Statistics 210A Lecture 11 Notes

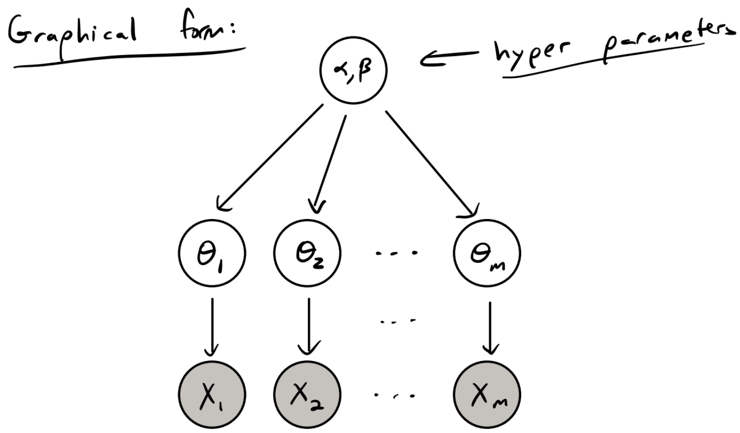
Daniel Raban

September 30, 2021

1 Hierarchical Bayesian Models and the James-Stein Estimator

1.1 Examples of hierarchical Bayesian models

Last time we talked about hierarchical Bayes models



Example 1.1. In our baseball model last time, we had the **hyperparameters** α, β with $\Theta \mid \alpha, \beta \sim \text{Beta}(\alpha, \beta)$ and $X_i \mid \Theta_i \sim \text{Binom}(n_i, \Theta_i)$.

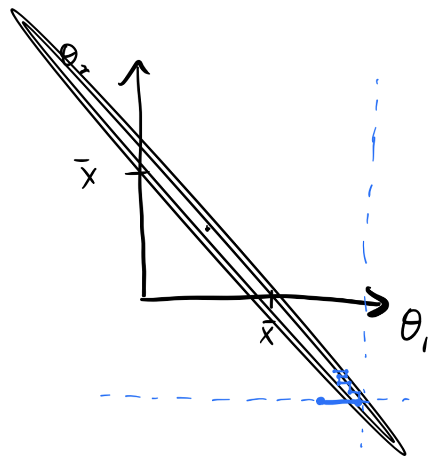
This was a directed graphical model with

$$p(\gamma, \theta_1, \dots, \theta_m, x_1, \dots, x_m) = p(\gamma) \prod_{i=1}^m p(\theta_i \mid \gamma) p(x_i \mid \theta_i).$$

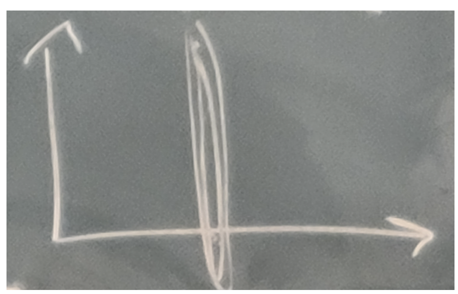
We also discussed **Markov chains** with kernels $Q(y \mid x)$; these had a **stationary distribution** π which satisfies $\int Q(y \mid x) \pi(x)\, dx$. A sufficient (but stronger) condition is **detailed balance**, which requires that $\pi(x) Q(y \mid x) = \pi(y) Q(x \mid y)$ for all x, y.

One particularly useful algorithm for sampling in hierarchical models is the **GIbbs sampler**, where we hold all the θ_i fixed except for one at a time and iteratively update our θ_is as we go. Here is an example of where things can go wrong with the Gibbs sampler.

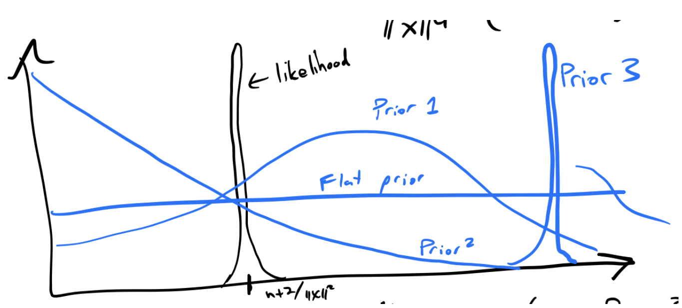
Example 1.2. Let $\Theta_1, \Theta_2 \overset{\text{iid}}{\sim} N(0,1)$ and $X_i \mid \Theta \overset{\text{iid}}{\sim} N(\Theta_1 + \Theta_2, 1)$ for $i = 1, \ldots, n$. If we do this, for large n, we will get a very highly correlated posterior distribution:



If we reparameterize the problem with $\beta_1 = \theta_1 + \theta_2$ and $\eta_2 = \theta_1 - \theta_2$, the parameters are much less dependent, so the Gibbs sampler will work better



Another issue would be when we have a bimodal distribution with the two modes having disjoint supports. Then the Gibbs sampler will not be able to jump from 1 of these modes to the other.



This can be a general problem with MCMC.

Example 1.3 (Gaussian hierarchical model)**.** Here is a Gaussian hierarchical model. Let $\tau^2 \sim \lambda(\tau^2)$ (e.g. $1/\tau^2 \sim$ Gamma), $\Theta_i \mid \tau^2 \overset{\text{iid}}{\sim} N(0, \tau^2)$, and $X_i \mid \tau^2\Theta \overset{\text{iid}}{\sim} N(\Theta_i, 1)$ for $i = 1, \dots, d$. The posterior mean is

$$\begin{aligned}
\mathbb{E}[\Theta_i \mid X] &= \mathbb{E}[\mathbb{E}[\Theta_i \mid X, \tau^2] \mid X] \\
&= \mathbb{E}\left[\frac{\tau^2}{\tau^2+1} X_i \mid X\right] \\
&= \underbrace{\left(\mathbb{E}\left[\frac{\tau^2}{1+\tau^2} \mid X\right]\right)}_{1-\mathbb{E}[\zeta \mid X]} X_i,
\end{aligned}$$

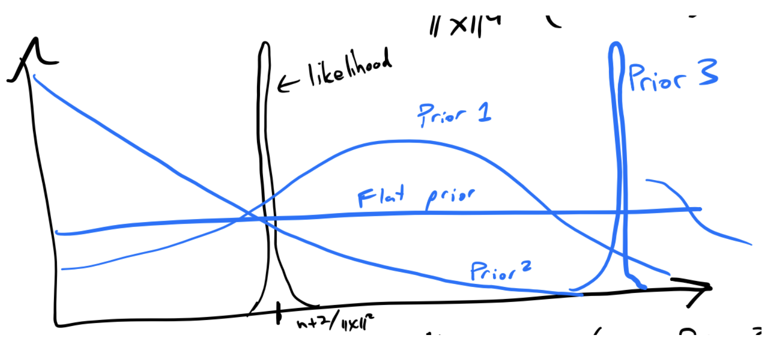
where $\zeta = \frac{1}{1+\tau^2}$. We can think of this as an *optimal shrinkage factor*.

If we marginalize out Θ, we get $X_i \mid \tau^2 \overset{\text{iid}}{N}(0, 1+\tau^2)$. If we think of this as just a problem of estimating τ^2, the sufficient statistic is

$$\begin{aligned}
\frac{\|X\|^2}{d} \mid \tau^2 &\sim \frac{1+\tau^2}{d}\chi_d^2 \\
&\mid (1+\tau^2, 2(1+\tau^2)^2/d),
\end{aligned}$$

where this notation means it is some distribution with mean $1+\tau^2$ and variance $2(1+\tau^2)^2/d$. The likelihood for τ^2 has a sharp peak near $\tau^2 = \frac{\|X\|^2}{d} - 1$ or, equivalently, near $\zeta = \frac{d}{\|X\|^2}$ (for large d).

For any reasonably open-minded prior (not prior 3 in the below figure), $\mathbb{E}[\zeta \mid X] \approx \frac{d}{\|X\|^2}$.



So

$$\mathbb{E}[\Theta_i \mid X] \approx \left(1 - \frac{d}{\|X\|^2}\right) X_i.$$

The moral is that if the prior doesn't matter so much, we can just try to estimate ζ directly from the data. This motivates the idea of **empirical Bayes** models: Write down a hierarchical model and just try to estimate a parameter like ζ using the data. In this way, we don't need to use the Gibbs sampler.

1.2 The James-Stein estimator

Empirical Bayes is a hybrid approach in which we treat the hyperparameters as fixed and treat the paramters as random.

Example 1.4. Think of τ^2 (or of ζ) as a fixed parameter, so we have $X_i \overset{\text{iid}}{N}(0, 1+\tau^2)$ and $\|X\|^2 \sim (1+\tau^2)\chi^2_d$. Then the UMVU estimator for τ^2 is

$$\widehat{\tau}^2 = \frac{\|X\|^2}{d} - 1, \quad \text{which gives} \quad \widehat{\zeta} = \frac{1}{1+\widehat{\tau}^2} = \frac{d}{\|X\|^2}.$$

This is not great because it can be negative. What if we took the UMVUE for ζ? Then we get the James-Stein estimator.

James and stein proposed that for $d \geq 3$,

$$\delta_{\text{JS}}(X) = \left(1 - \frac{d-2}{\|X\|^2}\right) X.$$

The interpretation is that $\frac{d-2}{\|X\|^2}$ is the UMVU estimator for ζ:

Proposition 1.1. *If* $Y \sim \chi^2_d = \text{Gamma}(d/2, 2)$ *with* $d \geq 3$*, then* $\mathbb{E}[1/Y] = \frac{1}{d-2}$.

Proof.

$$\begin{aligned}
\mathbb{E}\left[\frac{1}{Y}\right] &= \int_0^\infty \frac{1}{y} \frac{1}{2^{d/2}\Gamma(d/2)} y^{d-1} e^{-y/2}\, dy \\
&= \frac{2^{(d-2)}\Gamma((d-2)/2)}{2^{d/2}\Gamma(d/2)} \int_0^\infty \frac{1}{y} \frac{1}{2^{(d-2)/2}\Gamma(d/2)} y^{(d-2)/2-1} e^{-y/2}\, dy \\
&= \frac{1}{2} \cdot \frac{1}{(d-2)/2} \\
&= \frac{1}{d-2}.
\end{aligned}$$

□

Using the proposition,

$$\frac{\|X\|^2}{1+\tau^2} \sim \chi^2_d \implies \zeta^{-1}\, \mathbb{E}\left[\frac{1}{\|X\|^2}\right] = \frac{1}{d-2} \implies \widehat{\zeta} = \frac{d-2}{\|X\|^2}.$$

But the James-Stein estimator is more interesting than just this. Going back to a non-Bayesian model, suppose $X_j \sim N(\theta_j, 1)$ with $\theta \in \mathbb{R}^d$. Then for $d \geq 3$, X is inadmissible as an estimator of θ for the MSE. Say we have n observations:

Proposition 1.2 (James-Stein[1]). *Let* $X_i \overset{\text{iid}}{\sim} N_d(\theta, \sigma^2 I_d)$ *for* $i = 1, \ldots, n$ *with known* $\sigma^2 > 0$. *For*

$$\delta_{\text{JS}} = \left(1 - \frac{(d-2)\sigma^2/n}{\|\overline{X}\|^2}\right) \overline{X},$$

$$\text{MSE}(\theta, \delta_{\text{JS}}) < \text{MSE}(\theta, \overline{X})$$

for all $\theta \in \mathbb{R}^d$.

This says that if we have a bunch of unrelated experiments and we pool the observations together, we can get a better estimator for all of them by combining our observations.

Remark 1.1. We don't need to shrink around 0. For any $\theta_0 \in \mathbb{R}^d$,

$$\delta(X) = \theta_0 + \left(1 - \frac{d-2}{\|X - \theta_0\|^2}\right)(X - \theta_0)$$

renders X itself inadmissible for the mean squared error.

Next time, we will prove this result using Stein's lemma.

[1]This shocking result came out in the 50s, and no one was prepared for it.