Improving the Estimation of Density Functions Using Constraints

September 30, 2005

Enrique Castillo[1], **Roberto Mínguez**[2], **Ali S. Hadi**[3] **and Carmen Castillo**[4]

[1]*Department of Applied Mathematics and Computational Sciences,*

University of Cantabria, Spain.

[2] *Department of Applied Mathematics,*
University of Castilla-La Mancha, 13071 Ciudad Real, Spain,

and a Fulbright Scholar, Department of Social Statistics, Cornell University, USA.

[3] *Department of Mathematics, The American University in Cairo, Egypt.*

[4] *Grupo de Puertos y Costas, CEAMA, University of Granada, Spain.*

Abstract

This paper presents a method for estimating density functions that improves other existing methods by expressing the estimation issue as an optimization problem and incorporating extra information not considered by standard estimation methods. We include here three types of such information: bounds for the cumulative distribution function (cdf), bounds for the quantiles, and any restrictions on the parameters, as for example, those imposed by the support of the random variable under consideration. The method is quite general and can be applied to many estimation methods such as the maximum likelihood, the method of moments, the least squares, the least absolute values, and the minimax methods. The performances of the obtained estimates from several families of distributions are investigated for the maximum likelihood and the method of moments, using simulations. The simulation results show that for small sample sizes important gains can be achieved with respect to the case where the above information is ignored.

Key Words: Least absolute value, Least squares, Maximum likelihood, Method of moments, Minimax, Normal distribution, Uniform distribution, Weibull distribution.

1 Introduction

Consider a family of probability density functions (pdfs), $\{f(x;\boldsymbol{\theta})|\boldsymbol{\theta} \in \boldsymbol{\Theta}\}$, with the corresponding family of cdfs, $\{F(x;\boldsymbol{\theta})|\boldsymbol{\theta} \in \boldsymbol{\Theta}\}$, with support $\mathcal{S}$, where $\boldsymbol{\theta} = (\theta_1, \theta_2, \ldots, \theta_m)$ is an m-dimensional vector of parameters belonging to a parameter space $\boldsymbol{\Theta}$. We wish to estimate the parameter $\boldsymbol{\theta}$ based on an iid random sample $\boldsymbol{x} = (x_1, x_2, \ldots, x_n)$ drawn from a member of such a family. Several classical as well as robust methods are available. These include the maximum likelihood, the method of moments in its variants, the least squares, least absolute value, minimax, etc. See, for example, the classical books by Cramér (1946), Mann et al. (1974), and Bickel and Doksum (2000), or any standard or recent statistical inference/regression books such as Arthanari and Dodge (1993), Hogg and Craig (2000), Mukhopadhyay (2000), Casella and Berger (2001), Wackerly et al. (2001), Rao (2002), Miller and Miller (2003), and Silvapulle and Sen (2004).

The above methods have good asymptotic properties, but for small samples the qualities of the estimates degenerate and their performances become very poor. Small samples are frequently encountered in practice especially when modeling extreme events data, where the data are minima or maxima of monthly or yearly samples, hence the sample sizes are usually small. In such cases, one could use domain knowledge to improve the estimates. Common pieces of knowledge that can be easily incorporated into the estimation method are bounds on quantiles or cumulative distribution functions (cdfs).

In addition to incorporating bounds on the cdf and quantiles in the estimation process, an important feature of the proposed method is that it allows incorporating any restrictions on the parameter space. This occurs e.g. when the support of the random variable depends on the unknown parameters, as for example, when we have a uniform $U(\alpha, \beta)$ population, where we must have $\alpha \leq \min(x_1, x_2, \ldots, x_n)$ and $\beta \geq \max(x_1, x_2, \ldots, x_n)$. Incorporating such restrictions is not always possible in standard estimation methods.

Domain experts and/or statistical users who are familiar with the variables being analyzed can easily give bounds that will improve the performance of any estimators, even though these bounds may not be so precise. As a motivating example, we use the car speed data found in Castillo et al. (2005), p. 17, and given in Table 1. The data are the maximum car speeds registered at a mountain road for 200 dry weeks. Since the observed data are maxima, the maximal Weibull distribution is used to model these data and estimate the corresponding cumulative distribution function (cdf). The cdf of the maximal Weibull random variable is

$$F(x; \lambda, \delta, \kappa) = \exp\left[-\left(\frac{\lambda - x}{\delta}\right)^{1/\kappa}\right], \quad \text{if } x \leq \lambda, \kappa > 0, \tag{1}$$

This cdf depends on three parameters: a location parameter, λ, a scale parameter, δ, and a shape parameter κ.

Table 1: Car speed data and subsample used in the analysis.

51.6	56.6	63.3	61.1	66.8	54.5	60.9	56.9	46.3	51.0
59.5	64.8	65.7	56.1	**63.4**	**67.0**	**69.0**	**60.2**	**64.5**	**57.4**
62.6	**59.3**	**53.3**	**61.2**	**65.3**	**64.8**	**61.6**	**59.0**	**65.1**	**65.6**
60.6	**62.6**	**61.3**	**62.1**	61.1	67.8	54.9	54.4	67.4	63.2
60.0	53.1	51.8	62.6	60.1	65.0	68.9	68.2	66.1	64.7
57.1	63.3	57.9	55.7	66.9	42.0	55.4	63.3	55.7	48.7
61.6	57.6	58.5	64.0	59.0	49.5	52.4	57.3	59.4	56.7
53.4	55.0	55.0	64.4	56.2	65.0	57.0	56.6	59.5	62.7
63.1	62.9	61.9	59.7	43.2	52.0	48.1	58.2	62.3	64.2
58.8	61.8	61.7	64.3	66.8	66.6	65.4	59.6	56.2	60.5
57.0	66.7	55.9	57.8	55.0	58.8	53.3	67.6	65.2	64.0
62.7	58.8	63.3	61.7	61.7	57.8	60.8	57.2	56.8	47.4
51.4	57.1	60.2	63.1	57.3	52.2	62.4	63.4	60.8	54.4
57.5	65.0	67.4	60.8	55.7	55.4	50.0	59.9	65.2	67.7
61.6	64.2	55.9	63.5	54.3	62.7	65.7	61.8	55.9	57.3
49.7	56.2	54.4	48.1	64.7	61.0	59.2	60.5	57.3	53.1
55.8	62.1	64.4	56.0	64.5	60.5	55.2	62.2	58.8	67.3
59.3	62.4	61.2	59.5	55.5	63.4	67.4	56.1	61.9	64.8
54.7	53.2	50.3	56.9	59.3	52.1	60.1	42.6	68.0	62.6
64.5	57.1	65.8	64.7	51.3	51.9	62.3	53.8	67.2	65.2

As can be seen in Figure 1, the maximum likelihood method gives a very good estimate of the cdf. This is because the sample size here is relatively large ($n = 200$). To illustrate that the maximum likelihood can give a very bad fit when the sample size is small, we use a subsample of size 20 consecutive maxima (given in bold face in Table 1) to estimate the maximal Weibull cdf. The obtained cdf is also shown in Figure 1. Note that although this curve fits the 20 points rather well, it is very far from the curve based on all $n = 200$ points. This usually happens when one has small samples; in those cases, the chances of getting a non-representative sample and consequently a model far from the real one are large.

The maximum likelihood estimates can be substantially improved if one incorporates important information not present in the data set. For example, suppose that an engineer, based on the road and weather conditions, concludes that the 10% percentile cannot exceed 54 km/hour. Incorporating this simple information into the estimation problem can lead to substantial improvements. In this paper we present a method that improves the standard estimation methods by incorporating these types of information into the

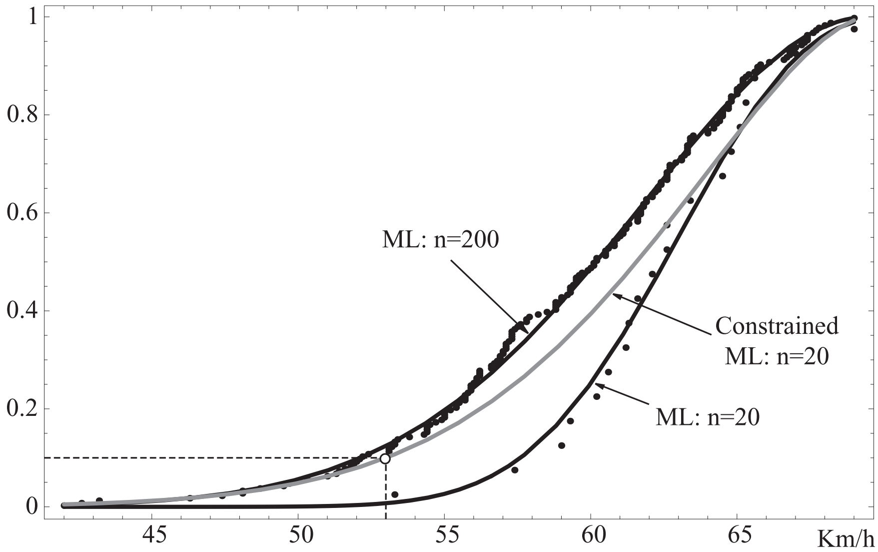

Figure 1: The Weibull models fitted to the whole sample and a subsample of size $n = 20$, and the constrained Weibull model after constraining the 10% percentile to 54 Km/hour.

estimation process. For example, as can be seen in Figure 1, the cdf obtained using the constrained maximum likelihood is much closer to the cdf based on $n = 200$ points than the one obtained by the unconstrained maximum likelihood using $n = 20$ points. This illustrates how the model estimated from small samples can be easily corrected by incorporating information not available in the data. These performance improvements are confirmed by the simulation results given later in the paper.

The proposed method is similar to Bayesian statistics in the sense that information not contained in the data is used in the estimation process. But in Bayesian statistics, the information is in the form of a prior distribution on the parameters and gives a posterior distribution on the parameters. The proposed method, however, incorporate the information above which is provided by either the domain expert or the data analyst.

The rest of the paper is organized as follows. Section 2 presents the proposed method in its general form, where it is shown that it can be used to improve many of the available estimation methods. In Section 3 the method is applied to the maximum likelihood method and in Section 4, to the method of moments. Three families of distributions (normal, uniform and Weibull) are used as illustrative examples. Section 5 extends the method to the case of infinitely many continuous bounds. Some conclusions are given in

Section 6.

2 The Proposed Method

Let $\hat{\boldsymbol{\theta}}$ be an estimator of a parameter vector $\boldsymbol{\theta}$ obtained using a given estimation method (such as the maximum likelihood or the method of moments) based on an iid random sample $\boldsymbol{x} = (x_1, x_2, \ldots, x_n)$ drawn from a family of pdfs, $\{f(x; \boldsymbol{\theta}) | \boldsymbol{\theta} \in \boldsymbol{\Theta}\}$, with support $\mathcal{S}$. Here we propose a method that improves the estimator $\hat{\boldsymbol{\theta}}$ by incorporating extra information in the estimation process. As we mentioned earlier, these information can be in the form of either bounds on the cdf, the quantile function or restrictions on the parameters imposed by the support of the random variable under consideration. The proposed method involves three simple steps:

1. Express the given estimation method as an optimization problem.

2. Add the available extra information as constraints.

3. Obtain the new estimators by solving the constrained optimization problem.

Step 1: Optimization Problems. The first step in the procedure consists of expressing the given estimation method as an optimization problem. We note here that many estimation methods are already stated as optimization problems. For example, the maximum likelihood, the least squares, the least absolute values, and minimax method. Other methods, such as the method of moments, can be expressed as optimization problems even though they are initially formulated as the solutions of a set of equations. This step is illustrated by its application to the following estimation methods:

1. The Maximum Likelihood Method. This method is already stated as an optimization problem, where the estimators are obtained by maximizing the log likelihood function:

$$\underset{\boldsymbol{\theta} \in \boldsymbol{\Theta}}{\text{Maximize}} \;\; \ell(\boldsymbol{\theta} | x_1, x_2, \ldots, x_n) = \sum_{i=1}^{n} \log f(x_i; \boldsymbol{\theta}). \tag{2}$$

2. The Method of Moments. Let

$$g_a^r(\boldsymbol{\theta}) = E(X - a(\boldsymbol{\theta}))^r, \quad r = 1, 2, \ldots, m,$$

be a set of moments of order r of the random variable X with respect to the point $a(\boldsymbol{\theta})$. For example, $g_0^r(\boldsymbol{\theta}) = E(X^r)$ is the rth population moment with respect to the origin and $g_\mu^r(\boldsymbol{\theta}) = E(X - \mu)^r$ is the rth moment with respect to the mean $\mu = E(X)$. Similarly, let

$$h_a^r(\boldsymbol{x}) = n^{-1} \sum_{i=1}^{n} (x_i - a(\boldsymbol{x}))^r, \quad r = 1, 2, \ldots, m,$$

be the corresponding set of sample moments of order r with respect to the point $a(\boldsymbol{x})$. For example, $h_0^r(\boldsymbol{x}) = n^{-1} \sum_{i=1}^{n} x_i^r$ is the rth sample moment with respect to the origin and

$$h_{\bar{x}}^r(\boldsymbol{x}) = n^{-1} \sum_{i=1}^{n} (x_i - \bar{x})^r$$

is the rth sample moment with respect to the sample mean $\bar{x} = n^{-1} \sum_{i=1}^{n} x_i$.

The method of moments estimates are obtained by solving the system of equations:

$$g_a^r(\hat{\boldsymbol{\theta}}) = h_a^r(\boldsymbol{x}), \quad r = 1, 2, \ldots, m. \tag{3}$$

The method of moments can be equivalently formulated as the optimization problem:

$$\underset{\boldsymbol{\theta} \in \boldsymbol{\Theta}}{\text{Minimize}} \sum_{r=1}^{\ell} \left[\frac{g_a^r(\boldsymbol{\theta})}{h_a^r(\boldsymbol{x})} - 1 \right]^2, \quad \ell \geq m. \tag{4}$$

In fact this is a generalization of the standard method of moments because with this formulation the number of moments, ℓ, can be larger than the number of parameters m to be estimated. In this paper we have not included examples where $\ell > m$ because our intention is to compare our results with the classical method of moments where $l = m$.

3. The Least Squares Method. The random variables X_i can be expressed as

$$X_i = \mu(\boldsymbol{\theta}) + \varepsilon_i, \quad i = 1, 2, \ldots, n,$$

where $\mu(\boldsymbol{\theta})$ is the mean of X, which depends on $\boldsymbol{\theta}$, and ε_i are random errors assumed to be independent with mean zero and constant variance. Then the least squares estimators of $\boldsymbol{\theta}$ are obtained by minimizing the sum of squared errors, that is,

$$\underset{\boldsymbol{\theta} \in \boldsymbol{\Theta}}{\text{Minimize}} \sum_{i=1}^{n} \varepsilon_i^2 = \sum_{i=1}^{n} [x_i - \mu(\boldsymbol{\theta})]^2 . \tag{5}$$

This method gives more attention to large errors.

4. The Least Absolute Values Method. The least absolute values estimators of $\boldsymbol{\theta}$ are obtained by minimizing the sum of the absolute errors instead of their squares, i.e.:

$$\underset{\boldsymbol{\theta}\in\boldsymbol{\Theta}}{\text{Minimize}} \sum_{i=1}^{n} |\varepsilon_i| = \sum_{i=1}^{n} |x_i - \mu(\boldsymbol{\theta})|, \tag{6}$$

thus treating all errors equally. Due to the presence of the non-differentiable absolute-value function, it is difficult to solve (6) using standard techniques. It can be shown, however, that the problem in (6) is equivalent to the following optimization problem (see, e.g., Arthanari and Dodge (1993), Castillo et al. (2001)):

$$\underset{\boldsymbol{\theta},\, \varepsilon_i}{\text{Minimize}} \;\; Z_{LAV} = \sum_{i=1}^{n} \varepsilon_i \tag{7}$$

subject to

$$x_i - \mu(\boldsymbol{\theta}) \le \varepsilon_i, \quad i = 1, \ldots, n, \tag{8}$$

$$\mu(\boldsymbol{\theta}) - x_i \le \varepsilon_i, \quad i = 1, \ldots, n. \tag{9}$$

We note that it is not necessary to add the set of constraints (added by some authors) $\varepsilon_i \ge 0$, because they are implied by (8) and (9).

5. The Minimax Method. The minimax method estimates the parameters $\boldsymbol{\theta}$ by minimizing the maximum error, that is,

$$\underset{\boldsymbol{\theta}}{\text{Minimize}} \;\max_i |x_i - \mu(\boldsymbol{\theta})|, \tag{10}$$

where now the function is non-differentiable due to the absolute and the maximum functions in it. Fortunately, the problem in (10) is equivalent to the optimization problem:

$$\underset{\boldsymbol{\theta},\, \varepsilon}{\text{Minimize}} \quad \varepsilon \tag{11}$$

subject to

$$x_i - \mu(\boldsymbol{\theta}) \;\le\; \varepsilon, \quad i = 1, \ldots, n, \tag{12}$$

$$\mu(\boldsymbol{\theta}) - x_i \;\le\; \varepsilon, \quad i = 1, \ldots, n, \tag{13}$$

where again the constraint ($\varepsilon \geq 0$) is implied by (12) and (13).

Note that the objective functions in (2), (4), and (5) have no constraints, but each of the objective functions in (7) and (11) has two sets of constraints.

Step 2: Incorporating Constraints. Once a given estimation method is stated as an optimization problem such as those in (2), (4), (5), (7)-(9), and (11)-(13), in Step 2 the available information is incorporated by imposing the following constraints on the optimization problem:

$$t_i^{(1)} \leq F(u_i; \boldsymbol{\theta}) \leq t_i^{(2)}, \qquad i = 1, 2, \ldots, p, \tag{14}$$

$$x_j^{(1)} \leq F^{-1}(v_j; \boldsymbol{\theta}) \leq x_j^{(2)}, \qquad j = 1, 2, \ldots, q, \tag{15}$$

$$g(\theta_k) \leq c_k(\boldsymbol{x}), \qquad k = 1, 2, \ldots, m, \tag{16}$$

where u_i are given values of the random variable X; v_j are given probability (percentile) values; $t_i^{(1)}$ and $t_i^{(2)}$ are bounds on the cdf at u_i; $x_j^{(1)}$ and $x_j^{(2)}$ are given bounds on the quantiles at v_j; and $c_k(\boldsymbol{x})$ are functions of the observed data. Note that the constraints in (16) represent restrictions on the parameters as, for example, those imposed by the cdf in cases where the support of the random variable depends on the parameters. These constraints are included only in cases where such restrictions exist. For example, when $X \sim U(\alpha, \beta)$, we must have $\alpha \leq x_{(1)}$ and $\beta \geq x_{(n)}$, where $x_{(1)}$ and $x_{(n)}$ are the sample minimum and maximum, respectively.

Note that the above constraints can be stated in a simpler form as

$$\begin{aligned} p_i &\leq F(z_i; \boldsymbol{\theta}), & i &= 1, 2, \ldots, r, \\ F(w_j; \boldsymbol{\theta}) &\leq q_j, & j &= 1, 2, \ldots, s, \\ g(\theta_k) &\leq c_k(\boldsymbol{x}), & k &= 1, 2, \ldots, m, \end{aligned} \tag{17}$$

where p_i and q_j are given probability values; and z_i and w_j are given values for the random variable.

Step 3: Solving the Constrained Problem. After expressing the estimation problem as a constrained optimization problem in Steps 1 and 2, we need to solve the constrained optimization problem to find the improved estimators. For this purpose, one can use any standard optimization package. The proposed method has been implemented in GAMS (General Algebraic Modeling System) (see Castillo et al. (2001)), but other alternatives could be used. GAMS is a very powerful software system especially designed for solving optimization problems (linear, non-linear, integer and mixed integer) of small to very large size. All the examples in this paper have been solved using the generalized reduced gradient method implemented in the CONOPT solver, which has good convergence properties for models with highly nonlinear constraints (for more details see, for example, Vanderplaats (1984), Bazaraa et al. (1993), and Drud (1996)).

Clearly, the method is quite general and can be applied to any estimation method that can be expressed as an optimization problem.

Assessing the Performance of Estimators. The proposed method is applied below to improve the maximum likelihood and method of moments estimators in three families of distributions (normal, uniform, and Weibull). The performance of the proposed method is compared with the corresponding standard methods and is assessed using simulated data generated from these three distributions. In each case, 10,000 samples of sizes $n = 5, 10, 20, 50, 100$ from the given population were simulated, and the performance of the methods based on the bias and the mean square error over the 10,000 samples, were analyzed.

3 Constrained Maximum Likelihood Methods

In this section, we apply the proposed method to the maximum likelihood method using the bounds for the cdf and/or the quantiles in (17). Then, the resultant constrained maximum likelihood method can be stated as:

$$\underset{\boldsymbol{\theta} \in \boldsymbol{\Theta}}{\text{Maximize}} \sum_{i=1}^{n} \log f(x_i; \boldsymbol{\theta}) \tag{18}$$

subject to the constraints in (17).

3.1 Asymptotic Distribution

It is clear that if the constraints are not asymptotically active, the asymptotic distribution coincides with that of the unconstrained maximum likelihood estimates. Otherwise, the asymptotic distribution of the parameter estimates and the dual variables can be obtained using the results of Aitchison and Silvey (Aitchison and Silvey, 1958). To this end, we write the active constraints in (17), which asymptotically coincide with probability one, as

$$h_j(\boldsymbol{\theta}) = 0; \quad j = 1, 2, \ldots, t. \tag{19}$$

The Karush-Kuhn-Tucker (KKT) conditions for the problem (18)–(19) are:

$$\boldsymbol{\ell}_{\boldsymbol{\theta}} + \boldsymbol{\lambda}^T \mathbf{H}_{\boldsymbol{\theta}} = \mathbf{0}, \tag{20}$$

$$h_j(\boldsymbol{\theta}) = 0, \quad j = 1, 2, \ldots, t, \tag{21}$$

where $\mathbf{H}_{\boldsymbol{\theta}}$ is an $s \times t$ matrix with elements $h_{ij} = \dfrac{\partial h_j}{\partial \theta_i}$, and $\boldsymbol{\ell}_{\boldsymbol{\theta}}$ is an $s \times 1$ vector with elements $\ell_i = \dfrac{\partial L}{\partial \theta_i}$.

Then, we have (see Aitchison and Silvey (Aitchison and Silvey, 1958)):

$$\begin{pmatrix} \hat{\boldsymbol{\theta}} - \boldsymbol{\theta}_0 \\ \frac{1}{n}\hat{\boldsymbol{\lambda}}_n \end{pmatrix} = \begin{pmatrix} \mathbf{P} & \mathbf{Q} \\ \mathbf{Q}^T & \mathbf{R} \end{pmatrix} \begin{pmatrix} \frac{1}{n}\boldsymbol{\ell}_{\boldsymbol{\theta}} \\ \mathbf{0} \end{pmatrix}, \tag{22}$$

where $\boldsymbol{\theta}_0$ is the true value, and

$$\begin{pmatrix} \mathbf{P} & \mathbf{Q} \\ \mathbf{Q}^T & \mathbf{R} \end{pmatrix} = \begin{pmatrix} \mathbf{B} & -\mathbf{H}_{\boldsymbol{\theta}} \\ -\mathbf{H}_{\boldsymbol{\theta}}^T & \mathbf{0} \end{pmatrix}^{-1}, \tag{23}$$

where $\mathbf{B}$ is an $s \times s$ matrix with elements $b_{ij} = -\dfrac{1}{n}\dfrac{\partial^2 L}{\partial \theta_i \partial \theta_j}$.

Asymptotically, we have

$$\begin{pmatrix} \hat{\boldsymbol{\theta}} - \boldsymbol{\theta}_0 \\ \frac{1}{n}\hat{\boldsymbol{\lambda}}_n \end{pmatrix} \sim N\left(\begin{pmatrix} \frac{1}{n}\mathbf{P}\boldsymbol{\ell}_{\boldsymbol{\theta}} \\ \frac{1}{n}\mathbf{Q}^T\boldsymbol{\ell}_{\boldsymbol{\theta}} \end{pmatrix}, \frac{1}{n}\begin{pmatrix} \mathbf{P} & \mathbf{0} \\ \mathbf{0} & -\mathbf{R} \end{pmatrix} \right). \tag{24}$$

In addition, we can test the null hypothesis H_0: θ_0 satisfies (19), using the acceptance region

$$-\frac{1}{n}\hat{\boldsymbol{\lambda}}^T \mathbf{R}^{-1} \boldsymbol{\lambda} \leq k, \quad \Pr[\chi_r^2 \leq k] = 0.95. \tag{25}$$

In addition, for $n \to \infty$, the system (20)–(21) gives the asymptotic solution when the parent population does not satisfy the constraints.

To illustrate the gain in the precision of the estimates using the above constraints, we consider here the normal, uniform, and Weibull distributions.

3.2 Application to the Normal Distribution

Suppose that we have a sample of size n coming from a normal population, $N(\mu, \sigma)$, and that we want to estimate the mean μ and the standard deviation σ. The log likelihood function to be maximized is

$$\ell(\alpha, \beta | x_1, \ldots, x_n) = -\frac{n}{2} - n \ln \sigma - \frac{1}{2\sigma^2} \sum_{i=1}^{n} (x_i - \mu)^2 , \tag{26}$$

which gives the following maximum likelihood estimates of μ and σ:

$$\hat{\mu} = \bar{x} \quad \text{and} \quad \hat{\sigma} = \sqrt{n^{-1} \sum_{i=1}^{n} (x_i - \bar{x})^2}. \tag{27}$$

To obtain the constrained maximum likelihood estimates, we first note that the support of X in the normal case does not depend on the parameters, hence the constraints in (16) do not exist in this case. Now, one way to obtain the CDF and quantile bounds, consists of using results from the literature on extreme value theory. It is well-known that the normal distribution belongs to domains of attractions of the Gumbel distribution for maxima and minima (see, for example, Galambos (1987), Coles (2001), and Castillo et al. (2005)). The cdf of the Gumbel distribution for maxima is

$$H(x; \lambda_{\max}, \delta_{\max}) = \exp\left(-\exp\left(\frac{\lambda_{\max} - x}{\delta_{\max}}\right)\right), \quad -\infty \leq x \leq \infty, \tag{28}$$

and the cdf of the Gumbel distribution for minima is

$$L(x; \lambda_{\min}, \delta_{\min}) = 1 - \exp\left(-\exp\left(\frac{x - \lambda_{\min}}{\delta_{\min}}\right)\right), \quad -\infty \leq x \leq \infty, \tag{29}$$

where $\lambda_{\max}, \delta_{\max}, \lambda_{\min},$ and $\delta_{\min}$ are the parameters associated with the maximal and minimal Gumbel distributions, respectively. Since the normal distribution belongs to the maximal and minimal Gumbel domain of attraction, then the normal cdf $F(z_i; \mu, \sigma)$ can be bounded by:

$$H(z_i; \lambda_{\max}, \delta_{\max}) \leq F(z_i; \mu, \sigma), \qquad i = 1, 2, \ldots, r, \tag{30}$$

$$F(w_j; \mu, \sigma) \leq L(w_j, \lambda_{\min}, \delta_{\min}), \qquad j = 1, 2, \ldots, s. \tag{31}$$

Thus, the constrained maximum likelihood (CML) estimators in this case are obtained by maximizing the log likelihood function in (26) subject to the constraints in (30)–(31).

As an illustrative example, we consider the following sample of size $n = 10$:

$$\begin{array}{rrrrr} 1.0044, & -0.7523, & 0.7560, & 1.2004, & 0.8435, \\ -1.4305, & 2.0993, & 0.6734, & 0.1690, & 1.6492, \end{array} \tag{32}$$

generated from $N(0, 1)$. As a particular example of the constraints in (30)–(31), we set $r = s = 6$, $\boldsymbol{z} = \boldsymbol{w} = (-3, -2, -1, 1, 2, 3)$ (i.e., plus and minus one, two and three standard deviations from the mean), $\lambda_{\max} = -0.05, \delta_{\max} = 1, \lambda_{\min} = 0$ and $\delta_{\min} = 1$. These values have been selected for illustration purposes, however, for real world applications, bounds should be selected based on experience and knowledge of the problem under study. With these values, the constraints in (30) and (31) become

$$\begin{array}{lcccl} 1 \times 10^{-8} & \leq & F(-3; \mu, \sigma) & \leq & 0.04857, \\ 0.00089 & \leq & F(-2; \mu, \sigma) & \leq & 0.12658, \\ 0.07534 & \leq & F(-1; \mu, \sigma) & \leq & 0.30780, \\ 0.70473 & \leq & F(\ 1; \mu, \sigma) & \leq & 0.93401, \\ 0.87921 & \leq & F(\ 2; \mu, \sigma) & \leq & 0.99938, \\ 0.95375 & \leq & F(\ 3; \mu, \sigma) & \leq & 1.00000. \end{array} \tag{33}$$

The upper and lower limiting points belonging to the maximal and minimal Gumbel distributions (dashed lines) are shown in Figure 2.

We used the standard ML and the CML with the constraints (33) to estimate the parameters and obtained the following estimates:

$$\begin{array}{lll} \text{ML:} & \hat{\mu} = 0.6214, & \hat{\sigma} = 1.0047, \\ \text{CML:} & \hat{\mu} = 0.4552, & \hat{\sigma} = 1.0126. \end{array} \tag{34}$$

Note that the CML estimate is closer to the $N(0, 1)$ distribution function than the ML estimate (see Figure 2).

To investigate the performance of the constrained maximum likelihood method, a simulation study is conducted. The results, based on 10,000 samples drawn from $N(0, 1)$ population, are shown in Table 2. These results show that the constrained ML method (CML) is more efficient than the standard ML for small sample sizes ($n = 5, 10, 20$) but these advantages disappear for large sample sizes ($n = 50, 100$). Note that the relative efficiency of the CML with respect to the standard ML method is large in spite of the very wide intervals given in (33), and that the relative improvement with respect to the

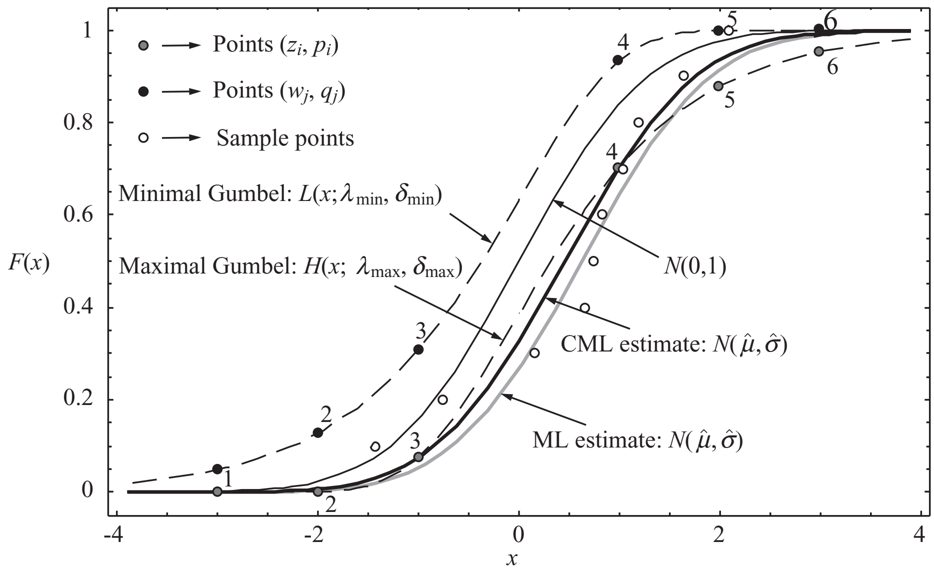

Figure 2: Illustration of the upper and lower limiting points belonging to the maximal and minimal Gumbel distributions (dashed lines), CDF of the $N(0,1)$ together with the ML and CML estimated cumulative distribution functions for the sample of size $n = 10$ in (32).

Table 2: Performance of the maximum likelihood (ML) and constrained maximum likelihood (CML) estimates for a normal parent and different sample sizes n.

n	Method	Bias($\hat{\mu}$)	MSE($\hat{\mu}$)	Bias($\hat{\sigma}$)	MSE($\hat{\sigma}$)
5	CML	–0.00967	0.07267	–0.05554	0.04604
	ML	0.00454	0.19961	–0.16224	0.11839
10	CML	–0.00799	0.05853	–0.04091	0.03314
	ML	–0.00111	0.09992	–0.07910	0.05464
20	CML	–0.00229	0.04049	–0.03010	0.02153
	ML	0.00015	0.04986	–0.04055	0.02640
50	CML	0.00361	0.01942	–0.01393	0.00980
	ML	0.00403	0.01982	–0.01450	0.01001
100	CML	0.00160	0.01002	–0.00678	0.00507
	ML	0.00165	0.01002	–0.00669	0.00507

standard ML method increases substantially with decreasing sample size. This shows that our method has the same asymptotic properties as the ML method.

3.3 Application to the Uniform Distribution

Suppose that we have a sample of size n coming from a uniform population, $U(\alpha, \beta)$, and that we wish to estimate the parameters α and β. In this case, the log likelihood function to be maximized is:

$$\ell(\alpha, \beta | x_1, \ldots, x_n) = -n \log(\beta - \alpha),$$

from which the maximum likelihood estimates of α and β can be found to be $\hat{\alpha} = x_{(1)}$ and $\hat{\beta} = x_{(n)}$, where $x_{(1)}$ and $x_{(n)}$ are the sample minimum and maximum, respectively.

The constrained maximum likelihood in this case becomes

$$\underset{\alpha,\beta;\alpha\leq\beta}{\text{Maximize}} \;\; \ell(\alpha, \beta | x_1, \ldots, x_n) = -n \log(\beta - \alpha) \tag{35}$$

subject to

$$\begin{array}{rcll} F(z_i; \alpha, \beta) & \geq & p_i, & i = 1, 2, \ldots, r, \\ F(w_j; \alpha, \beta) & \leq & q_j, & j = 1, 2, \ldots, s, \\ \alpha & \leq & x_{(1)}, & \\ x_{(n)} & \leq & \beta. & \end{array} \tag{36}$$

For illustration purposes, we have selected the following constraints:

$$\begin{array}{rcccl} 0.05 & \leq & \dfrac{0.1 - \alpha}{\beta - \alpha} & \leq & 0.30, \\ 0.80 & \leq & \dfrac{0.9 - \alpha}{\beta - \alpha} & \leq & 0.95, \\ \alpha & \leq & x_{(1)}, & & \\ & & x_{(n)} & \leq & \beta. \end{array} \tag{37}$$

Simulation results, based on 10,000 samples drawn from a uniform $U(0, 1)$, are shown in Table 3.

Note that the constrained maximum likelihood method (CML) clearly outperforms the standard maximum likelihood method for small sample sizes ($n = 5, 10, 20$), but the efficiency decreases with increasing sample size. Note also that the results for large samples ($n = 100$) are almost the same.

3.4 Application to the Weibull Distribution

To further illustrate the method, consider the case of estimating the parameters of a minimal Weibull distribution, which appears very frequently in practical problems when we observe data that represent minima values (see, e.g., Castillo et al. (2005)). The pdf of the minimal Weibull random variable, $W(\lambda, \delta, \kappa)$, is

$$f(x; \lambda, \delta, \kappa) = \frac{1}{\delta\kappa} \left(\frac{x - \lambda}{\delta} \right)^{1/\kappa - 1} \exp \left[- \left(\frac{x - \lambda}{\delta} \right)^{1/\kappa} \right], \quad x \geq \lambda \tag{38}$$

Table 3: Performance of the maximum likelihood (ML) and constrained maximum likelihood (CML) estimates for a uniform parent and different sample sizes n.

n	Method	Bias($\hat{a}$)	MSE($\hat{a}$)	Bias($\hat{b}$)	MSE($\hat{b}$)
5	CML	0.04798	0.00252	–0.04795	0.00252
	ML	0.16656	0.04778	–0.16773	0.04819
10	CML	0.04202	0.00209	–0.04176	0.00207
	ML	0.09100	0.01511	–0.09097	0.01519
20	CML	0.03287	0.00146	–0.03270	0.00145
	ML	0.04755	0.00433	–0.04726	0.00427
50	CML	0.01822	0.00057	–0.01840	0.00059
	ML	0.01928	0.00072	–0.01958	0.00076
100	CML	0.00965	0.00018	–0.00984	0.00019
	ML	0.00970	0.00019	–0.00988	0.00019

and the corresponding cdf is

$$F(x;\lambda,\delta,\kappa) = \begin{cases} 0, & \text{if } x < \lambda, \\ 1 - \exp\left[-\left(\dfrac{x-\lambda}{\delta}\right)^{1/\kappa}\right], & \text{otherwise.} \end{cases} \tag{39}$$

The $W(\lambda,\delta,\kappa)$ depends on three parameters: a location parameter, λ, a scale parameter, δ, and a shape parameter κ.

Suppose now that we have a sample of size n from a Weibull population, $W(\lambda,\delta,\kappa)$, and that we want to estimate the three parameters. In this case, the log likelihood function is

$$\ell(\lambda,\delta,\kappa|x_1,\ldots,x_n) = -n(\ln\kappa+\ln\delta) - \sum_{i=1}^{n}\left(\frac{x_i-\lambda}{\delta}\right)^{1/\kappa} + (1/\kappa-1)\sum_{i=1}^{n}\ln\left(\frac{x_i-\lambda}{\delta}\right), \tag{40}$$

and then, the constrained maximum likelihood estimates are the solution to the following optimization problem:

$$\underset{\lambda,\delta>0,\kappa>0}{\text{Maximize}}\ \ell(\lambda,\delta,\kappa|x_1,\ldots,x_n) = -n(\ln\kappa+\ln\delta) - \sum_{i=1}^{n}\left(\frac{x_i-\lambda}{\delta}\right)^{1/\kappa} + (1/\kappa-1)\sum_{i=1}^{n}\ln\left(\frac{x_i-\lambda}{\delta}\right)$$

subject to

$$\begin{array}{rl} p_i \le F(z_i;\lambda,\delta,\kappa), & i = 1,2,\ldots,r, \\ F(w_j;\lambda,\delta,\kappa) \le q_j, & j = 1,2,\ldots,s, \\ \lambda \le x_{(1)}, & \end{array} \tag{41}$$

where $x_{(1)}$ is the sample minimum and $F(x;\lambda,\delta,\kappa)$ is the cdf of $W(\lambda,\delta,\kappa)$ in (39).

To obtain the constrained ML estimates, we consider $\boldsymbol{z} = (1.85, 2.3, 3.3, 4.1)$ and $\boldsymbol{w} = (1.65, 2.1, 3, 3.7)$. With these values, the constraints in (41) become:

Table 4: Performance of the maximum likelihood (ML) and constrained maximum likelihood (CML) estimates for a Weibull parent and different sample sizes n.

n	Method	Bias($\hat{\kappa}$)	MSE($\hat{\kappa}$)	Bias($\hat{\delta}$)	MSE($\hat{\delta}$)	Bias($\hat{\lambda}$)	MSE($\hat{\lambda}$)
5	CML	0.01435	0.01457	0.14788	0.88455	–0.17046	0.76792
	ML	–0.03679	0.05229	–0.85557	1.20083	0.92786	1.65385
10	CML	0.02099	0.01452	0.10732	0.78413	–0.12698	0.68582
	ML	–0.00118	0.03947	–0.50784	1.01087	0.56413	1.29943
20	CML	0.01914	0.01251	0.07692	0.59047	–0.09215	0.51693
	ML	0.01494	0.02515	–0.20455	0.73897	0.22791	0.84505
50	CML	0.01782	0.00757	–0.00612	0.24192	–0.00494	0.20557
	ML	0.01577	0.01163	–0.06983	0.31721	0.07370	0.33316
100	CML	0.01271	0.00392	–0.03092	0.09708	0.02335	0.07966
	ML	0.00701	0.00645	–0.06744	0.14434	0.07530	0.17738

$$\begin{array}{rcl} F(1.65;\lambda,\delta,\kappa) \leq & 0.05 & \leq F(1.85;\lambda,\delta,\kappa), \\ F(2.10;\lambda,\delta,\kappa) \leq & 0.20 & \leq F(2.30;\lambda,\delta,\kappa), \\ F(3.00;\lambda,\delta,\kappa) \leq & 0.70 & \leq F(3.30;\lambda,\delta,\kappa), \\ F(3.70;\lambda,\delta,\kappa) \leq & 0.95 & \leq F(4.10;\lambda,\delta,\kappa), \\ & \lambda & \leq x_{(1)}. \end{array} \quad (42)$$

To investigate the performance of the constrained maximum likelihood estimators, we simulated 10,000 samples for each sample size $n = 5, 10, 20, 50$ and 100 from a Weibull $W(1, 2, 1/3)$.

The maximum likelihood estimators of λ, δ and κ are obtained numerically using GAMS. The simulation results are shown in Table 4. These results show that the constrained ML method (CML) is more efficient than the standard ML for all the samples but the relative efficiency is better for the smallest ones ($n = 5, 10, 20$).

4 Constrained Method of Moments

As mentioned above, a modified version of the method of moments can be formulated as the following optimization problem:

$$\underset{\boldsymbol{\theta}\in\boldsymbol{\Theta}}{\text{Minimize}} \sum_{r=1}^{\ell} \left[\frac{g_a^r(\boldsymbol{\theta})}{h_a^r(\boldsymbol{x})} - 1\right]^2, \quad \ell \geq m, \quad (43)$$

subject to

$$p_i \leq F(z_i;\boldsymbol{\theta}), \qquad i = 1, 2, \ldots, r,$$

$$F(w_j;\boldsymbol{\theta}) \leq q_j, \qquad j = 1, 2, \ldots, s,$$
$$\theta_k \leq c_k(\boldsymbol{x}), \qquad k = 1, 2, \ldots, m.$$

Note that the function (43) has been written in adimensional form, preventing numerical difficulties to occur and some moments from having much more weight than others, when $\ell \geq m$.

In the following sections we apply the method to data coming from normal, uniform, and Weibull parents.

4.1 Application to the Normal Distribution

The method of moments estimators in the normal case are the same as the maximum likelihood estimators given in (27). The constrained method of moments problem in this case can be formulated as:

$$\underset{\mu,\sigma>0}{\text{Minimize}} \ \left[\frac{\mu}{\bar{x}} - 1\right]^2 + \left[\frac{\sigma^2 + \mu^2}{h_0^2(\boldsymbol{x})} - 1\right]^2, \tag{44}$$

subject to the constraints in (30) and (31), where $h_0^2(\boldsymbol{x}) = n^{-1}\sum_{i=1}^{n} x_i^2$ and $\ell = m = 2$.

Using the same constraints in (33) with the simulated data, we obtain the simulation results in Table 5, which are based on 10,000 samples of sizes $n = 5, 10, 20, 50, 100$ from a normal $N(0, 1)$. The results show that the constrained MOM method (CMOM) is more efficient than the standard MOM for small sizes ($n = 5, 10, 20$) but these advantages disappear for large sample sizes ($n = 50, 100$). Note that the relative efficiency of the CMOM with respect to the standard MOM method is large in spite of the very wide intervals given in (33), and that the relative improvement with respect to the standard MOM method increases substantially with decreasing sample size.

Again this shows that our method and the standard method of moments are asymptotically equivalent if the constraints are really satisfied by the parent population.

4.2 Application to the Uniform Distribution

Suppose we have an iid sample of size n from the uniform distribution $U(\alpha, \beta)$. The first two moments with respect to the origin are

$$E(X) = \frac{\alpha + \beta}{2} \quad \text{and} \quad E(X^2) = \frac{\alpha^2 + \alpha\beta + \beta^2}{3}.$$

Table 5: Performance of the method of moments (MOM) and the constrained method of moments (CMOM) estimates for a normal parent and different sample sizes n.

n	Method	Bias($\hat{\mu}$)	MSE($\hat{\mu}$)	Bias($\hat{\sigma}$)	MSE($\hat{\sigma}$)
5	CMOM	–0.00607	0.08013	–0.05860	0.03697
	MOM	0.00454	0.19961	–0.16224	0.11839
10	CMOM	–0.00510	0.06453	–0.03836	0.02802
	MOM	–0.00111	0.09992	–0.07910	0.05464
20	CMOM	–0.00100	0.04321	–0.02682	0.01982
	MOM	0.00015	0.04986	–0.04055	0.02640
50	CMOM	0.00392	0.01974	–0.01301	0.00967
	MOM	0.00403	0.01982	–0.01450	0.01001
100	CMOM	0.00162	0.01015	–0.00585	0.00504
	MOM	0.00165	0.01002	–0.00669	0.00507

Then, the method of moments estimators of α and β are:

$$\hat{\alpha} = \bar{x} - \sqrt{3}\sqrt{h_0^2 - \bar{x}^2} \quad \text{and} \quad \hat{\beta} = \bar{x} + \sqrt{3}\sqrt{h_0^2 - \bar{x}^2}.$$

Setting $\ell = m = 2$, the optimization problem of the constrained method of moments becomes:

$$\underset{\alpha,\beta}{\text{Minimize}}\ Z = \left(\frac{\alpha+\beta}{2\bar{x}} - 1\right)^2 + \left(\frac{\alpha^2+\alpha\beta+\beta^2}{3h_0^2(\boldsymbol{x})} - 1\right)^2, \tag{45}$$

subject to the constraints in (36), where $h_0^2(\boldsymbol{x}) = n^{-1}\sum_{i=1}^n x_i^2$ are the first and second sample moments with respect to the origin, respectively, and $x_{(1)}$ and $x_{(n)}$ are the sample minimum and maximum, respectively.

To see the improvement in parameter estimation when using the constrained method of moments, we simulated 10,000 samples for each sample size $n = 5, 10, 20, 50$, and 100 from $U(0,1)$. Using the same constraints in (37) with the simulated data, the results are given in Table 6.

Note that the constrained method of moments outperforms both the maximum likelihood method (Table 3) and the standard method of moments (Table 6) for small sample sizes ($n = 5, 10, 20$), but is clearly inferior than the maximum likelihood method estimates for large samples ($n = 50, 100$). It is also important to note that the constrained MOM corrects the inconsistent estimates given by the standard MOM when $\hat{\theta} \notin \Theta$), that is, when $\hat{\alpha}_{MOM} > x_{(1)}$ or $\hat{\beta}_{MOM} < x_{(n)}$.

Table 6: A comparison of the performance of the standard method of moments estimates (MOM), and the constrained method of moments (CMOM) estimates for a uniform parent and different sample sizes n. The percentages of inconsistent estimates for the standard method of moments are shown.

						Inconsistent		
n	Method	Bias($\hat{\alpha}$)	MSE($\hat{\alpha}$)	Bias($\hat{\beta}$)	MSE($\hat{\beta}$)	α(%)	β(%)	α or β(%)
5	CMOM	–0.07316	0.02185	0.02413	0.00545			
	MOM	0.06941	0.03674	–0.07129	0.03676	12.53	12.19	24.72
10	CMOM	–0.05509	0.01512	0.02110	0.00458			
	MOM	0.03195	0.01540	–0.03261	0.01570	22.94	23.61	44.86
20	CMOM	–0.03795	0.00885	0.01787	0.00344			
	MOM	0.01586	0.00713	–0.01494	0.00712	31.11	30.52	56.10
50	CMOM	–0.02654	0.00394	0.01586	0.00193			
	MOM	0.00704	0.00276	–0.00538	0.00271	38.68	36.92	65.61
100	CMOM	–0.02228	0.00217	0.01376	0.00107			
	MOM	0.00359	0.00136	–0.00283	0.00134	42.11	40.60	69.89

4.3 Application to the Weibull Distribution

To further illustrate the method, consider the case of estimating the parameters λ, δ, and κ of the minimal Weibull distribution whose pdf is given in (38). Since we have three parameters to estimate, we need at least three moments to obtain the MOM estimates of the three parameters. We consider the first three moments with respect to the origin. These moments are:

$$\begin{aligned} g_0^1(\boldsymbol{\theta}) &= \lambda + \delta\Gamma(1+\kappa) = \mu, \\ g_0^2(\boldsymbol{\theta}) &= \lambda^2 + 2\lambda\delta\Gamma(1+\kappa) + \delta^2\Gamma(1+2\kappa), \\ g_0^3(\boldsymbol{\theta}) &= \lambda^3 + 3\lambda\delta\left[\lambda\Gamma(1+\kappa) + \delta\Gamma(1+2\kappa)\right] + \delta^3\Gamma(1+3\kappa). \end{aligned}$$

To obtain the constrained MOM estimates, we consider $\boldsymbol{z} = \boldsymbol{w} = (1.5, 2, 3, 4)$. With $\ell = m = 3$, the optimization problem then becomes:

$$\underset{\lambda,\delta>0,\kappa>0}{\text{Minimize}}\ Z = \left(\frac{\mu}{\bar{x}} - 1\right)^2 + \left(\frac{g_0^2(\boldsymbol{\theta})}{h_0^2(\boldsymbol{x})} - 1\right)^2 + \left(\frac{g_0^3(\boldsymbol{\theta})}{h_0^3(\boldsymbol{x})} - 1\right)^2, \tag{46}$$

subject to the constraints in (42), where $h_0^2(\boldsymbol{x}) = n^{-1}\sum_{i=1}^n x_i^2$ and $h_0^3(\boldsymbol{x}) = n^{-1}\sum_{i=1}^n x_i^3$.

To see the improvement in parameter estimation when using the constrained method of moments, we simulated 10,000 samples for each sample size $n = 5, 10, 20, 50$ and 100 from a Weibull population with parameters $\lambda = 1$, $\delta = 2$, and $\kappa = 1/3$.

Table 7: A comparison of the performance of the standard method of moments estimates (MOM), and the constrained method of moments estimates (CMOM) for a Weibull parent and different sample sizes n.

n	Method	Bias($\hat{\kappa}$)	MSE($\hat{\kappa}$)	Bias($\hat{\delta}$)	MSE($\hat{\delta}$)	Bias($\hat{\lambda}$)	MSE($\hat{\lambda}$)	Inconsistent λ(%)
5	CMOM	–0.01264	0.00234	0.09734	0.07845	–0.09630	0.05609	
	MOM	–0.03704	0.03363	–0.04633	1.89658	0.04221	1.92322	2.70
10	CMOM	–0.00669	0.00316	0.07720	0.10040	–0.07589	0.07538	
	MOM	–0.01497	0.03064	0.15113	1.69464	–0.15022	1.66388	2.31
20	CMOM	–0.00094	0.00379	0.05216	0.11672	–0.05194	0.09038	
	MOM	–0.00696	0.01991	0.20467	1.08618	–0.20802	1.01208	3.69
50	CMOM	0.00304	0.00377	0.02967	0.11123	–0.03091	0.08983	
	MOM	–0.00403	0.00890	0.10572	0.38653	–0.10812	0.34006	5.18
100	CMOM	0.00252	0.00300	0.01966	0.08445	-0.02156	0.07000	
	MOM	-0.00236	0.00451	0.04474	0.17421	-0.04643	0.14779	5.78

The results of the simulation are given in Table 7. Note that the constrained method of moments clearly outperforms standard method of moments for small sample sizes ($n = 5, 10, 20$), but showing a similar behavior for large samples ($n = 50, 100$). It is also important to notice that the constrained MOM corrects the inconsistent estimates given by the standard MOM (i.e., $\hat{\theta} \notin \Theta$), where $\hat{\lambda}_{MOM} > x_{(1)}$.

To investigate the behavior of the method further, 200 samples are generated for each of four sample sizes (5, 10, 20, and 100). For each of these samples, the density is estimated and plotted. Figure 3, shows the plots of the estimated densities using the standard method (lighter curves) and the constrained method (darker curves). It can be seen from these graphs that the variability of the constrained estimates is much less than that of the standard estimates specially for small samples. These variabilities are similar for large sample sizes.

5 The Case of Infinitely Many Continuous Constraints

In this case we assume that we have bounds for the cdf or all quantiles instead of a finite number of quantiles. Then, incorporating the available information by imposing these constraints the optimization problem becomes:

$$\underset{\boldsymbol{\theta} \in \boldsymbol{\Theta}}{\text{Minimize}} \; f(\boldsymbol{x}, \boldsymbol{\theta}), \tag{47}$$

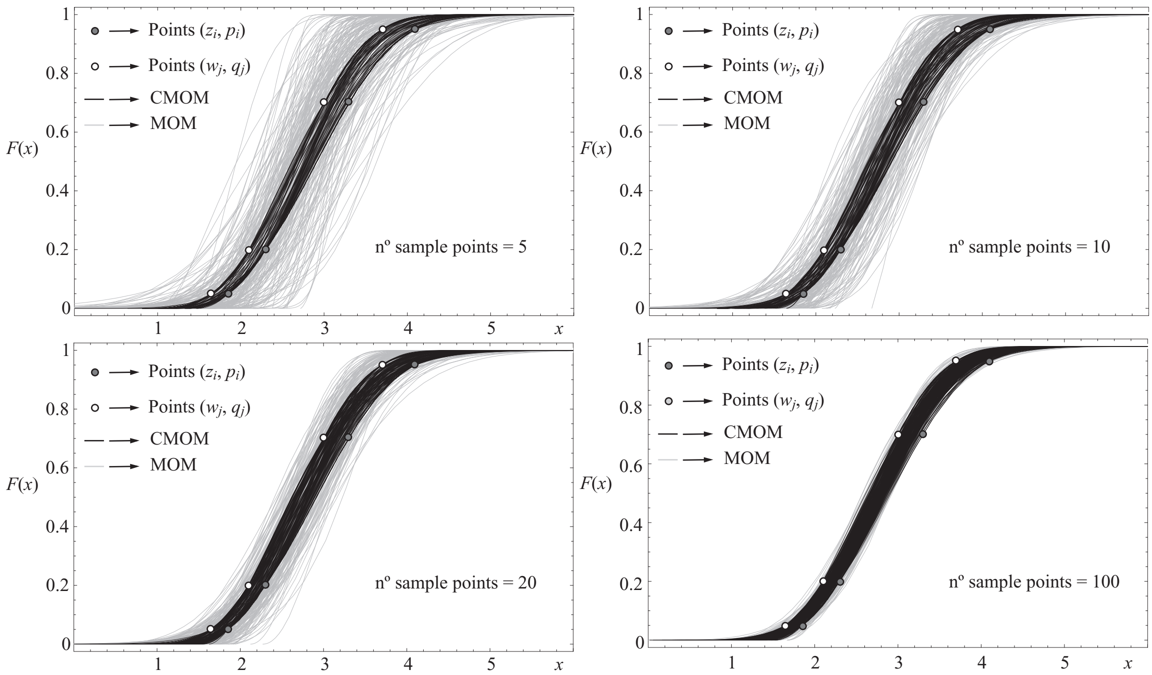

Figure 3: Illustration of the upper and lower limiting points in (42), together with the MOM and CMOM estimated cumulative distribution functions for 200 simulations and different sample sizes.

subject to

$$L(u) \leq F(u; \boldsymbol{\theta}) \leq H(u), \qquad \forall u \in \mathcal{S}, \tag{48}$$

$$\theta_k \leq c_k(\boldsymbol{x}), \quad k = 1, 2, \ldots, m, \tag{49}$$

where $L(u)$ and $H(u)$ are the lower and upper quantile bound functions. This way the estimation problem becomes an infinitely constrained optimization problem.

Condition (48) can be replaced by

$$\ell(p) \leq x(p; \boldsymbol{\theta}) \leq h(p), \quad \forall p \in [0, 1], \tag{50}$$

where $x(p; \boldsymbol{\theta})$ is the quantile function and $\ell(p)$ and $h(p)$ are the corresponding bound functions. They are the inverse functions of $p = L(u)$ and $p = H(u)$, respectively.

Figure 4(a) illustrates the constrained estimation problem showing the lower and upper cumulative distribution functions $L(x)$ and $H(x)$, respectively, together with one admissible cumulative distribution function $F(x)$.

The constraints (48) and (50) can be considered on all $\mathcal{S}$ or $[0, 1]$ or only in some subsets of them. Figure 4(b) illustrates one of these cases where the constraints have been

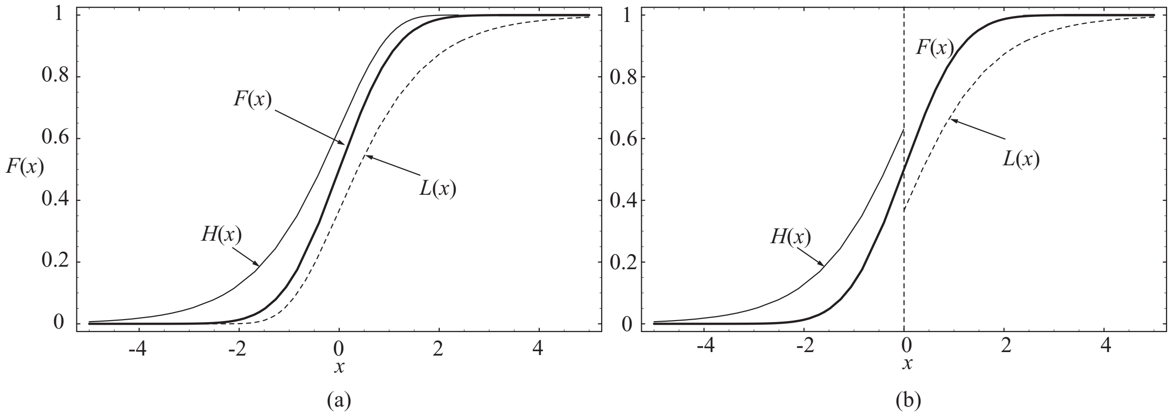

Figure 4: An illustration of the constrained estimation problem showing the lower and upper cumulative distribution functions $L(x)$, $H(x)$ and the admissible $F(x)$ cumulative distribution function $F(x)$: (a) constraints on all $\mathcal{S}$, and (b) constraints just on the tails.

considered only in the tails. This is a common case when one is dealing with extreme value models.

The problem (47)–(49) is difficult to solve because it has infinitely many constraints. However, it is equivalent to the following problem

$$\underset{\boldsymbol{\theta} \in \boldsymbol{\Theta}}{\text{Minimize}} \ f(\boldsymbol{x}, \boldsymbol{\theta}), \tag{51}$$

subject to

$$\underset{u \in \mathcal{S}}{\text{Maximum}} \ L(u) - F(u; \boldsymbol{\theta}) \ \leq \ 0, \tag{52}$$

$$\underset{u \in \mathcal{S}}{\text{Minimum}} \ H(u) - F(u; \boldsymbol{\theta}) \ \geq \ 0, \tag{53}$$

$$\theta_k \ \leq \ c_k(\boldsymbol{x}); \ k = 1, 2, \ldots, m, \tag{54}$$

which has only three constraints. However, the price we pay for this is that we obtain two complicated constraints (52) and (53), so called because standard optimization packages do not admit such a type of constraints, and they are difficult to implement because they involve other optimization problems.

One possibility for solving this problem consists of using the approximating hyperplane decomposition method (see, e.g., Castillo et al. (2005) or Conejo et al. (2005)) To this end, we use an iterative process in which we solve two types of problems: the master problem and the subproblems. We proceed as follows:

- **Step 0: Initialization.** Initialize the iteration counter $\nu = 1$ and set the value of the feasibility tolerance ϵ.

- **Step 1: Master problem solution for iteration ν.** The master problem is solved:

$$\underset{\boldsymbol{\theta} \in \boldsymbol{\Theta}}{\text{Minimize}} \; f(\boldsymbol{x}, \boldsymbol{\theta}), \tag{55}$$

subject to

$$\alpha^{(\ell)} + \sum_{j=1}^{m} \lambda_{1j}^{(\ell)} (\theta_j - \theta_j^{(\ell)}) \;\leq\; 0, \quad \ell = 1, 2, \ldots, \nu - 1, \tag{56}$$

$$\beta^{(\ell)} + \sum_{j=1}^{m} \lambda_{2j}^{(\ell)} (\theta_j - \theta_j^{(\ell)}) \;\geq\; 0, \quad \ell = 1, 2, \ldots, \nu - 1, \tag{57}$$

$$\theta_k \;\leq\; c_k(\boldsymbol{x}); \; k = 1, 2, \ldots, m, \tag{58}$$

obtaining the optimal value $\boldsymbol{\theta}^{(\nu)}$.

Note that the two constraints of the original problem in (52) and (53) have been replaced by linear approximations in (56) and (57).

- **Step 2: Subproblem solution.** Solve the two subproblems. The first is

$$\alpha^{(\nu)} = \underset{u \in \mathcal{S}, \, \boldsymbol{\theta} \in \boldsymbol{\Theta}}{\text{Maximum}} \; L(u) - F(u; \boldsymbol{\theta}) \tag{59}$$

subject to

$$\boldsymbol{\theta} = \boldsymbol{\theta}^{(\nu)}, \tag{60}$$

and the second is

$$\beta^{(\nu)} = \underset{u \in \mathcal{S}, \, \boldsymbol{\theta} \in \boldsymbol{\Theta}}{\text{Minimum}} \; H(u) - F(u; \boldsymbol{\theta}) \tag{61}$$

subject to

$$\boldsymbol{\theta} = \boldsymbol{\theta}^{(\nu)}, \tag{62}$$

¿From these problems we obtain the mismatches $\alpha^{(\nu)}$ and $\beta^{(\nu)}$ associated with constraints (52) and (53), respectively.

- **Step 3: Convergence checking.** If constraints (52) and (53) are feasible, that is, $\alpha^{(\nu)} \leq \epsilon$ and $\beta^{(\nu)} \geq -\epsilon$ then stop and return the estimates $\boldsymbol{\theta}^{(\nu)}$. Otherwise, let $\nu = \nu + 1$, go to Step 1 and repeat the process until convergence.

Note that at iteration $\nu = 1$ there is no hyperplane approximation (56) and (57) of constraints (52) and (53), respectively, so that the first master problem coincides with the ML method.

It should be noted that the problem in (55)–(58) is a relaxation of the problem in (51)–(54) in the sense that the functions (52) and (53) are approximated using cutting hyperplanes and they become more precisely approximated as the iterative procedure progresses, which implies that the problem in (55)–(58) reproduces more exactly the problem in (51)–(54) (see Kelley (1960)). Additionally, observe that cutting hyperplanes are constructed using the dual variable vector associated with constraints (60) and (62) in problems (59)–(60) and (61)–(62) (the subproblems), respectively.

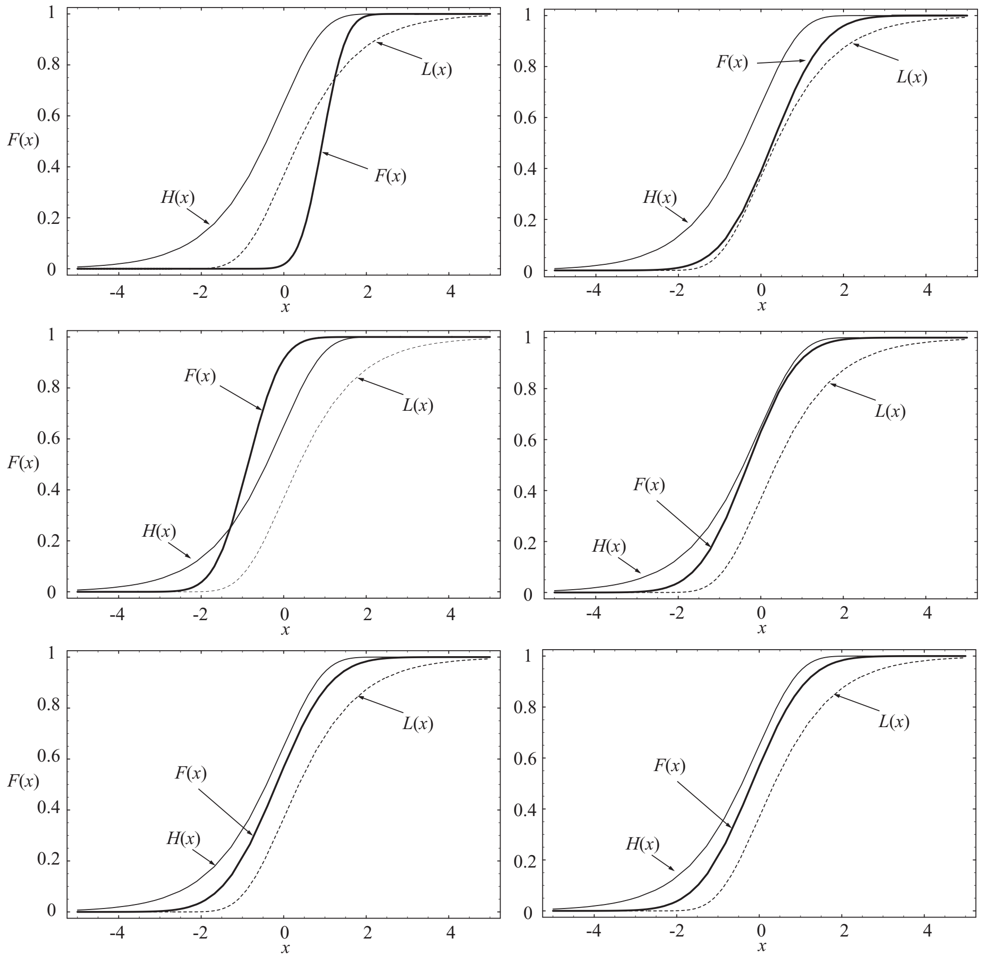

Figure 5: Illustration of some of the estimated densities and the lower and upper bounds. Left figures: unconstrained estimation. Right figures: Corresponding constrained estimation.

Figure 5 shows the resulting estimated cdfs for 4 particular samples of the standard (left side) and the constrained (right sides) methods. Comparing the corresponding figures (those located in the same places) one can see how the constrained method corrects the estimates. Note that the last ones (those in the lower part) are not corrected at all because the initial optimal solution satisfies the bounds.

5.1 Infinitely Constrained Maximum Likelihood Methods

In this section, we apply the proposed method to the maximum likelihood method subject to the bounds for the cdf and/or the quantiles in (48). Then, the resultant constrained maximum likelihood method can be expressed as the optimization problem:

$$\underset{\boldsymbol{\theta} \in \boldsymbol{\Theta}}{\text{Maximize}} \sum_{i=1}^{n} \log f(x_i; \boldsymbol{\theta}) \tag{63}$$

subject to (48).

To illustrate the gain in the precision of the estimates, we consider a sample of size n from a $N(\mu, \sigma)$ population. We wish to estimate the mean μ and the standard variation σ, but in this case we know that the cdf is bounded by

$$\begin{aligned} H(u) &= \exp\left(-\exp\left(\frac{\lambda_{\max} - u}{\delta_{\max}}\right)\right), \qquad -\infty < u < \infty, \\ L(u) &= 1 - \exp\left(-\exp\left(\frac{u - \lambda_{\min}}{\delta_{\min}}\right)\right), \quad -\infty < u < \infty, \end{aligned} \tag{64}$$

where $\lambda_{\max} = -0.05, \delta_{\max} = 1, \lambda_{\min} = 0$ and $\delta_{\min} = 1$ are the parameters associated with the maximal and minimal Gumbel distributions, respectively.

We have simulated 10,000 samples of size $n = 5, 10, 20, 50, 100$, from a normal $N(0, 1)$ population, and used both methods of estimations and obtained the results shown in Table 8, where it can be seen that the ICML method is more efficient than the standard ML for small sizes ($n = 5, 10, 20$) but these advantages disappear for large sample sizes ($n = 50, 100$). Note that the results are very similar to those in Table 3 but the estimates are more accurate because in this case the problem is more restricted. Note in Figure 6 that the ICML distribution function estimate is closer to the $N(0, 1)$ distribution function than the CML distribution function estimate.

Table 8: Performance of the maximum likelihood (ML) and the continuously constrained maximum likelihood (ICML) estimates for a normal parent and different sample sizes n (10,000 simulations).

n	Method	Bias($\hat{\mu}$)	MSE($\hat{\mu}$)	Bias($\hat{\sigma}$)	MSE($\hat{\sigma}$)
5	CCML	–0.00800	0.09375	–0.10922	0.10683
	ML	0.00454	0.19961	–0.16224	0.11839
10	CCML	–0.00635	0.06561	–0.06141	0.05125
	ML	–0.00111	0.09992	–0.07910	0.05464
20	CCML	–0.00112	0.04262	–0.03712	0.02593
	ML	0.00015	0.04986	–0.04055	0.02640
50	CCML	0.00387	0.01951	–0.01436	0.00999
	ML	0.00403	0.01982	–0.01450	0.01001
100	CCML	0.00165	0.01001	–0.00669	0.00507
	ML	0.00165	0.01002	–0.00669	0.00507

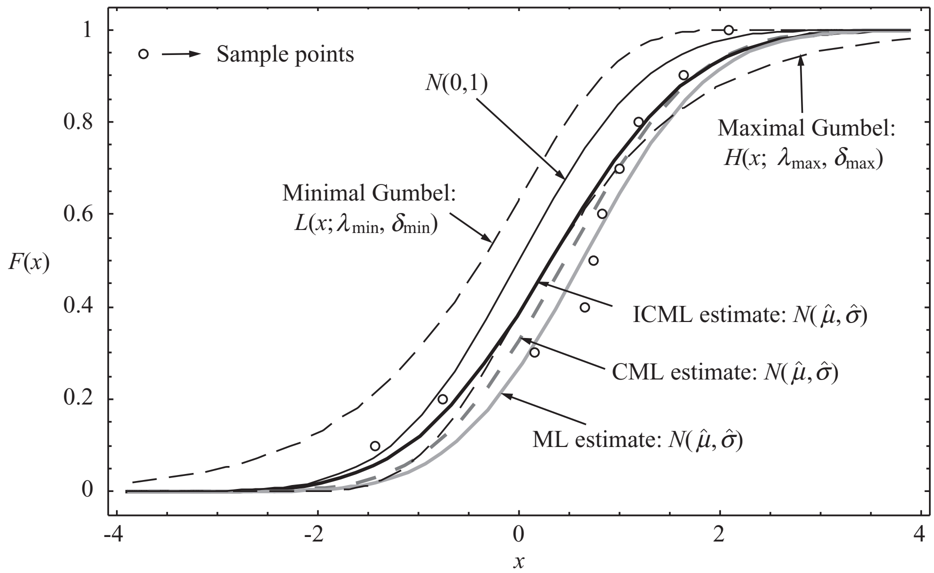

Figure 6: Illustration of the upper and lower limiting maximal and minimal Gumbel distributions (dashed lines), together with the ML, CML and ICML estimated cumulative distribution functions for a sample of size 10.

5.2 Infinitely Constrained Method of Moments Methods

In this section, we apply the proposed method to the method of moments subject to the bounds for the cdf and/or the quantiles in (48). Then, the resultant infinitely constrained method of moments can be expressed as the optimization problem:

$$\underset{\boldsymbol{\theta} \in \boldsymbol{\Theta}}{\text{Minimize}} \sum_{r=1}^{\ell} \left[\frac{g_a^r(\boldsymbol{\theta})}{h_a^r(\boldsymbol{x})} - 1 \right]^2, \quad \ell \geq m, \tag{65}$$

Table 9: Performance of the method of moments (MOM) and the continuously constrained method of moments (CCMOM) estimates for a normal parent and different sample sizes n (10,000 simulations).

n	Method	Bias($\hat{\mu}$)	MSE($\hat{\mu}$)	Bias($\hat{\sigma}$)	MSE($\hat{\sigma}$)
5	ICMOM	–0.00793	0.09791	–0.09148	0.11295
	MOM	0.00454	0.19961	–0.16224	0.11839
10	ICMOM	–0.00572	0.06849	–0.05264	0.05318
	MOM	–0.00111	0.09992	–0.07910	0.05464
20	ICMOM	–0.00089	0.04353	–0.03421	0.02635
	MOM	0.00015	0.04986	-0.04055	0.02640
50	ICMOM	0.00392	0.01958	–0.01415	0.01000
	MOM	0.00403	0.01982	–0.01450	0.01001
100	ICMOM	0.00165	0.01001	–0.00668	0.00507
	MOM	0.00165	0.01002	–0.00669	0.00507

subject to (48).

To illustrate the gain in the precision of the estimates, we consider a sample of size n from a normal $N(\mu, \sigma)$ population and we wish to estimate the mean μ and the standard deviation σ, but in this case we know that the cdf is bounded by the constraints in (64).

We have simulated 10,000 samples of size $n = 5, 10, 20, 50, 100$, from the $N(0, 1)$, and used both methods of estimations and obtained the results shown in Table 9, where it can be seen that the ICML method is more efficient than the standard ML for small sizes ($n = 5, 10, 20$) but these advantages disappear for large sample sizes ($n = 50, 100$).

6 Conclusions

In this paper we have introduced a new estimation method that allows incorporating extra information in the parameter estimation problem in terms of bounds for the cdf or the quantile functions and also in terms of constraints imposed by the support of the random variable of interest. The application of the method to two methods of estimation (the maximum likelihood method and the method of moments) and using three families of distributions (the normal, uniform, and Weibull) have been discussed. But the proposed method is general in the sense that it can be applied to improve any method of estimation as long as it can be expressed as an optimization problem.

The examples and simulations show that important gains can be achieved with respect to the case where this information is lacking for small sample sizes even when the bounds are not very precise. When the sample size is large, there is practically no gain.

References

Aitchison, J. and Silvey, S. D. (1958), "Maximum-Likelihood Estimation of Parameters Subject to Restraints," *The Annals of Mathematical Statistics*, 29(3):813–828.

Arthanari, T. S. and Dodge, Y. (1993), *Mathematical Programming in Statistics*, John Wiley & Sons, New York.

Bazaraa, M. S., Sherali, H. D., and Shetty, C. M. (1993), *Nonlinear Programming. Theory and Algorithms, Second Edition*, John Wiley & Sons, New York.

Bickel, P. J. and Doksum, K. A. (2000), *Mathematical Statistics: Basic Ideas and Selected Topics: Vol. I, Second Edition*, Prentice Hall, New York.

Casella, G. and Berger, R. L. (2001), *Statistical Inference, Second Edition*, Duxbury Press, Belmont, CA.

Castillo, E., Conejo, A., Castillo, C., Mínguez, R., and Ortigosa, D. (2005), "A Perturbation Approach to Sensitivity Analysis in Nonlinear Programming," *Journal of Optimization Theory and Applications*, 127(3).

Castillo, E., Conejo, A., Pedregal, P., García, R., and Alguacil, N. (2001), *Building and Solving Mathematical Programming Models in Engineering and Science*, John Wiley & Sons, New York.

Castillo, E., Hadi, A. S., Balakrishnan, N., and Sarabia, J. M. (2005), *Extreme Value and Related Models in Engineering and Science Applications*, John Wiley & Sons, New York.

Coles, S. (2001), *An Introduction to Statistical Modeling of Extreme Values*, Springer Verlag, New York.

Conejo, A., Castillo, E., Mínguez, R. and García-Bertrand, R. (2005), *Decomposition Techniques in Mathematical Programming. Engineering and Science Applications*, Springer Verlag, New York.

Cramér, H. (1946), *Mathematical Methods of Statistics*, Princeton University Press, Princeton, New Jersey.

Drud, A. (1996), CONOPT: A System for Large Scale Nonlinear Optimization, Reference manual for CONOPT subroutine library, ARKI Consulting and Development A/S, Bagsvaerd, Denmark.

Galambos, J. (1987), *The Asymptotic Theory of Extreme Order Statistics*, Robert E. Krieger Publishing Company, Malabar, Florida.

Hogg, R. V. and Craig, A. T. (2000), *Introduction to Mathematical Statistics*, Volume I and II, Sixth Edition. Prentice Hall, New York.

Kelley, J. E. (1960), "The Cutting Plane Method for Solving Convex Programs," *J. ,Soc. Ind. and Appl. Math.*, 8(4):703–712.

Mann, N. R., Schafer, R. E., and Singpurwalla, N. D. (1974), *Methods for Statistical Analysis of Reliability and Life Data*, John Wiley & Sons, New York.

Miller, I. and Miller, M. (2003), *John E. Freund's Mathematical Statistics with Applications, Seventh Edition*, Prentice Hall, New York.

Mukhopadhyay, N. (2000), *Probability and Statistical Inference*, Marcel Dekker, New York.

Rao, C. R. (2002), *Linear Statistical Inference and its Applications, Second Edition*, John Wiley & Sons, New York.

Silvapulle, M. J. and Sen, P. K. (2004), *Constrained Statistical Inference: inequality, order, and shape restrictions*, John Wiley & Sons, New York.

Vanderplaats, G. N. (1984), *Numerical Optimization Techniques for Engineering Design*, McGraw-Hill, New York.

Wackerly, D., Mendenhall, W., and Scheaffer, R. L. (2001), *Mathematical Statistics with Applications, Sixth Edition*, PWS-Kent Publishing Company, Boston, MA.