Bayesian Inference

Harvard Math Camp - Econometrics

Ashesh Rambachan

Summer 2018

Outline

Outline

Outline

Statistical Inference

Observe data x_i for $i = 1, \ldots, n$.

- Assume the data from from a random experiment, modeled by r.v. X with support $\mathcal{X}$.
- $\{x_i\}_{i=1}^n$ are realizations of X.
- Wish to use the data to learn something about $F_X(x)$

A **statistical model** is a set of probability distributions indexed by a parameter set.

$$\mathcal{F} = \{P_\theta(x) : x \in \mathcal{X}, \theta \in \Theta\}$$

- **Parametric** if P can be indexed with a finite dimensional parameter set. Otherwise, **non-parametric**.

Observe $\{x_i\}_{i=1}^n$ and wish to make inferences about θ.

Statistical Models: Examples

Example:the set of normal distributions with variance equal to one. Then, $\mathcal{X} = \mathbb{R}$, $\Theta = \mathbb{R}$ and

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

Wish to learn about θ.

Outline

Frequentists vs. Bayesians

Suppose we have a "good" statistical model.

$$F_X(x) \in \mathcal{F}$$

and 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.

- How should we model an unknown θ^* and how does that choice affect how inference should be conducted.

Frequentists

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 $F_{\theta^*}(x)$.

The random experiment is:

1. Nature draws the data $x_1, \ldots, x_n$ from $F_{\theta^*}(x)$.
2. We observe $x_1, \ldots, x_n$ and plugs them into our estimator, $\hat{\theta}(\cdot)$. Our estimate is $\hat{\theta}(x_1, \ldots, x_n)$.

Frequentists

Freqentists engage in the following thought experiment:

- Repeat the experiment many times. Each time, we obtain new data $x_1^b, \ldots, x_n^b$ and construct a new estimate, $\hat{\theta}(x_1^b, \ldots, x_n^b) = \hat{\theta}^b$.
- What properties will the **sampling distribution** of my estimator have?
  - As $n \to \infty$, what properties will the distribution of of my estimator have?

Frequentists focuses on the *behavior* of estimators in a **repeated random experiment**, where we want to understand the properties of $\hat{\theta}(\cdot)$ under the sampling distribution of the data.

Bayesians

Bayesians, model the unknown θ^* as a random variable itself, with its own distriution, $\Pi(\theta)$. This is the **prior distribution**.

- The prior encodes *prior information* about the parameter θ available prior to observing the data. This may come from prior experiments, observational studies or economic theory.

Bayesians

The random experiment then has an extra step:

1. Nature draws θ^* from the prior, $\Pi(\theta)$. This is unobserved.
2. Nature draws realizations $x_1, \ldots, x_n$ from the distribution $F_{\theta^*}(x)$. These are the data.
3. We observes $x_1, \ldots, x_n$ and plugs them into our estimator, $\hat{\theta}(\cdot)$. Her estimate is $\hat{\theta}(x_1, \ldots, x_n)$.

Bayesians

What is the point of the prior? **Bayes' rule**.

- Provides a logically consistent rule for combining prior information with the observed data.
- $x = (x_1, \ldots, x_n)$ and $f_\theta(x)$ is the density associated with distribution $F_\theta(x)$ and $\pi(\theta)$ is defined analogously.

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

  - **marginal density** of X: $f(x) = \int_\Theta f_\theta(x)\pi(\theta)d\theta$
  - **likelihood function**: $f_\theta(x)$
  - **posterior density**: $\pi(\theta|x)$

The posterior distribution of $\theta|x$ is the central object of interest in Bayesian inference.

Bayesians: Brief Aside

You will often see Bayes' rule written as

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

In English Bayes' rule says, "the posterior is proportional to the likelihood times the prior."

Bayesians

Uses the posterior distribution to make inferences about θ.

- E.g. the "posterior expectation of θ given the data x"

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

  is a common object of interest.
- Could also compute $Med(\theta|X), P(\theta < \tilde{\theta}|X)$ and so on.

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*.

Bayesians

Completely swept under the rug the very important question: How do we choose a prior distribution?

- Short answer: it's not easy! Requires a lot of careful thought.
- We'll pick this issue up at times in Ec 2120.
- If interested, check out Kasy & Fessley (2018) - "how should economic theory guide the choice of priors?"

Outline

Conjugate Priors

Once we have a prior distribution and a likelihood function, the only computational step is to use Bayes' rule.

- Sounds simple... But this can often be a mess.
- Lots of Bayesian statistics focues on doing this in a computationally feasible manner - MCMC, Variational Inference.

Important tool in bayesian inference: **conjugate priors**.

- Prior distribution is **conjugate** for a given likelihood function if the associated posterior distribution is in the same family of distributions as the prior.

We'll cover three useful conjugate priors that you will encounter.

Outline

The data

The data are $X = (X_1, \ldots, X_n)$.Conditional on θ, X_i are i.i.d. with

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

- σ^2 is fixed and assumed known.
- Define the **precision** as $\lambda_\sigma = 1/\sigma^2$.
- The parameter space is $\theta = \mathbb{R}$.

We observe realizations $x = (x_1, \ldots, x_n)$.

The likelihood

The likelihood function is

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

The prior

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

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

- Useful to define the **prior precision** as $\lambda_\tau = 1/\tau^2$.

So,

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

The posterior

The posterior distribution is given by Bayes' rule. This is a pain in the butt but the result is really nice.

Takes a deep breath

The posterior

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

The posterior

So,

$$\pi(\mu|x) \propto \exp\Big(-\frac{n\lambda_\sigma + \lambda_\tau}{2}(\mu - \frac{n\lambda_\sigma \bar{x} + \lambda_\tau m}{n\lambda_\sigma + \lambda_\tau})^2\Big)$$

and

$$\mu|x \sim N(\frac{n\lambda_\sigma \bar{x} + \lambda_\tau m}{n\lambda_\sigma + \lambda_\tau}, n\lambda_\sigma + \lambda_\tau).$$

The posterior

As I said: This was a pain in the butt. Is there an easier way?

Yes! Use our results for the multivariate normal distribution.

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

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

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

where I is a $n \times 1$ vector of ones.

The posterior

It then follows that

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

Exactly as before!

The posterior

Posterior mean:

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

Posterior precision:

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

Interpretation:

- Posterior mean is a weighted average of the sample mean and the prior mean in which the weights are the precisions.
- If λ_τ is large and the prior has a low variance, the prior mean receives a larger weight.
- "Shrinking" the posterior mean towards the prior

Machine learning aside

Machine learning aside:

$$Y_i = X_i\beta + \epsilon_i, \quad \beta|X \sim N(0,\Omega) \quad \epsilon_i|X,\beta \sim N(0,\sigma^2) i.i.d.$$

Joint likelihood of Y, β gives a ridge-type objective

$$\propto -\frac{1}{2\sigma^2}\sum_i (Y_i - \beta X_i)^2 - \frac{1}{2}\beta'\Omega\beta$$

Maximum a posteriori estimator: Ridge regression.

Can similarly motivate lasso using this Bayesian approach.

Outline

The data

D are $X = (X_1, \ldots, X_n)$.

- 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 $\Theta = [0, 1]$.

Observe realizations $x = (x_1, \ldots, x_n)$.

The likelihood

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

The prior distribution is a **beta distribution** with parameters $a, b > 0$.

- Support is over $[0, 1]$ with density

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

- 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

The posterior distribution is given by Bayes' rule.

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

The posterior distribution is also a beta distribution with parameters $a + n_1, b + n_0$.

The posterior

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 $\lambda = \frac{n}{a+b+n}$.

- The posterior mean is a convex combination of the sample mean n_1/n and the prior mean $a/(a+b)$.
- If $a + b$ is small relative to n, then most of the weight is placed on the sample mean.

Improper priors

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

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

Not a probability density as it integrates to ∞ over $[0, 1]$. Call this an **improper prior**.

But, the associated posterior distribution is well-defined.

- The posterior distribution is again a beta distribution but with parameters, n_1, n_0.
- Note
$$E[\theta|x] = \frac{n_1}{n} = \bar{x}$$
That is, the posterior conditional expectation coincides with the sample average

Outline

The data

Data are $X = (X_1, \ldots, X_n)$.

- Each X_i takes on discrete set of values $\{\alpha_j : j = 1, \ldots, J\}$.
- Conditional on θ, the X_i are i.i.d. with

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

- Parameter space is the unit simplex on $\mathbb{R}^J$ with

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

Observe realizations $x = (x_1, \ldots, x_n)$.

The likelihood

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

Prior distribution is a **Dirichlet distribution** with parameters $a_1, \ldots, a_J > 0$.

- Generalizes a generalization of the beta distribution.
- Its support is over the unit simplex in $\mathbb{R}^J$.
- Has density

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

The posterior

The posterior distribution is given by Bayes' rule.

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

The posterior distribution is also Dirichlet but with parameters $a_j + n_j$ for $j = 1, \ldots, J$.

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 $n_1, \ldots, n_J$.

Representing the posterior

Fact: we can represent the Dirichlet distribution using independent **gamma distributed** random variables.

- 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 $u > 0$.

- Useful property that if Q_j are independent gamma distributed with parameters (a_j, b), then

$$\sum_j Q_j \sim gamma(\sum_j a_j, b).$$

Representing the posterior

Suppose $Q_j \sim gamma(a_j, 1)$ for $j = 1, \ldots, J$ and $Q_1, \ldots, Q_j$ are independent. Let

$$S = \sum_{j=1}^{J} Q_j$$

and define

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

- Can show that $R \sim Dirichlet(a_1, \ldots, a_J)$.
- $J = 2$:

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

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

Representing the posterior

So, can represent the posterior distribution of θ as

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

where each Q_j are mutually independent gamma random variables with parameters $a = n_j + a_j - 1, b = 1$.

Component θ_j can be represented as

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

and so,

$$\theta_j|x \sim beta(n_j + a_j, \sum_{k \neq j} n_k + a_k)$$

Outline

Exchangeability and de Finetti's Theorem

So far, 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**, provides a justification.

- 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.

Exchangeability

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 infinite sequence of random variables is **exchangeable** if every finite subsequence is exchangeable.

Exchangeability

exchangeability is a weaker condition than i.i.d.

- If $X_1, \ldots, X_n$ are i.i.d., then the sequence is exchangeable.
- Elements of an exchangeable sequence are identically distributed but need not be independent.

Example: Polya's Urn

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,

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

de Finetti's Theorem: Binary Case

Let $X_1, X_2, \ldots$ be an exchangeable sequence of random variables that take on the values $\{0, 1\}$. 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 $F_\Theta(\theta) = \lim_{n\to\infty} P(\frac{1}{n}\sum_{i=1}^{n} X_i \leq \theta)$.

Interpretation

It is as if the sequence of Bernoulli random variables are i.i.d. conditional on Θ.

The distribution of Θ is determined by the limiting distribution of the sample frequency. We can view F_Θ as a prior distribution.

- One 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.