Hierarchical estimation of the extremal index

Mickey Warner

1 Abstract

This paper extends estimation of the extremal index to the Bayesian hierarchical setting based on likelihoods provided by Ferro and Segers (2003) and Süveges and Davison (2010). A comparison of the two likelihoods is made using coverage and root mean squared error (RMSE). The hierarchical model is applied to simulated data and CanCM4 climate model simulations.

2 Introduction

In extreme value analysis, a primary interest is in modeling the upper tail of a sequence of random variables $X_1, \ldots, X_n$ each with marginal distribution F. The standard approach begins by selecting a threshold u and assuming the exceedances $Y_i = X_i - u$ follow a generalized Pareto distribution having distribution function

$$H(y) = 1 - \left(1 + \xi \frac{y}{\sigma}\right)^{-1/\xi} \tag{1}$$

for $y > 0$ defined on $\{y : 1+\xi y/\sigma > 0\}$. The assumption is based on an approximation due to asymptotic theory for which at least two conditions must be satisfied: (1) the threshold u is high enough, and (2) the X_i's are independent.

When the second condition is not met (which is often the case in a time-series), inference can instead be based on clusters of exceedances. That is, because of dependence between observations, it could be expected that exceedances will arrive together in groups or clusters. Clusters can be formed by choosing a run parameter K and then grouping those exceedances which are seperated by no more than K non-exceedances. This is called runs declustering with the clusters assumed to be independent. Ferro and Segers (2003) provide a method for automatically declustering observations based on an estimate of the extremal index.

The extremal index, θ, appears in the following way (see Coles (2001)). For stationary process $X_1, X_2, \ldots$ with marginal distributions F, and $X_1^*, X_2^*, \ldots$ independent with marginal distributions F, let $M_n = \max(X_1, \ldots, X_n)$, and $M_n^* = \max(X_1^*, \ldots, X_n^*)$. Under suitable regularity conditions

$$P((M_n - b_n)/a_n \leq z) \to G_1(z)$$

as $n \to \infty$ for normalizing sequences $a_n > 0$ and b_n, where G_1 is a non-degenerate distribution function, if and only if

$$P((M_n^* - b_n)/a_n \leq z) \to G_2(z),$$

where

$$G_2(z) = G_1^{\theta}(z)$$

for $\theta \in (0, 1]$. The extremal index plays the important role of controlling the cluster size of exceedances by the following, loose, interpretion

$$\theta = (\text{limiting mean cluster size})^{-1}.$$

Therefore, if θ is known, we can form clusters such that our average cluster size is roughly θ^{-1}.

Our interest in this paper is to extend the estimators of θ provided by Ferro and Segers (2003) and Süveges and Davison (2010) to a hierarchical setting. In a time-series analysis, it is uncommon, and sometimes impossible, to have multiple realizations of a stochastic process over the same time domain. For example, we cannot go back in time, tweak a few variables, and observe new climatological data. With a computer model, there is no such constraint. Even for a deterministic model, variability can be induced by evaluating the model at a variety of input settings. Hence, we have several time-series that have their own extremal index, but it is believed that there is a commonality between them. The hierarchical model will also allow us to make inference on a climate simulation not yet observed.

Suppose we have observations $X_1, \ldots, X_n$. For a threshold u, the N exceedances $Y_i = X_i - u$ given $X_i > u$ occur at times $1 \leq j_1 < \cdots < j_N \leq n$. The observed interexceedance times are given by $T_i = j_{i+1} - j_i$ for $i = 1, \ldots, N-1$.

Estimation of the extremal index is based on the asymptotic distribution of interarrival times. Ferro and Segers (2003) produce the following log-likelihood for θ

$$\begin{aligned} l_1(\theta, p; \mathbf{T}) =& m_1 \log(1 - \theta p^{\theta}) + (N - 1 - m_1)\{\log(\theta) + \log(1 - p^{\theta})\} \\ &+ \theta \log(p) \sum_{i=1}^{N-1} (T_i - 1) \end{aligned} \tag{2}$$

where $m_1 = \sum_{i=1}^{N-1} \mathbb{1}(T_i = 1)$ and $p = F(u) = 1 - \overline{F}(u)$. Süveges and Davison (2010) define $S_i^{(K)} = \max\{T_i - K, 0\}$ and give this log-likelihood

$$l_2(\theta, p; \mathbf{S}^{(K)}) = (N - 1 - N_C) \log(1 - \theta) + 2N_C \log(\theta) - \theta(1 - p) \sum_{i=1}^{N-1} S_i^{(K)} \tag{3}$$

where $N_C = \sum_{i=1}^{N-1} \mathbb{1}(S_i^{(K)} \neq 0)$. Unless the prior for p under model (3) is very strong, the posterior for p will have undesirable properties. For all values of θ and for any threshold, $p \to 1$. We will then fix p at its empirical estimate. This model is a more general version of the one initially proposed by Süveges (2007).

3 Hierarchical model

Suppose we have R simulations from a climate model with observations from simulation i denoted $X_{i,1}, \ldots, X_{i,n}$. If we assume these simulations are indepedent from each other, then we expect there to be R unique extremal indices $\theta_1, \ldots, \theta_R$. However, since these all come from the same climate model, we may wish to assume that the θ_i come from a common distribution,

$$\theta_i \overset{iid}{\sim} Beta\left(\theta\nu, (1-\theta)\nu\right).$$

Under model (2), we place a similar prior on the p_i,

$$p_i \overset{iid}{\sim} Beta\left(p\tau, (1-p)\tau\right).$$

Since model (3) does not properly allow for the estimation of p_i, we fix each p_i to its empirical estimate at $\hat{p}_i = \sum_{j=1}^{n} \mathbb{1}(X_{i,j} \leq u)$.

The model is completed by choosing priors for θ, ν, p, and τ—the latter two parameters being required only for model (2). We assume

$$\begin{aligned}
\theta &\sim Beta(a_\theta, b_\theta) \\
\nu &\sim Gamma(a_\nu, b_\nu) \\
p &\sim Beta(a_p, b_p) \\
\tau &\sim Gamma(a_\tau, b_\tau)
\end{aligned}$$

with the hyperparameters chosen to be

θ:	$a_\theta = 1$	$b_\theta = 1/2$
ν:	$a_\nu = 1$	$b_\nu = 1/10$
p:	$a_p = 100\hat{F}$	$b_p = 100(1-\hat{F})$
τ:	$a_\tau = 1$	$b_\tau = 1/10$

where $\hat{F} = \sum_{i=1}^{R} \sum_{j=1}^{n} \mathbb{1}(X_{i,j} \leq u)$. Our parametrization for the gamma random variables are such that $X \sim Gamma(\alpha, \beta)$ has mean α/β. The prior values for θ attempt to mitigate some of the issues surrounding model (2)

By assuming indepedence between the simulations, we can construct the following log-likelihood

$$L = \sum_{i=1}^{R} l_m(\theta_i, p_i; \mathcal{D}_i) \tag{4}$$

where $m = 1, 2$ and $\mathcal{D}_i$ denotes the data from simulation i appropriate for the specified model m.

4 Simulation study

We apply the hierarchical model to simulations from a max-autoregressive process. For $i = 1, \ldots, R$, choose $\theta_i \in (0, 1]$. Let $W_{i,1} \ldots, W_{i,n}$ be independent unit Fréchet random variables, and define

$$\begin{aligned} Y_{i,1} &= W_{i,1}/\theta_i \\ Y_{i,j} &= \max\{(1-\theta_i)Y_{i,j-1}, W_{i,j}\} \qquad j = 2, \ldots, n \end{aligned} \tag{5}$$

then $Y_{i,\cdot}$ is stochastic process having extremal index θ_i. We let $R = 10$, $n = 1000$, and $\theta_1 = \cdots = \theta_R$. This construction is intended to somewhat mimic our situation with the climate model.

A single threshold is chosen based off a quantile of all simulations taken together. That is, if we wish to choose as a threshold the 0.95 quantile, then u is the solution to

$$0.95 = \sum_{i=1}^{R} \sum_{j=1}^{n} \mathbb{1}(Y_{i,j} \leq u)$$

It is certainly possible to select a threshold u_i for each sequence. While this is expected to yield better estimates for θ_i, it is not immediately clear how to apply model (1) in the hierarchical setting with different thresholds. What threshold should be used when calculating return levels for the population distribution? Or a new replicate Y^*? This, of course, leads to the broader question: If one threshold is to be used, how should it be chosen in the hierarchical setting?

Comparisons between models (2) and (3) are made using coverage and mean RMSE. With model (3) we perform the analysis for $K = 1, 5$.

Let θ_0 denote the true extremal index. R sequences of size n are simulated according to (5). The hierarchical model described in section 3 is fit. Posterior samples of θ, denoted $\theta^{(1)}, \ldots, \theta^{(M)}$, are obtained via MCMC. From these samples we calculate the 95% h.p.d. interval and RMSE,

$$\sqrt{1/M \sum_{k=1}^{M} (\theta^{(k)} - \theta_0)^2}.$$

This processes is repeated $B = 500$. Coverage is computed by finding the number of intervals containing the true value θ_0 and dividing by B. The mean RMSE is calculated as the mean of the RMSE's from the B sets of samples.

Results are shown in Figure 1. Likelihood (2) produces coverage at or near 1, which decreases toward the nominal 0.95 as threshold increases. The RMSE, on the other hand, is increasing with threshold. The pattern holds for the three θ_0's considered. These results are in agreement with those shown by Ferro and Segers (2003), Figure 2.

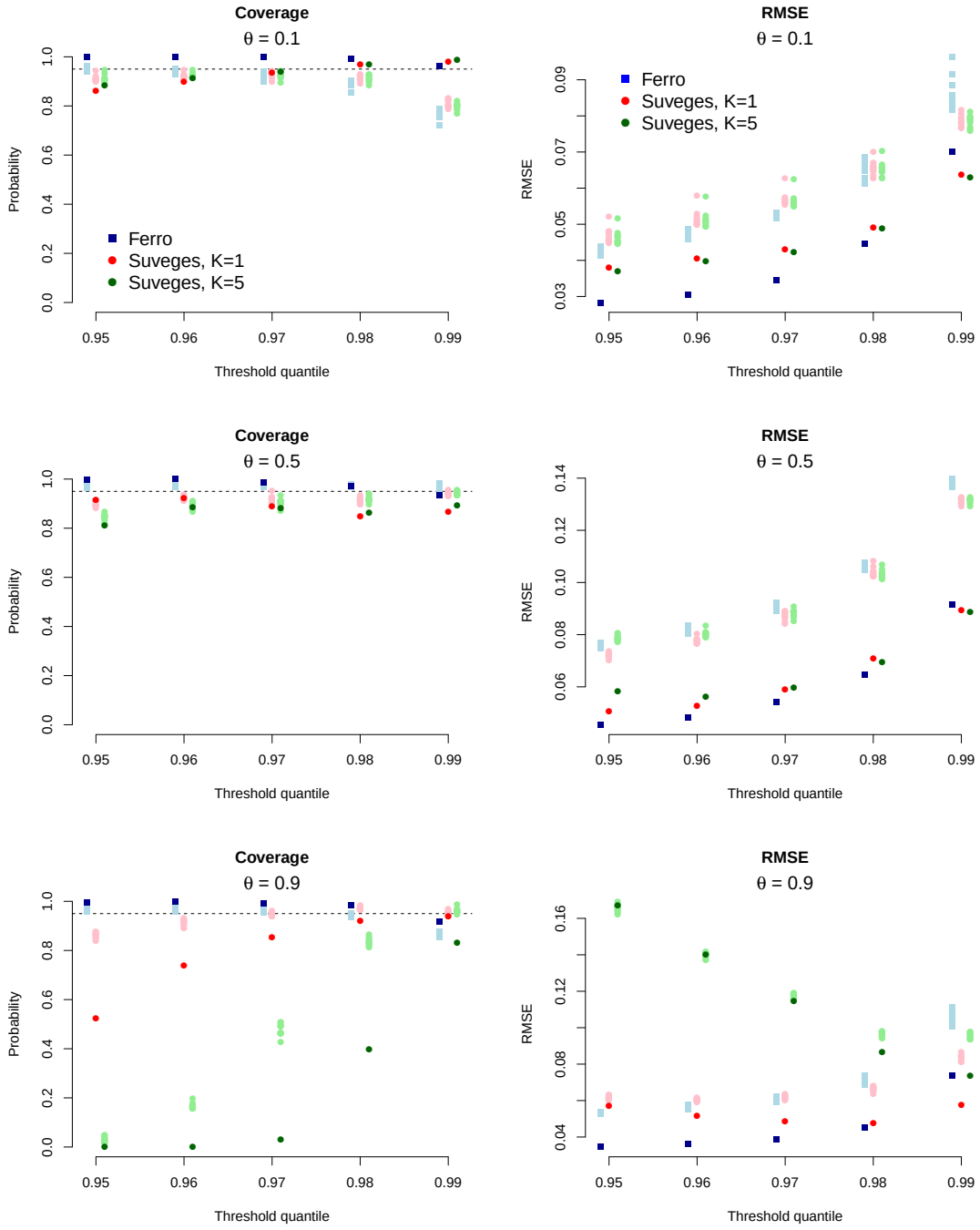

Figure 1: Coverage (left column) and mean RMSE (right) for the two likelihoods in the simulation study. Each row is based on a different extremal index, either 0.10, 0.50, or 0.90. The darker points represent the hierarchical mean (θ), the lighter points are from the individual sequences (θ_i). The nominal coverage probability (0.95) is marked by the dashed horizontal line.

In contrast, model (3) with $K = 1$ has coverage increasing with threshold toward 0.95. The RMSE is also increasing, but is greater than the RMSE for (2), except for $\theta_0 = 0.5$. For $\theta_0 = 0.9$, the RMSE does drop below that of (2) at high thresholds. (Waiting for more code to run, to include $K = 5$).

5 Climate model simulations

The Fourth Generation Coupled Global Climate Model (called CanCM4, here) produces a wide array of atmospheric conditions across the globe. Three experimental settings that are of particular interest are decadal, historical, and pre-industrial control runs. (Description of the type of simulations).

Decadal and historical simulations are run at $R = 10$ different input settings. To obtain $R = 10$ "replicates" for the control simulations, we randomly select ten non-overlapping 10-year periods. Our assumption is that each replicate is independent and has related extremal indexes within their respective simulation class.

We will look at daily winter precipitation and summer maximum temperature, both over California during the 1990s. The quantities used in the analysis are total precipitation (a weighted sum) and average maximum temperature (weighted). (EXPAND). This leads to a R univariate time series for each class of simulations. To each time-series is fit a dynamic linear model (DLM) having the first two harmonics. Anomalies are computed by taking the difference of the time-series and the smoothed predictions based on the DLM. For winter we look at only December, January, and February. For summer we have June, July, and August. Finally, we treat the time-series as though there is no gap between the seasons of interest. For example, 28 February is followed immediately by 1 December in the winter analysis. This completes our pre-processing. (After the processing, the sequences are assumed stationary).

A comparison of each simulation class and likelihood is found in Figures 2 and 3. (Not $K = 5$ yet). In every situation we see that the estimates for θ are increasing with threshold (Check: Is this happening in the simulation study?). For a given likelihood, the means for the extremal index in each simulation class are roughly the same. We do not observe this for a given class: there is clearly a discrepancy between the estimates provided by the two likelihoods. However, in all cases, the 95% probability intervals more or less overlap.

Where there is no overlap in intervals (see control runs for winter precipitation), we should consider how much of an issue this may be. Since the extremal index can be described as the reciprocal mean cluster size in the limit, there is a small practical difference between $0.63^{-1} = 1.58$ and $0.83^{-1} = 1.20$ when choosing to decluster the exceedances. The difference may be more pronounced when return levels are calculated. As θ decreases, non-overlapping posterior intervals become be much more concerning.

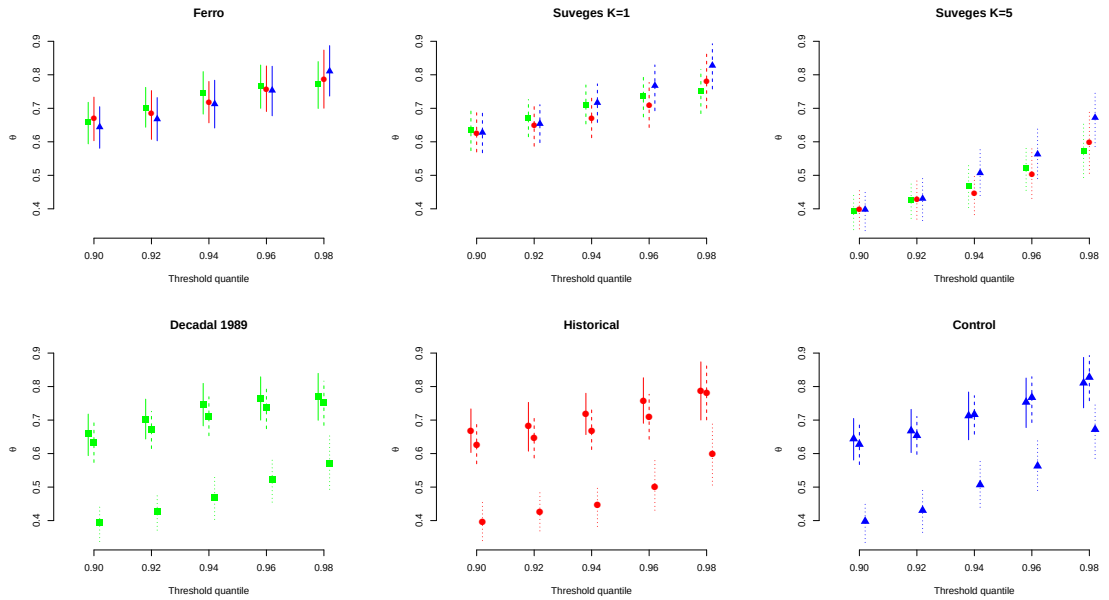

Figure 2: Both rows show the same information, but are arranged differently. The top row compares the extremal indexes of the climate models for a given likelihood. The bottom row compares the two likelihoods for each climate model. Solid lines (——) denote model M1, dashed lines (- - -) denote M2 and dotted lines (· · ·) denote M3. Squares (■) mark decadal runs, dots (●) mark historical runs, and triangles (▲) are control runs. The points are the posterior means and the lines are 95% h.p.d. intervals. The domain is California winter precipitation from 1990–1999.

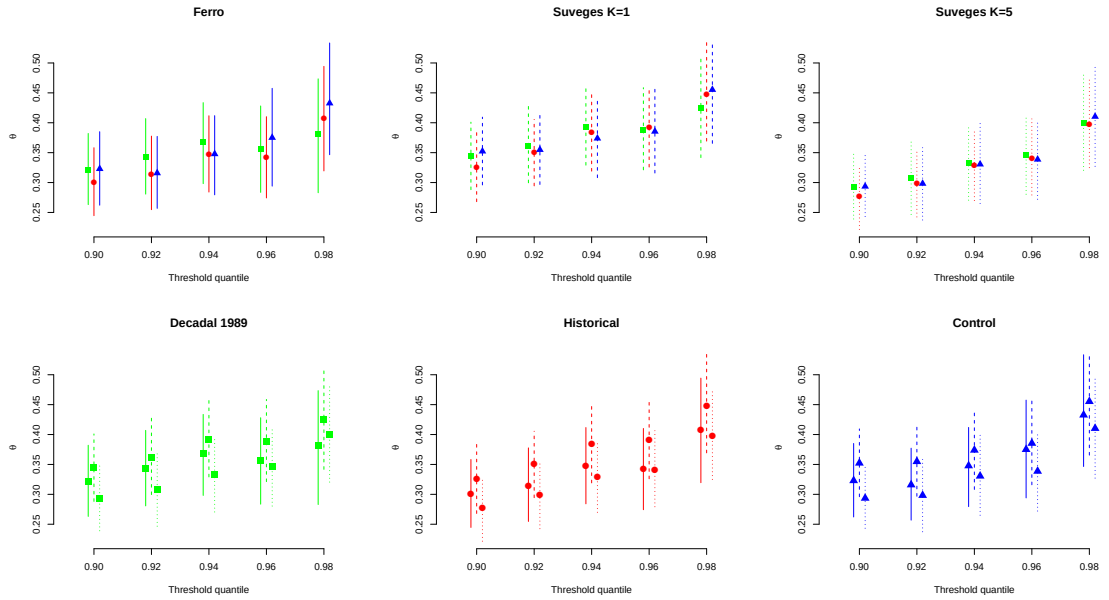

Figure 3: Same as in Figure 2, but the hierarchical model is applied to summer maximum temperature.

6 Conclusion

The simulation study seems to favor the intervals estimator (2) over the maximum likelihood estimator of Süveges and Davison (2010) for lower thresholds. This is problematic for the obvious reasons surrounding too-low thresholds. When working with small amounts of data, a lower threshold may be the only viable option.

Analysis of the CanCM4 simulations showed differences between the likelihoods in the estimates of the extremal index. There is also a trend for θ to increase with threshold (though, for model 2 this behavior was not evident in the simulations). This could suggest at least two things. First, that we need to pick a threshold sufficiently high before convergence to the true extremal index. Second, and possibly worse, that we are seeing the issues described in Ferro and Segers (2003) regarding model (2).

References

Coles, S. (2001), *An introduction to statistical modeling of extreme values*, vol. 208, Springer.

Ferro, C. A. and Segers, J. (2003), "Inference for clusters of extreme values," *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 65, 545–556.

Süveges, M. (2007), "Likelihood estimation of the extremal index," *Extremes*, 10, 41–55.

Süveges, M. and Davison, A. C. (2010), "Model misspecification in peaks over threshold analysis," *The Annals of Applied Statistics*, 4, 203–221.