On Reliability Growth Models Using Kolmogorov-Smirnov Bounds

LEV V. UTKIN[1*], FRANK P.A. COOLEN[2]

[1]*Department of Computer Science, St.Petersburg State Forest Technical Academy*
Institutsky per. 5, 194021 St. Petersburg, Russia
[2]*Department of Mathematical Sciences, Durham University*
Durham, DH1 3LE, England

(Received on …)

Abstract: An approach for constructing nonparametric imprecise growth models (regression models) is proposed. The approach is based on applying sets of probability distributions of the "noise" produced by means of Kolmogorov-Smirnov bounds. The corresponding growth models are constructed by minimizing the risk functional in the framework of predictive learning and by choosing "optimal" probability distribuions of the "noise" defining the minimax and minimin strategies. Numerical examples illustrate the proposed approach.

Keywords: *imprecise probabilities, reliability, lower and upper probability distributions, learning problem, risk functional, software reliability growth models.*

1. Introduction

Many models in statistics are used to link explanatory variables to a main variable of interest. Often, linear models are used, but a wide range of models has been developed each of which is suitable under specific circumstances. For example, generalized linear models may be useful in case of discrete variables, proportional hazards models may be useful in specific reliability or survival analysis scenarios, and accelerated lifetime models are important for practical testing of equipment with lifetimes which are too long for testing under normal conditions. While the use of such models in many application areas of statistics, including reliability, has been widely accepted, the implementation of statistical methods for fitting such models and for related inference remains a topic of wide interest and research activity [1]. Traditionally, strong modeling assumptions have enabled the use of a powerful inferential framework based on the normal distribution assumed for residual terms in the model, but often residuals are clearly not normally distributed which undermines the use of traditional methods for estimating the model parameters, such as least squares or maximum likelihood methods.

In this paper, we explore an alternative approach to fitting such models, reducing the assumptions for the residuals. We use the nonparametric Kolmogorov-Smirnov (KS) bounds for the cumulative distribution function (CDF) of the residuals, which provides approximate confidence intervals for the CDF corresponding to the underlying population from which the residuals are assumed to be sampled given the fitted model.

The main idea for constructing new reliability models is the following. Parameters of the models are computed by minimizing a risk functional (RF) defined by a certain loss function and by a probability distribution of the noise [2,3]. By having a small amount of training data or imprecise data, we construct a set of distributions or the P-box [4]. Then by using the set of probability distributions instead of a precise distribution, we can

* Corresponding author's email: lev.utkin@gmail.com

choose a single distribution from the set which minimizes or maximizes the RF. The probability distribution maximizing the RF corresponds to the *minimax strategy*. The distribution minimizing the RF corresponds to the *minimin strategy*. These cases can also be called the *pessimistic and optimistic statistical decisions*. The main feature here is that the chosen distributions as well as the bounds of the set of distributions depend on the unknown model parameters which have to be computed. After substituting the optimal distribution into an expression for the RF, we can compute the optimal parameters of the regression model by minimizing the risk measure over the set of values of parameters.

In summary, we can write the whole algorithm for developing new models:

1. A set of probability distributions is constructed on the basis of given training data by using KS confidence limits and nonparametric inference.

2. The pessimistic and optimistic probability distributions are determined. The largest and smallest risk measures as functions of the regression parameters are determined corresponding to the minimax and minimin strategies, respectively.

3. The model parameters are computed by minimizing the obtained risk measures.

As a result, we get two reliability growth models: pessimistic and optimistic.

2. Reliability growth model in the framework of nonparametric inference and predictive learning

The basic idea of nonparametric inference is to use data to infer an unknown quantity while making as few assumptions as possible. Given an independent and identically distributed (i.i.d.) sample $Z_1,...,Z_n$ having the CDF $F(z) = P(Z \le z)$ on the real line, we estimate F in the framework of nonparametric methods with the empirical CDF as the CDF that puts mass $1/n$ at each data point Z_i, i.e.,

$$F_n(z) = \frac{1}{n}\sum_{i=1}^{n} I(Z_i \le z), \quad I(Z_i \le z) \text{ is the indicator function.}$$

One of the ways for taking into account the scarcity of statistical data is to construct a set of probability distributions, namely, its bounds by using nonparametric estimation is KS confidence limits which can be regarded as distribution-free bounds about an empirical CDF. According to the KS statistic, a band of width $\pm d_{n,1-\gamma}$ around the empirical CDF will entirely contain $F(z)$ with confidence level $1-\gamma$. The expressions and tables for $d_{n,1-\gamma}$ can be found in [5]. Denoting $\nu = d_{n,1-\gamma}$ for short, we write the following bounds for the unknown CDF $F(z)$:

$$\max(F_n(z) - \nu), 0) \le F(z) \le \min(F_n(z) + \nu, 1) \qquad (1)$$

It can be seen from (1) that the left tail of the upper CDF is ν for $z \to -\infty$. However, the smallest possible value might be limited by some value depending on the analyzed application. The right tail of the lower CDF is $1-\nu$ for $z \to \infty$. In this case, we can also restrict the large values of the quantity under consideration.

3. Reliability growth model in the framework of predictive learning

According to Vapnik [3], the learning problem can be described as follows. Suppose there is a set of n independent observations $(\mathbf{x}_i, y_i)$, $i = 1,...,n$. We select the best available growth function $f(\mathbf{x}, \alpha_{\text{opt}})$ from the set of growth functions $f(\mathbf{x}, \alpha)$ parameterized by a set of parameters $\alpha \in \Lambda$ such that $y = f(\mathbf{x}, \alpha_{\text{opt}}) + z$, where $z \in \mathbb{R}$ is the independent

zero mean random error (noise) having the probability density function (PDF) $p(z)$ such that $\mathbb{E}z^2 < \infty$. The function $f(\mathbf{x}, \alpha_{\text{opt}})$ best approximates the system response. Here $\mathbf{x} \in \mathbb{R}^m$ is a multivariate input and $y \in \mathbb{R}$ is a scalar output.

The quality of an approximation is measured by the loss function $L(y - f(\mathbf{x}, \alpha)) = L(z)$, for instance, $L(z) = z^2$ or $L(z) = |z|$. The main goal of learning is to minimize the following RF:

$$R(\alpha) = \int L(z)\, p(z)\, \mathbf{d}z.$$

The minimization of the RF is carried out over the class of functions $f(\mathbf{x}, \alpha)$, $\alpha \in \Lambda$, i.e., the function $f(\mathbf{x}, \alpha_{\text{opt}})$ provides the minimum of $R(\alpha)$ such that $R(\alpha_{\text{opt}}) = \min_{\alpha \in \Lambda} R(\alpha)$. If the PDF $p(z)$ is unknown, then $R(\alpha)$ can be replaced by the so-called empirical RF

$$R_{\text{emp}}(\alpha) = \frac{1}{n} \sum_{i=1}^{n} L(y - f(\mathbf{x}, \alpha)).$$

It should be noted that some reliability growth models [6] have been studied in the framework of learning theory. However, these models just use the empirical RF and adopt it to the well-known software reliability models.

If a family of probability distributions is known, then a common technique for computing the best function $f(\mathbf{x}, \alpha_{\text{opt}})$ is the maximum likelihood estimation method [3].

The linear model assumes that the function f is of the form $f(\mathbf{x}, \alpha) = \alpha \mathbf{x}^{\mathsf{T}} + \alpha_{m+1}$.

4. The learning problem with a set of distributions

Now suppose that we do not know the precise probability distribution of the noise, but we know that this distribution belongs to a set $\mathcal{F}$ bounded by some lower $\underline{F}$ and upper $\overline{F}$ CDFs, where these bounds can be obtained as functions of the observations $(\mathbf{x}_i, y_i)$, $i = 1, ..., n$, for example using the KS bounds as presented in this paper, but also possibly following from other statistical inference methods. Let us introduce a set $\mathcal{R}$ of the corresponding PDFs, i.e., for every CDF from $\mathcal{F}$, there exists a PDF $p \in \mathcal{R}$. Then two possible strategies to derive an estimator are the minimax (pessimistic) strategy and the minimin (optimistic) strategy.

4.1 The minimax strategy

The minimax RF with respect to the minimax strategy is of the form:

$$\overline{R}(\alpha_{\text{opt}}) = \min_{\alpha \in \Lambda} \max_{p(z) \in \mathcal{R}} \int L(z)\, p(z)\, \mathbf{d}z.$$

It can be explained in a simple way. We do not know a precise PDF $p(z)$ and every distribution from $\mathcal{R}$ can be selected. Therefore, we should take the worst or pessimistic distribution providing the largest value of the RF. The minimax criterion appears as an insurance against the worst case because it aims at minimizing the expected loss in the least favorable case [7].

Now we consider the RF $\overline{R}(\alpha)$ in detail. The main assumption below is that the loss function L has one minimum for every $\alpha \in \Lambda$. This assumption is valid for different known types of loss function, for instance, the squared loss function has minimum 0 at

point $z = y - f(\mathbf{x}, \alpha)$. This implies that we have to find the upper expectation of the non-monotone function L having one minimum at point 0. It follows from [8, 9] that

$$\overline{R}_{\mathcal{F}}(\alpha) = \int_{-\infty}^{\overline{F}^{-1}(\tau)} L(z)\mathbf{d}\overline{F}(z) + \int_{\underline{F}^{-1}(\tau)}^{\infty} L(z)\mathbf{d}\underline{F}(z).$$

where τ is one of the roots of the equation

$$h\left(\overline{F}^{-1}(\tau)\right) = h\left(\underline{F}^{-1}(\tau)\right).$$

If the function L is symmetric about 0, i.e., $L(-z) = L(z)$ for all $z \in \mathbb{R}$, then the optimal value of τ does not depend on L and it is determined from the equation

$$\underline{F}^{-1}(\tau) + \overline{F}^{-1}(\tau) = 0. \tag{2}$$

The optimal CDF from the set $\mathcal{F}$ providing the upper bound $\overline{R}_{\mathcal{F}}(\alpha)$ is

$$F_U(z) = \begin{cases} \underline{F}(z), & z \le \overline{F}^{-1}(\tau), \\ \tau, & \overline{F}^{-1}(\tau) < z < \underline{F}^{-1}(\tau), \\ \overline{F}(z), & z \ge \underline{F}^{-1}(\tau). \end{cases}$$

The corresponding optimal PDF is

$$p_U(z) = \begin{cases} \mathbf{d}\underline{F}(z)/\mathbf{d}z, & z \le \overline{F}^{-1}(\tau), \\ 0, & \overline{F}^{-1}(\tau) < z < \underline{F}^{-1}(\tau), \\ \mathbf{d}\overline{F}(z)/\mathbf{d}z, & z \ge \underline{F}^{-1}(\tau). \end{cases}$$

Hence

$$\overline{R}(\alpha) = \int L(z)\, p_U(z)\, \mathbf{d}z = \int_{-\infty}^{\overline{F}^{-1}(\tau)} L(z)\mathbf{d}\overline{F}(z) + \int_{\underline{F}^{-1}(\tau)}^{\infty} L(z)\mathbf{d}\underline{F}(z). \tag{3}$$

The optimal values of parameters α can be found by minimizing $\overline{R}(\alpha)$ over $\alpha \in \Lambda$.

4.2 The minimin strategy

The minimin RF with respect to the minimin strategy is of the form:

$$\underline{R}(\alpha_{\text{opt}}) = \min_{\alpha \in \Lambda} \min_{p(z) \in \mathcal{R}} \int L(z)\, p(z)\, \mathbf{d}z.$$

It can be explained in the following way. Our goal is to minimize the risk measure $\underline{R}$. This implies that we should take a distribution p from $\mathcal{R}$ providing the minimum for the RF by arbitrary α. In other words, we have to find the lower expectation of the non-monotone function L having one minimum at point 0. It follows from [8, 9] that

$$\underline{R}(\alpha) = L(0)\left[\overline{F}(0) - \underline{F}(0)\right] + \int_{-\infty}^{0} L(z)\mathbf{d}\underline{F}(z) + \int_{0}^{\infty} L(z)\mathbf{d}\overline{F}(z).$$

The optimal CDF from the set $\mathcal{F}$ providing the lower bound for the expectation is

$$F_L(z) = \begin{cases} \underline{F}(z), & z \le 0, \\ \overline{F}(z), & z > 0. \end{cases}$$

The optimal CDF has a jump at point $z = 0$. The corresponding optimal PDF is

$$p_L(z)=\begin{cases}\mathbf{d}\underline{F}(z)/\mathbf{d}z, & z<0,\\ \left(\overline{F}(0)-\underline{F}(0)\right)\delta_0(z), & z=0,\\ \mathbf{d}\overline{F}(z)/\mathbf{d}z, & z>0.\end{cases}$$

Here $\delta_0(z)$ is the Dirac function which has unit area concentrated in the immediate vicinity of points 0. Note that $L(0)=0$. This implies that

$$\underline{R}(\alpha)=\int_{-\infty}^{0}L(z)\mathbf{d}\underline{F}(z)+\int_{0}^{\infty}L(z)\mathbf{d}\overline{F}(z). \tag{4}$$

5. Lower and upper risk functionals by nonparametric estimation

Let us consider the bounds for the empirical CDF (1) produced by KS confidence limits with the critical value of the test statistic $\nu=d_{n,1-\gamma}$. The lower and upper PDFs in this case totally depend on the value of ν. They are weighted sums of Dirac functions where the number of terms and their weights are different and depend on ν. Assume that $z_1<z_2<\ldots<z_n$. The following cases of ν can be considered.

Case 1: For $\nu<1/n$, $\nu<1/2$,

$$p_L(z)=\left(n^{-1}-\nu\right)\delta_{z_1}(z)+n^{-1}\sum_{i=2}^{n}\delta_{z_i}(z)+\nu\delta_q(z),\ q\to\infty,$$

$$p_U(z)=\nu\delta_r(z)+n^{-1}\sum_{i=1}^{n-1}\delta_{z_i}(z)+\left(n^{-1}-\nu\right)\delta_{z_n}(z),\ r\to-\infty.$$

Case 2: For $1/n\le\nu<2/n$, $\nu<1/2$,

$$p_L(z)=\left(2n^{-1}-\nu\right)\delta_{z_2}(z)+n^{-1}\sum_{i=3}^{n}\delta_{z_i}(z)+\nu\delta_q(z),\ q\to\infty,$$

$$p_U(z)=\nu\delta_r(z)+n^{-1}\sum_{i=1}^{n-2}\delta_{z_i}(z)+\left(2n^{-1}-\nu\right)\delta_{z_{n-1}}(z),\ r\to-\infty.$$

Case k: For $(k-1)/n\le\nu<k/n$, $\nu<1/2$,

$$p_L(z)=\left(kn^{-1}-\nu\right)\delta_{z_k}(z)+n^{-1}\sum_{i=k+1}^{n}\delta_{z_i}(z)+\nu\delta_q(z),\ q\to\infty,$$

$$p_U(z)=\nu\delta_r(z)+n^{-1}\sum_{i=1}^{n-k}\delta_{z_i}(z)+\left(kn^{-1}-\nu\right)\delta_{z_{n-k+1}}(z),\ r\to-\infty.$$

By having the lower and upper PDFs, we can try to find the optimal PDFs and substitute them into the expressions for the RF according to the minimax and minimin strategies.

5.1 The minimax strategy by nonparametric estimation

First, we find the optimal CDF and PDF in order to use (3). The optimal CDF has to satisfy equality (2). We again study different cases. It is important for us to know how many jumps the optimal CDF has on the left and on the right of 0. Moreover, it is important to know their sizes. Note that the values q and r are constant, they do not depend on α, and their contribution to the RF is constant. Therefore, by taking some finite values q and r, we have the opportunity not to consider these points. The RF will

be written below without the possible terms $\nu L(q)$ and $\nu L(r)$.

The optimal CDF coincides with the upper CDF for negative values of its argument. It also coincides with the lower CDF for positive values of its argument. Let $z_l < 0$ and $z_{l+1} > 0$ be two points of the CDF around 0. The following cases of the relationship of ν and n can be considered.

Case 1: $\nu < 1/n$, $\nu < 1/2$.

Subcase 1.1: $2\nu < 1/n$. In order to satisfy equality (2), the optimal CDF has a small jump of size $n^{-1} - 2\nu$ at point $z_{l+1} = \underline{F}^{-1}(\tau)$ (see Fig. 1, where the empirical CDF is presented by the dashed line having points depicted by stars, the lower and upper bounds are presented by the two thin lines and the optimal CDF by the thick line). This can be explained in the following way. Let us imagine a smoothed CDF instead of the step-wise lower and upper CDFs. By smoothing the CDFs, we replace all fixed (constant) intervals by increasing functions with a very small increment. The unique value τ of $F(z)$ for which equation (2) has a solution is shown in Fig. 1. Since $|z_l| > |z_{l+1}|$, then $z_{l+1} = \underline{F}^{-1}(\tau)$ and $\overline{F}^{-1}(\tau) > z_l$. Hence, the optimal CDF has a small jump at point z_{l+1}. By analyzing all jumps, we can conclude that the optimal CDF has l jumps of size n^{-1} on the left of 0 ($z < 0$), $n - l - 1$ jumps of size n^{-1} on the right of 0 ($z > 0$), and one jump, say the j-th jump, of size $n^{-1} - 2\nu$ closest to 0, which arises due to the transition from the upper CDF to the lower CDF around 0. We do not consider here two jumps with identical weights ν at points q and r. In sum, there are $n-1$ jumps of size n^{-1} and one jump of size $n^{-1} - 2\nu$. Hence,

$$\overline{R}(\alpha) = \int_{-\infty}^{\infty} L(z) \left(\left(n^{-1} - 2\nu\right) \delta_{z_j}(z) + n^{-1} \sum_{i=1, i \neq j}^{n} \delta_{z_i}(z) \right) \mathbf{d}z$$

$$= \left(n^{-1} - 2\nu\right) L(z_j) + n^{-1} \sum_{i=1, i \neq j}^{n} L(z_i).$$

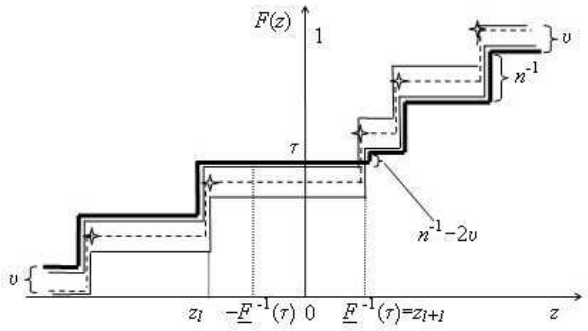


Figure 1: Case 1.1 of the relationship of ν and n

The following cases can be analyzed in the same way.

Subcase 1.2: $2\nu \geq 1/n$. The optimal CDF has $n-2$ jumps of size n^{-1} and one jump

of size $2n^{-1}-2\nu$, which arises due to the transition from the upper CDF to the lower CDF around 0 (see Fig. 2). Hence,

$$\overline{R}(\alpha)=\left(2n^{-1}-2\nu\right)L(z_j)+n^{-1}\sum_{i=1,i\neq j}^{n-1}L(z_i).$$

Case 2: $1/n\le\nu<2/n$, $\nu<1/2$.

Subcase 2.1: $2\nu<3/n$. The optimal CDF has $n-3$ jumps of size n^{-1} and one jump of size $3n^{-1}-2\nu$ closest to 0 (see Fig. 3). Hence,

$$\overline{R}(\alpha)=\left(3n^{-1}-2\nu\right)L(z_j)+n^{-1}\sum_{i=1,i\neq j}^{n-2}L(z_i).$$

Subcase 2.2: $2\nu\ge3/n$. The optimal CDF has $n-2$ jumps of size n^{-1} and one jump of size $2n^{-1}-2\nu$, which arises due to the transition from the upper CDF to the lower CDF around 0. Hence,

$$\overline{R}(\alpha)=\left(4n^{-1}-2\nu\right)L(z_j)+n^{-1}\sum_{i=1,i\neq j}^{n-3}L(z_i).$$

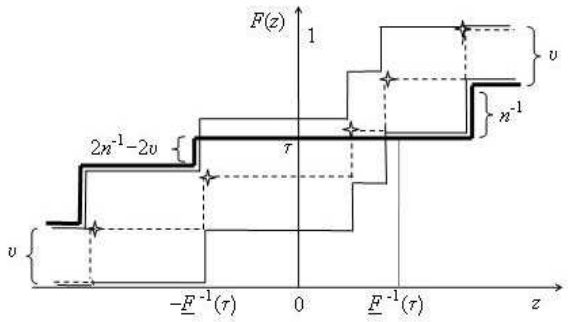


Figure 2: Case 1.2 of the relationship of ν and n

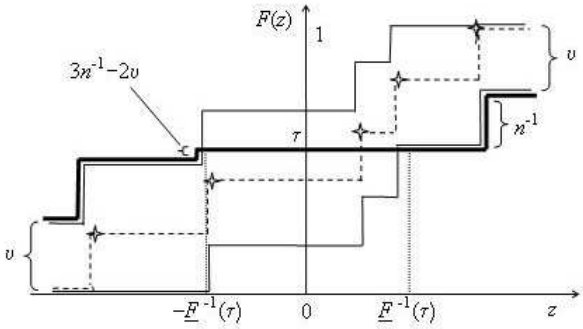


Figure 3: Case 2.2 of the relationship of ν and n

Case *k*: $(k-1)/n\le\nu<k/n$, $\nu<1/2$.

Subcase k.1: $2\nu<(2k-1)/n$. The optimal CDF has $n-2k+1$ jumps of size n^{-1} and

one jump of size $(2k-1)n^{-1}-2\nu$ closest to 0. Hence,

$$\overline{R}(\alpha)=\left((2k-1)n^{-1}-2\nu\right)L(z_j)+n^{-1}\sum_{i=1,i\neq j}^{n-2k+2}L(z_i).$$

Subcase k.2: $2\nu\geq(2k-1)/n$. The optimal CDF has $n-2k$ jumps of size n^{-1} and one jump of size $2kn^{-1}-2\nu$ closest to 0. Hence,

$$\overline{R}(\alpha)=\left(2kn^{-1}-2\nu\right)L(z_j)+n^{-1}\sum_{i=1,i\neq j}^{n-2k+1}L(z_i).$$

It can also be seen from the expressions for the RF that it is not necessary to separate the points lying on the left and on the right of 0. If we consider Subcase k.1, then it is important for us to select a point z_j from n points $z_1,...,z_n$ and a subset of $n-2k+1$ points whose indices are denoted by M_j such that $M_j\subset\{1,...,n\}$ and $j\notin M_j$. In other words, we select all possible subsets M_j for every $j=1,...,n$, and then we compute the values of risk measure $\overline{R}_{M_j}(\alpha)$ for every j and M_j. Hence,

$$\alpha_{\text{opt}}=\arg\min_{\alpha}\max_{j=1,...,n}\max_{M_j}\overline{R}_{M_j}(\alpha).$$

The same is valid for Subcase k.2, where every subset M_j has $n-2k$ elements.

Consider the case $L(z)=z^2$. In order to unify the next expressions, the weight of the j-th jump will be denoted w, i.e., $w=(2k-1)n^{-1}-2\nu$ for Subcase k.1 and $w=2kn^{-1}-2\nu$ for Subcase k.2. Then we write the following optimization problem for minimizing the risk measure by a certain value j and a subset M_j:

$$\overline{R}_{M_j}(\alpha)=n^{-1}\sum_{i\in M_j}\left(y_i-f(\mathbf{x}_i,\alpha)\right)^2+w\left(y_j-f(\mathbf{x}_j,\alpha)\right)^2\rightarrow\min_{\alpha}.$$

Let us study the simplest case of the reliability growth function $f(\mathbf{x},\alpha)=\alpha_1x+\alpha_2$, i.e., $\alpha=(\alpha_1,\alpha_2)$ and $\mathbf{x}=(x)$. Denote

$$G_j=\sum_{i\in M_j}x_i^2+nwx_j^2,\ H_j=\sum_{i\in M_j}x_i+nwx_j,\ Q_j=\sum_{i\in M_j}y_ix_i+nwy_jx_j,\ T_j=\sum_{i\in M_j}y_i+nwy_j.$$

Then we obtain the optimal parameters

$$\alpha_1=\frac{n(1-2\nu)Q_j-H_jT_j}{n(1-2\nu)G_j-H_j^2},\ \alpha_2=\frac{G_jT_j-H_jQ_j}{n(1-2\nu)G_j-H_j^2}.$$

By taking $j=1,...,n$ and all possible subsets M_j, we get a set of (α_1,α_2). By substituting every pair of the parameters into the expression for $\overline{R}_{M_j}(\alpha)$, we get a set of risk measures. The optimal values of (α_1,α_2) correspond to the largest value of $\overline{R}_{M_j}(\alpha)$.

It is interesting to note here that the larger the value of ν is (or the more imprecision there is), the smaller the number of points of training data is which contribute to the risk measure. If we use the precise empirical distribution, so the special case with $\nu=0$, we have Case 1.1 and

$$\overline{R}(\alpha)=n^{-1}L(z_j)+n^{-1}\sum_{i=1,i\neq j}^{n}L(z_i)=n^{-1}\sum_{i=1}^{n}L(z_i).$$

Of course, this special case gives the standard empirical RF. In particular, when $L(z)=z^2$, this special case is just the standard least-squares method. As a further special case, consider values of ν for which $n=2k$, i.e., $2\nu \geq (n-1)/n$. For such values, we have $\overline{R}_j(\alpha)=(1-2\nu)L(z_j)$ and $\alpha_{\text{opt}}=\arg\min_{\alpha}\max_{j=1,\ldots,n}L(z_j)$. In other words, in this case the model fitting is based only on single observations corresponding to each parameter value, which can be considered as a degenerate case which is not appropriate for fitting meaningful models, hence it is recommended not to use such large values of ν.

5.2 The minimin strategy by nonparametric estimation

If we assume $L(0)=0$, then the optimal CDF in (4) coincides with the lower CDF for $z<0$ and with the upper CDF for $z>0$, and it has an additional jump at point $z=0$. We again study different cases for the relationship of n and ν.

Case 1: $\nu<1/n$, $\nu<1/2$. The optimal CDF has l jumps of size n^{-1} and one jump of size $n^{-1}-\nu$ at point z_1 to the left of 0 ($z<0$), $n-l-2$ jumps of size n^{-1} and one jump of size $n^{-1}-\nu$ at point z_n to the right of 0 ($z>0$), and one jump, say the j-th jump, of size 2ν at point 0, which arises due to the transition from the lower CDF to the upper CDF at point 0. To summarize, there are $n-2$ jumps of size n^{-1}, two jumps of size $n^{-1}-\nu$ and one jump of size 2ν. Hence,

$$\underline{R}(\alpha)=\int_{-\infty}^{\infty}L(z)\left(\left(n^{-1}-\nu\right)\left(\delta_{z_1}(z)+\delta_{z_n}(z)\right)+n^{-1}\sum_{i=2}^{n-1}\delta_{z_i}(z)+2\nu\delta_0(z)\right)\mathrm{d}z$$

$$=\left(n^{-1}-\nu\right)\left(L(z_1)+L(z_n)\right)+n^{-1}\sum_{i=2}^{n-1}L(z_i).$$

Case 2: $1/n\leq\nu<2/n$, $\nu<1/2$. The optimal CDF has $n-4$ jumps of size n^{-1}, two jumps of size $2n^{-1}-\nu$ and one jump of size 2ν. Hence,

$$\underline{R}(\alpha)=\left(2n^{-1}-\nu\right)\left(L(z_1)+L(z_n)\right)+n^{-1}\sum_{i=2}^{n-3}L(z_i).$$

Case k: $(k-1)/n\leq\nu<k/n$, $\nu<1/2$. The optimal CDF has $n-2k$ jumps of size n^{-1}, two jumps of size $kn^{-1}-\nu$ and one jump of size 2ν. Hence,

$$\underline{R}(\alpha)=\left(kn^{-1}-\nu\right)\left(L(z_1)+L(z_n)\right)+n^{-1}\sum_{i=2}^{n-2k-2}L(z_i).$$

It can be seen from the above that it is not necessary to separate the points lying to the left and to the right of 0. It is important for us to select two points z_j and z_l from n points $z_1,\ldots,z_n$ and a subset of $n-2k$ points whose indices are denoted N_{jl} such that $N_{jl}\subset\{1,\ldots,n\}$ and $j,l\notin N_{jl}$. In other words, we select all possible subsets N_{jl} for every pair $j,l=1,\ldots,n$, and then we compute the corresponding values of risk measure $\underline{R}_{N_{jl}}(\alpha)$. Hence,

$$\alpha_{\text{opt}}=\arg\min_{\alpha}\min_{j=1,\ldots,n,\,l=1,\ldots,n}\min_{N_{jl}}\underline{R}_{N_{jl}}(\alpha).$$

Let us consider the case $L(z)=z^2$. We write the following optimization problem for

minimizing $\underline{R}_{N_{jl}}(\alpha)$:

$$\underline{R}_{N_{jl}}(\alpha) = n^{-1}\sum_{i\in N_{jl}}\left(y_i - f(\mathbf{x}_i,\alpha)\right)^2 + \left(kn^{-1}-\nu\right)\left(\left(y_j - f(\mathbf{x}_j,\alpha)\right)^2 + \left(y_l - f(\mathbf{x}_l,\alpha)\right)^2\right) \to \min_{\alpha}.$$

Let us take $f(\mathbf{x},\alpha) = \alpha_1 x + \alpha_2$ and denote

$$G_{jl} = \sum_{i\in N_{jl}} x_i^2 + \left(k-n\nu\right)\left(x_j^2 + x_l^2\right),\ \ H_{jl} = \sum_{i\in N_{jl}} x_i + \left(k-n\nu\right)\left(x_j + x_l\right),$$

$$Q_{jl} = \sum_{i\in N_{jl}} y_i x_i + \left(k-n\nu\right)\left(y_j x_j + y_l x_l\right),\ \ T_{jl} = \sum_{i\in N_{jl}} y_i + \left(k-n\nu\right)\left(y_j + y_l\right).$$

Then the optimal parameter values are

$$\alpha_1 = \frac{n\left(1-2\nu\right)Q_{jl} - H_{jl}T_{jl}}{n\left(1-2\nu\right)G_{jl} - H_{jl}^2},\ \ \alpha_2 = \frac{G_{jl}T_{jl} - H_{jl}Q_{jl}}{n\left(1-2\nu\right)G_{jl} - H_{jl}^2}.$$

By taking $j = 1,...,n$, $l = 1,...,n$, all subsets N_{jl}, and substituting every pair of the parameters into the expression for $\underline{R}_{N_{jl}}(\alpha)$, we get a set of risk measures. The optimal values of (α_1,α_2) correspond to the smallest value of $\underline{R}_{N_{jl}}(\alpha)$.

If we use the precise empirical distribution, i.e., $\nu = 0$, we have Case 1 and

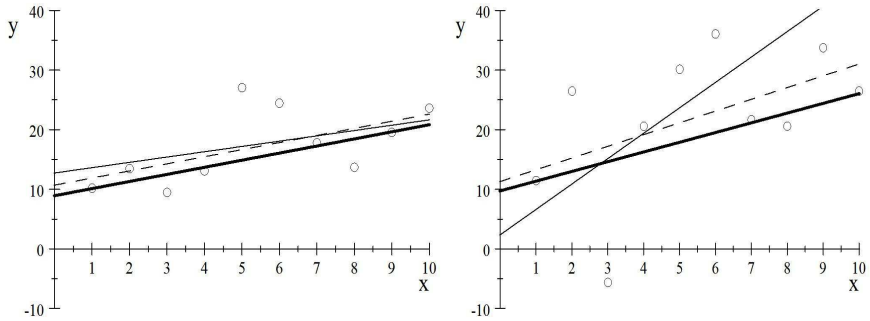
$$\underline{R}(\alpha) = n^{-1}L(z_1) + n^{-1}L(z_n) + n^{-1}\sum_{i=2}^{n-1} L(z_i) = n^{-1}\sum_{i=1}^{n} L(z_i).$$

We again get the standard empirical RF and the least-squares method in this case, as is easy to explain: If we have a precise probability distribution, the optimal probability distribution of course coincides with the precise distribution irrespective of the accepted strategy. The risk measure decreases as ν increases. If the value of ν is such that $n = 2k$, then $\underline{R}(\alpha) = 0$, which implies that there is no meaningful optimistic solution as discussed at the end of Section 5.1.

6. Linear regression example

Suppose that the "true" function f is known and is $f(\mathbf{x},\alpha) = 2x + 5$. We randomly generate additive normally distributed noise with expectation 0 and, in two separate cases, variances 4 and 16. We thus generate 10 observations points (x_i, y_i) and use $\nu = 0.368$ corresponding to confidence level $1-\gamma = 0.9$ [5]. We use the loss function $L(z) = z^2$.

Case a. Variance of the noise is 4. The noisy measurements are represented in Fig. 4(a) by circles. The computed parameter values according to the minimax strategy are $\alpha_{1\text{opt}} = 2.36$ and $\alpha_{2\text{opt}} = 2.07$, resulting in the thin line in Fig. 4(a). The minimin strategy gives parameter values $\alpha_{1\text{opt}} = 2.02$ and $\alpha_{2\text{opt}} = 5.51$ (the thick line). The standard least-squares method gives $\alpha_{1\text{opt}} = 2.18$ and $\alpha_{2\text{opt}} = 3.75$ (the thin dashed line).



(a) (b)

Figure 4: Fitted linear functions with noise variances 4 (a) and 16 (b)

Case b. Variance of the noise is 16. The parameter values according to the minimax strategy are $\alpha_{1\text{opt}} = 4.26$ and $\alpha_{2\text{opt}} = 2.37$ (the thin line in Fig. 4(b)) and according to the minimin strategy $\alpha_{1\text{opt}} = 1.62$ and $\alpha_{2\text{opt}} = 9.74$ (the thick line). The standard least-squares method gives $\alpha_{1\text{opt}} = 1.97$ and $\alpha_{2\text{opt}} = 11.29$ (the thin dashed line).

Fig. 4 illustrates our earlier theoretical arguments that the minimax strategy, for the chosen confidence level for the KS bounds and the chosen loss function, aims at minimizing a number of the largest (in absolute value) residuals, the number being determined by the chosen confidence level and the number of observations, together with one further residual which typically gets a lower weight. The minimin strategy, on the other hand, aims to minimize a number of the smallest (in absolute value) residuals, giving varying weights to these residuals. Hence, one could say that the minimax strategy focuses on the most scattered data while the minimin strategy focuses on the most concentrated data, both with regard to distance from a straight line (the model that is being fit). For example, in Fig. 4(b) the optimal minimin strategy minimizes the residuals corresponding to the x values 1 and 10 with unit weights and the residuals corresponding to the x values 7 and 8 with weights $k - n\nu = 0.32$. The model fit resulting from the minimax strategy minimizes the residuals corresponding to the x values 2 and 3 with unit weights and the residual corresponding to the x value 6 with weight 0.64. For the standard least-squares method, all residuals are taken into account equally.

7. Goel-Okumoto software reliability growth model

We now illustrate the application of our statistical approach to the well-known Goel-Okumoto software reliability growth model [10], which is based on the non-homogeneous Poisson process. According to this model, the mean number of failures occurring up to time x is given by the expression $f(x, \alpha) = \alpha_1(1 - e^{-\alpha_2 x})$. Here the parameter α_1 is interpreted as the number of initial faults in the software and the parameter α_2 is the fault detection rate which is related to the reliability growth rate in the testing process. Using results of Section 5, the system of equations for computing the parameters by the minimax strategy is

$$\sum_{i\in M_j}\left(y_i-\alpha_1(1-e^{-\alpha_2 x_i})\right)x_i e^{-\alpha_2 x_i}+nw\left(y_j-\alpha_1(1-e^{-\alpha_2 x_j})\right)x_j e^{-\alpha_2 x_j}=0,$$

$$\sum_{i\in M_j}\left(y_i-\alpha_1(1-e^{-\alpha_2 x_i})\right)(1-e^{-\alpha_2 x_i})+nw\left(y_j-\alpha_1(1-e^{-\alpha_2 x_j})\right)(1-e^{-\alpha_2 x_j})=0.$$

We introduce the notation

$$G_j(\alpha_2)=nwy_j x_j e^{-\alpha_2 x_j}+\sum_{i\in M_j} y_i x_i e^{-\alpha_2 x_i},$$

$$H_j(\alpha_2)=nw(1-e^{-\alpha_2 x_j})x_j e^{-\alpha_2 x_j}+\sum_{i\in M_j}(1-e^{-\alpha_2 x_i})x_i e^{-\alpha_2 x_i},$$

$$R_j(\alpha_2)=nwy_j(1-e^{-\alpha_2 x_j})+\sum_{i\in M_j} y_i(1-e^{-\alpha_2 x_i}),$$

$$Q_j(\alpha_2)=nw(1-e^{-\alpha_2 x_j})^2+\sum_{i\in M_j}(1-e^{-\alpha_2 x_i})^2.$$

Using this notation, the system of equations is given by

$$G_j(\alpha_2)-\alpha_1 H_j(\alpha_2)=0,\ R_j(\alpha_2)-\alpha_1 Q_j(\alpha_2)=0.$$

The optimal value for the parameter α_2 can be found from the equation

$$G_j(\alpha_2)Q_j(\alpha_2)-R_j(\alpha_2)H_j(\alpha_2)=0.$$

The parameter α_1 can then be computed from either one of two equations above by substituting the optimal value for α_2. The system of equations for computing the parameters by the minimin strategy can be obtained similarly.

We use a software reliability data set provided by the Shuttle Ground System for software for flight controllers at the Johnson Space Center [11]. All software errors have been divided into three types: critical errors, major errors and minor errors. We use only 'minor errors' and 8 tests for analyzing the proposed imprecise model (see Table 1), and we use the loss function $L(z)=z^2$.

Table 1: Software reliability data

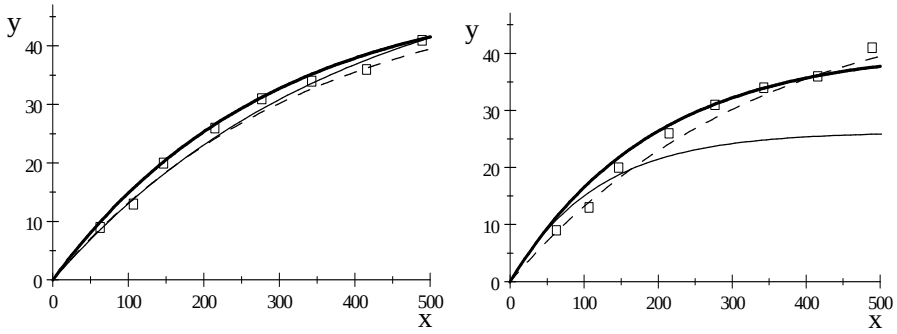
Number of test	Test (hours)	Numbers of errors	Number of test	Test (hours)	Numbers of errors
i	t_i-t_{i-1}	k_i	i	t_i-t_{i-1}	k_i
1	62.5	9	5	62	5
2	44	4	6	66	3
3	40	7	7	73	2
4	68	6	8	73.5	5

The cumulative measurements are presented in Fig. 5 by boxes. First, we take confidence level $1-\gamma=0.8$ leading to $v=0.358$. The optimal parameter values according to the minimax strategy are $\alpha_{1\text{opt}}=56.94$ and $\alpha_{2\text{opt}}=0.0027$ (the thin line in Fig. 5(a)), whilst the minimin strategy gives $\alpha_{1\text{opt}}=50.31$ and $\alpha_{2\text{opt}}=0.0035$ (the thick line). The standard least-squares method leads to $\alpha_{1\text{opt}}=50.82$ and $\alpha_{2\text{opt}}=0.0032$ (the thin dashed line).

Secondly, we take confidence level $1-\gamma=0.9$ and $v=0.408$. Actually, as there are only 8 observations in total, this high confidence level for the KS bounds implies that only

two residuals are considered in each of the minimax and minimin methods. The optimal parameter values according to the minimax strategy are $\alpha_{1\text{opt}} = 26.24$ and $\alpha_{2\text{opt}} = 0.0085$ (the thin line in Fig. 5(b)), whilst the minimin strategy gives $\alpha_{1\text{opt}} = 40.76$ and $\alpha_{2\text{opt}} = 0.0052$ (the thick line). Of course, the standard least-squares method is not affected by different confidence level for the KS bounds and gives the same optimal model fit as in Fig. 5(a) (the thin dashed line). The minimin strategy actually only uses the two residuals corresponding to the fifth and seventh x values, both with equal weight 0.736. The minimax strategy uses the residual corresponding to the third x value with unit weight and the residual corresponding to the second x value with weight 0.472. This actually results in a model fit that is far outside the majority of the observed points, and which therefore may be deemed not to be appropriate. This is a direct consequence of the two x values used in the optimal minimax model fit being both small, leading to a good fit for small x values but poor fit for larger x values as no information from such values has been taken into account. This mainly serves as a warning that, with only few observations, one cannot increase the confidence level of the KS bounds too much as the resulting model fits get determined by very small numbers of observations. Hence, one could say that in such cases 'robustness has been taken too far'.

It can be seen from the obtained results that the growth functions $f(x,\alpha)$ by the confidence level 0.8 are very close each other for both strategies and for the standard least-squares method. However, this probability is rather small for making decisions. By increasing the confidence level, we observe that the minimax strategy avoids some risk, provides a more cautious prediction and "compensates" the steep demands to the confidence level.



(a) (b)

Figure 5: Fitted reliability growth models using confidence levels 0.8 (a) and 0.9 (b)

8. Conclusion

The method presented in this paper achieves model fits for a variety of practically relevant scenarios, including nonparametric regression models for a wide range of loss functions. In traditional statistical regression models, typically a strong modeling assumption on the residuals is made, usually that these are independent and identically normally distributed with zero mean and equal variances. This assumption is then used for many inferences that go beyond fitting the model by estimating parameter values, enabling confidence intervals for the model parameters and for predictions. For the theory presented in this

paper, no such inferential framework has yet been established, this is a key topic for future research. It should be remarked here that many popular statistical methods that are widely used and advocated nowadays, for example many smoothing methods, also do not have a clear inferential framework beyond fitting of models and are mostly used to provide appropriate data summaries and insights into data structures.

The method is immediately applicable for multivariate models, as long as the dependent variable and hence the residuals are one-dimensional. However, the optimisation of the risk function will become more demanding, so for large-scale problems the implementation requires further development of efficient algorithms.

The use of KS bounds, together with the minimin and minimax criteria, fits with some elements of theories of imprecise probabilities [12], which has also proven to be successful in many reliability applications [13,14]. However, the minimin and minimax model fits are not extremes with regard to all possible model fits that would correspond to the same KS bounds combined with criteria 'in between' the two extremes of minimin and minimax, namely which would minimize another characteristic of the risk function than its minimum or maximum value over all distributions within the KS bounds. It would be of interest to explore such alternative characteristics, so different optimality criteria, which may lead to a family of model fits. It would be interesting to study what kind of criteria would lead to reasonably informative families of model fits, and which characteristics of the set of residuals corresponding to a specific criterion influence the model fit most.

The minimin method presented in this paper excludes a number of the smallest and largest values of residuals, where the number is determined by the chosen confidence level for the KS bounds. The effect of this has been illustrated in the examples. This neglecting of largest (in absolute value) residuals has conceptual links to some methods in robust statistics, where reduction of influence of possible outliers is also achieved by some procedures by effectively neglecting the most extreme values. The relation between the method presented in this paper and regression methods from robust statistics must be studied in detail, which is an important topic for future research that is likely to point to further ways in which this approach can be developed and applied. The minimax method presented also gives varying weights to residuals, further study is required to see if this has relations to existing concepts in robust statistics. Care must be taken if the actual number of observations that determine the minimax and minimin model fits become very small, as it may take `robustness too far' as commented on in the example in Section 7.

The presentation in the paper has been focused on simple models. However, the main concept can be applied to a wide class of regression-like models which provide a large number of research challenges and application opportunities in many areas of applied statistics including reliability. It will further be of interest to study if similar ideas can be combined with ideas underlying some popular semi-parametric models such as Cox' proportional hazards model. If this is possible, then the approach based on KS bounds may provide more robust methods for inference than the established methods. While the explicit optimization results presented in this paper, which enable implementation of the method, are derived for the KS bounds, it is an interesting topic for future research to explore the use of bounds resulting from alternative statistical inference approaches, for example P-boxes [4] or nonparametric predictive inference [15,16]. Such alternative approaches may have attractive specific inferential features, their links to fitting regression-type models as presented in this paper must be studied in great detail. If the implementation of methods such as presented in this paper would be sound, from

foundational perspective, when applied to such alternative inferential approaches, we expect that theoretical results to simplify the optimizations involved can be developed along similar lines to those presented in this paper.

References

[1]. Lawless JF. *Statistical Models and Methods for Lifetime Data*. New York: Wiley; 1982.

[2]. Hastie T, Tibshirani R, Friedman J. *The Elements of Statistical Learning: Data Mining, Inference and Prediction*. New York: Springer; 2001.

[3]. Vapnik V. *Statistical Learning Theory*. New York: Wiley; 1998.

[4]. Ferson S, Kreinovich V, Ginzburg L, Myers DS, Sentz K. *Constructing probability boxes and Dempster-Shafer structures*. Sandia National Laboratories; 2003. SAND2002-4015.

[5]. Johnson NL, Leone F. *Statistics and experimental design in engineering and the physical sciences*. vol.1. New York: Wiley; 1964.

[6]. Xing F, Guo P. *Support vector regression for software reliability growth modeling and prediction*. In: Lecture Notes in Computer Science. Berlin, Heidelberg: Springer; 2005. p. 925–930.

[7]. Robert CP. *The Bayesian Choice*. New York: Springer; 1994.

[8]. Utkin LV. *Risk analysis under partial prior information and non-monotone utility functions*. International Journal of Information Technology and Decision Making. 2007;6(4):625–647.

[9]. Utkin LV, Destercke S. *Computing expectations with continuous p-boxes: Univariate case*. International Journal of Approximate Reasoning. 2009;50(5):778–798.

[10]. Goel AL, Okomoto K. *Time dependent error detection rate model for software reliability and other performance measures*. IEEE Trans. Reliab. 1979:R-28:206–211.

[11]. Misra PN. *Software reliability analysis*. IBM Systems Journal. 1983:22(3):262–270.

[12]. Walley P. *Statistical reasoning with imprecise Probabilities*. New York: Chapman and Hall; New York, 1991.

[13]. Utkin LV, Coolen FPA. *Imprecise reliability: an introductory overview*. In: Levitin G (ed.), Computational Intelligence in Reliability Engineering, Volume 2: New Metaheuristics, Neural and Fuzzy Techniques in Reliability. Springer; Berlin, 2007:261-306.

[14]. Coolen FPA, Utkin LV. *Imprecise reliability*. In: Lovric M (ed.), International Encyclopedia of Statistical Science. Springer; Berlin, 2010: to appear.

[15]. Augustin T, Coolen FPA. *Nonparametric predictive inference and interval probability*. Journal of Statistical Planning and Inference. 2004:124:251-272.

[16]. Coolen FPA. *Nonparametric predictive inference*. In: Lovric M (ed.), International Encyclopedia of Statistical Science. Springer; Berlin, 2010: to appear.

Biographical Sketch

Lev V. Utkin is Professor of Computer Science at Saint Petersburg State Forest Technical Academy, Russia, where he is also the Vice-rector for Research from 2006. He completed his PhD at St.Petersburg Electrotechnical University in 1989 and his DSc at St.Petersburg State Institute of Technology in 2001. He has (co-)authored around 200 journal and conference papers on topics in Statistics, Reliability, Decision making and related fields.

Frank P.A. Coolen is Professor of Statistics at Durham University, UK. He completed his PhD at Eindhoven University of Technology, The Netherlands, in 1994. He serves on editorial boards of 4 international journals of Statistics and Reliability, and edited the Section on Reliability: Mathematical and Statistical Methods for Wiley's Encyclopedia of Quantitative Risk Analysis and Assessment (2008). He has (co-)authored around 150 journal and conference papers on topics in Statistics, Reliability and related fields. He is the main developer of Nonparametric Predictive Inference (www.npi-statistics.com).