Temporal Prediction Models from Marginal and Small Data Signal: the Otolith Data Example

Rajan Lamichhane* Norou Diawara* Cynthia Jones†

Abstract

Stochastic processes have applications in many areas such as oceanography and engineering. Special classes of such processes deal with time series of sparse data. Studies in such cases focus in the analysis, construction and prediction in parametric models.

Here, we assume several non-linear time series with additive noise components, and the model fitting is proposed in two stages. The first stage identifies the density using all the clusters information, without specifying any prior knowledge of the underlying distribution function of the time series. The effect of covariates is controlled by fitting the linear regression model with serially correlated errors. In the second stage, we partition the time series into consecutive non-overlapping intervals of quasi stationary increments where the coefficients shift from one stable regression relationship to a different one using a breakpoints detection algorithm. These breakpoints are estimated by minimizing the likelihood from the residuals. We approach time series prediction through the mixture distribution of combined error components. Parameter estimation of mixture distribution is done by using the EM algorithm. We apply the method to fish otolith data influenced by various environmental conditions and get estimation of parameters for the model.

Key Words: Bayesian, Regression models, reference distribution, likelihood, change points algorithm, mixture distribution.

1. Introduction

Stochastic processes for longitudinal data are fundamental in probability and statistics and have applications in many areas such as oceanography and engineering. Special classes of such processes deal with time series of sparse data. Studies in such cases focus on the analysis, construction and prediction in parametric models.

In this article, the prediction of time series is revisited and an application based on real data is given. The density uses all the clusters information, without specifying any prior knowledge of the underlying distribution function of time series. The effect of covariates is controlled by fitting the linear regression model with serially correlated errors. The change in stability of regression coefficients during the time course can be accounted by creating different breakpoints. We partition the time course into consecutive non-overlapping intervals where the coefficients shift from one stable regression relationship to a different one. These breakpoints are estimated by minimizing the residual sum of squares (RSS) using the algorithm described by Bai and Peron (2003) [6]. The algorithm in selecting the number of change points is based on a simple iterative step in which the maximum difference is less than a critical value of the difference of two consecutive values and is less than an optimal threshold chosen in a Bayesian framework. The partition algorithm fits a different probability model maximizing likelihood within each block interval.

Since different parts of data fit different models, forecasting depends not just on one model, but on all the relevant models. We develop a method based on mixture of different distributions to forecast in this type of models. The Expectation-Maximization (EM) algorithm, with initial values obtained from the empirical estimates, give the estimates of the

*Old Dominion University, Norfolk, Virginia, 4700 Elkhorn Ave., Norfolk, VA 23529.

†Old Dominion University, Norfolk, Virginia, 800 West 46th Street, Norfolk, VA 23508.

mixture distribution. Further improvement in the parameter estimation has been observed by using bootstrap re-sampling combined with EM algorithm. For simplicity, we name this method as Break Point Bootstrap Filtering (BPBF) method.
This paper is an extension of the ideas developed akin to the cited references and related work. It presents a novel concept in time series prediction and some supporting empirical evidence in terms of real data. The concept of using multiple break points based on minimum RSS or Bayesian Information Criteria (BIC) does not always create good intervals. Sometimes there are very few observations in some intervals and the estimates based on those observations are suspicious. In such cases, we improve the estimation of parameters by using block bootstrap. The block bootstrap is the most general method to improve the accuracy of bootstrap for time series data. By dividing the data into different blocks, it can preserve the original time series structure within a block. However, the accuracy of the block bootstrap is sensitive to the choice of block length, and the optimal block length depends on the sample size, the data generating process and the statistic considered. In our examples, we are using the approach proposed by Patton et al. (2009) [19] to identify the optimal block size. Varying block lengths that follow the geometric distribution are considered, and thus we avoid the problem of non-stationary by its construction (Politis and Romano (1994) [20]).

The paper is organized as follows. Section 2 presents the guidelines and theory of the different procedures in model fitting. The distributions of the models are specified, and our new method is provided. In Section 3, we apply our method to simulated and real data and get estimation of parameters as well as model forecasting. We conclude in Section 4 with some discussion.

2. Model Building

Partially observed time series models are studied under various conditions, e.g. State Space Models (Durbin and Koopman, (2001) [11]), Dynamic Models (West and Harrison (1997) [25]), and Hidden Markov Models (Cappe et al. (2005) [8]). All of these methods work if we have regular time series data where the model structure does not change locally. In other words, if the variance changes locally, then it is hard to build the model based on regular time series approach. However, there are cases where structural changes or breaks appear to affect models, for example in the evolution in key economic and financial time series such as output growth, inflation, exchange rates, interest rates and stock returns.[1] If data are collected over a long period of time, we are more likely to observe the structural change. This change could be the result of many possible factors such as institutional or technological changes, environmental changes, shifts in economic policy, or could even be due to large macroeconomic shocks such as the doubling or quadrupling of commodity prices experienced over the past decades.

One main goal that arises in the context of time-series forecasting of such models is to incorporate these different model structures to estimate the overall model parameters. We assume that if breaks have occurred in the past, surely they are also likely to happen in the future. Approaches that view breaks as being generated deterministically are not applicable when forecasting future events unless, of course, future break dates as well as the size of such breaks are known in advance. In most applications, this is not a plausible assumption and modelling of the stochastic process underlying the breaks is needed.

[1] A small subset of the many papers that have reported evidence of breaks in economic and financial time series includes Alogouskofis and Smith (1991) [1], Garcia and Perron (1996) [13], Koop and Potter (2001) [17], and Pastor and Stambaugh (2001) [18].

In this paper we provide a general framework for forecasting time series under structural breaks that is capable of handling the different above scenarios.

Regular time series linear model of responses $\boldsymbol{Y}$ based on predictors X can be defined as:

$$\boldsymbol{Y} = X\boldsymbol{\beta} + \boldsymbol{\zeta},$$

where the errors $\boldsymbol{\zeta}$'s are not independent and assume stationarity process.

Also, the lag h autocovariance for the $\boldsymbol{\zeta}$ is given by:

$$Cov(\zeta_t, \zeta_{t-h}) = Cov(\zeta_t, \zeta_{t+h}) = \gamma(h) = \sigma^2\rho_h,$$

and the $\boldsymbol{\zeta}$ follows an autoregressive moving average process of order (p, q), we denote as ARMA(p, q) which is:

$$\zeta_t - \phi_1\zeta_{t-1} - \phi_2\zeta_{t-2} - ... - \phi_p\zeta_{t-p} = Z_t + \theta_1 Z_{t-1} + ... + \theta q Z_{t-q},$$

with $\{Z_t\}$ being the white noise of the $\boldsymbol{\zeta}$ process and $\phi_1, ..., \phi_p$ and $\theta_1, ..., \theta_q$ are AR and MA components, respectively.

Also, we can further extend the model to autoregressive integrated moving average, ARIMA (p, d, q), where $\{\zeta_t\}$ satisfies a difference equation of the form

$$\phi(B)(1 - B)^d\zeta_t = \theta(B)Z_t,\ \{Z_t\}\ \sim WN(0, \sigma^2),$$

where $\phi(z)$ and $\theta(z)$ are polynomials of degrees p and q, respectively,
$\phi(z) \neq 0$ for $|z| \leq 1$, d is the difference indicator and B is the backshift operator.

For $d = 0$, an ARIMA(p, d, q) reduces to an ARMA(p, q) process.

If the structure of data is such that there is heterogeneous variance structure among different intervals then parameter estimates based on a regular time series model is very unrealistic. So we divide the data into different parts using multiple breakpoints. The foundation for estimating breaks in time series regression models was given by Bai (1994) [2] and was extended to multiple breaks by Bai ((1997a) [3] and (1997b) [4]) and Bai and Perron ((1998) [5] and (2003) [6]). The distribution function used for the confidence intervals for the breakpoints is given in Bai (1997b) [4]. The ideas behind this implementation are described in Zeileis et al. (2003) [26]. The break points are obtained by testing or assessing deviations from stability in the classical linear regression model

$$y_j = x_j^T\boldsymbol{\beta} + u_j,$$

where at time j, y_j is the observation of the dependent variable, $x_j = (1, x_{j1}, ..., x_{jk})^T$ is a $(k + 1) \times 1$ vector of observations of the independent variables, and u_j are $iid(0, \sigma^2)$, and $\boldsymbol{\beta}$ is the $(k + 1) \times 1$ vector of regression coefficients.

In many applications, it is reasonable to assume that there are m breakpoints, where the coefficients shift from one stable regression relationship to a different one. Thus, there are $m+1$ segments, $I_1, \cdots, I_{m+1}$ in which the regression coefficients are constant, and the model can be rewritten as:

$$y_j = x_j^T\boldsymbol{\beta_i} + u_j, \tag{1}$$

where $\boldsymbol{\beta_i}, i = 1, 2, \cdots, m + 1$ be the vector of regression coefficients within each segment, i denotes the segment index and $j = j_{i-1} + 1, ..., j_i$. In practice, the breakpoints are rarely given exogenously, but have to be estimated. They are estimated by minimizing the residual sum of squares (RSS) from equation (1). The algorithm for computing the optimal breakpoints given the number of breaks is based on a dynamic programming approach

based on the Bellman principle (Bellman (1952) [9]). The main computational effort is to compute a triangular RSS matrix, which gives the RSS for a segment starting at observation indexed j and ending at indexed $j^{'}$ with $j < j^{'}$. Also, the adjacent intervals separated by break points are significantly different.

Let I_i denote the i^{th} interval with density function $f_{ij}(y_{ij}, \theta_{\mathbf{i}})$ where $i = 1, 2, ..., m+1$ represents the number of intervals and $j = 1, ..., n_i$ represents the number of values within that interval and θ_i is the vector of time series parameters within each interval. Thus, we have $m + 1$ time series models and each model is based only on the data of corresponding interval. So our main challenge is to combine all this model information to create a common model that can be used for forecasting. Several studies have been done in the past to combine the multiple time series regression models. Qin (1993) [21], Qin and Lawless (1994) [22], Qin and Zhang (1997) [23], Gilbert (2000) [14], Zhang (2000) [27] and Fokianos et al. (2001) [12] worked on some semi-parametric methods. Recently, Kedem and Gagnon (2010) [16] further extended those ideas by showing the estimation of the probability distribution of a "reference" time series and using them in conditional prediction. All these aforementioned ideas use multiple time series regressions where different time series structures are related to different covariates but the ideas do not extend into the different time intervals.

2.1 Parameter estimation: EM algorithm and mixture of normal distributions

Let's assume that there are *m* break points, this gives us *m+1* time series intervals. Let I_i be the i^{th} interval with density function $f_i(y_{ij}, \boldsymbol{\theta}_i)$ where $i = 1, 2, ..., m+1$; $j = 1, \cdots, n_i$ and $\boldsymbol{\theta}_i$ is the vector of time series parameters within each interval.
Then,

$$\begin{gathered} y_{1,t_1}\text{=}f_1(\mathbf{z}_{1,t_1-1})\text{+}\zeta_{1,t_1},\, t_1 = 1, 2, ..., n_1\ . \\ \vdots \\ y_{(m+1),t_{m+1}}\text{=}f_{m+1}(\mathbf{z}_{m+1,t_{m+1}-1})\text{+}\zeta_{m+1,t_{m+1}}, \end{gathered}$$

where $t_{m+1} = t_m + 1, t_m + 2, ..., n_{m+1}$ and $\mathbf{z}_{i,t_i-1}$ contain past values of covariate time series possibly including even past values of y_{1,t_1}, ..., y_{m,t_m}, $y_{m+1,t_{m+1}}$.
The error sequence $\{\zeta_{i,t_i}\}$ is the sequence of iid random variables, $\zeta_{i,t_i} \sim g_i(y), i = 1, ..., m, m+1$.
We approach time series prediction through the mixture distribution of these error components. Noise from each of the intervals are combined to form combined noise:
$\boldsymbol{\zeta}$=$(\zeta_1, \zeta_2, ..., \zeta_n)$ =$\{(\zeta_{1,1}$, ... ,$\zeta_{1,n_1})$, ... ,$(\zeta_{i,1}$, ... ,$\zeta_{i,n_i})$, ... ,$(\zeta_{m+1,1}$, ... , $\zeta_{m+1,n_{m+1}})\}$,
$n = n_1 + ... + n_{m+1}$.
The joint density of finite mixtures is

$$g(y) = \sum_{i=1}^{m+1} p_i g_i(y),\ p_i \geq 0,\ i = 1, ..., m, m+1,$$

and

$$\sum_{i=1}^{m+1} p_i = 1.$$

Hence, the cumulative distribution function of combined data is

$$G(y) = \sum_{i=1}^{m+1} p_i G_i(y),$$

where $G_i(y)$ is the cumulative distribution function of vectors $\boldsymbol{\zeta}_i$, $\quad i = 1, ..., m, m+1$. Since, $y_{m,t+1}$=$f_m(\mathbf{z}_{m,t})$ +$\zeta_{m,t+1}$ and $\zeta_{m,t+1} \sim G$, we have the future probability approximation at $t+1$ conditional on $\mathbf{z}_{m,t}$ as:

$$\begin{aligned} P(y_{m,t+1} \leq y|\mathbf{z}_{m,t}) &= G(y - f_m(\mathbf{z}_{m,t}) \\ &\approx \hat{G}(y - f_m(\mathbf{z}_{m,t}) \\ &= \sum_{i=1}^{n} \hat{p}_i I(\zeta_i \leq y - f_m(\mathbf{z}_{m,t})), \end{aligned}$$

where $\hat{p}_i$ are the estimated weights using EM algorithm for each interval and I is the indicator function such that $\hat{G}(\zeta)$=$\sum_{i=1}^{n} \hat{p}_i I(\zeta_i \leq \zeta)$.
For ARMA(p, q) models in each interval, this conditional probability reduces to

$$P(y_{m,t+1} \leq y|\mathbf{z}_{m,t}) = \hat{G}(\sum_{i=1}^{n} \hat{p}_i (y - \sum_{j=1}^{p} \hat{\phi}_j y_{m,t+1-j} + \sum_{k=1}^{q} \hat{\theta}_k Z_{m,t+1-k})).$$

For each interval, without loss of generality, $\boldsymbol{\zeta}_i \sim N(0, \sigma_i^2)$.
For convenience, let's assume that i=2. Then, the density for the mixture of two Gaussian population is given as:

$$g_\zeta(\varsigma) = p\frac{1}{\sigma_1}\varphi\left(\frac{\boldsymbol{\varsigma} - \mu_1}{\sigma_1}\right) + (1-p)\frac{1}{\sigma_2}\varphi\left(\frac{\boldsymbol{\varsigma} - \mu_2}{\sigma_2}\right),$$

where φ is the cumulative distribution function of the standard normal distribution. We set indicators of which mixture component each observation belongs to as missing data and the EM algorithm will find the proportion of observations belonging to each normal distribution along with other unknown parameters for means and variances. $p, \mu_1, \mu_2, \sigma_1^2$ and σ_2^2 are the parameters to be estimated.
Let θ=$(p, \mu_1, \mu_2, \sigma_1^2, \sigma_2^2)$.
The indicator variable W_i can be treated as missing data such that:

$$W_i = \begin{cases} 1, & \text{if} \quad \zeta_i \text{ belongs to first interval} \\ 0, & \text{if} \quad \zeta_i \text{ belongs to second interval,} \end{cases}$$

where W_i is Bernoulli distributed with parameter p.
Therefore, the likelihood expression for complete data is given by:

$$L_n(\theta|\zeta, W) = \prod_{i=1}^{n} p^{W_i}(1-p)^{1-W_i} \frac{1}{\sigma_1^{W_i}}\varphi\left(\frac{\zeta_i - \mu_1}{\sigma_1}\right)^{W_i} \frac{1}{\sigma_2^{1-W_i}}\varphi\left(\frac{\zeta_i - \mu_2}{\sigma_2}\right)^{1-W_i}.$$

And the corresponding log-likelihood function for the density becomes:

$$\begin{aligned} l_n(\theta|\zeta, W) \quad = \quad & \sum_{i=1}^{n} W_i log(p) + \sum_{i=1}^{n}(1-W_i)log(1-p) - \frac{1}{2}\sum_{i=1}^{n} W_i log(2\pi\sigma_1^2) \\ & -\frac{1}{2\sigma_1^2}\sum_{i=1}^{n} W_i(\zeta_i - \mu_1)^2 - \frac{1}{2}\sum_{i=1}^{n}(1-W_i)log(2\pi\sigma_2^2) \\ & -\frac{1}{2\sigma_2^2}\sum_{i=1}^{n}(1-W_i)(\zeta_i - \mu_2)^2. \end{aligned}$$

From here, we apply the EM algorithm and find the expectation of W_i. The conditional distribution of W_i given $\boldsymbol{\zeta}$ is:

$$W_i|\zeta_i, \theta^{(k)} \sim Bin(1, p_i^{(k)}),$$

with

$$p_i^{(k)} = \frac{p^{(k)} \frac{1}{\sigma_1^{(k)}} \varphi(\frac{\varsigma_i - \mu_1^{(k)}}{\sigma_1^{(k)}})}{p(k) \frac{1}{\sigma_1^{(k)}} \varphi(\frac{\varsigma_i - \mu_1^{(k)}}{\sigma_1^{(k)}}) + (1 - p^{(k)}) \frac{1}{\sigma_2^{(k)}} \varphi(\frac{\varsigma_i - \mu_2^{(k)}}{\sigma_2^{(k)}})},$$

where $p^{(k)}$ is a set of known or estimated parameters at k^{th} step. The initial value $p^{(0)}$ can be obtained from the empirical distribution.
Hence, the conditional mean at k^{th} step is:

$$E(W_i|\zeta_i, \theta^{(k)}) = p_i^{(k)}.$$

By substituting $p_i^{(k)}$ for W_i, we obtain the expectation function as:

$$\begin{aligned} Q(\theta|\theta^{(k)}) &= \sum_{i=1}^{n} p_i^{(k)} log(p) + \sum_{i=1}^{n} (1 - p_i^{(k)}) log(1-p) - \frac{1}{2} \sum_{i=1}^{n} p_i^{(k)} log(2\pi\sigma_1^2) \\ &\quad - \frac{1}{2\sigma_1^2} \sum_{i=1}^{n} p_i^{(k)} (\zeta_i - \mu_1)^2 - \frac{1}{2} \sum_{i=1}^{n} (1 - p_i^{(k)}) log(2\pi\sigma_2^2) \\ &\quad - \frac{1}{2\sigma_2^2} \sum_{i=1}^{n} (1 - p_i^{(k)}) (\zeta_i - \mu_2)^2. \end{aligned}$$

Now, we maximize the expectation obtained in previous step. In the maximization step, we set the first derivative of $Q(\theta|\theta^{(k)})$ with respect to each parameter equal to zero and this results in the following equations for each parameter at the $(k+1)^{th}$ step:

$$p^{(k+1)} = \frac{1}{n} \sum_{i=1}^{n} p_i^{(k)},$$

$$\mu_1^{(k+1)} = \frac{\sum_{i=1}^{n} p_i^{(k)} \zeta_i}{\sum_{i=1}^{n} p_i^{(k)}}, \quad \mu_2^{(k+1)} = \frac{\sum_{i=1}^{n} (1 - p_i^{(k)}) \zeta_i}{\sum_{i=1}^{n} (1 - p_i^{(k)})},$$

$$\sigma_1^{(k+1)^2} = \frac{\sum_{i=1}^{n} p_i^{(k)} (\zeta_i - \mu_1^{(k+1)})^2}{\sum_{i=1}^{n} p_i^{(k)}} \quad , \quad \text{and} \quad \sigma_2^{(k+1)^2} = \frac{\sum_{i=1}^{n} (1 - p_i^{(k)}) (\zeta_i - \mu_2^{(k+1)})^2}{\sum_{i=1}^{n} (1 - p_i^{(k)})}.$$

The initial values of θ are again obtained from the empirical distribution.

2.2 Block Bootstrap

We use block bootstrap to generate bootstrap replicates of a statistic applied to time series. By dividing the data into several blocks, The original time series structure as well as the properties of original data generating process are preserved within a block.

Let $\{Y_t : t = 1, ..., n\}$ be time series data, then we construct bootstrap sample in the following steps:

1. Pick the optimal block size, l. The block size is chosen according to Patton et al. (Patton (2009) [19]).

2. Consider the overlapping blocks with varying block lengths. The optimal block size l is the mean of geometric distribution used to generate the block length. This avoids the problem of non-stationarity by construction (Politis and Romano (1994) [20]). For the overlapping method, we divide the data into $n - l + 1$ blocks, which block 1 being $\{Y_1, \cdots, Y_l\}$, block 2 being $\{Y_2, \cdots, Y_{l+1}\}, \cdots, etc.$

3. Resample the blocks randomly with replacement and generate bootstrap sample $\{Y_t^* : t = 1, \cdots, n\}$ by gluing blocks together in the order that they were sampled.

4. Calculate the estimator.

For simplicity, this combination of identification of breakpoints together with bootstrap is named as Breakpoints Bootstrap Filtering (BPBF) method.

3. Application and Estimation

We test the proposed methodology on simulated data and also apply it to actual data evaluated from fish otoliths. Otoliths are organs that detect sound and assist balance that are found in the inner ear. They are composed of calcium carbonate $(CaCO_3)$ and trace elements which reflect environmental conditions. Otoliths accrete daily bands for the first year of life and yearly bands thereafter (Jones (2002) [15] Chapter 2). Each band contains a fingerprint of the water chemistry to which the fish was exposed, and thus provides a chronology of changing habitat (Campana (1999)[7] and Dorval (2007)[10]). Our otolith data spans fish-birth years from 1967 to 2001. Otoliths were measured for $\delta^{18}O$ (the ratio of the stable isotopes ^{18}O:^{16}O), a measure of the oxygen isotopes contained in their $CaCO_3$, that mirrors water temperatures and origin. In our example, we use fish-otolith data collected from Lake Tasiat in eastern Canada, near the Arctic Circle. This region has experienced changing temperatures and precipitation that may reflect climate change. Covariates used are average precipitation (snow and rain), average temperature and average rainfall.

The simulation will allow us to justify our methodology. We simulate a combination of different ARMA models and use our proposed method for forecasting to a part of simulated data. Then, we validate our model by comparing the forecast result with remaining parts of the data.

3.1 Simulated Data

We simulate a time series data with different covariance structures in two different intervals. Combination of AR(1), and MA(2) is simulated. For the first interval we assume an AR(1) model, and for the second interval, an MA(2). The two models are generated with equal sample size, $n_1 = n_2 = 100$. For the AR(1) model, we use AR component $\phi_1 = 0.7$, zero mean with variance $\sigma_1^2 = 9$. The MA(2) components are $\theta_1 = 0.5$ and $\theta_2 = 0.4$, mean value of 2 and variance $\sigma_2^2 = 36$. Based on regular time series models, the best reasonable model with minimum AIC to fit the entire data is ARMA(1,2) with parameter estimates, $\hat{\sigma^2} = 26.72$, $\hat{\phi}_1 = 0.289$, $\hat{\theta}_1 = 0.433$ and $\hat{\theta}_2 = 0.398$. The AIC, BIC and log-likelihood for this model are $1233.61, 1251.96$, and -612.73, respectively.

Table 1: AICs for different ARMA (p, q) models for simulated data

$\downarrow AR \mid MA \longrightarrow$	0	1	2	3	4	5
0	NA	1294.756	1239.677	1240.411	1240.734	1238.576
1	1243.064	1244.882	1233.609	1239.873	1235.608	1237.507
2	1244.825	1245.34	1239.117	1235.608	1237.577	1239.526
3	1244.797	1242.76	1236.875	1237.66	1239.281	1233.936
4	1238.18	1240.125	1237.62	1237.101	1239.033	1235.583
5	1240.13	1242.18	1236.326	1235.514	1240.856	1238.491

We note that the forecast based on a usual time series model with minimum AIC can be improved. The model fit based on usual time series model in Figure 1 is unusually smooth and far from our expectation. In fact, its forecast does not explain the overall seasonal component.

We improve forecasting by identifying the break points where the data structures are different. Then, we fit different time series models for each intervals. The residuals from each intervals are combined and their joint density is estimated. For convenience, we assumed that these residuals are normally distributed. The parameters of mixture distribution are estimated by the EM algorithm. Further improvement in the parameter estimation is done by using parametric bootstrapping on the estimates obtained through EM algorithm. In our simulated data, the initial values for EM algorithm for mixture of two normal densities are taken as sample mean and variance of two error components. In our case, the two error components have zero sample means and variances are 13.73 and 320.82 for the first and second intervals, respectively. The estimated weights (proportion) using EM algorithm are 0.51 and 0.49 instead of 0.5 each. Bootstrap combined with EM algorithm gives the estimate of means and variances of the mixture distribution as $\hat{\mu_1} = -0.11$, $\hat{\mu_1} = 0.11$, ${\hat{\sigma_1}}^2 = 13.45$ and ${\hat{\sigma_2}}^2 = 324.73$, respectively. These estimates are used to generate the mixture distribution for forecasting. Model fit by using break points and forecasting is shown in Figure 1.

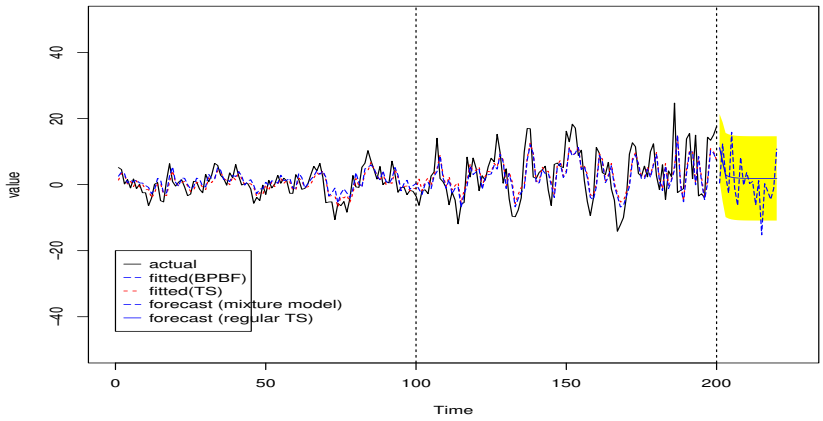

Figure 1: *Model fit for simulated data.*

For the first 100 data, AR(1) model with intercept -0.379, $\hat{\theta} = 0.494$, $\hat{\sigma}^2 = 10.47$, the AIC, BIC and log likelihood are 524.95, 532.76 and -259.47, respectively fit the data and

for the remaining 100 data, ARMA(2,2) with model parameter estimates, intercept of 5.46, $\hat{\sigma}^2 = 42.14$, $\hat{\theta_1} = 0.52$, $\hat{\theta_2} = 0.31$, $\hat{\phi_1} = 0.57$, $\hat{\phi_2} = 0.41$, the AIC, BIC and log likelihood of 672.32, 687.95 and -330.16 fit the data. Also, the root mean square errors for first 100 and last 100 observations are 3.24 and 6.49, respectively. Fitted model and forecasting based on the mixture model is given in Figure 1. Forecasting based on the mixture model with time and intercept adjustment looks much more reasonable compared to that based on regular time series model in Figure 1. We compared the similarity of cumulative distribution function (CDF) of simulated data and our mixture model. Kolmogorov-Smirnov tests shows that these two CDF are not significantly different (p=0.167) at 5 percent level of significance. Figure 2 compares the empirical CDF of simulated data with those from classical time series model and proposed mixture model.

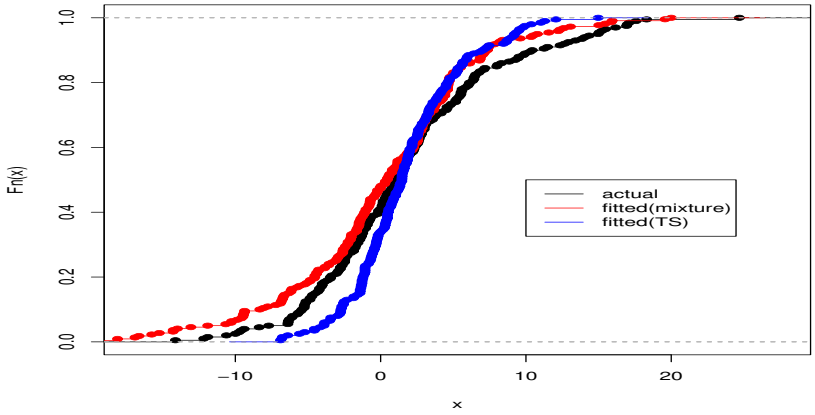

Figure 2: *Comparison of empirical CDFs for simulated data.*

3.2 Otolith Data

We implement our methodology using to Otolith data obtained from Lake Tasiat, Canada. The study of O18 (Oxygen Isotope $\delta^{18}O$) from fish otoliths is useful in estimating historical water temperature and weather. Lake Tasiat has information from years 1967 to 2000. Our main interest is to see the overall change of O18 over time and predict how it will behave in future. Average precipitation, average temperature and average rain are the available covariates which may cause changes in O18. Table 2, shows the summary of data from Lake Tasiat. The effects of these covariates are not significant in the linear regression model. The regression coefficients of average precipitation, average temperature and average rain are 0.0355, -0.00035 and 0.021, respectively.

Table 2: Mean and standard deviation of covariates.

Avg. O18		Avg. Temp.(^oC)		Avg. Rain (mm)		Avg. Prec. (mm)	
Mean	sd	Mean	sd	Mean	sd	Mean	sd
-12.62	0.33	-5.70	1.25	22.81	4.34	43.88	5.02

Figure 3 shows the original Lake data together with the best fitted model using regular time series model and our proposed model. We see that the model based on the usual time series does not fit the data very well so we use the breakpoints. The break points are

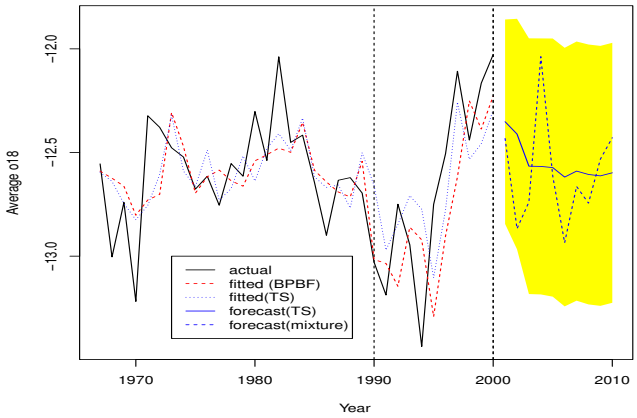

Figure 3: *Actual data and fitted model using classical time series model ARMA(3,2) and proposed model, BPBF.*

identified based on minimum BIC and RSS. Figure 4 shows the BIC and RSS for different breakpoints for Lake Tasiat. Based on minimum BIC, we use one breakpoint for Lake Tasiat. The data are divided into two groups: group 1 with first $1-23$ observations, and group 2 with and $24-34$ observations. Such consecutive groups are significantly different.

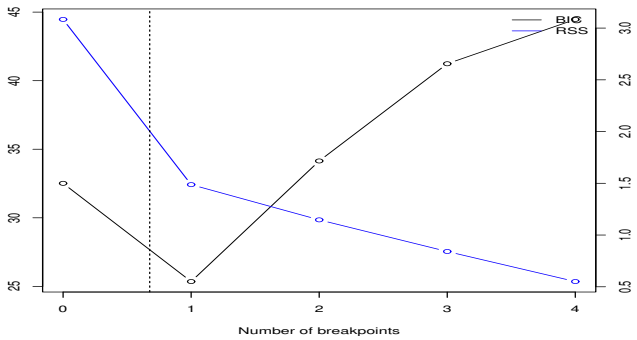

Figure 4: *Breakpoint identification for lake Tasiat using minimum BIC criterion.*

By using breakpoints, a MA(2) fits the first interval of Lake Tasiat data with parameters $\hat{\phi}_1 = 0.251$, $\hat{\phi}_2 = 0.519$, $\hat{\sigma^2} = 0.045$, AIC of 2.424 and log-likelihood of 9.20. An ARIMA$(0,1,1)$ fits the remaining data of Lake Tasiat. The fitted model parameters are $\hat{\theta}_1 = -0.281$, $\hat{\sigma^2} = 0.013$ and AIC of 1.94 and log-likelihood of 75.30.

Also, by using a classical time series model without the break points, ARMA$(3,2)$ fitted model was found as the best where parameters are $\hat{\phi}_1 = -0.628$, $\hat{\phi}_2 = -0.082$, $\hat{\phi}_3 = 0.319$, $\hat{\theta}_1 = 1.178$, $\hat{\theta}_2 = 1.000$, $\hat{\sigma^2} = 0.078$, AIC of 19.05 and log-likelihood of -2.20.

For forecasting Lake Tasiat, the mixture of normal distribution has estimates $\hat{p} = 0.671$, $\hat{\mu_1} = 0.004$, $\hat{\mu_2} = -0.01$, $\hat{\sigma_1}^2 = 0.053$ and $\hat{\sigma_2}^2 = 0.066$. In Figure 3, we can see the fitted model using break points and forecasting based on mixture model for Lake Tasiat. Figure 5 compares the empirical CDFs of Tasiat data with classical time series model and proposed mixture model. Our fitted model is significantly closer to the true nature of the small data.

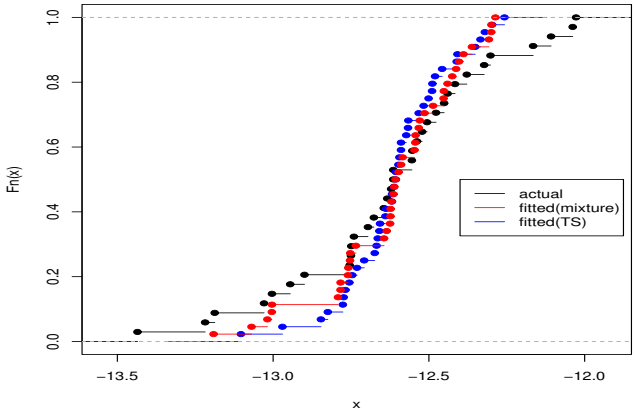

Figure 5: *Empirical cumulative distribution functions of Lake Tasiat together with classical time series model and the proposed mixture model.*

4. Conclusion

We have discussed methodology and analysis of time series data with locally changing variance (structural change) using breakpoints and bootstrap approaches. Breakpoints partition the time course into consecutive non-overlapping intervals where the coefficients shift from one stable regression relationship to a different one. Also, because there are limited observations in some intervals, we can use block bootstrapping to improve the parameter estimates. The optimal size of the blocks needed is chosen such that the RSS will be minimum. Once we fit the model for different intervals, such information is combined and used in the forecasting.

Forecasting partitioned data which has different model structures at different partitions is a challenging task. To our knowledge, there are no existing methods that discuss this problem. Our proposed method is different from other existing methods that are based on time series data where different covariates have different covariance structures. We have developed a new approach which advances previous concepts with new ideas for forecasting time series data that are subject to the structural breaks and non-equidistant time. Our approach is based on the mixture distribution where the parameters are estimated by using EM algorithm combined with bootstrapping. Our approach together with block bootstrapping performs very well when faced with small and sparse data sets as we have shown in our real example. Our approach is quite general and can be implemented in different ways other than those documented.

Further questions are being explored. One of the questions is related to the identification of optimal block size for block bootstrapping. Patton et al. (2009) [19] discussed the identification of optimal block sizes, but their approach still has some limitations. Another concern is related to finding a procedure of choosing initial value in EM algorithm for faster convergence.

References

[1] Alogoskoufis, G.S. and Smith, R. (1991), "The Phillips Curve, the persistence of Inflation, and the Lucas Critique: Evidence from Exchange Rate Regimes," *American Economic Review*, 81, 1254-1275.

[2] Bai, J. (1994), "Least Squares Estimation of a Shift in Linear Processes," *Journal of Time Series Analysis*, 15, 453-472.

[3] Bai, J. (1997a), "Estimating Multiple Breaks One at a Time," *Econometric Theory*, 13, 315-352.

[4] Bai, J. (1997b), "Estimation of a Change Point in Multiple Regression Models," *Review of Economics and Statistics*, 79, 551-563.

[5] Bai J., and Perron, P. (1998)," Estimating and Testing Linear Models With Multiple Structural Changes," *Econometrica*, 66, 47-78.

[6] Bai, J., and Perron, P. (2003), "Computation and Analysis of Multiple Structural Change Models," *Journal of Applied Econometrics*, 18, 1-22.

[7] Campana, S.E. (1999), "Chemistry and composition of fish otoliths: pathways, mechanisms and applications," *Marine Ecology Progress Series*, 188, 263-297.

[8] Cappè, O., Moulines, E. and Rydèn T. (2005), *Inference in Hidden Markov Models*, Springer, 2005.

[9] Bellman R. (1952), " On the Theory of Dynamic Programming," *Proceedings of the National Academy of Sciences*, 1952.

[10] Dorval, E., Jones, C.M., Hannigan, R., and J. and Van, J.M. (2007), "Relating otolith chemistry to surface water chemistry in a coastal plain estuary," *Canadian Journal of Fisheries Aquatic Sciences*, 64, 1-14.

[11] Durbin, J. and Koopman, S.J. (2001), *Time Series Analysis by State Space Methods*, Oxford University Press, 2001.

[12] Fokianos, K., Kedem, B., Qin, J. , and Short, D.A. (2001), "A Semiparametric Approach to the One-Way Layout," *Technometrics*, 43, 56-65.

[13] Garcia, R. and Perron, P. (1996), "An Analysis of the Real Interest Rate under Regime Shifts," *Review of Economics and Statistics*, 78, 111-125.

[14] Gilbert, P.B. (2000), "Large Sample Theory of Maximum Likelihood Estimation in Semiparametric Biased Sampling Models," *Annals of Statistics*, 28, 151-194.

[15] Jones, C.M. (2002), *Age and Growth*, , in Fisheries Science, [Editors] L.A. Fuiman and R.G. Werner, Blackwell Scientific, 30-63.

[16] Kedem, B. and Gagnon, R. (2010), "Semiparametric Distribution Forecasting," *Journal of Statistical Planning and Inference*, 140(2010) pp. 3734-3741.

[17] Koop, G. and Potter, S. (2001), "Are Apparent Findings of Nonlinearity Due to Structural Instability in Economic Time Series?," *Econometric Journal*, 4, 37-55.

[18] Pastor, L. and Stambaugh, R.F. (2001), "The Equity Premium and Structural Breaks," *Journal of Finance*, 56, 1207-1239.

[19] Patton, A. , Politis D. N., and White H. (2009), " "CORRECTION TO "Automatic block-length selection for the dependent bootstrap by D. Politis and H. White," *Econometric Reviews* 28(4), 372-375.

[20] Politis, D.N. and Romano, J.P. (1994), "The stationary bootstrap," *Journal of American Statistical Association*, 89: pp. 1303-1313.

[21] Qin, J. (1993), "Empirical Likelihood in Biased Sampling Problems," *Annals of Statistics*, 21, 1182-1186.

[22] Qin, J. and Lawless, J.F. (1994), "Empirical Likelihood and General Estimating Equations," *Annals of Statistics*, 22, 300-325.

[23] Qin, J. and Zhang, B. (1997), "A Goodness of Fit Test for Logistic Regression Models Based on Case-Control Data," *Biometrica*, 84, 609-618.

[24] Venkatraman and Olshen (2007), " bcp: an R Package for Performing a Bayesian Analysis of Change Point Problems," *Journal of Statistical Software*, 23 (3), pp. 1-13.

[25] West, M. and Harrison, J. (1997), *Bayesian Forecasting and Dynamic Models*, Springer, 1997 (2nd Ed.).

[26] Zeileis, A., Leisch, F., Hornik, K., Kleiber, C. (2003), "strucchange: An R Package for Testing for Structural Change in Linear Regression Models," *Journal of Statistical Software*, 7(2), 1-38.

[27] Zhang, B. (2000), "M-Estimation Under a Two Sample Semiparametric Model," *Scandinavian Journal of Statistics*, 27, 263-280.