Econ 2120: Section 4

Bayesian Inference

Ashesh Rambachan

Fall 2018

Outline

Setting it all up

Discrete-Dirichlet Model

- Representing the Posterior
- Predictive Distribution

Dogmatic Priors

The Bayesian Bootstrap

- The Best Linear Predictor
- The Bayesian Bootstrap

Outline

Setting it all up

Discrete-Dirichlet Model

- Representing the Posterior
- Predictive Distribution

Dogmatic Priors

The Bayesian Bootstrap

- The Best Linear Predictor
- The Bayesian Bootstrap

Recall: Inference Problem

The data are a realization of some random vector

$$D = (Y_1, \ldots, Y_n, Z_1, \ldots, Z_n),$$

where Y_i is a scalar outcome and Z_i is a vector of predictors.

Also write $D = (D_1, \ldots, D_n)$ with $D_i = (Y_i, Z_i)$.

Assume D is drawn from some distribution (unknown).

Specify set of distributions that contains D

$$D \sim P_\theta, \quad \text{for some} \quad \theta \in \Theta.$$

Θ is the **parameter space**.

Assume D_i i.i.d. $\implies$ can factor the joint distribution of D.

Recall: Probability models

The **probability model** is a map from the parameter space to a set of distributions

$$\theta \to P_\theta$$

Last section: We only assumed the data are i.i.d. But paid a price – only able to do inference using asymptotic approximations (which can be very poor in finite samples).

Bayesian Perspective: Specify a probability measure over Θ and exploit Bayes' Rule to perform inference – inference becomes conditional on the data.

> Let Π be a probability measure over Θ. This is **prior distribution**.

The probability model

So, the **state space** is

$$S = \Theta \times \mathcal{D} = \{(\theta, d) : \theta \in \Theta, d \in \mathcal{D}\}.$$

P_θ is a conditional distributional, $D|\Theta = \theta$.

We can write the joint distribution.

Notation: Θ is a random variable, D is a random variable.

$$P(\theta \in B, D \in A) = \int_B P_\theta(A)\pi(\theta)d\theta.$$

Bayes' Rule:

$$P(\theta \in B|D \in A) = \int_B P_\theta(A)\pi(\theta)d\theta/\int_\Theta P_\theta(A)\pi(\theta)d\theta.$$

Bayes' Rule

We will perform inference using the **posterior distribution** of $\theta|D = d$.

> This encodes all our uncertainty about θ given that we observed the data $D = d$.

Typically write

$$\pi(\theta|d) \propto f(d|\theta)\pi(\theta),$$

where we omit a constant that makes the posterior integrate to one ($f(d)$).

Outline

- Setting it all up
- Discrete-Dirichlet Model
  - Representing the Posterior
  - Predictive Distribution
- Dogmatic Priors
- The Bayesian Bootstrap
  - The Best Linear Predictor
  - The Bayesian Bootstrap

The data

Data are $D = (D_1, \ldots, D_n)$.

Each D_i takes on discrete set of values $\{\alpha_j : j = 1, \ldots, J\}$.

Conditional on θ, the D_i are i.i.d. with

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

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 $d = (d_1, \ldots, d_n)$.

The data

The values of D_i may be vectors and we will apply these results to inference for the linear predictor.

Think of

$$D_i = \begin{pmatrix} X_i \\ Y_i \end{pmatrix}, \quad \alpha_j = \begin{pmatrix} \alpha_{xj} \\ \alpha_y j \end{pmatrix}.$$

The likelihood

The likelihood function is

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

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

n_j for $j = 1, \ldots, J$ are **sufficient statistics** for the likelihood.

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

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

for $u_j > 0, \sum_{j=1}^{J} u_j = 1$.

The posterior

The posterior distribution is given by Bayes' rule.

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

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

Can consider the improper prior with $a_j \to 0$ for each $j = 1, \dots, J$.

> With this improper prior, the posterior distribution remains Dirichlet and has parameters $n_1, \dots, n_J$.

Outline

Setting it all up

Discrete-Dirichlet Model

- Representing the Posterior
- Predictive Distribution

Dogmatic Priors

The Bayesian Bootstrap

- The Best Linear Predictor
- The Bayesian Bootstrap

Gamma Random Variables

Recall: We can represent the Dirichlet distribution using Gamma-distributed random variables.

$Q \sim Gamma(a, l)$ with parameters $a > 0, l > 0$ if

$$f_Q(q) = \frac{l^a x^{a-1} e^{-lx}}{\Gamma(a)}$$

for $q > 0$ and 0 otherwise.

For $a = 1$, $Gamma(1, l) \sim Exponential(l)$.

You can show that the sum of i.i.d. exponential random variables is a Gamma distribution

$$\sum_{i=1}^{n} Exponential(l) \sim Gamma(n, l).$$

Properties of Gamma Random Variables

Fact 1: Suppose $X \sim Gamma(a_1, l)$, $Y \sim Gamma(a_2, l)$ with $X \perp Y$. Consider

$$U = X + Y, V = X/(X + Y).$$

$U \perp V$ and

$$U \sim Gamma(a_1 + a_2, l), V \sim Beta(a_1, a_2).$$

Properties of Gamma Random Variables

Fact 2: Suppose $X_1, X_2 \ldots$ are independent gamma random variables with $X_i \sim Gamma(a_i, l)$. Then,

$$\begin{aligned} X_1/(X_1+X_2) &\sim Beta(a_1, a_2) \\ \frac{X_1+X_2}{X_1+X_2+X_3} &\sim Beta(a_1+a_2, a_3) \\ &\vdots \\ \frac{X_1+\ldots+X_{K-1}}{X_1+\ldots+X_K} &\sim Beta(\alpha_1+\ldots+\alpha_{K-1}, \alpha_K) \end{aligned}$$

are independent.

Gamma Random Variables and the Dirichlet Distribution

Fact 3: Let $Q_j \sim Gamma(a_j, 1)$. If $Q_1, \ldots, Q_J$ are independent then

$$(Q_1/\sum_{j=1}^{J} Q_j, \ldots, Q_J/\sum_{j=1}^{J} Q_j) \sim Dirichlet(a_1, \ldots, a_J).$$

Fact 4: The j-th entry is

$$\frac{Q_j}{Q_1 + \ldots + Q_J}$$

. Can show that

$$\frac{Q_j}{Q_1 + \ldots + Q_J} \sim Beta(a_j, \sum_{k \neq j} a_k)$$

(by Fact 2).

Representing the posterior

So, we can represent the posterior for θ as

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

where $Q_j \sim Gamma(n_j + a_j, 1)$ for $j = 1, \ldots J$.

Moreover, each component θ_j has the representation

$$\theta_j|d \sim \frac{Q_j}{Q_j + \sum_{k \neq j} Q_j} = beta(n_j + a_j, \sum_{k \neq j} n_k + a_k).$$

Representing the Posterior

So,

$$E[\theta_j|d] = \frac{n_j + a_j}{\sum_{k=1}^{J} n_k + a_k}.$$

Can similarly write $V(\theta_j|d)$ using formulas from before.

Outline

Setting it all up

Discrete-Dirichlet Model

- Representing the Posterior
- Predictive Distribution

Dogmatic Priors

The Bayesian Bootstrap

- The Best Linear Predictor
- The Bayesian Bootstrap

Predictive distribution

Let's see what we can do with our posterior $\theta|d$.

Let's use it for prediction.

Suppose there is a new observation D_{n+1}. We want to predict it.

Our object of interest is $\gamma = P(D_{n+1} \in A|d)$

Note that

$$P(D_{n+1} \in A|\theta) = \sum_{j \in C} \theta_j$$

where $C = \{j : \alpha_j \in A\}$.

Predictive distribution

$P(D_{n+1} \in A|\theta) = \gamma(\theta)$ is just a function of θ. We derive its posterior distribution. That is,

$$\gamma(\theta)|d \sim ?$$

Turns out to be simple. Use the special case of $J = 2$ from earlier.

$$\theta_j|d \sim \frac{Q_j}{Q_j + \sum_{k \neq j} Q_j} \implies \sum_{j \in C} \theta_j \sim \frac{\sum_{j \in C} Q_j}{\sum_{k=1}^{J} Q_j}$$

where $\sum_{j \in C} Q_j \sim Gamma(\sum_{j \in C} n_j + a_j)$,
$\sum_{j \notin C} Q_j \sim Gamma(\sum_{j \notin C} n_j + a_j)$.

Predictive distribution

So,

$$\gamma(\theta)|d \sim \frac{\sum_{j\in C} Q_j}{\sum_{j=1}^{J} Q_j} \sim Beta(\sum_{j\in C} n_j + a_j, \sum_{j\notin C} n_j + a_j).$$

The conditional distribution of D_{n+1} given $(D_1, \ldots, D_n) = d$ is the **predictive distribution**.

> Notice that θ has been integrated out using the posterior distribution.
>
> We can use iterated expectations for this!

Predictive distribution

We have

$$\begin{aligned}
P(D_{n+1} \in A|d) &= E[1(D_{n+1} \in A)|d] \\
&= E[E[1(D_{n+1} \in A)|\theta, d]|d] \\
&= E[E[1(D_{n+1} \in A)|\theta]|d] \\
&= E[\gamma(\theta)|d] \\
&= \frac{\sum_{j \in C} n_j + a_j}{\sum_{j=1}^{J} n_j + a_j}.
\end{aligned}$$

Outline

Setting it all up

Discrete-Dirichlet Model

- Representing the Posterior
- Predictive Distribution

Dogmatic Priors

The Bayesian Bootstrap

- The Best Linear Predictor
- The Bayesian Bootstrap

Approximating continuous distributions

We can use the discrete-dirichlet model to approximate continuous distributions by letting $J \to \infty$.

In doing so, we need to be careful that the prior does not become **dogmatic**.

We want to ensure that the prior is “responsive to the data.”

We want to ensure that the posterior doesn’t just return the prior... that we actually learning from the data.

Illustrate what can go wrong with the predictive distribution as $J \to \infty$ if we aren’t careful.

Predictive distribution: $J \to \infty$

Recall:

$$\gamma = P(D_{n+1} \in A|d) = E[\gamma|w] = \frac{\sum_{j \in C} n_j + a_j}{\sum_{j=1}^{J} n_j + a_j}.$$

Suppose that $a_j = \epsilon > 0$ fixed for all j and let $J \to \infty$ while the data $d = (d_1, \ldots, d_n)$ is fixed.

Assume that

$$\frac{1}{J} \sum_{j \in C} 1 \to r$$

as $J \to \infty$. That is, the fraction of support points in A approaches r in the limit.

Predictive distribution: $J \to \infty$

Claim: Then,

$$P(D_{n+1} \in A|d) \to r.$$

Why? The prior is dogmatic for γ. The prior for γ is

$$\gamma \sim Beta(\sum_{j \in C} a_j, \sum_{j \notin C} a_j).$$

So,

$$E[\gamma] = \frac{\sum_{j \in C} a_j}{\sum_{j=1}^{J} a_j} = \frac{1}{J} \sum_{j \in C} 1 \to r,$$

$$V(\gamma) = \frac{E[\gamma](1 - E[\gamma])}{1 + \sum_{j=1}^{J} a_j} = \frac{E[\gamma](1 - E[\gamma])}{1 + \epsilon J} \to 0$$

As $J \to \infty$, for fixed ϵ, the prior distribution becomes concentrated around r.

What to do?

That's really bad! It means that even as you see data, your prior for γ is unchanging! You are NOT learning from your data.

> The prior over θ is **dogmatic** for γ. The prior on γ that π has all of its mass on r as the support points gets arbitrarily large.

To avoid this, we let $a_j \to 0$ for all j as $J \to \infty$.

> In the limit, this produces the improper Dirichlet distribution.
>
> So, if $n_j \geq 1$ for all j, the posterior will be a proper Dirichlet. But if we want J to approximate continuous distributions, we'll allow zero counts.

What to do?

So, if the count $n_k = 0$, the limiting posterior of θ_k as $J \to 0$ will become concentrated around 0.

> Why? $\theta_k | w \sim Beta(0, \sum_{j \neq k} n_j)$ and $E[\theta_k | w] = 0$. Since θ_k can only take values between $[0, 1]$, its expectation being zero implies that $P(\theta_k > 0 | w) = 0$ by Markov's Inequality.

$\implies$ Support points with $n_k = 0$ drop out of the posterior distribution, which is concentrated around support points with $n_j > 0$.

Why?

Gives us a way to take the tools we have (Discrete-Dirichlet) and apply it to continuous data.

- We use a limiting dirichlet prior and the posterior becomes concentrated around only support points on which we observe data.
- We'll next apply this to the linear predictor.

Outline

Setting it all up

Discrete-Dirichlet Model

- Representing the Posterior
- Predictive Distribution

Dogmatic Priors

The Bayesian Bootstrap

- The Best Linear Predictor
- The Bayesian Bootstrap

Outline

Setting it all up

Discrete-Dirichlet Model

- Representing the Posterior
- Predictive Distribution

Dogmatic Priors

The Bayesian Bootstrap

- The Best Linear Predictor
- The Bayesian Bootstrap

The Best Linear Predictor

Goal: Take framework we just set-up and apply it to the best linear predictor.

An observation is $D_i' = (X_i', Y_i)$, where X_i' is $K \times 1$ vector and Y_i is a scalar.

D_i takes on J possible values denoted

$$\alpha_j = \begin{pmatrix} \alpha_{xj} \\ \alpha_{yj} \end{pmatrix} \; j = 1, \ldots, J.$$

Assume conditional on θ, D_i are i.i.d. with

$$P(D_i = \alpha_j | \theta) = \theta_j.$$

The Best Linear Predictor

Given θ, the coefficients on the best linear predictor are defined as

$$\beta(\theta) = \arg\min_{c} E[(Y_i - X_i' c)^2 | \theta].$$

Equivalently,

$$\begin{aligned}\beta(\theta) &= E[X_i X_i' | \theta]^{-1} E[X_i Y_i | \theta] \\ &= \Big(\frac{1}{J} \sum_{j=1}^{J} \alpha_{xj} \alpha_{xj}' \theta_j \Big)^{-1} \Big(\frac{1}{J} \sum_{j=1}^{J} \alpha_{xj} \alpha_{yj} \theta_j \Big)\end{aligned}$$

KEY: β is a function of θ. The randomness of θ makes β random.

GOAL: Characterize the distribution of β using the posterior distribution of θ.

Distribution of β under posterior of θ

We know that

$$\theta|d \sim dirichlet(n_1 + a_1, \ldots, n_J + a_J)$$

and we can represent this as

$$\theta|d \sim (Q_1, \ldots, Q_J)/\sum_{j=1}^{J} Q_J,$$

where $Q_i \sim Gamma(a_j + n_j, 1)$.

Distribution of β under posterior of θ

Algorithm: Brute Force

```
For $b = 1, \ldots, B$:
    (1) Draw $Q_j^b \sim Gamma(a_j + n_j, 1)$ independently for $j = 1, \ldots, J$.
    (2) Form $\theta^b = (Q_1^b, \ldots, Q_J^b) / \sum_{j=1}^J Q_j^b$.
    (3) Compute $\beta^b = \beta(\theta^b)$ using

        $$\beta(\theta^b) = \Big(\frac{1}{J}\sum_{j=1}^{J} \alpha_{xj}\alpha'_{xj}\theta_j^b\Big)^{-1}\Big(\frac{1}{J}\sum_{j=1}^{J} \alpha_{xj}\alpha_{yj}\theta_j^b\Big)$$

    Store $\beta^b$.
```

The resulting values $\beta^1, \ldots, \beta^B$ are draws from the posterior distribution $\beta | d$.

Concerns?

This algorithm was for fixed J. What if X, Y are continuous?

> The **Bayesian Bootstrap** will help us in this case. It is just an computational tool that applies our results about limiting Dirichlet priors.

Outline

Setting it all up

Discrete-Dirichlet Model

- Representing the Posterior
- Predictive Distribution

Dogmatic Priors

The Bayesian Bootstrap

- The Best Linear Predictor
- The Bayesian Bootstrap

Allowing for continuous data

Letting $a_j \to 0$, we can represent the posterior distribution as

$$\beta | d \sim \arg\min_c \sum_{j:n_j>0} Q_j(\alpha_{yj} - \alpha_{xj}'c)^2$$

where $Q_j \sim Gamma(n_j, 1)$

Why don't we divide by the sum $\sum_j Q_j$? It is a constant that won't affect the minimization.

Can we remove the support points and write things in terms of the data? **Yes**.

The Bayesian Bootstrap

Let $V_i \sim \mathit{Gamma}(1,1) \sim \mathit{Exponential}(1)$ i.i.d. Then,

$$\begin{aligned}\sum_{i=1}^{n} V_i(y_i - x_i'c)^2 &= \sum_{j:n_j>0} \Big(\sum_{i:d_i=\alpha_j} V_i\Big)(\alpha_{yj} - \alpha_{xj}'c)^2 \\ &\sim \sum_{j:n_j>0} Q_j(\alpha_{yj} - \alpha_{xj}'c)^2\end{aligned}$$

So,

$$\beta|d \sim \arg\min_c \sum_{i=1}^{n} V_i(y_i - x_i'c)^2$$

or equivalently,

$$\beta|d \sim (\tilde{x}'\tilde{x})^{-1}(\tilde{x}'\tilde{y})$$

where $\tilde{x} = \begin{pmatrix}\sqrt{V_1}x_1' \\ \vdots \\ \sqrt{V_n}x_n'\end{pmatrix}$ and $\tilde{y} = \begin{pmatrix}\sqrt{V_1}y_1 \\ \vdots \\ \sqrt{V_n}x_n\end{pmatrix}$ is defined similarly.

The Bayesian Bootstrap

Algorithm: Bayesian Bootstrap

```
For $b = 1, \ldots, B$:
    (1) Draw $V_i^b \sim \textit{Exponential}(1)$ independently for $i = 1, \ldots, n$.
    (2) Compute $\beta^b$ using
            $\arg\min_c \sum_{i=1}^{n} V_i (y_i - x_i' c)^2.$
    Store $\beta^b$.
```

The resulting values $\beta^1, \ldots, \beta^B$ are draws from the posterior distribution $\beta | d$ associated with the limiting Dirichlet prior.