Adaptive Pareto Smoothed Importance Sampling

William Ruth

Université de Montréal

Who am I?

- PhD SFU (2023)
- Postdoc UdeM (present)

- Computational Statistics
  - Simulation and related methods
  - Latent variable models
- Infectious disease modelling

Topics

- Adaptive Pareto Smoothed Importance Sampling
- Multilevel Causal Mediation Analysis
- Modelling Tuberculosis in Foreign-Born Canadians

Topics

- **Adaptive Pareto Smoothed Importance Sampling**
- Multilevel Causal Mediation Analysis
- Modelling Tuberculosis in Foreign-Born Canadians

Outline

- Importance sampling
- Measuring performance
- Improving performance
    - Modifications
    - Optimization

Importance Sampling

- Need to compute an expected value
  - $\mathbb{E}_F \varphi(X)$
- Can't do the sum/integral
- Monte Carlo approximation
  - Simulating from F might be hard

Importance Sampling

- Introduce "proposal distribution", G:

$$\mathbb{E}_F \varphi(X) = \mathbb{E}_G \left[\varphi(X) \cdot \frac{f(X)}{g(X)} \right]$$
$$= \mathbb{E}_G \left[\varphi(X) \cdot w(X) \right]$$

- G can be nearly anything*
  - *Some choices will be better than others

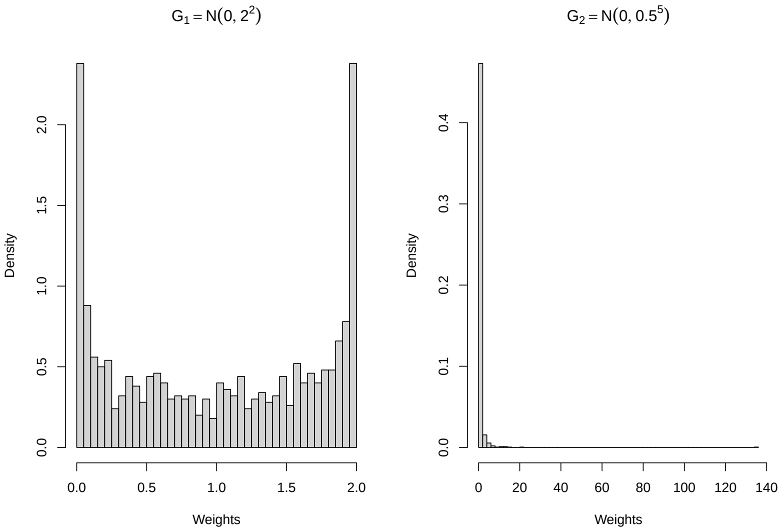
Example: Mystery Target

- f unknown, but can be evaluated
- $\varphi(X) = X^2$

- Try some proposals:
  - $G_1 \sim N(0, 2^2)$
  - $G_2 \sim N(0, 0.5^2)$

- Use $M = 1000$ samples from proposal
  - $\hat{\mathbb{E}}_1 = 0.99$, $\hat{\mathrm{SD}} = 1.97$
  - $\hat{\mathbb{E}}_2 = 1.10$, $\hat{\mathrm{SD}} = 2.32$

Example: Mystery Target



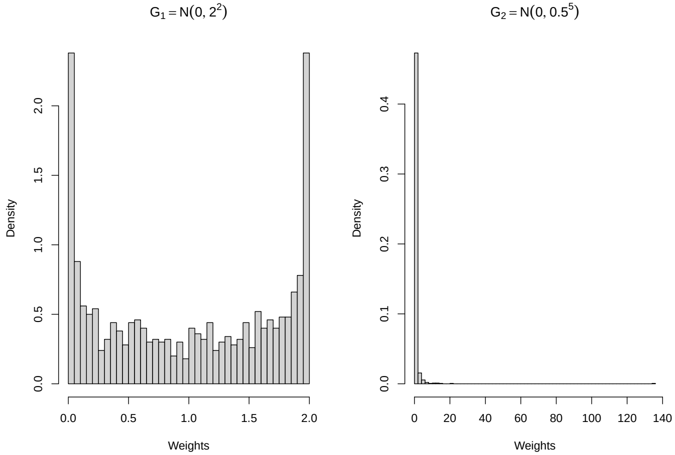
Importance Sampling

- G_1 weights look fine
- G_2 weights dominated by one large value

- We can make this difference precise
- "Effective Sample Size":

$$ESS = \frac{\left[\sum_i w(X_i)\right]^2}{\sum_i w(X_i)^2}$$

Example: Mystery Target



$ESS_1 \approx 662$ $\quad$ $ESS_2 \approx 54$

Importance Sampling

- Problem: Low ESS $\rightarrow$ hard to estimate means
- But ESS is based on means
  - (Chatterjee and Diaconis, 2018)

Improving IS

- Choose a good proposal
- Modify large weights
  - Truncated IS
  - Pareto Smoothed IS

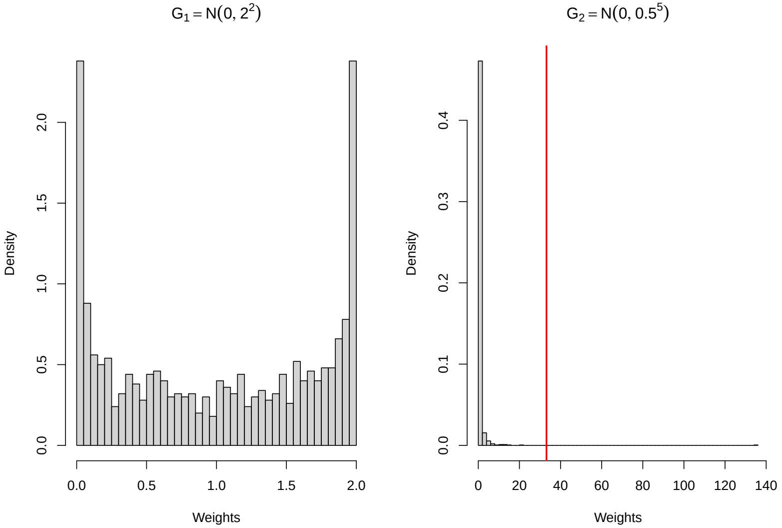
Improving IS

- Truncated Importance Sampling:
    - (Ionides, 2008)

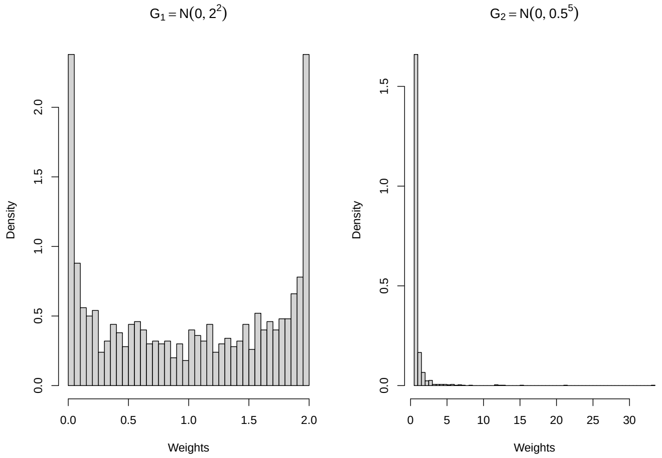
1. Choose a threshold
2. Apply hard thresholding to any large weights

- Still consistent for the target

Example: Mystery Target



Example: Mystery Target



$$ESS_1 \approx 662$$
$$ESS_1^{(\text{trunc})} \approx 662$$

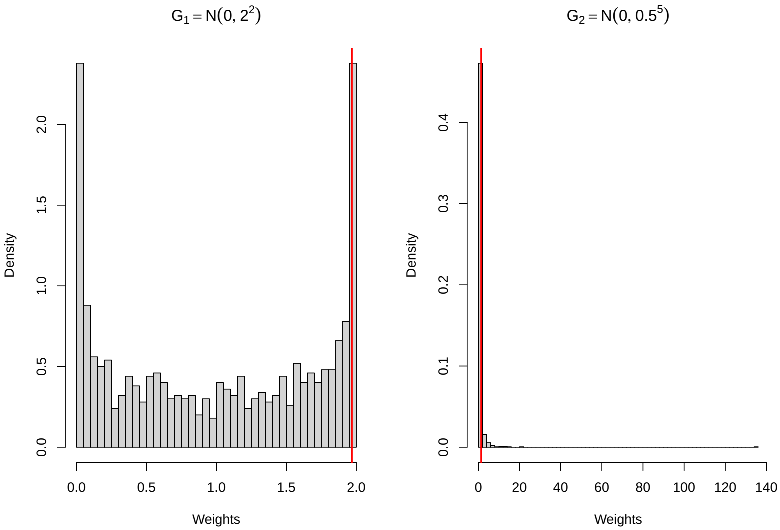
$$ESS_2 \approx 54$$
$$ESS_2^{(\text{trunc})} \approx 245$$

Improving IS

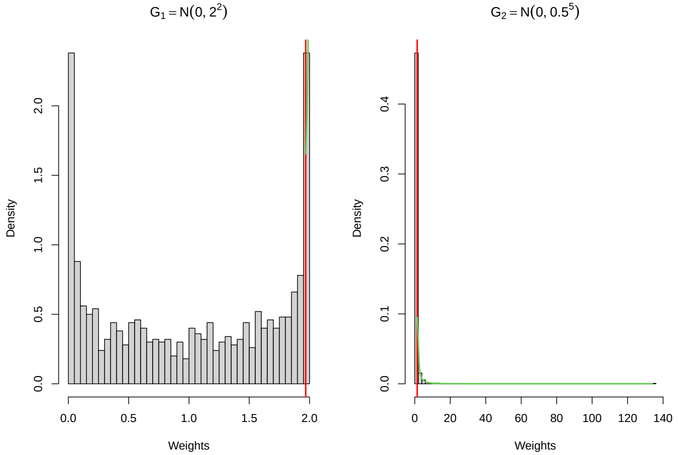
- Pareto Smoothed Importance Sampling:
  - (Vehtari et al., 2024)

1. Choose a threshold
   - Weights above threshold represent tail of their dist.
2. Approximate tail with Generalized Pareto Dist.
   - Fit GPD to weights above threshold
   - (Zhang and Stephens, 2009)
3. Replace large weights with quantiles of fitted GPD

Example: Mystery Target

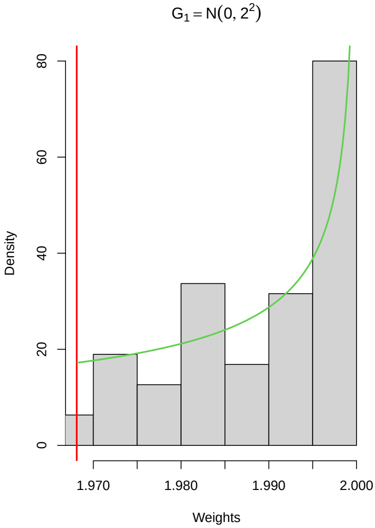


Example: Mystery Target

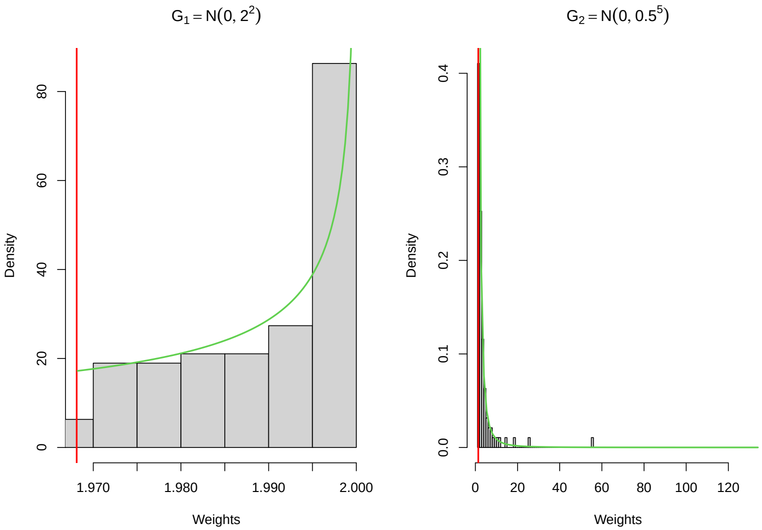


$\hat{k}_1 \approx -1.81$ $\hat{k}_2 \approx 0.72$

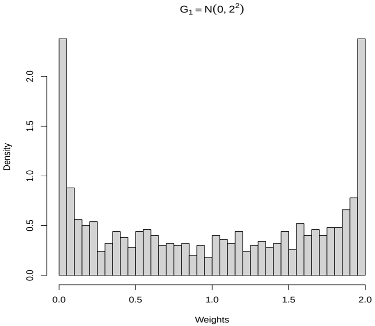
Example: Mystery Target

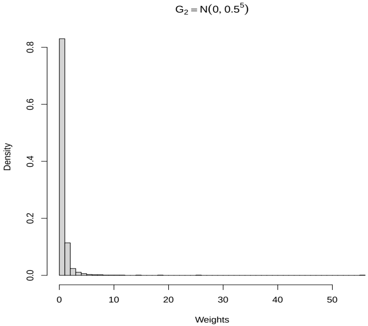


Example: Mystery Target



Example: Mystery Target





$$ESS_1 \approx 662$$
$$ESS_1^{(\text{trunc})} \approx 662$$
$$ESS_1^{(\text{PS})} \approx 662$$

$$ESS_2 \approx 54$$
$$ESS_2^{(\text{trunc})} \approx 245$$
$$ESS_2^{(\text{PS})} \approx 160$$

Adaptive IS

- Alternative approach: directly optimize ESS
- Adaptive Importance Sampling:
  - (Akyildiz and Míguez, 2021)

1. Choose a (parametric) family of proposals
2. Iteratively update the proposal to maximize ESS

Stochastic Approximation

- Actually, minimize a population-level analog:
  - $\rho = \mathbb{E}_G w^2(X) \approx \frac{N}{ESS}$
- If we had ρ, we would do gradient descent
  - $\theta_{k+1} = \theta_k - \alpha_k \nabla \rho(\theta_k)$
- Instead, do gradient descent on $\hat{\rho}$
  - $\hat{\theta}_{k+1} = \hat{\theta}_k - \alpha_k \nabla \hat{\rho}(\hat{\theta}_k)$
- **Stochastic approximation**

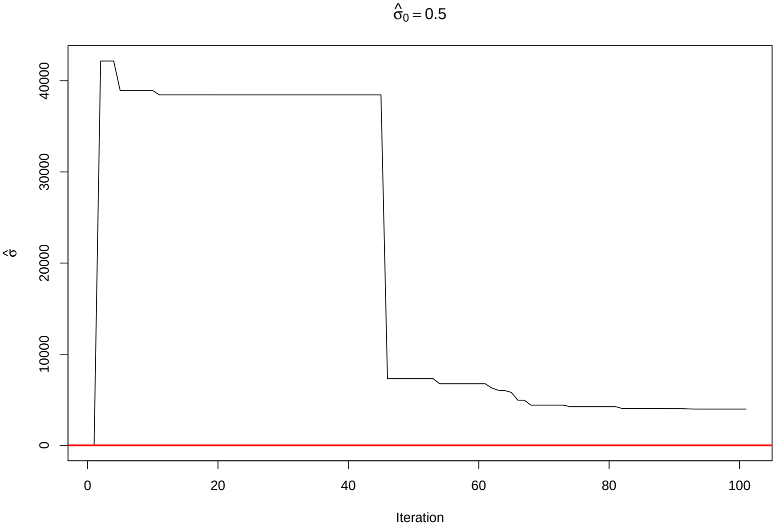
Stochastic Approximation

- Originally developed for root finding with noise
  - (Robbins and Monro, 1951)
- Quickly adapted for optimization
  - Use noisy evaluations for finite difference
  - (Kiefer and Wolfowitz, 1952)

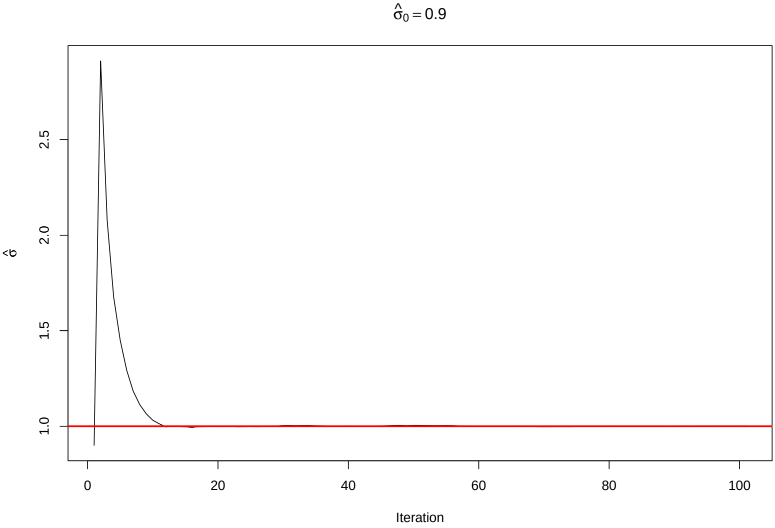
Stochastic Approximation

- Very well developed theory
- Step size $\to 0$
  - Called the "learning rate"
- Stochastic gradient descent
  - Popular in machine learning
  - Resample a (very) large dataset

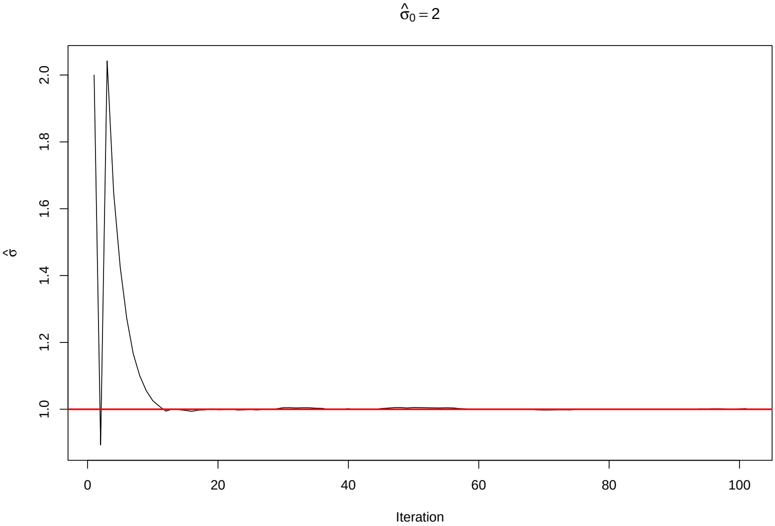
Example: Mystery Target



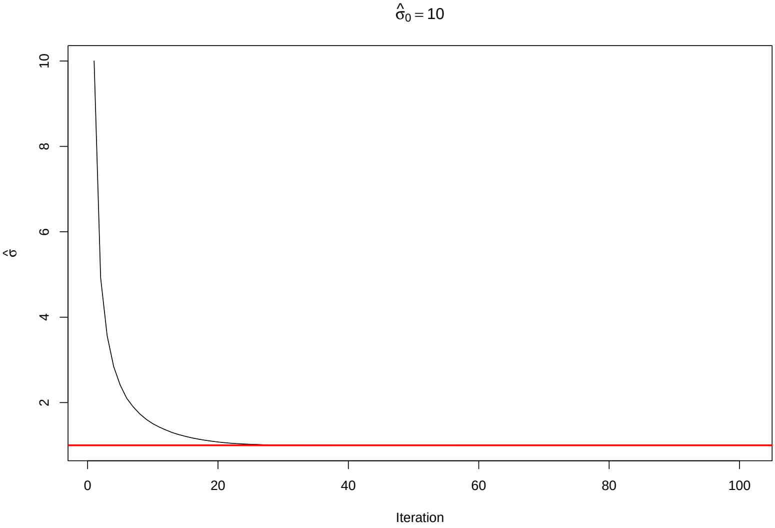
Example: Mystery Target



Example: Mystery Target



Example: Mystery Target



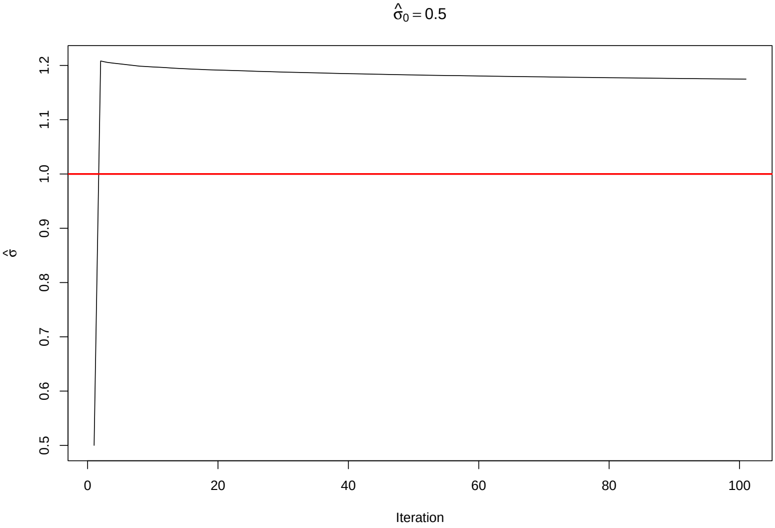
Our Method

- Recall: Be careful using IS means to diagnose IS
- Vehtari et al. give an alternative
  - Shape parameter of fitted tail distribution, $\hat{k}$

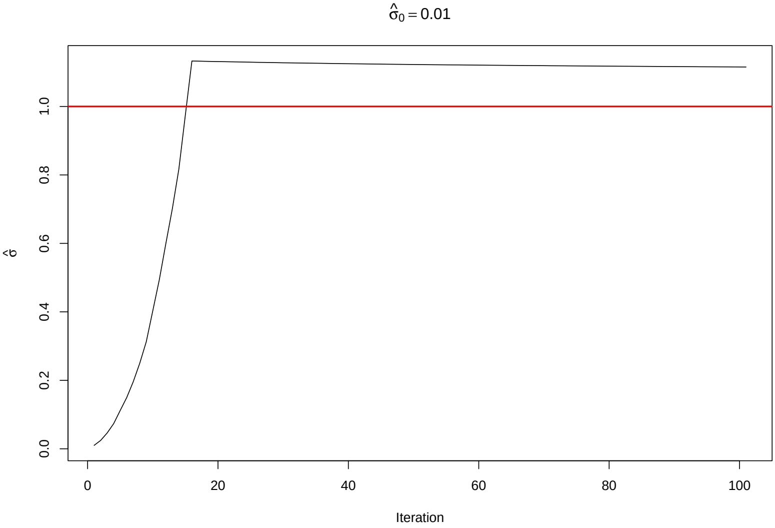
Our Method

- Use diagnostic as objective function
- Apply stochastic approximation to minimize $\hat{k}$
  - More precisely, its population analog: $k(\theta)$
- Use finite difference approximation to $\hat{k}'(\theta)$
  - This is subtle

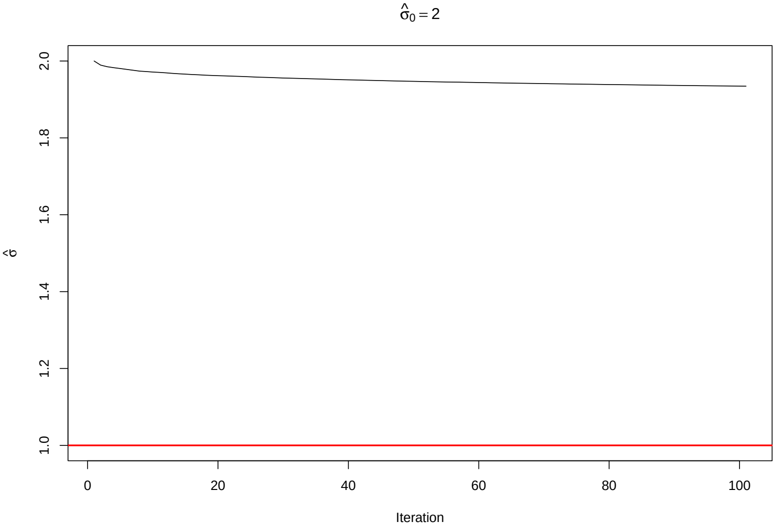
Example: Mystery Target



Example: Mystery Target



Example: Mystery Target



Our Method - Future Directions

- Refining the finite difference approximation
  - Generalize ESS version outside exponential families

- Analytical tail indices
- Convergence theory for stochastic approximation

- Applications
  - Latent variable models (e.g. GLMMs)
  - Bayesian inference in high-dimensions

Recap

- Importance sampling and extensions
  - Truncation
  - Pareto Smoothing
- Diagnostics for importance sampling
  - Effective sample size
  - Pareto tail index
- Adaptive importance sampling
  - Stochastic approximation

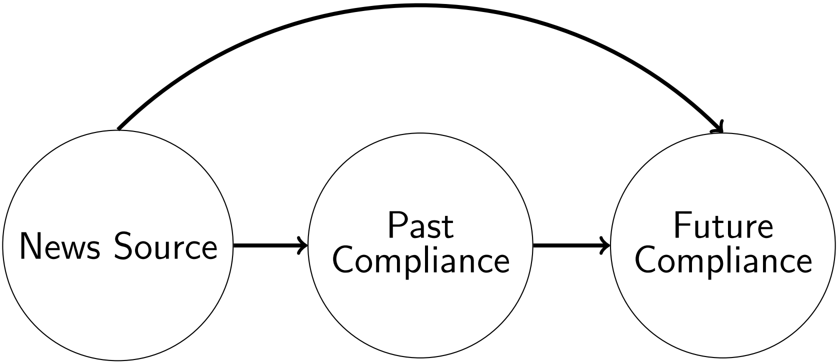
Topics

- Adaptive Pareto Smoothed Importance Sampling
- **Multilevel Causal Mediation Analysis**
- Modelling Tuberculosis in Foreign-Born Canadians

Example

- Goal: Understand adherence to restrictive measures
  - E.g. Lockdowns
  - Both past and future
- Influence of news source
  - How trustworthy?
- Disentangle influence on future from influence on past

Example



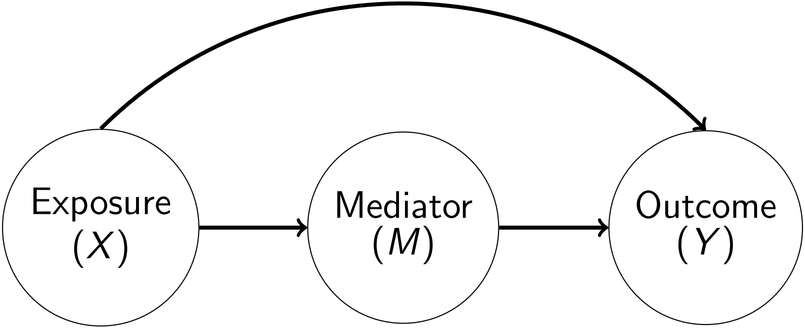
Example

Terminology

- Top path: Direct effect
- Center path: Indirect effect
- Combined: Total effect

- Exposure: X
- Outcome: Y
- Mediator: M

Mediation Analysis



Mediation Analysis

- Separate **Total Effect** of X on Y into
  - **Direct Effect**
  - **Indirect Effect**
- Define effects formally using counterfactuals

Mediation Analysis

- Under "identification assumptions", expected counterfactuals equal conditional expectations
- Use regression
  - Linear vs logistic
  - Single-level vs multi-level

Multi-Level Models

- Grouped data
    - E.g. By country
- Mediation effects differ by group
    - Mixed-effects regression

Multilevel Mediation Analysis

- Uncertainty quantification is messy
  - Bootstrap
  - Quasi-Bayesian Monte Carlo
  - δ-method

- Monte Carlo studies look promising

Topics

- Adaptive Pareto Smoothed Importance Sampling
- Multilevel Causal Mediation Analysis
- **Modelling Tuberculosis in Foreign-Born Canadians**

Tuberculosis

- Massive problem worldwide
  - Prevalence of 20-25%
  - (Cohen et al., 2019)
- Infection includes a latent period
  - Can last months or entire lifetime

Tuberculosis in Canada

- Relatively rare in Canada
- Mostly present in foreign-born and indigenous communities
  - We focus on foreign-born
- Immigration screens for active TB but not latent

Tuberculosis in Canada

- We model TB in foreign-born Canadians using a system of ODEs
    - Includes immigration and domestic transmission
- Some parameters from literature
- Others fit using data
- Not on track to meet our goal of 90% reduction by 2035

Acknowledgements

Collaborators:

- Payman Nickcki (UBC), Richard Lockhart (SFU)
- Bouchra Nasri (UdeM), Bruno Remillard (HEC), Rado Ramasy (UdeM), Rowin Alfaro (UdeM)
- Jeremy Chiu (SFU), Albert Wong (Langara)

Funding:

- CANSSI

Thank You

Some References

Akyildiz, Ö. D. and Míguez, J. (2021). Convergence rates for optimized adaptive importance samplers. *Statistics and Computing*, 31(12).

Chatterjee, S. and Diaconis, P. (2018). The sample size required in importance sampling. *The Annals of Applied Probability*, 28(2).

Cohen, A., Mathiasen, V. D., and Schön, T. (2019). The global prevalence of latent tuberculosis: a systematic review and meta-analysis. *European Respiratory Journal*, 53(3).

Ionides, E. L. (2008). Truncated importance sampling. *Journal of Computational and Graphical Statistics*, 17(2).

Kiefer, J. and Wolfowitz, J. (1952). Stochastic estimation of the maximum of a regression function. *The Annals of Mathematical Statistics*, 23(3).

Robbins, H. and Monro, S. (1951). A stochastic approximation method. *The Annals of Mathematical Statistics*, 22(3).

Vehtari, A., Simpson, D., Gelman, A., Yao, Y., and Gabry, J. (2022). Pareto smoothed importance sampling. *ArXiv*.

Vehtari, A., Simpson, D., Gelman, A., Yao, Y., and Gabry, J. (2024). Pareto smoothed importance sampling. *Journal of Machine Learning Research*, 25(72).

Zhang, J. and Stephens, M. A. (2009). A new and efficient estimation method for the generalized Pareto distribution. *Technometrics*, 51(3).