Statistics and Its Interface Volume 10 (2017) 711–725

Modeling the upper tail of the distribution of facial recognition non-match scores

Brett D. Hunter*, Daniel Cooley, Geof H. Givens, and J. Ross Beveridge

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. We develop a general model for the non-match distribution above u_τ, the $(1-\tau)$th quantile, borrowing ideas from extreme value theory. We call this model the GPD$_\tau$, as it can be viewed as a reparameterized generalized Pareto distribution (GPD). This novel model treats τ as fixed and allows us to estimate u_τ in addition to parameters describing the tail. Inference for both u_τ and the GPD$_\tau$ scale and shape parameters is performed via M-estimation, where our objective function is a combination of the quantile regression loss function and GPD$_\tau$ density. By parameterizing u_τ and the GPD$_\tau$ parameters in terms of available covariates, we gain understanding of these covariates' influence on the tail of the distribution of non-match scores. A simulation study shows that our method is able to estimate both the set of parameters describing the covariates' influence and high quantiles of the non-match distribution. We apply our 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.

Keywords and phrases: Generalized Pareto, M-estimation, Quantile regression.

1. INTRODUCTION

1.1 Facial recognition

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 [25]. 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. 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, certain properties of the images are known such as whether the images were taken indoors or outdoors, if the people in the images are wearing glasses, or metrics of image quality. Thus, the overall non-match distribution is actually a mixture of a number of distributions given covariates. Our primary aim is to model the tail of the non-match distribution given knowledge of the covariates, thereby understanding how covariates influence the tail of the resulting non-match distribution. We do not limit ourselves to modeling a single quantile (like the classification quantile), but instead develop a general model for the entire tail of the distribution above the $(1-\tau)$th quantile, where τ is a level of interest set by the researchers, and which is below any possible classification quantile. We will denote the $(1-\tau)$th quantile by u_τ. A major goal is to provide parametric estimates which are easily interpretable by our facial recognition collaborators. That is, parameters should link covariates to interpretable quantities like the location of u_τ, the 'scale' of the tail above u_τ, and the general shape/behavior of the tail. Furthermore, we do not want data from the bulk of the distribution to contaminate our inference about the tail.

*Corresponding author.

1.2 Threshold exceedance methods from extreme value theory

Extreme value analysis has well-developed methods for modeling threshold exceedances above a suitably high threshold. Let Y have distribution F_Y whose upper tail we wish to characterize. Typical extremes approaches fix a threshold u in order to estimate $\tau_u = P(Y > u)$ in addition to the parameters which characterize $F_{Y|Y>u}$. In this work, we do the reverse. We aim to model F above the $(1-\tau)$th quantile for fixed τ. That is, we wish to simultaneously estimate both u_τ (such that $F(u_\tau) = 1-\tau$) and the parameters which characterize $F_{Y|Y>u_\tau}$. As in typical extreme value analysis, we are interested in estimating very high quantiles, but we also extend our focus to the value of u_τ and how it is affected by available covariates.

[22] and [1] showed that if a distribution is in the domain of attraction of an extreme value distribution, 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

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

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 $\xi < 0$. ξ is the shape parameter and controls the nature of the tail. If $\xi > 0$ the tail is said to be heavy and the distribution's tail decays like a power function, if $\xi = 0$ then the distribution has an exponentially decaying tail, and if $\xi < 0$ the distribution has a bounded tail. Additionally, we denote the probability of a given observation exceeding the threshold u by τ_u, and this additional parameter is needed to calculate unconditional high quantiles.

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.

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 [4, Section 4.3.1]. However, threshold selection remains subjective and imprecise, and there has been some work to develop automated threshold selection methods, such as methods proposed by [7], [11], [24], and [18].

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. A widely used generalized-linear-model-like approach is to let the parameters of the distribution describing tail behavior be simple, often linear, functions of covariates [2, Chapter 7], [4, Chapter 6]. Several studies have employed models where the shape and scale parameters of the generalized Pareto distribution vary with covariates [2, 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 [23] can be used. [4, Section 7.6] suggests using the point process setting over a threshold exceedance model when working with time-varying thresholds, for example. [6] further detail the use of point process models involving covariates.

In Section 2 we develop a model for the upper τth proportion of a distribution, focusing on the more simple case where there are no covariates. In Section 3 we discuss our inference method via M-estimation, and discuss additional inferential challenges when parameters are themselves functions of covariates. In Section 4 we illustrate with an extensive simulation study, and in Section 5 we model a dataset of non-match facial recognition scores.

2. A MODEL FOR THE TAIL ABOVE THE $(1-\tau)$TH QUANTILE

Because we aim to model the upper tail corresponding to a *fixed* proportion τ, our approach cannot be viewed in the usual context of extreme value theory. Nevertheless, we borrow ideas from extremes, as extremes models provide a flexible framework which can describe tail behavior in just a few parameters which are largely interpretable. Our model, developed below, assumes that the GPD is a useful approximation for the tail above the $(1-\tau)$th quantile. Because τ is fixed, we cannot justify our model by a convergence result. In fact, there is no limiting distribution for the distribution above u_τ — one would have to know the distribution to 'correctly' model its upper τth proportion. Importantly, our approach follows the general practice of extremes of "letting the tail speak for itself". Our 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 our model is more complicated than traditional extremes studies because the threshold u_τ is estimated rather than being fixed at the outset, meaning that the exceedances vary with the parameter u_τ.

The three-types theorem [8, 10] states that as $n \to \infty$,

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

Assuming n is fixed and large enough for the above convergence to imply approximate equality, then for z a high

quantile of Y,

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

Assuming this approximation is appropriate for u_τ,

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

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

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

By conditioning we obtain

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

For our fixed n, defining $\sigma = a_n n^{-\xi}$ allows us to eliminate n, yielding

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

As this can be viewed as a reparametrization of the GPD, we refer to the conditional distribution given in (1) as the GPD_τ. Its density is given by

$$g_\tau(z; u_\tau, \sigma, \xi) = \frac{1}{\tau\sigma}\left(\xi\frac{z - u_\tau}{\sigma} + \tau^{-\xi}\right)^{-\frac{1}{\xi}-1} \quad (2)$$

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 (1, 2) does not depend on the threshold. For $\xi = 0$, both (1) and (2) should be interpreted as limits, and yield the exponential distribution just as with the standard GPD.

A version of the threshold stability property characterized by the generalized Pareto distribution is exhibited by GPD_τ. Suppose a random variable Y conditionally exceeding u_0 follows a GPD_τ with parameters u_0, σ, and ξ such that $P(Y > u_0) = \tau_0$. Then Y conditionally exceeding $u > u_0$ follows a GPD_τ with parameters u, σ, and ξ such that $P(Y > u) = \tau$, where $\tau = \left(\xi\left(u - u_0\right)/\sigma + \tau_0^{-\xi}\right)^{-1/\xi}$.

3. PARAMETER ESTIMATION

Given a set of observations, fitting the model from Section 2 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 [17, 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 (2) as u_τ is a parameter and the data exceeding this threshold is not fixed.

However, quantile regression [16] is a well-developed method for estimating quantiles and additionally modeling these quantiles' behavior in terms of covariates. It would seem quantile regression could be sensibly combined with the model in Section 2 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 (2) to create a likelihood. However, a disadvantage to this approach is that it would not propagate the uncertainty in the threshold. Instead, since both quantile regression and maximum likelihood are both M-estimators we create an objective function which combines the loss function from quantile regression and a 'likelihood' for estimating the GPD_τ parameters.

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

$$\begin{aligned} M_n\left(u_\tau, \sigma, \xi; \boldsymbol{y}\right) = &\sum_{i=1}^{n} q\left(u_\tau; y_i\right) \\ &+ \frac{1}{N}\sum_{i=1}^{n} \log g_\tau\left(u_\tau, \sigma, \xi; y_i\right) \mathbb{I}_{y_i \geq u_\tau}, \end{aligned} \quad (3)$$

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

$$q\left(u_\tau; y_i\right) = \tau\left(y_i - u_\tau\right)\mathbb{I}_{y_i < u_\tau} + \left(\tau - 1\right)\left(y_i - u_\tau\right)\mathbb{I}_{y_i \geq u_\tau} \quad (4)$$

arises from the loss function commonly used in quantile regression [16]. Thus, the objective function is the quantile regression objective function plus the *mean* log-'likelihood' contribution of the exceedances. We will perform M-estimation; that is, we seek the u_τ, σ, and ξ which maximize (3).

We provide some explanation of why the mean log-'likelihood' contribution is taken in (3) rather than the sum. In short, the mean likelihood is taken so that the GPD_τ piece has little influence on the estimate for u_τ. In the usual case, 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 (3), this second term's magnitude would increase with N, resulting in biased estimates for u_τ. In our investigations, the contribution from the GPD_τ piece has tended to be negative, thus, using the naive objection function (with a sum rather than mean) results in estimates

of u_τ which are biased high as the naive objective function favors values which result in too few exceedances. With the mean log-'likelihood', the second term of (3) converges to the mean log-likelihood contribution above u_τ rather than increasing with N. Since the mean is used, this piece's influence relative to the quantile regression piece lessens as the sample size grows (by design).

The objective function has the appealing property that only observations which exceed u_τ will influence the estimates of σ and ξ, because these parameters only appear in the mean log-'likelihood' piece. Importantly for a given u, the same values of σ and ξ which maximize the mean log-'likelihood' also maximize the standard log-likelihood.

An M-estimator is any estimate θ defined by minimizing $\sum_{i=1}^{n} \rho(x_i;\theta)$ [14, Section 3.2]. While there are some established sufficient conditions for M-estimator consistency, they are either hard to show or not widely applicable. [14] gives a set of five conditions that are sufficient for M-estimator consistency, but three of them rely on the existence of some unknown functions. [12] and [20] outline sufficient conditions that rely on convexity of the criterion function $\rho(x_i;T_n)$, but the objective function we minimize does not adhere to such a requirement.

Since scale and shape parameters only appear in the GPD$_\tau$ piece of (3), consistency of these parameters follows from standard extremes arguments. In order to prove consistency of u_τ, we show that for $u^* \neq u_\tau$ and as $n \to \infty$,

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

which is sufficient for showing consistency. We 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 term can be broken down into quantile regression and GPD pieces. The proof, given as supplementary material (http://intlpress.com/site/pub/pages/journals/items/sii/content/vols/0010/0004/s004), proceeds to show that the quantile regression difference inflates as n increases and the GPD difference is bounded below, so that the sum of the two must exceed 0 above some n.

3.1 Estimation and optimization

M-estimation provides a viable method to estimate the parameters in our model; however, it does not imply that optimization is straightforward. There are some practical modifications to such optimization to improve estimation.

The parameter u_τ appears in both the quantile regression and GPD$_\tau$ pieces of the objective function. As discussed earlier, because the quantile regression piece grows with n and the GPD$_\tau$ piece converges to a value, the quantile regression exerts far more influence on the estimate of u_τ (by design). However the imbalance of 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, we employ non-linear Gauss-Seidel iterization [9, Section 2.2.5]. Each iteration of our optimization has two steps: the first step optimizes the threshold parameter(s), and the second step optimizes the GPD parameters.

It is known that numerical maximum likelihood can produce very high estimates for ξ, particularly when sample size is small [5]. As our M-estimation method also requires numerical optimization, similar difficulties can arise. In initial tests of our simulation study (presented in Section 4), we found that a small number of the simulations would numerically converge to absurdly high estimates of ξ. Both [5] and [19] advocate penalized likelihood approaches which enforce ξ to take on reasonable values. Similar to [19], for our simulation study we construct a penalty via a shifted beta distribution centered at 0, which restricts the shape parameter to values in $[-0.5, 0.5]$. We think it is reasonable to assume that ξ is in this interval because if the shape is less than -0.5, the density evaluated at the upper endpoint exceeds 0 and if the shape is greater than 0.5, then the distribution does not have a finite variance. The shifted beta's log-density is

$$p(\xi) = \log\left(\frac{(0.5+\xi)^{\alpha-1}(0.5-\xi)^{\beta-1}}{B(\alpha,\beta)}\right), \tag{5}$$

where $B(\alpha,\beta)$ denotes the beta function. Throughout this study we set $\alpha = 2$ and $\beta = 2$ yielding a penalty symmetric about 0.

Implementation of the penalty must be done slightly differently for our M-estimator than in the likelihood setting. In the penalized likelihood setting, a penalty such as the one in (5) is added onto the log-likelihood, and because the 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 (3), since the magnitude of the 'likelihood' piece does not increase with sample size, we need to impose a penalty whose influence will decrease with sample size. Our penalized objective function is

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

Consequently, as $n \to \infty$, the estimate of ξ from the penalized objective function approaches the estimate of ξ from the unpenalized objective function. In the application in Section 5, we performed estimation both with and without this penalty term, and found that it had very little influence on the results.

$$M_n(\boldsymbol{y};u_\tau,\sigma,\xi) = q(u_\tau;\boldsymbol{y}) + \frac{1}{\sum_{i=1}^{n} w_i}p(\xi;\boldsymbol{y}) + \frac{1}{\sum_{i=1}^{n} w_i}\sum_{i=1}^{n} w_i \log g_\tau(u_\tau,\sigma,\xi;y_i) \tag{6}$$

3.2 Covariates

Our study is motivated by a desire to understand how covariate information affects the upper tail of a distribution. Let $\boldsymbol{X} = (X_1, \ldots, X_k)$ be a vector of covariate information. Taking a GLM-like approach, we assume

$$\begin{aligned} u_\tau &= f_{u_\tau}(\boldsymbol{X}, \boldsymbol{\beta}) \\ \sigma &= f_\sigma(\boldsymbol{X}, \boldsymbol{\gamma}) \\ \xi &= f_\xi(\boldsymbol{X}, \boldsymbol{\eta}). \end{aligned} \tag{7}$$

We continue to perform M-estimation finding the $\boldsymbol{\beta}, \boldsymbol{\sigma}$, and $\boldsymbol{\eta}$ which maximize the objective function.

However, there is an optimization issue which introducing covariates makes more complex. The issue is that as the value of u_τ changes such that data points are either included in or excluded from the set of exceedances, the objective function has a discontinuous jump which typically cause optimization programs to perform poorly. In a simple setting with no covariates for u_τ, the objective function has a local maximum at each observation and optimization of u_τ can be done by individually testing the discrete possible values [15]. Such a discrete approach is not possible when u_τ is a function of covariates. To improve the performance of the optimization, we introduce smoothness into the objective function. Essentially, rather than treating each observation as a unitary mass at a point, we center a kernel density at each observation. This introduces a weight into the objective function, where the weight corresponds to mass of the kernel which exceeds the threshold. Thus, whereas exceedances and non-exceedances were given respective weights of 1 and 0 before, now if the value of u_τ increases across an observation's value, the observation's contribution to M smoothly varies from 1 down to 0. This kernel smoothing also allows for the implementation of continuous covariates, which may have been adversely affected by the presence of discontinuous jumps in the objective function, into the model.

We use an isotropic kernel density with finite support and we 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. Using threshold stability of GPD_τ (and assuming this holds for values above $u_\tau - \delta$), one can show $\tau_\delta = \left(\tau^{-\xi} - \xi\delta/\sigma\right)^{-1/\xi}$, allowing us to fit observations above $u_\tau - \delta$ and still estimate u_τ. The bandwidth of the kernel involves a tradeoff: 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. In both the simulation study and the application, we use a uniform kernel with bandwidth 0.01, and sensitivity analysis performed on bandwidth selection had very little effect on results.

Thus, an objective function that can properly handle continuous covariates with the kernel density smoothing implemented is given in equation (6), where w_i are weights. These are defined such that $w_i = P(Y_i > u_\tau)$ for $Y_i \sim k_h(y_i)$ where $k_h(y_i)$ is the kernel density of the ith observation with bandwidth h.

In order to set the initial threshold parameters, we used a simple quantile regression fit. Shape and scale parameters are given initial values as in the ismev package in R [13].

4. SIMULATION STUDY

4.1 Set up

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

$$Y = 10 + 5X_1 + 20X_2 + \exp(1 + 0.02X_1)\, T_4, \tag{8}$$

where X_1 is a continuous variable with values spanning uniformly 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, we 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 model we fit captures the general behavior of the generating equation (8), it does not correspond exactly. For example, the true u_τ resulting from (8) is not linear. Also note that we fit a scale parameter using the categorical variable despite that it does not appear in the scaling term applied to the t-distributed random variable. Most importantly, the model we fit only models the tail using a very general model, whereas (8) specifies the entire distribution.

To obtain confidence intervals for both parameter estimates and estimated high quantiles, we perform a semiparametric paired bootstrap. Our procedure is as follows, where $(x_i, y_i), i = 1, \ldots, n$ denotes independent observations from (8):

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 we 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.

By using this semiparametric bootstrap process, we eliminate ties in the tail of the resampled data set, so that when it is used to fit our model, we have a better representation of the tail.

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.

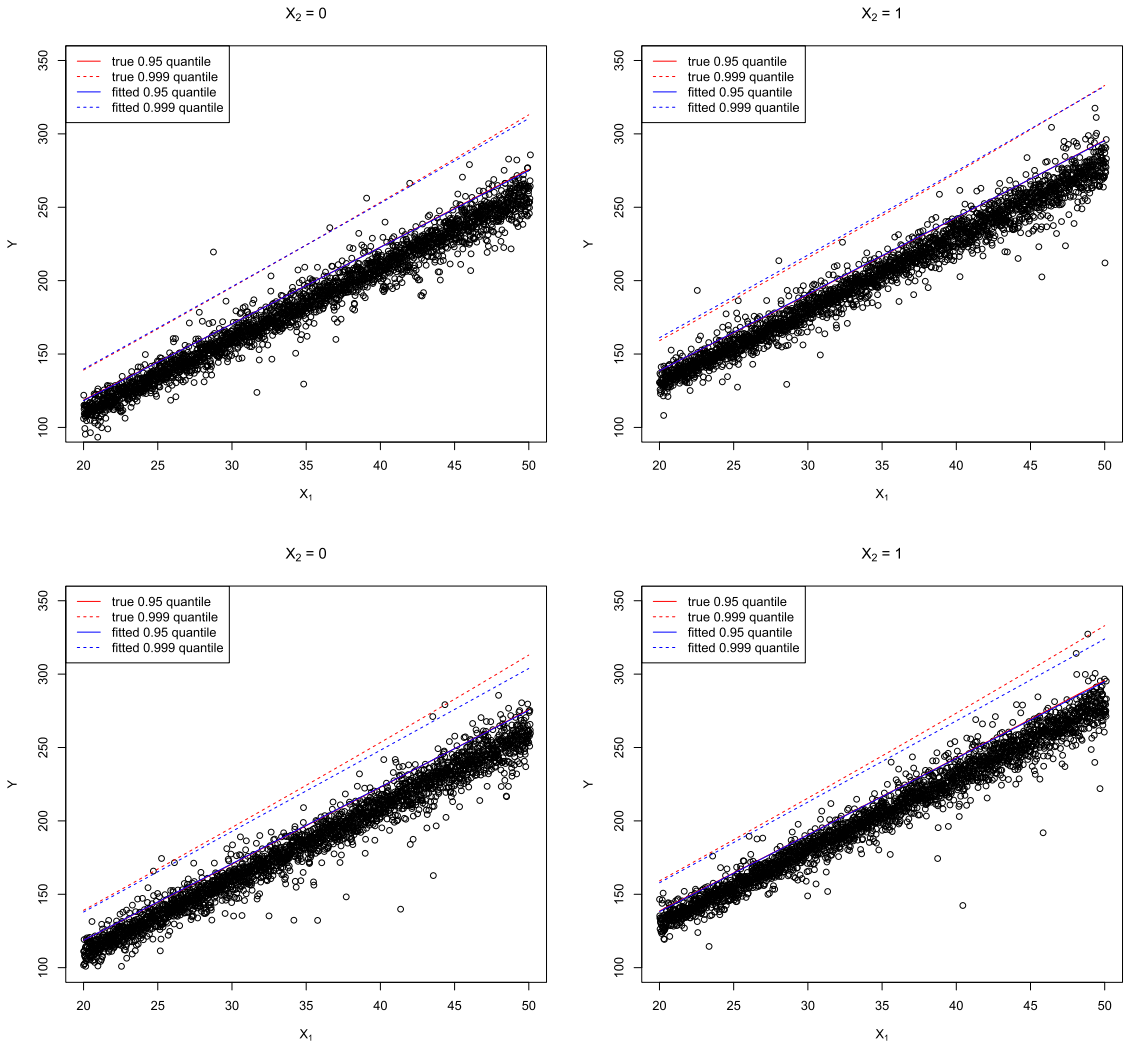

Figure 1. Fitted and true quantile against Monte Carlo generated data set 1 (top) and data set 2 (bottom).

We distribute the computing 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 could perform the process on each node in under 24 hours, and the system's queueing system allowed us to use up to four nodes at one time for a process of this length. Ultimately, we generated 504 Monte Carlo data sets with corresponding bootstraps.

4.2 Results

Figure 1 helps to illustrate the model fitting procedure by examining the fit for two separate simulated data sets. Shown are both the true and fitted 0.95 and 0.999 quantiles. These two particular instances were chosen because they reflect a good range of observed fits. The top panels show an instance where the fitted model mimics the truth quite well, where the bottom panels show some difference, but which still seems to capture the overall behavior reasonably well. We note that the fit shown for data set 2 was among the worst we observed.

Figure 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 (8). 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 2 shows the histograms for the scale parameter estimates. The positive estimates for γ_1 show that the fitted model is able to find the increasing scale with X_1, and we notice that the estimates for γ_2 are centered about 0 as they should be. The bottom panel of Figure 2 shows estimates of the shape

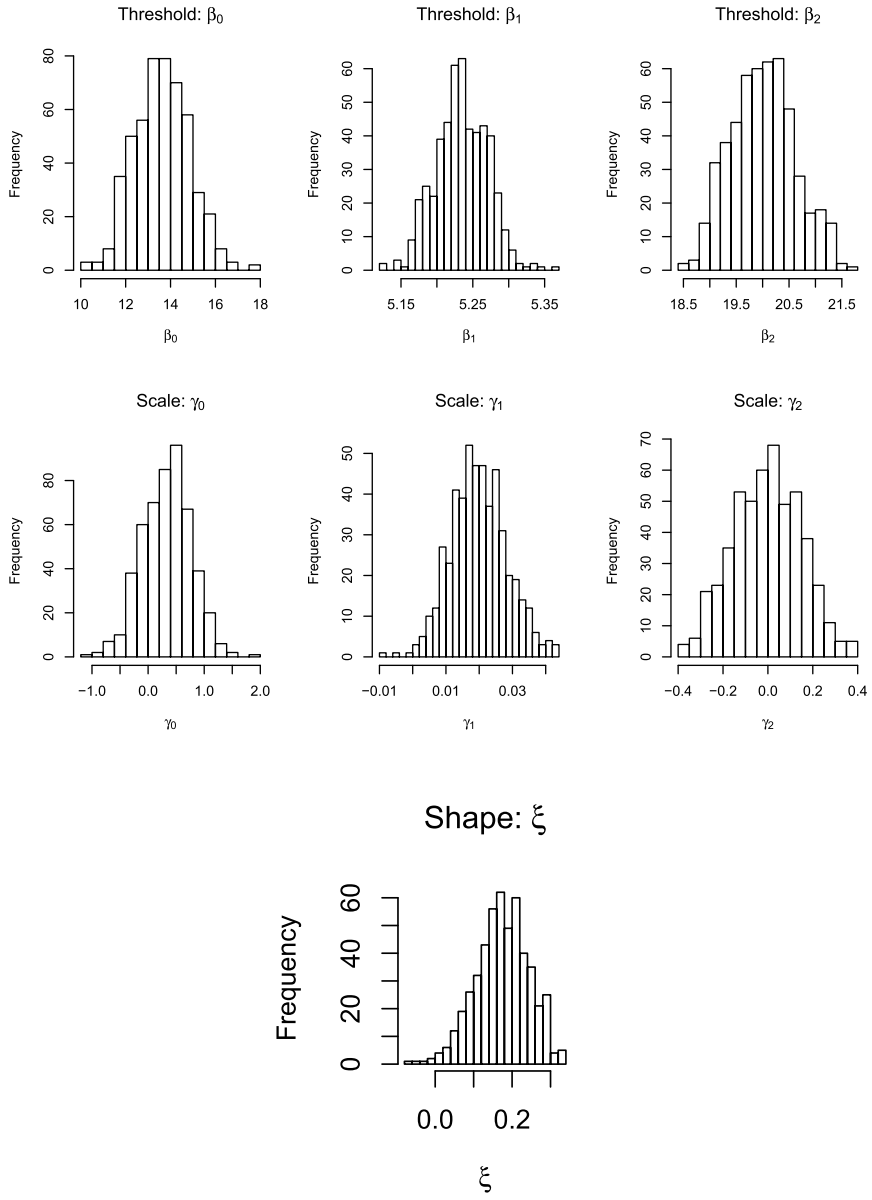

Figure 2. Histograms of threshold (top), scale (middle), and shape (bottom) parameter estimates from 504 Monte Carlo simulations.

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.

In contrast to the model parameter estimates which cannot be compared to truth due to the mismatch between generating and fitted models, we can compare the estimated quantiles to the true quantiles for specified covariate values. Histograms for the five quantiles of interest are given for two specific sets of covariates in Figure 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 our 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. We would expect only five observations above the 0.999 quantile for a data set of size $n = 5000$, and we find the performance quite reasonable.

We also assess the bootstrap method's ability to accurately account for estimation uncertainty. Table 1 shows the parameter estimates along with 95% bootstrap confidence intervals for the data set illustrated in the top panels of Figure 1. While we cannot assess coverage due to the mis-

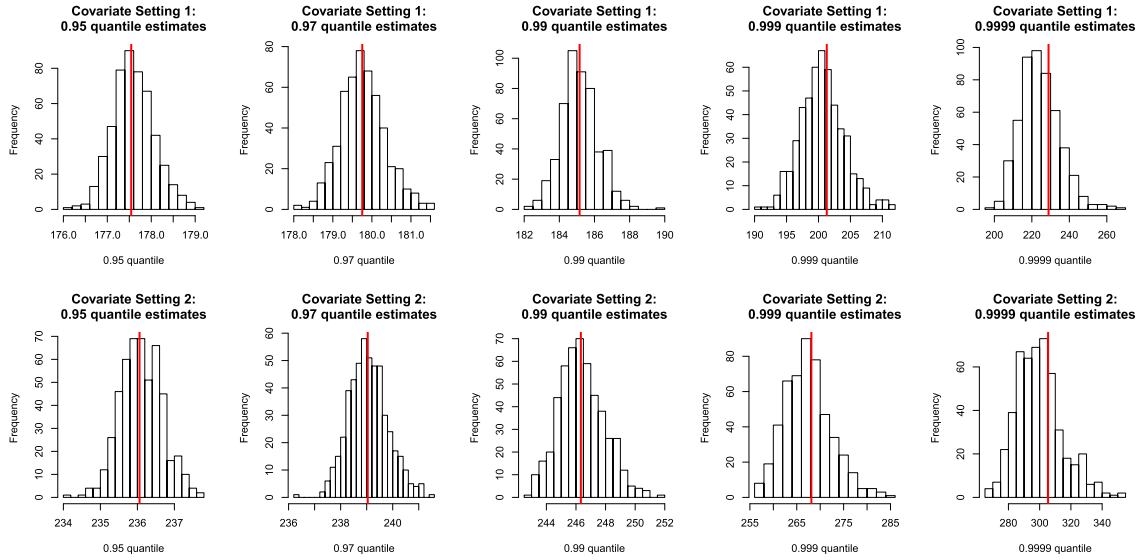

Figure 3. Histograms of quantile estimates from 504 Monte Carlo simulations evaluated for two covariate settings. The top row corresponds to $X_1 = 27.5$ and $X_2 = 1$ and the bottom row corresponds to $X_1 = 42.5$ and $X_2 = 0$. The vertical lines indicate the location of the true quantile.

Table 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 2. Quantile estimates and 95% bootstrap confidence intervals for GPD$_\tau$ and quantile regression (QR)

	Quantile	0.95: Setting 1	0.999: Setting 1	0.95: Setting 2	0.999: Setting 2
GPD$_\tau$	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

match between generating and fitted models, we do 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$_\tau$ row of Table 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$_\tau$ row of Table 3 for both covariate settings, and the coverage rate appears reasonable for the 0.95, 0.97, and 0.99 quantiles. Once again, we see that 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 2 and 3 also include QR rows, corresponding to estimates of the quantiles obtained using standard quantile regression methods. Table 2 shows that our new method and quantile regression yield similar estimates and 95% confidence intervals for the .95 quantile. Results for the .999, however, suggest that our method may be an improvement in generating confidence intervals for high quantiles, as our method's confidence intervals are narrower than those provided by quantile regression. Table 3 shows that the coverage rate of the confidence intervals are comparable for our method versus quantile regression for the .95, .97, .99, and .999 quantiles, whereas our method clearly outperforms quantile regression for the .9999 quantile.

Figure 4 plots the width of each of the 504 confidence intervals provided by our method against the confidence interval widths of quantile regression for the .95 and .999 quantiles. The plotted line shows a one-to-one relationship.

Table 3. 95% bootstrap confidence interval coverage rates and 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

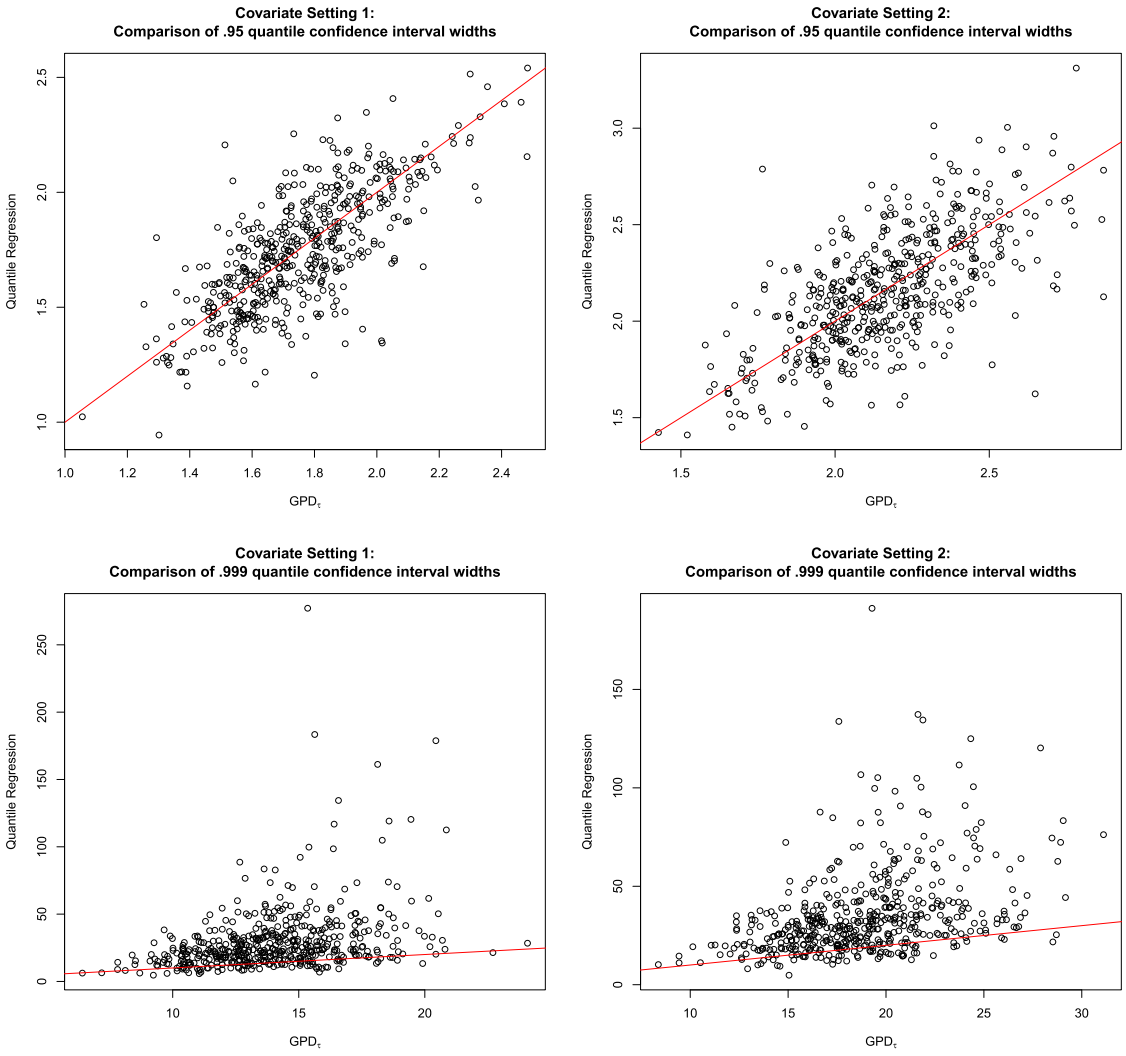

Figure 4. Comparison of 95% bootstrap confidence interval widths for GPD_τ versus quantile regression for the .95 quantiles (top row) and the .999 quantiles (bottom row). The line shows a 1:1 relationship.

The .95 quantile figures suggest that our method and quantile regression yield similar 95% confidence interval widths, whereas the .999 figures suggest that our method will produce a narrower confidence interval more often than quantile regression. Table 3 also shows the average width of the 95% confidence intervals for the different quantiles across both methods. While our 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 our method than in quantile regression, but our method also does a much better job in capturing the true .9999 quantile.

In summary, our simulation study shows that our method yields both interpretable parameter estimates and reasonable estimates for high quantiles. That the parameter estimates remain interpretable in our case of slight model mismatch is important as we turn our attention to the application, as one of the primary goals is to understand how covariates of the match scores influence the tail of the distribution. That the quantile estimates are reasonable is important for understanding approximate false discovery rates associated with some classification threshold.

5. FACIAL RECOGNITION APPLICATION

5.1 Data: non-match scores and covariates

We will fit 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 [21] to our model for $\tau = 0.05$. This data set consists of similarity scores yielded by an algorithm that compares still query and target images to each other, along with a set of covariates 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 we use, is considered to have average matching difficulty. The Bad partition contains 1,173,928 non-match pairs. To keep computational time manageable, we 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, we found it necessary to create new covariates from the ones given in many instances. Specifically, in addition to an age difference covariate, we 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, we will separate each of these categorical covariates into three binary covariates.

In addition to our newly created covariates, we will also use target and query FRIFM covariates when fitting the model. FRIFM is a continous measurement of picture quality, which is defined in Section 3.2 of [3]. FRIFM is expected to differ between any two images, so we will include the target/query FRIFM values separately in our model.

5.2 Exploratory data analysis and model choice

The empirical .999 quantile of the non-match scores is 4.093, thus this could be the classification threshold under current algorithms, regardless of covariates. The histograms in Figure 5 explore how the different covariates affect the tail and the probability of being incorrectly classified as a match. The top two rows of Figure 5 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 match are more likely to be classified as matches than those in which the categorical covariates do not match. 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. It appears the two FRIFM covariates don't have much of an effect on increasing the similarity score between two non match pairs.

We also explore the tail index parameter ξ for different covariates. We calculated the 95% confidence intervals given by fitting a GPD to data exceeding the fixed empirical .95 quantile for different subsets of the data. For all the subsets, $\hat{\xi}$ is roughly in the range from $-.1$ to .05, and there is a lot of overlap in the confidence intervals. 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. We conclude that we can adequately model the data with a common ξ parameter which is not a function of covariates. Further, if slight differences in true ξ values exist between the different groups, this will likely be compensated for by the flexibility in σ, allowing us to adequately capture tail behavior.

Based on the results of the exploratory analysis we will fit our model with $u_\tau = X\boldsymbol{\beta}$, $\sigma = X\boldsymbol{\gamma}$ 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 picture quality covariates. We once again use 0.01 as the kernel density bandwidth.

We distribute the computing differently on the cluster than we did in the simulation study. Optimization here is much more expensive than it was in our simulation study, as the sample size is much larger and we have many more parameters to estimate. We distribute the bootstrapping across nodes, running 24 bootstrap fits on each node at a time, resulting in 1008 bootstrap instances used to calculate confidence intervals.

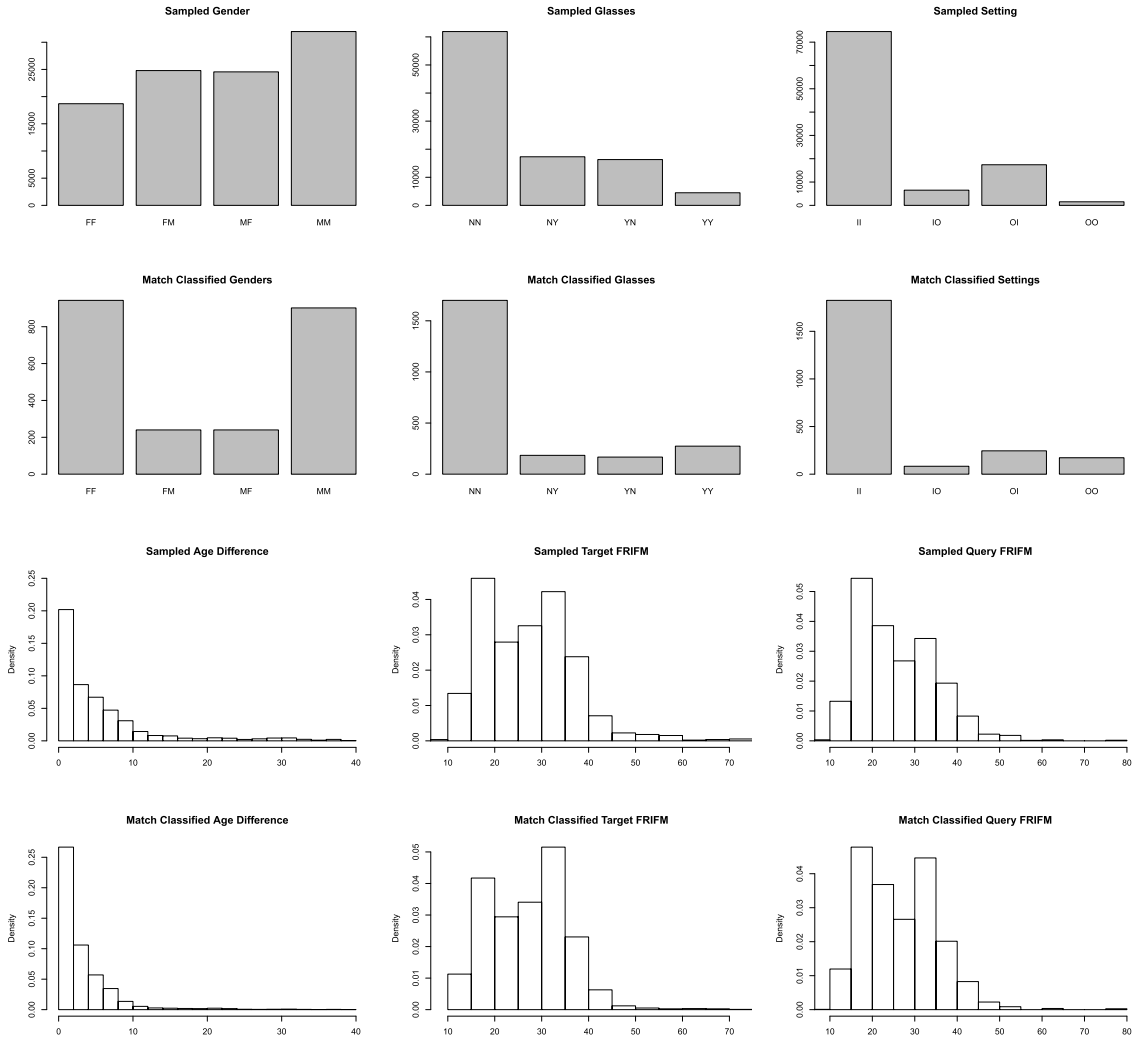

Figure 5. Top two rows are histograms showing breakdown of categorical variables in the overall sample (top row) and for pairs classified as matches (second row). Bottom two rows are histograms showing breakdown of numeric variables in the overall sample (third row) and for pairs classified as matches (bottom row).

5.3 Results

5.3.1 Parameter estimates and interpretation

The parameter estimates, along with bootstrap confidence intervals, are reported in Table 4. We 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.

We next interpret the β estimates describing how the continuous covariates effect 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.

Fewer of the scale parameter estimates are significant. Aside from γ_3, all the γ estimates which are significantly

Table 4. Parameter estimates for threshold parameters $\boldsymbol{\beta}$, scale parameters $\boldsymbol{\gamma}$, and tail 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	$(-0.21, 0.048)$	$(-0.23, -0.026)$	$(0.080, 0.49)$	$(-0.019, -0.012)$	$(-0.008, -0.000)$	$(0.004, 0.012)$	$(-0.021, 0.045)$

Table 5. Covariate vector used for each setting with corresponding probabilities of exceeding the algorithm's classification threshold

Covariate	Covariate Used						u	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

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 u_τ has larger scale than in the baseline FF case, despite the u_τ being lower for the MM case.

5.3.2 Covariate effect on tail and probability of false match classification

To get an idea of how different the tail behavior is for different covariate settings, we choose 16 covariate settings to investigate. For the first 8 settings, which are listed in Table 5, 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 9–16, 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 5 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, 2 units higher than any other that we tested. Also listed is the estimated probability that an observation with the listed covariates would have a match score exceeding overall empirical .999 quantile of 4.093. Settings 1, 3, and 4 all have an estimates for u_τ which exceed this level, meaning that our fitted model estimates that more 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.

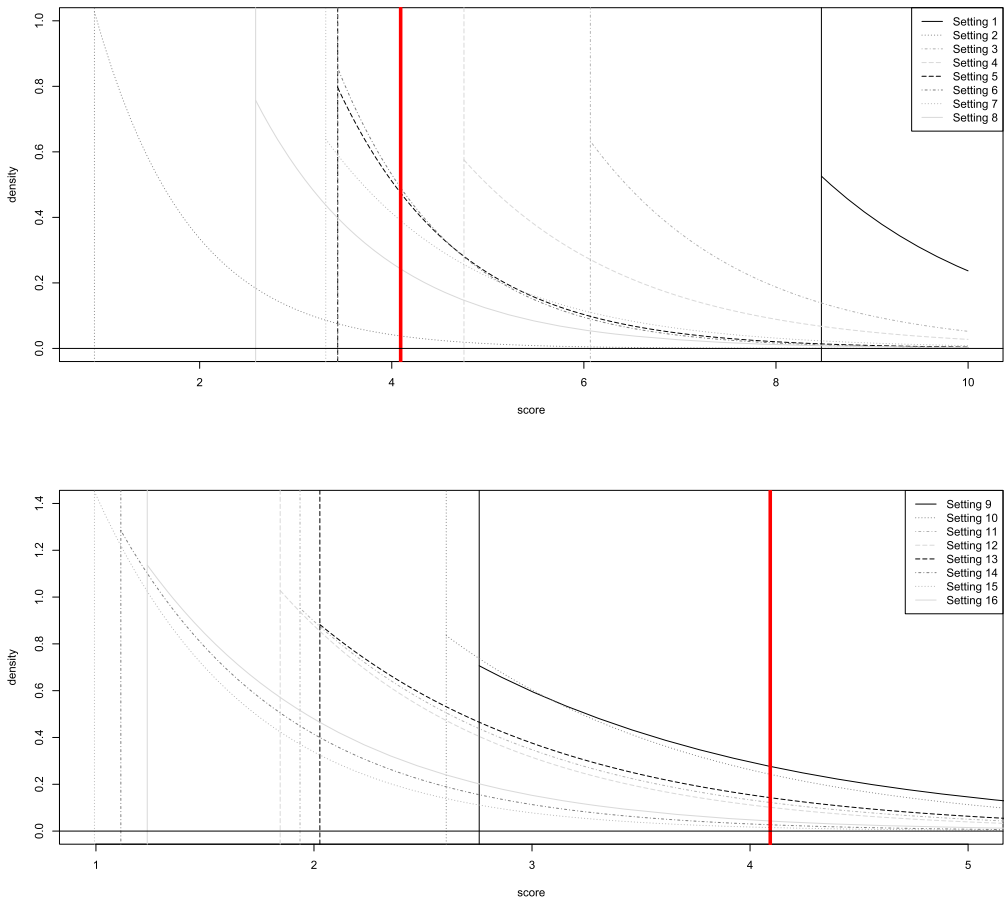

Figure 6. GPD$_\tau$ distributions for settings 1 through 8 (top) and 9 through 16 (bottom).

Figure 6 plots the estimated GPD$_\tau$ distributions for the 16 settings' values for comparison. The top panel shows settings 1–8, and the bottom panel 9–16. 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 6, such as the fact that the estimated distributions for settings 5 and 6 are very similar despite the fact that the settings themselves are 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$_\tau$ threshold gets bigger. 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. We also note that none of the GPD$_\tau$ distributions displayed in the bottom panel of Figure 6 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$_\tau$ thresholds above the match decision mark.

Table 6. Empirical .95 quantiles of the Bad partition compared to the predicted u for select settings

Covariate Setting	Bad Partition Empirical Quantile	u	95% Confidence Interval for u
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)

5.3.3 Model performance

Since we only used 100,000 of the non-match pairs in the GBU Bad partition, it is possible to compare the empirical .95 quantiles from the entire partition to our predicted u values. Table 6 compares such empirical quantiles to the predicted u 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, we are ignoring both target and query FRIFM effects, which are minimal.

In settings 7, 8, and 9, the Bad partition's .95 empirical quantile is contained within the 95% confidence interval for u. 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, our model predicts a u 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 our model does an admirable job in estimating the .95 quantile.

5.4 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.

6. DISCUSSION

We have proposed an approach to model the upper τth proportion of a distribution using a good parametric model, the GPD$_\tau$. Importantly, τ is fixed, which differs from threshold exceedance methods from extremes. Because our model is parametric, we are able to relate both u_τ and σ to covariates, and in turn to interpret how covariates influence the tail. Because our method assumes that the distribution is well approximated by a GPD above the $(1-\tau)$th quantile, it could only be used for relatively small values of τ. Inference is performed via M-estimation, and our objective function allows simultaneous estimation of u_τ and GPD$_\tau$'s parameters. Our simulation study shows that our GPD$_\tau$ is competitive with quantile regression methods for estimating high quantiles, and it may outperform quantile regression for extreme quantiles. Our facial recognition application represents a unique application with a different motivation than standard extremes analyses in disciplines such as hydrology or finance.

We demonstrate that covariates can have a dramatic effect on the upper $\tau = 0.05$ proportion of the distribution of non-match scores. In particular, the categorical covariate settings we investigated had a large effect on the location of u_τ and, to a somewhat lesser extent σ. For some of the covariate settings we investigated, we estimate that more than 5% of the observations exceed a classification threshold set at the .999 empirical quantile across covariate settings, implying that target/query pairs with these covariates will have a high false discovery rate. Future investigations could determine if there are other covariates that could prove more useful in describing the tail of the non-match distribution.

ACKNOWLEDGEMENTS

This research utilized the CSU ISTeC Cray HPC System supported by NSF Grant CNS-0923386. Daniel Cooley is partially supported by NSF Grant DMS-1243102.

Received 16 February 2016

REFERENCES

[1] Balkema, A. and De Haan, L. (1974). Residual life time at great age. *The Annals of Probability*, 2(5):792–804. MR0359049

[2] Beirlant, J., Goegebeur, Y., Segers, J., Teugels, J., Waal, D. D., and Ferro, C. (2004). *Statistics of Extremes: Theory and Applications*. Wiley, New York. MR2108013

[3] 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.

[4] Coles, S. (2001). *An Introduction to Statistical Modeling of Extreme Values*. Springer Verlag. MR1932132

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

[6] 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. MR1377836

[7] 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. MR1821820

[8] 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.

[9] Givens, G. H. and Hoeting, J. A. (2012). *Computational Statistics*, volume 710. John Wiley & Sons. MR3236433

[10] Gnedenko, B. (1943). Sur la distribution limite du terme maximum d'une série aléatoire. *The Annals of Mathematics*, 44(3):423–453. MR0008655

[11] 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. MR1841416

[12] Haberman, S. J. (1989). Concavity and estimation. *The Annals of Statistics*, 17(4):1631–1661. MR1026303

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

[14] Huber, P. J. (2011). *Robust Statistics*. Springer.

[15] Hunter, B. D. (2016). *Modeling the Upper Tail of the Distribution of Facial Recognition Non-match Scores*. PhD thesis, Department of Statistics, Colorado State University. MR3542343

[16] Koenker, R. (2005). *Quantile Regression*. Number 38. Cambridge university press. MR2268657
[17] Lehmann, E. L. and Casella, G. (1998). *Theory of Point Estimation*, volume 31. Springer. MR1639875
[18] 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. MR2785120
[19] 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.
[20] Niemiro, W. (1992). Asymptotics for m-estimators defined by convex minimization. *The Annals of Statistics*, pages 1514–1533. MR1186263
[21] 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.
[22] Pickands III, J. (1975). Statistical inference using extreme order statistics. *The Annals of Statistics*, pages 119–131. MR0423667
[23] 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. MR1041763
[24] 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.
[25] Zhao, W., Chellappa, R., Phillips, P. J., and Rosenfeld, A. (2003). Face recognition: a literature survey. *ACM computing surveys (CSUR)*, 35(4):399–458.

Brett D. Hunter
Department of Statistics
George Mason University
Fairfax, VA 22030
USA
E-mail address: bhunte11@gmu.edu

Daniel Cooley
Department of Statistics
Colorado State University
Fort Collins, CO 80523
USA
E-mail address: cooleyd@stat.colostate.edu

Geof H. Givens
Givens Statistical Solutions LLC
4913 Hinsdale Drive
Fort Collins, CO 80526
USA
E-mail address: geof@geofgivens.com

J. Ross Beveridge
Department of Computer Science
Colorado State University
Fort Collins, CO 80523
USA
E-mail address: ross@cs.colostate.edu