Detection of trend changes in time series using Bayesian inference

N. Schütz and M. Holschneider
Focus Area for Dynamics of Complex Systems, Universität Potsdam, Karl-Liebknecht-Str. 24, D-14476 Potsdam, Germany
(Dated: Received: April 19, 2011/ Accepted: date)

Change points in time series are perceived as isolated singularities where two regular trends of a given signal do not match. The detection of such transitions is of fundamental interest for the understanding of the system's internal dynamics. In practice observational noise makes it difficult to detect such change points in time series. In this work we elaborate a Bayesian method to estimate the location of the singularities and to produce some confidence intervals. We validate the ability and sensitivity of our inference method by estimating change points of synthetic data sets. As an application we use our algorithm to analyze the annual flow volume of the Nile River at Aswan from 1871 to 1970, where we confirm a well-established significant transition point within the time series.

PACS numbers: 02.50.Tt, 02.50.Cw, 05.45.Tp, 92.70.Kb

I. INTRODUCTION

The estimation of change points challenges analysis methods and modeling concepts. Commonly change points are considered as isolated singularities in a regular background indicating the transition between two regimes governed by different internal dynamics. In time series analysts focus on change points in observed data to reveal dynamical properties of the system under study and to infer on possible correlations between subsystems. Detecting trend changes within various data sets is under intensive investigation in numerous research disciplines, such as palaeo-climatology [1, 2], ecology [3, 4], bioinformatics [5, 6] and economics [7, 8].
In general, the detection of transition points is adressed via **(i)** regression [9] or **(ii)** spectral analysis methods [10], **(iii)** Bayesian approaches [11, 12] or **(iv)** recurrence network techniques [13, 14].
In this work we formulate transition points not only in terms of the underlying regular dynamics, but also as a transition in the heteroscedastic noise level. We use Bayesian inference to produce estimates for all relevant parameters. Our signal model is described by a regular mean undergoing a sudden change and a heteroscedastic fluctuation which undergoes as well a sharp transition at the same time point. Thus, in its simplest form, the observed signal $\boldsymbol{y}$ has a linear trend undergoing a break point θ at a time point $t_i = \theta$. The posterior density $p(\theta|\boldsymbol{y})$ of the change point given the signal enables us to derive the point estimate $\hat{\theta}$ as the most likely break point and its confidence bounds. By applying a sliding window, we formally localize the posterior density and the modelling of the subsignals as a linear trend is valid in first order. Consequently we investigate time series globally and locally for a generalized break point in the signal's statisitical properties.
In comparison to established methods (e.g. **(ii)** multiscale spectral analysis [10]) our technique is not restricted on a uniform time grid (e.g. as required for filtering methods). The majority of existing methods require additional approaches to interpret the confidence of the outcome (e.g. **(i)** bootstrapping, **(ii)** test statistics, **(iv)** introducing measures). Whereas our technique provides the confidence intervals of the estimates as a byproduct in a natural way. This, for us is actually the most convincing argument to approach the detection task via Bayesian inference since besides the parameter estimation on its own, we obtain a degree of belief about our assumed model and about the uncertainties in the parameters [15–17]. Existing techniques addressing Bayesian inference **(iii)** approach on the one hand the plain localization task of the singularity by treating the remaining model's parameter as hidden [18, 19]. On the other hand hierarchical Bayesian models are used [11] mainly based on Monte-Carlo-expectation-maximization (MEMC) algorithms for the estimation process [6, 12].
In contrast, we intend to achieve an insight in the parameter structure of the time series. We intend to detect multiple change points without enlarging the model's dimensionality, since this increases considerably the computational time. By addressing the general framework of linear mixed models (LMM) [20] we are able to factorize the joint posterior density into a family of parametrized Gaussians. This mirrors the separation of the linear from the non-linear parts and it simplifies considerably the explicit computation of the marginal distributions. Our technique will be applied to a hydrological time series of the river Nile, which exhibits a well known change point.

II. DEFINITION OF THE MODEL

In our modeling approach we consider two aspects of change points in a time series. On the one hand, a change point is commonly associated with a sudden change of local trend in the data. This indicates a transition point between two regimes governed by two different internal dynamics. On the other hand we assume that the systematic evolution of the local variability of the data around its average value undergoes a sudden transition at the change point. As we will show, both aspects can be combined into a linear mixed model with hyperparameters.

Our formulation allows the separation of the Gaussian from the intrinsic non-linear parts of the estimation problem, which besides clarifying the structure of the model, speeds up computations considerably.

A. Formulation of the linear mixed model

The simplest type of signal undergoing a change point at time θ can be expressed as

$$y(t) = \beta_0 + \beta_1|\theta - t|_- + \beta_2|\theta - t|_+ + \xi(t)\,. \quad (1)$$

Here we use the elementary Hockey sticks of first order defined through

$$|\theta - t|_- = \left(\zeta_-^\theta\right) = \begin{cases} \theta - t & \text{if } t \le \theta \\ 0 & \text{else} \end{cases}, \quad (2)$$

and

$$|\theta - t|_+ = \left(\zeta_+^\theta\right) = \begin{cases} \theta - t & \text{if } t \ge \theta \\ 0 & \text{else} \end{cases}. \quad (3)$$

Natural data series can in general not be modeled by such a simple behavior as given by these functions. Therefore we add some random fluctuations ξ around the mean behavior. These random fluctuations can be due to measurement noise as well as to some intrinsic variability, which is not captured by the low dimensional mean dynamics on both sides of the change point θ. For this fluctuating part of the signal we suppose that its amplitude is essentially constant around the change point. The intrinsic variability however may, like the mean behavior of the system itself, undergo a sudden change in its evolution of amplitude. Hence we consider stochastic fluctuations $\xi(t)$ whose amplitudes undergo a transition themselves according to

$$\mathrm{STD}(\xi(t)) = \sigma(1 + s_1|t - \theta|_- + s_2|t - \theta|_+)\,. \quad (4)$$

The scale factor σ could be the level of the measurement noise or some background level of the intrinsic fluctuations, whereas the constants $s_{1,2}$ describe the systematic evolution of the models intrinsic variability prior and after the change point measured in units of σ. Although clearly the fluctuating part may contain coherent parts, we assume that throughout this work, that the fluctuations are Gaussian random variables, wich at different time points are uncorrelated

$$\mathbb{E}(\xi(t)\xi(t')) = 0\,, \quad t \neq t'\,. \quad (5)$$

This clearly is an approximation and its validity can be questioned in concrete applications. However this assumption allows us to implement highly efficient algorithms for the estimation of the involved parameters. From now on we will call this fluctuating part simply "noise". A realization of such a time series is presented in Fig.**1**. Given a data set of n time points t_i, $i = 1, \ldots, n$, the observation vector $\boldsymbol{y} = [s(t_i)]^t \in \mathbb{R}^n$ can be written as follows

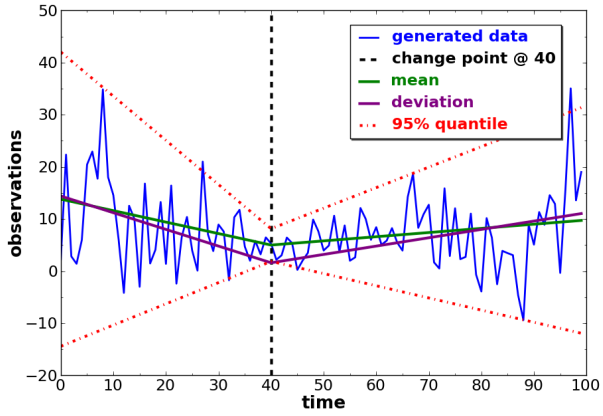

Figure 1: Realization of a synthetic time series of $n_{obs} = 100$ data points generated by Equ.(6) whereas the mean is parametrized by $F_\theta\boldsymbol{\beta} = 5{+}0.22{\cdot}\boldsymbol{\zeta}_-^\theta{+}0.08{\cdot}\boldsymbol{\zeta}_+^\theta$ and the deviation is modeled as $\sigma^2\Omega_{\theta,\boldsymbol{s}} = \left[1.6(1 + 0.2\cdot\boldsymbol{\zeta}_-^\theta + 0.1\cdot\boldsymbol{\zeta}_+^\theta)\right]^2$.

$$\boldsymbol{y} = F\boldsymbol{\beta} + \boldsymbol{\xi}\ . \quad (6)$$

Here the fixed effect vector $\boldsymbol{\beta} = (\beta_0, \beta_1, \beta_2)^T \in \mathbb{R}^3$ corresponds to the coefficients of the linear combination of the Hockey sticks modeling the mean behavior. The system matrix of the fixed effects, $F \in \mathbb{R}^{n\times 3}$, is then given by the sampling of the Hockey sticks $\boldsymbol{\zeta}_\pm^\theta$ defined in Equ.(2,3) at the observation points

$$F_\theta = \begin{pmatrix} 1 & \left(\zeta_-^\theta\right)_1 & \left(\zeta_+^\theta\right)_1 \\ \vdots & \vdots & \vdots \\ 1 & \left(\zeta_-^\theta\right)_n & \left(\zeta_+^\theta\right)_n \end{pmatrix}, \quad \left(\zeta_\pm^\theta\right)_i = \left(\zeta_-^\theta\right)(t_i)\,. \quad (7)$$

The noise $\boldsymbol{\xi} \in \mathbb{R}^n$ is a Gaussian random vector with zero mean and covariance matrix $\sigma^2\Omega \in \mathbb{R}^{n\times n}$,

$$\boldsymbol{\xi} \sim \mathcal{N}\left(0, \sigma^2\Omega\right). \quad (8)$$

The covariance itself is structured noise, which is parametrized by the two slope parameters $\boldsymbol{s} = (s_1, s_2)$ and the change point θ itself as

$$\left(\Omega_{\theta,s_1,s_2}\right)_{ij} = \left(\left[1 + s_1\left(\zeta_-^\theta\right)_j + s_2\left(\zeta_+^\theta\right)_j\right]^2\right)\cdot\delta_{ij}\,. \quad (9)$$

In conclusion, the probability density of the observations for fixed parameters (i.e. fixed effects, change point, slope parameters) can be written as

$$\boldsymbol{y} \sim \mathcal{N}\left(F\hat{\boldsymbol{\beta}}, \sigma^2\Omega\right). \quad (10)$$

The Likelihood function of the parameters given the data can then be written as

$$\mathcal{L}(\boldsymbol{\beta},\sigma,\boldsymbol{s},\theta|\boldsymbol{y}) = \frac{1}{(2\pi\sigma^2)^{\frac{n}{2}}\sqrt{|\Omega|}}\, e^{-\frac{1}{2\sigma^2}(\boldsymbol{y}-F\boldsymbol{\beta})^T\Omega^{-1}(\boldsymbol{y}-F\boldsymbol{\beta})}\,. \tag{11}$$

Note that the functional dependency of $\boldsymbol{\beta}$ is a Gaussian density. Clearly in the exponential $\boldsymbol{\beta}$ is of a quadratic form and since $\Xi = F^T\Omega^{-1}F$ is positive definite we may write

$$\frac{1}{(2\pi\sigma^2)^{\frac{n}{2}}\sqrt{|\Omega|}}\, e^{-\frac{\mathcal{R}^2}{2\sigma^2}}\, e^{-\frac{1}{2\sigma^2}(\boldsymbol{\beta}-\boldsymbol{\beta}^*)^T\Xi(\boldsymbol{\beta}-\boldsymbol{\beta}^*)} \tag{12}$$

where the mode of the Gaussian in $\boldsymbol{\beta}$ is the best linear unbiased predictor of the fixed effects (BLUP) [21]

$$\begin{aligned}\boldsymbol{\beta}^* &= \underset{\boldsymbol{\beta}\in\mathbb{R}^3}{\operatorname{argmin}}(\boldsymbol{y}-F\boldsymbol{\beta})^T\Omega^{-1}(\boldsymbol{y}-F\boldsymbol{\beta}) \\ &= (F^T\Omega^{-1}F)^{-1}F^T\Omega^{-1}\boldsymbol{y}\end{aligned} \tag{13}$$

and the residuum $\mathcal{R}$ measured in the Mahalanobis distance [22], induced by the covariance matrix Ω, is

$$\begin{aligned}\mathcal{R}^2 &= \min_{\boldsymbol{\beta}\in\mathbb{R}^3}(\boldsymbol{y}-F\boldsymbol{\beta})^T\Omega^{-1}(\boldsymbol{y}-F\boldsymbol{\beta}) \\ &= (\boldsymbol{y}-F\boldsymbol{\beta}^*)^T\Omega^{-1}(\boldsymbol{y}-F\boldsymbol{\beta}^*)\,.\end{aligned} \tag{14}$$

In addition, the profiled Likelihood function $\mathcal{L}(\boldsymbol{\beta}^*,\sigma,\boldsymbol{s},\theta|\boldsymbol{y})$ enables us to derive the profiled Likelihood estimator of the scale parameter σ

$$\hat{\sigma}^2 = \frac{\mathcal{R}^2}{n+1}\,, \tag{15}$$

which is auxiliary for the computation of the maximum of the Likelihood function.

B. Bayesian inversion

In the light of the Bayesian theorem, we can compute the posterior distribution $p(\boldsymbol{\beta},\sigma,\theta,\boldsymbol{s}|\boldsymbol{y})$ of the modeling parameters given the data $\boldsymbol{y}$ from the Likelihood function Eq. (11) by specifying the prior distribution of the parameters $p(\boldsymbol{\beta},\sigma,\theta,\boldsymbol{s})$, which encodes our belief about the parameters prior to any observation. Since we assume a priori no correlations between the parameters, the joint prior distribution can be factorized into the independent parts

$$p(\boldsymbol{\beta},\sigma,\theta,\boldsymbol{s}) = p(\theta)\cdot p(\boldsymbol{s})\cdot p(\sigma)\cdot p(\boldsymbol{\beta})\,. \tag{16}$$

In general, we do not have any a priori knowledge about these hyperparameters and thus we shall use flat and uninformative priors [23, 24]

$$p(\theta)\sim 1\,,\;\; p(\boldsymbol{s})\sim 1\,,\;\; p(\boldsymbol{\beta})\sim 1\,, \tag{17}$$

For the scale parameter σ we assume a Jeffrey's prior [25]

$$p(\sigma)\sim\frac{1}{\sigma}\,. \tag{18}$$

These statisitical assumptions enable us to compute the posterior density of the system's parameters given the data $\boldsymbol{y}$ as

$$p(\boldsymbol{\beta},\sigma,\theta,\boldsymbol{s}|\boldsymbol{y}) = C\cdot\mathcal{L}(\boldsymbol{\beta},\sigma,\theta,\boldsymbol{s}|\boldsymbol{y})\cdot\frac{1}{\sigma}\,. \tag{19}$$

The normalization constant C ensures that the right hand side actually defines a normalized probability density. From this expression, various marginal posterior distributions may be obtained by integrating over the parameters that shall not be considered. We are mostly interested in the posterior distribution of the possible change point locations θ. To produce the posterior distribution of this quantity, we have to marginalize out all other variables. It turns out that all but the integral over the noise slopes $\boldsymbol{s}$ may be carried out explicitely. Thanks to the Gaussian nature of the $\boldsymbol{\beta}$ dependency we obtain

$$p(\sigma,\theta,\boldsymbol{s}|\boldsymbol{y}) \sim \frac{\sigma^{1-n}}{\sqrt{|\Omega||F^T\Omega^{-1}F|}} e^{-\frac{1}{2\sigma^2}\mathcal{R}^2}\,, \tag{20}$$

and

$$p(\boldsymbol{\beta},\theta,\boldsymbol{s}|\boldsymbol{y}) \sim \frac{\left[(\boldsymbol{y}-F\boldsymbol{\beta})^T\Omega^{-1}(\boldsymbol{y}-F\boldsymbol{\beta})\right]^{-\frac{n}{2}}}{\sqrt{|\Omega|}}\,. \tag{21}$$

Further marginalization may be performed to yield

$$p(\theta,\boldsymbol{s}|\boldsymbol{y}) = \int d\sigma d\boldsymbol{\beta}\; p(\boldsymbol{\beta},\sigma,\theta,\boldsymbol{s}|\boldsymbol{y}) \tag{22}$$

$$= C'\cdot\frac{\mathcal{R}^{-(n-2)}}{\sqrt{|\Omega||F^T\Omega^{-1}F|}}\,. \tag{23}$$

Again C' is a constant, that ensures the normalization of the right hand side to a probability density. Finally the posterior marginal distribution of θ can be computed by numeric evaluation of the following integral

$$p(\theta|\boldsymbol{y}) = \int d\boldsymbol{s}\; p(\theta,\boldsymbol{s}|\boldsymbol{y})\,. \tag{24}$$

In the same way the numeric θ integral may be performed to elaborate the posterior information about the involved slope parameters $\boldsymbol{s}$ of the heteroscedastic behavior around the change point

$$p(\boldsymbol{s}|\boldsymbol{y}) = \int d\theta\; p(\theta,\boldsymbol{s}|\boldsymbol{y})\,. \tag{25}$$

III. VALIDATION THE METHOD

In order to validate the method's performance in an idealized setting we use synthetic time series to discuss its ability to estimate the model's parameters and to elaborate the sensitivity of the estimates to data loss. We generate the time series via the LMM Equ.(6) and infer on the change point by computing the global marginal posterior density Equ.(24), i.e. over the interval of all

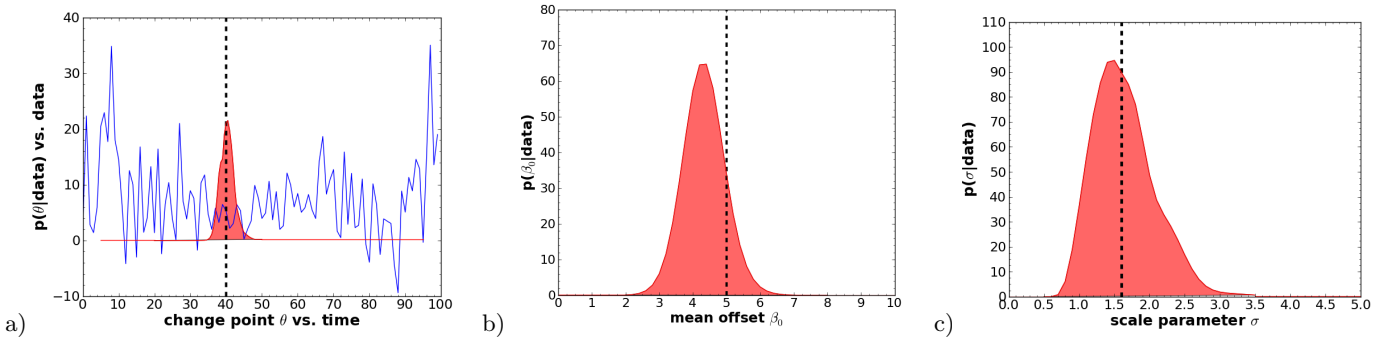

Figure 2: Normalized marginal posterior densities for the time series in Fig. **1**. The maxima indicate the most probable estimates of the a) change point $\hat{\theta} = 40.5$, b) fixed effect offset $\hat{\beta}_0 = 4.40$ and c) scale parameter $\hat{\sigma} = 1.50$. The dashed lines represent the true parameter values of the underlying model.

possible change point values θ. The location of the maximum of the marginal posterior density $[p(\theta|\boldsymbol{y})]_{\max}$ can be used as an estimator for the most probable location of a singularity $\hat{\theta}$. In case, the data contains more than one change point, the posterior distribution will exhibit multiple local maxima. This could therefore be used as an indicator for the existence of secondary change points in the time series. Although a more reasonable way would be to consider models with multiple change points this approach becomes quickly uncomputable due to exploding dimensionality. Thus we propose a local kernel based method to be able to apply our single change point model locally to multi change point data series.

A. Estimation of a single change point

To validate our technique, we apply it to the generated time series of Fig.**1** containing a single change point at $\theta = 40$. We compute all relevant two and one dimensional marginal distributions of the model's parameters using the formulas of the previous section. The marginal distributions provide Bayesian estimates for the change point θ, mean behavior $\boldsymbol{\beta}$, scale parameter σ and heteroscedastic behavior $\boldsymbol{s}$ of the data as the maxima of the one and two dimensional marginal distributions shown in Fig.**2**, **3**.

First note that due to the random nature of the observations, the posterior density too depends randomly on the actual series of observations. It is therefore not surprizing, that the locations of the maxima of the posterior does not exactly agree with the true parameter values. However, they are within a certain quantile of the posterior distribution. We automatically obtain confidence intervals or regions by considering those level intervals or contour-lines, that enclose a fixed percentage of the total probability. This yields a natural way of uncertainty quantification.

The estimated change point $\hat{\theta} = 40.5$ differs only little from the real value $\theta = 40.0$ within a relatively narrow and symmetric confidence interval $[35.7\,,\,45.9]$ (Fig.**2**a). Consequently we achieve to restrict the location of a probable singularity to a range $< 9\%$ of the time grid. The estimates of the mean behavior are obtained from Fig.**2**b, **3**a as $\hat{\boldsymbol{\beta}} = (4.40\,,\,0.206\,,\,0.096)$. The Bayesian estimates reproduce the real underlying mean model $\boldsymbol{\beta} = (5.0\,,\,0.22\,,\,0.08)$ convincingly. The two dimensional contour plot of the marginal density $p(\beta_1, \beta_2|\boldsymbol{y})$ of the mean slopes indicate an approximate symmetric confidence area of the most probable slope combinations (β_1, β_2) (red area in Fig.**3**a). The one dimensional projection $p(\beta_1|\boldsymbol{y})$ reveals a broader confidence interval for the estimation of $\hat{\beta}_1$ compared to $\hat{\beta}_2$. The scale parameter can be estimated as $\hat{\sigma} = 1.50$ from Fig.**2**c within the confidence interval $[0.806\,,\,2.84]$ unidirectional wider to growing σ-values and differs little from the true value $\sigma = 1.60$. The two dimensional contour plot of the marginal density of the deviation slope parameters $p(s_1, s_2|\boldsymbol{y})$ indicate a slight asymmetric confidence area of the most probable slope combinations (s_1, s_2) (red area in Fig.**3**b). The one dimensional projections $p(s_1|\boldsymbol{y})$ and $p(s_2|\boldsymbol{y})$ display unidirectional wider confidence bounds for the estimates $\hat{s_1} = 0.087$ to bigger and $\hat{s_2} = 0.167$ to smaller parameter values.

Table I: Estimated model of the synthetic signal of Fig. **1**

parameter	estimate	confidence $\geq 95\%$
$\hat{\theta}$	40.5	[35.7 , 45.9]
$\hat{\beta}_0$	4.40	[3.00 , 5.75]
$\hat{\beta}_1$	0.206	[0.035 , 0.390]
$\hat{\beta}_2$	0.096	[-0.015 , 0.189]
$\hat{\sigma}$	1.50	[0.806 , 2.84]
$\hat{s}_1$	0.087	[0.027 , 0.220]
$\hat{s}_2$	0.167	[0.050 , 0.380]

Thus for our realization, the marginal distributions of the heteroscedastic behavior (σ, s_1, s_2) indicate a broad range of probable parameter combinations compared to the mean behavior $\boldsymbol{\beta}$ or the change point θ. In Tab.**I** we summarize our point estimators and 95% confidence

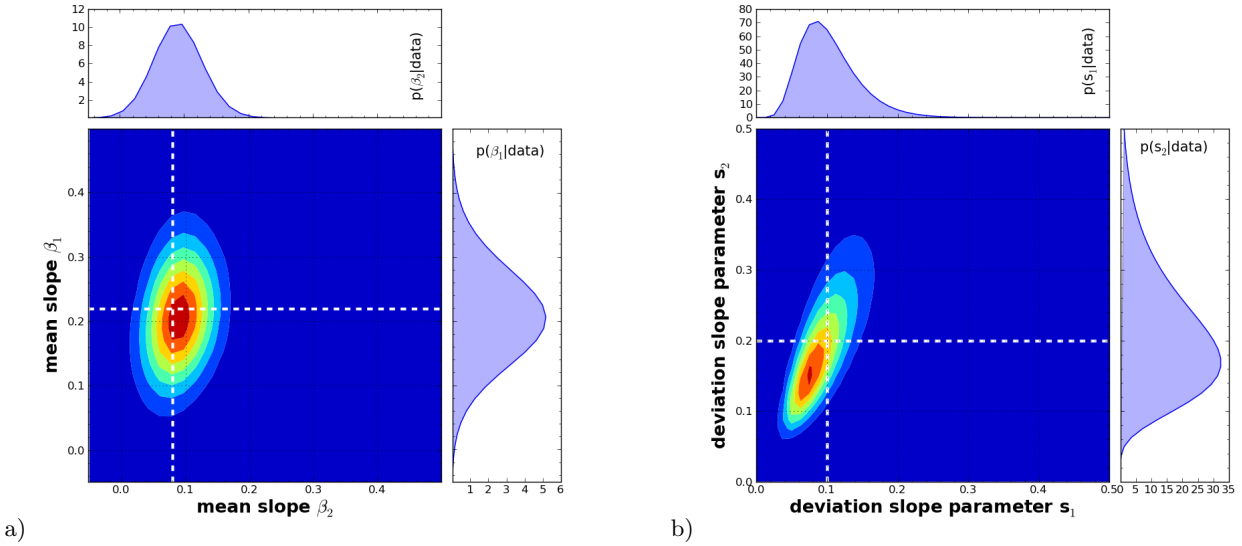

Figure 3: Normalized two dimensional marginal posterior densities for the time series in Fig.**1**. The maxima indicate the most probable estimates of the a) fixed effect slopes $(\hat{\beta}_1, \hat{\beta}_2) = (\,0.206\,,\,0.096\,)$ and b) deviation slope parameters $(\hat{s}_1, \hat{s}_2) = (\,0.087\,,\,0.167\,)$. Alongside the contour plots are presented the one dimensional projections of the posterior densities. The dashed lines represent the true values of the underlying model.

intervals for them based on our analysis.

1. *Sensitivity to data loss*

In real data, analysts have to deal with sparse and irregularly sampled data. Our technique does not require an uniform sampling grid of data points since from the beginning, it employs only the available data. As a validation for the sensitivity of our method to data loss, we randomly ignore stepwise 0% up to 87,5% of the time series modeled by a sequence of $n_{obs} = 200$ observations. The artifical time series undergo a change point $\theta = 80$ and are further parametrized by the mean $F_\theta \boldsymbol{\beta} = 12 + 0.24 \cdot \boldsymbol{\zeta}^\theta_- + 0.02 \cdot \boldsymbol{\zeta}^\theta_+$ and the deviation behavior $\sigma^2 \Omega_{\theta,\boldsymbol{s}} = \left[1.2(1 + 0.18 \cdot \boldsymbol{\zeta}^\theta_- + 0.04 \cdot \boldsymbol{\zeta}^\theta_+)\right]^2$. Leaving out randomly a defined percentage of the observations produces time series with random gaps and irregular sampling steps. For each of these random realizations consisting of n_{obs} data points we compute the posterior densities $p^i_{n_{obs}}(\theta|\boldsymbol{y})$ for $i = 1, \ldots, 50$ realizations.
The obtained averaged posterior densities $\langle p(\theta|\boldsymbol{y})\rangle_{n_{obs}}$ in the plane of the sample size n_{obs} are shown in Fig.**4**, indicating with their maxima the averaged most probable change points $\langle\hat{\theta}\rangle_{n_{obs}}$. Apparently the mean of the posterior densities differs from the true value, however still within the width of the distribution. The latter depends invers proportionaly on the square root of the sample size

$$\text{width}\left[\langle p(\theta|\boldsymbol{y})\rangle_{n_{obs}}\right] \propto \frac{1}{\sqrt{n_{obs}}} \quad . \tag{26}$$

At large numbers of sampling points n_{obs} the posterior converges towards a delta distribution located at the true parameter value $\theta = 80$. In any case, even for small data sets, as small as $n_{obs} = 25$, the non-flatness of the posterior clearly hints towards the existence of a change point in the time series. The investigation of the averaged marginal posterior densities in the plane of the remaining parameters reveals a broadening of the posterior distributions for $n_{obs} < 200$, as naturally expected due to information loss in the sub time series considered in the inference process.
Additionally we point out the efficiency of our method to infer on the explicit location of a singularity $\hat{\theta}^i_{n_{obs}}$ for every single time series of the previous setting. In Fig.**5** are presented the histograms of the global point estimators $\hat{\theta}^i_{n_{obs}}$ for every single realization $i = 1, \ldots, 50$. We observe that the particular global estimators $\hat{\theta}^i_{obs}$ are relatively robust to data loss and enable us to infer convincingly on the location of the singularity. Even considering only 50% of the full time series, i.e. $n_{obs} = 100$, produces global estimates that lie in the narrow interval $[76.0\,,\,83.5]$, representing $< 4\%$ of the full time grid.
However, for such a data-poor situation, local additional, less dominant maxima are likely to appear due to random fluctuations in the posterior, and more sofisticated techniqes are needed to assess the existence of single or multiple change points. One approach to clearify multimodial posterior densities is the computation of local posterior densities within a sliding window as presented in the following.

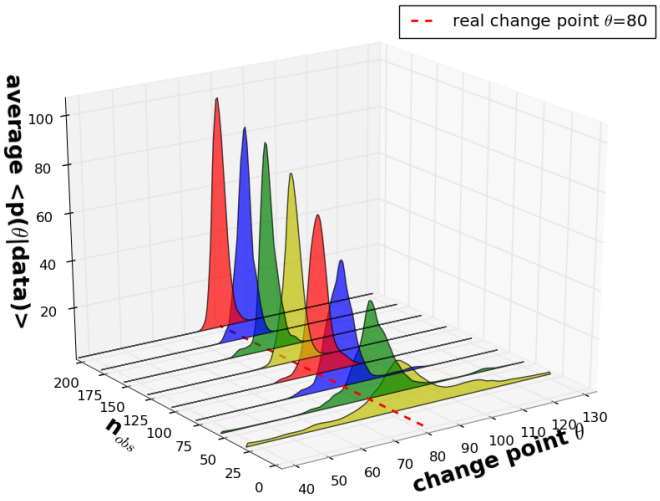

Figure 4: The global maxima of the averaged posterior densities $\langle p(\theta|\boldsymbol{y})\rangle_{n_{obs}}$ converge for increasing number of data points n_{obs} towards a delta distribution located at the true change point value $\theta = 80$.

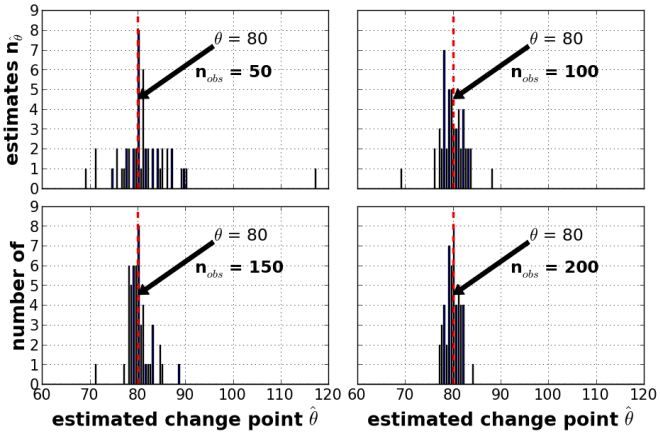

Figure 5: Histograms of the global change point estimators $\hat{\theta}^i_{obs}$ for $i = 1, \ldots, 50$ realizations and with respect to $n_{obs} = 50, 100, 150, 200$ data points from the setting of Fig.**4**. Even for $n_{obs} = 100$ nearly all global estimates $\hat{\theta}^i_{100}$ lie in the interval $[76.0\,,\,83.5]$, respectively $< 4\%$ of the full time grid.

2. *Local posterior density*

Long data sets are likely to contain more than one change point. So using our model globally may not be justified. However, locally our model assumption may still be valid. For this reason, we propose the following kernel based local posterior method. In addition this method allows us to treat very long data sets numerically more efficient since the computation scales with the the third power of the employed data points. Around each time point t we choose a data window $I_t = [t - \frac{T}{2}, t + \frac{T}{2}]$ of length T. Inside this window, we take as prior distribution for the change point location $p(\theta)$ a flat prior inside some subinterval of length a:

$$p(\theta) = \begin{cases} \frac{1}{a} & \text{for} \quad t - \frac{a}{2} \leq \theta \leq t + \frac{a}{2} \\ 0 & \text{else} \end{cases}, \quad 0 < a < T\,. \tag{27}$$

We then compute the local posterior $p_t(\theta|\boldsymbol{y}_{|I_t})$ around t based on the subseries in the data window $\boldsymbol{y}_{|I_t}$. This yields a posterior distribution of a possible change point within each window under the assumption that there is actually a singularity within the window. In order to compare different window locations, we need to quantify the credibility that there is a change point. Therefore we compute the maximum of the Likelihood within each window

$$f(t) = \max_{\theta \in [t-\frac{1}{a}, t+\frac{1}{a}], s_1, s_2 \in \mathbb{R}} \mathcal{L}(\boldsymbol{\beta}^*, \hat{\sigma}; \boldsymbol{y}_{|I_t}), \tag{28}$$

where $\hat{\sigma}$ and $\boldsymbol{\beta}^*$ are the estimators given by Eq.(15) and (13). The global distribution of change points θ given the full time series is then obtained as a weighted superposition in form of

$$p(\theta|\boldsymbol{y}) = C \cdot \int f(t)\, p_t(\theta|\boldsymbol{y}_{|I_t})dt\,, \tag{29}$$

whereas the constant C ensures the normalization to a probability density. In subdata sets with no change point, the credibility of the model fit is very low, in conclusion the Likelihood maxima is of very small value and local estimates are judged as negligible. By construction the method works for multiple change points as soon as they are separated by at least one data window. We demonstrate this by applying our algorithm first on a synthetic single change point time series. In Fig.**6** is shown the sum of the local posterior densities weighted by the maxima of the local Likelihood (dashed curve). The time series is one realization of the model in the previous Sect.III A 1 for a sequence of $n_{obs} = 200$ data points. Supplementary the applied window size $n_{obs} = 50$ and the sampling grid of the change points $n_{cp} = 30$ are presented for comparison. The sum of local posterior densities indicates the best model fit for windows covering the real change point $\theta = 80$ but is non-zero even between $[100, 121]$ suggesting that a change point model might be suitable for these singularity values as well.

A second quantity that may be used to produce relative credibility weights for the windows is given by the Bayes factor [26]. Besides the goodness of fit, the complexity of the assumed model has to be taken into account to assess the most capable model describing the data and thus performing the estimation. Thus we test the hypothesis of no change point, respectively a linear model $\mathcal{M}_{lin}$, against a change point model $\mathcal{M}_{cp}$ in form of the Bayes factor

$$BF(t) = \frac{p\left(\mathcal{M}_{lin}|\boldsymbol{y}_{|I_t}\right)}{p\left(\mathcal{M}_{cp}|\boldsymbol{y}_{I_t}\right)}\,. \tag{30}$$

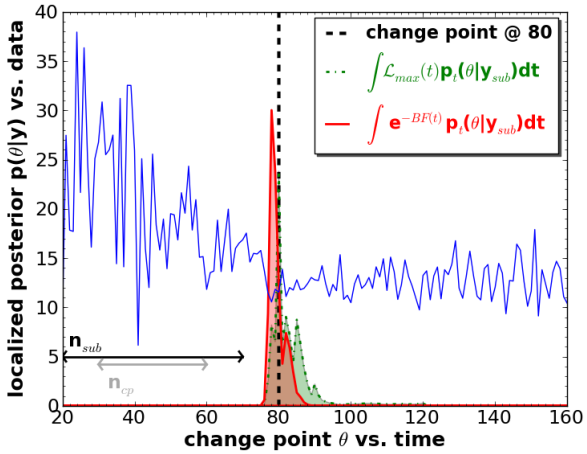

Figure 6: Normalized sum of local posterior densities weigted by the local Likelihood maxima (dashed) and with respect to the Bayes factor (solid), computed for sub time series of $n_{sub} = 50$ data points and a sampling grid of $n_{cp} = 30$ change points. The data is one realization of the time series defined in Sect.III A 1 for $n_{obs} = 200$ data points.

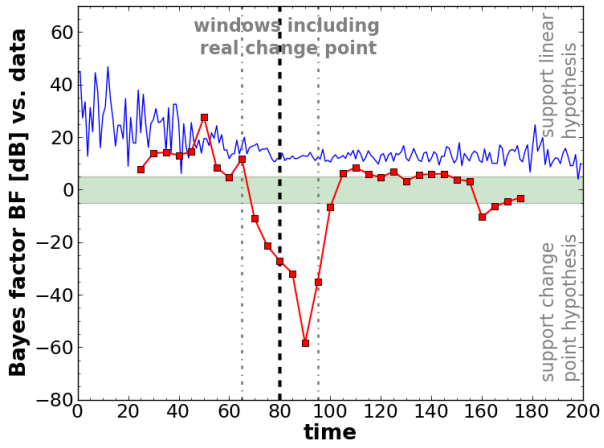

Figure 7: Local Bayes factor (squares) obtained for the time series in Fig.**6**. The shaded area encloses values whose support for none of the models is substantial (based on [26]). Values underneath this area strongly support a change point against a linear model, and vice versa for values above.

The dependency of the Bayes factor on a logarithmic scale is shown in Fig.**7** for the artifical time series of Fig.**6**. The Bayes factor in this test case favors the change point over the linear model for all local windows, for which the true change point is in the support of the inner prior distribution of θ. This local Bayes factor itself can be used as a diagnostic tool like the Likelihood weighted posterior, but we may also combine the techniques by using the BF as a window weighting function by setting $f(t) = e^{-BF(t)}$ in Eq.(29). In this form Eq.(29) corresponds therefore essentially to the total probability decomposition of the change point (cp)

$$\sum_{\text{windows}} p(\theta|\text{cp in window})\, p(\text{cp exists in window})\,. \tag{31}$$

For comparison of both kernel approaches we present in Fig.**6** additionally the sum of local posterior densities weighted by $e^{-BF(t)}$ (solid curve). The distribution weighted with respect to the Bayes factor are non-zero in the range between $[78, 89]$ whereas the one weighted by the maxima of the Likelihood is non-zero in $[78, 121]$. The long tail of the latter hints to less probable change point locations which are automatically rejected in the Bayes factor weighting. Furthermore we exemplify the algorithm on a synthetic multi change point time series shown in Fig.**8**. For clarity of presentation we plot the sum of posterior distributions weighted with the plain Bayes factor BF. We are able to infer on the true change point values $(\theta_1, \theta_2, \theta_3) = (40, 100, 160)$ via the estimators $(\hat{\theta}_1, \hat{\theta}_2, \hat{\theta}_3) = (38.9, 93.0, 162.9)$ within their intervals $([33.9, 47.8], [87.5, 109.1], [158.9, 167.0])$ of about 90% confidence. We obtain these intervals from a more detailed analysis of the partial sums of local posterior densities weighted by the factor e^{-BF} covering the estimated singularity locations.

The main advantage of this localization approach even in a single change point context is however the enormous speedup of the computations. For instance for a time series of $n_{obs} = 2000$ data points we pass from a global computation of the marginalized posterior density in $3h\,41min\,40s$ to a local one divided into 40 overlapping subdata sets of $n_{sub} = 100$ in $7min\,44s$, respectively

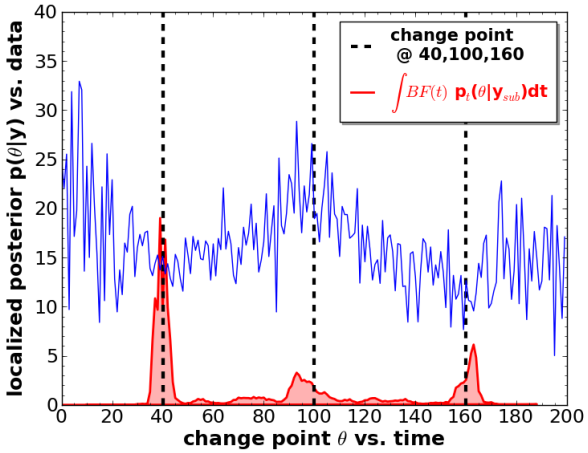

Figure 8: Normalized sum of local posterior densities weighted by the Bayes factor, computed for sub time series of $n_{sub} = 50$ data points and a sampling grid of $n_{cp} = 30$ change points. The parametrization of the mean is defined as $F\boldsymbol{\beta} = 14 + 0.2 \cdot \zeta_-^{40} + 0.1 \cdot \zeta_+^{40} - 0.25 \cdot \zeta_+^{100} + 0.3 \cdot \zeta_+^{160}$ and the deviation is modeled as $\sigma^2\Omega = \left[1.6(1 + 0.2 \cdot \zeta_-^{40} + 0.03 \cdot \zeta_+^{40} - 0.05 \cdot \zeta_+^{100} + 0.1 \cdot \zeta_+^{160})\right]^2$.

a speed up of about 95%. This is achieved using Python 2.6.5 on a Supermicro Intel(R) Core(TM)i7 CPU 920 @ 2.68GHz with 12GB RAM. In the context of complex multiple change point scenarios, as real time series mostly are, the localization approach of the posterior density $p(\theta|\boldsymbol{y})$ combined with the Bayes factor realizes a powerfull tool to scan the data seperately for single change points, as demonstrated in the following Sect.III B.

B. Annual Nile flow from 1871 to 1970

We demonstrate our technique by applying it on a time series including a known significant change point. For this purpose we analyze the annual Nile River flow measured at Aswan from 1871 to 1970 [27]. Several investigation methods have verified a shift in the flow levels starting from the year 1899 [4, 19, 27]. Historical records provide the fact, that this shift is attributed partly to weather changes and partly to the start of construction work for a new dam at Aswan. Since we expect a natural behavior of the underlying mean we generalize our previous model to undergo besides trend changes as well a sharp shift in the mean offset at the singularity θ. Therefore we modify the system matrix according to

$$F_\theta = \begin{pmatrix} (\varphi^\theta_-)_1 & (\zeta^\theta_-)_1 & (\zeta^\theta_+)_1 & (\varphi^\theta_+)_1 \\ \vdots & \vdots & \vdots & \vdots \\ (\varphi^\theta_-)_n & (\zeta^\theta_-)_n & (\zeta^\theta_+)_n & (\varphi^\theta_+)_n \end{pmatrix} , \quad (32)$$

whereas we define another type of Hockey sticks $\boldsymbol{\varphi}^\theta_-$ and $\boldsymbol{\varphi}^\theta_+$ referring to Eq.(2) and (3) not as linear but as constant. The general formulas of the Bayesian inference remain the same, with these new functions. First of all we compute the global posterior density $p(\theta, \boldsymbol{s}|\boldsymbol{y})$ as presented in Eq.(23). By initially guessing a reasonable sampling grid for the change point θ and the slope parameters $\boldsymbol{s}$ from the data, we clearly obtain significant maxima in the posterior projections $p(\theta|\boldsymbol{y})$ and $p(\boldsymbol{s}|\boldsymbol{y})$. Therefore we adjust the sampling grid to obtain finer posterior structures around the obvious maxima. We estimate the change point as $\hat{\theta} = 1898$ within a confidence interval $[1895, 1901]$ of over 95%. The slope parameters of the deviation are estimated as $(\hat{s}_1, \hat{s}_2) = (0.0065, -0.0015)$ within the 90% confidence intervals $\hat{s}_1$ in $[-0.0190, 0.0450]$ and $\hat{s}_2$ in $[-0.0065, 0.0855]$.

Prior the estimators $\hat{\theta}$ and $\hat{\boldsymbol{s}}$ we compute the posterior projections $p(\boldsymbol{\beta}, \theta, \boldsymbol{s}|\boldsymbol{y})$ and $p(\sigma, \theta, \boldsymbol{s}|\boldsymbol{y})$ formulated in Eq.(21) and (20). By minimizing the sampling grid of θ and $\boldsymbol{s}$ to its confidence intervals we are able to speed up the compuation and to estimate the remaining parameters $\boldsymbol{\beta}$ and σ. Finally we reveal from the global posterior distribution the most probable model plotted in Fig.**9** and listed in Tab.**II**.

Additionally we investigate the time series for local singularities by computing the sum of local posterior densities weighted by the Bayes factor as e^{-BF} (displayed in Fig.**9**)

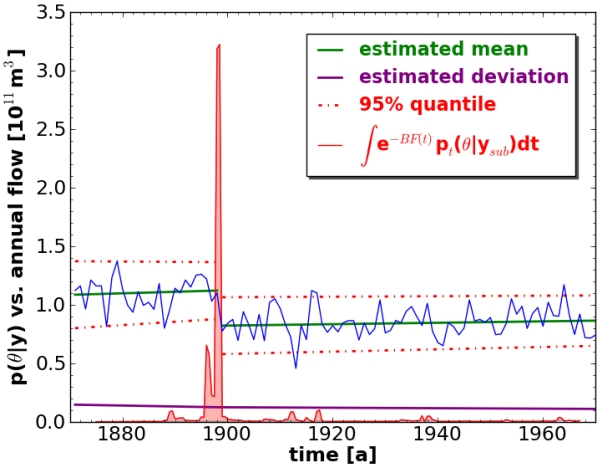

Figure 9: Annual Nile flow containing a known change point at $\theta = 1899$. The sum of localized posterior densities weighted with respect to the Bayes factor BF indicates a change point at $\hat{\theta} = 1898$ within its confidence interval $[1896, 1900]$ of about 90%. The estimated underlying model reveals the most dominant transition in the behavior of the mean.

Table II: Estimated model of the annual Nile flux.

parameter	estimate	confidence $\geq 90\%$
$\hat{\theta}$	1898	[1895 , 1901]
$\hat{\beta}_0$	1.12	[1.01 , 1.22]
$\hat{\beta}_1$	-0.0013	[-0.0082 , 0.0057]
$\hat{\beta}_2$	0.0006	[-0.0011 , 0.0024]
$\hat{\beta}_3$	0.82	[0.76 , 0.90]
$\hat{\sigma}$	0.124	[0.094 , 0.160]
$\hat{s}_1$	0.0065	[-0.0190 , 0.0450]
$\hat{s}_2$	-0.0016	[-0.0065 , 0.0855]

for the window sizes $n_{sub} = 50a$ of considered subseries. The change point sampling grid contains $n_{cp} = 30a$ in a resolution of $\Delta\theta = 0.5a$. Since most secondary maxima are $< 1\%$ we ignore them and therefore conclude on one global change point at $\hat{\theta} = 1898$ in the interval $[1896, 1900]$ of about 90% confidence. Note that we interpret the splitting of the global maximum as an artefact from the high resolution of the numerical change point sampling $\Delta\theta = 0.5a$.

In conclusion, we are able to confirm previous investigation techniques and auxiliary reveal further information from the parameter space of the multidimensional posterior density of the applied LMM.

IV. CONCLUSIONS

We introduce a general method for the detection of trend changes in heteroscedastic time series by describing the observations as a linear mixed model. The change

point is thereby considered as an isolated singularity in a regular background of a signal, assuming partial linear mean and deviation in the first order approach. By addressing the framework of linear mixed models we achieve to simplify the explicit computation of the marginal posterior distributions and thus reduce the computational time considerably. The formulation of the marginalized posterior densities of the model's parameters enables us to obtain inter alia the probability density of a change point given the data. Therefore the technique yields an insight in the parameter space of the underlying model, estimates these parameters and intrinsically provides a description of their confidence intervals.
We elaborate our technique for single change point models by infering on the relevant model parameters and discuss the sensitivity of the singularity estimator with respect to data loss. Additionally we present a kernel based approach to investigate more complex time series with multiple change points by localizing the posterior density and using the Bayes factor as a weighting function. Moreover we apply our algorithm on the annual flow volume of the Nile River at Aswan from 1871 to 1970. We confirm a well-established transition in the year 1899 by the estimated change point at 1898 within the interval $[\,1896\,,1900\,]$ of about 90% confidence. We specify the underlying model and identify the mean as the statistical property undergoing the most significant transition. We conclude by emphasizing that our algorithm depicts a powerfull tool to estimate the location of transitions in heteroscedastic time series and to infer on the underlying behavior in a partial linear approach, meanwhile reducing the computational time.

Acknowledgments

We thank M.H. Trauth for fruitful discussions and gratefully acknowledge financial support by DFG (GRK Nadi and GRK 1364) and the University of Potsdam.

[1] M.H. Trauth, J.C. Larrasoaña and M. Mudelsee, Quaternary Science Reviews **28**, (2009);
[2] M. Mudelsee and M.E. Raymo, Paleoceanography **20**, (2005);
[3] M.P. Girardin *et al.*, Global Change Biology **15**, (2009);
[4] P. Jong and J. Penzer, Journal of the American Statistical Association **93**, (1998);
[5] V.N. Minin and K.S. Dorman and Fang Fang and M.A. Suchard, Bioinformatics **21**, (2005);
[6] J.S. Liu and C.E. Lawrence, Bioinformatics **15**, (1999);
[7] P. Li and B.H. Wang, Physica A Statistical Mechanics and its Applications **378**, (2007);
[8] D.W.K. Andrews, Econometrica **61**, (1993);
[9] M. Mudelsee, European Physical Journal Special Topics **174**, (2009);
[10] L.R. Olsen, P. Chaudhuri and F. Godtliebsen, Computational Statistics and Data Analysis **52**, (2008);
[11] E. Moreno, G. Casella and A. Garcia-Ferrer, Stoch. Envron. Res. Risk Assess **19**, (2005);
[12] H. Liang, Bioinformatics **21**, (2009);
[13] R.V. Donner, Y. Zou, J.F. Donges, N. Marwan and J. Kurths, New Journal of Physics **12**, (2010);
[14] N. Marwan, J.F. Donges, Y.Zou, R.V. Donner and J. Kurths, Physics Letters A **373**, (2009);
[15] G. D'Agostini, Reports on Progress in Physics **66**, (2003);
[16] D.M. Bates and S. DebRoy, Journal of Multivariate Analysis **91**, (2004);
[17] A. Gelman, J.B. Carlin, H.S. Stern and D.B. Rubin, *Bayesian data analysis*, 2nd edition, Chapman & Hall/CRC Texts in Statistical Science, (2004);
[18] P. Fearnhead, Statistics and Computing **16**, (2006);
[19] A.B. Downey, *arXiv:0812.1237*, (2008);
[20] M.E. McCulloch, S.R. Searle and J.M. Neuhaus, *Generalized, Linear, and Mixed Models*, 2nd edition, Wiley, New York, (2008);
[21] G.K. Robinson, Statistical Science **6**, (1991);
[22] P.C. Mahalanobis, In Proceedings National Institute of Science **2**, (1936);
[23] R.E. Kass and A.E. Raftery, Journal of the American Statistical Association **90**, (1986);
[24] G. Wahba, Journal of the Royal Statistical Society. Series B (Methodological) **40**, (1978);
[25] H. Jeffreys, Royal Society of London Proceedings Series A **186**, (1946);
[26] R.E. Kass and A.E. Raftery, Journal of the American Statistical Association **90**, (1995);
[27] G.W. Cobb, Biometrika **65**, (1978);