Adaptive Pareto Smoothed Importance Sampling

William Ruth

Université de Montréal

Who am I?

- PhD SFU 2023
- Postdoc UdeM (present)

- Computational Statistics
  - Simulation
  - Simulation-based inference
- 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

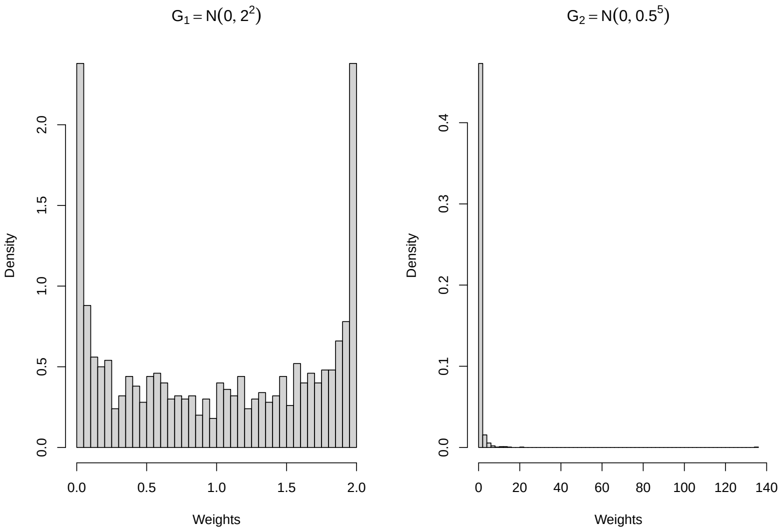
Example: Mystery Target

- 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
    - $\hat{\mathbb{E}}_1 = 0.99$, $\hat{\text{SD}} = 1.97$
    - $\hat{\mathbb{E}}_2 = 1.10$, $\hat{\text{SD}} = 2.32$

Example: Mystery Target



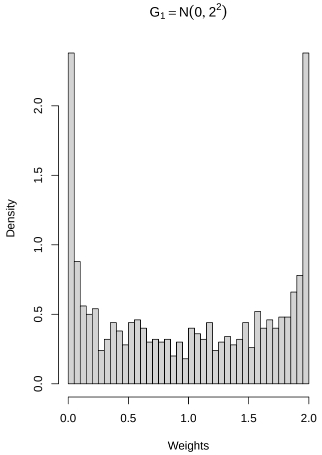
Importance Sampling

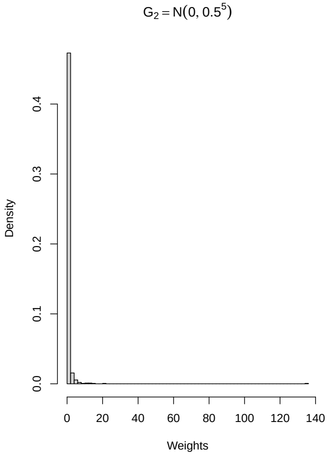
- G_1 weights look fine
- G_2 weights dominated by one large value

- 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





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

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

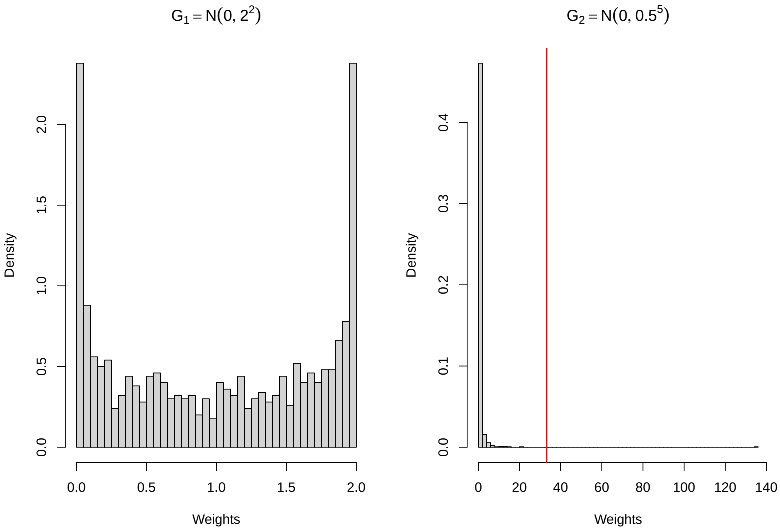
Improving IS

- Truncated Importance Sampling:
    - (Ionides, 2008)

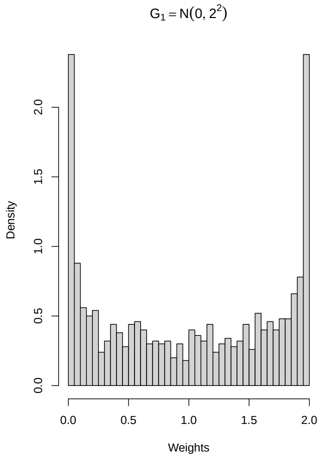
1. Choose a threshold
2. Apply hard thresholding to any large weights

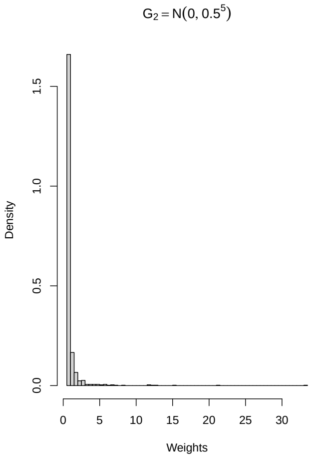
- Still consistent for the target

Example: Mystery Target



Example: Mystery Target





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

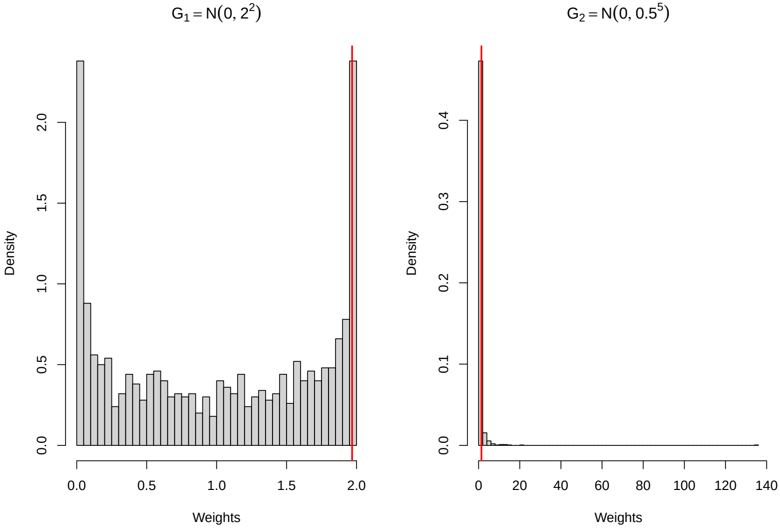
$$ESS_2 \approx 54$$
$$ESS_2^{(\text{trunc})} \approx 245$$

Improving IS

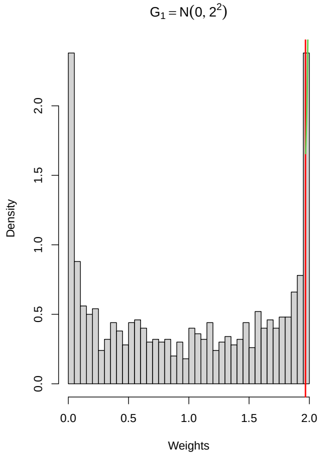
- Pareto Smoothed Importance Sampling:
    - (Vehtari et al., 2024)

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

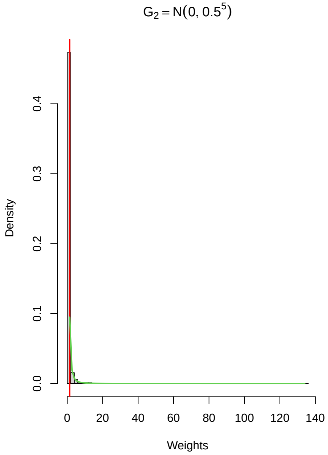
Example: Mystery Target



Example: Mystery Target

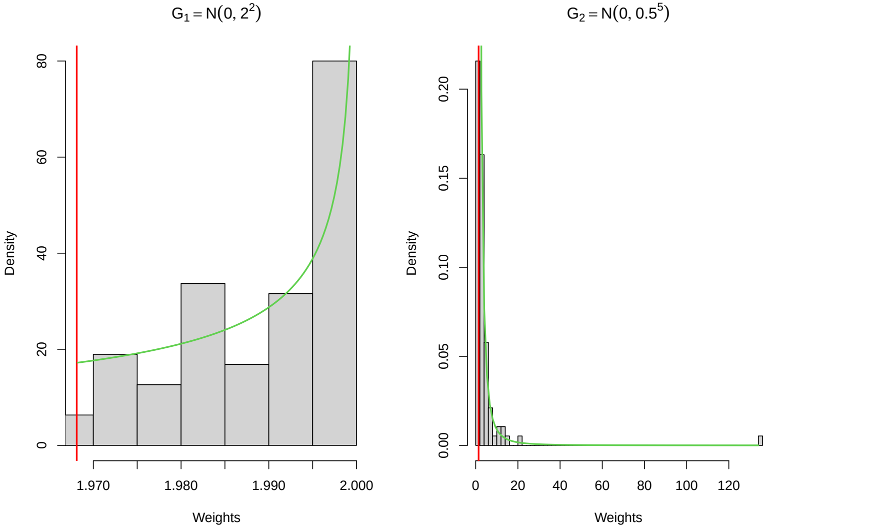


$\hat{k}_1 \approx -1.81$

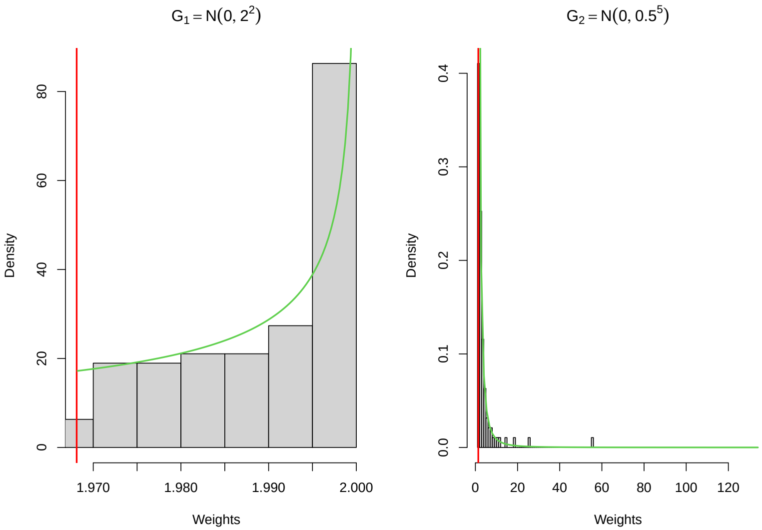


$\hat{k}_2 \approx 0.72$

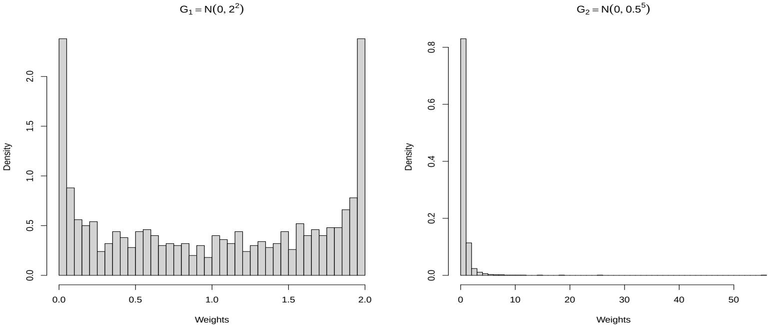
Example: Mystery Target



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

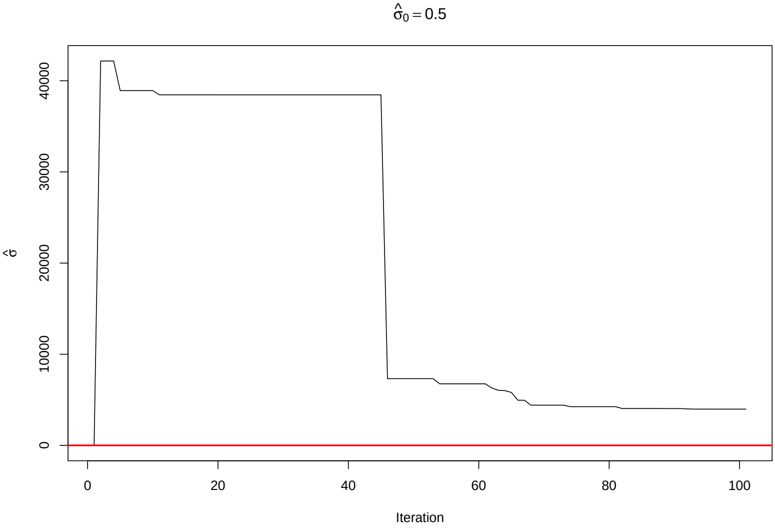
Stochastic Approximation

- 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)

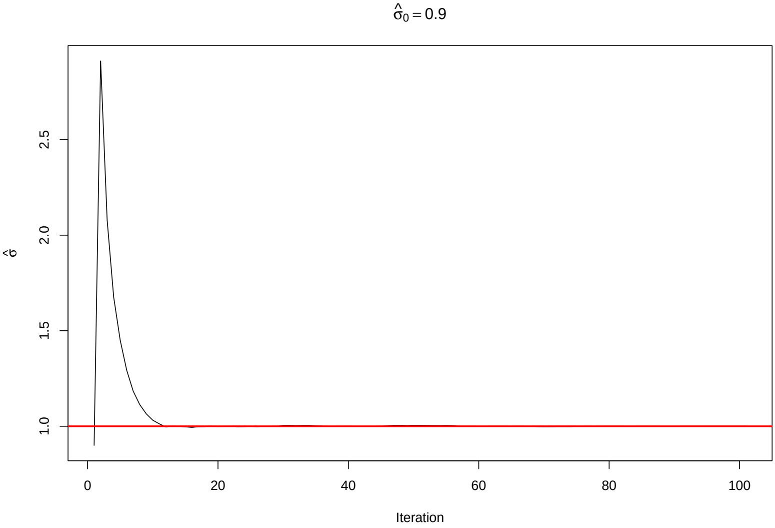
Stochastic Approximation

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

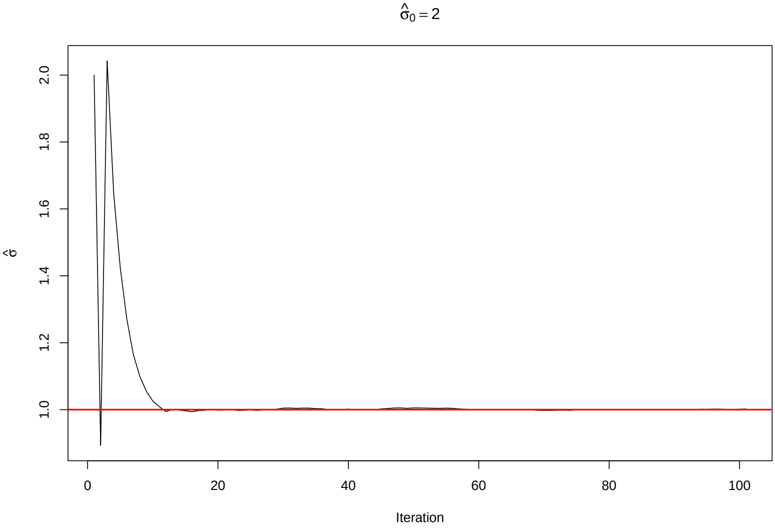
Example: Mystery Target



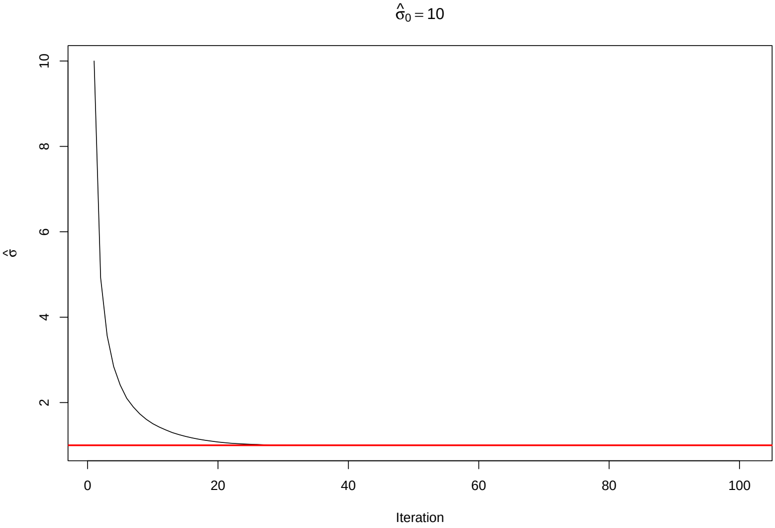
Example: Mystery Target



Example: Mystery Target



Example: Mystery Target



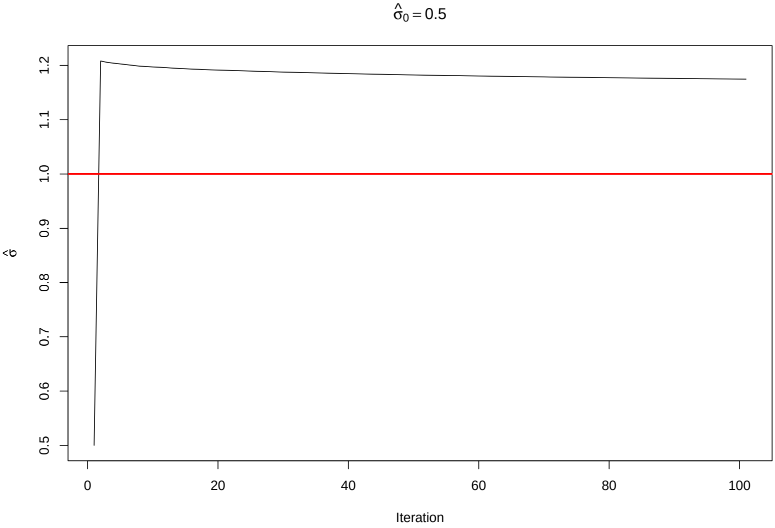
Our Method

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

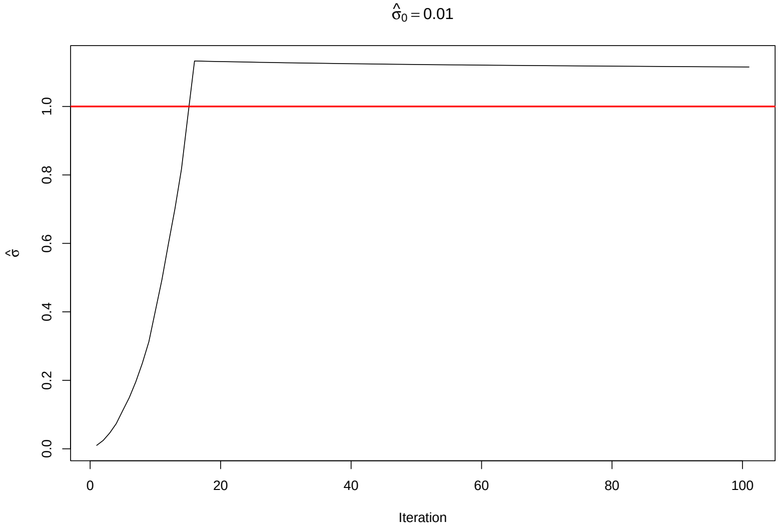
Our Method

- 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

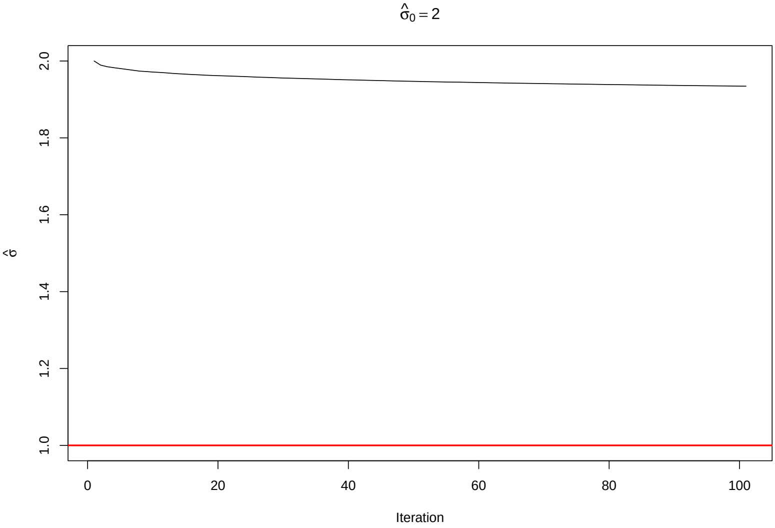
Example: Mystery Target



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

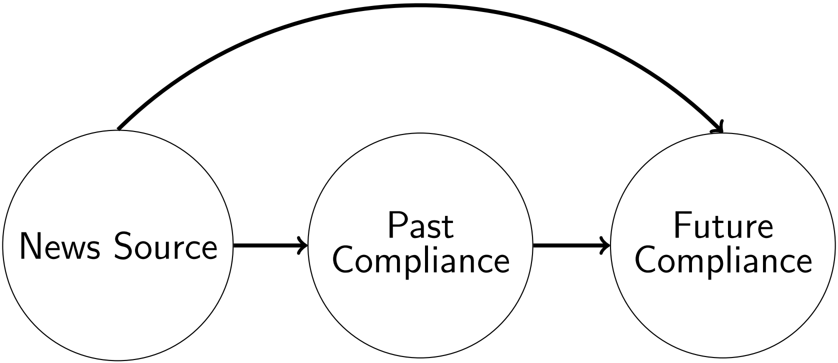
- Adaptive Pareto Smoothed Importance Sampling
- **Multilevel Causal Mediation Analysis**
- Modelling Tuberculosis in Foreign-Born Canadians

Outline

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



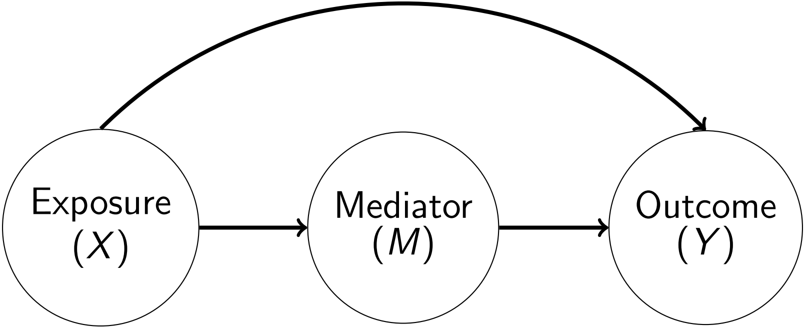
Example

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**

Traditionally, use regression

Mediation Analysis

Continuous outcome and mediator:

- $Y = \alpha_0 + \alpha_1 M + \alpha_2 X + \varepsilon_Y$
- $M = \beta_0 + \beta_1 X + \varepsilon_M$

Direct Effect: α_2

- "X in Y"

Indirect Effect: $\alpha_1 \cdot \beta_1$

- "M in Y" $\cdot$ "X in M"

Total Effect: $\alpha_2 + \alpha_1 \cdot \beta_1$

Mediation Analysis

Popular approach

- A bit outdated...

More popular: Causal mediation analysis

Causal Inference

Assume that X *causes* Y

Counterfactuals:

- What value would Y take if X were set to a particular level?
- Write Y_x for the value of Y when $X = x$
- If $X \neq x$ then Y_x is literally a "counterfactual"

Causal Inference

Example:

- Alice only reads scientific publications and will follow all lockdown mandates
- What if she instead only read Facebook?

- $Y_{Science}(Alice) = \text{follow}$
- $Y_{Facebook}(Alice) = \text{follow}$

Causal Inference

Example:

- Bob also only reads scientific publications and will follow all lockdown mandates, but is more susceptible to being influenced

- $Y_{Science}(Bob) = \text{follow}$
- $Y_{Facebook}(Bob) = \text{not follow}$

Causal Inference

- We only observe one outcome per individual
- Explore population-level effects by averaging
- Define mediation effects in terms of expected counterfactuals

Causal Inference

Total Effect: $\mathbb{E}(Y_{x'} - Y_x)$

- Effect on outcome when we change exposure from $X = x$ to $X = x'$

Other effects involve dependence on a mediator:

- Y_{xm}: Value of outcome when
  - Exposure (X) is set to x
  - Mediator (M) is set to m
- M_x: Value of mediator when
  - Exposure (X) is set to x
- "Nested Counterfactuals": Y_{xM_x} or $Y_{xM_{x'}}$

Causal Mediation Analysis

Controlled Direct Effect: $\mathbb{E}(Y_{x'm} - Y_{xm})$

- Effect of changing exposure with mediator held fixed

Natural Direct Effect: $\mathbb{E}(Y_{x'M_x} - Y_{xM_x})$

- Effect of changing exposure when we don't interfere with the mediator

Natural Indirect Effect: $\mathbb{E}(Y_{xM_{x'}} - Y_{xM_x})$

- Effect of changing which exposure value is seen by the mediator while holding fixed which exposure value is seen by the outcome

Causal Mediation Analysis

In our example

- Controlled Direct Effect: Effect of increasing news trustworthiness if the whole population followed guidelines in the past
- Natural Direct Effect: Effect of increasing news trustworthiness independent of any induced change in past compliance
- Natural Indirect Effect: Effect of changing past compliance if everyone only got news from Facebook

Causal Mediation Analysis

We can't measure all required counterfactuals

- E.g., Y_x or $Y_{x'}$, not both

Expected counterfactuals related to conditional expectations

- Under "identification" assumptions, $\mathbb{E}Y_x = \mathbb{E}(Y|X = x)$

Causal Mediation Analysis

How does causality change our analysis?

Still fit regression models, but include interaction terms between exposure and mediator

- $Y = \alpha_0 + \alpha_1 M + \alpha_2 X + \alpha_3 M \cdot X + \varepsilon_Y$
- $M = \beta_0 + \beta_1 X + \varepsilon_M$

Direct and indirect effects now depend on the levels of the exposure

Causal Mediation Analysis – Extensions

Discussion so far has involved continuous mediator and outcome

- What about binary?

Individuals might also be clustered

- E.g. Within countries

Causal Mediation Analysis – Extensions

- Handling binary variables is pretty straightforward
  - Instead of linear regression, use logistic regression
- Expected counterfactuals are now probabilities
- Might use risk-ratios or odds-ratios
- Dependence on regression coefficients becomes more non-linear

Causal Mediation Analysis – Extensions

- Clustered data more complicated
- Standard approach is multi-level modelling
  - I.e. Add random effects which vary across clusters
- Combined with categorical variables:
  - Generalized linear mixed models (GLMMs)

Clustered data more complicated

Mixed-Effects Models

The core idea is to augment our set of covariates

- Coefficients of these new covariates are random variables that vary across groups/clusters

In the linear setting:

- Old model: $Y = \alpha_0 + \alpha_1 X_1 + \ldots + \alpha_p X_p + \varepsilon$
- New model:
  $Y = \alpha_0 + \alpha_1 X_1 + \ldots + \alpha_p X_p + \mathbf{u_1 Z_1} + \ldots + \mathbf{u_q Z_q} + \varepsilon$

Mixed-Effects Models

The Z's are fixed, known covariates The u's are random variables

- I.e. Random effects

It's possible for the X's and Z's to overlap

- The coefficient on such a covariate has the form $\alpha_j + u_k$
- I.e. Mixed effect

Mixed-Effects Models

Extend to generalized linear models in the usual way

Linear predictor now has a random effects component

Mixed-Effects Models

Why bother?

- E.g. Measured some but not all levels of a categorical variable
- Estimate covariance matrix of random effects
- Test for non-zero variance of each random effect

"Predict" level of random effects for each group

- Conditional mean or conditional mode of random effects given response

Mixed-Effects Models

In our example:

- Data collected from 11 different countries
- Explicitly model inter-country variability
- Predict country-specific random effects
- Use country-specific coefficients in formulas for mediation effects
- Test for significant mediation effects within each country

Multilevel Mediation Analysis

Uncertainty quantification for mixed-effects models can be challenging

Strategies include:

- Bootstrap
- Quasi-Bayesian Monte Carlo
- δ-method

Multilevel Mediation Analysis

Uncertainty quantification for mixed-effects models can be challenging

Strategies include:

- Bootstrap
- Quasi-Bayesian Monte Carlo
- **δ-method**

Multilevel Mediation Analysis

- Mediation effects defined using nested counterfactuals – $Y_{xM_{x'}}$
- Expected nested counterfactuals expressed in terms of regression parameters
  - Coefficients and random effect covariances
- δ-method maps uncertainty in regression parameters to mediation effects

Multilevel Mediation Analysis

- Start with asymptotic covariance of regression parameters
  - Made possible by the *glmmTMB* package in `R`
- Pre- and post-multiply by Jacobian
  - Regression parameters to expected counterfactuals
  - Expected counterfactuals to mediation effects

Putting it All Together

- Define direct, indirect and total effects using counterfactuals
- Estimate these effects across groups using generalized linear mixed models
- Compute standard errors for estimated effects using the δ-method

Topics

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

Topics

- Give brief overview and mention directions for future research
- One of the profs in the department, Cristina Anton, does numerical SDEs. Mention potential collaboration

Acknowledgements

Collaborators:

- Payman (UBC), Richard (SFU)
- Bouchra (UdeM), Bruno (HEC), Rado (UdeM), Rowin (UdeM)
- Jeremy (SFU), Albert (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).

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).