Introduction to Bayesian Econometrics

Dario Caldara

Federal Reserve Board

ECON 597 - Georgetown University

The Scientific Method

Now this is the peculiarity of scientific method, that when once it has become a habit of mind, that mind converts all facts whatsoever into science. The field of science is unlimited; its material endless, every group of natural phenomena, every phase of social life, every stage of past or present development is material for science. The unity of all science consists alone in its method, not in its material.

Karl Pearson, *The Grammar of Science*, 1938

Scientific method relies mainly on two kinds of inferences:

- Deductive inference: Conclusion always follows the stated premises.
- Inductive inference: Reaching a general conclusion from specific examples.

Inductive Inference

- The fundamental problem of scientific progress, and fundamental one of everyday life, is that of learning from experience.
- Inductive inference: Making inference from past experience to predict future experience.

... than inference from past observations to future ones is not deductive. The observations not yet made might concern events either in the future or simply at places not yet inspected. It is technically called induction.

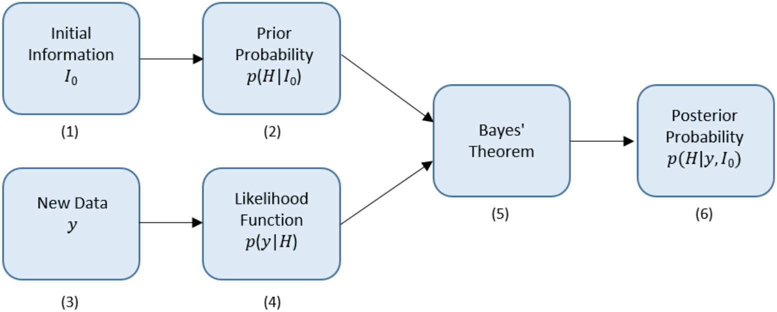
Harold Jeffreys, *Scientific Inference*, 1957

Bayesian Approach and Inductive Inference

- Bayesian approach to inference is making inductive inference.
- Bayesian inference describes the process of revising probabilities representing degrees of belief in propositions to incorporate new information.

1. Initial belief associated with a proposition H based on some initial information I_0: $\mathbf{p(H|I_0)}$.
   - I_0 comes from previous data, studies, theoretical considerations...
2. New data y bring information about the proposition: $\mathbf{p(y|H)}$.
3. Revision of initial belief to reflect the information in new data: $\mathbf{p(H|y, I_0)}$.

Schematic Representation of Bayesian Inference



A. Zellner, *An Introduction to Bayesian Inference in Econometrics*, 1971

Bayesian Inference Applied to a Parameter θ

- Replace proposition H with a parameter θ.
- $p(\theta|I_0)$: prior probability density function (pdf).
- $p(y|\theta)$: Likelihood function.
- $p(\theta|y, I_0)$: Posterior pdf.
- Employ posterior pdf to make probability statements:
    - The probability that $a < \theta < b$, where a and b are numbers.
- This procedure is operational and applicable in the analysis of a wide range of problems because it is central to the inductive process.

Bayesian Inference

The basis for Bayesian inference is the **The Bayes Rule**

Let A and B be two events. We have:

$$P(A \cap B) = P(A|B)P(B) = P(B|A)P(A)$$

$\Rightarrow$ Bayes Rule:

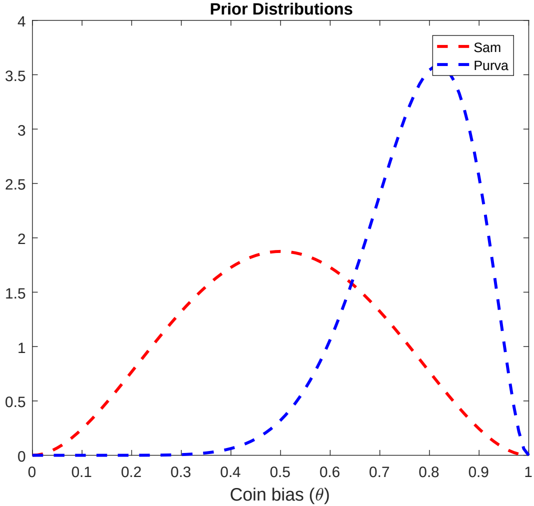
$$P(A|B) = \frac{P(B|A)P(A)}{P(B)} \propto P(B|A)P(A)$$

$P(A|B)$ and $P(B|A)$ are known as conditional probabilities.

$P(A)$ and $P(B)$ are known as marginal probabilities.

The Coin Bias Example: Prior Distributions

θ: coin bias.



The Coin Bias Example: Likelihood Function

Binomial distribution:

$$P(x) = \underbrace{\frac{N!}{x!(N-x)!}}_{constant} \theta^x (1-\theta)^{N-x}$$

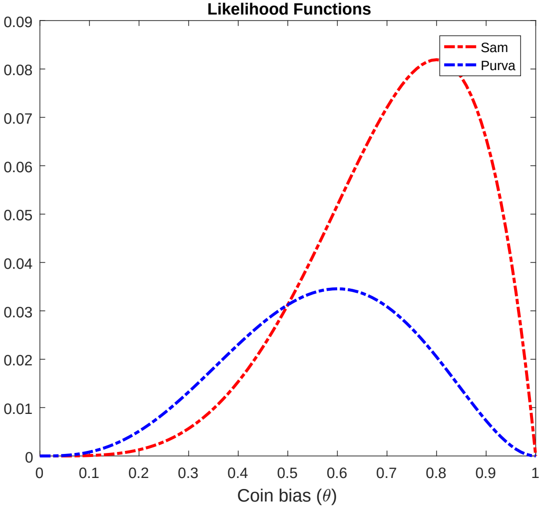
Likelihood function:

$$L(y|\theta) = \Pi_{i=1}^{N} \theta^x (1-\theta)^{N-x}$$

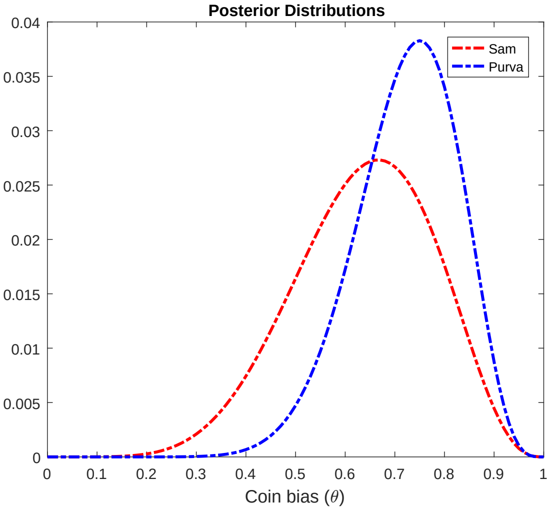
$N =$ number of tosses; $x = \{1 \equiv H; 0 \equiv T\}$;

The Coin Bias Example: Likelihood Functions

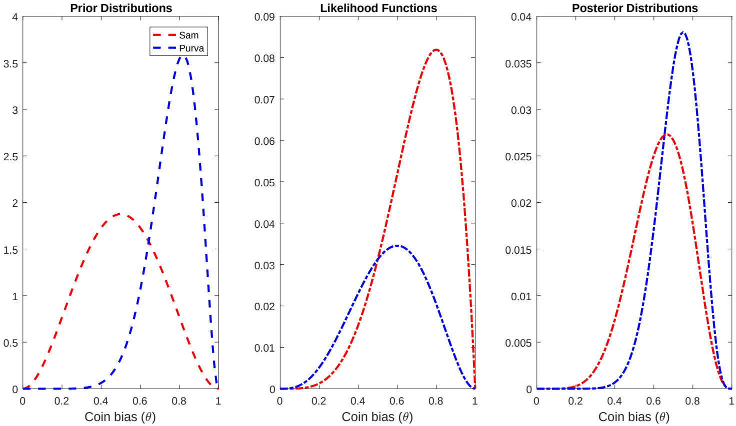
Sam tosses: [T H H H H]
Student 2 tosses: [H H T H T]



The Coin Bias Example: Posterior Distributions



The Coin Bias Example: Summing Up



Bayesian inference

i) Formulate a parametric model as a collection of probability distributions of all possible realization of the data (Y) conditional on different values of the model parameters $\theta \in \Theta$
Model: $p(Y|\theta)$

ii) Organize the belief about θ into a (prior) probability distribution over Θ.
Prior: $p(\theta)$

iii) Collect the data y and treat them as realizations of Y and insert them into the family of distributions.
Likelihood: $\mathcal{L}(y|\theta) = p(y|\theta)$

iv) Use the Bayes theorem to calculate the new belief about θ.
Posterior: $p(\theta|y) \propto \mathcal{L}(y|\theta)p(\theta)$

Inference about location

- T independent observations $\mathbf{y} = (y_1, y_2, ..., y_T)'$ drawn from a normal distribution with unknown mean μ and known variance σ^2

$$y_t = \mu + \varepsilon_t \qquad \varepsilon_t \sim \mathcal{N}(0, \sigma^2)$$

- Posterior for μ

$$\underbrace{p(\mu|\mathbf{y}, \sigma^2)}_{posterior} \propto \underbrace{p(\mu)}_{prior} \underbrace{p(\mathbf{y}|\mu, \sigma^2)}_{likelihood\ function}$$

- Before doing the algebra, recall that the probability density of the normal distribution is:

$$p(y_i|\mu, \sigma) = \frac{1}{\sqrt{2\pi\sigma^2}} exp\left[\frac{(y_i - \mu)^2}{2\sigma^2}\right]$$

Inference about location

- The likelihood function is given by:

$$p(\mathbf{y}|\mu,\sigma) = \Pi_{i=1}^{T} p(y_i|\mu,\sigma)$$

- In the example, the likelihood is

$$\begin{aligned} p(\mathbf{y}|\mu,\sigma) &= (2\pi\sigma^2)^{-T/2} \exp\left[-\frac{1}{2\sigma^2}\sum_{i=1}^{T}(y_i-\mu)^2\right] \\ &= (2\pi\sigma^2)^{-T/2} \exp\left[-\frac{1}{2\sigma^2}\left[(T-1)s^2 + T(\mu-\bar{\mu})^2\right]\right] \end{aligned}$$

where $\bar{\mu} = \frac{1}{T}\sum_{i=1}^{T} y_i$ is the sample mean and $s^2 = \frac{1}{T-1}\sum_{i=1}^{T}(y_i-\bar{\mu})^2$ is the sample variance. For this derivation we used the following result

$$\begin{aligned} \sum_{i=1}^{T}(y_i-\mu)^2 &= \sum_{i=1}^{T}\left[(y_i-\hat{\mu})-(\mu-\hat{\mu})\right]^2 \\ &= \sum_{i=1}^{T}(y_i-\hat{\mu})^2 + (\mu-\hat{\mu})^2 \end{aligned}$$

Inference about location

- **Which prior?** Let's assume we have no information on μ

$$p(\mu) \propto 1; -\infty < \mu < \infty$$

all values are equiprobable (improper prior)

- Likhelihood $\Longrightarrow$ Posterior

$$\underbrace{p(\mu|\mathbf{y},\sigma^2)}_{\textit{posterior}} \propto \underbrace{p(\mathbf{y}|\mu,\sigma^2)}_{\textit{likelihood function}}$$

$$p(\mu|\mathbf{y},\sigma^2) \propto (\sigma^2)^{-\frac{T+1}{2}} \exp\left[-\frac{1}{2}\frac{(T-1)s^2}{\sigma^2}\right] \underbrace{(\sigma^2)^{-1/2} \exp\left[-\frac{1}{2(\sigma^2/T)}(\mu-\bar{\mu})^2\right]}_{\mathcal{N}(\bar{\mu},\sigma^2/T)}$$

$$\Rightarrow \boxed{\mu|\mathbf{y},\sigma^2 \sim \mathcal{N}\left(\bar{\mu},\sigma^2/T\right)}$$

Inference about location

- **Which prior?** We can use as prior the posterior obtained from a dummy sample
- T_d independent (dummy) observations $\mathbf{y}_d$ (observed before sample), which we believe are drawn from the same distribution

$$p(\mu) = p(\mu|\mathbf{y}_d, \sigma^2) \sim \mathcal{N}\left(\bar{\mu}_d, \sigma^2 / T_d\right)$$

- The new posterior is:

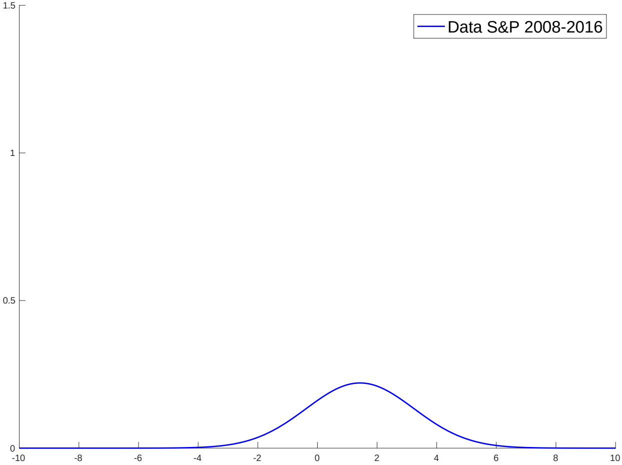
$$\begin{aligned} p(\mu|\mathbf{y}, \sigma^2) &\propto p(\mu)p(\mathbf{y}|\mu, \sigma^2) = p(\mathbf{y}_d|\mu, \sigma^2)p(\mathbf{y}|\mu, \sigma^2) = p(\mathbf{y}, \mathbf{y}_d|\mu, \sigma^2) \\ &\propto (2\pi\sigma^2)^{-(T_d+T)/2} \exp\left[-\frac{1}{2\sigma^2}\left(T_d(\mu - \bar{\mu}_d)^2 + T(\mu - \bar{\mu})^2\right)\right] \end{aligned}$$

$$\Rightarrow \boxed{\mu|\mathbf{y}, \sigma^2 \sim \mathcal{N}\left(\frac{1}{T + T_d}(T_d\bar{\mu}_d + T\bar{\mu}), \frac{1}{(T + T_d)}\sigma^2\right)}$$

S&P 500 returns – $p(\mu|\mathbf{y},\sigma)$

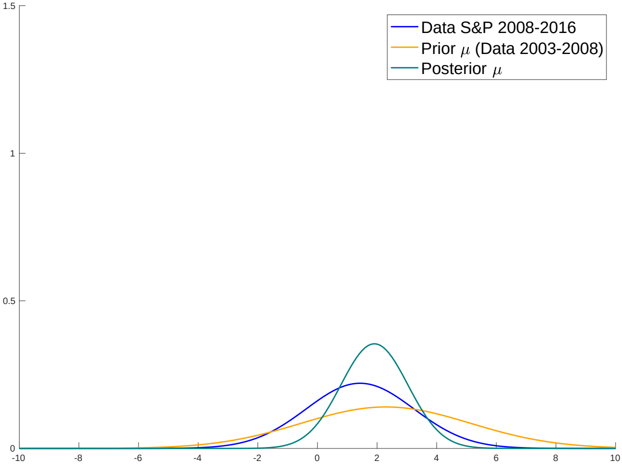
Likelihood function

Remember that μ is plotted on the x axis!



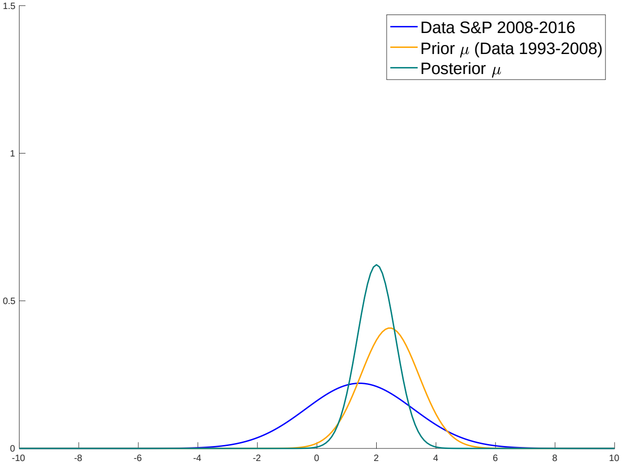
S&P 500 returns– $p(\mu|\mathbf{y}, \sigma)$

Remember that μ is plotted on the x axis!



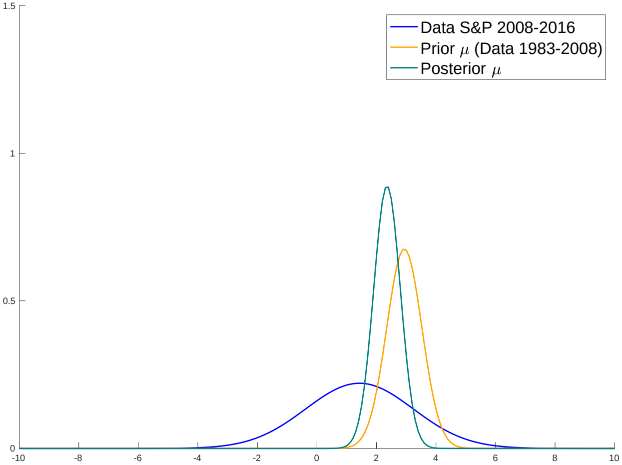
S&P 500 returns– $p(\mu|\mathbf{y}, \sigma)$

Remember that μ is plotted on the x axis!



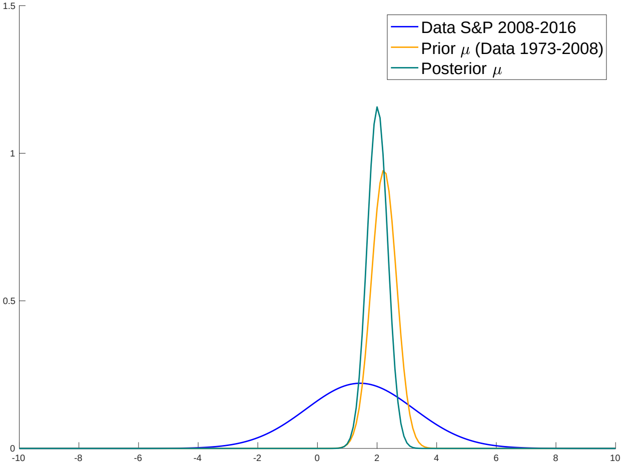
S&P 500 returns– $p(\mu|\mathbf{y}, \sigma)$

Remember that μ is plotted on the x axis!



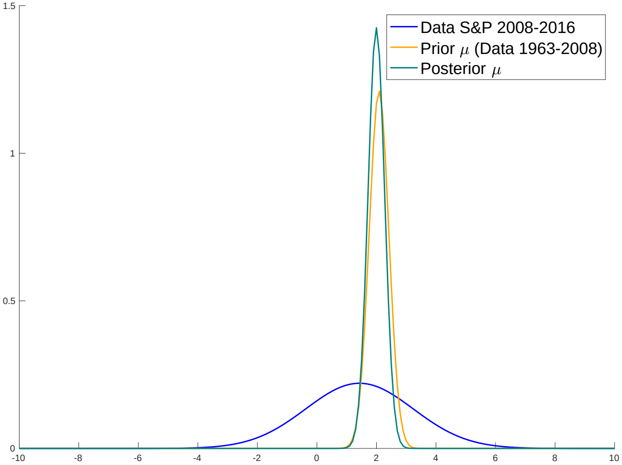
S&P 500 returns– $p(\mu|\mathbf{y}, \sigma)$

Remember that μ is plotted on the x axis!



S&P 500 returns– $p(\mu|\mathbf{y}, \sigma)$

Remember that μ is plotted on the x axis!



Inference about scale

- T independent observations $\mathbf{y} = (y_1, y_2, ..., y_T)'$ drawn from a normal distribution with known mean μ and unknown variance σ^2

$$y_t = \mu + \varepsilon_t \qquad \varepsilon_t \sim \mathcal{N}(0, \sigma^2)$$

- Posterior for σ^2

$$\underbrace{p(\sigma^2|\mathbf{y}, \mu)}_{\textit{posterior}} \propto \underbrace{p(\sigma^2)}_{\textit{prior}} \underbrace{p(\mathbf{y}|\sigma^2, \mu)}_{\textit{likelihood function}}$$

- Likelihood

$$\begin{aligned} p(\mathbf{y}|\mu, \sigma^2) &= (2\pi\sigma^2)^{-T/2} \exp\left[-\frac{1}{2\sigma^2}\sum_{i=1}^{T}(y_i - \mu)^2\right] \\ &= (2\pi\sigma^2)^{-T/2} \exp\left[-\frac{1}{2}\frac{Ts_\mu^2}{\sigma^2}\right] \end{aligned}$$

where $s_\mu^2 = \frac{1}{T}\sum_{i=1}^{T}(y_i - \mu)^2$

Inference about scale

- **Which prior?** Let's assume we have no information on $\log \sigma^2$ (improper prior)

$$p(\log \sigma^2) \propto 1 \Rightarrow p(\sigma^2) \propto \frac{1}{\sigma^2}; 0 < \sigma < \infty$$

- Likelihood $\Longrightarrow$ Posterior

$$\underbrace{p(\sigma^2|\mathbf{y},\mu)}_{posterior} \propto \underbrace{p(\mathbf{y}|\sigma^2,\mu)}_{likelihood\ function} \underbrace{\frac{1}{\sigma^2}}_{prior}$$

$$p(\sigma^2|\mathbf{y},\mu) \propto (2\pi\sigma^2)^{-\frac{T}{2}} \exp\left[-\frac{1}{2}\frac{Ts_\mu^2}{\sigma^2}\right](\sigma^2)^{-1} \propto (\sigma^2)^{-\frac{T+2}{2}} \exp\left[-\frac{1}{2}\frac{Ts_\mu^2}{\sigma^2}\right]$$

$$\Rightarrow \boxed{\sigma^2|\mathbf{y},\mu \sim \frac{Ts_\mu^2}{\chi^2_{(T)}} = \mathcal{IW}(Ts_\mu^2, T)}$$

In this univariate case:

$$\boxed{\sigma^2|\mathbf{y},\mu = \mathcal{IG}(T/2, Ts_\mu^2/2)}$$

Inverse Gamma Distribution

- Inverse Gamma Distribution:

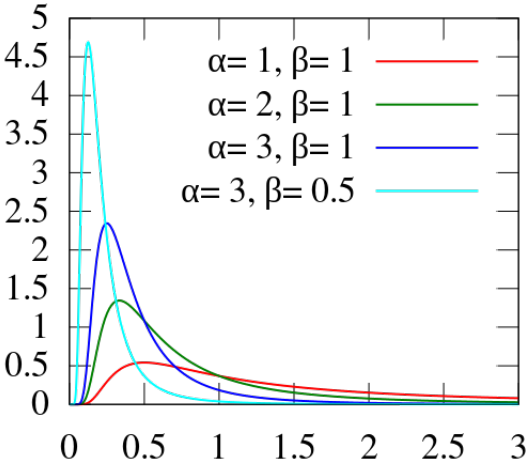
$$p(x|\alpha,\beta) = \frac{\beta^{\alpha} x^{-\alpha-1} exp[-\beta/x]}{\Gamma_1(\alpha)}$$

  where $\Gamma(\cdot)$ denotes the gamma function.

- With $x = \sigma^2$, $\alpha = T$, and $\beta = Ts_\mu^2$:

$$\begin{aligned} p(\sigma^2|T/2, Ts_\mu^2/2) &= \frac{(Ts_\mu^2/2)^{T/2}(\sigma^2)^{-T/2-1} exp[\frac{-Ts_\mu^2}{2\sigma^2}]}{\Gamma_1(T/2)} \\ &\propto \underbrace{\left(\frac{Ts_\mu^2}{2}\right)^{\frac{T}{2}}}_{constant} (\sigma^2)^{-T-1} exp\left[-\frac{1}{2}\frac{Ts_\mu^2}{\sigma^2}\right] \end{aligned}$$

Inverse Gamma Distribution



Chi-squared Distribution

- The Chi-squared distribution with T degrees of freedom is the distribution of a sum of the squares of T independent standard normal random variables.

- Let $Z_{1,t} \sim \text{i.i.d.}\mathcal{N}(0,1)$, define

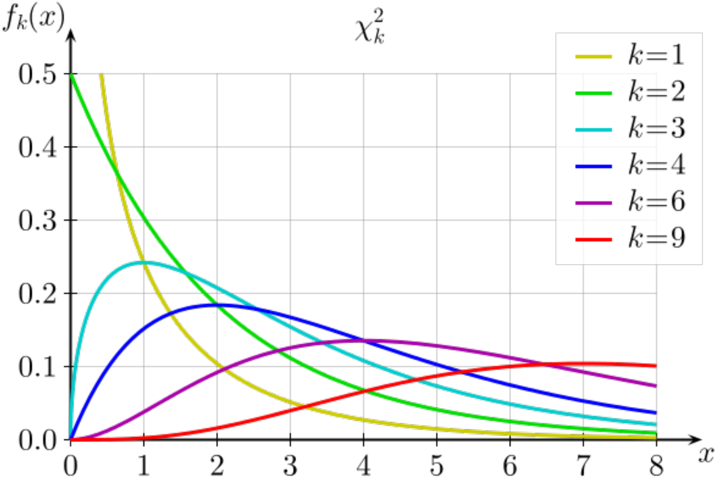
$$s = \sum_{t=1}^{T} Z_{1,t} Z'_{1,t}$$

- Then s has a Chi-squared distribution with T degrees of freedom

- Chi-square distribution:

$$p(s|T) = \frac{s^{T/2-1} exp[-s/2]}{2^{T/2}\Gamma(T/2)}$$

Chi-squared Distribution



The Wishart distribution:

Let $Z_t = (Z_{1,t}, ..., Z_{m,t}) \sim \mathrm{i.i.d.}\mathcal{N}(0, H)$, define

$$S = \sum_{t=1}^{T} Z_t Z_t'$$

Notice: S/T is the sample covariance matrix of Z_t based on a sample of size T.

Then we say that S has a Wishart distribution, with scale S and T degrees of freedom. We write: $S \sim \mathcal{W}(H, T)$

We say that Σ has Inverted Wishart distribution with scale Ψ and T degrees of freedom, and write $\Sigma \sim \mathcal{IW}(\Psi, T)$, if $\Sigma^{-1} \sim \mathcal{W}(\Psi^{-1}, T)$

$$p(\Sigma) \propto |\Psi|^{T/2} |\Sigma|^{-(T+m+1)/2} \exp\left\{-\frac{1}{2}\mathrm{tr}\left[\Sigma^{-1}\Psi\right]\right\}$$

Remark: In the univariate case ($m = 1$) we have the Gamma distributions and the Inverted Gamma distribution.

Joint inference on location and scale

- T independent observations $\mathbf{y} = (y_1, y_2, ..., y_T)'$ drawn from a normal distribution with unknown mean μ and unknown variance σ^2

$$y_t = \mu + \varepsilon_t \qquad \varepsilon_t \sim \mathcal{N}(0, \sigma^2)$$

- Prior $p(\mu, \log \sigma^2) \propto 1 \Rightarrow p(\mu, \sigma^2) \propto \frac{1}{\sigma^2}$
- $p(\mu, \sigma^2 | \mathbf{y}) \propto (2\pi\sigma^2)^{-T/2} \exp\left[-\frac{T}{2\sigma^2}\left[s^2 + (\mu - \bar{\mu})^2\right]\right](\sigma^2)^{-1}$

$$\propto \underbrace{(\sigma^2)^{-\frac{T+1}{2}} \exp\left[-\frac{1}{2}\frac{Ts^2}{\sigma^2}\right]}_{Ts^2/\chi^2_{(T-1)}} \underbrace{(\sigma^2)^{-1/2} \exp\left[-\frac{1}{2\,(\sigma^2/T)}(\mu - \bar{\mu})^2\right]}_{\mathcal{N}(\bar{\mu},\sigma^2/T)}$$

- Posteriors
  - $\mu | \sigma^2, \mathbf{y} \sim \mathcal{N}\left(\bar{\mu}, \sigma^2/T\right); \quad \sigma^2 | \mathbf{y} \sim \frac{(T-1)s^2}{\chi^2_{(T-1)}}; \quad \sigma^2 | \mu, \mathbf{y} \sim \frac{Ts^2_{\mu}}{\chi^2_{(T)}}$

Joint inference on location and scale

How to compute the marginal of μ, the joint of μ, σ^2

- Analytically
- Numerically
  1) Generate $(\sigma^2)^{(j)}$ by drawing from $p(\sigma^2|\mathbf{y})$
  2) Generate $\mu^{(j)}$ a drawing from $p\left(\mu|\mathbf{y}, (\sigma^2)^{(j)}\right)$
- Approximate moments and quantiles from the empirical distribution of the generated parameters.
- **Remark:** The algorithm can be used to approximate the distribution of any function of the parameters

Joint inference on location and scale

What if only the conditional posteriors were known

- **Gibbs sample**
  1) Generate $(\sigma^2)^{(j)}$ by drawing from $p(\sigma^2|\mathbf{y}, \mu^{(j-1)})$
  2) Generate $\mu^{(j)}$ a drawing from $p\left(\mu|\mathbf{y}, (\sigma^2)^{(j)}\right)$
- Starting from arbitrary $\mu^{(0)}$, repeating step (1) and (2) to obtain $\mu^{(j)}, \left(\sigma^2\right)^{(j)}, j = 1, ..., J$
- For large J we obtain independent draws the joint distribution $p(\mu, \sigma^2|\mathbf{y})$.
- Approximate moments and quantiles from the empirical distribution of the generated parameters, after discarding some initial draws.
- **Remark:** The algorithm can be used to approximate the distribution of any function of the parameters

Joint inference on location and scale

- **Informative Priors:** Use the posterior obtained from a dummy sample $\mathbf{y}_d$ of size T_d

$$p(\mu, \sigma) = p(\mu|\sigma)p(\sigma)$$

$$p(\mu|\sigma) = p(\mu|\mathbf{y}_d, \sigma) \sim \mathcal{N}\left(\bar{\mu}_d, \sigma^2 / T_d\right)$$

$$p(\sigma) = p(\sigma|\mathbf{y}_d) \propto \left(\sigma^2\right)^{-\frac{T_d+1}{2}} \exp\left(-\frac{(T_d-1)s_d^2}{2\sigma^2}\right)$$

- The **Posterior**

$$p(\mu, \sigma^2|\mathbf{y}) \propto p(\mathbf{y}|\mu, \sigma^2)p(\mathbf{y}_d|\mu, \sigma^2)\frac{1}{\sigma^2}$$

$$= p(\mathbf{y}, \mathbf{y}_d|\mu, \sigma^2)\frac{1}{\sigma^2} = p(\mu, \sigma^2|\mathbf{y}, \mathbf{y}_d)$$

This is proportional to the joint likelihood of the "actual" data and the "dummy" data:

Joint inference on location and scale

- The **Posterior**

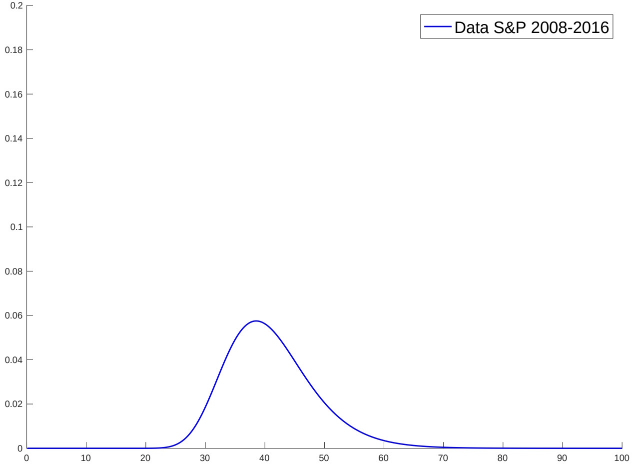
$$p(\mu, \sigma^2|\mathbf{y}) \propto p(\mathbf{y}|\mu, \sigma^2) p(\mathbf{y}_d|\mu, \sigma^2) \frac{1}{\sigma^2}$$

$$= p(\mathbf{y}, \mathbf{y_d}|\mu, \sigma^2) \frac{1}{\sigma^2} = p(\mu, \sigma^2|\mathbf{y}, \mathbf{y}_d)$$

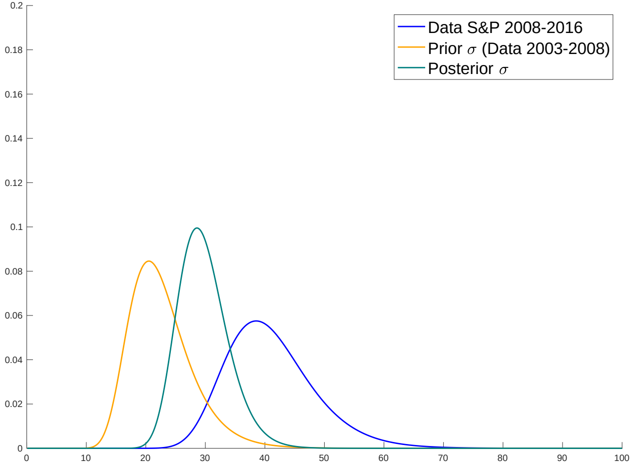
$$\mu|\mathbf{y}, \sigma^2 \sim \mathcal{N}\left(\frac{1}{T + T_d}(T_d \hat{\mu}_d + T\hat{\mu}), \frac{1}{(T + T_d)}\sigma^2\right) = \mathcal{N}\left(\hat{u}_*, \frac{1}{T_*}\sigma^2\right)$$

$$\sigma^2|\mathbf{y} \sim \left((T-1)s^2 + (T_d - 1)s_d^2\right) / \chi^2_{(T+T_d-1)} = \left((T_* - 1)s_*^2\right) / \chi^2_{(T_*-1)}$$

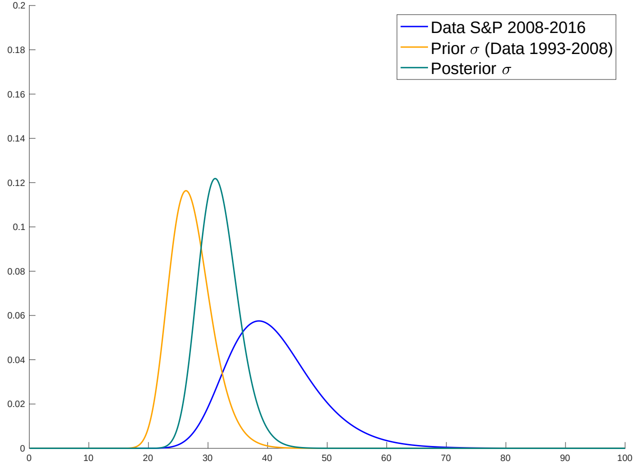
S&P 500 returns– $p(\sigma^2|\mathbf{y})$



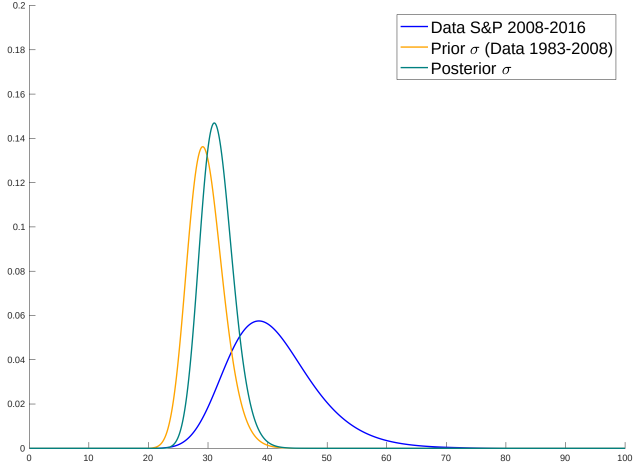
S&P 500 returns– $p(\sigma^2|\mathbf{y})$



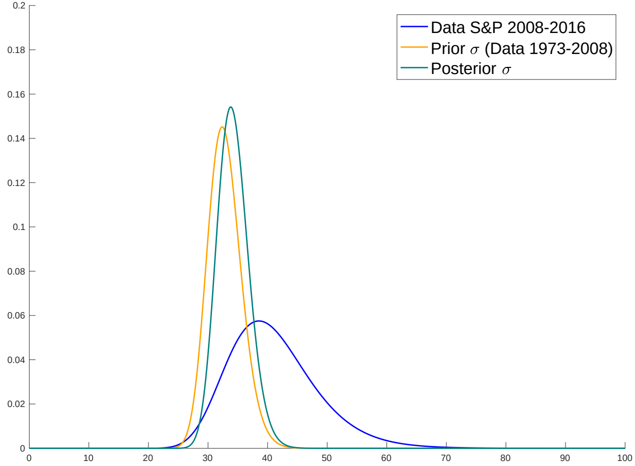
S&P 500 returns– $p(\sigma^2|\mathbf{y})$



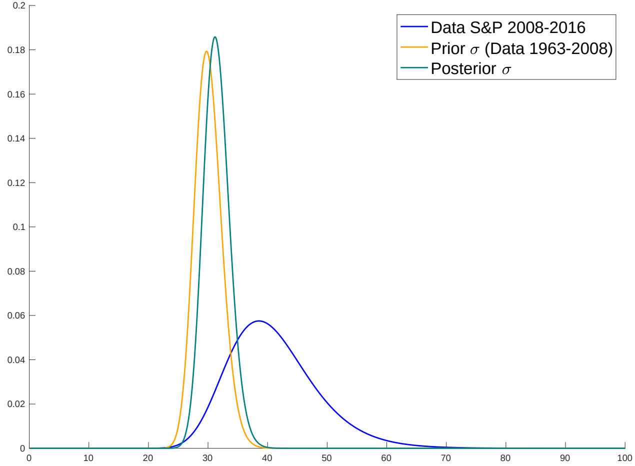
S&P 500 returns– $p(\sigma^2|\mathbf{y})$



S&P 500 returns– $p(\sigma^2|\mathbf{y})$



S&P 500 returns– $p(\sigma^2|\mathbf{y})$



Conjugate Priors

- Natural conjugate priors are the priors such that the posteriors distributions are the same distributional family of the priors
- In the Normal regression model the natural conjugate prior is the Normal-inverted Gamma (Wishart) prior
- It has the desirable property that the prior can be generated by a sample generated by the same model

General NIG(W) priors:

$$\beta|\sigma^2 \sim \mathcal{N}(\beta_0, V_0\sigma^2)$$

$$\sigma^2 \sim \mathcal{IW}(\nu_0\sigma_0^2, \nu_0)$$

Conjugate Prior

The Normal-inverted Wishart prior

$$\beta|\sigma^2 \sim \mathcal{N}(\beta_0, V_0\sigma^2)$$

$$\sigma^2 \sim \mathcal{IW}(\nu_0\sigma_0^2, \nu_0)$$

Can be implemented by using 'artificial' dummy observations

Idea: Need to find x_d and y_d such that:

$$(x_d' x_d)^{-1} x_d' y_d = \beta_0 \text{ and } (x_d' x_d)^{-1} = V_0$$

$$(y_d - x_d\beta_0)'(y_d - x_d\beta_0) = \nu_0\sigma_0^2 \text{ and } T_d - k = \nu_0$$

Conjugate Prior

The posterior is:

$$\beta | x, y, \sigma^2 \sim \mathcal{N}(\hat{\beta}, V\sigma^2)$$

$$\sigma^2 | x, y \sim \mathcal{IW}(\nu, \nu s^2)$$

where

- $\beta = (x'x + V_0^{-1})^{-1}(V_0^{-1}\beta_0 + x'y)$
- $V = (x'x + V_0^{-1})^{-1}$
- $\nu = T + \nu_0$
- $\nu s^2 = (y - x\hat{\beta})'(y - x\hat{\beta}) + \nu_0\sigma_0^2 + \left(\hat{\beta} - \beta_0\right)' V_0^{-1} \left(\hat{\beta} - \beta_0\right)$

Linear regression with conjugate priors

- The Model

$$\underset{T\times 1}{Y} = \underset{T\times k}{X} \underset{k\times 1}{\beta} + \underset{T\times 1}{\epsilon} \qquad \epsilon \sim N(0, \sigma^2 I_T)$$

- The Normal-inverted Wishart prior

$$\beta|\sigma^2 \sim \mathcal{N}(\beta_0, V_0\sigma^2); \quad \sigma^2 \sim \mathcal{IW}(\nu_0\sigma_0^2, \nu_0)$$

- The Posterior

$$\beta|x, y, \sigma^2 \sim \mathcal{N}(\hat{\beta}, V\sigma^2); \quad \sigma^2|x, y \sim \mathcal{IW}(\nu, \nu s^2)$$

  - $\hat{\beta} = (x'x + V_0^{-1})^{-1}(V_0^{-1}\beta_0 + x'y)$ and $V = (x'x + V_0^{-1})^{-1}$
  - $s^2 = \frac{1}{\nu}\left[y - x\hat{\beta})'(y - x\hat{\beta}) + \nu_0\sigma_0^2 + \left(\hat{\beta} - \beta_0\right)' V_0^{-1} \left(\hat{\beta} - \beta_0\right)\right]$
  - $\nu = T + \nu_0$

Conjugate Prior

Example: Simple prior:

$$\beta|\sigma^2 \sim \mathcal{N}\left(0, \frac{\sigma^2}{\tau^2} I_k\right)$$

for σ^2 given

- Can be implemented by using additional dummy observations.
- Need to find x_d and y_d such that:

$$(x_d' x_d)^{-1} x_d' y_d = 0$$

  and

$$x_d' x_d = \tau^2 I_k$$

- Set:

$$x_d = \tau I_k \qquad y_d = 0_{k \times 1}$$

Conjugate Prior

The posterior is:

$$\beta | x, y, \sigma^2 \sim \mathcal{N}(\hat{\beta}, V\sigma^2)$$

where

$$\hat{\beta} = (x'x + \tau^2 I_k)^{-1} x'y$$

and

$$V = (x'x + \tau^2 I_k)^{-1}$$

- τ is a tightness parameter, controls the weight we give to the prior.
- $\tau \to \infty \implies$ posterior $=$ prior ('dogmatic')
- $\tau \to 0 \implies$ OLS ('flat')

APPENDIX

Linear Regression: The Model

$$\underset{T\times 1}{Y} = \underset{T\times k}{X} \underset{k\times 1}{\beta} + \underset{T\times 1}{\epsilon} \qquad \epsilon \sim N(0, \sigma^2 I_T)$$

- The parameters of the models are:

$$\theta = (\beta', \sigma) \;\; and \;\; \Theta = \mathbb{R}^k \cup \mathbb{R}^+$$

- The probability of Y given the regressors X and the parameters θ is given by:

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

Linear Regression: The Model

We can rewrite as

$$p(Y|X,\beta,\sigma) \propto \left(\frac{1}{\sigma^2}\right)^{T/2} exp\left\{-\frac{1}{2\sigma^2}[\nu s^2 + (\beta-\hat{\beta})'X'X(\beta-\hat{\beta})]\right\}$$

where

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

$$s^2 = (Y - X\hat{\beta})'(Y - X\hat{\beta})/\nu$$

and

$$\nu = T - k$$

Linear Regression: The Prior

- Suppose that σ^2 is known.
- We assume a non-informative prior on the regression coefficient: uniform distribution over the real line

$$p(\beta_i|\sigma^2) = 1 \quad -\infty < \beta_i < \infty \quad \forall i = 1, ..., k$$

$$p(\beta|\sigma^2) \propto 1$$

- **Remark**: The prior is improper

$$\int_{-\infty}^{\infty} p(\mu) d\mu = \infty \neq 1$$

Linear Regression: The Posterior

- Collect the data y, x (realisation of Y, X). Use the data to evaluate the likelihood

$$\mathcal{L}(y|x, \beta, \sigma) = p(y|x, \beta, \sigma)$$

- ... and update the belief from the Bayes rule

$$\begin{aligned} p(\beta \mid y, x, \sigma^2) &\propto \mathcal{L}(y|x, \beta, \sigma) p(\beta, \sigma) \\ &\propto \left(\frac{1}{\sigma^2}\right)^{T/2} exp\left\{-\frac{1}{2\sigma^2}\left[\nu s^2 + (\beta - \hat{\beta})' x' x (\beta - \hat{\beta})\right]\right\} \end{aligned}$$

Linear Regression: The Posterior

- Up to a scale

$$p(\beta|y,x,\sigma^2) \propto \left|\frac{x'x}{\sigma^2}\right|^{1/2} exp\left\{-\frac{1}{2}\left[(\beta-\hat{\beta})'\frac{x'x}{\sigma^2}(\beta-\hat{\beta})\right]\right\}$$

- Hence we get

$$\beta|\sigma^2 \sim \mathcal{N}(\hat{\beta},(x'x)^{-1}\sigma^2)$$

- **Remark**: Same result as with Maximum likelihood

The Case of Unknown σ^2

- Let's consider now the case for unknown σ^2 and add a prior:

$$p(\log \sigma^2) \propto 1 \quad \Rightarrow \quad p(\sigma^2) \propto \frac{1}{\sigma^2}$$

- The Posterior Likelihood is

$$\begin{aligned}
p(\beta, \sigma^2 \mid y, x) &\propto \mathcal{L}(y|x, \beta, \sigma) p(\beta, \sigma) \\
&\propto \underbrace{\left|\frac{x'x}{\sigma^2}\right|^{1/2} exp\left\{-\frac{1}{2\sigma^2}(\beta - \hat{\beta})' x' x (\beta - \hat{\beta})\right\}}_{p(\beta|y,x,\sigma^2)} \\
&\quad \times \underbrace{\left(\frac{1}{\sigma^2}\right)^{\nu/2+1} exp\left\{-\frac{1}{2\sigma^2}\nu s^2\right\}}_{p(\sigma^2|x,y)} \\
&\propto \mathcal{N}(\hat{\beta}, (x'x)^{-1}\sigma^2) \times \mathcal{IW}(\nu s^2, \nu)
\end{aligned}$$

Bayesian Regression with Improper Prior

The Model:

$$\underset{T\times 1}{Y} = \underset{T\times k}{X} \underset{k\times 1}{\beta} + \underset{T\times 1}{\epsilon} \qquad \epsilon \sim N(0, \sigma^2 I_T)$$

The Priors:

$$p(\beta|\sigma^2) \propto 1$$

$$p(\sigma^2) \propto \frac{1}{\sigma^2}$$

The Posterior:

$$\sigma^2|x, y \sim \mathcal{IW}(\nu s^2, \nu)$$

$$\beta|x, y, \sigma^2 \sim \mathcal{N}(\hat{\beta}, (x'x)^{-1}\sigma^2)$$

The marginal posterior distribution of β

- We know
  - $p(\sigma^2|x,y)$: marginal posterior distribution of σ^2
  - $p(\beta|\sigma^2|x,y)$: conditional posterior of β distribution
- How to get
  - $p(\beta|x,y)$: marginal posterior distribution of β
  - $p(\beta,\sigma^2|x,y)$: joint posterior distribution
- Simulation:
  i) Generate $(\sigma^2)^{(j)}$ by drawing from $p(\sigma^2|x,y)$
  ii) Generate $\beta^{(j)}$ a drawing from $p\left(\beta|x,y,(\sigma^2)^{(j)}\right)$

Gibbs sampler
If instead only $p(\sigma^2|x,y,\beta)$ was available
i) Generate $(\sigma^2)^{(j)}$ by drawing from $p(\sigma^2|x,y,\beta^{(j-1)})$
ii) Generate $\beta^{(j)}$ a drawing from $p\left(\beta|x,y,(\sigma^2)^{(j)}\right)$

Starting from arbitrary $\beta^{(0)}$, repeating step i) and ii) to obtain $\beta^{(j)},\sigma^{(j)},j=1,\ldots,J$
For large J we obtain independent draws the joint distribution $p(\beta,\sigma^2|x,y)$.
Discard initial draws.

Bayesian Regression with Informative Priors

- Use the **posterior from a dummy sample** y_d, x_d of length T_d as a prior for the sample y, x of length T generated by the same model
- The **posterior** from the sample y_d, x_d using the **flat prior** is,

$$p(\sigma^2) = p(\sigma^2 | x_d, y_d) = \mathcal{IW}(\nu_d s_d^2, \nu_d)$$

$$p(\beta|\sigma^2) = p(\beta|x_d, y_d, \sigma^2) = \mathcal{N}(\hat{\beta}_d, (x_d' x_d)^{-1}\sigma^2)$$

where

$$\nu_d = T_d - k$$

$$\hat{\beta}_d = (x_d' x_d)^{-1}(x_d' y_d)^{-1}$$

and

$$s_d^2 = (y_d - x_d\hat{\beta}_d)'(y_d - x_d\hat{\beta}_d)/\nu_d$$

Informative Priors

Combining prior and likelihood, we obtain

$$
\begin{aligned}
\mathcal{L}(\beta,\sigma|y,x,y_d,x_d) &\propto \mathcal{L}(y|x,\beta,\sigma)p(\beta|x_d,y_d,\sigma^2)p(\sigma^2|x_d,y_d)\\
&\propto \underbrace{\left(\frac{1}{\sigma^2}\right)^{T/2} exp\left\{-\frac{1}{2\sigma^2}(y-x\beta)'(y-x\beta)\right\}}_{\mathcal{L}(y|x,\beta,\sigma)}\\
&\quad \underbrace{\times\left(\frac{1}{\sigma^2}\right)^{k/2} exp\left\{-\frac{1}{2\sigma^2}(\beta-\hat{\beta}_d)'x_d'x_d(\beta-\hat{\beta}_d)\right\}}_{p(\beta|x_d,y_d,\sigma^2)}\\
&\quad \underbrace{\times\left(\frac{1}{\sigma^2}\right)^{\nu_d/2+1} exp\left\{-\frac{1}{2\sigma^2}\nu_d s_d^2\right\}}_{p(\sigma^2|x_d,y_d)}\\
&\propto \left(\tfrac{1}{\sigma^2}\right)^{(T^*+1)/2} exp\left\{-\tfrac{1}{2\sigma^2}\left[(y^*-x^*\beta)'(y^*-x^*\beta)\right]\right\}
\end{aligned}
$$

Augmented data $y^* = (y_d'\ ,\ y')'$ and $x^* = (x_d'\ ,\ x')'$ and $T^* = T + T_d$

Informative Priors

Hence

$$\beta|\sigma, x, y \sim \mathcal{N}(\hat{\beta}, (x^{*\prime}x^{*})^{-1}\sigma^{2})$$

$$\sigma^{2}|x, y \sim \mathcal{IW}(\nu s^{2}, \nu)$$

where

$$\hat{\beta} = (x^{*\prime}x^{*})^{-1}x^{*\prime}y^{*}$$

$$s^{2} = \frac{1}{\nu}(y^{*} - x^{*}\hat{\beta})'(y^{*} - x^{*}\hat{\beta})$$

and

$$\nu = T^{*} - k$$

Informative Priors

$$\beta|\sigma, x, y \sim \mathcal{N}(\hat{\beta}, (x'x + x_d'x_d)^{-1}\sigma^2)$$
$$\sigma^2|x, y \sim \mathcal{IW}(\nu s^2, \nu)$$

where

$$\hat{\beta} = (x^{*'}x^*)^{-1}x^{*'}y^* = (x'x + x_d'x_d)^{-1}(x_d'y_d + x'y)$$
$$\nu s^2 = (y - x\hat{\beta})'(y - x\hat{\beta}) + (y_d - x_d\hat{\beta})'(y_d - x_d\hat{\beta})$$
$$\nu = T + T_d - k$$

Remark: If we had pooled the two samples and used a diffuse prior, the resulting posterior would had been exactly the same.