Dissertation

Modeling the Upper Tail of the Distribution of Facial Recognition Non-match Scores

Submitted by

Brett D. Hunter

Department of Statistics

In partial fulfillment of the requirements

For the Degree of Doctor of Philosophy

Colorado State University

Fort Collins, Colorado

Spring 2016

Doctoral Committee:

Advisor: Dan Cooley
Co-Advisor: Geof Givens

Piotr Kokoszka
Bailey Fosdick
Henry Adams

Copyright by Brett D. Hunter 2016

All Rights Reserved

Abstract

Modeling the Upper Tail of the Distribution of Facial Recognition Non-match Scores

In facial recognition applications, the upper tail of the distribution of non-match scores is of interest because existing algorithms classify a pair of images as a match if their score exceeds some high quantile of the non-match distribution. I construct a general model for the distribution above the $(1-\tau)$th quantile borrowing ideas from extreme value theory. The resulting distribution can be viewed as a reparameterized generalized Pareto distribution (GPD), but it differs from the traditional GPD in that τ is treated as fixed. Inference for both the $(1-\tau)$th quantile u_τ and the GPD scale and shape parameters is performed via M-estimation, where my objective function is a combination of the quantile regression loss function and reparameterized GPD densities.

By parameterizing u_τ and the GPD parameters in terms of available covariates, understanding of these covariates' influence on the tail of the distribution of non-match scores is attained. A simulation study shows that my method is able to estimate both the set of parameters describing the covariates' influence and high quantiles of the non-match distribution. The simulation study also shows that my model is competitive with quantile regression in estimating high quantiles and that it outperforms quantile regression for extremely high quantiles. I apply my method to a data set of non-match scores and find that covariates such as gender, use of glasses, and age difference have a strong influence on the tail of the non-match distribution.

Acknowledgements

I thank my advisors for their patience and guidance. Geof Givens introduced me to facial recognition, and his early mentorship was instrumental in the development of my research goals. Dan Cooley graciously agreed to serve as a co-advisor once it became clear that my research would be using extreme value theory, and he ably took over as my primary advisor following Geof's retirement. I am indebted to both.

The facial recognition data used in Chapter 7 was provided by Ross Beveridge, and he gave valuable feedback both during the formation of my project goals and in applying my method to the data set. I thank him for welcoming me as an unofficial member of the Computer Vision Group at Colorado State University and for serving as a committee member prior to my defense.

I thank my committee members for their attention and contributions: Piotr Kokoszka for providing counsel during a particularly difficult proof, Bailey Fosdick for her feedback on early chapters, and Henry Adams for valiantly agreeing to serve as a replacement committee member a week before my defense. I recognize that their time is valuable, so I appreciate the time they were willing to afford me.

This research utilized the CSU ISTeC Cray HPC System supported by NSF Grant CNS-0923386. I thank the system administrator, Richard Casey, for his attention following the 2015 system upgrade, during what I imagine was a stressful period for him.

I suffer from depression, the severity of which can be overwhelming, as it was at times during this research. I thank the people I've met through Counseling Services at CSU for their aid, receptiveness, and perspective.

Above all, I'd like to thank my parents, Arthur and Barbara, for their love and support. They believed in me when I was at my lowest and encouraged me when I was ready to quit. This dissertation would not exist without them.

Dedication

For my parents.

Table of Contents

List of Tables

List of Figures

CHAPTER 1

Facial Recognition Motivation

Facial recognition is the identification or verification of a person from a still image or video using a stored database of faces, and is used in law enforcement and surveillance, information security, and entertainment (Zhao et al., 2003). Facial recognition problems can be separated into identification or verification problems. In identification problems, an unknown face is submitted and the system reports back the determined identity. In verification problems the system must confirm or reject the claimed identity of the individual.

In both identification and recognition problems, facial recognition compares a query, an image of a person being examined, to a target, an image of a known individual of interest. The comparison of the two images is issued a score, with higher scores indicating a better match between the query and target. If the score exceeds a certain value, which we will term the "classification threshold", then the target/query pair is labeled as a match.

To make a meaningful determination of a classification threshold, one needs to understand the distribution of scores for target/query pairs known to be non-matches. Researchers have extensive databases of images of known individuals from which they can create target/query pairs of distinct individuals, and these can be subsequently scored providing draws from the distribution of possible non-match scores. Of particular interest is the upper tail of this distribution, as these are scores which indicate that the target/query pairs exhibit strong similarities. The bulk of this distribution is of little interest.

Currently, the two most commonly used classification thresholds are the empirical .99 or .999 quantiles of the non-match distribution. That is, the threshold is set so that the false match rate is 1-in-100 or 1-in-1000.

Figure 1.1 gives a heuristic representation of the decision making process. One can imagine the facial recognition procedure as consisting of two distributions: the non-match distribution, consisting of scores provided by pairs of images containing different subjects, and the much smaller match distribution, consisting only of scores from pairs of images containing the same subject. The mean of the match distribution should be higher than that of the non-match distribution. The vertical line represents the classification threshold, set so that the false match rate is .001. This classification threshold is chosen without taking available covariate information into account.

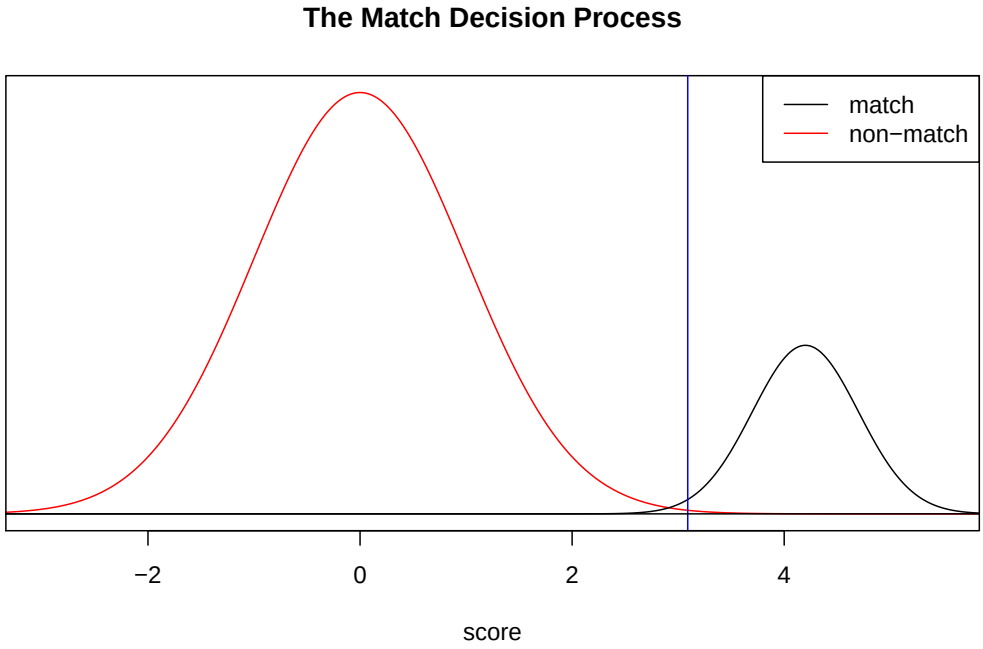

FIGURE 1.1. The vertical line is the classification threshold, set at the .999 quantile of the non-match distribution.

1.1. The Role of Covariates

Understanding factors that influence algorithm performance has been an important focus of some facial recognition work. To that end, both subject covariates and image covariates have been collected, and their relationship with the algorithm have been studied. The earliest

exploration of covariate effects usually used dataset partitioning (Gross et al., 2001), while more recent studies have used generalized linear mixed models (Beveridge et al., 2009). These studies are often concerned with identifying 'quality measures,' first introduced by Grother and Tabassi (2007), but adapted by (Beveridge et al., 2008) to represent covariates that are 'predictive of (algorithm) performance.' This is complicated by the fact that covariate effect is often dependent upon the algorithm used, although determining which covariate effects are consistent across algorithms has been studied (Givens et al., 2004; Lui et al., 2009; Givens et al., 2013).

Beveridge et al. (2008) break the covariates of interest into two categories: subject and image covariates. Subject covariates are specific to the person in the image, such as age, gender, or race, whereas image covariates are specific to the image quality, such as focus or size of the face. While usually only covariates that can be reliably and consistently measured are recorded, the question of whether the covariate can be changed when the image is taken is also a concern. Beveridge et al. (2008) calls these actionable covariates, and would include factors such as the size of the face in an image, and whether a person is wearing glasses. These are factors that can be controlled for in practical settings. Non-actionable covariates include gender, race, and age.

The aforementioned covariate studies have all been concerned with the verification rate. That is, they are concerned with the factors that lead to bigger or smaller similarity scores for matched pairs of images. The exploration of covariate effects on the non-match pairs has not been a focus.

It is important to stress that current algorithms do not make use of available covariate information which is included in a target/query pair. Although the identities of the people in

the target/query pair are unknown, covariates associated with the images are known. Thus, the overall non-match distribution is actually a mixture of a number of distributions given covariates. The classification threshold is set using the overall non-match distribution in most cases. This suggests that a situation such as that illustrated in Figure 1.2 is possible, where the overall match and non-match distributions are broken down into those with target images containing a male subject and those with target images containing a female subject. A situation similar to the one in Figure 1.2 would suggest that certain covariate values are more likely to result in false matches than others.

While there has been some exploration on the effect that the covariates have on choice of threshold (O'Toole et al., 2012), a flexible model for the non-match distribution that uses knowledge of the covariates has not yet been proposed. My primary aim is to develop such a model, thereby understanding how covariates influence the tail of the resulting non-match distribution.

1.2. Goals

The goal of this project is to model the upper tail of the non-match distribution given covariates. I will use a model which borrows ideas from extreme value theory, whose primary objective is to model the upper tail of a distribution. Unlike most extremes-based threshold exceedance approaches, I model the tail of a distribution above the $(1-\tau)$th quantile corresponding to a *fixed* proportion of observations exceeding that quantile, where τ is chosen as a level of interest by facial recognition researchers. Of particular interest is how covariates influence both u_τ and the tail of the distribution above this threshold. I adopt a model for the distribution above the threshold which is a reparameterization of the GPD, and because this distribution is parametric, it allows one to interpret how covariates affect the tail.

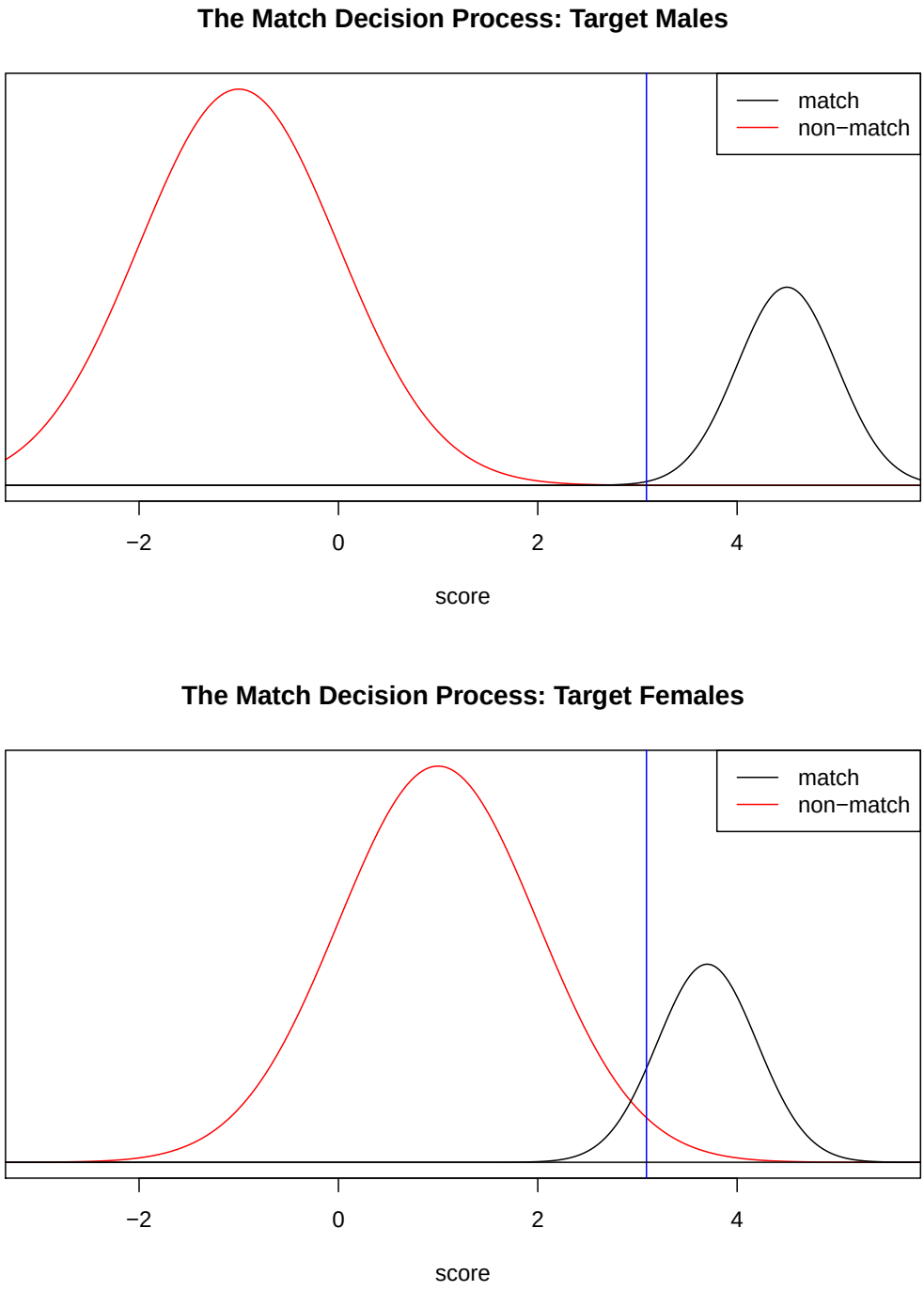

FIGURE 1.2. The vertical line is the classification threshold, set at the .999 quantile of the *overall* non-match distribution.

My approach will use all of the data to model u_τ, but will only use exceedances over this threshold for inference on the tail model. Inference for my model is more complicated than traditional extremes studies because the threshold u_τ is estimated rather than being fixed at

the outset, rendering standard likelihood-based approaches infeasible. In Chapter 2, I review key concepts from extreme value theory. I present my model in Chapter 3, with Chapters 4 and 5 used to discuss my inference method via M-estimation. Chapter 6 and 7 will illustrate the utility of my model with an extensive simulation study and application to a dataset of non-match facial recognition scores, respectively.

CHAPTER 2

A Brief Introduction to Extreme Value Theory

Extreme value theory is a branch of statistics that focuses on the unusually large (or small) levels of a data set. The goal of an extreme value analysis is often to extrapolate beyond the range of the data. For example, one might have 50 years of data, but need to make an estimate of the magnitude of an event that occurs once every 100 years on average. Extreme value theory is commonly used in fields such as hydrology, atmospheric science, finance, and insurance, where such rare events can have tremendous impact. Models in extreme value theory are derived using asymptotic arguments, of which there are two major approaches: the block maxima approach or the threshold exceedance approach.

2.1. Block Maxima Approaches

Classical extreme value theory focuses on the behavior of

$$D_n = \max\{Y_1, \ldots, Y_n\},$$

where $Y_1, \ldots, Y_n$ are independent random variables with common distribution function F. The distribution of D_n is F^n, but this is not helpful in practice if the distribution function F is unknown. While it might be possible to estimate F from observed data, small discrepancies in the estimate can lead to substantial discrepancies for F^n. Classical statistical methods for extremes treat F as unknown and estimate models of F^n using only block maxima data. That is, a block length is selected (like a year), the maximum of each block is extracted, and a model is fit only to this subset of block maxima.

2.1.1. The Extremal Limits. A result similar to the central limit theorem, which states that a normalized sample mean converges to a Gaussian distribution, is used to describe the limiting behavior of block maxima. The renormalization

$$\frac{D_n - b_n}{a_n} \tag{1}$$

as $n \to \infty$ is considered, where a_n and b_n are sequences of constants such that $a_n > 0$. Beirlant et al. (2004) describe this problem as two-fold: all possible limiting distributions of (1) must be determined and the distributions F for which there exist sequences a_n and b_n that lead to a limiting distribution must be categorized.

The limiting distributions problem has been solved by Fisher and Tippett (1928) and Gnedenko (1943). The three-types theorem (alternately called the extremal types theorem in Coles (2001)) states that if there exist sequences of constants $\{a_n > 0\}$ and $\{b_n\}$ such that

$$P\left(\frac{D_n - b_n}{a_n} \leq z\right) \xrightarrow{d} G_\xi(z) \text{ as } n \to \infty, \tag{2}$$

where G_ξ is a non-degenerate distribution function, then G_ξ belongs to one of three different families: the Gumbel, the Fréchet, or the Weibull. These three types of extreme value distributions are the only possible limits for distributions of the form given in (1). Thus, just as the central limit theorem states that if Y has a finite 2nd moment, then its suitably renormalized sample mean will converge to a Gaussian, the three-types theorem states that a renormalized maximum must converge to a Gumbel, Fréchet, or Weibull.

The Gumbel family, defined for all real numbers, has cumulative distribution function

$$G_\xi(z) = \exp\left(-\exp\left[-\frac{z-b}{a}\right]\right), \tag{3}$$

where a and b are scale and shape parameters, respectively, such that $a > 0$. The Fréchet family has cumulative distribution function

$$G_\xi(z) = \begin{cases} 0, & z \leq b \\ \exp\left(-\left(\frac{z-b}{a}\right)^{-\alpha}\right), & z > b \end{cases}. \tag{4}$$

In addition to scale and location parameters, the Fréchet has shape parameter α, where α must be positive. The cdf for the Weibull family is

$$G_\xi(z) = \begin{cases} \exp\left(-\left(-\frac{z-b}{a}\right)^{\alpha}\right), & z < b \\ 1, & z \geq b \end{cases}, \tag{5}$$

where α is once again a shape parameter restricted to positive values. These three families can be combined into a single family so that

$$G_\xi(z) = \exp\left(-\left(1 + \xi\frac{z-b}{a}\right)^{-\frac{1}{\xi}}\right), \tag{6}$$

with shape parameter ξ. The Gumbel family is interpreted as the limit of (6) as $\xi \to 0$.

The second part of the problem described by Beirlant et al. (2004) is the domain of attraction problem. This is concerned with determining which of the three extremal types will be the limiting distribution if Y follows a specific distribution F. For example, the uniform distribution is in the Weibull domain of attraction. That is, if $Y_1, \ldots, Y_n$ follow a uniform distribution, then $\exists\ \{a_n > 0\}$ and $\{b_n\}$ such that the limiting distribution of (1) is Weibull. Other distributions with bounded upper tails fall in the Weibull domain of attraction, so that a ξ in (6) is negative. Light tailed distributions, such as the Gaussian and

gamma distributions, are in the Gumbel domain of attraction with $\xi = 0$, whereas heavy tailed distributions such as the t, F, and Pareto distributions are in the Fréchet domain of attraction with $\xi > 0$.

2.1.2. Generalized Extreme Value Distribution. The three types of limits discussed in Section 2.1.1 correspond to the different forms of tail behavior for F. The earliest applications of the three-types theorem saw practitioners adopting one of the three families as the limiting distribution, and then estimating the relevant parameters of that distribution. This method is not ideal, as the choice of limiting family is accompanied by some uncertainty. However, the Gumbel, Fréchet, and Weibull families of the three-types theorem can be combined into a single family have models having common distribution function. This is known as the generalized extreme value (GEV) distribution.

The GEV distribution G with location parameter $\mu \in (-\infty, \infty)$, scale parameter $\tilde{\sigma} > 0$, and shape parameter $\xi \in (-\infty, \infty)$ defined on the set $\{z : 1 + \xi (z - \mu) / \tilde{\sigma} > 0\}$ is

$$GEV(z; \mu, \tilde{\sigma}, \xi) = \begin{cases} \exp\left(-\left(1 + \xi \frac{z-\mu}{\tilde{\sigma}}\right)^{-\frac{1}{\xi}}\right), & \xi \neq 0 \\ \exp\left(\exp\left(-\frac{z-\mu}{\tilde{\sigma}}\right)\right), & \xi = 0 \end{cases}. \tag{7}$$

The shape parameter ξ determines which of the three families of limiting distributions is used: $\xi > 0$ corresponds to the Fréchet family, $\xi < 0$ corresponds to the Weibull family, and $\xi = 0$ corresponds to the Gumbel family. Thus, the three-types theorem says

$$P\left(\frac{D_n - b_n}{a_n} \leq z\right) \xrightarrow{d} GEV(z) \text{ as } n \to \infty. \tag{8}$$

Figure 2.1 shows the behavior of the GEV under three different shape parameters, each corresponding to a different limiting distribution.

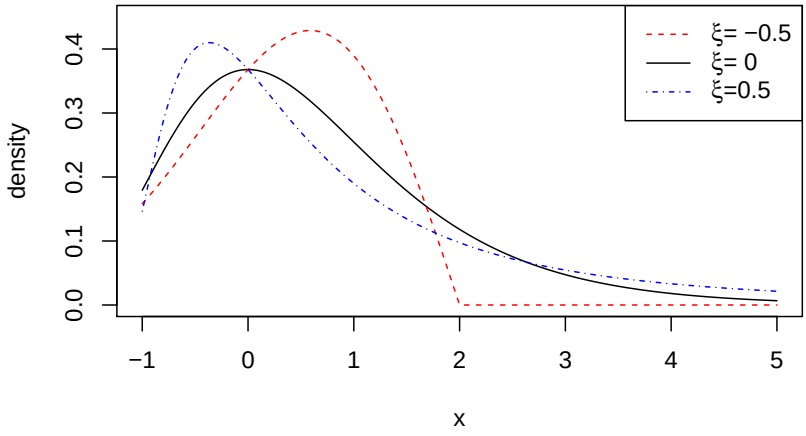

FIGURE 2.1. Plot of different GEVs. Each has $u = 0$ and $\sigma = 1$.

If n is large enough, then $GEV(z)$ is an approximation of (1), so that

$$P\left(\frac{D_n - b_n}{a_n} \leq z\right) \approx \exp\left(-\left(1 + \xi z\right)^{-\frac{1}{\xi}}\right). \tag{9}$$

This can be used to determine an approximation of the distribution of D_n. Let $z^* = a_n z + b_n$. Then

$$P\left(D_n \leq z^*\right) \approx \exp\left(-\left(1 + \xi\frac{z^* - b_n}{a_n}\right)^{-\frac{1}{\xi}}\right). \tag{10}$$

By treating a_n as the location parameter μ and b_n as the scale parameter $\tilde{\sigma}$, the GEV can be used as an approximate distribution of the D_n. In turn, the GEV can be used to fit a series of block maxima.

2.1.3. STATISTICAL PRACTICE WITH GEV. The GEV parameters are traditionally estimated using either numerical maximum likelihood methods or L-moments. The GEV

approach lends itself well to the estimation of return levels, which are extreme quantiles of the distribution of the (annual) maximum. Because of this, the block maxima approach is popular in environmental statistics.

2.2. Traditional Threshold Exceedance Methods

The block maxima approach has a weakness in that it discards many of the data points, some of which might be useful for describing extreme behavior. The block maxima approach is particularly problematic if several of the largest values are contained in the same block. The determination of the appropriate block size n can also cause problems. The block size choice represents a bias-variance trade-off so commonly seen in statistics: a block too small can result in the GEV approximation being poor, leading to bias in estimation and extrapolation; block size too large gives few maxima, leading to larger variance of the estimator, and consequently greater parameter uncertainty. Although the bias-variance trade-off remains, threshold based methods can be used to reduce the issue of wasting large values.

Let $Y_1, Y_2, \ldots$ be a sequence of i.i.d. random variables having distribution function F. The peaks over threshold method considers the Y_i exceeding some high threshold u as extreme events. While the distribution of threshold exceedances is known if F is known, this is not the case in practical applications. Thus, approximations that are broadly applicable for high values of the threshold are used, and asymptotic results lead to the on the generalized Pareto distribution.

2.2.1. The Generalized Pareto Distribution.

Pickands III (1975) and Balkema and De Haan (1974) showed that if a distribution is in the domain of attraction of the GEV, then the distribution of exceedances above a threshold u converges to a generalized Pareto distribution (GPD) as $u \to y_+$, where y_+ is the upper endpoint of the support of the

distribution. The GPD has a distribution given by

$$G(y;\sigma_u,\xi) = \begin{cases} 1-\left(1+\frac{\xi(y-u)}{\sigma_u}\right)^{-\frac{1}{\xi}}, & \xi \neq 0 \\ 1-\exp\left(-\frac{y-u}{\sigma_u}\right), & \xi = 0 \end{cases}, \tag{11}$$

where $\sigma_u > 0$ and depends on u, and $\xi \in (-\infty, \infty)$. G has support $y \geq u$ when $\xi \geq 0$ and $u \leq y \leq u - \sigma_u/\xi$ when $\xi < 0$. Additionally, the probability of a given observation exceeding the threshold u is denoted by τ_u, and this additional parameter is needed to calculate unconditional high quantiles.

The GPD exhibits a threshold stability property, in that once a GPD has been established for exceedances above a threshold u, the exceedances above all thresholds greater than u will also follow a GPD. That is, if $[Y|Y > u]$ is distributed GPD(σ_u, ξ), then $[Y|Y > u_0]$ for $u_0 > u$ is distributed GPD(σ_{u_0}, ξ), where $\sigma_{u_0} = \sigma_u + \xi\,(u_0 - u)$.

The GPD and GEV are closely related. The parameters of the GPD are uniquely determined by those of the associated GEV. In particular, the shape parameter ξ is unchanged between the two, and the scale parameter of the GPD is such that $\sigma_u = \tilde{\sigma} + \xi\,(u - \mu)$. Since the shape is unchanged, different values of ξ determine the nature of the GPD's tail. Positive shape indicates a heavy tail, negative shape a bounded tail, and shape of 0 an exponentially decaying tail. Figure 2.2 shows the tail behavior of the GPD under different shape parameter values.

Given an i.i.d. sample of size n, traditional threshold exceedance methods proceed by determining a threshold u above which a GPD approximation is reasonable. Only data exceeding this threshold are used to estimate σ_u and ξ, and τ_u is estimated by the observed proportion of exceedances.

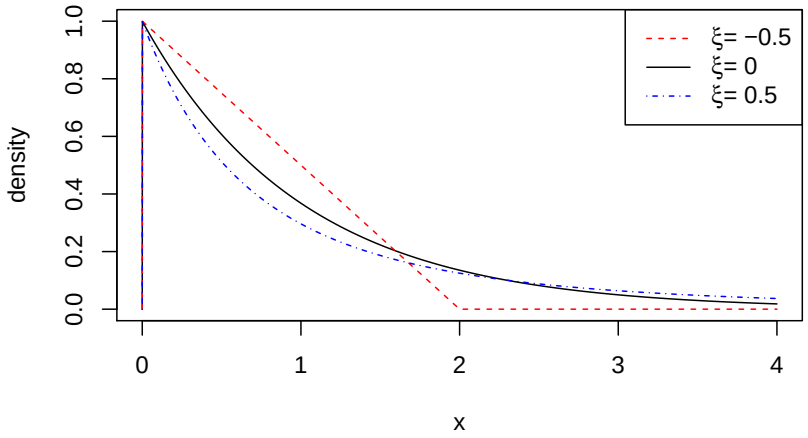

FIGURE 2.2. Plot of different GPDs. Each has $u = 0$ and $\sigma = 1$.

2.2.2. THRESHOLD SELECTION. Selecting an appropriate threshold is both important and difficult. If a chosen threshold is too low, then the GPD approximation will be poor, and estimates of high quantiles may be biased. If a chosen threshold is too high, then parameter estimates will have high variability due to inadequate sample size. Thresholds are commonly chosen using graphical methods such as mean exceedance plots and parameter stability plots, (Coles, 2001, Section 4.3.1). However, threshold selection remains subjective and imprecise.

Two types of diagnostic plots are often used for threshold selection. Figure 2.3 shows the mean exceedance plot (with 95% confidence intervals) for the data set of daily rainfall accumulations at a single location in south-west England, which is part of a larger data set detailed in Coles and Tawn (1996). If the distribution above a threshold u is exactly GPD, then the true mean exceedance $E[X|X > u]$ is a linear function of u. Empirical mean exceedance plots are therefore used to determine a u above which the mean exceedances appear to be linear and the GPD approximation is appropriate. Parameter stability plots for the same data set are given in Figure 2.4. The modified scale parameter is $\sigma^* = \sigma_u - \xi u$. The shape and modified scale parameter should both be constant above u_0 if the GPD

approximation is appropriate. When evaluating this data set, Coles (2001, Section 4.3.1) suggests that the mean exceedance plot appears to curve from $u = 0$ to about $u = 30$ before displaying a linear relationship until about $u = 60$. In my opinion, any curvature between $u = 5$ and $u = 30$ is unclear. I do agree that there is an approximately linear pattern between $u = 30$ and $u = 40$, but I find such a pattern occurring between $u = 40$ and $u = 60$ to be far more questionable. Meanwhile, the stability plots suggest that both the modified scale and shape parameters are constant for thresholds between $u = 0$ and $u = 40$. Taking both the mean exceedance and threshold stability plots into account suggest that a threshold of between 30 and 40 would be appropriate.

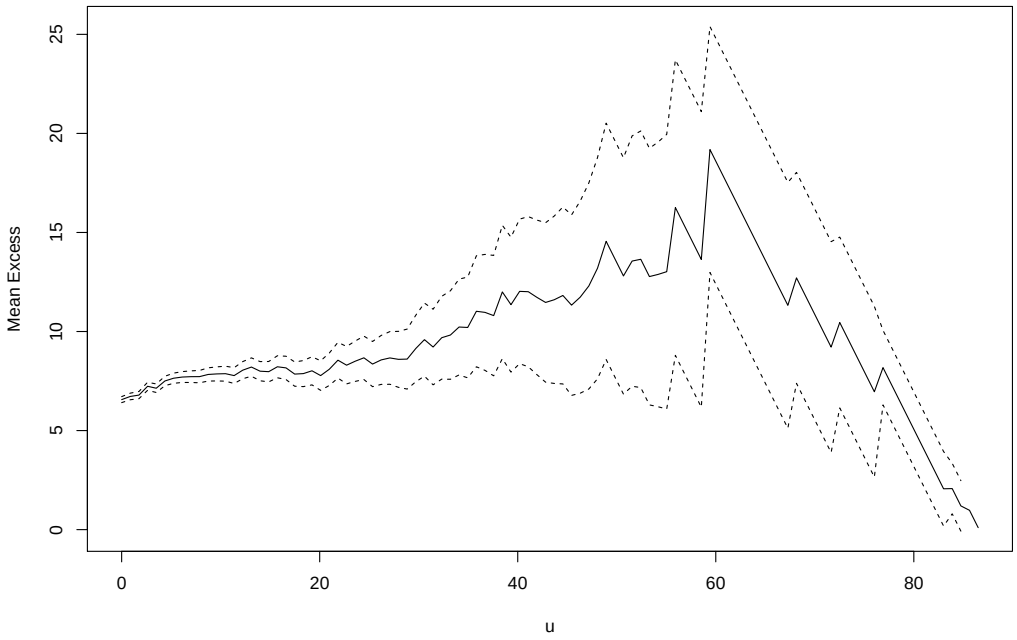

FIGURE 2.3. Mean exceedance plot for daily rainfall.

Figures 2.5 and 2.6 are the mean excess and stability plots for surge heights at a single location off south-west England (Coles, 2001, Example 1.10, Section 1.2). It is somewhat easier to see the desired patterns in these plots, but threshold selection is still imprecise. The mean exceedance plot shows a linear relationship between about 0.1 and 0.3. The

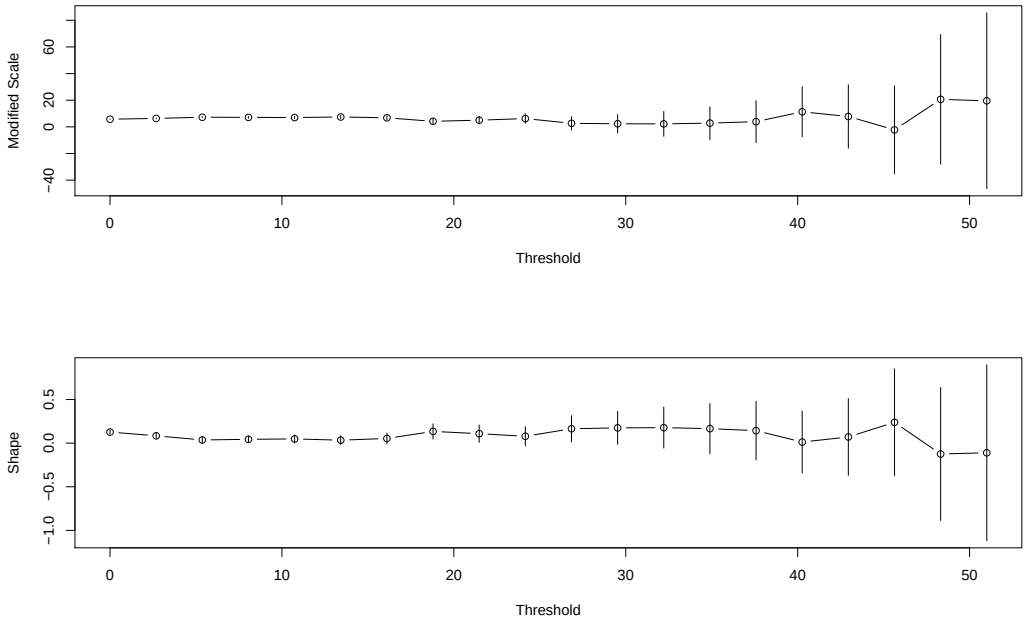

FIGURE 2.4. Stability plots for daily rainfall.

stability plots show that both the shape parameter and the modified scale parameter are fairly constant between about 0.1 and 0.3. Therefore, a threshold of between 0.1 and 0.3 would seem appropriate here.

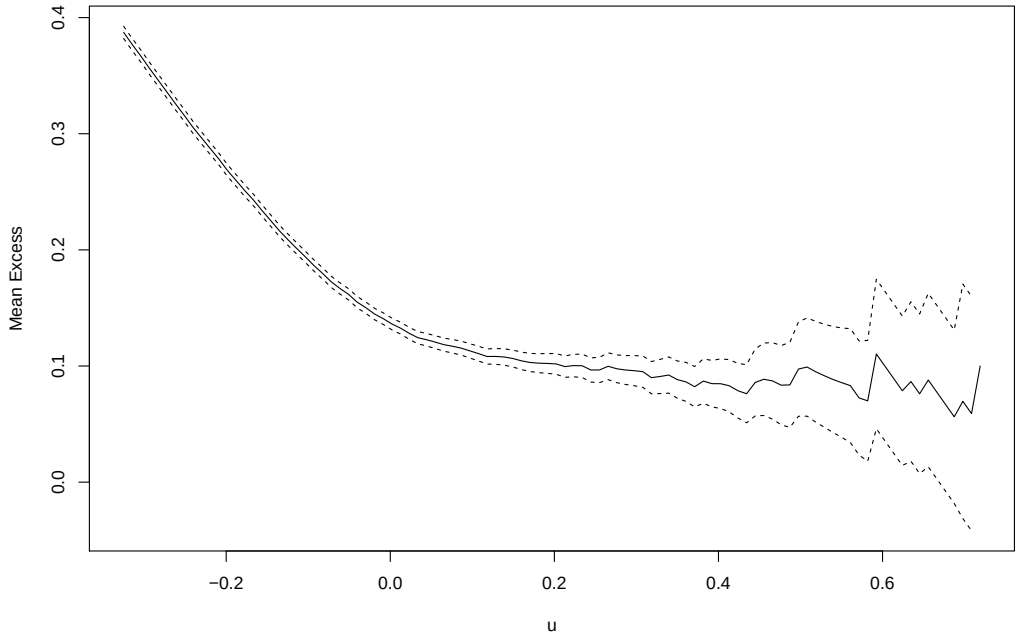

FIGURE 2.5. Mean exceedance plot for surge heights.

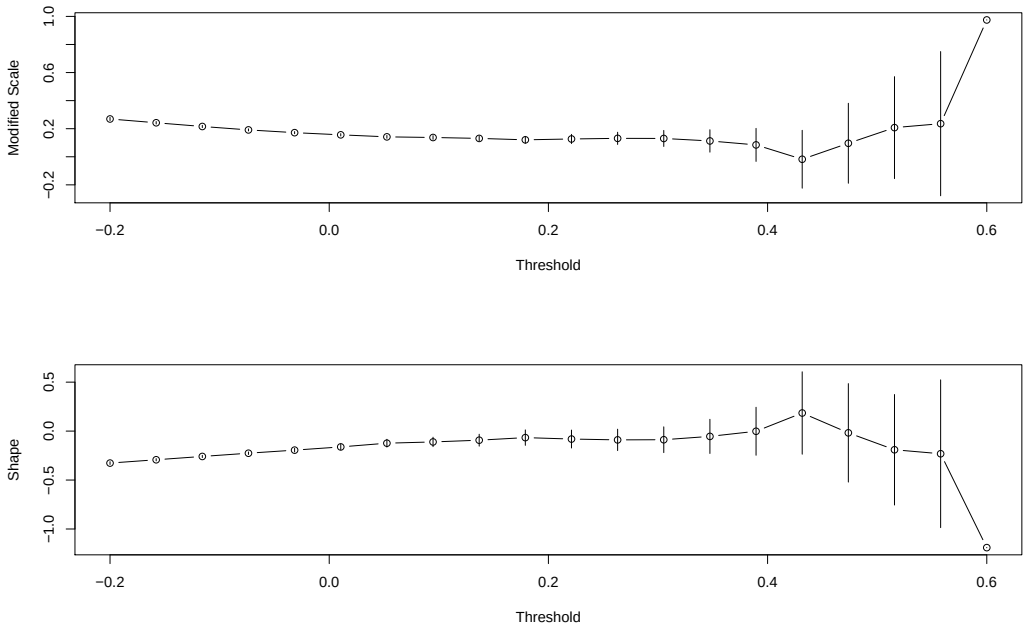

FIGURE 2.6. Stability plots for surge heights.

Motivated by the subjectivity of threshold selection methods, there has been some work to develop automated threshold selection methods, though these have thus far failed to replace the graphical methods. Guillou and Hall (2001) designed a method that uses the Hill estimator (Hill, 1975). Xiangxian and Wenlei (2009) created an algorithm for estimating the threshold that attempts to minimize the MSE while also using a Kolmogorov-Smirnov statistic to check if the GPD approximation fits the empirical distribution of excesses well. Bootstrap methods that attempt to minimize the MSE have been proposed by Caers and Dyck (1998) and Danielsson et al. (2001).

Alternative automated threshold selection methods, such as those of Behrens et al. (2004), Tancredi et al. (2006), and MacDonald et al. (2011), model the data below the threshold by fitting a parametric or more flexible model to the bulk of the distribution while fitting a GPD or other extreme value model above the threshold. As extremes methods wish to "let the tail speak for itself", a concern of any approach which uses non-extreme data is that the data in the bulk of the distribution could contaminate tail inference. These automatic threshold

selection methods are somewhat related to some of the approaches I used in modeling the upper τth proportion of a distribution, though my inference approach differs in that τ is fixed. As such, these methods will be revisited in Section 4.1.1.

2.2.3. Parameter Estimation. Once a proper threshold has been selected, the parameters of the GPD can be estimated. Numerical maximum likelihood is a commonly used option, as analytical maximization is not possible. Since maximum likelihood estimation requires a fixed set of data, this method requires that the threshold is selected before estimation, and the GPD is fit only to observations exceeding the threshold. If the threshold is not chosen before estimation, the exceedances are no longer fixed. This distinction will be important in Chapter 4.

An alternative to maximum likelihood estimation is the method of probability-weighted moments (PWM), first introduced by Hosking and Wallis (1987), though PWM implicitly assumes $\xi < 1$. Furthermore, the PWM method may result in estimates inconsistent with the observed data. This can occur if $\xi < 0$ if some of the observations fall above the estimate of the right endpoint. It has been suggested that PWM may outperform numerical maximum likelihood in cases of small sample size (Hosking and Wallis, 1987). Coles and Dixon (1999) respond that if a penalty similar to the assumption that $\xi < 1$ is applied, maximum likelihood is competitive with PWM. In the simulation study in Chapter 6, I will impose a penalty similiar to Coles and Dixon (1999).

2.2.4. Point Process Characterization. The threshold exceedances can also be modeled using a point process characterization. Let $Y_1, \ldots, Y_n$ be independent random variables with common distribution F such that equation (2) holds. Then the sequence of

point processes

$$N_n = \left\{ \left(\frac{i}{n+1}, \frac{Y_i - b_n}{a_n} \right); i = 1, \ldots, n \right\}, \tag{12}$$

for a_n and b_n appropriately chosen as in Section 2.1.1, converges to a Poisson process with intensity measure ν given by

$$\nu\left([t_1, t_2] \times [y, \infty)\right) = (t_2 - t_1)\left(1 + \xi\left(\frac{y - \mu}{\tilde{\sigma}}\right)\right)^{-\frac{1}{\xi}},$$

for $t_1 < t_2$.

The point process representation is advantageous in that is uses the GEV parameterization. Thus, all parameters are invariant to the threshold, whereas in the GPD representation, scale is dependent on threshold. Furthermore, the threshold exceedance rate forms part of the inference in the point process characterization. One of the advantages of the GPD$_\tau$ model that I will develop in Chapter 3 is that its parameters will not be functions of u_τ.

2.2.5. Use of Covariates. When covariate information is available, the data are no longer identically distributed across different covariate values. Regression methods for extremes allow the characteristics of the tail to change with covariates. Several studies have employed models where the shape and scale parameters of the generalized Pareto distribution vary with covariates (Beirlant et al., 2004, Section 7.4). It is less common for the threshold to vary with covariates in traditional methods. If it is desired that the threshold vary with covariates, the point process characterization detailed by Smith (1989) can be used.

Coles (2001, Section 7.6) suggests using the point process setting over a threshold exceedance model when working with time-varying thresholds, such as in the Wooster temperature data shown in Figure 2.7. This data consists of daily minimum temperatures (degrees Fahrenheit) in Wooster, Ohio for 1983 through 1987. A strong seasonal effect is apparent in Figure 2.7, and Coles (2001, Section 7.7) suggests using a threshold that gives an approximately uniform rate of exceedances over the course of the year, so that the threshold also varies with season.

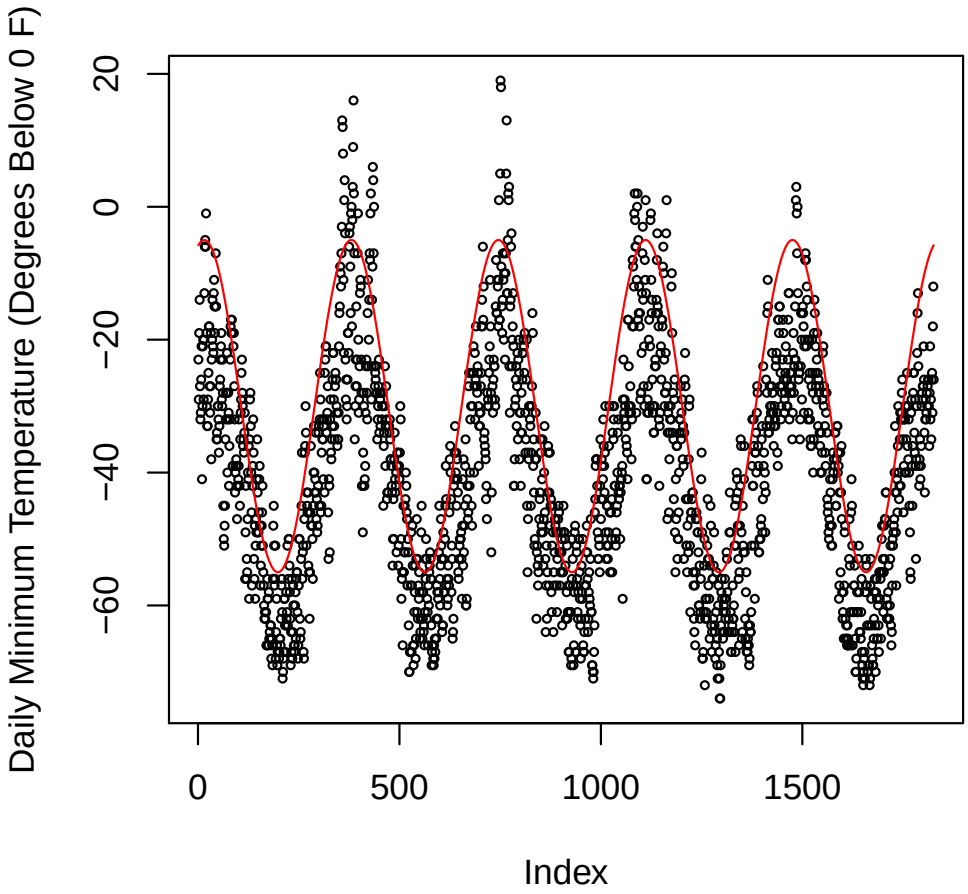

FIGURE 2.7. Wooster temperature data with time-varying threshold.

2.3. The Role of Extreme Value Theory in Facial Recognition

Extreme value theory has been used in some biometric recognition problems. Daugman (2006) uses the GEV as a model for the non-match distribution for iris comparisons. In this application, k different rotations of the same eye are compared to irises in a database. Only the best of these k rotations is kept in the non-match distribution. Shi et al. (2008) recognize that the tails of both the non-match and match distributions in a biometric system should follow a GPD. They choose a GPD threshold using a mean excess plot and fit a GPD tail to fingerprint comparison data.

Scheirer et al. (2010) theorizes that when comparing a single query image to all possible targets, the sampling of the top n similarity scores will result in a GEV. They attempt to justify the use of the method of block maxima to model the top n similarity scores by arguing that each of these top n scores is "likely to have been sampled from the extreme of their underlying portfolio," defining each portfolio as an independent subset of the overall non-match distribution. Since the independence assumption necessary for the three-types theorem to hold is suspect, the authors argue in Scheirer et al. (2011) that the maxima of the portfolios are exchangeable random variables, and thus the three-types theorem holds following Berman (1962). The fitted GEV is used to determine if the largest similarity score is an outlier, which would suggest it belongs in the match distribution, ultimately using the information gained to compare the performance of different algorithms.

CHAPTER 3

A New Model for the Tail Above the $(1-\tau)$th Quantile

In this chapter, I develop the GPD_τ distribution, a model for the upper τth proportion of a distribution. Because my aim is to model the upper tail corresponding to a *fixed* proportion τ, my approach cannot be viewed in the usual context of extreme value theory. Nevertheless, my argument justifying GPD_τ borrows ideas from extremes. In particular, the classical development of the GPD from the assumption that the distribution is in the domain of attraction of the GEV provides the framework for my model. For the GPD approximation to be valid, the fixed τ must be relatively small. In practice, τ would likely need to be less than 10%. Furthermore, the choice of τ has implications on how far into the tail one could make practical inference. The formal development follows.

Recall that the three-types theorem (Fisher and Tippett, 1928; Gnedenko, 1943) states that as $n \to \infty$,

$$P^n\left(\frac{Y-b_n}{a_n} \leq y\right) \to \exp\left[-\left(1+\xi y\right)^{-\frac{1}{\xi}}\right],$$

as stated in (9). Assuming n is fixed and large enough for the above convergence to imply approximate equality, then for z a high quantile of Y,

$$nP\left(Y > z\right) \approx \left(1+\xi\frac{z-b_n}{a_n}\right)^{-\frac{1}{\xi}}.$$

As long as this approximation is appropriate for u_τ, then

$$nP\left(Y > u_\tau\right) = n\tau \approx \left(1+\xi\frac{u_\tau-b_n}{a_n}\right)^{-\frac{1}{\xi}}.$$

Treating the approximation as an equality and solving for b_n yields $b_n = u_\tau - a_n/\xi \left[(n\tau)^{-\xi} - 1\right]$ so that for $z > u_\tau$,

$$nP\left(Y > z\right) \approx \left(\xi \frac{z - u_\tau}{a_n} + (n\tau)^{-\xi}\right)^{-\frac{1}{\xi}}.$$

Conditioning on $Y > u_\tau$ returns

$$P\left(Y > z | Y > u_\tau\right) = \frac{nP(Y > z, Y > u_\tau)}{nP(Y > u_\tau)} = \frac{\left(\xi \frac{z-u_\tau}{a_n} + (n\tau)^{-\xi}\right)^{-\frac{1}{\xi}}}{n\tau}.$$

Now assume (as in statistical practice) that n is fixed and define $\sigma = a_n n^{-\xi}$, which allows n to be eliminated, so that

$$P\left(Y > z | Y > u_\tau\right) = \frac{1}{\tau}\left(\xi \frac{z - u_\tau}{\sigma} + \tau^{-\xi}\right)^{-\frac{1}{\xi}}. \tag{13}$$

I will refer to the conditional distribution given in (13) as the GPD_τ, and its density is given by

$$g_\tau(z; u_\tau, \sigma, \xi) = \frac{1}{\tau\sigma}\left(\xi \frac{z - u_\tau}{\sigma} + \tau^{-\xi}\right)^{-\frac{1}{\xi}-1} \tag{14}$$

for $z \geq u_\tau$ when $\xi \geq 0$ and $u_\tau \leq z \leq u_\tau - \sigma\tau^{-\xi}/\xi$ when $\xi < 0$. Importantly, the scale parameter in (13) and (14) does not depend on the threshold. For $\xi = 0$, both (13) and (14) should be interpreted as limits just as with the standard GPD.

Because the scale and threshold parameters are independent in GPD_τ, σ is not equivalent to the scale of a standard GPD with the same threshold. However, σ can be used to find

the scale in the corresponding GPD. Let $[Y|Y > u_\tau] \sim \text{GPD}_\tau(u_\tau, \sigma, \xi)$. Then, using (13),

$$\begin{aligned} P(Y > z|Y > u_\tau) &= \frac{1}{\tau}\left(\xi\frac{z-u_\tau}{\sigma} + \tau^{-\xi}\right)^{-\frac{1}{\xi}} \\ &= \left(\frac{1}{\tau^{-\xi}}\right)^{-\frac{1}{\xi}}\left(\xi\frac{z-u_\tau}{\sigma} + \tau^{-\xi}\right)^{-\frac{1}{\xi}} \\ &= \left(\xi\frac{z-u_\tau}{\sigma\tau^{-\xi}} + 1\right)^{-\frac{1}{\xi}}. \end{aligned}$$

Let $\sigma_0 = \sigma\tau^{-\xi}$. Then,

$$P(Y > z|u > u_\tau) = \left(\xi\frac{z-u_\tau}{\sigma_0} + 1\right)^{-\frac{1}{\xi}}. \tag{15}$$

Since (15) is the GPD cumulative distribution function given in (11), then $[Y|Y > u_\tau] \sim \text{GPD}\left(u_\tau, \sigma\tau^{-\xi}, \xi\right)$.

The GEV can be shown to be the class of limiting distributions of the maximum of i.i.d. random variables as block size increases. The GPD is the limiting distribution of threshold exceedances as the threshold approaches the upper tail. Although GPD_τ does not have a similar asymptotic justification, it is nevertheless sensible to assume that it is a good approximation of the tail for sufficiently small τ. Furthermore, it has the benefit of being a general model for the tail, requiring only that the underlying distribution is in the domain of attraction of the GEV. Thus, GPD_τ can be used to model the tail of a distribution as long as τ is suitably small.

3.1. Threshold Stability

Notably, a version of the threshold stability property characterized by the generalized Pareto distribution is exhibited by GPD_τ. Let Y be a random variable. Consider a fixed τ

such that $P(Y > u_\tau) = \tau$. Assume $[Y|Y > u_\tau] \sim \text{GPD}_\tau(u_\tau, \sigma, \xi)$, so that for $y > u_\tau$,

$$P(Y > y|Y > u_\tau) = \frac{1}{\tau}\left(\xi\frac{y - u_\tau}{\sigma} + \tau^{-\xi}\right)^{-\frac{1}{\xi}}$$

by (13). Let $\tau^* < \tau$. Consider u_{τ^*} such that $P(Y > u_{\tau^*}) = \tau^*$, indicating that $u_\tau < u_{\tau^*}$. Then,

$$\begin{aligned} P(Y > y|Y > u_{\tau^*}) &= P(Y > y|Y > u_{\tau^*}, Y > u_\tau) \\ &= \frac{P(Y > y, Y > u|Y > u_{\tau^*})}{P(Y > u_{\tau^*}|Y > u_\tau)} \\ &= \frac{P(Y > y|Y > u_\tau)}{P(Y > u_{\tau^*}|Y > u_\tau)}. \end{aligned}$$

Thus, using (13),

$$\begin{aligned} P(Y > y|Y > u_{\tau^*}) &= \frac{\frac{1}{\tau}\left(\xi\frac{y-u_\tau}{\sigma} + \tau^{-\xi}\right)^{-\frac{1}{\xi}}}{\frac{1}{\tau}\left(\xi\frac{u_{\tau^*}-u_\tau}{\sigma} + \tau^{-\xi}\right)^{-\frac{1}{\xi}}} \\ &= \left(\frac{\xi\frac{y-u_\tau}{\sigma} + \tau^{-\xi}}{\xi\frac{u_{\tau^*}-u_\tau}{\sigma} + \tau^{-\xi}}\right)^{-\frac{1}{\xi}} \\ &= \left(\frac{\xi\left(\frac{y-u_{\tau^*}}{\sigma} + \frac{u_{\tau^*}-u_\tau}{\sigma}\right) + \tau^{-\xi}}{\xi\frac{u_{\tau^*}-u_\tau}{\sigma} + \tau^{-\xi}}\right)^{-\frac{1}{\xi}} \\ &= \frac{\left(\xi\frac{y-u_{\tau^*}}{\sigma} + \xi\frac{u_{\tau^*}-u_\tau}{\sigma} + \tau^{-\xi}\right)^{-\frac{1}{\xi}}}{\left(\xi\frac{u_{\tau^*}-u_\tau}{\sigma} + \tau^{-\xi}\right)^{-\frac{1}{\xi}}}. \end{aligned}$$

Now, let

$$\tau^* = \left(\xi\frac{u_{\tau^*} - u_\tau}{\sigma} + \tau^{-\xi}\right)^{-\frac{1}{\xi}} \tag{16}$$

so that

$$P\left(Y > y | Y > u_{\tau^*}\right) = \frac{1}{\tau^*}\left(\xi\frac{y - u_{\tau^*}}{\sigma} + \tau^{*-\xi}\right)^{-\frac{1}{\xi}}.$$

Notice that $[Y|Y > u_{\tau^*}] \sim \text{GPD}_{\tau^*}\left(u_{\tau^*}, \sigma, \xi\right)$ based on (13). Furthermore,

$$\begin{aligned} P\left(Y > u_{\tau^*}\right) &= P\left(Y > u_{\tau^*} | Y > u_\tau\right) P\left(Y > u_\tau\right) \\ &= \frac{1}{\tau}\left(\xi\frac{u_{\tau^*} - u_\tau}{\sigma} + \tau^{-\xi}\right)^{-\frac{1}{\xi}} \tau \\ &= \left(\xi\frac{u_{\tau^*} - u_\tau}{\sigma} + \tau^{-\xi}\right)^{-\frac{1}{\xi}}, \end{aligned}$$

and by (16),

$$\left(\xi\frac{u_{\tau^*} - u_\tau}{\sigma} + \tau^{-\xi}\right)^{-\frac{1}{\xi}} = \tau^*.$$

Therefore, if $P\left(Y > u_\tau\right) = \tau$ and $[Y|Y > u_\tau] \sim \text{GPD}_\tau\left(u_\tau, \sigma, \xi\right)$, then $[Y|Y > u_{\tau^*}] \sim \text{GPD}_{\tau^*}\left(u_{\tau^*}, \sigma, \xi\right)$ where $P\left(Y > u_{\tau^*}\right) = \tau^*$.

In the next chapter, I develop inference methods for the GPD_τ parameters u_τ, σ, and ξ. Inference for u_τ will require more than just the density in (14), as likelihood methods are unsuited for estimation of the parameters of GPD_τ.

CHAPTER 4

Parameter Estimation

Given a set of observations, fitting the model from Chapter 3 would entail obtaining estimates for the parameters u_τ, σ, and ξ. One estimation method used in traditional extremes threshold exceedance modeling is (numerical) maximum likelihood. Recall that a sample density considered as a function of the parameters for *fixed* observations is considered a likelihood (Lehmann and Casella, 1998, Section 6.3). For traditional GPD modeling, once the threshold is selected, the data exceeding the threshold are fixed and the generalized Pareto density can be used to construct a likelihood. Such an approach cannot be used with the density given in (14) as u_τ is a parameter and the data exceeding this threshold is not fixed. Thus, an alternative estimation method is needed.

A method that uses all data, both above and below u_τ, is needed to estimate u_τ. However, it is important that the data below u_τ does not influence the shape and scale parameters in GPD_τ. Thus, a piece in addition to GPD_τ is necessary in order to properly estimate all three parameters.

4.1. Unsuccessful Attempts to Estimate u_τ

The biggest difference between estimation here and in the traditional threshold exceedance setting is that u_τ must be estimated. My attempt to develop an objective function that could be used to accurately estimate u was met with several ineffective forms before achieving success. It is helpful to discuss some of these attempts, as well as the underlying reasons for their failure, as this will help to illustrate the usefulness of the estimator employed for the analysis.

4.1.1. Modeling the Bulk Distribution. Estimating u_τ requires the use of all of the data available, at least to the extent of utilizing information about the empirical quantile. Automatic threshold selection methods, as described in Chapter 2.2.2, have considered how to use data in the bulk of the distribution while not contaminating information in the tail. One suggested approach has been via mixture models, as described in Scarrott and MacDonald (2012, Section 6). These mixture models tend to model not only the tail of a distribution with a GPD, but also the bulk of the distribution below the threshold u, using density

$$f(y) = \begin{cases} f_1(y) \text{ for } y \leq u \\ \\ f_2(y) \text{ for } y > u \end{cases}, \tag{17}$$

where f_2 is assumed GPD. The choice of f_1 is the main way in which the different mixture model approaches differ. The major advantage of mixture models over other efforts in choosing u is that uncertainty in such a choice can be incorporated into one's inferences. However, extreme value models for the tail are attractive because they make no assumption about the underlying distribution, but the mixture model in (17) implies that a sensible model for the distribution below the threshold is necessary.

The approaches of Behrens et al. (2004), do Nascimento et al. (2012), and Frigessi et al. (2002) all use parametric or semiparametric bulk models for f_1. Model misspecification is an issue, as inference suffers if the chosen bulk distribution is inappropriate. Robustness is also a common issue, in that the bulk model often exerts influence on the tail. Additionally, each of these methods treats the GPD scale parameter and threshold as independent.

Use of nonparametric bulk models circumvent misspecification issues. Furthermore, the point process representation of the GPD can be used to overcome the scale and threshold dependence issue. The models of both Tancredi et al. (2006) and MacDonald et al. (2011) proceed in such a matter. Tancredi et al. (2006) use a "mixture of uniforms" density estimator for f_1, whereas MacDonald et al. (2011) use a symmetric kernel density estimator for f_1. Neither attempts to allow the threshold to vary with covariates.

Nonparametric models for the bulk would seemingly be an ideal solution, as they provide a model for the bulk which has very little, if any, influence on the tail. However, an issue does arise specifically because of the flexibility associated with a nonparametric model for f_1 in (17). Because the GPD is a limiting distribution for the tail above a high threshold, the nonparametric model f_1 is often found to fit the data better than the GPD f_2 very far into the tail (and often for the entire data set). Of course, an entirely nonparametric model is not useful for extrapolating further into the tail. In threshold selection contexts, the flexibility of the kernel coupled with the nature of the GPD often results in very high thresholds being selected. MacDonald et al. (2011) alleviate this issue by employing a Bayesian approach with an informative prior on u. The use of covariates would complicate the Bayesian prior specification.

I tried to use a similar model as that of MacDonald et al. (2011) by using (17) with

$$f_1 = \tau \frac{h\left(y;\lambda\right)}{H\left(y;\lambda\right)} \mathbb{I}_{y_u<u}$$

and

$$f_2 = \left(1-\tau\right) g_\tau\left(y;u,\sigma,\xi\right) \mathbb{I}_{y_i \geq u},$$

for g_τ as in (14) and kernel density h with bandwidth λ. This model allows one to directly construct a likelihood, but simulation studies showed that the estimates for u were biased high due to the flexibility of the kernel. Use of an empirical cdf in place of the kernel density exhibited the same issue. Ultimately, I found that it was not necessary to model the data in the bulk, so my subsequent approach will not have a model for f_1. However, I will no longer be able to directly use likelihood-based methods, as will be explained.

4.1.2. Using a Binomial Distribution to Model the Exceedances. Since only the number of observations that appear in the bulk versus the tail of the distribution is important, another approach I attempted was to use a binomial distribution to model whether an observation exceeded the threshold u. A GPD_τ was simultaneously used to model the exceedances. The objective function to be used for estimation was thus

$$M_n\left(u_\tau, \sigma, \xi; \mathbf{y}\right) = \log b\left(k; n, \tau\right) + \sum_{i=1}^{n} \log g_\tau\left(u_\tau, \sigma, \xi; y_i\right) \mathbb{I}_{y_i \geq u_\tau}, \tag{18}$$

where $k = \sum_{i=1}^{n} \mathbb{I}_{y_i \geq u_\tau}$ and b is the binomial probability mass function, so that

$$b\left(k; n, \tau\right) = \binom{n}{k} \tau^k \left(1-\tau\right)^{n-k}.$$

This model does not run the risk of having the bulk distribution's model overtake that of the tail, as the bulk is not modeled. At the same time, the binomial distribution has great influence on the choice of threshold. Simulation studies for this method proved it suitable for parameter estimation if no covariates were used. Since the scale and shape parameters appear only in the GPD_τ piece, the binomial piece has limited effect on the distribution of

the tail. Monte Carlo simulations suggested that the threshold is consistent. However, the introduction of covariates proved to be an issue.

While the binomial distribution works towards an estimated model that has about $\tau\%$ of the observations exceeding u_τ, it does not take into account the *location* of the exceedances. Simulation studies showed that this leads to identifiability issues, in that the optimization is satisfied as long as about $\tau\%$ of the observations exceed u_τ, regardless of where these exceedances occur. Most frequently, the estimates are chosen such that a majority of the exceedances occur at small values of the covariate, with very few, if any, occurring at high values. Figure 4.1 shows one such example, where the plotted line shows the estimated u_τ line such that $u_\tau = \beta_0 + \beta_1 x_i$.

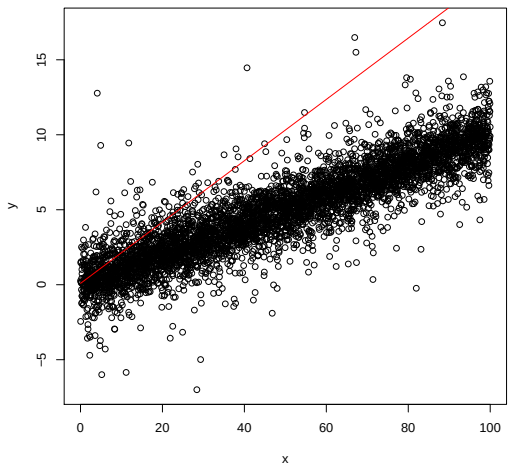

FIGURE 4.1. An example displaying the identifiability issue associated with the use of the binomial distribution in the objective function. The plotted line is the estimated u_τ line.

Quantile regression (Koenker, 2005) is a well-developed method for modeling a quantile as a function of covariates. In the next section, I show that substituting a quantile regression based function for the binomial distribution is a sensible alternative.

4.2. Quantile Regression Background

While standard least squares regression is used to estimate the conditional mean of a response variable, quantile regression is concerned with estimating the conditional quantiles of a response variable. First developed in Koenker and Bassett Jr (1978), the authors recognized that the quantile could be defined as the solution to a problem minimizing a sum of asymmetrically weighted absolute residuals (Hallock and Koenker, 2001). In other words, the unbiased θth quantile estimate is the solution to

$$\min_{\eta \in \Re} \sum \rho_\theta \left(y_i - \eta \right), \tag{19}$$

where ρ_θ is a loss function defined so that $\rho_\theta \left(z \right) = z \left(\theta - \mathbb{I}_{z<0} \right)$. Most importantly, the authors expressed the problem of finding the θth sample quantile as the solution to an optimization problem rather than through the sorting and ordering of the sample observations.

Koenker (2005, Chapter 1.4) defines the θth *conditional* quantile function as $Q_y \left(\theta | x \right) = \mathbf{x}^{\mathrm{T}} \beta \left(\theta \right)$, where $\hat{\beta} \left(\theta \right)$ is the solution to

$$\min_{\beta \in \Re^p} \sum \rho_\theta \left(y_i - \mathbf{x}^{\mathrm{T}} \beta \right). \tag{20}$$

So quantile regression estimates the sample quantile by replacing the scalar η in (19) with the parametric function $\eta \left(\mathbf{x}, \beta \right) = \mathbf{x}^{\mathrm{T}} \beta$. The minimization problem in (20) is usually efficiently solved through linear programming methods. Asymptotic theory for quantile regression is well developed in Koenker (2005, Chapter 4).

4.3. The Objective Function

Quantile regression can be sensibly combined with the model in Chapter 3 to obtain estimates for u_τ, σ, and ξ. A sequential approach could be employed, first estimating u_τ using quantile regression and then, treating u_τ as fixed, using (14) to create a likelihood. However, a disadvantage of this method is that it would not propagate the uncertainty in the threshold. Instead, since quantile regression and maximum likelihood are both M-estimators, I will use an objective function which combines the loss function from quantile regression and a 'likelihood' for estimating the GPD_τ parameters, which will allow the parameters to be estimated simultaneously.

Let $\boldsymbol{y} = (y_1, \ldots, y_n)^T$, where y_i are independent observations. The basic objective function I employ is

$$M_n\left(u_\tau, \sigma, \xi; \boldsymbol{y}\right) = \sum_{i=1}^{n} q\left(u_\tau; y_i\right) + \frac{1}{N} \sum_{i=1}^{n} \log g_\tau\left(u_\tau, \sigma, \xi; y_i\right) \mathbb{I}_{y_i \geq u_\tau}, \tag{21}$$

where $N = \sum_{i=1}^{n} \mathbb{I}\{y_i > u_\tau\}$ and

$$q\left(u_\tau; y_i\right) = \tau\left(y_i - u_\tau\right) \mathbb{I}_{y_i < u_\tau} + (\tau - 1)\left(y_i - u_\tau\right) \mathbb{I}_{y_i \geq u_\tau} \tag{22}$$

arises from (19). Thus, the objective function is the quantile regression objective function plus the *mean* log-"likelihood" contribution of the exceedances. I will perform M-estimation; that is, I seek the u_τ, σ, and ξ which maximize (21). The objective function has the appealing property that only observations which exceed u_τ will influence the estimates of σ and ξ.

One of my goals is to link the tail of a distribution to covariates, so each of the parameters may be a parametric function of covariates: $u_\tau = f_u(\boldsymbol{\beta}, X), \sigma = f_\sigma(\boldsymbol{\gamma}, X), \xi = f_\xi(\boldsymbol{\eta}, X)$.

This approach is analogous to regression approaches in extremes, which have been applied to both the generalized extreme value (GEV) distribution and the traditional GPD. The use of covariates will be discussed in more detail in Chapter 5.

It is worth providing some explanation of why the mean log-"likelihood" contribution is taken in (21) rather than the sum. Generally, a log-likelihood's magnitude increases with sample size; becoming increasingly negative (positive) if the likelihood contributions tend to be negative (positive). If the mean were replaced with the sum in (21), this second term's magnitude would increase with N, the number of exceedances above u_τ. This results in biased estimates for u_τ. In my investigations, the contribution from the GPD_τ piece tends to be negative, thus estimates of u_τ would be biased high as the naive objective function (with a sum rather than mean) would favor values which result in too few exceedances. With the mean log-"likelihood", the second term of (21) converges to the mean log-likelihood contribution above u_τ. Importantly for a given u, the same values of σ and ξ which maximize the mean log-"likelihood" also maximize the standard log-likelihood.

4.4. M-estimators

An M-estimator, or maximum likelihood type estimator, is any estimate θ defined by minimizing $\sum_{i=1}^{n} \rho(x_i; \theta)$ (Huber, 2011, Chapter 3). M-estimators are a very broad class of estimators, which includes the ordinary maximum likelihood estimator.

When ρ is a differentiable function with respect to θ, the minimization problem is equivalent to solving $\sum \psi(x_i; \theta) = 0$ for θ, where $\psi(x; \theta) = (\partial/\partial\theta)\rho(x; \theta)$. When such differentiation is possible, the M-estimator is said to be of ψ-type. Otherwise, the M-estimator is of ρ-type.

One of the advantages of the maximum likelihood estimator is that well known sufficient conditions exist establishing the consistency of the estimator. These conditions, often called regularity conditions, are satisfied in most reasonable problems (Casella and Berger, 1990). There are numerous consistency results for M-estimators of ψ-type where ψ is monotone (or ρ is convex) under general regularity conditions, such as those in Huber (2011, Chapter 3, Corollary 2.2), Haberman (1989), Niemiro (1992), and Hjort and Pollard (2011). However, the objective function in (21) does not adhere to such a convexity argument.

M-estimators where ψ is not monotone are called "redescending" (Maronna et al., 2006, Section 2.2). Consistency results for redescending M-estimators are more complicated than in monotone cases. Hampel et al. (1986, Chapter 2.5) places two assumptions on the population density and four assumptions on ψ to show consistency, which Shevlyakov et al. (2008) extends to three and five assumptions, respectively, but both of these results rely on symmetry of the population distribution, as do results in Freedman and Diaconis (1982). Results of Mizera (1994) require that ψ is unimodal. Jurecková and Sen (1996, Chapter 7) give consistency results for nonmonotone ψ that are sufficiently smooth in θ.

Huber (2011, Chapter 6) gives consistency under five assumptions for ρ-type M-estimators. Three of these conditions rely on the existence of unknown functions. The choice of such functions is not obvious.

4.5. Estimator Consistency

Since the requirements of the known results for consistency of redescending M-estimators cannot be shown for $M_n\left(u, \sigma, \xi; \boldsymbol{y}\right)$ as defined in (21), I will use a more direct approach. Assume that $[Y|Y > u_\tau] \sim \text{GPD}_\tau$ and $[Y|Y > u^*] \sim \text{GPD}_{\tau^*}$. Define u_τ such that $\text{P}(Y > u_\tau) = \tau$. Define $u^* \neq u_\tau$ such that $\text{P}(Y > u^*) = \tau^*$.

Since σ and ξ only appear in the GPD_τ portion of the objective function defined in (21), for any fixed $u > \min(u_\tau, u^*)$, let $(\hat{\sigma}_u, \hat{\xi}_u) = \text{argmax}_{(\sigma,\xi)} M_n(u, \sigma, \xi)$. Then $(\hat{\sigma}_u, \hat{\xi}_u) \xrightarrow{p} (\sigma, \xi)$, as these estimates correspond to the maximum likelihood estimates for a likelihood based on the log-density of the GPD_τ for exceedances over u.

Thus, I will focus on showing that the estimator

$$\hat{u}_n = \underset{u}{\text{argmax}}\, M_n\left(u, \hat{\sigma}_\tau, \hat{\xi}_\tau; \boldsymbol{y}\right), \tag{23}$$

where $\hat{\sigma}_\tau$ and $\hat{\xi}_\tau$ are as defined above, is consistent. I will first show that as $n \to \infty$,

$$P\left(M_n\left(u_\tau, \hat{\sigma}_\tau, \hat{\xi}_\tau; \boldsymbol{y}\right) - M_n\left(u^*, \hat{\sigma}^*, \hat{\xi}^*; \boldsymbol{y}\right) > 0\right) \to 1. \tag{24}$$

Note that plugging u^* into M_n creates a mismatch: the true probability that an observation exceeds u^* is τ^*, but M_n fixes this at τ.

The difference in (24), $M_n\left(u_\tau, \hat{\sigma}_\tau, \hat{\xi}_\tau; \boldsymbol{y}\right) - M_n\left(u^*, \hat{\sigma}^*, \hat{\xi}^*; \boldsymbol{y}\right)$, can be separated into a quantile regression difference and a GPD_τ difference. I will look at these two differences separately, but ultimately I will show that the quantile regression difference will grow with n, whereas the GPD_τ difference is bounded below.

LEMMA 4.5.1. *If u_τ is defined such that $\text{P}(Y \geq u_\tau) = \tau$, and $u^* \neq u_\tau$ so that $\text{P}(Y \geq u^*) = \tau^*$, then*

$$\text{P}\left(\sum_{i=1}^{n} q\left(u_\tau; y_i\right) - \sum_{i=1}^{n} q\left(u^*; y_i\right) < k\right) \to 0 \tag{25}$$

for any finite $k > 0$.

PROOF. Note that

$$\sum_{i=1}^{n} q\left(u; y_i\right) = \sum_{i=1}^{n} \left[\tau y_i - \tau u - y_i \mathbb{I}_{y_i \geq u} + u \mathbb{I}_{y_i \geq u}\right],$$

so that

$$\begin{aligned} \sum_{i=1}^{n} q\left(u_\tau; y_i\right) - \sum_{i=1}^{n} q\left(u^*; y_i\right) &= n\left[\tau u^* - \tau u_\tau + \frac{1}{n} u_\tau \sum_{i=1}^{n} \mathbb{I}_{y_i \geq u_\tau} - \frac{1}{n} u^* \sum_{i=1}^{n} \mathbb{I}_{y_i \geq u^*} \right. \\ &\qquad \left. + \frac{1}{n} \sum_{i=1}^{n} y_i \left(\mathbb{I}_{y_i \geq u^*} - \mathbb{I}_{y_i \geq u_\tau}\right)\right] \\ &=: nH\left(\boldsymbol{y}\right). \end{aligned} \tag{26}$$

In order to show (25), it is enough to show that $H\left(\boldsymbol{y}\right) \xrightarrow{p} C$ such that $C > 0$.

Define $y^* = \inf\left\{y_i \in (\min(u_\tau, u^*), \max(u_\tau, u^*)]\right\}$. Then

$$\begin{aligned} H\left(\boldsymbol{y}\right) &\geq \tau u^* - \tau u_\tau + \frac{1}{n} u_\tau \sum_{i=1}^{n} \mathbb{I}_{y_i \geq u_\tau} - \frac{1}{n} u^* \sum_{i=1}^{n} \mathbb{I}_{y_i \geq u^*} \\ &\qquad + \frac{1}{n} y^* \sum_{i=1}^{n} \left(\mathbb{I}_{y_i \geq u^*} - \mathbb{I}_{y_i \geq u_\tau}\right) \\ &=: H^*\left(\boldsymbol{y}\right). \end{aligned} \tag{27}$$

Taking the limit of $H^*(\boldsymbol{y})$ as $n \to \infty$, then

$$\begin{aligned} H^*(\boldsymbol{y}) &\xrightarrow{p} \tau u^* - \tau u_\tau + u_\tau \,\mathrm{E}[\mathbb{I}_{y_i \geq u_\tau}] - u^* \,\mathrm{E}[\mathbb{I}_{y_i \geq u^*}] \\ &\qquad + y^* \left(\mathrm{E}[\mathbb{I}_{y_i \geq u^*}] - \mathrm{E}[\mathbb{I}_{y_i \geq u_\tau}]\right) \\ &= u^* (\tau - \tau^*) + y^* (\tau^* - \tau) \\ &= (y^* - u^*)(\tau^* - \tau), \end{aligned} \tag{28}$$

by LLN.

Note that expression (28) must be positive. If $u_\tau > u^*$, then it follows that $\tau < \tau^*$ and $y^* > u^*$. If $u_\tau < u^*$, then it follows that $\tau > \tau^*$ and $y^* < u^*$. Thus, I have shown that $H(\boldsymbol{y}) \xrightarrow{p} C$ where $C > 0$, and (25) is therefore true. □

Lemma 4.5.2 and Proposition 4.5.3 will be helpful in determining the behavior of the GPD_τ difference, which is given in Lemma 4.5.4.

LEMMA 4.5.2. *If $\hat{\sigma} \xrightarrow{p} \sigma$ and $\hat{\xi} \xrightarrow{p} \xi$, then as $n \to \infty$,*

$$\frac{1}{n} \sum_{i=1}^{n} \log g_\tau \left(u, \hat{\sigma}, \hat{\xi}; y_i\right) \xrightarrow{p} \mathrm{E}[\log g_\tau (u, \sigma, \xi; y_i)]. \tag{29}$$

PROOF. When $\xi \neq 0$, showing (29) is equivalent to showing

$$\frac{1}{n} \sum_{i=1}^{n} \left[\log \left(\frac{1}{\tau \hat{\sigma}} \left(\hat{\xi} \frac{y_i - u}{\hat{\sigma}} + \tau^{-\hat{\xi}} \right)^{-\frac{1}{\hat{\xi}} - 1} \right) - \log \left(\frac{1}{\tau \sigma} \left(\xi \frac{y_i - u}{\sigma} + \tau^{-\xi} \right)^{-\frac{1}{\xi} - 1} \right) \right] \to 0 \tag{30}$$

as $n \to \infty$, since

$$\frac{1}{n} \sum_{i=1}^{n} \log \left(\frac{1}{\tau \sigma} \left(\xi \frac{y_i - u}{\sigma} + \tau^{-\xi} \right)^{-\frac{1}{\xi} - 1} \right) \xrightarrow{p} \mathrm{E} \left[\log \left(\frac{1}{\tau \sigma} \left(\xi \frac{y_i - u}{\sigma} + \tau^{-\xi} \right)^{-\frac{1}{\xi} - 1} \right) \right]$$

by LLN.

First, note that

$$\log\left(\frac{1}{\tau\sigma}\left(\xi\frac{y_i-u}{\sigma}+\tau^{-\xi}\right)^{-\frac{1}{\xi}-1}\right)=\log\left(\frac{\tau^{\xi}}{\sigma}\right)+\left(-\frac{1}{\xi}-1\right)\log\left(\tau^{\xi}\xi\frac{y_i-u}{\sigma}+1\right).$$

Thus, the left hand side of (30) is equal to

$$\begin{aligned}\log\left(\frac{\tau^{\hat{\xi}}}{\hat{\sigma}}\right) &+ \left(-\frac{1}{\hat{\xi}}-1\right)\frac{1}{n}\sum_{i=1}^{n}\log\left(\tau^{\hat{\xi}}\hat{\xi}\frac{y_i-u}{\hat{\sigma}}+1\right)\\ &- \log\left(\frac{\tau^{\xi}}{\sigma}\right)-\left(-\frac{1}{\xi}-1\right)\frac{1}{n}\sum_{i=1}^{n}\log\left(\tau^{\xi}\xi\frac{y_i-u}{\sigma}+1\right). \qquad (31)\end{aligned}$$

Since

$$\log\left(\frac{\tau^{\hat{\xi}}}{\hat{\sigma}}\right)\xrightarrow{p}\log\left(\frac{\tau^{\xi}}{\sigma}\right)$$

and

$$-\frac{1}{\hat{\xi}}-1\xrightarrow{p}-\frac{1}{\xi}-1$$

by the continuous mapping theorem, then (31) converges to 0 as long as

$$\frac{1}{n}\sum_{i=1}^{n}\left[\log\left(\tau^{\hat{\xi}}\hat{\xi}\frac{y_i-u}{\hat{\sigma}}+1\right)-\log\left(\tau^{\xi}\xi\frac{y_i-u}{\sigma}+1\right)\right]\xrightarrow{p}0. \qquad (32)$$

Working with the term inside the summation of expression (32) gives

$$\begin{aligned}\log\left(\tau^{\hat{\xi}}\hat{\xi}\frac{y_i-u}{\hat{\sigma}}+1\right)-\log\left(\tau^{\xi}\xi\frac{y_i-u}{\sigma}+1\right) &= \log\left(\frac{\tau^{\hat{\xi}}\hat{\xi}\frac{y_i-u}{\hat{\sigma}}+1}{\tau^{\xi}\xi\frac{y_i-u}{\sigma}+1}\right)\\ &= \log\left(\frac{\tau^{\hat{\xi}}\hat{\xi}\frac{y_i-u}{\hat{\sigma}}-\tau^{\xi}\xi\frac{y_i-u}{\sigma}}{\tau^{\xi}\xi\frac{y_i-u}{\sigma}+1}+1\right). \qquad (33)\end{aligned}$$

Notice that (33) converges to 0 as long as

$$\tau^{\hat{\xi}}\hat{\xi}\frac{y_i - u}{\hat{\sigma}} - \tau^{\xi}\xi\frac{y_i - u}{\sigma} \xrightarrow{p} 0,$$

which is true by the continuous mapping theorem. Therefore, (32) is true, (31) must converge to 0, and thus (30) is true.

If $\hat{\xi} \xrightarrow{p} 0$, then I need to show

$$\frac{1}{n}\sum_{i=1}^{n}\left[\log\left(\tau^{\hat{\xi}}\hat{\xi}\frac{y_i - u}{\hat{\sigma}} + 1\right)^{-\frac{1}{\hat{\xi}}-1} + \frac{y_i - u}{\sigma}\right] \xrightarrow{p} 0 \tag{34}$$

in place of (32). Using the fact that

$$\lim_{\xi\to 0}\left(1 + \tau^{\xi}\xi\frac{y_i - u}{\sigma}\right)^{-\frac{1}{\xi}-1} = \exp\left(-\frac{y_i - u}{\sigma}\right),$$

and the continuous mapping theorem, (34) follows.

□

Proposition 4.5.3.

$$\mathrm{E}[\log g_\tau(u, \sigma, \xi; y_i)] = \log\left(\frac{\tau^{\xi}}{\sigma}\right) - \xi - 1 \tag{35}$$

Proof. Focusing on the case where $\xi \neq 0$, note that the expected value in (35) is equal to the integral

$$\int_u^{y^+} \log\left(\frac{1}{\tau\sigma}\left(\xi\frac{y - u}{\sigma} + \tau^{-\xi}\right)^{-\frac{1}{\xi}-1}\right)\frac{1}{\tau\sigma}\left(\xi\frac{y - u}{\sigma} + \tau^{-\xi}\right)^{-\frac{1}{\xi}-1} dy, \tag{36}$$

which can be solved directly. Expression (36) is equal to

$$\int_u^{y^+} \left[-\log(\tau\sigma) + \left(-\frac{1}{\xi} - 1\right) \log\left(\xi \frac{y-u}{\sigma} + \tau^{-\xi}\right)\right] \frac{1}{\tau\sigma} \left(\xi \frac{y-u}{\sigma} + \tau^{-\xi}\right)^{-\frac{1}{\xi}-1} dx. \tag{37}$$

Using u-substitution, let $v = \xi \frac{y-u}{\sigma} + \tau^{-\xi}$ so that $dv = \frac{\xi}{\sigma}$. Thus, solving equation (37) is equivalent to solving

$$\int_{\tau^{-\xi}}^{v^+} \left[-\log(\tau\sigma) + \left(-\frac{1}{\xi} - 1\right) \log v\right] \frac{1}{\tau\xi} v^{-\frac{1}{\xi}-1} dv. \tag{38}$$

Now, using integration by parts on expression (38), choose $w = -\log(\tau\sigma) + \left(-\frac{1}{\xi} - 1\right) \log v$ and $dx = \frac{1}{\tau\xi} v^{-\frac{1}{\xi}-1} dv$, so that $dw = \left(-\frac{1}{\xi} - 1\right) \frac{1}{v} dv$ and $x = -\frac{v^{-\frac{1}{\xi}}}{\tau}$. Then (38) is equivalent to

$$\left(-\log(\tau\sigma) + \left(-\frac{1}{\xi} - 1\right) \log v\right) \left(-v^{-\frac{1}{\xi}} \tau^{-1}\right) \Bigg|_{\tau^{-\xi}}^{v^+} - \int_{\tau^{-\xi}}^{v^+} \left(-\frac{1}{\xi} - 1\right) v^{-\frac{1}{\xi}-1} \tau^{-1} dv,$$

which integrates to

$$\left(-\log(\tau\sigma) + \left(-\frac{1}{\xi} - 1\right) \log v\right) \left(-v^{-\frac{1}{\xi}} \tau^{-1}\right) + v^{-\frac{1}{\xi}} \tau^{-1} (1+\xi) \Bigg|_{\tau^{-\xi}}^{v^+}. \tag{39}$$

The upper bound of v^+ here will vary depending on whether $\xi > 0$ or $\xi < 0$.

First assume that $\xi > 0$. This indicates that $v^+ = \infty$. Solving for (39) gives

$$0 - \left[\left(-\log(\tau\sigma) + \left(-\frac{1}{\xi} - 1\right) \log\left(\tau^{-\xi}\right)\right) (-1) + 1 + \xi \right] = \log\left(\frac{\tau^\xi}{\sigma}\right) - \xi - 1.$$

Thus, if $\xi > 0$, then (35) is true.

Now assume that $\xi < 0$, so that $v^+ = 0$. First, adjust (39) to keep track of the negative shape. By letting $k = -\xi$ so that $k > 0$, (39) becomes

$$\left(-\log(\tau\sigma) + \left(\frac{1}{k} - 1\right)\log v\right)\left(-v^{\frac{1}{k}}\tau^{-1}\right) + v^{\frac{1}{k}-1}\tau^{-1}(1-k)\Big|_{\tau^k}^{0}, \tag{40}$$

and solving for (40) gives

$$\begin{aligned} 0 - \left[\left(-\log(\tau\sigma) + \left(\frac{1}{k} - 1\right)\log\left(\tau^k\right)\right)(-1) + 1 - k\right] &= \log\left(\frac{1}{\sigma\tau^k}\right) + k - 1 \\ &= \log\left(\frac{\tau^\xi}{\sigma}\right) - \xi - 1. \end{aligned}$$

Thus, (35) is also true when $\xi < 0$.

If $\xi = 0$, a similar process will show

$$\int_u^\infty \log\left(\frac{1}{\sigma}\exp\left\{-\frac{y-u}{\sigma}\right\}\right)\frac{1}{\sigma}\exp\left\{-\frac{y-u}{\sigma}\right\}dy = \log\left(\frac{1}{\sigma}\right) - 1.$$

□

LEMMA 4.5.4. *Let u_τ be defined such that $\mathrm{P}(Y \geq u_\tau) = \tau$, and $u^* \neq u_\tau$ such that $\mathrm{P}(Y \geq u^*) = \tau^*$. Let $(\hat{\sigma}, \hat{\xi}) = \mathrm{argmax}_{(\sigma,\xi)} M_n(u_\tau, \sigma, \xi)$ such that $(\hat{\sigma}, \hat{\xi}) \xrightarrow{p} (\sigma, \xi)$, and let $(\hat{\sigma}^*, \hat{\xi}^*) = \mathrm{argmax}_{(\sigma,\xi)} M_n(u^*, \sigma, \xi)$ such that $(\hat{\sigma}^*, \hat{\xi}^*) \xrightarrow{p} (\sigma^*, \xi^*)$. Then $\exists k > -\infty$ such that as $n \to \infty$,*

$$\mathrm{P}\left(\frac{1}{N}\sum_{i=1}^n \log g_\tau\left(u_\tau, \hat{\sigma}, \hat{\xi}; y_i\right)\mathbb{I}_{y_i \geq u_\tau} - \frac{1}{N^*}\sum_{i=1}^n \log g_\tau\left(u^*, \hat{\sigma}^*, \hat{\xi}^*; y_i\right)\mathbb{I}_{y_i \geq u^*} < k\right) \to 0, \tag{41}$$

for $N = \sum_{i=1}^n \mathbb{I}_{y_i \geq u_\tau}$ and $N^* = \sum_{i=1}^n \mathbb{I}_{y_i \geq u^*}$.

PROOF. Define $z_i = [y_i | y_i \geq u_\tau]$ and $z_i^* = [y_i | y_i \geq u^*]$. Then

$$\frac{1}{N}\sum_{i=1}^{n} \log g_\tau\left(u_\tau, \hat{\sigma}, \hat{\xi}; y_i\right) \mathbb{I}_{y_i \geq u_\tau} = \frac{1}{N}\sum_{i=1}^{N} \log g_\tau\left(u_\tau, \hat{\sigma}, \hat{\xi}; z_i\right).$$

Likewise,

$$\frac{1}{N^*}\sum_{i=1}^{n} \log g_\tau\left(u^*, \hat{\sigma}^*, \hat{\xi}^*; y_i\right) \mathbb{I}_{y_i \geq u^*} = \frac{1}{N^*}\sum_{i=1}^{N^*} \log g_\tau\left(u^*, \hat{\sigma}^*, \hat{\xi}^*; z_i^*\right).$$

Recognize that as $n \to \infty$, both N and $N^* \to \infty$. Lemma 4.5.2 says that

$$\frac{1}{N}\sum_{i=1}^{N} \log g_\tau\left(u_\tau, \hat{\sigma}, \hat{\xi}; z_i\right) \to \mathrm{E}[\log g_\tau\left(u_\tau, \sigma, \xi; z_i\right)]$$

and

$$\frac{1}{N^*}\sum_{i=1}^{N^*} \log g_\tau\left(u^*, \hat{\sigma}^*, \hat{\xi}^*; z_i^*\right) \to \mathrm{E}[\log g_\tau\left(u^*, \sigma^*, \xi^*; z_i^*\right)].$$

Thus, using Proposition 4.5.3,

$$\begin{aligned} \frac{1}{N}\sum_{i=1}^{N} \log g_\tau\left(u_\tau, \hat{\sigma}, \hat{\xi}; z_i\right) &- \frac{1}{N^*}\sum_{i=1}^{N^*} \log g_\tau\left(u^*, \hat{\sigma}^*, \hat{\xi}^*; z_i^*\right) \\ &\xrightarrow{p} \log\left(\frac{\tau^\xi}{\sigma}\right) - \xi - 1 - \left(\log\left(\frac{\tau^{\xi^*}}{\sigma^*}\right) - \xi^* - 1\right) \\ &= \log\left(\tau^{\xi-\xi^*}\frac{\sigma^*}{\sigma}\right) - (\xi - \xi^*). \end{aligned} \tag{42}$$

Choose $k < \log\left(\tau^{\xi-\xi^*}\frac{\sigma^*}{\sigma}\right) - (\xi - \xi^*)$. Then (41) holds. □

I can now show that (24) must be true.

THEOREM 4.5.5. *Let $u_\tau, u^*, \hat{\sigma}, \hat{\xi}, \hat{\sigma}^*$, and $\hat{\xi}^*$ be defined as in Lemma 4.5.4. Then as $n \to \infty$,*

$$\mathrm{P}\Big(M_n\left(u_\tau, \hat{\sigma}, \hat{\xi}; \boldsymbol{y}\right) - M_n\left(u^*, \hat{\sigma}^*, \hat{\xi}^*; \boldsymbol{y}\right) < 0\Big) \to 0. \tag{43}$$

PROOF. Let $k > -\infty$. Then

$$\begin{aligned}
\mathrm{P}\Big(M_n\left(u_\tau, \hat{\sigma}, \hat{\xi}; \boldsymbol{y}\right) &- M_n\left(u^*, \hat{\sigma}^*, \hat{\xi}^*; \boldsymbol{y}\right) < 0\Big) \\
&\leq \mathrm{P}\left(\sum_{i=1}^{n} q\left(u_\tau; y_i\right) - \sum_{i=1}^{n} q\left(u^*; y_i\right) < k\right) \qquad (44) \\
&\quad + \mathrm{P}\left(\frac{1}{n}\sum_{i=1}^{n} \log g_\tau\left(u_\tau, \hat{\sigma}, \hat{\xi}; y_i\right) \mathbb{I}_{y_i \geq u_\tau}\right. \\
&\quad \left. - \frac{1}{n}\sum_{i=1}^{n} \log g_\tau\left(u^*, \hat{\sigma}^*, \hat{\xi}^*; y_i\right) \mathbb{I}_{y_i \geq u^*} < -k\right). \qquad (45)
\end{aligned}$$

Lemma 4.5.1 says that (44) converges to 0. Lemma 4.5.4 says that (45) converges to 0. Therefore, (43) is true. □

Notice that Theorem 4.5.5 uses a specific u^*. I wish to replace u^* with $\hat{u}_n$ as defined in (23). Cases where $\hat{u}_n$ are more than a specific distance away from u_τ will be considered.

LEMMA 4.5.6. *Let $\hat{u}_n = \mathrm{argmax}_{u \in B_\delta^c} M_n\left(u, \hat{\sigma}, \hat{\xi}; \boldsymbol{y}\right)$, where $B_\delta = (u_\tau - \delta, u_\tau + \delta)$ for a fixed $\delta > 0$. Then*

$$\mathrm{P}\Big(M_n\left(u_\tau, \hat{\sigma}, \hat{\xi}; \boldsymbol{y}\right) - M_n\left(\hat{u}_n, \hat{\sigma}_{\hat{u}_n}, \hat{\xi}_{\hat{u}_n}; \boldsymbol{y}\right) < 0\Big) \to 0 \tag{46}$$

as $n \to \infty$.

PROOF. By Theorem 4.5.5, it is known that as $n \to \infty$,

$$\mathrm{P}\Big(M_n\left(u_\tau, \hat{\sigma}, \hat{\xi}; \boldsymbol{y}\right) - M_n\left(u^*, \hat{\sigma}_{u^*}, \hat{\xi}_{u^*}; \boldsymbol{y}\right) < 0\Big) \to 0$$

for a fixed $u^* \in B_\delta^c$. This implies that for any $\epsilon > 0$ there exists a n_ϵ such that if $n > n_\epsilon$,

$$\mathrm{P}\Big(M_n\left(u_\tau, \hat{\sigma}, \hat{\xi}; \boldsymbol{y}\right) - M_n\left(u^*, \hat{\sigma}_{u^*}, \hat{\xi}_{u^*}; \boldsymbol{y}\right) < 0\Big) < \epsilon.$$

Notice that n_ϵ is tied to the specific u^* selected. Call this $n_\epsilon(u^*)$, and define

$$n_{\epsilon, B_\delta^c} = \sup_{u \in B_\delta^c} n_\epsilon(u).$$

Then if $n > n_{\epsilon, B_\delta^c}$,

$$\mathrm{P}\Big(M_n\left(u_\tau, \hat{\sigma}, \hat{\xi}; \boldsymbol{y}\right) - M_n\left(\hat{u}_n, \hat{\sigma}_{\hat{u}_n}, \hat{\xi}_{\hat{u}_n}; \boldsymbol{y}\right) < 0\Big) < \epsilon,$$

$\forall \epsilon > 0$. Therefore, (46) holds. □

I can now establish consistency.

THEOREM 4.5.7. *Let $\hat{u}_n$ be defined as in* (23)*. Then $\hat{u}_n$ is a consistent estimator of u_τ.*

PROOF. Assume that for $\delta > 0$ and $c \neq 0$,

$$\mathrm{P}(|u_\tau - \hat{u}_n| > \delta) \to c.$$

This implies that there exists a $\gamma > 0$ such that every $n > 0$ has a $n^* > n$ such that

$$\mathrm{P}(|u_\tau - \hat{u}_{n^*}| > \delta) > \gamma.$$

If the event '$|u_\tau - \hat{u}_n| > \delta$' occurs, then $\hat{u}_n \in B_\delta^c$ and

$$M_n\left(\hat{u}_n, \hat{\sigma}_{\hat{u}_n}, \hat{\xi}_{\hat{u}_n}; \boldsymbol{y}\right) \geq M_n\left(u_\tau, \hat{\sigma}, \hat{\xi}; \boldsymbol{y}\right),$$

where $B_\delta = (u_\tau - \delta, u_\tau + \delta)$. So $\mathrm{P}(|u_\tau - \hat{u}_n| > \delta) > \gamma$ implies

$$\mathrm{P}\Big(\Big\{M_n\left(u_\tau, \hat{\sigma}, \hat{\xi}; \boldsymbol{y}\right) - M_n\left(\hat{u}_n, \hat{\sigma}_{\hat{u}_n}, \hat{\xi}_{\hat{u}_n}; \boldsymbol{y}\right) < 0\Big\} \bigcap \{\hat{u}_n \in B_\delta^c\}\Big) > \gamma. \tag{47}$$

However, (47) is a contradiction by Lemma 4.5.6, since

$$\begin{aligned}
\mathrm{P}\Big(\Big\{M_n\left(u_\tau, \hat{\sigma}, \hat{\xi}; \boldsymbol{y}\right) \quad - \quad & M_n\left(\hat{u}_n, \hat{\sigma}_{\hat{u}_n}, \hat{\xi}_{\hat{u}_n}; \boldsymbol{y}\right) < 0\Big\} \bigcap \{\hat{u}_n \in B_\delta^c\}\Big) \\
& \leq \mathrm{P}\Big(M_n\left(u_\tau, \hat{\sigma}, \hat{\xi}; \boldsymbol{y}\right) - M_n\left(\hat{u}_n, \hat{\sigma}_{\hat{u}_n}, \hat{\xi}_{\hat{u}_n}; \boldsymbol{y}\right) < 0\Big) < \epsilon,
\end{aligned}$$

for all $\epsilon > 0$. Thus, for all $\gamma > 0$,

$$\mathrm{P}(|u_\tau - \hat{u}_{n^*}| > \delta) < \gamma$$

as $n \to \infty$, and $\hat{u}_n$ is therefore a consistent estimator of u_τ. □

CHAPTER 5

Practical Optimization Considerations

Recall that I am interested in tying both the threshold u_τ and the GPD$_\tau$ parameters to covariates. Chapter 4 demonstrates that optimizing the objective function M_n defined in equation (21) leads to consistent estimates for u_τ, σ, and ξ in simple cases. Adding covariates makes things more difficult. This chapter will discuss some practical modifications to (21) to allow for the implementation of covariates. Further modifications to improve estimation are also discussed.

5.1. Implementing Covariates

Obtaining estimates for u_τ, σ, and ξ via the objective function in (21) requires numerical optimization. Allowing these three parameters to be functions of covariates (as in GLM settings) complicates the optimization. Furthermore, the objective function's unique treatment of the data set leads to unique optimization issues. Specifically, since the number of exceedances used to estimate the GPD$_\tau$ parameters changes with u_τ, treating u_τ as a parametric function of covariates adds a new layer of complexity to the model. As a result, covariate implementation must be carefully considered, with special consideration given to continuous covariates.

5.1.1. A Grid Search Method. The grid search method described in this section was developed during my investigation of equation (18), which uses the binomial probability mass function in place of the quantile regression piece of (21). The motivation for the grid search method is that the objective function has a discontinuous jump whenever u_τ, varying continuously, 'passes over' an observed data value, leading to its inclusion or exclusion in

the GPD piece of (18). This jump is exacerbated with the objective function which uses the binomial rather than the quantile regression piece, but still remains in the quantile regression formulation. The grid search method was only implemented in the case of discrete covariates. Its applicability in the presence of continuous covariates is dubious, as will be discussed.

For a given data set, let $Y_{(i)}$ and $Y_{(i+1)}$ be the ith and $(i+1)$th largest order statistics. Consider optimizing M_n from (18) for values of $u_\tau \in \left(Y_{(i)}, Y_{(i+1)}\right]$. It can be shown that the maximum value of M_n is attained when $u_\tau = Y_{(i+1)}$ (see Appendix A for proof). This allows for an optimization scheme based on a grid search method which can be used to implement categorical covariates.

Since the set of possible values for u_τ is finite, a grid search optimization scheme need only consider u_τ corresponding to observations in the given data set. By optimizing the shape and scale parameters at each possible u_τ, the set of parameters that maximize M_n are chosen as the estimates. Bivariate categorical covariates are implemented seamlessly by breaking the data into two separate sets. If $u_\tau = \beta_0 + \beta_1 X$, then the optimization scheme finds two values of u_τ: one for $X = 0$ and one for $X = 1$. These are used to determine the values of β_0 and β_1.

The implementation described can also be extended to allow multiple bivariate categorical covariates. If k covariates are used, then the algorithm searches a $k+1$ dimensional grid. Furthermore, categorical covariates with j categories are included by treating these as $j-1$ bivariate categorical variables, as is often done in regression settings.

While the grid search method can be unwieldy, especially if the sample size is large, practically limiting the observations considered as possible u_τ values to be within a certain range of the sample quantiles can hasten computation speed. Initial simulation studies using

this grid search optimization scheme suggest that it is a reasonable method of estimation if only categorical covariates are used.

An attempt to extend the grid search method to the most simple continuous covariate setting demonstrates its shortcomings. Assume that M_n is maximized if $u_\tau(y)$ passes through a data point. The logical extension of the grid search method is to fix a data point and then optimize β_0 and β_1 (in addition to the GPD_τ parameters) subject to u_τ passing through that data point. The process would be repeated for all sensible data points until one is convinced the maximum is achieved. Performing the repeated optimization that this procedure would require is clearly tedious, with multiple covariates compounding this issue. I chose not to proceed in this manner, seeking a more elegant and tractable estimation method instead.

5.1.2. A Kernel Smoothing Method. The grid search method was inspired by plots of the profile objective function of u_τ, which have 'sharktooth' appearances, an example of which is shown in Figure 5.1. I realized that the discontinuous jumps occurred at observations in the data set, which led to the proof in Appendix A. However, optimization is usually improved by smooth functions, so it is desirable to introduce smoothness into the objective function if the grid based method is not used. Use of the quantile regression form of the objective function, as in (21), works to smooth the profile objective function, but discontinuous jumps remain.

Instead of treating each observation as a unitary mass at a point, a kernel density, centered at each observation, is used to introduce a weight into the objective function. The weight corresponds to mass of the kernel which exceeds the threshold. Whereas exceedances and non-exceedances were previously given respective weights of 1 and 0, now if the value of u_τ

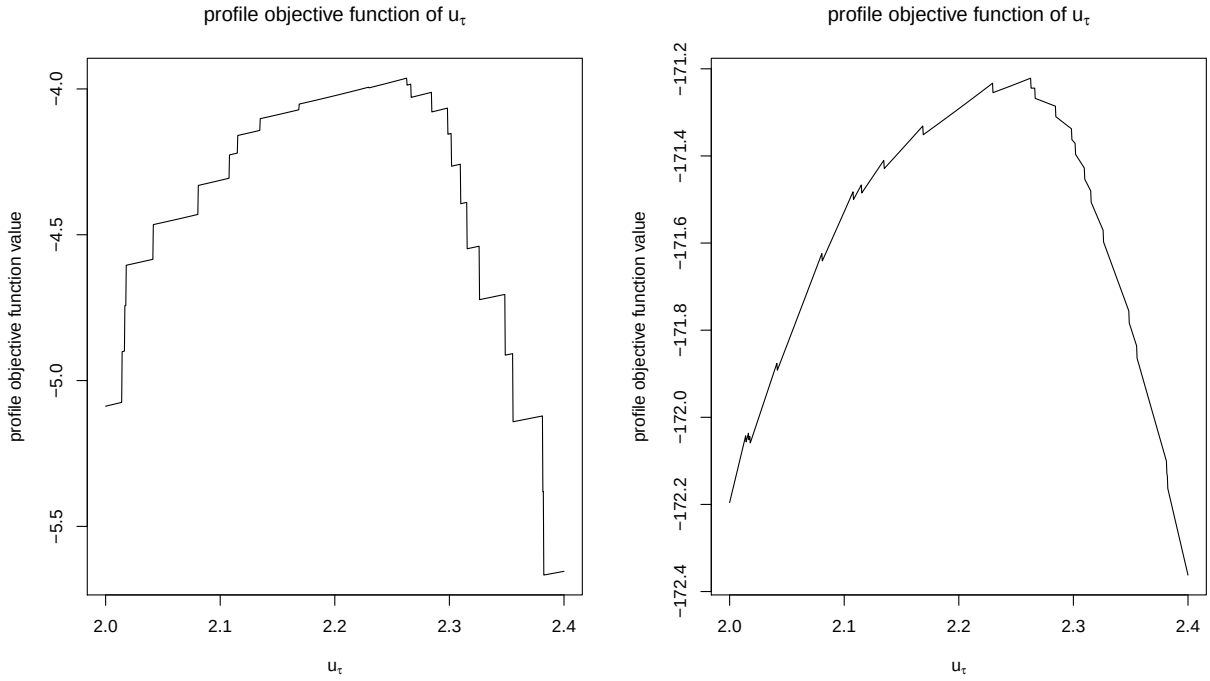

FIGURE 5.1. Example of the 'sharktooth' behavior exhibited by the profile objective function of u_τ. The first plot uses the objective function in (18), whereas the second uses the objective function in (21).

increases across an observation's value, that observation's contribution to M_n will smoothly vary from 1 down to 0.

An isotropic kernel density with finite support is used. Denote δ to be the radius of the kernel. Observations which exceed $u_\tau - \delta$ will contribute to the generalized Pareto portion of the objective function, which must be adjusted slightly to account for this change. Using the threshold stability property of GPD_τ (and assuming this holds for values above $u_\tau - \delta$), one can show $\tau_\delta = \left(\tau^{-\xi} - \xi\delta/\sigma\right)^{-1/\xi}$. Thus, u_τ can still be estimated, despite using observations that exceed $u_\tau - \delta$ in fitting GPD_τ.

The objective function with kernel density smoothing implemented is

$$M_n\left(u_\tau,\sigma,\xi;\boldsymbol{y}\right)=\begin{cases}\sum_{i=1}^{n}q\left(u_\tau;y_i\right)+\frac{1}{\sum_{i=1}^{n}w_i}\sum_{i=1}^{n}w_i\log g_\tau\left(u_\tau,\sigma,\xi;y_i\right), & \xi\neq 0\\ \sum_{i=1}^{n}q\left(u_\tau;y_i\right)+\frac{1}{\sum_{i=1}^{n}w_i}\sum_{i=1}^{n}w_i\log g\left(u_\tau,\sigma,\xi;y_i\right), & \xi=0\end{cases}, \tag{48}$$

where w_i are weights. These are defined such that $w_i = P\left(X_i > u_\tau\right)$ for $X_i \sim k_h\left(y_i\right)$ where $k_h\left(y_i\right)$ is the kernel density of the ith observation with bandwidth h. Note that only observations exceeding $u_\tau - \frac{h}{2}$ will contribute to the GPD_τ portion of (48), as all other observations will have weights of 0.

Some consideration must be given to the kernel bandwidth. Kernel densities are traditionally used to estimate probability density functions. The bandwidth controls the smoothness of the density estimate, with larger bandwidths yielding smoother densities (Givens and Hoeting, 2012, Chapter 10.2.2). The bandwidth is often used to balance the bias-variance trade-off in kernel density estimators: if bandwidth is too small, variance may be high, but if bandwidth is too large, estimates may be biased due to oversmoothing. While oversmoothing is not likely a concern in my approach, the bandwidth is still used to balance a trade-off: a wider bandwidth introduces more smoothness aiding the optimization, but too wide a bandwidth could introduce bias for estimates of σ and ξ as information about the tail becomes contaminated by observations in the bulk. Sensitivity analysis performed on bandwidth selection suggested that decreasing the bandwidth does not jeopardize optimization in my approach. Analysis in Chapters 6 and 7 therefore use small bandwidths.

In both the simulation study (Chapter 6) and the facial recognition application (Chapter 7), a uniform kernel is used. This kernel is chosen for simplicity. Simulations run using both

biweight and Epanechnikov kernels suggest that the choice of kernel is of relatively little importance as long as its support is finite.

This kernel-based weighting scheme further allows for the implementation of continuous covariates, as the choice of threshold no longer corresponds to one of the observations. Thus, u_τ, σ, and ξ in (48) may be parametric functions of covariates, both categorical and continuous, as in generalized linear modeling.

5.2. Avoiding Unreasonable Shape Estimates

It is well known that numerical maximum likelihood can produce bad estimates for ξ when sample size is small (Coles and Dixon, 1999). As my M-estimation method also requires numerical optimization, similar difficulties can arise. Both Coles and Dixon (1999) and Martins and Stedinger (2000) advocate penalized likelihood approaches which enforce ξ to take on reasonable values. Similar to Martins and Stedinger (2000), I construct a penalty via a shifted beta distribution centered at 0, which restricts the shape parameter to values in $[-0.5, 0.5]$. It is reasonable to assume that ξ is in this interval. If the shape is less than -0.5, the tail is not only finite, but the density evaluated at the upper endpoint exceeds 0, which would not mimic the behavior of the distribution of non-match scores. If the shape is greater than 0.5, then the distribution does not have a finite variance. Many application areas (such as the natural sciences) restrict ξ so that $-0.5 < \xi < 0.5$, and thus assume a finite second moment; both Hosking and Wallis (1987) and Coles and Dixon (1999) use such a restriction for practicality reasons. I am comfortable making an initial assumption that the non-match distribution has a finite second moment, but the behavior of ξ will be checked

in an exploratory analysis of the data. The shifted beta's log-density is

$$p(\xi) = \log\left(\frac{(0.5+\xi)^{\alpha-1}(0.5-\xi)^{\beta-1}}{B(\alpha,\beta)}\right), \tag{49}$$

where $B(\alpha,\beta)$ denotes the beta function. Throughout this study I set $\alpha = 2$ and $\beta = 2$ yielding a moderately peaked symmetric density about 0. This restricts the support of ξ to values between -0.5 and 0.5, but it also nudges all estimates slightly towards 0. Tuning α and β to specific cases may improve performance. For example, a left skewed penalty is likely preferred if the non-match distribution has a heavy tail.

In the penalized likelihood setting, a penalty such as the one in (49) is added onto the log-likelihood. Because log-likelihood's magnitude increases with sample size and the penalty does not, the influence of the penalty on the estimate of ξ decreases with sample size. With the objective function defined in (21), since the magnitude of the "likelihood" piece does not increase with sample size, a penalty whose influence will decrease with sample size is imposed. The penalized objective function is

$$M_n(u_\tau,\sigma,\xi;\boldsymbol{y}) = \sum_{i=1}^{n} q(u_\tau;y_i) + \frac{1}{\sum_{i=1}^{n} w_i} p(\xi) + \frac{1}{\sum_{i=1}^{n} w_i}\sum_{i=1}^{n} w_i \log g_\tau(u_\tau,\sigma,\xi;y_i). \tag{50}$$

While adding the penalty term will prove useful for the simulation study in Section 6, it will have very little influence on the results of the application. Consequently, as $n \to \infty$, the estimate of ξ from the penalized objective function approaches the estimate of ξ from the unpenalized objective function.

5.3. A Practical Optimization Scheme

Despite efforts to smooth the objective function, it is still possible that optimization performs poorly. These minor adjustments in optimization scheme will work to improve the provided parameter estimates.

5.3.1. Gauss-Seidel Iterization. The parameter u_τ appears in both the quantile regression and GPD_τ pieces of the objective function. Due to the fact that the quantile regression piece grows with n and the GPD_τ piece converges to a value, the quantile regression exerts far more influence on the estimate of u_τ (by design). However the imbalance in the magnitudes of the two pieces can lead to poor shape and scale estimates if the optimization scheme updates the three parameters all-at-once. In order to counteract this, I employ a non-linear Gauss-Seidel iterization (Givens and Hoeting, 2012, Section 2.2.5). Each iteration of the optimization has two steps. The first step optimizes the threshold parameter(s), whereas the second step optimizes the GPD parameters.

5.3.2. Reasonable Starting Values. Satisfactory performance of the numerical optimizer requires reasonable starting values, which merits some consideration. Threshold parameters are set using a simple quantile regression fit. Initial values for the shape and scale parameters are the GPD_τ equivalents to those used in the ismev package in R (Heffernan and Stephenson, 2012).

CHAPTER 6

A Simulation Study

In this chapter, a simulation study is performed in order to test the performance of my model in fitting the tail fo a distribution. 95% bootstrap confidence intervals for parameters and select quantiles are calculated. Estimated quantiles are compared to their true values, and my method's ability to estimate high quantiles is compared to standard quantile regression.

6.1. Generating Model and Bootstrapping Procedure

Monte Carlo data sets each with $n = 5000$ observations Y were generated according to the formula

$$Y = 10 + 5X_1 + 20X_2 + \exp(1 + 0.02X_1)\, T_4, \tag{51}$$

where X_1 is a continuous variable with values from 20 to 60, X_2 is binary, and T_4 is a t-distributed random variable with four degrees of freedom. The first three terms of the equation will effect the threshold, whereas the terms inside the exponential function will effect both the threshold and scale.

Using a kernel density bandwidth of 0.01, I fit a model that includes the continuous and categorical covariates in both the threshold and scale, such that $u_\tau = \beta_0 + \beta_1 X_1 + \beta_2 X_2$ and $\sigma = \exp(\gamma_0 + \gamma_1 X_1 + \gamma_2 X_2)$. Importantly, while the fitted model captures the behavior of the generating equation (51), it does not correspond exactly. For example, the true u_τ resulting from (51) is not linear. Also note that I fit a scale parameter using the categorical

variable even though it does not appear in the scaling term applied to the t-distributed random variable.

To obtain confidence intervals for both parameter estimates and estimated high quantiles, a semiparametric paired bootstrap is used. The procedure is as follows, where $(x_i, y_i), i = 1, \ldots, n$ denotes independent observations from (51):

(1) Resample with replacement from $\{(x_i, y_i), i = 1, \ldots, n\}$. Denote these resampled realizations as $(x_i^*, y_i^*), i = 1, \ldots n$.

(2) If $y_i^* \leq u_\tau(x_i^*)$, then $y_i^{**} = y_i^*$.

(3) If $y_i^* > u_\tau(x_i^*)$, then let y_i^{**} be drawn from a GPD_τ with fixed parameter values $\hat{\boldsymbol{\beta}}, \hat{\boldsymbol{\gamma}}$,and $\hat{\xi}$, and covariate value x_i^*, where $\hat{\boldsymbol{\beta}} = (\hat{\beta}_0, \hat{\beta}_1, \hat{\beta}_2)^T$ and $\hat{\boldsymbol{\gamma}} = (\hat{\gamma}_0, \hat{\gamma}_1, \hat{\gamma}_2)^T$.

(4) The model is fitted to the (x_i^*, y_i^{**}) realizations.

Use of this semiparametric bootstrap process eliminates ties in the tail of the resampled data set, improving the representation of the tail once the data set is fitted to my model.

Because optimization is computationally expensive, this process is performed on the CSU ISTeC Cray HPC System, a cluster computing environment composed of nodes each with 32 CPU cores and dedicated memory allocation. The computational process was distributed by running each Monte Carlo iteration and its bootstrap on an individual core, with 24 instances run on each node to prevent exceeding memory limits. The Cray performed the process on each node in under 24 hours, and the system's queuing system allowed for the use of up to four nodes at one time for a process of its length. Ultimately, I generated 504 Monte Carlo data sets with corresponding bootstraps.

6.2. Results

Figure 6.1 helps to illustrate the performance of the fitted model with regards to two separate simulated data sets. These two instances were chosen because they reflect a good range of observed fits. Shown are both the true and fitted 0.95 and 0.999 quantiles. The top panels show an instance where the fitted model mimics the truth quite well, whereas the bottom panels show some differences but still seem to capture the overall behavior reasonably well.

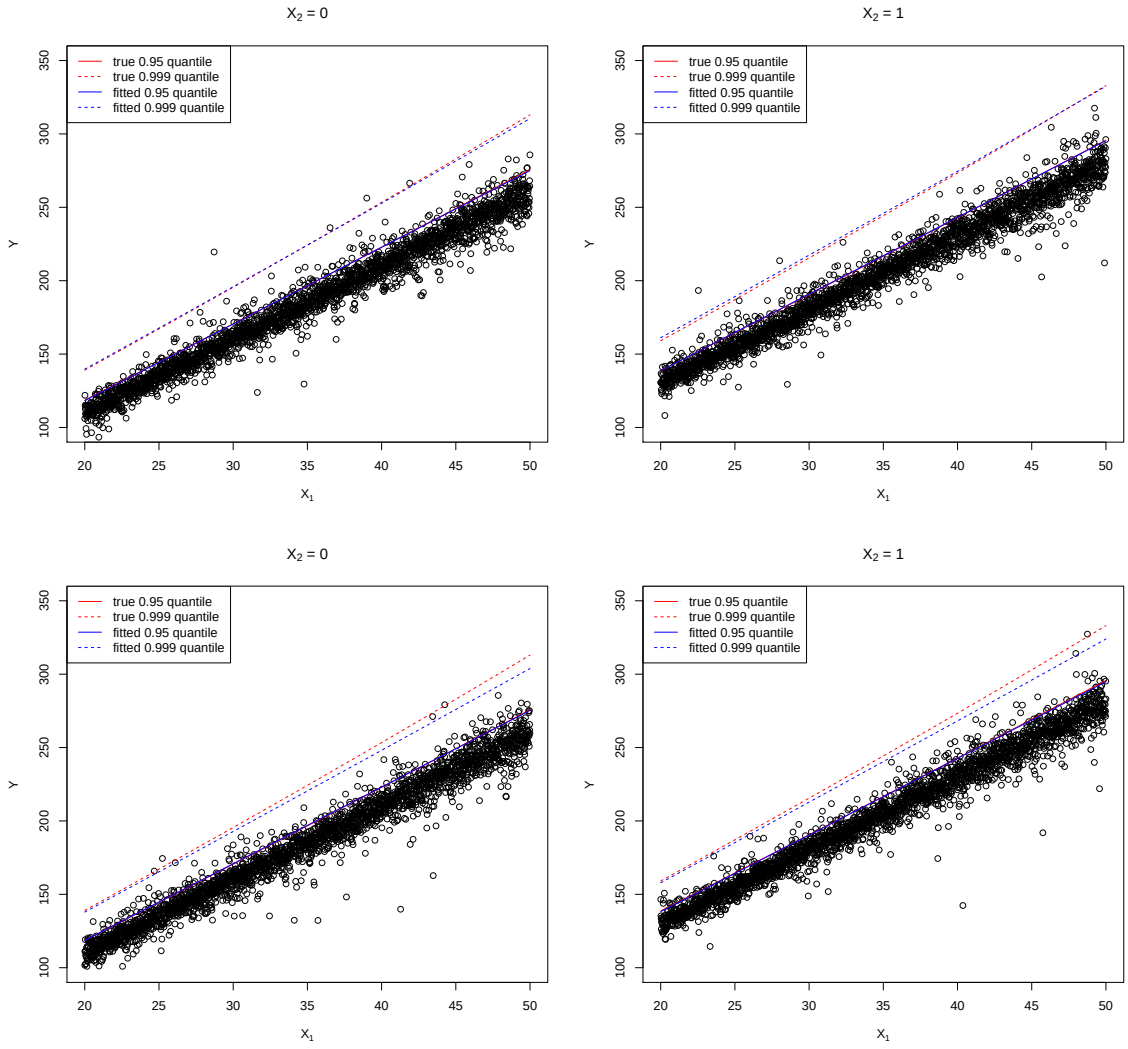

Figure 6.1. Fitted and true quantile against Monte Carlo generated data set 1 (top) and data set 2 (bottom).

Figure 6.2 shows histograms for the parameter estimates from the 504 Monte Carlo data sets. The top row shows estimates for the threshold parameters. Due to the mismatch between the generating equation and the fitted model, the estimates for $\boldsymbol{\beta}$ are not centered at the values in (51). Despite the mismatch, the threshold parameters remain very interpretable. Estimates of β_1 are slightly larger than 5, implying that the threshold grows at approximately this rate with a per unit increase in the continuous covariate X_1. Estimates of β_2 are approximately 20, also indicating the effect the binary covariate has on u_τ. The middle row of Figure 6.2 shows the histograms for the scale parameter estimates. The positive estimates for γ_1 show that the fitted model recognizes that scale increases with X_1. The estimates for γ_2 are properly centered about 0. The bottom panel of Figure 6.2 shows estimates of the shape parameter. The true shape for a GPD fit to the tail of a t-distribution with 4 degrees of freedom is 0.25. However, this parameter value is achieved as the sample size increases to infinity, and finite-sample estimates for ξ for a t-distribution tend to be lower than the asymptotic value. Additionally, the penalty could slightly nudge shape estimates toward 0.

In contrast to the model parameter estimates which cannot be compared to true values due to the mismatch between generating and fitted models, the estimated quantiles can be compared to the true quantiles for specified covariate values. Histograms for five quantiles of interest are given for two specific sets of covariates in Figure 6.3. The first set uses $X_1 = 27.5$ and $X_2 = 1$, whereas the second set uses $X_1 = 42.5$ and $X_2 = 0$. The line on each histogram indicates where the true quantile is located. Overall, the performance of my model in predicting the quantiles appears to be quite good. The estimates are relatively unbiased and roughly normally distributed. While some bias appears at the 0.9999 quantile, this is likely due to the underestimation of the shape parameter ξ. Only five observations

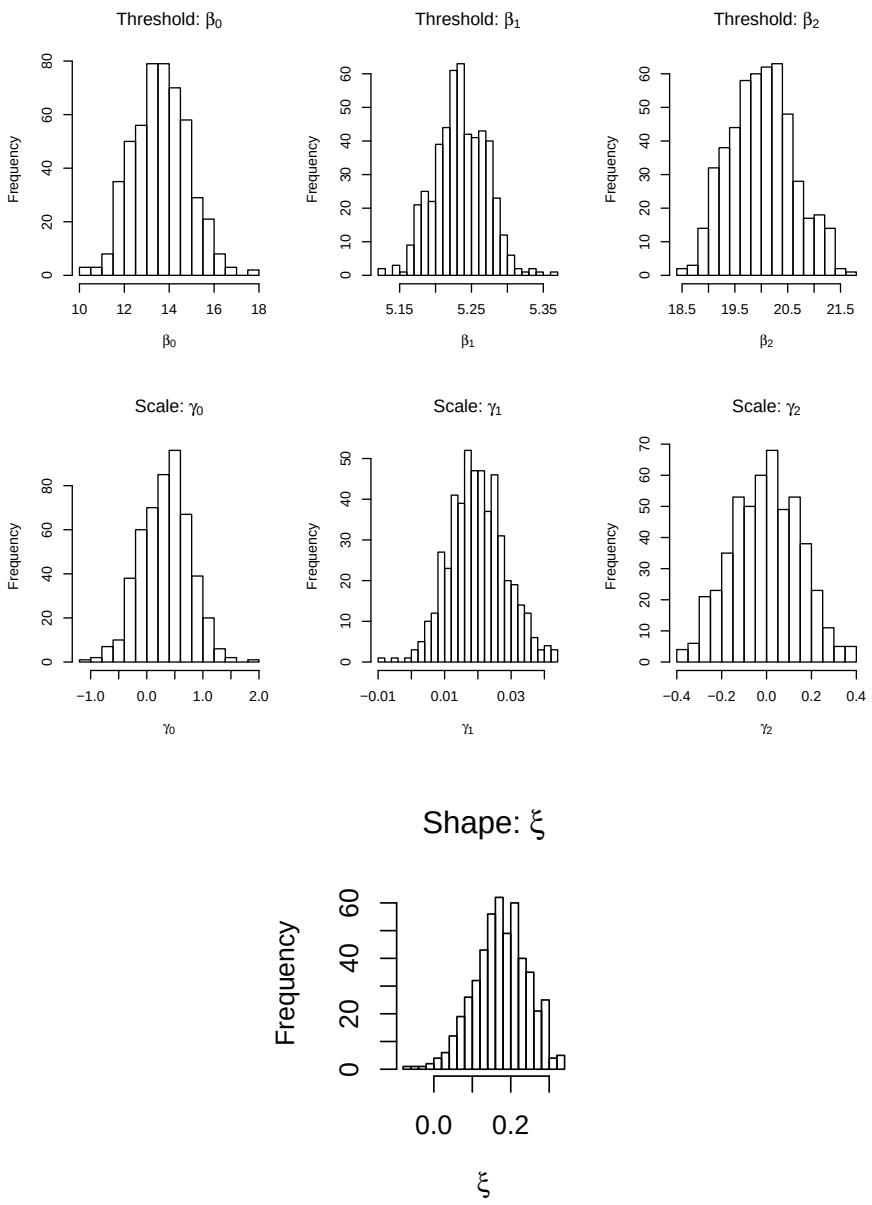

FIGURE 6.2. Histograms of threshold (top), scale (middle), and shape (bottom) parameter estimates from 504 Monte Carlo simulations.

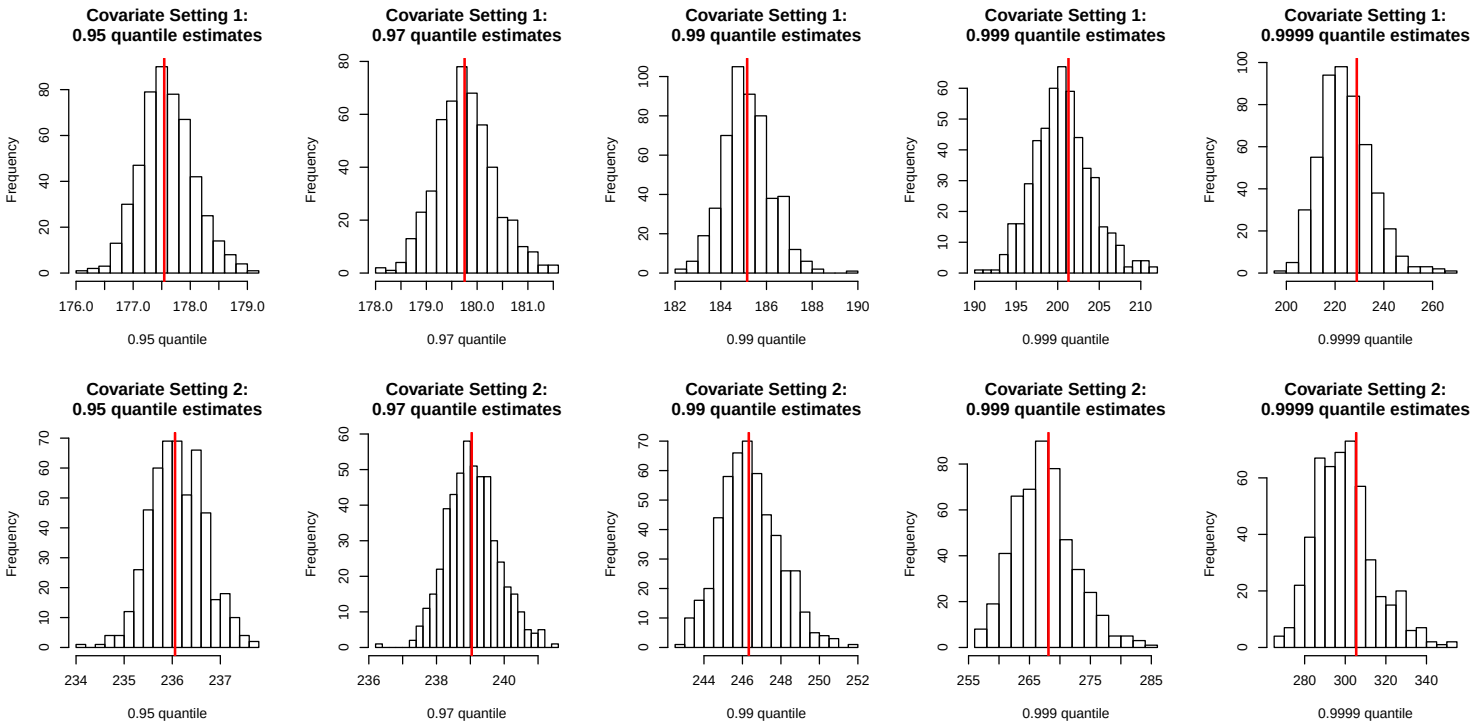

FIGURE 6.3. Histograms of quantile estimates from 504 Monte Carlo simulations evaluated for two covariate settings. The top row corresponds to $X_1 = 27.5$ and $X_2 = 1$ and the bottom row corresponds to $X_1 = 42.5$ and $X_2 = 0$. The vertical lines indicate the location of the true quantile.

are expected to occur above the 0.999 quantile for a data set of size $n = 5000$, so this performance is reasonable.

It is also worth assessing the bootstrap method's ability to accurately account for estimation uncertainty. Table 6.1 shows the parameter estimates along with 95% bootstrap confidence intervals for the data set illustrated in the top panels of Figure 6.1. While coverage cannot be assessed due to the mismatch between generating and fitted models, notice that the β estimates show relatively little uncertainty, while the confidence interval for ξ is relatively wide as is common for extremes studies. The GPD_τ row of Table 6.2 uses the same data set and shows selected quantile estimates and 95% bootstrap confidence intervals for the two covariate settings, along with the true quantile values in the last row. For this Monte Carlo simulation, the true quantile is contained in each of the confidence intervals. Bootstrap 95% confidence interval coverage rates for the entire simulation study are reported in the GPD_τ row of Table 6.3 for both covariate settings. Keeping in mind that there are only 504 confidence intervals considered, the coverage rate appears reasonable for the 0.95, 0.97, and 0.99 quantiles. Once again, performance deteriorates slightly in the 0.999 and 0.9999 quantiles, but the achieved coverage rate still yields a reasonable estimate of the uncertainty associated with these very high quantiles.

Tables 6.2 and 6.3 also include QR rows, which correspond to estimates of the quantiles obtained using standard quantile regression methods. Table 6.2 shows that my method and quantile regression yield similar estimates and 95% confidence intervals for the .95 quantile. Results for the .999 quantile, however, suggest that my method may be an improvement in generating confidence intervals for high quantiles, as its confidence intervals are narrower than those provided by quantile regression. Table 6.3 shows that the coverage rate of the

TABLE 6.1. Parameter estimates and 95% bootstrap confidence intervals.

Parameter	β_0	β_1	β_2	γ_0	γ_1	γ_2	ξ
Estimate	14.23	5.22	19.54	1.13	0.0131	0.0199	0.0403
Confidence Interval	(11.64, 16.40)	(5.16, 5.30)	(18.15,20.76)	(0.347, 1.965)	(-0.003,0.029)	(-0.0258,0.292)	(-0.103, 0.179)

TABLE 6.2. Quantile estimates and 95% bootstrap confidence intervals for GPD_τ and quantile regression (QR).

	Quantile	0.95: Setting 1	0.999: Setting 1	0.95: Setting 2	0.999: Setting 2
GPD_τ	Estimate	177.33	199.01	236.16	262.04
	Confidence Interval	(176.92, 178.40)	(193.42, 204.45)	(234.64, 236.52)	(255.77, 268.83)
QR	Estimate	177.80	211.12	235.83	269.41
	Confidence Interval	(176.81, 178.85)	(191.93, 219.17)	(234.69, 237.24)	(253.43, 286.99)
True Value		177.52	201.27	236.08	268.15

TABLE 6.3. 95% bootstrap confidence interval coverage rates and average widths for GPD_τ and quantile regression (QR).

		Quantile	0.95	0.97	0.99	0.999	0.9999
Setting 1	Coverage Rate (%)	GPD_τ	93.25	92.66	94.64	92.46	86.90
		QR	93.85	94.05	96.03	92.46	42.46
	Width	GPD_τ	1.752	2.189	4.251	14.020	42.105
		QR	1.743	2.439	5.302	29.682	35.492
Setting 2	Coverage Rate (%)	GPD_τ	95.44	93.65	92.86	91.87	88.10
		QR	95.83	94.84	94.25	93.85	49.80
	Width	GPD_τ	2.165	2.809	5.698	18.880	56.856
		QR	2.159	3.023	6.557	35.364	49.604

confidence intervals are comparable for my method versus quantile regression for the .95, .97, .99, and .999 quantiles, whereas my method clearly outperforms quantile regression for the .9999 quantile.

Figure 6.4 plots the width of each of the 504 confidence intervals provided by my method against the confidence interval widths of quantile regression for each of the quantiles. The plotted line shows a one-to-one relationship. The .95 quantile figures suggest that my GPD_τ method and quantile regression yield similar 95% confidence interval widths, whereas the .999 figures suggest that my method will produce a narrower confidence interval more often than quantile regression. Table 6.3 also shows the average width of the 95% confidence intervals for the different quantiles across both methods. While my method has larger average interval widths for the .95 quantile, the average widths are smaller for the .97, .99, and .999 quantiles. Interestingly, the .9999 quantile's mean interval width is actually larger for my method than

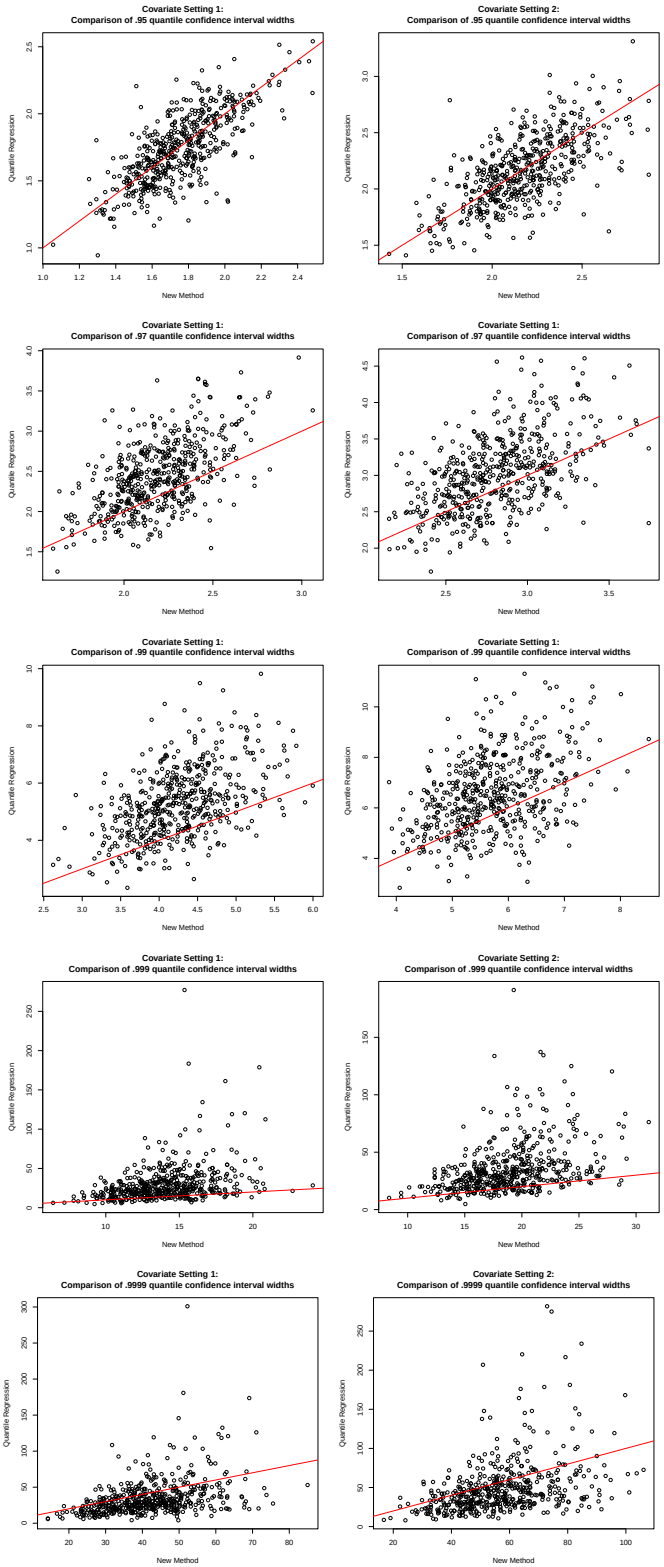

FIGURE 6.4. Comparison of 95% bootstrap confidence interval widths for my GPD_τ method versus quantile regression for the quantiles of interest. The plotted line shows a 1:1 relationship.

in quantile regression, but my method also does a much better job in capturing the true .9999 quantile.

6.3. Conclusion

In summary, the simulation study shows that my method yields both interpretable parameter estimates and reasonable estimates for high quantiles. My method also does a better job in capturing the uncertainty associated with large quantiles than standard quantile regression. That the quantile estimates are reasonable is important for understanding the approximate false discovery rate's association with some classification thresholds. The interpretability of the parameter estimates in this case of slight model mismatch will remain important as I turn my attention to the facial recognition application in Chapter 7, as one of its primary goals is to understand how covariates associated with the non-match scores influence the tail of the non-match distribution.

CHAPTER 7

Facial Recognition Application

7.1. Data: Non-match Scores and Covariates

A sample of the non-match pairs of the Bad partition of the Good, the Bad, and the Ugly (GBU) face challenge problem presented by Phillips et al. (2012) will be fit to my model for $\tau = 0.05$. This data set consists of similarity scores yielded by an algorithm that compares still query to target images. A set of covariates is attached to each image. The Good partition of the GBU data set contains images that are easy to match, whereas the Ugly partition contains images that are difficult to match. The Bad partition, which will be used, is considered to have average matching difficulty. The Bad partition contains 1,173,928 non-match pairs. To keep computational time manageable, I randomly selected 100,000 of these pairs to fit to the model.

Covariates in the GBU data set are assigned to each image. In the non-match setting, it is common for the covariates in the query and target images to be different. Thus, I found it necessary to create new covariates from the ones given in many instances. Specifically, in addition to an age difference covariate, I created new gender, glasses, and indoor or outdoor setting covariates so that each one had four categories based on the target/query pair. Gender, for example, would be classified as either female/female, female/male, male/female, or male/male. When fitting the model, these categorical covariates are separated into three binary covariates.

In addition to these newly created covariates, I will also use target and query FRIFM covariates when fitting the model. FRIFM is a continuous measurement of picture quality, which is defined in Section 3.2 of Beveridge et al. (2008). FRIFM is expected to differ

between any two images, so the target/query FRIFM values are included separately in the model.

7.2. Exploratory Data Analysis and Model Choice

The empirical .999 quantile of the randomly selected non-match scores is 4.093, thus this could be the classification threshold under current algorithms, regardless of covariates. The histograms in Figure 7.1 explore how the different covariates affect the tail and the probability of being incorrectly classified as a match. The top two rows of Figure 7.1 correspond to the categorical covariates, and the bottom two rows to the continuous covariates. The top row of each pair shows histograms for the entire sample, whereas the second row shows histograms for those non-match pairs in the sample that would exceed a classification threshold of 4.093. For many of these covariates, it is clear that the histograms differ, indicating that the value of the covariate affects the match score. Based on these histograms, it appears that images in which the categorical covariates are the same are more likely to be classified as matches than those in which the categorical covariates are not the same. Using gender as an example, a disproportionate amount of the target/query pairs which would be classified as matches were either MM or FF. Turning attention to the continuous covariates, it seems images comparing people with a smaller age difference are more likely to be classified as matches than those with large age differences. The two FRIFM covariates don't appear to have much of an effect on increasing the similarity score between two non-match pairs.

It is worth exploring how the tail index parameter ξ changes with different covariates. The 95% confidence intervals given by fitting a GPD to data exceeding the fixed empirical .95 quantile for different subsets of the data are calculated. For all the subsets, $\hat{\xi}$ is roughly in the range from -.1 to .05, and there is notable overlap in the confidence intervals. These

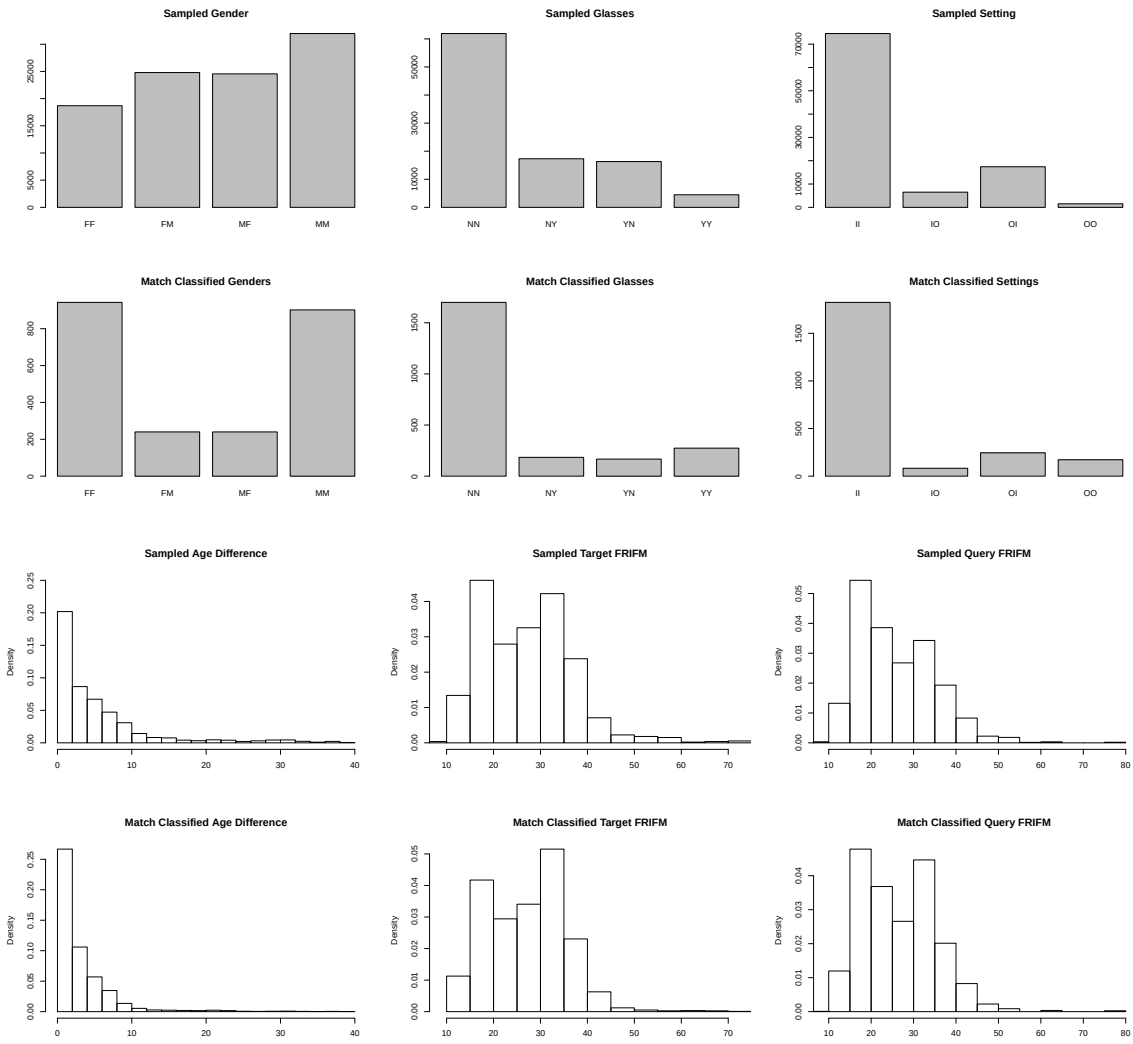

FIGURE 7.1. Top two rows are histograms showing breakdown of categorical variables in the overall sample (top row) and for pairs classified as matches (second row). Bottom two rows are histograms showing breakdown of numeric variables in the overall sample (third row) and for pairs classified as matches (bottom row).

intervals are displayed in Figure 7.2. Additionally, likelihood ratio tests performed on each of the six groupings of covariate subsets yielded large p-values when comparing the null model with common shape parameter to a model with a shape parameter that varies by subset, further suggesting that the use of a common ξ is appropriate. I conclude that use of a common ξ parameter, which is not a function of covariates, adequately models the data.

Further, if slight differences in true ξ values exist between the different groups, this will likely be compensated for by the flexibility in σ, satisfactorily capturing the tail behavior.

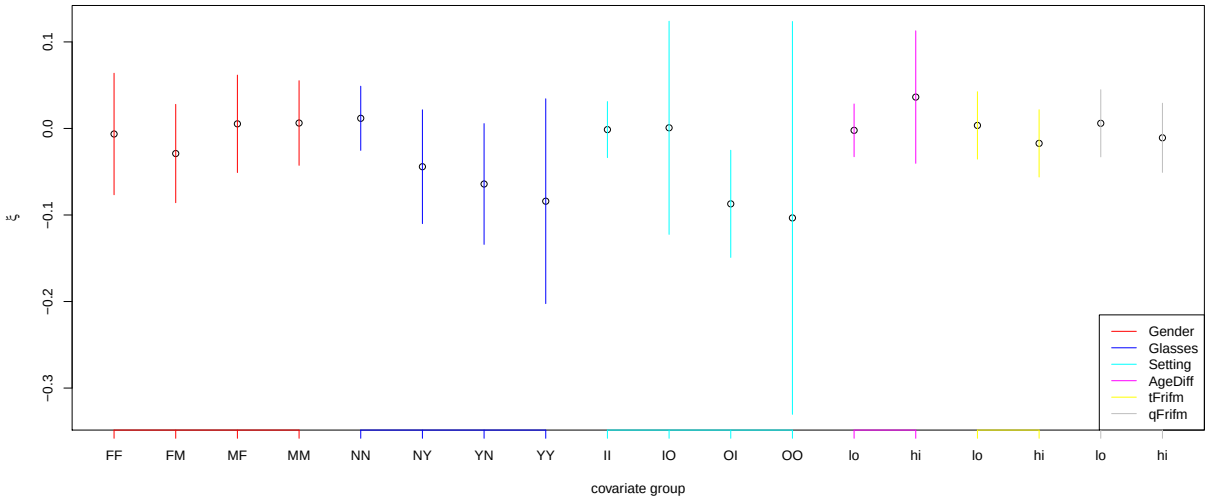

FIGURE 7.2. Shape point estimates and 95% confidence intervals.

Based on the results of the exploratory analysis, I will fit my model with $u_\tau = X\boldsymbol{\beta}$, $\sigma = \exp(X\boldsymbol{\gamma})$ where $\boldsymbol{\beta} = (\beta_0, \ldots, \beta_{12})^T$, $\boldsymbol{\gamma} = (\gamma_0, \ldots, \gamma_{12})^T$, and X is a design matrix with 13 columns. Coefficients 1 through 3 are indicators for the gender covariates, 4 through 6 are indicators for the glasses covariates, 7 through 9 are indicators for indoor or outdoor setting covariates, 10 corresponds to the age difference covariate, and 11 and 12 correspond to the two FRIFM covariates. I once again use 0.01 as the kernel density bandwidth.

Computing is distributed differently on the cluster than in the simulation study. Optimization here is much more expensive than it was in the simulation study, as the sample size is much larger and there are many more parameters to estimate. Bootstrapping is distributed across nodes, running 24 bootstrap fits on each node at a time, resulting in 1008 bootstrap instances used to calculate confidence intervals.

7.3. Results

7.3.1. Parameter Estimates and Interpretation. The parameter estimates, along with bootstrap confidence intervals, are reported in Table 7.1. I first interpret the parameters $\boldsymbol{\beta}$ which determine the threshold u_τ. All interpretations assume all other coefficients are being held constant.

The parameter estimates for the gender coefficients $\beta_1, \beta_2, \beta_3$ are all negative, suggesting that the non-match pairs containing two female subjects have the highest .95 quantile. The coefficients for the FM and MF categories are larger negative numbers indicating lower .95 quantiles for mixed-gender target/query pairs, likely reflecting an overall tendency for mixed gender scores to be lower. Parameter estimates for β_4, β_5, and β_6 indicate that target/query pairs where both subjects are wearing glasses have the highest .95 quantiles of four glasses categories, followed by cases where both subjects are not wearing glasses. The probability of being classified a match looks to increase fairly significantly if both pictures are taken outdoors. A non-match pair where both pictures are taken indoors is more likely to be classified as a match than pairs where the pictures are taken in different locations. Essentially, for all of the categorical covariates, u_τ is higher when there is agreement in the variable between the target and query.

I next interpret the β estimates describing how the continuous covariates affect u_τ. The negative estimate for the age difference covariate β_{10} indicates that as age difference increases the threshold u_τ decreases, thus non-match pairs with subjects that have similar ages have higher match scores. The FRIFM covariates β_{11} and β_{12} are both small in magnitude, although β_{12} is significantly different from zero.

TABLE 7.1. Parameter estimates for threshold parameters $\boldsymbol{\beta}$, scale parameters $\boldsymbol{\gamma}$, and shape parameter ξ.

Parameter	β_0	β_1: Gender FM	β_2: Gender MF	β_3: Gender MM	β_4: Glass NY	β_5: Glass YN	β_6: Glass YY
Estimate	4.252	-2.094	-2.046	-0.885	-0.761	-0.730	1.675
95% CI	$(4.09, 4.39)$	$(-2.19, -1.20)$	$(-2.15, -1.94)$	$(-0.99, -0.79)$	$(-0.86, -0.67)$	$(-0.83, -0.65)$	$(1.46, 1.86)$
Parameter	β_7: Setting IO	β_8: Setting OI	β_9: Setting OO	β_{10}: AgeDiff	β_{11}: tFRIFM	β_{12}: qFRIFM	-
Estimate	-0.438	-0.390	2.600	-0.041	-0.002	0.008	-
95% CI	$(-0.56, -0.31)$	$(-0.50, -0.31)$	$(2.26, 3.02)$	$(-0.045, -0.038)$	$(-0.005, 0.002)$	$(0.004, 0.012)$	-
Parameter	γ_0	γ_1	γ_2	γ_3	γ_4	γ_5	γ_6
Estimate	0.384	-0.096	-0.076	0.045	-0.176	-0.171	0.014
95% CI	$(0.067, 0.43)$	$(-0.11, 0.053)$	$(-0.093, 0.089)$	$(0.034, 0.20)$	$(-0.22, -0.023)$	$(-0.22, -0.011)$	$(-0.031, 0.23)$
Parameter	γ_7	γ_8	γ_9	γ_{10}	γ_{11}	γ_{12}	ξ
Estimate	-0.159	-0.187	0.224	-0.015	-0.003	0.008	-0.011
95% CI	$(-0.21, 0.048)$	$(-0.23, -0.026)$	$(0.080, 0.49)$	$(-0.019, -0.012)$	$(-0.008, -0.000)$	$(0.004, 0.012)$	$(-0.021, 0.045)$

Fewer of the scale parameter estimates are significant. Aside from γ_3, all the γ estimates which are significantly different from zero have the same sign as the estimate for the corresponding β, implying that an increase in u_τ tends to occur with an increase in the scale parameter σ. The significant positive estimate for γ_3 implies that when both query and target are male, the distribution above the threshold u_τ has larger scale than in the baseline FF case, despite the u_τ being lower for the MM case.

7.3.2. Covariate Effect on Tail and Probability of False Match Classification. To get an idea of how the non-match tail behaves under different covariate settings, I investigate 16 specific covariate settings. For the first 8 settings, which are listed in Table 7.2, the numeric variables are held constant, so that the age difference is 5, the target FRIFM is 25, and the query FRIFM is 25. For settings 9 through 16, categorical covariates are held constant, such that the non-match pairs both contain males, the target subject is not wearing glasses but the query subject is wearing glasses, and both pictures are taken indoors.

In addition to listing the settings, Table 7.2 lists the point estimates for u_τ and σ. It is clear that the covariates have noteworthy effect on these parameters. For instance, setting 1, which has all categorical covariates in agreement between query and target, has a much higher threshold and a scale parameter nearly double that of setting 2 which has all categorical covariates disagree. In fact, setting 1 has the highest threshold of any of the investigated settings, two units higher than any other that I tested. Also listed is the estimated probability that an observation with the listed covariates would have a similarity score exceeding the overall empirical .999 quantile of 4.093. Settings 1, 3, and 4 all have estimates for u_τ which exceed this level, meaning that the fitted model estimates that more

TABLE 7.2. Covariate values used for each setting with corresponding probabilities of exceeding the algorithm's classification threshold.

Covariate	Covariate Used						u_τ	σ	Prob
Setting	Gender	Glasses	Setting	AgeDiff	tFRIFM	qFRIFM			> 4.093
1	FF	YY	OO	5	25	25	8.473	1.963	> 0.05
2	FM	NY	IO	5	25	25	0.904	1.005	0.0018
3	FF	YN	OO	5	25	25	6.068	1.632	> 0.05
4	MF	NN	OO	5	25	25	4.752	1.795	> 0.05
5	FF	NY	II	5	25	25	3.437	1.297	0.0296
6	MF	YY	OI	5	25	25	3.437	1.206	0.0285
7	MM	NN	II	5	25	25	3.313	1.617	0.0332
8	MM	YN	II	5	25	25	2.583	1.363	0.0158
9	MM	NY	II	0	25	25	2.758	1.462	0.0194
10	MM	NY	II	0	40	10	2.607	1.236	0.0143
11	MM	NY	II	20	25	25	1.936	1.084	0.0063
12	MM	NY	II	20	10	10	1.845	1.004	0.0048
13	MM	NY	II	20	40	40	2.027	1.170	0.0079
14	MM	NY	II	40	25	25	1.114	0.804	0.0010
15	MM	NY	II	40	25	10	0.993	0.721	0.0005
16	MM	NY	II	40	25	40	1.235	0.908	0.0018

than 5% of observations with these covariates would be incorrectly classified as matches if this 4.093 were used as the classification threshold. Settings 1, 3, and 4 all compare images that were both taken outdoors.

Figure 7.3 plots the estimated GPD_τ distributions for the 16 settings' values for comparison. The top panel shows settings 1 through 8, and the bottom panel 9 through 16. The thick vertical lines in each figure represent the classification threshold of 4.093. Several of the aforementioned features are clearly illustrated with some distributions being entirely above the classification threshold. Differences in scales of the distributions are also evident. Other interesting aspects of the fitted model become evident in Figure 7.3, such as the fact that the estimated distributions for settings 5 and 6 are very similar despite the settings themselves being quite different. In the bottom panel, there is a noticeable distinction between settings 9 and 10, settings 11 through 13, and settings 14 through 16 which correspond to changes in age difference. As age difference gets smaller, the GPD_τ threshold gets bigger.

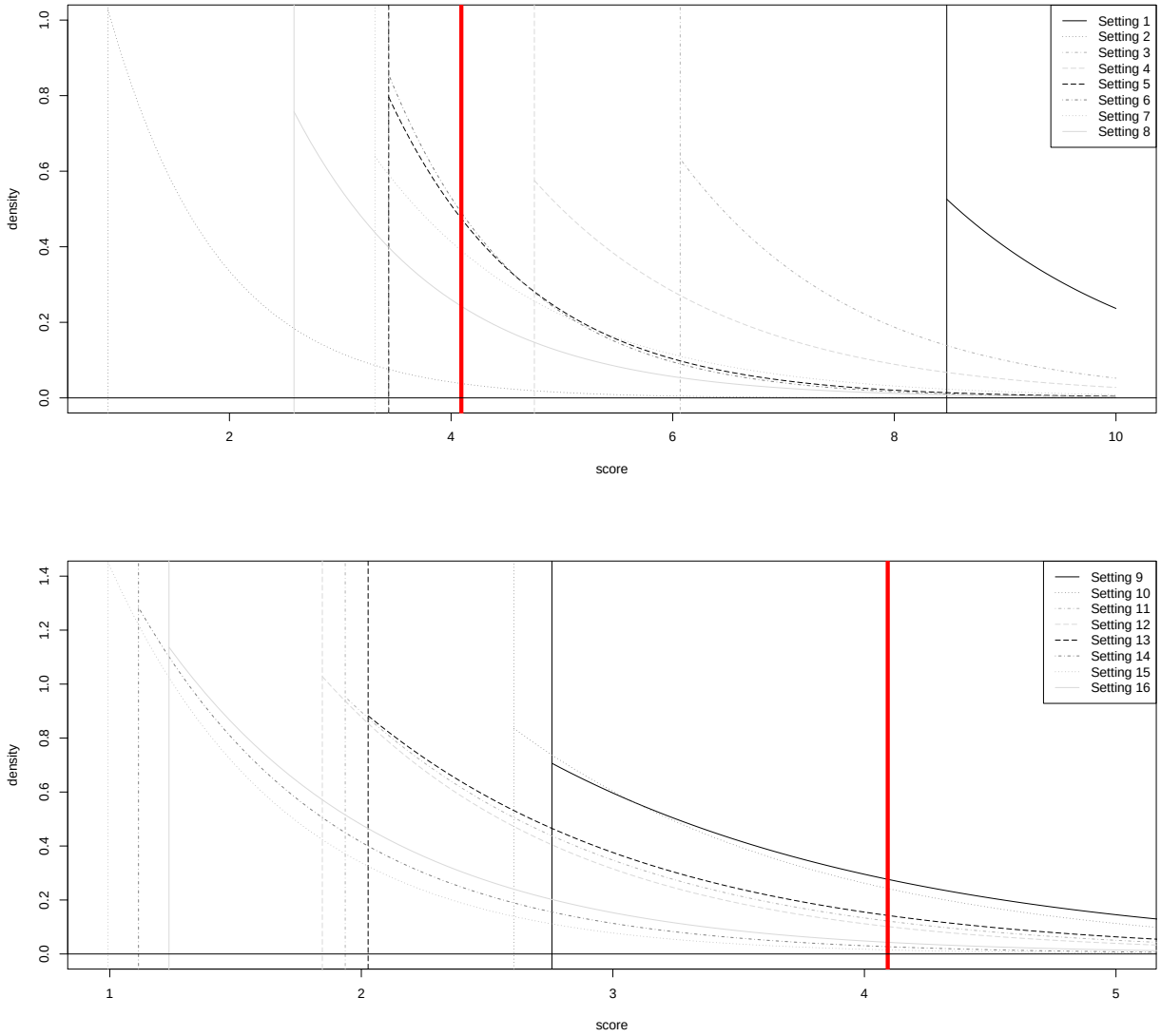

FIGURE 7.3. GPD_τ distributions for settings 1 through 8 (top) and 9 through 16 (bottom).

While changes in the target and query FRIFM do have an effect on the threshold placement, it's not as pronounced as the effect of age difference. Also note that none of the GPD_τ distributions displayed in the bottom panel of Figure 7.3 have thresholds that exceed the classification threshold. For all eight of these settings, the categorical covariates are fixed at settings which do not have the largest effect on the threshold, as the non-match pair is comparing two images of males taken indoors, where the target subject is not wearing glasses

but the query subject is wearing glasses. It appears that numeric covariates alone are not enough to push the GPD_τ threshold above the classification threshold.

7.3.3. Model Performance. Since only 100,000 of the non-match pairs in the GBU Bad partition are used, it is possible to compare the empirical .95 quantiles from the entire partition to the predicted u_τ values. Table 7.3 compares such empirical quantiles to the predicted u_τ and its confidence interval for select settings, chosen so that each setting had at least 100 observations in the sample of 100,000. Note that in order to find the empirical quantiles, I am ignoring both target and query FRIFM effects, which are minimal.

Table 7.3. Empirical .95 quantiles of the Bad partition compared to the predicted u_τ for select settings.

Covariate Setting	Bad Partition Empirical Quantile	u_τ	95% Confidence Interval for u_τ
5	2.974	3.437	(3.00, 3.88)
7	3.269	3.313	(2.85, 3.76)
8	2.175	2.583	(2.01, 3.11)
9	2.578	2.758	(2.22, 3.28)

In settings 7, 8, and 9, the Bad partition's .95 empirical quantile is contained within the 95% confidence interval for u_τ. The confidence interval for setting 5 does not include the .95 empirical quantile, though it is just below the lower bound. In this case, the .95 empirical quantile of the sample of 100,000 is 3.720, which suggests that the sample is a relatively poor representation of the Bad partition. More encouraging still, the model predicts a u_τ that lies between the two empirical .95 quantiles, suggesting that the model offsets this poor representation issue to some degree. Taking this into consideration, along with the performance for settings 7, 8, and 9, it appears that the model does an admirable job in estimating the .95 quantile.

7.4. Conclusion

In general, it appears that non-match pairs that compare images that are similar to each other in terms of subject gender, age, and use of glasses, as well as indoor or outdoor setting, have higher probabilities of being classified as matches. In some cases, such as situations where both images are taken outdoors, this probability far exceeds the 0.001 false accept rate that is applied to all non-match pairs when choosing the classification threshold. Furthermore, similarities in these situations are not created equal, as the algorithm is more likely to suggest two different female subjects are matches compared to two different male subjects. One way to lessen this probability of being incorrectly classified as a match is to control all images so that they are taken indoors and the subjects are not wearing glasses.

CHAPTER 8

Conclusion and Future Work

This dissertation presents a new method for modeling the tail of a distribution. The model is related to the peaks over threshold setting in extreme value theory. The utility of the model is demonstrated through both a simulation study and an application to a facial recognition setting. This chapter serves to review my work and to suggest future research directions.

8.1. Review

Chapter 1 describes the facial recognition setting which served as motivation for my work. An algorithm intended to determine if subjects in two different pictures are the same yields a set of non-match and match scores. Scores above some classification threshold are classified as matches, so that scores in the upper tail of the non-match distribution are false matches. Covariate information is not used to set the classification threshold. The purpose of my work is to determine whether the covariate information available can be used to predict whether a pair of images will lead to a false match.

In Chapter 2, I review standard statistical extreme value practices. The three-types theorem establishes the three possible limiting distributions of an n-block of maxima. These three domains of attraction carry over to the generalized Pareto distribution (GPD), which is the limiting distribution used in the peaks over threshold (POT) approach to modeling the tail of a distribution above some threshold u. Traditional POT methods fix a threshold u and estimate a probability of exceeding that threshold τ, but my new approach will fix an appropriate τ and estimate a threshold u_τ.

Chapter 3 establishes a new version of the generalized Pareto distribution, GPD_τ, which is a sensible model for the upper τth proportion of a distribution, so long as τ is relatively small. GPD_τ differs from the standard GPD in its aim, and it treats τ as fixed. The scale and threshold parameters of GPD_τ do not depend on each other as they do in the traditional GPD. Like the GPD, GPD_τ exhibits a type of threshold stability.

In Chapter 4, I use GPD_τ and quantile regression to develop an objective function to be used in order to estimate the threshold, shape, and scale parameters. Estimation is performed via M-estimation, and the consistency of the resulting estimators is proved.

Chapter 5 discusses the implementation of covariates into my model, as well as some practical considerations to improve optimization of the parameters. Covariate implementation is handled by introducing a kernel smoother to the objective function as a weight. A kernel with radius δ is centered at each observation, and all observations exceeding $u_\tau - \delta$ contribute to the GPD_τ portion of the objective function. Use of this kernel density appropriately handles issues arising from the fact that an observation may switch from appearing in the bulk distribution to appearing in the tail as optimization is performed. A penalty term is applied to the shape parameter to avoid unreasonable shape estimates. Because the objective function is not smooth, a Gauss-Seidel iterization scheme is used.

Chapter 6 presents a simulation study which verifies that my method can be used to reliably model the tail of some distribution, yielding unbiased quantile estimates. My method is comparable to quantile regression for estimating high quantiles, and outperforms quantile regression in estimating extreme quantiles.

Chapter 7 applies my model to the facial recognition data that motivated this work. I find that factors such as gender and indoor or outdoor setting strongly influence the location

of u_τ and, to a lesser extent, the scale σ of the tail. If classification were performed via the .99 or .999 empirical quantile across covariate levels, these factors would influence whether a pair of images are incorrectly classified as a match.

8.2. Future Work

Some of the suggestions I will make for future research are limited by the computational time of my method. Estimation using my method requires adequate computational resources, and many natural extensions of my work are liable to further increase computation time.

While use of a common shape parameter appears appropriate in the facial recognition setting, it is possible that covariates could change the tail behavior of a given distribution. Implementing covariates into the shape parameter is slightly more complicated than in the threshold or scale, as my method imposes a penalty on the shape parameter alone. Work to verify that the shape can adequately handle covariate implementation needs to be performed.

My work has treated all covariates as linearly related to the threshold and exponentially related to the scale, but it's theoretically possible for the covariates to have more complicated nonlinear relationships with the parameters. Implementation of higher order terms is likely straightforward. Interaction terms, which would be of particular interest in a facial recognition setting, could also be used. Use of such terms in my model should be verified for accuracy.

In the facial recognition application, I chose to use covariates that were of interest in previous facial recognition studies, foregoing a formal model selection process. Use of a stepwise procedure would increase the computational burden, as it requires repeated model fitting and bootstrapping in order to determine which covariates are significant. Use of likelihood ratio tests to determine covariate significance would be more tractable, eliminating the need

for bootstrapping, but these would require a likelihood, as would information criterion model selection methods such as AIC. The binomial objective function (18) discussed in Chapter 4.1.2 may prove useful here. It may be possible to treat this function as a log-likelihood once the parameters have been estimated, paving the way for a more efficient model selection method.

8.3. Conclusion

Importantly, my work clearly demonstrates that covariates can have a considerable effect on the upper tail of the distribution of non-match scores. GPD_τ, which produces a flexible model for the upper tail of a distribution, is an effective tool for capturing this covariate behavior, as GPD_τ's parameterization in terms of location (u_τ), scale (σ), and shape (ξ) yields interpretable covariate relationships. I look forward to improving my method's performance and applying it to a variety of different problems in the future.

References

Balkema, A. and De Haan, L. (1974). Residual life time at great age. *The Annals of Probability*, 2(5):792–804.

Behrens, C., Lopes, H., and Gamerman, D. (2004). Bayesian analysis of extreme events with threshold estimation. *Statistical Modelling*, 4(3):227.

Beirlant, J., Goegebeur, Y., Segers, J., Teugels, J., Waal, D. D., and Ferro, C. (2004). *Statistics of Extremes: Theory and Applications*. Wiley, New York.

Berman, S. M. (1962). Limiting distribution of the maximum term in sequences of dependent random variables. *The Annals of mathematical statistics*, pages 894–908.

Beveridge, J. R., Givens, G. H., Phillips, P. J., Draper, B., Lui, Y. M., et al. (2008). Focus on quality, predicting frvt 2006 performance. In *Automatic Face & Gesture Recognition, 2008. FG'08. 8th IEEE International Conference on*, pages 1–8. IEEE.

Beveridge, J. R., Givens, G. H., Phillips, P. J., and Draper, B. A. (2009). Factors that influence algorithm performance in the face recognition grand challenge. *Computer Vision and Image Understanding*, 113(6):750–762.

Caers, J. and Dyck, J. (1998). Nonparametric tail estimation using a double bootstrap method. *Computational statistics & data analysis*, 29(2):191–211.

Casella, G. and Berger, R. L. (1990). *Statistical inference*, volume 70. Duxbury Press Belmont, CA.

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

Coles, S. G. and Dixon, M. J. (1999). Likelihood-based inference for extreme value models. *Extremes*, 2(1):5–23.

Coles, S. G. and Tawn, J. A. (1996). Modelling extremes of the areal rainfall process. *Journal of the Royal Statistical Society. Series B (Methodological)*, pages 329–347.

Danielsson, J., de Haan, L., Peng, L., and de Vries, C. (2001). Using a bootstrap method to choose the sample fraction in tail index estimation. *Journal of Multivariate analysis*, 76(2):226–248.

Daugman, J. (2006). Probing the uniqueness and randomness of iriscodes: Results from 200 billion iris pair comparisons. *Proceedings of the IEEE*, 94(11):1927–1935.

do Nascimento, F. F., Gamerman, D., and Lopes, H. F. (2012). A semiparametric bayesian approach to extreme value estimation. *Statistics and Computing*, 22(2):661–675.

Fisher, R. A. and Tippett, L. H. C. (1928). Limiting forms of the frequency distribution of the larges or smallest members of a sample. *Proceedings of the Cambridge Philosophical Society*, 24:180–190.

Freedman, D. and Diaconis, P. (1982). On inconsistent m-estimators. *The Annals of Statistics*, pages 454–461.

Frigessi, A., Haug, O., and Rue, H. (2002). A dynamic mixture model for unsupervised tail estimation without threshold selection. *Extremes*, 5(3):219–235.

Givens, G., Beveridge, J., Phillips, P., Draper, B., Lui, Y., and Bolme, D. (2013). Introduction to face recognition and evaluation of algorithm performance. *Computational Statistics & Data Analysis*, 67:236–247.

Givens, G., Beveridge, J. R., Draper, B. A., Grother, P., and Phillips, P. J. (2004). How features of the human face affect recognition: a statistical comparison of three face recognition algorithms. In *Computer Vision and Pattern Recognition, 2004. CVPR 2004. Proceedings of the 2004 IEEE Computer Society Conference on*, volume 2, pages II–381. IEEE.

Givens, G. H. and Hoeting, J. A. (2012). *Computational statistics*, volume 710. John Wiley & Sons.

Gnedenko, B. (1943). Sur la distribution limite du terme maximum d'une série aléatoire. *The Annals of Mathematics*, 44(3):423–453.

Gross, R., Shi, J., and Cohn, J. (2001). Quo vadis face recognition? In *Third Workshop on Empirical Evaluation Methods in Computer Vision*.

Grother, P. and Tabassi, E. (2007). Performance of biometric quality measures. *Pattern Analysis and Machine Intelligence, IEEE Transactions on*, 29(4):531–543.

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

Haberman, S. J. (1989). Concavity and estimation. *The Annals of Statistics*, 17(4):1631–1661.

Hallock, K. and Koenker, R. (2001). Quantile regression. *Journal of Economic Perspectives*, 15(4):143–156.

Hampel, F. R., Ronchetti, E. M., Rousseeuw, P. J., and Stahel, W. A. (1986). *Robust statistics: the approach based on influence functions*. John Wiley & Sons.

Heffernan, J. E. and Stephenson, A. G. (2012). *ismev: An Introduction to Statistical Modeling of Extreme Values*. R package version 1.39.

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

Hjort, N. L. and Pollard, D. (2011). Asymptotics for minimisers of convex processes. *arXiv preprint arXiv:1107.3806*.

Hosking, J. R. and Wallis, J. R. (1987). Parameter and quantile estimation for the generalized pareto distribution. *Technometrics*, 29(3):339–349.

Huber, P. J. (2011). *Robust statistics*. Springer.

Jurecková, J. and Sen, P. K. (1996). *Robust statistical procedures: asymptotics and interrelations*, volume 311. John Wiley & Sons.

Koenker, R. (2005). *Quantile regression*. Number 38. Cambridge university press.

Koenker, R. and Bassett Jr, G. (1978). Regression quantiles. *Econometrica: journal of the Econometric Society*, pages 33–50.

Lehmann, E. L. and Casella, G. (1998). *Theory of point estimation*, volume 31. Springer.

Lui, Y. M., Bolme, D., Draper, B., Beveridge, J. R., Givens, G., Phillips, P. J., et al. (2009). A meta-analysis of face recognition covariates. In *Biometrics: Theory, Applications, and Systems, 2009. BTAS'09. IEEE 3rd International Conference on*, pages 1–8. IEEE.

MacDonald, A., Scarrott, C. J., Lee, D., Darlow, B., Reale, M., and Russell, G. (2011). A flexible extreme value mixture model. *Computational Statistics & Data Analysis*, 55(6):2137–2157.

Maronna, R. A., Martin, R. D., and Yohai, V. J. (2006). Robust statistics: Theory and methods. *J. Wiley*.

Martins, E. S. and Stedinger, J. R. (2000). Generalized maximum-likelihood generalized extreme-value quantile estimators for hydrologic data. *Water Resources Research*, 36(3):737–744.

Mizera, I. (1994). On consistent m-estimators: Tuning constants, unimodality and breakdown. *Kybernetika*, 30(3):289–300.

Niemiro, W. (1992). Asymptotics for m-estimators defined by convex minimization. *The Annals of Statistics*, pages 1514–1533.

O'Toole, A. J., Phillips, P. J., An, X., and Dunlop, J. (2012). Demographic effects on estimates of automatic face recognition performance. *Image and Vision Computing*, 30(3):169–176.

Phillips, P. J., Beveridge, J. R., Draper, B. A., Givens, G., O'Toole, A. J., Bolme, D., Dunlop, J., Lui, Y. M., Sahibzada, H., and Weimer, S. (2012). The good, the bad, and the ugly face challenge problem. *Image and Vision Computing*, 30(3):177–185.

Pickands III, J. (1975). Statistical inference using extreme order statistics. *the Annals of Statistics*, pages 119–131.

Scarrott, C. and MacDonald, A. (2012). A review of extreme value threshold es-timation and uncertainty quantification. *REVSTAT–Statistical Journal*, 10(1):33–60.

Scheirer, W., Rocha, A., Micheals, R., and Boult, T. (2010). Robust fusion: extreme value theory for recognition score normalization. In *Computer Vision–ECCV 2010*, pages 481–495. Springer.

Scheirer, W. J., Rocha, A., Micheals, R. J., and Boult, T. E. (2011). Meta-recognition: The theory and practice of recognition score analysis. *Pattern Analysis and Machine Intelligence, IEEE Transactions on*, 33(8):1689–1695.

Shevlyakov, G., Morgenthaler, S., and Shurygin, A. (2008). Redescending m-estimators. *Journal of Statistical Planning and Inference*, 138(10):2906–2917.

Shi, Z., Kiefer, F., Schneider, J., and Govindaraju, V. (2008). Modeling biometric systems using the general pareto distribution (gpd). In *SPIE Defense and Security Symposium*, pages 69440O–69440O. International Society for Optics and Photonics.

Smith, R. L. (1989). Extreme value analysis of environmental time series: an application to trend detection in ground-level ozone. *Statistical Science*, pages 367–377.

Tancredi, A., Anderson, C., and OHagan, A. (2006). Accounting for threshold uncertainty in extreme value estimation. *Extremes*, 9(2):87–106.

Xiangxian, Z. and Wenlei, G. (2009). A new method to choose the threshold in the pot model. In *Information Science and Engineering (ICISE), 2009 1st International Conference on*, pages 750–753. IEEE.

Zhao, W., Chellappa, R., Phillips, P. J., and Rosenfeld, A. (2003). Face recognition: A literature survey. *ACM computing surveys (CSUR)*, 35(4):399–458.

APPENDIX A

Grid Search Method Proof

THEOREM A.0.1. *Let M_n be defined as in (18). Consider u such that $Y_{(i)} < u \leq Y_{(i+1)}$. Let σ^* and ξ^* be the best estimates of σ and ξ so that for any u, $M_n(u, \sigma, \xi; \boldsymbol{y}) \leq M_n(u, \sigma^*, \xi^*; \boldsymbol{y})$. Then the objective function M_n is maximized when $u = Y_{(i+1)}$.*

PROOF. Assume that $M_n(u, \sigma^*, \xi^*; \boldsymbol{y}) > M_n\left(Y_{(i+1)}, \sigma^*, \xi^*; \boldsymbol{y}\right)$. The proof will be broken down into three cases.

Case 1: $\xi^* = 0$

For each j,

$$\begin{aligned} \frac{1}{\sigma}\exp\left(-\frac{Y_{(j)}-u}{\sigma}\right) &> \frac{1}{\sigma}\exp\left(-\frac{Y_{(j)}-Y_{(i+1)}}{\sigma}\right) \\ -\left(Y_{(j)}-u\right) &> -\left(Y_{(j)}-Y_{(i+1)}\right) \\ u &< Y_{(i+1)}, \end{aligned}$$

which is a contradiction. So $M_n\left(Y_{(i+1)}, \sigma^*, \xi^*; \boldsymbol{y}\right) \geq M_n(u, \sigma^*, \xi^*; \boldsymbol{y})$ in Case 1.

Case 2: $0 < \xi^* \leq 0.5$

Note that $\left(-\frac{1}{\xi^*}-1\right)$ is negative. So for each j,

$$\begin{aligned} \left(-\frac{1}{\xi^*}-1\right)\log\left(\xi^*\frac{Y_{(j)}-u}{\sigma}+\tau^{-\xi}\right) &> \left(-\frac{1}{\xi^*}-1\right)\log\left(\xi^*\frac{Y_{(j)}-Y_{(i+1)}}{\sigma}+\tau^{-\xi}\right) \\ \log\left(\xi^*\frac{Y_{(j)}-u}{\sigma}+\tau^{-\xi}\right) &< \log\left(\xi^*\frac{Y_{(j)}-Y_{(i+1)}}{\sigma}+\tau^{-\xi}\right) \\ Y_{(j)}-u &< Y_{(j)}-Y_{(i+1)} \\ Y_{(i+1)} &< u, \end{aligned}$$

which is a contradiction. Thus, $M_n\left(Y_{(i+1)}, \sigma^*, \xi^*; \boldsymbol{y}\right) \geq M_n(u, \sigma^*, \xi^*; \boldsymbol{y})$ in Case 2.

Case 3: $-0.5 \leq \xi^* < 0$

$\left(-\frac{1}{\xi^*} - 1\right)$ is positive here. For each j,

$$\begin{aligned}
\left(-\frac{1}{\xi^*} - 1\right) \log\left(\xi^* \frac{Y_{(j)} - u}{\sigma} + \tau^{-\xi}\right) &> \left(-\frac{1}{\xi^*} - 1\right) \log\left(\xi^* \frac{Y_{(j)} - Y_{(i+1)}}{\sigma} + \tau^{-\xi}\right) \\
\xi^* \frac{Y_{(j)} - u}{\sigma} &> \xi^* \frac{Y_{(j)} - Y_{(i+1)}}{\sigma} \\
Y_{(j)} - u &< Y_{(j)} - Y_{(i+1)} \\
Y_{(i+1)} &< u,
\end{aligned}$$

which is a contradiction. Thus, $M_n\left(Y_{(i+1)}, \sigma^*, \xi^*; \boldsymbol{y}\right) \geq M_n\left(u, \sigma^*, \xi^*; \boldsymbol{y}\right)$ in Case 3.

Thus, $M_n\left(Y_{(i+1)}, \sigma^*, \xi^*; \boldsymbol{y}\right) \geq M_n\left(u, \sigma^*, \xi^*; \boldsymbol{y}\right)$ for all possible values of ξ^*, and the objective function M_n from (18) for a fixed set of exceedances is maximized when u is equal to the smallest exceedance. □