Mixture Models for Prediction from Time Series, with Application to Energy Use Data

Najla M. Qarmalah, Jochen Einbeck and Frank P. A. Coolen

Abstract This paper aims to use mixture models to produce predictions from time series data. Given data of the form (t_i, y_i), $i = 1, \ldots, T$, we propose a mixture model localized at time point t_T with the k-th component as $y_i = m_k(t_i) + \varepsilon_{ik}$ with mixing proportions $\pi_k(t_i)$ such that $0 \leq \pi_k(t_i) \leq 1$ and $\sum_{k=1}^{K} \pi_k(t_i) = 1$, where K is the number of components. The $m_k(\cdot)$ are smooth unspecified regression functions, and the errors $\varepsilon_{ik} \sim N(0, \sigma^2)$ are independently distributed. Estimation of this model is achieved through a kernel-weighted version of the EM-algorithm, using exponential kernels with different bandwidths (neighbourhood sizes) h_k as weight functions. By modelling a mixture of local regressions at a target time point t_T but with different bandwidths h_k, the estimated mixture probabilities are informative for the amount of information available in the data set at the scale of resolution corresponding to each bandwidth. Nadaraya-Watson and local linear estimators are used to carry out the localized estimation step. For prediction at time point t_{T+1}, adequate methods are provided for each

Najla M. Qarmalah
Durham University, The UK,
✉ najla.qarmalah@durham.ac.uk

Jochen Einbeck
Durham University, The UK
✉ jochen.einbeck@durham.ac.uk

Frank P. A. Coolen
Durham University, The UK
✉ frank.coolen@durham.ac.uk

ARCHIVES OF DATA SCIENCE (ONLINE FIRST)
KIT SCIENTIFIC PUBLISHING
Vol. 2, No. 1, 2017

DOI 10.5445/KSP/1000058749/07
ISSN 2363-9881

local method, and compared to competing forecasting routines. The data under study give the energy use for Bolivia, Lebanon, and Greece from 1971 to 2011.

1 Introduction

Mixture models play an important role in the statistical analysis of data thanks to their flexibility to model a wide variety of random phenomena. They have been successfully employed in marketing and econometrics (Frühwirth-Schnatter, 2001) as well as biology and epidemiology (Green and Richardson, 2002), to name a few out of a huge number of fields of application.

One useful type of mixture models is the mixture of regression models. Mixtures of regression models are appropriate to use when the observations are from several subgroups with missing grouping identities, and in each subgroup, the response has a linear relationship with one or more other recorded variables. Many efforts have been made to extend such models as finite mixtures of generalized linear models which are comprehensively discussed by McLachlan and Peel (2004). Bayesian approaches for mixture regression models are summarized by Frühwirth-Schnatter (2006). Mixture models continue to be a topic of intense research activity, with special issues being edited in close succession (Böhning et al, 2014; Hinde et al, 2016). A large proportion of articles in those special issues discusses variants of mixture regression models, such as Poisson regression, spline regression, or regression under censoring.

Recently, mixtures of nonparametric regression models, which relax the linearity assumption on the regression functions, have gained particular attention. For example, Young and Hunter (2010) use kernel regression to model covariate-dependent proportions for mixtures of linear regression models, an idea which was further developed into a semi-parametric approach by Huang and Yao (2012). Huang et al (2013) have proposed a nonparametric finite regression mixture model where the mixing proportions, the mean functions, and the variance functions are all nonparametric, with application on the U.S. house price index (HPI) data. However, to our knowledge, there is no statistical method for prediction from time series based on mixture models and nonparametric regression. Nonparametric regression is a technique for modelling (possibly non-linear) trends in data. One approach to nonparametric regression is local modelling which locally estimates the mean function $m(t)$ using a set of parametric models. One of the most popular estimators of $m(t)$ is the Nadaraya-

Watson estimator or local constant regression estimator which is a special case of local polynomial regression (Fan and Gijbels, 1996).

The paper presented here aims to use a mixture of non-parametric regression models to produce predictions from time series data. Estimation of this model is achieved through a kernel-weighted version of the EM-algorithm, using exponential kernels with different bandwidths as weight functions. Nadaraya-Watson and local linear estimators are used to carry out the localized estimation step. In the first model, this forecast can be calculated directly from historical data as a local average of observed past values, with the size of the local neighborhood and the specific weights on the values defined by an exponential kernel. In the second model, the forecast is based on the fitted intercept and slope in the local neighborhood preceding the forecast point.

The rest of this paper is structured as follows. We present the main concepts in Sect. 2. In particular, in Sect. 2.1 we explain two popular estimators for nonparametric regression, which are the Nadaraya-Watson estimator and the local linear estimator. We define mixture models in Sect. 2.2, and we show in Sect. 3 how they can be used for prediction. In Sect. 4, we consider real data giving the energy use (kg of oil equivalent per capita) for Bolivia, Lebanon and Greece from 1971 to 2011 (recorded by the IEA[1]), and will compare our results to point forecasts obtained by Holt's exponential smoothing and ARIMA models. Finally, we provide conclusions in Sect. 5.

2 Main concepts

2.1 Non-parametric regression

In nonparametric regression models, restrictive assumptions on the functional form of the regression function are avoided. In simple words, we estimate the regression function by using data to find out more about it. An important special case of the general model is nonparametric simple regression, where there is only one predictor, that is $y_i = m(t_i) + \varepsilon_i$, where $m(\cdot)$ is the regression function and the errors ε_i are assumed to be normally distributed with mean 0 and constant variance σ^2.

Among the many ways to formulate an estimator $\hat{m}(\cdot)$ of $m(\cdot)$, an attractive technique is local fitting. The estimators of $m(\cdot)$ considered here depend on

[1] International Energy Agency, Available at: http://www.iea.org/

kernel regression, where localization is achieved through the use of a kernel (or weight) function W and a bandwidth parameter h, which controls the size of the local neighborhood and can be chosen to be constant or to depend on location. Kernel regression can be viewed as a method of computing weighted averages of the response variable in a fixed neighborhood around a target point, say t_T, the width of this neighborhood being governed by the bandwidth h. The exact form of weighting is determined by W that weights observations nearby t_T more heavily, and might be such that observations that are far away get zero weight (Fan and Gijbels, 1996).

2.1.1 Nadaraya-Watson estimator

The Nadaraya-Watson (or local constant) estimator is a special case of a larger class of kernel regression estimators which corresponds to a local constant least squares fit. It can be seen as a weighted local average of the response variables y_i. It shares this property with several other smoothing techniques (Fan and Gijbels, 1996).

Let $W_h = W(\cdot/h)/h$ be a weight function and $h > 0$ be a bandwidth. The Nadayara-Watson kernel regression estimator is given by

$$\hat{m}(t) = \frac{\sum_{i=1}^{n} W_h(t_i - t_T) y_i}{\sum_{i=1}^{n} W_h(t_i - t_T)}.$$

2.1.2 Local linear estimator

The local linear estimator applies the linear regression model locally to a fraction of the data around a given point t_T. Then the model is

$$y_i = \beta_0(t_T) + \beta_1(t_T)(t_i - t_T) + \varepsilon_i$$

where the dependence of the parameters β_0 and β_1 on t_T is emphasized. Then, the estimated regression curve at point t_T is (Fan and Gijbels, 1996)

$$\hat{m}(t_T) = \frac{\sum_{i=1}^{n} w_i y_i}{\sum_{i=1}^{n} w_i}$$

where $w_i = W_h(t_i - t_T)\,[S_{n,2} - (t_i - t_T)S_{n,1}]$ and $S_{n,j} = \sum_{i=1}^{n} W_h(t_i - t_T)(t_i - t_T)^j$.

2.2 Mixture models

Assume a random variable Y with density $f(y)$ is described as a finite mixture of K probability density functions $f_k(y)$, $k = 1, \ldots, K$, such that

$$f(y) = \sum_{k=1}^{K} \pi_k f_k(y)$$

with masses (or mixing proportions) $\pi_1, \ldots, \pi_K$ with $0 \leq \pi_k \leq 1$ and $\sum_{k=1}^{K} \pi_k = 1$. We refer to $f_k(\cdot)$, which may depend on a parameter vector θ_k, as the k-th component of the mixture of probability density functions.

3 Prediction using mixture models

In this section, mixture models are developed to forecast one-step-ahead and multi-step-ahead using two kinds of non-parametric estimators for the localized estimation step. The local constant estimator and local linear estimator are used. We refer to these models as mixture model using local constant smoothers (MLC) and as mixture model using local linear smoothers (MLL).

3.1 Mixture models using local constant kernel estimators (MLC)

For a time series of the form $(t, y) \in \{(t_i, y_i) : i = 1, \ldots, T\}$ we consider a localized mixture of K nonparametric regressions $m_k(t_i)$, $k = 1, \ldots, K$. At time point t_T, we define a locally constant model $m_k(t_i) \approx m_k(t_T)$, where the $m_k(t_T)$ play the role of parameters and are denoted as $\beta_k(t_T)$ henceforth. Then the model can be written as

$$y_i = \begin{cases} \beta_1(t_T) + \varepsilon_{i1}, & \text{with probability } \pi_1(t_T); \\ \vdots & \\ \beta_K(t_T) + \varepsilon_{iK}, & \text{with probability } \pi_K(t_T); \end{cases}$$

where K is a constant number of components, $\beta_1(t_T), \ldots, \beta_K(t_T)$ are constants which depend on the target point t_T, $\pi_k(t_T)$ is the proportion of the k-th component such that $0 \leq \pi_k \leq 1$ and $\sum_{k=1}^{K} \pi_k = 1$, and the errors $\varepsilon_{ik} \sim N(0, \sigma^2)$

are independently distributed. For ease of notation, we will often suppress the dependence of the parameters on t_T.
For given component k, we wish to obtain estimators of π_k, β_k and σ at time t_T. In the estimation step, the Expectation-Maximization (EM) algorithm is used which is a common method for mixture models. Let G be the random vector which draws a class $k \in 1,\ldots,K$, where

$$G_{ik} = \begin{cases} 1, & \text{if observation } i \text{ belongs to component } k; \\ 0, & \text{otherwise.} \end{cases}$$

We know that $P(G=k) = \pi_k$. Denoting

$$f_{ik} \equiv P(y_i|G=k) = \frac{1}{\sqrt{2\pi\sigma^2}} \exp\left(-\frac{(y_i-\beta_k)^2}{2\sigma^2}\right),$$

we also know that

$$P(y_i, G=k) = P(y_i|G=k)P(G=k) = f_{ik}\pi_k.$$

We introduce one-sided component-wise weight functions W_k anchored at t_T as follows:

$$W_k(t_i,t_T) = \begin{cases} \frac{\exp\left(\frac{t_i-t_T}{h_k}\right)}{h_k} & t_i - t_T \leq 0 \\ 0, & \text{otherwise.} \end{cases} \tag{1}$$

Assume now that, for an observation y_i, the value of G is known, i.e. we know to which of the K components the i-th observation belongs. This gives "complete" data $(y_i, G_{i1},\ldots,G_{iK})$, $i=1,\ldots,n$, with local probability

$$P(y_i, G_{i1},\ldots,G_{iK}) = \prod_{k=1}^{K} (f_{ik}\pi_k)^{G_{ik}W_k(t_i,t_T)}.$$

Then, the corresponding local likelihood function, which is called complete local likelihood, is

$$L^*(\theta|y_1,\ldots,y_T) = \prod_{i=1}^{T}\prod_{k=1}^{K} (f_{ik}\pi_k)^{G_{ik}W_k(t_i,t_T)}.$$

The log local likelihood function is

$$\ell^* = \log L^* = \sum_{i=1}^{T}\sum_{k=1}^{K} G_{ik}W_k(t_i,t_T)\log\pi_k + G_{ik}W_k(t_i,t_T)\log f_{ik}.$$

Interpreting the π_k as 'prior' probability of class membership, then posterior probabilities of class membership are produced via Bayes theorem, that is

$$r_{ik} = P(G_{ik} = 1) = \frac{\pi_k f_k(y_i)}{\sum_{\ell=1}^{K} \pi_\ell f_\ell(y_i)}. \tag{2}$$

Equation (2) is identical to the E-step of the EM-algorithm. The posterior probabilities r_{ik} using the current estimates of π_k, β_k and σ are then given as

$$r_{ik} = \frac{\pi_k \exp\left(-\frac{1}{2}(\frac{y_i-\beta_k}{\sigma})^2\right)}{\sum_{\ell=1}^{K} \pi_\ell \exp\left(-\frac{1}{2}(\frac{y_i-\beta_\ell}{\sigma})^2\right)}.$$

In the M-step, for the π_k, one needs to apply a Lagrange multiplier since $\sum_{k=1}^{K} \pi_k = 1$ by setting

$$\partial\left(\ell^* - \lambda(\sum_{k=1}^{K} \pi_k - 1)\right)/\partial\pi_k = 0, \quad k = 1,\ldots,K$$

and one obtains

$$\hat{\pi}_k = \frac{\sum_{i=1}^{T} r_{ik} W_k(t_i, t_T)}{\sum_{i=1}^{T}\sum_{k=1}^{K} r_{ik} W_k(t_i, t_T)}.$$

In addition, by setting $\partial\ell^*/\partial\beta_k = 0$ and $\partial\ell^*/\partial\sigma = 0$, the estimates are

$$\hat{\beta}_k = \frac{\sum_{i=1}^{T} r_{ik} W_k(t_i, t_T) y_i}{\sum_{i=1}^{T} r_{ik} W_k(t_i, t_T)},$$

$$\hat{\sigma}^2 = \frac{\sum_{i=1}^{T}\sum_{k=1}^{K} r_{ik} W_k(t_i, t_T)(y_i - \beta_k)^2}{\sum_{i=1}^{T}\sum_{k=1}^{K} r_{ik} W_k(t_i, t_T)}.$$

Forecasting using MLC

The m-step-ahead forecast equation is obtained by solving the minimisation problem which is

$$\hat{y}_{T+m} = \min_a \sum_{i=1}^{T}\sum_{k=1}^{K} r_{ik} W_k(t_i, t_{T+m})(y_i - a)^2.$$

Then, we have the following m-step-ahead forecast equation

$$\hat{y}_{T+m} = \frac{\sum_{i=1}^{T}\sum_{k=1}^{K} r_{ik} W_k(t_i, t_{T+m}) y_i}{\sum_{i=1}^{T}\sum_{k=1}^{K} r_{ik} W_k(t_i, t_{T+m})}.$$

3.2 Mixture models using local linear kernel estimators (MLL)

We have generalized the MLC model by using local linear kernel smoothing rather than local constant smoothing to carry out the localized estimation step. The kth regression function at time point t_T can be approximated as $m_k(t_i) \approx m_k(t_T) + m_k'(t_T)(t_i - t_T)$, motivating the localized model

$$y_i = \begin{cases} \beta_{01}(t_T) + \beta_{11}(t_T)(t_i - t_T) + \varepsilon_{i1}, & \text{with probability } \pi_1(t_T); \\ \vdots \\ \beta_{0K}(t_T) + \beta_{1K}(t_T)(t_i - t_T) + \varepsilon_{iK}, & \text{with probability } \pi_K(t_T); \end{cases}$$

where β_{0k} and β_{1k} are fixed unknown coefficients which depend implicitly on a fixed time t_T, and the errors $\varepsilon_{ik} \sim N(0, \sigma^2)$ are independently distributed. For given t_T, the data are weighted by exponential kernels for each component which is defined in (1). In the estimation step, the EM-algorithm is used to estimate the parameters π_k, β_{0k}, β_{1k} and σ for each component k. The posterior probabilities are found in the E-step as follows

$$r_{ik} = \frac{\pi_k \exp\left(-\frac{1}{2}\left(\frac{y_i - \beta_{0k} - \beta_{1k}(t_i - t_T)}{\sigma}\right)^2\right)}{\sum_{\ell=1}^{K} \pi_\ell \exp\left(-\frac{1}{2}\left(\frac{y_i - \beta_{0\ell} - \beta_{1\ell}(t_i - t_T)}{\sigma}\right)^2\right)}.$$

In the M-step, the estimators of π_k, β_{0k} ,β_{1k} and σ are

$$\hat{\pi}_k = \frac{\sum_{i=1}^{T} r_{ik} W_k(t_i - t_T)}{\sum_{i=1}^{T} \sum_{k=1}^{K} r_{ik} W_k(t_i - t_T)},$$

$$\hat{\beta}_{0k} = \frac{S_{k,T,2} S^*_{k,T,0} - S_{k,T,1} S^*_{k,T,1}}{S_{k,T,2} S_{k,T,0} - S^2_{k,T,1}}, \qquad \hat{\beta}_{1k} = \frac{S_{k,T,0} S^*_{k,T,1} - S_{k,T,1} S^*_{k,T,0}}{S_{k,T,2} S_{k,T,0} - S^2_{k,T,1}},$$

where $S_{k,T,j} = \sum_{i=1}^{T} W_k(t_i - t_T) r_{ik} (t_i - t_T)^j$ and $S^*_{k,T,j} = \sum_{i=1}^{T} W_k(t_i - t_T) r_{ik} y_i (t_i - t_T)^j$, and

$$\hat{\sigma}^2 = \frac{\sum_{i=1}^{T} \sum_{k=1}^{K} r_{ik} W_k(t_i - t_T)(y_i - \hat{\beta}_{0k} - \hat{\beta}_{1k}(t_i - t_T))^2}{\sum_{i=1}^{T} \sum_{k=1}^{K} r_{ik} W_k(t_i - t_T)}.$$

Forecasting using MLL
The m-step-ahead forecast equation is obtained by the fitted mixture as follows

$$\hat{y}_{T+m} = \sum_{k=1}^{K} \hat{\pi}_k \left[\hat{\beta}_{0k}(t_T) + \hat{\beta}_{1k}(t_T)(t_{T+m} - t_T) \right].$$

4 Examples

In this section real data examples are presented to investigate the performance of the MLC and MLL models in forecasting compared to other time series models such as Holt's exponential smoothing and ARIMA models. The data discussed in these examples come from the International Energy Agency (IEA) and represent the annual energy use (in kg oil equivalent per capita) between 1971 and 2011. Due to the nature of the data, which are restricted to the positive range and feature several countries with extremely large energy use, a log-transformation will be applied in all further analyses. While the full data set contains more than 130 countries, we choose three countries with representative patterns for this presentation. Figure 4 displays the time series of log energy use of Bolivia, Lebanon and Greece. It can be seen that the time series of Bolivia (left) has two features (which are shared by the large majority of countries in this data base): it shows an overall increasing linear trend, but still considerable variability. The other two time series illustrate extreme cases where one of these features is more pronounced: in the case of Lebanon (middle) we have very strong variability, and in the case of Greece (right) we have a very consistent linear trend with little variability.

The log energy use data of these countries are fitted at target points $t_T = 1990, \ldots, 2007$, in order to obtain m-step ahead forecasts ($m = 1, \ldots, 4$) for each time point t_T by different models. Hence, we have 18 forecasts for each model and forward lag. For the MLC model and the MLL model, $K = 2$ components are used to fit the data, where two different settings of bandwidths, $(h_1, h_2) = (1, 5)$ and $(h_1, h_2) = (1, 20)$, are considered, in order to capture different short- and long-term trends prevailing in these data sets.

To assess the performance of the forecasts using these models, we consider the sum of square relative error (SSRE) of forecasts and the sum of absolute relative error (SARE) of m-step ahead forecasts which are defined as

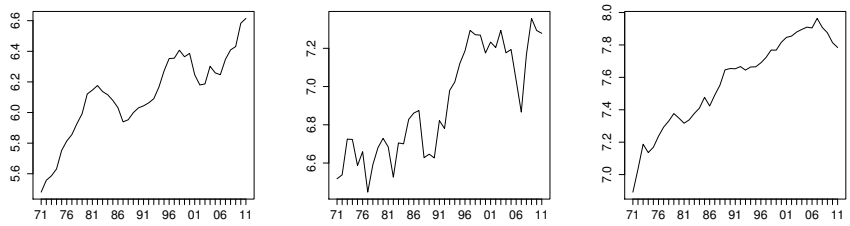

Fig. 1 Time series of Bolivia, Lebanon and Greece (from left to right). The horizontal axis denotes the calendar year (from 1971 to 2011), and the vertical axis gives the annual energy use (natural log of kg oil equivalent per capita).

$$SSRE(m) = \frac{\sum_{T=a}^{b} (\hat{y}_{T+m} - y_{T+m})^2}{\sum_{T=a}^{T=b} y_{T+m}^2}; \qquad SARE(m) = \frac{\sum_{T=a}^{b} |\hat{y}_{T+m} - y_{T+m}|}{\sum_{T=a}^{b} |y_{T+m}|};$$

where a is the first time point and b is the last time point, which for our analysis take the values $a = 1990$ and $b = 2007$, respectively.

Tables 1, 2 and 3 summarize the results[2] of m-step ahead forecasting ($m = 1,\ldots,4$) according to the SSRE and SARE criteria. From Table 1, we see that the MLC model has performed well for all forward lags, and has produced smaller errors than all other methods, except in the case when $h_2 = 20$ and $m = 1$. In this case MLL has shown a better performance than MLC, due to its ability to model the long-term linear trend. Further insight is provided by Fig. 4, which shows the time series of Bolivia, as well as the fitted parameters and predictions (top and bottom left), and the fitted mixture probabilities (top and bottom right) for $t_T, T = 1990,\ldots,2007$ for one-step ahead prediction from the MLC model. One can observe that the long-term component seems to become close to irrelevant for the MLC from around $t_T = 2002$ on, an effect which is not observed for the MLL (Fig. 4 right). In most cases, the proportion of the short-term component settles at about 80% which is plausible since recent information is considered more relevant. The additional information provided by the long-term component in the MLL model is useful for short-term prediction, but this advantage vanishes for $m > 1$ due to the increased variance.

For the Lebanon data, the errors in Table 2 are overall of larger magnitude than for Bolivia, due to the larger variability of the data themselves, but oth-

[2] All values in tables are to be divided by 1000.

Table 1 The SSRE and SARE of forecasting for Bolivia from 1991 to 2008.

Model	$SSRE(1)$	$SSRE(2)$	$SSRE(3)$	$SSRE(4)$	$SARE(1)$	$SARE(2)$	$SARE(3)$	$SARE(4)$
$(h_1,h_2)=(1,5)$								
MLC	0.09	0.08	0.09	0.09	7.11	7.34	7.78	7.87
MLL	0.12	0.24	0.48	0.87	8.64	11.56	18.38	25.49
$(h_1,h_2)=(1,20)$								
MLC	0.15	0.17	0.20	0.22	10.07	10.97	12.19	13.13
MLL	0.14	0.31	0.59	1.06	9.50	14.84	21.04	27.09
Holt	0.14	0.44	0.90	1.50	8.85	17.60	26.33	34.20
ARIMA	0.12	0.33	0.59	0.85	8.74	14.62	21.38	25.86

Table 2 The SSRE and SARE of forecasting for Lebanon from 1991 to 2008.

Model	$SSRE(1)$	$SSRE(2)$	$SSRE(3)$	$SSRE(4)$	$SARE(1)$	$SARE(2)$	$SARE(3)$	$SARE(4)$
$(h_1,h_2)=(1,5)$								
MLC	0.16	0.18	0.18	0.17	10.95	11.51	11.48	10.75
MLL	0.24	0.70	1.21	1.67	11.70	19.02	27.07	31.55
$(h_1,h_2)=(1,20)$								
MLC	0.30	0.32	0.35	0.35	15.95	16.62	17.13	16.83
MLL	0.24	0.60	1.05	1.40	12.24	17.76	24.38	29.05
Holt	0.34	0.69	0.95	1.07	15.26	21.25	26.88	28.19
ARIMA	0.31	0.71	1.05	1.26	14.05	20.94	26.92	27.66

Table 3 The SSRE and SARE of forecasting for Greece from 1991 to 2008.

Model	$SSRE(1)$	$SSRE(2)$	$SSRE(3)$	$SSRE(4)$	$SARE(1)$	$SARE(2)$	$SARE(3)$	$SARE(4)$
$(h_1,h_2)=(1,5)$								
MLC	0.03	0.02	0.02	0.02	5.11	4.43	3.92	3.64
MLL	0.01	0.03	0.07	0.14	2.71	3.53	5.68	8.32
$(h_1,h_2)=(1,20)$								
MLC	0.12	0.11	0.11	0.10	10.71	10.42	9.98	9.67
MLL	0.02	0.04	0.10	0.20	2.82	4.13	6.29	9.04
Holt	0.02	0.04	0.09	0.17	3.14	4.44	7.11	10.20
ARIMA	0.02	0.05	0.12	0.22	3.27	5.42	8.20	11.28

erwise the picture obtained previously is confirmed: MLC leads generally to favorable results, with the MLL becoming competitive only for $m = 1$ and a large long-term bandwidth. Holt and ARIMA can compete with the MLC model only for $m = 1$.

For the data from Greece the situation is different due to the nature of this time series which shows an increase which is close to linear. Here the ability to model a local linear trend can play a strong role at enhancing the prediction, and due to the stability of this trend, this continues to hold for forecast lags $m > 1$.

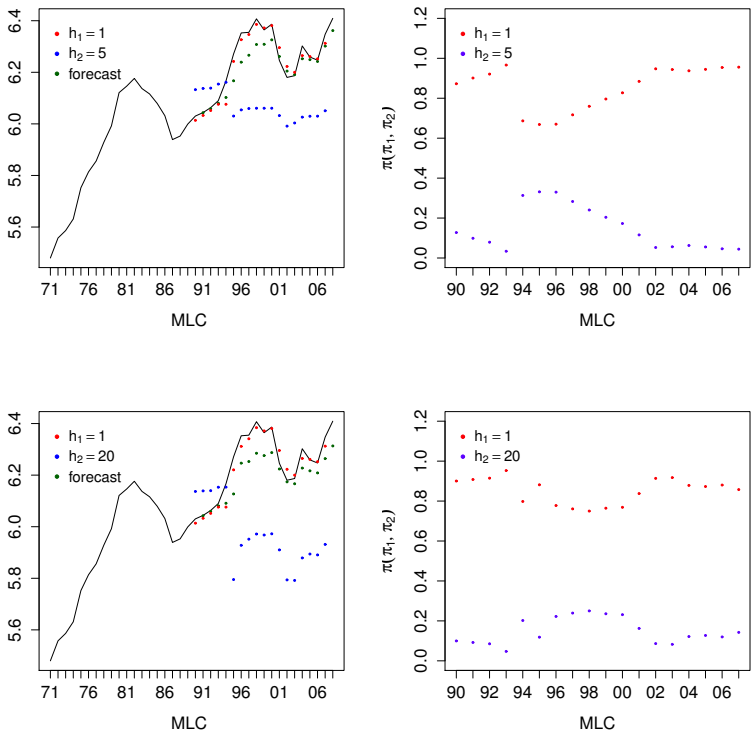

Fig. 2 For data from Bolivia, parameters $\hat{\beta}_k(t_T)$ fitted using MLC and resulting forecasts at $\hat{y}_{T+1}$ (left); and fitted parameters $\hat{\pi}_k(t_T)$ (right).

Summarizing, the examples have given some evidence for the superiority of the MLC method, especially for higher lags and smaller bandwidths. Remarkably, the performance of the MLC method almost does not depend on the forward lag. Here an apparent 'weakness' of the MLC method — namely the non-adaptability to linear trends — seems to turn into an advantage, as the technique does not 'learn' the direction of these local trends, and so avoids overshooting once the data take a turn. For the MLC method, the bandwidth choice $h_2 = 5$ produced generally better results than $h_2 = 20$. For the MLL this interpretation is less clear-cut, but it is right to say that our results, using MLL with $h_2 = 20$, where generally comparable to those obtained using the ARIMA and Holt methods. It appears that the MLL method is only recommendable when $m = 1$ and h_2 is large.

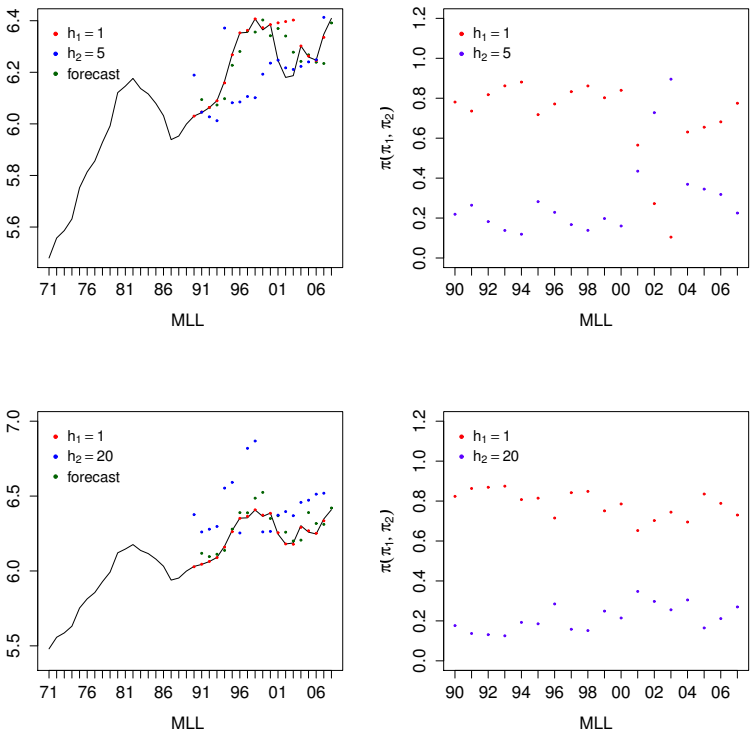

Fig. 3 For the time series from Bolivia, parameters $\hat{\beta}_{0k}(t_T)$ fitted using MLL and resulting forecasts at $\hat{y}_{T+1}$ (left); and fitted parameters $\hat{\pi}_k(t_T)$ (right).

5 Conclusion

This paper presents a novel approach to forecasting based on localized mixtures of nonparametric regressions. Nonparametric regression allows a forecast to be calculated directly from historical data as a local average of observed past values. In the first model which is named MLC, local constant estimators are used to carry out the localized estimation step. In the second model which is referred to as MLL, the MLC is generalized using local linear estimators. Estimation of these models is achieved through a kernel-weighted version of the EM-algorithm, using exponential kernels with different bandwidths as weight functions. In order to forecast, several approaches for prediction at time t_{T+m},

$m = 1, \ldots, 4$ from these models were investigated. The results suggest that mixture models can improve predictions from time series data compared to Holt's exponential smoothing and ARIMA models, though further forecasting methods should be investigated for this comparison. Currently, further consideration is given to optimal bandwidth choice for forecasting, and a simulation study to assess the accuracy of forecasting using MLC and MLL.

Acknowledgements We thank Prof. Guy Nason for his suggestion on forecasting based on fitted mixture models which led to improved insights and conclusions from our results. The first author of this manuscript is grateful to the Saudi Arabian Cultural Bureau in London for the financial support of this work.

References

Böhning D, Hennig C, McLachlan GJ, McNicholas PD (2014) The 2nd special issue on advances in mixture models. Computational Statistics & Data Analysis 71(C):1–2, DOI 10.1016/j.csda.2013.10.010

Fan J, Gijbels I (1996) Local polynomial modelling and its applications. CRC Press, Boca Raton

Frühwirth-Schnatter S (2001) Markov chain monte carlo estimation of classical and dynamic switching and mixture models. Journal of the American Statistical Association 96(453):194–209, URL `http://www.jstor.org/stable/2670359`

Frühwirth-Schnatter S (2006) Finite Mixture and Markov Switching Models. Springer, Berlin

Green PJ, Richardson S (2002) Hidden markov models and disease mapping. Journal of the American Statistical Association 97(460):1055–1070, DOI 10.1198/016214502388618870

Hinde J, Ingrassia S, Lin TI, McNicholas P (2016) The third special issue on advances in mixture models. Computational Statistics & Data Analysis 93(C):2–4, DOI 10.1016/j.csda.2015.08.014

Huang M, Yao W (2012) Mixture of regression models with varying mixing proportions: a semiparametric approach. Journal of the American Statistical Association 107(498):711–724, DOI 10.1080/01621459.2012.682541

Huang M, Li R, Wang S (2013) Nonparametric mixture of regression models. Journal of the American Statistical Association 108(503):929–941, DOI 10.1080/01621459.2013.772897

McLachlan G, Peel D (2004) Finite mixture models. Third ed. John Wiley & Sons, New York

Young DS, Hunter D (2010) Mixtures of regressions with predictor-dependent mixing proportions. Computational Statistics & Data Analysis 54(10):2253 – 2266, DOI 10.1016/j.csda.2010.04.002