• **Econometric model:** collection of probability distributions $p(Y|\theta)$ indexed by parameter $\theta \in \Theta$. Examples: VAR, DSGE model, ...

• The “easy” part: pick values for parameter vector $\theta \implies$ determine properties of model-simulated data $Y^{sim}(\theta)$.

• **Statistical inference:** observed data $Y^{obs} \implies$ determine suitable values for parameter vector $\theta$.

• **Basic Idea:** choose $\theta$ such that $Y^{sim}(\theta)$ look like $Y^{obs}$.

• **Goals:** estimates $\hat{\theta}$ as well as measures of uncertainty associated with these estimates.
Good Measures of Uncertainty are Important

NK Phillips Curve

\[ \tilde{\pi}_t = \gamma_b \tilde{\pi}_{t-1} + \gamma_f \mathbb{E}_t[\tilde{\pi}_{t+1}] + \kappa \tilde{M}C_t \]
• Econometric model generates a family of probability distributions $p(Y|\theta)$, $\theta \in \Theta$.

• Thought experiment: data are generated from the econometric model conditional on some “true” parameter $\theta_0$.

• The parameter vector $\theta$ is globally identifiable at $\theta_0$ if

$$p(Y|\theta) = p(Y|\theta_0) \quad \text{implies} \quad \theta = \theta_0.$$  

• **Treatment of $Y$:**
  • Pre-experimental perspective: the sample is not yet observed and condition needs to hold with probability one under the distribution $p(Y|\theta_0)$.
  • Post-experimental perspective: sample has been observed, parameter $\theta$ may be identifiable for some trajectories $Y$, but not for others.

• **Example:**

$$y_{1,t} | (\theta, y_{2,t}) \sim iidN(\theta y_{2,t}, 1), \quad y_{2,t} = \begin{cases} 0 & \text{w.p. } 1/2 \\ \sim iidN(0, 1) & \text{w.p. } 1/2 \end{cases}$$

With probability (w.p.) 1/2, one observes a trajectory along which $\theta$ is not identifiable because $y_{2,t} = 0$ for all $t$. 
• **Frequentist:**
  • pre-experimental perspective;
  • condition on “true” but unknown $\theta_0$;
  • treat data $Y$ as random;
  • study behavior of estimators and decision rules under repeated sampling.

• **Bayesian:**
  • post-experimental perspective;
  • condition on observed sample $Y$;
  • treat parameter $\theta$ as unknown and random;
  • derive estimators and decision rules that minimize expected loss (averaging over $\theta$) conditional on observed $Y$. 
• Suppose $Y_1$ and $Y_2$ are independently and identically distributed and

$$P_{\theta}^{Y_i}\{Y_i = \theta - 1\} = \frac{1}{2}, \quad P_{\theta}^{Y_i}\{Y_i = \theta + 1\} = \frac{1}{2}$$

• Consider the following coverage set

$$C(Y_1, Y_2) = \begin{cases} \frac{1}{2}(Y_1 + Y_2) & \text{if } Y_1 \neq Y_2 \\ Y_1 - 1 & \text{if } Y_1 = Y_2 \end{cases}$$

• **Pre-experimental perspective:** $C(Y_1, Y_2)$ is a 75% confidence interval. The probability (under repeated sampling, conditional on $\theta$) that the confidence interval contains $\theta$ is 75%.

• **Post-experimental perspective:** we are “100% confident” that $C(Y_1, Y_2)$ contains the “true” $\theta$ if $Y_1 \neq Y_2$, whereas we are only “50% percent” confident if $Y_1 = Y_2$. 
Bayesian Inference - Overview

Model of interest \((M_1)\) is assumed to be correctly specified, i.e. we believe the probabilistic structure is rich enough to assign high probability to the salient features of macroeconomic time series.

- Initial state of knowledge summarized in prior distribution \(p(\theta)\).
- Update in view of data \(Y\) to obtain posterior distribution \(p(\theta|Y)\):
  \[
p(\theta|Y, M_1) = \frac{p(Y|\theta, M_1)p(\theta|M_1)}{p(Y|M_1)}, \quad p(Y|M_1) = \int p(Y|\theta, M_1)p(\theta|M_1)d\theta.
  \]
- Make decisions that minimize posterior expected loss:
  \[
  \delta_* = \arg\min_{\delta \in D} \int L(h(\theta), \delta)p(\theta|Y, M_1)d\theta.
  \]
- Place probabilities on competing models and update:
  \[
  \frac{\pi_{1,T}}{\pi_{2,T}} = \frac{\pi_{1,0} p(Y|M_1)}{\pi_{2,0} p(Y|M_2)}.
  \]
Bayesian Inference – More Details

- **Ingredients of Bayesian Analysis:**
  - Likelihood function \( p(Y|\theta) \)
  - Prior density \( p(\theta) \)
  - Marginal data density \( p(Y) = \int p(Y|\theta)p(\theta)d\phi \)

- **Bayes Theorem:**
  \[
p(\theta|Y) = \frac{p(Y|\theta)p(\theta)}{p(Y)} \propto p(Y|\theta)p(\theta)
  \]

- **Implementation:** usually by generating a sequence of draws (not necessarily iid) from posterior
  \[
  \theta^i \sim p(\theta|Y), \quad i = 1, \ldots, N
  \]

- **Algorithms:** direct sampling, accept/reject sampling, importance sampling, Markov chain Monte Carlo sampling, sequential Monte Carlo sampling...
Warm-Up: Linear Regression / AR Models

- Consider AR(1) model:
  \[ y_t = y_{t-1} \theta + u_t, \quad u_t \sim iidN(0,1). \]

- Let \( x_t = y_{t-1} \). Write as
  \[ y_t = x_t' \theta + u_t, \quad u_t \sim iidN(0,1), \]
  or
  \[ Y = X\theta + U. \]

We can easily allow for multiple regressors. Assume \( \theta \) is \( k \times 1 \).

- Notice: we treat the variance of the errors as know. The generalization to unknown variance is straightforward but tedious.

- Likelihood function:
  \[ p(Y|\theta) = (2\pi)^{-T/2} \exp \left\{ -\frac{1}{2} (Y - X\theta)'(Y - X\theta) \right\}. \]
A Convenient Prior

• Prior:
  \[ \theta \sim \mathcal{N}(0_{k \times 1}, \tau^2 I_{k \times k}) \]
  \[ p(\theta) = (2\pi \tau^2)^{-k/2} \exp \left\{ -\frac{1}{2\tau^2} \theta'\theta \right\} \]

• Large \( \tau \) means diffuse prior.
• Small \( \tau \) means tight prior.
Deriving the Posterior

• Bayes Theorem:

\[ p(\theta | Y) \propto p(Y | \theta) p(\theta) \]
\[ \propto \exp \left\{ -\frac{1}{2} [(Y - X\theta)'(Y - X\theta) + \tau^{-2}\theta'\theta] \right\}. \]

• Guess: what if \( \theta | Y \sim N(\bar{\theta}_T, \bar{V}_T) \). Then

\[ p(\theta | Y) \propto \exp \left\{ -\frac{1}{2} (\theta - \bar{\theta}_T)'\bar{V}_T^{-1}(\theta - \bar{\theta}_T) \right\}. \]

• Rewrite exponential term

\[ Y'Y - \theta'X'Y - Y'X\theta + \theta'X'X\theta + \tau^{-2}\theta'\theta \]
\[ = Y'Y - \theta'X'Y - Y'X\theta + \theta'(X'X + \tau^{-2}I)\theta \]
\[ = (\theta - (X'X + \tau^{-2}I)^{-1}X'Y)' \left( X'X + \tau^{-2}I \right) \]
\[ \times (\theta - (X'X + \tau^{-2}I)^{-1}X'Y) \]
\[ + Y'Y - Y'X(X'X + \tau^{-2}I)^{-1}X'Y. \]
Deriving the Posterior

- Exponential term is a quadratic function of $\theta$.
- Deduce: posterior distribution of $\theta$ must be a multivariate normal distribution

$$\theta|Y \sim N(\bar{\theta}_T, \bar{V}_T)$$

with

$$\bar{\theta}_T = (X'X + \tau^{-2}I)^{-1}X'Y$$
$$\bar{V}_T = (X'X + \tau^{-2}I)^{-1}.$$

- $\tau \rightarrow \infty$:

$$\theta|Y \overset{approx}{\sim} N\left(\hat{\theta}_{mle}, (X'X)^{-1}\right).$$

- $\tau \rightarrow 0$:

$$\theta|Y \overset{approx}{\sim} \text{Pointmass at 0}$$
• Plays an important role in Bayesian model selection and averaging.

• Write

\[ p(Y) = \frac{p(Y|\theta)p(\theta)}{p(\theta|Y)} \]

\[ = \exp \left\{ -\frac{1}{2} \left[ Y'Y - Y'X(X'X + \tau^{-2}I)^{-1}X'Y \right] \right\} \]

\[ \times (2\pi)^{-T/2} |I + \tau^2X'X|^{-1/2} \]

• The exponential term measures the goodness-of-fit.

• \(|I + \tau^2X'X|\) is a penalty for model complexity.
• We will often abbreviate posterior distributions $p(\theta|Y)$ by $\pi(\theta)$ and posterior expectations of $h(\theta)$ by

$$
\mathbb{E}_\pi[h] = \mathbb{E}_\pi[h(\theta)] = \int h(\theta)\pi(\theta)d\theta = \int h(\theta)p(\theta|Y)d\theta.
$$

• We will focus on algorithms that generate draws $\{\theta^i\}_{i=1}^N$ from posterior distributions of parameters in time series models.

• These draws can then be transformed into objects of interest, $h(\theta^i)$, and under suitable conditions a Monte Carlo average of the form

$$
\bar{h}_N = \frac{1}{N} \sum_{i=1}^N h(\theta^i) \approx \mathbb{E}_\pi[h].
$$

• Strong law of large numbers (SLLN), central limit theorem (CLT)…
Posterior Sampler 1: Direct Sampling

- In the simple linear regression model with Gaussian posterior it is possible to sample directly.

- For \( i = 1 \) to \( N \), draw \( \theta^i \) from \( N(\bar{\theta}, \bar{V}_{\theta}) \).

- Provided that \( \mathbb{V}_\pi[h(\theta)] < \infty \) we can deduce from Kolmogorov’s SLLN and the Lindeberg-Levy CLT that

\[
\bar{h}_N \xrightarrow{a.s.} \mathbb{E}_\pi[h]
\]

\[
\sqrt{N} \left( \bar{h}_N - \mathbb{E}_\pi[h] \right) \implies N(0, \mathbb{V}_\pi[h(\theta)]).
\]
Posterior Sampler 2: Importance Sampling

• Approximate $\pi(\cdot)$ by using a different, tractable density $g(\theta)$ that is easy to sample from.

• For more general problems, posterior density may be non-normalized. So we write

$$
\pi(\theta) = \frac{p(Y|\theta)p(\theta)}{p(Y)} = \frac{f(\theta)}{\int f(\theta)d\theta}.
$$

• Importance sampling is based on the identity

$$
E_\pi[h(\theta)] = \int h(\theta)\pi(\theta)d\theta = \frac{\int_\Theta h(\theta)\frac{f(\theta)}{g(\theta)}g(\theta)d\theta}{\int_\Theta \frac{f(\theta)}{g(\theta)}g(\theta)d\theta}.
$$

• The ratio

$$
w(\theta) = \frac{f(\theta)}{g(\theta)}
$$

is called the (unnormalized) importance weight.
Importance Sampling

1. For $i = 1$ to $N$, draw $\theta^i \overset{iid}{\sim} g(\theta)$ and compute the unnormalized importance weights:

$$w^i = w(\theta^i) = \frac{f(\theta^i)}{g(\theta^i)}.$$

2. Compute the normalized importance weights

$$W^i = \frac{w^i}{\frac{1}{N} \sum_{i=1}^{N} w^i}.$$

An approximation of $\mathbb{E}_\pi[h(\theta)]$ is given by

$$\bar{h}_N = \frac{1}{N} \sum_{i=1}^{N} w^i h(\theta^i) = \frac{1}{N} \sum_{i=1}^{N} W^i h(\theta^i).$$
If \( \theta^i \)'s are draws from \( g(\cdot) \) then

\[
\mathbb{E}_\pi[h] \approx \frac{1}{N} \sum_{i=1}^{N} h(\theta^i) w(\theta^i), \quad w(\theta) = \frac{f(\theta)}{g(\theta)}.
\]
• Since we are generating iid draws from $g(\theta)$, it’s fairly straightforward to derive a CLT:

$$\sqrt{N}(\bar{h}_N - \mathbb{E}_\pi[h]) \implies N(0, \Omega(h)), \text{ where } \Omega(h) = \nabla_g[(\pi/g)(h - \mathbb{E}_\pi[h])].$$

• It can be shown that

$$\sqrt{N}(\bar{h}_N - \mathbb{E}_\pi[h]) \implies N(0, \Omega(h)), \text{ where } \Omega(h) = \nabla_g[(\pi/g)(h - \mathbb{E}_\pi[h])].$$

• Using a crude approximation (see, e.g., Liu (2008)), we can factorize $\Omega(h)$ as follows:

$$\Omega(h) \approx \nabla_\pi[h](\nabla_g[\pi/g] + 1).$$

The approximation highlights that the larger the variance of the importance weights, the less accurate the Monte Carlo approximation relative to the accuracy that could be achieved with an iid sample from the posterior.

• Users often monitor

$$ESS = N \frac{\nabla_\pi[h]}{\Omega(h)} \approx \frac{N}{1 + \nabla_g[\pi/g]}.$$
The posterior expected loss associated with a decision \( \delta(\cdot) \) is given by

\[
\rho(\delta(\cdot)|Y) = \int_{\Theta} L(\theta, \delta(Y)) p(\theta|Y) d\theta.
\]

A Bayes decision is a decision that minimizes the posterior expected loss:

\[
\delta^*(Y) = \arg\min_{\delta} \rho(\delta(\cdot)|Y).
\]

Since in most applications it is not feasible to derive the posterior expected risk analytically, we replace \( \rho(\delta(\cdot)|Y) \) by a Monte Carlo approximation of the form

\[
\bar{\rho}_N(\delta(\cdot)|Y) = \frac{1}{N} \sum_{i=1}^{N} L(\theta^i, \delta(\cdot)).
\]

A numerical approximation to the Bayes decision \( \delta^*(\cdot) \) is then given by

\[
\delta^*_N(Y) = \arg\min_{\delta} \bar{\rho}_N(\delta(\cdot)|Y).
\]
• **Point estimation:**
  
  • Quadratic loss: posterior mean
  
  • Absolute error loss: posterior median

• **Interval/Set estimation** $\mathbb{P}_\pi \{ \theta \in C(Y) \} = 1 - \alpha$:
  
  • highest posterior density sets
  
  • equal-tail-probability intervals
• Interpret point estimation as decision problem.

• Consider quadratic loss:

\[ L(\theta, \delta) = (\theta - \delta)^2 \]

• Optimal decision rule is obtained by minimizing

\[ \min_{\delta \in D} \mathbb{E}_\pi[(\theta - \delta)^2] \]

• Solution: \( \delta = \mathbb{E}_\pi[\theta] \), i.e., posterior mean.
Consistency of Posterior Mean

- **Consistency**: Suppose data are generated from the model \( y_t = x'_t \theta_0 + u_t \). Asymptotically the Bayes estimator converges to the “true” parameter \( \theta_0 \).

- Consider

\[
\bar{\theta}_T = (X'X + \tau^{-2}I)^{-1}X'Y
\]

\[
= \theta_0 + \left[ \left( \frac{1}{T} \sum x_t x'_t + \frac{1}{\tau^2} I \right)^{-1} - \left( \frac{1}{T} \sum x_t x'_t \right)^{-1} \right]
\]

\[
\times \left( \frac{1}{T} \sum x_t x'_t \right) \theta_0
\]

\[
+ \left( \frac{1}{T} \sum x_t x'_t + \frac{1}{\tau^2} I \right)^{-1} \left( \frac{1}{T} \sum x_t u_t \right)
\]

\[\xrightarrow{p} \theta_0\]

- Disagreement between two Bayesians who have different priors will asymptotically vanish.
• $H_0 : \theta \in \Theta_0$ versus $H_1 : \theta \in \Theta_1$.
• Decision space is 0 ("reject") and 1 ("accept").
• Loss function

$$L(\theta, \delta) = \begin{cases} 0 & \delta = \mathbb{I}\{\theta \in \Theta_0\} \quad \text{correct decision} \\ a_0 & \delta = 0, \ \theta \in \Theta_0 \quad \text{Type 1 error} \\ a_1 & \delta = 1, \ \theta \in \Theta_1 \quad \text{Type 2 error} \end{cases}$$

Note that the parameters $a_1$ and $a_2$ are part of the econometrician’s preferences.
• Optimal decision:

$$\delta(Y) = \begin{cases} 1 & \mathbb{P}_\pi\{\theta \in \Theta_0\} \geq \frac{a_1}{a_0 + a_1} \\ 0 & \text{otherwise} \end{cases}$$
• Posterior odds:

\[
\frac{P_\pi \{ \theta \in \Theta_0 \}}{P_\pi \{ \theta \in \Theta_1 \}}
\]

• Often, hypotheses are evaluated according to Bayes factors:

\[
B(Y) = \frac{\text{Posterior Odds}}{\text{Prior Odds}}
\]
• Set estimation is more difficult to cast into a decision problem \( \mapsto \) two-player game.

• **Bayesian credible set:** \( C_Y \subseteq \Theta \) is \( 1 - \alpha \) credible if

\[
\mathbb{P}^\theta_{Y}\{ \theta \in C_Y \} \geq 1 - \alpha
\]

\( \text{r.v.} \)

• A highest posterior density region (HPD) is of the form

\[
C_Y = \{ \theta : p(\theta|Y) \geq k_\alpha \} \quad \text{where } k_\alpha \text{ is chosen s.t. } \mathbb{P}^\theta_Y\{ \theta \in C_Y \} = 1 - \alpha.
\]

HPD regions have the smallest volume among all \( 1 - \alpha \) credible regions.

• HPD regions are often difficult to compute. Thus, Bayesians often report equal-tail probability credible intervals.

• **Recall definition of frequentist confidence set:**

\[
\mathbb{P}^Y_\theta\{ \theta \in C_Y \} \geq 1 - \alpha \quad \text{for all } \theta \in \Theta.
\]

\( \text{r.v.} \)
• **Example:**

\[
y_{T+h} = \theta^h y_T + \sum_{s=0}^{h-1} \theta^s u_{T+h-s}
\]

• *h*-step ahead conditional distribution:

\[
y_{T+h}\mid (Y_{1:T}, \theta) \sim N \left( \theta^h y_T, \frac{1 - \theta^h}{1 - \theta} \right).
\]

• Posterior predictive distribution:

\[
p(y_{T+h}\mid Y_{1:T}) = \int p(y_{T+h}\mid y_T, \theta)p(\theta\mid Y_{1:T})d\theta.
\]

• For each draw \(\theta^i\) from the posterior distribution \(p(\theta\mid Y_{1:T})\) sample a sequence of innovations \(u^i_{T+1}, \ldots, u^i_{T+h}\) and compute \(y^i_{T+h}\) as a function of \(\theta^i, u^i_{T+1}, \ldots, u^i_{T+h}\), and \(Y_{1:T}\).
Model Uncertainty

- Assign prior probabilities $\gamma_{j,0}$ to models $M_j$, $j = 1, \ldots, J$.
- Posterior model probabilities are given by
  \[
  \gamma_{j,T} = \frac{\gamma_{j,0} p(Y|M_j)}{\sum_{j=1}^{J} \gamma_{j,0} p(Y|M_j)},
  \]
  where
  \[
  p(Y|M_j) = \int p(Y|\theta_{(j)}, M_j) p(\theta_{(j)}|M_j) d\theta_{(j)}
  \]
- Log marginal data densities are one-step-ahead predictive scores:
  \[
  \ln p(Y|M_j) = \sum_{t=1}^{T} \ln \int p(y_t|\theta_{(j)}, Y_{1:t-1}, M_j) p(\theta_{(j)}|Y_{1:t-1}, M_j) d\theta_{(j)}.
  \]
- Model averaging:
  \[
  p(h|Y) = \sum_{j=1}^{J} \gamma_{j,T} p(h_j(\theta_{(j)})|Y, M_j).
  \]