Hierarchical models + Bayesian shrinkage

Paul Goldsmith-Pinkham

March 16, 2021

Today's topic: hierarchical modeling + shrinkage

- Already touched on shrinkage in the context of lasso
  - Today, we're going to provide a more general model structure for shrinkage and penalization

- Going to be studying a framework for thinking about two (simultaneous) features of our estimates:
  - Estimation uncertainty
  - True heterogeneity

- When considering two (or more) estimates, differences and variation in these estimates can be driven by either noise, or true variation
  - Goal today is to give a framework for considering these
  - Additional, highlight that there are improved methods that can be used for some estimands

Motivation for Bayesian approaches

- To motivate and give context to our discussion, will present the following three estimation problems.

- These are very different types of problems, but will all hopefully be more accessible following our discussion today.

Three examples - (1) Microcredit + pooling

- A set of experiments run in different locations, studying the efficiacy of microcredit on a variety of outcomes
  - Meager (2019)

- For every location l, we have a treatment, T, affecting our outcomes Y.

- We can use this to estimate $\hat{\tau}_l$, an unbiased estimate of the effect of microcredit in location l, τ_l
  - Note that these can be very different experiments
  - This gives a set of estimates, $\hat{\boldsymbol{\tau}} = \{\hat{\tau}_1, \ldots, \hat{\tau}_L\}$

- What are things we might want to say about these?
  - Are they similar to one another? If so, is that informative external validity?
  - Are they different? Is that because of estimation error, or heterogeneity?

- Could we improve on predictions more generally?
  - Chetty and Hendren (2017), Angrist et al. (2017), Goldsmith-Pinkham, Pinkovskiy and Wallace (2021)

Three examples - (2) Predictability and uncertainty

- Excess stock returns (above some risk free rate) are predictable based on historical data (e.g. the dividend yield). How does this affect our willingness to invest in stocks, depending on our horizon?
  - Barberis (2000)

- Our returns in period t are a predictable feature of lagged returns r_{t-1} and other characteristics x_{t-1} (plus some innovation):

$$z_t = \alpha + x_{t-1}B + \epsilon_t,$$

  where $z_t = (r_t, x_t)$ and $\epsilon \sim \mathcal{N}(0, \Sigma)$ (e.g. a VAR)

- These VAR parameters (α, B, Σ) are easily estimated using historical data
  - We can use this to model predictions about future values of r_{t+k}

- Riskiness in the investment will drive our decision in how much to invest
  - But crucially, riskiness is not just a function of Σ
  - The uncertainty of α and B play an important role

Three examples - (3) Estimation

- Recall from demand models that choice modeling can be complex
  - Important to allow for rich forms of choice substitution and preference heterogeneity
- Imagine we want to model labor supply of Uber drivers (Chen et al. (2019))
  - Specficially model the time-varying reservation wage for drivers

$$w_{it}^* = \mu_i(t) + \epsilon_{it}$$

  - w_{it}^* unobserved, but we see the decision to work, and the expected wage at a given time
- Crucially, a rich and flexible structure on ϵ_{it} is necessary, but makes things computationally complicated
  - Analgous to the multivariate logit / probit problem
- Can we use structure on the errors to make estimation feasible?

Bayes' rule and the Likelihood

- Recall from our (and your previous classes') discussion of the likelihood, that for a given model of i.i.d. data $(Y_i, X_i)_{i=1}^n$ with model parameters θ, we write down our likelihood:

$$L(\theta) = \prod_{i=1}^{n} f_\theta(Y_i, X_i)$$

- But really, this is just the joint probability of the data – e.g. $f(Y, X)$. If we are willing to be Bayesian, e.g., view the parameter θ as a random variable, we can even say $f(Y, X|\theta)$
  - Then, by Bayes' rule, we know we can write

$$f(\theta|Y, X) = \frac{f(Y, X|\theta) \times p(\theta)}{f(Y, X)}$$

- Note that this can be written as

$$f(\theta|Y, X) \propto f(Y, X|\theta) \times p(\theta),$$

because the denominator will only rescale the probability distribution to ensure that it is well-defined (integrates to 1)

Bayes' Rule and the Likelihood

$$f(\theta|Y,X) \propto f(YX|\theta) \times p(\theta),$$

- $f(\theta|Y,X)$ is known as the posterior of θ, while $p(\theta)$ is known as the prior. Tying this all together is the likliehood – $f(Y,X|\theta)$
- This is a simple application of Bayes' rule. However, it adds two crucial features to the likelihood:
  - The prior $p(\theta)$. Picking a parameter that maximizes $L(\theta) \times p(\theta)$ will give different estimators than just $L(\theta)$ except in special cases
  - Distribution of θ in finite samples. Recall that inference for MLE estiamtes of θ rely on asymptotic approximations of Normality

Simple Bernouilli example

- Let's start with a very simple example. We are considering a dataset of successes and failures (denoted by X_i)
  - This is a binomial, with total outcomes n and total failures $\sum_i X_i$
  - Failure rate is initially assumed to be identical p

- We know our joint distribution:

$$f(X|p) = p^{\sum_i X_i}(1-p)^{n-\sum_i X_i}$$

- We could solve for the MLE of p easily – that gives us $\hat{p}_{MLE} = \sum_i X_i / n$

- What does our *posterior* of p look like, given the data?
  - First neeed a prior – let's assume it's uniform over (0,1)

- Then, the posterior is

$$f(p|X) \propto f(X|p) \times 1 = p^{\sum_i X_i}(1-p)^{n-\sum_i X_i}$$

  - This is a *Beta* distribution with parameters $\alpha = \sum_i X_i + 1$ and $\beta = n - \sum_i X_i + 1$.
  - Posterior mean of Beta dist? $\alpha/(\alpha+\beta)$ Posterior mode? $\alpha - 1/(\alpha+\beta-2)$

Simple Bernouilli example

- Our prior was pretty uninformative / uninteresting
  - We can use a *conjugate* prior and get easily interpretable posteriors
  - E.g. conjugate prior for Bernouilli is a Beta distribution
  - Let's assume a Beta prior with $\alpha_p = 1$ and $\beta_p = 1$ (mean of 0.5)

- Then, the posterior is

$$f(p|X) \propto p^{\sum_i X_i + \alpha_p}(1-p)^{n - \sum_i X_i + \beta_p}$$

  - This is a *Beta* distribution with parameters $\alpha = \sum_i X_i + \alpha_p - 1$ and $\beta = n - \sum_i X_i + \beta_p - 1$.

- Hence our prior shifts our posterior estimates, but as n gets large, our posterior will converge to the MLE

Simple Normal example

- Now let's consider another example with Normal distributions instead
- We observe a set of observations X_i which we assume come from a normal distribution, with unknown mean and known variance $\sigma^2 = 1$
  - Knowing the variance just makes our life easier and is not necessary in general
  - Note that this is saying $X_i = \mu + \epsilon_i$, $\epsilon_i \sim \mathcal{N}(0, 1)$
- Recall that we need a prior to any Bayesian analysis. Our only unknown parameter is μ. What is our prior over this?
  - The conjugate prior for a Normal mean is also Normal
  - Let's assume a prior with mean zero, and variance 100
- Hence, we construct our posterior as:

$$f(\mu|X) \propto f(X|\mu) \times p(\mu) \propto \exp\left(-\sum_i \frac{(x_i - \mu)^2}{2\sigma^2}\right) \times \exp\left(-\frac{(\mu - 0)^2}{2 \times 100}\right)$$

- The trick is to combine terms, and then drop out any that do not involve μ. Then, complete the square.

Simple Normal example

- We are left with

$$f(\mu|X) \propto \exp\left(\frac{(\mu - \mu_{post})^2}{2\sigma^2_{post}}\right)$$

$$\mu_{post} = \frac{\sum_i x_i}{n + 1/100}$$

$$\sigma^2_{post} = \left(\frac{1}{100} + \frac{n}{1}\right)^{-1}$$

- In our setting, recall we had a prior at zero, with a large variance (so we really weren't very precise)
  - As a consequence, our posterior mean is shrunk towards zero, but not by a lot
  - If we make the prior more informative (smaller variance) or move the shrinkage point, that shifts both parameters
  - Again, as the data gets large, the prior (should) matter less

Shrinkage with more structure

- So far, our analysis focused on single parameter estimation
  - Now let's consider a set of mean parameters that we're interested in across locations
  - We're going to create hiearchichical structure on our parameters

- Consider means across l locations, where we observe observations within each location, and we model this as:

$$y_{il} \sim \mathcal{N}(\mu_l, \sigma^2)$$
$$\mu_l \sim \mathcal{N}(\mu, \eta^2)$$

(assuming σ^2 is known for simplicity)

- Note that we could write a simple version of this as:

$$y_{il} = \mu + \underbrace{(\mu_l - \mu)}_{\text{Variation from heterogeneity}} + \underbrace{(y_{il} - \mu_l)}_{\text{Sampling variation}}$$

where the two pieces on the right are random variables

Shrinkage with more structure

$$y_{il} \sim \mathcal{N}(\mu_l, \sigma^2)$$
$$\mu_l \sim \mathcal{N}(\mu, \eta^2)$$

- This is a hierarchical multivariate normal setting
  - A simple version! For example, we could allow for correlation across locations, or within locations
  - If you can dream it, you can try to estimate it

- The key details:
  - In this structure, like with the simple normal example, the estimate for mu_l will be shrunk towards the overall mean μ.
  - This shrinkage will be a function of the relative variance σ^2 and η^2
  - To do full Bayesian estimation here, recall that we need a prior for μ as well – this would be specified by the researcher
    - Implemnting this prior is the crucial distinction of going “Full Bayes”

Going fully hierarchical Bayes

Why does adding this structure help?

- Why would we want to add more structure in this way?

- Many potential reasons. A sample:
  1. Allow for shrinkage that accounts for covariates (Meager study).
     - Not only does this allow for shrinkage, but we can diagnose heterogeneity
  2. We can estimate posterior distributions of estimates, and consider predictive distributions of outcomes (Barberis study)
     - We have estimates of $p(\theta|X)$ (posterior), but can also estimate $p(X_{new}|X)$ (posterior predictive)
  3. We can estimate much more complicated statistical systems using computational techniques
     - E.g. complicated demand and supply systems! (Chen et al. (2019))
     - Let's discuss this next

Estimation of posterior using MCMC methods

- So far, much of what we have discussed in these hierachical models were either simple (e.g. assumed known variances) or conjugate (very restrictive assumptions on parametrization)

- None of this is necessary – Bayes' rule is well-defined irrespective of the parameterization choices, and θ can include many terms:

$$p(\theta|X) \propto f(X|\theta)p(\theta)$$

- The question, of course, is that if an analytic solution does not fall out as in our examples above, how does one generate a posterior distribution for θ?

- Key insight – we want to generate a chain of estimates $\theta_0, \theta_1, \ldots,$ such that after sufficiently large k , the estimates $\theta_k, \ldots$ are random draws from the posterior of θ.

- E.g. consider the challenge in the Bernouilli example
  - If I did not know the posterior parameters for the distribution (and the general form of the parameteric model), how do I algorithmically generate draws of θ such that they will line up?

Estimation using Hamiltonian Monte Carlo (hybrid MCMC)

- Beyond scope of this course, but intuitively the problem of searching over this space is just as challenging as our discusssions of finding MLEs using numerical methods

- When I learned all of this in 2005, Gibbs Sampler and Metropolis-Hastings algorithms were the default approaches
  - However, this is no longer the case! (Betancourt + Girolami (2013))

- One of the fastest algorithms to use now is known as Hamiltonian Monte Carlo
  - This method is much more complicated / challenging to implement that Gibbs or MH
  - But, open source community has solved this problem through *Stan* (RStan)
  - If you are going this route, use Stan

Downsides of Full Bayes

- What's the "cost" of all this?
- In most cases, it's the parameterization.
- We're typically quite uncomfortable with doing this
- e.g. consider the linear regression model – how do you want to parametrize the error terms? Can we approximate "robust" se? It's semi-parametric!
- Potential solutions?
  - Try to flexibly model the covariance structure
  - Assume the asymptotic distribution of the parameters
- The other "downside" – uncomfortable with priors
  - How do we know what to shrink to? Doesn't this create bias?

Middle ground

- If these are issues you have, there are middle ground solutions
- First, recall that much of the time, we assume our *estimates* are normal (that's how we do inference – asymptotic normality)
  - Classic result from Rubin (1981)
- We can just consider parameters

$$\hat{\tau}_l \sim \mathcal{N}(\tau_l, \hat{se}_l^2)$$
$$\tau_l \sim \mathcal{N}(\tau, \sigma_\tau^2)$$

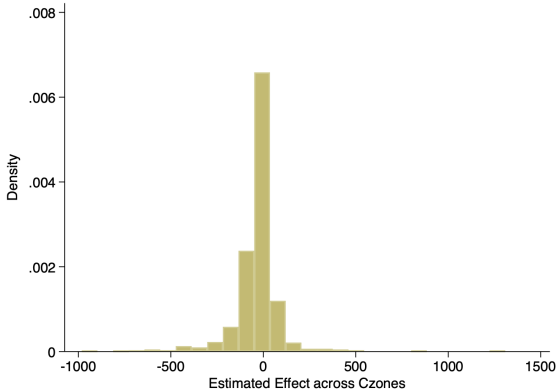
- Now we just need a prior on τ and σ_τ^2, and we can do full Bayes!
  - But we didn't assume anything about the DGP of the outcomes – we just focused on the normality of the estimates (and assumed unbiasedness)
- So the only "new" thing we're doing is assuming priors (but ignoring all other data)
  - How should we pick priors?

Cheating on priors with Empirical Bayes

- The problem is we don't know what prior to pick, a lot of the time. E.g., what should we shrink towards?
- The empirical Bayes approach suggests that the "traditional" estimator (e.g. the MLE) is good enough
  - Intuitively – shrink the values towards the overall mean
- This isn't really Bayesian but it seems to work – ignores variability in this estimate

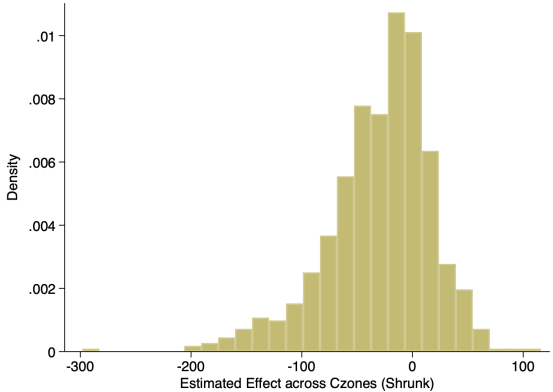
When can this be useful? Goldsmith-Pinkham et al. (2021)

- Paper studying Medicare impact on health insurance and credit outcomes
  - We use Regression Discontinuity (RD), but just assume we have correctly estimated causal effects of Medicare in each location $\hat{\tau}_l$
- These estimates are *noisy*



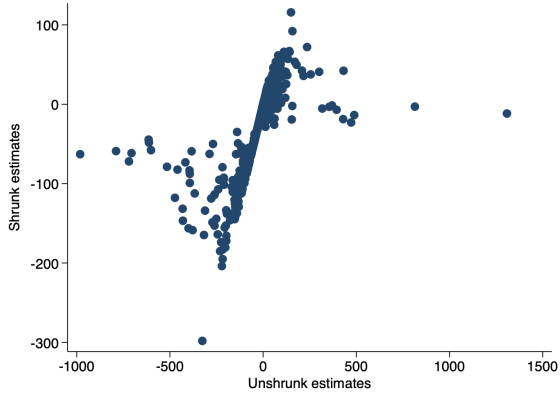
When can this be useful? Goldsmith-Pinkham et al. (2021)

- Paper studying Medicare impact on health insurance and credit outcomes
  - We use Regression Discontinuity (RD), but just assume we have correctly estimated causal effects of Medicare in each location $\hat{\tau}_l$
- These estimates are *noisy*
- Simple approach: shrink these estimates towards the overall mean of our estimates, weighted by how noisy these estimates are



When can this be useful? Goldsmith-Pinkham et al. (2021)

- Paper studying Medicare impact on health insurance and credit outcomes
  - We use Regression Discontinuity (RD), but just assume we have correctly estimated causal effects of Medicare in each location $\hat{\tau}_l$
- These estimates are *noisy*
- Simple approach: shrink these estimates towards the overall mean of our estimates, weighted by how noisy these estimates are
- This has substantial impact – makes our estimates more precise, but also create bias for the very noisy estimates



The shrinkage tradeoff

- Remember the key reason why shrinkage is so effective: improvements on mean squared error are gained by increasing bias

- The “shrunk” estimates have some bias introduced in order to massive reduce their noise

- In economics, we’ve traditionally worried about unbiasedness a lot
  - Important to identify what issues this can create
  - If something is so noisy that it’s not informative, a small amount of bias can be very useful
  - Especially with prediction problems!

- If we are trying to do prediction, forecast error (e.g. MSE) is typically much more important than unbiasedness

Having this structure gives intuition in many other settings

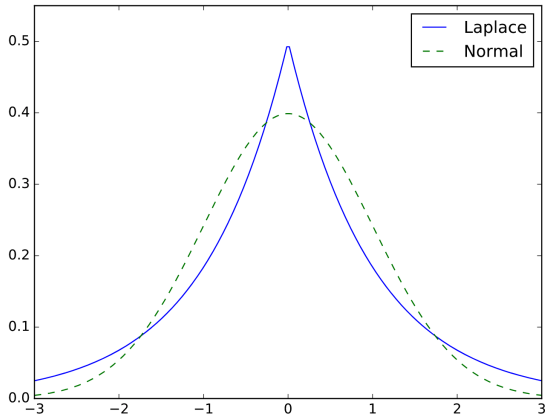
- This comes back to Lasso shrinkage
  - Lasso can be viewed as a Laplace prior on the beta in our regression methods

- Other regularization methods (e.g. ridge) use a less sharp-peaked prior (such as ridge, which is a Gaussian)

$$p(\beta|Y,X) \propto f(Y,X|\beta,\sigma_\epsilon^2)p(\beta)$$

$$\min_\beta n^{-1}\sum_{i=1}^{n}(Y_i - X_i\beta)^2 + \lambda\sum_{k=1}^{p}|\beta_k|$$

- Hence, the reasons behind Bayesian methods also motivate regularization methods and vice versa



Conclude (before examples)

- Adding structure can be very powerful, and is computationaly much easier to do now
- These techniques are widely applicable
  - Finance, IO, labor, public
- More generally, however, understanding how different methods incorporate shrinkage is valuable
- Decision tree for how to use these methods is non-obvious but for me, fall into three categories:
  1. I have many noisy estimates, and I want to exploit their joint info (at minimum by regressing towards their average)
  2. I have a complicated statistical problem that I need to generate an underlying distribution from
  3. I need to predict outcomes incorporating estimated parameter uncertainty
- You should consider Bayesian methods in this setting!

Example papers!

- Meager (2017)
- Angrist et al. (2017)
- Johannes, Lochstoer and Mou (2016)
- Cohen and Einav (2007)
- Chinco et al. (2021)
- Athey et al. (2018)
- Mackey et al. (2015)