Statistics 210A Lecture 10 Notes

Daniel Raban

September 28, 2021

1 Hierarchical Bayes

1.1 Recap: Choosing priors and conjugate priors

We've been talking about Bayesian statistics and estimation. Last time, we talked about 4 ways to choose a prior:

1. Prior or parallel experience

2. Subjective beliefs

3. Convenience prior

4. Objective prior (flat or Jeffreys)

We also gave examples of conjugate priors, where the posterior, $\lambda(\theta \mid x)$, comes from the same family as the prior, $\lambda(\theta)$.

Example 1.1. If $\Theta \sim \mathrm{Beta}(\alpha, \beta)$ and $X \mid \Theta \sim \mathrm{Binom}(n, \Theta)$, then $\Theta \mid X \sim \mathrm{Beta}(\alpha + X, \eta + n - X)$. The Bayes estimator for the mean squared loss is

$$\mathbb{E}[\Theta \mid X] = \frac{\alpha + X}{n + \alpha + \beta}.$$

1.2 Advantages and disadvantages of the Bayes approach

Here are some advantages of the Bayes approach to statistics.

1. **Appealing frequentist properties:** We will show later that Bayes estimators are always admissible. They also minimize average case loss.

2. **Estimator defined straightforwardly:** Compared to something like UMVU estimators, Bayes estimators are much easier to determine. We will see later that it is hard in general to find minimax estimators.

3. **Detailed output:** The posterior distribution gives a lot of information (although there is danger of overestimating the value of our posterior).

Here are some disadvantages.

1. **Difficult to choose prior:** There are many ways to choose a prior, and none of them is always better than the others.
2. **Calculations can be hard:** There is a significant amount of research on how to do the calculations for Bayesian statistics.
3. **Have to have opinions about everything:** If we don't have a parametric model, it may not make sense to come up with a prior.

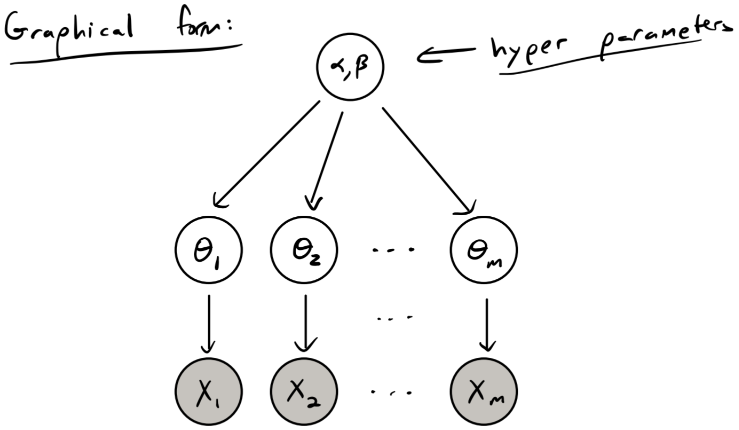
1.3 Hierarchical Bayes and graphical models

What if we want to solve a number of parallel problems at the same time?

Example 1.2. Suppose we want to predict a baseball batter's "true" batting average θ from n at bats. Let X denote the number of hits, with $X \sim \text{Binom}(n, \theta)$. The UMVU estimator is X/n. Most batting averages are between 10% and 30%, so if we observe $X = 4$ hits out of $n = 5$, we want to make sure we are not overestimating the player's batting average. We could use the convenience prior $\text{Beta}(\alpha, \beta)$, which requires us to pick α, β. How should we determine these values? The idea is that we should pool information across players $1, \dots, m$.

Here, $\alpha, \beta \sim \lambda(\alpha, \beta)$ are **hyperparameters**, which govern the distribution of the parameters. Then $\theta \mid \alpha, \beta \overset{\text{iid}}{\sim} \text{Beta}(\alpha, \beta)$, and $X_i \mid \theta, \alpha, \beta \overset{\text{ind}}{\sim} \text{Binom}(n_i, \theta_i)$.

Let's write this model in a graphical form:



This is called a **directed graphical model**. The graph above is a directed, acyclic graph, and it tells us how the joint density of these $2m+2$ random variables factorizes. If we have a graph (V, E), then the joint density factorizes as

$$p(z_1, \dots, z_m) = \prod_{i=1}^m p_i(z_i \mid \mathrm{Pa}(z_i)), \qquad \mathrm{Pa}(z_i) := (z_j : (j \to i) \in E).$$

For our model,

$$p(\alpha, \beta, \theta_1, \dots, \theta_m, x_1, \dots, x_m) = \lambda(\alpha, \beta) \prod_{i=1}^m p^\theta(\theta_i \mid \alpha, \beta) p^x(x_i \mid \theta_i).$$

1.4 Markov Chain Monte Carlo

This brings us to the idea of **Markov Chain Monte Carlo (MCMC)**: The posterior distribution is

$$\lambda(\theta \mid x) = \frac{p_\theta(x)\lambda(\theta)}{\int_\Omega p_\zeta(x)\lambda(\zeta)\, d\zeta},$$

where this integral is a high-dimensional integral (which may be difficult to calculate). An extremely successful computational strategy[1] is to set up a Markov chain whose stationary distribution is proportional to the numerator and then run the Markov chain for a long time to get samples from this distribution.

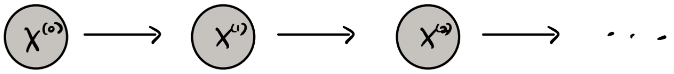
Definition 1.1. A (stationary) **Markov chain** with transition kernel $Q(y \mid x)$ and initial distribution $\pi_0(x)$ is a sequence of random variables $X^{(0)}, X^{(1)}, X^{(2)}, \dots$ such that

$$X^{(0)} \sim \pi_0(x), \qquad X^{(t+1)} \mid X^{(0)}, \dots, X^{(t)} \sim Q(\cdot \mid X^{(t)}).$$

We can think of this as

$$Q(y \mid x) = \mathbb{P}(X^{(t+1)} = y \mid X^{(t)} = x).$$

This is an example of a directed graphical model:



The marginal probability of $X^{(1)}$ is

$$\begin{aligned} \mathbb{P}(X^{(1)} = y) &= \int_{\mathcal{X}} \mathbb{P}(X^{(1)} = y \mid X^{(0)} = x)\pi_0(x)\ d\mu(x) \qquad \text{(for discrete random variables)} \\ &= \int_{\mathcal{X}} Q(y \mid x)\pi_0(x)\, d\mu(x). \end{aligned}$$

[1] This changed the general view of Bayesian statistics in the 90s.

Definition 1.2. If

$$\pi(y) = \int_{\mathcal{X}} Q(y \mid x)\pi(x)\, d\mu(x),$$

we say that π is **stationary** for the kernel Q.

A sufficient condition for π to be stationary is **detailed balance**:

$$\pi(x)Q(y \mid x) = \pi(y)Q(x \mid y) \qquad \forall x, y.$$

Proposition 1.1. *Detailed balance implies stationarity.*

Proof. If we have detailed balance,

$$\begin{aligned} \int_{\mathcal{X}} \underbrace{Q(y \mid x)\pi(x)}_{=\pi(y)Q(x|y)}\, d\mu(x) &= \pi(y) \underbrace{\int_{\mathcal{X}} Q(x \mid y)\, d\mu(x)}_{=1} \\ &= \pi(y). \end{aligned}$$

□

Theorem 1.1. *If a Markov chain with stationary distribution π is*

1. *Irreducible (for any x, y, it is possible to eventually get from x to y),*
2. *Aperiodic (the greatest common divisor of all the possible number of steps for any x to get back to itself is 1),*

then $\operatorname{dist}(X^{(t)}) \xrightarrow{t\to\infty} \pi$*, regardless of the initial distribution.*

1.5 The Gibbs Sampler

Suppose we have a generic parameter vector $\theta = (\theta_1, \ldots, \theta_d)$ and data X. Here is the algorithm:

```
Initialize $\theta = \theta^{(0)}$
For $t = 1, \ldots, T$,
    For $j = 1, \ldots, d$,
        Sample $\theta_j \sim \lambda(\theta_j \mid \theta_{\setminus j}, X)$.
    Record $\theta^{(t)} = \theta$.
```

Here are two variations on how we might do the inner loop:

1. Update a random coordinate $J^{(t)} \sim U\{1, \ldots, d\}$.
2. Update all coordinates in a random order.

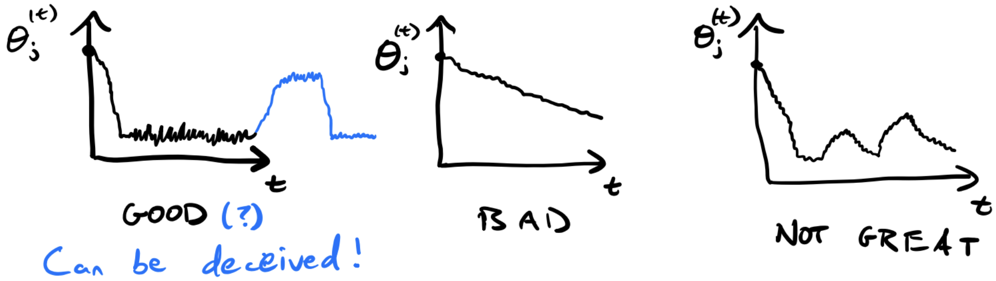
Why is this a good algorithm? If we have a directed acyclic graph, then

$$\lambda(\theta_j \mid \theta_{\setminus j}) \propto_{\theta_j} p(\theta_j \mid \theta_{\mathrm{Pa}(j)}) \prod_{i \in \mathrm{Pa}(j)} p(\theta_i \mid \theta_{\mathrm{Pa}(i)}).$$

In our example, $\theta_j \sim \mathrm{Beta}(\alpha + X_j, n + \alpha + \beta)$ is easy to sample. The α and β will be different every time we sample.

Check that the inner loops satisfies detailed balance, so the posterior distribution of the inner loop is the stationary distribution. This will give us the stationary distribution from the whole algorithm.

In practice, there can be issues:



Can be deceived!

Esp. for bimodal posterior