On Modelling RR Tails in Heart Rate Variability Studies: an Extreme Value Analysis

Sónia Gouveia[1,2], Manuel Scotto[3]

[1] Institute of Electronics and Informatics Engineering of Aveiro (IEETA), Univ of Aveiro, Portugal
[2] Center for R&D in Mathematics and Applications (CIDMA), University of Aveiro, Portugal
[3] CEMAT and Instituto Superior Técnico, Universidade de Lisboa, Portugal

Abstract

RR distributions with tails larger than the Gaussian have been proved to be an independent predictor of cardiac mortality in chronic heart failure patients. Within this context, extreme value theory provides a powerful tool to quantify the probability of a long RR occurrence, through the statistical characterization of the RR tail distribution. Here, tail characterization does not rely on the Gaussian assumption but by fitting the Generalized Pareto distribution (GPd) to the excesses above a properly chosen high threshold, and through the analysis of its corresponding tail index, denoted as γ. The new approach is illustrated with a 24-h RR recording from a normal subject and a Congestive Heart Failure (CHF) patient. Wavelet analysis allowed to reconstruct one signal containing the RR power traditionally related to respiratory rhythm ($\sim$ 0.25 Hz) and another to sympathetic baroreflex activity ($\sim$ 0.1 Hz). The fitted distributions for the normal subject do not reject the hypothesis of $\gamma = 0$ for both LF and HF while $\gamma > 0$ for the CHF patient. Thus, the CHF distributions are heavy-tailed, indicating a non-negligible probability that a very long RR interval can occur. In a forthcoming study, it will be assessed the impact of these preliminary findings in CHF mortality prediction.

1. Introduction

Many literature studies demonstrate that abnormal HRV measured over a 24-h period provides information on the risk of subsequent death in subjects with and without structural heart disease [1]. More recently, large deviations from a Gaussian RR distribution have been shown to be an independent predictor of cardiac death after acute myocardial infarction [2]. The RR distribution exhibits heavier tails than does the Gaussian and, thus, the probability of observing a long RR is higher than when assuming Gaussianity. Within this context, extreme value theory can be used to quantify the probability of a long RR occurrence, by means of the statistical characterization of the RR tail distribution. In this work, tail characterization does not rely on the Gaussian assumption but in fitting the Generalized Pareto distribution (GPd) to the excesses above a high threshold, and through the analysis of its corresponding tail index, γ. The GPd parameters are estimated by Peak-over-Threshold (POT) procedure and standard errors are approximated by reproducing POT in bootstrap replicates of the original exceedance values.

It is known that frequency domain analysis of HRV reflects the modulation of the autonomic nervous system (ANS) by means of the sympathetic and parasympathetic activities, through the power evaluation in low frequency (LF, 0.04-0.15 Hz) and high frequency (HF, 0.15-0.40 Hz) bands. In order to separate these activities, data was resampled at 2 Hz and wavelet analysis allowed to reconstruct one signal containing the RR power traditionally related to respiration ($\sim$ 0.25 Hz) and another to sympathetic baroreflex activity ($\sim$ 0.1 Hz). Then, the tail index was estimated separately for each of these components of 24-h HRV recordings.

2. Statistical Methods

This section provides a brief introduction to basic results on extreme value theory useful in the present stetting. In particular, the approaches to model sample maxima and exceedances of high thresholds are introduced. Furthermore, this section also describes the procedures to carry out parameters' estimation and statistical inference.

2.1. Extreme Value Theory

Let $X = (X_1, X_2, \ldots, X_n)$ be independent and identically distributed (i.i.d.) random variables (r.v's) with unknown underlying distribution function (d.f.) F and let $M_n(X) := \max(X_1, \ldots, X_n)$. If for some constants $a_n > 0$ and $b_n \in \mathbb{R}$,

$$\lim_{n\to\infty} P\left\{a_n^{-1}\left(M_n(X) - b_n\right) \leq x\right\} = G(x), \quad (1)$$

for some non-degenerate function $G(x)$, then F is in the domain of attraction of G ($F \in D(G)$, in short) and G must be the Generalized Extreme Value (GEV) distribution

$$G(x) \equiv G_\gamma(x) := \exp\left\{-\left[1+\gamma\left(\frac{x-u}{\sigma}\right)\right]^{-1/\gamma}\right\}, \quad (2)$$

for all x such that $1+\gamma(x-u)/\sigma > 0$, with location $u \in \mathbb{R}$, scale $\sigma > 0$ and shape parameter (also called tail index) $\gamma \in \mathbb{R}$. The GEV distribution has three possible forms depending on γ, namely the Weibull ($\gamma < 0$), Gumbel ($\gamma = 0$, read $G_0(x)$ as $\exp(-e^{-\frac{x-u}{\sigma}})$ for all $x \in \mathbb{R}$) and Fréchet ($\gamma > 0$) distributions. The Fréchet domain of attraction embraces heavy-tailed distributions with polynomially decaying tails, whereas all d.f's belonging to Weibull domain of attraction are light-tailed with finite right endpoint. The intermediate case $\gamma = 0$ is of particular interest not only because of the simplicity of inference within the Gumbel domain but also for the great variety of distributions ranging from moderately heavy (such as the lognormal distribution) to light (such as the Normal distribution) having finite right endpoint or not.

An alternative approach to model sample maxima, which is commonly known as the POT approach, is to considerer the excesses (or exceedances) above a sufficiently high threshold, say μ. In this case, the limiting distribution is the Generalized Pareto Distribution (GPD) defined as

$$H_\gamma(x) := \begin{cases} 1-\left(1+\gamma\frac{x}{\sigma^*}\right)^{-1/\gamma} & \gamma \neq 0 \\ 1-\exp(-\frac{x}{\sigma^*}) & \gamma = 0 \end{cases}, \quad (3)$$

where γ and σ^* are the shape and scale parameters, respectively, with $x > 0$ if $\gamma \geq 0$ and $\gamma < 0$ provided that $0 < x < -\sigma^*/\gamma$. This is the method which will be adopted throughout the paper. It is important to stress here that both GEV and GP distributions, although resulting from different approaches, share the same shape parameter γ and that the scale parameters are related by $\sigma^* = \sigma + \gamma(\mu - u)$. Traditionally, the threshold parameter is chosen before fitting. As expected, threshold choice hinges on balancing bias and variance. The threshold must be sufficiently high to ensure that the asymptotics underlying the GPD approximation stand true, thus reducing the bias. Nonetheless, the reduced sample size for high thresholds increases the variance of the parameter estimates.

2.2. Parameter Estimation

Several methods have been proposed in the literature for estimating the GPD parameters (see, e.g., [4, 5]). Roughly speaking, such methods can be grouped according to three broad categories: moments-based, regression-based and likelihood-based estimators. Briefly, the estimation procedure adopted in this work consists in the following two steps: firstly, the threshold parameter μ is chosen to be the lowest level where all the higher threshold in the sample mean excess function

$$e_n(\mu) = \frac{\sum_{i=1}^n (x_i - \mu)\, I(x_i > \mu)}{\sum_{i=1}^n I(x_i > \mu)}, \quad (4)$$

which represents the empirical counterpart of the mean excess function $e(\mu) = \mathrm{E}[X-\mu|X > \mu]$, are consistent with a straight line, once the sample uncertainty is accounted for. An upward trend in the line indicates a heavy-tailed distribution ([8]). Secondly, after fixing μ, the parameters σ^* and γ are estimated by maximum likelihood. Note that neither analytical estimates nor closed-form expressions for the expected Fisher information can be found and, thus, numerical procedures have to be employed. Statistical inference over the parameters was based on the asymptotic normality of maximum likelihood estimators. Standard error for the estimates are obtained by reproducing the same scheme to a set of bootstrap replicates, each obtained by resampling with replacement the original sample of exceedances. The maximum likelihood estimators are regular provided that $\gamma > -0.5$.

3. Experimental data and preprocessing

The new approach is illustrated with two 24-h RR recordings from PhysioBank [6]: a normal subject (nsr001) and a Congestive Heart Failure patient (chf201).

The normal subject was selected from the nsr2db database and corresponds to a 64 years old female. The CHF patient was selected from the chf2db database and is a 55 years old male with NYHA class III. In both nsr2db and chf2db databases, the original ECG recordings were digitized at 128 Hz, and the beat annotations were obtained by automated analysis with manual review and correction [6]. In this work, 24-h RR recordings were obtained from the temporal difference (in sec) between consecutive Normal beats (i.e. NN intervals) after resampled at 2 Hz.

The RR temporal variabilities traditionally related to the respiratory rhythm ($\sim$ 0.25 Hz) and to sympathetic baroreflex activity ($\sim$ 0.1 Hz) were obtained by multiresolution analysis through the computation of the maximal overlap discrete wavelet transformation (MODWT) [7]. Shortly, MODWT allows the decomposition of a given recording X_t into a sum of $J+1$ sub-series corresponding to each time-scale, that is

$$X_t = \sum_{j=1}^{J} D_{t,j} + S_{t,J}, \quad (5)$$

where D_j represents the time series with the wavelet detail $j = 1, 2, \ldots, J$ and S_J is the time series with the wavelet smooth i.e., the remaining parcel of the decomposition.

In this work, MODWT was implemented with Mallat's dyadic pyramid algorithm and Daubechies 4 mother wavelet [7]. By considering 2 Hz sampling resolution, the dyadic decomposition provided wavelet details D_j with frequency content around $1/2^{j-1}$ Hz. Thus, third and fourth MODWT scales (0.25 and 0.125 Hz), were considered as those representative of the power classically related to respiratory rhythm (~ 0.25 Hz) and to sympathetic baroreflex activity (~ 0.1 Hz).

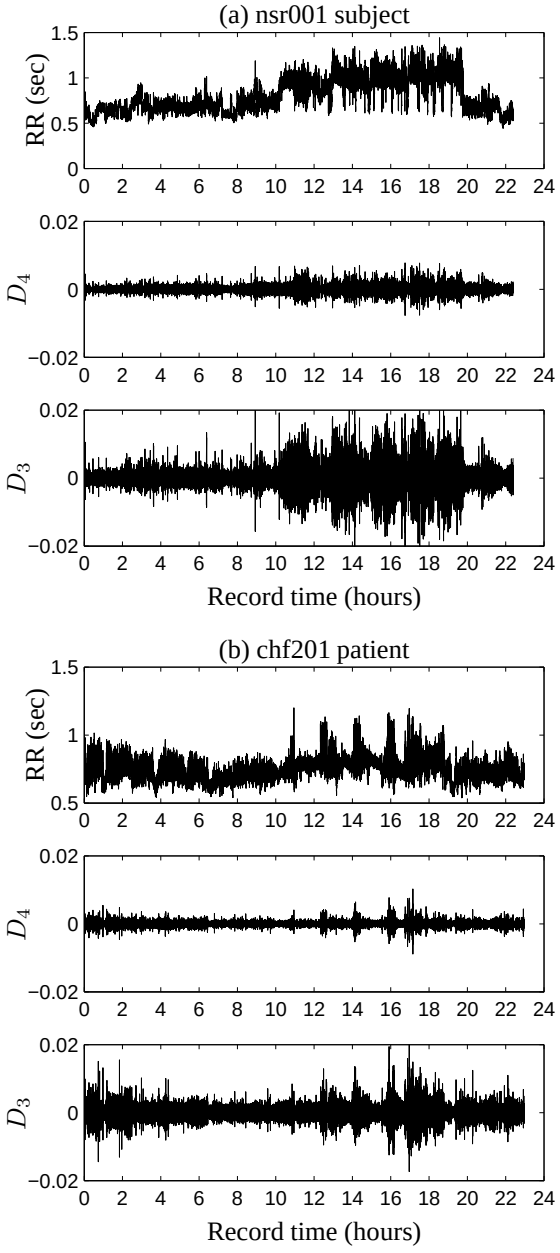

Figure 1. 24-hours RR recording and MODWT details classically related to sympathetic baroreflex activity (~ 0.1 Hz, D_4) and to respiratory rhythm (~ 0.25 Hz, D_3). Data from (a) nsr001 subject and (b) chf201 patient.

4. Results

Original 24-hour RR recordings for the normal subject and CHF patient are shown in Fig. 1. The MODWT details classically related to sympathetic baroreflex activity (D_4) and to the respiratory rhythm (D_3) are also presented. In concordance with previous studies, the RR variability in the CHF patient is typically lower than that observed for the normal subject [3]. Moreover, the variability of the D_3 detail is lower in the CHF patient than that evaluated in the normal subject, which has been pointed out as an evidence for decreasing parasympathetic activity during Congestive Heart Failure condition [3].

Figure 2 displays the sample mean excess function $e_n(\mu)$ against a set of threshold values μ. A closer look to Figure 2 reveals that for the four cases threshold parameters falling within the interval $I = [0.002, 0.01]$ are reasonable choices for μ. The procedure adopted to select a suitable point estimate for μ and also for the remaining GPd parameters comprises the following steps: (a) select a value for μ in I; (b) estimate the shape and scale param-

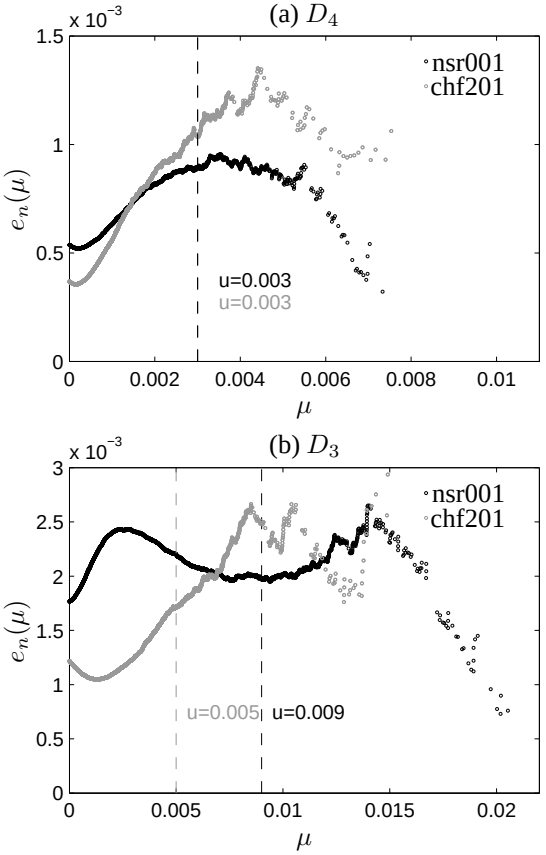

Figure 2. Mean excess plots obtained for the normal subject (nsr001) and the CHF patient (chf201), grouped by MODWT details D_4 (~ 0.1 Hz) and D_3 (~ 0.25 Hz).

eters by maximum likelihood; (c) perform a goodness-of-test fit for assessing the accuracy of the GPd model; (d) select another value for μ in I and repeat steps (b)-(c). Finally, select the value of μ which provides the better fit to the GPd model.

The results of the procedure described above are summarized in Table 1. It can be observed that the estimates of the tail index are larger for the CHF patient both in LF and HF. The fitted distributions for the normal subject exhibit $\gamma = 0$ for both LF and HF (-0.01 ± 0.03) and $0.04 \pm 0.02)$ while $\gamma > 0$ for the CHF patient (0.12 ± 0.06) and (0.15 ± 0.03), clearly supporting the idea that the CHF distributions are heavy-tailed.

Table 1. Estimate and standard error (ste) for the GPd parameters obtained for the normal subject (nsr001) and the CHF patient (chf201). Wavelet details D_4 and D_3 include respecively the LF and HF power of the HRV recording. The sample size n is also displayed.

filename	$\hat{\mu}$	$\hat{\gamma}$ (ste)	$\hat{\sigma^*}$ (ste)	n
nsr001 (D_4)	0.003	-0.002	0.0008	630
		(2.999e-02)	(1.996e-06)	
chf201 (D_4)	0.003	0.1152	0.0009	232
		(5.798e-02)	(1.999e-06)	
nsr001 (D_3)	0.009	0.040	0.0018	1066
		(2.327e-02)	(1.998e-06)	
chf201 (D_3)	0.005	0.152	0.0014	1091
		(2.752e-02)	(1.999e-06)	

5. Conclusion

In this work, the analysis of tail index in 24-h HRV recordings is introduced and illustrated in data from a normal subject and a CHF patient. The fitted distributions for the normal subject exhibit $\gamma = 0$ for both LF and HF $(-0.01 \pm 0.03$ and $0.04 \pm 0.02)$ while $\gamma > 0$ for the CHF patient $(0.12 \pm 0.06$ and $0.15 \pm 0.03)$. Thus, the CHF distributions are heavy-tailed, indicating a non-negligible probability that a very long RR interval occur. In a future study, the assessment of the impact of these preliminary findings in mortality prediction will be analyzed in detail.

Acknowledgements

This work was supported by the European Regional Development Fund (FEDER) through the COMPETE programme and by the Portuguese Government through the FCT, Fundação para a Ciência e a Tecnologia, in the scope of the projects UID/CEC/00127/2013 (IEETA, www.ieeta.pt) and project UID/MAT/04106/2013 (CIDMA, cidma.mat.ua.pt/). S. Gouveia acknowledges the postdoctoral grant by FCT (ref. SFRH/BPD/87037/2012).

References

[1] Huikuri HV, Makikallio T, Airaksinen KEJ, Mitrani R, Castellanos A, Myerburg RJ. Measurement of heart rate variability: a clinical tool or a research toy? J Am Coll Cardiol. 1999; 34:1878-1883.

[2] Hayano J, Kiyono K, Struzik ZR, Yamamoto Y, Watanabe E, Stein PK, Watkins LL, Blumenthal JA and Carney RM. Increased non-Gaussianity of heart rate variability predicts cardiac mortality after an acute myocardial infarction. Front. Physio. 2011; 65.

[3] Casolo G, Balli E, Taddei T, Amuhasi J, Gori C. Decreased spontaneous heart rate variability in congestive heart failure. Am J Cardiol 1989; 64:1162-1167.

[4] de Zea Bermudez P, Kotz S. Parameter estimation of the generalized Pareto distribution - PartI. J. Statist. Plann. Inference 2010; 140:1353-1373.

[5] Scarrott C, MacDonald A. A review of extreme value threshold estimation and uncertainty quantification. REVSTAT 2012; 10:33-60.

[6] Goldberger AL, Amaral LAN, Glass L, Hausdorff JM, Ivanov PCh, Mark RG, Mietus JE, Moody GB, Peng C-K, Stanley HE. PhysioBank, PhysioToolkit, and PhysioNet: components of a new research resource for complex physiologic signals. Circulation 2000; 101:e215-e220.

[7] Nowak JA, Ocon A, Taneja I, Medow MS, Stewart JM. Multiresolution wavelet analysis of time-dependent physiological responses in syncopal youths. Am J Physiol Heart Circ Physiol. 2009; 296:H171-9.

[8] Ghosh S, Resnick SI. When does the mean excess plot look linear? Stochastic Models 2011; 27:705-722.

Address for correspondence:

Sónia Gouveia
IEETA, Universidade de Aveiro
Campus Universitário de Santiago, 3810-193 Aveiro
E-mail address: sonia.gouveia@ua.pt