Ec402: Bayesian Basics

Christian Julliard

Department of Economics and FMG
London School of Economics

Outline

1. Bayesian Preliminaries
   - Bayes' Theorem
   - Mechanics of Priors and Posteriors
   - Likelihood Principle
2. Choosing Priors
3. Model Selection in a Bayesian World
   - Posterior Model Probabilities
   - Bayesian Model Averaging (BMA)

Outline

Bayes' Theorem

- The conditional probability of A given B is

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

  since we can exchange A and B we have
- **Bayes' Theorem:**

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

Parameters as random variables

- We observe the data Z_T ($Z_T := \{z_t\}_{t=0}^{T}$, $z_t \in R^n$), generated by a process with unknown parameters $\theta \in \Theta$, that is realizations from the distribution

$$p\left(Z_T|\theta\right)$$

- We are interested in, and would like to make probabilistic statements about, θ and functions ($g(.)$) of θ and Z_T (e.g. forecasts). That is, we would like to know

$$p\left(\theta|Z_T\right) \text{ and } p\left(g\left(\theta\right)|Z_T\right)$$

- **Remarks:**

Bayesians: θ is a random variable (not $\hat{\theta}$)
Frequentists: $\hat{\theta}$ is a random variable (θ is not)
But: given Z_T, the estimator $\hat{\theta}$ is simply a mapping from Z_T to R^m

From priors to posteriors

- From the Bayes' Theorem

$$p(\theta|Z_T) = \frac{p(Z_T|\theta)\,p(\theta)}{p(Z_T)}$$

- Since $p(Z_T)$ does not depend on θ we can focus on

$$p(\theta|Z_T) \propto \underbrace{\overbrace{p(Z_T|\theta)}^{\text{likelihood of the data}} \times \overbrace{p(\theta)}^{\text{prior}}}_{\text{posterior distribution}}$$

Note: to have the right hand side integrate to 1 and get a proper distribution, we simply have to divide it by

$$k := \int_{\Theta} p(Z_T|\theta)\,p(\theta)\,d\theta$$

- **Remark:** for any non dogmatic prior, as $T \rightarrow \infty$ the likelihood of the data will dominate the shape of the posterior
- **Theorem:** *If there is a consistent estimator $\hat{\theta}_T$ of $\theta \in \Theta$, if $g(\theta)$ is continuous and bounded above and below, and if $E\left[g\left(\theta\right)\right]$ exists, then $E\left[g\left(\theta\right)|Z_T\right]$ (the Bayesian posterior mean for $g\left(\theta\right)$) converges to $g\left(\theta_0\right)$, the true value, with probability one.*
  That is

$$\frac{1}{k}\int g\left(\theta\right)p\left(Z_T|\theta\right)p\left(\theta\right)d\theta \underset{T\longrightarrow\infty}{\longrightarrow} g\left(\theta_0\right) \text{ w.p. } 1$$

  (Proof: Schervish (1995), Theorem 7.78)

Note: the convergence could fail on sets with zero prior probability

The Likelihood Principle

- *"In drawing inference about θ after Z_T is observed, all relevant information is contained in the* ***likelihood function*** $p(Z_T|\theta)$*"* (BLR)
- That is: if our model gives the data the conditional pdf $p(Z_T|\theta)$, all we need to know about the data to form the conditional pdf $p(\theta|Z_T)$ is the likelihood function, and we need to know it only up to a factor of proportionality (follows form Bayes' Th.)
- In a nutshell: suppose we have a thermometer that we know will simply record correctly the temperature up to "30 degrees Celsius." Suppose that right now that thermometer is outside the window and reads 23 degrees Celsius. Should our reaction to this current outside reading be affected by our knowledge that the thermometer's measurements are truncated at higher temperatures? No.

Outline

Flat priors

THE CASE FOR

- Avoid subjective elements in scientific reporting
- *Seem* to represent ignorance or absence of prior prejudice

THE CASE AGAINST

- The scientific reporting argument make sense but depends on
  - the absence of strong economic prior beliefs
  - the parameter space being the “natural” one

But: flat priors do not represent “pure ignorance”:

- generally not “flat” for reparametrizations (think to the change of variable for a distribution)
- might have infinite discontinuities after reparametrization

Example of flat prior pitfall

- Suppose we have the model:
  $y_t = \alpha + \rho y_{t-1} + \varepsilon_t$, with flat prior $p\left(\alpha, \rho, \log \sigma^2\right) = 1$
- For $|\rho| < 1$, the unconditional mean and variance of y are

$$\mu = \frac{\alpha}{1-\rho};\ v^2 = \frac{\sigma_\varepsilon^2}{1-\rho^2}$$

- From the change of variable formula, the prior for the reparametrization in terms of μ ,ρ , $\log v^2$ is

$$\left|\det\left(\frac{\partial\left(\mu, \rho, \log v^2\right)}{\partial\left(\alpha, \rho, \log \sigma^2\right)}\right)\right|^{-1} p\left(\mu \cdot (1-\rho), \rho, \log v^2\right) = |1-\rho|$$

- A flat prior $\alpha, \rho, \log \sigma^2$ implies a non-flat prior for $\mu, \rho, \log v^2$

Note: The new prior puts a very low weight on $|\rho| \cong 1$, and this implies that the frequentist MLE (that is in this case the same thing as OLS) will be biased toward stationarity in small sample.

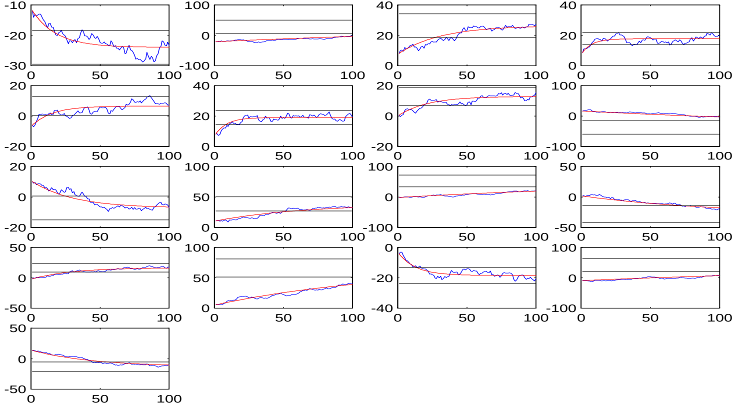

FIGURE 2. Initial Conditions Rogues' Gallery

Note: Rougher lines are Monte Carlo data. Smoother curved lines are deterministic components. Horizontal lines are 95% probability bands around the unconditional mean.

Results from 17 out of 100 OLS estimations of the model $y_t = \alpha + \rho y_{t-1} + \varepsilon_t$ using Monte Carlo data generated by $\{y_t | y_{t-1}\} \sim i.i.d.N\left(y_{t-1}, 1\right)$.

Deterministic component$= \left(1 - \hat{\rho}^t\right) \hat{\alpha} / \left(1 - \hat{\rho}\right) + \hat{\rho}^t y_0$

Source: Sims (2000)

General remarks on choosing priors

- Bauwens, Lubrano and Richard (1999) Ch.4 have an extensive discussion on prior distributions
- Non-flat priors can help a lot in estimating large models and correct small sample problems of the estimators
- Economic theory often give us guidance in choosing priors
- Always check how sensitive are the results to the prior!
- There is no mechanical rule for choosing priors that can give results that will always be reasonable
- **Priors are not a peculiar problem of Bayesian inference:** many classical procedures (e.g. ML) are equivalent to Bayesian inference with a flat prior (with all the related problems). The difference is that in the classical framework we cannot discuss these problems.

Outline

The problem

"All models are wrong, but some are useful."
Box (1976)

Suppose we have:

- **Data** Z ($Z := \{z_t\}_{t=0}^{T}$, $z_t \in R^n$)
- **Model 0** given by the likelihood $f(Z|\theta)$ with $\theta \in \Theta$
- **Model 1** given by the likelihood $g(Z|\psi)$ with $\psi \in \Psi$

Questions:

1. *How do we decide which model is more likely of being the data generating process of Z?*
2. *Suppose we are interested in some quantity Δ that is well-defined for every model (e.g. utility or simply a forecast). How do we compute $E[\Delta|Z]$? How do we measure uncertainty about $E[\Delta|Z]$?*

Posterior Model Probabilities

- define $m = 0$ if model 0 is true and $= 1$ otherwise
- we would like to know the joint distribution $v(\theta, \psi, m|Z)$ to obtain $\Pr(m = 0|Z)$.
- From Bayes' Th.

$$\Pr(m = 0|Z) = \frac{\Pr(Z|m{=}0)\Pr(m{=}0)}{\Pr(Z)} = \frac{\Pr(Z|m{=}0)\Pr(m{=}0)}{\Pr(Z|m{=}0)\Pr(m{=}0)+\Pr(Z|m{=})(1-\Pr(m{=}0))}$$

Note: the distribution of $Z|m$ is simply

$$f(Z|\theta)^{1-m} g(Z|\psi)^m$$

$\Rightarrow$ if we have *prior probability of* $m = 0$ $(\mu(\Theta))$, and priors over θ and ψ ($p(\theta)$ and $q(\psi)$), from Bayes Th. we can compute

posterior probability of $m = 0|Z$

$$\frac{\mu(\Theta)\int_\Theta f(Z|\theta)\,p(\theta)\,d\theta}{\mu(\Theta)\int_\Theta f(Z|\theta)\,p(\theta)\,d\theta + (1-\mu(\Theta))\int_\Psi g(Z|\psi)\,q(\psi)\,d\psi} \tag{1}$$

Bayesian Model Averaging (BMA)

When we are interested in $E[\Delta|Z]$, where Δ is some quantity that is well-defined for every model (e.g. utility), we generally proceed as follow:

1. we select a model using some statistics as selection criterion
2. given the selected model (say model 0) we behave as if

$$E[\Delta|Z] \equiv E[\Delta|Z, m = 0]$$

But:

- this is justified iff $Pr[m = 0|Z] \simeq 1$
- any approach that selects a single model and then makes inference conditionally on that model ignores uncertainty in model selection

Bayesian Model Averaging

If we have K competing models, the posterior mean and standard deviation of Δ are

$$\begin{aligned} E\left[\Delta|Z\right] &= \sum_{k=0}^{K} E\left[\Delta|Z, m=k\right] Pr\left(m=k|Z\right) \\ Var\left[\Delta|Z\right] &= \sum_{k=0}^{K} \left(Var\left[\Delta|Z, m=k\right] + E\left[\Delta|Z, m=k\right]^2\right) Pr\left(m=k|Z\right) \\ &\quad - E\left[\Delta|Z\right]^2 \end{aligned}$$

where $Pr\left(m=k|Z\right)$ is the posterior probability of model k

This result is based on the simple observation that

$$Pr\left(\Delta|Z\right) = \sum_{k=0}^{K} Pr\left(\Delta|Z, m=k\right) Pr\left(m=k|Z\right) \tag{2}$$