# Chapter 6

# Estimating Cluster Functionals

## 6.0 Introduction

In Chapter 4 techniques for identifying clusters were discussed. Chapter 5 addressed estimation of multivariate extreme-value distributions, which can be used to model cluster componentwise maxima from multivariate sequences. Simple empirical estimators were used in Chapter 4 to estimate more general cluster features (3.8). New estimators are introduced in this chapter that attempt to exploit the structure of clusters in univariate sequences identified by Theorem 3.3. In particular, it will be seen that clusters can be decomposed into a part for which there exists an asymptotically justified model, and a part for which no such model exists. Fitting the limiting model to the appropriate component of the observed clusters and combining this with an empirical description of the second component produces an estimation scheme that exploits all known cluster structure.

## 6.1 Clusters & Strings

Let $\{X_i\}_{i=1}^n$ be a sample from a stationary process of univariate random variables and suppose that a declustering scheme has identified $C$ clusters, $\mathcal{C}_i = \{X_j : j \in \mathcal{S}_i\}$ for $1 \leq i \leq C$, of exceedances of a threshold $u$, where $\mathcal{S}_i \subseteq \{1, \ldots, n\}$. Let $K_i$ be the number of points in $\mathcal{C}_i$ and re-label the points in the $i$-th cluster as $\{X_{i,j}\}_{j=1}^{K_i}$, where $X_{i,1} \geq \ldots \geq X_{i,K_i}$ for each $i$, and order the clusters so that $X_{1,1} \geq \ldots \geq X_{C,1}$. Define also $X_{C+1,1} = u$.

Denote the value of a cluster functional $c(\cdot)$ for the $i$-th cluster by

$$CF_i = c\{(X_{i,j} - u) : 1 \leq j \leq K_i\}.$$

The standard estimator for the expectation of the cluster functional is

$$\overline{CF} = C^{-1} \sum_{i=1}^{C} CF_i. \tag{6.1}$$

The distribution of the cluster functional can be estimated by choosing $CF_i = I[c\{(X_{i,j} - u) : 1 \leq j \leq K_i\} \leq x]$ for fixed $x$.

The estimators introduced in this chapter will require the computation of the string (3.3) associated with each cluster. Recall that a string characterises the position of cluster points relative to the cluster maximum. In order to compute strings, the marginal distribution of the exceedances must be specified. If this distribution is unknown then it can be estimated by fitting to threshold excesses the generalised Pareto distribution (2.3) with scale and shape parameters $\sigma$ and $\xi$.

It is convenient to work with excesses transformed to have standard exponential marginal distribution. Let $\Psi$ be this transformation and denote the transformed excesses by $\{Z_{i,j}\}_{j=1}^{K_i}$ for $1 \leq i \leq C$, where $Z_{i,j} = \Psi(X_{i,j} - u)$. If the marginal distribution has been estimated with the generalised Pareto model then

$$Z_{i,j} = \Psi(X_{i,j} - u) = \xi^{-1} \log \left\{ 1 + \xi \left( \frac{X_{i,j} - u}{\sigma} \right) \right\}, \tag{6.2}$$

taking the limit $Z_{i,j} = (X_{i,j} - u)/\sigma$ if $\xi = 0$.

For the standard exponential excesses, the string points for the $i$-th cluster are

$$Y_{i,j} = \exp(Z_{i,1} - Z_{i,j}) \quad \text{for } 1 \leq j \leq K_i. \tag{6.3}$$

Conversely, strings can be re-attached to cluster maxima to construct clusters:

$$Z_{i,j} = Z_{i,1} - \log Y_{i,j} \quad \text{for } 1 \leq j \leq K_i.$$

A consequence of the string definition (6.3) is that strings for different clusters are observed on different intervals, depending on the values of the cluster maxima, $\{Z_{i,1}\}_{i=1}^{C}$; the $i$-th string is observed on the interval $[1, \exp(Z_{i,1}))$.

A fundamental component of the estimators will be the value of the functional for a cluster constructed by attaching a string, $\{Y_{i,j}\}_{j=1}^{K_i}$, to a cluster maximum, $Z$. Denote these values by

$$CF_i(Z) = c\{\Psi^{-1}(Z - \log Y_{i,j})_+ : 1 \leq j \leq K_i\},$$

where the subscript $i$ identifies the attached string. Reference will also be made to the random variable

$$CF(Z) = c\{\Psi^{-1}(Z - \log Y_j)_+ : 1 \leq j \leq K\}$$

where $Z$ is standard exponential and $\{Y_j\}_{j=1}^{K}$ are points of the limiting string process in Theorem 3.3.

The standard estimator (6.1) is constituted from the $C$ original clusters, which are equivalent to the clusters formed by attaching the $i$-th observed string to the $i$-th observed maximum for each $i$. Therefore, the standard estimator can be re-written as

$$\overline{CF} = C^{-1} \sum_{i=1}^{C} CF_i(Z_{i,1}).$$

The estimators introduced in the following sections assume that the asymptotic structure of clusters described by Theorem 3.3 holds over the threshold $u$, in particular, that both the strings and the maxima are independent and identically distributed and that the strings and maxima are independent of one another. Given this structure, it is reasonable to assume that each observed string could have occurred with any of the observed cluster maxima. This idea motivates the SWOP (Strings With Observed Peaks) estimator described in Section 6.2. The SWAP (Strings With Asymptotic Peaks) estimator of Section 6.3 is an extension that exploits knowledge of the distribution of the cluster maxima. The SWOP and SWAP estimators are compared with the standard estimator in a simulation study in Section 6.6 and their use is demonstrated for the Wooster temperature series in Section 6.7.

## 6.2 SWOP Estimators

The limiting independence of strings and maxima motivates the construction of $C^2$ clusters by attaching each observed string to each observed cluster maximum; $C$ of these clusters will be the original, observed clusters. The cluster functional can be evaluated for each of the new clusters and combined to produce estimators such as

$$C^{-2} \sum_{i=1}^{C} \sum_{j=1}^{C} CF_j(Z_{i,1}) \tag{6.4}$$

for the expectation of the functional. It is hoped that constructing these additional clusters will result in better estimates. In effect, for each cluster maximum, not only the cluster shape that happened to coincide with it in the sample is observed, but also other cluster shapes that were not realised.

It was mentioned previously that the $i$-th string is observed only on the interval $[1, \exp(Z_{i,1}))$. Therefore, the cluster obtained from attaching the $i$-th string to a cluster maximum, $Z$, is defined on the interval $(Z - Z_{i,1}, Z]$. If the left-hand end of this interval is above zero, that is if $Z_{i,1}$ is less than $Z$, then the cluster will be truncated above the threshold. In effect, there will be an interval above zero in which the cluster is unobserved and the value of any cluster functional will be biased. For this reason it is inappropriate to attach the $i$-th string to cluster maxima greater than $Z_{i,1}$, and so estimator (6.4) is not valid. Discarding the truncated clusters leaves $C(C+1)/2$ clusters, yielding functional values $CF_j(Z_{i,1})$ for $1 \leq j \leq i \leq C$.

An estimator for the expectation of $CF(Z)$ conditional on $Z = Z_{i,1}$ is

$$i^{-1} \sum_{j=1}^{i} CF_j(Z_{i,1})$$

and so an estimator for the unconditional expectation of $CF(Z)$ is

$$\widehat{CF} = C^{-1} \sum_{i=1}^{C} i^{-1} \sum_{j=1}^{i} CF_j(Z_{i,1}),$$

which will be called the SWOP estimator. It is difficult to obtain analytically properties of the SWOP estimator because the string point process would need to be handled explicitly. Instead, its performance is evaluated in a simulation study in Section 6.6.

## 6.3 SWAP Estimators

The SWOP estimator of the previous section averages over only the observed (transformed) cluster maxima. In order to define strings, however, the marginal distribution must be known, or estimated with a generalised Pareto model. It seems wasteful to revert to the empirical marginal distribution once the strings have been computed. In this section, the assumption that the threshold excesses of the transformed cluster maxima are exponentially distributed is exploited. Instead of averaging over just the observed maxima, it is then possible to integrate over all possible cluster maxima above the threshold. This suggests the estimator

$$\int_0^\infty \left\{ C^{-1} \sum_{j=1}^{C} CF_j(z) \right\} e^{-z}\,\mathrm{d}z \tag{6.5}$$

for the expectation of $CF(Z)$.

Recall that attaching the $i$-th string to a maximum greater than $Z_{i,1}$ results in a truncated cluster and a biased value of the cluster functional. For the SWOP estimators, this was overcome by attaching to maxima only those strings that did not lead to truncated clusters; for each observed maximum there was at least one such string. The issue is more pressing here because of the integration over all maxima on $(0, \infty)$: on $(Z_{1,1}, \infty)$, all strings yield truncated clusters. A simple modification of the naive estimator (6.5) is not available.

The expectation of $CF(Z)$ can be decomposed as

$$E\{CF(Z)\} = E\{CF(Z) \mid Z \leq Z_{1,1}\}(1 - e^{-Z_{1,1}}) + E\{CF(Z) \mid Z > Z_{1,1}\}e^{-Z_{1,1}}.$$

The data provide an estimator (akin to the SWOP estimator of the previous section) for the first term:

$$\tilde{E}\{CF(Z) \mid Z \leq Z_{1,1}\}(1 - e^{-Z_{1,1}}) = \sum_{i=1}^{C} \int_{Z_{i+1,1}}^{Z_{i,1}} \left\{ i^{-1} \sum_{j=1}^{i} CF_j(z) \right\} e^{-z}\,\mathrm{d}z. \tag{6.6}$$

The difficulty lies in estimating the conditional expectation

$$E\{CF(Z) \mid Z > z\} \tag{6.7}$$

when $z = Z_{1,1}$. Any information on this quantity is incomplete in the sense that attaching any string to a cluster maximum $Z > Z_{1,1}$ will produce a cluster truncated above the threshold. Parametric modelling of the string point process is used to solve this problem. The conditional expectation (6.7) can then be approximated using Monte Carlo simulation or, in some cases, evaluated analytically. Another strategy that models the conditional expectation itself, as a function of $z$, will also be mentioned. This latter approach is less general because the model must be tailored to the functional of interest. Simulation studies have shown that the two strategies yield very similar results, so attention will be focussed on modelling the string process.

### 6.3.1 Modelling the string process

In applications, knowledge of how the string process behaves at unobserved levels could be available from an understanding of the mechanism generating the sequence. In this section, it is shown how such a model can be used to define a cluster-functional estimator. The idea is to choose a parametric model for the string process, fit it to the observed strings and assume that the tail of the model is a reasonable approximation to the unobserved string behaviour. Such a model enables simulation of clusters with maximum greater than $Z_{1,1}$ and so estimation of the expectation (6.7).

There is no theory to guide model choice in general, so the selection of the process with atom at $\{1\}$ and which is Poisson on $(1, \infty)$ with intensity $\lambda(\cdot)$ is arbitrary. The intensity function is chosen to decay according to a power law,

$$\lambda(y) = \beta y^{-\lambda} \quad \text{for } \lambda > 0 \text{ and } \beta \geq 0, \tag{6.8}$$

since this guards against the expectation (6.7) being infinite and yields tractable expressions for the expectation for some functionals. Denote the intensity measure on $[a, b) \subset (1, \infty)$ by

$$\Lambda[a,b) = \int_a^b \lambda(y)\,\mathrm{d}y = \begin{cases} \beta(1-\lambda)^{-1}\left(b^{1-\lambda} - a^{1-\lambda}\right) & \text{if } \lambda \neq 1, \\ \beta(\log b - \log a) & \text{if } \lambda = 1. \end{cases}$$

From a fit of the string model, an estimate of the expectation (6.7) can be obtained in the following way.

1. Generate a cluster maximum, $Z > Z_{1,1}$, by setting $Z = Z_{1,1} - \log U$, where $U$ is a uniform random variate on $(0, 1)$.

2. Generate the number, $K_0$, of model string points on $[\exp(Z_{1,1}), \exp(Z))$ from a Poisson distribution with mean $\Lambda[\exp(Z_{1,1}), \exp(Z))$.

3. Generate the $K_0$ string points, $\{Y_{0,k}\}_{k=1}^{K_0}$, from the model by setting

$$Y_{0,k} = \begin{cases} \left\{U_k e^{(1-\lambda)Z} + (1-U_k)e^{(1-\lambda)Z_{1,1}}\right\}^{1/(1-\lambda)} & \text{if } \lambda \neq 1, \\ \exp\{U_k Z + (1-U_k)Z_{1,1}\} & \text{if } \lambda = 1, \end{cases}$$

   where $U_1, \ldots, U_{K_0}$ are independent, uniform random variates on $(0,1)$.

4. Evaluate $CF(Z)$ for the string comprising the points $\{Y_{1,k}\}_{k=1}^{K_1}$ and $\{Y_{0,k}\}_{k=1}^{K_0}$.

5. Repeat these steps $N$ times to obtain $N$ values, $\{CF(Z_i)\}_{i=1}^{N}$, and the Monte Carlo estimator

$$\tilde{E}\{CF(Z) \mid Z > Z_{1,1}\}e^{-Z_{1,1}} = N^{-1}\sum_{i=1}^{N} CF(Z_i).$$

This computationally intensive procedure can sometimes be avoided if the cluster functional is additive. In that case the expectation of $CF(z)$ can be estimated by

$$\tilde{E}\{CF(z)\} = CF_1(z) + E\{CF_0(z)\},$$

where $CF_0(z)$ represents the value of the cluster functional obtained from attaching string points simulated from the fitted model on $[\exp(Z_{1,1}), \exp(z))$. For some functionals, the distribution of $CF_0(z)$ is tractable and it is possible to compute

$$\tilde{E}\{CF(Z) \mid Z > Z_{1,1}\}e^{-Z_{1,1}} = \int_{Z_{1,1}}^{\infty} \tilde{E}\{CF(z)\}e^{-z}\,\mathrm{d}z$$

analytically. Some examples for which this is the case are given later. Whichever method is used to estimate the conditional expectation (6.7), combining it with the estimator (6.6) yields the SWAP estimator,

$$\widetilde{CF} = \sum_{i=1}^{C} \int_{Z_{i+1,1}}^{Z_{i,1}} \left\{ i^{-1} \sum_{j=1}^{i} CF_j(z) \right\} e^{-z}\,\mathrm{d}z + \tilde{E}\{CF(Z) \mid Z > Z_{1,1}\} e^{-Z_{1,1}}, \quad (6.9)$$

for the expectation of $CF(Z)$.

It remains to fit the string model with intensity (6.8) to the $C$ observed strings, which will be done by maximum likelihood. Discounting the atoms at $\{1\}$, the $i$-th string is observed on the interval $(1, \exp(Z_{i,1}))$ and the Poisson process likelihood is

$$\prod_{i=1}^{C} \exp\left\{-\Lambda\left(1, e^{Z_{i,1}}\right)\right\} \prod_{j=2}^{K_i} \lambda(Y_{i,j}).$$

Maximising the likelihood is a one-dimensional problem since the estimators satisfy the relationship

$$\beta = \begin{cases} (1-\lambda) \sum_{i=1}^{C} (K_i - 1) \Big/ \sum_{i=1}^{C} \{e^{(1-\lambda)Z_{i,1}} - 1\} & \text{if } \lambda \neq 1, \\ \sum_{i=1}^{C} (K_i - 1) \Big/ \sum_{i=1}^{C} Z_{i,1} & \text{if } \lambda = 1. \end{cases}$$

### 6.3.2 Modelling the expectation

In the previous section an estimator for the expectation (6.7) when $z = Z_{1,1}$ was obtained by modelling the string process. Here, an alternative approach is taken that models the expectation itself as a function of $z$. Note first that the expectation can be estimated when $z = Z_{k,1}$, $2 \leq k \leq C+1$, using the observed clusters:

$$\tilde{E}\{CF(Z) \mid Z > Z_{k,1}\} = (k-1)^{-1} \sum_{i=1}^{k-1} CF_i. \tag{6.10}$$

To estimate the expectation at $z = Z_{1,1}$ for a particular functional, a parametric model for $E\{CF(Z) \mid Z > z\}$ will be chosen, fitted to the $C$ estimates (6.10) and then evaluated at $z = Z_{1,1}$. This approach will be illustrated for when the functional is cluster size, that is

$$CF_i(Z) = \sum_{j=1}^{K_i} I(Z - \log Y_{i,j} > 0),$$

and a suitable parametric model will be identified by examining the behaviour of the expectation for some simple processes.

Let $\{Z_i\}_{i\geq 1}$ be a process with standard exponential marginal distribution and define the conditional cluster-size distribution to be

$$\pi_n(j\,;\,z) = P\left\{ \sum_{i=1}^{r_n} I(Z_i > u_n) = j \,\middle|\, \max_{1\leq i\leq r_n}(Z_i) > u_n + z \right\}$$

for a threshold $u_n$, block size $r_n = o(n)$ and a constant $z \geq 0$. This is the cluster-size distribution conditional on the cluster maximum being bounded below by $u_n + z$. If

$$\pi(j\,;\,z) = \lim_{n\to\infty} \pi_n(j\,;\,z)$$

then $E\{CF(Z) \mid Z > z\} = \sum_{j\geq 1} j\pi(j\,;\,z)$, which will now be computed for a moving-maximum process (3.9).

Let $X_i = \max_{j\geq 1}(\alpha_j W_{i-j})$, where each $\alpha_j \geq 0$, $\sum_{i\geq 1}\alpha_i = 1$ and the $W_i$ are independent, standard Fréchet innovations. Denote by $\alpha_{(j)}$ the $j$-th largest of the $\alpha_j$, let $Z_i = -\log\{1-\exp(-1/X_i)\}$ be the transformed process with standard exponential margin and set $u_n = -\log\{1-\exp(-1/n)\}$. For large $n$, clusters tend to be formed from a single innovation, say $W_0$, giving

$$\begin{aligned}
\pi_n(j\,;\,z) &\sim P\left\{\sum_{i=1}^{r_n} I(\alpha_i W_0 > n) = j \,\middle|\, \max_{1\leq i\leq r_n}(\alpha_i W_0) > ne^z\right\} \\
&\sim P\left\{\frac{n}{\alpha_{(j)}} < W_0 \leq \frac{n}{\alpha_{(j+1)}} \,\middle|\, W_0 > \frac{ne^z}{\alpha_{(1)}}\right\} \\
&= \frac{\exp\{-\alpha_{(j+1)}/n\} - \exp[-1/\max\{n/\alpha_{(j)}, ne^z/\alpha_{(1)}\}]}{1-\exp\{-\alpha_{(1)}/(ne^z)\}}\, I\left\{e^z < \frac{\alpha_{(1)}}{\alpha_{(j+1)}}\right\}.
\end{aligned}$$

The limit is

$$\pi(j\,;\,z) = \frac{\min\{\alpha_{(j)}e^z, \alpha_{(1)}\} - \alpha_{(j+1)}e^z}{\alpha_{(1)}}\, I\left\{e^z < \frac{\alpha_{(1)}}{\alpha_{(j+1)}}\right\}.$$

Note that $\pi(\cdot\,;\,z) \to \pi(\cdot)$, the unconditional cluster-size distribution, as $z$ tends to zero.

Now let $k = k(z)$ satisfy

$$\alpha_{(1)}/\alpha_{(k)} < e^z \leq \alpha_{(1)}/\alpha_{(k+1)}. \tag{6.11}$$

Then, the conditional expected cluster size is

$$E\{CF(Z) \mid Z > z\} = \sum_{j \geq 1} j\pi(j\,;\,z) = k + \frac{e^z}{\alpha_{(1)}}\left\{1 - \sum_{j=1}^{k} \alpha_{(j)}\right\}.$$

This shows how the conditional expectation depends on $z$. Unfortunately, the moving-maximum process is too general to suggest a useful parametric model. Consider instead the special case of the max-autoregressive process, for which $\alpha_{(j)} = (1-\alpha)\alpha^{j-1}$. In this case, condition (6.11) becomes

$$\alpha^{1-k} < e^z \leq \alpha^{-k} \iff -z/\log\alpha \leq k < 1 - z/\log\alpha,$$

and if $\exp(z) = \alpha^{-k}$ then

$$E\{CF(Z) \mid Z > z\} = k + (1-\alpha)^{-1}\alpha^k e^z = (1-\alpha)^{-1} - z(\log\alpha)^{-1}.$$

This suggests the linear model

$$E\{CF(Z) \mid Z > z\} = \gamma + \beta z, \qquad \beta \geq 0 \text{ and } \gamma \geq 1. \tag{6.12}$$

Given the sample values (6.10), least-squares estimates of $\beta$ and $\gamma$ can be obtained by minimising

$$\sum_{k=2}^{C+1}\left[\tilde{E}\{CF(Z) \mid Z > Z_{k,1}\} - \gamma - \beta Z_{k,1}\right]^2.$$

A model has been proposed for the conditional expectation in light of results for the max-autoregressive process. To assess its wider applicability, one-thousand clusters with maximum over $z$, for each $z = 0, 1, \ldots, 10$, were generated from a Markov chain with logistic dependence structure. Plotting the mean cluster size against $z$ produced a near-linear curve (not shown): surprisingly, the model (6.12) is an excellent fit.

Let $\hat{\gamma} + \hat{\beta} Z_{1,1}$ be the estimator for the conditional expectation (6.7). Combining this with the estimator (6.6) produces an estimator for the expected cluster size:

$$\widetilde{CF} = \sum_{i=1}^{C} \int_{Z_{i+1,1}}^{Z_{i,1}} \left\{ i^{-1} \sum_{j=1}^{i} CF_j(z) \right\} e^{-z} \, \mathrm{d}z + (\hat{\gamma} + \hat{\beta} Z_{1,1}) e^{-Z_{1,1}}. \tag{6.13}$$

The linear model (6.12) seems to be a reasonable choice for expected cluster size. For other cluster functionals, different models may be required. The simulation study described later in this chapter was conducted for estimator (6.13) and its performance was found to be very similar to the SWAP estimator (6.9). Given the greater generality of the SWAP estimator, the method described in this section is not pursued any further.

## 6.4 Multiple Thresholds

Inferences about cluster functionals may be required at a range of thresholds $u$. For the SWOP and SWAP estimators, this means that estimates of the marginal distribution and, for SWAP, the string model must be obtained at each threshold. This could be achieved simply by re-estimating the generalised Pareto and string models at each threshold. More stable estimates are likely to be obtained, however, if the marginal transformation (6.2) is made at the lowest threshold over which the generalised Pareto model is believed to hold. If this threshold is $u_0$ then the parameters $(\sigma_0, \xi_0)$ can be estimated from the excesses over $u_0$. For thresholds $u_1 > u_0$, the generalised Pareto model still holds but with parameters $\xi_1 = \xi_0$ and $\sigma_1 = \sigma_0 + \xi_0(u_1 - u_0)$. Therefore, the parameters need to be estimated at only $u_0$, after which the parameters at all higher thresholds are determined.

A similar property holds for the string model. For clusters identified over the threshold $u_1$, the $j$-th string point of the $i$-th cluster is

$$\begin{aligned} Y_{i,j}^1 &= \left\{ \frac{1 + \xi_1(X_{i,1} - u_1)/\sigma_1}{1 + \xi_1(X_{i,j} - u_1)/\sigma_1} \right\}^{1/\xi_1} \\ &= \left\{ \frac{\sigma_1 + \xi_1(X_{i,1} - u_1)}{\sigma_1 + \xi_1(X_{i,j} - u_1)} \right\}^{1/\xi_1} \\ &= \left\{ \frac{\sigma_0 + \xi_0(X_{i,1} - u_0)}{\sigma_0 + \xi_0(X_{i,j} - u_0)} \right\}^{1/\xi_0} \\ &= Y_{i,j}^0. \end{aligned}$$

So, the string elements are unaffected by changing threshold; only the interval over which strings are observed is changed. This means that, at sufficiently high levels, the parameters $(\beta, \lambda)$ of the string model are expected to be the same at all thresholds. As for the parameters of the generalised Pareto distribution, therefore, the parameters of the string model can be estimated at one threshold and fixed at these values for all higher thresholds.

In summary, the cluster-functional estimators are implemented in the following way. First, set $u_0$ to be the lowest threshold above which the extremal behaviour of the process $\{X_i\}_{i\geq 1}$ is believed to be stable. Apply a declustering scheme at this threshold and identify the clusters $\{X_{i,j}\}_{j=1}^{K_i}$ for $1 \leq i \leq C$. Standard estimators can now be evaluated. Next, fit the generalised Pareto distribution to the excesses of $u_0$, obtaining parameter estimates $(\sigma_0, \xi_0)$. Transform the clusters to have standard exponential margins and determine the strings $\{Y_{i,j}\}_{j=1}^{K_i}$ for $1 \leq i \leq C$. SWOP estimators can now be evaluated. Fit the string model to these strings, obtaining parameter estimates $(\beta, \lambda)$. SWAP estimators can now be evaluated. For estimates at thresholds $u_1$ greater than $u_0$, re-apply the declustering scheme, use the parameters of the generalised Pareto distribution implied by the estimates at $u_0$ and, for the string model, use the parameter estimates at $u_0$. In the simulation study of Section 6.6, estimates obtained by re-estimating the generalised Pareto distribution at each threshold will also be considered.

## 6.5 Examples

In this section the forms of the standard, SWOP and SWAP estimators are made explicit for two cluster functionals: the cluster size and cluster excess (3.8).

### 6.5.1 Cluster size

The functional for cluster size has value $CF_i = K_i$ for the $i$-th cluster. Estimators will be reported for the expected cluster size, $1/\theta$, and the mass function, $\pi$, of the cluster-size distribution.

The standard estimators are

$$\begin{aligned}\bar{\theta}^{-1} &= C^{-1}\sum_{i=1}^{C} K_i,\\ \bar{\pi}(k) &= C^{-1}\sum_{i=1}^{C} I(K_i = k).\end{aligned}$$

Since cluster size is invariant to the scale on which it is computed, the SWOP estimators are

$$\begin{aligned}\hat{\theta}^{-1} &= C^{-1}\sum_{i=1}^{C} i^{-1}\sum_{j=1}^{i} K_{i,j},\\ \hat{\pi}(k) &= C^{-1}\sum_{i=1}^{C} i^{-1}\sum_{j=1}^{i} I(K_{i,j} = k),\end{aligned}$$

where $K_{i,j} = \sum_{k=1}^{K_j} I(Z_{i,1} - \log Y_{j,k} > 0)$.

Cluster size is an additive functional and the SWAP estimators can be written in closed form. First note that, for $z > Z_{1,1}$, the value, $CF_0(z)$, of the cluster functional obtained from attaching string points simulated from the string model on $[\exp(Z_{1,1}), \exp(z))$ is a Poisson random variable with mean

$$E\{CF_0(z)\} = \Lambda[e^{Z_{1,1}}, e^{z}) = \beta(1-\lambda)^{-1}\left\{e^{(1-\lambda)z} - e^{(1-\lambda)Z_{1,1}}\right\},$$

taking the limit, $\beta(z - Z_{1,1})$, when $\lambda = 1$. Now, $CF_1(z) = K_1$ for $z \geq Z_{1,1}$ and so

$$\int_{Z_{1,1}}^{\infty} \tilde{E}\{CF(z)\}e^{-z}\,\mathrm{d}z = K_1 e^{-Z_{1,1}} + \beta\lambda^{-1}e^{-\lambda Z_{1,1}}.$$

Furthermore,

$$\begin{aligned}
&\sum_{i=1}^{C}\int_{Z_{i+1,1}}^{Z_{i,1}} i^{-1}\sum_{j=1}^{i} CF_j(z)e^{-z}\,\mathrm{d}z \\
&\quad = \sum_{i=1}^{C} i^{-1}\sum_{j=1}^{i}\sum_{k=1}^{K_j}\int_{Z_{i+1,1}}^{Z_{i,1}} I(z-\log Y_{j,k}>0)e^{-z}\,\mathrm{d}z \\
&\quad = \sum_{i=1}^{C} i^{-1}\sum_{j=1}^{i}\sum_{k=1}^{K_j}\int_{m_{i+1,j,k}}^{m_{i,j,k}} e^{-z}\,\mathrm{d}z \\
&\quad = \sum_{i=1}^{C} i^{-1}\sum_{j=1}^{i}\sum_{k=1}^{K_j}(e^{-m_{i+1,j,k}}-e^{-m_{i,j,k}}),
\end{aligned}$$

where $m_{i,j,k} = \max\{Z_{i,1}, \log Y_{j,k}\}$, and the SWAP estimator for the expected cluster size is

$$\tilde{\theta}^{-1} = \sum_{i=1}^{C} i^{-1}\sum_{j=1}^{i}\sum_{k=1}^{K_j}(e^{-m_{i+1,j,k}}-e^{-m_{i,j,k}}) + K_1 e^{-Z_{1,1}} + \beta\lambda^{-1}e^{-\lambda Z_{1,1}}.$$

For the cluster-size distribution, define $Y_{j,k} = \infty$ for $k > K_j$ and $1 \le j \le C$. Then, for positive integer $k$,

$$\begin{aligned}
&\sum_{i=1}^{C}\int_{Z_{i+1,1}}^{Z_{i,1}} i^{-1}\sum_{j=1}^{i} I\{CF_j(z)\ge k\}e^{-z}\,\mathrm{d}z \\
&\quad = \sum_{i=1}^{C} i^{-1}\sum_{j=1}^{i}\int_{Z_{i+1,1}}^{Z_{i,1}} I(z-\log Y_{j,k}>0)e^{-z}\,\mathrm{d}z \\
&\quad = \sum_{i=1}^{C} i^{-1}\sum_{j=1}^{i}\int_{m_{i+1,j,k}}^{m_{i,j,k}} e^{-z}\,\mathrm{d}z \\
&\quad = \sum_{i=1}^{C} i^{-1}\sum_{j=1}^{i}(e^{-m_{i+1,j,k}}-e^{-m_{i,j,k}}).
\end{aligned}$$

For $z > Z_{1,1}$,

$$P\{CF_1(z) + CF_0(z) = k\} = \frac{I(k \geq K_1)}{(k-K_1)!} E\{CF_0(z)\}^{k-K_1} \exp[-E\{CF_0(z)\}]$$

and so the contribution to the SWAP estimator from the string model requires numerical approximation. If cluster sizes $k < K_1$ only are of interest, however, the contribution from the string model is zero and the SWAP estimator is

$$\tilde{\pi}(k) = \sum_{i=1}^{C} i^{-1} \sum_{j=1}^{i} \left\{ \left(e^{-m_{i+1,j,k}} - e^{-m_{i,j,k}}\right) - \left(e^{-m_{i+1,j,k+1}} - e^{-m_{i,j,k+1}}\right) \right\}.$$

### 6.5.2 Cluster excess

The functional for cluster excess has value $CF_i = \sum_{j=1}^{K_i}(X_{i,j} - u)$ for the $i$-th cluster. Estimators for the expected cluster excess, $A$, and the cluster-excess distribution, $F_A$, will be reported.

The standard estimators are

$$\begin{aligned}
\bar{A} &= C^{-1} \sum_{i=1}^{C} \sum_{j=1}^{K_i} (X_{i,j} - u), \\
\bar{F}_A(a) &= C^{-1} \sum_{i=1}^{C} I \left\{ \sum_{j=1}^{K_i} (X_{i,j} - u) \leq a \right\}.
\end{aligned}$$

The SWOP estimators are

$$\hat{A} = C^{-1}\sum_{i=1}^{C} i^{-1}\sum_{j=1}^{i}\sum_{k=1}^{K_j}\Psi^{-1}(Z_{i,1}-\log Y_{j,k})_+,$$

$$\hat{F}_A(a) = C^{-1}\sum_{i=1}^{C} i^{-1}\sum_{j=1}^{i} I\left\{\sum_{k=1}^{K_j}\Psi^{-1}(Z_{i,1}-\log Y_{j,k})_+ \le a\right\}.$$

The SWAP estimators require some working. The expectation $E\{CF_0(z)\}$ is computed first for $z > Z_{1,1}$. Let $K_0(z)$ represent the number of points simulated from the string model in $[\exp(Z_{1,1}), \exp(z))$. Conditional on $K_0(z)$, the simulated points are independent with density $\beta y^{-\lambda}/\Lambda[\exp(Z_{1,1}), \exp(z))$ on $[\exp(Z_{1,1}), \exp(z))$. Let $Y$ represent an arbitrary point simulated from this density and note that $z - \log Y > 0$ due to the support of $Y$. Now, $E\{K_0(z)\} = \Lambda[\exp(Z_{1,1}), \exp(z))$ and so

$$E\{CF_0(z)\} = \Lambda[e^{Z_{1,1}}, e^z)E\left\{\Psi^{-1}(z-\log Y) - u \mid K_0(z)\right\}.$$

Some algebra shows that this is equal to

$$\frac{\beta\sigma}{\xi(1-\lambda)(1-\xi-\lambda)}\left[\xi e^{(1-\lambda)z} + \left\{(1-\xi-\lambda)-(1-\lambda)e^{\xi(z-Z_{1,1})}\right\}e^{(1-\lambda)Z_{1,1}}\right],$$

taking the appropriate limit when $\xi = 0, \lambda = 1$ or $\xi + \lambda = 1$. Integrating yields

$$\int_{Z_{1,1}}^{\infty} E\{CF_0(z)\}e^{-z}\,\mathrm{d}z = \frac{\beta\sigma e^{-\lambda Z_{1,1}}}{\lambda(1-\xi)}$$

for $\lambda > 0$ and $\xi < 1$. The integral diverges when $\xi \ge 1$, which corresponds to the generalised Pareto distribution for the cluster maximum excess having infinite mean. Also,

$$\int_{Z_{1,1}}^{\infty} CF_1(z)e^{-z}\,\mathrm{d}z = \frac{\sigma}{\xi}\sum_{k=1}^{K_1}\left\{\frac{Y_{1,k}^{-\xi}}{1-\xi}e^{-(1-\xi)Z_{1,1}} - e^{-Z_{1,1}}\right\}$$

for $z > Z_{1,1}$ and $\xi < 1$, taking the limit when $\xi = 0$.

The contribution to the estimator for $z < Z_{1,1}$ is

$$\sum_{i=1}^{C} i^{-1}\sum_{j=1}^{i}\sum_{k=1}^{K_j} I_{i,j,k},$$

where

$$\begin{aligned} I_{i,j,k} &= \int_{Z_{i+1,1}}^{Z_{i,1}} \Psi^{-1}(z-\log Y_{j,k})_+ e^{-z}\,\mathrm{d}z \\ &= \frac{\sigma}{\xi}\left[\frac{Y_{j,k}^{-\xi}}{1-\xi}\left\{e^{-(1-\xi)m_{i+1,j,k}} - e^{-(1-\xi)m_{i,j,k}}\right\} - \left(e^{-m_{i+1,j,k}} - e^{-m_{i,j,k}}\right)\right], \end{aligned}$$

taking appropriate limits when $\xi = 0$ or $\xi = 1$. Therefore, the SWAP estimator for the expected cluster excess is infinity if $\xi \geq 1$ and

$$\tilde{A} = \sum_{i=1}^{C} i^{-1}\sum_{j=1}^{i}\sum_{k=1}^{K_j} I_{i,j,k} + \frac{\sigma}{\xi}\sum_{k=1}^{K_1}\left\{\frac{Y_{1,k}^{-\xi}}{1-\xi}e^{-(1-\xi)Z_{1,1}} - e^{-Z_{1,1}}\right\} + \frac{\beta\sigma e^{-\lambda Z_{1,1}}}{\lambda(1-\xi)}$$

if $\xi < 1$ with the appropriate limit at $\xi = 0$.

For the distribution function, the contribution from $z < Z_{1,1}$ is

$$\sum_{i=1}^{C} i^{-1}\sum_{j=1}^{i} I_{i,j},$$

where

$$I_{i,j} = \int_{Z_{i+1,1}}^{Z_{i,1}} I\left\{\sum_{k=1}^{K_j} \Psi^{-1}(z - \log Y_{j,k})_+ \le a\right\} e^{-z}\,\mathrm{d}z = e^{-m^*_{i+1,j}} - e^{-m^*_{i,j}},$$

$m^*_{i,j} = \min\{Z_{i,1}, z^*_j\}$ and $z^*_j$ satisfies

$$\sum_{k=1}^{K_j} \Psi^{-1}(z^*_j - \log Y_{j,k})_+ = a.$$

If only values $a < CF_1$ are of interest then the contribution from $z > Z_{1,1}$ is zero. Otherwise, $P\{CF_0(z) \le a\}$ requires numerical approximation.

## 6.6 Simulation Study

The performances of the standard, SWOP and SWAP estimators are compared through a simulation study. Three quantities are estimated: the extremal index, as the reciprocal of mean cluster size, the probability $\pi(1)$ of a cluster having size 1, and the mean cluster excess. The estimators are applied to 100 realisations of length 5000 from three max-autoregressive processes, with extremal indices 0.25, 0.5 and 0.75. The processes are generated with standard exponential margin, in which case the mean cluster excess has a finite limit as the threshold increases. In fact, the mean cluster excess is equal to the mean cluster size since the expected excess for a single exceedance is 1. The three quantities are estimated at ten thresholds, corresponding to the empirical $p$-quantiles for $p = 0.90, 0.91, \ldots, 0.99$, and the estimators are applied in three situations: when the margin is assumed known, when the margin is estimated at the lowest threshold, and when the margin is re-estimated at each threshold. The status of knowledge about the margin does not affect the standard estimator (6.1). The parameters of the string process model in the SWAP estimator are always estimated at only the lowest threshold. Intervals declustering is implemented at each threshold to identify clusters. The performance of the estimators is summarised by the Monte Carlo estimates of root mean square error (RMSE) reproduced in Figures 6.1 – 6.3.

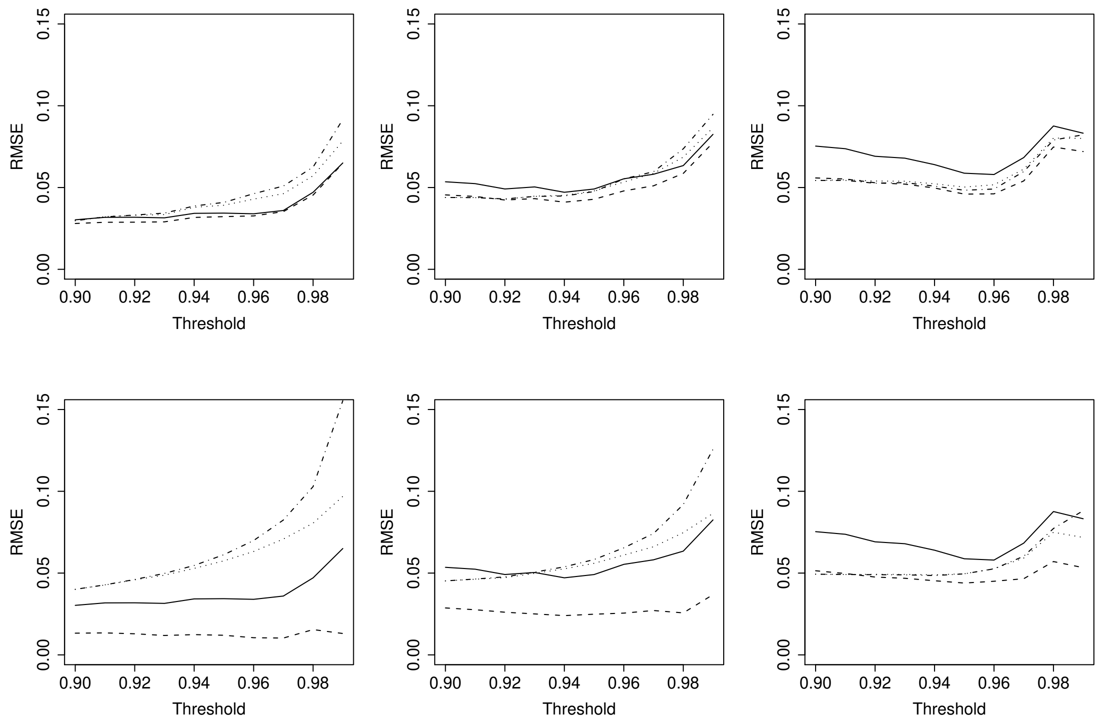


Figure 6.1: RMSE of seven estimators for the extremal index: standard (——), SWOP (above) and SWAP (below) with margin known (- - -), estimated (· · ·) and re-estimated (-·-·-). The true extremal index is 0.25 (left), 0.5 (middle) and 0.75 (right).

Consider first the relative performance of the standard and SWOP estimators, seen in the first rows of Figures 6.1 – 6.3. When the margin is known, the SWOP

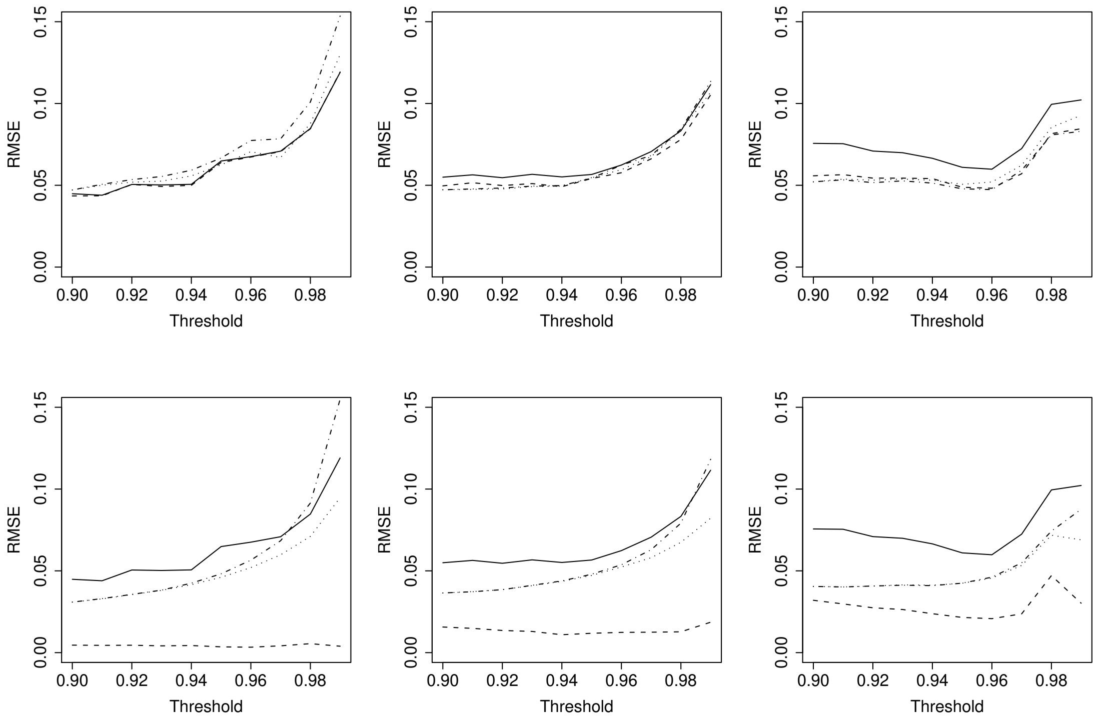


Figure 6.2: RMSE of seven estimators for $\pi(1)$: standard (——), SWOP (above) and SWAP (below) with margin known (- - -), estimated (· · ·) and re-estimated (-·-·-). The true extremal index is 0.25 (left), 0.5 (middle) and 0.75 (right).

estimator outperforms the standard estimator to a small degree in every case, and this superiority is greater when the extremal index is greater. The performance of the SWOP estimator deteriorates when the marginal distribution is estimated, but remains slightly superior to the standard estimator, at least when the extremal index is large. Bias and variance estimates (not shown) indicate that the SWOP estimators produce smaller clusters (larger extremal index and $\pi(1)$, smaller mean cluster excess) and, except when the margin is known, have greater variability than the standard estimator. Standard estimators tend to over-estimate clusters. When the margin is known, therefore, the superiority of the SWOP estimator is due to a reduction in variance and bias; when the margin is unknown, there is a

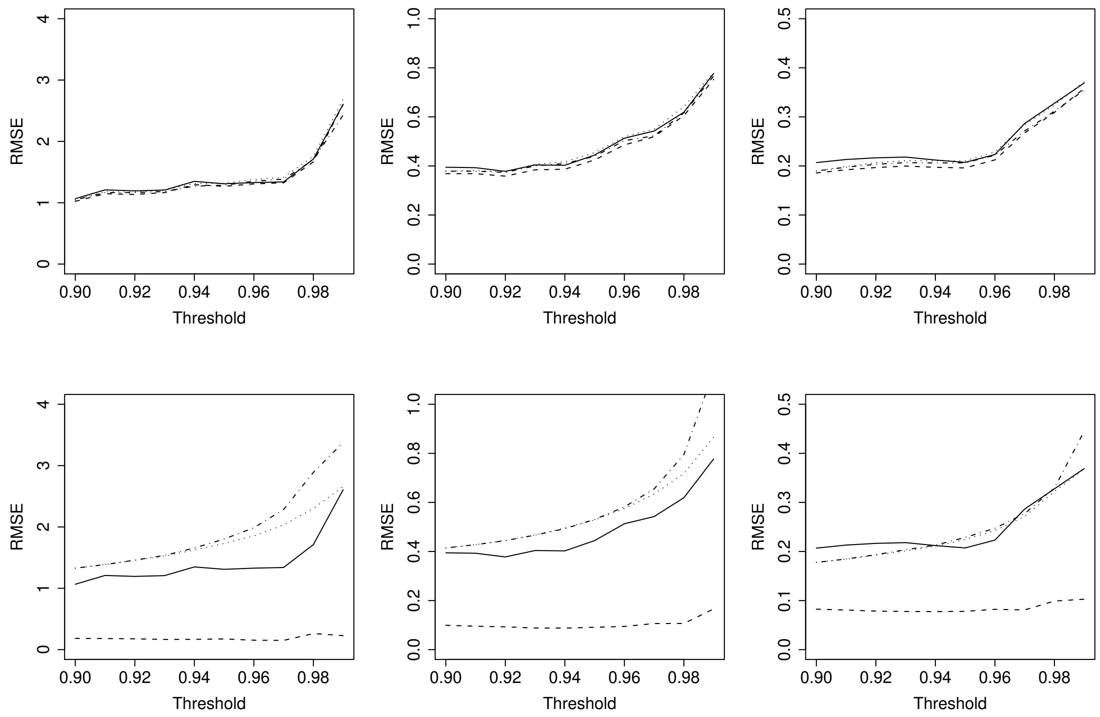


Figure 6.3: RMSE of seven estimators for the mean cluster excess: standard (——), SWOP (above) and SWAP (below) with margin known (- - -), estimated (· · ·) and re-estimated (-·-·-). The true extremal index is 0.25 (left), 0.5 (middle) and 0.75 (right).

trade-off between a reduction in bias and an increase in variance.

Consider now the performance of the SWAP estimators, seen in the second rows of Figures 6.1 – 6.3. When the margin is known, the SWAP estimator significantly outperforms the standard and SWOP estimators in every case. The performance is severely diminished when the margin is estimated, however, perhaps remaining superior only when the extremal index is large. Again, SWAP estimators produce smaller clusters than the standard estimator and, except when the margin is known, have greater variability.

The simulation study was repeated, replacing the max-autoregressive processes by

Markov chains with bivariate extreme-value logistic dependence structure (5.6). Although the extremal behaviours of the two processes are different, with the max-autoregressive clusters following a deterministic pattern and those of the Markov chains being stochastic, the results were qualitatively similar. The benefit of knowing the marginal distribution on the performance of the SWAP estimator was markedly reduced however.

When the marginal distribution is known, creating extra clusters by interchanging strings between cluster maxima has been found to reduce the bias and variance of estimators for some cluster functionals. When the margin is estimated, however, this procedure inflates the variance of the estimators, particularly when the clustering in the data is strong. Any error in the marginal estimation causes a mis-specification of the strings and, consequently, of the extra clusters. The effect is magnified for the SWAP estimator because strings are attached to all possible maxima over the threshold, not just the observed maxima, and because the strings are used to estimate the string model. The SWOP and SWAP estimators perform better when the clustering is weak because declustering becomes easier: with strong dependence, the scope for nominally independent clusters to be identified as a single cluster is greater, and all of the clusters formed from the corresponding string will be erroneous. The need for precise declustering is emphasised by the results for the Markov chains, for which declustering is more difficult compared to max-autoregressive processes.

The issues of marginal estimation and declustering appear to be of primary importance for improving the SWOP and SWAP estimators. No attempt was made to optimise these aspects in the simulation study, by careful selection of threshold for example, so this might be considered in more detail. Several methods were tried of reducing the influence of strings from clusters with large maxima, since these clusters tend to be larger and so are less likely to be declustered successfully. For example, the string from only the $i$-th cluster was used to define string points on $[\exp(Z_{i+1,1}), \exp(Z_{i,1}))$, and strings were created by drawing at random from the empirical distribution of all observed string points. Recall also that the string model for the SWAP estimator was fitted to all observed string points, but that the model is used to generate string points only above $\exp(Z_{1,1})$. The aim is to estimate just the tail of the string process, therefore, and so attempts were made to fit the model to only those string points over some threshold. None of these modifications to the creation of strings had a significant impact on the estimators.

## 6.7 Data Example

To conclude, the estimators are applied to the Wooster temperature series of Section 4.1.7. There, standard estimates of the extremal index, cluster-size distribution and the mean cluster excess were computed at several different thresholds. Here, the standard, SWOP and SWAP estimators are applied, re-estimating the parameters of the generalised Pareto and string models at each threshold. The estimates, with bootstrapped confidence intervals are displayed in Figure 6.4. The relative bias noted in the simulation is evident, with the standard estimator producing larger clusters. The variances of the standard and SWOP estimators are similar; the variance of the SWAP estimator is typically lowest. For thresholds above six degrees, the estimators are in broad agreement, with both the extremal index and $\pi(1)$ between about 0.5 and 0.7. These are similar to the values found with the Markov chain model of Smith *et al.* (1997).

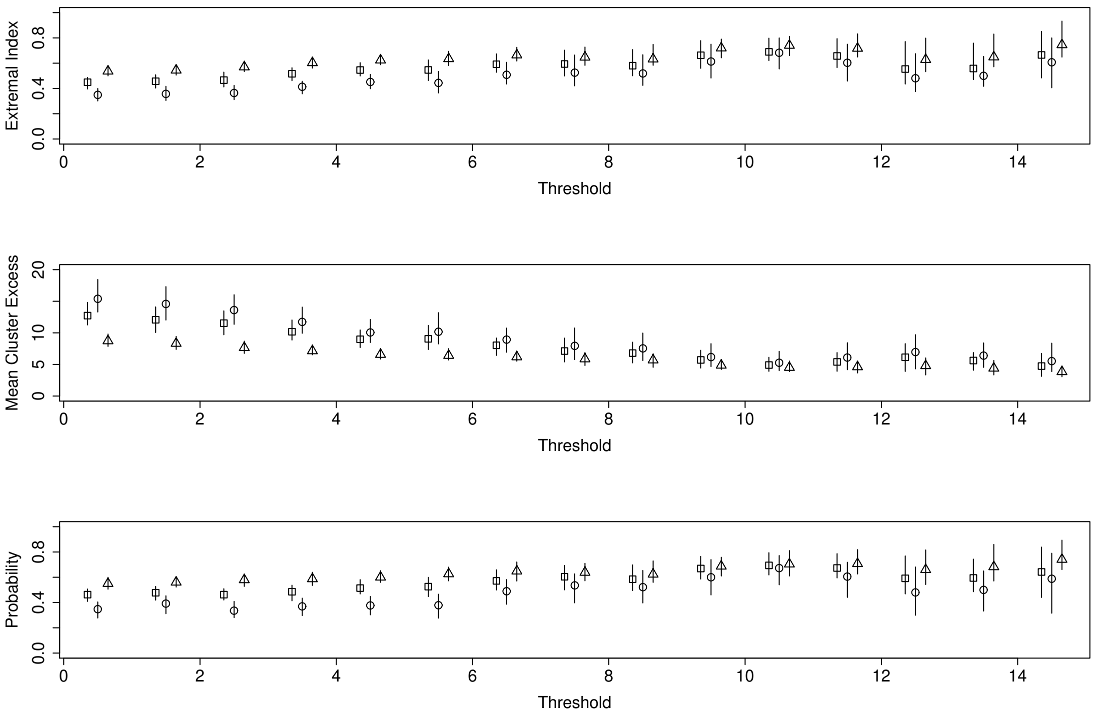


Figure 6.4: Standard (◯), SWOP (□) and SWAP (△) estimates of the extremal index (top), mean cluster excess (middle) and $\pi(1)$ (bottom) against threshold, with bootstrapped 90% confidence intervals (|).