Mixtures of Nonparametric Autoregressions

June 13, 2009

J. Franke, J. Stockis, J. Tadjuidje, W. K. Li,

Abstract

We consider data generating mechanisms which can be represented as mixtures of finitely many regression or autoregression models. We propose nonparametric estimators for the functions characterizing the various mixture components based on a local quasi maximum likelihood approach and prove their consistency. We present an EM algorithm for calculating the estimates numerically which is mainly based on iteratively applying common local smoothers and discuss its convergence properties.

AMS 2000 subject classification: Primary: 62G08; Secondary: 62M10
Key words and phrases: nonparametric regression, nonparametric autoregression, mixture, hidden variables, EM algorithm, kernel estimates, local likelihood

Corresponding author:
J. Franke, Dept. of Mathematics, University of Kaiserslautern,
D-67653 Kaiserslautern, Germany
Tel. +49-631-205-2741, Fax +49-631-205-3052, email franke@mathematik.uni-kl.de

Acknowledgement: The work was supported by the Deutsche Forschungsgemeinschaft (DFG) as well as by the *Computational Mathematical and Computational Modelling* $\mathbf{(CM)}^2$ funded by the state of Rhineland-Palatinate.

1 Introduction

We consider regressions and autoregressions which may be represented as a mixture of M different nonlinear models. The available data are of the form $(X_1, Y_1), \ldots, (X_N, Y_N)$, and we assume that they are part of a strictly stationary time series. For sake of simplicity, we restrict our considerations to one-dimensional variables $X_1, \ldots, X_N \in \mathbb{R}$; the generalization to higher-dimensional situations is straightforward. We assume that the data are generated by the following independent switching model

$$Y_t = \sum_{k=1}^{M} Z_{t,k}\{m_k(X_t) + \sigma\varepsilon_{t,k}\} \tag{1}$$

where the residuals $\varepsilon_{t,k}, t = 1, \ldots, N, k = 1, \ldots, M$, are i.i.d. random variables with mean 0 and variance 1, $m_1(x), \ldots, m_M(x)$ are the unknown regression functions of M regression models, and $\sigma^2 > 0$ is the residual variance. $Z_t = (Z_{t1}, \ldots, Z_{tM})^T$ are i.i.d. random variables which assume as values the unit vectors $e_1, \ldots, e_M \in \mathbb{R}^M$, i.e. exactly one of the Z_{tk} is 1, and the others are 0. Furthermore, we assume that Z_t is independent of $X_j, \varepsilon_{j,k}, j \leq t$. Let

$$\pi_k = \mathrm{pr}(Z_t = e_k) = \mathrm{pr}(Z_{tl} = 0 \text{ for } l \neq k), \qquad k = 1, \ldots, M,$$

be the probability that Y_t is generated from X_t using the k-th regression model, where $\pi_1 + \ldots + \pi_M = 1$. If, e.g., the $\varepsilon_{t,k}$ are standard normal variables with Φ denoting their distribution function, the conditional distribution function of Y_t given $X_t = x$ is

$$F(y|x) = \mathrm{pr}(Y_t \leq y | X_t = x) = \sum_{k=1}^{M} \pi_k \Phi\left(\frac{y - m_k(x)}{\sigma}\right), \tag{2}$$

and the conditional expectation of Y_t given $X_t = x$ is

$$E(Y_t|X_t = x) = \sum_{k=1}^{M} \pi_k \, m_k(x).$$

In particular, we allow for $X_t = Y_{t-1}$. In that case, we get a mixture of M nonparametric autoregressive processes of order 1:

$$Y_t = \sum_{k=1}^{M} Z_{t,k}\{m_k(Y_{t-1}) + \sigma\varepsilon_{t,k}\}. \tag{3}$$

In the special case, where the autoregression functions are all linear, i.e. $m_k(x) = \phi_{k0} + \phi_{k1}x, \quad k = 1, \ldots, M$, we get a mixture autoregressive model as considered by Wong and Li [12]. Conditions on $\pi_1, \ldots, \pi_M, m_1, \ldots, m_M$ for the existence of a stationarity solution of (3) have been given in a much more general context in [11]. Here, we only remark that some of the autoregressive dynamics characterized by $m_k(x)$ may be

explosive provided that they occur rarely enough, i.e. π_k is small enough.

The assumption of independent state variables Z_t is, of course, a considerable simplification, but the purpose of this paper is to present the main idea of combining nonparametrics, in particular local smoothers, and mixture models in a simple framework. We also present a real data set where the restricted model serves as a good approximation of the data generating process. In principle, however, nonparametric Markov switching models where the Z_t form a Markov chain with finite state space corresponding to the M different phases would be much more flexible and widely applicable. This will be a topic for consecutive research. Due to the same reason, we restrict ourselves to autoregressions of order 1 though the basic idea of estimating functions in a mixture of models can be transferred to, e.g., higher order autoregressions or ARCH-processes, compare Wong and Li [13] for the parametric case or Stockis et al. [11] for the general case.

In the next section, we present a local quasi maximum likelihood approach to deriving simultaneous estimates of the regression functions $m_1, \ldots, m_M$. Section 3 discusses an EM algorithm as an iterative numerical scheme for calculating those estimates which boils down to using common kernel estimates in the M-step. Section 4 illustrates the feasibility of this estimation procedure by applying it to some artificial an real data. Finally, in the technical appendix, we prove consistency of the estimates and have a look at the convergence properties of the EM algorithm.

2 Local likelihood estimates

In this paper, we do not restrict the functions m_k to particular parametric classes, but we assume only a certain degree of smoothness. Our goal is to derive simultaneous estimates for the parameters $\pi_1, \ldots, \pi_{M-1}, \sigma$ as well as for the regression functions $m_1(x), \ldots, m_M(x)$. Mark that π_M is only used as an abbreviation for $1 - \pi_1 - \ldots - \pi_{M-1}$ throughout the paper. For the homogeneous models, i.e. for $M = 1$, kernel estimates and, more generally, local polynomial estimates have been applied successfully to estimating regression and autoregression functions nonparametrically ([3], [4], [5], [7], [10]). We combine those ideas of local averaging with the approach of Wong and Li for getting estimates for parametric mixture models. If the data are generated by only one regression function ($M = 1$), a common nonparametric estimate for the function $m_1(x)$ is the Nadaraya-Watson kernel estimate

$$\hat{m}_1(x, h) = \frac{\sum_{t=1}^{N} K_h(x - X_t)\, Y_t}{\sum_{t=1}^{N} K_h(x - X_t)} \tag{4}$$

for some suitable bandwidth h. $K(u)$ is a kernel function satisfying

(K) $\quad K(u) \geq 0,\ K(-u) = K(u),\ \int K(u)du = 1$, and the support of K is compact.

These conditions could be relaxed, but again we prefer to keep this exposition as simple

as possible. $K_h(u) = \frac{1}{h}K(\frac{u}{h})$ denotes the rescaled kernel. $\hat{m}_1(x,h)$ can be interpreted as solution of a local weighted least-squares problem

$$\hat{m}_1(x,h) = \arg\min_{\mu\in\mathbb{R}} \sum_{t=1}^{N} K_h(x - X_t)(Y_t - \mu)^2$$

where the weights are specified by the kernel such that observations with $X_t \approx x$ have the largest influence on the estimate of the function at x. If the residuals $\varepsilon_{t,k}$ are normal random variables, then, equivalently, $\hat{m}_1(x,h)$ is also a local maximum likelihood estimate as, with $\varphi(u)$ denoting the standard normal density, it maximizes the local conditional log likelihood function

$$\sum_{t=1}^{N} K_h(x - X_t) \log \frac{1}{\sigma}\ \varphi\big(\frac{Y_t - \mu}{\sigma}\big)$$

with respect to μ for any $\sigma > 0$.

For the general case, we consider the corresponding Gaussian local conditional log likelihood

$$L(\vartheta|X,Y) = \sum_{t=1}^{N} K_h(x - X_t) \log \sum_{k=1}^{M} \frac{\pi_k}{\sigma}\ \varphi\big(\frac{Y_t - \mu_k}{\sigma}\big) \tag{5}$$

$\vartheta = (\pi_1, \ldots, \pi_{M-1}, \mu_1, \ldots, \mu_M, \sigma)^T \in \Theta$ denotes the partly local parameter where $\Theta \subseteq \mathbb{R}^{2K}$ is the set of admissible parameters satisfying $0 \le \pi_k,\ \ k = 1, \ldots, M-1, \quad \pi_1 + \ldots + \pi_{M-1} \le 1$ and $\sigma > 0$. We do not assume that the residuals $\varepsilon_{t,k}$ are normally distributed. Therefore, maximizing $L(\vartheta|X,Y)$ with respect to ϑ provides a local quasi maximum likelihood estimate $\widehat{\vartheta}_N$. In the appendix, we discuss conditions for the consistency of the estimate $\widehat{\vartheta}_N$ for $N \to \infty, h \to 0$.

3 The EM algorithm

Observing a mixture of nonparametric regressions or autoregressions like (1), we could treat it as M independent estimation problems if the Z_{tk} would be observable. By our assumptions, we would have M independent data sets

$$Y_t = m_k(X_t) + \sigma\varepsilon_{t,k}, \quad t \in T_k = \{n \le N;\ \ Z_{nk} = 1\},$$

$k = 1, \ldots, M$. The Nadaraya-Watson estimates for the functions m_k would be

$$\hat{m}_k^0(x,h) = \frac{\sum_{t\in T_k} K_h(x - X_t)\ Y_t}{\sum_{t\in T_k} K_h(x - x_t)} = \frac{\sum_{t=1}^{N} K_h(x - X_t)\ Y_t Z_{tk}}{\sum_{t=1}^{N} K_h(x - X_t) Z_{tk}}$$

as the Z_{tk} are either 1 or 0. The vector of function estimates $(\hat{m}_1^0(x,h_1), \ldots, \hat{m}_M^0(x,h_M))^T$ solves the weighted least-squares problem

$$\sum_{t=1}^{N}\sum_{k=1}^{M} (Y_t - \mu_k)^2 Z_{tk} K_h(x - X_t) = \min_{\mu_1,\ldots,\mu_M\in\mathbb{R}} !$$

As we do not observe the Z_{tk}, we follow the approach of Wong and Li (2000) instead, and approximate the hidden variables by their conditional expectations ζ_{tk}^0 given Y_t under the assumptions that the residuals $\varepsilon_{t,k}$ are standard normal variables. Let $\varphi(u)$ denote the standard normal density. If $Z_{tk} = 1$, then, conditional on $X_t = x$, the distribution of Y_t is $\mathcal{N}(m_k(x), \sigma^2)$. Therefore,

$$\begin{aligned}\zeta_{tk}^0 &= E\{Z_{tk}|Y_t = y\} = \mathrm{pr}\{Z_{tk} = 1|Y_t = y\} \\ &= \frac{\pi_k \frac{1}{\sigma}\varphi(\frac{y-m_k(X_t)}{\sigma})}{\sum_{l=1}^M \pi_l \frac{1}{\sigma}\varphi(\frac{y-m_l(X_t)}{\sigma})} \ .\end{aligned}$$

As we do not know the parameters π_k, σ and the regression functions $m_k(x)$, we apply the same kind of iterative EM-procedure as in Wong and Li (2001).

(a) **E-step:** Suppose that estimates $\hat{\pi}_1, \ldots, \hat{\pi}_M$, $\hat{\sigma}$ and approximations e_{tk} of the residuals $Y_t - m_k(X_t)$ are given. Then, the conditional expectations of the hidden variables Z_{tk} given Y_t are estimated by

$$\zeta_{tk} = \frac{\hat{\pi}_k \frac{1}{\hat{\sigma}}\varphi(\frac{e_{tk}}{\hat{\sigma}})}{\sum_{l=1}^M \hat{\pi}_l \frac{1}{\hat{\sigma}}\varphi(\frac{e_{tl}}{\hat{\sigma}})}, \ k = 1, \ldots, M, \ t = 1, \ldots, N.$$

(b) **M-step:** Suppose approximations ζ_{tk} for the hidden variables Z_{tk} are given. Then, we estimate the probabilities $\pi_1, \ldots, \pi_M$ by

$$\hat{\pi}_k = \frac{1}{N}\sum_{t=1}^N \zeta_{tk}, \ \ k = 1, \ldots, M.$$

We estimate the M regression functions by

$$\hat{m}_k(x, h_k) = \frac{\sum_{t=1}^N K_{h_k}(x - X_t)\ Y_t\zeta_{tk}}{\sum_{t=1}^N K_{h_k}(x - X_t)\ \zeta_{tk}}, \quad k = 1, \ldots, M,$$

and the residual variances by

$$\hat{\sigma}^2 = \frac{1}{N}\sum_{t=1}^N \sum_{k=1}^M e_{tk}^2\zeta_{tk},$$

where $e_{tk} = Y_t - \hat{m}_k(X_t, h_k)$ denote the sample residuals.

The estimates of the parameters and the regression functions are obtained by iterating these two steps until convergence.

Remark 1 *The final values of $\zeta_{tk}, k = 1, \ldots, M$, may be used for classifying the observations by the following common rule: Y_t is classified as belonging to state k iff $\zeta_{tk} = \max_{i=1,\ldots,M} \zeta_{ti}$.*

Remark 2 *$\hat{\pi}_k$ and $\hat{\sigma}^2$ are different from the natural estimates obained in the Appendix. However, they are asymptotically equivalent given the Z_{tk} are known.*

The EM-algorithm is a computationally simple numerical procedure for maximizing the Gaussian local conditional log-likelihood $L(\vartheta|X,Y)$ of (5). Under typical conditions, we prove in the appendix that it converges to a stationary point ϑ_0 of $L(\vartheta|X,Y)$. In practice, we may get different limit points corresponding to different local maxima of $L(\vartheta|X,Y)$ if we choose different initial values, but that is not unusual for maximum likelihood type procedures in situations with many parameters. Therefore, we recommend to apply the usual device of trying several starting values and compare the values of the target function $L(\vartheta|X,Y)$ for the various limits of the numerical procedure.

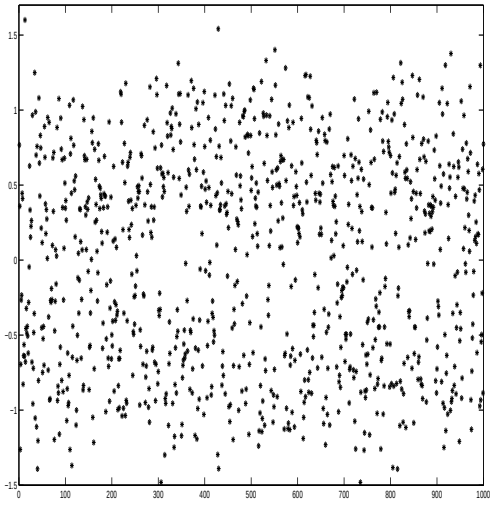

Figure 1: Simulated Data

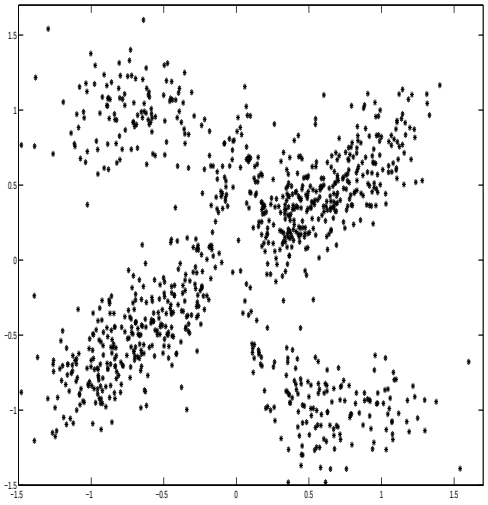

Figure 2: Scatter Plot Simulated Data

4 Numerical examples

4.1 A simulation

To illustrate the feasibility of the estimation procedure combined with the numerical procedure described above, we first consider some artificial data. We generate $N = 1000$ observations from a nonparametric AR(1)-mixture model (1), i.e. $X_t = Y_{t-1}$, with $M = 2$ components and standard normal innovations $\varepsilon_{t,k}$. We choose the state probabilities as $\pi_1 = 0.7, \pi_2 = 1-\pi_1 = 0.3$, the innovation variance as $\sigma^2 = 0.2$ and the two autoregressive functions as

$$m_1(x) = 0.7x + 2\varphi(10x), \quad m_2(x) = \frac{2}{1+e^{10x}} - 1,$$

where φ denotes the standard normal density. i.e. m_1 is a bump function and m_2 is a function of sigmoid shape. Figures 1 and 2 show the data and the corresponding scatter plot of Y_t against Y_{t-1}.

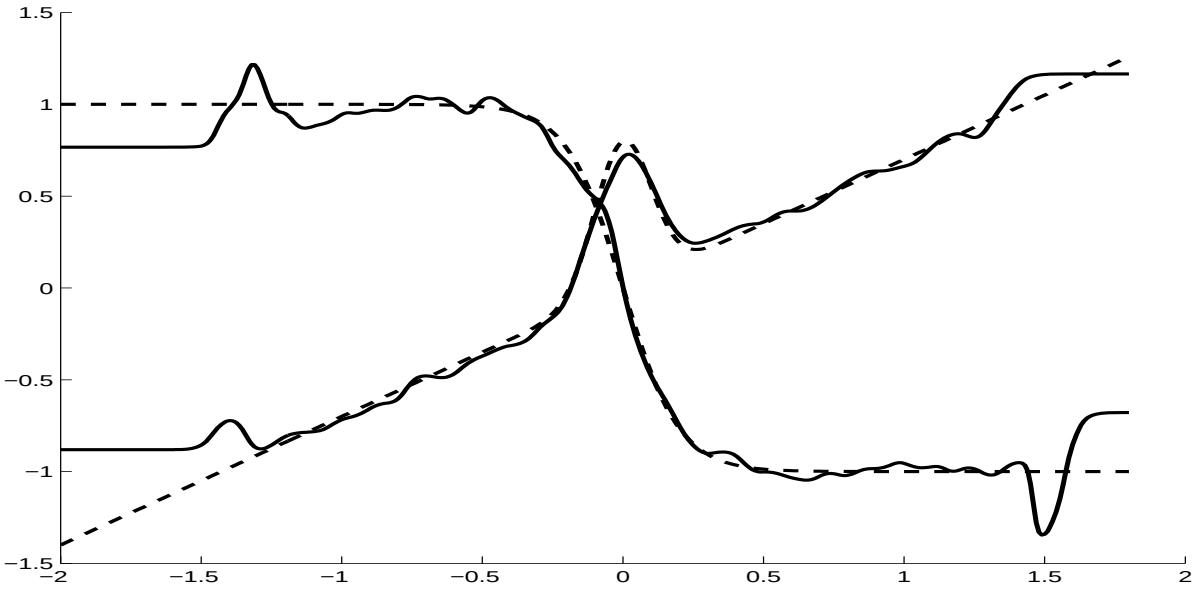

Figure 3: Estimated Trend Functions

We apply the EM-algorithm with bandwidth h chosen by an *opening the window* technique, i.e. by trying several bandwidths and deciding visually for a good compromise which is neither too smooth nor too rough. Of course, an automatic procedure would be desirable and will be the topic of future research. The estimation procedure yields for the parameters $\hat{\pi}_1 = 0.6990$ and $\hat{\sigma}^2 = 0.2004$. For the final bandwidths of the two kernel estimates, we get $\hat{h}_1 = 0.0465, \hat{h}_2 = 0.0393$. Figure 3 shows m_1, m_2 (dashed lines) and the respective kernel estimates (solid lines). Apart from some deviations at the boundaries which may be explained by scarceness of data in that region and by boundary effects, the quality of the estimates is rather good. Figure 4 shows the final values of $\max(\zeta_{t1}, \zeta_{t2}$ which, except for very few cases, are close to 1. The classification rule of Remark 1, therefore, mostly leads to a clear-cut decision.

4.2 An application to heart rate data

As a second example, we consider a set of data from a person suffering from a severe dysfunction of the rhythm of the heart. Y_t corresponds to the waiting time between two consecutive heart beats which is derived from the time lags between peaks in an electrocardiogram. Figures 5 shows the data where the sample size is $N = 2813$. Looking at the high degree of irregularity in the data, the assumption of independent state variables controlling the switching between phases seems to be plausible. Figure 6 shows the corresponding scatter plot. For a healthy person, the latter would show more or less

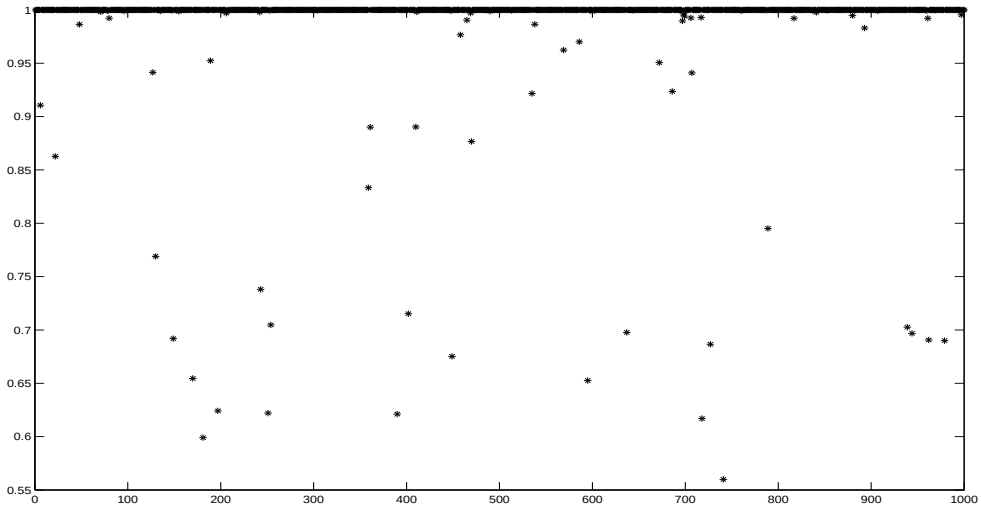

Figure 4: Maximum of the Estimated State Probabilities: Simulated Data

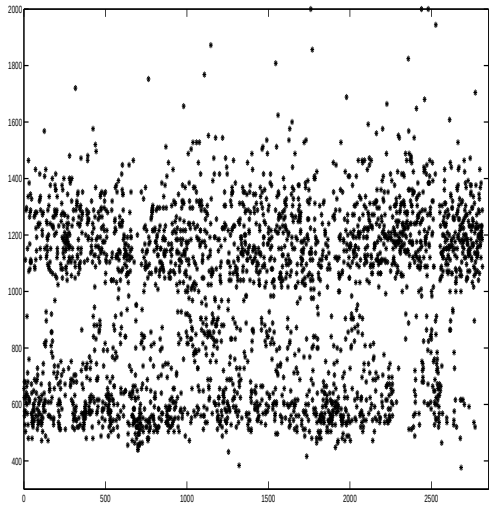

Figure 5: Heart Rate Data

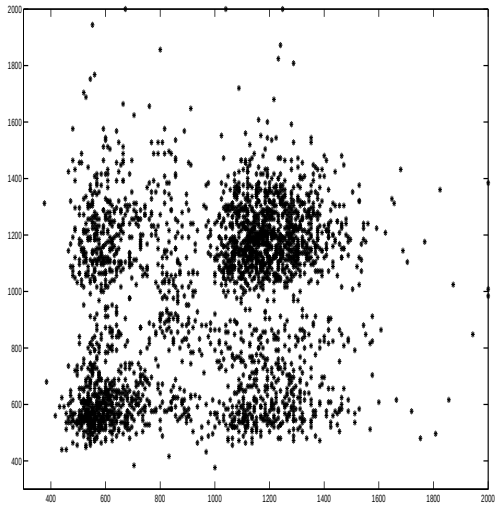

Figure 6: Scatter Plot

an ellipse with positive inclination due to the positive correlation between adjacent heart beats. The apparent clustering in Figure 6 does not only indicate the pathological nature of that data set, but also suggests the presence of several different phases.

We have fitted a mixture of $M = 3$ nonparametric AR(1)-processes to the data resulting in an estimate $\hat{\sigma} = 127.0838$ of the standard deviation of the innovations and in kernel estimates of the autoregressive functions shown in Figure 7. The dashed lines are more

or less constant corresponding to white noise with different means around 600 and 1200. The solid line shows a sigmoid function with positive inclination. We have used the rule of Remark 1 to classify the observations. The results are also shown in Figure 7 where the observations from the scatter plot of Figure 6 are now marked with different symbols corresponding to which of the three phases they are allocated. Figure 8 shows $\max(\zeta_{t1}, \zeta_{t2}, \zeta_{t3})$ which almost always are at least 0.5 in frequently considerably larger, i.e. there is a clear decision for one of the three phases in the large majority of cases.

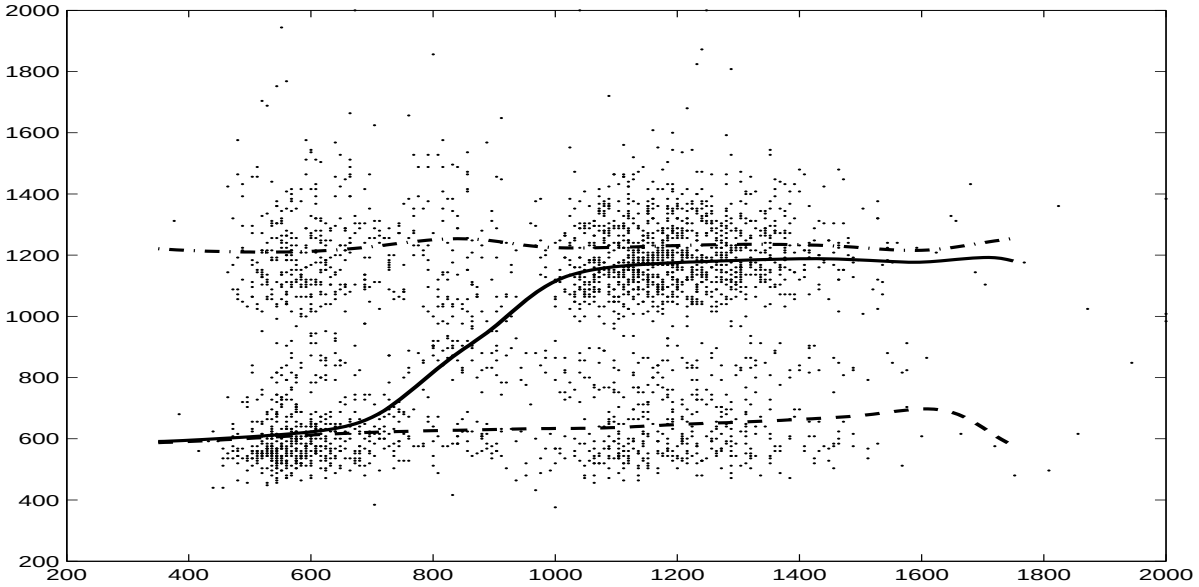

Figure 7: Scatter Plot and Functions Estimates: The upper dashed curve represents the first state trend function, the lower dashed the second state function and the third is represented by the solid curve.

We also have fitted a mixture model with 4 phases to the data which obviously did not lead to any improvement. The two upper function estimates in Figure 7 and the corresponding classification of observations remained largely unchanged. The third phase represented by the lower curve in Figure 7 was replaced by two kernel estimates which both where roughly constant and differed only slightly, i.e. they essentially estimated the same autoregressive function and represented the same data generated mechanism.

5 Appendix

5.1 Consistency of the local quasi maximum likelihood estimate

In this section, we discuss consistency of the local quasi maximum likelihood estimate $\widehat{\vartheta}_N$ of the parameters $\pi_1, \ldots, \pi_{M-1}, \sigma$ and of the regression functions $m_1(x), \ldots, m_M(x)$

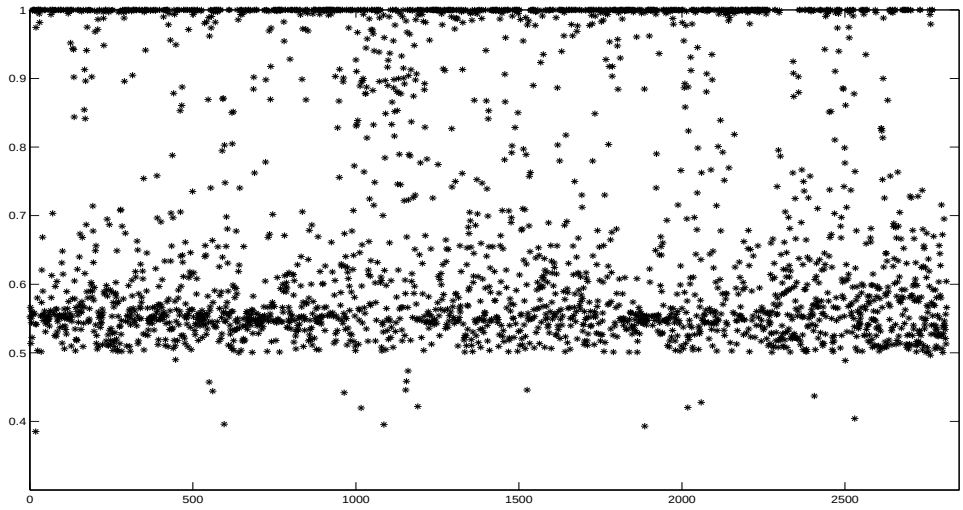

Figure 8: Maximum of the Estimated State Probabilities

at a fixed x of the mixture model (1). The local log likelihood, maximized by $\widehat{\vartheta}_N$, is of the general form

$$R_N^*(\vartheta) = \sum_{t=1}^{N} K_h(x - X_t)\rho(Y_t, \vartheta)$$

for some function $\rho : \mathbb{R} \times \Theta \to \mathbb{R}$. For convenience, we first study the general M-estimate, also called $\widehat{\vartheta}_N$, which maximizes $R_N^*(\vartheta)$ or, equivalently,

$$R_N(\vartheta) = \sum_{t=1}^{N} W_{Nt}\rho(Y_t, \vartheta) \quad \text{with } W_{Nt} = \frac{K_h(x - X_t)}{\sum_{j=1}^{N} K_h(x - X_j)}.$$

Under the assumptions, stated below, $R_N(\vartheta)$ will converge to

$$r(\vartheta) = E\{\rho(Y_1, \vartheta)|X_1 = x\}.$$

Our arguments are similar to those of Härdle and Tsybakov [6] who considered M-estimates in a location-scale regression model, but we cannot exploit the special structure of that setting. We assume

(A1) Θ is compact.

(A2) $\rho(y, \vartheta)$ is continuous in ϑ, and $E|\rho(Y_1, \vartheta)| < \infty$.

(A3) $r(\vartheta)$ is continuous and has a unique global maximum at $\vartheta_0 \in \Theta$.

(A4) $\rho_0(y, \vartheta) = \rho(y, \vartheta) - r(\vartheta)$ satisfies a uniform Lipschitz condition

$$|\rho_0(y, \vartheta) - \rho_0(y, \vartheta')| \le L(y)||\vartheta - \vartheta'||$$

for all $\vartheta, \vartheta' \in \Theta,\ y \in \mathbb{R}$ with some function $L \ge 0$ satisfying $EL(Y_1) < \infty$.

(A5) For $N \to \infty$ and $h \to 0$ such that $Nh \to \infty$,

$$\sum_{t=1}^{N} W_{Nt}\rho(Y_t,\vartheta) \xrightarrow[\text{p}]{} E\{\rho(Y_1,\vartheta)|X_1 = x\} = r(\vartheta) \quad \text{for all } \vartheta \in \Theta,$$

$$\sum_{t=1}^{N} W_{Nt}L(Y_t) \xrightarrow[\text{p}]{} E\{L(Y_1)|X_1 = x\}$$

Proposition 5.1 *Under the conditions* (**K**) *on the kernel and* **(A1)**, …,**(A5)**, *the general M-estimate* $\widehat{\vartheta}_N$ *is consistent for* ϑ_0, *i.e. for* $N \to \infty, h \to 0, Nh \to \infty$

$$\widehat{\vartheta}_N = \arg\min_{\vartheta\in\Theta} R_N(\vartheta) \xrightarrow[p]{} \vartheta_0 \quad \textit{for } N \to \infty.$$

Proof:

a) We first show uniform convergence of $R_N(\vartheta)$ to $r(\vartheta)$, i.e.

$$\sup_{\vartheta\in\Theta} |R_N(\vartheta) - r(\vartheta)| = \sup_{\vartheta\in\Theta} \sum_{t=1}^{N} W_{Nt}\rho_0(Y_t,\vartheta) \xrightarrow[\text{p}]{} 0. \tag{6}$$

As Θ is compact, we can choose $\delta > 0$, $\vartheta_1, \ldots, \vartheta_J \in \Theta$ such that Θ is covered by the δ-balls $\{\vartheta \in \Theta; ||\vartheta - \vartheta_j|| < \delta\}$, $j = 1, \ldots, J$. For arbitrary $\gamma > 0$, we have

$$\begin{aligned}
& \text{pr}\left\{ \sup_{\vartheta\in\Theta} |R_N(\vartheta) - r(\vartheta)| > \gamma \right\} \\
\leq\ & \text{pr}\left\{ \sup_{1\le j\le J} \sup_{\vartheta;||\vartheta-\vartheta_j||<\delta} \left| \sum_{t=1}^{N} W_{Nt}(\rho_0(Y_t,\vartheta) - \rho_0(Y_t,\vartheta_j)) + \sum_{t=1}^{N} W_{Nt}\rho_0(Y_t,\vartheta_j) \right| > \gamma \right\} \\
\leq\ & \text{pr}\left\{ \sup_{\vartheta,\vartheta';||\vartheta-\vartheta'||<\delta} \left| \sum_{t=1}^{N} W_{Nt}(\rho_0(Y_t,\vartheta) - \rho_0(Y_t,\vartheta')) \right| > \frac{\gamma}{2} \right\} \\
& + \text{pr}\left\{ \sup_{1\le j\le J} \left| \sum_{t=1}^{N} W_{Nt}\rho_0(Y_t,\vartheta_j) \right| > \frac{\gamma}{2} \right\} \\
\leq\ & \text{pr}\left\{ \sum_{t=1}^{N} W_{Nt}L(Y_t) - \ell > \frac{\gamma}{2\delta} - 2\ell \right\} + \sum_{j=1}^{J} \text{pr}\left\{ \left| \sum_{t=1}^{N} W_{Nt}\rho_0(Y_t,\vartheta_j) \right| > \frac{\gamma}{2} \right\}
\end{aligned}$$

where $\ell = E\{L(Y_1)|X_1 = x\}$ and where we have used that for $||\vartheta - \vartheta'|| < \delta$

$$\begin{aligned}
&\Big|\sum_{t=1}^{N} W_{Nt}\{\rho_0(Y_t,\vartheta) - \rho_0(Y_t,\vartheta')\}\Big| \\
\leq\ & \sum_{t=1}^{N} W_{Nt} L(Y_t)||\vartheta - \vartheta'|| + E\big\{L(Y_1)||\vartheta - \vartheta'||\,\big|X_1 = x\big\} \\
\leq\ & \delta\big[\sum_{t=1}^{N} W_{Nt} L(Y_t) - \ell\big] + 2\delta\ell.
\end{aligned}$$

For δ small enough, we have $\gamma/(2\delta) - 2\ell > 0$, and (7) converges to 0 by **(A5)**. (6) follows.

b) By **(A1), (A3)**, we have for arbitrary $\gamma > 0$

$$\Delta = r(\vartheta_0) - \max\{r(\vartheta);\ \vartheta \in \Theta, ||\vartheta - \vartheta_0|| \geq \gamma\} > 0.$$

As $R_N(\widehat{\vartheta}_N) \geq R_N(\vartheta_0)$, we have in the case where $||\widehat{\vartheta}_N - \vartheta_0|| \geq \gamma$

$$\begin{aligned}
\Delta &\leq r(\vartheta_0) - r(\widehat{\vartheta}_N) \\
&= r(\vartheta_0) - R_N(\vartheta_0) + R_N(\vartheta_0) - R_N(\widehat{\vartheta}_N) + R_N(\widehat{\vartheta}_N) - r(\widehat{\vartheta}_N) \\
&\leq 2\sup_{\vartheta\in\Theta} |R_N(\vartheta) - r(\vartheta)|,
\end{aligned}$$

and, therefore,

$$\mathrm{pr}(||\widehat{\vartheta}_N - \vartheta_0|| \geq \gamma) \leq \mathrm{pr}(\sup_{\vartheta\in\Theta} |R_N(\vartheta) - r(\vartheta)| \geq \frac{\Delta}{2}) \longrightarrow 0.$$

■

Conditions **(A1)**, **(A3)** are a bit restrictive, but typical for proving convergence of M-estimates in case that the criterion function has multiple local maxima in the limit. Essentially, they require to choose the set Θ of admissible parameters small enough such that it contains only one local (and then global) maximum of $r(\vartheta)$. Condition **(A5)** is nothing else but the consistency of the Nadaraya-Watson kernel estimates

$$\sum_{t=1}^{N} W_{Nt}\rho(Y_t,\vartheta) \quad \text{and} \quad \sum_{t=1}^{N} W_{Nt} L(Y_t)$$

for the conditional expectations

$$r(x,\vartheta) = E\{\rho(Y_1,\vartheta)|X_1 = x\} \quad \text{and} \quad \ell(x) = E\{L(Y_1)|X_1 = x\}$$

for arbitrary, but fixed ϑ. There are quite a number of results available guaranteeing this consistency under various sets of conditions on the functions $r(x,\vartheta), \ell(x)$, on the rate of the bandwidth h and on the dependence structure of the time series (X_t, Y_t). We only mention two of those results, the first one covering the regression case, the second one the case of α-mixing time series.

Lemma 5.1 *Let the kernel K satisfy the conditions* **(K)**, *let $(X_t, Y_t), t = 1, \ldots, N$, be independent identically distributed with*

$$E\ \rho^2(Y_1, \vartheta) < \infty,\ E\ L^2(Y_1) < \infty.$$

Let the density of X_t be continuous and positive in x, and let $r(x, \vartheta), \ell(x)$ be continuous in x. Then, for $N \to \infty, h \to 0$ such that $Nh \to \infty$

$$\sum_{t=1}^{N} W_{Nt}\rho(Y_t, \vartheta) \ \underset{p}{\longrightarrow}\ r(x, \vartheta),\ \sum_{t=1}^{N} W_{Nt}L(Y_t) \ \underset{p}{\longrightarrow}\ \ell(x).$$

The assertion follows immediately from Proposition 3.1.1 of [5].

Lemma 5.2 *Let the kernel K satisfy the conditions* (**K**), *let $(X_t, Y_t), t = 1, \ldots, N$, be strictly stationary and α-mixing with mixing coefficients α_t, satisfying for some $\delta > 0$ that $E\big\{|\rho(Y_1, \vartheta)|^{2+\delta}\big|X_1 = x'\big\}$ and $E\big\{L^{2+\delta}(Y_1)\big|X_1 = x'\big\}$ are uniformly bounded for x' in some neighbourhood of x and*

$$\sum_{t=1}^{\infty} t^{\gamma}\alpha_t^{\frac{\delta}{2+\delta}} < \infty \quad \textit{for some} \quad \gamma > \frac{\delta}{2+\delta}. \tag{7}$$

Moreover, let the joint density $f_t(u, v)$ of (X_1, X_{t+1}) as well as

$$E\big\{\rho^2(Y_1, \vartheta) + \rho^2(Y_t, \vartheta)\big|X_1 = x', X_t = x''\big\},\ \ E\big\{L^2(Y_1) + L^2(Y_t)\big|X_1 = x', X_t = x''\big\}$$

be bounded uniformly in $t \geq 1$ and in x', x'' in a neighbourhood of x, and let $r(x, \vartheta), \ell(x)$ be continuously differentiable in some neighbourhood of x. Then, for $N \to \infty, h \to 0$ such that $Nh \to \infty$, we have

$$\sum_{t=1}^{N} W_{Nt}\rho(Y_t, \vartheta) \ \underset{p}{\longrightarrow}\ r(x, \vartheta)\ ,\ \sum_{t=1}^{N} W_{Nt}L(Y_t) \ \underset{p}{\longrightarrow}\ \ell(x).$$

The assertion follows from the more general Theorem 2 of Masry and Fan [8] who showed mean-square consistency of local polynomial estimates. The Nadaraya-Watson kernel estimate correponds to the special case of a local constant fit. Mark that the mixing properties of (X_t, Y_t) immediately transfer to $(X_t, \rho(Y_t, \vartheta))$ and $(X_t, L(Y_t))$.

Now, we want to apply Proposition 5.1 to the special case where

$$\rho(y, \vartheta) = \log \sum_{k=1}^{M} \frac{\pi_k}{\sigma}\varphi(\frac{y - \mu_k}{\sigma}) = \log p_{\vartheta}(y). \tag{8}$$

We restrict the admissible parameters ϑ to a compact set Θ_0 satisfying in particular

$$0 < c_\pi \leq \pi_k\ ,\ |\mu_k| \leq C_\mu\ ,\ k = 1, \ldots, M,\ 0 < c_\sigma \leq \sigma \leq C_\sigma \quad \text{for all}\ \ \vartheta \in \Theta_0. \tag{9}$$

for suitable constants c_π, C_μ, c_σ and C_σ. Using the abbreviation

$$P_k(y) = \frac{1}{p_\vartheta(y)} \frac{\pi_k}{\sigma} \varphi(\frac{y-\mu_k}{\sigma}) \ , \ k = 1, \ldots, M,$$

we have, recalling that $\pi_M = 1 - \pi_1 - \ldots - \pi_{M-1}$,

$$\begin{aligned}
\frac{\partial}{\partial \pi_k} \rho(y, \vartheta) &= \frac{1}{\pi_k} P_k(y) - \frac{1}{\pi_M} P_M(y), \ k = 1, \ldots, M-1, \\
\frac{\partial}{\partial \mu_k} \rho(y, \vartheta) &= \frac{y-\mu_k}{\sigma^2} P_k(y) \ , \ k = 1, \ldots, M, \\
\frac{\partial}{\partial \sigma} \rho(y, \vartheta) &= \frac{1}{\sigma} \sum_{k=1}^{M} \left\{ \left(\frac{y-\mu_k}{\sigma} \right)^2 - 1 \right\} P_k(y).
\end{aligned}$$

Using (9) and $0 \leq P_k(y) \leq 1, \ k = 1, \ldots, M$, we conclude that ρ is continuously differentiable with derivatives bounded by $c_1 y^2 + c_2$ uniformly on Θ_0 where $c_1, c_2 > 0$ are suitable constants:

$$||\nabla \rho(y, \vartheta)|| \leq c_1 y^2 + c_2$$

and we immediately also have

$$||\nabla r(\vartheta)|| = ||E\{\nabla \rho(X_1, \vartheta) | X_1 = x\}|| \leq c_1 E\{Y_1^2 | X_1 = x\} + c_2.$$

Therefore,

$$||\nabla \rho_0(y, \vartheta)|| = ||\nabla \rho(y, \vartheta) - \nabla r(\vartheta)|| \leq c_1 (y^2 + E\{Y_1^2 | X_1 = x\}) + 2c_2 = L(y),$$

and **(A4)** is satisfied on Θ_0. We conclude, combining Proposition 5.1 and Lemma 5.2,

Theorem 1 *Let $Y_0, \ldots, Y_N$ be a sample of a stationary mixture of autoregressions satisfying (1) with $X_t = Y_{t-1}$ and with state probabilities $\pi_k = \pi_k^0, \ k = 1, \ldots, K$, and innovation variance $\sigma^2 = \sigma_0^2$. Let $\{Y_t\}$ be α-mixing with mixing coefficients satisfying (7) for some $\delta > 0$, and let $E|\varepsilon_{t,k}|^{4+2\delta} < \infty$. Moreover, let $E\{Y_1^4 | Y_0 = x', Y_t = x''\}$ be uniformly bounded in $t \geq 1$ and x', x'' in some neighbourhood of x. Assume, furthermore, that the autogression functions $m_1, \ldots, m_M$ are continuously differentiable in a neighbourhood of x, and that the density p of the innovations $\varepsilon_{t,k}$ is positive and continuous everywhere.*

Let the kernel K satisfy conditions **(K)**, *let $\Theta_0 \subseteq \Theta$ be compact, satisfying (9) and $(\pi_1^0, \ldots, \pi_{M-1}^0, m_1(x), \ldots, m_M(x), \sigma_0) = \vartheta_0 \in \Theta_0$. Furthermore, let Θ_0 be small enough such that*

$$\begin{aligned}
r(x, \vartheta) &= E\left\{ \log \sum_{k=1}^{M} \frac{\pi_k}{\sigma} \varphi\left(\frac{Y_1 - \mu_k}{\sigma} \right) | Y_0 = x \right\} \\
&= \sum_{l=1}^{M} \pi_l^0 \int \log \left[\sum_{k=1}^{M} \frac{\pi_k}{\sigma} \varphi\left(\frac{\sigma_0}{\sigma} z + \frac{m_k(x) - \mu_k}{\sigma} \right) \right] \ p(z) dz \qquad (10)
\end{aligned}$$

has a unique global maximum in Θ_0 *at* $\vartheta = \vartheta_0$. *Then,*

$$\widehat{\vartheta}_N = \arg\max_{\vartheta\in\Theta_0} \sum_{t=1}^{N} K_h(x - Y_{t-1}) \log \sum_{k=1}^{M} \frac{\pi_k}{\sigma}\varphi(\frac{Y_t - \mu_k}{\sigma}) \ \xrightarrow[p]{} \ \vartheta_0$$

for $N \to \infty,\ h \to 0$ *such that* $Nh \longrightarrow \infty$.

Proof: We have to check the assumptions of Proposition 5.1, where **(A1)**, **(A2)**, **(A3)** follow immediately from the special form (8) and from (9) and where we already have shown **(A4)**. It remains to check **(A5)**, i.e. the assumptions of Lemma 5.2.

We first remark that by monotonicity and concavity of the logarithm, we have

$$\begin{aligned} -\log\sqrt{2\pi\sigma^2} &= \log\sum_{k=1}^{M}\pi_k\frac{1}{\sigma}\varphi(0) \geq \rho(y,\vartheta) \\ &\geq \sum_{k=1}^{M}\pi_k \log\frac{1}{\sigma}\varphi\left(\frac{y-\mu_k}{\sigma}\right) = -\log\sqrt{2\pi\sigma^2} - \sum_{k=1}^{M}\pi_k\frac{(y-\mu_k)^2}{2\sigma^2}. \end{aligned}$$

Therefore, moments and conditional moments of $\rho(Y_t, \vartheta)$ exist and are bounded if this holds for the corresponding moments of Y_t^2 as long as $\vartheta \in \Theta_0$.

As p is positive, continuous and integrable, it is bounded, and, therefore, the conditional density of Y_1 given $Y_0 = x$ satisfies

$$0 < f_1(y|x) = \sum_{k=1}^{M}\frac{\pi_k^0}{\sigma_0}p\left(\frac{y - m_k(x)}{\sigma_0}\right) \leq c$$

for some $c > 0$ and all x, y. The same bound applies to the stationary density f of Y_1 as

$$f(y) = \int f(y|x)f(x)dx \leq c\int f(x)dx = c,$$

and, by iteration, we get that the conditional density $f_t(y|x)$ of Y_t given $Y_0 = x$ is also bounded by c, as

$$f_t(y|x) = \int f_{t-1}(y|u)f_1(u|x)du \leq \sup_u f_{t-1}(y|u)\cdot\int f_1(u|x)du = \sup_u f_{t-1}(y|u).$$

Then, for the joint density $f_t(x, y)$ of Y_0, Y_t, we have

$$f_t(x,y) = f_t(y|x)f(x) \leq c^2 \quad \text{for all} \quad t > 1,\ x, y \in \mathbb{R}.$$

It remains to show that for $\beta = 2\delta$

$$E\{|Y_1|^{4+\beta}|Y_0 = x'\}\ ,\quad E\{Y_1^4|Y_0 = x', Y_t = x''\}\ ,\ E\{Y_{t+1}^4|Y_0 = x', Y_t = x''\}$$

are uniformly bounded in $t \geq 1$ and x', x'' in a neighbourhood of x, where the second term is dealt with by assumption. The first property follows from

$$\begin{aligned} E\{|Y_1|^{4+\beta}|Y_0 = x'\} &= \int |y|^{4+\beta} f_1(y|x')dy \\ &= \sum_{k=1}^{M} \frac{\pi_k^0}{\sigma_0} \int |y|^{4+\beta} p\left(\frac{y - m_k(x')}{\sigma_0}\right) dy \\ &= \sum_{k=1}^{M} \pi_k^0 \int |m_k(x') + \sigma z|^{4+\beta} p(z)dz, \end{aligned}$$

using continuity of m_k and $E|\varepsilon_{t,k}|^{4+\beta} < \infty$. Analogously, we get the boundedness condition on

$$E\{Y_{t+1}^4|Y_0 = x', Y_t = x''\} = E\{Y_{t+1}^4|Y_t = x''\}.$$

Finally, the differentiability of of $r(x, \vartheta$ and $\ell(x)$ follow immediately from the representation (10) and from our assumptions on $m_1, \ldots, m_M$ and p. ■

We remark that the assumption on $E\{Y_1^4|Y_0 = x', Y_t = x''\}$ follows for fixed t immediately from the continuity and positivity assumptions on $m_1, \ldots, m_M$ and p. As, by the mixing assumption, $E\{Y_1^4|Y_0 = x', Y_{t+1} = x''\} \longrightarrow E\{Y_1|Y_0 = x'\}$ for $t \to \infty$, uniform boundedness with respect to t will typically be satisfied. To guarantee it, we would have to impose considerably stronger assumptions on the smoothness of $m_1, \ldots, m_M, p$ and on the mixing rate.

5.2 Convergence of the EM algorithm

In this section, we study the behaviour of the EM-algorithm for an increasing number p of iterations. We follow the terminology and notation of [2] and [14]. Recall the definition (5) of $L(\vartheta|X, Y)$ which we call the incomplete data log likelihood. Mark that it coincides with the corresponding quantity for the finite mixture models in example 4.3 of [2] up to the localizing kernel factors $K_h(x - X_t)$. Our goal is to maximize $L(\vartheta|X, Y)$ w.r.t. $\vartheta \in \Theta$ to get estimates of $\pi_1, \ldots, \pi_{M-1}, m_1(x), \ldots, m_M(x)$ and σ.

(5) is rather hard to maximize directly. If we would have observed the "complete" data $(Y_t, Z_t),\ t = 1, \ldots, N$, instead we could just maximize the corresponding complete data local conditional log likelihood

$$L(\vartheta|X, Y, Z) = \sum_{t=1}^{N} K_h(x - X_t) \sum_{k=1}^{M} Z_{tk} \log\{\pi_k \varphi_{\mu_k,\sigma}(Y_t)\}. \tag{11}$$

This is of a much simpler form as it separates into terms depending on $\pi = (\pi_1, \ldots, \pi_{M-1})^T$ and on $\mu = (\mu_1, \ldots, \mu_M)^T, \sigma$ resp.

$$L_1(\pi|Y, Z) = \sum_{t=1}^{N} K_h(x - X_t) \sum_{k=1}^{M} Z_{tk} \log \pi_k, \tag{12}$$

$$
\begin{aligned}
L_2(\mu,\sigma|Y,Z) &= -\frac{\log(2\pi\sigma^2)}{2}\sum_{t=1}^{N} K_h(x-X_t) \\
&- \frac{1}{2\sigma^2}\sum_{k=1}^{M}\sum_{t=1}^{N} K_h(x-X_t)Z_{tk}(Y_t-\mu_k)^2 \qquad (13)
\end{aligned}
$$

using $Z_{t1} + \ldots + Z_{tM} = 1$ and $\pi_1 + \ldots + \pi_M = 1$.

Remark 3 *Maximizing equation (12) as function of $\pi_k, k = 1, \cdots, M$ can be regarded as a constraint optimization problem. Therefore, an application of a Lagrange multiplier procedure yields*

$$
\hat{\pi_k} = \frac{\sum_{t=1}^{N} K_h(x-X_t)Z_{tk}}{\sum_{t=1}^{N} K_h(x-X_t)}. \qquad (14)
$$

By an application of Lemma 5.1, the Z_{tk} are i.i.d.,

$$
\hat{\pi}_k \longrightarrow \mathbb{E}Z_{tk}.
$$

Furthermore,

$$
\frac{\partial L_2(\mu,\sigma|Y,Z)}{\partial\sigma^2} = 0 \qquad (15)
$$

yields

$$
\hat{\sigma}^2 = \frac{\sum_{t=1}^{N}\sum_{k=1}^{M} K_h(x-X_t)e_{tk}^2 Z_{tk}}{\sum_{t=1}^{N} K_h(x-X_t)}
$$

for which an application of Lemma 5.2 implies

$$
\hat{\sigma}^2 \longrightarrow \sum_{k=1}^{M}\mathbb{E}(Z_{tk}(Y_t-\mu_k)^2) = \sum_{k=1}^{M}\mathbb{E}Z_{tk}\mathbb{E}(Y_t-\mu_k)^2.
$$

However, the Z_{tk} are not observable and therefore need to be estimated.

The basic idea of the EM algorithm is to replace $L(\vartheta|X,Y,Z)$ which contains the hidden variables Z_{tk} by its conditional expectation given only $Y = (Y_1,\ldots,Y_N)^T$ where the latter is calculated w.r.t. the parameter ϑ^* of a previous iteration. We get

$$
\begin{aligned}
Q(\vartheta|\vartheta^*) &= \mathrm{E}\{L(\vartheta|X,Y,Z)\ |Y,\vartheta^*\} \\
&= \sum_{t=1}^{N} K_h(x-X_t)\sum_{k=1}^{M}\mathrm{E}\{Z_{tk}|Y,\vartheta^*\}\log(\pi_k\varphi_{\mu_k,\sigma}(Y_t)) \\
&= \sum_{t=1}^{N} K_h(x-X_t)\sum_{k=1}^{M}\zeta_{tk}^*\log(\pi_k\varphi_{\mu_k,\sigma}(Y_t))
\end{aligned}
$$

where $\zeta_{tk}^* = \mathrm{E}\{Z_{tk}|Y, \vartheta^*\}$. Now, using this terminology, the EM-algorithm iterates between the following two steps

E-step: Given $\hat{\vartheta}^{(p)}$, determine $Q(\vartheta|\hat{\vartheta}^{(p)})$, i.e. determine $\zeta_{tk}^{(p)} = \mathrm{E}\{Z_{tk}|Y, \hat{\vartheta}^{(p)}\}$.

M-step: Set $\hat{\vartheta}^{(p+1)} = \arg\max_{\vartheta\in\Theta} Q(\vartheta|\hat{\vartheta}^{(p)})$.

The M-step defines a mapping $\hat{\vartheta}^{(p)} \longmapsto \hat{\vartheta}^{(p+1)} = M(\hat{\vartheta}^{(p)})$ which obviously satisfies $Q(M(\vartheta^*)|\vartheta^*) \geq Q(\vartheta^*|\vartheta^*)$ for all $\vartheta^* \in \Theta$. Therefore, our algorithm is a GEM algorithm in the sense of [2]. We set

$$\begin{aligned} H(\vartheta|\vartheta^*) &= Q(\vartheta|\vartheta^*) - L(\vartheta|X, Y) \\ &= \sum_{t=1}^{N} K_h(x - X_t) \left\{ \sum_{k=1}^{M} \zeta_{tk}^* \log[\pi_k \varphi_{\mu_k,\sigma}(Y_t)] - \log[\sum_{k=1}^{M} \pi_k \varphi_{\mu_k,\sigma}(Y_t)] \right\} \end{aligned}$$

where

$$\zeta_{tk}^* = \mathrm{E}\{Z_{tk}|Y, \vartheta^*\} = \frac{\pi_k^* \varphi_{\mu_k^*,\sigma^*}(Y_t)}{\sum_{l=1}^{M} \pi_l^* \varphi_{\mu_l^*,\sigma^*}(Y_t)}. \tag{16}$$

Correspondingly, we write

$$\zeta_{tk} = \mathrm{E}\{Z_{tk}|Y, \vartheta\} = \frac{\pi_k \varphi_{\mu_k,\sigma}(Y_t)}{\sum_{l=1}^{M} \pi_l \rho_{\mu_e,\sigma}(Y_t)}.$$

As $\zeta_{t1}^* + \ldots + \zeta_{tK}^* = 1$, we get

$$H(\vartheta|\vartheta^*) = \sum_{t=1}^{N} K_h(x - X_t) \sum_{k=1}^{M} \zeta_{tk}^* \log \zeta_{tk}.$$

By a corollary to Jensen's inequality, compare formula (1e6.6) of [9] with μ as the counting measure, we get that

$$\sum_{k=1}^{M} \zeta_{tk}^* \log \frac{\zeta_{tk}^*}{\zeta_{tk}} \geq 0$$

with equality iff $\zeta_{tk} = \zeta_{tk}^*,\ k = 1, \ldots, M$. It follows as in Lemma 1 of [2]

$$H(\vartheta^*|\vartheta^*) \geq H(\vartheta|\vartheta^*) \tag{17}$$

with equality iff $\zeta_{tk} = \zeta_{tk}^*,\ k = 1, \ldots, K$, for all t with $K_h(x - X_t) > 0$.

We conclude as in Theorem 1 of [2]

$$L(M(\vartheta^*)|Y) \geq L(\vartheta^*|Y) \quad \text{for all} \quad \vartheta^* \in \Theta \tag{18}$$

with equality iff both $Q(M(\vartheta^*)|\vartheta^*) = Q(\vartheta^*|\vartheta^*)$ and $\mathrm{E}\{Z_{tk}|Y, M(\vartheta^*)\} = \mathrm{E}\{Z_{tk}|Y, \vartheta^*\},\ k = 1, \ldots, M$, for all t with $K_h(x - X_t) > 0$.

(18) implies that in the course of the EM algorithm the incomplete data log likelihood increases monotonically, i.e. $L(\hat{\vartheta}^{(p+1)}|Y) \geq L(\hat{\vartheta}^{(p)}|Y),\ p \geq 0$. This implies a.s. convergence of the EM algorithm to a stationary point of $L(\vartheta|X, Y)$.

Theorem 2 *Let $N > K$ and $Y_s \neq Y_t$ for all $s \neq t$. Let h be chosen such that*

$$\min_{1\le t_1<\ldots<t_M\le N} \max_{t\notin\{t_1,\ldots,t_M\}} K_h(x - X_t) = \kappa > 0. \tag{19}$$

Then, all limit points of EM-sequences $\hat{\vartheta}^{(p)}$, starting in arbitrary $\hat{\vartheta}^{(0)}$ in the interior Θ^0 of Θ, are stationary points of $L(\vartheta|X,Y)$, i.e. $\nabla L(\vartheta|Y) = 0$, and $L(\hat{\vartheta}^{(p)}|Y)$ converges monotonically increasing to $L^* = L(\vartheta^*|Y)$ for some stationary point ϑ^*.

Proof: a) We first show that $L(\vartheta|X,Y)$ is bounded from above and converges to $-\infty$ for $\sigma \to 0$ uniformly in $\pi_1,\ldots,\pi_{M-1},\mu_1,\ldots,\mu_M$.

$$\begin{aligned}
L(\vartheta|X,Y) &= \sum_{t=1}^{N} K_h(x - X_t) \log\left(\sum_{k=1}^{M} \pi_k \frac{1}{\sqrt{2\pi\sigma^2}} e^{-\frac{(Y_t-\mu_k)^2}{2\sigma^2}}\right) \\
&= \sum_{t=1}^{N} K_h(x - X_t) \left\{-\frac{1}{2}\log(2\pi\sigma^2) + \log\left(\sum_{k=1}^{M} \pi_k e^{-\frac{(Y_t-\mu_k)^2}{2\sigma^2}}\right)\right\} \\
&\le -\frac{1}{2}\sum_{t=1}^{N} K_h(x - X_t)\log(2\pi\sigma^2) - \frac{1}{2\sigma^2}\sum_{t=1}^{N} K_h(x - X_t)\underline{e}_t^2
\end{aligned}$$

where, setting $\underline{e}_t^2 = \min_{k=1,\ldots,M}(Y_t - \mu_k)^2$, we have used monotonicity of log and exp and the fact, that $\pi_k, k = 1,\ldots,M$, sum up to 1. To get an upper bound for the second term on the right-hand side, we set $\eta = \frac{1}{2}\min\{|Y_t - Y_s|,\ 1 \le t < s \le N\} > 0$ a.s. Then, for each $k = 1,\ldots,M$, we have $|Y_t - \mu_k| < \eta$ for at most one $t = t_k$. Consequently, $\underline{e}_t^2 \ge \eta^2$ for all but at most M values of t. Therefore, with $\mathcal{T} = \{t;\ \underline{e}_t^2 \ge \eta^2\}$,

$$\begin{aligned}
L(\vartheta|X,Y) &\le -\frac{1}{2}\sum_{t=1}^{N} K_h(x - X_t)\log(2\pi\sigma^2) - \frac{\eta^2}{2\sigma^2}\sum_{t\in\mathcal{T}} K_h(x - X_t) \\
&\le -\frac{1}{2}\sum_{t=1}^{N} K_h(x - X_t)\log(2\pi\sigma^2) - \frac{\eta^2}{2\sigma^2}\max_{t\in\mathcal{T}} K_h(x - X_t) \\
&\le -\frac{1}{2}\sum_{t=1}^{N} K_h(x - X_t)\log(2\pi\sigma^2) - \frac{\eta^2\kappa}{2\sigma^2} \\
&\longrightarrow -\infty \quad \text{for } \sigma \to 0.
\end{aligned}$$

b) Remarking that L is continuous in Θ and differentiable in Θ^0, $Q(\vartheta|\vartheta^*)$ is continuous in ϑ and ϑ^*, and $H(\vartheta|\hat{\vartheta}^{(p)})$ is maximized over Θ at $\vartheta = \hat{\vartheta}^{(p)}$ by (17), we can apply the same arguments as in the proof of Theorem 2 of [14]. It only remains to show that $\Theta_{\hat{\vartheta}^{(p+1)}} \subseteq \Theta^0$ if $\hat{\vartheta}^{(p)} \in \Theta^0$ and that

$$\Theta_{\vartheta^*} = \{\vartheta \in \Theta;\ L(\vartheta|X,Y) > L(\vartheta^*|Y)\}$$

is compact for all $\vartheta^* \in \Theta$. The first property follows immediately from the iterative definition of $\hat{\pi}_k^{(p)}$, $k = 1, \ldots, M$, which are greater than 0 for all p and, therefore, also less than 1 for all p provided $0 < \hat{\pi}_k^{(0)} < 1$ for $k = 1, \ldots, M$. The compactness of Θ_{ϑ^*} follows from a), as L is continuous, L is uniformly bounded over $\{\vartheta \in \Theta;\ \sigma^2 \geq \delta\}$ for any $\delta > 0$ and $L(\vartheta|X, Y) < L(\vartheta^*|Y)$ for any ϑ with small enough variance component σ^2. ■

We remark that condition (19) is always satisfied if the support of the kernel K is $\mathbb{R}$ like for the Gaussian kernel. Otherwise, if K has compact support, we have to choose h large enough such that at least $M+1$ of the X_t are in the support of $K_h(x-.)$. Asymptotically for $N \to \infty$, this condition will hold anyhow, as the number of data in the support will be of the order Nh, which converges to ∞ under the usual consistency assumptions for kernel smoothers.

6 Conclusion

For a first simple example, we have illustrated that the local quasi maximum likelihood approach is applicable to mixtures of nonparametric regression and autoregression models. The EM algorithm provides a numerical method for calculating the function estimates which reduces to applying common local smoothers as part of an iterative scheme. The applications to artificial and real data look promising, but there are, of course, a lot of possible extensions and open questions to be adressed in future work. Apart from having a look at mixtures of more general models and allowing for Markovian instead of independent switching between states, automatic methods for choosing the smoothing parameter h as well as the number of states M are of prime interest. Also, the suitability of local polynomials and other local nonparametric function estimates for the mixture framwework has to be investigated.

References

[1] P. Bosq. *Nonparametric Statistics for Stochastic Processes*, 2nd ed. Lecture Notes in Statistics 110. Springer, Berlin-Heidelberg-New York, 1990.

[2] A.P. Dempster, N.M. Laird and D.B. Rubin. *Maximum Likelihood from incomplete data via the EM algorithm.* J. Royal Statist. Soc. B, 44, 1-38 (1977).

[3] J. Fan and I. Gijbels. *Local Polynomial Modelling and its Applications* Chapman and Hall, London, 1996.

[4] J. Fan and Q. Yao. *Nonlinear Time Series - Nonparametric and Parametric Methods* Springer, Berlin-Heidelberg-New York, 2005.

[5] W. Härdle. *Applied Nonparametric Regression.* Cambridge University Press, Cambridge, 1990.

[6] W. Härdle and A.B. Tsybakov. *Robust nonparametric regression with simultaneous scale curve estimation.* Ann. Statist. 16, 120-135 (1988).

[7] W. Härdle and P. Vieu. *Kernel regression smoothing of time series.* J. Time Series Anal., 13, 209-232 (1992).

[8] E. Masry and J. Fan. *Local polynomial estimation of regression functions for mixing processes.* Scand. Journal of Statistics 24, 165-179 (1997).

[9] C.R. Rao. *Linear Statistical Inference and its Applications*, 2nd ed. Wiley, New York, 1973.

[10] P. Robinson. *Nonparametric estimators for time series.* J. Time Series Anal., 4, 185-207 (1983).

[11] J.P. Stockis, J. Tadjuidje-Kamgaing and J. Franke. *On geometric ergodicity of CHARME Models.* J. Time Series Anal. (under revision).

[12] C.S. Wong and W.K. Li. *On a mixture autoregressive model.* J. Royal Statist. Soc. B 62, 95-115 (2000).

[13] C.S. Wong and W.K. Li. *On a mixture autoregressive conditional heteroscedastic model.* J. American Statist. Assoc. 96, 982-995 (2001).

[14] C.F.J. Wu. *On the convergence properties of the EM algorithm.* Ann. Statist. 11, 95-103 (1983).