Lecture 17 – Part 1
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:

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

© RS 2020 – Not to be posted/shared online without written authorization.

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:

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

© RS 2020 – Not to be posted/shared online without written authorization.

Bayesian Econometrics: Introduction

• We want to get $P(\theta|y) \propto P(y|\theta)$ x $P(\theta)$. 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 controversy starts when we model unknown 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.)

© RS 2020 – Not to be posted/shared online without written authorization.

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:*

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

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:

$$P(S|T = good) = \frac{P(S|T=good)\,P(S)}{P(T=good)} = \frac{\mathbf{.80} * .40}{\mathbf{.80} * .40 + .20 * .60} = .7272$$

© RS 2020 – Not to be posted/shared online without written authorization.

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

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

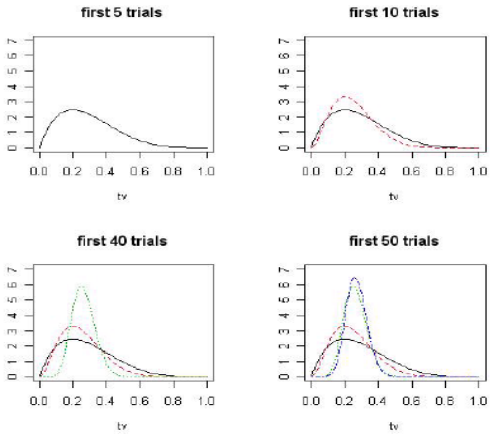
and after 40 trials the posterior is

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

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.

© RS 2020 – Not to be posted/shared online without written authorization.

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 success θ. That is, $y_1, y_2, ..., y_T \sim$ *i.i.d. Bin*(1,θ) –i.e., *Bernouilli*.

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, $\sum_{i=1}^{T} y_i$, is a sufficient statistic for θ & p($y_1, y_2, ..., y_T \,|\, \theta$). Moreover, $\sum_{i=1}^{T} y_i$ follows a *Bin(T, θ)*. These results are to be used later.

Likelihood: Normal

• Suppose $y_i \sim$ *i.i.d.* N(θ, σ²), 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(θ, σ²), 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,

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

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

© RS 2020 – Not to be posted/shared online without written authorization.

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 $\sum_{i=1}^{T} y_i \sim Bin(T, \theta)$. Suppose we observe $\{\sum_{i=1}^{T} y_i = s\}$. Suppose from our prior information, we assume $\theta \sim$ Beta(α, β). That is,

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

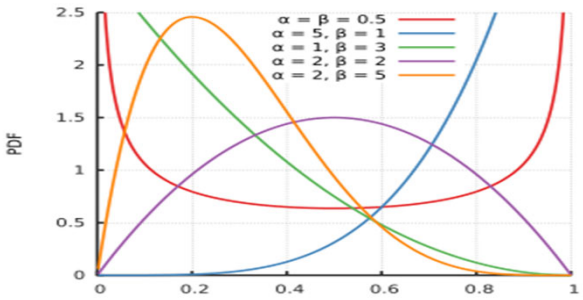
But, we could have assumed something different. For example, our prior information for θ tells us that all subintervals of [0,1] with the same length also have the same probability:

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

which leads to a uniform for θ $\Rightarrow$ 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] = $\alpha/(\beta + \alpha)$

Var[θ] = $\alpha\,\beta/[(\beta + \alpha)^2\,(\alpha + \beta + 1)]$ = E[θ] (1 - E[θ])/$(\alpha + \beta + 1)$

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

© RS 2020 – Not to be posted/shared online without written authorization.

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.

© RS 2020 – Not to be posted/shared online without written authorization.

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: Suppose 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?

© RS 2020 – Not to be posted/shared online without written authorization.

Priors: Diffuse Prior - Example

Example: Suppose $y_1, y_2, ..., y_T \sim$ *i.i.d. Bin*(1, θ). We know that $\sum_{i=1}^{T} y_i \sim$ *Bin(T, θ)*. Suppose we observe $\{\sum_{i=1}^{T} y_i = s\}$. Our prior information is not very good, it points towards a diffuse prior.

We formalize this information with a uniform distribution: P(θ)=1 for all θ ∈ [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(θ) = 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 ⇒ P(θ) ∝ $[I(\theta)]^{1/2}$, where I(θ) is the Fisher information for θ:

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

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

© RS 2020 – Not to be posted/shared online without written authorization.

Priors: Jeffreys' Non-informative Prior

Example: Suppose $y_1, y_2, ..., y_T \sim$ *i.i.d. Bin*(1, θ). Then, $\sum_{i=1}^{T} y_i \sim Bin(T, \theta)$, with a log-likelihood:

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

Then,

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

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

© RS 2020 – Not to be posted/shared online without written authorization.

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 $y_1, y_2, ..., y_T \sim$ *i.i.d. Bin*(1, θ). Then, $\sum_{i=1}^{T} y_i \sim$ *Bin*(*T*, θ). We observe $\{\sum_{i=1}^{T} y_i = s\}$. 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($s + \alpha$, $T + \beta - s$).

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.

© RS 2020 – Not to be posted/shared online without written authorization.

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:

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

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:

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

© RS 2020 – Not to be posted/shared online without written authorization.

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 $P(\boldsymbol{\varphi}) = P(\theta \mid h)\, P(h)$.

For $\theta \mid h$, we assume $P(\theta \mid h) \sim N(\theta_0, \sigma_0^2)$, 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 $\alpha = T/2$ & $\lambda = 1/(2\eta^2) = \Phi/2$ 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:

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

© RS 2020 – Not to be posted/shared online without written authorization.

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?

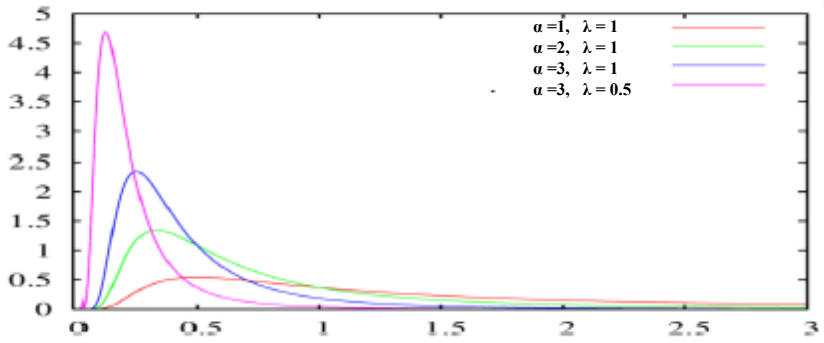
(1) $p(\sigma^2) = 0$ for $\sigma^2 < 0$.

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

(3) *Conjugate prior* ⇒ 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.

© RS 2020 – Not to be posted/shared online without written authorization.

Prior Information: Intuition

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

© RS 2020 – Not to be posted/shared online without written authorization.

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 ($\boldsymbol{y}$). We characterize this with the *posterior distribution*:

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

• The posterior is the basis of Bayesian estimation. It takes into account the data (say, $\boldsymbol{y}$ & $\mathbf{X}$) and our prior distribution (say, θ_0).

• $P(\theta|\boldsymbol{y})$ 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, $E[\theta|\boldsymbol{y}]$, is the *optimal Bayesian estimator* of θ.

© RS 2020 – Not to be posted/shared online without written authorization.

Posterior: Optimal Estimator

• We assume a loss function, $g(\theta, \hat{\theta})$, where $\hat{\theta}$ is an estimate. Let $\hat{\theta}$ solve the minimization 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: $g(\theta, \hat{\theta})=(\theta - \hat{\theta})^2$

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

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

• 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, $\sum_{i=1}^{T} y_i \sim$ *Bin*(T, θ). We observe $\{\sum_{i=1}^{T} y_i = s\}$.

• **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 recognize $P(\theta \mid Y=s)$ as a Beta (up to a constant), with $\alpha = (s + 1)$ & $\beta = (T - s + 1)$.

© RS 2020 – Not to be posted/shared online without written authorization.

Posterior: Example – Binomial-Uniform

Example (continuation):

We can derive $c(s)$ since $P(\theta \mid y)$ 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

• $P(\theta \mid y)$ is a pdf. For the simple case, the one parameter θ, it can be graphed. But, if $\boldsymbol{\theta}$ 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.

© RS 2020 – Not to be posted/shared online without written authorization.

Posterior: Presentation of Results - Example

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

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

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

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

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}:

$E[\theta \mid s = 15] = \alpha/(\beta + \alpha) = 16/27 =$ **.5927**

$\text{Var}[\theta \mid s = 15] = \alpha\,\beta/[(\beta + \alpha)^2\,(\alpha + \beta + 1)]$
$= 16*11/[(27)^2 *(28)] =$ **0.00862**

$P_{\theta \mid s}\,(0.55 > \theta > 0.65 \mid s = 15) =$ **0.4014529**. (=pbeta(.65,16,11)-pbeta(.55,16,11), in R)

• Check normal approximation, with N(**.5927**, sqrt(**.009**)), in R:

```
> pnorm(.65, .5927, sqrt(.009)) - pnorm(.55, .5927, sqrt(.009))
[1] 0.4007565
```

© RS 2020 – Not to be posted/shared online without written authorization.

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($s+1$, $T-s+1$). 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 $\{\sum_{i=1}^{T} y_i = s\}$.

• **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 $\text{Beta}(\alpha+s, \beta+T-s)$. (Not a surprise, we used a *conjugate prior*!). We 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($\alpha=1$, $\beta=1$).

© RS 2020 – Not to be posted/shared online without written authorization.

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 | $\sum_{i=1}^{T} y_i = s$] 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" ($\Rightarrow \approx$ "prior sample size.")
Then, the prior expectation is $\alpha/(\beta+\alpha)$.

• Role of T: As T grows $\Rightarrow$ Data dominates.
$\Rightarrow$ E[θ | $y = s$] $\approx s/T$
$\Rightarrow$ Var[θ | $y = s$] $\approx s/T^2 * [1 - (s/T)]$

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.

© RS 2020 – Not to be posted/shared online without written authorization.

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 degree of confidence in our prior.

• 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:

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

© RS 2020 – Not to be posted/shared online without written authorization.

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

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

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

• 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.

© RS 2020 – Not to be posted/shared online without written authorization.

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)
    }
}
```

© RS 2020 – Not to be posted/shared online without written authorization.

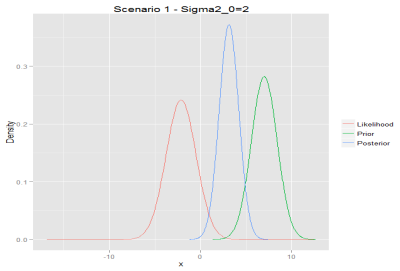
Posterior: Bayesian Learning

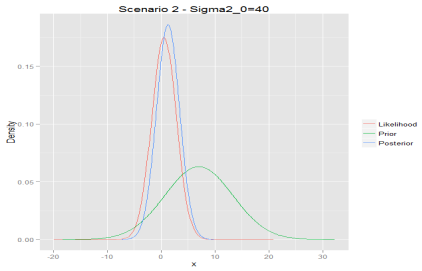
Example (continuation):

```
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, σ0^2 =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, σ0^2 =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 $\widehat{\mu_t}$. Let $m_1, m_2, ..., m_T$ be any numbers.

• Define

$$S = \sum_{i=1}^{T}(x_t - m_t)^2$$
$$\theta = 1 - [(T-2)\,\sigma^2/S]$$
$$m_t^* = \theta\,\widehat{\mu_t} + (1-\theta)\,m_t$$

Theorem: Under the previous assumptions,

$$E[\sum_{i=1}^{T}(x_t - m_t^*)^2] < E[\sum_{i=1}^{T}(x_t - \widehat{\mu_t})^2]$$

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.

© RS 2020 – Not to be posted/shared online without written authorization.

Predictive Posterior

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

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

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

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

Then,

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

$P(z \mid y)$ is the *predictive posterior distribution,* the distribution of new (unobserved) observations. It is equal to the conditional (over the posterior of $\theta \mid y$) 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:
$P(Z=good \mid S) = \mathbf{.95}$
$P(Z=good \mid S^C) = \mathbf{.10}$
(We have previously determined: $P(S \mid T = g) = \mathbf{0.72727}$.)

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 $T=good$:

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

Note: Z and T are conditionally independent.

© RS 2020 – Not to be posted/shared online without written authorization.

Predictive Posterior: Example 2

Example: We have $y_1, y_2, ..., y_{T=25} \sim$ *i.i.d. Bin*$(1, \theta)$. Let $\sum_{i=1}^{T} y_i = s$. We derive the predictive posterior of new data, Y*, as:

$P(Y^*=1 \mid y_1, y_2, ..., y_T) = E[\theta \mid y = s] = (\alpha+s)/(\alpha+\beta+T)$

$P(Y^*=0 \mid y_1, y_2, ..., y_T) = 1 - P(Y^*=1 \mid y = s] = (\beta+T-s)/(\alpha+\beta+T)$

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:

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

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: $y_j \sim N(\mu_j, \sigma^2_j)$.

© RS 2020 – Not to be posted/shared online without written authorization.

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

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

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

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

$\text{Cov}[\boldsymbol{\mu} \mid y_1, y_2, \ldots, y_T, \boldsymbol{\Sigma}] = \boldsymbol{\Lambda}_N = \mathbf{A}_N^{-1} = (\boldsymbol{\Lambda}_0^{-1} + N\, \boldsymbol{\Sigma}^{-1})^{-1}$

$\text{E}[\boldsymbol{\mu} \mid y_1, y_2, \ldots, y_T, \boldsymbol{\Sigma}] = \boldsymbol{\mu}_N = \mathbf{A}_N^{-1}\, \boldsymbol{b}_N = \boldsymbol{\Lambda}_N\, (\boldsymbol{\Lambda}_0^{-1}\, \boldsymbol{\mu}_0 + N\, \boldsymbol{\Sigma}^{-1}\, \bar{\boldsymbol{y}})$

• 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.

© RS 2020 – Not to be posted/shared online without written authorization.

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

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!

© RS 2020 – Not to be posted/shared online without written authorization.

Multivariate Models: MVN – IW Prior for Σ

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

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

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($\mathbf{\Sigma} \mid \mathbf{y}_1$,..., $\mathbf{y}_N$, $\boldsymbol{\mu}$)

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

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

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,

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

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

© RS 2020 – Not to be posted/shared online without written authorization.

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., $\log(\boldsymbol{S})\sim N(\boldsymbol{\xi}, \boldsymbol{\Lambda})$– 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

© RS 2020 – Not to be posted/shared online without written authorization.

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.

• 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.

© RS 2020 – Not to be posted/shared online without written authorization.

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.

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.

© RS 2020 – Not to be posted/shared online without written authorization.

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$\{s^2$ (**X**' **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.

Bayesian vs. Classical: Interpretation

• Asymptotics

– In classical statistics, we use the LLN and the CLT. Typical use of the LLN:

Consider a random sample, X_1, X_2, ..., X_N. Then,

as $N \to \infty$, $\bar{h}(X) \to E[h(X)]$.

– In Bayesian statistics, we use the LLN and the CLT too. But, with the size of the simulation, $M, \to \infty$. Typical use of the LLN:

Consider a random sample, θ^1, θ^2, θ^3, ..., θ^M. Then,

as $M \to \infty$, $\bar{h}(\theta) \to E[h(\theta)]$.

Note: In Bayesian statistics, the asymptotics are based on M (size of simulation determined by researcher) not on N (the sample size).

© RS 2020 – Not to be posted/shared online without written authorization.

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

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

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

- 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

$$(b_k - 0)/(sV_{kk}^{1/2}) = b_{T-k}$$ --the basis for hypothesis tests.

Linear Model: Bayesian Setup

• For the normal linear model, we assume $f(y_t \mid \mu_t, \sigma^2)$:

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

where $\mu_t = \beta_0 + \beta_1 \boldsymbol{x}_{1t} + \ldots + \beta_k \boldsymbol{x}_{kt} = \boldsymbol{x}_t \boldsymbol{\beta}$

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

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

$$f(\boldsymbol{\beta}, \sigma^2 \mid \boldsymbol{y}, \mathbf{X}) \propto \prod_{t=1}^{T} f(y_t \mid \mu_t, \sigma^2) \times f(\boldsymbol{\beta}, \sigma^2)$$

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

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

© RS 2020 – Not to be posted/shared online without written authorization.

Linear Model: Likelihood

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

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

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

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

$$\Rightarrow \text{TSS} = (\boldsymbol{y} - \mathbf{X}\mathbf{b})'(\boldsymbol{y} - \mathbf{X}\mathbf{b}) + (\beta - \mathbf{b})'\mathbf{X}'\mathbf{X}(\beta - \mathbf{b}) - 2(\beta - \mathbf{b})'\mathbf{X}'(\boldsymbol{y} - \mathbf{X}\beta)$$

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

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

Linear Model: Likelihood

• The likelihood can be factorized as:

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

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

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.

© RS 2020 – Not to be posted/shared online without written authorization.

Linear Model: Prior Distribution for β

• We choose a conjugate MVN prior for β: $f(\beta) \sim N(\boldsymbol{m}, \boldsymbol{\Sigma})$:

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

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

– $\boldsymbol{\Sigma}$ 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 $\boldsymbol{\Sigma}$).

• 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 an IG. Then, $\boldsymbol{h} = 1/\sigma^2$ 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_0, λ_0): $\alpha_0 = T/2$ 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,

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

© RS 2020 – Not to be posted/shared online without written authorization.

Linear Model: Joint Prior Distribution for θ

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

In this case, we can write P(θ) = P($\beta \mid h = \sigma^{-2}$) 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($\theta \mid y$) $\propto$ P($y \mid \theta$) P(θ).

Linear Model: Assumptions

• So far, we have made the following assumptions:

– **Likelihood**: Data is *i.i.d.* Normal: $y_t \sim N(\mu_t, \sigma^2)$ for $t = 1, ..., T$

– DGP for μ_t is known: $\mu_t = \beta_0 + \beta_1 x_{1t} + \ldots + \beta_k x_{kt} = x_t \beta$

– **X** is fixed.

– **Prior distributions**: $h = 1/\sigma^2 \sim \Gamma(\alpha_0, \lambda_0)$ & $\beta \sim N(m, \Sigma)$.

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

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

$$f(\Psi, \beta, \sigma \mid \mathbf{y}, \mathbf{X}) = f(\beta, \sigma \mid \mathbf{y}, \mathbf{X})\, f(\Psi \mid \mathbf{y}, \mathbf{X})$$

$$\Rightarrow f(\beta, \sigma \mid \mathbf{y}, \mathbf{X}) \propto f(\beta, \sigma)\, f(\mathbf{y} \mid \beta, \sigma, \mathbf{X})$$

© RS 2020 – Not to be posted/shared online without written authorization.

Linear Model: Joint Posterior Distribution for θ

• **Posterior**: *Likelihood* x *Prior*.

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

• Then, simple algebra delivers:

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

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 $P(\boldsymbol{\theta} \mid \boldsymbol{y})$. 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* $f(\boldsymbol{\beta} \mid \boldsymbol{y}, \mathbf{X})$ and $f(h \mid \boldsymbol{y}, \mathbf{X})$.

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

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

© RS 2020 – Not to be posted/shared online without written authorization.

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

• A little bit of algebra delivers:

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

$$\propto \exp\{\boldsymbol{\beta}'(\frac{\mathbf{X}'\mathbf{y}}{\sigma^2} + \Sigma^{-1}\mathbf{m}) - \frac{1}{2}\boldsymbol{\beta}'(\frac{\mathbf{X}'\mathbf{X}}{\sigma^2} + \Sigma^{-1})\boldsymbol{\beta}]\}$$

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

Then, we recognize the conditional posterior as proportional to an MVN with:

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

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

• That is, $f(\beta \mid \mathbf{y}, \mathbf{X}, h)$ is $N_k(\mathbf{m_n}, \Sigma_n)$.

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

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

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

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

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

© RS 2020 – Not to be posted/shared online without written authorization.

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 $\boldsymbol{m}$ 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 assumptions:
– We assumed an *i.i.d.* normal distribution for $(\boldsymbol{y}\,|\,\mathbf{X}, \sigma^2)$.
– We picked a normal prior for $\boldsymbol{\beta}$ ($\Rightarrow$ 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.

© RS 2020 – Not to be posted/shared online without written authorization.

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,

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

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, $\{\mathbf{y}', \mathbf{X}'\}$, of *T'* observations, which are independent of the current sample, $\{\mathbf{y}, \mathbf{X}\}$.

• 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 $\{\mathbf{y}', \mathbf{X}'\}$:

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

© RS 2020 – Not to be posted/shared online without written authorization.

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

• Again, 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):

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

© RS 2020 – Not to be posted/shared online without written authorization.

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,

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

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

© RS 2020 – Not to be posted/shared online without written authorization.

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 **β**, conjugate gamma prior for σ^2

Joint Posterior

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

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

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:

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

Multivariate t with mean **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.

© RS 2020 – Not to be posted/shared online without written authorization.

Presentation of Results

• $P(\theta \mid y)$ 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) $E(\theta \mid y) = \int \theta \, P(\theta|y) \, d\theta$ -- posterior mean

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

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

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

$$I = \int f(\theta) \, d\theta = \sum_{i=1}^{n} w_i \, \theta_i$$

Presentation of Results: MC Integration

• In the Math Review, we covered different numerical integration methods (trapezoid rule, Gaussian 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):

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

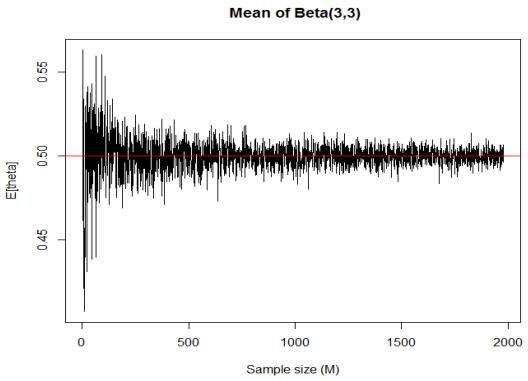
or via Monte Carlo methods (R Code):

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

© RS 2020 – Not to be posted/shared online without written authorization.

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.

© RS 2020 – Not to be posted/shared online without written authorization.

MC Integration: Plug-in estimator

• Idea: We start with a posterior, $P(\theta \mid y)$: $\pi(\theta) = P(y \mid \theta)\, P(\theta)/P(y)$.
We want to get moments of some function of θ, say

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

• If $\theta^{(M)} = \{\theta^1, \theta^2, \theta^3, \ldots, \theta^M\}$ 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, by CLT:

$$\bar{\theta} \sim N(\theta, \text{Var}[\theta \mid y]/M).$$

We approximate the $\sigma^2 = \text{Var}[\theta \mid y]$:

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

$\Rightarrow$ MC SE$[\bar{\theta}] = \sqrt{\hat{\sigma}^2/M}$.

We can select M to give us a desired precision relative to the posterior moment we are interested.

© RS 2020 – Not to be posted/shared online without written authorization.

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 $E[\theta \mid y]$ and its MC estimate to be less than 0.005 with high probability, we need to increase M such that

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

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 $P(\theta \mid y)$. To do this, we need $P(\theta \mid y)$ 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 $P(\theta \mid y)$ 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).

© RS 2020 – Not to be posted/shared online without written authorization.

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 $\theta^{(M)} = \{\theta^1, \theta^2, \theta^3, \ldots, \theta^M\}$ is a random sample from $q(\theta)$, then

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

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 $|\theta^m|$.

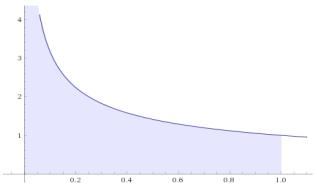
• IS can be turned into "importance sampling resampling" by using an additional resampling step based on the weights.

© RS 2020 – Not to be posted/shared online without written authorization.

MC Integration: IS - Example

Example: We want to use importance sampling (IS) to evaluate the integral $x^{-0.5}$ over the range [0,1] (we know the answer: 2):

```
# 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) )
```

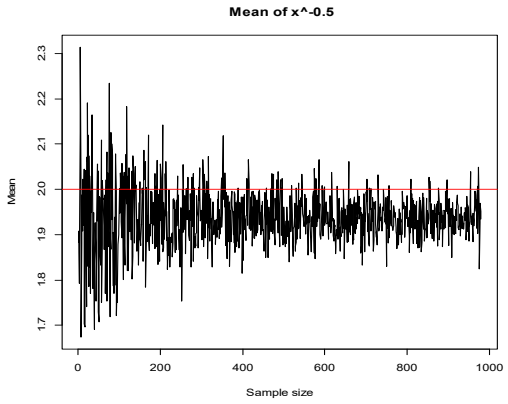


```
# Importance sampling Monte Carlo with an exponential
w <- function(x)   dunif(x, 0.001, 1)/dexp(x,rate=lambda) * pexp(1, rate=lambda)   # [pi(x)/q(x)]
h_f <- function(x)   x^(-0.5)                                                      # h(x)
X <- rexp(M,rate=lambda)
X <- X[X<=1]
Y.h <- w(X)*h_f(X)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

© RS 2020 – Not to be posted/shared online without written authorization.

MC Integration: IS – Importance weights

• If $\pi(\theta)$ is improper, $w(\theta^m)$ is normalized by $\sum_{m=1}^{M} w_m(\theta^m)$ (*normalized* w's.)

Example: Suppose $y \sim$ 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(θ) 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[θ | y] = **0.558493**.

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
```

© RS 2020 – Not to be posted/shared online without written authorization.