Model selection techniques for the frequency analysis of hydrological extremes: the MSClaio2008 R function

Alberto Viglione

Abstract

The frequency analysis of hydrological extremes requires fitting a probability distribution to the observed data to suitably represent the frequency of occurrence of rare events. The choice of the model to be used for statistical inference is often based on subjective criteria, or it is considered a matter of probabilistic hypotheses testing. In contrast, specific tools for model selection, like the well known Akaike Information Criterion (AIC) and Bayesian Information Criterion (BIC), are seldom used in hydrological applications. The paper of Laio et˜al. (2008) verifies whether the AIC and BIC work correctly when they are applied for identifying the probability distribution of hydrological extremes, i.e. when the available samples are small and the parent distribution is highly asymmetric. An additional model selection criterion, based on the Anderson-Darling goodness-of-fit test statistic, is proposed, and the performances of the three methods are compared trough an extensive numerical analysis. In this brief document, an application of the R function `MSClaio2008`, part of the package `nsRFA`, is provided.

Introduction

The problem of model selection can be formalized as follows: a sample of n data, $D = (x_1, \ldots, x_n)$, arranged in ascending order is available, sampled from an unknown parent distribution $f(x)$; N_m operating models, M_j, $j = 1, \ldots, N_m$, are used to represent the data. The operating models are in the form of probability distributions, $M_j = g_j(x, \hat{\theta})$, with parameters $\hat{\theta}$ estimated from the available data sample D. The scope of model selection is to identify the model M_{opt} which is better suited to represent the data, i.e. the model which is closer in some sense to the parent distribution $f(x)$.

Three different model selection criteria are considered here, namely, the Akaike Information Criterion (AIC), the Bayesian Information Criterion (BIC), and the Anderson-Darling Criterion (ADC). Of the three methods, the first two belong to the category of classical literature approaches, while the third derives from a heuristic interpretation of the results of a standard goodness-of-fit test (see Laio, 2004).

The R function `MSClaio2008`, part of the package `nsRFA`, is used on a data sample from the FEH database:

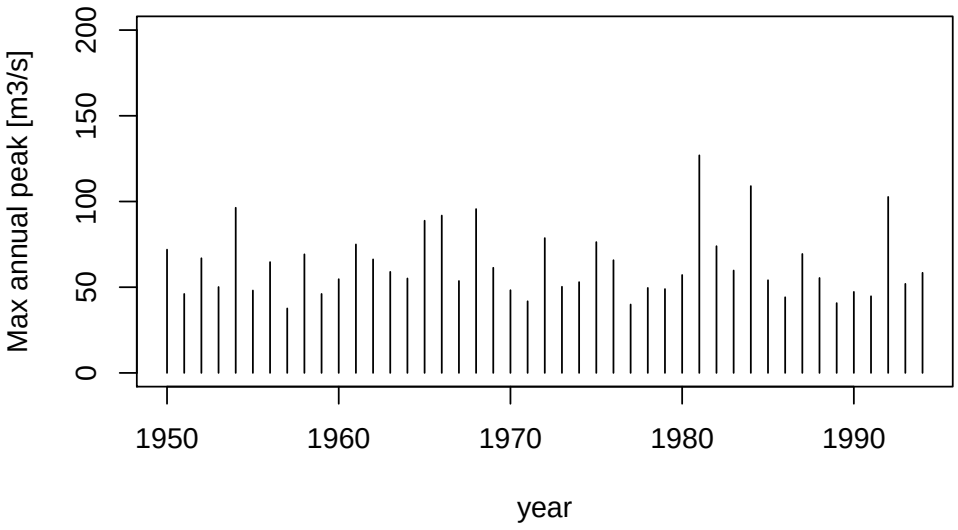
```
> data(FEH1000)
```

The data of site number 69023 are used here:

```
> sitedata <- am[am[,1]==69023, ]
```

whose series can be plotted with:

```
> serieplot(sitedata[,4], sitedata[,3], ylim=c(0,200),
+           xlab="year", ylab="Max annual peak [m3/s]")
```



Series of maximum annual flood peaks in station 69023.

Akalike Information Criterion

The Akaike information Criterion (AIC) for the j-th operational model can be computed as

$$AIC_j = -2\ln(L_j(\hat{\theta})) + 2p_j$$

where

$$L_j(\hat{\theta}) = \prod_{i=1}^{n} g_j(x_i, \hat{\theta})$$

is the likelihood function, evaluated at the point $\theta = \hat{\theta}$ corresponding to the maximum likelihood estimator of the parameter vector θ and p_j is the number of estimated parameter of the j-th operational model. In practice, after the computation of the AIC_j, for all of the operating models, one selects the model with the minimum AIC value, AIC_{min}. The application of the AIC method is performed by:

```
> MSC <- MSClaio2008(sitedata[,4], crit="AIC")
> MSC

------------------------
Akaike Information Criterion (AIC):
  NORM      LN  GUMBEL     EV2     GEV      P3     LP3
 400.7   388.4   386.5   383.6   385.6   381.6   384.9
```

Summarizing the choice is:

```
> summary(MSC)

Tested distributions:
[1]  NORM    LN      GUMBEL  EV2     GEV     P3      LP3
------------------------
Chosen distributions:
```

```
AIC
 P3
whose Maximum-Likelihood parameters are:
P3 parameters of x:  37.625  14.22391  1.842626
```

More information on the function `MSClaio2008` can be obtained by:

```
> help(MSClaio2008)
```

When the sample size, n, is small, with respect to the number of estimated parameters, p, the AIC may perform inadequately. In those cases a second-order variant of AIC, called AICc, should be used:

$$AICc_j = -2\ln(L_j(\hat{\theta})) + 2p_j\frac{n}{n-p_j-1}$$

Indicatively, AICc should be used when $n/p < 40$. The application of the AICc method is performed by:

```
> MSC <- MSClaio2008(sitedata[,4], crit="AICc")
> MSC

------------------------
Corrected Akaike Information Criterion (AICc):
  NORM      LN  GUMBEL     EV2     GEV      P3     LP3
 401.0   388.6   386.8   383.9   386.2   382.2   385.4

> summary(MSC)

Tested distributions:
[1]  NORM    LN      GUMBEL  EV2     GEV     P3      LP3
------------------------
Chosen distributions:
AICc
  P3
whose Maximum-Likelihood parameters are:
P3 parameters of x:  37.625  14.22391  1.842626
```

Bayesian Information Criterion

The Bayesian Information Criterion (BIC) for the j-th operational model reads

$$BIC_j = -2\ln(L_j(\hat{\theta})) + \ln(n)p_j$$

In practical application, after the computation of the BIC_j, for all of the operating models, one selects the model with the minimum BIC value, BIC_{min}. The application of the BIC method is performed by:

```
> MSC <- MSClaio2008(sitedata[,4], crit="BIC")
> MSC

------------------------
Bayesian Information Criterion (BIC):
  NORM      LN  GUMBEL     EV2     GEV      P3     LP3
 404.3   392.0   390.1   387.2   391.0   387.0   390.3
```

```
> summary(MSC)

Tested distributions:
[1]  NORM    LN      GUMBEL  EV2     GEV     P3      LP3
------------------------
Chosen distributions:
BIC
 P3
whose Maximum-Likelihood parameters are:
P3 parameters of x:  37.625  14.22391  1.842626
```

Anderson-Darling Criterion

The Anderson-Darling criterion has the form (see Laio et˜al., 2008; Di Baldassarre et˜al., 2008):

$$ADC_j = 0.0403 + 0.116 \left(\frac{\Delta_{AD,j} - \epsilon_j}{\beta_j} \right)^{\frac{\eta_j}{0.851}}$$

if $1.2\epsilon_j < \Delta_{AD,j}$,

$$ADC_j = \left[0.0403 + 0.116 \left(\frac{0.2\epsilon_j}{\beta_j} \right)^{\frac{\eta_j}{0.851}} \right] \frac{\Delta_{AD,j} - 0.2\epsilon_j}{\epsilon_j}$$

if $1.2\epsilon_j \geq \Delta_{AD,j}$, where $\Delta_{AD,j}$ is the discrepancy measure characterizing the criterion, the Anderson-Darling statistic:

$$\Delta_{AD,j} = -n - \frac{1}{n} \sum_{i=1}^{n} \left[(2i-1) \ln \left[G_j(x_i, \theta) \right] + (2n + 1 - 2i) \ln \left[1 - G_j(x_i, \theta) \right] \right]$$

and ϵ_j, β_j and η_j are distribution-dependent coefficients that are tabled by Laio (2004, Tables 3 and 5) for a set of seven distributions commonly employed for the frequency analysis of extreme events. In practice, after the computation of the ADC_j, for all of the operating models, one selects the model with the minimum ADC value, ADC_{min}. The application of the ADC method is performed by:

```
> MSC <- MSClaio2008(sitedata[,4], crit="ADC")
> MSC

------------------------
Anderson-Darling Criterion (ADC):
   NORM       LN   GUMBEL      EV2      GEV       P3      LP3
1.46149  0.32650  0.22244  0.02999  0.03839  0.04394  0.02894

> summary(MSC)

Tested distributions:
[1]  NORM    LN      GUMBEL  EV2     GEV     P3      LP3
------------------------
Chosen distributions:
ADC
LP3
whose Maximum-Likelihood parameters are:
P3 parameters of log(x):  3.511428  0.1452113  4.080334
```

The function `MSClaio2008` can be applied for all the distributions and all the criteria:

```
> MSC <- MSClaio2008(sitedata[,4])
> MSC

------------------------
Akaike Information Criterion (AIC):
  NORM      LN  GUMBEL     EV2     GEV      P3     LP3
 400.7   388.4   386.5   383.6   385.6   381.6   384.9
------------------------
Corrected Akaike Information Criterion (AICc):
  NORM      LN  GUMBEL     EV2     GEV      P3     LP3
 401.0   388.6   386.8   383.9   386.2   382.2   385.4
------------------------
Bayesian Information Criterion (BIC):
  NORM      LN  GUMBEL     EV2     GEV      P3     LP3
 404.3   392.0   390.1   387.2   391.0   387.0   390.3
------------------------
Anderson-Darling Criterion (ADC):
   NORM       LN   GUMBEL      EV2      GEV       P3      LP3
1.46149  0.32650  0.22244  0.02999  0.03839  0.04394  0.02894
```

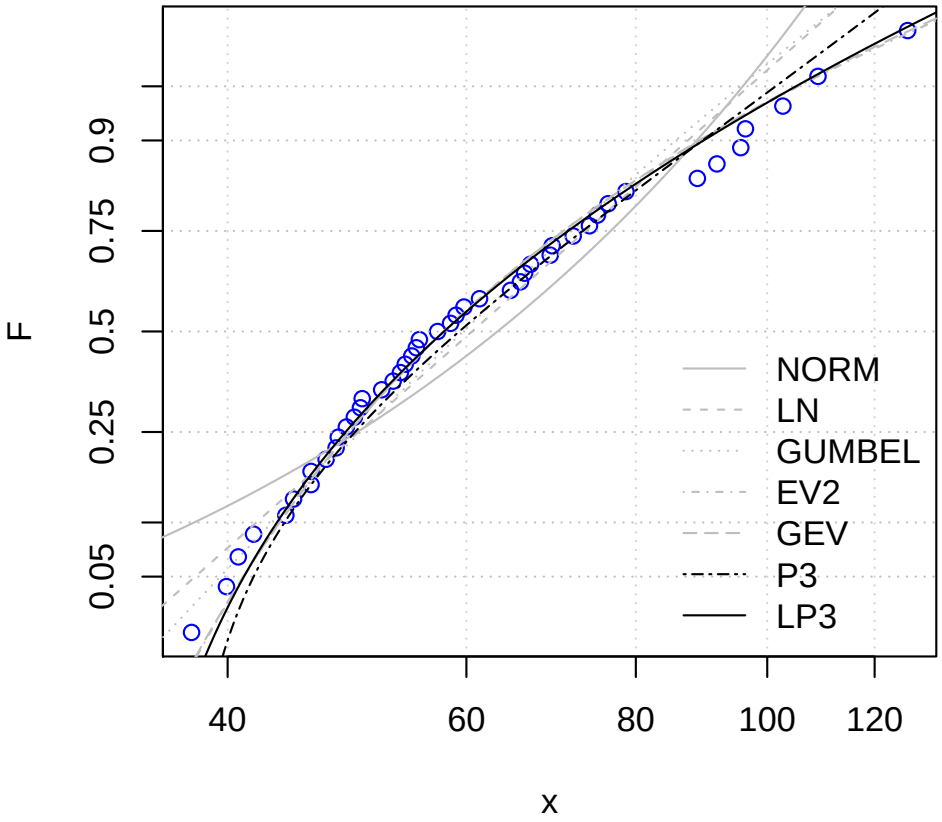
Summarizing the choices are:

```
> summary(MSC)

Tested distributions:
[1]  NORM    LN      GUMBEL  EV2     GEV     P3      LP3
------------------------
Chosen distributions:
 AIC  AICc   BIC   ADC
  P3    P3    P3   LP3
whose Maximum-Likelihood parameters are:
P3 parameters of x:  37.625  14.22391  1.842626
P3 parameters of log(x):  3.511428  0.1452113  4.080334
```

The candidate distributions and the selected ones can be plotted in a log-normal probability plot:

```
> plot(MSC)
```



Data (Weibull plotting position) and candidate distributions in lognormal probability plot. The distributions selected by one criterion, at least, are plotted in black, the others in gray.

References

Di Baldassarre, G., Laio, F., and Montanari, A. (2008). Design flood estimation using model selection criteria. Under review.

Laio, F. (2004). Cramer-von mises and anderson-darling goodness of fit tests for extreme value distributions with unknown parameters. *Water Resources Research*, 40:W09308, doi:10.1029/2004WR003204.

Laio, F., Di Baldassarre, G., and Montanari, A. (2008). Model selection techniques for the frequency analysis of hydrological extremes. Under review.