Introduction to Bayesian Estimation

How do Classical and Bayesian Analysis Differ?

Consider a simple model:

$y_t = \mu + \varepsilon_t$ where $t = 1, 2, \ldots, T$

$\varepsilon_t \sim N(0, \sigma^2)$

Assume σ^2 is known $\Longrightarrow$ we want to estimate μ

$\hat{\mu} = \frac{1}{T}\sum_{t=1}^{T} y_t$ $\qquad\qquad$ $\hat{\mu} \sim N(\mu, \frac{\sigma^2}{T})$

95% confidence interval: $\left[\hat{\mu} - 1.96\frac{\sigma}{\sqrt{T}}, \hat{\mu} + 1.96\frac{\sigma}{\sqrt{T}}\right]$

How do Classical and Bayesian Analysis Differ?

1. Classical analysis
   - μ is a fixed, unknown quantity $\Longrightarrow$ "true value"
   - The estimator $\hat{\mu}$ is a random variable and is evaluated via repeated sampling $\Longrightarrow$ the interval we constructed will contain the true value in 95% of cases if we estimate $\hat{\mu}$ for thousand different samples taken from a population with given μ and σ^2
   - The estimator $\hat{\mu}$ is "best" in the sense of having the highest probability of being close to the true μ
     $\Longrightarrow$ Probability is objective and is the limit of the relative frequency of an event.

How do Classical and Bayesian Analysis Differ?

2. Bayesian analysis
   - μ is treated as a random variable $\Longrightarrow$ it has a probability distribution
   - The distribution summarizes our knowledge about the model parameter $\Longrightarrow$ 2 sources of information:
     - Prior information (before seeing the data): subjective belief about how likely different parameter values are
     - Sample information: leads researcher to revise/update his prior beliefs
   - Probabilities are subjective and *not* necessarily related to the relative frequency of an event.
   - Explicit use of probabilities to quantify uncertainty.

Key Ingredients for Bayesian Analysis

1. Probabilities
   ⇨ Review some probability rules to derive Bayes' rule

2. Initial information
   ⇨ What is the reason for using prior information?
   ⇨ How to specify a prior distribution for parameters?

3. How to combine data and non-data (prior) information?
   ⇨ Bayesian estimation in practice

Some Rules of Probability

Consider two random variables: A and B

The rules of probability imply: $\boxed{p(\mathrm{A, B}) = p(\mathrm{A} \mid \mathrm{B})\, p(\mathrm{B})}$

Where
- $p(\mathrm{A, B})$ is the ***joint*** probability of A and B
- $p(\mathrm{A} \mid \mathrm{B})$ is the probability of A occurring ***conditional*** on B having occurred
- $p(\mathrm{B})$ is the ***marginal*** probability of B

Alternatively, we can reverse the roles of A and B so that:

$$\boxed{p(\mathrm{A, B}) = p(\mathrm{B} \mid \mathrm{A})\, p(\mathrm{A})}$$

Bayes' Rule

Equating the two expressions for the joint probability of A and B provides us with *Bayes' rule*:

$$\boxed{p(B|A) = \frac{p(A|B)p(B)}{p(A)}}$$

Let's map this rule into a simple regression model where we want to learn about a parameter $\boldsymbol{\theta}$ given the data $\mathbf{y}$:

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

A Closer Look at Each Component

Key object of interest: $\boldsymbol{p(\theta \mid y)}$

- $\boldsymbol{p(y)}$ $\Rightarrow$ marginal data density
  Since we are interested in learning about θ, we can ignore $p(y)$ since it does not involve θ.
- $\boldsymbol{p(\theta)}$ $\Rightarrow$ prior density
  It does not depend on the data y; instead, it contains non-data information about θ.
- $\boldsymbol{p(y \mid \theta)}$ $\Rightarrow$ likelihood function
  It is the density of the data conditional on the parameters.

The Posterior Distribution

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

"The posterior is proportional to the likelihood times the prior."

- The posterior summarizes all we know about θ *after* seeing the data.
  - ⇨ The posterior combines both data and non-data information.
- The equation can be viewed as an updating rule where the data allow us to update our prior views about θ.

Skills for Bayesian Inference

Bayesian inference requires a good knowledge of:

- Probability distributions
  - to formulate prior distributions
  - to generate draws from them
  - to analyze posterior distributions
- Numerical simulation techniques
  - Gibbs sampling
  - Metropolis-Hastings algorithm

More on Priors

- Two decisions with regard to priors:
  1. Family of the prior distribution
  2. Hyperparameters of the prior distribution
- *In principle* any distribution can be combined with the likelihood to form the posterior.
- *Conjugate priors*

  If a prior is conjugate, then the posterior has the same density as the prior. ⇨ Very convenient
- *Natural conjugate priors*

  Additional property: they have the same functional form as the likelihood function. ⇨ The prior can be interpreted as arising from earlier data analysis.

The Linear Regression Model

- Consider the linear regression model with K fixed regressors:

$$\mathbf{Y} = \mathbf{X}\boldsymbol{\beta} + \boldsymbol{\varepsilon} \qquad\qquad \boldsymbol{\varepsilon} \sim \boldsymbol{N}(\mathbf{0}, \sigma^2 \boldsymbol{I_T})$$

where $\mathbf{Y}$ and $\boldsymbol{\varepsilon}$ are $T \times 1$ vectors, $\mathbf{X}$ is a $T \times K$ matrix of exogenous variables and deterministic terms.

- Likelihood:

$$p(\mathbf{Y}|\boldsymbol{\beta}, \sigma^2) = \frac{1}{(2\pi\sigma^2)^{\frac{T}{2}}} \exp[-\frac{1}{2\sigma^2}(\mathbf{Y} - \mathbf{X}\boldsymbol{\beta})'(\mathbf{Y} - \mathbf{X}\boldsymbol{\beta})]$$

Bayesian Analysis

- Idea 1: The parameters $\boldsymbol{\theta} = [\boldsymbol{\beta}' \ \ \sigma^2]'$ are random variables with a probability distribution.
- Idea 2: A Bayesian estimate of this distribution combines prior beliefs and information from the data.
  - Step 1: Form prior beliefs about parameters (based on past experience or other studies) and express in the form of a probability distribution: $p(\boldsymbol{\theta})$
  - Step 2: Information contained in the data is summarized by the likelihood function: $L(\boldsymbol{\theta}|\mathbf{Y})$
  - Step 3: Bayes' Rule gives the posterior distribution of the parameters: $p(\boldsymbol{\theta}|\mathbf{Y}) \propto L(\boldsymbol{\theta}|\mathbf{Y})p(\boldsymbol{\theta})$

Example 1: Inference of β when σ^2 known

Prior distribution of β

$p(\boldsymbol{\beta}|\sigma^2) \sim N(\boldsymbol{\beta}_0, \boldsymbol{\Sigma}_0)$

Prior density: $(2\pi)^{-\frac{K}{2}}\,|\boldsymbol{\Sigma}_0|^{-\frac{1}{2}}\exp\left\{-\frac{1}{2}(\boldsymbol{\beta}-\boldsymbol{\beta}_0)'\boldsymbol{\Sigma}_0^{-1}(\boldsymbol{\beta}-\boldsymbol{\beta}_0)\right\}$

$$p(\boldsymbol{\beta}|\sigma^2) \propto \exp\left\{-\frac{1}{2}(\boldsymbol{\beta}-\boldsymbol{\beta}_0)'\boldsymbol{\Sigma}_0^{-1}(\boldsymbol{\beta}-\boldsymbol{\beta}_0)\right\}$$

Example: $\boldsymbol{\beta}_0 = \begin{bmatrix} 1 \\ -1 \end{bmatrix}$ and $\boldsymbol{\Sigma}_0 = \begin{bmatrix} 1 & 0 \\ 0 & 10 \end{bmatrix}$

Likelihood

$$L(\boldsymbol{\beta}|\sigma^2, \mathbf{Y}) \propto \exp\left\{-\frac{1}{2\sigma^2}(\mathbf{Y}-\mathbf{X}\boldsymbol{\beta})'(\mathbf{Y}-\mathbf{X}\boldsymbol{\beta})\right\}$$

Combining prior density and likelihood

$$p(\boldsymbol{\beta}|\sigma^2, \mathbf{Y}) \propto p(\boldsymbol{\beta}|\sigma^2)\, L(\boldsymbol{\beta}|\sigma^2, \mathbf{Y})$$

$$\propto \exp\left\{-\frac{1}{2}(\boldsymbol{\beta}-\boldsymbol{\beta}_0)'\boldsymbol{\Sigma}_0^{-1}(\boldsymbol{\beta}-\boldsymbol{\beta}_0) - \frac{1}{2\sigma^2}(\mathbf{Y}-\mathbf{X}\boldsymbol{\beta})'(\mathbf{Y}-\mathbf{X}\boldsymbol{\beta})\right\}$$

Posterior distribution of β

$$p(\boldsymbol{\beta}|\sigma^2, \mathbf{Y}) \sim N(\boldsymbol{\beta}_1, \boldsymbol{\Sigma}_1)$$

where

$$\boldsymbol{\beta}_1 = (\boldsymbol{\Sigma}_0^{-1} + \sigma^{-2}\mathbf{X}'\mathbf{X})^{-1}(\boldsymbol{\Sigma}_0^{-1}\boldsymbol{\beta}_0 + \sigma^{-2}\mathbf{X}'\mathbf{Y})$$

$$= (\boldsymbol{\Sigma}_0^{-1} + \sigma^{-2}\mathbf{X}'\mathbf{X})^{-1}(\boldsymbol{\Sigma}_0^{-1}\boldsymbol{\beta}_0 + \sigma^{-2}\mathbf{X}'\mathbf{X}\boldsymbol{b}) \text{ with } \mathbf{b}=(\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'\mathbf{Y}$$

$$\boldsymbol{\Sigma}_1 = (\boldsymbol{\Sigma}_0^{-1} + \sigma^{-2}\mathbf{X}'\mathbf{X})^{-1}$$

Example 2: Inference of σ^2 when β known

- Recall: $\varepsilon_i \sim N(0, \sigma^2) \quad \Rightarrow \quad W = \sum_{i=1}^{\nu} \varepsilon_i^2$

  then $W \sim \Gamma(\nu, \delta)$

with the density for the Gamma distribution given by:

$$p(W) = [\Gamma(\nu/2)]^{-1} \left[\frac{\delta}{2}\right]^{\nu/2} W^{(\frac{\nu}{2}-1)} \exp[\frac{\delta}{2} W]$$

where $E(W) = \frac{\nu}{\delta}$ and $Var(W) = \frac{\nu}{\delta^2}$

- Use this as a prior for the inverse of the variance σ^2 (also called the "precision"):

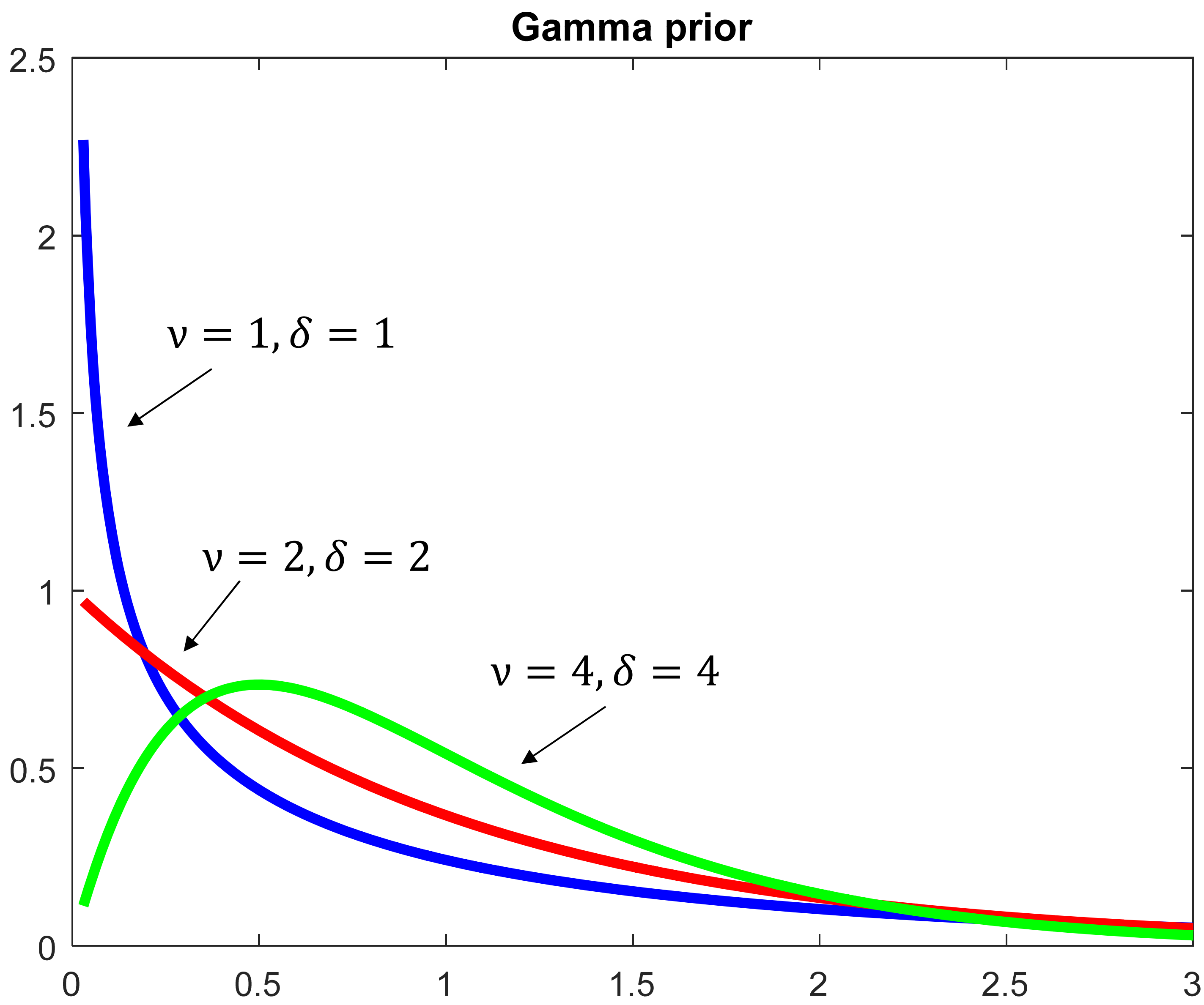
$$p(1/\sigma^2) \sim \Gamma(\nu_0, \delta_0)$$

Why Use this Prior?

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

2) flexible family (different shapes)

Gamma Distributions with Mean Unity



Why Use this Prior?

3) It is the "natural conjugate prior" given the likelihood, meaning that if the prior is $p(1/\sigma^2) \sim \Gamma(\nu_0, \delta_0)$, then the posterior turns out to be $p(1/\sigma^2|\mathbf{Y}) \sim \Gamma(\nu_1, \delta_1)$

- If prior were derived from earlier data analysis, it would have this form

  $\Rightarrow$ it is equivalent to having ν_0 observations with sum of squared residuals δ_0

- This prior makes analytical treatment of the problem tractable

Example 2: Inference of σ^2 when β known

Prior distribution of $1/\sigma^2$

$$p(1/\sigma^2|\boldsymbol{\beta}) \sim \Gamma(\nu_0, \delta_0)$$

Prior density: $p(1/\sigma^2|\boldsymbol{\beta}) \propto \left(\frac{1}{\sigma^2}\right)^{\frac{\nu_o}{2}-1} \exp\left(-\frac{\delta_0}{2\sigma^2}\right)$

Likelihood

$$L(1/\sigma^2|\boldsymbol{\beta}, \mathbf{Y}) \propto \frac{1}{(\sigma^2)^{\frac{T}{2}}} \exp\left\{-\frac{1}{2\sigma^2}(\mathbf{Y}-\mathbf{X}\boldsymbol{\beta})'(\mathbf{Y}-\mathbf{X}\boldsymbol{\beta})\right\}$$

Combining prior density and likelihood

$$p(1/\sigma^2|\boldsymbol{\beta}, \mathbf{Y}) \propto p(1/\sigma^2|\boldsymbol{\beta})\, L(\sigma^2|\boldsymbol{\beta}, \mathbf{Y})$$

$$\propto \left(\frac{1}{\sigma^2}\right)^{\frac{\nu_o}{2}-1} \exp\{-\frac{\delta_0}{2\sigma^2}\} \frac{1}{(\sigma^2)^{\frac{T}{2}}} \exp\left\{-\frac{1}{2\sigma^2}(\mathbf{Y}-\mathbf{X}\boldsymbol{\beta})'(\mathbf{Y}-\mathbf{X}\boldsymbol{\beta})\right\}$$

$$= (\frac{1}{\sigma^2})^{\frac{\nu_o}{2}+\frac{T}{2}-1} \exp\left\{-\frac{1}{2\sigma^2}[\delta_0 + (\mathbf{Y}-\mathbf{X}\boldsymbol{\beta})'(\mathbf{Y}-\mathbf{X}\boldsymbol{\beta})]\right\}$$

Posterior distribution of $1/\sigma^2$

$$p(1/\sigma^2|\boldsymbol{\beta}, \mathbf{Y}) \sim \Gamma(\nu_1, \delta_1)$$

where

$$\nu_1 = \nu_0 + T$$

$$\delta_1 = \delta_0 + (\mathbf{Y} - \mathbf{X}\boldsymbol{\beta})'\,(\mathbf{Y} - \mathbf{X}\boldsymbol{\beta})$$

What If All Parameters Are Unknown?

- Setting the prior: *joint* density for $\boldsymbol{\beta}$ and $1/\sigma^2$

$$p(\boldsymbol{\beta}, 1/\sigma^2) = p(\boldsymbol{\beta}|\sigma^2)\, p(1/\sigma^2)$$

$$\text{where} \quad p(\boldsymbol{\beta}|\sigma^2) \sim N(\boldsymbol{\beta}_0, \sigma^2\boldsymbol{\Sigma}_0)$$

$$p(1/\sigma^2) \sim \Gamma(\nu_o, \delta_0)$$

- Setting up the likelihood function

$$p(\mathbf{Y}|\boldsymbol{\beta}, \sigma^2) = \frac{1}{(2\pi\sigma^2)^{\frac{T}{2}}} \exp[-\frac{1}{2\sigma^2}(\mathbf{Y} - \mathbf{X}\boldsymbol{\beta})'(\mathbf{Y} - \mathbf{X}\boldsymbol{\beta})]$$

What If All Parameters Are Unknown?

- Calculating the *joint* posterior distribution

$$p(\boldsymbol{\beta}, 1/\sigma^2|\mathbf{Y}) \propto p(\mathbf{Y}, \boldsymbol{\beta}, 1/\sigma^2)$$

$$\propto \frac{1}{(\sigma^2)^{\frac{T}{2}}} \exp\left\{-\frac{1}{2\sigma^2}(\mathbf{Y}-\mathbf{X}\boldsymbol{\beta})'(\mathbf{Y}-\mathbf{X}\boldsymbol{\beta})\right\}$$

$$\left(\frac{1}{\sigma^2}\right)^{\frac{\nu_o}{2}-1} \exp\{-\frac{\delta_0}{2\sigma^2}\}$$

$$\left(\frac{1}{\sigma^2}\right)^{K/2} \exp\left\{-\frac{1}{2\sigma^2}(\boldsymbol{\beta}-\boldsymbol{\beta}_0)'\boldsymbol{\Sigma}_0^{-1}(\boldsymbol{\beta}-\boldsymbol{\beta}_0)\right\}$$

Posterior for $\boldsymbol{\sigma}^{-2}|\mathbf{Y}$

$$1/\sigma^2|\mathbf{Y} \sim \Gamma(\nu^*, \delta^*)$$

$$\nu^* = \nu_0 + T$$

$$\delta^* = \delta_0 + (\mathbf{Y} - \mathbf{Xb})'(\mathbf{Y} - \mathbf{Xb}) + (\mathbf{b} - \boldsymbol{\beta}_\mathbf{0})'\widetilde{\boldsymbol{\Sigma}}(\mathbf{b} - \boldsymbol{\beta}_\mathbf{0})$$

$$\mathbf{b} = (\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'\mathbf{Y}$$

$$\widetilde{\boldsymbol{\Sigma}} = \boldsymbol{\Sigma}_\mathbf{0}^{-1}(\boldsymbol{\Sigma}_\mathbf{0}^{-1} + \mathbf{X}'\mathbf{X})^{-\mathbf{1}}\,\mathbf{X}'\mathbf{X}$$

Posterior for $\boldsymbol{\beta}|\sigma^{-2}, \mathbf{Y}$

$$\boldsymbol{\beta}|\sigma^{-2}, \mathbf{Y} \sim \mathrm{N}(\boldsymbol{\beta}^{*}, \sigma^{2}\boldsymbol{\Sigma}^{*})$$

$$\boldsymbol{\beta}^{*} = \boldsymbol{\Sigma}^{*}(\boldsymbol{\Sigma}_{\mathbf{0}}^{-1}\boldsymbol{\beta}_{\mathbf{0}} + \mathbf{X}'\mathbf{Y})$$

$$\boldsymbol{\Sigma}^{*} = (\boldsymbol{\Sigma}_{\mathbf{0}}^{-1} + \mathbf{X}'\mathbf{X})^{-1}$$

- Diffuse prior: $\boldsymbol{\Sigma}_{\mathbf{0}} \rightarrow \infty \cdot \mathbf{I}_{K}$

$$\Rightarrow \boldsymbol{\Sigma}^{*} \rightarrow (\mathbf{X}'\mathbf{X})^{-1}$$

$$\Rightarrow \boldsymbol{\beta}^{*} \rightarrow (\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'\mathbf{Y}$$

= usual OLS formulas

Posterior for $\beta|\sigma^{-2}, \mathbf{Y}$

$$\boldsymbol{\beta}|\sigma^{-2}, \mathbf{Y} \sim \mathrm{N}(\boldsymbol{\beta}^*, \sigma^2\boldsymbol{\Sigma}^*)$$

$$\boldsymbol{\beta}^* = \boldsymbol{\Sigma}^*(\boldsymbol{\Sigma}_0^{-1}\boldsymbol{\beta}_0 + \mathbf{X}'\mathbf{Y})$$

$$\boldsymbol{\Sigma}^* = (\boldsymbol{\Sigma}_0^{-1} + \mathbf{X}'\mathbf{X})^{-1}$$

- Dogmatic prior: $\boldsymbol{\Sigma}_0 \rightarrow 0 \cdot \mathbf{I}_K$

$$\Rightarrow \boldsymbol{\Sigma}^* \rightarrow \mathbf{0}$$

$$\Rightarrow \boldsymbol{\beta}^* \rightarrow \boldsymbol{\beta}_0$$

posterior = prior

Posterior for $\beta|\sigma^{-2}, \mathbf{Y}$

$$\boldsymbol{\beta}|\sigma^{-2}, \mathbf{Y} \sim \mathrm{N}(\boldsymbol{\beta}^*, \sigma^2\boldsymbol{\Sigma}^*)$$

$$\boldsymbol{\beta}^* = \boldsymbol{\Sigma}^*(\boldsymbol{\Sigma}_0^{-1}\boldsymbol{\beta}_0 + \mathbf{X}'\mathbf{Y})$$

$$\boldsymbol{\Sigma}^* = (\boldsymbol{\Sigma}_0^{-1} + \mathbf{X}'\mathbf{X})^{-1}$$

- In general: $\boldsymbol{\beta}^*$ is a matrix-weighted average of $\boldsymbol{\beta}_0$ and $\widehat{\boldsymbol{\beta}}$, where weights depend on confidence in prior ($\boldsymbol{\Sigma}_0$) and strength of evidence from data ($\mathbf{X}'\mathbf{X}$)

Another Way to Interpret the Prior

- Suppose I had observed an earlier sample of $\tilde{T}$ observations:

$$\{\tilde{Y}_t, \widetilde{\boldsymbol{X}}_t\}_{\tilde{t}=1}^{\tilde{T}}$$

which were independent of the current observed sample:

$$\{Y_t, \boldsymbol{X}_t\}_{t=1}^{T}$$

Another Way to Interpret the Prior

- Then my OLS estimate based on all information would be:

$$\widehat{\boldsymbol{\beta}} = \left(\sum_{t=1}^{T} \mathbf{X}_t \mathbf{X}_t' + \sum_{\tilde{t}=1}^{\tilde{T}} \widetilde{\mathbf{X}}_t \widetilde{\mathbf{X}}_t'\right)^{-1} \left(\sum_{t=1}^{T} \mathbf{X}_t \mathbf{Y}_t + \sum_{\tilde{t}=1}^{\tilde{T}} \widetilde{\mathbf{X}}_t \widetilde{\mathbf{Y}}_t\right)$$

with variance (given σ^2) of:

$$Var(\widehat{\boldsymbol{\beta}}) = \sigma^2 \left(\sum_{t=1}^{T} \mathbf{X}_t \mathbf{X}_t' + \sum_{\tilde{t}=1}^{\tilde{T}} \widetilde{\mathbf{X}}_t \widetilde{\mathbf{X}}_t'\right)^{-1}$$

Another Way to Interpret the Prior

- Let $\boldsymbol{\beta}_0$ be the OLS estimate based on the prior sample *alone*:

$$\boldsymbol{\beta}_0 = \left(\sum_{\tilde{t}=1}^{\tilde{T}} \widetilde{\mathbf{X}}_t \widetilde{\mathbf{X}}_t'\right)^{-1} \left(\sum_{\tilde{t}=1}^{\tilde{T}} \widetilde{\mathbf{X}}_t \widetilde{\mathbf{Y}}_t\right)$$

and let $\sigma^2 \boldsymbol{\Sigma}_0$ denote its variance:

$$\boldsymbol{\Sigma}_0 = \left(\sum_{\tilde{t}=1}^{\tilde{T}} \widetilde{\mathbf{X}}_t \widetilde{\mathbf{X}}_t'\right)^{-1}$$

Another Way to Interpret the Prior

$$\widehat{\boldsymbol{\beta}} = \left(\sum_{t=1}^{T} \mathbf{X}_t \mathbf{X}_t' + \sum_{\tilde{t}=1}^{\tilde{T}} \widetilde{\mathbf{X}}_t \widetilde{\mathbf{X}}_t'\right)^{-1}$$

$$\left(\sum_{t=1}^{T} \mathbf{X}_t \mathbf{Y}_t + \sum_{\tilde{t}=1}^{\tilde{T}} \widetilde{\mathbf{X}}_t \widetilde{\mathbf{Y}}_t\right)$$

$$= \left(\sum_{t=1}^{T} \mathbf{X}_t \mathbf{X}_t' + \boldsymbol{\Sigma}_0^{-1}\right)^{-1}$$

$$\left(\sum_{t=1}^{T} \mathbf{X}_t \mathbf{Y}_t + \boldsymbol{\Sigma}_0^{-1} \boldsymbol{\beta}_0\right)$$

$\Longrightarrow$ identical to formula for posterior mean $\boldsymbol{\beta}^*$

Another Way to Interpret the Prior

$$Var(\widehat{\boldsymbol{\beta}}) = \sigma^2 \left(\sum_{t=1}^{T} \mathbf{X}_t \mathbf{X}_t' + \sum_{\tilde{t}=1}^{\tilde{T}} \widetilde{\mathbf{X}}_t \widetilde{\mathbf{X}}_t'\right)^{-1}$$

$$= \sigma^2 \left(\sum_{t=1}^{T} \mathbf{X}_t \mathbf{X}_t' + \boldsymbol{\Sigma}_0^{-1}\right)^{-1}$$

$$= \sigma^2 \boldsymbol{\Sigma}^*$$

$\Rightarrow$ for $\boldsymbol{\Sigma}^*$ the posterior variance defined earlier

Dummy Observations

- Augment original dataset with artificial observations that correspond to the prior

$$\underset{(T+k)\times 1}{\mathbf{y}^*} = \begin{bmatrix} y_1 \\ \vdots \\ y_T \\ \mathbf{P}^{-1}\boldsymbol{\beta}_0 \end{bmatrix} \qquad \underset{(T+k)\times k}{\mathbf{X}^*} = \begin{bmatrix} \mathbf{x}_0' \\ \vdots \\ \mathbf{x}_{T-1}' \\ \mathbf{P}^{-1} \end{bmatrix}$$

where $\mathbf{P}^{-1}$ is the Cholesky factor of $\boldsymbol{\Sigma}_{\mathbf{0}}^{-1} (= \mathbf{P}^{-1}\mathbf{P}^{-1\prime})$

$$\Longrightarrow \quad \boldsymbol{\beta}^* = \left(\sum_{t=1}^{T+k} \mathbf{x}_{t-1}^* \mathbf{x}_{t-1}^{*\prime}\right)^{-1} \left(\sum_{t=1}^{T+k} \mathbf{x}_{t-1}^* y_t^*\right)$$

Sources of Prior Information

- Observations of another dataset
  - Earlier time period
  - Different country
  - ➢ Question: How representative of sample/country we want to analyze?

Sources of Prior Information

- Solution: downweight these observations by κ

Use $\nu = \kappa \tilde{T}$, $\delta = \kappa \sum_{\tilde{t}=1}^{\tilde{T}} (y_{\tilde{t}} - \hat{\tilde{\boldsymbol{\beta}}}_i' \mathbf{x}_{\tilde{t}-1})^2$

$$\boldsymbol{\beta}_0 = \left(\sum_{\tilde{t}=1}^{\tilde{T}} \mathbf{x}_{\tilde{t}-1} \mathbf{x}_{\tilde{t}-1}' \right)^{-1} \left(\sum_{\tilde{t}=1}^{\tilde{T}} \mathbf{x}_{\tilde{t}-1} y_{\tilde{t}} \right)$$

$$\boldsymbol{\Sigma}_0 = \kappa^{-1} \left(\sum_{\tilde{t}=1}^{\tilde{T}} \mathbf{x}_{\tilde{t}-1} \mathbf{x}_{\tilde{t}-1}' \right)^{-1}$$

$\kappa = 1 \Rightarrow$ earlier data just as good as current

$\kappa = 0.5 \Rightarrow$ earlier gets half the weight of current

$\kappa = 0 \Rightarrow$ earlier data completely ignored

$\Longrightarrow$ κ summarizes how much you trust the other dataset (how many observations the prior is counted as)

Sources of Prior Information

- Typical time-series properties
  - Most variables are hard to forecast

    $\rightarrow$ most elements of $\boldsymbol{\beta}_0$ are zero
  - To the extent that variables do help, most recent values are likely to be more useful

$\Rightarrow$ Minnesota prior

(we will study in relation to VARs)

What About the Marginal Posterior for β?

- To make inference on $\boldsymbol{\beta}$, we need to know the *marginal* posterior:

$$p(\boldsymbol{\beta}|\mathbf{Y}) = \int_0^\infty p(\boldsymbol{\beta}, \frac{1}{\sigma^2}|\mathbf{Y}) d\frac{1}{\sigma^2}$$

- For this simple model under the natural conjugate prior analytical results can be obtained:

  $\boldsymbol{\beta}|\mathbf{Y} \sim$ multivariate Student t with $\nu_0 + T$ degrees of freedom, mean $\boldsymbol{\beta}^*$, and scale matrix $(\delta^*/\nu^*)\,\boldsymbol{\Sigma}^*$ as defined before

- BUT
  - » integration is hard
  - » with other prior distributions analytical derivation of joint and marginal posterior is not possible

Solution: Gibbs Sampling

- Suppose the parameter vector $\boldsymbol{\theta}$ can be partitioned as $\boldsymbol{\theta}' = (\boldsymbol{\theta}_1', \boldsymbol{\theta}_2', \boldsymbol{\theta}_3')$ with the property that $p(\boldsymbol{\theta}|\mathbf{Y})$ is of unknown form but

$$p(\boldsymbol{\theta}_1|\mathbf{Y}, \boldsymbol{\theta}_2, \boldsymbol{\theta}_3)$$
$$p(\boldsymbol{\theta}_2|\mathbf{Y}, \boldsymbol{\theta}_1, \boldsymbol{\theta}_3)$$
$$p(\boldsymbol{\theta}_3|\mathbf{Y}, \boldsymbol{\theta}_1, \boldsymbol{\theta}_2)$$

are of known form and we can easily sample from these *conditional* distributions (same idea works for 2, 4, or n blocks)

Gibbs Sampling: Theory

- What does that buy us?
  - Theory suggests that if we obtain many samples $\boldsymbol{\theta}_1^{(j)}, j \rightarrow \infty$ from $p(\boldsymbol{\theta}_1|\mathbf{Y}, \boldsymbol{\theta}_2, \boldsymbol{\theta}_3)$, then these will also be samples from the joint posterior $p(\boldsymbol{\theta}_1, \boldsymbol{\theta}_2, \boldsymbol{\theta}_3|\mathbf{Y})$ (see Geman and Geman, 1984; Casella and George, 1992)
  - The marginal posterior distribution $p(\boldsymbol{\theta}_1|\mathbf{Y})$ can be approximated by the *empirical* distribution of $\boldsymbol{\theta}_1$
    $\Rightarrow$ for example: estimate of mean for $\boldsymbol{\theta}_{1,i}$ is the sample mean of retained draws $\frac{1}{(D-D_0)}\sum_{j=D_0+1}^{D} \theta_{1,i}$

Gibbs Sampling: Implementation

(1) Start with arbitrary initial guesses

$\boldsymbol{\theta}_1^{(j)}, \boldsymbol{\theta}_2^{(j)}, \boldsymbol{\theta}_3^{(j)}$ for $j = 1$.

(2) Generate: $\boldsymbol{\theta}_1^{(j+1)}$ from $p(\boldsymbol{\theta}_1 | \mathbf{Y}, \boldsymbol{\theta}_2^{(j)}, \boldsymbol{\theta}_3^{(j)})$

$\boldsymbol{\theta}_2^{(j+1)}$ from $p(\boldsymbol{\theta}_2 | \mathbf{Y}, \boldsymbol{\theta}_1^{(j+1)}, \boldsymbol{\theta}_3^{(j)})$

$\boldsymbol{\theta}_3^{(j+1)}$ from $p(\boldsymbol{\theta}_3 | \mathbf{Y}, \boldsymbol{\theta}_1^{(j+1)}, \boldsymbol{\theta}_2^{(j+1)})$

(3) Repeat step (2) for $j = 1, 2, \ldots, D$

(4) Throw out first D_0 draws (for D_0 large) and use remaining $(D - D_0)$ draws for inference

Back to our Regression Model

- Idea: By sampling repeatedly from the conditional distributions $p(\boldsymbol{\beta}|\frac{1}{\sigma^2},\mathbf{Y})$ and $p(\frac{1}{\sigma^2}|\boldsymbol{\beta},\mathbf{Y})$, we can approximate the joint and marginal distributions of our parameters of interest

- Steps:
  1. Set priors and initial guess for σ^2
  2. Sample $\boldsymbol{\beta}$ conditional on $\frac{1}{\sigma^2}$
  3. Sample $\frac{1}{\sigma^2}$ conditional on $\boldsymbol{\beta}$
  4. Cycle through steps (2) and (3) a large number of times and keep only the last $(D - D_0)$ draws

Application 1

- Linear regression model with one exogenous variable:

$$y_t = x_t\beta + \varepsilon_t, \; t = 1, \dots, T \text{ and } \varepsilon_t \sim N(0, \sigma^2)$$

- <u>Gibbs sampling algorithm</u>:

(1) a. Set priors: $\beta \sim N(b_0, P_0)$ and $\frac{1}{\sigma^2} \sim \Gamma(t_0, R_0)$

Prior hyperparameters:

$b_0 = 0.5, P_0 = 10, t_0 = 0, R_0 = 0$

b. Set starting value for first iteration

$\sigma^{2,(0)} = 1$

Application 1

(2) At iteration j, conditional on draw $\sigma^{2,(j-1)}$, draw

$$\beta^{j} | \sigma^{2,(j-1)}, \boldsymbol{y} \sim N(b_1^{j-1}, P_1^{j-1})$$

where

$$P_1^{j-1} = (P_0^{-1} + \sigma^{2,(j-1)} \boldsymbol{x}'\boldsymbol{x})^{-1}$$

$$b_1^{j-1} = P_1^{j-1} (P_0^{-1} b_0 + \sigma^{2,(j-1)} \boldsymbol{x}'\boldsymbol{y})$$

(3) Conditional on draw β^{j}, draw

$$\frac{1}{\sigma^{2,(j)}} | \beta^{j}, \boldsymbol{y} \sim \Gamma(t_1, R_1^{j})$$

where

$$t_1 = t_0 + T$$

$$R_1^{j} = R_0 + (\boldsymbol{y} - \boldsymbol{x}\beta^{j})'(\boldsymbol{y} - \boldsymbol{x}\beta^{j})'$$

How to Take Draws

- Normal distribution

  To sample a $k \times 1$ vector $\boldsymbol{z}$ from $N(\mathbf{m}, \mathbf{V})$, generate $k \times 1$ draws $\boldsymbol{z}^{\mathbf{0}}$ from the standard normal distribution (randn in Matlab) and then apply the following transformation

$$\boldsymbol{z} = \mathbf{m} + \left[\left(\boldsymbol{z}^{\mathbf{0}}\right)' \cdot \mathbf{V}^{1/2}\right]' = \mathbf{m} + [randn(1, k) \cdot chol(\mathbf{V})]'$$

  - $\mathbf{A}$ is said to be a square root of $\mathbf{V}$ if the matrix product
  $$\mathbf{AA} = \mathbf{V}$$
  - For positive-definite matrices, one way to obtain the square root is the *Choleski decomposition* (chol in Matlab): $\mathbf{C} = chol(\mathbf{V})' \Rightarrow \ \mathbf{CC}' = \mathbf{V}$

How to Take Draws

- Inverse gamma distribution

$$\frac{1}{\sigma^2} \sim \Gamma(v, \frac{1}{\delta})$$

$$\sigma^2 \sim \Gamma^{-1}(v, \delta)$$

To sample a scalar s from an inverse gamma with degrees of freedom v and scale parameter δ, there are 2 options:

- generate T numbers from $\boldsymbol{s}^0 \sim N(0, 1)$ and apply the following transformation

$$s = \frac{\delta}{(\boldsymbol{s}^0)'\boldsymbol{s}^0}$$

- generate a draw $\bar{s}$ from a gamma with degrees of freedom v and scale parameter $\frac{1}{\delta}$ (gamrnd in Matlab) and compute

$$s = \frac{1}{\bar{s}}$$

Posterior Distribution