Time Series and Dynamic Models

Section 1
Intro to Bayesian Inference

Carlos M. Carvalho
The University of Texas at Austin

Outline[1]

1. Foundations of Bayesian Statistics
2. Bayesian Estimation
3. The Normal Model
4. Multivariate Normal
5. Linear Model

[1] Reference: *A First Course in Bayesian Statistical Methods* by Peter Hoff

Foundations

- *Statistical Inference:* Study of problems in which data has been generated in accordance with an unknown (random) process. The main goal is to come up with strategies that allow us to make statements about these processes.
- *Probability Models:* Simplified vehicle to seek the understanding of the unknown process... usually involves a set of rules, functional forms and parameters.
- *Probability:* Language used to address uncertainty about unknown quantities.
- *Information Set:* Main input in statistics... Data, scientific knowledge, assumptions, priors.

Probability

- *Definition:* OK!
- *Interpretation:* NOT...
  - Classical or Physical
  - Frequentist
  - Subjective

Probability

- *Classical:* Based on the concept of "equally likely outcomes".
  - Example: Coin Toss... Two outcomes, H and T... 1/2 each
  - No systematic method for assigning probabilities when outcomes are not equally likely

- *Frequentist:* Relative frequency of an outcome/event when the experiment is repeated a large number of times.

$$Pr(A) = lim_{n\to\infty}\frac{m}{n}$$

  - main shortcoming is that it only applies to problems when replication is possible.

- *Question:* How do we deal with statements like:
  - Will it snow today?
  - Will John get married to Mary?
  - Will the U.S. default on its debt?

Probability

- *Subjective:* *"Probability does not exist"* (De Finneti). The probability of an event A is a measure of someone's beliefs in the occurrence of A. Explicitly recognizes the subjective aspects of the scientific process

- *Example:* $A =$ "snow in Chicago"
  - For someone in Rio de Janeiro:

$$Pr(A|I_1) = 0.5$$

  - For someone in Peoria, IL

$$Pr(A|I_2) = \begin{cases} 0.75 & \text{if snow in Peoria} \\ 0.25 & \text{otherwise} \end{cases}$$

  - For someone in Chicago

$$Pr(A|I_3) = \begin{cases} 1 \\ 0 \end{cases}$$

Boy or Girl?

- One morning, on my way to work, I was stopped by a pregnant lady in Hyde Park. She was anxious to know the chance of her seventh baby to be male!
  - First instincts... $Pr(M|I_0) = 0.5$
  - Additional information: $I_1 = \{MMMMMF\}$
  - $Pr(A|I_0, I_1)$?
- *Probability Model:* $Pr(X_i = 1|\theta) = \theta$
  - Joint distribution
    $Pr(X_1 = 1, X_2 = 1, \ldots, X_6 = 0|\theta) = \theta^5(1-\theta)^1$
    What am I assuming here?
  - MLE: $\hat{\theta} = 5/6 = 0.83$

Alternative

$$
\begin{aligned}
Pr(X_7 = 1|I_1) &= \int_0^1 Pr(X_7 = 1, \theta|I_1)d\theta \\
&= \int_0^1 Pr(X_7 = 1|\theta, I_1)p(\theta|I_1)d\theta \\
&= \int_0^1 Pr(X_7 = 1|\theta)p(\theta|I_1)d\theta \\
&= \int_0^1 \theta p(\theta|I_1)d\theta \\
&= \mathrm{E}(\theta|I_1)
\end{aligned}
$$

- *Prior:* $p(\theta|I_0) = p(\theta) \rightarrow U(0,1)$
- *Posterior:* $p(\theta|I_1) \propto p(X_1, \ldots X_6|\theta) \times p(\theta)$

$Pr(X_7 = 1|I_1) = \mathrm{E}(\theta|I_1) = 0.75$ (Why?) $\rightarrow$ Shrinkage!

- *Bayesian Inference:* uses probability statements about unknown quantities conditional on the data and prior information as the basis for inference.

Representation Theorem

$$Pr(X_1, \ldots, X_n) = \int \theta^{S_n}(1-\theta)^{n-S_n} d\mu(\theta)$$

• de Finneti's Representation Theorem justify the above model by using one assumption: Exchangeability

• If there is an infinite sequence of exchangeable random quantities $\{X_n\}_{n=1}^{\infty}$ then there must be some random quantity θ such that the X_is are conditionally IID given θ. If the random variables are Bernoulli, θ can be taken to be the limit of proportions of successes in the first n observations. (Equivalent to strong law of large numbers)

• Exchangeability implies that θ is given an implicit meaning as a random variable rather than a fixed value.

• Provides a connection between two worlds: observables vs. mathematical constructs such as θ.

Conditioning and the Likelihood Principle

• *Likelihood Principle:* The likelihood function $L(\theta)$ contains all the relevant information about θ from the data. Moreover, two likelihoods contain the same information about the θ if they are proportional to each other.

• The Likelihood Principle makes explicit the natural conditional idea that *only* the actual observed x should be relevant to conclusions or evidence about θ. (Contrast with any sampling-based approach to inference)

• *Example...* Highlights that the Bayesian inference get to answers based only conditional on the observed data (and prior information) and not based on the distribution of estimators and test statistics that rely on non-observable quantities.

Some *Pro* and *Cons*

- Self-contained paradigm for statistical inference that will always quantify uncertainty via probabilistic statements conditionally on all available information. Uncertainty about all unknows is quantified... parameters, models, etc...

- Provide a coherent framework for the incorporation of prior information.

- Unified treatment a decision theory and inference.

- Equivalence of classically optimal and Bayes rules. Complete Class Theorem states that all "admissible" decision rules (estimators) correspond to a Bayes rule. This also includes all rules related to most powerful tests. Moreover, Bayesian estimators are consistent, asymptotic normal and efficient given very mild regularity conditions. Bayesian procedures are almost always equivalent to classical large sample strategies... advantages appear in finite sample situations.

Some *Pro* and *Cons*

• Operational advantages. (1) Conditioning on the observed data introduces great simplification in the analysis. No need to average over the data space. (2) The posterior distribution provides a simple mechanism to address a large number of questions simultaneously.

• It's the natural strategy for dealing with sequential inference and therefore perfect for dynamic predictive models.

• Prior specification can be viewed as a drawback.

• Requirement of a likelihood function...

• Computational challenges... Bayesians need to compute all sorts of integrals!

Bayesian Estimation

- Basic elements:
  - (1) Prior... $p(\theta)$
  - (2) Likelihood... $p(x_1, \ldots, x_n|\theta)$
  - (3) Posterior... $p(\theta|x_1, \ldots, x_n)$
  - (4) Predictive... $p(x_1, \ldots, x_n)$
  - (4) Posterior Predictive... $p(x_{n+1}|x_1, \ldots, x_n)$

Bayesian Estimation

- By a simple application of Bayes' Theorem:

$$\begin{aligned} p(\theta|X) &= \frac{p(\theta, X)}{p(X)} \\ &= \frac{p(X|\theta)p(\theta)}{p(X)} \\ &\propto p(X|\theta)p(\theta) \end{aligned}$$

- Predictive or marginal distribution of the data:

$$p(X) = \int p(\theta, X) d\theta = \mathrm{E}_\theta(p(X|\theta))$$

Bayesian Estimation

- Posterior Predictive:

$$
\begin{aligned}
p(x_{n+1}|X) &= \int p(x_{n+1}, \theta|X) d\theta \\
&= \int p(x_{n+1}|\theta, X) p(\theta|X) d\theta \\
&= \int p(x_{n+1}|\theta) p(\theta|X) d\theta \qquad \text{Why?} \\
&= \mathrm{E}_{\theta|X}(p(x_{n+1}|\theta))
\end{aligned}
$$

- Sequential nature of Bayesian analysis... notice that both the predictive and posterior predictive are NOT a function of the unobservable quantity θ

Bayesian Estimation

- Conjugate Families:
  - $N(\theta, \sigma^2)$ – Normal / IG
  - $Po(\lambda)$ – Gamma
  - $Ber(p)$ – Beta
  - $Exp(\lambda)$ – Gamma
  - etc...

The Normal Model

We start with the model

- $X \sim N(\theta, \sigma^2)$ where we assume knowledge of σ^2
- $p(\theta) = N(m_0, C_0)$

We then observe i.i.d. data $\{x_1, x_2, \ldots, x_n\}$

How should I estimate θ ?

The object of inference is always the posterior distribution!

$$p(\theta|x_1, x_2, \ldots, x_n) \propto p(x_1, x_2, \ldots, x_n|\theta)p(\theta)$$

The Normal Model

- It turns out that

$$p(\theta|x_1, x_2, \ldots, x_n) = N(m_1, C_1)$$

where

  - $C_1 = \left(\frac{n}{\sigma^2} + \frac{1}{C_0}\right)^{-1}$
  - $m_1 = C_1\left(\frac{n\bar{X}}{\sigma^2} + \frac{m_0}{C_0}\right)$

We can also write

$$m_1 = \left(\frac{nC_0}{nC_0 + \sigma^2}\right)\bar{X} + \left(\frac{\sigma^2}{nC_0 + \sigma^2}\right)m_0$$

a weighted combination of the prior and experimental information

What is the predictive for x_{n+1}?

The Normal Model

- Let's think about the result above...
  - what happens when $C_0 \to 0$?
  - what about $C_0 \to \infty$?

- Can you contrast this approach to inference to the frequentist, classical approach based on the sampling distribution?

$$\bar{X} \sim N\left(\theta, \frac{\sigma^2}{n}\right)$$

Multivariate Normal

Let a random p-vector $\mathbf{x}$ follow a multivariate normal distribution in p dimensions

$$\mathbf{x} \sim N_p(\mathbf{m}, \mathbf{V})$$

where $E(\mathbf{x}) = \mathbf{m}$ and $Var(\mathbf{x}) = \mathbf{V}$ (what are the dimensions of $\mathbf{V}$?). The density is defined as:

$$p(\mathbf{x}) = \{(2\pi)^p |\mathbf{V}|\}^{-1/2} \exp\left[-\frac{1}{2}(\mathbf{x}-\mathbf{m})'\mathbf{V}^{-1}(\mathbf{x}-\mathbf{m})\right]$$

Multivariate Normal

Suppose that we have the conformable partitions

$$\mathbf{x} = \begin{pmatrix} \mathbf{x}_1 \\ \mathbf{x}_2 \end{pmatrix}, \quad \mathbf{m} = \begin{pmatrix} \mathbf{m}_1 \\ \mathbf{m}_2 \end{pmatrix} \quad \text{and} \quad \mathbf{V} = \begin{pmatrix} \mathbf{V}_{11} & \mathbf{V}_{12} \\ \mathbf{V}_{21} & \mathbf{V}_{22} \end{pmatrix}$$

The marginal distributions are $\mathbf{x}_i \sim N(\mathbf{m}_i, \mathbf{V}_{ii})$

and the conditional $(\mathbf{x}_1|\mathbf{x}_2) \sim N(\mathbf{m}_{1.2}, \mathbf{V}_{1.2})$ where

$$\begin{aligned} \mathbf{m}_{1.2} &= \mathbf{m}_1 + \mathbf{V}_{12}\mathbf{V}_{22}^{-1}(\mathbf{x}_2 - \mathbf{m}_2) \\ \mathbf{V}_{1.2} &= \mathbf{V}_{11} - \mathbf{V}_{12}\mathbf{V}_{22}^{-1}\mathbf{V}_{21} \end{aligned}$$

Linear Model

Consider the standard multiple regression model... (assume knowledge of σ^2)

$$Y = X'\beta + \epsilon$$

where $\epsilon \sim N_n(0, \sigma^2 I_n)$
Assume the prior $\beta \sim N_p(m_0, C_0)$

The posterior is defined as $(\beta|Y) \sim N_p(m_1, C_1)$ where

$$\begin{aligned} C_1^{-1} &= \left(C_0^{-1} + X'X/\sigma^2\right)^{-1} \\ m_1 &= C_1\left[C_0^{-1}m_0 + X'Y/\sigma^2\right] \end{aligned}$$

Let's make sure this looks right...

- What about σ^2?
- and $p(Y)$?