Chapter 7

Inference for Multivariate Clusters

7.0 Introduction

In the previous chapter, new estimators for features of clusters in univariate processes were defined. These estimators decomposed clusters into maxima and strings, where a string defines the points in a cluster relative to the cluster maximum. In this chapter, the decomposition is extended to the multivariate setting. Combining the empirical description of strings with an extreme-value distribution for the maxima provides a model for the multivariate process, including its clustering behaviour, at levels outwith the data. Previous attempts at inference for multivariate processes have modelled only summaries of clusters, with estimates depending on the particular summary chosen.

One important problem in multivariate extremes is the estimation of failure probabilities, that is the probability that an observation falls in some rare region of the sample space. Such a failure region might, for example, characterise the conditions under which a sea-wall is breached, with respect to wave height, wave period, surge and tide. See Coles and Tawn (1994) for several other examples. Estimation of failure probabilities is reviewed in Section 7.1, where shortcomings of current methods are noted. The decomposition of multivariate clusters into maxima and strings is presented in Section 7.2 together with methods for fitting the model and estimating failure probabilities. A simulation study illustrating the benefits of the new approach is documented in Section 7.3 and a data example is provided in Section 7.4.

7.1 Estimating Failure Probabilities

Let $\{X_i\}_{i=1}^n$ be a stationary sequence of D-dimensional random variables $X_i = (X_{i1}, \dots, X_{iD})$ with unknown marginal distribution function F. Suppose that it is of interest to estimate the probability $P(X \in B)$ that an arbitrary observation X falls in some failure region $B \subset \mathbb{R}^D$.

Theorem 2.4 motivates a Poisson process model for points X_i on regions bounded away from the origin when those points are from an independent sequence. For stationary sequences, such a model is not applicable because extremes will form clusters, as described by Theorem 3.6. This problem can be avoided by filtering the sequence to obtain approximately independent observations. For example, Coles and Tawn (1994) partition the original data into blocks and retain only the componentwise maximum from each block, using arguments concerning the physical process generating the data to justify independence. Similarly, de Haan and Sinha (1999) retain only the componentwise maxima from subjectively identified storms.

Parametric and semi-parametric methods exist for estimating failure probabilities. Parametric methods rely on fitting the Poisson process model to the filtered data on some region bounded away from the origin, as described in Section 2.2. A model for the dependence structure, such as the logistic model (5.6) or the kernel estimator of Chapter 5, must be selected. The failure probability can be estimated by computing the probability that a point of the fitted Poisson process falls in B. This approach is developed by Coles and Tawn (1991, 1994) and Joe *et al.* (1992). A semi-parametric approach has been developed by de Haan and de Ronde (1998) and de Haan and Sinha (1999). This relies on a homogeneity property of the Poisson process measure to relate the failure probability to the probability that a point falls in a set $B_c = \{x : cx \in B\}$, where c shrinks B until it contains some of the observations. The latter probability can then be estimated by the proportion of observations in B_c.

Whether the parametric or semi-parametric approach is followed, the resulting failure probability estimate depends crucially on the filtering used to obtain independent observations. For example, consider componentwise maxima. Suppose that the original data have been partitioned into $n_b = \lfloor n/b \rfloor$ blocks of length b:

$$\{X_j : (i-1)b + 1 \leq j \leq ib\} \quad \text{for } 1 \leq i \leq n_b. \tag{7.1}$$

The block componentwise maxima, hereafter called the block maxima, are

$$X_i^* = (X_{i1}^*, \ldots, X_{iD}^*) \quad \text{for } 1 \leq i \leq n_b,$$

where $X_{id}^* = \max\{X_{jd} : (i-1)b + 1 \leq j \leq ib\}$. Note that a block maximum does not necessarily coincide with any actual observation in the block. Let the distribution of a block maximum be F^* with component distributions F_d^*. For failure regions satisfying

$$X_1 \in B \text{ or } X_2 \in B \implies (\max\{X_{11}, X_{21}\}, \ldots, \max\{X_{1D}, X_{2D}\}) \in B, \qquad (7.2)$$

the probability of a block maximum X^* falling in B is at least as large as the probability for an arbitrary observation X in the block. Basing inferences solely on the block maxima therefore leads to potential over-estimation of the failure probability.

The failure probability $P(X \in B)$ factorises as

$$P(X \in B \mid X^* \in B)P(X^* \in B) + P(X \in B \mid X^* \notin B)P(X^* \notin B), \qquad (7.3)$$

which is approximated by $P(X^* \in B)$ if only block maxima are used. The method described in the following section also begins by estimating $P(X^* \in B)$, but then estimates the correction factor $P(X \in B \mid X^* \in B)$. If B satisfies condition (7.2) then $P(X \in B \mid X^* \notin B)$ equals zero and nothing more is required. This is likely to be the case for most failure regions of interest. If condition (7.2) does not hold then estimates of the second term in factorisation (7.3) can also be obtained.

7.2 Modelling Multivariate Clusters

7.2.1 Block maxima

Let n_c of the n_b blocks (7.1) contain points in a region $A = \{x \in \mathbb{R}^D : x \not\leq u\}$ for some choice of high thresholds $u = (u_1, \ldots, u_D)$. For suitably large block length, these n_c collections of points can be considered independent clusters because of the limiting behaviour described by Theorem 3.6. This motivates the Poisson process of Theorem 2.4 as a model for block maxima in A if the component distributions F_d^*, $1 \leq d \leq D$, are standard Fréchet. If the component distributions are unknown then they can be estimated using generalised Pareto forms above thresholds and empirical distribution functions below thresholds, as detailed in Section 2.2. The transformation of the d-th component to standard Fréchet scale is $-1/\log F_d^*(\cdot)$. Define also the transformation

$$\Psi_d^*(\cdot) = -\log\{1 - F_d^*(\cdot)\} \tag{7.4}$$

of the d-th component to standard exponential scale.

7.2.2 Strings

As discussed in Section 7.1, modelling only block maxima can cause failure probabilities to be over-estimated. To obtain unbiased estimates it is necessary to retain information about the distribution of actual points within a block. For extreme failure regions, only the most extreme points in a block, such as those that fall in the region A, are relevant. Suppose again that n_c of the n_b blocks (7.1) contain points in A. Denote these n_c clusters of points by $\{X_j : j \in \mathcal{S}_i\}$, $1 \leq i \leq n_c$, where $\mathcal{S}_i$ contains the indices of the points in A for the i-th cluster. It was seen in Chapter 6 that for univariate clusters the distribution of cluster points relative to their cluster maximum is described by a string. An analogous decomposition for multivariate clusters is given below. This characterises the position of extreme points in a block relative to the block maximum and so provides the necessary information for estimating failure probabilities accurately.

Recall from Section 6.1 that the marginal distribution, F, must be specified in order to define strings associated with clusters. Let Ψ_d transform the d-th component to have standard exponential distribution: write $Z_{jd} = \Psi_d(X_{jd})$, $j \in \mathcal{S}_i$, for the transformed points of the i-th cluster and $Z^*_{id} = \Psi^*_d(X^*_{id})$ for the transformation of the block maximum X^*_i. The D-dimensional string point corresponding to observation X_j in the i-th cluster is defined to be $Y_{ij} = (Y_{ij1}, \ldots, Y_{ijD})$, where

$$Y_{ijd} = \exp(Z^*_{id} - Z_{jd});$$

the string associated with the i-th cluster is the collection $Y_i = \{Y_{ij} : j \in \mathcal{S}_i\}$.

This string definition is analogous to the univariate definition (6.3) for each component. The treatment of multivariate clusters in this way pre-empts theoretical developments: no multivariate extension of Theorem 3.3 has yet been published. The definition of strings seems natural nevertheless, and at least encompasses the structure of clusters in the M_4 process (3.11).

The choice of declustering region A means that clusters can include points below the threshold in up to $D-1$ components. In contrast to the univariate case, therefore, the transformation Ψ_d must be applicable above and below u_d. The method of Section 2.2 can be used, that is $\Psi_d(\cdot) = -\log\{1 - F_d(\cdot)\}$, where F_d is modelled with a generalised Pareto distribution above u_d and the empirical distribution function below u_d. Note that this transformation is based on all observations $\{X_i\}_{i=1}^n$, not just the block maxima that were used to estimate the transformation (7.4).

Identifying clusters by grouping observations into blocks does not accord with the theoretically motivated intervals declustering scheme proposed in Section 4.2. Intervals declustering considers points only in the extreme region A, however, and so yields only the n_c cluster maxima, not the n_b block maxima. If only the cluster maxima are available then estimation of the component distributions required for the Poisson process model is impossible without making parametric assumptions. The problem is the estimation below thresholds: empirical distributions are inappropriate because any points with one component below threshold have at least one other component above threshold; furthermore, points with components below thresholds could be sparse. For this reason, blocks declustering is used in this chapter. As a compromise, the block length b is chosen to yield the same number of clusters in A as would be identified by intervals declustering.

7.2.3 Failure probabilities

Block maxima have been modelled with a Poisson process and clusters of extreme points within blocks have been identified with strings. If a block maximum X^* is simulated from the fitted Poisson process and transformed to a block maximum $Z^* = \Psi^*(X^*)$ with standard exponential components then an observed string $Y_i = \{Y_{ij} : j \in \mathcal{S}_i\}$ can be attached to form cluster points Z_j with $Z_{jd} = Z_d^* - \log Y_{ijd}$. These points can then be transformed back to the original space by inverting Ψ^*. This enables the failure probability $P(X \in B)$ to be estimated in the following way.

First simulate a large number, n_s, of block maxima from the Poisson process model in a region B_0 containing the failure region B. This is perhaps most easily done in standard Fréchet space. Let $\tilde{B}$ be the failure region B transformed to standard Fréchet space using the transformations $-1/\log F_d^*(\cdot)$. Block maxima can then be simulated in a region $\tilde{B}_0 = \{x \in \mathbb{R}^D : x_1 + \ldots + x_D > r_0\}$ with r_0 chosen so that $\tilde{B} \subset \tilde{B}_0$ (Nadarajah, 1999). Clusters can be constructed by transforming to standard exponential space, attaching the n_c observed strings to each of the simulated block maxima and transforming back to the original space. The probability that a block maximum falls in $\tilde{B}_0$ is $1 - \exp\{-\Lambda(\tilde{B}_0)\} = 1 - \exp(-D/r_0)$, where Λ is the intensity measure of the Poisson process; see Coles and Tawn (1994). If n_m of the simulated block maxima fall inside B then a Monte Carlo estimate of $P(X^* \in B)$ is $(n_m/n_s)\{1-\exp(-D/r_0)\}$. To each maximum, n_c strings are attached, creating $n_m n_c$ blocks with block maximum in B. Nominally, there are b points in each block, but the strings only describe the extreme cluster points; the remaining points in a block are assumed to be negligible. If n_x is the total number of points in B summed over all $n_m n_c$ blocks with block maximum in B then an estimate of the correction factor $P(X \in B \mid X^* \in B)$ is $n_x/(b n_m n_c)$. If $P(X \in B \mid X^* \notin B)$ is non-zero then a similar Monte Carlo approximation can be found for the second term in the factorisation (7.3).

In Chapter 6, it was noted that attaching strings to maxima outside the range of the data leads to a truncation of clusters above the threshold. The same issue arises in the multivariate case and can cause under-estimation of the correction factor. One remedy is to model strings in censored regions, as was done for the SWAP estimator in Section 6.3, but this will not be attempted here. Instead, the performance of the method will be investigated with a simulation study, ignoring the issue of censoring.

7.3 Simulation Study

To illustrate the practical effect of estimating the correction factor, the model of the previous section is applied to data simulated from five M_4 processes (3.11).

Bivariate processes are used, with form

$$X_{id} = \max_{1 \le l \le L} \max_{1 \le j \le J} (\alpha_{ljd} Z_{l,i-j}) \quad \text{for } i \ge 1 \text{ and } d = 1, 2,$$

where the $Z_{l,i}$ are independent, standard Fréchet random variables. The coefficients are constrained to ensure that the component distributions of the X_i are also standard Fréchet, and $L = 1$ so that limiting clusters have only a single deterministic shape. The coefficients of the five processes are listed in Table 7.1 together with a sketch of the limiting cluster shape. The coefficients are such that cluster maxima tend to lie on the line $x_1 = x_2$ and are not members of clusters, except for process 1 where clusters comprise just single points. Processes 2 and 3 have two points per cluster, each of which contributes a coordinate to the cluster maximum. The distance between the cluster points and the cluster maximum is greater in process 3. Processes 4 and 5 are the same as processes 2 and 3 except that clusters have an additional point that does not contribute to the maximum.

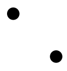
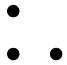
Process	J	$\{\alpha_{1j1}\}_{j=1}^{J}$	$\{\alpha_{1j2}\}_{j=1}^{J}$	Shape
1	1	$\{1\}$	$\{1\}$	
2	2	$\{1,2\}/3$	$\{2,1\}/3$	
3	2	$\{1,4\}/5$	$\{4,1\}/5$	
4	3	$\{1,1,2\}/4$	$\{2,1,1\}/4$	
5	3	$\{1,1,4\}/6$	$\{4,1,1\}/6$	

Table 7.1: Coefficients and cluster shapes for the five M_4 processes.

For each of the five processes, $N = 100$ sequences of length 1000 are simulated and a failure probability $P(X \in B)$ is estimated for each data-set. The failure region is $B = \{x \in \mathbb{R}^2 : x_1 + x_2 > v\}$ with v chosen so that the true failure probability is $p_0 = 1 - \exp(-2/v) = 0.001$. To estimate p_0, intervals declustering is applied with the region $A = \{x \in \mathbb{R}^2 : x \not\leq u\}$, where $u = (u_1, u_2)$ is defined by the empirical 95% quantiles of the two components. The block length is then chosen to produce the same number of clusters in A. Instead of fitting the full Poisson process model to the block maxima, the component distributions are estimated separately and the limiting dependence structure, for which block maxima fall on the line $x_1 = x_2$, is assumed known.

For each data-set, Monte Carlo estimates of $P(X^* \in B)$ and $P(X \in B)$ are computed by simulating one-thousand block maxima from the fitted Poisson process. Their performances as estimates of p_0 are summarised with two measures: relative bias and relative standard deviation, defined for estimates $\{\hat{p}_i\}_{i=1}^N$ as

$$\frac{1}{N}\sum_{i=1}^{N}\left(\frac{\hat{p}_i - p_0}{p_0}\right) \quad \text{and} \quad \sqrt{\frac{1}{N}\sum_{i=1}^{N}\left(\frac{\hat{p}_i - \bar{p}}{p_0}\right)^2},$$

where $\bar{p} = \sum_{i=1}^N \hat{p}_i / N$. The simulation results are presented in Tables 7.2 and 7.3. They show that $P(X^* \in B)$ vastly over-estimates p_0 and that this is greater when the distance from actual points to the block maximum is larger. Estimating the correction factor removes the bias almost completely and reduces the variance. The correction is equally effective for all five processes.

Process	1	2	3	4	5
$P(X^* \in B)$	1.32	6.09	7.11	6.57	9.14
$P(X \in B)$	0.02	0.09	0.11	-0.01	0.04

Table 7.2: Relative bias

Process	1	2	3	4	5
$P(X^* \in B)$	2.98	6.25	6.88	7.93	8.91
$P(X \in B)$	0.70	0.90	0.88	0.99	0.96

Table 7.3: Relative standard deviation

7.4 Data Example

In this section, the model for multivariate clusters is applied to an oceanographic, bivariate time series. The variables, both measured in metres, are surge (X_1) and significant wave height (X_2). Surge is the measured sea-level with the tidal component removed; significant wave height is approximately equal to the mean height of the highest one-third of the waves, and is hindcast. The data were recorded hourly at Christchurch on the south coast of England between 1 April 1978 and 31 March 1990. See Hawkes *et al.* (1998) for additional information and analysis. The raw data exhibit strong seasonality, removal of which is achieved here by using data from only the months December and January. Missing values are avoided by restricting attention to the period December 1980 to January 1987. The two variables are plotted separately in Figure 7.1, and jointly in Figure 7.2.

The first step is to select thresholds over which the two series exhibit stable extremal behaviour. To aid this selection, estimates of the extremal index, the mean excess and the parameters (σ, ξ) of the generalised Pareto distribution are plotted over a range of thresholds in Figures 7.3 and 7.4. The scale parameters of the

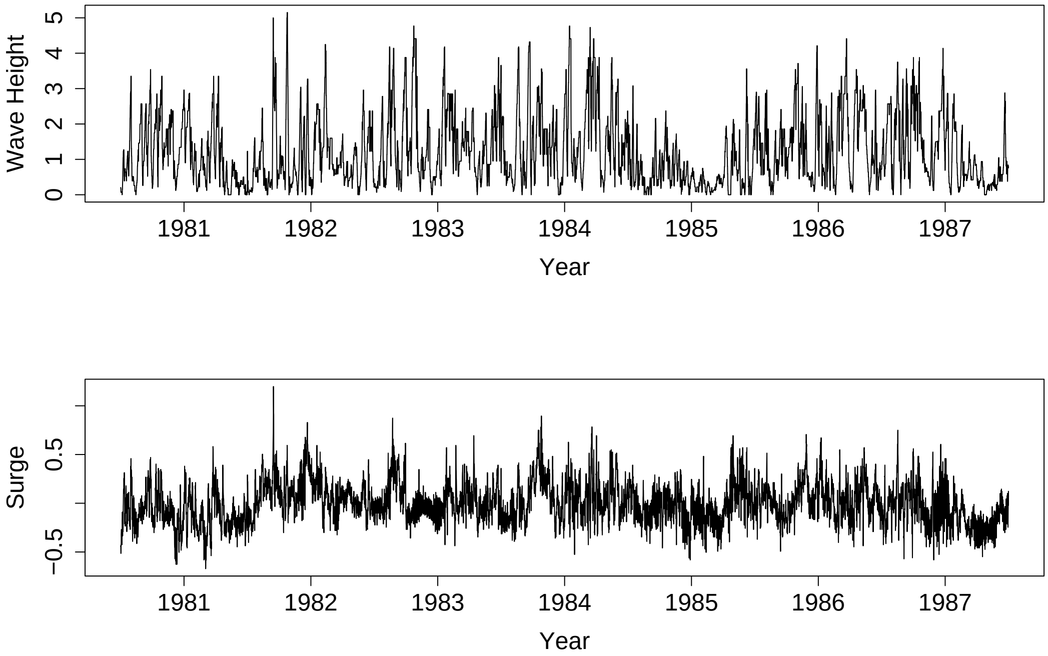


Figure 7.1: Significant wave height (above) and surge (below) recorded hourly in metres at Christchurch. The first observation in each year is labelled.

generalised Pareto distributions have been modified to $\sigma - \xi u$, since this quantity should be approximately constant across thresholds: see Section 6.4. For surge, stability appears to be achieved at $u_1 = 0.4$ if the rogue estimates of the generalised Pareto parameters at threshold 0.6 are ignored. For wave height, the estimates of the scale parameter are confusing. The threshold $u_2 = 3.0$ is tentatively selected. Diagnostic plots for the fit of the generalised Pareto distribution to the 402 surge exceedances and to the 625 wave exceedances are shown in Figure 7.5. Allowing for discreteness, the fit is acceptable for both variables.

Multivariate intervals declustering with extreme region $A = \{x \in \mathbb{R}^2 : x \not\leq u\}$ yields an extremal index estimate of 0.07 and 61 clusters. Blocks declustering also yields 61 clusters if the block length is $b = 82$. This partitions the data into 128

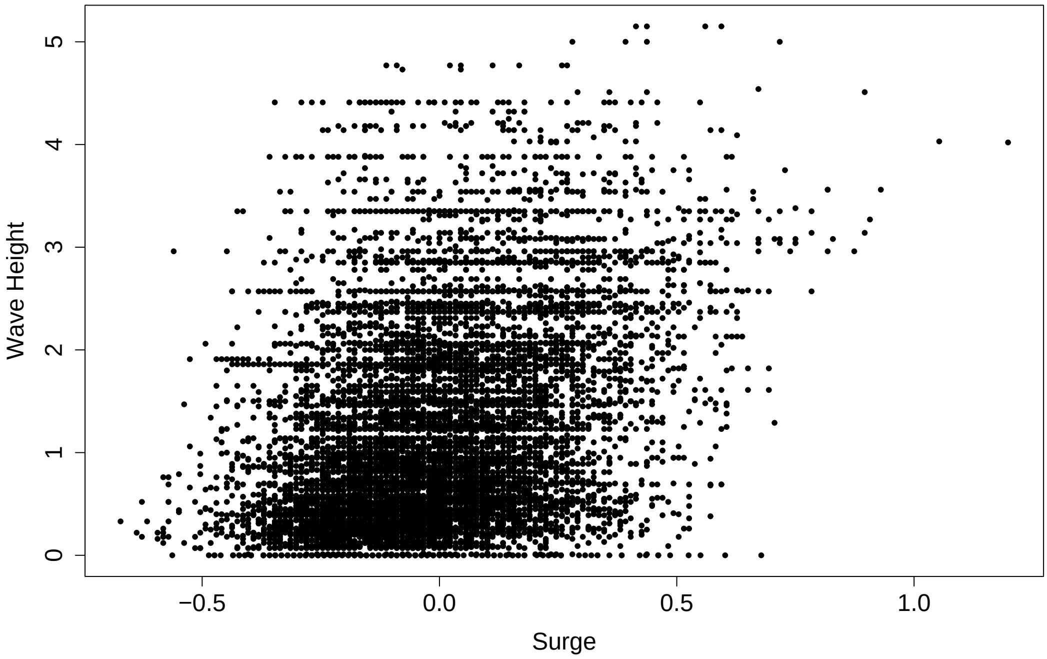


Figure 7.2: Significant wave height against surge, both recorded hourly in metres at Christchurch.

blocks, of which the maxima are plotted in Figure 7.6.

The Poisson process model is fitted to the block maxima with logistic dependence structure (5.6). The parameter estimates for the generalised Pareto components are $(0.20, 0.02)$ for surge and $(1.55, -0.67)$ for wave height; the dependence parameter estimate is 0.45.

The failure region $B = \{x \in \mathbb{R}^2 : x_1 + x_2/8 > v\}$ with $v = 2$ is used to illustrate the estimation of failure probabilities. This form of failure region approximates the conditions under which certain types of sea-wall are breached: see Coles and Tawn (1994). One-thousand cluster maxima are simulated from the fitted model in a region containing B, and superimposed on Figure 7.6. Note the upper limit of $u_2 - \sigma_2/\xi_2 = 5.32$ metres for the estimated significant wave height distribution. The

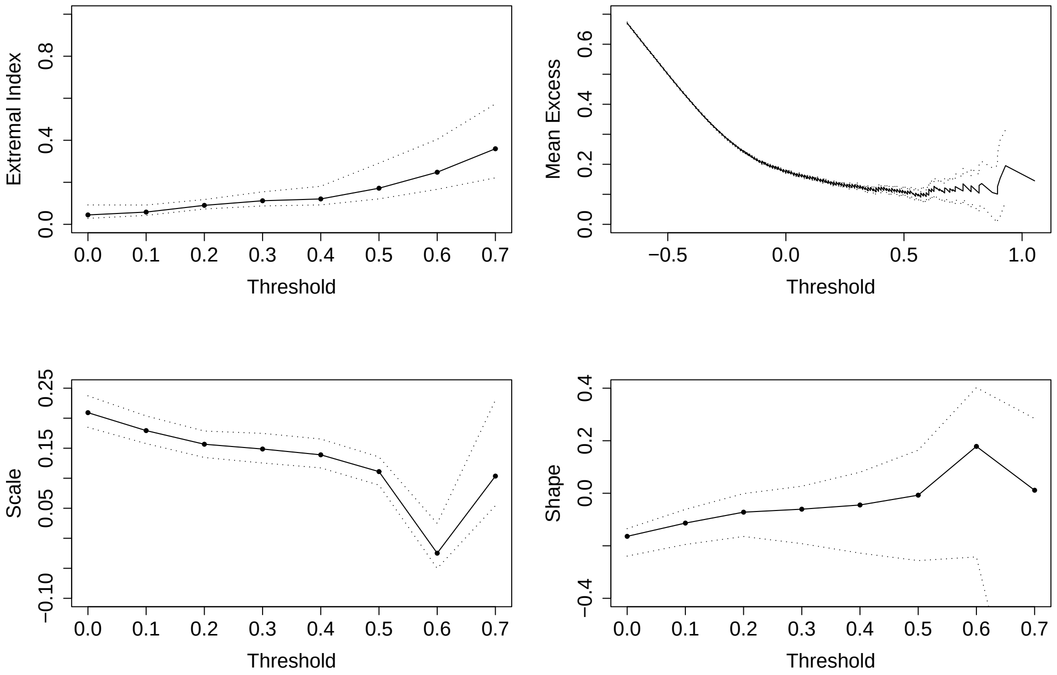


Figure 7.3: Estimates (–●–) for surge of the extremal index, mean excess, and parameters of the generalised Pareto distribution with bootstrapped 90% confidence intervals ($\cdots$).

estimate of $P(X^* \in B)$ is 0.0039. For $n_m n_c = 38\,918$ clusters with a maximum in B, 163 955 out of the nominal $bn_m n_c = 3\,191\,276$ points are in B. So the estimate of the correction factor is $163\,955/3\,191\,276 = 0.051$ and the estimate of $P(X \in B)$ is 0.00020.

The estimate of $P(X^* \in B)$ is in accordance with the data: there are 128 blocks in the sample, none of which has maximum in B, suggesting that the probability should be less than $1/128 = 0.0078$. There are 10 416 observations in the sample, and again none of these is in B, which suggests that $P(X \in B)$ should be less than $1/10\,416 = 0.0001$. Although the estimate is double this value, it is not at odds with the data because points will tend to occur in B in clusters. Indeed, the

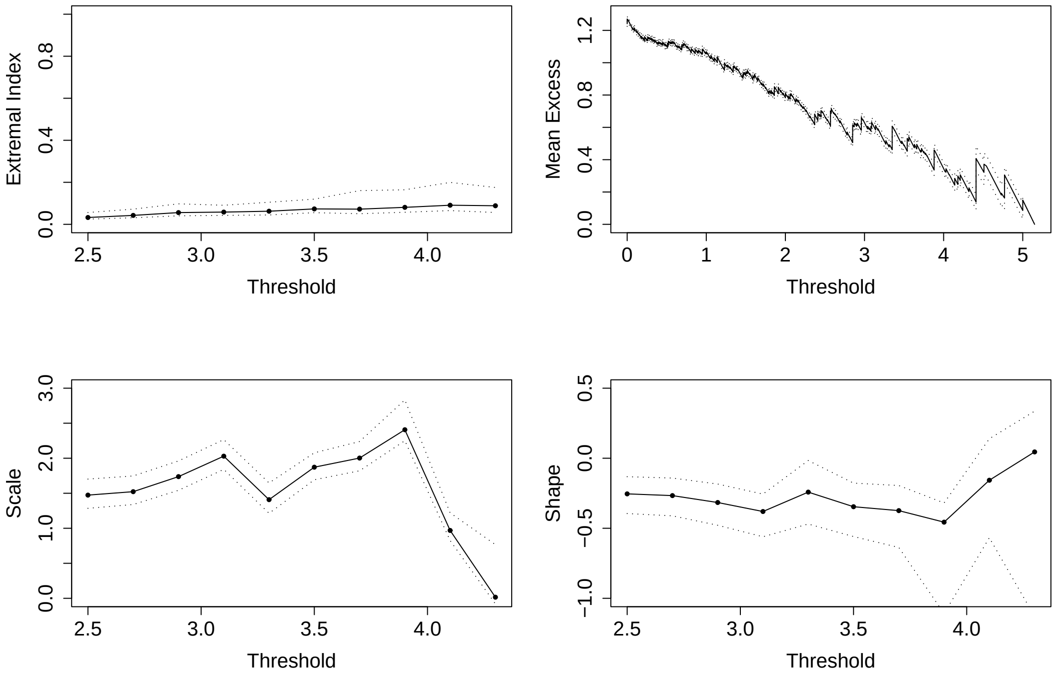


Figure 7.4: Estimates (–●–) for significant wave height of the extremal index, mean excess, and parameters of the generalised Pareto distribution with bootstrapped 90% confidence intervals $(\cdots)$.

average number of points in B among the simulated clusters that have a point in B is about 4.

7.5 Discussion

The technique described in this chapter allows failure probabilities to be estimated without prior filtering of the data, which can lead to over-estimation. Only simple failures, caused by a single observation falling in a failure region, have been considered, but failures might result from more complex events. For example, failure may occur only if two consecutive observations fall in the failure region. The probabilities of such failures cannot be estimated using filtered data, but are handled

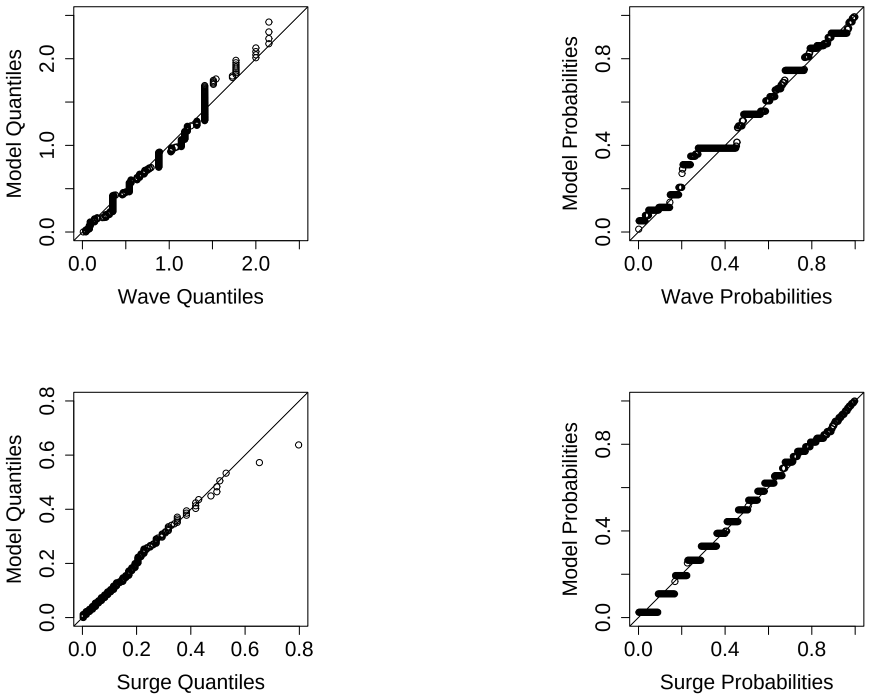


Figure 7.5: Quantile plots (left) and probability plots (right) for the generalised Pareto fit at the chosen thresholds for significant wave height (top) and surge (bottom).

naturally when clusters are modelled.

Obtaining confidence intervals for failure probabilities using the method of this chapter is complicated. In order to account for uncertainty in the block length, the strings and the Poisson process model, it is perhaps simplest to apply a block bootstrap to the original multivariate sequence. Using the block length identified by the declustering procedure would appear to be a natural choice for bootstrapping, but this issue has not been explored here.

Finally, estimates of the failure probability $P(X \in B)$ can be obtained from $P(X^* \in B)$ in another way if B has a particular form. Suppose that the se-

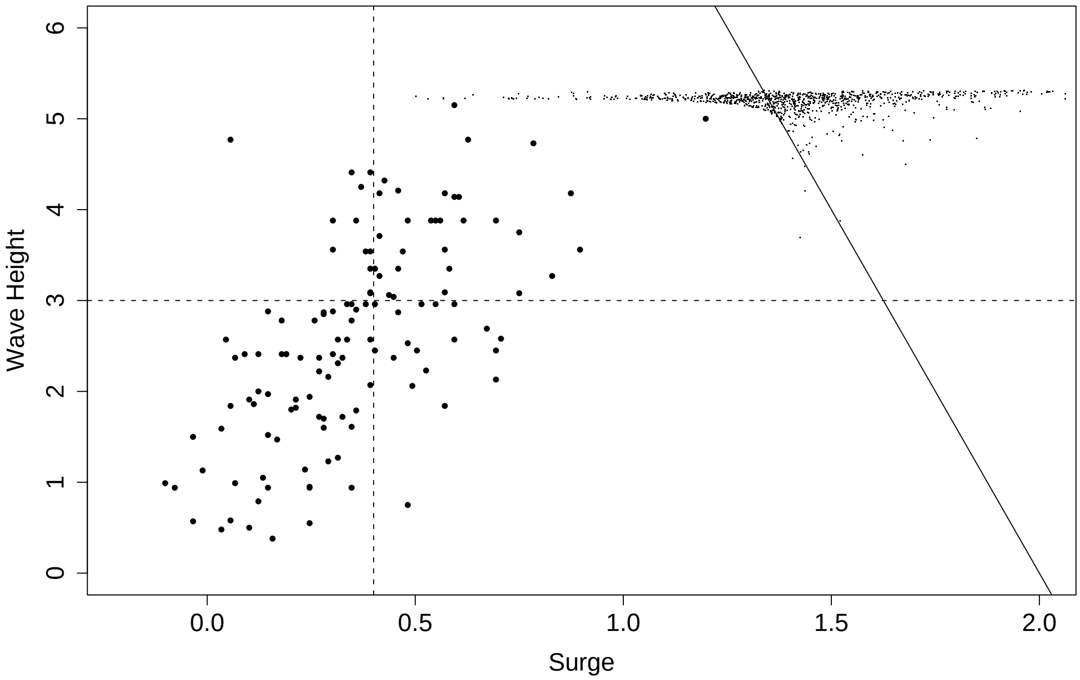


Figure 7.6: Block maxima (●), component thresholds (- - -), boundary of the failure region (——) and simulated cluster maxima (·).

quence has multivariate extremal index $\theta(\cdot)$ satisfying limit (3.2). Then, for $B = \{x \in \mathbb{R}^D : x \not\leq u\}$, and a suitably large block length b,

$$P(X^* \notin B) \approx P(X \notin B)^{\theta(u)b}.$$

Combining an estimate of $\theta(u)$, perhaps obtained using one of the methods described in Section 4.2, with an estimate of $P(X^* \notin B)$ yields an estimate of $P(X \in B)$. Whether or not such approximations can be exploited for general failure regions B remains an open question.