Insper

Semi-parametric inference for the means of heavy-tailed distributions

Matt Taddy
Hedibert Lopes
David Goldberg
Matt Gardner

Insper Working Paper
WPE: 362/2016

Semi-parametric inference for the means of heavy-tailed distributions

Matt Taddy, Hedibert Lopes, David Goldberg and Matt Gardner

Abstract

Heavy tailed distributions present a tough setting for inference. They are also common in industrial applications, particularly with Internet transaction datasets, and machine learners often analyze such data without considering the biases and risks associated with the misuse of standard tools. This article outlines a procedure for inference about the (possibly conditional) mean of a heavy tailed distribution that combines nonparametric inference for the bulk of the support with parametric inference – motivated from extreme value theory – for the heavy tail. We are able to derive analytic posterior conditional means and variances for the expected value of a heavy tailed distribution. We also introduce a simple and novel independence Metropolis Hastings algorithm that samples from the distribution for tail parameters via small adjustments to a parametric bootstrap, and through this algorithm are able to provide comparisons between our framework and frequentist semiparametric inference. We also provide a modeling extension that shrinks tails across distributions to an overall background tail. We illustrate on two examples: treatment effect estimation on a set of 72 A/B experiments, and the fitting of regression trees for prediction of user spending. Both use data from tens of millions of users of eBay.com.

1. Introduction

We refer to a data generating process (DGP) as *heavy tailed* when the distribution on exceedances beyond extreme thresholds cannot be bounded by an exponential distribution. Heavy tails are quite common in measures of user (e.g., a registered website user or a recognized device) activity on the internet (**??**). For example, Figure 1 illustrates spending, in US$ spent on bought merchandise, across samples of users on eBay.com. Each sample,[1] ranging in size from 10^5 to 10^8 users, corresponds to a treatment group in one of the 72 A/B experiments (randomized controlled trials) studied in Section 5. In our modal treatment group, there is less than 0.1% of users who spend more than $2000; however, these users account for 20% of the total spending.

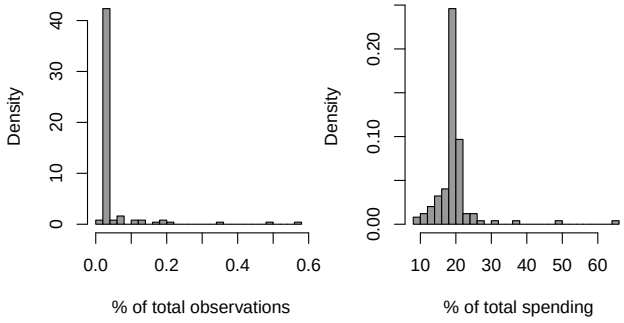


Figure 1. The proportion of observations (left) and of total spending (right) that is due to users spending greater than $2000 on eBay.com in each of 144 A/B experiment treatment groups.

Such heavy tails imply that observations in high percentiles are both high variance and will have a large influence on sample means. In some cases (including, it appears, 3/4 of the groups studied in Section 5) the tail variance will be infinite. Even when the variance is merely near-infinite, these heavy tails have important consequences for our inference.

- The learning rate for mean inference is slower than $\sqrt{n}$, such that the usual standard error estimators underestimate uncertainty (e.g., **?**).
- Common Gaussian error assumptions are invalid both in finite samples and asymptotically (**?**).
- Nonparametric bootstrap estimators of sampling uncertainty about the mean will fail: they are inconsistent for the true sampling distribution (**?**).

There are real practical implications for these issues. For example, the over-sized influence of large observations on the sample mean is well recognized by practitioners who

Preliminary work. Under review by the International Conference on Machine Learning (ICML). Do not distribute.

[1] These contain targeted user subsets and are not from current traffic. They are not representative of eBay's aggregated revenue.

measure on-line transactions. A common ad-hoc solution is to use Winsorization (**?**), wherein values above a threshold are replaced by that threshold. However, estimation is then sensitive to the Winsorization threshold and, due to the inconsistency of the nonparametric bootstrap, there are no obvious tools available for its optimal selection. Transaction distributions also include density spikes at zero and at other discrete points (e.g., 0.99, 100), making fully parametric modeling impractical. As another example, nonparametric learners such as random forests will tend to over-fit (and generally perform poorly) in the presence of extreme values (**?**).

This article resolves these issues by combining nonparametric inference for the bulk of a distribution with parametric inference for tail data above a fixed threshold. Related approaches have been proposed in the literature (see below), but we make a number of unique contributions.

- We are able to derive exact posterior moments for the distribution's mean. These expressions require only the posterior mean and variance for the expected tail exceedance, which we make available via both analytical and efficient computational approximation.
- For inference about the tail parameters, we present a novel independence Metropolis Hastings algorithm that samples from the posterior through adjustment of the results from a parametric bootstrap. The algorithm is trivial to code, and provides information about the distance between Bayesian and frequentist inference.
- Our procedure is closely related to a frequentist algorithm that combines nonparametric and parametric bootstraps. We are able to show that this algorithm is consistent for the true sampling distribution, and the theory provides guidance on the choice of threshold.
- We describe how to shrink individual tails towards an overall background distribution. This is used both for improved estimation for average treatment effects, but also in a novel algorithm for construction of regression trees on heavy tailed data.

Finally, we test and illustrate our work in two real applications: the analysis of A/B experiments and in building robust random forests. The resulting framework is very powerful, and performs better than the alternatives, but is also very simple. We hope that it will find users amongst the large community of analysts dealing with heavy tailed data who currently rely upon sensitive ad-hoc techniques.

1.1. Related Literature

A related Bayesian approach to extreme value analysis is proposed in **?**: they combine a Dirichlet process mixture model below a threshold with a GPD above. All parameters, including the value of the threshold itself, are sampled from their joint posterior in an MCMC algorithm. Our approach is more simple and scalable: we allow for analytic expression of many of the relevant posterior statistics of interest and require only a simple bootstrap-based sampler for the tail.

? describes estimation for the mean of a heavy tailed distribution that combines the sample mean below a threshold with the mean of a maximum likelihood estimated model above that threshold. The point estimates from this approach will be similar to ours, and will converge with enough data, but our approaches to uncertainty quantification are distinct. Johansson's asymptotic variance formulas depend upon unknown model parameters.

? provide a completely different approach to the problem, based upon sub-sampling. While the nonparametric bootstrap fails for heavy tails, Romano & Wolf show that algorithms based upon without-replacement sub-sampling can provide consistent approximations to the sampling distribution. We discuss and compare to their approach in our applications.

Finally, **?** estimate the tail distribution for small samples through exponential tilting of models fit on larger samples. While their approach is totally different from our Bayesian hierarchical shrinkage technique, both works share the strategy of using background datasets to inform difficult estimation for individual tails.

2. A Semiparametric model for heavy tailed data generating processes

Our inference strategy is built around the use of Dirichlet-multinomial sampling as a flexible representation for an arbitrary data generating process (DGP). In its standard application, this model treats the observed sample as a draw from a multinomial distribution over a large but finite set of support points. A Dirichlet prior is placed on the probabilities in this multinomial, and the posterior distribution over possible DGPs is induced by the posterior on these probabilities. The approach has a long history. It was introduced by Ferguson (**?**), it serves as the foundation for the Bayesian bootstrap (**?**), and it has been studied by numerous authors (**?????**).

Our work presents an extension of the standard Dirichlet-multinomial scheme. Consider a univariate random variable, say z. We assume the usual fully-nonparametric model below a certain fixed *threshold*, say u. That is, the prior DGP for $z < u$ is a Multinomial draw, with Dirichlet distributed probability, from a large-but-finite number of support points. At the same time, with some probability our realized z is instead drawn as $u + v$ where $v > 0$ is

a random *exceedance* from some distribution. Hence, the full DGP model is a mixture of Multinomial sampling on arbitrary support and a parametric tail distribution.

We model our tail exceedances as realizations from a generalized Pareto distribution (GPD)[2], with $\mathrm{p}(V < v) = 1 - (1 + \xi v/\sigma)^{-1/\xi}$ and density function on $v > 0$

$$\mathrm{GPD}(e;\ \xi,\ \sigma) = \frac{1}{\sigma}\left(1 + \xi\frac{v}{\sigma}\right)^{-(\frac{1}{\xi}+1)} \tag{1}$$

for tail index $\xi > 0$ and scale $\sigma > 0$. The generalized Pareto is a commonly applied tail model (**?????**) with theoretical justification as the limiting distribution for exceedance beyond large u for a wide family of processes (**???**). As $\xi \to 0$ the GPD converges to an exponential distribution, and for $\xi > 0$ the tails are heavier-than-exponential. For $\xi \geq 1/2$ the variance of v is infinite, and for $\xi \geq 1$ the mean is infinite. Thus our analysis will focus on GPD models with $\xi \in (0, 1)$, so that the tail is heavy enough to cause problems when bootstrapping[3] but not so heavy that the mean does not exist.

Combining the GPD and Dirichlet-multinomial sampling yields our semi-parametric prior for heavy tailed DGPs,

$$g(z;\boldsymbol{\theta}) = \frac{1}{|\boldsymbol{\theta}|}\sum_{l=1}^{L}\theta_l \mathbb{1}_{[z=\zeta_l]} + \frac{\theta_{L+1}}{|\boldsymbol{\theta}|}\mathrm{GPD}(z-u;\ \xi,\ \sigma)\mathbb{1}_{[z\geq u]} \tag{2}$$

where $\boldsymbol{\mathcal{Z}} = \{\zeta_1 \dots \zeta_L\}$, all elements less than u, is the support for the bulk of the DGP $g(z)$[4] and $\boldsymbol{\theta} = [\theta_1 \cdots \theta_{L+1}]'$ is a vector of random weights with $\theta_l \geq 0\ \forall\ l$.[5]

Observations are assumed drawn independently from (2) by first sampling l_i with probability θ_{l_i} and then assigning $z_i = \zeta_{l_i}$ for $l_i \leq L$ and otherwise drawing $z_i - u \sim \mathrm{GPD}$. A posterior over g is induced by the posterior over the model parameters: $\boldsymbol{\theta}$, ξ, and σ. Functionals of g, such as $\mathbb{E}_g f(z)$ for arbitrary function f and where $\mathbb{E}_g$ implies expectation over $z \sim g$, are thus random variables.

2.1. Inference on the sampling weights

A conjugate prior for the weights, $\boldsymbol{\theta}$, places independent exponential distributions on each element: $\theta_l \sim \mathrm{Exp}(a_l)$ for $l = 1, \dots, L+1$, where $\mathbb{E}[\theta_l] = a_l$, $\mathrm{var}(\theta_l) = a_l^2$, and we call $a_l > 0$ the prior 'rate'.[6] We use a single[7] rate parameter a, such that $\mathbf{a} = [a \cdots a]'$. After observing a sample $\mathbf{Z} = [z_1 \cdots z_N]'$, each weight remains independent in the posterior with exponential distribution $\theta_l \mid \mathbf{Z} \sim \mathrm{Exp}\left(a + \sum_{i=1}^{N}\mathbb{1}_{[l_i=l]}\right)$.

We focus on the limiting prior that arises as $a \to 0$; see Chamberlain and Imbens (**?**) and Taddy et al. (**??**) for discussion. This 'non-informative' limit yields a massive computational convenience: as $a \to 0$ the weights for unobserved support points converge to a degenerate random variable at zero: $p(\theta_l = 0|\mathbf{Z}) = 1$ if $l \neq l_i$ for any i. Our posterior for the DGP is then a multinomial sampling model with random positive weights on only the *observed data points* and on the tail model ($l_i = L + 1$).

To simplify notation, say $z_i < u$ for $i \leq m$ and $z_i \geq u$ for $i = m+1, \dots, m+n$ with $N = m + n$. We then overload notation and re-write $\boldsymbol{\theta} = [\theta_1, \dots, \theta_m, \theta_{m+1}]'$ as the posterior vector of weights on observations $z_1, \dots, z_m$ (all less than u) and on the tail. We can then write a posterior DGP realization, conditional upon GPD parameters, as

$$\begin{aligned} g(z)\ |\mathbf{z}, \xi, \sigma \qquad & \\ = \frac{1}{|\boldsymbol{\theta}|}\sum_{i=1}^{m}\theta_i\mathbb{1}_{[z=z_i]} + \frac{\theta_{m+1}}{|\boldsymbol{\theta}|}\mathrm{GPD}(z-u;\ \xi,\ \sigma)\mathbb{1}_{[z\geq u]}, & \\ \theta_i \overset{iid}{\sim} \mathrm{Exp}(1)\ \forall i \leq m \ \text{ and } \ \theta_{m+1} \sim \mathrm{Exp}(n). & \end{aligned} \tag{3}$$

This defines our posterior distribution over DGPs, with details on the GPD tail posterior deferred until Section 3.

2.2. Inference for the DGP mean

The mean of $g(z)$ is a random variable that can be written

$$\mu = \mathbb{E}z = \sum_{i=1}^{m}\frac{\theta_i z_i}{|\boldsymbol{\theta}|} + \frac{\theta_{m+1}}{|\boldsymbol{\theta}|}\left(u + \frac{\sigma}{1-\xi}\right). \tag{4}$$

Uncertainty about $\mathbb{E}z$ is induced by the posterior on weights $\boldsymbol{\theta}$ and on the mean exceedance $\lambda = \sigma/(1-\xi)$. Because u is fixed, we have $\boldsymbol{\theta} \perp\!\!\!\perp \lambda$.

It is easy to see that $\mathbb{E}\mu = \frac{1}{m+n}\sum_{i=1}^{m} z_i + \frac{n}{m+n}(u + \mathbb{E}\lambda)$, while the law of total variation yields posterior variance $\mathrm{var}\mu = \mathbb{E}[\mathrm{var}(\mu|\lambda)] + \mathrm{var}(\mathbb{E}[\mu|\lambda])$. Given the properties

[2] However, our development does not depend upon this specific tail model; you can replace the GPD with your preferred distribution while making use of the ideas in this paper.

[3] If you estimate $\hat{\xi} \approx 0$, then standard Bayesian bootstrap methods (e.g., **?**) should apply and there is no need to model a parametric tail.

[4] We will often suppress $\boldsymbol{\theta}$ and write $g(\cdot)$ for $g(\cdot;\ \boldsymbol{\theta})$ unless the weights need to be made explicit.

[5] $|\boldsymbol{\theta}|$ denotes $\sum_i |\theta_i|$, the L_1 norm.

[6] This is equivalent to the more common specification of a Dirichlet distribution on the normalized weights, written $\mathrm{Dir}(\boldsymbol{\theta}/|\boldsymbol{\theta}|; \mathbf{a}) \propto \prod_{l=1}^{L+1}(\theta_l/|\boldsymbol{\theta}|)^{a_l - 1}$.

[7] One could also use multiple rate parameters; e.g., a larger value a_l on θ_l associated with $\zeta_l = 0$ if zeros are common, or a larger a_{L+1} on θ_{L+1} corresponding to the tail data.

of the Dirichlet posterior on $\boldsymbol{\theta}/|\boldsymbol{\theta}|$, the first term is

$$\begin{aligned}\mathbb{E}[\mathrm{var}(\mu|\lambda)] &= \mathbb{E}\left[\frac{\sum_{i=1}^m (z_i-\mu_\lambda)^2 + n(u+\lambda-\mu_\lambda)^2}{(m+n)(m+n+1)}\right] \\ &= \frac{\sum_{i=1}^m (z_i-\mathbb{E}\mu_\lambda)^2 + n(u+\mathbb{E}\lambda-\mathbb{E}\mu)^2}{(m+n)(m+n+1)} \\ &\quad + \frac{n^2(m+n-2)\mathrm{var}(\lambda)}{(m+n)^2(m+n+1)} \end{aligned} \tag{5}$$

where $\mu_\lambda = \left[\sum_{i=1}^m z_i + n(u+\lambda)\right]/(m+n)$ and $\mathbb{E}\mu$ is the posterior expectation from above. The second term is

$$\mathrm{var}(\mathbb{E}[\mu|\lambda]) = \frac{n^2}{(m+n)^2}\mathrm{var}(\lambda) \tag{6}$$

and thus the full expression is

$$\begin{aligned}\mathrm{var}\mu =& \frac{\sum_{i=1}^m (z_i-\mathbb{E}\mu)^2 + n(u+\mathbb{E}\lambda-\mathbb{E}\mu)^2}{(m+n)(m+n+1)} \\ &+ \frac{2n^2(m+n-0.5)}{(m+n)^2(m+n+1)}\mathrm{var}(\lambda).\end{aligned} \tag{7}$$

As detailed below, $\mathbb{E}\lambda$ and $\mathrm{var}(\lambda)$ are available through either Laplace approximation or Markov chain Monte Carlo.

3. Inference for tail parameters

In this section we describe Bayesian modeling and posterior inference for the GPD parameters, ξ and σ, conditional upon the sample of exceedances $\{v_i = z_{m+i} - u\}_1^n$.

As discussed above, we are focusing on heavy tails with finite mean exceedences that correspond to $\xi \in (0,1)$. On this range, σ can take any positive value. A simple independent prior setup would then be

$$\begin{aligned}\pi(\sigma,\xi) &= \mathrm{Be}(\xi;a,b)\mathrm{Ga}(\sigma;c,d) \\ &\propto \xi^{a-1}(1-\xi)^{b-1}\sigma^{c-1}e^{-d\sigma}\end{aligned} \tag{8}$$

where $\mathrm{Ga}(\cdot\,;\,a,b)$ denotes a beta density with mean $a/(a+b)$ and $\mathrm{Ga}(\cdot\,;\,c,d)$ a gamma density with mean c/d, with $a,b,c,d > 0$. We will work with versions of this prior throughout, however our sampling algorithm is trivially adjusted to work for alternative specifications.

A useful version of (8) sets $a = b = 1$ and takes the noninformative limit $c,d \to 0$ to obtain

$$\pi(\sigma,\xi) = \frac{1}{\sigma}\mathbb{1}_{\xi\in(0,1)}, \tag{9}$$

the combination of a unit uniform on ξ and an improper uniform prior on $\log\sigma$. Following **?** and **?**, the posterior for GPD parameters will be proper under the prior in (9) given a minimum of three observations.[8] This noninformative default is used the absence of any background information.

[8] This holds under any prior specification that combines $\pi(\sigma) \propto \sigma^{-1}$ with an independent proper prior on ξ.

Finally, this specification combines with the GPD likelihood to yield a log posterior proportional to

$$\begin{aligned}l(\sigma,\xi) = &-n\log\sigma - \left(\frac{1}{\xi}+1\right)\sum_i \log\left(1+\xi\frac{v_i}{\sigma}\right) \\ &+ (a-1)\log\xi + (b-1)\log(1-\xi) + (c-1)\log\sigma - d\sigma,\end{aligned} \tag{10}$$

which simplifies considerably under the model in (9).

Maximization of (10) leads to MAP estimates of the parameters, say $[\hat{\xi},\hat{\sigma}]$. The related problem of MLE estimation for GPDs is well studied by **?** and his algorithm is easily adapted for fast MAP estimation within our domain $[\xi,\sigma] \in (0,1)\times\mathbb{R}^+$.

3.1. Laplace posterior approximation

The main object of interest is actually the posterior for the GPD mean, $\sigma/(1-\xi)$. We make the transformation

$$\lambda = \frac{\sigma}{1-\xi} \Leftrightarrow \sigma = \lambda(1-\xi), \tag{11}$$

with inverse Jacobian $|J| = 1-\xi$, to obtain the posterior

$$\begin{aligned}&\mathrm{p}(\lambda,\xi \mid \mathbf{v}) \propto \\ &\frac{\xi^{a-1}e^{-d\lambda(1-\xi)}}{\lambda^{n-c+1}(1-\xi)^{n-b-c+1}}\prod_i\left(1+\frac{\xi}{1-\xi}\frac{v_i}{\lambda}\right)^{-\left(\frac{1}{\xi}+1\right)}.\end{aligned} \tag{12}$$

Note that the MAP estimate for λ is just $\hat{\lambda} = \hat{\sigma}/(1-\hat{\xi})$. The Laplace approximation (**?**) to the *marginal* posterior distribution on λ is available as

$$\hat{\mathrm{p}}(\lambda \mid \mathbf{x}) = \mathrm{N}\left(\hat{\lambda}, -\nabla_{\lambda\lambda}^{-1}\big|_{[\hat{\lambda},\hat{\xi}]}\right), \tag{13}$$

where $\nabla_{\lambda\lambda}$ is the curvature of the log posterior [9] with respect to λ. Thus

$$\begin{aligned}\mathrm{var}(\lambda \mid \mathbf{x}) &= -\nabla_{\lambda\lambda}^{-1}\big|_{[\hat{\lambda},\hat{\xi}]} \\ &= -\hat{\lambda}^2\left[n-c+1+\left(\frac{1}{\hat{\xi}}+1\right)\sum_i\left(\hat{q}_i^2-2\hat{q}_i\right)\right]^{-1},\end{aligned} \tag{14}$$

with $\hat{q}_i = \hat{\xi}v_i/\left[(1-\hat{\xi})\hat{\lambda}+\hat{\xi}v_i\right]$, is the approximate posterior variance for λ.

3.2. Independence-MH via the parametric bootstrap

For small tail samples the Laplace approximation will dramatically underestimate posterior uncertainty. Instead, we propose a novel independence Metropolis Hastings (i-MH) algorithm (see, e.g., **?**) that uses a parametric bootstrap of

[9] Via $\partial\log\mathrm{p}(\lambda,\xi|\mathbf{v})/\partial\lambda = \left[(1/\xi+1)\sum_i q_i - n + c - 1\right]/\lambda - d(1-\xi)$ where $q_i = \xi v_i/\left[(1-\xi)\lambda+\xi v_i\right]$.

the MAP estimates as a proposal distribution in Markov Chain Monte Carlo. This approach is similar in spirit to the bootstrap reweighting of (**?**), but unlike that work it does not require an analytic expression for the sampling distribution of the statistics of interest. The algorithm proceeds as follows.

- Fit the MAP parameter estimates $[\hat{\xi}, \hat{\sigma}]$.
- Obtain B draws $[\hat{\xi}_b, \hat{\sigma}_b]$ from the parametric bootstrap:
  - Generate a sample of size n by simulating from the MAP estimated GPD model.
  - Obtain new MAP estimates $[\hat{\xi}_b, \hat{\sigma}_b]$ conditional upon this simulated sample.
- Estimate the bivariate bootstrap distribution, say $r(\xi, \sigma)$, via kernel smoothing on $\{\hat{\xi}_b, \hat{\sigma}_b\}_{b=1}^{B}$.
- For $b = 2 \dots B$, replace $[\hat{\xi}_b, \hat{\sigma}_b]$ with $[\hat{\xi}_{b-1}, \hat{\sigma}_{b-1}]$ with probability

$$1 - \min\left\{\frac{r(\hat{\xi}_{b-1}, \hat{\sigma}_{b-1})\exp[l(\hat{\xi}_b, \hat{\sigma}_b)]}{r(\hat{\xi}_b, \hat{\sigma}_b)\exp[l(\hat{\xi}_{b-1}, \hat{\sigma}_{b-1})]},\ 1\right\}$$

  where l is the log posterior objective in (10).[10]

In addition to being fast and simple, this algorithm offers a bridge between Bayesian and frequentist inference: if the acceptance probabilities are high, then there is little difference between the sampling distribution and the posterior.

4. A semiparametric heavy tailed bootstrap

This section studies a bootstrap algorithm that is closely related to our semiparametric Bayesian procedure. Consider frequentist inference about $Q_N = \sqrt{N}(\hat{\mu}_N - \mu)$ for a sample of N observations drawn from true distribution function $F(z)$, with $\int_0^\infty z dF(z) = \mu < \infty$ and where $\hat{\mu}_N$ denotes the MLE for μ based upon a size-N sample from F.

A frequentist bootstrap replaces $F \approx \widehat{F}_N$ and uses this to obtain $b = 1, \dots, B$ draws of $Q_N^b = \sqrt{N}(\hat{\mu}_N^b - \hat{\mu}_N)$ where $\hat{\mu}_N^b$ is the MLE based upon a size-N sample from $\widehat{F}_N$. The targeted sampling distribution, $G_N(q) = \mathrm{p}(Q_N < q)$, is then estimated as $\widehat{G}_N(q) = B^{-1}\sum_{b=1}^{B} \mathbb{1}_{[Q_N^b < q]}$.

Standard results on bootstrap consistency (**??**) require that $\widehat{G}_N$ converges in distribution (i.e., weakly) to G_∞ *uniformly* across all $\widehat{F}_N$ in a neighborhood, say $\mathcal{F}$, containing F and also $\widehat{F}_N$ for N big enough.[11] Convergence in probability for $\widehat{F}_N(z)$ to $F(z)$ $\forall z$ then implies consistency of $\widehat{G}_N$ in that, as $N \to \infty$, $\mathrm{p}(|\widehat{G}_N(q) - G_N(q)| < \epsilon) \to 0$ for all $q, \epsilon > 0$.

? shows that the nonparametric bootstrap – using the empirical distribution function (EDF) as $\hat{F}_N$ – is inconsistent for the distribution of the sample mean for data that has infinite variance. As explained by **?**, in this setting $\widehat{G}_N$ based upon samples from $\hat{F}_N$ does not converge uniformly to G_∞ because sums of the largest re-sampled observations, $\sum_{i=N-r}^{N} z_{(i)}^b$ for $r \geq 1$, can be dominated by repeats of the largest sample observation, $z_{(N)}$.

Instead, define a semiparametric bootstrap that takes the MLE tail parameters, $[\hat{\xi}_n, \hat{\sigma}_n]$, and for $b = 1, \dots, B$

- draw $m_b \sim \mathrm{Bin}(m/N, N)$ and set $n_b = N - m_b$;
- sample with replacement m_b observations z_i where $i \leq m$ (i.e., $z_i < u$), say $\{z_1^b, \dots, z_{m_b}^b\}$;
- generate $v_1^b \dots v_{n_b}^b$ from $\mathrm{GPD}(\hat{\xi}_n, \hat{\sigma}_n)$ and obtain the corresponding MLE, $\hat{\lambda}_{n_b}^b = \hat{\sigma}_{n_b}^b/(1 - \hat{\xi}_{n_b}^b)$; and
- set $\hat{\mu}_N^b = \left(\sum_{i=1}^{m_b} z_{i_b} + n_b(u + \hat{\lambda}_{n_b}^b)\right)/N$.

The results can then be applied in estimation of the sampling distribution, e.g., for $\sqrt{N}(\hat{\mu}_N - \mu)$ as approximately equal to the sampled distribution on $\sqrt{N}(\hat{\mu}_N^b - \hat{\mu}_N)$.

The distribution for $\hat{\mu}_N$ implied by our semiparametric bootstrap is the combination of three bootstrap estimators, for distributions on $\frac{1}{m}\sum_{i=1}^{N} z_i \mathbb{1}_{[z_i < u]}$, on m/N, and on $\hat{\lambda}_n$. Consistency of the nonparametric bootstrap for the first two statistics[12] can be established through standard arguments (**?**). To show consistency for our semiparametric bootstrap, we need to confirm that the parametric bootstrap using $\widehat{F}_N(z - u | z \geq u) = \mathrm{GPD}(\hat{\xi}_n, \hat{\sigma}_n)$ converges to the correct distribution for $\hat{\lambda}_n$.

? considers DGPs with distribution functions $F(z) = 1 - cz^{-1/\zeta}(1 + z^{-\delta}L(z))$, where $c, \delta > 0$ and $L(tz)/L(z) \to 1$ with $z \to \infty$ for $t > 0$. This defines a wide class of heavy tailed distributions, and for u_N large enough the distribution $F(z - u_N | z \geq u_N)$ approaches a $\mathrm{GPD}(\xi, \sigma_N)$ where $\sigma_N = u_N\xi$. Following the same steps as Johansson, which apply results from **?** on the asymptotic distribution for MLEs $[\hat{\xi}_n, \hat{\sigma}_n]$, you can show that for $F(z)$ with $\xi \in (0, 1)$ and $z^{-\delta}L(z)$ non-increasing, if $u_N = O(N^{\xi/(1+2\delta\xi)})$ then

$$\sqrt{n}(\hat{\lambda}_n - \mathbb{E}_F[z - u_N | z \geq u_N]) \to_d \mathrm{N}(0, q_n) \qquad (15)$$

where $q_n = \hat{\sigma}_N(1 + \xi)(1 - \xi + 2\xi^2)/(1 - \xi)^4$. Thus our bootstrap sample generator, $\mathrm{GPD}(\hat{\xi}_n, \hat{\sigma}_n)$, converges

[10] This 'rejection' probability – the probability of *not* moving states – is equal to one minus the familiar acceptance probability.

[11] In addition, the mapping $F \mapsto G_\infty$ must be continuous.

[12] The first is the mean from sampling a bounded domain, the second is the proportion of successes in a binomial trial of size N.

to $F(z - u_N | z \geq u_N)$ along a sequence of distributions with means $\hat{\lambda}_n$ that are asymptotically normal around the target of interest, $\mathbb{E}_F[z - u_N | z \geq u_N]$. From **?**, this is enough to establish consistency of this tail bootstrap, and hence of our full semiparametric bootstrap.

Intuitively, the parametric tail bootstrap succeeds here because our MLEs allow us to converge quickly to the 'true' GPD model; inference is then based upon *new* samples from this distribution and, unlike resamples from the EDF, these are not overly influenced by large order statistics in the original sample. From (15), this convergence holds so long as u_N is growing at the right rate. Theoretically, the true GPD tail that we converge to with u_N has $\sigma_N = \xi u_N$. One can compare this σ_N to $\hat{\sigma}_n$, the sample MLE for the same parameter, to see if they roughly agree. If not, u_N might need to increase. In practice, however, we also advocate repeating inference over a range of thresholds u and using results from the region where they stabilize.

We now return to a Bayesian framework for the remainder of this article, but the reader who is interested in frequentist inference may feel free to use the semiparametric bootstrap of this section instead. The two algorithms are very closely related: there is little difference between the Bayesian bootstrap and the nonparametric bootstrap on the bulk of the distribution (**?**) and the i-MH sampler of Section 3.2 explicitly agrees with the frequentist parametric tail bootstrap whenever acceptance probabilities are close to one. Given this connection, the discussion above indicates that our semiparametric Bayesian inference will also have good frequentist properties in large samples.

5. A/B Experiments

Our motivating application for these ideas involves A/B experiments: two independent heavy tailed samples are obtained, one from a group receiving a treatment and another from a group receiving a control. The object of interest is then

$$\gamma = \mu_1 - \mu_0 \tag{16}$$

where μ_1 is the mean of the treatment group and μ_0 the mean of the control group.

Our main inferential target is the average treatment effect in an *A/B* experiment: the difference in *average* response, say z, between a group that received some treatment (e.g., a change to the website, some marketing, a new search algorithm) and a control group that experiences the status quo. That is, for a binary treatment indicator d, where $d_i = 1$ for treated and $d_i = 0$ for untreated, we seek to estimate the treatment group means $\mu_d = \mathbb{E}[z|d]$ and the average treatment effect

$$\gamma = \mu_1 - \mu_0. \tag{17}$$

We wish to estimate and test for $\gamma \neq 0$ in a setting where each treatment group has $\mathrm{var}(z|d) = \infty$.

A standard analysis estimates γ with $\bar{z}_1 - \bar{z}_0$, the difference of sample means. A naive standard error for this estimate is $\sqrt{\mathrm{sse}(z_0)/N_0^2 + \mathrm{sse}(z_1)/N_1^2}$, where $N_d = m_d + n_d$ in the notation from above: m_d and n_d are the numbers of observations below and at-or-above the tail cutoff, respectively. The results for such an analysis are shown in Figure 2; out of our 72 experiments there are 8 significant p-values at a 10% False Discovery Rate. However, both point and standard-error estimates here will be sensitive to extreme values in the sample. Moreover, the standard error formula is incorrect for distributions with infinite variance, where the learning rate is less than $\sqrt{N_d}$.

there is a prior belief that the experiments do not make much of a difference in the tail. That is, we suspect that in most cases the users spending a lot of money will act the same regardless of what website experience they are met with. Unfortunately, for infinite variance data it is common that a few very large observations can have an outsized influence on the mean estimates (and those of γ). Thus we wish to accurately quantify uncertainty in the presence of heavy tails and avoid having them overly influence our inference.

5.1. Tail analysis results

Using a cutoff of $u = 2000$, we analyzed the exceedances across our 72 experiments. Four example experiments are in Figure 3; each analysis used 10,000 iterations of the I-MH sampler.

- Compared to results for the MCMC (I-MH) posterior samples, the Laplace approximation variance is far too small.
- The parametric bootstrap proposal distribution is very similar to the I-MH sampler. Indeed, 85% of our experiments the average acceptance probability was greater than 0.9. Thus our tail analysis is very similar to that of a parametric bootstrap.
- Posterior predictive fits look good. As predicted by extreme value theory, these improve even more for larger u (e.g., $u = 3000$) at the expense of smaller tail-sample-sizes.
- All tail indices ξ have 100% of posterior weight between 0 and 1. This implies that the exceedences have a finite mean but infinite variance.

5.2. Semiparametric Bayesian results

Combining the parametric tail analysis with nonparametric inference below the cutoff, we obtain a new set of effect and uncertainty estimates. The plot in Figure 4 shows inference

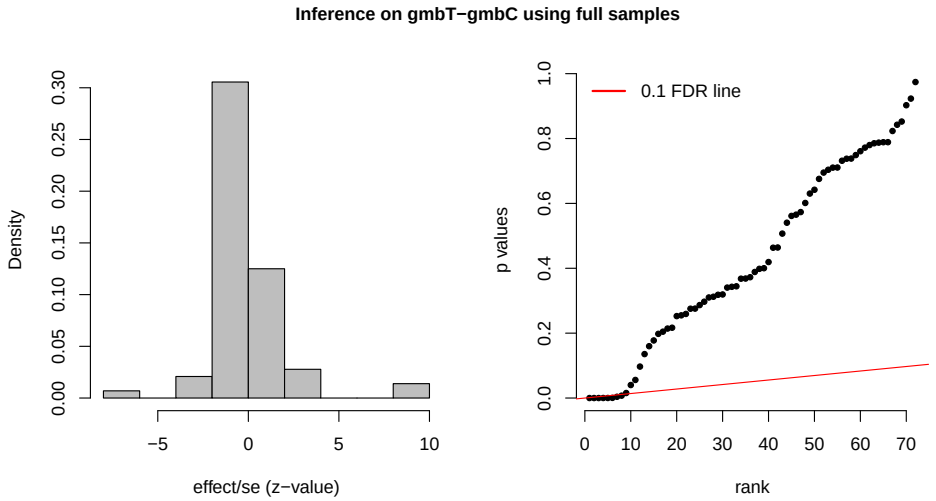


Figure 2. Inference using naive sample mean differences and standard errors.

analogous to our earlier results in Figure 2 (we are abusing concepts and reporting standardized effects and p-values based on the Bayesian posterior moments). The significant effect estimates are roughly similar: we'd make nearly the same decisions at 10% FDR. However, there are bigger differences than it first appears. In the next section we see, for example, that the naive confidence intervals can be far different from the correct posterior intervals.

5.3. Capping

We now investigate the effect of *capping* on estimation. In this procedure, one replaces response values above a certain 'cap' with that value. It is viewed as a variance reduction technique. Results for four of our experiments are shown Figure 5; in each the naive sample mean differences and standard errors are applied to the capped values, as well as to the raw uncapped values (red line), and we compare to the semiparametric Bayesian analysis (for $u = 2000$).

- The point estimates from higher capped values are near to the posterior mean than those from either low caps or the uncapped values.
- The capped value standard errors are always much smaller than the true posterior standard deviation. Note that these standard errors are correct for the mean of capped values; it is for the true overall mean – our object of interest – that they underestimate uncertainty.
- A safe-ish procedure appears to be to use a high cap to get the point estimate, but use the standard error calculated on the raw values to assess uncertainty. Note that, since we're learning at a rate $< \sqrt{N}$, these values still likely tend to underestimate uncertainty.

5.4. Sensitivity to cutoff

We consider how the semiparametric Bayesian analysis responds to changes in the tail cutoff, u. Figure 6 shows that results are quite stable to this choice; e.g., compare to figure 3 to see that the capping procedure is much more sensitive even over this limited range.

6. Extensions

There are some nice things we could do with this.

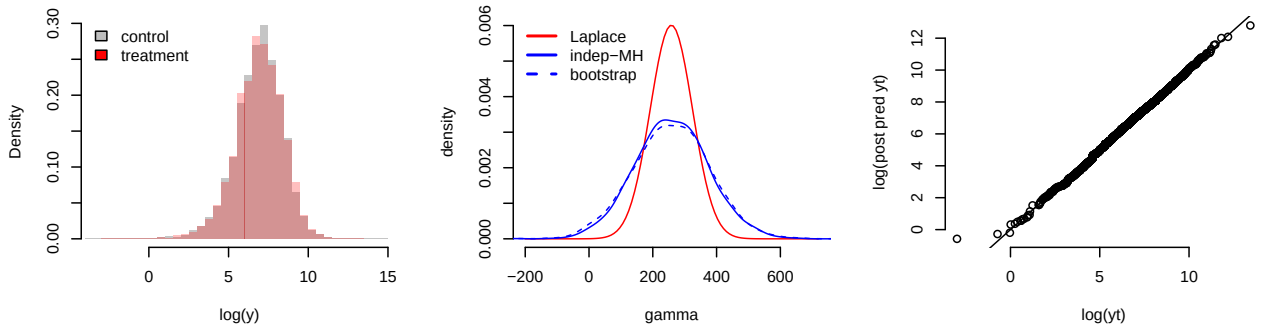
6.1. Hierarchical modeling

We (or rather eBay analysts) have a strong prior that the treatments do not make much of a difference in the tail. We could build this into the analysis by shrinking all of the tails to an overall mean. An efficient Empirical Bayes procedure would first calculate a GPD model for the aggregate tail, then use this to create a 'prior' model for the individual tail GPD parameters. You could also have things evolve in time, etc.

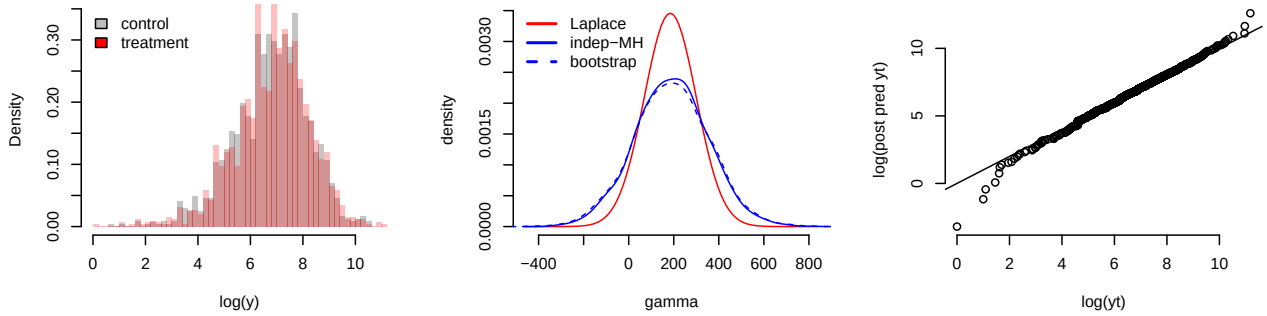
6.2. Heavy tailed regression

These same considerations apply to any regression model that we might want to fit where the response is heavy tailed (e.g., if the response is GMB). It would be good to look at both ordinary least-squares (OLS) and trees (CART, random forests) to see what we can improve.

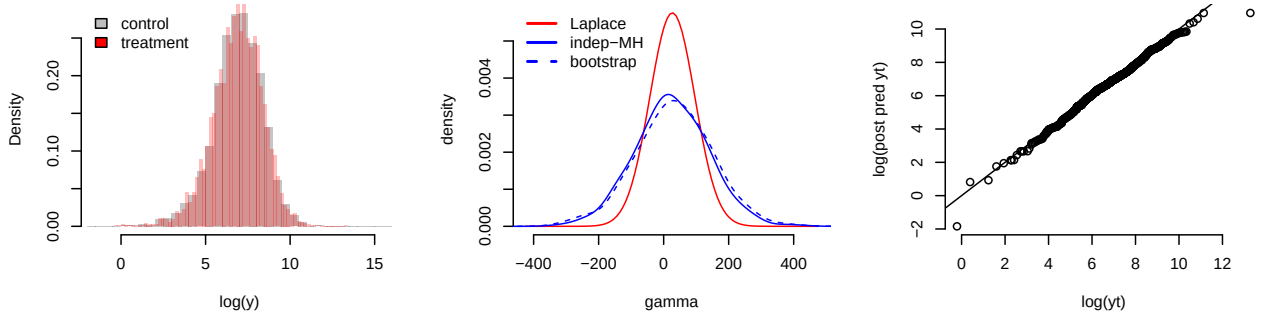
treatment 33203



treatment 32350



treatment 33444



treatment 32099

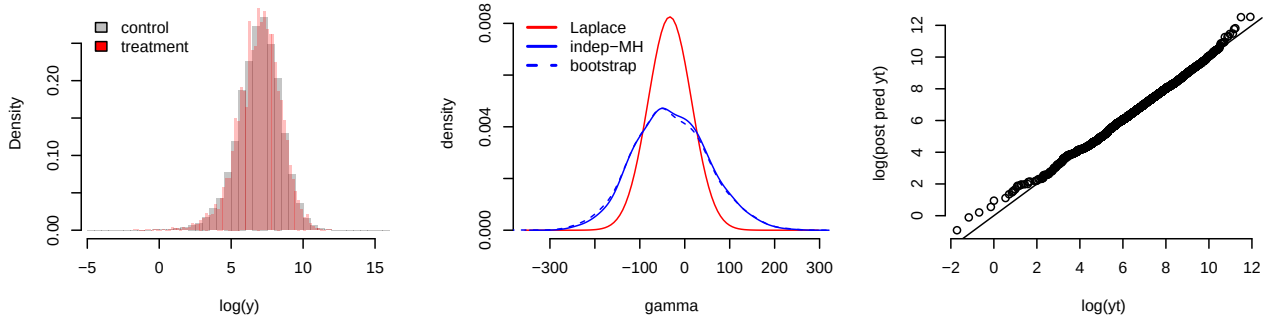


Figure 3. Tail analysis for exceedances beyone $z = 2000$. The left panel shows the histograms of exceedances by treatment group, the center panel shows both Laplace approximate and Independence MH posterior distributions for the treatment effect in the tail ($\lambda_1 - \lambda_0$), and the right plot compares quantiles for the observed tail data in the treatment group to those of a posterior predictive sample of the same size. The dashed lines in the center panel are the parametric bootstrap proposal distribution.

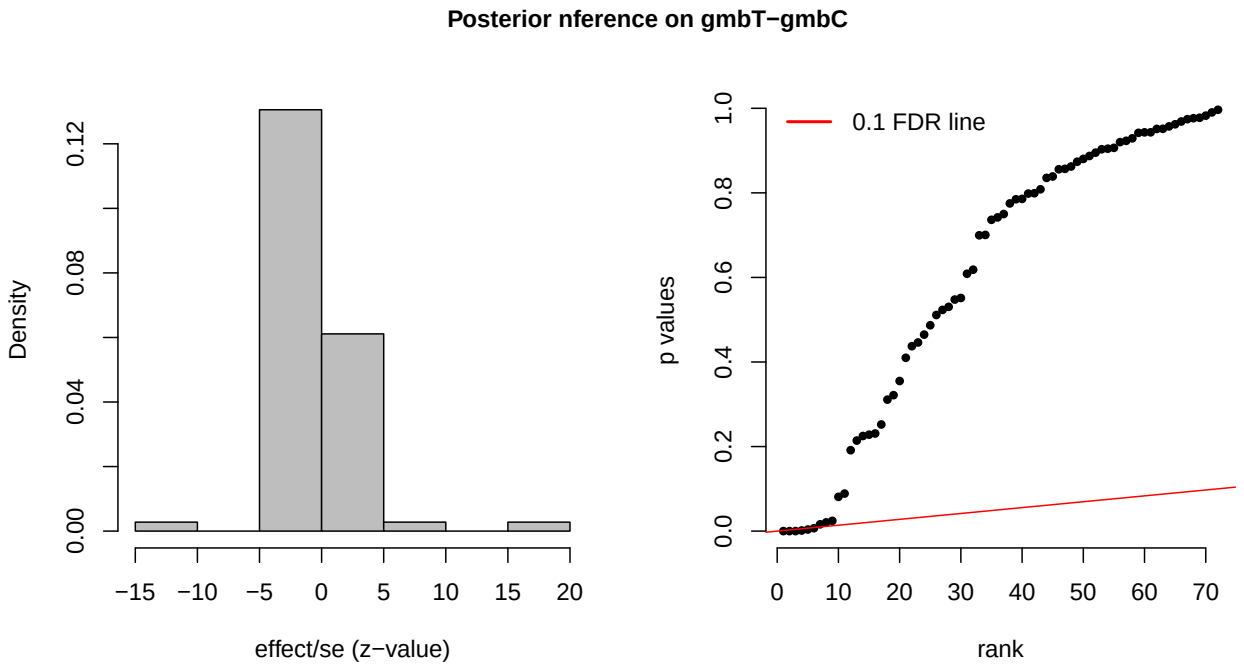


Figure 4. Posterior inference results when combining the parametric tail analysis with nonparametric inference below the cutoff at $u = 2000$. The 'effects' here are posterior means, while the 'se' are posterior standard deviations. For comparison with Figure 2, we report results on the the same scale of standardized effects and Normal distribution 'p-values'.

7. Conclusion

This is great Big Data analysis: avoiding modeling and difficulty for the easy bits (the middle of the distribution), and applying good solid statistical modeling on the hard bits (the tail). Since the posterior distribution and the parametric bootstrap for the tail models are virtually indistinguishable, you don't need to be Bayesian to buy the conclusions.

A. MAP estimation under the default prior

Maximization of the log posterior in (10) is unaffected by a re-parametrization[13], so we make the convenient translation $\alpha = 1/\xi$ and $\tau = \xi/\sigma$ to obtain the maximization objective

$$l(\alpha, \tau) = (n+1)(\log\alpha + \log\tau) - (\alpha+1)\sum_i \log(1+\tau v_i). \quad (18)$$

A similar replacement is advocated for maximum likelihood estimation in (**?**)[14]. The gradient with respect to α is

$$\nabla_\alpha = \frac{n+1}{\alpha} - \sum_i \log(1+\tau v_i). \quad (19)$$

This is solvable as a function of τ, so that at the maximum of (18) we can make the replacement

$$\hat{\alpha}(\tau) = \frac{n+1}{\sum_i \log(1+\tau v_i)}. \quad (20)$$

By substitution into (18), the resulting *profile* objective function is then proportional to

$$l(\tau) = \log[\tau\hat{\alpha}(\tau)] - \frac{1}{\hat{\alpha}(\tau)} \quad (21)$$

with gradient

$$\nabla_\tau = \frac{n+1}{\tau} - (\hat{\alpha}(\tau)+1)\sum_i \frac{v_i}{1+\tau v_i}. \quad (22)$$

Obtaining roots for ∇_τ over the domain $\tau > 0$ is straightforward. We just note that MAP $\hat{\tau}$ will be very small for many fat tailed datasets and thus you need to set an appropriately small convergence tolerance (10^{-8} is fine for our examples). Finally, the roots for (22) yield MAP estimates

$$\left[\hat{\xi},\ \hat{\sigma}\right] = \left[\frac{1}{\hat{\alpha}(\hat{\tau})},\ \frac{1}{\hat{\tau}\hat{\alpha}(\hat{\tau})}\right]. \quad (23)$$

[13] Note that we are just solving for a maximum, not translating the distribution to a different parameter space.

[14] Where $k = -\xi$.

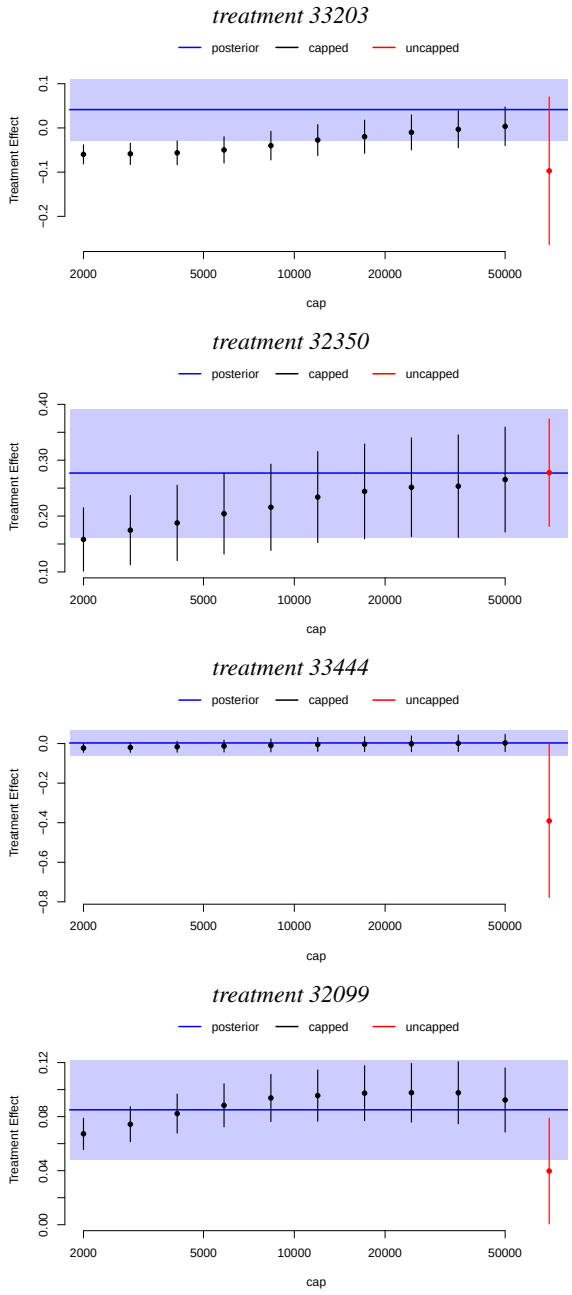


Figure 5. A study of capping (and naive uncapped analysis) in comparison to the semiparametric Bayesian posterior. In each case, intervals plotted are $\pm 1\mathrm{sd}$ or $\pm 1\mathrm{se}$, as appropriate.

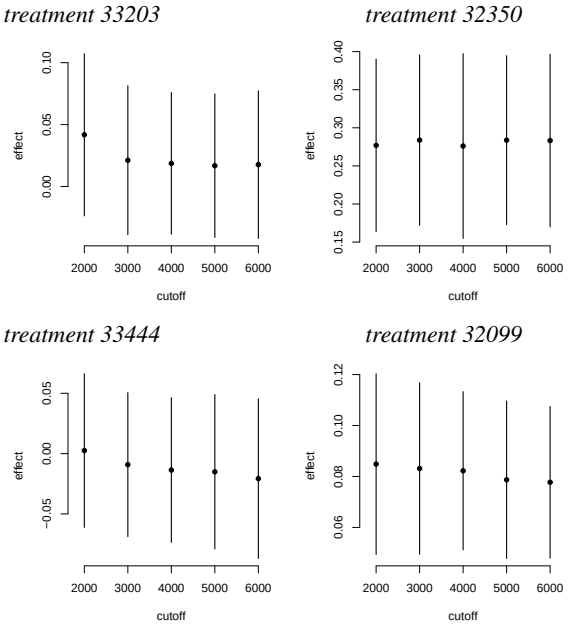


Figure 6. The posterior mean for γ, plus or minus 1 standard deviation, as a function of tail cutoffs $u = 2, 3, 4, 5$ thousand.