Robust estimation of the modified autoregressive index of film style

Nick Redfern

Abstract
The modified autoregressive index (mAR) describes the tendency of shots of similar length to cluster together in a motion picture but is not resistant to the influence of outliers if derived from the classical moment-based partial autocorrelation function. In this paper we calculate robust estimates of the modified autoregressive index based on outlier-resistant partial autocorrelation function based on the ranks of the shot length data and robust measure of scale. The classical, rank, and robust methods of determining mAR are compared for a sample of BBC news bulletins.

Keywords: autocorrelation, editing, film style, modified autoregressive index, robust estimation

1. Introduction

All statistical methods rely on sets of explicit and/or implicit assumptions that formalize what we know or believe about the process under analysis and represent mathematically convenient simplifications that are only approximately true. The body of knowledge relating to deviations from idealized assumptions in statistical methods is *robust statistics* (Hemel et al. 1986; Wilcox 2005; Maronna, Martin, & Yohai 2006), where *robustness* refers to 'the ability of a procedure or an estimator to produce results that are insensitive to departures from ideal assumptions' (Hella 2003: 17). This definition refers to the stability of statistical models and estimates when the true underlying distribution deviates slightly from the assumed model (distributional robustness) and to procedures or estimates that are not influenced by atypical observations (resistance to outliers). In this paper we are concerned with the effect of outliers on statistics of film style in the context of time series analysis and our focus will be on outlier-resistant methods.

Cutting, De Long, and Nothelfer (2010) proposed the modified autoregressive index (mAR) as a statistic of film style to describe the editing structure of motion pictures, calculating the value of the index by fitting a negative exponential function to the partial autocorrelation functions for the time series of film. Applying this method to a sample of 150 Hollywood films released between 1935 and 2005, they noted a tendency for films to increasingly cluster in packets of shots of similar length over time. However, the use of the mAR index to describe the editing style of a motion picture is questionable since the time series of shot lengths in a motion picture typically contains a number of extremely large observations. It is well known the classical autocovariance function and the autocorrelation and partial autocorrelation functions derived from it are not robust, and that just a single outlier can lead to significantly biased estimates of the autocorrelation and partial autocorrelation functions resulting in the incorrect specification of time series models both in terms of the type of model and the parameters of those models (Deutsch, Richards, & Swain 1990; Chan 1992, 1995). Consequently, the mAR index will not accurately describe the clustering of shots in a motion picture if it is based on such functions leading to flawed conclusions regarding the nature of film style.

This paper aims to derive robust estimates of the modified autoregressive index based on outlier-resistant estimates of the partial autocorrelation function. In section two we describe the time series functions compared in this study, and in section three we apply these methods to a sample of BBC news bulletins broadcast in April 2011.

2. Robust estimation of the autocorrelation and partial autocorrelation functions

The classical, moment-based autocovariance function for a weakly stationary time series $\mathbf{x} = (X_1, \dots, X_n)^{\mathrm{T}}$ is defined as

$$\gamma(h, \mathbf{x}) = \frac{1}{n}\sum_{i=1}^{n-h}(X_i - \bar{X})(X_{i+h} - \bar{X}) , \qquad (1)$$

where $\bar{X}$ is the mean and h is a lag operator specifying the distance between the observations X_i and X_{i+h}. The denominator in (1) is the total sample size (n), and so this function is biased and positive semidefinite. The autocovariance function when $h = 0$ is equal to the variance and standardising (1) by this value gives the autocorrelation function

$$\rho(h, \mathbf{x}) = \frac{\gamma(h, \mathbf{x})}{\gamma(0, \mathbf{x})}. \qquad (2)$$

The partial autocorrelation function ($\alpha(h, \mathbf{x})$) is the correlation between X_i and X_{i+h} with the linear dependence of the intervening lags removed, and can be calculated recursively using the Durbin-Levinson algorithm:

$$\alpha(h, \mathbf{x}) = \frac{\rho_h - \sum_{j=1}^{k-1}\rho_{h-j}\phi_{k-1,j}}{1 - \sum_{j=1}^{k-1}\rho_j\phi_{k-1,j}}, \qquad (3)$$

where $\phi_{a,b} = \phi_{a-1,b} - \alpha_a\phi_{a-1,a-b}$ and ρ_h is the autocorrelation function at lag h.

The presence of outliers in time series can seriously affect the structure of the autocorrelation function. The variance is greatly inflated in the presence of outliers since the square of their differences from the mean are very large, giving added weight to data points in the tails of the distribution at the expense of the mass of the data. Consequently, substituting the variance as the denominator in (2) depresses the autocorrelation function towards zero resulting in values of $\rho(h)$ that are too low and underestimate the strength of the correlation between X_i and X_{i+h} (Fajardo, Reisen, & Cribari 2009; Chatfield 2004: 27; Maronna, Martin, & Yohai 2006: 250-252). The finite sample breakdown point of both the mean and the variance is $1/N$%, and Ma and Genton (2000) showed the maximum temporal breakdown point of (1) for some lag h to be 0%, meaning that a just single outlier can destroy the information carried by this function. Since model identification for time series and parameter estimation depend heavily on the autocorrelation function, the presence of outliers may lead to the misidentification and misspecification of time series models.

As an alternative to the classical functions, a rank-based approach by Ahdesmäki et al. (2005; see Spangl 2008: 22) calculates the autocorrelation function of a time series as

$$\hat{\rho}_S(h, \mathbf{x}) = \frac{1}{n}\frac{12}{(n-h)^2 - 1}\sum_{i=1}^{n-h}\left(R_x(i) - \frac{n-h+1}{2}\right)\left(R'_x(i) - \frac{n-h+1}{2}\right), \qquad (4)$$

where $R_x(i)$ are the ranks of x_i in $S = \{x_t, t = 1, \dots n-h\}$ and $R'_x(i)$ are the ranks of x_{i+h} in $S' = \{x_{t+h}, t = 1, \dots n-h\}$. As an extension of Spearman's rank correlation statistic, $\hat{\rho}_S$ measures the monotonicity of the relationship between two observations and does not assume linearity. Because $\hat{\rho}_S$ is based on the ranks of the data it is resistant to the influence of outliers. This function

is biased but is directly comparable to the biased function based on (1), though it is not guaranteed to be positive semidefinite.

A robust method of estimating the autocorrelation function proposed by Ma and Genton (2000; see also Lévy-Leduc et al. 2010) based on a robust estimator of scale defines the autocovariance function of **x** as

$$\hat{\gamma}_Q(h, \mathbf{x}) = \frac{1}{4}\left[\left(Q_{n-h}^2(\mathbf{u}+\mathbf{v})\right) - \left(Q_{n-h}^2(\mathbf{u}-\mathbf{v})\right)\right], \tag{5}$$

where $\mathbf{u} = (X_1, \ldots, X_{n-h})^{\mathrm{T}}$ and $\mathbf{v} = (X_{1+h}, \ldots, X_n)^{\mathrm{T}}$. Q_n is the robust scale estimate proposed by Rousseeuw and Croux (1993):

$$Q_n = c \times \left\{\left|X_i - X_j\right| : i < j\right\}_k . \tag{6}$$

Q_n is the kth order statistic of the $\binom{n}{2}$ absolute pairwise differences, and which for large n equals 0.25. The value c is a consistency factor, which at the Gaussian distribution is 2.2191. The finite sample breakdown point of Q_n is 50%, and so $\hat{\gamma}_Q(h, \mathbf{x})$ has a maximum temporal breakdown point of 25% (Ma & Genton 2000: 673). This is the highest possible breakdown point in the context of autocovariance. The robust autocorrelation function is

$$\hat{\rho}_Q(h, \mathbf{x}) = \frac{Q_{n-h}^2(\mathbf{u}+\mathbf{v}) - Q_{n-h}^2(\mathbf{u}-\mathbf{v})}{Q_{n-h}^2(\mathbf{u}+\mathbf{v}) + Q_{n-h}^2(\mathbf{u}-\mathbf{v})}, \tag{7}$$

with $\left|\hat{\rho}_Q\right| \leq 1$. The value of $\hat{\rho}_Q$ is independent of the choice of c. To make (7) comparable to the biased classical estimator it is necessary to modify $\hat{\rho}_Q$ by $\frac{n-h}{n}$, although this function is not positive semidefinite.

3. The modified autoregressive index of BBC news bulletins

To determine the impact of outliers in shot length data on the mAR index and their subsequent influence on our interpretations of film style we calculate the indices based on the classical, rank, and robust partial autocorrelation functions for the 15 main BBC news bulletins broadcast at 1300, 1800, and 2200 from 11 April to 15 April 2011. We collected shot length data for these programmes by loading them into a non-linear editing programme and recording the duration of each shot in seconds. All broadcasts were recorded at 50 Hz, and shot length data is analysed at 25 frames-per-second. The 1300 and 1800 bulletins include the weather forecast, whereas 2200 does not; and to make the comparison between different bulletins direct these shots were discarded (including shots in which the newsreader announced the weather and interacted with the weather presenter before and after the forecast). The recap of a bulletin's main news story and the signoff by the newsreader that occur after the weather forecast in the 1300 and 1800 bulletins were retained as part of the data sets as these have corresponding shots in the bulletin at 2200. Each of the bulletins includes the headlines for the regional news programme to follow the nation broadcast and as a common feature of all the bulletins this data was also retained.

Table 1 presents the descriptive statistics for each bulletin, and shows that distribution of shot lengths in all cases are positively skewed with the median shot length substantially below the mean. To identify outliers in the time series of each bulletin we examined the box plot of the

residuals resulting from fitting a loess regression to each time series and this information is presented in Table 1.[1] Using this method, we identified between five and ten percent of the shots in each bulletin as outliers with a mean of 7.93% (SD = 1.65%).

Table 1 Descriptive statistics of BBC news bulletins broadcast 11 April 2011 to 15 April 2011, with the percentage of outliers identified in time series

	Length (s)	Shots	Mean (s)	Variance	Median (s)	Q_n (s)	Skewness	Outliers (%)
11/04/11: 1300	1714.6	216	7.9	116.8	4.6	3.1	5.2	9.7
11/04/11: 1800	1558.3	182	8.6	67.8	6.3	3.9	2.7	7.7
11/04/11: 2200	1602.5	197	8.1	74.9	5.3	4.0	3.4	5.1
12/04/11: 1300	1732.8	233	7.4	85.4	4.4	2.9	4.0	9.9
12/04/11: 1800	1621.7	238	6.8	57.1	4.0	2.6	3.5	8.4
12/04/11: 2200	1577.2	195	8.1	81.7	4.8	3.5	3.7	6.7
13/04/11: 1300	1697.6	233	7.3	77.3	4.2	3.1	3.6	10.3
13/04/11: 1800	1600.7	220	7.3	52.2	4.9	3.3	3.1	9.5
13/04/11: 2200	1562.3	214	7.3	73.6	4.5	2.8	4.0	9.3
14/04/11: 1300	1722.8	212	8.1	105.3	4.7	3.5	4.1	7.1
14/04/11: 1800	1557.7	203	7.7	50.1	5.0	3.5	2.6	4.9
14/04/11: 2200	1560.3	231	6.8	48.6	4.2	2.7	3.5	7.4
15/04/11: 1300	1709.2	190	9.0	163.6	5.6	3.7	4.9	7.9
15/04/11: 1800	1589.2	237	6.7	50.5	4.5	3.1	3.8	6.8
15/04/11: 2200	1560.9	233	6.7	63.3	4.6	3.3	5.4	8.2

Figure 1 presents three examples of time series from the sample with outliers identified. Shots identified as outliers in these bulletins are associated with specific aspects of the broadcasts' discourse structure and are typically static takes of people talking on screen including the kernel that introduces the main points of a news item, reporters' pieces-to-camera, studio discussions, and live two-way interviews between the presenter and a reporter on location (Montgomery 2007). Graphics with off-screen commentary may also be of very long duration though this is rarer. There is a strong tendency for outliers to cluster together due to their function in the discourse structure of a bulletin: in-studio discussions typically comprise a series of alternating shorter takes of a presenter asking questions (10-20s) and longer takes (>30s) of the interviewee's responses, while two-way interviews normally come at the close of a news item and are shortly followed by the kernel of the next item. For example, the cluster of outliers between shots 144 to 172 for the 1300 bulletin of 14 April in Figure 1 combines all these elements including a reporter's piece-to-camera, a kernel introducing a news item and the subsequent in-studio discussion, and the kernels for three of the next four items. There is no evidence that any outlier has an 'innovation' impact affecting the duration of subsequent shots. Because these types of shots are 'true outliers' that contain

[1] This approach is based on **R** code by Rob Hyndman available at http://stats.stackexchange.com/questions/1142/simple-algorithm-for-online-outlier-detection-of-a-generic-time-series, accessed 1 September 2012.

information about the decisions made by producers about the presentation of news items to the viewer it is desirable to retain this information as part of the data set without allowing their influence to distort our interpretation of the editing structure of the bulletins and this motivates the use of outlier-resistant methods in estimating the mAR index.

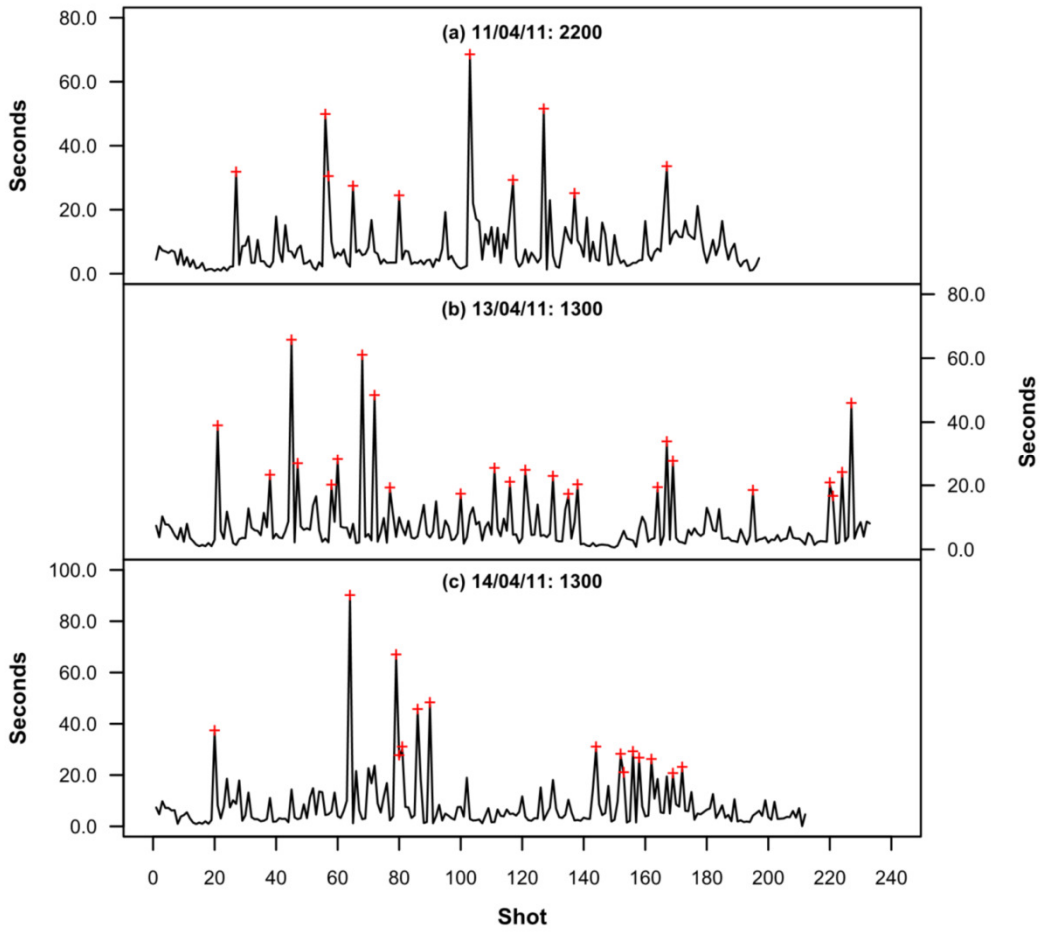

Figure 1 Time series with identified outliers for three BBC news bulletins

We calculated the classical, rank, and robust autocorrelation and partial autocorrelation functions for each time series having first linearly detrended the data. The rank and robust autocorrelation functions in (4) and (7) were calculated using unpackaged **R** functions by Spangl (2009). To ensure positive definiteness we arranged the robust autocorrelation functions as a toeplitz matrix and then found the nearest positive definite correlation matrix using the `nearPD` function in the **R** Matrix package (version 1.0-6), which implements Higham's (2002) algorithm. The partial autocorrelation functions were then calculated using the recursion in (3). To determine the autoregressive order of the local relations between shot lengths we follow Cutting, De Long, and Nothelfer (2010) in considering incremental positive partial autocorrelations as long as previous values remained positive and greater than the positive bound $2/\sqrt{N}$, where N is the sample size, thereby assuming a null hypothesis of white noise.

Table 2 presents the orders of autocorrelation and partial autocorrelation for each bulletin. The classical functions point to a white noise model with no evidence of autocorrelation among shot lengths for all but a few bulletins, and would therefore lead us to conclude the editing in BBC news

bulletins is, generally, a random process. This is contradicted by both the rank and robust functions, which indicate that editing patterns in these bulletins are not random with shot lengths correlated over short lags and that an AR(1) model is a more realistic model than white noise.

To take a specific example, Figure 2 shows the three autocorrelation functions for the 1800 bulletin broadcast on 11 April and it is immediately clear from there is a large difference between the time series structure identified by the classical autocorrelation function in the top panel and the two outlier-resistant functions. The lag-1 autocorrelations in particular for the three functions are very different: $\rho(1) = 0.1190$, $\hat{\rho}_S(1) = 0.4288$, and $\hat{\rho}_Q(1) = 0.5153$. The classical autocorrelation function in Figure 2 is not significantly different from zero for any lags (with the exception of lag 34) suggesting a white noise process with no linear relationship between the duration of shots. The rank and robust autocorrelation functions are significant up to lags 5 and 6, respectively, with the functions decaying slowly over a number of lags and we see from Table 2 the rank and robust partial autocorrelation functions indicate that a first order autoregressive model describes the structure of this bulletin rather than white noise.

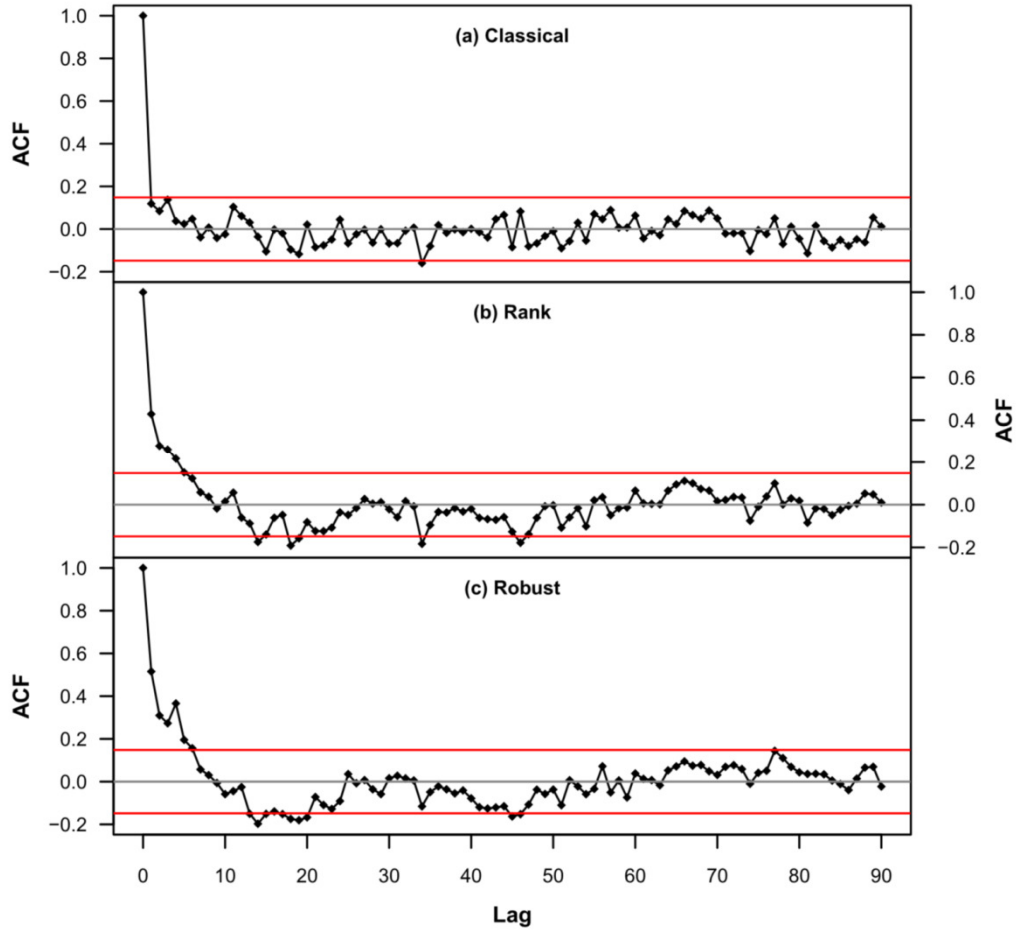

Figure 2 Three autocorrelation functions for lags 0 to 90 of the 1800 BBC News bulletin of 11 April 2011. The critical values are at $\pm 2/\sqrt{182} = \pm 0.1482$.

The next step is to calculate the modified autoregressive index for each bulletin. Following Cutting, De Long, and Nothelfer (2010), we fit the negative exponential function $1/[1 + h]^{\beta}$ to the classical and robust partial autocorrelation functions for lags 0 to 20 using nonlinear least squares

($df = 20$) with the value of mAR determined to be the intercept between the fitted function and a critical value of 0.1361 based on the median number of shots in a bulletin (N = 216).

Table 2 presents the modified autoregressive indexes based on each of the partial autocorrelation coefficients. The results clearly indicate a discrepancy between the values of the mAR based on the classical function and those based on the outlier-resistant functions. In the presence of outliers the classical mAR index is too low and does not accurately reflect the clustering of shots in these bulletins. The rank and robust mAR indices are broadly consistent with one another. We were unable to determine the value of mAR based on the classical function for the 1300 bulletin on 13 April as the series did not converge due to a singular gradient.

Table 2 Orders of autocorrelation (ACF), partial autocorrelation (PACF), and modified autoregressive index (mAR) based on classical and robust functions for BBC news bulletins

	Classical (ρ)			**Rank ($\widehat{\rho}_S$)**			**Robust ($\widehat{\rho}_Q$)**		
	ACF	**PACF**	**mAR**	**ACF**	**PACF**	**mAR**	**ACF**	**PACF**	**mAR**
11/04/11: 1300	0	0	0.45	1	1	1.09	1	1	1.32
11/04/11: 1800	0	0	1.09	5	1	2.21	6	1	2.33
11/04/11: 2200	1	1	1.04	4	1	1.95	4	3	2.33
12/04/11: 1300	0	0	0.57	2	1	1.76	3	1	1.89
12/04/11: 1800	1	1	0.99	1	1	1.67	6	1	2.29
12/04/11: 2200	0	0	0.73	1	1	1.42	2	1	1.89
13/04/11: 1300	0	0	***	1	1	1.67	4	1	2.15
13/04/11: 1800	1	1	1.07	1	1	1.48	1	1	1.29
13/04/11: 2200	0	0	0.79	1	1	1.33	1	1	1.28
14/04/11: 1300	0	0	0.93	1	1	1.79	4	1	2.03
14/04/11: 1800	1	1	0.98	1	1	1.35	2	1	1.51
14/04/11: 2200	0	0	0.53	1	1	1.39	3	1	2.05
15/04/11: 1300	0	0	0.36	1	1	1.61	1	1	1.84
15/04/11: 1800	0	0	0.70	1	1	1.28	1	1	1.52
15/04/11: 2200	0	0	1.01	2	1	2.29	2	1	2.40

Figure 3 shows the three partial autocorrelation functions and fitted negative exponential functions for the 1300 bulletin broadcast on 15 April 2011, and illustrates clearly how the effect of outliers on the classical partial autocorrelation function impacts the modified autoregressive index by depressing the function towards zero and leads us to underestimate the extent to which shots of similar length cluster together in this bulletin.

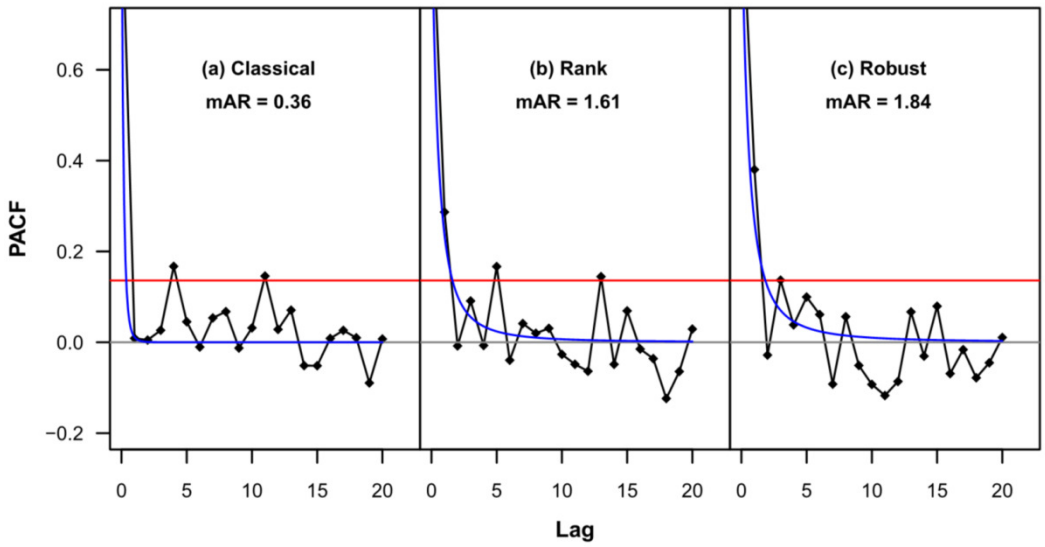

Figure 3 Partial autocorrelation functions (black) and fitted negative exponential functions (blue) for the 1300 BBC news bulletin of 15 April 2011. The critical value is at $2/\sqrt{216}$ = 0.1361. The ordinate has been truncated to 0.6.

4. Conclusion

The modified autoregressive index describes the tendency of shots of similar length in a motion picture to cluster together, but is not resistant to the influence of outliers if derived from the classical moment-based partial autocorrelation function. In this paper we estimated the mAR index using outlier-resistant partial autocorrelation functions based on the ranks of the data and on a robust measure of scale and showed that shot lengths in BBC news bulletins are correlated over short lags and that an AR(1) model is more appropriate than the random process identified by the classical-based mAR index.

In light of these results it is likely that the values of mAR given by Cutting, De Long, and Nothelfer (2010) underestimate the true clustering of shot lengths in Hollywood films.

Future development of these methods in the analysis of film style extends to robust spectral analysis. The power spectrum of a signal is the Fourier transform of its autocorrelation function, and if the structure of that function is severely biased by the presence of outliers those effects will be transferred to the power spectrum leading to the misidentification of the structure of the data. Applying a Fourier transform to the rank and robust autocorrelations functions used here we obtain robust estimates of the power spectral density (Spangl 2008: 19-37) and will allow us to test the further claim of Cutting, De Long, and Nothelfer that shot lengths in a motion picture follow a $1/f$ noise pattern.

References

Ahdesmäki M, Lähdesmäki H, Pearson R, Huttunen H, and Yli-Harja O 2005 Robust detection of periodic time series measured from biological systems, *BMC Bioinformatics* 6: 117.

Broersen PMT 2006 *Automatic Autocorrelation and Spectral Analysis*. London: Springer.

Chan WS 1992 A note on time series model specification in the presence of outliers, *Journal of Applied Statistics* 19: 117-124.

Chan WS 1995 Understanding the effects of time series outliers on sample autocorrelation, *Test* 4 (1): 179-186.

Chatfield C 2004 *The Analysis of Time Series: An Introduction*. Boca Raton, FL: Chapman & Hall/CRC.

Cutting JE, De Long JE, and Nothelfer CE 2010 Attention and the evolution of Hollywood film, *Psychological Science* 21 (3): 432-439.

Deutsch SJ, Richards JE, and Swain JJ 1990 Effects of a single outlier on ARMA identification, *Communications in Statistics: Theory and Methods* 19 (6): 2207-2227.

Fajardo FA, Reisen VA, and Cribari F 2009 Robust estimation in long-memory processes under additive outliers, *Journal of Statistical Planning and Inference* 139 (8): 2511-2525.

Hampel FR, Ronchetti EM, Rousseeuw PJ, and Stahel WA 1986 *Robust Statistics: The Approach Based on Influence Functions*. New York: John Wiley & Sons.

Higham NJ 2002 Computing the nearest correlation matrix – a problem from finance, *IMA Journal of Numerical Analysis* 22 (3): 329-343.

Lévy-Leduc C, Boistard H, Moulines E, Taqqu MS, and Reisen VA 2011 Robust estimation of the scale and of the autocovariance function of Gaussian short- and long-range dependent processes, *Journal of Time Series Analysis* 32 (2): 135-156.

Ma Y and Genton MG (2000) Highly robust estimation of the autocovariance function, *Journal of Time Series Analysis* 21 (7): 663-684.

Maronna R, Martin D, and Yohai V 2006 *Robust Statistics: Theory and Method*. Chichester: John Wiley & Sons.

Montgomery M 2007 *The Discourse of Broadcast News: A Linguistic Approach*. London: Routledge.

Rousseeuw PJ and Croux C (1993) Alternatives to the median absolute deviation, *Journal of the American Statistical Association* 88: 1273-1283.

Spangl B 2008 *On Robust Spectral Density Estimation*. Ph.D. Thesis, Vienna Technical University.

Spangl B 2009 [Robust-ts-commits] r9 - pkg/R (04 Mar 2009), http://lists.r-forge.r-project.org/pipermail/robust-ts-commits/2009-March/000000.html, accessed 23 March 2012.

Wilcox RR 2005 *Introduction to Robust Estimation and Hypothesis Testing*, Burlington, MA: Elsevier.