Semiparametric Exponential Families for Heavy-Tailed Data

William Fithian and Stefan Wager
Department of Statistics, Stanford University
{wfithian, swager}@stanford.edu

May 28, 2014

Abstract

We propose a semiparametric method for fitting the tail of a heavy-tailed population given a relatively small sample from that population and a larger sample from a related background population. We model the tail of the small sample as an exponential tilt of the better-observed large-sample tail using a robust sufficient statistic motivated by extreme value theory. Our method can be used to produce estimates of the small-population mean; in this case we give both theoretical and empirical evidence that our approach outperforms methods that do not use the background sample. We apply our method to both simulated data and a large controlled experiment conducted by an internet company, and demonstrate substantial efficiency gains over competing methods.

Keywords: Exponential family, extreme value theory, importance sampling, semiparametric estimation.

1 Introduction

We study the problem of efficiently estimating the tail of a distribution given a medium-sized sample from a heavy-tailed population of interest $X_i \sim F(x)$, $i = 1, \ldots, n$, and a much larger background sample from a qualitatively similar but non-identical population $Y_i \sim F_0(x)$, $i = 1, \ldots, N \gg n$. This setting arises frequently in modern statistical applications. As a motivating example, consider an internet company with millions of customers that wants to apply some treatment (a change to the site) for a random subset of users and estimate the effect of the treatment on revenue. Our goal is to understand the distribution of the smaller treatment sample, while all the other customers who did not see the treatment act as a background sample. In many cases, the distribution of revenue is heavy-tailed, say, 10% of the company’s revenue comes from the top 0.1% of its users, and so understanding the tail of $F$ is a crucial part of estimating functionals of interest such as the mean.

Estimating the tail of $F$ from the sample $\{X\}$ alone is difficult, as few of the $X$ values are directly relevant to estimating the tail. Direct approaches that only use the sample $\{X\}$ either suffer from highly variable tail estimates, or are forced to make strong parametric assumptions that can lead to substantial bias. In this paper, we use the background dataset to navigate this trade-off, and to produce stable yet accurate estimates of the tail of $F$. 
Our key assumption is that the tails of $F$ and $F_0$ are similar enough that we can model the tail of $F$ as a perturbation of the tail of $F_0$. If so, we can translate stable estimates of the tail of $F_0$ into good estimates of the tail of $F$. In the spirit of Efron and Tibshirani [1996], we model the tail of $F$ as an exponential tilt of the tail of $F_0$. For some threshold $t$ (a tuning parameter), define $G(x)$ to be the conditional tail law $L_F(X - t | X > t)$, and $G_0$ analogously. We then model $G$ using an exponential family with carrier measure $G_0$:

$$dG(x) = e^{\eta T(x) - \psi(\eta)} dG_0(x).$$

The performance of any such approach depends crucially on the choice of sufficient statistic $T(x)$ used to construct the exponential family. By exploiting results from extreme value theory, we derive a sufficient statistic that is specially tailored for fitting the tails of heavy-tailed distributions.

Our semiparametric framework is closely related to density ratio models [Fokianos et al., 2001, Fokianos, 2004, Huang and Rathouz, 2012, Tan, 2009], which are usually fit by empirical likelihood methods [Owen, 2001]. In particular, de Carvalho and Davison [2014] use a density ratio model for joint estimation of multivariate extremes in the presence of covariate information. Our sufficient statistic could also be used for density ratio modeling with heavy-tailed data.

We begin by presenting our method as a generic approach to estimating $F$, with an emphasis on the tail. In the second half of the paper, we then focus on the behavior of the mean estimate $\hat{\mu}$ induced by $\hat{F}$ and analyze its behavior both theoretically and empirically. Our estimates $\hat{F}$ could also be used for other purposes, such as large quantile estimation, semiparametric bootstrapping or density estimation.

## 2 A Semiparametric Approach to Tail Estimation

Our goal is to model the tail law $G$ of $X$. Direct approaches to fitting $G$ might specify a parametric model for it, and then estimate the relevant parameters using the sample $\{X\}$. We propose an alternative semiparametric approach. Assuming absolute continuity, we can always write $G$ in terms of the background tail $G_0$: $dG(x) = e^{\lambda(x)} dG_0(x)$ for some exponential tilting function $\lambda(x)$. Our method models $\lambda$ instead of fitting $G$ directly. The advantage of this approach is that we can specify a simple model for $\lambda$ while preserving idiosyncrasies of the carrier measure $G_0$ such as clustering or rounding effects. Given a background dataset $Y$, an estimate of $\lambda(x)$ directly translates to an estimate of $G$ by importance sampling:

$$d\hat{G}(x) = \frac{1}{\sum_{Y_i > t} e^{\lambda(Y_i - t)}} \sum_{Y_i > t} e^{\lambda(Y_i - t)} 1\{x = Y_i - t\} \Rightarrow e^{\lambda(x)} dG_0(x)$$

as the background-sample-size $N$ goes to infinity. Figure 1 illustrates our method in action on one realization of the simulation in Section 4.1. Although the background dataset has a noticeably different distribution from the population of interest, it still serves as a useful reference point for making sense of the sample tail, and our method captures the true tail law $G$ almost exactly.

### 2.1 Extreme Value Theory and a Sufficient Statistic

We model $\lambda(x)$ as a linear family $\lambda_\eta(x) = \eta \cdot T(x) - \psi(\eta)$ indexed by a parameter $\eta \in \mathbb{R}$, with a normalizing constant $\psi(\eta)$. The family of candidate distributions for $G$ is thus an exponential family with carrier $G_0$ and sufficient statistic $T$. Our choice of sufficient statistic controls the behavior and stability of the method, since maximum likelihood estimation in
Figure 1: An example of a tail fit produced by our method. The data is a realization of the simulation described in Section 4.1. Pictured are the true tail distribution $G$ of $X$, the observed tail histograms of $X_i$ and of $Y_i$, and the estimated tail law $\hat{G}$ obtained by reweighting the histogram of the $Y_i$ by $e^{\hat{\lambda}(Y_i - t)}$. Here, our method succeeds in using the small sample to tilt the background tail very close to the true tail law.

Exponential families operate by moment matching on $T$. A good sufficient statistic should capture relevant information about the tail of $G$ while remaining robust to the presence of very large observations. For example, the identity map $T(x) = x$ used by, e.g., Efron and Tibshirani [1996] would not be a good choice for us, because a few very large $X$ values could dominate the sufficient statistic.

Extreme value theory provides a flexible and powerful framework for modeling the tails of distributions; see, e.g., Beirlant et al. [2006] or De Haan and Ferreira [2006] for a review and Beirlant et al. [2012] for recent developments. Our sufficient statistic is motivated by a classical result: if $F$ is a heavy-tailed distribution with a regularly varying tail, then there is a sequence $\sigma_t$ and a constant $\gamma > 0$ such that, as $t \to \infty$,

$$L \left[ \frac{X - t}{\sigma_t} \mid X > t \right] \Rightarrow H_{\gamma,1}(x), \text{ where } H_{\gamma,\sigma}(x) = 1 - \left(1 + \frac{\gamma x}{\sigma} \right)^{-1/\gamma}. \quad (1)$$

The limiting distribution $H_{\gamma}$ is called a generalized Pareto distribution (GPD) with tail index $\gamma$.

Suppose that both our distribution of interest $F$ and the background $F_0$ both have regularly varying tails with the same tail index $\gamma > 0$. If the threshold $t$ is large enough for (1) to apply, we should expect the tail laws $G$ and $G_0$ to be well-approximated by GPDs with the same tail index $\gamma$ but with potentially different scales $\sigma$ and $\sigma_0$. Modeling different but related distributions as having common $\gamma$ and only allowing scale and location parameters to vary is not unusual; see for example the applications in Davison and Smith [1990] or in Coles [2001, Chapter 6].

If $G$ and $G_0$ were really GPDs, then the tilting function $\lambda(x)$ would be

$$\lambda(x) = \log(\sigma_0) - \log(\sigma) + \frac{1 + \gamma}{\gamma} \log(1 + \gamma x/\sigma) - \frac{1 + \gamma}{\gamma} \log(1 + \gamma x/\sigma_0).$$
If the two scales $\sigma$ and $\sigma_0$ are relatively close to each other, we can approximate $\lambda(x)$ as

$$\lambda(x) = \eta T(x) - \psi(\eta) + O\left((\sigma - \sigma_0)^2\right) \quad \text{where} \quad T(x) = \frac{x}{\sigma_0/\gamma + x},$$

while $\psi$ and $\eta$ only depend on $\sigma$, $\sigma_0$ and $\gamma$. This bound holds uniformly in $x$ because $T(x)$ is bounded. Thus, under extreme value theoretic conditions, a linear tilting function with a sufficient statistic of the form $T(x) = x/(\kappa + x)$ should closely replicate the true relative density.

Given our discussion so far, it might seem simpler to just fit the tail of $F$ directly as a GPD instead; this approach has in fact been advocated by, e.g., McNeil [1997]. Direct GPD modeling, however, is very vulnerable to model misspecification. A parametric GPD fit would, for example, ignore any discretization or grouping effects from $\hat{G}$, possibly giving a misleading picture of $G$. By contrast, we model $G$ as a perturbation of $G_0$, with (1) only motivating the direction of the perturbation. Thus, $\tilde{G}$ will reflect any local idiosyncrasies of the background $G_0$.

### 2.2 Our Method in Practice

We have proposed fitting the tail law $G = L_F(X - t \mid X > t)$ as an exponential tilt of the background tail $G_0$, with sufficient statistic $T(x) = x/(\kappa + x)$. Carrying out our proposal requires estimating the tilt parameter $\eta$, as well as choosing a bandwidth $\kappa$ and a threshold $t$.

Concerning $\eta$, if we had access to the full background distribution $G_0$, then the maximum likelihood estimate for $\eta$ would solve the moment-matching condition

$$\frac{1}{|X_i > t|} \sum_{X_i > t} T(X_i - t) = \int_t^\infty T(y - t) e^{\hat{\eta} T(y - t)} dG_0(y - t) \int_t^\infty e^{\hat{\eta} T(y - t)} dG_0(y - t).$$

In applications, $G_0$ is not exactly known. But, as shown by Owen [2007], we can obtain accurate estimates of $\hat{\eta}$ by logistic regression when the number of background observations $N$ is large. To do this, we first join the sample of interest and the background into a single dataset, assigning the former observations a label 1 and the latter ones a 0, and then perform a logistic regression on this dataset with an intercept $T(x - t)$ as the predictor. Then, as $N \to \infty$, Owen [2007] showed that the slope parameter of the logistic regression converges to the solution to (3). In our analysis, we assume that $N$ is large enough for the error in (3) to be negligible. This assumption is reasonable in our motivating internet applications, as we typically have access to an extremely large store of background data.

Second, our discussion from Section 2.1 suggests that setting $\kappa = \sigma_0/\gamma$ should be a good choice. In our experiments, we found that the simple approach of fitting $\gamma$ using the Hill estimator [Hill 1975] on the background dataset and $\sigma_0$ by maximum likelihood worked well. We obtained very similar results by taking $\kappa = t$, which is motivated by the asymptotic limit $\sigma_0/\gamma \sim t$.

Finally, we must choose a good threshold $t$. The most direct method, when possible, is to determine empirically what threshold works best in previously observed instances of the same problem. For example, the internet company may perform hundreds of experiments each day, and over time may learn which values of $t$ work well for different kinds of problems. In the absence of historical data, an alternative method is required. One option is to use off-the-shelf threshold selection rules such as the method of Guillou and Hall [2001] originally
intended to set the threshold for the Hill estimator. The Guillou-Hall procedure returns a number of observations $\hat{k}_{GH}$ to be used for tail index estimation; this value can be translated into a threshold $t_{GH}$ given by the $\hat{k}_{GH}$-th largest observation. In the next section, we show that if our goal is to estimate the mean of the $X$ then $t_{GH}$ grows to infinity at the correct asymptotic rate. However, in our experiments, the Guillou-Hall rule often picks larger-than-optimal thresholds. The weakness of the Guillou-Hall rule is that when $F$ and $F_0$ are very close to each other, we could use a fairly small threshold $t$ without suffering unduly high bias; however, $t_{GH}$ only looks at the $X$ sample, and so has no way of detecting this. Developing an adaptive threshold selection procedure that efficiently uses both samples $\{X\}$ and $\{Y\}$ presents an interesting avenue for further research.

\section{Asymptotic Theory for Mean Estimation}

Combining our semiparametric estimate $\hat{G}$ with the empirical law below $t$ yields the estimate

$$\hat{F}(x) = \frac{1}{n} \sum_{X_i \leq t} 1\{X_i \leq x\} + \frac{1}{n} \sum Y_i > t \frac{1}{e^{\lambda(x)}} \sum Y_i > t e^{\lambda(x)} 1\{Y_i \leq x\},$$

where $\hat{\lambda}(x) = \hat{\eta}T(x)$. Once we have $\hat{F}$, we can obtain plug-in estimates for, e.g., moments or quantiles of $F$. In this section, we provide results about the asymptotic behavior of the mean estimate $\hat{\mu}$ induced by $\hat{F}$. When estimating $\mu$, we always assume that $F$ and $F_0$ both have a shared tail index $0 < \gamma < 1$, which implies that they have finite means. Because our sufficient statistic $T(x) = x/(\kappa + x)$ is bounded, the limit of $\hat{F}$ as the background size $N$ goes to infinity also has a tail index $\gamma < 1$, and so $\hat{\mu}$ is well-defined in the large-$N$ limit. To simplify our analysis, we focus on this limit and assume that $N$ is large enough that the errors in $\hat{G}_0$ are negligeable. Our proofs are deferred to the appendix.

Throughout this section, we assume that $F$ has second-order regularly varying tails

$$1 - F(x) = Cx^{-1/\gamma} (1 + Dx^{-\beta} + o(x^{-\beta})) \text{ for some } \gamma \text{ and } \beta > 0, \tag{4}$$

and that the background distribution $F_0$ also satisfies this condition with the same $\gamma$ but possibly different values of $C_0$, $D_0$, and $\beta_0$. This second-order condition is discussed at length by De Haan and Ferreira [2006], the form given in (4) was introduced by Hall [1982]. All our results are in terms of asymptotic moments, i.e., the moments of the limiting Gaussian random variable. For Winsorization, we use a fixed threshold as discussed in Section 4.

\textbf{Theorem 1.} Suppose that the regularity condition (4) holds, that the background sample size $N$ grows faster than some function $A(n)$, and that $\gamma < 1$. Then our estimator is asymptotically normal for any threshold sequence satisfying $t(n) = o(n^{\gamma})$ and has asymptotic variance

$$n \text{var}_F[\hat{\mu}] = \text{pr}_F[X \leq t] \text{ var}[X | X \leq t] + \text{pr}_F[X > t] \frac{\text{cov}^2[T, X | X > t]}{\text{var}[T | X > t]} \tag{5}$$

$$+ \text{pr}_F[X \leq t] \text{pr}_F[X > t] \left( E_F[X | X > t] - E_F[X | X \leq t] \right)^2.$$

We can turn this result into a practical variance estimate by plugging in $\hat{F}$ for $F$ in (5). The delta method formula can give us more intuition about why our semiparametric estimate is more stable than the sample mean. Whenever the sample mean has finite variance,

$$n \left( \text{var}[\bar{X}] - \hat{\text{var}}[\mu]\right) = \text{pr}_F[X > t] \left( 1 - \text{corr}_F^2[T, X | X > t] \right) \text{var}_F[X | X > t].$$
In other words, our method works well if $T$ captures information relevant to estimating $\mu$ without being too correlated to $X$ itself.

Second, when the threshold sequence is properly specified, we show that our method achieves a better rate of convergence than Winsorization in the range $0.5 < \gamma < 1$ where $X$ has a finite mean but infinite variance.

**Theorem 2.** Suppose that $F$ and $F_0$ satisfy the second-order condition \(^4\) with the same $0.5 < \gamma < 1$, and that the background sample size $N$ grows faster than some function $A(n)$. Then,

$$E\left[(\hat{\mu}_S^* - \mu)^2\right] = O\left(n^{2\gamma - 2 + 4\gamma \min\{\beta, \beta_0\}}\right), \text{ while } E\left[(\hat{\mu}_W^* - \mu)^2\right] = O\left(n^{2\gamma - 2}\right). \quad (6)$$

Here, $S$ stands for our semiparametric method, $W$ stands for Winsorization, and the notation $\hat{\mu}^*$ denotes that both estimators are computed using optimal thresholds.

Finally, we show below that we can estimate the optimal threshold $t^*$ using the method of Guillou and Hall \cite{GuillouHall2001}, as discussed in Section 2.2.

**Corollary 3.** Under the conditions of Theorem 2 and assuming moreover that $\beta = \beta_0$, the rate from (6) remains valid if we replace $t^*$ with $\hat{t}^{GH}$, namely the threshold obtained by applying the Guillou-Hall method to the small sample $\{X\}$.

### 4 Examples and Experiments

In this section, we apply our method both to simulated data and to real data provided by an internet company. We focus on the problem of mean estimation, as this was the original motivation for our research. In our experiments our method comfortably outperforms the baselines, none of which can take advantage of the background sample. The main take-home point appears to be that there is useful information in the background sample that can considerably improve estimates of $\mu$ if properly exploited, and that our semiparametric estimator can achieve this goal given a good choice of $t$.

Our baselines are as follows: **Winsorization:** We consider a version of Winsorization that caps observations at a given threshold $t$ rather than at a predetermined quantile $\hat{\mu}_W = \frac{1}{n} \sum_{i=1}^{n} \min\{t, X_i\}$, as the fixed-$t$ version of Winsorization is more directly comparable to our method with fixed $t$. When we need to choose $t$ adaptively, we set $t$ to the second-largest observation as recommended by Rivest \cite{Rivest1994}. **Parametric GPD modeling:** Instead of using a semiparametric approach, we could have tried to fit a parametric GPD model directly to the tail of the $\{X\}$ by maximum likelihood \cite[e.g.,][]{McNeil1997, Johansson2003}; a related idea was also studied by Peng \cite{Peng2001}. We report results for the method of Johansson at its oracle threshold $t$, as well as for $t$ selected by the method of Guillou and Hall \cite{GuillouHall2001}. Results for the other parametric methods were similar.

#### 4.1 A Simulation Example

We begin by testing our method on data simulated from the log-gamma family: we drew $n = 1000$ and $N = 10^6$ values from the model $\log X_i \sim \text{Gamma}(k = 4, s = 0.45)$ and $\log Y_i \sim \text{Gamma}(k = 3, s = 0.45)$, where $k$ and $s$ are the shape and scale parameters. Log-gamma distributions have regularly varying tails with $\gamma = s$, but do not satisfy the
Table 1: Bias, Variance, and MSE for each method on log-Gamma simulation.

| Method                              | Variance (s.e.) | Bias$^2$ (s.e.) | MSE (s.e.) |
|-------------------------------------|-----------------|-----------------|------------|
| Semiparametric (Oracle $t$)         | 0.29 (0.00)     | 0.17 (0.00)     | 0.46 (0.01) |
| Semiparametric (Guillou-Hall)       | 0.32 (0.00)     | 0.27 (0.01)     | 0.59 (0.01) |
| Winsorized (Oracle $t$)             | 0.50 (0.01)     | 0.15 (0.01)     | 0.65 (0.01) |
| Winsorized ($k = 1$)                | 0.77 (0.02)     | 0.14 (0.01)     | 0.91 (0.02) |
| GPD Tail (Oracle $t$)               | 0.65 (0.01)     | 0.10 (0.00)     | 0.74 (0.01) |
| GPD Tail (Guillou-Hall)             | 3.44 (20.41)    | 0.16 (0.03)     | 3.60 (20.38) |

second-order condition $[4]$ for any $\beta > 0$; thus, this experiment may be seen as a reach case for our method. The different shape parameters $k$ cause the two distributions to have very different expectations: $\mathbb{E}Y = 6.0$ while $\mathbb{E}X = 10.9$.

Table 1 shows results for our method and baselines, both at optimal and adaptive threshold choices. Even with an adaptive threshold choice, our method beats the oracle-$t$ baselines. The MSE for all the methods is much better than the variance of the (unbiased) sample mean, which is 9.9. We show the bias-variance tradeoff for both our method and Winsorization in Figure 2a; we did not include the GPD curve as its behavior was quite erratic. The two methods are about equally biased at their respective optimal thresholds, but the semiparametric method has about 60% of the variance of Winsorization.

4.2 A Real-World Example

Finally, we present results from applying our method to a large multi-arm experiment conducted by Facebook. Working with this kind of dataset was the original motivation for our research. The goal of the experiment was to estimate the mean advertising revenue for two different website layouts. Because the estimand is revenue, we cannot ignore tail observations and obtaining an accurate estimate of the mean is crucial.

In practice, we would apply our method to cases where the control sample is much larger than the treatment sample and can act as a proper background. However, such a setup is not good for testing our method: if the treatment population is small, then we have no way of making sure our method provided the right answer. Thus, here, we started with two large experimental groups and then down-sampled one of them get get a small treatment sample on which to apply our method.

More specifically, the Facebook dataset has two large samples with about 5,000,000 observations each. In order to protect potentially sensitive information, we discarded users whose revenues were zero, then normalized the data so that the population of interest had mean 1. For the purposes of our experiment, these two large samples comprise the treatment and control populations, from which we sample smaller data sets with replacement. This approach allows us to estimate bias as well as variance for each procedure. We apply our method to $n = 200,000$ samples drawn with replacement from the first population (our population of interest $F$), and $N = 3,000,000$ from the second (the background population $F_0$). We then evaluate each method by comparing its mean estimate with the sample mean of the 5,000,000 original data points.

Figure 2b and Table 2 show results using our method as well as the baselines. The results are averaged across 10,000 trials. For a wide range of thresholds $t$, our method outperforms Winsorization and GPD modeling at their own optimal thresholds. These results suggest
Figure 2: Left panel (a): Performance of our method and Winsorization on the log-gamma simulation. The plotted circles are Monte Carlo estimates for semiparametric variance, squared bias, and MSE at the optimal threshold of our method; we obtained the rest of the variance curve using (5) and the rest of the bias curve by setting $\kappa = t$ and computing bias on larger $X$-samples. The right panel (b) shows MSE for three mean estimates on the Facebook dataset. The histogram of the $\{X\}$ is shown in gray.

that, by using our method, Facebook could have made good use of the available background information to considerably improve their mean revenue estimates.

Acknowledgments

We are grateful to Bradley Efron, Trevor Hastie, Robert Tibshirani and Guenther Walther for many helpful conversations, and to Facebook for allowing us to report results on their dataset. W. F. and S. W. are supported by NSF VIGRE grant DMS-0502385 and by a B. C. and E. J. Eaves Stanford Graduate Fellowship respectively.

References

Jan Beirlant, Yuri Goegebeur, Johan Segers, and Jozef Teugels. *Statistics of extremes: theory and applications*. John Wiley & Sons, 2006.

Jan Beirlant, Frederico Caeiro, and M Gomes. An overview and open research topics in statistics of univariate extremes. *REVSTAT-Statistical Journal*, 10(1):1–31, 2012.

S. Coles. *An introduction to statistical modeling of extreme values*. Springer Verlag, 2001.

A.C. Davison and R.L. Smith. Models for exceedances over high thresholds. *Journal of the Royal Statistical Society. Series B (Methodological)*, 52(3):393–442, 1990.

Miguel de Carvalho and Anthony C Davison. Spectral density ratio models for multivariate extremes. *Journal of the American Statistical Association*, 2014.

L. De Haan and A. Ferreira. *Extreme value theory: An introduction*. Springer, 2006.
Table 2: Variance, squared bias, and mean squared error for each method on Facebook data set. All numbers multiplied by 10^6 for display purposes.

| Method                                | Variance (s.e.) | Bias^2 (s.e.) | MSE (s.e.) |
|---------------------------------------|-----------------|---------------|------------|
| Semiparametric (Oracle t)             | 12.6 (0.2)      | 2.5 (0.1)     | 15.1 (0.2) |
| Semiparametric (t = 0.75 quantile)    | 27.8 (0.4)      | 1.4 (0.1)     | 29.2 (0.4) |
| Semiparametric (t = 0.9 quantile)     | 40.4 (0.6)      | 0.1 (0.0)     | 40.5 (0.6) |
| Semiparametric (Guillou-Hall)         | 66.9 (0.9)      | 2.5 (0.3)     | 69.3 (1.0) |
| Winsorized (Oracle t)                 | 72.1 (1.0)      | 1.9 (0.2)     | 74.0 (1.0) |
| Winsorized (k = 1)                    | 78.8 (1.2)      | 1.3 (0.2)     | 80.1 (1.2) |
| Pareto (Oracle t)                     | 69.8 (1.0)      | 2.0 (0.2)     | 71.9 (1.0) |
| Pareto (Guillou-Hall)                 | 642.4 (337.8)   | 0.1 (0.2)     | 642.5 (337.9) |

B. Efron and R. Tibshirani. Using specially designed exponential families for density estimation. *The Annals of Statistics*, 24(6):2431–2461, 1996.

Konstantinos Fokianos. Merging information for semiparametric density estimation. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 66(4):941–958, 2004.

Konstantinos Fokianos, Benjamin Kedem, Jing Qin, and David A Short. A semiparametric approach to the one-way layout. *Technometrics*, 43(1), 2001.

A. Guillou and P. Hall. A diagnostic for selecting the threshold in extreme value analysis. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 63(2):293–305, 2001.

P. Hall. On some simple estimates of an exponent of regular variation. *Journal of the Royal Statistical Society. Series B (Methodological)*, 44(1):37–42, 1982.

B. Hill. A simple general approach to inference about the tail of a distribution. *The Annals of Statistics*, 3(5):1163–1174, 1975.

Alan Huang and Paul J Rathouz. Proportional likelihood ratio models for mean regression. *Biometrika*, 99(1):223–229, 2012.

Joachim Johansson. Estimating the mean of heavy-tailed distributions. *Extremes*, 6(2):91–109, 2003.

Alexander J McNeil. Estimating the tails of loss severity distributions using extreme value theory. *Astin Bulletin*, 27(1):117–137, 1997.

A. Owen. Infinitely imbalanced logistic regression. *J. Mach. Learn. Res.*, 8:761–773, 2007.

Art B Owen. *Empirical likelihood*. CRC press, 2001.

Liang Peng. Estimating the mean of a heavy tailed distribution. *Statistics & Probability Letters*, 52(3):255–264, 2001.

L.P. Rivest. Statistical properties of Winsorized means for skewed distributions. *Biometrika*, 81(2):373–383, 1994.

Z Tan. A note on profile likelihood for exponential tilt mixture models. *Biometrika*, 96(1):229–236, 2009.
A Asymptotic Variance

Throughout the appendix, we use the notation

\[ \mu_1 = E[X|X \leq t], \mu_2 = E[X|X > t], \text{ and } p_2 = \Pr[X > t]. \]

Here, we show that our estimator \( \hat{\mu} \) is asymptotically normal, with asymptotic variance

\[
\tilde{\text{var}}[\hat{\mu}] \sim \left( \frac{\partial \hat{\mu}}{\partial T} \right)^2 \text{var}[T],
\]

where we use \( \tilde{\text{var}} \) to denote the asymptotic variance of \( \hat{\mu} \). The asymptotic variance of an asymptotically normal random variable is the variance of the limiting normal distribution.

We begin by showing asymptotic normality in the case where the threshold \( t \) remains fixed; this lets us motivate our formula without having to worry about technical conditions. We then extend our analysis to sequences of thresholds \( t(n) = o(n^\gamma) \).

A.1 Fixed-t analysis

Our method relies on an estimate for \( \hat{\eta} \) obtained using a logistic regression where the size \( N \) of the background sample is much larger than the size \( n \) of our sample of interest. As explained by Owen [2007], this lets us take a considerable shortcut in computing the distribution of our estimator \( \hat{\eta} \). Specifically, Owen’s result implies that, in the limit where \( N \) diverges to infinity but \( n \) remains fixed, the estimator \( \hat{\eta} \) obtained by logistic regression satisfies

\[
\frac{1}{\#\{X_i > t\}} \sum_{i: X_i > t} T(X_i - t) = \frac{\int_0^\infty T(y-t) e^{\hat{\eta} T(y-t) - \psi(\hat{\eta})} dG_0(y)}{\int_0^\infty e^{\hat{\eta} T(y-t) - \psi(\hat{\eta})} dG_0(y)},
\]

where \( t \) is the threshold at which we start using our semiparametric method and \( G_0 \) is the distribution of \( Y \) given that \( Y > t \). Thus, when \( N \) is large, we can in practice ignore the randomness of \( Y \), and use semiparametric exponential family theory to estimate the distribution of \( \hat{\eta} \). In the rest of this section, we assume that \( N \) is large enough that this approximation in fact holds.

We begin by recalling some standard results. Suppose that we have any exponential family

\[ g_\eta(x) = g_0(x) \exp[\eta T(x) - \psi(\eta)]. \]

Then, the maximum likelihood estimate \( \hat{\mu} \) is a function of the sufficient statistic \( T \), and by the delta-method \( \hat{\mu} \) is asymptotically normal with

\[
\tilde{\text{var}}[\hat{\mu}] \sim \left( \frac{\partial \hat{\mu}}{\partial T} \right)^2 \text{var}[T],
\]

In an exponential family, we can show that \( T = \frac{\partial \bar{\psi}}{\partial \eta} |_{\eta = \hat{\eta}} \), and so

\[
\frac{\partial T}{\partial \eta} = \frac{\partial^2 \psi}{\partial \eta^2} |_{\eta = \hat{\eta}} = \text{var}_{\eta}[T],
\]

\[
(1-p_2) \text{var}[X|X \leq t] + p_2 (1-p_2) (\mu_2 - \mu_1)^2
\]

\[
+ p_2 \frac{\text{cov}^2[T, X|X > t]}{\text{var}_{\eta}[T|X > t]}\]

\[
\rightarrow 1,
\]

where \( \tilde{\text{var}} \) to denote the asymptotic variance of \( \hat{\mu} \).
leading to \( \frac{\partial \hat{\eta}(T)}{\partial T} = \text{var}_{\hat{\eta}}^{-1}[T] \). Furthermore,

\[
\frac{\partial}{\partial \eta} E[\eta|X]\big|_{\eta=\hat{\eta}} = \int (T(x) - \psi'(\hat{\eta})) \cdot x \, dG(x) = \int (T(x) - E_0[T]) \cdot x \, dG(x) = \text{cov}_{\hat{\eta}}[X,T].
\]

The above discussion provides us with an estimate for the asymptotic variance of the upper mean \( \hat{\mu}_2 \) when the threshold \( t \) is fixed:

\[
\bar{\text{var}}[\hat{\mu}_2] = \frac{1}{p_2 n} \cdot \frac{\text{cov}_{\hat{\eta}}^2[T,X]}{\text{var}_{\hat{\eta}}[T]}. 
\]

We continue by noting that our lower mean estimate \( \hat{\mu}_1 \), our tail mean estimate \( \hat{\mu}_2 \), and the proportion of tail observations \( \hat{p}_2 \) are all asymptotically uncorrelated, and are asymptotically normal by the central limit theorem. Thus,

\[
\bar{\text{var}}[\hat{\mu}] = \bar{\text{var}}[(1 - \hat{p}_2) \hat{\mu}_1 + \hat{p}_2 \hat{\mu}_2] = (1 - p_2)^2 \cdot \bar{\text{var}}[\hat{\mu}_1] + p_2^2 \cdot \bar{\text{var}}[\hat{\mu}_2] + (\mu_2 - \mu_1)^2 \cdot \bar{\text{var}}[\hat{p}_2].
\]

The expression in (7) then follows immediately. Note that the sample size for estimating \( \hat{\mu}_1 \) and \( \hat{\mu}_2 \) are asymptotic to \( n(1 - \hat{p}_2) \) and \( np_2 \) respectively, canceling one factor in each of the first two terms.

### A.2 General Case

In this section we extend the above results and, under a regular variation assumption, establish the validity of our asymptotic variance formula to cases where the threshold sequence \( t(n) \) grows in \( n \). Specifically, suppose that \( X \) and \( Y \) both have regularly varying tails with a common tail index \( 0 < \gamma < 1 \), and that the second-order model (4) holds. Then, \( \hat{\mu} \) is asymptotically normal and our asymptotic variance formula (7) holds for any sequence of thresholds satisfying \( t(n) \to \infty \) and \( t(n) = o(n^\gamma) \). Our condition on the growth of \( t \) is equivalent to requiring \( np_2(t) \to \infty \). In other words, \( t \) must grow slow enough that the number of \( X_i \) exceeding the threshold tends to infinity, enabling our estimate of \( \hat{\mu}_2 \) to converge.

The second-order condition is not actually required, but simplifies the exposition. With a little extra work we can use Potter’s inequalities [see, e.g., De Haan and Ferreira, 2006] to establish our result under only first-order regular variation.

Given the second-order model (4), we can verify that

\[
\text{pr}[X > t] \asymp t^{-\frac{1}{\gamma}}, \quad E[X^2|X < t] \asymp t^{\frac{2\gamma - 1}{\gamma}}, \quad \text{and} \quad E[X - t|X > t] \asymp t.
\]

where \( f(t) \asymp g(t) \) means that \( f(t) \) and \( g(t) \) are of the same asymptotic order, i.e., \( \frac{f(t)}{g(t)} \to c \in (0, \infty) \) as \( t \to \infty \). Writing \( \hat{\mu}_1 = \mu_1 + \varepsilon_1, \hat{\mu}_2 = \mu_2 + \varepsilon_2, \) and \( \hat{p}_2 = p_2 + \varepsilon_3 \) and rearranging terms, we have

\[
\hat{\mu} - \mu = (1 - p_2)\varepsilon_1 + p_2\varepsilon_2 + (\mu_2 - \mu_1)\varepsilon_3 + (\varepsilon_2 - \varepsilon_1)\varepsilon_3.
\]
Equation (8) suggests that the first term has variance on the order of \( \frac{1}{n} \left( \frac{2\gamma - 1}{t} \right) \). Inflating (9) by this factor and compensating for the order of \( p_2 \) and \( \mu_2 - \mu_1 \), we obtain

\[
\sqrt{\frac{n}{t^{2\gamma - 1}}} \cdot (\hat{\mu} - \mu) = (1 - p_2) \cdot \sqrt{\frac{n}{t^{2\gamma - 1}}} \varepsilon_1 + \frac{p_2}{t^{1/\gamma}} \cdot \sqrt{\frac{nt^{-1/\gamma}}{t^2}} \varepsilon_2 + \frac{\mu_2 - \mu_1}{t} \cdot \sqrt{nt^{1/\gamma}} \varepsilon_3
\]

Making the substitutions

\[
 Z_1 = \sqrt{\frac{n}{t^{2\gamma - 1}}} \varepsilon_1, \quad Z_2 = \sqrt{\frac{nt^{-1/\gamma}}{t^2}} \varepsilon_2, \quad \text{and} \quad Z_3 = \sqrt{nt^{1/\gamma}} \varepsilon_3.
\]

(10) simplifies to

\[
\sqrt{\frac{n}{t^{2\gamma - 1}}} \cdot (\hat{\mu} - \mu) = (1 - p_2) \cdot Z_1 + \frac{p_2}{t^{1/\gamma}} \cdot Z_2 + \frac{\mu_2 - \mu_1}{t} \cdot Z_3 + \sqrt{\frac{t^{1/\gamma}}{n}} \left( Z_2 - t^{-1/\gamma} Z_1 \right) \cdot Z_3.
\]

If we can show that the \( Z_i \) converge weakly to independent normal random variables, then we will have established that the fourth term of (11) tends in probability to 0, and hence that the left-hand side is asymptotically normal with variance (7). Here, the coefficients multiplying \( Z_1, ..., Z_3 \) converge to finite limits as \( t \to \infty \), whereas the fourth coefficient converges to 0 because \( t = o(n^{\gamma}) \) by hypothesis.

Write \( Z_1 \) as

\[
 Z_1 = \sqrt{nt^{1/\gamma} t^{-1}} \left( \hat{\mu}_1 - \mu_1 \right) = \sum_{X_i < t} \sqrt{\frac{t^{1/\gamma}}{n}} t^{-1} (X_i - \mu_1).
\]

Since the magnitude of the summand is bounded by \( \sqrt{\frac{t^{1/\gamma}}{n}} \to 0 \), the Lindeberg condition is satisfied and so, by the central limit theorem for triangular arrays and (8), there is a sequence \( s_1 \) such that \( s_1 \approx 1 \) and \( s_1 Z_1 \) is asymptotically standard normal. Similarly, write \( Z_3 \) as

\[
 Z_3 = \sqrt{nt^{1/\gamma}} \left( \hat{\mu}_2 - \mu_2 \right) = \sum_{i=1}^n \sqrt{\frac{t^{1/\gamma}}{n}} \left( 1_{X_i > t} - \mu_2 \right).
\]

Once again the magnitude of the summand is bounded by \( \sqrt{t^{1/\gamma} n} \to 0 \), and so a similar argument as above applies. Finally we turn to asymptotic normality of \( Z_2 \). By our regular variation assumption \( \mathbb{L}\left[ X^t \right] \text{ and } \mathbb{L}\left[ Y^t \right] \) converge to Pareto limits both in law and in moments. Because, \( t = o(n^{\gamma}) \), the number of data points that can be used the fit \( \hat{\mu}_2 \) still grows to infinity, and so the same delta-method argument as in the fixed-\( t \) case goes through.

**B Optimal Rates of Convergence**

We start by deriving the risk of Winsorization; our semiparametric risk estimate will build on many of the same ideas. Here, we use the notation \( a(n) \approx b(n) \) to indicate that \( a(n)/b(n) \) converges to a finite non-zero limit.
B.1 A Warm-Up with Winsorization

Recall that the estimator we call the Winsorized estimate for the mean is
\[ \hat{\mu}_W(t) = \frac{1}{n} \sum_{i=1}^{n} \min\{X_i, t\}, \]
and so we see by (8) that
\[ \text{var}[\hat{\mu}_W(t)] \approx \frac{1}{nt^2} \gamma - \frac{1}{\gamma}, \]
and bias\(^2[\hat{\mu}_W(t)] \approx t^2 \gamma - \frac{1}{\gamma}. \]

To compute the variance estimate, we used the fact that \( E[X] \) is finite while var\([X]\) is infinite, and so
\[ \text{var}[\hat{\mu}_W(t)] \approx E[(\hat{\mu}_W(t))^2] \]
The mean-squared error \( \text{MSE} = \text{var} + \text{bias}^2 \) can then be minimized at a threshold \( t^* \approx n^\gamma \), giving us an optimal error
\[ \text{MSE}_W^* \approx n^{2\gamma - 2}, \]
as claimed in (6).

B.2 Semiparametric Risk

We are now ready to discuss the risk of our semiparametric method under the second-order assumption (4). As in Appendix A, we assume that \( N \) is large enough that the randomness of the \( Y \) doesn’t matter. By our delta-method estimate, for \( t = o(n^\gamma) \), the asymptotic variance of our method is
\[ \text{var}[\hat{\mu}_S(t)] \sim \text{var}[X; X < t] + \frac{\text{pr}[X > t]}{n} \frac{\text{cov}^2[T, X|X > t]}{\text{var}[T|X > t]} + \frac{(E[X|X > t] - E[X|X < t])^2}{n} \text{pr}[X > t](1 - \text{pr}[X > t]) \]
\[ \approx \frac{t^{2\gamma - 1}}{n} + \frac{t^{-1/\gamma} t^2}{1 + \frac{t^2}{n} t^{-1/\gamma}} \]
\[ \approx \frac{1}{n} t^{2\gamma - 1}, \]
where on the second line we used moment estimates from (8). The fact that \( \text{cov}[T, X|X > t] \approx t \) and \( \text{var}[T|X > t] \approx 1 \) can be verified by calculus because the scale parameter \( \sigma_0 \) grows proportionally to \( t \) \cite{Coles2001}.

So far, we have seen that given any shared threshold sequence, the variance of our semiparametric estimator decays at the same rate as that of Winsorization. Now, we will show that under the second order condition, the bias of our method decays faster than that of Winsorization, which will enable us to use smaller thresholds and achieve better risks.

For convenience, let \( \tilde{X} = X - t \) and \( \tilde{Y} = Y - t \), conditional respectively on \( X \) and \( Y \) exceeding the threshold \( t \). We consider the exponential tilts \( dG_\eta = e^{\eta T - \psi(\eta)}dG_0 \) with \( \eta \in \mathbb{R} \), and fit the distribution \( G \) of the \( \tilde{X} \) with \( G^*_\eta \), where \( \eta^* \) is the population MLE for \( \eta \). If we
write $\mu(\eta)$ for the mean of $dG_\eta$, we find that the bias of our semiparametric estimator is given by 

$$\tilde{\text{bias}} \left[ \hat{\mu}_S^{(t)} \right] = \text{pr}[X > t] \cdot \left( \mu(\eta^*) - E \left[ \bar{X} \right] + O \left( \frac{t}{n \text{pr}[X > t]} \right) \right),$$

where $\tilde{\text{bias}}$ is the bias of the center of limiting normal distribution of $\hat{\mu}_S^{(t)}$. Here, the main term is due to model misspecification arising from the fact that $\bar{Y}$ and $\bar{X}$ are only converging to the GPD. The remainder, which will turn out not to affect the decay rate of the asymptotic MSE, is due to higher-order curvature effects (i.e., the second-order term in the delta-method expansion).

We begin by establishing a tail bound. As a direct consequence of the second-order condition (4), we find that 

$$1 - G(t(1 + x)) = x^{-\frac{1}{\kappa}} \left( 1 + D_X t^{-\beta_x} \left( x^{-\beta_x} - 1 \right) + o(t^{-\beta_x}) \right).$$

This implies that 

$$1 - G \left( t \cdot \frac{\kappa}{1 - \tau} \right) = (1 + x)^{-\frac{1}{\kappa}} \left( 1 + D_X t^{-\beta_x} \left( (1 + x)^{-\beta_x} - 1 \right) + o(t^{-\beta_x}) \right),$$

or, in terms of the statistic $T = \frac{x}{\kappa + \tau}$,

$$\text{pr} \bar{X}[T > \tau] = \left( 1 + \frac{\tau}{1 - \tau} \right)^{-\frac{1}{\kappa}} \left( 1 + D_X t^{-\beta_x} \left( 1 + \frac{\tau}{1 - \tau} \right)^{-\beta_x} - 1 \right) + o(t^{-\beta_x})$$

for $\tau \in [0, 1)$. A similar expression holds for $\text{pr} \bar{Y}[T > \tau]$. Notice that the $o(t^{-\beta_x})$ term is bounded in $x$ as $x$ gets large; this means that we can use (12) to establish the convergence of moments. Now, recalling that $\kappa \sim t$, we can use the tail bound (12) to establish many useful relations.

- Because $\bar{X}$ and $\bar{Y}$ share the same tail bound with possibly different constants $D$ and $\beta$, we see that 

$$E \bar{X}[T] - E \bar{Y}[T] \approx t^{-\beta},$$

where $\beta = \min\{\beta_X, \beta_Y\}$.

- Recall that in an exponential family, the MLE $\eta^*$ is defined by the relation 

$$E \bar{X}[T] = E \bar{Y} \left[ e^{\eta^* T - \psi(\eta^*) T} \right].$$

We already know from Appendix [A] that 

$$\frac{dE \bar{Y} \left[ e^{\eta T - \psi(\eta) T} \right]}{d\eta} \bigg|_{\eta = 0} = \text{var} \bar{Y}[T],$$

which by (12) converges to a finite non-zero limit as the threshold $t$ goes to infinity. Thus, because of (13), we see that $\eta^* \approx t^{-\beta}$. 

14
Finally, doing some calculus, we can use (12) to show that
\[
\frac{E[\tilde{X}] - E[\tilde{Y}]}{E[\tilde{X}] - E[\tilde{Y}]} = \frac{(1 + \gamma)(1 + \gamma + \gamma \beta)}{(1 - \gamma)(1 - \gamma - \gamma \beta)} \cdot t + o(t).
\]

We are now ready to bound the bias of our method. By the same arguments as in Appendix A, we find that
\[
\frac{1}{t} \frac{d\mu(\eta)/d\eta}{dE_{\tilde{G}_n}[T]/d\eta}_{\eta=0} = \frac{1}{t} \frac{\text{cov}[\tilde{Y}, T]}{\text{var}[\tilde{Y}]} = \frac{(1 + \gamma)(1 + 2\gamma)}{(1 - \gamma)\gamma} + o(1),
\]
and that the second-order term also has a finite limit. Thus,
\[
\frac{1}{t} \left( \mu(\eta^*) - E[\tilde{X}] \right) = \frac{1}{t} \left( \mu(\eta^*) - E[\tilde{Y}] - E[\tilde{X}] - E[\tilde{Y}] \right) = \left( \frac{(1 + \gamma)(1 + 2\gamma)}{(1 - \gamma)\gamma} + o(1) \right) \cdot (\eta^* + O((\eta^*)^2)) - \left( \frac{(1 + \gamma)(1 + \gamma + \gamma \beta)}{(1 - \gamma)(1 - \gamma - \gamma \beta)} + o(1) \right) \cdot (E[\tilde{X}] - E[\tilde{Y}]) \lesssim t^{-\beta},
\]
and so finally
\[
\tilde{\text{bias}}(\hat{\mu}^{(t)}(\cdot)) \asymp \text{pr}[X > t] \cdot \left( t^{1-\beta} + O \left( \frac{t}{\text{pr}[X > t]} \right) \right) \lesssim t^{\frac{-\gamma - 1}{\gamma n}} + O \left( \frac{1}{t^{1-\gamma n}} t^{\frac{-\gamma - 1}{\gamma n}} \right).
\]

Putting all the pieces together, we get
\[
\tilde{\text{MSE}}(\hat{\mu}^{(t)}(\cdot)) \asymp \frac{1}{n} t^{\frac{2n - 2}{\gamma n}} + t^{\frac{2n - 2\gamma - 2}{\gamma n}} - 2\gamma - 2 + O \left( \frac{1}{t^{1-\gamma n}} t^{\frac{2n - 2\gamma - 2}{\gamma n}} \right),
\]
where the asymptotic mean-squared error \(\tilde{\text{MSE}}\) describes the limiting normal distribution of \(\hat{\mu}^{(t)}(\cdot)\). This is optimized with \(t^*(n) \asymp n^{\frac{\gamma}{1+2\beta}}\), which leads to
\[
\tilde{\text{MSE}}^*(\hat{\mu}^{(t)}(\cdot)) \asymp n^{\frac{2n - 2 - 2\gamma}{\gamma n}} \left( 1 + O \left( n^{\frac{\gamma}{1+2\beta}} \right) \right).
\]
Note that, for any \(\beta > 0\), our optimal threshold sequence in fact satisfies the relation \(t^*(n) = o(n^\gamma)\) which we assumed at the beginning.

**B.3 Proof of Corollary 3**

As shown above, achieving the optimal rate of convergence from Theorem 2 only requires that as \(n\) grows, our threshold \(t_n\) grows as \(t^*(n) \asymp n^{\frac{\gamma}{1+2\beta}}\). Guillou and Hall [2001] frame
their problem as choosing the number of order statistics $\hat{k}^{\text{GH}}$ with which to compute the Hill estimator. A choice of $\hat{k}^{\text{GH}}$ immediately implies a threshold, namely the $\hat{k}^{\text{GH}}$ largest observation $\hat{t}^{\text{GH}} = X_{n-\hat{k}^{\text{GH}} + 1, n}$. Under the assumption \([4]\), they show that $\hat{t}^{\text{GH}} / n^{\frac{2\gamma}{1+2\gamma}} \approx_n 1$. Now, we also know that $\mathbb{P}[X > t] \approx t^{-1/\gamma}$, which implies that

$$\hat{t}^{\text{GH}} \approx_n \left( n^{\frac{2\gamma}{1+2\gamma}} / n \right)^{-\gamma} = n^{\frac{2\gamma}{1+2\gamma}}.$$