Bayesian Inference

Empirical Analysis II, Econ 311: Topic 3

Prof. Harald Uhlig[1]

[1]University of Chicago
Department of Economics
huhlig@uchicago.edu

Winter 2019

Outline

Outline

The framework

- (Unknown) parameter $\theta \in \Theta$. Measure $\mu(d\theta)$.
- Observation $x \in X$. Measure $\nu(dx)$.
- Density $f(x \mid \theta)$ wrt ν.
- Likelihood function: $L(\theta \mid x) = f(x \mid \theta)$.
- Experiment on θ. Leads to an observation $x \sim f(x \mid \theta)$ for some known f, if it is carried out.
- Berger-Wolpert (1988).
- Christian P. Robert, *The Bayesian Choice*, Springer, 2nd edition, 2007.

Sufficiency

Definition

A function (“statistic”) T of x is sufficient, if the distribution of x conditional on $T(x)$ does not depend on θ.

Example: $x_i \sim \mathcal{N}(\mu, \sigma^2), i = 1, \ldots, n$, iid. $T(x) = [\bar{x}, s^2]$.

Principle

The Sufficiency Principle: Two observations x, y, which lead to the same value of a sufficient statistic T, $T(x) = T(y)$, shall lead to the same inference regarding θ.

Conditionality

Principle

The Conditionality Principle: If two experiments on θ are available, and if exactly one of these experiments is carried out with some probability p, then the resulting inference on θ should only depend on the selected experiment and the resulting observation.

The Likelihood principle

Principle

The Likelihood Principle:

- *The information brought about by an observation x about θ is entirely contained in the likelihood function $L(\theta \mid x)$.*
- *If two observations x_1 and x_2 lead to proportional likelihood functions,*

$$L(\theta \mid x_1) = cL(\theta \mid x_2), \text{ some } c > 0$$

then they shall lead to the same inference regarding θ.

Theorem

(Birnbaum 1962) The Likelihood Principle is equivalent to the Conditionality Principle and the Sufficiency Principle.

Implementation 1: Maximum Likelihood

- $\hat{\theta} = \arg\max_\theta L(\theta \mid x)$.
- For $\theta \in \mathbb{R}^n$, inference (i.e: standard errors, tests ...) per estimator $\hat{\mathcal{I}}$ of information matrix $\mathcal{I}(\theta)$, etc..

Implementation 2: Bayesian Inference

- Prior $\pi(\theta)$, a density wrt μ.
- Posterior

$$\pi(\theta \mid x) = \frac{L(\theta \mid x)\pi(\theta)}{\int_\Theta L(\theta \mid x)\pi(\theta)\mu(d\theta)}$$

- $m(x) = \int_\Theta L(\theta \mid x)\pi(\theta)\mu(d\theta)$: marginal distribution for x.
- Or:

$$\begin{aligned} \pi(\theta \mid x) &\propto L(\theta \mid x)\pi(\theta) \\ \log \pi(\theta \mid x) &= \log L(\theta \mid x) + \log \pi(\theta) - \log m(x) \end{aligned}$$

- Note: joint density is $f(x \mid \theta)\pi(\theta)$. Apply Bayes' rule,

$$P(A \mid E) = \frac{P(E \mid A)P(A)}{P(E \mid A)P(A) + P(E \mid A^c)P(A^c)}$$

Implementation 2: Bayesian Inference

- Prior $\pi(\theta)$, a density wrt μ.
- Posterior
$$\pi(\theta \mid x) = \frac{L(\theta \mid x)\pi(\theta)}{\int_\Theta L(\theta \mid x)\pi(\theta)\mu(d\theta)}$$
- $m(x) = \int_\Theta L(\theta \mid x)\pi(\theta)\mu(d\theta)$: marginal distribution for x.
- Or:
$$\begin{aligned} \pi(\theta \mid x) &\propto L(\theta \mid x)\pi(\theta) \\ \log \pi(\theta \mid x) &= \log L(\theta \mid x) + \log \pi(\theta) - \log m(x) \end{aligned}$$
- Note: joint density is $f(x \mid \theta)\pi(\theta)$. Apply Bayes' rule,
$$P(A \mid E) = \frac{P(E \mid A)P(A)}{P(E \mid A)P(A) + P(E \mid A^c)P(A^c)}$$

Implementation 2: Bayesian Inference

- Prior $\pi(\theta)$, a density wrt μ.
- Posterior

$$\pi(\theta \mid x) = \frac{L(\theta \mid x)\pi(\theta)}{\int_\Theta L(\theta \mid x)\pi(\theta)\mu(d\theta)}$$

- $m(x) = \int_\Theta L(\theta \mid x)\pi(\theta)\mu(d\theta)$: marginal distribution for x.
- Or:

$$\begin{aligned}\pi(\theta \mid x) &\propto L(\theta \mid x)\pi(\theta) \\ \log \pi(\theta \mid x) &= \log L(\theta \mid x) + \log \pi(\theta) - \log m(x)\end{aligned}$$

- Note: joint density is $f(x \mid \theta)\pi(\theta)$. Apply Bayes' rule,

$$P(A \mid E) = \frac{P(E \mid A)P(A)}{P(E \mid A)P(A) + P(E \mid A^c)P(A^c)}$$

Frequentist vs Bayesian Inference

- Frequentist:
  - Some true θ_0, unknown.
  - The observation $x \sim f(x \mid \theta_0)$ is random.
- Bayesian:
  - The observation $x \sim f(x \mid \theta_0)$ is given at inference time.
  - The "true" parameter $\theta_0 \sim \pi(\theta \mid x)$ is treated as random.

Consequences of the Likelihood Principle

Principle

Stopping Rule Principle: *If a sequence of experiments is directed by a stopping rule τ, which indicates when the experiments stop, then inference about θ shall depend on τ only through the resulting sample.*

Example 1: The conundrum of the experimenter

- Berger-Wolpert, example 19.1
- Experimenter has 100 observations $x_i \sim \mathcal{N}(\theta, 1)$ i.i.d., $\bar{x}_{100} = 0.2$.
- Frequentist test $H_0 : \theta = 0$ vs $H_1 : \theta \neq 0$. Reject at 5% level?
- Stopping rule 1: stop always. $\sqrt{100} \cdot 0.2 > 1.96$: reject.
- Stopping rule 2: if $\sqrt{100} \cdot \bar{x}_{100} \geq c$, stop and reject. If not, take another 100 draws, reject if $\sqrt{200} \cdot \bar{x}_{200} \geq c$.
- Critical value: $c = 2.18$. So, take another 100 draws.
  - Suppose $1.96 < \sqrt{200} \cdot \bar{x}_{200} < 2.18$. Don't reject ... but would have rejected, if the experimenter had not "paused" half-way through.
  - Suppose $\sqrt{200} \cdot \bar{x}_{200} > 2.18$. But: would the experimenter have kept going, if not? Suppose, this depends on whether the RA is available that day or not, which happens with some probability p. Etc.
- The conundrum is avoided by the stopping rule principle.

Example 1: The conundrum of the experimenter

- Berger-Wolpert, example 19.1
- Experimenter has 100 observations $x_i \sim \mathcal{N}(\theta, 1)$ i.i.d., $\bar{x}_{100} = 0.2$.
- Frequentist test $H_0 : \theta = 0$ vs $H_1 : \theta \neq 0$. Reject at 5% level?
- Stopping rule 1: stop always. $\sqrt{100} \cdot 0.2 > 1.96$: reject.
- Stopping rule 2: if $\sqrt{100} \cdot \bar{x}_{100} \geq c$, stop and reject. If not, take another 100 draws, reject if $\sqrt{200} \cdot \bar{x}_{200} \geq c$.
- Critical value: $c = 2.18$. So, take another 100 draws.
  - Suppose $1.96 < \sqrt{200} \cdot \bar{x}_{200} < 2.18$. Don't reject ... but would have rejected, if the experimenter had not "paused" half-way through.
  - Suppose $\sqrt{200} \cdot \bar{x}_{200} > 2.18$. But: would the experimenter have kept going, if not? Suppose, this depends on whether the RA is available that day or not, which happens with some probability p. Etc.
- The conundrum is avoided by the stopping rule principle.

Example 1: The conundrum of the experimenter

- Berger-Wolpert, example 19.1
- Experimenter has 100 observations $x_i \sim \mathcal{N}(\theta, 1)$ i.i.d., $\bar{x}_{100} = 0.2$.
- Frequentist test $H_0 : \theta = 0$ vs $H_1 : \theta \neq 0$. Reject at 5% level?
- Stopping rule 1: stop always. $\sqrt{100} \cdot 0.2 > 1.96$: reject.
- Stopping rule 2: if $\sqrt{100} \cdot \bar{x}_{100} \geq c$, stop and reject. If not, take another 100 draws, reject if $\sqrt{200} \cdot \bar{x}_{200} \geq c$.
- Critical value: $c = 2.18$. So, take another 100 draws.
  - Suppose $1.96 < \sqrt{200} \cdot \bar{x}_{200} < 2.18$. Don't reject ... but would have rejected, if the experimenter had not "paused" half-way through.
  - Suppose $\sqrt{200} \cdot \bar{x}_{200} > 2.18$. But: would the experimenter have kept going, if not? Suppose, this depends on whether the RA is available that day or not, which happens with some probability p. Etc.
- The conundrum is avoided by the stopping rule principle.

Example 2

- $\mathcal{B}(T, \theta)$: Binomial distribution for $x \in \{0, \ldots, T\}$,

$$f(x \mid \theta; T) = \binom{T}{x} \theta^x (1-\theta)^{T-x}$$

  $n = 1$: Bernoulli distribution, $x = 1$ with prob. θ.
- $x_t \sim \mathcal{B}(1, \theta)$ i.i.d.
- Let $x^{(T)} = \sum_{t=1}^{T} x_t$.
- Likelihood: $L\left(\theta \mid x^{(T)}\right) = f\left(x^{(T)} \mid \theta; T\right)$.
- Stopping rule 1: take 100 draws.
- Stopping rule 2: take draws, until $x^{(T)} = T/2$ or $T = 1000000$, whatever comes first.
- Suppose $T = 100$ and $x^{(T)} = T/2$. Stopping rule principle: Inference about θ does not depend on stopping rule.

Example 3

- Robert, p. 18.
- Observations $x_t \sim \mathcal{N}(\theta, 1)$ i.i.d..
- Stopping rule:

$$| \bar{x}_T | = | \frac{1}{T} \sum_{i=1}^{T} x_t | > \frac{1.96}{\sqrt{T}}$$

- (Careless) frequentist: always reject $H_0 : \theta = 0$ at 5% level?!
- Bayesian approach: does not. Shown elsewhere.

Significance Testing

- Berger-Wolpert, Example 30.
x =	0	1	2	3	4
$P(x \mid \theta_0)$	.75	.14	.04	0.037	0.033
$P(x \mid \theta_1)$	.70	.25	.04	0.005	0.005
- $P(x \geq 2 \mid \theta_0) = 0.11$. $P(x \geq 2 \mid \theta_1) = 0.05$.
- Observe $x = 2$. Significance-Testing: significant evidence against θ_1 at 5% level, but not against θ_0.
- Likelihood Principle: the evidence pro or against θ_0 is the same as pro or against θ_1.
- Jeffreys (1961): *"... a hypothesis which may be true may be rejected because it has not predicted observable results which have not occured."*

Significance Testing

- Berger-Wolpert, Example 30.
-

x =	0	1	2	3	4
P($x \mid \theta_0$)	.75	.14	.04	0.037	0.033
P($x \mid \theta_1$)	.70	.25	.04	0.005	0.005

- $P(x \geq 2 \mid \theta_0) = 0.11$. $P(x \geq 2 \mid \theta_1) = 0.05$.
- Observe $x = 2$. Significance-Testing: significant evidence against θ_1 at 5% level, but not against θ_0.
- Likelihood Principle: the evidence pro or against θ_0 is the same as pro or against θ_1.
- Jeffreys (1961): "*... a hypothesis which may be true may be rejected because it has not predicted observable results which have not occured.*"

Outline

The framework

- (Unknown) parameter $\theta \in \Theta \subset \mathbb{R}^m$.
- Observation $x \in \mathbb{R}^n$.
- Density $f(x \mid \theta)$ wrt dx.
- Likelihood function: $L(\theta \mid x) = f(x \mid \theta)$.
- Prior π wrt $d\theta$.
- Decision $\delta(x) \in \mathcal{D}$.
- Loss function $\mathcal{L}(\theta, \delta(x))$.
- Example: quadratic loss, $\mathcal{L}(\theta, \delta(x)) = \| \theta - \delta(x) \|^2$.
- Christian P. Robert, *The Bayesian Choice*, Springer, 2nd edition, 2007.

The framework

- (Unknown) parameter $\theta \in \Theta \subset \mathbb{R}^m$.
- Observation $x \in \mathbb{R}^n$.
- Density $f(x \mid \theta)$ wrt dx.
- Likelihood function: $L(\theta \mid x) = f(x \mid \theta)$.
- Prior π wrt $d\theta$.
- Decision $\delta(x) \in \mathcal{D}$.
- Loss function $\mathcal{L}(\theta, \delta(x))$.
- Example: quadratic loss, $\mathcal{L}(\theta, \delta(x)) = || \theta - \delta(x) ||^2$.
- Christian P. Robert, *The Bayesian Choice*, Springer, 2nd edition, 2007.

Risk

- Average loss / frequentist risk:

$$\mathcal{R}(\theta, \delta) = E_\theta \left[\mathcal{L}(\theta, \delta(x)) \right] = \int_X \mathcal{L}(\theta, \delta(x)) f(x \mid \theta) dx$$

- Bayesian perspective:
  - Posterior expected loss

$$\rho(\pi, \delta(x)) = E_\pi \left[\mathcal{L}(\theta, \delta(x)) \mid x \right] = \int_\Theta \mathcal{L}(\theta, \delta(x)) \pi(\theta \mid x) d\theta$$

  - Integrated risk

$$\begin{aligned} r(\pi, \delta) &= E_\pi [\mathcal{R}(\theta, \delta)] \\ &= \int_\Theta \int_X \mathcal{L}(\theta, \delta(x)) f(x \mid \theta) \pi(\theta) dx d\theta \\ &= \int_X \rho(\pi, \delta(x)) m(x) dx \end{aligned}$$

Admissibility

Definition

An estimator δ_0 is admissible, if there is no estimator δ_1, which dominates δ_0, i.e. which satisfies

$$\mathcal{R}(\theta, \delta_0) \geq \mathcal{R}(\theta, \delta_1)$$

and "$>$" for at least one value θ_0.

Bayes estimators

Definition

- A Bayes estimator associated with a prior distribution π and a loss function $\mathcal{L}$ is any estimator δ^π which minimizes $r(\pi, \delta)$

$$\delta^\pi(x) \in \arg\min_{d \in \mathcal{D}} \rho(\pi, d \mid x)$$

- The value $r(\pi) = r(\pi, \delta^\pi)$ is called the Bayes risk.

Bayes estimators are admissible

Proposition

If π is strictly positive on Θ, with finite Bayes risk and the risk function $\mathcal{R}(\theta, \delta)$ is a continuous function of θ for every δ, then the Bayes estimator δ^π is admissible.

Proposition

If the Bayes estimator associated with a prior π is unique, it is admissible.

See Propositions 2.4.22, 2.4.23 in Robert (2007).

Admissible estimators are Bayes estimators

Theorem

Suppose Θ is compact and $\mathcal{R}$ is convex. If all estimators have a continuous risk function, then, for every non-Bayes estimator δ', there is a Bayes estimator δ^π for some π, which dominates δ', i.e. the Bayes estimators constitute a complete class.

Theorem

Under some mild conditions, all admissible estimators are limits of sequences of Bayes estimators.

See Theorem 8.3.9 and Theorem 8.4.3 in Robert (2007).

The Inadmissibility of the MLE

- Zaman, Asad, *Statistical Foundations for Econometric Techniques,* Academic Press, 1996.
- Suppose that the MLE $\hat{\theta} \in \mathbb{R}^k$, $k \geq 3$ is distributed per

$$\hat{\theta} \sim \mathcal{N}(\theta, I_k)$$

- Quadratic loss function

$$\mathcal{L}(\theta, \delta) = (\delta - \theta)'(\delta - \theta)$$

- James-Stein estimator:

$$\delta_{JS}(\hat{\theta}) = \left(1 - \frac{k-2}{|| \hat{\theta} ||^2}\right) \hat{\theta}$$

Remark

The MLE $\hat{\theta}$ is inadmissible and is dominated by $\delta_{JS}(\hat{\theta})$.

Outline

Exponential Families

Definition

If there are real-valued functions $c_1, \ldots, c_k$ and d of θ and real-valued functions $T_1, \ldots, T_k, S$ on $\mathbb{R}^n$ and a set $A \subset \mathbb{R}^n$ such that

$$f(x \mid \theta) = \exp\left(\sum_{i=1}^{k} c_i(\theta) T_i(x) + d(\theta) + S(x)\right) \mathbf{1}_A(x) \tag{1}$$

for all $\theta \in \Theta$, then $\{f(\cdot \mid \theta) \mid \theta \in \Theta\}$ is called a k-parameter exponential family

Source: Bickel, P.J. and Doksum, K.A., *Mathematical Statistics*, Holden-Day Inc., California, 1977.

Remarks

- The vector $T(x) = (T_1(x), \ldots, T_k(x))$ is sufficient, and is called the natural sufficient statistic of the family.
- Many common probability distributions are exponential.
- Normal distribution $x \sim \mathcal{N}(\mu, \sigma^2)$:

$$f(x \mid \theta) = \exp\left(\frac{\mu}{\sigma^2}x - \frac{x^2}{2\sigma^2} - \frac{1}{2}\left(\frac{\mu^2}{\sigma^2} + \log(2\pi\sigma^2)\right)\right)$$

where

$$\begin{aligned}
c_1(\theta) &= \frac{\mu}{\sigma^2},\ T_1(x) = x \\
c_2(\theta) &= -\frac{1}{2\sigma^2},\ T_2(x) = x^2 \\
d(\theta) &= -\frac{1}{2}\left(\frac{\mu^2}{\sigma^2} + \log(2\pi\sigma^2)\right) \\
S(x) &= 0,\ A = \mathbb{R}
\end{aligned}$$

Conjugacy

Definition

If the prior π is a member of a parametric family of distributions, so that the posterior $\pi(\theta \mid x)$ also belongs to that family, then this family is called conjugate to $\{f(\cdot \mid \theta) \mid \theta \in \Theta\}$.

Conjugacy for exponential families

Proposition

The $(k+1)$*-st parameter exponential family*

$$\pi(\theta; (t_1, \ldots, t_{k+1})) = \exp\left(\sum_{j=1}^{k} c_j(\theta) t_j + t_{k+1} d(\theta) - \log \omega(t_1, \ldots, t_{k+1}) \right)$$

is conjugate to the exponential family (1). The posterior is given by

$$\pi(\theta \mid x) = \pi\left(\theta; (t_1 + T_1(x), \ldots, t_k + T_k(x), t_{k+1} + 1)\right) \tag{2}$$

Normal density, prior and posterior

- $f(x \mid \theta)$ given by $\mathcal{N}(\theta, \sigma^2)$.
- $\pi(\theta)$ given by $\mathcal{N}(\mu, \tau^2)$.
- Posterior $\pi(\theta \mid x)$ is given by $\mathcal{N}(\tilde{\mu}, \tilde{\tau}^2)$ where

$$
\begin{aligned}
\tilde{\tau}^{-2} &= \sigma^{-2} + \tau^{-2} \\
\tilde{\mu} &= \frac{\sigma^{-2}}{\sigma^{-2} + \tau^{-2}} x + \frac{\tau^{-2}}{\sigma^{-2} + \tau^{-2}} \mu
\end{aligned}
$$

- Precisions σ^{-2}, τ^{-2}
- Signal extraction.

Some distributions

- Poisson $\mathcal{P}(\theta), \theta > 0$: $E[x] = \theta$,

$$f(x \mid \theta) = e^{-\theta} \frac{\theta^x}{x!} \mathbf{1}_{\mathbb{N}}(x)$$

- Gamma $\mathcal{G}(\alpha, \beta)$: $E[x] = \alpha / \beta$,

$$f(x \mid \alpha, \beta) = \frac{\beta^\alpha}{\Gamma(\alpha)} x^{\alpha-1} \exp(-\beta x) \mathbf{1}_{[0,\infty)}(x)$$

  Note: $\chi^2_\nu = \mathcal{G}(\nu/2, 1/2)$.

- Beta $Be(\alpha, \beta)$, $\alpha > 0, \beta > 0$: $E[x] = \alpha/(\alpha+\beta)$,

$$f(x \mid \alpha, \beta) = \frac{\Gamma(\alpha+\beta)}{\Gamma(\alpha)\Gamma(\beta)} x^{\alpha-1}(1-x)^{\beta-1} \mathbf{1}_{[0,1]}(x)$$

Some distributions

- Poisson $\mathcal{P}(\theta), \theta > 0$: $E[x] = \theta$,

$$f(x \mid \theta) = e^{-\theta} \frac{\theta^x}{x!} \mathbf{1}_{\mathbb{N}}(x)$$

- Gamma $\mathcal{G}(\alpha, \beta)$: $E[x] = \alpha/\beta$,

$$f(x \mid \alpha, \beta) = \frac{\beta^\alpha}{\Gamma(\alpha)} x^{\alpha-1} \exp(-\beta x) \mathbf{1}_{[0,\infty)}(x)$$

Note: $\chi^2_\nu = \mathcal{G}(\nu/2, 1/2)$.

- Beta $Be(\alpha, \beta)$, $\alpha > 0, \beta > 0$: $E[x] = \alpha/(\alpha + \beta)$,

$$f(x \mid \alpha, \beta) = \frac{\Gamma(\alpha+\beta)}{\Gamma(\alpha)\Gamma(\beta)} x^{\alpha-1} (1-x)^{\beta-1} \mathbf{1}_{[0,1]}(x)$$

Some distributions

- Poisson $\mathcal{P}(\theta), \theta > 0$: $E[x] = \theta$,

$$f(x \mid \theta) = e^{-\theta} \frac{\theta^x}{x!} \mathbf{1}_{\mathbb{N}}(x)$$

- Gamma $\mathcal{G}(\alpha, \beta)$: $E[x] = \alpha/\beta$,

$$f(x \mid \alpha, \beta) = \frac{\beta^\alpha}{\Gamma(\alpha)} x^{\alpha-1} \exp(-\beta x) \mathbf{1}_{[0,\infty)}(x)$$

  Note: $\chi^2_\nu = \mathcal{G}(\nu/2, 1/2)$.

- Beta $Be(\alpha, \beta)$, $\alpha > 0, \beta > 0$: $E[x] = \alpha/(\alpha + \beta)$,

$$f(x \mid \alpha, \beta) = \frac{\Gamma(\alpha+\beta)}{\Gamma(\alpha)\Gamma(\beta)} x^{\alpha-1} (1-x)^{\beta-1} \mathbf{1}_{[0,1]}(x)$$

More priors and posteriors

$f(x \mid \theta)$	π	$\pi(\theta \mid x)$
Binomial $\mathcal{B}(n, \theta)$	Beta $Be(\alpha, \beta)$	Beta $Be(\alpha + x, \beta + n - x)$
Generalizes to Multinomial / Dirichlet		
Normal $\mathcal{N}(\mu, 1/\theta)$	Gamma $\mathcal{G}(\alpha, \beta)$	Gamma $\mathcal{G}(\alpha + 0.5, \beta + (\mu - x)^2/2)$
Gamma $\mathcal{G}(\nu/2, \theta)$	Gamma $\mathcal{G}(\alpha, \beta)$	Gamma $\mathcal{G}(\alpha + \nu/2, \beta + x)$
Poisson $\mathcal{P}(\theta)$	Gamma $\mathcal{G}(\alpha, \beta)$	Gamma $\mathcal{G}(\alpha + x, \beta + 1)$

Source: Robert (2007), Table 3.3.1

Jeffreys prior

- What is a good prior?
- Jeffreys prior: proportional to square root of determinant of information matrix,

$$\pi^*(\theta) \propto \det\left(\mathcal{I}(\theta)\right)^{1/2}, \ \mathcal{I}(\theta) = E_\theta\left[\frac{\partial \log f(x \mid \theta)}{\partial \theta}\left(\frac{\partial \log f(x \mid \theta)}{\partial \theta}\right)'\right]$$

- Jeffreys prior is flat, if $f(x \mid \theta)$ is $\mathcal{N}(\theta, \sigma^2)$.
- Jeffreys prior is invariant to reparameterizations. Suppose, $\psi = h(\theta)$ is 1-1, differentiable with differentiable inverse. Then

$$\det\left(\mathcal{I}(\theta)\right) = \det\left(\mathcal{I}(h(\theta))\right)\det(h'(\theta))^2$$

Outline

Non-conjugate priors

- Last 20 years: development of numerical methods to deal with non-conjugate distributions.
- Markov-Chain-Monte-Carlo (MCMC) methods.
- Metropolis-Hastings algorithm.
- Gibbs-sampling.
- Bayesian inference has been “unchained”.

The question

- Robert (2007).
- To avoid cluttered notation, we shall leave away the conditioning on the observations x, i.e. write $\pi(\theta)$ rather than $\pi(\theta \mid x)$.
- Assumption: the posterior can be written as a density $\pi(\theta)\lambda(d\theta)$ wrt to some measure λ. In slight abuse of notation, we also shall use $\pi(A)$ as the posterior probability for a set A.
- How can we sample from the posterior distribution?
- Typically of interest:

$$E[g(\theta)] = \int_{\Theta} g(\theta)\pi(\theta)\lambda(d\theta) \tag{3}$$

- Numerical integration methods.
- Monte-Carlo integration: calculate $E[g(\theta)]$ as some average of $g(\theta^{(j)}), j = 1, \ldots, n$, where $\theta^{(j)}$ are randomly drawn.
- Note in the calculations below: $\pi(\theta)$ needs to be known only up to a scaling constant.

Importance sampling

- Importance sampling:
- Choose a convenient approximating density $\phi(\theta)\lambda(d\theta)$.
- Take iid samples $\theta^{(j)}, j = 1, \ldots, n$ from it.
- Calculate weights

$$\omega_j = \frac{\pi(\theta^{(j)})}{\phi(\theta^{(j)})}$$

- evaluate integral (3) per weighted average,

$$\bar{g}_n = \frac{\sum_{j=1}^{n} \omega_j g(\theta^{(j)})}{\sum_{j=1}^{n} \omega_j} \tag{4}$$

- Drawback: works badly in high dimensions.

Importance sampling

- Importance sampling:
- Choose a convenient approximating density $\phi(\theta)\lambda(d\theta)$.
- Take iid samples $\theta^{(j)}, j = 1, \ldots, n$ from it.
- Calculate weights

$$\omega_j = \frac{\pi(\theta^{(j)})}{\phi(\theta^{(j)})}$$

- evaluate integral (3) per weighted average,

$$\bar{g}_n = \frac{\sum_{j=1}^n \omega_j g(\theta^{(j)})}{\sum_{j=1}^n \omega_j} \tag{4}$$

- Drawback: works badly in high dimensions.

Markov-Chain Monte Carlo (MCMC) methods

- Markov-Chain Monte Carlo (MCMC) method:
- find a Markov sequence $\theta^{(j)}, j = 1, \ldots, n$ with ergodic distribution $\pi(\theta)$.
- Evaluate integral (3) per sample average,

$$\bar{g}_n = \frac{1}{n} \sum_{j=1}^{n} g(\theta^{(j)}) \tag{5}$$

- $n\bar{g}_n$ is an additive process: adding $g(\theta^{(j)})$, where $\theta^{(j)}$ is Markov. Standard asymptotic theory is available for additive processes, and applies here.

Markov-Chain Monte Carlo (MCMC) methods

- Markov-Chain Monte Carlo (MCMC) method:
- find a Markov sequence $\theta^{(j)}, j = 1, \ldots, n$ with ergodic distribution $\pi(\theta)$.
- Evaluate integral (3) per sample average,

$$\bar{g}_n = \frac{1}{n} \sum_{j=1}^{n} g(\theta^{(j)}) \tag{5}$$

- $n\bar{g}_n$ is an additive process: adding $g(\theta^{(j)})$, where $\theta^{(j)}$ is Markov. Standard asymptotic theory is available for additive processes, and applies here.

Outline

The balance condition

- Consider a Markov chain in θ with transition kernel density $k(\theta' \mid \theta)$, i.e.

$$P(\theta' \in A \mid \theta) = \int_{\theta' \in A} k(\theta' \mid \theta)\lambda(d\theta')$$

  and $P(\theta' \in \Theta \mid \theta) = 1$, all θ.
- Balance condition:

$$k(\theta' \mid \theta)\pi(\theta) = k(\theta \mid \theta')\pi(\theta')$$

- Consequence: $\pi(\theta)$ is a stationary distribution.

The balance condition

- Consider a Markov chain in θ with transition kernel measure $k(d\theta' \mid \theta)$, i.e.

$$\mathrm{P}(\theta' \in A \mid \theta) = \int_{\theta' \in A} k(d\theta' \mid \theta)$$

  and $\mathrm{P}(\theta' \in \Theta \mid \theta) = 1$, all θ.

- Balance condition:

$$k(d\theta' \mid \theta)\pi(\theta)\lambda(d\theta) = k(d\theta \mid \theta')\pi(\theta')\lambda(d\theta')$$

- Consequence: $\pi(\theta)$ is a stationary distribution.

Metropolis-Hastings

- The Metropolis-Hastings algorithm:
- Target distribution: $\pi(\theta)$.
- Pick convenient proposal distributions with densities $q(\theta' \mid \theta)$ (wrt λ).
- Start from any θ_0
- Given $\theta^{(m)}$, generate $\xi \sim q(\xi \mid \theta^{(m)})$.
- Calculate the acceptance probability

$$\varrho(\xi \mid \theta^{(m)}) = \min \left\{ 1, \frac{q(\theta^{(m)} \mid \xi)\pi(\xi)}{q(\xi \mid \theta^{(m)})\pi(\theta^{(m)})} \right\}$$

- Take

$$\theta^{(m+1)} = \begin{cases} \xi & \text{with probability } \varrho(\xi \mid \theta^{(m)}) \\ \theta^{(m)} & \text{otherwise} \end{cases}$$

The random walk proposal distribution

- A popular proposal distributions: a random walk,

$$\xi = \theta^{(m)} + \epsilon$$

  where ϵ has a symmetric distribution around zero, e.g. normal with mean zero.
- Then,

$$\varrho(\xi \mid \theta^{(m)}) = \min\left\{1, \frac{\pi(\xi)}{\pi(\theta^{(m)})}\right\}$$

The kernel of Metropolis-Hastings

- Dirac measure $\delta_\theta(d\theta')$:

$$\int_A \delta_\theta(d\theta') = 1_{\theta \in A}$$

  Thus,

$$\int f(\theta')\delta_\theta(d\theta') = f(\theta)$$

- Kernel of the Metropolis-Hastings algorithm:

$$\begin{aligned} k(d\theta' \mid \theta) &= \varrho(\theta' \mid \theta) q(\theta' \mid \theta)\lambda(d\theta') \\ &\quad + \left(\int (1 - \varrho(\xi \mid \theta)) q(\xi \mid \theta)\lambda(d\xi)\right) \delta_\theta(d\theta') \end{aligned}$$

- One can check that the balance condition is satisfied.

Convergence properties

Theorem

- *If the chain* $(\theta^{(m)})$ *is irreducible, i.e., for any subset A with* $\pi(A) > 0$*, there is some M so that* $P_{\theta_0}(\theta_M \in A) > 0$*, then* $\pi(\theta)$ *is the stationary distribution of the chain.*
- *If, in addition, the chain is aperiodic, it is also ergodic with limiting distribution* $\pi(\theta)$ *for almost every initial value* θ_0*, i.e.*

$$\lim_{m\to\infty} \sup_A \mid P_{\theta_0}\left(\theta^{(m)} \in A\right) - \pi(A) \mid = 0 \,(a.s.)$$

Theorem 6.3.1 in Robert (2007)

An example

- $\theta \in \{a, b\}$, $\pi(a) = p, \pi(b) = 1 - p$, $p > 0.5$.
- $q(\theta' \mid \theta) = \alpha \in (0, 1)$, if $\theta \neq \theta'$ and $q(\theta' \mid \theta) = 1 - \alpha$, if $\theta = \theta'$. Symmetric. Thus, $\varrho(\xi \mid \theta^{(m)}) = \min\left\{1, \pi(\xi)/\pi(\theta^{(m)})\right\}$
- Describing the acceptance probabilities $\varrho(\xi \mid \theta)$:

	$\xi = a$	$\xi = b$
$\theta = a$	1	(1-p)/p
$\theta = b$	1	1

- Transition matrix

$$\mathbf{T} = \begin{bmatrix} 1 - \alpha\frac{1-p}{p} & \alpha\frac{1-p}{p} \\ \alpha & 1 - \alpha \end{bmatrix}$$

- Check that

$$[p, 1-p]\mathbf{T} = [p, 1-p]$$

Outline

Splitting the density: two cases

1. Auxiliary parameters / hierarchical structure: Suppose, that $\pi(\theta)$ can be written as

$$\pi(\theta) = \int \pi_1(\theta \mid \psi)\pi_2(\psi)d\psi$$

   such that the conditional distributions $\pi_1(\theta \mid \psi)$ and $\pi_2(\psi \mid \theta)$ are easy to draw from (Note: $\pi_2(\psi)$ is a marginal distribution).

2. Multivariate $\theta = (\theta_1, \theta_2)$, such that the conditionals $\pi_1(\theta_1 \mid \theta_2)$ and $\pi_2(\theta_2 \mid \theta_1)$ are easy to draw from.

The first case can be considered a version of the second case for the augmented parameter vector $\tilde{\theta} = (\theta, \psi)$.

Slightly more generally

$$\theta = (\theta_1, \ldots, \theta_r)$$

such that the conditionals

$$\pi_j(\theta_j \mid \theta_i, i \neq j), j = 1, \ldots, r$$

are easy to draw from.

The Gibbs-Sampler

The Gibbs-Sampler:

Given $\theta^{(m)} = (\theta_1^{(m)}, \ldots, \theta_r^{(m)})$, draw

1. $\theta_1^{(m+1)} \sim \pi_1(\theta_1 \mid \theta_2^{(m)}, \ldots, \theta_r^{(m)})$
2. $\theta_2^{(m+1)} \sim \pi_2(\theta_2 \mid \theta_1^{(m+1)}, \theta_3^{(m)}, \ldots, \theta_r^{(m)})$

   $\vdots$

r. $\theta_r^{(m+1)} \sim \pi_r(\theta_r \mid \theta_1^{(m+1)}, \ldots, \theta_{r-1}^{(m+1)})$

Ergodicity

Lemma

If $\pi_j(\theta_j \mid \theta_i, i \neq j) > 0$, $j = 1, \ldots, r$, and if the support of π is the Cartesian product of the support of the π_j, the resulting chain is ergodic with stationary distribution π.

See Robert (2007), Lemma 6.3.6, and p. 314.

Modifications

- If the conditional density for θ_j, say, is not easy to draw from, one may instead draw by taking a single Metropolis-Hastings step with that conditional density as target distribution.
- There are other possibilities too. The key is to keep $\pi(\theta)$ as stationary distribution.

Outline

Quantitative macroeconomics

- **D**ynamic **S**tochastic **G**eneral **E**quilibrium (**DSGE**) models.
- Typically: no solution in closed form.
- Log-linearization, solving for the stable roots.
- Numerical methods. “Toolkit”.
- Calibration.
- Estimation.
- Dynare

Dynare

- Dynare: a Matlab-based program, created by Michel Juillard with a community of scholars. Google-search for "Dynare", follow download and installation instructions.
- “addpath c:\dynare\4.1.0\matlab”
- Given (nonlinear) equations of a DSGE model, Dynare ...
  - solves for the steady state,
  - approximates the dynamics around the steady state
  - — first-order (“log-linearization”)
  - — higher-order
  - Simulates
  - Estimates, using MCMC methods.
- “dynare modelfile.mod”

Introduction to Dynare per example

- Source: Barillas-Colacito-Kitao-Matthes-Sargent-Shin, “Practicing Dynare,” draft, NYU 2007.
- a few corrections plus slight modification for Dynare 4.1.0
- A stochastic neoclassical growth (“real business cycle”) model.
- State the model. Pick parameters.
- Solve with Dynare, simulate data with Dynare.
- Estimate with Dynare, using the simulated data.

The model

- Social planner.
- Preferences

$$\max_{\{c_t,l_t\}_{t=0}^{\infty}} E\left[\sum_{t=1}^{\infty}\beta^{t-1}\frac{\left(c_t^{\theta}(1-l_t)^{1-\theta}\right)^{1-\tau}}{1-\tau}\right]$$

- Feasibility constraint:

$$c_t + k_t = e^{z_t}k_{t-1}^{\alpha}l_t^{1-\alpha} + (1-\delta)k_{t-1}$$

- Exogenous productivity:

$$z_t = \rho z_{t-1} + s\epsilon_t,\ \epsilon_t \sim \mathcal{N}(0,1)\, i.i.d.$$

FONCs

- Euler equation:

$$\frac{\left(c_t^{\theta}(1-l_t)^{1-\theta}\right)^{1-\tau}}{c_t} =$$

$$= \quad \beta E_t\left[\frac{\left(c_{t+1}^{\theta}(1-l_{t+1})^{1-\theta}\right)^{1-\tau}}{c_{t+1}}\left(\alpha e^{z_{t+1}} k_t^{\alpha-1} l_{t+1}^{\alpha}+1-\delta\right)\right]$$

- labor market:

$$\frac{1-\theta}{\theta}\frac{c_t}{1-l_t} = (1-\alpha)e^{z_t}k_{t-1}^{\alpha}l_t^{-\alpha}$$

Parameters: Calibration.

Parameter	Calibration
β	0.987
θ	0.357
δ	0.012
α	0.4
τ	2
ρ	0.95
s	0.007

GrowthApproximate_version02.mod: Simulation

```
periods 1000;
var c k lab z;
varexo e;

parameters bet the del alp tau rho s;

bet      = 0.987;
the      = 0.357;
del      = 0.012;
alp      = 0.4;
tau      = 2;
rho      = 0.95;
s        = 0.007;
```

GrowthApproximate_version02.mod: Simulation

```
model;

 (c^the*(1-lab)^(1-the))^(1-tau)/c = bet*
  ((c(+1)^the*(1-lab(+1))^(1-the))^(1-tau)/c(+1))*
  (1+alp*exp(z(+1))*k^(alp-1)*lab(+1)^(1-alp)-del);

  c=the/(1-the)*(1-alp)*exp(z)*
    k(-1)^alp*lab^(-alp)*(1-lab);

  k=exp(z)*k(-1)^alp*lab^(1-alp)-c+(1-del)*k(-1);

  z=rho*z(-1)+s*e;

end;
```

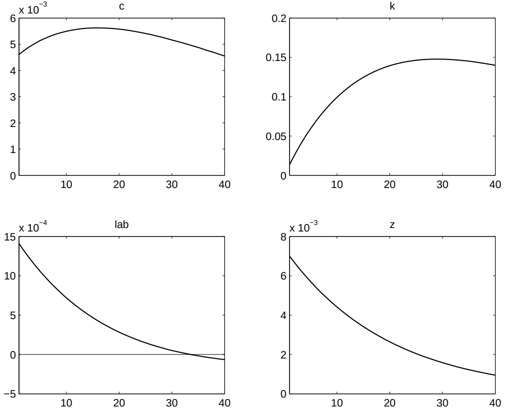
GrowthApproximate_version02.mod: Simulation

```
initval;
k   = 1;
c   = 1;
lab = 0.3;
z   = 0;
e   = 0;
end;
```

GrowthApproximate_version02.mod: Simulation

```
shocks;
var e;
stderr 1;
end;

steady;
stoch_simul(dr_algo=0,periods=1000,irf=40);

// datasaver('simudata',[]);
datasaver_version02('simudata',[]);
```

Impulse response functions



GrowthEstimate.mod: Estimation

```
var c k lab z;
varexo e;

parameters bet del alp rho the tau s;

bet     = 0.987;
the     = 0.357;
del     = 0.012;
alp     = 0.4;
tau     = 2;
rho     = 0.95;
s       = 0.007;
```

GrowthEstimate.mod: Estimation

```
model;

 (c^the*(1-lab)^(1-the))^(1-tau)/c=bet*
  ((c(+1)^the*(1-lab(+1))^(1-the))^(1-tau)/c(+1))*
  (1+alp*exp(z(+1))*k^(alp-1)*lab(+1)^(1-alp)-del);

  c=the/(1-the)*(1-alp)*exp(z)*
    k(-1)^alp*lab^(-alp)*(1-lab);

  k=exp(z)*k(-1)^alp*lab^(1-alp)-c+(1-del)*k(-1);

  z=rho*z(-1)+s*e;

end;
```

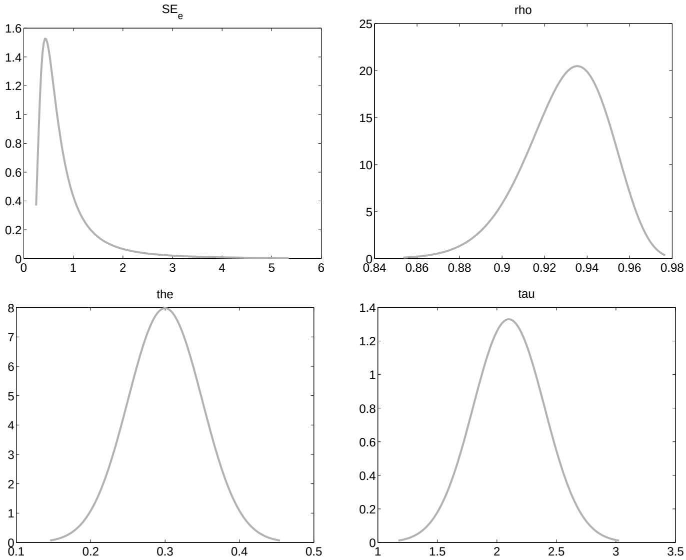
GrowthEstimate.mod: Estimation

```
initval;
k   = 1;
c   = 1;
lab = 0.3;
z   = 0;
e   = 0;
end;

shocks;
var e;
stderr 1;
end;
```

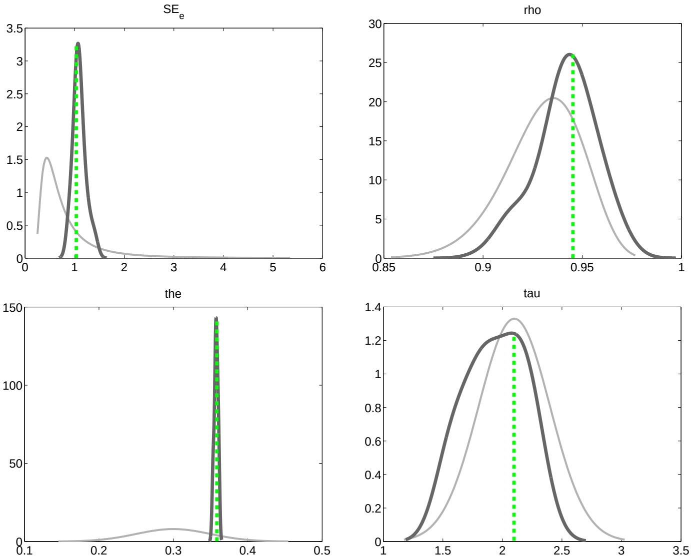
GrowthEstimate.mod: Estimation

```
estimated_params;
stderr e, inv_gamma_pdf, 0.95,30;
rho, beta_pdf,0.93,0.02;
the, normal_pdf,0.3,0.05;
tau, normal_pdf,2.1,0.3;
end;

varobs c;

estimation(datafile=simudata,mh_replic=1000,
  mh_jscale=0.9,nodiagnostic);
```

Priors


SE_e
0 0.2 0.4 0.6 0.8 1 1.2 1.4 1.6
0 1 2 3 4 5 6
rho
0 5 10 15 20 25
0.84 0.86 0.88 0.9 0.92 0.94 0.96 0.98
the
0 1 2 3 4 5 6 7 8
0.1 0.2 0.3 0.4 0.5
tau
0 0.2 0.4 0.6 0.8 1 1.2 1.4
1 1.5 2 2.5 3 3.5

Posteriors



SE_e
3.5
3
2.5
2
1.5
1
0.5
0
0
1
2
3
4
5
6
rho
30
25
20
15
10
5
0
0.85
0.9
0.95
1
the
150
100
50
0
0.1
0.2
0.3
0.4
0.5
tau
1.4
1.2
1
0.8
0.6
0.4
0.2
0
1
1.5
2
2.5
3
3.5

Smoothed Shocks