UNIVERSITY OF TROMSØ UIT

FACULTY OF SCIENCE AND TECHNOLOGY
DEPARTMENT OF MATHEMATICS AND STATISTICS

Scale-Space Methodology Applied to Spectral Feature Detection, Multinormality Testing and the *k*-sample Problem, and Wavelet Variance Analysis

Kristian Hindberg

A dissertation for the degree of Philosophiae Doctor
February 2012

Abstract

The number of scale-space statistical algorithms has been greatly increased over the last 15 years. The concept originated from computer vision, introduced in Lindeberg (1994). The seminal paper by Chaudhuri and Marron (1999) brought the scale-space concept into smoothing of curves and kernel density estimation through the SiZer tool. By using all relevant smoothing bandwidths, i.e., the "scale" part, SiZer allows the user to look for interesting features in the smoothed curves or density estimates simultaneously on all bandwidths. In the years following, a number of classical statistical problems were also included in the family of scale-space algorithms.

In this thesis, new scale-space algorithms for four such classical statistical problems are suggested. Paper II presents two closely related problems, addressed with highly similar approaches.

Paper I addresses spectral scale-space analysis. Peaks found in the estimated spectral density function of evenly sampled stationary signals are typically of great interest for scientists. A peak found at a given frequency translates to potential (hidden) periodicities in a data set. Therefore, algorithms to determine which spectral peaks that really are significant are important in real-world applications. The presented algorithm uses the infamous periodogram, for reasons explained later. The different Fourier frequencies are the "space" part of the algorithm, while the "scale" part is introduced through a smoothing parameter of an assumed prior distribution. By using the Integrated Nested Laplace Approximation (Rue et al., 2009), a full posterior distribution can be constructed, from which the needed p-values are found.

Unlike Papers I and III, Paper II presents a scale-space approach without introducing a prior distribution. Through two similar algorithms, two different questions are addressed: 1) Can a multivariate data set be considered to originate from some unspecified multivariate Gaussian distribution? 2) Can k multivariate data sets be considered to originate from some unspecified multivariate distribution? The "scale" part of both algorithms is connected to a weighted summation across neighboring dimensions. The number of dimensions that are summed across is given by the scale parameter. The "space" parameter is connected to the time or location index of the data series. The algorithms do not need to invert estimated covariance matrices, thereby they can handle the High Dimension Low Sample Size case, where most comparable methods fail.

Paper III brings the scale-space concept into long-range dependence and wavelet analysis. The basis of this third paper is the wavelet coefficients resulting from linear filtering of the data with localized wavelet filters of increasing widths. The variance of these coefficients forms the "wavelet variance". The "space" part is connected to the different wavelet filters/scales. As in Paper I, the "scale" part is connected to the smoothing parameter of the prior distribution. The degree of long-range dependence is fully characterized by the Hurst parameter. This parameter can be estimated through linear regression of the natural logarithm of the wavelet variance. Determining for which scales this regression should be done is not trivial, an issue which the presented algorithm addresses. A time-divided/local wavelet analysis for detecting non-stationarities in the data is also presented in Paper III.

Acknowledgements

These acknowledgements follow a deeply rooted personal tradition of expressing oneself in too many words. Readers frightened by the length of it should be enticed by the promise of a somewhat untraditional second half.

First of all, I would like to thank my two supervisors: Professor Fred Godtliebsen, my main supervisor here at the University of Tromsø, and my secondary supervisor, Professor Håvard Rue at the Norwegian University of Science and Technology. Fred, I am very grateful for being given the chance to do a Ph.D. with you as my supervisor. Throughout these years you have involved me in many really interesting projects, something that has given me the chance to improve my statistical and general knowledge immensely. I am looking forward to continue to work with you — starting Monday in the post-doc position. Thank you Håvard for having me as a guest scientist in 2007, and for introducing me to GMRFs and helping out with my research in general.

I would like to thank Steve Marron for giving me the chance to spend the autumn of 2008 at the University of North Carolina at Chapel Hill. During my stay at UNC, the seeds of Paper II of my thesis were sown, with the help of Fred and Jan Hannig. Thank you Jan for having me over for your traditional Thanksgiving dinner. The following spring semester was spent at the University of Washington/APL in the amazing city of Seattle. Here I was the guest of Professor Donald Percival, who took great care of me, and always had time to help me with what was the start of Paper III. Thanks Don for taking me to Mt. Townsend. Going for such a nice hike in the Olympic Mountains topped off my stay in Washington State. Hopefully I can return the favor when you come to Tromsø as a guest scientist.

On another note, Kristian thinks he has the best friends in the world, something which, together with Kristian's great family, was one of the main reasons that Kristian wanted to move back home after some time away. Since these friends of Kristian tend to trick him into way too much fun and shenanigans, Kristian undoubtedly blames them for him being somewhat late in completing his thesis. Now that Kristian has more free time, he will be the one tricking his friends into doing all of the fun things that the best city in the world has to offer. A special thanks goes out to Kristian's friends who have willingly or as a result of psychological extortion made Kristian dinner over the years.

A loving and great upbringing by Kristian's mom and dad is probably the main reason why Kristian, who back in elementary school had so much "energy" that he is still talked about at the school, was eventually transformed into a sensible guy who has gotten by quite ok in life. His mother somewhat mysteriously says, "Kristian, you and your siblings are more intentionally shaped by your father and me than you ever will realize...". Kristian does not yet know what this really means, but he does feel that this shaping must have worked quite well.

As Kristian grew up, he had a big brother who did quite well at school. Kristian, being the annoying little brother, of course wanted to beat his older brother at school. Kristian also has a little sister, who was an annoying little sister. She of course also wanted to beat her older brothers at school, something she did, by far. This Kristian did not like very much, something that made him work harder as he was realizing that his little sister was thrashing his grade-averages. So, thank you both Kent and Heidi for giving me great inspiration to work hard both early and later in life. Thanks Heidi for reading through parts of my thesis, and answering all of my stupid questions regarding the thesis process.

Thank you all for helping a soon 33.5 year old Kristian achieve the goal that the then 17 year old Kristian set himself back at Tromsdalen videregående skole. His long-time goal of becoming a lawyer (strongly influence by "L.A. Law") was quickly sat aside when the energetic and inspirational mathematics and physics teacher, Franck Pettersen, opened the wonderful world of science to him. Unfortunately, Franck passed away way too early, but his unique person and love for science lives on within all his students who have carried his energy with them through their lives in science. Thanks Franck, this probably would have been a thesis on fishery law without you there to inspire me :-)

List of Publications

I. Sigrunn H. Sørbye, **Kristian Hindberg**, Lena R. Olsen and Håvard Rue, "Bayesian multiscale feature detection of log-spectral densities", *Computational Statistics and Data Analysis*, vol. 53, num. 11, pp. 3746–3754, September 2009.

II. **Kristian Hindberg**, Jan Hannig and Fred Godtliebsen , "A novel scale-space approach for multinormality testing and the k-sample problem", Submitted to *Computational Statistics and Data Analysis*, January 2012.

III. **Kristian Hindberg**, Donald B. Percival, Tor Arne Øigård, Stilian A. Stoev, Fred Godtliebsen and Murad S. Taqqu , "A scale-space wavelet visualization tool for exploring non-stationarities in long-range dependent time series", *Unpublished manuscript.*

These publications are referred to by their roman letters in the following chapters.

Contents

Chapter 1

Introduction

The motivation for this thesis was to expand the collection of statistical scale-space methods. The concept of scale-space, or multi-scale, is fairly new. It is commonly considered to have been introduced within the computer vision field by Lindeberg (1994). Lindeberg uses two-dimensional Gaussian smoothing kernels to generate smoothed versions of the original image, where the "scale" part is connected to the variances of the smoothing kernels. For instance, a large degree of smoothing hides the details of the image, and instead focuses on the coarser structures in the image. No specific level of smoothing is considered to be the correct level. Different features might be present/detectable on different smoothed versions of the original image.

The underlying idea of the scale-space methodology is that different significant features of a data set might be connected to different scales. In general, the scale is connected to zooming in or out, or looking at the fine details or the coarse trends of the data. Generally, the scales that a feature is connected to, which will also be the best scales to detect/observe it at, will not be known. Without this prior knowledge, the normal approach is to guess or in some way estimate which scale is best. For instance, in classical nonparametric density estimation or smoothing schemes, some sort of bandwidth has to be chosen (Wand and Jones, 1995). By only looking at one bandwidth, one cannot detect features that are not detectable when using only this bandwidth.

The SiZer Methodology

The pioneering scale-space work within statistics was done by Chaudhuri and Marron (1999), with further theoretical justifications and improvements in Chaudhuri and Marron (2000) and Hannig and Marron (2006); Hannig and Lee (2006). Here, the SiZer (Significant Zero-crossings of the derivative) procedure for nonparametric, one-dimensional function estimation was introduced. SiZer is a tool for detecting significant trends of the true underlying signal viewed at different smoothing levels, also called scales. In this SiZer-paper, the concept of SiZer maps was introduced (see Erästö and Holmström (2005) for more on SiZer maps). A SiZer map is a two-dimensional image, graphically communicating relevant information. The first axis assigns the position or time of the data vector, while the

degree of smoothing is given by the second axis. As SiZer represents the scale-space idea in a comprehensible way, and since the presented methods of this thesis have strong ties to SiZer, the SiZer methodology is presented and demonstrated in the following paragraphs.

Let $x_i, i = 1, \ldots, n$ be a sample from some smooth univariate distribution $f(x)$. A kernel estimate of $f(x)$ is found as

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

where the bandwidth/smoothing parameter $h > 0$ is connected to the width of the smoothing kernel $K_h(x) = K(x/h)/h$ (Wand and Jones, 1995). The basic smoothing kernel $K(x)$ has to meet the requirements of a density function, and often the standard normal distribution is used.

In a nonparametric regression setting, the model considered can be described by

$$y_i = \mu(x_i) + e_i,$$

where $\mu(x)$ is the true regression function, $x_i, i = 1, \ldots, n$ are the explanatory variables, and the e_is are the independent error terms. The regression function is estimated through local linear regression, producing the estimate

$$\hat{\mu}_h(x) = \text{argmin}_a \sum_{i=1}^{n} \left[y_i - (a + b(x_i - x))\right]^2 K_h(x - x_i), \tag{1.2}$$

where the minimization is done jointly over a and b, but only a is kept.

Both of these two problems are curve estimation problems. In the normal setting, where one tries to estimate the true underlying function ($f(x)$ or $\mu(x)$), such nonparametric smoothing methods will produce biased estimates. In the SiZer methodology, this bias problem is not present since the target of the curve estimation is the true curve viewed at varying levels of resolution/smoothing.

In SiZer, the inference is done on the derivatives of the curves. Thereby, it is the derivatives of the curves viewed at different resolutions that have to be estimated. In the density estimation setting, the basic smoothing kernel in Equation (1.1) is replaced by the derivate of the kernel $K'(x)$, to produce the estimated first derivatives

$$\hat{f}'_h(x) = \frac{1}{n} \sum_{i=1}^{n} K'_h(x - x_i) = \frac{1}{nh} \sum_{i=1}^{n} K'\left(\frac{x - x_i}{h}\right). \tag{1.3}$$

In the regression setting, the first derivative is estimated by keeping the b parameter instead of the a parameter in Equation (1.2), given as

$$\hat{\mu}'_h(x) = \text{argmin}_b \sum_{i=1}^{n} \left[y_i - (a + b(x_i - x))\right]^2 K_h(x - x_i).$$

After estimating the first derivative at some location x and smoothing level h, a test for significance is done. If the estimated first derivative is found to be significantly positive/negative, this is marked with a blue/red pixel in the SiZer map.

As an example on how a SiZer analysis can be performed, a real-world data set is analyzed. The data consists of the maximal snow-depth of each year from 1920 to 2005 in Tromsø, Norway. The left panel of Figure 1.1 shows the data set as green dots. The solid blue lines are nonparametric curve estimates generated by using Equation (1.2) for a range of bandwidth parameters. There seem to be a varying number of features (peaks and dips) in the regression estimates for the different bandwidths. The goal of the SiZer analysis is to determine the features that are actually present, and not just sampling artifacts.

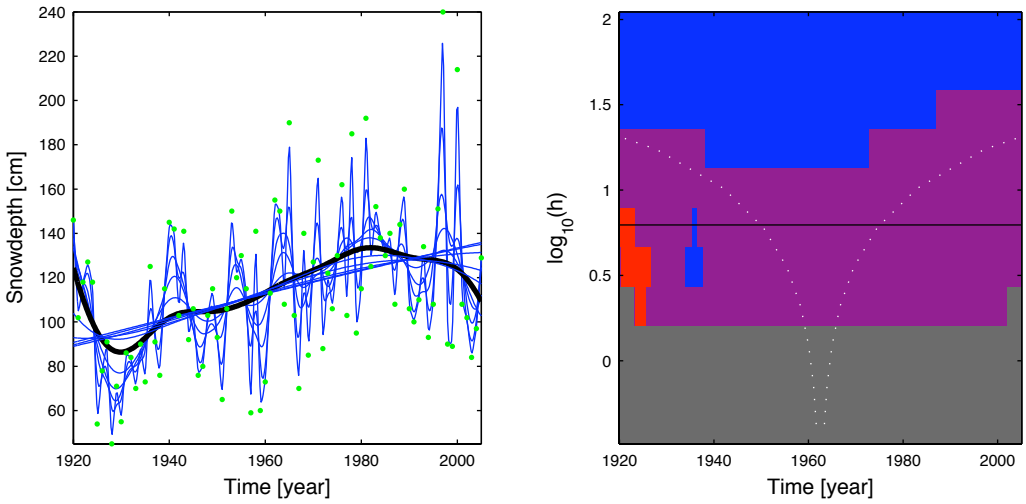

Figure 1.1: Left panel: Blue lines are curve estimates of the snow-depth data (green dots) using a Gaussian kernel in Equation (1.2). The black line corresponds to the result of using the bandwidth selected by the plug-in estimator from Ruppert et al. (1995). Right panel: SiZer map of the snow-depth data with a significance level of 0.05. The horizontal black line indicates the bandwidth selected by the plug-in method.

The right panel of Figure 1.1 shows the SiZer map, indicating for which time periods and scales (smoothing level) the slope of the nonparametric regression line is significantly different from zero. Red/blue pixels mark (location, scale) pairs where the slope is significantly decreasing/increasing. The choice of colors follow the original SiZer work, while they are swapped in Paper III. This is done since for some applications, such as temperature data sets, it seems more natural to use red for an increase, and blue for a decrease. The purple pixels indicate non-significant derivatives, while the gray pixels indicate that there are too few/sparse data points to do inference on the given location/scale. The horizontal distance between the two dotted white lines indicates the effective size of the smoothing parameter on the second axis. As can be seen, the period from around 1920 to 1940 of seemingly low snow-depths, shows up as a significant dip in the data when viewed on some

smoothing levels. This period of low snowfall is well-known to local meteorologists.

Scale-Space Methods

The concept of a SiZer map has been carried over to many other scale-space techniques. All the methods presented in this thesis include some sort of significance map closely related to the original SiZer map. One advantage of such maps is that they can communicate results for all scales and all positions simultaneously. A problem arises when the data for instance represent an image. Then the data are two-dimensional, making a natural extension of the two-dimensional SiZer map into a three-dimensional map. This problem was examined by Godtliebsen et al. (2004), where they chose to present significance maps connected to different smoothing levels as different slides of a time-lapse video. For the three papers of this thesis (Papers I-III, respectively), the first axis represents frequency, position (e.g. time), and wavelet scale number. For Papers I and III, the second axis represents the smoothing level. For Paper II, the second axis corresponds to the width of a summation window.

Papers I and III, the spectral feature detection and wavelet variance analysis papers, are more similar to each other than to Paper II (testing of multinormality and the k-sample problem). This is because Papers I and III apply a Bayesian scale-space approach. In general, when a Bayesian approach is taken, some sort of prior knowledge of the problem at hand is utilized in the procedure. Here, this is done by introducing a random walk smoothing prior distribution of either order one or two (Rue and Held, 2005). The chosen prior, with some smoothing parameter, determines how smooth realizations one can expect from the prior. Thus, when calculating the posterior distribution, a large smoothing parameter will give a smoother posterior distribution.

Using a smoothing prior in the scale-space field of signal analysis was introduced in the Posterior Smoothing method of Godtliebsen and Øigård (2005), and it was expanded and improved in Øigård et al. (2006). In the first paper, the quantiles were found through sampling the posterior distribution, while for Gaussian noise models the quantiles can be calculated exactly using the methods of the second paper. A basic problem with the general SiZer methodology is that it cannot handle data sets with few/sparse data points. Since the posterior distribution is used to find significant features, Posterior Smoothing does not run into such problems with sparse/few data. The ability to handle situations with few data points is essential for the methods in Papers I and III, and in particular for Paper III, where the number of data points "never" will be larger than 25.

The review paper by Holmström (2010a) thoroughly presents the evolution of the scale-space field and discusses many of the suggested methods. The BSiZer method was introduced in Erästö and Holmström (2005, 2007). Here, a different Bayesian approach on the SiZer methodology is taken. The review paper of Holmström (2010b) looks at the original BSiZer and extensions to the analysis of images and random fields. Other notable scale-space techniques include the extension of the SiZer methodology to dependent data in Park et al. (2004) for goodness-of-fit testing of time series models. This dependent SiZer method assumes that a times series follows a given time series model. The autocovariance function

connected to this model is incorporated in a scale-space test of this assumption. Scale-space bivariate density estimation was introduced in Godtliebsen et al. (2002), and Godtliebsen et al. (2004) looked at scale-space methods for detecting significant features in images, something which also is the focus of Holmström et al. (2011). Scale-space detection of significant changes of time-separated images is investigated in Holmström and Pasanen (2012). The detection and localization of non-stationarities and change-points is treated in scale-space manners in Kim and Marron (2006); Olsen et al. (2008a,b); Park et al. (2007). A graphical tool that tries to differentiate significant trends from dependence in the data is presented in Rondonotti et al. (2007).

In Chapter 2, Papers I-III will be presented and the results will be discussed. Ideas about future work will also be presented. Chapter 3 contains the conclusions of the thesis. Chapter 4 contains the three papers that make up the main body of the thesis.

Chapter 2

Results and Discussion

This chapter presents and discusses the three papers making up the thesis. For each paper, the problems and the suggested scale-space solutions/methods are presented and discussed. Some future work connected to each paper is also suggested.

2.1 Paper I - Spectral Feature Detection

Often scientists are interested in investigating whether or not a data set contains some periodic components. If detected, such periodic components can be connected to real-world phenomena. Mathematically, periodic components can be described by sinusoidal functions. Decomposing a signal into a set of weighted sinusoidal functions is essential what is done in spectral analysis. The spectral density function (SDF), or power spectrum, of a zero-mean stationary real-valued stochastic process x_t, $t \in \mathbb{Z}$, is defined by

$$f(\omega) \equiv \sum_{h=-\infty}^{\infty} \gamma(h) e^{-2\pi i \omega h}, \quad -1/2 \leq \omega \leq 1/2,$$

where the autocovariance function of the series, $\gamma(h) = E(x_t x_{t+h})$, $h \in \mathbb{Z}$ is assumed absolutely summable, i.e. $\sum |\gamma(h)| < \infty$. Frequencies connected to high $f(\omega)$ values are naturally of interest, since high SDF values indicate that the signal contains a strong periodic component connected to the frequency ω. Whether or not a peak is significant or just a sampling noise artifact might not always be obvious from an estimated SDF. Paper I introduces a scale-space approach to spectral feature detection. The idea is to detect significant peaks in the true underlying SDF viewed at different resolutions or scales. A peak is found if, for increasing frequencies, one goes from a frequency region of a significantly positive first derivative of the SDF, to a region with a significantly negative first derivative. Between these two significant regions, there can be a small region of non-significant first derivatives.

Estimating the SDF of some evenly sampled stationary signal is one of the classical statistical problems (Brockwell and Davis, 1991; Brillinger, 2001). The classical, but infamous, way of estimating the SDF of an evenly sampled time series is through calculating

the periodogram, defined as

$$I(\omega) \equiv \frac{1}{2n} \left| \sum_{t=1}^{2n} x_t e^{-2\pi i \omega t} \right|^2, \quad -1/2 \leq \omega \leq 1/2,$$

where $x_1, \ldots, x_{2n}$ denote the observed time series. In essence, the periodogram measures how well a set of different sinusoidal functions "fit" the data. The raw periodogram is known to be asymptotically unbiased, but not consistent, resulting in a highly fluctuating estimate. To smooth the periodogram is a common way of getting a consistent estimator of the SDF and makes it easier to interpret the output (Blackman and Tukey, 1958; Parzen, 1961; Priestley, 1981). The method presented in Paper I smooths the periodogram through a Bayesian approach.

As mentioned, the periodogram is generally far from the optimal way of estimating the SDF of a signal. The reason for it being used in Paper I, is that it produces uncorrelated sampling noise across the normal set of Fourier frequencies, given by $\omega_j = j/2n, j = 1, \ldots, n$. Having uncorrelated noise is essential for the posterior smoothing scheme of Paper I. The distribution of the log-transformed periodogram values is also easy to work with, even though the noise is non-Gaussian. By using an integrated Wiener process as the prior distribution, a Gaussian Markov Random Field (GMRF) that includes the first derivative of the log-transformed periodogram values can be constructed. The needed posterior marginals of the first derivatives can be very accurately approximated through the simplified Laplace approximation of Rue et al. (2009). These calculations have a low computational cost compared to Markov chain Monte Carlo methods. The "scale" part is introduced through the smoothing parameter of the prior.

The results are displayed in significance maps, illustrating for which smoothing levels and for which frequencies, peaks in the log-spectral density are detected as significant.

Future work

The periodogram requires the sampling to be even. Extensions of the periodogram to unevenly sampling were suggested by Lomb (1976) and Scargle (1982). By selecting a set of evenly spaced frequencies, similar to the Fourier frequencies, the estimated spectral density values will in general be correlated. By knowing the sampling pattern (in time or space), the correlation structure of the spectral estimates can be estimated.

For future work, the method of Paper I is planned to be extended to also handle uneven sampling. Currently, the implementation of the procedure do handle unevenly sampled data, but the spectral estimates coming out of the Lomb-Scargle procedure are (wrongly) assumed to be uncorrelated. The set of frequencies are selected based on an average sampling distance. To do it right, the correlations have to be incorporated into the posterior distribution somehow. Alternatively, one can select frequencies that are close to having zero correlation to all other frequencies. Generally, this will seldom be possible to do, except for when the sampling is close to even.

It might also be of interest to extend the procedure to the case of looking for areas with low spectral content, also called spectral dips. For communications signal processing, such

dips will represent frequency regions for which the amount of received power is low. The larger bias of the periodogram for low spectral power levels then has to be handled. This can be done by using a multitaper technique to generate the spectral estimates. Doing this, the grid of frequencies for which the spectral estimates will be uncorrelated will be coarser than the normal Fourier frequency grid (Percival and Walden, 1993).

2.2 Paper II - Multinormality Testing and the k-sample Problem

Scale-space approaches of two different classical statistical problems are presented in Paper II. The first problem is goodness-of-fit testing of multinormality (multivariate Gaussian distribution). It is very common to assume that some data set originates from a multinormal distribution. This is often assumed even though one might know it to be false, but the severity of making this false assumption depends on the problem at hand and the sensitivity to non-normal data of the statistical analysis being done on the data (Cox and Wermuth, 1994; Farrell et al., 2007; Looney, 1995).

There are a lot of different methods for testing for multinormality (Farrell et al., 2007; Mecklin and Mundfrom, 2004). In general, almost all of these tests at some point of the algorithms need to invert an estimated covariance matrix. If there are at least as many dimensions p as there are samples n, then the estimated covariance matrix is non-invertible. Hence, these methods fail in the case of $n \leq p$, also known as the High Dimension Low Sample Size (HDLSS) case. The presented scale-space method does not need to invert any covariance matrices, and thereby all combinations of sample and dimension sizes can be treated. Having this ability comes with a prize. The information that might lie in the covariance structure, when $n > p$, is not utilized by the presented method. This also is the case for the presented k-sample method.

The "scale" part of the method is connected to how many neighboring dimensions that are summed across. This summation window size, which makes up the vertical axis of the resulting significance map, is varied through the odd numbers from 1 to the number of dimensions. The "space" part is connected to the different dimensions of the data. As the summation is done across neighboring dimensions, changing the order of the dimensions around might change the results of the analysis. Therefore, the method should be restricted to data sets where there is a natural ordering of the dimensions, e.g. time series or one-dimensional spatial data.

The result of doing the summation is a vector of length n. This vector is then tested for univariate normality through the well-established Anderson-Darling test (Anderson and Darling, 1952, 1954; Lewis, 1961). A rejection/acceptance of normality is indicated with a red/blue pixel in the significance map at the coordinates corresponding to the summation window width and the dimension for which this window was centered on. This way of testing is not adequate to conclude that the data set in total might originate from some unspecified multinormal distribution, but a significance map dominated by blue pixels is a

strong indication on that being the case.

The k-sample problem is the other problem of this paper. A common question is whether or not k data sets of equal dimension size p might originate from some unknown discrete or continuous multivariate distribution. Also here there are a huge number of suggested algorithms. Most of these need to invert estimated covariance matrices, making them useless in the HDLSS case. Each of the k data sets are processed as in the testing for multinormality. So, for a given summation window size and center of summation, k vectors of potential unequal lengths are compared through the k-sample Anderson-Darling test (Pettitt, 1976; Scholz and Stephens, 1987).

As the presented methods have the special quality of being able to handle the HDLSS case, the two presented scale-space methods are compared to the only other methods that handle the HDLSS case. To the author's knowledge, only the methods of Liang et al. (2000, 2009) handles HDLSS for the multinormality case, while for the k-sample problem the methods of Friedman and Rafsky (1979); Hall and Tajvidi (2002); Henze (1988); Székely and Rizzo (2004) were used for comparison.

As shown in the paper, it is possible to "enhance" small mean-value differences when summing across dimensions, see the discussion at the start of the Results chapter concerning the motivational example. These mean-value differences can either be across data sets (the k-sample situation), or as the motivational example where the difference is connected to different modes in the data set. Unfortunately, there is nothing to gain by summing across dimensions when a difference is connected to the variance. A simple example explains this.

Assume that a data set consist of ten zero-mean, uncorrelated normally distributed three-dimensional vectors. The five first vectors have variance 1 for all dimensions, while the last five have variance 4 for all dimensions. When summing across the three dimensions (assuming equal summation weights of 1/3), the five first elements will be normally distributed with zero mean and variance equal to 1/3. The last five elements will on the other hand have a variance of 4/3. So, the variance has gone from 1 and 4, to 1/3 and 4/3, for the first and last five elements, respectively. Hence, the relative difference of the variances has not changed as a result of doing the summation.

The presented scale-space multinormality testing procedure could be extended to other distributions as long as the distribution is closed under summation. A distribution is thought of as being closed under summation if a sum of variables (with potentially different parameters) from this distribution has this same type of distribution, with parameters depending on the individual parameters. The normal distribution is an obvious example, and it is closed under summation also for variables that originate from a multinormal distribution with non-zero covariance elements. The χ^2, Poisson, Cauchy and Lévy distributions are other relevant examples. Unfortunately, these distributions require the variables to be independent for them to be guaranteed to be closed under summation.

Future work

Normalizing each dimension individually was considered for the paper. This might be relevant when there are large differences in the variance of the different dimensions. If

one of the dimensions has a clearly larger variance than the rest of the dimensions, this dimension might "dominate" when doing the summations. Figure 2.1 illustrates how large variances can influence the resulting significance map. The data set consists of 200 samples of 55 dimensions. All dimensions are multinormally distributed (as for the motivational example) except dimensions 11, 26 and 41. For these three dimensions, the 100 first dimensions have zero mean, while the other 100 samples have mean values of -2.5, 7, and 12, respectively. The figure clearly shows how large deviations from normality (large separation of the two modes) make the deviations, which originate in only one dimension, "spread" more upwards compared to when the deviations from normality are less clear. The rejection region of the highest scales is the result of the window picking up the non-normality at both dimension 26 and 41. When each dimension is normalized individually, the only remaining rejections are the scale 1 rejections of dimension 11, 26 and 41 (not shown).

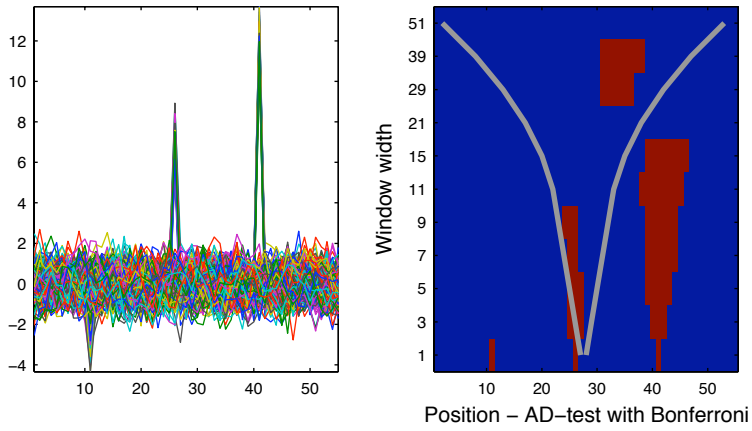

Figure 2.1: Left panel: 200 artificial signals of length 55. Right panel: Significance map of test for multinormality with a significance level of 0.05.

Introducing this into the suggested methods has been considered, but as of now it is not implemented. Doing so is quickly done, and will not change much at all in the paper. Part of the reason for not implementing it yet is that no examples were investigated where it was thought that a normalization of the data would significantly influence the results. These considerations were done early in the process of the work, so it might be worth looking into again before final publication of the paper.

2.3 Paper III - Log Wavelet Variance

Paper III is a redone version of Paper VI of the Ph.D thesis of Øigård (2004). A paper based on this manuscript was also submitted to Journal of Computational and Graphical Statistics in 2005. As pointed out by the reviewers, the needed changes to the paper were

quite massive. Based on the underlying idea presented in Øigård (2004) and the submitted paper, and with the permission of the original authors, a complete redo of the paper has been done, mainly by the two current first authors. One of the most prominent changes include the move from using the Discrete Wavelet Transform (DWT) to the Maximum Overlap DWT (Percival, 1995; Percival and Walden, 2000). The dependence assumptions of the wavelet variance estimates have also been totally changed. These changes resulted in the need to construct a completely new procedure to estimate the relevant quantiles of the posterior distribution.

Paper III suggests a wavelet-based scale-space analysis of time series. The wavelet coefficients of a time series indicate for which scales and locations the significant "energy" of the time series can be found (Percival and Walden, 2000). For instance, if a time series is dominated by changes on short time-scales, the wavelet coefficients connected to the lower scales will have larger absolute values than the coefficients connected to higher scales. If the time series is non-stationary, the distribution of the wavelet coefficients will change as a function of time/location for at least some of the scale parameters. The wavelet variance of the jth scale is defined as the variance of the wavelet coefficients of the jth scale. Increasing values of the scales parameter j corresponds to wider wavelets filters.

Stochastic processes X_t with long-range dependence (LRD) are known to have autocovariance sequences s_τ that decay so slowly that the sum $\sum_{\tau'} s_{\tau'}$ diverges. As the time lag τ of the ACVS increases, the ACVS can be described by

$$s_\tau = \mathrm{cov}(X_t, X_{t+\tau}) \sim C\tau^{-\gamma},$$

where $0 < \gamma < 1$, $\sim$ denotes asymptotic equivalence, and C is a real constant. The γ parameter is connected to the Hurst parameter H through $\gamma = 2 - 2H$. Thereby, the degree of LRD is fully characterized by this Hurst parameter.

Let the log wavelet variance be defined as the natural logarithm of the wavelet variance. If the time series has LRD, the log wavelet variance will follow a straight line for the upper scale numbers j. The slope of this line can be directly used to estimate the Hurst parameter of the time series (Abry and Veitch, 1998; Abry et al., 2003). Hence, the LRD of a time series will directly influence the slope of the upper part of the log wavelet variance. As demonstrated in Stoev et al. (2005), this linearity will in general not extend all the way down to the lowermost scale parameters. Therefore, it is important to have a way to determine for which scale parameters the log wavelet variance can be said to be linear. An automatic method to determine this is given in Veitch et al. (2003). This method is also partially used in the presented algorithm. Paper III presents a Bayesian scale-space method to investigate for which wavelet scale parameters this assumed linearity can be said to be true. The "scale" part of the algorithm is connected to the smoothing parameter of the random walk 2 prior.

As shown in Stoev et al. (2005), different forms of non-stationarity of a times series might influence the shape of the estimated log wavelet variance. To be able to detect some relevant non-stationarities, a local analysis is also suggested in Paper III. The time series is divided into non-overlapping windows of equal length. For each of these windows, the

log wavelet variance is estimated. By comparing these local log wavelet variance analyses to each other, one might detect non-stationarities in the signal. If the signal is stationary, the results of the different local analyses will not be different.

In the other papers of this thesis, the False Discovery Rate (FDR) correction for multiple testing (Benjamini and Hochberg, 1995) has been used along with the Bonferroni correction (Hochberg and Tamhane, 1987). The reason for it not being used in Paper III is that there are always very few data points (the wavelet variance estimates). For instance, with a mother wavelet filter of length 8, the data has to contain close to 235 million samples to have an uppermost scale number of $J_0 = 25$. Using an FDR correction on such few data points seems a bit unnecessary. To investigate the influence on the choice of the significance level, this level can either be varied, or the planned "significance maps" described in the paper can be used. Developing such "significance maps", as apposed to the current feature/summary maps, should be quite doable, and the authors are considering doing this before making the final version for submission to a journal.

Future work

As of now, non-overlapping windows are used in the local analysis. For each window, an increasing number of "boundary wavelet coefficients" are ignored in the analysis as the wavelet scale increases. These coefficients will be at the left/first part of each window. If important features of the data are located in this part of the data set, these features might not be detected when using non-overlapping windows. Changing the code to use overlapping windows is trivial, and adding this as an option for the user will be done. It will also be possible for the user to set the degree of overlap.

For the uppermost wavelet scale number J_0, the distributional assumptions presented in the paper are bad. When inspecting the distribution of the wavelet variance of the uppermost scale of simulated fGN data, the distribution cannot be fitted adequately with any χ^2 distribution. The way the degrees of freedom (DOF) is estimated relies on this χ^2 assumption of the wavelet variance. Thereby, the estimated DOF of the uppermost scale is also not reliable. This uppermost DOF is only used in the bias correction, which is also based on the assumed χ^2 distribution, of the estimated log wavelet variance. Therefore, the distributional assumption of the uppermost scale is inaccurate, and the uppermost scale is excluded from the analyses.

Finding a better description of the distribution of the wavelet variance of the uppermost scale has been investigated as part of this thesis, and some ad hoc methods for selecting a "better" DOF has been made. The general validity of this ad hoc method is questionable, and as there are few wavelet coefficients for the uppermost scale, the variance will be large compared to the other relevant scales. Therefore, none of these ad hoc methods are included in the presented algorithms. As long as the distribution of the log wavelet variance of the uppermost scale cannot be adequately approximated by a normal distribution, even a better knowledge of this uppermost distribution does not work well with the assumed multinormal likelihood/posterior distribution of the presented paper.

Chapter 3

Conclusions

Through the research and analyses in this thesis, the collection of statistical methods based on the scale-space methodology has been expanded. The methods presented are exploratory tools that should be used at an early stage of a statistical data analysis. In all three papers, the presented tools output some sort of two-dimensional maps, indicating for which level of smoothing/averaging (the "scale" parameter) and for which location/time/wavelet scale (the "space" parameter) the relevant null hypothesis is rejected.

The tools from Papers I, II and III allow the user to quickly see:

I) The location (frequency) and smoothing level for which the first derivative of the estimated spectral density function can be said to be significantly positive or negative, or not significantly different from zero. From this, peaks in the spectrum can be located, something that can be connected to real-world periodicies.

II) The location (time or sampling index, typically) and neighborhood (number of dimensions being averaged across) for which:

   i) A multivariate data set cannot be said to originate from some unspecified multivariate Gaussian distribution.

   ii) Two or more data sets cannot be said to originate from the same unspecified multivariate discrete or continuous distribution.

III) The location (wavelet scale number) and smoothing level for which the log wavelet variance cannot be described by a linear relationship as a function of the wavelet scale number. The suggested local log wavelet variance analysis can also detect some forms of non-stationarities in the data.

All the presented tools are easy to use, and the results are presented in two-dimensional images/maps that scientists with just basic statistical training can quickly interpret. The methods are implemented in MATLAB, and for the spectral method in Paper I, a user-friendly graphical user interface exists. Following publication of Papers II and III, user-friendly MATLAB code will be made available to the public.

Bibliography

Abry, P., Flandrin, P., Taqqu, M. S., Veitch, D., 2003. Theory and Applications of Long-range Dependence. Birkhäuser, Ch. Self-similarity and long-range dependence through the wavelet lens, pp. 527–556.

Abry, P., Veitch, D., January 1998. Wavelet analysis of long-range dependent traffic. IEEE Transactions on Information Theory 44 (1), 2–15.

Anderson, T. W., Darling, D. A., June 1952. Asymptotic theory of certain "goodness of fit" criteria based on stochastic processes. Annals of Mathematical Statistics 23 (2), 193–212.

Anderson, T. W., Darling, D. A., December 1954. A test of goodness of fit. Journal of the American Statistical Association 49 (268), 765–769.

Benjamini, Y., Hochberg, Y., 1995. Controlling the false discovery rate: A practical and powerful approach to multiple testing. Journal of the Royal Statistical Society. Series B (Methodological) 57 (1), 289–300.

Blackman, R. B., Tukey, J. W., 1958. The measurement of power spectra from the viewpoint of communications engineering. Bell System Technical Journal 37 (1/2), 185–282/485–569.

Brillinger, D. R., 2001. Time Series: Data Analysis and Theory. Society for Industrial and Applied Mathematics, Philadelphia, PA.

Brockwell, P. J., Davis, R. A., 1991. Time Series: Theory and Methods, 2nd Edition. Springer-Verlag, New York, NY.

Chaudhuri, P., Marron, J. S., September 1999. SiZer for exploration of structures in curves. Journal of the American Statistical Association 94 (447), 807–823.

Chaudhuri, P., Marron, J. S., April 2000. Scale space view of curve estimation. The Annals of Statistics 28 (2), 408–428.

Cox, D. R., Wermuth, N., 1994. Tests of linearity, multivariate normality and the adequacy of linear scores. Journal of the Royal Statistical Society. Series C (Applied Statistics) 43 (2), 347–355.

Erästö, P., Holmström, L., September 2005. Bayesian multiscale smoothing for making inferences about features in scatterplots. Journal of Computational and Graphical Statistics 14 (3), 569–589.

Erästö, P., Holmström, L., May 2007. Bayesian analysis of features in a scatter plot with dependent observations and errors in predictors. Journal of Statistical Computation and Simulation 77 (5), 421–434.

Farrell, P. J., Salibian-Barrera, M., Naczk, K., December 2007. On tests for multivariate normality and associated simulation studies. Journal of Statistical Computation and Simulation 77 (12), 1065–1080.

Friedman, J. H., Rafsky, L. C., July 1979. Multivariate generalizations of the Wald-Wolfowitz and Smirnov two-sample tests. The Annals of Statistics 7 (4), 697–717.

Godtliebsen, F., Marron, J. S. ., Chaudhuri, P., March 2002. Significance in scale space for bivariate density estimation. Journal of Computational and Graphical Statistics 11 (1), 1–21.

Godtliebsen, F., Marronb, J. S., Chaudhuri, P., November 2004. Statistical significance of features in digital images. Image and Vision Computing 22 (13), 1093–1104.

Godtliebsen, F., Øigård, T. A., February 2005. A visual display device for significant features in complicated signals. Computational Statistics and Data Analysis 48 (2), 317–343.

Hall, P., Tajvidi, N., June 2002. Permutation tests for equality of distributions in high-dimensional settings. Biometrika 89 (2), 359–374.

Hannig, J., Lee, T. C. M., March 2006. Robust SiZer for exploration of regression structures and outlier detection. Journal of Computational and Graphical Statistics 15 (1), 101–117.

Hannig, J., Marron, J. S., June 2006. Advanced distribution theory for SiZer. Journal of the American Statistical Association 101 (474), 484–499.

Henze, N., June 1988. A multivariate two-sample test based on the number of nearest neighbor type coincidences. The Annals of Statistics 16 (2), 772–783.

Hochberg, Y., Tamhane, A. C., 1987. Multiple Comparison Procedures. Wiley series in probability and mathematical statistics. Applied probability and statistics. John Wiley & Sons, New York, NY.

Holmström, L., September/October 2010a. BSiZer. Wiley Interdisciplinary Reviews: Computational Statistics 2 (5), 526–534.

Holmström, L., March/April 2010b. Scale space methods. Wiley Interdisciplinary Reviews: Computational Statistics 2 (2), 150–159.

Holmström, L., Pasanen, L., February 2012. Bayesian scale space analysis of differences in images, to be published in Technometrics, vol. 54, No. 1.

Holmström, L., Pasanen, L., Furrer, R., Sain, S. R., October 2011. Scale space multiresolution analysis of random signals. Computational Statistics and Data Analysis 55 (10), 2840–2855.

Kim, C. S., Marron, J. S. ., January 2006. SiZer for jump detection. Journal of Nonparametric Statistics 18 (1), 13–20.

Lewis, P. A. W., December 1961. Distribution of the Anderson-Darling statistic. Annals of Mathematical Statistics 32 (4), 1118–1124.

Liang, J., Li, R., Fang, H., Fang, K.-T., April 2000. Testing multinormality based on low-dimensional projection. Journal of Statistical Planning and Inference 86 (1), 129–141.

Liang, J., Tang, M.-L., Chan, P. S., September 2009. A generalized Shapiro-Wilk W statistic for testing high-dimensional normality. Computational Statistics and Data Analysis 53 (11), 3883–3891.

Lindeberg, T., 1994. Scale-Space Theory in Computer Vision. The Kluwer international series in engineering and computer science. Kluwer Academic, Boston, MA.

Lomb, N. R., February 1976. Least-squares frequency analysis of unequally spaced data. Astrophysics and Space Science 39 (2), 447–462.

Looney, S. W., February 1995. How to use tests for univariate normality to assess multivariate normality. The American Statistician 49 (1), 64–70.

Mecklin, C. J., Mundfrom, D. J., April 2004. An appraisal and bibliography of tests for multivariate normality. International Statistical Review 72 (1), 123–138.

Øigård, T. A., November 2004. Statistical analysis and representation of non-trivial signals and random processes. Ph.D. thesis, University of Tromsø, Norway.

Øigård, T. A., Rue, H., Godtliebsen, F., December 2006. Bayesian multiscale analysis for time series data. Computational Statistics and Data Analysis 51 (3), 1719—1730.

Olsen, L. R., Chaudhuri, P., Godtliebsen, F., March 2008a. Multiscale spectral analysis for detecting short and long range change points in time series. Computational Statistics and Data Analysis 52 (7), 3310–3330.

Olsen, L. R., Sørbye, S. H., Godtliebsen, F., March 2008b. A scale-space approach for detecting non-stationarities in time series. Scandinavian Journal of Statistics 35 (1), 119–138.

Park, C., Godtliebsen, F., Taqqu, M. S., Stoev, S. A., Marron, J. S., August 2007. Visualization and inference based on wavelet coefficients, SiZer and SiNos. Computational Statistics and Data Analysis 51 (12), 5994–6012.

Park, C., Marron, J. S., Rondonotti, V., October 2004. Dependent SiZer: goodness-of-git tests for time series models. Journal of Applied Statistics 31 (8), 999–1017.

Parzen, E., May 1961. Mathematical considerations in the estimation of spectra. Technometrics 3 (2), 167–190.

Percival, D. B., September 1995. On estimation of the wavelet variance. Biometrika 82 (3), 619–631.

Percival, D. B., Walden, A. T., 1993. Spectral Analysis for Physical Applications. Cambridge University Press, Cambridge, United Kingdom.

Percival, D. B., Walden, A. T., 2000. Wavelet methods for time series analysis. Cambridge series on statistical and probabilistic mathematics. Cambridge University Press, New York, NY.

Pettitt, A. N., April 1976. A two-sample Anderson-Darling rank statistic. Biometrika 63 (1), 161–168.

Priestley, M. B., 1981. Spectral Analysis and Time Series. Academic Press, London, UK.

Rondonotti, V., Marron, J. S., Park, C., 2007. SiZer for time series: A new approach to the analysis of trends. Electronic Journal of Statistics 1, 268–289.

Rue, H., Held, L., 2005. Gaussian Markov Random Fields: Theory and Applications. Vol. 104 of Monographs on statistics and applied probability. Chapman & Hall/CRC, Boca Raton, FL.

Rue, H., Martino, S., Chopin, N., April 2009. Approximate Bayesian inference for latent Gaussian models by using integrated nested Laplace approximations. Journal of the Royal Statistical Society. Series B (Methodological) 71 (2), 319–392.

Ruppert, D., Sheather, S. J., Wand, M. P., December 1995. An effective bandwidth selector for local least squares regression. Journal of the American Statistical Association 90 (432), 1257–1270.

Scargle, J. D., December 1982. Statistical aspects of spectral analysis of unevenly spaced data. Astrophysical Journal 263 (1), 835–853.

Scholz, F. W., Stephens, M. A., September 1987. K-sample Anderson-Darling tests. Journal of the American Statistical Association 82 (399), 918–924.

Stoev, S. A., Taqqu, M. S., Park, C., Marron, J. S., June 2005. On the wavelet spectrum diagnostic for hurst parameter estimation in the analysis of internet traffic. Computer Networks 48 (3), 423–445.

Székely, G. J., Rizzo, M. L., November 2004. Testing for equal distributions in high dimensions. InterStat (5), 1–16.

Veitch, D., Abry, P., Taqqu, M. S., December 2003. On the automatic selection of the onset of scaling. Fractals 11 (4), 377–390.

Wand, M. P., Jones, M. C., 1995. Kernel Smoothing. Vol. 60 of Monographs on Statistics and Applied Probability. Chapman & Hall/CRC, Boca Raton, FL.

Chapter 4

Papers I - III

Paper I:

Bayesian multiscale feature detection of log-spectral densities

Paper II:

A Novel Scale-Space Approach for Multinormality Testing and the k-Sample Problem

Paper III:
Scale-space wavelet variance

ISBN xxx-xx-xxxx-xxx-x