SCALE SPACE VIEW OF CURVE ESTIMATION

P. Chaudhuri
Indian Statistical Institute, Calcutta
and
J. S. Marron
University of North Carolina, Chapel Hill

ABSTRACT

Scale space theory from computer vision leads to an interesting and novel approach to nonparametric curve estimation. The family of smooth curve estimates indexed by the smoothing parameter can be represented as a surface called *the scale space surface*. The smoothing parameter here plays the same role as that played by the scale of resolution in a visual system. In this paper, we study in detail various features of that surface from a statistical viewpoint. Weak convergence of the empirical scale space surface to its theoretical counterpart and some related asymptotic results have been established under appropriate regularity conditions. Our theoretical analysis provides new insights into nonparametric smoothing procedures and yields useful techniques for statistical exploration of features in the data. In particular, we have used the scale space approach for the development of an effective exploratory data analytic tool called *SiZer*. SiZer is a graphical device for evaluating statistical significance of features (e.g. peaks and valleys) visible in a curve estimate by assessing the significance of zero crossings of the derivatives of that curve estimate at different levels of smoothing.

1 Introduction : Curve Estimation and Scale Space Theory

Curve estimation using nonparametric smoothing techniques is an effective tool for unmasking important structures from noisy data. Over the last couple of decades, nonparametric curve estimates have emerged as powerful exploratory and inferential tools for statistical data analysis [see e.g. Silverman (1986), Eubank (1988), Müller (1988), Härdle (1990), Rosenblatt (1991), Wahba (1991), Green and Silverman (1994), Wand and Jones (1995), Fan and Gijbels (1996), Simonoff (1996)]. Many different methodologies such as kernel, nearest neighbor, local polynomial, splines and wavelets have been investigated in the literature for construction of the nonparametric estimate $\hat{f}_h(x)$ of an unknown curve $f(x)$. Here the subscript h denotes the smoothing parameter associated with the curve estimate the nature of which varies depending on the methodology used (e.g. in the case of kernel smoothing it is the bandwidth). In the usual approach taken in the statistics literature, one focuses on the "true underlying function" $f(x)$, and an extensive amount of work has been reported on the estimation of $f(x)$ and on optimal choice of the smoothing parameter from the data and inferences about $f(x)$ based on confidence bands. A fundamental problem in nonparametric function estimation is that $E\{\hat{f}_h(x)\}$ is not necessarily equal to $f(x)$, so there is an inherent bias which creates special challenges. The problem does not appear in classical parametric statistics, where one tacitly assumes a "correct" parametric model for $f(x)$ with parameters that can be unbiasedly estimated.

In this paper, we study nonparametric curve estimation from the viewpoint of "scale space theory" from the computer vision literature [see e.g. Lindeberg (1994)]. We will focus simultaneously on a wide range of values for the smoothing parameter ($h \in H$, say) instead of trying to estimate the optimum amount of smoothing from the data. From the point of view of data analysis this is an effective strategy since different levels of smoothing may reveal different useful information [see e.g. the "family approach" in Marron and Chung (1997)]. When H is a subinterval of $(0, \infty)$ (e.g. the range of possible bandwidths in a kernel smoother) and x varies in an interval I of the real line $(-\infty, \infty)$, the family of smooth curves $\{\hat{f}_h(x) \mid h \in H\ , x \in I\}$ can be represented by a surface, the "scale space surface"shown in Figure 1, which models different features of the data visible at different levels of

smoothing that are comparable with variations in the scales of resolution in a visual system. This unconventional way of handling the curve estimation problem leads to an interesting reorientation of the bias problem mentioned in the preceding paragraph. We shift our attention from the "true underlying curve" $f(x)$ to the "true curves viewed at different scales of resolution", which is $E\{\hat{f}_h(x)\}$ as h varies in H and x varies in I. $E\{\hat{f}_h(x)\}$ is a "smoothed version" of the function $f(x)$ and can be viewed as the *theoretical scale space surface* if we consider $\hat{f}_h(x)$ as the *empirical scale space surface*. The empirical version here is by definition unbiased for the theoretical version.

We make $E\{\hat{f}_h(x)\}$ our target and focus our inference on it with the idea that it will enable us to extract relevant information available in the noisy data at a given level of smoothing. A large value of the smoothing parameter models "macroscopic or distant vision", where one can hope to resolve only large scale features. Similarly a small value of the smoothing parameter will model "microscopic vision" that can resolve small scale features provided that we have a sufficient amount of informative data. A detailed discussion of scale space philosophy and many related interesting examples can be found in Lindeberg (1994).

Figure 1a shows a simulated regression example, based on a target curve $f(x)$ (dashed line type), and an equally spaced design, $x_i = \dfrac{i-1}{n-1}$ for sample size $n = 201$, and data $Y_i = f(x_i) + \varepsilon_i$ (dots), where the ε_i's are independent $N(0, \sigma^2)$, with $\sigma = 0.2$. A family of Gaussian kernel local linear smooths $\left\{\widehat{f}_h(x) : h \in H\right\}$, indexed by the bandwidth, is overlaid on Figure 1a, as thin solid lines. The Ruppert, Sheather and Wand (1995) data driven choice of bandwidth is indicated as the thick solid line. The family shows the very wide range of smoothing being considered, from nearly the raw data (very wiggly thin line), to nearly the simple least squares fit line (the limit as the window width goes to infinity). Figure 1b, shows this same family of smooths $\left\{\widehat{f}_h(x) : h \in H\right\}$, arranged one behind the other in bandwidth order, to give the empirical scale space surface. Figure 1c shows the corresponding theoretical scale space surface $\left\{E\widehat{f}_h(x) : h \in H\right\}$, which is constructed by applying the same smoothing operations to $\{f(x_i) : i = 1, ..., n\}$, instead of to $\{Y_i : i = 1, ..., n\}$. Figure 1d shows the difference surface $\left\{\widehat{f}_h(x) - E\widehat{f}_h(x) : h \in H\right\}$, which is showing how noise is attenuated in scale space since it is the corre-

sponding family of smooths of $\{\varepsilon_i : i = 1, ..., n\}$.

[put Figure 1 about here]

FIGURE 1: *Simulated regression example showing scale space ideas. Figure 1a shows the target curve as the dashed line, the data as small dots, and a family of local linear smooths as thin solid lines, with the Ruppert-Sheather-Wand bandwidth highlighted as the heavy solid line. Figure 1b is the empirical scale space surface. Figure 1c is the theoretical scale space surface, i.e. smooths of the target curve. Figure 1d is the "noise surface", i.e. the difference between the surfaces shown in Figures 1b and 1c.*

The target curve has been selected to highlight an important question that arises in data analysis by smoothing methods: which features visible in a smooth are "really there?". The broad peak around $x = 0.55$ and the deep valley around $x = 0.85$ seem to be clearly discernible from the data. It is likely that the peak at $x = 0.2$ and the valley at $x = 0.3$ can be shown to be "statistically significant" as well. But what about the thinner peak at $x = 0.65$? This is much more questionable, since the corresponding sizes of the smooths are roughly comparable to the size of the spurious peaks just to the left. Note that the much thinner peak at $x = 0.75$ clearly does not have enough mass to be distinguishable from the background noise (even though it is a feature of the target curve). In Section 4 we discuss SiZer, a visualization which gives a convenient solution to this problem of which features are "really there", i.e. are "statistically significant".

Figure 1b shows how the scale space view of smoothing is looking at the data at a number of different resolutions. Figure 1c shows the corresponding multi-resolution views of the underlying target curve. These surfaces have a number of interesting properties, some of which are discussed in Section 2. Convergence results, which give a way of making precise the apparent approximations of the surfaces in Figures 1b and 1d, are derived in section 3. Figure 1d shows the "noise surface" that is displaying the variance part of the smoothing problem.

2 The Scale Space Surface

One of the prime objectives of nonparametric curve estimation is exploration of structures such as peaks and valleys. An important requirement, which the scale space surface should satisfy, is that as one moves from lower to higher levels of smoothing, structures (e.g. peaks and valleys) should disappear *monotonically*. In other words, the smoothing method should not introduce artifacts by creating "spurious structures" as we go from a finer to a coarser scale. This idea has been formalized as "causality" in the scale space literature [see e.g. Lindeberg (1994)], which is a property of the scale space surface, and it implies that the number of local extrema in the curve $\hat{f}_h(x)$ or $E\{\hat{f}_h(x)\}$ for a given h will be a decreasing function of h. The term "causality" was introduced to convey the idea that there should be a cause for structures appearing at coarser scales, in terms of finer scale structures. Causality, i.e. non-creation of new features with more smoothing, is visually apparent in Figures 1b, c and d.

Again assume that x varies in a subinterval I of $(-\infty, \infty)$ and h varies in a subinterval H of $(0, \infty)$. The kernel density estimator based on data $X_1, X_2, \ldots, X_n$, is

$$\hat{f}_h(x) \;=\; (nh)^{-1} \sum_{i=1}^{n} K\{(x - X_i)/h\} \;,$$

where $K(x)$ is the kernel function, which is usually taken to be a smooth density symmetric around zero. The fact that the number of peaks in a kernel density estimate based on a Gaussian kernel $K(x) = (1/\sqrt{2\pi}) \exp(-x^2/2)$ decreases monotonically with the increase in the bandwidth was proved in the statistics literature by Silverman (1981). Let us now consider the regression problem based on the data $(Y_1, X_1), (Y_2, X_2), \ldots, (Y_n, X_n)$. In this case, we have the Priestley-Chao estimate

$$(A) \qquad \hat{f}_h = (nh)^{-1} \sum_{i=1}^{n} Y_i K\{(x - X_i)/h\}$$

or the Gasser-Müller estimate

$$(B) \qquad \hat{f}_h(x) = \sum_{i=1}^{n} Y_i \int_{t_{i-1}}^{t_i} (1/h) K\{(x - s)/h\} ds \;,$$

where $-\infty = t_0 < X_1 < t_1 < X_2 < t_2 < \ldots < t_{n-1} < X_n < t_n = \infty$.

Observe that local extrema like peaks and valleys of the curve $\hat{f}_h(x)$ for fixed h are determined by the zero crossings of the derivative $\dfrac{\partial \hat{f}_h(x)}{\partial x}$. Similarly, points of inflexion are determined by the zero crossings of the second derivative $\dfrac{\partial^2 \hat{f}_h(x)}{\partial x^2}$. In general, zero crossings of the m-th order derivative $\dfrac{\partial^m \hat{f}_h(x)}{\partial x^m}$ for $m \geq 1$ can be used to identify structures in a smooth curve. We now state a theorem which gives an analog of Silverman's (1981) result for nonparametric regression problems. The proof is in Section 5.

Theorem 2.1 : *Assume that the scale space surface $\hat{f}_h(x)$ arises as in (A) or (B) above, and $K(x) = (1/\sqrt{2\pi})\exp(-x^2/2)$. Then for each fixed $h \in H$ and $m = 0, 1, 2, \ldots$, the number of zero crossings of the derivative $\dfrac{\partial^m \hat{f}_h(x)}{\partial x^m}$ will be a decreasing and right continuous function of h for all possible choices of the data $(Y_1, X_1), (Y_2, X_2), \ldots, (Y_n, X_n)$. Further, the same result holds for the m-th order derivative $\dfrac{\partial^m E\{\hat{f}_h(x)\}}{\partial x^m}$ of the theoretical scale space surface when we assume that the Y_i's are conditionally independently distributed given the X_i's, and E denotes the conditional expectation given $X_1, X_2, \ldots, X_n$.*

It will be appropriate to note here that for other versions of kernel based regression smoothers such as the Nadarya-Watson estimate and kernel weighted local polynomial estimates [see Wand and Jones (1995), Fan and Gijbels (1996) and Cleveland and Loader (1996) for useful discussion and historical background], which arise in the forms of ratios of two weighted averages, the "causality" (monotonicity) property may fail to hold on their scale space surfaces for certain data sets even if the Gaussian kernel is used. While discussing Silverman's (1981) result on kernel density estimates, Minnotte and Scott (1993) constructed some counter-examples to show that this monotonicity may fail to hold for certain non-Gaussian kernels including some compactly supported ones. However, they did not resolve the case of Cauchy kernel in a definite way. The example in Figure 2 demonstrates that the Cauchy kernel may not produce a scale space surface with the "causality" property.

The noncausality of the Cauchy kernel proved to be rather elusive, with trial and error simulation experiments not turning up a counterexample [similar to the experience reported by Minnotte and Scott (1993)]. This suggested that modes that were created with increasing bandwidth were rare and/or very small. To improve the magnification of our search method, we studied very small sets of parametrically indexed examples, which gave simple analytic representations for the derivative of the Cauchy kernel smooths. In particular, for regression with three data points, the number of roots of the derivative is the same as the number of real roots in a degree eight polynomial. Figure 2 shows one example where lack of causality, i.e. creation of additional modes, appeared for the Cauchy kernel. Figure 2a shows the three data points as circles, together with sample smooths using 3 bandwidths. Figure 2b shows the number of roots as a function of the bandwidth with the vertical overlaid lines corresponding to the three bandwidths in Figure 2a. Note the increase around $\log_{10}(h) = 0.4$, which implies that the number of modes in the smooth increases with h at that point. Note also that it is not clear in Figure 2a that the dashed and dot-dashed curves have two modes while the solid curve has three modes. Figures 2c and 2d show that this is actually the case, by successive zooming. The trimodality of the solid curve only becomes clear using the large amount of magnification shown in Figure 2d.

[put Figure 2 about here]

FIGURE 2: *Counterexample showing that the Cauchy kernel does not satisfy the causality property. A three point regression data set is shown as circles in Figure 2a, together with 3 Cauchy kernel smooths. Figure 2b shows the numbers of real roots of the derivative, as a function of the bandwidth , with the 3 smooths in Figure 2a represented as vertical bars with the same line types. Figure 2c and 2d are successive enlargements of the regions shown as boxes in Figures 2a and 2c respectively.*

Other examples we found were of similar very small magnitude, so we believe this noncausality of the Cauchy kernel is always small scale. We also were unable to find an example where all of the Y-values were positive, as in density estimation. So we conjecture that the Cauchy kernel may be causal

for density estimation [recall that Minnotte and Scott (1993) reported not finding a counterexample in that case].

For scale space surfaces arising as smooth convolutions of the form

$$(C) \qquad S(x,h) = f(x) * (1/h)K(x/h) = \int f(t)(1/h)K\{(x-t)/h\}dt \ ,$$

where $f(x)$ is a smooth function, Lindeberg (1994) gives a detailed discussion of the causality property and several interesting related results following Schoenberg (1950), Hirschmann and Widder (1955), Karlin (1968), Witkin (1983) and Koenderink (1984). A very interesting justification for the "causality" in the scale space surface $S(x,h)$ generated by the Gaussian kernel can be found in the scale space literature. If we accept that diffusion (e.g. heat diffusion) is a physical process that "destroys structures" over time and does not "create structures", and view the smoothing parameter as the time parameter in the diffusion process, the "causality" of the scale space surface can be reformulated in terms of the classical heat diffusion equation

$$(D) \qquad \frac{\partial S(x,\sqrt{t})}{\partial t} \ = \ (1/2)\frac{\partial^2 S(x,\sqrt{t})}{\partial x^2} \ .$$

The Gaussian kernel emerges as the Green's function solving (D). Here $h = \sqrt{t}$, so that time in the heat diffusion goes like the square of the bandwidth, i.e. the variance of the kernel window. For more formal mathematical details on the derivation of the heat equation in this context and its solution, readers are referred to Koenderink (1984) and Section 2.5 in Lindeberg (1994). Figure 3 provides visual insight into how solutions to the heat equation correspond to families of smooths.

The physical model for Figure 3 is a thin wire, with hot and cold spots at the beginning, and the heat dissipating over time (represented here by bandwidth). The color map in Figure 3b shows how the heat diffuses. The surface in Figure 3a is the corresponding solution to the heat equation. The starting values used were the raw data shown in Figure 1a. Figures 3a and 3b are both approximations, based on Gaussian kernel Nadaraya-Watson smooths.

[put Figure 3 about here]

FIGURE 3: *Figure 3a shows a family of Nadaraya-Watson smooths, similar to Figure 1b, for the data of Figure 1a, but now panels are shaded using a "temperature" color scale. Projection of these colors into the plane is shown in Figure 3b, which shows how (one dimensional) heat diffuses in time.*

The use of the heat equation as a paradigm for smoothing is quite well developed in some parts of the literature [see Weickert (1997) for good access to this work]. But for statisticians this approach provides a host of new answers to some old problems, e.g. boundary adjustments and corrections. Another such problem is: how should continuous convolution be discretized, as for nonparametric regression? There has been substantial debate concerning the Nadaraya-Watson (evaluate the kernel) vs. the Gasser-Müller (integrate the kernel over small rectangles) approaches. Many statisticians now prefer the local linear, for reasons made clear by Fan (1992,1993), although see e.g. Stone (1977) and Cleveland (1979) for much earlier insights in this direction. However, the heat equation approach gives a quite different resolution of this controversy, using the solution of a discrete analog of the heat equation. See section 3.6.2 of Lindebergh (1994) for details.

Silverman (1981) introduced the notion of "critical bandwidths", which are used to test for multimodality of densities. If $N(h)$ denotes the number of modes in a density estimate based on the Gaussian kernel with bandwidth h, we have already noted that $N(h)$ is a monotonically decreasing and right continuous function of h, and "critical bandwidths" are precisely the points of jump discontinuities of $N(h)$. Minnotte and Scott (1993) [see also Marchette and Wegman (1997)] introduced the notion of a "mode tree", which is a graphical tool that presents the locations of modes of a kernel density estimate at different bandwidths. We will now discuss some important connections between these statistical concepts and the geometry of the scale space surface. Suppose that we have a smooth scale space surface $\{\hat{f}_h(x) \mid x \in I, h \in H\}$ arising from a density estimation or a regression problem, and assume that "causality" holds for this surface. Consider the trajectories of the critical points on this surface given as

$$\left\{ (x, h, \hat{f}_h(x)) \mid x \in I, h \in H, \frac{\partial \hat{f}_h(x)}{\partial x} = 0 \right\} .$$

Then these trajectories trace the "mode tree" as well as the "antimode tree"

on the scale space surface (antimode = valley). Critical points $(x, h, \hat{f}_h(x))$ where $\dfrac{\partial^2 \hat{f}_h(x)}{\partial x^2} = 0$, are called degenerate critical points. Critical points where $\dfrac{\partial^2 \hat{f}_h(x)}{\partial x^2} \neq 0$, are called non-degenerate. Degeneracy of a critical point is a form of singularity on the surface where bifurcation of the trajectory may occur. The following theorem describes some interesting features of critical points on a scale space surface.

Theorem 2.2 : *A critical point $(x, h, \hat{f}_h(x))$ corresponds to a "critical bandwidth" only if it is a degenerate critical point. With the increase in the value of scale, the x co-ordinate of a non-degenerate critical point moves with a finite velocity along its trajectory (however, this drift velocity at a degenerate critical point may be infinite).*

Figure 4 shows a discretized version of these trajectories, for the same example as used in Figures 1 and 3. Red highlights the mode tree (subset of scale space consisting of local maximizers in the x direction), and yellow highlights the antimode tree (subset of scale space consisting of local minimizers in the x direction). The critical bandwidths are at the branch locations.

[put Figure 4 about here]

FIGURE 4: *Figure 4a shows the same family of smooths as in Figure 1b, with modes highlighted in red, and antimodes highlighted in yellow. Figure 4b shows the projection of the red mode locations and the yellow antimode locations into the plane, yielding the mode and antimode trees.*

Variations on mode and anti-mode trees have been developed in parallel in other literatures, see Muzy, Bacry and Arneodo (1994) for a wavelet version, and Wong (1993) for a neural net version.

3 Weak Convergence of Empirical Scale Space Surface and Its Derivatives

Though we have stated Theorem 2.2 in the preceding Section only for the empirical scale space surface $\{\hat{f}_h(x) \mid x \in I, h \in H\}$, analogous results hold

for the theoretical scale space surface $\{E\{\hat{f}_h(x)\} \mid x \in I, h \in H\}$, and the proofs will be virtually identical. Note that the "critical bandwidths" as well as the "mode tree" have their empirical and theoretical (or population) versions, where the former can be viewed as an estimate of the latter. We will now focus attention on statistical convergence of the empirical scale space surface and its derivatives to their theoretical counterparts. Consider first the density estimation problem based on i.i.d observations $X_1, X_2, \ldots, X_n$. Assume that $\hat{f}_h(x)$ is the usual kernel density estimate $(nh)^{-1} \sum_{i=1}^{n} K\{(x - X_i)/h\}$ and $E\{\hat{f}_h(x)\} = E[h^{-1}K\{(x - X_i)/h\}]$.

Theorem 3.1 : *Suppose that $X_1, X_2, \ldots, X_n$ are i.i.d random variables with a common distribution F_n, where as $n \longrightarrow \infty$, F_n converges weakly to a distribution F, and assume that I and H are compact subintervals of $(-\infty, \infty)$ and $(0, \infty)$ respectively. Let the smooth kernel $K(x)$ be such that for integer $m \geq 0$, the derivatives $\dfrac{\partial^m h^{-1} K(x/h)}{\partial x^m}$ and $\dfrac{\partial^{m+2} h^{-1} K(x/h)}{\partial h \partial x^{m+1}}$ both remain uniformly bounded as h varies in H and x varies in $(-\infty, \infty)$. Then as $n \longrightarrow \infty$, the 2-parameter stochastic process*

$$n^{1/2} \left[\frac{\partial^m \hat{f}_h(x)}{\partial x^m} - \frac{\partial^m E\{\hat{f}_h(x)\}}{\partial x^m} \right]$$

with $(h, x) \in H \times I$ converges weakly to a Gaussian process on $H \times I$ with zero mean and covariance function

$$cov(h_1, x_1, h_2, x_2) = COV \left(\frac{\partial^m h_1^{-1} K\{(x_1 - X)/h_1\}}{\partial x_1^m}, \frac{\partial^m h_2^{-1} K\{(x_2 - X)/h_2\}}{\partial x_2^m} \right) ,$$

where X has distribution F.

Let us next consider the regression problem based on independent observations $(Y_1, X_1), (Y_2, X_2), \ldots, (Y_n, X_n)$, and in this case we will assume that $\hat{f}_h(x)$ has the form $n^{-1} \sum_{i=1}^{n} Y_i W_n(h, x, X_i)$, where W_n is a smooth weight function that arises from the kernel function in usual kernel regression or kernel weighted local polynomial regression with bandwidth h. We will also set $E\{\hat{f}_h(x)\} =_{def} n^{-1} \sum_{i=1}^{n} E(Y_i|X_i) W_n(h, x, X_i)$ as before.

Theorem 3.2 : *Suppose that $(Y_1, X_1), (Y_2, X_2), \ldots, (Y_n, X_n)$ are i.i.d observations with a common bivariate distribution G_n such that we have*

$\sup_{n\geq 1}\sup_{x\in I} E_{G_n}\left\{|Y-E(Y|X=x)|^{2+\rho} \mid X=x\right\} < \infty$ *for some* $\rho > 0$, *and as in Theorem 3.1,* I *and* H *are compact subintervals of* $(-\infty,\infty)$ *and* $(0,\infty)$ *respectively. For integer* $m \geq 0$, *assume that as* $n \longrightarrow \infty$,

$$n^{-1}\sum_{i=1}^{n} VAR_{G_n}(Y_i|X_i)\frac{\partial^m W_n(h_1,x_1,X_i)}{\partial x_1^m}\frac{\partial^m W_n(h_2,x_2,X_i)}{\partial x_2^m}$$

converges in probability to a covariance function $cov(h_1,x_1,h_2,x_2)$ *for all* (h_1,x_1) *and* $(h_2,x_2) \in H \times I$, *and*

$$n^{-(1+\rho/2)}\left\{\max_{1\leq i\leq n}\left|\frac{\partial^m W_n(h,x,X_i)}{\partial x^m}\right|^{\rho}\right\}\sum_{i=1}^{n}\left\{\frac{\partial^m W_n(h,x,X_i)}{\partial x^m}\right\}^2 \longrightarrow 0$$

in probability for all $(h,x) \in H \times I$. *Also, assume that as* h *varies in* H *and* x *varies in* I, $VAR_{G_n}(Y_i|X_i)\left\{\frac{\partial^{m+2} W_n(h,x,X_i)}{\partial h \partial x^{m+1}}\right\}^2$ *will be uniformly dominated by a positive function* $M(X_i)$ *such that* $\sup_{n\geq 1} E_{G_n}\{M(X_i)\} < \infty$. *Then as* $n \longrightarrow \infty$, *the 2-parameter stochastic process*

$$n^{1/2}\left[\frac{\partial^m \hat{f}_h(x)}{\partial x^m} - \frac{\partial^m E\{\hat{f}_h(x)\}}{\partial x^m}\right]$$

with $(h,x) \in H \times I$ *converges weakly to a Gaussian process on* $H \times I$ *with zero mean and covariance function* $cov(h_1,x_1,h_2,x_2)$.

When $F_n \equiv F$ or $G_n \equiv G$ for all $n \geq 1$ (G being a fixed bivariate distribution), Theorems 3.1 and 3.2 yield the weak convergence of the empirical scale space surfaces and their derivatives under the standard i.i.d set up. On the other hand, if we take $F_n = \hat{F}_n$ i.e. the usual empirical distribution of the univariate data, or $G_n = \hat{G}_n$ i.e. the usual empirical distribution of the bivariate data, in view of the uniform strong consistency of the empirical distribution function based on i.i.d data (Glivenko-Cantelli theorem), we get the bootstrap versions of the weak convergence results. In that case, Theorems 3.1 and 3.2 imply that the weak convergence of the scale space surfaces (as well as their derivatives) to appropriate Gaussian processes will hold even under the bootstrap or resampled distributions. These results are useful in

setting up bootstrap confidence sets for theoretical scale space surfaces and their derivatives and also for carrying out bootstrap tests of significance for their features [see Chaudhuri and Marron (1997)].

Note that the conditions assumed on the kernel function in Theorem 3.1 are satisfied for many standard kernels including the Gaussian kernel. Similarly, the conditions assumed on the weight function in Theorem 3.2 are satisfied for many standard kernel regression estimates and kernel weighted local polynomial estimates for suitable distributions of (Y, X). Observe that a natural estimate for the covariance function in the case of density estimation is

$$\widehat{cov}(h_1, x_1, h_2, x_2) \;=\; n^{-1} \sum_{i=1}^{n} \frac{\partial^m h_1^{-1} K\{(x_1 - X_i)/h_1\}}{\partial x_1^m} \frac{\partial^m h_2^{-1} K\{(x_2 - X_i)/h_2\}}{\partial x_2^m}$$

$$- \; n^{-2} \left(\sum_{i=1}^{n} \frac{\partial^m h_1^{-1} K\{(x_1 - X_i)/h_1\}}{\partial x_1^m} \right) \left(\sum_{i=1}^{n} \frac{\partial^m h_2^{-1} K\{(x_2 - X_i)/h_2\}}{\partial x_2^m} \right) ,$$

which can be computed from the data in a straight forward way. Similarly, in the regression problem, a natural estimate for the covariance function is

$$\widehat{cov}(h_1, x_1, h_2, x_2) \;=\; n^{-1} \sum_{i=1}^{n} \widehat{VAR}(Y_i | X_i) \frac{\partial^m W_n(h_1, x_1, X_i)}{\partial x_1^m} \frac{\partial^m W_n(h_2, x_2, X_i)}{\partial x_2^m} ,$$

which too can be easily computed from the data once we have a suitable estimate for the conditional variance $VAR(Y_i|X_i)$.

We will now state the last theorem in this section, which is related to the behavior of the difference between the empirical and the theoretical scale space surfaces under the supremum norm on $H \times I$ and the uniform convergence of the empirical version to the theoretical one as the sample size grows.

Condition A1 : *In the set up of Theorem 3.1, let the smooth kernel* $K(x)$ *be such that for integer* $m \geq 0$*, the derivatives* $\dfrac{\partial^{m+1} h^{-1} K(x/h)}{\partial x^{m+1}}$ *and*

$\frac{\partial^{m+1} h^{-1} K(x/h)}{\partial h \partial x^m}$ *both remain uniformly bounded as* h *varies in* H *and* x *varies in* $(-\infty, \infty)$.

Condition A2 : *In the set up of Theorem 3.2, as* h *varies in* H *and* x *varies in* I*, both of*

$$VAR_{G_n}(Y_i|X_i) \left\{ \frac{\partial^{m+1} W_n(h, x, X_i)}{\partial x^{m+1}} \right\}^2$$

and

$$VAR_{G_n}(Y_i|X_i) \left\{ \frac{\partial^{m+1} W_n(h, x, X_i)}{\partial h \partial x^m} \right\}^2$$

are uniformly dominated by a positive function $M^*(X_i)$ *such that* $\sup_{n \geq 1} E_{G_n}\{M^*(X_i)\} < \infty$.

Theorem 3.3 : *Assume either the set up of Theorem 3.1 and Condition A1 or that of Theorem 3.2 and Condition A2. Then as* $n \longrightarrow \infty$, $\sup_{x \in I, h \in H} n^{1/2} \left| \frac{\partial^m \hat{f}_h(x)}{\partial x^m} - \frac{\partial^m E\{\hat{f}_h(x)\}}{\partial x^m} \right|$ *converges weakly to a random variable that has the same distribution as that of* $\sup_{x \in I, h \in H} |Z(h, x)|$. *Here* $Z(h, x)$ *with* $h \in H$ *and* $x \in I$ *is a Gaussian process with zero mean and covariance function* $cov(h_1, x_1, h_2, x_2)$ *as defined in Theorem 3.1 or Theorem 3.2 so that*

$$Pr\{\ Z(h, x) \text{ is continuous for all } (h, x) \in H \times I\ \} = 1\ ,$$

and consequently $Pr\left\{ \sup_{x \in I, h \in H} |Z(h, x)| < \infty \right\} = 1.$

It immediately follows from the preceding theorem that we have

$$\sup_{x \in I, h \in H} \left| \frac{\partial^m \hat{f}_h(x)}{\partial x^m} - \frac{\partial^m E\{\hat{f}_h(x)\}}{\partial x^m} \right| = O_P(n^{-1/2}) \text{ as } n \longrightarrow \infty\ .$$

This uniform $n^{1/2}$-consistency of the empirical scale space surfaces and their derivatives ensure convergence of the empirical versions of the "critical bandwidths" and the "mode tree" to their theoretical (or population) counterparts as the sample size grows.

Note that all the weak convergence results in this section have been established under the assumption that both of H and I are fixed compact subintervals of $(0, \infty)$ and $(-\infty, \infty)$ respectively. Compactness of the set $H \times I$ enables us to exploit standard results on weak convergence of a sequence of probability measures on a space of continuous functions defined on a common compact metric space. However, conventional asymptotics for nonparametric curve estimates allows the smoothing parameter h to shrink with growing sample size. There frequently one assumes that h_n is of the order $n^{-\gamma}$ for some appropriate choice of $0 < \gamma < 1$ so that the estimate $\hat{f}_{h_n}(x)$ converges to the "true function" $f(x)$ at an "optimal rate". This makes one wonder about the asymptotic behavior of the empirical scale space surface when h varies in $H_n = [an^{-\gamma}, b]$, where $a, b > 0$ are fixed constants. Extension of our weak convergence results along that direction will be quite interesting, and we leave it as a challenging open problem here.

4 Some Applications

In nonparametric curve estimation a question of fundamental importance is which of the observed features in an estimated curve are really significant, and which ones are spurious artifacts of random noise in the data. In the scale space literature, "blobs" in scale space surface are used as the primary tools for assessing significance of peaks observed in smooth curves at various levels of scale. Readers are referred to Lindeberg (1994) for detailed discussion on "blobs" and related mathematics. On the other hand, in the statistics literature on mode testing [see e.g. Good and Gaskins (1980), 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), Marchette and Wegman (1997), Minnotte (1997)], various statistical tests have been proposed for measuring the significance of modes in estimated curves. It will be interesting to note here that some of these tests, which are based on Silverman's "critical bandwidths", are comparable with the significance measures based on "lifetimes of blobs" (i.e. the ranges of the scale over which the "blobs" exist in the scale space surface). Similarly, other measures of "blob" significance that are obtained from the sizes and spatial extents of "blobs" have close connection with statistical tests based on Müller and

Sawitzki's "excess mass estimates".

We have already pointed out that features like peaks, valleys, points of inflexion, etc. of a smooth curve can be characterized in terms of zero crossings of derivatives. Hence the significance of such features, as discussed in Section 2, can be judged from statistical significance of zero crossings or equivalently the sign changes of derivatives. This idea has been successfully exploited by Chaudhuri and Marron (1997) in developing a simple yet effective tool called SiZer for exploring significant structures in curves. Let us now consider the null hypothesis $H_0^{(h,x)} : \frac{\partial^m E\{\hat{f}_h(x)\}}{\partial x^m} = 0$ for a fixed $x \in I$ and an $h \in H$. Then a statistical test can be carried out for this hypothesis based on $\frac{\partial^m \hat{f}_h(x)}{\partial x^m}$, and if $H_0^{(h,x)}$ is rejected, one can claim to have statistically significant evidence for $\frac{\partial^m E\{\hat{f}_h(x)\}}{\partial x^m}$ being positive or negative depending on the sign of $\frac{\partial^m \hat{f}_h(x)}{\partial x^m}$. Then each point of significant zero crossing of $\frac{\partial^m \hat{f}_h(x)}{\partial x^m}$ at a given level of scale (i.e. h) will be located between a pair of points $x_1, x_2 \in I$ such that there will be significant evidence in the data against both of $H_0^{(h,x_1)}$ and $H_0^{(h,x_2)}$, and $\frac{\partial^m \hat{f}_h(x_1)}{\partial x_1^m}$ and $\frac{\partial^m \hat{f}_h(x_2)}{\partial x_2^m}$ will have opposite signs. We have already seen in the preceding section that the process $n^{1/2}\left[\frac{\partial^m \hat{f}_h(x)}{\partial x^m} - \frac{\partial^m E\{\hat{f}_h(x)\}}{\partial x^m}\right]$ has a limiting Gaussian distribution, and this can be used to construct simultaneous asymptotic tests for the family of hypotheses $\{H_0^{(h,x)} | h \in H, x \in I\}$. We now state a theorem that highlights performance of such tests.

Theorem 4.1 : *Assume that either all the conditions in Theorem 3.1 and Condition A1 (in the case of density estimation) or those in Theorem 3.2 and Condition A2 (in the case of regression) hold. Let $q_{(1-\alpha)}$ be the $(1-\alpha)$-th quantile of the continuous distribution of $\sup_{x \in I, h \in H} |Z(h,x)|$, where $Z(h,x)$ is the Gaussian process on $H \times I$ with covariance function $cov(h_1, x_1, h_2, x_2)$ introduced in Theorem 3.3. Consider the statistical test that accepts the*

null hypothesis $H_0^{(h,x)} : \frac{\partial^m E\{\hat{f}_h(x)\}}{\partial x^m} = 0$ *if* $\left|\frac{\partial^m \hat{f}_h(x)}{\partial x^m}\right| \leq n^{-1/2} q_{(1-\alpha)}$ *and concludes significant evidence for* $\frac{\partial^m E\{\hat{f}_h(x)\}}{\partial x^m}$ *being positive or negative if* $\frac{\partial^m \hat{f}_h(x)}{\partial x^m} > n^{-1/2} q_{(1-\alpha)}$ *or* $< -n^{-1/2} q_{(1-\alpha)}$ *respectively. Then the asymptotic simultaneous level of this test for the entire family of hypotheses* $\{H_0^{(h,x)} | x \in I, h \in H\}$ *will be* α*. In other words, if the hypotheses* $H_0^{(h,x)}$ *are true for all* $(h,x) \in S \subseteq H \times I$*, all of them will be accepted by the test with asymptotic probability at least* $(1-\alpha)$ *as* $n \longrightarrow \infty$*. Further, this test will have the property that for any fixed* $h \in H$*, if* $\frac{\partial^m E\{\hat{f}_h(x)\}}{\partial x^m}$ *has* $k \geq 1$ *sign changes (i.e. k zero crossings) as x varies in I, the test will detect significant evidence for all of these k sign changes with probability tending to one as* $n \longrightarrow \infty$*.*

Finding the exact distribution of the supremum of the absolute value of a general Gaussian process is an almost impossible task, and the results available in the literature can only provide exponential bounds for the probability $Pr\left\{\sup_{x \in I, h \in H} |Z(h,x)| > \lambda\right\}$ for $\lambda > 0$ [see e.g. Adler (1990) for some detailed discussion]. Therefore in practice it may not be possible to evaluate the quantile $q_{(1-\alpha)}$ analytically, and one may have to use some approximation for it such as an estimate based on the bootstrap [see Chaudhuri and Marron (1997)]. Note that so long as such an approximation converges to the true quantile as the sample size grows to infinity, the asymptotic performance of the test described in the preceding theorem remains unaffected. One nice feature of the statistical test considered here is that it tries to detect the positions of significant zero-crossings in the derivative of the scale space surface in addition to the number of such zero crossings at different levels of the scale. Mode testing procedures considered in the literature however focus only on the number of modes of the curve, with little or no attention to their positions.

Figure 5a shows the SiZer map for the data shown in Figure 1a. Regions in scale space are shaded blue for significantly increasing, red for significantly decreasing, purple for unable to distinguish (i.e. the confidence interval for the derivative contains the origin), and gray for insufficient data in each

window. The SiZer map shows that the underlying regression is significantly increasing near $x = 0.1$, and near $x = 0.4$, and is significantly decreasing near $x = 0.25$ and $x = 0.7$. However the spikes in the regression curve, shown in Figure 1a, at $x = 0.65$ and $x = 0.75$ are not discernible from the data with this level of noise. Many more applications of SiZer are shown in Chaudhuri and Marron (1997).

[put Figure 5 about here]

FIGURE 5: *Figure 5a shows the SiZer map corresponding to the data and the family of smooths shown in Figure 1b. This shows which modes in the smooths are significant, and which are spurious sampling artifacts. Figure 5b shows the family surface with panels shaded according to the SiZer colors.*

A statistical test that has simultaneous asymptotic level α for all of the null hypotheses $H_0^{(h,x)}$ as x varies in I and h varies in H may turn out to be overly conservative in many finite sample situations. If necessary, one may consider tests that have simultaneous asymptotic level α only for the hypotheses $H_0^{(h,x)}$ as x varies in I for some fixed $h \in H$. In some sense, it may be quite reasonable to consider different levels of the scale separately and carry out separate tests for different curves corresponding to different values of h instead of pooling those curves together and conducting one simultaneous test for all of them. In this case, one has to use $(1-\alpha)$-th quantiles of the distributions of $\sup_{x \in I} |Z(h,x)|$ for different $h \in H$. Readers are referred to Chaudhuri and Marron (1997) for detailed discussion on different statistical tests for significant zero crossings of the derivative of the scale space surface and many illustrative examples that demonstrate their numerical implementation and performance.

5 Appendix : The Proofs

Proof of Theorem 2.1 : Let us denote the theoretical scale space surface $E\{\hat{f}_h(x)\}$ by $g_h(x)$. First observe that since E here means conditional expectation given $X_1, X_2, \ldots X_n$ and the Y_i's are assumed to be conditionally independent given the X_i's, in the case of Priestley-Chao estimate (A), we have

$$g_h(x) \;=\; (nh)^{-1} \sum_{i=1}^{n} E(Y_i|X_i) K\{(x - X_i)/h\} \;,$$

while in the case of Gasser-Müller estimate (B), we have

$$g_h(x) \;=\; \sum_{i=1}^{n} E(Y_i|X_i) \int_{t_{i-1}}^{t_i} (1/h) K\{(x - s)/h\} \;,$$

where $-\infty = t_0 < X_1 < t_1 < X_2 < t_2 < \ldots < t_{n-1} < X_n < t_n = \infty$ as before. Next observe that for Gaussian kernel $K(x) = (1/\sqrt{2\pi}) \exp(-x^2/2)$ and any integer $m \geq 0$, we have

$$\frac{\partial^m \hat{f}_{h_1}(x)}{\partial x^m} * K(x/h_2) = \frac{\partial^m \hat{f}_{\sqrt{h_1^2+h_2^2}}(x)}{\partial x^m} \text{ and } \frac{\partial^m g_{h_1}(x)}{\partial x^m} * K(x/h_2) = \frac{\partial^m g_{\sqrt{h_1^2+h_2^2}}(x)}{\partial x^m}$$

for all $h_1, h_2 > 0$ and both of Priestley-Chao and Gasser-Müller estimates. Here $*$ denotes usual convolution, and note that we are using the fact that $K(x/h_1) * K(x/h_2) \;=\; K\left(x/\sqrt{h_1^2 + h_2^2}\right)$. Now it follows from total positivity of Gaussian kernel and the variation diminishing property of functions generated by convolutions with totally positive kernels [see Schoenberg (1950), Karlin (1968)] that the number of sign changes in $\dfrac{\partial^m \hat{f}_h(x)}{\partial x^m}$ and that in $\dfrac{\partial^m g_h(x)}{\partial x^m}$ will both be monotonically decreasing function of h.

Suppose next that $\dfrac{\partial^m \hat{f}_{h_0}(x)}{\partial x^m}$ has $k \geq 0$ sign changes for some fixed $h_0 > 0$. Then arguing as in Silverman (1981), it is easy to see using the continuity of $\dfrac{\partial^m \hat{f}_h(x)}{\partial x^m}$ as a function of h and x that there exists $\epsilon > 0$ such that for all $h \in [h_0, h_0 + \epsilon)$ $\frac{\partial^m \hat{f}_h(x)}{\partial x^m}$ will have at least k sign changes. Hence the monotonic decrease in the number sign changes with increase in h implies that the number of sign changes in $\dfrac{\partial^m \hat{f}_h(x)}{\partial x^m}$ will be exactly equal to k for all $h \in [h_0, h_0 + \epsilon)$. An identical argument can be given for the number of sign changes in $\dfrac{\partial^m g_h(x)}{\partial x^m}$. This completes the proof of right continuity. $\square$

Proof of Theorem 2.2 : Let $(x_0, h_0, \hat{f}_{h_0}(x_0))$ be a non-degenerate critical point i.e. $\dfrac{\partial \hat{f}_{h_0}(x_0)}{\partial x_0} = 0$ and $\dfrac{\partial^2 \hat{f}_{h_0}(x_0)}{\partial x_0^2} \neq 0$. Then using the continuity of $\frac{\partial^2 \hat{f}_h(x)}{\partial x^2}$ as a function of h and x there will be an $\epsilon > 0$ and a $\delta > 0$ such that $\dfrac{\partial^2 \hat{f}_h(x)}{\partial x^2}$ will be non-zero and have the same sign for all $h \in (h_0 - \epsilon, h_0 + \epsilon)$ and $x \in (x_0 - \delta, x_0 + \delta)$. Further, in view of the implicit function theorem of calculus, ϵ and δ can be so chosen that for every $h \in (h_0 - \epsilon, h_0 + \epsilon)$ there will be a *unique* $x = x(h) \in (x_0 - \delta, x_0 + \delta)$ satisfying $\dfrac{\partial \hat{f}_h(x)}{\partial x} = 0$, and $(x(h), h, \hat{f}_h\{x(h)\})$ will automatically be a non-degenerate critical point. Now, if h_0 is a "critical bandwidth", it must correspond to a point of bifurcation $(x_0, h_0, \hat{f}_{h_0}(x_0))$ of the trajectory of critical points on the scale space surface. Hence, if ϵ and δ are sufficiently small, for all $h \in (h_0 - \delta, h_0]$, there will be more than one x's in $(x_0 - \delta, x_0 + \delta)$ satisfying $\dfrac{\partial \hat{f}_h(x)}{\partial x} = 0$. This contradicts the *uniqueness* of $x(h)$ and completes the proof of the first part of the theorem.

For the second part of the theorem let us observe that we have for $x = x(h) \in (x_0 - \delta, x_0 + \delta)$, $\dfrac{\partial \hat{f}_h(x)}{\partial x} = \dfrac{\partial \hat{f}_h\{x(h)\}}{\partial \{x(h)\}} = 0$ for all $h \in (h_0 - \delta, h_0 + \delta)$. Then the rule for differentiation of implicit functions leads to

$$\frac{\partial^2 \hat{f}_h(x)}{\partial h \partial x} \; + \; \frac{\partial^2 \hat{f}_h(x)}{\partial x^2} \frac{\partial x}{\partial h} \; = \; 0 \, ,$$

which implies that at $x = x(h)$, we will have

$$\frac{\partial x}{\partial h} \; = \; - \; \left\{ \frac{\partial^2 \hat{f}_h(x)}{\partial h \partial x} \right\} \left\{ \frac{\partial^2 \hat{f}_h(x)}{\partial x^2} \right\}^{-1} \, .$$

Finiteness of the drift velocity is now immediate. □

Proof of Theorem 3.1 : Let us first fix $(h_1, x_1), (h_2, x_2), \ldots, (h_k, x_k) \in H \times I$ and $t_1, t_2, \ldots, t_k \in (-\infty, \infty)$. Then observe that

$$n^{1/2} \sum_{i=1}^{k} t_i \left[\frac{\partial^m \hat{f}_{h_i}(x_i)}{\partial x_i^m} - \frac{\partial^m E\{\hat{f}_{h_i}(x_i)\}}{\partial x_i^m} \right] \; = \; Z_n \text{ (say)}$$

has zero mean, and its variance converges to $\sum_{i=1}^{k}\sum_{j=1}^{k} t_i t_j cov(h_i, x_i, h_j, x_j)$ as $n \longrightarrow \infty$ in view of the weak convergence of F_n to F and uniform boundedness of $\frac{\partial^m h^{-1} K(x/h)}{\partial x^m}$ as h varies in H and x varies in $(-\infty, \infty)$. Further, uniform boundedness of $\frac{\partial^m h^{-1} K(x/h)}{\partial x^m}$ implies that Lindeberg's condition holds for Z_n, and consequently its limiting distribution will be normal. This in turn implies using Cramer-Wold device that as $n \longrightarrow \infty$, the joint limiting distribution of

$$n^{1/2}\left[\frac{\partial^m \hat{f}_{h_i}(x_i)}{\partial x_i^m} - \frac{\partial^m E\{\hat{f}_{h_i}(x_i)\}}{\partial x_i^m}\right] \;=\; U_n(h_i, x_i) \;\; \text{(say)}$$

for $1 \leq i \leq k$ is multivariate normal with zero mean and $cov(h_i, x_i, h_j, x_j)$ as the (i,j)-th entry of the limiting variance covariance matrix for $1 \leq i, j \leq k$.

Next fix $h_1 < h_2$ in H and $x_1 < x_2$ in I. Then uniform boundedness of $\frac{\partial^{m+2} h^{-1} K(x/h)}{\partial h \partial x^{m+1}}$ implies that

$$\begin{aligned} & E_{F_n}\left\{U_n(h_2, x_2) - U_n(h_2, x_1) - U_n(h_1, x_2) + U_n(h_1, x_1)\right\}^2 \\ \leq \;& nE_{F_n}\left\{\frac{\partial^m \hat{f}_{h_2}(x_2)}{\partial x_2^m} - \frac{\partial^m \hat{f}_{h_2}(x_1)}{\partial x_1^m} - \frac{\partial^m \hat{f}_{h_1}(x_2)}{\partial x_2^m} - \frac{\partial^m \hat{f}_{h_1}(x_1)}{\partial x_1^m}\right\}^2 \\ & \leq C_1 (h_2 - h_1)^2 (x_2 - x_1)^2 \end{aligned}$$

for some constant $C_1 > 0$. It now follows from one of the main results in Bickel and Wichura (1971) that the sequence of processes

$$n^{1/2}\left[\frac{\partial^m \hat{f}_h(x)}{\partial x^m} - \frac{\partial^m E\{\hat{f}_h(x)\}}{\partial x^m}\right]$$

on $H \times I$ will have the tightness property. This completes the proof of the theorem. □

Proof of Theorem 3.2 : Once again let us begin by fixing $(h_1, x_1), (h_2, x_2), \ldots, (h_k, x_k) \in H \times I$ and $t_1, t_2, \ldots, t_k \in (-\infty, \infty)$. Then it is straight forward to verify by direct algebraic computation that the conditional distribution of

$$n^{1/2}\sum_{i=1}^{k} t_i \left[\frac{\partial^m \hat{f}_{h_i}(x_i)}{\partial x_i^m} - \frac{\partial^m E\{\hat{f}_{h_i}(x_i)\}}{\partial x_i^m}\right] \;=\; Z_n \text{ (say)}$$

given $X_1, X_2, \ldots, X_n$ has zero mean, and its variance is

$$n^{-1}\sum_{i=1}^{n}\sum_{j=1}^{n} t_i t_j \sum_{l=1}^{n} VAR_{G_n}(Y_l|X_l)\frac{\partial^m W_n(h_1, x_i, X_l)}{\partial x_i^m}\frac{\partial^m W_n(h_2, x_j, X_l)}{\partial x_j^m}\;,$$

which converges in probability to $\sum_{i=1}^{k}\sum_{j=1}^{k} t_i t_j cov(h_i, x_i, h_j, x_j)$ as $n \longrightarrow \infty$. Also, uniform boundedness of the conditional $(2+\rho)$-th central moment of Y given $X = x$ and the condition that

$$n^{-(1+\rho/2)}\left\{\max_{1\le i\le n}\left|\frac{\partial^m W_n(h, x, X_i)}{\partial x^m}\right|^{\rho}\right\}\sum_{i=1}^{n}\left\{\frac{\partial^m W_n(h, x, X_i)}{\partial x^m}\right\}^2 \;\longrightarrow\; 0$$

in probability as $n \longrightarrow \infty$ together imply that Lindeberg's condition holds for Z_n, and consequently its limiting distribution must be normal. Finally, it follows using Cramer-Wold device as in the proof of Theorem 3.1 that as $n \longrightarrow \infty$, the joint limiting distribution of

$$n^{1/2}\left[\frac{\partial^m \hat{f}_{h_i}(x_i)}{\partial x_i^m} - \frac{\partial^m E\{\hat{f}_{h_i}(x_i)\}}{\partial x_i^m}\right] \;=\; U_n(h_i, x_i) \;\text{ (say)}$$

for $1 \le i \le k$ is multivariate normal with zero mean and $cov(h_i, x_i, h_j, x_j)$ as the (i, j)-th entry of the limiting variance covariance matrix for $1 \le i, j \le k$.

Next fix $h_1 < h_2$ in H and $x_1 < x_2$ in I. Then the last condition assumed in the statement of the theorem implies that

$$\begin{aligned} & E_{G_n}\left\{U_n(h_2, x_2) - U_n(h_2, x_1) - U_n(h_1, x_2) + U_n(h_1, x_1)\right\}^2 \\ =\; & n^{-1}E_{G_n}n^{-1}\sum_{i=1}^{n} VAR(Y_i|X_i)\left\{\frac{\partial^m W_n(h_2, x_2, X_i)}{\partial x_2^m} - \frac{\partial^m W_n(h_2, x_1, X_i)}{\partial x_1^m}\right. \\ & \left. -\frac{\partial^m W_n(h_1, x_2, X_i)}{\partial x_2^m} - \frac{\partial^m W_n(h_1, x_1, X_i)}{\partial x_1^m}\right\}^2 \end{aligned}$$

$$\leq C_2(h_2-h_1)^2(x_2-x_1)^2 E_{G_n}\left\{n^{-1}\sum_{i=1}^{n} M(X_i)\right\} \leq C_3(h_2-h_1)^2(x_2-x_1)^2$$

for some constants C_2 and $C_3 > 0$. It now follows [see Bickel and Wichura (1971)] that the sequence of processes

$$n^{1/2}\left[\frac{\partial^m \hat{f}_h(x)}{\partial x^m} - \frac{\partial^m E\{\hat{f}_h(x)\}}{\partial x^m}\right]$$

on $H \times I$ will have the tightness property, and consequently the assertion in the theorem follows. □

Proof of Theorem 3.3 : Let us begin by observing that for (h_1, x_1) and (h_2, x_2) in $H \times I$, we have

$$E\left\{Z(h_2,x_2) - Z(h_1,x_1)\right\}^2$$

$$= \; cov(h_2,x_2,h_2,x_2) + cov(h_1,x_1,h_1,x_1) - 2cov(h_2,x_2,h_1,x_1)$$

$$\leq \; C_4\{(h_2-h_1)^2 + (x_2-x_1)^2\}$$

for some constant $C_4 > 0$. The above follows straight away from the fact that

$$E\left\{U_n(h_2,x_2) - U_n(h_1,x_1)\right\}^2 \; \leq \; C_4\{(h_2-h_1)^2 + (x_2-x_1)^2\}$$

for all $n \geq 1$ with some appropriate choice of C_4 if either Condition A1 or Condition A2 holds. Here U_n is as in the proof of Theorem 3.1 or 3.2 depending on whether we have a density estimation or a regression problem. Next consider the compact metric space $H \times I$ metrized by the pseudo metric

$$d\{(h_2,x_2),(h_1,x_1)\} \; = \; \left[E\left\{Z(h_2,x_2) - Z(h_1,x_1)\right\}^2\right]^{1/2} ,$$

which is nothing but the so called canonical metric associated with the Gaussian process $Z(h,x)$. Let $N(\epsilon)$ be the smallest number of closed d-balls with radius $\epsilon > 0$ in this metric space that are required to cover $H \times I$. So, $\log\{N(\epsilon)\}$ is the usual metric entropy of $H \times I$ under the metric d.

Note that for any $\epsilon >$ diameter$(H \times I)$, $N(\epsilon) = 1$ and $N(\epsilon) = O(\epsilon^{-2})$ for $0 < \epsilon \leq$ diameter$(H \times I)$. Hence, we must have $\int_0^\infty [\log\{N(\epsilon)\}]^{1/2} d\epsilon < \infty$. This ensures the continuity of the sample paths of the process $Z(h,x)$ as well as the finiteness of $\sup_{x \in I, h \in H} |Z(h,x)|$ with probability one [see Adler (1990, pp. 104–107)]. The proof of the theorem is now complete in view of the weak convergence of the centered and normalized empirical scale space process to the Gaussian process $Z(h,x)$ on $H \times I$ established in Theorems 3.1 and 3.2. □

Proof of Theorem 4.1 : First observe that if $H_0^{(h,x)}$ is true for all $(h,x) \in S \subseteq H \times I$, we have $\dfrac{\partial^m E\{\hat{f}_h(x)\}}{\partial x^m} = 0$ for all $(h,x) \in S$. Hence,

$$Pr\left\{ H_0^{(h,x)} \text{ is accepted for all } (h,x) \in S \right\}$$

$$= Pr\left\{ \left| \frac{\partial^m \hat{f}_h(x)}{\partial x^m} \right| \leq n^{-1/2} q_\alpha \text{ for all } (h,x) \in S \right\}$$

$$\geq Pr\left\{ \sup_{(h,x) \in H \times I} n^{1/2} \left| \frac{\partial^m \hat{f}_h(x)}{\partial x^m} - \frac{\partial^m E\{\hat{f}_h(x)\}}{\partial x^m} \right| \leq q_\alpha \right\}$$

$$\longrightarrow Pr\left\{ \sup_{(h,x) \in H \times I} |Z(h,x)| \leq q_\alpha \right\} = (1-\alpha)\,.$$

Note that the convergence in the last step asserted above follows from the weak convergence results established in Theorems 3.1 and 3.2, and this completes the proof of the first half of the theorem.

Next note that if for a fixed $h \in H$, $\dfrac{\partial^m E\{\hat{f}_h(x)\}}{\partial x^m}$ has $k \geq 1$ sign changes, we will be able to choose $x_1 < x_2 < \ldots < x_k < x_{k+1}$ in I such that $\dfrac{\partial^m E\{\hat{f}_h(x_i)\}}{\partial x_i^m} \neq 0$ for all $1 \leq i \leq k+1$, and $\dfrac{\partial^m E\{\hat{f}_h(x_i)\}}{\partial x_i^m}$ and $\dfrac{\partial^m E\{\hat{f}_h(x_{i+1})\}}{\partial x_{i+1}^m}$ will have opposite signs for all $1 \leq i \leq k$. Since $n^{-1/2} q_\alpha \longrightarrow 0$ as $n \longrightarrow \infty$, the second half of the theorem now follows from the fact that

$$\max_{1 \leq i \leq k+1} \left| \frac{\partial^m \hat{f}_h(x_i)}{\partial x_i^m} - \frac{\partial^m E\{\hat{f}_h(x_i)\}}{\partial x_i^m} \right|$$

$$\leq \sup_{(h,x)\in H\times I} \left| \frac{\partial^m \hat{f}_h(x)}{\partial x^m} - \frac{\partial^m E\{\hat{f}_h(x)\}}{\partial x^m} \right| = O_P(n^{-1/2}) \ ,$$

which has been observed following Theorem 3.3. □

REFERENCES

ADLER, R. J. (1990). *An Introduction to Continuity, Extrema and Related Topics for General Gaussian Processes*. IMS Lecture Note and Monograph Series, vol. 12.

BICKEL, P. J. AND WICHURA, M. J. (1971). Convergence criteria for multiparameter stochastic processes and some applications. *The Annals of Mathematical Statistics*, **42**, 1656–1670.

CHAUDHURI, P. AND MARRON, J. S. (1997). SiZer for exploration of structures in curves. Unpublished manuscript.

CLEVELAND, W. S. (1979). Robust locally weighted regression and smoothing scatterplots. *Journal of the American Statistical Association*, **74**, 829–836.

CLEVELAND, W. L. AND LOADER, C. (1996). Smoothing by local regression: principles and methods, in *Statistical Theory and Computational Aspects of Smoothing*. (Eds. Härdle, W. and Schimek, M. G.), Physica Verlag, Heidleberg, 10–49.

DONOHO, D. L. (1988). One sided inference about functionals of a density. *The Annals of Statistics*, **16**, 1390–1420.

EUBANK, R. L. (1988). *Spline Smoothing and Nonparametric Regression*. Marcel Dekker : New York.

FAN, J. (1992). Design adaptive nonparametric regression. *Journal of the American Statistical Association*, **87**, 998–1004.

FAN, J. (1993). Local linear regression smoothers and their minimax efficiency. *The Annals of Statistics*, **21**, 196–216.

FAN, J. AND GIJBELS, I. (1996). *Local Polynomial Modeling and Its Applications*. Chapman and Hall : London.

FISHER, N. I., MAMMEN, E. AND MARRON J. S. (1994). Testing for multimodality. *Computational Statistics and Data Analysis*, **18**, 499–512.

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.

GREEN AND SILVERMAN, B. W. (1994). *Nonparametric Regression and Generalized Linear Models*. Chapman and Hall : London.

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

HARTIGAN, J. A. AND HARTIGAN, P. M. (1985). The DIP test of multimodality. *The Annals of Statistics*, **13**, 70–84.

HARTIGAN, J. A. AND MOHANTY, S. (1992). The RUNT test for multimodality. *Journal of Classification*, **9**, 63–70.

HIRSCHMAN, I. I. AND WIDDER, D. V. (1955). *The Convolution Transform*. Princeton University Press, Princeton.

KARLIN, S. (1968). *Total Positivity*. Stanford University Press, Stanford.

KOENDERINK, J. J. (1984). The structure of images. *Biological Cybernatics*, **50**, 363–370.

LINDEBERG, T. (1994). *Scale Space Theory in Computer Vision*. Kluwer : Boston.

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.

MARCHETTE, D. J. AND WEGMAN, E. J. (1997). The filtered mode tree. *Journal of Computational and Graphical Statistics*, **6**, 143–159.

MARRON, J. S. AND CHUNG, S. S. (1997). Presentation of smoothers : the family approach. Unpublished manuscript.

MINNOTTE, M. C. (1997). Nonparametric testing of the existence of modes. *The Annals of Statistics*, **25**, 1646–1660.

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.

MÜLLER, H. G. (1988). *Nonparametric Regression Analysis of Longitudinal Data*. Lecture Note in Statistics and Probability, Springer Verlag.

MÜLLER, H. G. AND SAWITZKI, G. (1991). Excess mass estimates and tests for multimodality. *Journal of the American Statistical Association*, **86**, 738–746.

MUZY, J. F., BACRY, E. AND ARNEODO, A. (1994). The multifractal formalism revisited with wavelets. *International Journal of Bifurcation and Chaos*, **4**, 245–302.

ROSENBLATT, M. (1991). *Stochastic Curve Estimation*. NSF-CBMS Regional Conference Series in Probability and Statistics, vol. 3.

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.

SCHOENBERG, I. J. (1950). On Polya frequency functions, II : variation diminishing integral operators of the convolution type. *Acta Scientarum Mathematicarum Szeged*, **12B**, 97–106.

SILVERMAN, B. W. (1981). Using kernel density estimates to investigate multimodality. *Journal of the Royal Statistical Society, Series B*, **43**, 97–99.

SILVERMAN, B. W. (1986). *Density Estimation for Statistics and Data Analysis*. Chapman and Hall : London.

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

STONE, C. J. (1977). Consistent nonparametric regression. *The Annals of Statistics*, **5**, 595–620.

WAHBA, G. (1991). *Spline Models for Observational Statistics*. SIAM Lecture Note, Philadelphia.

WAND, M. P. AND JONES, M. C. (1995). *Kernel Smoothing*. Chapman and Hall : London.

WEICKERT J. (1997). *Anisotropic Diffussion in Image Processing*. Teubner : Stuttgart-Leipzig.

WITKIN, A. P. (1983). Scale space filtering. In *Proceedings of the 8th International Joint Conference on Artificial Intelligence* (Karlsruhe, West Germany), pp. 1019–1022.

WONG, Y. F. (1993). Clustering data by melting. *Neural Computation*, **5**, 89–104.

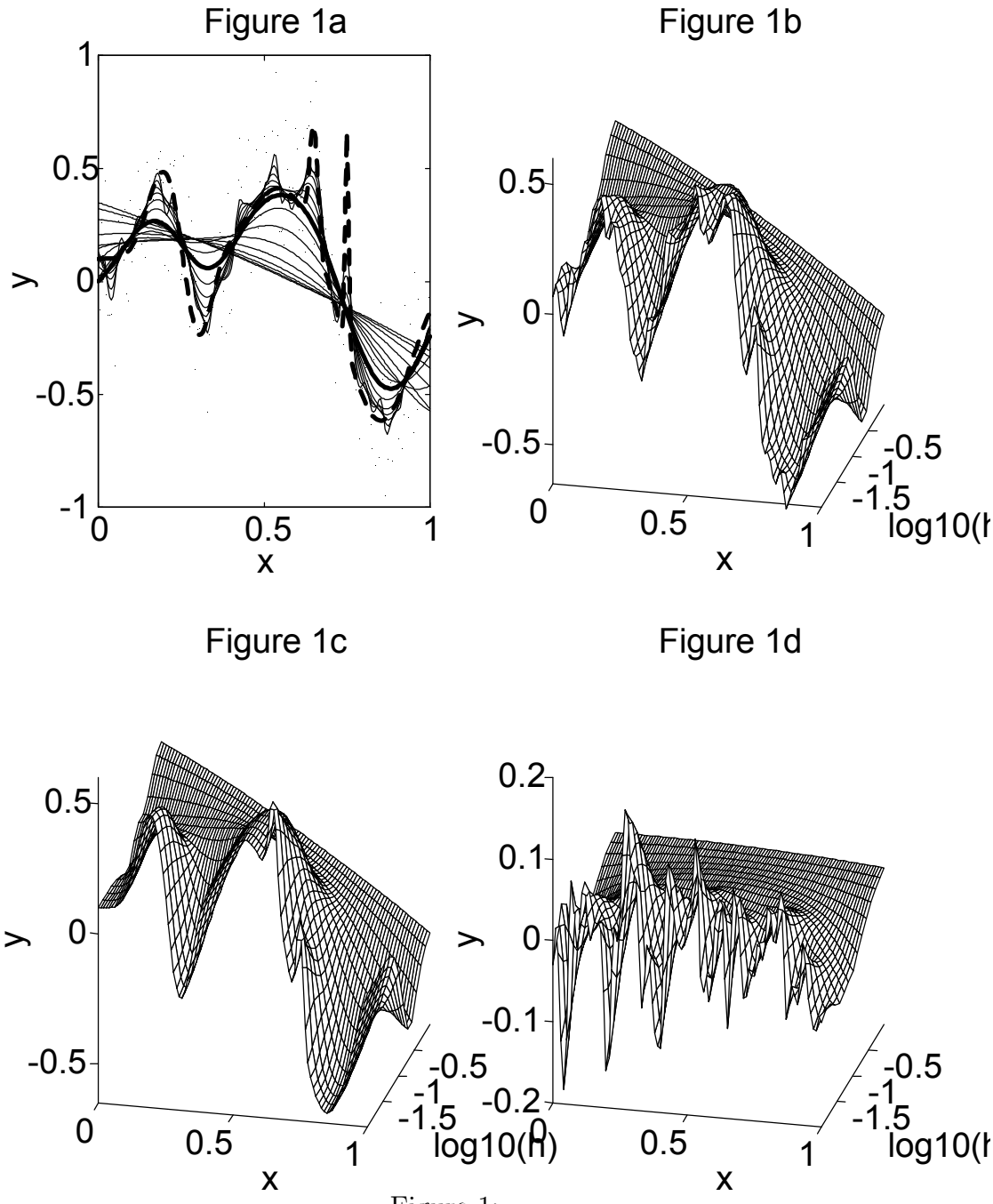
Figure 1a
Figure 1b
Figure 1c
Figure 1d
y
x
log10(h)
1
0.5
0
-0.5
-1
-1.5
0.2
0.1
-0.1
-0.2

Figure 1:

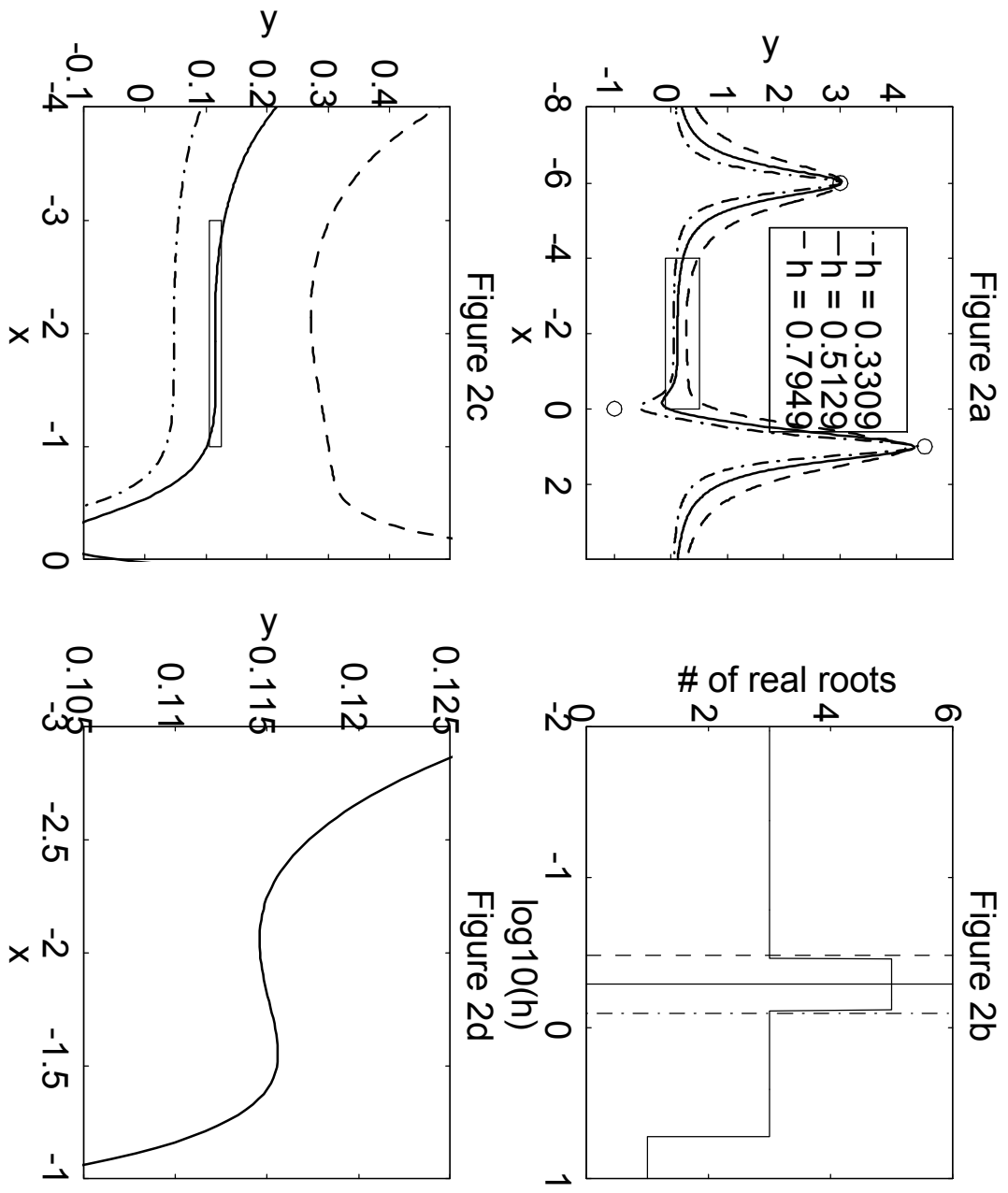
Figure 2a
y
x
h = 0.3309
h = 0.5129
h = 0.7949
Figure 2b
# of real roots
log10(h)
Figure 2c
y
x
Figure 2d
y
x

Figure 2:

Figure 3a

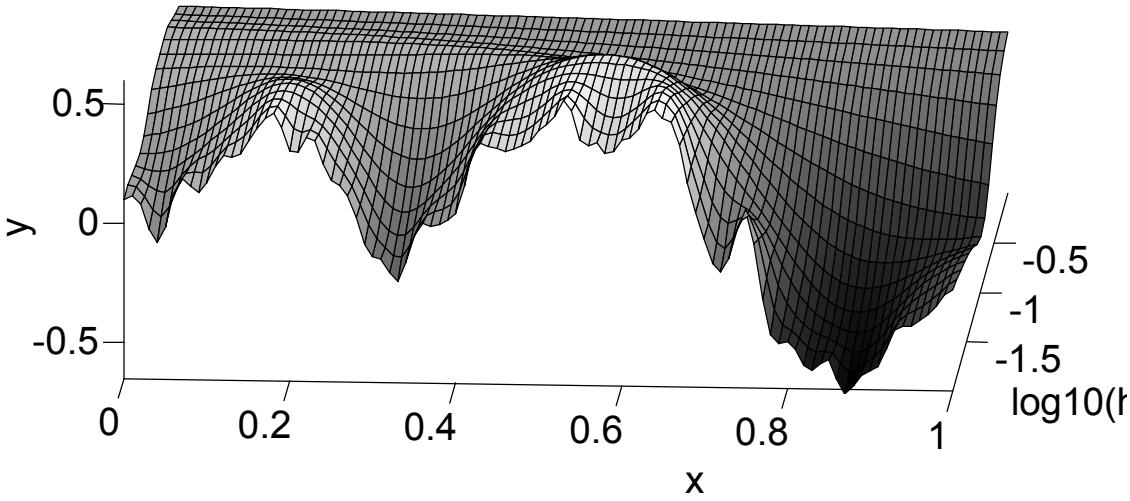

Figure 3b

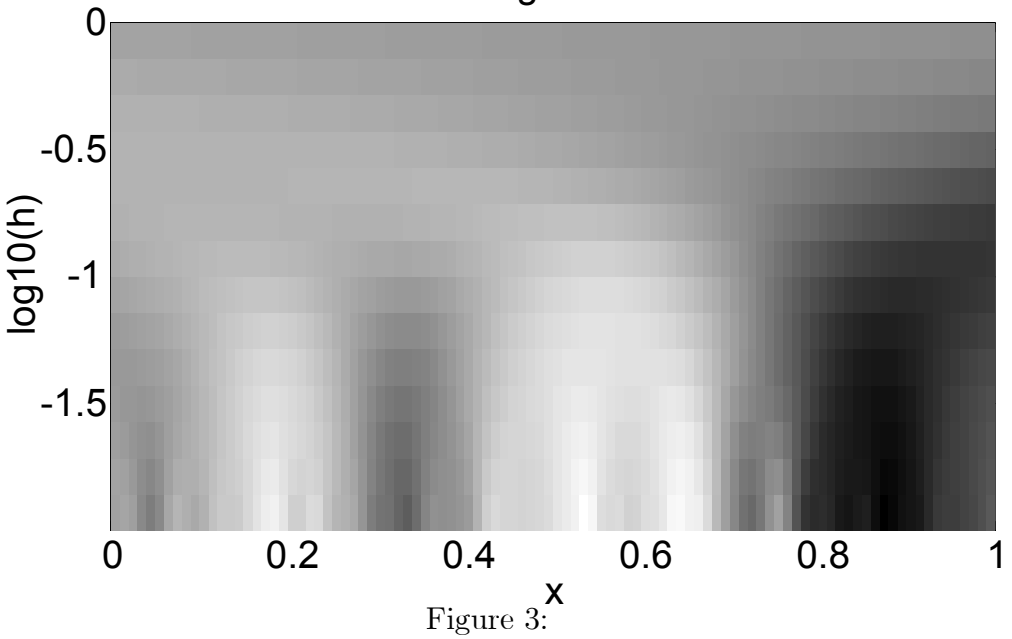

Figure 3:

Figure 4a

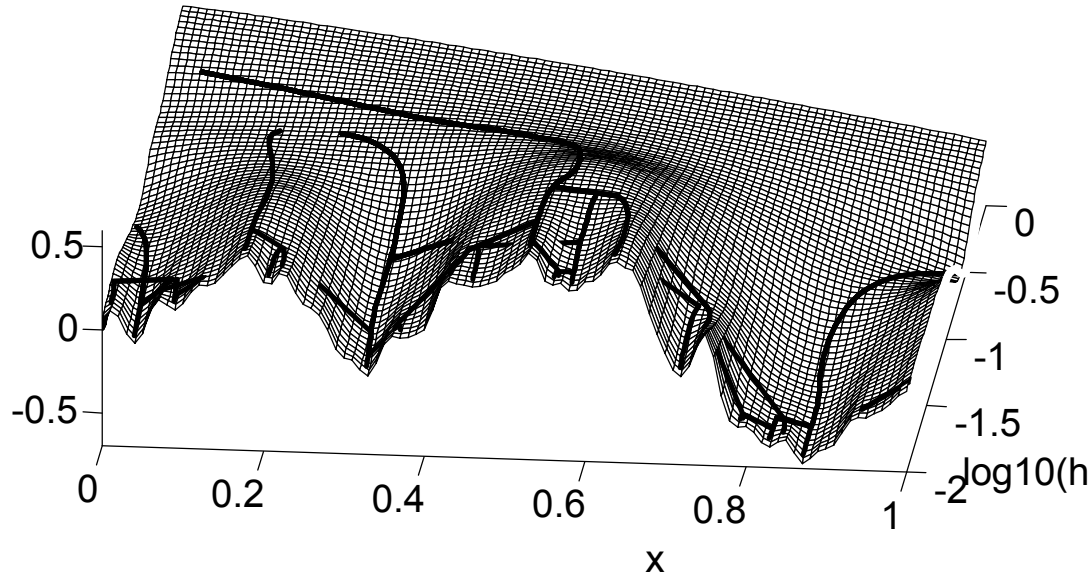

Figure 4b

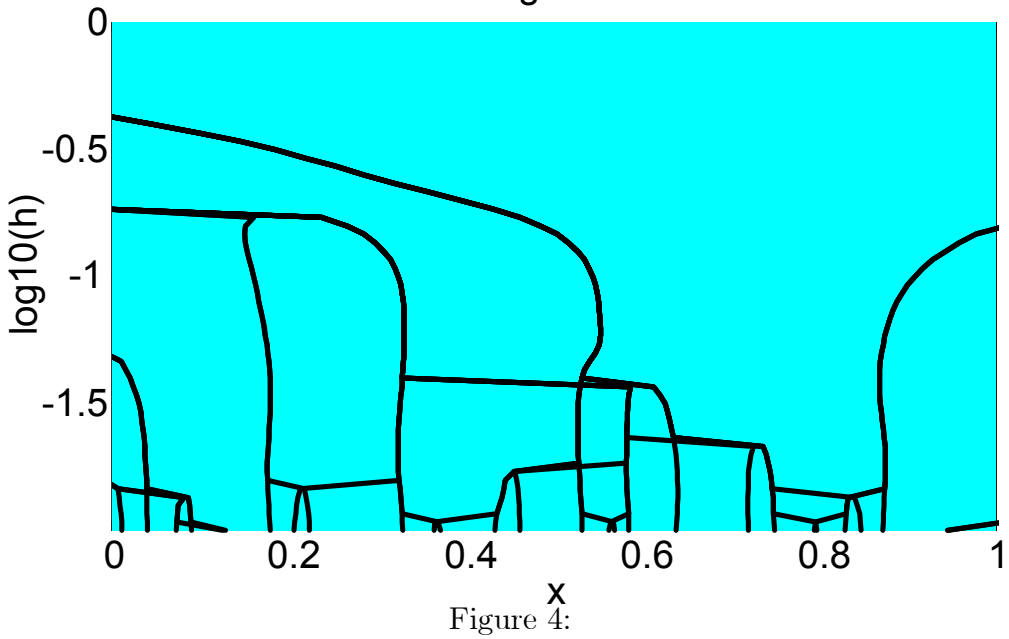

Figure 4:

Figure 5a

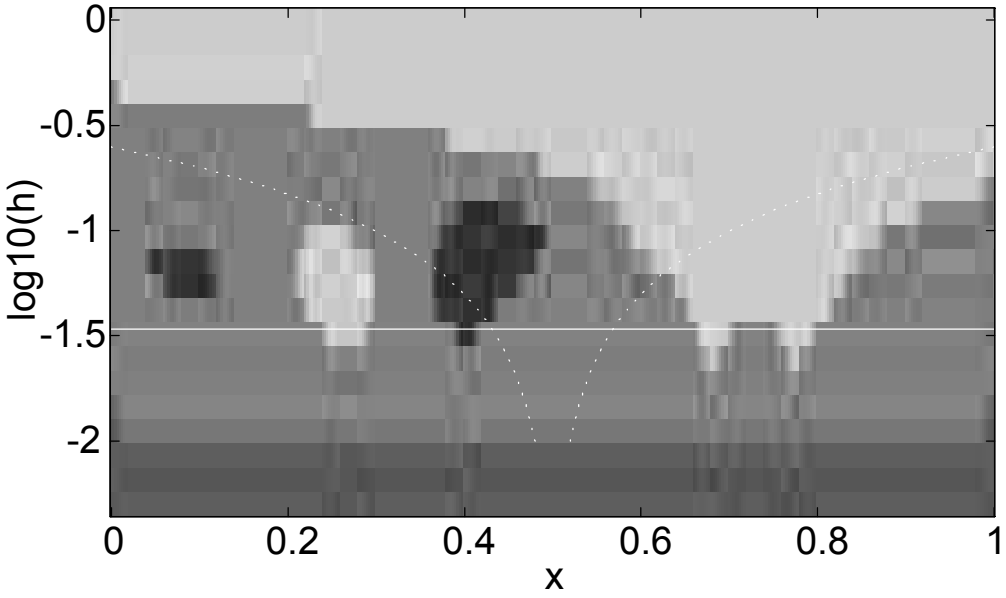

Figure 5b

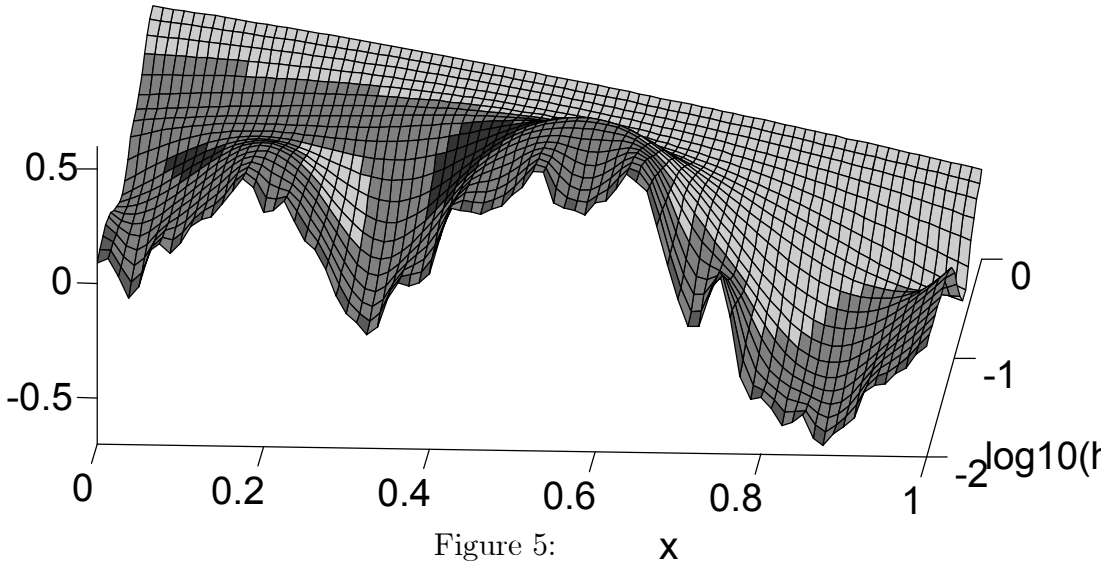

Figure 5: