EVERYTHING* I KNOW ABOUT IMPORTANCE SAMPLING

**That I can cover in 45ish minutes*

Daniel Simpson

Aki Vehtari, Andrew Gelman, Yuling Yao, Jonah Gabry

SUCH HEFT, SO IMPORTANCE

ONE OF THE FUNDAMENTAL TASKS OF STATISTICAL COMPUTATION

- One of the most common things that we need to do is compute the expectation of a random variable that has some probability distribution function p.

$$I_h = \mathbb{E}_{\theta \sim p}\left[h(\theta)\right] = \int_{\Theta} h(\theta) p(\theta)\, d\theta$$

- A whole bunch of the time, we only know $p(\theta)$ up to an unknown normalizing constant

$$p(\theta) = \frac{f(\theta)}{Z}$$

AN ELEGANT APPROXIMATION

- We can rarely compute our required expectations exactly, so we need to do something clever.
- The **Monte Carlo method** uses S **independent** samples from the distribution p to approximate the expectation as

$$I_h \approx \frac{1}{S} \sum_{s=1}^{S} h(\theta_i)$$

- This has two things going for it:

1. It is **unbiased**
2. It has finite, vanishing variance

$$\text{Var}(I_h) = \frac{\text{Var}\left(h(\theta)\right)}{S} = \mathcal{O}(S^{-1})$$

BUT MONTE CARLO IS RARELY PRACTICAL

- There are two big barriers to using Monte Carlo in practice:

  1. We usually can't easily draw samples from p
  2. We often only know f (that is, we don't know the normalizing constant for p)

- But that doesn't mean we give up hope.

IMPORTANCE SAMPLING ENTERS, RIDING A HORSE

- One way out of this problem is to replace the samples $\theta_s \sim p$ with draws from a different distribution g.
- Why? Well if we choose g carefully we will be able to sample from it.
- The problem is correcting for sampling from the wrong distribution.
- **IDEA**: Visit Monte Carlo upon a different integral

$$\int h(\theta)p(\theta)\,d\theta = \int h(\theta)\frac{p(\theta)}{g(\theta)}g(\theta)\,d\theta$$

BUT WHAT ABOUT THAT NORMALIZING CONSTANT?

- But what if we only know $f \propto p$?
- Well we can estimate the normalizing constant

$$Z = \int f(\theta)\, d\theta = \int \frac{f(\theta)}{g(\theta)} g(\theta)\, d\theta \approx \frac{1}{S} \sum_{s=1}^{S} r(\theta_s)$$

- Here $r(\theta) = \dfrac{f(\theta)}{g(\theta)}$
- The variance of Z will be $\mathrm{Var}(Z) = \dfrac{\mathrm{Var}_{\theta \sim g}\left[r(\theta)\right]}{S}$

SO WHAT DO WE KNOW?

- We can estimate integrals if we have a distribution g to sample for and the target distribution up to a constant $f \propto p$
- The **self-normalized importance sampling estimate is**

$$I_h \approx I_h^S = \frac{\sum_{s=1}^{S} h(\theta_s) r(\theta_s)}{\sum_{s=1}^{S} r(\theta_s)}, \qquad \theta_s \overset{\text{iid}}{\sim} g$$

- By the law of large numbers, $I_h^S \to I_h$ as long as

$$\mathbb{E}_{\theta \sim g}(r(\theta)) < \infty$$

SO DOES IT WORK?

YEAH?

- Well the good news first: Importance sampling has finite variance if two things happen:

  1. $\mathbb{E}_{\theta \sim p}(h(\theta)^2) < \infty$
  2. $\mathbb{E}_{\theta \sim g}(r(\theta)^2) < \infty$

- This seems good. It implies that

$$\Pr\left(|I_h^S - I_h| > \epsilon\right) \leq \frac{C}{S\epsilon^2}$$

- So we can always make S big enough to ensure that the error is below a prescribed threshold with very high probability.

YEAH NAH. (AKA MACKAY 2013 CHAPTER 29.2)

- Sadly, that was all just a dream.
- Easy example: Let p be uniform on the d-dimensional unit sphere and let g be a product of independent $N(0,\sigma^2)$.
- **Good:** The importance weights are bounded!!!!
- **Bad:** $\|\theta\|_2^2 \sim N(\sigma^2 d, \sigma^2\sqrt{2d})$
- So $r(\theta) = \exp\left(\frac{\|\theta\|_2^2}{2\sigma^2}\right) \approx \exp\left(\frac{d}{2} \pm \frac{\sqrt{2d}}{2}\right)$
- The ratio of the largest ratio to the average ratio will be around $\exp(\sqrt{2d})$ (About 1.4 million when $d = 100$)

WHAT DOES THIS MEAN?

- Lets say we order the samples so importance ratios so that they occur in increasing order (ie $r(\theta_1) \leq r(\theta_2) \leq \cdots$)
- If there is one ratio that is **much** larger than the others, we get

$$\frac{\sum_{s=1}^{S} h(\theta_s) r(\theta_s)}{\sum_{s=1}^{S} r(\theta_s)} \approx \frac{r(\theta_S) h(\theta_S)}{r(\theta_S)} = h(\theta_S)$$

- So if the ratios vary wildly in magnitude, then the self-normalized importance sampler only notices a few of the samples and will have **very** high variance.

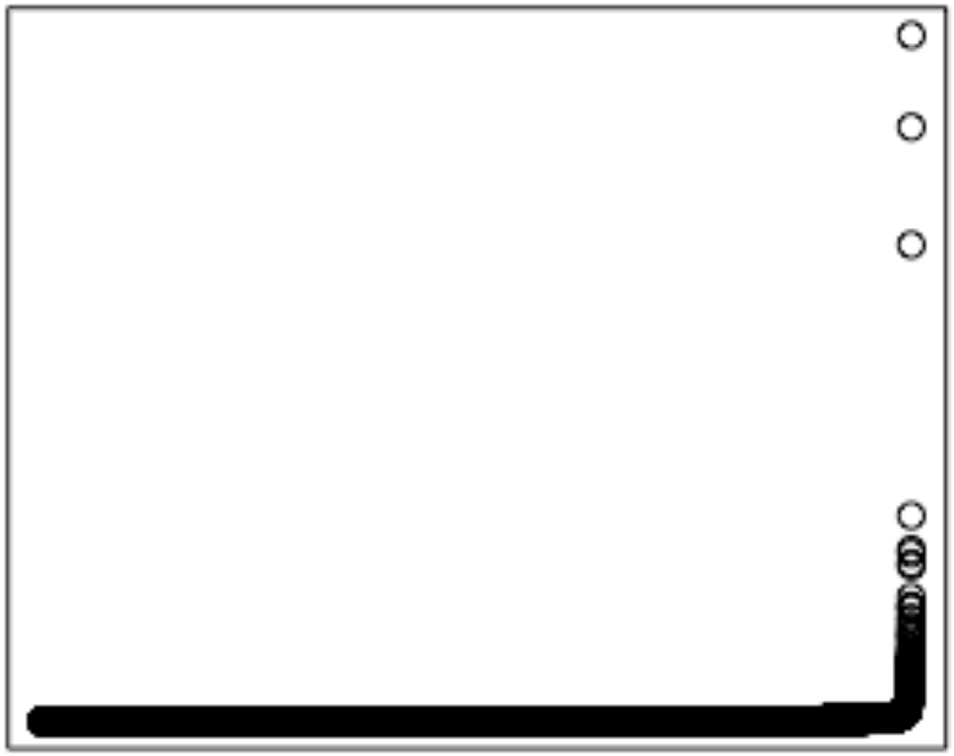
IS THERE A LESSON HERE?

- Probably there are two:

  1. Things are weird in high dimensions!
     - It's hard to make a good proposal in high dimensions.
     - "Importance ratios have bounded variance" is not a good criterion in even moderate dimensions.
  2. We need to care about the **distribution** of extreme importance ratios.
     - Asking for finite variance is a way to do this, but if we want this to work robustly we need more control.

WHAT DOES THE TAIL LOOK LIKE?

WHAT TO DO WHEN THE TAIL IS GIVING YOU TROUBLE



Importance ratios

Truncated Importance Sampling, Ionides, 2008

THIS FIXES A LOT OF PROBLEMS!

- Suddenly we are no longer unbiased because there is a modification to the extreme part of the integral
- But that bias is much much smaller than the estimate from the bulk and goes away asymptotically.
- We **always** have finite variance, regardless of the properties of $r(\theta)$
- And, **(under conditions)**, the **Truncated Importance Sampling (TIS)** estimator is asymptotically normal.

SO HOW MUCH DO WE HAVE TO TAKE OFF?

- Great question. (Actually, this is most of Ionides' paper)
- He suggests replacing the raw importance ratios with weights

$$w_s = \min\left(r(\theta_s), \tau_S\right)$$

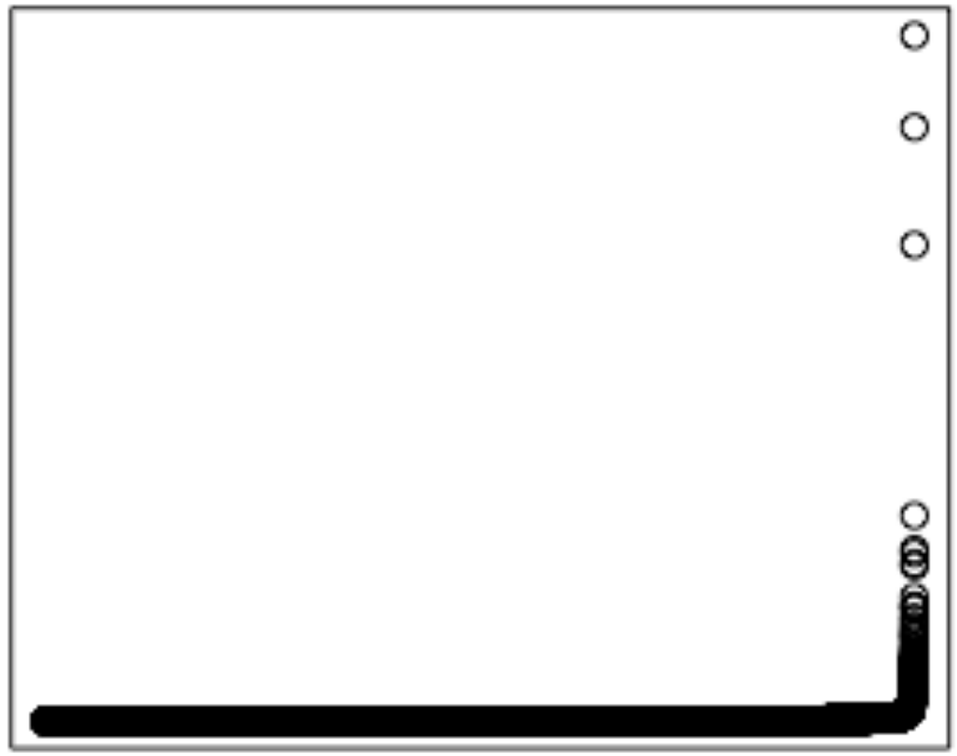
so the estimator is

$$I_h \approx I_h^S = \frac{\sum_{s=1}^S w_s h(\theta_s)}{\sum_{s=1}^S w_s}, \qquad \theta_s \overset{\text{iid}}{\sim} g$$

- The optimal choice of τ_S depends on the (unknown) distribution of $r(\theta_s)$
- He argues a default choice is $\tau_S = ZS^{1/2} \approx \bar{r}S^{1/2}$

ENTER AKI VEHTARI, ALSO RIDING A HORSE

- But there's obviously a problem.
- Ionides' theory only works with a threshold that's independent of the importance ratios
- And there is **a lot** of variation between problems
- Also, Aki is a Bayesian at heart, so he decided to just **model** the tail



DARLING I DON'T KNOW WHY I GO TO EXTREMES

ENTER EXTREME VALUE THEORY ON ANOTHER DAMN HORSE

- It turns out that the distribution of extreme events (like those bigger than a particular threshold) is a well studied thing.
- In particular, there are some classical limit results suggesting that

$$r(\theta) \mid r(\theta) > \tau \to GPD(\tau, \sigma, k), \quad \tau \to \infty$$

- Here *GPD* is the **Generalized Pareto Distribution,** which has pdf

$$\frac{1}{\sigma}\left(1 + k\frac{r-\tau}{\sigma}\right)^{-1/k-1}$$

- The important parameter here is k, which controls the heaviness of the tail.

A FIRST IDEA

- If you squint, you'll notice that $r(\theta)$ has k^{-1} finite moments.
- So maybe we can just not truncate if $k < 0.5$
- This is ok, but we're not really using any information about the tail

(This is not related to anything, but I'm required to put a picture of Celine on every talk because Canada and I had space.)

WHAT IF WE TAKE OUR MODEL SERIOUSLY?

- How about we **replace** our extreme weights with their modelled value.
- To so this, we sort θ_s so that $r(\theta_1) \leq r(\theta_2) \leq \cdots$
- This type of reasoning leads to "order statistics".
- If there are M samples with ratios bigger than τ, then the $(S - M + z)$th weight is

$$w_{S-M+z} = \tau + F^{-1}\left(\frac{z - 1/2}{M}\right),$$

where F is the CDF of the estimated Generalized Pareto Distribution

THIS ALMOST WORKS

- This *bias corrected* truncated importance sampling is almost a good idea.
- The trouble is finding a good default value of the threshold τ
- It turns out the performance is sensitive to this value: if it's not large enough, the GPD approximation will be bad
- Extremes theory suggests you want to use the largest $M = \mathcal{O}(S^{1/2})$ samples to estimate the GPD.
- Unfortunately, if the threshold is deterministic M is a random variable and this is a hard condition to guarantee.

WHEN IN DOUBT, FIX THE MARGINS

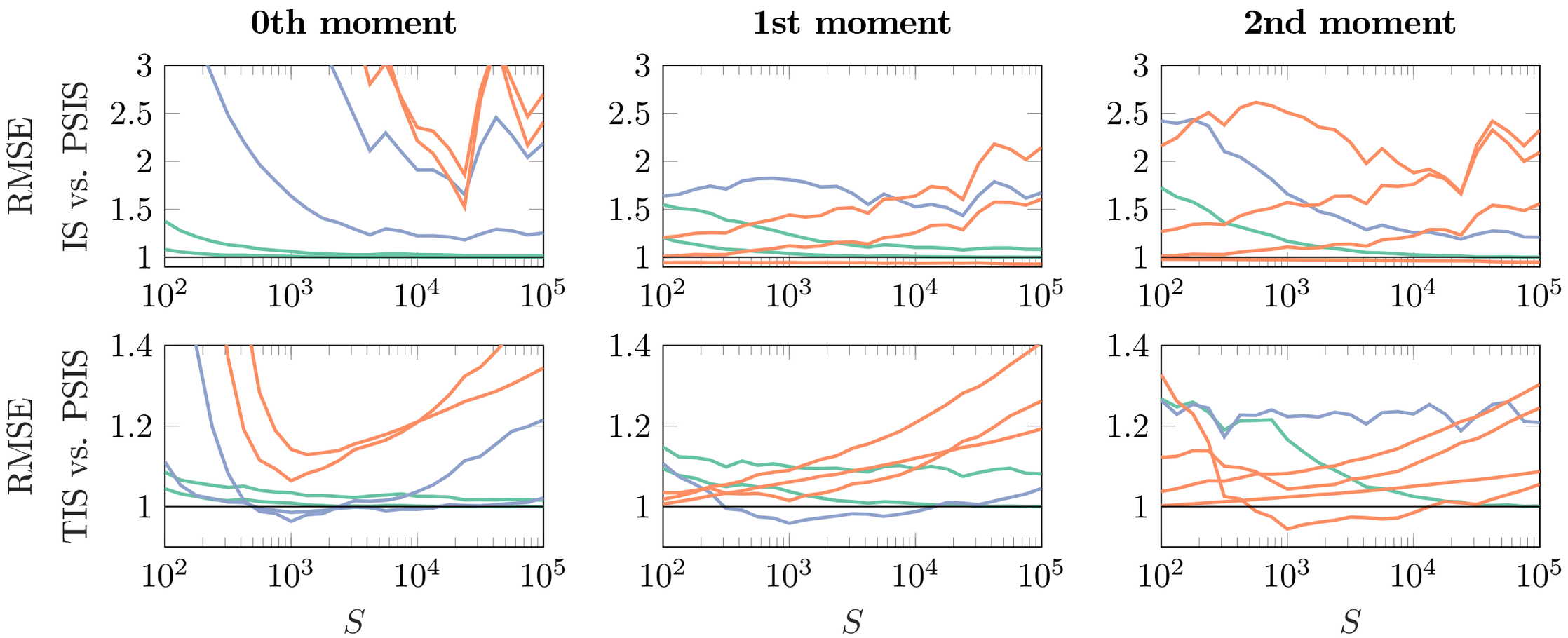
- A way through this is to fix $M = 3S^{1/2}$ and implicitly define the threshold to be $\tau = r_{S-M+1:S}$, the $(S - M + 1)$th order statistic.
- This leads to **Pareto Smoothed Importance Sampling (PSIS)**

$$I_h^S = \frac{1}{S}\sum_{s=1}^{S}\left(r(\theta_s) \wedge r_{(S-M+1):S}\right) h(\theta_s) + \frac{1}{S}\sum_{m=1}^{M} \tilde{w}_m h(\theta_{S-M+m}) \,.$$

- Written this way, it's clear that it is TIS with
  1. An adaptive threshold
  2. Bias correction

YES BUT DOES IT WORK?

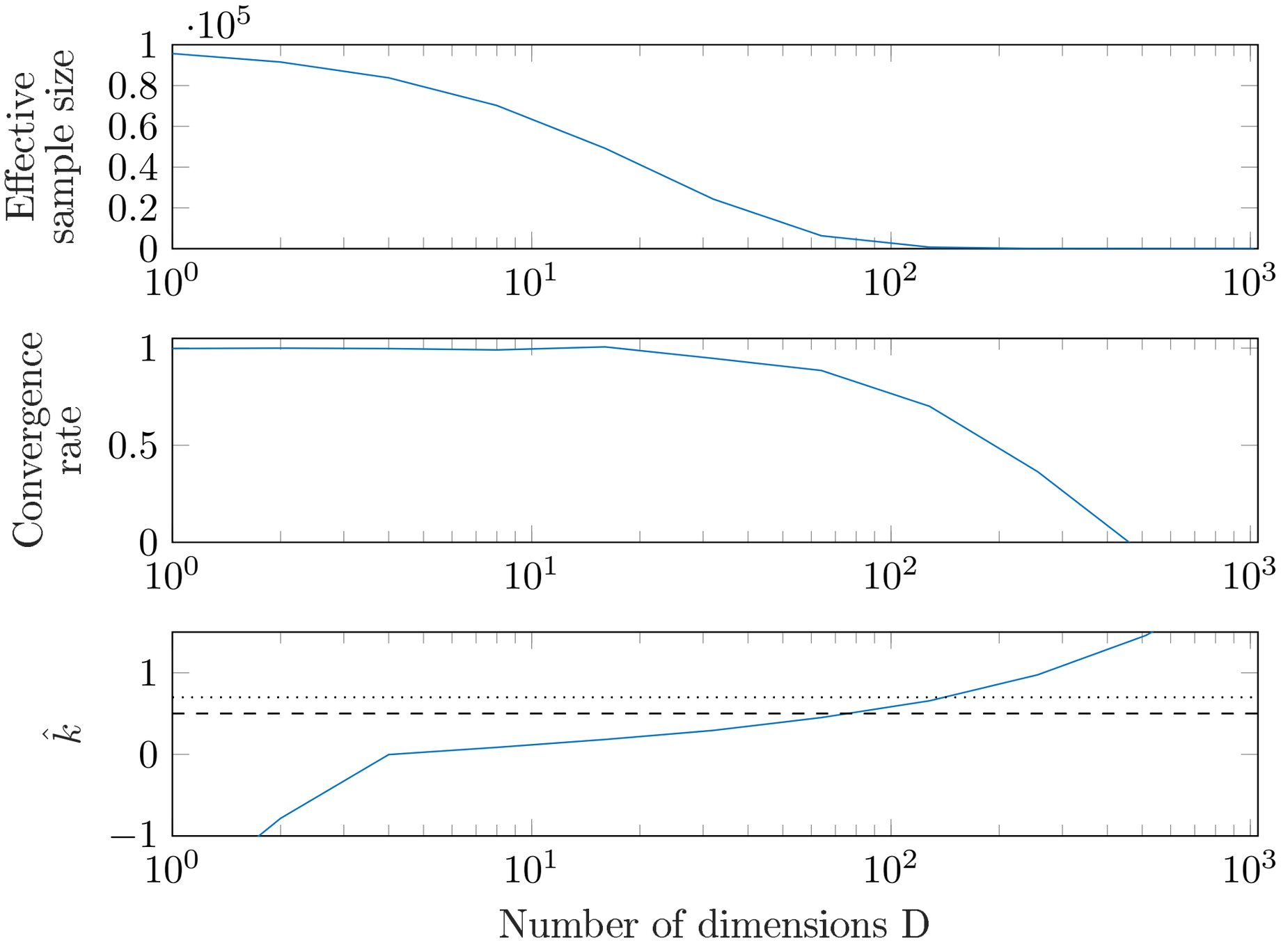
YES.


0th moment
1st moment
2nd moment
RMSE
IS vs. PSIS
TIS vs. PSIS
3
2.5
2
1.5
1
1.4
1.2
10^2
10^3
10^4
10^5
S

REALLY? YES.

- PSIS forms the basis of the popular (>350k downloads) loo package in R for leave-one-out cross validation
- We've also used it for a whole variety of other problems
- The key feature turns out to the tail parameter k which is more than just a nuisance parameter that needs to be estimated.
- The estimate of k, which we write as $\hat{k}$ or **k-hat**, is a great heuristic for identifying when PSIS will and will not work.

THE MAJESTY OF K-HAT



SO WHAT VALUE OF K-HAT SHOULD WE USE AS A THRESHOLD

- Aki did **extensive** simulations and the magic number turned out to be around 0.7.
- This also seems to hold for truncated importance sampling: even though the variance is finite, getting accuracy becomes very expensive.
- Is there a reason why this is true?

ENTER PERSI DIACONIS RIDING YET ANOTHER BLOODY HORSE

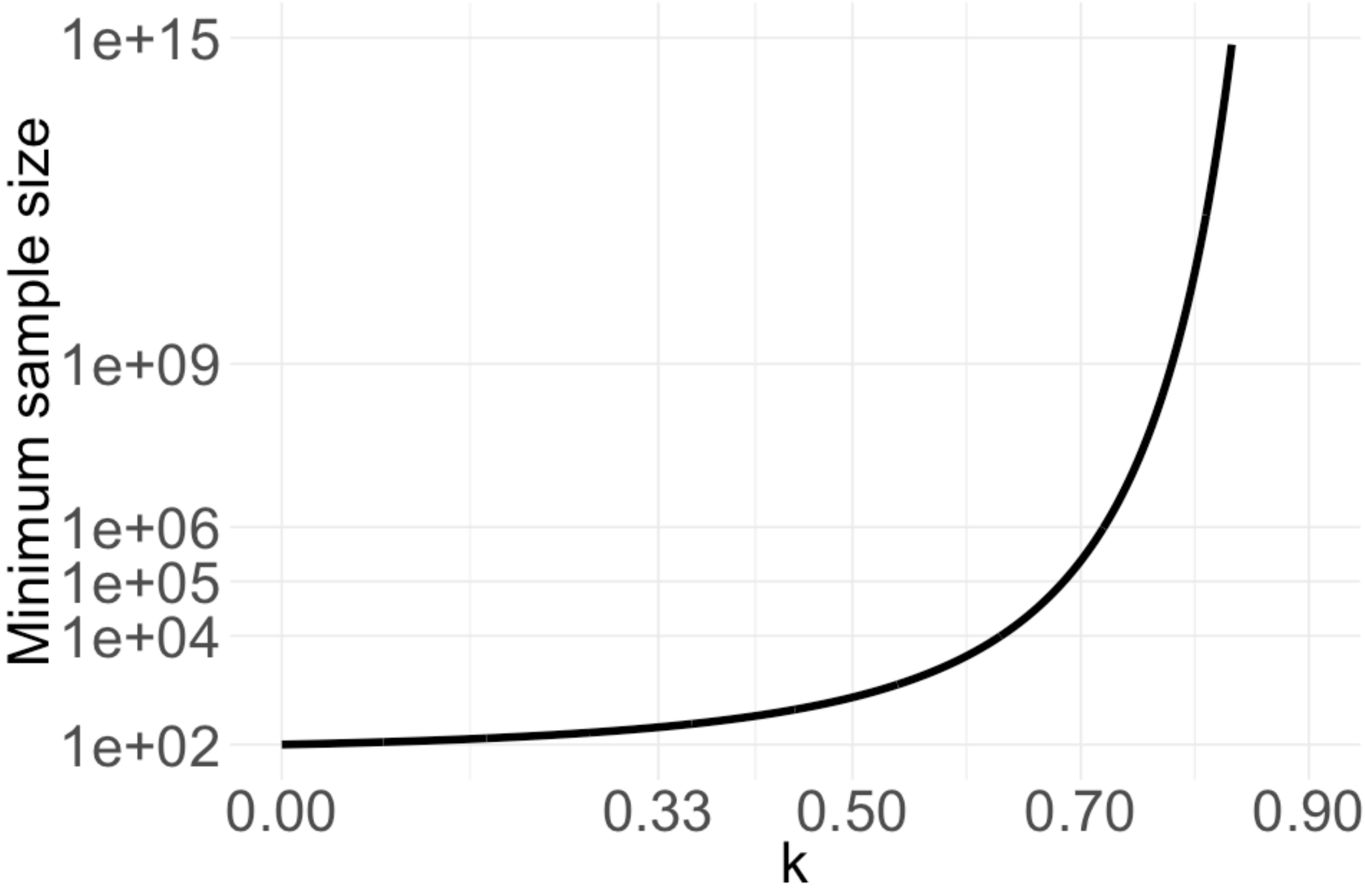
- What if we asked "how big should S be for a given accuracy?"
- Chaterjee and Diaconis (2018) give a good answer to that: It depends on how close g is to p in the Kullback-Leibler distance.
- In they showed that we need at least

$$\log S \geq \mathbb{E}_{\theta \sim g}\left[r(\theta)\log(r(\theta))\right]$$

samples to get any accuracy at all.

- So if this number is big we have exactly no hope of an accurate approximation
- We don't know the distribution of r, but we know the important bit is Pareto…

THE CHATERJEE/DIACONIS BOUND IF R IS EXACTLY PARETO



BUT OF COURSE WE DON'T KNOW K EXACTLY

- Recall that high-dimensional example again.
- It had bounded importance ratios, but the k-hat estimate was bigger than 0.7 when the dimension grew
- This indicates that k-hat can be viewed loosely as telling us how many moments a distribution that produces a **particular finite realization** of the importance ratios has.
- So what k-hat>0.5 is telling us in this context is that from the samples we have, it isn't clear that the ratios have finite variance.

BUT, LIKE, DOES IT ACTUALLY WORK

ENTER ME, ANNOYED TO BE ON A HORSE

- But extensive simulations, common sense, and 350k users do not make reviewers happy. There needed to be theory.
- This turns out to be hard.
- Why?
- Well, for one thing, all of the proofs are **much** easier when the threshold is deterministic and, therefore, the weights are either a) correct, or b) constant.
- Using an order statistic for the threshold makes the method work much better, but makes it **much** more complex as an object to do maths upon.

WHAT'S THE FIRST PROBLEM?

- Well, the importance sampler involves sums of $r(\theta_s)h(\theta_s)$, but the truncation only looks at $r(\theta_s)$
- So we need to make sure $h(\,\cdot\,)$ doesn't get **too big** as $r(\,\cdot\,)$ grows.
- We also need to make sure it doesn't throw a total tantrum at the very thought of infinity.
- This means we'll get conditions on

$$m_k(r) = \mathbb{E}(\,|h(\theta)|^k \mid r(\theta) = r)$$

- Thankfully, we can actually estimate these functions!

WHAT'S THE SECOND PROBLEM

- Counterexamples abide in the land of extremes.
- They usually look like one of two things:
    1. the distribution changes in some fundamental way at infinity.
    2. the asymptotic behaviour takes **ages** to kick in.
- There is very little we can do about this, other than assume that it doesn't happen.

SOME FUN-CHECKABLE ASSUMPTIONS ABOUT THE RATIOS

- If $r(\theta) \sim R$ has pdf $\tilde{r}$, then we need some (slightly equivalent conditions:

  1. R has $(1+\delta)$ moments
  2. $\lim_{z\to\infty} \frac{z\tilde{r}(z)}{1-R(z)} = c > 1$
  3. $\frac{1-R(\xi_M \pm \epsilon)}{1-R(\xi_M)} \overset{<}{>} 1 \mp cS^{-1/2}$

- These are "standard" conditions for anything that involves order statistics, but that doesn't mean we can check they hold.

AND A TONNE OF MOMENT CONDITIONS

- We need $h \in L^2(p) \cap L^2(g)$
- We need $\mathbb{E}(r^2_{S-M+1:S}(m_j(r_{S-M+1:S})) \vee 1) = o(S)$ uniformly in S for $j = 1,2$
- If **all of these things hold**, then PSIS is consistent, and has finite, vanishing variance. Just like a real boy!

YOU'RE BEING CONSPICUOUSLY QUIET ABOUT NORMALITY

- Because order statistics are not independent of the rest of the sample, PSIS is no longer that sum of independent random variables, which makes asymptotical normality less straightforward.
- There is exactly one paper that I could find that covers a similar enough case.

ENTER PHILIP S GRIFFIN, WHO THANKFULLY BOUGHT A HORSE IN 1987

Stochastic Processes and their Applications 29 (1988) 107–127
North-Holland

ASYMPTOTIC NORMALITY OF WINSORIZED MEANS

Philip S. GRIFFIN
Department of Mathematics, Syracuse University, Syracuse, NY 13244-1150, USA

Received 17 March 1987
Revised 2 February 1988

Let X_i be non-degenerate i.i.d. random variables with distribution function F, and let $X_{n1}, \ldots, X_{nn}$ denote the order statistics of $X_1, \ldots, X_n$. In trying to robustify the sample mean as an estimator of location, several alternatives have been suggested which have the intuitive appeal of being less susceptible to outliers. Here the asymptotic distribution of one of these, the Winsorized mean, which is given by

$$n^{-1}\left(s_n X_{n,s_n} + \sum_{i=s_n+1}^{n-r_n} X_{ni} + r_n X_{n,n-r_n+1}\right)$$

where $r_n \geq 0$, $s_n \geq 0$ and $r_n + s_n \leq n$, is studied. The main results include a necessary and sufficient condition for asymptotic normality of the Winsorized mean under the assumption that $r_n \to \infty$, $s_n \to \infty$, $r_n n^{-1} \to 0$, $s_n n^{-1} \to 0$ and F is convex at infinity. It is also shown, perhaps somewhat surprisingly, that if the convexity assumption on F is dropped then the Winsorized mean may fail to be asymptotically normal even when X_1 is bounded!

LOOK AT MY HORSE, MY HORSE IS AMAZING

The secret: Pareto Smooth $r(\theta)h(\theta)$ at both ends

$$\tilde{I}_h^S = \frac{1}{S}\sum_{s=M+1}^{S-M} h(\theta_s)r(\theta_s) + \frac{M}{S}h(\theta_M)r(\theta_M) + \frac{M}{S}h(\theta_{S-M+1})r(\theta_{S-M+1})$$
$$+ \frac{1}{S}\sum_{j=1}^{M}\left(k_R^{-1}\sigma_R\left[\left(1-\frac{j-1/2}{M}\right)^{-k_R}-1\right] - k_L^{-1}\sigma_L\left[\left(\frac{j-1/2}{M}\right)^{-k_L}-1\right]\right).$$

The following result is an application of the main result of Griffin (1988) to this estimator.

Theorem 3. *Let $F(\cdot)$ be the CDF of $h(\theta)r(\theta)$, $\theta \sim g$. Assume $R(z)$ is convex in some neighbourhood of $z = -\infty$ and $1 - R(z)$ in some neighbourhood of $z = \infty$ and that*

$$\lim_{z\to-\infty}\frac{|z|f(z)}{F(z)} = \frac{1}{k'_L}, \qquad \lim_{z\to\infty}\frac{zf(z)}{1-F(z)} = \frac{1}{k'_R},$$

for $0 < k_L, k_R < 1$, where $f(\cdot)$ is the density of F. Then there exists a sequence γ_S such that

$$c_S(\tilde{I}_h^S - I_h) \xrightarrow{d} N(0, \sigma^2),$$

where $c_S = \mathcal{O}(S^{1/2\wedge(3/4-k'/2)}$, where $k' = k'_L \vee k'_R$.

SHOW JUMPING

- But does asymptotic normality hold for PSIS as we use it.
- It's honestly not clear to me, but maybe if $h(\theta)$ isn't too exciting (eg if it's bounded)
- The problem turns out to be the bias correction term

$$\sum_{j=1}^{M} \tilde{w}_j h(\theta_{S-M+j})$$

- This looks like a sum of independent random variables, but it is **not**! (Recall θ_s is ordered!)
- This may look like it will go zero but, for fixed threshold, it has an asymptotically normal limit

WHAT WILL PROBABLY DO IT???

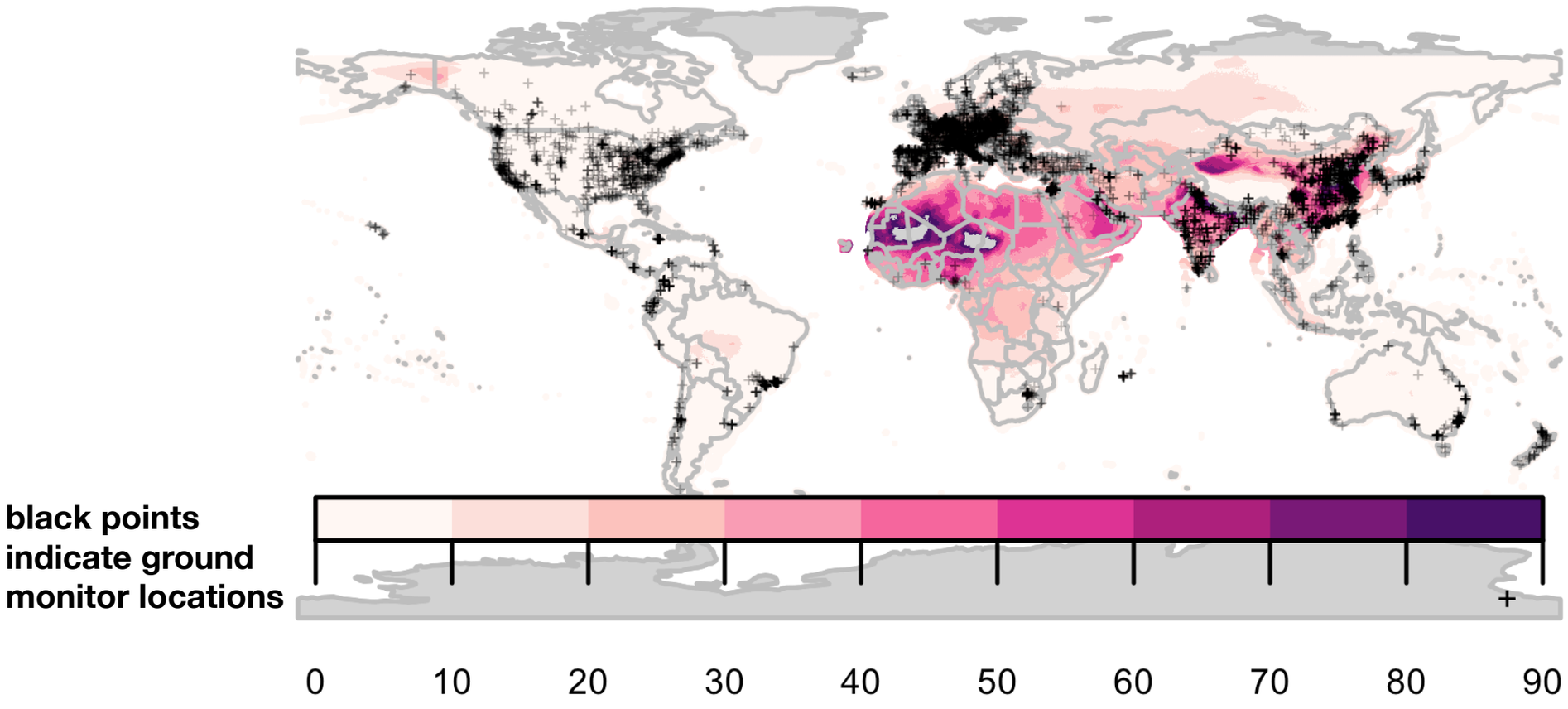
- Remember that the threshold is an order statistic, so it's random. But it is also a large (upper intermediate in the lingo) order statistic, so it it will go to infinity with S
- So we need to show that the convergence of that bias correction term (L-statistic of a concomitant in the lingo) to normal is **uniform** in the threshold. (There is a Berry-Esseen result that could help.
- **Then** we need to go carefully through Griffin's 15 pages of tightness calculations to make sure we didn't accidentally break it.
- But I think it's probably asymptotically normal.

BUT EVEN IF IT'S A BIT OF AN ARSE TO THEORY, PSIS IS REALLY USEFUL

WHEN KYLIE SAID "BREATHE" THIS WASN'T WHAT SHE WANTED

Goal Estimate global PM2.5 concentration

Problem Most data from noisy satellite measurements (ground monitor network provides sparse, heterogeneous coverage)



Satellite estimates of PM2.5 and ground monitor locations

POSTERIOR PREDICTIVE CHECKING

The *posterior predictive distribution* is the average data generation process over the entire model

$$p(\tilde{y}|y) = \int p(\tilde{y}|\theta)\, p(\theta|y)\, d\theta$$

POSTERIOR PREDICTIVES CAN BE USEFUL FOR MODEL COMPARISON

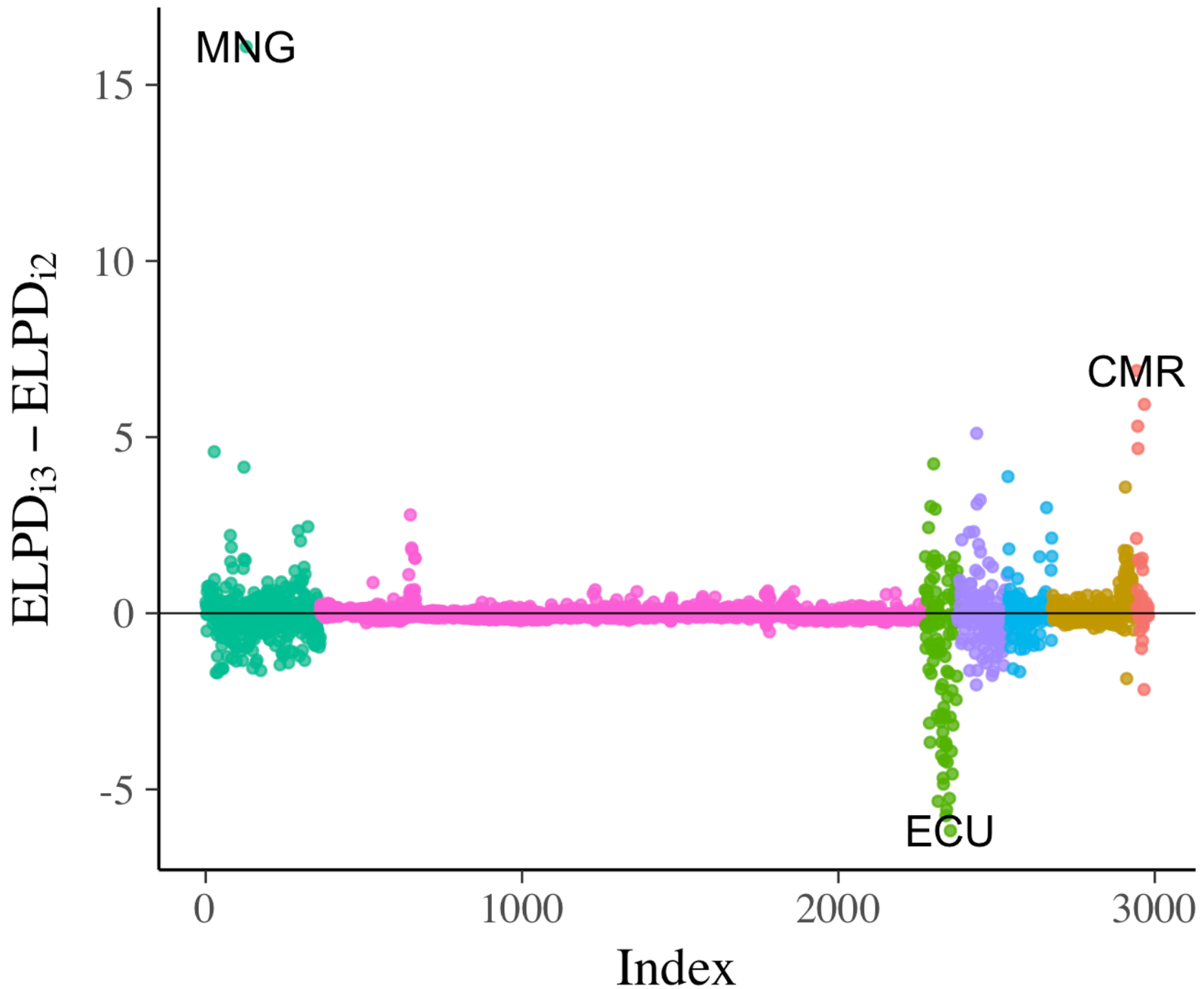
- One thing that can be worth looking at is the predictive distribution we would've had if one observation was missing

$$p(\tilde{y} \mid y_{-i}) \propto \int p(\tilde{y} \mid \theta) p(\theta \mid y_{-i})\, d\theta$$

- This can be computed with self-normalized importance sampling with proposal distribution $g(\theta) = p(\theta \mid y)$ and importance ratios

$$r(\theta) = \frac{1}{p(y \mid \theta)} \propto \frac{p(\theta \mid y_{-i})}{p(\theta \mid y)}$$

MORE THAN JUST COMPUTING A STATISTIC



IDEA: HOW MUCH DOES THE PREDICTIVE CHANGE?

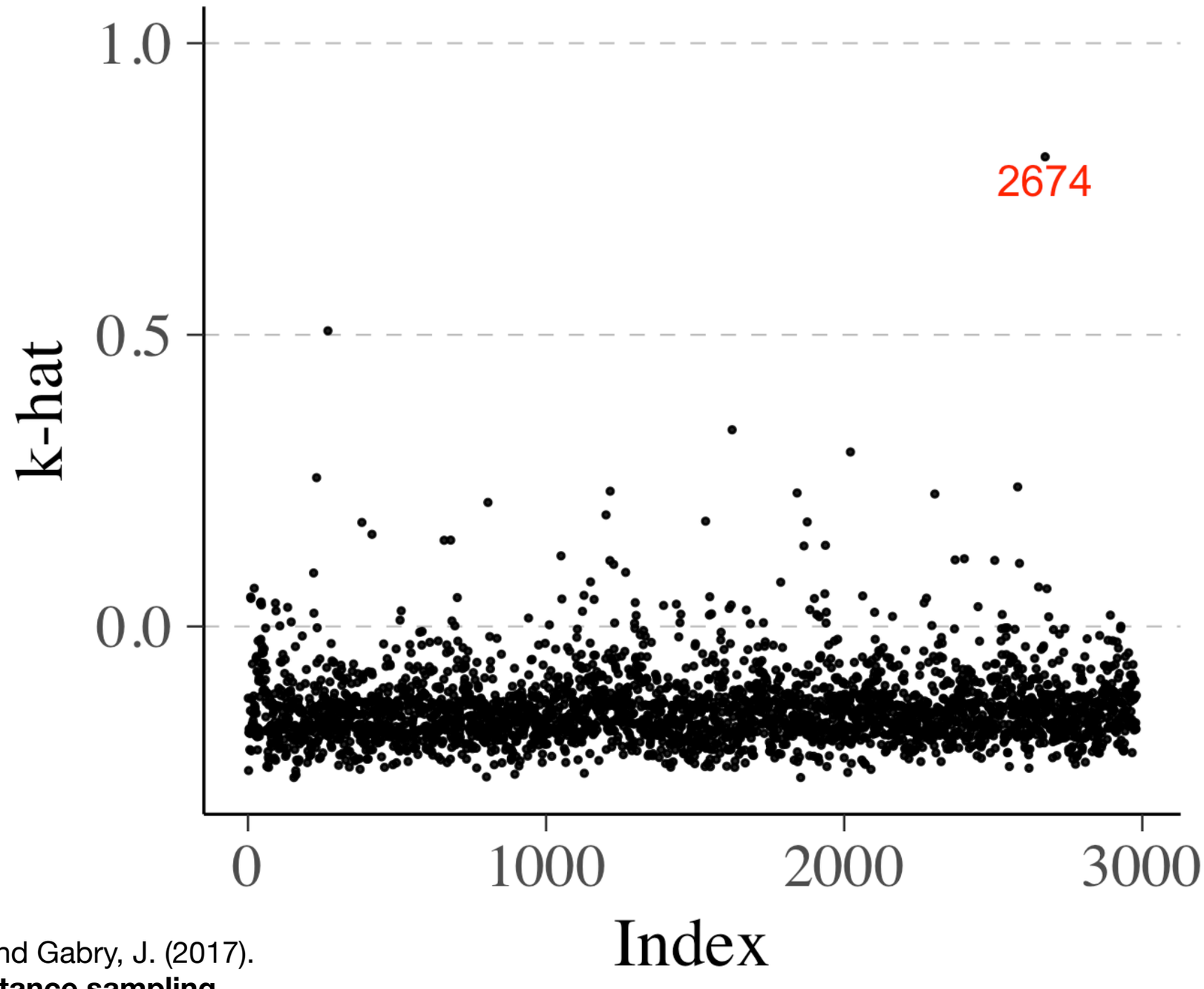
- One thing that is useful to look at is how much the posterior predictive distribution changes when a single data point is left out
- We can do this by looking at k-hat for

$$r(\theta) \propto \frac{p(\theta \mid y_{-i})}{p(\theta \mid y)}$$

- If k-hat is large, this means that adding the ith point greatly changes the posterior, so the inference is sensitive to this observation
- It is strongly related to leverage for linear models (Peruggia, 1997)

DIAGNOSTICS (K–HAT: A PREDICTIVE LEVERAGE)

Mongolia



Vehtari, A., Gelman, A., and Gabry, J. (2017).
Pareto smoothed importance sampling.
working paper arXiv: arxiv.org/abs/1507.02646/

A CONCLUSION WOULD BE NICE

WE'VE ONLY LOOKED AT PLAIN IMPORTANCE SAMPLING

- The Pareto smoothing technique is probably useful for a host of other things.
- Really any time that you use importance sampling and can live with some small bias as a trade off for reliability.
- The k-hat diagnostic has been super useful in a pile of situations (like diagnostics for variational inference, improving inverse probability weighting, stacking predictive distributions, and a bunch other things).
- Theory is annoying.