Supporting Information for "Similarity-Based Clustering for Patterns of Extreme Values"†

Miguel DE CARVALHO, Raphael HUSER, and Rodrigo RUBIO

† M. de Carvalho is Reader in Statistics, School of Mathematics, University of Edinburgh, UK (e-mail: *miguel.decarvalho@ed.ac.uk*). R. Huser is Associate Professor of Statistics, CEMSE Division, King Abdullah University of Science and Technology, Saudi Arabia (e-mail: *raphael.huser@kaust.edu.sa*). R. Rubio is Head of Data Analytics, BCI Bank, Chile. The research was partially funded by *Fundação para a Ciência e a Tecnologia* (Portuguese NSF) through the project UID/MAT/00006/2020, and by King Abdullah University of Science and Technology (KAUST).

1 Clustering performance measures employed in Section 3

In this appendix we give details on the precise definition of the measures used to assess the performance of our methods in Section 3 of the paper. Let $\Omega = \{\pi^1, \ldots, \pi^N\} \subseteq \Pi$ be a set of N points in the product-space $\Pi = \mathscr{C} \times (0, \infty)$, and let $P = \{C^{[1]}, \ldots, C^{[K]}\}$ and $P' = \{C'^{[1]}, \ldots, C'^{[K]'}\}$ be two partitions of Ω that we want to compare. Here, P partitions Ω into K subsets (i.e., clusters), while P' partitions Ω into K' (possibly different) subsets. Then, the Rand index $\mathcal{R}$ is defined as

$$\mathcal{R} = \frac{n_1 + n_2}{\binom{N}{2}} \in [0, 1],$$

where n_1 is the number of pairs of elements in Ω that are simultaneously in the same set in P and in the same set in P', and n_2 is the number of pairs of elements in Ω that are simultaneously in different sets in P and in different sets in P'. Roughly speaking, the Rand index measures the percentage of identical decisions made by two separate clustering algorithms leading to P and P', respectively. It takes values between 0 and 1, with $\mathcal{R} = 0$ indicating that the two algorithms do not agree on any pair of points, and $\mathcal{R} = 1$ indicating that they agree on all pairs. If the true data structure is known, then the Rand index can be used to assess the performance of a clustering algorithm, with higher Rand indices corresponding to better results.

To define the silhouette index, let $P = \{C^{[1]}, \ldots, C^{[K]}\}$ be a partition of

$$\Omega = \{\pi^1, \ldots, \pi^N\} \subseteq \Pi,$$

denoting the outcome of a similarity-based clustering algorithm (e.g., as in Section 2) based on the dissimilarity measure $D_\alpha : \Pi \times \Pi \mapsto [0, \infty)$ (e.g., from Example 1). Let $a(i)$ be the average dissimilarity of the ith point $\pi^i \in \Pi$ with respect to all other points classified within the same cluster. Mathematically, assuming that π^i belongs to the kth cluster, one has

$$a(i) = \frac{1}{N^{[k]} - 1} \sum_{j \in I^{[k]}, i \neq j} D_\alpha(\pi^i, \pi^j),$$

where $I^{[k]}$ is the index set corresponding to the kth cluster $C^{[k]}$ and $N^{[k]}$ is the cardinality of $C^{[k]}$, i.e., $N^{[k]} = |C^{[k]}|$. Furthermore, let $M(\pi^i, C^{[k']})$ denote the average dissimilarity of the ith point $\pi^i \in \Pi$ with respect to all points classified in a different cluster $C^{[k']}$, $k' \neq k$, i.e.,

$$M(\pi^i, C^{[k']}) = \frac{1}{N^{[k']}} \sum_{j \in I^{[k']}} D_\alpha(\pi^i, \pi^j),$$

where $I^{[k']}$ is the index set corresponding to the cluster $C^{[k']}$ and $N^{[k']} = |C^{[k']}|$. Let also $b(i)$ be the smallest of these average dissimilarities, i.e., $b(i) = \min_{k' \neq k} M(\pi^i, C^{[k']})$. Then, the silhouette index of the ith point π^i is defined as

$$s(i) = \frac{b(i) - a(i)}{\max\{a(i), b(i)\}} \in [-1, 1].$$

Table 1: Extremal index estimate $\widehat{\theta}$ for the stocks under analysis. The threshold is chosen as in the application in Section 4 of the main paper.

1	2	3	4	5	6	7	8	9	10	11	12	13
0.474	0.329	0.351	0.613	0.241	0.426	0.468	0.627	0.501	0.407	0.232	0.602	0.706
14	15	16	17	18	19	20	21	22	23	24	25	26
0.392	0.458	0.493	0.573	0.314	0.315	0.496	0.288	0.486	0.282	0.458	0.594	0.249

Intuitively, $s(i)$ is a normalized measure of adequacy of the ith point to its assigned cluster; high values indicate that π^i is a good match for its own cluster and a poor match for the other clusters. The cluster-mean silhouette index, and overall silhouette index are respectively defined as

$$s^{[k]} = \frac{1}{N^{[k]}} \sum_{i \in I^{[k]}} s(i), \qquad S = \frac{1}{K} \sum_{k=1}^{K} s^{[k]}.$$

The overall silhouette index S, used in Section 3 of the paper as a measure of clustering quality, lies within the interval $[-1, 1]$ and tells us if the estimated clusters are well separated.

2 Additional empirical reports supplementing Section 4

2.1 Temporal dependence analysis

We present here some supplementary diagnostics providing further details on temporal dependence for the case study in Section 4 of the paper (London Stock Exchange case study). We report tables and charts with estimates of the extremal index θ (Coles, 2001, Ch. 5), which measures the strength of dependence at extreme levels. In particular, we use an estimator based on inter-exceedance times series using the approach proposed by Ferro and Segers (2003)—using the functions `exi` and `exiplot` from the `R` package `evd`—and plot the extremogram, i.e., the empirical conditional probability $P(Y_{t+h} > u \mid Y_t > u)$ as function of the time lag $h = 0, 1, 2, \ldots$, and for a high threshold u (Davis and Mikosch, 2009)—using the `R` package `extremogram`. In Table 1, we report the extremal index estimates for each stock, and in Figure 1 we plot the estimates of these values over a grid of high thresholds u. Figures 2 and 3 show the extremograms for the negative loss returns for the stocks analyzed in the data application, using the threshold chosen in Section 4 of the main paper.

From Figures 1–3 one can see that extremal dependence in each time series is quite weak. This is exacerbated by considering that these diagnostics were computed from the original time series, which exhibit some degree of heteroscedasticity, therefore potentially suggesting stronger extremal dependence than there is in reality. This indicates that extremal dependence can be neglected in our real data application.

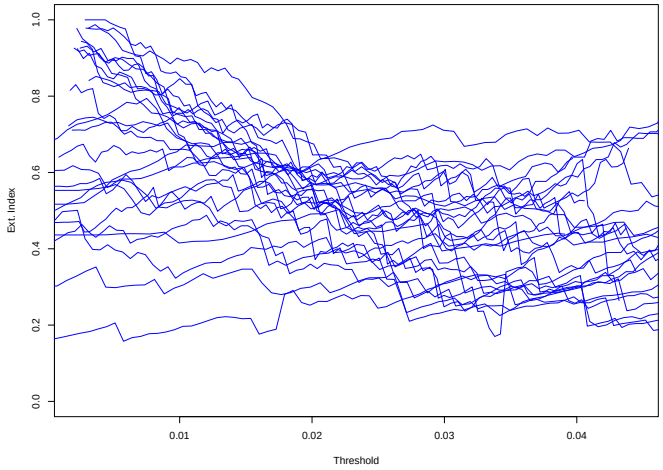

Figure 1: Extremal index estimate for the stocks under analysis for different values of the threshold u.

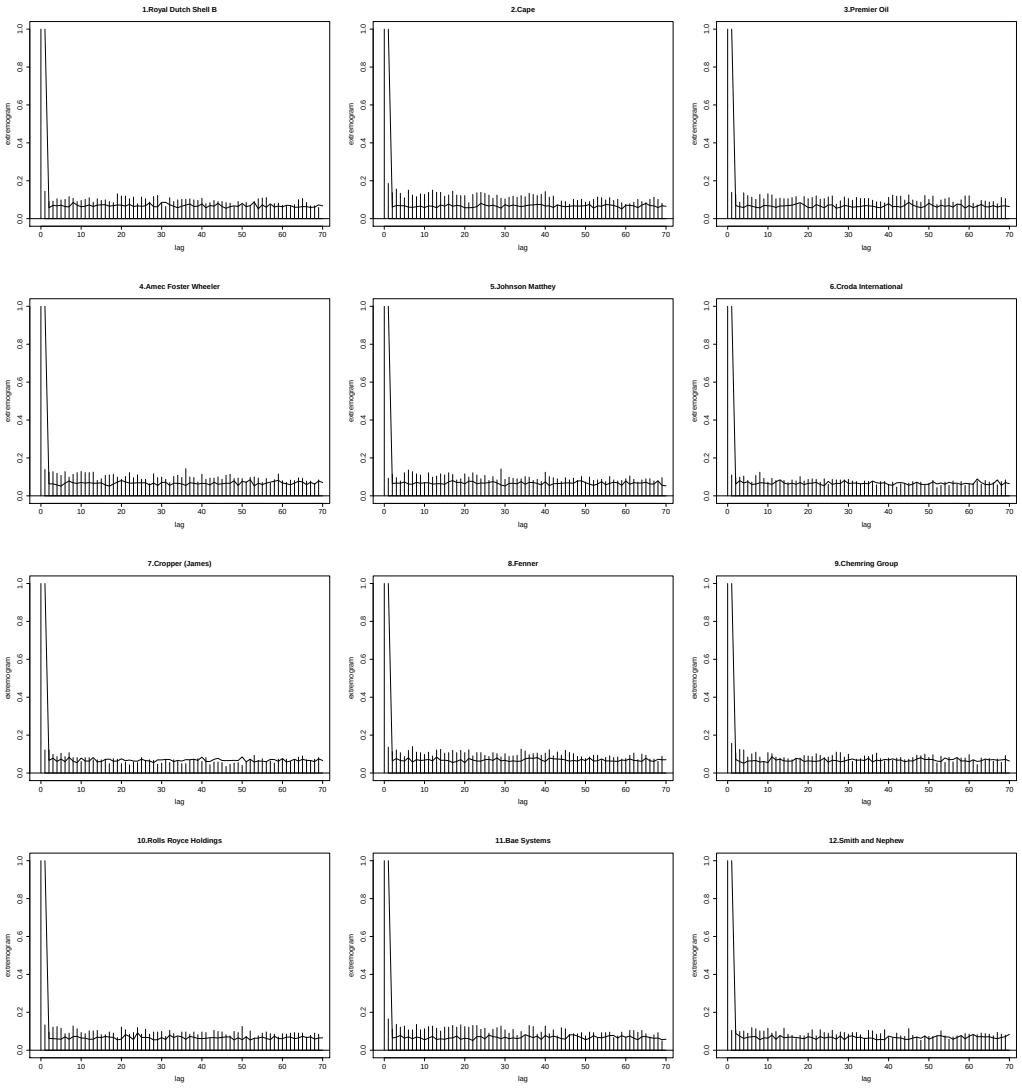

Figure 2: Empirical extremogram for stocks 1–12 (left to right, top to bottom). The threshold is chosen as in the application in Section 4 of the main paper.

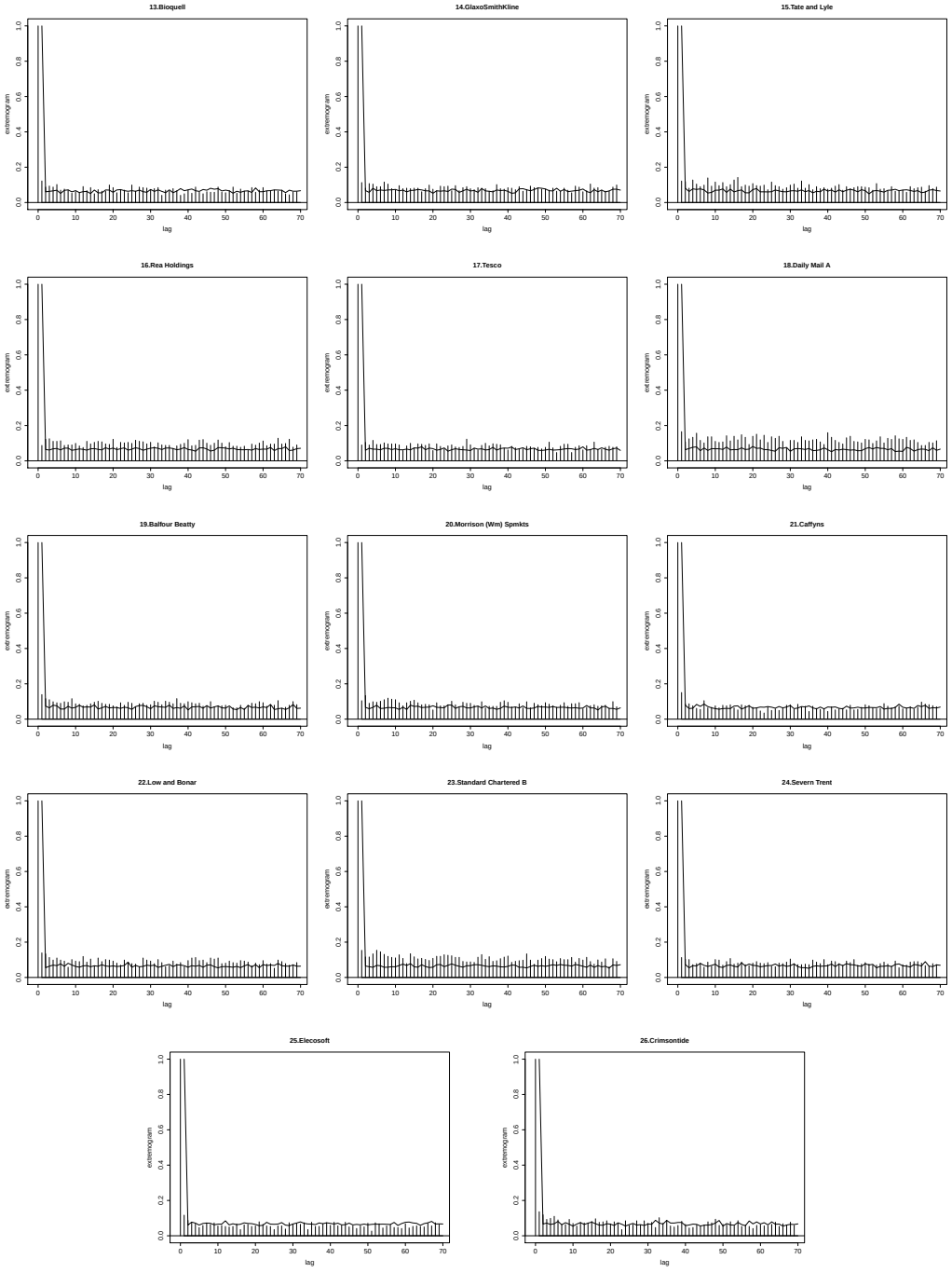

Figure 3: Empirical extremogram for stocks 13–26 (left to right, top to bottom). The threshold is chosen as in the application in Section 4 of the main paper.

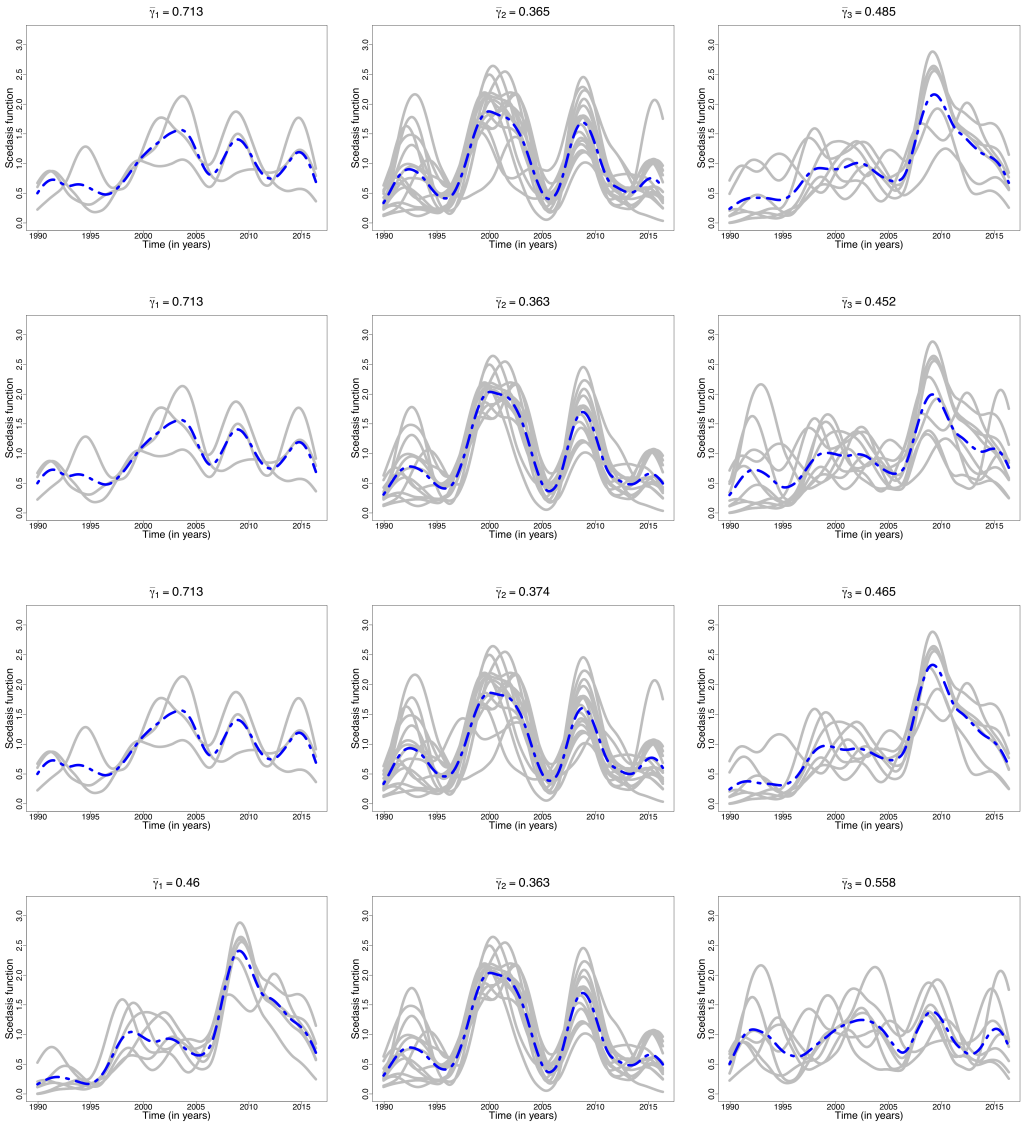

Figure 4: Scedasis functions (grey) partitioned per cluster, and cluster center scedasis functions (blue); clustering is obtained with the K-means algorithm for heteroscedastic extremes using $K = 3$ and with $\alpha = 0, 0.1, 0.3$ and 0.5, from top to bottom, respectively.

2.2 Sensitivity analysis

Here, we present a sensitivity analysis to assess how the clustering of stocks in the London Stock Exchange changes as a function of the tuning parameter α and the number of clusters K. We consider values of α in the grid $\{0, 0.1, \ldots, 0.9, 1\}$ and use $K = 3$ and $K = 5$. Figures 4 and 7 plot the partition of scedasis functions obtained using the K-means algorithm described in Section 2.3 of paper. Figure 4 shows that the clustering changes slightly but not significantly when using different values of α; from Figure 7, there are only noteworthy changes when $\alpha = 1$ (focus is fully on the dynamics of extreme losses) or $\alpha = 0$ (focus is fully on the magnitude of extreme losses). Figures 8 and 9 represent the partition of value-at-risk functions obtained using the K-geometric means algorithm described in Section 2.4 of paper.

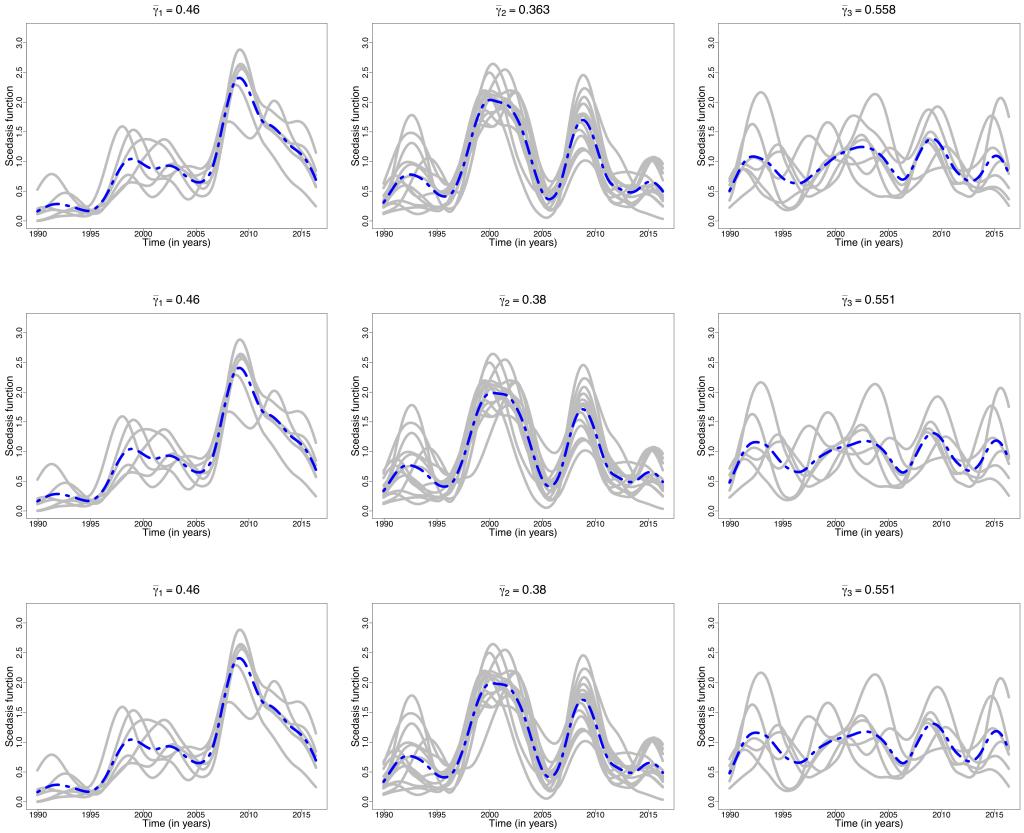

Figure 5: Scedasis functions (grey) partitioned per cluster, and cluster center scedasis functions (blue); clustering is obtained with the K-means algorithm for heteroscedastic extremes using $K = 3$ and with $\alpha = 0.7, 0.9$ and 1, from top to bottom, respectively.

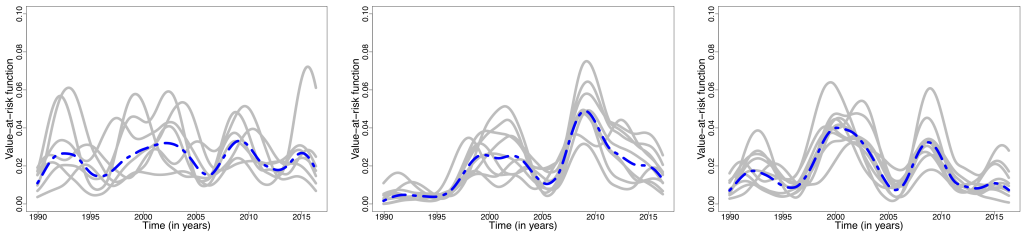

Figure 8: Value-at-risk functions (grey) partitioned per cluster, and value-at-risk cluster center function (blue), obtained with the K-geometric means algorithm for heteroscedastic extremes with $K = 3$ and $p = 0.95$.

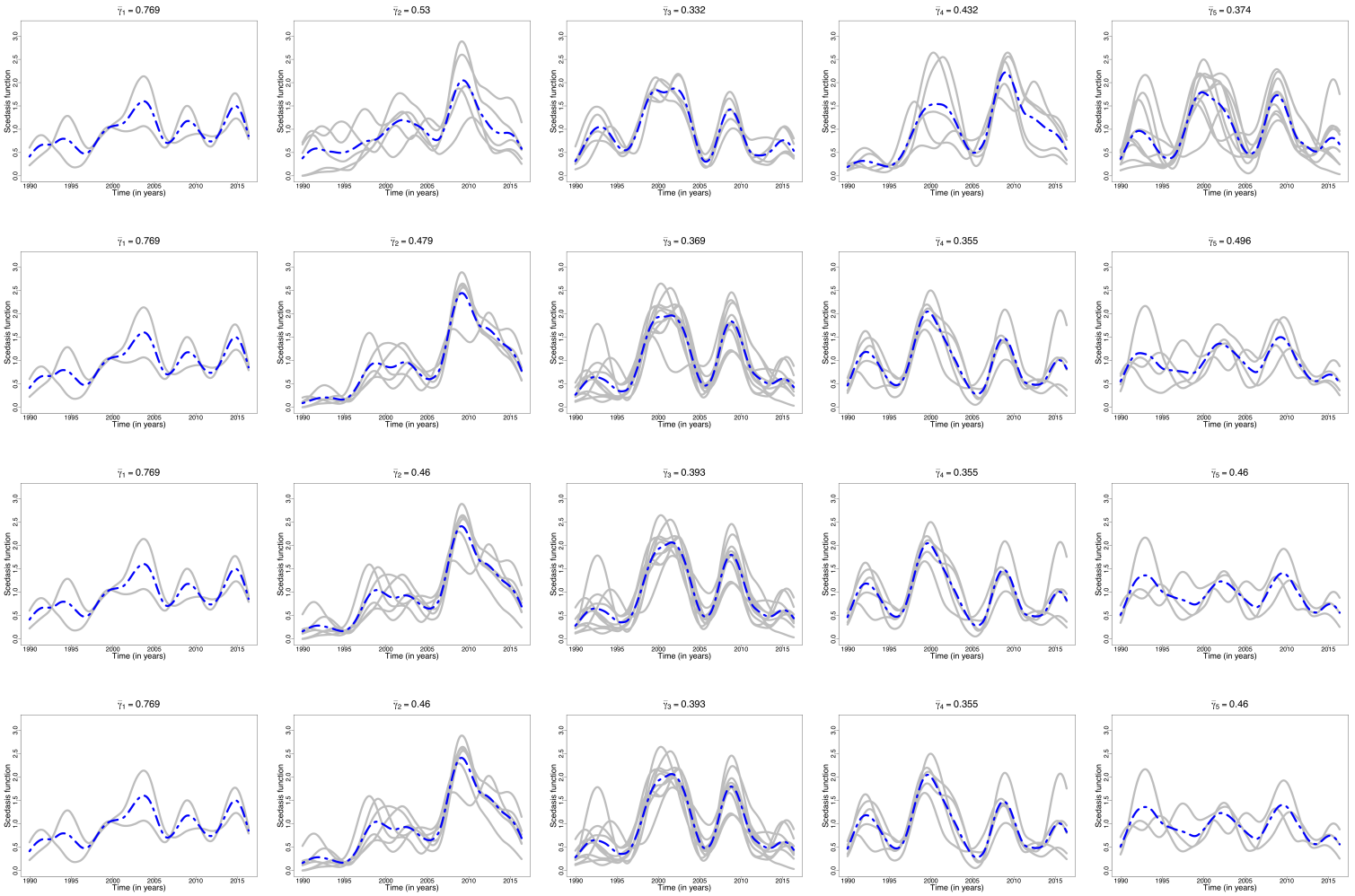

Figure 6: Scedasis functions (grey) partitioned per cluster, and cluster center scedasis functions (blue); clustering is obtained with the K-means algorithm for heteroscedastic extremes using $K = 5$ and with $\alpha = 0, 0.1, 0.3$ and 0.5, from top to bottom, respectively.

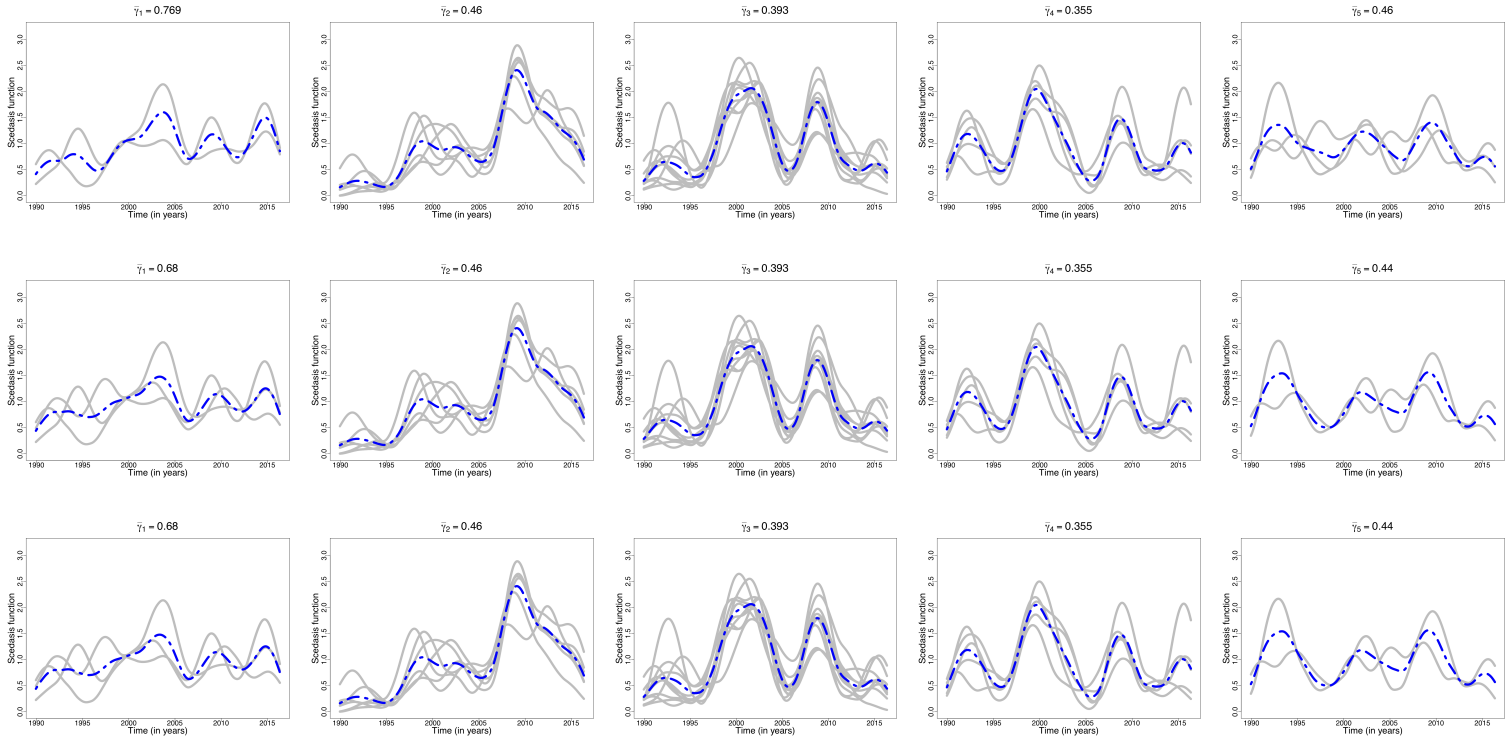

Figure 7: Scedasis functions (grey) partitioned per cluster, and cluster center scedasis functions (blue); clustering is obtained with the K-means algorithm for heteroscedastic extremes using $K = 5$ and with $\alpha = 0.7, 0.9$ and 1, from top to bottom, respectively.

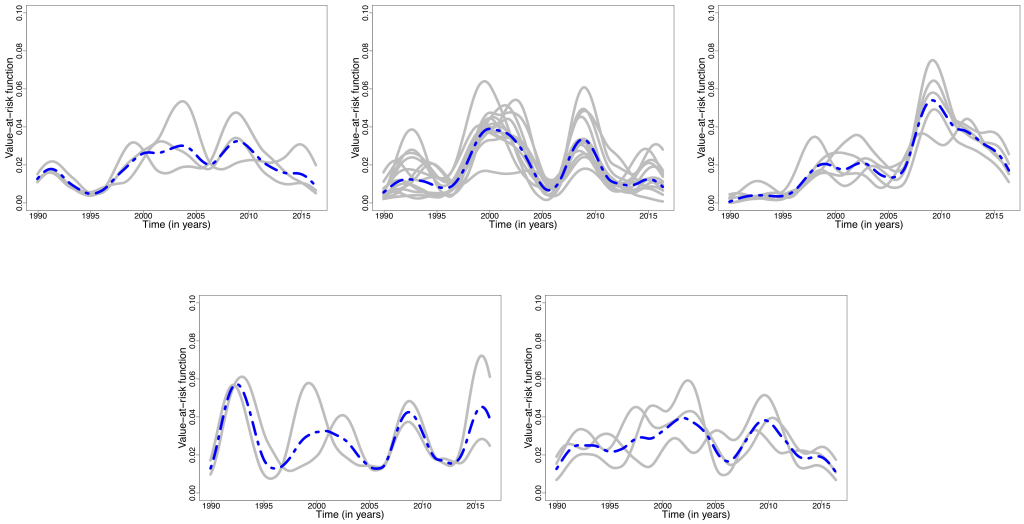

Figure 9: Value-at-risk functions (grey) partitioned per cluster, and value-at-risk cluster center function (blue); clustering is obtained with the K-geometric means algorithm for heteroscedastic extremes with $K = 5$ and $p = 0.95$.

2.3 Extended analysis

This section revisits the inquiries conducted in Sections 4.2–4.3 of the paper, but now considering all 139 stocks retained according the criteria described in Section 4.1. We recall that a main goal in Section 4.2 was to assess whether or not our estimated clusters of risk mirror the nine economic sectors of the London Stock Exchange, namely: oil and gas, basic materials, industrials, healthcare, consumer goods, consumer services, financials, utilities, and technology. Figures 10 and 11 respectively depict the K-means clustering with respect to the scedasis function and extreme-value index, and the K-geometric means clustering with respect to the time-varying value-at-risk, both using $K = 9$. Figure 10 reveals a fairly homogeneous distribution of number of curves within the estimated clusters, as opposed to the clustering displayed in Figure 11. The main finding from Section 4.2 in the paper also holds here: The partition of the stocks available from Figure 10 suggests that clusters of magnitude and frequency of extreme losses present no straightforward connection with the economic sectors of the corresponding stocks; the same finding holds if we examine the clusters of value-at-risk functions available from Figure 11. Figures 10 and 11 also corroborate the analysis reported in Section 4.3. In particular, as it can be seen from these figures, the modes of the scedasis and value-at-risk functions line up with periods of economic stress as dated by the business cycle chronologies from the National Bureau of Economic Research (NBER) and the Economic Cycle Research Institute (ECRI).

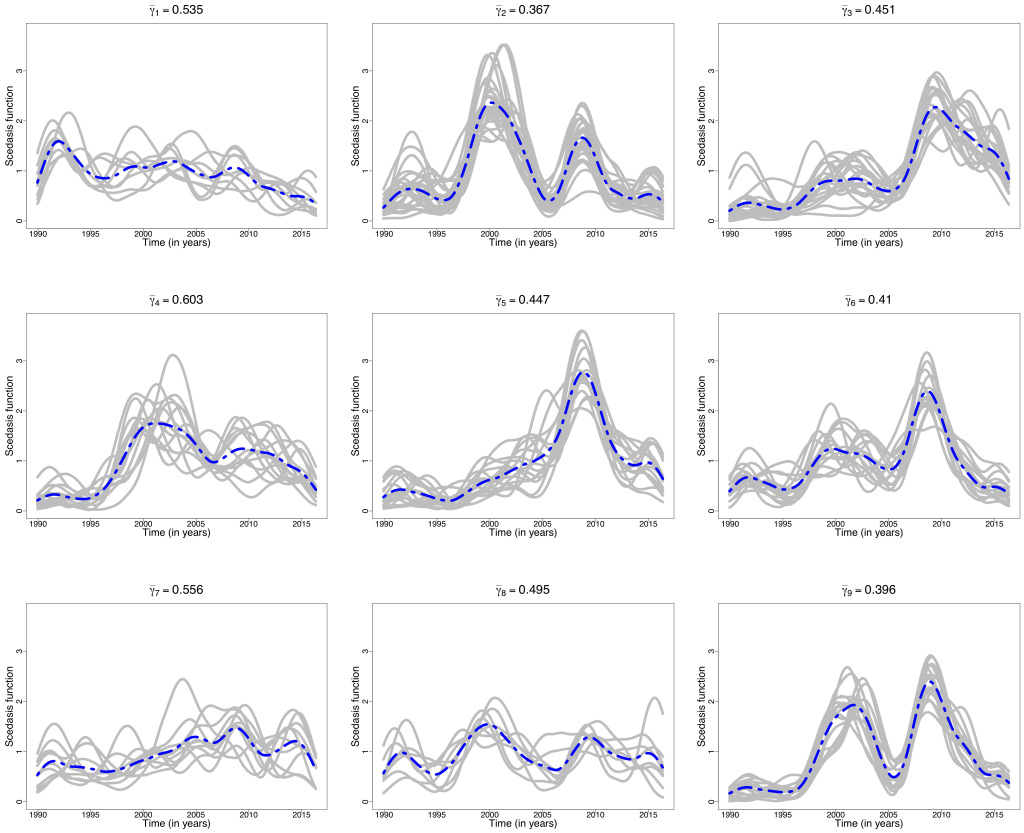

Figure 10: Extended analysis: Scedasis functions (grey) partitioned per cluster and cluster center scedasis functions (blue), obtained with the K-means algorithm for heteroscedastic extremes using $K = 9$ and with $\alpha = 0.5$.

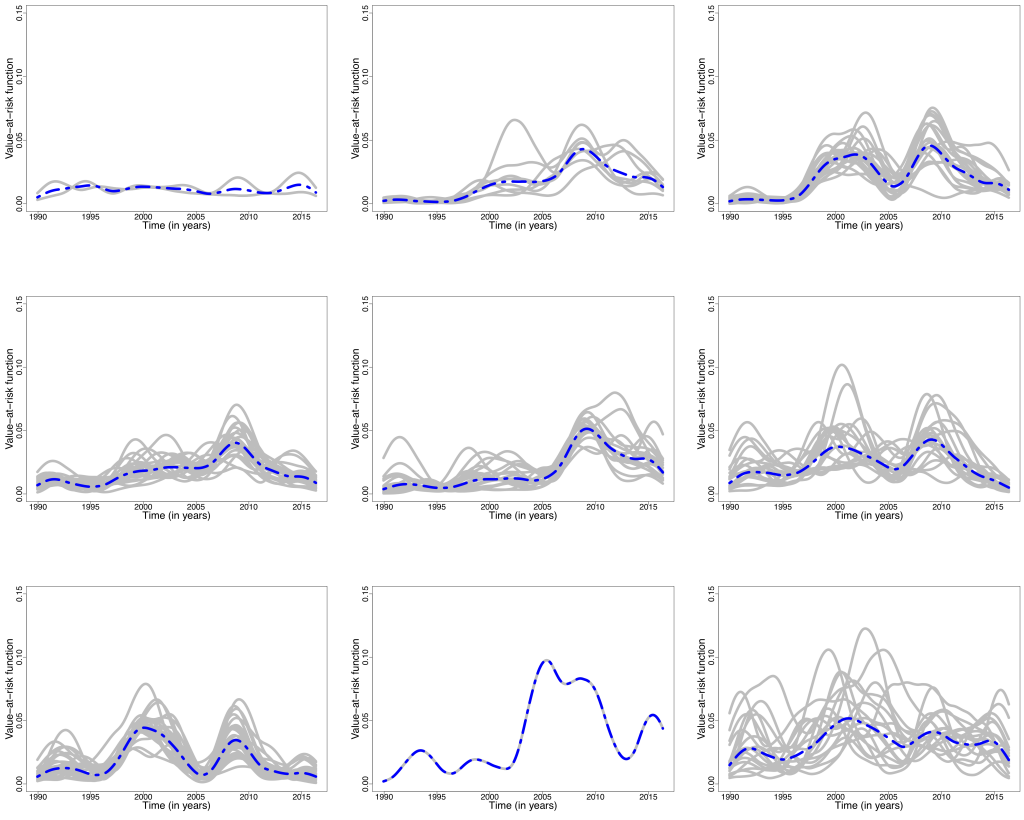

Figure 11: Extended analysis: Value-at-risk functions (grey) partitioned per cluster and value-at-risk cluster (blue), obtained with the K-geometric means algorithm for heteroscedastic extremes with $K = 9$ and $p = 0.95$.

3 R package `extremis`

Below we present the code used for replicating a one-shot experiment for fitting our K-means clustering method for heteroscedastic extremes. The core function is khetmeans from the R package `extremis`; the code below also requires the package `evd` for simulating the data.

```
## Load packages and set specifications
require(extremis)
require(evd)
set.seed(12)
T <- 5000
n <- 30

## True scedasis and extreme value index
c2 <- function(s)
    dbeta(s, 2, 5)
c3 <- function(s)
    dbeta(s, 5, 2)
gamma1 <- 0.7
gamma2 <- 1

## Simulate data
X <- matrix(0, ncol = T, nrow = n)
for(i in 1:5)
  for(j in 1:T)
    X[i,  j] <- rgev(1, c2(j / T), c2(j / T), gamma1)
for(i in 6:15)
  for(j in 1:T)
    X[i,  j] <- rgev(1, c2(j / T), c2(j / T), gamma2)
for(i in 16:20)
  for(j in 1:T)
    X[i,  j] <- rgev(1, c3(j / T), c3(j / T), gamma1)
for(i in 21:30)
  for(j in 1:T)
    X[i,  j] <- rgev(1, c3(j / T), c3(j / T), gamma2)
Y <- t(X)

## K-means for heteroscedastic extremes
fit <- khetmeans(Y, centers = 4)

##
## Rendering
## =========
## 1
## 2
## 3

plot(fit, c.c = TRUE, xlab = "w", ylab = "Scedasis Density",
     ylim = c(0, 4), main = "T = 5000",
     mgp = c(2, 0.5, 0))

## add lines with true scedasis
```

```r
grid <- seq(0, 1, length = 100)
lines(grid, c2(grid), type = 'l')
lines(grid, c3(grid), type = 'l')
```

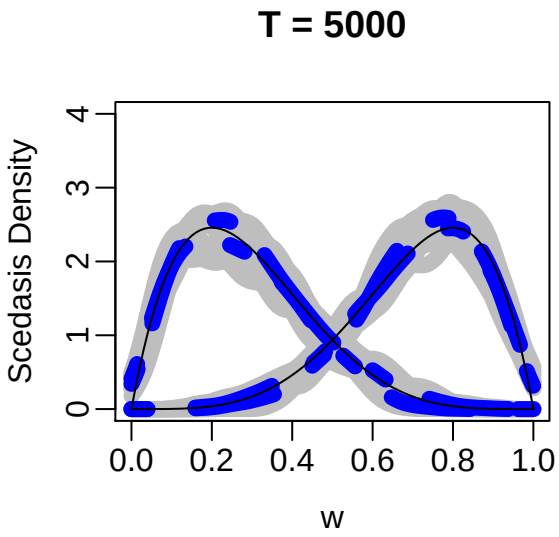

Figure 12: Reproducing part of Figure 2 (Scenario C, $T = 5000$) in the manuscript.

References

Coles, S. (2001) *An Introduction to Statistical Modeling of Extreme Values*. London: Springer.

Davis, R. A. and Mikosch, T. (2009) The extremogram: A correlogram for extreme events. *Bernoulli*, **15**, 977–1009.

Einmahl, J. H. J., L. de Haan and Zhou, C. (2016) Statistics of heteroscedastic extremes. *J. R. Statist. Soc., Ser.* B, **78**, 31–51.

Ferro, C. A. and Segers, J. (2003), Inference for clusters of extreme values. *J. R. Statist. Soc., Ser.* B, **65**, 545–556.

Tibshirani, R., Walther, G. and Hastie, T. (2001) Estimating the number of clusters in a data set via the gap statistic. *J. R. Statist. Soc., Ser.* B, **63**, 411–423.