SiZer for exploration of structures in curves

P. Chaudhuri & J. S. Marron

October 7, 1997

Abstract

In the use of smoothing methods in data analysis, an important question is often: which observed features are "really there?", as opposed to being spurious sampling artifacts. An approach is described, based on scale space ideas that were originally developed in computer vision literature. Assessment of SIgnificant ZERo crossings of derivatives, results in the SiZer map, a graphical device for display of significance of features, with respect to both location and scale. Here "scale" means "level of resolution", i.e. "bandwidth".

1. Introduction

Smoothing for curve estimation in statistics is a useful tool for dsicovering features in data. Some examples of this are shown in Figure 1. For many more such examples, see e.g. the monographs Silverman (1986), Eubank (1988), Härdle (1990), Wahba (1991), Scott (1992), Green and Silverman (1994), Wand and Jones (1995) and Fan and Gijbels (1996).

[put figure 1 about here]

FIGURE 1: *Examples of features revealed by smoothing. Figure 1a shows kernel density estimates, with three different bandwidths h, for the 1975 Income data. Figure 1b shows a scatterplot and local linear regression estimates, with three different bandwidths h, for the Fossil data, with the raw data shown as small circles.*

Figure 1a is an example of density estimation, where the typical goal is to present a density f which reveals structure in univariate data $X_1, ..., X_n$. The kernel approach involves centering small pieces of probability mass (having a Gaussian shape here) at each data point, using the formula given in (2.1) below. As seen, the window width h controls the amount of smoothing. The data here are $n = 7211$ family incomes (rescaled so that the mean is 1) for the year 1975, from the Family Expenditure Survey in the United Kingdom. See Schmitz and Marron (1992) for detailed discussion and analysis of this data. Note that the midrange bandwidth, $h = 0.05$ shows two prominent modes, perhaps an indication of an economic "class structure"? However, these modes can made to disappear simply by using the larger bandwidth $h = 0.2$. Also many more modes, which are likely to be spurious sampling artifacts, can be made to appear by using the smaller bandwidth $h = 0.0125$. Which modes are "really there"? The detailed analysis in Schmitz and Marron (1992) reveals that the two important modes are (perhaps surprisingly) important features of this data set. That analysis also reveals an interesting shift in the size of these modes over time.

Figure 1b is an example of scatterplot smoothing, also called nonparametric regression estimation, where bivariate data $(X_1, Y_1), ..., (X_n, Y_n)$ are "smoothed" (e.g. by a moving average) to give a curve which can be viewed as an estimated conditional mean $f(x) = E(Y|X = x)$. The smooths actually used here are local linear smooths, with Gaussian weights, explicitly defined in (2.2) below. These have some preferable properties as summarized for example in the monographs of Wand and Jones (1995), and Fan and Gijbels (1996). Again the window width is crucial to the smooth, with $h = 0.3$ and 4.8 representing, respectively, substantial undersmoothing and oversmoothing. The data, provided by T. Bralower of the University of North Carolina, reflect global climate millions of years ago, through ratios of Strontium isotopes found in fossil shells. The shells are dated by biostratigraphic methods, see Bralower, et. al. (1997), so the Strontium ratio can be studied as a function of time. Both the scatterplots and the smooths have relatively high ratio for fossils less than 105 million years old, have a substantial dip with a minimum near 115 million years ago, and then perhaps an increase for fossils around 120 million years old. These features are shown nicely by the larger bandwidth $h = 4.8$. However, at the dip, this bandwidth seems to be substantially oversmoothing, so there is a good chance that it could be smoothing away some features that are "really there". The bandwidth $h = 1.2$ seems closer to "a reasonable amount of smoothing", and note that this suggests additional possible features, such as an increase from 92 to 95 million years ago, and perhaps a dip around 98 million years ago. But the significance of at least this last dip is quite suspect, since a look at the data shows that it appears to be based on only two isolated observations.

Both examples in Figure 1 illustrate a major hurdle in the practical use of smoothing methods: which features observed in a smooth are "really there"? Data analysts who are familiar with smoothing methods are usually very good at answering this question (although even for them "gray areas" exist, where quantification would be helpful), when they have the time for a careful trial and error approach. However, such analysts don't always have lots of time, and even worse such skilled people are all too often just not available. In this paper we propose a graphical device, the SiZer map, with two important benefits. First, it speeds up the process of deciding "which features are really there" for the experienced analyst, while at the same time quantitatively resolving "gray area" problems. Second, it allows even inexperienced analysts to make inferences about which features are "really there".

Our approach involves a view of smoothing, and the statistical inference problem at hand, which is radically different from most of the literature. The traditional approach is to focus on a "true underlying curve", and do inference about that. In particular, much work has been done on choosing the bandwidth from the data, and many proposals have been made for inference based on confidence intervals/bands. For reasons discussed in detail in Section 6.2, such inference has not been very useful, especially for the problem of finding important features. The main problem is that curve estimators suffer inherently from a bias that is hard to deal with. This bias is not present in classical parametric statistics where one operates under the assumption that a parametric model is "truth". We believe this is why attempts to extend the classical notion of parametric confidence intervals to smoothing seem to have not yielded the same useful results.

Our methodology is motivated by "scale space" ideas from computer vision, see Lindeberg (1994) for an introduction and detailed discussion. Our approach departs from the classical in two ways. First we simultaneously study a very wide range of bandwidths, avoiding the classical need to choose a bandwidth. This idea is not foreign to good data analysts, who know well that "different useful information can be available at different levels of smoothing". The *family approach* of Marron and Chung (1997) is one way of tapping into this information, but does not address the key question of which features are "really present".

Our second departure from the classical view, again following scale space ideas from computer vision, is that we avoid the bias problem (in doing inference) by shifting the focus from the "true underlying curve" to "the true curve, viewed at varying levels of resolution". In particular, our inference focuses on "smoothed versions of the underlying curve", with the idea that this "contains all the information that is available in the data" when working with that bandwidth. Detailed discussion of this view of smoothing is given in Section 2.

In Section 3 our main inferential tool, the SiZer map, is developed. This studies features simultaneously over both location, and "scale", i.e. bandwidth, by using a color map, as shown in Figure 2. The idea is to highlight significant features, such as bumps, by displaying where (with respect to both location and scale) the curve significantly increases and decreases. Note that significant bumps will be at *zero crossings of the derivative* between regions of significant increase and decrease. The name "SiZer" is a shortening of "SIgnificant ZERo crossings of derivatives". The color scheme is blue (red) in locations where the curve is significantly increasing (decreasing, respectively), and the intermediate color of purple is used where the curve cannot be concluded to be either decreasing or increasing. Here the term "location" is used in the scale space sense of *both* "x-location" and also "bandwidth location". Gray is used to indicate regions where the data are too sparse to make statements about significance, because there are not enough points in each window, as defined precisely in Section 3.

[put figure 2 about here]

FIGURE 2: *Combination of family plots (parts a and b), and SiZer maps (parts c and d) for the data sets in Figure 1, using level of significance* $\alpha = 0.05$. *Figures 2a and 2c on the left are for the Income Data, and the important bandwidth* $h = 0.05$ *is highlighted in both plots. Figures 2b and 2d are for the Fossil Data, and again the important bandwidth* $h = 1.2$ *is highlighted in both plots. The dotted curves in the SiZer maps show "effective window widths" for each bandwidth, as intervals representing* $\pm 2h$ *(i.e.* ± 2 *standard deviations of the Gaussian kernel).*

Note that for both sets of data, the family approach (in the top panels of Figure 2), reveals potential interesting structure, in addition to lots of likely spurious structure. Perhaps the worst spurious structure is in the Fossil data, where the smallest bandwidth smooth actually leaves the range of the data, around 95 and 97 million years ago. This is caused by data sparsity in that region, and is an unappealing feature of the local linear smoother. See Hall and Marron (1997) for detailed discussion, and access to the literature on various fixes that have been proposed. The SiZer maps (in the bottom panels of Figure 2) for each, make it clear which structure seen in the family plots is "statistically significant", and which cannot be separated from the natural variability.

For the Income data (Figure 2c), at very coarse levels of resolution (i.e. large bandwidths), the smooths are significantly increasing (shaded blue), and then significantly decreasing (shaded red), meaning that these features are "really there" *at this level of resolution.* For bandwidths near $\log_{10}(h) = -1.3$, the two modes become apparent, and these are both seen to be statistically significant, because the shading changes from blue ($\uparrow$) to red ($\downarrow$) to blue ($\uparrow$) to red ($\downarrow$). Hence, SiZer gives the same answer as was known to be correct from the separate analysis discussed above. SiZer furthermore suggests that the other features that can be seen in the family of smooths (including the three small bumps near the broader peak in the smooth with the thickest width in Figure 2a) are just sampling artifacts because the color is purple in these regions. The gray areas in each lower corner are where the data are too sparse for SiZer to be effective, as described in detail in Section 3.

For the Fossil data (Figure 2d), at the coarsest levels of resolution (largest bandwidths), the smooth is not far from a simple least squares fit line (since the window is extremely large), although not the same, because SiZer shows significant decrease up to around 105 million years ago, and then no significant change. For bandwidths that are less grossly oversmoothed, e.g. the bandwidth $h = 4.8$ (note that $\log_{10}(4.8) = 0.68$) shown in Figure 1b, the estimate has no significant slope on the left, is significantly decreasing in the center, and significantly increases on the right. When one looks at finer levels of resolution (smaller bandwidths) the curve is seen to be significantly increasing at around 93 million years ago. However, the dip in the thick curve of Figure 2b, about 97 million years ago, is shown to be spurious, because this feature is in the gray area, where there is not enough data to conclude that this dip is "really there".

These examples demonstrate the great potential of SiZer as a tool for data analysis. More examples to this effect, that also illustrate potential pitfalls are given in Section 4.

Our main ideas can easily be adapted to many different types of smoothing methods, such as smoothing splines, regression splines, or wavelets. But in this paper we concentrate on kernel - local polynomial smoothers, because of their simplicity and interpretability, and because of their very direct connection to the scale space ideas from computer vision. There are a variety of other types of extensions of this methodology, that are probably worth pointing out. These are discussed in Section 5.

Other approaches to inference of this type are discussed in Section 6. An important competitor is formal mode tests, reviewed in Section 6.1.

2. Scale space viewpoint

In this section we introduce precise notation, and give some discussion of the scale space view of smoothing. For much more on this, see Chaudhuri and Marron (1997).

Kernel density estimation uses a random sample $X_1, ..., X_n$ from a smooth probability density $f(x)$, to estimate f through

$$\widehat{f}_h(x) = \frac{1}{n}\sum_{i=1}^{n} K_h\left(x - X_i\right), \tag{2.1}$$

where h is the "bandwidth", i.e. smoothing parameter, and K_h is the "h-rescaling" of the kernel function K, $K_h(\cdot) = \frac{1}{h}K\left(\frac{\cdot}{h}\right)$. The main idea is to "put probability mass $\approx \frac{1}{n}$ near each X_i". As shown in Figure 1a, the bandwidth controls the amount of smoothing, $\widehat{f}_h(x)$ is wiggly when h is small, and very flat when h is large. See for example Silverman (1986), Scott (1992) and Wand and Jones (1995) for discussion of many important properties and aspects of this estimator.

The local linear regression estimate uses a random sample $(X_1, Y_1), ..., (X_n, Y_n)$ to estimate the conditional expected value, i.e. the regression function,

$$f(x) = E\left(Y_i | X_i = x\right),$$

through

$$\widehat{f}_h(x) = \underset{a}{\arg\min} \sum_{i=1}^{n} \left[Y_i - (a + b(X_i - x))\right]^2 K_h\left(x - X_i\right), \tag{2.2}$$

where $\underset{a}{\arg\min}$ is interpreted to mean "minimize jointly over a and b, but use the a value". The main idea is that for each x, a line is fitted to the data, using K_h - weighted least squares. Again the bandwidth controls the amount of smoothness of $\widehat{f}_h(x)$, as shown in Figure 1b. See e.g. the monographs of Wand and Jones (1996) and Fan and Gijbels (1995) for discussion of many properties and important aspects of this estimator.

Scale space ideas from computer vision provide a viewpoint on kernel smoothing that is new to statisticians. The "scale space surface", which is the family of all kernel smooths indexed by the bandwidth h, is a model used in computer vision. The essential idea is that large h models "macroscopic (distant) vision" where "only large scale features can be resolved", and small h models "microscopic (zoomed in) resolution of small scale features". In particular, for a given function f (i.e. underlying signal) various amounts of "blurring of the signal" (at least some is present in any real visual system) are represented by the convolution $f * K_h$ for different values of h. In fact this family of convolutions becomes the focus of the analysis, with the idea that this is all that is available from a finite amount of data in the presence of noise. See Lindeberg (1994) for details, and a large amount of interesting discussion. This is very different from the classical statistical approach, where the focus is f.

Examples of "features" in curve estimation, include peaks and valleys. These can be characterized in several ways. In this paper we focus on zero crossings of the derivative. We say a zero crossing is "significant" when the derivative estimate is significantly different from zero on both sides, with opposite signs, e.g. as shown by blue and red areas in Figures 2c and 2d.

When these zero crossings of the smooth derivative estimates are studied across a range of bandwidths, the Gaussian kernel $K(x) = \frac{1}{\sqrt{2\pi}} e^{-x^2/2}$ has an important advantage over other kernels. In particular, for convolution smoothers, the number of zero crossings of the derivative smooth is always a decreasing function of h (not true for any other kernel used for kernel smoothing). In other words, only Gaussian blurring has monotonicity of features with respect to the amount of smoothing. Several ways to see this are given in Lindeberg(1994). Interesting related references in the statistical literature include Silverman (1981) and Minnotte and Scott (1993). Hence, only the Gaussian kernel is used in this paper.

The main point of this paper is the development of color maps as shown in Figures 2c and d, called "SiZer maps", which can be used for exploratory data analysis, that show regions in scale space (i.e. with respect to both x and h) where the derivative is significantly increasing and decreasing. As discussed in Section 6.2, classical approaches to significance of features, based on confidence bands, either are much too conservative for useful inference, or else are grossly invalid, because of bias problems. In this paper we take a novel approach to this old bias problem, by adopting the scale space point of view. In particular, instead of seeking confidence intervals for $f'(x)$, we seek confidence intervals for the scale space version $f_h'(x) \triangleq E\widehat{f}_h'(x)$ (for regression we take this E to be conditional on

$X_1, ..., X_n$). The center point of such intervals is automatically "correct", and the variance is estimated simply and effectively, as detailed below. From this point of view "significance" of any feature depends on the scale of resolution (i.e. on h), and must be interpreted in that way. E.g. in Figure 2c, we see that the bimodal structure is present at some levels of resolution, but disappears at coarser levels (i.e. there is only one mode at large bandwidths).

Note that this approach is rather different from traditional mode testing. In particular, the SiZer map not only counts the number of significant modes, at different levels of resolution, but also gives information about mode locations. There is a trade off though, in that the SiZer map tends to be more conservative than mode tests which specifically target the number of modes, see Section 4.

3. Development of SiZer

Our approach to the visual assessment significance of features such as peaks and valleys in a family of smooths $\left\{\widehat{f}_h(x) : h \in [h_{\min}, h_{\max}]\right\}$, is based on confidence limits for the derivative in scale space, $f_h'(x)$ (choice of $h_{\min}$ and $h_{\max}$ is discussed in section 3.1). Behavior at x and h locations is presented via the SiZer color map where blue (black in versions where only black and white are available) indicates locations where $\widehat{f}_h'(x)$ is significantly positive, red (white in black-white versions) shows where $\widehat{f}_h'(x)$ is significantly negative, and purple (gray in black-white versions) indicates where $\widehat{f}_h'(x)$ is not significantly different from zero.

Because repeated calculation of smoothers is required for such color maps, fast computational methods are very important. Binned (also called "WARPed") methods are natural for this, because the data need only be binned once. See Fan and Marron (1994) for detailed discussion of this, and other fast computation methods. The main idea is that calculation of $\widehat{f}_h'(x)$ becomes a rapidly computed discrete convolution when the data are approximated by bin counts on an equally spaced grid, which can result in speed savings of factors of 100 (for larger sample sizes). For the reasons discussed in Fan and Marron (1994) we use $g = 401$ grid points for most examples in this paper, although in some situations other values can be desirable as discussed below.

Confidence limits for $f_h'(x)$ are of the form

$$\widehat{f}_h'(x) \pm q \cdot \widehat{sd}\left(\widehat{f}_h'(x)\right), \tag{3.1}$$

where q is an appropriate quantile, and the standard deviation is estimated as discussed in section 3.1. An (x, h) location (in scale space) is called significantly increasing, decreasing, or not significant, when zero is below, above or within these confidence limits, respectively.

Candidates for calculation of the quantile q include:

1. Pointwise, Gaussian quantiles: $q_1(h) = q_1 = \Phi^{-1}\left(1 - \frac{\alpha}{2}\right)$.

2. Approximate simultaneous over x, Gaussian quantiles: based on "number of independent blocks", defined as q_2 below.

3. Bootstrap simultaneous over x, defined as q_3 below.

4. Bootstrap simultaneous over x and h, defined as q_4 below.

While Gaussian approximations worked quite well (because smoothers are local averages), the pointwise quantiles q_1 are not recommended. This is because this version of the SiZer map suggests that too many features are "significant", as shown in Figure 3.

[put figure 3 about here]

FIGURE 3: *For a simulated data set, of size $n = 100$, from the Marron-Wand density #3: Figure 3a shows the family approach, overlaid with the true underlying density (thick yellow curve); Figure 3b shows pointwise SiZer; Figure 3c shows simultaneous SiZer with no gray shading for sparse data regions; Figure 3d shows approximate simultaneous SiZer with sparse data regions shown in gray.*

Each panel in Figure 3 is for the same simulated data set of size $n = 100$ from the density #3 of Marron and Wand (1992). This density is shown as the heavy yellow curve in Figure 3a. It is a mixture of eight normals, intended to reflect much of the structure present in the log normal distribution: a single large peak, with a very long right tail. As shown in the family of smooths, based on the single data set, in Figure 3a, this density is challenging to estimate. In particular, small window widths are most appropriate near the peak to avoid smoothing that down to too low a level, but large bandwidths are more sensible in the tail to smooth out the spurious clusters that arise just by chance. The pointwise SiZer map, shown in Figure 3b, incorrectly indicates that some of these spurious clusters are "significant", e.g. the peaks near $x = -1.7$, -1.4 and 0.6. The problem is understood via the classical frequentist interpretation of Confidence Intervals:

looking at many replications should result in roughly proportion α intervals which do not "cover the true value". A natural solution to the problem is to adjust the length of the intervals to do "simultaneous inference", which is the goal of the other approaches to q mentioned above, which are discussed in detail below. The approximate simultaneous approach is shown in Figure 3c, where these spurious modes are now shaded correctly as purple. However, this version has a curious red stripe in the lower left corner, that we have not fully understood. We have not carefully analyzed this, because it is in a region in scale space where the data are very sparse. Both because of effects like this, and also because we don't trust confidence intervals that are based on too few points, regions in scale space where the data are too sparse for meaningful inference are grayed out. Based on the classical rule of thumb, a location gray is shaded gray when the "effective sample size in the window" (defined below) is less than 5. This gives the map shown in Figure 3d.

Our first suggestion, q_2, for approximate simultaneous confidence limits is based on the fact that when x and x' are sufficiently far apart, so that the kernel windows centered at x and x' are essentially disjoint, the estimates $\widehat{f}_h'(x)$ and $\widehat{f}_h'(x')$ are essentially independent, but when x and x' are close together, the estimates are highly correlated. The simultaneous confidence limit problem is then approximated by m independent confidence interval problems, where m reflects the "number of independent blocks". We estimated m through an "estimated effective sample size", defined for each (x, h) as

$$ESS(x,h) = \frac{\sum_{i=1}^{n} K_h\,(x - X_i)}{K_h\,(0)}.$$

Note that when K is a uniform (i.e. boxcar) kernel, $ESS(x, h)$ is the number of data points in the kernel window centered at x. For other kernel shapes, points are downweighted according to the height of the kernel function, just as they are in the averages represented by the kernel estimators. Next we choose m to be essentially the number of "independent blocks of average size available from our data set of size n",

$$m(h) = \frac{n}{avg_x ESS(x,h)}.$$

Now assuming independence of these $m(h)$ blocks of data, the approximate simultaneous quantile is

$$q_2 = q_2(h) = \Phi^{-1}\left(\frac{1 + (1-\alpha)^{1/m}}{2}\right).$$

The quantity ESS is also useful to highlight regions where the normal approximation implicit in (3.1) could be inadequate. This plays a role similar to np in the Gaussian approximation to the Binomial. So regions where $ESS(x,h) < n_0$ (we have followed the standard practice of $n_0 = 5$ at all points here) are shaded gray, to rule out spurious features, and also to indicate regions where the smooth is essentially based on sparse data as shown in Figure 3d. The above calculation of the block size $m(h)$ is modified to avoid problems with small ESS as

$$m(h) = \frac{n}{avg_{x \in D_h} ESS(x,h)},$$

where D_h is the set of x locations where the data are "dense",

$$D_h = \{x : ESS(x,h) \geq n_0\}$$

These approximate simultaneous confidence limits are somewhat crude, and also are only simultaneous over x, not h. To improve them, we explored several classical multivariate normal simultaneous confidence sets (both elliptical and rectangular). These are based on the standard principal component analysis. Unfortunately, they tended to be far too conservative, because the orientation of the usual confidence sets along the eigenvector directions gave a region that did not efficiently project back to confidence intervals for $f_h'(x)$ for each x. The projections, i.e. the resulting confidence intervals, tended to be far too long to find important features.

Simultaneous confidence sets which are hypercubes, whose edges are parallel to the axes, with lengths of the form (3.1), are much better oriented to reflect significance of our derivative estimates. The direct calculation of the probabilities of such rectangular sets in high dimensions is very difficult for these highly correlated normal distributions. Since simulation is the only tractable approach, it is natural to use the more direct method of the bootstrap (i.e. simulate from the empirical distribution of the data, instead of from the approximating Gaussian). For each bootstrap sample (i.e. random sample drawn with replacement from the data, see Efron and Tibshirani (1993) for an introduction to bootstrap ideas), we compute $\widehat{f}_h'(x)^*$ (again a fast implementation is crucial), and the standardized version

$$Z^*(x,h) = \frac{\widehat{f}_h'(x)^* - \widehat{f}_h'(x)}{sd(\widehat{f}_h'(x))}.$$

For each h, the bootstrap quantile $q_3 = q_3(h)$ that is simultaneous over x (where the data are reasonably dense) is the empirical quantile of $\max_{x \in D_h} |Z^*(x,h)|$ calculated over the bootstrap replications. Similarly, the bootstrap quantile q_4, that is simultaneous over both x and h is the empirical quantile of $\max_h \max_{x \in D_h} |Z^*(x,h)|$ taken over the bootstrap replications.

Study of many SiZer maps based on q_2, q_3 and q_4 showed that in many cases there was not a lot of difference between the "quick and approximate" quantile q_2 and the bootstrap quantile q_3 that is simultaneous over x. As expected, somewhat fewer features generally appeared as significant for q_4, the bootstrap value that is simultaneous over both x and h, although surprisingly often q_4 was quite similar to q_2 and q_3. The maps based on different choices for q were most similar for examples that were "homogeneous in x", meaning either equally spaced regression, or else density estimation examples where the local average height of the density is roughly homogeneous. This is because there is an implicit "homogeneity assumption" made by q_2 that is a reasonable approximation in this case.

An example where this homogeneity is lacking (thus giving interesting differences) is the Income data set, from Figures 2a and 2c. SiZer maps based on the bootstrap quantiles q_3 and q_4 are shown in Figure 4.

[put Figure 4 about here]

FIGURE 4: *SiZer maps for the Income Data. Based on* 1000 *bootstrap replications. Quantiles are:* q_3 *pointwise over* h *in the top panels, and* q_4 *simultaneous over* h *in the bottom panels. Significance levels are* $\alpha = 0.05$ *in the left panels,* $\alpha = 0.10$ *in the center panels and* $\alpha = 0.20$ *in the right panels.*

The SiZer map for q_3 with $\alpha = 0.05$, shown in Figure 4a is fairly similar to that for q_2 shown in Figure 2c, except the lower right red region above $x = 0.4$ is quite a bit thinner. That red region actually disappears for the fully simultaneous SiZer map based on q_4 with $\alpha = 0.05$ shown in Figure 4d. This shows that the q_4 SiZer map can be rather conservative, because it does not show that there are two significant modes here (at the level $\alpha = 0.05$), although these have been verified by other means. But when the level of significance is raised to $\alpha = 0.10$ as shown in Figure 4e, the red region reappears, so both modes are now statistically significant in this sense. Note that for both q_3 and q_4, as α increases, the red and blue regions grow, as expected.

Because the bootstrap versions of SiZer are much slower to compute, we suggest using q_2 for a first look at the data. This version of SiZer is called SiZer1 in our software, available from the URL:

`http://hotelling.stat.unc.edu/faculty/marron/marron_software.html`

But when there are any doubts (there should be more doubts in settings that are not homogeneous in x), we recommend that q_3 and q_4 (implemented in SiZer5 in our software) be used for verification. While q_4 is our only procedure that gives a rigorous test of significance of features, it is also generally somewhat conservative, so we recommend that features found in q_2 or q_3 SiZer maps, that don't appear in the q_4 version, be independently investigated by a "mode testing method", see Section 6.1. For example, the mode test of Fisher and Marron (1997) shows that the existence of two modes in the Income data can be established with $\alpha < 0.01$, by a test which focuses explicitly on number of modes. David Scott has pointed out that the SiZer map can be viewed as an "enhancement" of the mode tree of Minotte and Scott (1993) .

Bootstrap theory suggests improvement by a "studentized modification", see e.g. Hall (1992), (or other methods, see e.g. Efron and Tibshirani (1993)). Such methods have not been implemented here, because they involve recalculation of the variance estimate for each bootstrap sample, which would entail substantial computational cost.

3.1. Numerical Implementation

The bandwidth range $[h_{\min}, h_{\max}]$ can be chosen in several ways. One approach is a "broad range of smooths which should catch most interesting features", as developed in the "family approach to smoothing" in Marron and Chung (1997). Another approach is "a very wide range of smooths", which is determined more by the curve estimation setting, than the data. In the examples of this paper, we have used the latter, and we took $h_{\min}$ to be the smallest bandwidth for which there is no substantial distortion in construction of the binned implementation of the smoother, $h_{\min} = 2 * (\text{binwidth})$, and took $h_{\max}$ to be the range of the data.

3.1.1. Density Estimation Specifics

The main idea behind the calculation of $\widehat{sd}$ in this context, is that the derivative estimator $\widehat{f}'_h(x)$ is an average (of the derivative kernel functions), so we use the corresponding sample standard deviation,

$$\begin{aligned} \widehat{var}\left(\widehat{f}'_h(x)\right) &= \widehat{var}\left(n^{-1}\textstyle\sum_{i=1}^n K'_h(x - X_i)\right) \\ &= n^{-1}s^2\left(K'_h(x - X_1), ..., K'_h(x - X_n)\right), \end{aligned}$$

where s^2 is the usual sample variance of n numbers.

Details of the binned implementation of $\widehat{f}_h'(x)$ are similar to those given in Fan and Marron (1994), except that the kernel is now replaced by the derivative of the kernel. In particular, for the equally spaced grid of points $\{x_j : j = 1, ..., g\}$, let the corresponding bincounts (computed by some method, we have always used the "linear binning" described in Fan and Marron (1994)) be $\{c_j : j = 1, ..., g\}$. Then

$$\widehat{f}_h'(x_j) \approx n^{-1}\overline{S}_0'(x_j),$$

where

$$\overline{S}_0'(x_j) = \sum_{j'=1}^{g} \kappa_{j-j'}' c_{j'} \tag{3.2}$$

and

$$\kappa_{j-j'}' = K_h'(x_j - x_{j'}). \tag{3.3}$$

To similarly approximate $\widehat{sd}$, use

$$\widehat{sd}(x_j) = n^{-1/2}\sqrt{n^{-1}\sum_{j'=1}^{g}\left(\kappa_{j-j'}'\right)^2 c_{j'} - \left(\widehat{f}_h'(x_j)\right)^2}.$$

3.1.2. Regression estimation specifics

We prefer the local linear smoother (to a number of other sensible smoothers) because the derivative estimate is the simple and appealing slope of the local line. See e.g. Wand and Jones (1995) and Fan and Gijbels (1996) for further discussion. See Fan and Marron (1994) for a fast binned implementation of the local linear smoother.

Our proposed $\widehat{sd}$ is motivated by the fact that the derivative estimator is a weighted sum of the observed responses, and we essentially use the conditional (given $X_1, ..., X_n$) weighted sample variances,

$$var\left(\widehat{f}_h'(x)|X_1, ..., X_n\right) = var\left(n^{-1}\sum_{i=1}^{n} W_h(x, X_i)Y_i|X_1, ..., X_n\right) = \sum_{i=1}^{n}\sigma^2\left(Y_i|X_i\right)\left(W_h(x,X_i)\right)^2.$$

To estimate $\sigma^2\left(Y_i|X_i\right)$ we use a simple smooth of the residuals, e.g.

$$\widehat{\sigma}^2(Y|X = x) = \frac{\sum_{i=1}^{n}\widehat{e}_i^2 K_h(x - X_i)}{\sum_{i=1}^{n} K_h(x - X_i)}.$$

where $\widehat{e}_i = Y_i - \widehat{f}_h(X_i)$.

The quotient rule form of the derivative estimate, based on differentiating the local linear estimate is particularly unpleasant, so we have a strong personal preference for the derivative estimate based on the slope of the linear fit:

$$\widehat{f}_h'(x) = \arg\min_b \sum_{i=1}^{n} \left[Y_i - (a + b(X_i - x))\right]^2 K_h\left(x - X_i\right), \tag{3.4}$$

where a similar convention to that of (3.4) is used. An efficient binned approximation of this local linear derivative estimate $\widehat{f}_h'(x)$ is

$$\widehat{f}_h'(x_j) \approx \frac{\overline{T}_1(x_j) - \overline{T}_0(x_j)\overline{X}(x_j)}{\overline{S}_2(x_j) - 2\overline{S}_1(x_j)\overline{X}(x_j) + \overline{S}_0(x_j)\overline{X}(x_j)^2},$$

where the notations

$$\begin{aligned} \overline{S}_\ell(x_j) &= \textstyle\sum_{j'=1}^{g} \kappa_{j-j'} c_{j'} x_{j'}^\ell, \\ \overline{T}_\ell(x_j) &= \textstyle\sum_{j'=1}^{g} \kappa_{j-j'} Y_{j'}^\Sigma x_{j'}^\ell, \\ \overline{X}(x_j) &= \overline{S}_1(x_j)/\overline{S}_0(x_j), \end{aligned} \tag{3.5}$$

have been used together with

$$\kappa_{j-j'} = K_h(x_j - x_{j'}). \tag{3.6}$$

and $Y_{j'}^\Sigma$ for the bin sums of the Y_i.

A binned approximation to $\widehat{\sigma}^2(Y|X = x_j)$, based on calculations familiar from simple linear regression is:

$$\widehat{\sigma}^2(Y|X = x_j) \approx \left(1 - \widehat{\rho}(x_j)^2\right) \widehat{\sigma}(x_j)^2,$$

where

$$\begin{aligned} \widehat{\sigma}(x_j)^2 &= \tfrac{\overline{U}_0(x_j)}{\overline{S}_0(x_j)} - \left(\tfrac{\overline{T}_0(x_j)}{\overline{S}_0(x_j)}\right)^2, \\ \widehat{\rho}(x_j)^2 &= \left(\widehat{f}_h'(x_j)\right)^2 \left(\tfrac{\overline{S}_2(x_j) - 2\overline{S}_1(x_j)\overline{X}(x_j) + \overline{S}_0(x_j)\overline{X}(x_j)^2}{\widehat{\sigma}(x_j)^2 \overline{S}_0(x_j)}\right), \end{aligned} \tag{3.7}$$

using the notation (3.5) and

$$\overline{U}_0(x_j) = \sum_{j'=1}^{g} \kappa_{j-j'} Y_{j'}^{2\Sigma},$$

for $Y_{j'}^{2\Sigma}$ denoting the bin sums of the Y_i^2. Our binned approximation to the conditional variance is now

$$var\left(\widehat{f}_h'(x_j)|X_1, ..., X_n\right) \approx \frac{\overline{V}_2(x_j) - 2\overline{V}_1(x_j)\overline{X}(x_j) + \overline{V}_0(x_j)\overline{X}(x_j)^2}{\left(\overline{S}_2(x_j) - 2\overline{S}_1(x_j)\overline{X}(x_j) + \overline{S}_0(x_j)\overline{X}(x_j)^2\right)^2},$$

where

$$\overline{V}_\ell(x_j) = \sum_{j'=1}^{g} \kappa_{j-j'} c_{j'} x_{j'}^\ell \widehat{\sigma}^2(Y|X = x_{j'}),$$

and the notations (3.5), (3.6) and (3.7) have been used. This results in

$$\widehat{sd}(x_j) = \sqrt{var\left(\widehat{f}_h'(x_j)|X_1, ..., X_n\right)}.$$

4. More applications and examples

In this section, additional examples illustrating both the usefulness of SiZer, and also some potential pitfalls, are presented.

The Hidalgo Stamp data set was brought to the mode testing literature by Izenman and Sommer (1988). This is a univariate data set consisting of the thicknesses of stamps, issued in Mexico during the last century. These thicknesses have a remarkable amount of variability and clustering, which suggests a number of sources for the paper. An interesting philatelic question is to determine the number of paper sources, which was addressed by Izenman and Sommers (1988) via nonparametric density estimation. Figure 5 analyzes these data with the family approach and SiZer.

[put Figure 5 about here]

FIGURE 5: *Family plot (part a) and SiZer maps, based on 401 grid points (part b), 81 grid points (part c), and 201 grid points (part d), for the Hidalgo Stamp data. The Sheather Jones Plug In bandwidth is the thick curve in the family plot, and corresponds to the highlighted horizontal bar. The SJPI bandwidth suggests seven modes, but not all are "significant" from the SiZer point of view.*

The thick curve in Figure 5a is a kernel density estimate using the Sheather Jones Plug In bandwidth, as recommended e.g. in Jones, Marron and Sheather (1996a,b). This suggests seven modes in the data (i.e. at least seven sources for the paper), which agrees with the findings of Izenman and Sommers (1988), and some others. The SiZer map in Figure 5b shows that the two largest modes, at 0.072 mm and 0.079 mm are indeed significant, as is the mode at 0.1 mm. The mode at 0.09 mm is less certain, as SiZer finds a significant increase on the left, but no significant decrease on the right. Similarly for the mode at 0.11 mm, where there is only a significant decrease. SiZer completely misses the modes at 0.12 mm and 0.13 mm, but the existence of these is perhaps debatable. If one has a priori knowledge that no paper source has a very wide variance, then one may be able to believe these are actual modes. However if one accepts the possibility of a heavy tailed distribution, then the family plot suggests these could be just random clustering in such a heavy tail. Also note the thick density estimate is heavily into the gray region of the SiZer map, which says the data are very sparse in this region, which also casts doubt on these modes, from this point of view.

Note that at the finest level of resolution (smallest bandwidth), the SiZer map in Figure 5b suggests the existence of more "modes" between 0.068 and 0.083. This is caused by the data being heavily rounded, to 0.001 mm, which results in many replicate values in regions where the data are dense. When such rounded data are binned to 401 bins over this range, i.e. a binwidth of 0.0002, there are a number of bins which receive no observations. When these bincounts, which alternate between zero and very large numbers (because of the rounding) are smoothed with a very small bandwidth, one gets a kernel estimate which significantly increases and decreases, as shown. In this sense, these feature are "really there", although the only conclusion is that the data have been rounded. We have seen this same phenomenon in other data sets. A natural solution is that in Figure 5c, where the number of gridpoints is reduced to $g = 81$, which makes each rounded data value a bin center. Unfortunately the heavy rounding in the data entails a SiZer map which misses some of the most interesting levels of smoothing, such as the Sheather Jones Plug In bandwidth. The problem is fixed in Figure 5d, by going to $g = 201$. Note that this SiZer map has all the same important features as in Figure 5b.

Next we study the performance of SiZer in some simulation settings, which highlight the way that SiZer "displays the information available in the data". The first of these is shown in Figure 6, where we study the effect of increasing sample size n, i.e. increasing "information in the data", in density estimation.

[put figure 6 about here]

FIGURE 6: *Top panels are family plots, and bottom panels are corresponding SiZer maps, for kernel density estimates, based on simulated data, from the Marron and Wand density #9, "Trimodal", shown as the thick yellow curve in the family plots. Sample sizes are* $n = 100$ *in parts a and c,* $n = 1000$ *in parts b and e and* $n = 10,000$ *in parts c and f. The thick red curve in the family plots is the Sheather Jones*

Plug In bandwidth, which is the highlighted horizontal bar in the SiZer maps.

The family plot, in combination with the Sheather Jones Plug In bandwidth, for $n = 100$ suggests no significant modal structure in the data. This is also reflected in the SiZer map. There is just not enough data to resolve even the two larger modes that are present in the underlying density. For $n = 1000$, the situation is different, and now the two large modes are clearly present in the data. More interesting is the third central mode. It is not clear from the family plot if this is significant, the Sheather Jones Plug In bandwidth suggests this is dubious, and the SiZer map confirms this is not significant. For $n = 10,000$ the family plot shows that we have a great deal of information about this density, and can estimate it extremely well. The SiZer plot verifies this, showing that all three modes are clearly present.

There are also some interesting overall trends present, that can be expected in general. For example, as n grows, the gray area diminishes, and tends to be replaced by purple. The purple areas also tend to be eventually replaced by either red or blue, both from below, and also in the boundary regions.

Increasing information in regression is investigated in Figure 7, but this time the "information in the data increases" through decreasing the error variance, rather than increasing sample size.

[put Figure 7 about here]

FIGURE 7: *Top panels are family plots, and bottom panels are corresponding SiZer maps, for local linear regression estimates, based on* $n = 200$ *simulated data, shown as green dots, from an equally spaced design, and the regression curve shown as the yellow thick curve in the top panels. Simulated errors are independent Gaussian, with standard deviations:* $\sigma = 0.02$ *in parts a and d,* $\sigma = 0.18$ *in parts b and e, and* $\sigma = 0.66$ *in parts c and f. The thick red curve is the Ruppert, Sheather, Wand Direct Plug In bandwidth, which is highlighted in the corresponding SiZer maps.*

For the very low noise case, $\sigma = 0.02$, the data contain a lot of information about the underlying regression curve, so the Ruppert, Sheather, Wand bandwidth (see Ruppert, Sheather and Wand (1995) for detailed description) and the undersmoothed members of the family are all essentially the same as the target curve. The SiZer map shows that all features of the target curve are significant, for a wide range of different resolutions (i.e. bandwidths). Even the "flat spot" near $x = 0.6$, which is not easy to find in the smooths, shows up as purple. When the noise level is increased substantially to $\sigma = 0.18$, the family plot shows that the estimation problem is now harder, and the SiZer map shows fewer significant features. However, the regions of increase are still significant, and two regions of decrease still appear, although at different levels of resolution. Increasing the noise still further, to $\sigma = 0.66$, results in a very challenging estimation problem, and now SiZer does not indicate any of the decreases, and only one of the two increases as being significant. The family plot shows this is reasonable, because the noise level is so high. The Ruppert, Sheather, Wand bandwidth suggests a decrease (although it completely misses the small valley at $x = 0.8$), but it is not clear with this noise level that it is significant, and SiZer shows it is not.

SiZer is also useful even in settings where the underlying target curve is not smooth, and in fact is quite useful at highlighting "jumps". This is shown in Figure 8, where Donoho and Johnstone's Blocks function (famous from many papers on wavelets) is used as a regression target, but the added Gaussian noise is larger than is typical in wavelet examples. Note that the locations of each jump is highlighted by colored streaks (blue for up and red for down) that reach all the way to the bottom of the SiZer map. The streaks are caused by the fact that even at very small bandwidths, the estimates are significantly changing at these points. This phenomenon appeared in a number of other examples we have studied where the target curve has jump discontinuities. These indicated jumps could be used to construct a step function estimator with much better properties than the usual wavelet estimators for this example.

[put Figure 8 about here]

FIGURE 8: *Family plot and SiZer map, for local linear regression, based on $n = 1024$ simulated data, shown as green dots, from an equally spaced design, and the Donoho-Johnstone Blocks regression curve shown as the yellow thick curve in Figure 8a. Simulated errors are independent Gaussian, with standard deviations: $\sigma = 0.05$. The thick red curve in Figure 8a is the Ruppert, Sheather, Wand Direct Plug In bandwidth, which is highlighted in the SiZer map in Figure 8b.*

5. Future research directions

In this section we discuss a number of future research directions, that are motivated by SiZer.

5.1. Local likelihood

Geroge Terrell has pointed out that for density estimation, and for special types of regression such as logistic regression, symmetric confidence intervals such as those proposed here can be improved upon, using context specific information. We suggest a local likelihood approach to this. Local likelhood is a smoothing method which is more efficient than simple kernel methods in some cases, for example discrete response variables. See Tibshirani and Hastie (1987), Staniswallis (1989), Chaudhuri and Dewanji (1995) and Fan, Heckman and Wand (1995) for detailed discussion and more references. We anticipate that SiZer may be extended in a fairly straightforward way to this important smoothing context.

5.2. Handling dependency

In nonparametric regression, our current SiZer development assumes independent errors, which is not always realistic, for example in time series contexts. But SiZer has the potential to become an important tool in such contexts where "significance of trends" is often an important issue. We believe that such applications will require appropriate modeling of the error structure, e.g. by some ARMA of even long range dependent models, before useful inference can be done.

5.3. Testing other types of hypotheses

SiZer focuses on regions where the derivatives are significantly increasing and decreasing, but for some situations other aspects of the underlying curve, such as the second derivative, of even the curve itself could be more appropriate to study in this way. Variations of SiZer could also be used to address other problems, such as whether or not two curves are significantly different.

5.4. Other estimation settings

Smoothing is useful in other settings besides just density and regression estimation. For example, SiZer can be extended to estimation of the hazard function, and other functions appearing in survival analysis. Another interesting extension would be to various censored data contexts.

5.5. Local bandwidth selection

A separate potential application of SiZer is to the old field of location varying bandwidth selection. The need for this is demonstrated in Figure 9, where the family approach shows that one would prefer a smaller bandwidth on the right, where the underlying density has finer features, and a larger bandwidth on the left, where the density has less curvature. The Sheather Jones Plug In bandwidth does a reasonable job with the fatter peaks, but could be much improved on the smaller peaks. In particular, SiZer shows that the smaller peaks really are significant, but only at a finer level of resolution (smaller bandwidth). However, while the need for it has been clearly understood, data based local bandwidth selection has proven to be a very challenging problem. In particular the simulation study of Farmen (1996) and Farmen and Marron (1997) shows that most of the available methods do not fare much better overall than the simple global bandwidth chosen by the Sheather Jones Plug In method. A likely intuitive explanation for this is that local bandwidth selectors essentially require knowledge of the local curvature, which is very hard to estimate.

[put Figure 9 about here]

FIGURE 9: *Family plot and SiZer map for kernel density estimates, based on* $n = 1000$ *simulated data, from the Marron and Wand density #15, "Discrete Comb", shown as the thick yellow curve in the family plots. The thick red curve in the family plots is the Sheather Jones Plug In bandwidth, which is the highlighted horizontal bar in the SiZer maps.*

Note that the SiZer map gives some interesting visual cues as to how one might choose a local bandwidth function, which is described as a curve running across the map. For example, the Sheather Jones Plug In bandwidth could be used for $x \in (-3, 2)$, then the bandwidth curve could move down to around $\log_{10}(h) = -1.4$ for $x \in (2.3, 3)$. An interactive approach to local bandwidth selection could be based on tracing a "bandwidth curve" with a mouse on the SiZer map. The resulting local bandwidth smooth could be shown in another window. If the family has already been computed, computation of the local bandwidth smooth would be very fast, since it only needs interpolation among the family members. See Marron and Udina (1997) for a different approach to local bandwidth selection.

A natural question is: with SiZer, why do we need local bandwidth smoothing? The answer is that for presenting conclusions to non-experts (who are not interested in details behind the conclusions), a single location varying smooth will be very simple and attractive.

5.6. Higher dimensions

The problem of "which features are really present?" is also very important in smoothing settings of more than one dimension. In particular, the two dimensional case is "image analysis", which has a very large literature. An important problem with extending SiZer to higher dimensions is how to present the "map". The very simplest two dimensional version one might try is to study the magnitude of the gradient, and highlight scale space regions where this is significantly above zero. But now the "map" would be shaded regions in 3 dimensions, which is fairly challenging to visualize.

Other applications would likely result in the need to visualize even higher dimensional maps. For example, one could replace the magnitude of the gradient by directional derivatives. Another example is that in some cases it could be desirable to sue different bandwidths in different directions. Even with a two dimensional image, implementation of both ideas would result in a 6 or 7 dimensional map.

6. Other Approaches

6.1. Mode Testing

An older approach to analyzing which features in a smooth is mode testing. Here one formulates a null hypothesis of "few modes" (e.g. one), and then constructs a test which seeks strong evidence of the alternative of "more modes" (e.g. two). This approach goes back at least to Good and Gaskins (1980), and later references include Silverman (1981), Hartigan and Hartigan (1985), Donoho (1988), Müller and Sawitzki (1991), Hartigan and Mohanty (1992), Mammen, Marron and Fisher (1992), Minnotte and Scott (1993), Fisher, Mammen and Marron (1994), Cheng and Hall (1997) and Fisher and Marron (1997).

Such tests have an important place, even now that SiZer has been developed, because they are likely to have greater power than the inferences available from SiZer. This is because they focus directly on the question of modality, and also they are not hampered by trying to be simultaneous over all of scale space. However most available mode tests have the weakness that they only determine how many modes are present, and don't say where the modes are, or even which features in the smooth are "the modes". See Minnotte (1997) and Mammen, Marron and Udina (1997) for some interesting exceptions to this. The strength of SiZer is that it gives a much faster way of addressing the question of which modes are significant, and which are not. We believe that SiZer should be used mostly as an exploratory tool, with follow up analysis by explicit mode tests recommended in border line cases.

6.2. Why not conventional Confidence Bands?

In classical parametric statistics, a time honored approach to displaying variability is the confidence interval. Many attempts have been made to extend this idea to nonparametric curve estimation. There are two major hurdles to the effective use of this technique:

- Instead of a single real valued parameter, the quantity being estimated is now an entire curve. Furthermore inference about features will involves aspects of simultaneous inference.
- Unlike conventional parameter estimation, curve estimation necessarily involves an important bias component.

There is a large literature on attempts to address these problems in the context of smoothing. Good access is provided through the monograph Hall (1992).

A quick and simple approach has been suggested e.g. by Hastie and Tibshirani (1990), where one ignores both of the issues 1 and 2, and simply writes down standard confidence intervals which captures only the variability part of the error. This can only give a rather crude indications of which features are really present, because the intervals are both too short for valid inference, and also off center, because the bias is ignored.

A time honored approach to handling bias is to make it negligible by "undersmoothing", i.e. using a very small bandwidth. There are many papers that do this simply by assuming that asymptotically as the sample size grows, the bandwidth tends to zero faster than the optimal, which causes the bias to tend to zero at a faster rate. This still leaves open the problem of how the bandwidth should be chosen, and the fact that for any fixed set of data, any bandwidth is going to have at least a little bias. But even ignoring these problems, confidence intervals based on such bandwidths are not intuitively appealing, since they may be expected to be unnecessarily long, i.e. significant features can be missed.

Another approach is to try to estimate the bias, and adjust accordingly. An attempt at this presented in Härdle and Marron (1991) was "asymptotically successful", but gave incorrect coverage in simulations, as discussed in their Section 3. They also showed that the reason for the error was because the bias estimate was inefficient. Nychka provided an intuitive explanation of this with the statement "if you could estimate bias effectively, then you could get an improved estimate".

Nice insight into the failure of bias correction methods was developed in several papers by Hall, that is well summarized in Section 4.4 of Hall (1991). The approach taken there is to choose the bandwidth to make coverage probabilities as close as possible to the desired values. Asymptotic theory is developed for optimal bandwidths according to this criterion, and it is shown that when optimal bandwidths are used, simple undersmoothed bandwidths give shorter confidence intervals than if one attempts any type of bias correction.

This motivates a more careful look at undersmoothed bandwidths, and a natural question is: how long are the coverage optimal confidence intervals? Figure 10 shows an example addressing this point, using the explicit representation given just after (3.5) in Hall (1991).

[put Figure 10 about here]

FIGURE 10: *Underlying normal mixture density f in Figure 10a. SiZer map for a simulated data set of size $n = 500$ in Figure 10c. 100 replicates of kernel density estimates, using the coverage optimal bandwidth in Figure 10b, and the bandwidth that is MSE optimal at $x = 0.8$ in Figure 10d. These two bandwidths are highlighted in the SiZer map as a solid line for the MSE optimal, and a dashed line for the coverage optimal.*

The true underlying density shown in Figure 10a is the Gaussian Mixture density

$$0.425 \cdot N(0.35, 0.0144) + 0.425 \cdot N(0.575, 0.0144) + 0.15 \cdot N(0.8, 0.0009).$$

Here we study its estimation when $n = 500$ data points are used, and focus on the thinner peak, i.e. on estimation at $x = 0.8$. The practical effect of the coverage optimal bandwidth is shown in Figure 10b, where overlays of kernel density estimates for 100 independent replicates (i.e. regeneration of the $n = 500$ pseudo data points) are shown. The envelope of curves suggests that at this level of smoothing, there is not enough information in the data to establish the statistical significance of the thinner peak, since the top of the envelope near the valley point $x = 0.72$ is well above the bottom of the envelope at the peak, $x = 0.8$. Figure 10d investigates whether or not there is enough information in the data to resolve the second peak, by again overlaying 100 realizations of the density estimate, but this time with the bandwidth chosen to minimize the $MSE = E\left[\widehat{f}_h(x) - f(x)\right]^2$ (approximated by simulation) at the peak $x = 0.8$. This envelope of curves shows that at this level of resolution, their seems to be plenty of information in the data, and the second mode should be a significant feature. The SiZer map in Figure 10c finds both of the modes, and thus is using the information available in the data more effectively than confidence intervals with the coverage optimal bandwidth can do.

Note that even if it were possible to get effective classical confidence bands (seems doubtful in view of the above discussion), SiZer would still be a more powerful data analytic tool. This is because confidence bands need to focus on a single bandwidth, which (even when it can be well chosen from the data) can still miss features that appear at other levels of resolution.

References

[1] Bralower, T. J., Fullagar, P. D., Paull, C. K., Dwyer, G. S., and Leckie, R. M., (1997), Mid-Cretaceous Strontium-Isotope Stratigraphy of Deep-Sea Sections, *Geological Society of America Bulletin*, in press.

[2] Chaudhuri, P. and Dewanji, A. (1995) On a likelihood based approach in non-parametric smoothing and cross-validation, *Statistics and Probability Letters*, 22, 7-15.

[3] Chaudhuri, P. and Marron, J. S. (1997) Scale space view of curve estimation, unpublished manuscript.

[4] Cheng M. Y. and Hall, P. (1997) Calibration methods for excess mass and DIP tests of homogeneity, unpublished manuscript.

[5] Efron, B. and Tibshirani, R. J. (1993) *An Introduction to the Bootstrap,* Chapman and Hall, New York.

[6] Eubank, R. L. (1988) *Spline smoothing and nonparametric regression*, Dekker, New York.

[7] Fan, J. and Gijbels, I. (1996) *Local polynomial modeling and its applications*, Chapman and Hall, London.

[8] Fan, J., Heckman, N. E, and Wand, M. P. (1995) Local polynomial kernel regression for generalized linear models and quasi-likelihood functions, *Journal of the American Statistical Association*,90, 141-150.

[9] Fan, J. and Marron, J. S. (1994) Fast implementations of nonparametric curve estimators, *Journal of Computational and Graphical Statistics*, 3, 35-56.

[10] Farmen, M. (1996) The smoothed bootstrap for variable bandwidth selection and some results in nonparametric logistic regression, PhD Dissertation, North Carolina Institute of Statistics Tech. Report Series #2342.

[11] Farmen, M. and Marron, J. S. (1997) An assessment of finite sample performance of adaptive methods in density estimation, unpublished manuscript.

[12] Fisher, N. I., Mammen, E. and Marron, J. S. (1994) Testing for multimodality, *Computational Statistics and Data Analysis*, 18, 499-512.

[13] Fisher, N. I. and Marron J. S. (1997) Mode Testing Via the Excess Mass Estimate, unpublished manuscript.

[14] Good, I. J. and Gaskins, R. A. (1980) Density estimation and bump-hunting by the penalized maximum likelihood method exemplified by scattering and meteorite data (with discussion), *Journal of the American Statistical Association* ,75, 42-73.

[15] Green and Silverman (1994) *Nonparametric Regression and Generalized Linear Models*, Chapman and Hall, London.

[16] Härdle, W. (1990) *Applied Nonparametric Regression*, Cambridge University Press, Boston.

[17] Härdle, W. and Marron, J. S. (1991) Bootstrap simultaneous error bars for nonparametric regression, *Annals of Statistics*, 19, 778-796.

[18] Hall, P. (1991) Edgeworth expansions for nonparametric density estimators, *Statistics*, 2, 215-232.

[19] Hall, P. (1992) *The bootstrap and edgeworth expansion.* Springer: New York.

[20] Hall, P. and Marron, J. S. (1997) On the role of the ridge parameter in local linear smoothing, *Probability Theory and Related Fields*, 108, 495-516.

[21] Hartigan, J. A. and Hartigan, P. M. (1985) The DIP test of multimodality, *Annals of Statistics*, 13, 70-84.

[22] Hartigan, J. A. and Mohanty, S. (1992) The RUNT test from multimodality, *J. Classification*, 9, 63-70.

[23] Hastie, T. J. and Tibshirani, R. J. (1990) *Generalized Additive Models*, Chapman and Hall, London.

[24] Izenman, A. J. and Sommer, C. (1988) Philatelic mixtures and multimodal densities, *Journal of the American Statistical Association*, 83, 941-953.

[25] Jones, M. C., Marron, J. S. and Sheather, S. J. (1996a) A brief survey of bandwidth selection for density estimation, *Journal of the American Statistical Association*, 91, 401-407.

[26] Jones, M. C., Marron, J. S. and Sheather, S. J. (1996b) Progress in data-based bandwidth selection for kernel density estimation, *Computational Statistics*, 11, 337-381.

[27] Lindeberg, T. (1994) *Scale space theory in computer vision.* Kluwer: Boston.

[28] Mammen, E., Marron J. S. and Fisher, N. I. (1992) Some asymptotics for multimodality tests based on kernel density estimates, *Probability Theory and Related Fields*, 91, 115-132.

[29] Mammen, E., Marron, J. S. and Udina, F. (1997) Interactive mode testing, unpublished manuscript.

[30] Marron, J. S. and Chung, S. S. (1997) Presentation of smoothers: the family approach, unpublished manuscript.

[31] Marron, J. S. and Udina (1997) Interactive local bandwidth choice, unpublished manuscript.

[32] Marron, J. S. and Wand, M. P. (1992) Exact mean integrated squared error, *Annals of Statistics*, 20, 712-736.

[33] Minnotte, M. C. (1997) Nonparametric testing of the existence of modes, *Annals of Statistics*, 25, 1646-1660.

[34] Minnotte, M. C. and Scott, D. W. (1993) The mode tree: a tool for visualization of nonparametric density features, *Journal of Computational and Graphical Statistics*, 2, 51-68.

[35] Müller, D. W. and Sawitzki, G. (1991) Excess mass estimates and tests for multimodality, *Journal of the American Statistical Association* , 86, 738-746.

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

[37] Schmitz, H. P. and Marron, J. S. (1992) Simultaneous estimation of several size distributions of income, *Econometric Theory*, 8, 476-488.

[38] Scott, D. W. (1992) *Multivariate density estimation, theory, practice and visualization*, John Wiley: New York.

[39] Silverman, B. W. (1981) Using kernel density estimates to investigate multimodality, *Journal of the Royal Statistical Society, Series B*, 43, 97-99.

[40] Silverman, B. W. (1986) *Density estimation for statistics and data analysis*, Chapman and Hall: London.

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

[42] Tibshirani, R. J. and Hastie, T. J. (1987) Local likelihood estimation, *Journal of the American Statistical Association* ,82, 559-568.

[43] Wahba, G. (1991) *Spline Models for Observational Data*, SIAM, Philadelphia.

Wand, M. P. and Jones, M. C. (1995) *Kernel Smoothing*, Chapman and Hall: London.

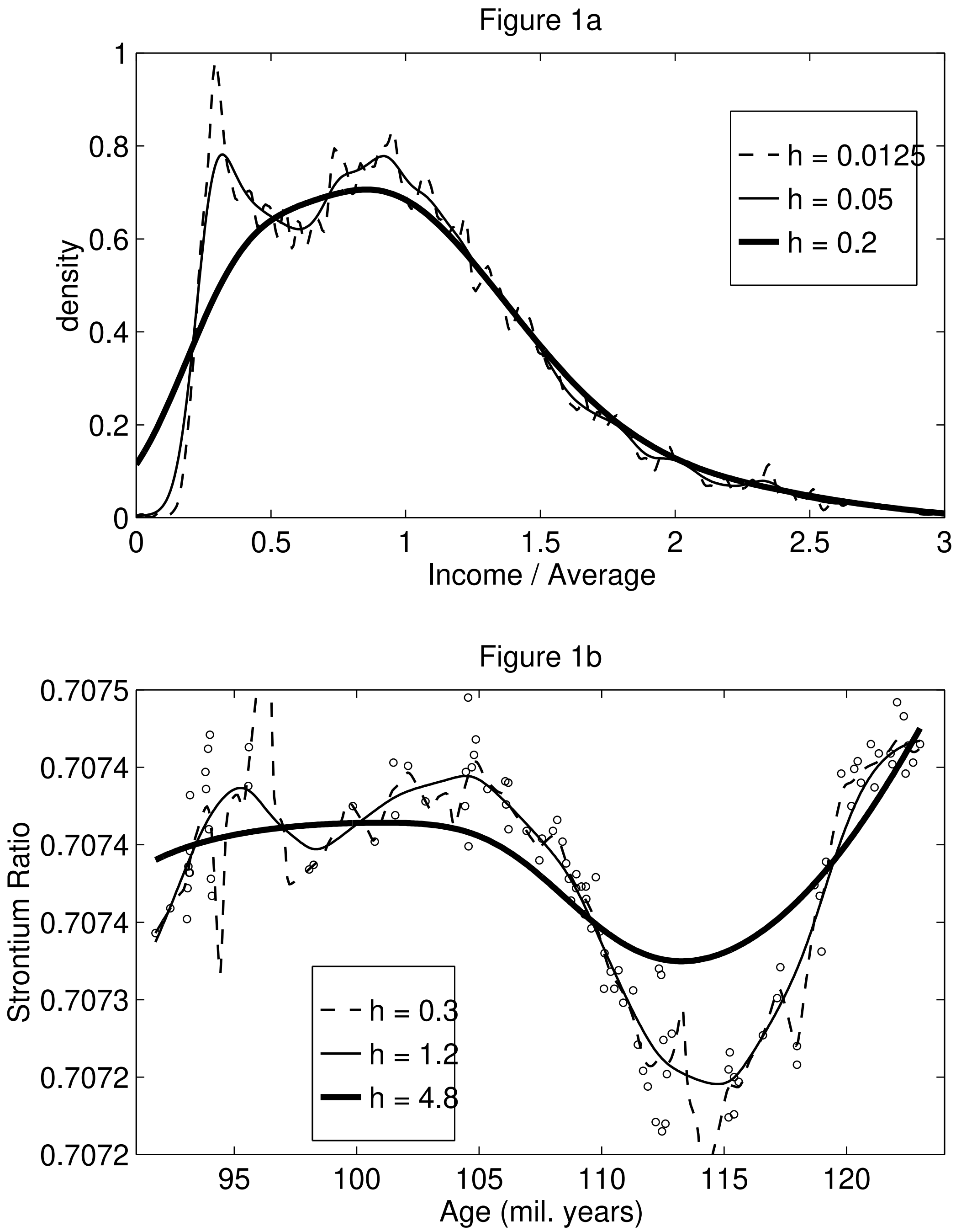
Figure 1a
density
Income / Average
- - h = 0.0125
h = 0.05
h = 0.2
0
0.2
0.4
0.6
0.8
1
0.5
1.5
2
2.5
3
Figure 1b
Strontium Ratio
Age (mil. years)
- - h = 0.3
h = 1.2
h = 4.8
0.7075
0.7074
0.7073
0.7072
95
100
105
110
115
120

Figure 6.1:

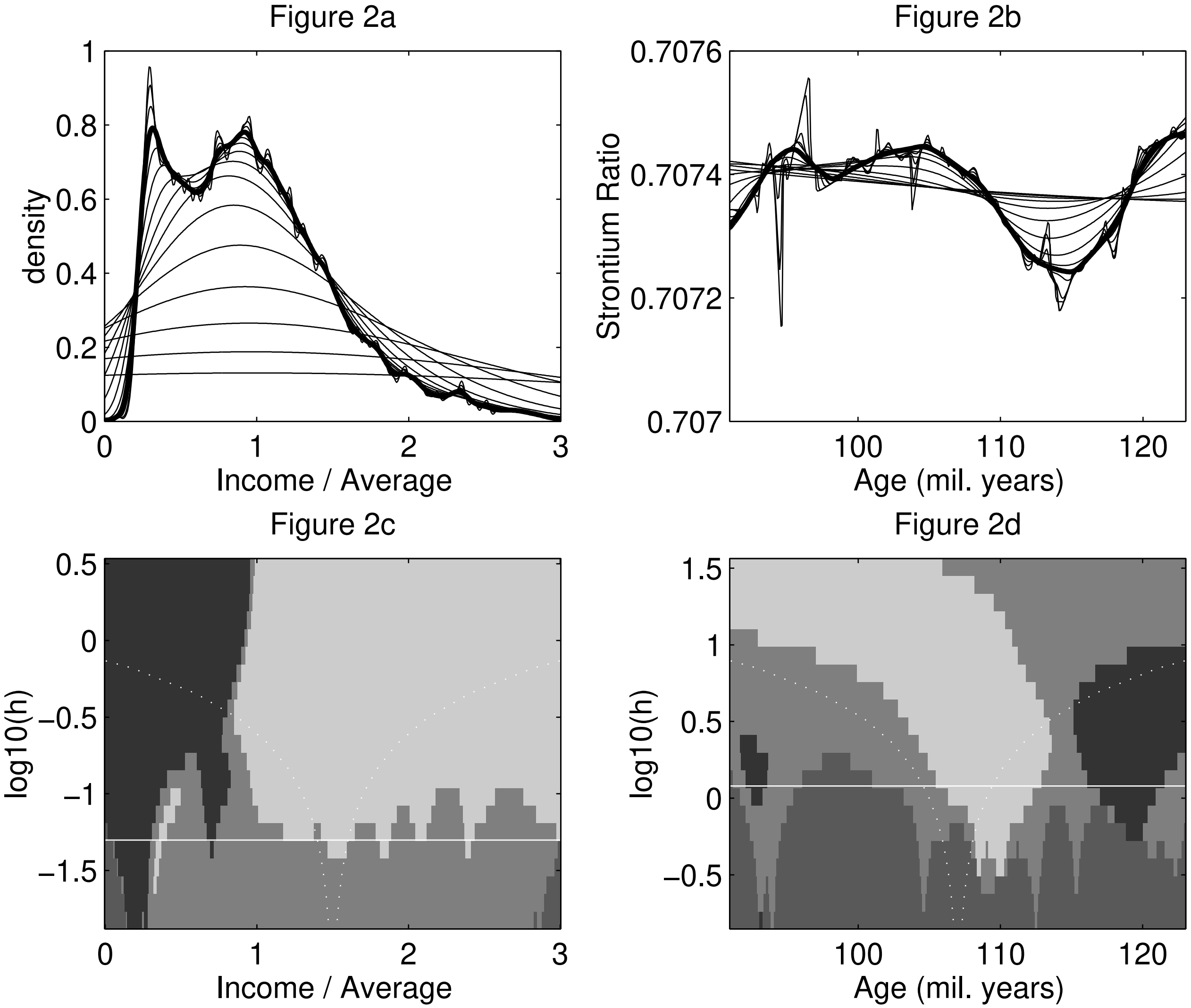

Figure 2a
density
Income / Average
Figure 2b
Strontium Ratio
Age (mil. years)
Figure 2c
log10(h)
Income / Average
Figure 2d
log10(h)
Age (mil. years)

Figure 6.2:

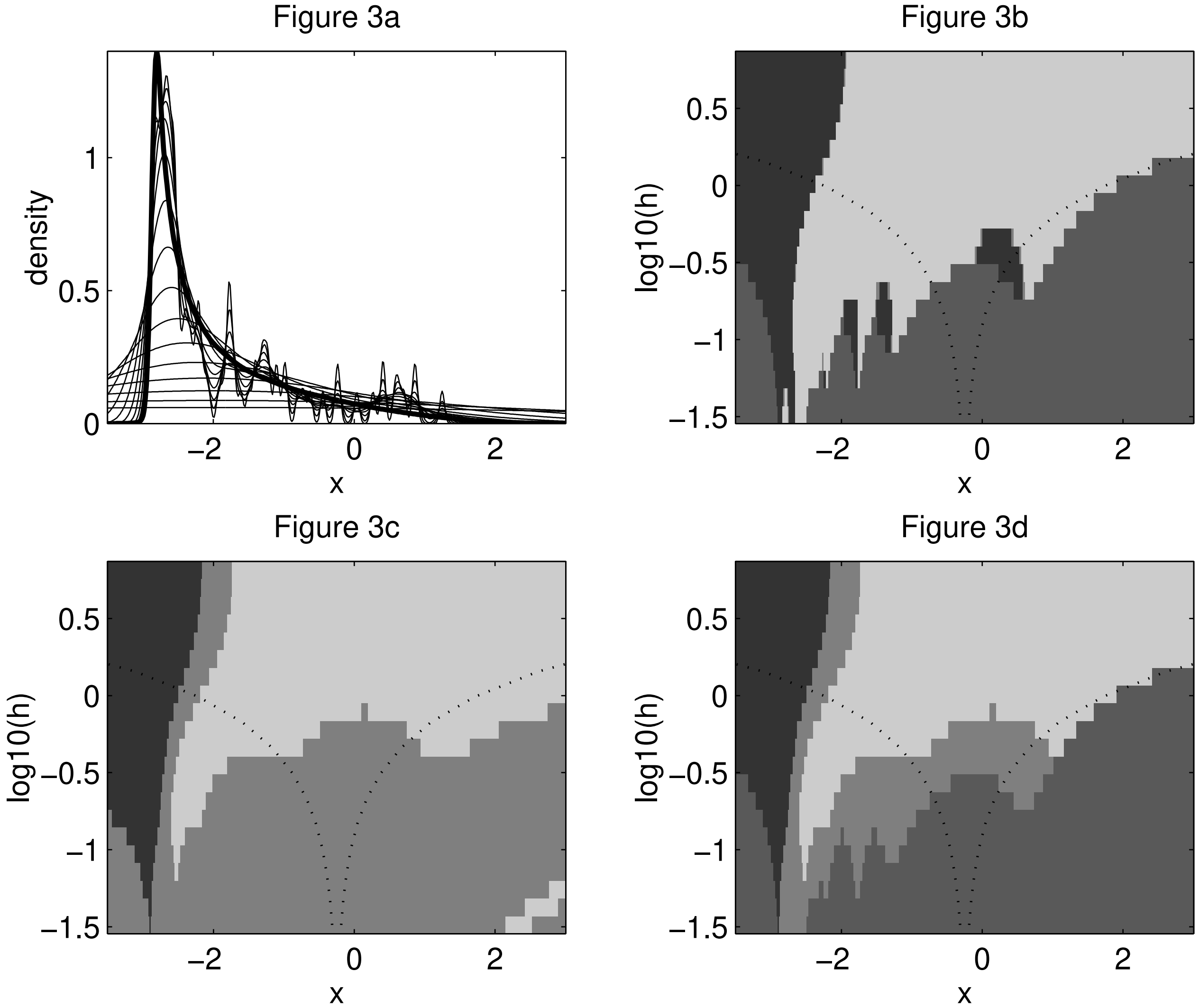

Figure 3a
density
x
Figure 3b
log10(h)
x
Figure 3c
log10(h)
x
Figure 3d
log10(h)
x

Figure 6.3:

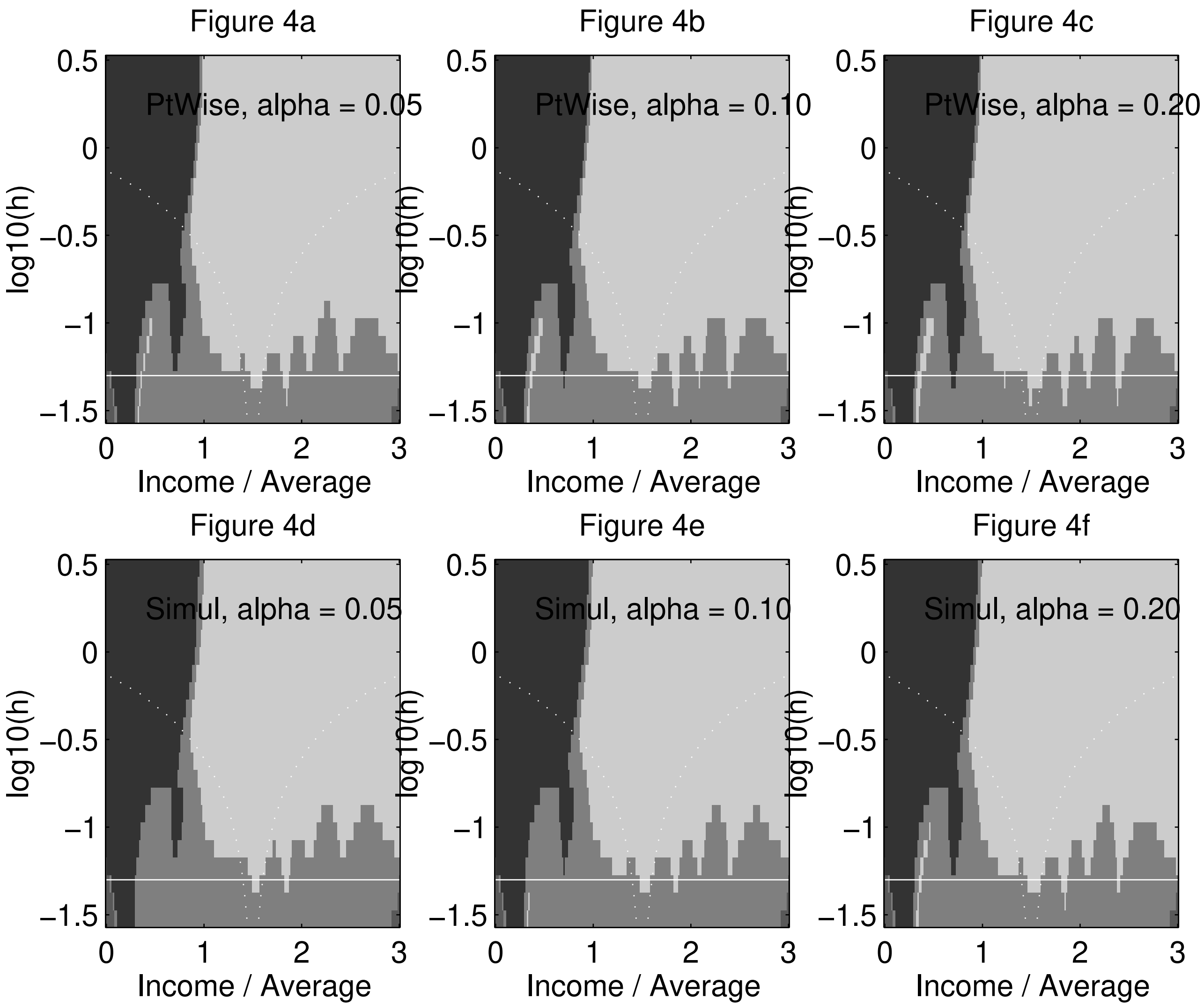
Figure 4a
PtWise, alpha = 0.05
Figure 4b
PtWise, alpha = 0.10
Figure 4c
PtWise, alpha = 0.20
Figure 4d
Simul, alpha = 0.05
Figure 4e
Simul, alpha = 0.10
Figure 4f
Simul, alpha = 0.20
log10(h)
Income / Average
0.5
0
−0.5
−1
−1.5
0
1
2
3

Figure 6.4:

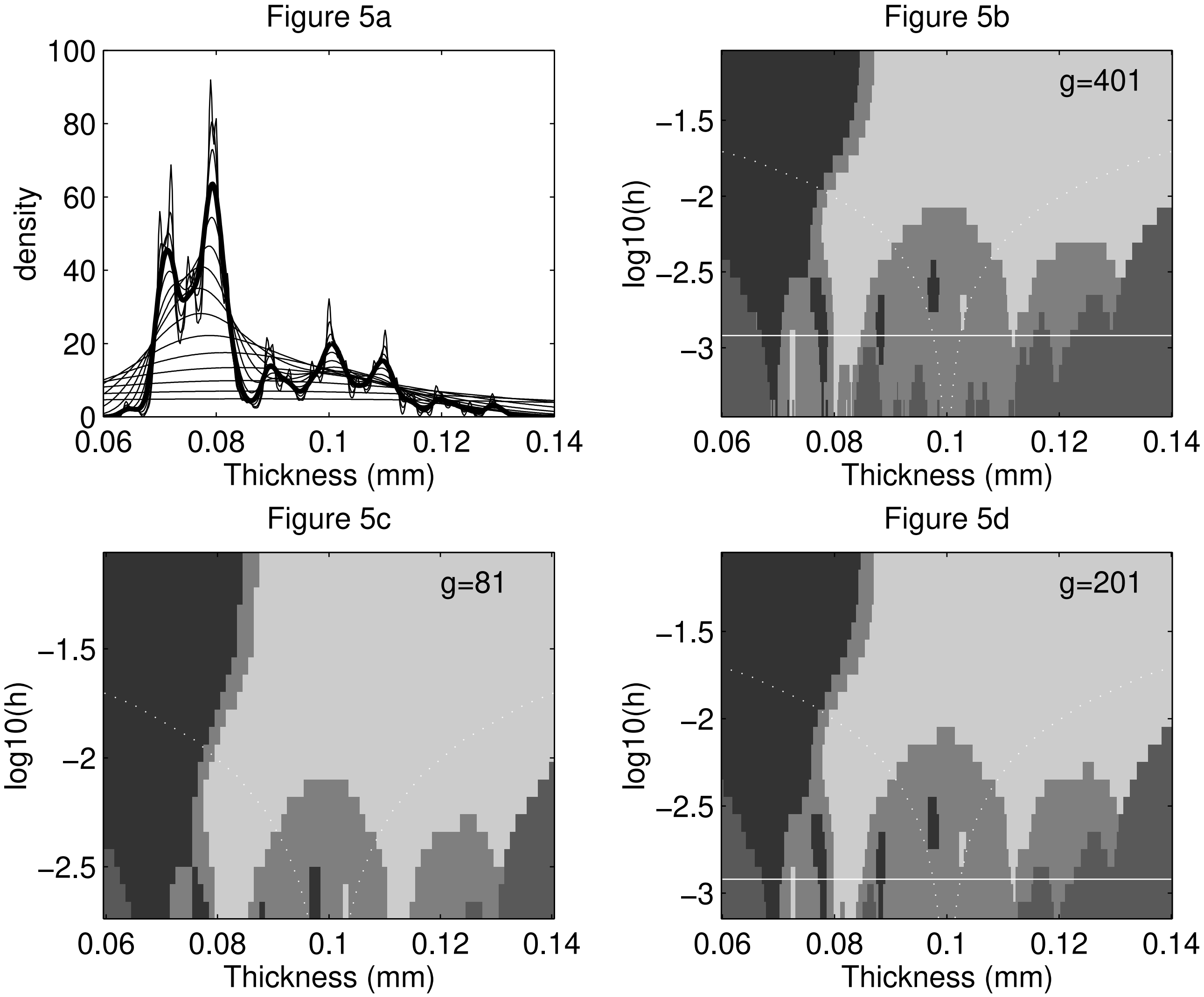
Figure 5a
density
Thickness (mm)
Figure 5b
g=401
log10(h)
Thickness (mm)
Figure 5c
g=81
log10(h)
Thickness (mm)
Figure 5d
g=201
log10(h)
Thickness (mm)

Figure 6.5:

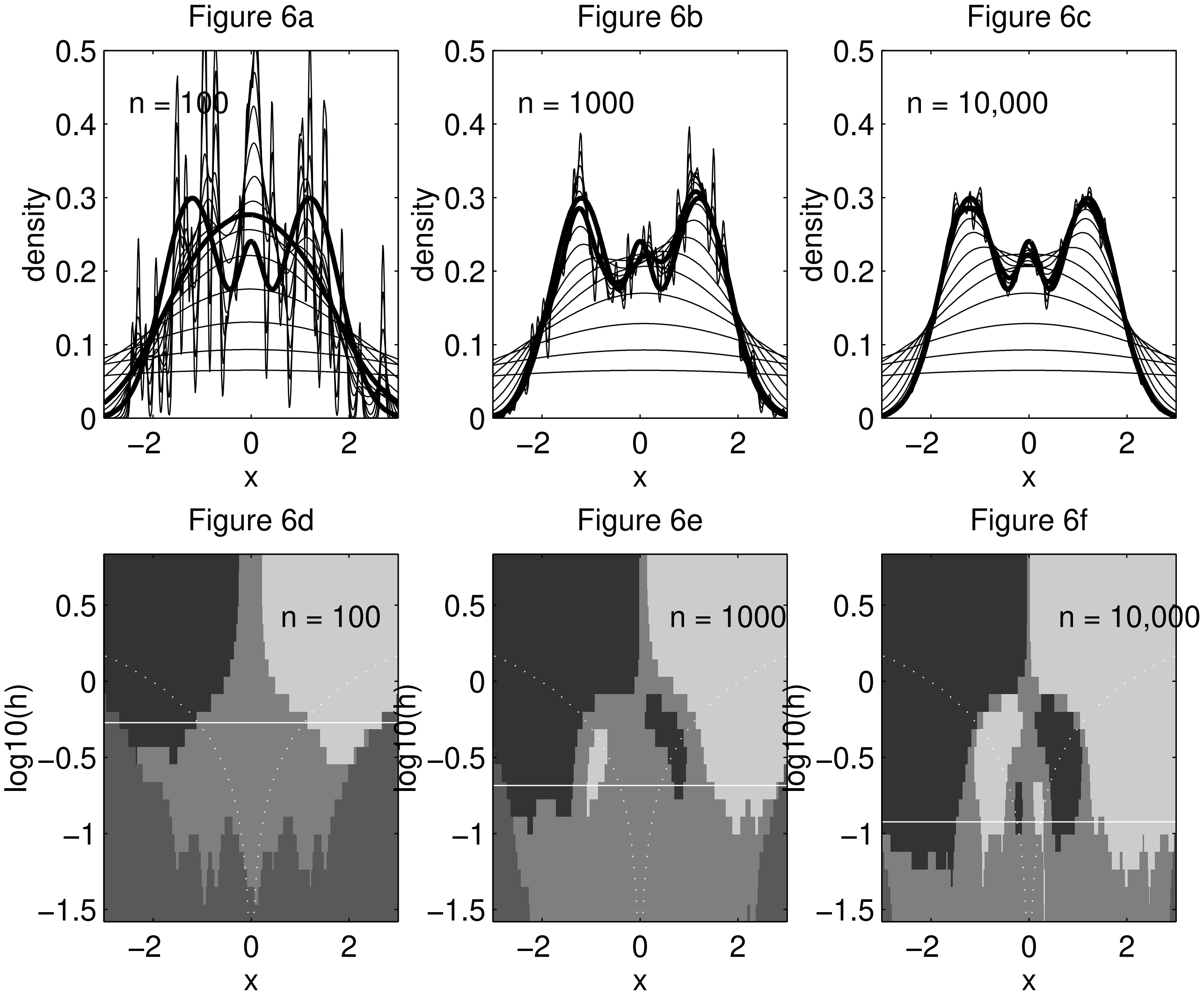

Figure 6a
Figure 6b
Figure 6c
n = 100
n = 1000
n = 10,000
density
x
Figure 6d
Figure 6e
Figure 6f
n = 100
n = 1000
n = 10,000
log10(h)
x

Figure 6.6:

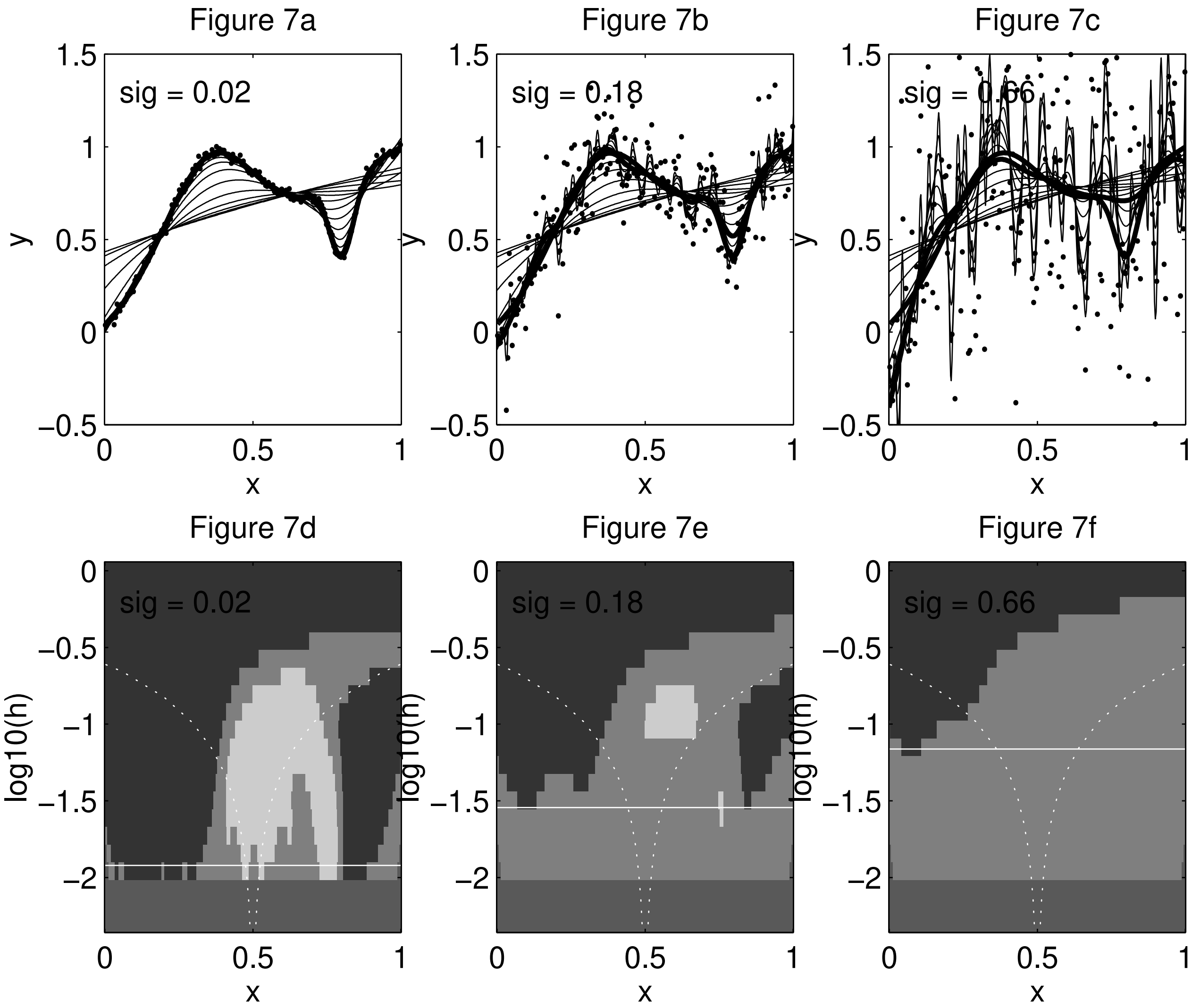
Figure 7a
sig = 0.02
Figure 7b
sig = 0.18
Figure 7c
sig = 0.66
Figure 7d
sig = 0.02
Figure 7e
sig = 0.18
Figure 7f
sig = 0.66
y
x
log10(h)

Figure 6.7:

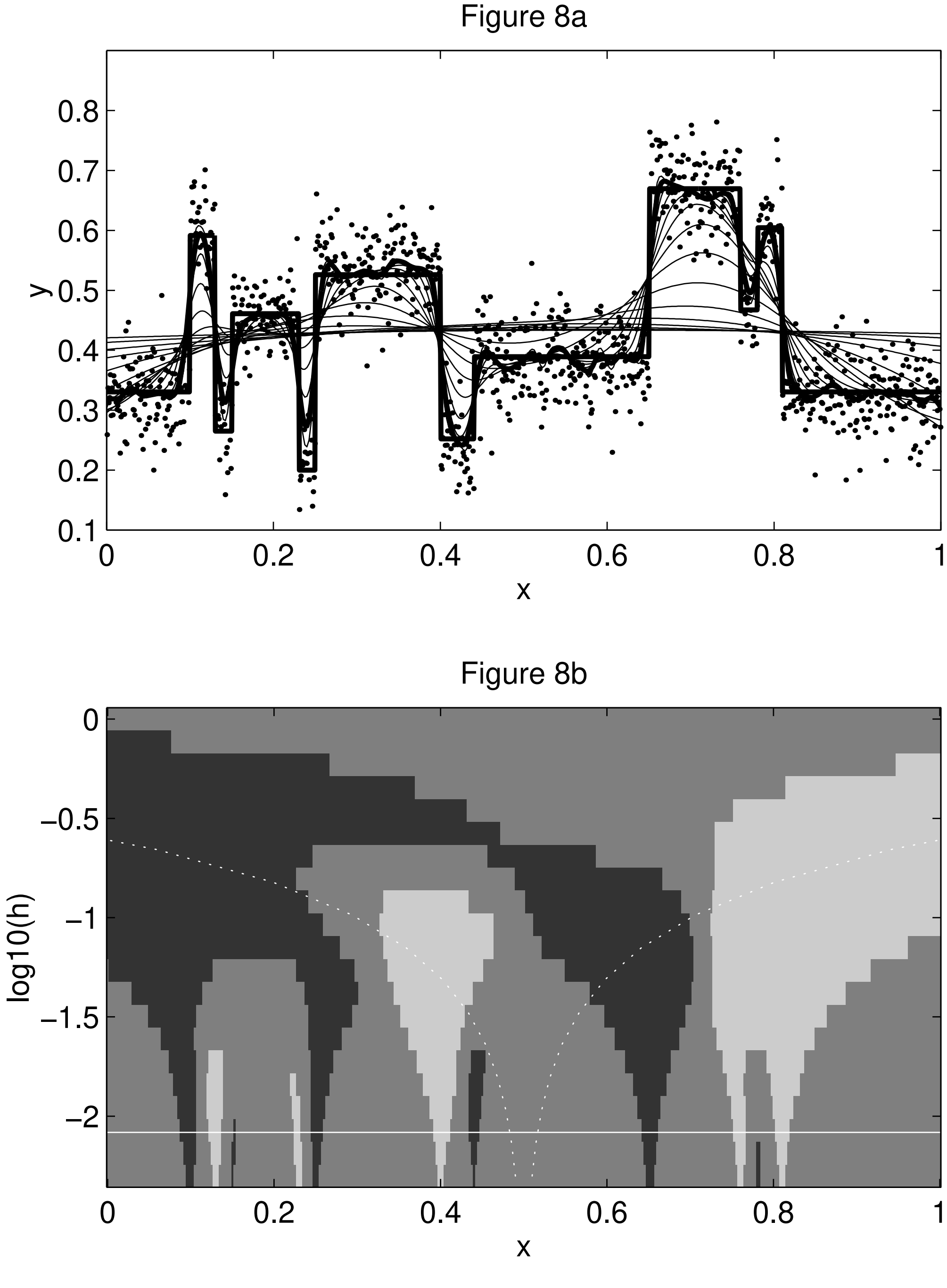
Figure 8a
0.8
0.7
0.6
y
0.5
0.4
0.3
0.2
0.1
0
0.2
0.4
0.6
0.8
1
x
Figure 8b
0
−0.5
log10(h)
−1
−1.5
−2
0
0.2
0.4
0.6
0.8
1
x

Figure 6.8:

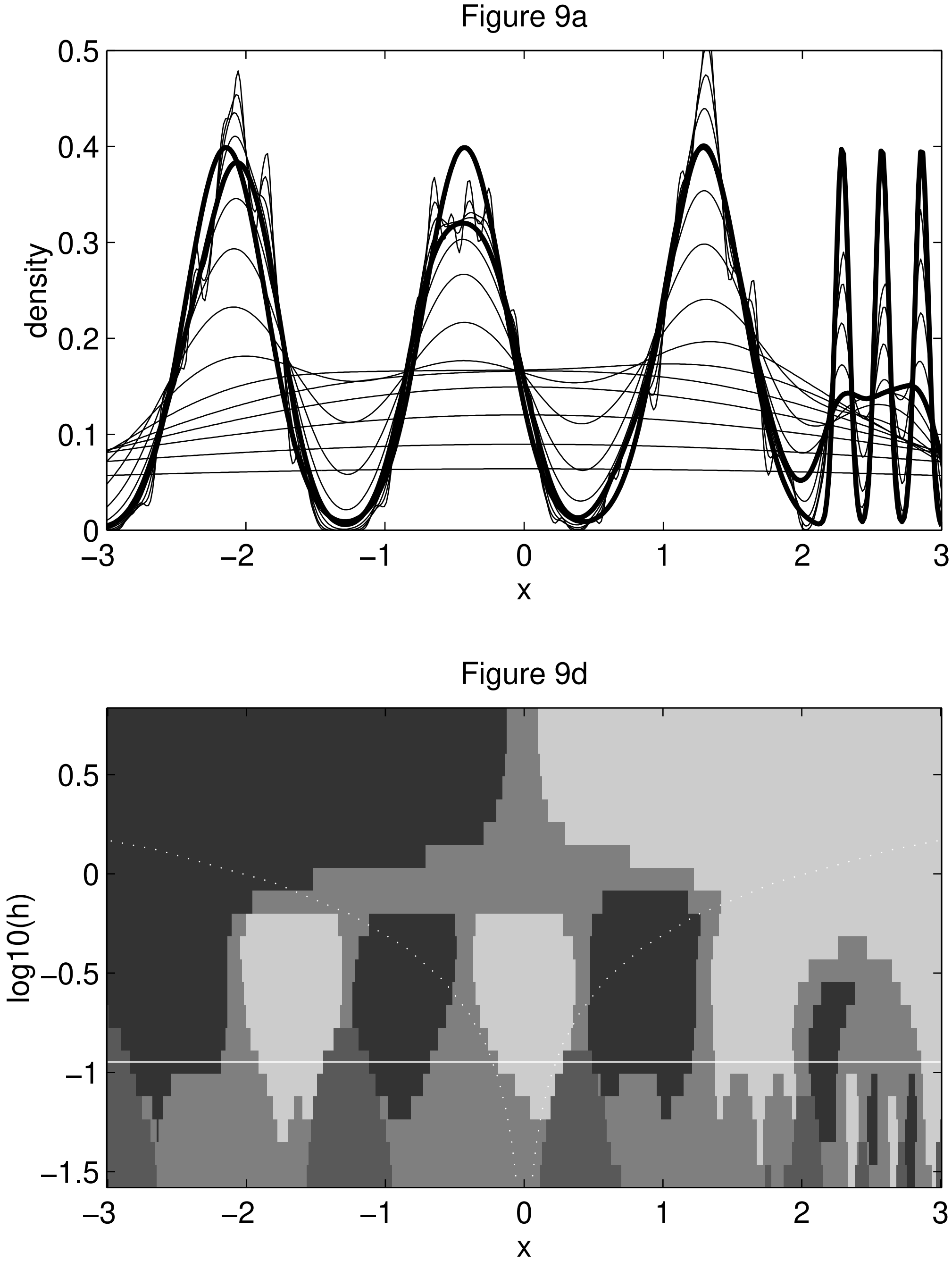
Figure 9a
0.5
0.4
0.3
0.2
0.1
0
density
−3
−2
−1
0
1
2
3
x
Figure 9d
0.5
0
−0.5
−1
−1.5
log10(h)
−3
−2
−1
0
1
2
3
x

Figure 6.9:

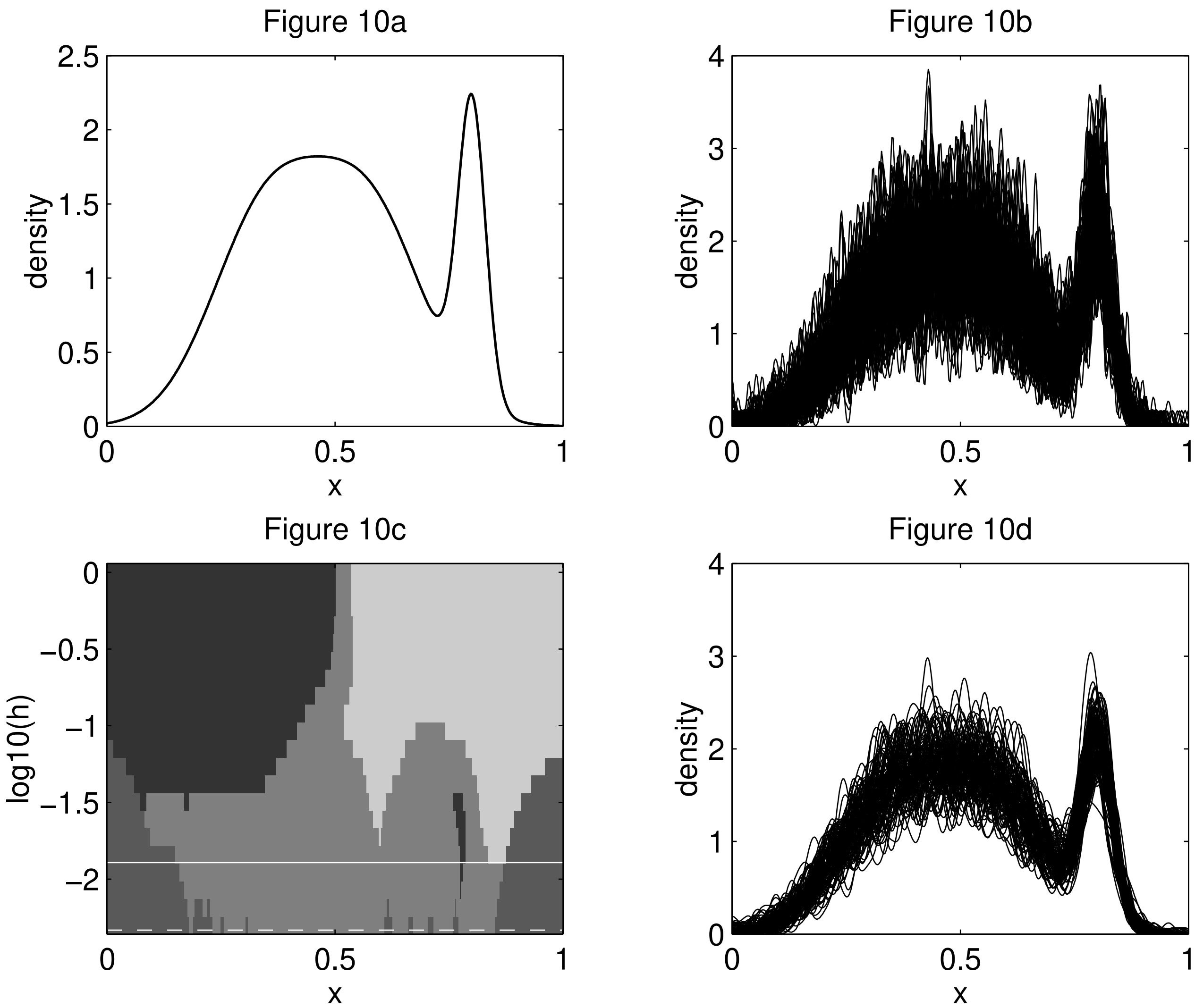
Figure 10a
density
x
Figure 10b
density
x
Figure 10c
log10(h)
x
Figure 10d
density
x

Figure 6.10: