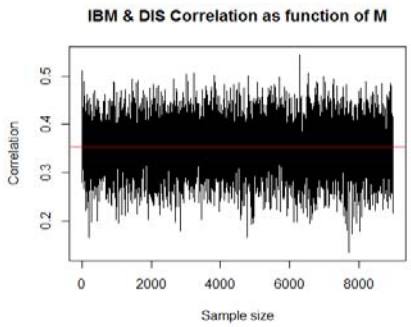
• Posterior (M=10,000, M_0=1,000) corr: .3546. (sample value: .3541)

95% CI: (0.2567, 0.4477)

MCMC: GS – MVN Example

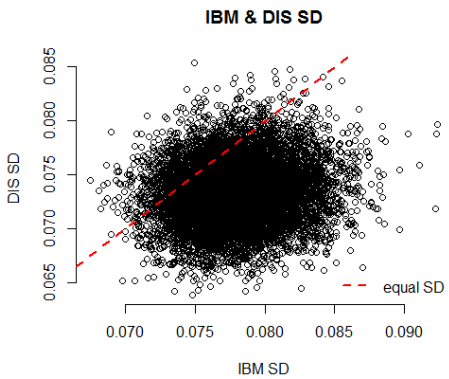
Example (continuation): We also check the histogram and traceplot, plotting the correlation as a funcion of M.





MCMC: GS – MVN Example

Example (continuation): We may also be interested in knowing which stock is more volatile. We can answer this question by constructing a 95% CI for SD_{IBM} - SD_{DIS}: (-0.00297, 0.012525).



IBM seems to be more volatile (88.12% of the time, IBM has higher SD. But, careful here: this does not mean the difference is 'economically' large).

MCMC: GS – Details

Note: The sequence $\{\theta^m\}_{m=1,...,M}$ is a Markov chain with transition kernel

$$\pi(\theta^{m+1} \mid \theta^k) = p(\theta_2^{m+1} \mid \theta_1^{m+1}; \mathbf{X}) \times p(\theta_1^{m+1} \mid \theta_2^m; \mathbf{X})$$

This transition kernel is a conditional distribution function that represents the probability of moving from θ^m to θ^{m+1}.

• Under general conditions, the realizations from a Markov Chain as $M \to \infty$ converge to draws from the ergodic distribution of the chain $\pi(\theta)$ satisfying

$$\pi(\theta^{m+1}) = \int \pi(\theta^{m+1} \mid \theta^m)\, \pi(\theta^m)\, d\theta^m$$

MCMC: GS – Diagnostics

• Convergence can be a problem for MCMC methods. It is important to check the robustness of the results before start using the output:

- Use different θ^0 (check traceplots for different sequences & GR)
- Use different M_0, M (may be use the "efective sample size," ESS)
- Plot θ^j as a function of j (check the auto-/cross-correlations in the sequence and across parameters).

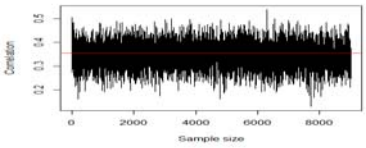
• Run Diagnostics Tests. There are many:

- Geweke (1992): A Z-test, comparing the means of the first 10% of sequence and the last 50% of the sequence.

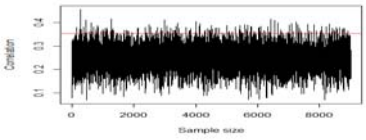
- Gelman and Rubin (GR, 1992): A test based on comparing different sequences, say N. The statistic is called *Shrink factor* is based on the ratio of the variance of the N posterior means sequences and the average of the posterior s^2 of the N sequences.

MCMC: GS – Diagnostics

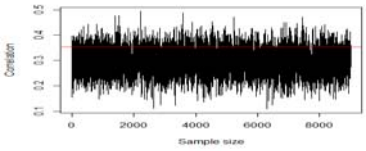
Example: Back to the monthly correlation between the returns of IBM and DIS from 1990-2016. We present traceplots for different θ^0:



$\mu_0 = (0, 0)$
$\Lambda_0 = S_0$ = rbind(c(.2^2, 0), c(0, .2^2))
$\Longrightarrow$ 95% CI for corr: (0.2469, 0.4406)



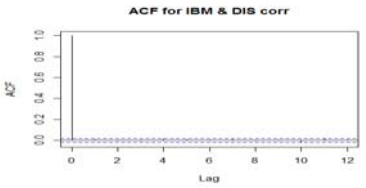
$\mu_0 = (0, 0)$
$\Lambda_0 = S_0$ = rbind(c(.6^2, -0.1), c(-0.1, .6^2))
$\Longrightarrow$ 95% CI for corr: (0.1460, 0.3500)



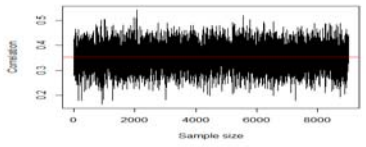
$\mu_0 = (.04, .04)$
$\Lambda_0 = S_0$ = rbind(c(.05^2, -0.1), c(-0.1, .05^2))
$\Longrightarrow$ 95% CI for corr: (0.1967, 0.3932)

MCMC: GS – Diagnostics

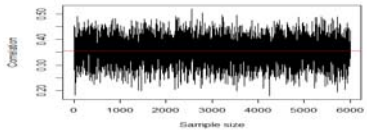
Example: We also present ACF and traceplots for different M_0:



$\Longrightarrow$ No autocorrelation problem.



$M_0 = 100$
$\Longrightarrow$ 95% CI for corr: (0.2565, 0.4478)



$M_0 = 4{,}000$
$\Longrightarrow$ 95% CI for corr: (0.2573, 0.4483)

MCMC: GS – Simplicity

• We sample from the known conditional posterior pdf of each parameter. We sample from them using a standard library function.

• Conjugacy is very helpful in this process. For example, if the conditional posterior distributions are all normal, gamma, or beta, then the GS makes sampling from the joint posterior easy.

• To take advantage of the simplicity of the GS, there are situations where drawing from the conditionals is not possible, but it may be possible to convert the problem into one where the GS works (see Albert and Chib's (1993) probit regression set up).

MCMC: GS – Limitations

Three usual concerns:

• Even if we have the full posterior joint pdf, it may not be possible or practical to derive the conditional distributions for each of the RVs in the model.

• Second, even if we have the posterior conditionals for each variable, it might be that they are not of a known form, and, thus, there is not a straightforward way to draw samples from them.

• Finally, there are cases where GS is very inefficient. That is, the "*mixing*" of the GS chain might be very slow, -i.e., the algorithm spends a long time exploring a local region with high density, and takes a very long to explore all regions with signicant probability mass.

GS – Example 1: Bivariate Normal

Draw a random sample from bivariate normal $\left[\begin{pmatrix}0\\0\end{pmatrix}, \begin{pmatrix}1 & \rho\\ \rho & 1\end{pmatrix}\right]$

(1) Direct approach: $\begin{pmatrix}v_1\\v_2\end{pmatrix}_r = \Gamma\begin{pmatrix}u_1\\u_2\end{pmatrix}_r$ where $\begin{pmatrix}u_1\\u_2\end{pmatrix}$ are two independent standard normal draws (easy) and $\Gamma=\begin{pmatrix}1 & 0\\ \theta_1 & \theta_2\end{pmatrix}$ such that $\Gamma\Gamma'=\begin{pmatrix}1 & \rho\\ \rho & 1\end{pmatrix}$. $\theta_1 = \rho$, $\theta_2 = \sqrt{1-\rho^2}$.

(2) Gibbs sampler: $v_1 \mid v_2 \sim N\left[\rho v_2, \sqrt{1-\rho^2}\right]$

$v_2 \mid v_1 \sim N\left[\rho v_1, \sqrt{1-\rho^2}\right]$

GS – Example 1: Bivariate Normal

• R Code

```
# initialize constants and parameters

M <- 5,000              # length of chain
burn <- 1,000           # burn-in length
X <- matrix(0, M, 2)    # the chain, a bivariate sample

rho <- -.75             # correlation
mu1 <- 0
mu2 <- 0
sigma1 <- 1
sigma2 <- 1
s1 <- sqrt(1-rho^2)*sigma1
s2 <- sqrt(1-rho^2)*sigma2
```

GS – Example 1: Bivariate Normal

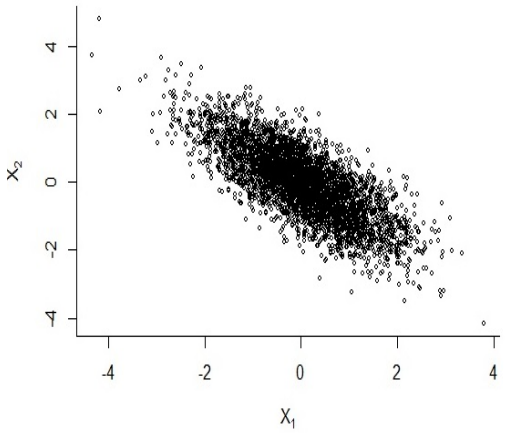
```
# generate the chain
   X[1, ] <- c(mu1, mu2)          #initialize
   for (i in 2:M) {
      x2 <- X[i-1, 2]
      m1 <- mu1 + rho * (x2 - mu2) * sigma1/sigma2
      X[i, 1] <- rnorm(1, m1, s1)
      x1 <- X[i, 1]
      m2 <- mu2 + rho * (x1 - mu1) * sigma2/sigma1
      X[i, 2] <- rnorm(1, m2, s2)
   }

   b <- burn + 1
   x <- X[b:M, ]

   # compare sample statistics to parameters
   colMeans(x)
   cov(x)
   cor(x)
   plot(x, main="", cex=.5, xlab=bquote(X[1]),
       ylab=bquote(X[2]), ylim=range(x[,2]))
```

GS – Example 1: Bivariate Normal

```
>    colMeans(x)
[1]  0.03269641 -0.03395135
>    cov(x)
          [,1]       [,2]
[1,]  1.0570041 -0.8098575
[2,] -0.8098575  1.0662894
>    cor(x)
          [,1]       [,2]
[1,]  1.0000000 -0.7628387
[2,] -0.7628387  1.0000000
```



GS – Example 2: CLM

In the CLM, we assume a normal prior for $\beta \sim N(\boldsymbol{m},\boldsymbol{\Sigma})$ and a gamma for $h=1/\sigma^2 \sim \Gamma(\alpha_0,\lambda_0)$. The joint posterior is complicated. We use the conditional posteriors to get the joint posterior:

$$f(\beta \mid \mathbf{y}, \mathbf{X}, h) \propto \exp\{-\frac{1}{2}[(\beta - m^*)' \Sigma^* (\beta - m^*)\}$$

$$f(h \mid \beta, \mathbf{y}, \mathbf{X}) \propto h^{T/2+\alpha_0-1} \exp\{-\frac{h}{2}(y - X\beta)'(y - X\beta) - h\lambda_0\}$$

where $m^* = \Sigma^{*-1}(h\mathbf{X'y}+h\Lambda m)$ and $\Sigma^*= (h\mathrm{X'X}+h\Lambda)^{-1}$. That is, we get a multivariate normal, for β with the usual mix of prior and sample info and a gamma for h, with parameters $(T/2+\alpha_0, (\mathbf{y}-\mathbf{X}\beta)'(\mathbf{y}-\mathbf{X}\beta)/2+\lambda_0)$.

• The GS samples back and forth between the two conditional posteriors.

GS – Example 2: CLM

Setup of GS in R (fill in the rest as exercise):

```
# Setup data (y,X)
set.seed(666)
T <- 50
beta.true <- c(1, 2);  sigma.true <- 5
X <- matrix(c(rep(1, T), rnorm(T, sd = sigma.true)), nrow = T)
y <- rnorm(T, X %*% beta.true, sigma.true)

# Specify the size of the burn-in and the number of draws thereafter; set the prior parameters:
burnin <- 100;  M <- 5000
m_0 <- matrix(c(0, 0), nrow = ncol(X))
Sig_inv <- diag(c(10^-2, 10^-2))
alpha_0 <- 0.1;  lamba_0 <- 0.1

# Pre-calculate some values outside the main MCMC loop:
p <- ncol(X)
pre_Sigma_star <- solve(crossprod(X) + Sig_inv)
pre_Mean <- pre_Sigma_star  %*% (crossprod(X, y) + Sig_inv %*% m_0)
pre_alpha <- alpha_0 + T/2 + p/2
```

GS – Example 2: CLM

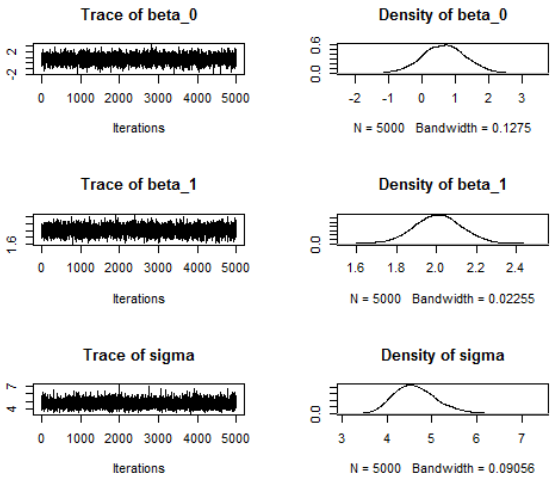
```
# Posterior means:
> colMeans(draws1)
  beta_0    beta_1     sigma
0.6637359 2.0072637 4.6392312

> # for comparison:
> summary(lm(y~X-1))
Call:
lm(formula = y ~ X - 1)


Coefficients:
   Estimate Std. Error t value Pr(>|t|)
X1   0.6595     0.6626   0.995    0.325
X2   2.0067     0.1180  17.011   <2e-16 ***
---
Residual standard error: 4.678 on 48 degrees of freedom
Multiple R-squared:  0.8577,    Adjusted R-squared:  0.8518
F-statistic: 144.7 on 2 and 48 DF,  p-value: < 2.2e-16
```

GS – Example 2: CLM

```
# Plot draws through coda's native plot method:
> plot(coda::mcmc(draws1), show.obs = FALSE)
```



GS – Example 3: Stochastic Volatility (SV)

The standard SV model (DGP):

$$y_t = \mu \exp(h_t/2)\upsilon_t \qquad \upsilon_t \sim i.i.d.N(0,1)$$
$$h_t = \omega + \phi(h_{t-1} - \omega) + \sigma\eta_t \qquad \eta_t \sim i.i.d.N(0,1)$$

• In general, we take y_t as log returns. It is common to think of $\mu \approx 1$. h_t is the log of the time-varying variance. We have 3 SV parameters to estimate $\theta = (\omega, \phi, \sigma)$ and the latent vector $\boldsymbol{h} = \{h_t, h_{t-1}, h_{t-2}, ..., h_1\}$

• We can rewrite the SV model in hierarchical form:

$$y_t \mid h_{t_t} \sim i.i.d.N(0, \exp(h_t))$$
$$h_t \mid h_{t-1}, \omega, \phi, \sigma^2 \sim N(\omega + \phi(h_{t-1} - \omega), \sigma^2)$$
$$h_1 \mid \omega, \phi, \sigma \sim N(\omega, \sigma^2/(1-\phi^2))$$

GS – Example 3: Stochastic Volatility (SV)

• Priors for θ. Assume independence of parameters:

1) $\omega \mid \phi, \sigma^2, \mathbf{h}, \mathbf{x} \sim$ Normal(ω_0, M_0) Usually vague: $\omega_0=0$, M_0=big ($\geq$100 for daily log returns.) $\Rightarrow$ Normal posterior

2) $\phi \mid \sigma^2, \omega, \mathbf{h}, \mathbf{x} \sim$ Normal(ϕ, $V_{\phi,0}$) $\Rightarrow$ Normal posterior

But, we may want to restrict $\phi \in (-1,1)$. Then,

$\phi \mid \sigma^2, \omega, \mathbf{h}, \mathbf{x} \sim$ Beta (adjusted): $(\phi+1)/2 \sim$ Beta(α_0,β_0)

$\Rightarrow E[\phi] = 2\alpha_0/(\alpha_0+\beta_0) - 1$ ($\alpha_0=20$; $\beta_0=2$ $\Rightarrow E[\phi]=.82$)

3) $\sigma^2 \mid \phi, \omega, \mathbf{h}, \mathbf{x} \sim$ IG $\Rightarrow$ IG posterior

Fruhwirth-Schnatter and Wagner (2010) propose $\Gamma(1/2, 1/2\, V_\sigma)$.

• Latent vector **h**. Note that the SV model implies:

$$\ln(y_t^2) = \omega + (h_t - \omega) + \log(\upsilon_t^2)$$

We know y_t. Then, if we draw $z_t = \log(\upsilon_t^2)$, we can draw $h_t \mid \phi, \omega, \mathbf{x}$. But, z_t is the $\log(\chi_1^2)$, a non standard pdf, but it can be approximated.

GS – Example 4: Logistic Regression

• A standard Bayesian logistic regression model (e.g., modelling the probability of a merger) can be written as follows:

$$y_i \sim Binomial\ (n_i, p_i)$$
$$logit\ (p_i) = X\beta$$
$$\beta_0 \sim N(0, m_0), \beta_1 \sim N(0, m_1)$$

• To use GS, from the complicated posterior, we write down the conditional posterior distributions, as usual. Say, for β_0:

$$p(\beta_0 \mid y, \beta_1) \propto p(y \mid \beta_0, \beta_1) p(\beta_0)$$
$$\propto \prod_i \left(\frac{\exp(\beta_0 + \beta_1 x_i)}{1 + \exp(\beta_0 + \beta_1 x_i)} \right)^{y_i} \left(\frac{1}{1 + \exp(\beta_0 + \beta_1 x_i)} \right)^{n_i - y_i} \times \frac{1}{\sqrt{m_0}} \exp(-\frac{\beta_0^2}{2m_0})$$
$$p(\beta_0 \mid y, \beta_1) \sim ?$$

GS – Example 4: Logistic Regression

• But, this distribution is not a standard distribution. It cannot be simply simulated from a standard library function. Thus, the GS cannot be used here.

• But, we can simulate it using MH methods.

• The R package MCMCpack can estimate this model. Also, Bayesian software OpenBUGS, JAGS, WinBUGS, and Stan, which link to R, can fit this model using MCMC.

See R link: https://cran.r-project.org/web/views/Bayesian.html

MCMC: Data Augmentation

• Situation: It is difficult or impossible to sample θ directly from the posterior, but there exists an unobservable/latent variable Y such that it is possible to conditionally sample $P(\theta | Y)$ and $P(Y | \theta)$.

• *Data augmentation* (DA): Methods for constructing iterative optimization or sampling algorithms through the introduction of unobserved data or latent variables.

• DA was popularized by Dempster, Laird, and Rubin (1977), in their article on the EM algorithm, and by Tanner and Wong (1987).

• A DA algorithm starts with the construction of the so-called *augmented data*, Y_{aug}, which are linked to the observed data, Y_{obs}, via a many-to-one mapping M: $Y_{aug} \longrightarrow Y_{obs}$.

MCMC: Data Augmentation

• Now, we have "complete data." To work with it, we require that the marginal distribution of Y_{obs} implied by $P(Y_{aug} | \theta)$ must be the original model $P(Y_{obs} | \theta)$. That is, we relate the "observed data" posterior distribution to the "complete data":

$$f(\theta | Y_{obs}, M) = \int f(\theta, Y_{aug} | Y_{obs}, M) dY_{aug} = \int f(\theta | Y_{aug}, Y_{obs}, M) f(Y_{aug} | Y_{obs}, M) dY_{aug}$$

• We introduce the RV Y_{aug} because it helps. We have a situation where a Gibbs sampler can be used to simulate $P(\theta | Y_{obs}, M)$. Two steps:

- Draw Y_{aug} from their joint posterior, $P(Y_{aug} | Y_{obs}, M)$
- Draw θ from its completed-data posterior: $P(\theta | Y_{obs}, Y_{aug}, M)$

Q: Under which conditions, inference from completed data and inference from observed data are the same?

MCMC: Data Augmentation – Example 1

Suppose we are interested in estimating the parameters of a censored regression. There is a latent variable:

$$Y_i^* = X_i\beta + \varepsilon_i, \qquad \varepsilon_i | X_i \sim iid\ \mathrm{N}(0, 1) \qquad i=1,2,...,K,....,L$$

• We observe $Y_i = \max(0, Y_i^*)$, and the regressors X_i. Suppose we observe $(L\text{-}K)$ zeroes.

• Suppose the prior distribution for β is $\mathrm{N}(\mu, \Omega)$. But, the posterior distribution for β does not have a closed form expression.

Remark: We view both the vector $\mathbf{Y}^* = (Y_1^*,...,Y_N^*)$ and β as unknown RV. With an appropriate choice of $P(\mathbf{Y}^* | \text{data}, \beta)$ and $P(\beta | \mathbf{Y}^*)$, we can use a Gibbs sample to get the full posterior $P(\beta, \mathbf{Y}^* | \text{data})$.

MCMC: Data Augmentation – Example 1

• The GS consists of two steps:

Step 1 (Imputation): Draw all the missing elements of $\mathbf{Y}^*$ given the current value of the parameter β, say β^m:

$$Y_i^* | \beta, \text{ data} \sim \mathrm{TN}(X_i\beta^m, 1; 0) \qquad (\text{an } L\text{x}1\text{vector!})$$

if observation i is truncated, where $\mathrm{TN}(\mu, \sigma^2; c)$ denotes a truncated normal distribution with mean μ, variance σ^2, and truncation point c (truncated from above).

Step 2 (Posterior): Draw a new value for the parameter, β^{m+1} given the data and given the (partly drawn) $\mathbf{Y}^*$:

$$p(\beta | \text{data}, \mathbf{Y}^*) \sim \mathrm{N}((\mathbf{X'X}+\Omega^{-1})^{-1}\ (\mathbf{X'Y}+\Omega^{-1}\mu), (\mathbf{X'X}+\Omega^{-1})^{-1})$$

MCMC: Data Augmentation – Example 2

Example: Incomplete univariate data

Suppose that $Y_1, ..., Y_L \sim$ Binomial $(1, \theta)$

Prior for $\theta \sim$ Beta(α,β)

Then, the posterior of θ is also Beta:

$$p(\theta|Y) \sim \text{Beta}(\alpha + \Sigma_{i \text{ to } L} Y_i , \beta + L - \Sigma_{i \text{ to } L} Y_i)$$

Suppose L-K observations are missing. That is, $Y_{obs} = \{Y_1, ..., Y_K\}$

Then, $p(\theta|Y_{obs}) \sim \text{Beta}(\alpha + \Sigma_{i \text{ to } K} Y_i , \beta + K - \Sigma_{i \text{ to } K} Y_i)$

Step 1: Draw all the missing elements of $\mathbf{Y}^*$ given the current value of the parameter θ, say θ^m.

Step 2: Draw a new value for the parameter, θ^{m+1} given the data and given the (partly drawn) $\mathbf{Y}^*$.

MCMC: Metropolis-Hastings (MH)

• MH is an alternative, and more general, way to construct an MCMC sampler (to draw from the posterior). The Gibbs sampler is a simplified version of the MH algorithm (so simplified, it does not look like it).

• It provides a form of generalized rejection sampling, where values are drawn –i.e., the θ's– from approximate distributions and "corrected" so that, asymptotically they behave as random observations from the *target distribution* –for us, the posterior.

• MH sampling algorithms sequentially draw candidate observations from a '*proposal*' distribution, conditional on the current observations, thus inducing a Markov chain.

MCMC: Metropolis-Hastings (MH)

• We deal with Markov chains: The distribution of the next sample value, say $y=\theta^{m+1}$, depends on the current sample valule, say $x=\theta^{m}$.

• In principle, the algorithm can be used to sample from any integrable function. But, its most popular application is sampling from a posterior distribution.

• The MH algorithm jumps around the parameter space, but in a way that the probability to be at a point is proportional to the function we sample from –i.e., the target function.

• Named after Metropolis et al. (1953), which first proposed it and Hastings (1970), who generalized it. Rediscovered by Tanner and Wong (1987) and popularized by Gelfand and Smith (1990).

MCMC: MH – Proposal Distribution

• We want to find a function $p(x, y)$ from where we can sample, that satisfies the *(time) reversibility condition* (*equation of balance*), a sufficient condition for stationarity of $\pi(.)$:

$$\pi(x)\, p(x, y) = \pi(y)\, p(y,x)$$

• The *proposal* (or *candidate-generating) density* is denoted $q(x,y)$, where

$$\int q(x,y)\, dy = 1.$$

Interpretation: When a process is at the point x ($=\theta^{k}$), the density generates a value y ($=\theta^{m+1}$) from $q(x,y)$. It tells us how to move from current x to new y. Another notation for $q(x,y) = q(y|x)$.

• Idea: Suppose P is the true density. We simulate a point y using $q(x,y)$. We 'accept' it only if it is "*likely*." If it happens that $q(x,y)$ itself satisfies the reversibility condition for all (x,y), we are done.

MCMC: MH – Proposal Distribution

• But, for example, we might find that for some (x,y):

$$\pi(x)\, q(x, y) > \pi(y)\, q(y,x) \qquad (*)$$

In this case, speaking somewhat loosely, the process moves from x to y too often and from y to x too rarely.

• We want balance. To correct this situation by reducing the number of moves from x to y with the introduction of a probability $a(x,y) < 1$ that the move is made:

$$a(x,y) = \text{probability of move from } x \text{ to } y.$$

If the move is not made, the process again returns x as a value from the target distribution.

• Then, transitions from x to y ($y \neq x$) are made according to

$$p_{MH}(x, y) = q(x,y)\, a(x,y) \qquad y \neq x$$

MCMC: MH – Algorithm Rejection Step

Example: We focus on a single parameter θ and its posterior distribution $\pi(\theta)$. We draw a sequence $\{\theta^1, \theta^2, \theta^3, ...\}$ from a MC.

- At iteration m, let $\theta = \theta^m$. Then, propose a move: θ^*. That is, generate a new value θ^* from a proposal distribution $q(\theta^m, \theta^*)$.
- Rejection rule:

Accept θ^* (& let $\theta^{m+1}=\theta^*$)	with (acceptance) probability $a(\theta^m,\theta^*)$
Reject θ^*	with probability 1-a (& set $\theta^{m+1}= \theta^k$).

We have define an acceptance function!

Note: It turns out that the acceptance probability, $a(x,y)$, is a function of $\pi(y)/\pi(x)$ –the *importance ratio*. This ratio helps the sampler to visit higher probability areas under the full posterior.

MCMC: MH – Probability of Move

• We need to define $a(x,y)$, the probability of move.

• In our example (*), to get movements from x to y, we define $a(y,x)$ to be as large as possible (with upper limit 1!). Now, the probability of move $a(x,y)$ is determined by requiring that $p_{MH}(x, y)$ satisfies the reversibility condition. Then,

$$\pi(x)\, q(x, y)\, a(x,y) = \pi(y)\, q(y,x)\, a(y,x) = \pi(y)\, q(y,x)$$
$$\Rightarrow a(x,y) = \pi(y)\, q(y,x)/[\pi(x)\, q(x, y)].$$

Note: If the example (*) is reversed, we set $a(x,y)=1$ and $a(y,x)$ as above.

• Then, in order for $p_{MH}(x, y)$ to be reversible, $a(x,y)$ must be

$a(x,y) = \min\{[\pi(y)\, q(y,x)]/[\pi(x)\, q(x, y)],1\}$ if $\pi(x)\, q(x, y)]>0$,

$= 1$ otherwise.

MCMC: MH – Probability of Move

• If $q(.)$ is symmetric, then $q(x,y) = q(y,x)$. Then, the probability of move $a(x,y)$ reduces to $\pi(y)/\pi(x)$ –the *importance ratio*. Thus, the acceptance function:

- If $\pi(y) \geq \pi(x)$, the chain moves to y.
- Otherwise, it moves with probability given by $\pi(y)/\pi(x)$.

Note: This case, with $q(.)$ symmetric, is called *Metropolis Sampling*.

• This acceptance function plays two roles:

1) It helps the sampler to visit higher probability areas under the full posterior –we do this through the ratio $\pi(y)/\pi(x)$.

2) It should explore the space and avoid getting stuck at one site –i.e., it can reverse its previous move. This constraint is given by the ratio by $q(y,x)/q(x,y)$.

MCMC: MH – At Work

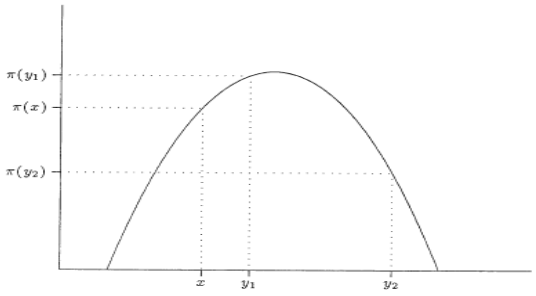


Figure 1. Calculating Probabilities of Move With Symmetric Candidate-Generating Function (see text).

• We consider moves from x (note that $q(x,y)$ is symmetric):

- A move to candidate y_1 is made with certainty –i.e., $\pi(y_1)>\pi(x)$.

⇒ We always say yes to an "up-hill" jump!

- A move to candidate y_2 is made with probability $\pi(y_2)/\pi(x)$.

Note: The $q(x,y)$ distribution is also called *jumping distribution*.

MCMC: MH – Transition Kernel

• In order to complete the definition of the transition kernel for the MH chain, we consider the possibly non-zero probability that the process remains at x:

$$r(x) = 1 - \int_R q(x,y)\, a(x,y)\, dx.$$

• Then, the transition kernel of the MH chain, denoted by $p_{MH}(x, y)$ is given by:

$$p_{MH}(x, y) = q(y,x)\, a(x,y)\, dy + [1 - \int_R q(x,y)\, a(x,y)\, dy]\, I_x(y).$$

where $I_x(y)$ is an indicator funtion = 1 if $x = y$ and 0 otherwise.

MCMC: MH Algorithm

• MH Algorithm

We know $\pi(\theta) = P(\theta|y) = P(y|\theta) \times P(\theta)/P(y)$, a complicated posterior. For example, from the CLM with Y_i *iid* normal, normal prior for $\boldsymbol{\beta}$ and gamma prior for h. $\Rightarrow \theta=(\boldsymbol{\beta},h)$.

Then, $$\frac{\pi(\theta^{m+1})}{\pi(\theta^{m})} = \frac{P(\theta^{m+1}|y)P(\theta^{m})}{P(\theta^{m}|y)P(\theta^{m+1})}.$$

P(y), the normalizing constant, plays no role. It can be ignored.

Assumptions:

- A symmetric proposal q(.) –i.e., $q(\theta^m, \theta^{m+1}) = q(\theta^{m+1}, \theta^m)$. Then,
$$a(\theta^m, \theta^{m+1}) = \pi(\theta^{m+1})/\pi(\theta^m).$$
- A starting value for θ: θ^0 (m=0).

MCMC: MH Algorithm

• MH Algorithm – Steps:

(1) Initialized with the starting value θ^0 (m=0):

(2) Generate θ^* from $q(\theta^m, .)$ and draw u from U(0, 1).

- If $u \leq a(\theta^m, \theta^*) = \pi(\theta^*)/\pi(\theta^m)$ $\Rightarrow$ set $\theta^{m+1} = \theta^*$.

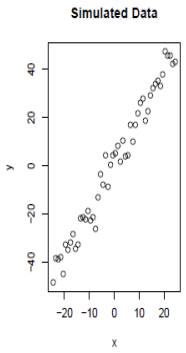
  Else $\Rightarrow$ set $\theta^{m+1} = \theta^m$.

(3) Repeat for $m = 1, 2, \ldots, M$.

• Return the values $\{\theta^{(M)}\} = (\theta^1, \theta^2, \ldots, \theta^m, \theta^{m+1}, \ldots, \theta^M)$.

MCMC: MH Algorithm – CLM Example

• (From Florian Hartig) Suppose we have the CLM with
Data: $Y_i = \alpha + \beta X_i + \varepsilon_i$, $\varepsilon_i \sim iid\, N(0,\sigma^2)$.
Priors: $\beta \sim U(0,10)$; $\alpha \sim N(m=0, \sigma_0^2=9)$; & $\sigma^2 \sim U(0.001,30)$
=> $\boldsymbol{\theta}=(\alpha,\beta,\sigma^2)$.



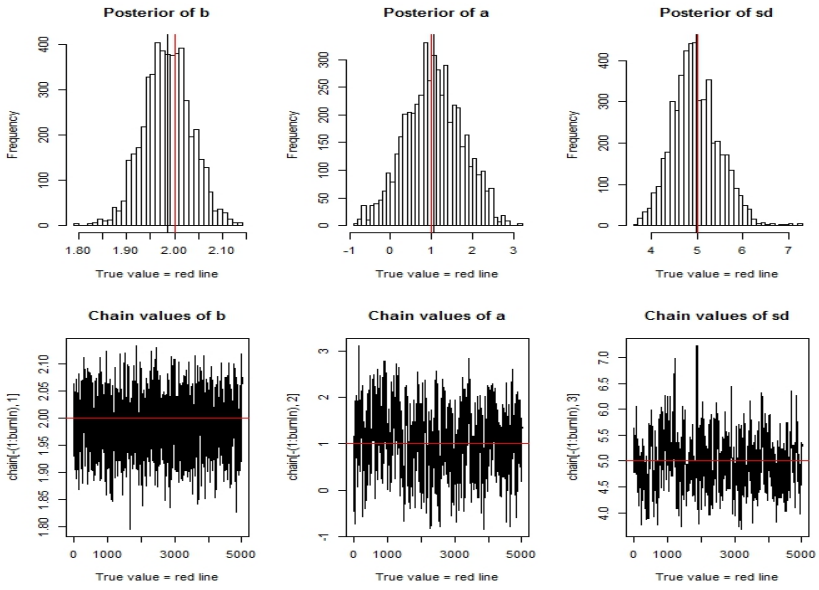
• We simulate the data, with $\alpha=1$, $\beta=2$, $\sigma=5$, & $T=50$.

• Proposal densities: 3 Normals with $\boldsymbol{\theta}^0=(2,0,7)$ & **SD**=(0.1,0.5,0.3).

• Iterations = 10,000 & Burn-in = 5,000.

• OLS: a=1.188 (0.68), b=1.984 (.047).

MCMC: MH Algorithm – CLM Example



MCMC: MH – Remarks about q(.)

• We pick proposal distributions, $q(x,y)$, that are easy to sample. But, remarkably, $q(x,y)$ can have almost any form.

• It is usually a good idea to choose $q(x,y)$ close to the posterior, $\pi(.)$.

• There are some (silly) exceptions; but assuming that the proposal allows the chain to explore the whole posterior and does not produce a recurrent chain we are OK.

• We tend to work with symmetric $q(x,y)$, but the problem at hand may require asymmetric proposal distributions; for instance, to accommodate a particular constraints in the model. For example, to estimate the posterior distribution for a variance parameter, we require that our proposal does not generate values smaller than 0.

MCMC: MH – Special Cases

• Three special cases of MH algorithm are:

1. Random walk metropolis sampling. That is,

$$y = x + z, \qquad z \sim q(z)$$

2. Independence sampling. That is,

$$q(x,y) = q(y).$$

3. Gibbs sampling. (We never reject from the proposals –i.e., the conditional posteriors!)

• Critical decision: Selecting the spread and location of $q(x,y)$. Note that difference choices deliver different *acceptance rates* –i.e., the fraction of candidate draws that are accepted).

Changing the spread and location of $q(x,y)$ to get a desired acceptance rate is called *tuning*.

MCMC: MH – Random Walk Metropolis

• This is a pure Metropolis sampling –see Metropolis et al. (1953).

- Let $q(x,y) = q(|y - x|)$ q is a multivariate symmetric pdf.

- $y = x + z$, where $z \sim q$. (It is called a *random walk* chain!)

• Typical RW proposals: Normal distribution centered around the current value of the parameter –i.e. $q(x,y) \sim N(x,s^2)$, where s^2 is the (fixed) proposal variance that can be *tuned* to give particular acceptance rates. Multivariate t-distributions are also used.

• The RW MH is a good alternative, usual default for the algorithm.

MCMC: MH Algorithm – RW Example

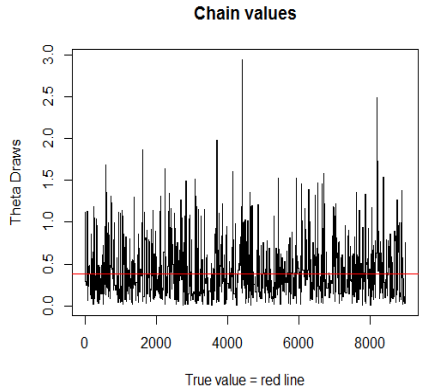
Example: (From P. Lam.) We use a RW MH algorithm to sample from a Gamma(1.7, 4.4) distribution with $q(x,y) \sim N(x, [SD=2]^2)$.

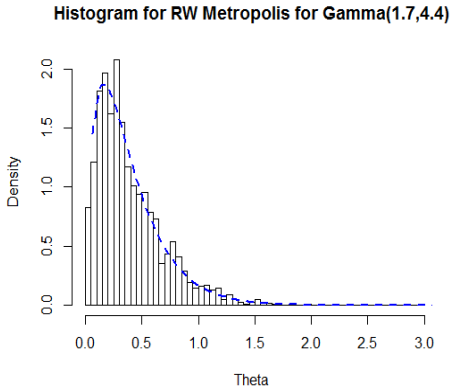
```
mh.gamma <- function(M.sims, start, burnin, cand.sd, shape, rate) {
        theta.cur <- start
        draws <- c()
        theta.update <- function(theta.cur, shape, rate)  {
              theta.can <- rnorm(1, mean = theta.cur, sd = cand.sd)                    # RW θ^(m+1)?
              accept.prob <- dgamma(theta.can, shape, rate)/dgamma(theta.cur,shape, rate)   # a()
               if (runif(1) <= accept.prob)  theta.can                                 # reject?
               else theta.cur
              }
       for (i in 1:M.sims) {
              draws[i] <- theta.cur <- theta.update(theta.cur, shape,rate)
       }
return(draws[(burnin + 1):M.sims])
}
 mh.draws <- mh.gamma(10000, start = 1, burnin = 1000, cand.sd = 2,shape = 1.7, rate = 4.4)
```

MCMC: MH Algorithm – RW Example

Example (continuation):

```
> mean(mh.draws)
[1] 0.3962097                                  # theoretical mean = 1.7/4.4 = 0.3863636
> hist(mh.draws, main="Histogram for RW Metropolis for Gamma(1.7,4.4)",xlab="Theta", breaks=50)
```





MCMC: MH – Independence Sampler

• The independence sampler is so called as each proposal is independent of the current parameter value. That is,

$q(x,y) = q(y)$ (an independent chain -see Tierney (1994).)

That is, all our candidate draws y are drawn from the same distribution, regardless of where the previous draw was.

This leads to acceptance probability $a(x, y) = \min\left(1, \frac{\pi(y)q(x)}{\pi(x)q(y)}\right)$.

Note that to determine $a(.)$, we use a ratio of importance weights.

• Distributions used for q(.): A Normal based around the ML estimate with inflated variance. A Multivariate-t.

MCMC: MH – Independence Sampler

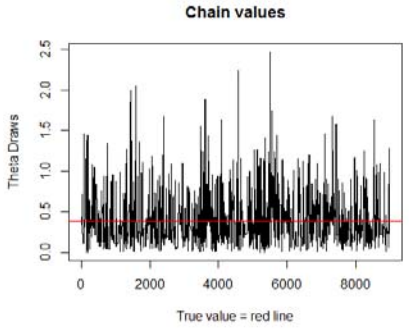
Example: We use an independence sampler MH algorithm to sample from a Gamma(1.7, 4.4) distribution with $q(y) \sim N(0, [SD=2]^2)$.

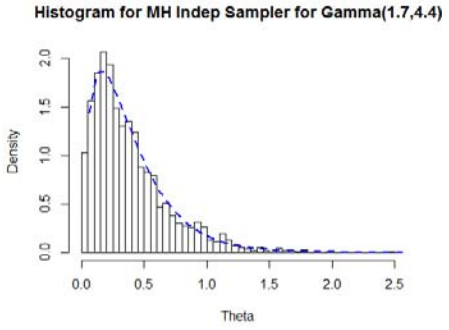
Code in R: Same code, but the theta.update function changes to

```
theta.update <- function(theta.cur, shape, rate)  {
        theta.can <- rnorm(1, mean = cand.mu, sd = cand.sd)                    # IS θ^{m+1}?
        accept.prob <- dgamma(theta.can, shape, rate)*dnorm(theta.cur, cand.mu, cand.sd)/
(dgamma(theta.cur,shape, rate)*dnorm(theta.can, cand.mu, cand.sd))             # a(.)
        if (runif(1) <= accept.prob)  theta.can                                # reject?
        else theta.cur
        }
```

MCMC: MH – Independence Sampler

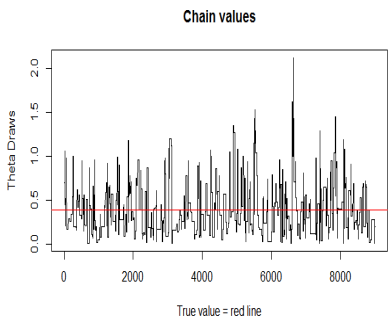
Example (continuation): Below, we draw the traceplot and histogram for the generated draws. The generated ED looks fine. That is, it can be used to calculate quantities of interest.

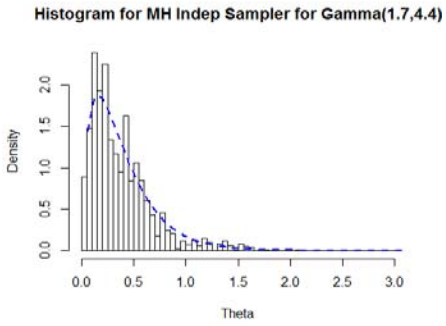




MCMC: MH – Independence Sampler

Example (continuation): Now, suppose we start with x_0=3 and use q(y) ~ N(2, [SD=6]^2). To check the chain, we show below the traceplot and histogram:





Note: No clear convergence. The chain seems stuck in some values.
$\Longrightarrow$ The chain may not be a good approximation to $\pi(\theta)$.

MCMC: MH – Independence Sampler

• The independence sampler can sometimes work very well but can also work very badly!

• The efficiency depends on how close the jumping distribution is to the posterior.

• Generally speaking, the chain will behave well only if the q(.) proposal distribution has heavier tails than the posterior and has similar shape to π(.) .

MCMC: MH – Acceptance Rates

• It is important to monitor the *acceptance rate* (the fraction of candidate draws that are accepted) of the MH algorithm:

- If the acceptance rate is too high, the chain is likely not *mixing* well -i.e., not moving around the parameter space enough.

- If it is too low, the algorithm is too *inefficient* -i.e., rejecting too many draws.

• In general, the acceptance rate falls as the dimension of $P(\theta|y)$ increases (especially, for highly dependent parameters) resulting in slow moving chains and long simulations.

• Simulation times can be improved by using the single component MH algorithm. Instead of updating the whole θ together, θ is divided in components -say, (β,h)-, with each component updated separately.

MCMC: MH – Acceptance Rates & Tuning

• What is high or low is algortihm specific. One way of finding a '*good*' q(.) is to choose a pdf that gives a particular acceptance rate.

• When we *scale* –i.e., adjust the scale parameters– q(.), say σ, to obtain a particular acceptance rate, we say "we are *tuning* the MH."

• In general, tuning is simple: proposal jump sizes are increased when acceptance rates are high and decreased when rates are low.

• The above mechanism suggests an *optimal* scale parameter –i.e., the proposal explores the parameter space efficiently.

MCMC: MH – Acceptance Rates & Tuning

• For RW MH, Roberts, Gelman and Gilks (1997), with Gaussian proposals and *i.i.d* components of θ, suggest as "good" rate:
- 45% for unidimensional problems.
- **23.4%** in the limit (some theoretical support for this result)
- 25% for 6 dimensions.

• 23.4% is often used in practice to tune q(.).

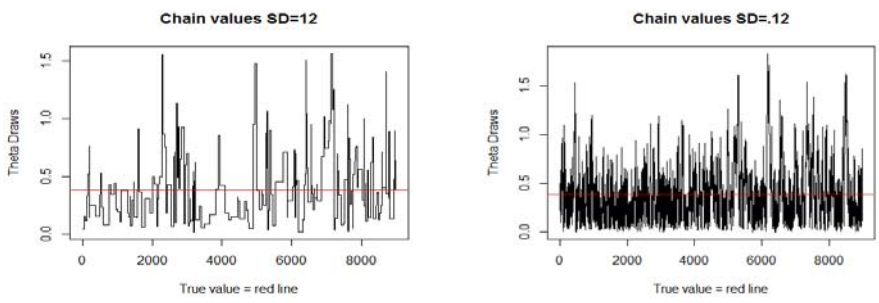
• There is a literature, however, –see, Bedard (2008)– arguing that in many cases 23.4% may be inefficient, for example, hierarchical models.

• For Independent MH, Muller (1993) suggests a 'good' rate is close to 100%.

MCMC: MH – Acceptance Rates & Tuning

Example: Back to the RW Metropolis algorithm to sample from a Gamma(1.7, 4.4) distribution with $q(x,y) \sim N(x, SD^2)$. Before, we used SD=2.

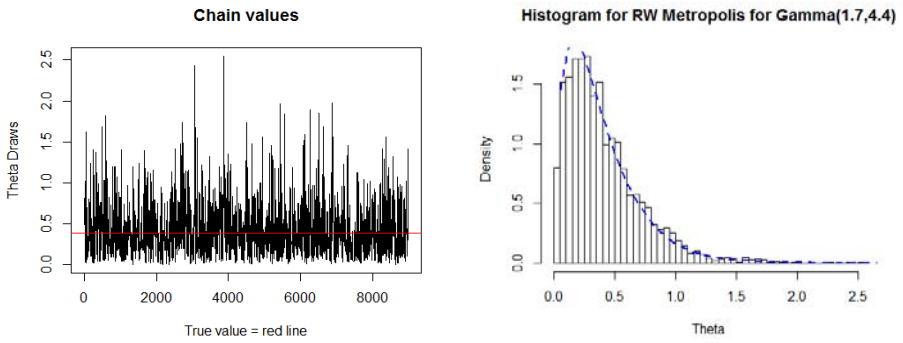
Now, we try two extreme SDs to illustrate the usual trade-off:
- SD=12 (too big), with acceptance rate 2.4% ⇒ inefficient.
- SD=.12 (too small), with acceptance rate 86% ⇒ not mixing well.



MCMC: MH – Acceptance Rates & Tuning

Example: If we use the Roberts et al's (1994) 23.4% acceptance rate as a target, then we adjust SD to get close to it. When SD=2, the acceptance rate was 14.2% (low).

Tuning SD to 1.2, we get a 23.8% accceptance rate. The ED generated looks better (see the traceplot and histogram below):



MCMC: MH – Tuning: Adaptive Method

• Adaptive Method (ad hoc)

- Before the burn-in, we have an *adaptation period*, where the sampler improves the proposal distribution. The adaptive method requires a desired acceptance rate, for example, 30% and tolerance, for example, 10% resulting in an acceptable range of (20%,40%).

- If we are outside the acceptable range, say we reject too much, we scale the proposal distribution, for example, by changing/reducing the spread (say, σ).

- Think of it as a method to find starting values.

• MLwiN uses an adaptive method to construct univariate Normal proposals with an acceptance rate of approximately 50%.

MCMC: MH – Adaptive Method Algorithm

• Run the MH sampler for consecutive batches of 100 iterations. Compare the number accepted, N with the desired acceptance rate, R. Adjust variance accordingly:

$$\text{If } N \leq R, \quad \Rightarrow \sigma_{new} = \sigma_{old} / (2 - \tfrac{N}{R})$$

$$\text{If } N > R, \quad \Rightarrow \sigma_{new} = \sigma_{old} \times (2 - \tfrac{100-N}{100-R})$$

• Repeat this procedure until 3 consecutive values of N lie within the acceptable range and then, *mark* (fixed) this parameter. Check other parameters.

• When all the parameters are marked the adaptation period is over.

Note: Proposal SDs are still modified after being marked until adaptation period is over.

MCMC: MH – Autocorrelation

• A usual problem in poor MCMC performance is high autocorrelation, or "stickiness" in the chain.

• Estimators based on MC samples (based on *independent* draws from the target) perform better than MCMC samples, which are correlated samples.

• The SE of both estimators is given by:

$\text{Var}_{\text{MC}}[\bar{\theta} \mid y] = \text{Var}[\theta \mid y]/M$

$\text{Var}_{\text{MCMC}}[\bar{\theta} \mid y] = \text{Var}[\theta \mid y]/M + 1/M \, \Sigma_{m \neq t} \, E\{(\theta^m - \theta_{\text{true}}) \, (\theta^t - \theta_{\text{true}})\}$

where $\bar{\theta}$ approximates $E[\theta \mid y] = \theta_{\text{true}}$. The 2nd term in $\text{Var}_{\text{MCMC}}[\bar{\theta} \mid y]$ is, in general, positive (& higher than $\text{Var}_{\text{MC}}[\bar{\theta} \mid y]$).

MCMC: MH – Autocorrelation

• Thus, we expect the MCMC approximation to be worse. The higher the autocorrelation, the less information we have in the chain. We may need a very large *M* to get enough information to estimate quantities of interest from $\pi(\theta)$.

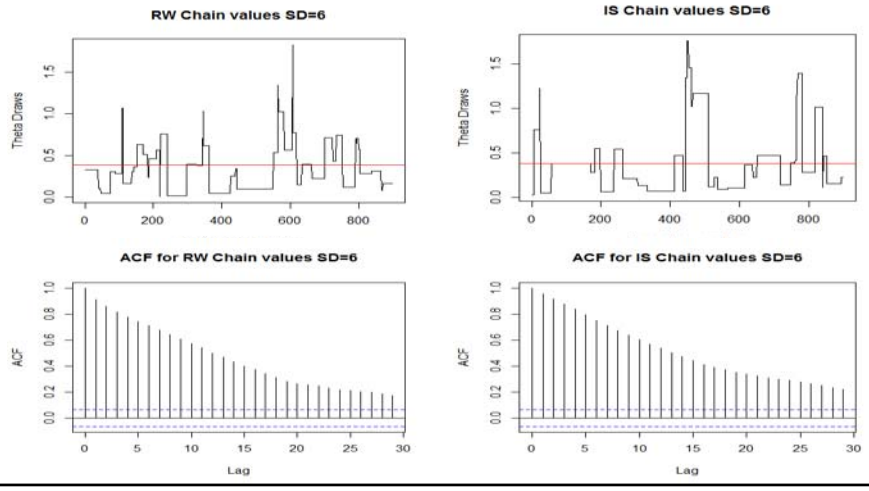
• That is, a chain with high autocorrelation moves around the parameter space Θ slowly, taking a long time to achieve the correct balance of samples to approximate $\pi(\theta)$.

• It is common practice to adjust the level of correlation by adjusting q(.), usually, tuning σ.

• Sample autocorrelation function (ACF) plots are used to determine how much autocorrelation is in the chain.

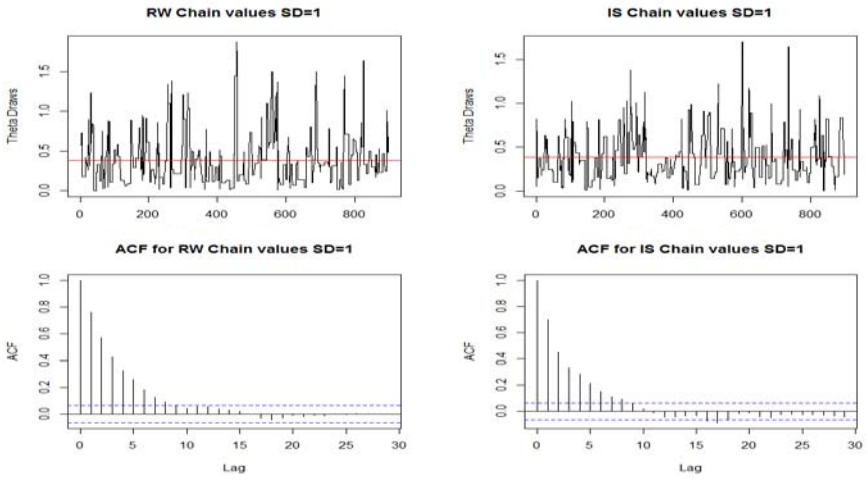
MCMC: MH – Autocorrelation: Tuning

Example: MH algorithm sampling from a Gamma(1.7, 4.4) pdf with q(.) ~ N(., SD^2). We plot the first 1,000 values & ACF (SD=6).



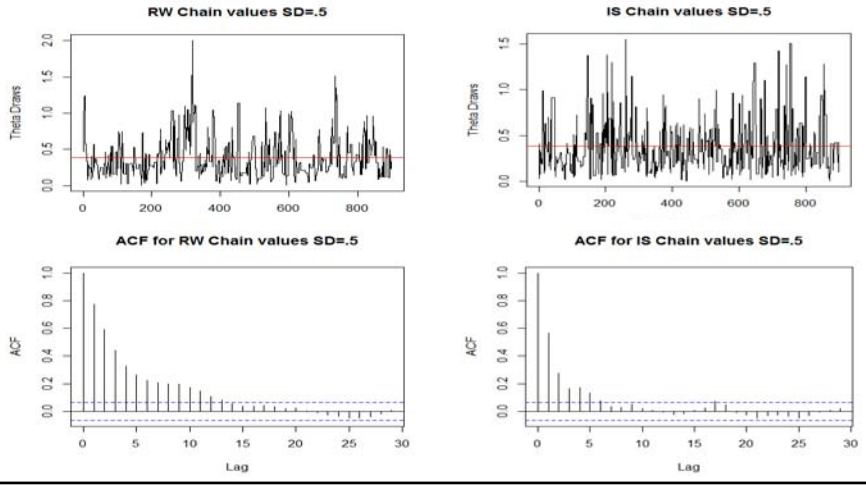
MCMC: MH – Autocorrelation: Tuning

Example (continuation): Now, we plot the first 1,000 values & ACF (SD=1).



MCMC: MH – Autocorrelation: Tuning

Example (continuation): Now, we plot the first 1,000 values & ACF (SD=.5).



MCMC: MH – Autocorrelation - Remarks

• In practice, we adjust the level of correlation by scaling q(.).

• By looking at the previous graphs, it is tempting to reduce σ to get a lower autocorrelation (and faster convergence). But, as $\sigma \to 0$, the acceptance rate goes to 1. That is, the chain never moves (see RW chain with SD=.5).

• That is, there is a trade-off when adjusting σ to control for low autocorrelation; we want σ to be:

- large enough so that the chain moves quickly throughout Θ.
- but not so large σ that the rejection rate is too high.

MCMC: MH – Diagnostics and Convergence

• Similar diagnostics tools as the ones discussed for the Gibbs Sampler.

• Convergence is a practical problem for MCMC methods.

- Converge can be slow ⇒ Let the MH algorithm run.
- There are some formal tests -see Robert and Casella (2004). In the usual complicated setups they tend to have a Type II error problem (accept convergence too much/too quickly) ⇒ Rely on graphs (traceplots & histograms, correlograms).

Practical advise: Run the algorithm until some iteration M^*, where it looks like the chain is stationary. Then, run it M more times to check! Discard the first M^* iterations. Keep the rest to approximate $\pi(\theta)$.

MCMC: MH – Remarks

• MH sampling produces a chain, $\{\theta^{(M)}\}$, with $\pi(.)$ as limiting distribution. The chain allows us to calculate quantities of interest of $\pi(.)$ (moments, C.I., etc.) when *i.i.d.* simulations cannot be used.

• Pros:

- We need less information about $\pi(.)$ than other methods.
- With little tuning, the algorithm works reasonably well.
- Large dimensions problems can be broken into sets of smaller ones.

• Cons:

- Results are only asymptotic (when $M \to \infty$).
- Convergence may be very slow. For practical purposes, the algorithm may not converge.
- Detecting slow convergence may be difficult.

Application 1: Bivariate Normal

• We want to simulate values from

$$f(x) \propto \exp\left\{-(1/2)\left(\mathbf{x'\Sigma^{-1}x}\right)\right\}; \quad \Sigma=\begin{pmatrix}1 & .9\\ .9 & 1\end{pmatrix}$$

• Proposal distribution: RW chain: $y = x + z$, $z \sim$ bivariate Uniform on $(-\delta_i,\delta_i)$, for i=1,2. (δ_i controls the spread)

To avoid excessive move, let δ_1=.75 and δ_2=1.

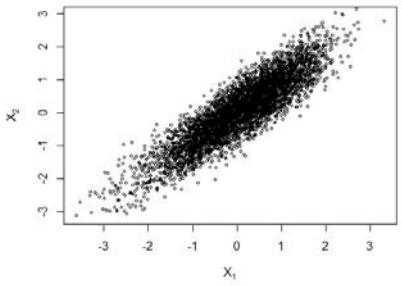
• The probability of move (for a symmetric proposal) is:

$$\alpha(x, y) = \min\left(1, \frac{\exp\{-\frac{1}{2}(y-\mu)'\Sigma^{-1}(y-\mu)\}}{\exp\{-\frac{1}{2}(x-\mu)'\Sigma^{-1}(x-\mu)\}}\right).$$

Application 1: Bivariate Normal (code in R)

```
    sigma00 <- matrix(.9, 2,2);     diag(sigma00) <- 1; inv_sig00 <- solve(sigma00)
    mu00 <- c(0,0);    m <- c(0,0)
# Posterior
target <- function(y){
    targ_l = exp(-.5*t(y-m)%*%inv_sig00 %*% (y-m))
    return(targ_l+.1^20)
}
# RW Chain
run_metropolis_MCMC <- function(startvalue, iterations){
    chain = array(dim = c(iterations+1,2))
    chain[1,] = startvalue
    for (i in 1:iterations){
        proposal = chain[i,] + c(runif(1,-.5,.5),runif(1,-1,1))
        probab = target(proposal)/target(chain[i,])

        if (runif(1,0,1) < probab){
            chain[i+1,] = proposal
        }else{
            chain[i+1,] = chain[i,]
        }
    }
    return(chain)
}
```



Application 2: The Probit Model (Greene)

- The Probit Model:

(a) $y_i^* = \mathbf{x}_i'\boldsymbol{\beta} + \varepsilon_i$

$\varepsilon_i \sim N[0,1]$

(b) $y_i = 1$ if $y_i^* > 0$, 0 otherwise

Consider estimation of $\boldsymbol{\beta}$ and y_i^* (data augmentation)

(1) If y* were observed, this would be a linear regression (y_i would not be useful since it is just $\text{sgn}(y_i^*)$.) We saw in the linear model before, $p(\boldsymbol{\beta} \mid y_i^*, y_i)$

(2) If (only) $\boldsymbol{\beta}$ were observed, y_i^* would be a draw from the normal distribution with mean $\mathbf{x}_i'\boldsymbol{\beta}$ and variance 1. But, y_i gives the sign of y_i^*. $y_i^* \mid \boldsymbol{\beta}, y_i$ is a draw from the truncated normal (above if y=0, below if y=1)

Application 2: The Probit Model (Greene)

• Gibbs sampler for the probit model:

(1) Choose an initial value for $\boldsymbol{\beta}$ (maybe the MLE)
(2) Generate y_i^* by sampling N observations from the truncated normal with mean $\mathbf{x}_i'\boldsymbol{\beta}$ and variance 1, truncated above 0 if $y_i = 0$, from below if $y_i = 1$.
(3) Generate $\boldsymbol{\beta}$ by drawing a random normal vector with mean vector $(\mathbf{X'X})^{-1}\mathbf{X'y}^*$ and variance matrix $(\mathbf{X'X})^{-1}$
(4) Return to 2 10,000 times, retaining the last 5,000 draws - first 5,000 are the 'burn in.'
(5) Estimate the posterior mean of $\boldsymbol{\beta}$ by averaging the last 5,000 draws.

(This corresponds to a uniform prior over $\boldsymbol{\beta}$.)

Aside: Generating Random Draws from *F*(X)

The inverse probability method of sampling random draws:
If F(x) is the CDF of random variable x, then a random draw on x may be obtained as $F^{-1}(u)$ where u is a draw from the standard uniform (0,1).
Examples:
Exponential: $f(x)=\theta\exp(-\theta x)$; $F(x)=1-\exp(-\theta x)$
$x = -(1/\theta)\log(1-u)$
Normal: $F(x) = \Phi(x)$; $x = \Phi^{-1}(u)$
Truncated Normal: $x=\mu_i + \Phi^{-1}[1-(1-u)*\Phi(\mu_i)]$ for $y=1$;
$x= \mu_i + \Phi^{-1}[u\Phi(-\mu_i)]$ for $y=0$.

Example: Simulated Probit

```
? Generate raw data
Sample  ; 1 - 1000 $
Create  ; x1=rnn(0,1) ; x2 = rnn(0,1) $
Create  ; ys = .2 + .5*x1 - .5*x2 + rnn(0,1) ; y = ys > 0 $
Namelist; x=one,x1,x2$
Matrix  ; xx=x'x ; xxi = <xx> $
Calc    ; Rep = 200 ; Ri = 1/Rep$
Probit  ; lhs=y;rhs=x$
? Gibbs sampler
Matrix  ; beta=[0/0/0] ; bbar=init(3,1,0);bv=init(3,3,0)$$
Proc    = gibbs$
Do for  ; simulate ; r =1,Rep $
Create  ; mui = x'beta ; f = rnu(0,1)
        ; if(y=1)  ysg = mui + inp(1-(1-f)*phi( mui));
           (else) ysg = mui + inp(     f *phi(-mui)) $
Matrix  ; mb = xxi*x'ysg ; beta = rndm(mb,xxi)
        ; bbar=bbar+beta ; bv=bv+beta*beta'$
Enddo   ; simulate $
Endproc $
Execute ; Proc = Gibbs $ (Note, did not discard burn-in)
Matrix  ; bbar=ri*bbar ; bv=ri*bv-bbar*bbar' $
Matrix  ; Stat(bbar,bv); Stat(b,varb) $
```

Application 2: Simulated Probit (Greene)

- MLE vs Gibbs Sampler

```
--> Matrix  ; Stat(bbar,bv); Stat(b,varb) $
+---------------------------------------------------+
|Number of observations in current sample =    1000 |
|Number of parameters computed here       =       3 |
|Number of degrees of freedom             =     997 |
+---------------------------------------------------+
+---------+-------------+---------------+--------+---------+
|Variable | Coefficient  | Standard Error |b/St.Er.|P[|Z|>z] |
+---------+-------------+---------------+--------+---------+
 BBAR_1         .21483281      .05076663     4.232   .0000
 BBAR_2         .40815611      .04779292     8.540   .0000
 BBAR_3        -.49692480      .04508507   -11.022   .0000
+---------+-------------+---------------+--------+---------+
|Variable | Coefficient  | Standard Error |b/St.Er.|P[|Z|>z] |
+---------+-------------+---------------+--------+---------+
 B_1            .22696546      .04276520     5.307   .0000
 B_2            .40038880      .04671773     8.570   .0000
 B_3           -.50012787      .04705345   -10.629   .0000
```

Application 3: Stochastic Volatility (SV)

• Gibbs Algorithm for Estimating SV Model –from K&S (2004).

$$\Delta r_t - (\hat{a}_0 + \hat{a}_1 r_{t-1}) \equiv RES_t$$

$$RES_t = \sqrt{h_t r_{t-1}^{2\alpha}}\,\varepsilon_t, \qquad \alpha = 0.5$$

$$\ln(h_t) = \omega + \phi_1 \ln(h_{t-1}) + \sqrt{\sigma_\eta^2}\,\eta_{t-1}$$

- In the SV model, we estimate the parameter vector and 1 latent variable: $\theta = \{\omega, \sigma_\eta, \phi_1,\}$ and $H_t = \{h_1,...,h_t\}$.
- Parameter set therefore consists of $\Theta = \{H_t, \theta\}$ for all t.

• Using Bayes theorem to decompose the joint posterior density as follows.

$$f(H_n, \theta) \propto f(Y_n | H_n) f(H_n |, \theta) f(\theta)$$

Application 3: Stochastic Volatility (SV)

$$f(H_n, \theta) \propto f(Y_n | H_n) f(H_n |, \theta) f(\theta)$$

• Next draw the marginals $f(H_t \mid Y_t, \theta)$, and $f(\theta \mid Y_t, H_t)$, using a Gibbs sampling algorithm:

Step 1: Specify initial values $\theta^{(0)} = \{\omega^{(0)}, \sigma_\eta^{(0)}, \phi^{(0)}\}$. Set $i = 1$.

Step 2:

Draw the underlying volatility using the multi-move simulation sampler –see, De Jong and Shephard (1995)–, based on parameter values from **step 1**.

- The multi-move simulation sampler draws H_t for all the data points as a single block. Recall we can write:

$$\ln(RES_t^2) = \ln(h_t) + \ln(r_{t-1}) + \ln(\varepsilon_t^2) \qquad \text{(A-1)}$$

Application 3: Stochastic Volatility (SV)

$$\ln(RES_t^2) = \ln(h_t) + \ln(r_{t-1}) + \ln(\varepsilon_t^2) \qquad \text{(A-1)}$$

where $\ln(\varepsilon_t^2)$ can be approximated by a mixture of seven normal variates –see Chib, Shephard, and Kim (1998).

$$\ln(\varepsilon_t^2) = z_t$$

$$f(z_t) = \sum_{i=1}^{7} f_N\left(z_i \middle| m_i - 1.2704, v_i^2\right); \qquad i = \{1,2,....7\} \qquad \text{(A-2)}$$

- Now, (A-1) can be written as

$$\ln(RES_t^2) = \ln(h_t) + \ln(r_{t-1}) + \left[z_t \middle| k_t = i\right] \qquad \text{(A-3)}$$

where k_t is one of the seven underlying densities that generates z_t.

- Once the underlying densities k_t, for all t, are known, (A-3) becomes a deterministic linear equation and along with the SV model can be represented in a linear state space model.

Application 3: Stochastic Volatility (SV)

- If interested in estimating α as a free parameter, rewrite (A-1) as

$$\ln(RES_t^2) = \ln(h_t) + 2\alpha \ln(r_{t-1}) + \ln(\varepsilon_t^2) \qquad \text{(A-1)}'$$

Then, estimate α approximating $\ln(\varepsilon_t^2)$ by a lognormal distribution. Once α is known, follow (A-3) and extract the latent volatility.

Step 3:

Based the on output from **steps 1** and **2**, the underlying k_t in (A-3) is sampled from the normal distribution as follows:

$$f\left[z_{t=i} \middle| \ln(y_t^2), \ln(h_t)\right] \propto q_i f_N\left(z_i \middle| \ln(h_t) + m_i - 1.2704, v_i^2\right) \quad i \le k \qquad \text{(A-4)}$$

For every observation *t*, we draw the normal density from each of the seven normal distributions $\{k_t = 1,2,..,7\}$. Then, we select a "k" based on draws from uniform distribution.

Application 3: Stochastic Volatility (SV)

Step 4:

Cycle through the conditionals of $\theta = \{\omega, \sigma_\eta, \phi_1\}$ as in Chib (1993), using output from **steps 1-3**. Recall that $f(\theta)$ can be decomposed as:

$$f(\theta|Y_n,H_n) \propto f(\omega|Y_n,H_n,\theta_{-\omega}) f(\sigma_\eta^2|Y_n,H_n,\theta_{-\sigma^2}) f(\phi|Y_n,H_n,\theta_{-\phi}) \qquad \text{(A-5)}$$

where θ_{-j} refers to the θ parameters excluding the jth parameter.

- The prior distributions and conditional posteriors (normal for ω and ϕ, inverse gamma for σ_η^2) are described in the previous SV example – see also Chib (1993). You need to specify the prior means and standard deviations.

Step 5: Go to **step 2**. (Now, Set $i = 2$.)

Conclusions (Greene)

- Bayesian vs. Classical Estimation
  - In principle, different philosophical views and differences in interpretation
  - As practiced, just two different algorithms
  - The religious debate is a red herring –i.e., misleading.
- Gibbs Sampler. A major technological advance
  - Useful tool for both classical and Bayesian
  - New Bayesian applications appear daily

Standard Criticisms (Greene)

- Of the Classical Approach
  - Computationally difficult (ML vs. MCMC)
  - It is difficult to pay attention to heterogeneity, especially in panels when N is large.
  - Responses: None are true. See, e.g., Train (2003, Ch. 10)

- Of Classical Inference in this Setting
  - Asymptotics are "only approximate" and rely on "imaginary samples." Bayesian procedures are "exact."
  - Response: The inexactness results from acknowledging that we try to extend these results outside the sample. The Bayesian results are "exact" but have no generality and are useless except for this sample, these data and this prior. (Or are they? Trying to extend them outside the sample is a distinctly classical exercise.)

Standard Criticisms (Greene)

- Of the Bayesian Approach
  - Computationally difficult.
  - Response: Not really, with MCMC and Metropolis-Hastings
  - The prior (conjugate or not) is a hoax. It has nothing to do with "prior knowledge" or the uncertainty of the investigator.
  - Response: In fact, the prior usually has little influence on the results. (Bernstein and von Mises Theorem)

- Of Bayesian 'Inference'
  - It is not statistical inference
  - How do we discern any uncertainty in the results? This is precisely the underpinning of the Bayesian method. There is no uncertainty. It is 'exact.