Journal of Hydrology 512 (2014) 498–505

Contents lists available at ScienceDirect

Journal of Hydrology

journal homepage: www.elsevier.com/locate/jhydrol

Correcting confidence intervals for quantiles of a heavy-tailed distribution: Case of the two-parameter Kappa distribution

Fahim Ashkar *, Salaheddine El Adlouni [1]

Department of Mathematics and Statistics, Université de Moncton, Moncton, NB E1A 3E9, Canada

ARTICLE INFO

Article history:
Received 8 September 2013
Received in revised form 31 January 2014
Accepted 10 March 2014
Available online 18 March 2014
This manuscript was handled by Andras Bardossy, Editor-in-Chief, with the assistance of Attilio Castellarin, Associate Editor

Keywords:
2-Parameter Kappa distribution
Quantiles
Confidence intervals
Maximum likelihood
Small samples

SUMMARY

In modeling hydrological phenomena, statistical distributions are commonly used as frequency models to fit hydrological data. The 2-parameter Kappa (KAP) distribution has been proposed to analyze precipitation, wind speed and stream-flow data. Estimates of distribution quantiles are important risk measures of the frequency of occurrence of extreme hydrological events, and the calculation of confidence intervals for these quantiles (CIQs) is also essential, as it provides a measure of the statistical error involved in the estimation. This study revisits the most frequently used method for calculating CIQs for the KAP distribution and proposes a method for improving their accuracy. The calculation of CIQs has traditionally been based on the large-sample assumption that the quantile estimators are normally distributed, but with small samples commonly available, this assumption is shown to be quite crude. It is shown that significantly more accurate CIQs are obtainable if the KAP quantile estimators are transformed to better fit a normal distribution, and then corrected for possible bias. The comparison among CIQs is done on the basis of their coverage probabilities of the true distribution quantile. The results of the comparison lead to improved methods for calculating CIQs for the KAP distribution, the application of which is illustrated through a hydrological example. Although the study restricts attention to the maximum likelihood (ML) fitting method, we anticipate that the drawn recommendations would apply to other fitting methods also.

1. Introduction

In modeling hydrological phenomena, statistical distributions are commonly used as frequency models to fit hydrological data. Some three-parameter frequency models such as the generalized extreme value (GEV) distribution have been widely used to fit data such as the annual maxima or minima of hydrological data series. However, no fits of any one particular distribution family to finite hydrological data series have been found to be always adequate.

Despite the demonstrated usefulness of frequency models with three or more parameters, they are not the ones to be always recommended. In fact, in many situations involving limited amounts of data, two-parameter models may be more appropriate because additional parameters may lead to "over-fitting" of the data. Moreover, two-parameter models are the ones typically used in certain types of modeling, such as in the Peaks-Over-Threshold (POT) approach to modeling hydrological extremes; e.g., for fitting flood peaks above a threshold. Such distributions have also been recommended for fitting low stream-flow data by the Deficit-Below-Threshold (DBT) approach (e.g., Ashkar et al., 2004). For these practical reasons, it is necessary to give two-parameter distributions the attention that they deserve in hydrologic modeling.

The 2-parameter Kappa distribution, which we shall simply denote by KAP, has the convenience of possessing closed algebraic expressions for its cumulative distribution function (CDF) and its quantile function. It belongs to an important group of two-parameter models widely used in hydrological modeling that: (i) have one scale parameter and one shape parameter, (ii) are defined on the interval $(0, +\infty)$, and (iii) are positively skewed. One feature of the KAP model that makes it particularly useful for modeling hydrologic extremes is that it belongs to the class of heavy-tailed distributions (El Adlouni et al., 2008). Ashkar et al. (2013) discuss some similarities and shape differences between the KAP probability density function (PDF) and that of other 2-parameter models such as the lognormal, gamma, Weibull, generalized Pareto and log–logistic. Studies such as those of Mielke (1973), Park et al. (2009) and Ashkar et al. (2004, 2013) have shown how the KAP model can be useful in fitting meteorological data, as well as precipitation or stream flow data above or below a threshold.

* Corresponding author. Tel.: +1 (506) 858 4312; fax: +1 (506) 858 4541.
E-mail addresses: ashkarf@umoncton.ca (F. Ashkar), salah-eddine.el.adlouni@umoncton.ca (S. El Adlouni).
[1] Tel.: +1 (506) 858 4253; fax: +1 (506) 858 4541.

http://dx.doi.org/10.1016/j.jhydrol.2014.03.025
0022-1694/

The KAP distribution with scale parameter β and shape parameter α will be denoted by KAP(α, β). The PDF and CDF of this distribution are respectively given by

$$f(x;\alpha,\beta) = (\alpha/\beta)[\alpha + (x/\beta)^{\alpha}]^{-(\alpha+1)/\alpha}; \quad 0 < x < \infty,\ \beta > 0,\ \alpha > 0 \tag{1}$$

$$F(x;\alpha,\beta) = (x/\beta)[\alpha + (x/\beta)^{\alpha}]^{-1/\alpha} \tag{2}$$

and its quantile function is given by

$$Q(p;\alpha,\beta) = F^{-1}(p;\alpha,\beta) = \beta p[\alpha/(1-p^{\alpha})]^{1/\alpha} \tag{3}$$

where $p = F(x;\alpha,\beta)$, $0 < p < 1$.

Many parameter estimation methods have been proposed to fit statistical distributions to data, but the maximum likelihood (ML) method stands out as a particularly important fitting method because it generally leads to efficient estimators with Gaussian asymptotic distributions (see, e.g. Hogg et al., 2004). The ML fitting method was also the one recommended for the KAP distribution by Park et al. (2009). A key goal in fitting distributions to data is to be able to estimate distribution quantiles or percentiles. It is also essential to provide a measure of the statistical error that is involved in the estimation. This is commonly done by constructing confidence intervals for the distribution's quantiles (CIQs), assuming that the chosen fitting model is the correct one (*model uncertainty* is also very important, but will not be considered in the present study). Hydrologists have traditionally used *large-sample theory* to construct such CIQs. However, it is shown in the present study that such CIQs are very inaccurate for right-tail quantile estimation in the case of a heavy-tailed distribution such as KAP. The main goal of this paper is to suggest an improvement to these classically obtained CIQs, under a KAP model, and to show that this improvement is both practical and useful. To do this, we will begin by presenting Fisher's information for the KAP model, and then review how this information has traditionally been used to derive large-sample CIQs.

The paper is organized as follows. Section 2 briefly discusses parameter estimation by ML for the KAP distribution and presents the basic asymptotic properties of the ML parameter estimators (MLEs) based on Fisher's information. Section 3 presents the approximate sampling distribution of the KAP quantile estimators. The most commonly used approach for calculating CIQs is then revisited, and an improvement to this approach is proposed. Then, through Monte Carlo (MC) simulation, the different approaches are compared in Section 4. In Section 5, the recommendations resulting from the MC simulations are put to use in a hydrological application. Finally, Section 6 presents the paper's main conclusions and briefly presents some future research ideas.

2. KAP MLEs and their asymptotic properties

Let $\mathbf{X}_n = \{X_i\}_{i=1}^n$ denote a random sample of size n from a KAP(α, β) distribution with PDF given in Eq. (1); i.e., the X_i's are independent and identically distributed (iid) with PDF $f(x;\alpha,\beta)$. The log-likelihood function is given by

$$\begin{aligned} l(\alpha,\beta|\mathbf{X}_n) &= \sum_{i=1}^{n} \ln f(X_i;\alpha,\beta) \\ &= n\ln\alpha - n\ln\beta - \frac{\alpha+1}{\alpha}\sum_{i=1}^{n}\ln[\alpha + (X_i/\beta)^{\alpha}] \end{aligned} \tag{4}$$

Maximizing the log-likelihood function $l(\alpha,\beta|\mathbf{X}_n)$ comes down to solving the following system of equations for the MLEs:

$$\begin{cases} \left.\frac{\partial l(\alpha,\beta|\mathbf{X}_n)}{\partial\alpha}\right|_{\substack{\alpha=\hat{\alpha}\\ \beta=\hat{\beta}}} = 0 \\ \left.\frac{\partial l(\alpha,\beta|\mathbf{X}_n)}{\partial\beta}\right|_{\substack{\alpha=\hat{\alpha}\\ \beta=\hat{\beta}}} = 0 \end{cases} \tag{5}$$

which, for the KAP distribution, comes down to solving the following system:

$$\begin{cases} \sum_{i=1}^{n}\ln\left[\hat{\alpha} + (X_i/\hat{\beta})^{\hat{\alpha}}\right] - \hat{\alpha}(\hat{\alpha}+1)\sum_{i=1}^{n}\frac{(X_i/\hat{\beta})^{\hat{\alpha}}\ln[X_i/\hat{\beta}]}{\hat{\alpha}+(X_i/\hat{\beta})^{\hat{\alpha}}} = 0 \\ n - (\hat{\alpha}+1)\sum_{i=1}^{n}\left[\hat{\alpha} + (X_i/\hat{\beta})^{\hat{\alpha}}\right]^{-1} = 0 \end{cases} \tag{6}$$

the solution of which has to be done numerically.

Let $(\dot{\alpha},\dot{\beta})$be a specific value of the parameter vector (α, β). Based on the log-likelihood function of Eq. (4), the observed (or sample-based) Fisher's information matrix, evaluated at $(\dot{\alpha},\dot{\beta})$,is given by

$$\mathbf{J}_n(\dot{\alpha},\dot{\beta}) = -\begin{pmatrix} \frac{\partial^2 l(\alpha,\beta|\mathbf{X}_n)}{\partial\alpha^2} & \frac{\partial^2 l(\alpha,\beta|\mathbf{X}_n)}{\partial\alpha\partial\beta} \\ \frac{\partial^2 l(\alpha,\beta|\mathbf{X}_n)}{\partial\alpha\partial\beta} & \frac{\partial^2 l(\alpha,\beta|\mathbf{X}_n)}{\partial\beta^2} \end{pmatrix}_{\substack{\alpha=\dot{\alpha}\\ \beta=\dot{\beta}}} = \begin{bmatrix} J_{11} & J_{12} \\ J_{21} & J_{22} \end{bmatrix}_{\substack{\alpha=\dot{\alpha}\\ \beta=\dot{\beta}}} \tag{7}$$

which is a random matrix, because it is a function of the random sample $\mathbf{X}_n$. The elements of this matrix $\mathbf{J}$ are mathematically developed in Appendix A.

By inverting the last term of Eq. (7), the inverse of Fisher's observed information matrix is obtained, which is key for calculating asymptotic CIQs:

$$\mathbf{J}_n^{-1}(\dot{\alpha},\dot{\beta}) = \begin{bmatrix} Var(\hat{\alpha}) & \mathrm{Cov}(\hat{\alpha},\hat{\beta}) \\ \mathrm{Cov}(\hat{\alpha},\hat{\beta}) & Var(\hat{\beta}) \end{bmatrix} \tag{8}$$

Note that another matrix of key importance in information theory is Fisher's (*expected*) information matrix, $\mathbf{I}_n(\dot{\alpha},\dot{\beta})$, which is defined as the *expected value* of the observed information matrix:

$$\mathbf{I}_n(\dot{\alpha},\dot{\beta}) = E[\mathbf{J}_n(\dot{\alpha},\dot{\beta})] \tag{9}$$

Unlike the matrix $\mathbf{J}_n(\dot{\alpha},\dot{\beta})$,this matrix $\mathbf{I}_n(\dot{\alpha},\dot{\beta})$ is not random because it is equal to the expectation of $\mathbf{J}_n(\dot{\alpha},\dot{\beta})$ over all possible random samples $\mathbf{X}_n$.

Both the observed and the expected information matrices (Eqs. (7) and (9), respectively) can be used as a basis for constructing CIQs. However, in a notable paper, Efron and Hinkley (1978) argue that the observed information matrix $\mathbf{J}_n(\dot{\alpha},\dot{\beta})$ should be favoured over the expected information matrix $\mathbf{I}_n(\dot{\alpha},\dot{\beta})$.

The authors' experience also supports this argument by Efron and Hinkley (1978). An additional practical advantage of using the matrix $\mathbf{J}_n(\dot{\alpha},\dot{\beta})$, instead of the matrix $\mathbf{I}_n(\dot{\alpha},\dot{\beta})$, is that the mathematical calculation of $\mathbf{I}_n(\dot{\alpha},\dot{\beta})$ is much more mathematically tedious than calculating $\mathbf{J}_n(\dot{\alpha},\dot{\beta})$.For these reasons, we will base our CIQ calculations on Eq. (7), rather than on Eq. (9).

Denoting by α_0 and β_0 the *true* parameter values of the distribution, the key asymptotic result upon which the calculation of CIQs is based, is

$$\begin{pmatrix} \hat{\alpha} \\ \hat{\beta} \end{pmatrix}^{n\to\infty} \sim N\left[\begin{pmatrix} \alpha_0 \\ \beta_0 \end{pmatrix}, \mathbf{J}_n^{-1}(\alpha_0,\beta_0)\right] \tag{10}$$

which states that the MLE vector $(\hat{\alpha},\hat{\beta})$ is asymptotically normally distributed with mean vector equal to (α_0, β_0) and covariance matrix equal to the inverse of Fisher's observed information matrix evaluated at (α_0, β_0).

3. Approximate CIQs

Based on the approximate normality of the MLEs presented in Eq. (10), it is now possible to obtain approximate distributions for various useful functions of the MLEs. One such function that is of key importance in hydrologic design and risk analysis is the p^{th} quantile estimator $Q_n(p;\hat{\alpha},\hat{\beta})$. Our main goal will be to obtain the best possible approximate distribution for $Q_n(p;\hat{\alpha},\hat{\beta})$, and to