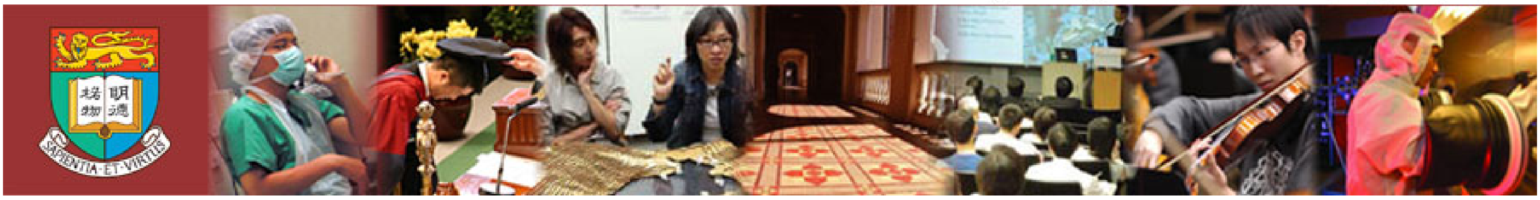

Title	**Mixtures of nonparametric autoregressions**
Author(s)	**Franke, J; Stockis, JP; TadjuidjeKamgaing, J; Li, WK**
Citation	**Journal Of Nonparametric Statistics, 2011, v. 23 n. 2, p. 287-303**
Issued Date	**2011**
URL	**http://hdl.handle.net/10722/134474**
Rights	**This is an electronic version of an article published in Journal of Nonparametric Statistics, 2011, v. 23 n. 2, p. 287-303. The article is available online at: http://www.tandfonline.com/doi/abs/10.1080/10485252.2010.539686**

Journal of Nonparametric Statistics
Vol. 00, No. 0, Month 2010, 1–17

Mixtures of nonparametric autoregressions

J. Franke[a]*, J.-P. Stockis[a], J. Tadjuidje-Kamgaing[a] and W. K. Li[b]

[a]*Department of Mathematics, University of Kaiserslautern, D-67653 Kaiserslautern, Germany;*
[b]*Department of Statistics and Actuarial Science, University of Hongkong, Hongkong, China*

(*Received 26 June 2009; final version received 08 November 2010*)

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 characterising 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.

Keywords: nonparametric regression; nonparametric autoregression; mixture; hidden variables; EM algorithm; kernel estimates; local likelihood

AMS Subject Classification: 62G08; 62M10

1. Introduction

We consider regressions and autoregressions which may be represented as a mixture of M different nonlinear models. We assume all over this paper that the available data $(X_1, Y_1), \ldots, (X_N, Y_N)$ are part of a strictly stationary time series. This includes the regression case, where (X_j, Y_j), $j = 1, \ldots, N$, are pairs of i.i.d. observations, as well as the autoregressive situation where $X_j = (Y_{j-1}, \ldots, Y_{j-p})$ consists of observations from the past of the stationary time series with current value Y_j. For the sake of simplicity, we restrict our considerations to one-dimensional variables $X_1, \ldots, X_N \in \mathbb{R}$, i.e. in the autoregressive case, to processes of order 1. We assume that the data are generated by the following independent switching model:

$$Y_t = \sum_{k=1}^{M} Z_{t,k}\{m_k(X_t) + \sigma_0 \varepsilon_{t,k}\}, \tag{1}$$

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 $\sigma_0^2 > 0$ is the residual variance. $Z_t = (Z_{t1}, \ldots, Z_{tM})^{\mathrm{T}}$ 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.

*Corresponding author. Email: franke@mathematik.uni-kl.de

ISSN 1048-5252 print/ISSN 1029-0311 online

DOI: 10.1080/10485252.2010.539686
http://www.informaworld.com

Furthermore, we assume that Z_t is independent of $X_j, \varepsilon_{j,k}, j \leq t$. Let

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

be the probability that Y_t is generated from X_t using the kth regression model, where $\pi_1^0 + \cdots + \pi_M^0 = 1$. If, e.g. the $\varepsilon_{t,k}$ are standard normal variables with Φ denoting their distribution function, then the conditional distribution function of Y_t given $X_t = x$ would be

$$F(y|x) = \mathrm{pr}(Y_t \leq y | X_t = x) = \sum_{k=1}^{M} \pi_k^0 \Phi\left(\frac{y - m_k(x)}{\sigma_0}\right). \tag{2}$$

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:

$$Y_t = \sum_{k=1}^{M} Z_{t,k}\{m_k(Y_{t-1}) + \sigma_0 \varepsilon_{t,k}\}. \tag{3}$$

In the special case, where the autoregression functions are all linear, i.e. $m_k(x) = \phi_{k0} + \phi_{k1}x,\ k = 1, \ldots, M$, we get a mixture autoregressive model as considered by Wong and Li (2000). Conditions on $\pi_1^0, \ldots, \pi_M^0,\ m_1, \ldots, m_M$ for the existence of a stationary solution of Equation (3) have been given in a much more general context in Stockis, Tadjuidje-Kamgaing, and Franke (2010). Here, we only remark that some of the autoregressive dynamics characterised by $m_k(x)$ may be explosive provided that they occur rarely enough, i.e. π_k^0 is small enough.

The assumption of independent state variables Z_t is motivated, e.g. by the following situation which is typical for mixture models: we consider independent data $(X_j, Y_j),\ j = 1, \ldots, N$, and we want to find a regression relation nonparametrically. The sample is not homogeneous, and the observations come from M different populations, such that, for each of them, we have to estimate a separate regression function $m_k(x) = E\{Y_t | X_t = x\}$. However, we do not know which observation comes from which population. Nevertheless, we want to estimate $m_1(x), \ldots, m_M(x)$ and, simultaneously, the asymptotic proportions $\pi_1^0, \ldots, \pi_M^0$ of the M subsamples in the total sample.

In case where the data come from a time series, assuming independence of the state variables is 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 time series 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 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 (2001) for the parametric case or Stockis et al. (2010) for the general case.

In the next section, we present a local quasi maximum likelihood approach to derive simultaneous estimates of all 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 and real data. Finally, in the technical appendix, we have a look at a more general model and, in that context, prove consistency of the local quasi maximum likelihood estimates and convergence of a related EM algorithm.

2. Local quasi maximum 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 $\pi_1, \dots, \pi_{M-1}, \sigma$ as well as for the regression functions $m_1(x), \dots, m_M(x)$. Mark that π_M is only used as an abbreviation for $1 - \pi_1 - \cdots - \pi_{M-1}$ throughout the paper, and it is not a free parameter. 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 (compare Robinson 1983; Härdle 1990; Härdle and Vieu 1992; Fan and Gijbels 1996; Fan and Yao 2005). As we consider distributions, we, moreover, rely on the general local likelihood regression approach of Tibshirani and Hastie (1987), compare also Fan, Farmen, and Gijbels (1998) and, for a survey, the book of Loader (1999). In particular, our approach is related to the work of Carroll, Ruppert, and Welsh (1998) who also consider essentially M-estimates of local parameters depending on an exogeneous variable Z which, however, in their case is continuous and observable.

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) $K(u) \geq 0,\ K(-u) = K(u),\ \int K(u)\,\mathrm{d}u = 1$, and the support of K is compact.

These conditions could be relaxed, but again we prefer to keep this exposition as simple as possible. $K_h(u) = (1/h)K(u/h)$ 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 $\varepsilon_{t,1}$ 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 maximises the local conditional log-likelihood function

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

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

For the general case $M \geq 1$, we consider the corresponding Gaussian local conditional quasi log likelihood

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

$\vartheta = (\pi_1, \dots, \pi_{M-1},\ \mu_1, \dots, \mu_M, \sigma)^{\mathrm{T}} \in \Theta$ denotes the partly local parameter where $\Theta \subseteq \mathbb{R}^{2K}$ is the set of admissible parameters satisfying the constraints $\sigma > 0, \pi_k \geq 0$ for $k = 1, \dots, M - 1$ and $\pi_1 + \cdots + \pi_{M-1} \leq 1$.

Mark that, throughout the paper, we do not assume normality of the residuals $\varepsilon_{t,k}$. They only have to satisfy some moment conditions and have a positive density, compare Section A.1. Therefore, maximising $L(\vartheta|X, Y)$ with respect to ϑ provides only a local quasi maximum Gaussian likelihood estimate $\hat{\vartheta}_N$.

As we use a Gaussian quasi likelihood, i.e. essentially a local least squares approach, the resulting estimates are not robust against outliers. If the distribution of the residuals $\varepsilon_{t,k}$ may be heavy-tailed, using general M-smoothers instead would be advisable, compare, e.g. Härdle and Gasser (1984) and Härdle and Tuan (1986). In that case, we have to replace φ in Equation (5) by the density of an appropriate heavy-tailed distribution standardised to mean 0 and variance 1 and sharing some regularity assumptions with the normal density. The theory of the appendix still holds. However, in general, we do no longer have explicit formulas for the local quasi maximum likelihood estimates like the Nadaraya–Watson estimates of, e.g. Equation (6). We have to consider numerical solutions which increases the computational load considerably.

3. The EM algorithm

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

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

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

$$\tilde{m}_k(x, h) = \frac{\sum_{t\in T_k} K_h(x - X_t)\, Y_t}{\sum_{t\in T_k} K_h(x - X_t)} = \frac{\sum_{t=1}^{N} K_h(x - X_t)\, Y_t Z_{tk}}{\sum_{t=1}^{N} K_h(x - X_t) Z_{tk}} \tag{6}$$

as the Z_{tk} are either 1 or 0. This vector of function estimates $(\tilde{m}_1(x, h), \dots, \tilde{m}_M(x, h))^{\mathrm{T}}$ is the solution of the weighted least-squares problem:

$$\text{Minimise} \quad \sum_{t=1}^{N}\sum_{k=1}^{M}(Y_t - \mu_k)^2 Z_{tk} K_h(x - X_t) \quad \text{w.r.t. } \mu_1, \dots, \mu_M \in \mathbb{R}!$$

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 which are calculated pretending (but not assuming) 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 $\mathcal{N}(m_k(x), \sigma_0^2)$. Therefore,

$$\begin{aligned}\zeta_{tk}^0 &= E\{Z_{tk}|Y_t, X_t\} = \mathrm{pr}\{Z_{tk} = 1|Y_t, X_t\}\\ &= \frac{\pi_k^0 (1/\sigma_0)\varphi(Y_t - m_k(X_t)/\sigma_0)}{\sum_{l=1}^{M}\pi_l^0(1/\sigma_0)\varphi(Y_t - m_l(X_t)/\sigma_0)}.\end{aligned}$$

As we do not know the parameters π_k^0 and σ_0 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 $\hat{\pi}_1, \dots, \hat{\pi}_M,\ \hat{\sigma}$ 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 and

X_t are estimated by

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

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

$$\hat{\pi}_k = \frac{1}{N} \sum_{t=1}^{N} \zeta_{tk}, \quad k = 1, \dots, M. \tag{7}$$

We estimate the M regression functions by

$$\hat{m}_k(x, h) = \frac{\sum_{t=1}^{N} K_h(x - X_t) Y_t \zeta_{tk}}{\sum_{t=1}^{N} K_h(x - X_t) \zeta_{tk}}, \quad k = 1, \dots, M, \tag{8}$$

and the residual variances by

$$\hat{\sigma}^2 = \frac{1}{N} \sum_{t=1}^{N} \sum_{k=1}^{M} e_{tk}^2 \zeta_{tk}, \tag{9}$$

where $e_{tk} = Y_t - \hat{m}_k(X_t, h)$.

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

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

The EM-algorithm is a computationally simple numerical procedure for maximising the Gaussian local conditional log likelihood $L(\vartheta | X, Y)$ of Equation (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.

4. Numerical examples

For fitting model (1) to the following data, we used a straightforward implementation of the EM algorithm described in Section 3 as a MATLAB 7.0 subroutine. On an up-to-date standard desktop PC, one step of the iteration took about 0.5 sec for the artificial data set with $N = 1000$ of Section 4.1, and about 5.4 sec for the heart rate data of Section 4.2 with $N = 2812$. Convergence to the shown estimates was achieved rather fast after 50–100 iterations depending on the starting values. Those results may, however, give a too optimistic view of the numerical efforts. In another numerical experiment with artificial data, not reported here, we considered two states with differing standard deviations $\sigma_1 \neq \sigma_2$, and in that case, the EM algorithm considerably needed more iterations (about 2000) to converge. For sample size 1000, the whole procedure took about 15 min of computation time.

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 $\pi_1^0 = 0.7$, $\pi_2^0 = 1 - \pi_1^0 = 0.3$, the

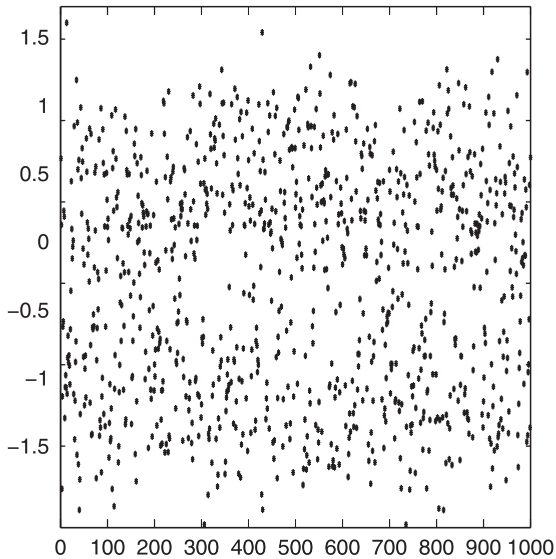

Figure 1. Simulated data.

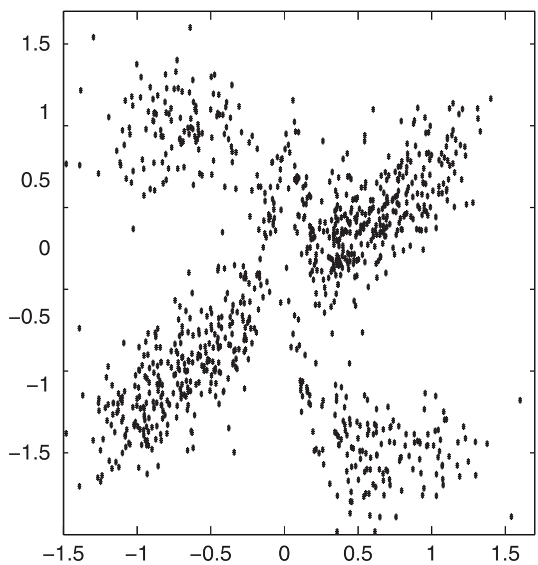

Figure 2. Scatter plot simulated data.

innovation variance as $\sigma_0^2 = 0.2$ 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}. **Q1**

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 $\hat{\sigma}^2 = 0.2004$.

Figure 3 shows m_1 and 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.

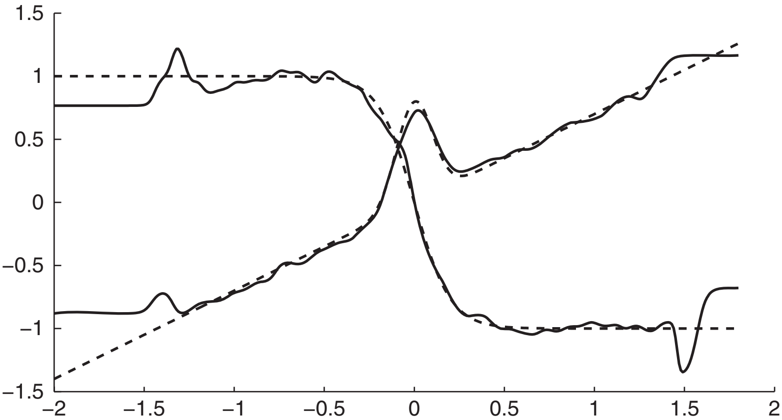

Figure 3. Estimated trend functions.

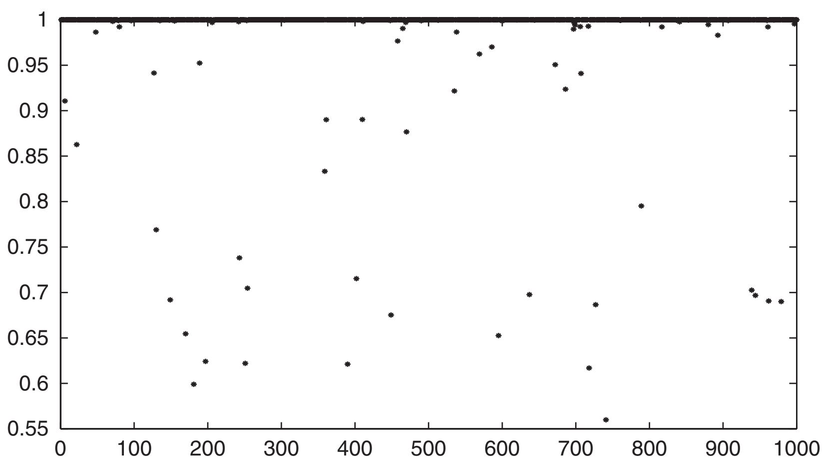

Figure 4. Maximum of the estimated state probabilities: Simulated data.

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. The data are available at the first author's homepage (www.mathematik.uni-kl.de/ ∼franke). Figure 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

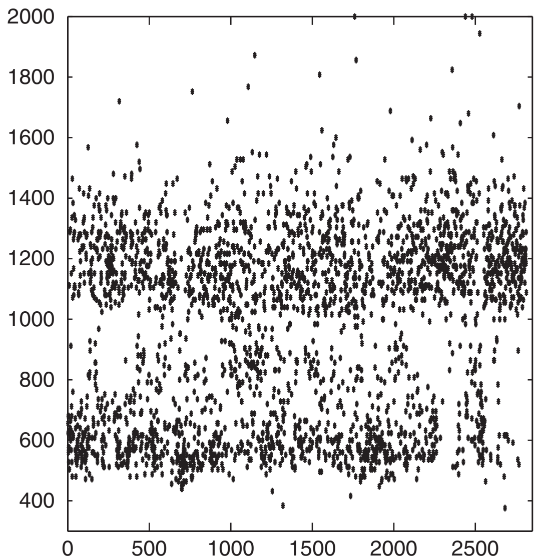

Figure 5. Heart rate data.

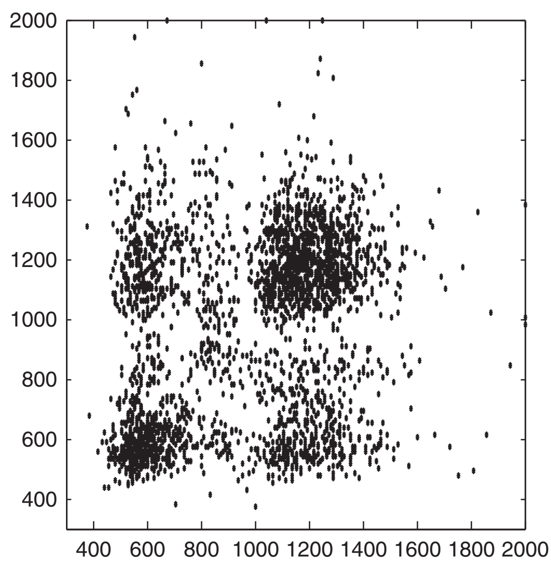

Figure 6. Scatter plot.

show more or less 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.

Figure 8 shows $\max(\zeta_{t1}, \zeta_{t2}, \zeta_{t3})$ which almost always are at least 0.5 and frequently considerably larger, i.e. there is a clear decision for one of the three phases in the large majority of cases.

We also have fitted a mixture model with four 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

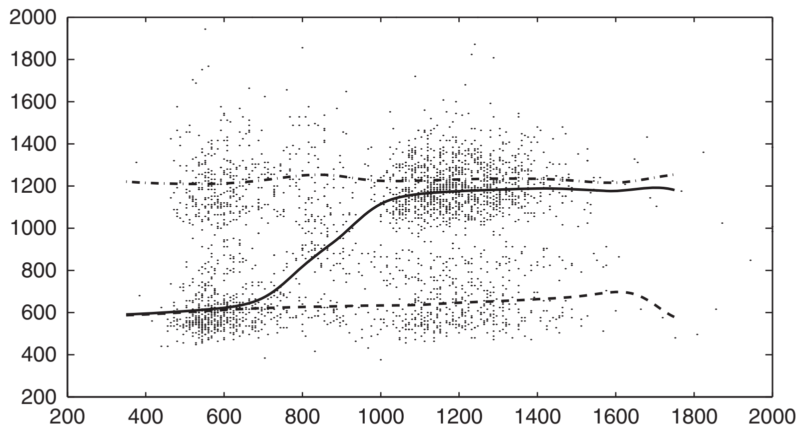

Figure 7. Scatter plot and functions estimates: The upper dashed curve represents the first state trend function, the lower dashed the second state function and the third is represented by the solid curve.

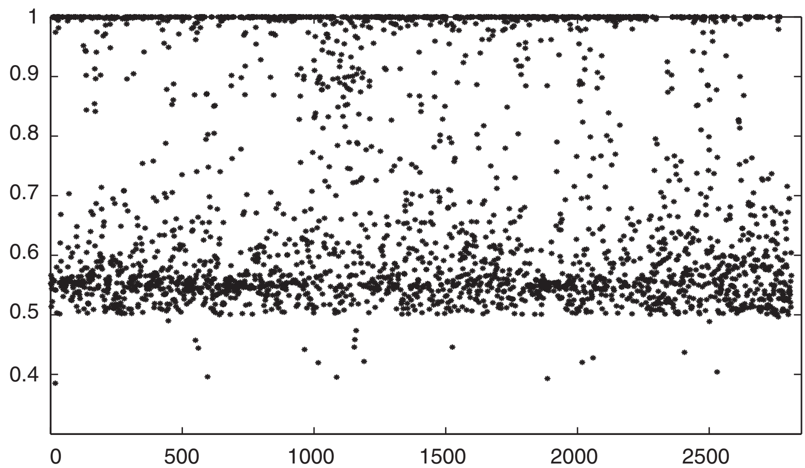

Figure 8. Maximum of the estimated state probabilities.

lower curve in Figure 7 was replaced by two kernel estimates which both were roughly constant and differed only slightly, i.e. they essentially estimated the same autoregressive function and represented the same data generating mechanism.

A similar observation has been made for the computer generated data where we have considered one more state in the estimation procedure than present in the mechanism used for generating the data.

5. 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 addressed in future work. Apart from having a look at mixtures of more general models and allowing for Markovian instead of independent switching between states, the asymptotics of the local parameter estimates and 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 framework has to be investigated.

Acknowledgements

We thank an anonymous referee for recommendations which led to a considerable improvement of the paper. The work was supported by the Deutsche Forschungsgemeinschaft (DFG) as well as by the *Center for Mathematical and Computational Modelling* $(CM)^2$ funded by the state of Rhineland-Palatinate.

References

Q2 Bosq, P. (1990), '*Nonparametric Statistics for Stochastic Processes*' (Vol. 110, 2nd ed.), Lecture Notes in Statistics, Berlin: Springer, 1990.

Carroll, R.J., Ruppert, D., and Welsh, A.H. (1998), 'Local Estimation Equations', *Journal of American Statistical Association*, 93, 214–227.

Dempster, A.P., Laird, N.M., and Rubin, D.B. (1977), 'Maximum Likelihood from Incomplete Data Via the EM Algorithm', *Journal of the Royal Statistical Society B*, 44, 1–38.

Fan, J., and Gijbels, I. (1996), *Local Polynomial Modelling and its Applications*, London: Chapman and Hall.

Fan, J., and Yao, Q. (2005), *Nonlinear Time Series – Nonparametric and Parametric Methods*, Berlin: Springer.

Fan, J., Farmen, M., and Gijbels, I. (1998), 'Local Maximum Likelihood Estimation and Inference', *Journal of the Royal Statistical Society B*, 60, 591–608.

Härdle, W., and Gasser, T. (1984), 'Robust Nonparametric Function Fitting', *Journal of the Royal Statistical Society B*, 46, 42–51.

Härdle, W., and Tuan, P-D. (1986), 'Some Theory of M Smoothing of time Series', *Journal of Time Series Analysis*, 7, 191–204.

Härdle, W., and Tsybakov, A.B. (1988), 'Robust Nonparametric Regression with Simultaneous Scale Curve Estimation', *Annals of Statistics*, 16, 120–135.

Härdle, W. (1990), *Applied Nonparametric Regression*, Cambridge: Cambridge University Press.

Härdle, W., and Vieu, P. (1992), 'Kernel Regression Smoothing of Time Series', *Journal of Time Series Analysis*, 13, 209–232.

Loader, C. (1999), *Local Regression and Likelihood*, Berlin: Springer.

Masry, E., and Fan, J. (1997), 'Local Polynomial Estimation of Regression Functions For Mixing Processes', *Scandinavian Journal of Statistics*, 24, 165–179.

Rao, C.R. (1973), *Linear Statistical Inference and its Applications*, 2nd ed., New York: Wiley.

Robinson, P. (1983), 'Nonparametric Estimators for Time Series', *Journal of Time Series Analysis*, 4, 185–207.

Stockis, J.P., Tadjuidje-Kamgaing, J., and Franke, J. (2008), 'A Note on the Identifiability of the Conditional Expectation for the Mixtures of Neural Networks', *Statistical Probability Letters*, 78, 739–742.

Stockis, J.P., Tadjuidje-Kamgaing, J., and Franke, J. (2010), 'On Geometric Ergodicity of Charme Models', *Journal of Time Series Analysis*, 31, 141–152.
Tibshirani, R., and Hastie, T. (1987), 'Local Likelihood Estimation', *Journal of American Statistical Association*, 82, 559–567.
Wong, C.S., and Li, W.K. (2000), 'On a Mixture Autoregressive Model', *Journal of Royal Statistical Society B*, 62, 95–115.
Wong, C.S., and Li, W.K. (2001), 'On a Mixture Autoregressive Conditional Heteroscedastic Model', *Journal of American Statistical Association*, 96, 982–995.
Wu, C.F.J. (1983). 'On the Convergence Properties of the EM Algorithm', *Annals of Statistics*, 11, 95–103.

Appendix

In the following, we consider a generalisation of the mixture model (1), allowing for a dependence of the innovation variance $s^2(X_t)$ and the state probabilities $\pi_k^0(X_t)$ on the current X_t:

$$Y_t = \sum_{k=1}^{M} Z_{t,k}\{m_k(X_t) + s(X_t)\varepsilon_{t,k}\}, \tag{A1}$$

where $\varepsilon_{t,k},\ t = 1, \dots, N,\ k = 1, \dots, M$, are i.i.d. random variables with mean 0 and variance 1, Z_t is conditionally independent of $X_s, Z_s, s < t, \varepsilon_{s,k}, s \le t$, given X_t, and

$$\text{pr}(Z_t = e_k | X_t = x) = \text{pr}(Z_{tl} = 0 \text{ for } l \ne k | X_t = x) = \pi_k^0(x), \quad k = 1, \dots, M,$$

with $\pi_1^0(x) + \cdots + \pi_M^0(x) = 1$.

A.1. *An auxiliary result on local M estimates*

For convenience, we first formulate an auxiliary result which we need for showing consistency of the local quasi maximum likelihood estimates of $\pi_k(x)$, $m_k(x)$ and $\sigma^2(x)$ of model (A1). We study the general local M-estimate $\hat{\vartheta}_N$ which maximises

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

for some function $\rho : \mathbb{R} \times \Theta \to \mathbb{R}$, 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

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

We assume that

(A1) Θ is compact.
(A2) $\rho(y, \vartheta)$ is continuous in ϑ, and $E|\rho(Y_1, \vartheta)| < \infty$.
(A3) $r(\vartheta)$ is continuous in ϑ 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 $Nh \to \infty$,

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

$$\sum_{t=1}^{N} W_{Nt}L(Y_t) \underset{\text{p}}{\longrightarrow} E\{L(Y_1) | X_1 = x\}.$$

Proposition A.1 *Under the conditions* (K) *on the kernel and* (A1), ..., (A5)*, the general M-estimate* $\hat{\vartheta}_N$ *is consistent for* ϑ_0*, i.e., for* $N \to \infty$, $h \to 0$, $Nh \to \infty$

$$\hat{\vartheta}_N = \arg\min_{\vartheta \in \Theta} R_N(\vartheta) \underset{p}{\longrightarrow} \vartheta_0 \quad \textit{for } N \longrightarrow \infty.$$

Proof We only sketch the main ideas, as the details are essentially the same as in proving a similar result by Härdle and Tsybakov (1988) on M-estimates in a location-scale regression model. First, a standard argument, covering the compact Θ by finitely many δ-balls, exploiting Lipschitz continuity (A4) and applying (A5), shows uniform convergence of $R_N(\vartheta)$ to $r(\vartheta)$, i.e.

$$\sup_{\vartheta \in \Theta} |R_N(\vartheta) - r(\vartheta)| = \sup_{\vartheta \in \Theta} \sum_{t=1}^{N} W_{Nt} \rho_0(Y_t, \vartheta) \underset{p}{\longrightarrow} 0.$$

Hence, $\hat{\vartheta}_N$ as the minimiser of $R_N(\vartheta)$ converges to the minimiser ϑ_0 of $r(\vartheta)$ using the identifiability assumption (A3). ■

Conditions (A1) and (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 $r(\vartheta)$. An identifiability condition is in particular necessary for the application to mixture models in the following subsection where $\rho(y, \vartheta)$ is the logarithm of a mixture density, compare Equation (A3). This density does not change, if we permute the numbering of the regimes, i.e. various different parameters lead to the same $\rho(y, \vartheta)$ and, then, $r(\vartheta)$. Additionally, if we have chosen M too large such that $m_k = m_j$ for some $k \neq j$, π_k and π_j will not be identifiable at all. To get a convergence result, we have to choose the parameter set Θ appropriately to exclude such ambiguities. For a more detailed discussion in a related context, compare Stockis, Tadjuidje-Kamgaing, and Franke (2008).

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 $r(x, \vartheta)$ and $\ell(x)$, on the rate of the bandwidth h and on the dependence structure of the time series (X_t, Y_t). In the case where (X_t, Y_t), $t = 1, \ldots, N$, are i.i.d., the assertion follows immediately from Proposition 3.1.1 of Härdle (1990) under the weak conditions that the second moments of $\rho(Y_1, \vartheta)$ and $L(Y_1)$ are finite and the density of X_t is continuous and positive in a neighbourhood of x. For time series, we use the following result under an α-mixing condition which follows from the more general Theorem 2 of Masry and Fan (1997), who showed mean-square consistency of local polynomial estimates.

Lemma A.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* $E\{|\rho(Y_1, \vartheta)|^{2+\delta}|X_1 = x'\}$ *and* $E\{L^{2+\delta}(Y_1)|X_1 = x'\}$ *are uniformly bounded for* x' *in some neighbourhood of* x *and*

$$\sum_{t=1}^{\infty} t^{\gamma} \alpha_t^{\delta/2+\delta} < \infty \quad for\ some\ \gamma > \frac{\delta}{2+\delta}. \tag{A2}$$

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

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

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

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

A.2. *Consistency of the local quasi maximum likelihood estimate*

For estimation in model (A1), we have to apply Proposition A.1 to the special case where

$$\rho(y,\vartheta) = \log \sum_{k=1}^{M} \frac{\pi_k}{\sigma} \varphi\left(\frac{y-\mu_k}{\sigma}\right) = \log p_\vartheta(y). \tag{A3}$$

is a Gaussian mixture quasi log likelihood. We restrict the admissible local parameters $\vartheta = (\pi_1, \dots, \pi_{M-1}, m_1, \dots, m_M, \sigma)$ to a compact set Θ_0 satisfying in particular

$$0 < c_\pi \le \pi_k, \quad |\mu_k| \le C_\mu, \quad k = 1, \dots, M, \quad 0 < c_\sigma \le \sigma \le C_\sigma \quad \text{for all } \vartheta \in \Theta_0. \tag{A4}$$

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

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

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

$$\frac{\partial}{\partial \pi_k} \rho(y,\vartheta) = \frac{1}{\pi_k} P_k(y) - \frac{1}{\pi_M} P_M(y), \quad k = 1, \dots, M-1,$$

$$\frac{\partial}{\partial \mu_k} \rho(y,\vartheta) = \frac{y-\mu_k}{\sigma^2} P_k(y), \quad k = 1, \dots, M,$$

$$\frac{\partial}{\partial \sigma} \rho(y,\vartheta) = \frac{1}{\sigma} \sum_{k=1}^{M} \left\{ \left(\frac{y-\mu_k}{\sigma}\right)^2 - 1 \right\} \quad P_k(y).$$

Using Equation (A4) and $0 \le P_k(y) \le 1,\ k = 1, \dots, M$, 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:

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

and we immediately also have

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

Therefore,

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

and (A4) is satisfied on Θ_0. We conclude, combining Proposition A.1 and Lemma A.2,

THEOREM A.3 *Let $Y_0, \dots, Y_N$ be a sample of a stationary mixture of autoregressions satisfying Equation* (A1) *with $X_t = Y_{t-1}$. Let $\{Y_t\}$ be α-mixing with mixing coefficients satisfying Equation* (A2) *for some $\delta > 0$, let the density p of the innovations $\varepsilon_{t,k}$ be positive and continuous everywhere, and $E|\varepsilon_{t,k}|^{4+2\delta} < \infty$. For given x, let $E\{Y_1^4 | Y_0 = x', Y_t = x''\}$ be uniformly bounded in $t \ge 1$ and x', x'' in some neighbourhood of x. Assume, furthermore, that the state probability functions $\pi_1^0, \dots, \pi_{M-1}^0$, the autoregression functions $m_1, \dots, m_M$ as well as the standard deviation function s are continuously differentiable in a neighbourhood of x, and that $s(u) \ge c_\sigma$ for all $u \in \mathbb{R}$ for some constant $c_\sigma > 0$.*

Let the kernel K satisfy conditions (K), let $\Theta_0 \subseteq \Theta$ be compact, satisfying Equation (A4) *and $(\pi_1^0(x), \dots, \pi_{M-1}^0(x), m_1(x), \dots, m_M(x), s(x)) = \vartheta_0 \in \Theta_0$. Furthermore, let Θ_0 be small enough such that*

$$r(x,\vartheta) = E\left\{ \log \sum_{k=1}^{M} \frac{\pi_k}{\sigma} \varphi\left(\frac{Y_1-\mu_k}{\sigma}\right) | Y_0 = x \right\}$$

$$= \sum_{l=1}^{M} \pi_l^0(x) \int \log \left[\sum_{k=1}^{M} \frac{\pi_k}{\sigma} \varphi\left(\frac{s(x)}{\sigma} z + \frac{m_k(x)-\mu_k}{\sigma}\right) \right] p(z)\, \mathrm{d}z \tag{A5}$$

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

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

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

Proof We have to check the assumptions of Proposition A.1, where (A1)–(A3) follow immediately from the special form (A3) and from Equation (A4) and where we already have shown (A4). It remains to check (A5), i.e. the assumptions of Lemma A.2.

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

$$-\log\sqrt{2\pi\sigma^2} = \log\sum_{k=1}^{M}\pi_k\frac{1}{\sigma}\varphi(0) \geq \rho(y,\vartheta)$$

$$\geq \sum_{k=1}^{M}\pi_k\log\frac{1}{\sigma}\varphi\left(\frac{y-\mu_k}{\sigma}\right) = -\log\sqrt{2\pi\sigma^2} - \sum_{k=1}^{M}\pi_k\frac{(y-\mu_k)^2}{2\sigma^2}.$$

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 $Y_0 = u$ satisfies

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

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

$$f(y) = \int f(y|u)f(u)\,\mathrm{d}u \leq c\int f(u)\,\mathrm{d}u = c,$$

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

$$f_t(y|u) = \int f_{t-1}(y|v)f_1(v|u)\,\mathrm{d}v \leq \sup_v f_{t-1}(y|v)\cdot\int f_1(v|u)\,\mathrm{d}v = \sup_v f_{t-1}(y|v).$$

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

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

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

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

are uniformly bounded in $t \geq 1$ and x', x'' in a neighbourhood of x, where the second term is dealt with by assumption. Using continuity of m_k, s and $E|\varepsilon_{t,k}|^{4+\beta} < \infty$, the first property follows from

$$E\{|Y_1|^{4+\beta}|Y_0 = x'\} = \int |y|^{4+\beta}f_1(y|x')\,\mathrm{d}y$$

$$= \sum_{k=1}^{M}\frac{\pi_k^0(x')}{s(x')}\int |y|^{4+\beta}p\left(\frac{y-m_k(x')}{s(x')}\right)\mathrm{d}y$$

$$= \sum_{k=1}^{M}\pi_k^0(x')\int |m_k(x')+s(x')v|^{4+\beta}p(v)\,\mathrm{d}v.$$

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 $r(x,\vartheta)$ and $\ell(x)$ follow immediately from the representation (A5) and from our assumptions on $\pi_1^0,\ldots,\pi_{M-1}^0, m_1,\ldots,m_M, s$ and p. ■

A.3. *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 Dempster, Laird, and Rubin (1977) and Wu (1983). Recall the definition (5) of $L(\vartheta|X,Y)$ which we call the incomplete data (quasi) log likelihood. Mark that it coincides with the corresponding quantity for the finite mixture models in Dempster et al. (1977, Example 4.3) up to the localising kernel factors $K_h(x - X_t)$. Our goal is to maximise $L(\vartheta|X,Y)$ w.r.t. $\vartheta \in \Theta$ to get estimates of $\pi_1^0(x),\ldots,\pi_{M-1}^0(x), m_1(x),\ldots,m_M(x)$ and $s(x)$.

Equation (5) is rather hard to maximise directly. If we would have observed the 'complete' data $(X_t, Y_t, Z_t),\ t = 1, \dots, N$, instead we could just maximise the corresponding complete data local conditional (quasi) log likelihood

$$L(\vartheta|X, Y, Z) = \sum_{t=1}^{N} K_h(x - X_t) \sum_{k=1}^{M} Z_{tk} \log\{\pi_k \varphi_{\mu_k,\sigma}(Y_t)\}. \tag{A6}$$

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

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

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

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

Maximising L_1 and L_2 yields explicit formulas for the solutions. Setting the partial derivatives of L_2 to 0, we get immediately

$$\hat{\mu}_k = \frac{\sum_{t=1}^{N} K_h(x - X_t) Z_{tk} Y_t}{\sum_{t=1}^{N} K_h(x - X_t) Z_{tk}}, \tag{A7}$$

$$\hat{\sigma}^2 = \frac{\sum_{t=1}^{N} \sum_{k=1}^{M} K_h(x - X_t) Z_{tk} e_{tk}^2}{\sum_{t=1}^{N} K_h(x - X_t)}, \quad e_{tk} = Y_t - \hat{\mu}_k. \tag{A8}$$

Maximising L_1 as function of $\pi_k, k = 1, \dots, M$, can be regarded as a constrained optimisation problem, and an application of a Lagrange multiplier procedure yields

$$\hat{\pi}_k = \frac{\sum_{t=1}^{N} K_h(x - X_t) Z_{tk}}{\sum_{t=1}^{N} K_h(x - X_t)}. \tag{A9}$$

Similar to Theorem A.3, we have under appropriate conditions for $N \to \infty$

$$\hat{\mu}_k \longrightarrow \mathbb{E}\{Y_t|X_t = x\} = m_k(x), \quad \hat{\sigma}^2 \longrightarrow \mathrm{var}\{Y_t|X_t = x\} = s^2(x), \quad \hat{\pi}_k \longrightarrow \mathbb{E}\{Z_{tk}|X_t = x\} = \pi_k^0(x).$$

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 $X = (X_1, \dots, X_N)^{\mathrm{T}}, Y = (Y_1, \dots, Y_N)^{\mathrm{T}}$ where the latter is calculated w.r.t. the parameter ϑ^* of a previous iteration. We get

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

where

$$\zeta_{tk}^* = \mathrm{E}\{Z_{tk}|X, Y, \vartheta^*\} = \frac{\pi_k^* \varphi_{\mu_k^*,\sigma^*}(Y_t)}{\sum_{l=1}^{M} \pi_l^* \varphi_{\mu_l^*,\sigma^*}(Y_t)}. \tag{A10}$$

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 $\zeta_{tk}^{(p)} = \mathrm{E}\{Z_{tk}|X, Y, \hat{\vartheta}^{(p)}\}$ from Equation (A10).

M-*step*: Set $\hat{\vartheta}^{(p+1)} = \arg\max_{\vartheta \in \Theta} Q(\vartheta|\hat{\vartheta}^{(p)})$, where the components $\hat{\pi}_1^{(p+1)}, \dots, \hat{\pi}_{M-1}^{(p+1)}, \hat{\mu}_1^{(p+1)}, \dots, \hat{\mu}_M^{(p+1)}, \hat{\sigma}^{(p+1)}$ of $\hat{\vartheta}^{(p+1)}$ are calculated from Equations (A7), (A8) and (A9), respectively, with $\zeta_{tk}^{(p)}$ replacing Z_{tk}.

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 Dempster et al. (1977). We set

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

using $\zeta_{t1}^* + \cdots + \zeta_{tK}^* = 1$, and writing

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

By a corollary to Jensen's inequality, compare formula (1e6.6) of Rao (1973) 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 Dempster et al. (1977)

$$H(\vartheta^*|\vartheta^*) \geq H(\vartheta|\vartheta^*) \tag{A11}$$

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 Dempster et al. (1977)

$$L(M(\vartheta^*)|X, Y) \geq L(\vartheta^*|X, Y) \quad \text{for all } \vartheta^* \in \Theta \tag{A12}$$

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

Equation (A12) implies that in the course of the EM algorithm the incomplete data log likelihood increases monotonically, i.e. $L(\hat{\vartheta}^{(p+1)}|X, Y) \geq L(\hat{\vartheta}^{(p)}|X, Y),\ p \geq 0$. This implies a.s. convergence of the EM algorithm to a stationary point of $L(\vartheta|X, Y)$.

THEOREM A.4 *Let $N > K$ and $Y_s \neq Y_t$ for all $s \neq t$. Let h be chosen such that*

$$\min_{1 \leq t_1 < \ldots < t_M \leq N} \ \max_{t \notin \{t_1, \ldots, t_M\}} K_h(x - X_t) = \kappa > 0. \tag{A13}$$

Then, all limit points of EM-sequences $\hat{\vartheta}^{(p)}$, starting in arbitrary $\hat{\vartheta}^{(0)}$ in the interior Θ° of Θ, are stationary points of $L(\vartheta|X, Y)$, i.e., $\nabla L(\vartheta|X, Y) = 0$, and $L(\hat{\vartheta}^{(p)}|X, Y)$ converges monotonically increasing to $L^* = L(\vartheta^*|X, Y)$ 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 $\pi_1, \ldots, \pi_{M-1}, \mu_1, \ldots, \mu_M$.

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

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 $\eta = \frac{1}{2} \min\{|Y_t - Y_s|,\ 1 \leq t < s \leq$

$N\} > 0$ a.s. Then, for each $k = 1, \ldots, M$, we have $|Y_t - \mu_k| < \eta$ for at most one $t = t_k$. Consequently, $\underline{e}_t^2 \geq \eta^2$ for all but at most M values of t. Therefore, with $\mathcal{T} = \{t; \underline{e}_t^2 \geq \eta^2\}$,

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

(b) Remarking that L is continuous in Θ and differentiable in Θ°, $Q(\vartheta|\vartheta^*)$ is continuous in ϑ and ϑ^*, and $H(\vartheta|\hat{\vartheta}^{(p)})$ is maximized over Θ at $\vartheta = \hat{\vartheta}^{(p)}$ by Equation (A11), we can apply the same arguments as in the proof of Theorem 2 of Wu (1983). It only remains to show that $\Theta_{\hat{\vartheta}^{(p+1)}} \subseteq \Theta^\circ$ if $\hat{\vartheta}^{(p)} \in \Theta^\circ$ and that

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

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 $L(\vartheta|X,Y) < L(\vartheta^*|X,Y)$ for any ϑ with small enough variance component σ^2. ■

We remark that condition (A13) is always satisfied if the support of the kernel K is $\mathbb{R}$ like for the Gaussian kernel. Otherwise, if K has a 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.

A.4. *Constant state probabilities and variances*

We now return to our original model (1), where $\pi_k^0(x) = \pi_k^0$, $k = 1, \ldots, M$, and $s^2(x) = \sigma_0^2$ do not depend on x. As this is a special case of Equation (A1), the results of the previous subsections remain valid. In Section 3, we have considered a different EM algorithm than the local one in Section A.2, taking into account explicitly the constancy of state probabilities and innovation variance. Could be done, but lengthy only simple heuristic argument why they are asymptotically equivalent **Q3** to first order of approximation.

For that purpose, we have a look at the case where the Z_t are observable, i.e. we consider the complete data quasi log likelihood. Maximising Equation (A6), we get the localised estimates $\hat{m}_k(x)$, $\hat{\sigma}^2(x)$ and $\hat{\pi}_k(x)$ given by Equations (A7), (A8) and (A9), respectively. By straightforward arguments similar to deriving Theorem A.3, but simpler as there are no hidden variables, we get consistency

$$\hat{\pi}_k(x) \longrightarrow \pi_k^0, \quad \hat{m}_k(x) \longrightarrow m_k(x), \quad k = 1, \ldots, M, \quad \hat{\sigma}^2(x) \longrightarrow \sigma_0^2 \quad \text{for } N \longrightarrow \infty$$

under the assumptions of Theorem A.3. Analogously replacing ζ_{tk} by Z_{tk} in Equations (7)–(9), we get

$$\tilde{\pi}_k = \frac{1}{N}\sum_{t=1}^{N} Z_{tk}, \quad k = 1, \ldots, M,$$

$$\tilde{m}_k(x) = \frac{\sum_{t=1}^{N} K_h(x-X_t)Y_t Z_{tk}}{\sum_{t=1}^{N} K_h(x-X_t)Z_{tk}}, \quad k = 1, \ldots, M,$$

$$\tilde{\sigma}^2 = \frac{1}{N}\sum_{t=1}^{N}\sum_{k=1}^{M} e_{tk}^2 Z_{tk}, \quad e_{tk} = Y_t - \tilde{m}_k(X_t).$$

We see immediately that the two estimates of m_k coincide: $\hat{m}_k(x) = \tilde{m}_k(x)$. From the consistency of those estimates, we conclude $e_{tk} = Y_t - \hat{m}_k(X_t) \to Y_t - m_k(X_t) = \sigma_0\varepsilon_{tk}$ if $Z_{tk} = 1$, and, hence, $e_{tk}^2 Z_{tk} \to \sigma_0^2\varepsilon_{tk}^2 Z_{tk}$, $k = 1, \ldots, M$, for all t. As only one of the Z_{tk}, $k = 1, \ldots, M$, is non-vanishing, $\tilde{\sigma}^2$ coincides asymptotically with an average of N i.i.d. random variables $\sigma_0^2\varepsilon_{tk}^2$ which converges to σ_0^2 as the ε_{tk} have mean 0 and variance 1. Finally, from the law of large numbers for the i.i.d. variables Z_{tk}, we have $\tilde{\pi}_k \to \pi_k$. Therefore, we have $\tilde{\pi}_k - \hat{\pi}_k(x) = o_p(1)$ and $\tilde{\sigma}^2 - \hat{\sigma}^2(x) = o_p(1)$ for all x.

To transform this heuristic argument into an exact proof that the numerical algorithm of Section 3 results in consistent estimates in case of model (1), we need some more refined asymptotics than just Theorem A.3. This will be a topic of future research.