Predicting future failure times by using quantile regression

Metrika (2023) 86:543–576
https://doi.org/10.1007/s00184-022-00884-z

Predicting future failure times by using quantile regression

Jorge Navarro[1] · **Francesco Buono**[2]

Received: 23 February 2022 / Accepted: 28 August 2022 / Published online: 27 September 2022

Abstract
The purpose of the paper is to study how to predict the future failure times in a sample from the early failures (type II censored data). We consider both the case of independent and dependent lifetimes. In both cases we assume identically distributed random variables. To predict the future failures we use quantile regression techniques that also provide prediction regions for them. Some illustrative examples show how to apply the theoretical results to simulated and real data sets.

Keywords Order statistics · Copula · Distorted distributions · Quantile regression · Type II censoring

1 Introduction

When in a study one works with a sample of several lifetimes, it is usual to have censored data. Thus, we might have just the exact values of the first r failures (or survival times). The other values are censored (type II censored data). This approach is of interest both in survival and reliability studies. An excellent review of the different situations about ordered and censored data was made in Cramer (2021).

Several tools have been developed to use these censored data. Also, some procedures have been studied to predict the unknown future failure times by using the exact values of the first r early failures. The main results can be seen in Basiri et al. (2016), Barakat et al. (2011), El-Adll (2011), Lawless and Fredette (2005) and in the references therein.

Jorge Navarro and Francesco Buono have contributed equally to this work

✉ Jorge Navarro
jorgenav@um.es

Francesco Buono
francesco.buono3@unina.it

[1] Facultad de Matemáticas, Universidad de Murcia, Campus de Espinardo, Murcia 30100, Murcia, Spain

[2] Università di Napoli Federico II, Via Cintia, Napoli 80126, Napoli, Italy

Recently, Barakat et al. (2022) and Bdair and Raqab (2022) proposed two solutions based on different pivotal quantities for samples of independent and identically distributed (IID) lifetimes with a common mixture of two exponential distributions. The results obtained in the second paper were based on a beta distribution (see next section).

The purpose of this paper is to extend these results in two ways. One way is to consider the IID case with the more general Proportional Hazard Rate (PHR) Cox model. To this end we will use the pivotal quantity proposed by Bdair and Raqab (2022). The other way is to consider also the case of dependent samples. We assume ID lifetimes but the general procedure can be applied to the general case as well (with more complicated expressions). In both cases, to provide such predictions we will use quantile regression (QR) techniques that can also be used to get prediction bands for them, see Koenker (2005) and Navarro et al. (2022). This approach allows us to get an accurate representation of the uncertainty associated to that predictions (especially when we estimate the upper extreme values). Some examples illustrate how to apply the theoretical findings.

The rest of the paper is organized as follows. The main results for the case of independent data are given in Sect. 2 while that for dependent data are in Sect. 3 The examples are placed in Sect. 4. Section 5 contains a simulation study about the coverage probabilities when we estimate the parameter of the PHR model. The conclusions and the main tasks for future research projects are explained in Sect. 6.

2 Independent data

Let $X_1, \ldots, X_n$ be a sample of independent and identically distributed (IID) random variables with a common absolutely continuous distribution function F and with a probability density function (PDF) $f = F'$ (a.e.). Let $\bar{F} = 1 - F$ be the reliability (or survival) function and let $X_{1:n} < \cdots < X_{n:n}$ be the associated ordered data (order statistics). The basic properties for them can be seen in Arnold et al. (2008) and David and Nagaraja (2003).

In many cases $X_1, \ldots, X_n$ are lifetimes (or survival times) of some items. So, in practice, sometimes we just have the first r early failure times $X_{1:n}, \ldots, X_{r:n}$ for some $r < n$. Then what we want is to predict the remaining lifetimes $X_{r+1:n}, \ldots, X_{n:n}$ from the early failures.

The results obtained in this section will be based on the following result extracted from Bdair and Raqab (2022) and the well known Markov property of the order statistics, see Arnold et al. (2008). For completeness, we include the proof here.

Proposition 2.1 (Bdair and Raqab 2022) *Let* $W_{r,s:n} = \bar{F}(X_{s:n})/\bar{F}(X_{r:n})$ *for* $1 \leq r < s \leq n$. *Then the distributions of the conditional random variables* $(W_{r,s:n} \mid X_{1:n} = x_1, \ldots, X_{r:n} = x_r)$ *and* $(W_{r,s:n} \mid X_{r:n} = x_r)$ *coincide with a beta distribution of parameters* $n - s + 1$ *and* $s - r$.

Proof The distributions coincide from Theorem 2.4.3 in Arnold et al. (2008). From expression (2.4.3) in that book (p. 23), the PDF of $(X_{s:n} \mid X_{r:n} = x_r)$ is

$$f_{s|r:n}(x_s \mid x_r) = c \left(\frac{\bar{F}(x_r) - \bar{F}(x_s)}{\bar{F}(x_r)} \right)^{s-r-1} \left(\frac{\bar{F}(x_s)}{\bar{F}(x_r)} \right)^{n-s} \frac{f(x_s)}{\bar{F}(x_r)}$$

for $1 \leq r < s \leq n$ and $x_r < x_s$, where c is the normalizing constant. On the other hand, if $\bar{G}$ is the reliability function of $(W_{r,s:n} \mid X_{r:n} = x_r)$, we get

$$\begin{aligned} \bar{G}\left(\frac{\bar{F}(x_s)}{\bar{F}(x_r)} \right) &= \Pr\left(W_{r,s:n} > \frac{\bar{F}(x_s)}{\bar{F}(x_r)} \mid X_{r:n} = x_r \right) \\ &= \Pr\left(\frac{\bar{F}(X_{s:n})}{\bar{F}(X_{r:n})} > \frac{\bar{F}(x_s)}{\bar{F}(x_r)} \mid X_{r:n} = x_r \right) \\ &= \Pr\left(X_{s:n} < x_s \mid X_{r:n} = x_r \right). \end{aligned}$$

Therefore, its PDF $g = -\bar{G}'$ satisfies

$$g\left(\frac{\bar{F}(x_s)}{\bar{F}(x_r)} \right) \frac{f(x_s)}{\bar{F}(x_r)} = f_{s|r:n}(x_s \mid x_r)$$

and so, by using the preceding expression for $f_{s|r:n}$, we obtain

$$g(w) = c(1-w)^{s-r+1} w^{n-s}$$

for $0 < w < 1$. Therefore, $(W_{r,s:n} \mid X_{r:n} = x_r)$ has a beta distribution with parameters $n - s + 1$ and $s - r$. □

The preceding proposition can be used to get the median regression curve to estimate $X_{s:n}$ from $X_{r:n} = x$ (or from $X_{1:n} = x_1, \ldots, X_{r:n} = x_r$). It can be stated as follows.

Proposition 2.2 *The median regression curve to estimate $X_{s:n}$ from $X_{r:n} = x$ is*

$$m(x) = \bar{F}^{-1}\left(q_{0.5} \bar{F}(x) \right), \tag{2.1}$$

where $q_{0.5}$ is the median of a beta distribution with parameters $n - s + 1$ and $s - r$.

Proof From the expressions obtained in the proof of the preceding proposition we get that

$$\Pr\left(X_{s:n} < x_s \mid X_{r:n} = x_r \right) = 0.5$$

is equivalent to

$$\bar{G}\left(\frac{\bar{F}(x_s)}{\bar{F}(x_r)} \right) = 0.5,$$

where $\bar{G}$ is the reliability function of a beta distribution with parameters $n-s+1$ and $s-r$. This expression leads to $\bar{F}(x_s) = q_{0.5}\bar{F}(x_r)$. Therefore, the first expression is equivalent to $x_s = \bar{F}^{-1}(q_{0.5}\bar{F}(x_r))$ which gives the expression for the median regression curve. □

Remark 2.1 The median $q_{0.5}$ of the beta distribution with parameters $\alpha = n-s+1 > 1$ and $\beta = s-r > 1$ can be approximated by

$$q_{0.5} \approx \frac{\alpha - 1/3}{\alpha + \beta - 2/3} = \frac{n-s+2/3}{n-r+1/3}.$$

By using this approximation, the mean regression curve obtained in (2.1) can be written as

$$m(x) = \bar{F}^{-1}\left(\frac{n-s+2/3}{n-r+1/3}\bar{F}(x)\right)$$

for $r+1 < s < n$. If we prefer to use the exact value of the median, we can use any statistical program to compute $q_{0.5}$. For example, the code in R to compute it is `qbeta(0.5,n-s+1,s-r)`.

Remark 2.2 Instead of the median approach, we could use the mean or the mode of the conditional random variable $(W_{r,s:n} \mid X_{r:n} = x_r)$ in Proposition 2.1. To use the mean, as it has a beta distribution, we get

$$E(W_{r,s:n} \mid X_{r:n} = x_r) = \frac{n-s+1}{n-r+1},$$

that is,

$$E(\bar{F}(X_{s:n}) \mid X_{r:n} = x_r) = \frac{n-s+1}{n-r+1}\bar{F}(x_r).$$

Therefore, $X_{s:n}$ can be predicted from $X_{r:n} = x$ with the curve

$$m_{mean}(x) = \bar{F}^{-1}\left(\frac{n-s+1}{n-r+1}\bar{F}(x)\right).$$

Note that this curve can be different to the classical regression curve $E(X_{s:n} \mid X_{r:n} = x_r)$. Analogously, if we use the mode of the beta distribution (i.e. the maximum likelihood estimator for $W_{r,s:n}$) we get

$$Mode(W_{r,s:n} \mid X_{r:n} = x_r) = \frac{n-s}{n-r-1}$$

for $s < n$ and $s > r+1$ (see p. 219, Johnson et al. 1995), that is, $X_{s:n}$ can be predicted from $X_{r:n} = x$ with the curve

$$m_{mode}(x) = \bar{F}^{-1}\left(\frac{n-s}{n-r-1}\bar{F}(x)\right).$$

In the case $s = r + 1$ with $r < n - 1$, the mode is obtained in the boundary and we can use this predictor with $m_{mode}(x) = x$. In the case $r + 1 < s = n$, the predictor $m_{mode}(x) = \bar{F}^{-1}(0)$ coincides with the right-end point of the support (we could use it when it is finite). Finally, in the case $r + 1 = s = n$ the mode is not unique since $(W_{r,s:n} \mid X_{r:n} = x_r)$ has a uniform distribution in $(0, 1)$ and we can replace this predictor with $m_{mode}(x) = \bar{F}^{-1}(0.5\bar{F}(x))$ (which coincides with the mean estimator).

The same approach can be used to determine prediction bands for these future failure times from the quantiles of a beta distribution. Thus, if we want to get a prediction interval of size $\gamma = \beta - \alpha$, where $\alpha, \beta, \gamma \in (0, 1)$ and q_α and q_β are the respective quantiles of the above beta distribution, then we use that

$$\Pr\left(\bar{F}^{-1}\left(q_\beta \bar{F}(x)\right) \leq X_{s:n} \leq \bar{F}^{-1}\left(q_\alpha \bar{F}(x)\right) \mid X_{r:n} = x\right) = \gamma. \tag{2.2}$$

For example, the centered 90% prediction band is obtained with $\beta = 0.95$ and $\alpha = 0.05$ as

$$C_{90} = \left[\bar{F}^{-1}\left(q_{0.95}\bar{F}(x)\right), \bar{F}^{-1}\left(q_{0.05}\bar{F}(x)\right)\right].$$

Sometimes one needs bottom (or lower) prediction bands starting at $X_{r:n} = x$. For example, the bottom 90% prediction band is obtained with $\beta \to 1$ and $\alpha = 0.1$ as

$$B_{90} = \left[x, \bar{F}^{-1}\left(q_{0.1}\bar{F}(x)\right)\right].$$

As we will see in the examples provided in Sect. 4, these prediction bands represent better the uncertainty in the prediction of $X_{s:n}$ from $X_{r:n}$. In particular the area of these bands will increase with $s - r$ (as expected). It is worth mentioning that the quantiles q_z of a beta distribution (including the median) are available in many statistical programs (for example, in R, q_z is obtained with the code `qbeta(z,a,b)`, with $a = n - s + 1$ and $b = s - r$ in our case). These quantiles could also be obtained from the procedure given in Van Dorp and Mazzuchi (2000). Moreover, in the following proposition we show that the exponential distribution is characterized in terms of its quantile regression curve. It is the unique distribution with quantile regression curves that are lines with slope equal to one.

Proposition 2.3 *Let X be a random variable with support $(0, \infty)$, with absolutely continuous reliability function $\bar{F}$ and with quantile regression curve of order α $m_\alpha(x) = \bar{F}^{-1}(q_\alpha \bar{F}(x))$. Then,*

$$m_\alpha(x) = x + c_\alpha \tag{2.3}$$

for all $x > 0$ and all $\alpha \in (0, 1)$, where c_α is a positive constant depending on α if, and only if, X is exponentially distributed.

Proof Suppose X is exponentially distributed with parameter $\theta > 0$. Then, $\bar{F}(x) = e^{-\theta x}$ and $\bar{F}^{-1}(x) = -\frac{1}{\theta}\log x$. Therefore, the quantile regression curve of order α is given by

$$m_\alpha(x) = \bar{F}^{-1}(q_\alpha \bar{F}(x)) = \bar{F}^{-1}\left(q_\alpha e^{-\theta x}\right) = x - \frac{1}{\theta}\log q_\alpha.$$

Conversely, suppose Eq. (2.3) holds. By choosing $x = 0$, we get $\bar{F}^{-1}(q_\alpha) = c_\alpha$, or, equivalently, $q_\alpha = \bar{F}(c_\alpha)$. By applying $\bar{F}$ to both sides of (2.3), it readily follows

$$\bar{F}(x) = \frac{\bar{F}(x + c_\alpha)}{q_\alpha},$$

that is,

$$\Pr(X > x) = \frac{\Pr(X > x + c_\alpha)}{\Pr(X > c_\alpha)} = \Pr(X > x + c_\alpha \mid X > c_\alpha),$$

for all $x > 0$ and all $c_\alpha > 0$ satisfying $c_\alpha = \bar{F}^{-1}(q_\alpha)$ for $\alpha \in (0, 1)$. As $\bar{F}$ is continuous, c_α can be any value in the support $(0, \infty)$. Hence, X satisfies the memoryless property which characterizes the exponential distribution. □

In practice, the common reliability function $\bar{F}$ is unknown. In some cases, it can be estimated from historical and complete data sets by using non-parametric estimators (e.g. we could use empirical or kernel estimators). In those cases, we just replace in the preceding expressions the exact unknown reliability function $\bar{F}$ with its estimation.

In other cases, we have a model for it. Say $\bar{F}$ is $\bar{F}_\theta$ with an unknown parameter θ. In those cases, we can use $X_{1:n} = x_1, \ldots, X_{r:n} = x_r$ to estimate θ. The associated likelihood function is

$$\ell(\theta) = \frac{n!}{(n-r)!}\bar{F}_\theta^{n-r}(x_r)\prod_{i=1}^{r} f_\theta(x_i)$$

(see, e.g., Arnold et al. 2008 or (5) in Bdair and Raqab 2022). By maximizing this function we get a good estimator for θ.

For example, we can assume the useful Proportional Hazard Rate (PHR) Cox model with $\bar{F}_\theta = \bar{F}_0^\theta$, where $\bar{F}_0$ is a known baseline reliability function and $\theta > 0$ is an unknown risk parameter. In this case,

$$\ell(\theta) = \frac{n!}{(n-r)!}\theta^r \bar{F}_0^{(n-r)\theta}(x_r)\prod_{i=1}^{r}\bar{F}_0^{\theta-1}(x_i)\prod_{i=1}^{r} f_0(x_i)$$

and so $L(\theta) = \log \ell(\theta)$ can be written as

$$L(\theta) = K + r\log\theta + (n-r)\theta\log\bar{F}_0(x_r) + (\theta - 1)\sum_{i=1}^{r}\log\bar{F}_0(x_i),$$

where K is a term that does not depend on θ. Hence, its derivative is

$$L'(\theta) = \frac{r}{\theta} + (n-r)\log \bar{F}_0(x_r) + \sum_{i=1}^{r} \log \bar{F}_0(x_i)$$

and the maximum likelihood estimator (MLE) for θ is

$$\widehat{\theta} = \frac{r}{-(n-r+1)\log \bar{F}_0(x_r) - \sum_{i=1}^{r-1} \log \bar{F}_0(x_i)}. \tag{2.4}$$

Thus, to get the point predictions and the prediction bands for $X_{s:n}$, we just replace in the preceding expressions $\bar{F}_\theta$ with $\bar{F}_{\widehat{\theta}}$. In the simulation study developed in Sect. 5, we will see the real coverage probabilities for the prediction bands obtained in this way.

Some well known distributions are included in the wide PHR model. For example, if $\bar{F}_0(t) = e^{-t}$ for $t \geq 0$ (exponential model), then (2.4) leads to

$$\widehat{\theta} = \frac{r}{(n-r+1)x_r + \sum_{i=1}^{r-1} x_i}. \tag{2.5}$$

Analogously, the mean $\mu = 1/\theta$ can be estimated with

$$\widehat{\mu} = \frac{1}{\widehat{\theta}} = \frac{n-r+1}{r} x_r + \frac{1}{r} \sum_{i=1}^{r-1} x_i \tag{2.6}$$

(a well known result, see e.g. Balakrishnan et al. 2007 or Cramer 2021). It can be written as $\widehat{\mu} = S_r/r$, where

$$S_r = (n-r)x_r + x_1 + \cdots + x_r$$

is known as the total time on test (TTT), see e.g. Khaminsky and Rhodin (1985) or David and Nagaraja (2003, p. 209).

The estimator obtained in that reference for μ by using the maximum (mode) of the joint likelihood function of μ and $X_{s:n}$ at $X_{r:n}$ is $\tilde{\mu} = S_r/(r+1)$. Thus the prediction obtained in Khaminsky and Rhodin (1985) for $X_{s:n}$ is

$$\tilde{m}(x_1, \ldots, x_r) = x_r + \frac{S_r}{r+1} \log\left(\frac{n-r}{n-s+1}\right).$$

Note that if $s = r+1$, then $\tilde{m}(x_1, \ldots, x_r) = x_r$ (i.e. the maximum is obtained in the boundary). Note that this prediction does not belong to the centered prediction intervals.

From (2.6) and Remark 2.2, we get the following (estimated) median regression curve to predict $X_{s:n}$ in the exponential distribution with unknown mean

$$\widehat{m}(x_1,\ldots,x_r) = x_r - \frac{S_r}{r}\log(q_{0.5}) \approx x_r + \frac{S_r}{r}\log\left(\frac{n-r+1/3}{n-s+2/3}\right)$$

for $r+1 \leq s \leq n$ in the first expression and $r+1 < s < n$ in the second. Similar curves can be obtained from Remark 2.2 by using the mean

$$\widehat{m}_{mean}(x_1,\ldots,x_r) = x_r + \frac{S_r}{r}\log\left(\frac{n-r+1}{n-s+1}\right) \tag{2.7}$$

for $r < s \leq n$ or the mode

$$\widehat{m}_{mode}(x_1,\ldots,x_r) = x_r + \frac{S_r}{r}\log\left(\frac{n-r-1}{n-s}\right) \tag{2.8}$$

for $r < s < n$. Another classical point predictor for $X_{s:n}$ in the one parameter exponential family is the Best Linear Unbiased (BLU) predictor given by

$$\widehat{m}_{BLU}(x_1,\ldots,x_r) = x_r + \frac{S_r}{r}\sum_{i=n-s+1}^{n-r}\frac{1}{i},$$

see, e.g., David and Nagaraja (2003, p. 209). These predictors are compared in Example 1.

Analogously, $\bar{F}_0(t) = 1/(1+t)$ for $t \geq 0$ (Pareto type II model), leads to

$$\widehat{\theta} = \frac{r}{(n-r+1)\log(1+x_r) + \sum_{i=1}^{r-1}\log(1+x_i)}. \tag{2.9}$$

We can use this estimator for θ to get the different regression curves as in the exponential case (see the examples in Sect. 4).

Finally, we note that the prediction regions obtained from different quantiles can be used to get bivariate box plots and goodness-of-fit tests for the assumed reliability function $\bar{F}$ (or $\bar{F}_\theta$), see Navarro (2020). For example, to get equal expected values as recommended in Greenwood and Nikulin (1996), we could consider the regions:

$$\begin{aligned}
R_1 &= \left[x, \bar{F}^{-1}\left(q_{0.75}\bar{F}(x)\right)\right],\\
R_2 &= \left[\bar{F}^{-1}\left(q_{0.75}\bar{F}(x)\right), \bar{F}^{-1}\left(q_{0.50}\bar{F}(x)\right)\right],\\
R_3 &= \left[\bar{F}^{-1}\left(q_{0.50}\bar{F}(x)\right), \bar{F}^{-1}\left(q_{0.25}\bar{F}(x)\right)\right],
\end{aligned}$$

and

$$R_4 = \left[\bar{F}^{-1}\left(q_{0.25}\bar{F}(x)\right), \infty\right].$$

Note that if $\bar{F}$ is correctly specified, then $\Pr(X_{s:n} \in R_i \mid X_{r:n} = x) = 1/4$ for $i = 1, 2, 3, 4$.

If we have many values for $X_{r:n}$ and $X_{s:n}$ we could consider these regions for these fixed values of r, s and n. If we just have some values from a sample of size n, we could consider the regions for different values of r and s (see Sect. 4). Resampling methods could be used as well. In all these cases we use the Pearson statistic

$$T = \sum_{i=1}^{4} \frac{(O_i - E_i)^2}{E_i},$$

where O_i and E_i are the observed and expected data in each region, respectively, with $E_i = N/4$ if we have N observations and we assume that $\bar{F}$ is correct (null hypothesis). Under this assumption, the asymptotic distribution of T as $N \to \infty$ is a chi-squared distribution with 3 degrees of freedom (2 if we use the MLE of the parameter θ). Under this null hypothesis, the associated P-value will be $\Pr(\chi^2_3 > T)$. Some illustrative examples will be provided later.

3 Dependent data

In this case we assume that we have a sample $X_1, \ldots, X_n$ of ID random variables with a common distribution function F but now we also assume that the data might have some kind of dependency. In many cases, this dependency is due to the fact that they share the same environment (for example, when they are components of the same system). The dependency will be modeled with a copula function C that is used to write their joint distribution as

$$\mathbf{F}(x_1, \ldots, x_n) = \Pr(X_1 \le x_1, \ldots, X_n \le x_n) = C(F(x_1), \ldots, F(x_n))$$

for all $x_1, \ldots, x_n$.

We assume that $\mathbf{F}$ is absolutely continuous and we again consider the ordered data $X_{1:n} < \cdots < X_{n:n}$ obtained from $X_1, \ldots, X_n$. However, in practice, we just have the first r values $X_{1:n} < \cdots < X_{r:n}$ and we want to predict $X_{s:n}$ for $s > r$.

Under dependency, the order statistics do not satisfy the Markov property, that is, the distributions of $(X_{s:n} \mid X_{1:n} = x_1, \ldots, X_{r:n} = x_r)$ and $(X_{s:n} \mid X_{r:n} = x_r)$ do not coincide. To get bivariate plots, we shall use the second one, that is, we will predict $X_{s:n}$ from $X_{r:n} = x$ for $r < s$. The other data can just be used to estimate the unknown parameters in the model. In this case, we might have unknown parameters both in F and in C. Also note that, we can obtain and compare different predictions for $X_{s:n}$ by using $X_{1:n}, \ldots, X_{r:n}$.

To get these predictions, we will use a distortion representation for the joint distribution of the random vector $(X_{r:n}, X_{s:n})$ as proposed in Navarro et al. (2022). Actually, the results for the case $n = 2, r = 1$ and $s = 2$ (paired ordered data) were already obtained there. Similar results were obtained in Navarro (2022b) for record values. The procedure is similar but, the expressions obtained here for $n = 4$ will be more complicated. They are based on the following two facts.

The first one is that there exists a distortion function D (which depends on r, s, n and C), such that the joint distribution $G_{r,s:n}$ of $(X_{r:n}, X_{s:n})$ can be written as

$$G_{r,s:n}(x, y) = \Pr(X_{r:n} \le x, X_{s:n} \le y) = D_{r,s:n}(F(x), F(y))$$

for all x, y. The distortion function $D_{r,s:n}$ is a continuous bivariate distribution function with support included in the set $[0, 1]^2$. Note that this representation is similar to the classical copula representation but that here $D_{r,s:n}$ is not a copula and that F does not coincide with the marginal distributions (i.e. the distributions of $X_{r:n}$ and $X_{s:n}$).

The second fact is that, from the results obtained in Navarro et al. (2022), we can obtain the median regression curve and the associated prediction bands to predict $X_{s:n}$ from $X_{r:n}$. The result can be stated as follows. It is Proposition 7 in Navarro et al. (2022). Throughout the paper we use the notation $\partial_i G$ for the partial derivative of function G with respect to its ith variable. Similarly, $\partial_{i,j} G$ represents the partial derivatives of G with respect to its ith and jth variables, and so on.

Proposition 3.1 *If we assume that both F and $D_{r,s:n}$ are absolutely continuous, then the conditional distribution of $X_{s:n}$ given $X_{r:n} = x$ is*

$$G_{s|r:n}(y \mid x) = \frac{\partial_1 D_{r,s:n}(F(x), F(y))}{\partial_1 D_{r,s:n}(F(x), 1)} \tag{3.1}$$

for $x < y$ such that $\partial_1 D_{r,s:n}(F(x), v) > 0$ and $\lim_{v \to 0^+} \partial_1 D_{r,s:n}(F(x), v) = 0$.

Sometimes, it is better to use the reliability functions instead of the distribution functions (see examples). We have similar results for them. For example, the joint reliability function of $X_1, \ldots, X_n$ can be written as

$$\bar{\mathbf{F}}(x_1, \ldots, x_n) = \Pr(X_1 > x_1, \ldots, X_n > x_n) = \bar{C}(\bar{F}(x_1), \ldots, \bar{F}(x_n))$$

for all $x_1, \ldots, x_n$, where $\bar{F} = 1 - F$ and $\bar{C}$ is another copula called *survival copula*. $\bar{C}$ can be obtained from C (and vice versa).

Analogously, the joint reliability function of $\bar{G}_{r,s:n}$ of $(X_{r:n}, X_{s:n})$ can be written as

$$\bar{G}_{r,s:n}(x, y) = \Pr(X_{r:n} > x, X_{s:n} > y) = \bar{D}_{r,s:n}(\bar{F}(x), \bar{F}(y)) \tag{3.2}$$

for all x, y. The distortion function $\bar{D}_{r,s:n}$ is also a continuous bivariate distribution function with support included in the set $[0, 1]^2$. It depends on r, s, n and C (or $\bar{C}$). From this expression, the conditional reliability function can be obtained as follows.

Proposition 3.2 *If we assume that both $\bar{F}$ and $\bar{D}_{r,s:n}$ are absolutely continuous, then the conditional reliability function of $X_{s:n}$ given $X_{r:n} = x$ is*

$$\bar{G}_{s|r:n}(y \mid x) = \frac{\partial_1 \bar{D}_{r,s:n}(\bar{F}(x), \bar{F}(y))}{\partial_1 \bar{D}_{r,s:n}(\bar{F}(x), 1)} \tag{3.3}$$

for $x < y$ such that $\partial_1 \bar{D}_{r,s:n}(\bar{F}(x), v) > 0$ and $\lim_{v \to 0^+} \partial_1 \bar{D}_{r,s:n}(\bar{F}(x), v) = 0$.

The preceding expressions can be used to solve the general case in which we want to predict $X_{s:n}$ from $X_{r:n}$ for $1 \leq r < s \leq n$. To show the procedure, we choose different cases for $n = 4$. In all these cases we assume that the joint distribution of (X_1, X_2, X_3, X_4) is exchangeable (EXC), that is, it does not change when we permute them. This is equivalent to the assumption that they are ID and C (or $\bar{C}$) is exchangeable.

In the first case, we choose $r = 1$ and $s = 2$. The result can be stated as follows.

Proposition 3.3 *If both $\bar{F}$ and $\bar{C}$ are absolutely continuous and $\bar{C}$ is EXC, then the conditional reliability function of $X_{2:4}$ given $X_{1:4} = x$ is*

$$\bar{G}_{2|1:4}(y \mid x) = \frac{\partial_1 \bar{C}(\bar{F}(x), \bar{F}(y), \bar{F}(y), \bar{F}(y))}{\partial_1 \bar{C}(\bar{F}(x), \bar{F}(x), \bar{F}(x), \bar{F}(x))} \tag{3.4}$$

for all $x < y$ such that $\partial_1 \bar{C}(\bar{F}(x), \bar{F}(x), \bar{F}(x), \bar{F}(x)) > 0$ and $\lim_{v \to 0^+} \partial_1 \bar{C}(\bar{F}(x), v, v, v) = 0$.

Proof The joint reliability function $\bar{G}_{1,2:4}$ of $(X_{1:4}, X_{2:4})$ satisfies

$$\bar{G}_{1,2:4}(x, y) = \Pr(X_{1:4} > x, X_{2:4} > y) = \Pr(X_{1:4} > x)$$

for all $x \geq y$, where

$$\begin{aligned} \Pr(X_{1:4} > x) &= \Pr(X_1 > x, X_2 > x, X_3 > x, X_4 > x) \\ &= \bar{C}(\bar{F}(x), \bar{F}(x), \bar{F}(x), \bar{F}(x)). \end{aligned}$$

Analogously, for $x < y$, we get

$$\begin{aligned} \bar{G}_{1,2:4}(x, y) &= \Pr(X_{1:4} > x, X_{2:4} > y) \\ &= \Pr(X_{1:4} > x, \max_{i=1,\ldots,r} X_{P_i} > y) \\ &= \Pr\left(\cup_{i=1}^{r} (\{X_{P_i} > y\} \cap \{X_{1:4} > x\})\right), \end{aligned}$$

where $X_{P_i} = \min_{j \in P_i} X_j$ and $P_1, \ldots, P_r$ are all the minimal path sets of $X_{2:4}$ (see, e.g., Navarro 2022a, p. 23). In this case they are all the subsets of $\{1, 2, 3, 4\}$ with cardinality 3. So $r = \binom{4}{3} = 4$. Hence, by applying the inclusion-exclusion formula and by using the exchangeable assumption we get

$$\begin{aligned} \bar{G}_{1,2:4}(x, y) &= \sum_{i=1}^{4} \Pr\left(X_{P_i} > y, X_{1:4} > x\right) - \sum_{i<j} \Pr\left(X_{P_i \cup P_j} > y, X_{1:4} > x\right) \\ &\quad + \sum_{i<j<k} \Pr\left(X_{P_i \cup P_j \cup P_k} > y, X_{1:4} > x\right) \\ &\quad - \Pr\left(X_{P_1 \cup P_2 \cup P_3 \cup P_4} > y, X_{1:4} > x\right) \\ &= 4 \Pr\left(X_1 > x, X_2 > y, X_3 > y, X_4 > y\right) \end{aligned}$$

$$
\begin{aligned}
&- 3\Pr(X_1 > y, X_2 > y, X_3 > y, X_4 > y) \\
&= 4\bar{C}\left(\bar{F}(x), \bar{F}(y), \bar{F}(y), \bar{F}(y)\right) - 3\bar{C}\left(\bar{F}(y), \bar{F}(y), \bar{F}(y), \bar{F}(y)\right)
\end{aligned}
$$

for $x < y$. Therefore, (3.2) holds for

$$
\bar{D}_{1,2:4}(u, v) = \begin{cases} \bar{C}(u, u, u, u) & \text{for } 0 \le u \le v \le 1; \\ 4\bar{C}(u, v, v, v) - 3\bar{C}(v, v, v, v) & \text{for } 0 \le v < u \le 1. \end{cases}
$$

Hence

$$
\partial_1 \bar{D}_{1,2:4}(u, v) = \begin{cases} 4\partial_1 \bar{C}(u, u, u, u) & \text{for } 0 \le u \le v \le 1; \\ 4\partial_1 \bar{C}(u, v, v, v) & \text{for } 0 \le v < u \le 1. \end{cases}
$$

Finally, we use (3.3) to get (3.4). □

In Example 4 we show how to apply this result to a specific case. In the following propositions we provide the expressions for the other cases. As the proofs are similar, they are omitted. Note that in Proposition 3.7 we use (3.1) instead of (3.3).

Proposition 3.4 *If both $\bar{F}$ and $\bar{C}$ are absolutely continuous and $\bar{C}$ is EXC, then the conditional reliability function of $X_{3:4}$ given $X_{1:4} = x$ is*

$$
\bar{G}_{3|1:4}(y \mid x) = \frac{3\partial_1 \bar{C}(\bar{F}(x), \bar{F}(x), \bar{F}(y), \bar{F}(y)) - 2\partial_1 \bar{C}(\bar{F}(x), \bar{F}(y), \bar{F}(y), \bar{F}(y))}{\partial_1 \bar{C}(\bar{F}(x), \bar{F}(x), \bar{F}(x), \bar{F}(x))} \tag{3.5}
$$

for all $x < y$ such that $\partial_1 \bar{C}(\bar{F}(x), \bar{F}(x), \bar{F}(x), \bar{F}(x)) > 0$ and $\lim_{v \to 0^+} 3\partial_1 \bar{C}(\bar{F}(x), \bar{F}(x), v, v) - 2\partial_1 \bar{C}(\bar{F}(x), v, v, v) = 0$.

Proposition 3.5 *If both $\bar{F}$ and $\bar{C}$ are absolutely continuous and $\bar{C}$ is EXC, then the conditional reliability function of $X_{4:4}$ given $X_{1:4} = x$ is*

$$
\bar{G}_{4|1:4}(y \mid x) = \frac{A_{4|1:4}(x, y)}{\partial_1 \bar{C}(\bar{F}(x), \bar{F}(x), \bar{F}(x), \bar{F}(x))}, \tag{3.6}
$$

for all $x < y$ such that $\partial_1 \bar{C}(\bar{F}(x), \bar{F}(x), \bar{F}(x), \bar{F}(x)) > 0$ and $\lim_{y \to \infty} A_{4|1:4}(x, y) = 0$, where

$$
\begin{aligned}
A_{4|1:4}(x, y) &= 3\partial_1 \bar{C}(\bar{F}(x), \bar{F}(x), \bar{F}(x), \bar{F}(y)) - 3\partial_1 \bar{C}(\bar{F}(x), \bar{F}(x), \bar{F}(y), \bar{F}(y)) \\
&\quad + \partial_1 \bar{C}(\bar{F}(x), \bar{F}(y), \bar{F}(y), \bar{F}(y)).
\end{aligned}
$$

Proposition 3.6 *If both $\bar{F}$ and $\bar{C}$ are absolutely continuous and $\bar{C}$ is EXC, then the conditional reliability function of $X_{4:4}$ given $X_{2:4} = x$ is*

$$
\bar{G}_{4|2:4}(y \mid x) = \frac{A_{4|2:4}(x, y)}{\partial_1 \bar{C}(\bar{F}(x), \bar{F}(x), \bar{F}(x), 1) - \partial_1 \bar{C}(\bar{F}(x), \bar{F}(x), \bar{F}(x), \bar{F}(x))} \tag{3.7}
$$

for all $x < y$ *such that* $\partial_1 \bar{C}(\bar{F}(x), \bar{F}(x), \bar{F}(x), 1) - \partial_1 \bar{C}(\bar{F}(x), \bar{F}(x), \bar{F}(x), \bar{F}(x)) > 0$ *and* $\lim_{y\to\infty} A_{4|2:4}(x, y) = 0$*, where*

$$A_{4|2:4}(x, y) = 2\partial_1 \bar{C}(\bar{F}(x), \bar{F}(x), \bar{F}(y), 1) - 2\partial_1 \bar{C}(\bar{F}(x), \bar{F}(x), \bar{F}(x), \bar{F}(y)) - \partial_1 \bar{C}(\bar{F}(x), \bar{F}(y), \bar{F}(y), 1) + \partial_1 \bar{C}(\bar{F}(x), \bar{F}(x), \bar{F}(y), \bar{F}(y)).$$

Proposition 3.7 *If both* F *and* C *are absolutely continuous and* C *is EXC, then the conditional distribution function of* $X_{4:4}$ *given* $X_{3:4} = x$ *is*

$$G_{4|3:4}(y \mid x) = \frac{\partial_1 C(F(x), F(x), F(x), F(y)) - \partial_1 C(F(x), F(x), F(x), F(x))}{\partial_1 C(F(x), F(x), F(x), 1) - \partial_1 C(F(x), F(x), F(x), F(x))},$$

for all $x < y$ *such that* $\partial_1 C(F(x), F(x), F(x), 1) - \partial_1 C(F(x), F(x), F(x), F(x)) > 0$.

We conclude this section by showing how to get predictions from more than one early failures. For example, if we want to predict $X_{3:4}$ from $X_{1:4} = x$ and $X_{2:4} = y$, we need a distortion representation for their joint reliability function as

$$\bar{G}_{1,2,3}(x, y, t) = \Pr(X_{1:4} > x, X_{2:4} > y, X_{3:4} > t) = \bar{D}(\bar{F}(x), \bar{F}(y), \bar{F}(t)),$$

where $\bar{F}$ is the common reliability function of X_1, X_2, X_3, X_4. Then their joint PDF is

$$g_{1,2,3}(x, y, t) = f(x)f(y)f(t)\ \partial_{1,2,3}\bar{D}(\bar{F}(x), \bar{F}(y), \bar{F}(t)),$$

where $f = -\bar{F}'$. Analogously, the joint reliability function of $X_{1:4}$ and $X_{2:4}$ is

$$\bar{G}_{1,2}(x, y) = \Pr(X_{1:4} > x, X_{2:4} > y) = \bar{D}(\bar{F}(x), \bar{F}(y), 1)$$

and their joint PDF

$$g_{1,2}(x, y) = f(x)f(y)\ \partial_{1,2}\bar{D}(\bar{F}(x), \bar{F}(y), 1).$$

Hence, the PDF of $(X_{3:4} \mid X_{1:4} = x, X_{2:4} = y)$ is

$$g_{3|1,2}(t \mid x, y) = \frac{g_{1,2,3}(x, y, t)}{g_{1,2}(x, y)} = f(t)\frac{\partial_{1,2,3}\bar{D}(\bar{F}(x), \bar{F}(y), \bar{F}(t))}{\partial_{1,2}\bar{D}(\bar{F}(x), \bar{F}(y), 1)}$$

and the reliability function is

$$\bar{G}_{3|1,2}(t \mid x, y) = \frac{\partial_{1,2}\bar{D}(\bar{F}(x), \bar{F}(y), \bar{F}(t))}{\partial_{1,2}\bar{D}(\bar{F}(x), \bar{F}(y), 1)}$$

for $x < y < t$, whenever $\partial_{1,2}\bar{D}(\bar{F}(x), \bar{F}(y), 1) > 0$ and $\lim_{v\to 0^+} \partial_{1,2}\bar{D}(\bar{F}(x), \bar{F}(y), v) = 0$.

A straightforward calculation shows that if the survival copula $\bar{C}$ is EXC, then

$$\begin{aligned}\bar{D}(u,v,w) = 12\bar{C}(u,v,w,w) - 12\bar{C}(u,w,w,w)\\ -6\bar{C}(v,v,w,w) + 7\bar{C}(w,w,w,w)\end{aligned}$$

and

$$\partial_{1,2}\bar{D}(u,v,w) = 12\bar{C}(u,v,w,w)$$

for $1 > u > v > w > 0$. Analogously,

$$\bar{D}(u,v,1) = 4\bar{C}(u,v,v,v) - 3\bar{C}(v,v,v,v)$$

and

$$\partial_{1,2}\bar{D}(u,v,1) = 12\bar{C}(u,v,v,v)$$

for $1 > u > v > 0$. Hence

$$\bar{G}_{3|1,2}(t \mid x,y) = \frac{\partial_{1,2}\bar{C}(\bar{F}(x),\bar{F}(y),\bar{F}(t),\bar{F}(t))}{\partial_{1,2}\bar{C}(\bar{F}(x),\bar{F}(y),\bar{F}(y),\bar{F}(y))} \tag{3.8}$$

for $x < y < t$ such that $\partial_{1,2}\bar{C}(\bar{F}(x),\bar{F}(y),\bar{F}(y),\bar{F}(y)) > 0$ and

$$\lim_{t\to\infty} \partial_{1,2}\bar{C}(\bar{F}(x),\bar{F}(y),\bar{F}(t),\bar{F}(t)) = 0.$$

The prediction is obtained by solving (numerically) $\bar{G}_{3|1,2}(t \mid x,y) = 0.5$ for given values of x and y. The prediction intervals are obtained in a similar way.

Proceeding as above we can also obtain the expression to predict $X_{4:4}$ from $X_{1:4} = x$ and $X_{2:4} = y$ as

$$\bar{G}_{4|1,2}(t \mid x,y) = \frac{2\partial_{1,2}\bar{C}(\bar{F}(x),\bar{F}(y),\bar{F}(y),\bar{F}(t)) - \partial_{1,2}\bar{C}(\bar{F}(x),\bar{F}(y),\bar{F}(t),\bar{F}(t))}{\partial_{1,2}\bar{C}(\bar{F}(x),\bar{F}(y),\bar{F}(y),\bar{F}(y))} \tag{3.9}$$

for $x < y < t$ such that $\partial_{1,2}\bar{C}(\bar{F}(x),\bar{F}(y),\bar{F}(y),\bar{F}(y)) > 0$ and

$$\lim_{t\to\infty} 2\partial_{1,2}\bar{C}(\bar{F}(x),\bar{F}(y),\bar{F}(y),\bar{F}(t)) - \partial_{1,2}\bar{C}(\bar{F}(x),\bar{F}(y),\bar{F}(t),\bar{F}(t)) = 0.$$

4 Examples

First we illustrate the IID case with simulated samples. In the first one we assume an exponential baseline distribution.

Example 1 We simulate a sample of size $n = 20$ from a standard exponential distribution. The ordered (rounded) sample values obtained are

0.00599 0.02454 0.04600 0.07663 0.08168 0.14609 0.24391 0.72400
1.30312 1.37244 1.37962 1.54357 1.71278 2.22949 2.24561 2.56783
2.61441 2.80786 3.90280 7.68743

If we want to predict $X_{2:20}$ from $X_{1:20}$ by assuming that $\bar{F}$ is known (or that it is estimated from a preceding sample), we use the quantile regression curve given in (2.1)

$$m(x) = \bar{F}^{-1}(q_{0.5}\bar{F}(x)) = x - \log(q_{0.5}) = x + 0.03648$$

where $q_{0.5} = 0.96418$ is the median of a beta distribution with parameters $n - s + 1 = 19$ and $s - r = 1$. Thus, we get the prediction for $X_{2:20}$ as

$$\widehat{X}_{2:20} = m(X_{1:20}) = m(0.00599) = 0.00599 + 0.03648 = 0.04247.$$

The real value is $X_{2:20} = 0.02454$. The 90% and 50% prediction intervals for this prediction are obtained from (2.2) as $C_{90} = [0.00869, 0.16366]$ and $C_{50} = [0.02113, 0.07895]$. The real value belongs to both intervals.

To see better what happens with these predictions we simulate $N = 100$ predictions of this kind, that is, 100 samples of size 20. The data are plotted in Fig. 1, left. There we can see that the prediction bands represent very well the dispersion of the majority of data (except some extreme values). In this sample, C_{50} contains 51 values and C_{90} contains 90 while 5 values are above the upper limit and 5 are below the bottom limit. Of course, if we test $H_0 : \bar{F}$ is correct vs $H_0 : \bar{F}$ is not correct, by using the four regions R_1, R_2, R_3, R_4 considered in Sect. 2, we get the observed values: 25, 30, 21, 24 and the T statistic value is

$$T = \frac{(25-25)^2}{25} + \frac{(30-25)^2}{25} + \frac{(21-25)^2}{25} + \frac{(24-25)^2}{25} = 1.68.$$

Thus, the P-value $P = \Pr(\chi^2_3 > 1.68) = 0.64139$ leads to the acceptance of the exponential distribution (as expected). Note that, in practice, it is not easy to perform this test because we need the first two values of several samples with the same size ($n = 20$ in this example). In this case, we could also use the standard quantile regression techniques (see Koenker 2005).

We do the same in Fig. 1, right, but for $n = 20, r = 12$ and $s = 13$. If we just use the initial sample, the prediction for $X_{13:20} = 1.71278$ from $X_{12:20} = 1.54357$ obtained with the median curve

$$m(x) = \bar{F}^{-1}(q_{0.5}\bar{F}(x)) = x - \log(q_{0.5}) = x + 0.08664$$

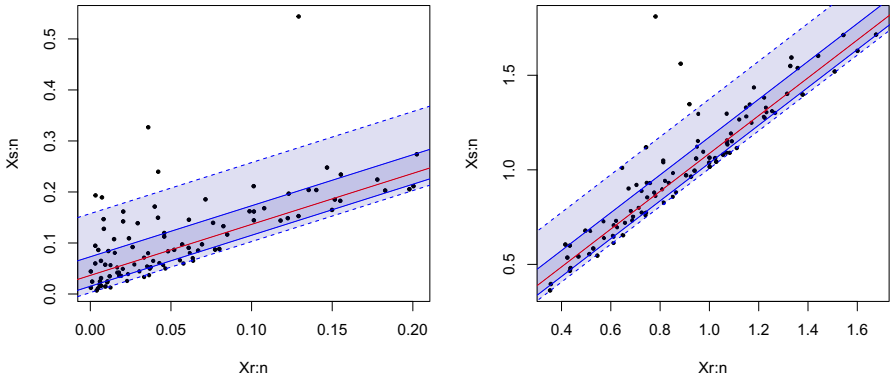

Fig. 1 Scatterplots of a simulated sample from $(X_{r:n}, X_{s:n})$ for $n = 20, r = 1$ and $s = 2$ (left) and $r = 12$ and $s = 13$ (right) for the exponential distribution in Example 1 jointly with the theoretical median regression curves (red) and 50% (dark grey) and 90% (light grey) prediction bands (colour figure online)

is $m(1.54357) = 1.63021$, where $q_{0.5} = 0.91700$ is the median of a beta distribution with parameters $n - s + 1 = 8$ and $s - r = 1$. The prediction intervals for this prediction are $C_{90} = [1.549986, 1.918041]$ and $C_{50} = [1.579534, 1.716861]$. Both intervals contain the associated order statistic. The 100 repetitions of this case plotted in Fig. 1, right, show the underlying uncertainty for that predictions. We remind that for the exponential distribution all the curves are lines with slope one (see Proposition 2.3). In this case, we have 92 values in C_{90}, 56 in C_{50} and 8 values out of C_{90} (4 above and 4 below). The T statistic is 2.24 and its associated P-value 0.52411 leads again to accept the (true) distribution F.

The predictions will be worse for more distant future values (i.e., the dispersion will be greater). To show this fact, in Fig. 2 we plot the prediction bands for $r = 12$, $s = 14$ (left) and $s = 20$ (right). However, of course, the coverage probabilities of these regions will be the same. The predictions obtained in the initial sample are $\widehat{X}_{14:20} = 1.76813$ and $\widehat{X}_{20:20} = 4.03254$, with prediction intervals $C_{90} = [1.59107, 2.17974]$ and $C_{90} = [2.70722, 6.59642]$, respectively. The real values are $X_{14:20} = 2.22949$ and $X_{20:20} = 7.68743$. Both values are out of the interval C_{90}. This situation is unexpected because, in this simulation, we get two of two failures in the prediction intervals. However, note that these events are not independent and that this is not the case in the other simulations.

In Fig. 3 we plot the predictions for $X_{s:20}$ (red line) jointly with the limits of the 90% (dashed blue lines) and 50% (continuous blue line) prediction intervals in the initial simulated sample from $X_{12:20}$ (left) for $s = 13, \ldots, 20$ and from the preceding data $X_{s-1:20}$ (right) for $s = 2, \ldots, 20$. In the left plot 2-out-of-8 exact points do not belong to the 90% prediction intervals while 5 are out of the 50% prediction intervals (the expected values are $8 \cdot 0.1 = 0.8$ and $8 \cdot 0.5 = 4$, respectively). In the right plot 4-out-of-19 points do not belong to 90% prediction intervals while 11 do not belong to the 50% prediction intervals (we expect 1.9 and 9.5, respectively).

Now, let us compare our method with the one based on the maximum likelihood predictions (MLP) proposed in Khaminsky and Rhodin (1985). If we know the exact

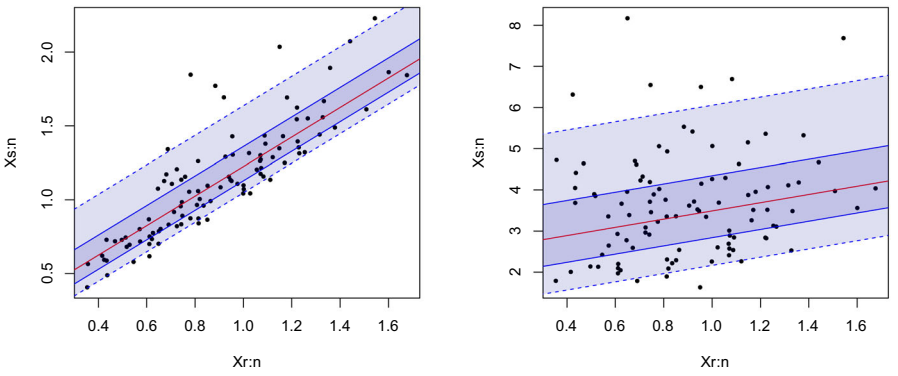

Fig. 2 Scatterplots of a simulated sample from $(X_{r:n}, X_{s:n})$ for $n = 20, r = 12$ and $s = 14$ (left) and $s = 20$ (right) for the exponential distribution in Example 1 jointly with the theoretical median regression curves (red) and 50% (dark grey) and 90% (light grey) prediction bands (colour figure online)

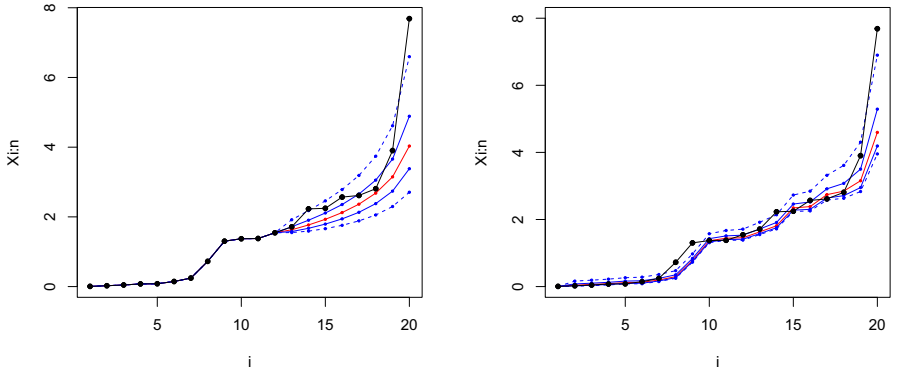

Fig. 3 Predictions (red) for $X_{s:n}$ from $X_{r:n}$ for $n = 20$, $r = 12$ and $s = 13, \ldots, 20$ (left) and $r = 1, \ldots, 19$ and $s = r + 1$ (right) for the exponential distribution in Example 1. The black points are the observed values and the blue lines are the limits for the 50% (continuous lines) and the 90% (dashed lines) prediction intervals (colour figure online)

distribution F, with pdf f, of the sample and the ordered values up to the index r, the MLP method suggests to predict the value of $X_{s:n} = x_s$, $s > r$, by resolving

$$\frac{f'(x_s)}{f^2(x_s)} + \frac{s - r - 1}{F(x_s) - F(x_r)} - \frac{n - s}{1 - F(x_s)} = 0.$$

In particular, for an exponential distribution this method leads to the prediction

$$\widehat{X}_{s:n} = x_r + \mu \log\left(\frac{n - r}{n - s + 1}\right)$$

for $X_{s:n}$ at $X_{r:n} = x_r$ when the mean μ is known. We can readily observe that if $s = r + 1$, the MLP method suggests to predict $X_{r+1:n}$ with the value assumed by the r-th order statistic.

For illustrative purposes, we consider again the cases $r = 12$ and $s = 14$ or $s = 20$. By using the MLP method with $\mu = 1$ the predictions are

$$\begin{aligned}\widehat{X}_{13:20} &= X_{12:20} = 1.54357,\\ \widehat{X}_{14:20} &= X_{12:20} + \log\frac{8}{7} = 1.67710,\\ \widehat{X}_{20:20} &= X_{12:20} + \log 8 = 3.62301,\end{aligned}$$

which are all worse than the ones obtained by using the method proposed here (1.63021, 1.76813 and 4.03254, respectively). The same conclusion holds in the other points, that is, for $s = 15, \ldots, 19$. However, in this case, the curve provided by the mean

$$\widehat{m}_{mean}(x_r) = x_r + \mu \log\left(\frac{n-r+1}{n-s+1}\right)$$

gives better predictions for $s = 13, \ldots, 16$ (and worse for $s = 17, \ldots, 20$).

In practice, we do not know the exact reliability function. If we just assume the exponential model $\bar{F}(t) = e^{-\theta t}$ for $t \geq 0$, with an unknown parameter $\theta > 0$, we can use (2.5) to estimate θ. With the above sample and $r = 12$ we get

$$\widehat{\theta}_{12} = \frac{12}{9X_{12:20} + \sum_{i=1}^{11} X_{i:20}} = 0.62188.$$

The exact value is $\theta = 1$. The estimation for the mean is $\widehat{\mu} = 1/\widehat{\theta}_{12} = 1.608017$, Replacing the exact reliability function $\bar{F}(t) = e^{-t}$ with $\bar{F}_{\widehat{\theta}}(t) = e^{-0.62188 \cdot t}$, we can obtain similar predictions for $X_{s:20}$ from $X_{12:20}$ as that obtained above. For example, for $s = 13$ we get the prediction $\widehat{X}_{13:20} = 1.6829$ for $X_{13:20} = 1.71278$. The estimated prediction intervals for this prediction are $\widehat{C}_{90} = [1.55388, 2.14572]$ and $\widehat{C}_{50} = [1.60140, 1.82222]$. Both intervals contain the exact value. However, in this case, as we estimate the parameter, we do not know the exact coverage probabilities for these intervals (see Sect. 5). The predictions from $X_{12:20}$ for $X_{s:20}$ and $s = 13, \ldots, 20$ are plotted in Fig. 4, left. The predictions for $s = 14$ and $s = 20$ are $\widehat{X}_{14:20} = 1.904668$ and $\widehat{X}_{20:20} = 5.545868$. The blue lines represent the prediction intervals. Note that all the exact values belong to the 90% intervals (dashed blue lines) and that three of them do not belong to the 50% intervals (blue lines).

We can compare these predictions with the ones obtained from the method proposed in Khaminsky and Rhodin (1985) or that obtained from the mean, the mode or the BLU (see Remark 2.2). In the first case we obtain the predictions

$$\begin{aligned}\widehat{X}_{13:20} &= X_{12:20} = 1.54357,\\ \widehat{X}_{14:20} &= X_{12:20} + \frac{9X_{12:20} + \sum_{i=1}^{11} X_{i:20}}{13}\log\frac{8}{7} = 1.74178,\\ \widehat{X}_{20:20} &= X_{12:20} + \frac{9X_{12:20} + \sum_{i=1}^{11} X_{i:20}}{13}\log 8 = 4.63014.\end{aligned}$$

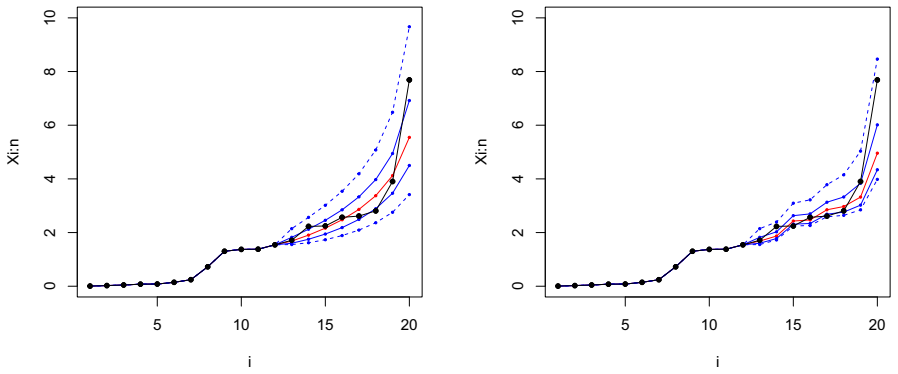

Fig. 4 Predictions (red) for $X_{s:n}$ from $X_{r:n}$ for $n = 20$, $r = 12$ and $s = 13, \ldots, 20$ (left) and $r = 12, \ldots, 19$ and $s = r + 1$ (right) for the exponential distribution in Example 1 with estimated θ at $X_{r:20}$. The black points are the exact values and the blue lines are the limits for the 50% (continuous lines) and the 90% (dashed lines) prediction intervals (colour figure online)

These predictions are worse than that obtained with the median. However, for the other values, now the predictions for $s = 17$ and $s = 18$ are better (and that for $s = 15, 16$ and $s = 19$ are worse). In the second case, the predictions obtained with the mean by using (2.7) are better than that obtained with the median for $s = 13, \ldots, 20$ and better to that obtained with the method proposed in Khaminsky and Rhodin (1985) for $s = 13, 14, 15, 16, 19, 20$. If we use the curve based on the mode given in (2.8) for $s = 13, \ldots, 19$ we obtain worse predictions than those obtained with the preceding cases. Finally, if we use the BLU predictor, we get better results in the cases $s = 14, 15, 16, 20$.

In Fig. 4, right, we provide the predictions for $X_{r+1:20}$ from all the preceding values $X_{1:20}, \ldots, X_{r:20}$ for $r = 12, \ldots, 19$ which are used to estimate θ. The estimations $\widehat{\theta}_r$ obtained for θ at $X_{r:20}$ and the predictions for $X_{s:20}$ are given in Table 1. Note that the estimations for θ are similar. The MLE for θ from the complete sample (which is not available under our assumptions) is $\widehat{\theta}_{20} = 20/(X_1 + \cdots + X_{20}) = 0.6113249$ which is very similar to our estimations for $r \geq 12$ (remember that the exact value is $\theta = 1$). In practice, when we work with real data, the stability of these predictions might confirm the assumed parametric model. Also note that all the exact values belong to the 90% prediction intervals but that 4 of them do not belong to the 50% prediction intervals (as expected). Surprisingly, in this case, the estimations obtained from $X_{12:20}$ seem to be better than that obtained from the preceding values but note that the lengths of the intervals in the first case are greater than the ones obtained in the second.

In the following example we perform a similar study by assuming a Pareto type II baseline distribution in the PHR model.

Example 2 We simulate a sample of size $n = 20$ from a Pareto type II distribution with parameter 2, i.e., with reliability function $\bar{F}(t) = 1/(1 + t)^2$ for $t \geq 0$. The ordered (rounded) sample values obtained are

0.01352 0.04743 0.05927 0.07542 0.16497 0.17561 0.22626 0.25210

Table 1 Predicted values $\widehat{X}_{r+1:n}$ and estimated centered prediction intervals $\widehat{C}_{50} = [l_r, u_r]$ and $\widehat{C}_{90} = [L_n, U_n]$ for $X_{r+1:n}$ from $X_{r:n}$ in a standard exponential distribution; $\widehat{\theta}_r$ is the estimate of θ based on $X_{1:n}, \ldots, X_{r:n}$ for $r = 12, \ldots, 19$

r	$\widehat{\theta}_r$	L_r	l_r	$\widehat{X}_{r+1:n}$	$X_{r+1:n}$	u_r	U_r
12	0.62188	1.55388	1.60140	1.68290	1.71278	1.82222	2.14572
13	0.62954	1.72442	1.77807	1.87007	2.22949	2.02736	2.39258
14	0.57692	2.24431	2.31260	2.42974	2.24561	2.62998	3.09493
15	0.61567	2.26227	2.33906	2.47078	2.56783	2.69595	3.21877
16	0.61599	2.58865	2.68459	2.84915	2.61441	3.13047	3.78366
17	0.64982	2.640723	2.76198	2.96997	2.80786	3.32553	4.15110
18	0.67312	2.845958	3.02155	3.32274	3.90280	3.83762	5.03313
19	0.65673	3.980903	4.34085	4.95825	7.68743	6.01370	8.46438

0.28100 0.32301 0.45214 0.46332 0.59333 0.80563 0.80966 0.86476
1.77393 2.43810 2.96807 6.05006

If we want to predict $X_{2:20}$ from $X_{1:20}$ by assuming that $\bar{F}$ is known, we use the quantile regression curve given in (2.1) to obtain

$$m(x) = \bar{F}^{-1}(q_{0.5}\bar{F}(x)) = \frac{1+x}{\sqrt{q_{0.5}}} - 1 = 1.01841x + 0.01841,$$

where $q_{0.5} = 0.96418$ is the median of a beta distribution with parameters $n - s + 1 = 19$ and $s - r = 1$. Thus, we get the prediction

$$m(X_{1:20}) = m(0.01352) = 0.03218$$

while the real value is $X_{2:20} = 0.04743$. The associated 90% and 50% prediction intervals are obtained from (2.2) as $C_{90} = [0.01490, 0.09666]$ and $C_{50} = [0.02123, 0.05118]$. The real value belongs to both intervals.

To see better what happens with these predictions we simulate $N = 100$ predictions of this kind. The data are plotted in Fig. 5, left. There we can see that the prediction bands represent very well the dispersion of the majority of data points (except some extreme values). In this sample, C_{50} contains 44 values and C_{90} contains 92 while 6 values are above the upper limit and 2 are below the bottom limit. Of course, if we test $H_0 : \bar{F}$ is correct by using the four regions considered in Sect. 2, we get the observed values: 25, 24, 20, 31 and the T statistic value

$$T = \frac{(31-25)^2}{25} + \frac{(20-25)^2}{25} + \frac{(24-25)^2}{25} + \frac{(25-25)^2}{25} = 2.48.$$

Thus, the P-value $P = \Pr(\chi^2_3 > 2.48) = 0.47892$ leads to the acceptance of the Pareto type II distribution (as expected).

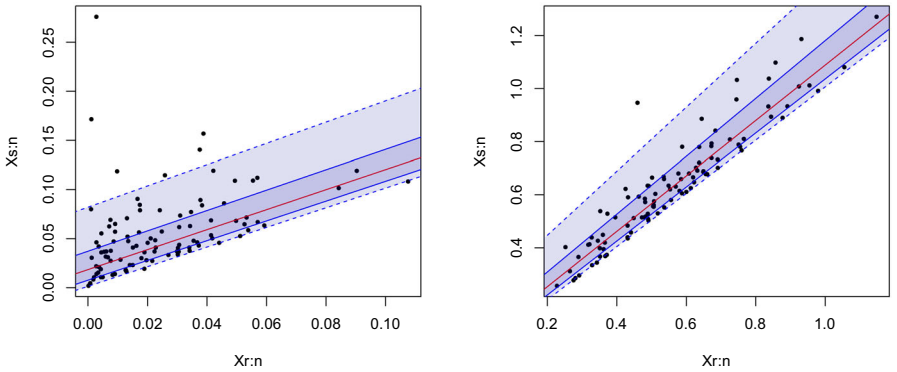

Fig. 5 Scatterplots of a simulated sample from $(X_{r:n}, X_{s:n})$ for $n = 20$, $r = 1$ and $s = 2$ (left) and $r = 12$ and $s = 13$ (right) for the Pareto distribution in Example 2 jointly with the theoretical median regression curves (red) and 50% (dark grey) and 90% (light grey) prediction bands (colour figure online)

We do the same in Fig. 5, right, but for $n = 20, r = 12$ and $s = 13$. Now the prediction for $X_{13:20} = 0.59333$ from $X_{12:20} = 0.46332$ obtained with the median curve is 0.52811, where $q_{0.5} = 0.91700$ is the median of a beta distribution with parameters $n - s + 1 = 8$ and $s - r = 1$. The prediction intervals for this prediction are $C_{90} = [0.46802, 0.76464]$ and $C_{50} = [0.48988, 0.59577]$. Both intervals contain the real value. The 100 repetitions of this case plotted in Fig. 5, right, show the underlying uncertainty for our predictions. In this case, we have 93 values in C_{90}, 55 in C_{50} and 7 values out of C_{90} (1 above and 6 below). The T statistic is 3.6 and its associated P-value 0.30802 leads again to accept the (real) distribution F.

The predictions will be worse for other future values. In Fig. 6 we plot the prediction bands for $r = 12$ $s = 14$ (left) and $s = 20$ (right). However, of course, the coverage probabilities of these regions will be the same. The predictions obtained are $\widehat{X}_{14:20} = 0.63721$ and $\widehat{X}_{20:20} = 4.07940$, with 90% prediction intervals $C_{90} = [0.49850, 1.01132]$ and $C_{90} = [1.61833, 17.30424]$, respectively. The real values are $X_{14:20} = 0.80563$ and $X_{20:20} = 6.05006$. Both values belong to the corresponding interval but the second interval is really wide.

In Fig. 7 we plot the predictions for $X_{s:20}$ (red line) jointly with the limits of the 90% (dashed blue lines) and 50% (continuous blue line) prediction intervals in the initial simulated sample from $X_{12:20}$ for $s = 13, \ldots, 20$ (left) and from the preceding value $X_{s-1:20}$ (right) for $s = 2, \ldots, 20$. In the left plot all the 8 exact points are in the 90% prediction intervals while 3 are out of the 50% prediction intervals (the expected values are 0.8 and 4, respectively). In the right plot 2-out-of-19 points do not belong to the 90% prediction intervals while 7 do not belong to the 50% (we expect 1.9 and 9.5, respectively).

Now, let us consider a more realistic scenario by just assuming the Pareto type II model $\bar{F}(t) = 1/(1+t)^{\theta}$ for $t \geq 0$, with an unknown parameter $\theta > 0$. We can use (2.9) to estimate θ and, with the above sample and $r = 12$, we get

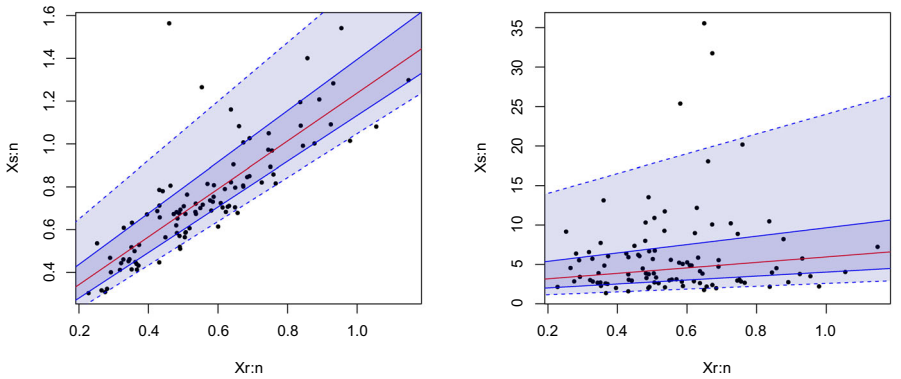

Fig. 6 Scatterplots of a simulated sample from $(X_{r:n}, X_{s:n})$ for $n = 20, r = 12$ and $s = 14$ (left) and $s = 20$ (right) for the Pareto distribution in Example 2 jointly with the theoretical median regression curves (red) and 50% (dark grey) and 90% (light grey) prediction bands (colour figure online)

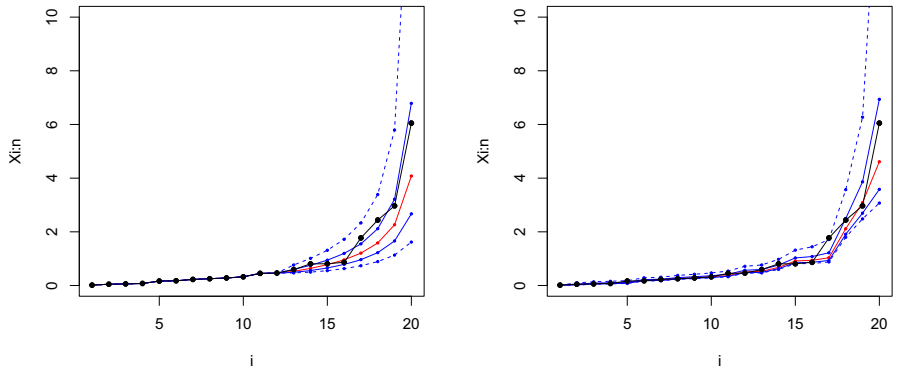

Fig. 7 Predictions (red) for $X_{s:n}$ from $X_{r:n}$ for $n = 20, r = 12$ and $s = 13, \ldots, 20$ (left) and $r = 1, \ldots, 19$ and $s = r + 1$ (right) for the Pareto distribution in Example 2. The black points are the exact values and the blue lines are the limits for the 50% (continuous blue lines) and the 90% (dashed blue lines) prediction intervals (colour figure online)

$$\widehat{\theta}_{12} = \frac{12}{9\log(1 + X_{12:20}) + \sum_{i=1}^{11} \log(1 + X_{i:20})} = 2.28120.$$

The exact value is $\theta = 2$. Replacing the exact survival function $\bar{F}(t) = 1/(1+t)^2$ with $\bar{F}_{\widehat{\theta}}(t) = 1/(1+t)^{2.28120}$ we can obtain similar predictions for $X_{s:20}$ from $X_{12:20}$ as the ones obtained above. For example, for $s = 13$ we get the prediction $\widehat{X}_{13:20} = 0.51997$ for $X_{13:20} = 0.59333$. The estimated prediction intervals for this prediction are $\widehat{C}_{90} = [0.46744, 0.72438]$ and $\widehat{C}_{50} = [0.48658, 0.57882]$. The exact value belongs to $\widehat{C}_{90}$ and it is out of $\widehat{C}_{50}$. However, in this case, as we estimate the parameter, we do not know the exact coverage probabilities for these intervals. In order to estimate them, we will perform a simulation study in Sect. 5. The predictions from $X_{12:20}$ for $X_{s:20}$ and $s = 13, \ldots, 20$ are plotted in Fig. 8, left, where the blue lines represent the prediction

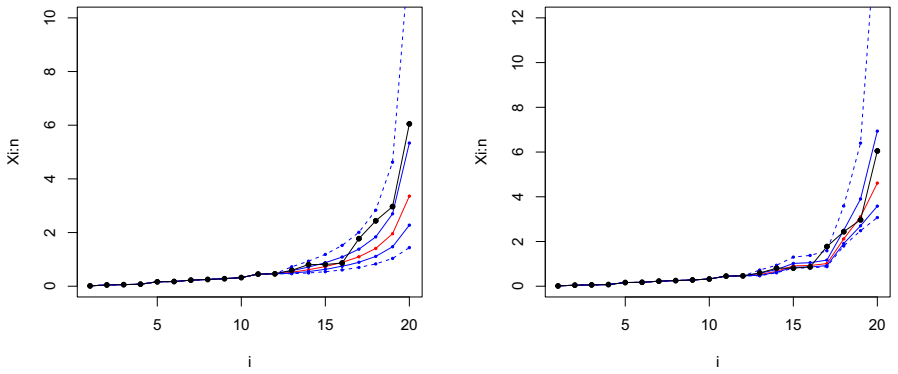

Fig. 8 Predictions (red) for $X_{s:n}$ from $X_{r:n}$ for $n = 20, r = 12$ and $s = 13, \ldots, 20$ (left) and $r = 12, \ldots, 19$ and $s = r + 1$ (right) for the Pareto distribution in Example 2 with estimated θ. The black points are the exact values and the blue lines are the limits for the 50% (continuous blue lines) and the 90% (dashed blue lines) prediction intervals (colour figure online)

Table 2 Predicted values $\widehat{X}_{r+1:n}$ and centered prediction intervals $\widehat{C}_{50} = [l_r, u_r]$ and $\widehat{C}_{90} = [L_r, U_r]$ for $X_{r+1:n}$ from $X_{r:n}$ in a Pareto type II distribution; $\widehat{\theta}_r$ is the estimate of θ at $X_{r:n}$ for $r = 12, \ldots, 19$

r	$\widehat{\theta}_r$	L_r	l_r	$\widehat{X}_{r+1:n}$	$X_{r+1:n}$	u_r	U_r
12	2.28120	0.46744	0.48658	0.51997	0.59333	0.57882	0.72438
13	2.18807	0.59868	0.62354	0.66709	0.80563	0.74427	0.93753
14	2.05372	0.81316	0.84828	0.91011	0.80967	1.02064	1.30257
15	2.19610	0.81814	0.85771	0.92758	0.86477	1.05319	1.37730
16	2.29218	0.87523	0.92420	1.01121	1.77394	1.16914	1.58537
17	1.98395	1.79795	1.91131	2.11655	2.43811	2.50148	3.58870
18	1.95382	2.48354	2.70077	3.10540	2.96807	3.90221	6.40057
19	2.00012	3.07115	3.58190	4.61159	6.05006	6.93582	16.74422

intervals. Note that all the exact values belong to the 90% intervals (dashed blue lines) and that six of them do not belong to the 50% intervals (continuous blue lines).

We do the same in Fig. 8, right, but, in this case, $X_{r+1:20}$ is estimated from the preceding value $X_{r:20}$ by using all the preceding data values $X_{1:20}, \ldots, X_{r:20}$ to estimate θ. The estimations $\widehat{\theta}_r$ for θ at $X_{r:20}$ and the predictions for $X_{r+1:20}$ with $r = 12, \ldots, 19$ are given in Table 2. Note that the estimations obtained for θ are similar. The MLE for θ from the complete sample (which is not available under our assumptions) is $\widehat{\theta}_{20} = 1.98527$ (remember that the exact value is $\theta = 2$). In practice, when we work with real data, the stability of these predictions might confirm the assumed parametric model. Also note that 2 of the exact values do not belong to the 90% prediction intervals and that 4 of them do not belong to the 50% prediction intervals (we expect 0.1 and 4).

In the following example we analyze a real data set by assuming that the original (not ordered) data values are independent (the ordered values are always dependent).

Example 3 Let us study the real data set considered in Bdair and Raqab (2022). They represent ordered lifetimes of 20 electronic components. The complete sample is

$$\begin{array}{cccccccccc} 0.03 & 0.12 & 0.22 & 0.35 & 0.73 & 0.79 & 1.25 & 1.41 & 1.52 & 1.79 \\ 1.80 & 1.94 & 2.38 & 2.40 & 2.87 & 2.99 & 3.14 & 3.17 & 4.72 & 5.09 \end{array}$$

Let us assume that we just know the first 12 failure times and that we want to predict the future failures.

If we assume an exponential distribution with unknown failure rate θ, then we estimate it from (2.5) at $X_{12:20}$ as

$$\widehat{\theta}_{12} = \frac{r}{(n-r+1)x_r + \sum_{i=1}^{r-1} x_i} = \frac{12}{9 \cdot 1.94 + 0.03 + \cdots + 1.8} = 0.43684.$$

From this value we obtain the point predictions and prediction intervals given in Fig. 9, left. For example, the prediction for $X_{13:20} = 2.38$ is $\widehat{X}_{13:20} = 2.13834$ with prediction intervals $C_{90} = [1.95468, 2.797216]$ and $C_{50} = [2.02232, 2.33668]$. The exact value belongs to C_{90} but not to C_{50}. As we can see, the predictions for the last values are not very good. However, all the exact values belong to the 90% prediction intervals and only 4-out-of-8 of them do not belong to the 50% prediction intervals (as expected). Note that this plot is similar to the plot obtained in Fig. 3, left, with a sample of size 20 from an exponential distribution. If we count the data in the four regions R_1, R_2, R_3, R_4 defined in Sect. 2, we get the observed data 3, 3, 1, 1 and the Pearson T statistic value is

$$T = \frac{(3-2)^2}{2} + \frac{(3-2)^2}{2} + \frac{(1-2)^2}{2} + \frac{(1-2)^2}{2} = 2.$$

We can approximate its distribution with a chi-squared distribution with 2 degrees of freedom (since we estimate a parameter), and then its associated P-value is $\Pr(\chi^2_2 > 2) = 0.36788$. So the exponential model cannot be rejected with the complete sample (by using this test).

To check the model with the first 12 values we could estimate θ and predict $X_{r+1:20}$ from $X_{r:20}$ for $r = 1, \ldots, 11$. The predictions can be seen in Fig. 9, right. The estimations $\widehat{\theta}_r$ for θ at $X_{r:20}$ for $r = 1, \ldots, 12$ are

$$\begin{array}{cccccc} 1.66667 & 0.86580 & 0.72993 & 0.63291 & 0.40323 & 0.45113 \\ 0.35461 & 0.36664 & 0.38894 & 0.38300 & 0.41969 & 0.43684 \end{array}$$

These estimations for θ are stable from $r = 5$ to $r = 12$. The MLE estimation with the complete sample is $\widehat{\theta} = 0.51666$. As we can see in that figure the predictions are accurate. Two and six exact points do not belong to the 90% and 50% prediction intervals, respectively. These numbers are close to the expected values ($0.1 \cdot 11 = 1.1$ and $0.5 \cdot 11 = 5.5$). So the exponential model cannot be rejected by using these first 12 values (the P-value obtained with the four regions in Sect. 2 is 0.20374).

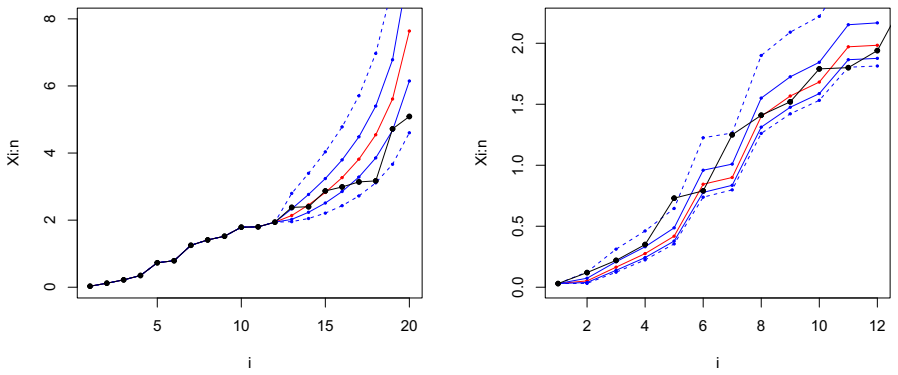

Fig. 9 Predictions (red) for $X_{s:n}$ from $X_{r:n}$ for $n = 20$, $r = 12$ and $s = 13, \ldots, 20$ (left) and $s = r + 1$ and $r = 1, \ldots, 11$ (right) for the real data set in Example 3 with estimated θ under an exponential model. The black points are the exact values and the blue lines are the limits for the 50% (continuous lines) and the 90% (dashed lines) prediction intervals (colour figure online)

In the same framework, if we assume a Pareto type II distribution with unknown parameter θ, then we estimate it from (2.9) as

$$\widehat{\theta} = \frac{r}{(n-r+1)\log(1+x_r) + \sum_{i=1}^{r-1}\log(1+x_i)} = 0.74311.$$

From this value we obtain the predictions and prediction intervals given in Fig. 10, left. The prediction obtained with the Pareto model for $X_{13:20} = 2.38$ is $\widehat{X}_{13:20} = 2.30358$. However, the predictions for the last values are very bad. In fact, 4 of the exact values do not belong to the 90% prediction interval and only 1 of them belongs to the 50% prediction interval. If we count the data in the four regions R_1, R_2, R_3, R_4 defined in Sect. 2, we get the observed data 7, 0, 1, 0 and the Pearson T statistic value is

$$T = \frac{(7-2)^2}{2} + \frac{(0-2)^2}{2} + \frac{(1-2)^2}{2} + \frac{(0-2)^2}{2} = 17.$$

By using again a chi-squared distribution with 2 degrees of freedom, its associated P-value is $\Pr(\chi_2^2 > 17) = 0.00020$ and so the Pareto type II model is rejected with the complete sample. Let us see what happens if we check the model with the first 12 values by estimating θ and predicting $X_{r+1:20}$ from $X_{r:20}$ for $r = 1, \ldots, 11$. The predictions can be seen in Fig. 10, right. The estimations $\widehat{\theta}_r$ obtained for θ at $X_{r:20}$ for $r = 1, \ldots, 12$ are

$$\begin{array}{cccccc} 1.61099 & 0.91625 & 0.80597 & 0.73482 & 0.53125 & 0.60464 \\ 0.53333 & 0.57068 & 0.61839 & 0.63802 & 0.70023 & 0.74311 \end{array}$$

In the figure, we can see that one and seven observed points do not belong to the 90% and 50% prediction intervals, respectively. In this case, the P-value obtained with the

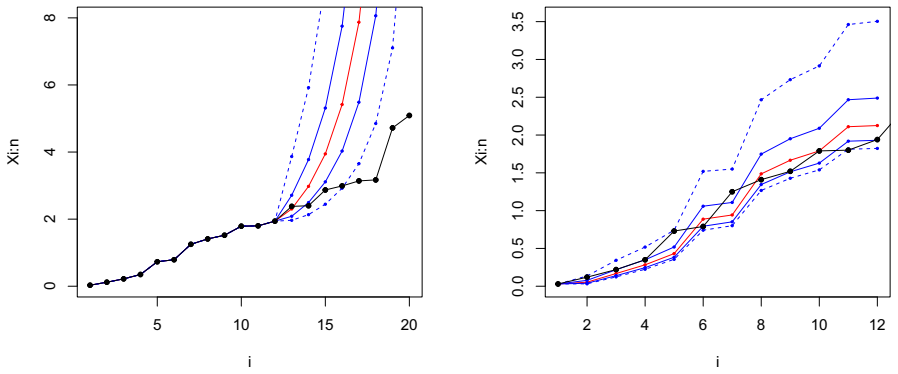

Fig. 10 Predictions (red) for $X_{s:n}$ from $X_{r:n}$ for $n = 20, r = 12$ and $s = 13, \ldots, 20$ (left) and $s = r + 1$ and $r = 1, \ldots, 11$ (right) for the real data set in Example 3 with estimated θ under a Pareto type II model. The black points are the exact values and the blue lines are the limits for the 50% (continuous lines) and the 90% (dashed lines) prediction intervals (colour figure online)

four regions is 0.06843 and the Pareto type II model could be rejected by using these 12 values.

We conclude this section with an example of four dependent data values. They can represent the values in a small sample but they could also be the lifetimes of the four engines in a plane. In the last case, it is very important to predict the future failure times!

Example 4 First we consider the case $r = 1, s = 2$ and $n = 4$, that is, we want to predict $X_{2:4}$ from $X_{1:4} = x$. Let us assume that (X_1, X_2, X_3, X_4) has the following Farlie-Gumbel-Morgenstern (FGM) survival copula

$$\bar{C}(u_1, u_2, u_3, u_4) = u_1u_2u_3u_4 + \theta u_1u_2u_3u_4(1 - u_1)(1 - u_2)(1 - u_3)(1 - u_4) \tag{4.1}$$

for $u_1, u_2, u_3, u_4 \in [0, 1]$ and $\theta \in [-1, 1]$. The independent case is obtained when $\theta = 0$. Then

$$\partial_1\bar{C}(u_1, u_2, u_3, u_4) = u_2u_3u_4 + \theta u_2u_3u_4(1 - 2u_1)(1 - u_2)(1 - u_3)(1 - u_4)$$

and

$$\lim_{v\to 0^+} \partial_1\bar{C}(\bar{F}(x), v, v, v) = \lim_{v\to 0^+} v^3 + \theta(1 - 2\bar{F}(x))v^3(1 - v)^3 = 0$$

for all x. Hence, from (3.4), we get

$$\bar{G}_{2|1:4}(y \mid x) = \frac{\bar{F}^3(y) + \theta\bar{F}^3(y)F^3(y)(1 - 2\bar{F}(x))}{\bar{F}^3(x) + \theta\bar{F}^3(x)F^3(x)(1 - 2\bar{F}(x))}$$

for all $x \le y$ such that $\bar{F}^3(x) + \theta\bar{F}^3(x)F^3(x)(1 - 2\bar{F}(x)) > 0$.

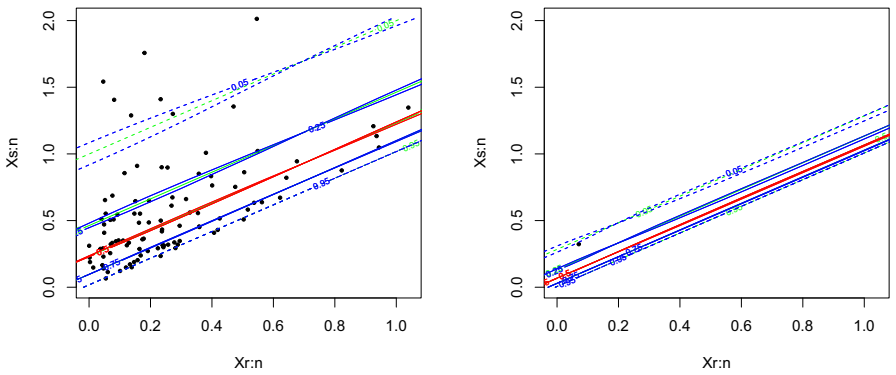

Fig. 11 Predictions (red) for $X_{s:n}$ from $X_{r:n}$ for $n = 4$, $r = 1$ and $s = 2$ for $\theta = -1, 0, 1$ (left) jointly with the values (black point) from 100 simulated samples from a standard exponential distribution and an FGM survival copula with $\theta = 1$ (see Example 4). The blue lines represent the limits for the 50% (continuous lines) and the 90% (dashed lines) prediction intervals. The green lines are the curves for the independent case. In the right plot we have the curves when the mean of the exponential distribution is estimated from the minimum $X_{1:4}$ in the first sample (colour figure online)

Unfortunately, we cannot obtain an explicit expression for its inverse. However, we can plot the level curves of this function to get the plots of the median regression curve (level 0.5) and the limits of the centered prediction regions C_{90} (levels 0.05, 0.95) and C_{50} (levels 0.25, 0.75). They are plotted in Fig. 11, left, jointly with the values obtained from 100 samples of size $n = 4$ from a standard exponential distribution and an FGM survival copula with $\theta = 1$. The (rounded) ordered values obtained in the first sample are

$$0.07086 \quad 0.32313 \quad 0.88360 \quad 1.66760.$$

The method used to generate these sample values is explained in the Appendix. Note that the data values are perfectly represented by these prediction regions. In the right plot we also provide the curves for $\theta = 0$ (green lines) and $\theta = -1$. As we can see the changes are really minor. This is due to the fact that the FGM copula gives a weak dependence relation. The curves might be more different in other dependence models (copulas).

If we assume that the parameters in the model are unknown, we might try to estimate them. Taking into account the preceding comments, instead of estimate θ, we could just plot the curves for the extremes values $\theta = -1, 1$. The exact curves will be between them. So we just need to estimate the parameter $\lambda = 1$ of the exponential distribution. For this purpose, in practice, we have just the sample minimum $X_{1:4}$. Its reliability function is

$$\bar{F}_{1:4}(t) = \bar{C}(\bar{F}(t), \bar{F}(t), \bar{F}(t), \bar{F}(t)) = \bar{F}^4(t) + \theta \bar{F}^4(t) F^4(t)$$

for $t \geq 0$. Hence, if $\bar{F}(t) = \exp(-t/\mu)$ with $\mu = 1/\lambda$, then the mean of $X_{1:4}$ is

$$\begin{aligned} E(X_{1:4}) &= \int_0^\infty (e^{-4t/\mu} + \theta e^{-4t/\mu}(1 - e^{-t/\mu})^4)dt \\ &= \int_0^\infty (1+\theta)e^{-4t/\mu} - 4\theta e^{-5t/\mu} + 6\theta e^{-6t/\mu} - 4\theta e^{-7t/\mu} + \theta e^{-8t/\mu} dt \\ &= \mu\left(\frac{1+\theta}{4} - \frac{4\theta}{5} + \theta - \frac{4\theta}{7} + \frac{\theta}{8}\right). \end{aligned}$$

Therefore, μ can be estimated with

$$\widehat{\mu} = \frac{X_{1:4}}{\theta + \frac{1+\theta}{4} - \frac{4\theta}{5} - \frac{4\theta}{7} + \frac{\theta}{8}}.$$

For $\theta = 1$, we get $\widehat{\mu} = 3.94366 X_{1:4}$ and for $\theta = -1$, $\widehat{\mu} = 4.05797 X_{1:4}$. In our first simulated sample, we obtain the value $X_{1:4} = 0.07086$ and so $\widehat{\mu} \in [0.27945, 0.28755]$. As we are assuming that θ is unknown, we can use the average of these two estimations to approximate μ with 0.2835. By using this value, we get the curves plotted in Fig. 11, right. Although the estimation for $\mu = 1$ is very bad (since we just have one data point) and the curves are far from the observed sample values (plotted in Fig. 11, left), note that the value $X_{2:4}$ belongs to the 90% prediction interval obtained from $X_{1:4}$.

Along the same lines and by using the same copula, we can consider other cases. If we want to predict $X_{3:4}$ from $X_{1:4} = x$, by using (3.5), we get

$$\bar{G}_{3|1:4}(y \mid x) = \frac{A_{3|1:4}(x, y)}{\bar{F}^3(x) + \theta \bar{F}^3(x) F^3(x)(1 - 2\bar{F}(x))}$$

for all $x \leq y$ such that $\bar{F}^3(x) + \theta \bar{F}^3(x) F^3(x)(1 - 2\bar{F}(x)) > 0$, where

$$\begin{aligned} A_{3|1:4}(x, y) = {} & 3\bar{F}(x)\bar{F}^2(y) + 3\theta \bar{F}(x)\bar{F}^2(y)F(x)F^2(y)(1 - 2\bar{F}(x)) \\ & - 2\bar{F}^3(y) - 2\theta \bar{F}^3(y)F^3(y)(1 - 2\bar{F}(x)). \end{aligned}$$

As in the preceding case, we plot the level curves of this function to get the plots of the median regression curve (level 0.5) and the limits of the centered prediction regions C_{90} (levels 0.05, 0.95) and C_{50} (levels 0.25, 0.75). They are plotted in Fig. 12, left, jointly with the values from 100 samples of size $n = 4$ from a standard exponential distribution and an FGM survival copula with $\theta = 1$. In this case we do not plot also the curves for $\theta = 0$ and $\theta = -1$ since the changes are again really minor.

Furthermore, if we want to predict $X_{4:4}$ from $X_{1:4} = x$, by using (3.6), we get

$$\bar{G}_{4|1:4}(y \mid x) = \frac{A_{4|1:4}(x, y)}{\bar{F}^3(x) + \theta \bar{F}^3(x) F^3(x)(1 - 2\bar{F}(x))}$$

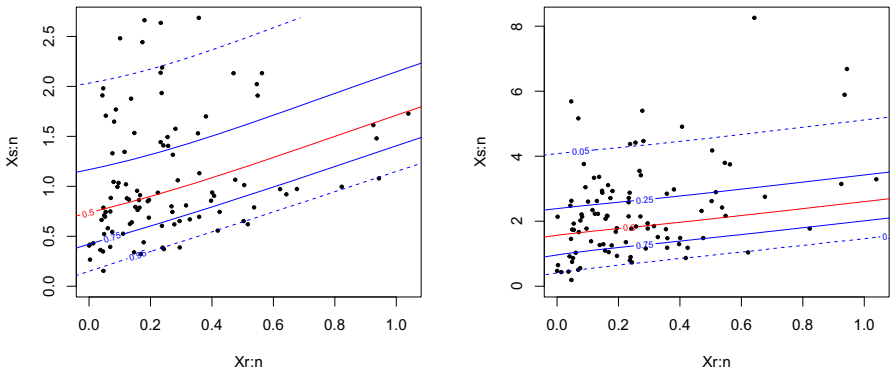

Fig. 12 Predictions (red) for $X_{s:n}$ from $X_{r:n}$ for $n = 4$, $r = 1, s = 3$ (left) and $s = 4$ (right) for $\theta = 1$ jointly with the values (black point) from 100 simulated samples from a standard exponential distribution and an FGM survival copula with $\theta = 1$ (see Example 4). The blue lines represent the limits for the 50% (continuous lines) and the 90% (dashed lines) prediction intervals (colour figure online)

for all $x \leq y$ such that $\bar{F}^3(x) + \theta \bar{F}^3(x) F^3(x)(1 - 2\bar{F}(x)) > 0$, where

$$\begin{aligned} A_{4|1:4}(x, y) = {} & 3\bar{F}^2(x)\bar{F}(y) + 3\theta \bar{F}^2(x)\bar{F}(y)F^2(x)F(y)(1 - 2\bar{F}(x)) \\ & - 3\bar{F}(x)\bar{F}^2(y) - 3\theta \bar{F}(x)\bar{F}^2(y)F(x)F^2(y)(1 - 2\bar{F}(x)) \\ & + \bar{F}^3(y) + \theta \bar{F}^3(y)F^3(y)(1 - 2\bar{F}(x)). \end{aligned}$$

We plot the level curves of this function to get the plots of the median regression curve (level 0.5) and the limits of the centered prediction regions C_{90} (levels 0.05, 0.95) and C_{50} (levels 0.25, 0.75). They are plotted in Fig. 12, right, jointly with the values from 100 samples of size $n = 4$ from a standard exponential distribution and an FGM survival copula with $\theta = 1$. Again, we do not plot the curves for $\theta = 0$ and $\theta = -1$ since the changes are again really minor.

Proceeding as above, we can predict $X_{2:4}$, $X_{3:4}$ and $X_{4:4}$ by using the median regression curve and we can obtain the limits of the centered prediction regions C_{90} and C_{50}. For example, in the first sample, the prediction obtained for $X_{2:4} = 0.32313$ from $X_{1:4} = 0.70086$ is $\widehat{X}_{2:4} = 0.29708$ with prediction intervals $C_{50} = [0.16582, 0.51384]$ and $C_{90} = [0.08788, 0.98928]$. Analogously, the prediction for $X_{3:4} = 0.88360$ from $X_{1:4} = 0.70086$ is $\widehat{X}_{3:4} = 0.79118$ with prediction intervals $C_{50} = [0.48511, 1.21566]$ and $C_{90} = [0.22238, 2.07476]$. Finally, the prediction for $X_{4:4} = 1.6676$ from $X_{1:4} = 0.70086$ is $\widehat{X}_{4:4} = 1.64681$ with prediction intervals $C_{50} = [1.05428, 2.46201]$ and $C_{90} = [0.50708, 4.14529]$. In Fig. 13, left, we plot these predictions (red) for $X_{2:4}$, $X_{3:4}$, $X_{4:4}$ from $X_{1:4}$ jointly with the exact values (black points) in the first simulated sample from a standard exponential distribution and an FGM survival copula with $\theta = 1$.

Finally, we can use more that one data point to predict future failures. For example, we can predict $X_{3:4}$ from $X_{1:4} = 0.07086$ and $X_{2:4} = 0.32313$ by using (3.8). With

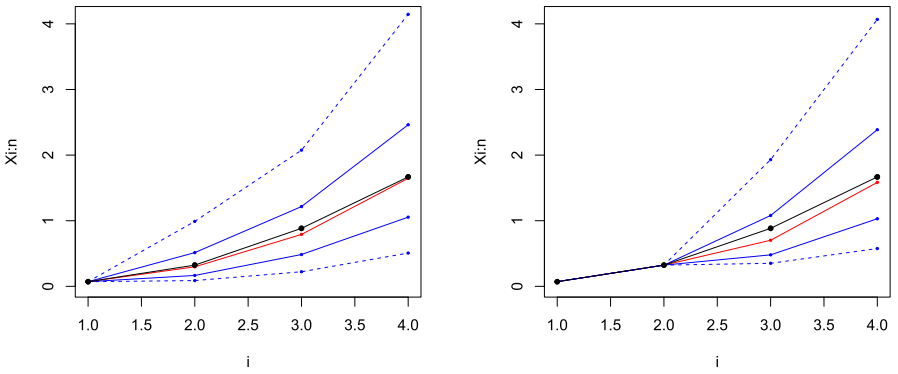

Fig. 13 Predictions (red) for $X_{s:n}$ from $X_{r:n}$ for $n = 4, r = 1$ and $s = 2, 3, 4$ (left) jointly with the exact values (black points) from a simulated samples from a standard exponential distribution and an FGM survival copula with $\theta = 1$ (see Example 4). The blue lines represent the limits for the 50% (continuous lines) and the 90% (dashed lines) prediction intervals. In the right plot we can see the predictions for $X_{3:4}$ and $X_{4:4}$ from $X_{1:4}$ and $X_{2:4}$ (colour figure online)

the FGM copula we get

$$\bar{G}_{3|1,2}(t \mid x, y) = \frac{1+\theta(1-2\bar{F}(x))(1-2\bar{F}(y))F^2(t)}{1+\theta(1-2\bar{F}(x))(1-2\bar{F}(y))F^2(y)} \cdot \frac{\bar{F}^2(t)}{\bar{F}^2(y)}$$

for $x < y \leq t$. By solving $\bar{G}_{3|1,2}(t \mid x, y) = 0.5$ we get the prediction $\widehat{X}_{3:4} = 0.70208$ for $X_{3:4} = 0.8836$. Analogously, we obtain the prediction intervals $C_{50} = [0.47975, 1.07961]$ and $C_{90} = [0.3509, 1.93089]$. In a similar way, we can predict $X_{4:4} = 1.58455$ from $X_{1:4} = 0.07086$ and $X_{2:4} = 0.32313$ by using (3.9) obtaining $\widehat{X}_{4:4} = 1.58455, C_{50} = [1.02971, 2.38660]$ and $C_{90} = [0.57592, 4.06791]$. The predictions are plotted in Fig. 13, right.

5 Simulation study

In this section we show a simulation study to estimate the coverage probabilities of the prediction regions when we estimate the parameter in the PHR model for samples of IID random variables.

First, let us assume the exponential model with parameter $\theta = 1$, $\bar{F}(t) = e^{-t}$ for $t \geq 0$. We generate N samples of size 20 and, by supposing that the parameter $\theta > 0$ is unknown, we use $X_{12:20}$ to predict $X_{s:20}, s = 13, 14, 20$. For each sample we use (2.5) with $r = 12$ to estimate θ, and we obtain $\widehat{\theta}_j$, $j = 1, \ldots, N$. Replacing the exact survival function $\bar{F}(t) = e^{-t}$ with $\bar{F}_{\widehat{\theta}_j}(t) = e^{-\widehat{\theta}_j}$, $j = 1, \ldots, N$, we can obtain predictions for $X_{s:20}$ from $X_{12:20}$ for each simulated sample. Our purpose is to estimate the coverage probabilities for the estimated prediction intervals $\widehat{C}_{90}$ and $\widehat{C}_{50}$ varying s and N. The results are listed in Table 3.

Furthermore, we perform the same study by choosing as baseline distribution for the PHR model the Pareto type II distribution, that is, $\bar{F}(t) = 1/(1+t)^{\theta}$ for $t \geq 0$.

Table 3 Number of observed values of $X_{s:20}$ in $\widehat{C}_{90}$ and $\widehat{C}_{50}$ for varying s (13, 14 and 20) and N (500, 1000 and 10000) for the exponential model

	$\widehat{C}_{90}$			$\widehat{C}_{50}$		
	N			N		
s	500	1000	10000	500	1000	10000
13	446	890	8869	259	512	4875
14	449	890	8705	233	483	4665
20	426	853	8309	210	441	4236

Table 4 Number of observed values of $X_{s:20}$ in $\widehat{C}_{90}$ and $\widehat{C}_{50}$ for varying s (13, 14 and 20) and N (500, 1000 and 10000) for the Pareto type II model

	$\widehat{C}_{90}$			$\widehat{C}_{50}$		
	N			N		
s	500	1000	10000	500	1000	10000
13	439	880	8812	247	472	4841
14	435	858	8711	251	491	4787
20	414	838	8373	202	416	4260

We choose $\theta = 2$. For each sample we use (2.9) with $r = 12$ to estimate θ, and we obtain $\widehat{\theta}_j$, $j = 1, \ldots, N$. Replacing the exact reliability function $\bar{F}(t) = 1/(1+t)^2$ with $\bar{F}_{\widehat{\theta}_j}(t) = 1/(1+t)^{\widehat{\theta}_j}$, $j = 1, \ldots, N$, we obtain predictions for $X_{s:20}$ from $X_{12:20}$ for each simulated sample.

The results about the coverage probabilities of the estimated prediction intervals are given in Table 4. As we can see, in both cases, the coverage probabilities are a little bit below of the expected values (for the exact model), especially when we predict the last value $X_{20:20}$ from $X_{12:20}$. Note that in both models (exponential and Pareto), we have some extreme upper values (especially in the Pareto model).

6 Conclusions

We have proved that quantile regression techniques can capture the underlying uncertainty in the prediction of future sample values. We consider both the cases of IID and DID samples. In both cases, if we know the underlying model (or it can be estimated from preceding right censored samples), the predictions are accurate. Even more, we can perform fit tests to confirm the model. If the model contains unknown parameters, then they should be estimated from the available values. In those cases, the predictions are not so accurate. However, the coverage probabilities obtained in the simulation study when the parameters are estimated are close to the expected ones. In the IID case we also provide some point predictors based on the median, the mean or the mode, that are compared with classical estimators. However, we recommend to use prediction intervals instead of point predictions since they represent better the uncertainty in the future failure times.

There are several tasks for future research. The main one could be the estimation of the unknown parameters in other parametric models (different to the PHR model) or

other dependence models (copulas). The expressions for the dependence case for $n > 4$ and/or for non-exchangeable joint distributions (copulas) should also be obtained following a procedure similar to the one proposed here. In this case, the estimation of the parameters in the model (both in the distribution and in the copula) is a more complex (but important) task.

Acknowledgements We would like to thank the anonymous reviewer for several helpful suggestions that have served to improve the earlier version of this paper. JN is partially supported by "Ministerio de Ciencia e Innovación" of Spain under Grant PID2019-103971GB-I00/AEI/10.13039/501100011033. FB is member of the research group GNAMPA of INdAM (Istituto Nazionale di Alta Matematica) and is partially supported by MIUR-PRIN 2017, project "Stochastic Models for Complex Systems", No. 2017 JFFHSH.

Funding Open Access funding provided thanks to the CRUE-CSIC agreement with Springer Nature.

Declarations

Conflict of interest. On behalf of all authors, the corresponding author states that there is no conflict of interest.

Appendix A: Sample from an FGM copula

Let us see how to generate a sample (U_1, U_2, U_3, U_4) from the survival copula $\bar{C}$. Then the sample from (X_1, X_2, X_3, X_4) with a common reliability $\bar{F}$ is obtained as $(\bar{F}(U_1), \bar{F}(U_2), \bar{F}(U_3), \bar{F}(U_4))$.

The joint distribution function of (U_1, U_2, U_3, U_4) is given in (4.1). Then, its joint PDF is

$$\bar{c}(u_1, u_2, u_3, u_4) = 1 + \theta(1 - 2u_1)(1 - 2u_2)(1 - 2u_3)(1 - 2u_4)$$

for $u_1, u_2, u_3, u_4 \in (0, 1)$. The joint distribution function of (U_1, U_2, U_3) is

$$\bar{C}_{1,2,3}(u_1, u_2, u_3) = \bar{C}(u_1, u_2, u_3, 1) = u_1 u_2 u_3$$

and so its joint PDF is $\bar{c}_{1,2,3}(u_1, u_2, u_3) = 1$ for $u_1, u_2, u_3 \in (0, 1)$. So they are IID and can be simulated just as independent uniform random variables.

The conditional PDF of U_4 given $U_1 = u_1, U_2 = u_2, U_3 = u_3$ is obtained as

$$\bar{c}_{4|1,2,3}(u_4 \mid u_1, u_2, u_3) = \bar{c}(u_1, u_2, u_3, u_4)$$

for $u_1, u_2, u_3, u_4 \in (0, 1)$. Therefore, its distribution function is

$$\begin{aligned}\bar{C}_{4|1,2,3}(u_4 \mid u_1, u_2, u_3) &= \int_0^{u_4} \bar{c}(u_1, u_2, u_3, z)dz \\ &= \int_0^{u_4} (1 + \theta(1-2u_1)(1-2u_2)(1-2u_3)(1-2z))dz \\ &= u_4 + \theta(1-2u_1)(1-2u_2)(1-2u_3)(u_4 - u_4^2)\end{aligned}$$

for $u_4 \in [0, 1]$. To get its inverse function we must solve

$$\bar{C}_{4|1,2,3}(u_4 \mid u_1, u_2, u_3) = q$$

for $0 < q < 1$, which leads to

$$\bar{C}^{-1}_{4|1,2,3}(q \mid u_1, u_2, u_3) = \frac{1 + A - \sqrt{A^2 + 1 + 2A(1-2q)}}{2A}$$

when $A \neq 0$, where $A = \theta(1-2u_1)(1-2u_2)(1-2u_3)$ (the other solution of the quadratic equation does not belong to the interval $[0, 1]$). In the simulation, as U_1, U_2, U_3 are independent random numbers in $(0, 1)$, then the event $A = 0$ has probability zero.

References

Arnold BC, Balakrishnan N, Nagaraja HN (2008) A first course in order statistics. SIAM, Philadelphia

Balakrishnan N, Kundu D, Ng HKT, Kannan N (2007) Point and interval estimation for a simple step-stress model with type-II censoring. J Qual Technol 39:35–47

Barakat HM, El-Adll ME, Aly AE (2011) Exact prediction intervals for future exponential lifetime based on random generalized order statistics. Comput Math Appl 61:1366–1378

Barakat HM, Khaled OM, Ghonem HA (2022) Predicting future lifetime for mixture exponential distribution. Commun Stat Comput Simul. 51:3533–3552. https://doi.org/10.1080/03610918.2020.1715434

Basiri E, Ahmadi J, Raqab MZ (2016) Comparison among non-parametric prediction intervals of order statistics. Commun Stat Theory Methods 45:2699–2713

Bdair OM, Raqab MZ (2022) Prediction of future censored lifetimes from mixture exponential distribution. Metrika 85:833–857. https://doi.org/10.1007/s00184-021-00852-z

Cramer E (2021) Ordered and censored lifetime data in reliability: an illustrative review. WIREs Comput Stat. https://doi.org/10.1002/wics.1571

David HA, Nagaraja HN (2003) Order statistics, 3rd edn. Wiley, Hoboken

El-Adll ME (2011) Predicting future lifetime based on random number of three parameters Weibull distribution. Comput Math Appl 81:1842–1854

Greenwood PE, Nikulin MS (1996) A guide to Chi-squared testing. Wiley, New York

Johnson NL, Kotz S, Balakrishnan N (1985) Continuous univariate distributions, vol 2. Wiley, Hoboken

Khaminsky KS, Rhodin LS (1985) Maximum likelihood prediction. Ann Inst Stat Math 37:507–517

Koenker R (2005) Quantile regression. Cambridge University Press, Cambridge

Lawless JF, Fredette M (2005) Frequentist prediction intervals and predictive distributions. Biometrika 92:529–542

Navarro J (2020) Bivariate box plots based on quantile regression curves. Depend Model 8:132–156. https://doi.org/10.1515/demo-2020-0008

Navarro J (2022a) Introduction to system reliability theory. Springer

Navarro J (2022b) Prediction of record values by using quantile regression curves and distortion functions. Metrika 85:675–706

Navarro J, Calì C, Longobardi M, Durante F (2022) Distortion representations of multivariate distributions. Stat Methods Appl. 31:925–954. https://doi.org/10.1007/s10260-021-00613-2

Van Dorp JR, Mazzuchi TA (2000) Solving for the parameters of a beta distribution under two quantile constraints. J Stat Comput Simul 67:189–201