Harvard Economics Math Camp 2018: Econometrics, Bayesian Inference

Ashesh Rambachan[1]

August 2018

[1] These notes are heavily based on materials from many textbooks (see references below) and draw heavily upon notes from other econometrics and statistics courses. I take ZERO credit and all errors are my own.

These notes provide a brief introduction into Bayesian inference. It is intended to provide a simple, very high-level framework for thinking about many of the tools that will be discussed in 2120. I draw heavily on Chapters 2-3 of *Computer Age Statistical Inference* by Bradley Efron and Trevor Hastie and Gary Chamberlain's lecture note 5 for Ec 2120 for these notes. [2]

[2] This is an excellent textbook and its pdf is available for free online!

DISCLAIMERS:

1. There is *absolutely no* expectation for you to read these notes prior to math camp. Maximize utility as you see fit.
2. You are *not* expected to master of this content before the fall. This is intended to provide a brief refresher on some basic concepts and preview some material that will be covered in the first year econometrics sequence. If some of the material is unfamiliar, *do not worry*.
3. These notes contain more content than we will have time to cover during math camp. This is intentional. Hopefully these notes can be a reference material for you throughout the year.

Contents

What is Bayesian Inference?

We begin by observing some data x_i for $i = 1, \ldots, n$ and assume that these data are the result of some random experiment. We model this random experiment with random variable X with support $\mathcal{X}$ and so, the data $\{x_i\}_{i=1}^n$ are realizations of X. We wish to use the data to learn something about the distribution of X, $F_X(x)$.

To do so, we construct a **statistical model**. A statistical model is a set of probability distributions indexed by a parameter set. That is, $\mathcal{F} = \{P_\theta(x) : x \in \mathcal{X}, \theta \in \Theta\}$ is a statistical model. A model is **parametric** if P can be indexed with a finite dimensional parameter set. Otherwise, it is **non-parametric**. The econometrician observes $\{x_i\}_{i=1}^n$ and wishes to make inferences about θ.

Example 0.1. *Suppose our statistical model is the set of normal distributions with variance equal to one. Then,* $\mathcal{X} = \mathbb{R}$, $\Theta = \mathbb{R}$ *and*

$$f_\theta(x) = \frac{1}{\sqrt{2\pi}} e^{-\frac{1}{2}(x-\theta)^2}.$$

Example 0.2. *Suppose our statistical model is the set of Poisson distributions. Then,* $\mathcal{X} = \mathcal{N}$, $\Theta = \mathbb{R}_+$ *and*

$$f_\theta(x) = e^{-\theta}\theta^x / x!.$$

So, both frequentists and bayesians begin with a probability model and wish to learn about the parameter θ. What makes them different? Go back to the definition of a statistical model. Suppose we have a "good" statistical model. That is, $F_X(x) \in \mathcal{F}$ and so, there exists some $\theta^* \in \Theta$ such that $F_X(x) = F_{\theta^*}(x)$. *The whole point of statistical inference is that* θ^* *is unknown.* I think the key difference between frequentists and bayesians is in how they model an unknown θ^* and what that, in turn, implies for how inference should be conducted.[3]

[3] If you ask some people, they will emphasize that the difference lies in how frequentists and bayesians interpret probability.

Frequentists assume that even though θ^* is unknown, we should view it as *fixed*. The data are modeled as random variables $X_1, \ldots, X_n$ drawn from the fixed, unknown distribution $F_{\theta^*}(x)$. Put in another way, frequentists model the random experiment as:

1. Nature draws realizations $x_1, \ldots, x_n$ from the distribution $F_{\theta^*}(x)$. These are the data.

2. The econometrician observes the data $x_1, \ldots, x_n$ and plugs them into her estimator, $\hat{\theta}(\cdot)$. Her estimate of $\hat{\theta}^*$ is $\hat{\theta}(x_1, \ldots, x_n)$.

With this in mind, frequentists then perform the following though experiment.

> Suppose I were to repeat the random experiment above many times. Each time I repeat the experiment, I obtain new data

$x_1^b, \ldots, x_n^b$ and construct a new estimate using my estimator, $\hat{\theta}(x_1^b, \ldots, x_n^b) = \hat{\theta}^b$. What properties will the **sampling distribution** of my estimator have? That is, as $B \to \infty$, what properties will the distribution of $(\hat{\theta}^1, \ldots, \hat{\theta}^B)$ have?

For this reason, Bradley Efron and Trevor Hastie note that *"behaviorism"* would be a better name for frequentism because it better focuses attention on the emphasis placed the *behavior* of estimators in a **repeated random experiment**.[4] An example of a desirable property for frequentists is **unbiasedness**. An estimator $\hat{\theta}(\cdot)$ is **unbiased** if $E[\hat{\theta}(X_1, \ldots, X_n)] = \theta^*$. Note that this expectation is taken over the sampling distribution of the estimator $\hat{\theta}$.

[4] Another way of thinking about this is: In frequentist calculations, θ^* is fixed and the data varies over different possible realizations conditional on θ^*.

Bayesians, on the other hand, prefer to model the unknown $\hat{\theta}^*$ as a random variable itself. $\hat{\theta}^*$ is a random variable that has its own distriution, $\Pi(\theta)$. This is called the **prior distribution**. The random experiment then has an extra step:

1. Nature draws θ^* from the prior distribution, $\Pi(\theta)$. This is unobserved.

2. Nature draws realizations $x_1, \ldots, x_n$ from the distribution $F_{\theta^*}(x)$. These are the data.

3. The econometrician observes $x_1, \ldots, x_n$ and plugs them into her estimator, $\hat{\theta}(\cdot)$. Her estimate is $\hat{\theta}(x_1, \ldots, x_n)$.

Clearly, the prior distribution will be an important part of Bayesian inference. How should we think about it? For now, think of the prior distribution as encoding *prior information* about the parameter θ available to the econometrician prior to observing the data. This may come from prior experiments, observational studies or economic theory.

What is the point of adding this additional layer? The payoff comes from the use of **Bayes' rule**. Bayes' rule provides a logically consistent rule for combining prior information with the observed data. Let $x = (x_1, \ldots, x_n)$ and let $f_\theta(x)$ denote the density associated with the distribution $F_\theta(x)$ and $\pi(\theta)$ is defined analogously. Bayes' rule tells us

$$\pi(\theta|x) = \frac{f_\theta(x)\pi(\theta)}{f(x)}$$

where $f(x) = \int_\Theta f_\theta(x)\pi(\theta)d\theta$ is the **marginal density** of X. $f_\theta(x)$ is the **likelihood function**. We call $\pi(\theta|x)$ the **posterior density** of θ and it is the key object of Bayesian inference. [5] The bayesian then uses the posterior distribution to make inferences about θ. For example, a common object of interest is the "posterior expectation of θ given the data x"

$$E[\theta|x].$$

[5] You will often see Bayes' rule written as

$$\pi(\theta|x) \propto f_\theta(x)\pi(\theta)$$

where $\propto$ means "is proportional to." In English Bayes' rule says, "the posterior is proportional to the likelihood times the prior."

However with the posterior distribution, the bayesian immediately answers all possible questions about θ. She could compute $E[\theta|x], Med(\theta|X), P(\theta < \tilde{\theta}|X)$ and so on. It is critical to note that in the posterior density, x is *fixed* at its realized value and θ varies over Θ. In this sense, bayesian inference is completely *conditional on the observed data*.

Conjugate Priors

As mentioned above, the choice of the prior distribution is the key step of bayesian inference. Once we have a prior distribution and a likelihood function, the only computational step is to use Bayes' rule to form the posterior. While it sounds simple, this can often be a mess unless we carefully choose the prior distribution for a given likelihood function.

As a result, an important tool in bayesian inference are **conjugate priors**. A prior distribution is **conjugate** for a given likelihood function if the associated posterior distribution is in the same family of distributions as the prior. The rest of this section covers some common conjugate priors that you will encounter throughout the first year econometrics sequence and in other areas of economics.

Normal-Normal model

The data are $X = (X_1, \ldots, X_n)$. We assume that conditional on θ, the X_i are i.i.d. with

$$X_i \sim N(\mu, \sigma^2)$$

σ^2 is fixed and assumed known. It is useful to define the **precision** as $\lambda_\sigma = 1/\sigma^2$. The parameter space is $\theta = \mathbb{R}$. Suppose we observe realizations $x = (x_1, \ldots, x_n)$. The likelihood function is then

$$\begin{aligned} f_\mu(x) &= f(x|\mu) \\ &= \Pi_{i=1}^n f(x_i|\mu) \\ &\propto \Pi_{i=1}^n \exp(-\frac{1}{2}\lambda_\sigma(x_i - \mu)^2) \\ &\propto \exp(-\frac{1}{2}\lambda_\sigma \sum_{i=1}^n (x_i - \mu)^2) \end{aligned}$$

The prior distribution for μ is also normal. We assume that

$$\mu \sim N(m, \tau^2).$$

Again, it is useful to define the **prior precision** as $\lambda_\tau = 1/\tau^2$. We have that

$$\pi(\mu) \propto \exp(-\frac{1}{2}\lambda_\tau(\mu - m)^2)$$

The posterior distribution is given by Bayes' rule. We have that[6]

$$\begin{aligned}
\pi(\mu|x) &\propto f_\mu(x)\pi(\mu) \\
&\propto \exp(-\frac{1}{2}\lambda_\sigma \sum_{i=1}^n (x_i-\mu)^2)\exp(-\frac{1}{2}\lambda_\tau(\mu-m)^2) \\
&\propto \exp\Big(-\frac{\lambda_\sigma}{2}\sum_{i=1}^n (x_i^2 - 2x_i\mu + \mu^2) - \frac{\lambda_\tau}{2}(\mu^2 - 2\mu m + m^2)\Big) \\
&\propto \exp\Big(-\frac{n\lambda_\sigma + \lambda_\tau}{2}\mu^2 + \frac{\lambda_\sigma \sum_{i=1}^n x_i + \lambda_\tau m}{2}\mu\Big) \\
&\propto \exp\Big(-\frac{n\lambda_\sigma + \lambda_\tau}{2}(\mu^2 - \frac{\lambda_\sigma \sum_{i=1}^n x_i + \lambda_\tau m}{n\lambda_\sigma + \sigma_\tau}\mu)\Big) \\
&\propto \exp\Big(-\frac{n\lambda_\sigma + \lambda_\tau}{2}(\mu^2 - \frac{n\lambda_\sigma \bar{x} + \lambda_\tau m}{n\lambda_\sigma + \lambda_\tau}\mu)\Big) \\
&\propto \exp\Big(-\frac{n\lambda_\sigma + \lambda_\tau}{2}(\mu^2 - \frac{n\lambda_\sigma \bar{x} + \lambda_\tau m}{n\lambda_\sigma + \lambda_\tau}\mu + (\frac{n\lambda_\sigma \bar{x} + \lambda_\tau m}{n\lambda_\sigma + \lambda_\tau})^2)\Big) \\
&\propto \exp\Big(-\frac{n\lambda_\sigma + \lambda_\tau}{2}(\mu - \frac{n\lambda_\sigma \bar{x} + \lambda_\tau m}{n\lambda_\sigma + \lambda_\tau})^2\Big)
\end{aligned}$$

And so, we have shown that the posterior distribution is also normally distributed with posterior mean

$$E[\mu|x] = \frac{n\lambda_\sigma \bar{x} + \lambda_\tau m}{n\lambda_\sigma + \lambda_\tau}$$

and posterior precision

$$\bar{\lambda}_\tau = n\lambda_\sigma + \lambda_\tau$$

What is the interpretation of the posterior mean? It is a weighted average of the sample mean and the prior mean in which the weights are the precisions. Therefore, if λ_τ is large and the prior has a low variance, the prior mean receives a larger weight. Alternatively, we can interpret this as "shrinking" the posterior mean towards the prior.[7]

We could have derived this using our results from the multivariate normal distribution. We have that

$$X|\mu \sim N(\mu, \sigma^2 I_n).$$

You can show that the marginal distribution of X is given

$$X \sim N(m, (\sigma^2 + \tau^2)I_n)$$

and that the joint distribution of X, μ is given by

$$\begin{pmatrix} X \\ \mu \end{pmatrix} \sim N(\begin{pmatrix} m \\ m \end{pmatrix}, \begin{pmatrix} (\sigma^2+\tau^2)I_n & \tau^2 l \\ \tau^2 l' & \tau^2 \end{pmatrix}$$

[6] The key to making this calculation easy is to remember that the posterior density is a function of μ. x, m are constants and so, we can drop them along the way.

[7] If you are familiar with machine learning jargon, you can write down Bayesian model similar to this to motivate Ridge regression.

where l is a $n \times 1$ vector of ones. It then follows that

$$\mu | X = x \sim N(m + \frac{\tau^2}{\sigma^2 + \tau^2} l' I_n (x - m), \tau^2 - \tau^2 (\sigma^2 + \tau^2)^{-1} \tau^2 l' l).$$

This is the same as the result we derived.

Exercise 0.1. *Use the properties of the multivariate normal distribution to derive the posterior distribution.*

Beta-Bernoulli model

The data are $X = (X_1, \ldots, X_n)$. We assume that conditional on θ, the X_i are i.i.d with

$$P(X_i = 1|\theta) = \theta, \quad P(X_i = 0|\theta) = 1 - \theta.$$

The parameter space is $\Theta = [0, 1]$. Suppose we observe realizations $x = (x_1, \ldots, x_n)$. The likelihood function is then

$$\begin{aligned} f_\theta(x) &= f(x|\theta) \\ &= P(X = x|\theta) \\ &= \Pi_{i=1}^n P(X_i = x_i|\theta) \\ &= \Pi_{i=1}^n \theta^{y_i} (1 - \theta)^{1 - y_i} \\ &= \theta^{n_1} (1 - \theta)^{n_0} \end{aligned}$$

where $n_1 = \sum_{i=1}^n y_i$ and $n_0 = \sum_{i=1}^n (1 - y_i) = n - n_1$. The prior distribution is a **beta distribution** with parameters $a, b > 0$. Its support is over $[0, 1]$ with density

$$\pi(\theta) \propto \theta^{a-1} (1 - \theta)^{b-1}.$$

The prior mean and variance are

$$E[\theta] = \frac{a}{a+b}, \quad V(\theta) = \frac{a}{a+b} \frac{b}{a+b} \frac{1}{a+b+1}.$$

The posterior distribution is given by Bayes' rule. We have that

$$\begin{aligned} \pi(\theta|x) &\propto f_\theta(x) \pi(\theta) \\ &\propto \theta^{a+n_1-1} (1 - \theta)^{b+n_0-1} \end{aligned}$$

Therefore, the posterior distribution is also a beta distribution and has parameters $a + n_1, b + n_0$. The posterior mean is then

$$E[\theta|x] = \frac{a + n_1}{a + b + n} = \lambda \frac{n_1}{n} + (1 - \lambda) \frac{a}{a+b}$$

where $\lambda = \frac{n}{a+b+n}$. In other words, the posterior mean is a convex combination of the sample mean n_1/n and the prior mean $a/(a+b)$.

Note that if $a + b$ is small relative to n, then most of the weight is placed on the sample mean.

What happens as $a, b \to 0$? The prior becomes

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

This is not a probability density as it integrates to ∞ over $[0,1]$. We call this an **improper prior**. However, the associated posterior distribution is well-defined. In this case, the posterior distribution is again a beta distribution but with parameters, n_1, n_0. For this improper prior,

$$E[\theta|x] = \frac{n_1}{n} = \bar{x}$$

That is, the posterior conditional expectation coincides with the sample average (i.e. the frequentist estimate of θ).

Multinomial-Dirichlet model

The data are $X = (X_1, \ldots, X_n)$. Each X_i takes on a discrete set of values $\{\alpha_j : j = 1, \ldots, J\}$. We assume that conditional on θ, the X_i are i.i.d. with

$$P(X_i = \alpha_j|\theta) = \theta_j \quad for j = 1, \ldots, J.$$

The parameter space is the unit simplex on $\mathbb{R}^J$ with

$$\Theta = \{\theta \in \mathbb{R}^J : \theta_j \geq 0, \sum_{j=1}^{J} \theta_j = 1\}.$$

We observe realizations $x = (x_1, \ldots, x_n)$. The likelihood function is

$$\begin{aligned} f_\theta(x) &= f(x|\theta) \\ &= \Pi_{i=1}^n P(X_i = x_i|\theta) \\ &= \Pi_{i=1}^n \Pi_{j=1}^J \theta_j^{1(x_i=\alpha_j)} \\ &= \Pi_{j=1}^J \theta_j^{n_j} \end{aligned}$$

where $n_j = \sum_{i=1}^n 1(x_i = \alpha_j)$ for $j = 1, \ldots, J$.

The prior distribution is a **Dirichlet distribution** with parameters $a_1, \ldots, a_J > 0$. The Dirichlet distribution is a generalization of the beta distribution. Its support is over the unit simplex in $\mathbb{R}^J$ and has density

$$\pi(u_1, \ldots, u_J) \propto \Pi_{j=1}^J u_j^{a_j - 1}.$$

The posterior distribution is then given by Bayes' rule. We have that

$$\begin{aligned} \pi(\theta|x) &\propto f_\theta(x)\pi(\theta) \\ &\propto \Pi_{j=1}^J \theta_j^{a_j+n_j-1}. \end{aligned}$$

The posterior distribution is also Dirichlet but with parameters $a_j + n_j$ for $j = 1, \ldots, J$. As in the Beta-Bernoulli model, we can consider the improper prior with $a_j \to 0$ for each $j = 1, \ldots, J$. With this improper prior, the posterior distribution remains Dirichlet and has parameters $n_1, \ldots, n_J$.

It turns out that we can represent the Dirichlet distribution using independent gamma distributed random variables. This is very useful in deriving several properties of the Dirichlet distribution and in simulations. The **gamma distribution** with **shape parameter** $a > 0$ and **scale parameter** $b > 0$ has density

$$g(u) \propto u^{a-1} \exp(-u/b)$$

with support over $u > 0$. The gamma distribution has the useful property that if Q_j are independent gamma distributed with parameters (a_j, b), then their sum $\sum_j Q_j \sim gamma(\sum_j a_j, b)$.

Suppose $Q_j \sim gamma(a_j, 1)$ for $j = 1, \ldots, J$ and $Q_1, \ldots, Q_j$ are independent. Let $S = \sum_{j=1}^{J} Q_j$. Define

$$R = (Q_1/S, \ldots, Q_J/S)$$

and one can show that $R \sim Dirichlet(a_1, \ldots, a_J)$. For the case $J = 2$, we have that

$$R = (Q_1/(Q_1 + Q_2), Q_2/(Q_1 + Q_2))$$

where $Q_1/(Q_1 + Q_2 \sim beta(a_1, a_2)$.

For the posterior distribution of θ, we can represent it as

$$\theta | x \sim \Big(\frac{Q_1}{\sum_{j=1}^{J} Q_j}, \ldots, \frac{Q_J}{\sum_{j=1}^{J} Q_j} \Big)$$

where each Q_j are mutually independent gamma random variables with parameters $a = n_j + a_j - 1, b = 1$. So a component θ_j can be represented as

$$\theta_j | x \sim \frac{Q_j}{Q_j + \sum_{k \neq j} Q_K}$$

and so, $\theta_j \sim beta(n_j + a_j, \sum_{k \neq j} n_k + a_k)$.

Exchangeability and de Finetti's Theorem

So far, we have assumed that there is some prior distribution π over θ and that conditional on θ, the observed data are i.i.d. de Finetti's Theorem, also known as the Representation Theorem, justifies this set-up. de Finetti's Theorem and related generalizations show that if a sequence of random variables $X_1, \ldots, X_n$ are **exchangeable**, then there exists a parameter θ and a prior distribution π for θ such that the elements of the sequence are i.i.d. conditional on θ. This is a powerful result.

Definition 0.1. *A finite sequence of random variables $X_1, \ldots, X_n$ is **exchangeable** if its joint distribution $F(\cdot)$ satisfies*

$$F(x_1, \ldots, x_n) = F(x_{p(1)}, \ldots, x_{p(n)})$$

*for all realizations $(x_1, \ldots, x_n)$ and all permutations p of $\{1, \ldots, n\}$. Any infinte sequence of random variables is **exchangeable** if every finite subsequence is exchangeable.*

Remark 0.1. *Note that exchangeability is a weaker condition than i.i.d. If $X_1, \ldots, X_n$ are i.i.d., then the sequence is exchangeable. However, the elements of an exchangeable sequence are identically distributed but need not be independent.*

Example 0.3. *Polya's Urn*

Consider an urn with b black balls and w white balls. Draw a ball and note its color. Replace the ball in the urn and add a additional balls of the same color to the urn. Let $X_i = 1$ if the i-th drawn ball is black and $X_i = 0$ if it is white. The sequence $X_1, X_2, \ldots$ is exchangeable. For example,

$$\begin{aligned} f(1,1,0,1) &= \frac{b}{b+w}\frac{b+a}{b+w+a}\frac{w}{b+w+2a}\frac{b+2a}{b+w+3a} \\ &= \frac{b}{b+w}\frac{w}{b+w+a}\frac{b+a}{b+w+2a}\frac{b+2a}{b+w+3a} \\ &= f(1,0,1,1) \end{aligned}$$

Theorem 0.1. *de Finetti's Theorem*

Let $X_1, X_2, \ldots$ be an exchangeable sequence. Then, there exists a random variable Θ with cdf $F_\Theta(\cdot)$ such that

$$f(x_1, \ldots, x_n) = \int_0^1 \theta^{n_1}(1-\theta)^{n-n_1} dF_\Theta(\theta)$$

where

$$n_1 = \sum_{i=1}^n x_i$$

and

$$\Theta = \lim_{n\to\infty} \frac{1}{n}\sum_{i=1}^n X_i$$

with $F_\Theta(\theta) = \lim_{n\to\infty} P(\frac{1}{n}\sum_{i=1}^n X_i \leq \theta)$.

It is as if the sequence of Bernoulli random variables are i.i.d. conditional on Θ. Moreover, the distribution of Θ is determined by the limiting distribution of the sample frequency. We can view F_Θ as a prior distribution. How do we interpret this? It provides us with a way to think about the prior distribution. By de Finetti's Theorem, the prior distribution F_Θ is determined by the limiting distribution of the sample frequency and so, we can view it as reflecting the researcher's subjective beliefs about the long-run frequency. de Finetti's Theorem generalizes in many ways. See, for instance, Diaconis (1988) for more results.

References

Diaconis, P. (1988). Recent Progress on de Finetti's Notions of Exchangeability. *Bayesian Statistics*.

Efron, Bradley and Trevor Hastie. (2016). *Computer Age Statistical Inference*.

Gelman et al. (2013). *Bayesian Data Analysis*.

Geweke, John. (2005). *Contemporary Bayesian Econometrics and Statistics*.

Lauritzen, Steffen. Exchangeability and de Finetti's Theorem. Lecture notes 2007.

Poirier, Dale. (2011). Exchangeability, Representation Theorem and Subjectivity. *The Oxford Handbook of Bayesian Econometrics*.