How to simulate rare events and why it is important

Z. Botev

University of New South Wales

Nov 2015

Introduction to Gender Pay Gap

- What is the cause of the gender wage gap? Is it the result of some malevolent discrimination in hiring, or perhaps mostly the result of the antagonism between work and family commitments?
- We revisit the famous *University of Michigan Panel Study of Income Dynamics* from 1975, the height of the women's movement in the USA.
- The sample consists of $m = 753$ married white women between the ages of 30 and 60, 428 of whom worked for a wage outside the home; 325 of the women worked zero hours.
- The dependent variable $\boldsymbol{y}$ represents the wife's annual hours of work. For the women who worked positive hours, the range is fairly broad, extending from 12 to 4950.

Wife's Annual Wage Data

The $d - 1 = 7$ explanatory variables $x_1, \ldots, x_7$ include [1]:

1. number of kids less than 6 years in the family
2. number of kids between 6 and 18 years old.
3. woman's age
4. woman's education in years
5. actual labor market experience in years
6. the non-wife income (household income minus wife's income)
7. actual labor market experience in years squared (will look for a nonlinear effect)

[1] T. A. Mroz, *Econometrica: Journal of the Econometric Society*, Vol. 55, No. 4 (Jul., 1987), pp. 765-799

Tobit (Bayesian) Model

- Assuming that the response is normal is problematic as some women work zero hours.
- To account for the zero working hours, the model for the response $\boldsymbol{y}$ includes censoring:

$$Y_i = \begin{cases} W_i, & \text{if } u_i < W_i \\ b_i, & \text{if } W_i \leq b_i \end{cases}, \qquad \boldsymbol{W} \sim \mathrm{N}(\mathrm{X}\boldsymbol{\beta}, \sigma^2 I)$$

  where X is the matrix with 7 predictors and $b_i = 0$.
- We need to infer the model parameters $(\boldsymbol{\beta}, \sigma)$ from the data.

Bayesian posterior

Given for the data $\boldsymbol{y}$ and with uninformative priors, say $p(\boldsymbol{\beta}) \propto 1$ and $p(\sigma) \propto \sigma^{-2}$, the posterior is then of the form:

$$f(\boldsymbol{\beta}, \sigma) =$$

$$\propto \exp\left(- \sum_{i:y_i>0} \left(\frac{(y_i - \boldsymbol{x}_i^\top \boldsymbol{\beta})^2}{2\sigma^2} + \ln \sigma \right) + \sum_{i:y_i=0} \ln \Phi((u_i - \boldsymbol{x}_i^\top \boldsymbol{\beta})/\sigma) \right) \times \sigma^{-2}$$

It is not clear how we can simulate from this monstrosity in order to perform inference for $(\boldsymbol{\beta}, \sigma)$.

Posterior simplificaton

- Let $\overline{\boldsymbol{y}}$ and $\underline{\boldsymbol{y}}$ be vectors that collect all $y_i > 0$ and $y_i = 0$, respectively.
- Denote the corresponding matrix with predictors via $\overline{\mathrm{X}}$ and $\underline{\mathrm{X}}$, respectively.
- Using a latent variable w_i for each $y_i = 0$, we can write $f(\boldsymbol{\beta}, \sigma, \boldsymbol{w}) \propto$

  $$\exp\left(-\frac{\|\overline{\boldsymbol{y}} - \overline{\mathrm{X}}\boldsymbol{\beta}\|^2}{2\sigma^2} - \frac{\|\boldsymbol{w} - \underline{\mathrm{X}}\boldsymbol{\beta}\|^2}{2\sigma^2} - (m+2)\ln\sigma\right)\mathbb{I}\{\boldsymbol{w} \leq \boldsymbol{0}\}$$

  so that the marginal of $(\boldsymbol{\beta}, \sigma)$ has the desired posterior pdf.
- Note that, conditional on $(\sigma, \boldsymbol{w})$, the distribution of $\boldsymbol{\beta}$ is

  $$\mathsf{N}(\mathrm{C}(\overline{\mathrm{X}}^\top \overline{\boldsymbol{y}} + \underline{\mathrm{X}}^\top \boldsymbol{w}), \sigma^2\mathrm{C}),$$

  where $\mathrm{C}^{-1} = \overline{\mathrm{X}}^\top\overline{\mathrm{X}} + \underline{\mathrm{X}}^\top\underline{\mathrm{X}}$.

Bayesian posterior

- Thus, to simulate from the posterior, it suffices to simulate from the marginal of $(\sigma, \boldsymbol{w})$, which is of the form:

$f(\sigma, \boldsymbol{w}) \propto$

$$\exp\left(-\frac{\|\boldsymbol{w}\|^2}{2\sigma^2}+\frac{(\overline{\mathrm{X}}^\top\overline{\boldsymbol{y}}+\underline{\mathrm{X}}^\top\boldsymbol{w})^\top\mathrm{C}(\overline{\mathrm{X}}^\top\overline{\boldsymbol{y}}+\underline{\mathrm{X}}^\top\boldsymbol{w})}{2\sigma^2}-\frac{\|\overline{\boldsymbol{y}}\|^2}{2\sigma^2}\right)\times\sigma^{d-m-2},$$

on the set of values satisfying $\boldsymbol{w} \leq \mathbf{0}$, where $\boldsymbol{w} \in \mathbb{R}^{325}$.

- Without the truncation condition $\boldsymbol{W} \leq \mathbf{0}$, simulating perfect $(\boldsymbol{W}, \sigma)$ from $f(\sigma, \boldsymbol{w})$ is feasible.
- Unfortunately, satisfying the condition $\boldsymbol{W} \leq \mathbf{0}$ by sampling from f will happen with probability

$$\mathbb{P}(\boldsymbol{W} \leq \mathbf{0}) = (2.17\ldots)\times 10^{-172}, \quad \boldsymbol{W} \sim f(\sigma, \boldsymbol{w})$$

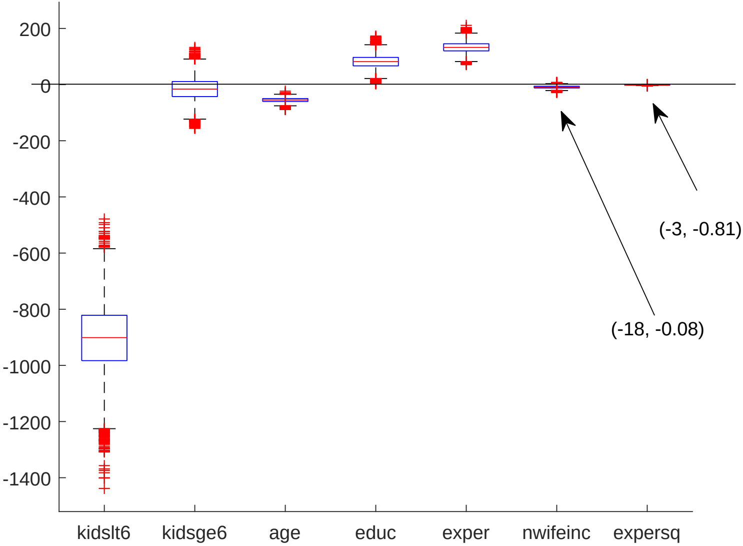
Posterior Simulation Challenges

- Naive Monte Carlo simulation faces an intractable rare-event simulation problem.
- Is **approximate** Markov chain Monte Carlo sampling the answer?
- Not really, we can do the **perfect** simulation provided we tackle the rare-event simulation problem.
- The idea is to apply a carefully crafted exponential twisting to $f(\sigma, \boldsymbol{w})$ so that the event $\boldsymbol{W} \leq \boldsymbol{0}$ is not rare.
- In our case, under the exponentially twisted measure we obtain

$$\widetilde{\mathbb{P}}(\boldsymbol{W} \leq \boldsymbol{0}) = 0.40\ldots$$

- This allows us to simulate from the Bayesian posterior using an acceptance-rejection scheme with success probability of at least 40%.

Perfect simulation for the first time



What's wrong with applying MCMC?

- This Tobit model has been previously dealt with MCMC. How is this approach better?
- MCMC is a heuristic solution, which seems to work empirically, but which has not been so far proven/shown to work in the mathematical sense for this class of problems.
- With MCMC, we can never be sure we sampled from the posterior accurately.
- With perfect rare-event simulation, we have a mathematical proof that the simulation from the posterior is exact.
- In summary, our perfect simulation approach is qualitatively different.

  The difference is between having or not having a certificate/guarantee that the algorithm has provided the correct answer.

Rare-events and high-dimensional integration

- Does rare-event simulation apply only to special Bayesian models?
- More generally, we may wish to integrate numerically a nonnegative function $f(\boldsymbol{x})$:

$$\int_{\mathbb{R}^d} f(\boldsymbol{x})\mathrm{d}\boldsymbol{x} = \int_0^\infty \mathbb{P}(f(\boldsymbol{X})/p(\boldsymbol{X}) > t)\mathrm{d}t$$

  where $\mathbb{P}$ is a measure with density p which dominates f.
- Thus any integration problem can be recast into the rare-event framework.
- The curse of dimensionality here is expressed in that typically

$$\mathbb{P}(f(\boldsymbol{X})/p(\boldsymbol{X}) > t) \downarrow 0, \quad d \uparrow \infty$$

Some technical details

The posterior which we wish to sample from can be rewritten as

$$f(\sigma, \boldsymbol{w}) \propto \exp\left(-\frac{(\boldsymbol{w}-\hat{\boldsymbol{w}})^\top(I-\underline{\mathrm{X}}\mathrm{C}\underline{\mathrm{X}}^\top)(\boldsymbol{w}-\hat{\boldsymbol{w}})}{2\sigma^2} - \frac{s^2}{2\sigma^2} - (m-d+2)\ln\sigma\right),$$

where $\boldsymbol{w} \leq \mathbf{0}$ and

$$\hat{\boldsymbol{w}} \overset{\mathrm{def}}{=} \underline{\mathrm{X}}(\overline{\mathrm{X}}^\top\overline{\mathrm{X}})^{-1}\overline{\mathrm{X}}^\top\overline{\boldsymbol{y}}$$

$$s^2 \overset{\mathrm{def}}{=} \overline{\boldsymbol{y}}^\top(\mathrm{I}-\overline{\mathrm{X}}(\overline{\mathrm{X}}^\top\overline{\mathrm{X}})^{-1}\overline{\mathrm{X}}^\top)\overline{\boldsymbol{y}}$$

Multivariate Student Distribution

It follows that the transformation

$$
\begin{aligned}
r &= s/\sigma \\
\boldsymbol{z} &= \mathrm{L}^{-1}(\hat{\boldsymbol{w}} - \boldsymbol{w})/\sigma \\
\text{where} \quad \mathrm{L}\mathrm{L}^{\top} &= \mathrm{I} + \underline{\mathrm{X}}(\overline{\mathrm{X}}^{\top}\overline{\mathrm{X}})^{-1}\underline{\mathrm{X}}^{\top} \text{ is the Cholesky decomposition} \\
\nu &\overset{\text{def}}{=} m - d - \dim(\underline{\boldsymbol{y}}) + 1 \\
\boldsymbol{l} &\overset{\text{def}}{=} \sqrt{\nu}\hat{\boldsymbol{w}}/s
\end{aligned}
$$

reveals that simulating from the posterior is equivalent to simulating form

$$
f(\boldsymbol{z}, r) = \frac{\exp\left(-\frac{\|\boldsymbol{z}\|^2}{2} - \frac{r^2}{2} + (\nu - 1)\ln r\right)\mathbb{I}\{\sqrt{\nu}\,\mathrm{L}\boldsymbol{z} \geq r\boldsymbol{l}\}}{\ell}
$$

Sequential decomposition

Due to the lower triangular structure of L, the region

$$\mathscr{R} = \{(r, \boldsymbol{z}) : r\,\boldsymbol{l} \le \sqrt{\nu} L\boldsymbol{z} \le r\,\boldsymbol{u}\}, \quad \boldsymbol{u} = \infty$$

can be decomposed into

$$\begin{aligned}
\tilde{l}_1(r) \overset{\text{def}}{=} \frac{r\, l_1}{\sqrt{\nu}}/L_{11} \le & z_1 \le \frac{r\, u_1}{\sqrt{\nu}}/L_{11} \overset{\text{def}}{=} \tilde{u}_1(r) \\
\tilde{l}_2(r, z_1) \overset{\text{def}}{=} \frac{r\, l_2 \nu^{-1/2} - L_{21} z_1}{L_{22}} \le & z_2 \le \frac{r\, u_2 \nu^{-1/2} - L_{21} z_1}{L_{22}} \overset{\text{def}}{=} \tilde{u}_2(r, z_1) \\
& \vdots \\
\underbrace{\frac{\frac{r\, l_d}{\sqrt{\nu}} - \sum_{i=1}^{d-1} L_{di} z_i}{L_{dd}}}_{\tilde{l}_d(r, z_1, \ldots, z_{d-1})} \le & z_d \le \underbrace{\frac{\frac{r\, u_d}{\sqrt{\nu}} - \sum_{i=1}^{d-1} L_{di} z_i}{L_{dd}}}_{\tilde{u}_d(r, z_1, \ldots, z_{d-1})}
\end{aligned}$$

Sequential Importance Sampling

- Let $\phi(\boldsymbol{z};\boldsymbol{\mu},\Sigma)$ denote the density of the $\mathrm{N}(\boldsymbol{\mu},\Sigma)$ distribution.
- Then, the decomposition above suggests the sequential importance sampling estimator

$$\hat{\ell} = \frac{f_\nu(R)\phi(\boldsymbol{Z};\mathbf{0},\mathrm{I}_d)}{g(R,\boldsymbol{Z})}$$

- with $(R,\boldsymbol{Z})$ distributed according to the importance sampling density on $\mathscr{R}$

$$g(r,\boldsymbol{z}) = g(r)g(\boldsymbol{z}\,|\,r) = g(r)g_1(z_1\,|\,r)\cdots g_d(z_d\,|\,r,z_1,\ldots,z_{d-1}).$$

- and R follows the χ_ν distribution with density

$$f_\nu(r) = \frac{\exp(-r^2/2+(\nu-1)\log r)}{2^{\nu/2-1}\Gamma(\nu/2)},\;\; r>0$$

Importance Sampling density

- $$g(r) = \frac{\phi(r; \eta, 1)}{\Phi(\eta)}, \ r > 0$$
  $$g_k(z_k \mid r, z_1, \ldots, z_{k-1}) = \frac{\phi(z_k; \mu_k, 1)\mathbb{I}\{\tilde{l}_k \leq z_k \leq \tilde{u}_k\}}{\Phi(\tilde{u}_k - \mu_k) - \Phi(\tilde{l}_k - \mu_k)}, \ k = 1, 2, \ldots$$
- In other words, if $\mathsf{TN}_{(a,b)}(\mu, \sigma^2)$ denotes the $\mathsf{N}(\mu, \sigma^2)$ distribution, truncated to the interval (a, b), then
  $$R \sim \mathsf{TN}_{(0,\infty)}(\eta, 1)$$
  $$Z_k \mid R, Z_1, \ldots, Z_{k-1} \sim \mathsf{TN}_{(\tilde{l}_k, \tilde{u}_k)}(\mu_k, 1), \quad k = 1, \ldots, d$$
- Let $\psi(r, \boldsymbol{z}; \eta, \boldsymbol{\mu}) = \ln \hat{\ell}$ denote the log-likelihood ratio.

Minimax Tilting

- All that remains is to choose the parameters $\eta, \boldsymbol{\mu}$ so that the estimator $\hat{\ell} = \exp(\psi(R, \boldsymbol{Z}; \eta, \boldsymbol{\mu}))$ has a well-behaved relative error.

- A simple way of selecting $(\eta, \boldsymbol{\mu})$ in our setting is to minimize the worst possible behavior of the likelihood ratio $\exp(\psi(r, \boldsymbol{z}; \eta, \boldsymbol{\mu}))$.

- In other words, we solve the optimization program

$$\inf_{\eta, \boldsymbol{\mu}} \sup_{(r, \boldsymbol{z}) \in \mathscr{R}} \psi(r, \boldsymbol{z}; \eta, \boldsymbol{\mu})$$

Saddle-point optimization

  - Note

$$\mathbb{V}\mathrm{ar}(\hat{\ell}) \leq \exp(2 \inf_{\eta,\boldsymbol{\mu}} \sup_{(r,\boldsymbol{z})\in\mathscr{R}} \psi(r,\boldsymbol{z};\eta,\boldsymbol{\mu})) - \ell^2$$

- Theorem (Parameter Selection)

  For $\nu \geq 1$ the saddle-point program

$$\inf_{\eta,\boldsymbol{\mu}} \sup_{(r,\boldsymbol{z})\in\mathscr{R}} \psi(r,\boldsymbol{z};\eta,\boldsymbol{\mu})$$

  has a unique solution, denoted $(r^*,\boldsymbol{x}^*;\eta^*,\boldsymbol{\mu}^*)$, which coincides with the solution of the convex optimization program:

$$\max_{r,\boldsymbol{z},\eta,\boldsymbol{\mu}} \psi(r,\boldsymbol{z};\eta,\boldsymbol{\mu})$$

$$\textit{subject to: } \partial\psi/\partial\eta = 0, \quad \partial\psi/\partial\boldsymbol{\mu} = \mathbf{0}, \quad (r,\boldsymbol{z}) \in \mathscr{R}\ .$$

Drawing one $(\boldsymbol{\beta}, \sigma)$ from posterior

```
Require: vectors $\boldsymbol{l}$, lower triangular $L$, and optimal $(r^*, \boldsymbol{z}^*; \eta^*, \boldsymbol{\mu}^*)$.
  repeat
      Simulate $R \sim \mathrm{TN}_{(0,\infty)}(\eta^*, 1)$
      for $k = 1, \ldots, d$ do
          Simulate $Z_k \sim \mathrm{TN}_{(\tilde{l}_k,\infty)}(\mu_k^*, 1)$
      Simulate $E \sim \mathrm{Exp}(1)$, independently.
  until $E \geq \psi(r^*, \boldsymbol{z}^*; \eta^*, \boldsymbol{\mu}^*) - \psi(R, \boldsymbol{Z}; \eta^*, \boldsymbol{\mu}^*)$
  Transform $(\boldsymbol{Z}, \boldsymbol{R})$ back into $(\boldsymbol{W}, \sigma)$
  Given $(\boldsymbol{W}, \sigma)$, simulate

      $\boldsymbol{\beta} \sim \mathrm{N}(\mathrm{C}(\overline{\mathrm{X}}^\top \overline{\boldsymbol{y}} + \underline{\mathrm{X}}^\top \boldsymbol{W}), \sigma^2 \mathrm{C}),$

  where $\mathrm{C}^{-1} = \overline{\mathrm{X}}^\top \overline{\mathrm{X}} + \underline{\mathrm{X}}^\top \underline{\mathrm{X}}$.
  return regression coefficients $\boldsymbol{\beta}$
```

Conclusions

- Rare-event simulation is not only important for modeling failures of electronic components or collapse of banks.
- It is a computational method that may tackle seemingly intractable high-dimensional integrals unrelated to rare-event phenomena, but common in statistical applications.
- Based on a limited data analysis we can proffer the following speculative prescription:

  To close the gender pay gap, women have to find husbands who are willing to share the burdens of child rearing.