# Likelihood Inference

Jesús Fernández-Villaverde
University of Pennsylvania

## Likelihood Inference

- We are going to focus in likelihood-based inference.

- Why?

  1. Likelihood principle (Berger and Wolpert, 1988).

  2. Attractive asymptotic properties and good small sample behavior (White, 1994 and Fernández-Villaverde and Rubio-Ramírez, 2004).

  3. Estimates parameters needed for policy and welfare analysis.

  4. Simple way to deal with misspecified models (Monfort, 1996).

  5. Allow us to perform model comparison.

Alternatives

- Empirical likelihood, non- and semi-parametric methods.
- Advantages and disadvantages.
- Basic theme in econometrics: robustness versus efficiency.
- One size does not fit all!

## The Likelihood Function (Fisher, 1921)

- We have observations $x_1, x_2, ..., x_T$.
- We have a model that specifies that the observations are realization of a random variable $X$.
- We deal with situations in which $X$ has a parametric density $f_\theta$ for all values of $\theta \in \Theta$.
- The likelihood function is defined as $l_x(\theta) = f_\theta(x)$, the density of $X$ evaluated at $x$ as a function of $\theta$.

Some Definitions

Definition of Sufficient Statistic: When $x \sim f_\theta(x)$, a function $T$ of $x$ (also called a statistic) is said to be sufficient if the distribution of $x$ conditional upon $T(x)$ does not depend on $\theta$.

Remark: Under the factorization theorem, under measure theoretic regularity conditions:

$$f_\theta(x) = g(T(x)|\,\theta)\, h(x|\,T(x))$$

i.e., a sufficient statistic contains the whole information brought by $x$ about $\theta$.

Definition of Ancillary Statistic: When $x \sim f_\theta(x)$, a statistic $S$ of $x$ is said to be ancillary if the distribution of $S(x)$ does not depend on $\theta$.

Experiments and Evidence

Definition of Experiment: An experiment $E$ is a triple $(X, \theta, \{f_\theta\})$, where the random variable $X$, taking values in $\Omega$ and having density $f_\theta$ for some $\theta \in \Theta$, is observed.

Definition of Evidence from an Experiment $E$ : The outcome of an experiment $E$ is the data $X = x$. From $E$ and $x$ we can infer something about $\theta$. We define all possible evidence as $Ev\,(E, x)$.

The Likelihood Principle

The Likelihood Principle: Let two experiments $E_1 = \left(X_1, \theta, \left\{f_\theta^1\right\}\right)$ and $E_2 = \left(X_2, \theta, \left\{f_\theta^2\right\}\right)$, suppose that for some realizations $x_1^*$ and $x_2^*$, it is the case that $f_\theta^1\left(x_1^*\right) = c f_\theta^2\left(x_2^*\right)$ ,then $Ev\left(E_1, x_1^*\right) = Ev\left(E_2, x_2^*\right)$

Intrepretation: All the information about $\theta$ that we can obtain from an experiment is contained in likelihood function for $\theta$ given the data.

How do we derive the likelihood principle?

Sufficiency Principle: Let experiment $E = (X, \theta, \{f_\theta\})$ and suppose $T(X)$ sufficient statistic for $\theta$, then, if $T(x_1) = T(x_2)$, $Ev(E, x_1) = Ev(E, x_2)$.

Conditionality Principle: Let two experiments $E_1 = \left(X_1, \theta, \left\{f_\theta^1\right\}\right)$ and $E_2 = \left(X_2, \theta, \left\{f_\theta^2\right\}\right)$. Consider the mixed experiment $E^* = \left(X^*, \theta, \left\{f_\theta^*\right\}\right)$ where $X^* = (J, X_J)$ and $f_\theta^*\left(\left(j, x_j\right)\right) = \frac{1}{2} f_\theta^j\left(x_j\right)$.

Then $Ev\left(E^*, \left(j, x_j\right)\right) = Ev\left(E_j, x_j\right)$.

Basic Equivalence Result

Theorem: The Conditionality and Sufficiency Principles are necessary and sufficient for the Likelihood Principle (Birnbaum, 1962).

Remark A slightly stronger version of the Conditionality Principle implies, by itself, the Likelihood Principle (Evans, Fraser, and Monette, 1986).

Proof: First, let us show that the Conditionality and the Sufficiency Principles $\Rightarrow$ Likelihood Principle.

Let $E_1$ and $E_2$ be two experiments. Assume that $f_\theta^1\left(x_1^*\right) = cf_\theta^2\left(x_2^*\right)$.

The Conditionality Principle $\Rightarrow Ev\left(E^*, \left(j, x_j\right)\right) = Ev\left(E_j, x_j\right)$.

Consider the statistic:

$$T(J, X_J) = \begin{cases} \left(1, x_1^*\right) \text{ if } J = 2, X_2 = x_2^* \\ (J, X_J) \text{ otherwise} \end{cases}$$

$T$ is a sufficient statistic for $\theta$ since:

$$P_\theta\left((J, X_J) = (j, x_j) | T = t \neq (1, x_1^*)\right) = \begin{cases} 1 \text{ if } (j, x_j) = t \\ 0 \text{ otherwise} \end{cases}$$

Now:

$$P_\theta\left((J, X_J) = (1, x_1^*) \,|T = (1, x_1^*)\right) =$$
$$= \frac{P_\theta\left((J, X_J) = \left(1, x_1^*\right), T = \left(1, x_1^*\right)\right)}{P_\theta\left(T = \left(1, x_1^*\right)\right)} =$$
$$= \frac{\frac{1}{2} f_\theta^1\left(x_1^*\right)}{\frac{1}{2} f_\theta^1\left(x_1^*\right) + \frac{1}{2} f_\theta^2\left(x_2^*\right)} = \frac{c}{1+c}$$

and

$P_\theta\left((J, X_J) = (1, x_1^*) \,|T = (1, x_1^*)\right) = 1 - P_\theta\left((J, X_J) = (2, x_2^*) \,|T = (1, x_1^*)\right)$

Since $T\left(1, x_1^*\right) = T\left(2, x_2^*\right)$, the Sufficiency Principle $\Rightarrow Ev\left(E^*, \left(1, x_1^*\right)\right) = Ev\left(E^*, \left(2, x_2^*\right)\right) \Rightarrow$ the Likelihood Principle.

Now, let us prove that the Likelihood Principle $\Rightarrow$ both the Conditionality and the Sufficiency Principles.

The likelihood function in $E^*$ is

$$l_{(j,x_j)}(\theta) = \frac{1}{2} f_\theta^j\left(x_j\right) \propto l_{x_j}(\theta) = f_\theta^j\left(x_j\right)$$

proportional to the likelihood function in $E_j$ when $x_j$ is observed.

The Likelihood Principle $\Rightarrow Ev\left(E^*, \left(j, x_j\right)\right) = Ev\left(E_j, x_j\right) \Rightarrow$ Conditionality Principle.

If $T$ is sufficient and $T(x_1) = T(x_2) \Rightarrow f_\theta(x_1) = df_\theta(x_2)$. The Likelihood Principle $\Rightarrow Ev(E, x_1) = Ev(E, x_2) \Rightarrow$ Sufficiency Principle.

Stopping Rule Principle

If a sequence of experiments, $E_1$, $E_2$, ..., is directed by a stopping rule, $\tau$, which indicates when the experiment should stop, inference about $\theta$, should depend on $\tau$ only through the resulting sample.

- Interpretation.
- Difference with classical inference.
- Which one makes more sense?

Example by Lindley and Phillips (1976)

- We are given a coin and we are interested in the probability of heads $\theta$ when flipped.

- We test $H_0 : \theta = \frac{1}{2}$ versus $H_1 : \theta > \frac{1}{2}$.

- An experiment involves flipping a coin 12 times, with the result of 9 heads and 3 tails.

- What was the reasoning behind the experiment, i.e., which was the stopping rule?

Two Possible Stopping Rules

1. The experiment was to toss a coin 12 times$\Rightarrow \mathcal{B}(12,\theta)$. Likelihood:

$$f_\theta^1(x) = \left( \begin{array}{c} n \\ x \end{array} \right) \theta^x (1-\theta)^{n-x} = 220\theta^9 (1-\theta)^3$$

2. The experiment was to toss a coin until 3 tails were observed$\Rightarrow$ $\mathcal{NB}(3,\theta)$. Likelihood:

$$f_\theta^2(x) = \left( \begin{array}{c} n+x-1 \\ x \end{array} \right) \theta^x (1-\theta)^{n-x} = 55\theta^9 (1-\theta)^3$$

- Note $f_\theta^1(x) = cf_\theta^2(x)$, consequently a LP econometrician gets the same answer in both cases.

Classical Analyses

Fix a conventional significance level of 5 percent.

1. Observed significance level of $x = 2$ against $\theta = \frac{1}{2}$ would be:

   $$\alpha_1 = P_{1/2}(X \geq 9) = f^1_{1/2}(9) + f^1_{1/2}(10) + f^1_{1/2}(11) + f^1_{1/2}(12) = 0.075$$

2. Observed significance level of $x = 2$ against $\theta = \frac{1}{2}$ would be:

   $$\alpha_2 = P_{1/2}(X \geq 9) = f^2_{1/2}(9) + f^2_{1/2}(10) + f^2_{1/2}(11) + f^2_{1/2}(12) = 0.0325$$

We get different answers: no reject $H_0$ in 1, reject $H_0$ in 2!

What is Going On?

- The LP tells us that all the experimental information is in the evidence.
- A non-LP researcher is using, in its evaluation of the evidence, observations that have *NOT* occurred.
- Jeffreys (1961): "...a hypothesis which may be true may be rejected because it has not predicted observable results which have not occurred."
- In our example $\theta = \frac{1}{2}$ certainly is not predicting $X$ larger than 9, and in fact, such values do not occur.

Savage (1962)

"I learned the stopping rule principle from Professor Barnard, in conversation in the summer of 1952. Frankly, I the thought it a scandal that anyone in the profession could advance an idea so patently wrong, even as today I can scarcely believe that some people resist an idea so patently right".

Limitations of the Likelihood Principle

- We have one important assumption: $\theta$ is finite-dimensional.
- What if $\theta$ is infinite-dimensional?
- Why infinite-dimensional problems are relevant?
  1. Economic theory advances: Ellsberg's Paradox.
  2. Statistical theory advances.
  3. Numerical advances.

## Infinite-Dimensional Problems

- Many of our intuitions from finite dimensional spaces break down when we deal with spaces of infinite dimensions.

- Example by Robins and Ritov (1997).

- Example appears in the analysis of treatment effects in randomized trials.

Model

- Let $(x_1, y_1), ..., (x_n, y_n)$ be $n$ i.i.d. copies of a random vector $(X, Y)$ where $X$ takes values on the unit cube $(0,1)^k$ and $Y$ is normally distributed with mean $\theta(x)$ and variance 1.

- The density $f(x)$ belongs to the class

$$\mathcal{F} = \left\{ f : c < f(x) \leq 1/c \text{ for } x \in (0,1)^k \right\}$$

where $c \in (0,1)$ is a fixed constant.

- The conditional mean function is continuous and $\sup_{x \in (0,1)^k} |\theta(x)| \leq M$ for some positive finite constant $M$. Let $\Theta$ be the set of all those functions.

Likelihood

- The likelihood function of this model is:

$$\mathcal{L}(f,\theta)=\left\{\prod_{i=1}^{n}\phi\left(y_i-\theta\left(x_i\right)\right)\right\}\left\{\prod_{i=1}^{n}f\left(x_i\right)\right\}$$

  where $\phi(\cdot)$ is the standard normal density.

- Note that the model is infinite-dimensional because the set $\Theta$ cannot be put into smooth, one-to-one correspondence with a finite-dimensional Euclidean space.

- Our goal is to estimate:

$$\psi=\int_{(0,1)^k}\theta(x)\,dx$$

## Ancillary Statistic is Not Irrelevant

- Let $X^*$ be the set of observed $x$'s.

- When $f$ is known, $X^*$ is ancillary. Why?

- When $f$ is unknown, $X^*$ is ancillary for $\psi$. Why? Because the conditional likelihood given $X^*$ is a function of $f$ alone, $\theta$ and $f$ are variation independent (i.e., the parameter space is a product space), and $\psi$ only depends on $\theta$.

Consistent Estimators

- When $f$ is unknown, there no uniformly consistent estimator of $\psi$ (Robins and Ritov, 1997).
- When $f$ is known, there are $n^{0.5}$-consistent uniformly estimator of $\psi$ over $f \times \theta \in \mathcal{F} \times \Theta$.
- Example: $\psi^* = \frac{1}{n} \sum_{i=1}^{n} \frac{y_i}{f(x_i)}$.
- But we are now using $X^*$, which was supposed to be ancillary!
- You can show that no estimator that respects the likelihood principle can be uniformly consistent over $f \times \theta \in \mathcal{F} \times \Theta$.

Likelihood Based Inference

- Likelihood Principle strongly suggests implementing likelihood-based inference.

- Two basic approaches:

  1. Maximum likelihood:

$$\widehat{\theta} = \arg\max_{\theta} l_x(\theta)$$

  2. Bayesian estimation:

$$\pi\left(\theta|X^T\right) = \frac{l_x(\theta)\,\pi(\theta)}{\int l_x(\theta)\,\pi(\theta)\,d\theta}$$

Maximum Likelihood Based Inference

- Maximum likelihood is well-know and intuitive.
- One of the main tools of classical econometrics.
- Asymptotic properties: consistency, efficiency, and normality.

Why Would Not You Use ML?

1. Maximization is a difficult task.

2. Lack of smoothness (for example if we have boundaries)

3. Stability problems.

4. It often violates the likelihood principle.

Classical Econometrics and the Likelihood Principle

- Consider the following example due to Berger and Wolpert (1984).

- Let $\Omega = \{1, 2, 3\}$ and $\Theta = \{0, 1\}$ and consider the following two experiments $E_1$ and $E_2$ with the following densities:

| | 1 | 2 | 3 |
|---|---|---|---|
| $f_0^1(x_1)$ | .9 | .05 | .05 |
| $f_1^1(x_1)$ | .09 | .055 | .855 |

and

| | 1 | 2 | 3 |
|---|---|---|---|
| $f_0^2(x_2)$ | .26 | .73 | .01 |
| $f_1^2(x_2)$ | .026 | .803 | .171 |

- Note: same underlaying phenomenon. Examples in economics? Euler Equation test.

Constant Likelihood Ratios

- Let $x_1 = 1$ and $x_2 = 1$.
- But $\left(f_0^1(1), f_1^1(1)\right) = (.9, .09)$ and $\left(f_0^2(1), f_1^2(1)\right) = (.26, .026)$ are proportional $\Rightarrow$LP$\Rightarrow$ same inference.
- Actually, this is true for any value of $x$; the likelihood ratios are always the same.
- If we get $x_1 = x_2$, the LP tells us that we should get the same inference.

## A Standard Classical Test

- Let the following classical test. $H_0$: $\theta = 0$ and we have the following test:

$$\begin{cases} accept \text{ if } x = 1 \\ reject \text{ otherwise} \end{cases}$$

- This test has the most power under $E_1$.

- But errors are different: Type $I$ error is 0.1 ($E_1$) against 0.74 ($E_2$) and Type $II$ error is 0.09 ($E_1$) against 0.026 ($E_2$).

- This implies that $E_1$ and $E_2$ will give very different answers.

What is Going On?

- Experiment $E_1$ is much more likely to proved useful information about $\theta$, as evidenced by the overall better error probabilities (a measure of ex ante precision).

- Once $x$ is at hand, ex ante precision is irrelevant.

- What matters is ex post information!

Is There an Alternative that Respects the Likelihood Principle?

- Yes: Bayesian econometrics.
- Original idea of Reverend Thomas Bayes in 1761.
- First modern treatment: Jeffreys (1939).
- During the next half century, landscape dominated by classical methods (despite contribution like Savage, 1954, and Zellner, 1971).
- Resurgence in the 1990s because of the arrival of McMc.

Basic Difference: Conditioning

- Classical and Bayesian methods differ basically on what do you condition on.

- Classical (or frequentist) search for procedures that work well ex ante.

- Bayesians always condition ex post.

- Example: Hypothesis testing.

Why Bayesian?

- It respects the likelihood principle.
- It can be easily derived from axiomatic foundations (Heath and Sudderth, 1996) as an if and only if statement.
- Coherent and comprehensive.
- Easily deals with misspecified models.
- Good small sample behavior.
- Good asymptotic properties.

Bayesian Econometrics: the Basic Ingredients

- Data $y^T \equiv \{y_t\}_{t=1}^T \in R^T$
- Model $i \in M$ :
  - Parameters set
$$\Theta_i \in R^{k_i}$$
  - Likelihood function
$$f(y^T|\theta, i) : R^T \times \Theta_i \rightarrow R^+$$
  - Prior Distribution
$$\pi(\theta|i) : \Theta_i \rightarrow R^+$$

## Bayesian Econometrics Basic Ingredients II

- The Joint Distribution for model $i \in M$

$$f(y^T|\theta, i)\pi(\theta|i)$$

- The Marginal Distribution

$$P\left(y^T|i\right) = \int f(y^T|\theta, i)\pi(\theta|i)\, d\theta$$

- The Posterior Distribution

$$\pi\left(\theta|y^T, i\right) = \frac{f(y^T|\theta, i)\pi(\theta|i)}{\int f(y^T|\theta, i)\pi(\theta|i)\, d\theta}$$

Bayesian Econometrics and the Likelihood Principle

Since all Bayesian inference about $\theta$ is based on the posterior distribution

$$\pi\left(\theta|Y^T, i\right) = \frac{f(Y^T|\theta, i)\pi\left(\theta|i\right)}{\int_{\Theta_i} f(Y^T|\theta, i)\pi\left(\theta|i\right) d\theta}$$

the Likelihood Principle always holds.

A Baby Example (Zellner, 1971)

- Assume that we have $n$ observations $y^T = (y_1, ..., y_n)$ from $\mathcal{N}(\theta, 1)$.

- Then:

$$\begin{aligned} f(y^T|\theta) &= \frac{1}{(2\pi)^{0.5n}} \exp\left[-\frac{1}{2}\sum_{i=1}^{n}(y_i-\theta)^2\right] \\ &= \frac{1}{(2\pi)^{0.5n}} \exp\left[-\frac{1}{2}\left(ns^2 + n\left(\theta-\theta'\right)^2\right)\right] \end{aligned}$$

  where $\theta' = \frac{1}{n}\sum y_i$ is the sample mean and $s^2 = \frac{1}{n}\sum\left(y_i-\theta'\right)^2$ the sample variance.

The Prior

- Prior distribution:

$$\pi(\theta) = \frac{1}{(2\pi)^{0.5}\,\sigma} \exp\left[-\frac{(\theta-\mu)^2}{2\sigma^2}\right]$$

  The parameters $\sigma$ and $\mu$ are sometimes called *hyperparameters*.

- We will talk in a moment about priors and where they might come from.

## The Posterior

$$
\begin{aligned}
\pi\left(\theta|y^T,i\right) &\propto \frac{1}{(2\pi)^{0.5n}}\exp\left[-\tfrac{1}{2}\left(ns^2+n\left(\theta-\theta'\right)^2\right)\right]\frac{1}{(2\pi)^{0.5}\,\sigma}\exp\left[-\frac{(\theta-\mu)^2}{2\sigma^2}\right] \\
&\propto \exp\left[-\frac{1}{2}\left(n\left(\theta-\theta'\right)^2+\frac{(\theta-\mu)^2}{\sigma^2}\right)\right] \\
&\propto \exp\left[-\frac{1}{2}\frac{\sigma^2+1/n}{\sigma^2/n}\left(\theta-\frac{\theta'\sigma^2+\mu/n}{\sigma^2+1/n}\right)^2\right]
\end{aligned}
$$

Remarks

- Posterior is a normal that with mean $\frac{\theta'\sigma^2+\mu/n}{\sigma^2+1/n}$ and variance $\frac{\sigma^2/n}{\sigma^2+1/n}$.
- Note the weighted sum structure of the mean and variance.
- Note the sufficient statistics structure.
- Let's see a plot: `babyexample.m`.

## An Asymptotic Argument

- Notice, that as $n \longrightarrow \infty$ :

$$\begin{aligned} \frac{\theta' \sigma^2 + \mu/n}{\sigma^2 + 1/n} &\rightarrow \theta' \\ \frac{\sigma^2/n}{\sigma^2 + 1/n} &\rightarrow 0 \end{aligned}$$

- We know, by a simple law of large numbers that, $\theta' \longrightarrow \theta_0$, i.e. the true parameter value (if the model is well specified) or to the pseudo-true parameter value (if not).

- We will revisit this issue.

Applications to Economics

- Previous example is interesting, but purely statistical.
- How do we apply this approach in economics?
- Linear regression and other models (VARs) are nothing more than small modifications of previous example.
- Dynamic Equilibrium models required a bit more work.
- Let me present a trailer of attractions to come.

## A Mickey Mouse Economic Example

- Assume we want to explain data on consumption:

$$C^T \equiv \{C_t\}_{t=1}^T$$

- Model

$$\max \sum_{t=1}^{T} \log C_t$$

s.t.

$$C_t \leq \omega_t$$

where $\omega_t \sim iid\ N(\mu, \sigma^2)$ and $\theta \equiv (\mu, \sigma) \in \Theta \equiv [0, \infty) \times [0, \infty)$.

- Model solution implies⇒Likelihood function

$$\{C_t\}_{t=1}^{T} \sim iid\ N(\mu, \sigma^2)$$

so

$$f(\{C_t\}_{t=1}^{T} | \theta) = \Pi_{t=1}^{T} \phi\left(\frac{C_t - \mu}{\sigma}\right)$$

- Priors

$$\mu \sim \text{Gamma}(4, 0.25)$$

$$\sigma \sim \text{Gamma}(1, 0.25)$$

so:

$$\pi(\theta) = G(\mu; 4, 0.25)\, G(\sigma; 1, 0.25)$$

Bayes Theorem

Posterior distribution

$$\begin{aligned}\pi\left(\theta|\left\{C_t\right\}_{t=1}^{T}\right) &= \frac{f(\left\{C_t\right\}_{t=1}^{T}|\theta)\pi\left(\theta|\right)}{\int_{\Theta} f(\left\{C_t\right\}_{t=1}^{T}|\theta)\pi\left(\theta\right)d\theta} \\ &= \frac{\Pi_{t=1}^{T}\phi\left(\frac{C_t-\mu}{\sigma}\right)G\left(\mu;4,0.25\right)G\left(\sigma;1,0.25\right)}{\int_{\Theta}\Pi_{t=1}^{T}\phi\left(\frac{C_t-\mu}{\sigma}\right)G\left(\mu;4,0.25\right)G\left(\sigma;1,0.25\right)d\theta}\end{aligned}$$

and

$$\int_{\Theta}\Pi_{t=1}^{T}\phi\left(\frac{C_t-\mu}{\sigma}\right)G\left(\mu;4,0.25\right)G\left(\sigma;1,0.25\right)d\theta$$

is the marginal likelihood.

Remarks

- Posterior distribution does not belong to any easily recognized parametric family:

  1. Traditional approach: conjugate priors$\Rightarrow$prior such that posterior belongs to the same parametric family.

  2. Modern approach: simulation.

- We need to solve a complicated integral:

  1. Traditional approach: analytic approximations.

  2. Modern approach: simulation.

Tasks in Front of Us

1. Talk about priors.

2. Explain the importance of posteriors and marginal likelihoods.

3. Practical implementation.

## Tasks in Front of Us

1. Talk about priors.

What is the Prior?

- The prior is the belief of the researcher about the likely values of the parameters.

- Gathers prior information.

- Problems:

  1. Can we always formulate a prior?

  2. If so, how?

  3. How do we measure the extent to which the prior determines our results?

## Proper versus Improper Priors

- What is a proper prior? A prior that is a well-defined pdf.

- Who would like to use an improper prior?

  1. To introduce classical inference through the back door.

  2. To achieve "non-informativeness" of the prior: why? Uniform distribution over $\Re$.

- Quest for "noninformative" prior.

## Some Noninformative Priors I: Laplace's Prior

- Principle of Insufficient Reason: Uniform distribution over $\Theta$.

- Problems:

  1. Often induces nonproper priors.

  2. Non invariant under reparametrizations. If we switch from $\theta \in \Theta$ with prior $\pi(\theta) = 1$ to $\eta = g(\theta)$, the corresponding new prior is:

$$\pi^*(\eta) = \left| \frac{d}{d\eta} g^{-1}(\eta) \right|$$

     Therefore $\pi^*(\eta)$ is usually not a constant.

Example of Noninvariance

- Discussion: is the business cycle asymmetric?
- Let $p$ be the proportion of quarters in which there GDP per capita grows less than the long-run average for the U.S. economy (1.9%).
- To learn about $p$ we select a prior $\mathcal{U}[0,1]$.
- Now, the odds ratio is $\kappa = \frac{p}{1-p}$.
- But the uniform prior on $p$ implies a prior on $\kappa$ with density $\frac{1}{(1+\kappa)^2}$.

## Some Noninformative Priors II: Unidimensional Jeffreys Prior

- Set $\pi(\theta) \propto I^{0.5}(\theta)$ where $I(\theta) = -E_\theta \left| \frac{\partial^2 \log f(x|\theta)}{\partial \theta^2} \right|$

- What is $I(\theta)$? Fisher information (Fisher, 1956): how much the model discriminates between $\theta$ and $\theta + d\theta$ through the expected slope of $\log f(x|\theta)$.

- Intuition: the prior favors values of $\theta$ for which $I(\theta)$ is large, i.e. it minimizes the influence of the prior distribution.

- Note $I(\theta) = I^{0.5}(h(\theta))(h'(\theta))^2$. Thus, it is invariant under reparametrization.

Our Example of Asymmetric Business Cycles

- Let us assume that number of quarters with growth rate below 1.9% is $\mathcal{B}(n, \theta)$.

- Thus:

$$f(x|\theta) = \left( \begin{array}{c} n \\ x \end{array} \right) \theta^x (1-\theta)^{n-x} \Rightarrow$$

$$\frac{\partial^2 \log f(x|\theta)}{\partial \theta^2} = \frac{x}{\theta^2} + \frac{n-x}{(1-\theta)^2} \Rightarrow$$

$$I(\theta) = n\left[\frac{1}{\theta} + \frac{1}{(1-\theta)}\right] = \frac{n}{\theta(1-\theta)}$$

- Hence: $\pi(\theta) \propto (\theta(1-\theta))^{-0.5}$.

## Some Noninformative Priors II: Multidimensional Jeffreys Prior

- Set $\pi(\theta) \propto [\det I(\theta)]^{0.5}$ where the entries of the matrix are defined as:

$$I_{ij}(\theta) = -E_\theta \left| \frac{\partial^2}{\partial\theta_i \partial\theta_j} \log f_\theta(x) \right|^{0.5}$$

- Note that if $f(x|\theta)$ is exponential (like the Normal):

$$f(x|\theta) = h(x) \exp(\theta x - \psi(\theta))$$

the Fisher information matrix is given by $I(\theta) = \nabla\nabla^t \psi(\theta)$. Thus

$$\pi(\theta) \propto \left( \prod_{i=1}^{k} \frac{\partial^2}{\partial\theta_i^2} \psi(\theta) \right)^{0.5}$$

## An Interesting Application

- Big issue in the 1980s and early 1990s was Unit Roots. Given:

$$y_t = \rho y_{t-1} + \varepsilon_t$$

  what is the value of $\rho$?

- Nelson and Plosser (1982) argued that many macroeconomic time series may present a unit root.

- Why does it matter?

  1. Because non-stationarity changes classical asymptotic theory.

  2. Opens the issue of cointegration.

Exchange between Sims and Phillips about Unit Roots

- Sims and Uhlig (1991), "Understanding Unit Rooters: A Helicopter Tour":

  1. Unit roots are not an issue for Bayesian econometrics.

  2. They whole business is not that important anyway because we will still have .

- Phillips (1991): Sims and Uhlig use a uniform prior. This affects the results a lot.

- Sims (1991): I know!

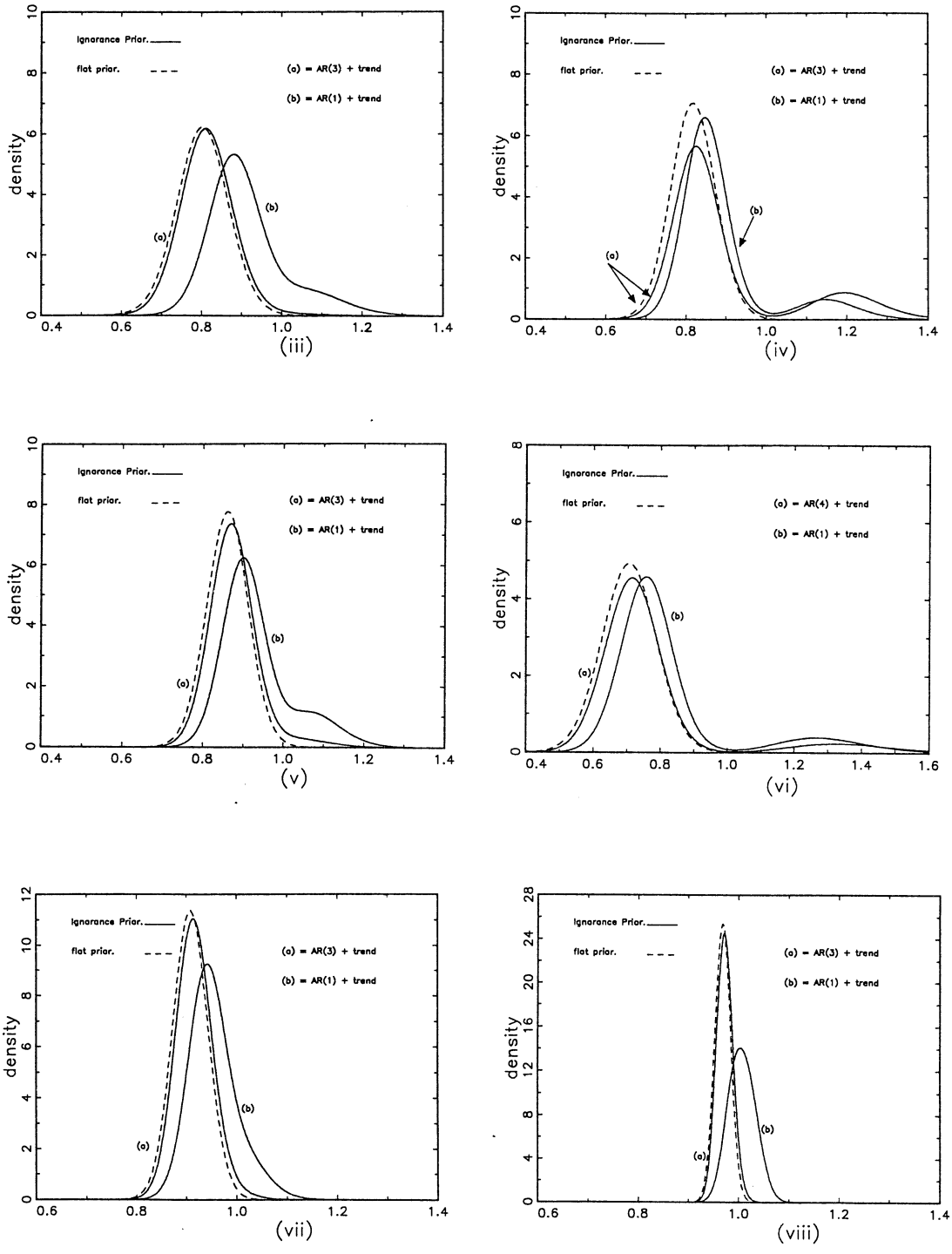
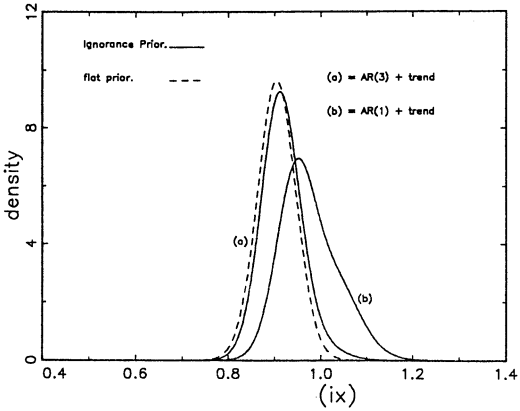
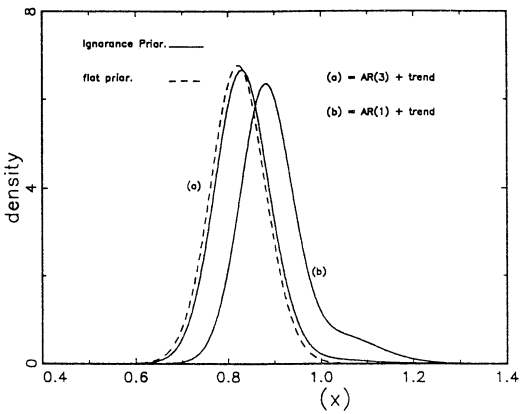
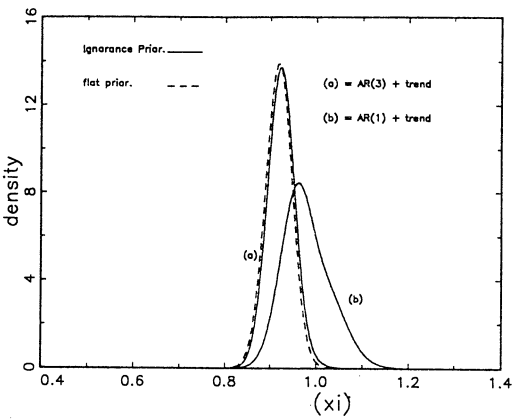
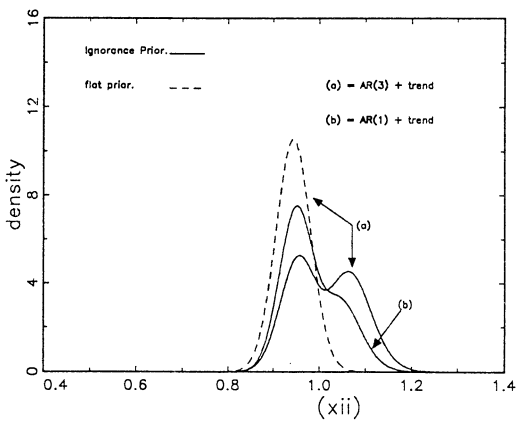


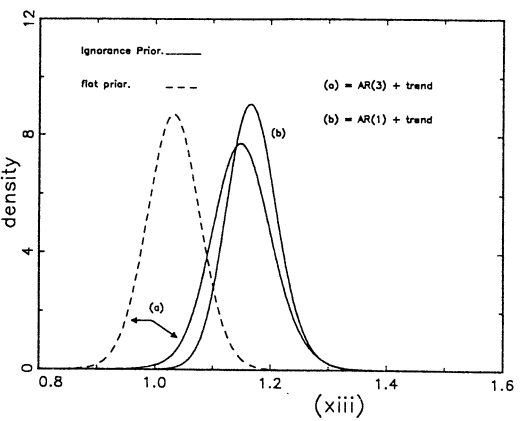

Figure 4. *continued.* (iii) Real per capita GNP: 1909–1970, (iv) Industrial production: 1860–1970, (v) Employment: 1890–1970, (vi) Unemployment rate: 1890–1970, (vii) GNP Deflator: 1889–1970, (viii) Consumer prices: 1860–1970.



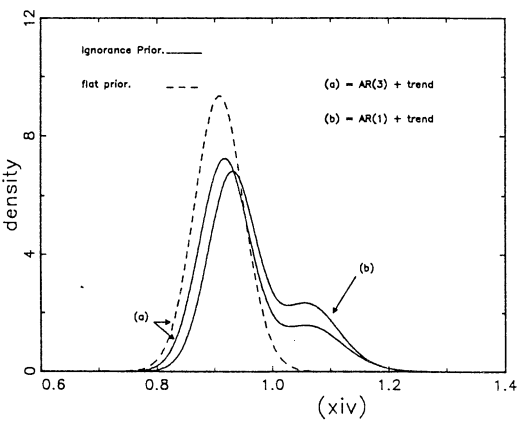




Figure 4. *continued*. (ix) Nominal wages: 1900–1970, (x) Real wages: 1900–1970, (xi) Money stock: 1889–1970, (xii) Velocity: 1869–1970, (xiii) Bond yields: 1900–1970, (xiv) Stock prices (SP500): 1871–1970.

Table IV. Posterior probabilities of stochastic nonstationarity

| Series | AR(1) + trend | | | | AR(3) + trend | | | | |
|---|---|---|---|---|---|---|---|---|---|
| | $P_J(\rho \geqslant 1)$ | $P_F(\rho \geqslant 1)$ | $P_J(\rho \geqslant 0{\cdot}975)$ | $P_F(\rho \geqslant 0{\cdot}975)$ | $P_J(\rho \geqslant 1)$ | $P_F(\rho \geqslant 1)$ | $P_J(\rho \geqslant 0{\cdot}975)$ | $P_F(\rho \geqslant 0{\cdot}975)$ | $P_{DJW}(\Lambda \geqslant 0{\cdot}975)$† |
| Real GNP | 0·193 | 0·023 | 0·242 | 0·054 | 0·012 | 0·002 | 0·019 | 0·005 | 0·003 |
| Nominal GNP | 0·361 | 0·092 | 0·485 | 0·203 | 0·074 | 0·021 | 0·141 | 0·063 | 0·020 |
| Real per capita GNP | 0·163 | 0·018 | 0·206 | 0·044 | 0·010 | 0·001 | 0·016 | 0·004 | 0·003 |
| Industrial production | 0·124 | 0·001 | 0·133 | 0·005 | 0·188 | 0·000 | 0·192 | 0·003 | 0·001 |
| Employment | 0·190 | 0·016 | 0·240 | 0·047 | 0·040 | 0·004 | 0·060 | 0·014 | 0·004 |
| Unemployment* | 0·126 | 0·000 | 0·129 | 0·001 | 0·086 | 0·000 | 0·087 | 0·000 | 0·002 |
| GNP deflator | 0·162 | 0·036 | 0·288 | 0·125 | 0·020 | 0·005 | 0·062 | 0·029 | 0·010 |
| Consumer prices | 0·601 | 0·272 | 0·880 | 0·713 | 0·176 | 0·082 | 0·652 | 0·528 | 0·196 |
| Nominal wages | 0·319 | 0·075 | 0·452 | 0·190 | 0·045 | 0·012 | 0·100 | 0·046 | 0·018 |
| Real wages | 0·103 | 0·011 | 0·140 | 0·031 | 0·014 | 0·001 | 0·021 | 0·005 | 0·003 |
| Money stock | 0·315 | 0·080 | 0·484 | 0·230 | 0·008 | 0·003 | 0·044 | 0·025 | 0·005 |
| Velocity | 0·353 | 0·051 | 0·483 | 0·168 | 0·537 | 0·073 | 0·642 | 0·204 | 0·592 |
| Bond yields | 0·999 | 0·968 | 0·999 | 0·992 | 0·996 | 0·764 | 0·998 | 0·892 | 0·617 |
| Stock prices | 0·301 | 0·028 | 0·385 | 0·092 | 0·215 | 0·017 | 0·278 | 0·059 | 0·040 |

* The penultimate four columns are based on an AR(4) + trend for this series, following Nelson and Plosser (1982).
† From Table 2 of DeJong and Whiteman (1989).

Criticisms of the Jeffreys Prior

- Jeffreys prior lacks a foundation in prior beliefs: it is only a trick.

- Often Jeffreys Prior it is not proper.

- It may violate the Likelihood Principle. Remember our stopping rule example? In the first case, we had a binomial. But we just derived that, for a binomial, the Jeffreys prior is $\pi^1(\theta) \propto (\theta(1-\theta))^{-0.5}$. In the second case, we had a negative binomial, with Jeffreys prior $\pi^2(\theta) \propto \theta^{-1}(1-\theta)^{-0.5}$. They are different!

- We will see later that Jeffreys prior is difficult to apply to equilibrium models.

## A Summary about Priors

- Searching for the right prior is sometimes difficult.
- Good thing about Bayesian: you are putting on the table.
- Some times, an informative prior is useful. For example: new economic phenomena for which we do not have much data.
- Three advises: robustness checks, robustness checks, robustness checks.

Tasks in Front of Us

1. Talk about priors (done).

2. Explain the importance of posteriors and marginal likelihoods.

Why are the Posterior and the Marginal Likelihood So Important?

- Assume we want to explain the following data $Y^T \equiv \left(Y_1', ..., Y_T'\right)'$ defined on a complete probability space $(\Omega, \Im, P_0)$.

- Let $M$ be the set of model. We define a model $i$ as the collection $S(i) \equiv \left\{f\left(T^T|\theta, i\right), \pi(\theta|i), \Theta_i\right\}$, where $f\left(T^T|\theta, i\right)$ is the likelihood, and $\pi(\theta|i)$ is a prior density $\forall i \in M$.

- Define Kullback-Leibler measure:

$$K\left(f^T(\cdot|\theta, i); p_0^T(\cdot)\right) = \int_{\Re^{m\times T}} \log\left(\frac{p_0^T\left(Y^T\right)}{f^T\left(Y^T|\theta, i\right)}\right) p_0^T\left(Y^T\right) d\nu^T$$

- The Kullback-Leibler measure is not a metric, because

$$K\left(f^T(\cdot|\theta,i)\,;p_0^T(\cdot)\right) \neq K\left(p_0^T(\cdot)\,;f^T(\cdot|\theta,i)\right)$$

but it has the following nice properties:

1. $K\left(f^T(\cdot|\theta,i)\,;p_0^T(\cdot)\right) \geq 0$.

2. $K\left(f^T(\cdot|\theta,i)\,;p_0^T(\cdot)\right) = 0$ iff $f^T(\cdot|\theta,i) = p_0^T(\cdot)$.

- Property 1 is obvious because $\log(\cdot) > 0$ and $p_0^T(\cdot) > 0$.

- Property 2 holds because of the following nice property of log function $\rightarrow \log\eta \leq \eta - 1$ and the equality holds only when $\eta = 1$.

The Pseudotrue Value

We can define the pseudotrue value as

$$\theta_T^*(i) \equiv \arg\min_{\theta \in \Theta_i} K\left(f^T(\cdot|\theta, i); p_0^T(\cdot)\right)$$

of $\theta$ that minimizes the Kullback-Leibler distance between $f^T(\cdot|\theta, i)$ and $p_0^T(\cdot)$.

A Couple of Nice Theorems

Fernández-Villaverde and Rubio-Ramírez (2004) show that:

- 1. The posterior distribution of the parameters collapses to the pseudo-true value of the parameter $\theta_T^*(i)$.

$$\pi\left(\theta|Y^T, i\right) \rightarrow^d \chi_{\left\{\theta_T^*(i)\right\}}(\theta)$$

2. If $j \in M$ is the closed model to $P_0$ in the Kullback-Leibler distance sense

$$\lim_{T\rightarrow\infty} P_{0T}\left(\frac{f^T\left(Y^T|i\right)}{f^T\left(Y^T|j\right)} = 0\right) = 1$$

Importance of Theorems

- Result 1 implies that we can use the posterior distribution to estimate the parameters of the model.

- Result 2 implies that we can use the bayes factor to compare between alternative models.

- Both for non-nested and/or misspecified models.

Limitations of the Theorems

- We need to assume that parameter space is finite dimensional.
- Again, we can come up with counter-examples to the theorems when the parameter space is infinite-dimensional (Freedman, 1962).
- Not all is lost, though...
- Growing field of Bayesian Nonparametrics: J.K. Ghosh and R.V. Ramamoorthi, *Bayesian Non Parametrics*, Springer Verlag.

Bayesian Econometrics and Decision Theory

- Bayesian econometrics is explicitly based on Decision Theory.

- Researchers and users are undertaking inference to achieve a goal:

  1. Select right economic theory.

  2. Take the optimal policy decision.

- This purpose may be quite particular to the problem at hand. For example, Schorfheide (2000).

- In that sense, the Bayesian approach is coherent with the rest of economics.

Parameter Estimation

- Loss function

$$\ell(\delta, \theta) : \Theta \times \Theta \rightarrow R^k$$

- Point estimate: $\widehat{\theta}$ such that

$$\widehat{\theta}\left(Y^T, i, \ell\right) = \arg\min_{\delta} \int_{\Theta_i} \ell(\delta, \theta)\, \pi\left(\theta | Y^T, i\right) d\theta$$

Quadratic Loss Function

If the loss function is $\ell(\delta, \theta) = (\delta - \theta)^2 \Rightarrow$ Posterior mean

$$\frac{\partial \int_R (\delta - \theta)^2 \pi\left(\theta | Y^T\right) d\theta}{\partial \delta} = 2 \int_R (\delta - \theta) \pi\left(\theta | Y^T\right) d\theta = 0$$

$$\widehat{\theta}\left(Y^T, \ell\right) = \int_R \theta \pi\left(\theta | Y^T\right) d\theta$$

Absolute Value Loss Function

If the loss function is $\ell(\delta, \theta) = |(\delta - \theta)| \Rightarrow$ Posterior median

$$\int_{-\infty}^{\infty} |\delta - \theta| \, \pi\left(\theta | Y^T\right) d\theta =$$

$$= \int_{-\infty}^{\delta} (\delta - \theta) \, \pi\left(\theta | Y^T\right) d\theta - \int_{\delta}^{\infty} (\delta - \theta) \, \pi\left(\theta | Y^T\right) =$$

$$= \int_{-\infty}^{\delta} P\left(\theta \leq y | Y^T\right) dy - \int_{\delta}^{\infty} P\left(\theta \geq y | Y^T\right) dy$$

Thus

$$\frac{\partial \int_{-\infty}^{\infty} |\delta - \theta| \, \pi\left(\theta | Y^T\right) d\theta}{\partial \delta} = P\left(\theta \leq \delta | Y^T\right) - P\left(\theta \geq \delta | Y^T\right) = 0$$

and

$$P\left(\theta \leq \widehat{\theta}\left(Y^T, \ell\right) | Y^T\right) = \frac{1}{2}$$

## Confidence Sets

- A set $C \subseteq \Theta$ is $1 - \alpha$ credible if:

$$P(\theta \in \Theta) \geq 1 - \alpha$$

- A Highest Posterior Density (HPD) Region is a set $C$ such that:

$$C = \left\{\theta : P\left(\theta | \, Y^T\right) \geq k_\alpha\right\}$$

  where $k_\alpha$ is the largest bound such that $C$ is $1 - \alpha$ credible.

- HPD regions minimize the volume among all $1 - \alpha$ credible sets.

- Comparison with classical confidence intervals.

Hypothesis Testing and Model Comparison

- Bayesian equivalent of classical hypothesis testing.
- A particular case of a more general approach: model comparison.
- We will come back to these issues latter.

Tasks in Front of Us

1. Talk about priors (done).

2. Explain the importance of posteriors and marginal likelihoods (done).

3. Practical implementation.

Three Issues

- Draw from the posterior $\pi\left(\theta|Y^T, i\right)$ (We would need to evaluate $f(Y^T|\theta, i)$ and $\pi\left(\theta|i\right)$).

- Use the Filtering Theory to evaluate $f(Y^T|\theta, i)$ in a DSGE model (We would need to solve the model).

- Compute $P\left(Y^T|i\right)$.

Numerical Problems

- Loss function (Compute expectations).

- Posterior distribution:

$$\pi\left(\theta|Y^T,i\right)=\frac{f(Y^T|\theta,i)\pi\left(\theta|i\right)}{\int_{\Theta_i}f(Y^T|\theta,i)\pi\left(\theta|i\right)d\theta}$$

- Marginal likelihood:

$$P\left(Y^T|i\right)=\int_{\Theta_i}f(Y^T|\theta,i)\pi\left(\theta|i\right)d\theta$$

## How Do We Integrate?

- Exact integration.
- Approximations: Laplace's method.
- Quadrature.
- Monte Carlo simulations.