# Fitting Tails by the Empirical Residual Coefficient of Variation: The ercv Package

*by Joan del Castillo, Isabel Serra, Maria Padilla and David Moriña*

**Abstract** This article is a self-contained introduction to the R package **ercv** and to the methodology on which it is based through the analysis of nine examples. The methodology is simple and trustworthy for the analysis of extreme values and relates the two main existing methodologies. The package contains R functions for visualizing, fitting and validating the distribution of tails. It also provides multiple threshold tests for a generalized Pareto distribution, together with an automatic threshold selection algorithm.

## Introduction and overview

Extreme value theory (EVT) is one of the most important statistical techniques for the applied sciences. A review of the available software on extreme value analysis appears in Gilleland et al. (2013). R software (R Core Team, 2017) contains some useful packages for dealing with EVT. The R package **evir** (Pfaff and McNeil, 2012) provides maximum likelihood estimation (MLE) at the same time for the block maxima and threshold model approaches. The R package **ismev** (Heffernan and Stephenson, 2018) allows fitting parameters of a generalized Pareto distribution depending on covariates and offers diagnostics such as qqplots and return level plots with confidence bands. The R package **poweRlaw** (Gillespie, 2015) enables power laws and other heavy tailed distributions to be fitted using the techniques proposed by Clauset et al. (2009).This approach had been used to describe sizes of cities and word frequency and is linked to the physics of phase transitions and to complex systems.

This paper shows that the R package **ercv** (del Castillo et al., 2017a), based on the coefficient of variation (CV), is a complement, and often an alternative, to the available software on EVT. The mathematical background is shown in Section Mathematical Background, including threshold models and the relationship between power law distribution and the generalized Pareto distributions (GPD), which is the relationship between the two different approaches followed by the aforementioned R packages **evir**, or **ismev**, and **poweRlaw**.

Section Exploratory data analysis with cvplot function introduces the tools for the empirical residual coefficient of variation developed in the papers del Castillo et al. (2014), del Castillo and Serra (2015) and del Castillo and Padilla (2016). Section Examples also shows the exploratory data analysis of nine examples, some of them from the R packages **evir** and **poweRlaw**, with the `cvplot` function, see Figure 1.

Section Estimation and Model diagnostics with Tm function explains the `Tm` function in the R package **ercv** that provides a multiple thresholds test that truly reduces the multiple testing problem in threshold selection and provides clearly defined $p$-values. The function includes an estimation method of the extreme value index. An automatic threshold selection algorithm provided by the `thrselect` function is explained in Section 2.5 to determine the point above which GPD can be assumed for the tail distribution.

Section Transformation from heavy to light tails (tdata) shows how the methodology developed in the previous sections can be extended with the `tdata` function to all GPD distributions, even with no finite moments. This technique is applied to the MobyDick example and to the Danish fire insurance dataset, a highly heavy-tailed, infinite-variance model. Finally, Section Fitting PoT parameters and tail plots (fitpot ccdfplot) describes the functions of the R package **ercv** that allow estimation of the parameters (`fitpot`) and drawing of the adjustments (`ccdfplot`) for the peak-over-threshold method.

## Mathematical Background

Extreme value theory is widely used to model exceedances in many disciplines, such as hydrology, insurance, finance, internet traffic data and environmental science. The underlying mathematical basis is now thoroughly established in Leadbetter et al. (1983), Embrechts et al. (1997), de Haan and Ferreira (2007), Novak (2012) and Resnick (2013). Statistical tools and methods for use with a single time series of data, or with a few series, are well developed in Coles (2001), Beirlant et al. (2006) and Markovich (2007).

### Threshold models

The first fundamental theorem on EVT by Fisher and Tippett (1928) and Gnedenko (1943) characterizes the asymptotic distribution of the maximum in observed data. Classical analyses now use the generalized extreme value family of distribution functions for fitting to block maximum data provided the number of blocks is sufficiently large. Another point of view emerged in the 1970's with the fundamental theorem by Pickands (1975) and Balkema and de Haan (1974). The Pickands-Balkema-DeHaan (PBdH) theorem, see McNeil et al. (2005, chap 7), initiated a new way of studying extreme value theory via distributions above a threshold, which use more information than the maximum data grouped into blocks.

Let $X$ be a continuous non-negative r.v. with distribution function $F(x)$. For any threshold, $t > 0$, the r.v. of the conditional distribution of threshold excesses $X - t$ given $X > t$, denoted as $X_t = \{X - t \mid X > t\}$, is called the *residual distribution* of $X$ over $t$. The cumulative distribution function of $X_t$, $F_t(x)$, is given by

$$1 - F_t(x) = (1 - F(x + t))/(1 - F(t)). \tag{1}$$

The quantity $M(t) = E(X_t)$ is called the *residual mean* and $V(t) = var(X_t)$ the *residual variance*. The plot of sample mean excesses over increasing thresholds is a commonly used diagnostic tool in risk analysis called ME-plot (`meplot` function in **evir** R package).

The *residual coefficient of variation* is given by

$$\mathrm{CV}(t) \equiv \mathrm{CV}(X_t) = \sqrt{V(t)}/M(t), \tag{2}$$

like the usual CV, the function $\mathrm{CV}(t)$ is independent under change of scale.

The PBdH theorem characterizes the asymptotic distributions of the residual distribution over a high threshold under widely applicable regularity conditions, see Coles (2001). The result essentially says that GPD is the canonical distribution for modelling excess over high thresholds. The probability density function for a $\mathrm{GPD}(\xi, \psi)$ is given by

$$g(x; \xi, \psi) = \begin{cases} \psi^{-1}(1 + \xi x/\psi)^{-(1+\xi)/\xi}, & \xi \neq 0, \\ \psi^{-1} \exp(-x/\psi), & \xi = 0, \end{cases} \tag{3}$$

where $\xi \in \mathbb{R}$ is called the *extreme value index* (evi) and $\psi > 0$ is a scale parameter, $0 \leq x \leq -\psi/\xi$ if $\xi < 0$, and $x \geq 0$ if $\xi \geq 0$. The value of $\xi$ determines the tail type. If $\xi < 0$, we say that the distribution is *light tailed*, if $\xi = 0$ we say it is *exponential tailed*. If $\xi > 0$ a GPD has finite moments of order $n$ if $\xi < 1/n$ and it is called *heavy tailed*. The mean of a GPD is $\psi/(1 - \xi)$ and the variance is $\psi^2/[(1 - \xi)^2(1 - 2\xi)]$ provided $\xi < 1$ and $\xi < 1/2$, respectively. Then, the coefficient of variation is

$$c_\xi = \sqrt{1/(1 - 2\xi)}, \tag{4}$$

the `cvevi` and `evicv` functions of the R package **ercv** correspond to this function and its inverse.

The residual distribution of a GPD is again GPD with the same extreme value index $\xi$, for any threshold $t > 0$, in fact

$$\mathrm{GPD}_t(\xi, \psi) = \mathrm{GPD}(\xi, \psi + \xi t). \tag{5}$$

Therefore, the residual CV for GPD is independent of the threshold and the scale parameter and is given by equation (4).

The probability density functions (3) are monotone decreasing (L-shaped) for $\xi > -1$, covering practically all the applications. Therefore, we are mainly concerned with the subset of data that indicate this behaviour. For example, if the dataset is concentrated in the centre and decreases on either side (bell-shaped) we will study the upper and lower part (changed sign) of the distribution separately, taking the median or some other location statistic as the origin.

### The power law distribution and GPD

The power law distribution is the model, introduced by Pareto,

$$p\,(x; \alpha, \sigma) = \frac{\alpha}{\sigma}\left(\frac{\sigma}{x}\right)^{\alpha+1}, \quad x > \sigma \tag{6}$$

where $\alpha > 0$ is the tail index and $\sigma > 0$ the minimum value parameter. The model corresponds to the distribution functions $F$ with the linear relation

$$log\,[1 - F\,(x)] = -\alpha\, log(x) + \alpha\, log\,(\sigma)\,, \tag{7}$$

see also Gillespie (2015).

Note that if $X$ is a r.v. with probability density function $p(x;\alpha,\sigma)$, given by (6), $Z = X - \sigma$ has probability density function

$$g(z;1/\alpha,\sigma/\alpha) = \frac{\alpha}{\sigma}\left(\frac{\sigma}{z+\sigma}\right)^{\alpha+1}, \quad z > 0, \tag{8}$$

that is, there is a one to one correspondence between power law distributions and GPD distributions with heavy tails ($\xi > 0$), where $\xi = 1/\alpha$ and $\sigma = \psi/\xi$. However, the two statistical models (3) and (6), with $\xi > 0$, are different since there is no unique transformation for all functions of the model (the transformation $Z = X - \sigma$ depends on the minimum value parameter $\sigma$ of the same variable $X$).

The MLE for model (6) leads to the Hill estimator and Hill-plot (`hill` function in **evir** R package). The support of the distributions in (6) depends on the minimum value parameter $\sigma$. Hence, the MLE has no standard regularity conditions and the minimum value parameter $\sigma$ is estimated with alternative methods, see Clauset et al. (2009) and its implementation in the **poweRlaw** R package by Gillespie (2015).

However, the support of the distributions in (3), with $\xi > 0$, does not depend on parameters and MLE existing for large samples provided $\xi > -1$ and is asymptotically efficient provided $\xi > -0.5$, see del Castillo and Serra (2015) and the references therein for details. The `gdp` function in the **evir** R package provides the MLE for (3).

Note that model (3) includes all the limit distributions (heavy or not) of the residual distribution over a high threshold and comes from a mathematical result (the PBdH theorem) and often (6) comes from empirical evidence of the linear relationship (7) and comparison with other models. Moreover, the linear relationship (7) is also obtained from the relationship between the parameters (8), see the `ccdfplot` function in Section Fitting PoT parameters and tail plots (fitpot ccdfplot).

### The residual CV approach

Gupta and Kirmani (2000) show that the residual CV characterizes the distribution in univariate and bivariate cases, provided threre is a finite second moment ($\xi < 1/2$). In the case of GPD, the residual CV is constant and is a one to one transformation of the extreme value index suggesting its use to estimate this index. The residual CV can also be expressed in terms of probabilities, rather than the threshold, through the inverse of the distribution function or the *quantile function* defined by $Q(p) = inf\,[x : F(x) \geq p]$, then the CV can be drawn, for $0 \leq p < 1$, for the threshold $t = Q(p)$, that is to plot the function $p \to CV(Q(p))$. This representation makes it possible to draw on the same scale for the $x$ axis the residual CV of distributions with different supports.

## Exploratory data analysis with `cvplot` function

In this section the `cvplot` function of the R package **ercv** is introduced as a graphical tool for use in a exploratory data analysis, through the nine examples described in Section 3.2. The `cvplot` function is essentially the empirical residual CV whose asymptotic distribution as a stochastic process is explained by del Castillo et al. (2014) and del Castillo and Padilla (2016).

### The empirical residual CV and confidence intervals.

Assume that the raw data consist of a sequence of independent and identically distributed measurements $x_1, ..., x_n$. Extreme events are identified by defining a high threshold $t$ for which the *exceedances* are $\{x_j : x_j > t\}$. Hence, we first identify a threshold $t$ such that its exceedances correspond to a constant residual CV (equivalently a GPD). We denote the ordered sample $x_{(1)} \leq x_{(2)} \leq \cdots \leq x_{(n)}$. The `cvplot` function provides the function $cv(t)$ of the sample coefficient of variation of the *threshold excesses* $(x_j - t)$ given by

$$t \to cv(t) = sd\{x_j - t \mid x_j > t\}/\text{mean}\{x_j - t \mid x_j > t\}, \tag{9}$$

in practice $t = x_{(k)}$ are the order statistics, where, $k$ $(1 \leq k \leq n)$ is the size of the sub-sample excluded. Hereinafter the graph of this function is called CV-plot. Figure 1 shows the CV-plots of nine examples (blue lines) that we comment on the next section.

Point-wise error limits for $cv(t)$ under GPD$(\xi,\psi)$ (provided $\xi < 1/4$) follow from the asymptotic distribution of the empirical residual CV, by del Castillo and Padilla (2016), in particular for a fixed

threshold $t$, the asymptotic confidence intervals in Figure 1 (solid orange lines) are obtained by

$$\sqrt{n(t)}(cv(t) - c_\xi) \xrightarrow{d} N(0, \sigma_\xi^2), \tag{10}$$

where $c_\xi$ is in (4), $n(t) = \sum_{j=1}^{n} 1_{(x_j > t)}$ . For an exponential distribution $(\xi = 0)$ , $c_0 = 1$ and $\sigma_0^2 = 1$, and for a uniform distribution $(\xi = -1)$ , $c_{-1} = 1/\sqrt{3}$ and $\sigma_{-1}^2 = 8/45$.

By default, if $\sqrt{2}$ is in the range of $y$'s then the `cvplot` function draws the line $y = \sqrt{2}$ (black dotted line), which corresponds to $\xi = 1/4$ (finite fourth moment). Hence, CV-plot larger than this value for high thresholds lead to very heavy tailed distribution and we suggest to switch to transformed data through function `tdata` (Section 2.6). Alternatively, finite moments can be checked by a confidence interval for the MLE estimator of evi, or the methods in the R package **RobExtremes** (Ruckdeschel et al., 2019) and the references cited therein can be used.

The CV-plot is an alternative tool to Hill-plot an to ME-plot. It has two advantages over ME-plot: first, it depends on a scale parameter and CV-plot does not; second, linear functions are defined by two parameters and the constants by only one. So the uncertainty is reduced from three to one single parameter. On the other hand, the Hill-plot can only be used for heavy tailed distributions.

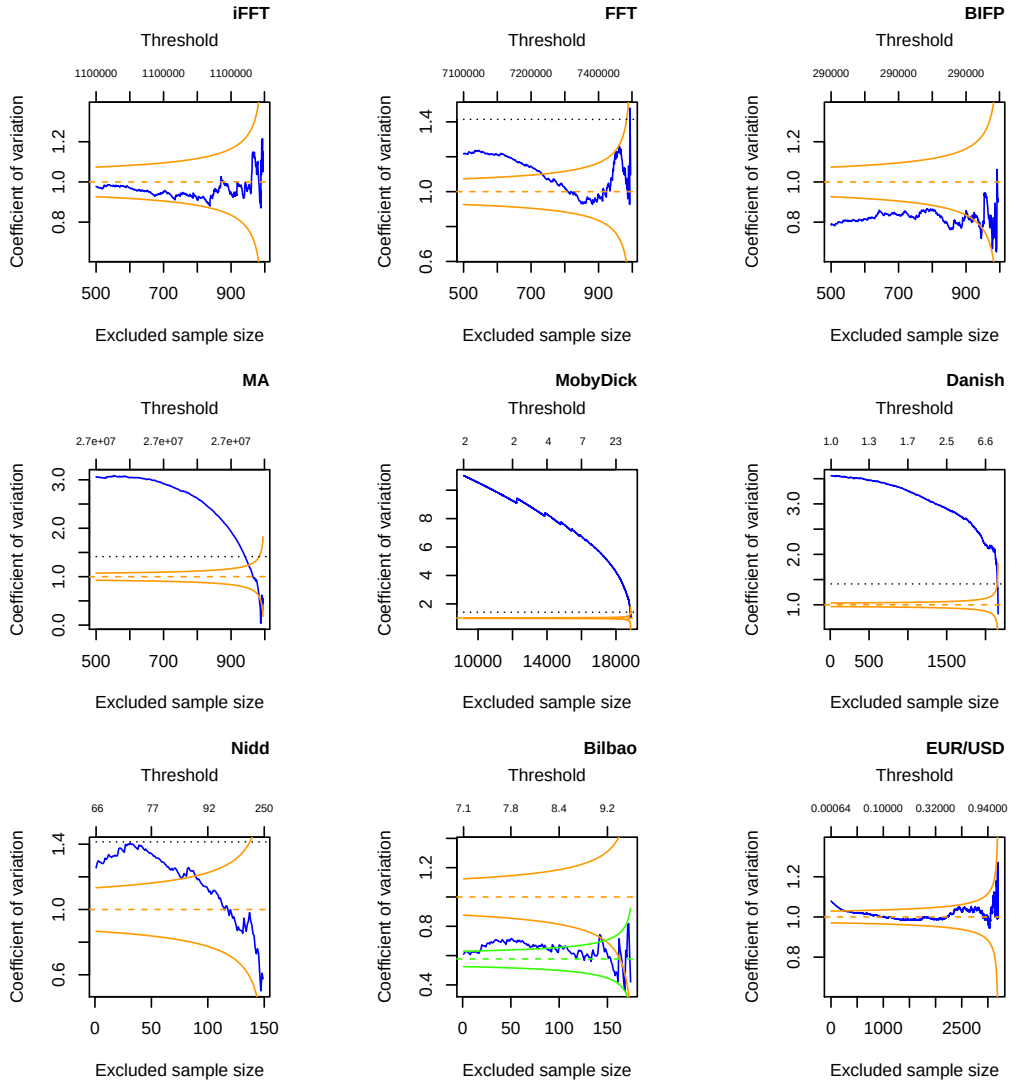


**Figure 1:** CV-plots stowed from left to right and top to bottom: four different types of execution time distributions of automotive applications, the frequency of words in the novel Moby Dick, Danish fire insurance data, River Nidd exceedances above value 65, Bilbao waves dataset and positive daily returns of euro/dollar exchange rates between 1999 and 2014.

### Examples

The use of the `cvplot` function and its options is described using nine examples. The first four (iFFT, FFT, BIFP and MA) correspond to different types of execution time distributions observed for a set of representative programs for the analysis of automotive applications. Three others are in R packages: MobyDick ("moby" in R package **poweRlaw**), Danish and Nidd ("danish" and '"nidd.thresh" in the R package **evir**). The Bilbao waves dataset (bilbao) was originally analysed by Castillo and Hadi (1997). EURUSD is the dataset of euro/dollar daily exchange rates between 1999 and 2016.

We collect samples with $n = 1,000$ observations for 4 of the 16 benchmarks in the EEMBC AutoBench suite (Poovey, 2007), which is a well-known suite for real-time systems that includes a number of programs used in embedded automotive systems. Hereinafter, these datasets will be called iFFT (idctrn), FFT (aifftr), BIFP (basefp) and MA (matrix), leaving the real names in parentheses, they correspond respectively to *Inverse Fast Fourier Transform, Fast Fourier Transform, Basic Integer and Floating Point* and *Matrix Arithmetic*, see Abella et al. (2017) and del Castillo et al. (2017b). The histograms of the four datasets are bell-shaped. Hence, when searching for L-shaped distributions, we start the exploratory data analysis of the upper part of the distribution by taking the median as origin. Note also that large samples increase the precision of the estimates, provided that the fitted model is validated. The CV-plots for these four datasets are obtained, for instance, with:

```
library("ercv")
data(iFFT)
cvplot(iFFT,thr=median(iFFT))
```

The plots in Figure 1 are stowed from left to right and top to bottom. For iFFT, the CV-plot is inside the confidence interval of the exponential distribution ($evi = 0$). Hence, it can be assumed that the CV is constant equal to 1 (dashed orange line). For FFT, the CV-plot is inside the confidence interval for the last 250 observations. For BIFP, the CV-plot looks like a constant with CV lower than 1, hence a light tailed GPD is suggested. For MA, the CV-plot suggests a heavy tailed distribution.

The following three CV-plots in Figure 1 are made from the MobyDick, Danish and Nidd datasets, which can be directly loaded from the R packages. The three plots are made with the default `cvplot` function options, but including title, for instance:

```
data("moby", package = "poweRlaw")
cvplot(moby,main="MobyDick")
```

The second row of Figure 1 shows three examples that suggest heavy tailed distributions. In the centre is MobyDick and on the right is the Danish fire insurance dataset, which is a highly heavy-tailed infinite-variance example used to illustrate the basic ideas of extreme value theory, see Embrechts et al. (1997), McNeil et al. (2005, Example 7.23) and Novak (2012, Example 9.8). Section Transformation from heavy to light tails (tdata) shows how to analyse these examples, with the `tdata` function, using the methodology developed in del Castillo and Padilla (2016).

Nidd is the dataset of high levels of the River Nidd above a threshold value of 65. Its CV-plot is always lower than $\sqrt{2}$, begins in the area of heavy tails and goes into the confidence interval of exponentially. The Bilbao waves dataset was originally analysed by Castillo and Hadi (1997). The Nidd and Bilbao datasets are two of the most commented examples of extreme values theory, which were also analysed by del Castillo and Serra (2015) from the MLE point of view.

By default, the `cvplot` function draws a 90% confidence interval of CV-plot from exponential distribution ($evi = 0$). The *evi* parameter of the function provides confidence intervals of the corresponding GPD ($evi < 1/4$). The *conf.level* parameter allows for changing confidence levels. Both *evi* and *conf.level* may be a vector. For light tailed distributions, as is presumably the case with the wave levels, it is also advisable to draw a confidence interval from the uniform distribution ($evi = -1$). Hence, the Bilbao CV-plot in Figure 1 has confidence intervals for exponential (orange) and uniform (green) distributions.

```
data(bilbao)
cvplot(bilbao,evi = c(0,-1),main="Bilbao")
```

EURUSD is the data frame object of the euro/dollar daily exchange rates between 1999 and 2016, including the financial crisis of 2007-08, which was obtained from the R package **quantmod** (Ryan, 2016). Various parts of the EURUSD series have been studied by several authors, see Gomes and Pestana (2007) and del Castillo and Padilla (2016). The last plot in Figure 1 shows the CV-plot of the positive log-returns of the euro/dollar daily prices, obtained from

```
data("EURUSD")
prices<-ts(EURUSD$EUR.USD,frequency=365,start=1999)
```

```
#plot(prices,col="blue",main="euro/dollar daily prices(1999-2016)")
return <- 100*diff(log(prices));
pos.return <- subset(return, return >0);
cvplot(pos.return,main="pos.returns EUR/USD 1999-2016")
```

The dynamics of the daily return can be described by a GARCH(1,1) model. One might then hope that for sufficiently high values of $t$ the subset of daily returns that are above $t$ is so well separated in time that independence can reasonably be assumed. Then, the CV-plot clearly shows that the tail of the distribution looks like an exponential.

## Estimation and Model diagnostics with `Tm` function

Following the exploratory analysis, we would like to confirm or deny some of the previous observations. It is known that in order to make optimum decisions, it is necessary to quantify the uncertainty of information extracted from data. Statistics provides mechanisms to ensure a controlled probability of error, but there is always the risk of misuse for multiple testing, especially in EVT where quite small changes can be greatly magnified on extrapolation. The asymptotic distribution of the residual coefficient of variation for GPD as a random process indexed by the threshold by del Castillo and Padilla (2016) provides pointwise error limits for CV-plot, used in the last section, and a *multiple thresholds test* that truly reduces the multiple testing problem, hence, the $p$-values are clearly defined.

Using the building blocks given by (10) the multiple threshold test $T_m$ (the `Tm` function of the R package **ercv**) for a (supplementary) number of thresholds $m$ as large as necessary for practical applications is constructed from

$$T_m(\xi) = n \sum_{k=0}^{m} p^k (cv(q_k) - c_\xi)^2, \tag{11}$$

where $c_\xi$ is in (4), $q_k$ are the empirical quantiles corresponding to probabilities $1 - p^k$ and probability $p$ is chosen so that $n\, p^m \approx omit$, where $omit$ is the smaller sample size used to calculate CV. This statistic can be used to test whether a sample is distributed as a GPD with parameter $\xi$.

The `Tm` function makes it possible to see whether the 75 largest values of Nidd can be assumed to be exponentially distributed.

```
data("nidd.thresh",package = "evir")
Tm(nidd.thresh,evi=0, nextremes = 75)

nextremes   cvopt   evi    tms   pvalue
       75   1.000   0.000  0.981 0.310
```

The `Tm` function provides $tms = Tm(evi)/(m+1)$, which is stable on vary the number of thresholds $m$, the $p$-value says that it can not be rejected exponenciality (the number of simulations can be increased with *nsim*). Moreover, by default the `Tm` function assumes that the parameter $\xi$ is unknown ($evi = NA$), then the *cvopt* is estimated as the value $\widetilde{c}_{\widetilde{\xi}}$ such that achieves the minimum of $T_m(\xi)$, and reversing (4) provides an estimator $\widetilde{\xi}$.

The following code shows that the assumption of constant CV (GDP) is rejected for the complete sample.

```
Tm(nidd.thresh)

nextremes   cvopt  evi    tms   pvalue
      154   1.225  0.167  1.214 0.030
```

It is rejected that Bilbao dataset is uniform distributed. However, It can not be rejected GPD as the following code shows

```
Tm(bilbao,evi=-1,nsim=1000)

nextremes   cvopt     evi    tms     pvalue
      179   0.577   -1.000   0.629   0.003

Tm(bilbao,nsim=1000)

nextremes   cvopt     evi    tms     pvalue
      179   0.650   -0.685   0.254   0.172
```

The confidence interval for the parameter estimation $evi = -0.685$ can be obtained with

```
cievi(nextremes=length(bilbao),evi=-0.685)

   5%     95%
-0.778 -0.549
```

Using a small threshold, $(0.1\%)$, the `Tm` function shows that the positive and negative returns of the euro/dollar between 1999 and 2016 can be assumed exponentially distributed.

```
Tm(pos.return,m=50,evi=0,thr=0.1,nsim=1000)

nextremes   cvopt   evi   tms    pvalue
     2207   1.000  0.000  0.392  0.780

neg.return <- -subset(return, return <0);
Tm(neg.return,m=50,evi=0,thr=0.1,nsim=1000)

nextremes   cvopt   evi   tms    pvalue
     2187   1.000  0.000  1.160  0.231
```

The last statement with Intel(R) Core(TM) i5-6200U CPU @ 2.30GHz takes *elapsed=4.73* (R> proc.time()).

## Threshold selection algorithm (`thrselect`)

There are two different approaches to the question of threshold choice. The first approach is to regard the free choice of the threshold as an advantageous feature of the procedure. By varying the threshold, the data can be explored, and if a single estimate is needed it can be obtained by subjective choice. It may well be that such a subjective approach is in reality the most useful one.

The other, to some extend opposing, view is that there is a need for an automatic method whereby the threshold is chosen by the data. It is fairer to use the word automatic rather than objective for such a method, because there are arbitrary decisions involved in the choice of the method itself. Nevertheless, it is of course the case that conditional on the automatic method being used, the threshold is indeed objective. Automatic methods need not be used in an uncritical way; they can of course be used as a starting point for fine tuning.

The `thrselect` function in the R package **ercv** starts with the $T_m(\xi)$ calculation (11) where the number of thresholds $m$ must be fixed by the researcher. This determines the thresholds where the CV is calculated, $0 = q_0 < q_1 < \cdots < q_m$, which are fixed throughout the procedure. We accept or reject the null hypothesis for the shape parameter using all the thresholds. If the hypothesis is rejected, the threshold excesses $\left(x_j - q_1\right)$ are calculated for the sub-sample $\left\{x_j > q_1\right\}$. The previous steps are repeated, but removing one threshold, to accept or reject the null hypothesis that the sample comes from a GPD with parameter $\xi$, see del Castillo and Padilla (2016).

If we apply the function `thrselect` on the Nidd dataset the code shows

```
DF <- thrselect(nidd.thresh,m=10, nsim=1000)

    m nextremes threshold   rcv   cvopt   evi     tms    pvalue
5   6        63     87.85   1.193 1.073   0.0656  0.408  0.102
```

This means that the algorithm need 5 steps to achieve a $p$-value larger than 0.10 and it is using in this step $m = 6$ thresholds. Then, constant CV can be accepted for the last 63 extremes over the threshold 87.85, with the CV $cvopt = 1.0728$ and the corresponding $evi = 0.0656$.

The output of `thrselect` is in the data frame DF, the printed values are in DF$solution and DF$options provides complementary information that can be used for a more personal approach.

```
print(DF$options,digits=4)

    m nextremes threshold   rcv  cvopt      evi     tms    pvalue
1  10       154     65.08 1.2486 1.2249  0.166758 1.33553  0.023
2   9       123     74.38 1.4082 1.2183  0.163112 1.47158  0.012
3   8        99     77.80 1.3163 1.1634  0.130594 0.93927  0.034
4   7        79     81.40 1.2587 1.1175  0.099606 0.64548  0.064
5   6        63     87.85 1.1933 1.0728  0.065559 0.40795  0.102
```

```
6   5        50     92.82 1.1328 1.0320  0.030493 0.24415  0.217
7   4        40     99.14 1.0714 0.9945 -0.005584 0.12917  0.457
8   3        32    107.94 1.0054 0.9619 -0.040406 0.05888  0.609
9   2        26    115.93 0.9006 0.9396 -0.066323 0.04218  0.637
10  1        21    131.87 0.9473 0.9667 -0.034986 0.01755  0.597
```

## Transformation from heavy to light tails (tdata)

It is possible to extend the previous methodology based on CV to all distributions, even without finite moments. For CV-plots above the straight line $y = \sqrt{2}$ , like the three examples in the second row of Figure 1, the datasets are transformed by the strictly increasing function that applies $(0, \infty)$ to $(0, \sigma)$,

$$y(x) = \sigma x/(x + \sigma),$$

where $\sigma > 0$, using the tdata function in the R package **ercv**.

This technique is founded on the following result: if $X$ is a random variable $GPD(\xi, \psi)$ distributed and $\xi > 0$, then for $\sigma = \psi/\xi$ the transformed random variable $Y = y(X)$ is $GPD(-\xi, \psi)$ distributed. Furthermore, the converse is also true, as evidenced by applying the inverse transformation $x(y) = \sigma y/(\sigma - y)$, see also del Castillo and Padilla (2016). The $\sigma > 0$ parameter is estimated by tdata, using MLE with the internal function egpd, (see del Castillo and Serra (2015)) or may be provided by the researcher as a preliminary estimate.

The CV-plots for Danish and MobyDick transformed by tdata function are obtained with:

```
data("danish",package = "evir")
tdanish<- tdata(danish)
cvplot(tdanish,main="transformed Danish")
tmoby<- tdata(moby)
cvplot(tmoby,main="transformed MobyDick")
```

The CV-plots in Figure 2 for the transformed datasets are more stable than the original CV-plots in Figure 1 and actually look light tailed. The CV-plot of the transformed MobyDick has a sawtooth profile because the original dataset only takes positive integer values and the smaller values have a high frequency (among the 18,855 values, 1 appears 9,161 times, 2 appears 3,085, ... ). In order to use a GPD approach for this example we assume that the data correspond to positive values rounded to the nearest integer.

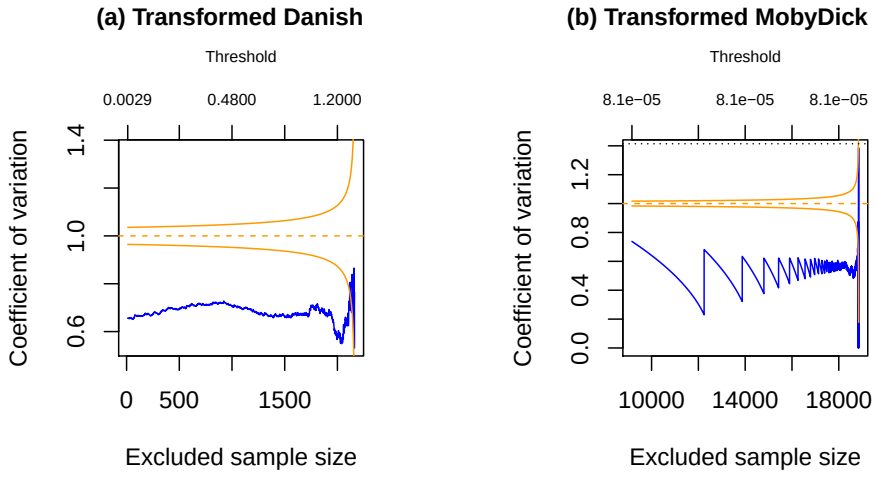


**Figure 2:** CV-plots under *tdata* transformation of Danish fire insurance data and frequencies of words in the novel Moby Dick.

The Tm function rejects GPD for the complete transformation of MobyDick. The same result is obtained with the transformation of the dataset on the thresholds 2 and 3. However, GPD is not rejected on threshold 4, hence the frequencies of words that appear four or more times in the novel Moby Dick (4,980 observations) can be approximated by a GPD distribution with $evi = 0.982$, as the following code shows (changing the sign of *evi*):

```
t4moby<-tdata(moby,thr=4)
Tm(t4moby,m=50,nsim=1000)

nextremes   cvopt    evi     tms    pvalue
     4980   0.581   -0.982   0.198  0.293
```

The Danish example was studied by del Castillo and Padilla (2016). The results obtained are validated by the `Tm` function after the transformation `tdata`

```
Tm(tdanish,m=20,nextremes = 951,omit = 8, nsim = 1000)

nextremes   cvopt     evi     tms      pvalue
      951   0.676    -0.595   0.256    0.253
```

Applying the `thrselect` function to Danish after the transformation by `tdata` we obtain

```
DF<-thrselect(tdanish,m=30,nsim=1000)

    m  nextremes  threshold  rcv    cvopt     evi    tms     pvalue
19 12    116      1.283      0.589  0.6747   -0.598  0.265   0.11
```

The automatic algorithm chooses the threshold 1.283 (116 extremes) with the estimate $evi = 0.598$ (changing the sign of *evi*) really close to the previous one $evi = 0.595$. The result is different from that obtained by McNeil et al. (2005) by MLE $evi = 0.50$ (109 extremes). However the *cievi* function shows that $evi = 0.50$ can not be rejected, as shown by the confidence interval provided by the following code (changing the sign of *evi* again),

```
cievi(116,evi=-0.596)

    5%    95%
-0.714 -0.440
```

In the next section we will discuss these results with new features of the R package **ercv**.

## Fitting PoT parameters and tail plots (`fitpot` `ccdfplot`)

The tools described in the previous sections provide an asymptotic model for threshold exceedances over a high quantile, the so-called *peak-over-threshold* (PoT) method, see McNeil et al. (2005). The PoT method is based on determining a high enough threshold from which the distribution of the observations above this value, adjusted to zero, approaches to a GPD distribution. Then, given a threshold $t$, for $x > t$ the *complementary cumulative distribution function* (ccdf) is estimated by

$$1 - \hat{F}(x) = \hat{p}_t \left(1 - G\left(x - t; \hat{\xi}_t, \hat{\psi}_t\right)\right) \tag{12}$$

where $G(x; \xi, \psi)$ is the cumulative distribution function of the GDP, whose probability density function was introduced in (3), and $\left(\hat{\xi}_t, \hat{\psi}_t\right)$ are their estimated parameters for the $n_t$ threshold exceedances over $t$ adjusted to zero, from a sample of size $n$ with $\hat{p}_t = n_t/n$. Alternatively, given $n_t$ the estimated parameter is $t$.

The `ppot` function is the cumulative distribution function for the PoT method. That is, given an estimate of the four parameters in (12), $\left(\hat{\xi}, \hat{\psi}, \hat{t}, \hat{p}\right)$, the right hand part of (12) is provided by $1-\texttt{ppot}\left(x, \left(\hat{\xi}, \hat{\psi}, \hat{t}, \hat{p}\right)\right)$. The `qpot` is the quantile function for the PoT method that assigns to each probability $p$ attained by `ppot` the value $x$ for which $\texttt{ppot}(x) = p$, given the same vector of four parameters. The `qpot` function can be used in the estimation of high quantiles, that in terms of risk is expressed as the *value at risk* (VaR). For a small $p$, $VaR_p = q$ if and only if $1 - F(q) = p$. Hence, if $\varepsilon < \hat{p}$,

$$VaR_\epsilon = \hat{t} + qpot\left(\left(1 - \varepsilon/\hat{p}\right), \left(\hat{\xi}, \hat{\psi}, \hat{t}, \hat{p}\right)\right).$$

The `fitpot` function of the R package **ercv** provides an estimate of the four parameters in (12) that allow approximating the empirical cumulative distribution function of a dataset. It is assumed that the threshold $t$, or the number of extremes, has been chosen based on the tools of the previous sections. By default `fitpot` uses MLE. However, since parameter $\xi$ (evi) can be estimated minimizing (11) by the `Tm` function, this value can be entered into the function `fitpot` and then it uses MLE by the restricted model to a single parameter. From now on this method of estimation will be called CV method.

The two methods of estimation of `fitpot` applied to Danish explain the differences between the results obtained by us and by other researchers, which we have discussed in the previous section, as we can see with the code

```
fit1<-fitpot(danish,nextremes =116);fit1                #MLE

  evi       psi     threshold   prob
0.446      7.462    9.200      0.054

fit2<-fitpot(danish,evi=0.598,nextremes =116);fit2      #CV

  evi       psi     threshold   prob
0.598      6.450    9.200      0.054
```

Naturally, different estimation methods provide different estimates, but the question of identifying the best approach still remains. To clarify this point, we can use the `ccdfplot` function, which draws the empirical complementary cumulative distribution function with the approximations provided by the parameters estimated by `fitpot`. The `ccdfplot` function allows to draw several approaches at several scales. The approximation is linear in the log-log scale for datasets with heavy tails, although it is linear in log scale for datasets with exponential tails (log = "y", by default). To draw the approach on natural scale the option *log* = "" has to be used.

The plots of Figure 3 have been obtained with `ccdfplot` function applied to Danish data with the estimates obtained by MLE (orange) and CV method (green) on logarithmic and double logarithmic scales, with

```
ccdfplot(danish,pars=list(fit1,fit2),main="Danish (log scale)")
ccdfplot(danish,pars=list(fit1,fit2),log="xy",main="Danish (log-log
scale)")
```

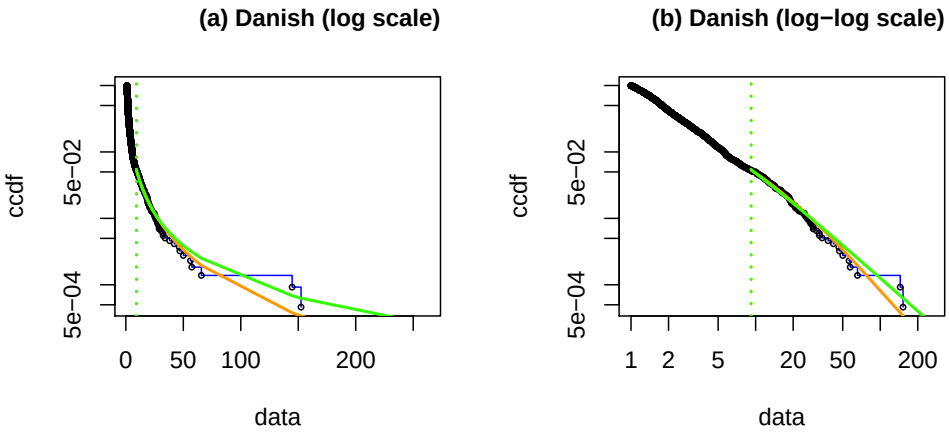


**Figure 3:** Complementary cumulative distribution function of Danish fire insurance data adjusted by MLE and CV methods. in log scale and log-log scale.

Figure 3 shows that both adjustments are reasonable. The CV method is not worse than MLE, perhaps less optimistic or more realistic. The previous PoT approach can be validated using the Clauset et al. (2009) point of view.

Based on the four parameters estimated by `fitpot` $(\hat{\xi}, \hat{\psi}, \hat{t}, \hat{p})$ for heavy tailed models (evi > 0), the linear relationship (7) can be obtained for the dataset values over the threshold, with the new threshold $\hat{\sigma} = \hat{\psi}/\hat{\xi}$ and the probability 1, see the following code.

```
fit1<-as.numeric(fit1$coeff);sg1<- fit1[2]/fit1[1];sg1
fit2<-as.numeric(fit2$coeff);sg2<- fit2[2]/fit2[1];sg2
exDanish<-danish[danish>fit1[3]]-fit1[3]  #origin to zero
exDanish1<- exDanish+sg1                  #origin to sg1
exDanish2<-exDanish+sg2                   #origin to sg2
exfit1<-c(fit1[1],fit1[2],sg1,1)
exfit2<-c(fit2[1],fit2[2],sg2,1)
ccdfplot(exDanish, pars=c(exfit1),log="xy",main="adjusted by MLE")
ccdfplot(exDanish2, pars=c(exfit2),log="xy",main="adjusted by CV")
```

The Figure 4 plot (a) shows the linear relationship (7) for the 116 upper extremes of Danish adjusted by MLE. Changing the previous *fit1* by *fit2* the linear relationship is obtained by the CV method and is

shown in plot (b). Notice that the linear relationship (7) begins at the threshold $sg1 = 16.727$ for MLE and at a threshold $sg2 = 10.787$ for the CV method, so we can not overlay them in the same graph. The goodness of fit can now be measured by the correlation between the logarithm of the complementary empirical distribution function, $log(1 - F_n)$ and the logarithm of the data, $\log(x + sg)$, where $(x + sg)$ are the 116 upper extremes of Danish, adjusted to sigma. The results are $correlation = -0.981$ using MLE, plot (a), and $correlation = -0.990$ using CV-method, plot (b).

We can also calculate the threshold $th$ having a maximum correlation between $log(1 - F_n)$ and $log(x + th)$, obtaining $th = 6.996$ and $correlation = -0.992$. Thus, the correlation on which the goodness of the CV method adjustment is based on is very close to the best that can be obtained by this procedure, which is in line with Clauset et al. (2009) and the **poweRlaw** R package by Gillespie (2015) (although here the estimation of evi is different). This shows that the methodology provided by the R package **ercv** complements and connects the contributions of **evir** (Pfaff and McNeil, 2012) and **poweRlaw** by Gillespie (2015).

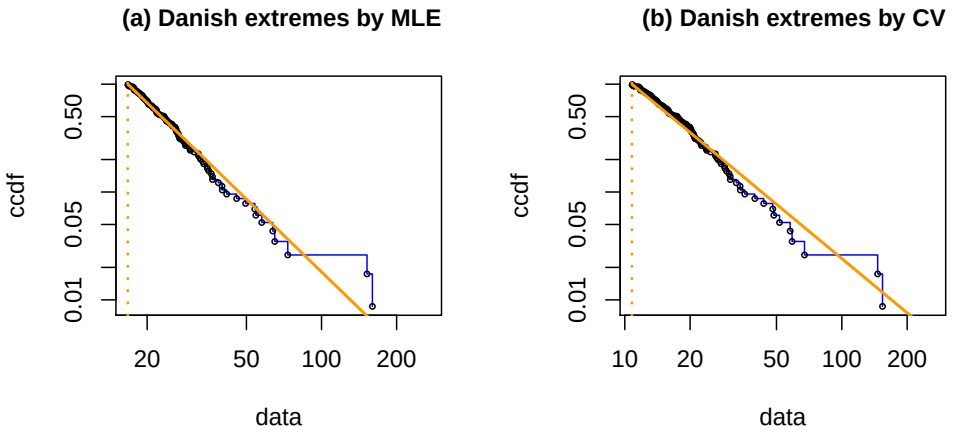


**Figure 4:** The linear relationship for the 116 upper extremes of Danish fire insurance data adjusted by MLE and CV method.

## Acknowledgements

This work was supported by the Spanish Ministry of Economy and Competitiveness under Grant: Statistical modelling of environmental, technological and health risks, MTM2015-69493-R. David Moriña acknowledges financial support from the Spanish Ministry of Economy and Competitiveness, through the María de Maeztu Programme for Units of Excellence in R&D (MDM-2014-0445) and from Fundación Santander Universidades.

## Bibliography

J. Abella, M. Padilla, J. del Castillo, and F. Cazorla. Measurement-based worst-case execution time estimation using the coefficient of variation. *ACM Transactions on Design Automation of Electronic Systems (TODAES)*, 22(4):72:1–72:29, 2017. [p60]

A. Balkema and L. de Haan. Residual life time at great age. *Annals of Probability*, 2:792–804, 1974. [p57]

J. Beirlant, Y. Goegebeur, J. Segers, and J. L. Teugels. *Statistics of Extremes: Theory and Applications*. John Wiley & Sons, Chichester, UK, 2006. [p56]

E. Castillo and A. S. Hadi. Fitting the generalized pareto distribution to data. *Journal of the American Statistical Association*, 92:1609 – 1620, 1997. [p60]

A. Clauset, C. R. Shalizi, and M. Newman. Power-law distributions in empirical data. *SIAM Review*, 51(4):661–703, 2009. [p56, 58, 65, 66]

S. Coles. *An Introduction to Statistical of Extremes Values*. Springer-Verlag, London, 2001. [p56, 57]

L. de Haan and A. Ferreira. *Extreme Value Theory: An Introduction*. Springer-Verlag, New York, 2007. [p56]

J. del Castillo and M. Padilla. Modelling extreme values by the residual coefficient of variation. *Statistics and Operations Research Transactions*, 40:303–320, 2016. [p56, 58, 60, 61, 62, 63, 64]

J. del Castillo and I. Serra. Likelihood inference for generalized pareto distribution. *Computational Statistics & Data Analysis*, 83:116–128, 2015. [p56, 58, 60, 63]

J. del Castillo, J. Daoudi, and R. Lockhart. Methods to distinguish between polynomial and exponential tails. *Scandinavian Journal of Statistics*, 41:382–393, 2014. URL https://doi.org/10.1111/sjos.12037. [p56, 58]

J. del Castillo, D. Moriña, and I. Serra. **ercv***: Fitting Tails by the Empirical Residual Coefficient of Variation*, 2017a. URL https://CRAN.R-project.org/package=ercv. R package version 1.0.0. [p56]

J. del Castillo, M. Padilla, J. Abella, and F. Cazorla. Execution time distributions in embedded safety-critical systems using extreme value theory. *International Journal of Systems Control and Information Processing*, 9(4):348–361, 2017b. [p60]

P. Embrechts, C. Klüppelberg, and T. Mikosch. *Modelling Extremal Events for Insurance and Finance*. Springer-Verlag, Berlin, 1997. [p56, 60]

R. Fisher and L. Tippett. Limiting forms of the frequency distribution of the largest or smallest member of a sample. *Mathematical Proceedings of the Cambridge Philosophical Society*, 24(2):180–190, 1928. [p57]

E. Gilleland, M. Ribatet, and A. Stephenson. A software review for extreme value analysis. *Extremes*, 16:103–119, 2013. [p56]

C. Gillespie. Fitting heavy tailed distributions: The powerlaw package. *Journal of Statistical Software*, 64(2):1–16, 2015. [p56, 58, 66]

B. V. Gnedenko. Sur la distribution limite du terme maximum d'une serie aleatoire. *The Annals of Mathematics*, 44(3):423–453, 1943. [p57]

M. I. Gomes and D. Pestana. A sturdy reduced-bias extreme quantile (var) estimator. *Journal of the American Statistical Association*, 102:280–292, 2007. [p60]

R. Gupta and S. N. U. A. Kirmani. Residual coefficient of variation and some characterization results. *Journal of Statistical Planning and Inference*, 91:23–31, 2000. [p58]

J. E. Heffernan and A. G. Stephenson. *Ismev: An Introduction to Statistical Modeling of Extreme Values*, 2018. URL https://CRAN.R-project.org/package=ismev. [p56]

R. Leadbetter, G. Lindgren, and H. Rootzén. *Extremes and Related Properties of Random Sequences and Processes*. Springer-Verlag, New York, 1983. [p56]

N. Markovich. *Nonparametric Analysis of Univariate Heavy-Tailed Data*. John Wiley & Sons, Chichester, UK, 2007. [p56]

A. McNeil, R. Frey, and P. Embrechts. *Quantitative Risk Management: Concepts, Techniques and Tools*. Princeton Series in Finance, New Jersey, 2005. [p57, 60, 64]

S. Novak. *Extreme Value Methods with Applications to Finance*. CRC Press, Boca Raton, 2012. [p56, 60]

B. Pfaff and A. McNeil. **evir***: Extreme Values in R*, 2012. URL https://CRAN.R-project.org/package=evir. R package version 1.7-3. [p56, 66]

J. Pickands. Statistical inference using extreme order statistics. *The Annals of Statistics*, 3:119–131, 1975. [p57]

J. Poovey. *Characterization of the EEMBC Benchmark Suite*. North Carolina State University, Raleigh, NC, 2007. [p60]

R Core Team. *R: A Language and Environment for Statistical Computing*. R Foundation for Statistical Computing, Vienna, Austria, 2017. URL https://www.R-project.org/. [p56]

S. Resnick. *Extreme Values,Regular Variation,and Point Processes*. Springer-Verlag, New York, 2013. [p56]

P. Ruckdeschel, M. Kohl, and N. Horbenko. *RobExtremes: Optimally Robust Estimation for Extreme Value Distributions*, 2019. URL http://robast.r-forge.r-project.org/. Contributions by S. Desmettre, G. Kroisandt, E. Massini, D. Pupashenko and B. Spangl; R package version 1.2.0. [p59]

J. A. Ryan. **quantmod***: Quantitative Financial Modelling Framework*, 2016. URL https://CRAN.R-project.org/package=quantmod. R package version 0.4-7. [p60]

*Joan del Castillo*
*Departament de Matemàtiques, Universitat Autònoma de Barcelona (UAB)*
*Edifici C, E-08193 Barcelona*
*Spain*
castillo@mat.uab.cat

*Isabel Serra*
*Centre de Recerca Matemàtica (CRM)*
*Edifici C, E-08193 Barcelona*
*Spain*
iserra@crm.cat

*Maria Padilla*
*Departament de Matemàtiques, Universitat Autònoma de Barcelona (UAB)*
*Edifici C, E-08193 Barcelona*
*Spain*
mpadilla@mat.uab.cat

*David Moriña*
*Barcelona Graduate School of Mathematics (BGSMath), Departament de Matemàtiques, Universitat Autònoma de Barcelona (UAB)*
*Edifici C, E-08193 Barcelona*
*Spain*
dmorina@mat.uab.cat