Alternative Approaches to Significant Zero Crossings (SiZer) Method for Feature Detection in Non-Parametric Univariate Kernel Density Estimation

Boyan Semerdjiev

A Thesis in

The Department of Mathematics and Statistics

Presented in Partial Fulfillment of the Requirements for the Degree

of Master of Science (Mathematics)

at Concordia University

Montreal, Quebec, Canada

July 2014

© Boyan Semerdjiev, 2014

CONCORDIA UNIVERSITY

School of Graduate Studies

This is to certify that the thesis prepared

By: Boyan Semerdjiev

Entitled: Alternative Approaches to Significant Zero Crossings (SiZer) Method for Feature Detection in Non-Parametric Univariate Kernel Density Estimation

and submitted in partial fulfillment of the requirements for the degree of

Master of Science (Mathematics)

complies with the regulations of the University and meets the accepted standards with respect to originality and quality.

Signed by the final Examining Committee:

________________ Chair (Dr. Yogendra Chaubey)

________________ Examiner (Dr. Cody Hyndman)

________________ Supervisor (Dr. Arusharka Sen)

Approved by

_________________ Chair of Department or Graduate Program Director

_________________ Dean of Faculty

_________________ 2014

ABSTRACT

Alternative Approaches to Significant Zero Crossings (SiZer) Method for Feature Detection in Non-Parametric Univariate Kernel Density Estimation

Boyan Semerdjiev

This work presents several methods for feature detection in density estimation of univariate data. Two versions of the original Significant Zero Crossings for Derivatives (SiZer) method and two other SiZer approaches for signal detection with Euler and Hadwiger characteristics are explored. The latter are based on approximating level-crossing probabilities by expected number of upcrossings. In addition, a method for two-sample density comparison is proposed, based on the discussed SiZer methodologies. Estimating a single best bandwidth parameter value is difficult. Therefore, all signal detection and comparison approaches utilize the concept of scale-space and color maps, allowing consideration of curve smoothing at multiple bandwidth levels simultaneously. Finally, the proposed methodologies do not compete and are therefore not compared. Instead, they complement each other and combining the observations from all of them together allows for better statistical inference of the data set.

Contents

List of Figures

Chapter 1 - Introduction to SiZer

One of the most central problems in statistics is the estimation of a model from a data sample. Curve estimation, or curve smoothing, has been extensively researched and many findings have been reported in literature. Wand and Jones (1995) present a convenient method for model estimation, using a symmetric kernel. Given a sample of n iid random variables $X_i, i = 1,2,\dots,n$, with a common probability density function $f(\cdot)$, the estimator at a point x and a bandwidth level h is given as

$$\hat{f}(x,h) = \frac{1}{n}\sum_{i=1}^{n}\frac{1}{h}\,K\left(\frac{x-X_i}{h}\right), \tag{1.1}$$

where $K(\cdot)$ is a symmetric kernel, defined on $(-\infty,+\infty)$, and $\int_{-\infty}^{+\infty}K(x)dx = 1$. For the remainder of this work, the Gaussian kernel is used: $K(x) = \frac{1}{\sqrt{2\pi}}e^{-\frac{x^2}{2}}$. The bandwidth parameter h is a positive number that determines the estimated density $\hat{f}$. The larger the h, the coarser the estimation is, which results in oversmoothing of the curve. A common issue in such a case is missing of important features of the density, such as location of peaks and valleys. Alternatively, small bandwidths result in undersmoothing, or picking too many features of the data set, including noise and spurious inputs. Therefore, the choice of the smoothing parameter is vital for estimation of the correct features of a data set. It is important to distinguish which components represent a signal, and which represent noise.

In this work we try to detect features of probability density functions of several univariate data sets, using the concept of SiZer color maps, presented in section 1.1. In this chapter we also present a slightly improved method for approximating correct tail quantiles of the test-statistic distribution, used in SiZer. Chapter 2 presents alternative methods for approximating the tail probability. One method uses the Hadwiger characteristic when the limiting Gaussian process is assumed to be stationary. Another method uses the number of derivative upcrossings without assuming stationarity. Chapter 3 presents a comparison of two estimated densities with a methodology derived from the original SiZer and the upcrossings counting method. In chapter 4 we present the conclusions and the future work.

The methods presented in chapters 2 and 3 are easily extended to other nonparametric curve estimation problems, such as nonparametric regression, hazard rate, etc.

Section 1.1 - Original SiZer – The Concept

It is never known in advance which bandwidth value will detect the features of a curve accurately. This choice is frequently data-dependent. Here we present a new approach for feature detection by considering a broad range of bandwidth values simultaneously. Originally developed by Chaudhuri and Marron (1999), the SiZer (Significant Zero Crossings for Derivatives) tool provides an overall picture of the significance of features over multiple bandwidths. Instead of overlaying smoothed curves for a range of h parameter values on the same plot, the concept of scale-space is presented, using a color map. A two-dimensional color map is a plot with location on the x-axis and the bandwidth values on the y-axis. The h values are plotted on the logarithmic scale for more accurate visualization. The color map is especially useful for analyzing univariate data sets and estimating their pdfs $f(x)$. The color map is a grid of $m_x$by $n_h$boxes, where the data range is grouped into m_x values and a bandwidth range is chosen and broken down to n_h locations. Then at every point (x, h) of the grid, the value $\widehat{f}\ (x, h)$ is estimated and the following hypothesis is tested:

H_0: $f'(x, h) = 0$

H_1: $f'(x, h) \neq 0$,

where $f'(x, h) = E[\hat{f}'(x, h)]$. If H_0 is not rejected, the slope of the smoothed curve at that point is found to be insignificant and the respective box is colored in purple. If H_0 is rejected, then the density is increasing or decreasing at that location x at bandwidth h. If $\hat{f}'(x, h)$ is significantly positive, then the (x, h) bin is colored in blue, signifying an increasing region of the probability density function. Significantly negative $\hat{f}'(x, h)$ values result in red coloring and a significantly decreasing PDF. Thus, according to Chaudhuri and Marron (2000), the calculated theoretical scale-space surface $E\left(\hat{f}'(x, h)\right) = f'(x, h)$ represents $f'(x)$ at *resolution h*. Since the entire scale-space is divided into a large number of small-sized boxes, each box can be assumed to represent the entire set of (x, h) values in the neighborhood of the points of estimation – with x ranging between $[x_{min}, x_{max}]$ and h between $[h_{min}, h_{max}]$, where h_{min} and h_{max} are both on the logarithmic scale. The resulting color map of blue, purple and red colored boxes provides general overview of where density is increasing or decreasing. Instead of focusing on a single bandwidth, we can conclude that a signal at a given location is significantly increasing or decreasing if the blue or red coloring is solid for a range of h values, or alternatively, the signal is not present in the case of a purple region, or even with blue or red regions over very few bandwidths.

Next, the testing procedure for significance of $\hat{f}'(x, h)$ is presented, as proposed by Chaudhuri and Marron (1999). To test H_0 at each location (x, h), a confidence interval is established for each $\hat{f}'(x,h)$ and it is observed whether zero is included. The confidence interval limits are $\hat{f}'(x,h) \pm q \cdot \widehat{SD}\left(\hat{f}'(x,h)\right)$. Following (1.1) by taking the derivative, we obtain

$$\hat{f}'(x,h) = \frac{1}{n}\sum_{i=1}^{n}\frac{1}{h^2}\,K'\left(\frac{x-X_i}{h}\right) \tag{1.2}$$

$$\widehat{SD}\left(\hat{f}'(x,h)\right) = \sqrt{var\left(\hat{f}'(x,h)\right)}$$

and q is the cutoff quantile. The confidence interval containing zero is equivalent to $\left|\frac{\hat{f}'(x,h)}{\widehat{SD}\left(\hat{f}'(x,h)\right)}\right| < q$. Otherwise, values of $\hat{f}'(x,h)$ are indicative of the signal trend, depending whether they are positive or negative.

The appropriate quantile determination is significant for developing the SiZer approach. Most of this work is focused on determining the cutoff value of $\hat{f}'(x,h)$ to determine significance. Chaudhuri and Marron (1999) have presented several approaches for determining q. At the α significance level, the simplest one is $q_1(h) \equiv q = \Phi^{-1}\left(1-\frac{\alpha}{2}\right)$, known as the pointwise Gaussian quantile. However, this formula derivation of q assumes that the errors are independent and can lead to overestimation of significant features. An improved approach for handling dependencies is determining $q(h)$ simultaneously over many locations of x. The proposed formula is

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

where

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

and $ESS(x,h)$ is the effective sample size at point (x,h), given by

$$ESS(x, h) = \frac{\hat{f}(x,h)}{\frac{1}{h}K(0)}.$$

The concept of effective sample size can be intuitively best understood when the rectangular kernel $K(x) = \frac{1}{2}I(-1 \leq x \leq 1)$ is used. Then $ESS(x, h)$ is equal to the number of data points X_i in the interval $(x - h, x + h)$. However, when using the Gaussian kernel, data points closer to x carry more weight in the ESS determination, which can also result in a non-integer number. In our implementation and simulation of the original SiZer methodology, $q_2(h)$ is used as the cutoff quantile.

Section 1.2 - Implementation of the original SiZer and Results

Two data sets are simulated with the original SiZer approach. The first one contains 5000 iid random variables $X_1 \dots X_{5000}$, randomly generated from the pdf $f(x) = 0.4 \cdot |N(0,0.3)| + 0.3 \cdot (|N(0,0.4)| + 1) + 0.3 \cdot (|N(0,0.2)| + 3)$. Then the sample X_i is smoothed using the SiZer analysis and the features of the theoretical probability density function above are detected on a color map. For the rest of this work, this data sample will be referrer to as the 'half-normal'. The same analysis is done for a 'full-normal' sample $X_1 \dots X_{5000}$, randomly generated from the pdf $f(x) = 0.4 \cdot N(0.5,0.3) + 0.3 \cdot N(2,0.4) + 0.3 \cdot N(3.5,0.2)$. The range of the x values, the points where the densities are estimated vary from 0 to 4, equally cut into 200 bins. The h values range from 0.01 to 1 on the logarithmic scale, divided into 40 intervals. This range is comparable with the values recommended by Chaudhuri and Marron (1999), where the lower bandwidth values are approximately equal to the bin width and the higher range is in the order of the entire interval size of x. The significance level α is assumed to be the standard 0.05. The theoretical density curves, from which we generate the samples, are shown on figures 1.1 and 1.2. The estimated derivatives as a function of (x, h) are given on figures 1.3 and 1.4. The simulated color maps are displayed on figures 1.5 and 1.6.

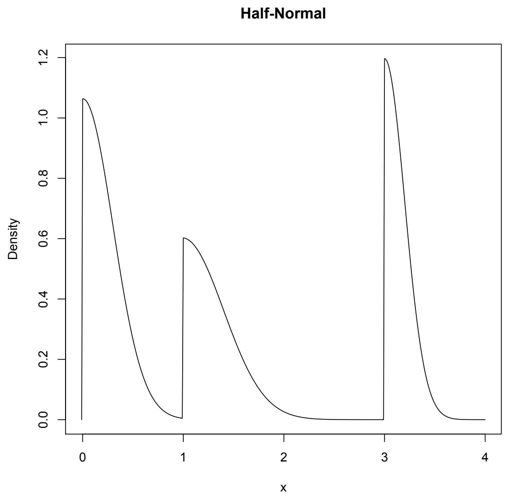


Figure 1.1 – pdf of half-normal data set

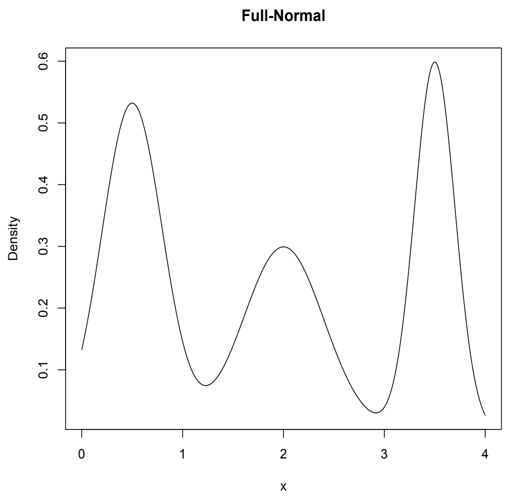


Figure 1.2 – pdf of full-normal data set

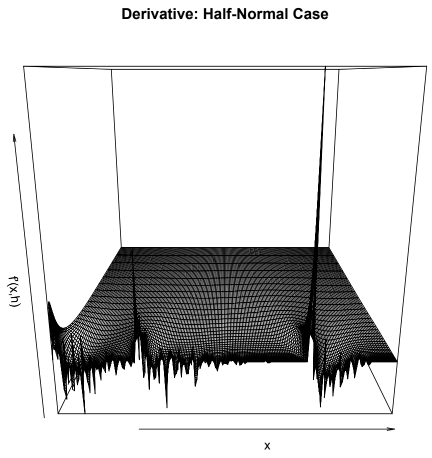


Figure 1.3 – Estimated Derivatives: half-normal

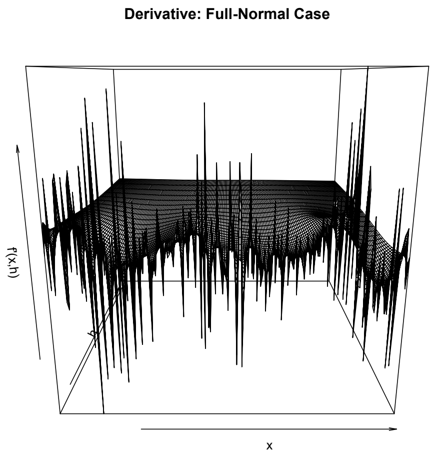
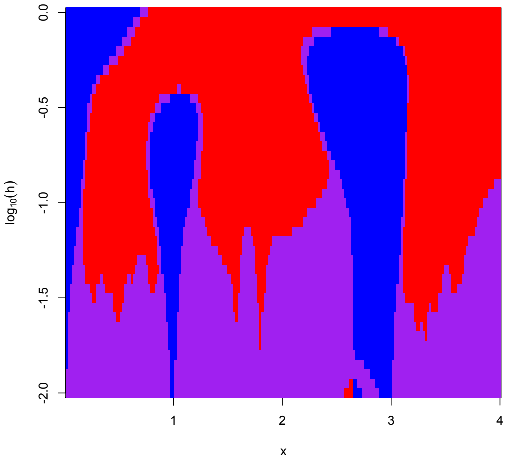


Figure 1.4 – Estimated Derivatives: full-normal



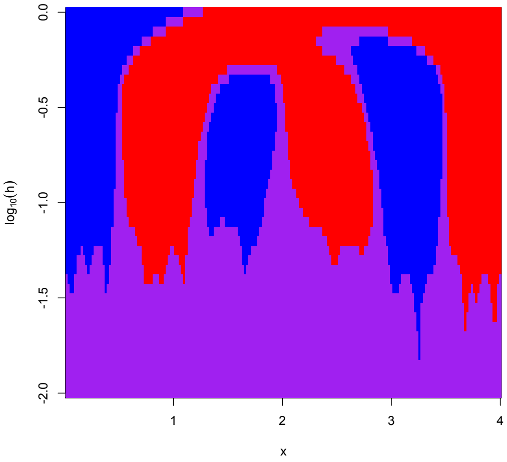


Figure 1.5 – Color map of half-normal data set

Figure 1.6 – Color map of full-normal data set

The color maps clearly demonstrate the features of both half-normal and full-normal density functions over a broad range of bandwidths. For the half-normal data, figure 1.1 shows that the density is rising a lot faster than decreasing and this is reflected in its color map on figure 1.5. The blue regions, or the darkest spots on the color map if the paper is printed on a black-and-white printer, are detected over a broader range of values, where decreasing of the density is not as easy to detect, judging by the smaller bandwidth ranges of observation of the red regions, or the brightest spots on the color map, if the paper is printed on a black-and-white printer. It is also observed that all correct features are observed for the middle range of bandwidths, between -1 and -0.5 on the logarithmic scale, or between 0.1 and 0.3 for *h*. Extreme low bandwidths result in wiggly curve estimation and most of the signal is inconclusive, as detected by the predominantly purple region, or the middle shade of gray if the paper is printed on a black-and-white printer, in the lower part of figure 1.5. In the very top part of figure 1.5, for large bandwidth values, oversmoothing occurs and the three half-normal segments are increasingly merged together, resulting in the mostly red region at the extreme top. Figure 1.3 shows the estimated derivative values. The derivative plot also shows that features are best detected around mid-bandwidth values.

The full-normal data set density on figure 1.2 and its color map on figure 1.6 also agree with each other and exhibit similar behavior to the half-normal density and color map. Intervals of increasing and decreasing are clearly visible for the mid-range bandwidth values, as seen by the alternating blue-red regions. Undersmoothing results in inconclusive purple area, and oversmoothing tends to also merge the three normal curves as one, resulting in a single blue-then-red region. Most importantly, in both data sets the regions of increasing and decreasing are visible over a large mid-range of bandwidths, which reliably proves the signal presence is correctly estimated at each location. Figure 1.4 displays the estimated derivative plot for the full-normal data set and it confirms the data set features.

Section 1.3 - More Advanced SiZer implementation

Similarly to the originally developed SiZer method, an improved method for computing the quantiles is proposed by Hannig and Marron (2006). Let g be the number of bins and the quantiles at every bandwidth level h are denoted by $z_3(h)$. Estimation of the quantiles is based on the fact that

$$P\left(max_{i=1..g}\hat{f}'(x_i,h) - q_3(h)\cdot\widehat{SD}\left(\hat{f}\,(x_i,h)\right) \leq 0\right) = 1-\tfrac{\alpha}{2},$$

as g simultaneous tests need to be performed to test all H_0's. Then

$$1-\tfrac{\alpha}{2} = \Phi\big(z_3(h)\big)^{\theta\cdot g},$$

resulting in

$z_3(h) = \Phi^{-1}\left(1 - \frac{\alpha}{2}\right)^{\frac{1}{\theta \cdot g}}$, as also proposed by Hannig and Lee (2006). Here the parameter θ is called ‘cluster index’ and is characteristic to determining the cutoff values of the derivative. It is given by

$\theta = 2\Phi\left(\sqrt{3\log g} \cdot \frac{\widehat{\Delta}}{2h}\right) - 1$, and $\widehat{\Delta}$ is the width of each bin, equal to $\frac{x_{max} - x_{min}}{g}$. To obtain the formula for the cluster index, Hannig and Marron (2006) make a comparison between the distribution of the maximum value among zero-mean, unit-variance Gaussian random variables and the limiting Gumbel distribution. They simplify a derived formula to estimate more conveniently the maximum value of a stationary series of g Gaussian random variables to $\Phi(x)^{\theta g}$, that is,

$$P\left(max_{i=1..g} T_i \leq x\right) = \Phi(x)^{\theta g} ,$$

and θ is derived from the original limiting distribution.

Using this new approach, new values of the quantiles are obtained and new color maps are created for both the half-normal and full-normal data sets. The plots can be seen on figures 1.7 and 1.8. Comparing them closely with the plots of the originally proposed SiZer implementation on figures 1.5 and 1.6, we can observe that the difference between the two approaches is not significant.

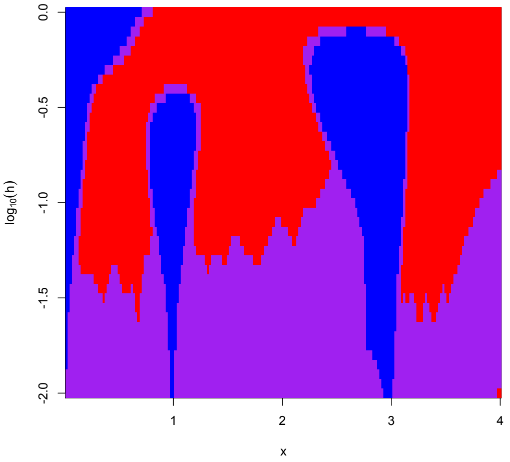


Figure 1.7 – Color map of half-normal data set, advanced quantile calculation

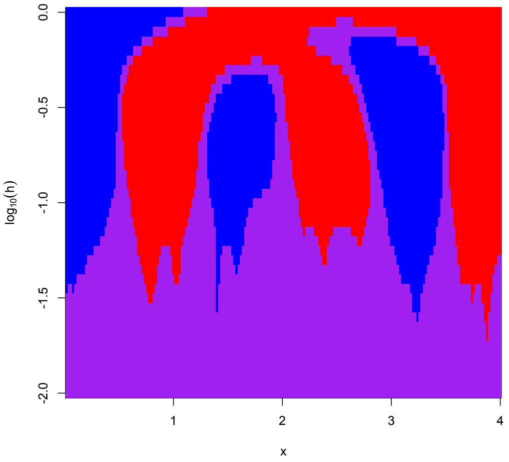


Figure 1.8 – Color map of full-normal data set, advanced quantile calculation

Upon a closer look at the color maps, the improved quantile calculation achieves slightly smoother boundary between the red-purple and the blue-purple zones. In addition, the spurious red and blue boxes at the lowest bandwidth are eliminated. The rise of the half-normal signal is detected over a broader range of bandwidths compared to the full-normal case, due to the significantly sharper jumps, as can be also verified from the probability density functions. Finally, whether the advanced approach is better than the original one remains open to interpretation, but nevertheless, it seems more stable between bandwidths with the current data and the results look a little less bandwidth-dependent.

Chapter 2 – Detecting a Peak or a Valley Using the Tail Probability of the Maximum Value of the Derivative

Concurrently with the original SiZer methods, outlined in chapter 1, a second methodology has been developed. Siegmund and Worsley (1995) and Adler and Taylor (2007) present two similar approaches, where the extreme value of the estimated derivatives over a certain region is compared with a theoretically estimated tail probability to determine whether a signal is significantly increasing or decreasing. Using the binning methodology of the original SiZer, groups of bins are grouped into rectangular clusters and the most extreme derivative value is elected for comparison with the cutoff critical value. The results depend on the level of granularity of the grouping and the two presented methods follow the same concept, however, they differ in the procedure for estimating the cutoff parameter for comparison with the extreme derivative.

Section 2.1 - Testing for Unknown Signal in a Stationary Gaussian Random Field

The first methodology is based on Siegmund and Worsley (1995) who consider a random field $X(t,\sigma)$ with noise W, which is a stationary Gaussian random field. It is given by

$$X(t,\sigma) = \frac{1}{\sqrt{\sigma}} \int k\left(\frac{y-t}{\sigma}\right)\left[\frac{\xi}{\sqrt{\sigma_0}} f\left(\frac{y-t_0}{\sigma_0}\right) dy + dW(y)\right], \tag{2.1}$$

where $f(\cdot)$ is a square integrable function, $k(\cdot)$ is a symmetric kernel and the signal amplitude, location and scale are ξ, t_0 and σ_0, respectively.

For each location t of the one-dimensional spectrum, the presence of signal with an amplitude ξ needs to be tested over a range of bandwidths $\sigma_{min} < \sigma < \sigma_{max}$, that is:

$H_0: \xi = 0$

$H_1: \xi \neq 0$

H_0 is rejected if $max_{t,\sigma}|X(t,\sigma)| > b(\sigma_{min}, \sigma_{max})$, the cutoff value.

In our problem, we have the random field $\hat{f}'(x,h)$, which may be written as, under $H_0: f'(x,h) = E\left[\hat{f}'(x,h)\right] = 0$,

$$\hat{f}'(x,h) = \frac{1}{n}\sum_{i=1}^{n}\frac{1}{h^2}K'\left(\frac{x-X_i}{h}\right) = \frac{1}{h^2}\int K'\left(\frac{x-y}{h}\right)F_n(dy) \approx \frac{1}{h^2}\int K'\left(\frac{x-y}{h}\right)\left[W\big(F(dy)\big) - W\big(F(\infty)\big)F(dy)\right],$$

where $F_n(\cdot)$ is the empirical distribution function of $X_1, \dots, X_n$, $F(\cdot)$ is the cdf of the latter and $W(\cdot)$ is the Wiener process. The approximation above is valid because of the Central Limit Theorem. Comparing with the originally developed SiZer, σ plays the role of the parameter h and $X(t,\sigma)$ is replaced by $\hat{f}'(x,h)$. Clustering together of several bins is performed, for example, every 10 values of x horizontally and every 4 values of h vertically form one new region on the color map and the extreme value of $\hat{f}'(x,h)$ is selected among the original 40 values. The extreme value is the one to compare against the cutoff parameter $b(h)$. With a slight abuse of notation, we will use interchangeably σ and h to signify the bandwidth. We will also often refer to the derivative $\hat{f}'(x,h)$ as the output $X(t,\sigma)$, since this is what we compare to the cutoff value $b(h)$. Note that $\hat{f}'(x,h)$, although approximately a Gaussian random field by the Central Limit Theorem, is not stationary. Nevertheless, we use the Siegmund and Worsley (1995) method here as a first approximation. The stationarity assumption is removed in the next section.

For determination of the cutoff value, the Euler and Hadwiger characteristics of excursion sets are used. Let the excursion set $A_b = \{(t, \sigma): |X(t, \sigma)| \geq b\}$ be the collection of all points (t, σ) where $|X(t, \sigma)|$ is above b, as proposed in (Siegmund & Worsley, 1995) and (Worsley, 1995). Also, let $\psi(A_b)$ be the Hadwiger characteristic of the excursion set A_b. It counts the number of connected components, or 'blobs', minus the number of 'holes' (Chamandy et al., 2008). For the scope of this work, the Euler and Hadwiger characteristics serve identical purposes (Siegmund & Worsley, 1995). To further illustrate the concept of Euler characteristic and excursion sets, we refer to figure 2.1, taken from Hasofer (1978). A convex region at the bottom of a compact subset of R^n brings a contribution of +1, while a concave region at the bottom brings a contribution of -1 to the Euler characteristic, as illustrated at the top of figure 2.1. For each subset, the Euler characteristic is calculated as the difference between the number of positive and negative contributions.

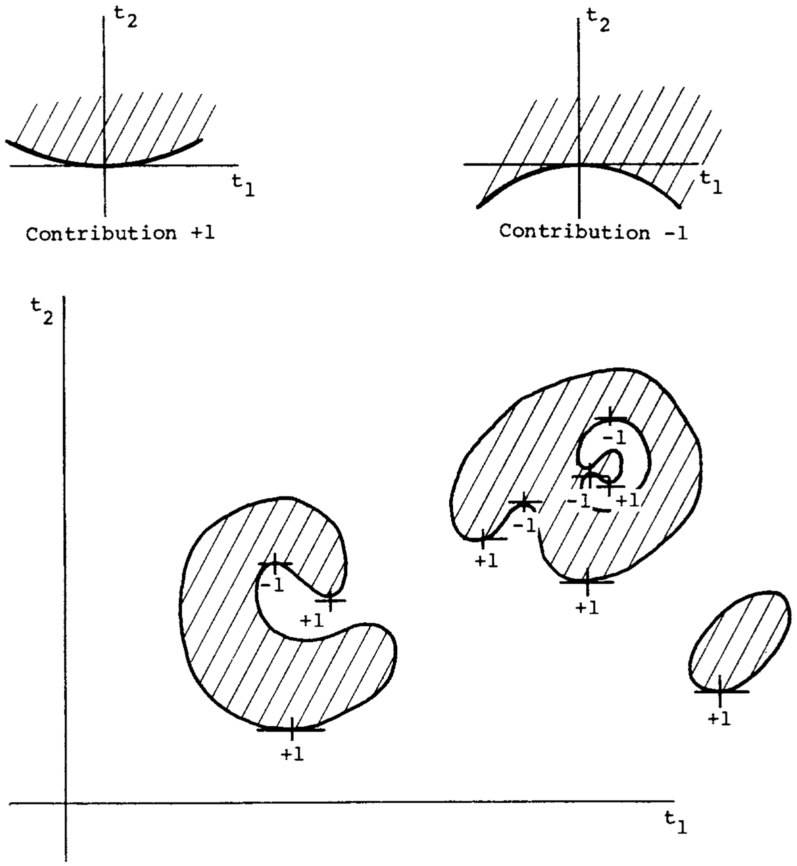


Figure 2.1 – Euler Characteristic Calculations for Various Geometries of Excursion Sets

In the univariate case, the Euler, or Hadwiger characteristic is reduced to the number of upcrossings of the signal at level b. As illustrated on figure 2.2, the Euler characteristic is equal to the number of disjoint sets of $X(t, \sigma)$ above level b for fixed σ.

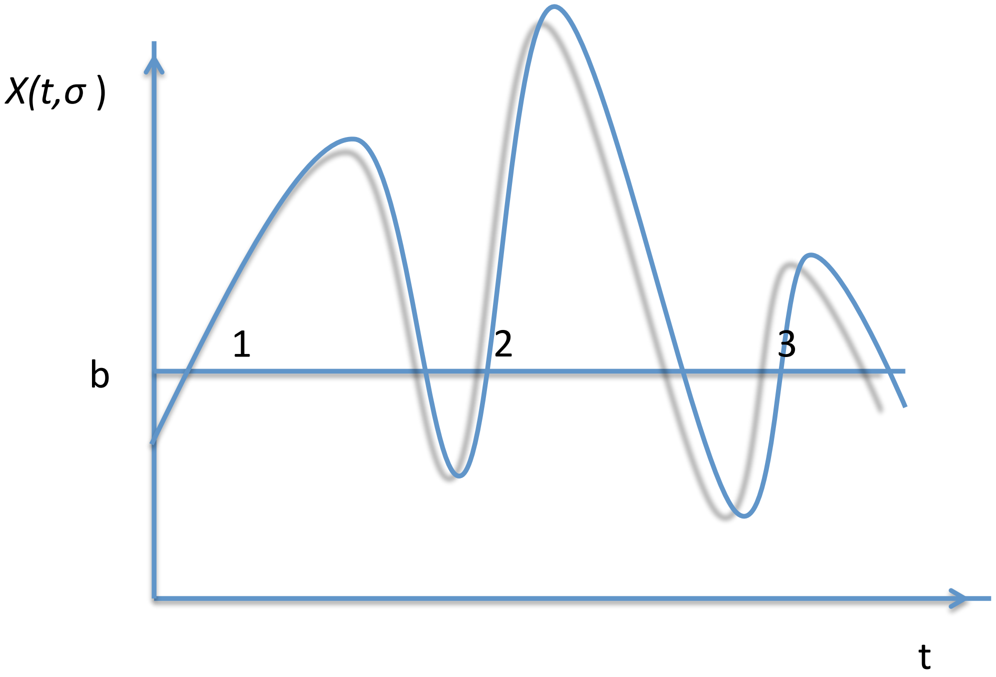


Figure 2.2 – Upcrossings Example in One Dimension. In this case $(\psi(A_b) = 3$.

Next, an expression for $E(\psi(A_b))$ is established. For large values of b, no holes will be present in A_b. If $|X_{max}(t,\sigma)| < b|$, then the excursion set is empty and $\psi(A_b) = 0$. If $|X_{max}(t,\sigma)| > b$, for sufficiently large b, only one region will be present in A_b and $\psi(A_b) = 1$. Therefore, for sufficiently large cutoff values,

$$P(|X_{max}(t,\sigma)| \geq b) \approx E\big(\psi(A_b)\big) = \alpha \tag{2.2}$$

For example, it can be claimed at the 95% confidence level that the signal is present if $|X_{max}(t,\sigma)| > b$, where b has been estimated, such that $E\big(\psi(A_b)\big) = 0.05$. Then for the purposes of our implementation, let max $|\hat{f}'(x,h)|$ be the extreme value of all derivatives from bins in a cluster of predefined dimensions. If $\max\left|\hat{f}'(x,h)\right| < b$, then a signal is not present and the entire set of bins, or boxes, are colored purple. Otherwise, a positive extreme derivative represents increasing signal and the bins are colored blue and a negative derivative represents a decreasing signal and the bins are colored red, as in the originally developed SiZer. As proposed by Siegmund and Worsley (1995), for a given cluster, $\psi(A_b) = \psi_V + \psi_E + \psi_B$, where ψ_V is the contribution from the interior of the cluster, ψ_E is the contribution from the edges, i.e. the lowest and highest t, or x values, and ψ_B is the contribution from the base, on the edge of the low side of the h value. Considering these three factors that determine the Hadwiger characteristic, in the case of univariate data a formula for its expectation, equal to α has been derived (Siegmund & Worsley, 1995):

$$E\big(\psi(A_b)\big) = \frac{|C|\big(\sigma_1^{-1}-\sigma_2^{-1}\big)(\lambda\kappa)^{\frac{1}{2}}b\phi(b)}{2\pi} + \frac{\left(\frac{|C|}{2}\right)\big(\sigma_1^{-1}+\sigma_2^{-1}\big)\lambda^{\frac{1}{2}}\phi(b)}{(2\pi)^{\frac{1}{2}}} + \psi(C)\frac{\log\left(\frac{\sigma_2}{\sigma_1}\right)\kappa^{\frac{1}{2}}\phi(b)}{(2\pi)^{\frac{1}{2}}} +$$

$$\psi(C)\big(1-\Phi(b)\big), \qquad (2.3)$$

where $|C|$ is the cluster width, or $x_{max} - x_{min}$. σ_1 and σ_2 are h_{min} and h_{max} respectively, the limits of the cluster bandwidth, where σ_1 and σ_2 are again on the logarithmic scale. $\kappa = \frac{N}{2}$, where N is the data dimension and $\lambda = \frac{1}{2} I_{N\times N}$. As N=1 for univariate data, both κ and λ are equal to $\frac{1}{2}$. $\psi(C) = 1$, since the cluster is rectangular with no holes and as b is large, a single region above b is expected. $\phi(b)$ is the standard normal pdf, evaluated at b. Then the formula in (2.3) simplifies to

$$E\big(\psi(A_b)\big) = \frac{|C|\big(\sigma_1^{-1}-\sigma_2^{-1}\big)b\phi(b)}{4\pi} + \frac{|C|\big(\sigma_1^{-1}+\sigma_2^{-1}\big)\phi(b)}{4\sqrt{\pi}} + \frac{\log\left(\frac{\sigma_2}{\sigma_1}\right)\phi(b)}{2\sqrt{\pi}} + \big(1-\Phi(b)\big). \qquad (2.4)$$

To find $b(h)$, the right side of (2.4) is evaluated and b is adjusted until $E\big(\psi(A_b)\big)$ reaches a value, sufficiently close to 0.05, within 0.001 of it. The value of b starts at a random number and if $E\big(\psi(A_b)\big)$ is high, b is decreased. If the expectation is low, b is increased. Every time

b is changed half-way in the direction of the closest previously attempted value, which guarantees the completion of the estimation. Finally, the elected extreme values of the derivatives are compared with the array of $b(h)$ and the color map is constructed. The half-normal and full-normal data sets are smoothed and presented in figures 2.3 and 2.4.

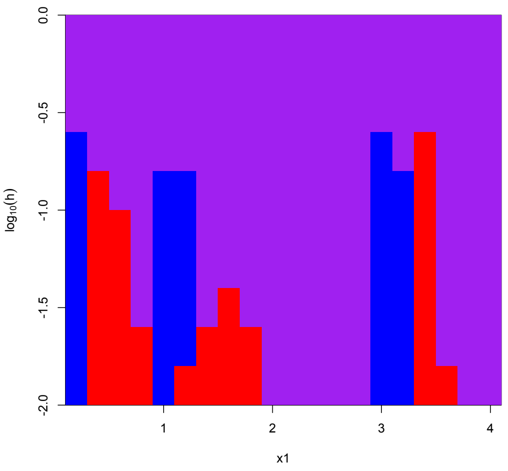


Figure 2.3 – Half-Normal data set, maximum derivative

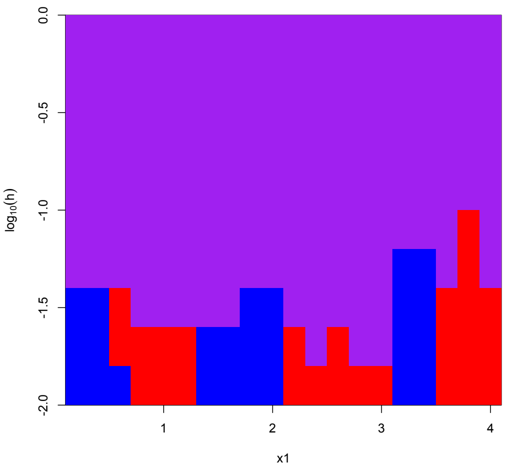


Figure 2.4– Full-Normal data set, maximum derivative

Both the half-normal and full-normal data features are detected correctly. For the half-normal color map the blue regions are located around $x = 0$, $x = 1$ and $x = 3$. They are followed by red regions. For the full-normal color map, the blue regions occur before $x = 0.5$, before $x = 2$ and before $x = 3.5$. They are also followed by red regions. The intervals of increasing density for the half-normal data are expectedly narrower, as the signal rises more rapidly than falls. This methodology of testing a signal via the extreme value of the derivative results in signal detection at lower bandwidths, compared to the original SiZer approach. At high bandwidths oversmoothing prevents signal detection, due to loss of sensitivity. Comparing the half-normal and the full-normal data sets, it can also be observed that the half-normal signal can be detected up to larger bandwidths. Consistent with the idea proposed by Chaudhuri and Marron (2000), as bandwidth increases, the features disappear and new features are not generated. These findings are also supported by the purple upper parts of the color maps.

Section 2.2 - Testing for Signal with a non-Stationary Gaussian Random Field

Formulae for expected Euler characteristics for general random fields have been developed in Hasofer (1978) and Adler and Taylor (2007, Chapter 11.1). Here the random field is not assumed to be stationary Gaussian. A Gaussian field is uniquely determined by its mean and variance. Recall that $\hat{f}'(x,h)$ was given in (1.2), and $\hat{f}''(x,h)$ is given by

$$\hat{f}''(x,h) = \frac{1}{n}\sum_{i=1}^{n}\frac{1}{h^3}K''\left(\frac{x-X_i}{h}\right), \tag{2.5}$$

where $K''(x) = \frac{1}{\sqrt{2\pi}}e^{-\frac{x^2}{2}}(x^2-1)$. By the Central Limit Theorem,

$\frac{1}{\sqrt{n}}\sum_{i=1}^{n}\frac{1}{h^2}K'\left(\frac{x-X_i}{h}\right) \to N\left(0,\sigma_1^2(x,h)\right)$, under our null hypothesis, and

$\frac{1}{\sqrt{n}}\sum_{i=1}^{n}\frac{1}{h^3}K''\left(\frac{x-X_i}{h}\right) \to N\left(\mu(x,h),\sigma_2^2(x,h)\right)$

Here $\hat{f}'(x,h)$ is assumed to have zero mean, according to the null hypothesis, and the mean of the second derivative is estimated from the sample:

$$\mu(x,h) = \mu\left(\hat{f}''(x,h)\right) = \frac{1}{nh^3}\sum_{i=1}^{n}\left(K''\left(\frac{x-X_i}{h}\right)\right)$$

The variances and covariance of the two normal distributions are also estimated from the samples.

$$\sigma_1(x,h) = \sqrt{var\left(\hat{f}'(x,h)\right)}$$

$$\sigma_2(x,h) = \sqrt{var\left(\hat{f}''(x,h)\right)}$$

The covariance is estimated as

$$\sigma_{12}(x,h) = \frac{1}{nh^5} cov\left(K'\left(\frac{x-X_i}{h}\right), K''\left(\frac{x-X_i}{h}\right)\right)$$

The correlation is

$$\rho(x,h) = \frac{\sigma_{12}(x,h)}{\sqrt{\sigma_1(x,h)\sigma_2(x,h)}}$$

Here too, we consider $P(max_{x,h}\hat{f}'(x,h) > u)$ in order to test H_0: $f'(x,h) = 0$. However, we no longer use the result of Siegmund and Worsley (1995), which is based on stationarity.

To determine the values of $u(h)$ over a range of bandwidths, a few steps are necessary. Adler and Taylor (2007) (p. 263) define the number of upcrossings of level u by a function f in the interval $(0, T)$ by

$$N_u^+(0,T) = \#\{t \in (0,T): f(t) = u, f'(t) > 0\} \tag{2.6}$$

That is, for $\varepsilon \to 0^+$, $f(t) = u$ and $f(t-\varepsilon) < u$. The number of upcrossings is a convenient way of counting the elements of an excursion set A_u, similar to the one developed by Siegmund and Worsley (1995). Furthermore, as shown by Adler and Taylor (2007),

$$E\left(N_u^+(0,T)\right) = \int_0^T \int_0^\infty y\, p_t(u,y)\, dy\, dt, \tag{2.7}$$

where $p_t(\cdot)$ is the joint pdf of $f(t)$ and $f'(t)$ over $(0, T)$. Since our null hypothesis tests f' for significance, in our implementation and simulation $p_t(\cdot)$ represents the joint density of $\hat{f}'$ and $\hat{f}''$. From section 2.1, the expected Euler characteristic, which reduces to the expected number of upcrossings from (2.7) in one dimension, approximates the tail probability of the maximum value of the derivative. If the boundary point 0 is included, as the base value for the beginning of each cluster, the expected Euler characteristic becomes

$$E\big(\rho(A_u)\big) = E\big(N_u^+(0,T)\big) + P(f(0) \geq u)$$

To estimate the cutoff value $u(h)$ for every bandwidth, the entire interval is separated into bins, as outlined in chapter 1 and let T_1 and T_2 be the lower and upper boundaries of a bin. Then $\Delta T = T_2 - T_1$. Modifying (2.7) for our simulation,

$$E\big(N_u^+(T_1,T_2)\big) = \int_{T_1}^{T_2} \int_0^{\infty} y\, p_t(u, y-\mu)\, dy\, dt, \tag{2.8}$$

where $p_t(\cdot)$ is the joint pdf of $f'(x,h)$ and $f''(x,h)$. For bounds T_1 and T_2 in the close neighborhood of a given x, the expected number of upcrossings from T_1 and T_2 is derived in the following theorem:

Theorem 1:

$$E\big(N_u^+(T_1,T_2)\big) =$$

$$|\Delta T| \left[\frac{\sigma_2\sqrt{1-\rho^2}}{2\pi\sigma_1} \exp\left[-\left(\frac{u^2}{2\sigma_1^2} + \frac{\left(\mu+\rho\frac{\sigma_2}{\sigma_1}u\right)^2}{2\sigma_2^2(1-\rho^2)} \right) \right] + \frac{\rho\frac{\sigma_2}{\sigma_1}u+\mu}{\sqrt{2\pi}\sigma_1} \exp\left(-\frac{u^2}{2\sigma_1^2} \right) \left[\Phi\left(\frac{\mu+\rho\frac{\sigma_2}{\sigma_1}u}{\sigma_2\sqrt{1-\rho^2}} \right) \right] \right] \tag{2.9}$$

Proof:

Let $f(\cdot)$ be the joint pdf of $f'(x,h)$ and $f''(x,h)$.

Let $f'(x,h) \sim N(0,\sigma_1)$, $f''(x,h) \sim N(\mu,\sigma_2)$ and $\rho(x,h)$ be the correlation between them. An expression for $E\big(N_u^+(T_1,T_2)\big)$ is derived, starting from (2.8). After change of variable, replacing $y-\mu$ with v, we obtain

$$E\big(N_u^+(T_1,T_2)\big) = \int_{T_1}^{T_2} \int_{-\mu}^{\infty} (v+\mu) f(u,v)\, dv\, dt$$

$$= \int_{T_1}^{T_2} \int_{-\mu}^{\infty} (v+\mu) \frac{1}{2\pi\sigma_1\sigma_2\sqrt{1-\rho^2}} \exp\left[-\frac{1}{2(1-\rho^2)} \left(\frac{u^2}{\sigma_1^2} + \frac{v^2}{\sigma_2^2} - 2\rho\frac{uv}{\sigma_1\sigma_2} \right) \right] dv\, dt$$

$$= \int_{T_1}^{T_2}\int_{-\mu}^{\infty}(v+\mu)\,\frac{1}{2\pi\sigma_1\sigma_2\sqrt{1-\rho^2}}\exp\left[-\frac{1}{2\sigma_2^2(1-\rho^2)}\left(v^2-2\mathrm{v}\rho\frac{u\sigma_2}{\sigma_1}+\frac{u^2\sigma_2^2}{\sigma_1^2}\right)\right]dv\,dt$$

$$= \int_{T_1}^{T_2}\int_{-\mu}^{\infty}(v+\mu)\,\frac{1}{2\pi\sigma_1\sigma_2\sqrt{1-\rho^2}}\exp\left[-\frac{1}{2\sigma_2^2(1-\rho^2)}\left(v^2-2\mathrm{v}\rho\frac{u\sigma_2}{\sigma_1}+\rho^2\frac{\sigma_2^2}{\sigma_1^2}u^2-\rho^2\frac{\sigma_2^2}{\sigma_1^2}u^2+\right.\right.$$

$$\left.\left.\frac{u^2\sigma_2^2}{\sigma_1^2}\right)\right]dv\,dt$$

$$= \int_{T_1}^{T_2}\frac{e^{-\frac{u^2}{2\sigma_1^2}}}{2\pi\sigma_1\sigma_2\sqrt{1-\rho^2}}\int_{-\mu}^{\infty}(v+\mu)\exp\left[-\frac{1}{2\sigma_2^2(1-\rho^2)}\left(v-\rho\frac{\sigma_2}{\sigma_1}u\right)^2\right]dv\,dt$$

Replacing $y = v-\rho\frac{\sigma_2}{\sigma_1}u$, then $v = y+\rho\frac{\sigma_2}{\sigma_1}u$ and $dy = dv$, we get

$$= \int_{T_1}^{T_2}\frac{e^{-\frac{u^2}{2\sigma_1^2}}}{2\pi\sigma_1\sigma_2\sqrt{1-\rho^2}}\int_{-\mu-\rho\frac{\sigma_2}{\sigma_1}u}^{\infty}\left(y+\rho\frac{\sigma_2}{\sigma_1}u+\mu\right)e^{-\frac{y^2}{2\sigma_2^2(1-\rho^2)}}\,dy\,dt$$

$$= \int_{T_1}^{T_2}\frac{e^{-\frac{u^2}{2\sigma_1^2}}}{2\pi\sigma_1\sigma_2\sqrt{1-\rho^2}}\int_{-\mu-\rho\frac{\sigma_2}{\sigma_1}u}^{\infty}y\,e^{-\frac{y^2}{2\sigma_2^2(1-\rho^2)}}+\left(\rho\frac{\sigma_2}{\sigma_1}u+\mu\right)e^{-\frac{y^2}{2\sigma_2^2(1-\rho^2)}}dy\,dt$$

$$= \int_{T_1}^{T_2}\frac{e^{-\frac{u^2}{2\sigma_1^2}}}{2\pi\sigma_1\sigma_2\sqrt{1-\rho^2}}\int_{-\mu-\rho\frac{\sigma_2}{\sigma_1}u}^{\infty}(A+B)\,dy\,dt \tag{2.9.1}$$

From equation (2.9.1), we have $\int_{-\mu-\rho\frac{\sigma_2}{\sigma_1}u}^{\infty}A\,dy = \int_{-\mu-\rho\frac{\sigma_2}{\sigma_1}u}^{\infty}y\,e^{-\frac{y^2}{2\sigma_2^2(1-\rho^2)}}\,dy$

Letting $w = -\frac{y^2}{2\sigma_2^2(1-\rho^2)}$, $dw = -\frac{y}{\sigma_2^2(1-\rho^2)}dy$, $y\,dy = -\sigma_2^2(1-\rho^2)$, the above integral is

$$\int_{-\frac{\left(\mu+\rho\frac{\sigma_2}{\sigma_1}u\right)^2}{2\sigma_2^2(1-\rho^2)}}^{-\infty}-\sigma_2^2(1-\rho^2)e^{w}\,dw$$

$$= \sigma_2^2(1-\rho^2)\exp\left[-\frac{\left(\mu+\rho\frac{\sigma_2}{\sigma_1}u\right)^2}{2\sigma_2^2(1-\rho^2)}\right] \tag{2.9.2}$$

From equation (2.9.1), we have $\int_{-\mu-\rho\frac{\sigma_2}{\sigma_1}u}^{\infty}B\,dy = \int_{-\mu-\rho\frac{\sigma_2}{\sigma_1}u}^{\infty}\left(\rho\frac{\sigma_2}{\sigma_1}u+\mu\right)e^{-\frac{y^2}{2\sigma_2^2(1-\rho^2)}}\,dy$

$$= \left(\rho\frac{\sigma_2}{\sigma_1}u+\mu\right)\sqrt{2\pi(1-\rho^2)}\sigma_2\int_{-\mu-\rho\frac{\sigma_2}{\sigma_1}u}^{\infty}\frac{e^{-\frac{y^2}{2\sigma_2^2(1-\rho^2)}}}{\sqrt{2\pi(1-\rho^2)}\sigma_2}\,dy$$

$$= \left(\rho\frac{\sigma_2}{\sigma_1}u + \mu\right)\sqrt{2\pi(1-\rho^2)}\sigma_2\left[1 - \Phi\left(-\frac{\mu+\rho\frac{\sigma_2}{\sigma_1}u}{\sigma_2\sqrt{1-\rho^2}}\right)\right]$$

$$= \left(\rho\frac{\sigma_2}{\sigma_1}u + \mu\right)\sqrt{2\pi(1-\rho^2)}\sigma_2\left[\Phi\left(\frac{\mu+\rho\frac{\sigma_2}{\sigma_1}u}{\sigma_2\sqrt{1-\rho^2}}\right)\right] \qquad (2.9.3)$$

Combining the results of (2.9.2) and (2.9.3) into (2.9.1), we obtain

$$E\left(N_u^+(T_1,T_2)\right) =$$

$$\int_{T_1}^{T_2}\frac{e^{-\frac{u^2}{2\sigma_1^2}}}{2\pi\sigma_1\sigma_2\sqrt{1-\rho^2}}\left[\sigma_2^2(1-\rho^2)\exp\left[-\frac{\left(\mu+\rho\frac{\sigma_2}{\sigma_1}u\right)^2}{2\sigma_2^2(1-\rho^2)}\right] + \right.$$

$$\left.\left(\rho\frac{\sigma_2}{\sigma_1}u + \mu\right)\sqrt{2\pi(1-\rho^2)}\sigma_2\left[\Phi\left(\frac{\mu+\rho\frac{\sigma_2}{\sigma_1}u}{\sigma_2\sqrt{1-\rho^2}}\right)\right]\right]dt$$

$$= |T|\left[\frac{\sigma_2\sqrt{1-\rho^2}}{2\pi\sigma_1}\exp\left[-\left(\frac{u^2}{2\sigma_1^2} + \frac{\left(\mu+\rho\frac{\sigma_2}{\sigma_1}u\right)^2}{2\sigma_2^2(1-\rho^2)}\right)\right] + \frac{\rho\frac{\sigma_2}{\sigma_1}u+\mu}{\sqrt{2\pi}\sigma_1}\exp\left(-\frac{u^2}{2\sigma_1^2}\right)\left[\Phi\left(\frac{\mu+\rho\frac{\sigma_2}{\sigma_1}u}{\sigma_2\sqrt{1-\rho^2}}\right)\right]\right].$$

This proves *Theorem 1* and equation (2.9).

Finally, the expected Euler characteristic over the entire scale-space for a fixed bandwidth h is can be approximated by

$$E\left(\rho\left(A_u\left(f'(0,T)\right)\right)\right) \approx \Psi\left(\frac{u}{\sigma_1(x=0,h)}\right) + \sum_{j=1}^{g}\left[|\Delta T|\left[\frac{\sigma_2\sqrt{1-\rho^2}}{2\pi\sigma_1}\exp\left[-\left(\frac{u^2}{2\sigma_1^2} + \frac{\left(\mu+\rho\frac{\sigma_2}{\sigma_1}u\right)^2}{2\sigma_2^2(1-\rho^2)}\right)\right] + \right.\right.$$

$$\left.\left.\frac{\rho\frac{\sigma_2}{\sigma_1}u+\mu}{\sqrt{2\pi}\sigma_1}\exp\left(-\frac{u^2}{2\sigma_1^2}\right)\left[\Phi\left(\frac{\mu+\rho\frac{\sigma_2}{\sigma_1}u}{\sigma_2\sqrt{1-\rho^2}}\right)\right]\right]\right], \qquad (2.10)$$

where $\Delta T\ = \frac{T}{g}$, g is the number of data points x, where f' and f'' are evaluated and σ_1, σ_2 and ρ are evaluated g times, or at every location x in the summation.

For the simulation, the expected Euler characteristic is set to 0.05 and the cutoff values $u(h)$ are estimated iteratively, similar to the b values in section 2.1. The color maps of the half-normal and full-normal data sets are presented on figures 2.5 and 2.6.

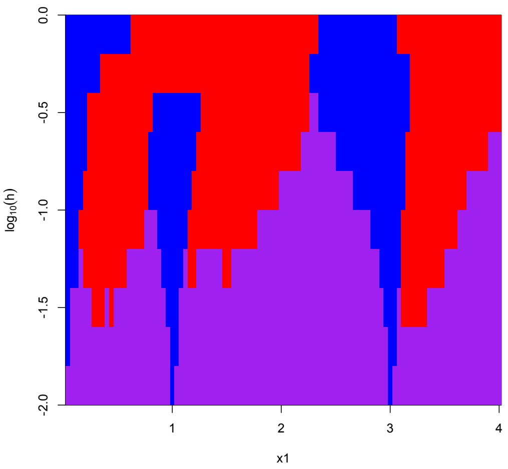


Figure 2.5 – Half-Normal data set, non-Stationary Gaussian Approach

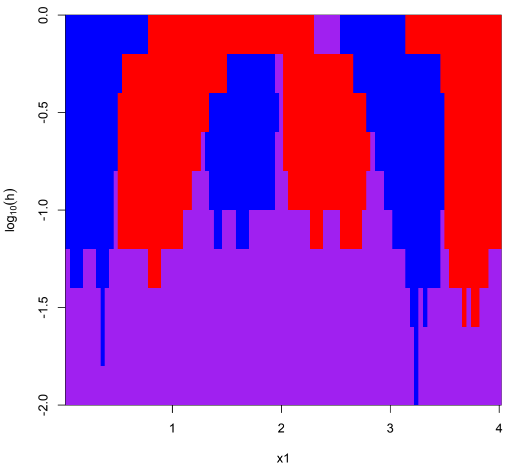


Figure 2.6 – Full-Normal data set, non-Stationary Gaussian Approach

The density curve features are correctly estimated around middle-range bandwidths (0.1 to 0.4). Compared with the cutoff values $b(h)$ in section 2.1, the cutoff values $u(h)$ are estimated to be significantly larger at low bandwidths and significantly smaller at high bandwidths, resulting in the different color maps, when comparing figures 2.3 and 2.5 for the half-normal data set, and figures 2.4 and 2.6 for the full-normal data set. This results in a purple color map area and undetected signal for low bandwidths, accurate mid-range estimation and oversmoothed signal at upper bandwidths, where the second increase of the signal is masked by a predominant decreasing red region on the color map.

Chapter 3 - Two-Sample Univariate Density Comparison

As an extension to the univariate density estimation, the upcrossings method for estimating non-Gaussian noise developed by Adler and Taylor (2007) can be extended to compare two univariate data samples. Park and Kang (2008) have extended the original Significant Zero Crossings for Derivatives method into ‘Significant Zero Crossings for Differences’ in order to compare two or more regression curves. In this work we adapt the regression comparison with the proposal by Adler and Taylor (2007) to compare the probability densities of two univariate curves f_1 and f_2. If $f_1(t)$ is the PDF estimated from the sample $X_{11} \dots X_{1n_1}$ and $f_2(t)$ is the PDF estimated from the sample $X_{21} \dots X_{2n_2}$, let

$$H_0\colon f_1(t) = f_2(t)$$

$$H_1\colon f_1(t) \neq f_2(t)$$

Again with a slight abuse of notation, H_0 is equivalent to $f_1(x,h) - f_2(x,h) = 0$ for every bin (x,h). To determine whether $(f_1 - f_2)(x,h)$ is significantly different from zero, compare it with the cutoff value $u(h)$, as in chapter 2, obtained iteratively with (2.10), where $f'(x,h)$ is replaced with $(f_1 - f_2)(x,h)$ *and* $f''(x,h)$ is replaced with $(f_1' - f_2')(x,h)$. In (2.7) $p_t(\cdot)$ is the joint density of $(f_1 - f_2)(x,h)$ and $(f_1' - f_2')(x,h)$. Therefore, the same formula can be used as in (2.10), but with modified $\mu(x,h)$, $\sigma_1(x,h)$, $\sigma_2(x,h)$ and $\rho(x,h)$. Here both $(f_1 - f_2)(x,h)$ and $(f_1' - f_2')(x,h)$ have means of zero, therefore, $\mu(x,h) = 0$ for all (x,h). Estimating the variances and covariance from the samples is done as follows.

$$\left(\widehat{f_1} - \widehat{f_2}\right)(x,h) = \frac{1}{n_1 h}\sum_{i=1}^{n_1} K\left(\frac{x - X_{1i}}{h}\right) - \frac{1}{n_2 h}\sum_{i=1}^{n_2} K\left(\frac{x - X_{2i}}{h}\right)$$

$$\left(\widehat{f_1'} - \widehat{f_2'}\right)(x,h) = \frac{1}{n_1 h^2}\sum_{i=1}^{n_1} K'\left(\frac{x - X_{1i}}{h}\right) - \frac{1}{n_2 h^2}\sum_{i=1}^{n_2} K'\left(\frac{x - X_{2i}}{h}\right)$$

$$\sigma_1^2 = var\left(\widehat{f_1} - \widehat{f_2}\right) = \frac{var\left(K\left(\frac{x-X_{1i}}{h}\right)\right)}{n_1 h^2} + \frac{var\left(K\left(\frac{x-X_{2i}}{h}\right)\right)}{n_2 h^2}$$

$$\sigma_2^2 = var\left(\widehat{f_1'} - \widehat{f_2'}\right) = \frac{var\left(K'\left(\frac{x-X_{1i}}{h}\right)\right)}{n_1 h^4} + \frac{var\left(K'\left(\frac{x-X_{2i}}{h}\right)\right)}{n_2 h^4}$$

$$\sigma_{12} = cov\left(\widehat{f_1} - \widehat{f_2}, \widehat{f_1'} - \widehat{f_2'}\right) = \frac{cov\left(K\left(\frac{x-X_{1i}}{h}\right), K'\left(\frac{x-X_{1i}}{h}\right)\right)}{n_1 h^3} + \frac{cov\left(K\left(\frac{x-X_{2i}}{h}\right), K'\left(\frac{x-X_{2i}}{h}\right)\right)}{n_2 h^3}$$

The final point before the simulation and analysis is that purple segments of the color map demonstrate insignificant differences between the two estimated densities. Blue regions signify $f_1(x, h) > f_2(x, h)$ and red regions show that $f_1(x, h) < f_2(x, h)$. It is important to note that rapidly alternating red-blue boxes also signify insignificant differences, as the pattern cannot be established in a stable manner. The entire range of the bandwidths also needs to be inspected when considering signal presence at a given location.

Two simulations are performed to assess the proposed density comparison methodology. Firstly, a sample generated from the half-normal density, or f_1, is compared with a sample generated from the half-normal curve, shifted by 0.3 units to the right, f_2. Secondly, two identical samples, generated from the same half-normal curve f_1 are compared. The pdfs are

$$f_1(t) \;=\; 0.4 \cdot |N(0{,}0.3)| + 0.3 \cdot (|N(0{,}0.4)| + 1) + 0.3 \cdot (|N(0{,}0.2)| + 3)$$

$$f_2(t) \;=\; 0.4 \cdot (|N(0{,}0.3)| + 0.3) + 0.3 \cdot (|N(0{,}0.4)| + 1.3) + 0.3 \cdot (|N(0{,}0.2)| + 3.3)$$

Each group is represented by a sample of size 5000. The differences $(\widehat{f_1} - \widehat{f_2})(x, h)$ are shown on figures 3.1 and 3.2. The resulting color maps are displayed on figures 3.3 and 3.4.

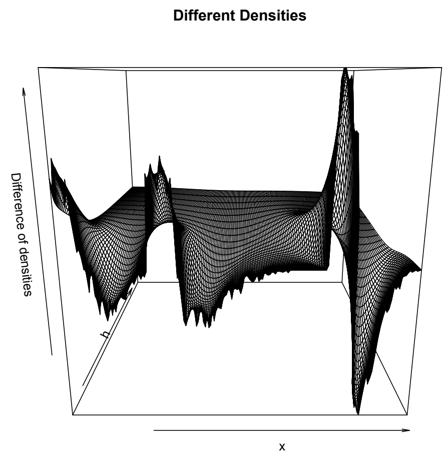


Figure 3.1 – Difference of Different Densities

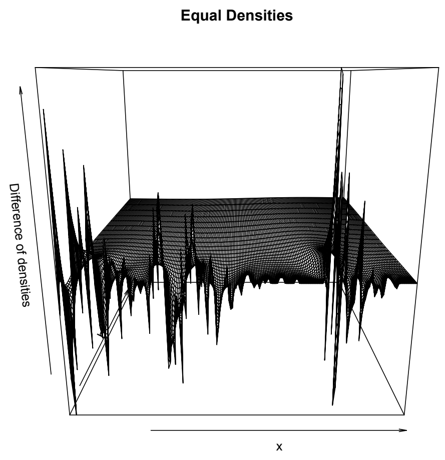


Figure 3.2 – Difference of Equal Densities

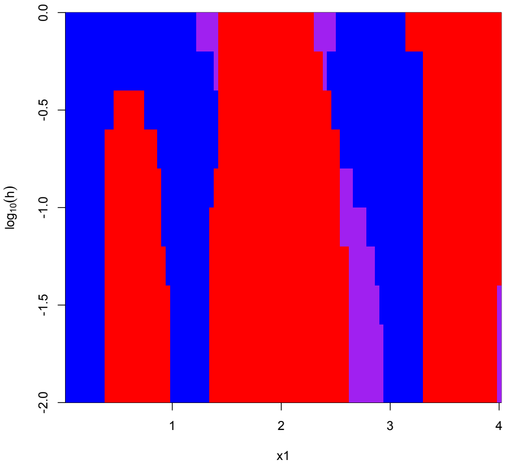


Figure 3.3 – Comparing Different Densities

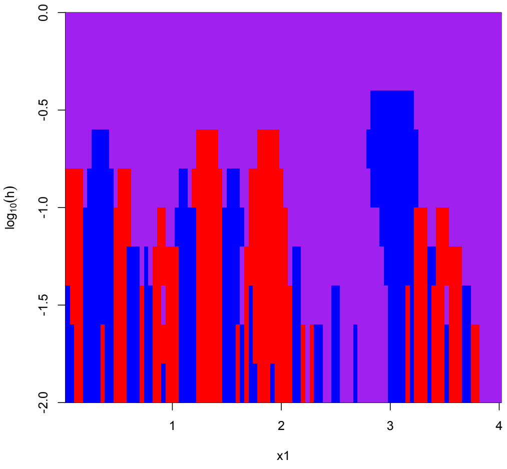


Figure 3.4 – Comparing Equal Densities

The color maps correctly identify both the differences in the densities and identical densities, as well. When comparing f_1 with f_2 on figure 3.3, distinct blue and red regions are present and are mostly pronounces in low to mid-range bandwidth. Sincef_2 lags f_1 by 0.3, the blue bands in the interval [0,0.3], [1,1.3] and [3,3.3] outline the superiority of the f_1 value in these regions. Immediately after 0.3, 1.3 and 3.3 the red regions prevail over most bandwidths, showing the superiority of f_2 over f_1. Toward the high end of the bandwidth, the first red region is not sufficiently pronounced due to oversmoothing. The 3D perspective of the difference of densities on figure 3.1 clearly outlines the regions of dominance, precisely as demonstrated with the color map.

In the case of comparison of theoretically equal densities, no large blocks of dominant values are present, as shown on figure 3.4. While in the upper part of the bandwidth range the color map is expectedly purple, over the low range of h, rapidly alternating red and blue bands appear over short intervals. These spurious outputs can occur due to undersmoothing, and mostly due to the low cutoff values, as calculated for $u(h)$. However, when observing a location x over the entire bandwidth range, it becomes clear that the red and blue bands are accidental noise. The 3D perspective of the difference of densities on figure 3.2 shows wiggly behavior at low bandwidths and very oversmoothed behavior at high bandwidths. As outlined with the color map, it can not be concluded that one density is significantly different from the other.

Chapter 4 - Conclusions and Future Work

Four different approaches to feature detection in probability density estimation and one for density comparison were presented. First, the original SiZer approach, second, SiZer with an advanced quantile estimation using a cluster index, third, testing for a signal with a stationary Gaussian process and fourth, testing for a signal with a non-stationary Gaussian process. The last method was also adapted to compare two signals. All presented results accurately describe all characteristics of the simulated data sets. The presented methods appear to more easily detect sharp peaks than smooth ones. While it is not possible to objectively determine which of the four approaches is superior, a far more important conclusion is reached. Using all known approaches together to analyze a data set provides a more powerful tool for data analysis. When some methodologies complement the weakness of others, viewing the data from different angles provides a more complete picture of the correct features of a given data set.

A future challenge is to improve feature detection methods to distinguish more subtle differences between narrower signals. One possibility for future work is to improve the Gaussian approximation of the number of upcrossings via an Edgeworth series expansion. Another direction of future research is to develop feature detection techniques for bivariate and multivariate curve estimation. Godtliebsen et al. (2002) did a similar study and the color map is three-dimensional and looks more complex. In smoothing bivariate data the upcrossings methodology will not apply and gradients need to be used for data analysis. Hasofer (1978) has derived the following expected Euler characteristic, which can be reduced for two-dimensional data $\boldsymbol{t} = (t_1, t_2)$ to

$$E[\varphi_u(S)] = -\lambda(S) \int_{-\infty}^{+\infty} \int_0^{+\infty} x_2 |D| \phi(u; 0, x_2, z) dx_2 \, dz \, , \tag{4.1}$$

where $\phi(x; y_1, y_2, z)$ is the joint pdf of $X(\boldsymbol{t})$, $\partial X(\boldsymbol{t})/\partial t_1$, $\partial X(\boldsymbol{t})/\partial t_2$ and $\partial^2 X(\boldsymbol{t})/\partial t_1^2$, $|D| = \partial^2 X(\boldsymbol{t})/\partial t_1^2$ and $\lambda(S)$ is the Lebesgue measure of the region S. Identically to the one-dimensional case, the Euler characteristic is calculated by using contributing points by either +1 or -1, as follows. For a point to be a contributor, $X(\boldsymbol{t}) = u$, $\partial X(\boldsymbol{t})/\partial t_1 = 0$ and $\partial X(\boldsymbol{t})/\partial t_2 > 0$. Then $\partial^2 X(\boldsymbol{t})/\partial t_1^2 < 0$ signifies a local maximum and contributes to the characteristic with +1 and $\partial^2 X(\boldsymbol{t})/\partial t_1^2 > 0$ signifies a local minimum along t_1 and contributes with -1 for the Euler characteristic. Once the bivariate case is developed, the potential for a robust application in 2D and 3D imagery is expected to be significant.

References

Adler, R. J., & Taylor, J. E. (2007). *Random Fields and Geometry.* New York, NY: Springer Monographs in Mathematics.

Chamandy, N., Worsley, K. J., Taylor, J., & Gosselin, F. (2008). Tilted Euler Characteristic Densities for Central Limit Random Fields, With Application to "Bubbles". *The Annals of Statistics, 36,* 2471-2507.

Chaudhuri, P., & Marron, J. S. (1999). SiZer for Exploration of Structures in Curves. *Journal of the American Statistical Association, 94,* 807-823.

Chaudhuri, P., & Marron, J. S. (2000). Scale Space View of Curve Estimation. *The Annals of Statistics, 28,* 408-428.

Godtliebsen, F., Marron, J. A., & Chaudhuri, P. (2002). Significance in Scale Space for Bivariate Density Estimation. *Journal of Computational and Graphical Statistics, 11,* 1-21.

Hannig, J., & Lee, T. (2006). Robust SiZer for Exploration of Regression Structures and Outlier Detection. *Journal of Computational and Graphical Statistics, 15,* 101-117.

Hannig, J., & Marron, J. S. (2006). Advanced Distribution Theory of SiZer. *Journal of the American Statistical Association, 101,* 484-499.

Hasofer, A. M. (1978). Upcrossings of Random Fields. *Advances in Applied Probability, 10,* 14-21.

Park, C., & Kang, K. (2008). SiZer Analysis for the Comparison of Regression Curves. *Computational Statistics & Data Analysis, 52,* 3954-3970.

Siegmund, D. O., & Worsley, K. J. (1995). Testing for a Signal with Unknown Location and Scale in a Stationary Gaussian Random Field. *The Annals of Statistics, 23,* 608-639.

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

Worsley, K. J. (1995). Estimating the Number of Peaks in a Random Field Using the Hadwiger Characteristic of Excursion Sets, With Application to Medical Images. *The Annals of Statistics, 23,* 640-669.