• **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 \( \theta \) \( \Rightarrow \) determine properties of model-simulated data \( Y^{sim}(\theta) \).

• **Statistical inference:** observed data \( Y^{obs} \) \( \Rightarrow \) 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

\[ \tilde{\pi}_t = \gamma_b \tilde{\pi}_{t-1} + \gamma_f \mathbb{E}_t[\tilde{\pi}_{t+1}] + \kappa \tilde{MC}_t \]
Model Misspecification is a Concern

Prior for misspecification parameters $\Phi^\Delta$: Shape of contours determined by Kullback-Leibler distance.

$\Phi^*(\theta)$: Cross-equation restriction for given value of $\theta$

subspace generated by the DSGE model restrictions
• We want to determine the effect of a policy change.

• Policy effect depends on model parameters.

• Can we learn the model parameters from the observed data?

• Thought experiment: suppose model is “true” and we observe an infinite amount of data from the model. What can we learn?
• 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 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 $Y_1 = Y_2$. 
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.

- Desirable to let the model-implied probability distribution \(p(Y|\theta_0, M_1)\) determine the choice of the objective function for estimators and test statistics to obtain a statistical procedure that is efficient (meaning that the estimator is close to \(\theta_0\) with high probability in repeated sampling).

- Maximum likelihood (ML) estimator

\[
\hat{\theta}_{ml} = \arg\max_{\theta \in \Theta} \log p(Y|\theta, M_1).
\]

- Minimize discrepancy between sample statistics \(\hat{m}_T(Y)\) and model-implied population statistics \(E[\hat{m}_T(Y)|\theta, M_1]\):

\[
\hat{\theta}_{md} = \arg\min_{\theta \in \Theta} Q_T(\theta|Y) = \|\hat{m}_T(Y) - E[\hat{m}_T(Y)|\theta, M_1]\|_{W_T}.
\]
Model of interest ($M_1$) is assumed to be misspecified or incompletely specified.

- Example: suppose a DSGE model only has a monetary policy shock. Then,

$$\frac{1}{\kappa_p(1 + \nu)x_{\epsilon_R}/\beta + \sigma_R} \hat{R}_t - \frac{1}{\kappa_p(1 + \nu)x_{\epsilon_R}} \hat{\pi}_t = 0,$$

which is clearly violated in the data.

- Need reference model $M_0$, e.g., VAR, under which to evaluate sampling distribution of $Y$.

- Concept of “true” value is no longer sensible $\Rightarrow$ pseudo-optimal parameter value:

$$\theta_0(Q, W) = \arg\min_{\theta \in \Theta} Q(\theta|M_0),$$

where

$$Q(\theta|M_0) = \left\| \mathbb{E}[\hat{m}_T(Y)|M_0] - \mathbb{E}[\hat{m}(Y)|\theta, M_1] \right\|_W.$$
Bayesian Inference

**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:
  \[
  \delta_* = \arg\min_{\delta \in \mathcal{D}} \int L(h(\theta), \delta)p(\theta|Y, M_1)d\theta.
  \]
- Place probabilities on competing models and update:
  \[
  \frac{\pi_{1,T}}{\pi_{2,T}} = \frac{\pi_{1,0}}{\pi_{2,0}} \frac{p(Y|M_1)}{p(Y|M_2)}.
  \]
Model of interest ($M_1$) is assumed to be misspecified or incompletely specified.

- Derive posterior distributions under a more flexible reference model $M_0$, e.g., VAR. Then choose $\theta$ to minimize discrepancy between implications of $M_0$ and DSGE model $M_1$.

- Use DSGE model $M_1$ to generate a prior distribution for a more flexible reference model $M_0$. (see next slide)

- Rather than using posterior probabilities to select among or average across two DSGE models, one can form a prediction pool, which is essentially a linear combination of two predictive densities:

$$
\lambda p(y_t | Y_{1:t-1}, M_1) + (1 - \lambda) p(y_t | Y_{1:t-1}, M_2).
$$

The weight $\lambda \in [0, 1]$ can be determined based on

$$
\prod_{t=1}^{T} \left[ \lambda p(y_t | Y_{1:t-1}, M_1) + (1 - \lambda) p(y_t | Y_{1:t-1}, M_2) \right].
$$
Using a DSGE Model as Prior for a VAR

\[ \Phi^*(\theta) : \text{Cross-equation restriction for given value of } \theta \]

Prior for misspecification parameters \( \Phi^\Delta \): Shape of contours determined by Kullback-Leibler distance.

Subspace generated by the DSGE model restrictions
Using a DSGE Model as Prior for a VAR - Weight on Model Restrictions

Frank Schorfheide
Introduction to (Bayesian) Inference
Using a DSGE Model as Prior for a VAR - Weight on Model Restrictions

Frank Schorfheide
Introduction to (Bayesian) Inference
• Macroeconomists/econometricians have been criticized for relying on models that abstract from financial intermediation / frictions.

• With hindsight it turned out that financial frictions were important to understand the Great Recession. But are they also important in normal times?

• We need tools that tell us in real-time when to switch models...

• Linear prediction pool:

  \[
  \text{Density Forecast}_t = \lambda_t \cdot \text{Forecast from “Normal” Model}_t \\
  + (1 - \lambda_t) \cdot \text{Forecast from “Fin Frictions” Model}_t
  \]

• Determine weight \( \lambda_t \) in real time based on historical forecast performance.
Pooling “New” and “Old” Models

Relative forecasting performance changes over time

“Old” Smets-Wouters Model vs. “New” DSGE with Financial Frictions

It’s easy to see with hindsight which model we should have used.
Pooling “New” and “Old” Models

Time-Varying Weight $\lambda_t$ (Posterior Distribution) on “New” DSGE with Financial Frictions

It’s more difficult to determine the best model in real time...
Pooling “New” and “Old” Models


Techniques for determining the best model in real time are available.
Bayesian Inference

- **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
  $\theta^i \sim p(\theta|Y), \; i = 1, \ldots, N$

- **Algorithms:** direct sampling, accept/reject sampling, importance sampling, Markov chain Monte Carlo sampling, sequential Monte Carlo sampling...
• Consider AR(1) model:

\[ y_t = y_{t-1}\phi + u_t, \quad u_t \sim iidN(0, 1). \]

• Let \( x_t = y_{t-1} \). Write as

\[ y_t = x_t'\phi + u_t, \quad u_t \sim iidN(0, 1), \]

or

\[ Y = X\phi + U. \]

We can easily allow for multiple regressors. Assume \( \phi \) 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:

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

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

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

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

- **Guess:** what if \(\phi|Y \sim N(\bar{\phi}_T, \bar{V}_T)\). Then
  \[
p(\theta|Y) \propto \exp \left\{ -\frac{1}{2} (\phi - \bar{\phi}_T)'\bar{V}_T^{-1}(\phi - \bar{\phi}_T) \right\}.
  \]

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

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

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

with

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

- $\tau \rightarrow \infty$:

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

- $\tau \rightarrow 0$:

$$
\phi | Y \overset{approx}{\sim} \text{Pointmass at } 0
$$
Marginal Data Density

- 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 $p(\phi|Y)$ by $\pi(\phi)$ and posterior expectations of $h(\phi)$ by

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

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

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

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

Strong law of large numbers (SLLN), central limit theorem (CLT)…
In the simple linear regression model with Gaussian posterior it is possible to sample directly.

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

Provided that $\nabla_{\pi} [h(\phi)] < \infty$ we can deduce from Kolmogorov’s SLLN and the Lindeberg-Levy CLT that

$$
\overline{h}_N \xrightarrow{a.s.} E_{\pi}[h]
$$

$$
\sqrt{N} \left( \overline{h}_N - E_{\pi}[h] \right) \implies N(0, \nabla_{\pi} [h(\phi)]).
$$
Decision Making

- 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:
  \[ \delta^*(Y) = \arg\min_d \rho(\delta(\cdot)|Y). \]

- 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
  \[ \delta^*_N(Y) = \arg\min_d \bar{\rho}_N(\delta(\cdot)|Y). \]
• 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
Point Estimation

- 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:** 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
\[
\bar{\theta}_T = (X'X + \tau^{-2}I)^{-1}X'Y \\
= \theta_0 + \left[ \left( \frac{1}{T} \sum x_t x_t' + \frac{1}{\tau^2 T} I \right)^{-1} - \left( \frac{1}{T} \sum x_t x_t' \right)^{-1} \right] \\
\times \left( \frac{1}{T} \sum x_t x_t' \right) \theta_0 \\
+ \left( \frac{1}{T} \sum x_t x_t' + \frac{1}{\tau^2 T} I \right)^{-1} \left( \frac{1}{T} \sum x_t u_t \right) \\
\xrightarrow{p} \theta_0
\]

- 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

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

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:

\[
\frac{\mathbb{P}_\pi \{ \theta \in \Theta_0 \}}{\mathbb{P}_\pi \{ \theta \in \Theta_1 \}}
\]

• Often, hypotheses are evaluated according to Bayes factors:

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

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

$$\mathbb{P}_Y^{\theta} \left\{ \theta \in C_Y \right\} \geq 1 - \alpha$$

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

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

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:**

$$\mathbb{P}_\theta^{Y} \left\{ \theta \in C_Y \right\} \geq 1 - \alpha \text{ for all } \theta \in \Theta.$$
• Example:

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

• \( h \)-step ahead conditional distribution:

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

• Posterior predictive distribution:

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

• For each draw \( \theta^i \) from the posterior distribution \( p(\theta | Y_{1:T}) \) 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:
  \[ p(h|Y) = \sum_{j=1}^{J} \gamma_{j,T} p(h_j(\theta_{(j)})||Y, M_j). \]