Bayesian Hierarchical Models and Microcredit

Based on Meager

(2019, AEJ Applied)

- Tom Glinnan

Outline of the Paper

- Economics has a tonne of focus on *internal validity* - how well effects are identified within a particular study. But how well do these generalize?
- If expanding microcredit is very effective in one country at one time period with one particular set of institutions, what about the situation where you actually want to do your policy?
- This *external validity* is **as important** for policymaking (arguably more)
- How to get at it $\rightarrow$ combining the results of many studies. Here: (only) 7 RCTs
- Main tool used - a Bayesian Hierarchical Model
- I'll also introduce Bayesianism more generally

Bayesianism I

Some Philosophy: what does it mean for something to be random? Two main schools:

- *Frequentist:* If we run the experiment n times,

$$\mathbb{P}(A) = \lim_{n \to \infty} \frac{1}{n} \ \#\text{Trials where } A \text{ happens}$$

- *Bayesian:* $\mathbb{P}(A)$ is how likely we think A is to happen, with our beliefs being updated by Bayes' rule

Think of a coin toss. *Randomness is a model*

Bayesianism II

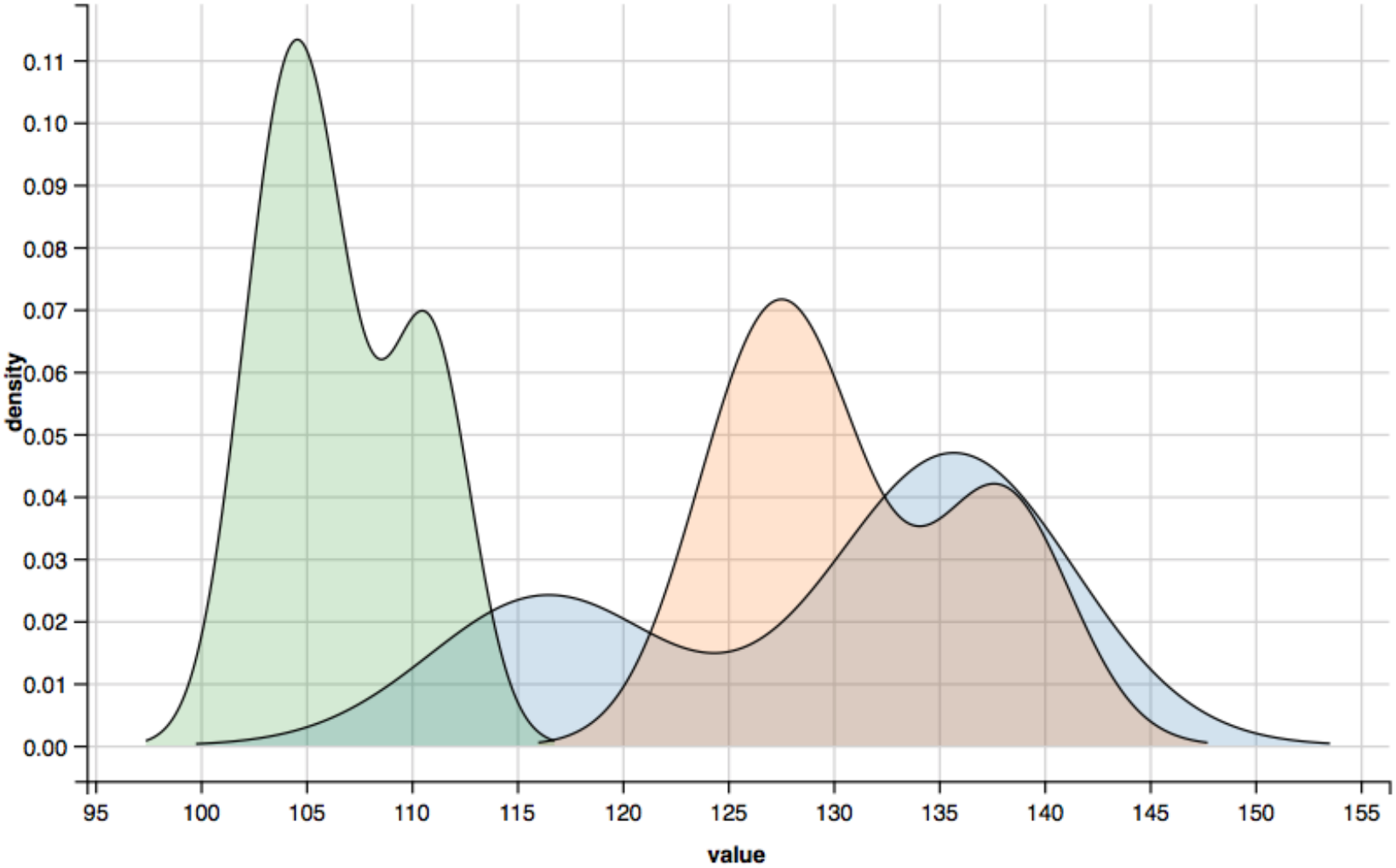
Why does this matter? Say we have a model like one we're used to:

$$y_i = x_i' \beta_0 + \varepsilon_i \qquad \mathbb{E}(x_i \varepsilon_i) = 0$$

This is just specifying a relationship between rvs (x_i, y_i)

- **If** probability is just long run averages, it doesn't make sense to talk about a pdf over β. There's one β ('the truth' β_0) and that's it
- But **if** probability is just rational belief, then it does make sense to talk about pdfs over β (ie your beliefs about β, even though there's only one true β_0)

Bayesianism III



Different data $x = (x_i, y_i) \implies$ Different beliefs about β

Bayesianism IV

If we're Bayesians we could get point estimators $\hat{\beta}$, but why throw information away? Our entire beliefs about β_0 are summed up in $p(\beta|x)$

Maybe helpful: Bayesian methods keep x fixed and produce a pdf over β (the posterior distribution). Frequentist methods keep β fixed and ask what happens if you had got different x

Aim: $p(\beta|x)$

Bayesianism V

Stats 101 Bayes Theorem:

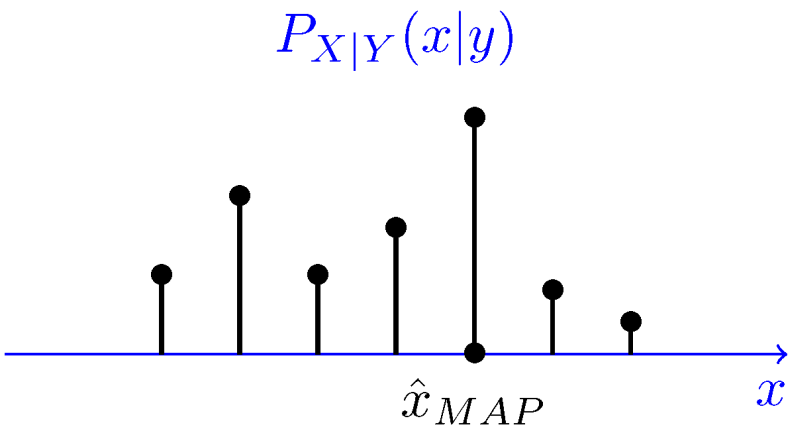
$$p(\beta|x) = \frac{p(x|\beta) \cdot p(\beta)}{p(x)}$$

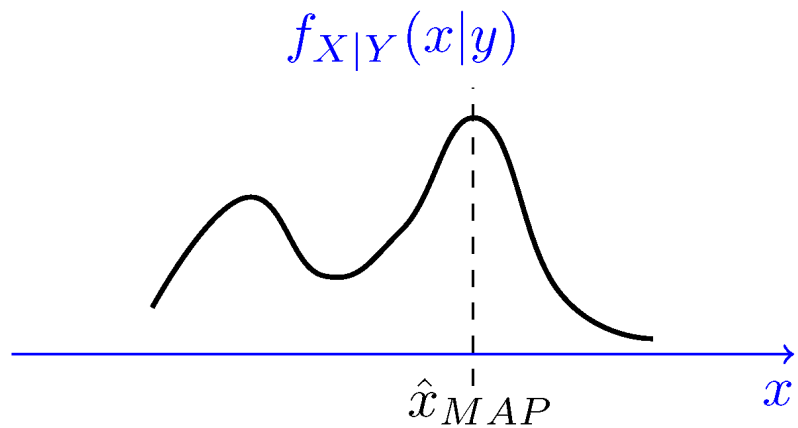
Also known as:

$$\text{Posterior} = \frac{\text{Likelihood} \times \text{Prior}}{\text{A Normalizing Constant}} \propto \text{Likelihood} \times \text{Prior}$$

Cookbook: 1) model the likelihood and the prior 2) calculate the posterior (analytically or by computer methods) 3) report the posterior, or functionals of it

The MAP





If we insist on one number as an estimator, we can choose eg the most likely thing in the posterior: the *Maximum a Posteriori*

Calculation I

- Many quantities of interest can be written as $\lambda = \mathbb{E}_{\text{Posterior}} f(x)$. Often the posterior itself is very high-dimensional

- There are many procedures to estimate functionals directly without estimating all of the posterior

Three main methods: Grid-Based, Monte Carlo, and MCMC

Calculation I

Method 1: **Grid-Based**

- Draw a grid on the space of x and work out the posterior at each point to get b samples x_i. Then as $b \to \infty$ make grid finer, and use estimate $1/b \sum f(b)$
- Consistent but 1) curse of dimensionalty 2) very slow even in low dimensions 3) if your grid doesn't cover the support, it's inconsistent
- Intuition for second point: if $f(x_i)$ is very low then it counts little towards the estimate. Would be better to sample more from regions with large $f(x_i)$, then make an adjustment *ex post*

Calculation II

Method 2: **Monte Carlo**

- Consider a distribution Q on the space of x. Draw a random value from Q and take that as your x_i. Then it turns out a good estimate is a weighted average of the $f(x_i)$
- Intuitively, we want to sample most from where f is highest, so ideally want to choose $Q = f$. However we don't know f - we're trying to work it out!
- Therefore often choose $Q =$ uniform, $Q =$ prior, etc.

Calculation II

Method 3: **Markov Chain Monte Carlo (MCMC)**

- A way of creating a Markov Chain over the space of x_i which has the same distribution as f
- By sampling à la Monte Carlo from this we effectively get $f = Q$, so peak efficiency
- However, the correlation between the samples can cause reductions in efficiency $\rightarrow$ a trade-off between MCMC and MC

Calculation III

Pseudocode: (Metropolis-hastings) MCMC:

1. Pick an initial β_1 to start with. Then for $b = 1, ..., B$ large
   - 1.1 Draw from a proposal distribution Q over β. This suggests some β'
   - 1.2 Calculate the posterior at β' as:

$$p(\beta'|x) \propto p(x|\beta') \cdot p(\beta')$$

   - 1.3 Compute the transition probability $t \in [0, 1]$ from the ratio of $p(\beta'|x)$ and $p(\beta_b|x)$
   - 1.4 Draw a random number $u \sim U[0, 1]$. If $u \leq t$ then set $\beta_{b+1} = \beta'$, else $\beta_{b+1} = \beta_b$. Save β_{b+1} to memory
2. Finally we define

$$\hat{\lambda} = \frac{1}{B} \sum_{b=1}^{B} f(\beta_b)$$

Calculation IV

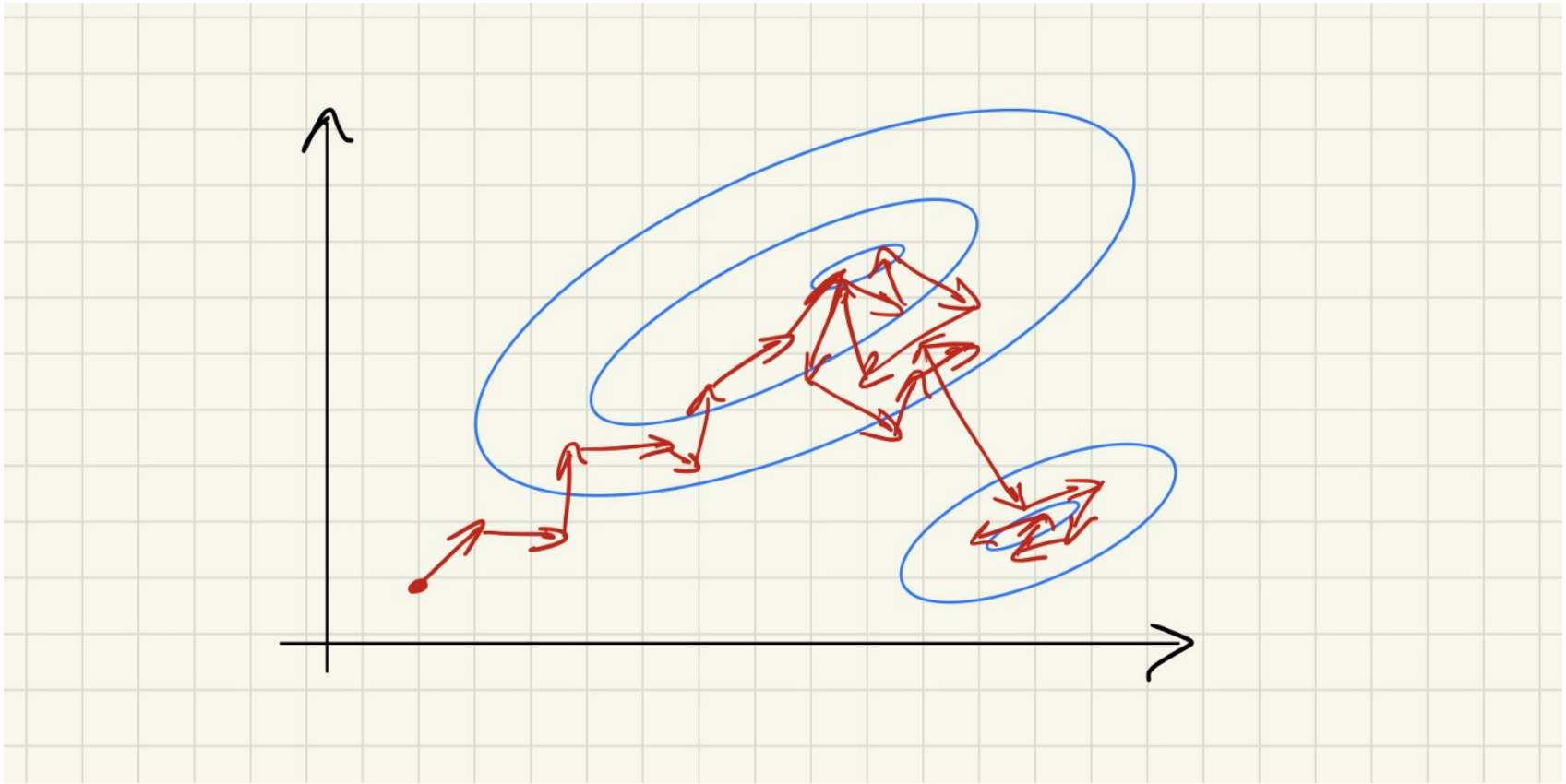
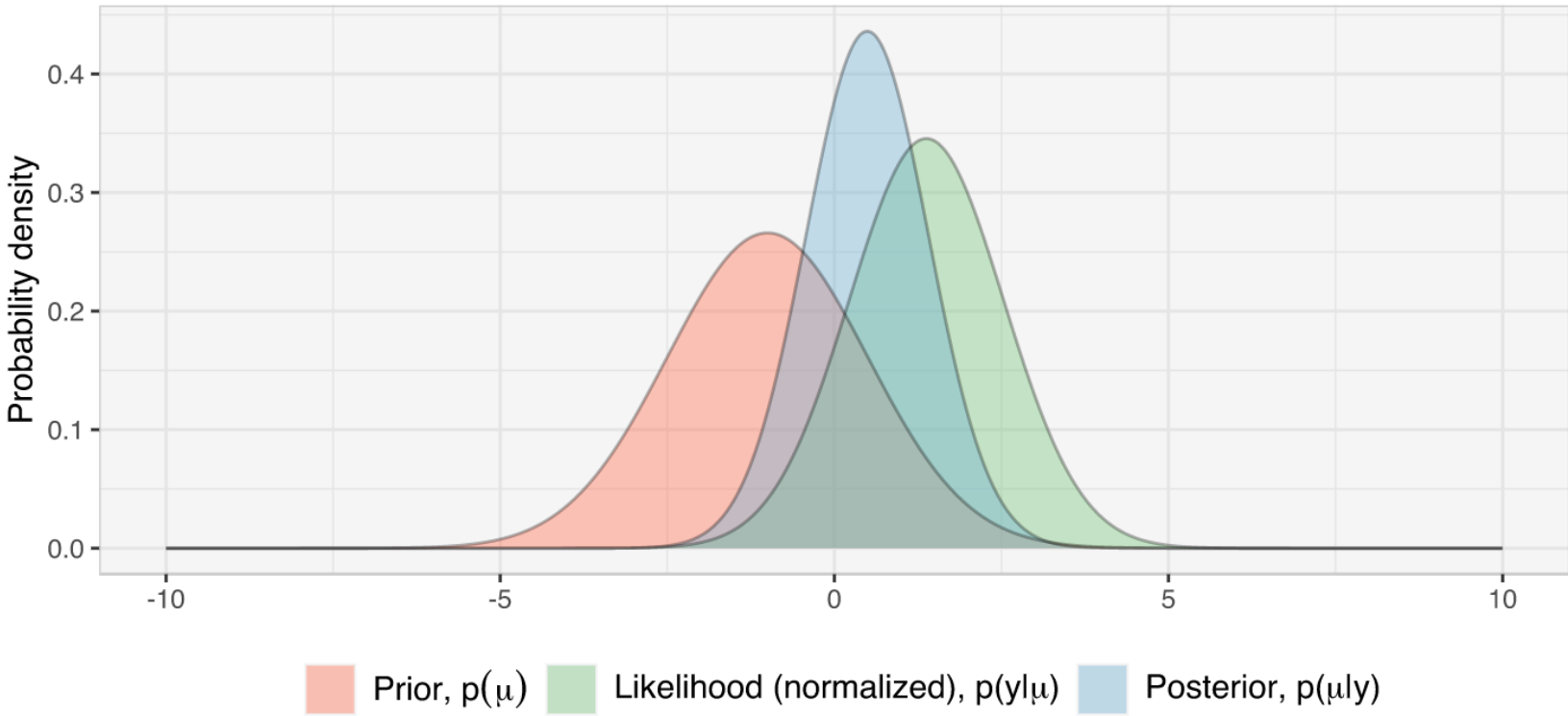


Figure: MCMC 'Exploring the Posterior'

Bayesian Estimation I

- Estimators such as the MAP can be analysed in the normal (frequentist) way
- Nearly always they end up being some kind of average of the mode of the prior and the mode of the likelihood (MLE!)
- As such, they bias estimators. Intimately connected to the idea of **shrinkage**. For example, the Lasso can be interpreted as a MAP estimator with a certain prior around 0 (and hence it shrinks estimates towards 0)
- This biasing helps with finite sample performance (eg smaller MSE) but disappears as $n \to \infty$ so we still have consistency

Bayesian Estimation II



The posterior is an average between prior and likelihood; between new and old knowledge

Bayesian Estimation III

The downside of Bayesian Estimation: choosing priors. Do we choose:

- One which reflects previous scientific beliefs (eg in macro)?
- One which is 'least informative' in some sense?
- One which the result is least sensitive to?
- One which which gives the same result no matter how you parametize the relationship between the observable rvs in your model?

Before you judge: whenever you run a Probit / Logit you're making very strong assumptions too!

Hierarchical Models I

The idea is of a hierarchical model is that studies should be thought about in two 'layers':

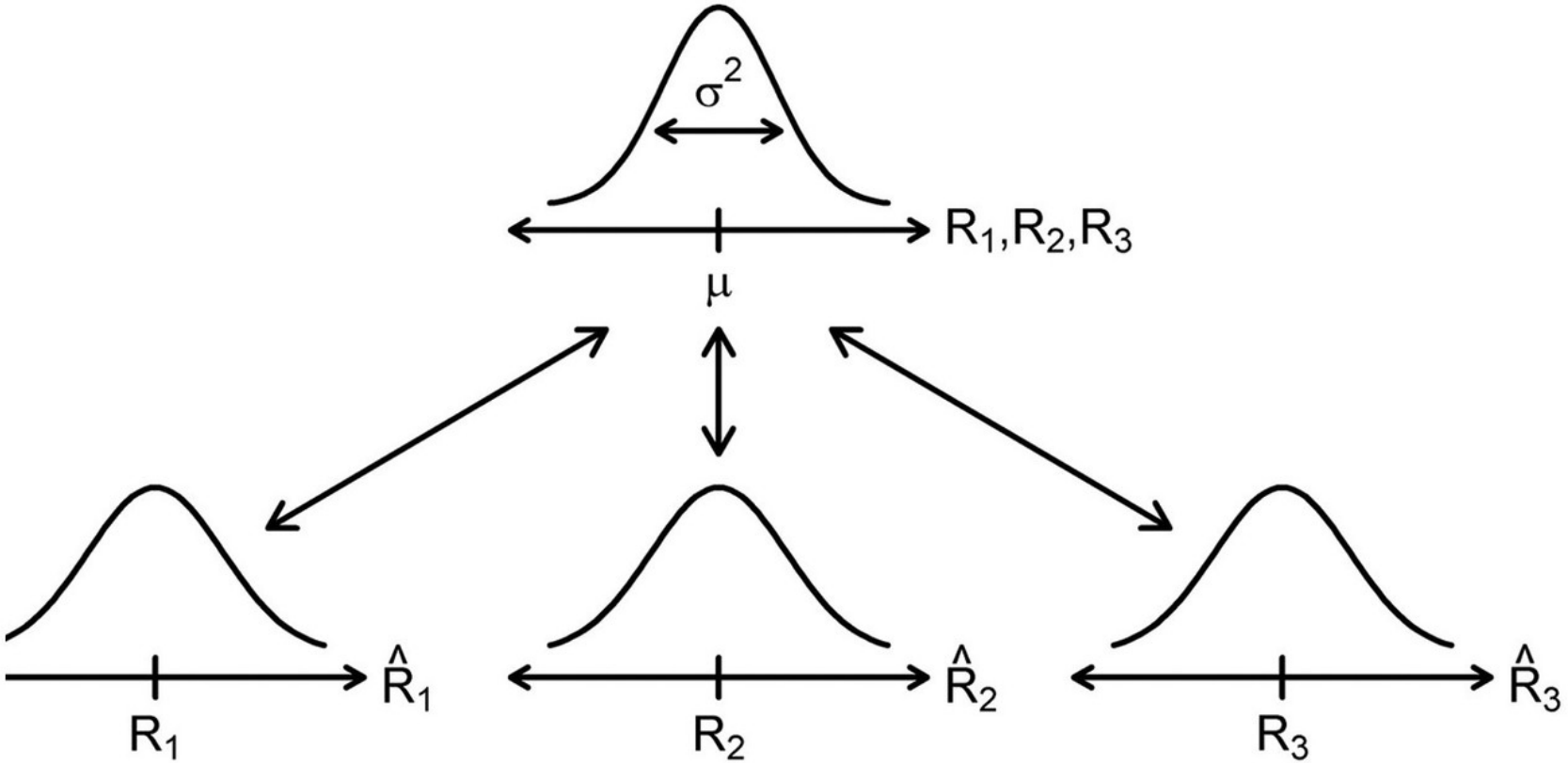
- In each experiment k, the estimated ATE / ITT etc $\hat{\tau}_k \sim f_k(\tau_k)$
- The τ_k themselves are drawn from some $g(\tau)$

The first point models any specific study, and the second models how different studies relate to each other. Internal and external validity

We see $\hat{\tau}_k$ for every k, and are interested in things about g - as this tells us what should happen with future models

The **external** stage is like the **prior**, the **internal** stage is like the **likelihood**. Generates a posterior over population parameters - a link between $\{\hat{\tau}_k\}$ and τ

Hierarchical Models II



Data from experiments $\{\hat{R}_1, \hat{R}_2, \hat{R}_3\}$ are used to find out things about the 'top level' distribution, such as μ, σ^2

Hierarchical Models III

Meager (2019)'s main specification is[1]:

$$y_{ik} \sim N(\mu_k + \tau_k T_{ik}, \sigma_y^2)$$
$$\begin{pmatrix} \mu_k \\ \tau_k \end{pmatrix} \sim N\left[\begin{pmatrix} \mu \\ \tau \end{pmatrix}, \begin{pmatrix} \sigma_\mu^2 & \sigma_{\tau\mu} \\ \sigma_{\tau\mu} & \sigma_\tau^2 \end{pmatrix}\right]$$

Where i is an observation, k is a study, μ_k is a control mean, τ_k is the (eg) ITT, y_{ik} is one of several outcomes (eg business profits) and T_{ik} is a treatment dummmy

Therefore: 1) different treatment and control means for every study 2) μ_k and τ_k can be correlated

[1] Eqn (2) in the paper

Hierarchical Models IV

This isn't the only way to compare studies. More classical methods include:

- Running pooled OLS estimates on all the data together with study-level fixed effects
- Assuming all τ_k are the same, or that all $\hat{\tau_k} = \tau_k$

Despite the parametric assumptions made here - it still nests pretty much all classical models as a special case[2]. She also runs a model forcing $\sigma_{\tau\mu} = 0$ for robustness

[2] Technical idea: OLS for β equivalent to MLE with normal errors

The Case for Bayesianism Here

Ultimately, we just want a model connecting studies together in a way that adds just the right amount of structure

- The classical Bias-Variance tradeoff

We want to have considerable flexibility, but only have $|k| = 7$

- The regularization provided by Bayesian Models allows us to estimate these with such a small sample

The priors aren't very controversial

- Using normal priors for errors $\approx$ OLS

Data

Data is from 7 field experiments on microcredit in 7 countries (Bosnia, Ethiopia, India, Mexico, Mongolia, Morocco, the Philippines)

All RCTs, with variation in who they treated, group / individual level, focused on women, interest rate of the loan, etc

Evaluated models like above

- Some forcing correlations to be 0, some not
- Some with covariates added, some not
- Running pooled OLS alongside everything to highlight differences in the approach, and for more robustness

Estimates of τ

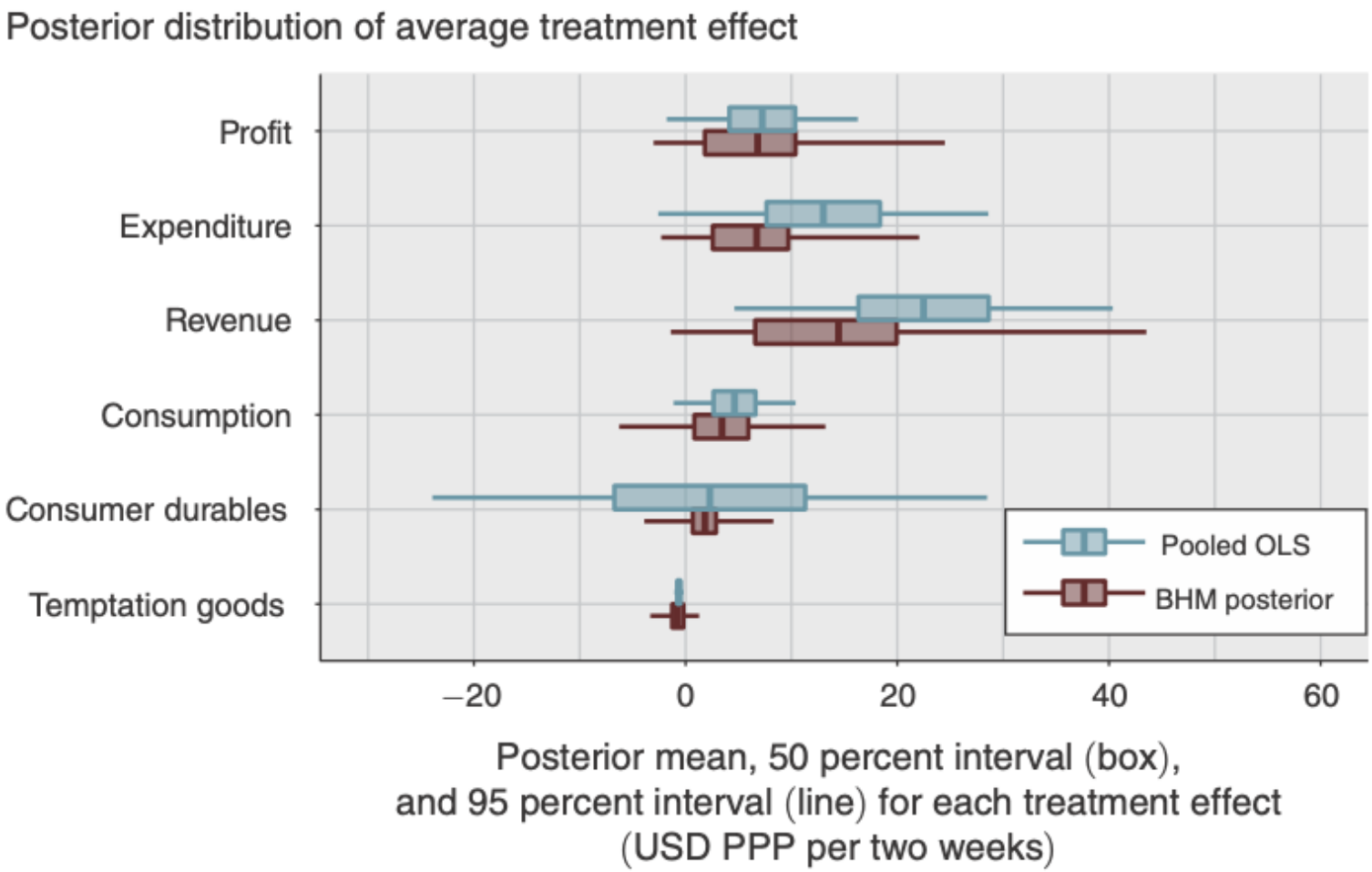


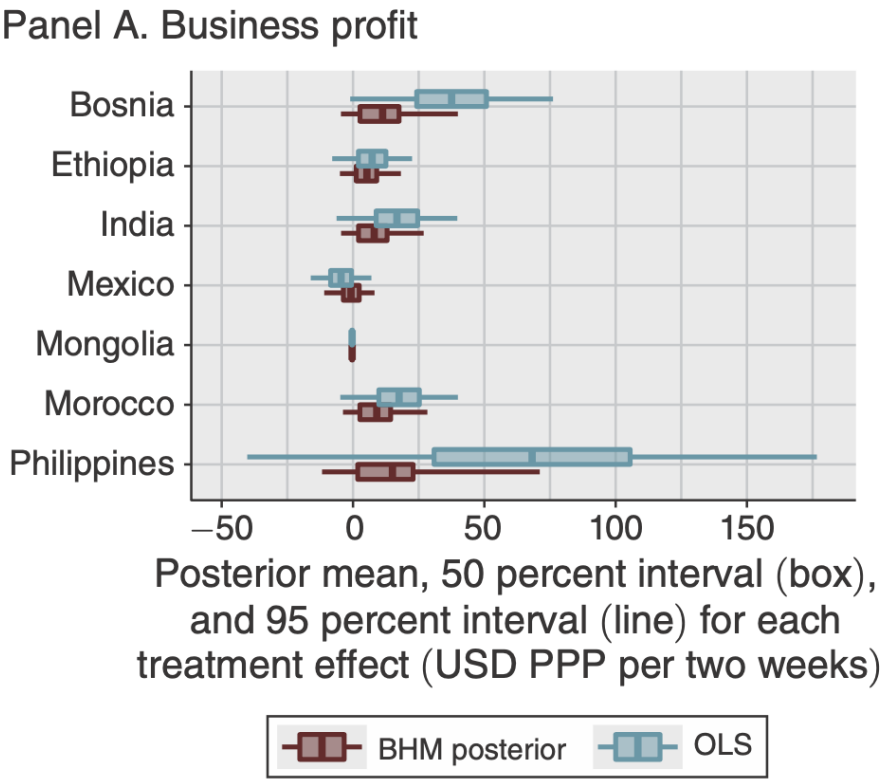
FIGURE 1. GRAPH OF POSTERIORS FOR EACH τ FROM THE MAIN SPECIFICATION OF THE JOINT BAYESIAN HIERARCHICAL MODEL (BHM), WITH THE FULL POOLING OLS INTERVALS FOR COMPARISON

τ summary

- *Summary:* very small but positive τ. Mean increase in profit is \$7 USD PPP every two weeks. Unlikely to change poverty much[3]
- Standard hypothesis tests give joint statistical significance; but doesn't hold up in this more flexible model
- Bayesian methods also give us better estimates of the individual τ_k - shrink them all towards $\hat{\tau}$ - so potentially more accurate (especially if there is low σ_τ^2)

[3] Though might due to poverty traps etc

Heterogeneity in τ_k[4]



- Much less heterogeneity in τ_k that maybe thought
- 60% of variation in $\hat{\tau}_k$ is explained by *sampling variation*, rather than genuine heterogeneity
- But lots of caveats about how generalizable things are

[4] Fig 3.1

Heterogeneity by Previous Business Ownership

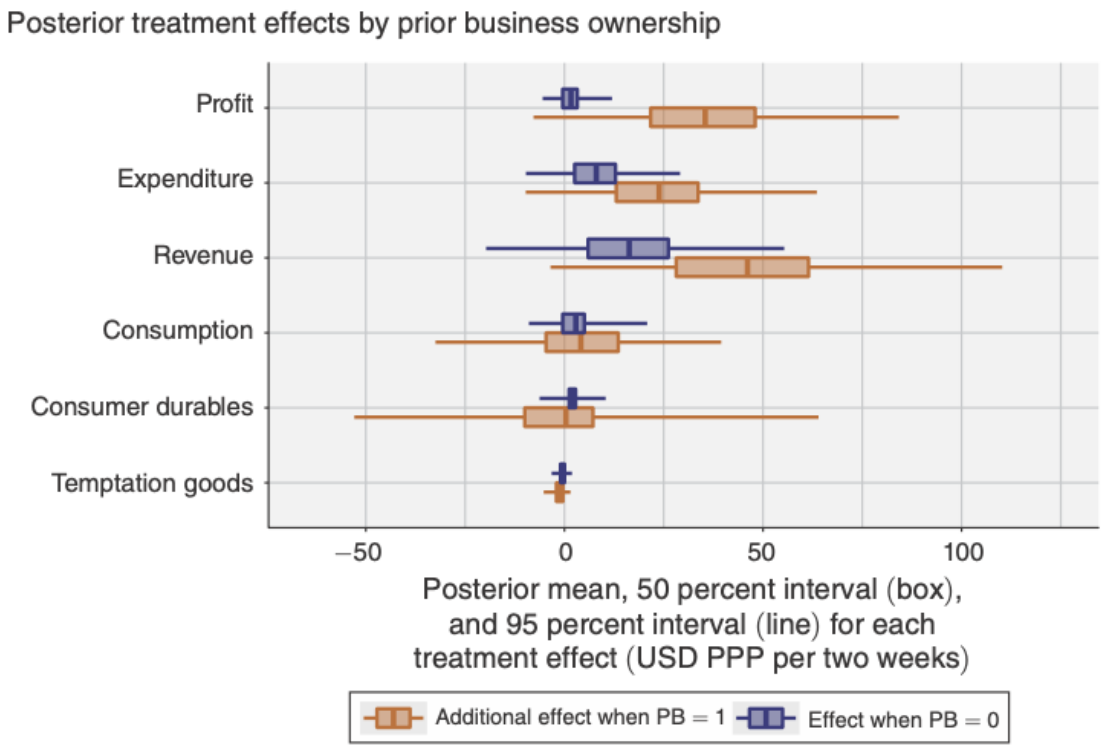
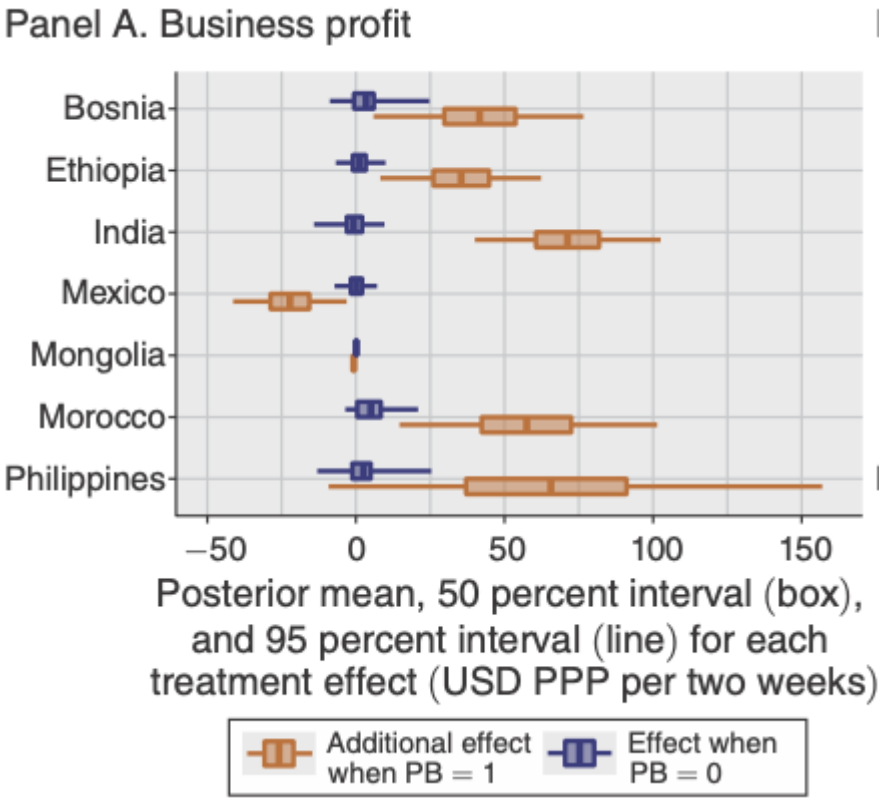


FIGURE 7. POSTERIOR DISTRIBUTIONS OF τ FOR ALL OUTCOMES SPLIT BY PRIOR BUSINESS OWNERSHIP

Grameen Bank claims 'Beggars can turn to business'. Substantial doubt

Profit Heterogeneity by Previous Business Ownership[5]



Those with previous business experience have

- Higher average profits
- But more hetrogenous profits

[5] Fig 9.1

What drives differences in τ_k?

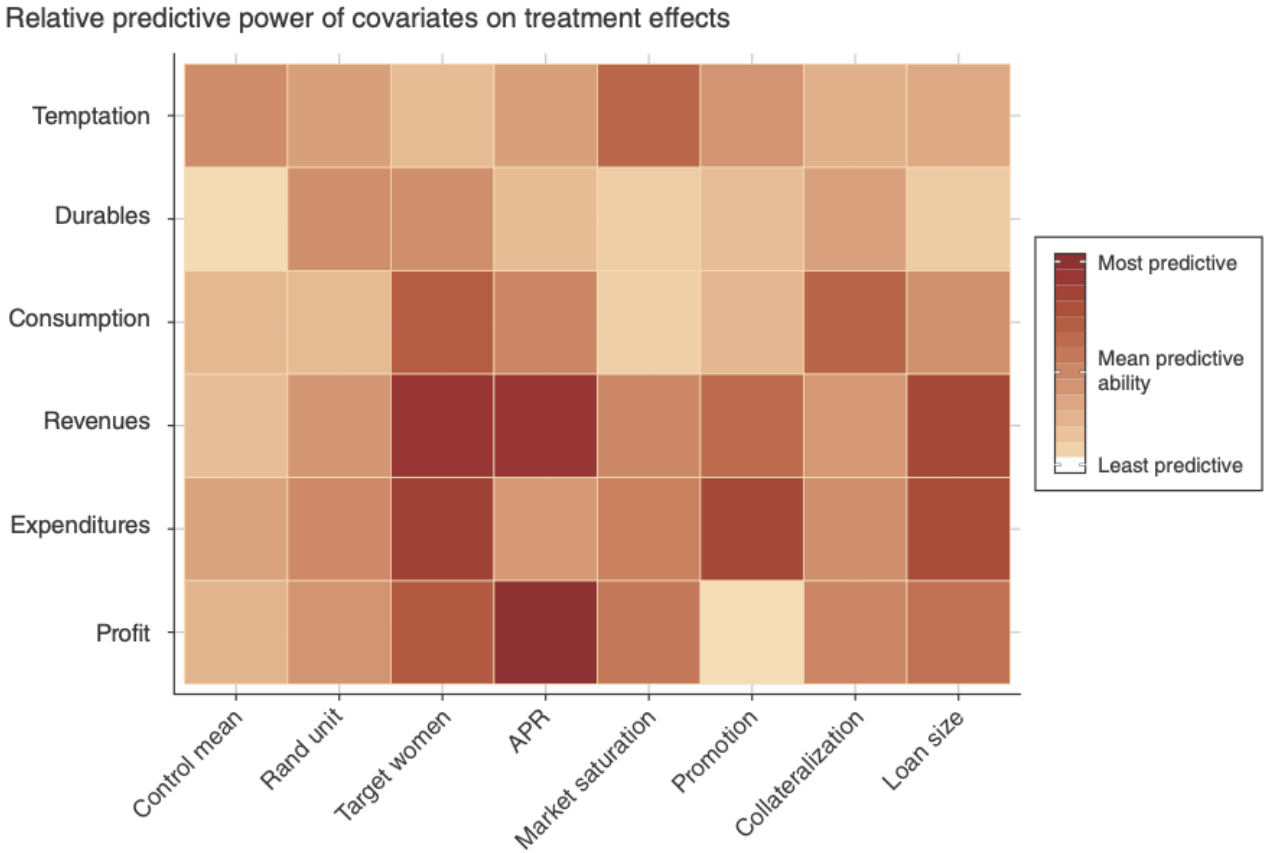


FIGURE 10. ABSOLUTE MAGNITUDE OF THE RIDGE REGRESSION COEFFICIENTS FOR ALL OUTCOMES AND COVARIATES

From a Ridge Regression of τ_k on covariates. Economic conditions are more predictive for variation in τ_k than experiment details (eg sampling method)

Conclusion

- Bayesian estimators $\approx$ shrinkage. Very useful when small amounts of data
- Hierarchical Models are estimated using Bayesian techniques. Allows you to compare studies both to get potentially more accurate estimates of each τ_k[6], and of population features such as τ
- These studies show microcredit can be somewhat beneficial (but hetrogenous) for those who own businesses already; not so much those who don't
- These results are quite robust over situations - and when they differ it's mostly because of Economic conditions

[6] By shrinkage, if you have good priors!