Kauermann:

Edge Preserving Smoothing by Local Mixture Modelling

Sonderforschungsbereich 386, Paper 255 (2001)

Online unter: http://epub.ub.uni-muenchen.de/

Edge Preserving Smoothing by Local Mixture Modelling

Göran Kauermann

University of Glasgow *

14th July 2001

Abstract

Smooth models became more and more popular over the last couple of years. Standard smoothing methods however can not cope with discontinuities in a function or its first derivative. In particular, this implies that structural changes in data may be hidden in smooth estimates. Recently, Chu, Glad, Godtliebsen & Marron (1998) suggest local M estimation as edge preserving smoother. The basic idea behind local M estimation is that observations beyond a jump are considered as outliers and down-weighted or neglected in the estimation. We pursue a different, but related idea here and treat observations beyond a jump as tracing from a different population which differs from the current one by a shift in the mean. This means we impose locally a mixture model where mixing takes place due to different mean values. For fitting we apply a local version of the EM algorithm. The advantage of our approach shows in its general formulation. In particular, it easily extends to non Gaussian data. The procedure is applied in two examples, the first concerning the analysis of structural changes in the duration of unemployment, the second focusing on disease mapping.

KEYWORDS: Disease Mapping, Edge Preserving Smoothing, EM Algorithm, Generalized Mixed Models, Nonparametric Maximum Likelihood Estimation, Smoothing.

*Department of Statistics & Robertson Centre, Boyd Orr Building, University of Glasgow, Glasgow G12 8QQ

1 Introduction

In recent years a considerable number of papers dealt with models where a general smooth structure is disturbed by one or several change points or discontinuities. For statisticians a major focus of research has thereby been on testing the existence and location of change points in smooth models, see e.g. Müller (1992), Hall & Titterington (1992) or Müller & Stadtmüller (1999). Beside these inferential considerations, the fitting of models with edges and change points has been considered by several authors, see e.g. McDonald & Owen (1986) or Wu & Chu (1993). Recently, Foxall (2000) and Aitkin & Fox (2000) discuss the use of Artificial Neural Networks for the estimation of discontinuous functions. In image analysis a frequently used edge preserving smoother is the sigma filter, see e.g. Lee (1983) or Godtliebsen (1991). Chu, Glad, Godtliebsen & Marron (1998) enlighten the sigma filter by borrowing results from robust statistics and suggest local M estimation in order to stabilise the performance of the sigma filter. Further discussions of local M estimation can also be found in Winkler, Aurich, Hahn, Martin & Rodenacker (2000).

The general method running behind local M estimation is a local robust estimate which allows for outliers from a general smooth structure. Basically this means, when fitting the value of a function on one side of a jump, observations beyond the jump are treated as outliers and hence they are neglected or down-weighted in the local mean estimate. The approach presented here is based on a similar idea. We consider observations beyond a jump as tracing from a different population which differs from the local one by a shift in the mean. In this respect the location of a

jump can be interpreted as cut-point between two populations and the jump height is the difference in the population means. Beside jumps in the mean function we also consider bends, i.e. jumps in the first derivative. In this case we fit locally a mixture model with different intercept and slope parameters in the different populations. Estimation in both cases is carried out by a local version of the EM algorithm.

Mixture models are generally discussed in Böhning (1999), Aitken (1999) or Titterington, Smith & Makov (1985) and references given in there. They are also known as mixture of experts as introduced in Jacobs, Jordan, Nowlan & Hinton (1991) (see also Jacobs, Peng & Tanner, 1997). A 'classical' mixture model thereby assume that observations trace from different subpopulations with different mean and/or different slope parameters. The *a priori mixing distribution* is thereby usually assumed to be the same for all observations. Rosen, Jian & Tanner (2000) recently extended this by modelling the *a priori mixing distribution* to depend parametrically on some covariates. The approach suggested here can be seen as a smooth generalisation of their work by allowing the mixing distribution as well as other parameters to depend smoothly but nonparametrically on some covariates.

Unlike most methods for edge preserving smoothing, the approach of local mixture modelling can directly be applied to non Gaussian response data simply by embedding the approach in the framework on Generalised Mixture Models. This appears as advantage, but we emphasise at this point that for Gaussian response data local mixture modelling as edge preserving fitting routine can hardly compete with available edge preserving smoothers like e.g. the sigma filter or local M esti-

mation. This is basically due to the required computational effort which arise from local applications of the EM algorithm. On the other hand the simple generalisation of the routine towards non Gaussian data makes the approach appealing and easily applicable for a variety of data situations. In the paper we consider two examples. The first shows the performance of the routine applied to discrete survival data giving the length of unemployment. As second example we consider disease mapping where smoothing takes place spatially.

The paper is organised as follows. In Section 2 we introduce local mixture modelling and discuss feature of the local EM estimation. Section 3 demonstrates the applicability of our approach in simulations and examples. Section 4 gives some extensions to estimate discontinuities in the slope, i.e. jumps in the first derivative.

2 Smooth Generalised Mixture Models

2.1 Local Mixture Modelling

The major ingredient of our modelling approach is the generalised smooth model

$$\mu = E(y|x) \quad = \quad h\{\gamma(x)\} \tag{1}$$

where $h(\cdot)$ is a known link function and $\gamma(\cdot)$ is an unknown but smooth function varying with the covariate x. For simplicity of presentation we assume that x is univariate and scaled such that it takes values between 0 and 1. The response y for given x is assumed to be distributed according to the exponential family distribution

$$f(y|\eta) \quad = \quad \exp\left[\{y\theta - k(\theta)\}/\phi\right] \tag{2}$$

where $\theta = \theta(\eta)$ is the natural parameter, $k(\cdot)$ is the cumulant generating function with $\partial k(\theta)/\partial\theta = \mu$, ϕ is the dispersion parameter and $\eta = \gamma(x)$ is the predictor. For ease of notation we restrict ourselves in the following to natural link functions $h(\cdot)$, i.e. we assume $\theta = \gamma(x)$. Smooth estimation in model (1) by means of local fitting is treated e.g. in Fan & Gijbels (1996), Fan, Farmen & Gijbels (1998) or Carroll, Wang, Simpson, Stromberg & Ruppert (1998).

We extend model (1) by incorporating unsmooth jump effects. This is done by adding a discontinuous step function $\Delta(x)$ to $\gamma(x)$, where

$$\Delta(x) \quad = \quad \sum_{k=1}^{q-1} \delta_k 1_{[t_{k-1},t_k)}(x), \tag{3}$$

with $0 = t_0 < t_1 < \ldots t_{q-1} < 1$ being the discontinuity or jump points and $1_{[t_{k-1},t_k)}(x)$ is the indicator function taking value 1 if $t_{k-1} \leq x < t_k$ and 0 otherwise. The coefficients δ_k thereby give the jump heights. Both, the jump heights as well as the jump locations are unknown and have to be estimated from the data. The resulting discontinuous model has the form

$$E(y|x, t_1, \ldots t_k) \quad = \quad h\{\gamma(x) + \Delta(x)\}. \tag{4}$$

Let $t(x) = \{1_{[t_0,t_1)}(x), \ldots, 1_{[t_{q-2},t_{q-1})}(x), 1_{[t_{q-1},1]}(x)\}^T$ be the indicator functions organised as vector and let $\xi(x) = \{\gamma(x) + \delta_1, \ldots, \gamma(x) + \delta_{q-1}, \gamma(x)\}^T$ be the vector valued smooth function resulting from $\gamma(x)$ shifted by the jumps. This allows to rewrite model (4) to

$$\mu = E\{y|x, t(x)\} \quad = \quad h\{t(x)^T \xi(x)\}. \tag{5}$$

The indicator vector $t(x)$ is unknown and we make use of a Bayesian perspective and assume that $t(x)$ is random. A natural distributional assumption is to consider $t(x)$ as multinomially distributed, i.e.

$$t(x) \quad \sim \quad M(\pi), \tag{6}$$

with $\pi(x) = \{\pi_1(x), \ldots \pi_q(x)\}$ as cell probabilities $\pi_k(x) = P\{t(x) = e_k|x\}$, where e_k is the k-th unit vector, i.e. e_k consitsts of zeros except of a 1 at the k-th position. The parameters $\pi(x)$ give the *a prior probability* for $t(x)$ which may vary smoothly with x. For simplicity of notation we frequently write π only, i.e. neglect the dependence on x. Estimation of $\pi(x)$ and $\xi(x)$ is now carried out by a local EM algorithm.

2.2 Local EM Estimation

Let $w_{0i} = K\{(x_0 - x_i)/h\}$ denote some kernel weights with $K(\cdot)$ as unimodal, symmetrical kernel function and h as bandwidth. The basic idea is to fit model (5) locally at point x_0 by assuming $\gamma(x) \approx \gamma(x_0) =: \gamma_0$ for x close to x_0 where locally here refers to the kernel weights w_{0i}. This means we fit locally the model

$$E\{y_i|x_i, t(x_i)\} \quad = \quad h\{t(x_i)^T \xi_0\}, \tag{7}$$

with $\xi_0 = (\gamma_0 + \delta_1, \ldots, \gamma_0 + \delta_{q-1}, \gamma_0)$. Marginalization over the unknown indicator vector $t(x)$ provides the marginal local likelihood function

$$l_{(x_0)}(\theta_0) \quad = \quad \sum_i w_{0i} l_i(\theta_0) \tag{8}$$

where $\theta_0 = (\xi_0^T, \pi_0^T)$ is the vector of parameters, $\pi_0 = \pi(x_0)$ and $l_i(\theta_0) = \log f(y_i; \theta_0)$ is the marginal log likelihood contribution with $f(y_i; \theta_0) = \sum_{k=1}^q \pi_{0k} f(y_i|\xi_{0k})$. Direct

maximisation of (8) is complicated since $l_i(\theta_0)$ depends in a clumsy way on the parameters ξ_0 and π_0. Instead, the EM algorithm presents itself as an alternative (see e.g.Böhning, 1999, Aitken, 1999 or Friedl & Kauermann, 2000). The major difference between the local version applied here and the standard EM is that smoothing is involved and additional smoothing weights w_{0i} occur. The EM algorithm provides a simple M step by maximising a weighted generalised linear model while the E step results by simple density multiplication. We refer to the appendix for more details and a discussion on the choice of starting values.

Posterior Mode and Posterior Mean Estimates

The local EM algorithm provides an estimate for the parameter vector $\theta_0 = (\xi_0^T, \pi_0^T)^T$. For the unknown step function in (4), this implies that $\Gamma(x_0)$ at point x_0 takes value ξ_{0k} with probability π_{0k} for $k = 1, \ldots, q$. A prediction for the value of $\Gamma(x_0)$ can now be obtained from the *posterior mean*

$$\widehat{\Gamma}_{Mean}(x_0) := \sum_{k=1}^{q} \widehat{\xi}_{0k} \widehat{\pi}_{0k|y_0}$$

or the *posterior mode* estimate

$$\widehat{\Gamma}_{Mode}(x_0) := \widehat{\xi}_{0l} \text{ with } l = \arg\max_k \{\widehat{\pi}_{0k|y_0}\}$$

where $\widehat{\pi}_{0k|y_0}$ is a plug in estimate of the *posterior probability*

$$\pi_{0k|y_0} = \frac{f(y_0|\xi_{0k})\pi_{0k}}{\sum_{l=1}^{q} f(y_0|\xi_{0l})\pi_{0l}}. \tag{9}$$

In simulations we experienced that the *posterior mode estimate* performs generally more stable and it is therefore preferred subsequently.

3 Examples

3.1 Simulation

Before applying the local EM algorithm to examples we first demonstrate its behaviour and performance in simulations. We draw data from the normal response model $y_i = \mu(x_i) + \epsilon_i$, $i = 1, \ldots, 150$, with $\mu(x_i) = \gamma(x_i) + \Delta(x_i)$ as seen from Figure 1 and ϵ_i independently $N(0, \sigma^2)$ distributed with $\sigma = 0.025$. Figure 1 demonstrates the performance of the local EM estimate in comparison with a local M estimate. For both fits we used a Gaussian kernel with bandwidth $h = 0.1$. Starting values for the EM algorithm were chosen as suggested in the appendix. The local M estimate is defined by $\widehat{\mu}(x_0) = argmin_\mu \sum_i w_{0i}\rho(y_i - \mu)$ where $\rho(t) = 1 - exp\{-1/2(t/\sigma_y)^2\}$ with σ_y set equal to σ. It appears that both estimates behave comparable. We investigate the variation of the estimates by repeating the simulation 300 times. The resulting simulation region for the estimates is shown in Figure 2. The performance is rather stable. The small jump at 0.6 is detected in more than 50 % of the simulations, the remaining jumps are also clearly found. The local M estimate behaves similar, except that it does not uncover the small jumps at 0.6. This could be corrected by choosing a smaller value for σ_y, which however makes the estimate rather unstable and jagged, so that we did not pursue this setting.

3.2 Unemployment Data

In the following example we considers unemployment data taken from the German socio economic panel. The data have been analysed previously in Kauermann &

Tutz (2001) with the focus on smooth modelling. Here we investigate possible discontinuities in the model. Participating households in the socio economic panel record on a yearly basis if and how long they have been unemployed in the last observation period (1 year). We consider households where at least one member has been unemployed and restrict the analysis to German citizens. The focus of interest is on estimating the chances to return to professional life. Let y_i denote the reported duration of unemployment by the i-th individual. We consider the discrete survival model

$$P(y_i \leq x+1 | y_i \geq x) \quad = \quad h\{\gamma(x) + \Delta(x)\}$$

with $h(\cdot)$ as logit function and $x = 1, 2, \ldots$ giving the duration of unemployment in months. The fitted curves are shown in Figure 3 separately for males and females. For comparison we also show a smooth fit based on a fitting a model like (1) to the data, i.e. setting $\Delta(x)$ to zero (smooth fitting and variance estimation is carried out using local likelihood methods, see e.g. Fan, Farmen & Gijbels, 1998 or Kauermann & Tutz, 2000). The curves plotted are calculated for bandwidth $h = 5$. The structure of the curves however remains basically unchanged for different, reasonable settings of the bandwidth.

Overall, the chance for returning to professional life is decreasing over time. The smooth structure is however disturbed at month 3 where a jump occurs in both genders. This means a relatively large number of unemployed individuals report to return to professional life exactly after 3 months of unemployment. A second smaller jump occurs after 12 months but interestingly enough only for males. Interpretation

of such bumps can be a delicate issue due to political implications. To us, three explanations appear plausible for the observed phenomena. First, the peaks can mirror calendar a effect or reporting bias, meaning that even if an individual is unemployed less than or longer than 3 months, the time of unemployment is rounded quarter-year-wise in the questionnaire of the panel. A second explanation for the 3 month peak could be that unemployed people require some time to get orientated on the job market, since fluent transition of jobs does not lead to unemployment. The time period to get orientated is thereby about three months. A final possible explanation for the 3 month peak is that the jump occurs due to seasonal workers who are often unemployed over winter months. In general, further investigation on an individual basis seems necessary to find a valid explanation. The discontinuous pattern is however concealed in a smooth fit and is discovered by discontinuous smoothing.

3.3 Spatial Smoothing and Disease Mapping

The second example considered concerns spatial smoothing in form of disease mapping. For a general overview about statistical methods in disease mapping we refer to Lawson (2001), Lawson, Biggeri, Böhning, Lesaffre, Viel & Bertollini (1999) or Böhning (1999) and references given there. Nonparametric Maximum Likelihood estimation respectively mixture modelling in the context of disease mapping was suggested by Clayton & Kaldor (1987) and further developed among others by Schlattmann & Böhning (1993). We investigate data giving the mortality rate

from oral cancer for males in Germany in the years 1986–1990. The data have been analysed before by Knorr-Held & Raßer (2000) with Bayesian methods using MCMC methods. Their focus has been on finding clusters of districts with the same risk. Figure 4 shows the observed standard mortality rates (SMR) for 544 German districts. The SMR is thereby defined by the observed death counts y_i divided by the expected cases e_i, where the expected cases are calculated on the basis of an external population stratified for different ages (see Clayton & Hills, 1993 for more details and Knorr-Held & Raßer, 2000 for an exact description of the strata used in the considered data). There appear a considerable amount of variation and no smooth structure can be seen from the raw SMR. Fitting a local mixture model yields the results shown in Figure 5. The local weights are thereby constructed from the distances t_{ij} between districts i and j. As distance we thereby define the smallest number of borders one has to pass when going from district i to j. Hence, direct neighbours have distance $t_{ij} = 1$ and we define $t_{ii} = 0$. Weights used to obtain the fits in 5 were chosen as $w_{ij} = (8 - t_{ij})/8$ for $t_{ij} \leq 3$ and $w_{ij} = 0$ for $t_{ij} > 3$. We experienced other weights with larger neighbourhoods, i.e. larger bandwidths as well, but the results were rather similar. The local mixed model (4) used in this example has the form

$$E\{y_i|t(x_i)\} \quad = \quad h\{v_i + \gamma(i) + \Delta(i)\}$$

where $h(\cdot) = \exp(\cdot)$, $v_i = \log(e_i)$ is serving as given offset and $\gamma(i)$ is a spatially smooth function over the districts. In Figure 5 we plotted the relative risk $RR_i = \exp\{\widehat{\gamma}(i) + \widehat{\Delta}(i)\}$, where values $RR_i > 1$ indicate that the fitted mortality is above

the expected and vice versa. The starting values were chosen as suggested in the Appendix. Sensitivity on the starting values was checked, but didn't prove to be apparent.

In general, regions with higher risk are the north-eastern part of Germany as well as the south-western part. A similar finding is reported in Knorr-Held & Raßer (2000) who name tobacco smoking and alcohol consumption as the potential risk factors for oral cancer in the north-eastern part (see also Becker & Wahrendorf, 1997). For the south-western part the high risk extends over to France (see e.g. Blot, Devesa, McLaughlin & Fraumeni, 1994). Beside the general smooth structure, some districts distinctly distinguish from their neighbours. These are in particular larger cities, as e.g. Hamburg (HH), Kiel (KI) and the western part of Berlin (WB). In the western part of Germany the cities Krefeld (KR), Düsseldorf (D) and Cologne (K for Köln) and east of them the industrial area along the River Ruhr (Ruhrgebiet) also show a higher risk. The local mixture model can cope with these discontinuities while fitting generally a smooth surface.

Simulation

It appears worthwhile at this point to investigate the performance of the estimate used in this example above. In particular we investigate, if the discontinuities observed for some of the cities occur only due to a variation in the expected mortality rate v_i. We simulate data from the model $E(y_i|x_i, v_i) = h\{v_i + \gamma(x_i) + \Delta(x_i)\}$, $i = 1, \ldots, 150$ and x_i equally spaced on $[0, 1]$. The expected log moralities v_i serve

as given offsets in the estimation and they are drawn once from a contaminated normal distribution as seen from Figure 6. The fitted relative risk $\exp\{\widehat{\gamma}(x_i) + \widehat{\Delta}(x_i)\}$ and the true curve are shown in Figure 6. Clearly, there is no indication that the extreme values of v_i disturb the reconstruction of the discontinuous structure. This also shows in the larger simulation based on 150 simulations also shown in Figure 6.

4 Extensions and Discussion

4.1 Jumps in the Derivative

Local Linear Mixture Modelling

In the same fashion as local constant smoothing can be extended to local linear smoothing (see e.g. Fan & Gijbels, 1996), local EM estimation can readily be extended to local linear EM estimation. The basic idea behind local linear smoothing is that $\gamma(x)$ is expanded linearly about x_0, i.e. $\gamma(x) \approx \gamma(x_0) + \gamma'(x_0)(x_0 - x)$. Transferring this to model (4) and (7) yields the local linear mixture model

$$E\{y|x_i, t(x_i)\} \quad = \quad h\left\{t(x_i)\xi_0 + (x_i - x_0)\gamma_0'\right\} \tag{10}$$

with $\gamma_0' = \gamma'(x_0)$. Model (10) can be seen as mixture model with random intercept and parametric slope parameter γ_0' corresponding to the covariate $x_i - x_0$.

The next step is now to extend (10) to accommodate jumps in the first derivative. Assume therefore that the smooth function $\gamma(x)$ in (1) is disturbed by jumps and/or bends. We model this by adding two discontinuous step functions to both, $\gamma(x)$ and

its first derivative. This yields the model

$$E\{y|x,t(x)\} \quad = \quad h\{\gamma(x) + \Delta_1(x) + \int_0^x \Delta_2(u)du\} \tag{11}$$

where $\Delta_l(x)$ for $l = 1, 2$ are step functions as defined in (3). We rewrite model (11) to

$$E\{y|x,t(x)\} \quad = \quad h\{t(x)^T\xi(x) + \int_0^x t(u)^T\delta_2 du\}$$

with $t(x)$ and $\xi(x)$ defined as above and $\delta_2 = (\delta_{21}, \ldots, \delta_{2q-1}, 0)^T$ as vector containing the jump heights in $\Delta_2(x)$. Local linear mixture modelling can now be used to fit (11). We expand $\Gamma(x) = \gamma(x) + \Delta_1(x) + \int_0^x \Delta_2(u)du$ about x_0, where $x_0 \in (0,1)\backslash\{t_1, \ldots, t_q\}$ which yields $\Gamma(x) \approx t(x_0)\xi_0 + (x - x_0)\{\gamma'(x_0) + t(x)^T\delta_2\}$. This provides the local linear mixture model

$$E\{y_i|x_i,t(x_i)\} \quad = \quad h\{t(x_i)^T\xi_0 + (x_i - x_0)t(x_i)^T\xi_0'\} \tag{12}$$

where $\xi_0 = \xi(x_0)$ and $\xi_0' = \{\gamma'(x_0) + \delta_{21}, \ldots, \gamma'(x_0) + \delta_{2q-1}, \gamma'(x_0)\}^T$. In contrast to model (10), in (12) both, the intercept as well as the slope parameter are tracing from mixtures. Following a Bayesian framework (6) we model that $(t(x)^T\xi_0, t(x)^T\xi_0')$ to come from a discrete valued distribution with masspoints $\{(\xi_{01}, \xi_{01}'), \ldots, (\xi_{0q}, \xi_{0q}')\}$ and masses $(\pi_{01}, \ldots, \pi_{0q})$.

Fitting of (12) can again be carried out with a local EM algorithm. The only practical difficulty occurring is the choice of appropriate starting values. We give some guidelines in the Appendix.

Simulation

We demonstrate the procedure with a simulation study. We draw data from a normal distribution model as seen from Figure 7 (left plot). We set bandwidth $h = 0.05$ and use $q = 5$ with the starting values as suggested in the Appendix. The local linear mixture model clearly detects the bend at 0.3. Also, the first bend at about 0.2 is oversmoothed. The smooth structure of the function between 0.3 and 0.8 is fitted smoothly and the jump at 0.8 is detected. In Figure 7 (right plot) we also show the posterior mode estimate for the estimated first derivative, i.e. $\widehat{\xi}'_{0l}$ with $l = \arg\max\{\widehat{\pi}_{0k|y_0}\}$ and $\pi_{0k|y_0}$ as defined in (9) but for the local linear mixture model. The estimate clearly mirrors the jump at 0.3 in the first derivative while the quadratic structure between 0.5 and 0.8 is nicely reproduced in a linear fit for the first derivative. We run a simulation study to investigate the overall performance. The results are shown in 8. The conclusion remains unchanged in that the local linear mixture modelling detects the jumps in the function and/or in the derivative except of the fist obtuse bend at 0.2.

4.2 Discussion

In the paper we showed how mixture models and smoothing models can be combined to cope for discontinuous effects in a function. The procedure is in particular applicable to non normal response data, as it allows for simple estimation using the EM algorithm. The examples demonstrate the practical impact one achieves when allowing a generally smooth structure to have some breakpoints.

Inferential arguments using local mixture models in order to test for jumps are somewhat weak still. This is mainly due to non standard asymptotic behaviour of a likelihood ratio statistics in mixture models with different numbers of components (see e.g. Aitken, 1999). An ad hoc solution to circumvent the technical difficulties is to rely on bootstrap procedures. This however will demand for further computational effort. To us, there seems room and need for further research in this area.

Acknowledgements

The first parts of this work were prepared while the author was member of the Collaborative Research Centre (Sonderforschungsbereich) 386 at the Ludwig-Maximilians-Universität München (LMU), funded by the Deutsche Forschungsgemeinschaft. Support in various aspects during this period is gratefully acknowledged. The author would also like to thank Günter Rasser, LMU München, for his help and enlightening discussions related to the oral cancer data.

A Algorithmic Details

Local EM algorithm

The following algorithm results directly from the parametric EM algorithm as applied in mixed models (see e.g. Böhning, 1999, Aitken, 1999 or Friedl & Kauermann, 2000). Let $f(y_i, t_i; \theta)$ denote the joint density of y_i and $t_i := t(x_i)$, where the dependence on x is notationally neglected. Note that $\log f(y_i, t_i; \theta) = \log g(t_i|y_i; \theta) + \log f(y_i; \theta)$ where $g(\cdot|\cdot)$ is the conditional density of t_i given y_i. Taking expectation

on both sides of the equation and using kernel weights provides the relation

$$\sum_{i=1}^{n} w_{0i} Q_i(\theta, \theta^*) \;=\; \sum_{i=1}^{n} w_{0i} H_i(\theta, \theta^*) + l_{(x_0)}(\theta)$$

where $Q_i(\theta, \theta^*) = \int \log f(y_i, t_i; \theta) dG(t_i|y_i; \theta^*)$ and $H_i(\theta, \theta^*) = \int \log g(t_i|y_i; \theta) dG(t_i|y_i; \theta^*)$. Using arguments as applied in the standard EM setting one finds $H_i(\theta, \theta^*) \geq H_i(\theta^*, \theta^*)$ so that increasing $Q_{(x_0)}(\theta, \theta^*) = \sum_{i=1}^{n} w_{0i} Q_i(\theta, \theta^*)$ provides an increase of the marginal likelihood. Inserting the multinomial distribution for t_i allows now to rewrite $Q_{(x_0)}(\theta, \theta^*)$ to $Q_{(x_0)}(\theta, \theta^*) = \sum_{i=1}^{n} \sum_{k=1}^{q} w_{0i} \omega_{ik}(\theta^*) \{\log f(y_i|\xi_{0k}) + \log \pi_k\}$ where $\omega_{ik}(\theta^*) = f(y_i|\xi_k^*)\pi_k^* / \sum_l f(y_i|\xi_l^*)\pi_l^*$ with $\theta^* = ({\xi^*}^T, {\pi^*}^T)^T$. The E step of the EM algorithm is now pursued by the calculation of the weights $\omega_{ik}(\theta^{(t)})$, while the M step follows by finding $\theta^{(t+1)}$ which maximises $Q_{(x_0)}(\theta^{(t+1)}, \theta^{(t)}) = \max_\theta Q_{(x_0)}(\theta, \theta^{(t)})$. It appears that $\xi^{(t+1)}$ is updated by solving the weighted score function

$$0 \;=\; \sum_{i=1}^{N} \sum_{k=1}^{q} w_{0i} \omega_{ik}(\theta^{(t)}) s_i(\xi_k^{(t+1)}), \tag{13}$$

where $s(\eta) = \partial \log f(y_i|\eta)/\partial \eta$ is the standard score as found in generalised linear models. Hence, the solution of (13) can be calculated using standard software for generalised linear models by simply incorporating weights. Finally we update π by $\pi_k^{(t+1)} = \sum_{i=1}^{N} \sum_{k=1}^{q} w_{0i} \omega_{ik}(\theta^{(t)}) / \sum_{i=1}^{N} w_{0i}$. In practice the local EM algorithm is not required to be carried out for all observed values of x, but instead it can be calculated over a grid of points spanning the observed x values

Starting Values

The application of the EM algorithm requires the specification of (i) a number of masspoints and (ii) starting values for the parameters ξ_0 and π_0. Since local esti-

mation is pursued, only a fraction of the data is used for fitting. This implies, that not all jumps occurring in the data show in the local kernel window. Hence locally q can be less than the total number of jumps. In our examples and simulations we experienced that choosing $q = 3$ (or $q = 5$) provides an appropriate performance. This holds as long as the jumps are not clustered and well separated. Secondly, starting values $\pi_0^{(0)}$ and $\xi_0^{(0)}$ for π_0 and ξ_0 should be chosen such that the EM algorithm converges desirably quickly to the (global) maximum of (8). We suggest the starting values based on the following considerations. Observations beyond a jump are considered as tracing from a different population which differs from the local one by a shift δ in the mean, say. For $q = 3$, for instance, we therefore suggest the starting value $\xi_0^{(0)} = (\widehat{\xi}_0^{*(0)} - \delta, \widehat{\xi}_0^{*(0)}, \widehat{\xi}_0^{*(0)} + \delta)$ with $\delta > 0$ as shift in the mean and $\widehat{\xi}_0^{*(0)}$ as a rough estimate for the local mean of the data. For instance if estimation at point $x_0 = x_i$ is carried out one may use $\widehat{\xi}_0^{*(0)} = h^{-1}(y_i)$ and take shift δ as a multiple of the standard deviation of y, e.g. $\delta = 3\sigma$. As starting values for the probability π we suggest to give the initial mean estimate $\widehat{\xi}_0^{*(0)}$ a higher probability than the shifted values, e.g. for $q = 3$ we set $\pi_0^{(0)} = (\varepsilon^{(0)}, 1 - 2\varepsilon^{(0)}, \varepsilon^{(0)})$ for some small, positive $\varepsilon^{(0)}$ (with $0 < \varepsilon^{(0)} < 1/2$), e.g. $\epsilon^{(0)} = 0.1$. Hence the setting of the starting values mirror the situation where $t(x_i)\xi_0$ takes the general mean $\widehat{\xi}_0^{*(0)}$ with high probability $1 - 2\varepsilon^{(0)}$ while the extreme values (outlieres) $\widehat{\xi}_0^{*(0)} \pm \delta$ occur with small probability $\varepsilon^{(0)}$. In general it is advisable to fit the model also with different starting values to investigate their impact on the fit.

Starting Values for Local Linear Mixture Model

For local linear mixture as in (12) we suggest a similar choice. Choosing $q = 5$, for instance, and set

$$\begin{aligned} \xi_0^{(0)} &= (\widehat{\xi}_0^{*(0)} - \delta, \widehat{\xi}_0^{*(0)} - \delta, \widehat{\xi}_0^{*(0)}, \widehat{\xi}_0^{*(0)} + \delta, \widehat{\xi}_0^{*(0)} + \delta) \\ \xi_0^{\prime(0)} &= (\widehat{\xi}_0^{\prime *(0)} - \delta', \widehat{\xi}_0^{\prime *(0)} + \delta', \widehat{\xi}_0^{*(0)}, \widehat{\xi}_0^{\prime *(0)} - \delta', \widehat{\xi}_0^{\prime *(0)} + \delta') \end{aligned}$$

where $(\widehat{\xi}_0^{*(0)}, \widehat{\xi}_0^{\prime *(0)})$ result from a local linear fit of the data. The coefficients δ and δ' give the shift in the starting values for the intercept and first derivative. Moreover, as starting value for π_0 we suggest $\pi_0 = (\varepsilon^{(0)}, \varepsilon^{(0)}, 1 - 4\varepsilon^{(0)}, \varepsilon^{(0)}, \varepsilon^{(0)})$ for some small $\varepsilon^{(0)}$ with $(0 < \varepsilon^{(0)} < 1/4)$. The setting mirrors the situation where we give large probability to a local linear fit and low probability to shifts of this fit.

Estimation of the Dispersion Parameter

For normal response models the dispersion parameter ϕ has to be estimated in order to apply the EM algorithm. Assuming independent observations variance estimation can be carried out by the difference based estimator as suggested in Gasser, Sroka & Jennen-Steinmetz (1986). Let $\epsilon_i = a_i y_{i-1} + b_i y_{i+1} - y_i$ with $a_i = (x_{i+1} - x_i)/(x_{i+1} - x_{i-1})$ and $b_i = (x_i - x_{i-1})/(x_{i+1} - x_{i-1})$, for $i = 2, \ldots n-1$ and ordered x values $x_i < x_{i+1}$. The variance is then estimated by a weigted mean of ϵ_i^2 with weights $1/(a_i^2 + b_i^2 + 1)$. In the situation where the underlying function has jumps, it is preferable to substitute the weighted mean by a robust weighted mean, e.g. by a trimmed mean, in order to avoid biased variance estimates.

References

Aitken, M. (1999). A general maximum likelihood analysis of variance components in generalized linear models. *Biometrics* **55**, 218–234.

Aitkin, M. and Fox, R. (2000). Statistical modelling of artificial neural networks using the multi-layr perceprton. (manuscript).

Becker, N. and Wahrendorf, J. (1997). *Atlas of Cancer Mortality in the Federal Republic of Germany 1981–1990*. Berlin: Springer Verlag.

Blot, W. J., Devesa, S., McLaughlin, J., and Fraumeni, J. F. (1994). Oral and pharyngeal cancers. In R. Doll, K. Fraumeni, & C. Muir (Eds.), *Cancer Surveys: Trends in Cancer Incidence and Mortality*, Volume 19/20, pp. 23–42. New York: Cold Spring Harbor Laboratory Press.

Böhning, D. (1999). *Computer-Assisted Analysis of Mixtures and Applications. Meta-Analysis, Disease Mapping and others*. Boca Raton: Chapman & Hall / CRC.

Carroll, R. J., Wang, S., Simpson, D. G., Stromberg, A. J., and Ruppert, D. (1998). The sandwich (robust covariance matrix) estimator. Technical Report, Preprint.

Chu, C. K., Glad, I. K., Godtliebsen, F., and Marron, J. S. (1998). Edge-preserving smoothers for image processing (with discussion). *J. Amer. Statist. Assoc.* **93**, 526–541.

Clayton, D. and Hills, M. (1993). *Statistical Models in Epidemiology*. Oxford University Press.

Clayton, D. and Kaldor, J. (1987). Empirical bayes estimates of age-standardized relative risks for use in disease mapping. *Biometrics* **43**, 671–681.

Fan, J., Farmen, M., and Gijbels, I. (1998). Local maximum likelihood estimation and inference. *Journal of the Royal Statistical Society, Series B* **B 60**, 591–608.

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

Foxall, R. (2000). *Statistical Modelling of Artificial Neural Networks*. Ph. D. thesis, University of Newcastle-upon-Tyne.

Friedl, H. and Kauermann, G. (2000). Standard errors for EM estimates in variance component models. *Biometrics* **56**, 761–767.

Gasser, T., Sroka, L., and Jennen-Steinmetz, C. (1986). Residual variance and residual pattern in nonlinear regression. *Biometrika* **73**, 625–633.

Godtliebsen, F. abd Spjøtvoll, E. (1991). Comparison of statistical methods in MR imaging. *International Journal of Imaging Systems and Technology* **3**, 33–39.

Hall, P. and Titterington, D. M. (1992). Edge-preserving and peak-preserving smoothing. *Technometrics* **34**, 429–440.

Jacobs, R., Jordan, M., Nowlan, S., and Hinton, G. (1991). Adaptive mixtures of local experts. *Neural Computing* **10**, 79–87.

Jacobs, R., Peng, F., and Tanner, M. (1997). A bayesian approach to model selection in hierarchical micture-of-expert architecture. *Neural Networks* **10**, 231–241.

Kauermann, G. and Tutz, G. (2000). Local likelihood estimation in varying-coefficient models including additive bias correction. *Journal of Nonparametric Statistics* **12**, 343–371.

Kauermann, G. and Tutz, G. (2001). Testing generalized linear and semiparametric models against smooth alternatives. *Journal of the Royal Statistical Society, Series B* **63**, 147 – 166.

Knorr-Held, L. and Raßer, G. (2000). Bayesian detection of clusters and discontinuities in disease maps. *Biometrics* **56**, 13–21.

Lawson, A. (2001). *Statistical Methods in Spatial Epidemiology*. Chichester, UK: Wiley.

Lawson, A., Biggeri, A., Böhning, D., Lesaffre, E., Viel, J.-F., and Bertollini, R. E. (1999). *Disease Mapping and Risk Assessment for Public Health*. Chichester, UK: Wiley.

Lee, J. S. (1983). Digital image smoothing and the sigma filter. *Computer Vision, Graphics and Image Processing* **24**, 255–269.

McDonald, J. A. and Owen, A. B. (1986). Smoothing with split linear fits. *Technometrics* **28**, 195–208.

Müller, H.-G. (1992). Change-points in nonparametric rgeression analysis. *Ann. Statist.* **20**, 737–761.

Müller, H.-G. and Stadtmüller, U. (1999). Discontinuous versus smooth regression. *Ann. Statist.* **27**, 299–337.

Rosen, O., Jian, W., and Tanner, M. (2000). Mixture of marginal models. *Biometrika* **87**, 391–404.

Schlattmann, P. and Böhning, D. (1993). Mixture models and disease mapping. *Statistics in Medicine* **12**, 1943–1950.

Titterington, D. M., Smith, A. F. M., and Makov, U. E. (1985). *Statistical Analysis of Finite Mixture Distributions*. Wiley, New York.

Winkler, G., Aurich, V., Hahn, K., Martin, A., and Rodenacker, K. (2000). Noise reduction in images: Some recent edge-preserving methods. *Journal of Pattern Recognition and Image Analysis* **9**, 749–766.

Wu, J. S. and Chu, C. K. (1993). Nonparametric function estimation and bandwidth selection for discontinuous regression functions. *Statistica Sinica* **3**, 557–576.

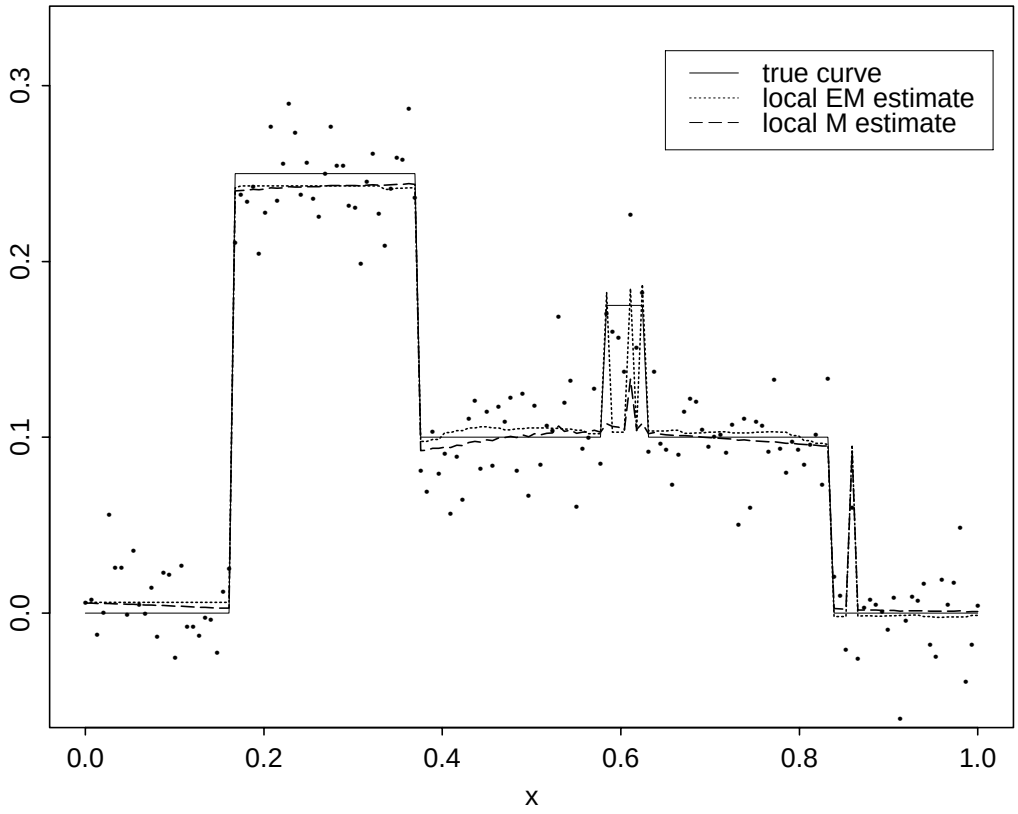

Figure 1: Local EM Posterior Mode estimate and local M estimate

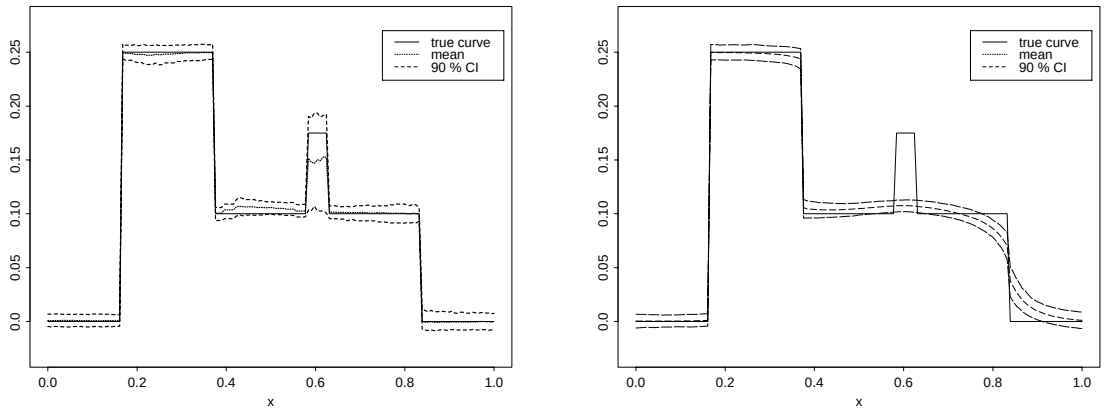

Figure 2: Simulations intervals (90 %) and simulation median for local EM estimates (left plot) and local M estimation (right plot) based on 300 simulations.

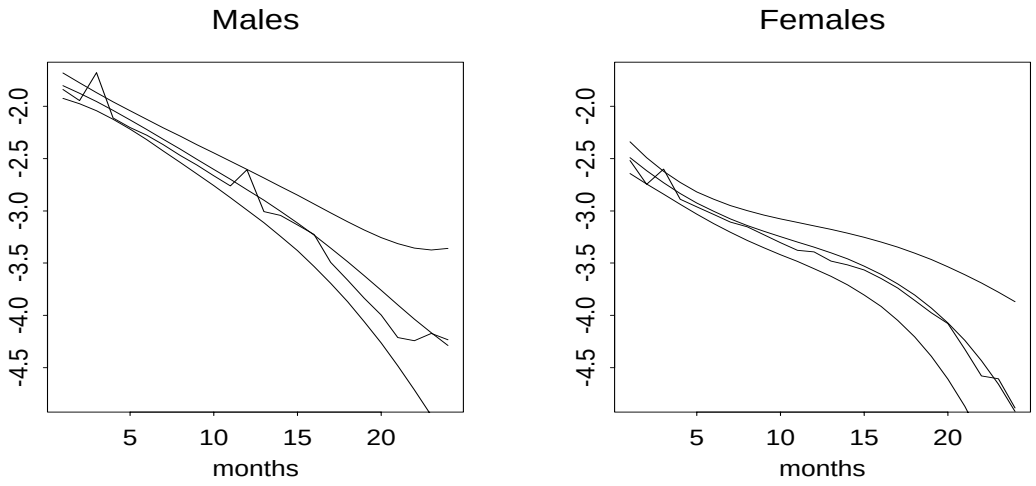

Figure 3: Estimated log odds of the Probability to return to professional life.

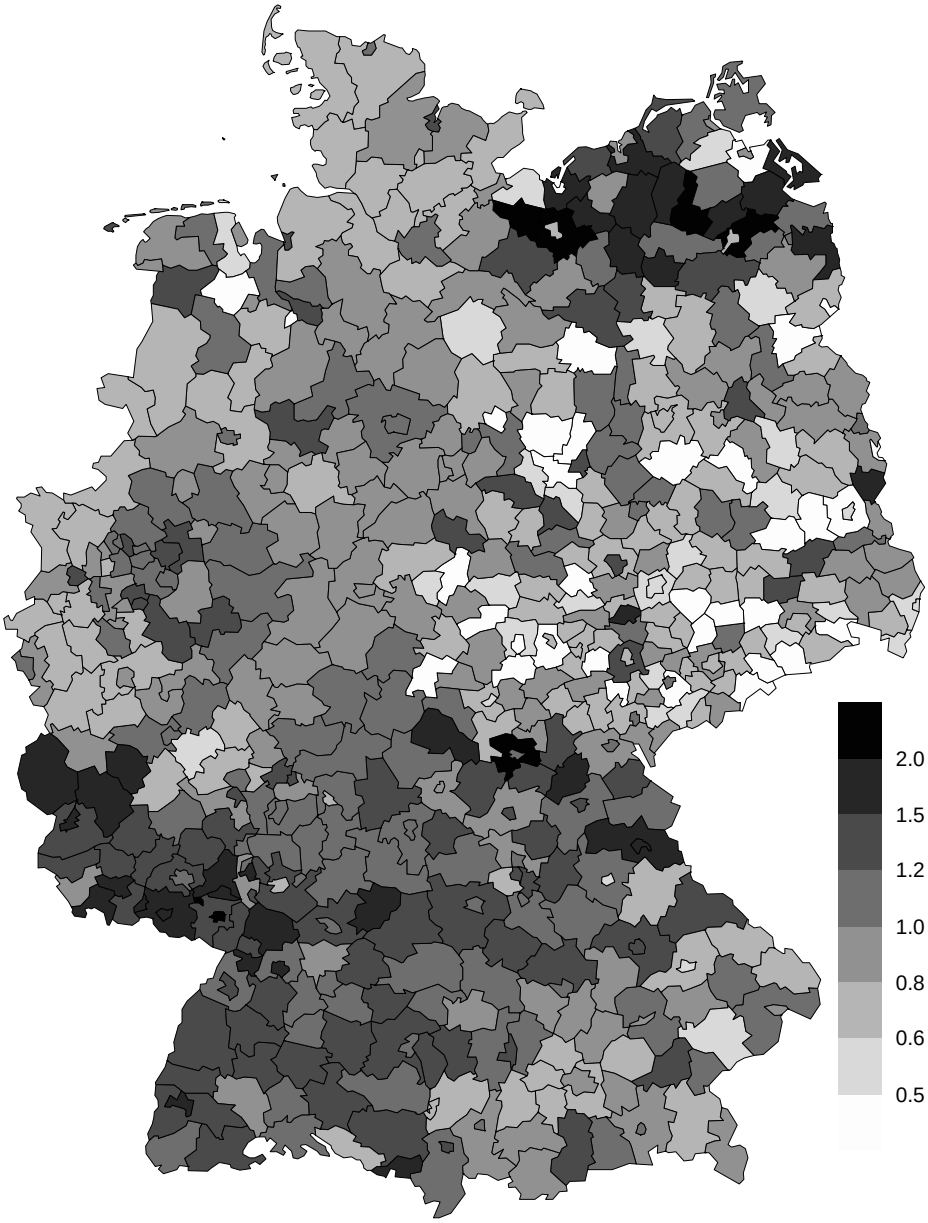

Figure 4: Standardized Mortality Rate for German districts

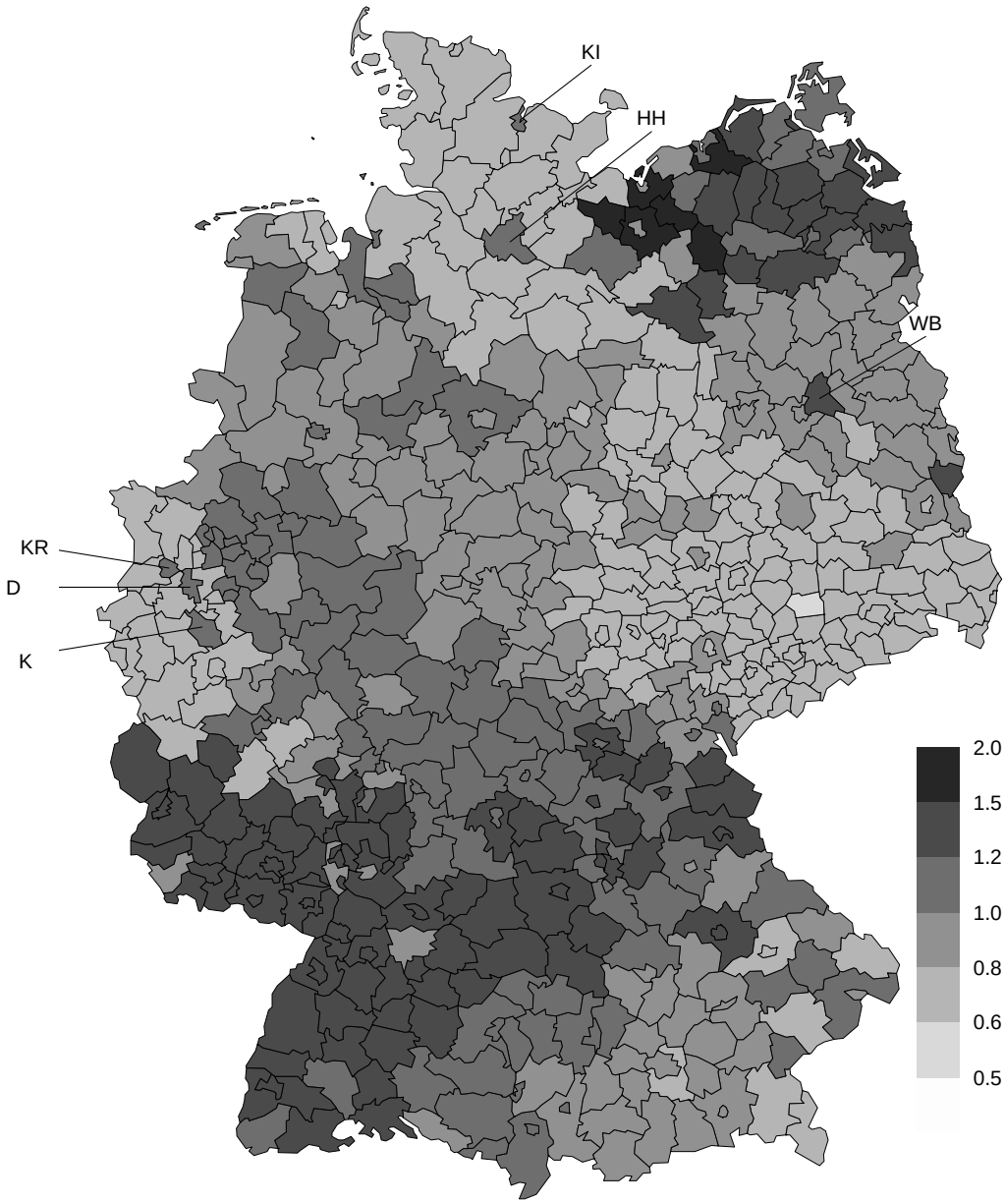

Figure 5: Local EM estimates for relative risk for German districts.

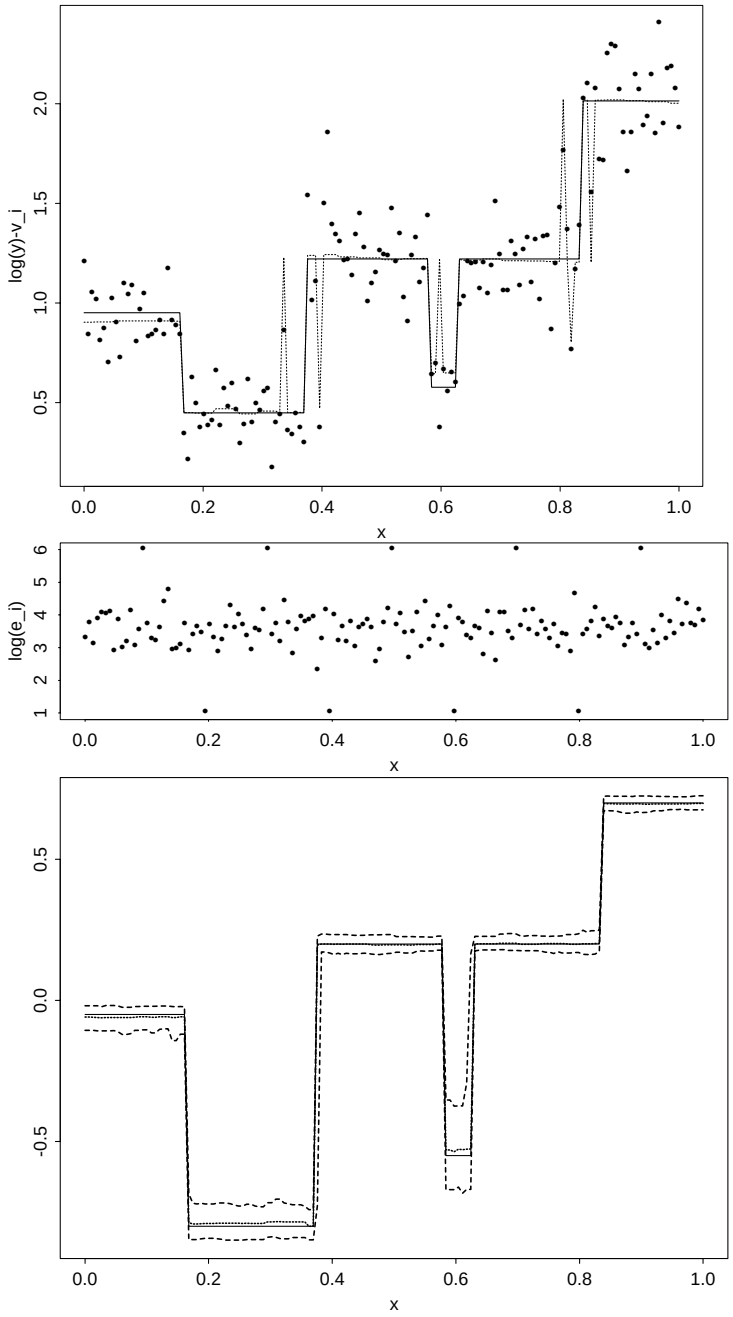

Figure 6: Simulated and fitted Poisson data from model $E(y_i|x_i, v_i) = h\{v_i + \gamma(x_i) + \Delta(x_i)\}$ (upper plot) with v_i drawn from a contaminated normal model (middle plot). Bottom plot gives simulation confidence intervals based on 150 simulations.

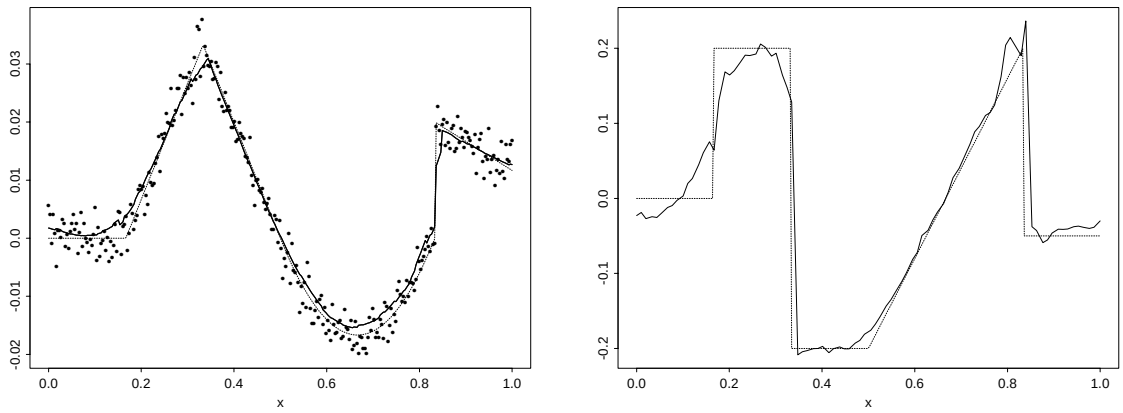

Figure 7: Posterior mode estimate of mean function (left plot) and its first derivative (right plot) based on local linear mixture modelling. Dashed curves gives true function

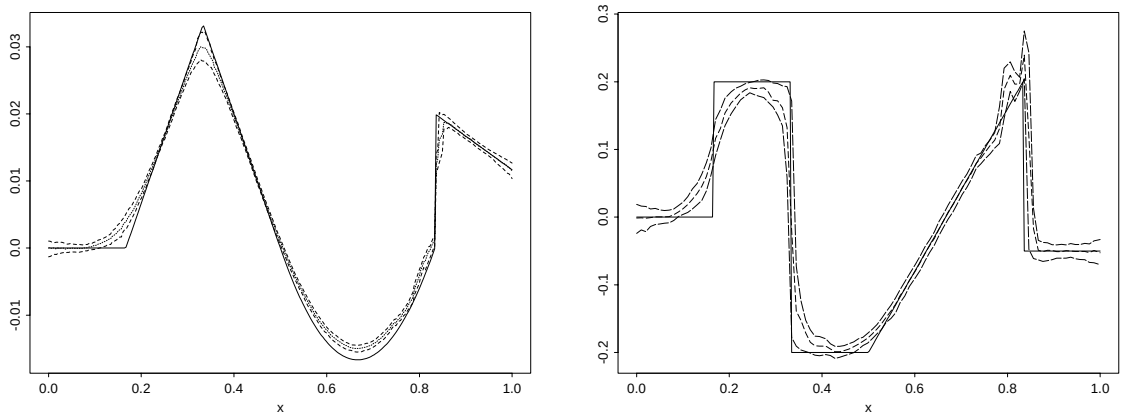

Figure 8: Simulation confidence intervals (90%) and simulation mean for posterior mode estimate of mean function (left plot) and its first derivative (right plot) based on local linear mixture modelling. Dashed curves gives true function