Introduction to Bayesian Inference - 3

Bayesian Estimation and Prediction

Eric Jacquier

Boston University Questrom School of Business

MF 840: Financial Econometrics

Spring 2021

1. Basic Framework

2. Application to the Linear Model

3. Estimation

4. Prediction

5. Model Comparison (aka testing, Bayesians don't test)

3. Estimation: Choice of a Location Estimator

3.1 Estimation in classical vs Bayes frameworks

- So far, we have the posterior density of the unknown, random parameters.

  We have only done the "statistical part of the job", characterized our knowledge about the parameter.

  We have not decided what "numbers" to report for the parameter.

  In Bayesian inference, this is a separate part of the process.

- Contrast with classical inference where estimation and statistics are "mixed"

  E.g., OLS, GLS, MLE give point estimates (the "number" to report), and also their statistical properties, often using asymptotic approximations.

- Option pricing model can be used for pricing or hedging – different purposes.

  Hedging and pricing errors are different non-linear functions of the parameters (parameter: volatility).

  Should we minimize sum of squared pricing errors or of squared hedging errors?

  $\text{Min}(C_i^m - C_i^{BS}(\sigma))^2$ / $\text{Min}(\Delta^m - \Delta^{BS}(\sigma))^2$

  We would find two different "numbers" for the parameter estimate!

  Does this make sense?

3.2 Decision Theory to choose what point estimate to report

Review MF793 Lecture Note 3, section 2

- Rational decision on what parameter value to "report" depends on an error loss function, the loss of making an error.

location number

P(θ|D)

Loss function for a choice θ* of the random parameter θ: $L(\theta, \theta^*)$:

Cost of making an error by using that number

Compute the expected loss: $$EL(\theta^*) = \int L(\theta, \theta^*) p(\theta | D) d\theta$$

… The integral is over the random true parameter θ

… We minimize it with respect to θ*, a choice of value on the posterior of θ.

- Decision and information theorists adopted Bayesian Inference early because of the clean rational framework that

  separates statistical information processing (getting the posterior) from decision making (what point estimate to report).

- Mean, Median, Mode are all optimal location estimates to report for a specific Loss function

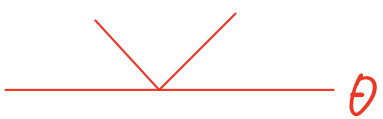
- **Quadratic** Loss: write it for a vector of parameters θ

$E_\theta[(\theta - \theta^*)' W (\theta - \theta^*)]$, where W is a PDS weighting matrix

$= \int (\theta - \theta^*)' W (\theta - \theta^*)\, p(\theta \mid D)\, d\theta$

$0 = \int -2W(\theta - \theta^*)\, p(\theta \mid D)\, d\theta$ => **$\theta^* = E(\theta|D)$**

- **Absolute** Loss: write it for a scalar parameter θ



$E_\theta[|\theta - \theta^*|] = \int |\theta - \theta^*|\, p(\theta|D)\, d\theta$

$= \int_{-\infty}^{\theta^*} (\theta^* - \theta)\, p(\theta|D)\, d\theta + \int_{\theta^*}^{\infty} (\theta - \theta^*)\, p(\theta|D)\, d\theta$

$0 = \int_{-\infty}^{\theta^*} p(\theta|D)\, d\theta - \int_{\theta^*}^{\infty} p(\theta|D)\, d\theta$

⇨ **$\theta^* = \text{Median}(\theta|D)$**

- 0/1 Loss: $L(\theta,\theta^*) = I_{\theta^*=\theta}$

  Different cost of making an error if you under-or-overestimate.

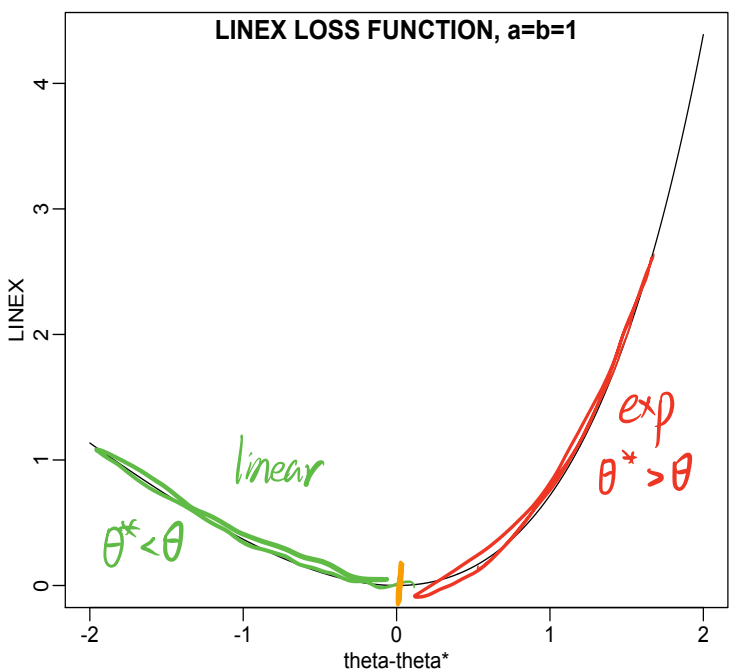
$\theta^* = \text{Mode}(\theta|D)$

θ random

- Hal Varian's **LINEX** linear-exponential Loss:

exponential linear

$$L(\theta, \theta^*) = b\left[e^{a(\theta^*-\theta)} - a(\theta^* - \theta) - 1\right], \qquad b > 0$$



Point estimate? $\frac{\partial E(L(\theta,\theta^*))}{\partial \theta^*}$

$$0 = a\, e^{a\theta^*} E\left(e^{-a\theta}\right) - a$$

$$\Rightarrow \theta^* = \frac{-Log\left(E\left(e^{-a\theta}\right)\right)}{a}$$

no analytical solution

Example: $p(\theta|D) \sim N(\mu,\sigma)$

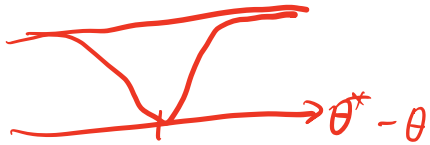
$$\theta^* = \frac{-Log\left(e^{-a\mu+0.5a^2\sigma^2}\right)}{a}$$

like a risk aversion

$$\theta^* = \mu - 0.5a^2\sigma^2$$

Certainty equivalent

- Bounded Loss: $L(\theta, \theta^*) = a\left[1 - e^{-b(\theta^* - \theta)^2}\right],\ a, b > 0$



- By Definition, the Bayesian point estimate is optimal given a choice of Error Loss.

  Randomness is around the true parameter posterior distribution.

  The data is fixed, it is **this** sample *y*.
  Bayesians don't care what could happen for other hypothetical samples.

One can always find a Bayesian point estimate minimizing EL for a given sample

- Frequentist world: optimality results like Gauss-Markov for OLS under ideal assumptions, Cramer-Rao lower bound for MLE is approximate (large sample).

  Randomness is around the estimator due to randomness of the data.

  Frequentists also look at a loss function $L(\theta, \hat{\theta})$, it is called the **risk function**.

  The frequentist loss function is random because $\hat{\theta}$ is random, not θ!

  $\hat{\theta}$ is random because y is random (other samples)

  Integration must be done over y, the is not trivial, there may be some θ left:

从许多θ中找到θ to min L(θ,θ)很难

One can not always find a frequentist estimate that minimizes classical risk over random samples

- The Bayesian and Classical (frequentist) approaches are a bit apples and oranges, we can't compare them strictly but there is one result:

E(θ|D) for my sample $y^{(1)}$

bayesian : optimal for every sample already classic worry about $\{ y^{(2)}, \vdots, y^{(n)}$

1. The Bayesian estimate is optimal by definition for the given sample *y* considered and the loss function chosen.

2. Then minimized Loss can then be considered random over random samples *y*.

classic : E(θ|D) ~ random over samples

Since the Bayesian estimate minimizes expected loss for every sample,
it must be admissible over random sampling when viewed as a classical estimate.

- Enough with fundamental quarrels, let's get practical again

4 Unconditional Prediction in Bayesian Inference

4.1 Regression example

$y = x\beta + \varepsilon$ $\quad p(\beta|\sigma, D)\,p(\sigma|D)$

Need to predict $y_f = x_f' \beta + \varepsilon_f$.

- Conditional prediction: $p(y_f \mid \beta, \sigma, D, x_f) \sim N(x_F' \beta, \sigma)$

  What β to use, $\bar{\bar{\beta}}$ the posterior mean? What σ ?

  Integrated out some to get true ~

- But: $\mathrm{Var}(y_F \mid \sigma, D, x_F) = V(x_F'\beta + \varepsilon_F \mid D, x_F)$

  $= x_F' V(\beta \mid D)\, x_F + \sigma^2 = \sigma^2(1 + x_F'(X'X)^{-1} x_F)$

  diffuse prior $\hookrightarrow \sigma^2(x'x)^{-1}$

  $E(y_F \mid \sigma, D, x_F) = x_F' E(\beta|D) = x_F'\bar{\bar{\beta}}$

  We just informally integrated out β ! We have $p(yf \mid D, \sigma, x_f)$

- Just remains σ.

$$p(y_f|D) = \int \underbrace{\frac{1}{\sigma} e^{\frac{-(y_f - x_f'\bar{\bar{\beta}})^2}{2V(y_f|\sigma,D)}}}_{p(y_f|\sigma,D)} \underbrace{e}_{p(\sigma|D)} = \int \frac{1}{\sigma^{v+1}} e^{\frac{-A}{2\sigma^2}}\, d\sigma$$

We should always use the **unconditional predictive** density .. to make predictions:

By definition of a marginal density:

$$p(y_F \mid D, x_F) = \int p(y_F \mid \beta, \sigma, D, x_F)\; p(\beta, \sigma \mid D)\, d\beta\, d\sigma$$

posterior

- Let's integrate σ out. Using the integral result of the posterior (when we went from β|σ to β), we have:

$$y_F \mid D, x_F \propto [\nu s^2 + (y_F - x_F'\bar{\beta})' V^{-1} (y_F - x_F'\bar{\beta})]^{-\frac{\nu+1}{2}} \qquad [1]$$

Student-t

Where $V = (1 + x_F'(X'X)^{-1}x_F)$ diffuse priors

- Done …. But it's not always that simple

- When it's not simple, we will integrate by simulation, let's see an example

4.2 Multistep Forecasts, using Monte Carlo simulations

HW3

Aug 17, 2015 , 8124 8131

[2 ... T] [1 ... T-1]

AR(1): $Y_t = \alpha + \beta Y_{t-1} + \varepsilon_t$ $t = 1, \ldots T$

$P(\alpha, \beta | \sigma, D)$

$P(\sigma | D)$

Multi-step forecasts

$Y_{T+1} = \alpha + \beta Y_T + \varepsilon_{T+1}$	$Y_{T+1} \mid D$	$\sim$ Student-t per [1] above
$Y_{T+2} = \alpha + \beta Y_{T+1} + \varepsilon_{T+2}$	$Y_{T+2} \mid Y_{T+1}, D$	$\sim$ Student-t

NO! $(Y_{T+2} | \hat{Y}_{T+1}, D)$ is not correct, it does not reflect the whole uncertainty of Y_{T+2}.

yes! $P(Y_{T+2} \mid D) = \int P(Y_{T+2} | Y_{T+1}, D) P(Y_{T+1} | D) dY_{T+1}$ No Analytical solution!

Student-T

Integrating by simulation will be very easy though.

- Integrating by Monte-Carlo simulation

	$\sigma\|D$	$\alpha,\beta\|\sigma,D$	ε_{T+1}	Y_{T+1}	$\alpha+\beta Y_{T+1}=E(Y_{T+2}\|Y_{T+1})$	ε_{T+2}	Y_{T+2}
Draw	(1)	(1)	(1)	compute	compute	(1)	compute
(2)	..	..	..		..	..	..
10000	(S)	(S)	(S)	(S)	(S)	(S)	(S)

- What is our first (the **seed**) draw?

  σ is the **only** parameter for which we have a marginal distribution, its posterior denstiy is the parent distribution of all our simulations

- What are the draws of the second column, $p(\alpha,\beta\,|D)$ or $p(\alpha,\beta\,|\sigma,D)$?

  $p(\alpha,\beta|D)=\int p(\alpha,\beta\,|\sigma,D)\;\; p(\sigma|D)d\sigma$ p(σ|D): like a weight in the summation

- S draws:

  $\frac{1}{S}\sum_1^S Y_{T+2}^{(n)}$ estimates $E(Y_{T+2}\,|\,D)$, How about the last column?

Then what does $\frac{1}{S}\sum_1^S EY_{T+2}\,|Y_{T+1}$ do? (sample average of column 5)

- How do we estimate V(β|D)? Quantiles of (β|D)? or of Y_{T+2}|D