Stat210A: Theoretical Statistics **Lecture Date: October 02, 2014**

Bayesian Awesomeness, Gaussian Conjugacy, Hierarchical and Empirical Bayes

Lecturer: Michael I. Jordan *Scribe: Cathy Wu*

1 Bayesian recursion

Bayesian recursion is one of the most important parts of the Bayesian paradigm. Most of the stuff we've talked about so far doesn't have a recursive relationship, but we're going to look at such a setup now.

Suppose we first collect m data points, followed by collecting $n-m$ more data points. Take

$$X_i|\theta \text{ i.i.d.}, i = 1, \cdots, n, 0 < m < n$$

Recall: We're in the Bayesian setting now, so parameters are also RV.

Then the overall posterior is

$$\begin{aligned}
p(\theta|X_1,\cdots,X_m) &\propto p(X_1,\cdots,X_m|\theta)p(\theta)\\
p(\theta|X_1,\cdots,X_m,X_{m+1},\cdots,X_n) &\propto p(X_1,\cdots,X_n|\theta)p(\theta)\\
&= p(X_1,\cdots,X_m|\theta)p(X_{m+1},\cdots,X_n|\theta)p(\theta)\\
&= \underbrace{p(X_{m+1},\cdots,X_n|\theta)}_{\text{likelihood}}\underbrace{p(\theta|X_1,\cdots,X_m)}_{\text{prior}}
\end{aligned}$$

And this is a very natural way of looking at the situation of adding new data.

2 Gaussian conjugacy

Take $X_i|\mu,\sigma^2 \overset{iid}{\sim} \mathcal{N}(\mu,\sigma^2)$

$$p(x|\mu,\sigma^2) = \frac{1}{(2\pi\sigma^2)^{n/2}}e^{-\frac{1}{2\sigma^2}\sum_i(x_i-\mu)^2} \propto (\sigma^2)^{-n/2}e^{-\frac{c}{\sigma^2}}$$

With respect to the variable σ^2, this turns out to be the inverse gamma distribution, i.e. $\sigma^2 \sim IG(a,b)$

2.1 Inverse gamma

Let $X \sim Ga(a,b)$, and consider $Z = X^{-1} \implies X = Z^{-1} \implies \left|\frac{dx}{dz}\right| = \frac{1}{Z^2}$

Then

$$p(z) = p(x(z))\left|\frac{dx}{dz}\right| = \frac{b^a}{\Gamma(a)}(z^{-1})^{a-1}e^{-b/z}\frac{1}{z^2} = \frac{b^a}{\Gamma(a)}z^{-a-1}e^{-b/z}$$

Posterior update: $\sigma^2|\mu, X$

$$b \to b + \frac{1}{2}\sum(x_i - \mu)^2$$
$$a \to a + \frac{n}{2}$$

Bayesian updates are just these 2 lines!

Remark 1. a represents the cumulative strength of the prior. As I keep adding points, I add in wisdom from past data points. The posterior then shrinks with $1/a$.

Remark 2. b represents a measure of dispersion. If my data is spread out, I add a larger term to the dispersion.

What about the other way around, posterior with respect to μ where σ^2 is fixed? (Note: related homework problem pending.) What should be my conjugate prior for μ? Gaussian.

$$p(\mu) \propto e^{-\frac{1}{2\tau^2}(\mu-\mu_0)^2}$$

For σ^2 fixed, draw $\mu \sim \mathcal{N}(\mu_0, \tau^2)$, posterior update is simple as before.

Finally, for σ^2, μ random, $\sigma^2 \sim IG$, then what distribution is $\mu|\sigma^2$ drawn from? (Recall: need to couple priors.)

3 Hierarchical and empirical Bayes

Reading: Section 11.1 Keener

Two of the most appealing things about the Bayesian paradigm: 1) recursion, and 2) you can build hierarchies. Let's now discuss the latter.

We've seen some simple hierarchies, but you can keep going and build really complicated ones. Here we examine a 3-level hierarchy. Take

$$X_i|\theta_i \sim \mathcal{N}(\theta_i, 1)$$
$$\theta_i \overset{iid}{\sim} \mathcal{N}(0, \tau^2)$$

Note that this situation is a bit odd in that we have one parameter for each data point. We can imagine that the Gaussian distributions represent variables that are obeying some physical laws. A full Bayesian model would additionally have

$$\tau^2 \sim \Lambda_\tau$$

Example inference: compute $p(\theta_i|X)$. note that it's missing the other RVs, which means they were integrated out at some point.

Now, we want a blend of Frequentist and Bayesian thinking. Let's assume τ^2 fixed, which is convenient since Frequenists wouldn't know how to address τ^2 as a RV.

Compute $p(\theta_i|X)$, which is possible from "Gaussian-ity". Since we multiply Gaussians, a common technique for computing the overall Gaussian is completing the square. However, here is an alternative approach (see also the notes that Mike will post on this topic).

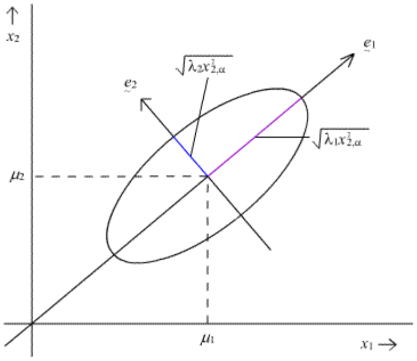
x_2
e_1
e_2
$\sqrt{\lambda_2 \chi^2_{2,\alpha}}$
$\sqrt{\lambda_1 \chi^2_{2,\alpha}}$
μ_2
μ_1
x_1

Figure 1: Given the covariance, we would expect that the mean of x_2 to increase after conditioning.

3.1 Multivariate Gaussian computation

$$X = \begin{bmatrix} X_1 \\ X_2 \end{bmatrix} \sim \mathcal{N}\left(\begin{bmatrix} \mu_1 \\ \mu_2 \end{bmatrix}, \begin{bmatrix} \Sigma_{11} & \Sigma_{12} \\ \Sigma_{21} & \Sigma_{22} \end{bmatrix} \right)$$

The core question is how to compute the conditional (which is also Guassian), see Figure 1 for an example multivariate Gaussian.

$$E[X_2|X_1] = \mu_2 + \Sigma_{21}\Sigma_{11}^{-1}(x_1 - \mu_1) \tag{1}$$

which is just a small regression problem. And

$$Var(X_2|X_1) = \Sigma_{22} - \Sigma_{21}\Sigma_{11}^{-1}\Sigma_{12} \tag{2}$$

The derivation for this involves the Schur complement (and the notes will be made available).

Example Take

$$\theta \sim \mathcal{N}(0, \tau^2), X = \theta + \epsilon, \epsilon \sim \mathcal{N}(0, 1)$$

Then,

$$E(\theta) = 0, \quad E(X) = E(E(X|\theta)) = E(\theta) = 0$$

And (assuming independent noise),

$$Var(\theta) = \tau^2, \quad Var(X) = Var(E(X|\theta)) + E(Var(X|\theta)) = \tau^2 + 1$$
$$E(X\theta) = E(\theta + \epsilon)\theta = E(\theta^2) = \tau^2$$

We stack the respective means and variances of X and θ and apply Equations 1 and 2.

$$\begin{bmatrix} X \\ \theta \end{bmatrix} \sim \mathcal{N}(\begin{bmatrix} 0 \\ 0 \end{bmatrix}, \begin{bmatrix} \tau^2 + 1 & \tau^2 \\ \tau^2 & \tau^2 \end{bmatrix})$$

$$E(\theta|X) = 0 + \frac{\tau^2}{\tau^2 + 1} x = \frac{\tau^2}{\tau^2 + 1} x$$

$$Var(\theta|X) = \tau^2 - \tau^2(\tau^2 + 1)^{-1}\tau^2 = \frac{\tau^2}{\tau^2 + 1}$$

General strategy: manipulate problem into matrix form, then plug into Equations 1 and 2. This is nicer than ≈2 pages of density calculations.

3.2 Back to the posterior

Compute $p(\theta_i|x) = p(\theta_i|x_i)$

$$p(\theta_i|x) = p(\theta_i|x_i) = \mathcal{N}(\frac{\tau^2}{\tau^2+1}X_i, \frac{\tau^2}{\tau^2+1})$$

Now we need a loss function to relate the Bayesian and Frequentist thinking. Consider (compound) loss:

$$L(\theta,\delta) = \sum_{i=1}^{p}(\theta_i - \delta_i(x))^2$$

We are working towards the *James-Stein phenomenon*, which says that using *all* of the data can give you a better estimator of the parameters than the sample mean.

If we were to minimize the Bayes risk, the best thing to do is to take the posterior mean (from last time), i.e.

$$\delta_{Bayes,i}(x) = \frac{\tau^2}{\tau^2+1}x_i = (1-\frac{1}{\tau^2})x_i \tag{3}$$

This is what a Bayesian statistician would return and is called a shrinkage factor (in the above example, shrinks from the prior mean towards 0).

Now let τ^2 be unknown.

3.3 Hierarchical Bayesian approach

Treat τ^2 as random, say $\tau^2 \sim \Lambda_\tau$, then include this in all calculations:

$$\delta(X) = E[\theta|X] = E[E[\theta|X,\tau]|X] = E[(1-\frac{1}{\tau^2+1}X)|X] = E[(1-\frac{1}{\tau^2+1})|X]X$$

Since can write down the posterior for τ (from the prior and likelihood), the Bayesian approach is done at this point.

Remark 3. One might ask: well, what's the distribution of τ^2. The Bayesian would say that they tried a bunch and their results didn't change very much. Indeed, at higher and higher levels, the distributions of hyper-hyper-parameters matter less and less.

3.4 Frequentist approach - estimate τ^2 (empirical Bayes)

What "empirical Bayes" means is that we are Bayes up to some level but at some point we stop being Bayes and turn Frequenist. In this case, we ignore the third level of the distribution hierarchy and instead use a point estimate for τ. Hence, there will be more spread for Bayesian techniques.

For instance, let's take the UMVU estimate of τ^2+1 where

$$X_i \overset{iid}{\sim} \mathcal{N}(O, \tau^2+1)$$

which is $\frac{1}{p}\sum_{i=1}^{p} X_i^2$.

This is the starting point for the Frequentist. We have data and a parameter. Then, combining with the Bayes estimator (3) yields

$$\delta_{heuristic,i}(X) = 1 - \frac{1}{\frac{1}{p}\sum_{j=1}^{p} X_j^2} X_i \tag{4}$$

Note that this estimator also has a shrinkage factor, which comes from the Bayes estimator.

Alternative approach (*James-Stein*): estimate $\frac{1}{1+\tau^2}$ by $\frac{p-2}{\sum x_j^2} \implies \delta_{JS}(X) = (1 - \frac{p-2}{\sum X_j^2})X_i$

Note/Preview: We will explain why this is a reasonable estimator later (not today). With this estimator, when $p \geq 3$, we get shrinkage like Bayesian methods. Can we get a biased estimator that's better? This estimator is an example where this is the case. It can be shown that for $p \geq 3$, this estimator dominates the maximum likelihood estimator. It also turns out this estimator is actually the Bayesian estimator under a specific prior.

Remark 4. Lesson: there are linkages in statistics!

4 Priors

The last topic with Bayesian statistics: where do priors come from? 2 schools of thought:

Subjective school This was a backlash to Neymann, Peterson, Fisher, which resulted in some weird estimator. If you just allow priors, then things get cleaner; also, it's natural: people have priors. There was quite a bit of behavioral stuff, e.g. how do you elicit priors from real people. This approach seems reasonable for small problems. For high dimensional problems, this becomes infeasible. But there weren't high dimensional problems back then, so there was a bunch of support for these approaches.

Objective school I want priors, since they do clean stuff up. But, I don't know where they come from, so I'm going to choose priors in such a way that they make as little impact on the rest of the inference machinery as possible. We're going to use Frequentist ideas, but we're going to get Bayes estimators.

5 Next time

Jeffreys prior (from Harold Jeffreys, physicist): class of priors that is invariant under reparameterization.

Also of note are *reference priors* (which we won't cover). If I have data X and prior θ. Bayesian can write down $p(X,\theta)$, $p(\theta)$, $p(\theta|X)$. The Bayesian would be willing to imagine data and compute the following:

$$E_\theta[p(\theta|X)\log(\frac{p(\theta)}{p(\theta|x)})]$$

which is the *expected KL-divergence* or *mutual information.* The reference prior maximizes the mutual information with respect to $p(\theta)$. In the one-dimensional case, the Jeffreys prior and the reference prior are the same.