Kauermann, Küchenhoff:

Modelling Data from Inside of Earth: Local Smoothing of Mean and Dispersion Structure in Deep Drill Data

Sonderforschungsbereich 386, Paper 269 (2002)

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

Modelling Data from Inside of Earth: Local Smoothing of Mean and Dispersion Structure in Deep Drill Data

Göran Kauermann*
University of Glasgow

Helmut Küchenhoff†
Ludwig-Maximilians-Universtität München

13th December 2001

Abstract

In this paper we analyse data originating from the German Deep Drill Program. We model the amount of 'cataclastic rocks' in a series of measurements taken from deep drill samples ranging from 1000 up to 5000 meters depth. The measurements thereby describe the amount of strongly deformed rock particles and serve as indicator for the occurrence of cataclastic shear zones, which are easily speaking areas of severely 'ground' stones due to movements of different layers in the earth crust. The data represent a 'depth series' as analogue to a 'time series', with mean, dispersion and correlation structure varying in depth. The general smooth structure is thereby disturbed by peaks and outliers so that robust procedures have to be applied for estimation. In terms of statistical modelling technology we have to tackle three different peculiarities of the data simultaneously, that is estimation of the correlation structure, local bandwidth selection and robust smoothing. To do so, existing routines are adapted and combined in new 'two stage' estimation procedures.

KEYWORDS: Local Bandwidth, Local Smoothing, Robust Smoothing, Smoothing Correlated Data.

*Department of Statistics & Robertson Centre, Boyd Orr Building, University of Glasgow, Glasgow G11 8QQ, UK

†Department für Statistik, Akademiestraße 1, 80799 München, Germany

1 Introduction

1.1 The data

During the course of the German Continental Deep Drilling Program (Kontinentales Tiefbohrprogramm der Bundesrepublik Deutschland, KTB) two scientific boreholes were drilled into the crystalline crust near the village of Windischeschenbach (Bavaria, Germany). One of the holes was drilled down to 4 km depth and the second down to 9.1 km depth. The scientific aims of the project are manifold including e.g. the evaluation of geophysical structures and phenomena, investigation of the thermal structure of the continental crust, in-situ investigation of rock fluids and their contribution to formation of ore deposits, and elucidation of the structure and evolution of the Earth's crust and many more.

The KTB project was one of the most extensive and expensive research program in geosciences ever undertaken in Germany. Data coming from this project have been extensively analysed before, mainly in the geophysical literature. The budget amounted to approximately 380,000,000 EURO from 1982 until the completion of drilling activities in 1995. For further description of the project and results we refer to Emmermann & Lauterjung (1997). Updated information about the project as well as data access is available on the project web-page http://icdp.gfz-potsdam.de/html/ktb/.

Technically, investigations in the KTB field laboratory on-site revealed about 68 variables of lithological components, petrophysical properties and geochemical composition from drill cuttings. Drill cutting samples are a mixture of material from the drilled sections comprising a segment of several meters depth. They are transported by a special liquid ("drill mud") to the surface during circulation. The

drill hole conditions vary widely with depth. Bore-hole diameter (caliper), hydraulic pressure, strength of the rocks, and drilling rate can influence the drill cuttings sample until it reaches the surface.

One current issue of research is the frequent occurrence of cataclastic shear zones which are important structures in the upper and middle crust of the Earth. In such zones the minerals are severely deformed due to the movement of greater blocks and layers. They can represent seismic reflectors, zones of high electrical conductivity, fluid pathways, and in consequence, exhibit secondary mineralisation. The structure and composition of cataclastic shear zones were investigated by laboratory investigations on drill core samples e.g. by Zulauf, Palm, Petschick & Spies (1999). An important variable addressing cataclastic shear zones is the 'amount of cataclastic rocks' (CATR) in the drill cuttings, observed by microscopic investigations of the samples. Simply speaking this variable measures the proportion of strongly deformed rock particles in the sample at a certain depth. In Winter, Adelhardt, Jerak & Küchenhoff (2002) the relationship between this variable and other geophysical and geochemical variables has been analysed with respect to two different lithologies: gneiss and metabasite. In this paper we analyse CATR in the depths between 1000 and 5000 meters. The data thereby form a non equidistant "depth series" consisting of 2748 observations with depth differences between two measurements ranging from 0.5 to 6 meters. The main focus of our analysis is how the mean structure of CATR changes with depth. In particular we are interested in possible relations to lithology, technical operations, etc. Interesting features in the data are step-like variations and peaks which may yield relevant information.

1.2 Modelling Approach

The data considered have a number of peculiarities which demand for careful consideration. First, the data are not free of outliers and moreover the mean structure, even though mainly smooth, does contain peaks or spikes. This makes smooth estimation complicated, as first observations corresponding to spikes or outliers have to be classified. Secondly, observations are correlated, which has to be considered in a data driven bandwidth selection for the smooth fit being applied. Finally, the complexity of the mean structure changes with depth so that a global smoothing parameter appears not adequate and local bandwidth selection has to be employed. Even though each of these three problems is treated in different depth in the literature it is not straightforward to apply the suggested routines in a data situation where all three peculiarities occur at the same time. We will adapt some of the available routines as well as develop new ones to cope with the data constellation in our example. The emphasis is on practicability and rigorous theoretical justification is not given, even though simulations are shown which support our approach.

Local smoothing is a widely developed area and treated extensively in the literature in the last couple of years. We refer exemplary to Fan & Gijbels (1996) or Simonoff (1996). Local robust smoothing has been recently used in Chu, Glad, Godtliebsen & Marron (1998) as edge preserving smoother. Local approaches for smoothing correlated data have been suggested by Chiu (1989), Altman (1990) or Hart (1991). An general overview is given by Opsomer, Wang & Yang (2001). The central problem occurring for correlated errors is that data driven bandwidth selection requires the knowledge or estimation of the correlation structure. This can lead to substantial problems in practice since nonparametric fitting of the mean structure induce strong correlation in the fitted residuals even if the true residuals

are independent. We refer to Opsomer, Wang & Yang (2001) for a nice illustration of this problem. On top of that, methods based on fitted residuals or frequency periodograms are hardly robust enough to cope with outliers in the data or jumps and peaks in the mean function. We will therefore go a different route by employing a two step procedure. First we estimate the correlation structure directly from the empirical correlations of consecutive data pairs y_i, y_{i+1}. In particular, we consider small clusters of data and calculate their empirical pairwise correlation. In a second step a robust local smoother is applied to diminish the variability in the raw pairwise correlations and down-weight outliers. The idea is discussed in more detail in a forthcoming PhD thesis by Christina Yap from the Department of Statistics, University of Glasgow. In general, the approach shows some similarities to variogramm estimates (see e.g. Diggle, Liang & Zeger 1994).

Beside the occurrence of correlation we observe a locally varying amount of complexity in the mean function. This implies that when using a global bandwidth, overfitting occurs for some areas of the data while other areas suffer from undersmoothing. To avoid this we use local bandwidths. For independent observations local bandwidth selection is treated e.g. in Staniswalis (1989), Fan & Gijbels (1995) or Herrmann (1997) and references given in these papers. The basic idea is to derive locally the optimal bandwidth by asymptotic arguments which are then used for plug in procedures or local estimates. The asymptotic results thereby typically require that the locations at which the measurements are taken become infinitely dense and the response variables are measured independently. Both prerequisites are not met in our data example. A first proposal for local bandwidth selection in dependent data is given in Yao & Tong (1998). We here make use of the idea proposed in Fan & Gijbels (1995) but adopt it for dependent data. Moreover, we do

not rely on asymptotic results but optimise a local Akaike criterion. Again, a two step estimation is applied, i.e. first we optimise the Akaike criterion locally, leading to rather variable results due to the small amount of information used. Secondly we smooth the raw local optimal bandwidths to achieve a smoothly varying local bandwidth estimate.

The methods applied in this paper are somewhat heuristically in nature and not proved to be optimal in a rigorous mathematical sense. We emphasise however, that available routines for which a theoretical framework is known are tailored for special data constellations and require particular regularity requirements, none of which is met in our data example. Despite of a formal theoretical justification we think however there is practical justification of our routines as we investigate them in simulation studies, some of which are reported here.

The paper is organised as follows. In Section 2 we consider the correlation and dispersion structure in the data. The mean structure is analysed in Section 3. Section 4 gives some conclusions.

2 Modelling Correlation and Dispersion

2.1 Correlation Structure

Measurements of the response variable give proportions of cataclastic rocks, i.e. they take values between 0 and 1. Instead of working with the proportions directly, we consider the empirical logit of the observed proportions, where we shrink observed values to the interval $[0.01, 0.99]$ so that the logit is properly defined. We define by y the corresponding empirical logit of the measurement and model the mean of y for given depth measurement x by

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

with $\mu(x)$ as unknown but fixed function in x. At this point we do not postulate that $\mu(\cdot)$ is generally smooth in x, as we allow for breakpoints, peaks or spikes and outliers. We assume however that $\mu(\cdot)$ is smooth except of a limited number of discontinuity points. Beside the mean structure (1) we allow the variance of y also to depend on the covariate x, that is

$$Var(y|x) \quad = \quad \sigma^2(x). \tag{2}$$

with $\sigma^2(x)$ as smooth function. Finally, we also impose a smooth model for the correlation structure. We embed this by assuming the correlation of y to mirror an AR(1) process. Letting y_i denote the measurement at x_i, $i = 1, \ldots, n$, we set

$$Cor(y_i, y_{i+1}) \quad = \quad \rho(\tilde{x}_i)^{|x_i - x_{i+1}|} \tag{3}$$

where $\rho(\cdot)$ is considered as smooth function in x and $\tilde{x}_i$ as center of x_i and x_{i+1}, i.e. $\tilde{x}_i = (x_i + x_{i+1})/2$.

Estimation of $\rho(x)$ is a peculiar problem, as we have to decompose $y(x)$ in its functional and its stochastic component. To avoid problems resulting from using fitted residuals (see e.g. Opsomer, Wang & Yang 2001) we suggest to estimate $\rho(x)$ without any prior specification of a mean model. The only assumption we make is that the mean model is smooth at least in most parts, even though it may contain jumps or outliers. The procedure we suggest consists of two steps. First, for estimation of $\rho(x_0)$ at a particular point x_0 we consider data pairs (y_j, y_{j+1}) in the neighbourhood $\mathcal{D}(x_0, r) = \{j : x_j \in [x_0 - r, x_0 + r]\}$ around x_0 with r as bandwidth defining the size of the neighbourhood. The value of r is chosen rather small, e.g. such that the number of elements in $\mathcal{D}(x_0, r)$ takes a value lying between 10 and 50. The idea is now to use the empirical correlation of these data pairs to obtain a first rough estimate for $\rho(x_0)$. The assumption implicitly used thereby is that

all y_j with $j \in \mathcal{D}(x_0, r)$ have approximately the same mean. This is true as long as the bandwidth r is small and $\mu(x)$ is sufficiently smooth. We will however see empirically that the smoothness assumption is only conceptionally needed here and the procedure is easily robustified to cope with outliers, discontinuities or spikes in the mean function.

We define the empirical correlation of the data pairs (y_j, y_{j+1}) in the standard form as

$$\check{\rho}_{\mathcal{D}(x_0,r)} = \frac{\sum_{j \in \mathcal{D}(x_0,r)} (y_j - \bar{y}_{(0)})(y_{j+1} - \bar{y}_{(1)})}{\sqrt{\sum_{j \in \mathcal{D}(x_0,r)} (y_j - \bar{y}_{(0)})^2} \sqrt{\sum_{j \in \mathcal{D}(x_0,r)} (y_{j+1} - \bar{y}_{(1)})^2}} \tag{4}$$

with $\bar{y}_{(l)} = \sum_{j \in \mathcal{D}(x_0,r)} y_{j+l}/n_0$ for $l = 0, 1$. Since data points are not equidistant, we also have to adjust for unequal spacing of the x measurements. It is not difficult to show that with (3) the expectation of (4) is approximately

$$E\{\check{\rho}_{\mathcal{D}(x_0,r)}\} \approx \frac{1}{n_0} \sum_{j \in \mathcal{D}(x_0,r)} \rho(x_0)^{d_j} \tag{5}$$

with $d_j = |x_{j+1} - x_j|$ as difference of two measurement locations. From (5) a moment based estimate for $\rho(x_0)$ is then easily obtained by solving the empirical version of (5), i.e. by solving

$$\check{\rho}_{\mathcal{D}(x_0,r)} = \sum_{j \in \mathcal{D}(x_0,r)} \check{\rho}(x_0)^{d_j} \tag{6}$$

for $\check{\rho}(x_0)$. The solution of (6) is analytically available only if d_j is constant, i.e. if x values are equidistant. If this is not the case a simple Newton algorithm easily allows to solve (6) after a very few steps.

It will be seen that the resulting estimate $\check{\rho}(x_0)$ are rough and rather variable. This has basically two reasons. First, bandwidth r defining the neighbourhood

$\mathcal{D}(x_0, r)$ has been chosen as small value so that only a few observations give information about $\check{\rho}(x_0)$. Secondly, if $\mu(x)$ is not smooth but discontinuous at or close to x_0 the basic assumption of smoothness used in (5) is violated and $\check{\rho}(x_0)$ will be biased due to the discontinuity on $\mu(x)$. It is therefore necessary to proceed with a second estimation step in order to smooth the values of $\check{\rho}(x_0)$. Let therefore $x_{01}, \ldots, x_{0N}$ be a grid of points at which the empirical estimates $\check{\rho}(x_{0l})$ are calculated, $l = 1, \ldots, N$. We smooth the values by a robust estimate in order to down-weight extreme values of $\check{\rho}(x_{0l})$. To be more specific we make use of the local M estimate as suggested in Chu, Glad, Godtliebsen & Marron (1998) and calculate

$$\widehat{\rho}(x_0) \quad = \quad \arg\min \sum_{l=1}^{N} K\Big(\frac{x_0 - x_{0l}}{h_x}\Big) m\Big(\frac{\check{\rho}(x_{0l}) - \widehat{\rho}(x_0)}{h_y}\Big) \tag{7}$$

where $K(\cdot)$ is a kernel function with h_x as bandwidth and $m(\cdot)$ is a robust distance function with h_y as bandwidth. For $m(y) = y^2$ standard kernel smoothing results while a robust version is obtained if $m(y)$ is bounded for $y \to \pm\infty$ (see Hampel, Ronchetti, Rousseeuw & Stahel 1986). We use $m(y) = 1 - \exp(-y^2)$ below. Additional consideration of the resulting estimates is sometimes required to check that the 'global' optimum of (7) has been achieved and not a 'local' version.

2.2 Estimation of the Dispersion

In a similar fashion as above we estimate the dispersion function $\sigma^2(x)$. We first calculate the rough empirical variance at x_0 as $\check{\sigma}^2_{\mathcal{D}(x_0,r)} = \sum_{j \in \mathcal{D}(x_0,r)} (y_j - \bar{y}_0)^2/(n_0 - 1)$. Let $P_0 = I_{n_0} - 1_{n_0} 1_{n_0}^T / n_0$, where I_{n_0} is the n_0 dimensional identity matrix and 1_{n_0} is the n_0 dimensional unit vector, where n_0 is the number of elements in $\mathcal{D}(x_0, r)$. Moreover, with R_0 we denote the correlation matrix with entries $\rho(x_0)^{|x_j - x_l|}$ for $j, l \in \mathcal{D}(x_0, r)$. Assuming now $\sigma^2(x)$ to be sufficiently smooth we get with simple

calculation

$$E(\check{\sigma}^2_{\mathcal{D}(x_0,r)}) \approx \sigma^2(x_0)\mathrm{tr}(P_0R_0P_0). \quad (8)$$

A rough estimate $\check{\sigma}^2(x_0)$ is then obtained by solving the empirical version of (8), i.e. $\check{\sigma}^2(x_0) = \check{\sigma}^2_{\mathcal{D}(x_0,r)}/\mathrm{tr}(P_0\widehat{R}_0P_0)$, where $\widehat{R}_0$ is a plug in estimate of the correlation calculated above. Rough estimates $\check{\sigma}^2(x_{0l})$ are now calculated for a series grid of points $x_{01}, \ldots, x_{0N}$ from which the final estimate $\widehat{\sigma}^2(x)$ is achieved by robust smoothing in complete analogy to (7).

2.3 Simulations and Application to Deep Drill Data

Simulation

Before applying the procedure to the deep drill data we want to experience its performance at simulated data. In Figure 1 (upper plot) we show normally distributed data tracing from a quadratic mean model with equidistant x values. The pairwise correlation increases linearly from value 0.4 at x=0 to 0.8 at $x = 1000$. The left panel in Figure 1 shows the rough estimates $\check{\rho}(x_{0l})$ for different values of bandwidth r ranging from $r = 5$, 10, 20, 30, 50 to 100, top downwards. The second step estimate $\widehat{\rho}(x)$ resulting from (7) is shown as standard kernel smoother (wicked curve) and as robust local M estimate (dotted curve). The true correlation function is shown a solid line. We see generally a promising behaviour of $\widehat{\rho}(x)$ as long as the bandwidth r is reasonably sized. For small values the correlation is somewhat under-estimated while in contrast for large values of r it is over-estimated. The first effect can be corrected by a small sample adjustment applied to $\check{\rho}(x_{0l})$, which is however not further considered in this paper. The reason for over-estimation in the later case is that correlation is calculated for observation pairs from a too wide range of x values. Consequently the implicit assumption of a joint mean level is violated and empirical

correlations are biased.

In the right panel we show the rough estimates $\sqrt{\breve{\sigma}^2}(x_{0l})$ accompanied by a standard kernel smooth (wicked curve) and a local M estimate (dotted curve). Moreover we fit the rough estimates by a global M estimate, that is a robust estimate using all data points. This is shown as dashed line. As reference the true dispersion of value 0.3 is given a straight line. The estimates appears to perform reasonably well, except for large bandwidths r.

We run the procedure again in a second simulated data set, this time with discontinuous effects, as shown in Figure 2. The performance is very much the same and hardly disturbed by the breakpoints. This demonstrates the robustness of the routine towards breakpoints and outliers. Finally, we investigate how the procedure copes with independent observations in case of varying dispersion. The results are shown in Figure 3. The performance is very much the same as in the first two simulations.

Deep Drill Data

We now apply the routine to the deep drill data. In Figure 4 (left panel) we show the fitted correlation structure for bandwidths $r = 10$, 30 and 100, top downwards. Standard kernel estimates show as wicked curves and local M estimates are given as dotted curves. The correlation increase in the first 2000 meters and keeps a constant level of 0.8 between 3000 and 5000 meters depth. The pattern can nicely be explained by technical features of the drill experiment. With a special liquid the rock samples are washed from the drill location to the surface. Due to this washing the rocks to some extend technically mix up with remaining rocks from higher level locations. This puts an additional source of autocorrelation to the data

which increases as the distance of transport increases.

In the right hand panel of Figure 4 we show the fitted dispersion for the three different values of bandwidth r. The curves represent a standard kernel estimate (wicked line), a local M estimate (dotted line) and a global linear parametric fit given as dashed line. Clearly, the dispersion decreases with depth, which goes along with the technical explanation just given. Based on the experience we collected in the simulations we consider the results for $r = 30$ as appropriate choice, even though overall differences for the different bandwidths are marginal.

3 Modelling of the Mean Structure

3.1 Detecting Spikes and Outliers

The above results can now be used to check for spikes and outliers in the data. We therefore consider pairwise differences $y_{i+1} - y_i$. Assuming $\mu(x)$ to be smooth we have $E(y_{i+1} - y_i) \approx 0$, as $x_{j+1} - x_j$ is small. If however $\mu(x)$ has a spike or an observation is an outlier this shows in significantly large or small values of $y_{i+1} - y_i$. Note that by assuming smoothness for $\sigma(x)$ and $\rho(x)$ one gets

$$\begin{aligned} Var(y_{i+1} - y_i) &= \sigma^2(x_i) + \sigma^2(x_{i+1}) - 2\rho(\tilde{x}_i)^{d_i}\sigma(x_i)\sigma(x_{i+1}) \\ &\approx \sigma^2(x_i)\{2 - \rho(x_i)^{d_i}\}. \end{aligned} \tag{9}$$

Inserting plug in estimates in (9) yields the standardised differences $z_i = (y_i - y_{i+1})/\sqrt{Var}(y_{i+1} - y_i)$ which are plotted in Figure 5. As horizontal line we include thresholds based on assumed standard normality of z_i. The dotted line corresponds to a local 99% level, the solid line gives an overall 90% level, based on a Bonferoni adjustment. In Figure 6 we plot the data and indicate the locations detected as breakpoints on a local 99% level by vertical bars.

The detected breakpoints are not necessarily identical with peaks in the data but they may mark changes of trends. Individual breakpoints can be related to technical operations which occurred in the drilling, e.g. in depths 1100 to 1300 meters bore-hole breakouts (extended caliper) took place; at 3000 meters the bore-hole diameter was reduced due to casing of the upper part. Technical explanations however do not hold for all detected breakpoints. For instance the breakpoints found at 4000 and 4900 m can clearly be related to prominent cataclastic shear zones.

3.2 Fitting the Mean Structure

We now finally examine the mean structure (1) in the data. In a first step we exclude observations (and their neighbours) which were classified as outliers or peaks above as these observations were found not to match to the smooth model (1). Based on the remaining observations we apply a simple kernel smoother. The bandwidth is thereby chosen data driven. Remember that appropriate bandwidth selection in correlated data is known to be an important step and the correlation generally forbids to use standard routines like e.g. cross validation. Since we however have already estimated the correlation structure, even without mean structure specification, it is now straight forward to apply an Akaike criterion based on a factorisation of the likelihood. We assume $y_1 \sim N\Big(\mu(x_1), \sigma^2(x_1)\Big)$ and for $1 \leq i < n$ we set $y_{i+1}|y_i \sim N(\mu(x_{i+1}|y_i), \sigma^2(x_{i+1}|y_i))$ with $\mu(x_{i+1}|y_i) = \mu(x_{i+1}) + \rho^{d_i}(x_i)\Big(y_i - \mu(x_i)\Big)$ and $\sigma^2(x_{i+1}|y_i) = \sigma^2(x_{i+1})\Big(1 - \rho^{2d_i}(x_i)\Big)$. The Akaike criterion is then defined as

$$AIC(h) = -2\log\{\phi(y_i)\prod_{i=1}^{n-1}\phi(y_{i+1}|y_i, h)\} + 2df(h) \tag{10}$$

with $\phi(\cdot, h)$ as normal distribution density with moments as specified above and fitted mean value dependent on h. The degree $df(h)$ of the model is thereby defined in the standard way as $df = \sum_{i=1}^{n} 1/\sum_{j=1}^{n} K\{(x_i - x_j)/h\}$ with $K(\cdot)$ as kernel

function (see e.g. Hastie & Tibshirani 1990). The resulting Akaike curve is shown in Figure 7. There is a clear minimum at $df = 100$ which corresponds to the bandwidth $h = 40$.

The resulting fit is shown in Figure 6. Apparently, the fit does not look satisfactory, since the estimate is rather jagged in some areas. Note that this happens despite of the fact that we explicitly considered the correlation structure in the bandwidth selection. The data however demand for local bandwidth selection since the complexity of the mean structure apparently changes with depth. We make use of a procedure comparable to Fan & Gijbels (1995) by choosing the bandwidth by locally optimising the Akaike criterion. Note that the elements in (10) decompose additively to

$$AIC(h) \approx 2\sum_{i=1}^{n-1}\Big(-\log\phi(y_{i+1}|y_i,h) + df_i(h)\Big) \tag{11}$$

by neglecting the contribution of the first observation and $df_i(h) = 1/\sum_j K\{(x_i - x_j)/h\}$. A local Akaike criterion is now achieved by replacing the sum in (11) by a locally weighted sum, that is

$$AIC(x_0,h) \approx 2\sum_{i=1}^{n-1} K\Big(\frac{x_0 - x_i}{h_{AIC}}\Big)\Big(-\log\phi(y_{i+1}|y_i,h) + df_i(h)\Big) \tag{12}$$

with h_{AIC} as bandwidth giving the locality of the criterion. In practice we set h_{AIC} as small value and in the example we use the global optimal bandwidth resulting from (10). We define with $\check{h}_{x_0}$ the minimiser of (12). In practice $\check{h}_{x_0}$ can be obtained from a grid search over different bandwidths. Locally optimal bandwidths are now calculated over a grid of points $x_{01}, \ldots, x_{0N}$. It is obvious that the resulting outcomes $\check{h}_{x_0}$ will be rather variable as only a small number of observations is used for local bandwidth selection. This can also be seen from Figure 8 where we plot the raw estimates $\check{h}_{x_0}$ for the deep drill data. As previously we therefore apply a second

step to smooth $\check{h}_{x_0}$ yielding a smooth bandwidth curve $\widehat{h}(x)$. We use simple kernel smoothing as well as local M smoothing. The final bandwidth estimate $\widehat{h}(x)$ is then used for further fitting. This means we estimate the mean structure with the locally chosen bandwidth $\widehat{h}(x)$. The result is shown in Figure 9. The positive effect of local bandwidth usage becomes obvious.

The data are now appropriately smoothed, meaning high complexity areas are fitted with a lower bandwidth, while low complexity areas get a larger bandwidth. The first point appears that peaks are more clearly pronounced, the latter shows in smooth fits for the remaining regions. The peaks in the local bandwidth fit now clearly mark shear zones, at e.g. 1170, 1550, 1980, and 4000 meters. In addition, some of the high gradients can be explained by changes in lithology, e.g., in the sections from 3150 to 3400 m and 4380 to 4420 m. Moreover, the peak at 4800 meters can be related to a enhanced occurrence of the radioactive element Thorium.

Simulation

Finally, we want to demonstrate the effect of local bandwidth selection in a simulation study. In Figure 10 we show data resulting from a model with varying complexity over the range of x. The data are correlated and we use the true correlation structure in the following calculation. We first minimise the global Akaike criterion yielding the estimate shown as dotted curve in the upper plot. Clearly, undersmoothing become apparent in the flat areas of the function while oversmoothing occurs in the peaks. We now calculate raw local bandwidths $\check{h}_{x_{0l}}$ for a range of grid points $x_{01}, \ldots, x_{0N}$, shown in the bottom plot of Figure 10. We use local kernel estimation and local M estimation to smooth the raw bandwidths. The latter robust estimate is taken to resmooth the mean structure using a local bandwidth which is

shown as solid curve in the upper plot. The positive effect becomes apparent, as undersmoothing in the flat regions is avoided and the structure in the peak regions is satisfactory fitted.

4 Conclusion

The analysis presented in this paper gave interesting new insight in the structure of depth related drill cuttings. The trend of increasing autocorrelation with depth could clearly be related to technical aspects of drill cuttings sampling. The breakpoint analysis gave at some depths hints for possible relations to bore-hole conditions whereas at other depths changes in lithology or shear zones were represented. Generally, the interpretations hold for distinct peaks only and a unique explanation seems yet not possible. This will be subject of further interdisciplinary research in this area. The analysis of the mean structure with its peaks uncovered cataclastic shear zones. Some additional links to lithological changes could be observed locally and the effect of overlapping of several influence variables seems to be present. This can be resolved by multivariate analysis taking other variables into account (e.g., Winter et al., 2002). In general, mean structure analysis and breakpoint analysis are procedures which give a new methodical input to the analysis of drill hole data.

In terms of statistical technology the paper demonstrated the analysis of a dataset with various peculiarities. We were faced with correlated errors, spikes and outliers as well as the necessity to determine the bandwidth locally. Each of the three topics is separately well developed in the literature. Their application to our data problem seemed however questionable since for instance estimation of correlation based on fitted residual can hardly cope with spikes or outliers. We demonstrated how the problems occurring in the data could be handled by two step estimation routines.

First rough local estimates were calculated which were then smoothed by a second smooth fit. Further investigation is required to justify this approach on a theoretical basis. The simulation conducted (only a part of them are shown here in the paper) let us however hope that we were doing the right things.

Acknowledgements

Küchenhoff's research was supported by a grant from the German Research Foundation. We would like to thank Helmuth Winter for useful discussions and help in interpreting our findings.

References

Altman, N. S. (1990). Kernel smoothing of data with correlated errors. *Journal of the American Statistical Association* **85**, 749–759.

Chiu, S.-T. (1989). Bandwidth selection for kernel estimate with correlated noise. *Statistics & Probability Letters* **8**, 347–354.

Chu, C. K., Glad, I. K., Godtliebsen, F., and Marron, J. S. (1998). Edge-preserving smoothers for image processing (with discussion). *Journal of the American Statistical Association.* **93**, 526–541.

Diggle, P. J., Liang, K.-Y., and Zeger, S. L. (1994). *Analysis of longitudinal data.* Oxford University Press.

Emmermann, R. and Lauterjung, J. (1997). The german continental deep drilling program ktb: Overview and major results. *Journal of Geophysical Research* **102(B8)**, 18179–18201.

Fan, J. and Gijbels, I. (1995). Data-driven bandwidth selection in local polynomial fitting: Variable bandwidth and spatial adaptation. *Journal of the Royal Statistical Society, Series B* **57**, 371–394.

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

Hampel, F. R., Ronchetti, E. M., Rousseeuw, P. J., and Stahel, W. A. (1986). *Robust Statistics: The Approach Based on Influence Functions.* Wiley.

Hart, J. D. (1991). Kernel regression estimation with time series error. *Journal of the Royal Statistical Society, Series B* **53**, 173–187.

Hastie, T. and Tibshirani, R. (1990). *Generalized Additive Models.* London: Chapman and Hall.

Herrmann, E. (1997). Local bandwidth choice in kernel regression estimation. *Journal of Computational and Graphical Statistics* **6**, 35–54.

Opsomer, J., Wang, Y., and Yang, Y. (2001). Nonparametric regression with correlated errors. *Statistical Science* **16**, 134–153.

Simonoff, J. S. (1996). *Smoothing Methods in Statistics.* New York: Springer Verlag.

Staniswalis, J. G. (1989). The kernel estimate of a regression function in likelihood–based models. *Journal of the American Statistical Association.* **84**, 276–283.

Winter, H., Adelhardt, S., Jerak, A., and Küchenhoff, H. (2002). Characterization of cataclastic shear zones of the ktb deep drill hole by regression analysis of drill cuttings data. *Geophys. J. Internat.*, (to appear).

Yao, Q. and Tong, H. (1998). Cross-validatory bandwidth selections for regression estimation based on dependent data. *Journal of Statistical Planning and Inference* **68**, 387–415.

Zulauf, G., Palm, S., Petschick, R., and Spies, O. (1999). Element mobility and volumetric strain in brittle and brittle-viscous shear zones of the superdeep well ktb (germany). *Chemical Geology* **156**, 135–149.

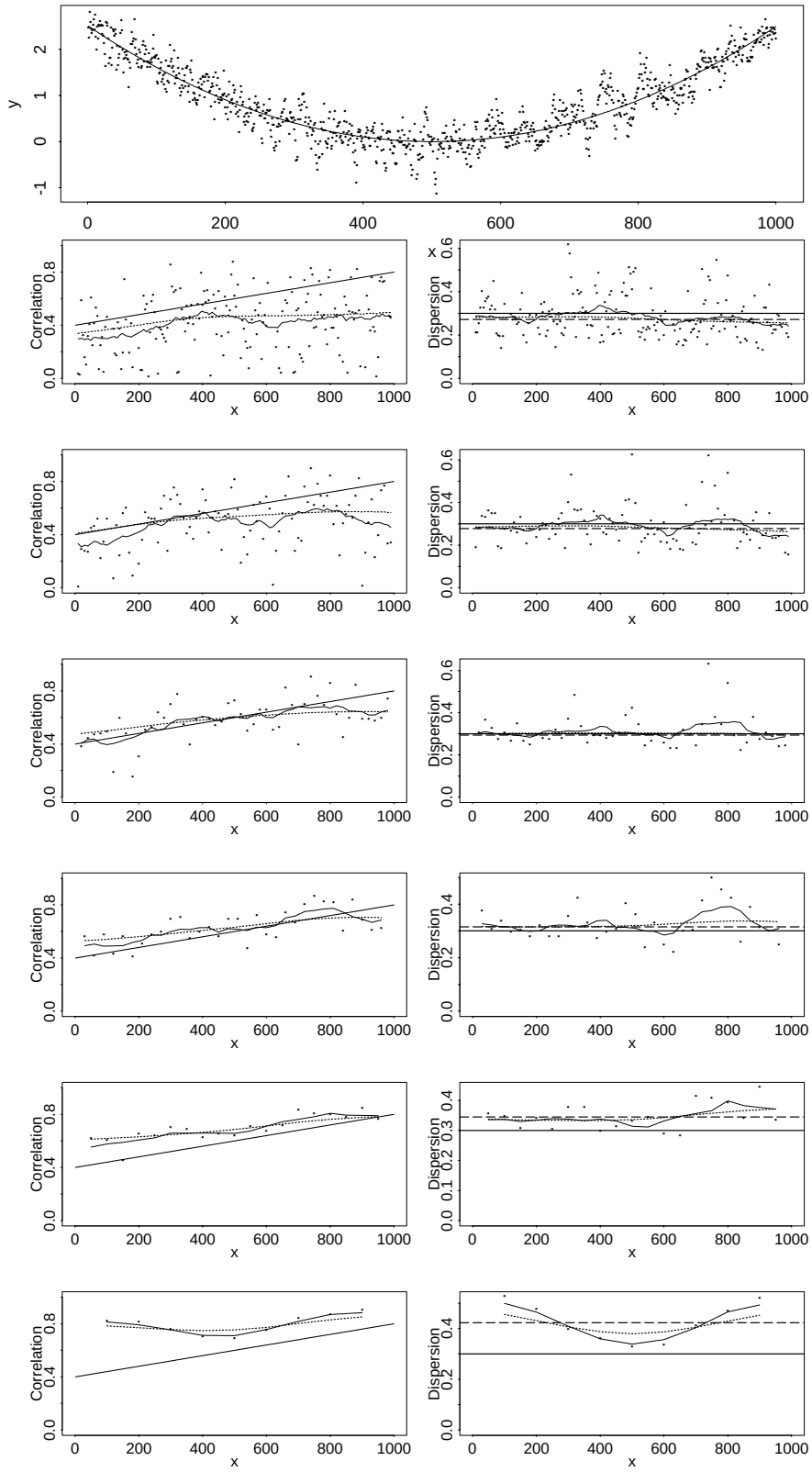

Figure 1: Simulated correlated data (upper plot) and estimated correlation structure (left column) and dispersion structure (right column) for different bandwidths r (from top down: $r = 5,\ 10,\ 20,\ 30,\ 50,\ 100$) Wicked lines gives standard kernel smoother, dotted lines shows local M smoother, solid lines shows true curves. Dashed line in right panel gives a robust parametric fit.

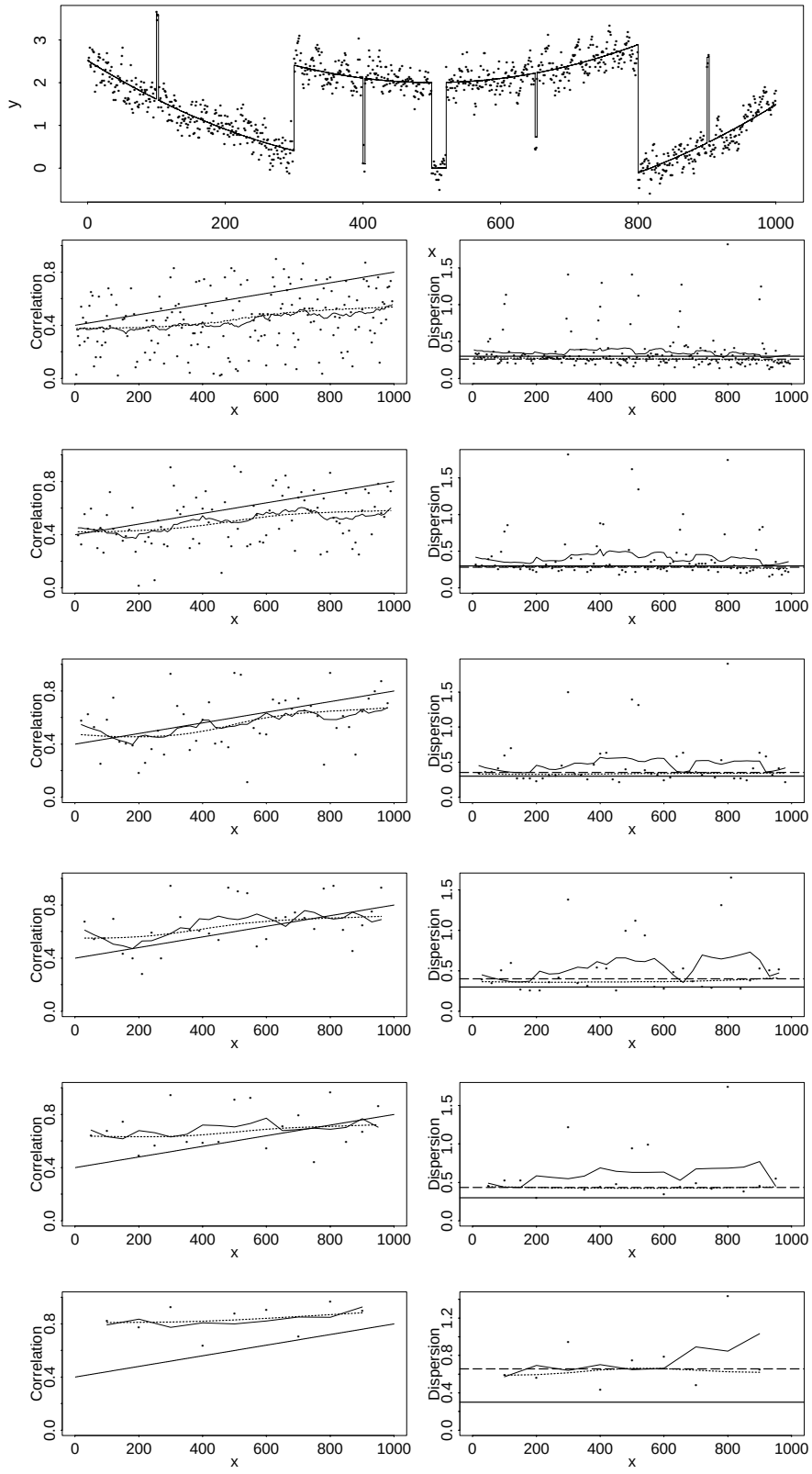

Figure 2: Simulated correlated data (upper plot) and estimated correlation structure (left column) and dispersion structure (right column) for different bandwidths r (from top down: $r = 5,\ 10,\ 20,\ 30,\ 50,\ 100$) Wicked lines gives standard kernel smoother, dotted lines shows local M smoother, solid lines shows true curves. Dashed line in right panel gives a robust parametric fit.

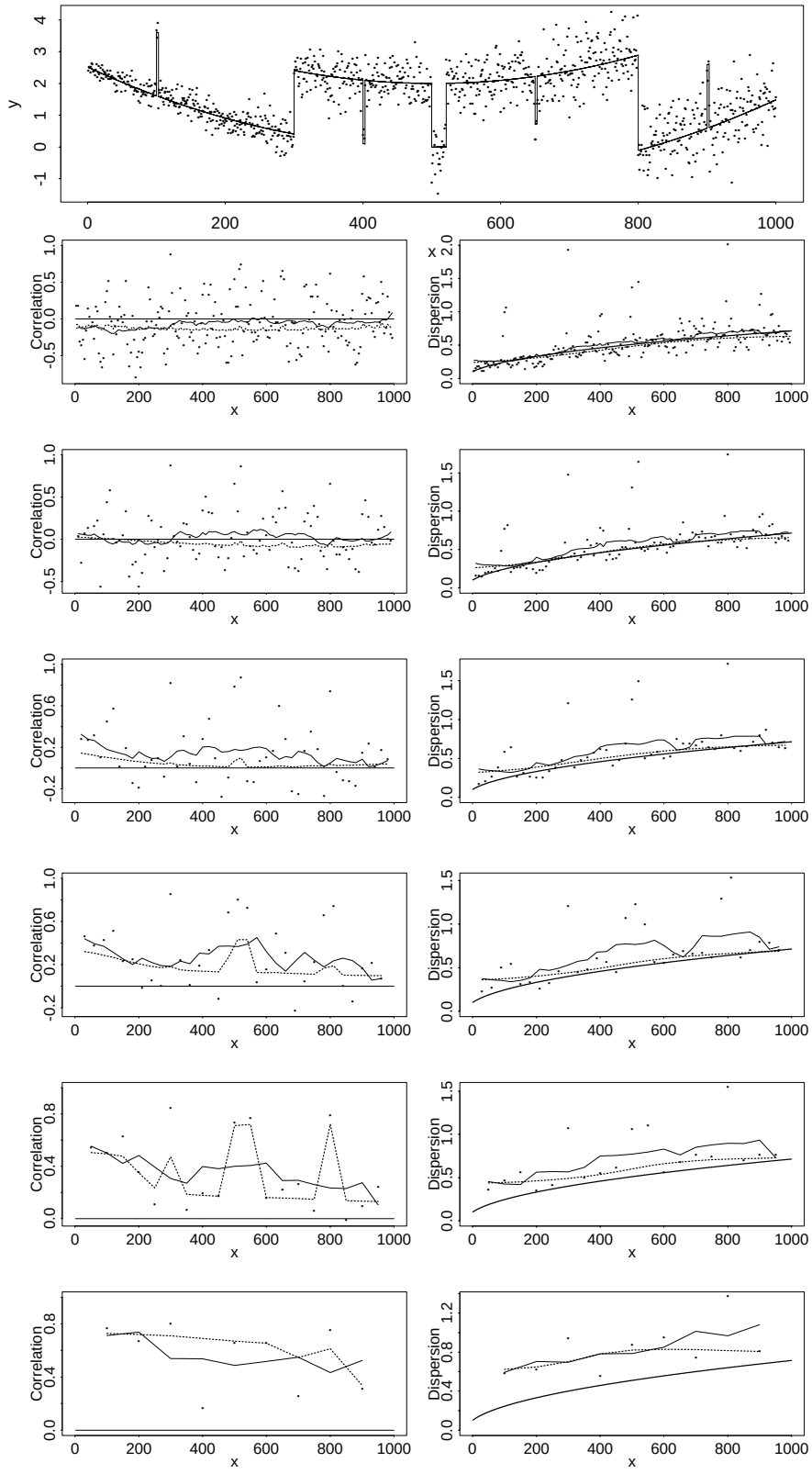

Figure 3: Simulated independent data (upper plot) with increasing dispersion. Estimated correlation structure (left column) and dispersion structure (right column) for different bandwidths r (from top down: $r = 5, 10, 20, 30, 50, 100$) Wicked line gives standard kernel smoother, dotted line shows local M smoother. Solid line in left panel shows true value 0, solid line in right hand panel give true curve $\sqrt{\sigma^2(x)}$

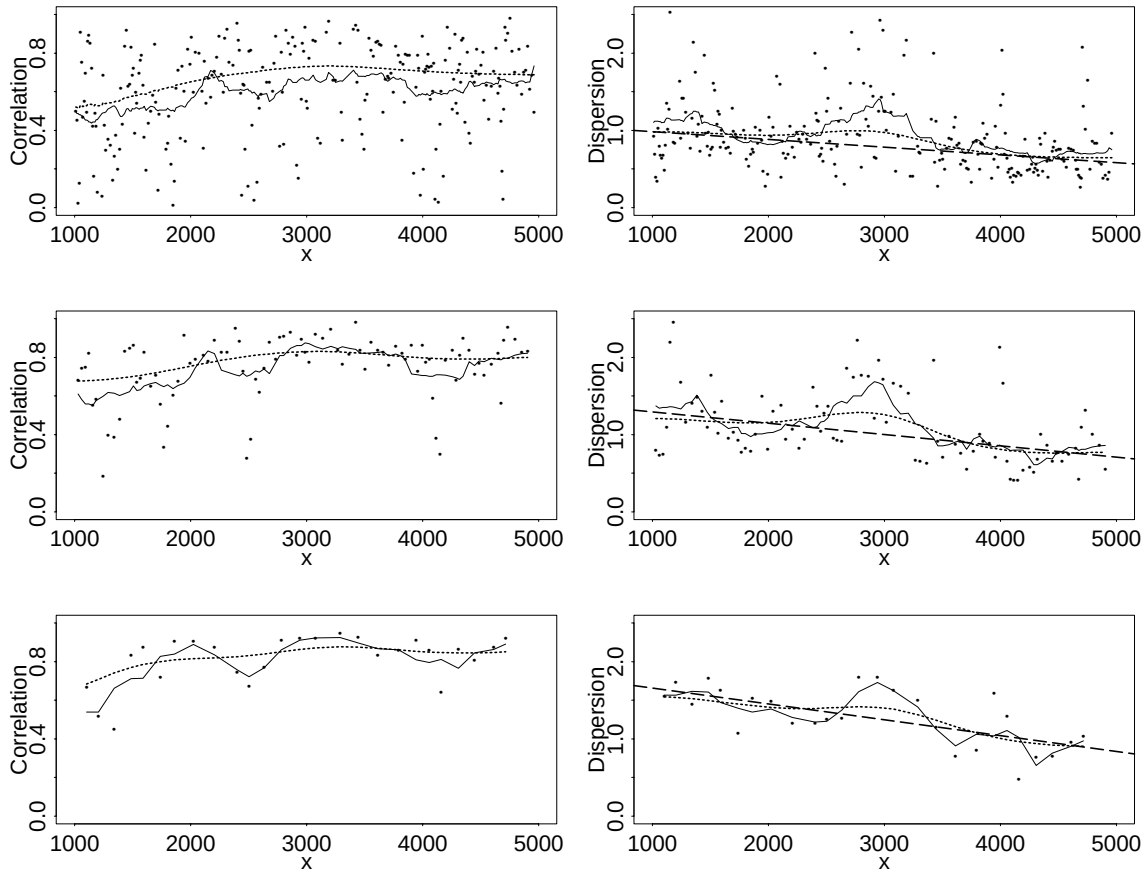

Figure 4: Estimated correlation (left panel) and dispersion structure (right panel) for different bandwidths (from top down $r = 10, 30, 100$). Wicked line gives standard kernel smoother, dotted line shows local M smoother. Dashed line in right panel is a robust parametric fit with linear slope.

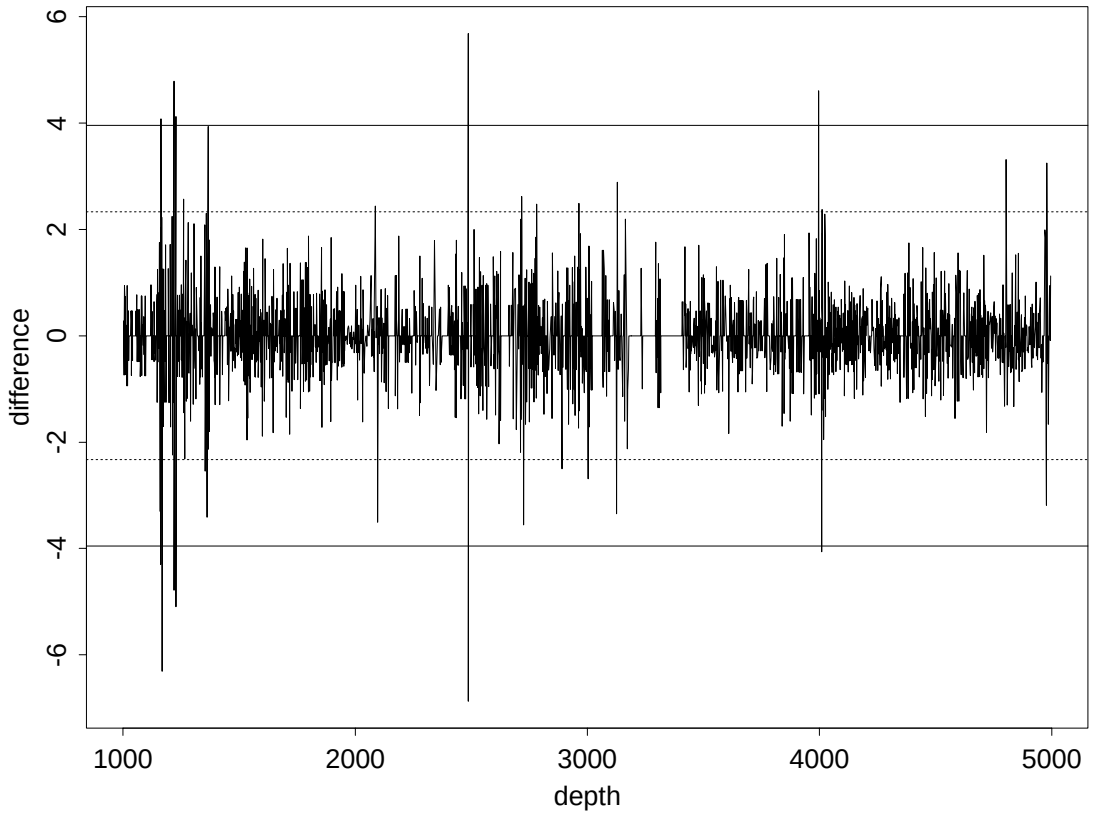

Figure 5: Standardised difference estimates with local threshold based on 99% threshold (dotted line) and global threshold based on 90%

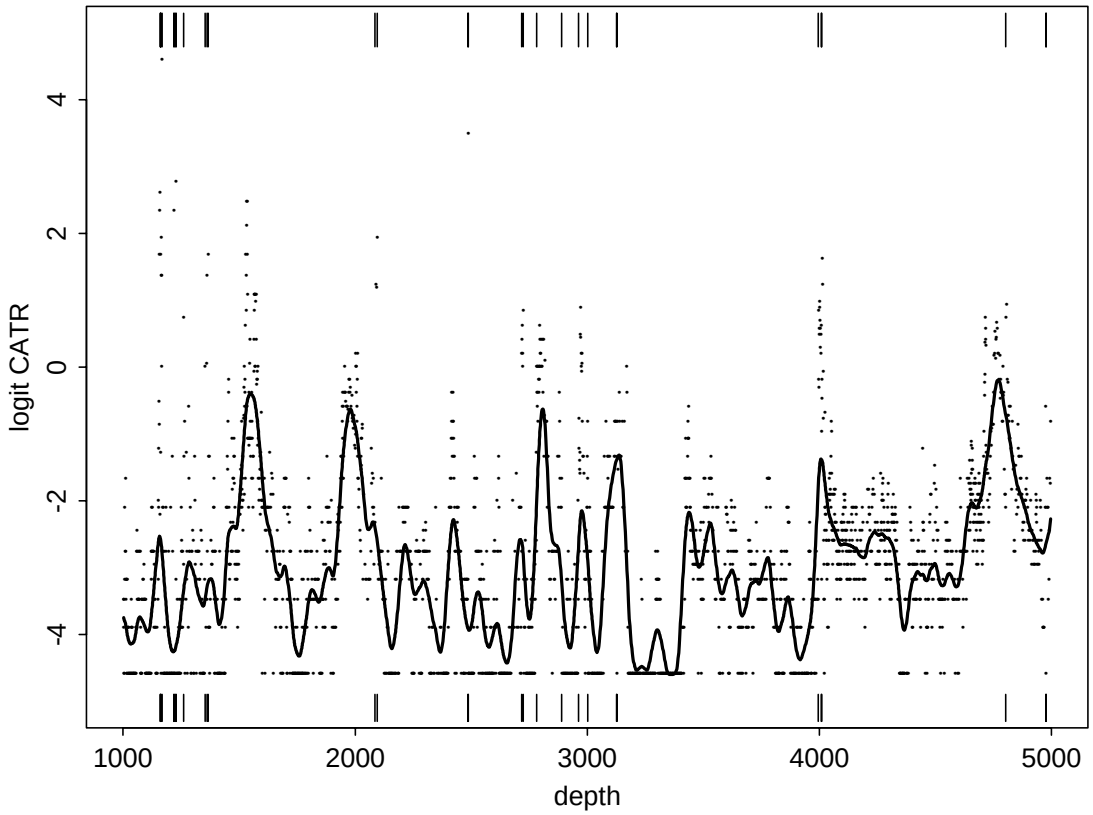

Figure 6: Deep drill data with significant breakpoints

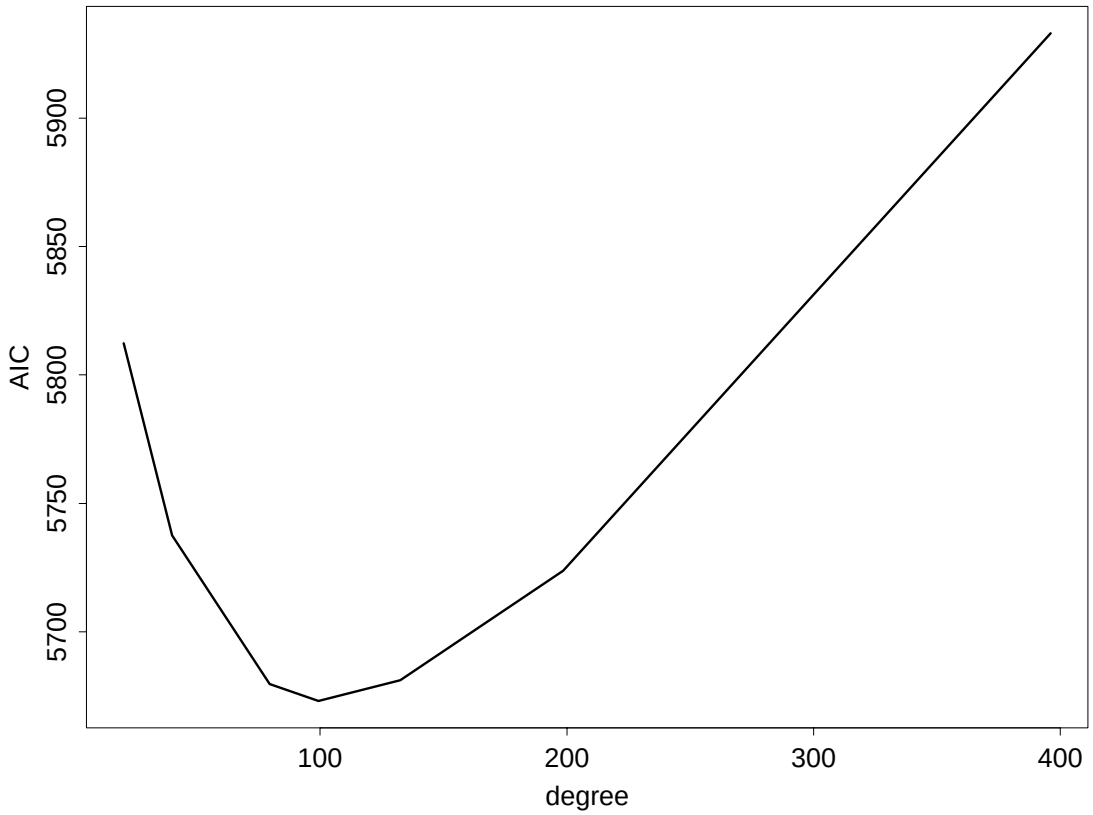

Figure 7: Akaike criterion for deep drill data

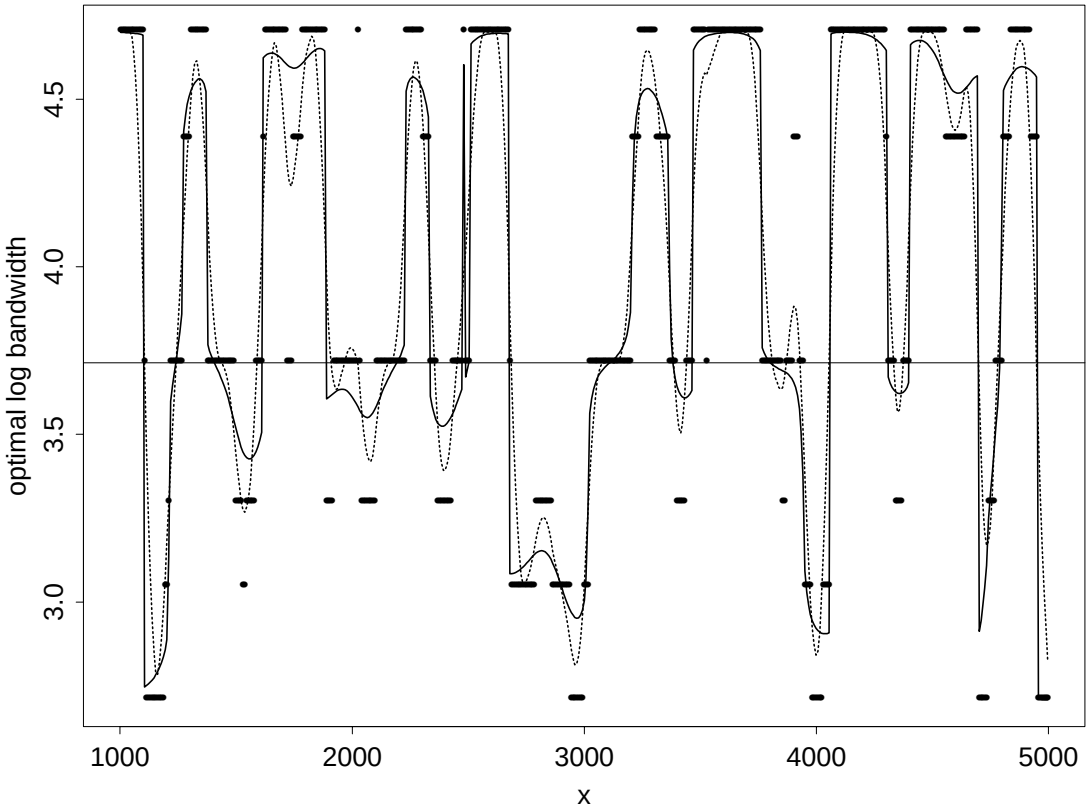

Figure 8: Local bandwidths $\check{h}_{x_{0l}}$ and their corresponding kernel smooth (dotted line) and local M smooth (solid line). Optimal global bandwdith (log(40)) is shown as horizontal line.

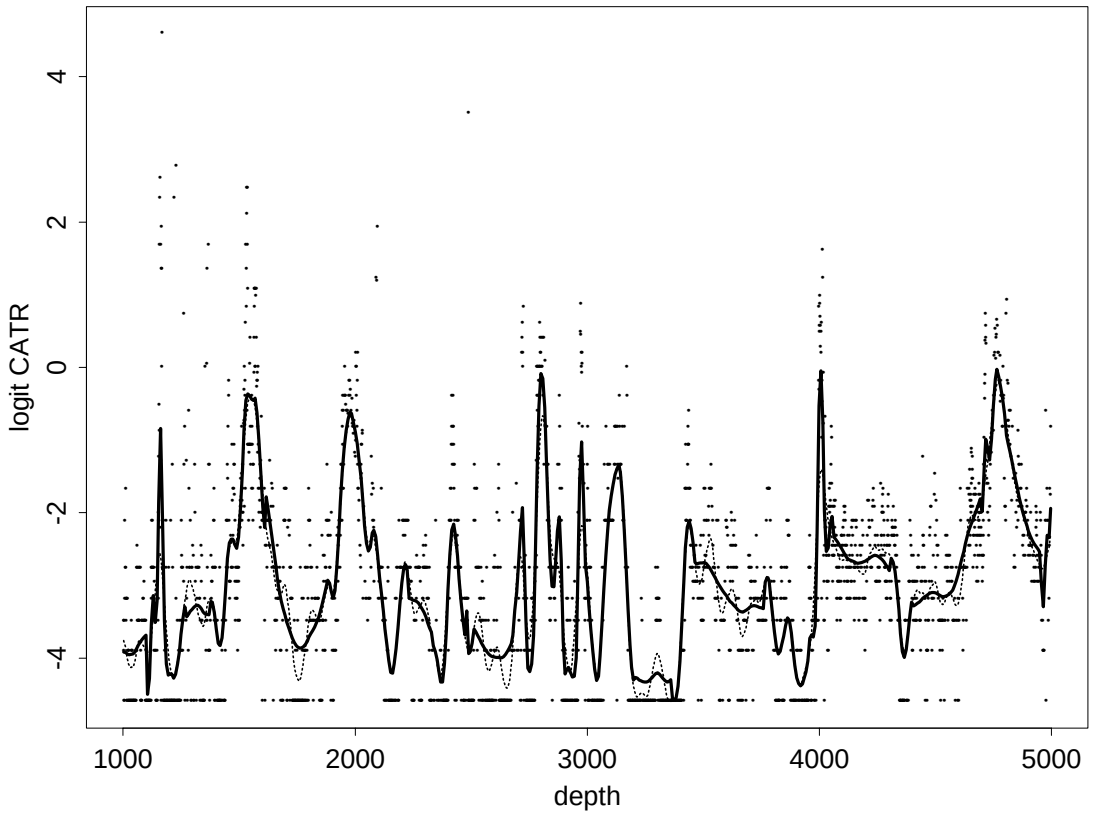

Figure 9: Smooth Mean Structure with local bandwidth (solid curve) and global bandwidth (dotted curve)

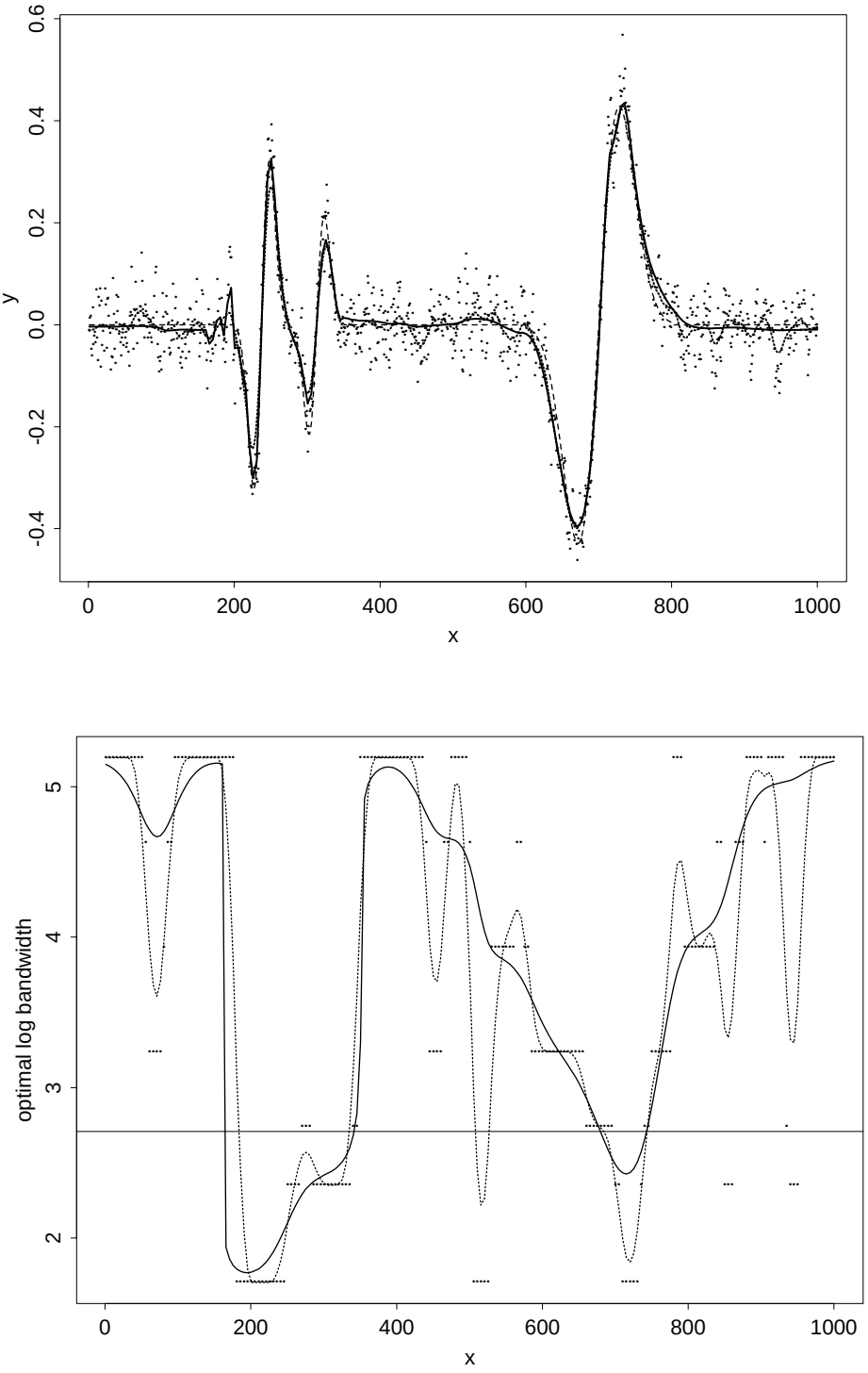

Figure 10: Smooth mean structure (upper plot) with local bandwidth (solid curve) and global bandwidth (dotted curve). True curve is given as dashed line. Local optimal bandwidths $\check{h}_{x_{0l}}$ (lower plot) with kernel smooth (dotted line) and local M smooth (solid line). Global optimal bandwidth is indicated as horizontal line