LECTURE 3
A Bayesian motivation for the proceduralist approach

There are typically two ways to validate a statistical estimation procedure

(1) Show the procedure is consistent.
(2) Show there is a Bayesian procedure that produces the same results.

3.1. Consistency

What is meant by a consistent estimator in the context of regression is the following:

Definition (Consistency)**.** *In regression an estimator $\hat{f}$ selected from a class of functions $\mathcal{F}$ is consistent if* $\forall \varepsilon > 0$

$$\lim_{n\to\infty} \sup_{\rho} \mathbb{P}_D \left\{ I[\hat{f}] > \inf_{f\in\mathcal{F}} I[f] + \varepsilon \right\} = 0,$$

where ρ is the joint distribution of the data, and $D \equiv \{(X_i, Y_i)\}_{i=1}^n \overset{iid}{\sim} \rho(x, y)$ are data sampled iid from the joint distribution.

Later in the course (in a few lectures) we will discuss why criterion (2) is valid. We will now develop a Bayesian interpretation for the procedure developed in the previous lecture.

3.2. Likelihoods

The first step in any Bayesian formulation is to state the likelihood model for the data. A more formal statement of the previous sentence falls under what is called the Likelihood principle, which can be paraphrased as all the evidence in a sample relevant to model parameters is contained in the likelihood function.

Our model so far has been

$$Y_i = f(X_i) + \varepsilon_i, \quad \varepsilon_i \overset{iid}{\sim} \mathrm{N}(0, \sigma^2),$$

with $f \in \mathcal{F}$ and $\mathcal{F} = \left\{ f \,|\, f(x) = \beta^T x \right\}$. There are two sets of parameters in this model: the vector β and the variance of the error σ^2 or $\theta = \{\beta, \sigma^2\}$. The likelihood function can be stated as

$$\mathrm{Lik}(D; \theta) = \prod_{i=1}^{n} \frac{1}{\sigma\sqrt{2\pi}} \exp\left(-\frac{\|y_i - \beta^T x_i\|^2}{2\sigma^2} \right).$$

An important idea above the above likelihood is the idea of a sufficient statistic $T(x)$. This means that once the sufficient statistic is computed the data can be thrown out with no loss of information. An example is given the likelihood for a univariate normal with known variance σ^2 the sample mean $t(x_1, ...x_n) = n^{-1} \sum_i x_i$ is a sufficient statistic. A sufficient statistic can be thought of as compressing the information in a data set. The standard way to check if a statistic is sufficient is via what is called the Neyman-Fisher Factorization Criteria:

Definition (Neyman-Fisher Factorization)**.** *If a density has the following factorization*

$$f(x_1, ..., x_n; \theta) = g(t(x_1, ..., x_n), \theta)\, h(x_1, ..., x_n),$$

then $t(x_1, ..., x_n)$ is a sufficient statistic. The above suggests we can decouple the sufficient statistic t from the data $x_1, ...x_n$.

There are a class of likelihood functions we will often work with because they have nice properties and they are useful in modeling the error or noise in the data. This class is called the exponential family and the above normal likelihood is one example.

Definition (Exponential family)**.** *A density $f(x \mid \theta)$ belongs to the exponential family if the density function as the following form*

$$f(x \mid \theta) = h(x)\, g(\theta)\, \exp\left(\eta(\theta)^T \cdot T(x) - A(\theta)\right),$$

where $T(x)$ are the sufficient statistics of the data, $\eta(\theta)$ is a function (sometimes the identity of the parameters), $h(x)$ and $g(\theta)$ serve to normalize the density.

A wide variety of likelihood models belong to the exponential family including the multivariate normal, binomial, multinomial, Poisson, and exponential densities.

3.2.1. Univariate normal

We now show that the univariate normal belongs to the exponential family.

$$f(x; \mu, \sigma) = \frac{1}{\sqrt{2\pi\sigma^2}} \exp\left(-(x-\mu)^2/(2\sigma^2)\right)$$

$$\eta = \left(\frac{\mu}{\sigma^2}, -\frac{1}{2\sigma^2}\right)^T, \quad h(x) = \frac{1}{\sqrt{2\pi}}$$

$$T(x) = \left(x, x^2\right)^T, \quad g(\eta) = \frac{\mu^2}{2\sigma^2} + \ln|\sigma|$$

3.2.2. Bernoulli

We now consider the Bernoulli distribution. For a Bernoulli random variable $x \sim \mathrm{Be}(\pi)$ where π is the mean parameter of the random variable X. The exponential family distribution is as follows

$$\begin{aligned}
\mathrm{Be}(x \mid \pi) &= \pi^x (1-\pi)^{1-x} \\
&= \exp\left[\log(\pi^x (1-\pi)^{1-x})\right] \\
&= \exp\left[x \log \pi + (1-x)\log(1-\pi)\right] \\
&= \exp\left[x(\log \pi - \log(1-\pi)) + \log(1-\pi))\right] \\
&= \exp\left[x \log\left(\frac{\pi}{1-\pi}\right) + \log(1-\pi)\right]
\end{aligned}$$

On comparing the above formula with the exponential family form, we have

$$\begin{aligned} h(x) &= 1 \\ T(x) &= x \\ \eta &= \log\left(\frac{\pi}{1-\pi}\right) \\ g(\eta) &= \log\left(\frac{1}{1-\pi}\right) \\ &= \log(1+\exp(\eta)) \end{aligned}$$

3.3. Maximum a posteriori estimation

From the derivations in the previous section if we assume the standard linear regresion model with know variance we can specify the likelihood as

$$\text{Lik}(D;\beta) = \prod_{i=1}^{n} \frac{1}{\sigma\sqrt{2\pi}} \exp\left(-\frac{\|y_i - \beta^T x_i\|^2}{2\sigma^2}\right).$$

We can also specify a prior on the effect size parameters β motivated by the James-Stein or shrinkage model

$$\pi(\beta) = \frac{1}{(2\pi)^{p/2}\gamma^{1/2}} \exp\left(-\frac{1}{2}\beta^T(\tau_0^2\mathbf{I}_p)^{-1}\beta\right),$$

By Bayes' rule the posterior probability on β is

$$\text{Post}(\beta \mid D) \propto \left[\prod_{i=1}^{n} \frac{1}{\sigma\sqrt{2\pi}} \exp\left(-\frac{\|y_i - \beta^T x_i\|^2}{2\sigma^2}\right)\right] \times \frac{1}{(2\pi)^{p/2}\gamma^{1/2}} \exp\left(-\frac{1}{2}\beta^T(\tau_0^2\mathbf{I}_p)^{-1}\beta\right).$$

We can write the negative of the log of the posterior as

$$L = (2\sigma^2)^{-1} \sum_{i=1}^{n} \|y_i - \beta^T x_i\|^2 + \frac{1}{2\tau_0^2}\beta^T\beta,$$

which can be rewritten as

$$\begin{aligned} L &= \frac{1}{n}\sum_{i=1}^{n} \|y_i - \beta^T x_i\|^2 + \frac{\sigma^2}{n\tau_0^2}\beta^T\beta \\ &= \frac{1}{n}\sum_{i=1}^{n} \|y_i - \beta^T x_i\|^2 + \lambda_n\beta^T\beta, \end{aligned}$$

where the regularization parameter λ_n is now a function of the sample size n and has an interpretation of the ration of the variance of the noise over the variance of the prior. We can minimize L to obtain what is called the maximum a posteriori (MAP) estimator

$$\hat{\beta} = \arg\min_{\beta} \left[\frac{1}{n}\sum_{i=1}^{n} \|y_i - \beta^T x_i\|^2 + \lambda_n\beta^T\beta\right],$$

which is nothing but the shrinkage estimator of the previous section and

$$\hat{\beta} = (\mathbf{X}^T\mathbf{X} + \lambda_n n\mathbf{I})^{-1}\mathbf{X}^T Y.$$

3.4. Conjugate priors

When the prior and the posterior have the same form, we say that the prior is a conjugate prior for the corresponding likelihood. For all distributions in the exponential family, we can derive a conjugate prior. Let the prior be $p(\eta \mid \tau)$, where τ denotes the hyper-parameters. The posterior can be written as:

$$p(\eta \mid X) \propto p(X \mid \eta)\, p(\eta \mid \tau)$$

The likelihood of the exponential family is:

$$p(X \mid \eta) = \left(\prod_{i=1}^{n} h(x_i) \right) \exp \left(\eta^T \sum_{i=1}^{n} T(x_i) - n g(\eta) \right)$$

Assume the prior has the form

$$p(\eta \mid \tau) \propto \exp \left\{ \eta^T \tau - \tau_0 g(\eta) \right\},$$

where we observe an inner product structure between the hyper-parameters and the parameter η. Then the posterior can be written as:

$$\begin{aligned} p(\eta \mid X) &\propto p(X \mid \eta)\; p(\eta \mid \tau) \\ &\propto \exp \left(\eta^T \sum_{i=1}^{n} T(x_i) - n g(\eta) \right) \exp \left(\eta^T \tau - \tau_0 g(\eta) \right) \\ &= \exp \left\{ \eta^T \left(\sum_{i=1}^{n} T(x_i) + \tau \right) - (n + \tau_0) g(\eta) \right\} \end{aligned}$$

The posterior has the same exponential family form as the prior, and the posterior hyper-parameters are adding the sum of the sufficient statistics to hyper-parameters of the conjugate prior. The exponential family is the only family of distributions for which the conjugate priors exist. This is a convenient property of the exponential family because conjugate priors simplify computation of the posterior. We can do algebra instead of calculus, integration.

3.4.0.1. *Algebraic perspective.* A family of priors is conjugate if it is closed under sampling, This means that a family $\mathcal{F}$ of distributions over $\eta \in \Theta$ is closed under sampling with respect to a sampling distribution $p(x \mid \eta)$ if and only if for any sample $p(\eta) \in \mathcal{F}$ it holds that $p(\eta \mid x) \in \mathcal{F}$.

An interesting observation is under mild conditions conjugate priors can be characterized by the following posterior linearity condition

$$\mathbb{E}\Big[\mathbb{E}(X \mid \eta) \mid X = x \Big] = ax + b.$$

USEFUL PROPERTIES OF THE MULTIVARIATE NORMAL*

3.1. Conditionals and marginals

For Bayesian analysis it is very useful to understand how to write joint, marginal, and conditional distributions for the multivariate normal.

Given a vector $x \in \mathbb{R}^p$ the multivariate normal density is

$$f(x) = \frac{1}{(2\pi)^{p/2}|\Sigma|^{1/2}} \exp\left(-\frac{1}{2}(x-\mu)^T\Sigma^{-1}(x-\mu)\right).$$

Now split the vector into two parts

$$x = \begin{bmatrix} x_1 \\ x_2 \end{bmatrix}, \quad \mu = \begin{bmatrix} \mu_1 \\ \mu_2 \end{bmatrix}, \quad \text{of size } \begin{bmatrix} q \times 1 \\ (p-q) \times 1 \end{bmatrix},$$

and

$$\Sigma = \begin{bmatrix} \Sigma_{11} & \Sigma_{12} \\ \Sigma_{21} & \Sigma_{22} \end{bmatrix}, \quad \text{of size } \begin{bmatrix} q \times q & q \times (p-q) \\ (p-q) \times q & (p-q) \times (p-q) \end{bmatrix}.$$

We now state the joint and marginal distributions

$$x_1 \sim \mathrm{N}(\mu_1, \Sigma_{11}), \quad x_2 \sim \mathrm{N}(\mu_2, \Sigma_{22}), \quad x \sim \mathrm{N}(\mu, \Sigma),$$

and the conditional density

$$x_1 \mid x_2 \sim \mathrm{N}\left(\mu_1 + \Sigma_{12}\Sigma_{22}^{-1}(x_2 - \mu_2), \Sigma_{11} - \Sigma_{12}\Sigma_{22}^{-1}\Sigma_{21}\right).$$

The same idea holds for other sizes of partitions.

3.2. Conjugate priors

3.2.1. Univariate normals

3.2.1.1. *Fixed variance, random mean.* We consider the parameter σ^2 fixed so we are interested in the conjugate prior for μ:

$$\pi(\mu \mid \mu_0, \sigma^2) \propto \frac{1}{\sigma_0} \exp\left(-\frac{1}{2\sigma_0^2}(\mu - \mu_0)^2\right),$$

where μ_0 and σ^2 are hyper-parameters for the prior distribution (when we don't have informative prior knowledge we typically consider $\mu_0 = 0$ and σ^2 large).

The posterior distribution for $x_1, ..., x_n$ with a univariate normal likelihood and the above prior will be

$$\text{Post}(\mu \mid x_1, ..., x_n) \sim \text{N}\left(\frac{\sigma_0^2}{\frac{\sigma^2}{n} + \sigma_0^2} \bar{x} + \frac{\sigma^2}{\frac{\sigma^2}{n} + \sigma_0^2} \mu_0, \left(\frac{1}{\sigma_0^2} + \frac{n}{\sigma^2} \right)^{-1} \right).$$

3.2.1.2. *Fixed mean, random variance.* We will formulate this setting with two parameterizations of the scale parameter: (1) the variance σ^2, (2) the precision $\tau = \frac{1}{\sigma^2}$.

The two conjugate distributions are the Gamma and the inverse Gamma (really they are the same distribution, just reparameterized)

$$\text{IG}(\alpha, \beta) : f(\sigma^2) = \frac{\beta^\alpha}{\Gamma(\alpha)} (\sigma^2)^{-\alpha - 1} \exp(-\beta (\sigma^2)^{-1}), \quad \text{Ga}(\alpha, \beta) : f(\tau) = \frac{\beta^\alpha}{\Gamma(\alpha)} \tau^{\alpha - 1} \exp(-\beta \tau).$$

The posterior distribution of σ^2 is

$$\sigma^2 \mid x_1, ..., x_n \sim \text{IG}\Big(\alpha + \frac{n}{2}, \beta + \frac{1}{2} \sum (x_i - \mu)^2 \Big).$$

The posterior distribution of τ is not surprisingly

$$\tau \mid x_1, ..., x_n \sim \text{Ga}\Big(\alpha + \frac{n}{2}, \beta + \frac{1}{2} \sum (x_i - \mu)^2 \Big).$$

3.2.1.3. *Random mean, random variance.* We now put the previous priors together in what is called a Bayesian hierarchical model:

$$\begin{aligned} x_i \mid \mu, \tau &\overset{iid}{\sim} \text{N}(\mu, (\tau)^{-1}) \\ \mu \mid \tau &\sim \text{N}(\mu_0, (\kappa_0 \tau)^{-1}) \\ \tau &\sim \text{Ga}(\alpha, \beta). \end{aligned}$$

For the above likelihood and priors the posterior distribution for the mean and precision is

$$\begin{aligned} \mu \mid \tau, x_1, ..., x_n &\sim \text{N}\Big(\frac{\mu_0 \kappa_0 + n \bar{x}}{n + \kappa_0}, (\tau(n + \kappa_0))^{-1} \Big) \\ \tau \mid x_1, ..., x_n &\sim \text{Ga}\Big(\alpha + \frac{n}{2}, \beta + \frac{1}{2} \sum (x_i - \bar{x})^2 + \frac{n}{n+1} \frac{(\bar{x} - x_i)^2}{2} \Big). \end{aligned}$$

3.2.2. Multivariate normal

Given a vector $x \in \mathbb{R}^p$ the multivariate normal density is

$$f(x) = \frac{1}{(2\pi)^{p/2} |\Sigma|^{1/2}} \exp\left(-\frac{1}{2} (x - \mu)^T \Sigma^{-1} (x - \mu) \right).$$

We will work with the precision matrix instead of the covariance and we will consider the following Bayesian hierarchical model:

$$\begin{aligned} x_i \mid \mu, \Lambda &\overset{iid}{\sim} \text{N}(\mu, (\Lambda)^{-1}) \\ \mu \mid \Lambda &\sim \text{N}(\mu_0, (\kappa_0 \Lambda)^{-1}) \\ \Lambda &\sim \text{Wi}(\Lambda_0, n_0), \end{aligned}$$

the precision matrix is modeled using the Wishart distribution

$$f(\Lambda; V, n) = \frac{|\Lambda|^{(n-d-1)/2} \exp(-.5 \text{tr}(\Lambda V^{-1}))}{2^{nd/2} |V|^{n/2} \Gamma_d(n/2)}.$$

For the above likelihood and priors the posterior distribution for the mean and precision is

$$\begin{aligned}
\mu \mid \Lambda, x_1, ..., x_n &\sim \mathrm{N}\Big(\frac{\mu_0\kappa_0 + n\bar{x}}{n+\kappa_0}, (\Lambda(n+\kappa_0))^{-1}\Big) \\
\Lambda \mid x_1, ..., x_n &\sim \mathrm{Wi}\bigg(n_0 + \frac{n}{2}, \Lambda_0 + \frac{1}{2}\left[\bar{\Sigma} + \frac{\kappa_0}{\kappa_0 + n}(\bar{x} - \mu_0)(\bar{x} - \mu_0)^T\right]\bigg).
\end{aligned}$$