STUDIES IN TREND DETECTION OF SCATTER PLOTS WITH VISUALIZATION

PANU ERÄSTÖ

Academic Dissertation for the Degree of Doctor of Philosophy

To be presented, with the permission of the Faculty of Science
of the University of Helsinki, for public criticism in Auditorium XII, the Main
Building (Fabianinkatu 33), on January 20th 2006, at 12 o'clock.

Department of Mathematics and Statistics
Faculty of Science
University of Helsinki

ISBN 952-91-9795-0 (Paperback)
ISBN 952-10-2865-3 (PDF)
Yliopistopaino
Helsinki 2005

Cover: Three views with different scales on an image showing a wall fern (*Polypodium vulgare*). Photo: Panu Erästö, Porvoo, 2005.

Acknowledgements

I wish to express my sincere gratitude to my supervisor Lasse Holmström for his guidance and support during my doctoral studies. Through his incessant questioning, he has taught me the value of rigor and hard work. I am also indebted to Petri Koistinen for fruitful discussions and for having introduced me to the subject together with Lasse. I also warmly acknowledge the extensive collaboration with Atte Korhola and Jan Weckström in the application field. Furthermore, I would like to thank all the people of the former Rolf Nevanlinna Institute for the supportive and motivating atmosphere - Elja Arjas, Tommi Härkänen, Pia Kemppainen-Kajola, Dario Gasbarra and Matti Taskinen, just to name a few.

Lunches and other breaks with Petri, Fabian, Mikko and Simopekka have also been important, and not only in terms of physical nutrition. In addition, I thank all my friends for support and for showing me that there is life outside of the office. Particularly the meetings with Matti and Jukka have been vital for me. I want to dedicate this thesis to my family: Antti, Pirjo, Tytti, Sari and Kuutti.

I gratefully acknowledge the financial support from the Academy of Finland, the graduate schools of Stochastics, Mathematical Analysis and Logic, and SIIDA.

Helsinki, December 2005

Panu Erästö

List of original publications

This thesis is based on the following six original articles which are referred to in the text by the format 'Paper' followed by a Roman numeral:

Paper I: ***Using Smoothing to Reconstruct the Holocene Temperature in Lapland***. L. Holmström, P. Erästö, P. Koistinen, J. Weckström, and A. Korhola. In E. J. Wegman and Y. Martinez, editors. *Proceedings of the 32nd Symposium on the interface of computing science and statistics*, New Orleans, pages 425-438. Interface Foundation of North America, 2000.

Paper II: ***A Quantitative Holocene Climatic Record from Diatoms in Northern Fennoscandia***. A. Korhola, J. Weckström, L. Holmström, and P. Erästö. *Quaternary Research*, 54:284-294, 2000.

Paper III: ***Making Inferences about Past Environmental Change Using Smoothing in Multiple Time Scales***. L. Holmström, and P. Erästö. *Computational Statistics & Data Analysis*, 41(2):289-309, 2002.

Paper IV: ***Bayesian Multiscale Smoothing for Making Inferences about Features in Scatter Plots***. P. Erästö and L. Holmström. *Journal of Computational and Graphical Statistics*, 14(3):569-589, 2005.

Paper V: ***Bayesian analysis of features in a scatter plot with dependent observations and errors in predictors***. P. Erästö and L. Holmström. Submitted.

Paper VI: ***Selection of Prior Distributions and Multiscale Analysis in Bayesian Temperature Reconstructions Based on Fossil Assemblages***. P. Erästö and L. Holmström. To appear in *Journal of Paleolimnology.*

Contents

1 Foreword

The ever increasing amount of complex data gathered in scientific, technological and business applications has created a need for new flexible data analysis methods. Exploration of data in order to discover its main features has become a vital step in many data management tasks as merely storing the data is not adequate for most present-day purposes. In this thesis mathematical tools for explorative data analysis of scatter plots are developed. In particular, ways to make statistically reasonable inferences about features in the data will be considered. As an important application we will analyze past temperatures reconstructed on the basis of samples of modern and past environmental indicators and responses.

Since exploration of data should not be left only to highly trained specialists we try to provide results of inferences in a form that should be accessible for nonspecialists, too. This is done through graphics that summarize the results of inference in an easily understandable manner. Thus, one goal of this thesis is to develop visualization techniques that could facilitate exploratory data analysis for these not extensively trained in statistics.

2 Smoothing

We will concentrate on making inferences about two-dimensional scatter plot data (y_i, x_i), $i = 1, \dots, n$. Let us denote $\mathbf{y} = [y_1, \dots, y_n]^T$ and $\mathbf{x} = [x_1, \dots, x_n]^T$. Unless otherwise stated, we consider a fixed design, where the values of the explanatory variables x_i are known and fixed. They could, however also be treated as random. Two examples of such fixed design scatter plot data are stock price vs. time, and temperature vs. time.

In real world situations it is usually reasonable to assume that the observed responses y_i are somewhat corrupted versions of the reality, that is, to assume that

$$Y_i = m(x_i) + \varepsilon_i, \qquad i = 1, \dots, n, \tag{1}$$

where the ε_i's are the errors often assumed independent and that $\mathbb{E}(\varepsilon_i) = 0$, $\mathrm{Var}(\varepsilon_i) = \sigma^2$. Sometimes, like in this work, the assumptions about the errors are weakened, giving up in particular their independence and making more general distributional assumptions about the observations.

In the above formulation, the target of interest is the mean function m for which $\mathbb{E}(Y_i) = m(x_i)$. Estimation of m is a natural means for extracting interesting phenomena in the data. Sometimes the estimate can also be used to calculate responses for new values of explanatory variables (prediction). Assessing of m from a given set of observations is a function estimation problem. We consider here one dimensional regression, where the only assumption is that m is a mapping $m : \mathbb{R} \to \mathbb{R}$ with some regularity properties assumed when needed. Two other function estimation problems are classification, where $m : \mathbb{R}^k \to \{1, 2, \ldots, c\}$, and density estimation, where $m(x) \geq 0$ and $\int_{\mathbb{R}^k} m(x)dx = 1$, see e.g. [9,41].

When studying the behavior of the unknown function m the first step often is to consider a graphical representation of the data. The estimation of m can be difficult in particular when the data exhibit large variation and the influences of the random errors in the observations are unclear. A common approach for making inferences about m is to smooth the data. A plot of the smooth can be a satisfactory first step for most exploratory data analysis purposes and the smooth can also be used for prediction. There exists a large family of methods for two-dimensional scatter plot smoothing but probably the three most popular are kernel regression, smoothing splines, and orthogonal decomposition methods, especially wavelets. For more details on scatter plot smoothing, see e.g. [6,19,42,43]. In this study we will consider kernel regression and smoothing splines.

2.1 Kernel regression

In kernel regression, a special instant of kernel smoothing methods, a weight function $\kappa(x, \cdot, \cdot)$ is placed at the estimation point x and the observations y_i are weighted according to $\kappa(x, x_i, h)$, where $h > 0$ is a smoothing parameter that controls the width of the weighting kernel. For an illustration, see Figure 1. Letting $h \to \infty$ leads to global smoothing while $h = 0$ corresponds to interpolation.

The basic idea of kernel regression is that after giving weights $\kappa(x, x_i, h)$ to the observations Y_i, the fit $\hat{m}(x; h)$ at x is calculated using a small class of simple functions. Usually the fit is found in the sense of least squares which is the also criterion used implicitly in the following. Note that estimation at x is carried out locally

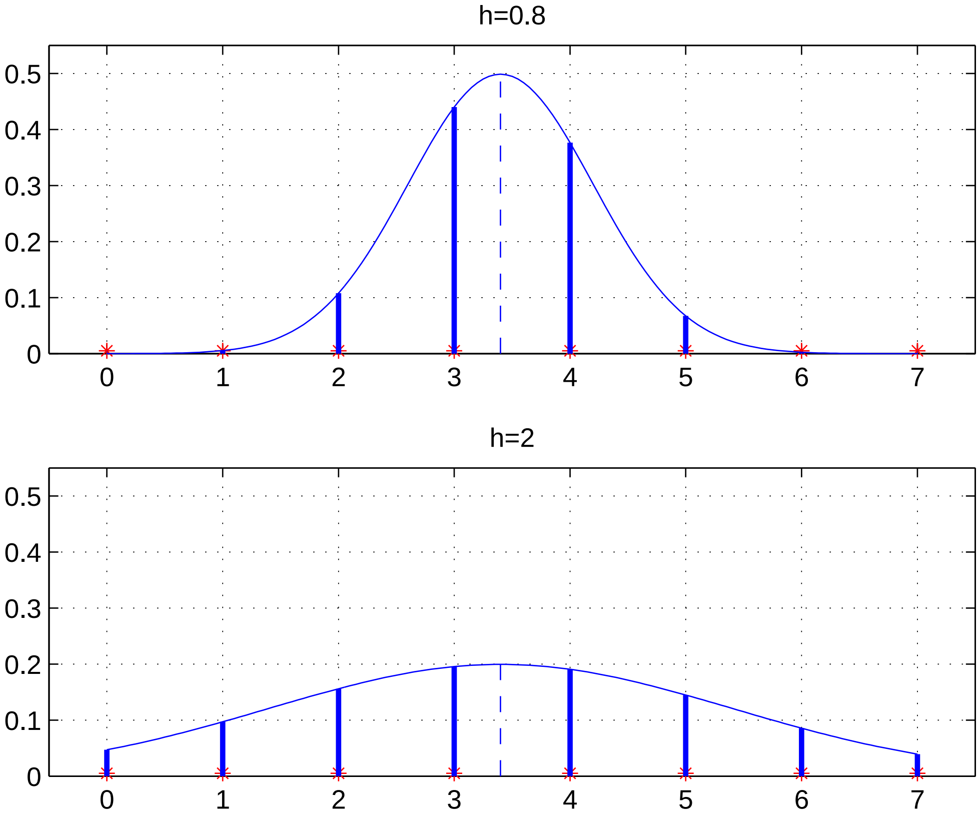

Figure 1: Weighting of the observations at $x_i = i, i = 0, \ldots, 7$ near the estimation point $x = 3.4$ with two different smoothing parameters using a Gaussian weight function. The length of the thick solid lines between the weight function and the horizontal axis correspond to the weights of the observations Y_i. In the upper panel $h = 0.8$ while in the lower panel $h = 2$.

with the degree of locality determined by h. This differs from global estimation of m where one single estimation procedure is performed and the fit at x is calculated by substituting in the value x into the global fitted function. The weight function κ is usually selected so that

$$\kappa(x, x_i, h) = K(|x_i - x|/h) = K((x_i - x)/h)$$

and this is also what we will do. Note that the weight of Y_i can also be value of some functional of $K((x_i - x)/h)$ that depends on the particular estimator. This is the case for example in Gasser-Müller estimator [10]. The Gaussian probability density function

$$K((x_i - x)/h) = \frac{1}{\sqrt{2\pi h^2}} e^{-\frac{(x_i - x)^2}{2h^2}} \tag{2}$$

is commonly used as a weight function although it is not the best choice in the sense of the mean integrated squared error $\mathbb{E} \int \{\hat{m}(x; h) - m(x)\}^2 dx$, see e.g. [43]. For other kernel functions, see [28, 43]. The class of functions from which the estimate is calculated can in principle be anything, but it is practical to choose some relatively small set of functions for which the computational burden stays moderate. The most popular choice is the class of polynomial functions of certain order p, in particular constant ($p = 0$) or linear functions ($p = 1$). With $p = 0$ one obtains the classical Nadaraya-Watson estimator,

$$\hat{m}_{\mathrm{NW}}(x; h) = \frac{\sum_{i=1}^{n} K((x_i - x)/h) Y_i}{\sum_{i=1}^{n} K((x_i - x)/h)},$$

where $\hat{m}_{\mathrm{NW}}(x; h)$ is the best constant fit at x that is, the weighted average of observations, see [32, 44]. For other local constant estimators, see e.g. [7, 43]. Note that the simple running mean smoother can be seen as an another example of local constant regression with the weight function giving symmetrically equal weights to observations inside the smoothing window.

Another popular choice is local linear regression ($p = 1$), where a first order polynomial $t \mapsto b + a(t - x)$ is fitted using weighted observations. The resulting estimator $\hat{m}_{\mathrm{LLR}}(x; h)$ at x is the constant $\hat{b}$ of the fitted polynomial and it can be written in a closed form as

$$\hat{m}_{\mathrm{LLR}}(x; h) = \frac{1}{n} \sum_{i=1}^{n} \frac{[s_2(x; h) - s_1(x; h)(x_i - x)] \, K((x_i - x)/h) Y_i}{s_2(x; h) s_0(x; h) - s_1(x; h)^2},$$

where

$$s_q(x;h) = \frac{1}{n}\sum_{i=1}^{n}(x_i - x)^q K((x_i - x)/h).$$

The upper panel of Figure 2 shows two local linear smooths of a simulated data set with $h = 0.5$ and $h = 60$. The data are

$$Y_i = 5\sin(X_i) + 0.5X_i + \varepsilon_i, \tag{3}$$

with $X_i = i, i = 0, \ldots, 50$, and independent $\varepsilon_i \sim \mathrm{N}(0, 7^2)$. Note that the local linear estimate converges to the (global) least squares linear fit as $h \to \infty$, since the weights become equal regardless of the distance from the estimation point.

2.2 Smoothing splines

Smoothing splines are usually defined by finding the function $\hat{m}_{\mathrm{SS}}(\cdot;\lambda)$ that minimizes

$$(\mathbf{y} - \mathbf{m})^T(\mathbf{y} - \mathbf{m}) + \lambda \int_a^b (m^{(d)})^2. \tag{4}$$

Here m is assumed to have an absolutely continuous derivative $m^{(d-1)}$, $a \leq x_1 < \cdots < x_n \leq b$, $\lambda \geq 0$, and we have used the shorthand notation $\mathbf{m} = [m(x_1), \ldots, m(x_n)]^T$. The common choice is $d = 2$ and it is also used throughout this work. In the lower panel of Figure 2 two smoothing splines are fitted to the data set (3) using values $\lambda = 0.6$ and $\lambda = 50$.

For the analysis of piecewise constant or discrete signals we extend the term smoothing spline to cover also a discrete version of the above where, with $d = 2$, we minimize

$$(\mathbf{y} - \mathbf{m})^T(\mathbf{y} - \mathbf{m}) + \lambda\mathbf{m}^T\mathbf{C}^T\mathbf{C}\mathbf{m}, \tag{5}$$

and where $\mathbf{C}$ is a matrix that defines second order differencing [see section 2.1 of Paper IV].

Both of the above minimization problems can also be seen as regularization procedures, where the λ-term controls the regularity of the solution [34]. This is practical since in many cases the unregulated procedure would produce numerically

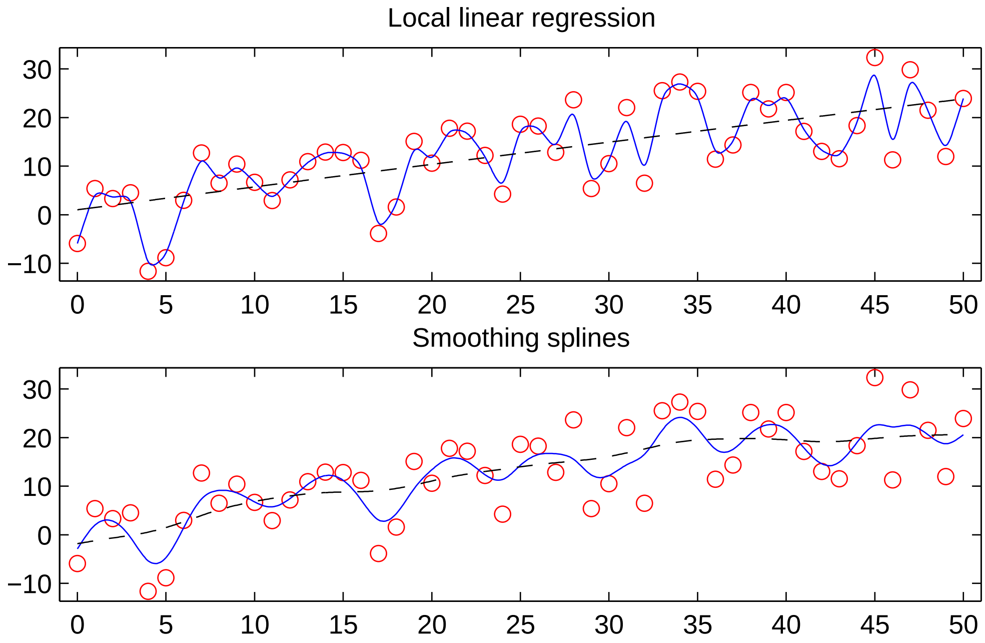

Figure 2: Smoothing of the simulated scatter plot data (3) shown as circles using two different methods and two levels of smoothing. Panel A: Smoothing with local linear regression using Gaussian weights and the smoothing parameter values $h = 0.5$ (solid curve) and $h = 60$ (dashed curve). Panel B: Smoothing splines using the values $\lambda = 0.6$ (solid curve) and $\lambda = 50$ (dashed curve).

unstable solutions with unrealistically large variation and with poor prediction performance. Two other popular methods that use such a regularization idea are ridge regression and the penalized likelihood method; see e.g. [18] and [17].

2.3 Smoothing parameter selection

Both of the above smoothing methods use a smoothing parameter, h in local regression and λ in the smoothing spline. The smoothing parameter controls the roughness of the smooth and therefore has a crucial influence on the inference about m. This is graphically shown in Figure 2. Sometimes, when we have prior information about the behavior of m, especially about its roughness or the number and/or the locations of its maxima and minima, a subjective trial-and-error procedure for the selection of the parameter may suffice. However, when no information is available or when one does not want to use it, we are usually forced to use some sort of automatic data driven smoothing parameter selection. Especially for kernel methods, there exists a huge number of results on the best choice of h in different configurations and contexts; see e.g. [17, 21, 37, 43] and the references therein. Most of the results are based on asymptotic analysis and thus might not work in practice. Another, more pragmatic solution, is to use cross-validation (CV), where one part of the sample at a time is used as a prediction set and the parameter value that provides the best prediction error averaged over the different prediction sets is chosen. This approach is reasonable albeit sometimes rather computationally intensive. It may also happen that the value of λ produced is very inconsistent with our prior beliefs about the underlying m. An example of this can be found in Section 4 of [Paper IV]. See, also [33].

3 The Bayesian paradigm

Local linear regression and kernel methods in general, as well as smoothing splines are usually regarded as nonparametric methods. An alternative to this is the parametric approach where distributional assumptions about the data and other unknown variables are stated in a parametric form. In function estimation, in addition to the parameters of the distribution of the data, the variables of interest can be the vector $\mathbf{m}$ of the values $m(x_i)$ or the parameters of the function m, for instance the slope and the constant term of a linear fit. In this work we are mainly interested in the function values $m(x_i)$. The overall inference about the parameters is based on their posterior distribution which arises from combining the data likelihood and the prior information. The likelihood describes the assumed relationship between the parameters of the data generating process and the observed data themselves. In the prior, our beliefs about the possible and probable values of the unknown quantities are encoded in a parametric distributional form. The combination of the likelihood and the prior is done according to Bayes rule that gives the posterior distribution of the parameters $\boldsymbol{\theta}$ given the observed data $\mathbf{y}$ as

$$p(\boldsymbol{\theta}|\mathbf{y}) = \frac{p(\mathbf{y}|\boldsymbol{\theta})p(\boldsymbol{\theta})}{p(\mathbf{y})}.$$

Thus, here $p(\boldsymbol{\theta}|\mathbf{y})$ denotes the posterior, $p(\mathbf{y}|\boldsymbol{\theta})$ the likelihood, $p(\boldsymbol{\theta})$ the prior and $p(\mathbf{y}) = \int p(\mathbf{y}|\boldsymbol{\theta})p(\boldsymbol{\theta})d\boldsymbol{\theta}$, the unconditional distribution of the data which serves as a normalizing factor. Such a parametric approach is called Bayesian, due to the above formula used to calculate the posterior. The normalizing factor is ignored when possible and the posterior distribution is often written using the proportionality sign $\propto$ as

$$p(\boldsymbol{\theta}|\mathbf{y}) \propto p(\mathbf{y}|\boldsymbol{\theta})p(\boldsymbol{\theta}),$$

because the posterior distribution as a function of $\boldsymbol{\theta}$ does not depend on $p(\mathbf{y})$. In real world problems with many unknowns an explicit formula rarely exist for $p(\boldsymbol{\theta}|\mathbf{y})$ and we have to resort to numerical methods to draw a sample from it. Then various Monte Carlo methods are typically used, see e.g. [11,14,35,36].

One of the fundamental ideas of this work is to use a penalty approach similar to smoothing splines in the prior structure for $\mathbf{m}$. In the discrete case we take

$$p(\mathbf{m}|\lambda) \propto \lambda^{\frac{n-2}{2}} \exp\left[-\frac{\lambda}{2}\sum_{i=2}^{n-1}\left(\frac{m_{i+1}-m_i}{x_{i+1}-x_i} - \frac{m_i - m_{i-1}}{x_i - x_{i-1}}\right)^2\right], \qquad (6)$$

where we have used the shorthand notation $m_i = m(x_i)$. For the errors $\boldsymbol{\varepsilon} = [\varepsilon_1, \ldots, \varepsilon_n]^T$ we make the common assumption $\boldsymbol{\varepsilon} \sim \mathrm{N}(0, \boldsymbol{\Sigma})$, where $\boldsymbol{\Sigma}$ is a positive definite covariance matrix of size $n \times n$. Defining $\mathbf{C}$ so that the exponent in (6) is $-(1/2)\lambda\mathbf{m}^T\mathbf{C}^T\mathbf{C}\mathbf{m}$ and assuming (1), the posterior can be written as

$$p(\mathbf{m}|\mathbf{y}, \boldsymbol{\Sigma}, \lambda) \propto \exp\left[-\frac{1}{2}\left((\mathbf{y}-\mathbf{m})^T\boldsymbol{\Sigma}^{-1}(\mathbf{y}-\mathbf{m}) + \lambda\mathbf{m}^T\mathbf{C}^T\mathbf{C}\mathbf{m}\right)\right].$$

Note that maximizing this is equivalent to minimizing a weighted form of (5). The minimizer is in fact the Bayesian maximum a posteriori (MAP) estimate when normal likelihood and the smoothing prior (6) are used [22]. Of course, the Bayesian approach provides more information since, in addition to a point estimate, it in fact provides the whole distribution of $\mathbf{m}|\mathbf{y}, \boldsymbol{\Sigma}, \lambda$. When the prior (6) is used the distribution is in fact multivariate normal. In the Bayesian approach the smoothing parameter λ and the covariance matrix $\boldsymbol{\Sigma}$ can also be treated as random. In standard smoothing spline literature one uses a homoscedastic model with the actual value of the variance usually fixed, ignored or estimated from the data [5,40]. In the random design case the explanatory variables can be treated as random variables, too (see [Paper V]).

4 Multiscale analysis and SiZer

As discussed above, the selection of a smoothing parameter is not a straightforward task. In fact, in many applications the use of one optimal value of a smoothing parameter may be insufficient. The underlying signal may exhibit interesting behavior in many different scales and a single smooth may not capture all features. Also, the features always have some characteristic scale or a range of scales at which they are reasonably defined. An example of the use of different scales is given in Figure 3 where three different views of an object are shown. At the largest scale one detects only the main features of the underlying signal, while decreasing the scale more details can be seen. At the smallest scale single local features such as individual leaves can be seen.

In practical data analysis, the scale at which the behavior of the unknown function should be investigated can in fact be one of the unknowns posing a problem that is not mathematically well defined. During the last few decades the problem of scale selection has been studied intensively in computer science, especially in computer vision. An important goal has been to provide visual signal processing abilities to machines and robots to facilitate visually guided navigation and object recognition [26].

A natural solution to the problem of scale selection is offered by multiscale analysis, where the unknown objects are investigated with different resolutions. On a very general level this approach is thought to have an interesting connection to the mammalian vision mechanism [20, 45, 46]. In the multiscale approach the signal is investigated with a large number of different scales ranging from the smallest reasonable to the largest at which the features are reasonable defined. More generally, the object of multiscale analysis can also be the outputs of functions, algorithms etc. [13]. If necessary, the selection of one or a few scales can be made after multiscale analysis.

During the last few years methods based on the multiscale approach have gained a lot of popularity in statistics and various fields of applications. This is in particular due to new useful visualization tools. Early ideas in multiscale data analysis include

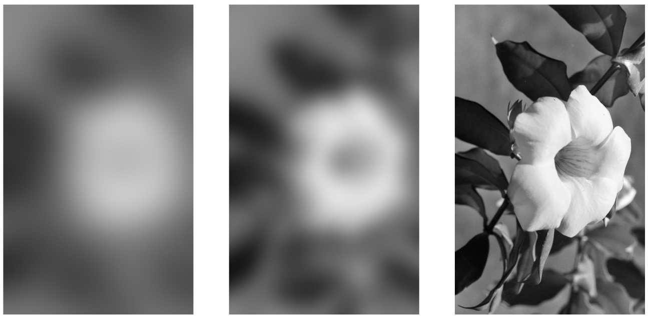

Figure 3: Three different views on an object. The leftmost image uses a very large scale, while in the center image a moderate scale has been used. The original image is on the right.

mode hunting and mode trees [8, 27, 29–31]. Recently, the multiscale idea has also been succesfully applied to classification and pattern recognition [12, 13].

Above we have used the terms scale and resolution in the same meaning. Although common usage in the literature, this is strictly speaking slight abuse of terms. The term 'resolution' is used in the somewhat similar idea of multiresolution analysis of wavelets and it was also employed in the early versions of multiscale representations of images, such as quad-trees and pyramids; see e.g. [19] and [2, 24]. Resolution in those contexts is a quantity that can be changed only in discrete steps and it can in principle be increased indefinitely. This, however, is not suitable for our purposes because we define the scale, or resolution, as the effective length of the smoothing window, a quantity assumed to be continuous.

4.1 Scale space analysis

Scale space analysis used in computer vision is one particular application of the general idea of multiscale analysis. It uses multilevel smoothing with a continuous smoothing parameter and with a constant sampling resolution in the explanatory variables at all scales. In our own approach we also require that the smooths change continuously with scale, that is, the smooths with different scales should be consistent with each other. In scale space analysis one assumes a convolution type smoothing of the form

$$\hat{m}(\cdot, h) = g(\cdot, \mathbf{X}, h) * \mathbf{y}, \tag{7}$$

where $*$ denotes the convolution operation between $\mathbf{y}$ and a weight function g which is similar to κ in kernel smoothing [26]. It can be shown that the Gaussian weight function (2) is the unique weight function that produces a decreasing number of modes in the smooth with increasing h [25].

The definition (7) covers many types of smoothing, in particular local polynomial regression. With this in mind, local linear regression with a Gaussian weight function and a large set of smoothing parameters can be viewed as an instance of scale space analysis. This is what is done in the SiZer method discussed next.

4.2 SiZer

One of the mainstays of this thesis is the idea of SiZer (Significant Zero Crossings of Derivatives), where the trends in the smooth are inspected using various different levels of smoothing [3,4]. Investigation of trends is a reasonable approach to finding the features in the data as they are the natural indicators of changes in the signal and they also directly identify the local minima and maxima. SiZer combines the analysis of trends with the idea of scale space smoothing and summarizes inferences about the slopes of the smooths in what is called a SiZer map. From this map inferences can be made by simple visual inspection making analysis of features accessible also to persons without a degree in statistics.

A crucial advantage of SiZer over traditional smoothing based approaches to finding features in data is that SiZer avoids the bias problem present in nonparametric function estimation. This is achieved by analyzing the features of the smooths and not those of the "true" underlying function [3,43]. Note the small but fundamental difference to smoothing the original data and focusing on the features of the resulting smooth.

The original SiZer of Chaudhuri and Marron uses local linear regression smoothing. In our Bayesian approach to multiscale smoothing we use a roughness penalty approach similar to smoothing splines (cf. Sections 3 and 2.2). The basic idea of investigating the trends of the smooths and visualizing the results with a map remains the same. This Bayesian version of SiZer is called BSiZer and the associated visualization is called the BSiZer map. Next we describe the basic steps in drawing a BSiZer map. For more details on map constructions, see [3] and Papers IV and V.

First, a level of credibility $\alpha > 0.5$ is chosen. A slope (a difference quotient in the discrete case) of the smooth at smoothing level $\lambda > 0$ is said to be significantly positive (negative) if its probability of being positive (negative) is at least α. Significantly positive or negative slopes are then marked with color blue or red, respectively. Nonsignificant features are marked with color gray. This is done with a range of values of λ and the pixels thus colored constitute the BSiZer map. Inferences about the significant trends in the smooths can then be made by simple visual inspection of colors. Note that the choice of colors is, of course, just a matter of convenience and in some applications a different convention is perhaps more intuitive [Paper II]. In Paper IV we further proposed to use tints of red and blue colors to encode the magnitude of the posterior probability of the signs of the slopes instead of binary style decisions. A BSiZer map of the simulated data set (3) is shown in the lower panel of Figure 4. In this example a continuous m is assumed and the penalty term $\lambda \mathbf{m}\mathbf{C}^T\mathbf{C}\mathbf{m}$ of the discrete case is replaced with $\lambda \mathbf{m}\mathbf{C}^T\mathbf{R}^{-1}\mathbf{C}\mathbf{m}$, where matrix $\mathbf{R}$ is defined on p.13 of [17]. At the smallest values of λ the map hints at a periodic nature of the underlying curve. The overall increasing trend is also clearly visible with the largest scale of smoothing. In the map, pointwise or independent features are shown, that is, the inferences may not be valid simultaneously for all values of the explanatory variable with the selected level of credibility α. For details on different hypothesis testing procedures and related concepts, see e.g. [3, 38] and Papers IV and V.

In the SiZer and BSiZer maps, for each each level of smoothing, there is a certain smoothing method dependent degree of localness at which the inferences can be thought to be made. In many applications this information can be of great value in the inference phase of the data analysis problem. For local linear regression with

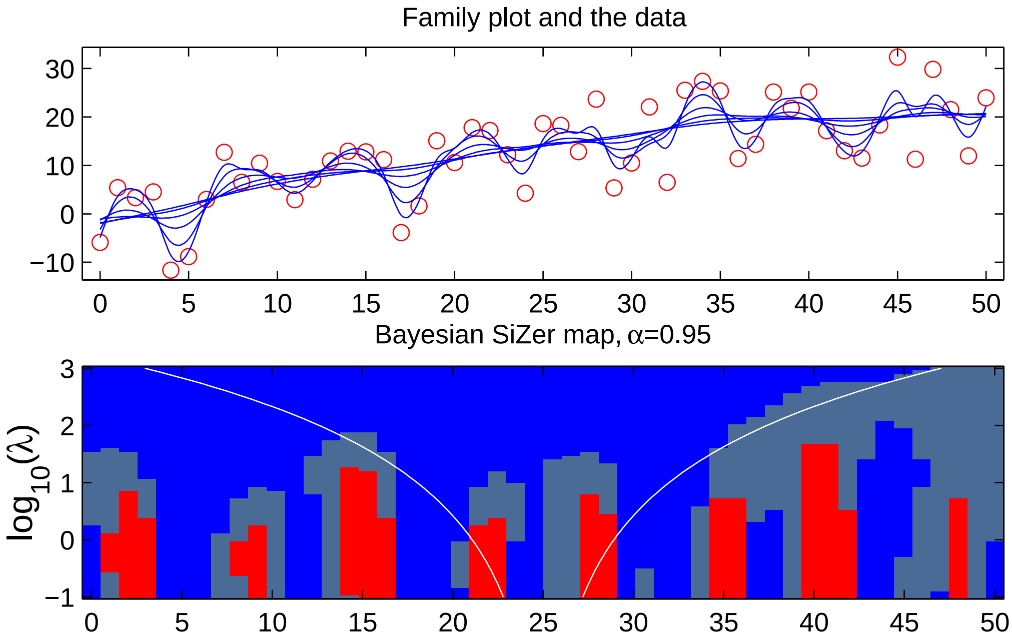

Figure 4: Upper panel: Family plot with 6 different smooths together with the original data shown as circles. Lower panel: The corresponding BSiZer map with credibility level $\alpha = 0.95$. The horizontal distance between the white curves shows the width of the effective smoothing window for different values of λ.

the Gaussian weight function the degree of localness is usually defined as $4h$. For smoothing splines the definition is it a bit more difficult due to the implicit nature in which the smoother is defined. However, here asymptotic results on the equivalence of spline smoothing and kernel regression can be used [39]. The effective degree of localness, that is, the effective smoothing window width at each level of smoothing is often depicted as the horizontal distance between two curves drawn in the BSiZer map. An example is shown in the lower panel of Figure 4. Also the so-called family plot, a plot of smooths with different levels of smoothing, is usually shown together with the map to facilitate the analysis of features in different scales. Such a plot is shown in the upper panel of Figure 4 together with the original data.

In addition to signs of slopes or difference quotients, the SiZer idea can also be used to make inferences about other properties of the signal such as the smooth itself or its higher order derivatives. One straightforward extension of the current work would be to test the significance of the magnitude of the slopes. A mouse-driven GUI could provide a convenient tool for inference by allowing one to select a magnitude of threshold and then drawing the corresponding map. Other extensions are discussed in [3]. One drawback of SiZer is that the visualization of inferences of higher dimensional data becomes difficult. Some approaches are suggested in [15,16]. The idea of level set trees could also be used in the generalization to the multivariate case [23].

5 Applications

The range of applications for the proposed methods is wide, extending from different areas of science and technology to business and other areas. The fixed design regression configuration described above can also be extended to the random design case as well as to observations with correlated error structures [Paper V]. In addition, the methods can easily be extended to situations with more than one observation of each $m(x_i)$. A central application in the development of the methods of this thesis has been the analysis of the reconstructed temperatures during the Holocene, a roughly 10000 year period in the Earth's history extending from the end of the last ice age to the present. Here the features of the smooths of the past temperature can be interpreted as representing temperature changes in different time scales.

Understanding how temperature in the past has changed over different time scales provides perspective for evaluating the present change of our environment.

Another interesting question is to consider the temperature reconstruction method itself. This is done in Papers I, II and VI. The typical set-up is that modern temperatures together with modern and historical values of a suitable response variable are first collected. In our case the geographical area of interest has been northwest Finnish Lapland and chironomid or diatom assemblages have been used as response variables. Other possible indicators are e.g. tree rings. The modern responses and their dependence on the temperature are assumed to be similar to past responses and their dependence on past temperatures. Other variables affecting the responses are ignored or assumed to be constant. Of course, the environmental variable under interest could be other than temperature, e.g. pH or alkalinity. For more details on the reconstruction problem see [1].

6 Summaries of the original papers

6.1 Paper I, Using Smoothing to Reconstruct the Holocene Temperature in Lapland

A nonparametric inverse approach is used to regress the past temperatures in the Finnish Lapland. First a dimension reduction is performed for the pre-processed explanatory variables using principal component analysis or partial least squares. Then local linear regression based on modern training data is used to predict values of past temperatures based on past values of the explanatory variable. The best reduced dimension and smoothing parameter in local linear regression are sought using leave-one-out cross-validation. The responses used are relative taxon abundances in diatom assemblages. The reconstructions obtained are in line with other studies. Besides assuming a similar temperature dependence for the modern and the past response variables, no other environmental assumptions are made.

6.2 Paper II, Quantitative Holocene Climatic Record from Diatoms in Northern Fennoscandia

The SiZer method of Chaudhuri and Marron is applied to the analysis of regressed Holocene temperatures with diatom taxon abundances as the response variable. The climatological events suggested by the SiZer map coincide well with events considered in other studies but which usually discuss them without proper statistical analysis.

6.3 Paper III, Making Inferences about Past Environmental Change Using Smoothing in Multiple Time Scales

Modifications of the original SiZer are discussed in the context of different assumptions about the modern training set and the prediction set of past responses. Different confidence intervals are derived when the training and prediction sets are alternatively fixed or random.

6.4 Paper IV, Bayesian Multiscale Smoothing for Making Inferences about Features in Scatter Plots

A Bayesian version of SiZer (BSiZer) for the regression case is proposed assuming a fixed design and independent errors. The smoothing parameter of the prior can be treated either as fixed or random. The conceptually simple Bayesian approach leads to nice closed form posteriors without any approximations. The data analysis examples demonstrate the practical usefulness of the new method. A new sampling based greedy algorithm for calculating the simultaneous confidence bands of difference quotients is proposed. Use of tints of red and blue in the BSiZer map is used to indicate the credibility of significant trends. A Matlab BSiZer software package is made publicly available.

6.5 Paper V, Bayesian analysis of features in a scatter plot with dependent observations and errors in predictors

In this paper BSiZer is extended to non-fixed designs and correlated errors. For explanatory variables, a normal likelihood together with an application specific prior distribution on their true values is assumed. General covariance matrix $\mathbf{\Sigma}$ is used for the error structure. For random $\mathbf{\Sigma}$ an inverse Wishart prior distribution is used. Further, a modification for a continuous regression function is described. An additional degree of simultaneousness in the construction of the BSiZer map is also considered. Matlab software is provided.

6.6 Paper VI, Selection of Prior Distributions and Multiscale Analysis in Bayesian Temperature Reconstructions Based on Fossil Assemblages

Two Bayesian temperature reconstruction models proposed earlier, the Poisson and the multinomial models, are compared. With the used data, both proposed models seem to be sensitive to the selection of prior for the past temperatures. In addition, the use of modern temperatures as informative priors of the past is questioned and a non-informative smoothing prior is proposed. Bayesian SiZer methodology is applied directly to the posterior distribution of the past temperatures. Software for this "model within BSiZer" approach is provided.

References

[1] H. J. B. Birks. Quantitative palaeoenvironmental reconstructions. In D. Maddy J. S Brew, editor, *Statistical modelling of Quaternary Science Data, Technical Guide 5*, pages 161–254. Quaternary Research Association, Cambridge, 1995.

[2] P. J. Burt. Fast filter transforms for image processing. *Computer Vision, Graphics, and Image Processing*, 16:20–51, 1981.

[3] P. Chaudhuri and J. S. Marron. SiZer for Exploration of Structures in Curves. *Journal of the American Statistical Association*, 94 (447):807–823, 1999.

[4] P. Chaudhuri and J. S. Marron. Scale space view of curve estimation. *Annals of Statistics*, 28:408–428, 2000.

[5] C. de Boor. *Practical Guide to Splines.* Springer Verlag, 1978.

[6] R. Eubank. *Nonparametric Regression and Spline Smoothing.* Statistics, a Series of Textbooks and Monographs. Marcel Dekker, second edition, 1999.

[7] J. Fan and I. Gijbels. *Local Polynomial Modelling and Its Applications*, volume 66 of *Monographs on Statistics and Applied Probability.* Chapman & Hall, 1996.

[8] N. I. Fisher, E. Mammen, and J. S. Marron. Testing for Multimodality. *Computational Statistics and Data Analysis*, 18:499–512, 1994.

[9] K. Fukunaga. *Introduction to Statistical Pattern Recognition.* Academic Press, second edition, 1990.

[10] T. Gasser and H.-G. Müller. Kernel estimation of regression functions. In T. Gasser and M. Rosenblatt, editors, *Smoothing Techniques for Curve Estimation*, pages 23–68. Springer-Verlag, Heidelberg, 1979.

[11] A. Gelman, J. B. Carlin, H. S. Stern, and D. B. Rubin. *Bayesian data analysis.* Texts in Statistical Science. Chapman & Hall/CRC, second edition, 1995.

[12] A. Ghosh, P. Chaudhuri, and D. Sengupta. Multi-scale kernel discriminant analysis. In D. P. Mukherjee and S. Pal, editors, *Proceedings of 5th International Conference on Advances in Pattern Recognition ICAPR-03*, pages 89–93, Kolkata, India, 2003. Allied Publishers.

[13] A. Ghosh, P. Chaudhuri, and D. Sengupta. Classification Using Kernel Density Estimates: Multi-scale Analysis and Visualization. To appear in Technometrics, 2005.

[14] W. R. Gilks, S. Richardson, and D. J. Spiegelhalter. *Markov Chain Monte Carlo in Practice.* Chapman & Hall, 1995.

[15] F. Godtliebsen, J. S. Marron, and P. Chaudhuri. Significance in scale space for bivariate density estimation. *Journal of Computational and Graphical Statistics*, 11:1–22, 2002.

[16] F. Godtliebsen, J. S. Marron, and P. Chaudhuri. Statistical significance of features in digital images. *Image and Vision Computing*, 22:1093–1104, 2004.

[17] P. J. Green and B. W. Silverman. *Nonparametric Regression and Generalized Linear Models. A roughness penalty approach*, volume 58 of *Monographs on Statistics and Applied Probability.* Chapman & Hall, 1994.

[18] A. E. Hoerl and R. W. Kennard. Ridge regression: Biased estimation for nonorthogonal problems. *Technometrics*, 12, 1970.

[19] W. Härdle, G. Kerkyacharian, D. Picard, and A. Tsybakov. *Wavelets, Approximation, and Statistical Applications*, volume 129 of *Lecture Notes in Statistics*. Springer, 1998.

[20] J. Jones and L. Palmer. An Evaluation of the Two-dimensional Gabor Filter Model of Simple Receptive Fields in Cat Striate Cortex. *Journal of Neurophysiology*, 58:1233–1258, 1987.

[21] M. C. Jones, J. S. Marron, and S. J. Sheather. A brief survey of bandwidth selection for density estimation. *Journal of the American Statistical Association*, 91:401–407, 1996.

[22] G. Kimeldorf and G. Wahba. A correspondence between Bayesian estimation on stochastic processes and smoothing by splines. *Annals of Mathematical Statistics*, 41:495–502, 1970.

[23] J. Klemelä. Visualization of Multivariate Density Estimates with Level Set Trees. *Journal of Computational and Graphical Statistics*, 13(3):599–620, 2004.

[24] A. Klinger. Pattern and search statistics. In J. S. Rustagi, editor, *Optimizing Methods in Statistics*. Academic press, 1971.

[25] J. J. Koenderink. The structure of images. *Biological cybernetics*, 50:363–370, 1984.

[26] T. Lindeberg. *Scale-Space Theory in Computer Vision*. Kluwer Academic Publishers, 1994.

[27] D. J. Marchette and E. J. Wegman. The Filtered Mode Tree. *Journal of Computational and Graphical Statistics*, 6:143–159, 1997.

[28] J. S. Marron and D. Nolan. Canonical kernels for density estimation. *Statistics & Probability Letters*, 7:195–199, 1988.

[29] M. C. Minnotte. Nonparametric testing of the existence of modes. *Annals of Statistics*, 25:1646–1660, 1997.

[30] M. C. Minnotte, D. J. Marchette, and E. J. Wegman. The bumpy road to the mode forest. *Journal of Computational and Graphical Statistics*, 7:239–251, 1998.

[31] M. C. Minnotte and D. Scott. The Tree Mode: a tool for visualization of nonparametric density estimates. *Journal of Computational and Graphical Statistics*, 2:51–68, 1993.

[32] E. A. Nadaraya. On estimating regression. *Theory of Probability and its Applications*, 9:141–142, 1964.

[33] B. U. Park and J. S. Marron. Comparison of data-driven bandwidth selectors. *Journal of the American Statistical Association*, 85:66–72, 1990.

[34] J. O. Ramsay and B. W. Silverman. *Functional Data Analysis.* Springer, 1997.

[35] C. P. Robert. *The Bayesian Choice.* Springer, 1994.

[36] C. P. Robert and G. Casella. *Monte Carlo Statistical Methods.* Springer, 1999.

[37] D. Ruppert, S. J. Sheather, and M. P. Wand. An effective bandwidth selector for local least squares regression. *Journal of the American Statistical Association*, 90:1257–1270, 1995.

[38] M. P. Salganik, M. P. Wand, and N. Lange. Comparison of Feature Significance Quantile Approximations. *Australian & New Zealand Journal of Statistics*, 46(4):569–582, 2004.

[39] B. W. Silverman. Spline smoothing: The equivalent variable kernel method. *The Annals of Statistics*, 12:898–916, 1984.

[40] B. W. Silverman. Some Aspects of the Spline Smoothing Approach to Nonparametric Regression Curve Fitting. *Journal of the Royal Statistical Society B*, 47(1):1–52, 1985.

[41] B. W. Silverman. *Density Estimation for Statistics and Data Analysis*, volume 26 of *Monographs on Statistics and Applied Probability.* Chapman & Hall, 1986.

[42] G. Wahba. *Spline Models for Observational Data.* Society for Industrial and Applied Mathematics, 1990.

[43] M. P. Wand and M. C. Jones. *Kernel Smoothing*, volume 60 of *Monographs on Statistics and Applied Probability.* Chapman & Hall, 1995.

[44] G. S. Watson. Smooth regression analysis. *Sankhya Series A*, 26:359–372, 1964.

[45] H. R. Wilson. Psychophysical evidence for spatial channels. In O. J. Braddick and A. C. Sleigh, editors, *Physical and Biological Processing of Images.* Springer-Verlag, New York, 1983.

[46] R. A. Young. The Gaussian Derivative Model for Spatial Vision: I. Retinal mechanisms. *Spatial Vision*, 2:273–293, 1987.