HAL
open science

Change-points detection for variance piecewise constant models

Giada Adelfio

To cite this version:

Giada Adelfio. Change-points detection for variance piecewise constant models. Communications in Statistics - Simulation and Computation, 2011, 41 (04), pp.437-448. 10.1080/03610918.2011.592248 . hal-00759673

HAL Id: hal-00759673

https://hal.science/hal-00759673v1

Submitted on 2 Dec 2012

HAL is a multi-disciplinary open access archive for the deposit and dissemination of scientific research documents, whether they are published or not. The documents may come from teaching and research institutions in France or abroad, or from public or private research centers.

L'archive ouverte pluridisciplinaire **HAL**, est destinée au dépôt et à la diffusion de documents scientifiques de niveau recherche, publiés ou non, émanant des établissements d'enseignement et de recherche français ou étrangers, des laboratoires publics ou privés.

Change-points detection for variance piecewise constant models

Journal:	*Communications in Statistics - Simulation and Computation*
Manuscript ID:	LSSP-2011-0068
Manuscript Type:	Original Paper
Date Submitted by the Author:	14-Feb-2011
Complete List of Authors:	adelfio, giada; University of Palermo, Dipartimento Scienze Statistiche e Matematiche
Keywords:	Change-points, changes in variation, cumulative segmentation
Abstract:	A new approach based on the fit of a generalized linear regression model is introduced for detecting change-points in the variance of heteroscedastic Gaussian variables, with piecewise constant variance function. This approach overcome some limitations of both exact and approximate well known methods that are based on successive application of search and tend to overestimate the real number of changes in the variance of the series. The proposed method just requires the computation of a gamma GLM with log-link, resulting in a very efficient algorithm even with large sample size and many change points to be estimated.
Note: The following files were submitted by the author for peer review, but cannot be converted to PDF. You must view these files (e.g. movies) online.	
picking.zip	

SCHOLARONE™
Manuscripts

For Peer Review Only

CHANGE-POINT DETECTION FOR VARIANCE PIECEWISE CONSTANT MODELS

GIADA ADELFIO

Abstract. A new approach based on the fit of a generalized linear regression model is introduced for detecting change-points in the variance of heteroscedastic Gaussian variables, with piecewise constant variance function. This approach overcome some limitations of both exact and approximate well known methods that are based on successive application of search and tend to overestimate the real number of changes in the variance of the series. The proposed method just requires the computation of a gamma GLM with log-link, resulting in a very efficient algorithm even with large sample size and many change points to be estimated.

For Peer Review Only

1. Introduction

In this paper we propose a breakpoint detection procedure for changes in variation assuming that the variance function can be described by a piecewise constant function with segments delimited by unknown change-points.

In the literature testing about changes in mean in a Gaussian model has been studied by many authors (Chernoff and Zacks, 1964; Gardner, 1969; Hawkins, 1992; Worsley, 1979). Muggeo (2003) defined a segmented procedure fitting piecewise terms in regression models where one or more change-points are true parameters of the model, introducing a simple linearization technique.

1991 *Mathematics Subject Classification.*

Key words and phrases. Change-points, changes in variation, cumulative segmentation.

The problem of variance change-point detection has been widely considered in the literature. Some papers focuss on autoregressive time-series models (Wichern *et al.*, 1976; Wang and Wang, 2006; Zhao *et al.*, 2010, e.g.), most neglecting the problem of multiple change-points in part because of the difficulty in handling computations.

A general test statistic for detecting change-points in multidimensional stochastic processes with unknown parameters is proposed by Gooijer (2006). Vostrikova (1981) proposed a Binary Segmentation procedure (Scott and Knott, 1974) to detect the number of change-points in a multidimensional random process, also proving its consistency, with the computational advantage of detecting both the number of change-points and their position simultaneously. Inclán and Tiao (1994) used the cumulative sum of squares method to tackle the multiple variance change-points issue, in a way similar to the Binary Segmentation approach. Chen and Gupta (1997) proved the potential of the Bayesian Information Criterion and Binary Segmentation method in terms of computational cost and in the selection framework, avoiding some of the limitations of Inclán and Tiao (1994) approach that may identify too many changes and require some controls of the researcher to avoid cycling indefinitely.

The problem of detecting multiple change-points in large datasets has been considered by Killick and Eckley (2011a). The authors introduced the exact PELT (Pruned Exact Linear Time) approach that minimizes a cost function over possible numbers and locations of change-points characterized by a linear computational cost.

A likelihood-ratio method to detect changes in gamma distribution parameter for general values of the scale parameter has been adapted in Jandhyala *et al.* (2002).

For Peer Review Only

While Hawkins (1977) derived the exact null distribution of the likelihood ratio statistic for the case of detecting a change in the mean of a sequence of normally distributed random variables, Worsley (1986) showed the application of the same procedure for the gamma case, focussing on the exponential distribution.

In this paper a BIC-type method is used to test for abrupt changes in variance according to a stepwise function in a sequence of heteroscedastic Gaussian family random variables, investigating whether the variance of the observations has changed at unknown time points by a simpe GLM formulation of the problem.

The rest of the paper is organized as follows. In section 2 a brief introduction to the estimation and selection model for general mean variation is provided, while an extension to the variance piecewise constant changing model is introduced in section 3. As an example of general interest a well known study of U.S. stock market price volatility during 1971-1974 is provided in section 4. Section 5 presents results from a simulation study to assess the performance of the proposed approach in detecting and selecting unknown change-points in the variance model. Finally, discussion and conclusion are reported in section 6.

2. A general model: from changes in mean to changes in variance

Let $\{(x_i, y_i)\}_i^n$ be the observed data, where y_i is the outcome and x_i represents the observed sample for $i = 1, 2, \ldots, n$ occasions. Let us assume that $y_i = \mu_i + \epsilon_i$, with μ_i is for instance a sinusoidal function representing the observed signal and $\epsilon_i \overset{id}{\sim} \mathcal{N}(0, \sigma_i^2)$, with σ_i^2 a variance function approximated by a piecewise constant regression function with $K_0 + 1$ segments (see figure 1).

For Peer Review Only

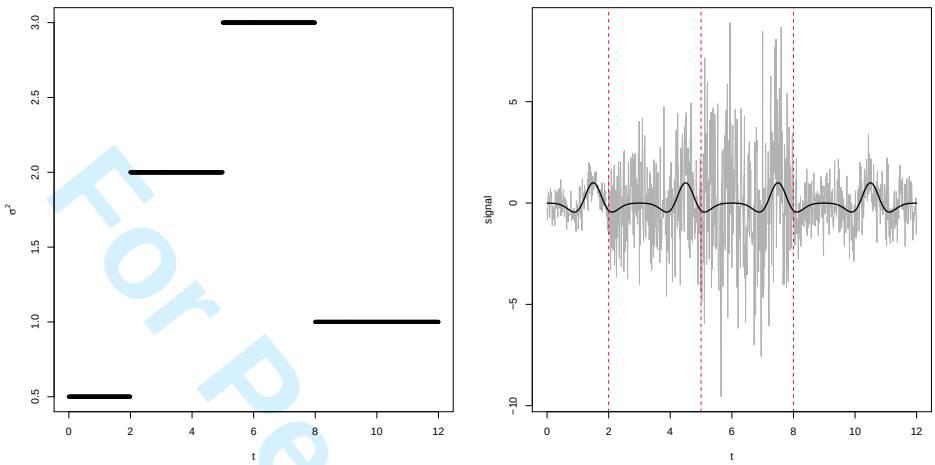

FIGURE 1. An example of variance with jump points and corresponding signal

For simplicity, the model for a change in variance after the k^*-th observation is:

$$y_i = \begin{cases} \mu_i + \lambda\epsilon_i & 1 \leq i \leq k^* \\ \mu_i + \tilde{\lambda}\epsilon_i & k^* < i \leq n \end{cases}$$

with $\lambda, \tilde{\lambda}$ and k^* unknown and $H_0 : \lambda = \tilde{\lambda}$ vs $H_A : \lambda \neq \tilde{\lambda}$.

Taking advantage of a generalized linear model formulation of the investigated problem, the test for stepwise changes in variance of a sequence of Gaussian random variables may be transformed equivalently to the case of testing for changes in mean on the squared residuals from an estimated linear model that accounts for the mean behavior of the observed signal. The estimation of the mean signal $\hat{\boldsymbol{\mu}}$ can be carried out by using a common smoothing procedure, e.g. fitting a cubic smoothing spline to the supplied data since this does not seem to influence results significantly.

This framework leads to model variance change-points on the basis of the estimation of a GLM where the response is the squared sum

For Peer Review Only

of Gaussian heteroscedastic variables. Since in this case we wish to model a strictly positive, continuous and typically skewed dispersion, the gamma is the obvious choice.

In particular, following a suggestion of Smyth *et al.* (2001) we fit a gamma GLM with a log-link function, with response given by the squared studentized residuals $s_i = (y_i - \hat{y}_i)^2/w_i$, with $\hat{\mathbf{y}} = \hat{\boldsymbol{\mu}}$ and weights $w_i = 1 - h_i$, with h_i the i-th diagonal element of the hat matrix H.

According to this approach, testing H_0 against H_A means that we are looking for a change in the mean of the residuals from a fitted linear model.

The proposed approach can be considered as a wider version of the *cumSeg* model proposed in Muggeo and Adelfio (2011) to detect multiple change-points in the mean of the gene expression levels in genomic sequences by a least squares approach. The authors assume that the datum $y_i, \forall i$ is defined as the sum of the signal μ_i and noise $\epsilon_i \overset{iid}{\sim} \mathcal{N}(0, \sigma^2)$ and that μ_i is approximated by a piecewise constant regression function with $K_0 + 1$ segments, that is:

$$y_i = \beta_1 + \delta_1 I(x_i > \psi_1) + \ldots + \delta_{K_0} I(x_i > \psi_{K_0}) + \epsilon_i \tag{2.1}$$

In (2.1), $I(\cdot)$ is the indicator function, such that $I(x) = 1$ if x is true, $\boldsymbol{\psi}$ represents the K_0 locations of the changes on the observed phenomenon, β_1 is the mean level for $x_i < \psi_1$ and $\boldsymbol{\delta}$ is the vector of the differences in the mean levels at the change-points.

The authors proceed taking the cumulative sums of equation (2.1) in order to get a convenient modelling expression that faces the discontinuities at the change-points ψ_k assuming a piecewise linear or segmented relationship.

For Peer Review Only

3. A gamma model for residuals

In this paper a variation of the *cumSeg* approach is presented to obtain estimates of the number and location of changes in variation when the variance function is described by a piecewise constant function with unknown jump points.

Let us assume that the sequence of squared residuals s_i $\forall i$ has a one-parameter exponential family distribution with density:

$$f(s_i) = \exp[\{s_i a(\mu_i) - b(\mu_i)\}/\phi_i + c(s_i, \phi_i)] \quad i = 1, \ldots, n \tag{3.1}$$

where a, b and c are known functions which specify the distribution and ϕ_i is a known dispersion parameter.

In equation (3.1) we assume the model with $a(\mu_i) = -\frac{1}{\mu_i}$ and $b(\mu_i) = \log(\mu_i)$, since s_i is a sequence of gamma variables with parameters $(\frac{1}{\phi_i}, \lambda_i)$ such that $\mu_i = \frac{\lambda_i}{\phi_i}$ and with dispersion parameter fixed to $\phi = 2$. In other words, we proceed assuming that testing for changes in the variance of a sequence of Gaussian variables is the same as testing for change-points in the rate parameter λ_i of gamma variables.

Given the link function $g(\cdot) = \log(\cdot)$, the model for changes in $g(\mu_i)$ can be formulated slightly modifying model (2.1) as follows:

$$g(\mu_i) = \beta_1 + \delta_1 I(x_i > \psi_1) + \ldots + \delta_{K_0} I(x_i > \psi_{K_0}). \tag{3.2}$$

The basic statistical problem with model (3.2) consists in the identification of the number of change-points K_0, using standard generalized linear models to fit the change point model, resulting in a very efficient algorithm even with large sample size and many change points to be estimated.

If the cumulative sum of s_i is still distributed as a variable with density belonging to the exponential family, expressed like (3.1), with

For Peer Review Only

$g(\theta_i) = g(E[z_i])$ where $z_i = \sum_j^i s_j$, taking the cumulative sums of the right side of equation (3.2) provides

$$g(\theta_i) = \beta_1 x_i + \delta_1 (x_i - \psi_1)_+ + \ldots + \delta_{K_0}(x_i - \psi_{K_0})_+ \tag{3.3}$$

where for each k and i, $x_i = i = \sum_j^i 1$, $\eta_i = \sum_j^i \epsilon_j$, and $(x_i - \psi_k)_+ = \sum_j^i I(x_j > \psi_k) = (x_i - \psi_k) I(x_i > \psi_k)$. Although model (3.3) has the same parameters of equation (3.2), equation (3.3) assumes a *piecewise linear* or *segmented* relationship differently from (3.2), namely a regression function continuous at the change-points ψ_k. This approach, then has the advantage of an efficient estimating approach via the algorithm discussed in Muggeo (2003, 2008) and fitting iteratively the generalized linear model:

$$g(\theta_i) = \beta_1 x_i + \sum_k \delta_k \widetilde{U}_{ik} + \sum_k \gamma_k \widetilde{V}_{ik}^-, \tag{3.4}$$

where $\widetilde{U}_{ik} = (x_i - \tilde{\psi}_k)_+$, $\widetilde{V}_{ik}^- = -I(x_i > \tilde{\psi}_k)$. The parameters β_1 and $\boldsymbol{\delta}$ are the same of equations (3.2) and (3.3), and the $\boldsymbol{\gamma}$ are 'working' coefficients useful for the estimation procedure (Muggeo, 2003).

At each step the working model (3.4) is fitted and new estimates of the change points are obtained via

$$\hat{\psi}_k = \tilde{\psi}_k + \hat{\gamma}_k / \hat{\delta}_k.$$

iterating the process up to convergence.

Then, the structure of the algorithm proceeds following the one proposed in Muggeo and Adelfio (2011) up to the convergence returning $K^*(< K)$ values and producing the fitted model

$$g(\hat{\theta}_i^*) = \hat{\beta}_1 + \hat{\delta}_1 V_{i1} + \ldots + \hat{\delta}_{K^*} V_{iK^*}, \tag{3.5}$$

where $V_{ik} = I(x_i > \hat{\psi}_k)$ for $k = 1, 2, \ldots, K^*$.

For Peer Review Only

The only difference now is that in the present approach the squared residuals are modelled as the response of a gamma GLM with logarithmic link function, still observing that if the number of change-points K_0 is known, the procedure converges in a few of iterations returning 'reasonably often' exactly K_0 estimated change-points.

Otherwise this makes some difference in selecting the number of significant change-points, that still reduces to selecting the significant variables among $V_1, \ldots, V_{K^*}$, with K^* the estimated number of change-points from model 3.4.

We solve this variable selection problem by using the *lars* algorithm by Efron *et al.* (2004). To make it useful in our procedure, assuming gamma distributed variables, we provide as the response variable of the *lars* algorithm its logarithmic normalizing transform, to still take advantage of the efficient cost of a single least squares computation that returns the solutions for the entire path. It could be useful to notice that the *lars* algorithm is just used as a computational efficient selection procedure, and once the entire path is provided, a gamma GLM is again estimated, from the only-intercept model to the full model when every variable V_k is included, to get the right likelihood values that need to compute any goodness of fit criterion penalized for the model dimension.

Thus the fitted optimal model with $\hat{K} \leq K^*$ change-points is selected by the generalized Bayesian Information Criterion (BIC) defined by:

$$BIC_{C_n} = -2 \log L + edf \log(n) C_n \tag{3.6}$$

where L is the likelihood function, *edf* is the actual model dimension quantified by the number of estimated parameters (including the intercept, the $\boldsymbol{\delta}$ and $\boldsymbol{\psi}$ vectors) and C_n is a known constant. Wang *et al.*

For Peer Review Only

(2009) discuss the use of $C_n > 1$ when the number of parameters is not fixed but it diverges as $n \to \infty$. Muggeo and Adelfio (2011) defined a generalized BIC based on Gaussian distributed iid errors, with $C_n = \log\log n$ that appears the most suitable value. In the present case it seems reasonable assuming a value $C_n = 2\log\log n$, from the BIC expression obtained for not Gaussian variables.

Main steps of the algorithm:

Let $y_i = \mu_i + \epsilon_i$ with ϵ_i normal distributed heteroscedastic zero mean errors; looking for changes in variance of $\mathbf{y}$, the proposed algorithm can be summarized as follows:

(1) Provide an estimate of $\hat{\mathbf{y}} = \hat{\mu}$ to account for the mean behavior of the observed signal (a simple average of data or a cubic smoothing spline can be considered).

(2) Compute $s_i = (y_i - \hat{y}_i)^2/(1 - h_i)$ for each i, such that $s_i \sim Gamma(\frac{1}{\phi_i}, \lambda_i)$ and look for change-points in the rate parameter λ_i of these gamma variables as in eq. (3.2).

(3) Compute the cumulative sum of $s_i,\ i = 1, \ldots, n$ such that, assuming that the squared residuals are modelled as the response of a gamma GLM with log-link, a regression function continuous at the change-points can be considered, as in eq. (3.3).

(4) Iteratively fit the generalized linear model in eq. (3.4).

(5) Select the number of significant change-points K^* by using the *lars* algorithm, applying a logarithmic normalizing transform to the response variable.

(6) Once the entire path is provided, estimate again a gamma GLM, in order to obtain the likelihood values necessary to compute the goodness of fit criterion penalized for the model dimension.

For Peer Review Only

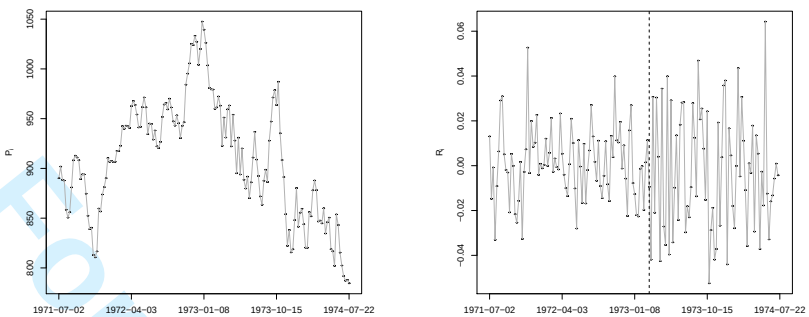

FIGURE 2. Weekly closing values P_i of the Dow-Jones Industrial Average from July 1, 1971 to August 2, 1974 and rates of return R_i. The vertical line in the plot on the right corresponds to the highly significant change identified by Hsu (1979) in late March 1973.

(7) Select the optimal model with $\hat{K} \leq K^*$ change-points by the generalized Bayesian Information Criterion (BIC) defined in (3.6).

The algorithm has been completely implemented by software R (R Development Core Team, 2006).

4. Application

As an example of application we consider 162 weekly closing values P_t of the Dow-Jones Industrial Average from July 1, 1971 to August 2, 1974 (Hsu, 1979). The rates of return are $R_i = (P_{i+1} - P_i)/P_i, (i = 1, ..., 161)$. It is assumed that $R_1, ..., R_{161}$ are independent normal random variables with constant mean and a variance which may change after an unknown time (see plots in fig. 2).

The proposed method well identifies as a point of change the 89-th observation. For comparison we also consider two different methods for multiple change-points using exact or approximate approaches,

For Peer Review Only

PELT (Killick and Eckley, 2011a) and Binary Segmentation respectively (Scott and Knott, 1974) implemented in *changepoint* Package (Killick and Eckley, 2011b) of R (R Development Core Team, 2006).

Both PELT and Binary Segmentation proceed iteratively looking for segments of the original sequence of random variables into multiple subsequences, detecting changes and continuing the process until no more changes are found in any of the subsequences. While the latter proceeds in an approximate way, PELT is an exact approach that attains linear computational cost and leads to a more accurate segmentation of data than Binary Segmentation (Killick and Eckley, 2011a).

Using BIC as the penalty selection criterion (Chen and Gupta, 1997), both methods can still identify the 89-th observation as a point of change in variance, but they seem to overestimate the real number of change-points, since they find also other points almost at the end of the serie as possible changes (155-th and 160-th observations for Binary Segmentation and 155-th and 159-th observations for PELT), reflecting one limitation of segmentation-based approaches.

5. Simulation

A simulation study is carried out to assess the performance of the proposed approach in identifying the correct number of change-points, that in our simulations is fixed to $K_0 = (1, 2, 3)$. In particular executing 1000 runs for each scenario, we first compare the selection procedure of the proposed approach based on the criterion in equation (3.6) and the *cumSeg* one defined in Muggeo and Adelfio (2011), where a generalized BIC criterion is based on Gaussian distribution assumption of the random signal, and evaluate the sensitivity of results. We consider

For Peer Review Only

also the Binary Segmentation and PELT methods introduced in section 4 for multiple changes in variance. Also the case when $K_0 = 0$ is investigated, that represents a crucial situation to compare different approaches.

Performance results assessed in terms of empirical mean and Mean Squared Error (MSE) are reported in tables 1 and 2.

For performance comparison between *cumSeg* and other well known procedures such as CBS (circular binary segmentation, (Olshen *et al.*, 2004; Venkatraman and Olshen, 2007)), CGHseg (Picard *et al.*, 2005) and LB (the lasso-based discussed by Huang *et al.* (2005)) see Muggeo and Adelfio (2011).

The true variance used for simulation is $\sigma_i^2 = 0.5 + 7.5I(i > 0.2) + 2.5I(i > 0.6) + I(i > 0.8)$ when $K_0 = 3$, $\sigma_i^2 = 0.5 + 7.5I(i > 0.2) + 2.5I(i > 0.6)$ when $K_0 = 2$, $\sigma_i^2 = 0.5 + 2.5I(i > 0.3)$ when $K_0 = 1$, and $\sigma_i^2 = 0.5i$ when $K_0 = 0$ (see for instance figures 3 and 4), while the true mean is a sinusoidal signal with expression $\mu_i = \sin(3\pi i)$, assuming that $x_i = i$ for the sake of simplicity.

For the simulated scenarios the gamma distribution assumption for squared residuals to test changes in variance of heteroscedastic Gaussian signal seems reasonable, both in terms of the mean number of selected change-points and their mean squared errors. Indeed, although the alternative approaches for multiple changes in variance can detect in mean the right number of change-points even for small sample sizes, their performance can not be considered satisfying, since they tend to overestimate the number of real changes with an increasing MSE for increasing sample sizes.

For Peer Review Only

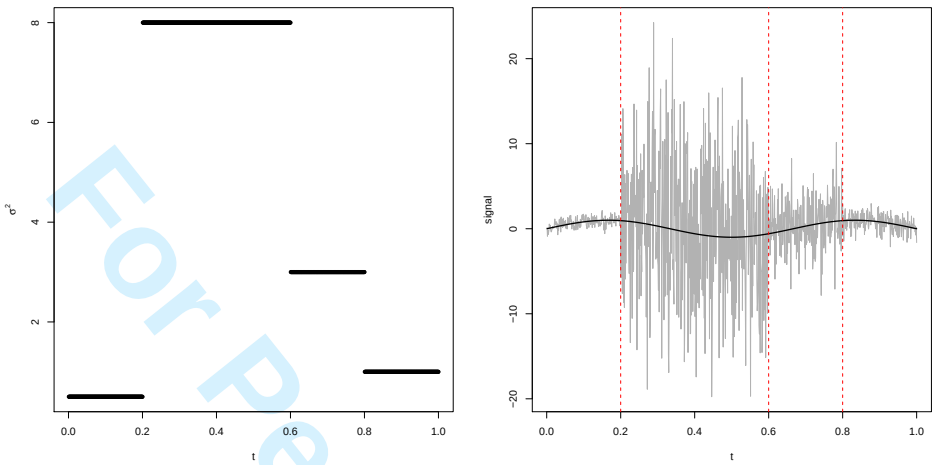

FIGURE 3. An example of variance with jump points and corresponding signal used in simulations for $K_0 = 3$

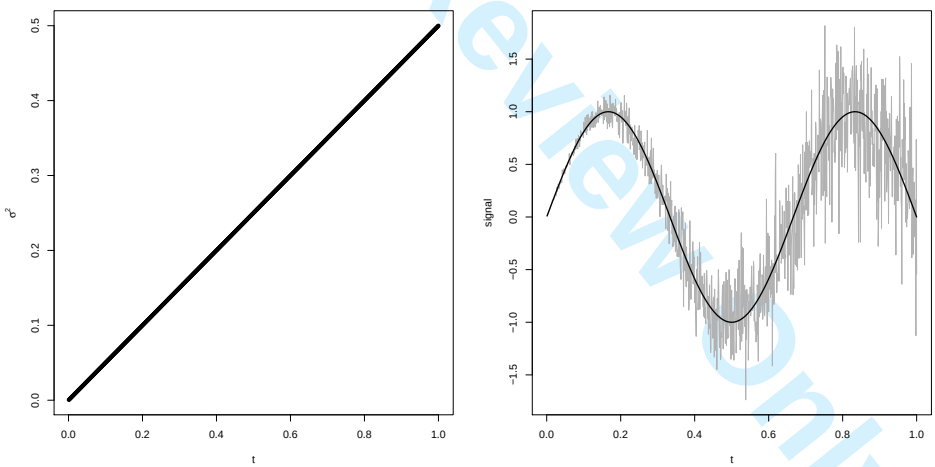

FIGURE 4. An example of variance linearly increasing and corresponding signal used in simulations for the null case

For the null case the proposed method outperforms the other approaches, always identifying no change-points with zero standard errors, although *cumSeg* for Gaussian heteroscedastic distribution outperforms its competitors (Muggeo and Adelfio, 2011). This confirms

For Peer Review Only

the importance in defining this procedure that is necessary to extend the change-points selection model based on cumulative segmented estimation approach to generalized linear regression models.

In analogous way as in Huang *et al.* (2005) we assess empirically the performance of the proposed algorithm with respect to the estimated locations of the $\hat{K}$ change-points $\hat{\psi}_k$. We select $K_1 = \min\{\hat{K}, K_0\}$ pairs $(\hat{\psi}_1, \psi_1^0), \ldots, (\hat{\psi}_{K_1}, \psi_{K_1}^0)$, that is we find for each estimated change-point $\hat{\psi}_k$ the closest ψ, and we call this ψ_k^0. Then, we just assign a score to each estimate $\hat{\psi}_k$ according to a simple rule:

$$\omega_k = \begin{cases} 1 & \text{if } \hat{\psi}_k \in [\psi_k^0 - 2, \psi_k^0 + 2] \text{ and } K_0 = \hat{K} \\ 0 & \text{if } \hat{\psi}_k \notin [\psi_k^0 - 2, \psi_k^0 + 2] \text{ and } K_0 = \hat{K} \\ -0.5|K_0 - \hat{K}| & \text{if } K_0 \neq \hat{K} \end{cases} \tag{5.1}$$

Therefore the overall score for each simulation is simply the sum of the scores defined in (5.1), such that $\varpi = \sum_k \omega_k$. When all the true change points are correctly identified, the overall score is $\varpi = K_0$, otherwise when some ψ_k^0 is mis-located and/or $\hat{K} \neq K_0$, the overall score will be $\varpi < K_0$. For each simulated scenario, the performance index is the average over the 1000 runs of these overall scores. We consider the same previous simulation settings: $K_0 = (1, 2, 3)$ and three sample sizes $n = (100, 500, 1000)$. Table 3 reports the performance indices for our procedure in comparison to the others introduced in the paper; actually for these approaches the penalty used is the asymptotic one (see Killick and Eckley (2011b) for more details) since the BIC penalty provides very unperforming and incomparable results with respect the proposed gamma-BIC selection.

This specific test reflects the 'conservative' nature of the proposed method with respect the others, since it tends to do not over-estimate

For Peer Review Only

the number of real change-points, approaching K_0 as the sample size increases in each scenario. Otherwise while the Gaussian-BIC approach always provides lower values of the performance index ϖ, both the Binary Segmentation and PELT methods reflect a quite inefficient behavior finding too many spurious points as the sample size increases.

6. Conclusion

In this paper a simple and efficient method to detect and select unknown change-points in regression models with piecewise constant models has been developed.

The approach is a generalization of the *cumSeg* procedure proposed by Muggeo and Adelfio (2011) for any regression model with a linear predictor, since testing for changes in GLM gamma with $\phi = 2$ equals testing for changes in the variance of a sequence of Gaussian random variables.

The framework is based on quite efficient algorithm to estimate the change-points and a variation of lars procedure adapted to the GLM case, to discard the spurious ones on the basis of a generalized version of the BIC. The proposed approach just require the fitting of a generalized linear model to the change-point model, resulting in a very efficient algorithm even with n large and many change-points to be estimated.

Simulations have shown good performance of the proposed approach, such as a sample size increases MSE decreases.

There are many circumstances in which testing for change in variance is crucial, such as stock markets records, waveforms of earthquakes, etc. In this regard, we think that one of the main advantages of the proposed method is just its wide application fields, also related to its capability in detecting more of one single change-point in a variance function

For Peer Review Only

with jump-points via a very efficient computationally procedure. For the mean issue just few other approaches in literature do the same, e.g. Huang *et al.* (2005), Tibshirani and Wand (2008).

Although simulations suggest a precise choice for C_n, the point how to define a general C_n is still an interesting question open to further discussion.

Acknowledgements

I would like to thank Dr. Vito M. R. Muggeo for his fruitful and insightful comments.

References

Chen, J. and Gupta, A. K. (1997). Testing and locating variance changepoints with application to stock prices. *Journal of the American Statistical association*, **92**, 739–747.

Chernoff, H. and Zacks, S. (1964). Estimating the current mean of a normal distribution which is subjected to changes in time. *The Annals of Mathematical Statistics*, **35**, 999–1018.

Efron, B., Hastie, T., Johnstone, I., and Tibshirani, R. (2004). Least angle regression. *Annals of Statistics*, **32**, 407–489.

Gardner, L. A. (1969). On detecting changes in the mean of normal variates. *The Annals of Mathematical Statistics*, **40**(1), 116–126.

Gooijer, J. G. D. (2006). Detecting change-points in multidimensional stochastic processes. *Comput. Stat. Data Anal.*, **51**(3).

Hawkins, D. M. (1977). Testing a sequence of observations for a shift in location. *Journal of the American Statistical Association*, **72**, 180–186.

For Peer Review Only

Hawkins, D. M. (1992). Detecting shifts in functions of multivariate location and covariance parameters. *Journal of Statistical Planning and Inference*, **33**(2), 233 – 244.

Hsu, D. A. (1979). Detecting shifts of parameter in gamma sequences with applications to stock price and air traffic flow analysis. *Journal of the American Statistical Association*, **74**, 31–40.

Huang, T., Wu, B., Lizardi, P., and Zhao, H. (2005). Detection of DNA copy number alterations using penalized least squares regression. *Bioinformatics*, **21**, 3811–3817.

Inclán, C. and Tiao, G. C. (1994). Use of cumulative sums of squares for retrospective detection of changes of variance. *Journal of the American Statistical Association*, **89**, 913–923.

Jandhyala, V. K., Fotopoulos, S. B., and Hawkins, D. M. (2002). Detection and estimation of abrupt changes in the variability of a process. *Computational Statistics and data analysis*, **40**, 1–19.

Killick, R., F. P. and Eckley, I. (2011a). An exact linear time search algorithm for multiple changepoint detection. *Submitted*.

Killick, R. and Eckley, I. A. (2011b). *changepoint: Contains funcions that run various single and multiple changepoint methods*. R package version 0.3.

Muggeo, V. M. R. (2003). Estimating regression models with unknown break-points. *Statistics in Medicine*, **22**, 3055–3071.

Muggeo, V. M. R. (2008). Segmented: an R package to fit regression models with broken-line relationships. *R News*, **8**(1), 20–25.

Muggeo, V. M. R. and Adelfio, G. (2011). Efficient change point detection for genomic sequences of continuous measurements. *Bioinformatics*, **27**, 161–166.

For Peer Review Only

Olshen, A. B., Venkatraman, E. S., Lucito, R., and Wigler, M. (2004). Circular binary segmentation for the analysis of array-based dna copy number data. *Biostatistics*, **5**, 557–572.

Picard, F., Robin, S., Lavielle, M., Vaisse, C., and Daudin, J. (2005). A statistical approach for array CGH data analysis. *BMC Bioinformatics*, **6**(27).

R Development Core Team (2006). *R: A Language and Environment for Statistical Computing*. R Foundation for Statistical Computing, Vienna, Austria. ISBN 3-900051-07-0.

Scott, A. J. and Knott, M. (1974). A cluster analysis method for grouping means in the analysis of variance. *Biometrics*, **30**, 507–512.

Smyth, G. K., Huele, A. F., and Verbyla, A. P. (2001). Exact and approximate reml for heteroscedastic regression. *Statistical Modelling*, **1**(3), 161–175.

Tibshirani, R. and Wand, P. (2008). Spatial smoothing and hot spot detection for cgh data using the fused lasso. *Biostatistics*, **9**, 18–29.

Venkatraman, E. S. and Olshen, A. B. (2007). A faster circular binary segmentation algorithm for the analysis of array cgh data. *Bioinformatics*, **23**, 657–663.

Vostrikova, L. (1981). Detecting 'disorder' in multidimensional random processes. *Sov. Math. Dokladys*, **24**, 5559.

Wang, H., Li, B., and Leng, C. (2009). Shrinkage tuning parameter selection with a diverging number of parameters. *Journal of Royal Statistical Society B*, **71**, 671–683.

Wang, L. and Wang, J. (2006). Change-of-variance problem for linear processes with long memory. *Statistical Papers*, **47**, 279–298.

For Peer Review Only

Wichern, D. W., Miller, R. B., and Hsu, D.-A. (1976). Changes of variance in first-order autoregressive time series models-with an application. *Journal of the Royal Statistical Society. Series C (Applied Statistics)*, **25**(3), 248–256.

Worsley, K. J. (1979). On the likelihood ratio test for a shift in location of normal populations. *Journal of the American Statistical Association*, **74**(366), 365–367.

Worsley, K. J. (1986). Confidence regions and tests for a change-point in a sequence of exponential family random variables. *Biometrika*, **73**, 91–104.

Zhao, W., Tian, Z., and Xia, Z. (2010). Ratio test for variance change point in linear process with long memory. *Statistical Papers*, **51**(2), 397–407.

Dip. di Scienze Statistiche e Matematiche "S. Vianelli", University of Palermo
E-mail address: adelfio@unipa.it

For Peer Review Only

TABLE 1. Empirical Mean (m) and Mean Squared Error (mse) of the detected number of change-points over 1000 runs in three sample sizes and three variance structures ($K_0 = (1, 2, 3)$) according to four different approaches :(a) = Gaussian BIC; (b) = gamma BIC (proposed approach); (c) = Binary Segmentation; (d) =PELT.

		$K_0 = 3$			
n		(a)	(b)	(c)	(d)
100	m	0.493	1.903	2.435	2.488
	mse	7.824	1.343	0.585	0.540
500	m	2.504	2.544	3.489	3.439
	mse	0.805	0.462	0.603	0.505
1000	m	2.701	3.141	3.947	3.919
	mse	0.644	0.315	1.255	1.187
		$K_0 = 2$			
n		(a)	(b)	(c)	(d)
100	m	0.532	1.156	2.009	1.379
	mse	3.734	0.714	0.019	0.749
500	m	2.566	2.021	2.386	2.393
	mse	0.913	0.02	0.408	0.421
1000	m	2.565	2.002	2.812	2.850
	mse	0.815	0.002	0.848	0.954
		$K_0 = 1$			
n		(a)	(b)	(c)	(d)
100	m	0.286	0.601	1.018	1.025
	mse	0.991	0.399	0.018	0.029
500	m	1.249	1.008	1.327	1.379
	mse	0.411	0.008	0.391	0.507
1000	m	1.317	1.006	2.133	2.238
	mse	0.465	0.006	1.894	2.179

For Peer Review Only

TABLE 2. Empirical Mean (m) and Mean Squared Error (mse) of the detected number of change-points over 1000 runs in three sample sizes and for $K_0 = 0$ according to four different approaches:(a) = Gaussian BIC; (b) = gamma BIC (proposed approach); (c) = Binary Segmentation; (d) =PELT.

		$K_0 = 0$			
n		(a)	(b)	(c)	(d)
100	m	0.573	0	1.002	1.037
	mse	1.338	0	1.006	1.115
500	m	2.201	0	5.191	5.374
	mse	6.220	0	28.054	29.384
1000	m	3.497	0	8.389	8.319
	mse	14.142	0	71.468	69.946

For Peer Review Only

TABLE 3. Performance of the proposed procedure with respect to the estimated locations of the change points. The entries in the Table refer to the index for three sample sizes, $K_0 = (1, 2, 3)$ and four different approaches:(a) = Gaussian BIC; (b) = gamma BIC (proposed approach); (c) = Binary Segmentation; (d) =PELT.

	$K_0 = 3$			
n	(a)	(b)	(c)	(d)
100	0.294	1.500	2.132	2.220
500	1.992	2.668	2.756	2.781
1000	2.066	2.756	2.527	2.543
	$K_0 = 2$			
n	(a)	(b)	(c)	(d)
100	0.256	1.260	1.986	1.984
500	1.690	1.990	1.807	1.804
1000	1.716	2.000	1.594	1.579
	$K_0 = 1$			
n	(a)	(b)	(c)	(d)
100	0.170	0.601	0.991	0.987
500	0.782	0.996	0.837	0.810
1000	0.775	0.997	0.434	0.384

For Peer Review Only

Reply to Referee report 1

dear Referee,
Many Thanks for reviewing my paper. I have read with attention all your comments and suggestions and I have modified the paper accordingly. I really think the paper is much more clear and, in general, greatly improved.

Below there are our point-by-point responses to your comments.
Thank you very much for your work.

Major Comments

I really do not think that the current paper can be considered just an application of the cumSeg procedure proposed by Muggeo and Adelfio (2011), not just because it is already published, but mostly because it provides an extension of the method for changes in mean there presented for testing changes in variance when changes follow a stepwise function.

I really think that although the current work is greatly based on the previous one, it is an useful extension and generalization of the previous approach with the advantage of requiring just the fitting of a generalized linear model to the change-point model, resulting in a very efficient algorithm even with large sample size and many change-points to be estimated.

For Peer Review Only

Reply to Referee report 2

dear Referee,
Many Thanks for reviewing my paper. I have read with attention all your comments and suggestions and I have modified the paper accordingly. I really think the paper is much more clear and, in general, greatly improved.

Below there are our point-by-point responses to your comments.
Thank you very much for your work.

Major Comments

1. I have modified the abstract adding more details of the proposed approach. Also the introduction now provides more indications of different approaches that are in the literature, with the objective of showing throughout the paper the advantages of our approach with respect some of these.

   In particular I have focussed on the well known Binary Segmentation and PELT approaches, that are both used for detecting multiple changes in variance with good computational performance, providing some comparison both by an application to a well known case study and by simulations.

2. I have given some introduction about the cumSeg method (Muggeo and Adelfio, 2011) to make the reader understand the current work more easily. I really do not think that the current paper can be considered just an application of the cumSeg procedure proposed, because it provides an extension of the method for changes in mean there presented for testing changes in variance when changes follows a stepwise function.

   I really think that although the current work is greatly based on the previous one, it is an useful extension and generalization of the previous approach with the advantage of requiring just the fitting of a generalized linear model to the change-point model, resulting in a very efficient algorithm even with n large and many change-points to be estimated.

3. I have added a summary of the procedure in section 3 for a better comprehension of the proposed approach.

4. I have modified the simulation scenarios, taking into account more cases for the number of real change-points and more estimation approaches, in order to compare results.

   I have also assessed empirically the performance of the proposed algorithm with respect to the estimated locations of the change-points following a test procedure that focus not just on the ability of the method in estimating the right number of changes in variance but also on the precision in finding their real locations.

   I think that this paper can be considered as a first presentation of this new method that seems to be efficient and very simple in its application: this requires just the possibility of fitting an usual GLM and allows the detection of multiple changes in variance characterized by discontinuous steps, avoiding any sequential segmentation of data that need an excessive criticism of the researcher.

5. An application of the method is provided in section 4.

For Peer Review Only