

## A Constrained Optimization Method for Fitting Prediction Models

Enrique Castillo<sup>1</sup>, Ali S. Hadi<sup>2</sup> and José María Sarabia<sup>3</sup>

**Abstract.** This paper proposes a new method for fitting distribution functions to sample data when prediction is of primary interest (for example, when the goal of the analysis is to estimate the quantiles, rather than the parameters). The proposed method is designed to produce fitted models that have two desirable properties: (a) the least squares line, obtained when the predicted order statistics are regressed on the observed order statistics, has an intercept of 0 and a slope of 1, and (b) the estimators satisfy an optimality criterion with respect to a given objective function (e.g., a minimum loss function or a maximum likelihood) subject to the constraints in (a). The rationale for imposing the constraints in (a) is given. The estimators are shown to be asymptotically unbiased, consistent, and normal. The method is applicable regardless of the number of unknown parameters as long as the cumulative distribution function is invertible. The method is illustrated with application to several distribution functions. A simulation study indicates that the proposed method has a good performance and outperforms other commonly used methods when a goodness-of-fit criterion is used. Finally, the method is used to fit a generalized Pareto distribution to a real-life data set.

**Key Words:** Anderson-Darling statistic; constrained least squares; constrained maximum likelihood; goodness-of-fit criteria; ideal fit property, Kolmogorov-Smirnoff statistic; least absolute value.

### 1. Introduction

Consider a random sample  $(x_1, \dots, x_n)$  drawn from a population with cumulative distribution function (cdf)  $F(x; \theta)$ , which depends on an

---

<sup>1</sup> Department of Applied Mathematics and Computational Sciences, University of Cantabria, SPAIN.

<sup>2</sup> Department of Statistics, Cornell University, USA.

<sup>3</sup> Department of Economics, Universtiy of Cantabria, SPAIN.

unknown, possibly vector-valued, parameter  $\theta$ . Of interest is fitting  $F(x; \theta)$  to the sample data. In many engineering applications, for example, the objective of data analysis is to estimate the distribution function (particularly extreme quantiles) which are useful in determining design values. For example, Van Vledder et al. (1993) indicate that the statistical analysis of extreme wave data is an important tool in the determination of the design wave height for coastal and offshore structures. The aim of their analysis consists of finding the distribution function which best fits the observed data to determine the design wave height which will be exceeded with a certain probability in the lifetime of a structure. Many other applications where the estimation of the distribution function is of interest can be found in Davidson (1984), Tiago de Oliveira (1984), D'Agostino and Stephens (1986), and Hosking and Wallis (1987), and in the references therein.

The fitting of a distribution function to the observed data usually starts with estimating the parameter  $\theta$  using an appropriate estimation method such as maximum likelihood method (MLM), least squares method (LSM), method of moments, etc. The next step depends on the goal of the analysis, which can be either parameter estimation or prediction, for example. If the goal is parameter estimation, then the quality of the estimators are judged using several criteria such as unbiasedness, consistency, efficiency, etc. On the other hand, if the goal of the analysis is prediction then the quality of the fit is judged by goodness-of-fit criteria such as maximum correlation between predicted and observed values, minimum loss function (such as the sum of absolute or squared errors), etc. These and other goodness-of-fit criteria for comparing models are discussed in Muir and El-Sharawi (1986) and Van Vledder et al. (1993). In this paper we are concerned with prediction, rather than parameter estimation, as is often the case in practice.

While using the value of a loss function is reasonable for judging the quality of the fitted model, the sole use of the correlation coefficient between predicted and observed values can actually be misleading. This is because an estimation method can produce a very good straight line relationship between predicted and observed values, yet the fitted line is far from the ideal relationship, namely, a line with an intercept of 0 and a slope of 1. We illustrate this point using the following real-life data.

This data set is known as the Kodiak data. According to Van Vledder et al. (1993), the Kodiak data set consists of all peak storm wave heights with significant wave height exceeding 6m and is shown in Table I.

These represent the 78 storms which have occurred during the period from January 1, 1956 to December 31, 1975 in Kodiak Island. One purpose of the analysis is to estimate the distribution function. The data set has been thoroughly analyzed by each of the eight authors in Van Vledder et al. (1993). We shall report the results of these and other analyses of the Kodiak data in Section 5.

One of the distribution functions that were fit to the data is the Minimal Weibull Family (MW):

$$F(x; \lambda, \sigma, \kappa) = 1 - \exp \left[ - \left( \frac{x - \lambda}{\sigma} \right)^\kappa \right], \quad (1)$$

where  $x \geq \lambda$  and  $\sigma, \kappa > 0$ . The scatter plot of predicted versus observed order statistics obtained by the least squares method is shown in Figure 1(a), where the line with intercept 0 and slope 1 is superimposed on the scatter plot. Although the correlation coefficient between predicted and observed order statistics is very high (0.989), the least squares line has an intercept of -2.056 and a slope of 1.203. As can be seen from the scatter of point, the least squares line is significantly far from the ideal line (the t-statistics for testing zero-intercept and unit slope are -13.3 and 10.0, respectively). Thus, the use of correlation coefficient alone in this case is misleading.

We now fit the three-parameter generalized Pareto distribution (GP3) to the data using the methods proposed in this paper. The GP3 family is given by

$$F(x; \lambda, \sigma, \kappa) = 1 - \left[ 1 - \left( \frac{x - \lambda}{\sigma / \kappa} \right) \right]^{1/\kappa}, \quad (2)$$

where  $\lambda, \sigma / \kappa$ , and  $\kappa$  are location, scale, and shape parameters, respectively. The GP3 is usually used to model exceedences over a threshold (see Pickands, 1975).

The scatter plot of predicted versus observed order statistics obtained by the proposed method is shown in Figure 1(b). The proposed method produced, by design, a least squares line with intercept and slope of 0 and 1, respectively. In this case the correlation coefficient is high (0.997) but not misleading.

We refer to the case where the least squares line between the predicted and observed values has a zero intercept and a unit slope as the *ideal fit property*. It is intuitively clear that, for

Significant Wave Heights Exceeding 6 m.									
6.2	8.8	6.6	6.9	7.8	6.3	11.7	7.2	7.4	9.9
8.9	7.5	7.0	6.7	9.2	6.2	6.3	8.1	6.3	7.2
6.3	6.0	8.4	6.8	9.3	6.7	6.5	7.2	8.5	6.9
6.6	9.4	8.2	6.3	7.6	7.3	8.6	7.4	7.1	6.0
6.3	6.0	7.7	6.6	6.5	6.9	7.7	8.2	6.7	7.4
6.4	6.1	7.1	6.5	8.5	8.8	9.1	8.0	6.3	9.1
6.6	6.7	7.2	10.2	7.0	10.1	7.8	6.1	6.3	8.6
7.1	10.0	8.0	6.1	8.4	7.4	8.2	8.1		

Table I: The Kodiak data.

prediction purposes, it is desirable for the fitted model to satisfy two properties: (a) the ideal fit property and (b) an optimality with respect to an appropriate objective function (e.g., a minimum loss function). In this paper we introduce a fitting method which produces a fitted model with these two properties.

In Section 2 we give the fundamentals of the proposed method. In Section 3 we describe in detail four specific cases of the proposed method and derive the asymptotic variances of the resultant estimators. The method is applied in Section 4 to several examples of distribution functions. In section 5 we give a real-life example of application. Some simulation results are presented in Section 6. A summary and concluding remarks are given in Section 7.

## 2. Fundamentals of the Proposed Estimation Method

Let  $x_{i:n}$  be the observed  $i$ th order statistic from a sample of size  $n$  and  $y_{i:n}(p_{i:n}; \theta)$  be the predicted value of the  $i$ th order statistic, where

$p_{i:n} = (i - \gamma) / (n + \delta)$  is a plotting position formula with  $0 \leq \gamma \leq 1$ ;  $0 \leq \delta \leq 1$ . In this paper we have used  $\gamma = 0.5$  and  $\delta = 0$ . This predicted value can be obtained by inverting the cdf as follows

$$y_{i:n}(p_{i:n}; \theta) = F^{-1}(p_{i:n}; \theta); \quad i = 1, \dots, n. \quad (3)$$

The predicted value  $y_{i:n}(p_{i:n}; \theta)$  depends on the unknown parameter  $\theta$ . For simplicity of notation, we write  $y_{i:n}(\theta)$  instead of  $y_{i:n}(p_{i:n}; \theta)$ . To obtain an estimate of  $y_{i:n}(\theta)$ , we need to estimate  $\theta$ . Intuitively, the relationship between the predicted and observed order statistics should be linear; that is; given the observed value  $x_{i:n}$ , we can write  $y_{i:n}(\theta)$  as

$$y_{i:n}(\theta) = \alpha + \beta x_{i:n} + \varepsilon_{i:n}, \quad (4)$$

where  $\varepsilon_{i:n}$  is a random error with mean zero and variance  $\sigma^2$ .

Alternatively, one can write (4) as

$$x_{i:n} = \gamma + \delta y_{i:n}(\theta) + \varepsilon_{i:n}. \quad (5)$$

The predicted values are regressed on the observed values in (4) and the converse is true in (5). However, the observed data are realizations of random variables, hence they can be considered fixed once they are known. That is, the fitting process is conditional on the observed data. For this reason, we use (4) to illustrate the proposed method, although it is applicable to either of the two equations in (4) and (5).

The least squares estimators of  $\alpha$  and  $\beta$  in model (4) are given by

$$\hat{\alpha} = \bar{y}(\theta) - \beta \bar{x}, \quad (6)$$

and

$$\hat{\beta} = \frac{\sum_{i=1}^n [y_{i:n}(\theta) - \bar{y}(\theta)][x_{i:n} - \bar{x}]}{\sum_{i=1}^n [x_{i:n} - \bar{x}]^2}, \quad (7)$$

where  $\bar{x}$  is the mean of  $x_{1:n}, \dots, x_{n:n}$  and  $\bar{y}(\theta)$  is the mean of  $y_{1:n}(\theta), \dots, y_{n:n}(\theta)$ .

The ideal fit property requires that the relationship between the predicted and observed order statistics can be written as *predicted = observed + random error*, which implies that in (4), we must have  $\hat{\alpha} = 0$  and  $\hat{\beta} = 1$ . Therefore, according to the ideal fit property, if  $\hat{\theta}$  is an estimate of  $\theta$ , then it must satisfy

$$y_{i:n}(\hat{\theta}) = x_{i:n} + \varepsilon_{i:n}. \quad (8)$$

From (6) and (7), the ideal fit property constraints  $\hat{\alpha} = 0$  and  $\hat{\beta} = 1$  imply that  $\hat{\theta}$  must satisfy

$$\sum_{i=1}^n y_{i:n}(\hat{\theta}) = \sum_{i=1}^n x_{i:n} \quad (9)$$

and

$$\sum_{i=1}^n x_{i:n} y_{i:n}(\hat{\theta}) = \sum_{i=1}^n x_{i:n}^2. \quad (10)$$

Thus, (9) and (10) are the necessary and sufficient conditions for the fitted model (4) to have a zero intercept and a unit slope.

In addition to satisfying the ideal fit property constraints in (9) and (10), it is desirable to choose  $\hat{\theta}$  in such a way that the predicted and observed values are as close to each other as possible. This calls for an optimization of an appropriate objective function subject to the constraints imposed by the ideal fit property. There are several possibilities for the choice of an objective function. Examples are:

1. Loss functions: Find  $\hat{\theta}$  that minimizes the loss function

$$\sum_{i=1}^n h(y_{i:n}(\hat{\theta}) - x_{i:n}),$$

subject to (9) and (10), where  $h(\cdot)$  is a suitable loss function. Examples of loss functions are:

- Constrained least squares (CLS): Minimize

$$\sum_{i=1}^n h(y_{i:n}(\hat{\theta}) - x_{i:n}) = \sum_{i=1}^n [y_{i:n}(\hat{\theta}) - x_{i:n}]^2,$$

subject to (9) and (10).

- Constrained least absolute value (CLAV): Minimize

$$\sum_{i=1}^n h(y_{i:n}(\hat{\theta}) - x_{i:n}) = \sum_{i=1}^n |y_{i:n}(\hat{\theta}) - x_{i:n}|,$$

subject to (9) and (10).

- Constrained weighted least squares (CWLS): Minimize

$$\sum_{i=1}^n w_i h(y_{i:n}(\hat{\theta}) - x_{i:n}) = \sum_{i=1}^n [F(x_{i:n}; \hat{\theta})(1 - F(x_{i:n}; \hat{\theta}))]^{-1} [y_{i:n}(\hat{\theta}) - x_{i:n}]^2,$$

subject to (9) and (10).

- Constrained weighted least absolute value (CWLAV): Minimize

$$\sum_{i=1}^n w_i h(y_{i:n}(\hat{\theta}) - x_{i:n}) = \sum_{i=1}^n [F(x_{i:n}; \hat{\theta})(1 - F(x_{i:n}; \hat{\theta}))]^{-1} |y_{i:n}(\hat{\theta}) - x_{i:n}|,$$

subject to (9) and (10).

2. Constrained maximum likelihood (CML): Maximize the likelihood function itself

subject to (9) and (10).

3. Goodness-of-fit statistics: Optimize a goodness-of-fit statistic such as:

- Constrained correlation coefficient (CCC): Maximize the correlation coefficient subject to (9) and (10).
- Constrained Kolmogorov-Smirnoff statistic: Minimize the Kolmogorov-Smirnoff statistic subject to (9) and (10).
- Constrained Anderson-Darling statistic (CAD): Minimize the Anderson-Darling statistic subject to (9) and (10).

The above loss functions and goodness-of-fit statistics are also useful for testing the goodness-of-fit of the selected families to a given set of data. The standard deviations of these statistics can be estimated by different methods (analytical, bootstrap, etc.) and then we can derive acceptance intervals for the test.

For a discussion of estimation methods such as the ML and the LS, see any mathematical statistics books such as Bickel and Doksum (1977) and Casella and Berger (1990) or Koenker and Bassett (1978). References for unconstrained optimization and for nonlinear least squares are Kennedy and Gentle (1980), Seber (1989), and Ratkowski (1990). For a discussion of Kolmogorov-Smirnoff statistic, Anderson-Darling statistic, and other goodness-of-fit criteria see, for example, Lawless (1982) and D'Agostino and Stephens (1986).

In this paper we use the CLS criterion function. Thus, the proposed fitting method can be summarized as follows: Find  $\hat{\theta}$  that minimizes

$$\sum_{i=1}^n [y_{i:n}(\theta) - x_{i:n}]^2 \quad (11)$$

subject to:

$$\sum_{i=1}^n y_{i:n}(\theta) = \sum_{i=1}^n x_{i:n}, \quad (12)$$

and

$$\sum_{i=1}^n x_{i:n} y_{i:n}(\theta) = \sum_{i=1}^n x_{i:n}^2. \quad (13)$$

In the next section we present four specific cases of this method and derive the corresponding estimators as well as their asymptotic means and variances. We conclude this section with some remarks:

1. When  $\theta$  is a scalar-valued parameter, we impose only one of the two conditions in (12) and (13) depending on whether  $\theta$  is a location or a scale parameter (see Sections 3.1 and 3.2). When  $\theta$  contains two or more parameters both conditions are imposed.
2. When  $\theta$  contains one or two parameters, the feasible region determined by the constraints is actually a single point. This single point is, then, the solution to the optimization problem in (11); see Sections 3.1-3.3. In this case the choice of optimization criteria is immaterial. On the other hand, when  $\theta$  consists of more than two parameters, as in Section 3.4, the feasible region is no longer a single point and the objective function (11) is optimized over the feasible region.

### 3. Four Specific Cases

In this section we give the details of the proposed method in four specific cases. In the first two cases,  $\theta$  contains only a single parameter; in the third,  $\theta$  contains two parameters; and in the last,  $\theta$  contains three or more parameters.

#### 3.1 Location Families

Suppose now that  $\theta$  consists of only one location parameter  $\lambda$ , say; that is,  $F(x; \lambda)$  is a location family of distributions. In this case  $F(x; \lambda)$  can be expressed as

$$F(x; \lambda) = F(x - \lambda; 0). \quad (14)$$

Thus, the predicted order statistics are given by

$$y_{i:n}(\lambda) = \lambda + F^{-1}(p_{i:n}; 0) = \lambda + w_{i:n}, \quad i = 1, \dots, n, \quad (15)$$

where  $w_{i:n} = F^{-1}(p_{i:n}; 0)$ .

To obtain an estimator of  $\lambda$  we minimize (11) subject to only the zero-intercept condition (12) because this is a one-parameter location family. Using (15), condition (12) leads to

$$\sum_{i=1}^n [\lambda + w_{i:n}] = \sum_{i=1}^n x_{i:n}. \quad (16)$$

Solving (16) for  $\lambda$ , we obtain

$$\hat{\lambda} = \bar{x} - \bar{w} \quad , \quad (17)$$

where  $\bar{w}$  is the mean of  $w_{1:n}, \dots, w_{n:n}$ . For the purpose of obtaining the mean and variance of  $\hat{\lambda}$ , we write (17) as

$$\begin{aligned} \hat{\lambda} &= \lambda + \frac{1}{n} \sum_{i=1}^n (F^{-1}(u_{i:n}; 0) - F^{-1}(p_{i:n}; 0)) \\ &= \lambda + \frac{1}{n} \sum_{i=1}^n (v_{i:n} - w_{i:n}) \quad , \end{aligned} \quad (18)$$

where  $v_{i:n} = F^{-1}(u_{i:n}; 0)$  and  $u_{i:n}; i = 1, \dots, n$ , are the order statistics of a sample of size  $n$  drawn from a uniform  $U[0,1]$ . This leads to the mean and variance of  $\hat{\lambda}$  which are given by

$$\begin{aligned} E(\hat{\lambda}) &= \lambda + E(x_0) - \frac{1}{n} \sum_{i=1}^n F^{-1}(p_{i:n}; 0), \\ V(\hat{\lambda}) &= \frac{1}{n} V(F^{-1}(U; 0)) = \frac{1}{n} V(X_0), \end{aligned} \quad (19)$$

where  $X_0 \sim F(x; 0)$  and  $U \sim U(0,1)$ . From (19) we have

$$\lim_{n \rightarrow \infty} E(\hat{\lambda}) = \lambda; \quad \lim_{n \rightarrow \infty} V(\hat{\lambda}) = 0. \quad (20)$$

Therefore,  $\hat{\lambda}$  is an asymptotically unbiased and consistent estimator of  $\lambda$ . The mean square error (MSE) of  $\hat{\lambda}$  is

$$MSE(\hat{\lambda}) = \frac{V(X_0)}{n} + \left[ E(X_0) - \frac{1}{n} \sum_{i=1}^n F^{-1}(p_{i:n}; 0) \right]^2. \quad (21)$$

Furthermore, if  $0 < V(X_0) < \infty$ , by the central limit theorem, we have

$$\sqrt{n}(\hat{\lambda} - \lambda) \xrightarrow{D} N(0, V(X_0)), \quad (22)$$

as  $n \rightarrow \infty$ .

### 3.2 Scale Families

Here we assume that the family  $F(x; \theta)$  depends on one scale parameter  $\sigma$ , say. In this case, the cdf can be expressed as

$$F(x; \theta) = F(x/\sigma; 1), \quad (23)$$

and the predicted order statistics can be written as

$$y_{i:n}(\sigma) = \sigma F^{-1}(p_{i:n}; 1) = \sigma w_{i:n}; \quad i = 1, \dots, n, \quad (24)$$

where now  $w_{i:n} = F^{-1}(p_{i:n}; 1)$ . To estimate the scale parameter  $\sigma$ , we minimize (11) subject to only the unit-slope condition (13) because the cdf is a one-parameter scale family. Using (24), condition (13) leads to

$$\sum_{i=1}^n x_{i:n} \sigma w_{i:n} = \sum_{i=1}^n x_{i:n}^2. \quad (25)$$

Solving (25) for  $\sigma$  we obtain the estimator of  $\sigma$ ,

$$\hat{\sigma} = \frac{\sum_{i=1}^n x_{i:n}^2}{\sum_{i=1}^n x_{i:n} w_{i:n}} \quad (26)$$

To obtain the asymptotic distribution of  $\hat{\sigma}$ , let  $v_{i:n} = F^{-1}(u_{i:n}; 1)$  and write  $\hat{\sigma}$  as

$$\hat{\sigma} = \sigma \frac{\sum_{i=1}^n [v_{i:n}]^2}{\sum_{i=1}^n v_{i:n} w_{i:n}}. \quad (27)$$

We also make use of the fact that, as  $n \rightarrow \infty$ ,

$$\sqrt{n}((u_{1:n}, \dots, u_{n:n}) - \mu) \xrightarrow{D} N_n(0, \Sigma), \quad (28)$$

where,  $\mu = (p_{1:n}, p_{2:n}, \dots, p_{n:n})^T$  and  $\Sigma$  is a variance-covariance matrix with elements

$$\sigma_{i:j} = \begin{cases} p_{i:n}(1-p_{j:n}), & \text{if } i \leq j, \\ p_{j:n}(1-p_{i:n}), & \text{otherwise.} \end{cases} \quad (29)$$

Now consider the function

$$g(p_{1:n}, \dots, p_{n:n}) = \sigma \frac{\sum_{i=1}^n v_{i:n}^2}{\sum_{i=1}^n v_{i:n} w_{i:n}}. \quad (30)$$

Taking derivatives with respect to  $p_{i:n}$  we obtain

$$\frac{\partial g}{\partial p_{i:n}} = \sigma \frac{v_{i:n} v'_{i:n} \left[ 2 \sum_j v_{j:n} w_{i:n} - v_{i:n} \sum_j v_{i:n}^2 \right]}{\left[ \sum_j v_{j:n} w_{i:n} \right]^2}, \quad (31)$$

where  $v'_{i:n} = \partial v_{i:n} / \partial u_{i:n}$ . Let  $G$  be the column vector with elements  $\partial g / \partial p_{i:n}$ ,  $i = 1, 2, \dots, n$ , then as  $n \rightarrow \infty$  we have

$$\sqrt{n}(\hat{\sigma} - \sigma) \xrightarrow{D} N(0, G^T \sum G). \quad (32)$$

Therefore,  $\hat{\sigma}$  is an asymptotically unbiased and consistent estimator of  $\sigma$ .

### 3.3 Location Scale Families

Let us now consider the two-parameter family

$$F(x; \lambda, \sigma) = F((x - \lambda) / \sigma; 0, 1), \quad (33)$$

where  $\lambda$  and  $\sigma$  are location and scale parameters, respectively. Then, the predicted order statistics can be written as

$$y_{i:n}(\lambda, \sigma) = \lambda + \sigma F^{-1}(p_{i:n}; 0, 1) = \lambda + \sigma w_{i:n}, \quad (34)$$

where  $w_{i:n} = F^{-1}(p_{i:n}; 0, 1)$ . The zero-intercept and unit-slope conditions are

$$\lambda + \sigma \bar{w} = \bar{x}, \quad (35)$$

$$\lambda \sum_{i=1}^n x_{i:n} + \sigma \sum_{i=1}^n x_{i:n} w_{i:n} = \sum_{i=1}^n x_{i:n}^2, \quad (36)$$

which lead to the estimators

$$\hat{\lambda} = \bar{x} - \hat{\sigma} \bar{w}$$

$$\hat{\sigma} = \frac{\sum_{i=1}^n x_{i:n}^2 - n\bar{x}^2}{\sum_{i=1}^n x_{i:n} w_{i:n} - n\bar{x} \bar{w}}. \quad (37)$$

To derive the asymptotic means and variances of and  $\hat{\sigma}$  it is convenient to write (37) as

$$\hat{\lambda} = \lambda + \sigma(\bar{v} - \bar{w}),$$

$$\hat{\sigma} = \sigma \frac{\sum_{i=1}^n v_{i:n}^2 - n(\bar{v})^2}{\sum_{i=1}^n v_{i:n} w_{i:n} - n\bar{v} \bar{w}} = \frac{S_v^2}{S_{v,w}}, \quad (38)$$

where  $S_v^2$  is the variance of  $v_{1:n}, \dots, v_{n:n}$  and  $S_{v,w}$  is the covariance of  $v_{1:n}, \dots, v_{n:n}$  and  $w_{1:n}, \dots, w_{n:n}$ . The asymptotic joint distribution of both estimators is normal and its parameters can be obtained, as before, by the  $\delta$ -method. Let

$$G = \begin{bmatrix} \partial \hat{\lambda} / \partial p_{1:n} & \dots & \partial \hat{\lambda} / \partial p_{n:n} \\ \partial \hat{\sigma} / \partial p_{1:n} & \dots & \partial \hat{\sigma} / \partial p_{n:n} \end{bmatrix}. \quad (39)$$

As  $n \rightarrow \infty$  we have

$$\sqrt{n}((\hat{\lambda}, \hat{\sigma})^T - (\lambda, \sigma)^T) \xrightarrow{D} N((0,0)^T, G \Sigma G^T). \quad (40)$$

### 3.4. Multiparameter Families

Finally, we consider the case where  $\theta$  consists of three or more parameters. Suppose that  $\theta = (\lambda, \sigma, \kappa)$ , where  $\lambda$  and  $\sigma$  are location and scale parameters, respectively, and  $\kappa$  is possibly a vector-valued parameter. The cdf can then be written as

$$F(x; \lambda, \sigma, \kappa) = F((x - \lambda) / \sigma; 0, 1, \kappa),$$

from which it follows that the predicted values are

$$y_{i:n}(\lambda, \sigma, \kappa) = \lambda + \sigma F^{-1}(p_{i:n}; 0, 1, \kappa) = \lambda + \sigma w_{i:n}(p_{i:n}, \kappa),$$

where  $w_{i:n} = F^{-1}(p_{i:n}; 0, 1, \kappa)$ . Thus, (11), (13) become: Minimize

$$\sum_{i=1}^n (y_{i:n}(\hat{\lambda}, \hat{\sigma}, \hat{\kappa}) - x_{i:n})^2 \quad (41)$$

subject to

$$\hat{\lambda} + \hat{\sigma} \bar{w}(\hat{\kappa}) = \hat{x}, \quad (42)$$

and

$$\hat{\lambda} \sum_{i=1}^n x_{i:n} + \hat{\sigma} \sum_{i=1}^n x_{i:n} w_{i:n}(\hat{\kappa}) = \sum_{i=1}^n x_{i:n}^2, \quad (43)$$

which lead to the estimators

$$\hat{\lambda} = \bar{x} - \hat{\sigma} \bar{w}(\hat{\kappa}),$$

$$\hat{\sigma} = \frac{\sum_{i=1}^n x_{i:n}^2 - n\bar{x}^2}{\sum_{i=1}^n x_{i:n} w_{i:n}(\hat{\kappa}) - n\bar{x} \bar{w}(\hat{\kappa})}, \quad (44)$$

where  $\hat{\kappa}$  minimizes (41). The estimators of  $\lambda$  and  $\sigma$  can be expressed as

$$\hat{\lambda} = \lambda + \sigma[\bar{v}(\kappa) - \bar{w}(\kappa)], \quad (45)$$

and

$$\hat{\sigma} = \sigma \frac{\sum_{i=1}^n w_{i:n}^2(\kappa) - n(\bar{w}(\kappa))^2}{\sum_{i=1}^n v_{i:n}(\kappa) w_{i:n}(\kappa) - n\bar{v}(\kappa)\bar{w}(\kappa)} = \sigma \frac{S_{v(\kappa)}^2}{S_{v(\kappa)w(\kappa)}}. \quad (46)$$

Again, the asymptotic variances can be obtained using the  $\delta$ -method.

#### 4. Some Applications

In this section we illustrate the proposed method by its application to several families of distributions. We note, however, that these are only a few examples of applications and that the method is applicable to other families of distributions.

##### 4.1 The Two-Parameter Uniform Distribution

The cdf of the two-parameter uniform distribution  $U(\lambda, \lambda + \sigma)$ , is

$$F(x; \lambda, \sigma) = \frac{x - \lambda}{\sigma}; \quad \lambda \leq x \leq \lambda + \sigma, \quad (47)$$

from which it follows that the predicted order statistics can be expressed as

$$y_{i:n}(\lambda, \sigma) = \lambda + \sigma p_{i:n}. \quad (48)$$

Since this is a two-parameter family, the two constraints (12) and (13) give rise to the estimators

$$\hat{\lambda} = \bar{x} - \hat{\sigma} \bar{p},$$

$$\hat{\sigma} = \frac{\sum_{i=1}^n x_{i:n}^2 - n\bar{x}^2}{\sum_{i=1}^n x_{i:n} p_{i:n} - n\bar{x} \bar{p}}, \quad (49)$$

where  $\bar{p}$  is the mean of  $p_{1:n}, \dots, p_{n:n}$ . For comparison, to be used later, we remind the reader that the maximum likelihood estimates (MLE) of  $\lambda$  and  $\sigma$  are

$$\hat{\lambda} = x_{1:n}; \quad \hat{\sigma} = x_{n:n} - x_{1:n}. \quad (50)$$

## 4.2 The Two-Parameter Exponential Distribution

The cdf of the two-parameter exponential distribution is

$$F(x; \lambda, \sigma) = 1 - \exp\left\{-\left(\frac{x - \lambda}{\sigma}\right)\right\}; \quad x \geq \lambda. \quad (51)$$

Due to its important properties such as lack of memory, stability with respect to truncations, or constant expected residual life, the exponential distribution is used in several fields such as reliability theory, stochastic models, etc.

From (51), the predicted order statistics can be written as

$$y_{i:n}(\lambda, \sigma) = \lambda - \sigma \log(1 - p_{i:n}). \quad (52)$$

Thus, the two constraints in (12) and (13) give the estimators

$$\hat{\lambda} = \bar{x} + \hat{\sigma} \frac{1}{n} \sum_{i=1}^n \log(1 - p_{i:n}),$$

$$\hat{\sigma} = \frac{\sum_{i=1}^n x_{i:n}^2 - n\bar{x}^2}{\sum_{i=1}^n (1 - p_{i:n})\bar{x} - \sum_{i=1}^n x_{i:n} \log(1 - p_{i:n})}. \quad (53)$$

For comparison purposes, the MLE of  $\lambda$  and  $\sigma$  are

$$\hat{\lambda} = x_{1:n}; \quad \hat{\sigma} = \bar{x} - x_{1:n}. \quad (54)$$

### 4.3 The Two-Parameter Generalized Pareto Distribution

With this example we illustrate how the proposed method can be used to estimate parameters even though they are not location and scale parameters.

The cdf of the two-parameter generalized Pareto distribution  $GP2(\sigma, \kappa)$  is

$$F(x; \sigma, \kappa) = 1 - \left[ 1 - \left( \frac{\kappa x}{\sigma} \right) \right]^{1/\kappa}, \quad (55)$$

where  $\sigma$  and  $\kappa$  are scale and shape parameters, respectively. Note that when  $\kappa = 1$ , the  $GP2$  becomes a uniform distribution.

The Generalized Pareto Distribution (GPD) was introduced by Pickands (1975) to model exceedences over a threshold. It has since been used by many authors to model data, such as annual maximum floods, tensile strength data, etc.

The predicted order statistics are

$$y_{i:n}(\sigma, \kappa) = \sigma \left[ 1 - (1 - p_{i:n})^\kappa \right] / \kappa = \sigma w_{i:n}(\kappa), \quad (56)$$

where  $w_{i:n}(\kappa) = \left[ 1 - (1 - p_{i:n})^\kappa \right] / \kappa$ . According to (12)-(13), we get

$$\sigma \sum_{i=1}^n w_{i:n}(\kappa) = \sum_{i=1}^n x_{i:n}, \quad (57)$$

and

$$\sigma \sum_{i=1}^n w_{i:n}(\kappa) x_{i:n} = \sum_{i=1}^n x_{i:n}^2. \quad (58)$$

From (57), we can write  $\sigma$  as a function of  $\kappa$  as follows:

$$\sigma(\kappa) = \frac{\sum_{i=1}^n x_{i:n}}{\sum_{i=1}^n w_{i:n}(\kappa)}. \quad (59)$$

Substituting (59) in (58), we eliminate  $\sigma$  and obtain

$$\frac{\sum_{i=1}^n x_{i:n}}{\sum_{i=1}^n w_{i:n}(\kappa)} - \frac{\sum_{i=1}^n x_{i:n}^2}{\sum_{i=1}^n w_{i:n}(\kappa) x_{i:n}} = 0. \quad (60)$$

which is an equation of only one variable  $\kappa$ . Thus, an estimate  $\hat{\kappa}$  can be obtained by solving (60) in  $\kappa$  using the bisection method, for example. The estimate  $\hat{\kappa}$  can then be substituted in (59) to obtain an estimate of  $\sigma$ ,

$$\hat{\sigma} = \frac{\sum_{i=1}^n x_{i:n}}{\sum_{i=1}^n w_{i:n}(\hat{\kappa})}. \quad (61)$$

#### 4.4 The Three-Parameter Generalized Pareto Distribution

The cdf of the three-parameter generalized Pareto distribution  $GP3(\lambda, \sigma, \kappa)$  is defined in (2). Like the  $GP2$ , the  $GP3$  becomes a uniform distribution when  $\kappa = 1$ . The predicted order statistics are

$$y_{i:n}(\lambda, \sigma, \kappa) = \lambda + \sigma [1 - (1 - p_{i:n})^\kappa] / \kappa = \lambda + \sigma w_{i:n}(\kappa), \quad (62)$$

where  $w_{i:n}(\kappa) = (1 - (1 - p_{i:n})^\kappa) / \kappa$ . According to (11)-(13), the quantity to be minimized is

$$Q = \sum_{i=1}^n [\hat{\lambda} + \hat{\sigma} w_{i:n}(\hat{\kappa}) - x_{i:n}]^2, \quad (63)$$

subject to

$$\hat{\lambda} + \hat{\sigma} \bar{w}(\hat{\kappa}) = \bar{x}, \quad (64)$$

and

$$\hat{\lambda} \sum_{i=1}^n x_{i:n} + \hat{\sigma} \sum_{i=1}^n x_{i:n} w_{i:n}(\hat{\kappa}) = \sum_{i=1}^n x_{i:n}^2, \quad (65)$$

From (62) and (65), it follows that

$$\hat{\lambda} = \bar{x} - \hat{\sigma} \bar{w}(\hat{\kappa}), \quad (66)$$

and

$$\hat{\sigma} = \frac{\sum_{i=1}^n x_{i:n}^2 - n\bar{x}^2}{\sum_{i=1}^n x_{i:n} w_{i:n}(\hat{\kappa}) - n\bar{x} \bar{w}(\hat{\kappa})}. \quad (67)$$

Substitution of (66) and (67) in (63), makes Q a function of only one variable  $\hat{\kappa}$ . Thus we first find  $\hat{\kappa}$  which minimizes (63), then substitute this value of  $\hat{\kappa}$  in (66) and (67) to obtain  $\hat{\lambda}$  and  $\hat{\sigma}$ .

## 5. Example of Real-Life Data

We now return to the Kodiak data set (given in Table I and described in Section 1) and use it to illustrate the above methodology. The data set has been thoroughly analyzed by each of the eight authors in Van Vledder et al. (1993) and by Castillo, Hadi, and Sarabia (1995).

The authors in Van Vledder et al. (1993) estimate distribution functions and use some goodness-of-fit criteria as the Kolmogorov-Smirnoff, Anderson-Darling, Chi-square, correlation coefficient, etc., to decide which of the candidate families of distributions is the most adequate. The adequate selection of distribution functions for obtaining accurate wave height values associated with given return periods is important because the cost of the work is very sensitive to these values.

The following models have been fit to the data:

### 1. The Truncated Weibull Family (TW):

$$F(x; \lambda, \sigma, \kappa) = 1 - \exp \left[ - \left( \frac{x - \lambda}{\sigma} \right)^\kappa + \left( \frac{u - \lambda}{\sigma} \right)^\kappa \right], \quad (68)$$

where  $x \geq u$ , and  $\sigma, \kappa \geq 0$ .

## 2. The Gumbel Family (G):

$$F(x; \lambda, \sigma) = \exp \left[ -\exp \left( \frac{x - \lambda}{\sigma} \right) \right], \quad (69)$$

where  $-\infty < x < \infty$  and  $\sigma > 0$ .

## 3. The Minimal Weibull Family (MW) defined in (1).

## 4. The Three-Parameter Generalized Pareto Family $GP3$ defined in (2).

Some of the authors in Van Vledder et al. (1903) fit more than one model and use more than one method of fitting. The first 12 models in Table II are the results of the models proposed by the authors in Van Vledder et al. (1993). Model 13 is proposed by Castillo, Hadi, and Sarabia (1995), who fit the  $GP3$  to the data using the elemental percentile method (EPM). Models 14 and 15 are obtained by applying the CLS and CLAV criteria to the  $GP3$ .

To judge the overall goodness-of-fit, we use the average absolute error,

$$AAE = n^{-1} \sum_{i=1}^n |x_{i:n} - \hat{x}_{i:n}|, \quad (70)$$

and the average squared error,

$$ASE = n^{-1} \sum_{i=1}^n (x_{i:n} - \hat{x}_{i:n})^2, \quad (71)$$

where  $\hat{x}_{i:n}$  is the  $i$ th predicted order statistic.

This example shows that the sole use of the correlation coefficient  $r$  between the predicted and observed order statistics as a goodness-of-fit criterion can be misleading. This fact can be seen in Table II by comparing  $r$ ,  $AAE$ ,  $ASE$ , the intercept  $\hat{\alpha}$ , the slope  $\hat{\beta}$ , the t-statistic  $t(\alpha = 0)$  for testing  $\alpha = 0$ , and the t-statistic  $t(\beta = 1)$  for testing  $\beta = 1$ . All models produce large values of  $r$ , yet the models produce substantially different least square

lines when the predicted order statistics are regressed on the observed order statistics. Many of these methods give least squares lines with intercepts and a slopes significantly different from 0 and 1, respectively (see, for example, Models 5, 6, 11 and 12 in Table II). The models also produce substantially different values of  $AAE$  and  $ASE$ . Thus, while a small value of  $r$  indicates a bad fit, a large value does not necessarily indicate a good fit. These phenomena have also been observed in the simulated data of Section 2. By design, the proposed method gives lines with zero-intercepts and unit-slopes. Despite the fact that the method is a constrained optimization and the other methods are not constrained, it produces the smallest values of  $AAE$  and  $ASE$ .

In addition, using the bootstrap method, we have estimated the mean and standard deviations of the  $ASE$  and  $AAE$  statistics for the  $CLS$  and  $CLAV$  cases, respectively, and we have obtained the following results:

- **CLS method** :  $E[ASE] = 0.0146$ ;  $VAR[ASE] = 0.0082^2$ ,
- **CLAV method** :  
 $E[AAE] = 0.0830$ ;  $VAR[AAE] = 0.0232^2$ ,

which when compared with the values  $ASE = 0.009$  and  $AAE = 0.061$  in rows 14 and 15, respectively, in Table II, one can conclude that the  $GP3$  model is adequate (cannot be rejected) for the Kodiak data.

## 6. Simulations

In this section we give the results of some simulations to illustrate the quality of the proposed estimators. To judge the overall goodness-of-fit, we use the average absolute error  $AAE$  and the average square error  $ASE$  defined in (70) and (71), respectively. Note that in (70) and (71), one can use the actual percentile from the simulated distribution rather than the sample order statistics. However, we chose the sample order statistics because in practical situations we do not know the true distribution function and we have to use goodness-of-fit measure based on the observed sample, to compare two or more different distributions from the same or different families.

We have performed four simulation experiments: One for each of the distributions discussed in Section 4. The results, which are based on 1000 replications, are shown in Table III, for the Uniform and Exponential

distributions; and in Tables IV and V for the two generalized Pareto distributions.

For the Uniform  $U[0,1]$  and Exponential  $E[0,1]$  distributions, the results indicate the following:

- Both the AAE and the ASE are better for the proposed method than for the MLE in the case of the Uniform distribution.
- For the Exponential distribution the ASE is better for the proposed method than for the MLE, while the AAEs are comparable.

The MLM are not satisfactory in the case of the generalized Pareto distribution for both the two- and three-parameter cases (see, e.g., DuMouchel (1983), Davidson (1984), Smith (1984, 1985) and Hosking and Wallis (1987)). The probability weighted moments (PWM) is proposed by Hosking, Wallis, and Wood (1985) to estimate the parameters and quantiles of the generalized Pareto distribution. We compare the proposed method with the PWM method. Table IV shows the simulation results based on 1000 replications for the  $GP2$  distribution. The results indicate that the proposed method out performs the PWM in terms of AAE and ASE.

For the  $GP3$  case, the results are shown in Table V. The results are seen to converge to zero for positive values of  $\kappa$ . For negative values, the AAE and ASE values become stable as the sample size increases. This is not surprising because in this case, the variance is infinity, we measure errors in the random variable scale and larger percentiles are involved.

## 7. Summary and Concluding Remarks

A method for fitting distribution functions to sample data, based on two desirable properties: (a) the least squares line, obtained when the predicted order statistics are regressed on the observed order statistics, has an intercept of zero and a slope of 1, and (b) the estimators satisfy an optimality criterion with respect to a given objective function (e.g., a minimum loss function), has been described. The estimation methods have been shown to be consistent and asymptotically normal distributed. Simulation studies have been presented showing the good properties of the method when compared to other well recognized methods such as the maximum likelihood and the

probability weighted moments methods. An example of application has been used to illustrate the method and show how it can be used for testing the quality of different families.

	Mod el	Method	$\hat{\lambda}$	$\hat{\sigma}$	$\hat{\kappa}$	r	AAE	ASE	$\hat{\alpha}$	$\hat{\beta}$	t-test ( $\alpha = 0$ )	t-test ( $\beta = 1$ )
1	TW	MLM	0	5.854	2.693	0.997	0.06	0.01	0.13	0.98	2.0	-2.0
2	G	LSM	8.050	1.040	-	0.992	1.15	1.35	0.56	1.08	4.6	4.8
3	G	MOM	8.060	1.129	-	0.992	1.21	1.53	-0.07	1.17	-0.5	9.8
4	MW	LSM	5.81	1.86	1.4	0.996	0.07	0.01	-0.01	1.00	-0.1	0.1
5	MW	MLM	6.7	1.578	1.386	0.996	0.64	0.44	1.70	0.86	25.5	-16.2
6	G	MLM	7.665	0.776	-	0.992	0.61	0.44	2.08	0.80	22.8	-16.3
7	MW	MLM	5.9	1.74	1.324	0.995	0.08	0.01	0.04	0.99	0.5	-0.6
8	MW	LSM	5.81	1.86	1.4	0.996	0.07	0.01	-0.01	1.00	-0.1	0.1
9	G	MOM	8.02	1.06	-	0.992	1.13	1.31	0.39	1.10	3.1	6.0
10	MW	LSM	5.805	1.862	1.4	0.996	0.07	0.01	-0.02	1.00	-0.2	0.2
11	G	LSM	7.616	0.876	-	0.992	0.62	0.41	1.31	0.91	12.7	-6.8
12	MW	LSM	4.145	3.188	2	0.989	0.54	0.39	-2.06	1.20	-13.3	10.0
13	GP3	EPM	5.967	2.015	0.349	0.996	0.07	0.02	0.38	0.94	5.0	-5.7
14	GP3	CLS	5.994	1.919	0.271	0.997	0.06	0.01	0.00	1.00	0.0	0.0
15	GP3	CLAV	5.969	1.989	0.297	0.997	0.06	0.01	0.00	1.00	0.0	0.0

MLM = Maximum Likelihood Method

LSM = Least Squares Method

MOM = Method of Moments

EPM = Elemental Percentile Method

CLAV = Constrained Least Absolute value

CLS = Constrained Least Squares

Table II: Results from fitting several models to the Kodiak data.

Sample Size	Statistic	Uniform Distribution		Exponential Distribution	
		CLS	MLE	CLS	MLE
50	<i>AAE</i>	0.0287	0.0431	0.1179	0.1150
	<i>ASE</i>	0.0013	0.0031	0.0373	0.0494
100	<i>AAE</i>	0.0202	0.0311	0.0908	0.0858
	<i>ASE</i>	0.0007	0.0016	0.0252	0.0327
200	<i>AAE</i>	0.0146	0.0222	0.0665	0.0623
	<i>ASE</i>	0.0003	0.0008	0.0152	0.0196
500	<i>AAE</i>	0.0091	0.0141	0.0429	0.0400
	<i>ASE</i>	0.0001	0.0003	0.0075	0.0093

Table III: Simulation results based on 1000 replicates from the Uniform  $U[0,1]$  and the Exponential  $E[0,1]$  distributions.

Sample Size	Statistic	$\kappa = -0.4$		$\kappa = 0.4$	
		CLS	PWM	CLS	PWM
50	<i>AAE</i>	0.2357	0.2812	0.0511	0.0527
	<i>ASE</i>	0.4452	2.2890	0.0055	0.0062
100	<i>AAE</i>	0.1975	0.2128	0.0363	0.0377
	<i>ASE</i>	0.3871	1.5521	0.0029	0.0034
200	<i>AAE</i>	0.1783	0.1770	0.0256	0.0267
	<i>ASE</i>	0.4718	1.5311	0.0014	0.0017
500	<i>AAE</i>	0.1379	0.1232	0.0165	0.0173
	<i>ASE</i>	0.4112	0.9408	0.0006	0.0008

Table IV: Simulation results based on 1000 replications from the Generalized Pareto two-parameter distribution  $GP2(1, \kappa)$ .

		$\kappa = -0.4$	$\kappa = 0.4$
Sample Size	Statistic	CLS	CLS
50	AAE	0.2594	0.0450
	ASE	0.2685	0.0035
100	AAE	0.2353	0.0326
	ASE	0.2512	0.0019
200	AAE	0.2138	0.0233
	ASE	0.2563	0.0010
500	AAE	0.2018	0.0150
	ASE	0.2979	0.0004

Table V: Simulation results based on 1000 replications from the Generalized Pareto Three-parameter distribution  $GP3(0,1,\kappa)$ .

## Appendix

## References

- Bickel, P.J., and Doksum, K. A. (1977) *Mathematical Statistics: Basic Ideas and Selected Topics*. San Francisco: Holden Day.
- Casella, G., and Berger, R. L. (1990) *Statistical Inference*. Belmont, CA: Wadsworth.
- Castillo, E., Hadi, A. S., and Sarabia, J. M. (1995) "Statistical Analysis of Extreme Waves." *Offshore Mechanics and Arctic Engineering* Vol. II: Safety and Reliability, 33-40.
- D'Agostino, R.B., and Stephens, M. A. (1986) *Nonlinear Regression*. New York: Marcel Dekker.
- Davidson, A. C. (1984) "Modeling Excesses Over High Thresholds, with an Application", In: *Statistical Extremes and Applications* (ed.) J. Tiago de Oliveira, NATO ASI Series, D. Reidel, Dordrecht, 461-482.
- DuMouchel, W. (1983) "Estimating the Stable Index  $\alpha$  in Order to Measure Tail Thickness." *The Annals of Statistics* 11: 1019-1036.
- Hosking, J. R. M. and Wallis, J. R. (1987) "Parameter and Quantile Estimation for the Generalized Pareto Distribution." *Technometrics* 29: No. 3, 339-349.
- Hosking, J. R. M., Wallis, J. R. and Wood, E. F (1985) "Estimation of the Generalized Pareto Distribution by the Method of Probability-Weighted Moments" *Technometrics* 27: 251-261.
- Kennedy, W. J., Jr. and Gentle, J. E. (1980) *Statistical Computing*. New York: Marcel Dekker.
- Koenker, R. and Bassett, G. (1978), "Regression Quantiles." *Econometrica* pp. 46, 33-50.
- Lawless, J. F. (1982) *Statistical Models for Lifetime Data*. New York: John Wiley and Sons.
- Muir, L. R. and El-Shaarawi, A. H. (1986) "On the Calculation of Extreme Wave Heights: A Review." *Ocean Engineering* 13: 93-118.

- Pickands, J. (1975), "Statistical Inference Using Extreme Order Statistics." *The Annals of Statistics* 3: 119-131.
- Ratkowski, D. A. (1990) *Handbook of Nonlinear Regression Models*. New York: Marcel Dekker.
- Seber, G. A. F. (1989) *Nonlinear Regression*. New York: John Wiley and Sons.
- Smith, R. L. (1984) "Threshold Methods for Sample Extremes" In: *Statistical Extremes and Applications* (ed.) J. Tiago de Oliveira), NATO ASI Series, D. Reidel, Dordrecht, 621-638.
- Smith, R. L. (1985) "Maximum Likelihood Estimation in a Class of Nonregular Cases." *Biometrika* 72: 67-90.
- Tiago de Oliveira, J. (1984) *Statistical Extremes and Applications*. NATO ASI Series, D. Reidel, Dordrecht.
- Van Vledder, G., Goda, Y., Hawkes, P., Mansard, E., Martin, M. J., Mathiesen, M., Peltier, E. and Thompson, E. (1993) "Case Studies of Extreme Wave Analysis: A Comparative Analysis." *Proceedings of Waves '93 Conference*: 978-992.

