

# Bayesian multiscale analysis for time series data

Tor Arne Øigård<sup>a,\*</sup>, Håvard Rue<sup>b</sup> Fred Godtlielsen<sup>c</sup>

<sup>a</sup>*Department of Physics and Technology,  
University of Tromsø, NO-9037 Tromsø, Norway*

<sup>b</sup>*Department of Mathematical Sciences, The Norwegian University for Science and  
Technology, NO-7491 Trondheim, Norway*

<sup>c</sup>*Department of Statistics, University of Tromsø, NO-9037 Tromsø, Norway*

---

## Abstract

In this paper we reconsider a recently proposed Bayesian multiscale tool for exploratory analysis of time series data and suggests numerous important improvements. The improvements are both in the model itself, the algorithms to analyse it and how to display the results. The consequence is that exact results can be obtained in real time using only a tiny fraction of the CPU-time previously needed to get approximate results. We present analysis of both real and synthetic data illustrate our new approach.

Multiscale analysis for time series data is an useful tool in applied time series analysis, and with our new model and algorithms, it is also possible to do such analysis in real time.

*Key words:*

SiZer, Gaussian Markov random fields, multiscale analysis, time series analysis, statistical inference, sparse matrices

---

## 1 Introduction

Smoothing for curve estimation in statistics is a useful tool for making inferences and discovering features in data. While investigating the data at a fixed

---

\* Corresponding author. Fax: +47-776-44765

*Email addresses:* [torarne@phys.uit.no](mailto:torarne@phys.uit.no) (Tor Arne Øigård),  
[hrue@math.ntnu.no](mailto:hrue@math.ntnu.no) (Håvard Rue), [fred@stat.uit.no](mailto:fred@stat.uit.no) (Fred Godtlielsen).

level of smoothing often may be satisfactory, it is sometimes worthwhile to consider a whole family of smooths simultaneously. A single smooth may not be able to capture all features of the underlying signal, and important structures of the signal may exist in many different scales. At large scales one might only see the main features of the underlying signal. By decreasing the scale or resolution more details can be seen. At the smallest scales local features appear. Such a multiscale approach is well known in computer vision, see e.g. Lindeberg (1994), and it has been introduced to nonparametric function estimation in the form of the SiZer methodology (SIGNificant ZERo crossings of derivatives) in Chaudhuri and Marron (1999) which combines multiscale ideas with statistical inference. In SiZer, trends of smoothed estimates are inspected using various degree of smoothing. Investigation of trends is a reasonable approach to finding the features in the data since they are the natural indicators of changes in the signal, and they also directly identify the local minima and maxima (Erästö, 2005). The appealing property of SiZer is that the methodology combines the analysis of trends with the idea of multiscale smoothing, and summarises inference about the slopes of the smooths in what is called a SiZer map (Erästö, 2005). This makes analysis of features and trends in time series data accessible also to persons without a degree in statistics since inferences is made by visual inspection of the SiZer map.

Although SiZer has proven very successful in a large set of applications, it is clear that it has weaknesses for complicated signals, see Godtlielsen and Øigård (2005) where a new Bayesian multiscale method was introduced. The new method turned out to be very successful for exploratory analysis of structures in applications where the observed signal contains complicated and rapidly changing structures. The Gaussian approximation needed for SiZer to work was replaced with sampling from the posterior distribution. By doing this one can, in contrast to SiZer, do meaningful inference also at very small scales. In Godtlielsen and Øigård (2005), scale was introduced through a smoothing parameter in the prior distribution of the true underlying curve. Samples from the posterior distribution were drawn using Markov Chain Monte Carlo (MCMC). Although a such approach was shown to be highly successful for signals containing complicated structures, it suffered from some major drawbacks. Generating samples by using MCMC is computationally inefficient. Furthermore, Gaussian approximations were used to find the quantiles for deciding the confidence limits when testing for zero-crossings of the derivative. For large scales these quantiles turns out to be inaccurate, and because of this, the output of the posterior smoothing method gave unreliable results at these scales.

In this paper we reconsider the approach taken by Godtlielsen and Øigård (2005), and suggest numerous important improvements in the model, the algorithms and in how to display the results. The changes are as follows.

**Model** We suggest using a continuous integrated Wiener process as the prior model for the underlying curve, which makes it possible to define and compute the properties of the derivative at any location. This is comparison with Godtliebsen and Øigård (2005), which suggested using the Wiener-process as the prior model, but the derivatives does not exists so finite difference alternatives had to be applied.

**Algorithms** We show how to derive exact algorithms to compute the posterior (Gaussian) marginals for the derivatives using algorithms for band-matrices. This allows for exact testing of zero-crossings of the derivative of the underlying continuous curve. The algorithms only require  $\mathcal{O}(n)$  flops, hence the costs is linear in the length of the time series. The simulation based approach using MCMC taken by Godtliebsen and Øigård (2005), is both costly in comparison with an exact algorithm and does not provide reliable estimates when the smoothing is severe. This is due to slow convergence.

**Display of the results** We compute the *effective sample size* (ESS) exact, and we use this to provide an approach to determine the interesting range of the smoothing level automatically. This is important for applied analysis when such methods are implemented as a black-box for the user, as it allows for a unified display of the results with an easy and clear interpretation. The approach taken in Godtliebsen and Øigård (2005) was only approximative and is not suitable as a basis for a fully automatic procedure.

The paper is organised as follows. Section 2 describes the statistical model for the Bayesian multiscale exploratory analysis. In section 3 we present the computational aspects of the method and describe how to obtain exact solutions of the posterior mean and the posterior marginal variances. In Section 4 the assessment of significant structures is discussed, and in Section 5 we show how to define the effective smoothing window and how to find a precise range of the bandwidths. Section 6 presents results obtained from synthetic and real data sets. A discussion is given in Section 7.

## 2 Statistical model and assumptions

Let  $\mathbf{y} = (y(t_1), \dots, y(t_n))^T$  be  $n$  (conditionally) independent observations, where each  $y(t_i)$  is an observation of the true underlying continuous curve  $\mathbf{x}(t)$  at position  $t_i$ ,  $x(t_i)$ . Assume for simplicity that the positions  $\{t_i\}$  are ordered so that  $t_i < t_j$  for all  $i < j$ , and  $t_1 > 0$ .

Similar to Godtliebsen and Øigård (2005), we assume that  $\{y(t_i)\}$  is conditionally independent Gaussian random variables,

$$y(t_i) | x(\cdot) \sim \mathcal{N}(x(t_i), \sigma_i^2) \tag{1}$$

where  $\sigma_i^2$  are the observational variances. Although the observation variance can depend on  $i$ , the most common cases are that  $\sigma_i^2 = \sigma^2$  for all  $i$ , or that  $\sigma_i^2 = w_i \sigma^2$  for known or (more) easily estimated weights  $\{w_i\}$ . We will use the robust method by Godtliebsen and Øigård (2005) to estimate  $\sigma^2$  in the examples later on.

The underlying true curve  $x(\cdot)$  is assumed to be continuous and smooth in some sense. A natural choice for a prior model, is to consider  $x(t)$  as a realisation of an integrated Wiener process. More specifically,

$$x(t) \mid \{\beta, x(0) = 0\} = \frac{1}{\sqrt{\beta}} \int_0^t (t-h) dW(h) \quad (2)$$

where  $W(h)$  is a standard Wiener-process and  $\beta$  the precision. The integrated Wiener-process has a tight connection to cubic splines, as the posterior expectation of  $x(\cdot)$  conditioned on the observations  $\mathbf{y}$ , equals the cubic smoothing spline as the precision for the initial condition,  $x(0)$ , gets increasingly diffuse Wahba (1978). However, as will be clear in a moment, we will make use of a Bayesian approach to access the posterior marginal variance of our estimate of  $x(\cdot)$  and also its derivative  $x'(\cdot)$ .

Let us now consider the posterior distribution for the underlying true curve at the positions  $\{t_i\}$ ,  $\mathbf{x} = (x(t_1), \dots, x(t_n))^T$ , conditioned on the observations  $\mathbf{y}$  and the initial condition  $x(0) = 0$ . Since  $\mathbf{x} \mid \{x(0) = 0, \beta\}$  is Gaussian, we get

$$p(\mathbf{x} \mid \mathbf{y}, \beta, x(0) = 0) \propto p(\mathbf{x} \mid x(0) = 0) \prod_{i=1}^n p(y(t_i) \mid x(t_i)) \quad (3)$$

where  $y(t_i) \mid x(t_i)$  is Gaussian and given in Eq. 1, and  $\mathbf{x} \mid x(0) = 0$  is multivariate Gaussian with conditional covariance

$$\begin{aligned} \text{Cov}(x(t), x(s) \mid x(0) = 0) &= \frac{1}{\sqrt{\beta}} \int_0^t (s-h)(t-h) dh \\ &= \frac{1}{\sqrt{\beta}} \left( -\frac{1}{6} t^3 + \frac{1}{2} s t^2 \right), \quad 0 < s \leq t \end{aligned} \quad (4)$$

and zero conditional mean as  $E(x(t) \mid x(0) = 0, \beta) = 0$ . The posterior in Eq. (3) is then multivariate Gaussian, and we can for this reason compute directly (Gaussian) marginals for each  $x(t_i)$  for fixed  $\beta$ . These marginals are what is required later on. However, two issues remain. The first is that the covariance matrix defined from Eq. (4) does not allow for fast computations but requires a computational cost of  $\mathcal{O}(n^3)$ . This is a serious drawback for our multiscale problem. The second, is that we need to integrate out the boundary condition on  $x(0)$  after placing a diffuse prior on it.

A more efficient computational scheme can however be derived which requires

only  $\mathcal{O}(n)$  flops. The idea is to augment the model with also the derivatives to construct a model with Markov properties Jones (1981); Wecker and Ansley (1983). Let  $x'(t)$  be the derivative of the integrated Wiener-process at time  $t$ . It might then be shown, that the augmented model for both  $x(t)$  and  $x'(t)$ , has the property that

$$\begin{pmatrix} x(t_{i+1}) \\ x'(t_{i+1}) \end{pmatrix} \Big| \left\{ \begin{pmatrix} x(s) \\ x'(s) \end{pmatrix}, s \leq t_i \right\}, \beta \sim \mathcal{N} \left( \begin{pmatrix} 1 & \delta_i \\ 0 & 1 \end{pmatrix} \begin{pmatrix} x(t_i) \\ x'(t_i) \end{pmatrix}, \frac{1}{\beta} \begin{pmatrix} \delta_i^3/3 & \delta_i^2/2 \\ \delta_i^2/2 & \delta_i \end{pmatrix} \right) \quad (5)$$

where  $\delta_i = t_{i+1} - t_i$ , see for example Rue and Held (2005, Sec. 3.5). Hence, augmenting with the derivatives at each point makes the process Markov and fast computations are possible to derive using the Kalman-filter/smoothen Jones (1981); Wecker and Ansley (1983); Kohn and Ansley (1987). However, we need to place diffuse initial distributions on  $x(0)$  and  $x'(0)$  to avoid the influence on the initial conditions for the integrated Wiener-process. This makes the  $x(t)$  (and also  $x'(t)$ ) process improper which complicates the algorithms, see for example Kohn and Ansley (1987) for an approach to get around this problem.

We take here a slightly different view of Eq. (5) and consider the augmented vector of length  $2n$

$$\boldsymbol{\eta} = (x(t_1), x'(t_1), \dots, x(t_n), x'(t_n))^T \quad (6)$$

as a member of a more general class of models, namely as a *Gaussian Markov random field* (GMRF), refer to Rue and Held (2005) for a thorough description. A GMRF  $\boldsymbol{\eta}$  is a finite dimensional vector which is Gaussian distributed obeying some Markov or conditional independence properties. These properties are usually given as follows; For some  $i \neq j$ , then  $\eta_i$  and  $\eta_j$  are conditionally independent given the rest. This is expressed as

$$\eta_i \perp \eta_j \mid \boldsymbol{\eta}_{-ij} \quad (7)$$

where  $\boldsymbol{\eta}_{-ij}$  denote all elements of  $\boldsymbol{\eta}$  except  $\eta_i$  and  $\eta_j$ . The consequence of Eq. (7) to hold, is that that  $ij$ 'th element of the precision (inverse covariance) matrix  $\mathbf{Q}$ , is zero. The reverse statement is also true. The Markov properties for the integrated Wiener-process is implicitly defined through Eq. (5), that conditioned on  $x(t_i)$  and  $x'(t_i)$ , then the past and the future is independent.

The precision matrix for the GMRF  $\boldsymbol{\eta}$  (with zero mean) can be derived from Eq. (5). The result is as follows, see Rue and Held (2005, Ch. 3.5) for details. Define the matrices



precision matrix  $\mathbf{H} = \mathbf{Q} + \mathbf{D}$  where

$$\mathbf{D} = \begin{pmatrix} \frac{1}{\sigma_1^2} \mathbf{D}_1 & & & \\ & \frac{1}{\sigma_2^2} \mathbf{D}_2 & & \\ & & \ddots & \\ & & & \frac{1}{\sigma_n^2} \mathbf{D}_n \end{pmatrix}, \quad (10)$$

$\mathbf{D}_i = \begin{pmatrix} 1 & 0 \\ 0 & 0 \end{pmatrix}$ ,  $i = 1, \dots, n$ , and mean  $\boldsymbol{\mu} = \mathbf{H}^{-1} \mathbf{b}$  where

$$\mathbf{b} = \begin{pmatrix} y(t_1)/\sigma_1^2 \\ 0 \\ y(t_2)/\sigma_2^2 \\ 0 \\ \vdots \\ \vdots \\ y(t_n)/\sigma_n^2 \\ 0 \end{pmatrix}. \quad (11)$$

Note that the Markov properties remains unchanged when conditioning on the observed data so  $\mathbf{H}$  is still a band-matrix with band-width  $b_w = 3$ . This is important for the computational aspects discussed next.

### 3 Computational aspects

We will now discuss the computational aspects of the tasks we need to solve producing the output of the exploratory analysis in Section 6. The properties we need to compute, are

- the posterior mean, and
- the posterior marginal variance for  $\boldsymbol{\eta}$ .

We will take advantage of band-structure of the precision matrix of  $\boldsymbol{\eta}$  to construct algorithms that cost only  $\mathcal{O}(n)$  flops, i.e. which are linear in the main dimension. We will use numerical methods for band-matrices which we find both easier to explain and to implement, compared to algorithms based on the Kalman-filter with extensions, as discussed by Kohn and Ansley (1987). (Refer to Knorr-Held and Rue (2002, Appendix) for a discussion of these two

approaches.) Using algorithms for band-matrices (or sparse-matrices in general) is a great advantage, as these algorithms applies directly to our problem and there is no extra complications due to the improper prior model for  $\boldsymbol{\eta}$ . Further, they can directly be applied to non-dynamic problems like spatial models and they can easily be extended to deal with linear constraints. For a more detailed account of how numerical methods for sparse matrices can be of advantage for GMRFs, see Rue and Held (2005).

As the dimension of  $\mathbf{H}$  is  $2n \times 2n$ , let  $N = 2n$  in the following to simplify the discussion.

### 3.1 The Cholesky factorisation

The first step is to compute the Cholesky triangle  $\mathbf{L}$  of the posterior precision matrix for  $\boldsymbol{\eta}$ , so that  $\mathbf{H} = \mathbf{L}\mathbf{L}^T$ . Here,  $\mathbf{L}$  is a lower-triangular matrix. The benefit of the banded structure of  $\mathbf{H}$ , is that  $\mathbf{L}$  will be banded as well with same (lower) band-width as  $\mathbf{H}$ , i.e.  $b_w = 3$ . This is a well known property of band-matrices. For a statistical proof of this property, see Rue and Held (2005, Sec. 2.4.1). The computational algorithms can now take advantage of the fact that  $L_{ij} = 0$  whenever  $i - j > b_w$ . For algorithms, Golub and van Loan (1996) can be consulted.

Computing the Cholesky factorisation of a band-matrix is a standard task in numerical linear algebra and most numerical algebra libraries provide such functionality. The computational cost is  $N(b_w^2 + 3b_w)$  and is linear in  $N$ .

### 3.2 Computing the posterior mean

The lower triangular Cholesky triangle  $\mathbf{L}$  is the basis for computing the posterior mean of  $\boldsymbol{\eta}$ . The posterior mean is given as the solution of

$$\mathbf{H}\boldsymbol{\mu} = \mathbf{L}\mathbf{L}^T\boldsymbol{\mu} = \mathbf{b}. \quad (12)$$

To solve Eq. (12), we first compute  $\mathbf{v}$  from  $\mathbf{L}\mathbf{v} = \mathbf{b}$ , and then  $\boldsymbol{\mu}$  from  $\mathbf{L}^T\boldsymbol{\mu} = \mathbf{v}$  to obtain the posterior mean. Note that computing  $\mathbf{v}$  from  $\mathbf{L}\mathbf{v} = \mathbf{b}$  is done using forward-substitution, while computing  $\boldsymbol{\mu}$  is done in using backward-substitution. The computational cost to solve Eq. (12) is  $4Nb_w$ .

### 3.3 Computing the posterior marginal variances

The computation of the marginal variances for  $\boldsymbol{\eta}$  is somewhat more involved than computing the posterior mean. We will therefore describe this procedure in slightly more detail. Our exposition is motivated from the statistical derivation in Rue and Martino (2006, Sec. 2) valid for sparse precision matrices in general, but the discussion simplifies in our case since  $\mathbf{H}$  is a band-matrix with band-width  $b_w = 3$ . Denote by  $\boldsymbol{\Sigma}$  the inverse of  $\mathbf{H}$ , where the posterior marginal variances are found on the diagonal of  $\boldsymbol{\Sigma}$ .

Let us start with the algorithm to sample from a zero mean GMRF with precision matrix  $\mathbf{H}$  and Cholesky triangle  $\mathbf{L}$ . Let  $\mathbf{z}$  be a vector of  $N$  independent standard Gaussian random variables, then the solution of

$$\mathbf{L}^T \boldsymbol{\eta} = \mathbf{z} \quad (13)$$

has the correct distribution Rue (2001), since

$$\text{Cov}(\mathbf{x}) = \mathbf{L}^{-T} \mathbf{I} \mathbf{L}^{-1} = (\mathbf{L} \mathbf{L}^T)^{-1} = \mathbf{H}^{-1}. \quad (14)$$

As  $\mathbf{L}$  is lower triangular, then Eq. (13) equals

$$\eta_i = \frac{z_i}{L_{ii}} - \sum_{k=i+1}^{\min\{i+b_w, N\}} \frac{L_{ki}}{L_{ii}} \eta_k, \quad i = N, \dots, 1. \quad (15)$$

Note that Eq. (15) defines a non-homogeneous auto-regressive process of order  $b_w$  backward in “time”  $i$ , with the correct distribution. Multiply both sides of Eq. (15) with  $\eta_j$  where  $j \geq i$ , and taking expectation, we obtain

$$\Sigma_{ij} = \frac{\delta_{ij}}{L_{ii}^2} - \sum_{k=i+1}^{\min\{i+b_w, N\}} \frac{L_{ki}}{L_{ii}} \Sigma_{kj}, \quad j \geq i, \quad i = N, \dots, 1. \quad (16)$$

Here,  $\delta_{ij}$  is one if  $i = j$  and zero otherwise. A close look at Eq. (16) will convince the reader that we can compute all elements in  $\boldsymbol{\Sigma}$  if we compute them in the correct order; The outer loop runs from  $i = N, \dots, 1$  while the inner loop runs from  $j = N, \dots, i$ . Recall that  $\boldsymbol{\Sigma}$  is symmetric so we only need to compute  $N(N+1)/2$   $\Sigma_{ij}$ 's.

Since we only are interested in the diagonal of  $\boldsymbol{\Sigma}$ , we might be able to reduce the costs by not computing all  $\Sigma_{ij}$ 's, but only those that are needed to compute  $\Sigma_{NN}, \Sigma_{N-1, N-1}, \dots, \Sigma_{11}$ . It turns out, that we only need to compute  $\Sigma_{ij}$  from Eq. (16) for all  $ij$ 's such that  $0 \leq j - i \leq b_w$ , to compute the diagonal of  $\boldsymbol{\Sigma}$ . To see this, imagine that we solve  $\Sigma_{ij}$  for all  $i$  and  $j$  such that  $0 \leq j - i \leq b_w$  in the same order as above; The outer loop runs from  $i = N, \dots, 1$  while the inner loop runs from  $j = \min\{N, i + b_w\}, \dots, i$ . It is clear from Eq. (16), that

to compute  $\Sigma_{ij}$  we need some  $\Sigma_{kj}$ 's where  $i < k \leq \min\{i + b_w, N\}$ . If we in this process *only* make use of, or require *previously* computed  $\Sigma_{kj}$ 's, then this modification is valid and will give us the diagonal of  $\Sigma$ . It is easy to verify that this is indeed true, as

$$i < k \leq i + b_w \quad \text{and} \quad i \leq j \leq i + b_w \quad \implies \quad -b_w \leq k - j \leq b_w. \quad (17)$$

We summarise these findings in the following algorithm to compute the marginal variances:

```

for  $i = N$  to 1 do
  for  $j = \min\{i + b_w, N\}$  to  $i$  do
    Compute  $\Sigma_{ij}$  from Eq. (16)
  endfor
endfor

```

It is implicit in this algorithm that we also have to use that  $\Sigma$  is symmetric. The computational cost is  $Nb_w^2/2$ , and it is linear in  $N$  and of the same order as factorising  $\mathbf{H}$ .

#### 4 Assessment of significant structures

Significant structures in the curves are detected by exploring the derivative of  $\mathbf{x}$  at different scales. For a given scale  $\beta$ , we test, for each point  $t_i$ , whether  $x'(t_i)$  is different from zero. Recall that  $x'(t_i)$  is identical to  $\eta_{2i}$ , see Eq. 6. Our decision rule is developed using  $p(x'(t_i)|\mathbf{y}, \beta)$  and the fact that, for the chosen statistical model in Eq. (1),  $p(x'(t_i)|\mathbf{y}, \beta)$  is Gaussian. If the derivative of  $\mathbf{x}$  at point  $t_i$  is zero, we expect the absolute value of  $E(x'(t_i)|\mathbf{y}, \beta)$ , to be large compared to  $SD(x'(t_i)|\mathbf{y}, \beta)$ . More precisely, for a given value of  $\beta$ , our decision rule is to claim that the derivative at point  $t_i$  is “significantly different” from zero when

$$\left| \frac{E(x'(t_i) | \mathbf{y}, \beta)}{SD(x'(t_i) | \mathbf{y}, \beta)} \right| > q(\alpha/2), \quad (18)$$

where  $q(\alpha/2)$  is the ordinary Gaussian  $\alpha/2$  quantile, and  $\alpha$  is the level of the test.

For the chosen statistical model, the quantities  $E(x'(t_i)|\mathbf{y}, \beta)$  and  $SD(x'(t_i)|\mathbf{y}, \beta)$  can be found exact as described in Section 3.

## 5 Effective Smoothing Window

One useful feature shown in the original SiZer map in Chaudhuri and Marron (1999), is to display (for each scale) the width of the smoother as the horizontal scale containing the *effective sample size* (ESS) number of observations. We will now define ESS for our smoother. Assume for simplicity that the observational variance  $\sigma_i^2$  is constant and equal to  $\sigma^2$  for all  $i$ . We define the ESS through

$$\text{Var}(x(t) \mid \mathbf{y}, \beta) = \frac{\sigma^2}{\text{ESS}(t, \beta)}. \quad (19)$$

Hence, at position  $t$  for fixed  $\beta$ , the smoother has the same effect to the posterior marginal variance, as the average of ESS independent observations. Note that ESS depend in general on the position  $t$ , and since the prior variance of  $x(t)$  does not exist, it does not enter or influence Eq. (19). The ESS can be given a graphical interpretation as the width of the horizontal scale covering ESS observations, i.e. the width of the smoother for each scale.

The limiting behaviour of  $\text{ESS}(t, \beta)$  for our smoother are as follows;

$$\text{ESS}(t, \beta \rightarrow \infty) \leq n, \quad (20)$$

while

$$\text{ESS}(t, \beta \rightarrow 0^+) \geq \begin{cases} 1 & \text{if } t = t_j \text{ for some } j \\ 0 & \text{otherwise.} \end{cases} \quad (21)$$

We will take advantage of this limiting behaviour of  $\text{ESS}(t, \beta)$  to automatically define a range for the smoothing parameter  $\beta$ , as  $[\beta_{\min}, \beta_{\max}]$ . Let  $t_*$  be the middle-most location where there is an observation, and define  $\beta_{\min}$  as the value of  $\beta$  so that  $\text{ESS}(t_*, \beta_{\min}) \approx 2$ , while  $\beta_{\max}$  is the value of  $\beta$  so that  $\text{ESS}(t_*, \beta_{\max}) \approx 0.95n$ .

On the smallest smoothing scale (low value of  $\beta$ ), we smooth roughly over two observations, whereas on the largest smoothing scale (high value of  $\beta$ ) we smooth over near all observations. This is done to ensure that fine structures can be found at small scales whereas for high scales the curve should be more or less linear. A such procedure is a great advantage over the one used in Godtliebsen and Øigård (2005), where  $\beta_{\min}$  and  $\beta_{\max}$  was mainly chosen to be to small and to large, respectively.

## 6 Case studies

We will in this section present results obtained using our approach applied to both synthetic and real data sets. Our two first examples reanalyse examples

in Godtlibsen and Øigård (2005), whereas the last shows an example where our approach fails. A program of the procedure written in C can be obtained from the authors.

### 6.1 Synthetic data

Our first example is a very complicated signal. The number of observed data points is  $n = 200$ , and the noise is independent and identically  $\mathcal{N}(0, 0.25^2)$  distributed. Figure 6.1a show the observations. Note that the data has a complicated structure with rapidly changing observations. Also, the frequency of the oscillations is changed for  $t_i = 0.1$ . Figure 6.1b shows a family of smooths, i.e. the estimate of the underlying curve is viewed at different levels of smoothing. This figure is called a family plot. Each curve in the family plot corresponds to one particular choice of the smoothing parameter  $\beta$ . The observations are shown as circles ( $\circ$ ).

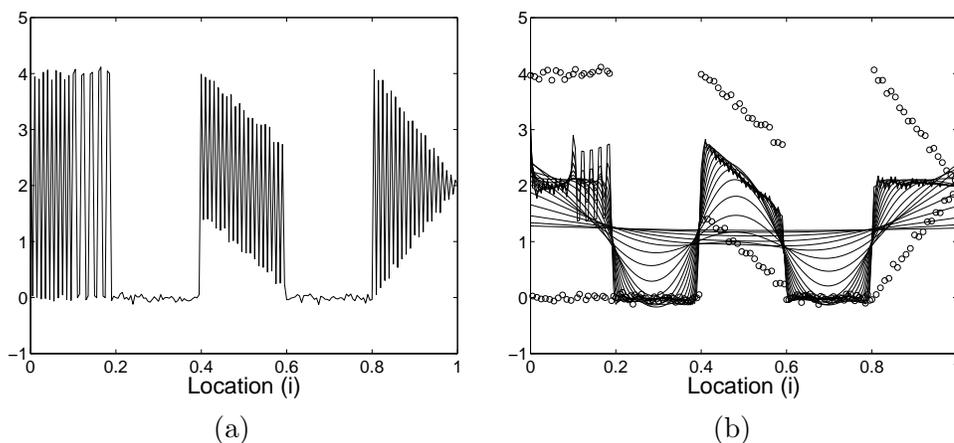


Fig. 1. (a) Observations, and (b) A selection of smoothed estimates for  $x(t_i)$  for the synthetic data set.

The feature map is the output of the Bayesian multiscale analysis method, and is given as a function of location and degree of smoothing. A significantly positive derivative is flagged as black while a significantly negative derivative is flagged as white. The colour light grey is used at locations where the derivative is not found to be significantly different from zero. The feature maps is displayed in Figure 2.

We chose to show the feature map as a function of location and *normalised ESS*. A normalised ESS of 0 corresponds to  $\beta_{\min}$  such that  $\text{ESS}(t_*, \beta_{\min}) \approx 2$ , whereas a normalised ESS of 1 corresponds to  $\beta_{\max}$  such that  $\text{ESS}(t_*, \beta_{\max}) \approx 0.95n$ . This is a natural choice since we chose the bandwidth  $\beta$  implicitly by ESS. A further advantage, is that this approach gives a unified visual

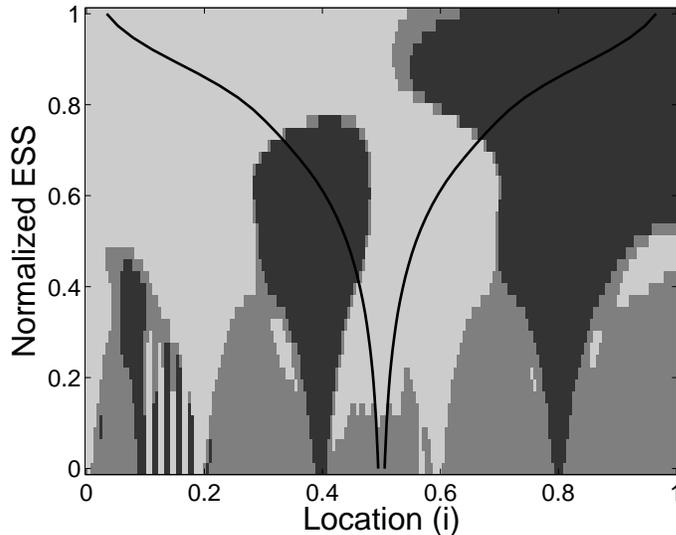


Fig. 2. Feature map obtained using the proposed method on the synthetic data. Here  $\alpha = 0.05$ .

impression of all feature maps in general. The value of the ESS is also shown as the horizontal distance between the two black lines.

Our improved model with its exact algorithms completely solve all the problematic issues reported by Godtliebsen and Øigård (2005).

- There were problems due to inaccurate MCMC-estimates of the quantiles (due to slow convergence) used for testing the derivative, which lead to “noisy” feature maps. This is no longer an issue since we compute exact results. Further, false non-existing structures instead is no longer detected as for a high normalised ESS, since we essentially fit a linear trend to the observed data in that case.
- The MCMC-approach used by Godtliebsen and Øigård (2005) had a further disadvantage that is CPU was substantial. They reported using 270s to compute their feature map similar to the one shown in Figure 2. Our new exact algorithms used only 0.06s to compute the feature map in Figure 2 and is exact. This allow for real-time analysis.

The interpretation of the results in Godtliebsen and Øigård (2005) is similar but not equal to our improved approach. They used finite difference to estimate derivatives, like

$$x'(t_i) \approx \frac{x(t_{i+1}) - x(t_i)}{\delta_i} \quad (22)$$

and they computed their feature maps based on this finite difference is significant or not. A such approach is sensible if the underlying curve is smooth relative to  $\delta_i$  and gives similar or near identical results for moderate to high normalised ESS, whereas the results may differ somewhat for low normalised ESS. We do prefer to use the exact derivatives as our test criterion since it is

both appealing and in agreement with the SiZer methodology.

There may be situations where neighbour difference like in Eq. 22 is preferable to derivatives, to produce feature maps. This include the detection of very fine structures interpreted as change in differences and not change in derivatives. A such approach can be applied to this example. We note that our exact algorithms can also handle this option; We can compute the mean and variance of  $x(t_{i+1}) - x(t_i)$ , since the covariance between  $x(t_{i+1})$  and  $x(t_i)$  is computed (they are neighbours) to obtain the marginal variances.

## 6.2 Ice core data

This example shows an analysis of an ice core data set from Antarctica containing  $n = 1992$  data points. The data points are describing permittivity as a function of depth (and hence as a function of time). Glaciologists use such data sets to gain information about past climate. Understanding changes in past climate can be utilised to give better prediction of future climate. An important aspect with this data set is the interpretation as a function of scale. On the one hand, glaciologists are very interested in a correct detection of the peaks in the data series shown in Figure 3. The observations are shown as

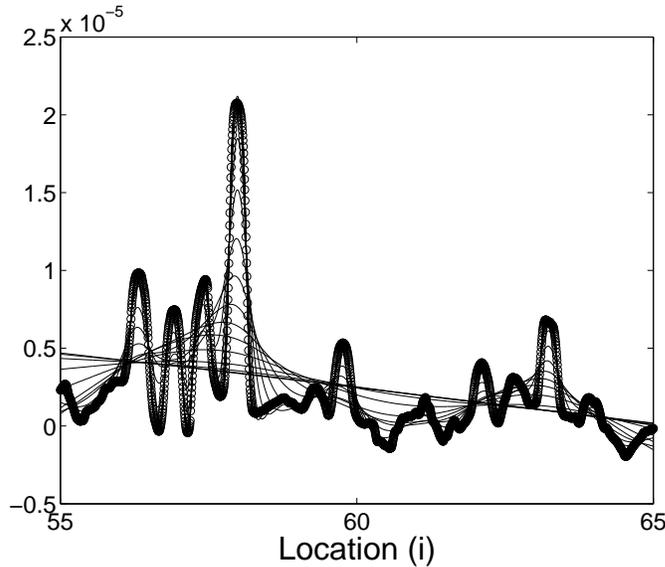


Fig. 3. Family plot: A selection of smoothed estimates of  $x(t_i)$  for the ice core data.

circles ( $\circ$ ), and the lines show the family of smoothed estimates of  $x(t_i)$ . Such peaks correspond to volcanic eruptions. It is of crucial importance to detect these peaks correctly so that the observed permittivity plot as a function of depth can be calibrated with respect to time. The point being, of course, that glaciologists know from other sources when recent volcanic eruptions have

taken place. Calibration of the ice core data can therefore be obtained by utilising this knowledge when they interpret the ice core data sets. On the second hand, glaciologists are interested in longer areas of the ice core data sets where the intensity is high or low. These two situations refer to time periods where the volcanic activity has been high and low, respectively.

Figure 4 displays the feature map obtained using our approach. The results are similar to those presented in Godtlielsen and Øigård (2005), which was affected with the same problems as for the previous example. The CPU time for our exact algorithm, was only 0.43s, which demonstrate that real time analysis is now possible even for such rather long time series as in this example.

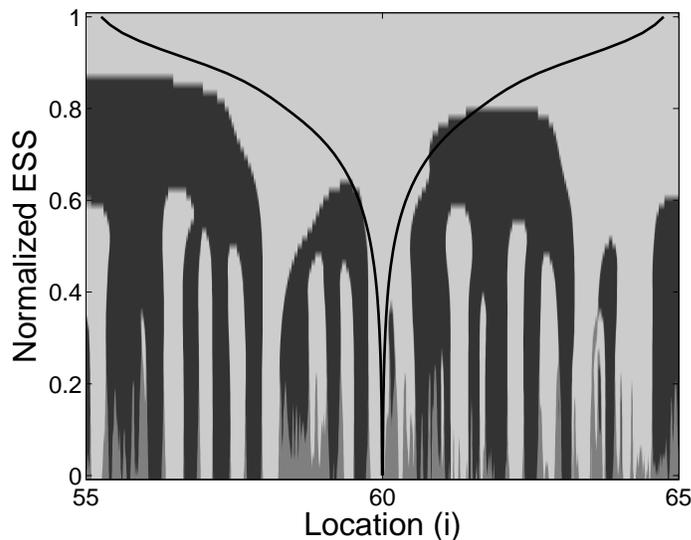


Fig. 4. Feature map obtained using the proposed method on the ice core data. Here  $\alpha = 0.05$ .

### 6.3 Cauchy distributed data

This example is motivated by a comment from on the referees, which encourage us to look at an example where our approach fails. Samples from a standard Cauchy distribution will void the Gaussian assumption of the observational noise and detect false features in the underlying curve. The samples are shown in Figure 5, and we see that the behaviour of the data is highly impulsive and far from Gaussian. The corresponding feature map is shown in Figure 6, and we see that we have many spurious detections due to all the sharp spikes. Ideally the feature map should have been flat, but the noise process has to fat tails compared to the Gaussian distribution.

We can change the likelihood model to be able to analyse also Cauchy data, but the posterior is no longer Gaussian. Despite that a MCMC approach is

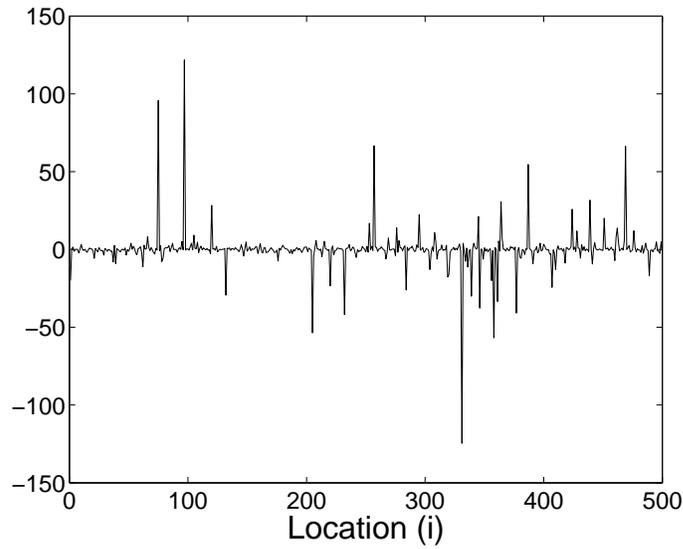


Fig. 5. Samples from a standard Cauchy distribution.

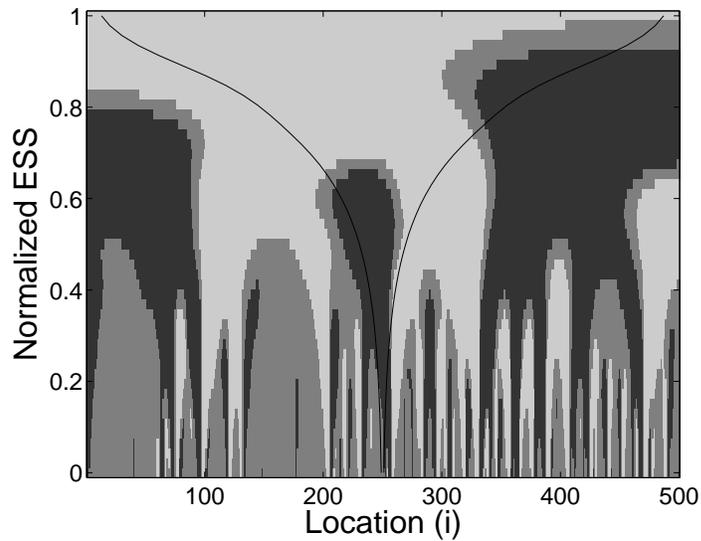


Fig. 6. Feature map obtained using the proposed method on realisations from a standard Cauchy distribution. Here  $\alpha = 0.05$ .

always possible it will have the same problems, if not worse, than reported by Godtliebsen and Øigård (2005).

## 7 Discussion

We have in this paper reconsidered the approach taken by Godtliebsen and Øigård (2005) to do Bayesian multiscale analysis of time series data. The approach is significantly improved in all aspects; The model itself, the algorithms to analyse it and how to present the results. Our improved approach allows for

exact and *real-time* Bayesian multiscale analysis for time series data, which is an useful tool in applied analysis.

A useful extension would be to allow for non-Gaussian observations, like Poisson and Binomial counts. A non-Gaussian likelihood will make the posterior non-Gaussian hence our algorithm cannot be used. Although an MCMC-approach will always be possible, it will likely have the same drawbacks as those reported by Godtliebsen and Øigård (2005). There are however promising new research along the lines of Rue and Martino (2006), which may provide near exact results in real-time.

### *Acknowledgements*

We are grateful to Dr. Lars Karlöf for giving us access to the ice core data.

### **References**

- Chaudhuri, P. and Marron, J. S., 1999. SiZer for exploration of structures in curves. *Journal of the American Statistical Association*, 94 807-823.
- Cressie, N. A. C., 1993. *Statistics for spatial data*, 2nd edition. John Wiley, New York.
- Erästö, P, Studies in trend detection of scatter plots with visualization. Doctor of Philosophy thesis. (Dept. of Mathematics and Statistics, University of Helsinki, Finland, 2005).
- Godtliebsen, F. and Øigård, T. A., 2005. A visual display device for significant features in complicated signals. *Computational Statistics & Data Analysis*, 48 317-343.
- Golub, G. H. and van Loan, C. F., 1996. *Matrix Computations*. Johns Hopkins University Press, Baltimore, 3rd edition.
- Jones, R. H., 1981. Fitting a continuous time autoregression to discrete data. *Applied Time Series Analysis, II* (Tulsa, Okla., 1980), pages 651–680. Academic Press, New York.
- Knorr-Held, L. and Rue, H., 2002. On block updating in Markov random field models for disease mapping. *Scandinavian Journal of Statistics*, 29 597-614.
- Kohn, R. and Ansley, C. F., 1987. A new algorithm for spline smoothing based on smoothing a stochastic process. *SIAM Journal of Scientific and Statistical Computing*, 8 33-48.
- Lindeberg, T., 1994. *Scale-Space theory in computer vision*. Kluwer, Dordrecht.
- Rue, H. and Martino, S., 2006. Approximate Bayesian inference for hierarchical Gaussian Markov random fields models. To appear in *Journal of Statistical Planning and Inference*.

- Rue, H. and Held, L., 2005, Gaussian Markov Random Fields: Theory and Applications. Chapman & Hall, London.
- Rue, H., 2001. Fast Sampling of Gaussian Markov random fields. Journal of the Royal Statistical Society, Series B, 63 325-338.
- Wahba, G., 1978. Improper priors, spline smoothing and the problem of guarding against model errors in regression. Journal of the Royal Statistical Society, Series B, 40 364-372.
- Wecker, W. E. and Ansley, C. F., 1983. The signal extraction approach to non-linear regression and spline smoothing. Journal of the American Statistical Association, 78 81-89.