

A scale space approach for estimating the characteristic feature sizes in hierarchical signals

Leena Pasanen^{a*}, Tuomas Aakala^b and Lasse Holmström^a

^a Research Unit of Mathematical Sciences, University of Oulu, Oulu, Finland

^b Department of Forest Sciences, University of Helsinki, Helsinki, Finland

* leena.pasanen@oulu.fi

Abstract

The temporal and spatial data analyzed in e.g. ecology or climatology are often hierarchically structured, carrying information in different scales. An important goal of data analysis then is to decompose the observed signal into distinctive hierarchical levels and to determine the size of the features that each level represents. Using differences of smooths, scale space multiresolution analysis decomposes a signal into additive components associated with different levels of scales present in the data. The smoothing levels used to compute the differences are determined by the local minima of the norm of the so-called scale-derivative of the signal. While this procedure accomplishes the first goal, the hierarchical decomposition of the signal, it does not achieve the second goal, the determination of the actual size of the features corresponding to each hierarchical level. Here we show that the maximum of the scale-derivative norm of an extracted hierarchical component can be used to estimate its characteristic feature size. The feasibility of the method is demonstrated using an artificial image and a time series of a drought index, based on climate reconstructions from long tree ring chronologies.

Environmetrics; Image analysis Smoothing; Time series; Visualization

1 Introduction

Questions involving temporal and spatial data are common in ecology, environmental sciences and many other fields. The signals contained in such data are often hierarchically structured, carrying information in different temporal or spatial scales (Wu, 2013). For example, climate data often contain oscillations in several frequencies and satellite images can contain pixel-wise variation, variation at the landscape level emerging from different human land uses, as well as large scale gradients due to latitudinal variation in climatic conditions. Importantly, in addition to analyzing these scales and the consequent scale-dependent variation, knowledge of the actual sizes of the temporal or spatial features associated with each scale may

This article has been published in final form at <https://onlinelibrary.wiley.com/doi/full/10.1002/sta4.195>

offer important information about the patterns themselves and help link the observed patterns to the underlying physical processes (Delcourt & Delcourt, 1988; Skøien et al., 2003).

We propose a two-step procedure for the analysis of such hierarchical data. First, the signal is decomposed into components that represent its variation in different levels of hierarchy and second, the characteristic feature size of each hierarchical level is determined.

In the first step, scale space multiresolution analysis (Holmström et al., 2011), an instance of statistical scale space methodology, is applied. Conventional scale space analysis considers smooths of a signal, interpreting each smooth to provide information about the object of interest at a particular scale. The concept of a scale space has its origin in computer vision literature (Witkin, 1983, 1984; Lindeberg, 1994). The idea was first introduced to statistics by Chaudhuri & Marron (1999) and it has subsequently developed into a versatile ensemble of techniques with many applications (Holmström & Pasanen, 2017). In traditional scale space analysis, increasing smoothing progressively suppresses smaller scale data features, thus revealing increasingly coarse structures in the data. However, according to hierarchy theory (O’Neill, 1986), hierarchically structured data contains signals at specific, distinguishable scales (see also the overview of Wu (2013) and the references therein). Hence, if the goal is to analyze the signal in its different levels of hierarchy, the features underlying the observed data need to be extracted at particular scales, a task not easily accomplished with traditional scale space analysis. On the other hand, scale space multiresolution analysis is ideally designed for this task since it extracts the scale-dependent features of the signal using differences of smooths where a higher level smooth is subtracted from a lower level smooth, thus isolating the signal components corresponding different hierarchical levels.

A successful discovery of the hierarchical structure in the data naturally requires that the smoothing levels employed in the multiresolution analysis are chosen appropriately. To facilitate this crucial step, Pasanen et al. (2013) proposed to consider the so-called scale-derivative of the signal, that is, the derivative of the smooth of the signal with respect to the logarithm of the smoothing parameter. They showed that the smoothing level sequence for the multiresolution analysis can usefully be selected as the local minima of the norm of the scale-derivative.

Scale space multiresolution analysis has been shown to successfully resolve both temporal and spatial signals into their scale-dependent components (see Pasanen et al., 2017; Aakala et al., 2018; Fang et al., 2017; Lehmann et al., 2017; Heersink et al., 2013), but it does not provide direct information about the typical sizes of the signal features in the components. In fact, all one can say is that the hierarchical component extracted as the difference of two smooths has features that are larger than those smoothed out by the smaller scale smoother and smaller than those suppressed by the larger scale smoother. Other statistical scale space methods suffer from similar vagueness when the signal is analyzed with several smoothing levels and typically for each level simply the effective width of the smoothing kernel is thought to provide insight about the sizes of the signal features. Here, we complement scale space multiresolution analysis with an explicit estimate for the characteristic size of features of the extracted components.

The characteristic feature size of a multiresolution component will be determined by its “characteristic scale”, a quantity for which various definitions have been proposed in the literature, both for temporal and spatial data. For example, approaches based on wavelets (Keim & Percival, 2015), temporal and spatial variograms (Skøien et al., 2003) and quadtrees (Rehrauer et al., 1999) have been suggested, and recently Sangireddy et al. (2017) proposed a scale space based approach specifically designed for the definition of the characteristic length scale for high-resolution topographic data. However, in computer

vision oriented scale space research, the idea of a characteristic scale has in fact been of interest already much earlier. In that context, the characteristic scale of an object in an image is the scale in which it is most easily perceived by the observer (Lindeberg, 1998; Luo et al., 2007). More formally, it was proposed that the scale of a “blob”, a region in a digital image which is darker or brighter than its neighborhood, can be estimated as the scale for which a suitable function, such as the scale-normalized Laplace operator, attains an extremum over the scales. This idea has been widely used for example for blob and interest point detection (see e.g. Mikolajczyk & Schmid, 2004, and the references therein) and also in medical imaging for defining feature sizes for example in clustered microcalcifications in digital mammograms (Netsch & Peitgen, 1999) and lung nodule size in CT imaging (Diciotti et al., 2010). As the scale-derivative can be shown to approximate the scale-normalized second derivative (Laplace in a spatial setting; see Section S1 of the supporting information and also Lindeberg (2015)), we adopt Lindeberg’s definition for the characteristic scale and estimate it using the maximum of the scale-derivative norm. In computer vision and medical imaging, the characteristic scale is actually defined for each image feature separately. A different approach was taken in Luo et al. (2007) where the characteristic scale of a whole satellite image is determined using the scale-normalized total variation. Since our goal is to estimate the characteristic scale that represents the hierarchical component viewed over the entire time interval or spatial area considered in the analysis, and not the scale of each individual feature that comprises the component, we follow the approach of Luo et al. and maximize the scale-derivative norm of the whole component.

Lindeberg’s definition of a characteristic scale adopted here depends on the particular smoother used because, depending on the smoothing method applied, similar smoothing parameter values can produce very different levels of smoothing. In this article, we consider kernel and roughness penalty smoothers. On the other hand, characteristic feature size is a property of the data, independent of the smoother used in the analysis, and therefore the estimated characteristic scale needs to be transformed into a characteristic feature size. In Netsch & Peitgen (1999) and in Diciotti et al. (2010), the scale is transformed to size using the size of the Laplacian of the Gaussian smooth at the characteristic scale. On the other hand, in Luo et al. (2007) the scale is transformed into size using the characteristic scales of signals with a known period. Such a scale-period approach actually resembles the scale-pseudo-frequency transformation applied with wavelets (Torrence & Compo, 1998).

In time series analysis, periodicities in the signal are often of interest making the scale-period transformation a natural approach for estimating the characteristic size. However, for images a more natural interpretation of a characteristic size might be simply a typical feature size, and hence approaches that do not rely on periodicity might be more appropriate. Here, we will examine both the scale-period transformation and transformations based on the size of the smoother at the characteristic scale.

As a real-world example, we analyzed the climatic drought index constructed from climate reconstructions based on tree ring chronologies (see Aakala et al., 2018, for details). The low-frequency variability in northern European climate has been documented in a number of studies (e.g. Hurrell & Van Loon, 1997), and is well shown to be present in tree ring based reconstructions at multiple time scales (Esper et al., 2002; Sirén & Hari, 1971). Over the long term, this type of climatic variability has been shown to correlate with solar cycles (Kasatkina et al., 2007) that influence the flux density of galactic cosmic radiation, which in turn may influence cloudiness and climate patterns, and eventually also influence the properties of tree rings (Dengel et al., 2009). We will apply the scale space multiresolution analysis with size estimation to examine the presence of these periods in the data.

The rest of the article is organized as follows. The scale space multiresolution decomposition is summarized in Section 2. The estimation of the characteristic scale is introduced in Section 3 and the alternative approaches for converting characteristic scale to size are considered in Section 4. In Section 5, we demonstrate the feasibility of our approach for an artificial image and the drought index time series. Finally, in Section 6, we discuss our findings.

2 The scale space multiresolution decomposition

Let \mathbf{x} be a signal, e.g. a time series or a vectorized digital image, and let \mathbf{S}_λ denote a smoothing operator (a matrix) that defines a smooth $\mathbf{S}_\lambda \mathbf{x}$ of \mathbf{x} . The parameter $\lambda > 0$ determines the smoothing level, the larger the value of λ , the smoother the result. We use here Nadaraya-Watson kernel regression (Nadaraya, 1964; Watson, 1964) (for one dimensional data see also Wand & Jones (1994), and for the general multivariate case see e.g. Klemelä (2014)) and roughness penalty smoothers (see Appendix A). We seek an expansion of \mathbf{x} into additive hierarchical components that contain features in different temporal or spatial scales. The components are estimated using differences of smooths of \mathbf{x} .

Thus, let us denote $\mathbf{S}_0 \mathbf{x} := \mathbf{x}$ and consider a sequence of smoothing levels $0 = \lambda_1 < \dots < \lambda_L < \infty$. The signal \mathbf{x} can then be decomposed as

$$\mathbf{x} = \sum_{i=1}^{L-1} (\mathbf{S}_{\lambda_i} - \mathbf{S}_{\lambda_{i+1}}) \mathbf{x} + (\mathbf{S}_{\lambda_L} \mathbf{x} - \bar{\mathbf{x}}) + \bar{\mathbf{x}} = \sum_{i=1}^{L+1} \mathbf{z}_i, \quad (1)$$

where $\mathbf{z}_{L+1} = \bar{\mathbf{x}}$ is the mean of the signal and $\mathbf{z}_1, \dots, \mathbf{z}_L$ are the scale-dependent hierarchical components we are interested in. The component \mathbf{z}_i can be interpreted as the detail in \mathbf{x} which is smoothed out when smoothing is increased from λ_i to λ_{i+1} . We therefore refer to the smoothing levels λ_i also as the scale breaking points. If \mathbf{x} can be viewed as a random vector, Bayesian inference can be used to infer the time intervals/regions where \mathbf{z}_i differs credibly from zero (Holmström et al., 2011; Pasanen et al., 2013).

A successful hierarchical decomposition (1) requires that the smoothing levels λ_i are selected carefully. To help make the selection, Pasanen et al. (2013) introduced the concept of a scale-derivative,

$$\frac{\partial \mathbf{S}_\lambda \mathbf{x}}{\partial \log \lambda} := \mathbf{D}_\lambda \mathbf{x},$$

that measures the change in the smoothed signal with respect to the logarithmic change in the scale (here \log denotes the natural logarithm). They used the scale-derivative with a one dimensional roughness penalty smoother, for which a simple analytic expression of the scale-derivative can be derived (see Pasanen et al., 2013). For other smoothers, a useful analytic expression might not be available, but the derivative can always be approximated using a difference quotient,

$$-\frac{(\mathbf{S}_\lambda - \mathbf{S}_{\gamma\lambda}) \mathbf{x}}{\log \gamma} \approx \mathbf{D}_\lambda \mathbf{x},$$

where $\gamma > 1$. In the experiments of this article, we used for γ values between 1.02 and 1.07.

For time series data, Pasanen et al. (2013) introduced also the so-called scale-derivative map to visualize $\mathbf{D}_\lambda \mathbf{x}$. It is a plot with time on the horizontal axis, $\log_{10}(\lambda)$ on the vertical axis, and the color of the pixels

reflects the value of the scale-derivative. The middle row of Figure 4 displays examples of scale-derivative maps. The scale-dependent components of the signal \mathbf{x} show as oscillating bands of blue and red and they can be extracted as differences of smooths corresponding to the sequence of smoothing levels defined by the local minima of the function $\lambda \mapsto \|\mathbf{D}_\lambda \mathbf{x}\|$ (cf. Pasanen et al., 2013). This method of choosing the smoothing levels can be applied also to images. More insight into the character of the components \mathbf{z}_j can be obtained by examining also the norms of their scale-derivatives.

3 Estimation of the characteristic scale

Let K be the standard Gaussian density function and consider a continuous signal $u(t)$, $t \in \mathbb{R}$, and its convolution smooth $L(t, h^2)$, where h^2 is the variance of Gaussian kernel,

$$L(t, h^2) = \int_{-\infty}^{\infty} u(t - z) \frac{K(z/h)}{h} dz.$$

Smoothing reduces rough small scale features in the signal and hence suppresses derivatives. It follows that visualization of a derivative of a smooth can be difficult because its magnitude diminishes with increasing level of smoothing. As a remedy, Lindeberg (1998) used normalized (dimensionless) coordinates $\xi = t/h$ (cf. Florack et al., 1992) and defined the normalized derivative and second derivative for one dimensional data as

$$\frac{\partial L(t, h^2)}{\partial \xi} = h \frac{\partial L(t, h^2)}{\partial t}, \quad \frac{\partial^2 L(t, h^2)}{\partial^2 \xi} = h^2 \frac{\partial^2 L(t, h^2)}{\partial t^2}.$$

Similarly, the normalized gradient and Laplace operators for two dimensional data can be defined as $h\nabla$ and $h^2\nabla^2$. The smoothing parameter values that maximize a scale-normalized derivative of the Gaussian convolution of the signal reflect the scales at which spatial or temporal variations take place in the data (Lindeberg, 1998). In case of digital images, the normalized Laplace operator can be used for scale determination as the Laplace provides differential entities useful for blob detection (see e.g. Marr, 1982). For time series, the second normalized derivative can be used.

In Section S1 of the supporting information we show that for the Gaussian convolution smoother the scale-derivative is actually proportional to the second normalized derivative and for two dimensional signals to the normalized Laplace operator. Hence, apart from edge effects, an approximate correspondence between the scale-derivative and the second normalized derivative and the Laplace exists also for Gaussian kernel smoothers such as the Nadaraya-Watson (NW from now on) smoother. On the other hand, Silverman (1984) showed that a spline smoothers approximate kernel smoothing with a roughly Gaussian type kernel and hence also they can be thought to resemble Gaussian convolution. The Gaussian convolution smoothing parameter h and the spline smoothing parameter λ are related through an approximate equation that, in the case of an equispaced grid with grid length d and a smoothing spline that uses a second derivative roughness penalty, is of the form $h \propto d^{1/4} \lambda^{1/4}$ so that $\lambda \propto d^{-1} h^4$ (see Silverman, 1984; Green & Silverman, 1993). Reasoning as in Section S1 of the supporting information, the exponent of λ can be seen to affect the connection between the scale-derivative and the scale-normalized derivative only through a multiplicative constant. This makes the connection between the scale-derivative and the second normalized derivative plausible also for the one dimensional spline smoothers. Further, since the correspondence between the one dimensional spline and kernel smoothers extends to two dimensions

(Furrer & Nychka, 2007), such a connection can be assumed to hold also for images. The above reasoning motivates our use of the scale-derivative for both NW and roughness penalty smoothers in defining the characteristic scales in one and two dimensional data.

While the method we propose for discovering the characteristic scales in hierarchical data bears resemblance to the approach often taken in computer vision literature, there are two important differences. First, instead of analyzing the observed signal \mathbf{x} as such, we first decompose it into its scale-dependent components \mathbf{z}_i (cf. (1)) and then estimate the characteristic scale of each component. Second, we do not estimate the scale of each individual feature present in a component \mathbf{z}_i since our goal is to suggest a characteristic scale that represents the whole component. Therefore, to define the characteristic scale for the whole component, we maximize the norm of its scale-derivative, defining the characteristic scale of \mathbf{z}_i as the value of λ that maximizes the function $\lambda \mapsto \|\mathbf{D}_\lambda \mathbf{z}_i\|$.

4 From characteristic scale to characteristic size

In the previous section, we described how to use the scale-derivative to find the characteristic scale of a signal. Here we establish the relationship between the characteristic scale and characteristic feature size. We consider first characteristic periods and then more general features.

4.1 Characteristic period

One and two dimensional sine waves with different wavelengths can be used to estimate the relationship between the characteristic scale and the period of the signal. For each sinusoidal wave, the characteristic scale is first estimated as the maximum of the scale-derivative norm and then the correspondence between the scale and the period is inferred using regression. See Section S2 of the supporting information for further details.

For the NW smoother, the scale and the period turn out to have a linear relationship (See Table 1 and Figure 1). When the scale-normalized second derivative is used with a Gaussian convolution smoother, the theoretical relationship between the scale h and the period p of an one dimensional sine wave is $p = (2\pi/\sqrt{2})h \approx 4.44h$ (Lindeberg, 1998). Considering the results reported in Table 1, the relationship inferred with the scale-derivative norm and the NW smoother, both for one and two dimensional sine waves, is close to the theoretical one.

For one dimensional sine waves, the standard spline smoother considered e.g. in Green & Silverman (1993) that uses a second derivative roughness penalty turns out to have an estimated scale-period relationship of the form $p \propto \lambda^{1/4} d^{1/4}$, where $d = t_{i+1} - t_i$. This is as expected since, as noted in Section 3, in the case of equispaced one dimensional data, the equivalent kernel smoothing parameter h satisfies $h \propto \lambda^{1/4} d^{1/4}$ and, on the basis of the results obtained for the NW smoother, there is likely to be at least an approximate linear correspondence between h and p . We however prefer to work with a roughness penalty smoother for which the smoothness of the result depends only on the value of the smoothing parameter and not the resolution of the data and for such a smoother, the inferred relationship is of the form $h \propto \lambda^{1/4}$. This grid length independent smoother is defined in the Appendix A. The estimated scale-period relationship for the one dimensional roughness penalty smoother is shown in Table 1 (See also Figure 1). For the two

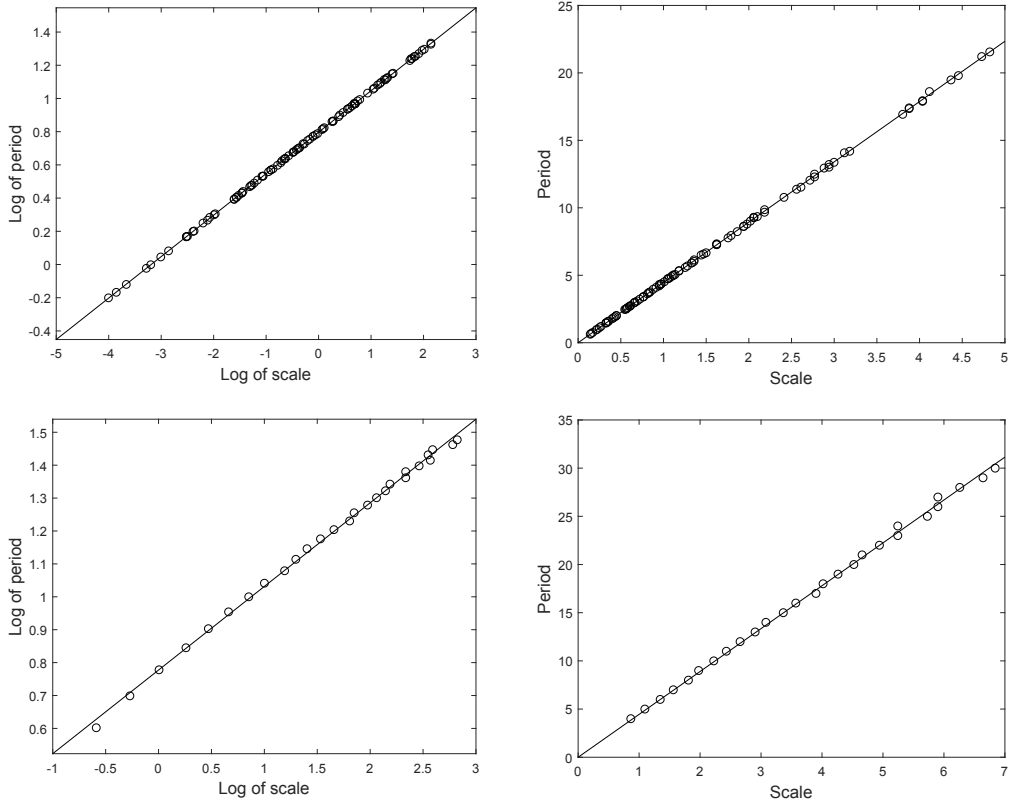


Figure 1: The correspondence between the characteristic scales and the true periods estimated with linear regression. Upper panels: the one dimensional sine waves. Lower panels: the two dimensional sine waves. Left and right: the roughness penalty and the NW smoothers, respectively. For the roughness penalty smoothers, note the logarithmic scale. For the NW smoothers the intercept in the linear regression was fixed at 0. See Section S2 in the supporting information for further information.

dimensional sines, the estimated relationship is close to the one obtained in the one dimensional case (Table 1).

4.2 Characteristic size

The characteristic scale of a two dimensional Gaussian blob can be shown to match its standard deviation (Lindeberg, 1998). This observation can be interpreted to correspond to the so-called matched filter theorem (see Rosenfeld & Kak, 1982) which states that a signal corrupted by additive white noise is best detected by smoothing with a filter whose shape matches that of the noiseless signal (cf. Worsley et al., 1996). This motivates estimation of the characteristic size directly by using the width of the Gaussian density corresponding to the characteristic scale.

Thus, denoting by K the standard Gaussian, for NW kernel regression we estimated the characteristic size corresponding to characteristic scale h as 2δ where δ is solved from the equation

$$K(\delta/h) = K(0)\kappa, \tag{2}$$

and where $0 < \kappa < 1$ is fixed. The solution of (2) is $\delta = \sqrt{-2\log(\kappa)}h$ and in our experiments the values $\kappa = 0.02, 0.05$ were used (cf. Table 1). This resembles the approach often used in statistical scale space methods where the effective kernel width is shown as a space between a pair of whiskers on a map that summarizes statistical inference (e.g. Chaudhuri & Marron, 1999, 2000).

For a scale-dependent hierarchical component, another plausible definition of a feature is a contiguous interval (or area) where the value of the signal is consistently positive or negative. Since the overall mean of a component is close to zero, it is natural to estimate the size of a feature by taking $\kappa = 0.5$ in (2) (cf. Table 1). This approach corresponds to the concept of “full width at half maximum” (FWHM) often used e.g. in medical imaging to define the width of a feature that has no clear boundaries (e.g. Epstein, 2007).

In the third approach, we mimic the method of Netsch & Peitgen (1999) and use the zero crossings of the scale-derivative of a Gaussian density. Thus, for a characteristic scale h , we solve δ from the equation

$$\frac{\partial K(\delta/h)/h^r}{\partial \log h} = 0, \tag{3}$$

giving $\delta = 2^{(r-1)/2}h$, where r is the dimension of the data. The characteristic feature size is then $2\delta = 2^{(r+1)/2}h$. The resulting scale-size relationships are shown in the last column of Table 1.

In the case of other smoothing methods, such as the roughness penalty smoother considered here, the width of the smoother can be approximated using its impulse response. In the one dimensional case, consider the discrete signal $\mathbf{x} = [x_1, \dots, x_{2n}]'$ observed at times $t_1 < t_2 < \dots < t_{2n}$, where $x_i = 0$ if $i = n$ and $x_i = 0$ otherwise. The impulse response of a smoother \mathbf{S}_λ is then defined as $\mathbf{S}_\lambda \mathbf{x}$ and the size of the smoother can be defined as the length of the longest interval $t_{n+\delta} - t_{n-\delta}$ for which $(\mathbf{S}_\lambda \mathbf{x})_{n+\delta} > \kappa(\mathbf{S}_\lambda \mathbf{x})_n$. When the signal is sparsely sampled in time, we interpolate the smooth on a denser grid to improve the accuracy of the size estimate. Table 1 shows the resulting scale-size relationships derived using linear regression with the characteristic scales found in Section 4.1. Here the approach corresponding to (3) uses the impulse response of the scale-derivative $\mathbf{D}_\lambda \mathbf{x}$. For all methods based on the impulse response, the size is related to scale through $\lambda^{1/4}$ (Table 1). An analogous approach works for image data and produces similar results (Table 1).

In the estimation of the size of a periodic feature, the transformations obtained using scale-period regression (Section 4.1) are of course the most accurate. Comparing the formulas in the first column of Table 1 to the other transformations in that table, in the case of the NW smoother, the FWHM appeared to be closest to the scale-period transformation. On the other hand, in the case of the roughness penalty smoother, in one and two dimensional cases, the smoother kernel thresholds of 0.02 and 0.05 were closest to the scale-period transformation. However, the one based on the scale-derivative does not seem to be competitive, under and over estimating the period for one and two dimensional data, respectively.

		scale-period	$\kappa = 0.02$	$\kappa = 0.05$	$\kappa = 0.5$	scale-derivative
NW	1-D	$4.5h$	$5.6h$	$4.9h$	$2.4h$	$2.0h$
	2-D	$4.5h$	$5.6h$	$4.9h$	$2.4h$	$2.8h$
Roughness penalty	1-D	$6.3\lambda^{0.25}$	$6.3\lambda^{0.25}$	$5.9\lambda^{0.25}$	$2.8\lambda^{0.25}$	$2.5\lambda^{0.25}$
	2-D	$6.0\lambda^{0.25}$	$7.2\lambda^{0.24}$	$6.4\lambda^{0.25}$	$2.4\lambda^{0.26}$	$4.1\lambda^{0.24}$

Table 1: The scale-size transformations. 1D: the length of a feature. 2D: the diameter of a feature.

5 Experiments

5.1 Checkerboard image

Here, we re-analyze the checkerboard test image of Holmström et al. (2011). The composite image consists of Gaussian iid noise, three components of blurred squares of different sizes, and an intensity gradient (first and second row of Figure 2). We used both a thin plate spline type roughness penalty smoother corresponding to (A2) and a Gaussian NW smoother. The scale-derivative norms are shown in Figure 3 and in the roughness penalty and NW analyses they have four and three local minima, respectively. For the roughness penalty smoother, the shape of the scale-derivative norms of the extracted components is unimodal with no bumps in the tails, suggesting that the components are well extracted. On the other hand, for the NW smoother, the scale-derivative norms of the components have somewhat bumpy tails, indicating possible mixing between the individual components.

The scale space multiresolution decompositions are shown in the third and fourth rows of Figure 2. For both smoothers, the first component contains the noise and the second and third correspond to the two smallest scale true image components. In the case of the roughness penalty smoother, the largest squares and the gradient have been divided into two components. As the scale-derivative norms of the components suggested, there is some mixing between the components in the case of the NW smoother. The smallest scale components contains traces of the second component, the second component contains traces of the third and first, and so on. In contrast, the features in the roughness penalty components appear to be better separated into distinct scale categories.

The side lengths of the black and white squares in the true components are 1, 7, 30, and 100 pixels, respectively. Of the techniques discussed in Section 4, the goal of both the FWHM method using $\kappa = 0.5$ and the scale-derivative method based on (3) is to estimate these sizes directly, while the scale-period and blob detection approaches based on $\kappa = 0.05, 0.02$ estimate the distance between the local maxima, which for the different size squares are 2, 14, 60, and 200 pixels, respectively. The results are shown in Table 2. For both smoother types, most of the size estimates appear to be slight underestimates. Still, the size estimates match the extracted squares relatively well (see the red circles showing the FWHM estimates in Figure 2). The slight underestimation is perhaps due to the fact that the extracted components appear to be somewhat smaller than the corresponding truths. The approach based on the scale-derivative filter, that had difficulties in estimating the period of the sine waves, appears to have a better performance here.

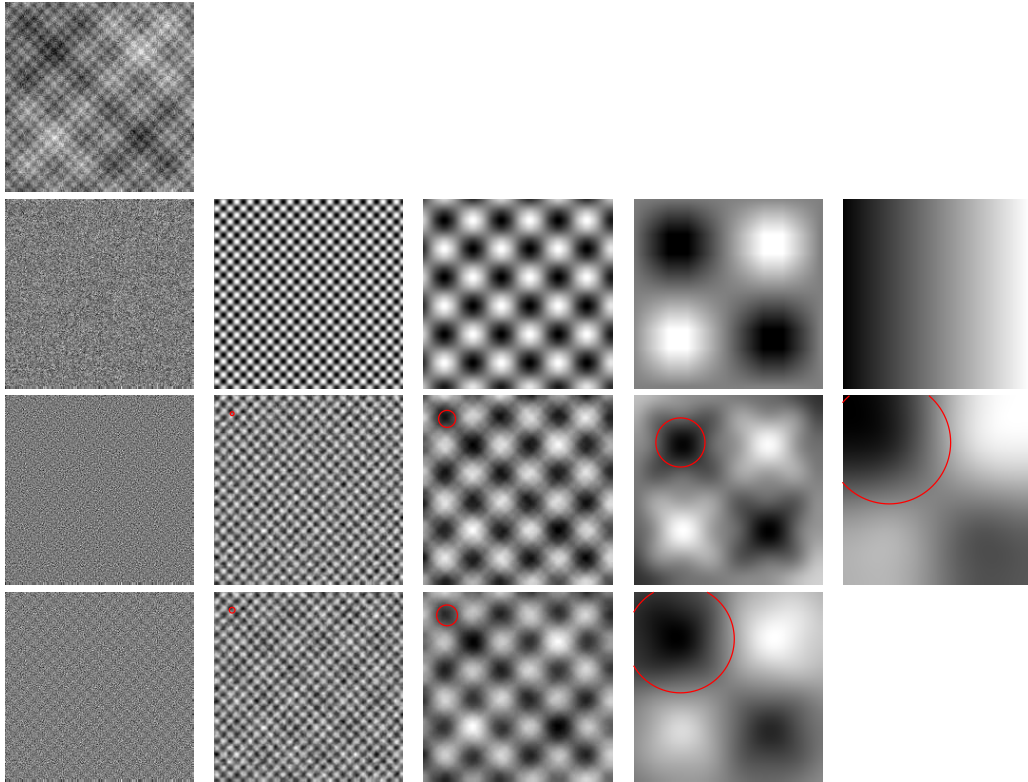


Figure 2: The decompositions of the checkerboard test image. Top row: the test image. Second row: the true components. Third and fourth row: components extracted with roughness penalty and NW smoothers, respectively. Each component image is individually scaled. FWHM ($\kappa = 0.5$) size estimates are shown as red circles.

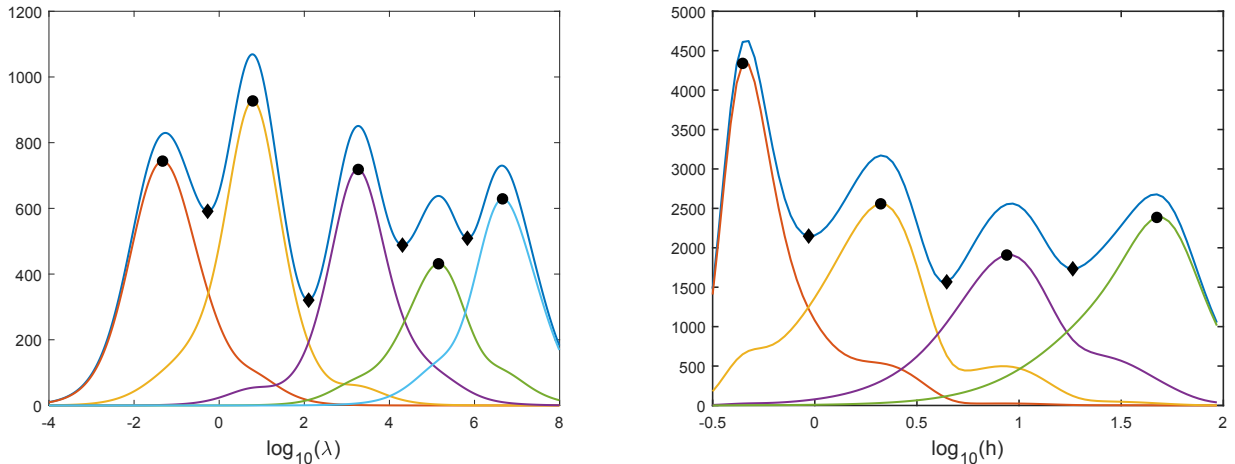


Figure 3: The scale-derivative norms of the checkerboard test image (dark blue line) and those of the extracted components (lines with other colors). Left and right: the roughness penalty and NW smoothers, respectively. The local minima of $\lambda \mapsto \|\mathbf{D}_\lambda \mathbf{x}\|$ are shown with black diamonds and the maxima of $\lambda \mapsto \|\mathbf{D}_\lambda \mathbf{z}_i\|$ are shown with black circles for each \mathbf{z}_i .

	RP	NW	RP	NW	RP	NW	RP	NW
True period	2		14		60		200	
scale-period	2.8	2.0	9.4	9.5	39	39	117, 277	213
$\kappa = 0.02$	3.5	2.5	11	12	44	49	125, 285	264
$\kappa = 0.05$	3.0	2.2	10	10	42	43	124, 295	231
True size	1		7		30		100	
$\kappa = 0.5$	1.1	1.1	3.8	5.0	17	21	52, 129	113
scale-derivative	2.0	1.2	6.3	5.9	25	24	71, 162	132

Table 2: The true and estimated feature sizes of the checkerboard test image, using a roughness penalty smoother (RP) and a Nadaraya-Watson smoother (NW).

5.2 Drought index

As a real-world example, we analyzed the climatic drought index that was originally used to explain the variation in the occurrence of forest fires in Finland. The index was constructed as a linear combination of precipitation and temperature time series, obtained from climate reconstructions based on tree ring chronologies (see Aakala et al., 2018, for details). The drought index time series has an annual resolution and it covers the years 1554-1900.

The time series (posterior mean), the roughness penalty and NW based scale-derivative maps and norms are shown in Figure 4. In the case of the roughness penalty and NW smoothers, the scale-derivative norms have four and two local minima, respectively. For the roughness penalty smoother, the two smallest minima appear to be located between oscillating bands of red and blue in the scale-derivative map,

suggesting scale breaking points between hierarchical levels. However, the largest local minimum appears to be located on top of an oscillating band, and is therefore omitted from the analysis. In the NW analysis, both minima are well located between oscillating bands, but the scale-derivative map suggests that an additional scale break is needed at around $10^{0.5}$. Both scale-derivative maps indicate that the smallest scale component could be further decomposed into two components, but that the period and amplitude of the additional component might change in time. In the roughness penalty analysis, the extra breaking point is located at the dent in scale-derivative norm at around $10^{0.64}$ and in the NW analysis at a slope change in the scale-derivative norm at 10^0 .

The periods of the hierarchical components estimated using the scale-period transformations (Table 1) are 3.9, 11.9, 26.7, 118.3 and 436 years for the roughness penalty smoother and 2.2, 6.8, 24.2, 147 and 475 years for the NW smoother. Apart from the two shortest periods, the estimated periods are relatively similar. The difference in the shortest periods is mostly caused by the apparent nonstationarity in the oscillations at those scales that makes the estimates sensitive to the location of the scale breaking points.

The extracted components are shown in Figure 5. Using Bayesian analysis, we determined the time intervals where a component is credibly positive or negative. The posterior model used is described in detail in Aakala et al. (2018). The posterior probability required for credibility was 0.95 and inference was performed jointly over the whole time interval using the method of Highest Pointwise Probabilities first proposed by Erästö & Holmström (2005). In both analyses, the first and second components are only partially credible. For the third component with a roughly 25-year period, especially those features that appear after the early 1700s tend to be credible. The components with an estimated wavelength of roughly 120 years and 150 years appear credible for most of the time period analyzed. The last components show the deviations from the mean. According to the roughness penalty analysis, it was less dry than on average before about 1670 and after about 1870, and it was drier than on average between about 1720 and about 1850. In the NW analysis, it was drier after about 1700 and less dry prior to about 1680. In the largest scales, the difference between the NW and roughness penalty analyses is mostly due to their different boundary behavior: the NW smooth tends to the mean whereas the roughness penalty smoother to the linear regression as the smoothing level increases.

When seeking explanations for these variations, we compare to the solar cycles and the flux density of galactic cosmic radiation that have shown to occur with principal periodicities of 11-years (Schwabe), 22-years (Hale), 33-year (Bruckner) and 80-100-year (Gleissberg) where the connection between the 22-year cycle is observed to be the strongest one and the 11-year cycle is not always present (See Kasatkina et al., 2007, and the references therein). Also the North Atlantic Oscillation (NAO) can cause periodicity of period 7.3-8 years (Rogers, 1984) (cf. Helama & Lindholm, 2003). The roughness penalty smoother was able to detect the Schwabe, Hale and Gleissberg cycles with reasonably accurate period estimates. In addition, just as documented, also in our results the Hale cycle was more prominent than the Schwabe cycle. The NW smoother appeared to detect the NAO cycle instead of the Schwabe cycle in addition to the Hale and Gleissberg cycles.

For comparison, we applied also multitaper spectral analysis to the data (Thomson, 1982). This method was used also e.g. in Kasatkina et al. (2007). The results are shown in Section S3 of the supporting information. The multitaper method detects the Hale, Schwabe and NAO cycles, as well as oscillations with shorter periods (e.g. 4.8 and 3.7 years). However, it has difficulties in detecting Gleissberg cycles. Furthermore, unlike the scale-derivative map, the multitaper method cannot detect possible nonstationarities in the oscillations.

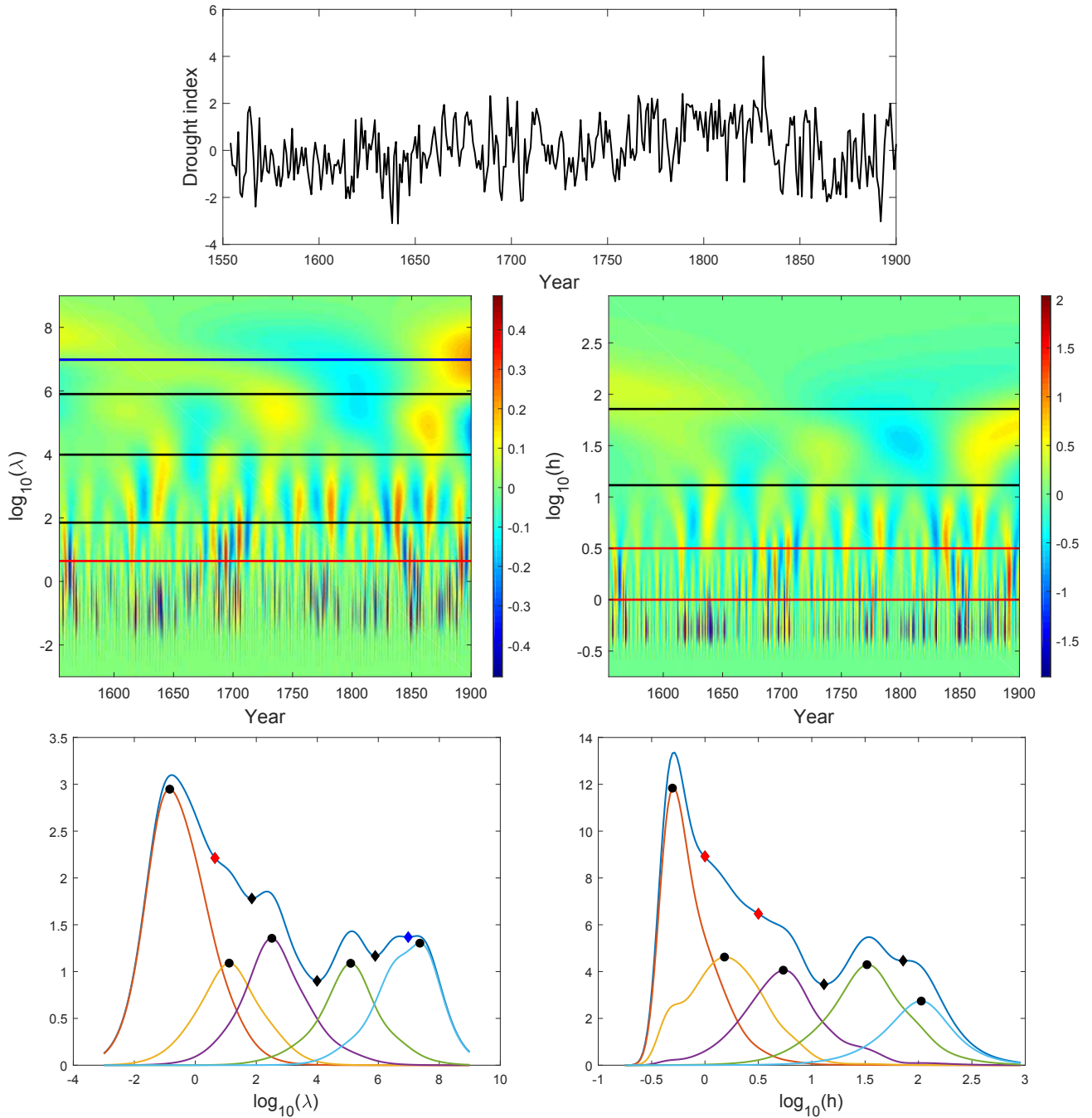


Figure 4: Multiresolution analysis of the drought index. Top row: the data. Second row: the scale-derivative maps. The scale breaks obtained from the local minima of the scale-derivative norms are shown with black horizontal lines and other scale breaks are shown with red lines. Local minimum of the scale-derivative norm not used in the decomposition is marked with a blue line. Bottom row: the scale-derivative norms of the signal (dark blue line) and those of the extracted components (lines with other colours). The local minima are marked with black diamonds. Other scale breaks and local minima not used in the analysis are marked with red and blue diamonds, respectively. The maxima of the scale-derivative norms of the components are marked with black circles. Left and right: the roughness penalty and NW smoothers, respectively.

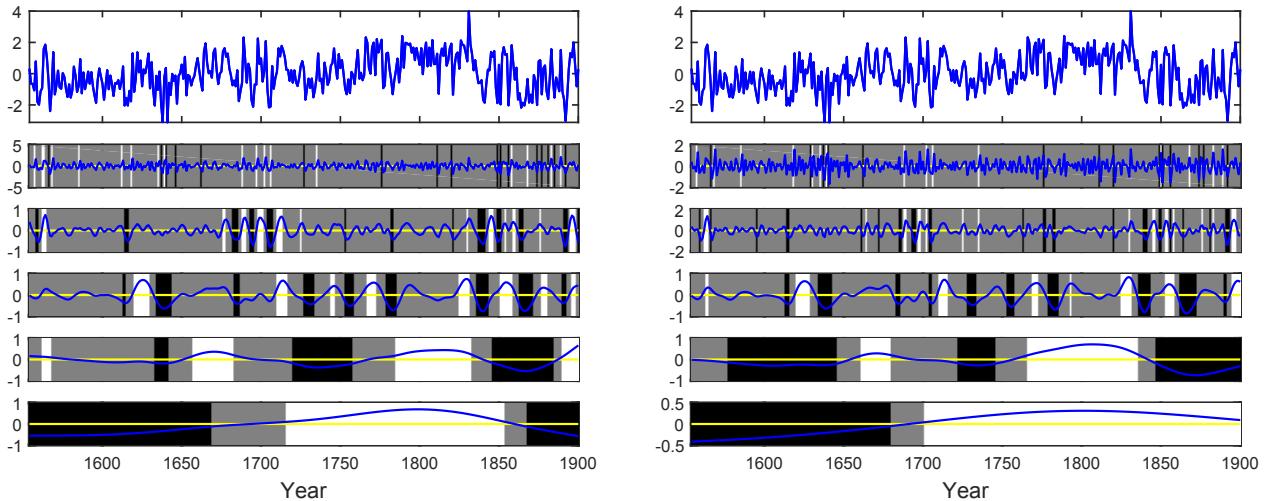


Figure 5: The components extracted from the drought index time series and their credibility analysis. The background for the time intervals where a component is credibly positive, negative or neither is white, black and gray, respectively. Left and right: the roughness penalty and NW smoothers, respectively. See the text for more details.

6 Discussion

The focus of this manuscript was on the estimation of feature sizes in hierarchical signals. In the proposed approach, an observed signal is first decomposed into scale-dependent components that correspond to the hierarchy levels in the signal. In the second step, the characteristic size is estimated using the connection between the scale-derivative and normalized second derivative, ergo by merging concepts from statistical and computer vision oriented scale space literature. Finally, the characteristic scales were transformed into characteristic sizes either by using the scale-period relation of periodic signals or size of the smoother at a characteristic scale.

We used both kernel smoothers and roughness penalty smoothers. While the kernel method was exemplified by the Nadaraya-Watson smoother, other smoothers such as local-linear regression could be used as well. The scale-derivative norms of the extracted components revealed that the components extracted using the NW smoother often contained traces of larger and smaller scales. This was the case in particular for the checkerboard test image. Interestingly, this phenomenon did not appear with the roughness penalty smoothers.

The scale-size relation was estimated both assuming that the data contain a periodicity, and also directly from the width of the smoother at the characteristic scale. We observed that all the proposed approaches produced relatively similar results. Importantly, with all size estimation approaches, the characteristic size was obtained simply by scaling the characteristic scale. In the case of roughness penalty smoothers, the scaling was for the fourth root of the characteristic scale.

There are, of course, many alternative approaches for decomposing a signal into its constituent components, such as wavelets (see e.g. Torrence & Compo, 1998, and the references therein), and singular

spectrum analysis (SSA) (Elsner & Tsonis, 1996). However, compared to such methods, the strength of scale space multiresolution analysis is in its highly intuitive and transparent way of dealing with the hierarchical structures in the data.

In our experiments with the drought index time series, spectral analysis (exemplified here by the multitaper method) appeared more sensitive in the detection of high frequency oscillations, whereas the proposed method might be more sensitive in the detection of low frequency components in the data. The drought index also highlighted the major strength of the scale-derivative approach when compared with spectral methods, namely its ability to discover temporal changes in the periodicities and amplitudes of the hierarchical components in a time series.

We wish also to point out that size estimation using the combination of scale-derivative map and the maxima of the scale-derivative norm is not merely an accessory for the scale space multiresolution analysis, but it can be applied in data analysis more generally. As an example, in SSA analysis a signal is decomposed into components that are not strictly periodic. In that context, Holmström & Launonen (2013) estimated the period of the components using the Lomb-Scargle periodogram, but scale-derivative analysis might provide an alternative approach.

In our examples, we demonstrated the power of the proposed approach to reveal hierarchical structures in the data and to estimate their characteristic sizes. The focus was on time series and while we demonstrated the method with two-dimensional signals only using a simulated example, we believe that scale space multiresolution analysis and feature size estimation would make up a powerful toolkit also for analyzing spatial data. An example of a field where this would be particularly useful is landscape ecology, where the landscapes are assumed to consist of hierarchical structures that form patch mosaics, and where the aim is often to extract the hierarchical variation and inspect the characteristic sizes of the features (Wu, 2013, 1999; Dungan et al., 2002). Our experiments in analyzing canopy cover in forest matrices using the methods proposed here show strong promise also in a real-world two dimensional data setting (Kulha et al., 2018).

7 Acknowledgment

This study was funded by the Academy of Finland under grant no. 276022 and by the Kone Foundation.

A Roughness penalty smoothers

Let $\mathbf{x} = [x_1, \dots, x_n]'$ be a signal such as a time series or a vectorized digital image. In this article we consider the roughness penalty smoother defined as

$$\mathbf{S}_\lambda \mathbf{x} = \underset{\mathbf{u}}{\operatorname{argmin}} \{ \|\mathbf{x} - \mathbf{u}\|^2 + \lambda \mathbf{u}^T \mathbf{Q} \mathbf{u} \} = (\mathbf{I} + \lambda \mathbf{Q})^{-1} \mathbf{x}, \quad (\text{A1})$$

where $\mathbf{x}^T \mathbf{Q} \mathbf{x}$ in some sense measures the roughness of \mathbf{x} and \mathbf{I} is the $n \times n$ identity matrix.

For time series data, we assume that the values x_i are observed at equispaced times t_i with $t_{i+1} - t_i = d$ for all $i = 1, \dots, n - 1$. We take $\mathbf{Q} = \mathbf{C}^T \mathbf{C}$ where $\mathbf{C} \in \mathbb{R}^{(n-2) \times n}$ is the matrix of second forward differences,

that is, $\mathbf{C}\mathbf{u} = \mathbf{w}$ with $w_j = (u_{j+2} - 2u_{j+1} + u_j)/d^2$, $j = 1, \dots, n - 2$. As $\lambda \rightarrow \infty$, $\mathbf{S}_\lambda \mathbf{x}$ converges to the linear regression line.

Next, suppose \mathbf{x} is an image or, more generally, a planar random field on a regular equispaced grid with n grid points. For two grid point locations (pixels) s and t , write $s \sim t$, if they are neighbors. The matrix \mathbf{Q} we use in this case is defined by

$$\mathbf{x}^T \mathbf{Q} \mathbf{x} = \sum_t \left(\sum_{s \sim t} x_s - 4x_t \right)^2, \quad (\text{A2})$$

where the inner summation is over all unordered pairs of neighboring locations. In order to have four neighbors also at a boundary location t , the boundary values of \mathbf{x} are actually extended beyond the original grid. This Neumann boundary condition modifies \mathbf{Q} accordingly. As $\lambda \rightarrow \infty$, $\mathbf{S}_\lambda \mathbf{x}$ converges to the mean of \mathbf{x} . For details, see Holmström et al. (2011).

References

- Aakala, T, Pasanen, L, Helama, S, Vakkari, V, Drobyshev, I, Seppä, H, Kuuluvainen, T, Stivrins, N, Wallenius, T, Vasander, H & Holmström, L (2018), ‘Multiscale variation in drought controlled historical forest fire activity in the boreal forests of eastern Fennoscandia,’ *Ecological Monographs*, **88**(1), pp. 74–91, doi:10.1002/ecm.1276.
- Chaudhuri, P & Marron, JS (1999), ‘SiZer for exploration of structures in curves,’ *Journal of the American Statistical Association*, **94**(447), pp. 807–823.
- Chaudhuri, P & Marron, JS (2000), ‘Scale space view of curve estimation,’ *The Annals of Statistics*, **28**(2), pp. 408–428.
- Delcourt, HR & Delcourt, PA (1988), ‘Quaternary landscape ecology: Relevant scales in space and time,’ *Landscape Ecology*, **2**(1), pp. 23–44.
- Dengel, S, Aeby, D & Grace, J (2009), ‘A relationship between galactic cosmic radiation and tree rings,’ *New Phytologist*, **184**(3), pp. 545–551, doi:10.1111/j.1469-8137.2009.03026.x.
- Diciotti, S, Lombardo, S, Coppini, G, Grassi, L, Falchini, M & Mascalchi, M (2010), ‘The LoG characteristic scale: A consistent measurement of lung nodule size in CT imaging,’ *IEEE Transactions on Medical Imaging*, **29**(2), pp. 397–409.
- Dungan, JL, Perry, JN, Dale, MRT, Legendre, P, Citron-Pousty, S, Fortin, MJ, Jakomulska, A, Miriti, M & Rosenberg, MS (2002), ‘A balanced view of scale in spatial statistical analysis,’ *Ecography*, **25**(5), pp. 626–640, doi:10.1034/j.1600-0587.2002.250510.x.
- Elsner, JB & Tsonis, AA (1996), *Singular spectrum analysis: A new tool in time series analysis*, New York, Plenum press.
- Epstein, CL (2007), *Introduction to the mathematics of medical imaging*, SIAM.
- Erästö, P & Holmström, L (2005), ‘Bayesian multiscale smoothing for making inferences about features in scatter plots,’ *Journal of Computational and Graphical Statistics*, **14**(3), pp. 569–589.

- Esper, J, Cook, ER & Schweingruber, FH (2002), ‘Low-frequency signals in long tree-ring chronologies for reconstructing past temperature variability,’ *Science*, **295**(5563), pp. 2250–2253, doi:10.1126/science.1066208.
- Fang, K, Chen, D, Ilvonen, L, Frank, D, Pasanen, L, Holmström, L, Zhao, Y, Zhang, P & Seppä, H (2017), ‘Time-varying relationships among oceanic and atmospheric modes: A turning point at around 1940,’ *Quaternary International*, doi:https://doi.org/10.1016/j.quaint.2017.09.005.
- Florack, LM, ter Haar Romeny, BM, Koenderink, JJ & Viergever, MA (1992), ‘Scale and the differential structure of images,’ *Image and Vision Computing*, **10**(6), pp. 376 – 388, doi:https://doi.org/10.1016/0262-8856(92)90024-W.
- Furrer, EM & Nychka, DW (2007), ‘A framework to understand the asymptotic properties of kriging and splines,’ *Journal of the Korean Statistical Society*, **36**(1), pp. 57–76.
- Green, PJ & Silverman, BW (1993), *Nonparametric regression and generalized linear models: A roughness penalty approach*, CRC Press.
- Heersink, DK, Furrer, R & Mooney, MA (2013), ‘Intelligent compaction and quality assurance of roller measurement values utilizing backfitting and multiresolution scale space analysis,’ *arXiv preprint arXiv:1302.4631*.
- Helama, S & Lindholm, M (2003), ‘Droughts and rainfall in south-eastern Finland since AD 874, inferred from scots pine ring-widths,’ *Boreal Environment Research*, **8**(2), pp. 171–183.
- Holmström, L & Launonen, I (2013), ‘Posterior singular spectrum analysis,’ *Statistical Analysis and Data Mining*, **6**(5), pp. 387–402, doi:10.1002/sam.11195.
- Holmström, L & Pasanen, L (2017), ‘Statistical scale space methods,’ *International Statistical Review*, **85**(1), pp. 1–30, doi:10.1111/insr.12155.
- Holmström, L, Pasanen, L, Furrer, R & Sain, SR (2011), ‘Scale space multiresolution analysis of random signals,’ *Computational Statistics & Data Analysis*, **55**(10), pp. 2840 – 2855, doi:10.1016/j.csda.2011.04.011.
- Hurrell, JW & Van Loon, H (1997), ‘Decadal variations in climate associated with the North Atlantic Oscillation,’ *Climatic Change*, **36**(3), pp. 301–326, doi:10.1023/A:1005314315270.
- Kasatkina, E, Shumilov, O & Krapiec, M (2007), ‘On periodicities in long term climatic variations near 68° N, 30° E,’ *Advances in Geosciences*, **13**, pp. 25–29.
- Keim, MJ & Percival, DB (2015), ‘Assessing characteristic scales using wavelets,’ *Journal of the Royal Statistical Society: Series C (Applied Statistics)*, **64**(2), pp. 377–393.
- Klemelä, J (2014), *Multivariate nonparametric regression and visualization : with R and applications to finance*, Wiley series in computational statistics, Wiley-Interscience, Hoboken, New Jersey.
- Kulha, N, Pasanen, L, Holmström, L, De Grandpré, L, Kuuluvainen, T & Aakala, T (2018), ‘At what scales and why does forest structure vary in naturally dynamic boreal forests? An analysis of forest landscapes on two continents,’ Submitted.

- Lehmann, A, Höflich, K, Post, P & Myrberg, K (2017), ‘Pathways of deep cyclones associated with large volume changes (LVCs) and major Baltic inflows (MBIs),’ *Journal of Marine Systems*, **167**, pp. 11 – 18, doi:<https://doi.org/10.1016/j.jmarsys.2016.10.014>.
- Lindeberg, T (1994), *Scale-Space Theory in Computer Vision*, Kluwer Academic Publishers.
- Lindeberg, T (1998), ‘Feature detection with automatic scale selection,’ *International Journal of Computer Vision*, **30**(2), pp. 79–116, doi:10.1023/A:1008045108935.
- Lindeberg, T (2015), ‘Image matching using generalized scale-space interest points,’ *Journal of Mathematical Imaging and Vision*, **52**(1), pp. 3–36, doi:10.1007/s10851-014-0541-0.
- Luo, B, Aujol, JF, Gousseau, Y, Ladjal, S & Maitre, H (2007), ‘Resolution- independent characteristic scale dedicated to satellite images,’ *IEEE Transactions on Image Processing*, **16**(10), pp. 2503–2514, doi:10.1109/TIP.2007.906004.
- Marr, D (1982), *Vision*, W. H. Freeman and Company.
- Mikolajczyk, K & Schmid, C (2004), ‘Scale & affine invariant interest point detectors,’ *International Journal of Computer Vision*, **60**(1), pp. 63–86, doi:10.1023/B:VISI.0000027790.02288.f2.
- Nadaraya, E (1964), ‘On estimating regression,’ *Thory Probab. Appl.*, **10**, pp. 186–190.
- Netsch, T & Peitgen, HO (1999), ‘Scale-space signatures for the detection of clustered microcalcifications in digital mammograms,’ *IEEE Transactions on medical imaging*, **18**(9), pp. 774–786.
- O’Neill, RV (1986), *A hierarchical concept of ecosystems*, no. 23 in Monographs in population biology, Princeton University Press.
- Pasanen, L, Laukkanen-Nevala, P, Launonen, I, Prusov, S, Holmström, L, Niemelä, E & Erkinaro, J (2017), ‘Extraction of sea temperature in the Barents Sea by a scale space multiresolution method – prospects for Atlantic salmon,’ *Journal of Applied Statistics*, **44**(13), pp. 2317–2336, doi: 10.1080/02664763.2016.1252731.
- Pasanen, L, Launonen, I & Holmström, L (2013), ‘A scale space multiresolution method for extraction of time series features,’ *Stat*, **2**(1), pp. 273–291.
- Rehauer, H, Seidel, K & Datcu, M (1999), ‘Characteristic scale detection in remote-sensing data,’ in *Geoscience and Remote Sensing Symposium, 1999. IGARSS ’99 Proceedings. IEEE 1999 International*, vol. 1, pp. 116–118 vol.1, doi:10.1109/IGARSS.1999.773419.
- Rogers, JC (1984), ‘The association between the north atlantic oscillation and the southern oscillation in the northern hemisphere,’ *Monthly Weather Review*, **112**(10), pp. 1999–2015.
- Rosenfeld, A & Kak, AC (1982), *Digital Picture Processing*, Academic Press, Inc., Orlando, FL, USA, 2nd edn.
- Sangireddy, H, Stark, CP & Passalacqua, P (2017), ‘Multiresolution analysis of characteristic length scales with high-resolution topographic data,’ *Journal of Geophysical Research: Earth Surface*, **122**(7), pp. 1296–1324.

- Silverman, BW (1984), ‘Spline smoothing: The equivalent variable kernel method,’ *Ann. Statist.*, **12**(3), pp. 898–916, doi:10.1214/aos/1176346710.
- Sirén, G & Hari, P (1971), ‘Coinciding periodicity in recent tree rings and glacial clay sediments,’ *Rep. Kevo Subartic Res.*, pp. 155 – 157.
- Skøien, JO, Blöschl, G & Western, A (2003), ‘Characteristic space scales and timescales in hydrology,’ *Water Resources Research*, **39**(10).
- Thomson, DJ (1982), ‘Spectrum estimation and harmonic analysis,’ *Proceedings of the IEEE*, **70**(9), pp. 1055–1096, doi:10.1109/PROC.1982.12433.
- Torrence, C & Compo, GP (1998), ‘A practical guide to wavelet analysis,’ *Bulletin of the American Meteorological society*, **79**(1), pp. 61–78.
- Wand, M & Jones, M (1994), *Kernel Smoothing*, Chapman & Hall/CRC Monographs on Statistics & Applied Probability, Taylor & Francis.
- Watson, GS (1964), ‘Smooth regression analysis,’ *The Indian Journal of Statistics Series A*, **26**, pp. 359–372.
- Witkin, AP (1983), ‘Scale-space filtering,’ in *8th International Joint Conference of Artificial Intelligence*, pp. 1019–1022.
- Witkin, AP (1984), ‘Scale-space filtering: A new approach to multi-scale description,’ in *Acoustics, Speech, and Signal Processing, IEEE International Conference on ICASSP '84.*, vol. 9, pp. 150–153, doi:10.1109/ICASSP.1984.1172729.
- Worsley, KJ, Marrett, S, Neelin, P & Evans, AC (1996), ‘Searching scale space for activation in pet images,’ *Human Brain Mapping*, **4**(1), pp. 74–90, doi:10.1002/(SICI)1097-0193(1996)4:1<74::AID-HBM5>3.0.CO;2-M.
- Wu, J (1999), ‘Hierarchy and scaling: Extrapolating information along a scaling ladder,’ *Canadian Journal of Remote Sensing*, **25**(4), pp. 367–380, doi:10.1080/07038992.1999.10874736.
- Wu, J (2013), ‘Hierarchy theory: An overview,’ in Rozzi, R, Pickett, S, Palmer, C, Armesto, JJ & Callicott, JB (eds.), *Linking Ecology and Ethics for a Changing World: Values, Philosophy, and Action*, Springer Netherlands, Dordrecht, pp. 281–301, doi:10.1007/978-94-007-7470-4.

Online Supplement to:
A scale space approach for estimating the characteristic feature sizes in
hierarchical signals

Leena Pasanen^a, Tuomas Aakala^b and Lasse Holmström^a

^a Research Unit of Mathematical Sciences, University of Oulu, Oulu, Finland

^b Department of Forest Sciences, University of Helsinki, Helsinki, Finland

S1 Scale-derivative vs. scale-normalized derivative

The convolution $L(t, h^2)$ of the function $u(t)$, $t \in \mathbb{R}$ and the Gaussian kernel is the solution of the heat equation

$$\frac{\partial L(t, h^2)}{\partial h^2} \propto \frac{\partial^2 L(t, h^2)}{\partial t^2}, \quad (\text{S1})$$

with initial condition $L(t, 0) = u(t)$ (see e.g. Serov, 2017). It follows that

$$\frac{\partial L(t, h^2)}{\partial \log h^2} \propto h^2 \frac{\partial^2 L(t, h^2)}{\partial t^2}. \quad (\text{S2})$$

Note, that equation (S2) holds also if, instead of $\log h^2$, the derivative is taken with respect to $\log h$, since by (S1),

$$\frac{\partial L(t, h^2)}{\partial h} \propto 2h \frac{\partial^2 L(t, h^2)}{\partial t^2}$$

and therefore the exponent of h only adds a multiplicative constant to the equation,

$$\frac{\partial L(t, h^2)}{\partial \log h} \propto 2h^2 \frac{\partial^2 L(t, h^2)}{\partial t^2}.$$

This establishes the connection between the scale-derivative and the second normalized derivative with respect to time.

Similarly, in the two dimensional case, we consider a continuous signal $u(\mathbf{t})$, $\mathbf{t} \in \mathbb{R}^2$ and its smooth $L(\mathbf{t}, h^2)$. Again, using the heat equation,

$$\frac{\partial L(\mathbf{t}, h^2)}{\partial \log h} \propto h^2 \nabla^2 L(\mathbf{t}, h^2),$$

showing the relationship between the scale-derivative and the normalized Laplace of the smooth.

S2 Characteristic period

Here we describe how the scale-period relationship was estimated using one and two dimensional sine waves.

In the one dimensional case, we considered 100 sinusoidal waves of the form

$$\mathbf{x} = \sin(2\pi\mathbf{t}/p),$$

where \mathbf{t} is an equispaced n -point grid of $[0, b]$ and where both b and n were uniformly sampled from the intervals $[1, 250]$ and $[120, 1000]$, respectively. The period p was randomly generated on the condition that the number of complete cycles within each wave was between 8 and 30 and that each wave had at least 15 grid points.

In the two dimensional case, we considered 100×100 images \mathbf{x} of spatial sine waves

$$\mathbf{x}(i, j) = \sin(2\pi i/p) + \sin(2\pi j/p),$$

where $i = 1, \dots, 100$ and $j = 1, \dots, 100$, and $p = 4, \dots, 30$.

For the NW smoother, there appeared to be a linear dependence between the scale h and the true period p (cf. Figure 1 in main article). Since $p = 0$ should correspond to $h = 0$, we used linear regression without an intercept term and concluded that

$$p \approx 4.47h \quad \text{and} \quad p \approx 4.45h,$$

for the one and two dimensional sines, respectively.

For the roughness penalty smoothers ((A1) in the main article), the estimated relationships between the true period p and the smoothing parameter λ were found to be

$$p \approx 6.3\lambda^{0.25} \quad \text{and} \quad p \approx 6.0\lambda^{0.25},$$

for the one and two dimensional sines, respectively.

S3 Analyzing the drought index with the multitaper method

The multitaper method is a nonparametric spectral density estimation technique that aims to reduce both the bias and variance in the estimate (Thomson, 1982). See e.g. the review article Babadi and Brown (2014) for further information. To estimate the spectral density of the drought index time series, we applied multitaper analysis with Slepian tapers as implemented in the Matlab function `pmtm` (MATLAB, 2017). Experimenting with different numbers of tapering functions, the value of 7 seemed to work best, detecting oscillations with periods of 512, 22.3, 12.8 and 7.2 years, as well shorter periods, such as 4.8 and 3.7 years (the left panel of Figure S1). These include the Schwabe, Hale and NAO oscillations but not a periodicity with a roughly century cycle that could be connected to the Gleissberg oscillation (See Section 5.2 in the main article for further information). Reducing the number of tapers to 2 produces a peak in the spectral estimate at 171 years but such a spectral estimate is clearly too noisy for making reliable inferences (Figure S1).

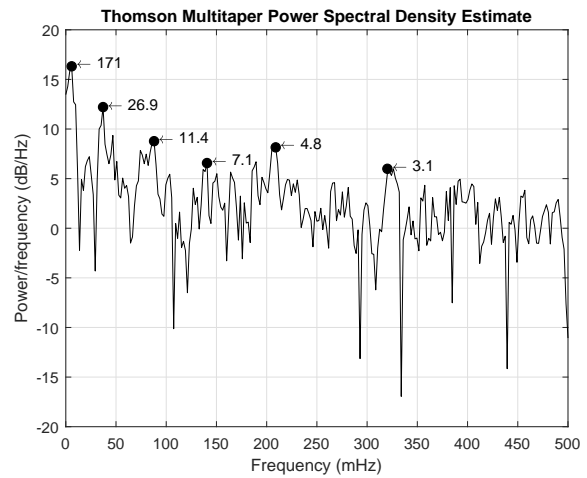
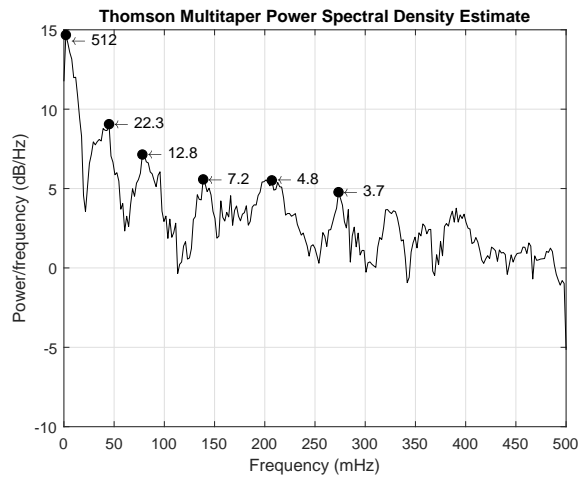


Figure S1. Multitaper spectral analysis of the drought index based on 7 (left) and 2 (right) Slepian tapers. The most prominent peaks in the power spectrum estimates are marked with black dots, together with their associated periods.

References

- Babadi, B. and Brown, E. N. (2014). A review of multitaper spectral analysis. *IEEE Transactions on Biomedical Engineering*, 61(5):1555–1564.
- MATLAB (2017). Signal processing toolbox 7.5. The MathWorks, Natick, MA, USA.
- Serov, V. (2017). *Fourier Series, Fourier Transform and Their Applications to Mathematical Physics*, volume 197 of *Applied mathematical sciences*. Springer, New York.
- Thomson, D. J. (1982). Spectrum estimation and harmonic analysis. *Proceedings of the IEEE*, 70(9):1055–1096.