Scalable Semiparametrics for Fat Tails

Matt Taddy – http://taddylab.com

Microsoft Research and Chicago Booth

with Hedibert Lopes (Insper) and Matt Gardner (eBay)

paper is at http://arxiv.org/abs/1602.08066

Big Data

The sample sizes are enormous.

- We'll see 21 and 200 million today.
- Data can't fit in memory, or even storage, on a single machine.

The data are super weird.

- Internet transaction data distributions have a big spike at zero and spikes at other discrete values (e.g., 1 or $99).
- Big tails (e.g., $12 mil/month eBay user spend) that matter.
- The potential feature space is unmanageably large.
- We cannot write down or measure believable models.

Both 'Big' and 'Strange' beg for nonparametrics.

Distribution-free Bayesian nonparametrics

Find some *statistic of the DGP* that you care about:

- derive from first principles, e.g. moment conditions
- *an algorithm* that we know works, e.g. CART
- think about geometric projections, e.g. OLS

Call this statistic $\boldsymbol{\theta}(g)$ where $g(\mathbf{z})$ is the DGP (e.g., for $\mathbf{z} = [\mathbf{x}, y]$).

Then you write down a flexible model for the DGP g, and study properties of the posterior on $\boldsymbol{\theta}(g)$ induced by the posterior over g.

A flexible model for the DGP

Say $\mathbf{z} = [\mathbf{x}, y]$ is a single *independent* data point.

Each data point assumes one of a *finite* number of possible values, $[\boldsymbol{\zeta}_1 \dots \boldsymbol{\zeta}_L]$, with probabilities proportional to $[\boldsymbol{\theta}_1 \dots \boldsymbol{\theta}_L]$.

$$g(\mathbf{z}) = \frac{1}{|\boldsymbol{\theta}|} \sum_{l=1}^{L} \theta_l \mathbb{1}_{[\mathbf{z}=\boldsymbol{\zeta}_l]}$$

We complete specification with a conjugate prior on the weights:

$$\frac{\boldsymbol{\theta}}{|\boldsymbol{\theta}|} \sim \mathrm{Dir}(a) \propto \frac{1}{|\boldsymbol{\theta}|^{L(a-1)}} \prod_l \theta_l^{a-1} \quad \text{where} \quad a, \theta_l > 0.$$

This is the Dirichlet-multinomial sampling model (Ferguson 1973).

Now you've observed some data, say $\mathbf{Z} = \{\mathbf{z}_1 \ldots \mathbf{z}_n\}$.
(say every $\mathbf{z}_i = [\mathbf{x}_i, y_i]$ is unique).

The posterior over weights has $\theta_l \overset{ind}{\sim} \mathrm{Exp}\left(a + \mathbb{1}_{[\zeta_l \in \mathbf{Z}]}\right)$.

A convenient limiting case

$a \to 0$ leads to $\mathrm{p}(\theta_l = 0) = 1$ for $\zeta_l \notin \mathbf{Z}$.

In this case, we can focus on only the *observed support* and write the posterior for our DGP

$$g(\mathbf{z}) = \tfrac{1}{|\boldsymbol{\theta}|} \sum_{i=1}^{n} \theta_i \mathbb{1}[\mathbf{z} = \mathbf{z}_i], \quad \theta_i \overset{iid}{\sim} \mathrm{Exp}(1).$$

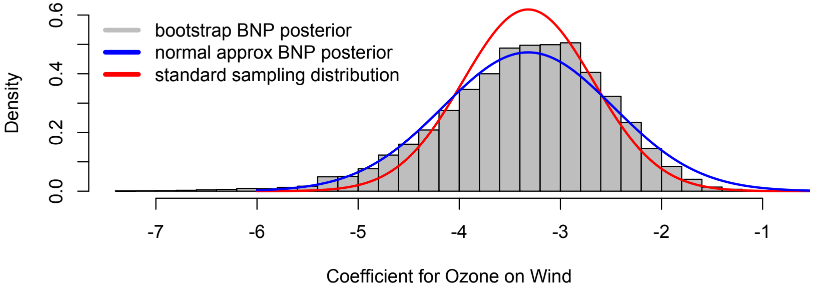
This is just the Bayesian bootstrap. (Rubin 1981)

Example: **Ordinary Least Squares**

Population OLS is a posterior functional

$$\boldsymbol{\beta} = (\mathbf{X}'\boldsymbol{\Theta}\mathbf{X})^{-1}\mathbf{X}'\boldsymbol{\Theta}\mathbf{y}$$

where $\boldsymbol{\Theta} = \mathrm{diag}(\boldsymbol{\theta})$. This is a random variable. (sample via BB)



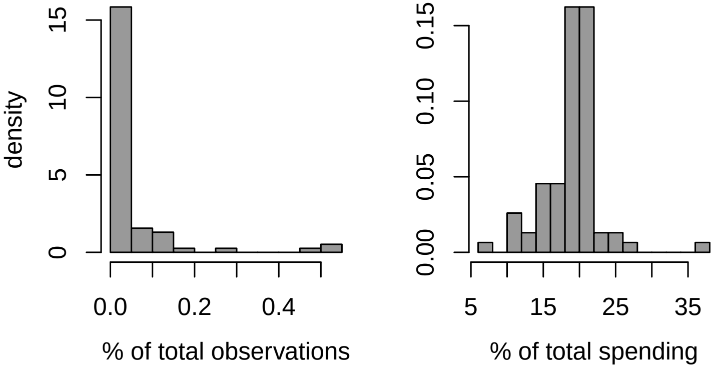
We've had a bunch of success with this strategy (and especially, taking advantage of analytic approximations to the BNP posterior).

- Bayesian forest alternative to Random Forests, and the scalable Empirical BF (T, Chen, Yu, Wyle 2015 ICML)
- BNP theory for regression adjustement in treatment effect estimation (T, Gardner, Chen, Draper 2016 JBES)
- BFs for heterogeneous treatment effects (T+ 2016 JBES)

But, at heart, we are essentially [nonparametric] bootstrapping. And sometimes the bootstrap fails:

e.g., model selection, high dimensions, and heavy tails.

Internet transaction data is super heavy tailed.



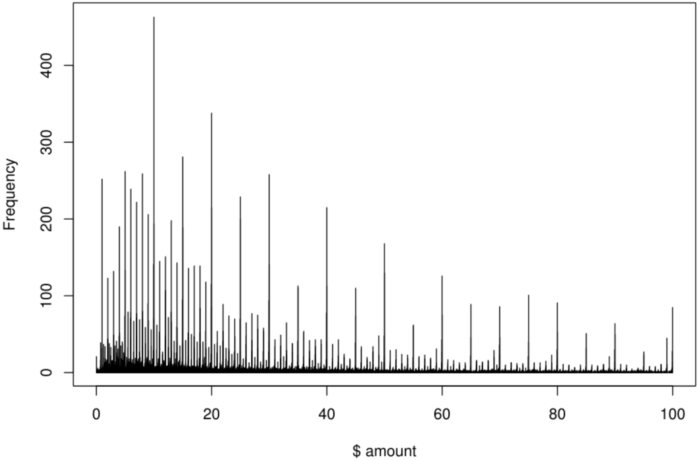
The standard bootstrap fails for infinite variance distributions.

We should suspect our Bayesian bootstrap will have similarly bad frequentist properties. e.g., the posterior will not be consistent.

Lots of spikes!

14,521 unique $ amounts from 50,000 users.

78.18% of the $ amounts are less than $100.



So the distribution below the tail is also screwy.

Semi-parametrics: only model where you really need it.

Add a parametric tail above some threshold u:

$$\mathrm{p}(z) = \frac{1}{|\boldsymbol{\theta}|}\sum_{l=1}^{L}\theta_l \mathbb{1}[z=\zeta_l] + \frac{\theta_{L+1}}{|\boldsymbol{\theta}|}\mathrm{GPD}(z-u;\ \xi,\ \sigma)\mathbb{1}[z \geq u]$$

where GPD is the generalized Pareto $\mathrm{p}(y) = \frac{1}{\sigma}\left(1+\xi\frac{y}{\sigma}\right)^{-(\frac{1}{\xi}+1)}$.

A nice simple prior is $\pi(\sigma,\xi) = \frac{1}{\sigma}\xi^{a-1}(1-\xi)^{b-1}\mathbb{1}_{\xi\in(0,1)}$.

You can set $a=b=1$ with enough data, but it is cooler to use background info. It is usually a bad idea to add information on σ.

Inference

Say $\mu = \mathbb{E}[z]$ and $\lambda = \mathbb{E}[z - u | z \geq u]$.

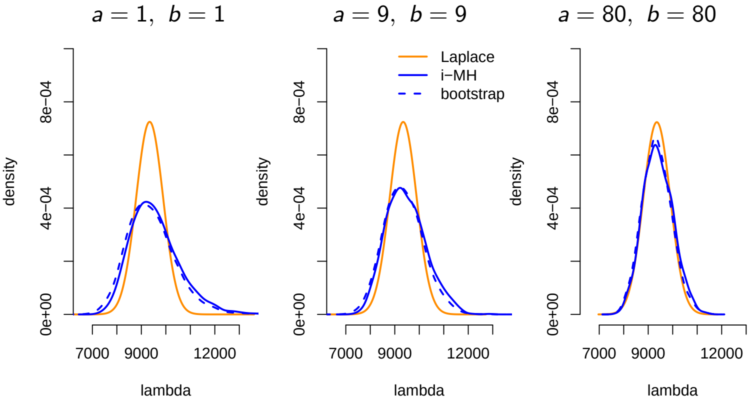
$\mathbb{E}[\mu]$ and $\mathrm{var}(\mu)$ are available in closed form up-to $\mathbb{E}[\lambda]$ and $\mathrm{var}[\lambda]$.

For these GPD tail parameters, we introduce a new independence MH sampler based upon the parametric bootstrap:

- Fit the MAP parameter estimates $[\hat{\xi}, \hat{\sigma}]$ and obtain B draws $[\hat{\xi}_b, \hat{\sigma}_b]$ from the parametric bootstrap for this MAP.
- Get a kernel estimate, say $r(\xi, \sigma)$, for the bootstrap density, and use this in your acceptance probability calculations.

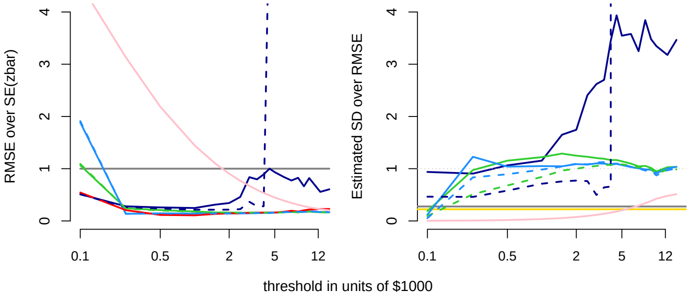
Or, you can use a laplace approximation.

Inference for the mean exeedance $\lambda = \sigma/(1-\xi)$, beyond $u = \$9000$, in one of our treatment groups.



These results give all we need for inference about the overall mean.

Performance over 200 resamples of $N = 50,000$ from 10^7 total.



– Bayes a,b=1 **– Bayes a,b=9** **– Bayes a,b=80**

– Winsorization **– FW tilting** **– sample mean** **– N/2 subsampling**

Consistency and thresholds

The *semiparametric bootstrap* is consistent if $u_N = O(N^{\xi/(1+2\delta\xi)})$.
Which is cool, because the nonparametric bootstrap is not.

In this limit, $\sigma_N = \xi u_N$. So, one can increase u until this holds.
Indeed, it holds roughly in our examples for good performing u.

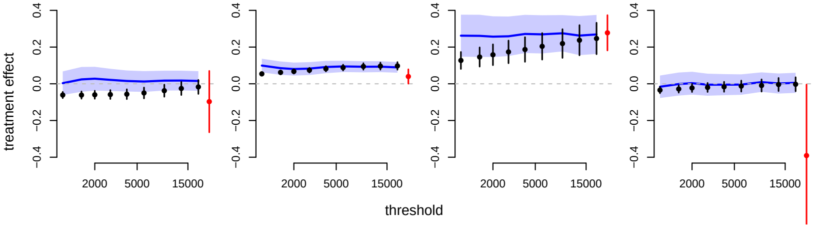
In practice, results are robust to a range of tail thresholds.

This also suggests why informative priors on σ are not useful:
it is already informed by both the threshold and the tail index.

Semi-parametrics in treatment effect estimation

Semiparametric inference makes a big difference in A/B analysis. Looking at ATEs, $\mu_1 - \mu_2$, in a few example experiments:

sp-bayes, capped, naive



Such care in uncertainty quantification is essential whenever you need to understand the full posterior (or sampling) distribution.

e.g., bandit learning with heavy-tailed rewards is another application.

Efficient Big Data analysis

We're reserving difficult modeling for the difficult parts of our learning problem, while taking advantage of nonparametric consistency and large numbers of observations for the easy bits

In general, we statisticians and machine trainers should be thinking about what portions of the 'model' are hard or easy to learn.

Once we figure this out, we can use a little bit of the data to learn the easy stuff and direct our full data at the hard stuff.

This is *the* future for Big Data.

thanks!