Segmentation of Time Series with Heteroskedastic Components

Derquenne, Christian
Electricité de France, R&D
1, avenue du Général de Gaulle
92141 Clamart Cedex, France
E-mail: christian.derquenne@edf.fr

1 Context and issues

The time series are decomposed into several types of changes: trend, seasonality, volatility and noise. They may be more or less regular as the application domain. Behavioral changes that characterize these series are mainly of several types: peak (price of energy in tense situation, but on a very short period), jumps in level or trend (or separation of gathering data stream), jumps variability (yield of the FTSE 100). Modeling of these series is very delicate and requires a lot of experience in the application domain. It may be interesting to detecting changes in behavior for many applications in the pre-treatment or not: construction of sub-models in each segment, stationnarized series using segmentation, building of symbolic curves to achieve a clustering of curves, modeling of multivariate time series, etc. Many segmentation methods [3,4,5,6,7] have been and are developed to address various problems in economics, finance, human sequencing, meteorology, energy management, etc. Most of these methods rely on the use of dynamic programming to reduce drastically the number of segmentations possible because it would obviously be totally illusory to calculate them all. Indeed, the number of segmentations for a series of length T and S fixed number of segments is $\binom{T-1}{S-1}$ whereas the set of all segments $S = 1, T$, the total number of segments increases to 2^{T-1}. The complexity of these algorithms is in general $O(T^2)$. These methods of detecting break points are designed to solve three problems [5]: (i) detecting a change in the mean, with a constant variance, (ii) detecting the change in variance with a constant mean (iii) detecting changes in the overall distribution of the phenomenon, without distinguishing changes in level, variability and distribution errors.

We introduced a method [2], which not only reduces the complexity compared to other methods, but mainly to propose solutions segmentation of the series containing segments increasing, decreasing, constant and different dispersions. Our method is original in its approach as it moved, in stages, a decision support for data segmentation. It contains two main phases: data preparation with a first data segmentation and modeling of segments using a Gaussian heteroskedastic linear model by successive adaptations. Each of the two phases is repeated a few times depending on the degree of smoothing applied to the data. The degree of smoothing can vary from 1 to T theory. The empirical complexity is $O(T\sqrt{T})$ and the theoretical complexity is $O(T^2)$. This method has been tested on many series and has provided encouraging results on both simulated data to assess the quality of reconstruction of the series: detection and modeling segments, but mainly on real data, especially in the area of price formation energy market.

But for all the segmentation methods that are based on a dynamic programming approach or an exploratory approach as ours, it appears that the segmentation quality may be lacking in the detection of contiguous segments when levels (or constant slopes) were statistically similar but have different variances. In this case only one segment will be detected, then there are two structurally. Therefore, we propose in this paper, a new method improves the previous one. This new approach has three phases. The first is to establish a proper transformation of data to obtain a new set characterizing the temporal evolution of the dispersion of observations, the second phase amounts to segmenting this new series with the same principle as the method [2] to obtain segments of dispersion, and the third phase reapplies [2] but taking into account the distribution of segments of dispersion, especially in the construction of the heteroscedastic linear model. To test our approach, we then conducted a comparative study with dynamic programming algorithms proposed in [6]. As we can see the quality of results was considerably improved. Finally, we propose to extend the comparisons with other methods, as well as future researches for the generalization of two theorems introduced in this article.

2 The proposed method

2.1 The model and its inference

Let's be a time series $(Y_t)_{t=1,T}$, we assume that it decomposes according to the Gaussian heteroskedastic linear model (or variance components) [8] as follows:

$$(1) \quad Y_t = \sum_{s=1}^{S} (\beta_0^{(s)} + \beta_1^{(s)} t + \sigma_s \epsilon_t) 1_{[t \in \tau_s]}$$

where $\beta_0^{(s)}$, $\beta_1^{(s)}$ and $\sigma_s > 0$ are respectively the parameters of level, slope and dispersion for the segment τ_s and ϵ_t follows a standard Normal. Finally, the number of observations per segment τ_s is denoted T_s with $\sum_{s=1}^{S} T_s = T$. Each segment τ_s contains the set of values: Y_t for $t = U_{s-1} + 1$ to U_s, where $U_s = U_{s-1} + T_s$, finally $U_S = T$. There are so $3S$ parameters to be estimated, knowing the number of segments S is unknown. To estimate the Gaussian heteroskedastic linear model, several estimators are available: ordinary least squares (OLS), maximum likelihood (ML) and restricted maximum likelihood or residual (REML).

2.2 The general process of segmentation

The approach introduced in [2] contains two phases: the first is the preparation of data, while the second is to establish successive models based on the model (1). Data preparation consists of three steps: smoothing, differentiation and counting. Smoothing aims to summarize the time series so as to keep only the strong trends in the series. For this, we chose to use the moving median as it is much more robust than the moving average. The degree of smoothing, denoted j $(j > 1)$ is the number of observations in the moving median $m_j(t)$ for $t = 1, T - j + 1$. Over j increases, unless the irregularity of the data is taken into account. The differentiation step can detect trends in the series on which the moving median was used. The differentiation must be high enough to appear differences in trend, but not too much not to miss. We chose to consider the property of the moving median with a difference at time t and time $k = j/2$ if j is even, and $k = (j+1)/2$ if j is odd. The counting step uses the results of stage of differentiation which has established a series of differences in positive, negative or zero, the number of values of the same sign is reasonably based on the degree of smoothing. Indeed, the lower it is, the more chances are that the size of the series of the same sign is small. Each serie will correspond to an initial segment. The first segment $\tau_{j,1}^{(0)}$ contains $T_{j,1}^{(0)}$ observations of the same sign, then the second segment $\tau_{j,2}^{(0)}$ include the $T_{j,2}^{(0)}$ observations of the same sign but different from that of $\tau_{j,1}^{(0)}$, etc. At the end of the process, we get a vector of segments $(\tau_{j,1}^{(0)}, ..., \tau_{j,s}^{(0)}, ..., \tau_{j,S}^{(0)})$, size $(T_{j,1}^{(0)}, ..., T_{j,s}^{(0)} ..., T_{j,S}^{(0)})$ and $\sum_{s=1}^{S} T_{j,s}^{(0)} = T$. However, this initial exploratory segmentation usually contains too many segments, especially the degree of smoothing is small. The objective is then to sum (simplify) the best (with the least possible loss of information) that prior segmentation. For this model (1) is proposed in the modeling phase. It takes place in several steps of simplification of the proposed model from it. Each step is divided itself into several stages based on the following inference:

(i) model estimated by REML, (ii) test of homoskedastity on overall series, (iii) test for equality of each slope coefficient associated in its segment, (iv) and for the constants, (v) test of homoskedasticity of each pair of successive segments, (vi) test for equality of the pairs of slope coefficients and (vii) test for equality for couples constants. Each stream contains not then a more simplified than the previous step in preserving the quality of the statistical model. The rule is based on minimum values from the REML estimator, the origin of the estimated final model, information criteria BIC, AIC and R^2 fit, and that the MAPE (Mean Absolute Percentage of Errors).

2.3 A new approach to pre-estimate of the dispersion

As indicated, the method [2] has provided encouraging results on different types of time series whether simulated or real, compared with methods based on dynamic programming. But for all the segmentation methods developed that are based on a dynamic programming approach or on an exploratory approach as ours, it appears that the segmentation quality may be lacking in the following scenario. Upon detection of contiguous segments: levels (constant or linear) are close but have statistically different variances, in this case only one segment will be detected instead of two. The new method proposed here has a particular aim to overcome this problem. This new approach has three main phases. The first is to establish a proper transformation of data to obtain a new set characterizing the temporal evolution of the dispersion of observations, the second phase amounts to segmenting this new series with the same principle as the method [2] to obtain segments of dispersion, and the third phase reapplies [2] but taking into account the distribution of segments of dispersion, especially in the construction of the Gaussian heteroskedastic linear model.

2.3.1 Steps 1 and 2: Transformation characterizing the volatility time series and the first segmentation

The goal is to build a new time series to exhibit volatility of time series observations that are assumed to be governed by the model (1). The transformation is the most natural differentiation of order 2, such that: $Z_t = (1-B)^2 Y_t$ where Y_t is the original time series. It is then possible to apply two operators on Z_t: $U_t =\mid Z_t \mid$ or $V_t = Z_t^2$. The two following theorems can obtain the variance σ^2 from these transformations.

Theorem 1: Let Y_t a Gaussian process i.i.d. indexed be in time with mean $\beta_0 + \beta_1 t$ and variance σ^2, as $Y_t = \beta_0 + \beta_1 t + \sigma\epsilon_t$, where ϵ_t is a standard normal, then $\sigma = \frac{\sqrt{\pi}}{2\sqrt{3}}\mathbb{E}(\mid Y_t - 2Y_{t-1} + Y_{t-2} \mid)$.

Demonstration: Let Y_t a Gaussian process i.i.d. be indexed in time with mean $\beta_0+\beta_1 t$ and variance σ^2, asking $Z_t = Y_t - 2Y_{t-1} + Y_{t-2}$ then $Z_t = \sigma(\epsilon_t - 2\epsilon_{t-1} + \epsilon_{t-2})$ is zero mean and variance $6\sigma^2$.

Calculate the distribution $U_t =\mid Z_t \mid=\mid Y_t - 2Y_{t-1} + Y_{t-2} \mid$. It raises the cumulative distribution function associated with U_t: F_{U_t}, such that $F_{U_t}(u_t) = \mathbb{P}[U_t < u_t] = \mathbb{P}[\mid Z_t \mid< u_t] = \mathbb{P}[-u_t < Z_t < u_t] = 2\mathbb{P}[Z_t < u_t] - 1 = 2F_{Z_t}(u_t) - 1$. The density function associated is: $f_{U_t} = 2f_{z_t}$.

Finally, calculate the expectation of U_t : $\mathbb{E}(U_t) = \frac{2}{\sqrt{6}\sigma\sqrt{2\pi}}\int_0^{+\infty} u_t e^{-\frac{u_t^2}{12\sigma^2}} du_t$. We put $s_t = \frac{u_t^2}{12\sigma^2}$ then $ds_t = \frac{u_t}{6\sigma^2}du_t$. Therefore, we get $\mathbb{E}(U_t) = \frac{2\sqrt{3}\sigma}{\sqrt{\pi}}\int_0^{+\infty} e^{-s_t} ds_t = \frac{2\sqrt{3}\sigma}{\sqrt{\pi}}[-e^{-s_t}]_0^{+\infty} = \frac{2\sqrt{3}\sigma}{\sqrt{\pi}}$, QED.

The first theorem provides an estimate of σ, as $\hat{\sigma}_U = \frac{\sqrt{\pi}}{2\sqrt{3}(T-2)}\sum_{t=3}^{T} \mid y_t - 2y_{t-1} + y_{t-2} \mid$.

Theorem 2: Let Y_t a Gaussian process i.i.d. indexed in time be with mean $\beta_0 + \beta_1 t$ and variance σ^2, as $Y_t = \beta_0 + \beta_1 t + \sigma\epsilon_t$, where ϵ_t is a standard normal, then $\sigma^2 = \frac{\mathbb{E}((Y_t - 2Y_{t-1} + Y_{t-2})^2)}{6}$.

Demonstration: Let Y_t a Gaussian process i.i.d. indexed in time be with mean $\beta_0+\beta_1 t$ and variance σ^2, asking $Z_t = Y_t - 2Y_{t-1} + Y_{t-2}$ then $Z_t = \sigma(\epsilon_t - 2\epsilon_{t-1} + \epsilon_{t-2})$ is zero mean and variance $6\sigma^2$.

Calculate the distribution $V_t = Z_t^2 = (Y_t - 2Y_{t-1} + Y_{t-2})^2$. It raises the cumulative distribution function associated with V_t: F_{V_t}, such that $F_{V_t}(v_t) = \mathbb{P}[V_t < v_t] = \mathbb{P}[Z_t^2 < v_t] = \mathbb{P}[-\sqrt{v_t} < Z_t < \sqrt{v_t}] = 2\mathbb{P}[Z_t < \sqrt{v_t}] - 1 = 2F_{Z_t}(\sqrt{v_t}) - 1$. The density function associated is: $f_{V_t} = \frac{1}{\sqrt{v_t}} f_{Z_t}$.

Finally, calculate the expectation of V_t: $\mathbb{E}(V_t) = \frac{1}{\sqrt{6}\sigma\sqrt{2\pi}} \int_0^{+\infty} v_t^{1/2} e^{-\frac{v_t}{12\sigma^2}} dv_t$. We put $w_t = \frac{v_t}{12\sigma^2}$ then $dw_t = \frac{dv_t}{12\sigma^2}$. Consequently, we have: $\mathbb{E}(V_t) = \frac{12\sigma^2}{\sqrt{\pi}} \int_0^{+\infty} w_t^{1/2} e^{-w_t} dw_t$. We use integration by parts, putting: $r_t = w_t^{1/2}$, so $dr_t = \frac{1}{2} w_t^{-1/2} dw_t$ and $ds_t = e^{-w_t} dw_t$ then $s_t = -e^{-w_t}$. The integral takes the following form: $\frac{12\sigma^2}{\sqrt{\pi}}([-w_t^{1/2} e^{-w_t}]_0^{+\infty} + \frac{1}{2}\int_0^{+\infty} w_t^{-1/2} e^{-w_t} dw_t) = \frac{6\sigma^2}{\Gamma(1/2)} \int_0^{+\infty} w_t^{-1/2} e^{-w_t} dw_t = 6\sigma^2$. Indeed, The integral divided by $\Gamma(1/2)$ is equal to one because its corresponds to cumulative distribution function on the entire domain of $\gamma(1/2)$ distribution, QED.

This second theorem gives an estimator of σ^2, as $\hat{\sigma}_V^2 = \frac{1}{6(T-2)} \sum_{t=3}^{T} (y_t - 2y_{t-1} + y_{t-2})^2$.

The results obtained using the previous two theorems are essential for the second phase because they allow to appear in each observed series u_t or v_t, levels of dispersion of segments of the candidate series. Indeed, the segmentation approach explained in paragraph 2.2. is then applied, either on the series u_t or on the series v_t. At the end of the process, the segmentation will provide a selected set of segments characterized by the model (1). The straight line segments obtained can be constant, increasing or decreasing, which provide additional information on the behavior of interesting data that may be heteroskedastic even on a segment.

Let now $\tau_1^\sigma, ..., \tau_{S_1}^\sigma$ be, S_1 the segments of dispersion obtained previously on the series u_t. Then the segment τ_s^σ provides an estimate of T_s values of u_t such that $\hat{u}_t = \hat{\alpha}_0 + \hat{\alpha}_1 t$ for $t \in \tau_s^\sigma$.

2.3.2 Step 3: Second segmentation taking into account the dispersion

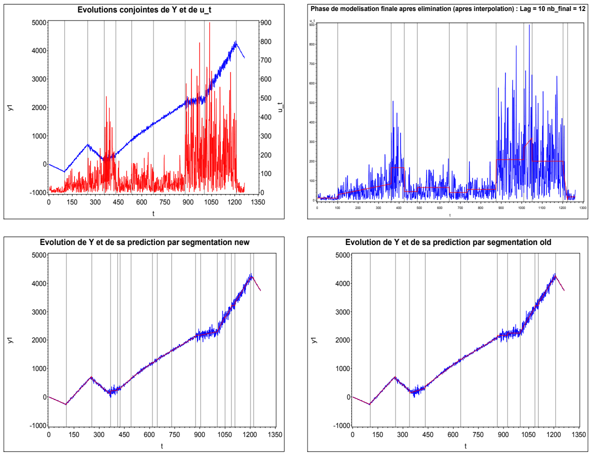
This third and last phase aims to provide a final segmentation of the original time series Y_t taking into account the dispersion of data which is estimated using phase 2. For this, two solutions are possible. Each value Y_t is standardized by $\hat{u}_t$ in order to eliminate the scattering effect from the beginning of segmentation, that is to say, the smoothing step in the phase data preparation. Either $\hat{u}_t$ are only integrated in the modeling phase to structure the dispersion matrix of Gaussian heteroskedastic linear model (1). This means using a diagonal matrix of weights when estimating model parameters using the REML estimator. $\hat{v}_t$ is used in the same way that $\hat{u}_t$.

3 Application

We applied the new method on the same set of simulated data we used in [2] to compare the results are output. We chose 10 segments, depending on model (1). For each of 10 segments, the number of observations, the values of the coefficients β_0 and β_1 and standard-deviation σ are generated randomly. Furthermore, we compared our results with those obtained using dynamic programming algorithms developed in [6]. They can detect multiple breakpoints in a time series. These are estimated by minimizing a penalized contrast function. The types of changes detected are: mean with constant variance, variance with constant mean, mean and variance (named DCPC3 in Figure 2a), distribution (DCPC4, 2b) and in the spectrum. In addition, a version was developed as Bayesian approach on the types of change (BDCPC3 and BDCPC4, the figures 2c and 2d). In this case, the moments of failure are exhibited by minimization of a posterior distribution. The mode of this posterior distribution is the minimum estimate of the penalized contrast.

We use the quantities u_t in phases 1 and 2, and the estimated values of these weights as the Gaussian heteroskedastic linear model in Phase 3. Figure 1a shows the simulated series (blue), the segments generated and the associated temporal evolution of u_t (red) end of phase 1. It may be noted that u_t are not constant on average, meaning that significant segments of different level of dispersion exist. Figure 1b provides the results of the first segment on the u_t (phase 2) on which 12 segments of dispersion were detected. We can see that all segments are not constant, because the second and ninth segments are growing. Figure 1c shows the final segmentation, which consists of 18 segments in which all the breaks seem to be detected. It was the same for the method [2] segmentation in which 12 segments had been identified (Figure 1d). Visually, the respective qualities of the two segmentations are comparable. However, although this is not very visible, additional segments of the new approach can best cut variation. For example, the fourth segment of the method [2] which is divided into two sub-segments with the new approach, a variation was actually appear heterogeneous series (more variability in the first sub-segment in the second). By cons, figures 2a-2d, which provide the results of the leak detection by dynamic programming are much less satisfactory. Indeed, in Figure 2a (DCPC3), the first three and last breaks are not detected by the algorithm, even if there are only seven segments. This problem extends to DCPC4 and BDCPC4 that should be more precise, since this type of detection freer, more that 16 segments. Finally BDCPC3, which gives a prohibitive number of segments (28 segments), also fails to properly identify the first three breakpoints and the last, like the three other methods of dynamic programming.

The quality of the raw signal recovery and adequacy of segmentations estimated by six methods to the segmentation generated is provided in Table 1. For this, we find that the values of MAPE of our old and new methods are very similar to that of the segmentation generated, whereas this is not the case for three methods of dynamic programming in four. It is the same for the median (MED) distributions of percentage absolute errors. The next column provides the percentage of errors less than 10%, on which appears the most efficient BDCPC4 (89.79%), although the MAPE is very high (59.85%). Finally, the last two columns provide the number of segments of each method identified the same place as those of the segmentation generated. Again, our two approaches are most effective on this dataset than the other four, with 6 segments found, against 0, 2 or 3 for methods of dynamic programming. These results are complemented by the error of the estimated distance of the segments to the segments generated. The method proposed in this paper obtains a relatively low percentage error (16.26%) compared to the former (24.34%) and the four other approaches, the best is that BDCPC3 still 35.28% error.



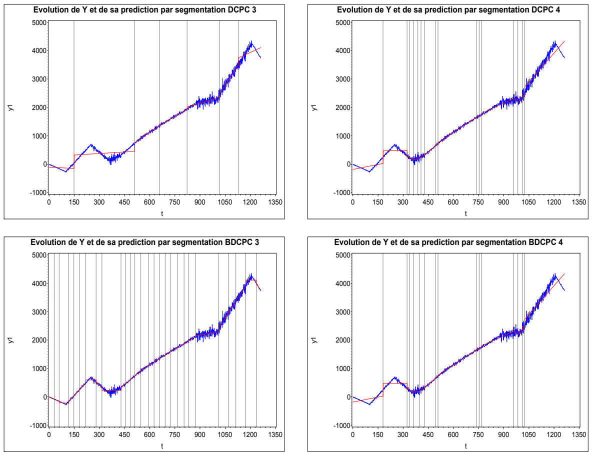


Figure 1: a,b,c,d : New and old approaches ———- Figure 2: a,b,c,d : Dynamic programming

Method	**Nb. seg.**	**MAPE (%)**	**MED (%)**	**<10%**	**B. seg.**	**Err. seg. (%)**
Generated data	10	9.90	2.32	87.32	n.a.	n.a
Method [2]	12	11.06	2.37	87.00	6	24.34
New method	18	9.89	2.30	86.29	6	16.26
DCPC3	7	75.80	4.48	67.31	0	36.77
DCPC4	17	59.72	4.00	70.29	3	52.00
BDCPC3	28	12.77	2.55	83.68	2	35.28
BDCPC4	16	59.85	4.03	89.79	3	48.94

Table 1: Results of the six methods

4 Contributions, reviews, applications and future directions

The method proposed here, which is used to segment a time series, aims to improve a process introduced in [2]. It had achieved encouraging results on simulated data and real data. It competes strongly approaches based on dynamic programming. The proposed improvement is to produce a signal from the raw data representing their dispersion through two theorems, and then perform an initial segmentation [2] of it to take segments of dispersion, and finally to include these as weights in a second segmentation [2] during the phase of successive models. On the simulated example, this new approach allows both to improve the old method, but also shows that it is more efficient than dynamic programming approaches [6]. This method is particularly interesting for applications in which signals change process. For the future directions, we will compare our method to that developed in [1] which uses a cross-validation. Finally, the two theorems introduced, in this paper, to a Gaussian signal will be generalized to other distribution of probability in our future research.

Bibliography

[1] Arlot, S. & Celisse, A. (2010): Segmentation of the mean of heteroscedastic data via cross-validation, *Statistics and Computing*, pp. 1-20.

[2] Derquenne, C. (2011): An Explanatory Segmentation Method for Time Series, *in Proceedings of Compstat 2010*, Y. Lechevallier & G. Saporta (eds.), 1^{st} Edition, pp. 935-942.

[3] Guédon, Y. (2008): Exploring the segmentation space for the assessment of multiple change-point models. Institut National de Recherche en Informatique et en Automatique, *Cahier de recherche 6619.*

[4] Hébrail G., Hugueney B., Lechevallier Y., Rossi F. (2010): Exploratory analysis of functional data via clustering and optimal segmentation. *Neurocomputing 73(7-9)*: pp. 1125-1141.

[5] Lavielle, M. and Teyssière, G. (2006): Détection de ruptures multiples dans des séries temporelles multivariées. *Lietuvos Matematikos Rinikinys*, Vol **46**.

[6] Lavielle, M. (2009): Detection of Changes using a Penalized Contrast (the DCPC algorithm), *http:www.math.u-psud.fr∼lavielleprogrammes_lavielle.html.*

[7] Perron, P. and Kejriwal, M. (2006): Testing for Multiple Structural Changes in Cointegrated Regression Models. Boston University, *C22.*

[8] Rao, CR. and Kleffe, J. (1988): *Estimation of variance components and applications.* North Holland series in statistics and probability, Elsevier.