LABORATORY FOR INFORMATION & DECISION SYSTEMS

Sensitivity of MCMC-based analyses to small-data removal

Tin (Stan) Nguyen

Thesis Defense

May 7, 2024

Introduction

[Angelucci et al. 2015] is a randomized controlled trial (RCT), examining effect of microcredit, in Mexico

If we run MCMC on a Bayesian model, microcredit may be seen as reducing profit (“hurting”)

For policymaking, we want to know if findings *generalize* beyond our data

Idea: If conclusion changes after removing small data, we might be concerned about generalization [Broderick et al. 2020]

Our work shows: by removing 16 out of 16560 households, microcredit appears “helpful”

Problem: It is too computationally expensive to check every data subset

Idea: Approximate dropping worst-case data

Problem: Existing approximations e.g. [Broderick et al. 2020] does not apply to MCMC

Our contributions:

No approximation for MCMC → We extend [Broderick et al. 2020] to MCMC-based conclusions

MCMC analyses have Monte Carlo noise → We quantify this variability

Experiments analyzing the quality of the approximation in real data

Roadmap

Another reason to care about dropping data

How expensive is brute-force approach?

Setup for dropping data

Our approximation: (linear approximation + MCMC estimate) & confidence interval
– We show it is fast

Experiments from economics and ecology
– Our approximation performs well in a simple model
– Performance is mixed in a complex model

Another reason to care about dropping data

Problem: Do conclusions from a data analysis generalize?

Idea: Use standard generalization checks - confidence intervals (CI), p-values

Example: If CI is entirely < 0, analyst makes generalization i.e. at large, effect is negative

Problem: Real data deviates from the standard CI / p-value's working assumption of i.i.d.-ness

Hope: Deviations are small so that CI reflects generalization & conclusion holds

Idea: Validate this hope by checking if conclusion holds under deviations

A realistic deviation: a small data fraction α is missing

If removing α fraction changes conclusions, we might be worried about generalization

Definition of small is subjective (like a p-value threshold): our default is α = 1%

How expensive is brute-force approach?

There is a combinatorial explosion in leaving out every possible subset and re-run

An economist might be worried if removing 0.1% could change their conclusion

Dropping every 0.1% of microcredit data means enumerating over 10^54 things

If each run takes 1 minute, exhaustive search still takes > 10^48 years

Existing works [Broderick et al. 2020, Shiffman et al. 2023, Moitra et al. 2022, Freund et al. 2023] do not apply to MCMC

Setup for dropping data

For data $(\text{microcredit access}^{(n)}, \text{profit}^{(n)})_{n=1}^{N}$

E.g. $\text{profit}^{(n)} \sim \text{Gaussian}(\mu + \theta \times \text{microcredit access}^{(n)}, \sigma^2)$.

Log likelihood of the n-th data point is $L_n(\beta)$

A *prior* $p(\beta)$ encodes domain information

Posterior density is proportional to

$p(\beta) \prod_{n=1}^{N} \exp(L_n(\beta))$

A quantity of interest: ϕ.

MCMC draws ($\beta^{(1)}, \beta^{(2)}, \ldots, \beta^{(S)}$): $\mathbb{E}[\beta_d] \approx \frac{1}{S} \sum_{s=1}^{S} \beta_d^{(s)}$

Data weights: $(w_1, w_2, \ldots, w_N) =: w$

Weighted posterior has density proportional to $p(\beta) \prod_{n=1}^{N} \exp(w_n L_n(\beta))$

$w_n = 0$: n-th observation is dropped

Quantity of interest: $\phi(w)$.

Small-data sensitivity is a constrained optimization problem. WLOG, assume $\phi(\mathbf{1}) < 0$

Feasible set is $W_\alpha := \{w \in \{0,1\}^N : \sum_{n=1}^{N} (1 - w_n) \leq N\alpha\}$

If $\max_{w \in W_\alpha} \phi(w) > 0$, we might worry about generalization

Method part I: Taylor series & MCMC estimates

Goal: Fast approx. of worst-case posterior mean* $\max_{w \in W_\alpha} \phi(w)$

For estimating equations, [Broderick et al. 2020] sidesteps brute-force with a linear approximation

Idea to use linear approximation is still relevant beyond estimating equations

We replace posterior mean with a Taylor series: $\phi(w) - \phi(\mathbf{1}) \approx \sum_{n=1}^{N} (w_n - 1) \frac{\partial \phi}{\partial w_n}\Big|_{w=\mathbf{1}}$

While [Diaconis et al. 1986, Ruggeri et al. 1986, Gustafson 1996, Giordano et al. 2023, etc.] have known that derivatives are covariances, this relationship has not been used for small-data sensitivity

We know from past works: $\frac{\partial \phi}{\partial w_n}\Big|_{w=\mathbf{1}} = \mathrm{Cov}_{\mathbf{1}}(\beta_d, L_n)$

We estimate covariances:

We estimate linear approximation, $\sum_{n=1}^{N} (w_n - 1) \frac{\partial \phi}{\partial w_n}\Big|_{w=\mathbf{1}} \approx \sum_{n=1}^{N} (w_n - 1) \hat{\psi}_n$, and optimize

$\max_w \sum_{n=1}^{N} (w_n - 1) \hat{\psi}_n = \max_w \left(- \sum_{w_n = 0} \hat{\psi}_n \right) \longrightarrow$ **Algorithm:** Sort; Remove most extreme values

Our approximation is fast

Time complexity is O(N x S + N x log N) if we do not need to compute log likelihoods

In one analysis, while MCMC takes 12 hours, our approximation takes only two minutes

* Our method applies to other quantities of interest, too

Method part II: Quantify uncertainty

Our approximation encounters a type of error not faced by previous works: Monte Carlo noise

Goal: Estimate variability due to MCMC randomness

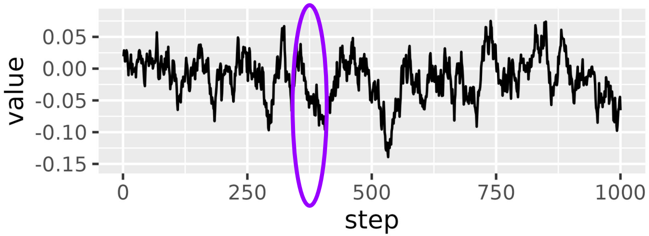
Our estimate is a function of random sample $\hat{\Delta}(\beta^{(1)}, \beta^{(2)}, \ldots, \beta^{(S)})$

If draws were i.i.d., use bootstrap [Efron 1979]

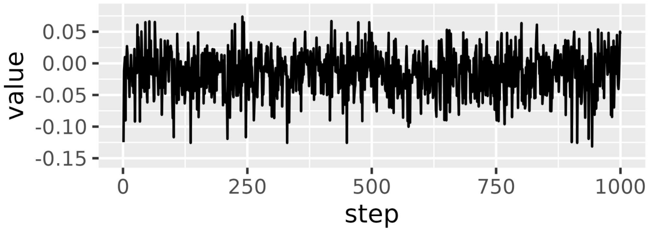
- Resample from $\beta^{(1)}, \beta^{(2)}, \ldots, \beta^{(S)}$: $(\beta^{*(1)}, \beta^{*(2)}, \ldots, \beta^{*(S)})$
- Use spread of $\hat{\Delta}(\beta^{*(1)}, \beta^{*(2)}, \ldots, \beta^{*(S)})$ as confidence interval

Generally, sample has time series dependence & bootstrap is expected to underperform

Original

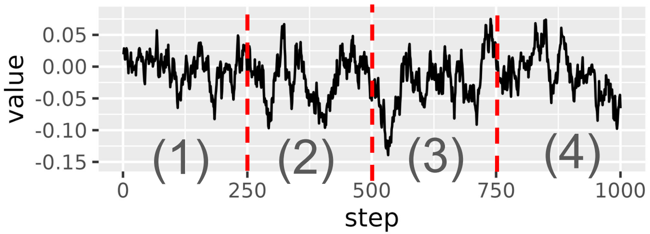


Bootstrap

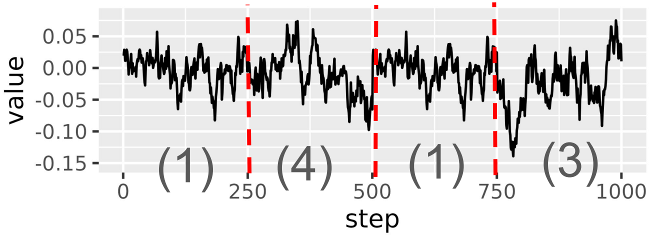


We use block bootstrap [Carlstein 1986] to handle time series dependence

Original



Block bootstrap



This resampling scheme has one parameter: the block length

On a simple model, our approximation works well

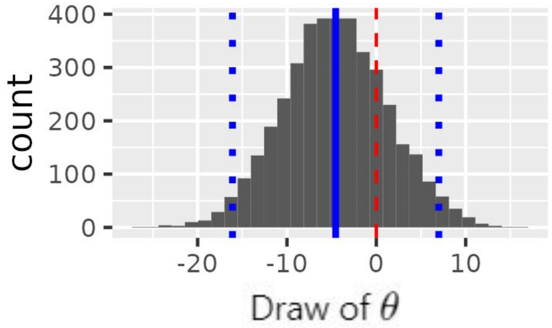
We consider a variant of analysis from [Meager 2019] & [Meager 2022] *

$$\text{profit}^{(n)} \sim \text{Gaussian}(\mu + \theta \times \text{microcredit access}^{(n)}, \sigma^2)$$

We define wide priors and estimate effect with MCMC

Running MCMC takes 3 minutes

Microcredit might have a negative effect, but it is not conclusive



It takes 2 seconds to assess sensitivity

Prediction range (bars) contain the refit (x)

We predict sign change after removing 0.10%

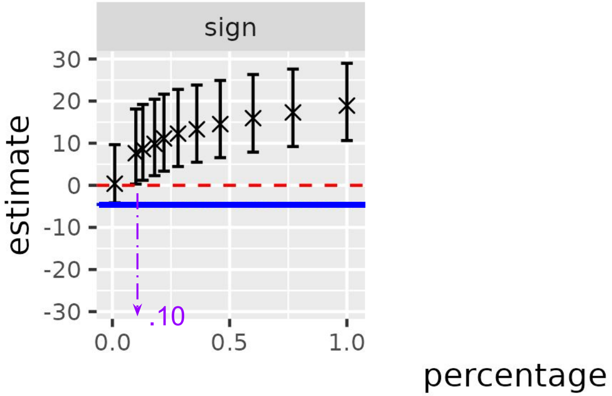
Refit confirms prediction

Each refit takes 3 minutes

We predict sig. change after removing 0.36%

Refit confirms prediction

We are not able to predict if a positive and sig. effect is possible



* [Meager 2022] also analyzes microcredit using different data and a more complex Bayesian model. Our paper contains a sensitivity analysis of that model, too

Performance on a complex model is mixed

[Senf et al. 2020] regresses ``tree death'' on ``water balance''

Linear predictor involves many parameters

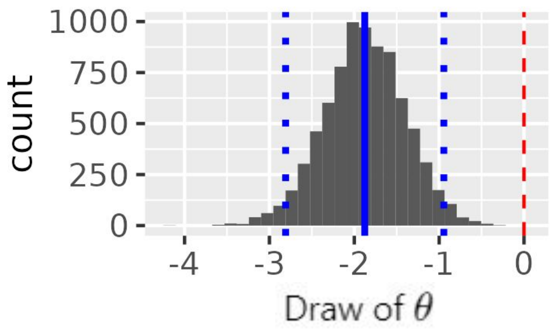
Population: $\mu + \theta \times \text{water balance}^{(n)}$

Regional: $\mu_r^{(\text{region})} + \theta_r^{(\text{region})} \times \text{water balance}^{(n)}$

~ 6000 regional parameters are organized hierarchically

Running MCMC takes 12 hours

Water balance has (-) and sig. association



It takes only 2 minutes to assess sensitivity

We predict sign change at 0.17%

Each refit takes 12 hours

Change actually happens at 0.22%

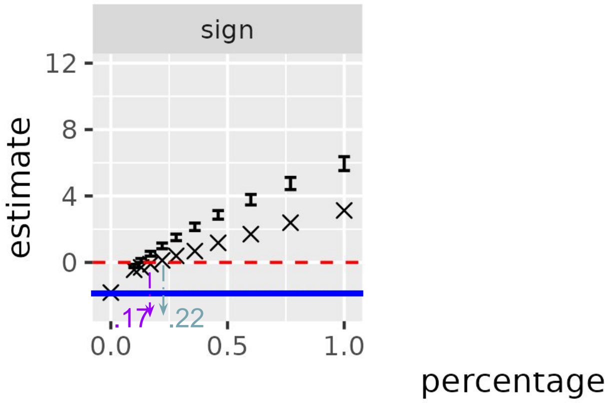
We predict sig. change at 0.10%

Change does happen

Our method predicts (+) and sig. link at 0.17%

Change actually happens at 1%

Prediction (bars) is more extreme than realized by refitting (x)



Confidence interval quality across MCMC randomness

Ideal: How often does confidence interval (CI) contain worst-case quantity of interest?

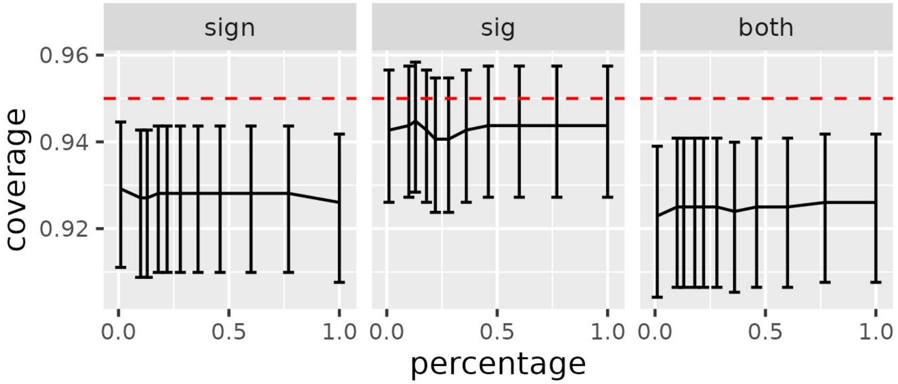
Partial answer: How often does CI contain result of linear approx.? $-\sum_{n \in I} \mathrm{Cov}_{\mathbf{1}}(\beta_d, L_n)$

We estimate CI coverage with another level of Monte Carlo

We run 960 Markov chains → averaging gives high-quality est. of $-\sum_{n \in I} \mathrm{Cov}_{\mathbf{1}}(\beta_d, L_n)$

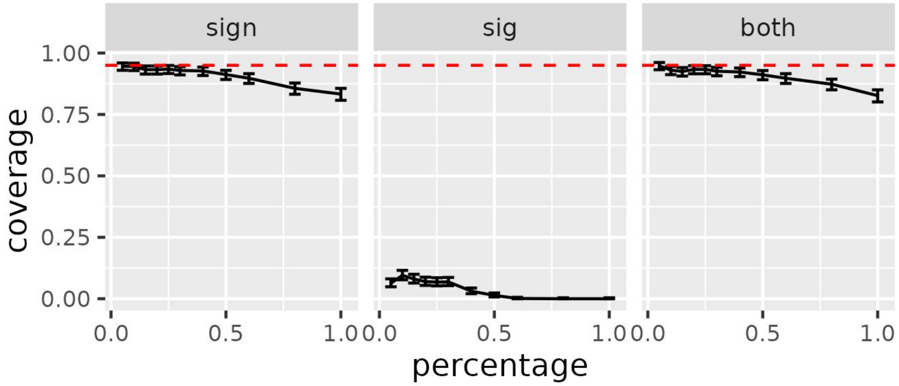
We run 960 Markov chains → averaging gives high-quality est. of CI coverage

In simple model, confidence interval (CI) contains ground truth with adequate frequency



Estimate of coverage is very close to nominal 95%

In complex model, CI can have very poor coverage of ground truth*



Estimate of coverage can be very far from nominal 95%

*We subsample 2,000 observations from the original ~80,000 observations

Why is the coverage in the complex model poor?

(Recall)

We resample blocks from $\beta^{(1)}, \beta^{(2)}, \ldots, \beta^{(S)}$ to generate $(\beta^{*(1)}, \beta^{*(2)}, \ldots, \beta^{*(S)})$

We use interquantile range of $\hat{\Delta}(\beta^{*(1)}, \beta^{*(2)}, \ldots, \beta^{*(S)})$ as confidence interval

Calculation of $\hat{\Delta}$ involves a sort i.e. $\hat{\psi}_{(1)} \leq \hat{\psi}_{(2)} \leq \ldots \leq \hat{\psi}_{(N)}$ & $\hat{\Delta}$ = negative of sum of extremes

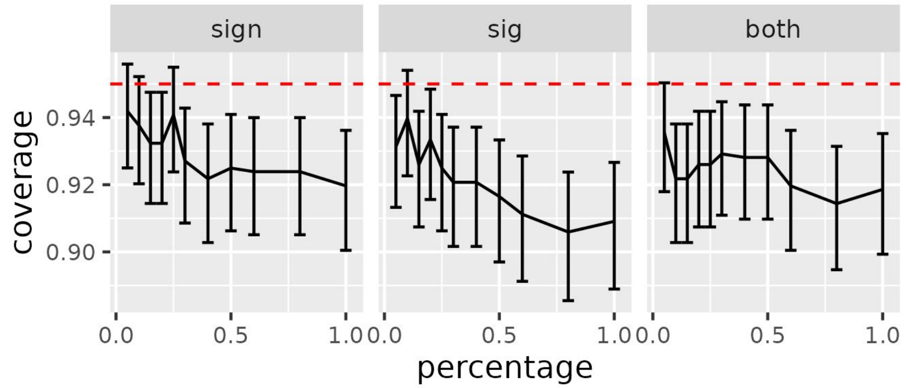
Sorting is non-smooth

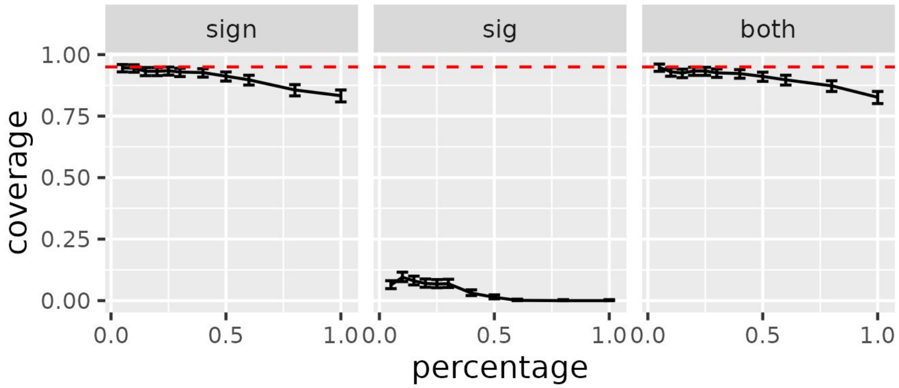
Suspicion: sorting creates complex dependencies that cause poor coverage

To test, we consider a version of $\hat{\Delta}$ that does not involve sorting i.e. $\sum_{n \in I} \hat{\psi}_n$ for fixed I

If CI from resampling $\sum_{n \in I} \hat{\psi}_n$ covers $\sum_{n \in I} \mathrm{Cov}_{\mathbf{1}}(\beta_d, L_n)$ well, we attribute issue to sorting

CI for ``fixed-indices" is adequate → *Severe underperformance is due to sorting*





Future work

- Set problem-dependent block length
- Extend to posterior quantiles
- Identify the source of difficulty in complex models (many params. or hierarchy?)

Summary

We have developed & tested a fast approximation for the removal of worst-case small data in MCMC-based analyses

We will arXiv this work soon!

Thesis theme: Faster methods for Bayesian unsupervised learning

Existing works aim to speed up Bayes through parallelism

Problem: They struggle due to so-called label-switching problem

Solution: I use a representation that evades the problem to derive fast & accurate estimates

Tin Nguyen, Brian L. Trippe, Tamara Broderick (2022). Many processors, little time: MCMC for partitions via optimal transport couplings. In *AISTATS 2022*.

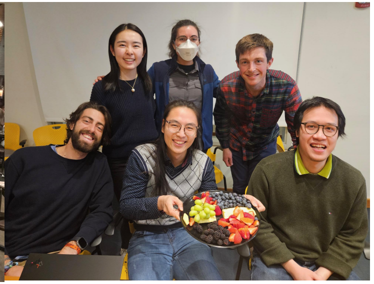
Bayesian nonparametrics posit a countable infinity of latent traits

Problem: Computers cannot learn a countable infinity of things

Solution: I derive accurate and easy-to-use finite approximations

Tin Nguyen, Jonathan Huggins, Lorenzo Masoero, Lester Mackey, Tamara Broderick (2023). Independent finite approximations for Bayesian nonparametric inference. Bayesian Analysis Advance Publication.

I dedicate this thesis to you!

… Broderick lab …



… my family …

… my friends …





References

Manuela Angelucci, Dean Karlan, Jonathan Zinman, Kerry Brennan, Ellen Degnan, Alissa Fishbane, Andrew Hillis, Hideto Koizumi, Elana Safran, Rachel Strohm, Braulio Torres, Asya Troychansky, Irene Velez, Glynis Startz, Sanjeev Swamy, Matthew White, Anna York, and Compartamos Banco. Microcredit impacts: Evidence from a randomized microcredit program placement experiment by compartamos banco. American Economic Journal: Applied Economics, 7:151–82, 2015.

Cornelius Senf, Allan Buras, Christian S. Zang, Anja Rammig, and Rupert Seidl. Excess forest mortality is consistently linked to drought across Europe. Nature Communications, 11, 12 2020.

Tamara Broderick, Ryan Giordano, and Rachael Meager. An automatic finite-sample robustness metric: Can dropping a little data change conclusions? 2020

Miriam Shiffman, Ryan Giordano, and Tamara Broderick. Could dropping a few cells change the takeaways from differential expression? 2023

Ankur Moitra and Dhruv Rohatgi. Provably auditing ordinary least squares in low dimensions. 2022

Daniel Freund and Samuel B. Hopkins. Towards practical robustness auditing for linear regression. 2022

Bradley Efron. Bootstrap methods: Another look at the jackknife. Annals of Statistics, 1979.

Edward Carlstein. The use of subseries values for estimating the variance of a general statistic from a stationary sequence. Annals of Statistics, 1986.

Rachael Meager. Understanding the average impact of microcredit expansions: A bayesian hierarchical analysis of seven randomized experiments. American Economic Journal: Applied Economics, 11:57–91, 2019.

Terry C. Jones, Guido Biele, Barbara M¨uhlemann, Talitha Veith, Julia Schneider, J¨orn Beheim-Schwarzbach, Tobias Bleicker, Julia Tesch, Marie Luisa Schmidt, Leif Erik Sander, Florian Kurth, Peter Menzel, Rolf Schwarzer, Marta Zuchowski, J¨org Hofmann, Andi Krumbholz, Angela Stein, Anke Edelmann, Victor Max Corman, and Christian Drosten. Estimating infectiousness throughout sars-cov-2 infection course. Science. 2021.

Rachael Meager. Aggregating distributional treatment effects: A bayesian hierarchical analysis of the microcredit literature. American Economic Review, 2022.

Tenelle Porter, Diego Catal´an Molina, Andrei Cimpian, Sylvia Roberts, Afiya Fredericks, Lisa S. Blackwell, and Kali Trzesniewski. Growth-mindset intervention delivered by teachers boosts achievement in early adolescence. Psychological Science, 2022.

Persi Diaconis and David Freedman. On the consistency of bayes estimates. The Annals of Statistics, 1986.

Fabrizio Ruggeri and Larry Wasserman. Infinitesimal sensitivity of posterior distributions. The Canadian Journal of Statistic, 1993

Paul Gustafson. Local sensitivity of posterior expectations. The Annals of Statistics, 1996

Ryan Giordano and Tamara Broderick. The bayesian infinitesimal jackknife for variance. 2023.